Predicting the Sale – Price of a Building Unit Sold in the New York City Property Market Using Various Building Characteristics

I. Background

Property in New York City (NYC) is extremely expensive, and one might naively conclude that the expensive price tags would hinder the property market. Nevertheless, the Big Apple real-estate market continues to entice investors, and tens of thousands of properties are bought and sold each year. As more investors look to buy a piece of NYC, the prices will keep rising. According to a report from CNBC, three NYC neighborhoods were listed among the top 15 'Most Expensive Places to Buy a Home in the US.' Having said this, we can claim that the NYC property market is a saturated market. If you are a follower of the *Efficient Market Hypothesis*, you would believe that there is no way to make an above average return by investing in a saturated market. However, if you are a follower of Howard Mark's philosophy on investing, you believe that no market is perfectly efficient; and that it can be possible to make above average returns in a saturated market—through means of some insight that isn't available to others, such as being able to predict the price of an asset. It is true that the *Efficient Market Hypothesis* and Howard Mark's philosophy are geared towards the stock market, but I believe that the principles of investing in any market are ubiquitous. Being able to accurately predict a price in a market—whether it is the stock market, the commodities market, or the *real-estate market*—can result in surreal profits. In this study, we examine an annual dataset that lists all NYC property sales; this particular dataset is for the 2016-2017 fiscal year, and it can be found here. The scope of this project was to use the dataset to identify market trends, and ultimately build a powerful model through statistical methods that can predict the sales-price of a property—given the values of the predictor variables.

II. Approach

The dataset contains variables that describe the type of the building, as well as its location. The information provided by these variables can be utilized to model and predict the sales price of building(s) in the NYC real-estate market.

III. Methods/Results

a. Exploratory Data Analysis (EDA) and Variable Selection

The NYC Property Sales dataset is a rolling (annual) list of all property sales in NYC. It is generally published by the NYC government website, but the dataset for this study was acquired through Kaggle. The dataset lists all property sales in NYC from August 1st, 2016 to August 31st, 2017. The dataset contains various fields such as borough, neighborhood, block, and lot, which describe the building(s) location. Other dataset fields such as tax-class and building-class-category describe the building(s). Initially the dataset contained 22 columns—or 21 features and the *Sale Price* column. Four of these columns—*Unnamed, Apartment #, Address,* and *Ease-ment*—were discarded right away because they were either an index column, an empty column, or filled with NaN values. 18 columns that remained are listed in **Table 1**.

Column/Variable	**Description**	**Type**
Sale_Price	Value in USD at which the property was sold.	Numerical
Sale_Date	Date on which the property was sold.	Date
Land_square_feet	Inside area of the building	Numerical
Gross_square_feet	Outside area of the building	Numerical
Residential_Units	Number of residential units in the sale	Numerical
Commercial_Units	Number of commercial units in the sale	Numerical
Total_Units	Residential_Units + Commercial_Units	Numerical
Year_Built	Year, the build was built	Year
Borough	The NYC Borough	Categorical*
Block	Specific Block in the Borough	Numerical
Lot	Specific Lot in the Specific Block	Numerical
Zip	Zip Code	Numerical
Neighborhood	Neighborhood of the Building Location	Categorical
Building_Class_Category	Type of Building	Categorical
Building_Class_At_Time_of_Sale	Type of Building at Sale	Categorical
Building_Class_At_Present	Type of Building Currently	Categorical
Tax_Class_At_Time_of_Sale	Tax Class of Building at Sale	Categorical*
Tax_Class_At_Present	Tax Class of Building Currently	Categorical
***Categorical: These values were given as Integers but are actually categorical variables.**		

Table 1

Though it is possible to build a model with all 18 variables, these variables were first assessed to understand their possible significance—or the lack thereof—in predicting the property sales price.

The initial dataset consisted of 84,548 observations. Out of these, 956 rows were duplicates of other observations and thus, were subsequently removed from the dataset—bringing the total down to 83,592 observations. Many of these observations had ***NaN*** values for some of the numerical variables. One method to deal with ***NaNs*** is to replace them with the mean. However, there is a disadvantage to this method. Replacing one, two, or even ten(s) of rows with the mean might not impact the ultimate model. This dataset had thousands of missing values and replacing them with the mean would certainly have impacted the final results. Fortunately, with a large dataset we can afford to drop the rows with ***NaN*** values. Rows that had ***NaN*** values for the numerical variables listed above were dropped—truncating the dataset to a total of 47,844 observations.

Before any further cleaning and processing of the data, the numerical variables were used to create a correlation matrix. (**Figure 1**)

