

# Zillow Analysis

Vishal Aiely

Sehee Hwang

Matt Mohandiss

Marc Muszik

Selena Xue

## I. INTRODUCTION

Home ownership is often thought of as a hallmark of prosperity and a staple of the American Dream. Almost everyone takes part in the real estate market, whether as a buyer, seller, renter, or landlord. Home ownership is an investment expected to appreciate over time and recently the rate of return is rising faster than ever. That leaves many to navigate the most expensive real estate market since before the Great Recession. While we cannot solve the affordable housing crisis on our own, we can mitigate its impact and navigate the market by using the power of data.

The objective of this project is to use digital archaeological techniques to retrieve data from Zillow and analyze trends in the housing market. The team analyzed the correlation between a property's attributes and its price. For example, we consider the price per square foot in relation to the number of bedrooms and bathrooms. We also investigated the changes in neighborhood prices if historical data is available.

## II. MOTIVATION

The real estate market can be difficult to follow as it is always in flux – housing prices and real estate trends fluctuate wildly. Understanding the market is important as real estate is a substantial investment for the average American. By recognizing correlations in the real estate market, we believe people can more easily navigate the market and make more well-informed decisions. Our homes are the most expensive item we'll likely ever own, so it is crucial that buyers understand the true value of a property.

The data we collect will help us answer important questions that pertain to the decision to buy. Which neighborhoods have the best value and which are the most overpriced? What renovation will add the most value to my property? Is now a good time to buy? These are difficult questions to answer, and many attempt to do so without a quantifiable, data-driven approach. Our goal is to bring that data-driven approach to these questions in order to provide answers that give us real-world insight.

## III. DISCUSSION OF DATA

The most detailed, up-to-date and readily available information on the housing market comes from Zillow.com. Zillow has comprehensive data about many of the most important attributes that go into a property, and makes that data available to us via their website. We collected any and all attributes that affect a property's value, like the location, number of bedrooms, bathrooms, square footage, lot size, and year remodeled, among others. We also collected

the 'Zestimate', or more interestingly, how the list price compares. The data was stored in a cloud database after being read and parsed from a web scraper.

Another data stream we collected is historical price changes in the U.S. This data was made available by the U.S. Census Bureau in a CSV file. The data included the prices of houses sold per U.S. region from the 1960s to present. We used this data to analyze historical trends and see if our Zillow follows that trend.

## IV. EXPECTED OUTCOME

After gathering and analyzing the data, we expected to see several trends. The primary outcomes that we expected to discover were two-fold: first, what makes certain properties worth different than others and second, what causes the overall increase in housing prices across different regions. Beyond that, we hoped to understand more regarding the cause of increased price and why some regions may see more of a rise than others. Some of the physical features of a home that we thought would play a role would be more bedrooms, more bathrooms, or the presence of a pool. Other external influences we expected were the quality of nearby schools, crime rates of the area, and trends of houses in the same neighborhood. Some trends we expected to see from a buyer's standpoint was an increase in non-primary home purchases, as well as an increase in the difference between the closing price and the estimated price prior to bidding. These outcomes may help build an understanding of the market with respect to current economic events such as COVID-19, shortage of housing, the relocation of individuals from higher priced regions to lower priced regions. In our project, we hoped to provide people a better understanding of the housing market and how it evolves over time.

## V. PROJECT OBJECTIVES

The objectives of this project were to build the Zillow Scraper, to implement data storage and collection, and to create an analysis based on the data. Our goal was to develop a web scraping algorithm that would allow us to collect useful listing information from Zillow without getting blocked. For data collection, we collected several attributes that can affect a property's value, such as the location, number of bedrooms, bathrooms, square footage, lot size, and year remodeled, among others. For data analysis, we analyzed trends such as listing prices in cities, number of bedrooms in listings from various cities, median lot size by state, land prices in various states, trends in home type, median house prices, listings among seasons, and other regression analyses.

