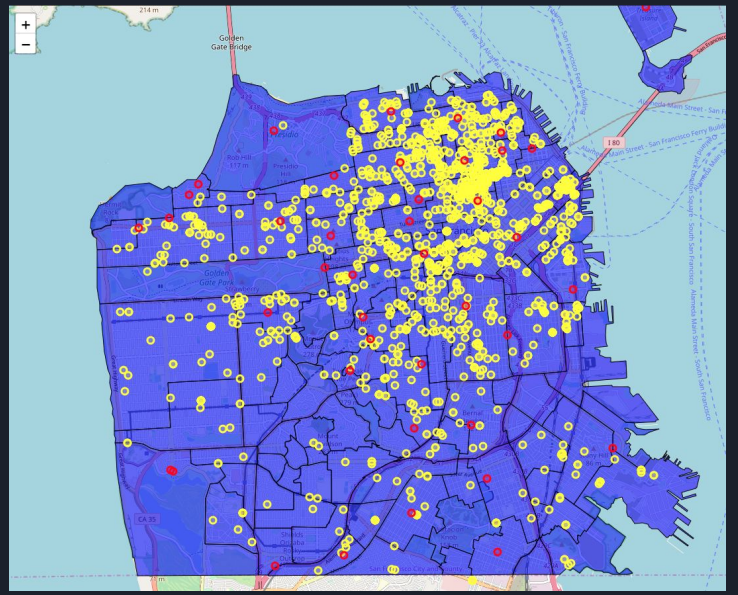
Apartment Hunting with Data Science in SF

Brian Moore

Introduction

- Moving to a new city is tough, newcomers don't really understand the real estate market.
- In competitive markets like San Francisco, competition is tough and underpriced apartments are quickly snatched up.
- Data Science tools can help newcomers overcome two problems:
  - **Problem #1:** Identifying 'Quality' neighborhoods well suited to their individual needs.
  - **Problem #2:** Identifying fairly priced, overpriced and underpriced apartments in desirable neighborhoods.



Data acquisition and cleaning

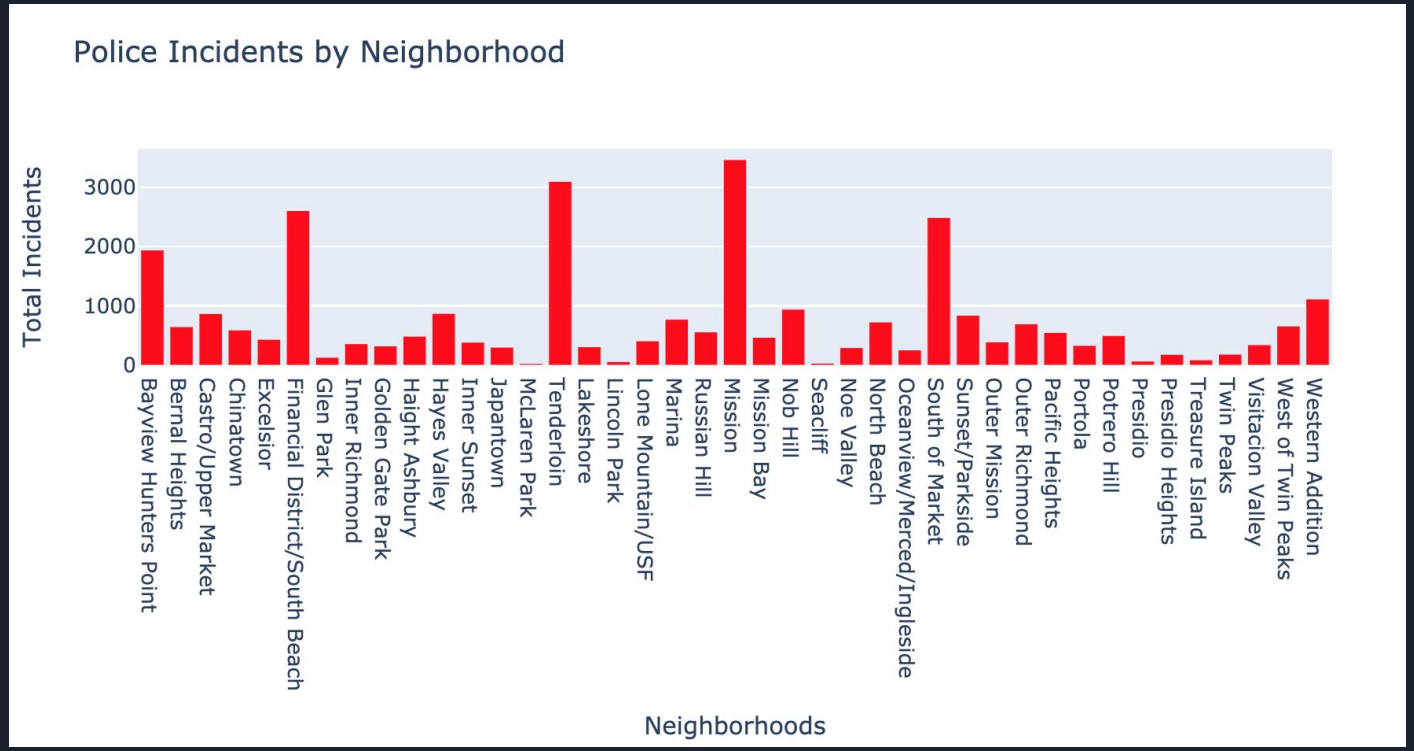
Three datasets were used to complete this analysis:

