Finding the Ideal San Francisco Apartment – Week 5

Brian Moore

1. Background and Introduction

Few changes in life are simultaneously as frightening, uncertain and exciting as moving to a new city for work or school, yet, historically, upwards of 40% of Americans moved to a new city over the course of their lifetime. While moving is often new and exciting, the process of selecting a new place to live is extremely challenging and poses serious financial risks, even in the best of circumstances.

Primarily, this is because you're not necessarily familiar with your surroundings before you arrive in a new city, yet, you're expected to make critical, long term decisions regarding housing and transportation with limited time and information. As a result, you might unknowingly rent an apartment that's priced fairly and is located close to work, but, lacks quality community spaces like parks, restaurants, and suffers from high rates of petty crime. Issues such as these are compounded when renters move to complex and competitive markets like the San Francisco apartment market. Renters looking to enter this market face a high stakes uphill battle with two primary challenges:

> Problem #1: Identifying 'Quality' neighborhoods well suited to their individual needs.
> Problem #2**:** Identifying fairly priced, overpriced and underpriced apartments in desirable neighborhoods.

Given time and resources, renters can overcome the challenges above, however, data science and machine learning tools can provide renters with the answers they need to make informed decisions about where they choose to inside a much narrower time horizon.

2. Purpose and Audience

The purpose of this project is to provide new renters and those unfamiliar with the San Francisco real estate market with advanced analytics and resources, allowing them to:

1. Determine which neighborhoods are best suited to their individual needs.
2. Determine if a potential apartment or house is over or underpriced.

3. Data acquisition and cleaning

To provide answers to both problems listed above, I will be utilizing data from the following sources:

Problem #1 – "Neighborhood Quality" Data: Determinizing neighborhood quality is far more nuanced, however, for the sake of simplicity, I've decided to utilize data from two sources to help me objectively identify desirable and undesirable neighborhood characteristics. First, I'll be utilizing a comprehensive crime dataset provided by the City of San Francisco outlining "police incident activity" segmented by location, type of crime, date, and time. Given the importance of living in a safe neighborhood, I will use "police incident activity" (and lack thereof) as a crude proxy to determine neighborhood desirability. I'm also going to be relying on the assumption that Neighborhoods with less police incidents are safer than those with more police activity. Next, I'll be utilizing location data from the Foursquare API to build a comprehensive dataset of venues around potential candidate apartments. The venue data from Foursquare will be used to identify neighborhoods with high concentrations of 3rd places" or public spaces where I can meet friends and interact with the community. The Foursquare dataset contains over 4,000 venues and outlines:

1. Venue Name
2. Venue Genre
3. Venue Location (lat, lon)
4. Venue Neighborhood

Problem #2 – Apartment Listing Data Sources: Before determining whether apartments are under/overpriced, we must begin with a rich dataset of available apartments located in the city of San Francisco. After speaking with locals, I determined that Craigslist is the go-to resource due to its high rate of quality listings. As a result, all apartment listing data originates from scraped craigslist listings. Specifically, I built a scraper designed to gather and store over 2,000 current listings. Once scraped, the raw apartment data will then be cleaned, analyzed and loaded into a Random Forest ML classification model to determine under or overpriced apartments in the region.

Other data sources: San Francisco Neighborhood GeoJSON data is also utilized throughout this analysis to accurately determine neighborhood names and boundaries.

4. Methodology: Understanding Neighborhood quality

Once the data was gathered, I put myself in the shoes of someone moving to San Francisco. With over 2000 apartment listings to choose from, how would I begin narrowing down my options? Assuming I've definitively decided on moving to San Francisco but had no knowledge of the local neighborhoods in San Francisco, how would I go about choosing a neighborhood that fits my needs? As stated previously, determining "neighborhood quality" is extremely nuanced and subjective. Often times, the subjective characteristics that make a neighborhood ideal for one individual might be completely unsuitable for another. As a result, I'll be focusing on data that uses objective measures, such as: 1)The presence or lack of crime. 2) The presence or lack of community venues or "3rd Places".

Focusing first on locating an area free of major crime, I'll begin by analyzing a comprehensive crime dataset published by the city of San Francisco. Once I've narrowed down a set of safe candidate neighborhoods, I'll then cluster neighborhoods by availability of locations / venues that improve my quality of life, such as: access to Parks, coffee shops and gyms. In other words, "3rd Places" where I can meet and interact with members of the community outside of my home or work.

