Rental price vs. Neighborhoods - Coursera Data Science Capstone Project

Introduction

Renting a place in a city is always the first issue faced by employees, students and tourists new to the city. Finding a right place matching living requirement and financial budget is the common goal for this issue. Here, the Neighborhoods of New York City will be explored in this scope by employing data science techniques and methods.

In this capstone project, the composition of nearby venues in a neighborhood is used as the features and the relationship between these neighborhood features and rental price will be investigated. How different types of venues influence the price will be studied. To accomplish this analysis, venue information will be collected and processed as independent variables, while the rental price will be gathered and considered as a dependent variable. A fitted correlation will be generated as a guide to estimate the rental price.

Data

Two sets of data are necessary to conduct this project, the rental price and the neighborhood venues. Before gathering these two sets of data, the neighbourhood list of New York City is firstly collected from Wikipedia websites by using Python scraping tool BeautifulSoup in combination with the geographical information of New York City given in this course. Also, the rental price will be obtained from the Rentcafe website by leveraging BeautifulSoup. A small portion of the DataFrame is shown as below.

	Borough	Neighborhood	Average rental price / $	Latitude	Longitude
0	Manhattan	Marble Hill	1694	40.876551	-73.910660
1	Manhattan	Inwood	2225	40.867684	-73.921210
2	Manhattan	Washington Heights	2243	40.851903	-73.936900
3	Manhattan	Randalls and Wards Islands	2336	0.000000	0.000000
4	Manhattan	East Harlem	3334	40.792249	-73.944182

The neighborhood venues can be obtained by implementing the FourSquare API, which has been demonstrated in the previous practice. Part of the queried result is shown in the below DataFrame.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Manhattan	Marble Hill	40.876551	-73.91066	Metro North - Marble Hill Train Station	40.874698	-73.912052	Train Station
1	Manhattan	Marble Hill	40.876551	-73.91066	Marble Hill Youthmarket	40.874519	-73.910394	Farmers Market
2	Manhattan	Marble Hill	40.876551	-73.91066	marble hill pharmacy	40.875050	-73.909195	Pharmacy
3	Manhattan	Marble Hill	40.876551	-73.91066	MTA Subway - 225th St/Marble Hill (1)	40.874486	-73.909589	Metro Station
4	Manhattan	Marble Hill	40.876551	-73.91066	Marble Hill Playground	40.877765	-73.907994	Playground

These data form the foundation of this project, and the correlation between the rental prices and the neighborhood feature will be explored.

Methodology

First, the rental prices are examined alone for brief understanding. Figure 1 displays the corresponding average rental price of neighborhoods in New York City. It can be seen clearly that the rental prices vary from neighbourhood to neighbourhood. It will be worthy to investigate the factors driving this variation.

