UCL DEPARTMENT OF GEOGRAPHY

YEAR 2020-21

MODULE CODE:	**GEOG0115**
MODULE NAME:	**Introduction to Social and Geographic Data Science**
COURSE PAPER TITLE:	**New York City Property Price Prediction**
WORD COUNT:	**2497**

Your essay, appropriately anonymised, may be used to help future students prepare for assessment. Double click this box to opt out of this ☒

GEOG0115 Introduction to Social and Geographic Data Science

New York City Property Price Prediction

Introduction

New York City involves the most expensive and competitive real estate markets in the United States, and has been one of the highest-yield markets over the world. Even during the previous national recovery period, the real estate market in New York was still thriving (Goff, 2002). As mentioned by NYCREC (2018), the scarcity effect has created a consistently valuable real estate market for New York City, with enhanced value and desirability of properties affected by limited availability. That is because New York City consists of many islands, where the real estate is limited by the land and the height is determined by zoning laws. The promising investment prospect of the real estate markets in New York City strengthens the necessary and significance of the related research. Therefore, this essay will observe the spatial patterns of New York City real estate market and provide guidance or insights for the potential investment.

Generally, there are plenty of theories and models about property value prediction from economic and geographic perspectives. Among them, Gibbons and Machin (2008) have emphasized the influence of three policy-relevant factors on housing prices, including transport accessibility, education quality and security conditions. Similarly, Orford (2002), also mentioned these factors and other locational externality effects, considering the accessibility to various services and living quality. However, those models all require the external information that is difficult to obtain and would become the potential barrier of real estate market prediction in future. Therefore, this essay aims to only involve the inherent property-level characteristics of the real estate to investigate the current New York City property price, identify influence factors and make predictions for housing price.

Data and Methodology

The applied main dataset, Rolling Sales Data, is from New York City government, which contains the transaction information of New York City real estate market from October 2019 to September 2020 (NYC, 2020). The dataset consists of neighborhood, unit, year built, square feet, sale price and other information. In order

to integrate the statistical data with geography, another dataset (ZIP code boundary data) will also be used, containing the latitude and longitude values of observations (NYC, 2018).

The research on New York City real estate market will apply both statistical and spatial approaches, to identify whether there is useful knowledge that could be observed from a year's worth of transactions. At the beginning, descriptive statistics (such as boxplot analysis) will be calculated to remove the outliers and have a brief overview of the property price in New York City. After that, it is necessary to map the sale price by merging the location data with statistical data, in order to explore general underlying spatial patterns. Then, machine learning linear predictive models will be built for five boroughs to provide better predictions for unseen data points properties. To analyze the accuracy of the models, r-squared and residual standard error values are all calculated. Besides, sale price distribution maps for test datasets and its predictions are plotted and compared to visually evaluate the accuracy of the models. Based on the report of the linear models, the influence degree of related variables could be recognized. To further verify these relationships, principal component regression models for boroughs are applied.

Descriptive analysis

New York City has five boroughs, and their relative positions can be recognized in Figure 1. Since each borough has specific social, economic, political and cultural conditions, property price might probably be affected accordingly. Corresponding to the color of each borough in Figure 1, Figure 2 represents the unit property price level of them. It is obvious that unit house price in Manhattan is the highest one with larger price span, while that in Bronx is the lowest one with the median value of just over 250$ per square feet. Generally, most data points could be included with the unit price of less than 2000$, which could be suggested as the requirement of effective data points.

Figure 1: Relative position of five boroughs of New York City (Source: Author's own)

Figure 2: Unit property price in New York City boroughs (Source: Author's own)

After filtering out the duplicate data, outliers, temporarily unused columns and missing values, the New York City Rolling Sales Data and the ZIP code boundary data with spatial components can be merged. By locating the recorded houses, the geographical variation of house price in New York City can be visualized. As shown in Figure 3, crimson colors represent high housing price, while light colors are the opposite. Manhattan is the borough with the most expensive houses with over 5 million dollars each, which might be related to the high unit property price. Following that, western part of Brooklyn and northern part of Queens also have higher house prices, around 2 million dollars each. Besides, Bronx southwest has amount of overpriced houses, which seems influenced by neighboring Manhattan regions.