Scraping Crime Data: Sourcing raw crime data in San Francisco is luckily quite easy. The city of San Francisco maintains an extensive database outlining all police activity within a five year period. Given the size of the dataset, I choose to pull details on all crimes taking place within the city limits beginning on Jan 1, 2020. Each record in the database is an isolated "Police Indecent" and represents an event where police response to a criminal activity. Each entry in the database also includes rich geographical information, neighborhood data, crime type, etc. Cleaning this data required very little effort and yielded interesting results.

Exploring Crime Data: After cleaning, I began analyzing the dataset by simply grouping and graphing the number of events that took place within the boundaries of each neighborhood. It quickly became clear that the majority of police activity within the city was heavily concentrated in just five neighborhoods: Bayview, the financial district, Tenderloin, The mission district and South of market district.

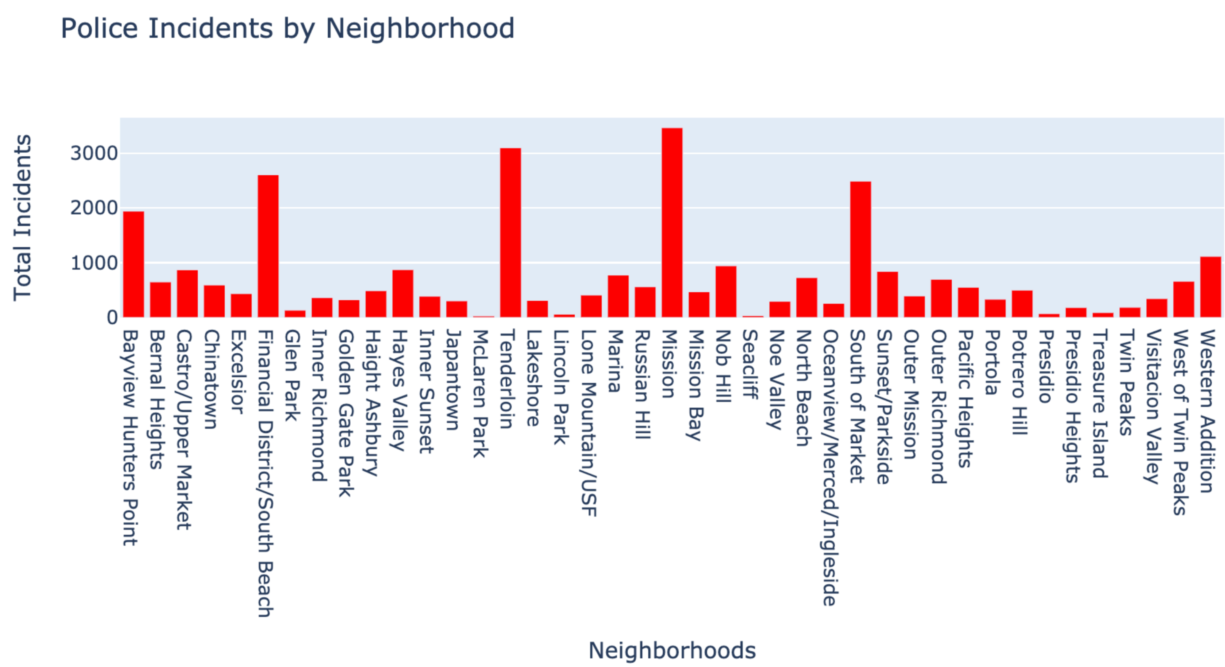


Figure 1: Crime Events by Neighborhood

Graphically, it's easy to identify neighborhoods with high rates of crime, however, this doesn't tell the whole story because crime simply doesn't end at the boundary of one neighborhood, nor begin at the boundary of another. As a result, I've utilized the Google Maps API to make an interactive heatmap (Figure 2) outlining precisely where police incidents have occurred. This data helps us identify good areas even in neighborhoods with statistically high levels of crime. Unfortunately, due to Google's restrictions on the Maps API, I will not be including the code used to create this map in my final project notebook. Based on crime levels alone, it looks like: **Pacific Heights, Height, Noe Valley, and Russian Hill look like great candidates for a safe living environment.**