## VI. ZILLOW SCRAPER

The Zillow Scraper is a tool that we built in order to gather all of the necessary housing data from Zillow and to store it. We first needed to know either the zip code or city and state in order to get data on listings in the designated area. We settled on using zip codes rather than cities to parse the data. Details on this choice are explored in the data collection section. To ensure that we do not add duplicate listings into our database, the team created a collection within Firestore that tracks all of the zip code and whether they have been scraped. Whenever we have completed scraping a specific zip code, the database is updated so that the zip code is marked as scraped. The code below describes how we ensure we do not include previously scraped zip codes in our current session.

---

```
codes = [z.to_dict() for z in zipcodes
         .start_at(db_ref.get())
         .limit(amount+1).get()
         if not z.to_dict()['Scraped']]
```

---

Once we have a specific zip code to search, we construct the URL that we use to make a request for the necessary listings data. Zillow.com uses a specific URL format when looking for a region. The region, which includes the city or zip code followed by "," and the state, can be appended as a route to the URL. This URL will fetch the data for the first page of listings in that area, and we can append "/page\_p", where page is the page number to fetch more listings in the area. The scraper keeps requesting pages until the status code returned is no longer successful. We also used a wrapper around the requests library, named requests, to raise an error when the requests was unsuccessful or blocked by CAPTCHA.

---

```
url = "https://www.zillow.com/"
region = f"{city_or_zip},{state}"
search_url = f"{url}{region}"
requests(f"{search_url}{page}_p/", headers)
```

---

We initially attempted to get data by visiting the individual pages for each of the listings. While this option would have resulted in more detailed data on each of the listings, the time it would take to get the data would be greatly increased. Instead we pulled data from the card listing shown for each of the listings on each of the search results page. The data here still contained substantial information such as price, lot size, bedrooms, and days on the market. We gathered this data by looking through the HTML sent back to the client after each request. The HTML contained this data inside of a script tag where the data was represented in JSON. This also allows us to get multiple listings while only making one 'GET' request rather than one for each listing, which greatly reduces our chances of getting blocked by Zillow.

The team originally attempted to run the program locally on personal computers to gather data. However, this resulted in all team members getting a response from Zillow.com

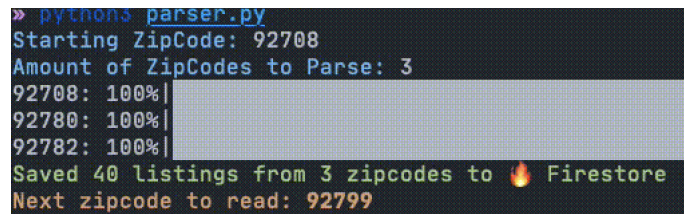
A terminal window with a dark background and light blue text. The text shows the execution of a Python script named 'parser.py'. It starts with 'Starting ZipCode: 92708', followed by 'Amount of ZipCodes to Parse: 3'. Then it shows progress for three zip codes: '92708: 100%', '92780: 100%', and '92782: 100%'. The next line says 'Saved 40 listings from 3 zipcodes to Firestore' with a small orange flame icon. The final line is 'Next zipcode to read: 92799'.

Fig. 1. Parser Running on 4 Zip codes

requiring a CAPTCHA on each request. To avoid this, we attempted to run our code on a Google Colab notebook. This method prevented our program from getting blocked by Zillow.com, allowing us to make all the requests needed to fill the database.

## VII. DATA COLLECTION

We collected data by running the Zillow Scraper on all zip codes in the US and US territories. The reason for using zip codes is that they reduce ambiguity in using city names (e.g., multiple cities with the same name, or one city with multiple names/spellings) and they cover a smaller area than cities, reducing the risk of too many results. Additionally, many cities have more results than could fit in the maximum number of search results, which would lead us to miss data.