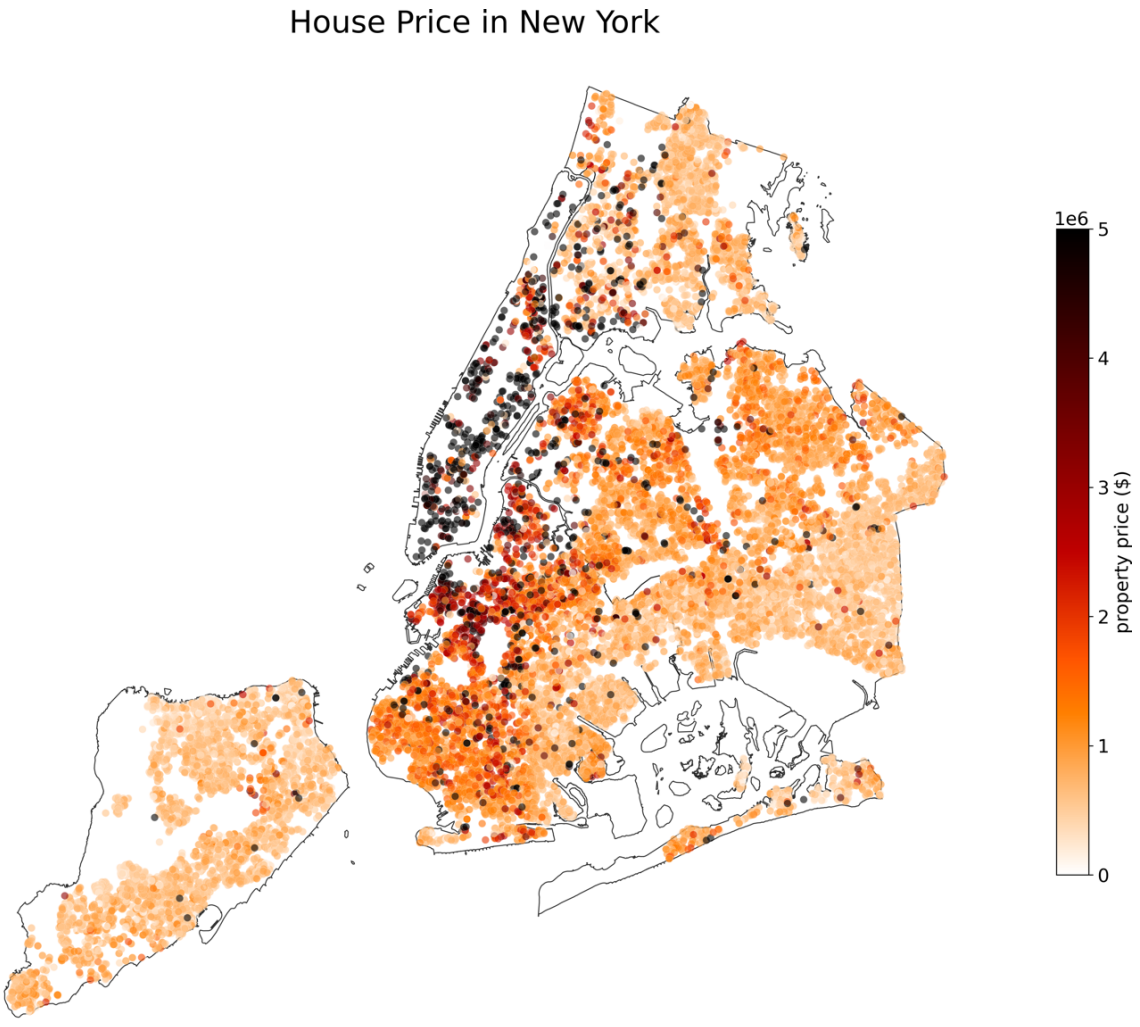


Figure 3: Geographical variation of house sale price in New York City
(Source: Author's own)

Linear predictive model

Because of the obvious difference between boroughs, the research decided to build 5 ordinary least squares (OLS) regression models for 5 boroughs of New York City, in order to identify detailed relationships between house price and influence factors. Before building the models, correlationship between independent variables and linear relationship between the dependent and independent variables have been tested, preventing from involving unrelated factors or multicollinearity. Considering the model fit evaluation, r-squared and root mean squared error (RMSE) are applied. Higher r-squared value means the model fits better, while lower RMSE value means less difference between original dataset and predicted dataset. According to Table 1, the predictive model for Bronx is the most accurate model with around 60 percent accuracy, but the model for Staten Island does not perform well because of the limited related variables. Besides, considering the RMSE for five boroughs, the value for Manhattan is much higher than others, which means the predicted values for certain points would differ greatly from the actual values. In contrast, that value for other boroughs seems relative satisfactory.

