SAN FRANCISCO NEIGHBORHOOD CLASSIFICATION

Jaime Landazuri

1 Introduction

San Francisco is one of the most expensive cities in the United States of America, and with the rising minimum wages, many clients are stepping back in opening any food related business in town. So, a real estate company is developing an app which include also a segment to attract the attention of those clients. Entrepreneurs in the food industry do not feel confident with the information that the real estate give to them, therefore creating a classification of best neighborhood where to start a new business will help to improve this situation.

2 Data

This project uses data from 4 sources: San Francisco Police Department, Foursquare, RentCafé, and DataSF's. Since the real estate company will only work with clients that want to rent, we are not including average buying prices for each neighborhood. RentCafé is a nationwide internet listing service (ILS) that enables renters to easily find apartments and houses for rent throughout the United States, and because they provide a table with the average rent price for each neighborhood, we will use their data. From Foursquare we can get all the venues related to food and get data about the interest in food. Using an API on DATASF, we get the information about business end dates of restaurants by each neighborhood. We also count the average incidents by neighborhood in San Francisco using the data from the San Francisco Police Department.

2.1 Data Cleaning

San Francisco Police Department provide a dataset of all the daily incidents by location. The dataset is huge, so we need to get only the information we need. We grouped all the incidents by neighborhood and count them monthly. We also used this data set to get the average geolocation of each neighborhood in San Francisco.

Foursquare provides the information about interests about some location. We classify the requests related to food service (venue category), and group them by neighborhood.

San Francisco Data portal is callable by an API and has the record of every business starting date, and end date by location. We counted all the businesses that opened and finished by month and neighborhood.

RentCafé already provides a table of averages rent by neighborhood, so we only need to scrape the data from the website.

2.2 Feature Selection

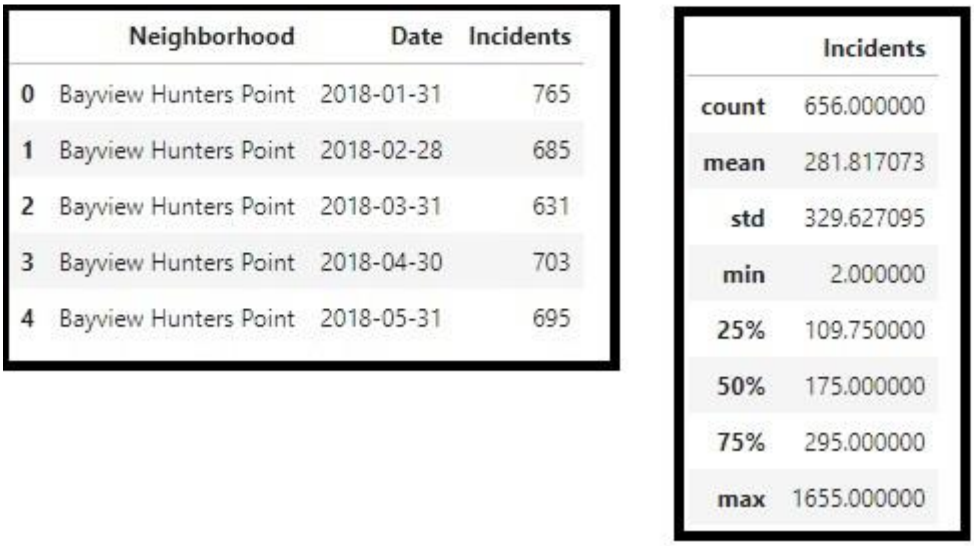
The dataset from the police department had 195142 rows with 34 features, after cleaning we got two tables. The table with the coordinates of each neighborhood has 35 rows, and the table of monthly incidents by neighborhood 656. For the business count for neighborhood we got a dataset of 236,905 rows and after doing the data preprocessing, we created two tables: monthly count of starting business by neighborhood of 3379 rows, and another table of monthly count of closing businesses by neighborhood of 663 rows. The average rent table has 84 rows. This table include more neighborhoods than the other tables, so we only use the neighborhoods listed in the coordinates table. From Foursquare we create a table with the proportion of food related venues to the total venues by neighborhood.

3 Methodology

In this project, I will use the KNN algorithm to classify the neighborhoods. To use the algorithm, we need to do some modifications in the tables. I will explain the modification for each table.

3.1.1 Incidents Table

The original table has 34 columns. I first transform the date column to datetime format, and then group the incidents by neighborhood on a monthly basis and count them. Only 3 columns are going to be used: neighborhood, date (month) and the sum of incidents.



