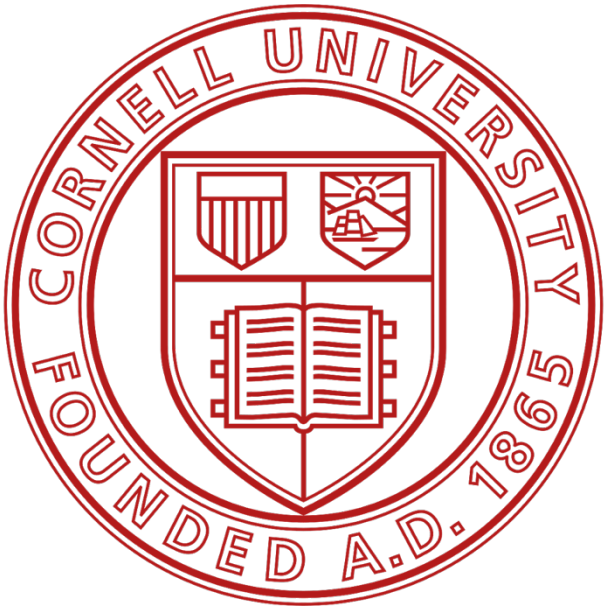



CORNELL UNIVERSITY

DEPARTMENT OF STATISTICAL SCIENCE

# Final Report on Comparing Housing Sales Factors in Manhattan and Brooklyn in 2018

**Author**: Shiyun(Shirly) Wang (sw983)

**Advisor**: Ziye Zhang

Dec 15, 2019

# TABLE OF CONTENTS

## Executive Summary

### CPR5850 Final Project

CPR5850 is a small-sized seminar-based class, which perfectly combines economics theories for US and China housing markets and data techniques necessary to conduct empirical studies on urban housing markets. It focuses on urban residential markets and provides rich hands-on experience with data analysis using Python and Jupyter Notebook. While learning about housing demand and supply, housing markets, and experimenting with data collection and modeling, students would get curious about some specific questions and initiate their own projects.

In general, the final project should display students' ability to research on a question, collect data, manipulate housing data and conduct qualitative and quantitative analysis.

# 1 INTRODUCTION

## 1.1 Background

As I'm graduating in December 2019 and will start to work in New York City in 2020, I started to search for information about housing market in New York City and areas around in this semester. I think about renting an apartment in the midtown of Manhattan or ones in the downtown Brooklyn. During the process of comparing the housing prices in Manhattan and Brooklyn, I had some intuitive sense on some factors affecting the housing prices, but it's still hard to generate the exact patterns or trends among so many different factors and data, and offer any quantitative proof. Thus, my research question came as what factors affect the housing prices in these two regions and how much influence they impose on the housing prices. Combined with the regression models mentioned and interpretations we

generated in class, I think by running a regression model on some of the attributes to make predictions on housing prices should be a good solution for my question. Since the housing data is highly correlated with geography and location, I decided to use gwr package in python and also geopanda for data frame, which I'll explain more details later in this report paper.

### 1.2 Social Background

HOUSING has been the long-term focus of urban development and social and economic policies. Besides the new graduates like me, I believe there are millions of people who are interested in housing markets for various purposes, such as to investment in housing markets, to buy or sell for normal households, to make a decision on renting a house or buying one, or to conduct academic research, etc.

Homeownership creates benefits not only for individuals and households, but also for the whole communities. If so, with the obvious benefits, why people still hesitate a lot on purchasing a house or apartment. The answer should be that despite the benefits of homeownership and the motivations to own a house, affordability has become a critical issue. In my opinion, people's concerns over the affordability of housing come from two major facts: housing cab be the largest expenditure component in the budgets of most individuals or families, and many metropolitan areas, like New York city, have experienced a stark increase in housing prices and rentals. Affordability can be affected by several different factors, including housing price, housing quality, level of household income, borrowing ability of a household, public policies and so on. So, a quantitative analysis and explanation on factors influencing housing prices would offer people a more rational way to make a decision on how much they would like to pay for a house and the better location or time for them to purchase a house.

