Exploring San Francisco Neighborhoods

Through the lens of police incident reports, citizen reported 311 cases, and housing prices from 2018 to 2020

Introduction

According to Wikipedia, San Francisco is the cultural, commercial, and financial center of Northern California, the 16th most populous city in the United States, and the 4th most populous in California. San Francisco has some of the highest median home prices in the world.

With this backdrop, we want to analyze the different neighborhoods of San Francisco. Specifically, we will look at the following data of each neighborhood in the time frame spanning from January 2018 up to and including September 2020, aggregated by month:

1. Police activity: Represented by San Francisco Police Department incident reports, retrieved from Police Department Incident Reports: 2018 to Present | DataSF | City and County of San Francisco
2. Citizen involvement: The city of San Francisco provides a customer service center called SF 311, which opens cases on anything from requesting general information to reporting issues requiring nonemergency, government involvement (i.e. noise complaints, abandoned vehicles, excessive litter). Data was retrieved from 311 Cases | DataSF | City and County of San Francisco
3. Housing sales prices: Data was retrieved from Redfin's Downloadable Housing Market Data and aggregated by Redfin Region. In order to map Redfin Regions to the official San Francisco neighborhoods used in the two City-provided datasets, we also needed:
    a. Official San Francisco dataset containing Analysis Neighborhoods referenced by the San Francisco Police Department incident reports dataset
    b. Official San Francisco dataset containing SF Find Neighborhoods referenced by the 311 cases dataset

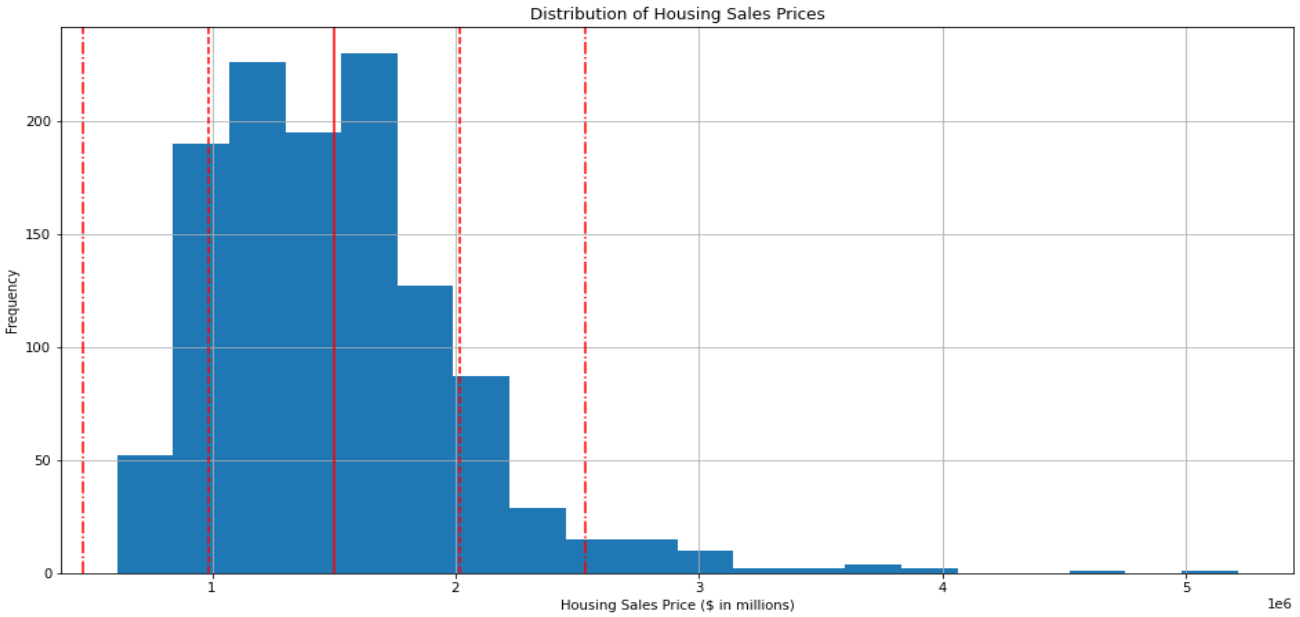
Problem Statement

We want to explore the relationship between housing prices and police incident reports and 311 cases. That is, using supervised learning, does a model exist such that housing prices can be predicted from features derived from the police incident reports and 311 cases?

