FOURSQUARE PROJECT

10/19/2021

Living in Cincy

In this project, the data provided by the Foursquare API related to venues in the city of Cincinnati, Ohio, USA are explored. The main objective is to get characteristics of its neighborhoods and assess which ones are best suited to the lifestyle of those who will live in this city. To achieve this, a complete study is done using important steps of a data science project (from data collection to the evaluation of the applied model).

Foursquare Project

LIVING IN CINCY

1- INTRODUCTION

Life is a constant change. One day you are just well, doing your routine activities and suddenly you are caught by surprise with an opportunity in another city, maybe in another state, another country and even another continent. So, you stop to think and realize that you will have to make important decisions, but you have no idea what awaits you in this new "world". To assist in decision making, currently, there are many tools that help us to better understand several different scenarios, like the one mentioned just now. 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. The applied tools will be discussed throughout this document and all processes required for this project were done using the Python programming language in Anaconda's Jupyter Lab environment.

1.1 - About the city

This is a wonderful city with a beautiful view of the Ohio River and it can offer a little of everything: varied types of restaurants, many tourist sites such as museums, parks, festivals, in addition to the imposing Roebling Suspension Bridge (Roebling Bridge). Intimately known as “Cincy”, it has elegant neighborhoods, lots of green areas and a lot of German influence. Next to Roebling Bridge are the two largest sports stadiums in the city, one for baseball (The Great American Ball Park) and the other for American football (Paul Brown Stadium) which are very traditional sports in the United States of America. Its oldest and leading university, the University of Cincinnati, was founded in 1819 as Cincinnati College and has an annual enrollment of more than 44,000 students, making it the second largest university in Ohio. There are 52 neighborhoods in this city, and the venues of these neighborhoods are the object of study in this project.

> **“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**

2 – 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.

2.2 – Problem Statement

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

2.2 – 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.

3 – DATA REQUIREMENTS

Firstly, it is necessary to obtain a list with the neighborhoods of the city to be explored, in this case Cincinnati. After obtaining this list, it is necessary to obtain the coordinates of each neighborhood contained in the list. With the neighborhood coordinates located, the next step is to make a request to the Foursquare API, which will return the venues within a given proximity radius (in this case 1 Km) from each neighborhood coordinate informed. All these data will be used for further analysis.

4 – DATA COLLECTION

These steps following what was defined in the data requirements topic.

4.1 Web Scraping

So, firstly, web scraping was applied to the following Wikipedia page, that contains the list of neighborhoods and districts of Cincinnati <https://en.wikipedia.org/wiki/List_of_Cincinnati_neighborhoods>. After this process, a dataframe was created with the lists containing the 5 main districts and their respective neighborhoods (52 neighborhoods in total).

4.2 Getting the neighborhoods coordinates

Now that the list of neighborhoods was obtained, it was necessary to obtain the coordinates of each neighborhood. For this, a function that uses the geocoder library with ArcGis as data provider was used. More details of this library can be found at <https://pypi.org/project/geocoder>. This data was added to dataframe.

4.2.1 Visualizing the neighborhoods in a Folium map

Using the coordinates, it's possible to visualize the points of each neighborhood in a leaflet map using the Folium, a library used for visualizing geospatial data. Look at the Figure 1:

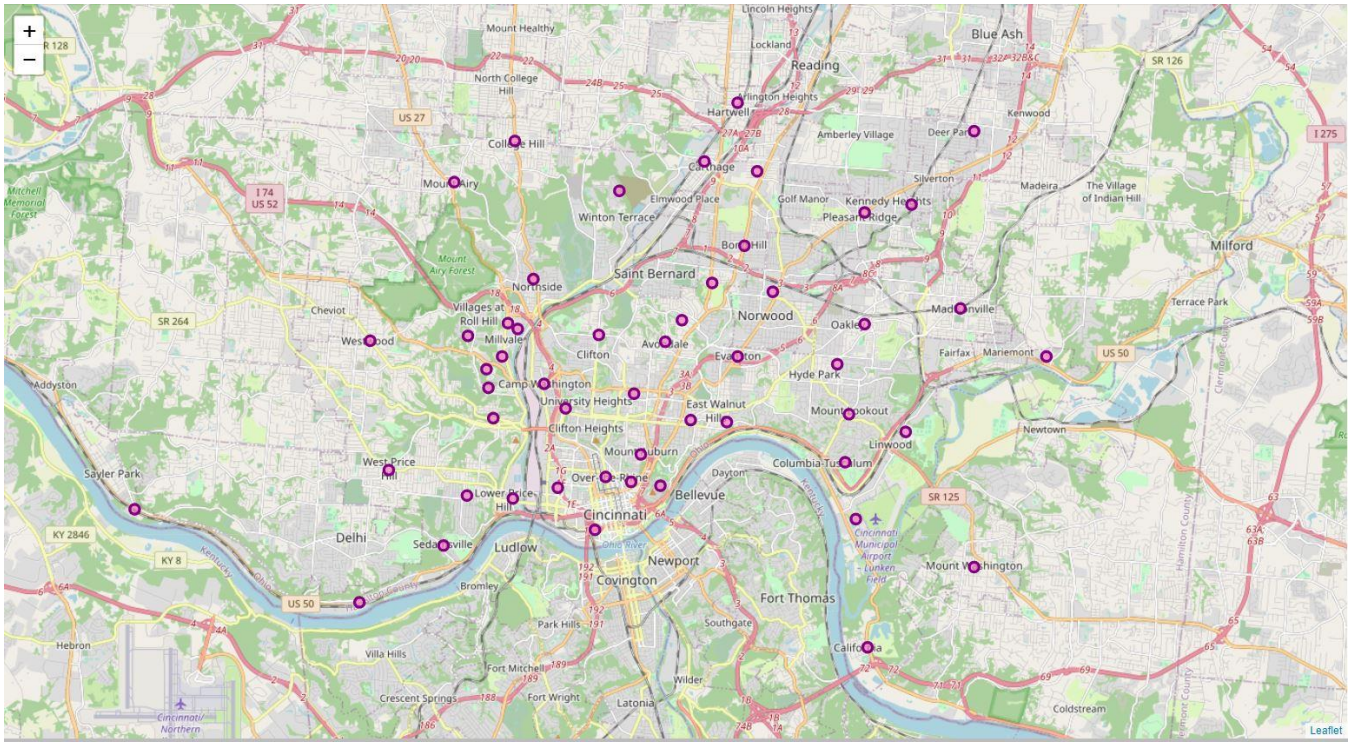


Figure 1: Cincinnati's neighborhood points in a leaflet map using Folium

It is a great way to visualize the distribution of neighborhoods, venues and many other information in a map as well. This same visualization was applied after clustering to see the distribution of clusters.