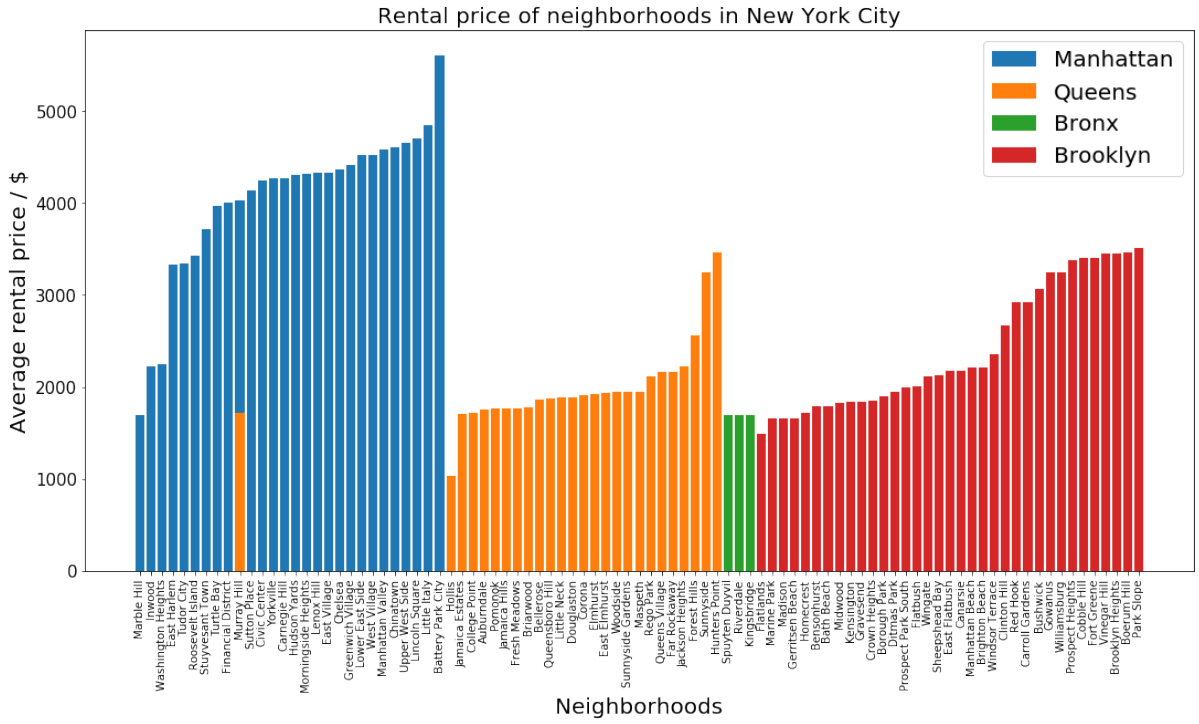

Figure 1 Average rental price of each neighborhood in New York City.

The primary statistical information about the rental prices of given boroughs is represented by the boxplot in Figure 2. In general, Manhattan borough is the most expensive to live in terms of the renting, while the Bronx and Queens boroughs are cheaper.

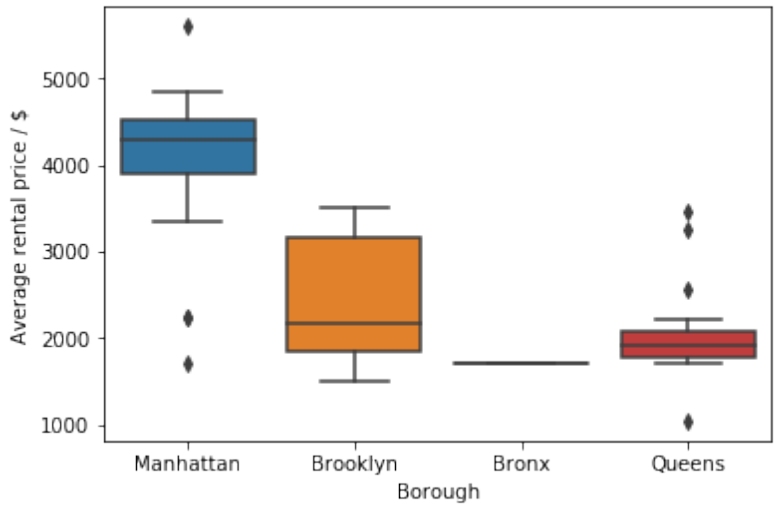

Figure 2 Boxplot of average rental price of each borough

Second, when looking at the venues of these neighbourhoods, one observation is that the numbers of venues in particular neighbourhood found by FourSquare fluctuates considerably, which may be one of the factors driving the rental price variation. On the other hand, the venue records returned by the FourSqaure present one major disadvantage, i.e., the original category description of the venues that have 251 categories in total are too specific to be used for searching patterns.