Table 1: Accuracy test for 5 linear regression models (Source: Author's own)

	r-squared	root mean squared error (le6)
Bronx	0.593	0.945
Manhattan	0.422	4.667
Queens	0.372	2.033
Brooklyn	0.355	1.055
Staten Island	0.278	1.144

In order to further evaluate the models and have an overview of their performance visually, the research has visualized the housing price data. As shown in Figure 4, the house price variation of model test dataset over space (on the left) and the predicted house price variation of the same points (on the right) have been represented. Because of the satisfactory RMSE values as mentioned above, the overall distributions of houses with different prices seem quite similar for actual and predicted conditions. However, when comparing the maps carefully, it is apparent that the house price in Queens and Brooklyn has been overvalued on the whole.

Besides, several cheaper houses (for around 2 million dollars) in Manhattan would be estimated to be expensive ones (for over 3 million dollars), which corresponds to the calculated higher root mean square error. The unsatisfactory performance for Manhattan model might possibly be relevant to the inadequate valid data, because there are plenty of data points in this borough have no value for the selected independent variables of OLS models. Apart from that, the difference for Bronx and Staten Island could not be identified easily. The reason for Bronx case might be the high model fit (with the highest r-squared), while the reason for Staten Island might be related to the overall low housing price and spatial homogeneity.

