Analyzing Rent Prices in an Area & Neighboring Venues

Author – Amit Kamat

Table of contents

Table of Contents

I. Introduction

One of the most important tasks someone moving to a new city is to identify accommodation. The aim of this project is to aid recent graduates, moving to a new city in selecting a suitable property to rent.

In this project, we explore the neighborhoods of Atlanta in order to extract the correlation between the real estate value and its surrounding venues. We shall employ the Foursquare location data to compare various neighborhoods in the city of Atlanta.

It is common for owners or agents to advertise their properties being in the proximity of some interesting venues like supermarkets, restaurants, businesses etc; displaying convenience of the location in order to raise the property's value. We shall be analyzing this correlation and derive some highlights.

This project could be used as a buyers' guide by the following:

- Students moving to a new city, interested to live in a happening neighborhood.
- Property agents who can optimize the value of their properties.
- Real estate planners, aiding them to decide what kind of venues around their products can help them earn a big profit.
- Other budding Data Scientists who are curious about exploring trends in the real estate market and conduct their own studies.

The inspiration for this project came from my ordeal of moving to a new city for work. I was unable to find a source of information that was concise and lucid in order to make a well-informed decision when selecting a property to rent. Hence, I decided to try to undertake my own study and construct a frame to base future studies upon.

II. Data Description

Atlanta was chosen as the pilot city for this project. The real estate prices by neighborhoods were scrapped from the following website:

https://www.rentcafe.com/average-rent-market-trends/us/ga/atlanta/

The data covers all the neighborhoods in the city of Atlanta. Note that this data describes the average rent for a studio apartment in a given neighborhood. This data is one of the driving factors in the study. Along with this data, we shall be using geodata of Atlanta pulled from a file stored in cloud:

https://www.dropbox.com/s/edecety6jpkbtfv/atlanta-postal.csv?dl=0

Due to the lack of open data available I decided to construct a couple of dictionaries mapping the latitudes and longitudes to their neighborhoods.

Finally, in order to explore venues in and around the neighborhoods in question, we use the Foursquare API.

The process of cleaning the data and processing it includes the following steps:

1. Scrapping the rentcafe website fo the average rent prices for a condo in Atlanta by neighborhood.
2. Pull the geographic data of the neighborhoods from the cloud file and concatenating the two dataframes.
3. Construct two dictionaries for the neighborhoods, storing their coordinates and then appending the dictionaries to the previous dataframe.
4. The coordinates of each neighborhood are then passed to the Foursquare API . We then obtain a list of venues in the pre-determined radius.
5. We then count the number of occurences of each type of venue in a neighborhood. We turn each venue type into a column with their occurrence as the value, using one hot encoding.
6. We then standardize the average price by removing the mean and scaling to unit variance.

The resultant data frame is shown in Fig. (1):

	Neighborhood	Yoga Studio	Accessories Store	Acupuncturist	Adult Boutique	African Restaurant	American Restaurant	Animal Shelter	etarian Vegan taurant	Video Game Store	Video Store	Vietnamese Restaurant	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	StandardizedAvgPrice
0	Ansley Park	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	1.501525
1	Ardmore	1	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	1.501525
2	Atlantic Station	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	2	2.528886
3	Bolton	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	-0.197047
4	Brookwood	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	-1.349975

Figure (1): Final Dataset

The dataset has 42 samples and 264 features. The number of features may vary for different runs due to Foursquare API may recommend different venues at different points in time.

Since the number of features is much bigger than the number of samples, we will run into problems during the analysis processing.

III. Methodology

The project is based on the pretense that surrounding venues influence the price of properties. We shall be using regression techniques where dependent variable will be standardized average prices. The regression model will provide us with a coefficient list. These coefficients give us an idea about the type of venues that influence the prices of a property.

The first order of events following the accumulation of data will be visualization of the data on a choropleth map. This map shows us the varying average rent pricing across the map in terms of color shader. The map is shown in Figure (2) below.

