IBM Capstone Project for Data Sciences Exploring the desert: Analysis of venues density in Tucson - AZ

Maria Cecília Rodrigues do Prado

1.Introduction - Tucson, AZ, EUA

- Scattered city
- Desert weather

Data

2.1.Neighborhood names

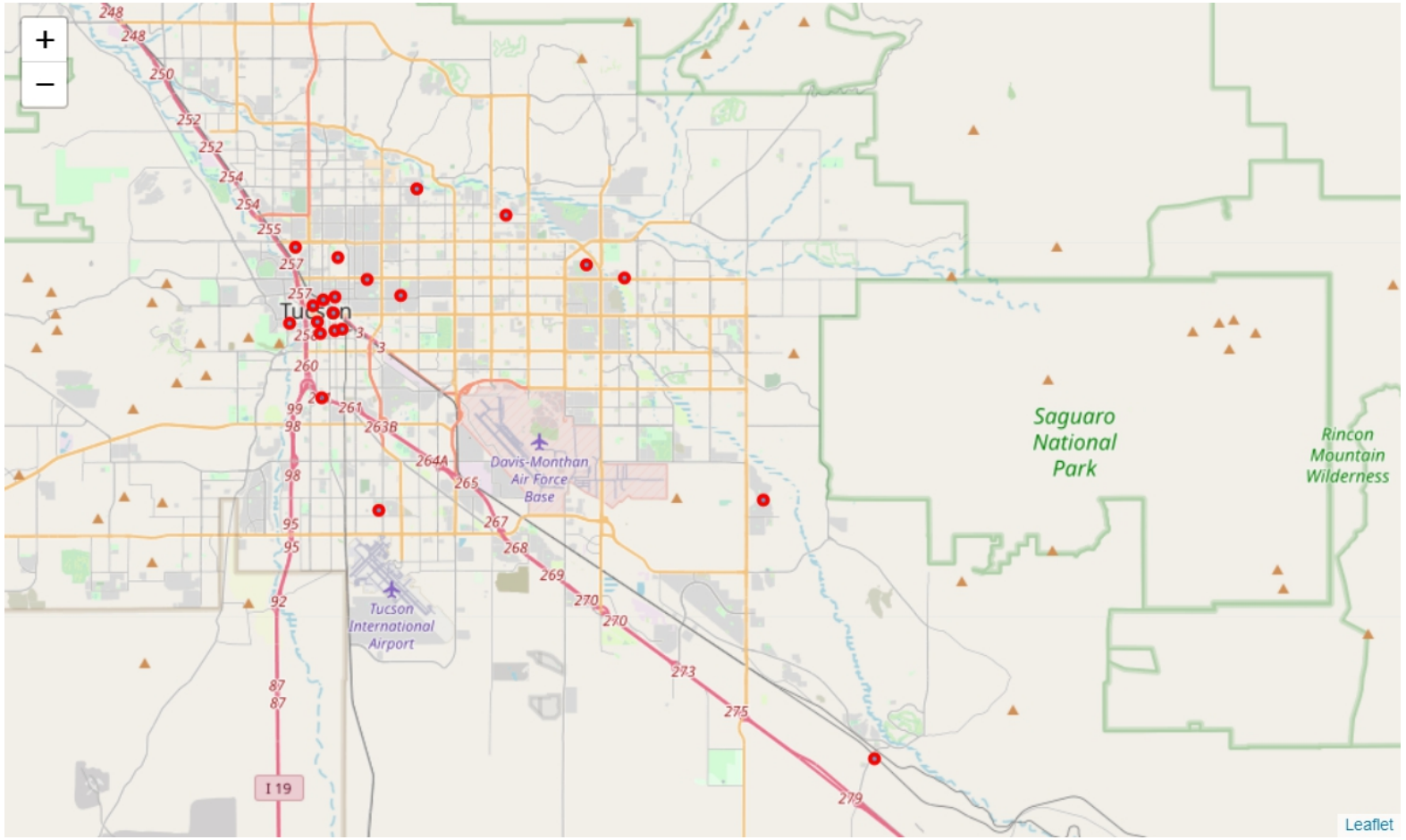
The website city-data.com provides a range of information for US neighborhoods, including race and of inhabitants, household income, house values, education, means of transportation and many other information. From this source, we obtained the names of Tucson’s neighborhoods.

