THE BATTLE OF NEIGHBORHOODS

Applied Data Science Capstone Project

BY: M JAWAD BASHIR

Introduction

As more than half of the world's population lives in urban areas, and the proportion is expected to increase to 70 percent by 2050.

Incomes are five times higher in the more urbanized countries and infant mortality rates are less than a third in the more urbanized countries.

Urbanization is crucial to generate employment, wealth and productivity growth, and drive national economic development. Here, by comparing the Foursquare Venue Category data of high population density cities to low density cities, critical features might emerge and shed light on the direction of city development.

Data Acquisition

The population density data of US cities could be obtained from governing website (https://www.governing.com/gov-data/population-density-land-area-citiesmap.html). This data contains the population density (persons in square miles), population in 2016, and land area (in square miles) for 754 US cities.

- Latitude and longitude coordinates of these cities
- All the venues surrounding the geographic coordinates of these cities, and the
- venue data would be the Venue Categories, such as ATM, Accessories Store and etc.
- The detailed list can be found on the foursquare website (https://developer.foursquare.com/docs/resources/categories). These information will be collected from Foursquare API.

Methodology

• **Acquire population density data of US cities:**

Web scraping the data of US cities from governing website using the Python library BeautifulSoup and requests, and convert the data to a Pandas dataframe.

• **Visual representation of population density:**

Use Pandas bar plot to present the population density of US cities.

• **Acquire the geographic coordinate of the US cities:**

Use Python library GeoPy to obtain the latitude and longitude coordinates of these US cities, and add the coordinates to the dataframe.

• **Spatial visualization of population density of US cities:**

Use Python library Folium, mark the US cities on US map and color the markers based on the population density of the city to generate geographical insights from the population density data.

• **Collect the Nearby Venues:**

Collect the nearby venues based on the geographic coordinate of US cities using Foursquare API, and then convert the categorical data to venue composition data with one hot encoding.

• **Correlation between the venue frequency and population density:**

Use Python library Pandas dataframe.corr() function to find the pairwise correlation of venue composition and population density.

Results

Figure 1. Basic statistics of population density data from governing website.

Mean, standard deviation, minimum, first quartile (Q1), median, third quartile (Q3), and maximum of population density (persons in square miles), population in 2016, and land area (in square miles) of the 754 US cities.

	Population_Density	Population	Land_Area
count	754.000000	7.540000e+02	754.000000
mean	4242.729443	1.646172e+05	55.015915
std	4323.792554	3.973563e+05	95.695024
min	172.000000	5.007700e+04	1.000000
25%	2076.000000	6.417050e+04	19.000000
50%	3128.500000	8.669450e+04	31.500000
75%	4720.000000	1.380125e+05	54.750000
max	54138.000000	8.537673e+06	1705.000000

Figure 1.

Figure 2. Histogram of population density.

(A) The distribution of population density in 10 bins.

(B) The destribution of cities with population density between 0 to 10000 in 100 bins (100 people per bin).