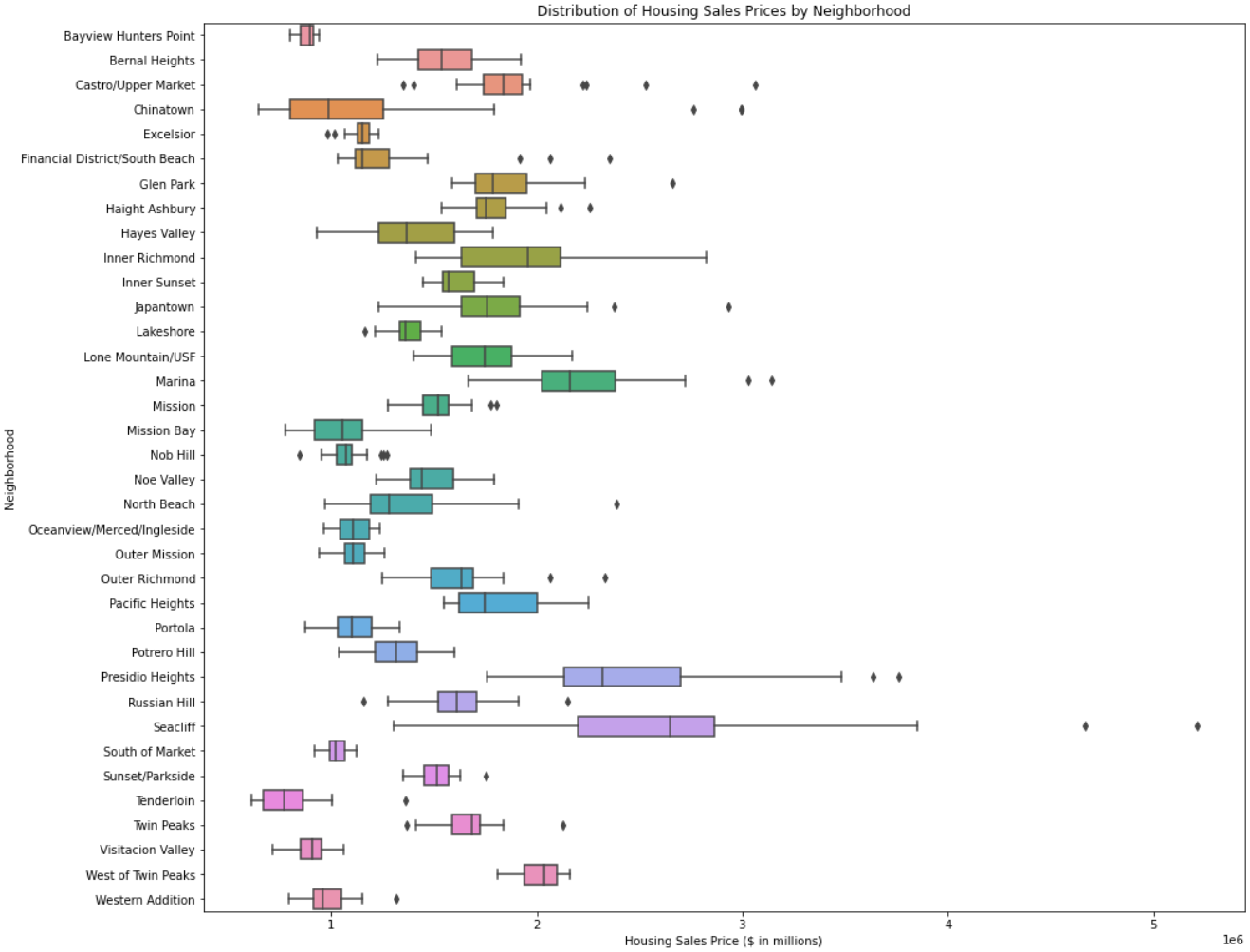
Barring such a relationship, can we use unsupervised learning (clustering) to understand how the neighborhoods differ from one another?