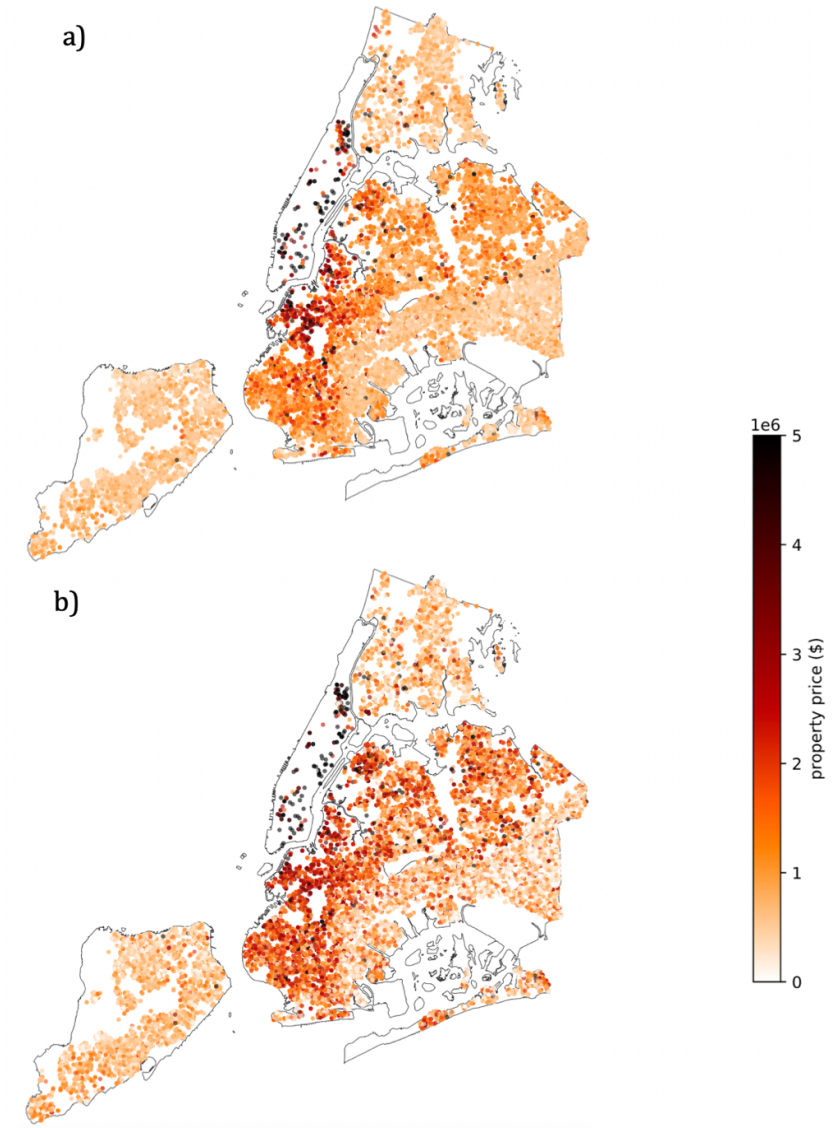


Figure 4: Comparison between actual and predicted house price of model test dataset, where a) represents actual values; b) represents predicted values (Source: Author's own)

The influence of related variables for house price can be recognized in the linear regression model report (see Table 2). The characteristics concerned in the models include gross square feet, land square feet, year built, residential units, commercial units and location elements, which have all been evaluated as having a linear relationship with house sale price for specific boroughs. Considering the process, standardization approach has been applied for input data to ensure the consistent influence weight of different attributes and further reduce the multicollinearity (Dabbura, 2018). In that case, estimates in the result report could only reflect the relative coefficients. According to the significance codes for selected features, most of them are statistically significant (p-value <0.01), apart from those of Manhattan. Generally, the most relevant elements are gross square feet and land square feet, which are all positive influence factors. Considering other factors, residential units amount represents positively significant effects on the house price in Bronx, accounting for 0.511. Besides, commercial units amount represents positive effects on the price in Queens and Staten Island, accounting for 0.105 and 0.214. In addition, the location of the properties could be seen as an importance influence factor, which reflects the spatial heterogeneity within boroughs.

Table 2: Results of linear predictive models, showing the relationship between house price and related characteristics (Source: Author's own)

	estimate & significance for house price (boroughs):				
variables	Bronx	Manhattan	Queens	Brooklyn	Staten
(intercept)	0.000	0.000	0.000	0.000	0.000
gross square feet	0.197***	0.655***	0.344***	0.368***	0.311***
land square feet	0.163***	0.027	0.284***	0.158***	0.180***
year built	-0.022	0.063	-0.013	0.013	0.105***
residential units	0.511***	-0.039	-	0.053**	-
commercial units	-	-0.092 .	0.105***	-	0.214***
latitude	-0.039**	-0.148**	0.035***	0.171***	-
longitude	-0.0073	-	-0.121***	-0.182***	0.059***

significance codes: '***'0.001, '**'0.01, '*'0.05, '.'0.1

Principal component regression model

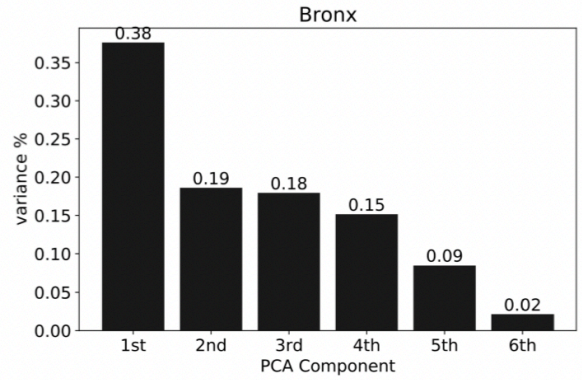
To further evaluate the relationships between housing price and various attributes, the principal component analysis (PCA) has also been conducted, which applies the dimensions reduction concept to compress the information from existing features and create principal components (Bork & Møller, 2018). Table 3 shows the accuracy test results of the models for New York City boroughs, which perform similar to the OLS linear regression models. According to the results, the accuracy of model for Bronx ranks the first, followed by that of Manhattan, Queens and Brooklyn. Besides, the RMSE values for PCA models are also similar to the OLS models with slight difference.

