**Using Machine Learning and Foursquare Data to Find Ideal Locations to Open A Gas Station in The City of Miami, Florida**

*This project is done for a capstone course of the Coursera*[*IBM Data Science Professional specialization*](https://www.coursera.org/specializations/ibm-data-science-professional-certificate)*.*

*Python for Machine Learning, Foursquare API, Web Scraping*

**Introduction**

This project is completed as part of IBM’s Data Science Professional Certificate course offered by Coursera.org. One of the purposes of this capstone course is to understand what data scientists go through in real life when they need to come up with a data-driven solution to a given problem. Therefore, for this project, a scenario was assumed where an entrepreneur wants to open a gas station in the City of Miami, Florida. Given the presence of tourists throughout the year, Miami city is a great location for a gas station business. The city of Miami is divided into many neighborhoods. These neighborhoods have hundreds of venues and among these venues, there exist some gas stations as well. This project will look for neighborhoods where there is little or no presence of gas stations to find an untapped neighborhood’s market of Miami. Through using Foursquare’s geographic data and one of the machine learning tools, this project aims to help the entrepreneur finding a suitable location for the gas station.

**Business Problem**

One crucial decision for opening a gas station is to find a suitable location and much of the business success relies on it. The entrepreneur needs to know where exactly in Miami city he should set up the station. Intuition wouldn’t help him that much and that’s why the decision needs to be taken based on what data says. The output of this project will help him find a suitable location. By using data science methods and techniques along with machine learning algorithms such as K-means clustering, this project aims to give solutions to answer the business question: In the city of Miami, if an entrepreneur wants to open a gas station, where should he open it?

**Target Audience**

The analysis would help any automotive fuel company or an entrepreneur from the city of Miami who considers opening a gas station franchise among the city’s neighborhood.

**Data Description and Data Sources**

Data used in this project is collected from several sources. A brief description of the sources of data are given below:

* The list of neighborhoods along with their geographical coordinates from the city of Miami, Florida (“List of neighborhoods in Miami”, 2020).

Image for post

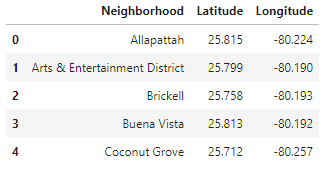


Table 1: List of Miami neighborhood scraped from Wikipedia

* Foursquare API to explore venue types surrounding each neighborhood of the city of Miami. The query was made for the number of venues in each category within a 1000m radius around each neighborhood (“Documentation — Foursquare Developer”, 2020).

**Methodology**

The list of neighborhoods of Miami is extracted from Wikipedia through web scraping. For web scraping, panda’s Html web scraping method is used as it is easier to fetch tabular data from a web page into a data frame. The data frame was cleaned and prepared for visualizing the map as it contains some unnecessary columns. The final data frame contains a list of neighborhoods name and their corresponding coordinates.

Now, a map of Miami is created using a folium library package to verify all coordinates are correct and it pulls out the right map for further analysis. In the next step, the Four Square’s API is used to pull out the list of top 100 venues within a radius of 1000 meters. To pull the data, a personal developer account from the Four Square’s sandbox tier is used.

Initially, the venues are pulled out only for one neighborhood called ‘Brickell’ to see how it works. The venues extracted from each neighborhood are in the JSON file type which then is transformed through normalizing. Also, the function is written to find the category of each venue. Finally, through filtering the category of each row, the data frame is extracted in tabular format which contains the names of each venue including their categories and coordinates. Now, this time the nearby venues of all neighborhoods listed above are extracted through Four Square API and organized in a data frame including venue latitude and longitude.

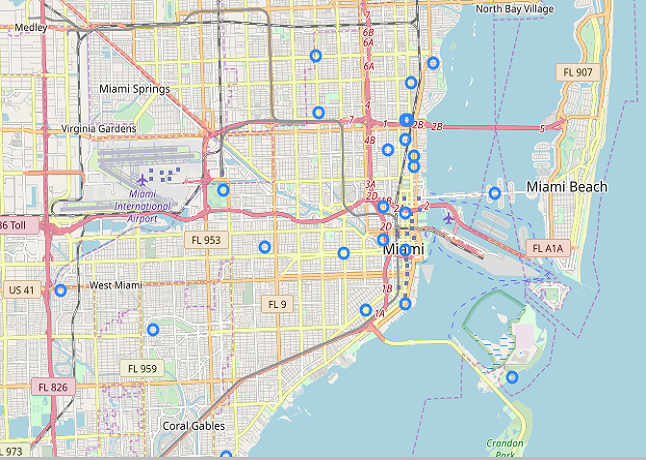
As the target is to find the best possible location of Gas Station, it is checked whether there is any presence of existing Gas Station and it is found there exists gas station within the list of venues. Then, each neighborhood is analyzed further using one-hot encoding and grouping the rows by neighborhood. The data frame is organized for further analysis by taking the mean on the frequency of occurrence of each venue category.

Finally, the K-means clustering is used to perform the clustering of the neighborhoods. The reason for using K-means clustering is because of its simplicity and conformity with this project. The neighborhoods of Miami are clustered into 3 clusters based on the frequency of occurrence for “Gas Station”. The ideal location for a Gas Station can be identified based on the result derived through the concentration of clusters.

**Results**

The folium library package is used to create the following map of Miami with the coordinates of neighborhoods of Miami. This is done to check whether all the coordinates are correct.