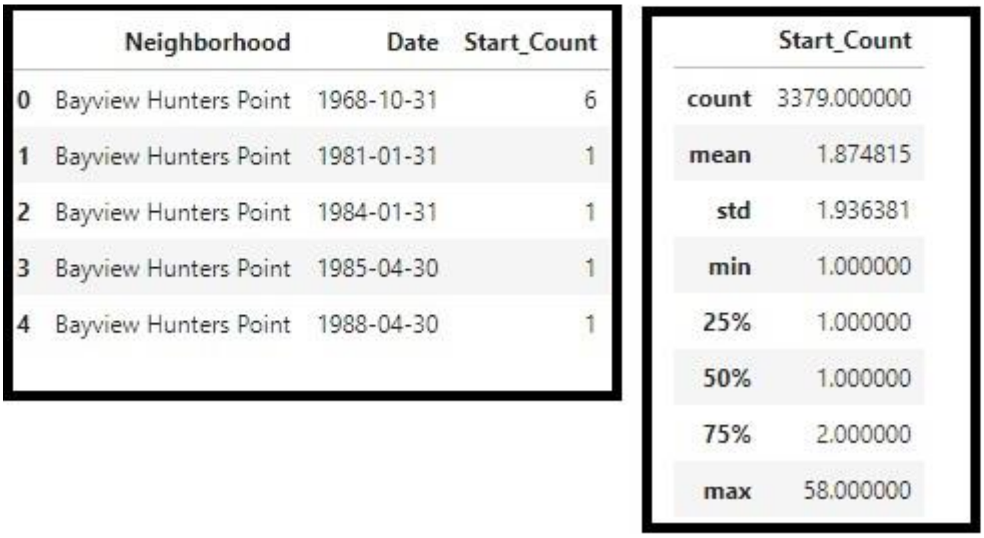
	Neighborhood	Date	Incidents
0	Bayview Hunters Point	2018-01-31	765
1	Bayview Hunters Point	2018-02-28	685
2	Bayview Hunters Point	2018-03-31	631
3	Bayview Hunters Point	2018-04-30	703
4	Bayview Hunters Point	2018-05-31	695

	Incidents
count	656.000000
mean	281.817073
std	329.627095
min	2.000000
25%	109.750000
50%	175.000000
75%	295.000000
max	1655.000000

Figure 3-1 Left: Head of the incidents table. Right: Statistics of the incidents' column.

3.2 Business Table

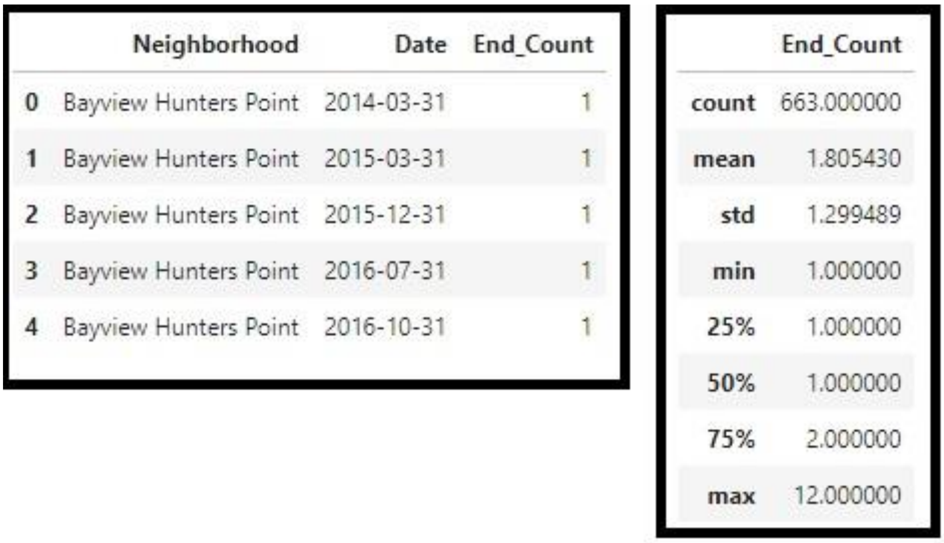
From San Francisco database, the table has 36 features. First from the NAICS code I only selected the rows containing Food Service in the NAICS code description. There were 19 unique categories listed. Then I transformed the date column to a Datetime type. I dropped all the row without any information of closing date for the closing business table. Then created two tables by grouping on a monthly basis by neighborhoods how many businesses closed and open.



	Neighborhood	Date	Start_Count
0	Bayview Hunters Point	1968-10-31	6
1	Bayview Hunters Point	1981-01-31	1
2	Bayview Hunters Point	1984-01-31	1
3	Bayview Hunters Point	1985-04-30	1
4	Bayview Hunters Point	1988-04-30	1

	Start_Count
count	3379.000000
mean	1.874815
std	1.936381
min	1.000000
25%	1.000000
50%	1.000000
75%	2.000000
max	58.000000

Figure 3-2 Left: Starting business table head. Right: Statistics of table.



