Analyzing Migration Trends through Credit Card and Foot Traffic Data

---

A Master Qualifying Project submitted to the Faculty of Worcester Polytechnic Institute in partial fulfillment of the requirements for the Degree of Bachelor of Science.

Project Team:
Katherine Deyette
Lena Sophia Thompson
Hunter Trautz

Project Advisor:
Professor Robert Sarnie
Department of Business

Project Advisor:
Professor Wilson Wong
Department of Computer Science

Abstract

Through a collaboration with a prominent investment firm, a machine learning model was built to predict human migration patterns to inform real estate investments. The Agile Scrum methodology was applied in managing this project. Software development was primarily done using Python and Azure Databricks. Our work focused on two sets of data, which we attempted to combine using various statistical methodologies. Although this was not feasible, the data sets were still useful training sets for our model. The machine learning algorithm developed has the potential to be a great asset to the alternative investment firm, as it is able to predict the population within the United States to a high degree of confidence, even with the limited amount of initial training data.

Acknowledgments

We would like to thank the people who helped make our project possible. Firstly, we would like to thank our sponsor, the alternative investment firm that we worked at. They have given us invaluable learning experiences and our first taste of the FinTech world. We would also like to thank Garrett Stephenson and Jake Needleman for their consultation on leadership. A very special thanks to Connor Dietz for his support and consultation throughout the writing process. Finally, we would like to thank our advisors Professor Robert Sarnie and Professor Wilson Wong for their guidance and advice throughout the whole FinTech experience.

Executive Summary

Over the course of seven weeks, we were tasked with creating a machine learning algorithm with the goal of being able to predict human migration within the United States. We began our project by analyzing the data sets that would serve as the training data of the machine learning model to ensure the dependability and validity of the sets. We used the data sets in several data validation exercises such as normalization. This analysis was performed using Python in Azure Databricks Notebooks. Our sponsor intends to use the model we developed as a tool in their real estate investment portfolio. While the work we did took the 2019 SARS-CoV-2 pandemic into account and was affected by it, the alternative investment firm intends to use our model as a foundation for building a long term strategy, even after the pandemic is over.

The software, analysis strategies and recommendations developed were presented as a resource to be used by the alternative investment firm. In the future, new data sources may be integrated to continue to build out the model. These new data validation techniques and insights may eventually complement existing investment strategies as time progresses and the data sets are improved. Our analysis of the data sets concluded that while they are not able to be combined, they are still useful training sets individually, we have verified that the data is sound. This discovery enabled us to continue with the project and develop the machine learning algorithm that would predict human migration across the United States.

Our team was able to create a Random Forest machine learning model that can predict the future migration patterns of humans in the United States, and will inform the alternative investment firm's future financial investments. The model predicted the population of the top two hundred counties in the United States using the training data provided to it. The training data consisted of the population of each county from 2018 to 2019, as given by the United States Census, and the number of bars and full service restaurants from SafeGraph for those years. Lastly, the model outputs its prediction for the population of each county based on the number of bars and restaurants recorded in SafeGraph so far in 2020. We found that the accuracy of the predictions created by our model were extremely promising when compared to one other online source that produced predictions for the same counties.

To best optimize our time during the seven weeks we had to complete the project, we used the Agile Scrum methodology which would help keep us on track yet also enabled us to adapt to any challenges should they happen to appear. We learned how to adapt to unpredictable project changes and we will continue to use our developed skill of flexibility as we go on with our careers. These skills are even applicable outside of a software development context, and we considered how our learnings are relevant to other engineering disciplines. This capstone project saw the integration of business management and software development paradigms, and the following report presents our process and findings throughout it.

Authorship

Section	Primary Author	Primary Editor
Abstract	Lena Sophia Thompson	
Acknowledgements		
Executive Summary		
1. Introduction	All	
2. Background	Lena Sophia Thompson	Katherine Deyette, Hunter Trautz
2.1 Customers		
2.2 Real Estate Investments		
2.3 Associated Risks		
2.4 Risk Management		
2.5 Domain Research	Katherine Deyette Hunter Trautz	Lena Sophia Thompson
2.6 Machine Learning		
2.7 Data Science and Data Engineering		
3. Software Management Methodology	Lena Sophia Thompson	
3.1 Development Methodology	Hunter Trautz	Lena Sophia Thompson
3.1.1 Risk Management	Lena Sophia Thompson	
3.2 Alternative Methodology		
3.2.1 Systems Engineering		
4. Software Development Environment	Katherine Deyette, Hunter Trautz	Lena Sophia Thompson
4.1 Project Management Software		
4.2 Programming Environment		
4.2.1 Azure Databricks		

	Normalization			
	8.3 Data Analysis			
		8.3.1 Yodlee vs. Google Trends and Population		
		8.3.2 SafeGraph vs. Yodlee		
		8.3.3 SafeGraph vs. Population		
		8.3.4 Developing a Model with SafeGraph		
	8.4 Benefits of the Analysis			
9. Assessment			Lena Sophia Thompson	
	9.1 Business Learnings			
	9.2 Technical Learnings		Katherine Deyette, Hunter Trautz	Lena Sophia Thompson
	9.3 Accomplishments			
	9.4 Mechanical Engineering Connections		Lena Sophia Thompson	
10. Future Work			Katherine Deyette, Hunter Trautz	Lena Sophia Thompson
11. Conclusion				
Report Compilation and Formatting			Lena Sophia Thompson	

Table of Contents

Table of Tables

Table of Figures

1. Introduction

The FinTech industry has become a significant field, due to its attempt to automate financial processes that would otherwise require greater expenditure of company resources. The inclusion of data analytics can benefit companies such as investment firms, especially those specializing in real estate. Through the understanding of migration trends, provided by data sets such as credit-card transactions and geolocation data, investment firms can improve their investment decisions. The analysis done can be used to signal high-valued properties that will provide the best return on investment for the firm. Specifically, the firm is investigating the use of historical migration trends to predict which regions have emerging value. Though this strategy of predicting future hotspot regions is certainly not new to real estate investors, the 2019 SARS-CoV-2 pandemic caused significant and observable inflection points in migration patterns. These inflection points present valuable opportunities to develop new prediction models.

The alternative investment firm recently acquired key data that could be used to represent migration patterns of humans within the United States. Migration trends were derived from credit card transaction data, and customer visitation frequency at "places of interest" (POIs) collected using geolocation data. From this data, the firm observed that there were significant and anomalous changes in migration patterns due to the pandemic. This project was an example of the alternative investment firm's risk cultural mindset that favors novel and alternative strategies. Their intent is not just to leverage these emerging patterns for short term predictions, but also to use it as a foundation for building a long term strategy, even after the pandemic is over.

The initial task was to determine if the given data sets could be combined to form a unified migration metric over time. Towards that goal, the alternative investment firm wanted us to derive a more precise representation of historic customer visitation frequency by location.

Then, using either this new metric or one of the original data sets as a basis, a prediction model of migration patterns in the United States would be built. A machine learning model was a prime candidate, due to its ability to learn and improve as more migration data is acquired, yielding confident predictions that can be used in the firm's real estate acquisition techniques. Python in Azure Databricks was used to develop the prerequisite analysis and machine learning model.

The software, analysis strategies, a machine learning model, and recommendations developed were presented as a resource to be used by the investment firm. Furthermore, these new data validation and analysis techniques and insights may eventually complement existing investment strategies as time progresses and the data sets are improved. Agile project management and a multitude of software development techniques were successfully implemented throughout this project to coordinate with the alternative investment firm, complete our analysis, and develop our prediction model

2. Background

The alternative investment firm is a privately held investment firm that manages approximately $39 billion in assets through cutting edge credit and real estate strategies. They focus on research, investment, protecting capital, money with opportunity, and their multi-strategy to be a successful company. With many locations in Europe, Asia, and the United States, the firm employs around five hundred people. The alternative investment firm was founded in 1988 and launched their distressed debt, convertible, and merger arbitrage as their first investment strategy. In 1993, the firm expanded into the real estate industry by purchasing the distressed debt of a bankrupt real estate company which then led them to begin investing in opportunistic real estate. Looking for international real estate opportunities, they then considered opening a location in Asia, where there was a booming growth-market. The Asian markets presented the alternative investment firm with new opportunities for high risk-adjusted returns, which led them to apply their United States opportunistic real estate investing model. In 2011, the alternative investment firm's mortgage investment trust was listed on the New York Stock Exchange. This trust is externally managed and advised by a subsidiary of the alternative investment firm, which was formed as a publicly traded real estate investment trust that invests in, acquires and manages a diversified portfolio of residential mortgage assets, other real estate-related securities, financial assets and real estate.

2.1 Customers

The alternative investment firm's customers consist of wealthy investors looking to grow their money by contributing to a fund that is low risk and high return on investment. Their competitors are in the liquid real estate sector, privately owned alternative investment management, and distressed securities and private equity. In 2017, the alternative investment firm introduced their liquid credit strategy. It was introduced to invest in adjacent areas of credit markets—predominantly High Yield and Investment Grade—with an aligned, fundamental approach and conservative, downside protection orientation. Recently, on August 14, 2020, the alternative investment firm filed their 13F form, which is their quarterly report revealing that they currently own 2,000,000 shares of FinTech Acquisition Corp. valued at $20,460,000. The value of their investment has grown 22.84 percent during last quarter alone. FinTech Acquisition Corp is a special purpose acquisition company that has no operations but plans to go public with the intention of acquiring or merging with a company utilizing the proceeds of its initial public offering.

2.2 Real Estate Investments

Just like any other investments, real estate investments do not have a guaranteed return and take time to have an increased value. With real estate, investments are in physical assets; they are tangible investments. Stocks and other market securities are just that, securities and contracts that represent ownership of some fractional value of a market entity. While most other

kinds of investments have capital gain taxes, real estate investments also have associated real estate taxes. Additionally, real estate investments can have development costs if the intention is to increase the value before selling. However, despite all of these additional costs, physical assets are less likely to suddenly become worthless in value. In other words, real estate investments are considered to be less volatile than stocks and other intangible investments.

2.3 Business Risks

Making investments in an atypical market, for example, due to the SARS-CoV-2 pandemic, carries high financial risk. Additionally, there is an innovation risk as valuable resources are being spent to develop a new investment strategy. New investment strategies may be complex, more sensitive to market conditions, or may have mechanisms of action that are difficult to understand. Any of these factors could mean that the time and money put towards the venture would be wasted. However, if the alternative investment firm makes poor real estate investments based on an unproven strategy or software, customers of the firm could, in the worst case, reconsider their involvement with the company. As with any investment, real estate does not carry a guaranteed return, though it is more stable. However, the 2019 SARS-CoV-2 pandemic caused financial markets, including real estate, to become more volatile than historic trends (Trading Economics, 2020).

Independent of the financial application of this project, any outsourced software development introduces more risk. In this project, the alternative investment firm is contracting development work to us. Whenever new people are brought onto a project, there are many integration factors that must be considered. Companies have existing culture, methodologies, and standards. Onboarding new members brings the risk of them not producing work in line with those standards, which can make it difficult for existing employees to continue work. There is also a training overhead if new members do not have necessary background knowledge, or do not know how to use company tools. Training in particular introduces potential risk vectors as it can be difficult to assess how well it has gone. If it was not successful, or if more training time is needed, more valuable resources must be allocated to account for it.

The background for this project is highly technical, and the investment strategy under development is itself a large risk. However, it should be noted that the alternative investment firm is mitigating some of the risk by removing much of the monetary resource requirements associated with new project members. Though the investment strategy poses potentially large financial risks, the only major upfront cost to the alternative investment firm is the time spent with us by our sponsor liaison.

A project like this also has technical, operational, and security risks. The alternative investment firm is giving a third party, us, access to confidential financial data. Though we would never do anything malicious or in breach of non-disclosure agreements, there is always the potential information misuse from the perspective of the alternative investment firm. Additionally, giving third parties access to sensitive databases brings the possibility of accidental deletion or otherwise destructive accessing of the database. Handling sensitive personal

information also means strict privacy laws must be observed. Furthermore, these privacy laws can change rapidly, so it is possible that mistakes could be made, or information in breach of these laws is inadvertently shared. It is unlikely that the alternative investment firm will not have backups of this data and strict security measures in place, though it is still important to note.

2.4 Project Risks

The primary mechanism the alternative investment firm has in place to manage risk is their Risk Management Group (RMG). This department deals in customer information reporting, and acts as a throttle to the operations of the front office. While the RMG is responsible for ensuring the front office is not taking risks that are too large, they do act as a final authority on their actions, and allow them a high degree of autonomy. This check and balance system facilitates the firm's risk cultural mindset that risks are an opportunity to make money. For example, the firm has used the distress investing strategy, where shares of a company currently deemed worthless on the market are purchased with the belief that said company will eventually regain its former value. Additionally, the firm has had success purchasing debt, generating over half a billion in revenue from debt of a prominent global financial services firm purchased at 0.1 on the dollar. In general, the alternative investment firm does not just target long term investment strategies, but also frequently employs alternative investment strategies.

The data represented just a small sample of the population of the United States, and it is not continuously updated in real time. Both of these factors could potentially make the data unreliable. The alternative investment firm was able to offer some insight into the data, mentioning that it was not necessarily conclusive as the data has only been collected for a few years and had some inaccuracies. Additionally, the data may have already been out of date as it could no longer be updated in real time. In this way, it was important for us to evaluate whether or not the data was actionable, but with the risk that inputs were not as accurate as we would have liked, their conclusions from these data sets may be inaccurate.

2.5 Domain Research

In order to draw conclusions from migration patterns of humans, we needed the data that track their location. The alternative investment firm had already identified two types of data sets that could provide that information involving a data set focused on points of interest and a data set that focuses on purchases. It is important that both data sets undergo multiple layers of normalization and a process that involves filtering out potential data outliers and that their associated vendors are dependable and reputable companies. A data set that focuses on purchases has data on whether the purchase was made online or offline and from whatever retailer the purchase was made. If a purchase was made offline at a specific location, there should be a correlating data point, which is collected through cell phone geolocation. Through combining these types of data sets, we would be able to identify individual user transactions and user habits. An identifying factor of gentrification of an area is the opening of low initial cost industries, such as bars. Another good indication of foot traffic is grocery store visits, since food and other

essentials are constantly needed by the general population. These insights gave us insight in researching population migration trends.

2.6 Machine Learning

2.6.1 Ensemble Learning Method

The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms. This learning method has two types of techniques, the Boosting Technique and the Bootstrap Aggregation, i.e., the Bagging, Technique. The Bagging Technique involves random sampling with replacement, allowing the developer to understand the bias and variance of the data through multiple, smaller, random samples of the data set. Each of these smaller samples are used as training data, creating multiple decision trees that run independently of each other. The outputs of each decision tree are aggregated at the end into one model, without any preference for any tree (Medium, 2020a). An example of how trees are developed can be seen in the following section.

Combining multiple decision trees into one model can lead to some problems. One is that the decision trees for each of the models are sensitive to the data set it is trained on, leading to predictions that are significantly different from each other. Another problem with using multiple decision trees is that they require a lot of computation, yielding longer run times, especially with the increasing number of decision trees as the data set increases in size (Medium, 2020a).

