NYC RealEstate

Fredrick Jones, Jian Quan Chen, Tilon Bobb

2024-04-24

Contents

Abstract

This study uses real estate transaction data to investigate the factors influencing property values in New York City (NYC). The dataset includes a variety of parameters that were gathered from public records and real estate listings, including property location, kind, size, and sale price. To understand the distribution and interrelationships of the dataset, exploratory data analysis (EDA) is the first systematic stage in the study technique. Data preparation is a step that comes next in order to encode categorical variables, handle missing values, and create new features. Stepwise regression, generalized linear models (GLM), robust regression, and conventional linear regression are all included in the design of regression models. The goal of developing predictive models that offer robustness against outliers while generalizing effectively to new data is achieved through the use of goodness-of-fit measures and diagnostic tests for residual analysis as the basis for model selection.

Keywords

NYC Real Estate, Data Analysis, Regression Modeling, Housing Market Trends, Neighborhood Analysis

Introduction

The New York real estate market is one of the most dynamic and influential sectors of urban development. With every real estate transaction representing a tangible real estate exchange, the NYC real estate market is a barometer of the city's socioeconomic landscape. In this report, we take an in-depth look at the NYC market through insights from a diverse dataset of New York real estate sales records (https://www.kaggle.com/datasets/datasciencedonut/current-nyc-property-sales).

Background and Motivation

The database contains a wealth of data spanning two decades, providing a detailed chronicle of real estate transactions in the five boroughs of New York - Manhattan, the Bronx, Brooklyn, Queens and Staten Island. Our analysis is based on the foundation created by this dataset, which provides insight into the multifaceted dynamics of NYC real estate sales.

The motivation is to identify the factors that influence NYC real estate prices, so our research is based on a variety of analytical techniques. and methods. From data analysis to regression modeling, we seek to uncover the complex interplay of variables that affect real estate. By examining real estate sales trends, identifying key predictors of sales prices, and assessing the impact of factors such as property size, location, and tax bracket, we aim to shed light on the mechanics behind the NYC real estate world.

As we delve into the depths of the NYC real estate market, our analysis aims to provide stakeholders with actionable insights on investors and from decision makers to real estate developers and potential buyers. By uncovering the drivers of real estate prices and delineating market trends, we aim to provide decision makers with the information they need to navigate the complexities of the real estate ecosystem.

Literature Review

Michael Gaynor's Project

Michael Gaynor conducted a project to explore the NYC real estate market using SQL queries and Tableau visualization techniques. His investigation aimed to answer four main questions:

1. Which of the five boroughs is the most expensive?
2. Which of the five boroughs have the most sales?
3. What type of properties sell the most in each of the 5 boroughs?
4. What property type influences sales?

Gaynor's approach involved understanding the task, prepping the data, analyzing the data using SQL queries and Tableau visualization, and presenting the findings through an interactive dashboard. Through his analysis, Gaynor discovered insights such as the most expensive borough, the borough with the most property sales, and the types of properties that sell the most in each borough.

Comparison and Evaluation

Gaynor's research utilized the same NYC real estate dataset to address similar questions as our investigation. However, there are significant differences between Gaynor's approach and our own project:

- **Methodology**: Gaynor primarily used SQL queries and Tableau visualization tools for data analysis, while our investigation utilized R programming language. Our approach involved a combination of data preprocessing, exploratory data analysis (EDA), statistical modeling, and visualization techniques implemented in R.

- **Presentation of Findings**: Gaynor's project focused on presenting the results through an interactive dashboard created in Tableau. In contrast, our investigation may present the findings through various formats such as tables, charts, and narratives within the R Markdown document.

- **Data Preprocessing**: While Gaynor mentioned data cleaning and preprocessing, the details of these steps were not extensively discussed. In our investigation, we employed specific techniques such as handling missing values, outlier detection and removal, and data transformation using R packages like dplyr and tidyr.

- **Statistical Modeling**: Our investigation may involve the application of statistical models such as linear regression, generalized linear models, or machine learning algorithms to explore relationships between variables and predict real estate prices. Gaynor's project did not explicitly mention the use of statistical modeling.

Advantages and Drawbacks

The advantages of Gaynor's approach include:

- Comprehensive analysis of the NYC real estate market using SQL and Tableau.
- Clear presentation of findings through interactive visualization.

However, there may be some drawbacks to Gaynor's approach, such as:

- Reliance on SQL and Tableau tools may limit accessibility for researchers unfamiliar with these technologies.
- Lack of detailed explanation of data cleaning and preprocessing steps.

Methodology

Data Preparation

The first part of our analysis consisted of importing the dataset that included data on property sales in New York City. After we uploaded the dataset, we conducted an exploratory data analysis to understand its organization, factors, and any problems that required attention.

During the exploratory data analysis, we faced one of the first hurdles with the discovery of missing values in multiple columns. To tackle this problem, we methodically pinpointed the variables with incomplete data and assessed the percentage of missing values in each instance. After careful deliberation, we made the choice to eliminate data points containing incomplete information, as they comprised less than 5% of the entire data set. This method enabled us to keep a large part of the data while reducing the effect of missing values on future analyses.

After addressing missing data, we focused on the distribution of numerical variables in the dataset. We noticed that many variables showed noticeable skewness, suggesting possible departures from normal distribution. To tackle this problem, we utilized Tukey's method for identifying and eliminating outliers. Our goal was to enhance the robustness of future analyses by improving the distributional properties of numeric variables through the identification and exclusion of outliers from the dataset.

Additionally, we examined the connections between variables using correlation analysis. This included computing correlation coefficients for pairs of numerical variables in order to evaluate the magnitude and orientation of their relationships. During this examination, we found multiple variables that showed strong positive relationships with each other, along with variables that had weaker or negative relationships. These results offered valuable information on potential factors that could predict real estate prices and guided the choice of variables for inclusion in regression analysis.

In addition, we analysed changes over time by graphing the time-based pattern of property values in New York City. This examination showed a rising pattern in mean sale prices over time, with variations in certain time frames. Through the visualization of time-based trends, we achieved a better comprehension of the fluctuations within the real estate market of New York City, pinpointing potential influences on property price fluctuations.

Regression Modeling

After completing thorough data preparation and exploratory analysis, we focused on the main objective of our research: using regression modeling to forecast real estate prices in New York City. The method we used involved carefully building linear regression models, using different predictor variables to understand the complex factors influencing real estate prices. A linear model was chosen here because it provides a simple and interpretable relationship between the features and the continuous target variable, allowing us to understand how each feature influences the price.