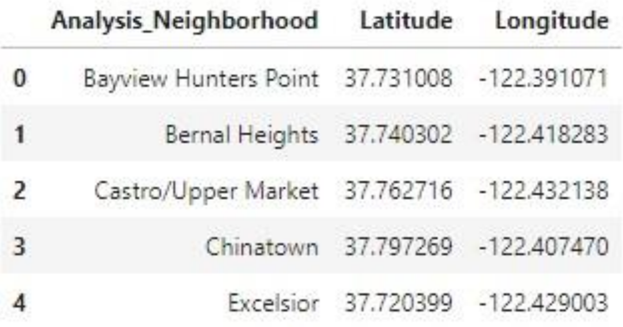
	Neighborhood	Date	End_Count
0	Bayview Hunters Point	2014-03-31	1
1	Bayview Hunters Point	2015-03-31	1
2	Bayview Hunters Point	2015-12-31	1
3	Bayview Hunters Point	2016-07-31	1
4	Bayview Hunters Point	2016-10-31	1

	End_Count
count	663.000000
mean	1.805430
std	1.299489
min	1.000000
25%	1.000000
50%	1.000000
75%	2.000000
max	12.000000

Figure 3-3 Left: Closing business table head. Right: Statitics of tabe.

3.3 Coordinates Table

I used the incidents table to get the latitude and longitude for each neighborhood. Because the original table is very extensive, some preprocessing was necessary. First I drop any row without coordinates data. Then limit to only use the first 300 rows. Then grouped by neighborhood and averaging the values in Latitude and Longitude. From this we obtained a table of 35 neighborhoods with its latitudes and longitudes.

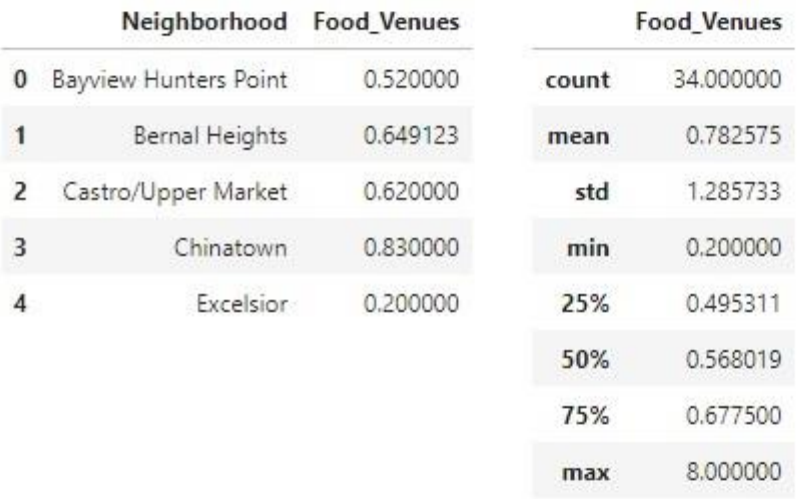


	Analysis_Neighborhood	Latitude	Longitude
0	Bayview Hunters Point	37.731008	-122.391071
1	Bernal Heights	37.740302	-122.418283
2	Castro/Upper Market	37.762716	-122.432138
3	Chinatown	37.797269	-122.407470
4	Excelsior	37.720399	-122.429003

Figure 3-4 Coordinate table head.

3.4 Foursquare Venues

Using the latitude and longitude from the above table, we retrieve the information from Foursquare from each neighborhood. We retrieved a total of 297 unique categories of venues. So we only choose those related to food. We passed from a table of 2080 rows to one of only 1243, but with information related to food services. Then we grouped by neighborhood and count. And for the final data we divided the food related venues to the total count of venues.