2.2.The Foursquare API

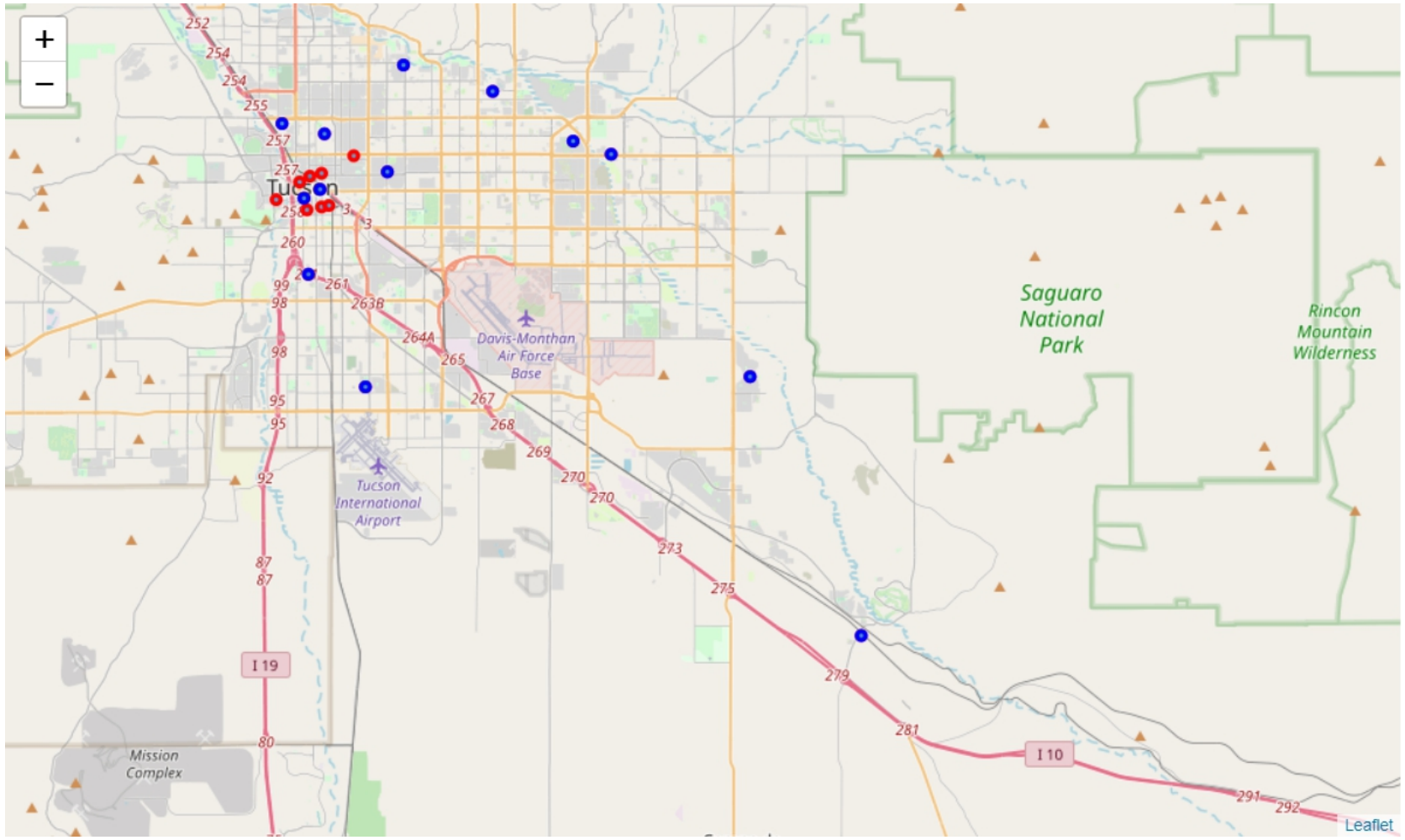
The Foursquare API provides information about millions of venues and users from all over the world. With a developer’s account, it is possible to search venues based on its location, category, name and many other features. In this study, we are interested in venues of all categories around specific geographic coordinates - the coordinates of our neighborhoods.

