CAPSTONE PROJECT- THE BATTLE OF NEIGHBORHOODS

Fabiana Rossi – January 2021

INTRODUCTION

During the last couple of decades, California has been a leader state in air pollution, with high ozone levels. Most of the cities in California state report days of unhealthy air quality [1], and more than ninety percent of Californians live in counties with unhealthy air [2]. Though a lot of effort has been put into addressing this problem and some advances have been achieved, air pollution represents an unmet public need which dates back to 1943 [3]. Poor air quality has a direct influence in health, having a myriad of effects influencing the nervous system [4], lungs [5], cardiovascular system [3, 6] (among others), and increases the risk of all-cause mortality [7]. Consequently, more efficient strategies should be developed to mitigate air pollution.

The objective of this project was to analyze California cities contamination and determine whether the percentage of parks and green areas (relative to all venues in each city) had a relation with the air quality data. In order to perform a more complete analysis, cities will be clustered according to air quality index information and the relation between percentage of parks/green areas and air pollution will be studied within each cluster.

Since it has been demonstrated that population density has both benefits and costs for air quality [8], agglomeration will also be considered in the analysis.

For this purpose, air quality data will be retrieved from the United States Environmental Protection Agency (US EPA) [1], and California demographic data will be obtained from Wikipedia [9, 10] and Kaggle [11]. After cleaning and preparing the data, cities will be clustered according to their air quality index, and correlation analysis will be performed within each cluster.

It is expected that the results derived from this study could be of general public interest. If a positive correlation existed between the number of parks and air quality, irrespective of population density in each cluster, then it could be speculated that parks could mitigate air contamination and therefore city planning should consider the development of parks and green areas in the future.

METHODOLOGY

The objective of this project was to analyze California cities contamination and determine whether the percentage of parks (relative to all venues in each city) within each cluster had a relation with the air quality data.

As mentioned before, this information came from the United States Environmental Protection Agency (US EPA), Wikipedia and Kaggle.

After cleaning and preparing the data, cities were clustered according to their air quality index, and correlation analysis was performed within each cluster.

Data collection consisted in five parts, as follows:

- part 1= Californian cities population density data
- part 2= Californian cities latitude and longitude data
- part 3= air quality data
- part 4= final dataframe and preliminary analysis
- part 5= California cities venues data

1- California cities population density data

California cities population data was imported from Wikipedia [https://en.wikipedia.org/wiki/List_of_largest_California_cities_by_population]. The data was collected as a dataframe.

California cities land area data was then imported from Wikipedia [https://en.wikipedia.org/wiki/List_of_largest_California_cities_by_land_area] in order to calculate population *density*, since population alone might not represent faithfully cities size and transit (which might be more related to contamination). The data was collected and obtained as a dataframe.

Data from each dataframe was cleaned and prepared, and then dataframes were merged (*inner join using 'City')*. The resulting dataframe was used to calculate population density (population per land area).

2- Latitude and Longitude data

In order to draw each city in a California map and to get cities venues, latitude and longitude data was downloaded from https://www.kaggle.com/camnugent/california-housing-feature-engineering?select=cal_cities_lat_long.csv. Then, data was cleaned and merged to the population density dataframe in a new dataframe called "new_df" which contained city names, latitude and longitude data, and population density, as shown in Figure 1.

	City	Latitude	Longitude	Population density
0	Bakersfield	35.35	-119.03	2565.584719
1	Fremont	37.52	-122.00	3111.096774
2	Fresno	36.78	-119.79	4750.455764
3	Hesperia	34.40	-117.31	1309.849521
4	Irvine	33.68	-117.77	4347.972769

Figure 1: Each city name, latitude, longitude and population density were condensed in the "new_df" dataframe. The figure shows the five first rows of the "new_df" dataframe.

3- Air Quality Index Data

California Air Quality data (AQI) was downloaded from United States Environmental Protection Agency (US EPA): https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report.

The data corresponds to the number of days where air quality was considered Good(G), Moderate(M), Unhealthy for sensitive groups (USG), unhealthy(U) or Very unhealthy (VU) according to US EPA standards during 2019. Data was cleaned and prepared as a dataframe in which the number of days with different air quality was relativized for the number of days analyzed for each city.