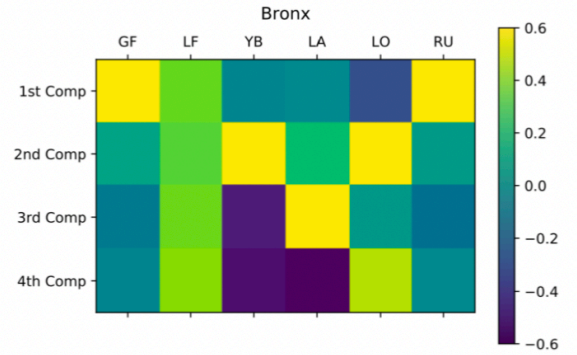
Table 3: Accuracy test for 5 PCA regression models (Source: Author's own)

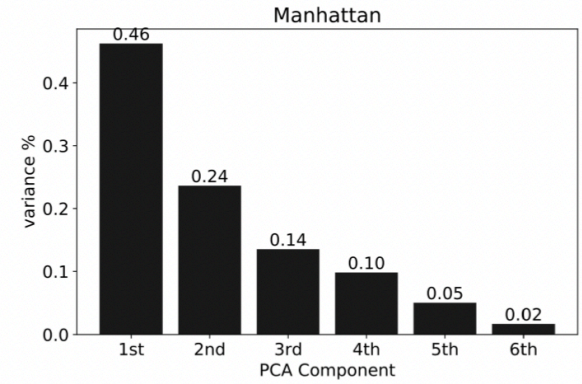
	r-squared	root mean squared error (le6)
Bronx	0.567	0.945
Manhattan	0.396	4.667
Queens	0.373	2.032
Brooklyn	0.362	1.049
Staten Island	0.177	1.142

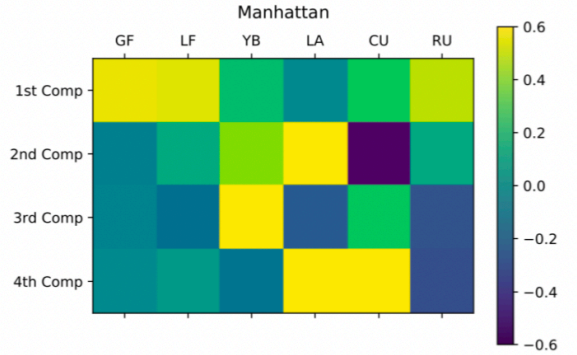
The influence factors have been recognized by OLS models, which have also been verified by PCA models. For PCA analysis, principal components could be identified from the same characteristics mentioned in OLS models, including gross square feet (GF), land square feet (LF), year built (YB), residential units (RU), commercial units (CU), latitude (LA) and longitude (LO). Different principal components could capture the variance that is independent to each other, where one principal component is able to represent several features in a lower dimension (Bork & Møller, 2018). Among them, the first principal component always could possibly capture the maximum variance of the data. As Figure 5 represents, first principal component for 5 boroughs can all capture over 30 percent variance. Variance representation of other PCA components and the composition of every component have also been displayed in Figure 5. Although it is relatively difficult to interpret, the similar significance of these attributes can be noticed. For example, the significance of GF

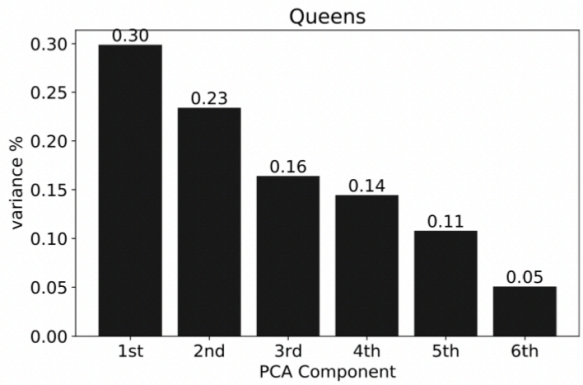
for housing price (mainly captured by the first component) could be recognized in each borough. Similarly, the number of residential units for Bronx and Brooklyn seem highly relevant to the housing price (captured by the first component), as well as the commercial elements for Queens and Staten Island (captured by the third and fourth components).

