Analyzing Housing Market Trends with Data and External Factors

Project Team:

Alex Bolduc (*aebolduc@wpi.edu*)

Reagan Brunelle (*rmbrunelle2@wpi.edu*)

Aleksander Proko (*aproko@wpi.edu*)

Yueting Zhu (*yzhu8@wpi.edu*)

Project Advisor

Professor Wilson Wong

Department of Computer Science

Professor Robert Sarnie

Business School

Professor Marcel Y. Blais

Department of Mathematical Sciences

Project Co-Advisor

Professor Michael Elmes

Business School

Abstract

In this project we identified and analyzed various trends within the housing market in the United States to assist an alternative investment firm in identifying opportunities for investment. Using data science technologies–Databricks, Apache Spark, and pandas–we wrote software and queries to parse the various datasets at our disposal and create important interactive visualizations to be used by the firm in the future. We looked at hurricanes and other influential factors as we developed and implemented a forecasting model for median sale price of some of the top housing markets in the United States. The combination of our interactive visualizations and the forecasting model is significant for our sponsors in their pursuit of investments and returns.

Executive Summary

Throughout our project we were assigned with identifying useful trends in the real estate market for our sponsors to further analyze and use as needed. For this task, we were presented with different MLS datasets provided by sponsors from which we extracted and used the data to create useful visualizations in a PowerBI report and mathematical models such as ARIMA and STL. The results from our research and analysis were presented to sponsors on a weekly basis, while updates were reported daily.

We began identifying trends in the real estate market from different hurricane datasets. We looked at market indicators such as listing prices, closing prices, days on market, number of listings sold over list price, and other indicators during different years and locations where hurricanes hit the coast of the United States. After extracting and analyzing the MLS data, we created visualizations for a few different hurricanes and a few different indicators, which were presented to sponsors.

We then moved away from analyzing trends based on hurricanes and started looking into general factors that influence the real estate market, trying to extract potential signals from the datasets and possibly identifying trends in the market that would serve the investment firm for making potential investment decisions based on the research and analysis. While conducting research in specific areas proposed by sponsors, we also looked at the national level, so we could compare the results and examine findings. Furthermore, to diversify the sources of information we used Redfin, Zillow and Realtor.com public datasets to compare our results with. This way we avoided relying solely on the MLS datasets and also noticed any major variations that could have come from mistakes in analysis or defects in data.

Lastly, we used the findings and information collected throughout the project to create an interactive PowerBI report that includes graphs and maps of market indicators among the top 200 metropolitan statistical areas in the country, which allow different departments of the alternative investment firm to review and further analyze the information presented in a platform that accommodates all departments. We used the developed mathematical methods to test for stationarity and work around seasonality of the data, while also forecasting the market. To keep track of our work and assign tasks to team members, we utilized Jira as our agile development tool. Agile Scrum Methodology helped us plan our weekly sprints, while adapting to changes in our project. Overall, we developed new skills and accumulated new knowledge throughout our experience with this project, which we believe will serve us in our future careers.

Acknowledgement

We would like to thank everyone involved in the completion of this project for their guidance and support throughout. Notably, our advisors: Professors Blais, Elmes, Sarnie, and Wong for their assistance in preparing for the project term and their continued support during the project term to ensure the team was making good progress and staying on track. We also want to bring attention to the individuals at the investment firm we worked with for the duration of the project. We were able to take time each day to touch base with our project lead from the firm, who would propose ideas for future project requirements as well as helping us with any issues that we were faced with. He played a vital role in the completion of this project.

Table of Contents

List of Figures

List of Tables

Authorship

Section	Main Author(s)	Main Editor(s)
Cover Page	Alex Bolduc	Aleksander Proko Reagan Brunelle
Abstract		
Executive Summary	Aleksander Proko	Reagan Brunelle
1.0 Introduction	Alex Bolduc	
2.0 Research		
2.1 Company Background	Alex Bolduc Aleksander Proko	Aleksander Proko Alex Bolduc
2.2 Hurricane Effects on Real Estate	Aleksander Proko	Alex Bolduc Reagan Brunelle
2.3 Real Estate Market		
2.4 Business Risks		Yueting Zhu
2.4.1 Risks from Natural Disasters		Alex Bolduc Reagan Brunelle
2.4.2 Risks from Other Factors		
3.0 Methodology		
3.1 Software Development Methodology	Everyone	Aleksander Proko Alex Bolduc
3.2 Mathematics Methodology		
3.2.1 Mann-Whitney U Test	Yueting Zhu	Alex Bolduc Aleksander Proko
3.2.2 Augmented Dickey-Fuller Test		
3.2.3 ARIMA Model		Alex Bolduc Reagan Brunelle
3.2.4 STL Model		
3.2.5 ACF graph		
3.2.6 PACF Graphs		
4.0 Software Development Environment	Yueting Zhu Reagan Brunelle	Aleksander Proko Reagan Brunelle
4.1 Data Processing Infrastructure		
4.2 Project Management & Logistics		
4.3 Data Sources		Yueting Zhu
5.0 Software Requirements		
5.1 Software Requirements Gathering	Alex Bolduc	Reagan Brunelle
5.2 Functional Requirements	Reagan Brunelle	Alex Bolduc
5.3 User Stories and Epics	Everyone	Everyone
6.0 Design		
6.1 High Level Architecture	Alex Bolduc	Reagan Brunelle
7.0 Software Development		

	7.1 Sprint One	Everyone	Everyone
	7.2 Sprint Two		
	7.3 Sprint Three		
	7.4 Sprint Four		
	7.5 Sprint Five		
8.0 Findings			
	8.1 Sprint One and Two Findings	Everyone	Everyone
	8.2 Sprint Three Findings		
	8.3 Sprint Four Findings		
	8.4 Social Impact of Real Estate	Aleksander Proko	Alex Bolduc Reagan Brunelle
	8.4.1 Hurricane Effects in Shaping Demographics		
	8.4.2 The Social and Ethical Perspective of Real Estate		
		8.4.2.1 Rent Control	
	8.4.3 Company's ESG efforts		
	8.5 Business Values and Project Risk Management		Reagan Brunelle
	8.5.1 Risk Mitigation		Alex Bolduc
	8.5.2 Business Risks, Rewards and Values of the Project		
9.0 Assessment		Reagan Brunelle	Alex Bolduc
	9.1 Business Takeaways	Aleksander Proko	
	9.1.1 Risk Culture		
	9.1.2 Additional Risks		
	9.2 Team's Learnings		
10. Future Work		Reagan Brunelle	Yueting Zhu
11. Conclusion		Alex Bolduc	Reagan Brunelle

Table 1: Authorship Table

1. Introduction

Real estate is an incredibly complex market where sometimes even the smallest things can have major effects. As an investment firm that specializes in the real estate industry, our sponsor is very interested in understanding different trends in the market and what may be causing them. We analyzed hurricanes to see if they had any consistent effects on local housing markets, and looked more broadly at housing market trends nationwide to assist our sponsor in identifying areas of investment.

We worked with various real-estate transaction datasets in order to conduct our analyses. Initially, we were given access to a dataset that contained individual property sale data aggregated from many of the local “Multiple Listing Services” (MLS) from around the country. This dataset contained all individual property sales that occurred in the covered areas, which gave us a very fine-grained view of how different housing markets were performing in recent years. We also utilized public datasets from Zillow, Redfin, and Realtor.com. We used these datasets to both verify and compare against the MLS dataset to ensure that its trends aligned with those that these real-estate companies had also identified.

After making our initial analyses, we were tasked with creating an interactive report in PowerBI that an analyst could use as a starting point for research into various housing markets and regions. We identified important housing market metrics and created interactive visuals which show how these metrics changed over time for any given zip code throughout the United States.

Finally we also designed and implemented a mathematical model based on median housing price in the top 200 metropolitan statistical areas (MSAs) in the country. The model

based on the STL model allowed us to make reasonable predictions about the direction of these hot markets in the US.

2. Research

2.1 Company Background

We are working with a New York City based alternative investment firm. When the firm was founded they began with *three* initial investment strategies:

1. *Distressed Debt,* which is “the process of investing capital in the existing debt of a financially distressed company, government, or public entity”. This form of investment allocates resources for good companies that are struggling to tackle debt. This way the firm benefits by either being repaid by the distressed company in the case of bankruptcy, or, in the best case scenario, the distressed company goes through restructuring and the firm becomes an equity or debt holder [13].
2. *Convertible Arbitrage* which is “a form of arbitrage related to convertible bonds, also called convertible notes or convertible debt”. Convertible bonds can be converted into shares of the underlying company, which gives investors the advantage of profiting from the bond’s conversion price and the current price of the company’s stock [14].
3. *Merger Arbitrage* which is “a type of arbitrage related to merging entities, such as two publicly traded businesses”. This form of arbitrage involves the acquiring company and the target, where the acquiring company must purchase the outstanding shares of the target company at a premium price, which enables the potential for profit if the premium purchasing price is lower than the current price of the company’s shares [14].

Over a period of 10 years they doubled their number of strategies by adding *real estate, private equity,* and *leveraged loans* to their portfolio. Starting in the early 2000s the firm expanded globally, opening offices in major cities across the United States, Europe, and Asia. As of 2022 they currently have offices in 14 different cities across the globe and employ over 600 employees.

The alternative investment firm primarily invests in credit and real estate. They work with both corporate and structured credit and real estate opportunities to provide relief to organizations and individuals accordingly. Their strategy is to watch market cycles closely and use machine learning and deep learning techniques to predict flows of alternative markets and employment. This enables the firm to make many safe and predictable bets as long as enough data is gathered. Their approaches to real estate range from "the less intensive – improving existing operations, leasing, and renovations – to more intensive activity such as major repositioning, change of use, and, occasionally, ground-up development". This strategy has been used throughout the company's history to great success. Since 1993, the firm has acquired approximately $25 billion in properties in the US. Since 2005, when they started their real estate business in Asia, they have accumulated more than $10 billion of property.

There are several competitors within the finance space that the investment firm participates in. They are a relatively small company compared to big players like JPMorgan or Blackstone, managing only $50 billion in assets compared to $2.5 trillion for JPMorgan and $951 billion for Blackstone. The firm competes more on the level of smaller asset management firms like Oaktree Capital and Canyon Partners, which manage about $163 billion and $23 billion respectively [6,7,8,9].

The firm's customers range from high net worth individuals looking to increase their wealth to entities like corporations, foundations, and pension funds who want to keep their money in secure investments with relatively high returns.

2.2 Hurricane Effects on Real Estate

An important aspect of real estate investing is analyzing trends within the market. This includes fluctuations in price, listings, insurance rates, and more. Hurricanes present a significant effect on these trends, which is why they present a critical analysis space for corporations involved in real estate.

The most recent hurricane to hit the south coast of the U.S. at the time of writing was Hurricane Ian. Analysts began predicting its potential damage costs before the hurricane actually hit the coast, and the predictions were very similar to the actual costs. Core Logic, a California based corporation that provides information on financials and real estate, predicted $41 to $70 billion in insured and uninsured losses from Hurricane Ian. That prediction was quite accurate after Ian became one of the strongest hurricanes recorded in the U.S., resulting in damages of more than $40 billion [1,2,3].

The many neighborhoods of Florida range from simple $100,000 homes to multi million dollar mansions, and although most homes are insured, many homeowners refuse to purchase home insurance to save on costs, which becomes a huge problem when it comes to recovering from the hurricanes. After Hurricane Ian, home pending sales declined by 32% nationwide and more than 40% in highly affected areas. New listings were also down significantly. Jon Schneyer, when asked why homes would not be insured in Florida, in a conversation for a Core Logic podcast, explained the situation saying: "First of all, it could be uninsurable. It could be a

type of structure or property that insurance companies won't even write policies for." These cases happen when houses are either in a condition that does not qualify them for insurance, or the location of the houses are in a high risk area where insurance companies do not see a potential benefit to insure the house. On special occasions, homeowners can reach a specialty insurer to get a policy, but it is usually too expensive to consider. The other issue that Jon explained is that insurance prices are higher for areas that do not require insurance by law, which is why a lot of homeowners decide to not purchase insurance. A typical case is flood insurance. If the house is in a Special Flood Hazard Area as defined by FEMA, flood insurance is required by law, but if the house is located outside that area, flood insurance is not required. So, homeowners outside the Special Flood Hazard Area avoid paying flood insurance, which, in cases of hurricanes that affect their houses, becomes a burden for them [2].

Most houses that were able to withstand the storm were built exceeding the building code requirements, which raises the dilemma of whether the government should add regulations on the building code among the highly affected areas. Adding new regulations on the building code ensures more safety and less logistical costs and disturbances in cases of hurricanes and other natural disasters, but these regulations would add more costs to homeowners and new developers, since more materials and reinforcements will be needed for homes, resulting in increased prices. These regulations also tend to change the demographics of the communities they are implemented in, as they limit who can live in the area. For example, after Hurricane Michael in 2018, Mexico Beach, Florida imposed changes to the building code for homes, increasing requirements to withstand strong winds from wind speed of 130 mph to 140 mph. The new requirements that this change imposed contributed, in part, to an increase in average home prices of $271,000 in 2019 to $453,000 in 2021 [5]. Price increases like this are why many

inhabitants who leave to seek safety from hurricanes do not come back to their homes. This issue contributes to a slower recovery of the community or even no recovery at all. According to a report by brokerage Redfin, “62% of U.S. residents who plan to buy or sell a home in the next year are hesitant to move to an area with climate risk.” On the other side, investors with great capital express interest in potential investments. The nice weather and the amendments that the coastal line offers still remain attractive for investors, keeping the demand high [5].

2.3 Real Estate Market

The real estate market over the past few years has been gaining significant attention. One reason being the COVID-19 pandemic, but also due to other factors such as changes in interest rates, price fluctuations, and supply and demand. Although there is some instability in the economy and in the housing market, indicators can still be compared year over year. According to Redfin data, the number of new home listings is down about 18% in 2022 compared to the previous two years [26]. Figure 2.3.1 below shows the yearly new listings in the U.S. for the last 3 years.

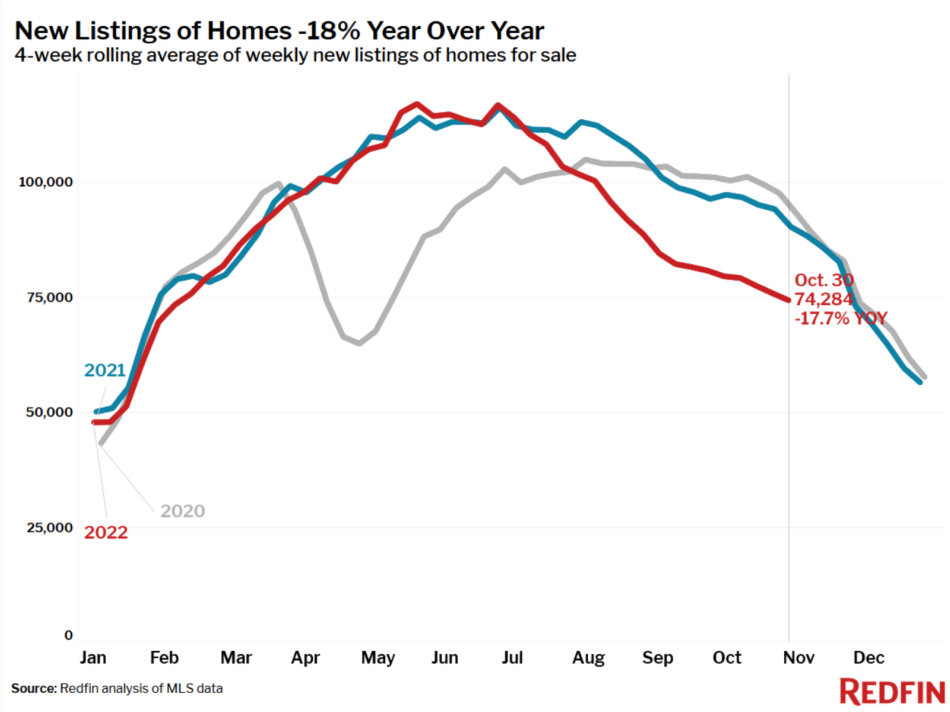

Figure 2.3.1: Redfin new listings of homes year over year

Compared to previous years, 2022 has shown a decrease in new listings as a result of inflation and increased prices. Buyers are hesitant to purchase at a time of high prices and high mortgage rates (48.2% higher mortgage payments compared to previous two years), while sellers are also hesitant to list, because of the low demand in the market [26].

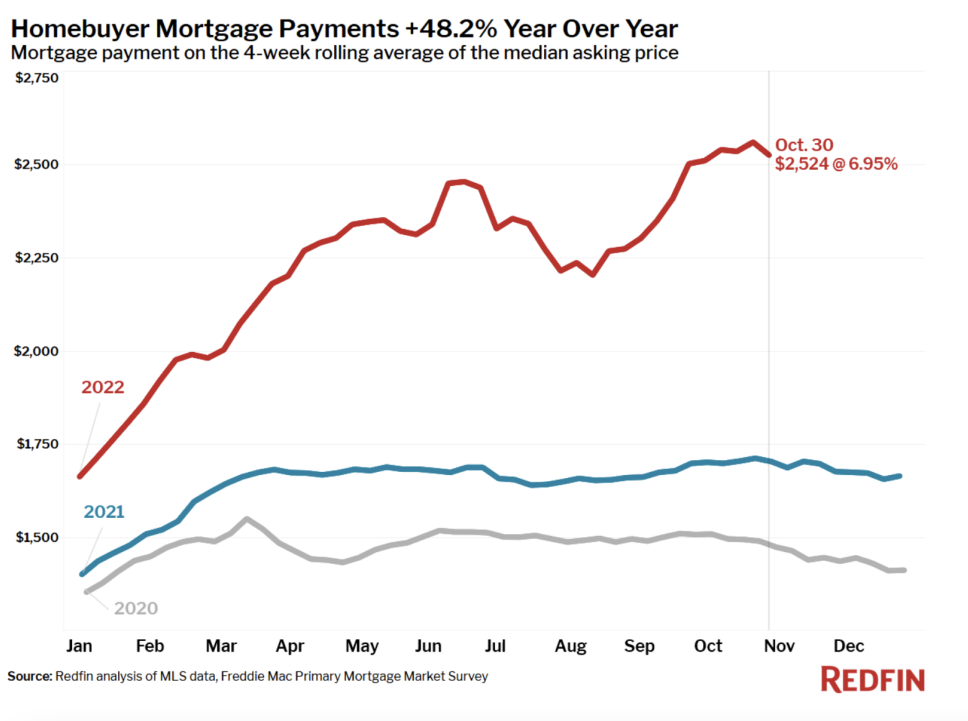

Figure 2.3.2: Redfin homebuyer mortgage payments year over year

This situation has not stopped some particular cities and states from having a boost in the housing sector. Due to factors like low prices, low taxes, remote jobs, or even natural disasters, the demand waves have favored a few places over others. Florida remains one of the top states where real estate is booming, with numerous cities that have resisted the negative effects of natural disasters, and are now top locations for homebuyers and real estate investors. Other states that have also seen booms include Texas, Nevada, Utah, and Colorado [25].

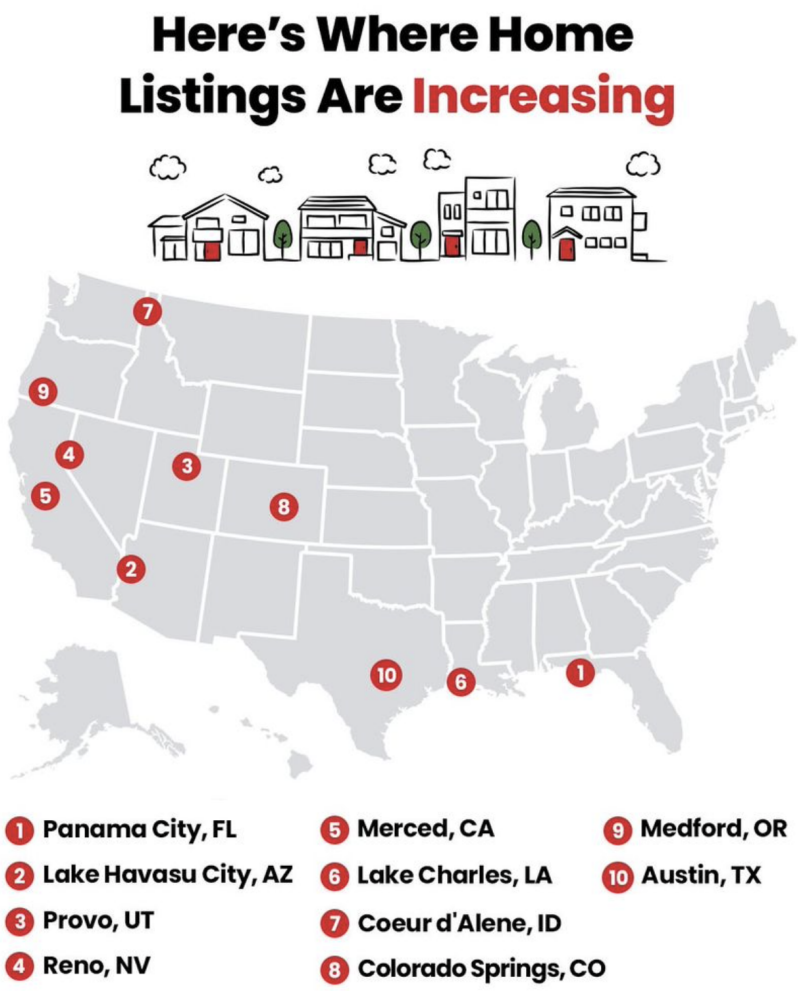

Figure 2.3.3:Realtor.com Top 10 cities with biggest spikes in homes for sale

According to a Realtor.com article, the cities in Figure 2.3.3 above are seeing considerable growth in new listings, which indicates that the real estate market in these areas is incrementally growing. The hurricane situation comes back again in some of these cities, in particular Panama City, which was hit by hurricane Michael in 2018. About four years after the hurricane, the real estate market in the city is still growing due to major investments that took place. Similarly, in Lake Charles, Hurricane Laura in August 2020 and Hurricane Delta in October 2020 caused significant damages in the city, but now real estate is booming because of investors interested in flipping houses and people selling their houses and leaving the area.