With the two sets of data ready, further analysis can be performed to discover correlation between them. Since the numbers of venues spread over a wide range, a scatter plot was generated to show how it relates to rental prices graphically in Figure 3. Recalling the ranking of rental expensiveness represented by the boxplot of average rental prices in the four boroughs, this scatter plot roughly shows the sense that people living in expensive boroughs actively use FourSquare. In the following paragraphs, the statistical correlation between the rental price and some features of the neighbourhoods will be checked. Different fitting methods, including simple linear regression, multiple linear regression and polynomial fitting in the Sci-kit learn library, will be employed, and the performance of these methods is evaluated accordingly. In addition, these neighbourhood will be clustered by machine

learning also in the Sci-kit learn library to examine potential correlations between rental price and clustering results.

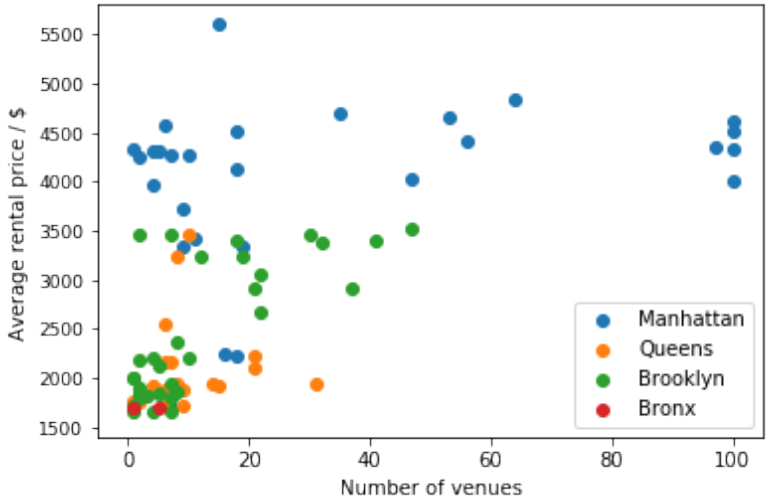

Figure 3 Average rental price vs. Venue numbers.

Result

As mentioned before, the venue categories are too detailed, so the venues are recategorized into more inclusive categories. According to the name of the original description, three big categories can be identified, Restaurant for those names having restaurant keyword, Entertainment for those names having pub, club and bar keyword, and Miscellaneous Convenience for the rest.

	Borough	Neighborhood	Average rental price / $	Latitude	Longitude	Restaurant	Entertainment	Miscellaneous Convenience
0	Manhattan	Marble Hill	1694	40.876551	-73.910660	0	0	7
1	Manhattan	Inwood	2225	40.867684	-73.921210	4	1	13
2	Manhattan	Washington Heights	2243	40.851903	-73.936900	3	0	13
3	Manhattan	East Harlem	3334	40.792249	-73.944182	6	2	11
4	Manhattan	Tudor City	3338	40.746917	-73.971219	0	0	9

This categorisation provides three more features of each neighbourhood. Including the total number of venues, four features of each neighbourhood are paired with the corresponding average rental price individually to conduct linear regression. The fitted results are exhibited in Figure 4 and the statistical coefficients to evaluate the fitting is listed in Table 1.

Table 1 The statistical coefficients

	R-sqaured	Pearson coefficient	p-value
Price vs. Counts	0.291	0.540	1.418e-7
Price vs. Restaurant	0.275	0.524	8.248e-8
Price vs. Entertainment	0.154	0.393	0.0001
Price vs. Miscellaneous	0.290	0.538	3.170e-9

From the statistical coefficients, the positive correlation between rental prices and these four features are demonstrated, especially for the Counts, Restaurant and Miscellaneous Convenience of which the p-values are extremely small. However, the R-squared values are far from one, indicating that linear regression cannot find meaningful function to describe their relationships.