4.3 Getting venues from Foursquare API

Finally, a request is made to the Foursquare API to get the venues that are close to the informed neighborhoods with a maximum radius of 1 km and limited to 200 venues per coordinate. From the coordinates of the 52 neighborhoods of Cincinnati, 1477 venues were obtained, distributed in 256 different categories generating a new dataframe. In the exploratory data analysis stage, some relevant information about these categories was extracted and presented. This API can return a lot of information based on the given coordinates, but in this project, we only got the venues' names, their coordinates, category and distance from the given coordinates. To use this API, it is necessary to create a login, password and generate a token for access. To get more information about the usage of Foursquare API, access the following link <https://developer.foursquare.com/>.

5 – EXPLORATORY DATA ANALYSIS

In this phase, the dataframe generated after the Foursquare API request was used to discover relevant information about the database and generate graphs to facilitate this visualization. But first, a relevant information obtained about the venues was that there are 1477 different venues, but only 1251 unique points in the dataset. This issue occurred because the radius was set to 1 Km to obtain the data from Foursquare API. Therefore, these 226 points overlap from one neighborhood to another but, it is known that the maximum distance for each venue is 1 kilometer considering the center point of each neighborhood, so it is easily accessible anyway. Let's analyze the next charts and figures to look for new insights.

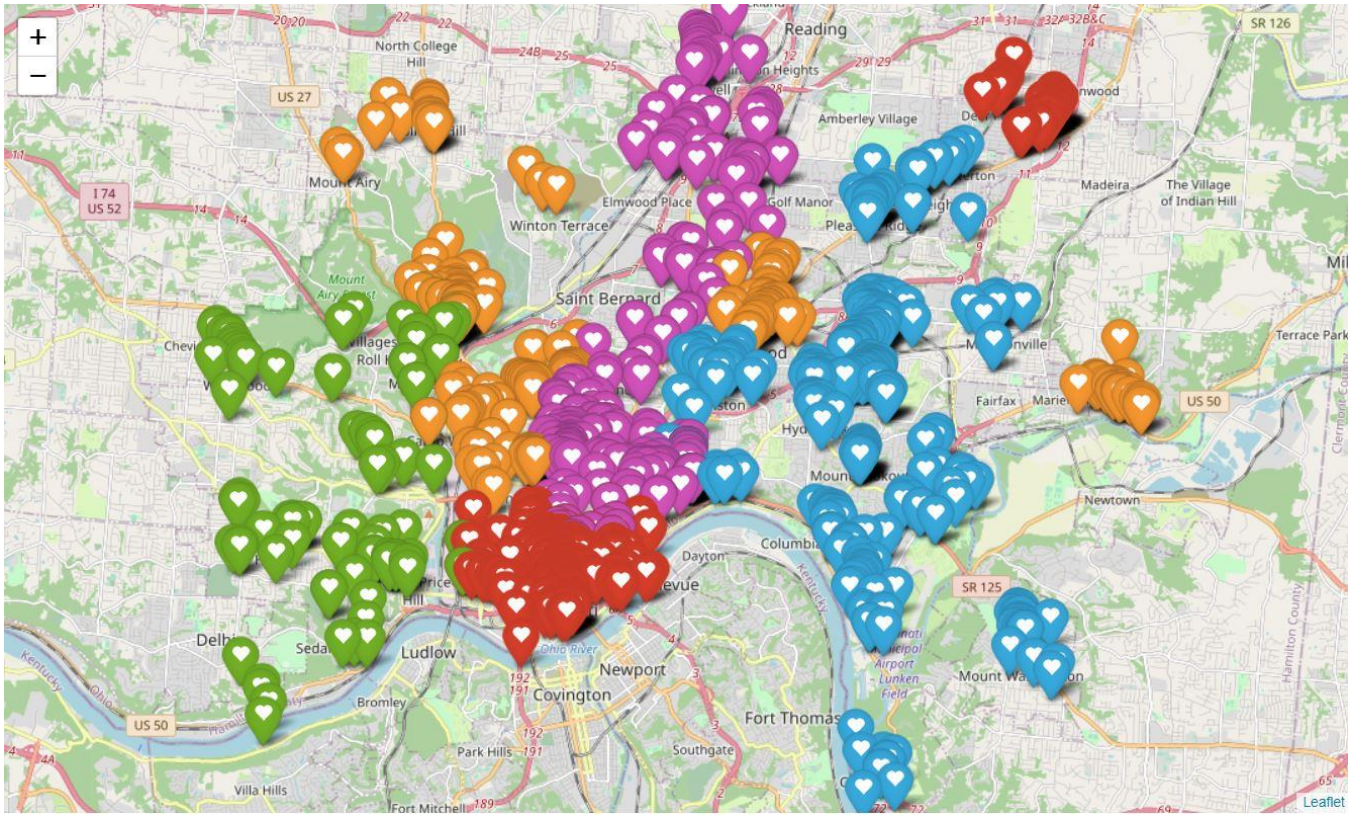


Figure 2: The venues distributed over the 5 main districts

In this figure we can visualize the venues over the 5 districts of the city. The red icons are venues in the district one, the blue ones are in the district 2, the green ones are in the district 3, the purple ones are in the district 4 and finally the orange ones are in the district 5. It shows that district 1 is probably the smallest and district 2 is the biggest.