## 2 DATA PREPARATION

## 2.1 Data Collection

Initially, I got the inspiration of where to get the data with both housing attributes and geographic information from our class. We talked about housing sales for Manhattan in 2009 and I found that dataset is exactly what I need. So, I logged onto NYC's government website and found out the data I need in Department of Finance. I downloaded the dataset for Manhattan and Brooklyn in 2018.

## 2.2 Data Framing and Cleaning

After I downloaded the data, I need to frame my data to the format required for running the regression model. The very first questions came to my mind is that since housing is a normal good, is it appropriate to make the total housing prices the response variable in my model? It is a common sense that the larger the house is, the higher the price should be. So with the suggestions from Prof. Zhang, I decided to make unit price of the housing as my response variable. Unit price is calculated by two existing columns in the dataset—'Sale Price' divided by 'Land Square Feet'. When I looked at the newly-created column 'Unit Price', there were several infinite numbers and extreme values. That made me go back to 'Sale Price' and 'Land Square Feet' columns to filter out extreme values such as 0, infinity, and some illogical small values for 'Sale Price'. After cleaning the data, the newly-created 'Unit Price' looked better.

Here I would like to mention the reason why I did not make categorical data for 'ZIP CODE', 'BLOCK' and 'LOT'. It seems that these three variables are like the location ID for housings in New York and it should be made as dummy variables to be used in modeling. However, after I checked with the distributions of these three variables on map, it's not hard to see that the numbers are logically distributed. Surely, New York City and Brooklyn are

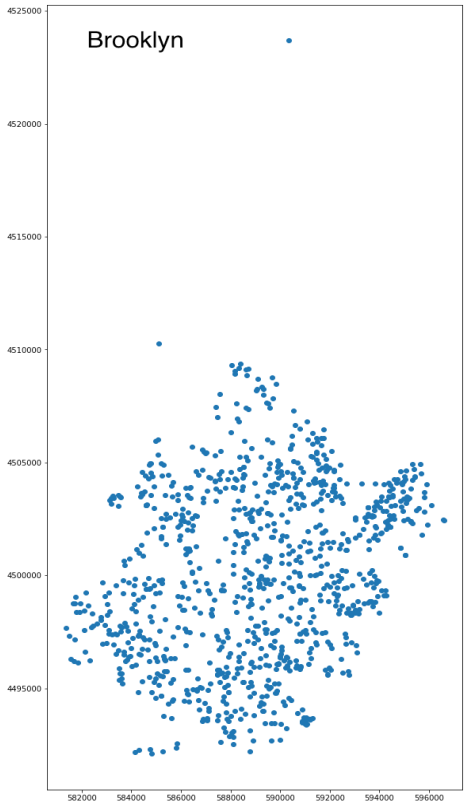
rationally designed and numbered, so there's no need to exceptionally make categorical variables for them and this can be proved in the plots I display in part 4.

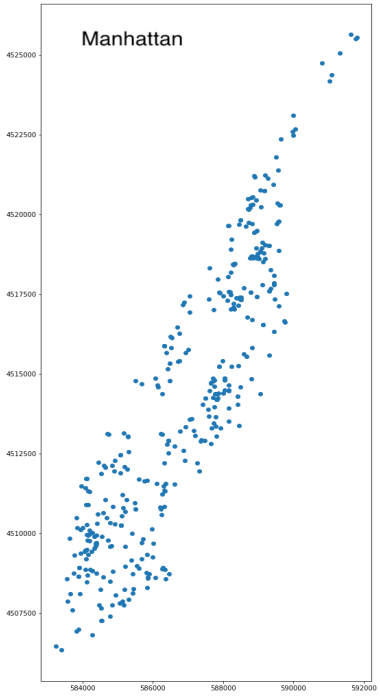
The next important step is to transform the normal text address in the dataset to geocode, which would be used later in the gwr regression model, since I care about the geographic location of the housing. For API preparation, to use the Google API, we need to prepare two things: Google API account and Google API library for Python. With these two things done, I used gmaps.geocode to geocode the addresses in the data frame. However, with approximately 20,000 data points in Brooklyn and Manhattan datasets respectively, it's almost impossible to geocode all of the addresses on my personal laptop. My solution is to geocode 1000 random data points from the datasets. In order to get unbiased, randomly selected data, I used the sample() in pandas package to generate the data points. After geocoding all these 1000 addresses in the dataset, we can see that the geocoding result is a list containing two dictionary elements, representing two possible geocoding results corresponding to the original address. From those two possible results, it's not hard to see that only one of the results refers to my target location—Manhattan or Brooklyn. The other result is wrong simply because there's no specification of the state and the city of the address. I need to remove all whitespaces at the end of the addresses and add ", Manhattan, NY" or ", Brooklyn, NY" to the end, which was accomplished by function "address_refine" created by my own. After applying this function to each address, I extracted the geographic coordinates—longitude and latitude from the result.

