IBM Data Science Professional Certificate

# Capstone Project - The Battle of Neighborhoods

Yash Borade

FEB , 2020

# 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.
- Critically, economists and urbanists have found the connection between urbanization and economic development.
- 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.

# INTRODUCTION

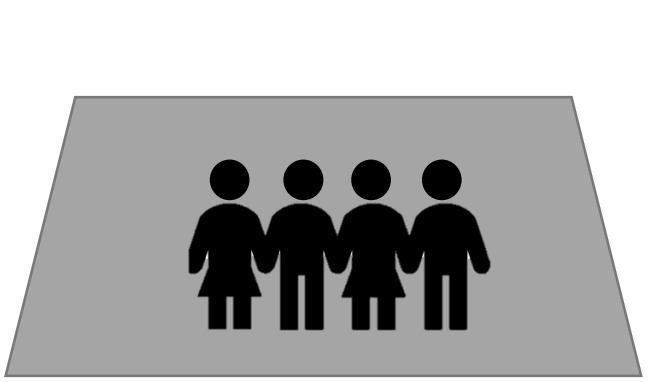
- According to United Nations report, more than half of the world's population lives in urban areas, and the proportion is expected to increase to 70 percent by 2050.
- Critically, economists and urbanists have found the connection between urbanization and economic development.

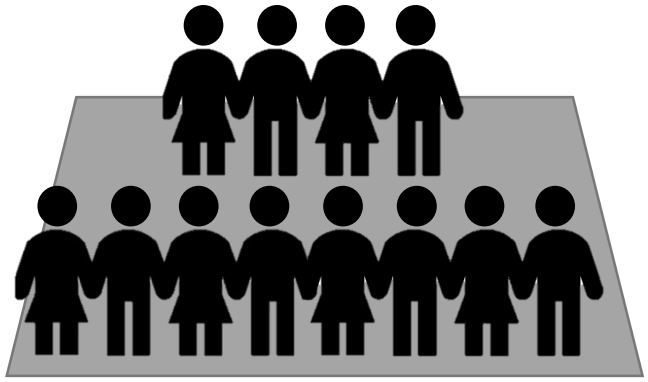
UNITED NATIONS
DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS

https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html

# INTRODUCTION

- 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).

# DATA ACQUISITION

- All the venues surrounding the geographic coordinates of these cities will be collected from Foursquare API.

- The detailed list of venue categories can be found on the foursquare website . https://developer.foursquare.com/docs/resources/categories

# METHODOLOGY

- **Acquire population density data of US cities:**
  - Python library BeautifulSoup and requests
- **Acquire the geographic coordinate of the US cities**
  - Python library Geopy
- **Spatial visualization of population density of US cities**
  - Python library Folium
- **Collect the Nearby Venues**
  - Foursquare API

# METHODOLOGY

- **Correlation between the venue frequency and population density:**
  - Pandas.dataframe.corr()
- **Predict population density with Machine Learning approaches**
  - Python library sklearn K Nearest Neighbor, Decision Tree, and Logistic Regression
  - Evaluated by accuracy score and F1 score.

# RESULTS

| | 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. Basic statistics of population density data from governing website

