**Predicting the rent price by the numbers of different venues available**

**near each New York neighborhood**

Anjin Xianyu

August 11th, 2019

**1. Introduction**

* 1. **Background**

Living in New York is never a financially easy task. Renting a home can cost a significant portion of one’s salary. According one report from StreetEasy, in 2016, some renters in New York paid 65.2 percent of their salary for renting an apartment. However, there are a lot of neighborhoods available in New York. The rents in some neighborhoods are definitely more expensive than the others. It would be wise to think carefully for which neighborhoods one should choose to live in. Choosing living in the Manhattan might bring your home to a variety of stores and restaurants you want to go for your after-work life closer, but it also cost you more. Would it be a better choice to live in the nearby counties such as Brooklyn and Queens to save some rent? Whether the convenience and infrastructure in the Manhattan neighborhoods actually paid off the costly rent for living in New York can be a very interesting question.

* 1. **Problem**

In this project, I want to predict the amount of rent based on the living quality in the neighborhoods. The living quality in a neighborhood here simply means the number of a variety of venues available. Here, I assume in all neighborhoods, there are a similar proportion of different types of housing, such as studio, one bedroom apartment and so on. Thus, the median rent of one neighborhood can reflect how pricy to live in this neighborhoods. Multiple linear regression will be used to predict the median rent price by using the number of different venues near the neighborhoods.

* 1. **Interest**

Renters in New York would be very interested in knowing whether their rent actually worth for the quality of life their neighborhoods provide to them. They may also want to move to neighborhoods with better living quality but with less rent. Furthermore, it would be also very beneficial for the rental buildings and landlords to decide whether they asked for a fair price based on the living quality of neighborhoods.

**2. Data acquisition and cleaning**

**2.1 Data sources**

The median rent data of New York is extracted from this website <https://streeteasy.com/blog/data-dashboard/>. Furthermore, the numbers of venues near each neighborhood are extracted from Foursqare API (<https://foursquare.com/>)

**2.2 Data cleaning**

The median rent price for each New York neighborhoods are available for from 2010 to 2019. However, considering the venue data from Foursquare API reflects the current situation. It might not be suitable to use outdated rent prices in this analysis. Therefore, only the median rent prices for the last 12 months were used. Averages of the last 12 months were calculated for all neighborhoods as the median rent price for each neighborhood.

After that, to determine the GPS coordinates of all New York neighborhoods, I pull out such data from <https://geo.nyu.edu/catalog/nyu_2451_34572>. In this dataset, there are actually a few missing neighborhoods. To find out the GPS coordinates of these missing neighborhoods, I used the geopy packages to extract these coordinates. Moreover, there are also some inconsistencies of the names of neighborhoods. I manually examine these names and reassign them to the closest match. Then, I plotted a map for these neighborhoods and their assigned GPS coordinates. No obvious problem has been observed. After this step, I combine the median rents of all New York neighborhoods with their GPS coordinates. The next step is to pull out the number of different venues available near each neighborhood. To accomplish this goal, I used Foursqare API to extract all venues 500 meters around this neighborhood GPS coordinates. The reason I used 500 meters is this appears to be a reasonable walking distance to me. The number of all venues then can be attached to each neighborhood. This makes the dataset I used for the exploratory analysis and the regression model. For example, the neighborhood A has 2 coffee shop, 3 restaurants and so on, with a median rent $2000.

**3. Exploratory analysis**

**3.1 Case study of the neighborhood with the highest rent and the lowest rent.**

In the dataset, The Central Park South neighborhood has the highest rent with $7481, while the Van nest neighborhood has the lowest rent with $1487. A close look at the 5 most common venue types for these two neighborhoods are quite different (Table 1). In the central Park South, there are a lot of hotels, boutique, jewelry stores, American restaurants and Italian restaurants, while in Van Nest, there are a lot of Deli and Pizza places. At the first glance, the life styles indeed differ in these two neighborhoods.

|  |  |  |
| --- | --- | --- |
| Table 1. Comparison of number of venues in the most expensive and cheapest neighborhoods | | |
|  | Central Park South(median rent: $7481) | Van Nest ($1481) |
| 1 | Hotel(13) | Deli/Bodega(7) |
| 2 | Boutique(7) | Pizza Place(5) |
| 3 | Jewelry Store(6) | Bus Station (2) |
| 4 | American Restaurant(5) | Liquor Store(1) |
| 5 | Italian Restaurant(5) | Chinese Restaurant(1) |

**3.2 Relationship between the median rent and the total number of available venues**

From the previous comparison, it seems there might be a difference between the total numbers of venues in the first place. Thus, I added all the venue numbers and plotted it against the rent median for each neighborhoods.