Then, to do a spatial analysis on a data frame, I need to create a geo data frame by using GeoPandas in python. For the geometry list, I defined the coordinates got before as Point geometric objects using function Point() *from* shapely.geometry. To create the geo data frame, I combined a data frame, a coordinate reference system(in our case, "epsg: 4326"), and the geometry list. A new column "Geometry" was added to the original data frame. The

geographic coordinates obtained from Google Geocoding API are in WGS84, the true coordinates on the spheroid. Thus, I projected them into projected coordinates (flat plane, x-y in meter). By importing Proj and transform functions from the library pyproj, I projected input geographic coordinate system used by Google API, i.e., WGS84 (epsg: 4326) into output projection coordinate system-WGS 84 / UTM zone 18N (epsg: 32618). To project the coordinates of all addresses together, I first defined a new function proj which converts a coordinate tuple into a projected coordinate tuple. Then, I wrote a loop over all coordinate tuples in the data and apply the function proj to each of them.

In the final stage of data framing and cleaning, plotting is important to help us to make sure about data quality and to understand our data. The plots for data points for Brooklyn and Manhattan are attached below. It is not hard to see the basic shapes of Brooklyn and Manhattan from the plots below and thus the data quality is basically good.




## 3 Multiscale Geographically Weighted Regression (MGWR)

### 3.1 Introduction to MGWR

Geographically weighted regression (GWR) is a spatial statistical technique — a local form of linear regression that is used to model spatially varying relationships. GWR is an improvement of traditional global regression models, as global regression models can be limited when spatial processes vary with spatial context. GWR captures process spatial heterogeneity by allowing coefficients to vary across space. To do this, An ensemble of local linear models are calibrated at any number of locations by 'borrowing' nearby data. which provides a set of location-specific coefficient estimates for each relationship in the model. Hence, the coefficient estimates are allowed to vary spatially, as well as a calculated bandwidth parameter that provides intuition about the geographic scale of the processes. A further extension to this framework allows each relationship to vary according to a distinct spatial scale parameter, and is therefore known as multiscale GWR.

In this project, I mainly used the mgwr, a Python-based implementation of MGWR that explicitly focuses on the multiscale analysis of spatial heterogeneity. MGWR does not impose a restriction on its assumptions as GWR, since MGWR allows relationship between the response and a covariate to vary both locally and regionally or not vary at all. Through Eliminating the restriction that all relationships should vary at the same spatial scale, we can largely minimize over-fitting, reduce bias, and mitigate collinearity due to similar functional transformations. Therefore, MGWR is highly recommended as the default local model when using GWR to investigate spatial heterogeneity and scale.

### 3.2 MGWR Functionality

As mentioned before, compared with traditional GWRA, MGWR provides an extension that allows each variable to be associated with a distinct bandwidth by recasting GWR as a model such that:

$$y = \sum_{j=1}^{k} f_j + \epsilon,$$

where fj is a smoothing function applied to the jth explanatory variable that is characterized by distinct bandwidth parameter.