# RESULTS

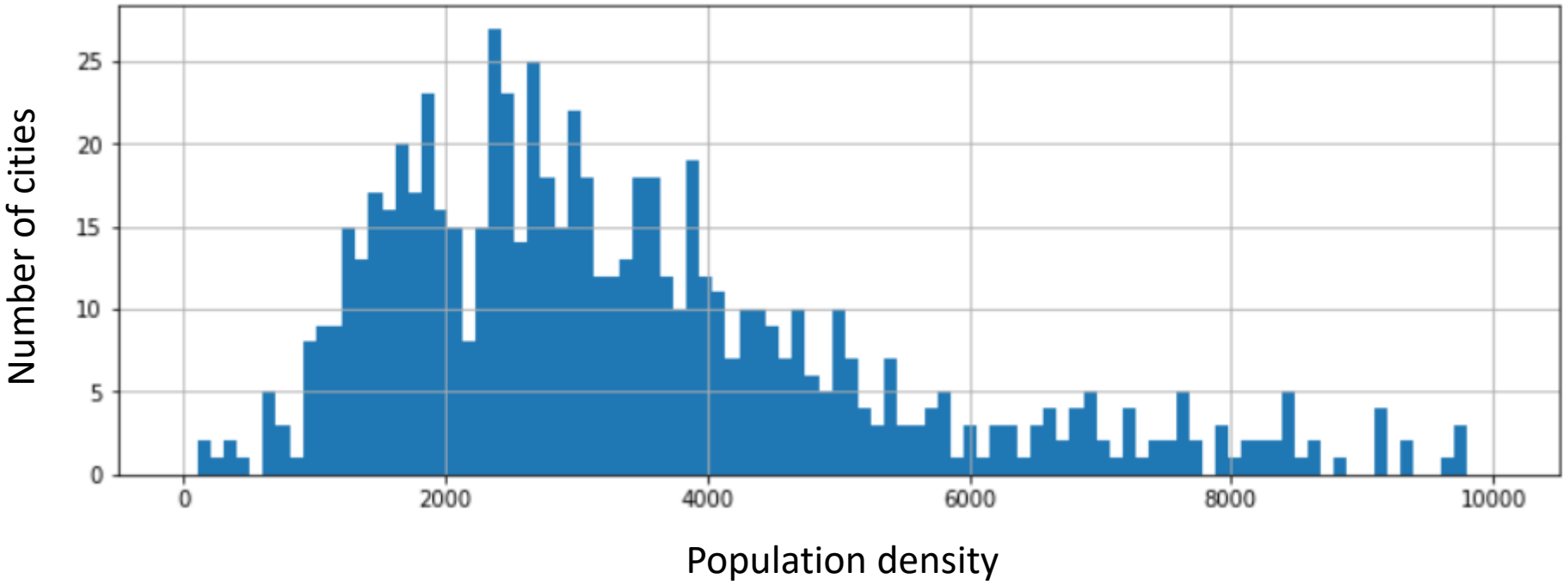



Figure 2. Histogram of population density (0-10000 persons in square miles).

