Izaak King

Professor Eleish

Data Science

November 4, 2025

"Assignment 5"

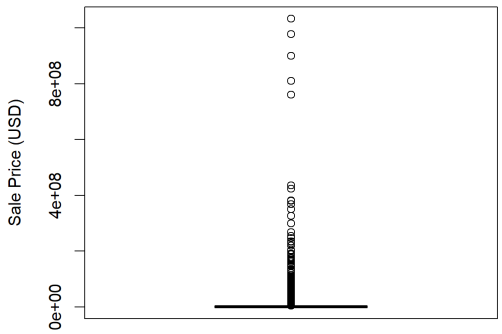
1. For this question I will perform my analysis on a derived dataset from the "Manhattan" borough.

    a. For this data set I plan to look for trends or patterns in how sales prices vary across attributes of the house data like both the land and gross square footage. I will also explore how sales date affects the sales price, and how when the house was built affects the sales prices. I will start with exploratory data analysis, using histograms and boxplots to look at sales price and its distribution. Then, I will use scatterplots to visualize correlations between the mentioned data attributes. Then, I will move to using regression models to predict Sale Price based on these attributes. I will also ensure the data is cleaned properly, with outliers and missing data handled gracefully.

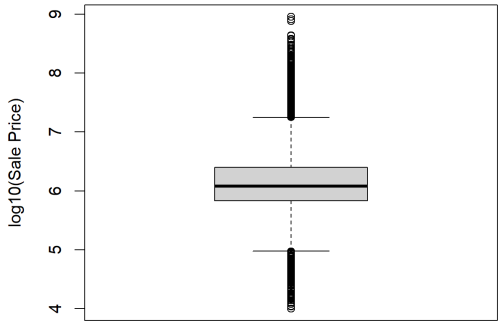
    b. To begin exploratory data analysis I began by looking at the Sale Price attribute. Particularly, I looked at the data manually and saw clear outlier data in prices < 10,000, with some prices being 0, or only a couple hundred. I also saw some outliers on the other end, with very large Sale Price > 900,000,000, and no other attributes filled out. Boxplots were created before and after outliers were removed

and log cleaning was applied, to demonstrate the outliers relative to the other data points.

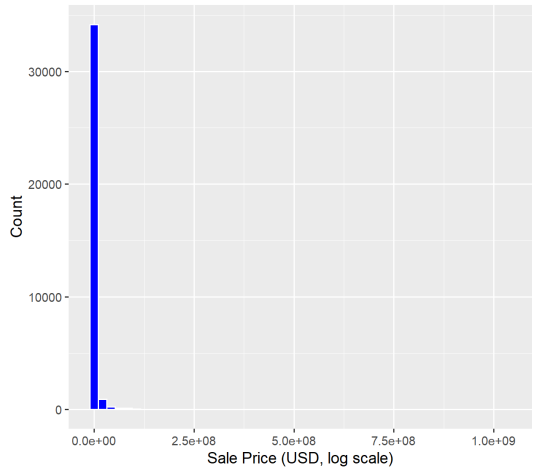
I. Boxplot of Sale Prices (Before Cleaning):