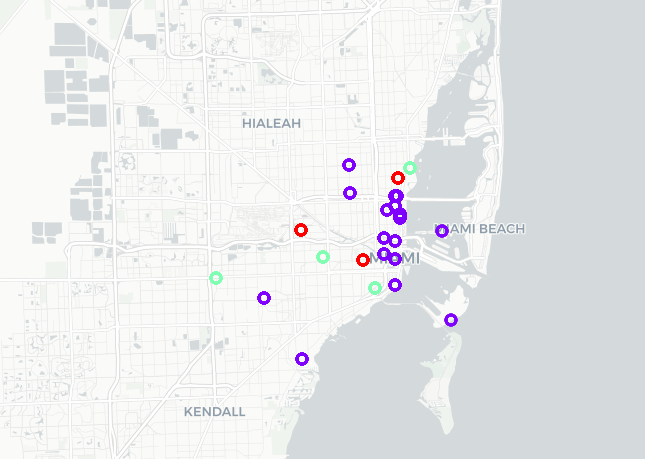
Image for post



Map 1: City of Miami with marked neighborhoods

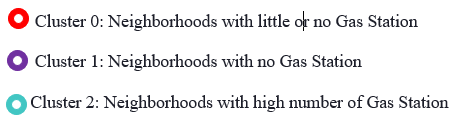
Running K-means clustering, the following map is created based on the frequency of occurrence for ‘Gas Station’.

Image for post



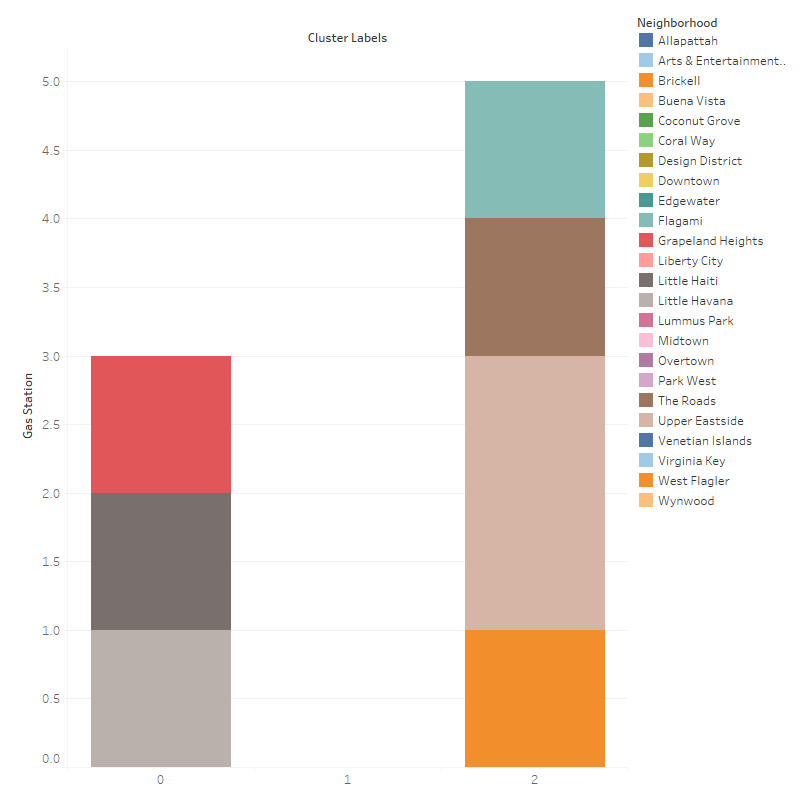
Map 2: City of Miami neighborhoods with clusters based on the frequency of Gas Station

Image for post



For further analysis, using Tableau, neighborhoods are visualized in the following chart based on the presence of ‘Gas Station’ and cluster labels.

Image for post



Graph 1: Bar Chart of Clusters showing the presence of Gas Station within the neighborhood

It shows that cluster 2 which contains the neighborhood ‘Upper Eastside’ has the highest number of the gas station. In the same cluster group, ‘Flagami’, ‘The roads’, ‘West Flagler’ neighborhood also has a gas station. From the cluster 0, neighborhoods like ‘Grapeland Heights’, ‘Little Haiti’ and ‘Little Havana’ also has a gas station.

Neighborhoods in cluster 1 have a vacuum of the gas station. This cluster include places like ‘Allapattah’, ‘Arts& Entertainment district’, ‘Brickel’, ‘Buena Vista’, ‘Coconut Grove’, ‘Coral Way’, ‘Design District’, ‘Downtown’, ‘Liberty City’, ‘Lummus Park’, ‘Midtown’, ‘Over town’, ‘Park West’, ‘Venetian Island’, ‘Virginia key’. These are the ideal places to set up a gas station.

**Recommendations**

As shown by the analysis done in this project, it is recommended to open a gas station within the neighborhoods which are included in cluster 1. Some neighborhoods from cluster 0 could also be good for setting up a gas station. However, this might depend on some other factors like population, distance and neighborhood size, etc.

**Limitations and Future Implications**

This project only considers the frequency of gas stations within the neighborhoods of Miami city for choosing the ideal location. However, several other factors might influence the recommendation like population size, the purchasing power of consumers, distance from the community, number and type of vehicle owned by the people in those neighborhoods, etc. considering those factors are beyond the scope of this project. Nevertheless, these factors could lay the foundation of newer insights for future research on this topic.

**Conclusion**

To conclude, this project uses Four Square’s data to pull out the venues from the neighborhood. However, the number of venues pulled out has limitations as the ‘Free tier’ is used for this project. If the ‘Start-ups’ or ‘Enterprise’ data can be used, it would produce a better result which might have an impact on taking more accurate decisions.