Analysis of Residential Real Estate Pricing in New York City

Team 57: James Barrett, Rui Dong, Aruna Kumari Gudivada, Huyen Nguyen, Qianwen Zhang

I. Introduction and Motivation

While real estate is the most visible investment class, it is surprising that it may also be the hardest asset to price. The most popular pricing method currently used in the market is to use comparables, such that the price of a property is manually determined by reviewing prices of other similar properties. This method is often subject to human bias as it involves a lot of judgement. Also, the manual nature increases the chance of error due to the limited number of properties can be sampled in each valuation. Often, the price is set too high, resulting in either the house sitting on the market for an extended time or buyer offers that are well below the listing price. Other times, the price is set too low, which results in either a seller "leaving money on the table," or a bidding war which drives the price far above the listing price. Consequently, this causes inefficiencies for both sellers and buyers.

II. Problem Definition

Literature Survey

Being able to reasonably estimate the fair value of a residential property is crucial for the seller to determine the right listing price and for the buyer to identify if the property is a sound investment. Several studies have examined the current practice of real estate valuation. Robey et al. [6] pointed out the three most popular traditional approaches to value real estate, which include the income approach, sales comparison approach and the cost approach. However, according to Ling et. al [4], the income approach is more suitable for commercial real estate while the cost approach is more prone to errors due to the dependency on the appraiser's estimation of the cost of replacing the property net accrued depreciation. The sales comparison approach estimates the property's value through comparing it to a few comparable properties recently sold. However, as discussed by Pagourtzi E et.al.[5], this approach involves several judgments from the appraiser, including identifying comparable sales, adjusting the valuation to account for differences between the subject and the comparable, and assigning weight to the different comparable to estimate the value of the subject. Another major shortage of the sales comparison approach is its dependency on the availability, accuracy, completeness, and timeliness of the transaction data. The common drawback for all three approaches is their reliance on the subjective assumptions of the appraiser.

Utilizing regression analysis is emerging as one of the non-traditional approaches to valuation of real estate. The major benefit of this approach, as discussed by Benjamin et. al [1], is how it can eliminate the human bias factor in the appraiser's selection of comparable and that the direction and analysis are solely statistically determined by the model. However, regression also has its own limitations. Kubus [3] raised awareness on the impact of redundant or irrelevant variables on the model stability and accuracy. Additionally, Isakson [2] pointed out two major limitations of using multiple regression analysis in real estate appraisal, which include model specification and the regression results robustness. To remediate these limitations, he recommends using a large sample size and conducting several statistical tests to determine the reliability of the model.

Proposed Method

With this project, our group proposes a solution wherein we will use current real estate listing data, along with demographic, educational and income data of the different zip codes, to build a regression model which will predict the listing price range of residential properties in New York City. By using a statistical model, we hope to reduce the human bias factor in property valuation, which is a key weakness of the existing approaches, and streamline the process of setting a listing price and resulting in efficient pricing of properties. Through the

model, we also want to understand which factor has the most significant impact on the price of a residential property in New York. To remediate the limitations of regression model discussed in the literature review, we will try several different models and use a sufficiently large data set from Redfin with slightly over 24,000 data points to increase the robustness of the model. A variable selection process using methods like Stepwise Regression, LASSO Regression, Elastic Net Regression, and Ridge Regression, along with analysis on collinearity, will be adopted to ensure only relevant variables are considered in the final model and enhance model specification.

The big differentiator of our method compared to most of the current real estate value estimation regression models is the consideration of additional demographic variables, namely, education level, race, gender, and age distribution; most other studies consider property features and only a few demographic variables such as income. The impact from those additional demographic variables on estimated real estate values are not only interesting due to the potential sociological insights they may provide; they also ensure we have a more complete list of factors to consider within our modeling process. Lastly, what really differentiates our product is an interactive choropleth map to help sellers and buyers visualize the overall market and how their property is valued compared to other properties on sale. This will provide tremendous value to both sellers and buyers in setting and identifying appropriate values of properties, relative to their peers.

III. Data Acquisition and Data Cleaning

There are two primary data sources that we use for this project: (1) Residential property listing inventory (i.e., condo, co-op, multifamily, single family, townhouse) in New York city as of 9/29/2022 and (2) the demographic, income, and education data for different zip codes in New York state.

The housing data was downloaded from Redfin.com and consisted of ~24,000 data points. We then replaced data points having N/A values in (1) lot size for condo and co-op and (2) HOA fee for single-family and multi-family with a 0, as lot size concept is not applicable to condo and co-op and single-family and multi-family properties do not typically have HOA fees. We then removed data points with N/A values and those that were empty. Additionally, we discarded variables that were not of interest (e.g., address, open house date/time, etc.).

We downloaded ACS data on education, demographic, and income for different New York zip codes directly from the Census Bureau website. We used the 5-year estimate data due to its increased statistical reliability for less populated areas and the availability of data by zip code, which is not available for the 1-year data. We removed any data point that did not have ACS data for its zip code.