Exploratory Data Analysis

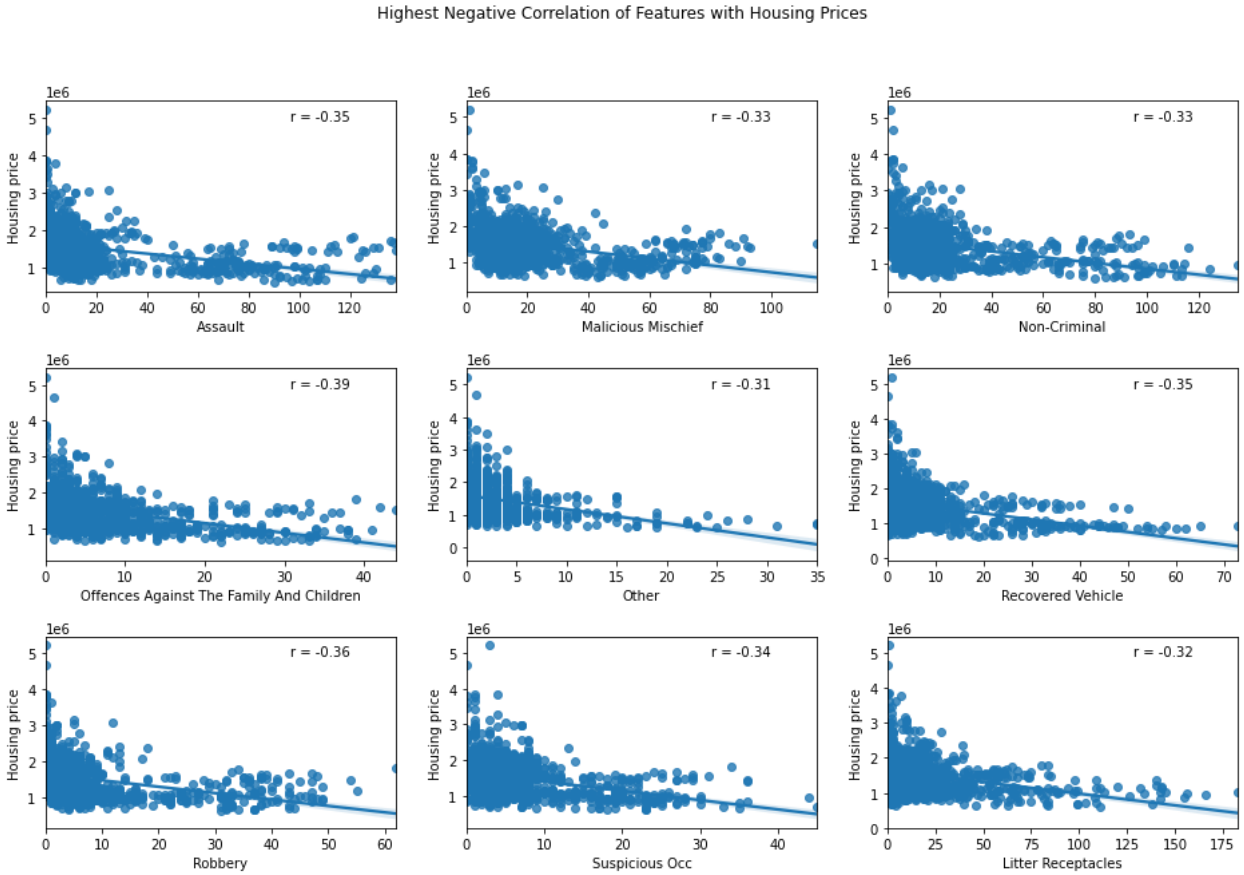
We first looked at the overall distribution of housing prices in San Francisco. We can see below that the distribution skews to the right and has a mean housing price of approximately $1.5M, with a standard deviation of approximately $0.5M.



Then, we looked at the distribution of housing prices across the different neighborhoods.



We can see from above that neighborhoods can differ greatly in their housing sales price from one another. Next, we identified the features with the strongest negative correlation with housing prices.



Most of these features correspond to some sort of crime (Offences Against The Family and Children, Assault, Malicious Mischief, Robbery, and even Recovered Vehicle). Interestingly, Litter Receptacles appears to indicate that lack of care in the neighborhood's maintenance correlates with a reduction in housing prices.