4- Final data frame

Air Quality data was merged to cities dataframe (that contained population density and Lat/long data), as shown in Figure 2:

	City	Latitude	Longitude	Population density	G%	M%	USG%	U%	VU%
0	Bakersfield	35.35	-119.03	2565.584719	0.326027	0.413699	0.238356	0.016438	0.005479
1	Fresno	36.78	-119.79	4750.455764	0.380822	0.438356	0.175342	0.005479	0.000000
2	Los Angeles	34.11	-118.41	8490.667805	0.180822	0.569863	0.169863	0.076712	0.002740
3	Oakland	37.77	-122.22	7760.412186	0.695890	0.279452	0.024658	0.000000	0.000000
4	Redding	40.57	-122.37	1553.523490	0.854396	0.145604	0.000000	0.000000	0.000000

Figure 2: Data frame containing each city name, latitude, longitude, population density and percentage of good (G), Moderate(M), Unhealthy for sensitive groups (USG), unhealthy(U) or Very unhealthy (VU) days, according to US EPA standards during 2019. The figure shows the five first rows of the dataframe.

As a preliminary analysis prior to clustering, population density was plotted for each category against air quality (Figure 3).

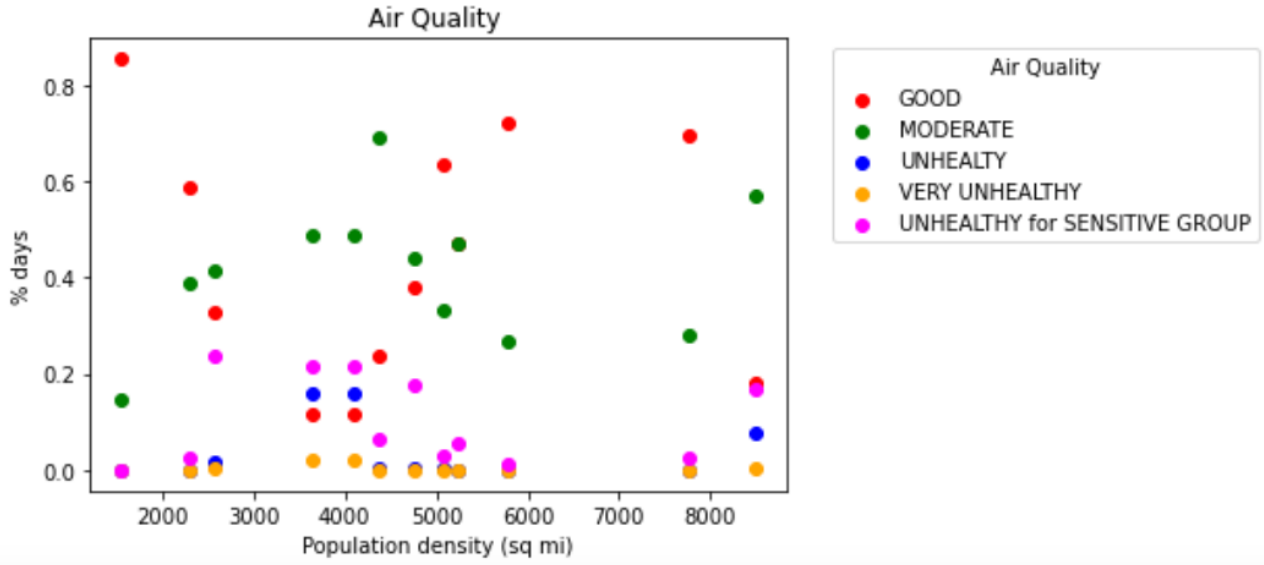

Figure 3: Scatter plot of the percentage of days of a given air quality category against population density (sq mi).

Unexpectedly, no correlation between % of good days (or the other categories) and population density was observed. Consequently, population density was excluded from the clustering analysis.

5- California venues

Next, California state map was uploaded, and cities were superimposed in the map. Using Forthsquare, venues were obtained for each of the analyzed cities.

RESULTS

Before analyzing the data, cities in California state were clustered according to their air quality characteristics using k-means clustering. Firstly 'k' selection was performed for k-means clustering, using Elbow method, which indicated that k=2 should be used for the analysis (Figure 4).

