Predicting Rent Prices in Atlanta Neighborhoods using Nearby Venues

Alon Baruch

December 31st, 2020

Introduction

Business Problem:

Using data on the venues present in different neighborhoods of Atlanta, GA combined with average rent of each neighborhood. We 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.

Background:

The neighborhoods that comprise the city of Atlanta are rich with history. Atlanta was the home of many civil rights leaders such as John Lewis, Ralph David Abernathy, Dr. Martin Luther King Jr, and many more. The neighborhoods they resided in are decorated with tributes to their historic neighbors. However the neighborhoods of Atlanta have also been shaped by its less glamorous history. During the American Civil War Atlanta was a crucial stronghold of the confederacy due to the multiple railroads that intersected within the city making the city crucial for sending supplies to confederate troops throughout the South East. In 1864 during the Union Army's "March to the Sea " led by General William Sherman the city of Atlanta was burnt down (with some help from the retreating confederates) and it's residents displaced. After the Civil War ended in 1865, Atlanta began its rebuilding process which was accelerated due to the city's extensive railroad network. This rebuilding led to a boom in the city's population and the city of Atlanta was made the capital city of Georgia soon after. The city experienced significant growth through the late 1800's and early 1900's however racial tensions and Jim Crow laws would make a lasting mark on the city. In 1906 the Atlanta Race Riots forced black residents out of the booming Downtown Atlanta into the Sweet Auburn neighborhood which remains a predominantly black neighborhood to this day. Further redlining and segregation led to some very noticable differences throughout the neighborhoods such as roads being renamed as they pass through historically "Black" or "White" neighborhoods (i.e. Boulevard becoming Monroe and Moreland becoming Braircliff). After sustained growth through the 20th century including the construction of the world's busiest airport in south Atlanta, several interstates that pass through the city, establishing universities within the city such as Georgia Tech, Morehouse, Emory University. The huge growth of Atlanta throughout the 20th century was capstoned with the 1996 Summer Olympics being hosted in the city. Current day Atlanta is composed of an incredibly diverse, well-educated, and young population which has grown over 24% in the last decade making it one of the fastest growing cities in the U.S.

Data Acquisition and Processing

Data Sources:

The data that will be used to approach this problem will come from RentCafe for the average rent per neighborhood which was acquired using Selenium’s Chromedriver, Geopy for the latitude and longitude data, and finally the Foursquare API to get venue data.

Data Processing:

The data processing required for this project involved collecting all the data needed, removing null values and identifying outliers or incorrect data points, and finally binning neighborhoods into relative rent values. Step one was to gather the rent data per neighborhood using a web scraper. The next step was to get the latitude and longitude data for each neighborhood using geopy and combining it with the neighborhood names and average rent. Since geopy was unable to locate many of the neighborhoods returned from our web scraper we had to shrink the size of our data significantly to use only the neighborhoods which we had latitude and longitude values for. After dropping all the null location values I then calculated each neighborhood's distance from the center of Atlanta to identify any incorrect location values. After running a statistical analysis on the neighborhood’s distance, I noticed that the median distance was only 0.071846 latitude/longitude points however the max distance was 199.116. I removed all outliers by setting the cutoff at 1.5 times the 75th percentile. This resulted in dropping 31 neighborhoods from my dataset since they had incorrect location values.

Once I had all the neighborhoods as well as their average rent and location data, I was able to bin the neighborhoods into 4 groups ranging from most expensive to least expensive.

These groups are critically important because I can now use a classification algorithm to estimate which group a neighborhood would fall into. After the neighborhoods had been binned, I used Foursquare API to collect venue information for each neighborhood. Once the raw venue data was collected I calculate the frequency of each venue by converting the data into a one hot encoded data frame and then group by Neighborhood and calculating average values for each venue in a neighborhood. This would return what percentage of venues in a Neighborhood are of each venue type.

