**Battle of Cities**

**Viswanathan Appakkudal Ramani**

**Coursera Learner**

**Battle of Cities**

This note is part of the IBM Applied Data Science Capstone project. This course is Part-4 of 4-Course Specialization. The capstone project involves using Python and libraries such as numpy, pandas, matplotlib, seaborn, and folium. The Project is 'Battle of Cities'.  
Lets say, you want to relocate to another city, due to work reasons. You adore your current city, and lounge for same/similar city in your new State. This project involves finding the right city using the location data offered by Foursquare API.

# Introduction

I am a city person, and I want to explore new cities. However, due to work reasons, I allowed to relocate only among these 4 states (New York, New Jersey, Connecticut, and Pennsylvania). I want to find a city of my choice to relocate. What would it be?  
Let me list out, what I expect a city to-have and not-to-have:

1. Coffee
2. Convenience stores
3. Restaurants
4. Food trucks / Fast food joints
5. No/Limited Pubs and Bars

## Data

In order to get the list of cities in New York, New Jersey, Connecticut, Pennsylvania, we use the wiki page <https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population>. The wiki page lists all the cities in Unites States. We will take only the cities in New York, New Jersey, Connecticut, and Pennsylvania.  
We will use Foursquare API to get trending venues in each city. Then, we will compare my requirements with the output we get after performing data cleaning to get a top city.

1. **Methodology**

In order to perform this project, the following aspects were done:

1. Web Scraping

#### Data Cleaning

1. Feature Engineering
2. Data Visualization
3. Data Analytics
4. Modeling

**3.1) Web Scraping.**

##### Web scraping is the process of collecting structured web data in an automated fashion. Python has Beautful Soup library to perform web scraping. I am considering 4 states in United States (New York, New Jersey, Connecticut, and Pennsylvania). I need to extract all the cities from these states. We could get list of all stats form "[https://en.wikipedia.org/wiki/List\_of\_United\_States\_cities\_by\_population"](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population%22).

##### **3.2) Data Cleaning**

After scraping data from wiki page, we need to clean data. The data had unwanted texts such as super-scripts, special characters, which ought to be removed. The project requires only the names of States, and Cities, and hence we remove all other columns from the dataframe. Since, our target is only four states, we keep only the cities within the states New York, New Jersey, Connecticut, and Pennsylvania.

**3.3) Feature Engineering**

To perform data analytics, we also require latitude and longitudes. We use geocoder library to get latitude and longitudes of every city from the dataframe. We concatenate the emerging location dataframe with our original dataframe.

**3.4) Data Visualization**

After getting the latitudes and longitudes, we visualize the locations on a map using Folium library. We use circle markers to locate the cities, and popups to display the name of the city. It gives a rough idea, of where the cities are located, and also gives a way through which we could assume what could essentially be popular in a location. For instance, if the surrounding area of the map is more vegetative, we can assume that it might be popular on groceries, farmer’s market. On the other hand, if it has the least vegetative cover, then, we can assume that it would be more commercial area.

**3.5) Data Analytics**

We use the Foursquare API to get list of 100 trending venues in every city. We make a new dataframe with cities and their corresponding popular venues. We perform one hot encoding and find the mean occurrence of every venue at a given location through which we can find, what’s the most trending in each city. We then pick the top 10 trending locations based on the mean scores and make a table of the city and top 10 trending venues.

**3.6) Modeling**

Then, we pick my favorite venues, and perform weighted voting. If city have the locations I prefer, they get higher points, and if the city has venues that I dislike, they get negative points. And if the venue, that I like is the 1st trending venue, it gets the highest point, and scores decreases cumulatively, as the number on trending venue decreases. Finally, we get the total scores of every city and infer that, the one with maximum score, is my favorite location.

1. **Results**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **1th Most Common Venue** | **2th Most Common Venue** | **3th Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** | **6th Most Common Venue** | **7th Most Common Venue** | **8th Most Common Venue** | **9th Most Common Venue** | **10th Most Common Venue** | **Total** |
| Elizabeth | Pizza Place | Discount Store | Coffee Shop | Theater | Burger Joint | South American Restaurant | Peruvian Restaurant | Lounge | Fried Chicken Joint | Latin American Restaurant | 17.5 |
| New York City | Coffee Shop | Sandwich Place | Hotel | Gym | Burger Joint | Cocktail Bar | Italian Restaurant | Plaza | Pizza Place | Park | 17.5 |

From the analysis, we find that two cities New York City, and Elizabeth are awarded with the same score. It’s an interesting result, as one city has high population density, while the other has lower density. However, the venues match out.

From the above table we could see that, New York City’s most trending place is Coffee. Hence, I would choose NYC over Elizabeth.

1. **Conclusion**

Thus, from the analysis stated above, I could very well say that New York City will be my next stop to relocate to. It has almost everything, that I desire. Be it, Coffee shops, Restaurants, Pizza joints, it looks like a perfect destination for living.

##### .