Leading our regression modeling was the incorporation of important predictor variables that were considered to have a significant impact on real estate prices. These variables included a wide range of factors, each providing valuable perspectives on the intricate fabric of the New York City real estate market. Included in these predictors were the quantity of housing units in a property, offering an insight into its size and ability to house residents. The categorization of taxes for properties at various points highlighted their financial status and legal ramifications, revealing insights into the larger economic and legal environments in which these properties function.

Furthermore, the year in which the building was constructed was identified as a crucial factor in predicting real estate values, giving a historical perspective on how they change over time. We aimed to capture temporal trends and identify any seasonal or cyclical patterns that could affect pricing dynamics by including the sale date of properties in our models. Moreover, factors like total area and land area were essential in evaluating the physical size and spatial characteristics of properties, enhancing our comprehension of their inherent value.

By conducting thorough regression analysis, we discovered significant statistical connections between these predictor variables and property prices, revealing the complex interaction of factors that influence pricing decisions in the real estate market of New York City. Nevertheless, even though our models were strong, the adjusted R-squared values suggested that there might be additional variability that was not accounted for, indicating the presence of hidden factors that were not included in our analysis.

To validate our model, we used a train/test (80/20) split strategy, where we reserved a portion of our data for testing the model's performance. We evaluated our model using metrics such as Mean Squared Error (MSE) and R-squared value. We also performed residual diagnostics to check the assumptions of linear regression, including linearity, independence, homoscedasticity, and normality.

Experimentation and Results

Explorartory Data Analysis & Data Preparation

Glipmse of the dataset

```
## Rows: 1,603,826
## Columns: 21
## $ BOROUGH                        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ NEIGHBORHOOD                   <chr> "ALPHABET CITY", "ALPHABET CITY", "ALPH~
## $ BUILDING.CLASS.CATEGORY        <chr> "01 ONE FAMILY DWELLINGS", "02 TWO FAMI~
## $ TAX.CLASS.AT.PRESENT           <chr> "1", "1", "2B", "2B", "2", "2A", "2A", ~
## $ BLOCK                          <int> 374, 377, 373, 373, 376, 377, 377, 379,~
## $ LOT                            <int> 46, 1, 16, 17, 54, 52, 52, 25, 45, 47, ~
## $ EASE.MENT                      <chr> "", "", "", "", "", "", "", "", "", "",~
## $ BUILDING.CLASS.AT.PRESENT      <chr> "A4", "S2", "C1", "C1", "C4", "C2", "C2~
## $ ADDRESS                        <chr> "347 EAST 4TH STREET", "110 AVENUE C", ~
## $ APARTMENT.NUMBER               <chr> "", "", "", "", "", "", "", "", "", "",~
## $ ZIP.CODE                       <dbl> 10009, 10009, 10009, 10009, 10009, 1000~
## $ RESIDENTIAL.UNITS              <dbl> 1, 2, 10, 10, 20, 5, 5, 7, 10, 10, 29, ~
## $ COMMERCIAL.UNITS               <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, ~
## $ TOTAL.UNITS                    <dbl> 1, 3, 10, 10, 20, 5, 5, 8, 10, 10, 29, ~
## $ LAND.SQUARE.FEET               <dbl> 2116, 1502, 2204, 2204, 2302, 2168, 216~
## $ GROSS.SQUARE.FEET              <dbl> 4400, 2790, 8625, 8625, 9750, 3728, 372~
## $ YEAR.BUILT                     <dbl> 1900, 1901, 1899, 1900, 1900, 1900, 190~
## $ TAX.CLASS.AT.TIME.OF.SALE      <int> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ BUILDING.CLASS.AT.TIME.OF.SALE <chr> "A4", "S2", "C1", "C1", "C4", "C2", "C2~
## $ SALE.PRICE                     <dbl> 399000, 2999999, 16800000, 16800000, 15~
## $ SALE.DATE                      <chr> "2022-09-29 00:00:00", "2022-09-15 00:0~
```

Assessing missing values

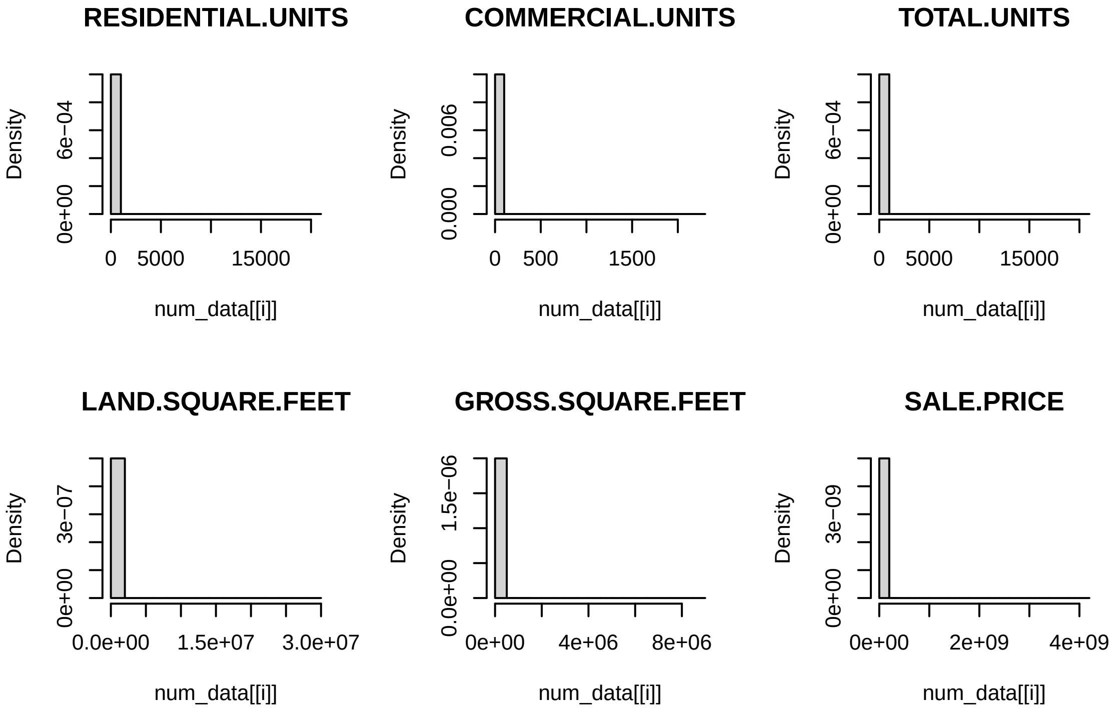
```
##
## ********** Number of missing values for each column **********

##                  ZIP.CODE         RESIDENTIAL.UNITS          COMMERCIAL.UNITS
##                        36                     76753                    113609
##               TOTAL.UNITS          LAND.SQUARE.FEET         GROSS.SQUARE.FEET
##                     70733                    119630                    119629
##                YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE                SALE.PRICE
##                     24024                         1                         1
```