Zip codes were found using publicly available data. We began with a total of 42.5k zip codes throughout the US and its territories. We then omitted zip codes that belong to PO boxes, military land, and organizations. This left us with approximately 30k zip codes that were used in the scraper to collect data.

Data storage was initially done using Google Firestore, a No-SQL non-relational database backed by the Google Cloud Platform. This type of data storage is the simplest to implement because our data is in JSON format, making it directly compatible with the No-SQL structure of Firestore. The downside of Firestore is the limited number of read and write queries per day. Our testing brought us very close to the read and write limit of the free tier, which led us to consider BigQuery as an alternative.

BigQuery is a standard relational database, also using the Google Cloud Platform. BigQuery, however, is not limited by the number of reads and writes to the database, but by the size of the data transferred. This proved to be much less restrictive for our purposes. The tabular format of BigQuery required the extra step of processing the collected JSON format into a single-layer dictionary with the same schema as the database. This extra processing increased the complexity of using BigQuery but paid off when it was time to retrieve the data because it was already in a tabular format. Converting JSON data into tabular data was always necessary for analysis but now is done before storing it in the database.

We used BigQuery to store the collected data and also keep track of the Zillow Scraper's progress. We maintained a table of all zip codes, and updated a column of True/False values whenever the scraper had finished a specific zip code. This method allows the scraper to be concurrent, resistant to

unexpected shutdowns, and limits the amount of duplicate data collected.

## VIII. DATA SUMMARY

As of December 9, 2021, we collected approximately 240,000 Zillow listings from about 20,000 zip codes. The listings we collected corresponded to various property types including undeveloped lots, single-family homes, condos, town-homes, multi-family homes, manufactured homes, and apartments (listed in order of most to least number of listings). Since the majority of the listings ( 150,000 listings) we collected were either undeveloped lots or family homes, we focused on these property types in our data analysis.

## IX. RESULTS

After data analysis, we reached three conclusions. First, we found that Knoxville housing prices are mostly concentrated around \$200k-\$300k. While this range is higher than the median house price for Tennessee, it is relatively low compared to many other states. Second, we discovered that less populated suburban areas have listings with more bedrooms. This information may be useful to families that are desire a home with more bedrooms. Third, we concluded that housing prices and household income are correlated. Our data from Zillow shows that housing prices are lower in the southern region of the United States. Similarly, the U.S. Census data shows that household income is lower in the South and higher in Northeast or West. This led us to believe that housing prices and income could be correlated.

The following charts and figures explore the data we collected and help visualize these findings.

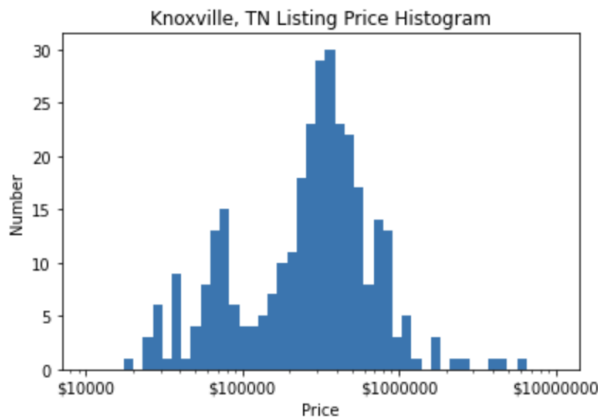


Fig. 2. This graph shows the current listing prices for Knoxville, TN. The average price of a listing in Knoxville is in the range of \$200,000 - \$300,000.

