New York Property Sales Analysis

Non-Technical Summary – New York Property Sales Analysis

Project Goal:

The objective of this analysis project is to explore and analyze the NYC property sales dataset, focusing on specific building class categories, and provide insights into the factors influencing property prices. The project involves data preprocessing, exploratory data analysis, feature engineering, and encoding categorical variables for further analysis. By applying LASSO and RIDGE regression models as well as CA, I aim to build predictive models that can accurately estimate property prices and identify significant predictors.

Data Source and Schema:

The dataset represents a comprehensive collection of property sales in the New York City real estate market. It contains information about the location, address, type, sale price, and sale date of each building unit sold. With a total of 84,549 instances, we will use 70% of the data for training and 30% for testing purposes. The dataset encompasses 22 fields and provides a comprehensive record of building and apartment sales in New York City.

Project Modules:

Importing Libraries
Data Preprocessing
Data Visualization
Correspondence Analysis
Regression
Ordinary Least Square Regression, Lasso and Ridge Regression
Visualizations

Methods Used: Supervised Learning:

Linear Regression: Predicts a continuous dependent variable based on labeled training data.
LASSO Regression: Extends linear regression with feature selection using a regularization term.
Ridge Regression: Extends linear regression with regularization to handle multicollinearity.

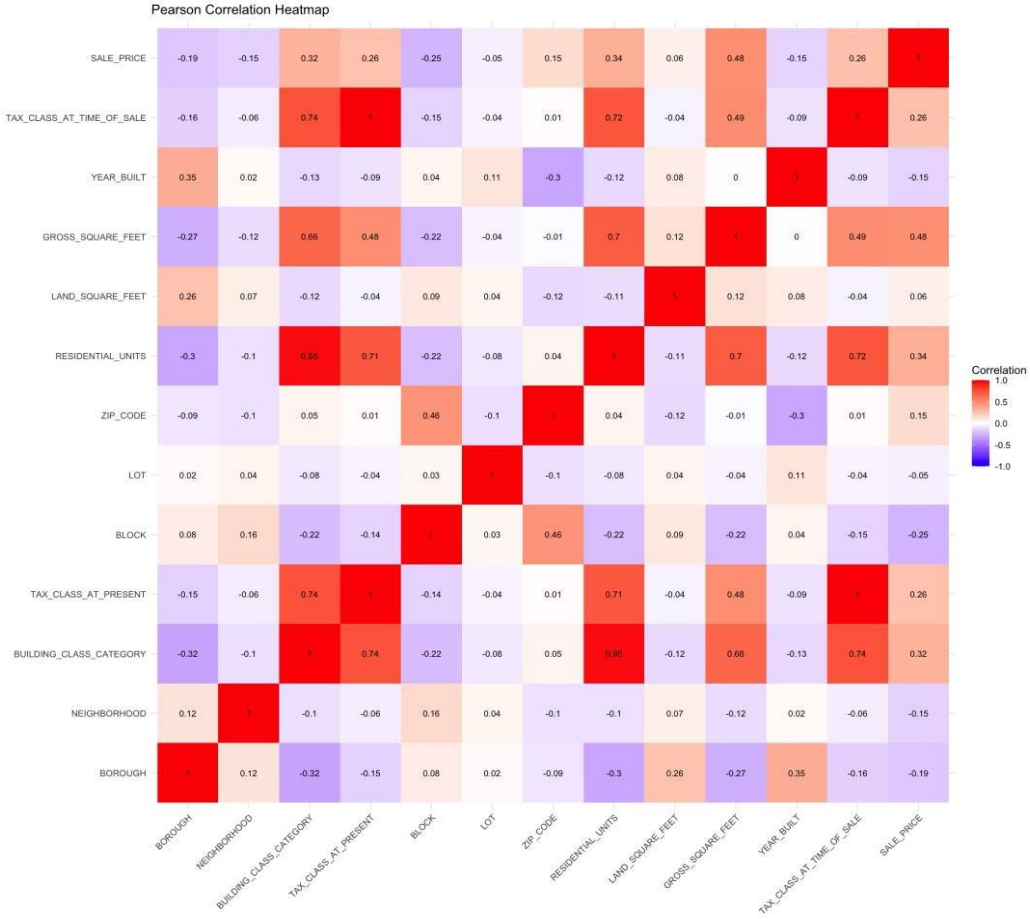
Data Pre-processing:

1. Deletion of Unnecessary Columns and Cleansing: Columns such as EASE-MENT,X and SALE DATE, which were either empty or irrelevant to the analysis, were dropped from the dataset. Duplicates were also removed, and the effectiveness of this step was verified by checking the dataset's shape and obtaining descriptions of each column.
2. Categorizing and Numerical Variable Specification: To ensure proper data representation, the format of the columns was adjusted to reflect the appropriate data types. Initially, all columns were of the 'object' type.
3. Conversion of Tax Class Data Type: The Tax class column was sliced and converted into numerical variables. Specifically, values such as 1A, 1B, 1C, and 1D were changed to 1 based on information provided in an article referenced from {https://www1.nyc.gov/assets/finance/downloads/pdf/brochures/class_1_guide.pdf }.
4. Selection of Significant Sales Price Values: In the training data, only Sales Price values greater than 50,000 were considered. This decision was based on the observation that there were numerous insignificant sales prices. This threshold was chosen as the minimum significant price for a house in the New York state, ensuring that only relevant data was included in the analysis.
5. Calculation of New Feature "Building Age": Through analysis, it was determined that converting the YEAR BUILT column to the BUILDING AGE would provide more meaningful information. The BUILDING AGE was calculated by subtracting the YEAR BUILT from the reference year of 2017. Subsequently, the YEAR BUILT column was dropped from the dataset .

VARIABLE SELECTION:

1. For our analysis, we have selected the following variables that are relevant and provide valuable insights into the properties sold in the New York City property market:
2. Borough: This categorical variable is categorized into five levels representing different boroughs (1, 2, 3, 4, 5). It helps us understand the geographic location and its impact on property prices.
3. Block: This numerical variable represents the block number. It provides information about the specific location of the property and its influence on sale prices.
4. Building Class at Present: This categorical variable is categorized into 24 levels, indicating the class of the building at present. It offers insights into the types of properties sold and their impact on sale prices.
5. Gross Square Feet: This numerical variable represents the total area of the building. It provides information on the size of the property and its influence on sale prices.
6. Tax Class at Time of Sale: This categorical variable is categorized into three levels, indicating the tax class at the time of sale. It helps us understand the tax implications and their impact on property prices.
7. Sale Price: This numerical variable represents the price at which the building was sold. It is the dependent variable in our analysis and serves as the target variable for predicting property prices.
8. Building Age: This numerical variable indicates the age of the building in years. It provides insights into the impact of building age on property prices.
9. These selected variables form the core set of features for our analysis, enabling us to explore the factors influencing property prices in the New York City real estate market.

