



WPI

U.S. Migration Patterns

Tracking migration patterns of people in result of the COVID-19 Pandemic

Project Team:

Xiaowei Chen xchen10@wpi.edu

Garrett Stephenson ghstephenson@wpi.edu

Luke Trujillo ltrujiillo@wpi.edu

Project Co-Advisor

Professor Wilson Wong

Department of Computer Science

Project Co-Advisor

Professor Robert Sarnie

Department of Business

This report represents the work of WPI undergraduate students submitted to the faculty as evidence of completion of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the projects program at WPI, please see <http://www.wpi.edu/academics/ugradstudies/project-learning.html>

Acknowledgements

We would like to thank our project sponsors for dedicating their time to guiding us and teaching us about data patterns within the alternative investment industry during these unprecedented and uncertain times.

Thank you to our project advisors at Worcester Polytechnic Institute, Professor Wilson Wong and Robert Sarnie for their willingness to show patience and support through the entirety of the project.

Finally, we would like to thank Worcester Polytechnic Institute for allowing us to continue working in a safe and healthy environment during a year like none other than we have seen before.

Thank you,

Xiaowei Chen

Garrett Stephenson

Luke Trujillo

Table of Contents

ACKNOWLEDGEMENTS	II
TABLE OF FIGURES	IV
TABLE OF TABLES	IV
ABSTRACT	V
EXECUTIVE SUMMARY	VI
1. INTRODUCTION	1
1.1 ALTERNATIVE INVESTMENT FIRM	1
1.2 OPPORTUNITY	1
1.3 PROJECT GOALS AND SCOPE.....	1
2. BACKGROUND	2
2.1 ALTERNATIVE INVESTMENT FIRM INDUSTRY	2
2.2 GEOMETRIC DATA POLYGONS	2
2.3 SOFTWARE DEVELOPMENT ENVIRONMENT	3
2.4 PROJECT MANAGEMENT TOOLS.....	5
2.5 PROJECT RISK ANALYSIS	5
3. METHODOLOGY	8
3.1 PROJECT MANAGEMENT	8
3.2 SOFTWARE DEVELOPMENT METHODOLOGY	8
3.3 REQUIREMENTS GATHERING	10
4. BUSINESS AND PROJECT RISK MANAGEMENT	13
4.1 RISK AND REWARD	13
4.2 RISK CULTURE	13
4.3 OPERATIONAL RISKS	14
4.4 MARKET RISKS	14
4.5 CHANGE MANAGEMENT RISKS	15
5. DESIGN	16
5.1 QUERYING THE DATASET	16
5.2 JUPYTER LAB	17
5.3 MAPBOX	17
5.4 POWER BUSINESS INTELLIGENCE	17
6. IMPLEMENTATION	20
6.1 SPRINT 1	20
6.2 SPRINT 2	21
6.3 SPRINT 3	22
6.4 SPRINT 4	24
6.5 SPRINT 5	25
6.6 WEEKLY BURNDOWN.....	26
7. FINDINGS	28
7.1 POPULATION TRAFFIC OBSERVATIONS.....	28

7.2 BIAS IN DATASET.....	29
7.3 MAPPING.....	29
8. TESTING.....	35
9. FUTURE WORK.....	36
10. LEARNING ASSESSMENT.....	37
10.1 BUSINESS LEARNINGS.....	37
10.2 SOFTWARE DEVELOPMENT LEARNINGS.....	38
10.3 WHAT WE WOULD DO DIFFERENTLY.....	39
11. CONCLUSION.....	40
WORK CITED.....	41

Table of Figures

FIGURE 1 RISK MITIGATION LAYOUT.....	6
FIGURE 2 SPRINT PLANNING.....	9
FIGURE 3 KANBAN BOARD.....	10
FIGURE 4 SCRUM DIARY.....	11
FIGURE 5 RISK CULTURE CYCLE (HILLSON, 2013).....	13
FIGURE 6 NEW YORK CITY MARKET RISK.....	14
FIGURE 7 SYSTEM ARCHITECTURAL OVERVIEW DIAGRAM.....	16
FIGURE 8 POWERBI SLICERS.....	18
FIGURE 9 SPRINT 1 USER STORIES.....	20
FIGURE 10 SPRINT 2 USER STORIES.....	21
FIGURE 11 SPRINT 3 USER STORIES.....	22
FIGURE 12 SPRINT 4 USER STORIES.....	24
FIGURE 13 SPRINT 5 USER STORIES.....	25
FIGURE 14 USER STORIES COMPLETED.....	26
FIGURE 15 STORY POINTS REMAINING.....	27
FIGURE 16 HOURS REMAINING.....	27
FIGURE 17 NET POPULATION GROWTH HEAT MAP.....	30
FIGURE 18 NET POPULATION GROWTH HEAT MAP.....	31
FIGURE 19 GEOGRAPHIC POINT MAP.....	32
FIGURE 20 NEW YORK.....	32
FIGURE 21 ANIMATION SHOWN.....	33
FIGURE 22 TOOL TIP.....	33

Table of Tables

TABLE 1 POWERBI DATA DISPLAYED.....	19
TABLE 2 TOP 10 CITIES WITH HIGHEST POPULATION TRAFFIC.....	28
TABLE 3 TOP 10 CITIES BEING MOVED OUT OF.....	28
TABLE 4 TOP 10 CITIES WITH THE HIGHEST NET POPULATION TRAFFIC.....	29
TABLE 5 TOP 5 REGIONS WITH THE NEGATIVE NET POPULATION TRAFFIC.....	30

Abstract

Due to the Global Pandemic, COVID-19, migration patterns for humans in the United States are changing from what they have been in the past. The company we worked for during our project was an alternative investment firm, and some of their biggest investments are in real estate. Our goal was to visualize alternative investment data to help our sponsors view migration patterns within the United States. Our team used Agile Scrum methodologies to develop a Jupyter Notebook and Power Business Intelligence (Power BI) dashboard to provide visuals of migration patterns from March 2020 to November 2020. The notebook and analytics tool we created visualizes the most popular areas within the United States where people are moving into and out of. We developed a dashboard that also utilizes filters and animation to show key demographics like age and median income of the population movement. According to the displayed migration patterns, the data will allow our sponsors to strategize where to invest in next. By using census block group data, we were able to predict foot traffic at different home improvement and appliance stores within the U.S.

Executive Summary

The investment firm focuses on a large range of portfolios including credit, private equity, real estate, and multi strategy investments. With the global pandemic, the firm is changing processes in many of their portfolios. However, their real estate portfolio must change rapidly and adapt to using business intelligence tools to view migration trends and strategize which regions to invest in next. The change in human behavior because of COVID-19 makes analyzing migration patterns important to know due to people staying distant and isolating from densely populated areas. The hot spots in the United States five years ago, or even one year ago, like New York City, Los Angeles, Houston, and Atlanta are no longer seeing the same population net growth that they were before COVID. We worked with the investment firm to create an interactive tool to determine where the population was moving.

We developed a Jupyter Notebook, an open-sourced software that accommodates dozens of computer languages, with visualizations that showed migration patterns within the U.S. We did this through the aggregation of over seven million entries of census block data and cellular data stored in an internal company database. The assumption was made that if a person travels over 225 miles to visit a home appliance or improvement store, they most likely moved close to the surrounding location of that point of interest during a time that it was recommended to stay in place. The stores we looked closely at were Home Depot, Lowe's, Ace Hardware, IKEA, HomeGoods, Walmart, Target, and Bed Bath & Beyond. By doing this, we were able to calculate migration patterns to and from cities.

Using the dataset generated in the notebook, we visualized it further using a PowerBI dashboard. We created an animation showing the change of migration traffic over time starting in March 2020 to November 2020. Through colored maps, we showed major hot spots throughout the country representing which regions were experiencing net growth and loss.

Over the course of the project, our group made use of the Airtable project management software, which supports User Story Cards, to implement the Agile Scrum methodology. We were able to utilize different views of our six one-week sprints, including a grid view and a Kanban workflow. Daily scrum meetings within the team and with the sponsors allowed us to create user stories for each sprint. We went through weekly scrum review sessions to demonstrate our progress with our sponsors which enabled us to receive client feedback. Once we reviewed this feedback, we were able to plan upcoming sprints while continuously improving our notebook and analytical dashboard.

We were able to display interactive maps including many animations that showed a range of several different demographics and a time span. The maps could be filtered to show median income, average age, and distance traveled. From these views we were able to conclude that suburbs outside of major cities were seeing the most foot traffic, therefore, more people were moving outside dense cities and typically away from the country's coasts. Our sponsors were able to use this information and algorithm to view migration trends and predict population movement for the future.

1. Introduction

1.1 Alternative Investment Firm

We worked with an alternative investment firm with offices all around the globe. The firm manages billions of dollars across a range of credit and real estate strategies. Since their start, they have invested in pension funds, corporations, endowments, foundations, sovereign wealth funds and high net worth individuals. The company uses analytics, predictive software, and technology to find investment candidates through the detecting mispricing of securities and assets.

Our sponsors use quantitative and qualitative research as their cornerstone to establish good investment decisions. Within the company, industry specialists and consultants analyze equity and debt markets to create investment opportunities that generate attractive risk-adjusted returns. They credit their success to acquiring assets that are underperforming and employing their skills to improve property performances.

1.2 Opportunity

The pandemic of COVID-19 has caused an economic upheaval, and because of this real estate owners and renters are moving to new residencies. Due to the ease of transmission of the virus, families and individuals in urban areas are finding themselves uncomfortable living in highly populated buildings and areas, taking public transportation, and feeling overcrowded. The pandemic has caused people to look elsewhere to live and the investment firm wanted to view population migration patterns within the United States. In the long run, this will protect their investment capital by allowing them to predict patterns in population movement by analyzing past trends. With the unpredictability of the future of this virus and migration movement, it is important to analyze current migration trends to help make smarter investment decisions based on population hot spots.

1.3 Project Goals and Scope

We researched and gathered alternative data alongside our sponsor's research team to see where the United States population migrated to due to COVID-19. The goal for our team was to gather alternative investment data from different resources into a program where we analyzed and visualized such information. We created a notebook and application that provided visual data showing population migration patterns pre and post COVID-19.

To analyze this data, we developed a neighborhood recognition algorithm focused on the United States. Specifically, this algorithm showed a picture of people moving into and out of different census block groups. Using alternative location data, financial information based on geography, we created a program that will be included in the investment firm's core systems to visualize what neighborhoods look like now, looked like in the past, and what they will look like in the future. They will be able to then invest in prospective neighborhoods whose population and economy is growing so they will see a high return in their investment.

2. Background

2.1 Alternative Investment Firm Industry

2.1.1 *Alternative Investments*

An alternative investment relies on tangible assets rather than having a continuous income of cash and finances. This industry invests in private equity, venture capital, hedge funds, real property and other commodities that are often illiquid (Chen, 2020). Alternative investment companies are looking to create higher returns beyond their cash, stocks and bonds. Money coming from alternative investments are also taxed less, which makes them more favorable (Kennon).

2.1.2 *Real Estate*

The real estate portion of this investment firm's portfolio dominates their business. Investing in real estate can allow for "absolute returns, hedging against unexpected inflation, diversification against stocks and bonds, steady cash flows, and income tax advantages" (CAIA, 2016). Investing in real estate is often one of the biggest risks that is found in a portfolio due to the constant change in human behavior.

With COVID-19, the real estate market is changing rapidly, and being able to track movements in human behavior to then make inferences about the future, is a key part of a strategic plan to predict which areas would potentially have high investing returns. Therefore, this project was immensely important to our sponsors.

2.2 Geometric Data Polygons

Our team did research on geometric polygons, or simplified boundaries, to have an understanding on the data and information we would be looking at daily. The Census Bureau generates this tracking data which is what we used for the design of our notebook to show migration trends. The points of interest we used all had their own geometric polygon which allowed us to track foot traffic.

2.2.1 *Census Bureau*

The United States Census Bureau is responsible for collecting data on the American people and economy. The information shows the overall population of the United States but breaks it down further into different groups to analyze changes over time. The Bureau collects information on age, sex, race, and economic status; and looks at birth, death, and immigration rates. From this information, demographic trends are predicted. Both the Economic Census and Census of Governments are collected every five years to measure the economy at a national level, broken down into geographic areas, and provides comprehensive data about all the states, counties, cities and towns throughout the nation (Bureau, 2019).

2.2.2 *Geospatial Data*

Geospatial data provides information about events, spread of infectious diseases, landscapes, and populations. This data uses "location information, attribute information, and temporal information" to create a clear picture of what is happening on Earth's surface. It provides information for public and private sectors to get an idea of different geographical landscapes (Stock & Guesgen, 2016). The most interesting thing about geospatial data is studying the ontology as to why humans are living where they are, which is what the investment firm is looking to understand.