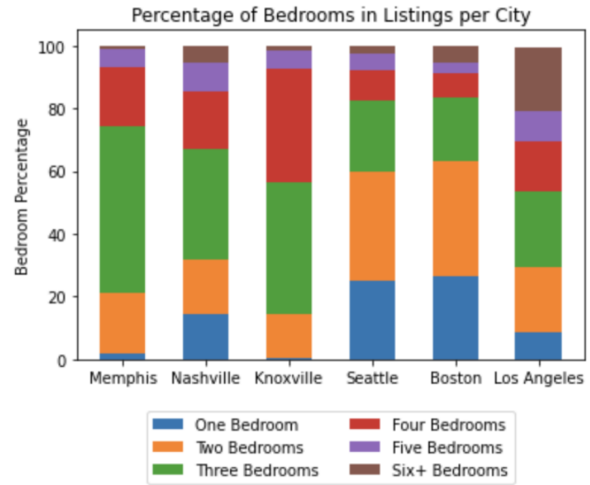


Fig. 3. This graph shows the percentage of the number of bedrooms per listings in different cities across the United States.

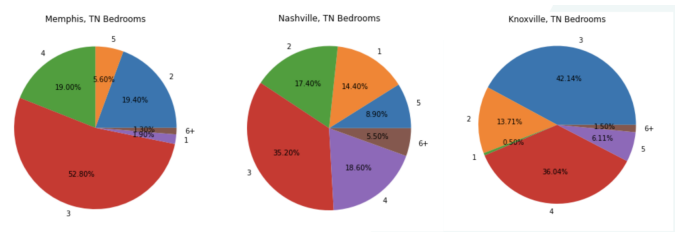


Fig. 4. This graph shows the percentage of number of bedrooms in the major cities of Tennessee.

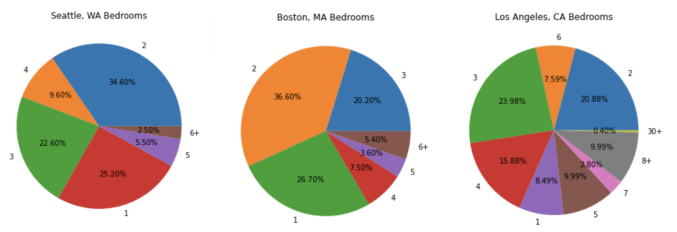


Fig. 5. This graph shows the percentage of number of bedrooms in Seattle, Boston, and Los Angeles.

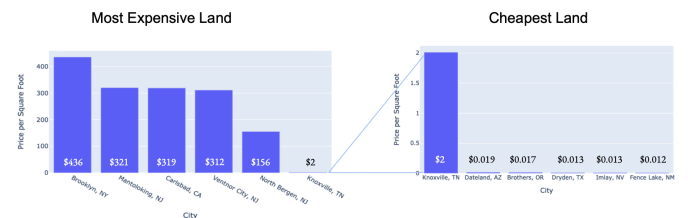


Fig. 6. The above graph shows the huge discrepancy in land costs throughout the country. The most expensive land is in Brooklyn, NY at a median cost of \$436 per square foot, while the cheapest land is \$.012 per square foot in Fence Lake, NM. Knoxville has a median cost of \$2 per square foot for comparison.

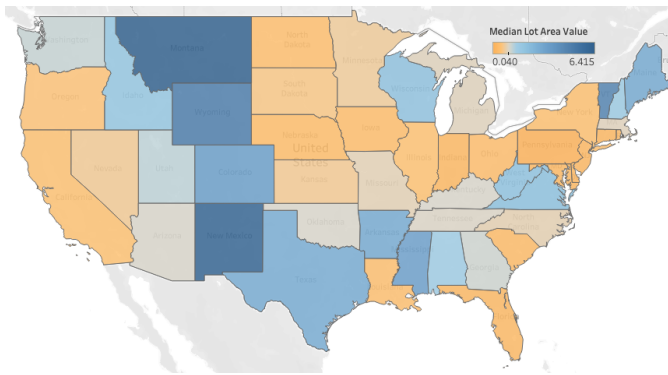


Fig. 7. Keeping with the theme of land availability, this plot shows the median lot size available by state. The most land is available in the region around Montana, Wyoming, Colorado and New Mexico. Some states in the Northeast and Midwest tend to have smaller median lot sizes.