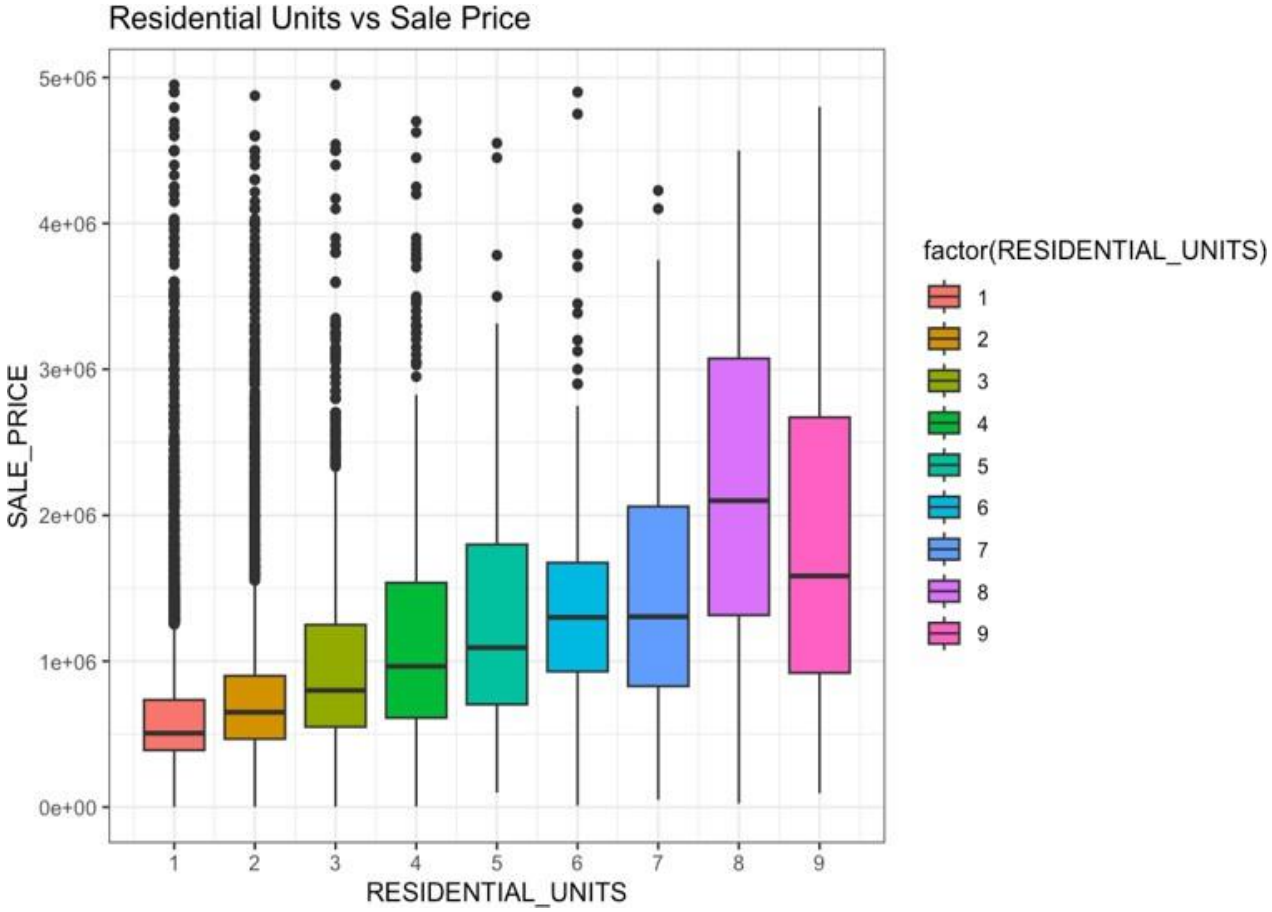
Pearson Correlation Map Findings:

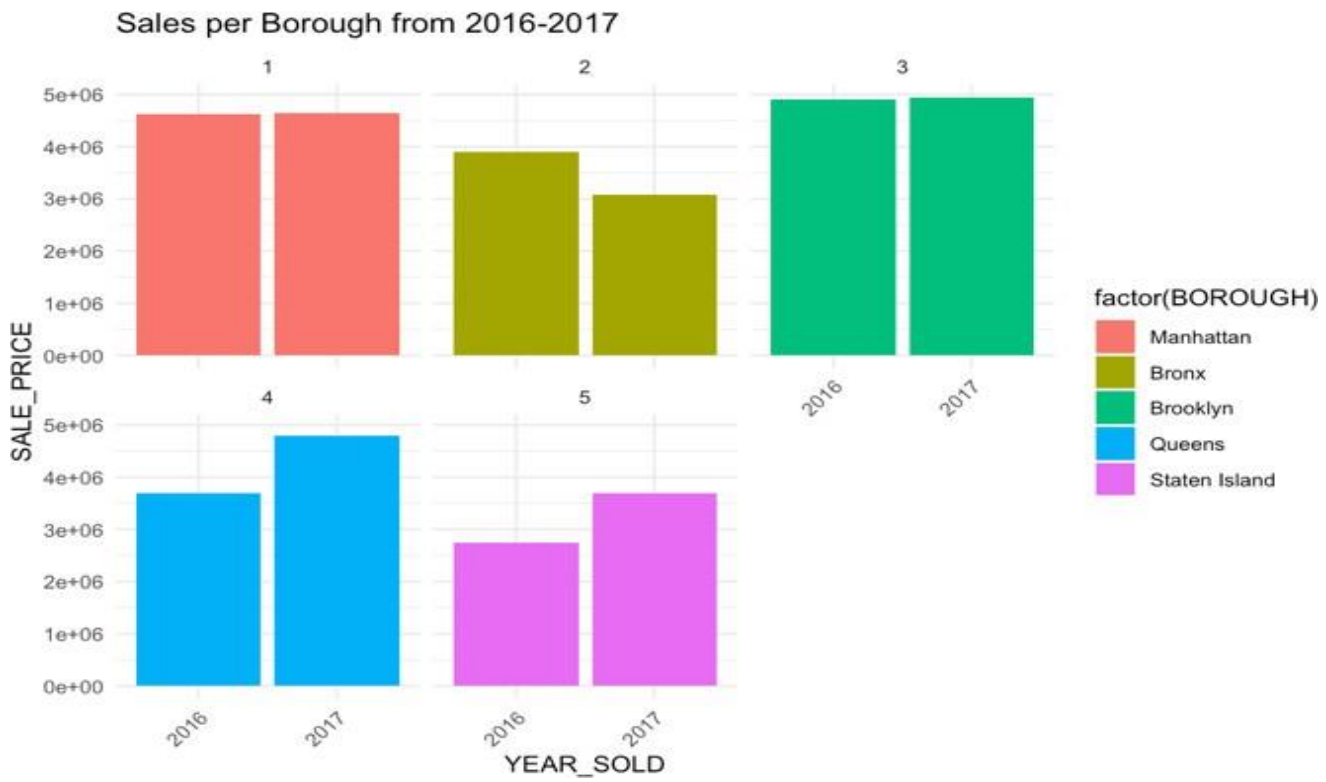


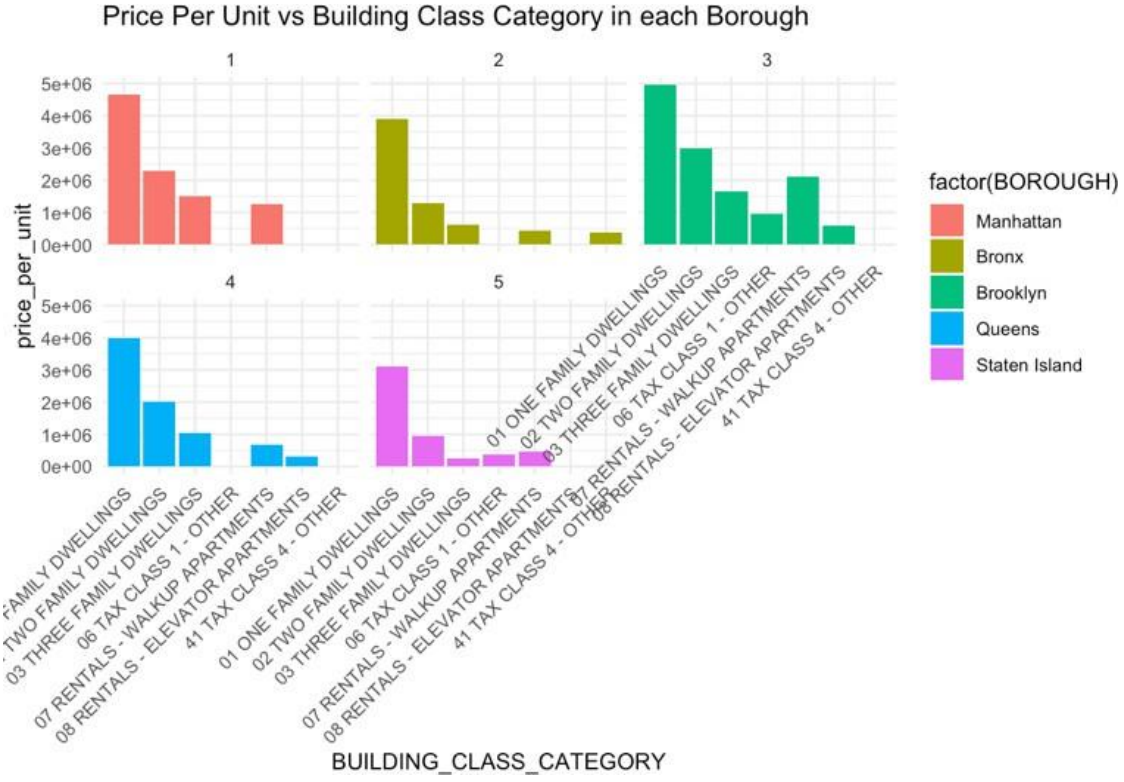
Based on the Pearson correlation map analysis, and the other figures above we chose a few features for our machine learning model and we have identified several key findings:

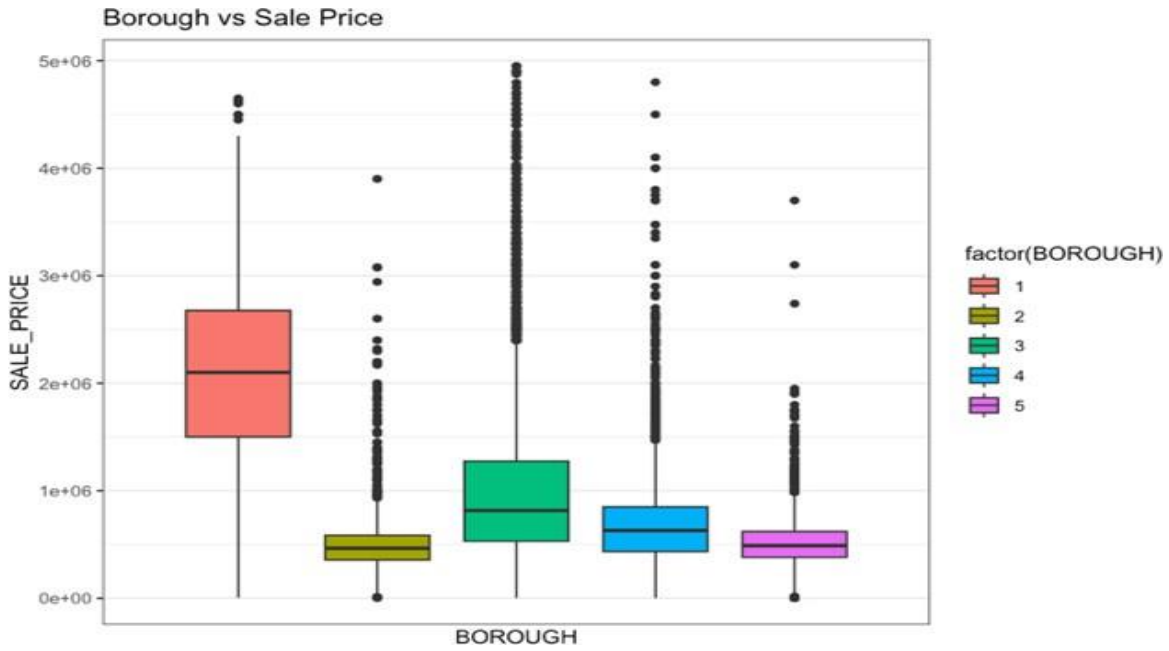
- Borough and Neighborhood: The variables Borough (values 1 to 5)and Neighborhood (using label encoder to transform strings to numbers)show a strong correlation with the target variable (property prices). This suggests that the geographic location of the property has a significant impact on its price.
- Building Class Category (OneHotEncoder because number of attributes is low): The Building class category also demonstrates a strong correlation with the target variable. This indicates that the type or class of the building plays a crucial role in determining its price.
- Block(numerical), Residential Units(values 1 to 9), and Gross Square Feet(numerical): The variables Block, Residential Units, and Gross Square Feet exhibit significant correlations with the property prices. This implies that factors such as the specific block number, the number of residential units in the building, and the total area of the building contribute to the variation in property prices.
- Tax Class at Present: It is important to note that the variable Tax Class at Present could potentially cause data leakage or bias in the analysis.
- Considering these findings, Borough, Neighborhood, Building Class Category, Block, Residential Units, and Gross Square Feet are identified as important predictors for property prices.