Methodology

Webscrap to get neighborhood names

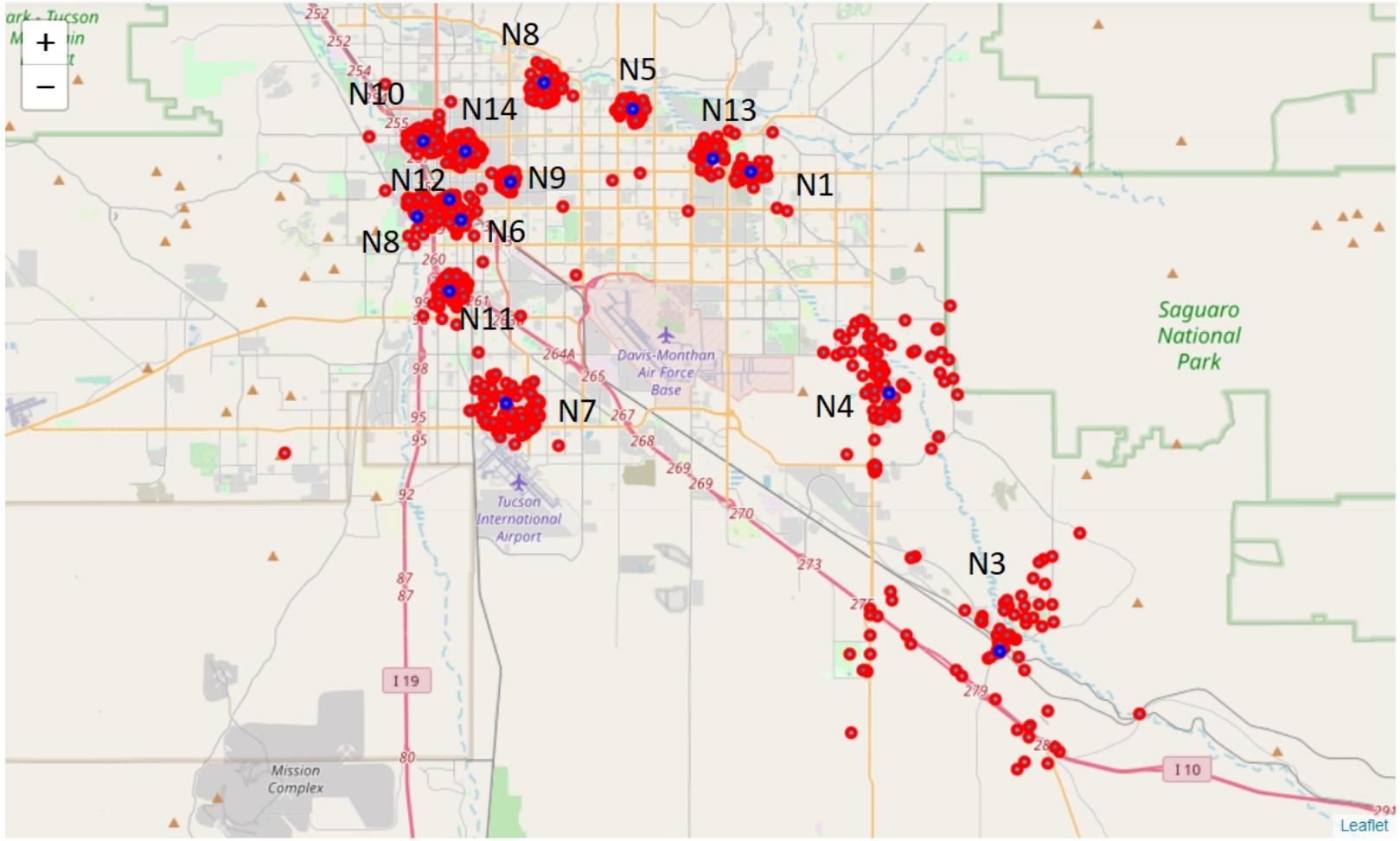


Clustering the neighborhoods



- We wanted the points more evenly distributed, so we chose the number of clusters instead of performing any metrics test.