Another factor that contributes to the growing real estate market in these areas is the coronavirus pandemic. Employees switching from in person to remote or hybrid work started looking for more spacious houses outside the crowded and expensive states like New York and California, contributing to price increases in the peripheral areas. Reno, Nevada has been one of the "hot spots" for Californians moving from expensive cities like Los Angeles. Sarah Scattini, a real estate agent with Re/Max Premier Properties in Reno said: "You'd stick a sign in the yard, and almost immediately you'd have multiple offers and then be in contract within five days. We were just jamming". In Austin, Texas the situation is different. The high number of listings comes as a result of high property taxes. Because Texas is a no income tax state, a wide proportion of the state's revenue comes from property taxes. In the past 24 months there has been a huge appreciation for homes in Texas, forcing homeowners and investors to either raise rents or sell their houses in order to make profit [25].

Figure 2.3.4: Redfin survey of factors influencing people's decision on home locations

Among other factors influencing the housing sector, affordability, safety, and cost of living are among the top considerations for people, according to a Redfin survey. This explains the relocation of many people from major cities like New York or Los Angeles to more peripheral areas and cities [27].

2.4 Business Risks

2.4.1 Risks from Natural Disasters

There are many risks that arise from natural disasters. For investors, other than property damage and even property losses that could incur significant costs, there are also indirect costs associated with properties involved in natural disasters. These indirect costs include insurance insolvency, market distress, property value depreciation, among others. Insurance insolvency was seen during Hurricane Andrew in the early 1990s, when insurance companies were not able to pay the volume of claims they were receiving, therefore going bankrupt and leaving homeowners with the burden of paying everything themselves. Market distress is often a major consequence of a natural disaster, along with destruction that affects tourism and infrastructure, contributing to a decreased property value. This value can also be affected on a long term basis, if natural disasters consistently affect an area, it becomes a red spot for investors as the risk of potential damages is higher. Therefore, property values begin to depreciate over time as demand decreases in the area.

Another form of risk that comes from natural disasters is the risk of missing out on potential investments. While some areas are constantly affected by hurricanes and the risk of investment is significant, there are areas where a hurricane is less likely to reoccur. These areas can achieve great recovery, which creates great potential for high return on investment. Without proper research, the risk of missing out on these opportunities is considerable.

2.4.2 Risks from Other Factors

The housing sector involves a lot of industries and influences that come together to make it count for a considerable percentage of a country's GDP. There are multiple ways to get involved in this sector, from construction workers all the way to governments or private banks.

Because so many people are involved in the sector, risks are as great as the industry is. From an investor's perspective, risks start from simply a faulty analysis. Market analysis is a crucial part of the business because it is what generates new opportunities and keeps investors up to date on new regulations, events or simply market trends that indicate an opportunity or prevent a possible future risk. Analysis has to be thorough and carefully considered before decisions are made [39].

Choosing the wrong location is another risk. The location of the investment can be the major driver of the price, whether it is rental price or the overall value of the house. Before making an investment decision, the area should be evaluated. For rental properties, usually areas near shopping centers, businesses, schools and other institutions can be rented out for a higher price compared to properties in more peripheral areas. The opposite is often common for single family homes, which can be slightly more expensive in peripheral areas compared to crowded and noisy city centers. Locations also vary from city to city and state to state, which include even more risk factors, because different states might have different laws affecting housing, resulting in potential restrictions.

Market predictability is another risk factor. Generally the real estate market shows continuous historical trends, but it is not always predictable. After the pandemic, the market has become even more unpredictable, as a result of quickly changing prices and rates shifting the demand up and down accordingly. Inflation rates suggest a recession may be imminent, but

according to historical trends a recession happens every 4 years, which has not been the case in recent years. About 14 years after the last 2008 recession, the cycle has not repeated itself, which has caused disruptions in market forecasting. This unpredictable market can be very risky for investors if a certain budget is not allocated or if there is no diversity in investment locations and types [40].

3. Methodology

3.1 Software Development Methodology

Throughout this project we will be doing all software development through the Agile Scrum development methodology. Specifically, Agile focuses on delivering working software quickly, and being able to adapt to changes effectively. Both are crucial requirements for this project, given the fast-paced environment and short timeline [15].

Agile is the product of a meeting between 17 members of the software industry who had separately gone down various roads to improve the industry's approach to software development. There were representatives from various emerging methodologies including "Extreme Programming, SCRUM, DSDM, Adaptive Software Development, Crystal, Feature-Driven Development, Pragmatic Programming, and others." This group of industry minds came together and agreed upon a set of twelve principles titled *Manifesto for Agile Software Development* which is commonly referred to now as *The Agile Manifesto* [16]. The 12 principles laid out by the manifesto focus on efficiency, ownership, and flexibility and set the foundation for what would become the leading methodology in the software development industry.

While Agile is a set of principles for how development should be done it does not prescribe any specific processes to fulfill those principles. That is where scrum fits into the picture. Scrum is a detailed set of roles, meetings, and processes, based on the Agile principles, that a team adheres to to assist in their software development and get things done [18]. Scrum defines three main roles inside of a scrum team: the scrum master, the product owner, and the

developer(s). These roles do not imply any form of hierarchy within the team but instead define different responsibilities for different team members to ensure that things move smoothly [17].

The scrum master is the team member that is responsible for ensuring that the team is correctly and completely following the scrum process. They work with the team to facilitate the various meetings prescribed by the scrum methodology. During these meetings the scrum master often guides the team, trying to avoid and mitigate potential blockers that may arise. As the scrum master the individual is often beholden to the tenets of scrum, but is also flexible to the needs and suggestions of their team to improve the process if necessary [19].

The product owner is responsible for defining and prioritizing product requirements. They are also responsible for streamlining the progress and execution of priorities, while maintaining the business perspective of the project [17].

Developers are team members equipped with the right skills for the technical aspects of the project. Their task is to utilize their programming and software development expertise to build, design and test software applications and systems, as needed to accomplish the goals of the project. Developers should work closely with product owners in order to maintain a smooth progress flow [17].

The scrum methodology separates the development process into units called sprints. Sprints are set periods of time that a team is able to plan out and complete a set amount of work. Similar to a workout at the gym, sprints are sets of a whole workout. Going into your workout, there are expectations set for the total number of sets and the number of repetitions in each set. Essentially, sprints allow the team to set deliverable expectations for a defined period of time and then allows the team to look back and reflect on how well, or how poorly the previous sprint went [18]. In this project we will be using sprints of one week, while in the industry sprints of

two weeks are more common. Throughout the sprint there are four "ceremonies" or meetings that the team will undergo: sprint planning, daily standup, sprint review, and sprint retrospectives.

Sprint planning is the first meeting of the sprint where the team sets expectations for the sprint. The team goes through and identifies product requirements to address in that sprint and creates user stories to represent these requirements. User stories are a way of representing a requirement in the product that is written as if it were from the perspective of an individual who needs that requirement [20]. For example, if the user of an application needed to be able to add two numbers together a user story could be: "as a user I want to be able to input two numbers and receive the sum of those numbers." User stories can also come from the perspective of a developer or project manager such as "as a project manager I want to have metrics available so that I can see where we are losing users from the website." Also during the sprint planning meeting the team will discuss and agree upon the time and work it should take to fulfill the requirements. This is done by assigning a point value to the user stories that indicate a relative timeframe for completion of the story [21].

These user stories are then organized into what is called the product backlog. The product backlog is a sorted list of user stories that indicate the upcoming priorities for the team. Product owners are tasked with curating the team's backlog and sorting it according to priority. The higher priority the story is, the higher it should be in the backlog. Stories in the backlog are then taken according to their priorities into the sprint backlog. Similar to the product backlog, the sprint backlog is a list of user stories that are set to be completed within the coming sprint. During the sprint planning meeting stories are pulled from the top of the product backlog into the

sprint backlog to closely define the scope of the sprint and identify what is and is not to be worked on [58].

Sprint review is the evaluation meeting that occurs at the end of each sprint. This is where a form of internal audit takes place in the team and the progress of the sprint is reviewed in comparison with the expectations set at the beginning. This is an important reflective meeting that yields improvement strategies for future sprint planning [22].

Sprint retrospective is similar to the review. At the end of each sprint the team meets to reflect on the progress and evaluate improvement strategies. The retrospective is a more formal way of evaluation, compared to the review, because in this phase the team utilizes the collected information from the review and makes improvement plans on the technical and communication parts of the agile development [23].

The most regular part of Scrum is the daily stand-up meetings (also referred to as scrum meetings), which are typically held towards the beginning of the workday. These meetings are meant to be efficient and quick, and should not last much more than 20 minutes. The term "stand-up" comes from the practice of everyone standing up for the entire meeting. The idea is that this encourages the developers to keep the meeting short so they can sit down again, however, not all implementations of scrum actually involve standing up during stand-up meetings. These meetings are run by the scrum master and are meant for checking in on the progress of other developers, as well as directing and unblocking them. No work is to be done during stand-up, and it is important for the scrum master to recognize when to stop tangents and push for deeper conversations to be held outside of stand-up [24].

3.2 Mathematics Methodology

3.2.1 Mann-Whitney U Test

The Mann-Whitney U test is a non-parametric statistical test for identical distribution, commonly used to compare two sample means from the same population. The simple idea of this test is to fit statistic T to a U-distribution developed for the test, which outputs a probability between 0 and 1, so that the probability can be compared to a specific threshold. If the probability is less than the threshold, the null hypothesis, that "two populations that datasets are coming from are equal", can be rejected, and if it is larger it cannot be rejected. To expand on this method, the statistic U is in fact the "number of times that y precedes an x", in which x and y are two samples drawn from datasets respectively. If $P(U < \overline{U}) = \alpha$ under the null hypothesis, the test is significant and the null hypothesis of identical distributions of x and y will be rejected [44].

Delving into the history of Mann-Whitney U test, we found that the test method originated from what "Wilcoxon proposed in the *Biometrics Bulletin* in December, 1945." However, the U test added more points of distribution, while the original Wilcoxon's test "gave only a few points of the distribution of his statistics." [45]

In order to compare the different datasets that were given to the team we utilized the Mann-Whitney U test to determine whether trends of the monthly mean and median listing prices suggested by the datasets appeared to be the same. With parameter choice, the team used $\alpha = 0.01$ as the threshold of the test, since the dataset is relatively large with over 200 data points. The team then compared trends of monthly mean and median listing price in three

regions, Arizona, Florida, and the United States. In all three cases, the Mann-Whitney U test was analogously applied.

3.2.2 Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller Test is a statistical testing method for determining whether within a time series with n observations from y_1 to y_n is generated by the model $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$, where α is a constant and ε_t follows $N(0, \sigma^2)$, has ρ equals to 1. The null hypothesis of the test is $H_0: \rho = 1$, meaning the time series is stationary since the rate of change of y_t, Δy_t, is normally distributed:

$$\Delta y_t = \alpha + (\rho - 1) y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \alpha + \varepsilon_t \text{ when } \rho = 1$$

The alternative hypothesis is $H_1: p < 1$, meaning the time series is not stationary. Because the central limit theorem does not apply to the term y_{t-1}, we cannot simply do a T-test on $(\rho - 1)$ using a T distribution. On the other hand, David A. Dickey and Wayne N. Fuller calculated the distribution of the least squares estimator for $(\rho - 1)$ under the null hypothesis, so we can compare the T-statistics to the Dickey-Fuller distribution on a certain threshold [46].

In this project, the team is using the "adfuller" module implemented in Statsmodels which takes a list of y and outputs key statistics and results from the Dickey-Fuller test.

3.2.3 ARIMA Model

The ARIMA (Autoregressive integrated moving average) model is a machine learning model that is used for forecasting time series and also for separating noise from signal in data. The ARIMA regression process can be broken into two parts, the ARMA (Autoregressive Moving Average) model, and an operation on the dataset called differencing or differentiation [50, 52].

The ARMA model is a concatenation of two linear regression models, Autoregressive (AR) and Moving Average (MA) models. The AR model is a regression of the variable against itself. An example of autoregressive model of order p, AR(p), can be written as $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$, in which c is a constant to be predicted and ε_t is the white noise following a normal distribution $N(0, \sigma^2)$. The AR model is a "long term memory model", since it depends on every past time stamp, more or less, to make the prediction. Here is a simple example with an AR(1) model:

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

$$y_{t-1} = c + \phi_1 y_{t-2} + \varepsilon_{t-1}$$

If we substitute y_{t-1} into $y_t = c + \phi_1 y_{t-1}$, we get:

$$y_t = c + \phi_1 (c + \phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t$$

From the above equation, we can tell that y_t depends on y_{t-2}. This model is usually restricted to stationary data, inferring that it does not perform the best for the datasets that have a

trend, seasonality, or both. The partial autocorrelation graph can help identify the best parameters.

Rather than forecasting the future values using regression methods like the AR model, the MA (Moving Average) model relies on the past errors to make predictions. A typical MA(q) model looks like $y_t = c + \varepsilon_t + \theta\varepsilon_{t-1} + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q}$, where ε_t is white noise. This model has a "short term memory" meaning that the model does not account for errors made from the past when going enough further into the past. Below is a simple demonstration using the simplest MA(1) model

$$y_t = c + \varepsilon_t + \theta\varepsilon_{t-1}$$

$$y_{t+1} = c + \varepsilon_{t+1} + \theta\varepsilon_t$$

From the model we notice that the prediction for tomorrow, y_{t+1}, is no longer dependent on the error made yesterday, ε_{t-1}. This model is powerful, because it can reduce the effect of "outlier mistakes" made from the past, which provides us with a more robust model. Similar to the AR model, it is best to make sure that the datasets to be predicted are stationary when implementing the MA model. If we combine these two models, AR and MA, we will get an ARMA model. Below gives a simple example for the ARMA(1, 1) model will be:

$$y_t = c + \phi_1 y_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t$$

The ARIMA model is simply adding a technique called differencing over the ARMA model, with the purpose of performing a transformation of non-stationary time series into

stationary time series [51]. If we define a price at a time t as P_t, then the price at the last time stamp is P_{t-1}. The first difference $\Delta P_t = (1 - B)X_t$, where B is the backwards operator, equals $P_t - P_{t-1}$. Oftentimes, the first or second differences of a non-stationary series is stationary [54]. In this project, the ARIMA model only takes the first difference for all time series, and of course this is customizable and it can be changed anytime in the code.

By combining an ARMA model and differencing, we can turn a non-stationary time series into stationary through differencing, then apply the ARMA model. This model is called an ARIMA model.

3.2.4 STL Model

STL (Seasonal-Trend decomposition) is a filtering procedure for decomposing a time-series into additive components of variation (trend, seasonality and the remainder) by the application of Locally Weighted Scatterplot Smoothing (LOESS) models [53]. The model can be split into several parts: overall trend, seasonal fluctuation, and residuals.

The model uses LOESS smoothing to fit a curve to the time series. Suppose we have a time series with length n. We break the whole series down to smaller chunks called windows, where the length of each window is customizable. A window of length q at point p_i is defined as the set of q points that are closest to p_i on the t-axis. For each window w, we will fit a curve either quadratically or linearly using least squares with weight $v_i(p)$ for each point p in the range of window w. Below is the weight function of each point for each p [55]

$$W(u) = (1 - u^3)^3 \text{ for } 0 \leq u < 1$$

$$\text{or } W(u) = 0 \text{ for } u \geq 1.$$

Let $\lambda_q(p)$ be the qth farthest p_i from p. The neighborhood weight for any p_i is

$$v_i(p) = W\left(\frac{|p_i - p|}{\lambda_q(p)}\right).$$

Besides the weight applied on the location of p on the t-axis, a y-axis weight, ρ_i, is also applied. For example, if p_i have variances $\sigma^2 k_i$ along the y-axis where the k_i are known, then ρ_i might be $1/k_i$ [55]. To incorporate these weights on the y-axis, we can simply multiply the weights $\rho_i v_i(p)$ to form our final weights for each window for each point.

From here we will notice that while the size of the window w gets larger, meaning there are fewer sections that we are fitting, the trend will be smoother and vice versa. Moreover, if substantial curvature is not detected in the dataset, we should use linear regression to fit instead of quadratic regression, as the latter may lead to overfitting [55].

After completing the fitting for all windows, we simply concatenate all fits in all windows to form a general trend. Here is a fit from an outside example, dark line represents linear fit and blue line represents quadratic fit:

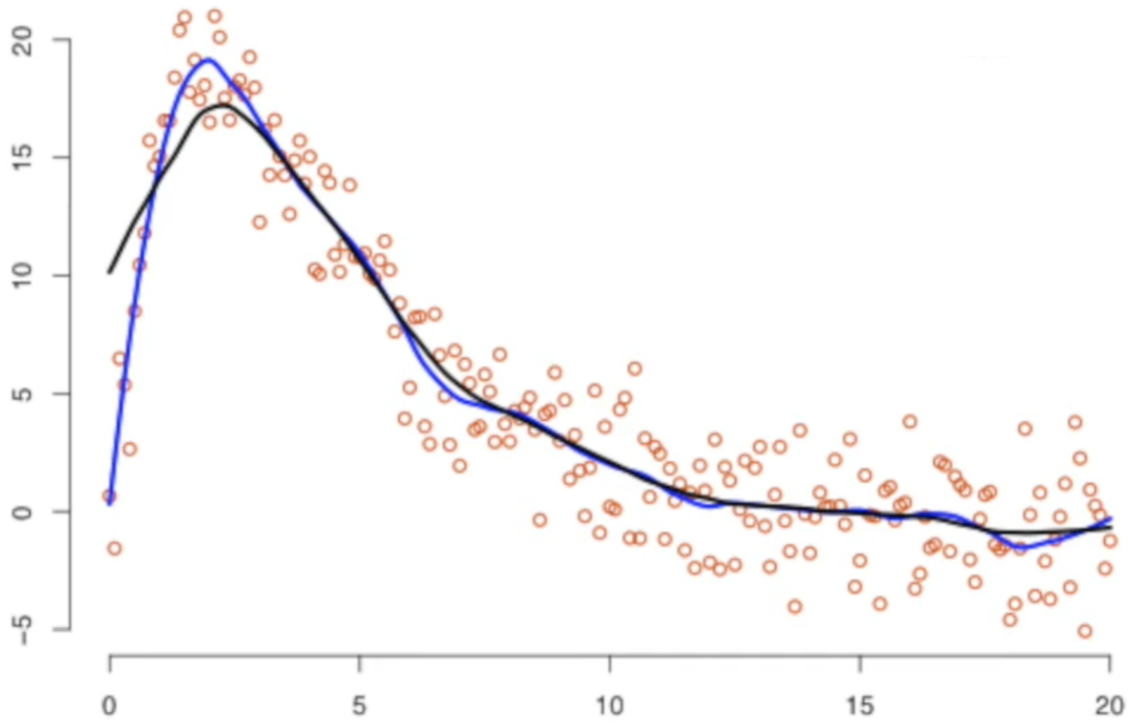

Figure 3.2.4: Quadratic fit and linear fit for a dataset

As we can see in the graph above, the blue line, quadratic fit, fits the data better than the dark line, linear fit. The reason behind the scene is simple, because the dataset is quadratic in this case, quadratic fit will fit better than its linear counterpart.

To account for seasonality, we first have to take away the trend by subtracting the original time series by the fitted trend. Then, cycle-subseries smoothing is performed, and each detrended series is smoothed by LOESS with $q = n_{(s)}$ and $d = 1$. The next step is to apply a low-pass filtering of smoothed cycle-subseries, where the filter consists of a moving average of length $n_{(p)}$, followed by another moving average of length $n_{(p)}$, followed by a moving average of length 3, followed by a loess smoothing with $d = 1$ and $q = n_{(1)}$. In this project, we are using $n_{(p)} = 12$, since our time series is based on months [55].

After the seasonality component is calculated, we can remove both trend and seasonality from the model, and the remainder should be stationary since trend and seasonality should capture the non-stationary part of the time series.

There are several reasons why we are choosing this decomposition model. First, due to inflation and monetary policy, the seasonal component of the series will change as time progresses.he STL can deal with seasonality change over time, and the rate of change is customizable. Second, the smoothness of the trend can be customizable as mentioned in the earlier paragraph. Third, the decomposition is robust to outliers. This is particularly useful because we observed in the dataset that data in particular areas like New York City and Nashville are lacking or even completely missing from our datasets, leading to extremely biased data points when calculating median close price. Because in some areas in the old times, where the MLS database was not fully covered, there were only few sales recorded, these recorded data points are not representative to the whole market since there are not enough sales samples [56].

3.2.5 Autocorrelation Function Graph

The autocorrelation function (ACF) graph is “a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals” [47]. In this project, the team extracted useful information from the ACF not only to determine the parameters of the ARIMA model but more importantly to reassure that the time series data is stationary. This was important because autocorrelation “can affect the validity of inferential statements associated with conventional hypothesis tests and confidence intervals” [48]. This extra step helped the team in performing a more appropriate statistical analysis, and provided a practical visual representation to identify the stationarity of a time series.