We then combined the above datasets into one by zip code (using VLOOKUP in Excel). We also changed variable names for clarity, resulting in improved interpretability and easier coding. As of the time of this progress report, we have finished the data cleaning process.

IV. Exploratory Data Analysis

Once we had clean data, we sought to better understand the distribution of data within each attribute as well as any relationships between attributes. For reference, below is a list of the attributes with descriptions:

1. **Price** – the property listing price (in USD)
2. **PropertyType**– a factor (Condo, Co-op, Multifamily, Single Family, Townhouse)

3. **Borough** – a factor (Bronx, Brooklyn, Manhattan, Queens, Staten Island)
4. **NumBeds** – number of bedrooms
5. **NumBaths** – number of bathrooms
6. **sqft** – interior square footage
7. **LotSize** – size of lot (in sqft)
8. **Age** – how old the property is
9. **DaysOnMarket** – how many days listed
10. **HOAfee** – monthly homeowners association fees
11. **ZipMedianIncome** – median income for zip code
12. **Percent_pop.25YRwBdegree** – percent of population >= 25yrs old with a bachelor's degree
13. **Percent_pop.65yr** – percent of population >=65yrs old
14. **Percent_White** – percent of population that is white
15. **Percent_Asian** – percent of population that is Asian
16. **Percent_BlackAA** – percent of population that is African American
17. **Percent_Native** – percent of population that is Native American
18. **SexRatio** - # of males / # of females

Histograms & Boxplots

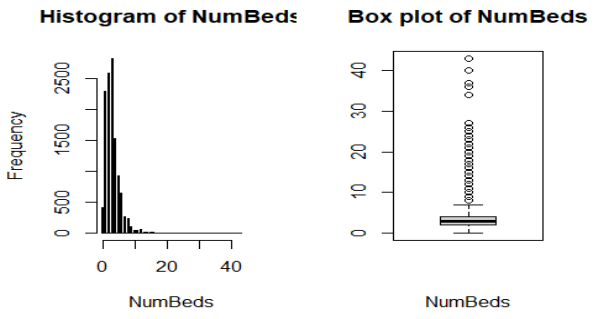


Figure 1. Histogram and Boxplot - NumBeds

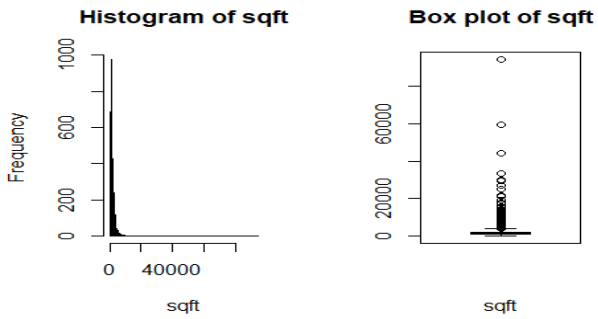


Figure 2. Histogram and Boxplot - sqft

By using a histogram and boxplot, we can see the distribution of values for each variable. This helps us to better understand how data may be skewed and alert us the presence of potential outliers. Above is a sample of the boxplot and histogram for the Number of Beds and sqft variables (we created these for all variables).

Correlation Matrix

Figure 3 shows the correlation between different attributes, with blue indicating a positive correlation, red indicating a negative correlation, and the fullness of the pie and darkness of color representing the strength (a full dark circle represents a strong correlation This matrix provides insights into the relationships between our predicting variables and the response variable (price) by looking down the first column on the left. It can also help us understand where we may expect to find issues with collinearity. As we can see, certain pairs of

variables exhibit strong correlation; for example, NumBaths and NumBeds exhibit strong positive correlation, while Percent_BlackAA and Percent_White show a strong negative correlation.

