Predicting Rent Values in Atlanta

Alon Baruch - IBM Data Science Professional Certificate

Business Problem

Using data on the venues present in different neighborhoods of Atlanta, GA combined with average rent of each neighborhood, I will train a classification model to predict rent prices of a neighborhood. This problem will require the use of Foursquare API to get local venues as well as data about rent and neighborhood location which will be acquired through various sources. This report should interest developers and landlords wondering where to build or how much to charge for rent as well as renters who are looking for a place to rent and are interested in how much is fair rent. Finally this report will also interest city officials who will be able to gain insight on certain venues' effects on neighborhood rent.

Data Acquisition

Data Sources

- Average Rent by Neighborhood: RentCafe
- Location Data for Neighborhoods: Geopy
- Venue Information and location : Foursquare API

Data Processing

Step 1) Scraping RentCafe for average rent by neighborhood

Step 2) Using Geopy to find location data for each neighborhood

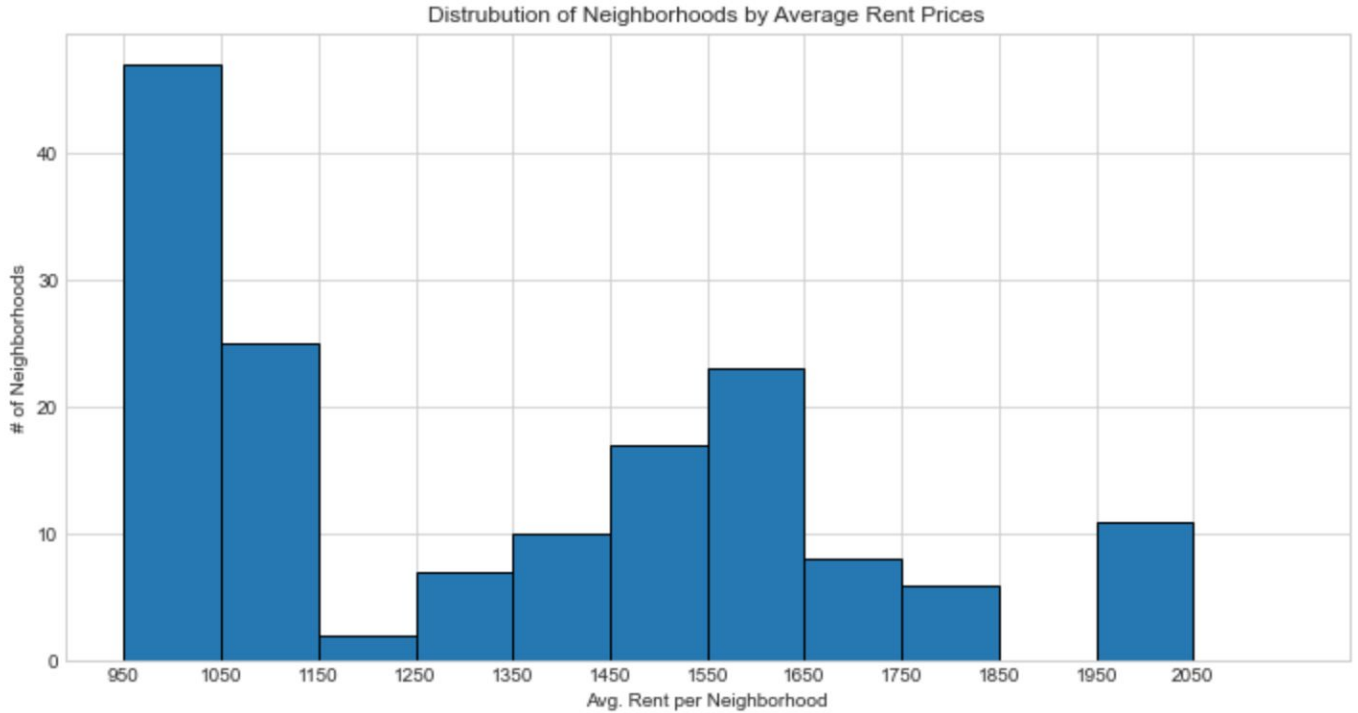
Step 3) Remove missing/incorrect location data using outlier detection