2.2.3 Geo-fencing

Predefined fences are put in place in different geographical areas to act as “perimeters” around a certain point of interest. These points can often be businesses, schools, entertainment services, stores and more. Each of these points has the potential to have their own fence. Once a person carrying a cell phone walks into this designated area, their phone may send a notification about what is around them. Geo-fencing allows these places to send direct advertising to a person's phone based off their location in proximity to them. A person's phone often connects to Bluetooth, Wi-Fi, and Global Positioning System, or GPS. Therefore, smart devices, if allowed by the user, are always location aware. Geo-fencing places these boundaries on areas thus creating different and separated geographies. Data like pollution, foot traffic, etc. can be analyzed within each boundary. Geo-fencing is desirable for an investment firm because they can analyze what types of areas migrating people are the most attracted to (Pongpaichet, Singh, Jain, & Pentland, 2013).

2.2.4 Census Block Groups

Census block groups are the smallest geographical blocking that The United States Census Bureau creates to collect statistical data. The Census Bureau maps these blocks by streets, roads, railroads, rivers, and other physical and cultural features. This data collection helps mostly with large urban areas, where data can be collected block by block and provide detailed housing information to allow cities to modify building codes, establish zoning ordinances, and prepare plans for improvements. To invest, it is important to understand which cities and areas are the most popular and growing, which typically indicates increasing property values, but census blocking groups take this data even further and show which streets are becoming more popular (Bureau, 2019). Comparing this data to census economic data is important because it shows which areas are growing, potentially increasing property value, and why.

2.3 Software Development Environment

2.3.1 Python

The programming language we used for this project was Python. Python is an “interpreted, object-oriented, high-level programming language with dynamic semantics” (Python). It is favored by many people because of its increased productivity, simplicity, and easy-to-learn syntax. Python is often used in software development companies such as “gaming, web frameworks and applications, language development, prototyping, graphic design applications” and more (Mindfire Solutions, 2017).

Anaconda is an open-source distribution of Python. Anaconda is coined “The World’s Most Popular Data Science Platform” for its widespread popularity but also its out-of-the-box capabilities and simplified package management. Throughout the entire scope of the project, Anaconda was used for Python development.

Visual Studio is the IDE that we used for development during this project. IDE stands for integrated development environment which assists with writing and compiling code. It is popular because it can “combine common activities of writing software into a single application: editing source code, building executables, and debugging” (IDE). Visual Studio supports many languages like C, JavaScript, Python, etc.

2.3.2 Pandas

Pandas is a widely popular Python library with powerful data analysis and data manipulation tools. Its primary feature is the use of data frames, which are identical to a traditional database table. This allows for data to be easily grouped, aggregated, sorted, and merged.

The set of tables that we pulled data from were all unindexed. Using Pandas, the data was aggregated and grouped to provide useful analytics within a reasonable amount of time.

2.3.3 Jupyter Notebook

Jupyter Notebook is a web-based IDE that allows users to write code, visualize data, model statistics and share documents for professionals. Jupyter Notebook is favored by many data scientists for several reasons. All notebooks are saved as text files and thus are easy to share between users. It also has a built-in tool called nbconvert to convert notebooks to other formats like HTML or PDF. The extensions that Jupyter Notebook has also makes it a very convenient tool. For example, people can write Python code directly in Jupyter with some third-party extensions and magic commands which are commands specific to the IPython shell (Kumaraswamy, 2018).

2.3.4 PostGIS

PostGIS is a PostgreSQL extension where PostgreSQL is an object-relational database. It adds support for geographic objects allowing for location queries to be run on the database. The location awareness is a special feature of PostGIS, and especially useful for this project as analyzing locations was a requirement.

2.3.5 Mapbox API

Mapbox is a powerful application programming interface (API) that has customizable and flexible mapping capabilities. The Mapbox API is a RESTful API meaning that it is accessed through JSON-serialized GET requests. For easier access, the mapboxgl_notebook and mapboxgl_jupyter Python packages were used. This allowed for the API to be accessed in an object-oriented way making it easier for maps to be generated.

We primarily used the Mapbox API to create geographic point maps marking the different points of interest and their associated traffic. The maps were displayed in Jupyter output cells.

2.3.6 Power Business Intelligence (PowerBI)

PowerBI, according to its official definition, is the tool that “connects to and visualizes any data using the unified, scalable platform for self-service and enterprise business intelligence (BI) that’s easy to use and helps you gain deeper data insight” (PowerBI). PowerBI has secure data encryption and is also easy to integrate with other applications, personalized, quickly retrievable and efficient. It is an excellent front-end tool and creates a clear picture for projects that focus heavily on back-end development.

2.3.7 GitHub Source Control

For source control, our team decided to use GitHub. GitHub is a repository hosting service with many features (Finley, 2012). The open-source command line tool, and repository hosting service allowed our team to develop code on their own individual branches and push and pull code into the main branch where our code merged to create our final product. This proved to be necessary since this project was done remotely and allowed for collaboration and contribution from each member.

2.4 Project Management Tools

2.4.1 *Airtable*

Our team decided to use Airtable, a cloud based collaborative tool, that allowed us to create Kanban workflows and spreadsheets. Please see section 3.2 Software Development Methodology for more information on Kanban. Airtable permitted each team member to view and organize weekly sprint tasks. Using Airtable, members could check off functionality and tasks to ensure that user story requirements were met. Airtable was used in every meeting to track what work had been accomplished through different views, a spreadsheet, and a Kanban diagram.

This tool also enabled us to calculate user story points and hours worked. We needed this to calculate our team velocity, and to associate risks with each story. Details about our sprints and user stories can be found in the Implementation chapter. We initially anticipated using GitHub because of our previous experience with it. However, we quickly switched to Airtable to keep track of our sprints, making our sprints visually pleasing and easy to show and read during our daily scrums and sprint meetings with each other, our advisors, and sponsors.

2.4.2 *Communication*

At the start of our project, the alternative investment firm connected us to a remote virtual machine so we could access their data and communication software. Our sponsors created email accounts for us so we could communicate via Microsoft Office. We used Microsoft Teams for instant messaging communication and video chat. We also used Microsoft Outlook to schedule meetings and send lengthy messages and reports

To communicate with only our team members daily, we used an instant messaging mobile app, GroupMe. Not being in person often, this allowed us to step away from our computers and still be in communication. We also took advantage of our WPI's school license with Zoom to meet virtually every day with each other and often with our project advisors.

2.5 Project Risk Analysis

For this project we determined risks in advance so we would be able to mitigate them when they came along. Risks for each sprint can be seen in the Implementation chapter later in the report.

2.5.1 *Types of Risk*

We assumed that we would encounter four diverse types of risk during this project: technical, inexperience, time, and dependency. To prepare for these risks, we kept track of what type of risk we thought we would encounter, why it was a risk, and briefly described risk mitigation. Figure 1 refers to the layout that we created with our project management software to view and handle these risks. After sprint one, we created plans on how we could mitigate these risks and be ready when they occurred. The different risks that occurred are listed and described below.

As a user, I want to		Risk	Type of Ri...	Risk Explained	Risk Mitigation
SPRINT COMPLETED ▼ Sprint 4		Count 11			
30	Be able to zoom in and out on heat map and continue to see color and density of foot traffic	Medium ...	Technical Inexperience	Not sure if PowerBI will allow for zooming and keeping color density	May have to push to sprint 5, but will stay in communication with sponsor
31	Have tool tips appear when hovering over different CBG's on the heatmap	Medium ...	Technical Inexperience	PowerBI may not have this functionality, and be able to do this.	Going to talk to sponsor and see if he wants us to continue PowerBI and/or plot the heatmap with ...
32	See home city and state data that is pulled from a more complete dataset to create better statistics	Low Risk	Technical	Same code, but just pulling from a more complete dataset to represent the statistics ...	Communicate with sponsor since he knows where to look to find the right data
33	Also have a heat map of where people are moving from	Low Risk	Dependent	Waiting on getting a more accurate dataset	This should be relatively easy. Has already been done before, but is reversed. Make sure we are usin...

Figure 1 Risk Mitigation Layout

Technical – these risks were associated with the virtual computer provided, access to the firm’s entire data set, software limitations, and problems with inaccurate coding and data.

Inexperience – this subset of technical risk realized that all team members did not have the same experience with programming and software environments. This was resolved by giving members of the team cross functional roles.

Time - the entirety of the project was only seven weeks, so the team had to work quickly. Each sprint and story had time allotted to it to ensure a successful sprint. If there were blockers or a story that required too much effort in just one sprint, it was communicated to the sponsors and moved into another sprint or discussed further.

Dependency – this risk came about once the project was started and the team realized future steps relied on previous steps being completed because much of this project was linear.

2.5.2 Changing trends with COVID-19

Due to the unprecedented times with the ongoing pandemic, our group was prepared to adapt to any change that may displace us over the length of the project. At the beginning of this project all members of our team were able to be on campus and meet in person in a safe manner to work together. We anticipated that Worcester Polytechnic institute may have changed their school regulation, due to rising infection case numbers, forcing us to only work from home.

We were prepared to work remotely using Zoom and Microsoft Teams as a way of communication to complete the project. Similarly, the firm was prepared to work with us virtually, and they were dedicated to our term long project being completely online under these circumstances.

Regarding our own personal health, the team went through weekly health screenings and COVID-19 testing. If someone had tested positive for the virus or felt unwell, we would have made sure to continue our team meetings virtually and quarantine if necessary. About halfway through the project, WPI instituted heightened campus restrictions so we discontinued meeting in person.

2.5.3 Technology Experience & Usefulness

With a limited amount of time, the biggest risk that may have occurred was software compatibility risk. This risk comes with creating any software technology. Once we gained remote access to the investment firm, we worked with some unfamiliar software, so it took some learning. Part of our weekly sprints was

dedicated to learning modern technology and familiarizing ourselves with how to make our program compatible for our sponsors at the firm.

3. Methodology

3.1 Project Management

Due to the fast nature of the MQP at WPI, it was imperative that our team adapted to constant change in our work environment. This unpredictable change required agile approaches, and evolutionary development to complete the major task at hand. To complete this project, our team went through regular, continuous rounds of feedback and development. Our style followed the Agile Scrum process to continuously check our work and provide high quality software to our sponsors. Members of our team and the investment firm came together in daily scrums to discuss progress and give feedback.

We implemented the Agile framework often used in delivering a software product to a client. What makes agile work so well in this seven-week term long project is the ability to adapt to fast changes. It was beneficial for letting us grow, and learning and trying new techniques. If failures occurred, we were able to correct them quickly and finish the project on time for the client. For this framework to work, we needed to have clear communication between the client, product owner, and development team.

3.2 Software Development Methodology

3.2.1 Scrum

Scrum is an Agile framework used in product development. It is a feature-driven framework that delivers functional updates quicker than other methodologies. Our team decided that we would operate on a one-week sprint schedule. This duration of sprint allowed us enough time to make meaningful additions while not allowing much time without new software iterations.

This process held members of the team accountable for different roles. Our team was made up of three separate roles which included the Product Owner, Scrum Master, and Development Team (Ruhe & Wohlin, 2016). To make the most out of these roles, each member assumed one role but was expected to have responsibility in each.

The Product Owner was responsible for creating and communicating business objectives to the clients, Scrum Master and Development Team, through the organization of the product backlog (Vanderjack, 2015). Our product backlog was a collection of user stories, or deliverable goals created through our sponsors requests, which came to be implemented during the scope of the overall project (Vanderjack). Backlogs and user stories were created by understanding and capturing the clients wants, needs and ideas. The Product Owner was in control of how many points and stories were allowed into each sprint to make sure the team was not overwhelmed, but still worked to its full capacity. The Product Owner worked with the Scrum Master to remove any potential blockers that the development team may have experienced.

The Scrum Master ensured that development was organized in a way that allowed for agile principles. A key component that came with this role was developing the points and hours for each story and task (Vanderjack). This aspect of the role allowed the team to debate how many points and stories would be allowed in each sprint. This maximized our effort devoted towards the product development.

The Development Team created and tested programs that were written and brought the client's product to life by completing each user story (Vanderjack). The Development Team constantly communicated with the whole team to have a clear vision of what work could be completed within sprints and expressed when problems and roadblocks occurred.

3.2.2 User Stories

User Stories, that were first developed through active listening and open dialogue with project sponsors. Once the team felt like they had an understanding on what the project sponsors wanted, user stories were started with “As a user, I want to” and finished with deliverables that would be later shown to the sponsors. Stories that were created but not put into a specific sprint were put into the product backlog to mold the scope of our project. If our team completed all the user stories in the sprint backlog, we took more stories from the product backlog and improved our work.