# RESULTS

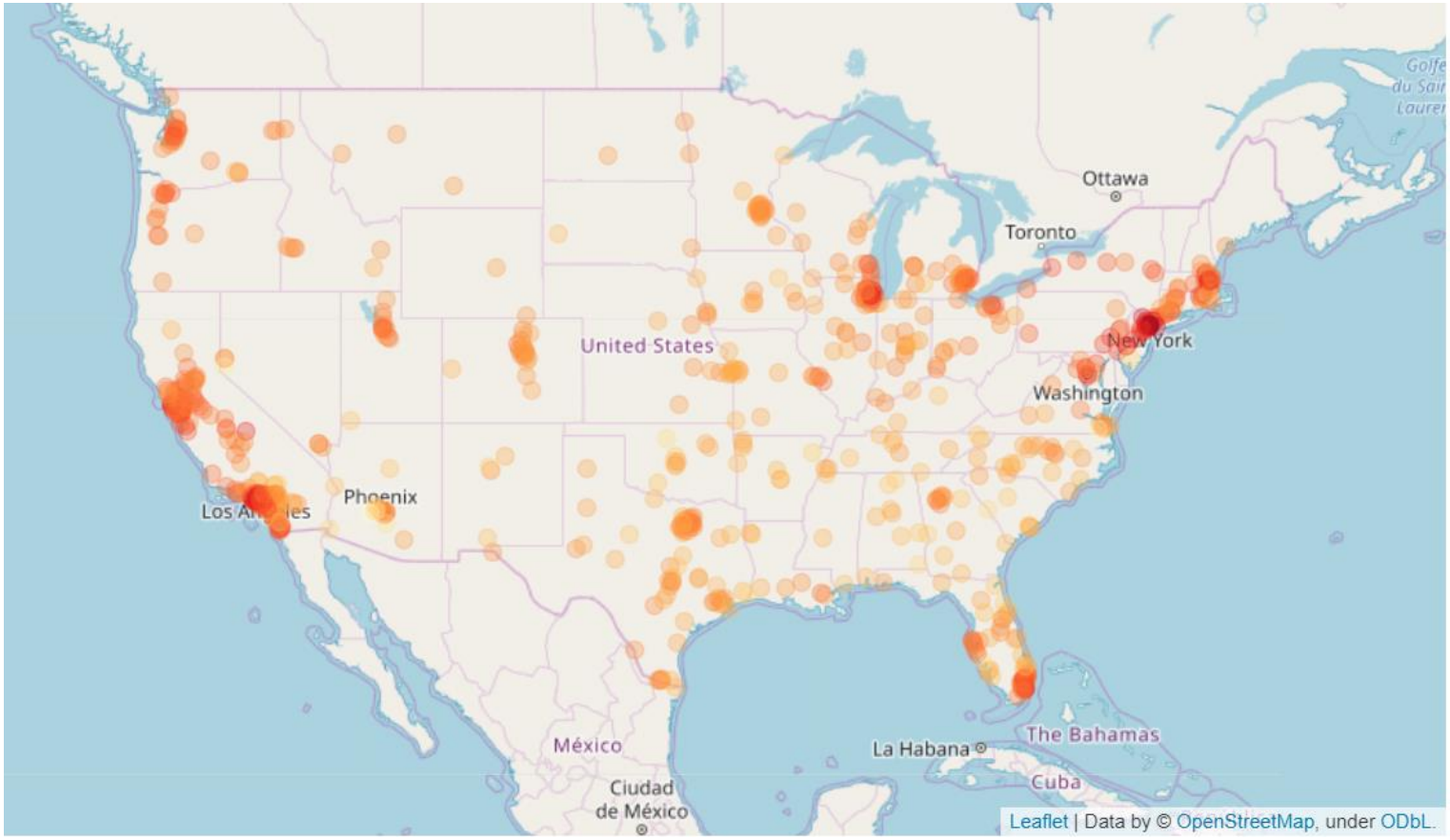



Figure 3. Spatial visualization of population density of US cities.

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.