Modeling – Predicting Housing Prices

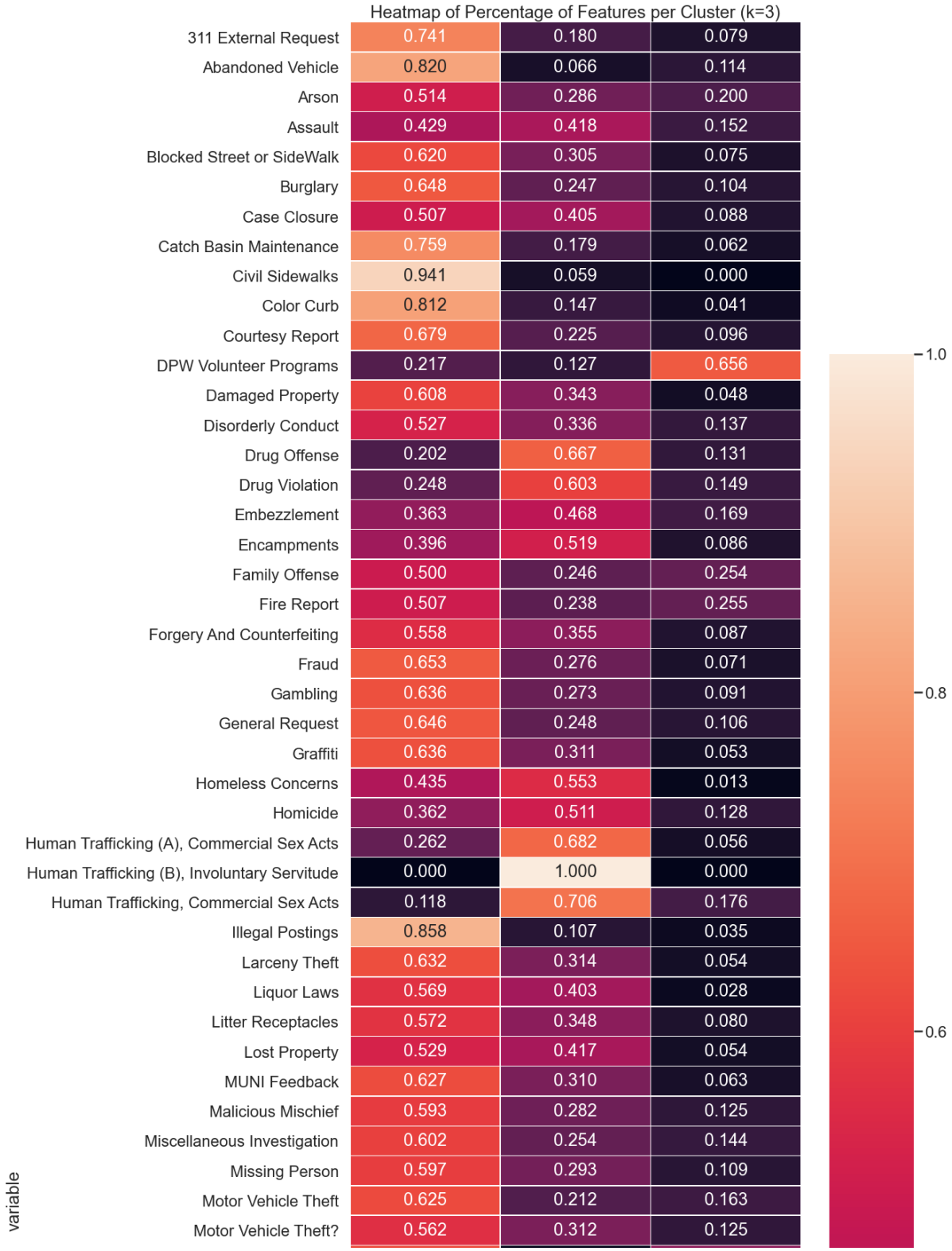
We built different machine learning models to predict San Francisco housing prices based on police incident reports and 311 cases, but our best model, which ended up using Min Max Scaler, PCA with n_components = 7, and a Random Forest Regressor of max_depth = 10 and n_estimator = 31, was still quite inaccurate at predicting the housing prices, with a R2 score of ~35% on the test set. Why?