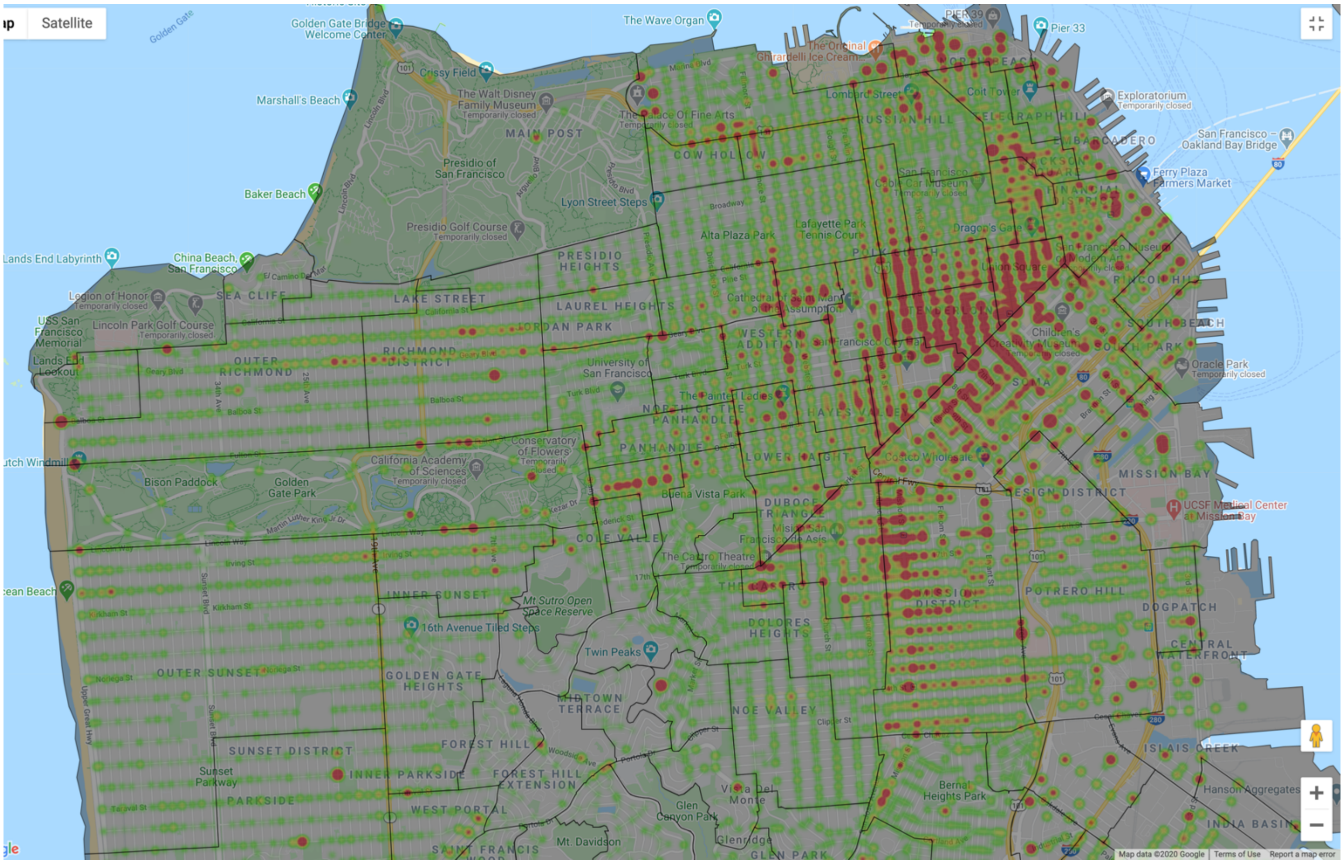


Figure 2: SF Crime Heatmap (Google Maps API)

<u>Scraping / Cleaning Foursquare Data:</u> With the crime dataset analyzed, we'll now layer venue data from foursquare. To pull venue data, I passed a data frame containing the list of neighborhoods from the crime dataset into a for loop designed to query Foursquare for 100 venues within the boundaries of each neighborhood located within the city of San Francisco. The data was then grouped by neighborhood yielding over 3,300 individual venues. After grouping, the data was OneHot Encoded. The resulting dataframe displays the types of venues and their density grouped by neighborhood.