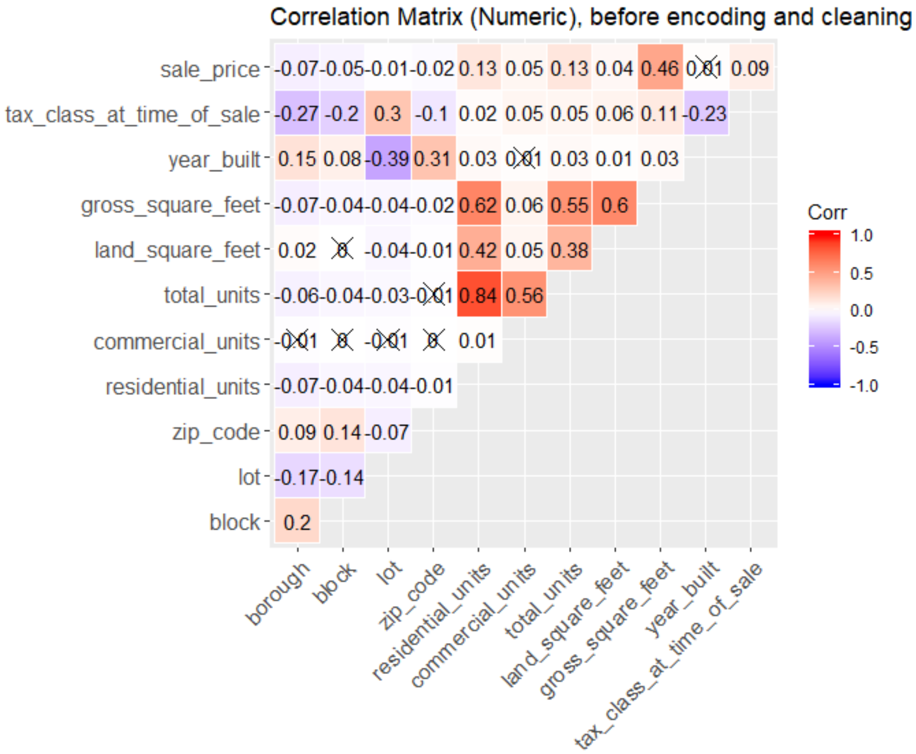


Figure 1

The correlation matrix shows that no particular variable—apart from *gross square feet*—has a notable correlation with the property sales price. It is possible that some of these variables might be significant predictors when included in the ultimate model. Nevertheless, it was perceived that it would be futile to carry all the variables along to the final model, because some variables provide information that is covered by another predictor(s). The predictors that were dropped—and the reasoning—is provided in **Table 2**.

Following the removal of the listed variables, the dataset was left with six predictor variables and the sales price column. The correlation matrix of these variables is shown in **Figure 2**. Despite the low correlations, I was confident in these predictors to build an effective sales price predicting model.

The next step was to look at the distributions of the various numerical variables.

Variable	Reason
Zip-Code	The location is aptly covered by ***Borough***
Block	The location is aptly covered by ***Borough***
Lot	The location is aptly covered by ***Borough***
Neighborhood	The location is aptly covered by ***Borough***
Building_Class_At_Time_Of_Sale	Aptly covered by Building_Class_Category
Building_Class_At_Present	Present class does not impact sales price
Tax_Class_At_Present	Present class does not impact sales price
Residential_Units	Values were mostly 0s or 1s
Commercial_Units	Values were mostly 0s or 1s
Total_Units	No correlation to sales price, but collinear to square footage
Sales_Date	Property buying isn’t seasonal, and times-series analysis was outside the scope of this study