Exploratory Data Analysis (Visualizations) :




Sales per Borough from 2016-2017
1
2
3
4
5
SALE_PRICE
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
2016
2017
YEAR_SOLD
factor(BOROUGH)
Manhattan
Bronx
Brooklyn
Queens
Staten Island

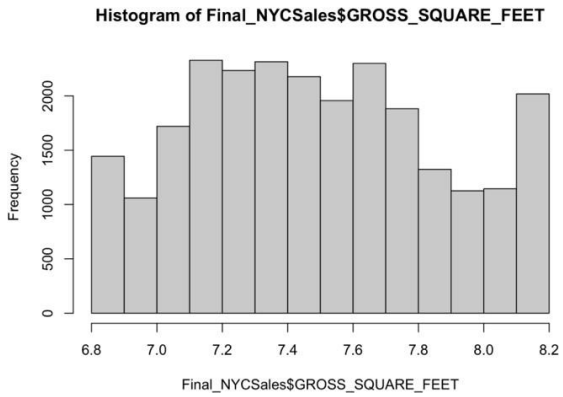

Price Per Unit vs Building Class Category in each Borough
1
2
3
4
5
price_per_unit
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
01 ONE FAMILY DWELLINGS
02 TWO FAMILY DWELLINGS
03 THREE FAMILY DWELLINGS
06 TAX CLASS 1 - OTHER
07 RENTALS - WALKUP APARTMENTS
08 RENTALS - ELEVATOR APARTMENTS
41 TAX CLASS 4 - OTHER
BUILDING_CLASS_CATEGORY
factor(BOROUGH)
Manhattan
Bronx
Brooklyn
Queens
Staten Island

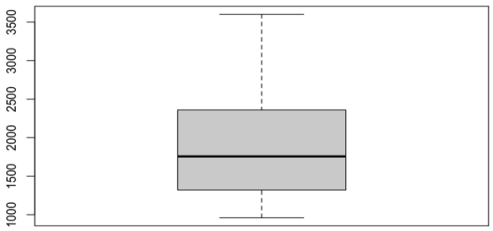


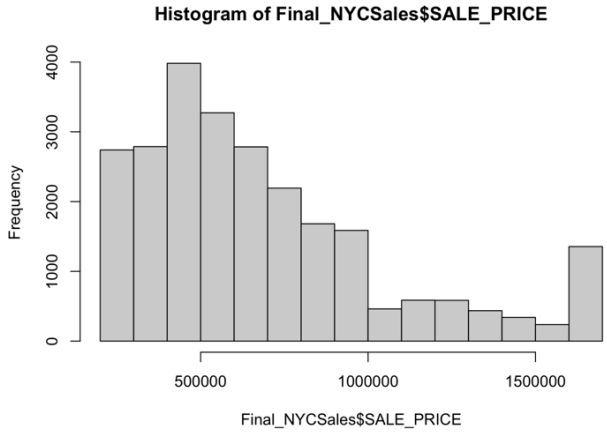
Outlier Analysis and Log Transformation:

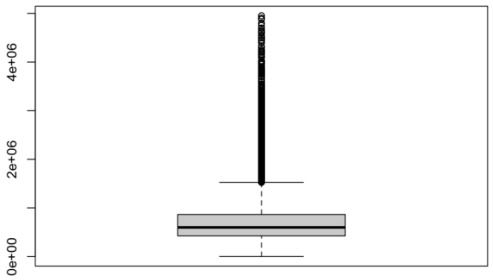
Before and after removing outliers.

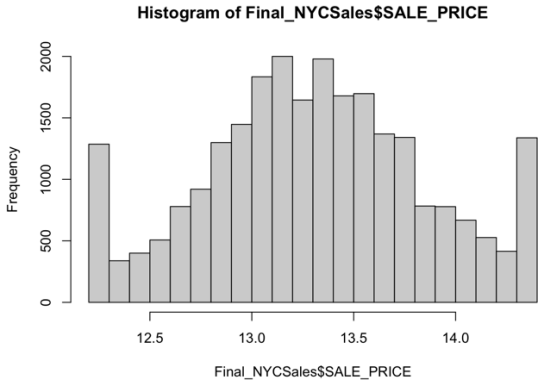
The below histogram of gross_square_feet and sale_price after transformation, we have used log transformation over here.

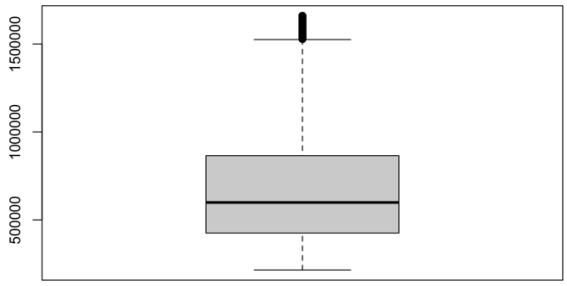












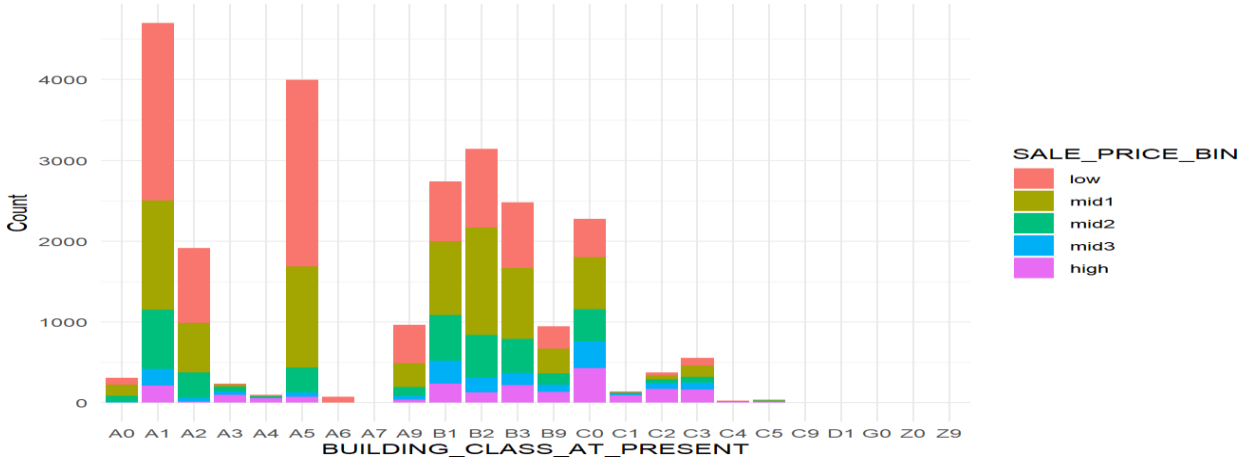
CORRESPONDENCE ANALYSIS:

The correspondence analysis was performed using the variables - BOROUGH and BUILDING_CLASS_AT_PRESENT and SALE_PRICE_BIN. This analysis aimed to explore the relationships between these categorical variables. The variable SALE_PRICE_BIN was divided into five categories representing different price ranges: Category 1 (low), Category 2 (mid1), Category 3 (mid2), Category 4 (mid3), and Category 5 (high). These categories were determined based on specific sale price thresholds.