# RESULTS

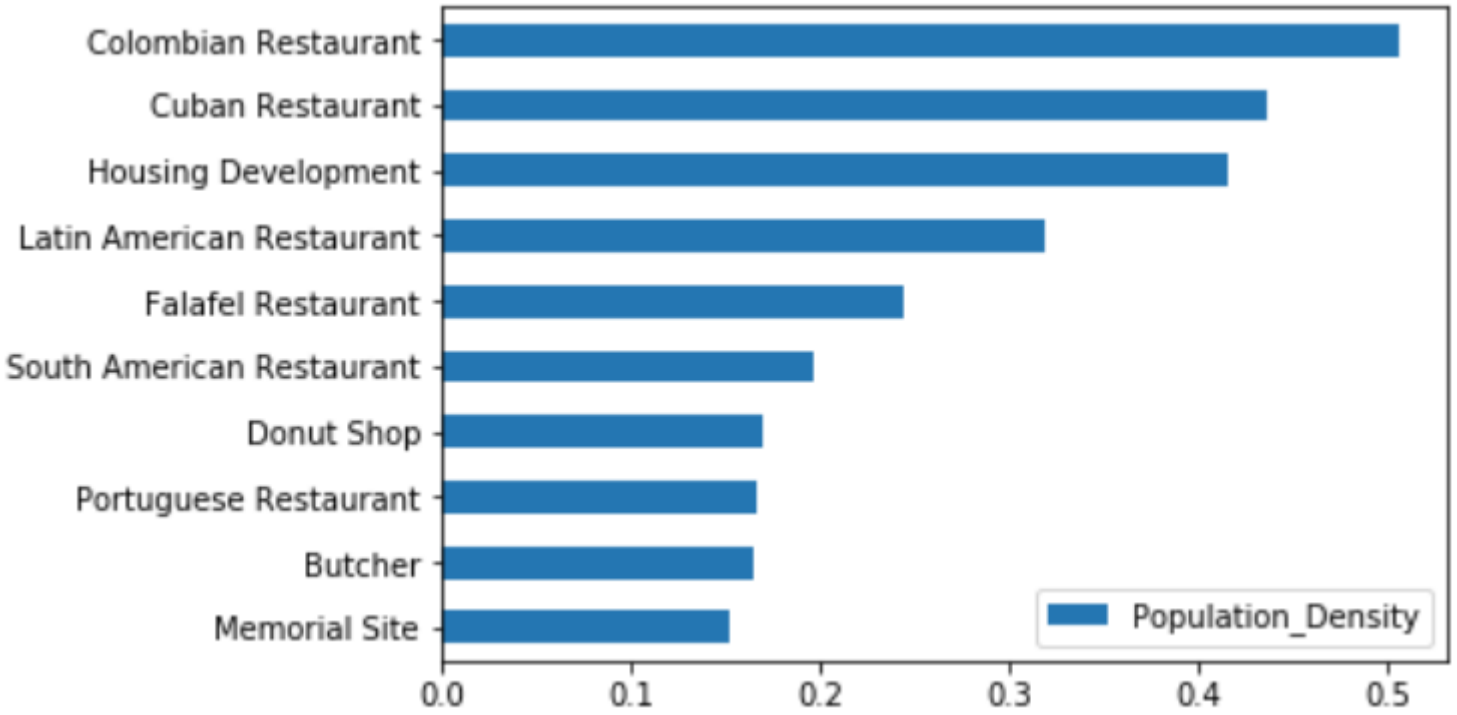



Figure 4. (A) Top 10 venue categories that are positively correlated with population density.

# RESULTS

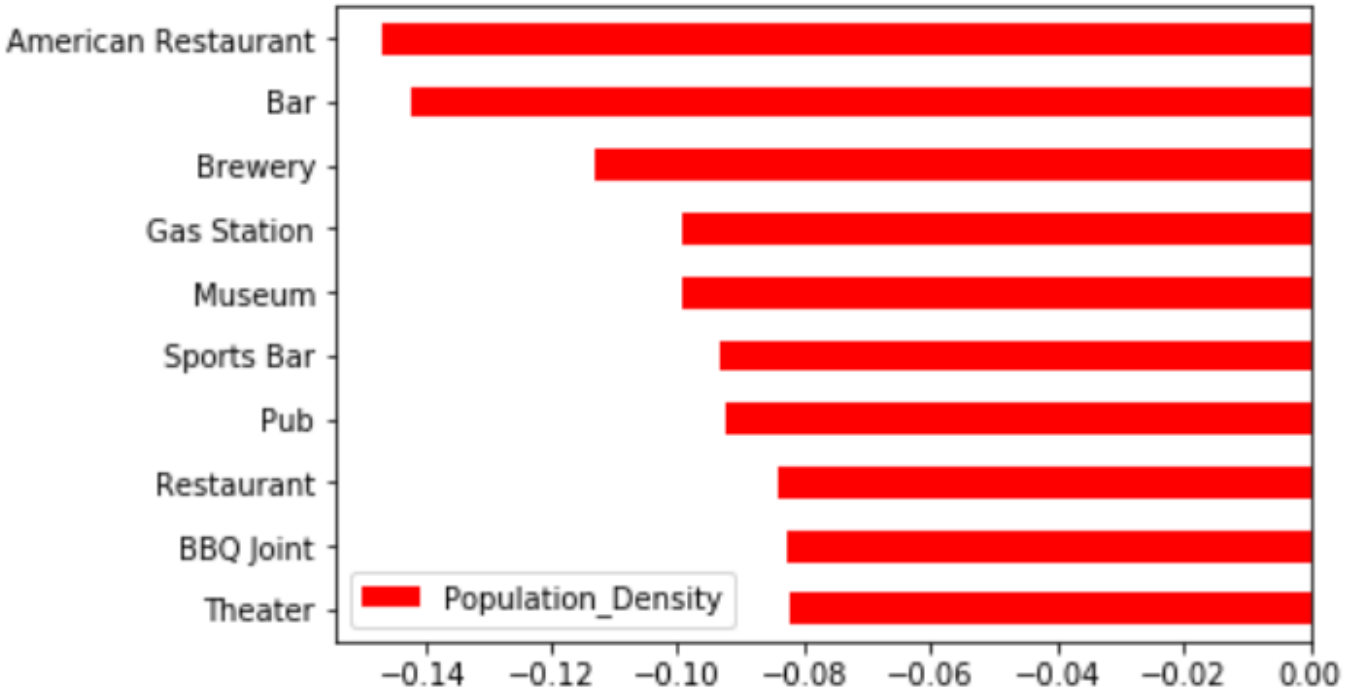



Figure 4. (B) Top 10 venue categories that are negatively correlated with population density.

# RESULTS

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

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.

# DISCUSSION

- 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.

# DISCUSSION

- The accuracy suggests the composition of venues can partially explain the population density, but other components might also exist.

# 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.