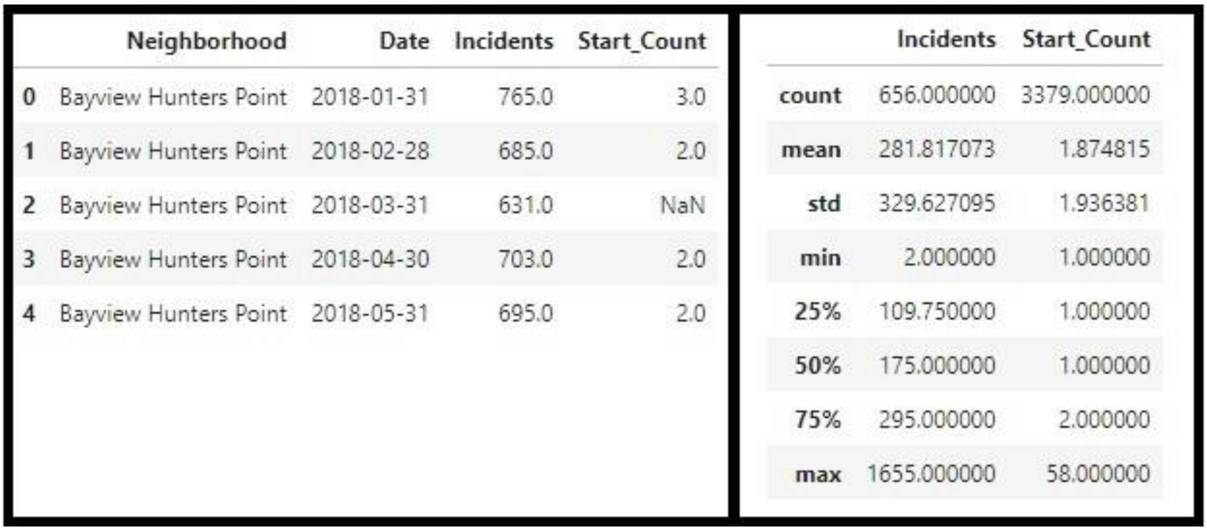
	Neighborhood	Food_Venues
0	Bayview Hunters Point	0.520000
1	Bernal Heights	0.649123
2	Castro/Upper Market	0.620000
3	Chinatown	0.830000
4	Excelsior	0.200000

	Food_Venues
count	34.000000
mean	0.782575
std	1.285733
min	0.200000
25%	0.495311
50%	0.568019
75%	0.677500
max	8.000000

Figure 3-5 Left: Food related proportion table head. Right: Statistics of table.

3.5 Creating only one table

The merging was by stages. First we merged the incidents table with the closing business table using an outer method. We obtained a table “df1” of 3711 rows and 4 features.

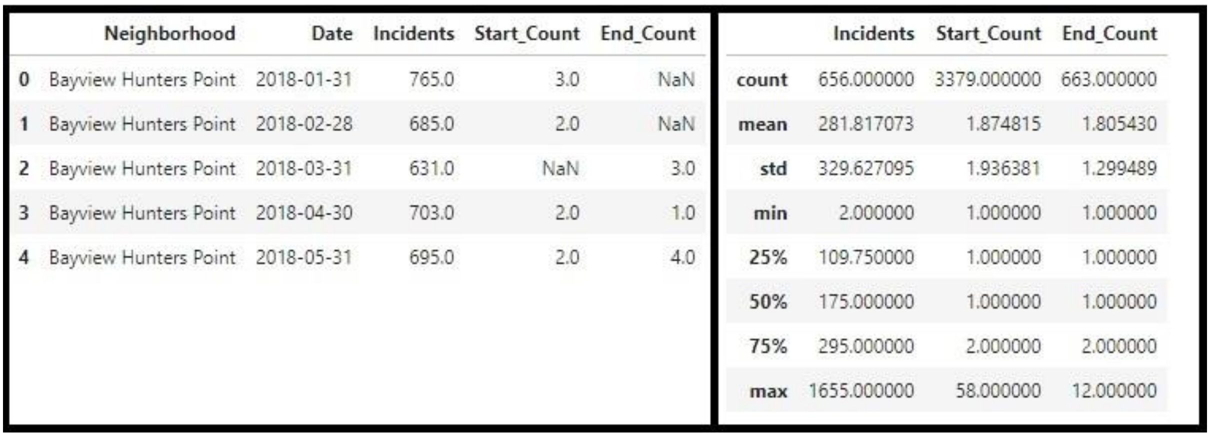


	Neighborhood	Date	Incidents	Start_Count
0	Bayview Hunters Point	2018-01-31	765.0	3.0
1	Bayview Hunters Point	2018-02-28	685.0	2.0
2	Bayview Hunters Point	2018-03-31	631.0	NaN
3	Bayview Hunters Point	2018-04-30	703.0	2.0
4	Bayview Hunters Point	2018-05-31	695.0	2.0

	Incidents	Start_Count
count	656.000000	3379.000000
mean	281.817073	1.874815
std	329.627095	1.936381
min	2.000000	1.000000
25%	109.750000	1.000000
50%	175.000000	1.000000
75%	295.000000	2.000000
max	1655.000000	58.000000

Figure 3-6 Left: df1 head. Right: df1 statistics.

Then we merged this table with the starting business table to create table df2 of 3883 rows and 5 features. We used the outer method.