In my project, I manually selected the independent or exploratory variables, which are of my personal interest, and they are 'BLOCK', 'LOT', 'ZIP CODE', 'RESIDENTIAL UNITS', 'COMMERCIAL UNITS', 'TOTAL UNITS', 'LAND SQUARE FEET', 'GROSS SQUARE FEET', 'YEAR BUILT', and 'TAX CLASS AT TIME OF SALE'.

### *3.2.1 Standardizing Variables*

In order to compare different bandwidths obtained from an MGWR model in the next step, it is necessary to scale both the dependent and the independent variables so that they are in the same scale and scattered in the same range of variation. Here, I used preprocessing.scale() function from python's sklearn to scale the data before moving to the next step.

### *3.2.2 Bandwidth Selection*

MGWR uses a back-propagation fitting algorithm for model tuning, which processes through sequentially tuning a set of univariate GWR models based on the partial residuals from the previous iteration until the MGWR model converges to a certain values in the interval we set. Two primary differences arise in how an MGWR model is specified and calibrated in mgwr when compared to GWR.

To select bandwidth, I first imported pysal library and used Sel_BW function. A Sel_BW object is necessary for obtaining model results because parameter estimation occurs simultaneously with bandwidth selection and therefore, the Sel_BW object must be passed to

an MGWR object in order to carry out MGWR calibration. The bandwidths for my Brooklyn model and Manhattan are 139.0 and 222.0 respectively.

### 3.3 Model Fit

Since all variables and bandwidths were ready to use, I fit the model by inputting coordinates, response variable and independent variables, used the bandwidths calculated before and set kernel to 'bisquare'. GWR results, diagnostic information and summary statistics for parameter estimates are displayed by summary.

### 3.4 Results

Though it displays an R2 and adjusted R2 to assess model fit accuracy for MGWR, we still need to use a model fit criterion to account for model complexity, such as the AICc:

$$AIC_c = 2n \log_e \left(\frac{RSS}{n}\right) + n \log_e (2\pi) + n\left\{\frac{n + tr(\boldsymbol{S})}{n - 2 - tr(\boldsymbol{S})}\right\},$$

where *n* is the number of observations, S is the hat matrix, and RSS is the residual sum of squares. The results for Brooklyn is attached below:

```
Geographically Weighted Regression (GWR) Results
------------------------------------------------------------------------------
Spatial kernel:                                     Adaptive bisquare
Bandwidth used:                                              139.000

Diagnostic information
------------------------------------------------------------------------------
Residual sum of squares:                                     397.404
Effective number of parameters (trace(S)):                   143.271
Degree of freedom (n - trace(S)):                            856.729
Sigma estimate:                                                0.681
Log-likelihood:                                             -957.538
AIC:                                                        2203.618
AICc:                                                       2252.659
BIC:                                                        2911.665
R2:                                                            0.603
Adjusted R2:                                                   0.536
Adj. alpha (95%):                                              0.004
Adj. critical t value (95%):                                   2.898

Summary Statistics For GWR Parameter Estimates
------------------------------------------------------------------------------
Variable                   Mean        STD        Min     Median        Max
-------------------- ---------- ---------- ---------- ---------- ----------
X0                       -0.159      0.371     -1.982     -0.112      1.340
X1                       -0.218      0.395     -2.206     -0.122      0.956
X2                       -0.119      0.285     -1.236     -0.065      1.073
X3                       -0.007      0.128     -0.499      0.000      0.475
X4                      -26.358    807.894 -25548.122      0.144    198.847
X5                       -2.933     91.629  -2897.502      0.059     22.625
X6                       26.751    809.115   -199.188      0.106  25587.088
X7                       -0.691      0.572     -3.314     -0.601      0.329
X8                        0.242      0.649     -3.112      0.355      2.245
X9                       -0.204      0.495     -2.376     -0.041      0.700
X10                       0.220      0.465     -2.810      0.102      5.010
==============================================================================
```

## 4 Results Interpretation

By running the summary of the both models for Manhattan and Brooklyn, the parameter estimates and also the correponding variables in the models are attached below.