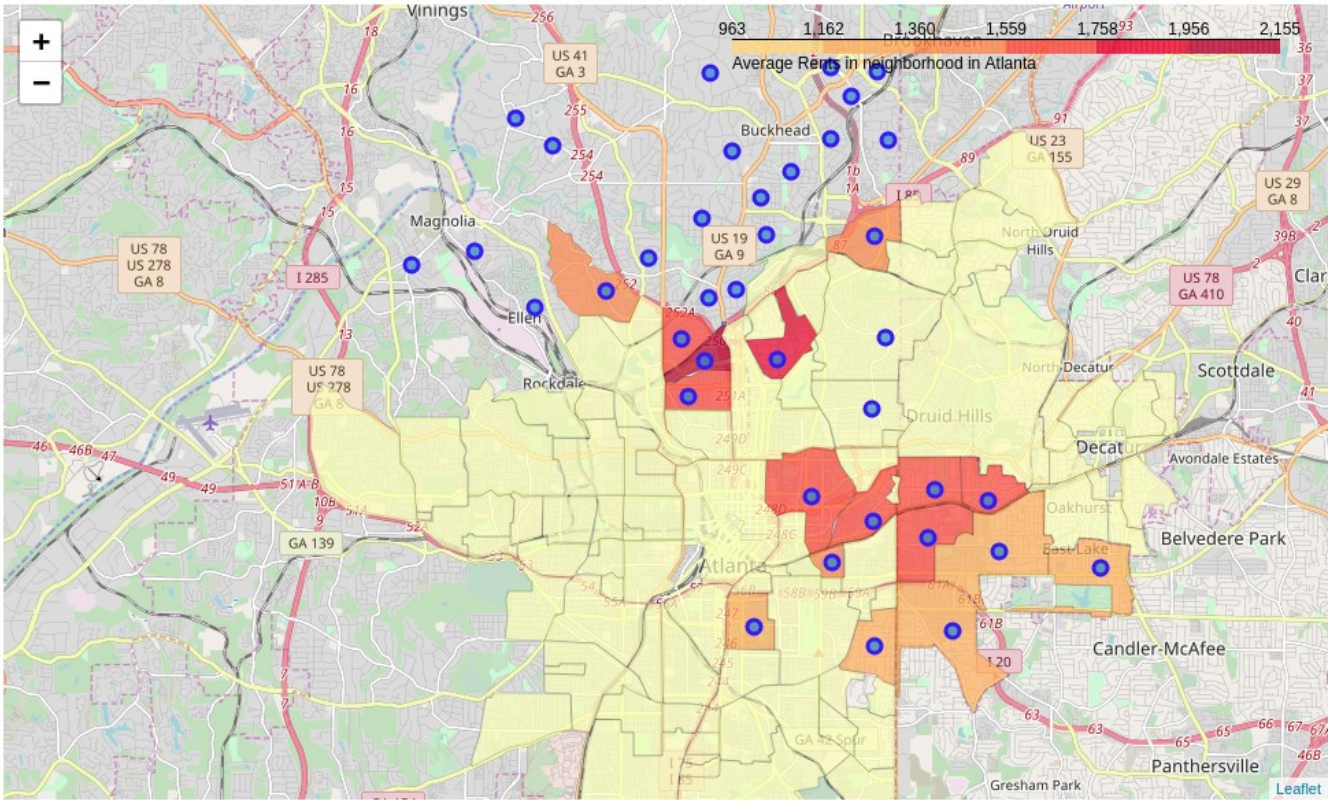

Figure (2): Average Rents in neighborhood in Atlanta

The darker the shade is, the higher the average rent price is in the neighborhood. The average prices seem to be higher in the areas encompassing Georgia Tech university at the upper side and the areas near the railway tracks near the lower right. Intuitively, this data and distribution makes sense.

IV. Results

Upon using Linear regression, we get a very low R-squared and a low MSE value. Thus, our model has not performed very well. This could very well be credited to the fact that the number of features overwhelm the number of samples that we have. Figure (3) shows the MSE and R squared values obtained and also the coefficients corresponding to the venue types.