We can drop missing values since there is less than 5% of dataset missing values hence safe to drop all missing values

Distribution Plot

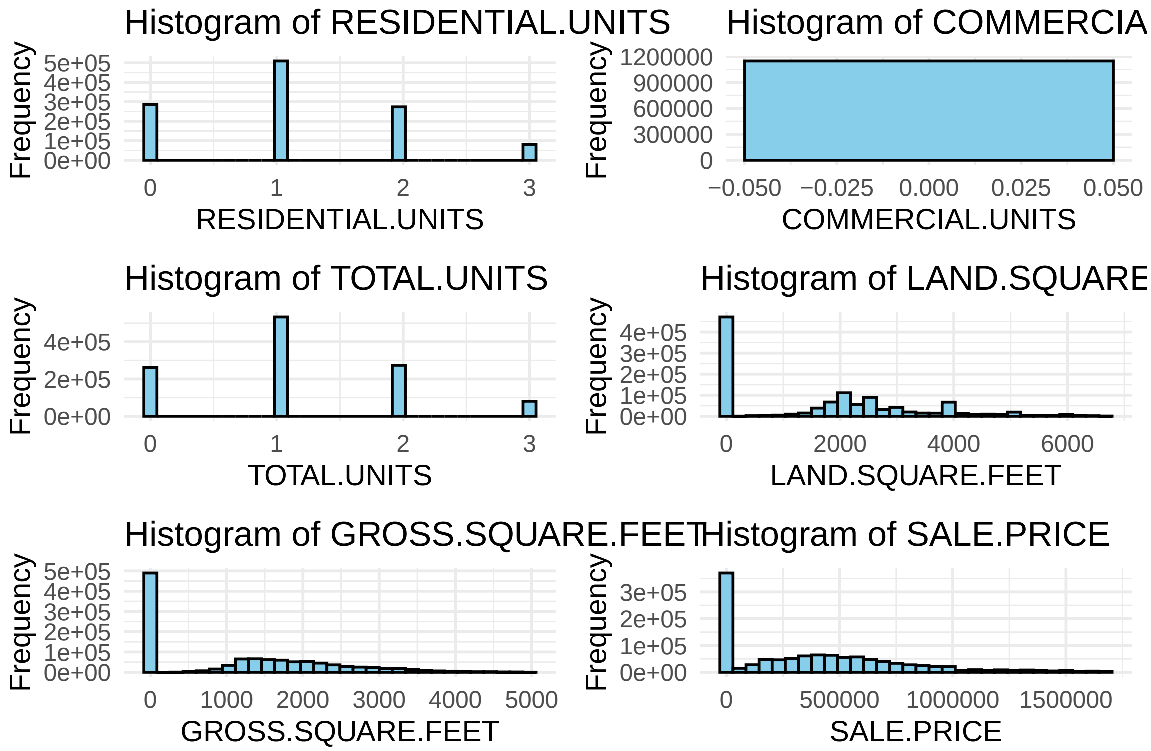
All numeric variables are heavily skewed to the right, hence a clear indication of outliers

All distributions exhibit a highly skewed pattern, with a single bar extending vertically at the rightmost end of the x-axis, suggesting the presence of potential outliers with extremely high unit counts.



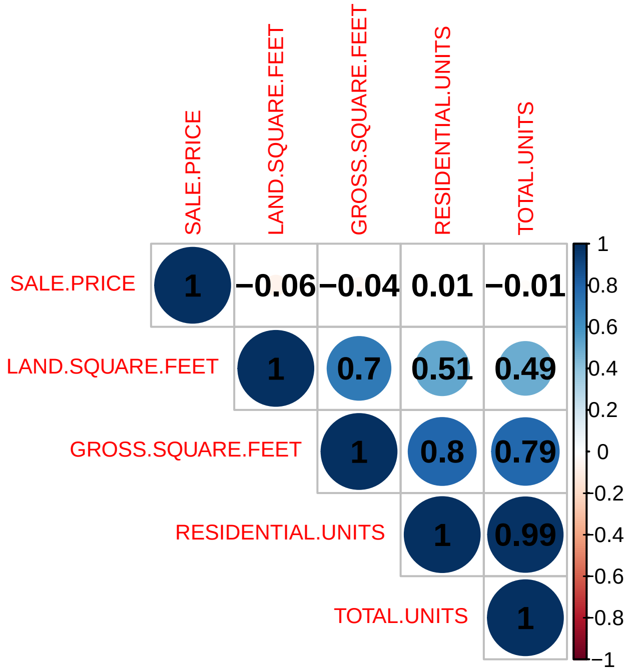
Removing Outliers

Distribution after applying an outlier removal technique, such as the Interquartile Range (IQR) method. The blue bars show a somewhat more balanced distribution, with the most extreme outliers removed, resulting in a narrower range of residential unit counts. There was a clear improvement of distribution after removal of outliers



Correlation plot

The darker blue circles indicate a stronger positive correlation, while the lighter blue and red circles represent weaker or negative correlations. For example, there is a strong positive correlation between gross square feet and land square feet, as well as between residential units and total units. However, sale price has a weak or slightly negative correlation with most of the other variables, suggesting that higher sale prices may not necessarily be associated with larger property sizes or more units.