<u>Feature and Model Selection – Clustering Neighborhoods</u>: By incorporating foursquare data into our analysis, we can better understand the venues or community spaces that shape the characteristics of candidate neighborhoods in San Francisco. For example, I personally value residential areas with high concentrations of three types of venues: Parks, Gyms and Coffee Shops. As a result, I'll be passing the OneHot Encoded Foursquare data into a KMeans classifier designed to identify and cluster neighborhoods with similar densities of: Parks, Coffee Shops and Gyms. Given the range and diversity of neighborhoods in San Francisco, specified a total of seven clusters, allowing the algorithm plenty of freedom to cluster various neighborhoods differently. Once the neighborhoods were clustered, I plotted the neighborhoods on a map and examined their characteristics, shown in figure 3 below. Once plotted, I examined the resulting clusters and found that clusters 2 (blue) and 0 (red)

contained the highest quantities of park venues with adequate densities of gyms as well as coffee shops. **Cluster 2 Neighborhoods consisted of: Haight, Twin Peaks, Golden gate park, and Castro/ Upper Market. Cluster 0 Neighborhoods contained: Visitacion Valley, Russian Hill, Presido, and others.** Based on these results, I'm confident that any of the neighborhoods mentioned above will contain a variety of quality parks, coffee shops and gyms. With a viable list of candidate neighborhoods identified, it's now time to explore the 2nd part of the analysis: finding fairly priced apartments located in the neighborhoods above.