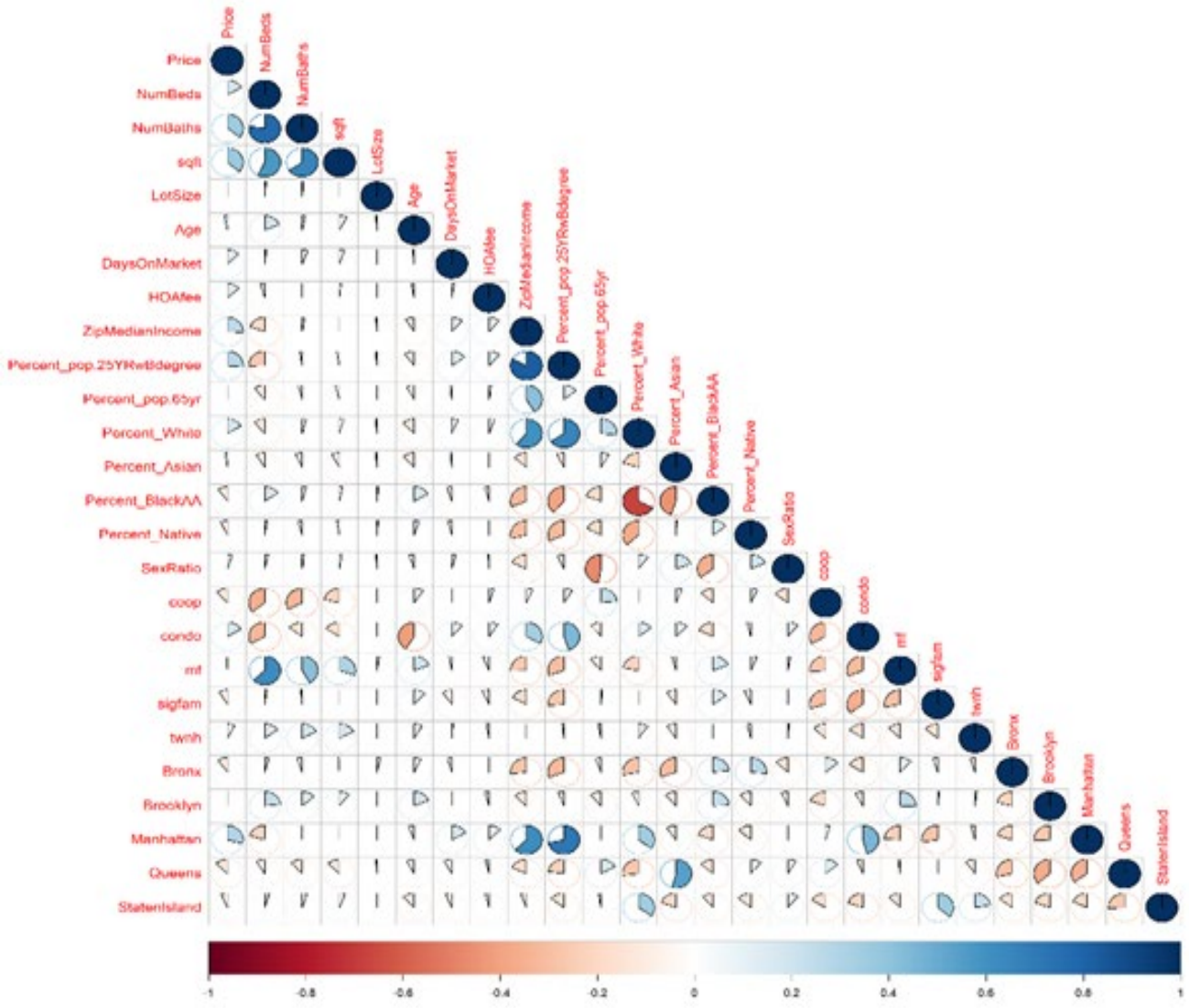


Figure 3. Correlation Matrix

V. Model Building

A linear regression model was fitted on the entire dataset, which used all the available attributes as predictors and Price as the response variable. The model was checked against the linearity assumptions to determine if a linear regression model is potential.

<u>Checking Assumptions</u>

The diagnostic plots ion Figure 4 suggests that while there is a linear relationship between the predictor variables and the response variable and that the residuals are sufficiently normally distributed, heteroskedasticity and outliers exist.

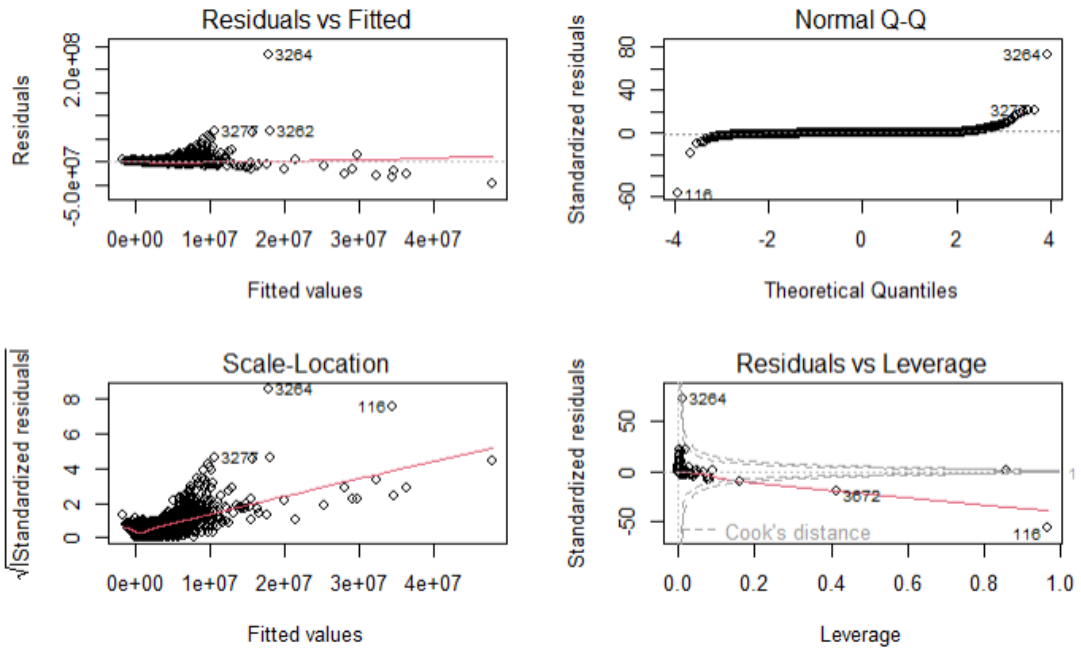
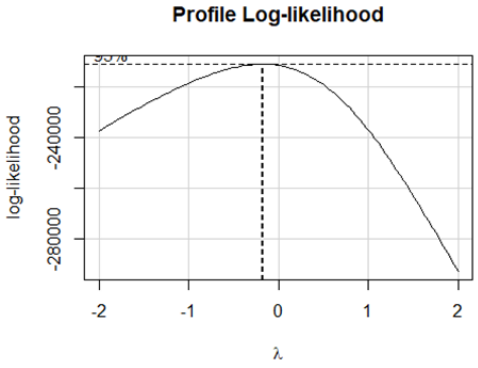