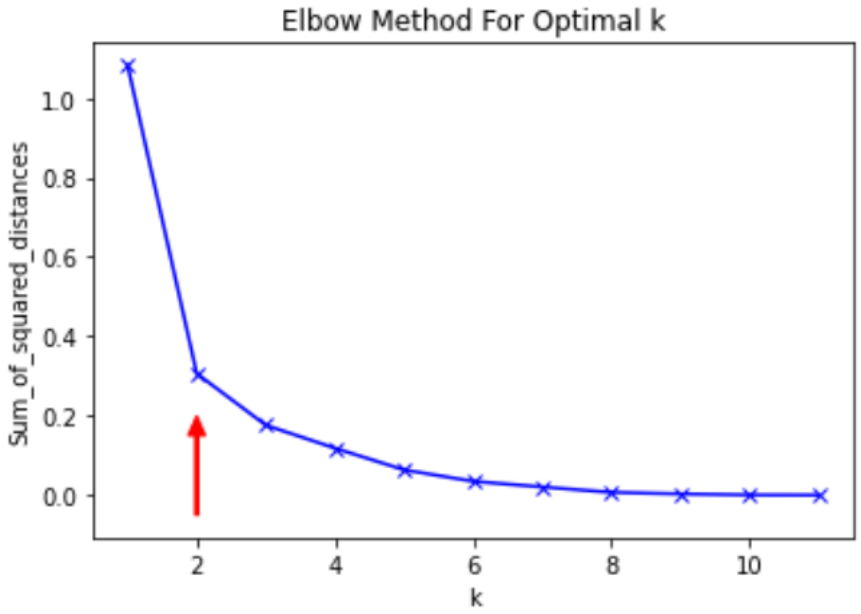

Figure 4: Determination of optimum k value for k-means cluster analysis, using Elbow Method.

Next, data was clustered and cities from each cluster were superimposed on a California map (Figure 5).

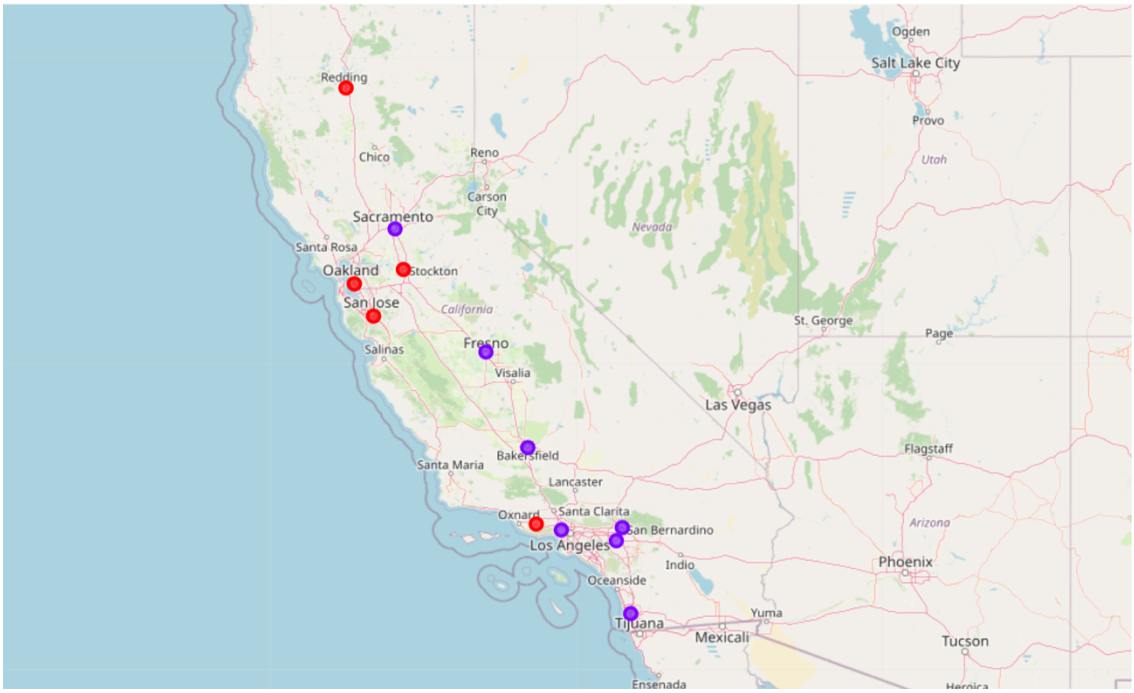

Figure 5: California state map with superimposed cities of each cluster (cluster 0 = RED, cluster 1= VIOLET).