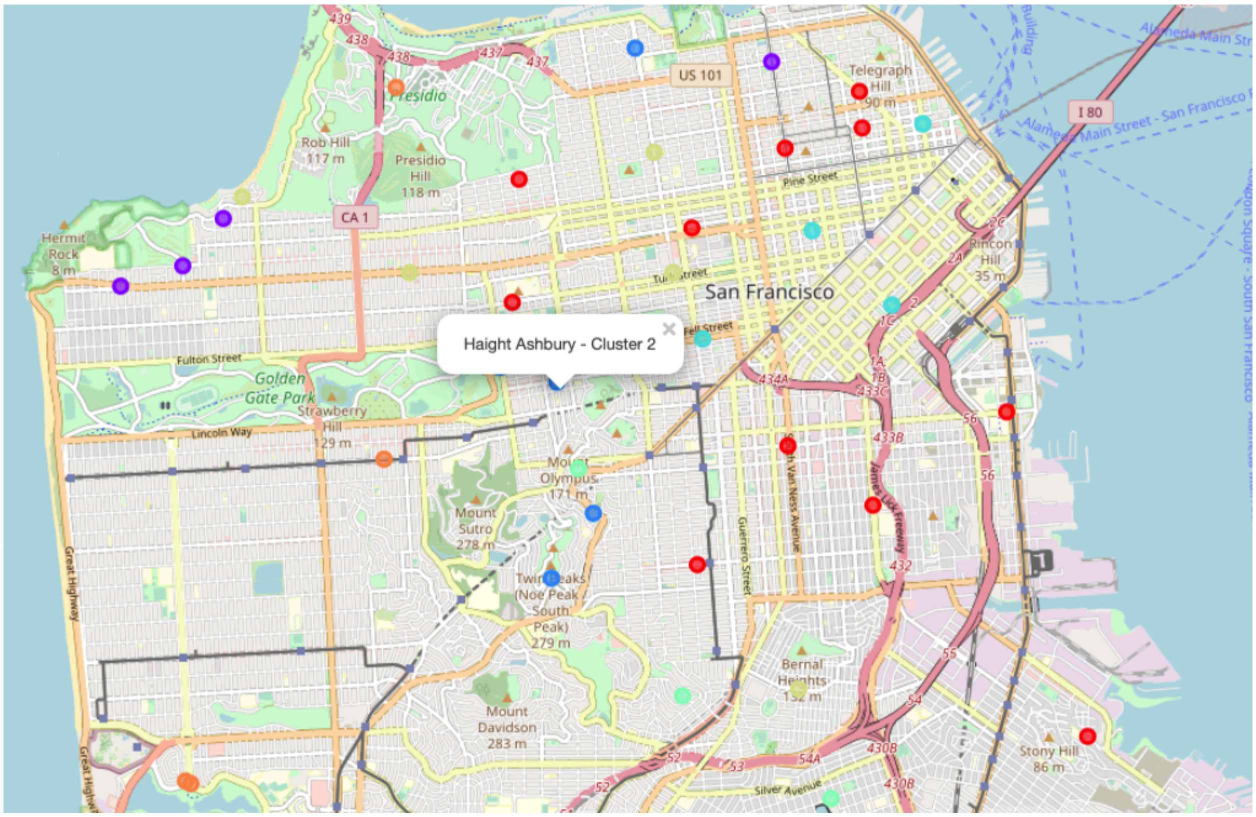


Figure 3: Foursquare Neighborhood Clusters

5. Methodology: Understanding Apartment Pricing in San Francisco

Scraping / Cleaning Apartment Listing Data: Current apartment data was scraped from Craigslist Housing into a data frame consisting of over 2000 current apartment listings. In the process of scraping, my search criteria was designed to capture as many listings as possible, as a result, the search was limited to rental properties the city of San Francisco with prices between $1000 and $10,000 per month. Once the scrape was complete, the resulting data frame was extremely messy and contained may columns that ultimately discarded (eg: listing photos). I began cleaning the data frame by dropping all unnecessary columns and rows with null data in the following columns: Location, Area (Sq/Ft), or prices. This resulted in a final data frame containing around 1100 cleaned listings (roughly half of the original 2000) and contained the following columns:

- Listing id

- Listing URL
- Listing Price
- Area (SqFt)
- Bedrooms, Bathrooms,
- Location (lat, lon), Neighborhood
- Date Created

<u>Exploring Apartment Listing Data</u>: After cleaning the data, I plotted the Apartment Listings data frame using Folium Maps and overlaid a GeoJSON file containing San Francisco neighborhood boundaries. Yellow points on the map indicate an available apartment, red points on the map display the neighborhood name.