Category	Range	Category level
1	SALE_PRICE <= 503000	Low
2	503001 <= SALE_PRICE <= 791000	Mid1
3	791001 <= SALE_PRICE <= 1079000	Mid2
4	1079001 <= SALE_PRICE <= 1367000	Mid3
5	SALE_PRICE > 1367000	High

BUILDING_CLASS_AT_PRESENT:

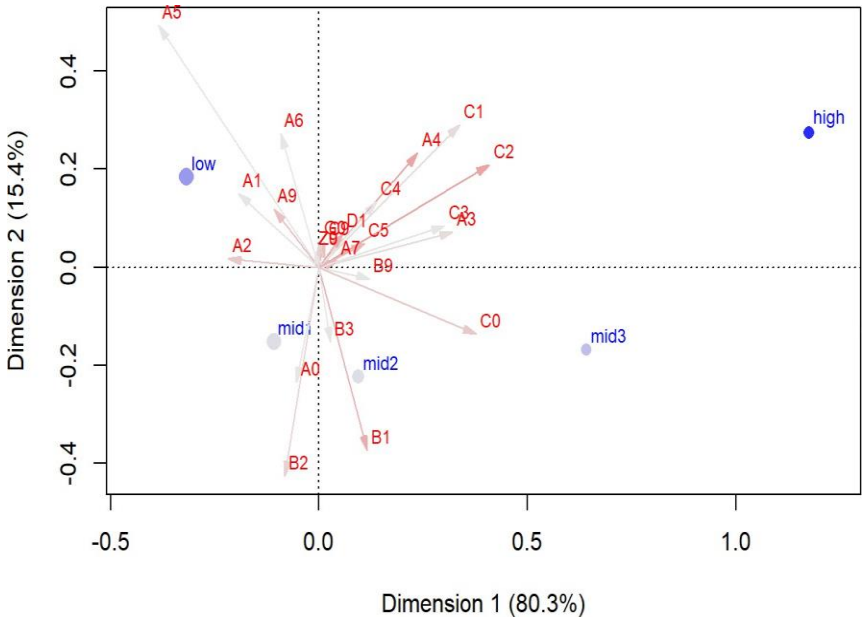
The variable BUILDING_CLASS_AT_PRESENT represents different classes of buildings. By conducting correspondence analysis between BUILDING_CLASS_AT_PRESENT and SALE_PRICE_BIN, insights can be gained regarding the associations or patterns between building classes and the categorized sale prices. This analysis aims to identify any tendencies where certain building classes tend to be associated with higher or lower sale prices compared to others.



The output of the correspondence analysis provides information on principal inertias (eigenvalues), which indicate the amount of variance explained by each dimension. In this case, the analysis resulted in four dimensions, with the first dimension explaining 80.3% of the total variance. The analysis also presents information on the rows and columns. In the rows section, each category of SALE_PRICE_BIN is described, including its mass (number of observations), quality (contribution to total variance), inertia (inertia on each dimension), and correlations with the dimensions. Similarly, in the columns section, each category of BUILDING_CLASS_AT_PRESENT is described, including its mass, quality, inertia, and correlations with the dimensions.

Building Class	Description
A0	One-family dwellings (attached or semi-detached)
A1	One-family dwellings (semi-detached)
A2	Two-family dwellings (attached)
A3	Multi-family walk-up apartments (3 or more units)
A4	Multi-family elevator apartments (3 or more units)
A5	Cooperative units (apartment buildings)
A6	Condominium units (apartment buildings)
A7	Office buildings
A9	Open space (non-buildable land)
B1	Store buildings (primarily retail)
B2	Store buildings (mixed use)
B3	Store buildings (primarily residential)
B9	Other stores (hotel, garage, etc.)
C0-C9	Walk-up apartments (mostly 3 to 6 stories)
D1	Elevator apartments (mostly 7 or more stories)
G0	Miscellaneous mixed-use buildings
Z0	Open space with a residential building
Z9	Miscellaneous structures of any zoning class

```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %    cum%   scree plot
##  1      0.187036  80.3  80.3  ********************
##  2      0.035913  15.4  95.7  ****
##  3      0.005787   2.5  98.2  *
##  4      0.004236   1.8 100.0
##         -------- -----
##  Total: 0.232972 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  low |  381  994  221 | -317 745 205 |  184 249 357 |
## 2 | mid1 |  317  821   58 | -109 279  20 | -152 542 203 |
## 3 | mid2 |  152  766   50 |   95 116   7 | -223 649 211 |
## 4 | mid3 |   65  880  140 |  641 824 144 | -168  57  51 |
## 5 | high |   84  995  531 | 1176 943 624 |  274  51 177 |
##
## Columns:
##      name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1  |   A0 |   12  795   14 | -209 170   3 | -401 625  55 |
## 2  |   A1 |  188  873   38 | -193 782  37 |   66  91  23 |
## 3  |   A2 |   77  949   40 | -340 948  47 |   12   1   0 |
## 4  |   A3 |    9  967   86 | 1432 958 103 |  141   9   5 |
## 5  |   A4 |    4  959   56 | 1642 810  56 |  704 149  54 |
## 6  |   A5 |  160  982  159 | -417 748 148 |  234 235 243 |
## 7  |   A6 |    3  914   20 | -728 340   8 |  945 574  74 |
## 8  |   A7 |    0  776    2 | 2101 763   2 |  280  14   0 |
## 9  |   A9 |   39  932   13 | -240 761  12 |  114 172  14 |
## 10 |   B1 |  109  926   35 |  152 309  13 | -214 617 140 |
## 11 |   B2 |  126  846   39 | -100 136   7 | -228 711 182 |
## 12 |   B3 |   99  672    6 |   38  96   1 |  -92 576  24 |
## 13 |   B9 |   38  955   12 |  269 949  15 |  -22   7   1 |
## 14 |   C0 |   91  962  121 |  540 938 142 |  -85  23  18 |
## 15 |   C1 |    6  967  108 | 1959 846 114 |  740 121  85 |
## 16 |   C2 |   15  999  141 | 1448 951 167 |  323  47  43 |
## 17 |   C3 |   22  990   74 |  874 976  90 |  107  15   7 |
## 18 |   C4 |    1  960   19 | 1925 813  19 |  819 147  18 |
## 19 |   C5 |    1  957   10 | 1220 921  12 |  240  36   2 |
## 20 |   C9 |    0  621    1 |  400 155   0 |  693 466   3 |
## 21 |   D1 |    0  874    4 | 2719 681   3 | 1447 193   5 |
## 22 |   G0 |    0  931    0 |   53   5   0 |  710 926   3 |
## 23 |   Z0 |    0  909    0 | -734 332   0 |  968 577   1 |
## 24 |   Z9 |    0  909    0 | -734 332   0 |  968 577   1 |
```