The point system we used for our stories was a one-to-one ratio with hours and points. The more time something took, the more points were allotted. Doing this gave us a clear view of what time and effort was needed to complete the sprint. Stories were ranked in the order of highest importance which helped with our linear progression.

Although the team used GitHub as our source code repository, Airtable was used for project management. This led for a clear view of sprint planning and a Kanban flow of stories. When the Product Owner and team agreed that it was time to bring stories into a sprint, stories were moved from the backlog and were assigned to the development team depending on which member of the team wanted to complete the task. From there, the team was to discuss functionality, tasks, risks, risk type and mitigation, story points and hours of each story. Refer to figures 2 and 3 to see an example of our dashboards over the course of the project

As a user, I want to	Date Started	Date Completed	Rank	Status	Assignee	Sprint Completed	Time..
SPRINT COMPLETED							
Count 8							
Sprint 2							
View a map of the chosen stores within The United States	11/6/2020	11/6/2020	1	Completed	Garrett Stephenson Luke Trujillo	Sprint 2	4
View the foot traffic for each chosen store in The United States	11/6/2020	11/9/2020	2	Completed	Xiaowei Chen	Sprint 2	4
View home visits from customers 50 plus miles out from chosen stores POI	11/6/2020	11/9/2020	3	Completed	Luke Trujillo	Sprint 2	3
Tell the median income of people coming from CBG's outside the POI's 50 mile parameter	11/9/2020	11/10/2020	4	Completed	Xiaowei Chen	Sprint 2	3

Figure 2 Sprint Planning

As shown in Figure 3, we were easily able to drag stories from our backlog into progress. When a story was completed, we once again dragged the story to the Completed column, and a cycle of pulling from the backlog was created. This view provided a quick snapshot on what was being worked on and who was working on what.

3.2.3 Kanban

Our Kanban, or pull system, is another view and system we used for this project. This just-in-time method used for lean manufacturing allowed us to work on the project in a linear fashion (Ríos-Mercado & Ríos-Solís, 2012). In our Kanban view, we could “pull” stories from our backlog when one had previously been finished. Although we used Agile Scrum, the Kanban boards helped us visualize our tasks and helped with organization and efficiency. Our project progressed steadily because of this view we were able to create.

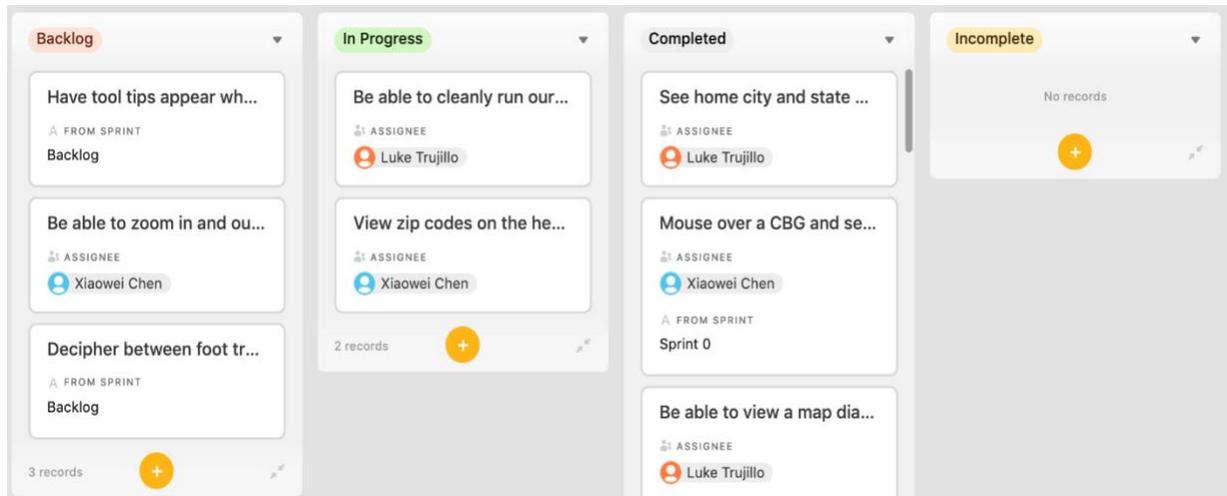


Figure 3 Kanban Board

3.3 Requirements Gathering

At the beginning of our term, our team set up our daily meeting schedule with each other, our sponsors, and our advisors. To implement Agile in this project, it was especially important to be able to communicate daily, especially being done remotely. We gathered our requirements from our sponsors with the methodology discussed below

3.3.1 Daily Scrum Meetings

Team Scrum Meetings

The team met every morning of the week to discuss progress on the sprint. We kept a scrum diary which gave us a guideline on how to move the conversation. Every morning, each member of the team discussed what they had completed since our last meeting, what they were working on today, and what they were having trouble with. We logged this in a scrum diary that gave us the opportunity to help each other with questions and think about doing things differently. Figure 4 was the format we used to help us refer to conversations on previous stories and sprints.

November 22, 2020

Luke:

- Completed:
- Working on:
- Trouble with:

Xiaowei:

- Completed:
- Working on:
- Trouble with:

Garrett:

- Completed:
- Working on:
- Trouble with:

Figure 4 Scrum Diary

[Sponsor Scrum Meetings](#)

Microsoft Teams was used for the team to have video stand ups with our sponsors every afternoon to discuss our progress. We showed them our code, and they often suggested improvements for us to work on to shape the product to their liking. It also served as a time where we communicated any blockers or risks that we saw in our upcoming stories. This kept the sponsors continuously thinking about innovative ideas and changes they wanted to see and kept the team ready to agilely adapt to any adjustments.

[3.3.2 Sprint Planning](#)

At the beginning of each sprint, our team met and decided on which user stories we would work on for the weekly sprint. In this meeting the Product Owner collaborated with the Development Team to negotiate what user stories were moved from the product backlog and developed. We did this by communicating with one another about what our ideas were moving forward with a combination of input from our sponsors at the alternative investment firm.

[3.3.3 Sprint Review](#)

Before each sprint review with the sponsors, the team met to discuss how the sprint went and prepared a summary of the sprint to present to them. The sponsors received a dashboard summary in form of an email to give them a chance to review the sprint beforehand. This included all the user stories, velocity of work, and data. With all this information present, the team reviewed the results of the sprint with our project sponsors and demonstrated the user stories that were successfully implemented during this sprint.

Due to the constant virtual communication and demonstration, there was nothing for our sponsors to be surprised about. This clear communication, allowed for us to quickly review the work, analyze the data, and begin discussing the next steps forward.

3.3.4 Sprint Retrospective Meeting

In addition to the sponsor meeting and review at the end of the sprint, the team held a Sprint Retrospective meeting where the team analyzed how things went during the previous sprint and brought any outstanding communication, team, or technical issues to the attention of the entire team. This meeting allowed us to analyze what worked well, and what could be improved for both the group and individuals.

4. Business and Project Risk Management

4.1 Risk and Reward

In any investment there are always risks but investing in real estate can be especially volatile due to the unpredictability in the real estate market. Real estate can seem to be rather remunerative until it is not. For example, the Great Recession of 2008. Investors continued to predict an unchanging market that would continue to grow in an upward linear direction. Excessive borrowing, too much risk taking, lack of regulation, or inadequate regulation on the financial industry led to a steep decline in the United States economy. This unpredicted recession led to economic failure for millions of U.S. Citizens.

COVID –19 is another unpredicted event that has changed the lives of everyone on Earth. COVID-19 has affected people in ways unimagined. With this airborne virus, people have been asked to stay far apart from one another. Because of this, mainstream media and housing data has shown that a large amount of the population is moving out of overpopulated urban areas. Human behavior is changing and so is the real estate market. Therefore, the risk taking the time and allocating the resources to track human movement to predict what may happen in case of another recession, and or epidemic is important. Hopefully, this will help the investment firm come out on top in the future. If the investment company can analyze and predict movement of the population in the future the return will be rewarding.

If this project is not successful, the investment firm will lose potential future profits based on investing in popular areas to live in. They could continue to use their old ways of investing and continue to invest in urban areas; however, they would miss a large market that is moving away from crowded cities.

Ultimately, relying on this project and future advancements in code and analysis should allow the investment firm to see country hot spots and improve their real estate investment efficiency. Although they may lose revenue from previous investments, they should be able to increase revenue in suburbia's and areas outside of major cities based off population tracking.

4.2 Risk Culture

For the investment firm, to survive, they must take risks. Taking previously accumulated data and identifying patterns to predict what may occur in the future is a great peril. Science shows that COVID and its possible mutations may be around in the future, but human behavior is constantly changing, and people could be drawn back to large and crowded cities. The investment firm needed to take a chance in this predicative thinking to be prepared for the future.

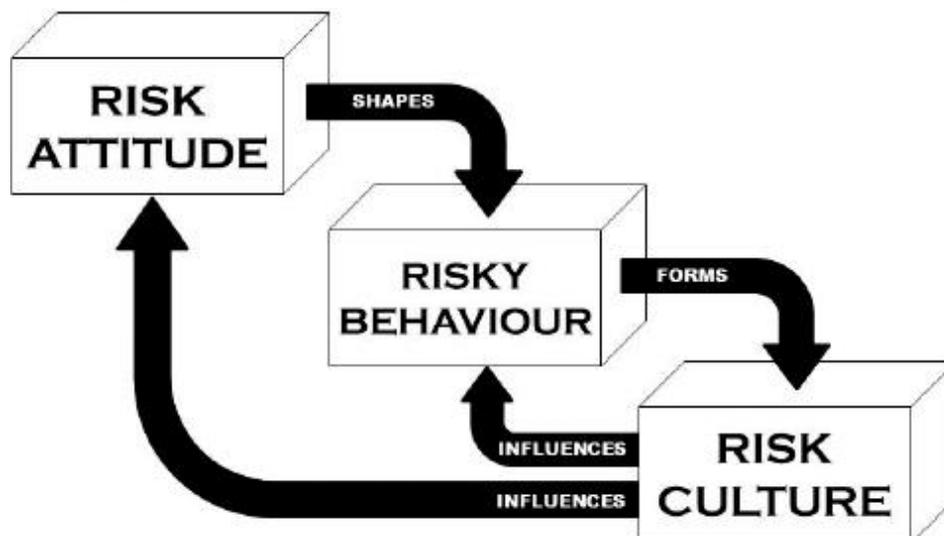


Figure 5 Risk Culture Cycle (Hillson, 2013)

The company has control of the data they currently have and how they analyze it. They can pull from different data sets to see where people are moving from and moving to, but that does not mean that the data will not change the next day. It will be important for the investment firm to continue updating our work to stay current and continue assessing risks.

4.3 Operational Risks

With our virtual machines, we were only given limited access and processing ability. Because of this, we kept our dataset small. Overtime, to analyze more accurate data, the investment firm will be able to take our notebooks and expand upon them, including more points of interest and data points to look at.

There are not many operational risks besides the possibility of strict privacy data laws. This could potentially get in the way of our sponsors future investments. This project was done by looking at people's locations based on each person's mobile home phone and census block group location. Increased privacy laws would increase the limitations on the complete story the data curated shows.

4.4 Market Risks

With changing conditions during the pandemic, properties around the country are seeing price fluctuations that are not been seen previously. Cities in 2019 that were considered some of the most expensive to live in are seeing decreases in value, while cities and suburbs that were not as expensive to live in in 2019 are now seeing price increases. This shift in the market shows that the range between rent prices in places considered to be of greater value and areas of lesser value are seeing a more even distribution. According to Forbes, since March 2020, the top five rental markets including San Francisco, New York City, Boston, San Jose, and Oakland have been seeing a significant downward trajectory in their median rent since the beginning of the pandemic in the United States (Richardson, 2020).