Figure 4. Diagnostic Plot for Linearity Assumptions (pre-transformation)

Transformation of Response Variable (if needed)



We then used the Box-Cox function in R to identify the optimal transformation on the dependent variable. Our Box-Cox plot shows an optimal lambda value of approximately zero, which suggests that we should log transform our response variable (Price). Additionally, this transformation will ensure the model intercept will not be negative, which was an issue prior to the transformation.

We then fit a new model which used the transformed dependent variable. The transformation seems to have helped heteroskedasticity issue significantly as indicated from the below Residuals vs Fitted plot. The slope of Scale-Location has also appeared less steep compared to before the transformation. The normality fit seems to have improved, with a smaller upper tail as indicated by Normal Q-Q. Outliers/Leverage points still appear, and we will address those next.

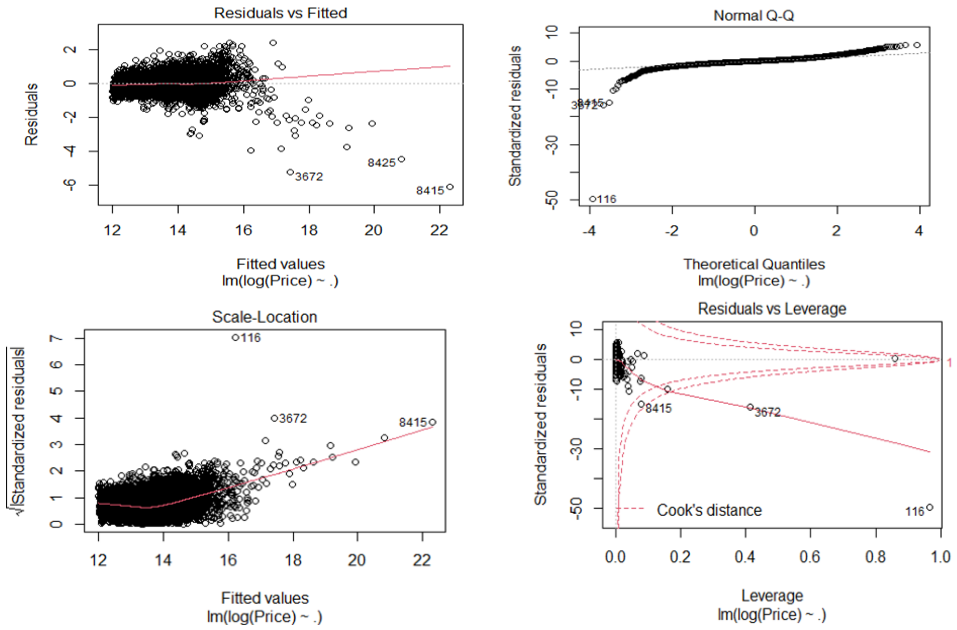
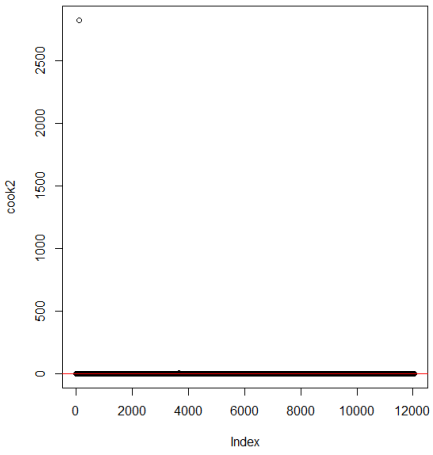


Figure 5. Diagnostic Plot for Linearity Assumption (post-transformation)

Removal of High-Leverage Points

Plotting the Cook's distance, we see that a few points have values greater than 1 (data points 116 and 3672), which we then removed from the model. Also note this confirms what we see in the Residuals vs Leverage plot above.



