IBM Data Science Professional Certificate

Capstone Project - The Battle of Neighborhoods

Teng-Kuei Hsu

May, 2019

INTRODUCTION

According to United Nations report, more than half of the world's population lives in urban areas, and the proportion is expected to increase to 70 percent by 2050 (1). Critically, economists and urbanists have found the connection between urbanization and economic development. Harvard University economist Edward Glaeser points out, "incomes are five times higher in the more urbanized countries and infant mortality rates are less than a third in the more urbanized countries." (2)

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-cities-map.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.

- **Predict population density with Machine Learning approaches**

  1. Use Python library sklearn to apply K Nearest Neighbor, Decision Tree, and Logistic Regression to predict the population density based on the composition of venues.
  2. The cities were classified as high population density (>6000 persons in square miles) and low density (<2000 persons in square miles).
  3. Data will then be split into training (80%) and testing set (20%).
  4. Results were evaluated based on accuracy score and F1 score.

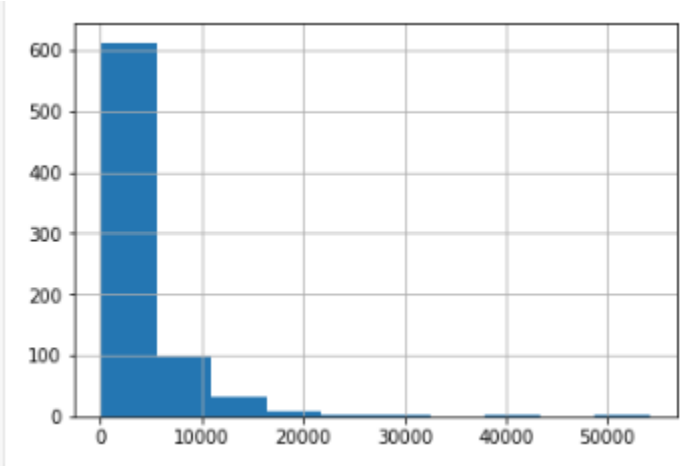
RESULTS

- **Figure 1. Basic statics 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.

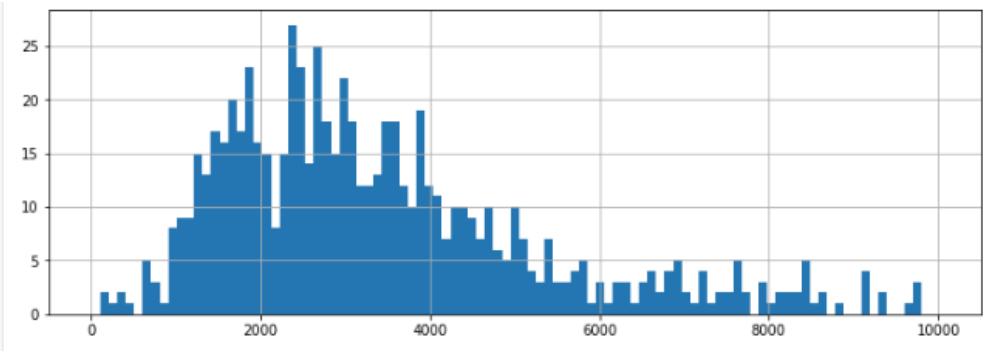
	Population_Density	Population	Land_Area
count	754.000000	7.540000e+02	754.000000
mean	4242.729443	1.646172e+05	55.015915
std	4323.792554	3.973563e+05	95.695024
min	172.000000	5.007700e+04	1.000000
25%	2076.000000	6.417050e+04	19.000000
50%	3128.500000	8.669450e+04	31.500000
75%	4720.000000	1.380125e+05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

- **Figure 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).