Cluster characterization

The mean population density in each cluster was analyzed using boxplots. Results indicated that this variable mean value is similar between cities from each cluster (Figure 6, left).

Next, each city latitude was plotted in a boxplot, where each category corresponded to a different cluster value. Despite differences were not statistically different (*t-student test*), the mean latitude for cities in cluster 0 was higher than the mean latitude from cluster 1 (Figure 6, right).

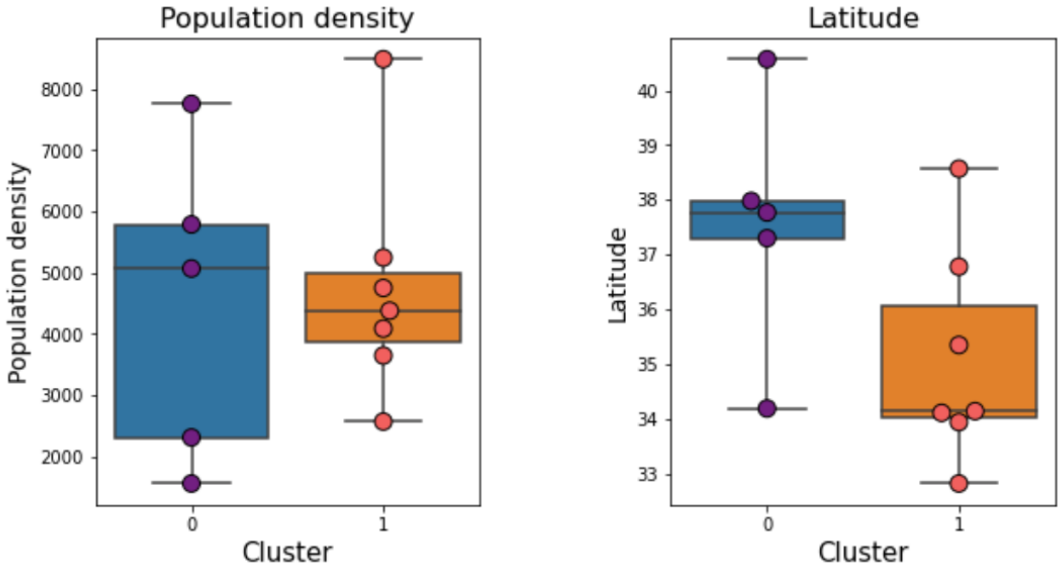

Figure 6: Boxplot graphs for population density distribution (left) and latitude (right) in each cluster.

Next, air quality was analyzed within each cluster, and color-coded markers were used in order to visualize if there was a relation between air quality and population density (Figure 7).

Air quality data included the percentage of days with good(G), moderate(M), Unhealthy for sensitive groups (USG), unhealthy(U) or very unhealthy (VU) categories of air quality.