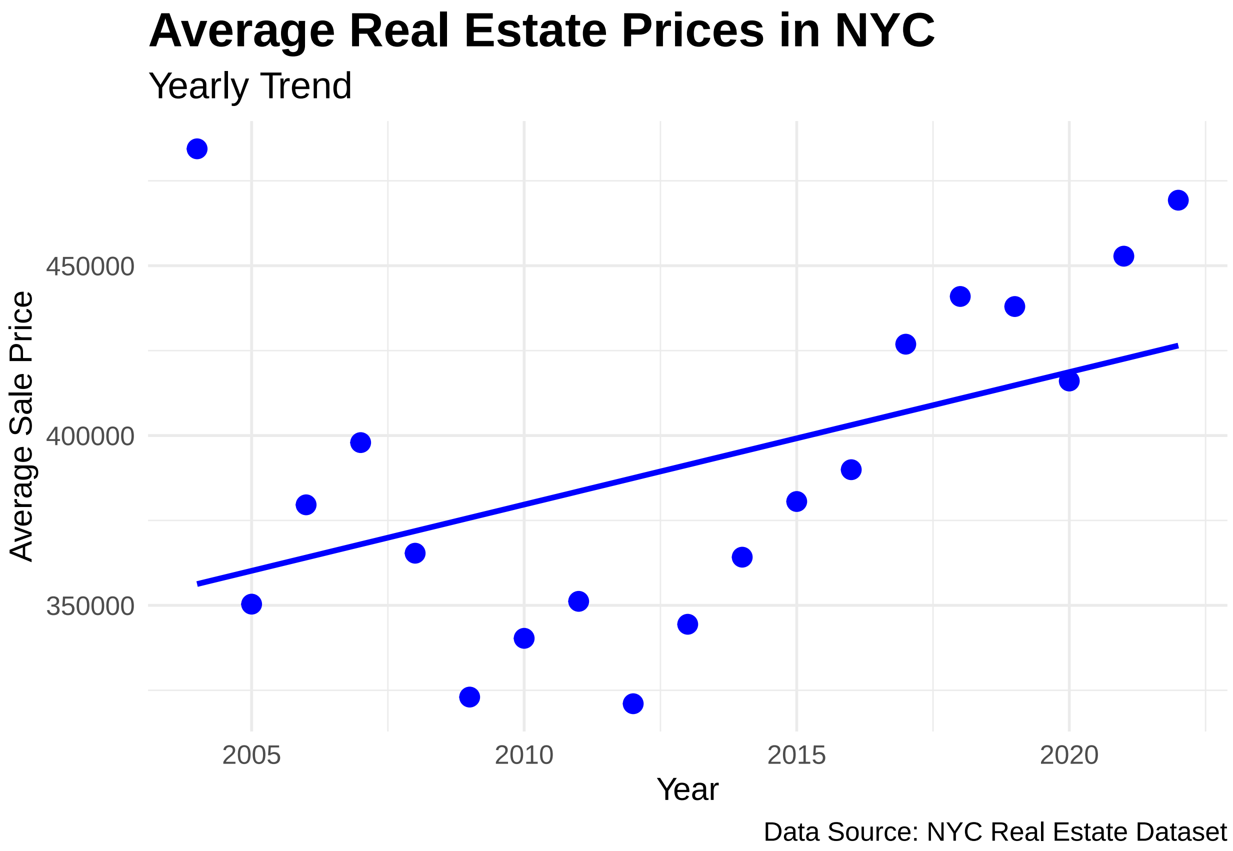
Transforming Categorical Variables to Factors

Since the BUILDING.CLASS.CATEGORY, TAX.CLASS.AT.PRESENT, BUILDING.CLASS.AT.PRESENT, TAX.CLASS.AT.TIME.OF.SALE columns are an ordinal categorical variable we can transform them into a factor.

```
## 'data.frame':    1149281 obs. of  4 variables:
##  $ BUILDING.CLASS.CATEGORY  : Factor w/ 124 levels "                                            ",..
##  $ TAX.CLASS.AT.PRESENT     : Factor w/ 11 levels " ","  ","1","1A",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ BUILDING.CLASS.AT.PRESENT: Factor w/ 129 levels " ","  ","A0",..: 7 12 17 17 16 12 7 12 16 13 ...
##  $ TAX.CLASS.AT.TIME.OF.SALE: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

Real Estate Prices Over Time

Over the years, there is a clear upward trend, indicating that the average real estate prices in NYC have been steadily increasing. The line exhibits a consistent upward slope, with prices rising from around $350,000 in 2005 to over $450,000 by 2020. Although there are some fluctuations in specific years, the overall trajectory demonstrates a significant increase in real estate prices in NYC over the 15-year period depicted in the graph.



Linear Regression Model - Key Factors Influencing Real Estate Prices

The linear regression model suggests that various factors significantly influence real estate prices in New York City. Notably, residential units, tax class, year built, sale date, tax class at time of sale, gross square feet, and land square feet all demonstrate statistically significant relationships with sale prices. However, the model's adjusted R-squared value of 0.06181 indicates that only about 6.164% of the variability in sale prices is explained by these factors. Additionally, the residuals' distribution reveals a considerable spread, indicating potential heteroscedasticity or unaccounted-for factors in the model.

```
##
```

```
## Call:
## lm(formula = SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT +
##     YEAR.BUILT + SALE_DATE + TAX.CLASS.AT.TIME.OF.SALE + GROSS.SQUARE.FEET +
##     LAND.SQUARE.FEET, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -775555 -321274  -42602  227912 1618482
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.318e+05  1.264e+04  18.336  < 2e-16 ***
## RESIDENTIAL.UNITS           4.011e+04  7.895e+02  50.797  < 2e-16 ***
## TAX.CLASS.AT.PRESENT       -2.034e+05  1.181e+04 -17.225  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1      -3.294e+05  1.239e+04 -26.589  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1A     -2.628e+05  1.252e+04 -20.997  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1B     -4.261e+05  1.261e+04 -33.787  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1C     -9.937e+04  1.560e+04  -6.370 1.90e-10 ***
## TAX.CLASS.AT.PRESENT1D     -2.155e+05  2.702e+04  -7.976 1.52e-15 ***
## TAX.CLASS.AT.PRESENT2      -1.828e+05  1.100e+04 -16.610  < 2e-16 ***
## TAX.CLASS.AT.PRESENT2C     -9.346e+04  1.132e+04  -8.259  < 2e-16 ***
## TAX.CLASS.AT.PRESENT3      -4.767e+05  1.904e+05  -2.504   0.0123 *
## TAX.CLASS.AT.PRESENT4      -3.147e+05  1.233e+04 -25.520  < 2e-16 ***
## YEAR.BUILT                  5.086e+00  8.548e-01   5.950 2.68e-09 ***
## SALE_DATE                   1.854e+01  2.098e-01  88.365  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE2  9.617e+04  7.229e+03  13.304  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE3  4.405e+03  1.840e+05   0.024   0.9809
## TAX.CLASS.AT.TIME.OF.SALE4 -1.612e+05  6.608e+03 -24.393  < 2e-16 ***
## GROSS.SQUARE.FEET           1.398e+01  7.652e-01  18.265  < 2e-16 ***
## LAND.SQUARE.FEET            1.936e+01  4.296e-01  45.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367800 on 919407 degrees of freedom
## Multiple R-squared:  0.06181,    Adjusted R-squared:  0.06179
## F-statistic:  3365 on 18 and 919407 DF,  p-value: < 2.2e-16
```

The stepwise regression process, with a starting AIC of 23565539, selected a final model with predictors including residential units, tax class, year built, sale date, tax class at the time of sale, gross square feet, and land square feet. This model, fitted using linear regression, reveals statistically significant relationships between these predictors and sale prices, as indicated by the low p-values and the coefficients' significance levels. However, the adjusted R-squared value remains low at 0.06164, suggesting that this model explains only a small portion of the variability in sale prices.