2.6.2 The Random Forest Regression Model

The random Forest Regression Model uses the Bagging Technique since the decision trees run independently and have no interaction between the trees as they are being built, until the results are aggregated. Each decision tree returns either a statistical mode of the classes for classification, or a mean prediction for regression. The model aggregates the results of the decision trees through two main methods. The first being that the number of features in the training data set that can be split at each node of the decision tree is limited by a hyperparameter, which is some percentage of the total number of features. This prevents the model from relying on an individual feature in an attempt to prevent outliers from the individual predictions of each decision tree. The second method of aggregation the Random Forest Regression Model undergoes is ensuring that each tree draws from a random sample of the original data set when generating the splits at each node of the tree, adding an additional layer of randomness and preventing some overfitting (Medium, 2020a). An example of the Random Forest Regression Model is seen in *Figure 2.1.*

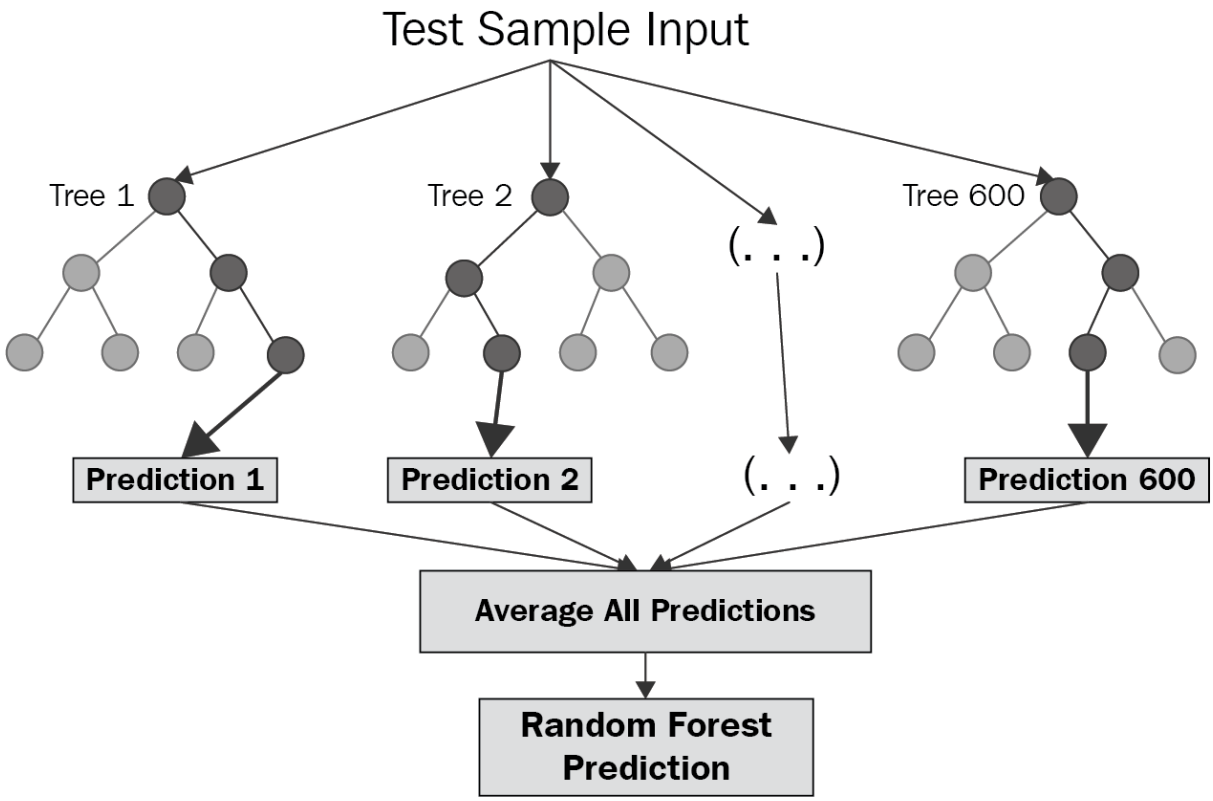

Figure 2.1. Decision Trees and Random Forest Regression Model (Medium, 2020a)

2.7 Data Science and Data Engineering

Data engineering is a separate but closely related field to data science. In fact, data scientists and data engineers often work together. The difference between the two is that data engineers are responsible for cleaning or scrubbing the data while the data scientists are responsible for taking that data and producing statistical analysis that is valuable to their company (Furbush, 2018). Data engineers typically come from a programming background due to the knowledge of distributed systems required. This is opposed to the heavily math and statistics background required by data scientists to create advanced analytics (Anderson, 2018). The development team took on both the role of data scientists and data engineers during the project term.

3. Project Management Methodology

3.1 Development Methodology

Agile project management is an iterative approach to managing software development projects that focuses on continuous product releases and incorporating customer feedback with every iteration (Atlassian, 2020). Scrum is a framework for Agile that software and technology industries utilize due to its structure of roles, meetings, and rules (Cprime, 2020). Teams usually consist of about seven people, and the goal of the team is to deliver a releasable product at the end of each "sprint," the name given to the fixed-length iterations (Scrum Reference Card, 2014). As shown in *Figure 3.1*, each team uses a "product backlog" to manage the tasks they have to complete. The backlog consists of user stories, which represent features that are important to the product's user base and appear in the prototype. At the beginning of each sprint during the "planning meeting," the team chooses a reasonable amount of achievable and pertinent user stories to complete by the end of the sprint and moves these to the sprint backlog. The team will assign story points to each user story using the Fibonacci numbers for scale (1, 2, 3, 5, 8, 13, 21) depending on the difficulty of the task.

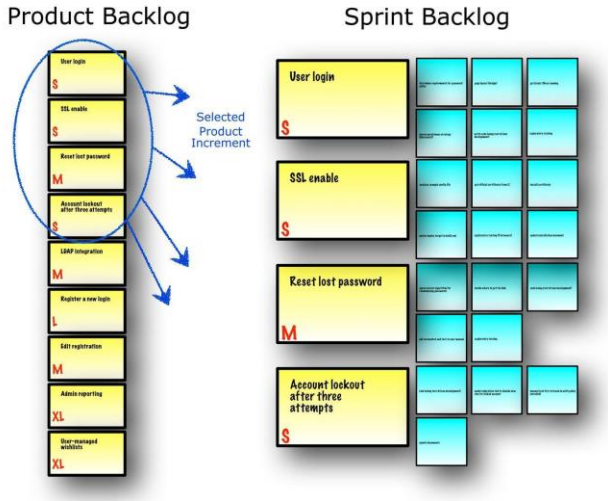

Figure 3.1. Agile Scrum Product and Sprint Backlogs (Scrum Reference Card, 2014)

Each member of the team has one of three roles:

- Developer
- Product owner
- Scrum master

The product owner considers the interests of stakeholders by creating and constantly re-prioritizing the product backlog. The scrum master is in charge of leading the team's daily meetings, or "daily scrums", and is responsible for removing obstacles that arise during the discussion. At this meeting, each member must take turns speaking about what tasks they completed the previous day, what tasks they will complete today, and any issues that they are

currently experiencing. The rest of the team members are primarily developers and focus on collaborating to finish the user stories for that sprint. At the end of each sprint, the team presents and demonstrates their product to the stakeholders in the "scrum review meeting" to receive crucial feedback. This feedback will then be converted by the product owner and scrum master into new product backlog items for the product owner to prioritize. Lastly, the "sprint retrospective meeting," led by the scrum master, officially concludes each sprint in which the team reflects on its own process and how they can improve for the next iteration. It is crucial that the scrum master creates a safe environment where the team members feel comfortable expressing their concerns and suggestions for improvement. The process begins all over again at the start of the next iteration with another sprint planning meeting as shown below in *Figure 3.2*.

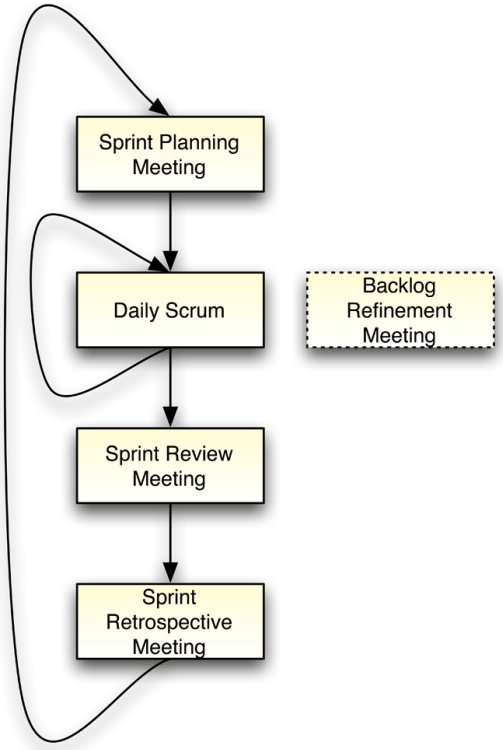

Figure 3.2. Scrum Flow (Scrum Reference Card, 2014)

An alternative and less structured project management framework for Agile is called Kanban. Unlike Scrum, Kanban does not have set length sprints but instead focuses on a continuous flow of delivery. The workflow follows a linear pipeline style, meaning user stories there are currently in progress must be completed before moving onto another task. There are also no roles within teams and continuous re-prioritization of user stories is encouraged even in the middle of the development period (Atlassian, 2020). One downside of being able to add tasks during a sprint is that it could cause a bottleneck in the team's ability to deliver cohesive software. A bottleneck occurs when there are too many individual user stories in progress at one time (Pipefy, 2020). The project team will maintain a Kanban board via Trello in order to visualize the project workflow and identity any bottlenecks should they occur.

The project team will utilize a mixture of the Agile Scrum methodology and Kanban to ensure the creation of a prototype that meets the needs of the project sponsor. The concrete deadlines and meeting structure of Agile Scrum suits the limited seven-week timeline of the project while Kanban will allow the team to adapt to the constantly changing needs of the sponsor at every stage of development.

3.1.1. Risk Management

We broke down the project risks into these categories based upon severity:

Human Capital - A risk that relates to the person hours or energy of the team.
Organizational - A risk that involves time or planning of the project.
Technical - A risk that involves software.
(Colors are those used throughout our risk management reporting)

For each sprint, we identified and categorized the associated risks, assigned a probability and current risk factor status for each one, and detailed the measures we took to mitigate them. We presented this information for each sprint using the format in the table below.

Table 3.1 Sprint Risk Management Template

Name	Description	Risk Category	Probability	Risk Status	Mitigation
[Risk 1 Name]	[Risk 1 desc.]	Technical	Medium	Low	[Risk 1 mitigation measures]
[Risk 2 Name]	[Risk 2 desc.]	Human Capital	Low	Low	[Risk 2 mitigation measures]
[Risk 3 Name]	[Risk 3 desc.]	Organizational	High	Low	[Risk 3 mitigation measures]
...	...	...	...	...	...

Both probability and Risk Status may have values of Low, Medium, and High, which represents an agreed upon severity by us and the alternative investment firm. The table also includes space for a description of each risk, as well as a name to keep track of risk status between sprints.

Each sprint risk management table we created helped us identify the present and future risks we would be facing. This helped us plan out our next steps for the coming week in addition to helping us realize what some former problem areas were so that we would be aware of them and fix it in the future. For example, “time” was a risk as we only had seven weeks to complete the project. This meant that the impact was high and the risk status was high as it was a risk that was unable to be changed and was constantly happening. One way that we used these tables to our advantage was by using them as a planning aid when assigning number values to each user story.

3.2 Alternative Methodology

3.2.1 Systems Engineering

One of our team members has a Mechanical Engineering concentration, so an analysis was done on how our work in this project is applicable to the physical engineering disciplines. In engineering design and manufacturing, more so than ever, designs of products and manufacturing processes are becoming increasingly data-driven. While this project is not concerned with developing a product or manufacturing process based on data analysis, the statistical methods developed in this project are applicable. Correlation and regression modeling are a key tool for identifying the causes of recorded failure modes of a system. The work we did to develop an optimized development strategy is similar to how the Six Sigma method would approach a process defect.

When designing something using engineering principles, such as a mechanical one, there are various methodologies that can be used. Systems Engineering (F.J. Looft) is one such set of guidelines. The basic process is:

1. Identify and Meet with Stakeholders
2. Develop Performance Metrics
3. Formalize Requirements
4. Outline System Context
5. Quantify/Qualify Risk
6. Operational Concept Development
7. Architect the Solution

These steps do not necessarily need to be performed in order, and some may not be necessary at all depending on the project. For this project, the majority of these steps have already been completed by our sponsor. Regardless of the application domain, these steps provide a solid foundation when developing a system or product. What follows is a breakdown of each of these steps to highlight the things that need to be considered before a project can begin. Most importantly, this systems engineering approach, though tailored to mechanical system design, is applicable to this project, too, as we will discuss in section 9.4 Mechanical Engineering Connections.

Any project has a number of parties with vested interests in it. The first critical step in designing a system is identifying these parties, or stakeholders, and determining how important each of their needs for the project are. It is important to keep the stakeholders updated with progress from the start so any issues can be addressed immediately.

Performance metrics are how the success of a project is measured. In other words, based on the needs of each stakeholder, is the product or system performing as intended? The formalization of requirements is a direct continuation of defining performance metrics; it establishes the specifications that the project must meet.

Regardless of what field one is working in, it is important to recognize how the project fits in with the context of its environment, and who and what it will affect. For example, building a wind farm might cause negative ecological effects, which must be considered alongside the

positive environmental impacts it will ultimately have. With any project, it is important to consider what risks are being taken by investing in it. Though it is impossible to anticipate every risk that may be encountered over the course of a project, more time spent on this early stage step will prevent unforeseen roadblocks or consequences down the line. In engineering, a common tool typically used is Failure Modes and Effects Analysis (FMEA). FMEA is a tool for identifying potential failure modes of a system and a framework for identifying the effects and severity of these failures (ASQ, 2020).

Operational concept development is the process of taking the initial requirements and then outlining how the product or system will be used. An example guiding question for this is: how will users interact with the product? Many times, the concept development of a project is already defined by the company or upper management, and this was no different for our project. We were given a goal by the company and were guided towards a solution.

Architecting a solution is the phase in which the actual engineering and design happens. This is where the product or system is developed, taking into account the previously determined requirements, risks, and use cases. Only after all of the previous steps are completed should one begin to architect a solution. The reason for this is that many physical engineering projects require designers and firms to navigate lengthy legal and safety validate processes. It could be costly if these steps need to be revisited during the implementation phase. Where Systems Engineering focuses on providing a framework for project setup, Agile provides a framework for incremental project completion. Though they are different and are intended for different project domains, Systems Engineering still shares many similarities with Agile. This is why we have included information on Systems Engineering. It provides important concepts to keep in mind when tackling a large project, and oftentimes software development projects fail due to short-sightedness, despite the fact that this could be avoided.

Identifying and meeting with stakeholders is an implicit step at the beginning of any project when using Agile Scrum. From there, both the performance metrics and requirements phase in Systems Engineering are similar to the planning phase of each sprint in Agile Scrum. There are a set of tasks that need to be completed for a sprint to be considered successful, and developers must know when they have been successful. This is especially critical as short sprints demand efficient use of time.