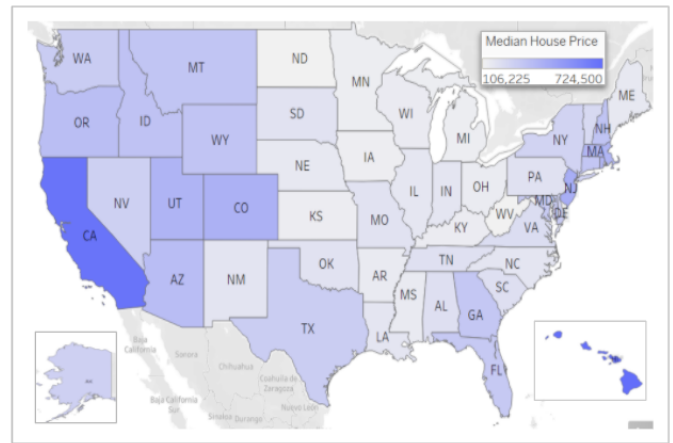


Fig. 8. This map depicts the variance in median house prices (based from Zillow listings) across the United States. Generally, it appears that the northeast and western regions of the U.S. have higher house prices.

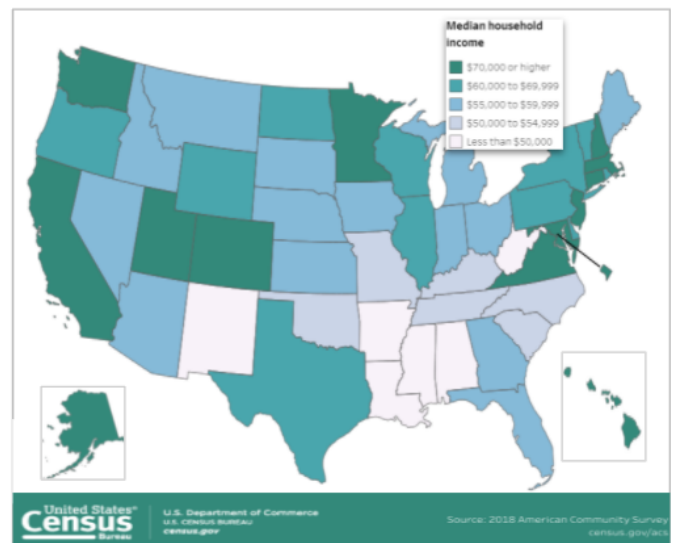


Fig. 9. This map depicts the variance in median household income (based from the 2018 U.S. Census) across the United States. It seems that the northeast and western regions tend to have higher incomes relative to the rest of the U.S. .

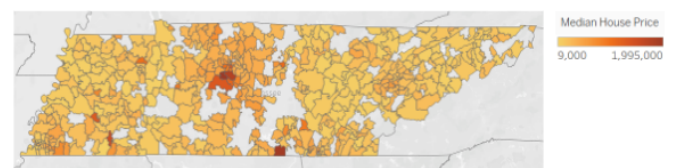


Fig. 10. This map shows the variance in median house prices (based from Zillow listings) across Tennessee. The areas with higher prices are located in Tennessee's major cities such as Nashville, Memphis, and Chattanooga.

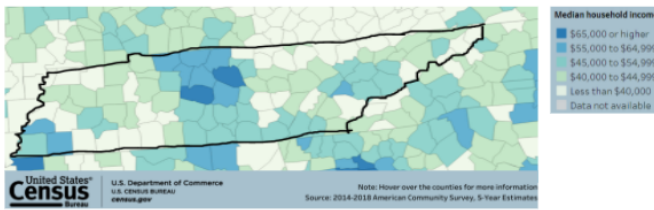


Fig. 11. This map shows the variance in median household income (based from the 2018 U.S. Census) across Tennessee by county. The counties with higher household incomes are located in Tennessee's major cities such as Nashville, Memphis, and Knoxville.