Addressing Collinearity

	GVIF	Df	GVIF^(1/(2*Df))
PropertyType	6.673394	4	1.267778
Borough	12.798412	4	1.375291
NumBeds	4.195650	1	2.048329
NumBaths	3.731776	1	1.931780
sqft	2.446215	1	1.564038
LotSize	1.002517	1	1.001258
Age	1.427344	1	1.194715
DaysOnMarket	1.047124	1	1.023291
HOAfee	1.762649	1	1.327648
ZipMedianIncome	5.235038	1	2.288020
Percent_pop.25YRwBdegree	8.707438	1	2.950837
Percent_pop.65yr	2.735124	1	1.653821
Percent_White	16.548211	1	4.067949
Percent_Asian	6.234002	1	2.496798
Percent_BlackAA	11.285336	1	3.359365
Percent_Native	1.394267	1	1.180791
SexRatio	2.049248	1	1.431519

We have also calculated the VIF for all predictors, the result of which showed that the variables "Borough," "Percent_White," and "Percent_BlackAA" exhibit collinearity with other variables as their VIF is higher than the threshold we set, which is 10. As a result, we removed these variables from the model.

We would like to note here that after addressing outliers and reducing collinearity, the models adjusted R-squared value increased from 75.33% to 77.36%.

Variable Selection

We used several variable selection methods, including Stepwise Regression, LASSO Regression, Elastic Net Regression, and Ridge Regression. Each variable selection methods resulted in new models with differing predictors (and predictor values). The results are summarized below:

Note: Data was first scaled so we could use LASSO, Elastic Net and Ridge regression techniques.

1. **Stepwise Regression**: The final output omits variable "LotSize."
2. **LASSO Regression**: We trained a 10-fold cross validation Lasso model to find the optimal lambda value and then fit the model. The result indicated that all variables will be kept with a different set of coefficients.
3. **Elastic Net Regression**: We then constructed an Elastic Net Regression model using similar steps as LASSO with an alpha set at 0.5. The output indicated that all variables will be kept with a different set of coefficients.
4. **Ridge Regression**: The result also indicated that we will keep all variables with a different set of coefficients. This is characteristic of Ridge regression, which doesn't yield a sparse model since it can shrink predictors close to 0, but not to 0.

Model Selection and Evaluation

The models resulted from the variable selection process above are further optimized through Monte Carlo cross-validation to identify the model that has the highest predictive power. Several metrics have been used, including adjusted R-squared, Mean Squared Error (MSE), and AIC/BIC. The result is summarized in the table below. The Stepwise model was selected as it has the lowest MSE and relatively high Adjusted R-squared.

	Linear Regression	**Stepwise**	**Lasso**	**Elastic net**	**Ridge**
MSE	0.1729	0.1728	0.1757	0.1742	0.1852
Adjusted R-squared	0.7745	0.7745	0.7714	0.7728	0.7583
AIC	10133.5529	10131.7236	-5480.4654	-5489.0363	-5385.2320
BIC	10269.8429	10260.7687	-5364.4620	-5370.0945	-5263.3519

Figure 6 below shows a summary of the selected model, where we can see that all selected variables are statistically significant at alpha level of 0.001. As the response variable was log-transformed, the estimated coefficients of the predictors were exponential-transformed for ease of interpretation (see Figure 7). We can see that most of the variables have a positive linear relationship with the price, except for "Percent_pop.65yr" and "Percent_Native". Compared to Co-op, all other property types have a higher price, leading by Townhouse, which is an expected result. Of all features of the property and demographic characteristics, the number of baths seems to have the highest impact on the price.

```
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               1.113e+01  1.060e-01 105.033  < 2e-16
## PropertyTypeCondo                         8.177e-01  1.368e-02  59.757  < 2e-16
## PropertyTypeMulti-Family                  1.040e+00  1.933e-02  53.800  < 2e-16
## PropertyTypeSingle Family Residential     9.468e-01  1.505e-02  62.933  < 2e-16
## PropertyTypeTownhouse                     1.114e+00  2.265e-02  49.201  < 2e-16
## NumBeds                                   5.260e-02  3.596e-03  14.629  < 2e-16
## NumBaths                                  9.628e-02  4.975e-03  19.354  < 2e-16
## sqft                                      6.485e-05  3.940e-06  16.457  < 2e-16
## Age                                       7.113e-04  1.237e-04   5.749 9.27e-09
## DaysOnMarket                              1.047e-04  2.155e-05   4.857 1.21e-06
## HOAfee                                    1.734e-04  3.491e-06  49.654  < 2e-16
## ZipMedianIncome                           1.404e-06  1.492e-07   9.413  < 2e-16
## Percent_pop.25YRwBdegree                  1.459e-02  4.345e-04  33.583  < 2e-16
## Percent_pop.65yr                         -7.602e-03  7.706e-04  -9.865  < 2e-16
## Percent_Asian                             6.452e-03  2.885e-04  22.363  < 2e-16
## Percent_Native                           -4.299e-02  1.370e-02  -3.137 0.001711
## SexRatio                                  8.165e-03  2.107e-03   3.875 0.000107
##
## (Intercept)                              ***
## PropertyTypeCondo                        ***
## PropertyTypeMulti-Family                 ***
## PropertyTypeSingle Family Residential    ***
## PropertyTypeTownhouse                    ***
## NumBeds                                  ***
## NumBaths                                 ***
## sqft                                     ***
## Age                                      ***
## DaysOnMarket                             ***
## HOAfee                                   ***
## ZipMedianIncome                          ***
## Percent_pop.25YRwBdegree                 ***
## Percent_pop.65yr                         ***
## Percent_Asian                            ***
## Percent_Native                           **
## SexRatio                                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4106 on 9618 degrees of freedom
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.7741
## F-statistic:  2064 on 16 and 9618 DF,  p-value: < 2.2e-16
```