Brooklyn

| Variable | Est. | SE | t(Est/SE) | p-value |
|---|---|---|---|---|
| X0 | 0.000 | 0.028 | 0.000 | 1.000 |
| X1 | -0.350 | 0.030 | -11.638 | 0.000 |
| X2 | 0.002 | 0.028 | 0.080 | 0.936 |
| X3 | 0.053 | 0.029 | 1.830 | 0.067 |
| X4 | 0.005 | 849393.254 | 0.000 | 1.000 |
| X5 | 0.161 | 96332.933 | 0.000 | 1.000 |
| X6 | 0.171 | 850680.573 | 0.000 | 1.000 |
| X7 | -0.250 | 0.047 | -5.319 | 0.000 |
| X8 | 0.144 | 0.082 | 1.756 | 0.079 |
| X9 | -0.100 | 0.028 | -3.611 | 0.000 |
| X10 | 0.083 | 0.035 | 2.350 | 0.019 |

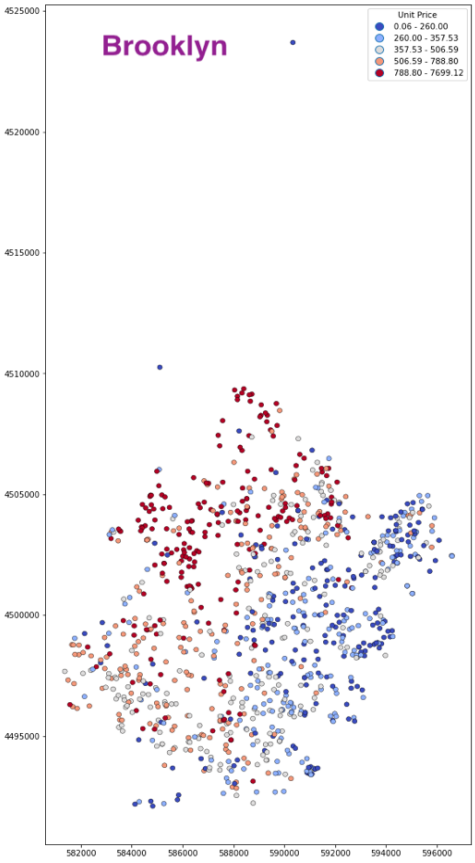
Manhattan

| Variable | Est. | SE | t(Est/SE) | p-value |
|---|---|---|---|---|
| X0 | -0.000 | 0.047 | -0.000 | 1.000 |
| X1 | -0.231 | 0.051 | -4.542 | 0.000 |
| X2 | 0.049 | 0.047 | 1.027 | 0.305 |
| X3 | 0.038 | 0.050 | 0.750 | 0.453 |
| X4 | 0.124 | 3536415.638 | 0.000 | 1.000 |
| X5 | 0.065 | 2525385.416 | 0.000 | 1.000 |
| X6 | -0.033 | 4284448.920 | -0.000 | 1.000 |
| X7 | -0.188 | 0.089 | -2.117 | 0.034 |
| X8 | 0.285 | 0.078 | 3.666 | 0.000 |
| X9 | -0.103 | 0.048 | -2.151 | 0.031 |
| X10 | 0.100 | 0.054 | 1.858 | 0.063 |

```
X0
X1   'BLOCK',
X2   'LOT',
X3   'ZIP CODE',
X4   'RESIDENTIAL UNITS',
X5   'COMMERCIAL UNITS',
X6   'TOTAL UNITS',
X7   'LAND SQUARE FEET',
X8   'GROSS SQUARE FEET',
X9   'YEAR BUILT',
X10  'TAX CLASS AT TIME OF SALE'
```

First, by looking at the parameter estimates in Brooklyn alone, it's easy to see that 'BLOCK' , 'LAND SQUARE FEET' and 'TOTAL UNITS' have a comparatively large impacts on unit price of housings. More specifically, we can conclude from the charts above that as the block of housings get farther away from Manhattan—center of New York, the unit price gets lower and a larger number of total units of the housing means a high unit price. However, the estimate for 'LAND SQUARE FEET' looks weird and make me feel not convincing at first. After carefully looking at the other parameters, I think there might be a correlation between 'LAND SQUARE FEET' and 'GROSS SQUARE FEET'. As suggested in the dataset, gross square feet refers to the total square footage of the usable areas a building has, not including both non-usable areas. It makes sense that the larger the gross square feet a housing has, the higher the unit price should be. But when a housing property

