Foursquare Project

Living in Cincy

Leila Fabiola Ferreira

Summary

Introduction

In this project, we will explore data from Cincinnati (city belonging to the state of Ohio in the United States of America) obtained from Foursquare API. These steps could be applied on any other city, as long as data about it is available to explore.

> **"Cincinnati is a beautiful city; cheerful, thriving, and animated. I have not often seen a place that commends itself so favorably and pleasantly to a stranger at the first glance as this does."**
> **— Charles Dickens, American Notes for General Circulation**

Business Understanding

Business Understanding

- When we are looking for a new place to live, either for intentional change or for reasons of necessity and even for a trip for tourism or business, it is interesting to find a place that suits our lifestyle and especially that this place allows us to have easy access to certain types of places of interest, often with a simple walk. Therefore, the definition of venues in this project were obtained with a maximum distance of 1 kilometer from the central point of each neighborhood.

Business Understanding

PROBLEM STATEMENT

- The main question is: Based on this data, is it possible to group neighborhoods using the categories of their commercial establishments or places in order to visualize which ones would be more suitable for living or traveling?

Business Understanding

OBJECTIVE

- The data obtained from the Foursquare API about the neighborhoods are not labeled, so the clustering technique was used to see if it is possible to identify patterns in each group founded and thus describe them. After that, we can analyze which group is most suitable to live in and then look for a house or apartment to rent in this group of neighborhoods.

Data Requirements

Steps

Get the list of neighborhoods

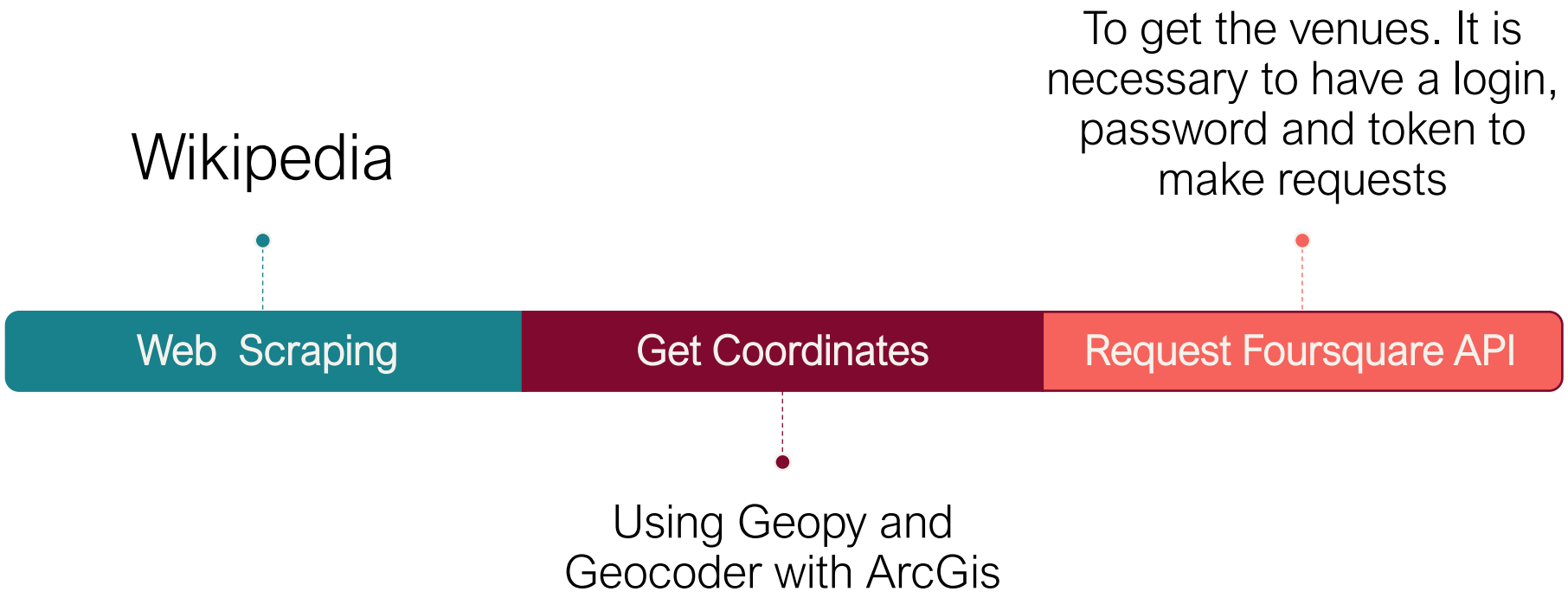
Get the coordinates of each neighborhood