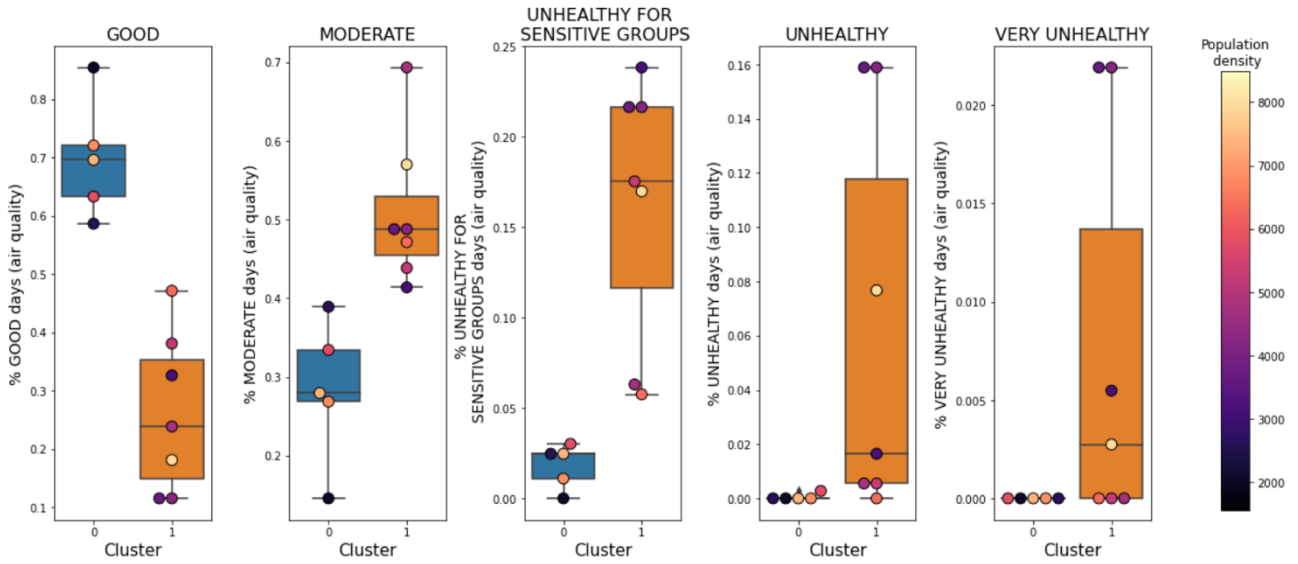

Figure 7: Boxplot graphs for percentage of days in the different air quality categories in each cluster. Values are color-coloded ranging from high population density (in yellow) and low population density (in black).

These results indicate that cluster "0" has more percentage of GOOD days, and less percentage of MODERATE/UNHEALTHY FOR SENSITIVE GROUPS/UNHEALTHY OR VERY UNHEALTHY days, compared to cluster "1". Overall, cluster "0" seems to have less air pollution than the other cluster.

Color-coded markers were used for each value in the boxplots, where yellow indicated high population density cities and black indicated low population density cities. No clear correlation was observed between air quality and population density in neither of the clusters, for all the air quality categories that were analyzed.

In order to compare venues characteristics in each cluster, higher-order categories were created. As an example: 'Mexican Restaurant' - 'Sushi Restaurant'- 'Chinese Restaurant'- 'Moroccan Restaurant', etc. were included in a new 'RESTAURANT' category. As a consequence, venues data from each city was recategorized into 10 categories (instead of 80 different, which were the ones originally present in our Fourthsquare data). The percentage of these new higher-order (or 'general') categories (relative to all the venues in a given city) was analyzed for each cluster. Data is shown in Figure 8.