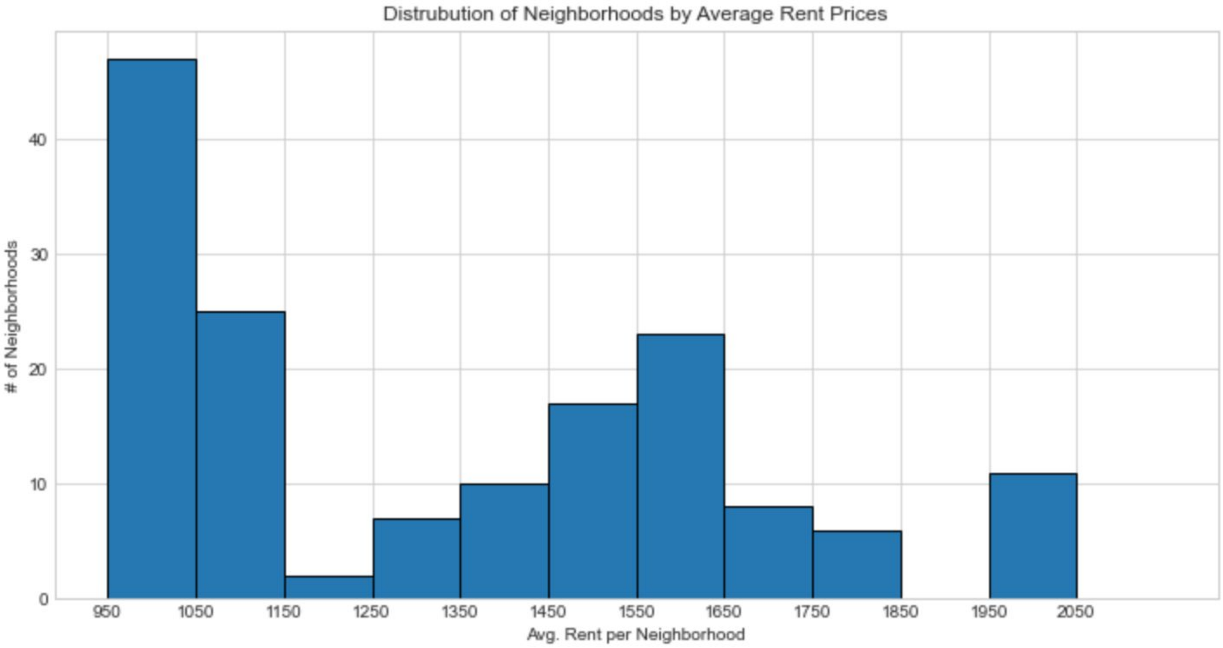
Feature Selection:

In order to determine which variables (venue types) would be most useful for predicting neighborhood rent I ran a correlation analysis on the data frame finding each venue type's correlation with Neighborhood rent. To limit the number of variables in my model I set the cutoff at the 75th percentile and removed any venue types that were not in the top 75% of venue types in terms of magnitude of correlation (the absolute value of correlation). This left me with 70 venue types compared with the 281 venue types I started with.

The reason I went with correlation magnitude instead of venue frequency (which was my original plan) was that some venues are not very frequent but have huge impacts on rent prices. Venues such as sports stadiums, historical landmarks, and large parks would not be considered in a frequency based selection while they are considered in correlation based selection.