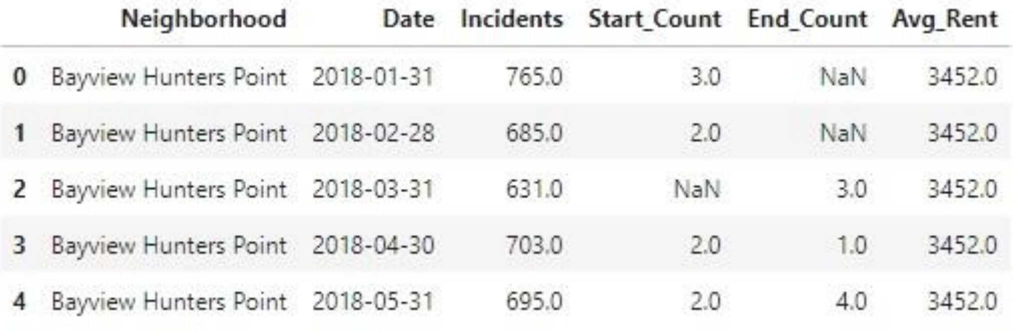


	Neighborhood	Date	Incidents	Start_Count	End_Count
0	Bayview Hunters Point	2018-01-31	765.0	3.0	NaN
1	Bayview Hunters Point	2018-02-28	685.0	2.0	NaN
2	Bayview Hunters Point	2018-03-31	631.0	NaN	3.0
3	Bayview Hunters Point	2018-04-30	703.0	2.0	1.0
4	Bayview Hunters Point	2018-05-31	695.0	2.0	4.0

	Incidents	Start_Count	End_Count
count	656.000000	3379.000000	663.000000
mean	281.817073	1.874815	1.805430
std	329.627095	1.936381	1.299489
min	2.000000	1.000000	1.000000
25%	109.750000	1.000000	1.000000
50%	175.000000	1.000000	1.000000
75%	295.000000	2.000000	2.000000
max	1655.000000	58.000000	12.000000

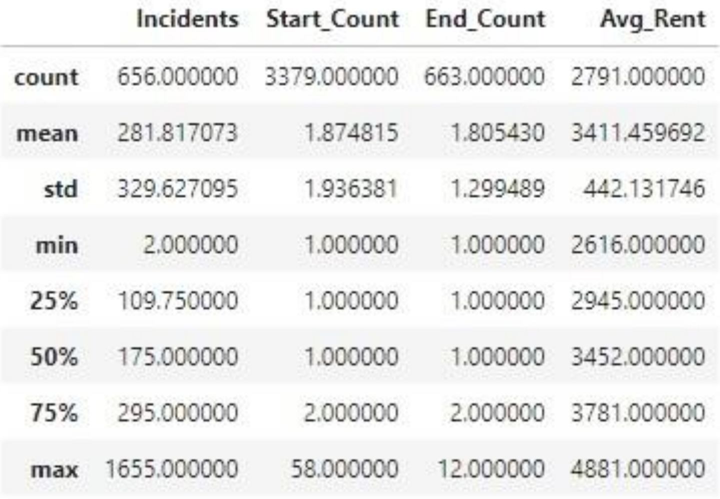
Figure 3-7 Left: df2 table head. Right: df2 statistics.

Then we merged df2 with san Francisco rent table to a new dataframe, df3, with 3883 rows and 6 features.



	Neighborhood	Date	Incidents	Start_Count	End_Count	Avg_Rent
0	Bayview Hunters Point	2018-01-31	765.0	3.0	NaN	3452.0
1	Bayview Hunters Point	2018-02-28	685.0	2.0	NaN	3452.0
2	Bayview Hunters Point	2018-03-31	631.0	NaN	3.0	3452.0
3	Bayview Hunters Point	2018-04-30	703.0	2.0	1.0	3452.0
4	Bayview Hunters Point	2018-05-31	695.0	2.0	4.0	3452.0

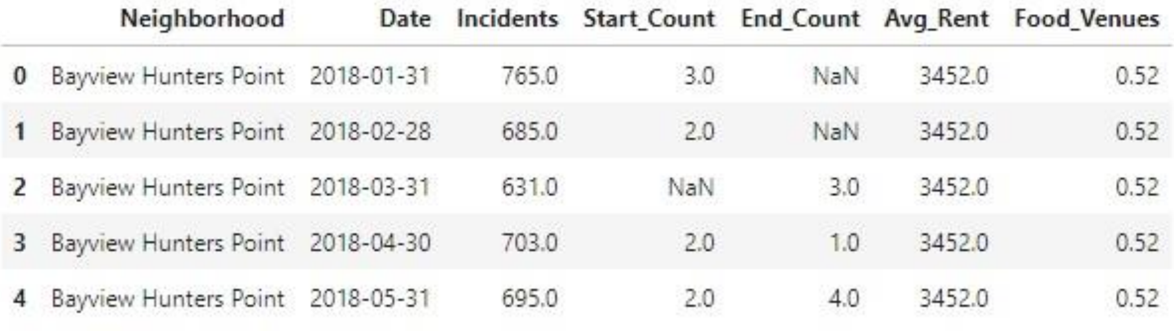
Figure 3-8 Df3 table head.