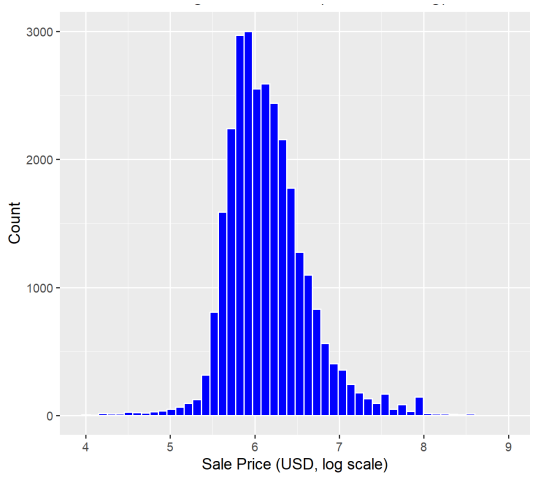

II. Boxplot of Log10 Sale Price (After Cleaning):




III. Distribution of Manhattan Sale Prices (Before cleaning):



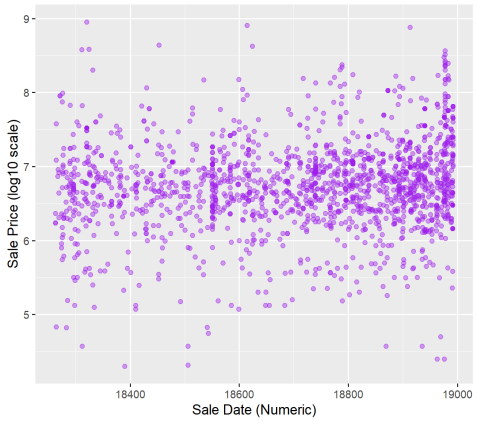
IV. Distribution of Log10 Sale Price (After Cleaning):



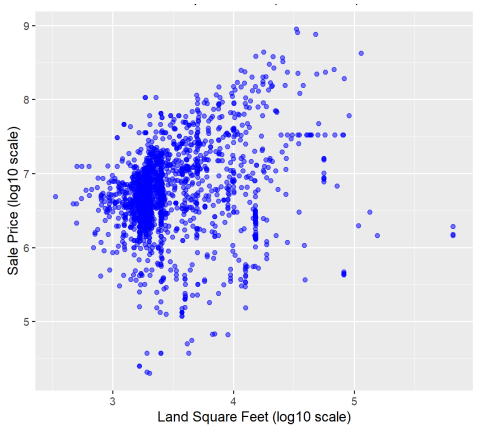

To build onto this data exploration, I created scatterplots to look at the distribution of Sale Price vs Sale Date, Sale Price vs Gross Square Feet, Sale Price vs Land Square Feet, and Sale Price vs Year Built. For cleaning, incomplete data points with NA values were first removed from our dataset. Next, I filtered the Gross and Land Square Feet attributes, ensuring they were all stored as numeric values, rather than strings with commas. I also performed this technique on the sale dates, so we can have numeric values, rather than dates, for analysis. With this, I was able to create basic plots to identify outliers in our data, removing points with Year Built < 1850,  Land Square Feet not within (0, 1,000,000), Gross Square Feet not within (0, 1,000,000) . Next I applied the same log10 cleaning we did to the Sale Price to the square feet, to combat the large range of values. Overall this gives clean data which we can further analyze with clean visualizations. Upon creating scatterplots, there appears to be a general correlation between Sale Price vs Land Square

Feet as well as Sale Price vs Gross Square Feet, but the other distributions look pretty spread. Further analysis will determine if any of these attributes can be used as a good predictor for sale price.

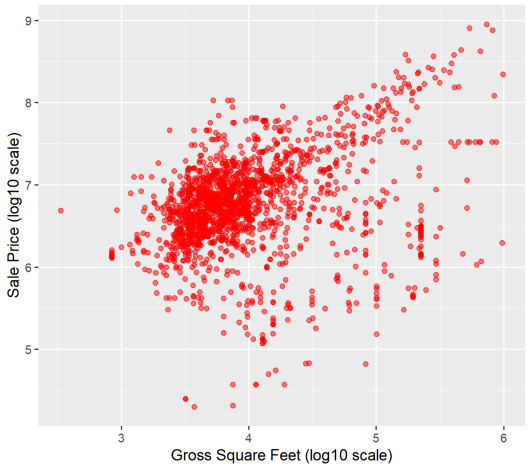
I. Sale Price (Log10) vs. Sale Date(Numeric):