Contracts Signed Year-over-year

Properties under \$2 million are selling again while pricy real estate lingers

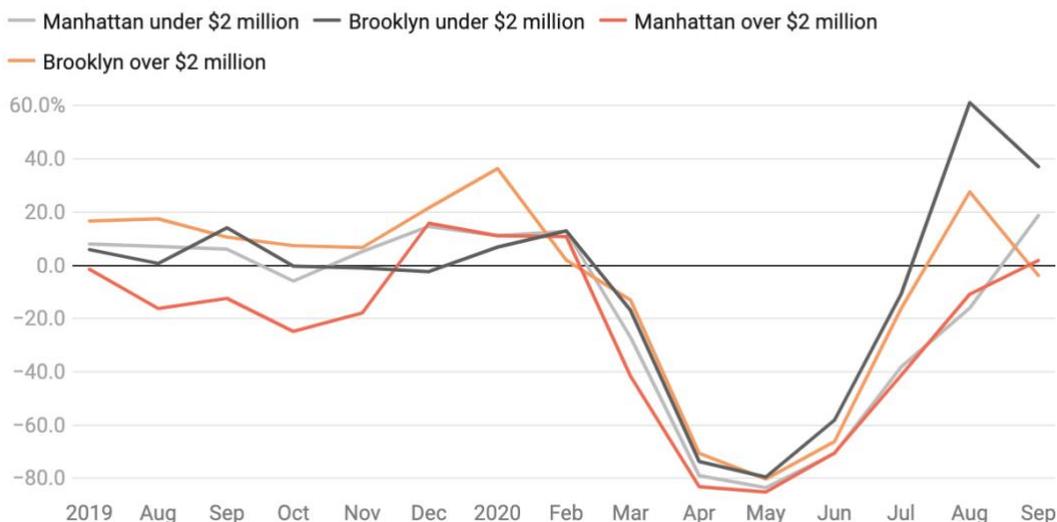


Figure 6 New York City Market Risk

Figure 5, from Curbed.com (Andrews, 2020), shows the norm of contract signing in New York City. In March, contract signing took a nosedive considering a peak of COVID and the city lockdown. During this drop, people were moving elsewhere, and the point of this project was to find out who and where.

4.5 Change Management Risks

The purpose of this project was to be used in longevity as time moves forward. The idea is for the notebook to be expanded to see movement patterns from the past and be able to predict future outcomes. This project will be used within the data science team we worked with to be a leading indicator for understanding neighborhoods and migration patterns. We gave our final deliverables over to the investment firm, but the project needs to adapt with time to remain relevant in the future of the company.

5. Design

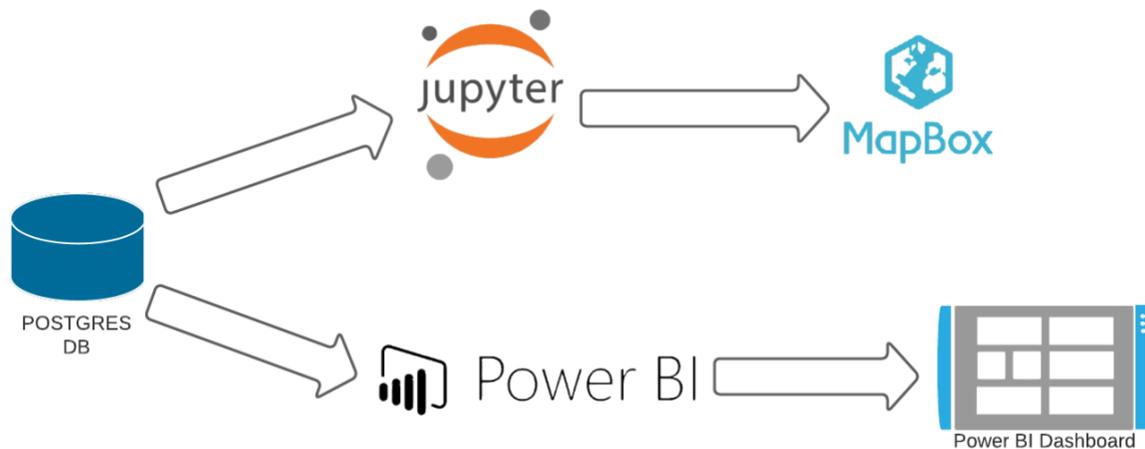


Figure 7 System Architectural Overview Diagram

As shown in the diagram above, all data was sourced from an internal POSTGRES database with over 15 million entries of migration data. This data was then queried using Python’s pyodbc library in the Jupyter Notebook. The data was aggregated, and metrics were calculated using Python’s pandas library. This data was loaded into Power BI and used to generate animations with configurable sliders. The Jupyter Notebook, with the dataset aggregated, was used to natively display Mapbox interactive maps inside of the notebook output.

5.1 Querying the Dataset

For the entire scope of the project, all the data that was used to make our migration maps was sourced from an internal POSTGRES database. The Jupyter Notebook, which contained the code for querying the dataset had configuration options that affected the output of the query. The first option was the location name of various businesses. For the duration of the project, the location names that were used were: The Home Depot, Lowe’s, Bed Bath & Beyond, IKEA, Target Corporation, Ace Hardware, and Walmart. These location names were intended to be common locations that someone who recently moved would frequent and were chosen by the project sponsor. The second configuration option was the start date for the dataset. Given the project was focused on analyzing migration trends after the start of the COVID-19 pandemic, the start date used was March 1st 2020. The final configuration option relating to the database query was the datapoint limit. The size of the tables queried was over 15 million entries and Jupyter Lab consistently crashed after querying around 7 million rows. Therefore, to prevent this issue a datapoint limit configuration option was created.

5.1.1 Limitations of Database

Majority of the tables did not use indexing or database standards which severely increased the query time. In order to create our models, it required the merge of three different database tables with entries totaling over 15 million. The lack of relational optimization and indexing served as a bottleneck for the rate of development throughout our entire project. Our team was only granted read-only access so simple optimizations through the creation of “Views” was unavailable to us. In order to overcome and

minimize this issue, nearly all aggregation and processing occurred on the local machine using the pandas library.

The data we were provided was essentially the number of cellular devices from one census block group that visited a given location on a weekly basis. This data was joined with other tables to translate the home census block group location to a city and state. Then population traffic was computed for a given location by using the following equation:

$$\text{Population Traffic} = \frac{\text{Number of Devices that Visited the Location from a CBG}}{\text{Total Number of Devices Being Tracked in the CBG}} \times \text{Population of the CBG}$$

Then the data was aggregated together to compute the total population traffic that a location has experienced.

5.2 Jupyter Lab

Jupyter Lab was the development environment used for generating interactive Mapbox maps and displaying tables of information. Using the pandas library, the data queried from the database was aggregated. After, the population traffic and distance travelled was computed using lambda functions. By using Jupyter Lab, we created data frames in individual cells reducing the amount of coupling in the notebook. Then, using a Python library for interfacing with Mapbox the aggregated data was supplied to Mapbox and an interactive JavaScript map appeared in the output view of a cell.

5.3 Mapbox

The Mapbox graphs served as one of the primary visualizations of the dataset. The location of the configured businesses were plotted on a point map. There were three of these maps generated and each map was colorized based upon the value of a different column. The columns that one map was colorized on were: Median Household Income, Median Age, and Population Traffic. There was additionally one more point map generated which plotted the cities that people moved from for a given city and state. This map was colorized based upon the Population Traffic.

5.4 Power Business Intelligence

For the built-in heatmap of PowerBI, there were two rendering types. One is the contour map mode, and the other is the real heat map mode, and we relied on the heat map mode. In this mode, red represents the highest level of population growth while blue means the lowest level of population growth. Therefore, the darker the red is, the more people are coming in; the lighter the blue is, the fewer people are coming in.

The geographic point map is one sub type of the Mapbox visual. In this map, the darker the blue is, the more people are coming in. Different from the heatmap, the geographic point map does support showing up the negative values, and thus the darker the red is, the more people are leaving. Yellow represents 0 net population growth. The colors to represent the population growth in geographic point map are reversed from the colors used for heatmap as the project sponsors asked us to do so.

The built-in heatmap supports animation itself by giving it a "group" value. For example, if we set the group value to be the date, the map will animate its population growth based on different dates. The geographic point map does not support animation itself, so we imported a visual from file. The visual was called Play Axis, and once we set the value, the animation was played based off of that specific value.

The slicer is a tool that we used to filter the data (see Figure 8 below). The slicer has two modes, depending on what kind of data is put in. If we put age into the slicer field, it will automatically provide a range slicer as the age values are continuous and hard to split down. We can select what range of data we want depending on the parameters. We can also select a certain region to get more specific. For example, if we only want data with people aging from 25-30, we can narrow down the age slicer, and the map will only show up data within this age range.



Figure 8 PowerBI Slicers

The PowerBI table is a tool we used to display the data (see Table 1 below). We put various types of data into the table field, to aid the visualization of our maps. The PowerBI table supported many spreadsheet capabilities like sorting and filtering which allowed us to manipulate the data to directly see the data shown on the map.

City	State	Net Population Traffic	Average of distance_in	Average of age_in	A
Chicago	IL	-57238	715.35	39.81	
Houston	TX	-48651	650.43	37.34	
Miami	FL	-40324	971.67	38.04	
Atlanta	GA	-39164	564.12	39.76	
Brooklyn	NY	-38133	833.08	42.17	
Fort Lauderdale	FL	-33328	1,011.22	38.72	
Saint Louis	MO	-28517	543.28	39.00	
Los Angeles	CA	-27850	1,185.64	40.38	
Philadelphia	PA	-22756	889.12	39.68	
Milwaukee	WI	-21307	688.72	37.61	
Total		5741052	703.25	40.23	

Table 1 PowerBI Data Displayed

6. Implementation

To implement our user stories, we gathered our requirements from our client through our scrum meetings with them. We then used our sprint planning meeting to put these stories together to create our sprint backlog.

6.1 Sprint 1

As a user, I want to	Rank	Risk	Type of Ri...	Risk Explained	Risk Mitigation
4 View a map of the Home Depots within The United States		Medium ...	Technical	Garrett does not have that much experience with programming but will work with Luke to complete	
5 View the foot traffic for each Home Depot in The United States		Low Risk	Technical	Database pulling the write data	
6 Tell the median income of people coming from CBG's outside the POI's 50 mile parameter		Low Risk	Technical		
7 See popular CBG's 50 miles plus that have visited Home Depots colored darker		Low Risk	Technical	Pulling the right data	
8 View home visits from customers 50 plus miles out from POI		Low Risk	Technical	Pulling the right data, and how to show it	

Figure 9 Sprint 1 User Stories

Sprint Review

The first sprint began on October 28th and our main goal was to be able to view each Home Depot within the United States (see Figure 9). From there, we gathered data that showed the amount of foot traffic for each store and then determined which foot traffic was coming from a census block group from more than 50 miles away. We also looked at the median income from each block group. Doing this allowed us to see money movement patterns of people who shopped at Home Depot for potential moving projects.

We predicted risks related to technical issues in connecting to the database and not having administrative access. Because of this, the data took a while to load as we were programming but that did not have any major effects. The project manager took on a user story so there was an inexperience risk, however, this was mitigated with help of the Development Team.

Sprint Retrospective Meeting

What worked well

- Work seemed equally split up for the Development Team.
- Daily communication with sponsors allowed for adjustments to user stories.

Improvements

- Have more organization of code in our notebooks. This will help moving forward with collaborating on code through GitHub.
- After sprint one there was now a backlog, so we could pull from that to work more hours because it seemed like the estimated hours was low.
- Added more stories in the next sprint

Story Points Completed (Velocity): 90
 Hours Worked: 90
 Percentage of Story Points Completed: 100%

6.2 Sprint 2

As a user, I want to	Rank	Risk	Type of Ri...	Risk Explained	Risk Mitigation
9 View a map of the chosen stores within The United States	1	Low Risk		Will have to take the code we wrote for Sprint 1 and Home Depot and apply it to different stores.	
10 View the foot traffic for each chosen store in The United States	2	Low Risk		Will have to take the code we wrote for Sprint 1 and Home Depot and apply it to different stores.	
11 View home visits from customers 50 plus miles out from chosen stores POI	3	Low Risk		Will have to take the code we wrote for Sprint 1 and Home Depot and apply it to different stores.	
12 Tell the median income of people coming from CBG's outside the POI's 50 mile parameter	4	Low Risk		Will have to take the code we wrote for Sprint 1 and Home Depot and apply it to different stores.	
13 Be able to see a table that provides information about where people are moving to	5	Medium ...	Technical Time	This should be one of the last stories that is finished. Need data from all the other stores th...	Key is getting the correct data from different store databases
14 Be able to see a table that provides information about what CBG's people are moving out of	6	Medium ...	Technical	This should be one of the last stories that is finished. Need data from all the other stores th...	Key is getting the correct data from different store databases
15 See a tool tip showing distance traveled	From B...	Medium ...	Time	Started on 11/9	May move to sprint 3
16 See data starting in March 2020	From B...	Low Risk	Time	Started on 11/9	

Figure 10 Sprint 2 User Stories

Sprint Review

Sprint 2 began on November 5th and the team set out to expand the data collected and visualize it differently (see Figure 10). In sprint 1 the team looked specifically at Home Depots and foot traffic coming to the point of interest from 50 miles or more within the United States. This visualized where people and money moved to. The goal was to expand the code and data and visualize more than just Home Depot, and view foot traffic coming from more than 225 miles away. The geometrical map now showed different stores including not only Home Depot, but IKEA, Target, Walmart, Lowe's, Bed Bath & Beyond Inc., HomeGoods, and Ace Hardware. All stores were assumed to be visited when people moved to a new location. There is a mix between furniture, hardware, and decorative stores.