Table 2

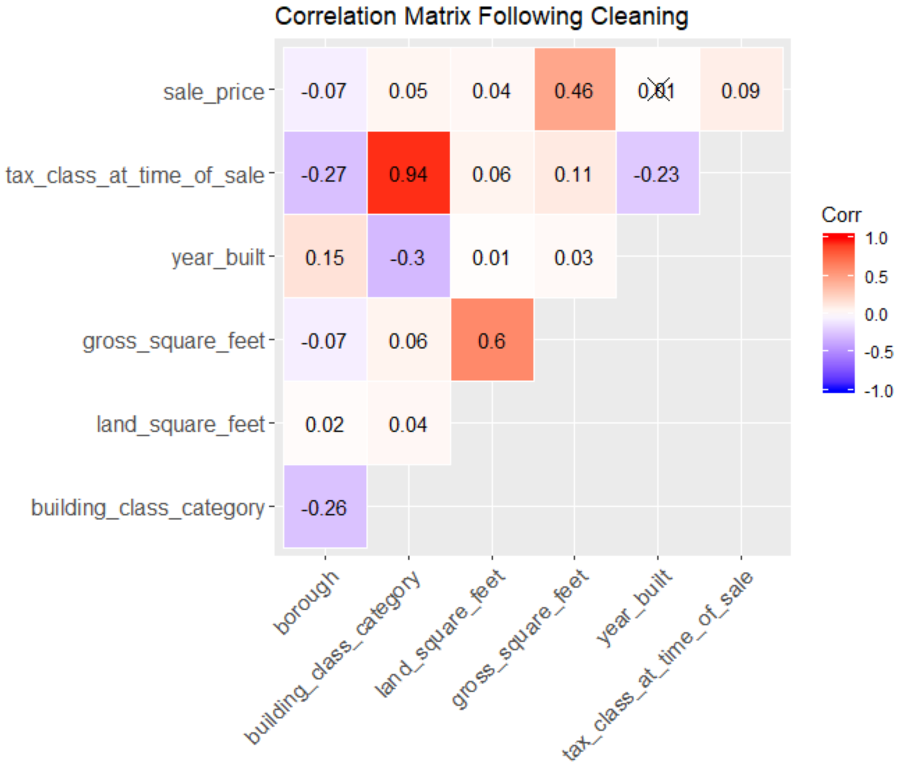


Figure 2

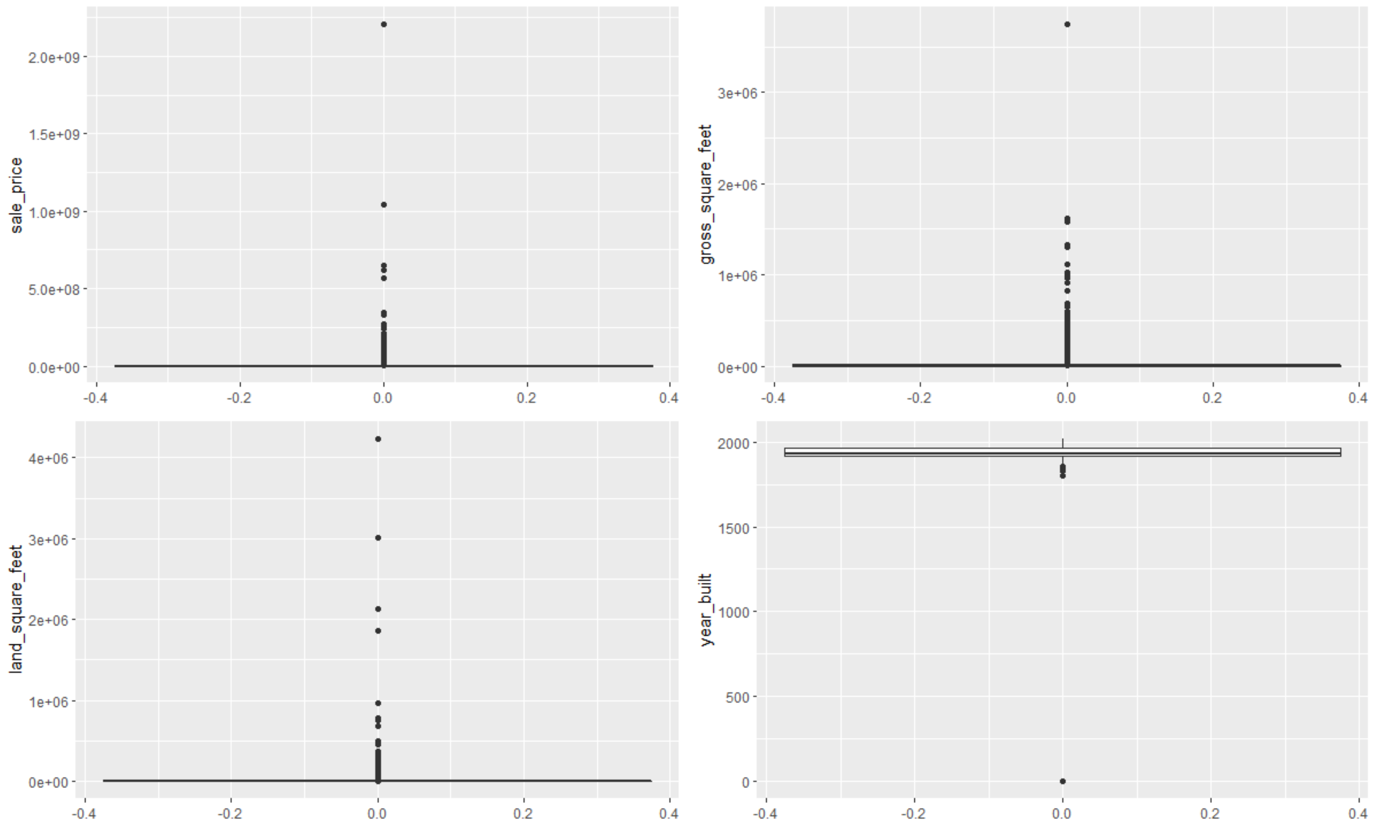


Figure 3

The distributions of the numerical variables are shown in **Figure 3**. It's evident that the distributions are heavily skewed and contain outliers. Looking at the raw values, it was discovered that the sales price column had many rows with 0s or other values which did not reflect a true sale. A $0 sales prices was for property transfers—such as from parents to children—rather than a property sale. The gross and land square feet columns also had many rows with 0s, which also do not represent a true sale. There was also an observation with a value of 0 for the year built variable. It's likely that these rows simply lacked information for these features but including these values would still weaken our model. In an effort to limit the skewness of our dataset and make it more representative of true property sales, a subset of the data was selected using the following conditions: sales price greater than $100,000, gross square feet and land square feet greater than 0 ft^2, and year built greater than 1880. This subset contained a total of 28,246 observations.

Despite selecting a subset and addressing the outliers, the distributions of our numerical variables were still skewed as seen in **Figure 4**. Since, the year-built column only had one outlier—removing it resolved the skewed distribution. The other variables, including the sales price, were still skewed. To resolve this is issue, a log-transformation, along with standard scaling, was applied to the numerical variables. The resulting distributions of the numerical variables are shown in **Figure 5**.