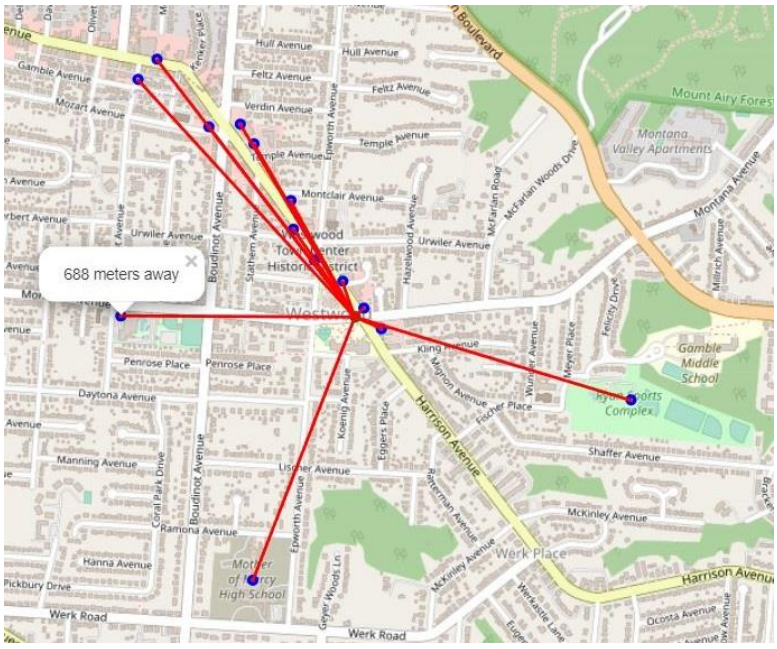


Figure 3: Distances between the venues and the center point in Westwood

The Figure 3 shows that it is possible draw lines based on distances between a central point of a neighborhood and its venues. For instance, in this figure we can see the distances of all venues in Westwood. The next bar plot shows the top 15 categories of venues. This graph allows us to see the most common venues' categories in Cincinnati, but not looking at the neighborhoods separately.

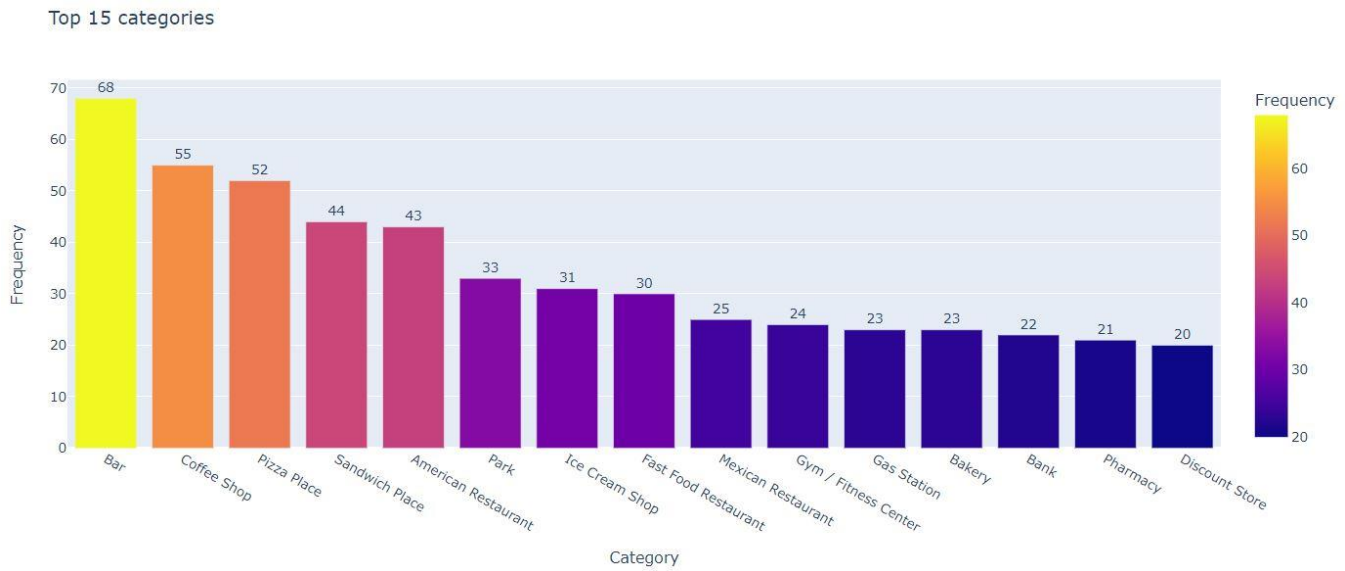


Figure 4: Top 15 venues' categories in Cincinnati

Great, interest and useful places to go right? The next bar plot is similar to the previous bar plot, but shows the top 15 restaurant categories in the city. Likewise, this graph allows us to see the most common restaurant categories in Cincinnati, and also not look at the neighborhoods separately