When each sprint ends, the user must identify the risks, analyze the possible impact said risk has both in the future and in the present. This requirement is outlined by the Agile Scrum methodology. However, the Systems Engineering approach requires the user to identify the problems and possible risks *before* the project even starts. The contrast between these two methods of management provide strengths and weaknesses. For example, the Agile methodology allows for the users to break down the risks and then analyze and complete the tasks to neutralize the risks during the process. System Engineering on the other hand allows for the users to see where the issues may lay during the project and then enable the user to predict and solve problems before they even happen. However, the weakness with Agile is that because it is not predictive, problems that arise may present to be too big of an issue to be addressed in one sprint

which is already in progress. This is the reason why sprints have retrospectives. Conversely, Systems Engineering does not account for problems that arise during a project which can make it difficult for a project to adapt to the problematic environment. When these two methods are combined, they are able to use their strengths to their advantage while also cancelling out their weaknesses.

4. Software Development Environment

4.1 Project Management Software

The project team used Trello for organizing and maintaining the product and sprint backlogs. The application is developed by Atlassian, who are known for their industry-leading products that are used by software engineers and companies across the globe. The interface of the application is intuitive and mimics that of a physical Agile Scrum board so that anyone can be comfortable using it regardless of their experience. Trello was a crucial tool for us throughout the duration of the project that allowed us to visualize the progress as well as prioritize the group's tasks. An alternative project management software that we could have used is Github Project Boards. The majority of software companies utilize Github Enterprise for version control of their applications. In fact, GitHub is the largest code host in the world recording 57 million active users as of December 2020 (Github, 2013). Github, in addition to providing tools for version control, also allows teams to create and maintain Kanban boards right on the project repository page. Github source control and Github Project Boards would have been our choice for project management had our work not needed to be completed on Azure Databricks to access the alternative investment firm's Azure Data Lake for data analysis.

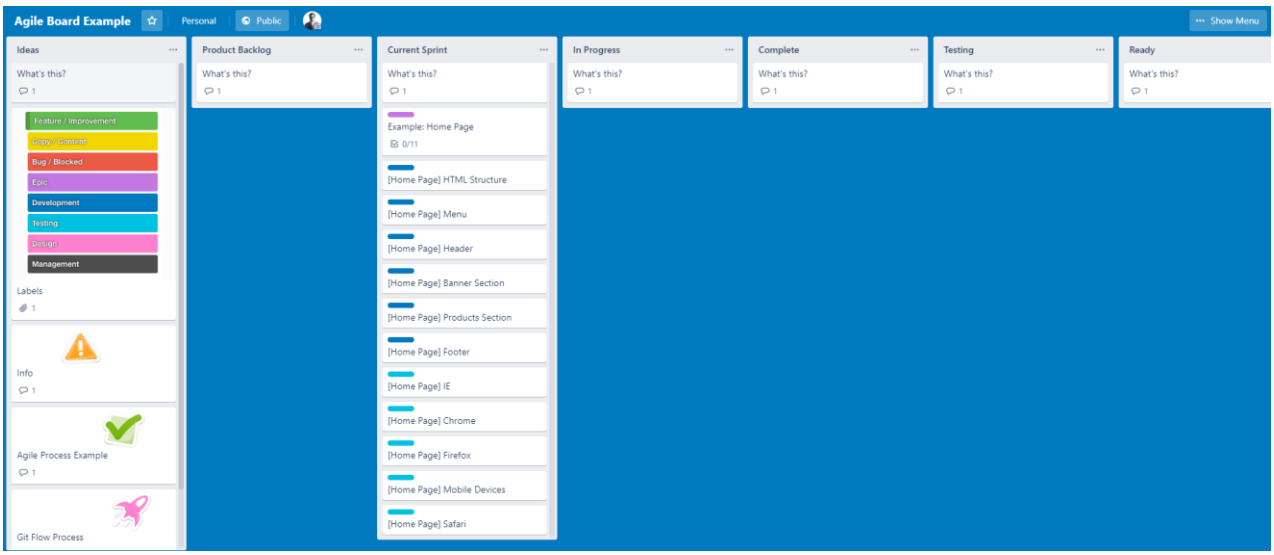

Figure 4.1. Agile Scrum Example Trello Board (Scrum Reference Card, 2014)

4.2 Programming Environment

4.2.1 Azure Databricks

Azure Databricks is a collaborative Apache Spark web-environment that interfaces with data sources stored in Azure Data Lake. In Databricks one can write Spark Notebooks using Python, SQL, Scala, R, or Java to access and manipulate data tables for the analysis. The

environment used in this project was hosted on a Python cluster with access to the Plotly library and other libraries to create and annotate graphs from the output of the notebook's queries.

4.2.2 VSCode

For any programming done outside of a PySpark Notebook, the project team used Visual Studio Code (VSCode) version 1.51 as an IDE. This was mostly a matter of personal preference as both of our team's programmers have had past experience using it and were knowledgeable with it. It is worth noting that VSCode provides a vast number of helpful extensions that are user friendly. Some examples are Microsoft's Python extension, Azure Data Lake Tools, and Azure Machine Learning.

4.2.3 Pandas Python Framework

Pandas (version 1.1.4) was utilized by the project team due to the improved performance and flexibility provided by the framework's DataFrame object compared to Apache Spark's DataFrame object. An additional bonus was that using Pandas DataFrames gave us direct access to the Plotly Express API instead of having to import the Plotly framework each time a graph was made. Furthermore, Pandas DataFrames can also be converted into NumPy arrays, which is the only data type accepted by the machine learning framework SkLearn, while Spark DataFrames cannot.

4.2.4 SkLearn Python Framework

Our linear regression model was built using the SkLearn framework (version 0.24). Population data used to train the model was retrieved from the USCensus.gov, and the number of unique grocery stores by county was found using SafeGraph. The model outputs a confidence interval (0-1) which was calculated by comparing its predicted population for each county to the true population for 2020 as given in the United States Census. It also outputs the square root of the mean squared error showing on average how much the predicted population was off from the true value.

4.2.5 Azure Data Lake

Azure Data Lake is the cloud data repository used by the alternative investment firm to store their SQL/PostgreSQL databases. There are many benefits associated with utilizing Azure Data Lake to host their many databases. These benefits include the reliability offered by a well-renowned company like Microsoft and the seamless integration of other data sources into one platform to proceed with visualization techniques assigned individually to each analysis. The data provided to us by the alternative investment firm for this project include United States Census data, SafeGraph geolocation data, Yodlee credit card transaction data, and miscellaneous data sets acquired from scrapping public websites including Google Trends.

5. Software Requirements

5.1 Software Requirement Gathering Strategy

The requirements for this project were determined by the advisors at the alternative investment firm, with whom we met daily. The project's focus was adjusted after the daily scrum meeting each day, taking into consideration the needs and suggestions given by the advisors. The first area of project work involved confirming the authenticity of SafeGraph and Yodlee data sets independently and then together.

5.2 Functional and Nonfunctional requirements

The primary functional requirement of this project was to analyze population trends, specifically of counties, states, and cities in the United States. In order for our machine learning model to meet the performance requirements, it needed to be trained on SafeGraph data obtained from business industries that have at least a R^2 value of 0.8 when comparing the number of points of interest (POIs) from each county to the population. This was so that the alternative investment firm would have the ability to make reliable business decisions.

A constraint of the project was that we had to use the data sets provided by the alternative investment firm. These data sets consist of credit card transaction data (Yodlee), geolocation data (SafeGraph), and population data (United States Census). However, we were free to incorporate additional data sets included in their research. Another constraint was that the project team had to access and perform all analysis of this data in Azure Databricks as it is stored on the alternative investment firm's Azure Data Lake server.

5.3 User Stories

Items that are placed in the product backlog are known as user stories because they are written in what is known as user story format. The format is as follows:

As a <type of user>, I want to <some goal> so that <reason for goal>

Epics are oversized versions of user stories that typically can be broken off into smaller tasks.

Table 5.1 User Stories

Sprint	User Story	Points
Epic: Project Organization/Business Involvement		
1	As a student, I want to review the project outline in order to have an idea on how to write an initial draft of the final paper	8
2	As a professor, I would like to see a Weekly Sprint Recap so that I can know the latest progress of the project.	2
2	As a student, I would benefit from meeting with the other alternative investment firm's team to discuss progress and compare experiences	1
4	As a student, I would like to revise the paper's outline based on the software development updates provided by the development team in order to attempt creating a polished overview for our sponsors.	8
6	As a student, I would like to complete the business portion of the paper in order to provide a sound business purpose to my team's developmental analysis to future readers	13
6	As a vendor/associate of AG, I would like to have a copy of the team's final presentation and polished final paper in order to have the team's findings organized for seamless handoff in order to continue the team's recommended next steps, as stated in the presentation and paper	8
Epic: Data Organization/Processing		
1	As an associate of the firm, I would like the content of each PySpark notebook to be broken up and organized by content to ensure clarity of the analysis done	5
1	As a developer, I need to make sure that normalization occurs before comparing the two data sets to ensure consistent and reliable analysis between two different data sets	8
1	As a developer, I would like to utilize Plotly for charting and presenting the team's findings so that they are accessible to non-developers.	2
2	As a database vendor, I want all of the analysis done consolidated into a single PySpark notebook presentation in search of the validity of the data in order to deliver a concise analysis	2
2	As a developer, I want to familiarize myself with the environments listed in order to be well-versed for my sponsor	3

3	As a database vendor, I want all of the analysis done consolidated into a single PySpark notebook presentation in search of the validity of the data in order to deliver a concise analysis	5
6	As a developer, I would like to migrate ML scripts into PySpark/Databricks so that it does not require a local development environment	5
6	As an investor, I would like a technical presentation regarding the team's 7 week analysis in order to continue their research in validating the data sets, leading to confidence in our population trends and predictions	13
Epic: Yodlee Validation		
2	As a vendor, I would like to see a comparison between online vs.offline shopping for Yodlee data in order to assist in making sound business investments	5
2	As an investor, I would like to see if Hardware Stores (based on NAICS 444130) is a strong indicator of migration patterns in order to make sound real estate investments	8
3	As a database vendor, I would like to see the correlation between the average credit card transaction amount vs.the number of Yodlee users in order to see if factors other than inflation influence the number of Yodlee users	8
3	As an investor, I want to be able to see which supermarkets exist in only one state in order to create credit card transaction analysis against the state's population growth	5
3	As a database vendor, I would like to see the correlation between the average credit card transaction amount vs.the number of Yodlee users in order to see if factors other than inflation influence the number of Yodlee users	8
3	As a database investor, I want to understand why Yodlee users are being double-counted, so getting the monthly aggregated data from the vendor is necessary in order to compare credit card transaction users vs.the population of Texas	5
3	As an investor, I want to be able to see which supermarkets exist in only one state in order to create credit card transaction analysis against the state's population growth	8
Epic: Yodlee vs. SafeGraph		
1	As an investor, I want to compare foot traffic between different regions for each several locations in order to make sound regional investments	21
2	As an investor, I would like to see the comparison between the number of grocery stores (number of H-E-B and number of stores in general) vs.population by county in TX in order to see if we can predict population growth by the number of grocery stores	13
2	As a database vendor, I want to compare the number of Yodlee users to the number of SafeGraph users in order to check the validity of both Yodlee and SafeGraph	8
3	As a database vendor, I would like to check the validity of both Yodlee and SafeGraph by comparing the number of Yodlee users to the number of SafeGraph users and make sure only the SafeGraph users above the rolling average of users by location are included in order to reduce unnecessary "noise" and outliers	13
5	As an investor, I want to see if the correlation between Yodlee and SafeGraph users would be improved by normalizing using the home_cbg column in order to validate using both data sets to	8

	predict migration patterns	
Epic: SafeGraph Validation		
3	As an investor, I would like to see the comparison between the number of grocery stores (number of H-E-B and number of stores in general) vs.population by county in TX in order to see if we can predict population growth by the number of grocery stores	8
2	As an investor, I would like to understand the 2019 SafeGraph user drop-offs to confirm the authenticity of the data set.	5
Epic: Machine Learning		
4	As a database vendor, I want our users to update POI queries to use current_poi tables/yr instead of monthly_patterns in order to accommodate our normalization techniques	5
4	As a real estate investor, we would like to predict the population of a state by county or city using ML correlation matrices for different industries trained using 2018-2019 sg_poi data in order to predict population and make sound real estate investments	21
4	As a real estate investor, I would like to know how accurate the models are by calculating confidence intervals for the ML regression model in order to determine if the analysis will be useful for my investments.	8
4	As a real investor, I would like to know how the model works by having the team create documentation for the ML regression model in order to maintain the model after the project has ended or for a future project team.	3
4	As a real estate investor, I would like to see if the model works on other industries/states by expanding the number of industries and locations used in the regression model in order to predict population in other regions	3
5	As an investor, I would like have confidence in the model being used to predict population and investigating its overfitting issue in order to have more confidence in the regression model	8
5	As an investor, I would like to use bars, restaurants, gyms to predict population because of the low cost barrier	8
6	As an investor, I would like to use the regression model to predict the population of the United States top two hundred counties so I can predict populations of the top counties in order to make sound business investments.	21

6. Design

6.1 Software Frameworks and Architecture

Our final deliverable was an Apache Spark Notebook, containing Python scripts and SQL queries, that interfaced with the alternative investment firm's SafeGraph table from Azure Data Lake to train a Random Forest Regression Model. Our model was created using the SkLearn Python framework's RandomForestRegressor function and is trained on the number of bars and full-service restaurants from the top two hundred counties in the United States during the years 2018 and 2019. After the model is trained, it makes a prediction of each county's population based on the number of bars and full-service restaurants recorded in SafeGraph during 2020. SkLearn then outputs a graph showing the accuracy of each of its predictions as well as the mean squared error value. The rest of the notebook contains graphs that further analyze the accuracy of the predictions and highlight discoveries we made regarding the SafeGraph data set. The graphs were created using the Python framework Plotly, for which we had to convert our data frames using the Python framework Pandas. A diagram of the software architecture can be found below in *Figure 6.1*.