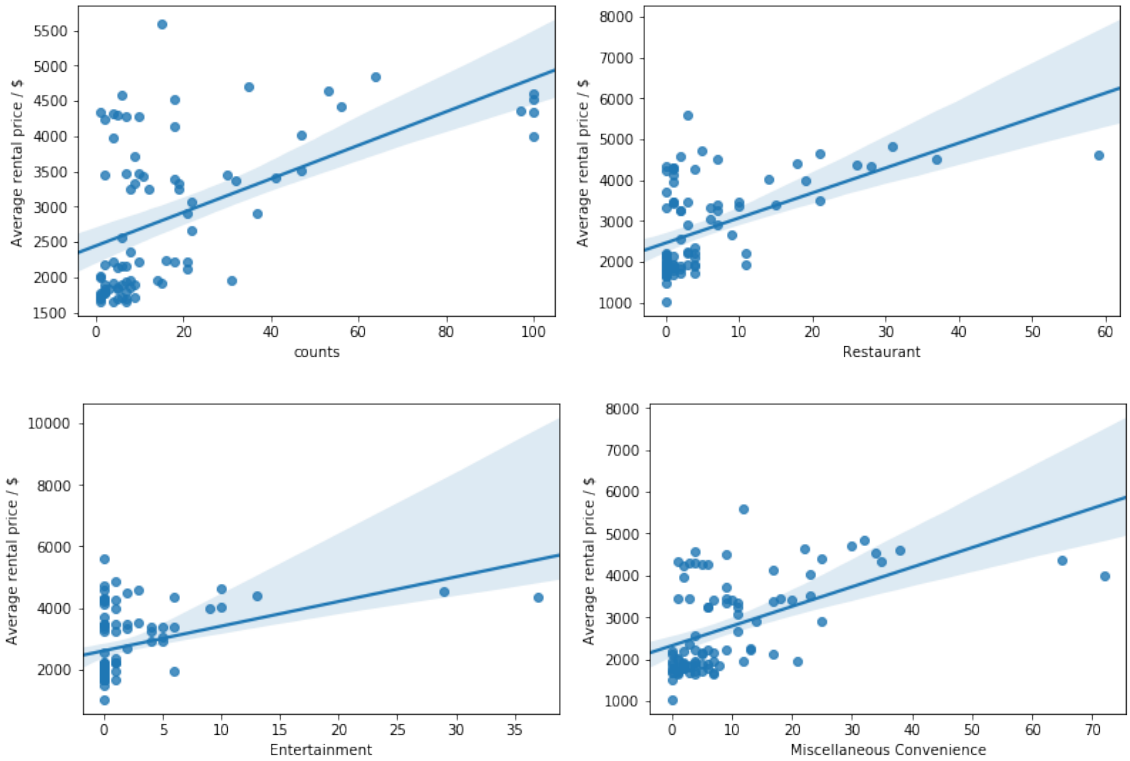

Figure 4 Linear regression fits on four pairs of data.

As the fitting outcomes of the simple linear regression deviate from the actual result substantially, the multiple variable linear regression is further implemented to search for potential solution. The Restaurant (x_R), Entertainment (x_E) and Miscellaneous Convenience (x_M) are the three independent variables to predict the rental price (y_P). The regression gives $y_P = 29.578 * x_R + 12.271 * x_E + 27.365 * x_M + 2333.948$. According to the magnitude of the

coefficients of the three variables, Restaurant and Miscellaneous Convenience have stronger impact on the average rental prices than the Entertainment.

Due to the difficulty to visualise the actual fitting result, the direct comparison between the predicted values and the actual values is used instead to evaluate the precision of this fitting. In Figure 5, the distributions of the predicted and actual rental prices are presented. Large mismatch between these two distributions can be clearly found, pointing out that the multiple linear regression is also not the accurate way to describe the correlations. Further, the R-squared value is determined as 0.324, which is relatively low and conforms the above observation.