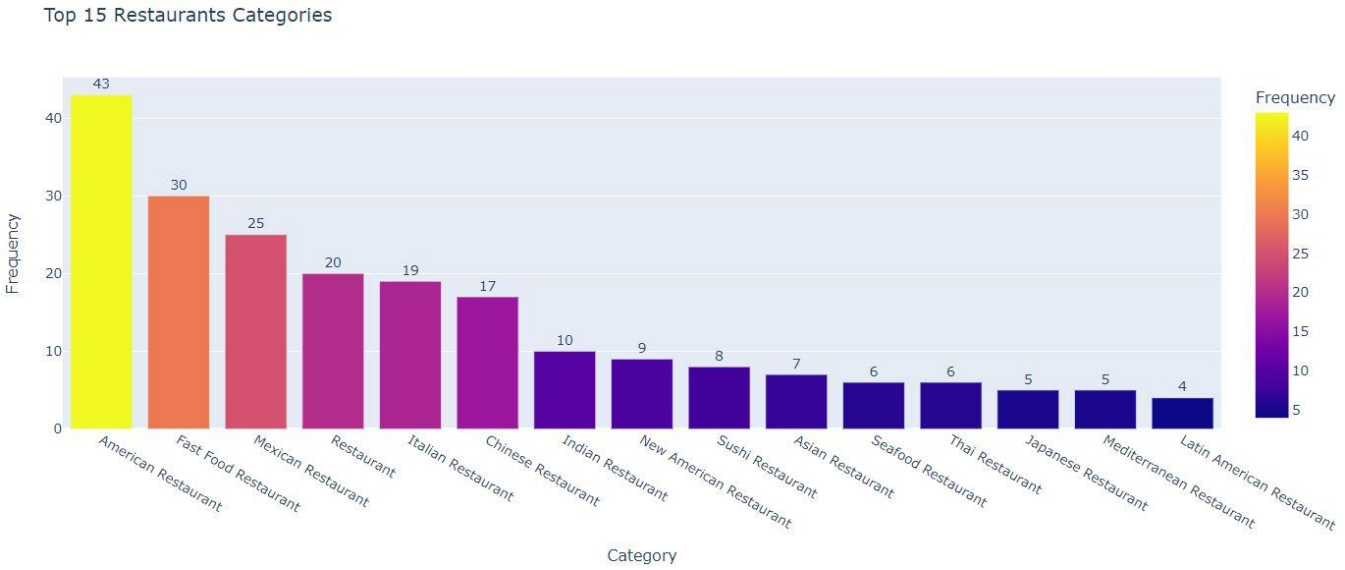


Figure 5: Top 15 restaurants categories in Cincinnati

The predominant kind of restaurant in Cincinnati is the American Restaurants, following by Fast Food Restaurants and Mexican Restaurants. There are some categories of restaurants that have only one occurrence. Now let's look at the main neighborhoods, considering the number of locations in each one.

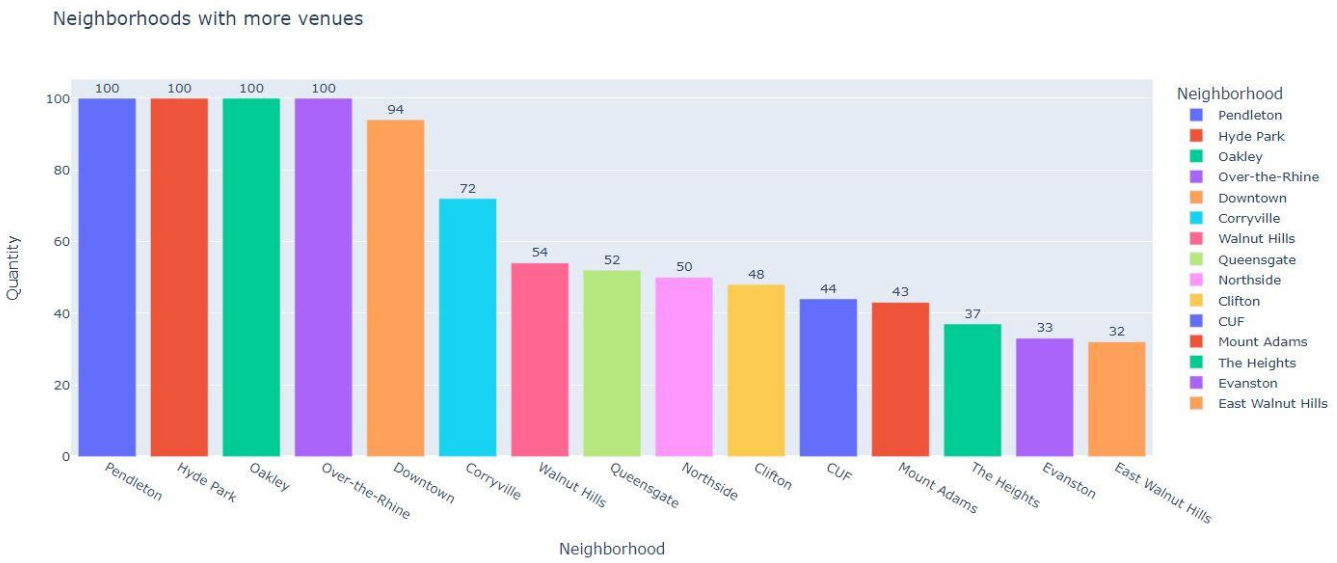


Figure 6: The neighborhoods with more venues

This graph shows the neighborhoods with more venues, and we can use this information to compare with the groups extracted in the clustering phase. will the number of venues in each neighborhood be crucial for clustering? Finally, the next plot shows a district ranking based on the number of venues in each district.