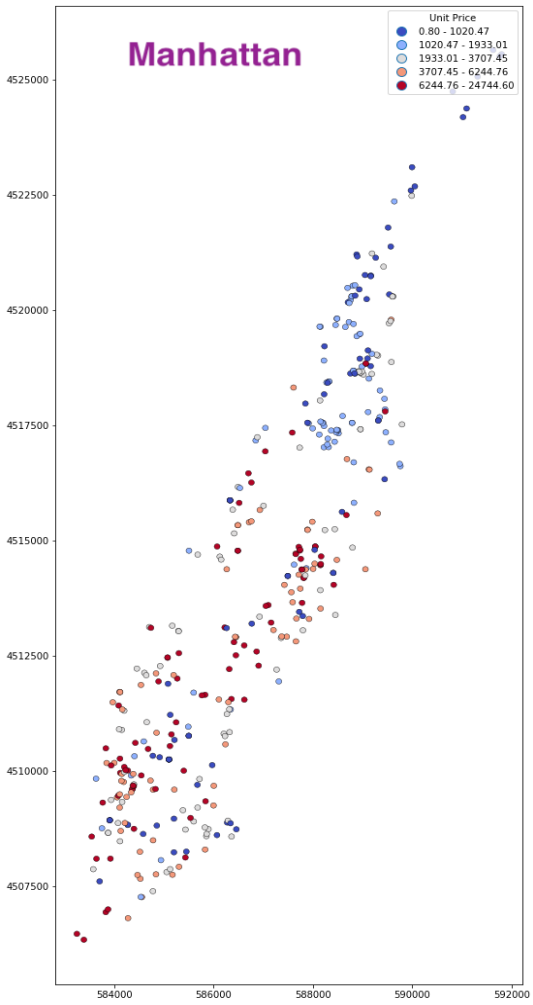
has a large gross area but a relatively small land area, it may negatively affects customers' willingness to buy this house and thus the unit price gets lower. The plot for unit price distribution in Brooklyn is attached below on the left and it's obvious that the housings closer to Manhattan have higher price. The relative positions between Manhattan and Brooklyn are shown on the right below.





Then, by looking at the parameter estimates in Manhattan alone, 'BLOCK', 'GROSS SQUARE FEET', 'LAND SQUARE FEET' and 'RESIDENTIAL UNITS' have more impacts on unit price of housings. From both the parameter estimates, we can conclude that for the housings that are farther away from the midtown of Manhattan, they have lower unit price. For 'GROSS SQUARE FEET' and 'LAND SQUARE FEET', there's the same problem as those in Brooklyn model. The larger the gross square feet a housing has, the higher the unit price should be. But when a housing property has a large gross area but a

relatively small land area, it may negatively affects customers' willingness to buy this house and thus the unit price gets lower. As for residential units, the more residential units, the higher the unit price that housing would have. The plot for unit price distribution in Manhattan is attached below and it's obvious that the housings near midtown and east village have higher price, which conform to our common sense.




As for the comparison between Manhattan and Brooklyn, the block of the housing has a larger effect on unit price in Brooklyn than in Manhattan, which may be explain in the way that the general housing price on Manhattan is higher than Brooklyn, so the difference of unit prices in Manhattan is not as large as Brooklyn, where unit prices vary tremendously if a housing is very far away from Manhattan or to say downtown Brooklyn. Another reason

might be Brooklyn is much larger than Manhattan, so the distance can affect the unit price in a more obvious way. And 'TOTAL UNITS' in Brooklyn matters more than that in Manhattan, which implies housing properties with different total units in Manhattan do not have as much difference in unit price as in Brooklyn. It may be caused by the fact that the housing density is much denser in Manhattan, so the housing properties with different total units may not have obviously different square feet, whereas the housing properties with different total units can have very obviously different square feet and that causes a big difference in unit price. And finally, due to the correlation between 'GROSS SQUARE FEET' and 'LAND SQUARE FEET', it's hard to directly compare the estimates in Manhattan and Brooklyn.