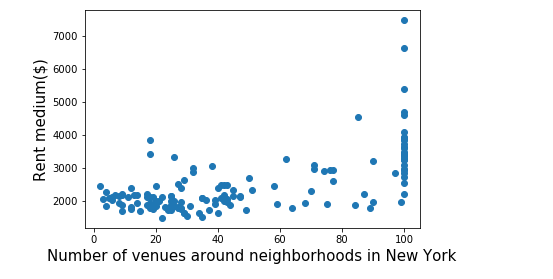
It turns out Foursquare set a hard limit for 100 venues per location, despite that I requested a limit of 500. Regardless, the correlation coefficient is 0.63, which suggests there is a weak correlation between the total venue number and the median rent.

Figure 1: the relationship between number of venues around neighborhoods in New York and the total number of venues in each neighborhoods.

**3.2 Relationship between the median rent and the public transports**

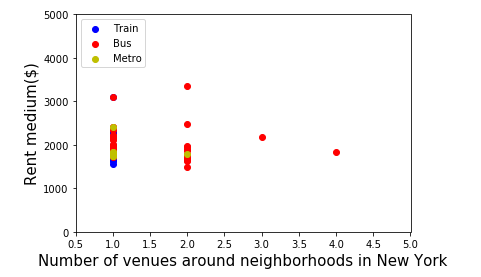
After I examined all venue names in the dataset, I noticed there are a few transport related venues, such as Bus line, Bus station, Bus stop, Train, Train station and Metro station. It is understandable that people may want to live near the public transports for easy commuting. Therefore, I decided to combine Bus line, Bus station and Bus stop as Bus related, Train and Train station as Train related, and Metro station as Metro Station. Surprisingly, the results showed that there is no obvious relationship between the public transports with the rent median. This might be due to there is a highly developed public transport systems in New York, leading no obvious difference between the rich and poor neighborhoods in terms of convenience to the closest public transport.

Figure 2: the relationship between number of venues around neighborhoods in New York and the median rent for each neighborhood

**3.3 Correlation between each variable against the median rent**

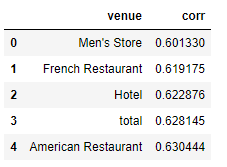
Before applying any model to fit the data, I checked the correlation between each variable against the median rent, to see which variable has the highest positive or negative correlation. The top variable is the number of American restaurants in the neighborhood, with the correlation coefficient as high as 0.63. The following correlated venues are total hotel, French restaurant, Men’s store, Boutique and coffee shop, with the correlation coefficient varying from 0.63 to 0.57.

Table 2: Top 5 venues with the highest correlation with the median rent



Table 3: Top 5 venues with the smallest correlation with the median rent

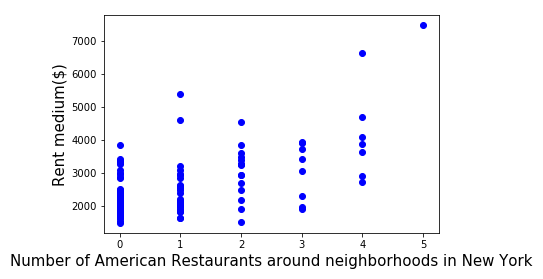
Then, I plotted the median rent of each neighborhood against the number of American restaurant (i.e. the most correlated) or Egyptian restaurants (i.e. the least correlated). The results show that there is indeed a weak correlation between the number of American restaurants and median rent, while due to most of the neighborhoods have 0 Egyptian restaurants, there is not much correlation at all. By checking a few other venues with low correlation coefficient (figures now shown here), there is only very few of them in New York, which means their relationship with the rent median of the neighborhoods is unknown due to most of the neighborhoods simply do not have any of them.