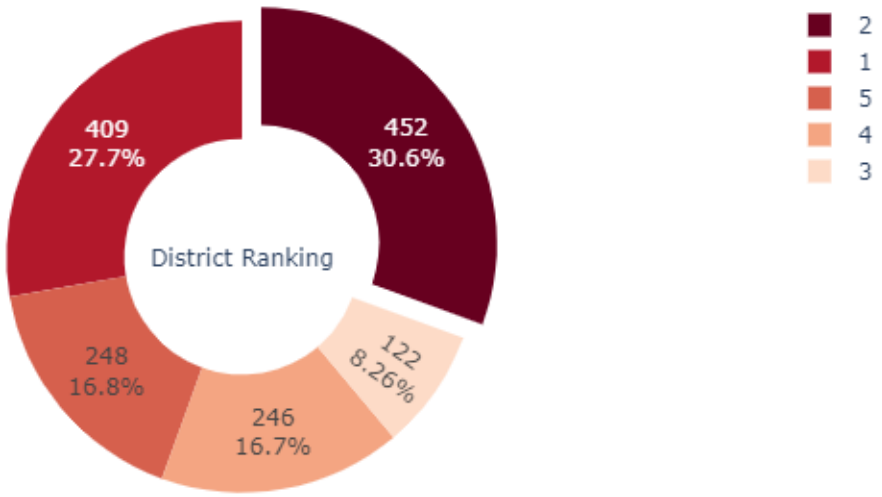


Figure 7: District Ranking

6 – 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. So, first the categorical data needs to be transformed in numerical data, because machine learning algorithms works with numbers. To do this, the get_dummies function of Pandas library was applied in the category column. This process transforms the category column in many other columns, and each column represent a unique category (in this case 256 new columns because the dataset has 256 different categories). Now each data cell contains only "0" or "1", which represents whether or not that neighborhood has that specific category. Next, all occurrences were added to obtain the total number of occurrences for each category per neighborhood. After analyze the 256 different categories, it was seen that many categories were very similar and could form a new larger category. Then the categories were reorganized into 10 new major categories and a few others were dropped as not relevant. The 10 new categories were classified as: Foods, HappyHour, College_School, Workout, Transport, Shopping, Sports, Nightlife, Services and Entertainment, generating a new dataset with 52 neighborhoods in the rows and 10 categories in the columns, where each column brings the number of occurrences of that category per neighborhood. Finally, the data were normalized using the StandardScaler function from the Sklearn library. This function is used to standardizes the features by removing the mean and scaling to unit variance, so, the data will have the same weight for the machine learning algorithm. This type of function is essential because if the dataset contains data in different scales, such as height in meters and price in dollars, for example, it will be standardized to the same scale. To read more about this function look at this link < https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. The next boxplot shows the data before standardization.

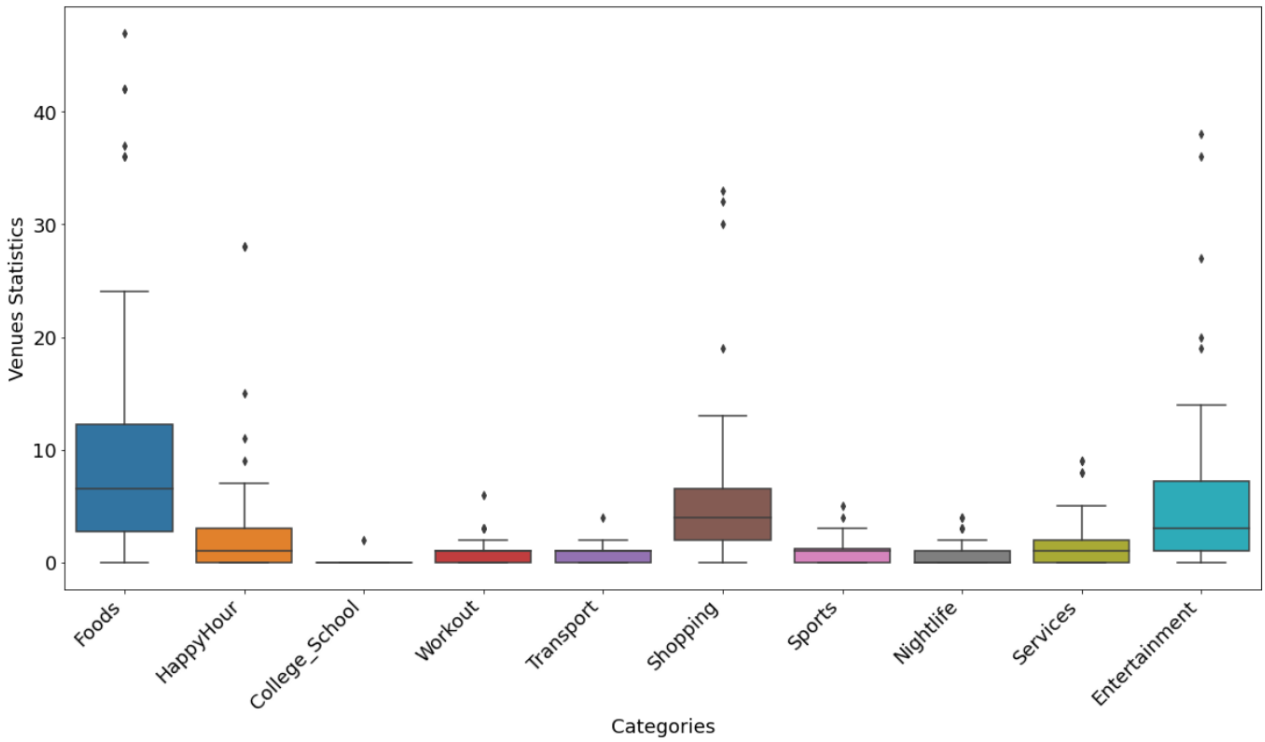


Figure 8: Statistical information before standardization

In this boxplot, we can visualize the statistical data about the ten categories. The Foods and Entertainment categories have the highest number of occurrences and College_School the lowest. The next boxplot is about the normalized data.