Request Foursquares API

Manipulate the data

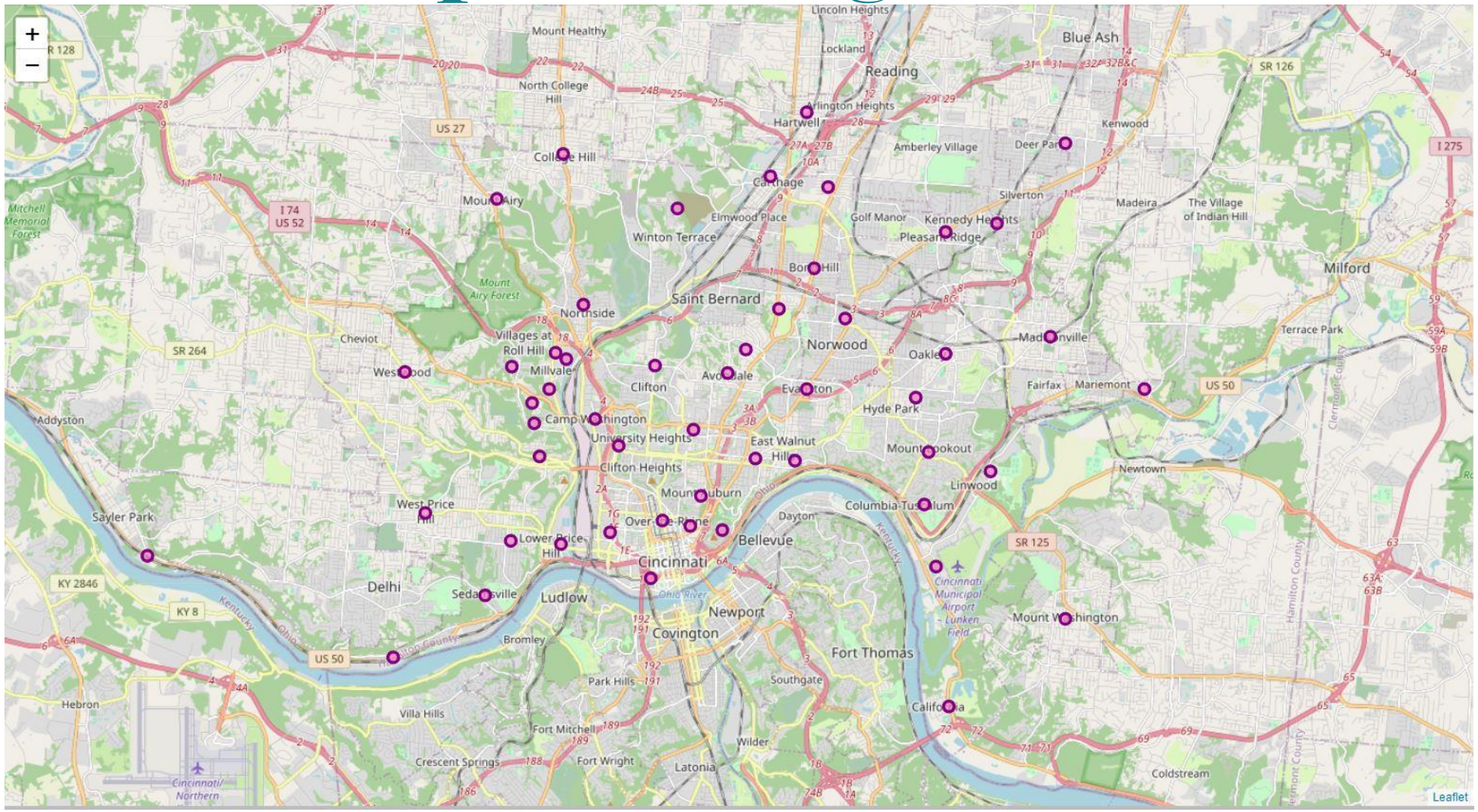
Data Collection

Data Collection Timeline

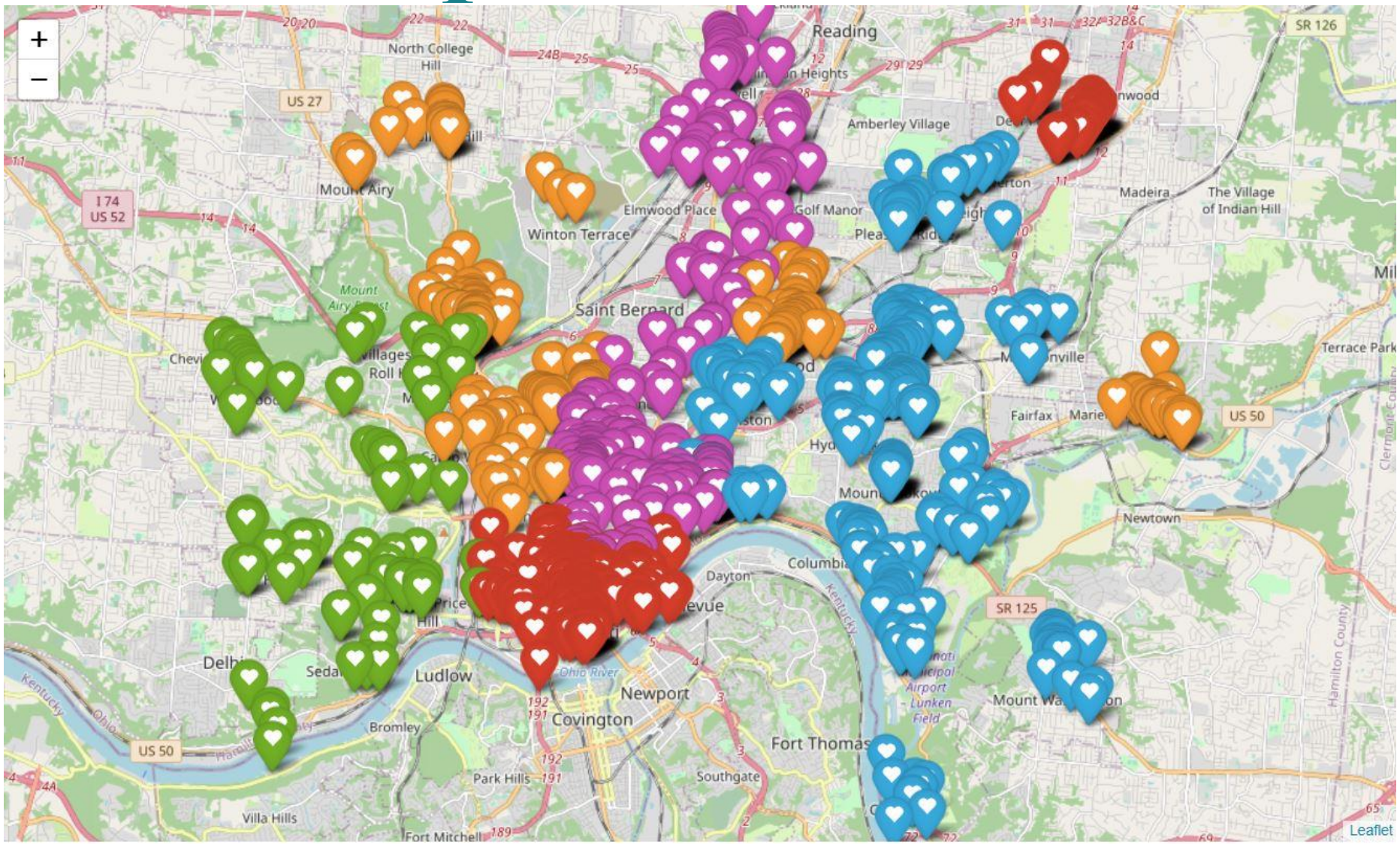


Exploratory Data Analysis

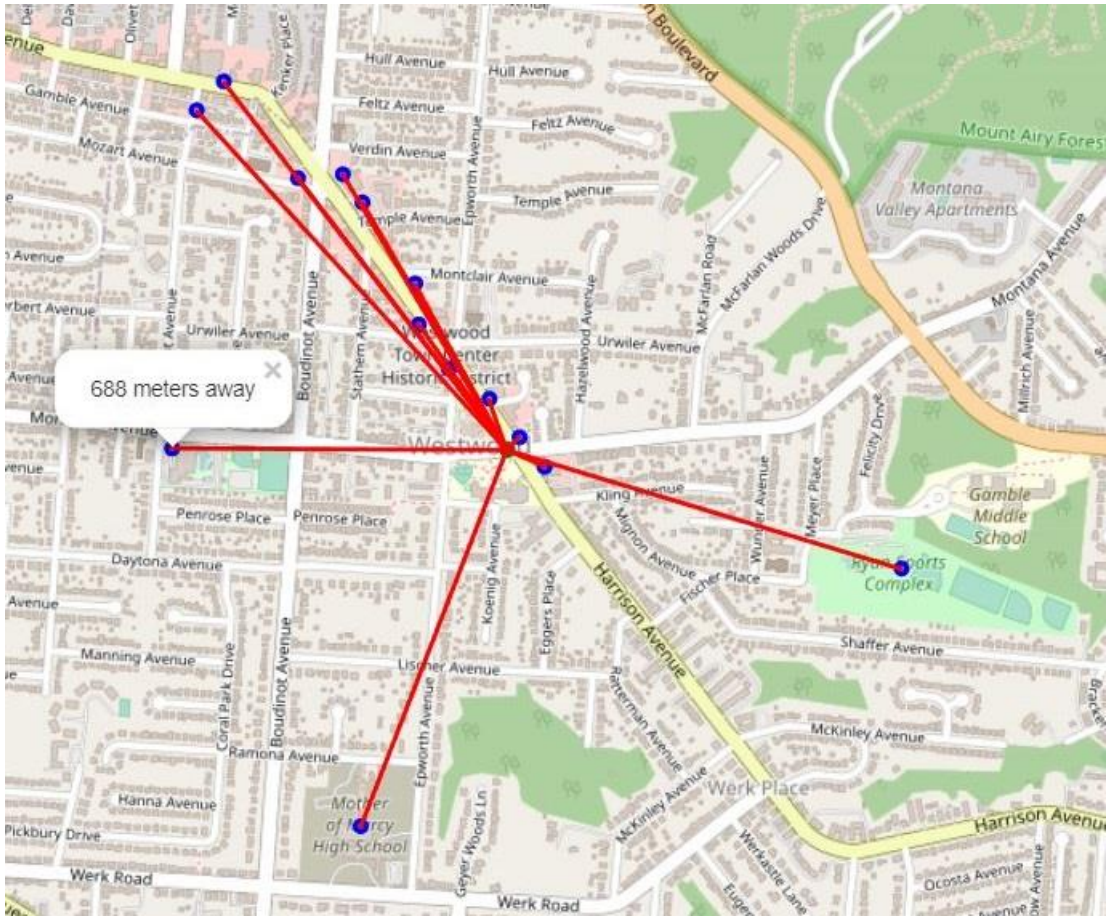
