Capstone Project- NYC Neighborhood Rents

IBM Data Science Professional Certificate by Coursera

Gage Abell

INTRODUCTION

In New York City, accessibility and proximity to public services are highly valued. A local park can be a very attractive addition to a densely populated neighborhood. Similarly, a nearby subway station can significantly shorten one's daily commute. These luxuries are often desired by renters; however, they come at a cost. Not all neighborhoods have beautiful parks or a subway station right around the corner.

When renters seek a residential rental property in the city, it's important that they understand how much an amenity or service like these can affect rents. Likewise, landlords should be conscious of the value placed on amenities and services surrounding their property to ensure their asking price reflects the market value.

In this capstone project, I aim to answer the question, how does proximity to subway stations and public parks affect the median rent (asking price) in New York City neighborhoods?

DATA

In order to conduct my analysis, I will need to source location data for New York City and separate it into geographical boundaries defined by the city’s boroughs. Additionally, I will need information about the proximity of subway stations and public parks to residents in each neighborhood. Fortunately, much of this data is available from the NYU Furman Center which provides historical data on the city's neighborhood populations.

I will be using information from 2017 as it is the most recent year that has the most complete data within the database.

-The proximity to a subway station will be defined by the percentage of residential units that are within a half mile walk of a station entrance for the New York City Subway, Long Island Rail Road, PATH, Amtrak, Metro-North Railroad, or Staten Island Railway.

-The proximity to a public park will be defined by the percentage of residential units that are within a quarter mile of a park.

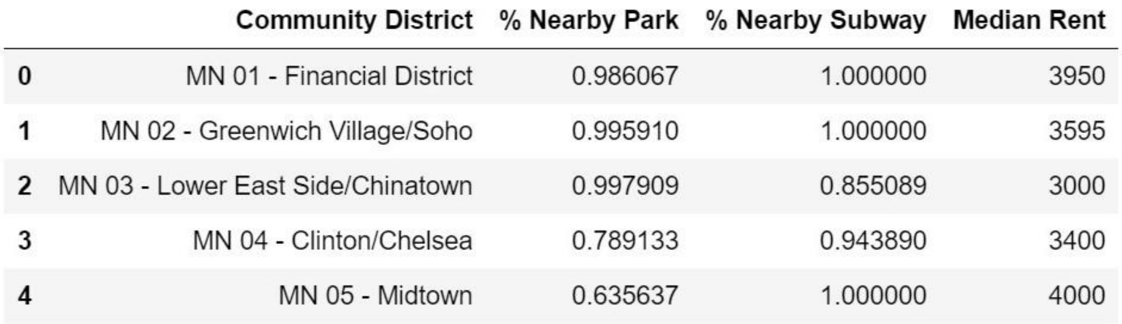
-Median rent will be defined as the median rent that landlords ask for housing units available for rent (in USD).

METHOOLOGY

Before I can begin making any insights or correlations, I must ensure that all my data is properly imported and cleaned to contain only relevant information. The relevant pieces of information for my analysis include-

- ‘Community District’- Names of NYC boroughs
- ‘% Nearby Park’
- ‘% Nearby Subway’
- ‘Median Rent’

Once each unique column is imported and cleaned, they will be combined into a single data frame, ‘df_combined’.



	Community District	% Nearby Park	% Nearby Subway	Median Rent
0	MN 01 - Financial District	0.986067	1.000000	3950
1	MN 02 - Greenwich Village/Soho	0.995910	1.000000	3595
2	MN 03 - Lower East Side/Chinatown	0.997909	0.855089	3000
3	MN 04 - Clinton/Chelsea	0.789133	0.943890	3400
4	MN 05 - Midtown	0.635637	1.000000	4000

Next, I will make initial observations by using data visualization techniques included in the matplotlib library. By creating a simple scatter plot, I am able to plot ‘% Nearby Park’ (*Figure 1*) and ‘% Nearby Park’ (*Figure 2*) against ‘Median Rent’ to make general observations.