The data that was pulled was put into a geometric map was also put into a table that sponsors could easily read and view showing popular places where people moved from and to. The team pulled stories from the backlog to create this table with filters so the sponsors could see different demographics like age, income, and mile range.

Sprint Retrospective Meeting

Even though the heat map story was worth a lot of points for the sprint, due to the unfamiliarity and amount of effort, it held a low priority. The sponsors wanted different visuals to be created eventually, but it was more important to get stories with higher priority done in this sprint.

Also, the team experienced our own blocker, that we quickly overcame on the first day of sprint 3. The team came to believe that the data that we produced was inaccurate for people moving out of CBG's. The team spent the beginning of the next sprint testing and debugging the code. Because of this, what we thought had been completed we moved to sprint 3 due to inaccuracies, thus driving our velocity down.

What worked well

- Work in sprint 2 was more evenly divided compared to sprint 1
- Pulled from an abundant backlog at end of the sprint
- Communicated with sponsors/team about moving user stories to sprint 3

Improvements

- Made better estimates for the hours worked at the beginning of the sprint
- Maximized the amount of effort and time in the duration of sprint

Story Points Completed (Velocity): 78

Hours Worked: 99

Percentage of Story Points Completed: 79%

6.3 Sprint 3

	As a user, I want to	Rank	Risk	Type of Ri...	Risk Explained	Risk Mitigation
17	View which cities are gaining people	1	Low Risk	Dependent	We already have the colors and data. Just showing net gain or loss now.	Find the net growth of each city
18	View which cities are losing people	2	Low Risk	Dependent	Based on what data we already have	Find the net growth of each city
19	See the movement in and out of the city with arrows	3	Low Risk	Technical	Need to learn how to draw arrows on map. New to us	Find the net growth of each city. Communicate with sponsor, he knows where to look
20	See a map for a specific cities showing where people are moving to	4	Low Risk	Technical	Just taking the data we already have and breaking it down into the city level	Find the net growth of each city. Communicate with sponsor, he knows where to look
21	View map with age distribution	5	Low Risk	Technical	We already have this age distribution in a table. Now just need to show it on a separate map.	Luke & Xiaowei will help Garrett
22	View map with wealth distribution	6	Low Risk	Technical	We already have this age distribution in a table. Now just need to show it on a separate map.	Luke & Xiaowei will help Garrett
23	Be able to filter table by age	From B...	Low Risk	Dependent	Table needed to first be finished	Testing data to accomplish previous sprint
24	Be able to filter table by income	From B...	Low Risk	Dependent	Table needed to first be finished	Testing data to accomplish previous sprint
25	Be able to filter table by mile range	From B...	Low Risk	Dependent	Table needed to first be finished	Testing data to accomplish previous sprint
26	Be able to filter table by state	From B...	Low Risk	Dependent	Table needed to first be finished	Testing data to accomplish previous sprint
27	See a colorized map of population movement	From B...	Medium ...	Time	Started on 11/9	May move to sprint 3
28	View a heat map of the most popular CBG's that people are moving to	From P...	High Risk	Technical Inexperience	Have not had any experience with doing this. Paul mentioned that he may have some code fo...	Potentially will push this to a later sprint until we figure out exactly how to present. (Not high priority)
29	Tell person that moved to city X came from city Y	From P...	Medium ...	Time Technical	Started 11/10	Move to sprint 3 now that we have realized that there may be inaccuracies with data that is being pulled

Figure 11 Sprint 3 User Stories

Sprint Review

This third sprint began on November 12th, which was a day later than we usually began our sprints, due to conflict in scheduling with team sponsors. Our team quickly adapted and prepared ourselves for upcoming agenda (see Figure 11). Instead of wasting a day of work, we worked on stories had not been completed in sprint 2 and pulled from our backlog which contributed to filters within a table that was shown to our sponsors.

The theme of this sprint was making the map previously created more dynamic in showing the data. Instead of showing dots with different opaqueness, and density on the map, the goal was to show both a heat map with darker colors representing different movement densities. We created arrows that showed movement across country, but also movement within specific cities. This map was also able to be viewed by age and wealth distribution to ascertain movement of different demographics.

Sprint Retrospective Meeting

The only blocker that we experience was that PowerBI may not be able to be downloaded onto our virtual desktop. We experienced this issue in the first two sprints, not having admin access to all databases and all software the was needed. However, we were able to user PowerBI on our own desktops.

What worked well

- With respect to development, it worked well for Luke and Xiaowei to work independently on two separate things, and there was more productivity
- We finished all our intended stories and were able to pull from the backlog without pushing anything to the next sprint

Improvements

- Wrote a more detailed functionality and requirements section for each user story

Story Points Completed (Velocity): 96

Hours Worked: 84

Percentage of Story Points Completed: 114%

6.4 Sprint 4

Rank	Risk	Type of Ri...	Risk Explained	Risk Mitigation
6	Medium ...	Technical Inexperience	Not sure if PowerBi will allow for zooming and keeping color density	May have to push to sprint 5, but will stay in communication with sponsor
From B...	Medium ...	Technical Inexperience	PowerBI may not have this functionality, and be able to do this.	Going to talk to sponsor and see if he wants us to continue PowerBI and/or plot the heatmap with ...
1	Low Risk	Technical	Same code, but just pulling from a more complete dataset to represent the statistics ...	Communicate with sponsor since he knows where to look to find the right data
2	Low Risk	Dependent	Waiting on getting a more accurate dataset	This should be relatively easy. Has already been done before, but is reversed. Make sure we are usin...
3	Low Risk	Technical	Went over adding slicers together with Paul so this should not be that big of a risk	Stay in communication with sponsor over PowerBI
4	Low Risk	Technical	Went over adding slicers together with Paul so this should not be that big of a risk	Stay in communication with sponsor over PowerBI
5	Low Risk	Technical	Went over adding slicers together with Paul so this should not be that big of a risk	Stay in communication with sponsor over PowerBI
7	Low Risk	Technical Inexperience		
From B...	Low Risk	Technical Inexperience	This feature is actually a built in feature in PowerBI	Create the correct date timeline
From B...	Medium ...	Technical	Sponsor had mentioned using different colors red meaning growth and blue meaning loss. ...	Going to show the heat map as we have been doing so far and worry about blue and red colors for sprin...
From B...	Low Risk			

Figure 12 Sprint 4 User Stories

Sprint Review

Sprint 4 began on November 18th, so the team was back to starting sprints on Wednesday, unlike the previous sprint where it was started a day late. As our project approached its final days, the number of points in this sprint were significantly fewer (see Figure 12). We communicated and displayed our work daily to continuously adjust the different maps we showed. The sponsors wanted the map to be multi-dimensional to analyze different demographics and movement including age, income, and different mile range.

Once we were able to display a small range of data on a heat map, the goal was to then pull additional data up to the current day. This allowed for the animations of the heat map to show movement over an extended period and showed before and after large spikes in COVID-19.

Sprint Retrospective Meeting

In our sprint planning, we predicted that there may be some technical issues with PowerBI. These issues involved displaying features that can be seen above in user stories that were not completed. To mitigate this risk, we discussed having continuous communication with our sponsors and made them aware if features were not able to be added. After trying to work these risks out, we were ready to push those stories to sprint 5, but our sponsors were comfortable leaving these two stories out of the final product. Because of this, we were able to pull stories from our backlog that included animation, and several types of maps both showing net growth or loss.

Also, the members of the scrum team had started working on separate things and a lot more effort had to be put into the heatmap and PowerBI rather than the dataset due to a lack of experience. Limitations

of PowerBI did not allow for the Development Team to truly collaborate, but that allowed members of the team not working on the dashboard to put effort towards drafting our report.

What worked well

- Clear communication with the sponsors was key to make them aware of features that were not able to be added
- Was able to pull larger amounts of data into PowerBI as our notebook became more accurate
- Demonstrations of our progress had been longer in this sprint, and we were able to work with sponsors step by step to bring this together

Improvements

- The stories could have been more equally divided amongst the Development Team now that two members were working on separate deliverables.

Story Points Completed (Velocity): 57

Hours Worked: 42

Percentage of Story Points Completed: 135%

6.5 Sprint 5

	As a user, I want to	Rank	Risk	Type of Ri...	Risk Explained	Risk Mitigation
41	Be able to cleanly run our notebook on my own machine	1	Low Risk	Technical		Will have to work with sponsor to ensure full compatability
42	View zip codes on the heat map to show specifics areas and their net growth	2	Low Risk	Technical		Sponsor knows how to narrow down to zip codes

Figure 13 Sprint 5 User Stories

Sprint Review

Sprint 5 started on November 30th and only had two stories (see Figure 13). Our backlog had been cleared out at the end of sprint 4 and this sprint was about the finishing touches on our work. The day our sprint began we met with our sponsors and decided what else needed to be done to have a clean transfer of deliverables at the end of the week. The stories were completed later that day and our sponsors wanted us to dedicate the remainder of the sprint to analyzing the data and maps and coming up with our own conclusions before we handed them the deliverables.

Sprint Retrospective Meeting

The beginning of this sprint started off slowly because we were waiting to create more stories with our sponsors during our daily scrum. We were pleased to find out that sponsors were impressed with our progress and what we had done so far.

What worked well

- Being able to finish this sprint early allowed us to focus on analyzing the data and working on the report and presentation for sponsors and advisors
- Early sponsor feedback allowed us to have confidence in the work we had accomplished

- Asked sponsors for recommendations on how they want the data displayed for their team helped mold demonstrations

Improvements

- We should have tried to steadily work on the design and analysis of the paper as our sprints continued

Story Points Completed (Velocity): 12

Hours Worked: 12

Percentage of Story Points Completed: 100%

6.6 Weekly Burndown

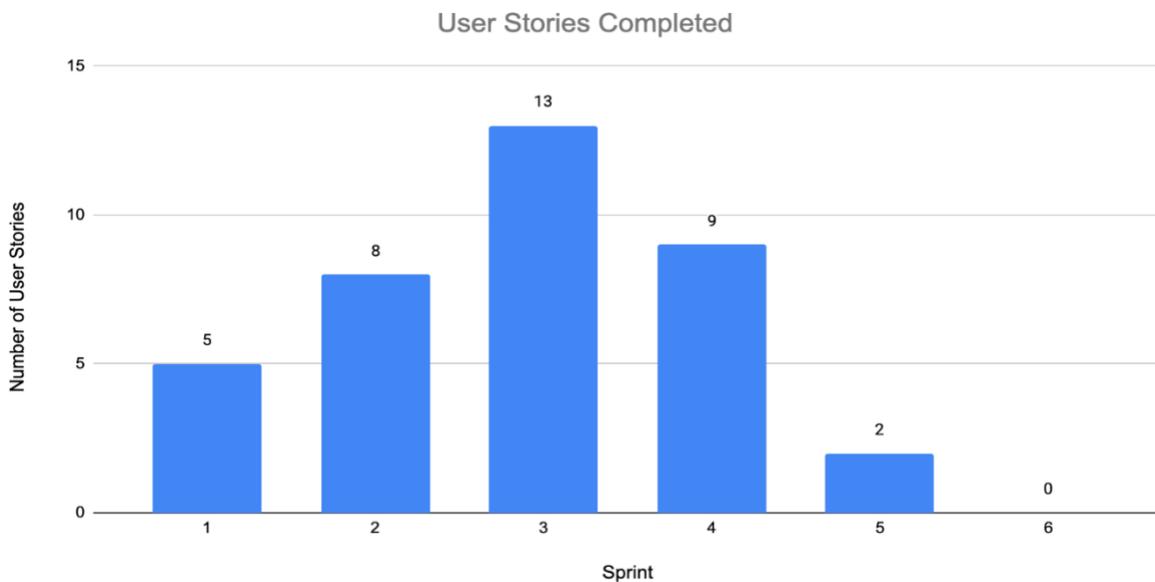


Figure 14 User Stories Completed

Figure 14 above shows the number of user stories that we completed. As you can see the first two weeks there was a buildup of the number of stories we could include in our demonstrations because we were still learning how to work with our sponsors. We had the most stories in sprint 3 after we felt comfortable with the databases we used and most of project was developed at that time. Afterward, in sprint 4 and 5 our focus was about perfecting our development and tuning it to our sponsor's requirements.

