USA Restaurant Mania!

Introduction:

When a business is thinking about creating either a new chain of restaurants or adding more locations of an existing chain, it is a great strategic idea to know the lay of the land already for a given city or groups of cities. Knowledge on what types of restaurants are already present and the frequency of those in general is a key factor in making a decision. Additionally, understanding any patterns or relationships between the cities could help drive smart decisions on either which cuisines the company wants to introduce to those cities or which cuisines they should focus on for those closely related cities.

Business Plan:

The top 300 USA cities are listed on Wikipedia. However, it does not include the restaurant venue information along with any additional analytics such as the most common cuisines for each city. Further, one cannot easily see any relationships at all between the cities. If this information was available, the company could make more objective and informed decisions on where to create additional restaurants and what type of cuisine they should offer and serve.

The main purpose of this project is to create the dataset described above. First, a dataset will be created by web scraping the top 300 USA cities with their respective locations. Next, another dataset will be created listing the closest 100 restaurants from the each city center (if available) by leveraging the coordinates. Finally, a common clustering algorithm will be used to visually show relationships between the cities based upon the most frequently available cuisine types based upon the information found.

Data Description:

This section contains descriptions of the data that will be used to analyze the problem of determining where to create new restaurants and what type of cuisine should be severed. The data is to be collected from two main sources.

1. **Top 300 USA Cities by Population:**

   First the table is extracted from the website (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) by using a web scraping technique. This dataset includes the City name, the state, the population, the latitude and longitude coordinates.

2. **100 Closest Restaurants around each city:**

   In order to retrieve the closest named restaurants (up to 100) for each of the cities, the coordinates from the Top 300 USA Cities dataset are leveraged to trigger the Foursquare API for each city. The query is to bring back the results in a JSON file format. The results are then parsed further and reshaped in order to display the city, the restaurant name, type, and location for each data point.

3. **Clustering Results:**

   The above dataset described as 100 Closest Restaurants around each city is massaged and manipulated further in such a way to create the top 10 most frequent types of restaurants for each city. This dataset is then used to perform the k-means clustering algorithm in order to generate a new dataset containing the cluster assignments that will be used to visually establish relationships between the cities based upon the top 10 most frequent restaurant types. The visual distribution is displayed by using the USA Map with different colored dots representing how the cities are related to one another based upon restaurant cuisine type distributions. The resulting clustering results dataset will contain the City name, the cluster it belongs to, and the top 10 restaurant types. Deeper analysis of this dataset will provide insight on both how similar each city is to one another and what the top 10 most frequent restaurant types there are within or around the city.