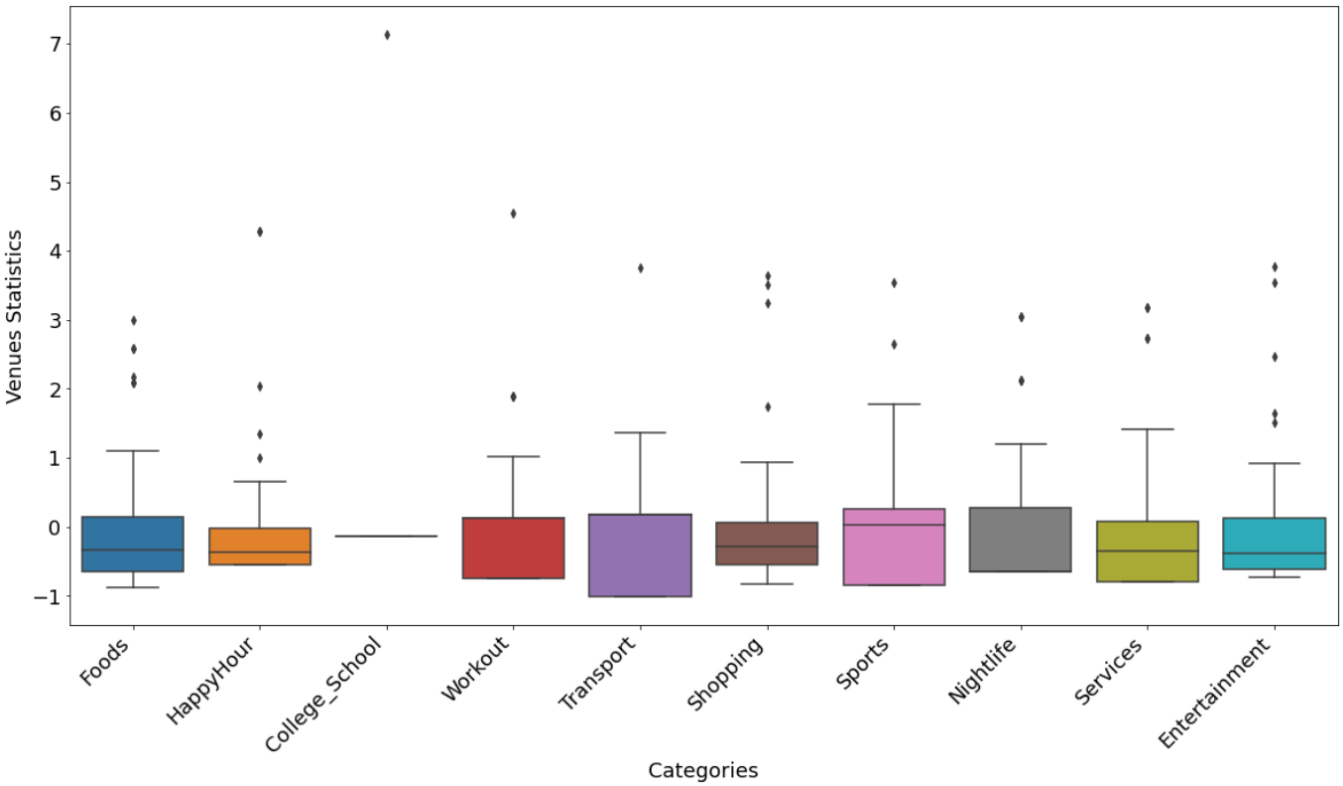
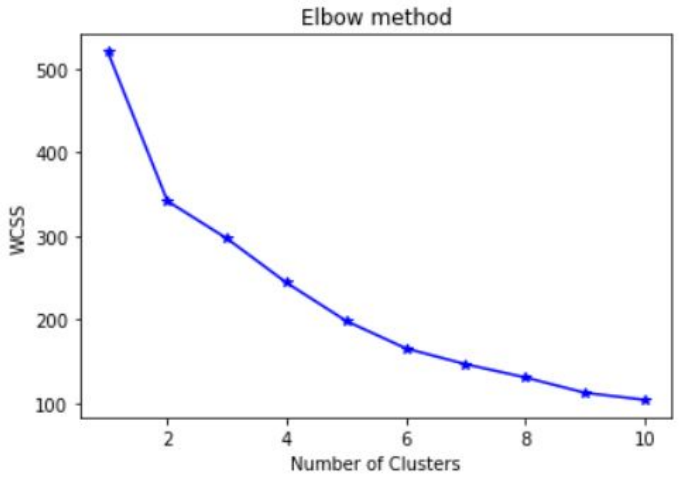


Figure 9: Statistical information after standardization

Now, it is possible to see that the categories are on the same scale, even with their differences, the range of all categories is usually close to -1 and 1. A new dataframe was generated with the changes made in the feature engineering phase.

7 – CLUSTERING WITH 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. 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. The ideal number of clusters is the value in which there is no significant decrease in the sum of squared distances from one k to the next. In this project, both methods were applied and presented good results for a k with a value 2, that means, two clusters. The next figures show the elbow method and silhouette method results.




```
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
```

Figure 10: Elbow and Silhouette Methods Results

After setting the k to 2 and apply the K-Means algorithm, the clusters are shown in a map, to see the neighborhoods of each cluster. Let's see the leaflet map of clusters using Folium.