# 5 CONCLUSION

## 5.1 Regression Model Conclusion

From the quantitative analysis above, I can partially answer my research question in the beginning—what factors affect the housing prices in Brooklyn and Manhattan respectively, and how much influence they impose on the housing prices. 'BLOCK' , 'LAND SQUARE FEET' and 'TOTAL UNITS' have more impacts on Brooklyn's housing unit price, while 'BLOCK', 'GROSS SQUARE FEET', 'LAND SQUARE FEET' and 'RESIDENTIAL UNITS' impact Manhattan's housing unit price more.

The difference factors affecting housing unit price between Manhattan and Brooklyn mainly lies in the 'BLOCK' and 'TOTAL UNITS'. The block of the housing has a larger effect on unit price in Brooklyn than in Manhattan, and that means the difference of unit price for housing in downtown Brooklyn and suburban Brooklyn is larger than that in midtown Manhattan and suburban Manhattan. On the other hand, 'TOTAL UNITS' in Brooklyn

matters more than that in Manhattan, which means housing properties with different total units in Manhattan do not have as much difference in unit price as in Brooklyn.

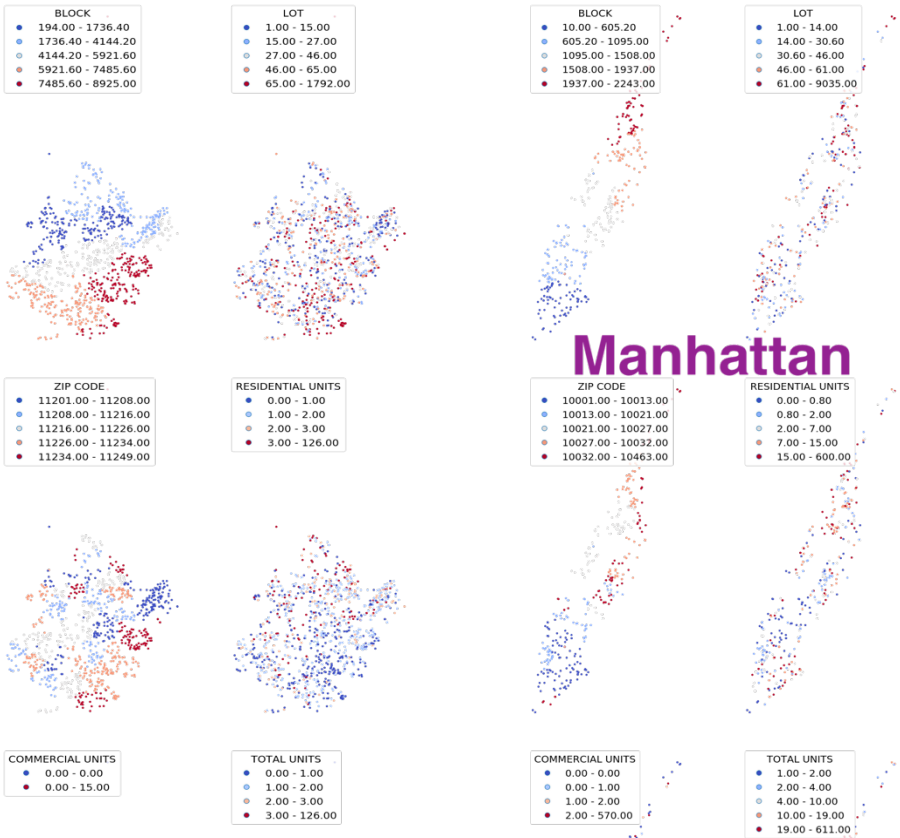
### 5.2 Discussion and Further Research

By looking at the parameter estimates in the output for two models, it's not hard to figure out the correlation between 'GROSS SQUARE FEET' and 'LAND SQUARE FEET', and thus it's hard to directly compare the estimates of these two parameters in Manhattan and Brooklyn and make any implications further. Another obvious flaw in the output is the p value for the estimates. Nearly half of the p values are high if we set significance level equal to 0.1 or 0.05, as usually in the statistical modeling. So, for the p values in the chart, I'm not sure if there's any problem in the data itself or the normal significance level I should use in housing market research.