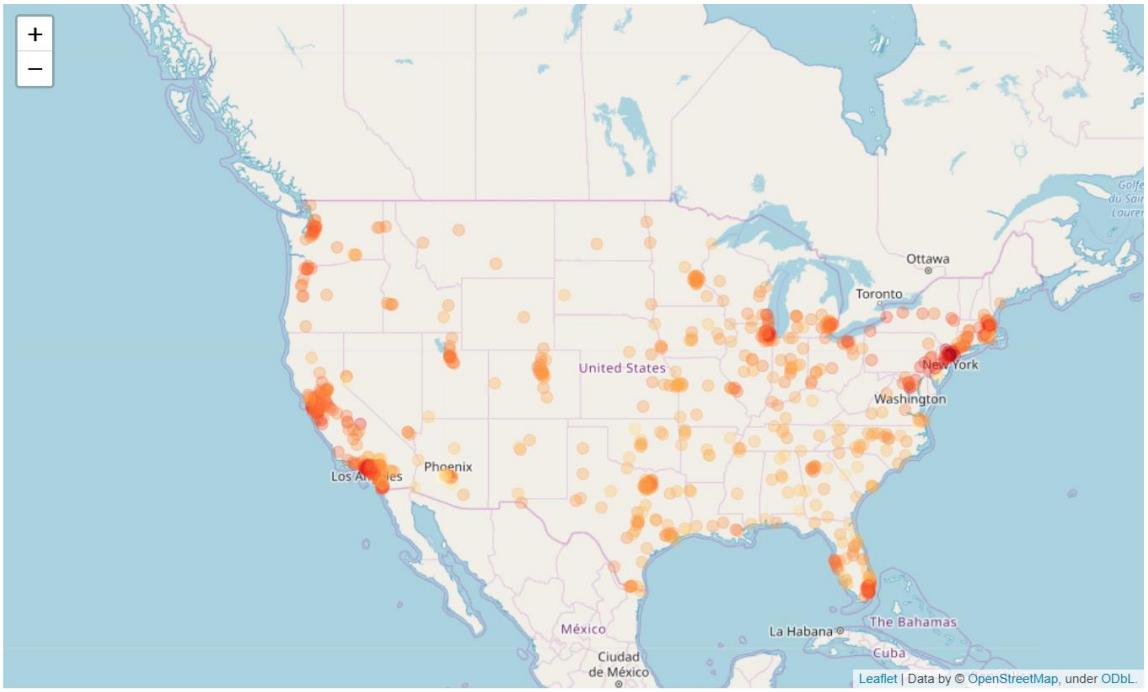
(A)



(B)

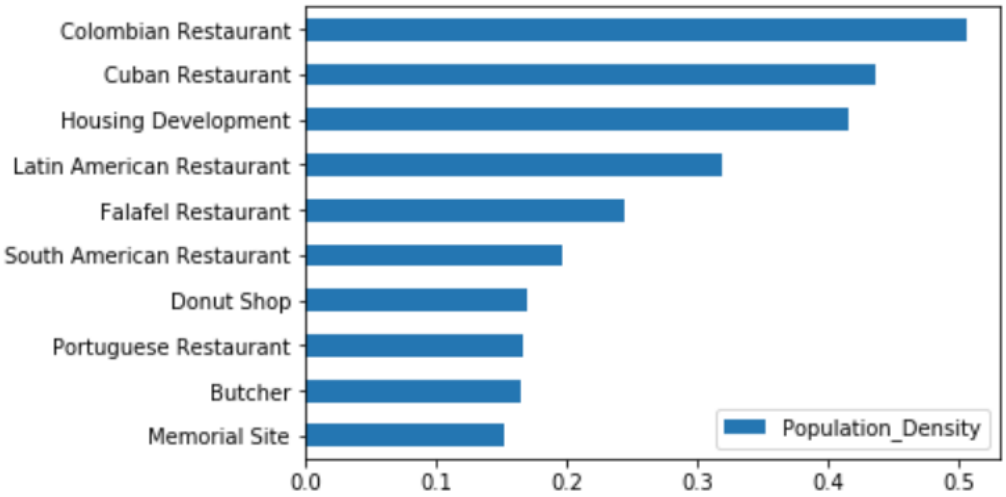


- **Figure 3. Spatial visualization of population density of US cities.** The population density of the city is correlated to the color intensity of the marker. Higher population density is indicated by a darker shade of red.

