THE BATTLE OF NEIGHBOURHOODS

BY –ZAK

Introduction and Business Problem

A CEO of a company is interested in starting a gaming arcade in the best locality of all the cities in United states. He defines a best locality based on the following constraints,

- Population density of a locality
- Per Capita Income
- Population of each location
- Venues in each locality

We have to suggest the best locality to setup a gaming arcade in the United States.

Data

- List of all the cities in United States with population density and coordinates:
  https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population
- List of all the cities in United States with Per Capita Income :

https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income

- Using Foursquare API to get the following
  - List of all venues in each city
  - List of all venues in each locality in the selected city

Methodology

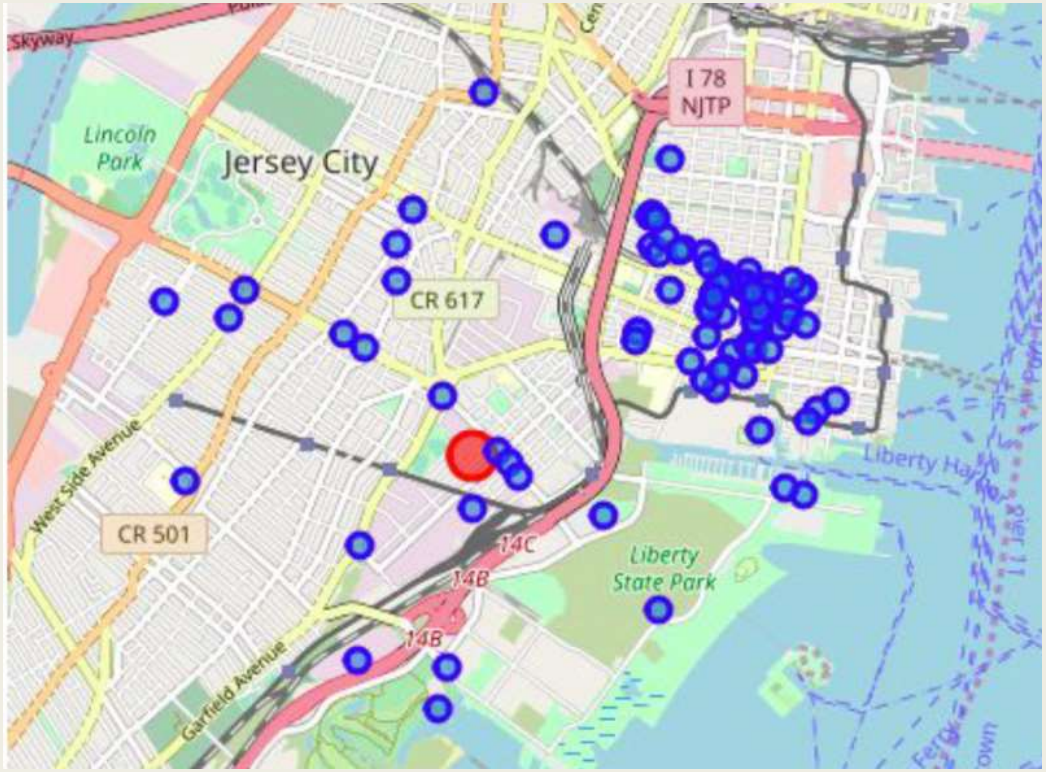
In order to do the analysis and suggest the best location, following steps were followed:

- The Wikipedia page (https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population and https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income) was scraped using the BeautifulSoup library to build a pandas dataframe listing the cities, states, coordinates, area, per capita income and population density. The data frame was cleaned and processed appropriately.
- The Foursquare API is then used to get the venues in each city of United State, based on the categories of each venue as decided by the CEO, we have assigned weights to each of them and got the city that has the maximum weight.
- We will now use K means to cluster the venues based on the category and get the coordinates of the cluster that has maximum weight which is also our preferred location to setup a gaming arcade.

Plot of all cities of USA that were extracted:

Blue dot indicate the city canter in the country map of United States.

Map of venues in the Jersey City:



Based on the constraints we have chosen, we found that Jersey city would be the better place to start.

Result

The circle indicates the *best place* to start an arcade in the **Jersey City**

Recommendation/ Improvements:

In the Foursquare API, we have queried the Venues of a locality by specifying the LIMIT and Radius of our choice. We have chosen less LIMIT as the number of API calls that can be done using a free account in Four Square are less.

- We can increase the limit for more accurate results.
- We can increase the Radius for more venue results from each city.

In the venue categories we are choosing only few out of 2000 that are available to give weights and identify the best cluster. Hence, assigning weights must be done relatively for each category and then considering more number of venue categories would actually yield a better output.