	Incidents	Start_Count	End_Count	Avg_Rent
count	656.000000	3379.000000	663.000000	2791.000000
mean	281.817073	1.874815	1.805430	3411.459692
std	329.627095	1.936381	1.299489	442.131746
min	2.000000	1.000000	1.000000	2616.000000
25%	109.750000	1.000000	1.000000	2945.000000
50%	175.000000	1.000000	1.000000	3452.000000
75%	295.000000	2.000000	2.000000	3781.000000
max	1655.000000	58.000000	12.000000	4881.000000

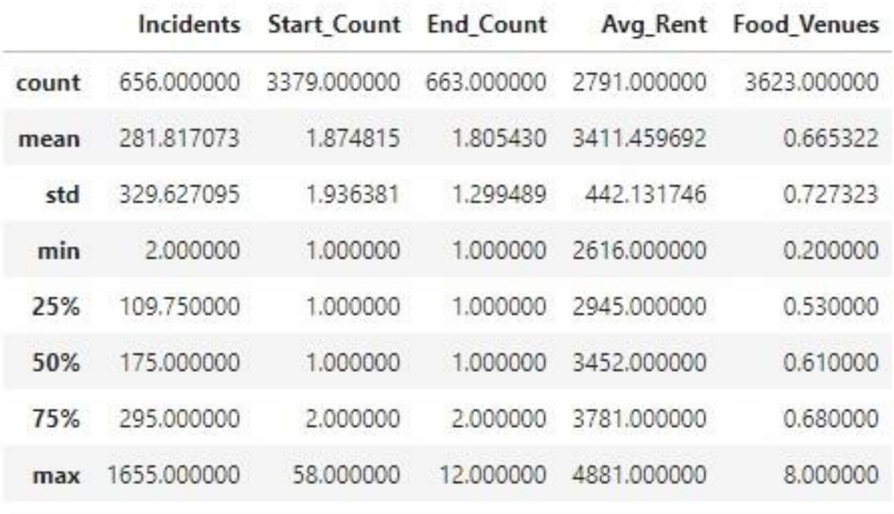
Figure 3-9 Df3 statistics.

And finally, df4 is the result of merging df3 with table of proportion of venues related to food by neighborhood.



	Neighborhood	Date	Incidents	Start_Count	End_Count	Avg_Rent	Food_Venues
0	Bayview Hunters Point	2018-01-31	765.0	3.0	NaN	3452.0	0.52
1	Bayview Hunters Point	2018-02-28	685.0	2.0	NaN	3452.0	0.52
2	Bayview Hunters Point	2018-03-31	631.0	NaN	3.0	3452.0	0.52
3	Bayview Hunters Point	2018-04-30	703.0	2.0	1.0	3452.0	0.52
4	Bayview Hunters Point	2018-05-31	695.0	2.0	4.0	3452.0	0.52

Figure 3-10 Df4 table head



	Incidents	Start_Count	End_Count	Avg_Rent	Food_Venues
count	656.000000	3379.000000	663.000000	2791.000000	3623.000000
mean	281.817073	1.874815	1.805430	3411.459692	0.665322
std	329.627095	1.936381	1.299489	442.131746	0.727323
min	2.000000	1.000000	1.000000	2616.000000	0.200000
25%	109.750000	1.000000	1.000000	2945.000000	0.530000
50%	175.000000	1.000000	1.000000	3452.000000	0.610000
75%	295.000000	2.000000	2.000000	3781.000000	0.680000
max	1655.000000	58.000000	12.000000	4881.000000	8.000000

Figure 3-11 Df4 statistics.