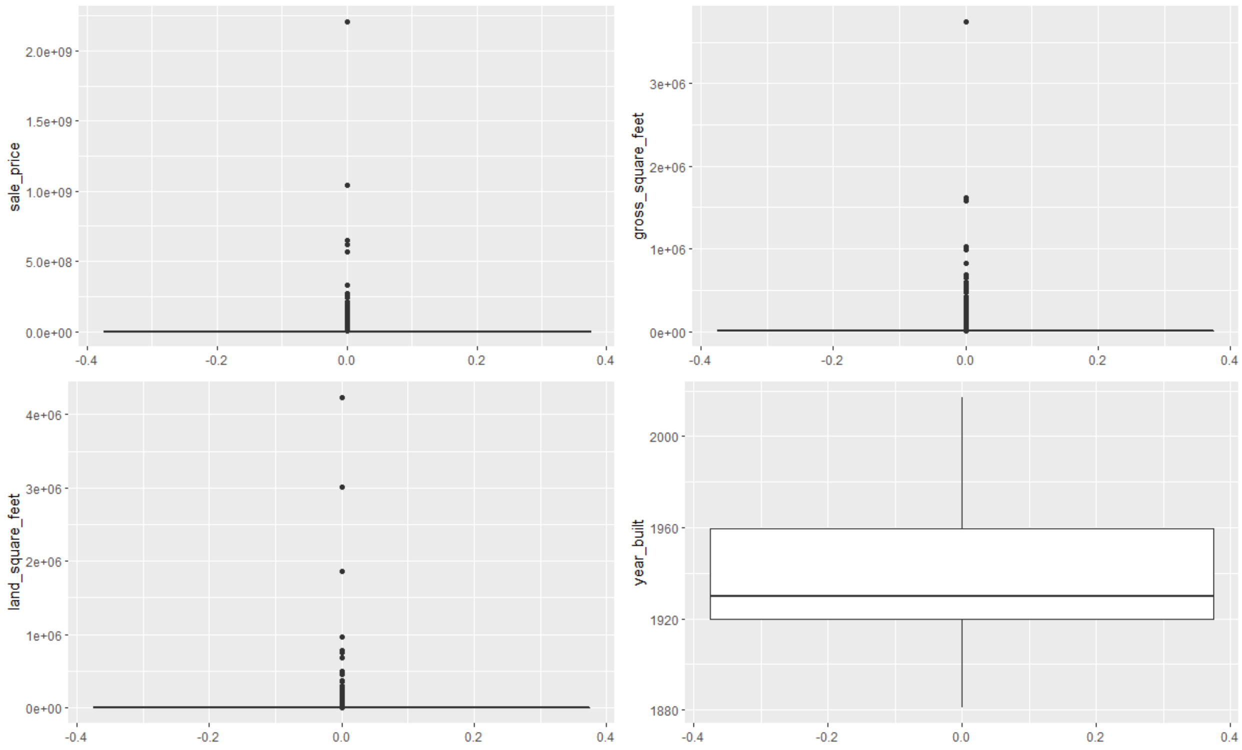


Figure 4

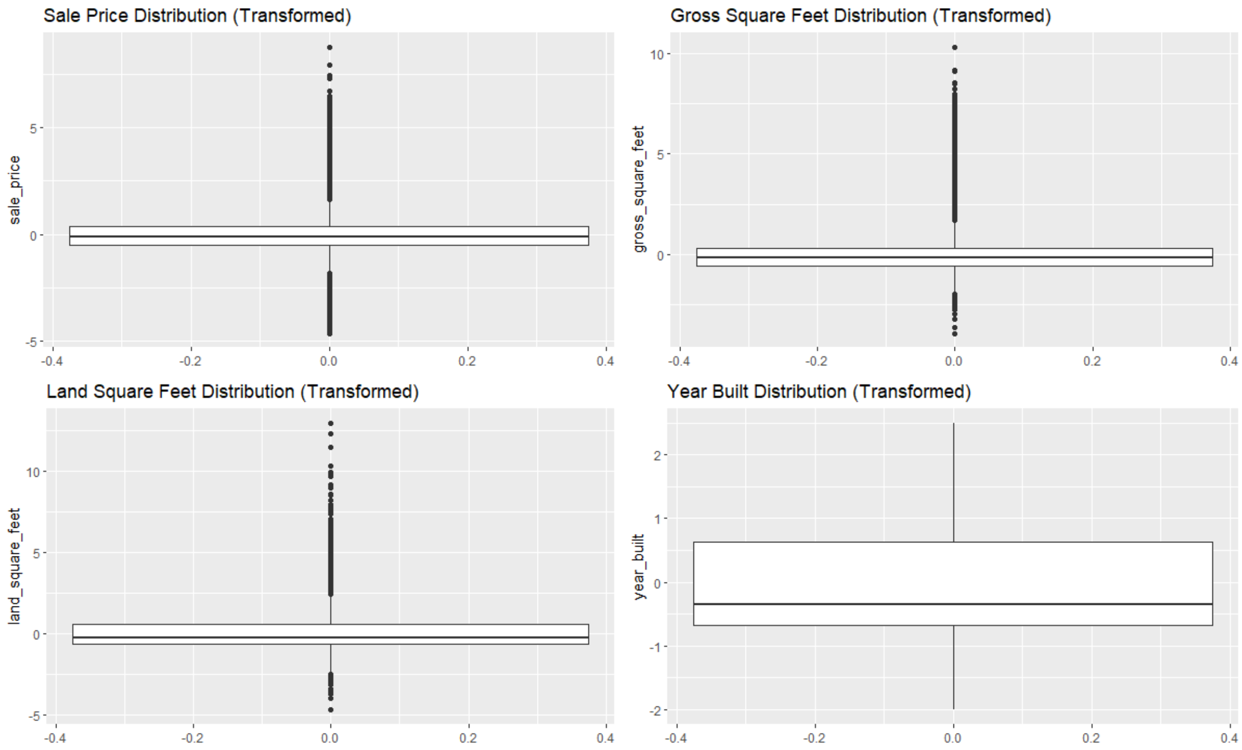


Figure 5

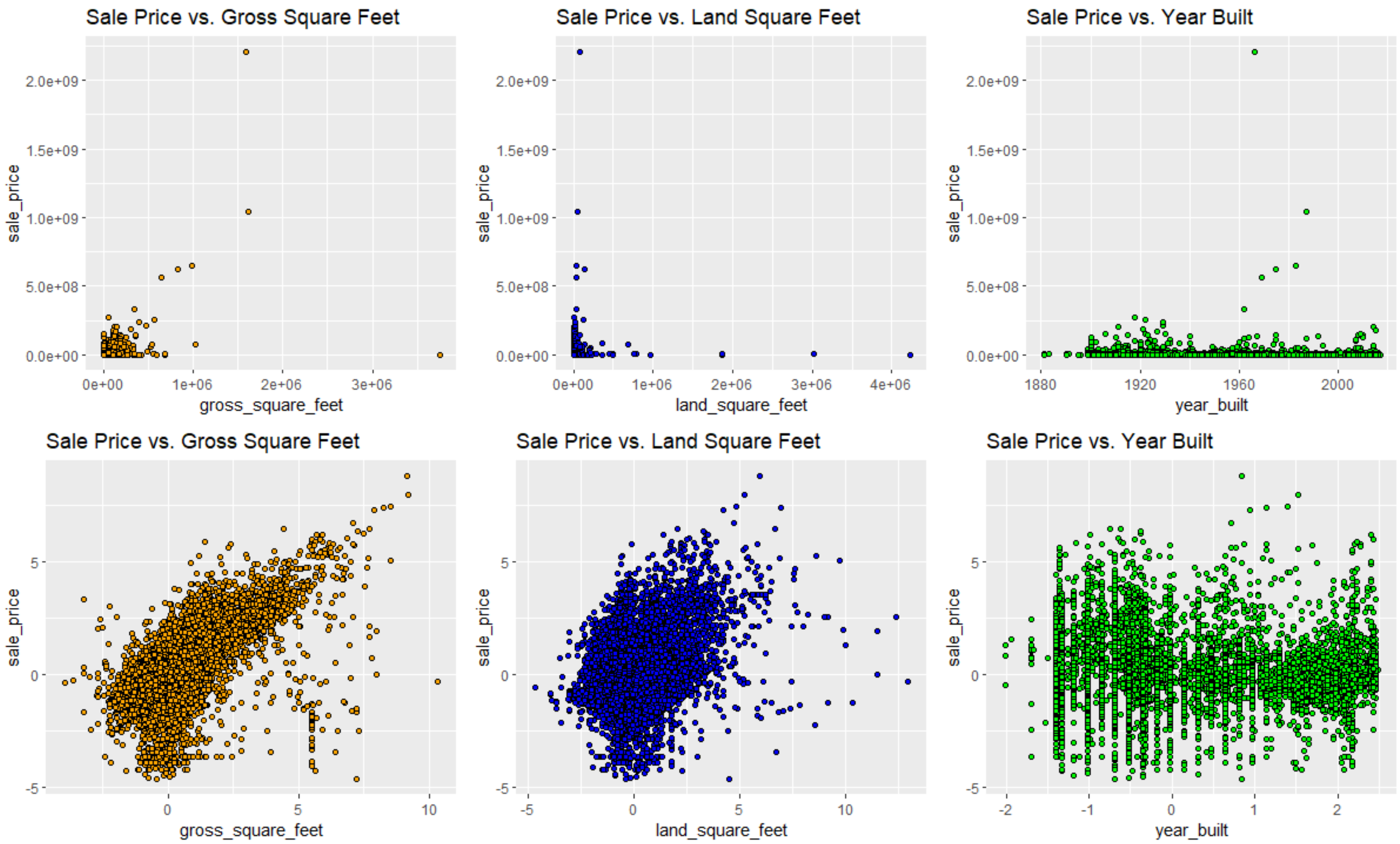


Figure 6

The effect of the transformation is also shown in **Figure 6**, where the top row shows the relationship between the sales price against the three numerical variables prior to the transformation, and the bottom row shows the same relationship following transformation. Constructing a correlation matrix with the transformed variables is likely to show better results than in **Figure 1** and **Figure 2.**

Apart from assessing the relationship between sales prices and the numerical variables, it's important to do the same for the categorical variables. **Figure 7** was used to study the distributions of the sales price for the different categories of the borough and tax-class variables. These box plots aptly show that price distributions vary in regard to borough and tax-class. Borough 1 tends to have a higher sales price than the other four boroughs—this is likely because borough 1 refers to Manhattan. Similarly, it's also evident that tax-class 1 generally has a lower sales price than tax-classes 2 and 4. **Figures 6 and 7** provide confidence in our decision to utilize these variables as we approach the final stage of modeling.

The categorical variables were label encoded, meaning each category was assigned a specific number to make the variables machine readable. These variables can now be treated as numerical variables. Unfortunately, label encoding has a major draw-back because it can lead to priority issues, where labels with higher values maybe considered to have a higher priority than labels with lower values—(i.e 0 < 1 < 2).