Statsmodels is of aid as it is a Python module which provides classes and functions to fit data to different statistical models, which gives estimates of the parameters. Therefore, the major function of the graph in this project is to reassure that the residuals, after STL detrending, are stationary, because if the residuals are not stationary, the ACF graph will show a gradually decreasing trend.

3.2.6 PACF Graphs

PACF (Partial Autocorrelation) is a conditional correlation between two variables under the assumption that we know and take into account the values of some other set of variables [49]. Similar to the ACF graph, if the PACF graph shows an apparent decreasing trend, it is an indicator of non-stationarity in the time series. Consequently, the PACF graph provided another way for the team to diagnose stationarity in a time series.

4. Software Development Environment

The alternative investment firm that we worked with on this project has significant existing infrastructure and tooling that they use to process data. We were given full access to this infrastructure and utilized the same tools and workflows that the firm uses so that our work could be easily used by the team in the future.

4.1 Data Processing Infrastructure

4.1.1 Azure Databricks

Databricks is a service provided by Microsoft Azure that allows teams to easily query and work with large datasets. We used a web-based interface for databricks that allowed the team to write both SQL and Python code in an IPython-style notebook to conduct various analyses on the firm's datasets. These notebooks are useful for reporting a combination of code, text, and visualization. One of the advantages of using Databricks was that it is built on top of Apache Spark which optimizes performance of the data queries and runs them on a cloud-based computer cluster to assist with large queries on large datasets.

Version Number: 11.2

4.1.2 Apache Spark

"Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters." [28] Spark handles data processing with multiple programming languages including SQL and Python. In this project, the team

utilized Spark through its integration with Databricks, and also through the python library PySpark.

Version Number: 3.3.1

4.1.3 Pandas

Pandas is a Python library that is very commonly used in the field of data science and analysis. It serves as a toolbox "providing high-performance, easy-to-use data structures and data analysis tools" [29] for Python. The pandas dataframe, which is the most common data structure provided by pandas, serves as the Python equivalent of a table and allows for easy and efficient manipulation of all aspects of the data in the dataframe. Also, the pandas dataframe is widely supported throughout the Python data science community allowing the team to use the dataframe with other libraries including PySpark. We also utilized its UDF (user defined function) to define our own function for column operations to optimize the runtime of our code.

Version Number: 1.5.1

4.1.4 Power BI

Power BI is an interactive dashboard provided by Microsoft with direct connections to Azure databases. The dashboard created in Power BI is based on all data available in the Azure database, which does not require the team to make extra efforts on connections and accessibility. Tables can be prepared in Azure Databricks based on the specific fields that the team wants to visualize in Power BI. The team members can then connect those tables to the dashboard to monitor real time changes in the tables. Common features that the team will be using in Power BI are line graphs, tables, slicer, and map visuals.

Version Number: 2.109. 1021.0

4.2 Project Management & Logistics

4.2.1 Jira

Jira is an industry standard project management tool used by software development teams that use the agile methodology as we did in this project. We used Jira to organize and keep track of the user stories and epics that were generated throughout the project [33]. It provided a very polished interface that includes many of the important features that are required for agile scrum such as the sprint board, backlog, and support for assigning story points. It also came with extra features that also assisted in our project management. Features like automatic burndown chart generation were very helpful to the team in analyzing how the team performed in each sprint.

Version Number: Jira Software 9.3

4.2.2 Cisco Webex

Cisco Webex is a teleconferencing app that is used widely within the investment firm. We used Webex to conduct our daily standup meetings with our sponsor and additional check-in presentations. Its easy to use video calling features allowed us to conduct effective remote standup meetings and presentations where we could share content with our sponsors which was vital in ensuring the successful completion of the project.

Version Number: 42.10

4.3 Data Sources

4.3.1 Multiple Listing Service

The Multiple listing service (MLS) is “a private offer of cooperation and compensation by listing brokers to other real estate brokers.” Corelogic is a financial service firm that has aggregated data from raw MLS datasets into a more standardized dataset that our sponsors have provided to us. This dataset is a large list of offer records which allow the team to extract market trends and features. Not all MLS datasets are present in Corelogic’s aggregation meaning that some locations across the country are missing data. Additionally, some local-level policies have prevented MLS from gathering data in certain locations. This dataset uses about 125 out of the 600+ total MLS datasets. This includes most states and most major MLS areas, so it provides a good representation of all MLS data. The Corelogic MLS dataset is helpful for analysis and for creating visualizations for the housing market overall [35].

4.3.2 Zillow

The team is using a free websource housing price dataset from company Zillow, “ the most-visited real estate website in the United States” [36]. Zillow is used by most real estate agents and homeowners wanting to list their homes for sale, but also by home buyers who can easily check for new listings and filter them by a variety of options such as size, price, and type of house. Zillow can also be used by analysts who can access the data that Zillow provides from the listings they receive daily. The data can be used to analyze housing market trends. “The Zillow Home Value Index (ZHVI) captures both the level and appreciation of home values across

a wide variety of geographies and housing types", which the team uses to verify different datasets [37].

4.3.3 Redfin

Our sponsors provided us with two datasets from Redfin with housing market prices and trends. One of the datasets contains zip code level data of aggregate measures monthly on a three month rolling average. The other is a county level dataset which also has the same aggregate measures weekly with no rolling average. On both datasets, we analyzed home sales, median days on market, new listings, median sale price, and median list price. We compared the Redfin trends with the MLS dataset for verification purposes, and proceeded to generate 90 day rolling averages at the zipcode level and a monthly rolling average with the weekly data at the county level [42].

4.3.4 Realtor.com

Our sponsors also provided us with a dataset from Realtor.com to verify trends in the MLS dataset. This dataset focuses on monthly real estate trends, including median days on market, median listing price, and new listing count. This simplified the process for us to calculate monthly rolling averages by postal code, to compare with MLS. We split up city and state to allow for querying on those levels as well.

5. Software Requirements

5.1 Software Requirements Gathering

Throughout the project, software requirements were very fluid and changed often over the working period. Our daily standup meetings with our sponsor allowed us to keep up to date with what they were expecting from us and allowed us to plan our sprints according to those requirements. Sometimes this meant having to stray away from our original plan from the sprint planning meeting in order to satisfy the requirements given to us by the sponsor.

5.2 Functional Requirements

The large-scale functional requirements for this project focus on providing data visualizations to our sponsors for analysts to use for predictions and further trend identifications. These visualizations show real estate trends for prices, listing counts, and days-on-market in notable locations with real estate fluctuations in recent years, and they utilize the MLS, Zillow, Redfin, and Realtor datasets.

Meeting these requirements took numerous tasks including preparing dataset tables for PowerBI, generating PowerBI graphs, including slicers for ease of analysis, and verifying the Zillow, Redfin, and Realtor datasets with MLS. Database tables needed to be prepared for each dataset with the desired property fields for analysis along with any cleaning or calculations for additional columns. Once the tables were ready, the additional datasets were compared to MLS on price, listings, and days-on-market to verify that each dataset provided common and reliable results. After verifying the datasets, they were visualized in PowerBI using slicers to display data for specific locations, projects, and properties for each dataset, typically with line graphs. A final

PowerBI report was put together with all the datasets and desired slicer options to provide the analysts with an easy way to utilize our findings.

5.3 User Stories and Epics

Sprint	User Story	Points
	Epic: Final MQP Report	
1	As a reader I want an up to date introduction and abstract so that I can have an understanding of the contents of the MQP report.	1
1	As a project manager I want to identify possible risk factors and possible mitigations so that we can avoid "them" as we continue.	2
2	As a reader and developer, I want to include information on the design chapter of the MQP report.	3
3	As a project manager and investor, I want to identify risks related to investments in the real estate industry.	2
3	As a project manager, I want to identify business risks and risk mitigation methods related to real estate investments, derived from previous analysis.	3
3	As a reader I want to see a section on the software development environment used in the project	1
3	As a reader, I want to be able to see the research and analysis findings presented to sponsors.	2
3	As a reader, I want to see some social impact of the real estate industry, derived from research.	1
3	As an investor and project manager, I want to collect information on past historical events that will yield potential data and information to be linked with current and future trends.	3
3	As a reader I want to have information in the paper on the current state of the real estate market, including indicators and potential trends.	2
2	As a reader, I want to have thorough and formal information on	0

	chapters of the MQP report.	
4	As a reader, I want to read about the design and architecture of the project.	1
4	As a developer and project manager, I want to have a conclusion chapter for the project.	1
4	As an ISC concentration student, I want to include Innovation for Social Change concentration findings on the project.	1
4	As a project manager, I want to have business related takeaways under the assessment chapter of the report.	2
4	As a developer and data science student, I want to include the ARIMA model findings in the project report.	1
4	As a project manager, I want to have a complete risk management section in the project report.	1
4	As a reader I want to be able to identify different tables and figures on the project report.	1
4	As a reader I want to be able to see all the user stories and epics up to date in a separate chapter of the project report.	1
4	As a developer, I want to write about the software applications and environment that is used throughout the project.	2
4	As a data science student, I want to write on the project report about the mathematical method called the U Test.	1
4	As a developer, I want to edit and keep the software requirements section of the paper up to date.	1
4	As a developer and project manager, I want to write an assessment of the project reflecting the work completed throughout the term.	1
4	As a project manager, I want to include the last sprint's findings presented to sponsors, into the project report.	0
4	As a data science student, I want to write a project report about the mathematical method of Autocorrelation/Partial Autocorrelation.	1
4	As a data science student, I want to write on the project report about the mathematical method used in the project, called Augmented Dicky-Fuller Test.	1

4	As a data science student, I want to write on the project report about the mathematical method used in the project, called Time Series.	1
4	As a data science student, I want to write on the project report about the mathematical method used in the project, called STL.	1
4	As a data science student, I want to write on the project report about the mathematical method used in the project, called ARIMA.	1
5	As a sponsor, I want to have a section on the paper mentioning how the remote access was given to students.	1
5	As a business student, I want to write an executive summary of the whole project report in the form of a business report.	2
5	As a sponsor, I want to keep the name of the company private and outside of the project report.	2
5	As a project manager, I want to include the last sprint's findings presented to sponsors, into the project report.	2
5	As a reader, I want to have an up to date intro and abstract which applies to the report.	1
5	As a reader, I want to have the right page numbering on the paper to easily navigate through.	1
5	As the writers of the report, we want to have a bibliography with all the references used throughout the project.	2
5	As a developer, I want to include a description of sprint and product backlog in the appertaining section of the report.	1
5	As writers of the paper, we want to acknowledge everyone who made the project possible, by dedicating a section for them in the paper.	1
5	As writers of the paper, we want to have a conclusion chapter for the project.	1
5	As writers, we want to include a future work section on the paper to represent the work that can be continued in the future.	2
5	As a writer, I want to address the advisors' feedback and comments and make the necessary changes to the paper.	3
5	As writers of the paper, we want to have a thorough report, therefore we want to go over the paper and address any issues or clarifications that need attention.	5

	Epic: Trend Identification	
1	As a reader I would like to have background information on how major hurricanes have impacted local housing markets.	2
2	As an investor I want to have information regarding trends in local housing markets in response to major hurricanes so that I can possibly identify similar trends in the most recent hurricane.	5
1	As a developer I want to identify columns in the MLS dataset that would be useful in identifying trends in the housing market so that we can narrow down the dataset and focus on those columns	3
1	As a developer I want to identify what columns or derived fields will be used as output so that I can structure my queries for those fields	3
1	As a developer I want to identify key import tables that will be useful for my queries in the future.	3
1	As a reader and project manager I want visualizations of the identified housing trends so that they are easier to understand and identify.	2
2	As an investor, I want to find information that can produce potential opportunities for investments in the real estate market after hurricane destruction.	5
2	As a developer, I want to analyze the housing market's average price change over time after hurricanes, to assist in identifying possible trends in the market.	2
3	As a developer I want to create visualizations of Redfin data in PowerBI.	2
3	As a developer and project manager, I want to create a list of cities that present potential trends in the real estate market.	1
3	As a developer I want to develop an ARIMA model for predicting prices like listing or closing prices, based on the data from previous times.	2
3	As a developer I want to compare Realtor.com data with the MLS data in order to see how they line up and identify possible trends.	0
3	As a developer I want to perform a statistical test on the datasets from Zillow and MLS to show how they line up and if they follow the same trend.	2

4	As a developer I want to compare Realtor.com data with the MLS data in order to see how they line up and identify possible trends.	3
	Epic: Tool Familiarization	
1	As a developer I want to familiarize myself with how to interact with the MLS dataset through Databricks so that I can work with it efficiently in the future	1
	Epic: Extracting Signal	
2	As a developer, I want to investigate listing prices in areas or regions with considerable growth fluctuations.	1
2	As a developer, I want to investigate closing costs in areas or regions with considerable growth fluctuations.	1
2	As a developer, I want to investigate days on market for listings in areas or regions with considerable growth fluctuations.	1
2	As a project manager, I want to be able to identify areas where the real estate market has recorded considerable fluctuations over time.	2
2	As a developer and project manager, I want to compare Zillow valuations to MLS factors like listing price and close price to see how they line up.	2
2	As a developer, I want to investigate the number of listings in areas or regions with considerable growth fluctuations.	1
3	As a developer, I want to make a table that has all the analysis that has been done in Databricks for the MLS data on a monthly basis to compare with Zillow data. (For zip codes with sponsor's properties)	3
3	As a developer I want to be able to generate a graph of the year over year growth by quarter of the percent of homes that were sold over listing price in a given zip code.	1
3	As a developer I want to be able to generate a graph which shows the year over year growth of the median "sale price to list price" ratio in a given zip code.	1
3	As a developer I want to make sure the research findings are thorough by creating visualizations for 1, 3 and 5 mile radius to be able to see the differences and similarities.	1

3	As a developer I want to create visualizations of various indicators grouped together for specific cities in order to diversify information and reduce risk.	2
3	As a developer, I want to distinguish the relationship between popularity and price in a specific area provided with a zip code.	2
3	As a developer I want to verify that there are similar trends in the Redfin, MLS, and Zillow datasets so that they can be used alongside each other for comparisons and analyses	1
3	As a developer, I want to format Redfin data in a way that is compatible to be inserted into PowerBI.	3
3	As a developer I want to include Building Permit Surveys into my research with MLS data prices.	0
	Epic: Presentations for Sponsors	
2	As a project manager and developer, I want to prepare a presentation where I can present my findings and show my work to sponsors.	3
2	As a developer, I want to create visualizations that are easy to read and understand from sponsors.	1
2	As a developer and project manager, I want to have visualizations for Hurricane Ian to draw conclusions and report to sponsors.	1
2	As a developer and project manager, I want to select the best findings to include in the presentation for sponsors.	1
3	As a developer and project manager I want to present my findings from my research and analysis to the sponsors to review and receive feedback.	4
	Epic: PowerBI Dashboard	
4	As a developer, I want to combine my findings with those from my peers to include them into one final report.	2
4	As a developer, I want to format the PowerBI dashboard in accordance with the sponsor's desires.	1
4	As a developer, I want to aggregate the MLS dataset for the top 200 MSAs, as required by sponsors.	2

4	As a developer, I want to create a map in PowerBI showing the findings in an interactive map.	2
5	As a developer, I want to collect useful examples that show trends in the PowerBI report.	1
5	As a developer, I want to integrate the aggregate model data into the PowerBI report.	1
5	As a developer, I want to be able to filter the PowerBI hotness map by state.	1
5	As a developer, I want to create tables representing changes as a result of federal rates increases in the PowerBI report.	1
4	**Epic: Forecasting Model**	
4	As a data science student, I want to develop an STL model for Closing Price.	3
4	As a data science student I want to apply the developed STL model to the top 200 MSAs.	1
5	As a developer, I want to turn the model into Pandas UDF.	1
5	As a developer, I want to aggregate the forecasting model output to a single table.	2
5	As a developer, I want to backtest the forecasting model.	3

Table 5.3: User Stories and Epics

6. Design

6.1 High Level Architecture

Our technology stack largely involved technologies that are commonly used for data science and big data management. We were given access to the investment firm's instance of databricks to query and interact with the firm's Azure Data Lake which holds many of their datasets. Databricks, running on top of Apache Spark, gave us an efficient way to query data through both Python and Structured Query Language (SQL). On top of that we were able to install various third-party python libraries such as pandas, statsmodels, and plotly. The overall architecture of the data processing pipeline that we used throughout the project is shown in figure 6.1 below.

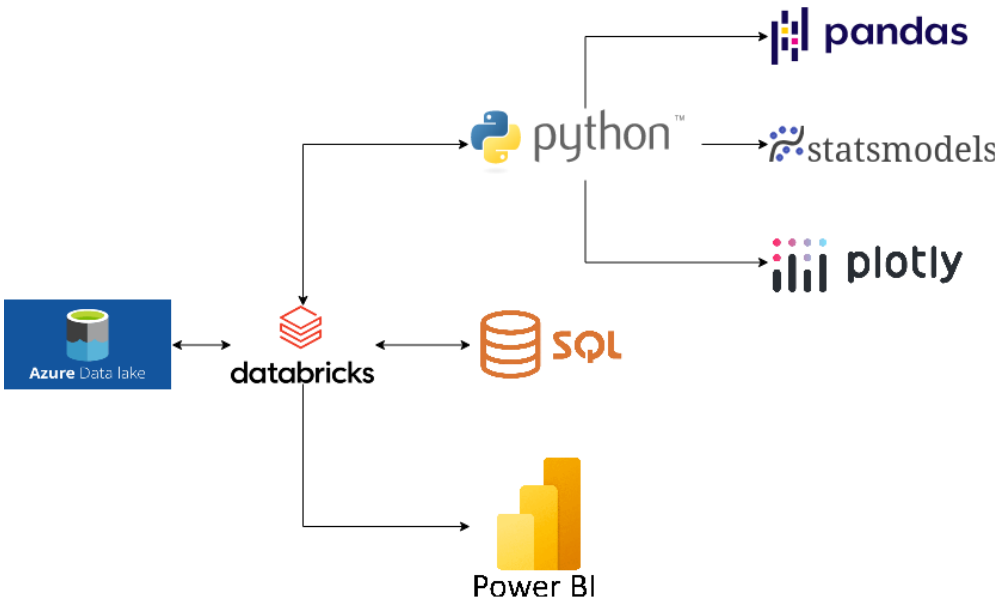

Figure 6.1: Architectural diagram displaying technologies and libraries used

Figure 6.1 shows a high level view of the different technologies we used throughout the project and how they interacted with each other. Databricks, at the center of the diagram, was where most of the development work occurred. We were able to utilize its easy-to-use web interface to query the firm's databases which gave us necessary access to large amounts of data quickly. Within databricks we were able to take advantage of both SQL, for querying data from the databases, and python which we used to conduct the various analyses. With python we were able to query the databases through PySpark which would return the data in a convenient data structure which could be easily converted into a pandas dataframe, a modern standard for working with data in python. Being such a well-known standard within the python community these data frames were able to work well with the other libraries that we used in our analyses. Plotly, our data visualization library, was well suited to take in data from the data frames and output the various charts and graphs that we needed to produce. The dataframes also worked well with the statsmodels library which assisted us in using the various statistical models and applying them correctly and efficiently.