3.6 Kmeans algorithm

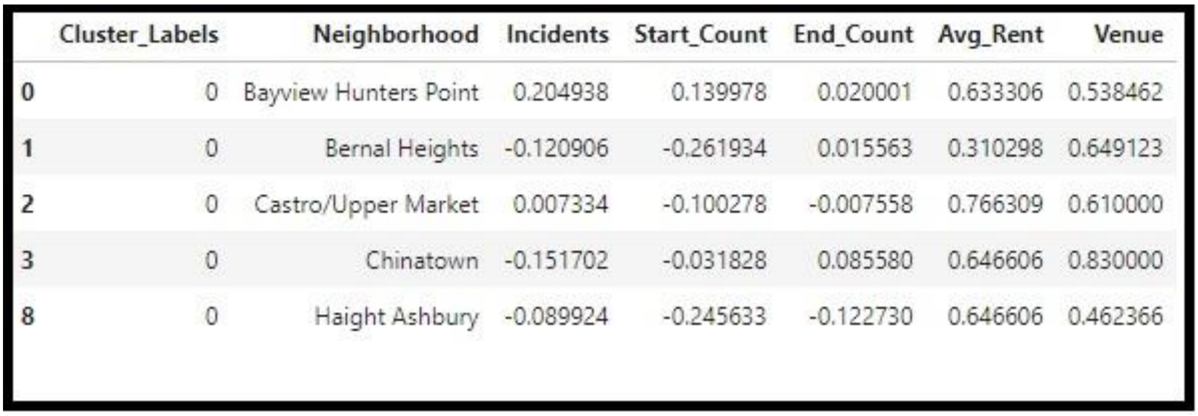
Now that we have all the features in one table, we can start preparing the data for the KNN algorithm of classification. First we need to standardize the features in the dataset, then we grouped by neighborhood and place the mean value of each feature. After this we could use the algorithm and created 4 clusters.

4 Results

We created 4 clusters that are shown on the tables.

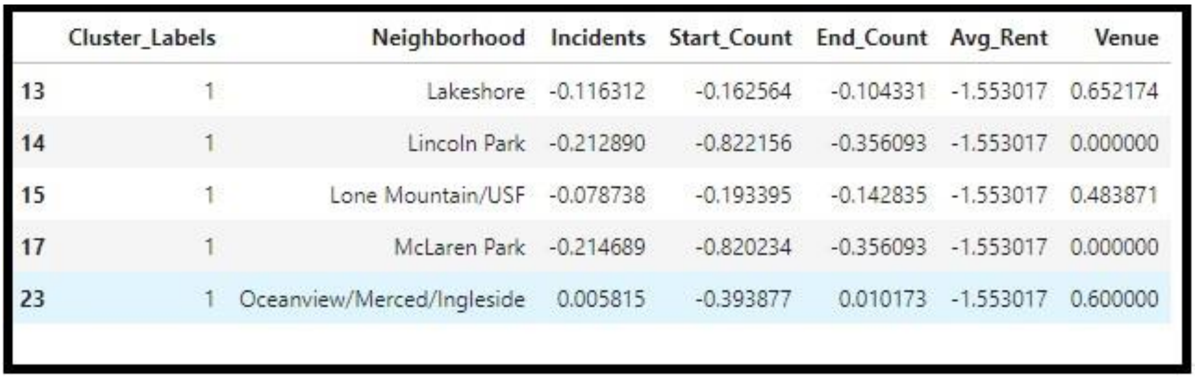
Table 4-1 Number of Neighborhoods by cluster

Cluster	Number of Neighborhoods
0	19
1	9
2	2
3	11



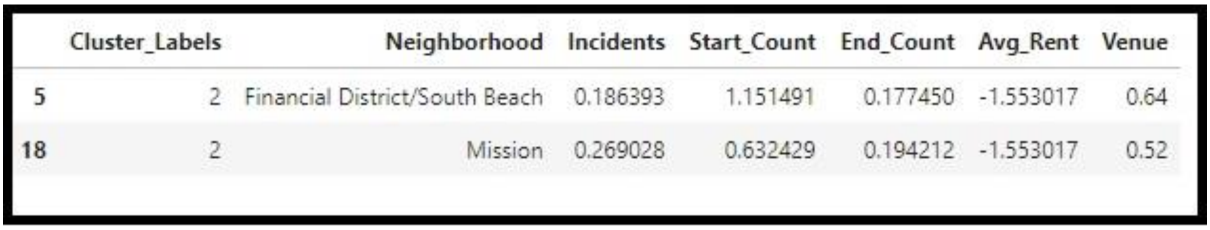
	Cluster_Labels	Neighborhood	Incidents	Start_Count	End_Count	Avg_Rent	Venue
0	0	Bayview Hunters Point	0.204938	0.139978	0.020001	0.633306	0.538462
1	0	Bernal Heights	-0.120906	-0.261934	0.015563	0.310298	0.649123
2	0	Castro/Upper Market	0.007334	-0.100278	-0.007558	0.766309	0.610000
3	0	Chinatown	-0.151702	-0.031828	0.085580	0.646606	0.830000
8	0	Haight Ashbury	-0.089924	-0.245633	-0.122730	0.646606	0.462366