Figure 6. Final Model Summary

```
round((exp(coef(finalmod))-1)*100, digits=4)

##                                (Intercept)                                PropertyTypeCondo
##                                  6837687.1591                                       126.5353
##             PropertyTypeMulti-Family PropertyTypeSingle Family Residential
##                                     182.9524                                       157.7562
##                PropertyTypeTownhouse                                    NumBeds
##                                     204.7682                                         5.4010
##                             NumBaths                                       sqft
##                                      10.1072                                       0.0065
##                                  Age                                DaysOnMarket
##                                       0.0712                                       0.0105
##                                HOAfee                             ZipMedianIncome
##                                       0.0173                                       0.0001
##                Percent_pop.25YRwBdegree                           Percent_pop.65yr
##                                       1.4700                                      -0.7573
##                           Percent_Asian                              Percent_Native
##                                       0.6472                                      -4.2083
##                                SexRatio
##                                       0.8198
```

Figure 7. Exponential Transformation of Coefficients

The selected model has been fitted on the entire data set to predict the fair listing price for the residential properties. The predicted values are then compared with an interval of +/- 15% of the existing listing price on Redfin to identify properties that are over-, fairly-, or undervalued. An interval was chosen instead of the listing price to adjust for the intentional over- or underpriced strategy when properties listed (e.g., sellers intentionally list a property at a very low price to attract many buyers). The analysis shows that ~32% of the properties in NYC are overly priced (Above Interval) and ~30% are underpriced (Below Interval). Of which, Condos and Co-ops have a higher percentage of properties being overpriced compared to other types.