Figure 3: the relationship between number of American restaurants around neighborhoods in New York and the median rent for each neighborhood

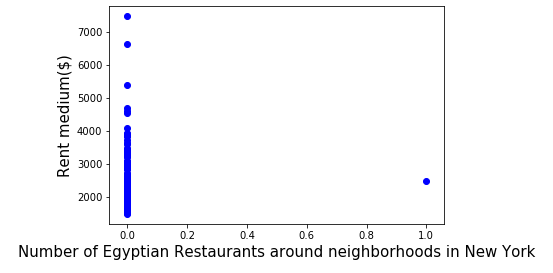


Figure 4: the relationship between number of American restaurants around neighborhoods in New York and the median rent for each neighborhood

**Regression model**

**Multiple linear regression model**

Considering the independent variables are continuous and the dependent variable is continuous as well, I applied the multiple linear regression model to the data.

By applying the data to the multiple linear regression model, I get an average of $444 deviation between the predicted value and the true value of the median rent. From the test set, the dots of the predicted median rent vs the true median rent are distributed around y=x , suggesting this model at least is predicting the median rent based on the number of venues in the neighborhoods. Despite the variance score, 0.39 is not very high, the $444 average deviation looks reasonable to me.

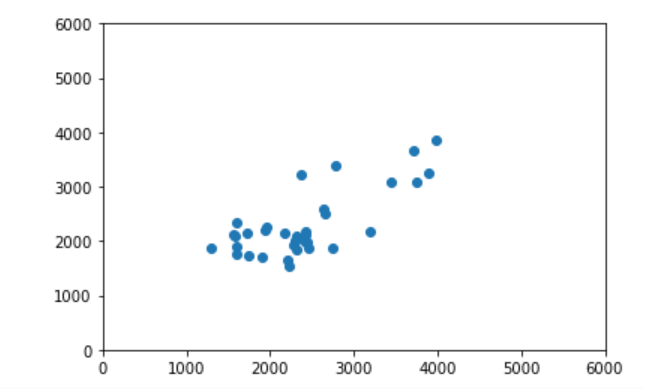


Figure 5: the relationship between the predicted median rent vs the true median rent($)

By examining the coefficients for each venue, the college cafeteria, Campground, and food court become the first three contributors to the median price. However, some of them do not really have high frequency of appearance. I am wondering what will happen if we only use more commonly appearing venues in New York to fit this multiple linear regression model.

By feeding the model with venue appearance at least 1 to 10 times, the performances of the models actually go down from $444 average deviation to $2012 average deviation, along with variance score dropping dramatically. Thus it may actually require more venues to get a better model fitting result instead of removing low frequency venues.

If only the high appearance venues (i.e. more than 10) are taken into consideration, the food court, the playground and the hotel contributes the top three venues providing the most rent increase per occurrence. In the exploratory analysis, American restaurant has the highest correlation with the median rent. In this model, every American restaurant contributes to $98, while every Egyptian restaurant with the lowest correlation the median rent only contribute $2 to the median rent. Despite high correlation does not necessarily mean higher coefficient in the multiple linear regression, the fact that the venues with higher correlation has higher coefficient imply the model is functional following common sense.

**Decision tree**

While the independent and dependent data in this dataset are both continuous, decision tree model can still be used if I turn these continuous variables into discrete variables. The numbers of venues are turned into 0, 1 and 2, standing for none, one venue and many venues, while the median rent is turned into low, median and high. The result shows that it has about 0.61 as its accuracy, which is much higher than random picking 0.33. However, such accuracy is not more useful than the multiple linear regression model.