Step 4) Bin Neighborhoods into 4 groups based on rent values

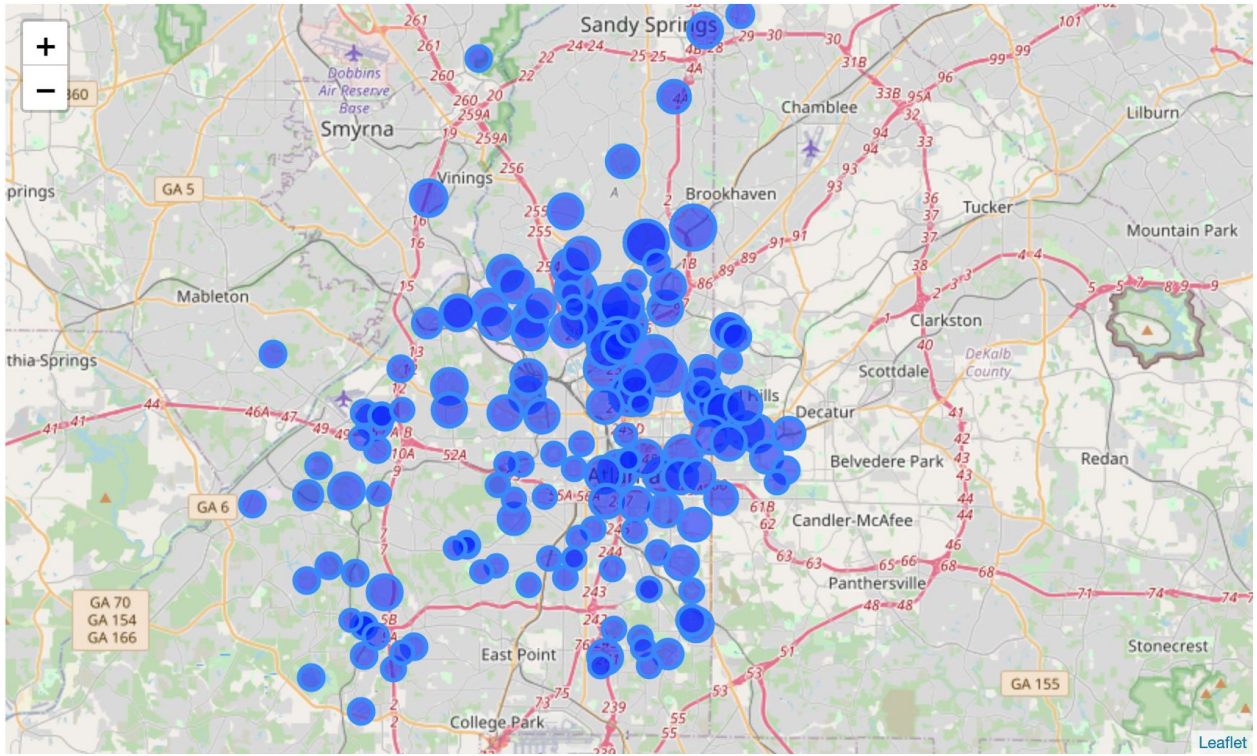
Step 5) Use Foursquare API to get venues in each Neighborhood