Folium Map with Neighborhoods



Folium Map with District Venues



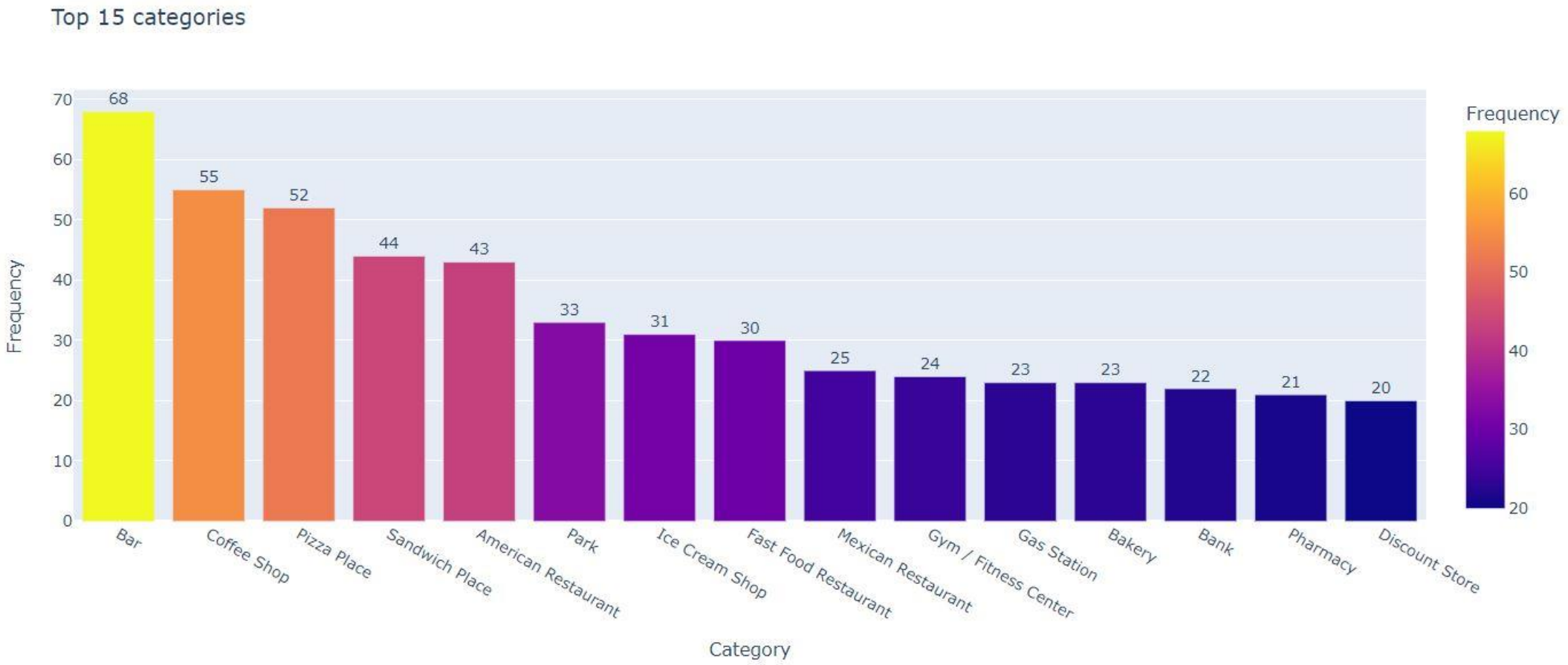
Distances Between Venues and Neighborhood



In this example, it is showing the Westwood distances.

