# Predicting Property Prices in New York City

Group 2 - Fredrick Jones, Jian Quan Chen, Tilon Bobb

# 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.

# Introduction

- The New York real estate market is one of the most dynamic and influential sectors of urban development
- **Motivation**: Identify the factors that influence NYC real estate prices
- **Goal**: Create a model that predicts NYC real estate prices
- **Data Source**: New York City's (NYC) Department of Finance Property Sales
    - Contains real estate transactions from the five boroughs of New York
    - Contains 20 years of data from 2003 to 2023
- The dataset we are analyzing contains 21 columns and 1,603,826 rows
- The response variable will be the Sale Price of the property

# Glimpse of the Data

```
Rows: 1,603,826
Columns: 21
$ BOROUGH                       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ NEIGHBORHOOD                  <chr> "ALPHABET CITY", "ALPHABET CITY", "ALPHABET CITY", "ALPHABET CITY", "ALPHABET CITY", "ALPHABET CITY", "ALPHABET CITY", …
$ BUILDING.CLASS.CATEGORY       <chr> "01 ONE FAMILY DWELLINGS", "02 TWO FAMILY DWELLINGS", "07 RENTALS - WALKUP APARTMENTS", "07 RENTALS - WALKUP APARTMENTS…
$ TAX.CLASS.AT.PRESENT          <chr> "1", "1", "2B", "2B", "2", "2A", "2A", "2B", "2B", "2B", "2", "2", "2", "2", "2B", "2", "2", "2", "2", "2", "2", "2", "…
$ BLOCK                         <int> 374, 377, 373, 373, 376, 377, 377, 379, 389, 389, 390, 390, 392, 392, 393, 396, 396, 399, 400, 400, 401, 402, 402, 402,…
$ LOT                           <int> 46, 1, 16, 17, 54, 52, 52, 25, 45, 47, 50, 57, 33, 37, 37, 26, 27, 47, 39, 50, 54, 25, 55, 56, 33, 34, 1, 23, 46, 49, 4…
$ EASE.MENT                     <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "",…
$ BUILDING.CLASS.AT.PRESENT     <chr> "A4", "S2", "C1", "C1", "C4", "C2", "C2", "C7", "C4", "C4", "C1", "C7", "C7", "C7", "C7", "C4", "C4", "C7", "C7", "C7",…
$ ADDRESS                       <chr> "347 EAST 4TH STREET", "110 AVENUE C", "326 EAST 4TH STREET", "328 EAST 4TH STREET", "719 EAST SIXTH STREET, 1B", "271 …
$ APARTMENT.NUMBER              <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "",…
$ ZIP.CODE                      <dbl> 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, 10009, …
$ RESIDENTIAL.UNITS             <dbl> 1, 2, 10, 10, 20, 5, 5, 7, 10, 10, 29, 10, 17, 25, 8, 20, 20, 32, 30, 28, 18, 14, 10, 10, 4, 8, 76, 32, NA, NA, NA, NA,…
$ COMMERCIAL.UNITS              <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 5, 2, 0, 0, 3, 2, 1, 0, 0, 0, 0, 1, 2, 8, 1, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
$ TOTAL.UNITS                   <dbl> 1, 3, 10, 10, 20, 5, 5, 8, 10, 10, 29, 11, 18, 30, 10, 20, 20, 35, 32, 29, 18, 14, 10, 10, 5, 10, 84, 33, NA, NA, NA, N…
$ LAND.SQUARE.FEET              <dbl> 2116, 1502, 2204, 2204, 2302, 2168, 2169, 4337, 2271, 2248, 3900, 1600, 1909, 3839, 1920, 2778, 2831, 7460, 4807, 4592,…
$ GROSS.SQUARE.FEET             <dbl> 4400, 2790, 8625, 8625, 9750, 3728, 3728, 14347, 6985, 6985, 18099, 5320, 8441, 18836, 7485, 10740, 10900, 19600, 24506…
$ YEAR.BUILT                    <dbl> 1900, 1901, 1899, 1900, 1900, 1900, 1900, 1920, 1900, 1910, 1900, 1910, 1900, 1900, 1920, 1920, 1920, 1900, 1920, 1900,…
$ TAX.CLASS.AT.TIME.OF.SALE     <int> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 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", "C7", "C4", "C4", "C1", "C7", "C7", "C7", "C7", "C4", "C4", "C7", "C7", "C7",…
$ SALE.PRICE                    <dbl> 399000, 2999999, 16800000, 16800000, 158822, 0, 0, 1635000, 5250000, 5650000, 235251, 10, 0, 11600000, 21919, 11800000,…
$ SALE.DATE                     <chr> "2022-09-29 00:00:00", "2022-09-15 00:00:00", "2022-08-04 00:00:00", "2022-08-04 00:00:00", "2022-09-27 00:00:00", "202…
```