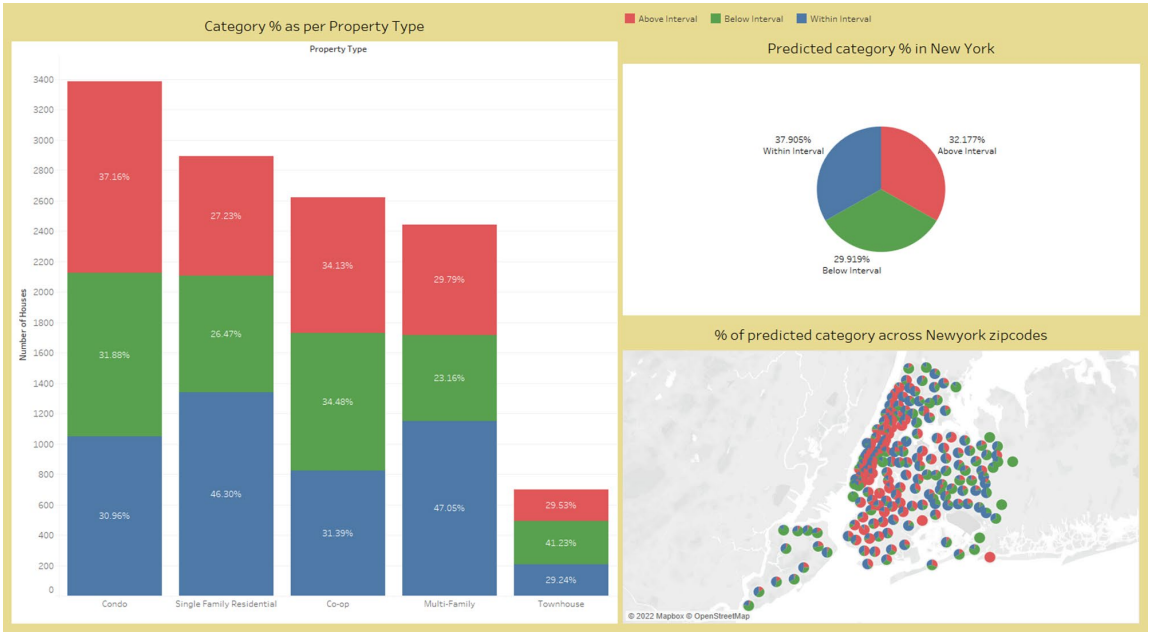


Figure 8. Summary of model output by property type

The result of the analysis has also been visualized in an interactive choropleth map (see Figure 9 – the map is also published in this link: [Team57's Project | Tableau Public]), where the green dots and red dots represent properties that are underpriced and overpriced compared to the market, respectively. The map shows that majority of the residential properties in Manhattan and Brooklyn areas are overpriced, as opposed to in Bronx or Staten Island. A more balanced state is observed in Queens where properties in areas closed to Manhattan tend to see more overpriced properties. This is in line with our expectation and understanding of NYC as properties in Manhattan and Brooklyn are usually listed at a premium due to location advantages.

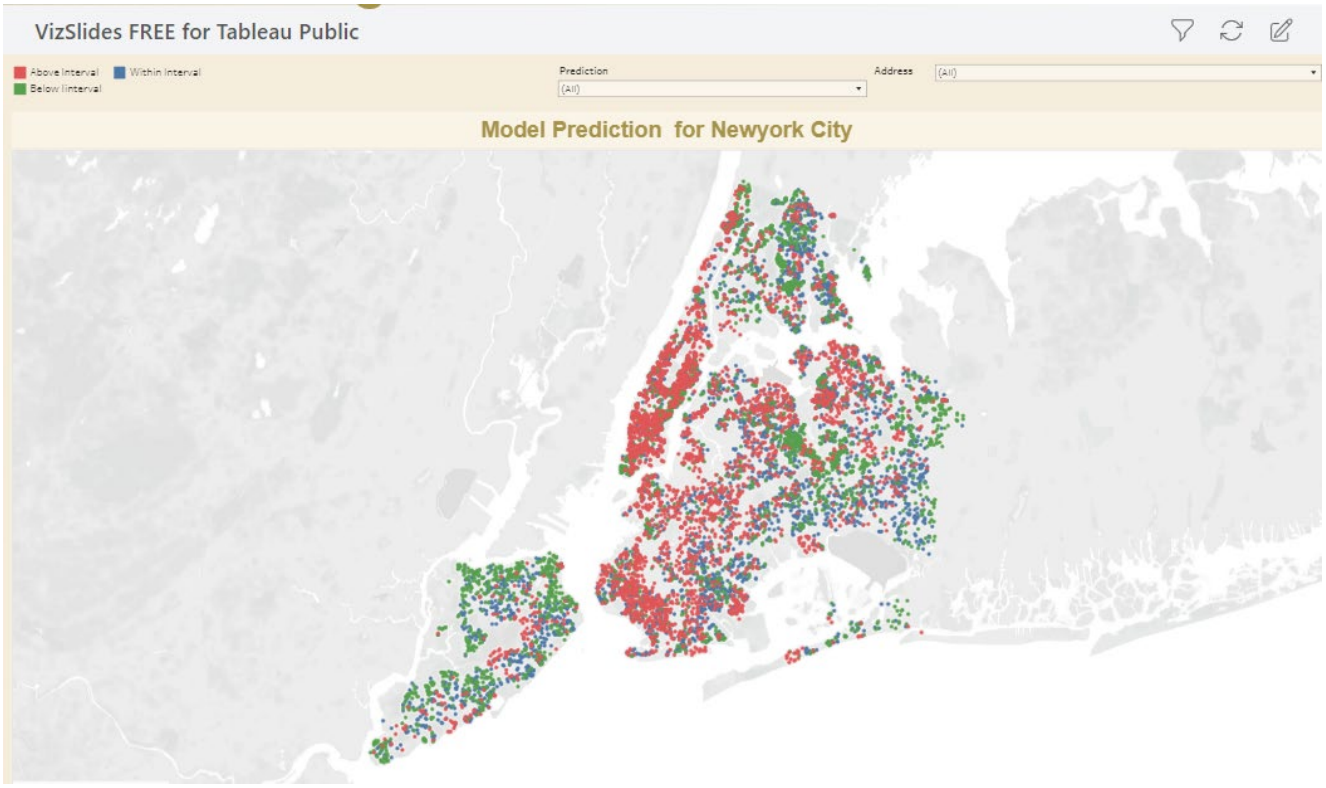


Figure 9. Overview of model output on choropleth map

Additionally, the map also allows users to filter for a certain address and see how the property is priced based on the color of the dot that represents the property. They are also able to filter for only inventories that are either over-priced, within interval or underpriced.