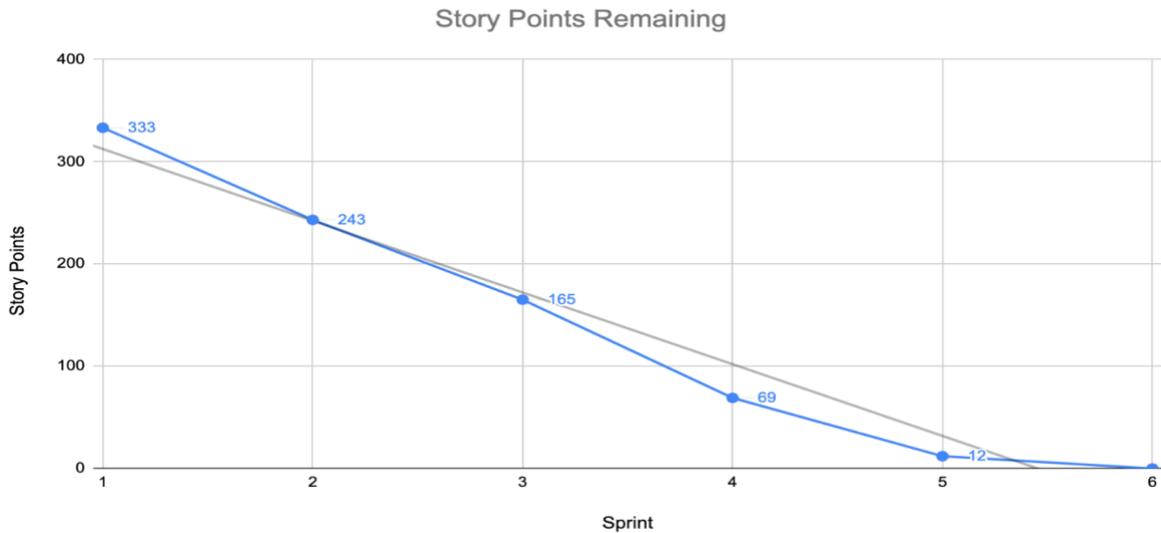


Figure 15 Story Points Remaining

Figure 15 shows that our group worked at a steady rate work done and the more our project progressed the less overall work we had to put into our development. As we neared our final sprints, there were fewer sprints to complete because the project was coming to an end. Therefore, we shifted our focus from software development to our report and presentation for our sponsors.

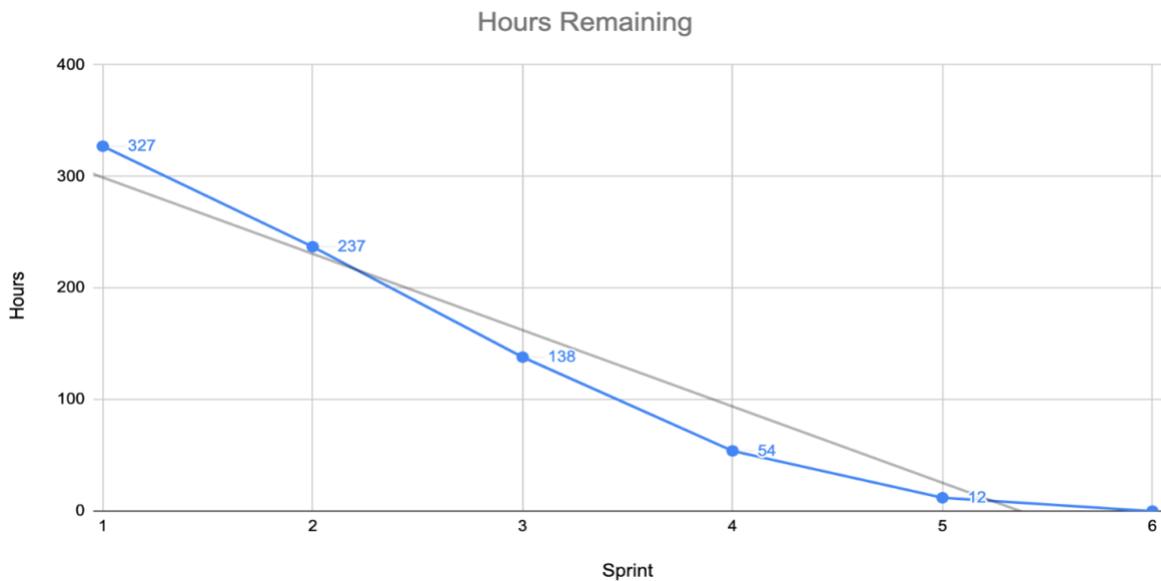


Figure 16 Hours Remaining

Figure 16 shows a similar story because the ratio between our story points and hours were 1 to 1. The more our software was developed to our sponsors wants we worked less because they wanted fewer major changes and more minor fixes.

7. Findings

7.1 Population Traffic Observations

Through the aggregation of over seven million rows of data from an internal Postgres database, our software was able to accurately provide quantitative metrics for modeling population traffic to and from cities within the United States. Our analytic goal was to calculate which cities had the highest number of people moving to them during this global pandemic. Table 2 below shows the cities with most people moving to them.

Location	Population Traffic	Median Household Income	Median Age
Panama Beach City, FL	288,741	\$52,337	40.3
Myrtle Beach, SC	188,221	\$54,847	41.6
Sevierville, TN	152,415	\$51,250	40.9
Las Vegas, NV	147,876	\$60,017	37.2
Destin, FL	141,491	\$56,005	39.8
Gulf Shores, AK	139,549	\$52,313	40.6
Kissimmee, FL	135,399	\$58,125	39.4
San Antonio, TX	112,342	\$54,295	36.9
Orange Beach, AL	107,517	\$54,053	40.8
Orlando, FL	99,649	\$58,438	39.5

Table 2 Top 10 Cities with Highest Population Traffic

From a quick glance, it can be observed that Florida overall has had the most movement to various cities. After consulting with the project sponsor, it was concluded that the most likely explanation of this mass migration to Florida during the pandemic was from “snowbirds”, a common term for people who move south during the colder seasons. Our team aggregated the dataset to calculate which cities were being moved out of the most since the start of the pandemic. Table 3 below shows the top ten cities that has the most people moving out of it.

Location	Population Traffic	Median Household Income	Median Age
Houston, TX	-134,766	\$53,661	32.2
San Antonio, TX	-92,294	\$64,388	33.2
Las Vegas, NV	-84,696	\$58,393	36.4
Miami, FL	-68,810	\$55,975	39.2
Chicago, IL	-64,963	\$50,526	34.3
Dallas, TX	-63,513	\$53,398	32.3
El Paso, TX	-62,645	\$50,694	29.0
Colorado Springs, CO	-57,755	\$71,645	32.6
Fort Worth, TX	-55,593	\$70,319	32.6
Atlanta, GA	-55,067	\$52,945	34.0

Table 3 Top 10 Cities Being Moved Out Of

The net population traffic is a calculated statistic where the number of people moving to a city and the number of people moving from that same city are summed giving the net population traffic. This reduces the effect of inaccuracies caused by vacationers on the overall dataset. When looking at the trends of the data, net population traffic was the indicator of growth or loss. The table below outlines the top 10 cities with the highest net population traffic.

Location	Net Population Traffic
Panama City Beach, FL	283,561
Myrtle Beach, SC	176,934
Sevierville, TN	143,612
Destin, FL	139,783
Gulf Shores, AL	137,338
Kissimmee, FL	119,730
Orange Beach, AL	105,730
Branson, MO	76,483
Las Vegas, NV	63,180
Santa Rosa Beach, FL	62,811

Table 4 Top 10 Cities with the Highest Net Population Traffic

7.2 Bias in Dataset

The data provided to generate the population traffic of cities inherently had some constraints, limitations, and biases. In the data provided, the home location of a person changes after spending six weeks in the different CBG location. This would skew the data because if someone went on vacation for six weeks, it would record two entries of movement. The first entry would be that they moved from their household to their vacation location and then the second entry would be that they moved from their vacation location back to their household. This particularly skewed the data in states such as Florida which are notorious for having seasonal vacationers or “snowbirds”. This vacationer effect is best highlighted in the population traffic of San Antonio, TX where the city is both one of the highest-ranking cities of population growth and population loss. This indicated that many people vacationed there for over six weeks and then left. This example reinforces why net population traffic was used as our leading metric for observing migration patterns in the United States.

Another source of potential error in our data was caused by the distance radius that someone needs to travel in order to be considered a mover. For the duration of the project, we decided that if someone’s home CBG was more 225 miles away from the CBG of the location being visited then that would indicate that someone had moved. In certain rural locations, locations that require far distance travel to get to one of the point of interest we selected could have led to inaccuracies and false indications of movement.

7.3 Mapping

Our team created different maps and animations by pulling data provided by our Jupyter Notebook.

7.3.1 Net Population Growth Heat Map

The red in the heat map represents the areas that have the most amount of people moving into them. The more opaque shades indicates that there are less people moving into them. No color at all indicates that those areas overall are experiencing a net loss of people.

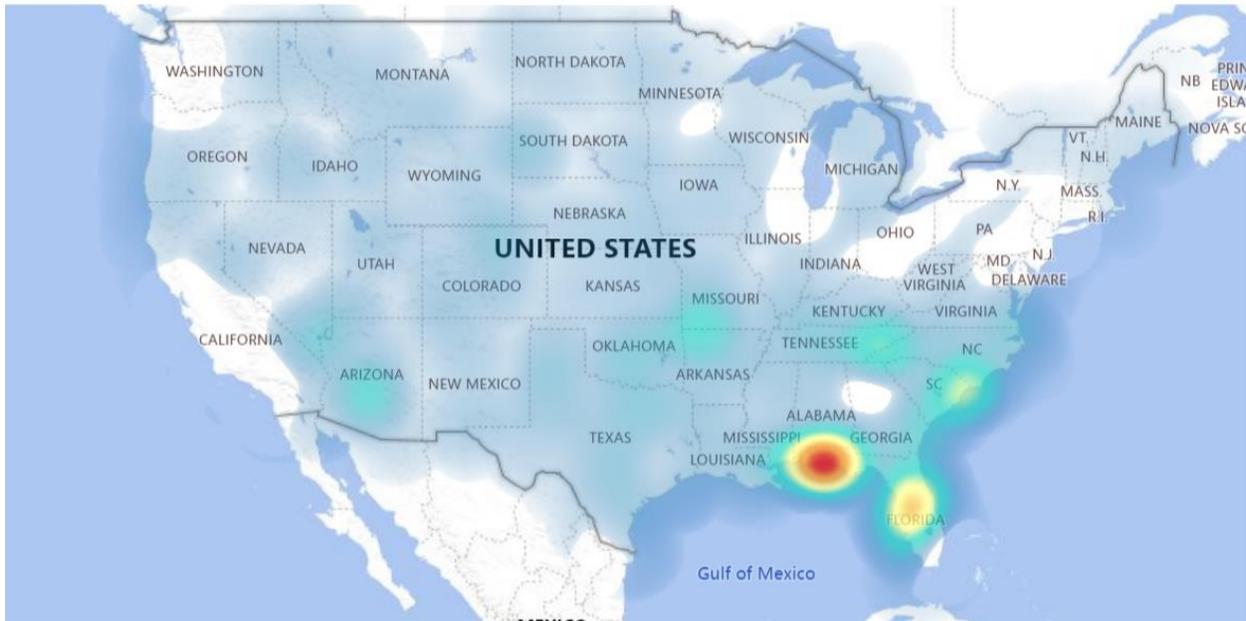


Figure 17 Net Population Growth Heat Map

According to the heat map, the 5 states with the most positive net population traffic growth are Florida, Texas, South Carolina, Alabama and Tennessee, which means that in these states there are many more people coming in than leaving in total from March 1st to November 23rd.

The obvious regions, which are in white, are the ones experiencing the most negative population foot traffic growth. These regions are California, New York, Ohio, Illinois and Washington which can be seen in the table shown below.

Region	Net Population Traffic
California	-108,150
New York	-60,454
Ohio	-24,569
Illinois	-20,448
Washington	-19,741

Table 5 Top 5 Regions with the Negative Net Population Traffic

However, even if a state has a huge positive net population growth, the cities can still have many people leaving it. For example, even though Florida has a total population growth of 1,511,937, the city Miami of Florida has a population traffic growth of -40,324. A lot of people may be moving into Florida, but we saw that major cities had negative population growth.

The heatmap is also able to play an animation based on date. By default, the map displays the net population traffic growth for all the months counted. Therefore, to see the visualization of the population growth on separate days, it is necessary to play the animation. The following picture shows the population growth on March 30th, 2020.