6.2 Entity Relationship Diagrams

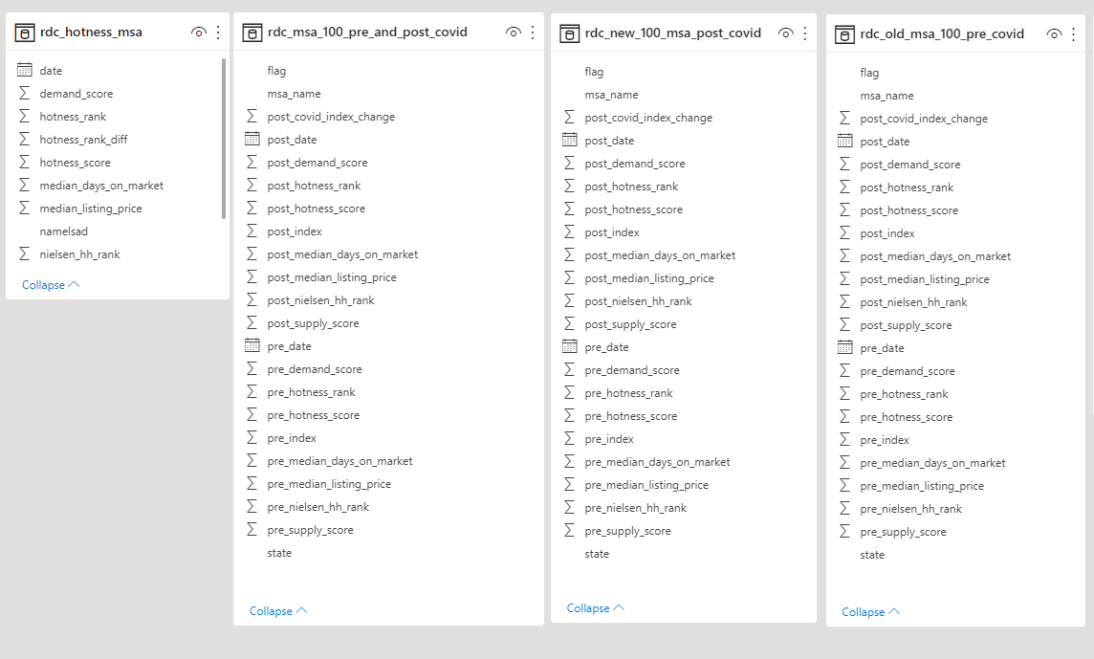

Figure 6.2.1: First Entity Relationship Diagram for Realtor.com MSA Rankings

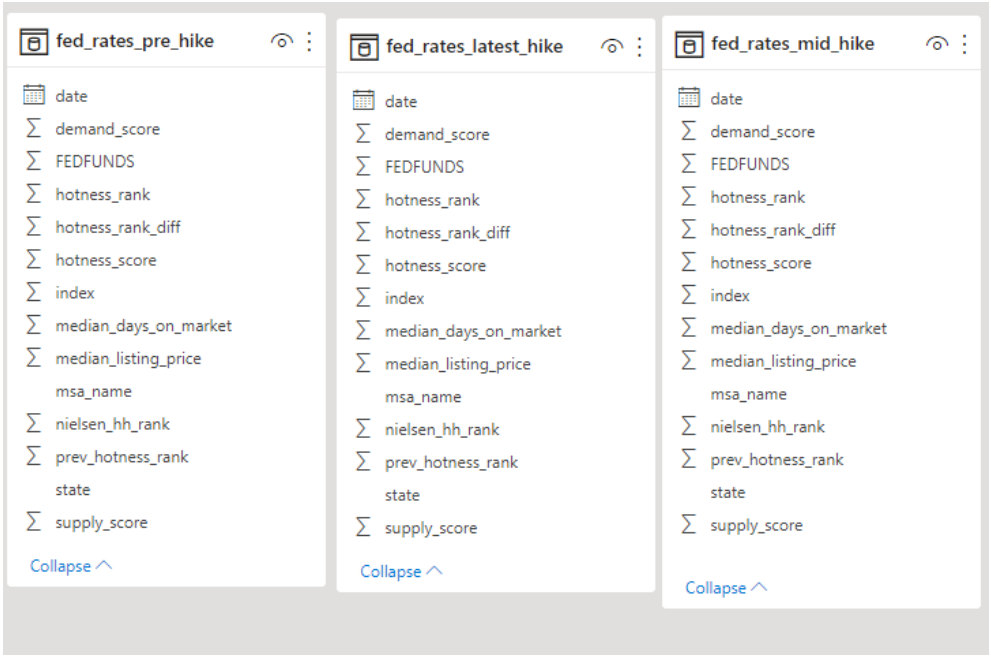

Figure 6.2.2: Second Entity Relationship Diagram for Realtor.com MSA Ranking

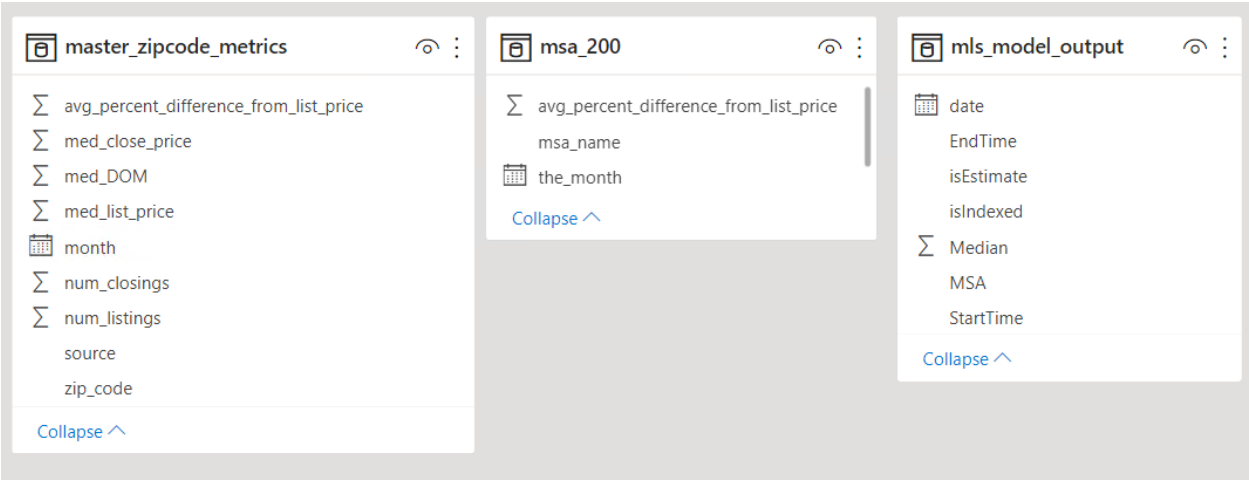

Figure 6.2.3: Entity Relationship Diagram for Combined MLS, Realtor.com, and Redfin Analyses

7. Software Development

7.1 Sprint One

During the first week of the project, an agile board was developed and user stories were created for everyone to work on according to requirements set by the sponsors. Using Jira to manage our agile process, the following chart represents the first sprint:

Completion confirmation	User Story	Points
	Epic: Final MQP Report	
Yes	As a reader I want an up to date introduction and abstract so that I can have an understanding of the contents of the MQP report.	1
Yes	As a project manager I want to identify possible risk factors and possible mitigations so that we can avoid “them” as we continue.	2
	Epic: Trend Identification	
Yes	As a reader I would like to have background information on how major hurricanes have impacted local housing markets.	2
No	As an investor I want to have information regarding trends in local housing markets in response to major hurricanes so that I can possibly identify similar trends in the most recent hurricane.	5
Yes	As a developer I want to identify columns in the MLS dataset that would be useful in identifying trends in the housing market so that we can narrow down the dataset and focus on those columns	3
Yes	As a developer I want to identify what columns or derived fields will be used as output so that I can structure my queries for those fields	3
Yes	As a developer I want to identify key import tables that will be useful for my queries in the future.	3
Yes	As a reader and project manager I want visualizations of the identified housing trends so that they are easier to understand and identify.	2

	Epic: Tool Familiarization	
Yes	As a developer I want to familiarize myself with how to interact with the MLS dataset through Databricks so that I can work with it efficiently in the future	1
	Total Points Completed	17

Table 7.1: Sprint One User Stories

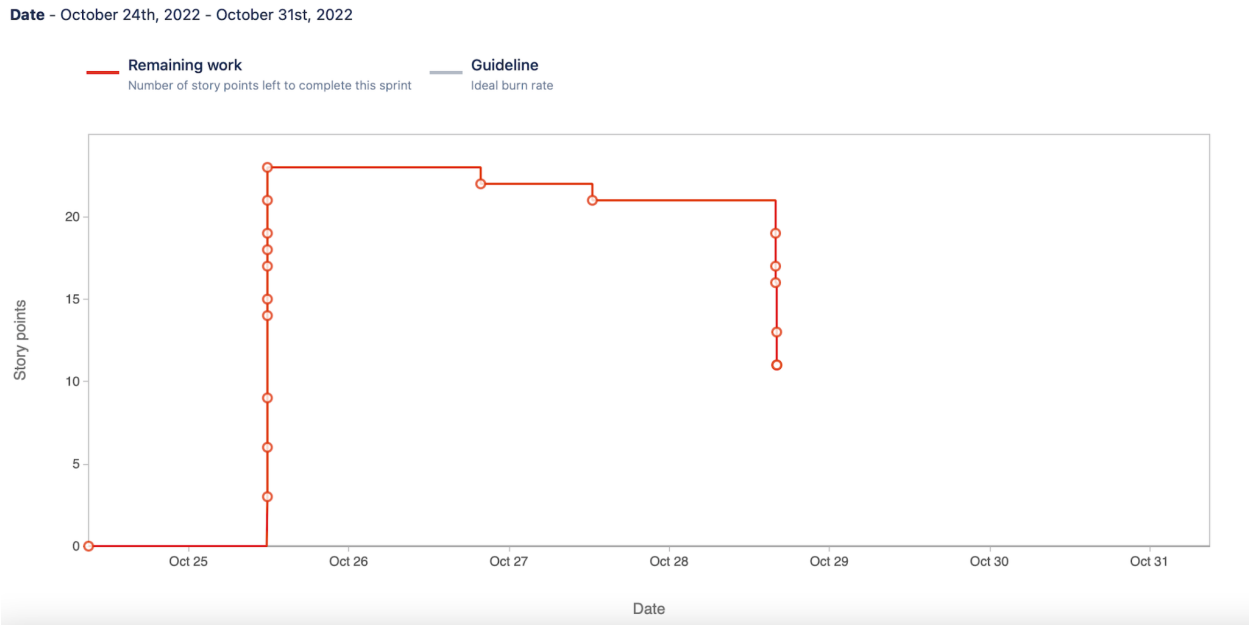

Figure 7.1: Sprint One Burndown Chart

The first sprint was very much experimental. We began familiarizing ourselves with the MLS dataset and the remote access. Our main goal was to explore trends in the housing market in response to hurricanes. We looked at a few different sectors of housing, such as listings, cancellations, and home prices during periods when different hurricanes hit the coast of the U.S. We were not able to find any particularly strong trends, most of the data was inconsistent from one hurricane to another. This makes us think that there are other factors that come into play when it comes to hurricanes and the housing market. From external research, we found that

insurance plays a significant role, government regulations on the building code, and also geographic importance for strategic investments. This last one is important as to whether big investors would want to get involved with real estate investments in a particular area after a hurricane. There might be great potential, but there are also risks associated. We carried some tasks over to the second sprint to continue to explore those ideas.

Retrospective:

We believe this first sprint taught us a lot, whether it was technical learning or industry learning, the challenges presented contributed to our learning. We were able to advance ahead with the final paper by adding new background information and reworking previously written sections according to advisors' feedback. We are now familiar with databricks and can work with the tool confidently. From a project management point of view, attendance at the meetings was good and we received positive feedback from sponsors and advisors.

Because this first sprint was more about familiarizing ourselves with the software and the data, we found it challenging to break down the tasks and assign particular work to team members. The overall goal of the sprint was also not very clear from the sponsors, challenging us to find areas of interest among an incredibly broad database. To ease these challenges, we decided to implement action items on the second sprint. Having a better understanding of the objective, we could create more detailed tasks for team members, clarifying duties. We hope these challenges will turn into learnings.

7.2 Sprint Two

During the second week of the MQP, the agile board was updated with a new sprint to evaluate and keep track of new work progress. From the Jira agile software, the following user stories, representing the second sprint were extracted:

Completion confirmation	User Story	Points
	Epic: Final MQP Report	
Yes	As a reader and developer, I want to include information on the design chapter of the MQP report.	3
No	As a reader, I want to have thorough and formal information on chapters of the MQP report.	2
	Epic: Trend Identification	
Yes	As an investor I want to have information regarding trends in local housing markets in response to major hurricanes so that I can possibly identify similar trends in the most recent hurricane.	5
Yes	As an investor, I want to find information that can produce potential opportunities for investments in the real estate market after hurricane destruction.	5
Yes	As a developer, I want to analyze the housing market's average price change over time after hurricanes, to assist in identifying possible trends in the market.	2
	Epic: Presentations for Sponsors	
Yes	As a project manager and developer, I want to prepare a presentation where I can present my findings and show my work to sponsors.	3
Yes	As a developer, I want to create visualizations that are easy to read and understand from sponsors.	1
Yes	As a developer and project manager, I want to have visualizations for Hurricane Ian to draw conclusions and report to sponsors.	1

Yes	As a developer and project manager, I want to select the best findings to include in the presentation for sponsors.	1
	Epic: Extracting "Signal"	
Yes	As a developer, I want to investigate listing prices in areas or regions with considerable growth fluctuations.	1
Yes	As a developer, I want to investigate closing costs in areas or regions with considerable growth fluctuations.	1
Yes	As a developer, I want to investigate days on market for listings in areas or regions with considerable growth fluctuations.	1
Yes	As a project manager, I want to be able to identify areas where the real estate market has recorded considerable fluctuations over time.	2
Yes	As a developer and project manager, I want to compare Zillow valuations to MLS factors like listing price and close price to see how they line up.	2
Yes	As a developer, I want to investigate the number of listings in areas or regions with considerable growth fluctuations.	1
	Total Points Completed	29

Table 7.2: Sprint Two User Stories

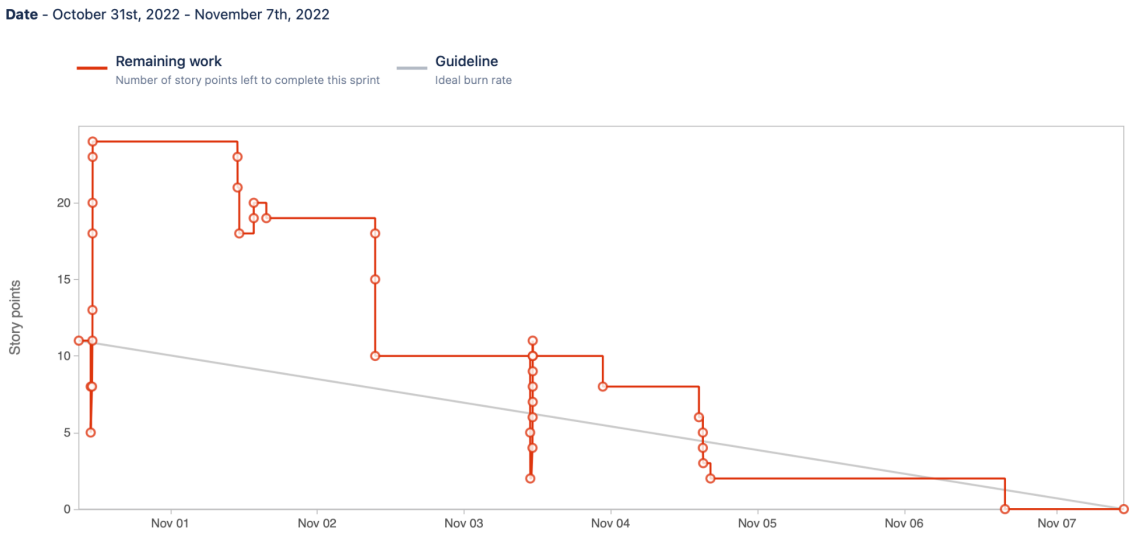

Figure 7.2: Sprint Two Burndown Chart

This second sprint, we researched deeper into the hurricane topic as suggested by the sponsors. We used the MLS data and outside resources to prepare a presentation for the sponsors to show our findings and challenges, and receive feedback from them. After our research, we were able to see some minor trends of hurricane effects on the housing market. We analyzed the information based on four major hurricanes: Michael, Katrina, Harvey and the most recent one, Ian. From our research we found that both the number of listings and the number of closings experienced a sharp decrease during the week of the hurricanes. There was no trend identified for the median and mean of price change. However, there were increases in the median number of days a house was on the market, and across the hurricanes we looked at there was always a quick recovery back to normal within a couple weeks. The recoveries depend on the location of the region and the importance of the location.

Even though there is significant damage from hurricanes and many regions are marked as risky areas, there is still a high demand from investors in these areas due to their investment potential. For example, regions of Florida, which are more commonly affected by hurricanes, are still a “hot spot” for investors, due to other factors such as low taxes, great weather, and even

politics. The problem with the increased demand after hurricanes, is that middle to low income individuals experience a hard time recovering, especially those without homeowner's insurance, which can be a whole new case study. These individuals are leaving the area for safety purposes during hurricanes and most of them do not return, creating a change in demographics, due to the posed "restrictions" on who can live in the area after big investors come in. Hurricane research is a broad area of investigation, which can lead to great investment potential and great case studies. After the presentation, however, the sponsors decided to change the direction of the project and move away from hurricane effects on real estate, focusing more on identifying signals that suggest fluctuations in the housing sector. These signals will then be used by the sponsors for further investigation and research, potentially deriving possible investment opportunities, or avoiding risky areas of investment.

Retrospective:

We implemented our learnings and feedback from the first sprint into this second sprint. Our agile board was more detailed and we were able to better split up user stories. Workflow was better spread and more objective oriented. We also decided to rotate the responsibility of leading meetings with advisors and sponsors, which went really well. Overall we made good progress on the project and learning.

A challenge that we found during this sprint was the mid-week presentation for the sponsors, which changed the direction of the project for the remainder of the week, causing some disruption on the agile planning and forcing us to adjust the agile board to meet the new objectives. To address this challenge we planned on adjusting the agile schedule accordingly if future direction changes occur in the middle of the week. We also found it challenging catching up with some information presented from sponsors during stand ups, for which we decided to

take notes during meetings to be able to refer to as we work independently. Lastly, from the first research topic we noticed some improvement areas from the research outside the MLS data, and decided to have a more connected outside research with the MLS data.

7.3 Sprint Three

During this third week, we changed the timeline of the sprints from Monday-Friday to Thursday-Wednesday as a result of changes in the project direction after presentation day on Wednesday. Because of this change this sprint is stretched out to a week and a half, which results in a higher number of points compared to previous sprints.

Completion confirmation	User Story	Points
	Epic: Final MQP Report	
Yes	As a project manager and investor, I want to identify risks related to investments in the real estate industry.	2
Yes	As a project manager, I want to identify business risks and risk mitigation methods related to real estate investments, derived from previous analysis.	3
Yes	As a reader I want to see a section on the software development environment used in the project	1
Yes	As a reader, I want to be able to see the research and analysis findings presented to sponsors.	2
Yes	As a reader, I want to see some social impact of the real estate industry, derived from research.	1
Yes	As an investor and project manager, I want to collect information on past historical events that will yield potential data and information to be linked with current and future trends.	3
Yes	As a reader I want to have information in the paper on the current state of the real estate market, including indicators and potential trends.	2
	Epic: Extracting Signal	
Yes	As a developer, I want to make a table that has all the analysis that has been done in Databricks for the MLS data on a monthly basis to	3

	compare with Zillow data. (For zip codes with sponsor's properties)	
Yes	As a developer I want to be able to generate a graph of the year over year growth by quarter of the percent of homes that were sold over listing price in a given zip code.	1
Yes	As a developer I want to be able to generate a graph which shows the year over year growth of the median "sale price to list price" ratio in a given zip code.	1
Yes	As a developer I want to make sure the research findings are thorough by creating visualizations for 1, 3 and 5 mile radius to be able to see the differences and similarities.	1
Yes	As a developer I want to create visualizations of various indicators grouped together for specific cities in order to diversify information and reduce risk.	2
Yes	As a developer, I want to distinguish the relationship between popularity and price in a specific area provided with a zip code.	2
Yes	As a developer I want to verify that there are similar trends in the Redfin, MLS, and Zillow datasets so that they can be used alongside each other for comparisons and analyses	1
Yes	As a developer, I want to format Redfin data in a way that is compatible to be inserted into PowerBI.	3
No	As a developer I want to include Building Permit Surveys into my research with MLS data prices.	2
	Epic: Trend Identification	
Yes	As a developer I want to create visualizations of Redfin data in PowerBI.	2
Yes	As a developer and project manager, I want to create a list of cities that present potential trends in the real estate market.	1
Yes	As a developer I want to develop an ARIMA model for predicting prices like listing or closing prices, based on the data from previous times.	2
No	As a developer I want to compare Realtor.com data with the MLS data in order to see how they line up and identify possible trends.	3

Yes	As a developer I want to perform a statistical test on the datasets from Zillow and MLS to show how they line up and if they follow the same trend.	2
	Epic: Presentations for Sponsors	
Yes	As a developer and project manager I want to present my findings from my research and analysis to the sponsors to review and receive feedback.	4
	Total Points Completed	39

Table 7.3: Sprint Three User Stories

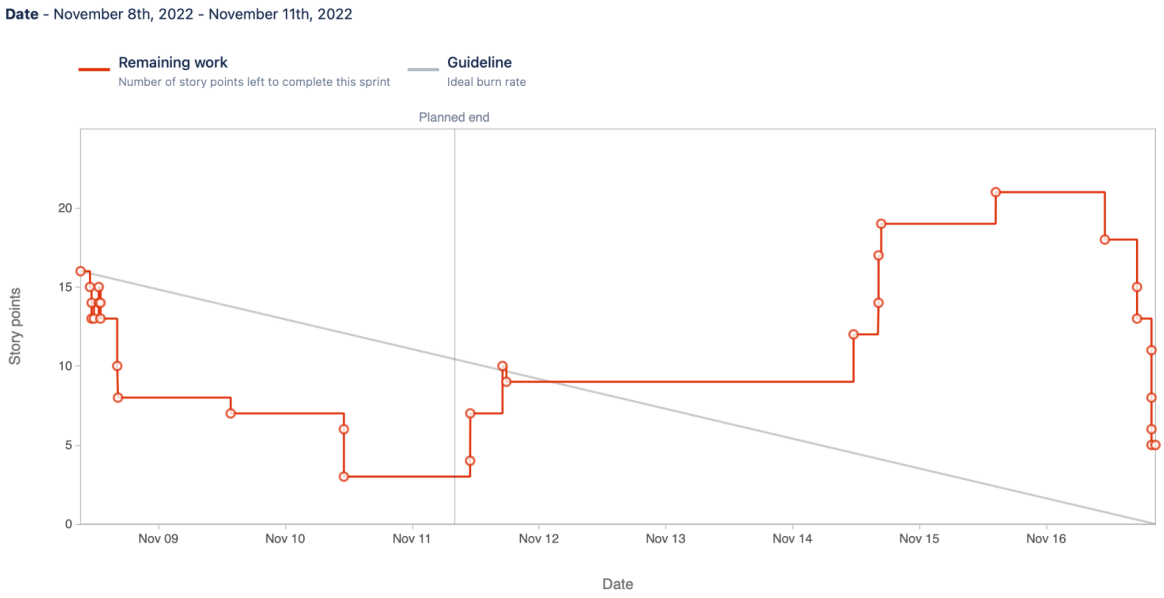

Figure 7.3: Sprint Three Burndown Chart

During this third sprint, we decided to switch the timeline as previously mentioned. After last week's presentation with the sponsors, we added new tasks to the agile board and continued with the sprint throughout the next week. We pivoted away from hurricanes and looked at factors influencing the real estate market, trying to identify potential trends. We were also given datasets from Zillow, Redfin, and Realtor.com to assist in our analysis. We performed some statistical

analysis to determine the similarities and differences in the data between different providers. From the results we concluded that some of the data follow the same trend, but most of the data from MLS and Zillow failed to pass the Mann-Whitney U Test, which we used for this particular issue. This shows that there are differences in the way MLS and other platforms collect and use data and the risk of relying on only one platform can be significant. After last week's presentation, we started developing a PowerBI report that will combine all the visualizations we have used so far through MLS and other platforms into one large report.