Venues from Foursquare API



4 different searches:

- radius = 3000m

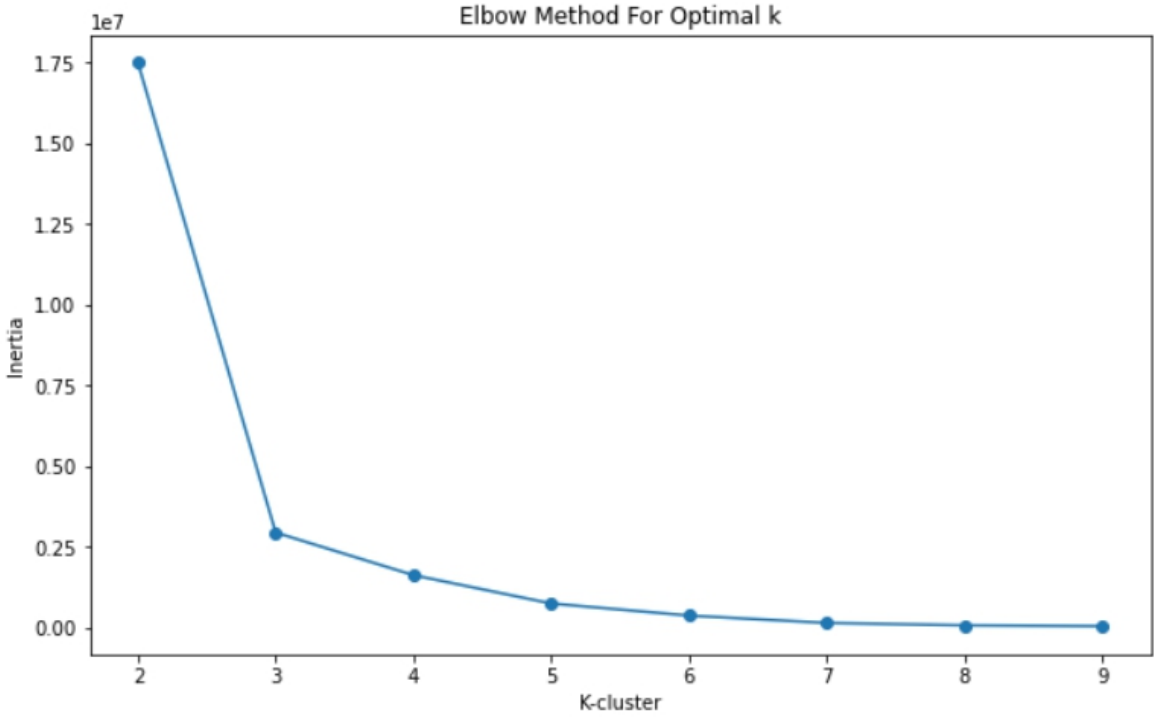
- radius = 1000m

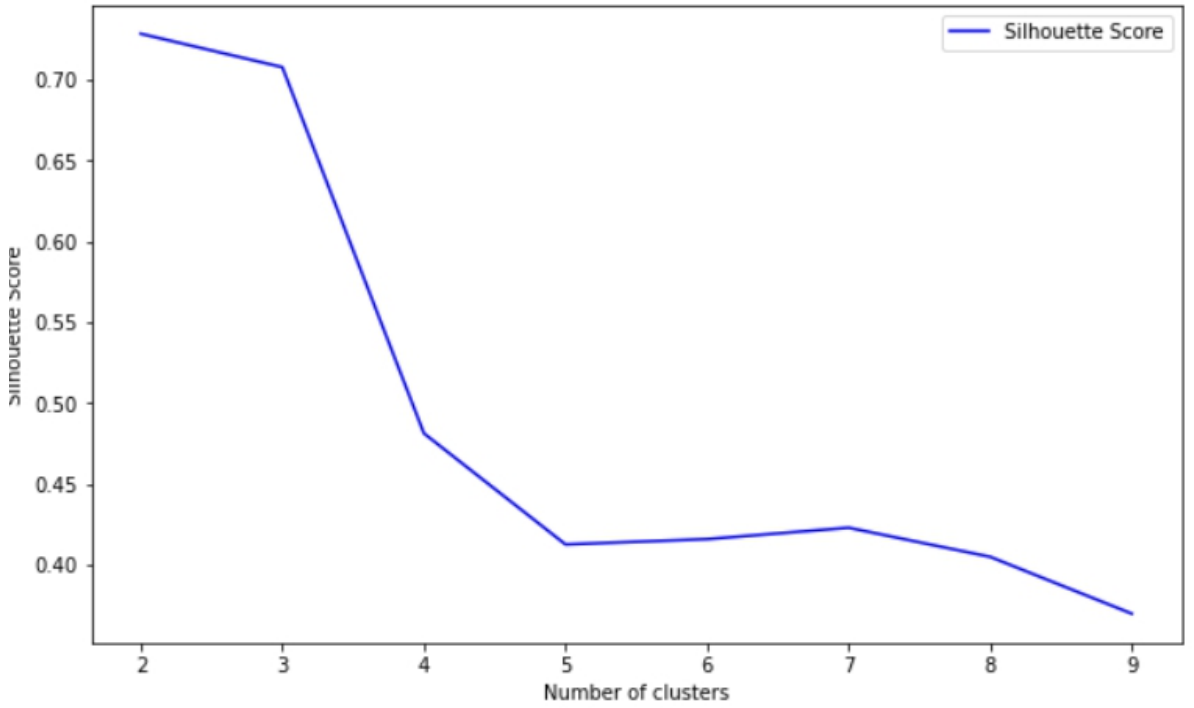
- radius = 500m

- radius = 100m

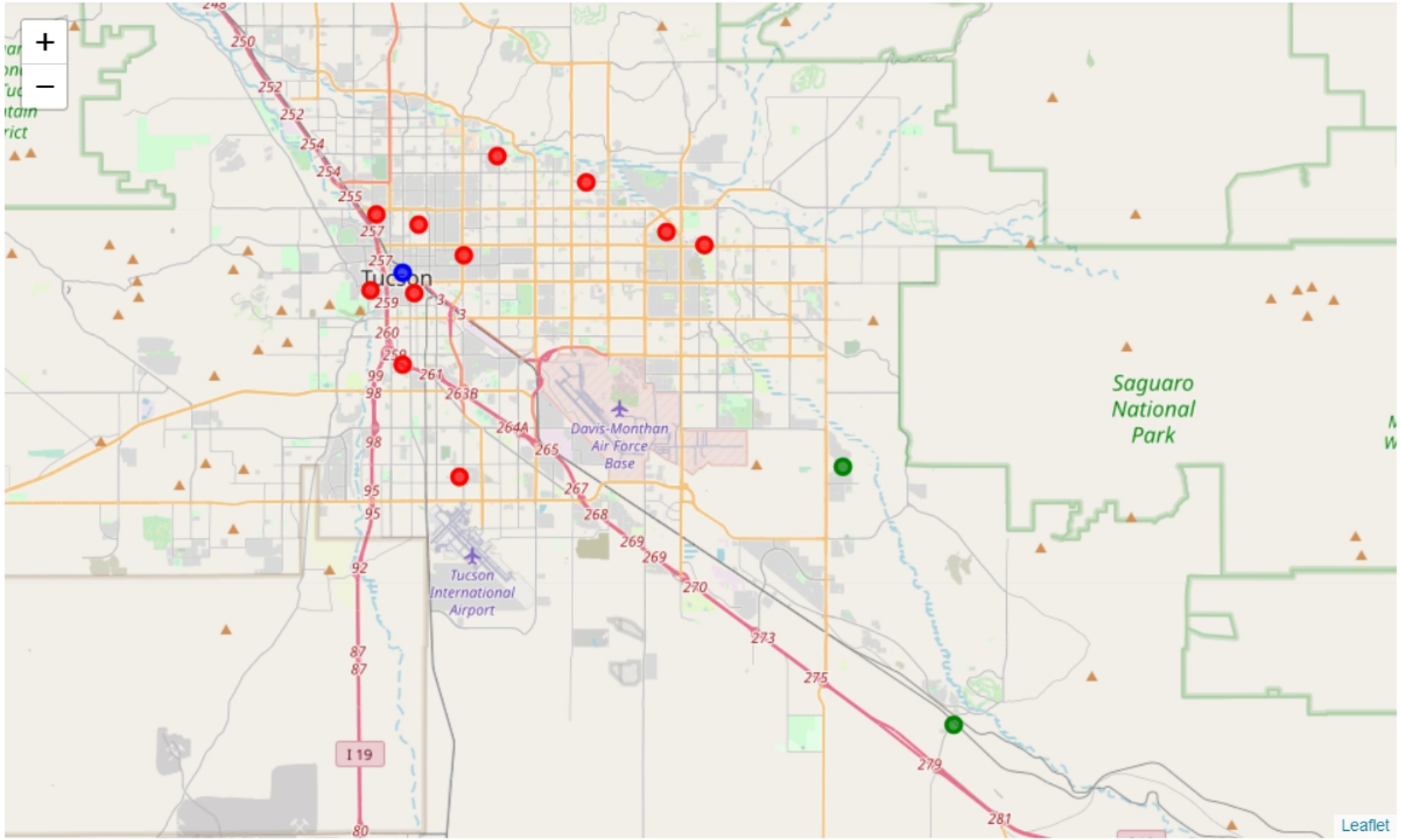
Clustering the results

Optimal value of K = 3

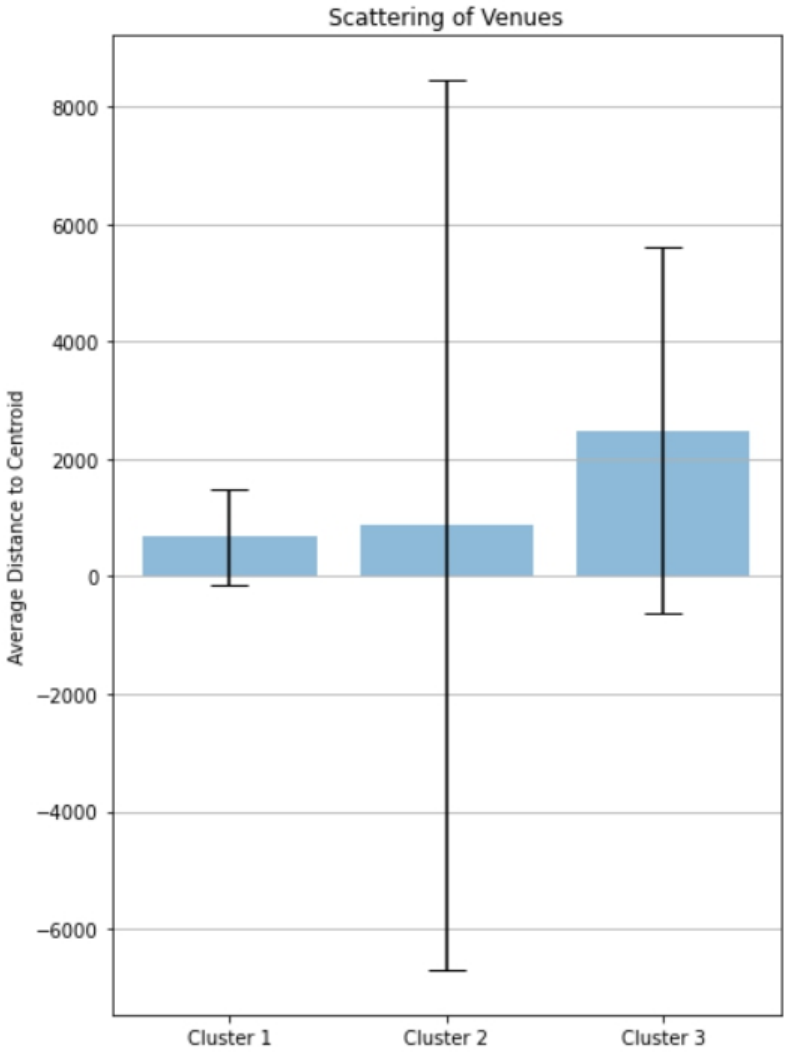




Final Clusters



Final Clusters



- Cluster 1 - Higher density
- Cluster 2 - Intermediate average distance and outliers
- Cluster 3 - Lower density

Conclusion

We identified how venue density is geographically spread over the city. There were three main cluster conformations: low mean and low standard deviation, low mean and very high standard deviation and high mean and high standard deviation.

The first configuration describes venues that are near each other and in a homogeneous manner. The second describes venues that are mostly close to each other, but some relevant venues located very distant from the center. The third configuration describe venues that are spread apart.

Further Analyzis

For further analysis, we recommend to identify and treat the outlier venues, deciding carefully wether these venues should be included or not based on their relevancy. Moreover, we recommend to include venues cathegory to further understand the geographical distribution of venues within the city and with that, provide information support for better decisions regarding this issue.