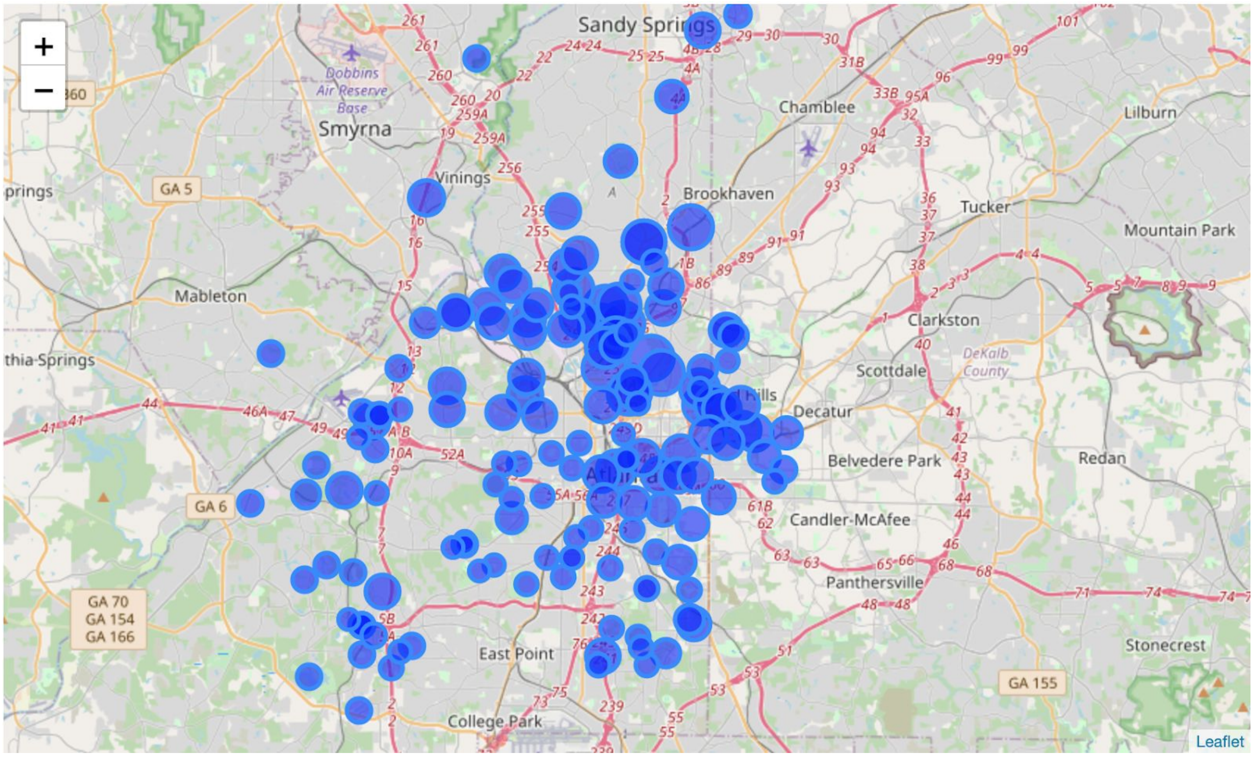
Exploratory Data Analysis

Neighborhood Rent Visualizations:



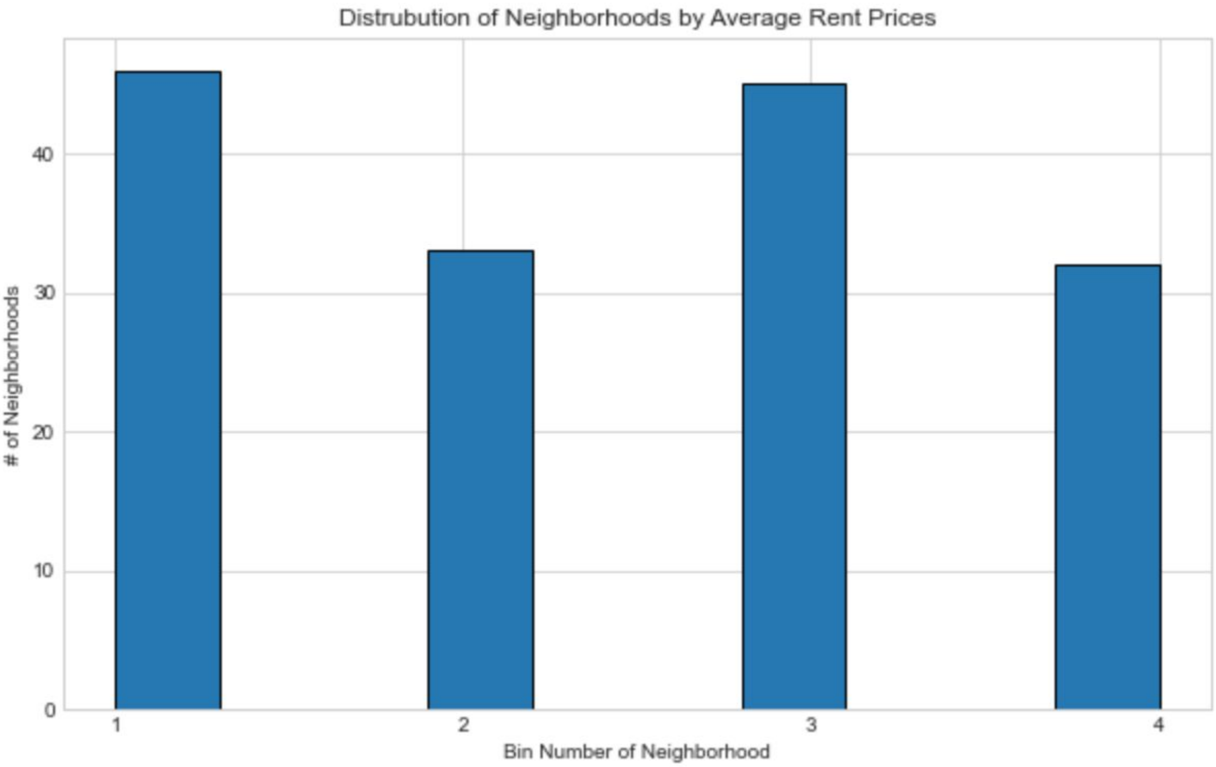
The histogram above shows the rent prices of neighborhoods on the x axis and the number of neighborhoods in those price ranges on the y axis. With the graphic it is easily seen that the mode of our data is in the $950 to $1050 range. We can also visually see that there are no neighborhoods in the $1850 to $1950 range.