Retrospective:

This sprint was well executed. We received good feedback from sponsors during two presentations that happened this sprint. Changing the timeline of the sprint will be beneficial to us, because we present our progress on Wednesdays and receive feedback from the sponsors which we can then take into the next sprint and plan accordingly. This way we do not have to rush objectives and rewrite new ones in the middle of the sprint. Progress is looking good and we are producing results that seem beneficial to us and sponsors.

A challenge that we found during this sprint was the familiarization with the new timeline and the need to change scheduled meetings in our agendas. Furthermore, we were introduced with a new platform like PowerBI, which gave us more challenges with installation and familiarization. To address the challenges with PowerBI, we spent some extra time with the tech team, which were able to solve any issues and concerns we had. Lastly, during our first presentation last week, we encountered some challenges coordinating and presenting our findings. To address this we decided to meet a few minutes before future presentations and go over the presentation addressing any issues that come up.

7.4 Sprint Four

During this sprint, we extended the sprint length again, due to the holidays break in the middle of the sprint. This extension will contribute to more sprint points, compared to a regular sprint week.

Completion confirmation	User Story	Points
	Epic: Final MQP Report	
Yes	As a reader, I want to read about the design and architecture of the project.	1
Yes	As a developer and project manager, I want to have a conclusion chapter for the project.	1
Yes	As an ISC concentration student, I want to include Innovation for Social Change concentration findings on the project.	1
Yes	As a project manager, I want to have business related takeaways under the assessment chapter of the report.	2
Yes	As a developer and data science student, I want to include the ARIMA model findings in the project report.	1
Yes	As a project manager, I want to have a complete risk management section in the project report.	1
Yes	As a reader I want to be able to identify different tables and figures on the project report.	1
Yes	As a reader I want to be able to see all the user stories and epics up to date in a separate chapter of the project report.	1
Yes	As a developer, I want to write about the software applications and environment that is used throughout the project.	2
Yes	As a data science student, I want to write on the project report about the mathematical method called the U Test.	1
Yes	As a developer, I want to edit and keep the software requirements section of the paper up to date.	1

Yes	As a developer and project manager, I want to write an assessment of the project reflecting the work completed throughout the term.	1
No	As a project manager, I want to include the last sprint's findings presented to sponsors, into the project report.	2
Yes	As a data science student, I want to write on the project report about the mathematical method of Autocorrelation/Partial Autocorrelation.	1
Yes	As a data science student, I want to write on the project report about the mathematical method used in the project, called Augmented Dicky-Fuller Test.	1
Yes	As a data science student, I want to write on the project report about the mathematical method used in the project, called Time Series.	1
Yes	As a data science student, I want to write on the project report about the mathematical method used in the project, called STL.	1
Yes	As a data science student, I want to write on the project report about the mathematical method used in the project, called ARIMA.	1
	Epic: Trend Identification	
Yes	As a developer I want to compare Realtor.com data with the MLS data in order to see how they line up and identify possible trends.	3
	Epic: PowerBI Dashboard	
Yes	As a developer, I want to combine my findings with those from my peers to include them into one final report.	2
Yes	As a developer, I want to format the PowerBI dashboard in accordance with the sponsor's desires.	1
Yes	As a developer, I want to aggregate the MLS dataset for the top 200 MSAs, as required by sponsors.	2
Yes	As a developer, I want to create a map in PowerBI showing the findings in an interactive map.	2
	Epic: Forecasting Model	
Yes	As a data science student, I want to develop an STL model for Closing Price.	3

Yes	As a data science student I want to apply the developed STL model to the top 200 MSAs.	1
	Total Points Completed	33

Table 7.4: Sprint Four User Stories

Figure 7.4: Sprint Four Burndown Chart

Throughout this sprint we continued with the new schedule implemented last week. We had to extend the sprint as a result of the Thanksgiving break, in order not to have a short 3 day sprint. We added more stories as we worked on finalizing the paper and producing a PowerBI report, together with forecasting models for the sponsors. We did not continue with research through the MLS data sets, but instead we converted our results from the previous weeks into a PowerBI report that the sponsors will use for further analysis.

Retrospective:

This sprint was also well executed, given also the distraction from the break. We were able to adapt to the changes quickly and work our way through completing our tasks, continuing to make good progress on our objectives. Our sponsors appeared to be satisfied with the overall progress.

A challenge that we faced during this sprint was working on the paper and on the forecasting model and PowerBI dashboard at the same time. Because we are heading towards the end of the project, we are simultaneously trying to finish up our work, which increases the challenge. To overcome these challenges, we are trying to prioritize work by the due dates, so we can complete our tasks in a chronological order. We also saw slight declines in the activity of the agile board during this sprint, so we will focus on reporting our work more actively in Jira.

7.5 Sprint Five (Final)

During this sprint we worked on finalizing our project in accordance with the requirements from our advisors and our sponsors. To keep track of our progress, we continued to utilize our Jira board, the results of which are shown in the table below:

Completion confirmation	**User Story**	**Points**
	Epic: Final MQP Report	
Yes	As a sponsor, I want to have a section on the paper mentioning how the remote access was given to students.	1
Yes	As a business student, I want to write an executive summary of the whole project report in the form of a business report.	2

Yes	As a sponsor, I want to keep the name of the company private and outside of the project report.	2
Yes	As a project manager, I want to include the last sprint's findings presented to sponsors, into the project report.	2
Yes	As a reader, I want to have an up to date intro and abstract which applies to the report.	1
Yes	As a reader, I want to have the right page numbering on the paper to easily navigate through.	1
Yes	As the writers of the report, we want to have a bibliography with all the references used throughout the project.	2
Yes	As a developer, I want to include a description of sprint and product backlog in the appertaining section of the report.	1
Yes	As writers of the paper, we want to acknowledge everyone who made the project possible, by dedicating a section for them in the paper.	1
Yes	As writers of the paper, we want to have a conclusion chapter for the project.	1
Yes	As writers, we want to include a future work section on the paper to represent the work that can be continued in the future.	2
Yes	As a writer, I want to address the advisors' feedback and comments and make the necessary changes to the paper.	3
Yes	As writers of the paper, we want to have a thorough report, therefore we want to go over the paper and address any issues or clarifications that need attention.	5
	Epic: Forecasting Model	
Yes	As a developer, I want to turn the model into Pandas UDF.	1
Yes	As a developer, I want to aggregate the forecasting model output to a single table.	2
Yes	As a developer, I want to backtest the forecasting model.	3
	Epic: PowerBI Dashboard	

Yes	As a developer, I want to collect useful examples that show trends in the PowerBI report.	1
Yes	As a developer, I want to integrate the aggregate model data into the PowerBI report.	1
Yes	As a developer, I want to be able to filter the PowerBI hotness map by state.	1
Yes	As a developer, I want to create tables representing changes as a result of federal rates increases in the PowerBI report.	1
	Total Points Completed	34

Table 7.5: Sprint Five (Final) User Stories

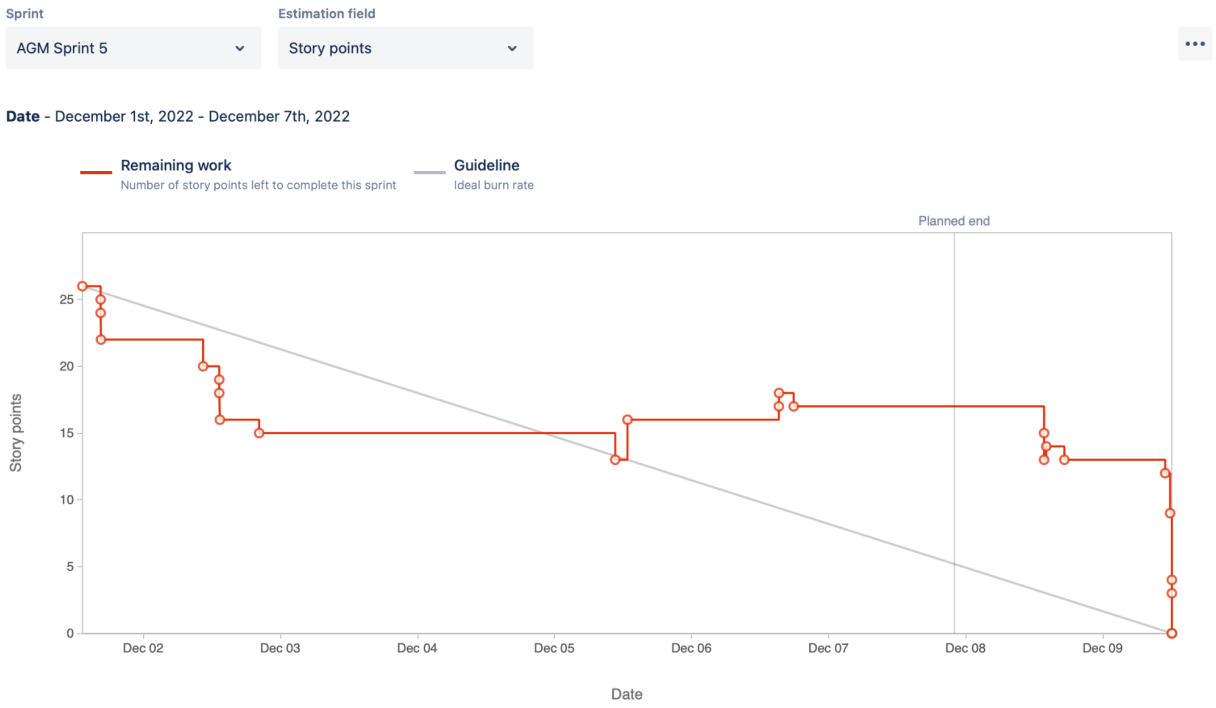

Figure 7.5: Sprint Five (Final) Burndown Chart

Throughout this final sprint, we focused on developing our PowerBi report, the forecasting model and finalizing our project report. Because of continuous incoming feedback

from sponsors and advisors, we decided to close the last sprint on Friday instead of Wednesday, in order to be able to address every issue before finishing the last sprint.

Retrospective:

We successfully completed our tasks for this last sprint, making overall good progress on everything. We were able to receive feedback on our final project report draft from the advisors, which we found very useful and used the feedback to align our progress with the requirements from sponsors and advisors. A challenge that we faced throughout the sprint was the final presentation preparation, while finalizing our work. To address this, we have decided to prioritize tasks and also get together to work on our final conclusions and findings that we believe will be applicable to include in the final presentation.

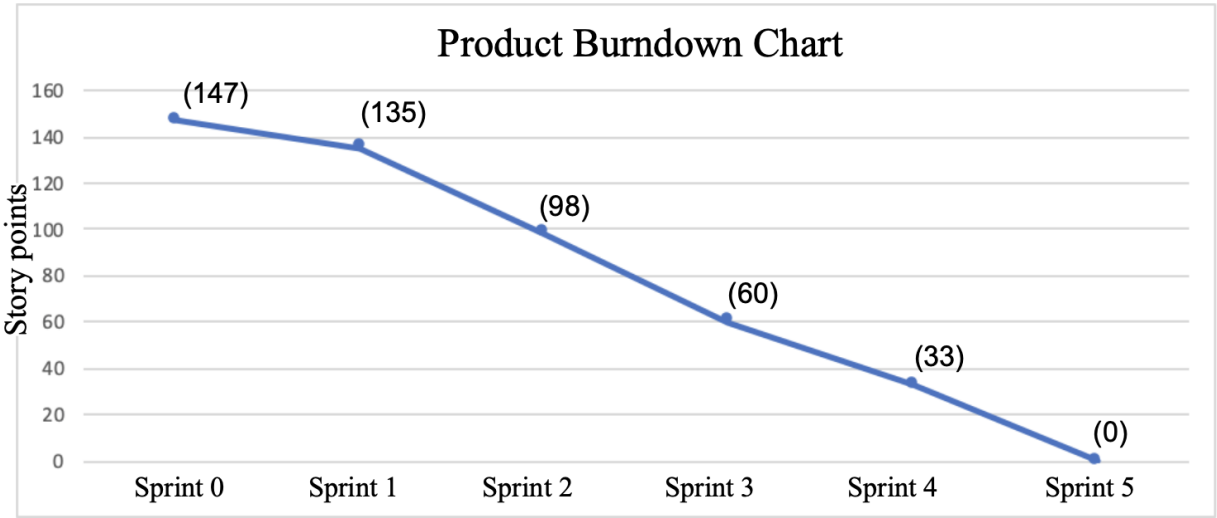

Figure 7.5.1: Final Product Burndown Chart

8. Findings

8.1 Sprint One and Two Findings

Analyzing the number of new listings from the MLS dataset provided, we were able to create the following visualizations:

Figure 8.1.1: MLS number of new listings during Hurricane Katrina (2005)

Figure 8.1.2: MLS number of new listings during Hurricane Michael (2018)

From these graphs, we can see a trend of a decreasing number of new listings during the week when hurricanes occur. During the hurricane season, homeowners are of course hesitant to list their homes for sale, as the results of the hurricane aftermath are uncertain, which is why such a decrease in the number of new listings happens. This trend is also seen among other hurricanes, listed in the table below:

Michael (FL)	-88.5%
Katrina (La.)	-96.7%
Harvey (TX)	-59.56%
Ian (FL)	-51.22%
Average Change	-73.99%

Table 8.1.1: Percent changes in new listings during hurricanes

On average, there is about a 75% decrease in the number of new listings during the hurricane period, as shown in the table drawn from the MLS dataset. What seems interesting from the visualizations is the recovery period after the hurricanes, which usually appears to be relatively quick. The quick recovery comes as a result of increased demand from investors, who express interest in buying and flipping damaged houses.

The number of closings appears to have a significant decrease during the hurricane period and a quick recovery afterwards, similar to the number of listings, because of the same reasons. The visualizations below represent the number of closings during hurricane Katrina and Harvey respectively:

Figure 8.1.3: MLS number of closings during Hurricane Katrina (2005)

Figure 8.1.4: MLS number of closing during Hurricane Harvey (2017)

The table below shows the percent decrease in number of closings during the week of hurricanes. Compared to the number of new listings, this is a slightly smaller decrease on average, which comes as a result of a small decrease in the number of closings for hurricane Ian.

Michael (FL)	-85.81%
Katrina (La.)	-99.53%
Harvey (TX)	-46.75%
Ian (FL)	-20.15%
Average Change	-63.06%

Table 8.1.2: Percent changes in closings during hurricanes

In terms of prices of homes, there appears to be no significant drop or increase during the periods of hurricanes, indicating that the prices in the market remain the same. Depending on other factors that are involved after the hurricane periods, prices can experience some changes. An example is the increase of prices of homes as a result of extended building code regulations for homes after Hurricane Michael in 2018. Similar extended regulations are expected to take place shortly after Hurricane Ian, as a result of significant differences in damages seen from houses exceeding the current building code requirements and those following the current

building code requirements. These factors are not included in the MLS dataset, but can influence prices and other areas of market analysis.

Another area analyzed with the MLS dataset was the days on market for houses in areas impacted by hurricanes over years. As seen in the charts below, there appears to be a slight increase in the days on market in most cases, as a result of the shock and distress that hurricanes bring. It takes a while to return back to normal after a hurricane, which contributes to increasing days on market for homes.

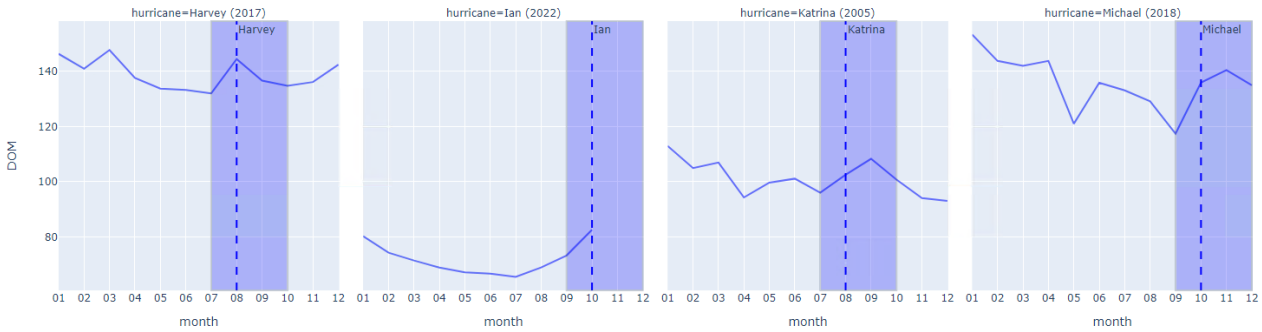

Figure 8.1.5: MLS days on market during Hurricane Harvey, Ian, Katrina and Michael

8.2 Sprint Three Findings

During the second and third sprint, we looked for trends in the real estate market overall. Using visualizations derived from the MLS datasets, we found that similar to Redfin and Zillow, the MLS datasets showed overall decreases in the number of new listings in 2022 compared to the previous two years. The MLS visualization of the number of new listings showed slightly more aggressive decreases compared to Redfin's graph, which is shown in section 2.4.

Similar to the country level, states like Florida and Arizona showed decreases in the number of listings and closings overall. To mitigate risk and diversify information, we compared the visualizations of MLS with those from Zillow, which is shown in the figures below:

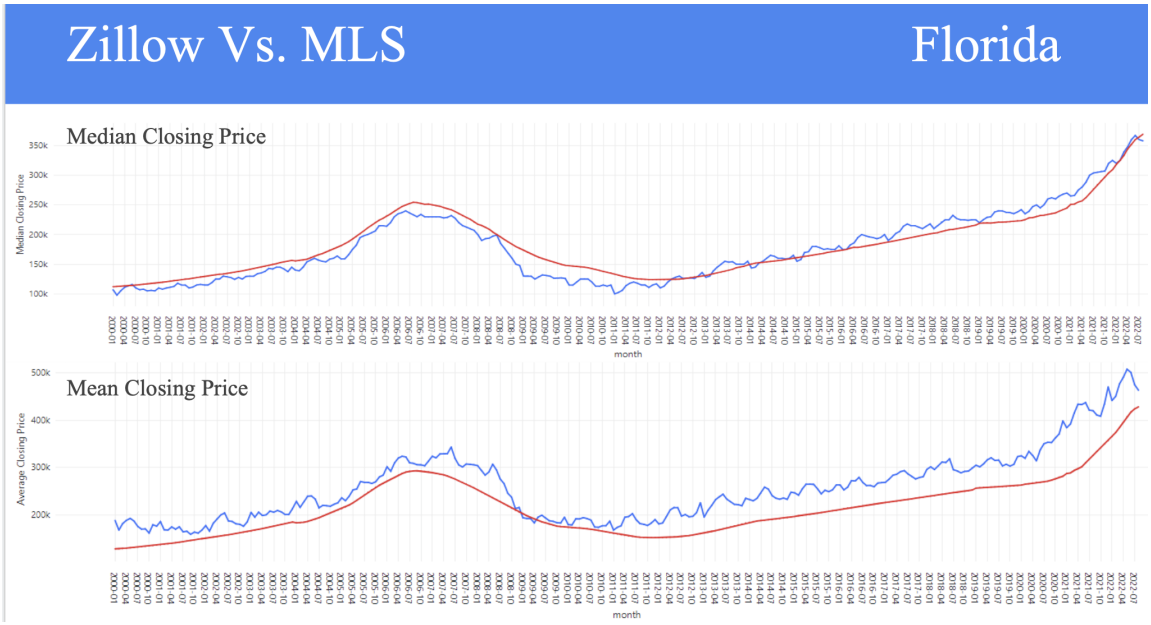

Figure 8.2.1: Zillow vs MLS data on Median and Mean Closing Prices in Florida

As seen in the figure, the trend between Zillow and MLS dataset seems to be the same for Florida. The same relationship appears for Arizona as well:

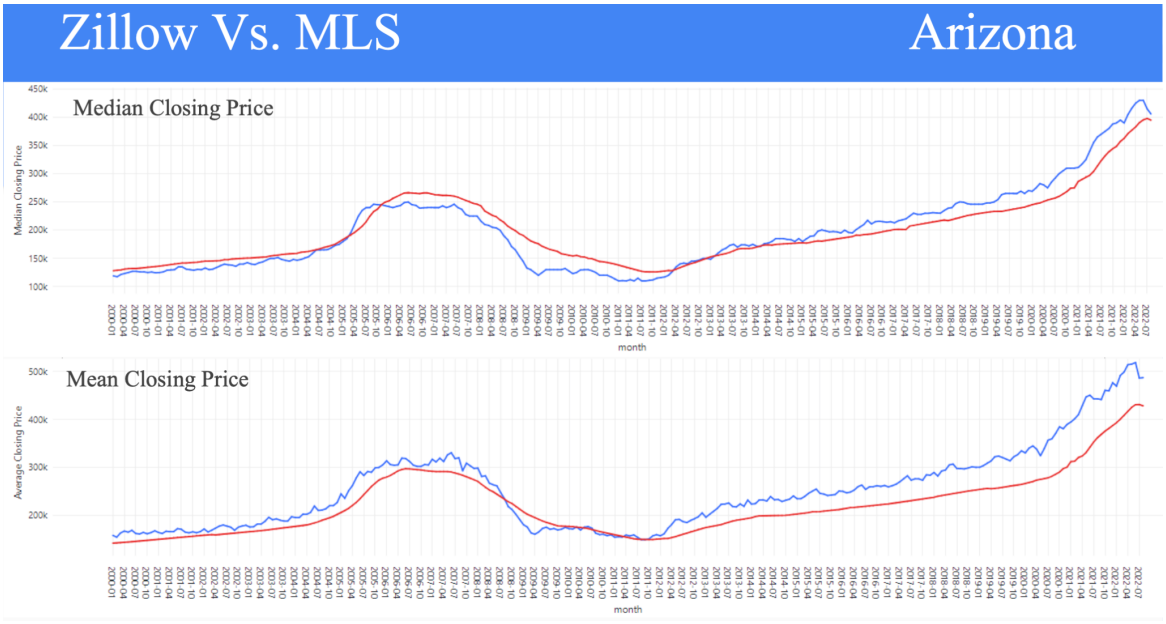

Figure 8.2.2: Zillow vs MLS data on Median and Mean Closing Prices in Arizona

At the national level, the trend seems to be the same, although it seems to be drifting slightly away from each other towards the end:

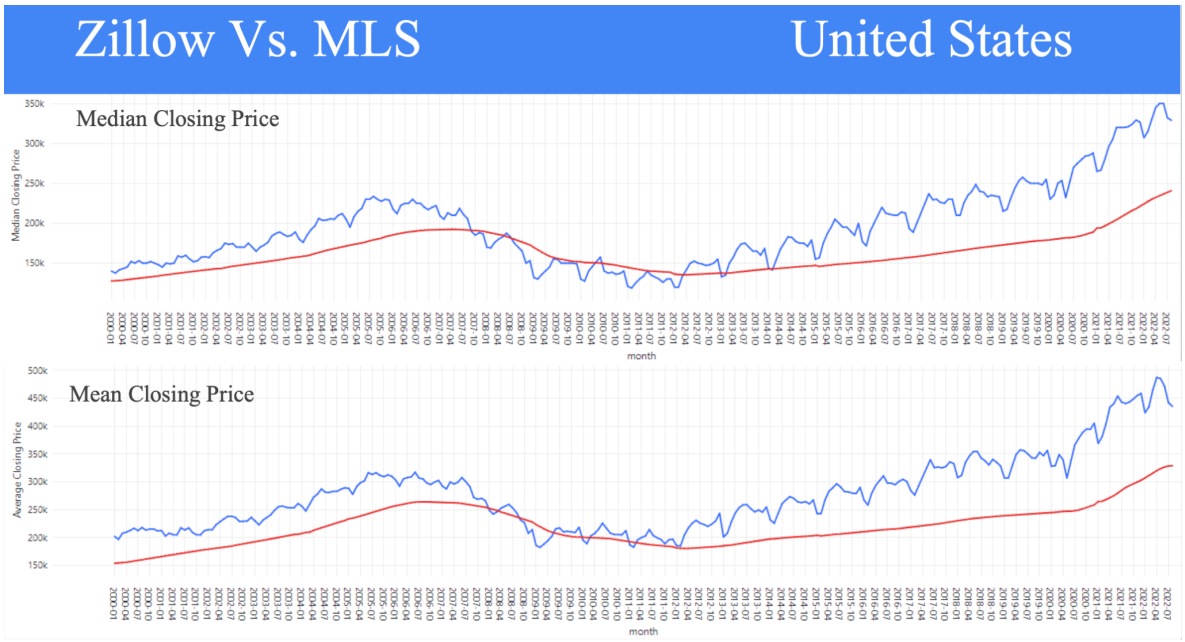

Figure 8.2.3: Zillow vs MLS data on Median and Mean Closing Prices in the United States

Because of the uncertainty whether the MLS dataset and Zillow data follow the same trend, we decided to do a Mann-Whitney U Test, which is a test used to determine if two samples derive from the same population [41]. The results from the test appear as follow:

	Mean Differences (in thousands)		Median Differences (in thousands)
AZ Mean	4.97e-06	AZ Median	0.102
FL Mean	1.13e-12	FL Median	0.045
US Mean	1.17e-55	US Median	1.76e-47

Table 8.2.4: Mann-Whitney U Test results

The null hypothesis for this Mann-Whitney test is that two datasets, Zillow and the MLS, have the same trend, while the alternative hypothesis is that two datasets do not have the same

trend. The test results for each pair is shown in Table 8.2.1. We did a total of 6 tests with both means and medians in three different areas, Arizona, Florida, and the United States as a whole. From the test, we concluded that only the median closing price for Arizona and Florida follow the same trend between Zillow data and MLS data, while the others, although they appear to follow the same trend on the graphs above, fail to pass the Mann-Whitney U Test.

Next, we implemented an ARIMA model to check for seasonality and convert non-stationary data to stationary. The model that we have is not complicated enough to meet the industry standard, so sponsors suggested an alternative model, the STL decomposition, that has been tested and used previously during similar projects this week.

We also looked at more specific areas where our sponsors currently own properties. These areas show the same decreasing trend in the number of listings during the year of 2022, similar to the national level trend. Something that we found interesting was the percentage of homes sold over the listing price. The figures below show the percentage of homes sold over-priced in three different areas where our sponsor holds properties. The data is graphed within a one mile radius, three mile radius and five mile radius.

Figure 8.2.5: MLS Percentage of Homes Sold Over List Price Year Over Year

Starting towards the end of 2020 and the beginning of 2021, there is a large spike in the percentage of homes sold over-priced, a trend which appears to be aggressively decreasing in 2022, which *indicates a decrease in demand and therefore possible price negotiations between homeowners and buyers.*

The ARIMA model is starting in sprint 3, simultaneously. At the beginning of the ARIMA modeling, the team took a look at the overview of several regions, and we found that the time series is non-stationary like most of the real world datasets.

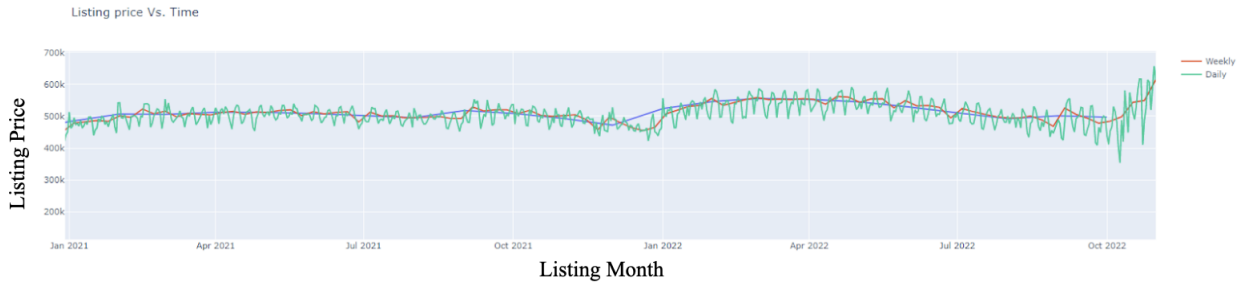

Figure 8.2.6: Overview of median price over listing month

From here, after we normalized the data for reducing computational power required, the first step that we took was addressing the non-stationarity within the dataset so that we could apply our model. We first took the difference of the dataset, an often used technique to achieve stationarity by excluding the general trends.

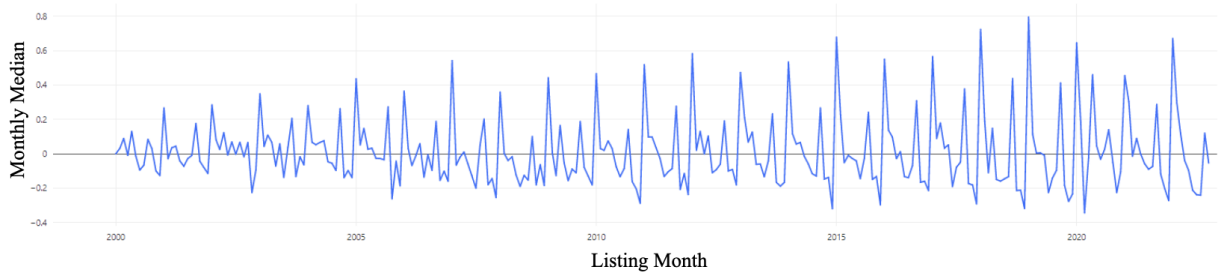

Figure 8.2.7: Monthly change of median of closing price over listing month, first differencing

After taking the first difference of the dataset, we can still identify some obvious non-stationary factors, seasonality and moving volatility. Therefore, the next step is to remove volatility by dividing each month's monthly change in price by standard deviation of that month. Here is the processed data up to this point:

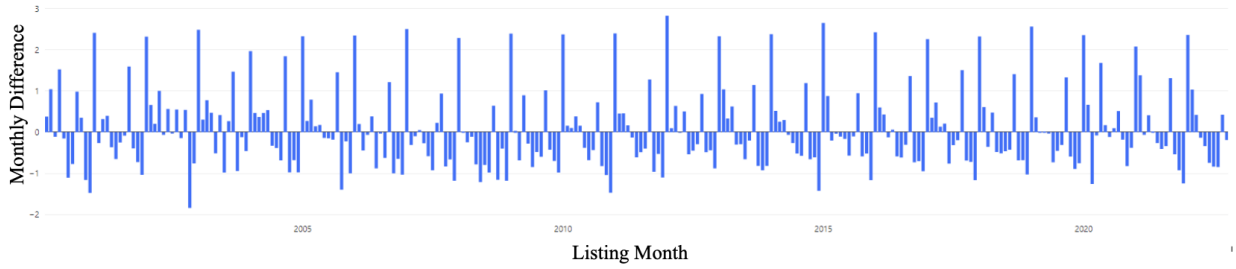

Figure 8.2.8: Data after first differencing and divided by standard deviation

From here, we can clearly see that volatility of the data has been drastically reduced, leaving peaks and valleys almost at the same level. Then, we were trying to tackle the seasonality of the time series by subtracting each month's monthly change by the average of that month. Here is the processed data up to this point:

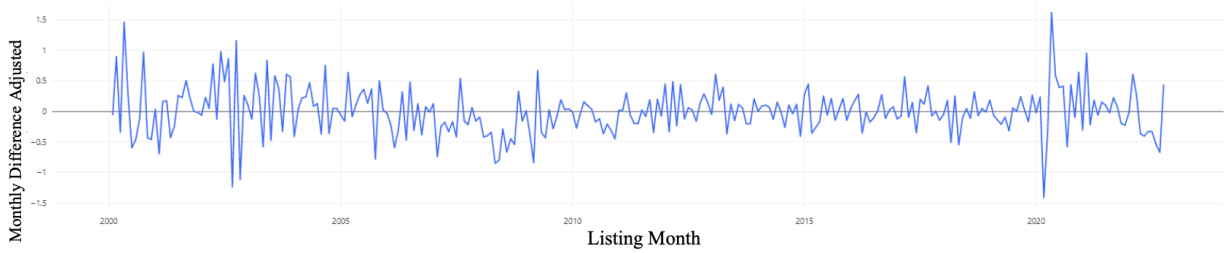

Figure 8.2.9: Data after first differencing, divided by standard deviation, and subtracted by the average of change

Up to this point, we have visual evidence that suggests that our data is stationary. However, a robust test is needed to confirm this conclusion. The team used the Augmented Dicky-Fuller Test, with the threshold of 0.05, in which the null hypothesis is that the dataset is non-stationary. In the result, we can tell that the P value is well below 0.05, so we can reject the null hypothesis and conclude statistically that the dataset is stationary. The test result is shown below:

```
Monthly Difference adjusted stats result:
1. adf: -3.9461903098519886
2. P value: 0.0017200209991774936
3. Lags: 8
4. Observation used: 264
5. Critical Value: {'1%': -3.455365238788105, '5%': -2.8725510317187024, '10%': -2.5726375763314966}
```

Figure 8.2.10: Result for Dicky-Fuller Test

Finally, we can conclude that the processed dataset shown in figure 8.2.8 is stationary, meaning that it is appropriate to implement the model. The next step is to estimate the parameters of the ARIMA model. In this project, we used computer algorithms to determine the best model parameters by trying out all kinds of model combinations.

SARIMAX Results

Dep. Variable:	Monthly_difference	**No. Observations:**	243
Model:	ARIMA(2, 0, 1)	**Log Likelihood**	-95.582
Date:	Tue, 15 Nov 2022	**AIC**	201.164
Time:	05:57:39	**BIC**	218.629
Sample:	0 - 243	**HQIC**	208.198
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.0109	0.051	-0.212	0.832	-0.112	0.090
ar.L1	0.6380	0.078	8.145	0.000	0.484	0.792
ar.L2	0.2914	0.052	5.593	0.000	0.189	0.393
ma.L1	-0.8404	0.074	-11.283	0.000	-0.986	-0.694
sigma2	0.1284	0.009	13.756	0.000	0.110	0.147

Ljung-Box (L1) (Q):	0.08	**Jarque-Bera (JB):**	25.34
Prob(Q):	0.77	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.32	**Skew:**	-0.24
Prob(H) (two-sided):	0.00	**Kurtosis:**	4.51

Figure 8.2.11: Result for determine parameters for the ARIMA model

The output in Figure 8.2.11 shows the result for a function provided by statsmodels that tries to fit different model parameters on the dataset, which helps in determining the best model to fit. The function determines the best model by minimizing the AIC (Akaike Information Criterion), a single number score that can be used to determine which of multiple models is most likely to be the best model for a given data set [65]. The reason for using AIC scores rather than others is that AIC score accounts for not only the accuracy of the fit, but also the simplicity of the model to prevent overfit. In this case, the dataset that we are trying to fit is the modified median price data, and the best model parameters indicated is ARIMA (2, 0, 1) with a minimum AIC of 201.164.

The final step is to input the data to the model which completes the process. The figure 8.2.12 is the plotted output of the fitted model (red curve) and the real data (blue curve). The model can be used in visualization for the model on the original dataset, the median price over

time in this project. Because the model is fitting the data that have been processed, we have to propagate the model fit back into the original data to visualize or draw conclusions. This backpropagation will be adding the monthly average to restore the seasonality, multiply the monthly standard deviation to restore the volatility, and lastly undo the differencing. However, the team did not continue on this model since the sponsors suggested some better and industry standard models at this point. If built successfully, this model can be used to generate predictions for the housing market which helps sponsors make wiser business decisions.

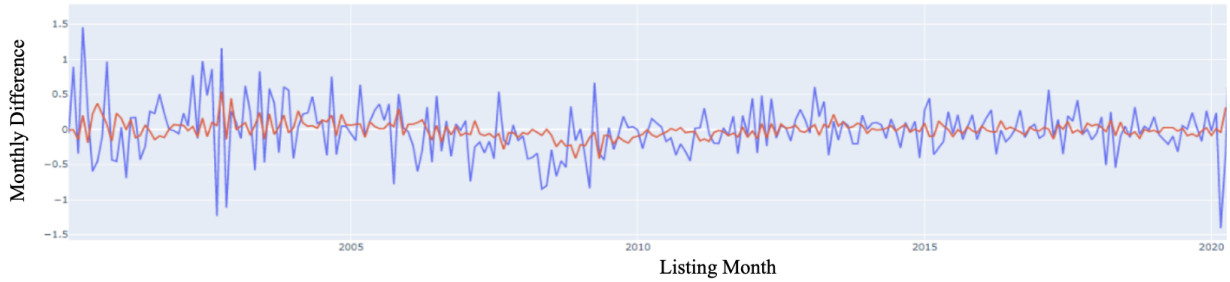

Figure 8.2.12: Plotted table of the model

8.3 Sprint Four Findings

During our fourth sprint, we focused on developing a Power BI report condensing all of our previous findings into one report. The visualizations in the report allow the user to choose a specific zip code and see the historical trends of the important housing factors we identified compared between the MLS, Redfin, and Realtor datasets. Figure 8.3.1 below illustrates an example, showing the new listing, and days on market trends for Worcester, Massachusetts.

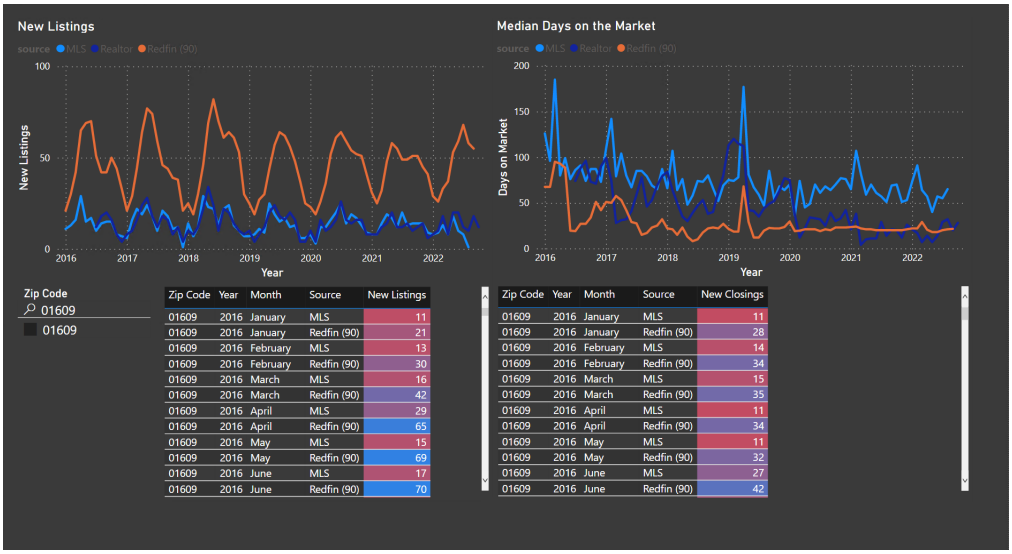

Figure 8.3.1: Power BI Visualizations of New Listings and Days on Market for Worcester, MA

In the Power BI report, we also created an interactive map, showing the year over year change in average percent difference between sale price and list price. The map shows the top 200 MSAs in the United States based on sales volume and allows users to select a specific month and year to get the results. Figure 8.3.2 below shows an example of the map during July 2021.

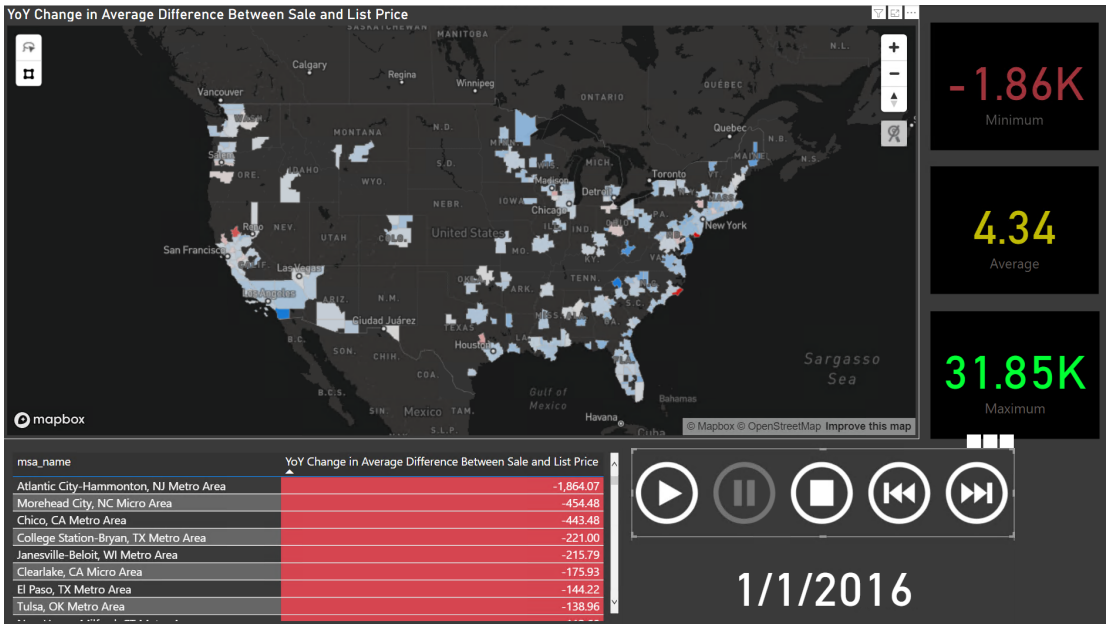

Figure 8.3.2: Map of YoY Change in Average Difference Between Sale and List Price in top 200 MSAs

We also continued to develop the ARIMA model, including an STL model for the ARIMA to better capture seasonality and trends, while also achieving a prediction. The sponsor gave a clearer objection and specified specific models and decomposition methods for the team. We were required to use STL (Seasonal and trend Decomposition using Loess) to decompose the data into three components: seasonality, trend, and residuals. Because the detrending methods used in the past week would not fulfill the new requirements, we started a new pipeline for median price prediction. Similar to last week, we started the model with the overview of the data:

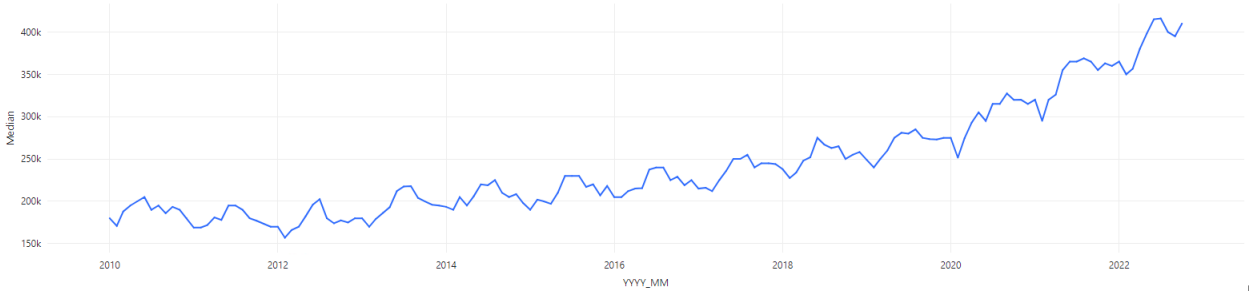

Figure 8.3.3: Overview of median price over listing month

To account for the non-stationarity for this time series, we performed the STL decomposition to break it into seasonality, trend, and residuals. The result is shown below:

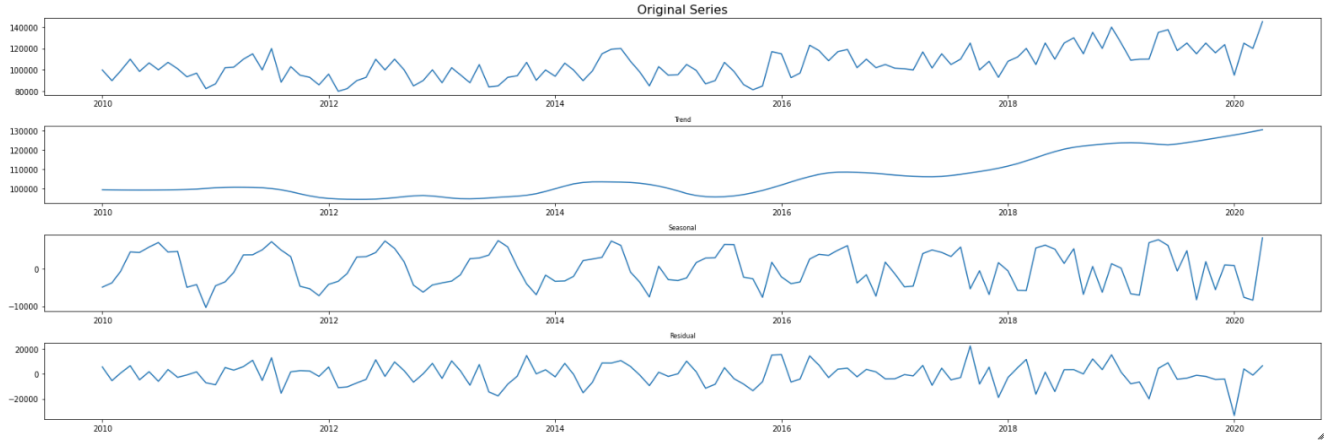

Figure 8.3.4: the STL decomposition result

Figure 8.3.4 above shows the breakdown for the time series. After breaking down the time series, we still needed to ensure that the residuals are stationary in order to proceed. Again, the Augmented Dick-Fuller Test was applied, and the result is shown below:

```
Monthly stats result:
1. adf: -11.582261404704099
2. P value: 2.9236457245940043e-21
3. Lags: 0
4. Observation used: 151
5. Critical Value: {'1%': -3.4744158894942156, '5%': -2.880878382771059, '10%': -2.577081275821236}
The dataset is stationary, please proceed!
```

Figure 8.3.5: Result for Augmented Dicky-Fuller Test

With p value less than 0.05, we can reject the null hypothesis that the residuals are not stationary. The next step is to feed the decomposition into the ARIMA model.

Figure 8.3.6 below shows the actual vs predicted median prices of homes in Worcester, MA. The blue curve represents the actual values, while the orange curve represents the predicted values. From the figure, we can see the separation of the predicted and actual values during the pandemic months. The predicted values fit well with the actual values throughout the years, until 2020, when the pandemic began. Although the trend remains the same, the median predicted prices are lower than the actual prices after 2020, showing the disruption that the pandemic caused in the real estate market.

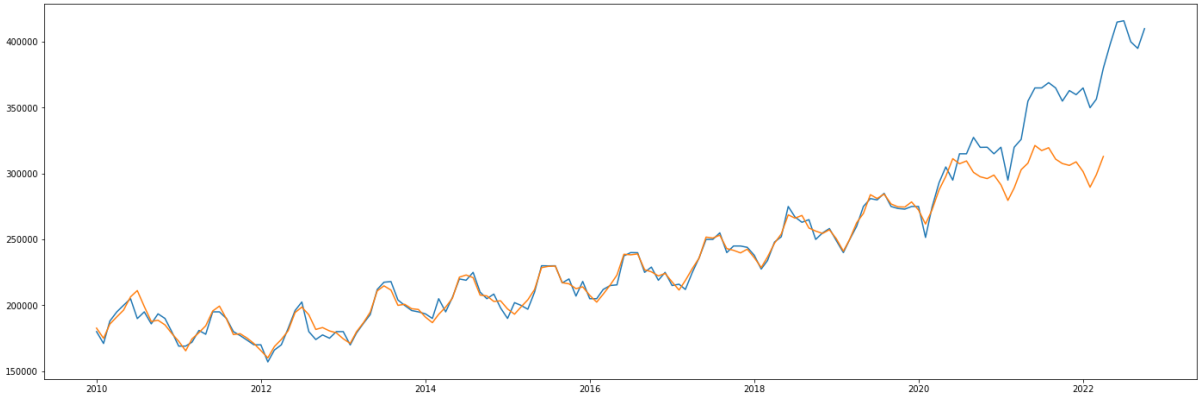

Figure 8.3.6: Actual vs Predicted Median Prices of Homes in Worcester, MA

Figure 8.3.7 below shows another visualization of the actual vs predicted median prices of homes in Worcester, MA, but different from the previous figure, this model continues the prediction up to year 2024. Based on the predicted values, the median prices of homes in Worcester, MA will continue to increase considerably, following the momentum created during the past couple years.

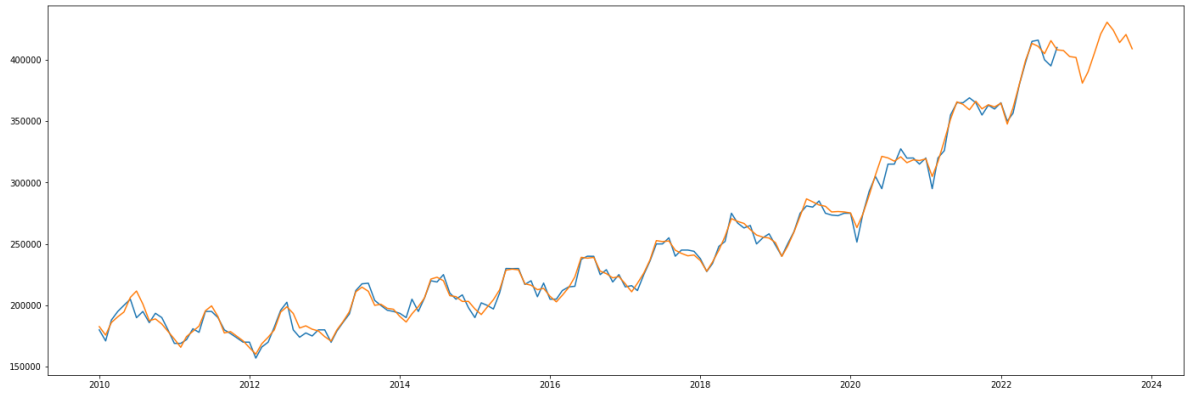

Figure 8.3.7: Actual vs Predicted Median Prices of Homes in Worcester, MA with Forecasting Included

The team also performed a similar analysis on the "real value" of the closing price using the Case-Shiller Index. The Case-Shiller Index measures the change in the value of the US residential housing market by tracking the purchase prices of single-family homes [63]. The Case-Shiller Index modified close price, the "real value", is calculated by the median closing price of that month divided by Case-Shiller Index value that month. This was done to help isolate and identify price movements that occurred outside of normal market movement. Throughout the COVID-19 pandemic, for example, the housing market increased more rapidly in comparison to past years. By normalizing the data according to the Case-Shiller Index it gave us an indicator of price movement that is either outpacing or underperforming the COVID housing market. Below is a Case-Shiller Index Modified median example:

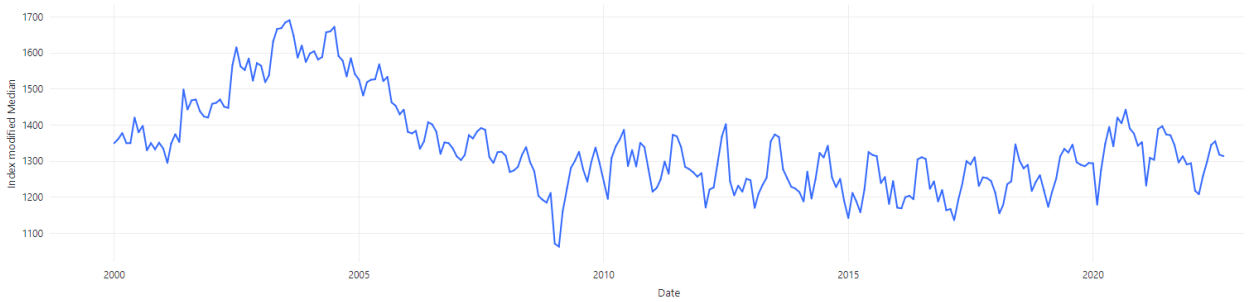

Figure 8.3.8: Case Shiller Index Modified Median vs. time / "real value" vs. time

After the model ran through all 200 top metropolitan statistical areas (MSAs), the team found that there were a fair amount of data points missing. Almost all the missing data points happened between 2000 and 2005, usually leading to a large gap of 6 months to 1 year. The predictive model failed because of the many instances of non-consecutive data with a large gap. Because of this missing pattern in the data, we developed a method to account for this. First, gaps of the missing data are generally large. Second, those missing data points always happened in earlier years when MLS was not fully developed. The team's solution was to find the latest 3-month gap and record its time, then remove all data points before this time. This did not significantly influence our prediction since the model itself is small and the removed data points and time stamps were very old, between 2000 and 2005. Here is an example of the dataset and model before and after cleaning (Orlando-Kissimmee):

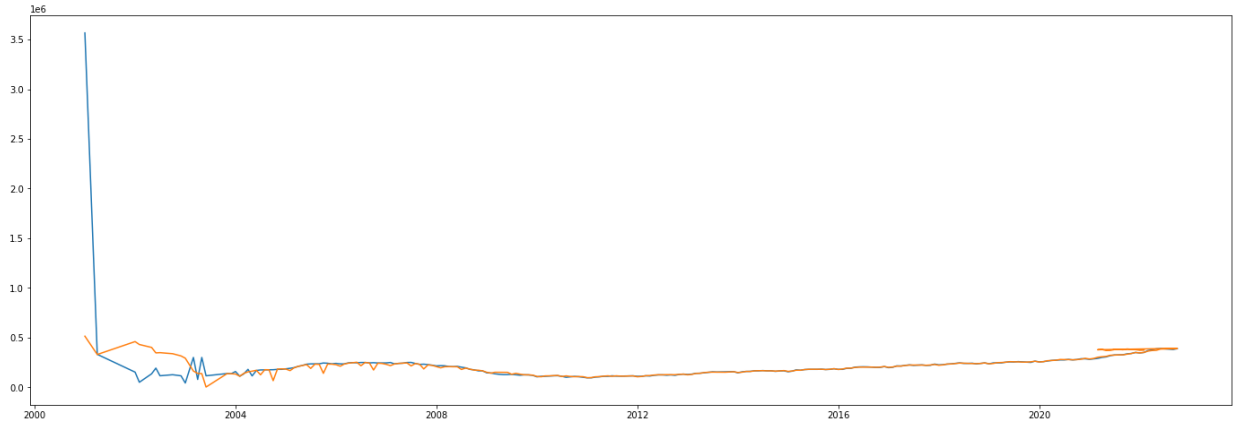

Figure 8.3.9: Dataset and model before removing the non consecutive part

Figure 8.3.10: Dataset and model after removing the non consecutive part

8.4 Social Impact of Real Estate

The real estate industry has been evolving into a competitive sector of the overall economy, accounting for 15-18% of GDP in the United States [38]. The impact of real estate in society has been continuously beneficial for a significant number of people, especially investors, who have created multi million dollar businesses with the different types of real estate investment opportunities. On the other hand, a considerable number of people have not had

similar experiences, due to supply and demand, which determines house prices and interest rates, imposing barriers to entering the real estate market.

8.4.1 Hurricane Effects in Shaping Demographics

Hurricanes can have a major influence on the real estate industry in the United States. Homes in certain areas of the country can range from one-hundred thousand dollar homes to multi-million dollar homes. The housing markets in which these disparities exist respond differently when it comes to the aftermath of a hurricane depending on the value of homes. The lower priced houses are built in accordance with the building code requirements that exist in the area, while the more expensive ones usually exceed those building code requirements. After a hurricane, the homes built exceeding the building code requirements show incredible resistance to the impact of hurricanes, being an example of how homes should be built in the area. After evaluating these differences, oftentimes local authorities impose new building code requirements, pushing homeowners or new construction investors to spend more in reinforcing materials for homes in order to meet the new code requirements, similar to the situation with Hurricane Michael in 2018, as described in chapter two. This incentive contributes to increasing home prices in the area, creating a challenge for many people trying to rebuild their destroyed houses after a hurricane. The situation becomes even more challenging when homeowners do not have insurance. Flood insurance in particular is a form of insurance that is not required for most homeowners, unless the house is located in a special hazard area as defined by FEMA. For those homes outside these special areas, flood insurance and other specific insurances are more expensive, which is why many homeowners try to avoid paying extra for insurance, ending up in worse situations.

Because of new building code requirements pushing prices up and challenges with insurances, many people leave the area and do not come back. This creates new supply from those who sell their homes, but also increases demand, because of the interest that investors express in investing in these areas, especially in convenient areas of Florida, where taxes are low, the weather is favorable throughout most of the year and the attractions that the region offers are persuasive. As middle to low income class people leave the area as a result of increased prices, big investors who come in the area for investment create an inequality, because they set a new standard of people who can afford to live in these areas. If this event continuously happens, then the demographics of the population living in these areas and the areas nearby starts changing.

8.4.2 The Social and Ethical Perspective of Real Estate

While looking for trends in the real estate market, we found that social factors have an impact in the real estate industry. These factors influence the real estate demand and therefore set pricing standards in particular areas. According to a Redfin survey, affordability, safety and cost of living are among the top factors that influence people's decisions on where to live. But other factors such as strong gun control laws, legal abortion and voting rights protections, make up significant portions of this survey [27]. As an investor, these additional factors should be evaluated, in order to make the best decisions when it comes to real estate. Apart from the house size, year built, and other common housing elements, these social factors can be of significant importance for producing demand, and thus driving returns on investments.

Among the factors listed above, affordability seems to be the most emphasized factor, not only from homebuyers and investors, but also from analysts, economists and writers. Housing prices have had an incremental increase over the years, making the sector a safe place for

investors seeking return on investment. But these continuous rises in prices have made it difficult for middle and low income class individuals to enter the market. The debate is ongoing as "there are people who obviously advocate for affordable housing, but there are other people who worry about their property values" [59]. Not only are the prices increasing, but with increasing property values, rents also follow the trend and go up, significantly raising the challenges for middle to low income individuals.

According to Redfin's data center, the average monthly rent for a single family residential property in the United States is around $2,000 while the mortgage payment with a 15% down payment, stands at around $2200, just $200 more than rent [60]. This is great for landlords who rent out properties, because they only pay about 10% of the mortgage themselves, but on the other side, the lessee is paying 90% of the landlord's mortgage, which is an incredibly high percentage. The question that arises from an ethical perspective is: who is at fault? Are landlords and investors morally drifting or is it just how the economy has unfolded with the evolution of capitalism? Clearly the answers from these questions will vary in accordance with the advocacy of the person being asked.

In the United States, about 40% of houses are mortgage free, which implies that homeowners who rent out those houses are able to make around 80-90% profit, depending on the contracts [61]. If these homeowners would rent out their apartments considering a smaller profit, the prices for rent would come down considerably, as it would decrease the demand for expensive rental properties, by creating supply for more affordable options. But considering the purpose of renting out apartments, which is for as much profit as possible, landlords justify their rising prices based on rising home values. From a utilitarian perspective, the landlords have the advantage, as currently the landlord-renter ratio stands at 66-34% respectively [62]. Until the

culprit is found together with a solution, the debate will still be ongoing and renters will still pay 90% of the homeowner's mortgage.

8.4.2.1 Rent Control

The role of governments in business is viewed as unethical by many economists and supporters of the free market, although it is oftentimes necessary, especially for industries like pharmaceuticals or electrical, which have to be restricted in order not to become monopolies or get involved in unethical behavior when maximizing profits [11]. The case of government involvement in restricting rent prices, known as rent control, is a topic that is yet unclear whether it is ethical and necessary to have. Rent control usually prevents homeowners from changing the prices of rent. This way a constant price is maintained for the duration of the whole contract. As of 2022, there are six states that currently have rent control policies: California, New York, New Jersey, Maryland, Oregon, and Minnesota.

Supporters of rent control base their arguments on the protection that rent control gives tenants. In the absence of restrictions, landlords can engage in price gouging. "Price gouging is defined by an increase in rents that (I) cannot be attributed to increased maintenance costs, (II) exceeds rents at which landlords receive a 'fair' return on investment." Although the free market competition can usually impose natural restrictions to the prices of rent for landlords, supporters of rent control are concerned of potential collusions among landlords [11].

Those opposing rent control are oftentimes described as part of the underlying ethical theory of Sedgwick's egoism. The theory states : "the rational agent regards the quantity of consequent pleasure for himself as alone important in choosing between alternatives of action" [10]. People against rent control, other than supporting the free market, argue that rent control is

bad for the overall economy, because they create apartment shortages, housing deterioration and less potential investments.

The issue of rent control will be continuously a moral dilemma, even if the inconsistencies that exist in this case are removed. Until a sort of optimal solution is reached, landlords and tenants have to remain cautious and adaptable to government intervention in rent prices.

8.4.3 Company's ESG efforts

The alternative investment firm we are working with has recently joined a company that helps investment firms strengthen their ESG (Environmental, Social and Governance) environment. The partnership is very recent, which makes it hard to collect information on the progress, but through the technologies offered by the company they joined, the investment firm will be able to track and analyze their ESG data, achieving a more sustainable and inclusive form of capitalism, which is the mission of the joined company.

Generally, companies invest in ESG with the purpose of screening investments in compliance with corporate policies, while also encouraging the company to act responsibly. "Environmental criteria consider how a company safeguards the environment, including corporate policies addressing climate change, for example. Social criteria examine how it manages relationships with employees, suppliers, customers, and the communities where it operates. Governance deals with a company's leadership, executive pay, audits, internal controls, and shareholder rights [64]." Part of the social criteria is also the company's ethical behavior and their support of rights, diversity and inclusion.

The benefits of a firm investing in ESG principles are improvements to the reputation of the company, increased attraction of investors who are actively being more cautious of the ESG principles that companies follow and also investing in ESG holds the company accountable for any drifting behavior from the principles, that may have greater consequences in the future. This depends on the evaluation and whether it is measurable and realistic, which is why partnering with an ESG company is the best route to evaluate and improve the company's ESG principles. On the other hand, as with all types of investments, ESG investing has its disadvantages. It comes at a premium which companies pay, making it oftentimes expensive for companies. Around “74% of respondents said that valuation/price was ‘very or extremely important to them’”, which imposes a restriction on many companies trying to invest in ESG [64].

8.5 Business Values and Project Risk Management

8.5.1 Risk Mitigation

Factors that influence the real estate industry bear risks with them, therefore detailed analysis is crucial to avoid risks. When investing in real estate, diversification is the best risk mitigation option. From diversification of market analysis methods, to resources and even investment types and locations, this is probably the only way to create elasticity and resist market fluctuations.

In our team, we used the MLS datasets provided by our sponsors to draw conclusions from our analysis and research. Then we compared the data to other platforms like Zillow or Redfin, providing the same or at least statistically similar data, in order to see how our findings lined up with the findings from these platforms. This shows an example of mitigating risk

through diversification of resources. By verifying our main datasource with other data sources, the analysis became much more thorough and risk was reduced.

Another risk that influences real estate is the geographic locations of the properties and the politics that surround them. States operate differently, therefore some can be more interesting to real estate customers than others, based on taxes, employment rates, new developments and even factors like safety and politics. To get the best return on investment, real estate investors have to carefully consider the political landscape of an investment at certain periods of time. Working with our sponsors, we noticed the geographical spread of investment properties that our sponsors hold. This gives them the elasticity to mitigate risks, in cases of a downturn in certain areas where their properties are located. An example goes back to a potential hurricane situation. If a certain number of properties in a hurricane area experience significant damages, then the cash flow coming from investments in other states can be used to subsidize the other properties that were affected by the hurricane.

8.5.2 Business Risks, Rewards and Values of the Project

The real estate market is a widely studied field, for which many companies collect data to analyze and share with the public or with private parties. Risks of this project are present anywhere from the collection of the data from sources, to analysis of the data and forecasting. The amount that is allocated for investments derived from this project's findings, could be referred to, from a monetary value perspective, as an indicator of the amount of risk that the project maintains.

If the project is successful, the company will use the results and findings to make strategic investment decisions in the future. If the forecasting based on the research and analysis

of the project is successful, then the corresponding investments will add to the company's assets and therefore increase revenue.