Step 6) Calculate venue type frequency per neighborhood

Exploratory Data Analysis

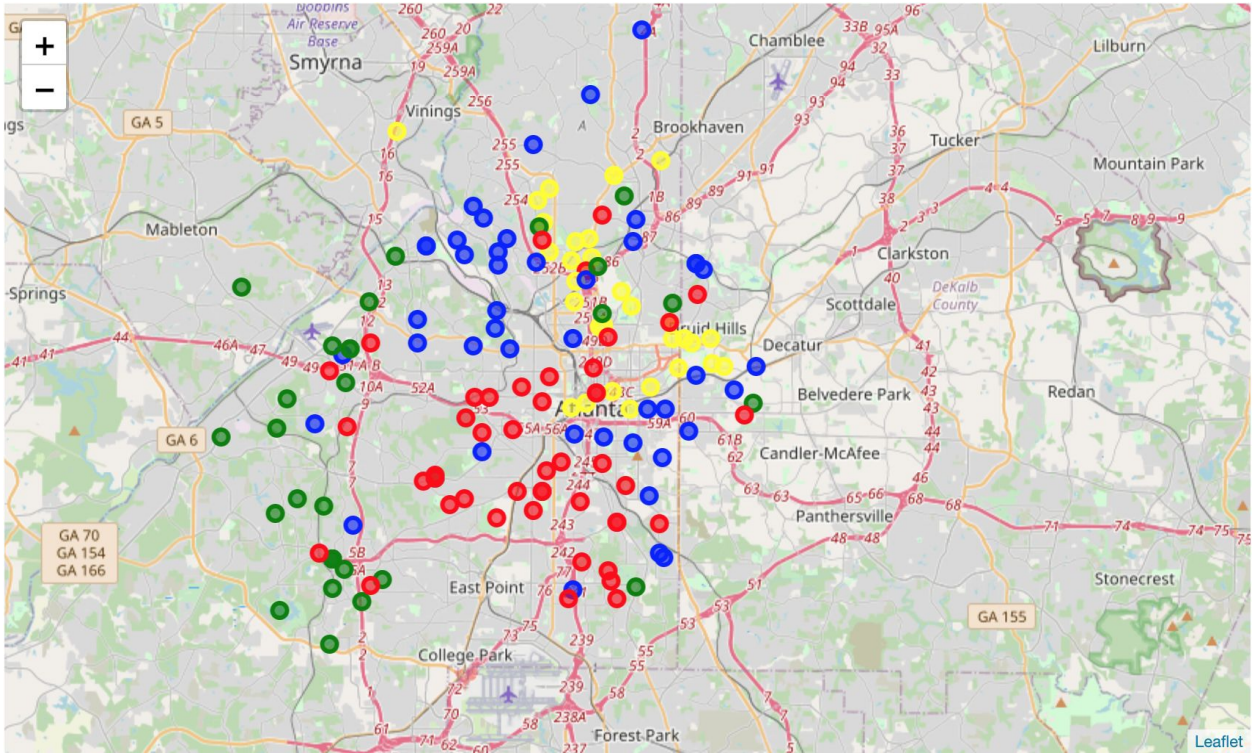


Exploratory Data Analysis



Each bubble represents a neighborhood with the size of the bubble being proportional with rent prices in neighborhood over average rent in the city

