SAN FRANCISCO CLASSIFICATION

Jaime Landazuri

INTRODUCTION

- A real estate company is developing an app which include also a segment to attract the attention of clients that want to open a restaurant.
- There is a lack in information for them, therefore creating a classification of best neighborhood where to start a new business will help to improve this situation.

DATA

- This project uses data from 4 sources:
  - San Francisco Police Department
  - Foursquare
  - RentCafé
  - DataSF's.

DATA

- 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, and because they provide a table with the average rent price for each neighborhood, we will use their data.
- Foursquare has data related to 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.

DATA CLEANING

- San Francisco Police Department provide a dataset of all the daily incidents by location.
- 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.

DATA CLEANING

- 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.

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.

FEATURE SELECTION

- For the business count for neighborhood we got a dataset of 236,905 rows.
- 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.

METHODOLOGY

	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

METHODOLOGY

	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

METHODOLOGY

	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

METHODOLOGY

	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

METHODOLOGY
FINAL TABLE

	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

METHODOLOGY
FINAL TABLE STATS

	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

RESULTS
CLUSTERS 0 AND 1

	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

	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

RESULTS
CLUSTERS 2 AND 3

	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

	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

CONCLUSIONS

- **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.
-

CONCLUSIONS

- **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.

CONCLUSIONS

- **Cluster 2**
- Financial District/South Beach and Mission are the two neighborhoods of this cluster. These are not good recommendation for a restaurant.

CONCLUSIONS

- **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.