```
## Start:  AIC=23565539
## SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT + YEAR.BUILT +
##     SALE_DATE + TAX.CLASS.AT.TIME.OF.SALE + GROSS.SQUARE.FEET +
##     LAND.SQUARE.FEET
##
##                             Df  Sum of Sq        RSS      AIC
## <none>                                    1.2439e+17 23565539
## - YEAR.BUILT                 1 4.7898e+12 1.2439e+17 23565572
## - GROSS.SQUARE.FEET          1 4.5133e+13 1.2443e+17 23565870
## - TAX.CLASS.AT.TIME.OF.SALE  3 1.4858e+14 1.2454e+17 23566630
## - LAND.SQUARE.FEET           1 2.7490e+14 1.2466e+17 23567566
## - RESIDENTIAL.UNITS          1 3.4909e+14 1.2474e+17 23568113
## - TAX.CLASS.AT.PRESENT      10 5.2173e+14 1.2491e+17 23569367
## - SALE_DATE                  1 1.0564e+15 1.2544e+17 23573312
```

```
##
## Call:
## lm(formula = SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT +
##     YEAR.BUILT + SALE_DATE + TAX.CLASS.AT.TIME.OF.SALE + GROSS.SQUARE.FEET +
##     LAND.SQUARE.FEET, data = train_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -775555 -321274  -42602  227912 1618482
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.318e+05  1.264e+04  18.336  < 2e-16 ***
## RESIDENTIAL.UNITS           4.011e+04  7.895e+02  50.797  < 2e-16 ***
## TAX.CLASS.AT.PRESENT       -2.034e+05  1.181e+04 -17.225  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1      -3.294e+05  1.239e+04 -26.589  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1A     -2.628e+05  1.252e+04 -20.997  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1B     -4.261e+05  1.261e+04 -33.787  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1C     -9.937e+04  1.560e+04  -6.370 1.90e-10 ***
## TAX.CLASS.AT.PRESENT1D     -2.155e+05  2.702e+04  -7.976 1.52e-15 ***
## TAX.CLASS.AT.PRESENT2      -1.828e+05  1.100e+04 -16.610  < 2e-16 ***
## TAX.CLASS.AT.PRESENT2C     -9.346e+04  1.132e+04  -8.259  < 2e-16 ***
## TAX.CLASS.AT.PRESENT3      -4.767e+05  1.904e+05  -2.504   0.0123 *
## TAX.CLASS.AT.PRESENT4      -3.147e+05  1.233e+04 -25.520  < 2e-16 ***
## YEAR.BUILT                  5.086e+00  8.548e-01   5.950 2.68e-09 ***
## SALE_DATE                   1.854e+01  2.098e-01  88.365  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE2  9.617e+04  7.229e+03  13.304  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE3  4.405e+03  1.840e+05   0.024   0.9809
## TAX.CLASS.AT.TIME.OF.SALE4 -1.612e+05  6.608e+03 -24.393  < 2e-16 ***
## GROSS.SQUARE.FEET           1.398e+01  7.652e-01  18.265  < 2e-16 ***
## LAND.SQUARE.FEET            1.936e+01  4.296e-01  45.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367800 on 919407 degrees of freedom
## Multiple R-squared:  0.06181,    Adjusted R-squared:  0.06179
## F-statistic:  3365 on 18 and 919407 DF,  p-value: < 2.2e-16
```

The next model fitted using generalized linear regression with a Gaussian family and an identity link function maintains predictors including residential units, tax class, year built, sale date, tax class at the time of sale, gross square feet, and land square feet. The coefficients and their significance remain consistent with the previous models. The null and residual deviances provide additional information on the goodness of fit, with the residual deviance being slightly lower than the null deviance, suggesting some level of model improvement.

```
##
## Call:
## glm(formula = SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT +
##     YEAR.BUILT + SALE_DATE + TAX.CLASS.AT.TIME.OF.SALE + GROSS.SQUARE.FEET +
##     LAND.SQUARE.FEET, family = gaussian(link = "identity"), data = train_set)
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.318e+05  1.264e+04  18.336  < 2e-16 ***
## RESIDENTIAL.UNITS           4.011e+04  7.895e+02  50.797  < 2e-16 ***
## TAX.CLASS.AT.PRESENT       -2.034e+05  1.181e+04 -17.225  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1      -3.294e+05  1.239e+04 -26.589  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1A     -2.628e+05  1.252e+04 -20.997  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1B     -4.261e+05  1.261e+04 -33.787  < 2e-16 ***
## TAX.CLASS.AT.PRESENT1C     -9.937e+04  1.560e+04  -6.370 1.90e-10 ***
## TAX.CLASS.AT.PRESENT1D     -2.155e+05  2.702e+04  -7.976 1.52e-15 ***
## TAX.CLASS.AT.PRESENT2      -1.828e+05  1.100e+04 -16.610  < 2e-16 ***
## TAX.CLASS.AT.PRESENT2C     -9.346e+04  1.132e+04  -8.259  < 2e-16 ***
## TAX.CLASS.AT.PRESENT3      -4.767e+05  1.904e+05  -2.504   0.0123 *
## TAX.CLASS.AT.PRESENT4      -3.147e+05  1.233e+04 -25.520  < 2e-16 ***
## YEAR.BUILT                  5.086e+00  8.548e-01   5.950 2.68e-09 ***
## SALE_DATE                   1.854e+01  2.098e-01  88.365  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE2  9.617e+04  7.229e+03  13.304  < 2e-16 ***
## TAX.CLASS.AT.TIME.OF.SALE3  4.405e+03  1.840e+05   0.024   0.9809
## TAX.CLASS.AT.TIME.OF.SALE4 -1.612e+05  6.608e+03 -24.393  < 2e-16 ***
## GROSS.SQUARE.FEET           1.398e+01  7.652e-01  18.265  < 2e-16 ***
## LAND.SQUARE.FEET            1.936e+01  4.296e-01  45.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.3529e+11)
##
##     Null deviance: 1.3258e+17  on 919425  degrees of freedom
## Residual deviance: 1.2439e+17  on 919407  degrees of freedom
## AIC: 26174759
##
## Number of Fisher Scoring iterations: 2
```

The final model, fitted using robust linear regression (rlm), estimates the intercept at $287676.91. Each additional residential unit increases the sale price by $33,621.92. For the tax class at present, each category shows significant negative impacts on the sale price, with Tax Class 1B having the largest effect, reducing the price by $431,324.29. A one-unit increase in the year built is associated with a $8.8062 increase in sale price. Similarly, each day increment in the sale date adds $14.22 to the sale price. Other variables, such as gross square feet and land square feet, also exhibit significant positive effects on the sale price. The residual standard error, indicating the model's accuracy, is $427,900.

As for the coefficients, they do make sense in the real world. For instance, the positive coefficient for RESIDENTIAL.UNITS suggests that properties with more residential units tend to have higher sale prices, all else being equal. This aligns with the expectation that larger properties (with more units) would generally sell for more.The negative coefficients for the different TAX.CLASS.AT.PRESENT variables suggest that the tax class of a property at the time of sale can negatively impact its sale price. This could be because properties in higher tax classes tend to be more expensive to own and maintain, which could lower their appeal to potential buyers.The positive coefficient for YEAR.BUILT indicates that newer properties tend to sell for more than older properties. The GROSS.SQUARE.FEET and LAND.SQUARE.FEET variables also have positive coefficients, suggesting that larger properties are higher in prices, which aligns with general real estate market expectations.