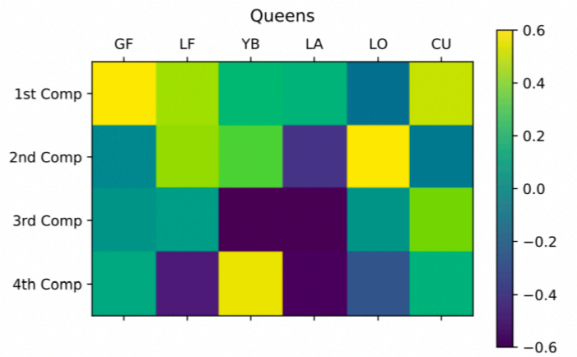












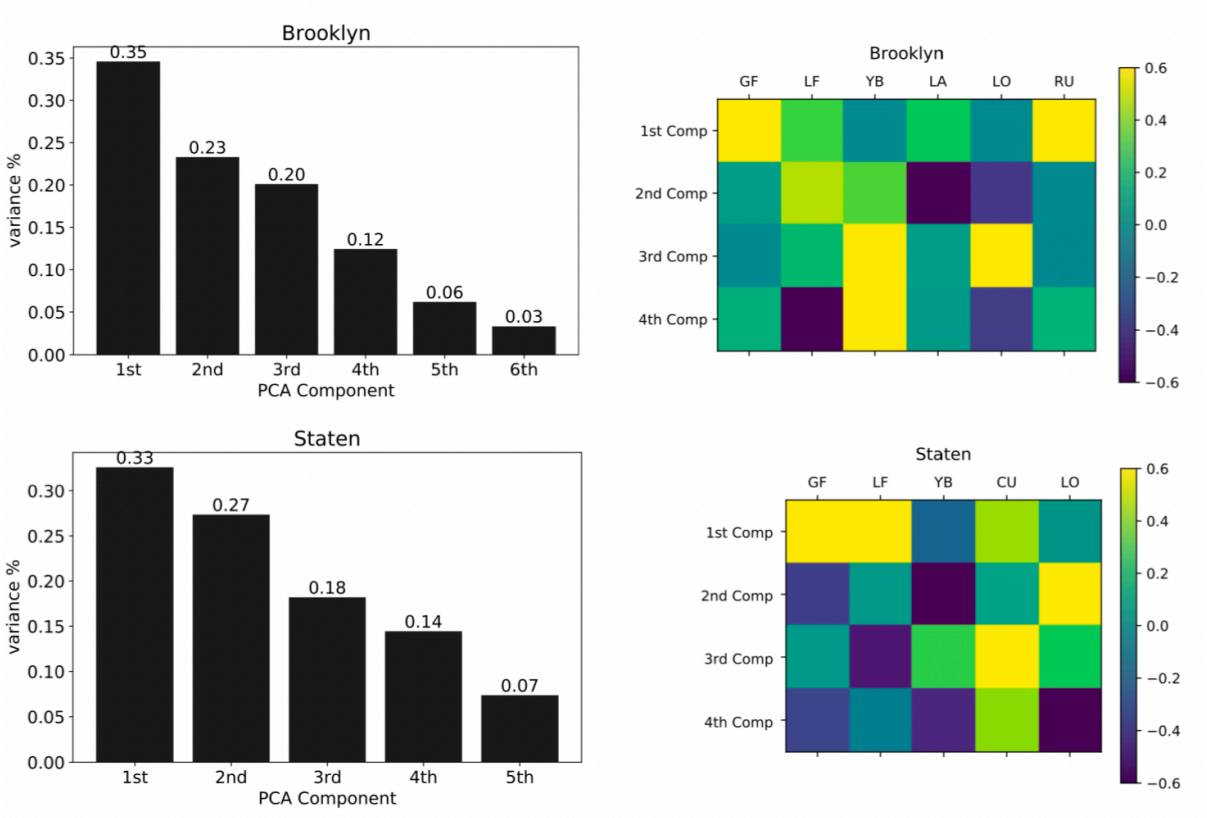


Figure 5: PCA results for boroughs in New York City, showing the variance captured by each PCA component (left) and the influence of each variable in the first to fourth components where yellow represents greater impact (right)
(Source: Author's own)

Discussion

In the descriptive analysis part, the essay has indicated the differences between boroughs in New York City. From Figure 3, the geographical variation of house sale price has presented a spatial heterogeneous pattern over the whole region. Especially, areas in Manhattan and surrounding Manhattan within its neighboring boroughs all have relatively higher housing price, forming the cluster. It is inevitable, since Manhattan is the administrative, economic and cultural zenter of the world, where United Nations Headquarters and the Wall Street are located.