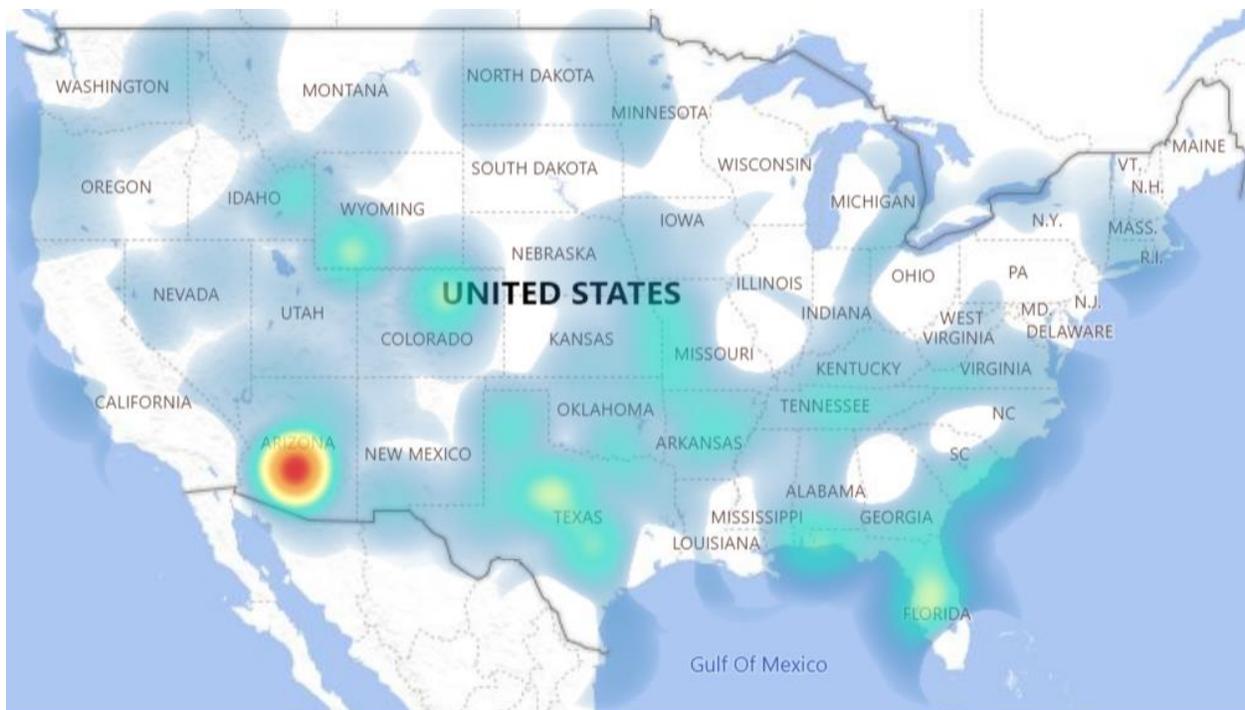


Figure 18 Net Population Growth Heat Map

An overall assumption from Figure 18 is that people are moving from both the west and east coasts and moving to the south. Overall, Florida had the biggest net population growth since COVID-19. However, this image shows the growth as of March 30th, which in that case Arizona was experiencing the most net population growth at the time.

7.3.2 Geographic Point Map

The Geographic Point Map we created similarly shows the net population growth (see Figure 19). We aggregated locations to allow each point to represent cities which allowed us to view growth in more specific areas. A blue point represents net growth, red represents net negative growth, and yellow represents a close to zero net growth. The colors to represent the population growth in geographic point map are reversed from the colors used for heatmap as the project sponsors asked us to do so.

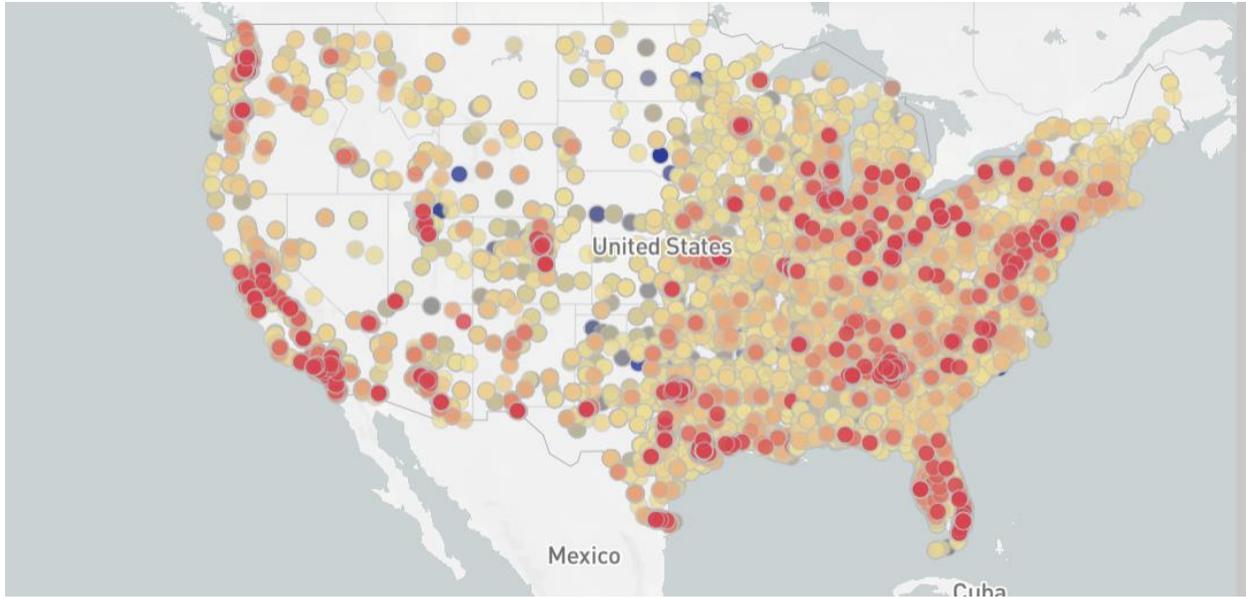


Figure 19 Geographic Point Map

The view supports the state filter to select states, so it only displays the data for the selected states. The following picture (see Figure 20) shows the net population traffic growth in New Jersey and New York since March 1st, 2020 until November 23rd, 2020.

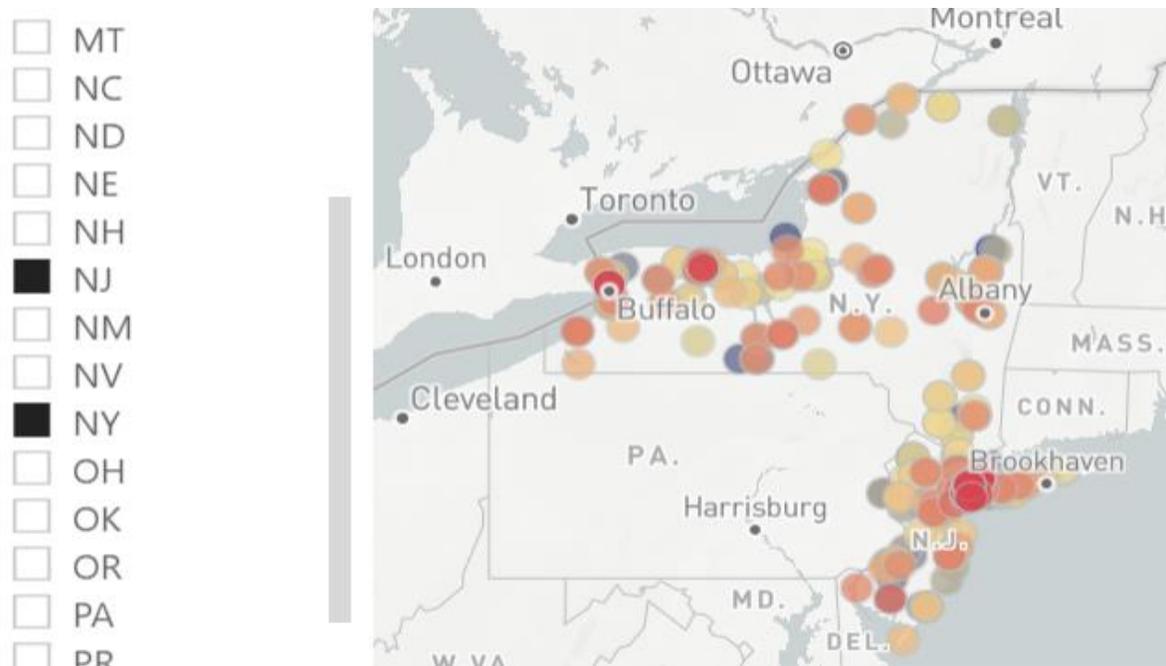


Figure 20 New York

The animation function works for the geographic point map as well. The following picture (see Figure 21) shows the population traffic growth in New Jersey and New York on May 04, 2020.

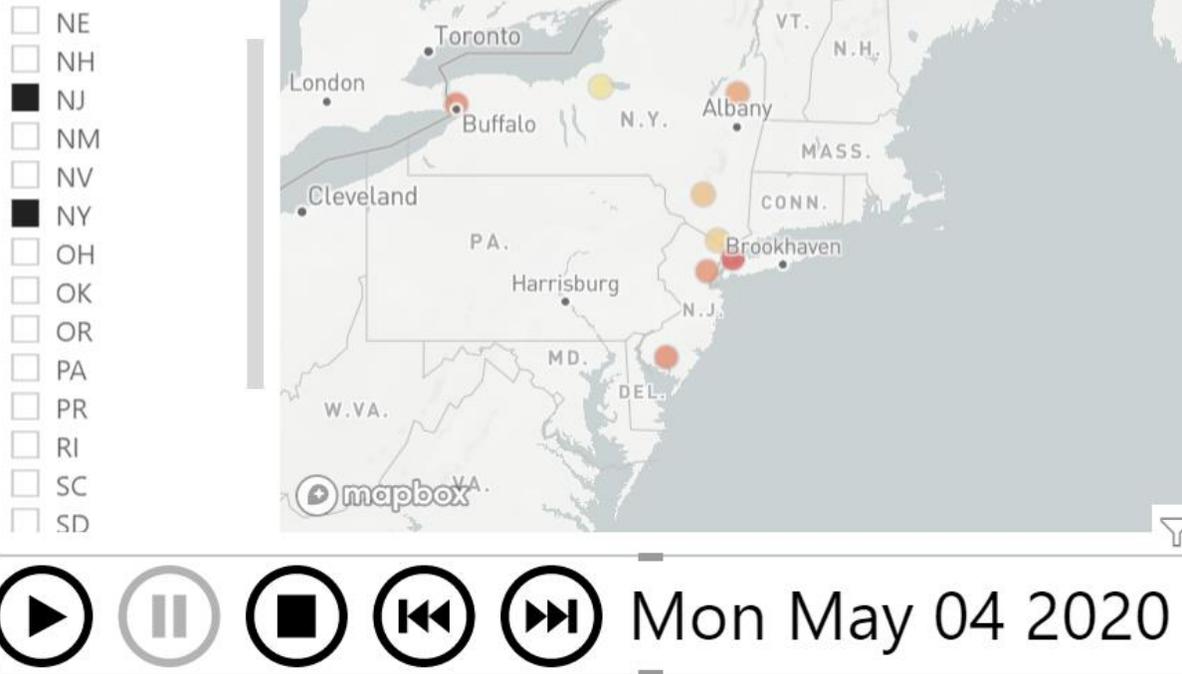


Figure 21 Animation Shown

This view supported more functionality than the heat map, including the tooltip. The tooltip included as many parameters as we want: net population traffic growth, people's age, people's income, the average distance that people traveled, etc. (see Figure 22).

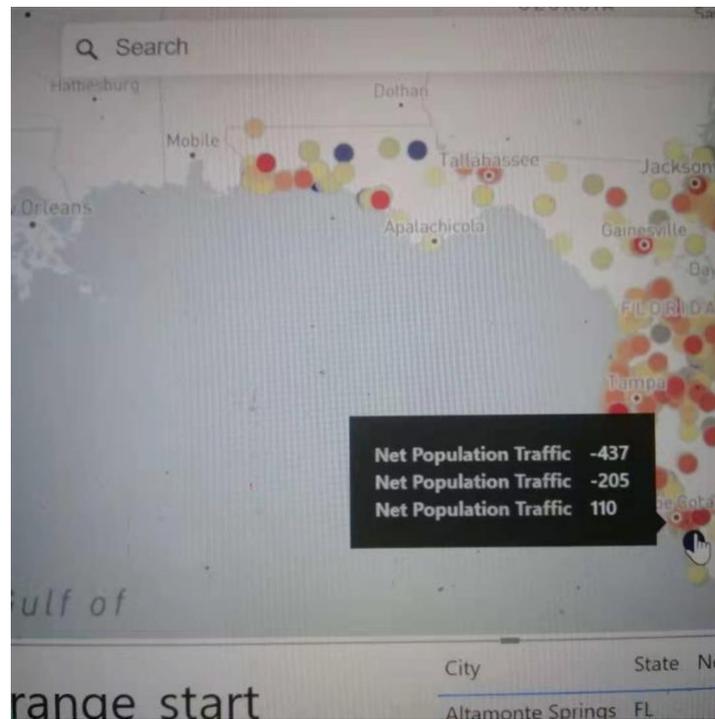


Figure 22 Tool Tip

7.3.3 Outliers

In the heatmaps, we found that CBG's near airports close to a city were experiencing positive net growth. We believe this to be an outlier because of vacationers. In our dataset, we included stores like Walmart and Target, which can be popular stores amongst these travelers. These stores include more of everyday goods rather than appliances for home projects. We had to be mindful of this when analyzing our maps.