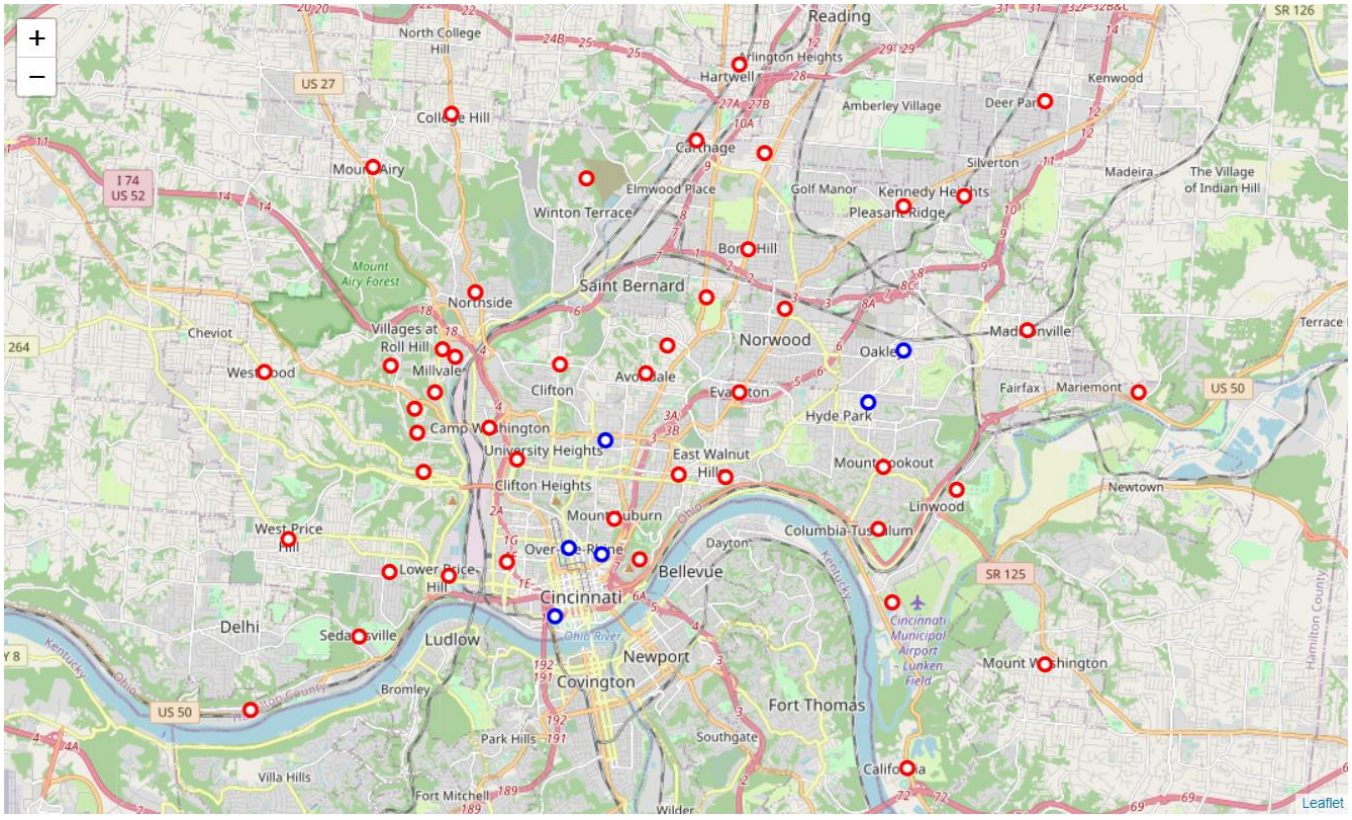


Figure 11: Clusters in a Folium Map

In this map, it is possible to visualize the two clusters formed by K-Means. Red is the largest and covers most neighborhoods. Blue is the smallest and only covers 6 neighborhoods.

8 – ANALYSIS OF GENERATED CLUSTERS

Now, we can analyze each cluster and try to understanding their patterns. Two cluster are generated using K-Means, they are labeled as Cluster_0 and Cluster_1.

8.1 Cluster_0

This cluster covers 46 neighborhoods and most of these neighborhoods are located in the most peripheral regions of the city. The neighborhoods in the list are: Avondale, Bond Hill, CUF, California, Camp Washington, Carthage, Clifton, College Hill, Columbia-Tuluscum, East End, East Price Hill, East Walnut Hills, East Westwood, English Westwood, Evanston, Hartwell, Kennedy Heights, Linwood, Lower Price Hill, Madinsonville, Millvale, Mount Adams, Mount Airy, Mount Auburn, Mount Lookout, Mount Washington, North Avondale, North Fairmount, Northside, Paddock Hills, Pleasant Ridge, Queensgate, Riverside, Roselawn, Sayler Park, Sedamsville, South Cumminsville, South Fairmont, Spring Grove Village, The Heights, The Villages of Roll Hill, Walnut Hills, West End, West Price Hill, Westwood and Winton Hills. 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.