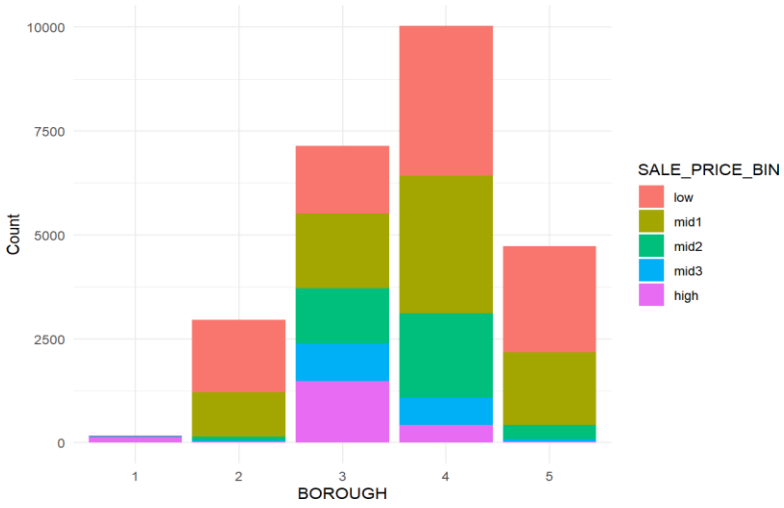


The first dimension explains 80.3% of the variance, capturing the majority of the relationships. "Mid3" and "high" have a stronger association with the first dimension, indicating higher influence on the building class and sale price relationship. Similarly, "A3" and "A4" have a stronger association with the first dimension, influencing the sale price and building class relationship. Specific categories like "mid3" tend to be associated with lower sale prices (negative correlation with "A0" and "A1"), while "high" is associated with higher sale prices (positive correlation with "A3," "A4," and "A5"). The plot visually represents these relationships, providing insights into building class and sale price patterns in specific New York City boroughs. The cumulative explained variance of the first two dimensions is 95.7%.

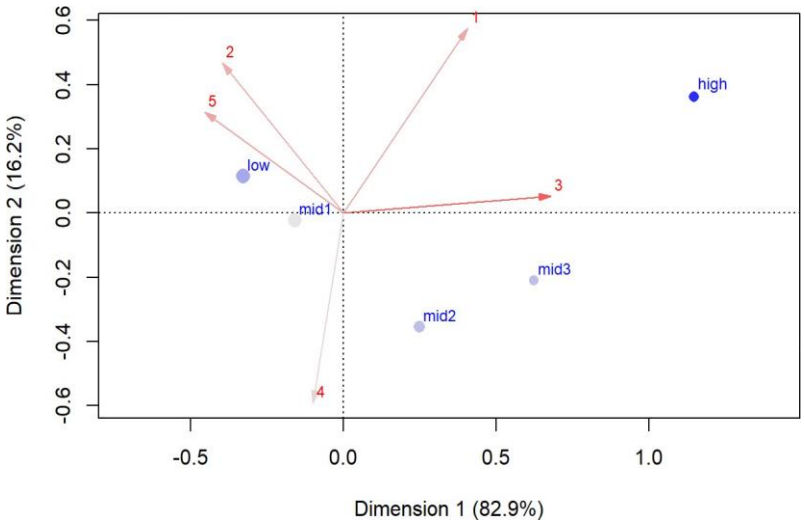
BOROUGH:

In the context of New York City, the borough numbers 1, 2, 3, 4, and 5 represent the five boroughs that make up the city. Each borough is a distinct administrative division with its own characteristics and local government. Here's a breakdown of what each borough represents:

Borough	Description
Manhattan(1)	The most densely populated borough located at the center of New York City.
Bronx(2)	Located north of Manhattan, it is primarily situated on the mainland.
Brooklyn(3)	Located on the western end of Long Island, it is the most populous borough.
Queens(4)	Located on Long Island, east of Manhattan, it is the largest borough in terms of land area.
Staten Island(5)	Located in the southwestern part of the city, it is separated from the rest of the city by the waters of the New York Harbor




```
##
## Principal inertias (eigenvalues):
##
##  dim    value      %    cum%   scree plot
##  1      0.194921  82.9  82.9  **********
##  2      0.038046  16.2  99.1  ****
##  3      0.002176   0.9 100.0
##  4      2.9e-050   0.0 100.0
##         -------- -----
##  Total: 0.235173 100.0
##
##
## Rows:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |  low |  381 1000  196 | -328 890 210 |  115 109 132 |
## 2 | mid1 |  317  998   35 | -159 978  41 |  -22  19   4 |
## 3 | mid2 |  152  983  123 |  249 326  48 | -353 657 498 |
## 4 | mid3 |   65  947  127 |  623 852 130 | -208  95  74 |
## 5 | high |   84  999  520 | 1148 909 570 |  362  90 291 |
##
## Columns:
##     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |    1 |    7  977  195 | 2141 703 165 | 1337 274 330 |
## 2 |    2 |  118  996  167 | -510 784 158 |  265 212 218 |
## 3 |    3 |  285  994  384 |  561 993 461 |   20   1   3 |
## 4 |    4 |  401  972   67 |  -69 122  10 | -182 850 349 |
## 5 |    5 |  189 1000  187 | -462 914 207 |  141  86  99 |
```



The first dimension explains 82.9% of the total variance, indicating that it captures the majority of the relationships between the BOROUGH variable and SALE_PRICE categories. The second dimension explains an additional 16.2% of the variance, contributing to a cumulative explained variance of 99.1%. The BOROUGH categories are represented in the rows. Categories such as "mid2" and "high" have a stronger association with the first dimension, indicating a higher influence on the relationship between BOROUGH and SALE_PRICE categories.The SALE_PRICE categories

are represented in the columns. Categories such as "1" and "5" have a stronger association with the first dimension, suggesting a higher influence on the relationship between SALE_PRICE and BOROUGH categories. Based on the correlations and contributions, specific BOROUGH categories are closely associated with certain SALE_PRICE categories. For example, "mid2" has a negative correlation with "1" and "3," indicating that it tends to be associated with lower SALE_PRICE categories. Conversely, "high" has a positive correlation with "1" and "5," suggesting a tendency for higher SALE_PRICE.