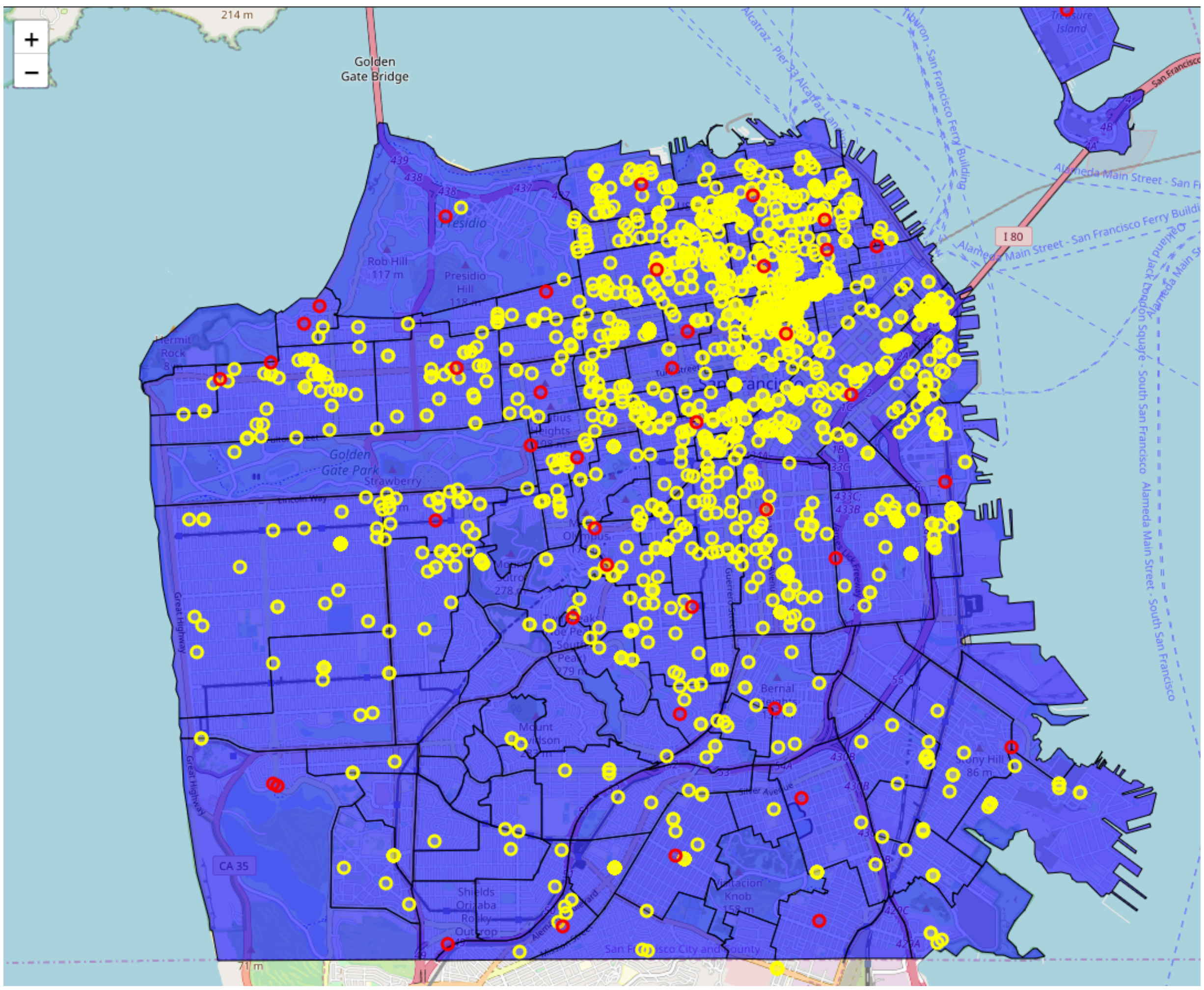


Figure 4: SF Craigslist Listings

Once the data was plotted, It immediately became clear that most available apartments listed were clustered in denser and often more unsafe neighborhoods. My hypothesis was confirmed when I graphed the total number of listings, sorted by neighborhood. Figure 6 (shown below) indicates that neighborhoods with the highest available listings (nob hill and SOMA) both contain high levels of crime. However, both Haight and Russian Hill, neighborhoods identified in the first section of this analysis, were listed as two of the neighborhoods containing a higher proportions of listings.

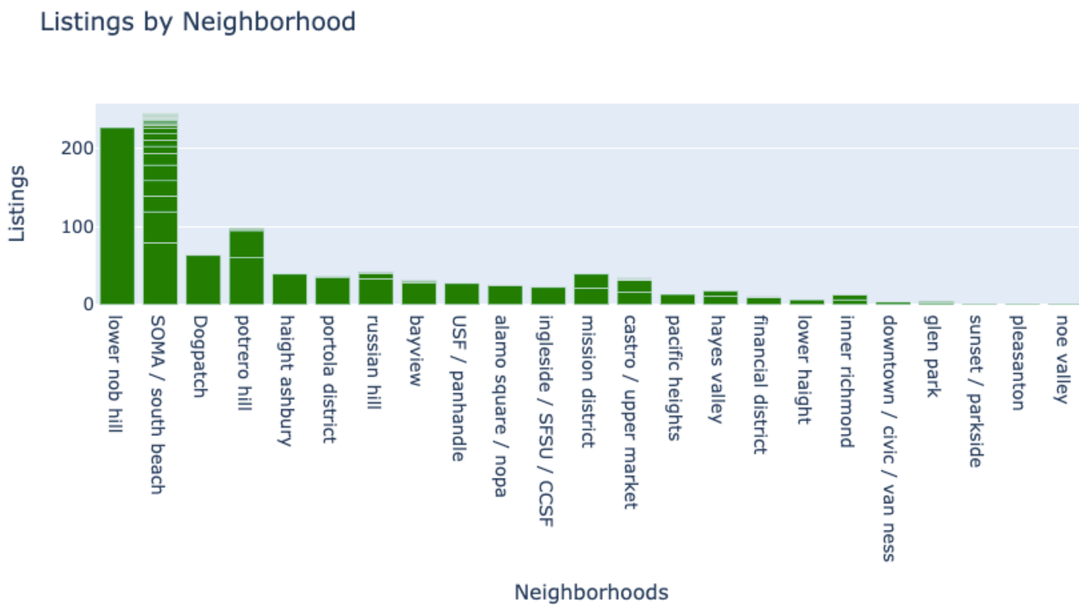


Figure 5 – Listings by Neighborhood

Figure 7 (shown below) illustrates the mean, median, upper as well as lower pricing bounds across all neighborhoods in the city. This is an essential graph that provides valuable context and clarity in the rental market. Examining the neighborhoods discussed earlier, we can observe that neighborhoods like: Russian Hill, Haight and Pacific heights contain listings with a mean rental prices greater than that of the San Francisco Market as a whole.