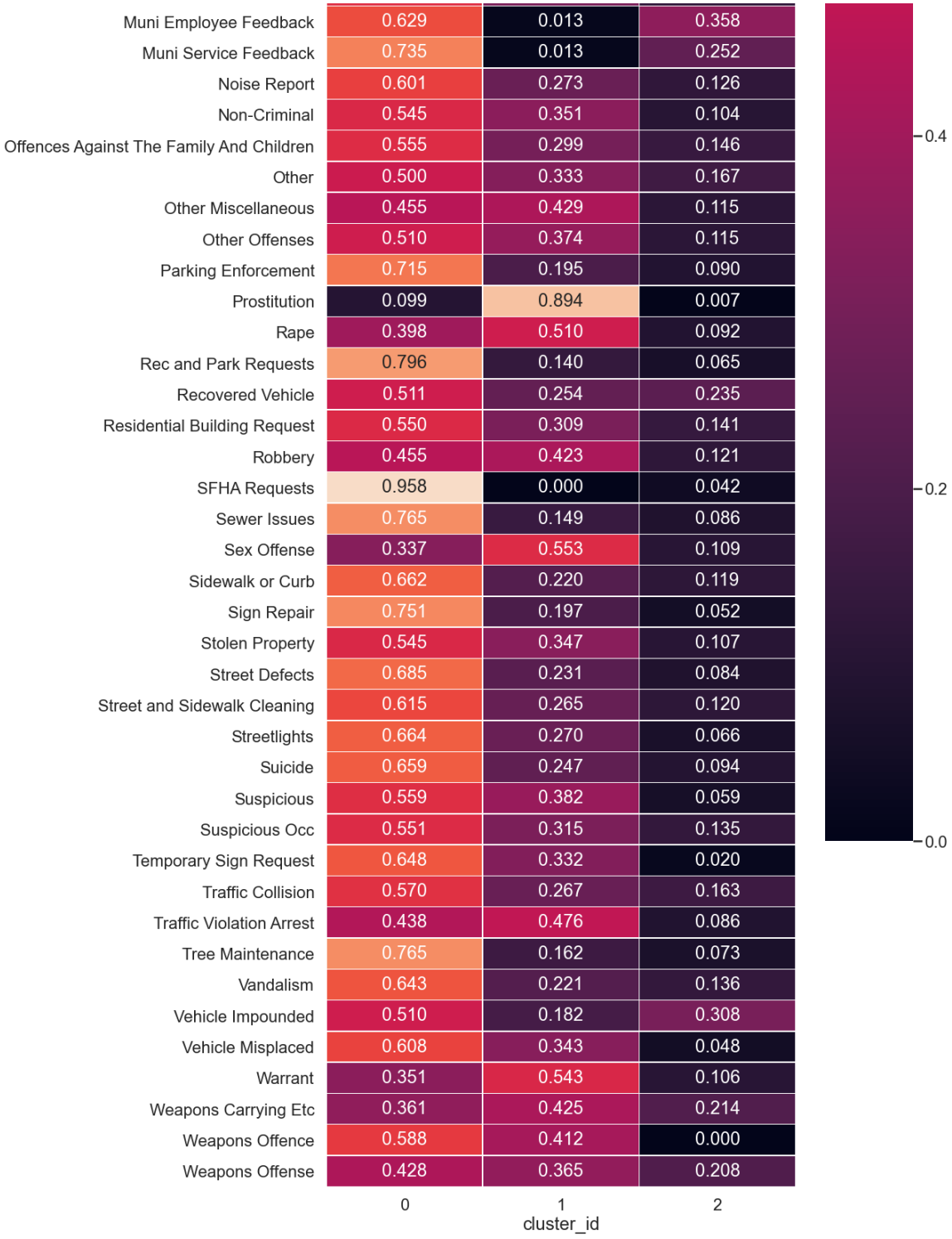
1. **Are outliers disrupting the model?** Perhaps; the effectiveness of `MinMaxScaler`, which boxes the range of the features in, shows that the features' large range has an effect on the modeling. However, we did try removing "outliers", i.e. features that have the greatest range and most disrupt the distribution, and there was negligible effect on model performance.
2. **Is this the curse of dimensionality?** We might have too many features and not enough data. However, if this were the main detractor of model performance, we might have seen a greater improvement when we limited the number of PCA components used, and we did not see that.
3. **Do police incident reports and 311 cases actually influence housing prices?** Our initial hypothesis was that we could develop a model to predict housing prices in San Francisco based on police incident reports and 311 cases. Perhaps this is not the case and our initial hypothesis is invalid. Could neighborhoods actually be homogeneous?

Thus, we shifted to an unsupervised learning approach.