We also provided seven other notebooks that were created during the course of the project term that contain analysis done on both the Yodlee and SafeGraph data sets independently and when trying to join the two. Visualizations of the data flow for each data validation exercise can be seen in *Figures 6.2, 6.3, and 6.4*

Figure 6.1. Project Team's Software Development Stack

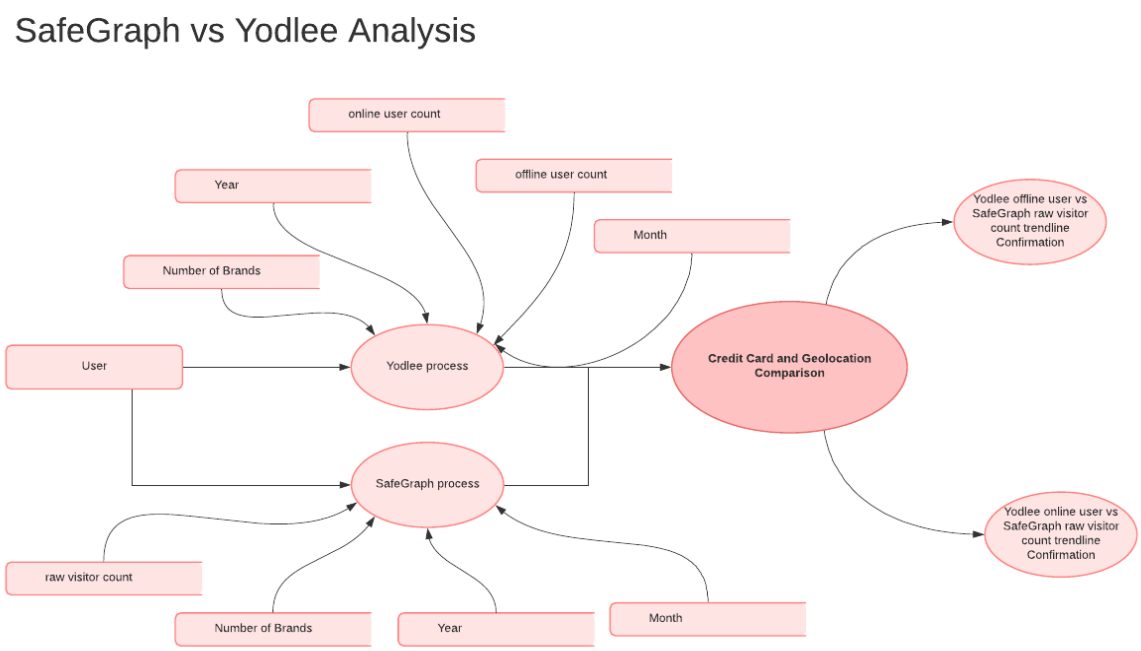

Figure 6.2. Data Flow of SafeGraph vs. Yodlee Analysis

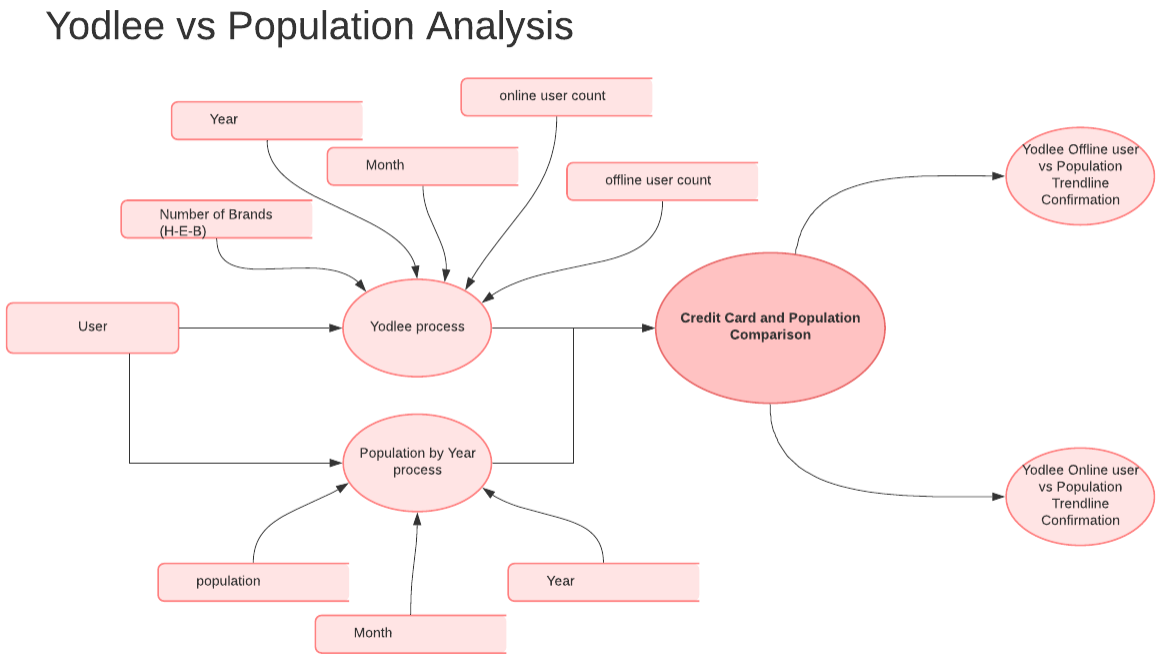

Figure 6.3. Data Flow of Yodlee vs. Population Analysis

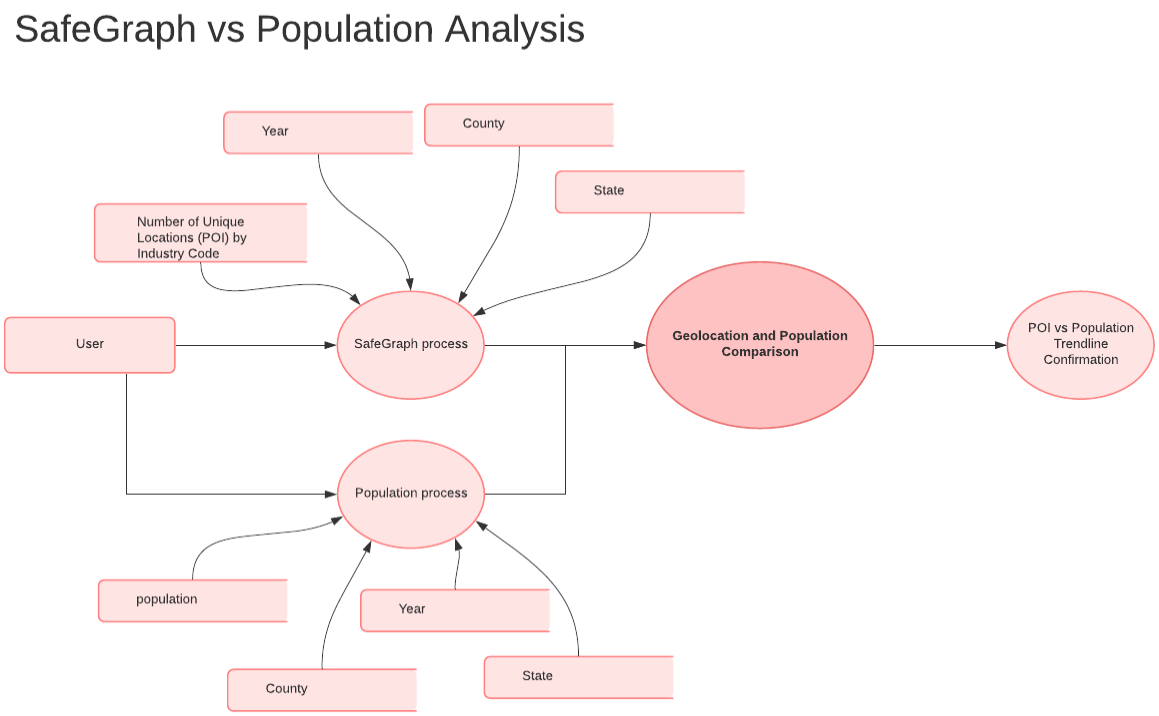

Figure 6.4. Data Flow of SafeGraph vs. Population Analysis

7. Development

7.1 Acclimation of Data - Sprint 1

Table 7.1 Sprint 1 User Stories

Completed During Sprint?	User Story	Points
	Epic: Project Organization/Business Involvement	
Yes	As a student, I want to review the project outline in order to have an idea on how to write an initial draft of the final paper	8
	Epic: Data Organization/Processing	
Yes	As an associate of the firm, I would like the content of each PySpark notebook to be broken up and organized by content to ensure clarity of the analysis done	5
Yes	As a developer, I need to make sure that normalization occurs before comparing the two data sets to ensure consistent and reliable analysis between two different data sets	8
Yes	As a developer, I would like to utilize Plotly for charting and presenting the team's findings so that they are accessible to non-developers.	2
	Epic: Yodlee vs. SafeGraph	
Yes	As an investor, I want to compare foot traffic between different regions for each several locations in order to make sound regional investments	21
No	As a database vendor, I want to compare the number of Yodlee users to the number of SafeGraph users in order to check the validity of both Yodlee and SafeGraph	╳
No	As a developer, I want to familiarize myself with the environments listed in order to be well-versed for my sponsor	╳
	Points Completed	**44**

Table 7.2 Sprint 1 Risks

Name	Description	Risk Category	Probability	Impact	Risk Status	Mitigation
Knowledge	Quick onboarding period, insufficient amount of time to get over learning curve with the data sets	Technical	Medium	Medium	Low	Over the past few weeks, Hunter and Katie have become accustomed to the way Yodlee and SafeGraph work as well as Azure Databricks through daily tasks
Personal	Member is not a fully functioning capacity	Human Capital	High	High	Medium	Sophia is still recovering; she is doing her best
Budget	Too many points into the Trello Sprints, daily new assigned tasks prevent us from working on already existing tasks	Human Capital	Medium	High	Medium	We are having a bit of a challenge trying to plan as the tasks are daily, but we are trying to project what future tasks would be and then plan accordingly
Time	Seven week time constraint, given daily tasks	Organizational	Medium	High	High	As the project is progressing, tasks are becoming more complicated and knowing the time constraint yet not the entire picture, we are doing our best with attempting to set realistic goals
Environmental	Acclimation to new remote environment is more difficult to connect with mentors	Technical	Medium	Medium	Low	We have settled into a good routine with our tech advisor and are meeting with each other daily. Our business advisor pops in when necessary

The project team focused on familiarizing themselves with the individual data sets and looking into ways to compare the Yodlee and SafeGraph data. The work done with the SafeGraph data included visualizing the foot traffic of hardware stores such as Home Depot, Lowes, and Ace Hardware using NAICS code, which is an industry code given to businesses by the Federal Statistical Agencies. This data had to be normalized using the SafeGraph normalization device panel to accurately represent the visitor counts for each store. We looked at this data on a national view and then organized it by region to look for differences. For Yodlee, we compared the amount of online and offline credit transactions recorded for the years 2019 and 2020, noting a huge drop off in physical transactions at the beginning of the 2019 SARS-CoV-2 pandemic and a spike in the number of online purchases. Yodlee has data from 2014-2020 so there was plenty of data to compare to 2020 which has been an outlier in terms of statistical trends. Next, we began to compare the two data sets by finding supermarkets that exist in both and are currently working to find ways to join the SafeGraph and Yodlee tables.

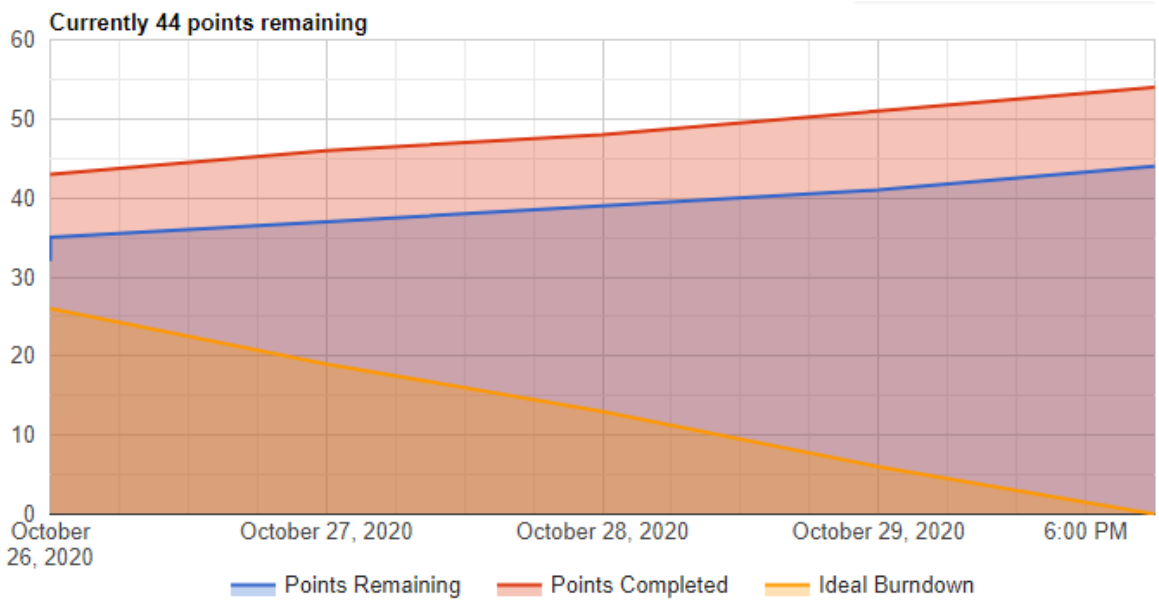

Figure 7.1 Sprint 1 Burndown Chart

7.1.1 Retrospective 1 for Sprint 1

As this was our first week working with the new data sets and software, we had a lot of growing to do. As the week went on, we developed a better understanding of the environments with the alternative investment firm's guidance. Though few concrete deliverables were produced this week, much of the work done has set us up to be successful for the future of the project. The primary thing we need to work on is keeping consistent and detailed documentation, as well as improving the quality of code.

7.2 Yodlee vs. Population Data - Sprint 2

Table 7.3 Sprint 2 User Stories

Completed During Sprint?	User Story	Points
	Epic: Project Organization/Business Involvement	
Yes	As a student, I want to review the project outline in order to have an idea on how to write an initial draft of the final paper	1
Yes	As a professor, I would like to see a Weekly Sprint Recap so that I can know the latest progress of the project.	2
	Epic: Data Organization/Processing	
Yes	As a database vendor, I want all of the analysis done consolidated into a single PySpark notebook presentation in search of the validity of the data in order to deliver a concise analysis As a developer, I want to familiarize myself with the environments listed in order to be well-versed for my sponsor	2
Yes	As a database vendor, I want all of the analysis done consolidated into a single PySpark notebook presentation in search of the validity of the data in order to deliver a concise analysis As a developer, I want to familiarize myself with the environments listed in order to be well-versed for my sponsor	3
	Epic: Yodlee vs. SafeGraph	
Yes	As a database vendor, I want to compare the number of Yodlee users to the number of SafeGraph users in order to check the validity of both Yodlee and SafeGraph As an investor, I would like to see the comparison between the number of grocery stores (number of H-E-B and number of stores in general) vs.population by county in TX in order to see if we can predict population growth by the number of grocery stores	8
Yes	As a database vendor, I want to compare the number of Yodlee users to the number of SafeGraph users in order to check the validity of both Yodlee and SafeGraph As an investor, I would like to see the comparison between the number of grocery stores (number of H-E-B and number of stores in general) vs.population by county in TX in order to see if we can predict population growth by the number of grocery stores	13
	Epic: SafeGraph Validation	
Yes	As an investor, I would like to understand the 2019 SafeGraph user drop-offs to confirm the authenticity of the data set.	5
	Points Completed	**47**

Table 7.4 Sprint 2 Risks

Name	Description	Risk Category	Probability	Impact	Risk Status	Mitigation
Knowledge	Quick onboarding period, insufficient amount of time to get over learning curve with the data sets	Technical	Medium	Medium	Low	Over the past few weeks, Hunter and Katie have become accustomed to the way Yodlee and SafeGraph work as well as Azure Databricks through daily tasks
Personal	Member is not a fully functioning capacity	Human Capital	High	High	Medium	Sophia is still recovering; she is doing her best
Budget	Too many points into the Trello Sprints, daily new assigned tasks prevent us from working on already existing tasks	Human Capital	Medium	High	Medium	We are having a bit of a challenge trying to plan as the tasks are daily, but we are trying to project what future tasks would be and then plan accordingly
Time	Seven week time constraint, given daily tasks	Organizational	Medium	High	High	As the project is progressing, tasks are becoming more complicated and knowing the time constraint yet not the entire picture, we are doing our best with attempting to set realistic goals
Environmental	Acclimation to new remote environment is more difficult to connect with mentors	Technical	Medium	Medium	Low	We have settled into a good routine with our tech advisor and are meeting with each other daily. Our business advisor pops in when necessary

This week we continued to keep documentation for the notebooks and deliverables created during Sprint 2. This was important to the project as we wanted the process to be replicable. We did this by separating notebooks into meaningful divisions and categories, doing so has helped them gain a better understanding of the environments and correlations between the data. Additionally, we found that analyzing offline and online shopping for the Yodlee data set was important to consider when looking at the migration patterns. We began comparing the population growth from 2015-2019 for a specific supermarket in Texas, called H-E-B. Using both Yodlee, we found that there was a correlation between offline credit card transactions at HEB and the population of Texas. This was our first real data comparison that showed correlation to migration patterns. To look for further correlation within the Texan population, we began investigating with the SafeGraph data set next. To visualize the data, we began using scatterplots, trendlines, and line graphs through the Plotly framework.