```
##
## Call: rlm(formula = SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT +
##     YEAR.BUILT + SALE_DATE + TAX.CLASS.AT.TIME.OF.SALE + GROSS.SQUARE.FEET +
##     LAND.SQUARE.FEET, data = train_set)
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -689242 -310366  -23897  245241 1632902
##
## Coefficients:
##                             Value        Std. Error   t value
## (Intercept)                  287676.9105   11978.4159      24.0163
## RESIDENTIAL.UNITS             33621.9236     748.1168      44.9421
## TAX.CLASS.AT.PRESENT        -227795.1558   11190.2849     -20.3565
## TAX.CLASS.AT.PRESENT1       -307244.5736   11740.3930     -26.1699
## TAX.CLASS.AT.PRESENT1A      -254717.7267   11860.2448     -21.4766
## TAX.CLASS.AT.PRESENT1B      -431324.2937   11950.6843     -36.0920
## TAX.CLASS.AT.PRESENT1C      -131863.1663   14781.9210      -8.9206
## TAX.CLASS.AT.PRESENT1D      -206206.0243   25599.0244      -8.0552
## TAX.CLASS.AT.PRESENT2       -202761.9840   10426.5495     -19.4467
## TAX.CLASS.AT.PRESENT2C      -105398.3560   10722.8404      -9.8293
## TAX.CLASS.AT.PRESENT3       -485274.8784  180412.3223      -2.6898
## TAX.CLASS.AT.PRESENT4       -313200.2492   11684.0645     -26.8058
## YEAR.BUILT                        8.8062       0.8099      10.8727
## SALE_DATE                        14.2285       0.1988      71.5717
## TAX.CLASS.AT.TIME.OF.SALE2   90266.0791    6849.8345      13.1778
## TAX.CLASS.AT.TIME.OF.SALE3   19981.2049  174334.5489       0.1146
## TAX.CLASS.AT.TIME.OF.SALE4 -166583.2088    6261.5066     -26.6043
## GROSS.SQUARE.FEET                 5.4616       0.7250       7.5328
## LAND.SQUARE.FEET                 18.1803       0.4071      44.6634
##
## Residual standard error: 427900 on 919407 degrees of freedom
```

Model evaluation

The models chosen for this analysis - Linear Regression, Stepwise Regression, Generalized Linear Model (GLM), and Robust Linear Model - each offer unique advantages. The Linear Regression model serves as a simple and interpretable baseline model. The Stepwise Regression model was used to perform automatic feature selection, potentially improving the model by removing irrelevant features. The GLM was chosen to handle potential non-normal distribution of residuals and the Robust Linear Model was used to mitigate the influence of outliers.

To evaluate the performance of each model, we can look at their performance on the test data set. Below is a table that summarizes the R-squared, RSE, and AIC value of each model:

```
##            Model R_Squared_Test RSE_Test AIC_Test
## 1   Linear Model     0.06100972 368458.6 26174759
## 2 Stepwise Model     0.06100972 368458.6 26174759
## 3            GLM     0.06100972 368458.6 26174759
## 4      Robust LM     0.05656965 369328.7 26179144
```

In terms of performance, the linear , stepwise Regression, and GLM models all achieved an R-squared value of approximately 0.061 on the test data, indicating that they explain about 6.1% of the variance in the target variable. The Robust Linear Model had a slightly lower R-squared value of 0.057. The Residual Standard Error (RSE) was similar across all models at 368458.6. Lastly the AIC was the same in the linear , stepwise Regression, and GLM models at 26174759 but higher in the Robust LM at 26179144. Since the results are identical, we can do a residual analysis to see which model's assumptions are best met.

Based on this residual plot(See supplemental figure), there seems to be no clear patterns or trends and the points appear randomly scattered around the horizontal line at zero. This validates the assumption of homoscedasticity and randomness of the residuals, which are key assumptions of linear regression models.

Given these results, the linear Regression, stepwise Regression, or GLM model have similar performance and can be recommended as models to predict property prices in NYC. However, the stepwise model may be preferred due to its simplicity and interpretation. Again, it is worth noting that the relatively low R-squared values suggest there may be other important predictors not included in the models, so further feature engineering or model selection work could potentially improve the results.

Discussion and Conclusion

In our study, we used a comprehensive dataset from the Department of Finance in New York City, aiming to investigate the complex dynamics of property sales over the past two decades. We looked into many variables and applied different linear models to predict property values.

Our exploration revealed that a multitude of factors, including the number of residential units, the tax class, the year of construction, the date of sale, the tax class at the time of sale, the gross square footage, and the land square footage, all wield a statistically significant influence over real estate prices in the bustling metropolis of New York City. However, the adjusted R-squared values of our models suggested that these factors collectively only explained about 6.1% of the variability in sale prices. This finding suggests that there are other influential predictors that are currently not included in our models.

In terms of performance on the test data, the linear model, stepwise model, and GLM all demonstrated similar efficacy, with an R-squared value of approximately 0.061 and a RSE of 368458.6. On the other hand, the robust linear model, exhibited a slightly lower R-squared value of 0.057, suggesting it might be marginally less effective at explaining the variance in the target variable compared to the other models.

One limitation of our analysis is the relatively low R-squared values, these indicate that a significant portion of the variance in the target variable is not explained by the models. This could be due to missing predictors, non-linear relationships, or interactions between predictors that were not accounted for in the models. Another potential limitation is the exclusion of certain economic factors such as mortgage rates at the time of sale, which could have a substantial impact on property prices. In addition, the dataset contained locations of the property sold, an important factor in determining real estate prices. However, we were not able to incorporate this into our linear model since the location data was categorical.

Looking ahead, we recommend exploring other potential predictors and considering the use of other types of models, such as non-linear models, to enhance the predictive power of our analysis. Also, it might be beneficial to collect more data, especially on factors that we suspect might be influencing real estate prices but were not available in our current dataset, such as mortgage rates at the time of sale.

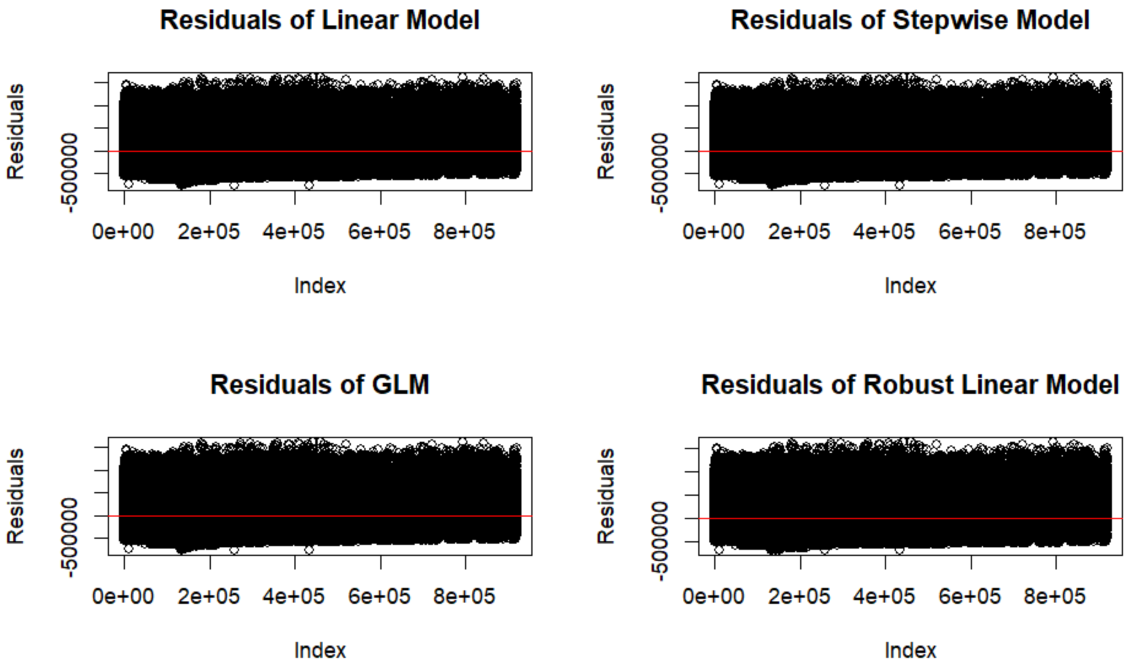
In conclusion, our models provide some insights into the factors influencing real estate prices in New York City. However, the low R-squared values in our model show that these variables represent only a small portion in this complex problem of real estate prices. While our current models have their limitations, they represent a good starting towards a more comprehensive understanding of real estate prices.

References

1. DataScienceDonut. (2024). Current NYC Property Sales. Kaggle. Retrieved April 24, 2024, from https://www.kaggle.com/datasets/datasciencedonut/current-nyc-property-sales
2. Gaynor, M. (2022). Analysis of NYC Property Sales. Medium. Retrieved April 24, 2024, from https://medium.com/@mgaynor228/analysis-of-nyc-property-sales-9af7686aa2ca

Appendix:

Supplemental figure



Residuals of Linear Model

Residuals of Stepwise Model

Residuals of GLM

Residuals of Robust Linear Model

Code

```
## Loading Required libraries
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
library(ggplot2)
library(tidyr)
library(kableExtra)
library(corrplot)
library(skimr)
library(dplyr)
library(Hmisc)
library(reshape2)
library(tidyr)
library(MASS)
library(treemap)
library(randomForest)
library(lubridate)
library(forecast)
library(caret)
library(readxl)

#Clear all
rm(list = ls())
```

```r
options(scipen = 999)
knitr::opts_chunk$set(echo=FALSE,warning = FALSE, message = FALSE)
nyc_data <- read.csv("C:/Users/Jian/Desktop/DATA 621 -Business Analytics and Data Mining/Final Project/

#head(nyc_data)
glimpse(nyc_data)
# Check for missing values
missing_values <- colSums(is.na(nyc_data))

missing_columns <- names(missing_values[missing_values > 0])

# Print the number of missing values for each column
cat("\n********** Number of missing values for each column **********\n")
print(missing_values[missing_values > 0])
clean_nyc <- na.omit(nyc_data)
# Numeric variables
numeric_vars <- c("RESIDENTIAL.UNITS", "COMMERCIAL.UNITS", "TOTAL.UNITS",
                  "LAND.SQUARE.FEET", "GROSS.SQUARE.FEET", "SALE.PRICE")


num_data <-clean_nyc[, numeric_vars]
# Set up the plot layout
par(mfrow = c(2, 3))

for (i in 1:length(names(num_data))){
  hist(num_data[[i]], main=names(num_data)[i], breaks=20, prob=TRUE)
}
# Function to remove outliers based on Tukey's method
remove_outliers <- function(data, variable) {
  q1 <- quantile(data[[variable]], 0.25)
  q3 <- quantile(data[[variable]], 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  filtered_data <- data[data[[variable]] >= lower_bound & data[[variable]] <= upper_bound, ]
  return(filtered_data)
}

# Apply the function to each numeric variable in clean_nyc
for (var in numeric_vars) {
  clean_nyc <- remove_outliers(clean_nyc, var)
}

library(gridExtra)

num_data <- clean_nyc[, numeric_vars]

# Create histograms for each numeric variable
hist_plots <- lapply(numeric_vars, function(var) {
  ggplot(data = num_data, aes(!!sym(var))) +
    geom_histogram(fill = "skyblue", color = "black", bins = 30) +
    labs(title = paste("Histogram of", var),
         x = var,
         y = "Frequency") +
    theme_minimal()
})

# Arrange the plots in a grid
grid.arrange(grobs = hist_plots, ncol = 2)

# Drop COMMERCIAL.UNITS variable
clean_nyc <- clean_nyc[, !names(clean_nyc) %in% "COMMERCIAL.UNITS"]
library(treemap)

# Categorical variables
categorical_vars <- c("NEIGHBORHOOD", "BUILDING.CLASS.CATEGORY",
                      "TAX.CLASS.AT.PRESENT", "BUILDING.CLASS.AT.PRESENT",
                      "TAX.CLASS.AT.TIME.OF.SALE", "BUILDING.CLASS.AT.TIME.OF.SALE")

