CAPSTONE PROJECT: THE BATTLE OF NEIGHBORHOODS

Salvador Roca Guillén

1. INTRODUCTION

a. BACKGROUND

A close friend of us is running a fast food business in Sacramento and, as his company is generating great profit, he has the intention to make his business grow by **expanding to San Francisco**.

b. PROBLEM AND INTEREST

As he knows we have studied a comprehensive Data Science course endorsed by IBM, he is asking us to perform an deep analysis with the objective of **determining the best zones of the city to establish the first restaurant** in the city.

2. DATA ACQUISITION AND CLEANING

a. DATA SOURCES

To perform this analysis, we will need some information:

- **San Francisco geographical data**, which can be found in the link below. With this information, we will be able to identify the different zones of the city and represent the data in a visual manner by generating a Choropleth maps: https://data.sfgov.org/Geographic-Locations-and-Boundaries/San-Francisco-ZIP-Codes/srq6-hmpi
- **Population by ZIP code**; it is interesting to select the zones with the higher population density in order to generate a higher demand. This information can be found in the next webpage: https://www.zip-codes.com/city/ca-san-francisco.asp
- **Data of all different businesses in each zone of San Francisco**. We will filter the data to consider only the type of business that can be a competitive threat for the restaurant. The main source to obtain this information **Foursquare database**.

b. DATA CLEANING

The raw data downloaded from the mentioned sources is clean and coherent enough to continue to feature selection.

c. Feature Selection

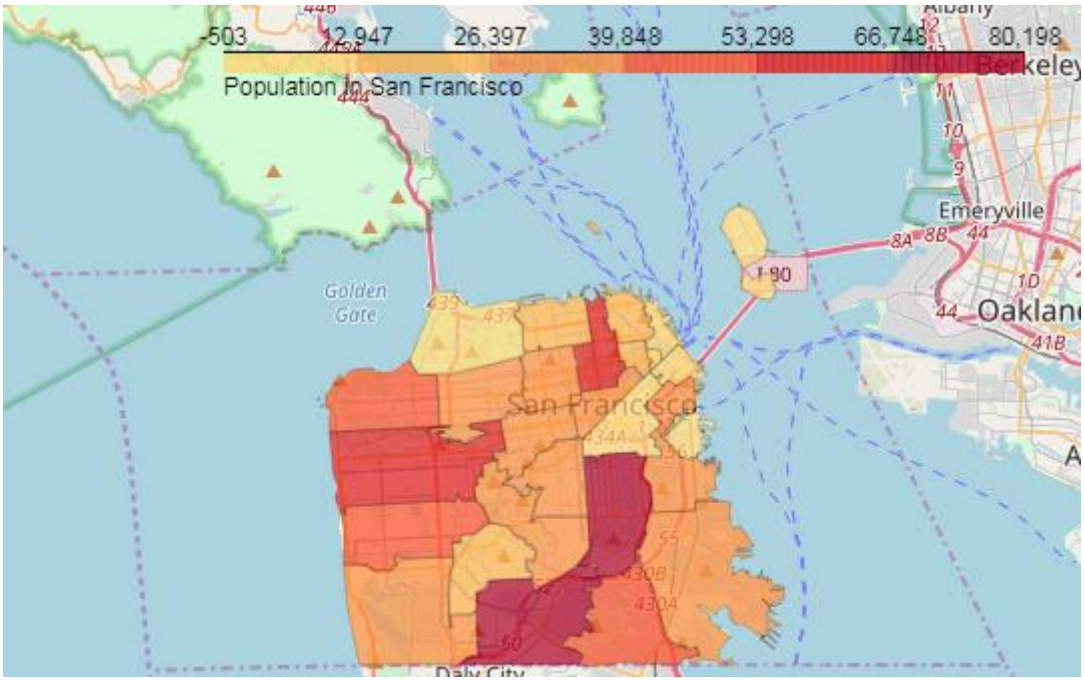
From San Francisco geographical data, we will need the multipolygons associated to each ZIP code in the .json file. This will allow us to generate a Choropleth map representing the population density of each zone.