8. Testing

Through conversations with our sponsors, we discovered early on that we would not be going through regular testing trials. The benefit of having a daily scrum with our sponsors allowed us to continuously demonstrate our progress. Having daily feedback enabled us to continue working and make edits to our notebook that visually mapped our deliverables. As more precise stories were created through these meetings, our deliverables started to take form.

We discovered that the investment company had previously done some of their own analysis. Because of this, they compared our visualizations with theirs to ensure accuracy and confirmed that our data matched up with theirs. Also, the other MQP team working with them was doing a similar project using credit card purchases. Therefore, between our sponsors work, our work, and the other team's, the investment firm can compare data to create a leading indicator for understanding neighborhoods and population movement.

9. Future Work

Our deliverables that helped for understanding migration movement concluded with a Jupyter notebook providing population movement based on foot traffic, median income and age. Our PowerBI heat maps showed hot spots where people were moving into and out of. There were some visualizations that PowerBI did not have the capability of providing. Different zoomed in views of regions on the heat map with different software may be helpful in the future to get exactly what our sponsors wanted. PowerBI's heatmap was helpful for seeing the United States overall, but zooming features were absent when wanting to look at specific areas.

Due to the limits of our virtual machines provided by the investment company; we only had enough memory to show 7 million data points. Once our notebook transitioned into the company's Databricks notebook, they were able to run the data over a larger horizon. For instance, we started running the data from March 1, 2020 to present day. In the future, our sponsors will be able to look at months and years prior to when the pandemic was in full swing in the United States. Their data frame should also have the capability to look at more home improvement and appliance retailers besides the few that we selected on which to perform data analysis.

In the future, our sponsors could use this data and our deliverables to predict how people will react to future epidemics. If a similar pandemic were to come along again, they will be able to compare people's migration patterns from 2020 to the next pandemic. Based on what happened in 2020, they can make inferences about human behavior and which regions would create large revenue returns.

10. Learning Assessment

10.1 Business Learnings

10.1.1 Leadership

What our team realized, was that although there were assigned “roles”, everyone was a leader in their own way. To get stories completed fast and on time, team members had to take leadership of stories they oversaw. The Project Manager of the team was by no means anyone's supervisor. However, the Project Manager was an organizer of thoughts and ideas that could be put together to complete the overall project. Deliverables were made by the team with the assistance of our sponsors and any story that was created by the Project Manager had the ability to be adjusted by the Development Team. Leadership, in the case of this project, does not command but listens, collaborates, and creates guidelines to satisfy the client’s needs.

10.1.2 Industry

Something to note regarding the industry and using the Agile framework, is that no time should be wasted. Our team was lucky enough to be offered the opportunity to work with this investment firm. From the beginning, we understood that our sponsors did not have to meet with us every day. They were gracious enough to meet with us, therefore we made sure that we were prepared for every meeting and made effective use of our time. Doing scrums allowed our team to learn how to process ideas quickly and demonstrate proficiency.

10.1.3 Culture

Our team was fortunate enough to work in a collaborative and educational environment. We were encouraged to experiment with data that we were using to get a better grasp on the information with which we were dealing. This allowed us to learn things for ourselves and made us feel comfortable to ask for help when needed. This culture motivated us to try different things and not be afraid of failing quickly and fixing our mistakes with feedback.

10.1.4 Working Remotely

Working from home is something we had not anticipated. From looking at previous reports of students who had worked with this investment company, we were concerned about the amount of hours they had spent on sprints. We came to the realization that they had the opportunity to go into the office every day and sit in an office or cubicle. This not only provided fewer life distractions but also granted them access to in-person relationships with one another and members of the investment firm. We quickly came to an understanding that working from home may require less time than going into the office. In an office, one assumes you are there for an eight-hour workday. However, at home we were able to break sprints into individual hours by assignment for the team. The part that we missed was making those personal relationships with people within the office. In many meetings, we were not able to see each other's faces due to technical reasons, which is a key element to getting to know one another. Because of this, the project was always business.

10.1.5 Time Management

This aspect of the project was one of the most important given the circumstances of the pandemic. Working from home allowed us to do other things like participate in activities and remain active. This was certainly a benefit, because it was important to have a healthy balance.

The user stories' estimated hours and potential risks really allowed for a sizeable schedule for each member of the team. The quick weekly sprints never slowed down but our understanding of our own capability increased as time went on and we were able to work with full effort and balance that with other activities. While this led to the sponsors being happy, it also led us as a team to never feeling overwhelmed.

10.1.6 Team Management

It was apparent that allowing team members to work on tasks that they found fascinating and fun created a healthy and interactive work environment for everyone. This created a positive space where we could have controlled conversations with one another expressing our opinions and ideas towards the work that had been accomplished. Overall, we were able to come to a point where we all understood each other and that led to synergy in our team.

10.2 Software Development Learnings

10.2.1 Technology

It was our developers' first-time using python and postgres as data science tools. The project allowed them to enhance their ability to process data using a Python data frame with a Jupyter notebook on a large scale. They both learned how to use the Python data frame to successfully organize data by using filtering, grouping, and merging data frames hundreds of times.

Because of this experience, Anaconda proved to be a great tool for launching a Jupyter notebook. It allowed us to use Jupyter's features and run commands from its internal shell prompt. All team members gained the ability to draw maps with the mapboxgl library in python as well.

10.2.2 Business Analytic Tools

PowerBI was the external software that the Development Team had to learn to create our deliverables. It is a powerful tool, as it allowed us to draw many different maps and animations that were needed for our project. We were able to gather the data, and then we were able to produce heatmaps, choropleth maps, circle maps, and animations to view the results of that data and show it in a useful way. Filtering the maps with tables and slicers was a way we were able to narrow down exact demographics and dates that we were looking for. This interactive tool was something that we had not used before and seeing the data come to life was exciting.

10.2.3 Query Efficiency

The Postgres database used for the scope of the project was not optimized nearly enough as it should have been. The reasoning behind this is because given the relatively sudden onset of the pandemic, there was not enough time to design and build a large database with optimized tables and columns. None of the tables that were used were indexed. This severely inflated the amount of time that a query took to execute. In addition, the tables used were not uniformly normalized which prompted inefficiencies. Given that our team was only allowed read-only access to the database, there was very little that we could do in terms of optimizing the query run time. This major roadblock slowed the development progress in the beginning of the project. From this, our team learned the importance of normalization and indexing when designing a database, especially a database with an extremely large number of entries.

10.3 What we would do differently

In the beginning of the project, it seemed hard to develop the software simultaneously. Many stories relied on a previous story being completed by the Development Team. This created non-value-added time for other members of the team. We realized we worked most efficiently when members of the Development Team worked on separate tasks that could be brought together in the end. If we could do things differently, we would have the Development Team work on two separate deliverables to complete the overall project.

11. Conclusion

The alternative investment firm we worked with wanted to improve their real estate portfolio with the changing market that COVID-19 effected. People's movement patterns have changed because of the virus, and for the firm to continue growing its investments they thought it was necessary to invest in a neighborhood recognition algorithm which we provided. We set out to track where populations were moving to by using Census Bureau data provided by our sponsors. The code we developed with Jupyter Notebook produce its own maps, but we pulled that code into PowerBI to have a more interactive experience with the data which the company was looking for.

Given this opportunity by the investment firm, we able to create a key indicator for neighborhood recognition and migration patterns. Our Jupyter notebook code generated data that provided net growth of different census block groups that will help in predicting future movement patterns. Our PowerBI application layer will allow for investors to visualize real time movement of people within the United States.

We were able to display interactive maps including many animations. Our maps were able to show a range of several different demographics and time span. The maps produced could be filtered to show specific median income, average age, and distance traveled. This allowed our sponsors to view the maps differently based on what investments they want to make.

From our analysis, we were able to conclude that suburbs outside of major cities are seeing the most foot traffic, therefore, more people are moving outside dense cities. From our heatmap, people are also moving away from the coast and more inland, with exception of Florida which sees a lot of traffic with vacationers, and from fewer COVID-19 restrictions. These were general findings, but with the filters we created, our sponsors can look to target the key demographics they are looking for.

This project occurred during unprecedented times, and we were forced to learn and adapt in an environment none of us had been in before. Due to this, the Agile Scrum framework proved necessary to develop and deliver such complex deliverables. We were challenged with using software we had never used before. However, we constructed meaningful code to develop relevant maps and animations to contribute to the investment firm's real estate portfolio under new team dynamic circumstances.

Through a lot of communication, hard work, daily Scrum meetings, sprint planning meetings, sprint reviews with our sponsors, and sprint retrospectives we were able to achieve our project goal. With the completion of this project, our sponsors will be using our software programs to perform their analysis and make better real estate investments in the future.

Work Cited

- Andrews, J. (2020, October 14). Everything You Need to Know About NYC's Housing Market. Retrieved December 08, 2020, from <https://www.curbed.com/article/nyc-real-estate-housing-rent-buy-manhattan-brooklyn.html>
- Bureau, U. (2019, October 25). U.S. Census Bureau at a Glance. Retrieved November 13, 2020, from <https://www.census.gov/about/what/census-at-a-glance.html>
- Bureau, U. (n.d.). What Is the 2020 Census? Retrieved December 08, 2020, from <https://2020census.gov/en/what-is-2020-census.html?cid=20394%3A+census++legal%3Asem.ga%3Ap%3Adm%3Aen%3A>
- CAIA, A., Black, K. H., Chambers, D. R., & Kazemi, H. B. (2016). Alternative investments : Caia level ii. ProQuest Ebook Central <https://ebookcentral-proquest-com.ezpxy-web-p-u01.wpi.edu>
- Chen, J. (2020, August 29). Alternative Investment Definition. Retrieved November 13, 2020, from https://www.investopedia.com/terms/a/alternative_investment.asp
- Finley, K. (2012, July 14). What Exactly Is GitHub Anyway? Retrieved November 13, 2020, from <https://techcrunch.com/2012/07/14/what-exactly-is-github-anyway/>
- Hillson, D. (2013). The A-B-C of risk culture: how to be risk-mature. Paper presented at PMI® Global Congress 2013—North America, New Orleans, LA. Newtown Square, PA: Project Management Institute.
- Kennon, J. (n.d.). Alternative Investments for Beginners. Retrieved November 13, 2020, from <https://www.thebalance.com/alternative-investments-for-beginners-4144347>
- Pongpaichet, S., Singh, V. K., Jain, R., & Pentland, A. (. (2013). Situation fencing. *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia - PDM '13*. doi:10.1145/2509352.2509401
- Python Development Services. (n.d.). Retrieved December 08, 2020, from <https://www.mindfiresolutions.com/python-development.htm>
- Python For Beginners. (n.d.). Retrieved December 08, 2020, from <https://www.python.org/about/gettingstarted/>
- Richardson, B. (2020, August 04). Covid Effect On Rental Markets: Pricy Cities Become Cheaper, Cheaper Cities Become Costly. Retrieved December 08, 2020, from <https://www.forbes.com/sites/brendarichardson/2020/08/04/covid-effect-on-rental-markets-pricy-cities-become-cheaper-cheaper-cities-become-costly/?sh=312a52bd6314>
- Rios, R., & Ríos-Solís, Y. (2012). Just-in-Time Systems (1st ed. 2012.). Springer New York. <https://doi.org/10.1007/978-1-4614-1123-9>

Ruhe, G., & Wohlin, C. (2014). *Software Project Management in a Changing World* (1st ed. 2014.). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-55035-5>

Stock, K., & Guesgen, H. (2016). Geospatial Reasoning With Open Data. *Automating Open Source Intelligence*, 171-204. doi:10.1016/b978-0-12-802916-9.00010-5

Vanderjack, B. (2015). *The Agile edge: managing projects effectively using Agile Scrum*. In *The Agile edge* (First edition.). Business Expert Press.

What Is an IDE? (n.d.). Retrieved December 08, 2020, from <https://www.codecademy.com/articles/what-is-an-ide>

What is Power BI: Microsoft Power BI. (n.d.). Retrieved December 08, 2020, from <https://powerbi.microsoft.com/en-us/what-is-power-bi/>