Mapping Atlanta Neighborhoods:

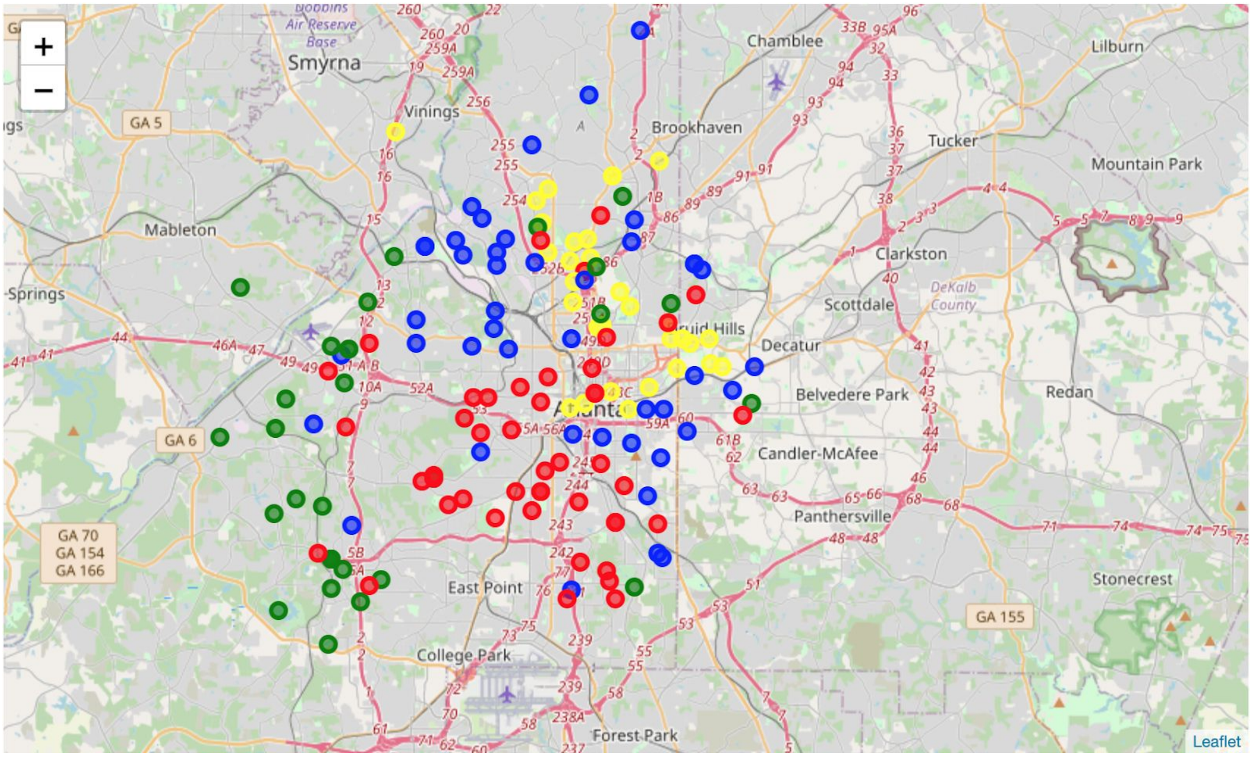


In the map above, each neighborhood is marked by its location and the size of each marker is relative to how expensive a neighborhood is. Larger bubbles represent more expensive neighborhoods while smaller bubbles represent more affordable neighborhoods. This map is useful to visualize where in Atlanta are people of different socio-economic status living. There is a large collection of expensive neighborhoods north of Ponce de Leon Ave. With a few expensive neighborhoods south of Ponce De Leon but still north of I-20 and East of I-85. These physical barriers have represented dividing lines between rich and poor neighborhoods in Atlanta.