# Create treemaps for each categorical variable
treemap_plots <- lapply(categorical_vars, function(var) {
  treemap(clean_nyc, index = var, vSize = "SALE.PRICE", title = paste("Treemap of", var))
})

# Output the treemaps
for (plot in treemap_plots) {
  plot
}

#print(sum(any(is.na(clean_nyc))))

num_data <- as.data.frame(num_data)

# Drop COMMERCIAL.UNITS variable
num_data <- num_data[, !names(num_data) %in% "COMMERCIAL.UNITS"]


# Remove observations with missing, NaN, and infinite values
clean_data <- num_data[complete.cases(num_data) & !is.infinite(rowSums(num_data)), ]

# Calculate correlation matrix
correlation_matrix <- cor(clean_data)

# Plot correlation matrix
corrplot(correlation_matrix, method = "circle", type = "upper", order = "hclust",
         addCoef.col = "black", tl.cex = 0.7, cl.cex = 0.7)
cat_vars <- c("BUILDING.CLASS.CATEGORY", "TAX.CLASS.AT.PRESENT", "BUILDING.CLASS.AT.PRESENT", "TAX.CLASS

# Convert categorical variables to factors
for (var in cat_vars) {
  clean_nyc[[var]] <- factor(clean_nyc[[var]])
}

# Verify the transformation
str(clean_nyc[cat_vars])
# Convert SALE_DATE to Date format
```

```r
clean_nyc$SALE_DATE <- as.Date(clean_nyc$SALE.DATE)


# Group data by year and calculate average sale price per year
yearly_prices <- clean_nyc %>%
  mutate(year = lubridate::year(SALE_DATE)) %>%
  group_by(year) %>%
  summarise(avg_price = mean(SALE.PRICE))

# Create a line plot of average sale price over time (yearly)
ggplot(yearly_prices, aes(x = year, y = avg_price)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue", linetype = "solid", size = 1) +  #smoother lin
  geom_point(color = "blue", size = 3) +
  labs(title = "Average Real Estate Prices in NYC",
       subtitle = "Yearly Trend",
       x = "Year",
       y = "Average Sale Price",
       caption = "Data Source: NYC Real Estate Dataset") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 18),
        plot.subtitle = element_text(size = 14),
        plot.caption = element_text(size = 10),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10))
set.seed(123)
trainIndex <- createDataPartition(clean_nyc$SALE.PRICE, p = 0.8, list = FALSE)
train_set <- clean_nyc[trainIndex, ]
test_set  <- clean_nyc[-trainIndex, ]
# regression analysis
lm_model <- lm(SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT + YEAR.BUILT + SALE_DATE + TAX.CLAS
summary(lm_model)
# Perform stepwise regression
stepwise_model <- step(lm_model)

# Summary of the stepwise model
summary(stepwise_model)
# Fit GLM with different error distribution and link function
glm_model <- glm(SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT + YEAR.BUILT + SALE_DATE + TAX.Cl
                 data = train_set,
                 family = gaussian(link = "identity"))
summary(glm_model)
library(MASS)

# Fit robust linear regression model
lm_model_robust <- rlm(SALE.PRICE ~ RESIDENTIAL.UNITS + TAX.CLASS.AT.PRESENT + YEAR.BUILT + SALE_DATE +
summary(lm_model_robust)
# Make predictions using the test set
lm_predictions <- predict(lm_model, newdata = test_set)
stepwise_predictions <- predict(stepwise_model, newdata = test_set)
glm_predictions <- predict(glm_model, newdata = test_set)
robust_predictions <- predict(lm_model_robust, newdata = test_set)
```

```
# Calculate R-squared for each model
r_squared_lm_test <- 1 - sum((test_set$SALE.PRICE - lm_predictions)^2) / sum((test_set$SALE.PRICE - mea
r_squared_stepwise_test <- 1 - sum((test_set$SALE.PRICE - stepwise_predictions)^2) / sum((test_set$SALE
r_squared_glm_test <- 1 - sum((test_set$SALE.PRICE - glm_predictions)^2) / sum((test_set$SALE.PRICE - me
r_squared_robust_test <- 1 - sum((test_set$SALE.PRICE - robust_predictions)^2) / sum((test_set$SALE.PRIC

# Calculate RSE for each model
rse_lm_test <- sqrt(mean((test_set$SALE.PRICE - lm_predictions)^2))
rse_stepwise_test <- sqrt(mean((test_set$SALE.PRICE - stepwise_predictions)^2))
rse_glm_test <- sqrt(mean((test_set$SALE.PRICE - glm_predictions)^2))
rse_robust_test <- sqrt(mean((test_set$SALE.PRICE - robust_predictions)^2))

# Calculate AIC for each model
aic_lm <- AIC(lm_model)
aic_stepwise <- AIC(stepwise_model)
aic_glm <- AIC(glm_model)
aic_robust <- AIC(lm_model_robust)
# Create a data frame
model_performance_test <- data.frame(
  Model = c("Linear Model", "Stepwise Model", "GLM", "Robust LM"),
  R_Squared_Test = c(r_squared_lm_test, r_squared_stepwise_test, r_squared_glm_test, r_squared_robust_te
  RSE_Test = c(rse_lm_test, rse_stepwise_test, rse_glm_test, rse_robust_test),
  AIC_Test = c(aic_lm, aic_stepwise, aic_glm, aic_robust)
)

# Print the data frame
print(model_performance_test)

# Calculate residuals for each model
lm_residuals <- residuals(lm_model)
stepwise_residuals <- residuals(stepwise_model)
glm_residuals <- residuals(glm_model)
robust_residuals <- residuals(lm_model_robust)

# Create a 2x2 plot layout
par(mfrow = c(2, 2))

# Plot residuals for Linear Model
plot(lm_residuals, main = "Residuals of Linear Model", ylab = "Residuals")
abline(h = 0, col = "red")

# Plot residuals for Stepwise Model
plot(stepwise_residuals, main = "Residuals of Stepwise Model", ylab = "Residuals")
abline(h = 0, col = "red")

# Plot residuals for GLM
plot(glm_residuals, main = "Residuals of GLM", ylab = "Residuals")
abline(h = 0, col = "red")

# Plot residuals for Robust Linear Model
plot(robust_residuals, main = "Residuals of Robust Linear Model", ylab = "Residuals")
abline(h = 0, col = "red")
knitr::include_graphics("D:/Documents/GitHub/DATA-621-Group-2/Final Project - Housing Prices/resdiual a
```