REGRESSION(OLS Model):

The OLS Model is a statistical method that estimates the relationship between a dependent variable and independent variables. It assumes a linear relationship and minimizes the squared differences between observed and predicted values. Assumptions include linearity, independence, constant variance, and normality of residuals. The model was highly significant ($p < 2.2e-16$), indicating a strong association. It explained 39.6% of the variability in sale price, suggesting a moderate impact of independent variables. The adjusted R-square is 0.396, and the residual standard error is 0.4141.

```
#OLS MODEL
# Fit the initial linear regression model
OLS_MODEL <- lm(SALE_PRICE ~ BOROUGH + BLOCK + GROSS_SQUARE_FEET + TAX_CLASS_AT_TIME_OF_SALE + BUILDING_CLASS_AT_
PRESENT + BUILDING_AGE , data = Final_NYCSales)
summary(OLS_MODEL)
```

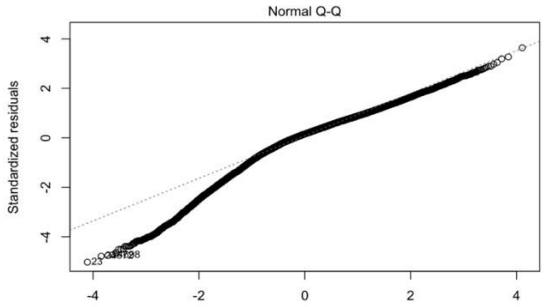
```
##
## Call:
## lm(formula = SALE_PRICE ~ BOROUGH + BLOCK + GROSS_SQUARE_FEET +
##     TAX_CLASS_AT_TIME_OF_SALE + BUILDING_CLASS_AT_PRESENT + BUILDING_AGE,
##     data = Final_NYCSales)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.06546 -0.20546  0.06476  0.27462  1.50675
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.001e+01  8.822e-02 113.488  < 2e-16 ***
## BOROUGH2                        -8.141e-01  3.391e-02 -24.008  < 2e-16 ***
## BOROUGH3                        -2.306e-01  3.321e-02  -6.944 3.91e-12 ***
## BOROUGH4                        -2.727e-01  3.372e-02  -8.089 6.31e-16 ***
## BOROUGH5                        -7.117e-01  3.432e-02 -20.734  < 2e-16 ***
## BLOCK                           -4.509e-05  8.746e-07 -51.557  < 2e-16 ***
## GROSS_SQUARE_FEET                5.714e-01  1.019e-02  56.078  < 2e-16 ***
## TAX_CLASS_AT_TIME_OF_SALE2       3.680e-01  2.073e-01   1.775 0.075930 .
## TAX_CLASS_AT_TIME_OF_SALE4      -2.332e-02  4.149e-01  -0.056 0.955172
## BUILDING_CLASS_AT_PRESENTA1     -1.727e-01  2.465e-02  -7.006 2.51e-12 ***
## BUILDING_CLASS_AT_PRESENTA2     -9.076e-02  2.558e-02  -3.549 0.000388 ***
## BUILDING_CLASS_AT_PRESENTA3      1.731e-01  3.688e-02   4.695 2.68e-06 ***
## BUILDING_CLASS_AT_PRESENTA4     -1.518e-01  4.911e-02  -3.090 0.002001 **
## BUILDING_CLASS_AT_PRESENTA5     -2.756e-01  2.477e-02 -11.125  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTA6     -4.657e-01  5.384e-02  -8.651  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTA7      3.078e-01  2.940e-01   1.047 0.295148
## BUILDING_CLASS_AT_PRESENTA9     -2.348e-01  2.750e-02  -8.538  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTB1     -2.870e-01  2.579e-02 -11.129  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTB2     -2.722e-01  2.541e-02 -10.712  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTB3     -2.010e-01  2.548e-02  -7.886 3.25e-15 ***
## BUILDING_CLASS_AT_PRESENTB9     -2.725e-01  2.792e-02  -9.759  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTC0     -3.157e-01  2.664e-02 -11.849  < 2e-16 ***
## BUILDING_CLASS_AT_PRESENTC1     -4.305e-01  2.118e-01  -2.033 0.042080 *
## BUILDING_CLASS_AT_PRESENTC2     -6.070e-01  2.099e-01  -2.892 0.003836 **
## BUILDING_CLASS_AT_PRESENTC3     -6.921e-01  2.095e-01  -3.303 0.000956 ***
## BUILDING_CLASS_AT_PRESENTC4     -5.533e-01  2.254e-01  -2.455 0.014099 *
## BUILDING_CLASS_AT_PRESENTC5     -6.576e-01  2.195e-01  -2.997 0.002733 **
## BUILDING_CLASS_AT_PRESENTC9     -9.374e-01  2.790e-01  -3.360 0.000780 ***
## BUILDING_CLASS_AT_PRESENTD1     -2.868e-01  3.598e-01  -0.797 0.425321
## BUILDING_CLASS_AT_PRESENTG0     -5.158e-01  1.868e-01  -2.761 0.005760 **
## BUILDING_CLASS_AT_PRESENTZ0     -5.187e-01  4.149e-01  -1.250 0.211170
## BUILDING_CLASS_AT_PRESENTZ9             NA         NA      NA       NA
## BUILDING_AGE                    -1.140e-03  1.072e-04 -10.640  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4141 on 24997 degrees of freedom
## Multiple R-squared:  0.3968, Adjusted R-squared:  0.396
## F-statistic: 530.4 on 31 and 24997 DF,  p-value: < 2.2e-16
```

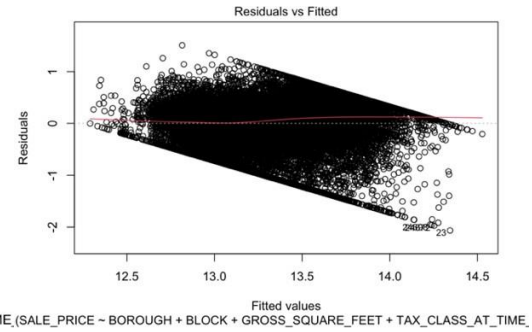
Important variables such as borough, block, gross square feet, and building age significantly influenced sale price. However, some building class at present categories did not have a statistically significant impact on sale price.

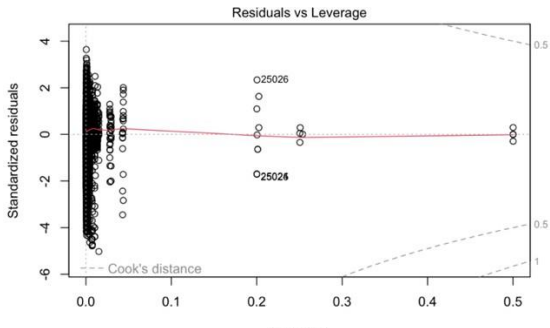
The following are the key findings of the study:

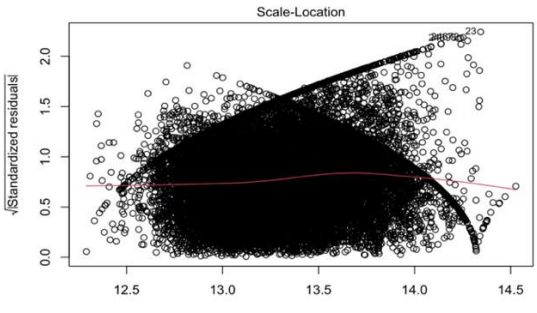
Sale price is significantly influenced by borough, block, gross square feet, and building age.Building class at present does not have a statistically significant impact on sale price.The model can be used to predict sale price with moderate accuracy.