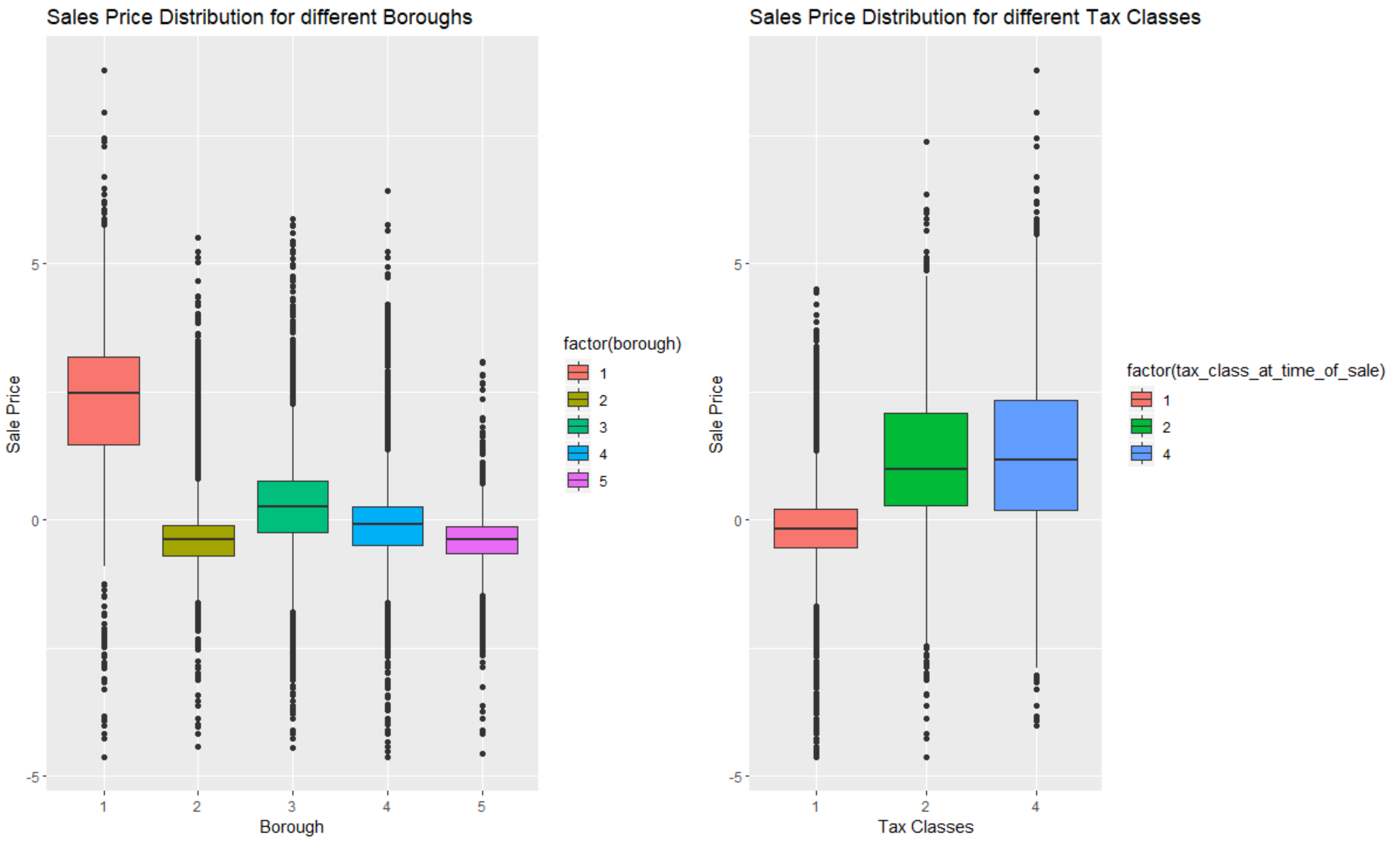


Figure 7

Thus, ***one-hot-encoding*** was used to numerically represent the categorical variables, where each category is converted into a binary column—a 1 means that the row-observation belongs to the category and a 0 means that it does not. The borough, tax-class, and building class category variables were one-hot-encoded. The borough and tax-class variables had five and three categories each respectively, but the building class category variable had 47 different categories. To avoid a sparse matrix of 0s, only the top eight building class categories were selected to be one-hot-encoded.

b. Modeling

Several models were trained and evaluated to determine the method that leads to the most accurate sales price predictions. Since this is a regression problem, the criterion used to evaluate the models was RMSE. The first model was a linear regression model. The optimal linear regression model was selected using stepwise selection via the 'lm()' and 'step()' functions. Both forward and backward stepwise regression methods were used. The criterion used for subset selection was the Bayes' Information Criterion (BIC), because we aimed to penalize complex models. The backward selection method—where predictors from the full model are eliminated at each step to find a reduced model that optimally explains the data—led to the minimum BIC. The BIC at each step of the backward stepwise regression method is shown in **Figure 8**. The method led a model of 11 variables—the parameter estimates of which are shown **Figure 9**.