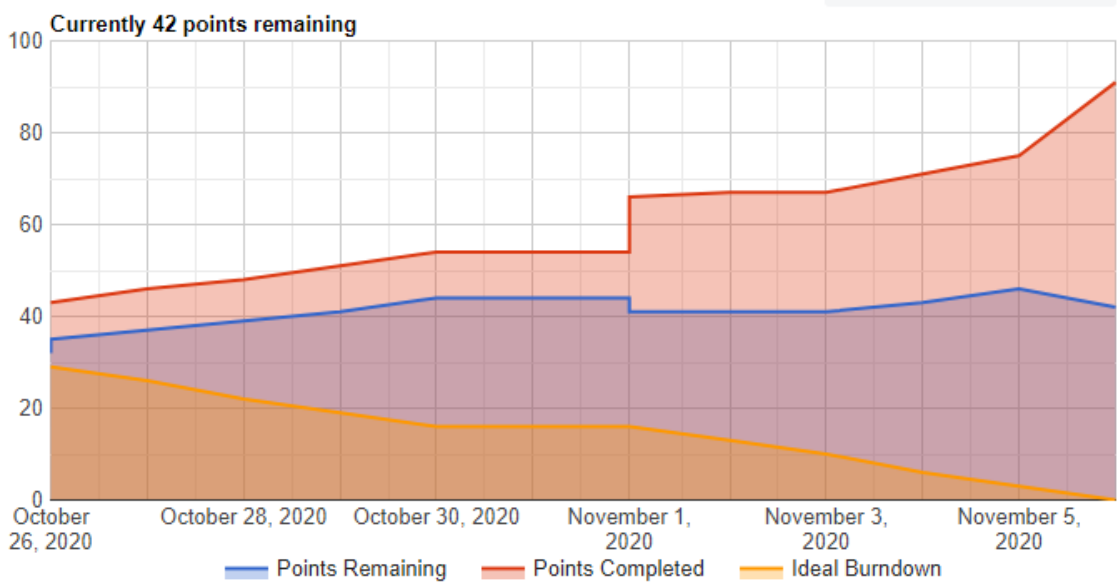

Figure 7.2 Sprint 2 Burndown Chart

7.2.1 Retrospective 2 for Sprint 2

Using the Agile Management style, we were able to produce consistently high-quality work, however, our communication and documentation throughout this process had some room for improvement. As one of the members had no prior experience with Python coding, we needed more explanations for the written code, and as such we looked to improve the documentation for Sprint 3. We learned that it was important to document their work so that anyone can read it and understand it as well as try to document it from different perspectives as this will provide insight for future readers.

7.3 Finding correlations between Yodlee and SafeGraph - Sprint 3

Table 7.5 Sprint 3 User Stories

Completed During Sprint?	User Story	Points
	Epic: Data Organization/Processing	
Yes	As a database vendor, I want all of the analysis done consolidated into a single PySpark notebook presentation in search of the validity of the data in order to deliver a concise analysis	5
	Epic: Yodlee Validation	
Yes	As an investor, I want to be able to see which supermarkets exist in only one state in order to create credit card transaction analysis against the state's population growth	5
Yes	As a database investor, I want to understand why Yodlee users are being double-counted, so getting the monthly aggregated data from the vendor is necessary in order to compare credit card transaction users vs.the population of Texas	5
Yes	As a database vendor, I would like to see the correlation between the average credit card transaction amount vs.the number of Yodlee users in order to see if factors other than inflation influence the number of Yodlee users	8
Yes	As a database vendor, I would like to see the correlation between the average credit card transaction amount vs.the number of Yodlee users in order to see if factors other than inflation influence the number of Yodlee users	8
Yes	As an investor, I want to be able to see which supermarkets exist in only one state in order to create credit card transaction analysis against the state's population growth	8
	Epic: Yodlee vs. SafeGraph	
Yes	As a database vendor, I would like to check the validity of both Yodlee and SafeGraph by comparing the number of Yodlee users to the number of SafeGraph users and make sure only the SafeGraph users above the rolling average of users by location are included in order to reduce unnecessary "noise" and outliers	13
	Epic: SafeGraph Validation	
Yes	As an investor, I would like to see the comparison between the number of grocery stores (number of H-E-B and number of stores in general) vs.population by county in TX in order to see if we can predict population growth by the number of grocery stores	8
	Epic: Machine Learning	
No	As a real estate investor, we would like to predict the population of a state by county or city using ML correlation matrices for different industries trained using 2018-2019 sg_poi data in order to predict population and make sound real estate investments	╳
No	As a database vendor, I want our users to update POI queries to use current_poi tables/yr instead of monthly_patterns in order to accommodate our normalization techniques	╳

Points Completed	**60**

Table 7.6 Sprint 3 Risks

Name	Description	Risk Category	Probability	Impact	Risk Status	Mitigation
Knowledge	Additional data sets included to aid our analysis as well as new python machine learning libraries (Keras)	Technical	Medium	Medium	Low	Over the past few weeks, Hunter and Katie have become accustomed to the way Yodlee and SafeGraph work as well as Azure Databricks through daily tasks
Personal	Member is not a fully functioning capacity	Human Capital	High	High	Medium	Sophia is still recovering; she is doing her best and was not able to participate this week
Budget	Too many points into the Trello Sprints, daily new assigned tasks prevent us from working on already existing tasks	Human Capital	Medium	High	Medium	We are having a bit of a challenge trying to plan as the tasks are daily, but we are trying to project what future tasks would be and then plan accordingly
Time	Seven week time constraint, given daily tasks	Organizational	Medium	High	High	As the project is progressing, tasks are becoming more complicated and knowing the time constraint yet not the entire picture, we are doing our best with attempting to set realistic goals

Environmental	Acclimation to new remote environment is more difficult to connect with mentors	Technical	Medium	Medium	Low	We have settled into a good routine with our tech advisor and are meeting with each other daily. Our business advisor pops in when necessary

This week we continued to keep documentation for the notebooks and deliverables created during Sprint 3. We found that this was especially important as the alternative investment firm deemed it fit to present our findings to the SafeGraph and Yodlee vendors. We focused on the supermarket company H-E-B, which only exists in Texas because it allows them to perform analysis we could not with national supermarkets. Specifically, we compared the population growth for that specific state to the number of offline/online credit card transactions. We also compared the number of unique H-E-B locations in SafeGraph by county to the population by county over 2018-2019 and found a strong correlation.

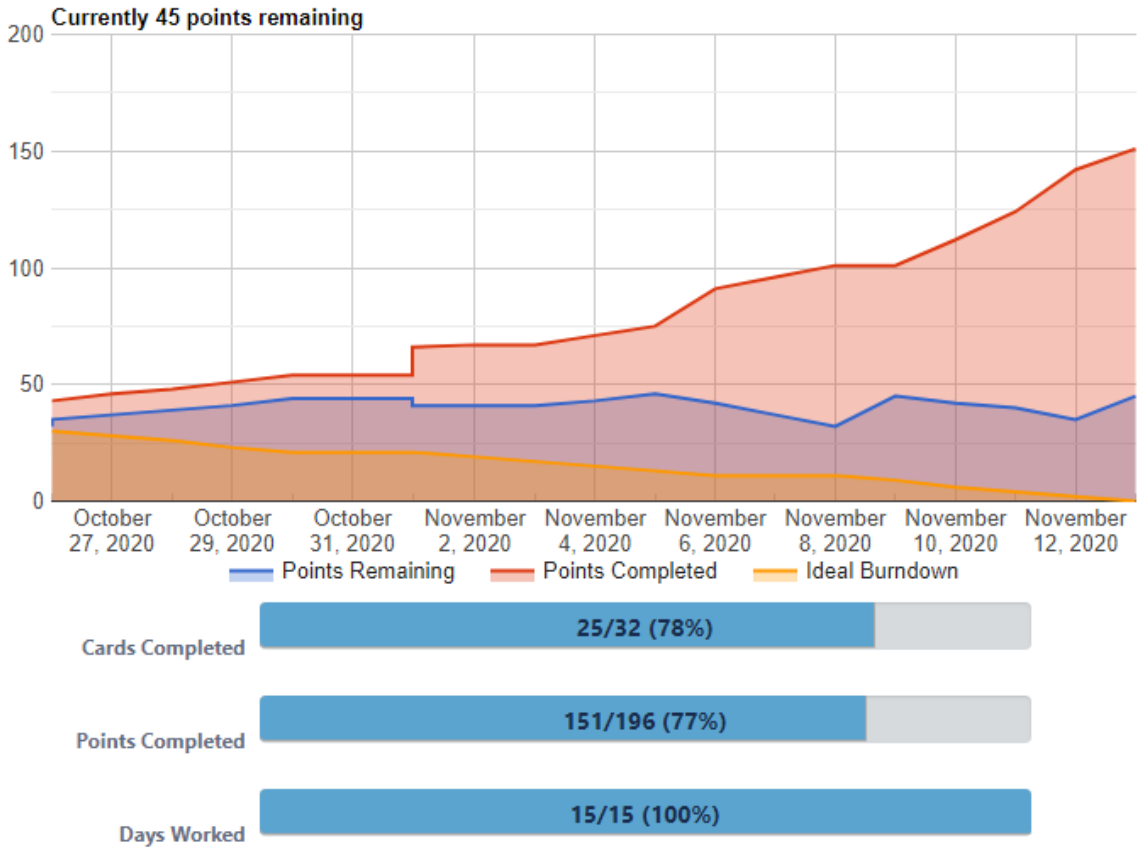

Figure 7.3 Sprint 3 Burndown Chart

7.3.1 Retrospective 3 for Sprint 3

We improved the way we communicated and tracked their progress through their documentation. The two members of the group responsible for developing the software of the project provided more detail. This enabled the other team member to understand the process more. We found that more emphasis should be placed upon the planning for the entire project, however, as the tasks are given to us daily without the larger picture being considered, it was challenging to do so.

7.4 Finding correlations between POI (Number of Stores) and SafeGraph - Sprint 4

Table 7.7 Sprint 4 User Stories

Completed During Sprint?	User Story	Points
	Epic: Machine Learning	
Yes	As a real investor, I would like to know how the model works by having the team create documentation for the ML regression model in order to maintain the model after the project has ended or for a future project team.	3
Yes	As a real estate investor, I would like to see if the model works on other industries/states by expanding the number of industries and locations used in the regression model in order to predict population in other regions	3
Yes	As a database vendor, I want our users to update POI queries to use current_poi tables/yr instead of monthly_patterns in order to accommodate our normalization techniques	5
Yes	As a real estate investor, I would like to know how accurate the models are by calculating confidence intervals for the ML regression model in order to determine if the analysis will be useful for my investments.	8
Yes	As a real estate investor, we would like to predict the population of a state by county or city using ML correlation matrices for different industries trained using 2018-2019 sg_poi data in order to predict population and make sound real estate investments	21
No	As an investor, I want to see if the correlation between Yodlee and SafeGraph users would be improved by normalizing using the home_cbg column in order to validate using both data sets to predict migration patterns	╳
	Points Completed	**48**

Table 7.8 Sprint 4 Risks

Name	Description	Risk Category	Probability	Risk Status	Mitigation
Knowledge	Quick onboarding period, insufficient amount of time to get over learning curve with the data sets	Technical	Medium	Low	Over the past few weeks, Hunter and Katie have become accustomed to SkLearn machine learning models in order to predict population
Personal	Member is now back at almost full capacity	Human Capital	Low	Low	Sophia has now been cleared to return to daily life and normal tasks, so she is ramping back up to a normal workload. Catching up on the work from the last week is the reasoning behind a "medium" risk status, but is going smoothly as Hunter and Katie have documented the work well
Budget	Daily new assigned tasks prevent us from working on already existing tasks, however, some newer tasks are more long term which allows for better planning	Human Capital	Medium	Low	Tasks have gotten more challenging and take multiple days to complete, this enables a bit more planning but since the tasks are still somewhat daily, we have to take it a day at a time.
Time	Seven week time constraint, given daily tasks	Organizational	Medium	Low	We believe we will have sufficient time to complete the project report and deliverables by the end of the term. This is partially due to our Trello burndown and also how we have personally evaluated the situation
Environmental	SkLearn	Technical	Medium	Low	This was a new technology that we leveraged to produce our machine learning model. It took some time to get used to the nuances but we were able to get acclimated quickly because it uses Python syntax.

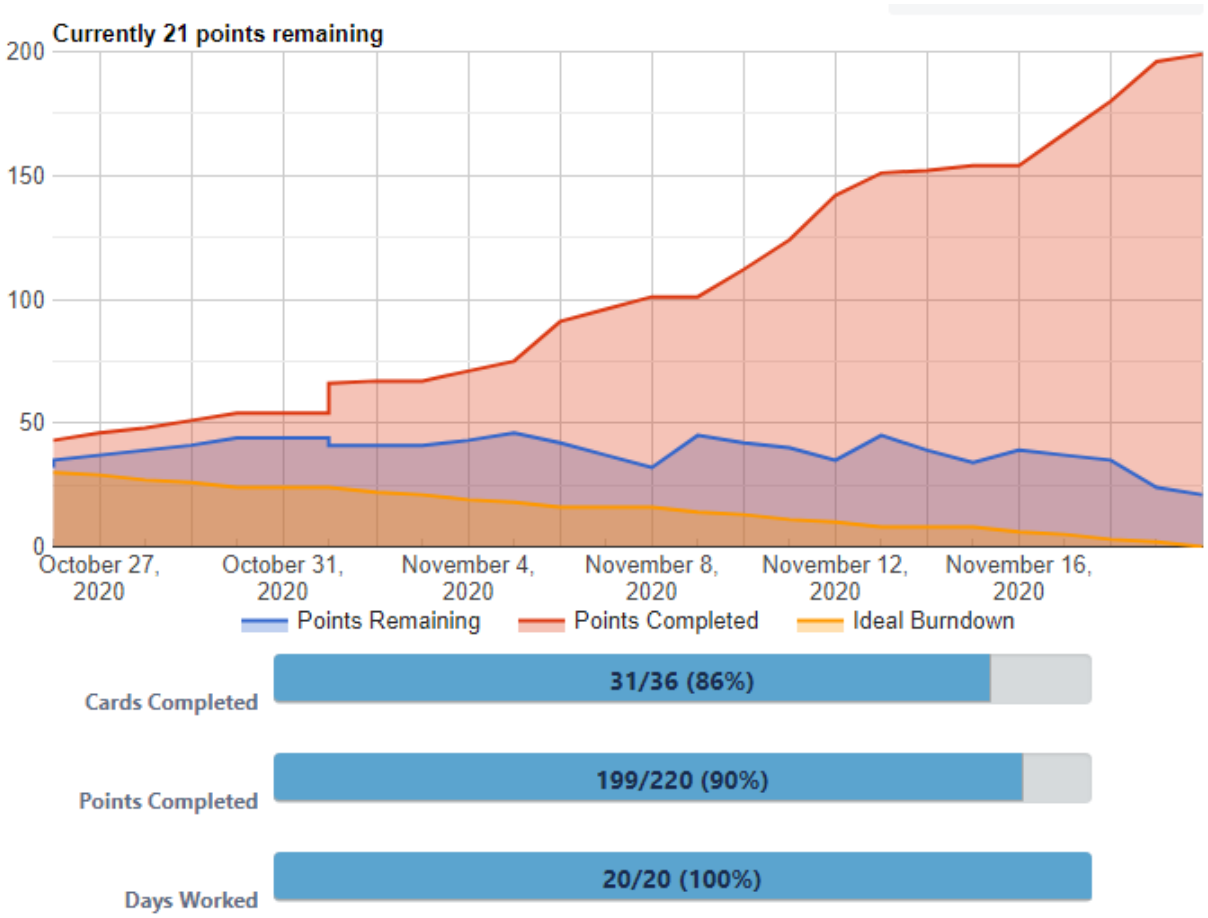

Figure 7.4 Sprint 4 Burndown Chart

This week we began work on one of our primary deliverables, a machine learning regression model that is able to predict the population of states by county by utilizing the existing linear correlation between population and the unique number of grocery stores stored in SafeGraph. The model was created with SkLearn and is trained using SafeGraph and United States Census data from 2018 to 2019. It then uses the 2020 SafeGraph data to produce predictions for the population for each county of the specified state. We also calculated a confidence interval for the results model by comparing its predictions to a third party population prediction found online. The model was tested on states with increasing and decreasing populations yet the confidence interval remained above 92%.