- **Dataset #1:** Craigslist Apartment listings (2000 Listings)
- **Dataset #2:** Foursquare Venue data (3300 Listings)
- **Dataset #3:** Crime data provided by The City of San Francisco (333,000 Crimes)

Cleaning the data :

- Craigslist data was cleaned by dropping listings without numerical data leaving a clean dataset of 1000 listings.
- Crime data was restricted to a smaller subset containing only crimes that took place after Jan 1, 2020.

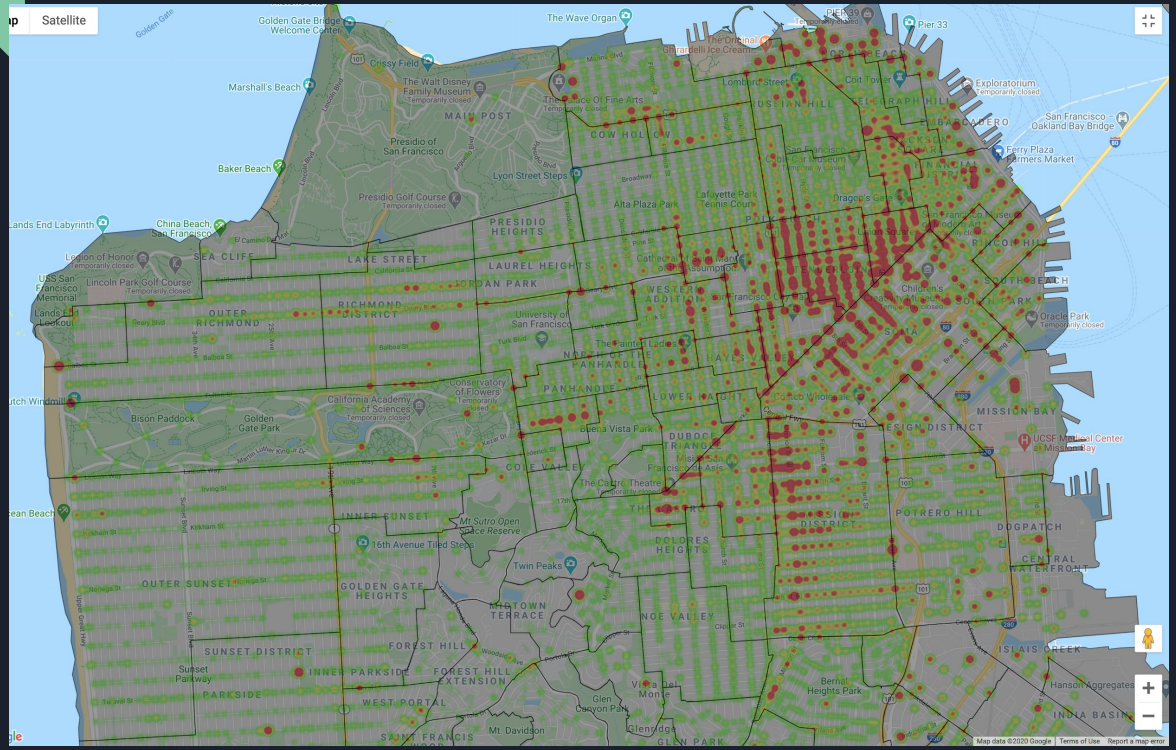
Finding a Safe Neighborhood



Crime data from the city of San Francisco shows most crime takes place in just five Neighborhoods:

- Bayview
- Financial Dist.
- Tenderloin
- Mission
- SOMA

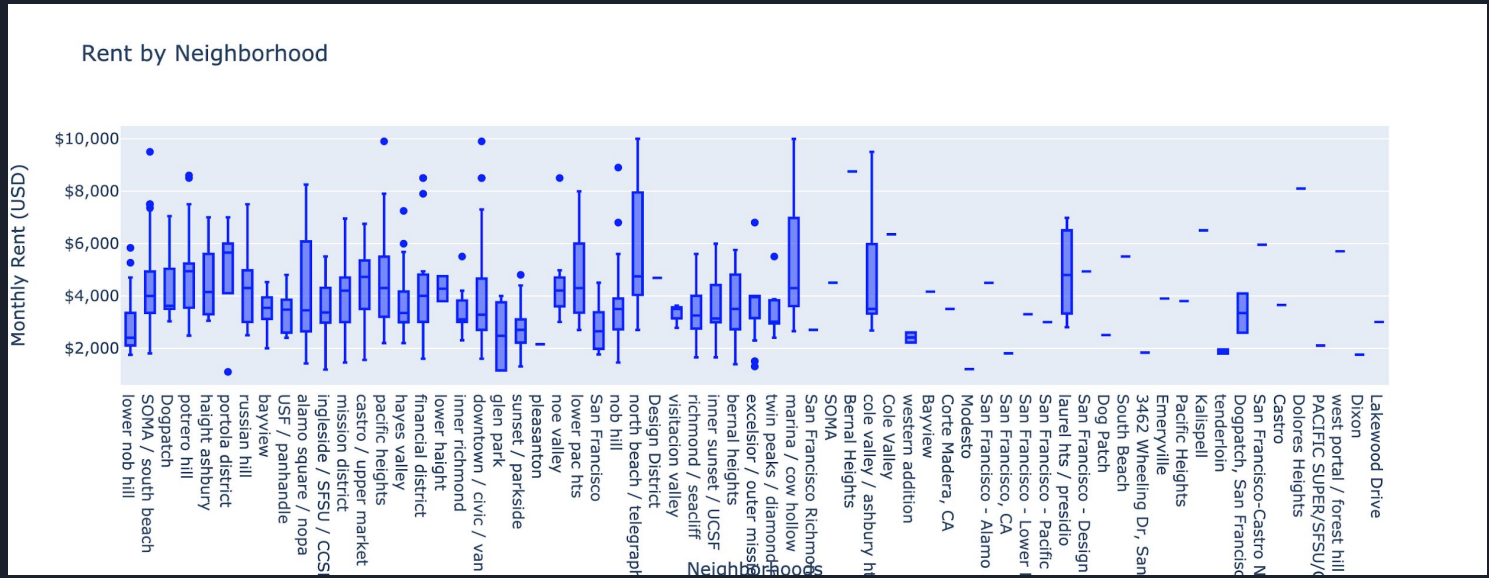
Finding a Safe Neighborhood



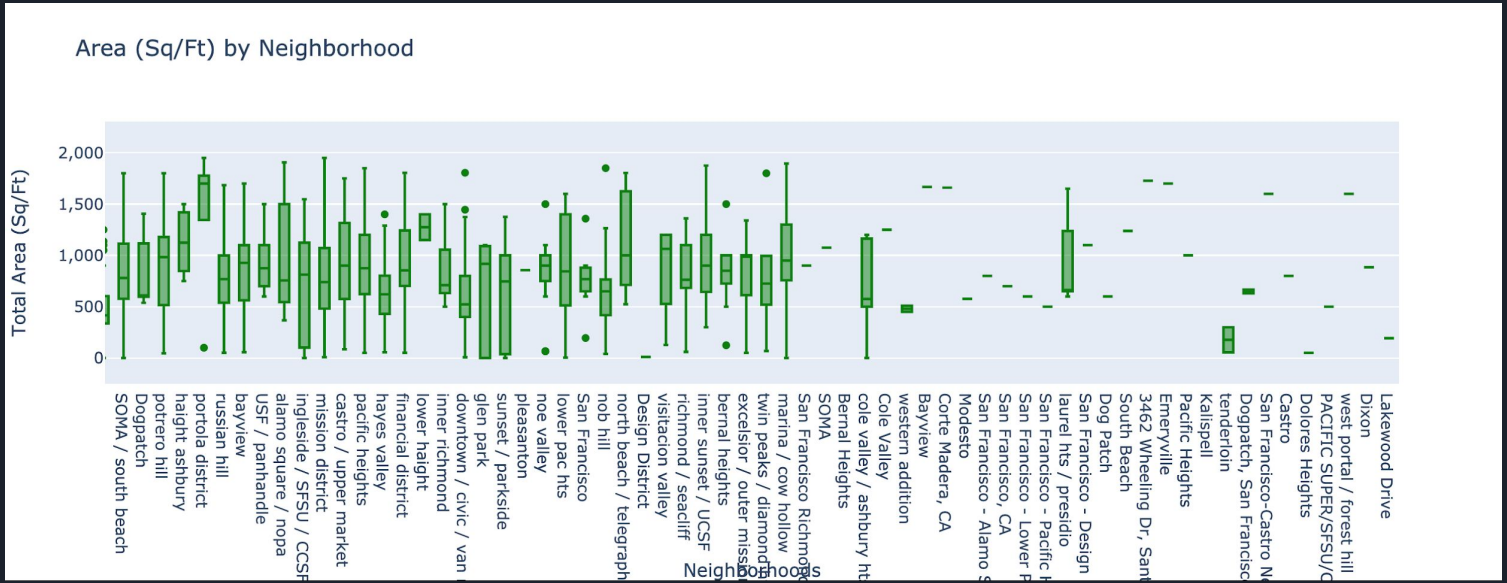
Using a heatmap, we can see that crime is not strictly limited by neighborhood boundaries.