# Methodology

Exploratory Data Analysis (EDA) & Data Preparation

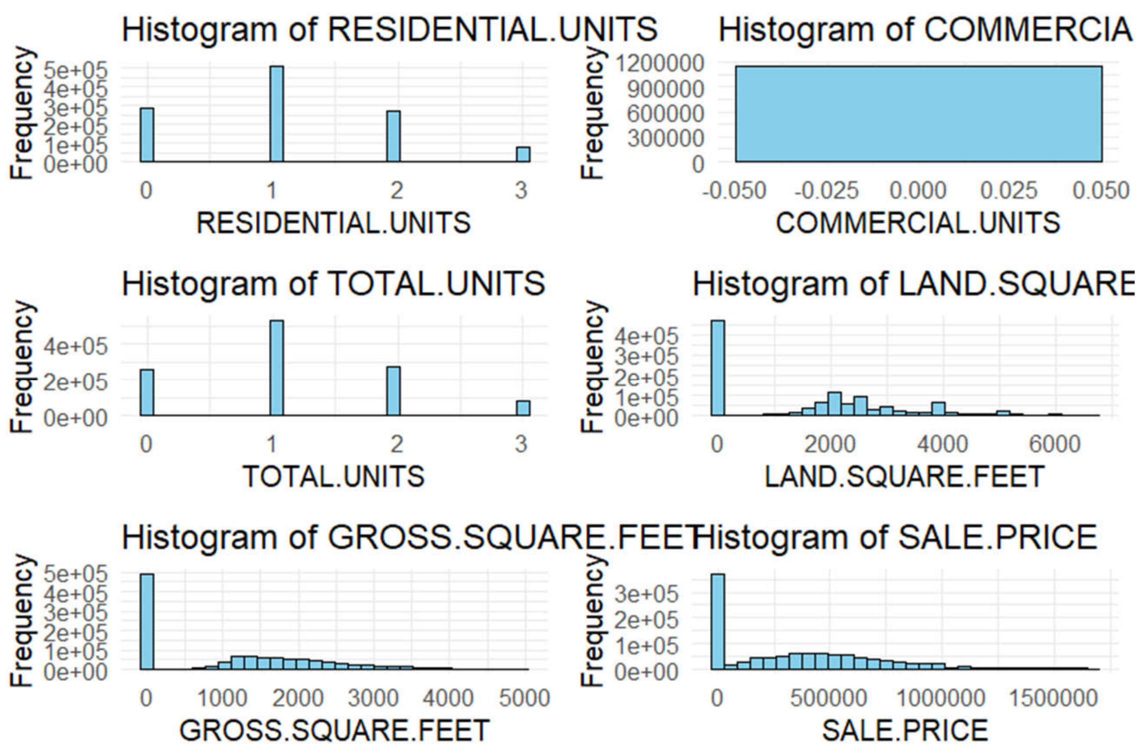
- Identify columns with missing values
  - There were 9 columns with missing values
  - The missing values represented less than 5% of the total observations
  - Dropped ALL missing values
- Distribution Plots of the numeric variables
- Identifying and removing outliers using the Interquartile range method
- Correlation Analysis
- Property values over time