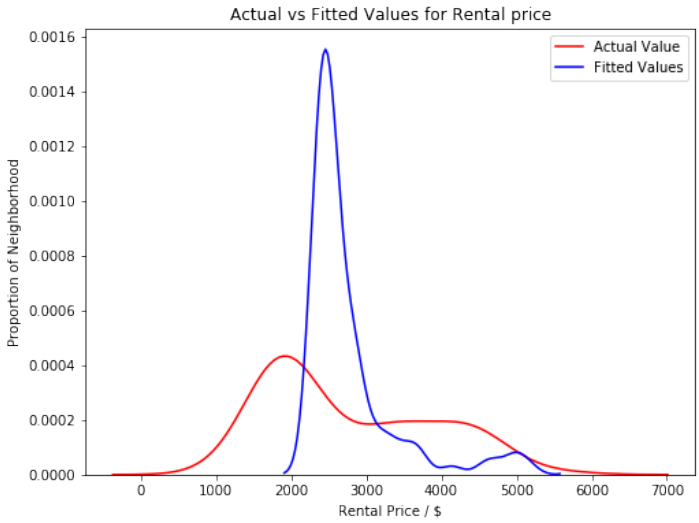

Figure 5 Model evaluation on multiple variable linear regression fitting.

According the previous analysis, the linear regression can be hardly found between the rental prices and the three variables, so the polynomial fitting is applied. To avoid overfitting, only the degree 2 polynomial fitting is used. A similar comparison between the predicted values by polynomial fitting and the actual values is exhibited in Figure 6. The mismatch is still obviously identified, but the deviation between the two sets of values is less remarkable. This demonstrates the polynomial fitting has a stronger tendency to pass through every data point. The R-squared value (0.395) of this polynomial fitting is slightly higher than the linear regression.

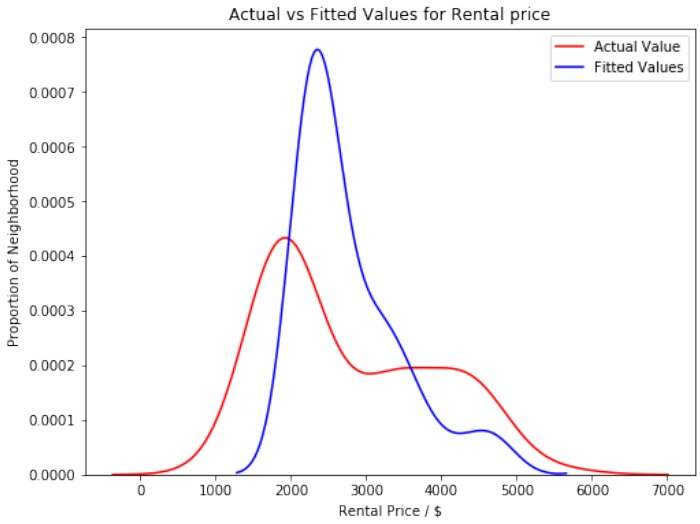

Figure 6 Model evaluation on multiple variable polynomial fitting

The fitting methods discussed above indicate that the three selected variables are positively correlated to the average rental prices to some degree. However, there is no function capable to describe the correlation accurately. Although increasing the degree generally improves the deviation, but this may introduce overfitting.

By using the Restaurant, Entertainment and Miscellaneous Convenience as the features, clustering can be performed. The K-means method is employed here to cluster these neighborhoods. The rental price ranges of those clusters that are represented by boxplots are further used to investigate whether the clusters can be used as indicators for the variation of rental prices. Ideally, the overlaps of the price ranges between different clusters should be minimised if the clusters greatly differ from each other with respect to the average rental price. After examining the above criteria for different cluster numbers, the result of three clusters turns out to be a relatively good choice.

	Cluster Label	Borough	Neighborhood	Average rental price / $	Latitude	Longitude	Restaurant	Entertainment	Miscellaneous Convenience
0	0	Manhattan	Marble Hill	1694	40.876551	-73.910660	0	0	7
1	0	Manhattan	Inwood	2225	40.867684	-73.921210	4	1	13
2	0	Manhattan	Washington Heights	2243	40.851903	-73.936900	3	0	13
3	0	Manhattan	East Harlem	3334	40.792249	-73.944182	6	2	11
4	0	Manhattan	Tudor City	3338	40.746917	-73.971219	0	0	9