Figure 4-1 Cluster 0 head



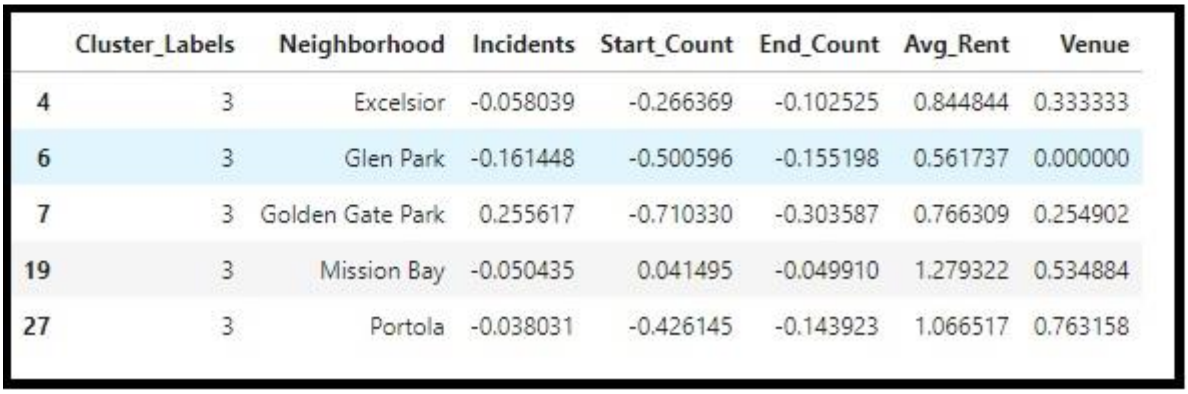
	Cluster_Labels	Neighborhood	Incidents	Start_Count	End_Count	Avg_Rent	Venue
13	1	Lakeshore	-0.116312	-0.162564	-0.104331	-1.553017	0.652174
14	1	Lincoln Park	-0.212890	-0.822156	-0.356093	-1.553017	0.000000
15	1	Lone Mountain/USF	-0.078738	-0.193395	-0.142835	-1.553017	0.483871
17	1	McLaren Park	-0.214689	-0.820234	-0.356093	-1.553017	0.000000
23	1	Oceanview/Merced/Ingleside	0.005815	-0.393877	0.010173	-1.553017	0.600000

Figure 4-2 Cluster 1 head



	Cluster_Labels	Neighborhood	Incidents	Start_Count	End_Count	Avg_Rent	Venue
5	2	Financial District/South Beach	0.186393	1.151491	0.177450	-1.553017	0.64
18	2	Mission	0.269028	0.632429	0.194212	-1.553017	0.52

Figure 4-3 Cluster 2 head



	Cluster_Labels	Neighborhood	Incidents	Start_Count	End_Count	Avg_Rent	Venue
4	3	Excelsior	-0.058039	-0.266369	-0.102525	0.844844	0.333333
6	3	Glen Park	-0.161448	-0.500596	-0.155198	0.561737	0.000000
7	3	Golden Gate Park	0.255617	-0.710330	-0.303587	0.766309	0.254902
19	3	Mission Bay	-0.050435	0.041495	-0.049910	1.279322	0.534884
27	3	Portola	-0.038031	-0.426145	-0.143923	1.066517	0.763158

Figure 4-4 Cluster 3 head

5 Discussion

Based on the values for each cluster, it is very difficult to evaluate what each cluster means, but we can try to draw some conclusions using the describe method. We could a better model if it were not for the fact that we could not get a monthly based rent for San Francisco, and the venues by date.

6 Conclusions

We classify the city in 4 clusters using Kmeans and these are the groups:

6.1 Cluster 0

Bayview Hunters Point, Bernal Heights, Castro/Upper Market, Chinatown, Haight Ashbury, Hayes Valley, Inner Richmond, Inner Sunset, Japantown, Marina, Nob Hill, Noe Valley, North Beach, Outer Mission, Pacific Heights, Russian Hill, South of Market, Tenderloin and Western Addition are the neighborhoods that form this cluster. This is the best cluster for opening a restaurant, the mean values for the interest in food is the higher and there are many restaurants opening in this area.

6.2 Cluster 1

Lakeshore, Lincoln Park, Lone Mountain/USF, McLaren Park, Oceanview/Merced/Ingleside, Outer Richmond, Seacliff, Sunset/Parkside and West of Twin Peaks form cluster 1. This is the second recommendation, but here the interest is lower and there are more closing restaurants.

6.3 Cluster 2

Financial District/South Beach and Mission are the two neighborhoods of this cluster. These are not good recommendation for a restaurant.

6.4 Cluster 3

Excelsior, Glen Park, Golden Gate Park, Mission Bay, Portola, Potrero Hill, Presidio, Presidio Heights, Treasure Island, Twin Peaks and Visitacion Valley are part of the last cluster. This is the worst group to open a restaurant.