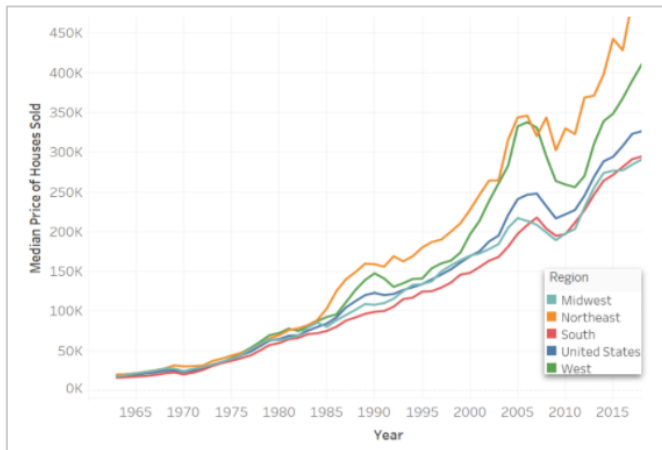


Fig. 12. This graph shows the change in the price of houses sold by U.S. region from the 1960s to present. Clearly, house prices have increased drastically over the past decades. However, it is important to note that the northeast and western regions of the U.S. have consistently had higher housing prices relative to other U.S. regions.

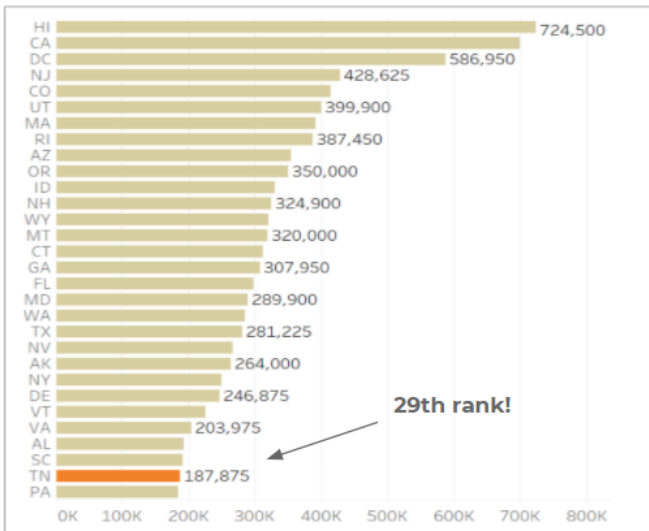


Fig. 13. This graph shows a ranking of the Top 30 states in terms of median house prices (based on Zillow listings). Most of the top 15 states on this list are located in the northeast or western region of the U.S. Tennessee ranks as the 29th highest state on this list with a median house price of roughly \$188,000.

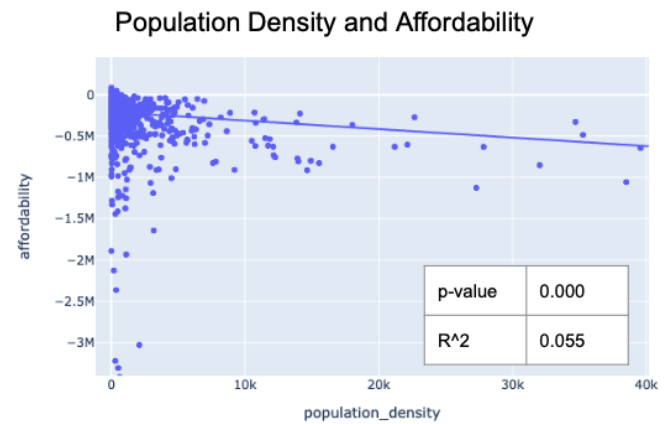


Fig. 14. We built regression models to explore some of the relationships with our collected data and demographic data from the Census. A very common assumption with real estate prices is that it is more unaffordable to live where the population is more dense. We test that assumption here, calculating the affordability of a city as the difference between the median home cost and the median income of that city. The plot shows that there is a significant correlation; the higher the population density, the more unaffordable that city becomes. The small r-squared value shows that this accounts for very little of the variation; there are other variables that account for the variation as well.