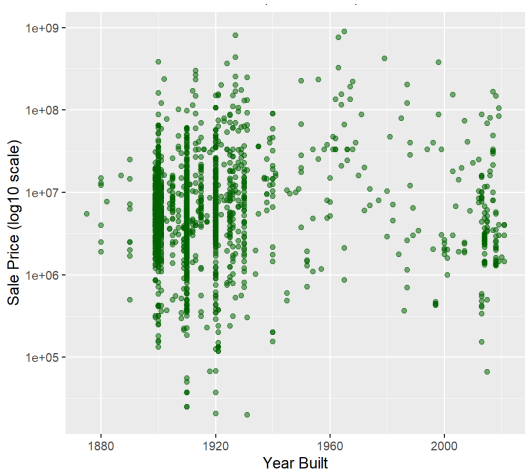

II. Sale Price (Log10) vs Land Square Feet (Log10):



III. Sale Price (Log10) vs Gross Square Feet (Log10):




IV. Sale Price (Log10) vs Year Built:




c. Now, I will conduct regression analysis on the Manhattan Borough to predict the Sale Price using our other attributes. I will be using the cleaned data, with log10 for the sale prices and square footages. As mentioned, I also removed major outlier data points removed, including points with attributes Sale Price not within (10,000, 900,000,000), Year Built < 1850, Land Square not within (0, 1,000,000), or Gross Square Feet not within (0,

1,000,000), removed. The first regression models I will analyze will be simply one variables affect on Sale Price. Then, I will move to a more complex approach, using multiple variables and how they together predict the Sale Price. Summaries will be included for analysis of performance.

I. Sale Date (Numeric) as a predictor for Sale Price (Log10):

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.921e-01  1.215e+00   0.652    0.515
SALE.DATE.NUM 3.191e-04  6.496e-05   4.913 9.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5993 on 1737 degrees of freedom
Multiple R-squared:  0.01371,   Adjusted R-squared:  0.01314
F-statistic: 24.14 on 1 and 1737 DF,  p-value: 9.793e-07
```

II. Land Square Feet (Log10) as a predictor for Sale Price (Log10):

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.41114    0.13031   41.52   <2e-16 ***
LOG.LAND.SQUARE.FEET 0.38636    0.03702   10.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5854 on 1737 degrees of freedom
Multiple R-squared:  0.059,     Adjusted R-squared:  0.05846
F-statistic: 108.9 on 1 and 1737 DF,  p-value: < 2.2e-16
```

III. Gross Square Feet (Log10) as a predictor for Sale Price (Log10):

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.44595    0.10086   54.00   <2e-16 ***
LOG.GROSS.SQUARE.FEET 0.32799    0.02488   13.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5754 on 1737 degrees of freedom
Multiple R-squared:  0.09097,   Adjusted R-squared:  0.09045
F-statistic: 173.8 on 1 and 1737 DF,  p-value: < 2.2e-16
```

IV. Year Built as a predictor for Sale Price (Log10):

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.5601765  0.9112121   6.102 1.29e-09 ***
YEAR.BUILT  0.0006264  0.0004744   1.320    0.187
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6032 on 1737 degrees of freedom
Multiple R-squared:  0.001003,  Adjusted R-squared:  0.0004276
F-statistic: 1.743 on 1 and 1737 DF,  p-value: 0.1869
```

V. Land Square Feet (Log10) and Gross Square Feet (Log10) as predictors for Sale Price (Log10):

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.4453680  0.1281906  42.479  < 2e-16 ***
LOG.LAND.SQUARE.FEET  0.0004491  0.0613528   0.007    0.994
LOG.GROSS.SQUARE.FEET 0.3277444  0.0419459   7.813 9.56e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5755 on 1736 degrees of freedom
Multiple R-squared:  0.09097,   Adjusted R-squared:  0.08993
F-statistic: 86.87 on 2 and 1736 DF,  p-value: < 2.2e-16
```

VI. Sale Date (Numeric) and Year Built as predictors for Sale Price (Log10):

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.377e-01  1.506e+00  -0.224    0.823
SALE.DATE.NUM  3.181e-04  6.495e-05   4.898 1.06e-06 ***
YEAR.BUILT     5.981e-04  4.713e-04   1.269    0.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5992 on 1736 degrees of freedom
Multiple R-squared:  0.01462,   Adjusted R-squared:  0.01349
F-statistic: 12.88 on 2 and 1736 DF,  p-value: 2.801e-06
```

VII. Land Square Feet (Log10), Gross Square Feet (Log10), Sale Date (Numeric) and Year Built as predictors for Sale Price (Log10):