We are in the process of updating our SafeGraph vs. Yodlee comparison with a different normalization technique to the SafeGraph data to see if the correlation between the two data sets improves.

We also wrote the preliminary outline to the alternative investment firm's thesis presentation.

7.4.1 Retrospective 4 for Sprint 4

During Sprint 4, we were able to plan a bit more as there were more well-defined methodology and end goals. Significant work was completed on the machine learning model which marks a big step forward for the project. The results of the training were promising and produced a great base model that can be extended. The next task for us was finishing up the machine learning model and how to wrap up the project and make sure that the comparisons of the model match the other predicted model we were using.

It was also quite informative and helpful to go over the completed sections of the drafts of the final paper with both Professors Wong and Sarnie.

7.5 Predicting Populations by County with a Regression Model - Sprints 5 and 6

Table 7.9 Sprints 5 and 6 User Stories

Completed During Sprint?	User Story	Points
	Epic: Project Organization/Business Involvement	
Yes	As a vendor/associate of AG, I would like to have a copy of the team's final presentation and polished final paper in order to have the team's findings organized for seamless handoff in order to continue the team's recommended next steps, as stated in the presentation and paper	8
Yes	As a student, I would like to complete the business portion of the paper in order to provide a sound business purpose to my team's developmental analysis to future readers	13
	Epic: Data Organization/Processing	
Yes	As a developer, I would like to migrate ML scripts into PySpark/Databricks so that it does not require a local development environment	5
Yes	As an investor, I would like a technical presentation regarding the team's 7 week analysis in order to continue their research in validating the data sets, leading to confidence in our population trends and predictions	13
	Epic: Yodlee vs. SafeGraph	
Yes	As an investor, I want to see if the correlation between Yodlee and SafeGraph users would be improved by normalizing using the home_cbg column in order to validate using both data sets to predict migration patterns	8
	Epic: Machine Learning	
Yes	As an investor, I would like have confidence in the model being used to predict population and investigating its overfitting issue in order to have more confidence in the regression model	8
Yes	As an investor, I would like to use bars, restaurants, gyms to predict population because of the low cost barrier	8
Yes	As an investor, I would like to use the regression model to predict the population of the United States top two hundred counties so I can predict populations of the top counties in order to make sound business investments.	21
	Points Completed	**84**

Table 7.10 Sprints 5 and 6 User Stories

Name	Description	Risk Category	Probability	Risk Status	Mitigation
Knowledge	PySpark conversion (ML model from local environment to data bricks)	Technical	Medium	Low	Setting aside the time to do the conversion
Time	Wrapping up everything in time, development wrapped up on 12/4	Organizational	Medium	Low	Stopped work on the ML system on 12/4

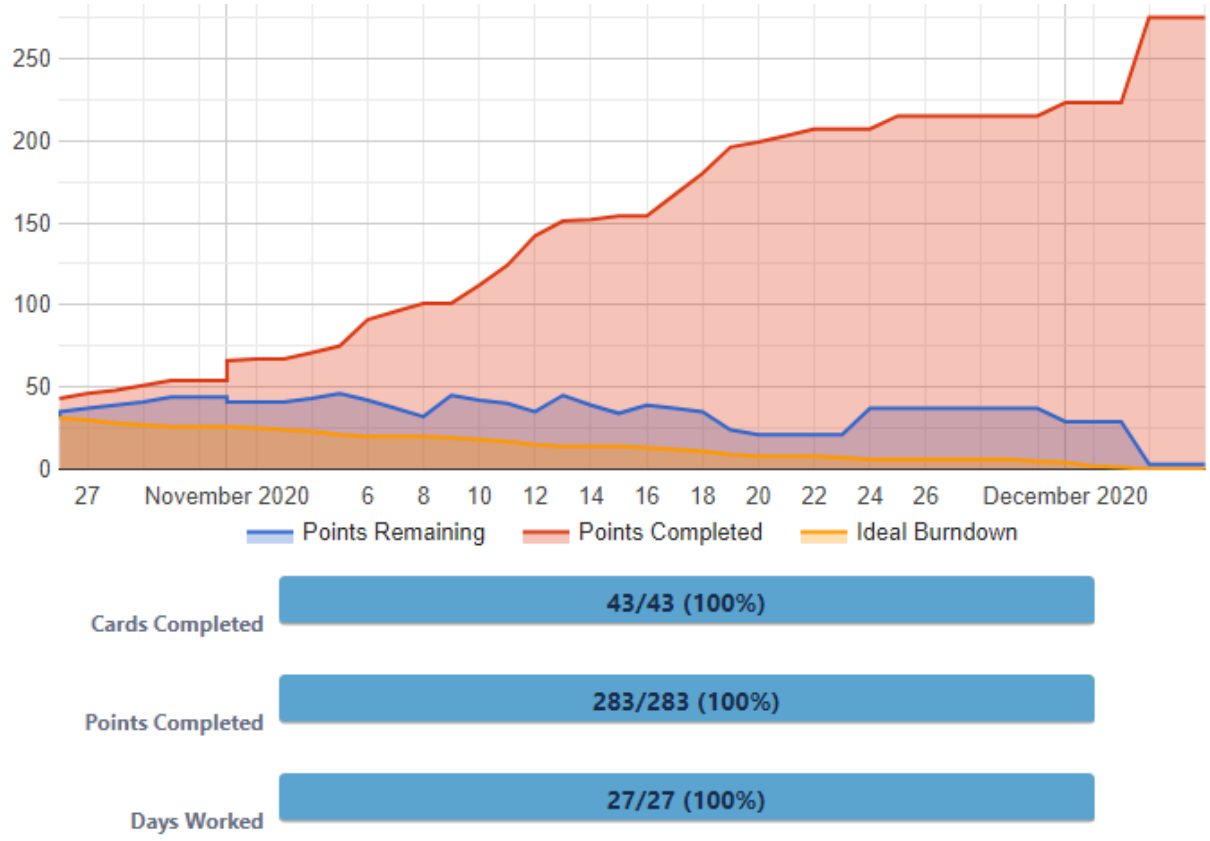

Figure 7.5 Sprints 5 and 6 Burndown Chart

For this Sprint overview, we combined weeks 5 and 6 due to the 2-day work week during week 5. We have implemented our model to the top 200 counties in the US, with a mean squared error of ~230,000 people when checked against the "true" 2020 population found on https://worldpopulationreview.com/states/. Hunter and Katie presented a draft of the final presentation to the alternative investment firm and they were very pleased with the outcome. The final presentation took place the following Wednesday, so we had until then to perfect the presentation.

The group has continued working on the final paper and are waiting on feedback from the professors. We will continue to include up-to-date information into our paper until then.

7.5.1 Retrospective for Sprints 5 and 6

The development team finalized the Random Forest Regression model. The presentation we gave to the alternative investment firm went well and we received constructive criticism from the alternative investment firm. The paper was progressing well as the end of the term approached.

8. Discussion

The alternative investment firm gave us two main goals to complete during the duration of the MQP which are listed below:

1. Determine if the available data sets can be used to identify migration patterns through a series of data validation exercises.
2. With the data analysis and validation completed, determine how the data sets can be used to find and predict migration patterns.

The development team used the software environments, provided to us through the alternative investment firm, to accomplish these goals.

8.1 Computer Programs and Models Used

When researching correlations between data sets, the development team used Plotly, a Python library, in order to plot and find linear regressions. When developing the model used in predicting population by county, we used the SkLearn library, which is another Python library, in order to access the Random Forest Regression model. We were successful in both graphing and developing models and giving a high level of accuracy even with the limited data in both Yodlee and SafeGraph.

The analysis tests and final analysis were displayed using various forms of graphs and charts. The majority of the findings are displayed through scatter plots with the trendlines of population predictions by state, county, and then by major city. We constructed a final slide show presentation that was delivered to the alternative investment firm which highlights the primary takeaway from our research, the methodology used, and the next steps for the continuation of the project by another team.

8.2 Data Processing - SafeGraph and Yodlee Normalization

Before we could discover migration trends, we performed simple queries in tandem with research on the data vendors to familiarize the group with the data sets.

Although the Yodlee data set had been updated with credit card transaction data daily since 2014, it provides no data on the transaction location. This made it impossible to count the number of users in a specific location by vendor, unless the vendor operated exclusively in one location. The SafeGraph data set was much better suited for this project as it provided the exact location a transaction was made based on date and vendor.

We found that both data sets required modifications before being used. The SafeGraph data needed a second pass of normalization by the panel, which is the number of devices in a census block group. This was done using SQL code similar to this:

```
raw_visitor_counts * (sum(raw_visitor_counts)/number_devices_residing)
```

The number_devices_residing, the normalization panel, is found in the vw_home_panel table. We attempted a different normalization technique on SafeGraph in an effort to be more accurate in determining the true visitor count. This involved the similar calculations, done by separating the visitors into the panels the devices originated from and then totaling each result. Yodlee also required partitioning by date and brand to reduce server load.

The SafeGraph and Yodlee data sets can only be compared for the years of 2018, 2019, and 2020 because SafeGraph did not exist before then. Both data sets have a column for the brand name, the name of the company, and a timestamp column that contains the date. However, joining the data sets is not simple because Yodlee is structured on a daily basis while SafeGraph is structured by week. Also, Yodlee does not provide geolocation data associated with the credit card data, therefore, if we wanted to compare online and offline users from brands in a specific region, we must look at businesses that are geographically isolated so we can be sure of where the credit card transactions are coming from.

8.3 Data Analysis

8.3.1 Yodlee vs. Google Trends and Population

While investigating seasonal shopping trends for different types of business industries using only the Yodlee data, we discovered the Supermarket H-E-B, which exists only in the state of Texas.

We began attempting to validate the Yodlee data set by comparing it against the store's online popularity as rated by Google Trends. As we can see in *Figure 8.1*, Yodlee online users experience a spike at the same time as the online popularity does, which was at the same time H-E-B acquired its online delivery service, Favor. A second spike can be seen between Yodlee online users and H-E-B's popularity when their mobile application is launched. Yodlee offline users experience a drop in users during 2020, taking place at the height of the 2019 SARS-CoV-2 pandemic.

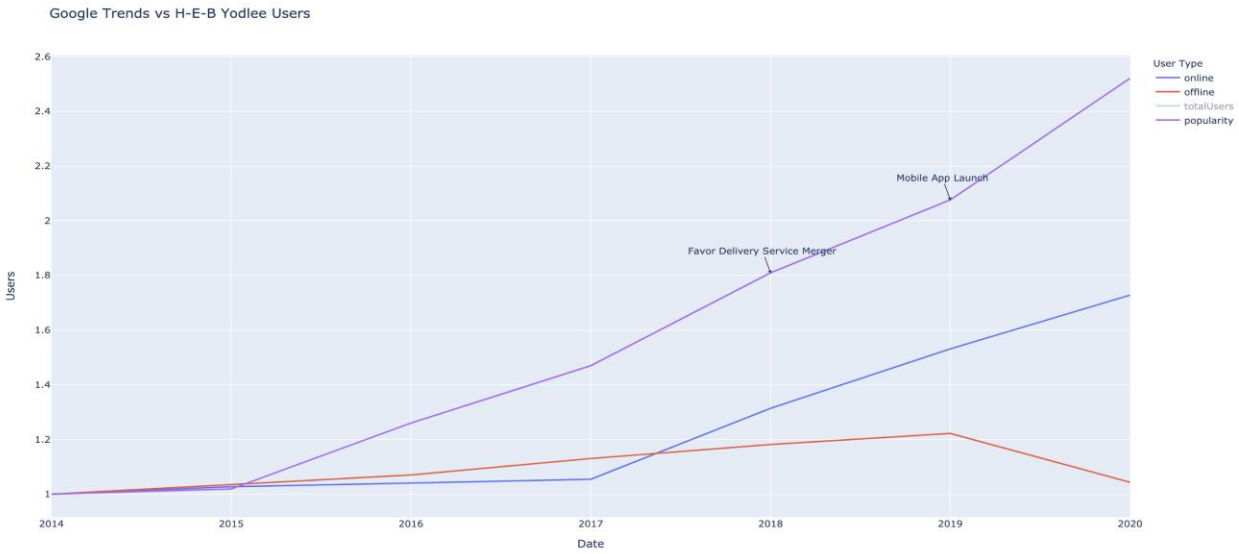

Figure 8.1 Google Trends vs. H-E-B Yodlee Users

Due to H-E-B's store locations being limited to only one state, the alternative investment firm suggested we compare the number of offline and online credit card transactions to the total population of Texas from 2014 to 2019. Performing this analysis on other businesses using the Yodlee data would be difficult because the data set lacks location data. This comparison showed a very strong correlation over a linear trendline, with a coefficient of determination of 0.97 over all five years, as seen in *Figure 8.2*. This result was particularly interesting to the alternative investment firm because it could mean they would be able to predict the population of Texas before the United States Census data is released using only the Yodlee data set.