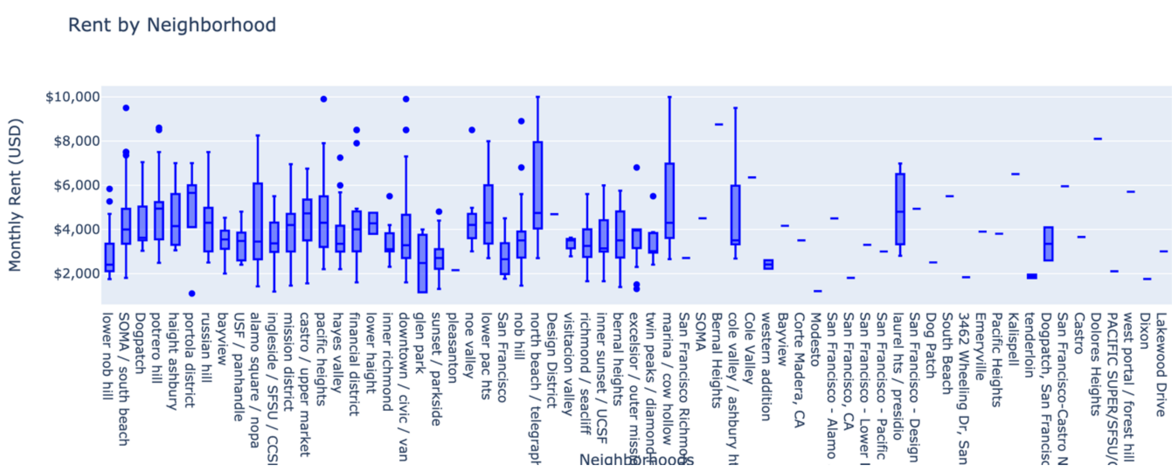


Figure 6 – Pricing by Neighborhood

Figure 8 (shown below) illustrates the mean, median, upper as well as lower area or square footage bounds across all neighborhoods in the city. This graph highlights an interesting relationship between price and area that will be further explored in figure 9.

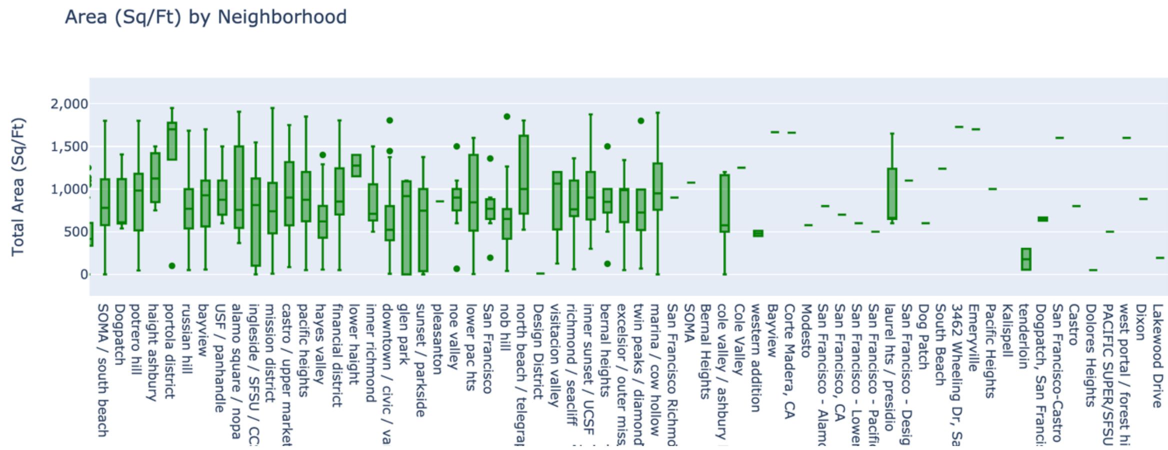


Figure 7 – Pricing by Neighborhood

Figure 9 represents a correlation matrix containing all numerical columns for each apartment listing. Of note here is the strong relationship between price and bedrooms as well as the relationship between area and price as well as bedrooms. Based on these relationships, it's clear that Area, bedrooms and even Longitude and Latitude (location) have an effect on price. With the initial findings out of the way, it's time to build a model that can identify underpriced as well as overpriced apartments in San Francisco.