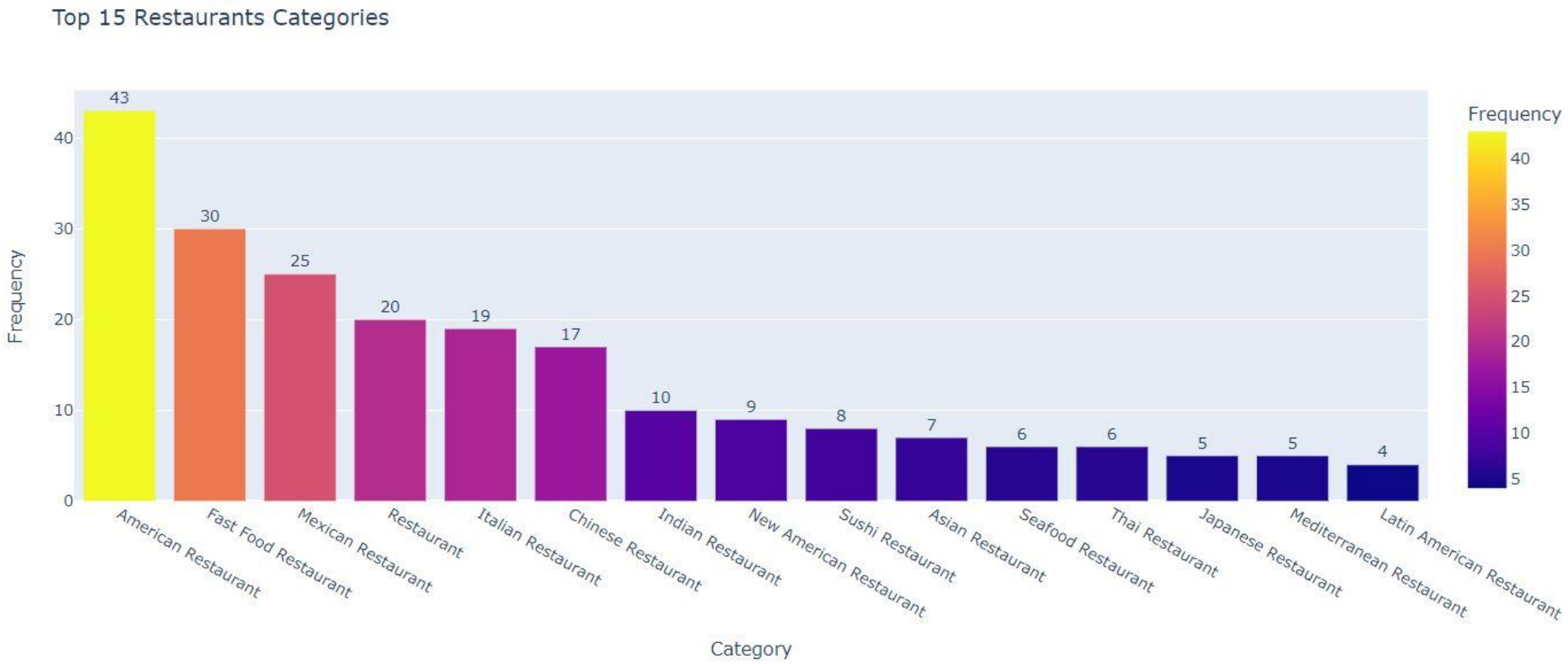
Bar Graphs

TOP 15 VENUES



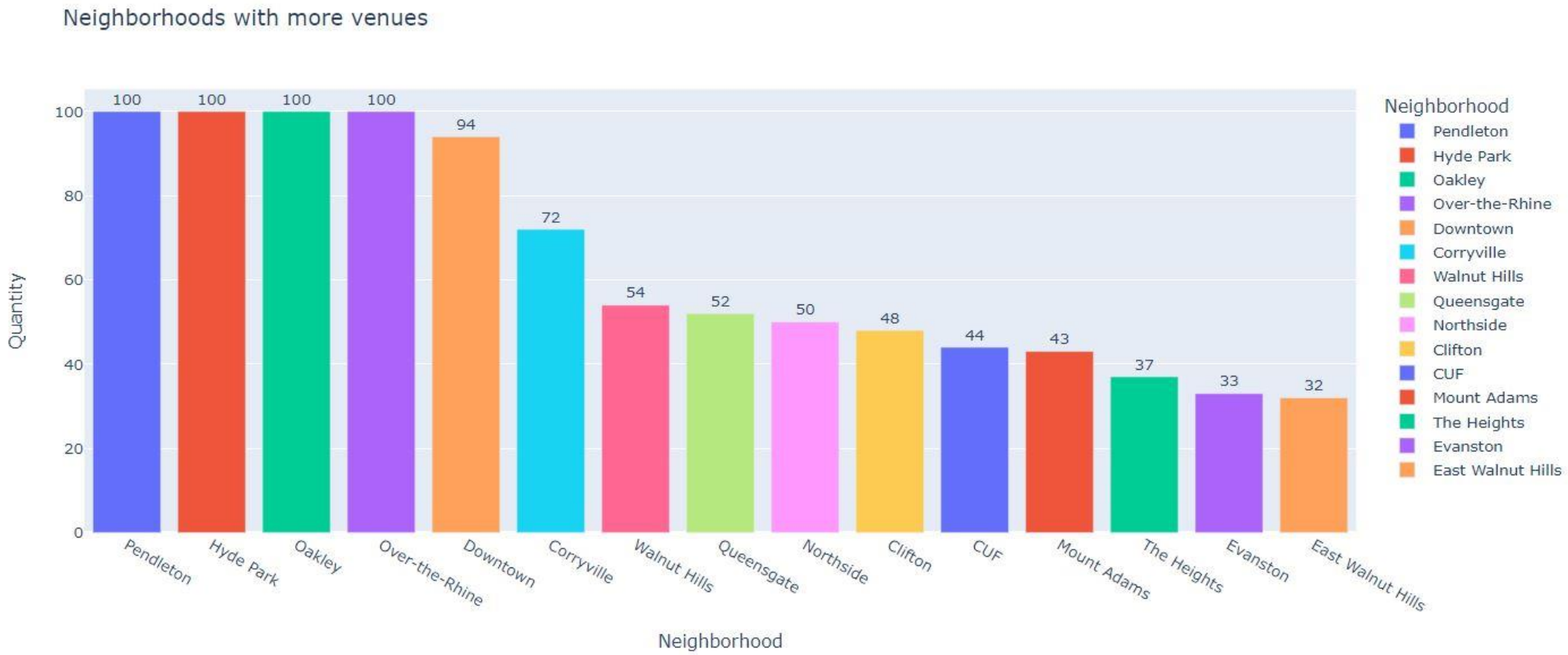
Bar Graphs

TOP 15 RESTAURANTS

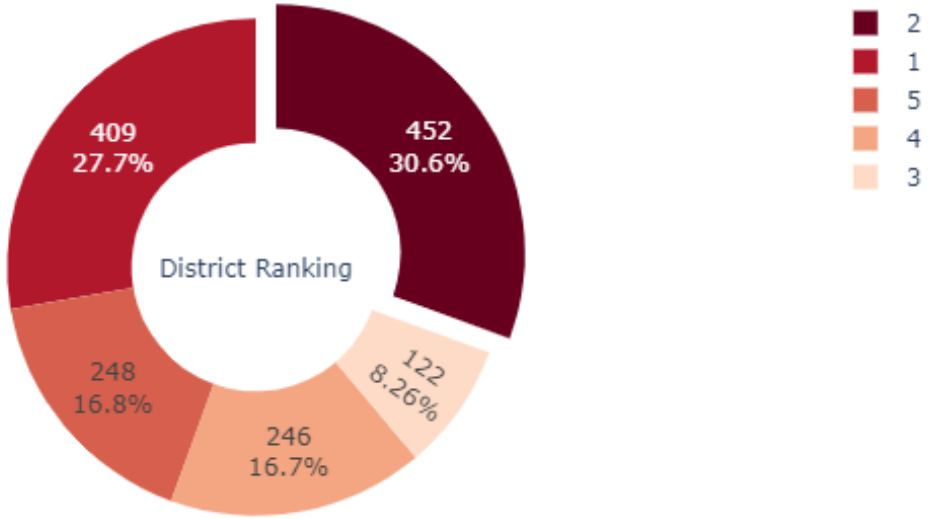


Bar Graphs

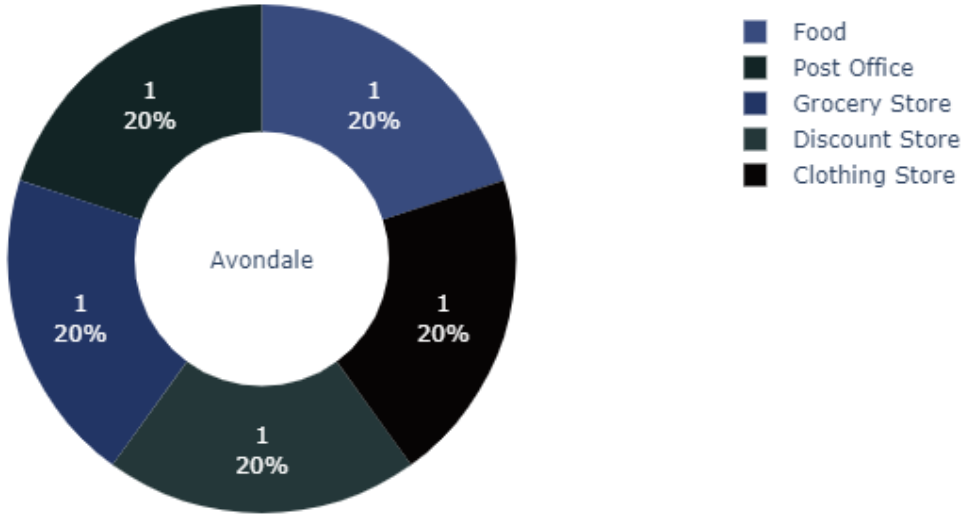
TOP 15 NEIGHBORHOODS



District Ranking



Top 5 venues per Neighborhood



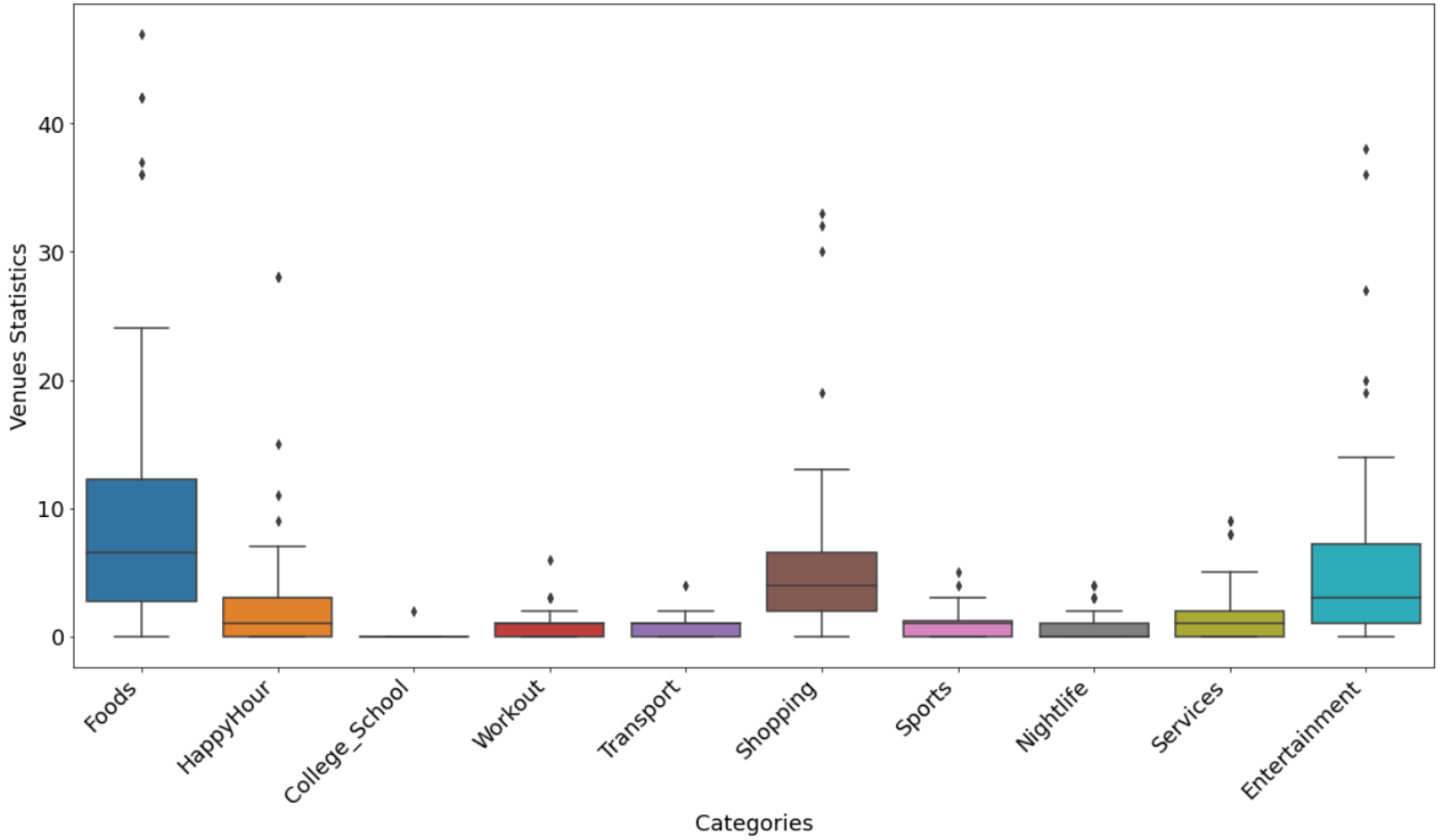
Feature Engineering

Feature Engineering

After extract, transform, load and explore the data, an important step before applies any machine learning model is the feature engineering. This step is essential to prepare the data and make its format compatible with machine learning algorithms and thus improve the model's performance.