Population by ZIP code can be easily obtained from the given source in .csv format, and for our analysis we can get rid of all columns different than Population and ZIP Code.

Each call to Foursquare database will provide a bunch of information. The features we need to focus on are Venue Name, Venue Category and Venue Postal Code (or ZIP code).

3. Exploratory Data Analysis

With the help of a Choropleth map, we will see the population of each zone:



The ZIP codes with highest Population density are represented in the next table:

ZIP Code	Population
94112	79407
94110	69333
94122	56023
94109	55984
94116	43698

Once we have this overview on population density, it is needed to get the number of restaurants in each ZIP Code. By consulting Foursquare database and applying the proper filters, we will get our general dataframe to start analyzing the data.

	ZIP Code	Restaurants Count	Population
0	94102	24	31176
1	94103	17	2717
2	94104	24	406
3	94105	24	5846
4	94107	21	26599
5	94108	27	13768
6	94109	21	55984
7	94110	23	69333
8	94111	36	3713
9	94112	25	79407
10	94114	17	31124
11	94115	24	33021
12	94116	26	43698
13	94117	18	39169
14	94118	24	38319
15	94121	31	41203
16	94122	35	56023
17	94123	21	23088
18	94124	23	33996
19	94127	23	19289
20	94129	8	3183
21	94130	1	288
22	94131	18	26881
23	94132	8	28129
24	94133	29	26237
25	94134	15	40798
26	94158	15	4792

4. ANALYSIS: K-MEANS CLUSTERING

The selection of the best zones to open a restaurant will be made according to a k-means clustering, which will identify up to 5 different groups (clusters) of zone types. The resulting clusters are presented in tables below:

- **Cluster 1**

	ZIP Code	Restaurants Count	Population	Cluster Labels
0	94102	24	31176	0
4	94107	21	26599	0
10	94114	17	31124	0
17	94123	21	23088	0
19	94127	23	19289	0
22	94131	18	26881	0
23	94132	8	28129	0
24	94133	29	26237	0

- **Cluster 2**

	ZIP Code	Restaurants Count	Population	Cluster Labels
1	94103	17	2717	1
2	94104	24	406	1
3	94105	24	5846	1
5	94108	27	13768	1
8	94111	36	3713	1
20	94129	8	3183	1
21	94130	1	288	1
26	94158	15	4792	1

- **Cluster 3**

	ZIP Code	Restaurants Count	Population	Cluster Labels
6	94109	21	55984	2
16	94122	35	56023	2

- **Cluster 4**

	ZIP Code	Restaurants Count	Population	Cluster Labels
11	94115	24	33021	3
12	94116	26	43698	3
13	94117	18	39169	3
14	94118	24	38319	3
15	94121	31	41203	3
18	94124	23	33996	3
25	94134	15	40798	3

- **Cluster 5**

	ZIP Code	Restaurants Count	Population	Cluster Labels
7	94110	23	69333	4
9	94112	25	79407	4

5. Result Summary and Conclusion

a. Summary

To simplify the decision-making process, we will summarize data from each cluster by using key indicators. The key indicator in our case will be a ratio comparing the mean population vs the mean number of restaurants of each cluster.

Cluster	Restaurants Count mean	Population mean	Ratio
1	20.125	26565.375000	1320.018634
2	19.000	4339.125000	228.375000
3	28.000	56003.500000	2000.125000
4	23.000	38600.571429	1678.285714
5	24.000	74370.000000	3098.750000

b. Discussion and Conclusion

The best option for this business case is to open the first restaurant in **Cluster 5**, as it presents the **highest *Population vs Restaurants ratio***. This cluster presents, by far, the highest population density, and it is not the cluster with the highest number of restaurants, which justifies the calculated ratio and the **decision to open the restaurant in either ZIP Code 94110 or 94112**.