In order to overcome the imbalance of the data-set, we use a feature selection technique in the form of PCA. However, the results of PCA do not vastly improve upon the Linear Regression model, as we can see from Figure (4).

```
R2-score: -0.44458195300790426
Mean Squared Error: 0.7526282864335427
Max positive coefficients: [0.43120925 0.41331943 0.36817285 0.32298654 0.28814906 0.26241
 0.25760279 0.24507346 0.22226638 0.18529904]
Venue types with most postive effect: ['Other Great Outdoors' 'Leather Goods Store' 'Print Shop' 'Locksmith'
 'Jewelry Store' 'Financial or Legal Service' 'Kitchen Supply Store'
 'Art Museum' 'Residential Building (Apartment / Condo)'
 'Paper / Office Supplies Store']
Max negative coefs: [-0.92307063 -0.61780383 -0.55661865 -0.49736444 -0.42924259 -0.37928339
 -0.33276326 -0.30156432 -0.23355192 -0.22716792]
Venue types with most negative effect: ['Toy / Game Store' 'Tennis Court' 'Cemetery' 'Asian Restaurant'
 'Movie Theater' 'Grocery Store' 'Martial Arts Dojo'
 'Brazilian Restaurant' 'Liquor Store' 'Whisky Bar']
Min coefs: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Venue types with least effect: ['Racetrack' 'Convention Center' 'Gymnastics Gym'
 'Construction & Landscaping' 'Hardware Store' 'Soup Place' 'Hotel Pool'
 'Bridge' 'Public Art' 'Recreation Center']
```

Figure(3): Linear Regression Evaluation

```
Max positive coefs: [0.01825622 0.01650884 0.01649235 0.01635703 0.01609322 0.01581053
 0.01581053 0.01581053 0.01581053 0.01581053]
Venue types with most positive effect: ['Performing Arts Venue' 'Indian Restaurant' 'Hotel' "Men's Store" 'Lake'
 'Bistro' 'Exhibit' 'Concert Hall' 'Indie Theater' 'Field']
Max negative coefs: [-0.0041051  -0.00373014 -0.00368489 -0.00344844 -0.00344279 -0.00340825
 -0.00332525 -0.00332525 -0.00324038 -0.00318918]
Venue types with most negative effect: ['Discount Store' 'Gas Station' 'Outdoors & Recreation' 'Soccer Field'
 'Wings Joint' 'Baseball Field' 'Construction & Landscaping' 'Tree' 'Pier'
 'Park']
Min coefs: [-7.37972446e-05  1.83326843e-04  1.83326843e-04  1.97812115e-04
 -2.04667062e-04 -2.16613424e-04 -2.35739526e-04  3.42677905e-04
 -4.11949782e-04 -4.20068244e-04]
Venue types with least effect: ['Bank' 'Piercing Parlor' 'Costume Shop' 'Other Repair Shop' 'Supermarket'
 'Tea Room' 'Martial Arts Dojo' 'Mobile Phone Shop' 'Liquor Store'
 'Wine Shop']
```

Figure(4): PCA Evaluation

The PCA results tell us a more sensible story. Venues such as theaters, restaurants and stores will be in and around more expensive neighborhoods. Areas neighboring repair shops, fields and gas stations will have a lesser cost.

V. Conclusion

I faced several challenges and the most difficult of these was the unavailability of public data-set. None of the data-sets or even website had enough data. Moreover, while combining data from various sources, a lot of inconsistency arises, which can lead to dropping of samples, that otherwise could have enriched our results. After the construction of a dataframe, we need to decide if further analysis needs to be done and to perform transformations and standardization. Choosing a suitable technique is quite important as shown by this report.

Unfortunately, we could not design a suitable model that show a positive relationship between neighborhoods and average property prices in the area. However, we do gain some meaning full insights from the data and the models.