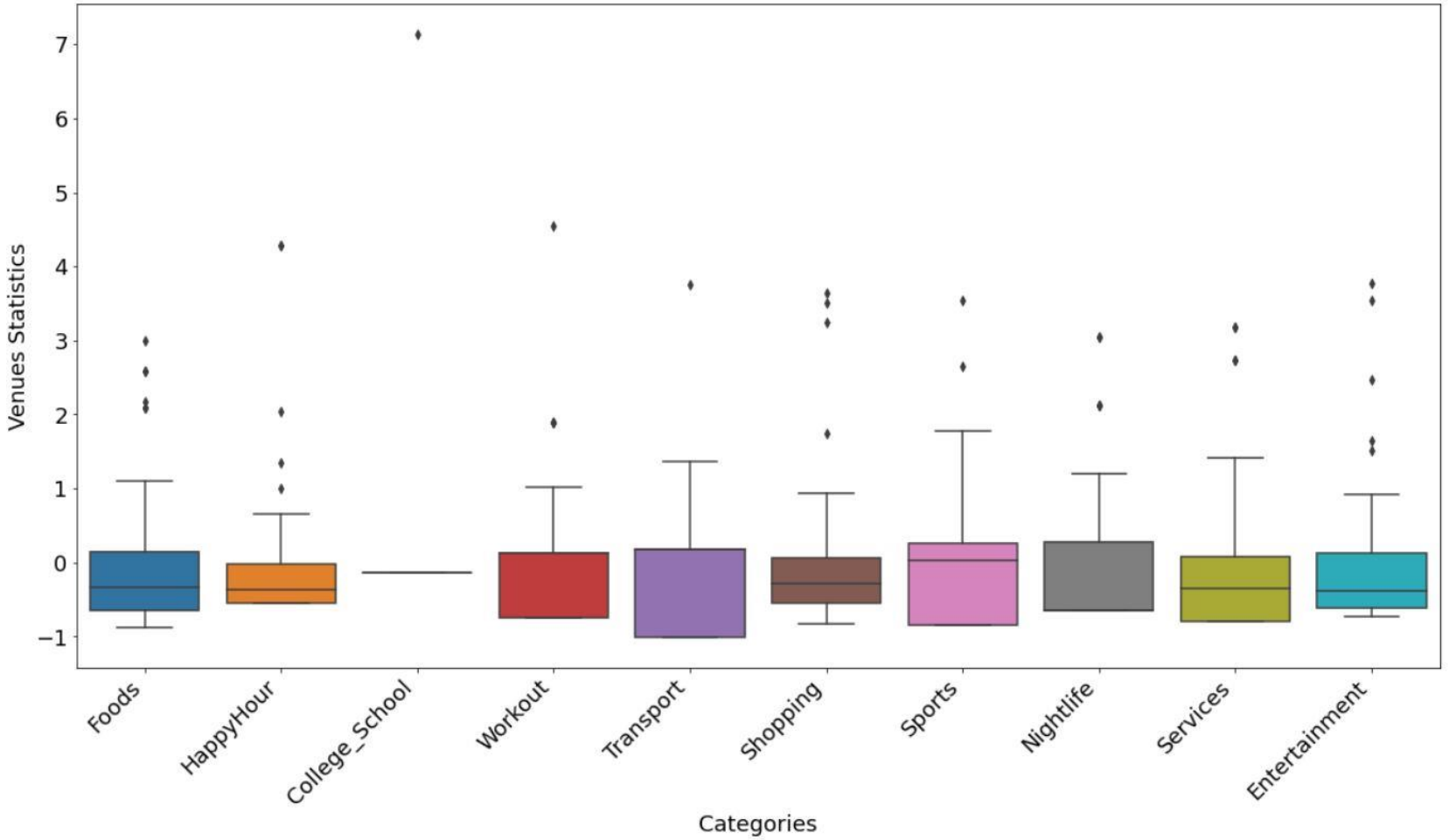
Feature Engineering

Before Standardization



Feature Engineering

After Standardization



Clustering with K-Means

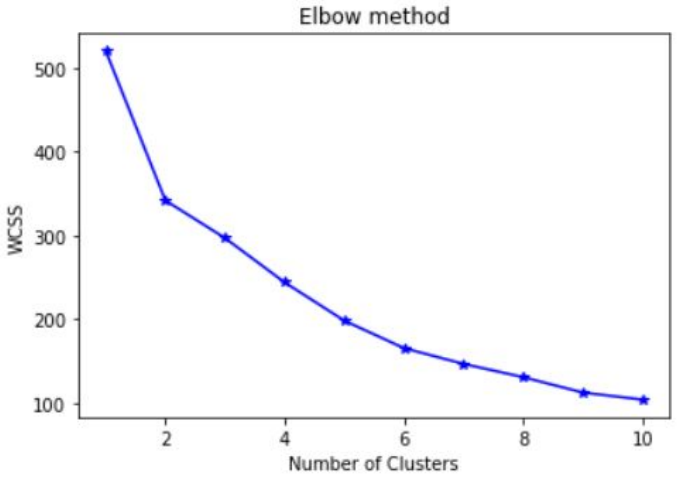
Applying K-Means

This is the data modeling phase. To cluster the neighborhood, the K-Means algorithm was applied. K-Means is an unsupervised algorithm, meaning it works with unlabeled data. The purpose of this algorithm is to find similarities and group the data according to the number passed in variable k.

Elbow and Silhouette Methods

Finding the best k value

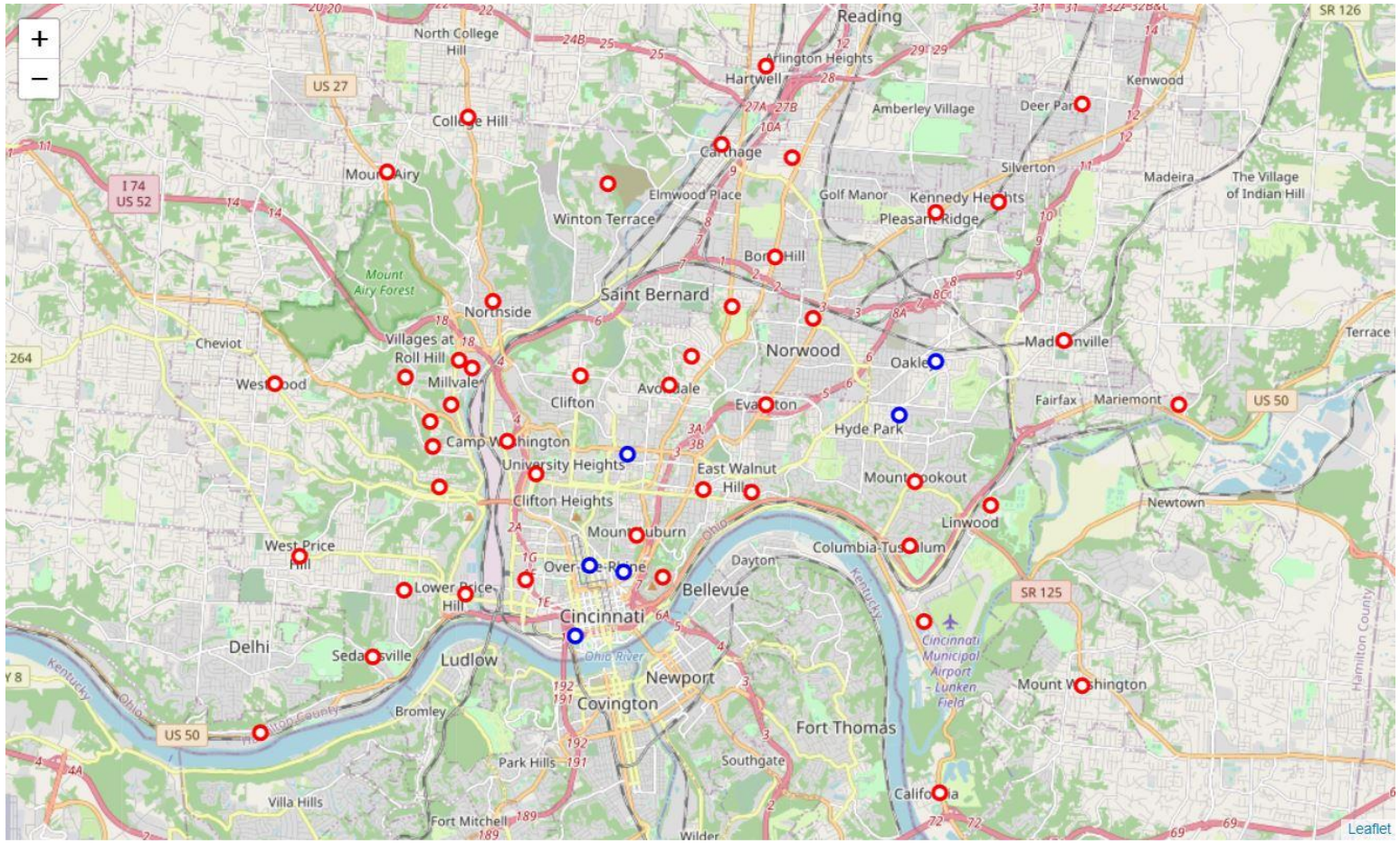
To find the best k value, two methods are commonly used. The silhouette calculates a value ranging between 1, -1 and measures the similarity of an object to its own cluster compared to other clusters. When the calculated value is high, it indicates that the object is well matched with its own cluster and poorly matched with neighboring clusters. The elbow method calculates the sum of the squared distances of the samples to the nearest cluster center using a predefined range of values for k clusters.




```
Sillhouette Method Results
N_cluster: 2, score: 0.5260105170007969
N_cluster: 3, score: 0.2638619489636931
N_cluster: 4, score: 0.43293805927952383
N_cluster: 5, score: 0.23326308075205043
N_cluster: 6, score: 0.25532775029902594
N_cluster: 7, score: 0.19523291430020973
N_cluster: 8, score: 0.2422403063112023
N_cluster: 9, score: 0.21040364159400393
N_cluster: 10, score: 0.21714338297378413
```