Considering the results of OLS and PCA models for these boroughs, the similarity could be seen, which has double verified the findingts. As represented in Figure 4, it is possible to successfully predict the housing price in New York City just based on the annual transaction records. In comparison with geographical variance of actual sale prices, although the mapping of predicted data could capture the overall spatial pattern, it would overvalue the properties in some regions. Besides, the similar pattern in the boroughs, such as Staten Island, might because of the lower housing price on the whole not the accuracy. As mentioned by Kiprop (2018), Staten Island, as the least populated and suburban borough even not served by subway system, has been felt to be politically neglected. Considering the property-level influence factors, property area and land use conditions have been identified as the most relevant characteristics, similar for OLS and PCA models.

However, the accuracy of the OLS and PCA models is not perfect enough. Models for Bronx perform better than others with approximate 0.6 for r-squared, while others only account for around 0.4. As mentioned above, large numbers of missing values might be one reason. According to Austin et al. (2012), missing values and outliers would probably result in serious deviation and false prediction. Additionally, since this research aims to find out whether it is possible to make predictions for New York City's housing prices just using property-level information, it is promising to improve the models by adding other independent variables in the future. As mentioned at the beginning, previous studies have explored other external influence factors for local housing sale price, such as transportation, education and security elements. After adding those elements, r-squared and RMSE values would be better, which should be concerned in further research.

In addition, the analysis also has limitations because of the influence of different conditions and policies through COVID-19 pandemic period (from 2020 to 2021). According to Santarelli (2021), New York City as one of the epicenters of the pandemic has facilitated the urban-to-rural population flow, bidding up the property prices in suburban regions gradually. In contrast, those sticking around in the city could get an opportunity to purchase better houses with lower sale prices. Therefore, after the market recovery, the accuracy of the prediction should be further evaluated.

Conclusion

In conclusion, the research has successfully made general predictions for housing price in New York City, using inherent property-level characteristics. At the beginning, the research has described the existing situation of New York City housing price, using boxplot and point data plotting approaches. Manhattan and its surrounding areas could be seen as the cluster center for overpriced houses. After that, linear predictive models and PCA models for 5 boroughs have been built respectively, in order to judge the accuracy and significance of the relationships between housing price and property-level characteristics. From the similar results of OLS and PCA models, it is confident to propose that predicting the overall pattern for housing price based on these selected variables, including area information, year built, land use and location elements, is possible but the details are not that accurate. In order to improve the accuracy of the forecasts, adding relevant external elements mentioned in previous research is required. Besides, the occurrence of COVID-19 pandemic might affect the current housing price, therefore, the research results should be further verified.

Reference list

Austin, E. et al. (2012). 'A framework for identifying distinct multipollutant profiles in air pollution data', *Environment International*, 45, pp. 112-121.

Bork, L. & Møller, S. (2018). 'Housing Price Forecastability: A Factor Analysis', *Real estate economics*, 46 (3), pp. 582-611.

Dabbura, I. (2018). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a (Accessed: 20 January 2021).

Gibbons, S., and Machin, S. (2008). 'Valuing school quality, better transport, and lower crime', *Oxford Review of Economic Policy,* 24 (1), pp. 99-119.

Goff, L. (2002). 'A hell of a town; Despite a glut of sublet space, New York's real estate market is doing better than other cities', thanks to lower vacancy rates and higher rents', *Crain's New York business*, 18 (42), p. 29.

Kiprop, V. (2018). *The boroughs of New York City-NYC boroughs map.* Available at: https://www.worldatlas.com/articles/the-boroughs-of-new-york-city.html (Accessed: 20 January 2021).

NYC (2018). *ZIP code boundaries.* Available at: https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u (Accessed: 20 January 2021).

NYC (2020). *Rolling sales data*. Available at: https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page (Accessed: 20 January 2021).

NYCREC (2018). *A brief history of New York City's real estate market*. Available at: https://medium.com/@teamnycrec/a-brief-history-of-new-york-citys-real-estate-market-841a724439ca (Accessed: 20 January 2021).

Orford, S. (2002). 'Valuing Locational Externalities: A GIS and Multilevel Modelling Approach', *Environment and planning. B, Planning & design*, 29 (1), pp. 105-127.

Santarelli, M. (2021). *New York real estate market: prices | trends | forecast 2021.* Available at: https://www.noradarealestate.com/blog/new-york-real-estate-market/ (Accessed: 20 January 2021).