	Cluster Label	Borough	Neighborhood	Average rental price / $	Latitude	Longitude	Restaurant	Entertainment	Miscellaneous Convenience
9	1	Manhattan	Murray Hill	4022	40.748303	-73.978332	14	10	23
19	1	Manhattan	Greenwich Village	4415	40.726933	-73.999914	18	13	25
24	1	Manhattan	Upper West Side	4654	40.787658	-73.977059	21	10	22
25	1	Manhattan	Lincoln Square	4706	40.773529	-73.985338	5	0	30
26	1	Manhattan	Little Italy	4845	40.719324	-73.997305	31	1	32

	Cluster Label	Borough	Neighborhood	Average rental price / $	Latitude	Longitude	Restaurant	Entertainment	Miscellaneous Convenience
8	2	Manhattan	Financial District	4005	40.707107	-74.010665	19	9	72
17	2	Manhattan	East Village	4334	40.727847	-73.982226	28	36	36
18	2	Manhattan	Chelsea	4359	40.744035	-74.003116	26	6	65
21	2	Manhattan	West Village	4524	40.734434	-74.006180	35	28	37
23	2	Manhattan	Chinatown	4609	40.715618	-73.994279	59	3	38

In spite of the overlap of the price ranges is tremendous, the median price of each cluster splits significantly, shown by the Figure 7. Further looking into the three variables of each cluster, the amount of the venues represented by the value of the three variable is positively correlated to the median prices, which shows consistency with the observation of the fitting result.

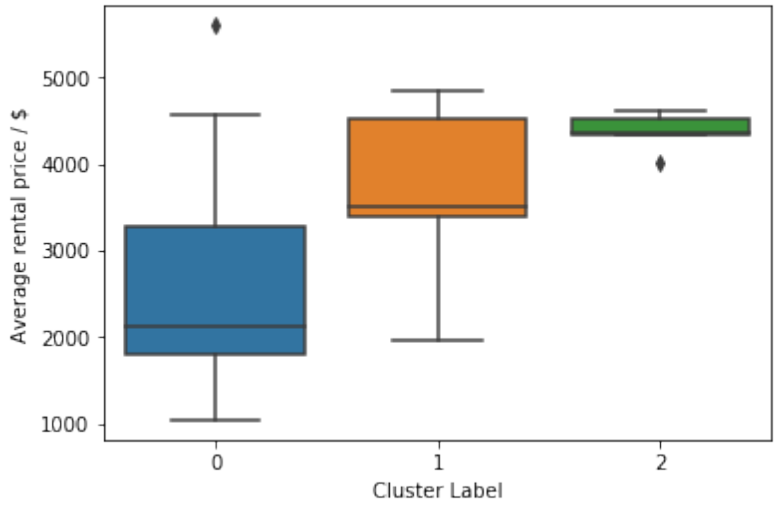

Figure 7 Boxplot of the average rental price

From the above analysis, it demonstrates that no empirical function can be found to accurately describe the relationship between the rental prices and the found venues in corresponding neighbourhoods. Nevertheless, the results do show the activeness of FourSquare usage, i.e., the amount of venues found by FourSquare, is positively correlated to the rental price. Among the found venues, venues belonging to Restaurant and Miscellaneous Convenience have stronger influence on increasing the rental price.

From another perspective, the level of the rental price also implies the wealthy condition of the people living in the corresponding neighbourhood. Wealthier people can afford higher rental price, suggesting that they possibly have more money to spend on other activities, such as dining outside, joining recreations and visiting different venues. In general, these people have high inclination to use application like FourSquare to record their life, contributing to the sufficient amount of result stored by FourSquare, and eventually to the positive correlation between the rental price and the venues.

Conclusion

In a nutshell, how the venues impact the rental prices is studied by some data science methodologies in this project. It proves that the increasing convenience by having more Restaurant and Miscellaneous Convenience types of venues can elevate the rental price of a neighborhoods.

However, this project also suggests the importance of the structure of data used for analysis. Better recommendation can be achieved by classifying the venues into less detailed but distinct categories.