## House Price and Median income

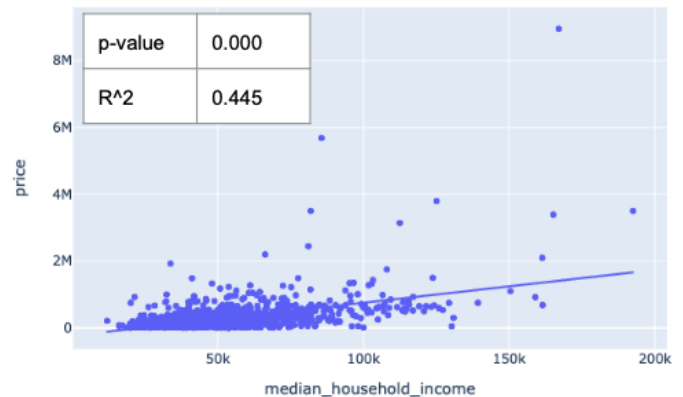


Fig. 15. Another common assumption is that cities with expensive real estate also have affluent residents. This correlation shows that relationship, and the regression model shows that there is a significant correlation. The r-squared shows that the median income accounts for a decent amount of variation in median home price, about 45 percent. Note these plots do not make an assumption of causality; we cannot conclude that higher-paid families raise housing costs, or that residents can expect a pay raise in a rising real estate market.

## X. CHALLENGES AND FUTURE WORK

The largest challenge the team faced in this project was creating a working Zillow scraper. Since all team members were unfamiliar with web scraping, it took us several weeks to develop a Zillow scraper that would collect necessary information without getting blocked by CAPTCHA. Moreover, since the scraper was not fully completed until early November, we were only able to scrape Zillow listings over a two week span. While we still collected thousands of listings, the team hoped to scrape even more data in order to create

a fully representative view of the housing situation in the U.S. In the future, further research could consider a possible correlation between housing prices and the time of year.

## XI. ORG CHART

Timeline:

Week	Milestones
Week 1-2	<ul style="list-style-type: none"> <li>• Complete scraper</li> <li>• Successfully scrape the Zillow web page</li> </ul>
Week 3	<ul style="list-style-type: none"> <li>• Put web scraper into a working function</li> <li>• Working function accepts a zip code</li> </ul>
Week 4	<ul style="list-style-type: none"> <li>• Integrate Firebase and BigQuery for data storage</li> <li>• Stretch goal: comparing for-sale vs rental properties</li> </ul>
Week 5	<ul style="list-style-type: none"> <li>• Cleaning Data</li> <li>• Stretch goals: New attributes/features: enrich data with <ul style="list-style-type: none"> <li>• Population growth, school district, crime rates, walkability, etc.</li> </ul> </li> </ul>
Week 6	<ul style="list-style-type: none"> <li>• Analyzing Data, Visualizing Data, Creating models, Finding Correlations</li> <li>• Writing the final project report</li> </ul>

Responsibilities:

### A. Vishal Aiely

- Help develop web scraping algorithm
- Integrate Google Cloud data storage
- Assist in visualization of results from data

### B. Sehee Hwang

- Help develop web scraping algorithm
- Find different data sets to analyze historical trends
- Assist in visualization of results from data

### C. Matt Mohandiss

- Work with API sources to gather data
- Integrate Google Cloud data storage
- Analyze results, visualize data

### D. Marc Muszik

- Build code to parse html response
- Experimented with ways to rotate ip addresses
- Implemented function as GCP Cloud Function
- Set up BigQuery database to house data
- Analyzed results and created visuals

### E. Selena Xue

- Help develop web scraping algorithm
- Find different data sets to analyze historical trends
- Analyze results, model visualization