In the below plot, we can see the qq plot shows us that the data is not completely normal there are quiet a few outliers. From the Residuals vs fitted there is a clear set of boundary as set by the outlier analysis. We can also the modeling providing certain predictability based on the predictors utilized although there are many outliers we must also refer to the output metrics of the model.



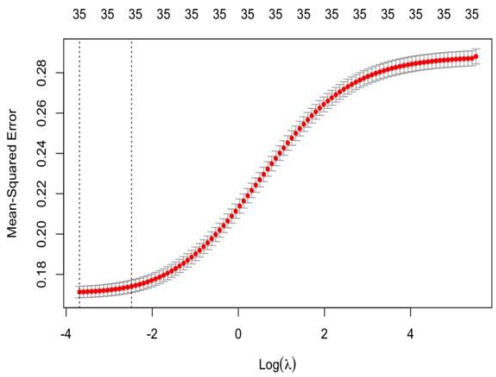


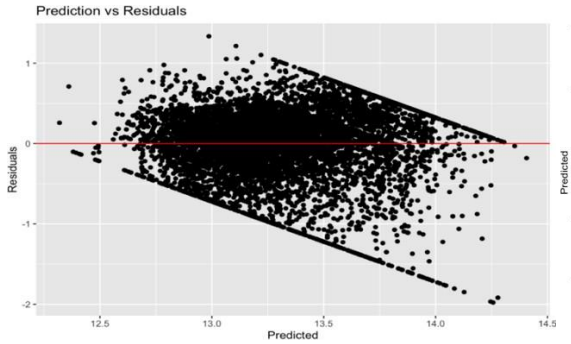


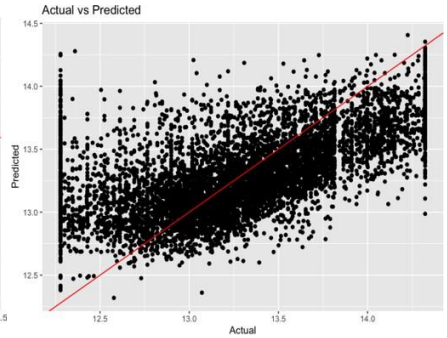


RIDGE REGRESSION:

Ridge regression is a regularization technique used in linear regression to reduce variance by adding a penalty term. It helps when there is multicollinearity in the data. The data is split into a training set (70%) and a test set (30%). Cross-validation is used to find the optimal lambda value, which is determined to be 0.02495994. The model is trained using this lambda value and the training data. Predictions are made on the test data. Model evaluation includes visualizing predicted vs. actual values and assessing residuals. The root mean squared error (RMSE) is 0.417293689493776, indicating prediction accuracy. The R-squared value is 0.364829192699975, showing the proportion of variance explained. The adjusted R-squared value, accounting for predictors, is 0.361854352842421. The model performs well overall, with accurate predictions, although slightly overfitting is observed based on the adjusted R-squared value.

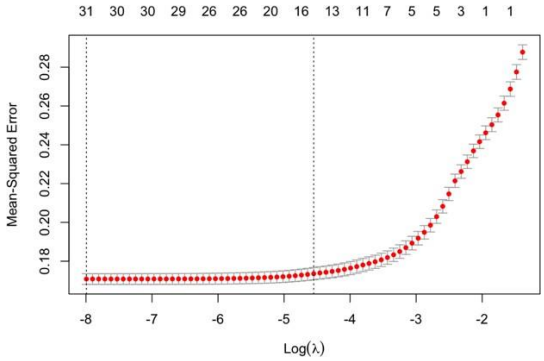


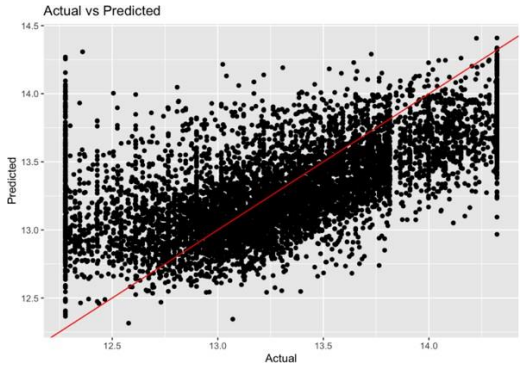


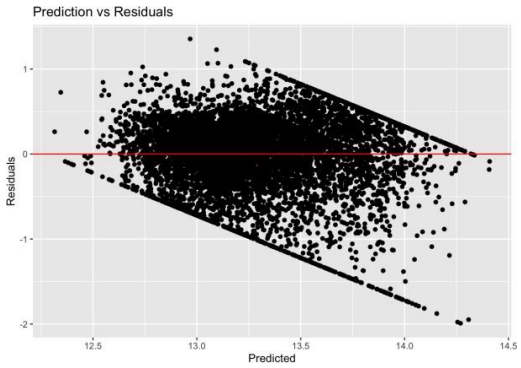


Lasso Regression:

Lasso Regression combines variable selection and regularization to enhance linear regression models. It involves steps such as data splitting, optimal lambda selection through cross-validation, model training using the chosen lambda value, prediction on test data, visualization of model performance, residual analysis, calculation of root mean squared error (RMSE), R-squared, and adjusted R-squared values. The optimal lambda was determined as 0.000338, balancing complexity and accuracy. The Lasso Regression model performed well, with an RMSE of 0.417539523681832, R-squared of 0.364080594301084, and adjusted R-squared of 0.361102248362444, indicating good prediction accuracy and explanatory power. Overall, Lasso Regression proved to be a valuable tool in our analysis.







Conclusion

The New York Property Sales Analysis successfully identified key factors influencing property prices across New York City, leveraging a comprehensive dataset and advanced analytical techniques. By integrating data preprocessing, exploratory data analysis, and regression modeling, the project provided meaningful insights into the real estate market dynamics.

Key predictors such as borough, building class, gross square feet, and building age were identified as significant contributors to property valuation. Statistical models like OLS, LASSO, and Ridge Regression demonstrated moderate accuracy in predicting sale prices, with Ridge and LASSO offering additional robustness through regularization. Correspondence analysis further uncovered relationships between building classes, sale price categories, and boroughs, revealing patterns that align with the distinct characteristics of NYC's real estate landscape.

Despite its successes, the project acknowledges areas for enhancement, such as addressing outliers and improving the handling of categorical variables to further refine model performance. These findings not only provide valuable insights for real estate stakeholders but also serve as a foundation for future studies to build more sophisticated predictive tools and decision-support systems.

In conclusion, this analysis bridges data-driven insights with practical implications for the NYC property market, offering a valuable resource for understanding and navigating its complexities.