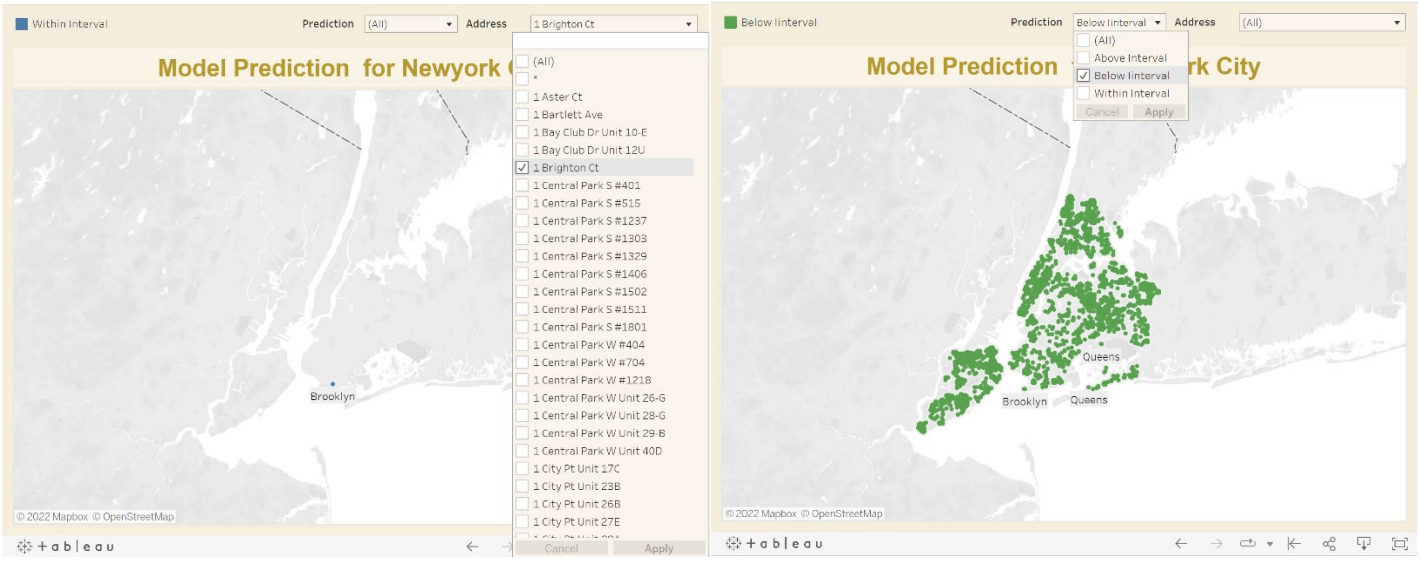


Figure 10. Overview of the filtering feature on choropleth map

VI. Conclusion and Discussion

Our project was motivated by the need of a pricing method for residential properties that can minimize the limitations of the existing approaches, especially in reducing the human bias factor. This has been achieved through having a regression model using real-life real estate and demographic data from Redfin and US Census Bureau. In addition to the model, our final deliverables also include an interactive choropleth map that helps users visualize the overall market at a point in time. For the buyers, the map can help them quickly identify a potential buy in the area of interest. The model can then be used to estimate the fair listing price of that property, thus facilitating the negotiation process. For the sellers, they can use the model to derive a base estimated price for their property, while utilizing the map to understand how the surrounding properties are priced and subsequently adjust the estimated price based on their need (e.g., decrease the price to attract more buyers). Additionally, our model also indicates that the number of baths, out of all property features, has the most significant impact on the price of the property. This insight is helpful, especially for the sellers as if they want to significantly increase the value of a property, they might want to consider creating another bathroom or half-bathroom.

As the next step for the project, we will work on the automatic feed of the input data into our model through using API. While regression model has several advantages, including reduction in bias, we recognize the need for a rigorous monitoring plan to identify any model drift, especially in situation of new dataset or change in economic environment which impacts how the real estate market behaves. Where a significant model drift is identified, model recalibration should be considered. We will continue monitoring the model and its output for the NYC market in the near future and hope to extend our project to other cities and areas of the US.

REFERENCES

1. Benjamin, J., Guttery, R., & Sirmans, C. (2004). Mass appraisal: An introduction to multiple regression analysis for real estate valuation. *Journal of Real Estate Practice and Education*, *7*(1), 65–77. https://doi.org/10.1080/10835547.2004.12091602

2. Isakson, H. R. (2001). Using Multiple Regression Analysis in Real Estate Appraisal. *Appraisal Journal*.

3. Kubus, M. (2016). Assessment of predictor importance with the example of the real estate market. *Folia Oeconomica Stetinensia*, *16*(2), 29–39. https://doi.org/10.1515/foli-2016-0023

4. Ling, D. C., & Archer, W. R. (2021). *Real estate principles a value approach*. McGraw-Hill.

5. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance*, *21*(4), 383–401. https://doi.org/10.1108/14635780310483656

6. Robey, S. L., McKnight, M. A., Price, M. R., & Coleman, R. N. (2019). Considerations for a regression-based real estate valuation and appraisal model: A pilot study. *Accounting and Finance Research*, *8*(2), 99. https://doi.org/10.5430/afr.v8n2p99