As for the possible bias caused by lack of some significant variables, I think the information about if the housing property is second-hand may lead to a bias in year built. Second-hand housing properties should have lower unit prices and on the other hand whether second-hand could be highly correlated with the year when housing was build. So for the further research on this topic, I'd like to collect more information on if the housing properties are second-hand.

## APPENDIX

### Descriptive Analysis of Independent Variables


BLOCK
194.00 - 1736.40
1736.40 - 4144.20
4144.20 - 5921.60
5921.60 - 7485.60
7485.60 - 8925.00
LOT
1.00 - 15.00
15.00 - 27.00
27.00 - 46.00
46.00 - 65.00
65.00 - 1792.00
BLOCK
10.00 - 605.20
605.20 - 1095.00
1095.00 - 1508.00
1508.00 - 1937.00
1937.00 - 2243.00
LOT
1.00 - 14.00
14.00 - 30.60
30.60 - 46.00
46.00 - 61.00
61.00 - 9035.00
Manhattan
ZIP CODE
11201.00 - 11208.00
11208.00 - 11216.00
11216.00 - 11226.00
11226.00 - 11234.00
11234.00 - 11249.00
RESIDENTIAL UNITS
0.00 - 1.00
1.00 - 2.00
2.00 - 3.00
3.00 - 126.00
ZIP CODE
10001.00 - 10013.00
10013.00 - 10021.00
10021.00 - 10027.00
10027.00 - 10032.00
10032.00 - 10463.00
RESIDENTIAL UNITS
0.00 - 0.80
0.80 - 2.00
2.00 - 7.00
7.00 - 15.00
15.00 - 600.00
COMMERCIAL UNITS
0.00 - 0.00
0.00 - 15.00
TOTAL UNITS
0.00 - 1.00
1.00 - 2.00
2.00 - 3.00
3.00 - 126.00
COMMERCIAL UNITS
0.00 - 0.00
0.00 - 1.00
1.00 - 2.00
2.00 - 570.00
TOTAL UNITS
1.00 - 2.00
2.00 - 4.00
4.00 - 10.00
10.00 - 19.00
19.00 - 611.00
Brooklyn
LAND SQUARE FEET
304.00 - 1800.00
1800.00 - 2000.00
2000.00 - 2292.00
2292.00 - 2730.60
2730.60 - 133000.00
GROSS SQUARE FEET
234.00 - 1459.20
1459.20 - 2000.00
2000.00 - 2584.00
2584.00 - 3344.40
3344.40 - 120689.00
LAND SQUARE FEET
548.00 - 1699.00
1699.00 - 2177.40
2177.40 - 2605.00
2605.00 - 6249.40
6249.40 - 83092.00
GROSS SQUARE FEET
1160.00 - 3758.00
3758.00 - 6309.00
6309.00 - 10286.00
10286.00 - 27603.60
27603.60 - 754487.00
YEAR BUILT
0.00 - 1907.00
1907.00 - 1920.00
1920.00 - 1930.00
1930.00 - 1945.00
1945.00 - 2018.00
TAX CLASS AT TIME OF SALE
1.00 - 1.00
1.00 - 4.00
YEAR BUILT
0.00 - 1900.00
1900.00 - 1910.00
1910.00 - 1914.40
1914.40 - 1922.20
1922.20 - 2018.00
TAX CLASS AT TIME OF SALE
1.00 - 1.00
1.00 - 2.00
2.00 - 4.00

**Python Code**

See my attached python jupyter notebook file.

**Data Resource**

https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page