Binning Neighborhoods:



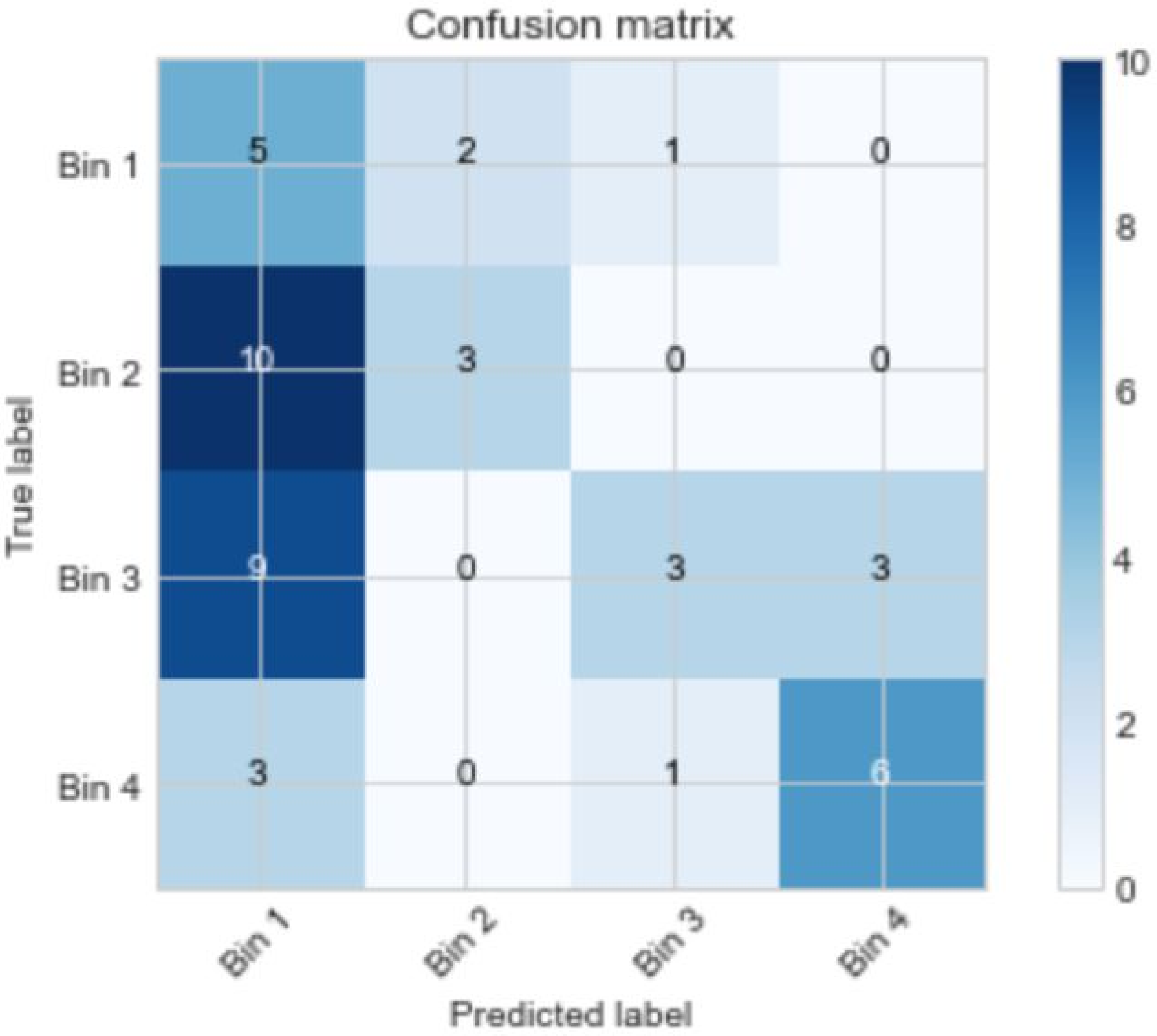
The histogram above shows how many neighborhoods have been placed into each bin. Bin 1 being the Neighborhoods with the lowest rent values and Bin 4 being the Neighborhoods with the highest rent values. Another way to visualize this is through a folium map



In the map above, Neighborhoods with the highest rent values are colored Yellow, the Neighborhoods with the second highest rent values are colored Blue. Third highest in Green and the Neighborhoods with lowest rent values are in Red.

Model Building

While building my classification model I tried Support Vector Machines with many different kernels as well as Logistic Regression with several different optimizers. The model that I found worked best was a Logistic Regression using the liblinear optimizer. Using this model I managed a F1 score of 0.379 and a Jaccard Similarity Score of 0.243. Below is the confusion matrix of using this model



As is evident above, the model had a difficult time distinguishing neighborhoods that belonged in Bins 2 and 3 versus Neighborhoods that belonged in Bin 1. I believe that much of this error can be attributed to the fact that Bin 1 was the largest bin while bins 2 and 3 were the smallest. Furthermore, while I do analyze venue types I do not analyze anything else about this venue so a cheap bar and the most luxurious and expensive bar in Atlanta will both be analyzed as simply a bar. In the future I would like to add venue prices and reviews as well as how busy they are in order to help the model make more informed predictions.