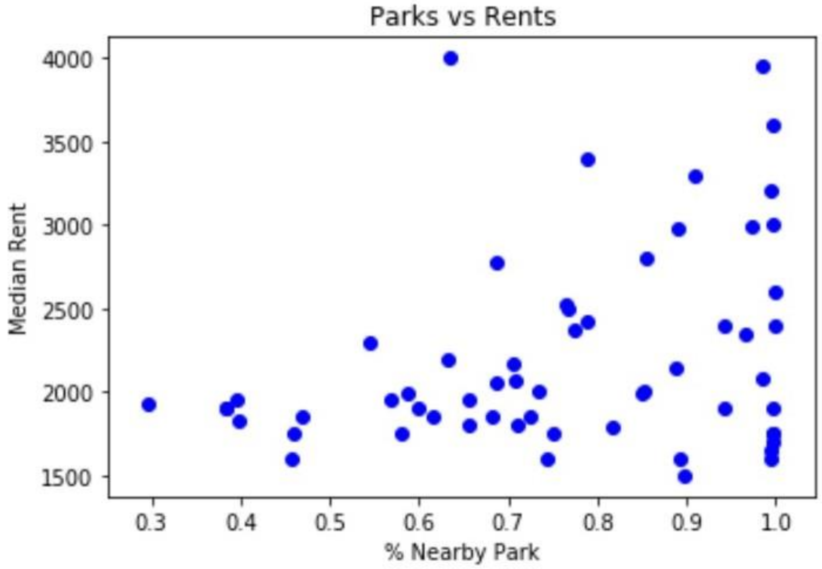


Figure 1

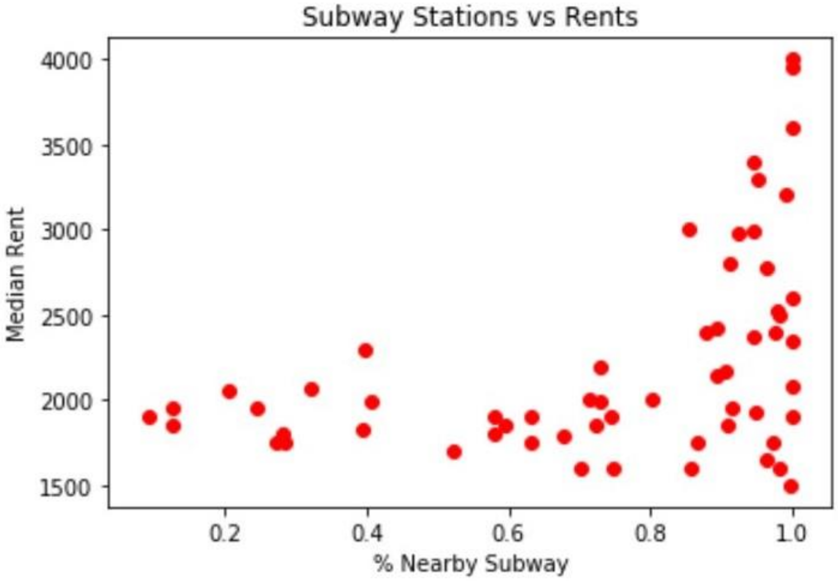


Figure 2

In both scatter plots, there seems to be a positive correlation. However, there is not a clear pattern that shows that a higher x-value definitely correlates to a higher median rent. Specifically, in Figure 2, we see that both the minimum and maximum y-values occur when 100% of residential units are nearby a subway station. Similarly, Figure 2 shows a similar pattern as some of the lowest and highest x-values occur when 100% of residential units are nearby a park.

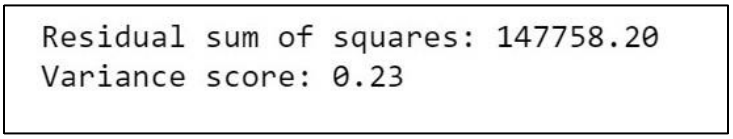
Multiple Linear Regression

I chose to use multiple linear regression, a machine learning method, because I needed to compare multiple independent variables ('% Nearby Subway', '% Nearby Park') to a single dependent variable ('Median Rent').

This method relies on splitting the dataset into a 'training set' and a 'testing set'. The machine takes a defined portion of the dataset, in my case 80%, to train itself and to learn to recognize patterns. The remaining 20% is called the 'testing set'. This will test the model's ability to make accurate predictions based on known values of the original dataset.

Once the dataset has been split and passed through the algorithm, the linear regression model will assign coefficients that correspond the independent variables. The resulting linear equation will be the basis for making future predictions.

To test the accuracy of the model, I calculated the residual sum of squares and the variance score which evaluates how well the model can predict the dependent variable. The evaluation method works because I can test it by comparing the results of its prediction against the actual values included in the test set.



```
Residual sum of squares: 147758.20
Variance score: 0.23
```

RESULTS AND DISCUSSION

In my analysis I aimed to understand correlations and impacts of nearby neighborhood services (subway stations and parks) on the median rents in New York City neighborhoods. After creating a multiple linear regression model based on two independent variables, I found that the model had a very low variance score; thus, my prediction was far from perfect. At a score of just 0.23 (1 being a perfect prediction), this is not a reliable way to make accurate predictions of the median rental prices. Additionally, using the ordinary least squares method, the model attempted to minimize the error between the actual output and the predicted output. However, the residual sum of squares remained very high (147,758.30).

If I could improve the model, I would start by choosing different variables that may have a more direct correlation or bigger impact on rental prices. Subway stations and public parks can only give so much insight and are likely not among the driving factors that can make accurate predictions. Clearly, factors such as income would have a significant impact on the rental prices in a neighborhood. Although, in this capstone, I attempted to use uncommon variables to generate similar patterns and insights.

CONCLUSION

In conclusion, this project has made me better understand the importance of taking time to formulate a clear goal and plan from the very beginning. Even though many of the machine learning methods taught in this course can be used to conduct very insightful analysis or create amazingly accurate models, I think one of the most important lessons I learned is that these tools are only useful when properly applied to answer the question at hand.

In this project, I could have included additional independent variables to try and give the model more training. However, if I included more data that didn't correlate any better than my other variables, I would run the risk of overfitting my model and the additional data wouldn't be helpful. This is why I believe spending the time to understand exactly what data to use for a model and which methods will do the best at solving the specific question is so important.