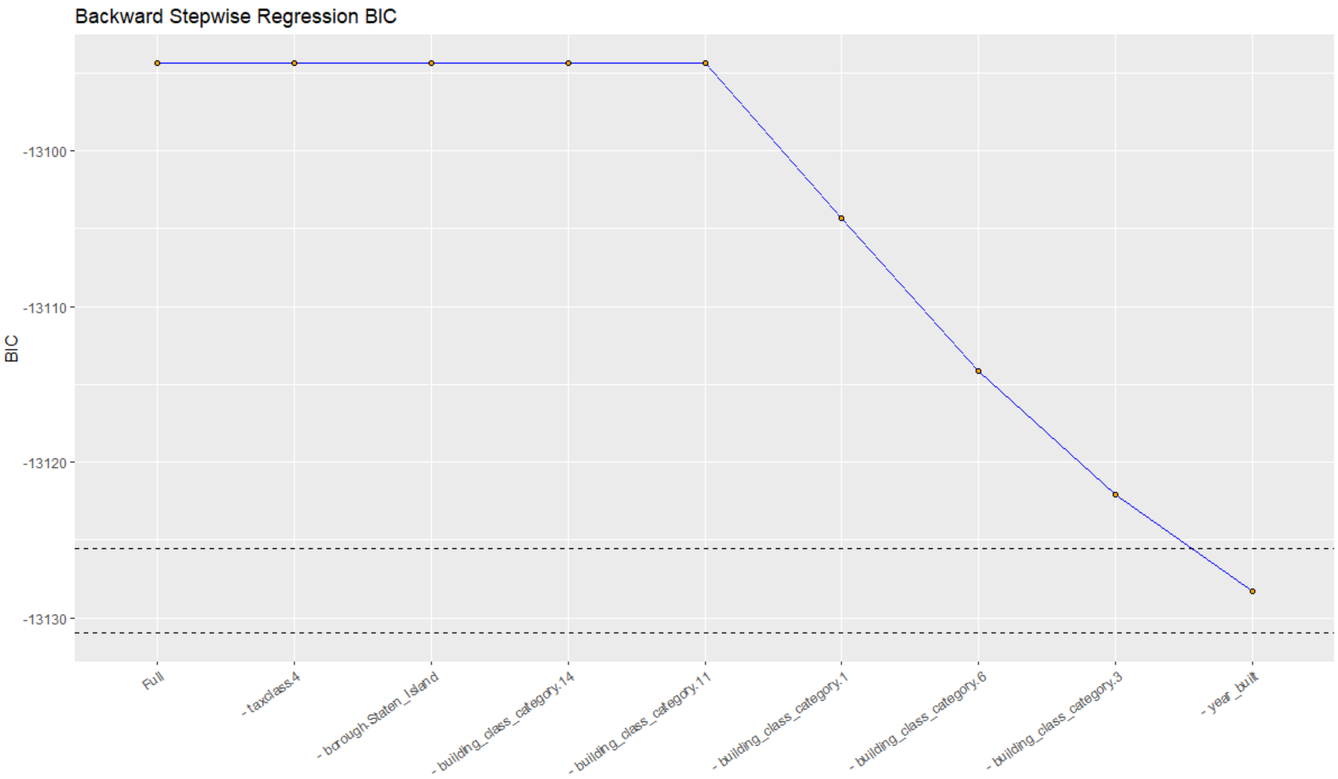
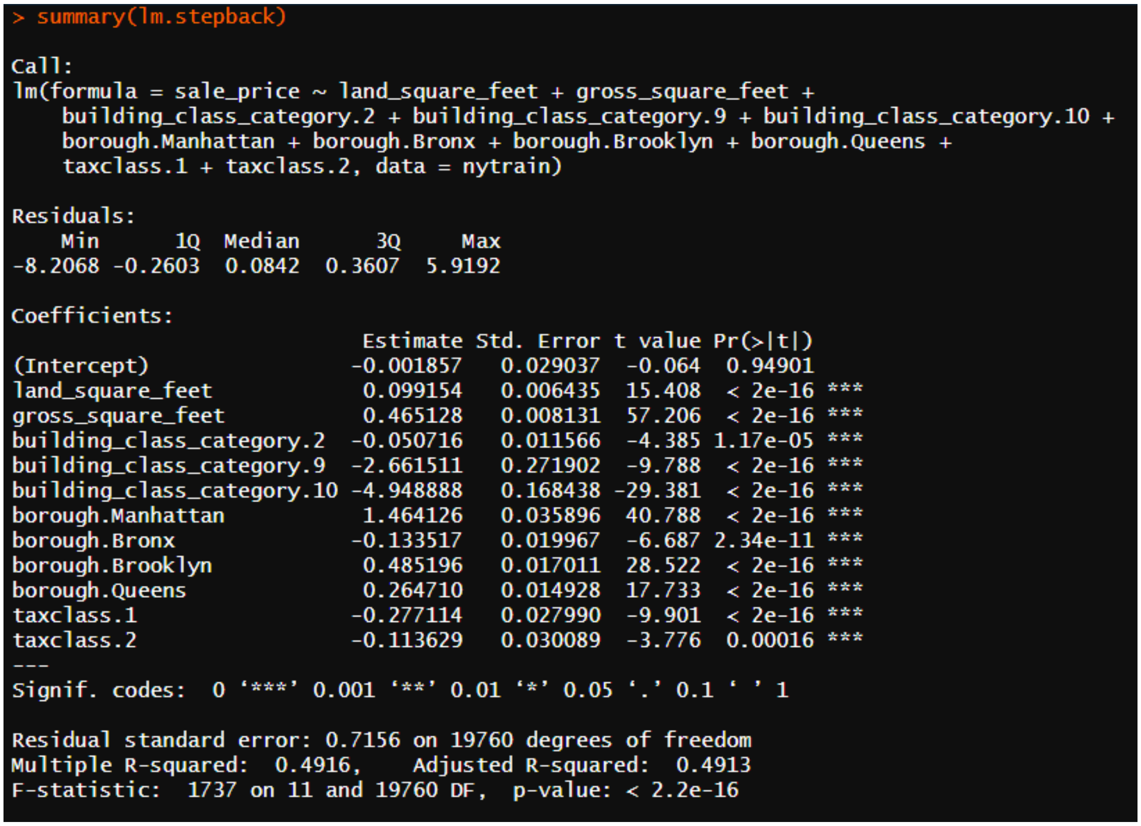


Figure 8



```
> summary(lm.stepback)

Call:
lm(formula = sale_price ~ land_square_feet + gross_square_feet + 
    building_class_category.2 + building_class_category.9 + building_class_category.10 + 
    borough.Manhattan + borough.Bronx + borough.Brooklyn + borough.Queens + 
    taxclass.1 + taxclass.2, data = nytrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.2068 -0.2603  0.0842  0.3607  5.9192 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -0.001857   0.029037  -0.064  0.94901    
land_square_feet           0.099154   0.006435  15.408  < 2e-16 ***
gross_square_feet          0.465128   0.008131  57.206  < 2e-16 ***
building_class_category.2 -0.050716   0.011566  -4.385 1.17e-05 ***
building_class_category.9 -2.661511   0.271902  -9.788  < 2e-16 ***
building_class_category.10 -4.948888   0.168438 -29.381  < 2e-16 ***
borough.Manhattan          1.464126   0.035896  40.788  < 2e-16 ***
borough.Bronx             -0.133517   0.019967  -6.687 2.34e-11 ***
borough.Brooklyn           0.485196   0.017011  28.522  < 2e-16 ***
borough.Queens             0.264710   0.014928  17.733  < 2e-16 ***
taxclass.1                -0.277114   0.027990  -9.901  < 2e-16 ***
taxclass.2                -0.113629   0.030089  -3.776  0.00016 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7156 on 19760 degrees of freedom
Multiple R-squared:  0.4916,	Adjusted R-squared:  0.4913 
F-statistic:  1737 on 11 and 19760 DF,  p-value: < 2.2e-16
```

Figure 9

Note that all parameters in **Figure 9** are significant at an alpha of 0.0001—so no further variables were removed from the model. Notice the high parameter estimate for the borough.Manhattan variable, which matches our observation from **Figure 7**—it is generally more expensive to buy a property in Manhattan than in the other NYC boroughs. A residual analysis was done—comparing the model residuals against the model predictors—to determine if a polynomial fit was needed for any of the predictors. The residual analysis showed no need for a polynomial fit.