Modeling – Clustering Neighborhoods

We used an unsupervised learning approach to create clusters of our features (derived from police incident reports and 311 cases) and compare them to the Neighborhood labels to determine accuracy of the model. We used K-Means Clustering and found that we should choose the number of clusters to be 3. Doing this revealed the following heatmap of our features across the 3 clusters:





We can treat cluster_id = 0 as the "catch-all" cluster, or the cluster representing the majority of San Francisco neighborhoods. Meanwhile, cluster_id = 1 has the majority of the following:

- Drug Offense
- Drug Violation
- Embezzlement
- Encampments
- Homeless Concerns

- Homicide
- Human Trafficking
- Prostitution
- Rape
- Sex Offense
- Warrant

We can see that some of those features likely strongly correlate with each other (i.e. Prostitution and Sex Offense), but it appears that this cluster of neighborhoods captures more serious incidents.

Cluster_id = 2 has the following identifiers:

- Vast majority of DPW Volunteer Programs
- Nearly no Encampments or Homeless Concerns when compared with the other two clusters.

To clarify the "DPW Volunteer Programs": this is a Category of 311 requests. San Francisco's Department of Public Works runs various volunteer initiatives, focusing on keeping streets clean and removing graffiti. The 311 cases of this type are issued when, for example, a volunteer needs additional cleaning supplies or paint from the city to remove graffiti.

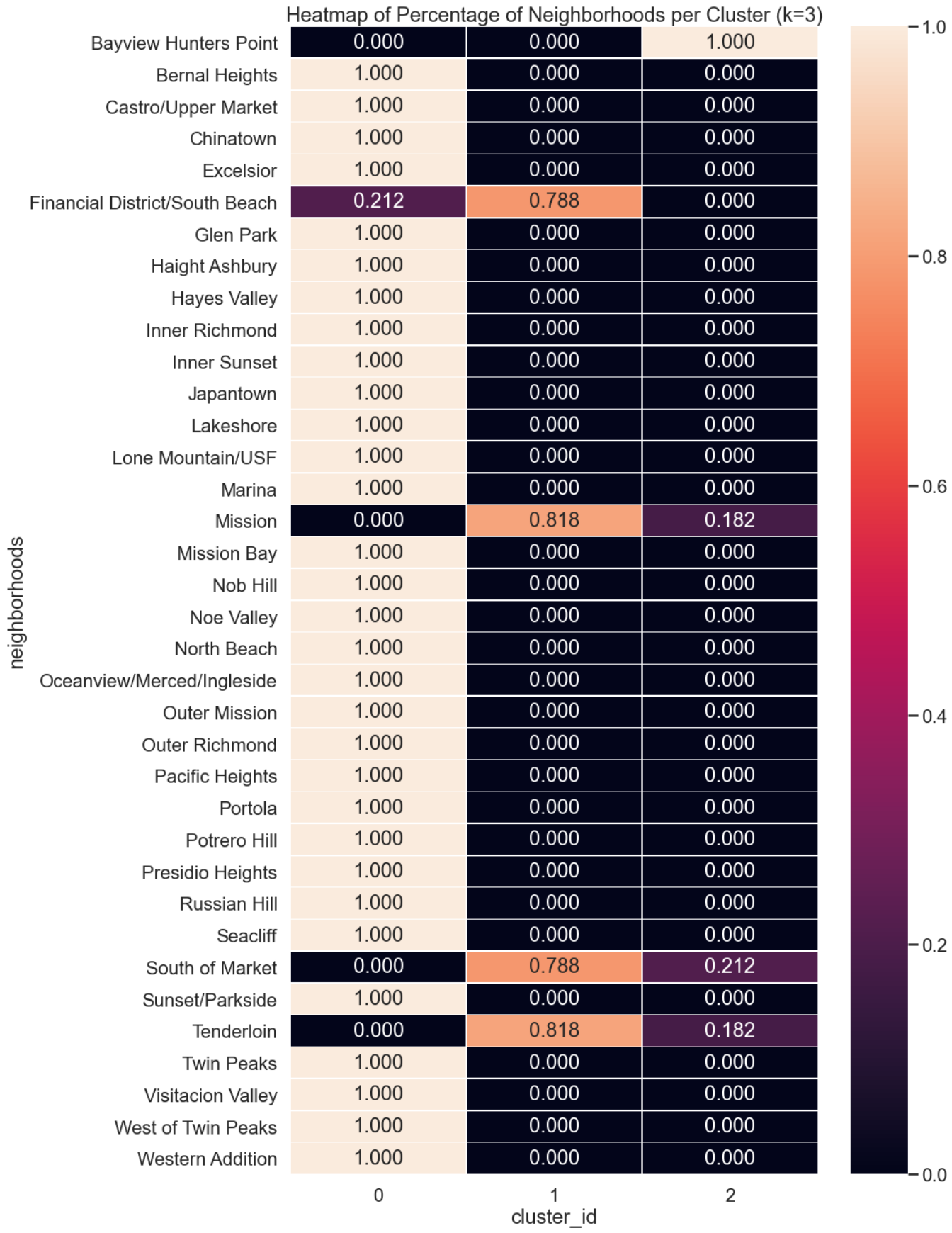
It is somewhat surprising that this cluster has nearly no Encampments or Homeless Concerns, when those issues are so pervasive across the rest of the city.