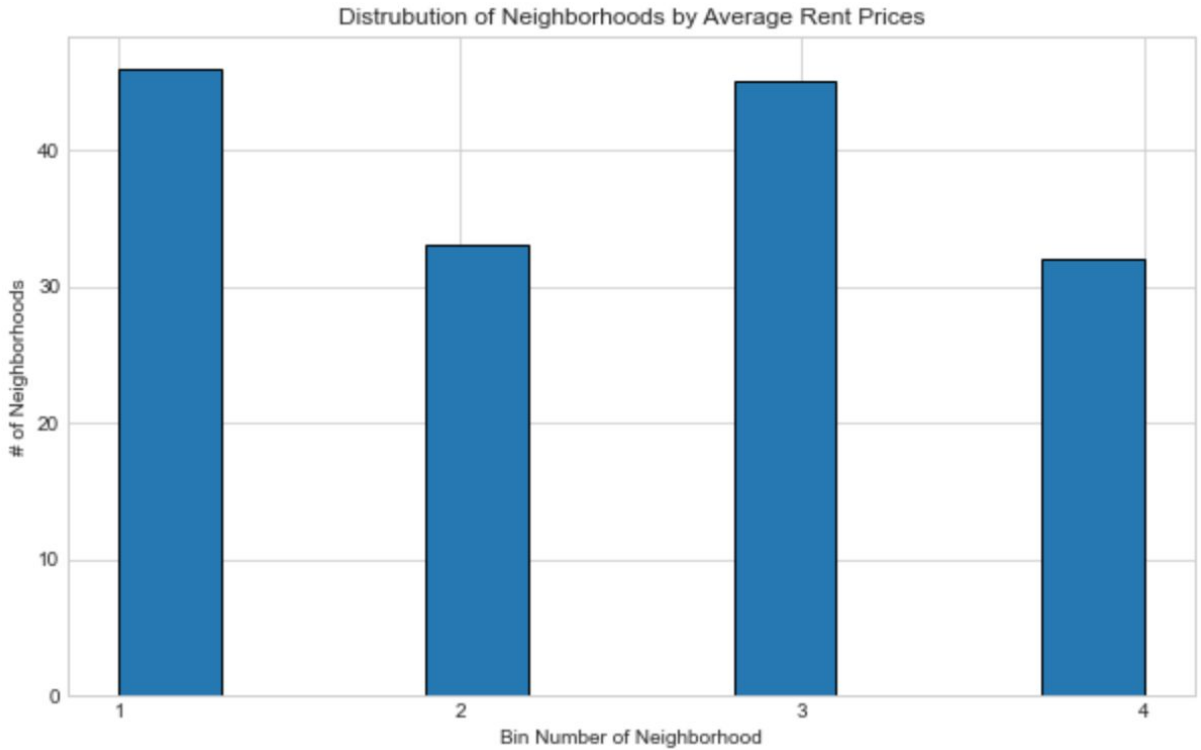
Exploratory Data Analysis

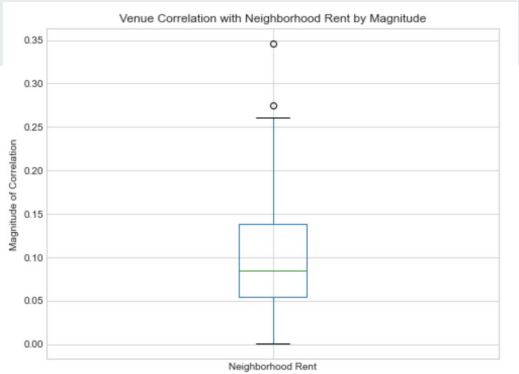


Each bubble represents a neighborhood with the color of the bubble representing which bin the neighborhood belongs in

Yellow = Most expensive
Blue = 2nd Most expensive
Green = 3rd Most expensive
Red = Least Expensive

Exploratory Data Analysis





Feature Selection

In order to select which variables will be used in the classification model I will find the magnitude of correlation between every venue type and the rent per neighborhood keeping only the top 25% of them.

```
rent_corr.describe()
```

```
count    281.000000
mean       0.099693
std        0.063318
min        0.000192
25%        0.054791
50%        0.084201
75%        0.138766
max        0.346016
Name: Neighborhood Rent, dtype: float64
```

```
Clothing Store              0.346016
Mediterranean Restaurant    0.274595
Miscellaneous Shop          0.260622
Shopping Mall               0.248173
Yoga Studio                 0.245517
                               ...
Art Gallery                 0.007318
Shopping Plaza              0.006813
Jewelry Store               0.006540
Sports Bar                  0.000562
Performing Arts Venue       0.000192
Name: Neighborhood Rent, Length: 281, dtype: float64
```

Model Building

The most accurate model I built was a logistic regression using the liblinear numeric optimizer. This model had a F1 score of 0.379 and a Jaccard Similarity Score of 0.243

This model struggled when discerning between bins 2 and 3 versus bin 1. One way to improve this is by adding venue prices and reviews as well as how busy they are in order to help the model make more informed predictions. This way the model will be able to discern between high quality and low quality locations in the same venue type.