Finding Under/ Overpriced apartments

Rent varies widely by neighborhood and in often influenced by crime and venue selection.



Data acquisition and cleaning

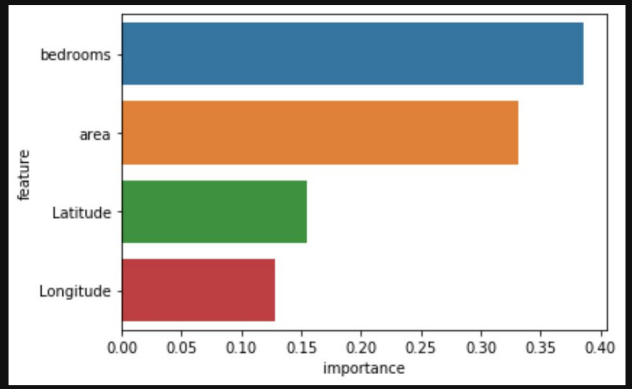


Regardless of price, apartments are quite small in San Francisco.

Random Forest - Model Performance

	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
```

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
```

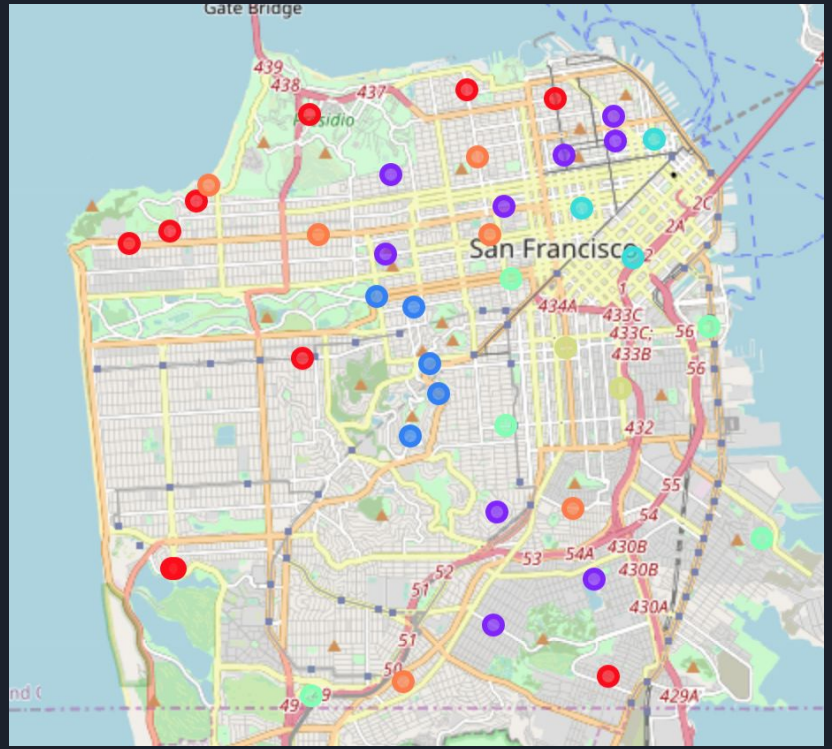
This model performed adequately and yielded usable predictions.

Kmeans Classifier - Model Performance

- **Targeted Venues:** Gyms, Parks, Coffeeshops

- Each cluster represents higher or lower densities of each venue
- **Highest density:** Blue Clusters (Cluster 2, Cluster 0)
- **Lowest Density:** Green clusters (Cluster 3)

- **Conclusions:** We'll target apartments in Cluster 2 or 0



Findings and Conclusion

1. **Findings:**
    a. By combining and analyzing data from three sources, we can accurately determine ideal neighborhoods and identify under/overpriced apartments.
    b. Neighborhoods identified include: Russian hill, pacific heights, Height and presidio.
2. Future Directions
    a. Integrate more data from other sources into existing datasets from Craigslist.
    b. Import data from craigslist and Zillow.
    c. Quantify each neighborhood's access to public transportation.
    d. Develop a unifying visual to make presentation to end users more clear.