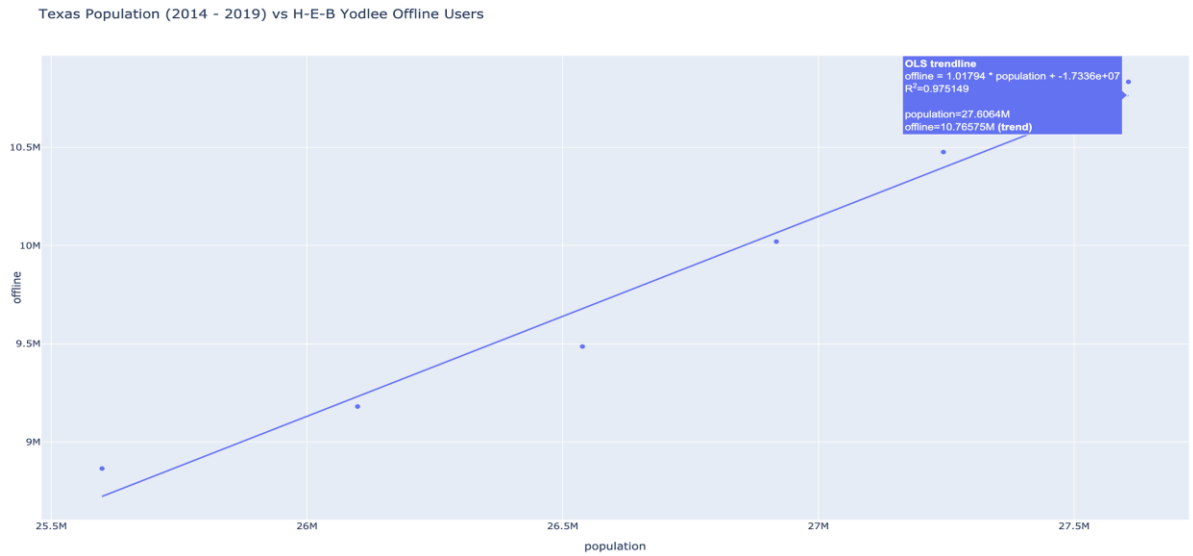

Figure 8.2 Texas Population (2014 - 2019) vs. H-E-B Yodlee Offline Users

8.3.2 SafeGraph vs. Yodlee

This led to us focusing on supermarket H-E-B again, and comparing the number of SafeGraph visitors to the number of Yodlee users to see if there was any correlation between the two data sets. Since we already knew that there was an existing correlation between Yodlee users and population, a correlation between SafeGraph and Yodlee might mean that SafeGraph could be used to predict population as well.

The correlation was initially very low and was not being improved, despite the implementation of a low-pass filter on the SafeGraph visitor count, which attempted to remove stores that could negatively influence the trendline by having too few visitors. Regardless of our best efforts to normalize and restructure the data, using both SafeGraph normalization techniques described previously, the correlation between Yodlee and SafeGraph was consistently low. The final correlation found after all normalization techniques were performed, is shown in *Figure 8.3*.

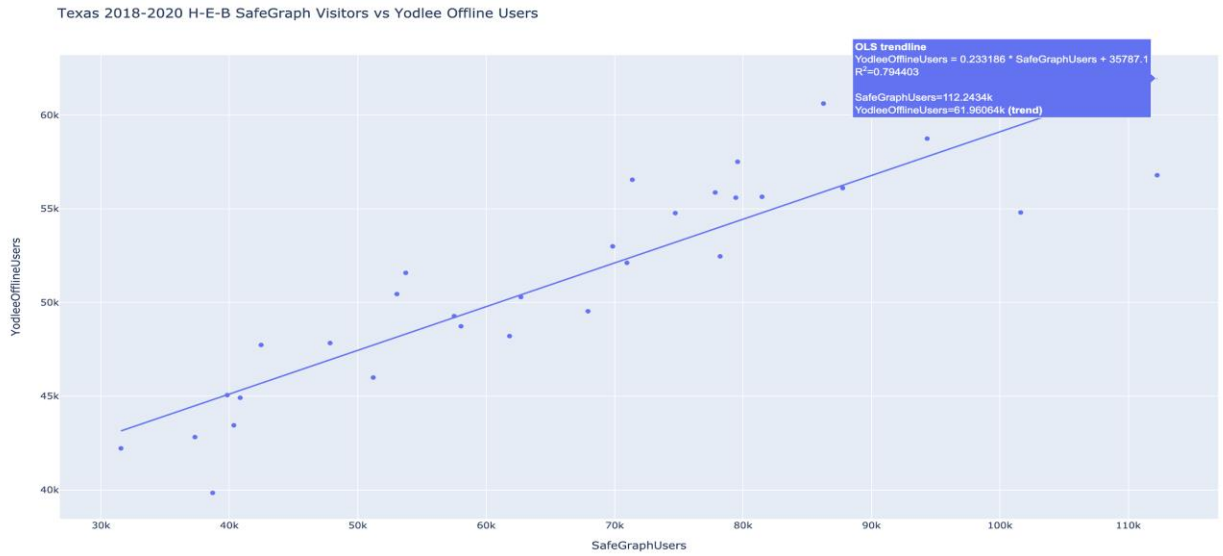

Figure 8.3 Texas 2018-2020 H-E-B SafeGraph Visitors vs. Yodlee Offline Users

The alternative investment firm decided to meet with the Yodlee vendor and notified us that with the current aggregation practices, it would be impossible to compare these two data sets due to the company's differing internal normalization techniques. This is due to the potential over counting of users if we were to aggregate the daily Yodlee data by month in order to compare it to the monthly aggregated SafeGraph data. Although finding a meaningful correlation would have been nice, it was still valuable information for the firm to know in the future that the comparison of these two data sets would be impossible.

8.3.3 SafeGraph vs. Population

Even though Yodlee data has a long history of being compared to SafeGraph, it does not provide data by location, such as number of stores or POI by county. This led us to focus only on the SafeGraph data so that we would no longer be limited to only looking at H-E-B.

We used POI, the number of bars and restaurants, instead of raw_visitor_count to eliminate the need to normalize against the panel of devices, and chose bars and restaurants as the industry since they have a lower cost barrier in opening, compared to an industry such as grocery stores, giving a better indication of population growth. SafeGraph locations can be sorted by industry code, also known as NAICS code so that all brands that fall under a certain business industry, in this case bars and restaurants can be selected for consistent analysis.

In *Figure 8.4*, the graph compares change in POI and population percent growth beginning with 2018. As depicted, regardless of the percent change in population growth, the majority of the change POI for 2019 (in blue) remains at 0.

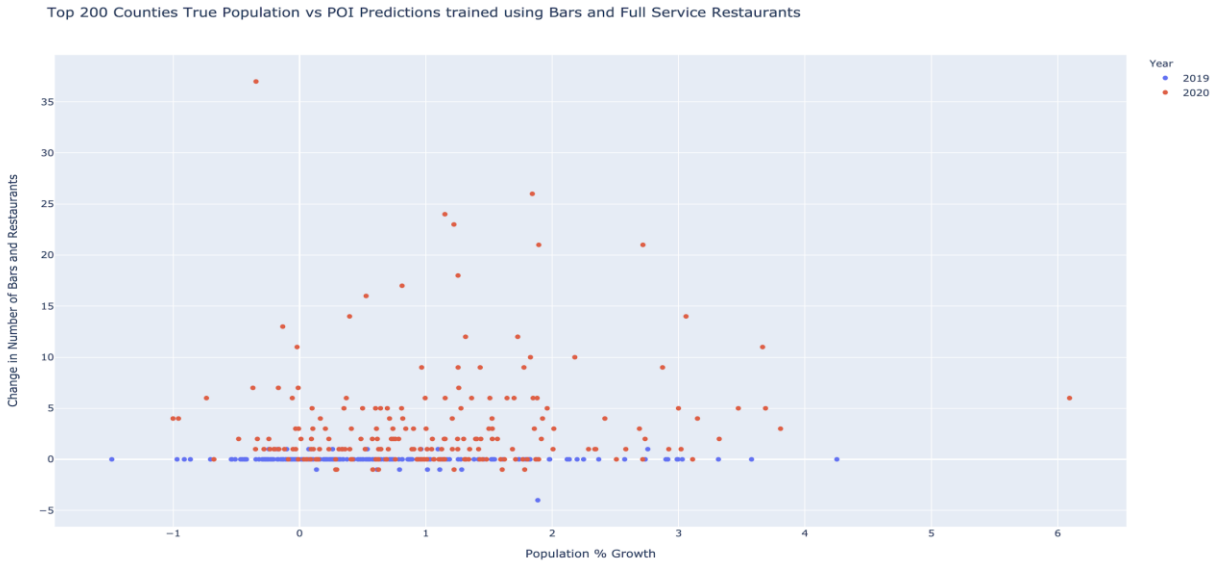

Figure 8.4 Top 200 Counties True Population vs. POI Predictions trained using Bars and Full Service Restaurants

In *Figure 8.5*, the significant spike in the number of bars and restaurants can be seen from 2019 to 2020, in red, compared to the little to no spike from 2018 to 2019, in blue.

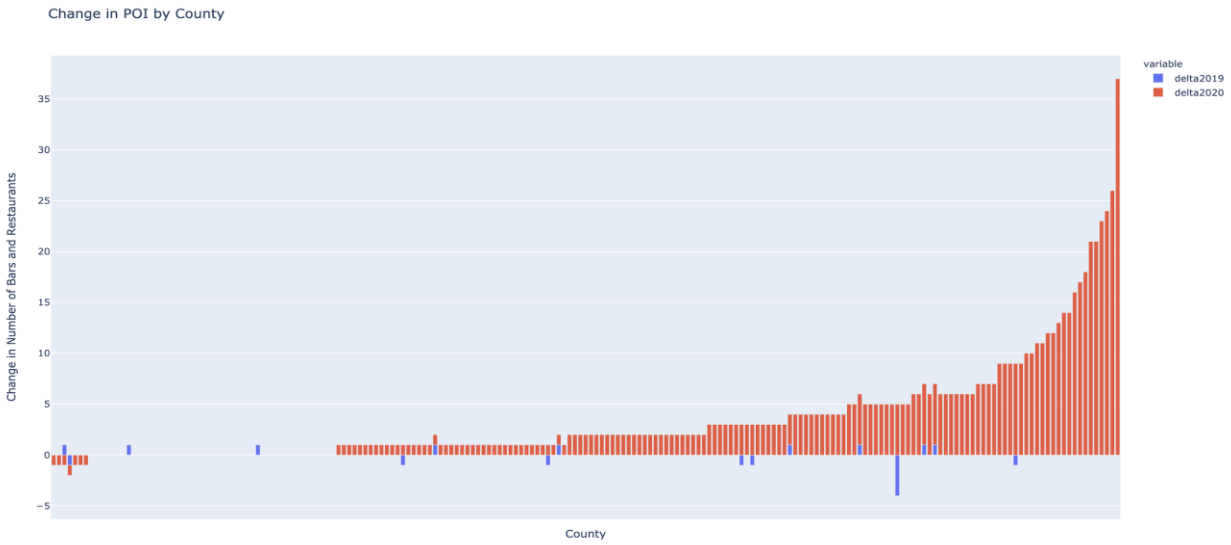

Figure 8.5 Changes in POI by County

Normally, we would assume the drop in the number of bars and restaurants due to the pandemic but that is not what is shown, leading to the conclusion that SafeGraph has evolved their product, leading to the spikes in new brand locations in 2020.

8.3.4 Developing a Model with SafeGraph

We narrowed down the SafeGraph data by state, county, and major city to count the number of POIs for each grouping and then created queries that returned year, county or city, state, and POIs for the years 2018 and 2019. We created two models, one that would predict the population for 2020 if the correlation between POI and population was linear and another one that would handle nonlinear correlations; the linearity was based on 2018 and 2019 data.

To test the accuracy of their predictions, we collected 2020 population projections from https://worldpopulationreview.com/states/ and compared it to the populations given by the model. Our accuracy in predicting population using the grocery store industry code was between 75% and 85% for the states of Michigan, Ohio, New York, and Texas. This was significant because it demonstrates that the model is accurate using states that have an increasing and decreasing population.

Industries such as bars require fewer resources to open compared to grocery stores, providing more of an indication of gentrification in a county. Bars and restaurants mimic population behavior since they are directly influenced by attendance and economic health. The trendline between the number of brand locations and population by county changed due to the group focusing on a different industry code, which necessitated a different model to train the data. We used a Random Forest Regression model, a supervised learning algorithm that uses ensemble learning methods for regression, since this model performed the best with the new bar industry data set (Medium, 2020). Our model was made up of decision trees that returned mean predictions, since our project's goal was to use the model to predict populations with regression

and not classification. This also led to reviewing accuracy between the models in order for us to choose the best model for each data set. We calculated accuracy by checking how many times the true 2020 population, the predicted value given by the website mentioned previously, was between their prediction range. The prediction range is found by using the model's prediction as the midpoint and then adding and subtracting the mean squared error.

Once we were comfortable with the model chosen for bars, the Random Forest Regression model, the research was expanded to a more national level. This required confidence in the previous queries ran for individual states since running a query for counties for each state would require more processing power. We then ran a query for the number of bars, county name and state, and year for the top two hundred counties with the highest populations in the United States from 2018 and 2019 and ran the model to predict the populations for 2020. The command line used to run the Random Forest Regression Model script was as follows:

python3 rfModel [output title] [training data] [test data] [2020 true pop]

- rfModel: name of python3 script
- Output title (string): name of model's output csv file
- Training data (csv): 2018 and 2019 training data (year, county name, POI)
- Test data (csv): 2020 test data (year, county name, POI)
- 2020 true pop (csv): Worldpopulationreview.com's 2020 U.S. county population prediction

The accuracy is depicted in *Figure 8.6*. As depicted in the graph below, the majority of the counties lie fairly close to the line of accuracy, even with only two years of training data. Around 76% of data fall within the 80th percentile, so we are confident that the model will only improve as time progresses.

We also considered the model to be successful due to the high accuracy and a relatively low mean squared error. The counties chosen are listed on this website: https://worldpopulationreview.com/us-counties.

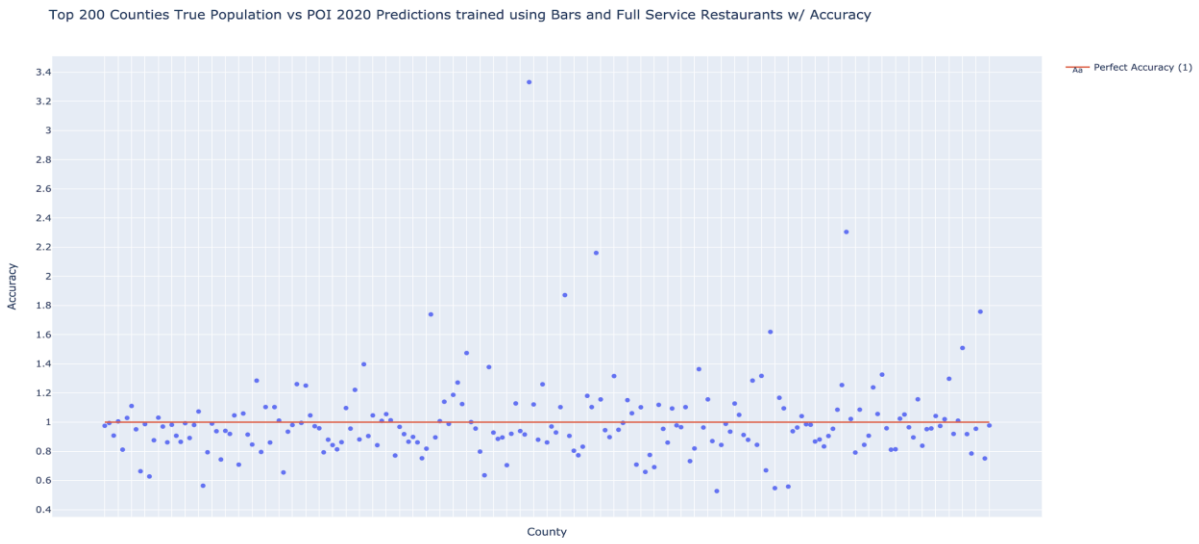

Figure 8.6 Top 200 Counties True Population vs. POI 2020 Predictions training using Bars and Full Service Restaurants w/ Accuracy

8.4 Benefits of the Analysis

The alternative investment firm benefits from the analysis developed in several ways. The validation of the Yodlee data set against Google Trends, a reliable data source, confirms that Yodlee can be confidently used to track the success of businesses that make new acquisitions, as seen with the H-E-B delivery service acquisition. The high correlation between the H-E-B grocery store visits and the population of Texas validates the Yodlee data set, as well as proves that population prediction could be done using other regionally isolated brand names from the Yodlee data set.