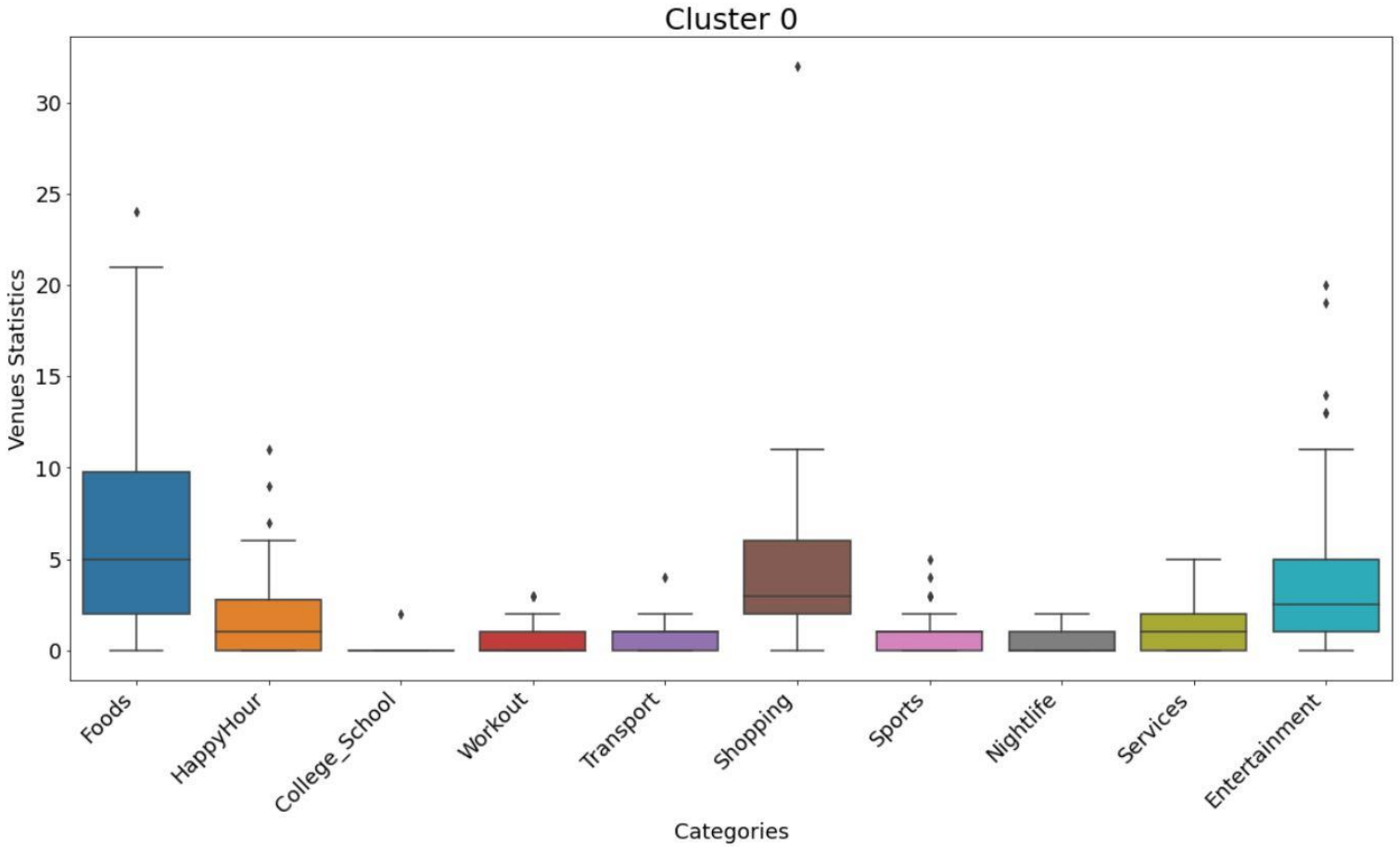
Map with clusters

Cluster_0: Red
Cluster_1: Blue



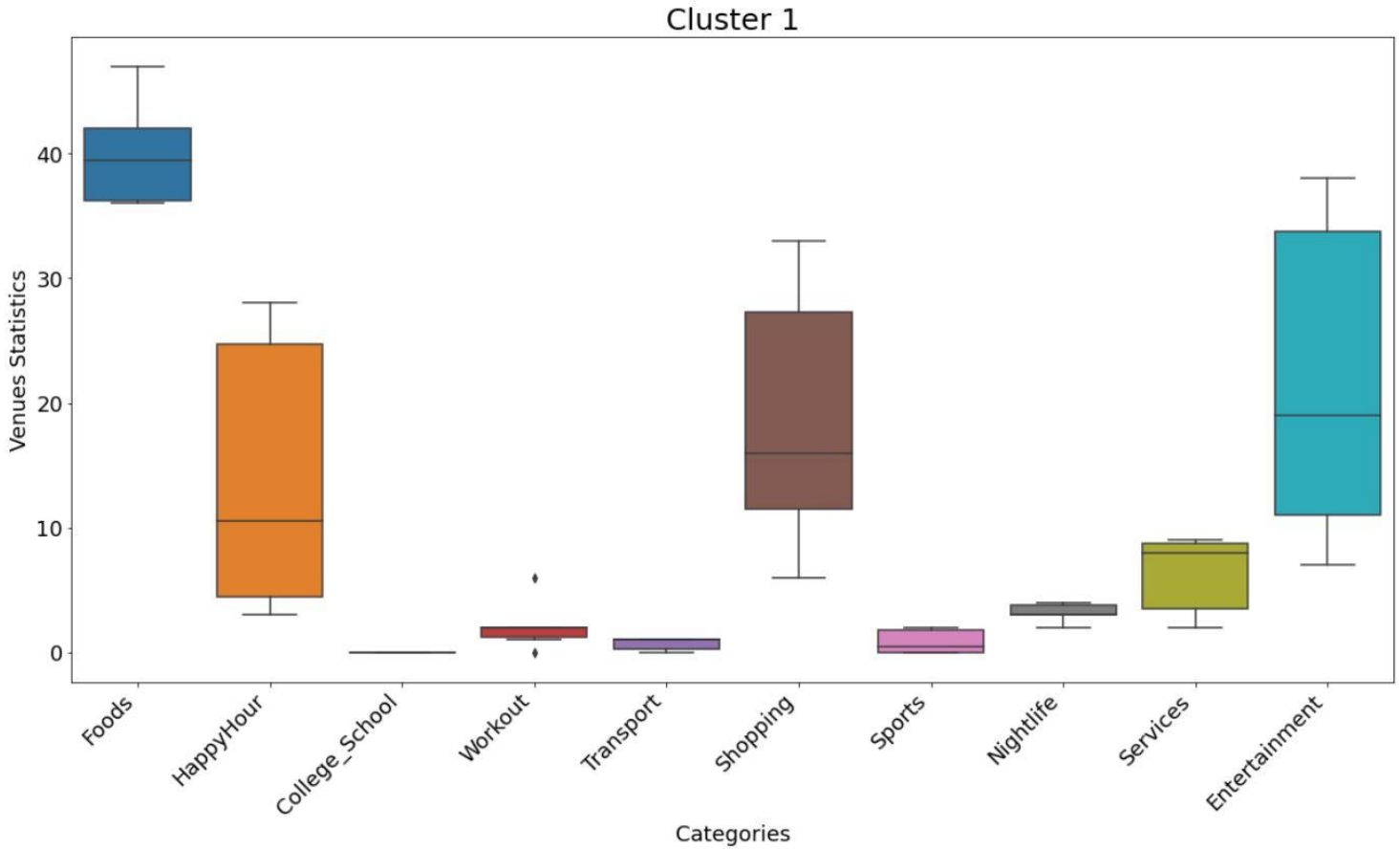
Analyzing the Clusters

Cluster_0



This cluster covers 46 neighborhoods and most of these neighborhoods are located in the most peripheral regions of the city. In these neighborhoods the average amount of each category is usually lower, these neighborhoods are more quiet and probably predominantly residential. The next boxplot of this cluster shows this information. The predominant venue category are “Foods”, but the average of occurrences is still low like the other ones.

Cluster_1



This cluster covers only 6 neighborhoods and they are located in the downtown region of the city. This cluster is formed by busier neighborhoods, with more options for restaurants, commerce, services and entertainment.

Discussion

Discussions

There are many other possibilities that can be explored and improved from this data and also work them together with other data, such as real estate rental values in the studied neighborhoods or using crime rates as well, but for this, it is necessary that texist available data or the possibility of creating a database for this.