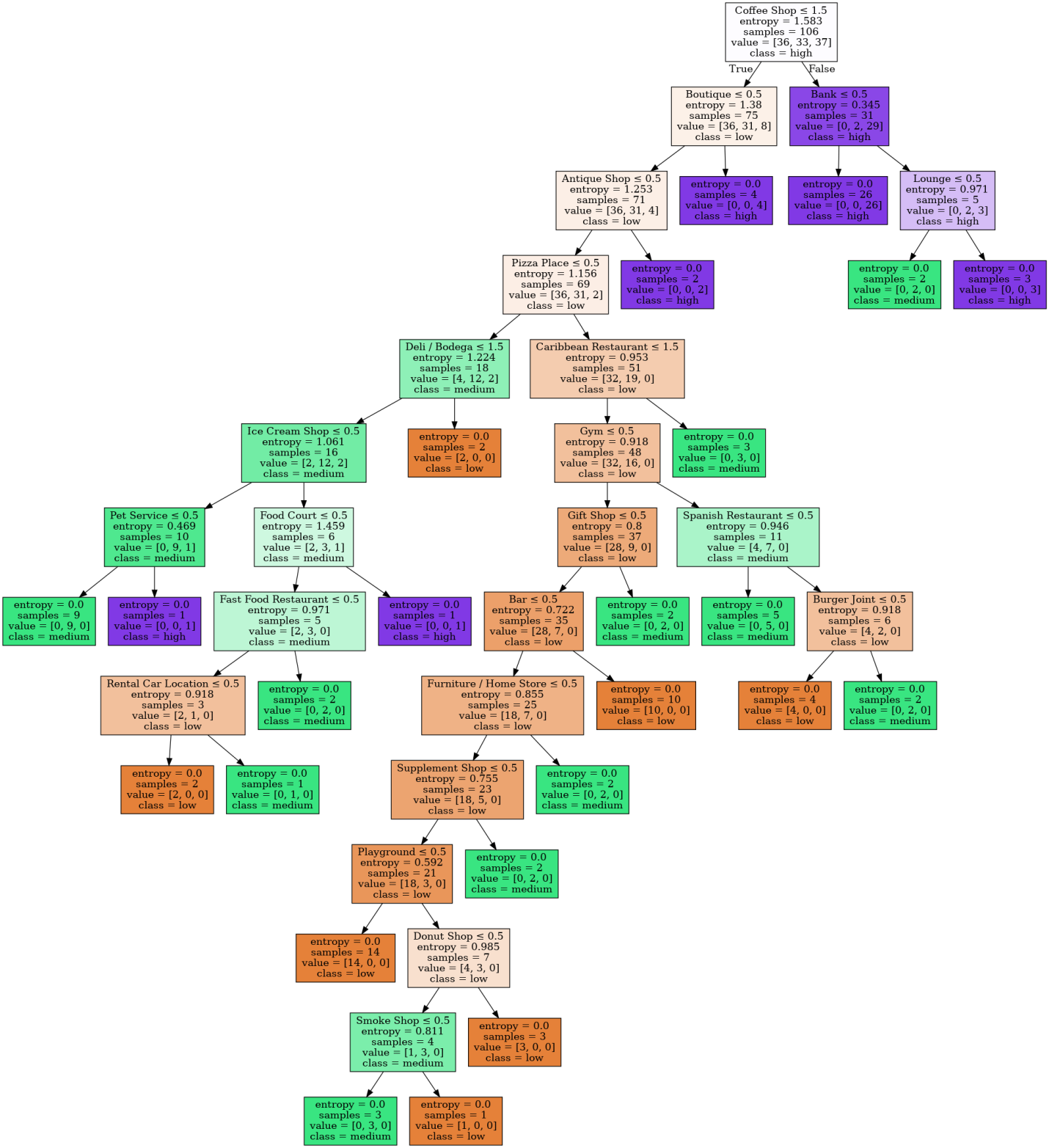


Figure 6: visualization of decision tree

**Discussion**

The attempt to predict the median rent of a neighborhood in New York based on the number of venues in this neighborhood is partially successful. By using the multiple linear regression model, an average of $444 rent price deviation is achieved. Since removing the venues types dramatically reduces the accuracy of this model, to further improve the model performance, more venue types should be included. Furthermore, considering there is only ~150 neighborhoods in the New York city, an expansion of data by including more cities might be useful to improve model performance by increasing training set. However, this could also introduce more errors considering dwellers in different cities might have different taste.

In addition, Foursquare does not allow more than 100 venues per location to be collected. Such limitation might also cause a drop in model performance. To overcome this issue, another location information source, if available, can be used.

Alternative methods including decision tree might not be very useful for this question. By arbitrarily compressing the median rent and number of venues into discrete numbers, information is lost during this process. An attempt by using decision tree to predict the median rent only achieves accuracy 0.61. Despite that this accuracy is much higher than random picking 0.33, it is not useful for potential users to extract information from such model. Further attempt by categorize the median rent into more steps (very low, low, medium, high, very high, and etc) even further reduce the accuracy.

In my multiple linear regression model, the venues appeared in neighborhood with high median rent also have higher coefficient, which providing extra confidence that this model is functional.

**Conclusion**

The multiple linear regression model I used here can predict the median rent of New York neighborhoods based on the number of venues in the neighborhoods with a reasonable accuracy. Further attempts to include extra information about neighborhoods might increase the model performance. This model and future improved versions of this model can provide insight to the New Yorker and the real estate owner to access and estimate the proper rent.