Using the same variables as selected via the backward stepwise regression method, three more models were trained and evaluated. These include the ridge regression model, the lasso regression model, and the random-forest regression model. Several parameters were tested for each of the models. A lambda of 0.01 led to the minimum BIC for both the ridge and lasso regression models. The random-forest model performed optimally with the ‘ntree’ parameter set to 500. The test RMSE values of these four models are summarized in **Table 3**. The random forest regression model has the lowest test RMSE.

Model	**Test RMSE**
Linear Regression	0.6847841
Ridge Regression	0.6849772
Lasso Regression	0.6877079
Random Forest Regression	0.6716746

Table 3

In **Figure 10** the sales price predictions are shown plotted against their actual values. Ideally, the points in this plot should be close to the diagonal line with a constant variance or scatter.

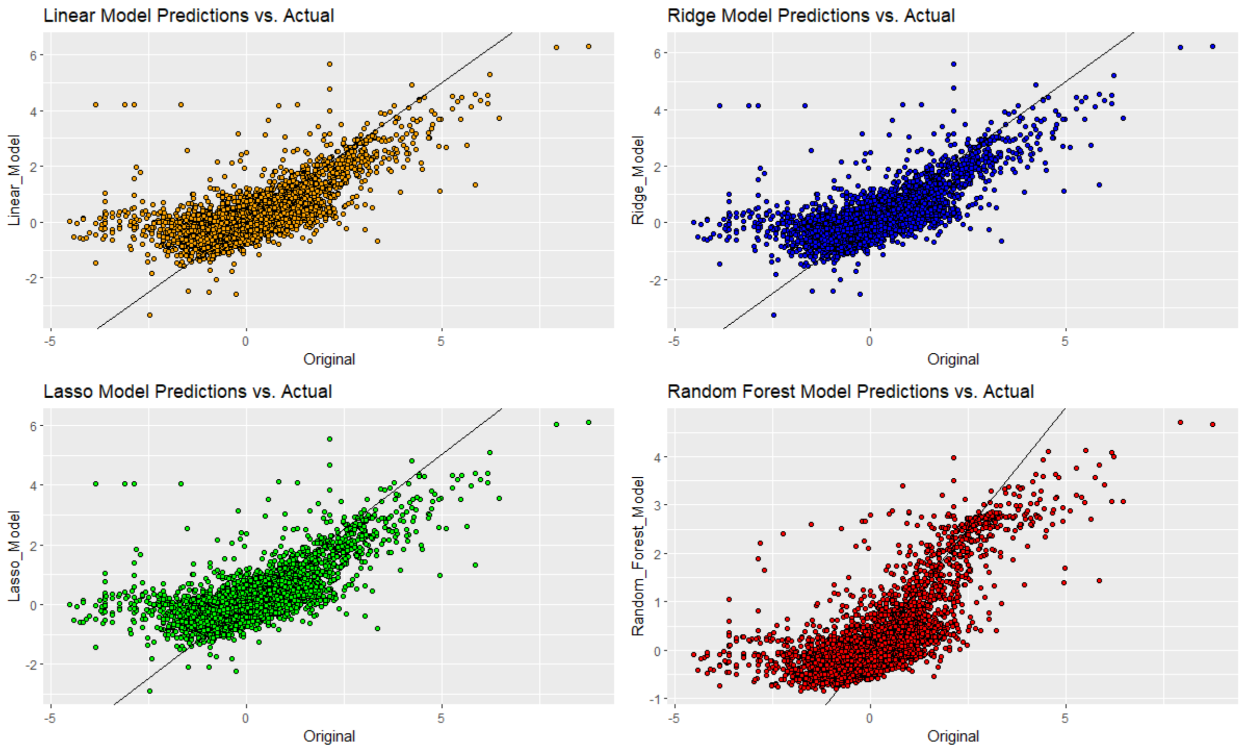


Figure 10

IV. Discussion and Limitations

Considering the test RMSE values and the scatter plots of the sales_price predictions against the observed, we can claim that our models performed adequately. No one model is significantly better than the others, but the random forest regression model does have the lowest test RMSE. We performed extensive EDA prior to modeling, and in the process eliminated many variables. The predictors we modelled with—gross square feet, land square feet, year built, borough, tax class, and building class categories—all have strong estimates and low p-values, which highlights their importance in predicting the sales price of a property in NYC. Though our models are able to predict the sales price, they do have room to improve extensively. If one of these models is to serve as our "second level investor insight" in Howard Mark's theory, it may not give us the edge against the other investors looking to deal in NYC's real-estate market. A possible reason for such an average performance could be that we eliminated more variables than we needed to. For example, the total number of units might provide information that the current variables do not, and lead to an improved performance. The random forest model, especially, might perform better with more variables. Our first course of action should be to revisit the EDA portion of the study and explore the inclusion of other variables into the models, to determine if they improve performance. Another reason could stem from the dataset inherently. As noted previously, many values were ***NaNs*** or zeros and we had to drop these observations from our dataset. A more complete dataset could lead to models that perform better—and possibly deliver the insight we are looking for. The dataset also does not include variables that are often considered when purchasing property—number of bedrooms, number of bathrooms, etc. These variables were not accounted for in the dataset, and perhaps their inclusion can lead to more complete models. Future studies should consider these observations and suggestions.