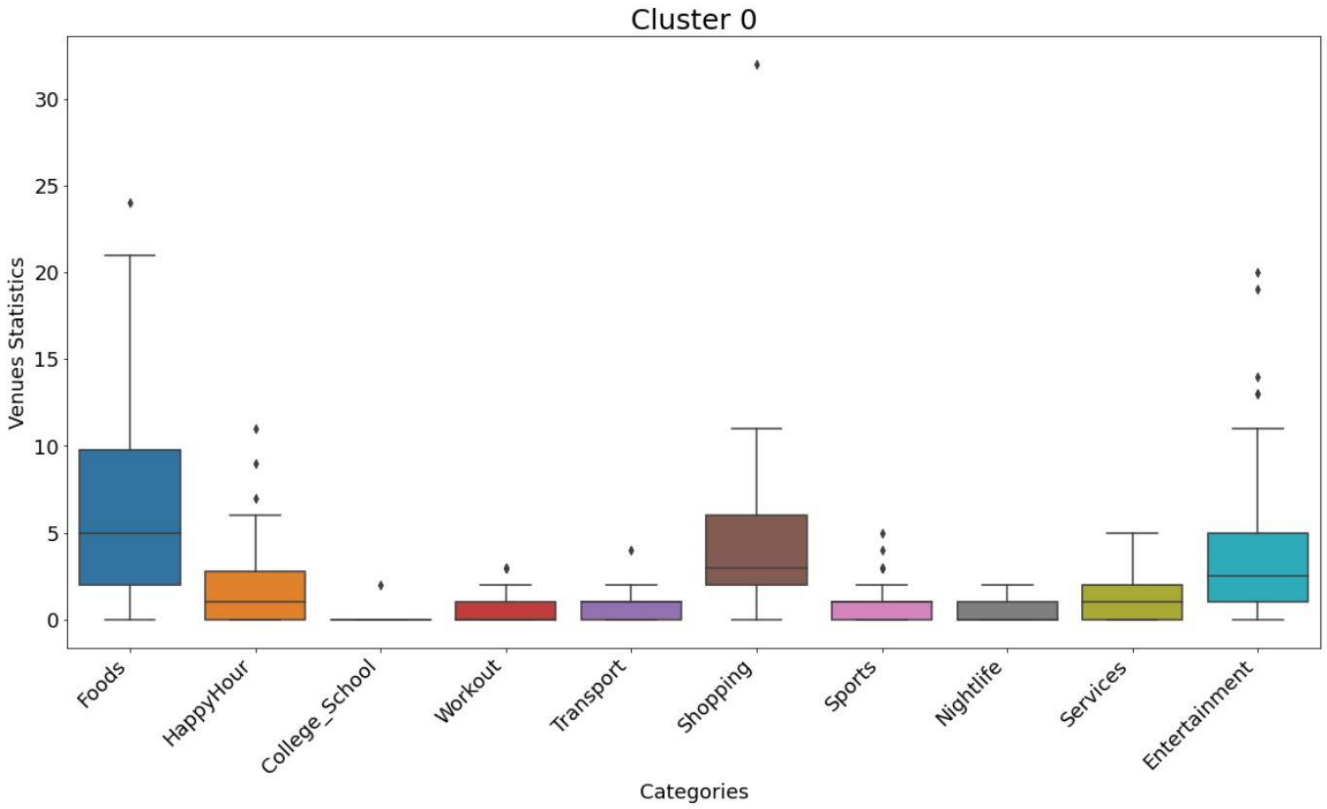


Figure 12: Cluster 0 statistics

8.2 Cluster_1

This cluster covers only 6 neighborhoods and they are located in the downtown region of the city. The neighborhoods in the list are: Corryvile, Downtown, Hyde Park, Oakley, Over-the-Rhine and Pendleton. This cluster is formed by busier neighborhoods, with more options for restaurants, commerce, services and entertainment. The next figure shows the boxplot of this cluster.

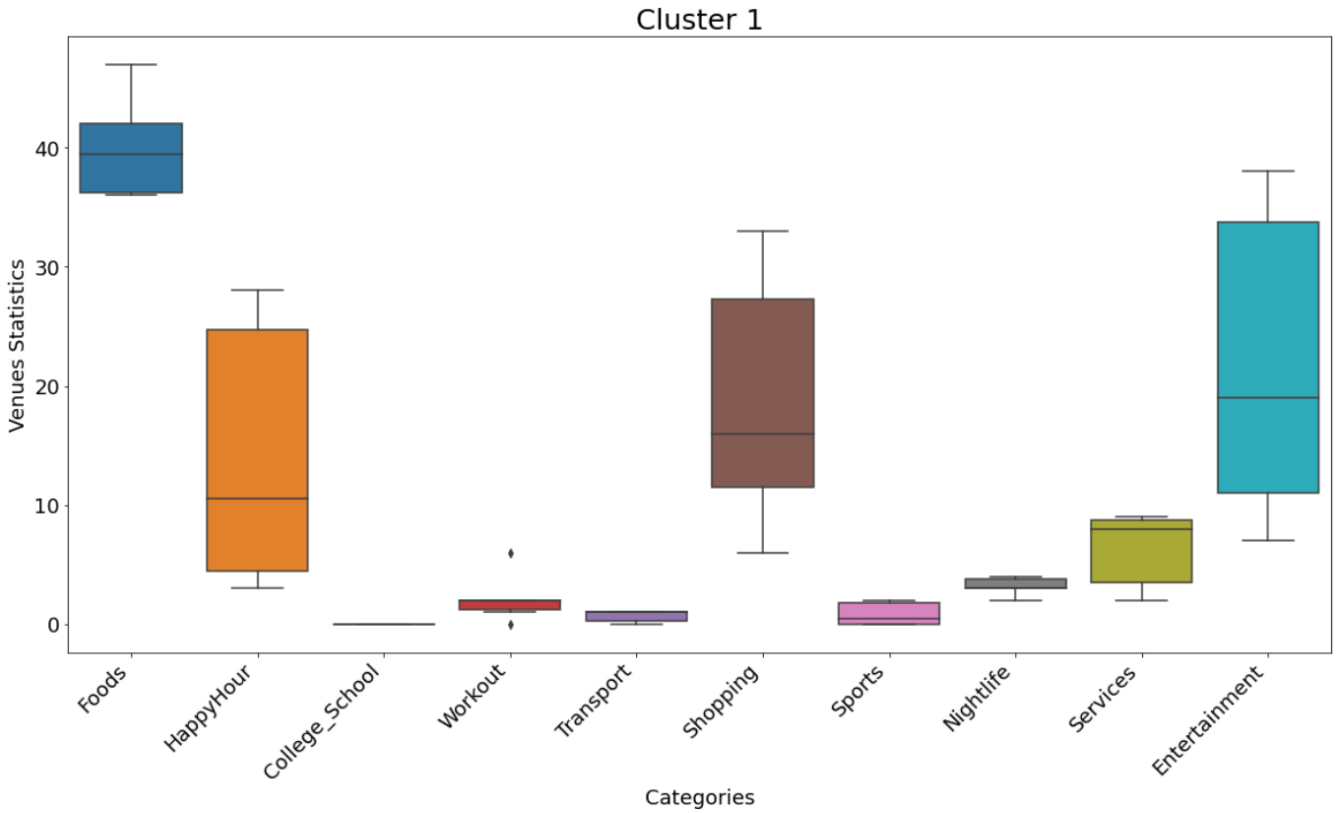


Figure 13: Cluster 1 statistics

It is clear the difference between Cluster_0 and Cluster_1 looking at their boxplots. Cluster_1 contains neighborhoods with lots of places to eat, to have a happy hour, to shop, to find services and have fun with different entertainments.

9 – DISCUSSION

To achieve the main objective of this project, which was to group the neighborhoods of Cincinnati based on the venues close to each neighborhood within a radius of 1 km and thus analyze the patterns of each formed group, it was necessary to follow the essential steps of a data science project. Firstly, understanding the problem to be solved and then defining what data would be needed, collecting data from various sources such as Wikipedia through Web Scraping and data obtained from the Foursquare API, analyzing, exploring and manipulating the data to suit the project's needs and finally, applying and evaluating a Machine Learning model, in this case the K-Means algorithm to cluster the unlabeled data and thus obtain the patterns of each formed cluster. 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.

10 – 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. In my case, I prefer busier neighborhoods with plenty of nearby places to go with a short walk and 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.

Foursquare Project Report developed by Leila Fabiola Ferreira