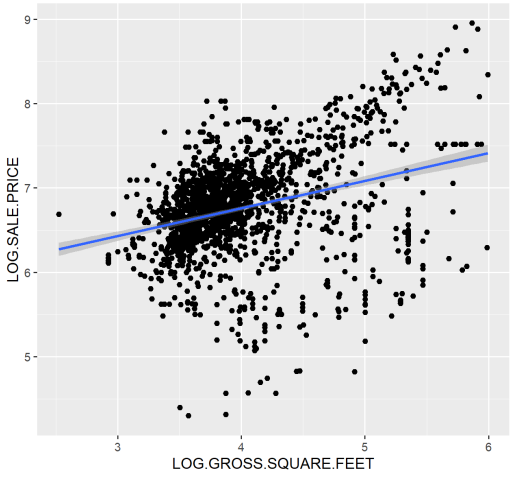
```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.514e+00  1.494e+00   1.683 0.092513 .
LOG.LAND.SQUARE.FEET   7.049e-02  6.636e-02   1.062 0.288278
LOG.GROSS.SQUARE.FEET  3.324e-01  4.183e-02   7.946 3.43e-15 ***
SALE.DATE.NUM          3.467e-04  6.189e-05   5.602 2.46e-08 ***
YEAR.BUILT            -1.989e-03  5.215e-04  -3.813 0.000142 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5683 on 1734 degrees of freedom
Multiple R-squared:  0.1147,    Adjusted R-squared:  0.1126
F-statistic: 56.15 on 4 and 1734 DF,  p-value: < 2.2e-16
```

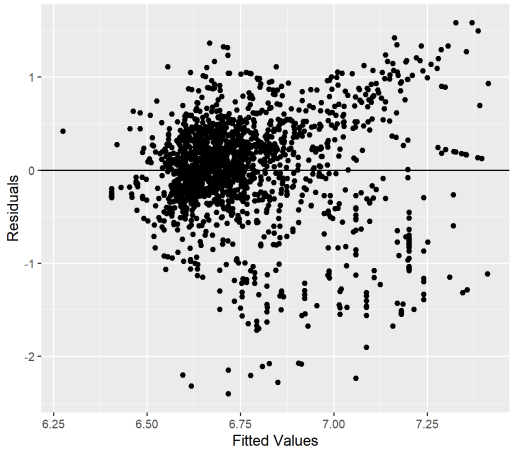
Based on these summaries, our best predictor variable alone is Gross Square Feet. Our best overall performing model uses all 4 attributes, Land Square Feet (Log10), Gross Square Feet (Log10), Sale Date (Numeric) and Year Built as predictors for Sale Price (Log10). I will create plots for these two models.

I. Gross Square Feet (Log10) predicting Sale Price (Log10):

Scatterplot with regression line:

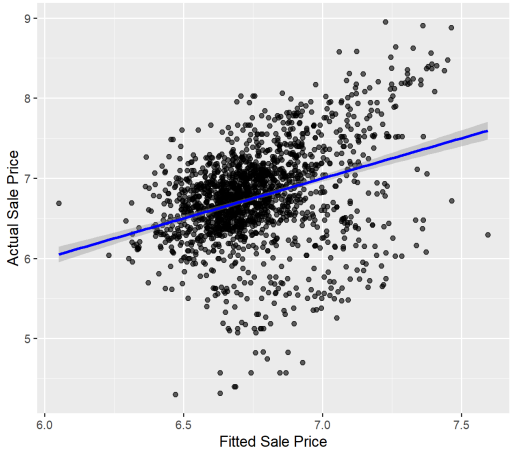



Residual vs Fitted Values:

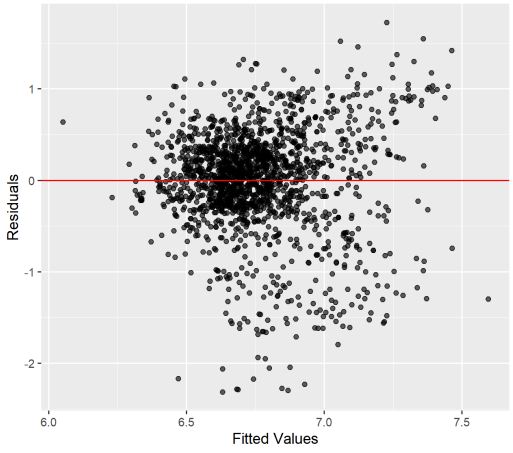


II. Land Square Feet (Log10), Gross Square Feet (Log10), Sale Date (Numeric) and Year Built as predictors for Sale Price (Log10):