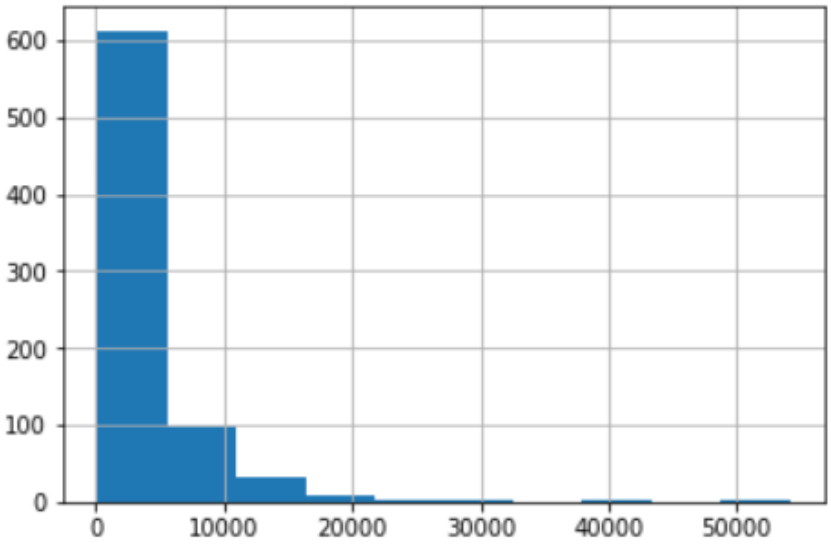

Figure 2-A

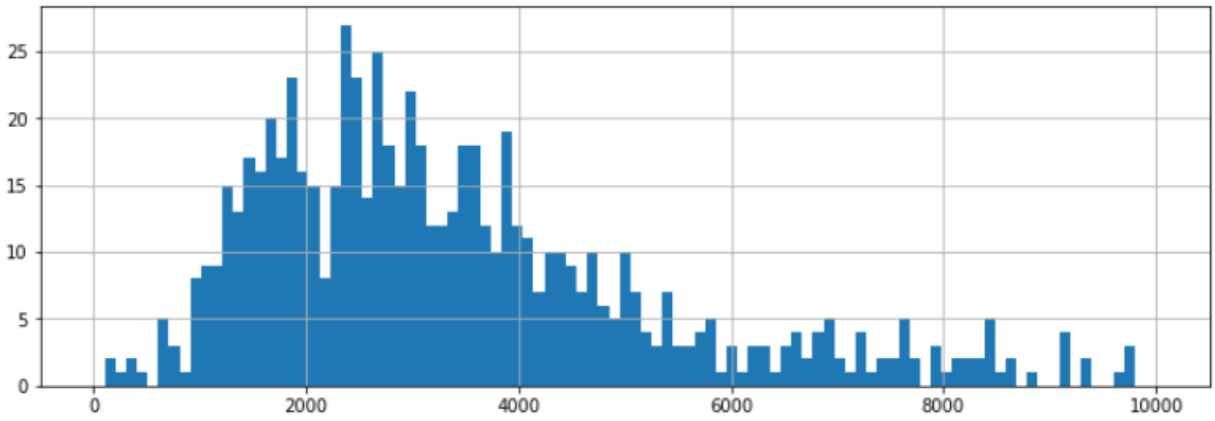

Figure 2-B

Figure 3. Spatial visualization of population density of US cities.

The population density of the city is correlated to the color intensity of the marker. Higher population density is indicated by a darker shade of red.

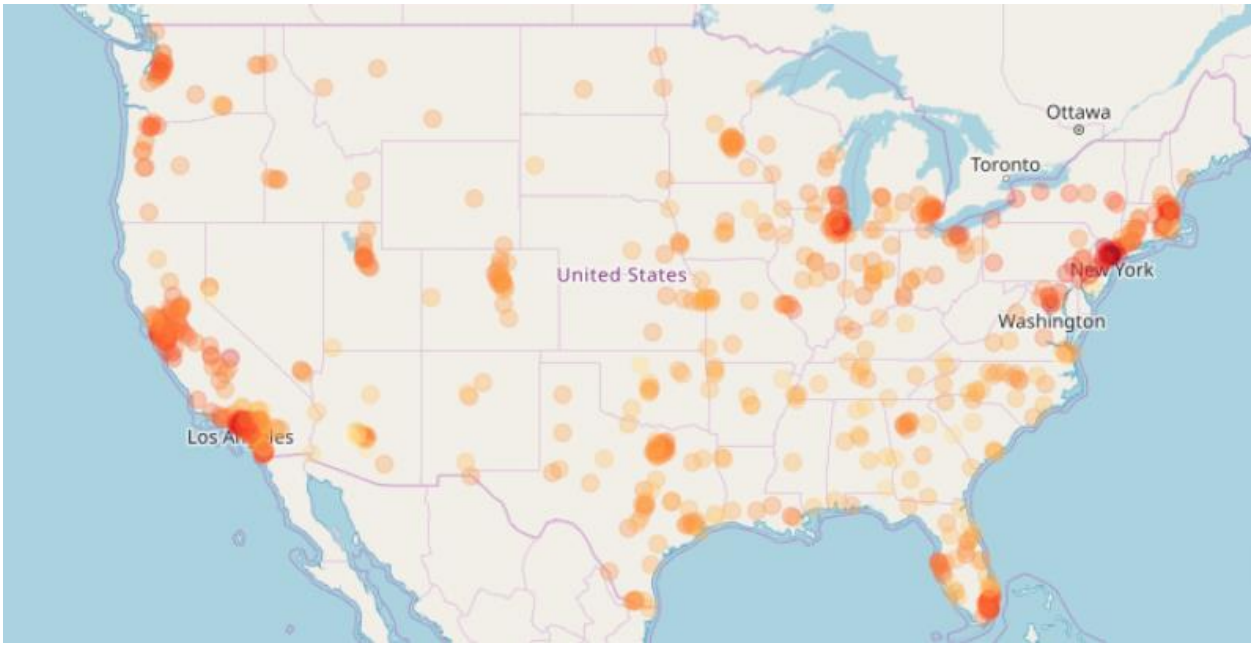

Figure 3.