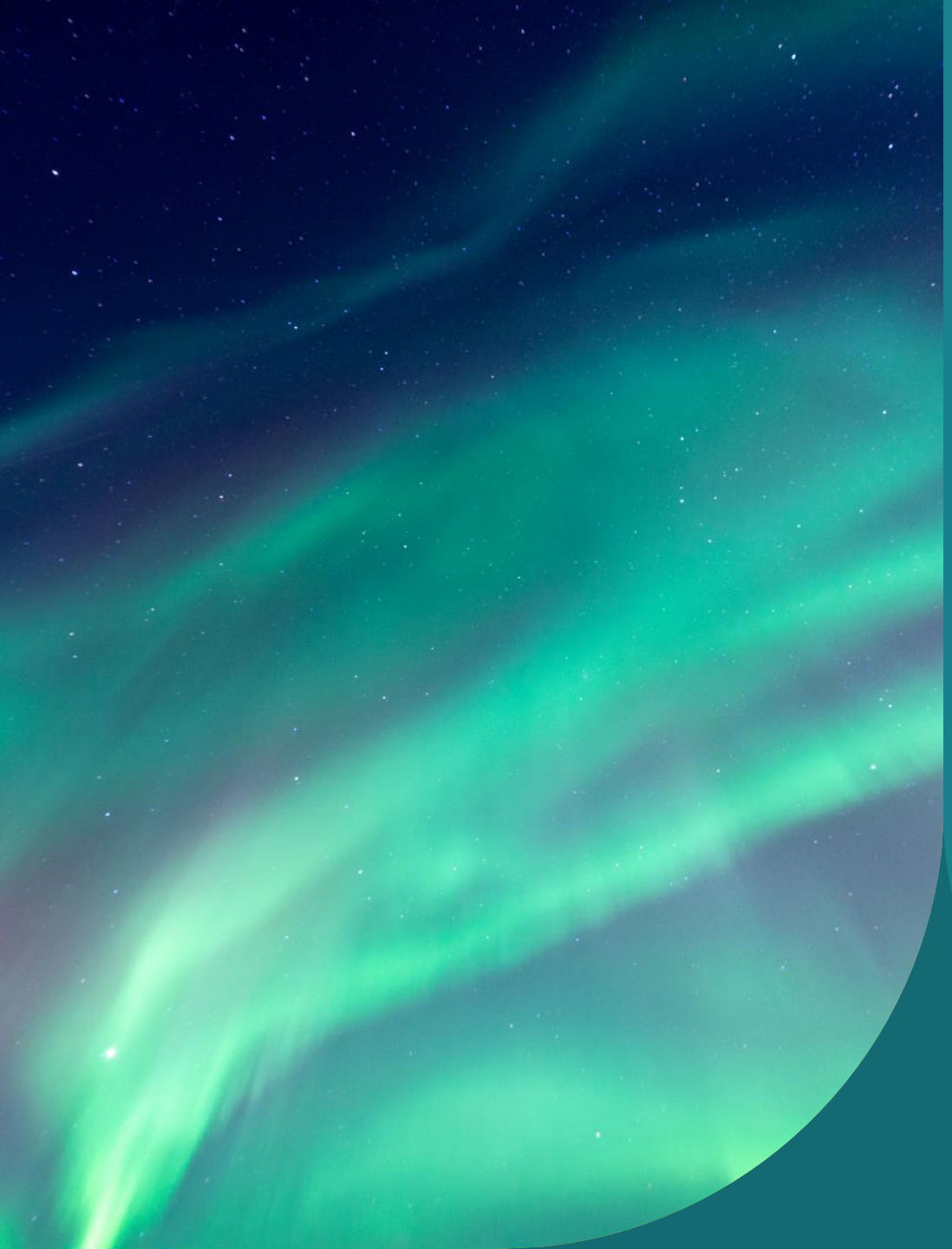
Conclusion

Conclusion

This project was completed achieving the main objective, which was to learn more about Cincinnati neighborhoods using their closest venues obtained in the Foursquare API. Thus, the two clusters formed brought relevant information about each group of neighborhoods and this will certainly be useful to select one of these neighborhoods to live.

Summary

I would prefer busier neighborhoods with plenty of nearby places to go with a short walk, so in this case, I certainly would choose one neighborhood of the Cluster_1 to look for a house or apartment for rent. We saw that the main differences between Cluster 0 and 1 are related to the number of nearby places of foods, happy hour, shopping, services and entertainment categories. The K-Means algorithm proved to be effective in selecting each neighborhood to cluster based on their venues.



Thank You

Leila Fabiola Ferreira

leila.ferreira@al.infnet.edu.br

https://www.linkedin.com/in/leila-fabiola-ferreira-31675163/

https://github.com/leilaff89/foursquare_project