Predicted vs Actual Sale Price:




Residual vs Fitted Values:



d. Now I will train and evaluate supervised learning models with the data. I will continue to use the same cleaned dataset, with log cleaning and general outliers removed. For the models, I will train k-NN and Random Forest models based on the quantitative variables of sales price, and square footage attributes. It's important to note I will use k = 38, the number of neighborhoods for our dataset. To evaluate these models, I will analyze the confusion matrix, calculating the accuracy. This will give an estimate of which models perform best.

I. k-NN model based on Sale Price (Log10):

```
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 366"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.21"
```

II. k-NN model based on Square Feet (Log10) and Gross Square Feet (Log10):

```
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 426"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.245"
```

III. k-NN model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10):

```
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 520"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.299"
```

IV. Random Forest model based on Sale Price (Log10):

```
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 1196"
> print(paste("RF Model 0 Accuracy:", round(accuracy, 3)))
[1] "RF Model 0 Accuracy: 0.688"
```

V. Random Forest model based on Square Feet (Log10) and Gross Square Feet (Log10):

```
> print(paste("Amount Correct:", correct1))
[1] "Amount Correct: 1720"
> print(paste("RF Model 1 Accuracy:", round(accuracy1, 3)))
[1] "RF Model 1 Accuracy: 0.989"
```

VI. Random Forest model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10):

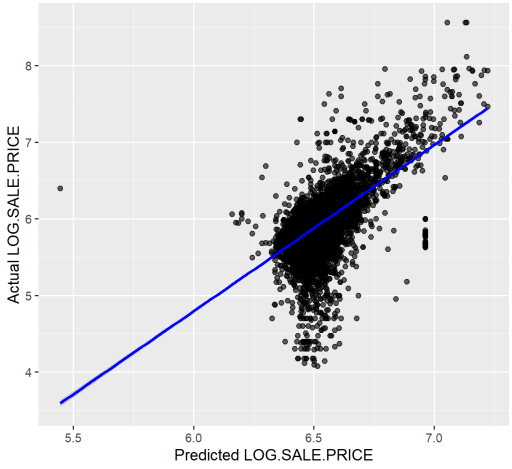
```
> print(paste("Amount Correct:", correct2))
[1] "Amount Correct: 1738"
> print(paste("RF Model 2 Accuracy:", round(accuracy2, 3)))
[1] "RF Model 2 Accuracy: 0.999"
```

Thus, clearly based on these accuracy reports, the Random Forest models performed much better than the k-NN models. The most accurate model, based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10), was incredibly accurate, getting only 1/1739 data points incorrect. It appears that area (Square Feet (Log10) and Gross Square Feet (Log10)) were the most significant quantitative variables in this, as their Random Forest also yielded a very accurate 98.9%.

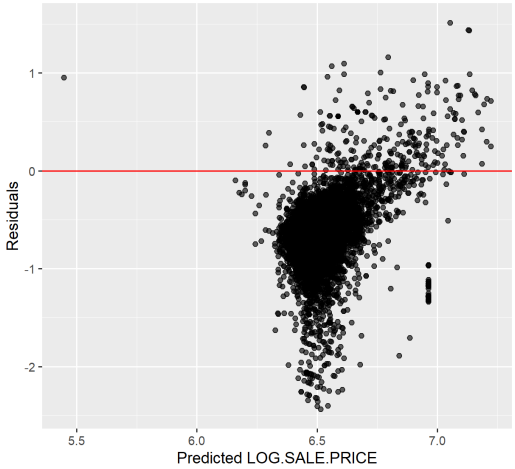
2. For this question I will perform my analysis on another derived dataset, this from the "Queens" borough.
    a. I will start by applying the two best performing regression models from 1c to predict the sale price of data, this time in the new borough. Thus, we will use the apply models using predictor variable Gross Square Feet alone and using predictor variables of all 4 attributes, Land Square Feet (Log10), Gross Square Feet (Log10), Sale Date (Numeric) and Year Built. We will plot the predictions and residuals to see how well our models generalize to the new dataset.