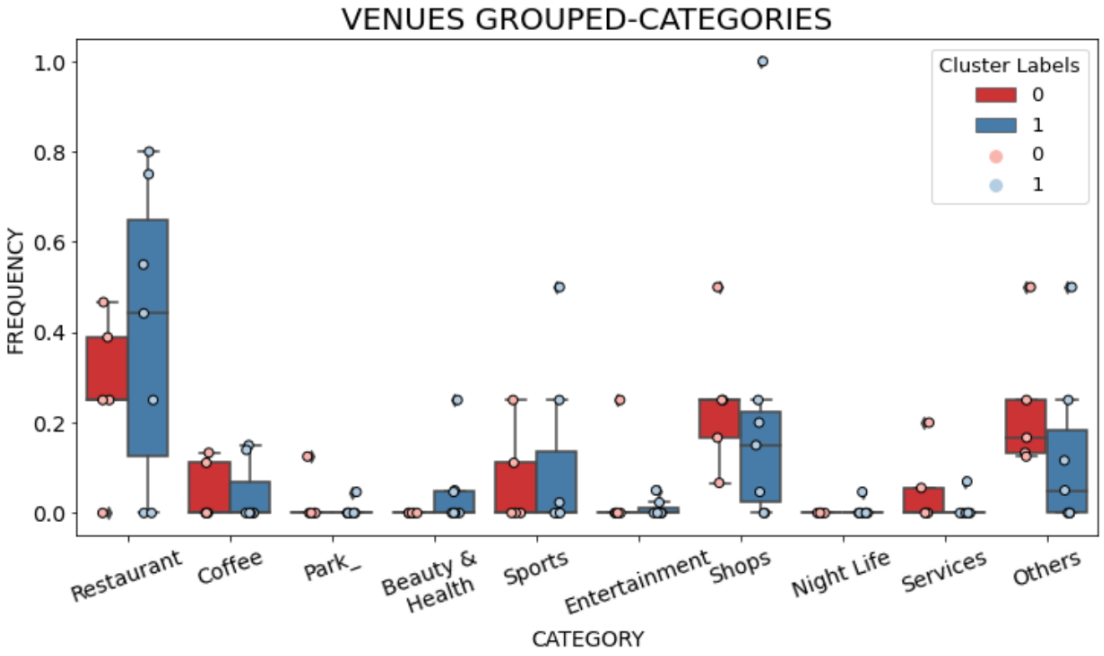

Figure 8: Venues grouped-categories ('general'/higher-order categories) in each cluster.

The comparison between each cluster indicated that both clusters' venues percentages were similar, except for 'SHOPS', as cluster 0 had a higher proportion of venues in this category than cluster 1. More data should be considered if this analysis were repeated for statistical analysis.

Number of parks correlation to Air Quality data, within each cluster

Finally, since the idea behind this analysis was to determine if there was a correlation between either of the clusters air quality data and the number of parks, 'PARKS' high-order category was analyzed. Unfortunately, it seems that the parks data obtained from Foursquare (at least with this user version) does not thoroughly explore "real" data, as 0 parks were obtained for all the cities except for Sacramento (there was also a venue that belonged to 'park' in cluster 1, but this was a 'Skate park' and was therefore dismissed). In this respect, no further analysis could be performed.

DISCUSSION and CONCLUSION

Our analysis allowed the identification of 2 clusters which differed mainly in the extent of air pollution. Cities in cluster 0, which are mainly located at higher latitudes than cities in cluster1, have a higher proportion of 'Good' air quality days. Population density was a factor that did not seem to affect air quality, and venues characteristics were similar for both clusters.

Unfortunately, the number (or proportion) of parks in each cluster and its relation to air quality could not be studied in the present analysis due to insufficient data. It would be worthwhile to perform a more detailed analysis in the future with a larger set of Californian cities.

REFERENCES

[1] https://www.epa.gov/outdoor-air-quality-data/air-quality-index-report
[2] California State University (CSU) Chancellor's Office. (2017, June 19). California named state with the worst air quality (again). ScienceDaily. Retrieved January 4, 2021 from www.sciencedaily.com/releases/2017/06/170619092749.htm
[3] https://en.wikipedia.org/wiki/Pollution_in_California
[4] Genc S, Zadeoglulari Z, Fuss SH, Genc K. The adverse effects of air pollution on the nervous system. J Toxicol. 2012;2012:782462. doi:10.1155/2012/782462
[5] Holgate ST. 'Every breath we take: the lifelong impact of air pollution' - a call for action. Clin Med (Lond). 2017 Feb;17(1):8-12. doi: 10.7861/clinmedicine.17-1-8. PMID: 28148571; PMCID: PMC6297602.
[6] https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm
[7] Zhao S, Liu S, Hou X, Sun Y, Beazley R. Air pollution and cause-specific mortality: A comparative study of urban and rural areas in China. Chemosphere. 2021 Jan;262:127884. doi: 10.1016/j.chemosphere.2020.127884. Epub 2020 Aug 11. PMID: 33182102.
[8] Borck, Rainald; Schrauth, Philipp (2020). Population density and urban air quality. Regional Science and Urban Economics, (), 103596–. doi:10.1016/j.regsciurbeco.2020.103596
[9] https://en.wikipedia.org/wiki/List_of_largest_California_cities_by_population
[10] https://en.wikipedia.org/wiki/List_of_largest_California_cities_by_land_area
[11] www.kaggle.com/camnugent/california-housing-feature-engineering?select=cal_cities_lat_long.csv