Let's tie the above heatmap of the features per cluster to the neighborhood labels (see below). Recall, we treated cluster_id = 0, which contained the majority of the neighborhoods, as a catch-all cluster, or in other terms, what one could consider "normal" for San Francisco. This left the following neighborhoods as "outliers":

- Bayview Hunters Point
- Financial District/South Beach
- Mission
- South of Market
- Tenderloin

Bayview Hunters Point falls entirely in cluster_id = 2. From our earlier analysis of features, it appears that this neighborhood has graffiti and street cleaning issues, but nearly no homeless concerns.

Financial District/South Beach has approximately 80% in cluster_id = 1 and 20% in cluster_id = 0. The rest of the neighborhoods, Mission, South of Market, and Tenderloin, also have approximately 80% in cluster_id = 1 but the remaining 20% in cluster_id = 2. These neighborhoods, over the past decade, have been undergoing drastic changes in neighborhood makeup due to startups, construction (new buildings and renovation of old buildings), and an influx of jobs and people. This analysis shows that, although many aspects of these neighborhoods have improved tremendously, work must still be done to clear out problematic issues.


Heatmap of Percentage of Neighborhoods per Cluster (k=3)
neighborhoods
Bayview Hunters Point 0.000 0.000 1.000
Bernal Heights 1.000 0.000 0.000
Castro/Upper Market 1.000 0.000 0.000
Chinatown 1.000 0.000 0.000
Excelsior 1.000 0.000 0.000
Financial District/South Beach 0.212 0.788 0.000
Glen Park 1.000 0.000 0.000
Haight Ashbury 1.000 0.000 0.000
Hayes Valley 1.000 0.000 0.000
Inner Richmond 1.000 0.000 0.000
Inner Sunset 1.000 0.000 0.000
Japantown 1.000 0.000 0.000
Lakeshore 1.000 0.000 0.000
Lone Mountain/USF 1.000 0.000 0.000
Marina 1.000 0.000 0.000
Mission 0.000 0.818 0.182
Mission Bay 1.000 0.000 0.000
Nob Hill 1.000 0.000 0.000
Noe Valley 1.000 0.000 0.000
North Beach 1.000 0.000 0.000
Oceanview/Merced/Ingleside 1.000 0.000 0.000
Outer Mission 1.000 0.000 0.000
Outer Richmond 1.000 0.000 0.000
Pacific Heights 1.000 0.000 0.000
Portola 1.000 0.000 0.000
Potrero Hill 1.000 0.000 0.000
Presidio Heights 1.000 0.000 0.000
Russian Hill 1.000 0.000 0.000
Seacliff 1.000 0.000 0.000
South of Market 0.000 0.788 0.212
Sunset/Parkside 1.000 0.000 0.000
Tenderloin 0.000 0.818 0.182
Twin Peaks 1.000 0.000 0.000
Visitacion Valley 1.000 0.000 0.000
West of Twin Peaks 1.000 0.000 0.000
Western Addition 1.000 0.000 0.000
0
1
2
cluster_id
1.0
0.8
0.6
0.4
0.2
0.0

Ideas for further research

1. The police incident report and 311 case data were aggregated by month in order to conform with the Redfin housing prices data set. With regards to the neighborhood clustering model, since we do not utilize the Redfin data, we can increase the granularity of the police incident report and 311 case data (e.g. aggregate by day or by week instead of by month).
2. We can also increase the time frame of the data. This data was retrieved in October of 2020, so there will be additional months of data to consume for the model by now. We are limited in that the police incident report data is only available from January 2018 onwards.
3. If we want to continue to tune our predictive housing price model, we can look into adding additional features related to the properties. For example, we can include price per square foot, lot sizes, types of property (single family homes vs condominiums). If a source were available, we could also look into rental prices. Although most rental rates are not commonly available, San Francisco does provide Section 8 housing for low income families and their rental rates should be available internally to the city.