# Methodology

Regression Modeling

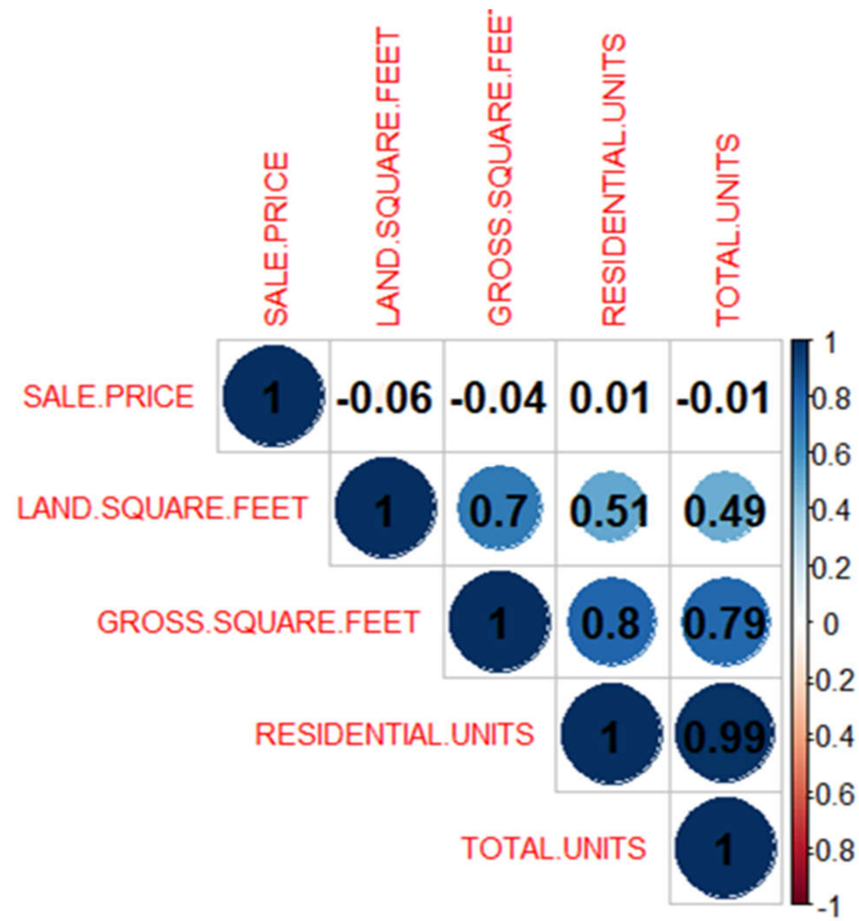
- Use regression modeling to forecast real estate prices in New York City
- Build linear regression models using different predictor variables
- Predictor variables were selected based on their impact on the real estate prices
- Linear models were chosen here because they provide a simple and interpretable relationship between the features and the continuous target variable
- Used a train/test (80/20) split strategy to train and evaluate our linear model
- Performance measured by Mean Squared Error (MSE),R-squared value, Akaike information criterion (AIC)