Figure 4. Correlation between venue categories and population density.

(A) Top 10 venue categories that are positively correlated with population density.

(B) Top 10 venue categories that are negatively correlated with population density.

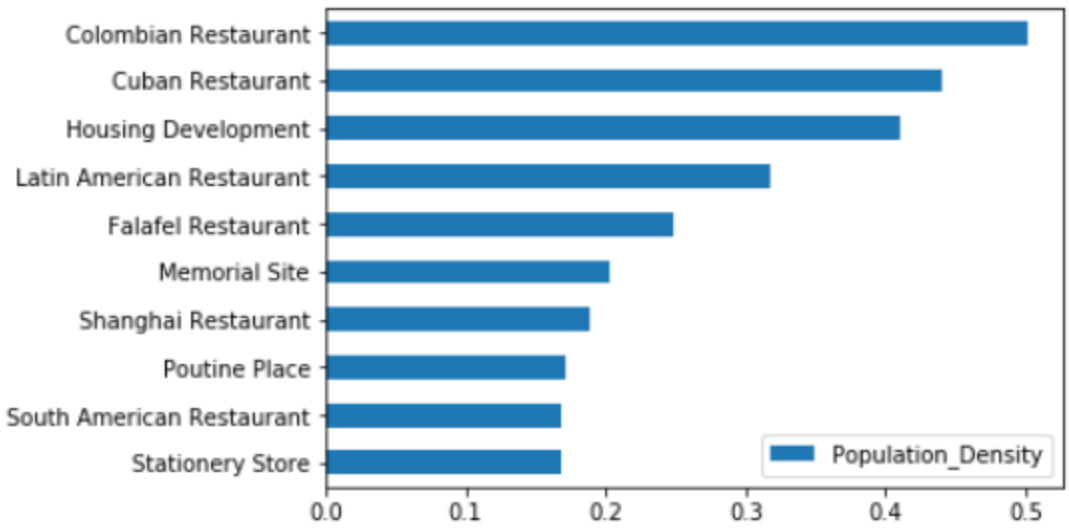

Figure -4 (A)

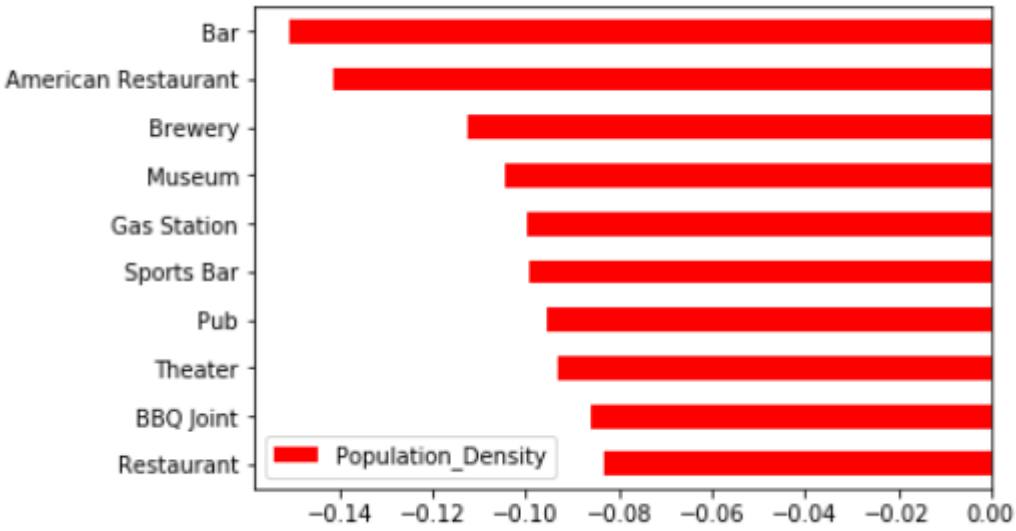

Figure-4 (B)

Conclusion

The diverse restaurant types that are positively correlated with population density come from different cultures, indicated population diversity. Population diversity has been known to lead to economic growth, this might explain why urbanization is connected to economic development. The accuracy suggests the composition of venues can partially explain the population density, but other components might also exist.