The alternative investment firm also benefits from the analysis done when comparing the Yodlee data set against the SafeGraph data set, by understanding the necessity of a weekly aggregation of the Yodlee data to prevent double counting in the number of Yodlee users and for a more synchronized comparison against the SafeGraph data set. Currently, Yodlee is a daily aggregation. Once that is acquired, the alternative investment firm can begin analysis to see if migration trends have an influence on brands using Yodlee as an overview for regionally isolated brands and then using SafeGraph to zoom in geographically.

The conclusion found from the SafeGraph and population comparisons benefit the alternative investment firm through the evidence of the gradual improvement of the vendor's product, yielding a more accurate reading on the number place ID's of points of interests in 2020, compared to 2018 and 2019.

The firm primarily benefits from the regression model, which gives the firm the ability to predict population within a reasonable error range. The firm is considering acquiring the Veraset data set, due to the unreliability in counting the number of distinct SafeGraph place ID's in the

SafeGraph data for 2018 and 2019. The goal of the acquisition of the new data is to provide the firm a more accurate history by giving access to tracking migration by device ID instead of place ID. Specific information about devices between regions would now be available through Veraset instead of the statistical model that Yodlee and SafeGraph provide. These data trends from the Veraset data set can also be used in the regression model in an attempt at predicting with higher accuracy.

9. Assessment

9.1 Business Learnings

Over the course of this project, we learned a number of things in key areas that will be invaluable as we begin our careers. Firstly, we learned how to adapt to the online nature of the work and how to work together remotely. Secondly, we learned about the workings of the financial industry. Lastly, we learned about the importance of credible data. While we learned many other things during the course of this project, these three main topics best represent the skills that we will be able to take with us into our future careers.

We learned about communication in the industry through their daily standups and also learned how to work with a project manager that provides autonomy. In addition, we learned about professionalism, channels of communication, and the level and frequency of communication typical of prominent financial institutions. Due to the 2019 SARS-CoV-2 pandemic, we had to develop remote working skills. It was critical that we met frequently and used the time efficiently. Agendas ensured we were focused while sprint summaries helped to catalogue the work done within the week. The remote requirements of the pandemic also necessitated quality documentation from the beginning, which was invaluable throughout the project. For the first leg of the project, one of the members was unavailable to meet regularly and so, documentation and thorough communication was paramount to ensure a smooth onboarding after the project had begun.

We also learned a lot about the financial industry and the real estate market. We learned how much of investments are data driven, and the methods behind them. We were able to utilize the alternative investment firm's development environment to perform statistical analysis and receive advice from senior developers who write this sort of code every day. Through this, we gathered insights into how we approach these financial analysis problems. While it is difficult to predict the real estate market alone, we found that analyzing activity patterns such as customers going to grocery stores or restaurants were a good indication of human migration.

Though the technical details of this project were highly specific, we still learned some skills that will be useful to us in the future. When aggregating data sources, it is important to validate the data *before* testing it against other data sources. It is also critical to know when to stop using data given by vendors if it no longer adequately suits the project requirements. Failure to do so results in valuable time lost trying to finagle the data into the application.

These three things are what we feel best represent our abstract learning during the course of this project. We found that communication was incredibly important, especially given the circumstances that everyone was working remotely. Coming from an engineering background, it was invaluable to learn how to develop a technical engineering project in the context of the financial industry. Finally, we learned to verify our sources before starting a project as this can prevent spending too much time on ultimately unimportant tasks. This was an amazing opportunity to work with such a large firm, and we have no doubt that these experiences will give us a head start as we move to the next chapters of our lives.

9.2 Technical Learnings

The development team was involved in a significant amount of real world technical experience. Through working in the firm's technical work environment, data scrubbing and working with machine learning models, we were able to experience being data engineers and data scientists at a leading FinTech company.

The technical work environments, involving Azure Databricks and PySpark notebooks, was an overall positive experience. It provided a faster development environment compared to the other environments the development team has experienced before this MQP.

The significance of resource sharing and communication was also experienced by the development team during the final phases of the project. The final phase of the project required a large amount of computing resources from the development cluster and since the firm's development team used the same cluster, it was important to communicate when we were running significant queries as to not interfere with other project timelines.

Data scrubbing was a consistent theme experienced throughout the whole project term. In order to accurately find trends between two separate data sets, it was important to pass the data through several normalization techniques, especially when the machine learning model was developed. Working with machine learning models showed us how training models can be used to predict results and to improve the current investment models of the firm

Through these major experiences, we were able to assist the alternative investment firm in validating their data as well as giving us a valuable MQP. We find these technical learnings valuable, as we do not believe we would have encountered these types of experiences in a school environment.

9.3 Accomplishments

In this project, our team was able to create a machine learning model that can predict the future migration patterns of humans in the United States, and will inform the alternative investment firm's future financial investments. To accomplish this, we met a few major milestones. We successfully implemented the Agile Scrum project management method. Agile Scrum and the sprint records we created were especially useful when we onboarded one team member that could not be present for the initial leg of the project. It also provided concrete guidelines to tackle the two major technical challenges of this project. The first being the initial

data analysis to build a training set, and the second being the development of our final machine learning model. The alternative investment firm and future project teams can use this, along with the recommendations we have laid out, to continue the work we have done.

After considering many management styles, we decided that the Agile Scrum method would provide us the greatest flexibility, taking into account the incremental nature of the tasks we were given. Understanding and implementing Agile Scrum early on allowed us to keep pace with the tasks given to us in our daily stand-ups. We were successful at prioritizing our tasks each week which then in turn enabled us to incorporate additional tasks that were given to us at our daily stand-ups into our sprints.

A significant accomplishment of ours was onboarding a member mid-project, something we did not initially expect to do. One of our team members was injured, and it was necessary for them to take some time off to recover. When they returned, we had to get them up to speed quickly so work could continue. Onboarding anyone during any part of a project is difficult as it takes time for them to learn how new programs work, what is currently being done on the project, and what the project trajectory and timeline look like. Not only does it take time for the new person, it also requires that the team teach the new person, which can decrease time that they spend on working on the project themselves. Once they were up to speed, we were able to jump right back to where we left off.

An important step towards the development of our machine learning model was validating, normalizing, and then analyzing the two primary data sets, Yodlee and SafeGraph. The result then served as the training data for the regression model. Different normalization techniques were suggested to us by the alternative investment firm, so we had to try them and analyze the results to determine which produced the most optimal results. The final method was a simple scaling approach. However, after multiple normalization attempts, we determined that the data sets were not compatible, and no common metric could be derived. Both data sets were pre-normalized in different ways and had different time periods in which the data was collected.

The Future Work section of this paper constitutes part of our recommendations for the alternative investment firm and what they might do if they would like to continue this project. Additionally, when we presented our findings to them, we also detailed what some next steps would be for them moving forward. We believe these recommendations will enable the alternative investment firm and future project teams to build off of the things we have accomplished over the course of the project. As this project carried potentially great financial value, we took great care in forming our recommendations, just as we did when conducting the other parts of the project.

The work done for this project constitutes successfully meeting the primary project goals. The primary goal of this project was to validate the individual data sets and then join them to identify migration patterns in the United States from credit card and geolocation data, and then use that analysis as the foundation for an investment strategy. Even with the discovery of the temporary inability to combine the data sets for analysis, we were able to achieve the goal

through analysis of the individual data sets, and making a machine learning model that can help predict the migration patterns of the population of the United States.

9.4 Mechanical Engineering Connections

Though this is primarily a software development project, there are many parallels to other engineering disciplines. In the background section, we provided a description of the Systems Engineering design process. Systems Engineering can be thought of as an alternative to the incremental Agile Scrum development cycle. However, there are many similarities between the two, and in many cases they can complement each other.

Firstly, for any project, it is important to meet with the people who are either funding or sponsoring the project. For us, that meant we needed to meet with the alternative investment firm. In these meetings, we outlined what the next sprint would look like. After each sprint, and many other times in between, we met with the alternative investment firm.

During these sprints, the alternative investment firm tasked us with achieving two goals: they needed the team to evaluate and understand the data sets to make sure they would be applicable and useful to the project and once that was completed, be able to predict the migration patterns of humans through the analysis of the data. We were brought on to this project with the needs of the alternative investment firm already decided, so in effect, we picked up an already existing contract to continue the work and after the team leaves, the work will be continued. With any project the success is measured by a certain set of predefined goals. To gauge how well goals have been met, we needed to define performance metrics. An example of a measure of performance for a mechanical system might be: does the car reach one hundred miles per hour? For this project, an example is: was the team able to perform operations on Azure Data Lake programmatically?

In terms of system context, we had to consider the effects of the SARS-CoV-2 pandemic and where people may be migrating to. As SARS-CoV-2 is not an event where we could compare recent data to, we had to take this into account when considering the system. System context has an additional meaning for this project. Our work was to develop something to be used for private financial gain. This is not a bad thing, but it does set a different precedent for how the work should be completed.

A large consideration of any product development is how the users will use it. This is the operational concept development phase in Systems Engineering. An analogous software development paradigm is Test Driven Development, where a unit test for the software is written first, and then the software is developed to implement the intended functionality. More generally, the interface of any piece of software or API must be considered.

The final process in the Systems Engineering approach is architecting a solution. We used our knowledge of data analysis and machine learning to create a solution which can be applied to projects after we conclude ours. We were only able to get to our solution through following all the other steps of Systems Engineering as the final solution is the last step in the process.

Systems Engineering is just one example of how the work in this project is relevant to other disciplines of engineering. Other topics of discussion could include service quality, where, for example, we needed to document the process of creation and design in order to help diagnose the problems that will inevitably arise along the way. While this project does not completely encompass the methodologies necessary to manage all engineering projects, it does have a similar framework that can be applied to manufacturing, product design, and even other business projects outside of engineering.

10. Future Work

If we were to continue the analysis, we would add a data set that provides the location data of an individual device. SafeGraph aggregates location data based on SafeGraph place ID which correlates to the locations of businesses that users visited. These businesses are known in SafeGraph as POIs, or "places of interest." Each POI contains a variety of data columns including the raw visitor count, other businesses that visitors commonly went to, its census block group, and the state where it is located. While the data provided by SafeGraph does not identify customers, it still can prove useful when looking for overall patterns or trends in the movement of people throughout the United States.

We also recommend considering using Veraset instead of SafeGraph to train the machine learning regression model. Veraset, a sister company of SafeGraph, provides location data of a device which gives us more specifics about migration data instead of just estimates. We believe the acquisition of this data will help the alternative investment firm to make more informed investment decisions. It will provide added insight into migration data on a more detailed level than previously possible.

In continuing the development of our migration prediction model, we see a few potential applications outside the immediate scope of this project. Regarding the ongoing SARS-CoV-2 pandemic at the time of this project, migration predictions could be used to determine preemptively which demographics or geographical areas should be vaccinated first. Large natural disasters often cause similar inflection points in migration patterns, so the model could be used to efficiently target key areas to restock food and aid supplies. Bringing the scope back to real estate, the model could be used to target areas with housing shortages. This information can then be used to more efficiently plan new housing developments, allowing for lower cost living.

Big data permeates across all sectors of the FinTech industry, and there will always be ways we can use it to our advantage. As we continue to develop our careers, this skill of recognizing where a given resource is useful, and where it falls short is important. Especially in management, it is critical to know how to quickly assess the utility of something to decide if it should be pursued further. Even so early in our careers, we have already seen project resources depleted due to the sunk cost fallacy.

11. Conclusion

The alternative investment firm benefits from the integration of data engineering and data analytics into their real estate investment strategies. Through our efforts in validating the data and using machine learning techniques we have enabled the investment firm to confidently rely on their predictions instead of the standard yearly market reports utilized by the majority of investors.

Through our validation exercises, we have concluded that the Yodlee data set is validated, due to the significance in the trends found between H-E-B's Favor acquisition and the online popularity and between the H-E-B offline users and the Texas population. We also concluded that SafeGraph is a valid independent source of data due to the observable trends showing a decrease in the number of in store visitors during the beginning of the 2019 SARS-CoV-2 pandemic. SafeGraph was then used in our machine learning efforts, involving the Random Forest Regression Model, which was able to accurately predict the population of the top two hundred counties in the United States to a high degree of confidence. This model will improve as time goes on and more data is acquired from SafeGraph. We presented our analysis to the alternative investment firm, which will be compounded with another layer of validation after the firm acquires the correct aggregation of the data sets and incorporates Veraset.

It is evident, through our work done at the sponsor company, that the importance of data validation in the FinTech industry. Confidence in the data drives an investment firm's willingness to utilize it in the decisions that they make on behalf of their clients. The FinTech industry benefits from incorporating data analysis into their investment decisions, reducing the unpredictable aspect of making investments and separating the need to rely on traditional data sets used by competing firms.

12. References

Anderson, J. (2018). *Data engineers vs. data scientists*. Retrieved from https://www.oreilly.com/radar/data-engineers-vs-data-scientists/

ASQ. (2020). *What is FMEA? Failure Mode & Effects Analysis*. Retrieved from https://asq.org/quality-resources/fmea

Atlassian. (2020). *Agile Project Management*. Retrieved from https://www.atlassian.com/**agile**/project-management

Atlassian. (2020a). *Kanban vs Scrum*. Retrieved from https://www.atlassian.com/agile/kanban/kanban-vs-scrum

Cprime. (2020). *What is Agile? What is Scrum?* Retrieved from https://www.cprime.com/resources/what-is-agile-what-is-scrum/

F.J. Looft, Systems Engineering for Capstone Projects. Worcester: WPI, 2018.

Furbush, J. (2018). *Data engineering: A quick and simple definition.* Retrieved from https://www.oreilly.com/content/data-engineering-a-quick-and-simple-definition/

Github. (2013). *Github Users Search*. Retrieved from https://github.com/search?q=type:user&type=Users

Medium. (2020). *Random Forest Regression*. Retrieved from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Medium. (2020a). *Random Forest and Its Implementation*. Retrieved from https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Pipefy. (2020). *Solving bottlenecks with the Kanban method. Retrieved from https://www.pipefy.com/blog/the-kanban-method/*

Scrum Reference Card. (2014). *Scrum Reference Card*. Retrieved from http://scrumreferencecard.com/scrum-reference-card/

Trading Economics. (2020). *United States Nahb Housing Market Index | 1985-2020 Data | 2021-2022 Forecast*. Retried from tradingeconomics.com/united-states/nahb-housing-market-index.

Trello. (2020). *Agile Board Example*. Retrieved from https://trello.com/b/WELsX9Hm/agile-board-example