# Distribution Plot

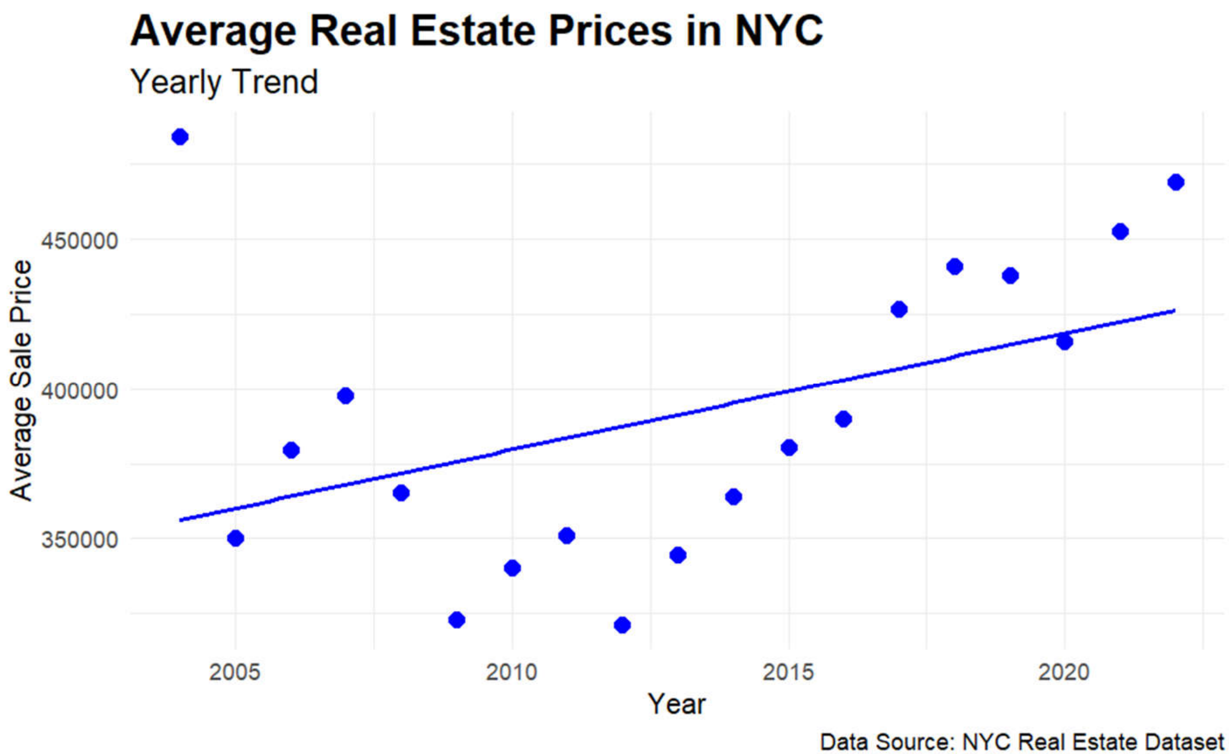


# Correlation Plot

- There is a strong positive correlation between gross square feet and land square feet, as well as between residential units and total units.
- 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.



# Real Estate Prices Over Time



# Experimentation and Results

# Linear 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 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

# Stepwise Model

```
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
```

- AIC of 23565539
- Stepwise regression process 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 reveals statistically significant relationships between these predictors and sale prices, as indicated by the low p-values and the coefficients' significance levels
- The adjusted R-squared value remains low at 0.06164

# Generalized Linear Model (GLM)

```
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
```

- Fit GLM with different error distribution and link function
- 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

# Robust Linear Model (RLM)

```
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
```

- 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
- Each day increment in the sale date adds $14.22 to the sale price
- Gross square feet and land square feet, also exhibit significant positive effects on the sale price

# Robust Linear Model (RLM)

```
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
```

Do the coefficients make sense?

- Yes, the positive coefficient for RESIDENTIAL.UNITS suggests that properties with more residential units tend to have higher sale prices
- Yes, 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
- Yes, the positive coefficient for YEAR.BUILT indicates that newer properties tend to sell for more than older properties
- Yes 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

# Model Evaluation

| Model | R Squared | RSE | AIC |
|---|---|---|---|
| Linear Model | 0.06100972 | 368458.6 | 26174759 |
| Stepwise Model | 0.06100972 | 368458.6 | 26174759 |
| GLM | 0.06100972 | 368458.6 | 26174759 |
| RLM | 0.0565965 | 369328.7 | 26179144 |

- In terms of performance, the linear, stepwise Regression, and GLM models all achieved an R-squared value of approximately 0.061, a RSE of 368458.6, and an AIC of 26174759
- The RLM performed worse of the linear models with a R-squared of 0.0566, a RSE of 369328.7, and an AIC of 26179144
- Given these results, the linear Regression, stepwise Regression, or GLM model can be recommended as models to predict property prices in NYC

# Discussion and Conclusion

- 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

# Discussion and Conclusion

- 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
- Limitations:
    - Relatively low R-squared values
    - Exclusion of certain economic factors such as mortgage rates
    - Did not include location data as it 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
- Our models provide some insights into the factors influencing real estate prices in New York City.
- 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


- DataScienceDonut. (2024). Current NYC Property Sales. Kaggle. Retrieved April 24, 2024, from https://www.kaggle.com/datasets/datasciencedonut/current-nyc-property-sales

- Gaynor, M. (2022). Analysis of NYC Property Sales. Medium. Retrieved April 24, 2024, from https://medium.com/@mgaynor228/analysis-of-nyc-property-sales-9af7686aa2ca