Our team has been working with the data provided and produced insights from research and analysis of the data, comparing them to other data providers like Zillow or Redfin. These insights are beneficial for the company not only to make investment decisions, but also to know the current state of the market, potential trends and signals on a national level and/or more specific areas.

9. Assessment

Throughout the course of this project, we stuck to the good AGILE practices layed out in our methodology. This included daily stand-up meetings where we shared progress, got feedback from a member of our sponsor's team, and determined the direction for the day. We also conducted sprint planning, review and retrospective meetings for each sprint, which consisted of assigning and ranking user stories, evaluating progress, and noting potential improvements for the next sprint. We had to adjust the start and end days of our sprints to line up with our mid-week presentation check-ins with our sponsors, which ensured that our sprint planning meetings would accurately cover the stories that needed to be completed for each sprint. Sticking with AGILE helped us to adequately compile useful fields from the datasets that our sponsors provided us with and to create visualizations that would best help the sponsors, as well as implement mathematical and statistical models to the data. Regular contact with our sponsors helped us get a good understanding of the project and adjust when necessary.

Early in the project, our goals and tasks were fluid, as we started by focusing on the effects of hurricanes on real estate and then moved to looking at real estate hot spots across the US. Large-scale changes in our project's direction made it difficult to determine an end goal. At first, we expected to determine correlations and make predictions, but we realized that putting our efforts into providing visualizations and models of the datasets that we had would best help with data analysts. Our team was new to some of the software tools that we were using, but we picked them up quickly (namely, Databricks and PowerBI). Connecting Databricks tables to PowerBI through Azure was necessary to visualize the data properly. PowerBI enabled us to

show our results in an organized report, which was hard to do as nicely in Databricks. Overall, we are confident that our results will be of aid to our sponsors.

9.1 Business takeaways

9.1.1 Risk Culture

During our project, working with the sponsors, we were able to notice the risk culture that the company maintains. We were able to access data and do our own research, coming up with results to present to upper level management. Although we were able to easily access information and use information, risk prevention and mitigation practices were always implemented.

One of the most noticeable risk preventative practices was the diversification of information and resources seen throughout the whole duration of the project. When presented with findings from one source of information, the sponsors required a comparison of the findings with other relevant resources that provided similar information. This risk preventative method is great because it exposes potential mistakes throughout research or simply the special cause variations that could be noticed in between different resources, indicating risk.

Another practice that we noticed throughout the project is the use of our findings. We noticed that there are levels of “checkpoints” in the company’s management, where information from research is digested in accordance with the roles and departments, where the information is presented. We presented our findings on a daily basis to our team manager, then weekly to upper level management. After the end of the project, the results will be analyzed further from upper level management and will be modified and digested as needed. This hierarchical type of risk

mitigation method is great, because information is passed through different layers of security and is analyzed from different levels of management, bringing their perspective and critical analysis, before the information is used for decision making.

When it comes to investments, the company considers all possible risks and they try to avoid areas of high investment risks, by choosing safer investment strategies. One example is when investing in distressed debt. There are two types of debt investments which companies use for investment: subordinated or junior debt and senior debt [12]. A managing director in distressed and corporate special situations at the company, says the firm invests primarily in senior debt, which has a higher priority of repayment compared to junior debt. The priority that senior debt maintains compared to others, is what makes it a safe investment, minimizing the risk of potential losses in cases of bankruptcy of the firm invested in.

9.1.2 Additional Risks

The project we have been working on will be used in the future from the company to make investment decisions based on the research findings. This automatically implies financial risks, followed by reputational risks, because, as with every investment, the possibility of a downturn is never zero. If the company experiences multiple financial downturns, their reputation could be damaged. Market risks are also present in this project. As we have seen in the past couple years, the real estate market has been the center of attention for many investors, due to large fluctuations throughout the pandemic and after. Although these fluctuations in the market generate great opportunities for investments, the market risks that arise with them are significant.

9.2 Team's Learnings

Throughout our project, we were presented with insights from the real estate industry, for which we now have developed a deeper understanding. From the various analyses and research conducted throughout the project, we learned some of the major components of the real estate market; the connections and relationships between the housing sector and social, economic and political factors; and we learned about the ways that companies involved in the industry operate when it comes to investigating opportunities and investing.

We also learned about different types of investments that investment firms are involved in, such as distressed debt investments and equity investments. Speaking with a managing director of distressed investments at the firm gave us a better overview of distressed debt investments and the benefits of investing in one type of debt over the others from a risk perspective.

From the conversation, we also saw how the company leaders communicate and inform each other. The conversation was possible thanks to a managing director, who was able to arrange a virtual meeting in just a couple days from the notification moment. Overall, the employees of the firm have always been available and quick to respond, which shows great communication throughout the company. A great technique that the company uses for quick and concise communication is email allies based on the different departments, which saves a lot of time among employees when requesting or sharing information throughout the company.

From a technical perspective, we developed new skills and understandings, when applying software and data to derive business findings for the alternative investment firm. We used a variety of methodologies and data types throughout the project from Databricks to PowerBI reports, as well as developed mathematical models to support our findings and make

forecasts. We learned how to navigate through big datasets and filter down what was needed to make the project successful. By doing so, we were exposed to data science and data analytics experiences in a real world environment outside our academic domains.

Apart from learnings coming directly from the company, we also developed new skills working remotely and with hands-on experience. The difficulties of working remotely provided great time and team management skills. We all had to adjust our schedules to make time for individual work, group work and meetings with advisors and sponsors, while adapting to changes in our sprint retrospectives and plannings. A crucial part of our team work was the flexibility to changes and the quick adaptability to the new schedules or project paths. Remote access was also new to us, but we quickly adapted to the environment with the help of our sponsors. Accessing files and data from the company's software was a bit challenging at the beginning, but as we continued using the remote access everyday, we learned how to properly navigate through the platform and the firm's files.

Overall, completing the project has enriched us with a variety of new experiences and skills that will directly apply to our careers in the future. From time and team management skills, to organizational, communication and reporting to management. This overall experience has given us close insights on how the investment firms operate and how the market is analyzed. Upon completion, we look forward to carrying these learnings over and applying them to our career in the future.

10. Future Work

Continued work on this project would open up doors for integrating more datasets, increasing the accuracy of our findings, and developing further analyses. Additional datasets can provide more data points that MLS does not cover and also provide other data fields to supplement MLS records with. Obtaining more data points is a trivial way to increase the accuracy of our findings and help the analysts make strong investment decisions.

Additional work can also be done on the hurricane analysis portion of our project. Specific trends that would be good to look into are insurance rates/policies and government regulations in relation to hurricanes and how they affect price and demand. This opens new doors for our sponsors to make investment decisions on aside from the surface level real estate price and listing trends. More findings can also be put together from the data that we were provided in the Redfin hotness dataset. This could be utilized more to find the factors and demographics that correlate with booming real estate markets. Some ideas to investigate are the number of businesses nearby, crime rates, construction, traffic, and distance to water and other amenities like gasoline and grocery stores.

Further trend identification can also be done on the impacts of the COVID-19 pandemic. Comparison of pre and post COVID trends can show expected returns to “normal” as well as prolonged market changes since the pandemic took effect. Data visualizations pinned around these time periods would be of aid to the sponsors.

On the predictive model using the ARIMA, it might need some final fine tuning with all parameters to get the optimum result. We can compare the R-square statistics of different models in different areas and find out the most accurate parameters for the model.

Overall, the housing market is broad and there are many ways to analyze real estate investment opportunities. Identifying common locations and policies for good investments by listings and prices is a surefire way to identify scalable and meaningful investments.

11. Conclusion

Given Corelogic's MLS dataset, we were able to identify key housing market indicators and trends in the housing market. We were able to compare this dataset to another housing dataset from Zillow in order to statistically validate the trends that we saw in the dataset. This allowed us to conclude that we could safely use the MLS dataset in our deliverables as a representative measure of the housing market.

Within the MLS dataset we identified five key housing indicators. The housing indicators that we identified were both listing and closing price, number of new listings, the ratio between the closing price and its original listing price, and the median number of days that a listing stayed on the market. After identifying the indicators, we created a series of interactive visualizations in a PowerBI dashboard. Within the dashboard we included visualizations of the historical performance of these indicators from three of the datasets available to us: Corelogic's MLS dataset, Redfin's zip code-level dataset, and Realtor.com's zip code-level dataset. This dashboard is an incredibly valuable research tool for the analysts from the alternative investment firm that we worked with to form investment theses.

In addition to the housing indicators that we identified, we were able to create a forecasting model for the median sale price of various housing markets. Although the model fails to catch the trends during the COVID period, it catches the majority of the seasonal components of the data. Therefore, it is an useful indicator of the median market closing price in respect to seasonality.

All in all we can say that the analyses that we have done, and the results that have been produced as a result of them, are of significant use to the alternative investment firm. They will

be able to utilize our work in the future to reliably assess potential investment decisions and bring greater returns to their investors.

References

1 - *Many Floridians Hit Hardest by Hurricane Ian Can't Afford to Rebuild*. (n.d.). Time. Retrieved October 28, 2022, from https://time.com/6223550/hurricane-ian-recovery-florida-too-expensive/

2 - Smith, M. B. (2022, October 19). *Weathering the Storm: Will Hurricane Ian Upend the Florida Insurance Landscape?*CoreLogic®. https://www.corelogic.com/intelligence/weathering-the-storm-will-hurricane-ian-upend-the-florida-insurance-landscape/

3 - CoreLogic. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=CoreLogic&oldid=1106434427

4 - *Hurricane Ian exacerbates housing market slowdown in Southwest Florida | Seeking Alpha*. (n.d.). Retrieved October 25, 2022, from https://seekingalpha.com/news/3893282-hurricane-ian-exacerbates-housing-market-slowdown-in-southwest-florida

5 - *Hurricane Ian: How Storms Are Transforming Florida's Coastal Property Market*. (n.d.). WSJ. Retrieved October 28, 2022, from https://www.wsj.com/story/hurricane-ian-how-storms-are-transforming-floridas-coastal-property-market-b942395d

6 - *About Us | J.P. Morgan Asset Management*. (n.d.). Retrieved October 28, 2022, from https://am.jpmorgan.com/us/en/asset-management/adv/about-us/

7 - *Blackstone*. (n.d.). Retrieved October 28, 2022, from https://www.blackstone.com/the-firm/

8 - *OakTree Capital*. (n.d.). Retrieved October 28, 2022, from https://www.oaktreecapital.com/about

9 - *Paragon Commercial Group Closes $43 Million Sale of South Bay Grocery Anchored Retail Center | Canyon Partners*. (n.d.). Retrieved October 28, 2022, from https://www.canyonpartners.com/strategies/real-estate/press-releases/paragon-commercial-group-closes-43-million-sale-of-south-bay-grocery-anchored-retail-center/

10 - Phillips, D. (2013, December 9). *Sidgwickian ethics – an overview*. Revue d'études benthamiennes. Retrieved December 7, 2022, from https://journals.openedition.org/etudes-benthamiennes/669

11 - Somchith, S. (2018, March 21). *Business in ethics – the ethics of rent control*. LinkedIn. Retrieved December 7, 2022, from https://www.linkedin.com/pulse/business-ethics-rent-control-sydney-somchith/

12 - Tarver, E. (2022, November 7). *Subordinated debt. vs. senior debt: What's the difference?* Investopedia. Retrieved December 8, 2022, from https://www.investopedia.com/ask/answers/061615/what-difference-between-subordinated-debt-and-senior-debt.asp

13 - *What Is Distressed Debt Investing? | HBS Online*. (2021, August 5). Business Insights Blog. https://online.hbs.edu/blog/post/distressed-debt-investing

14 - *What Is Arbitrage? 3 Strategies to Know*. (2021, July 20). Business Insights Blog. https://online.hbs.edu/blog/post/what-is-arbitrage

15 - *Manifesto for Agile Software Development*. (n.d.). Retrieved September 20, 2022, from http://agilemanifesto.org/

16 - *History: The Agile Manifesto*. (n.d.). Retrieved September 20, 2022, from http://agilemanifesto.org/history.html

17 - Atlassian. (n.d.-b). *Agile Scrum Roles*. Atlassian. Retrieved September 20, 2022, from https://www.atlassian.com/agile/scrum/roles

18 - Atlassian. (n.d.-d). *Scrum Sprints: Everything You Need to Know*. Atlassian. Retrieved September 20, 2022, from https://www.atlassian.com/agile/scrum/sprints

19 - Atlassian. (n.d.-i). *What is a Scrum Master?* Atlassian. Retrieved September 20, 2022, from https://www.atlassian.com/agile/scrum/scrum-master

20 - Atlassian. (n.d.-h). *User Stories | Examples and a Template*. Atlassian. Retrieved September 26, 2022, from https://www.atlassian.com/agile/project-management/user-stories

21 - Atlassian. (n.d.-e). *Sprint Planning*. Atlassian. Retrieved October 28, 2022, from https://www.atlassian.com/agile/scrum/sprint-planning

22 - Atlassian. (n.d.-g). *Three steps to better sprint reviews*. Atlassian. Retrieved October 28, 2022, from https://www.atlassian.com/agile/scrum/sprint-reviews

23 - Atlassian. (n.d.-a). *Agile retrospectives: Use the past to define the future*. Atlassian. Retrieved October 28, 2022, from https://www.atlassian.com/agile/scrum/retrospectives

24 - Atlassian. (n.d.-f). *Standups for agile teams*. Atlassian. Retrieved October 28, 2022, from https://www.atlassian.com/agile/scrum/standups

25 - *Spoiled for Choice: 10 Cities Where the Number of Homes for Sale Is Dramatically Rising*. (2022, August 23). Real Estate News & Insights | Realtor.Com®. https://www.realtor.com/news/trends/2022-where-inventory-increasing-the-most/

26 - Ellis, T. (2022, November 4). *Housing Market Update: Demand Declines Ease as Mortgage Rates Steady Around 7%*. Redfin Real Estate News. https://www.redfin.com/news/housing-market-update-demand-declines-ease/

27 - Anderson, D. (2022, November 4). *Redfin Survey: Housing Affordability Is On Voters' Minds As They Head to the Polls*. Redfin Real Estate News. https://www.redfin.com/news/midterm-election-housing-survey-2022/

28 - *Apache Spark™ - Unified Engine for large-scale data analytics*. (n.d.). Retrieved November 11, 2022, from https://spark.apache.org/

29 - pandas documentation — pandas 1.5.1 documentation. (n.d.). Retrieved November 11, 2022, from https://pandas.pydata.org/docs/index.html

30 - *Spark NLP*. (n.d.). Retrieved November 11, 2022, from https://nlp.johnsnowlabs.com/docs/en/production-readiness

31 - This is where the real action in artificial intelligence takes place. (n.d.). *Washington Post*. Retrieved November 11, 2022, from https://www.washingtonpost.com/news/the-switch/wp/2016/06/09/this-is-where-the-real-action-in-artificial-intelligence-takes-place/

32 - *Project Jupyter*. (n.d.). Retrieved November 11, 2022, from https://jupyter.org

33 - Atlassian. (n.d.-c). *Jira | Issue & Project Tracking Software*. Atlassian. Retrieved November 11, 2022, from https://www.atlassian.com/software/jira

34 - Webex. (n.d.). *Webex App*. Webex. Retrieved November 11, 2022, from https://www.webex.com/all-new-webex.html

35 - *Multiple Listing Service (MLS): What Is It?* (2012, January 11). Www.Nar.Realtor. https://www.nar.realtor/nar-doj-settlement/multiple-listing-service-mls-what-is-it

36 - Corporate - About. (n.d.). *Zillow*. Retrieved November 11, 2022, from https://www.zillow.com/z/corp/about/

37 - Zillow Home Value Index Methodology, 2019 Revision: What's Changed & Why. (2019, December 19). *Zillow Research*. https://www.zillow.com/research/zhvi-methodology-2019-highlights-26221/

38 - *Housing's Contribution to Gross Domestic Product*. (n.d.). Retrieved November 14, 2022, from https://www.nahb.org/news-and-economics/housing-economics/housings-economic-impact/housings-contribution-to-gross-domestic-product

39 - *Is Real Estate Investing Safe?* (n.d.). Investopedia. Retrieved November 14, 2022, from https://www.investopedia.com/articles/investing/122415/why-real-estate-risky-investment.asp

40 - Duggan, W. (2022, October 28). *Is the U.S. headed for another recession?* Forbes. Retrieved December 1, 2022, from https://www.forbes.com/advisor/investing/is-a-recession-coming/

41 - *Mann Whitney U Test (Wilcoxon Rank Sum Test)*. (n.d.). Retrieved November 17, 2022, from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html

42 - *Data Center Metrics Definitions*. (n.d.). Redfin Real Estate News. Retrieved November 21, 2022, from https://www.redfin.com/news/data-center-metrics-definitions/

43 - *Realtor.com Real Estate Data and Market Trends for Download*. (n.d.). Realtor.Com Economic Research. Retrieved November 21, 2022, from https://www.realtor.com/research/data/

44 - *Mann-Whitney U Test: Assumptions and Example*. (n.d.). Informatics from Technology Networks. Retrieved November 28, 2022, from

http://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

45 - Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

46 - Dickey, D. & Fuller, Wayne. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. JASA. Journal of the American Statistical Association. 74. 10.2307/2286348.

47 - *What Is Autocorrelation?* (n.d.). Investopedia. Retrieved November 28, 2022, from https://www.investopedia.com/terms/a/autocorrelation.asp

48 - Huitema, Bradley & Laraway, Sean. (2006). Autocorrelation.

49 - *2.2 Partial Autocorrelation Function (PACF) | STAT 510*. (n.d.). PennState: Statistics Online Courses. Retrieved November 28, 2022, from https://online.stat.psu.edu/stat510/lesson/2/2.2

50 - Hyndman, R. J., & Athanasopoulos, G. (n.d.). *Forecasting: Principles and Practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

51 - Sangarshanan. (2019, April 7). *Time series Forecasting — ARIMA models*. Medium. https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06

52 -*Introduction to ARIMA models*. (n.d.). Retrieved November 28, 2022, from https://people.duke.edu/~rnau/411arim.htm

53 - Sanchez-Vazquez, M. J., Nielen, M., Gunn, G. J., & Lewis, F. I. (2012). Using seasonal-trend decomposition based on loess (STL) to explore temporal patterns of pneumonic lesions in finishing pigs slaughtered in England, 2005–2011. *Preventive Veterinary Medicine*, *104*(1), 65–73. https://doi.org/10.1016/j.prevetmed.2011.11.003

54 - *ARIMA Processes*. (n.d.). Retrieved December 1, 2022, from https://www.youtube.com/watch?v=lpXeed6Y6uM&list=PLf7Of__eXS8ILDGH3RDHZj932sa0TaJkS&index=10

55 - Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Statistic Sweden*, *6*, 3–73. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-de

56 - Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, *6*(1), 3–33.

57 - BERACHA, E., & PRATI, R. S. (n.d.). *How Major Hurricanes Impact Housing Prices and Transaction Volume* (1st ed., Vol. 33, p. 12). Retrieved October 25, 2022, from https://cre.org/wp-content/uploads/2017/04/Hurricanes.pdf

58 - Atlassian. (n.d.). *The product backlog: Your ultimate to-do list*. Atlassian. Retrieved December 1, 2022, from https://www.atlassian.com/agile/scrum/backlogs

59 - Constant, P. (2021, August 21). *A real estate CEO appraises the home ownership inequality problem in the US - and how to fix it*. Business Insider. Retrieved December 5, 2022, from https://www.businessinsider.com/how-to-fix-homeownership-inequality-problem-in-us-2021-8

60 - *Data center*. Redfin Real Estate News. (2022, November 18). Retrieved December 6, 2022, from https://www.redfin.com/news/data-center/

61 - Richardson, B. (2022, October 12). *Nearly 40% of homes in the U.S. are free and clear of a mortgage*. Forbes. Retrieved December 6, 2022, from https://www.forbes.com/sites/brendarichardson/2019/07/26/nearly-40-of-homes-in-the-us-are-free-and-clear-of-a-mortgage/?sh=32a5e5a647c2

62 - U.S. Census Bureau. (2022, November 2). *Quarterly residential vacancies and homeownership, third ... - census.gov*. QUARTERLY RESIDENTIAL VACANCIES AND HOMEOWNERSHIP, THIRD QUARTER 2022. Retrieved December 6, 2022, from https://www.census.gov/housing/hvs/files/currenthvspress.pdf

63 - Fontinelle, A. (2022, July 13). *Understanding the case-shiller housing index*. Investopedia. Retrieved December 8, 2022, from https://www.investopedia.com/articles/mortgages-real-estate/10/understanding-case-shiller-index.asp

64 - Team, T. I. (2022, November 22). *What is environmental, social, and governance (ESG) investing?* Investopedia. Retrieved December 9, 2022, from https://www.investopedia.com/terms/e/environmental-social-and-governance-esg-criteria.asp#toc-what-is-environmental-social-and-governance-esg-investing

65 - Zajic, A. (2022, November 29). *What is akaike information criterion (AIC)?* Built In. Retrieved December 9, 2022, from https://builtin.com/data-science/what-is-aic

Appendix A

-Remote Access

During the PQP term, we began contacting our sponsors, scheduling weekly meetings to prepare for the beginning of the project. During this phase the remote access was requested and the team members completed a background check form for the sponsors to investigate further. After the background check was complete from the sponsors, we were given an email and access to company computers remotely, using the app “Workspot”.

The remote access enabled us to work with the MLS dataset, PowerBI report and other necessary platforms used, while communicating directly with sponsors and other employees of the company. Most of our work was completed through remote access, which made it easier for everyone at the company to review and give us feedback.