Figure 8 – Feature Relationship Matrix

<u>Feature and Model Selection – Predicting apartment prices</u>: The final step in my analysis is centered around helping newcomers to the San Francisco apartment rental market identify both underpriced and overpriced apartments. To do so, I considered using many clustering

machine learning algorithms but ended up using: SciKitLearn's Random Forrest regressor to classify and predict apartment prices based on the data already obtained and stored in the craigslist data frame. Given limited time, I chose only four features to pass into the model: # of bedrooms, area (sq/Ft), Longitude and Latitude. Given the four inputs, the model examines all 1100 listings to determine the relationship between price and four variables. As indicated below, the relationship between bedrooms, area as well as location proved to have a significant impact on price. As shown in figure 10 below, the model places the heaviest weight on the number of bedrooms, closely followed by area. Location proved to play a lesser, but still important role in the pricing model. Finally, given the data provided, I'm confident that this model can provide useful predictions given it's score of .72. That being said, there is unquestionably room for improvement.

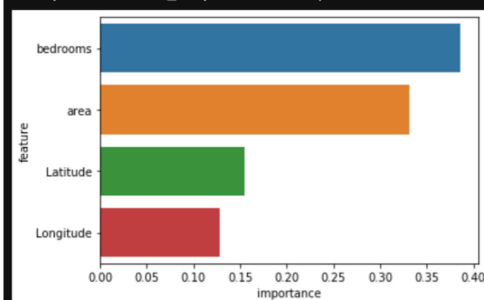
Pricing Model Test

```
address = '2790 Pine Street San Francisco, CA 94115'

geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print(latitude, longitude)

price_pred = model_pr.predict([[latitude, longitude, 775, 2.0]])
print('The predicted price is: ${}'.format(float(price_pred)))
```

```
37.78701471428571 -122.44167157142857
The predicted price is: $3829.37
```

	feature	importance
3	bedrooms	0.385814
2	area	0.331803
0	Latitude	0.154543
1	Longitude	0.127840

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c309c9390>
```




```
model_pr.score(x_test_pr, y_test_pr)
```

```
0.7204249631066979
```

Figure 9 – Random Forrest feature importance graph & pricing test

<u>Predicting apartment prices – Random Forrest Model Results:</u> To test the model, I chose an apartment at random from Zillow, not Craigslist, to ensure that my model was not tested using the data I scraped earlier. Given some refinement to the number of regressors, my pricing model successfully and accurately determined the apartment's rent, within $100 of the listing on Zillow. Given these results, I'm confident that the Random Forrest classifier can help newcomers to the San Francisco market identify both over and underpriced apartments.

6. Results

<u>Problem #1: Identifying 'Quality' neighborhoods:</u> Beginning by targeting quality neighborhoods, I utilized two data sources (Crime and Venue data) to screen quality neighborhoods from poor quality neighborhoods. As a result of cleaning, filtering and combining the datasets**, I discovered that neighborhoods such as: Russian Hill, Pacific Heights, Height and Presidio** are safe and contain an abundance of venues like parks, gyms and coffee shops. As a result, I'll be targeting my search in these areas.

<u>Problem #2: Identifying fairly priced, overpriced and underpriced apartments</u>: Thanks to the Random Forrest Classification model, I was able to uncover a list of approximately 10 underpriced apartments in each of the neighborhoods listed above. Assuming I was actually in the market for a new apartment, the results yielded should be more than enough to narrow my search to find the perfect apartment.

7. Conclusion

Though simplistic, the methods and data discussed above successfully helped me sort through thousands of apartment listings in the city of San Francisco. By utilizing venue and crime data, we were able to objectively distinguish good quality neighborhoods from poor quality neighborhoods. As a result, identifying underpriced apartments in quality neighborhoods vastly improved the process of searching for new apartments in the city of San Francisco.

8. Future Directions

Given the results of data and models used, I'm confident that it will provide useful results. Regardless, I view my attempts here simply as a proof of concept. To take my analysis to the next level, I'll need to begin with more metrics that signal neighborhood quality. For instance, I'll need to quantify each neighborhood's access easy access to public transportation as well as integrate more venue data from Foursquare and other sources. I'll also need to integrate this data into the Random Forrest apartment pricing to more accurately determine how neighborhood data affects apartment pricing. Finally, I'll need to develop a method to visualize the data concisely in one map to make findings as clear as possible to end users.