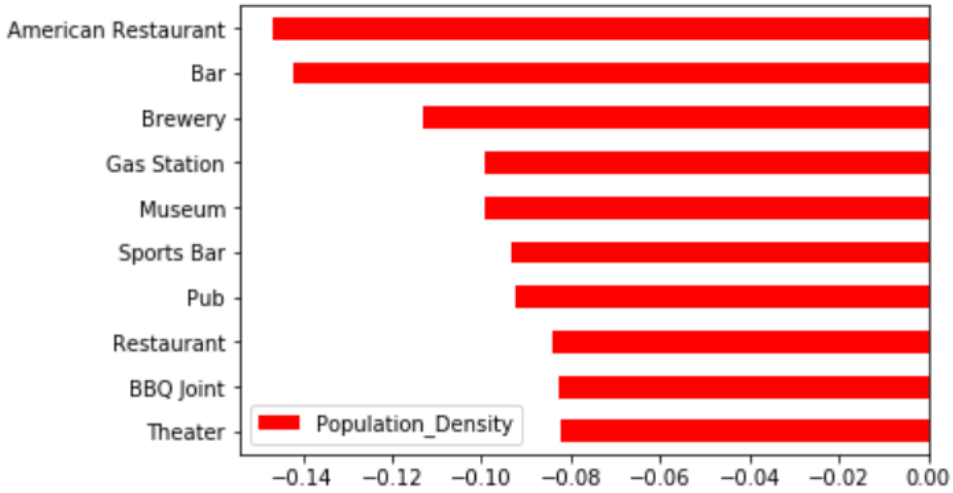


- **Figure 4.** Correlation between venue categories and population density. (A) Top 10 venue categories that are positively correlated with population density. (B) Top 10 venue categories that are negatively correlated with population density.

(A)



(B)



- **Figure 5.** The accuracy and F1-score of K Nearest Neighbor, Decision Tree, and Logistic Regression in predicting the population density based on the composition of venues.

	Accuracy	F1-score
KNN (Best K=2)	0.733	0.714
Decision Tree	0.767	0.696
Logistic Regression	0.667	0.643

DISCUSSION

In this study, I analyzed the relationship between population density and the composition of venue categories. By analyzing the correlation between venue composition and population density, I found the venues such as “Colombian Restaurant”, “Cuban Restaurant”, “Latin American Restaurant”, “Falafel Restaurant” are positively correlated with population density, while the “American Restaurant” is negatively correlated. The diverse restaurant types that are positively correlated with population density come from different cultures, indicated population diversity.

The population diversity encourages innovation as different people look at things from varying perspectives. Diversity also stimulates intellectual growth since a multicultural learning environment would enable people to interact and learn from people that are different from them. Critically, population diversity has been known to lead to economic growth, this might explain why urbanization is connected to economic development.

Also, by applying machine learning approaches including K Nearest Neighbor, Decision Tree, and Logistic Regression, I found the population density can be predicted based on the composition of venues in the cities, with accuracy ranging 0.6 to 0.8. The accuracy suggests the composition of venues can partially explain the population density, but there are other components might be involved in.

CONCLUSION

In conclusion, these data suggest the population diversity is positively correlated with population density in the US cities, which reflected on the diversity of restaurants in high population density cities. Since the population diversity leads to economic growth, this might partially explain how urbanization is connected to economic growth.

REFERENCE

1. 2018 Revision of World Urbanization Prospects. United Nations Department of Economic and Social Affairs (2018) https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html
2. Cities are making us more human. The European Magazine (2011) https://www.theeuropean-magazine.com/420-glaeser-edward/421-humans-cities-and-the-environment