        I. Gross Square Feet (Log10) as a predictor for Sale Price (Log10):

Predicted vs Actual Sale Price:

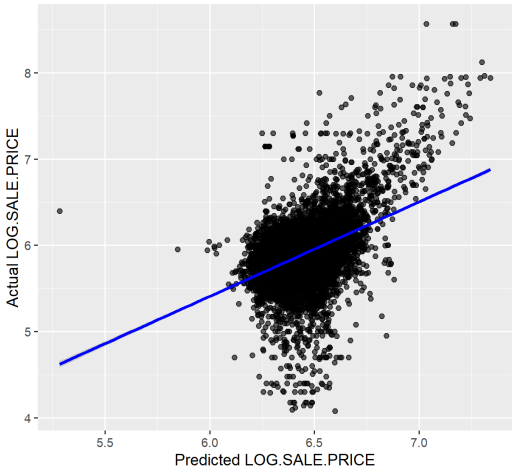


Residual vs Fitted Values:

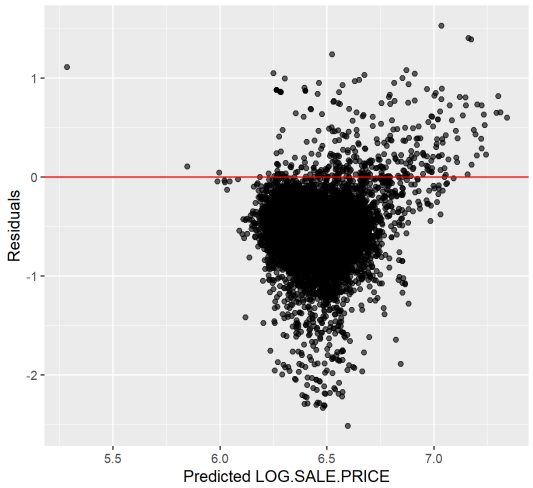



II. Land Square Feet (Log10), Gross Square Feet (Log10), Sale Date (Numeric) and Year Built as predictors for Sale Price (Log10):

Predicted vs Actual Sale Price:



Residual vs Fitted Values:




After applying the Manhattan-trained models to the Queens dataset, the predicted versus actual plots consistently show that most points lie below the prediction line, indicating that the models overestimate the prices of Queens properties. The residual plots confirm this, with residuals mostly negative. Overall, this shows that these models do not generalize well to Queens. One reason could be that sale prices in Queens are typically lower than those in Manhattan. Additionally, differences in property sizes, year built, and sale dates may also affect the models' performance.

b. Next, I will apply the classification models from 1d to the data from the Queens borough. We will use the best performing k-NN model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10), as well as the best performing Random Forest model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10). We will compare the confusion matrices and see how these models' accuracy compares to it's performance on the Manhattan Borough.

I. k-NN model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10):

```
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 193"
> print(paste("KNN Model 2 Accuracy on Queens:", round(accuracy, 3)))
[1] "KNN Model 2 Accuracy on Queens: 0.01"
```

II. Random Forest model based on Sale Price (Log10), Square Feet (Log10) and Gross Square Feet (Log10):

```
> print(paste("Amount Correct:", correctRF))
[1] "Amount Correct: 282"
> print(paste("RF Model 2 Accuracy on Queens:", round(accuracyRF, 3)))
[1] "RF Model 2 Accuracy on Queens: 0.015"
```

Similarly to the regression models, the classification models did not generalize well to the new dataset for the Queens borough. The best kNN model had of about 29.9%, meanwhile when applied to the Queens data, it has accuracy of about 1%. The Random Forest, which performed remarkably well with 99.9% accuracy only had an accuracy of 1.5%. This

poor performance is likely because the models were trained on the Manhattan neighborhood labels, which have different qualities and sizes than those in Queens. Thus, the models could not be used to correctly classify the different neighborhood in the Queens borough. Furthermore, differences in sale prices and property sizes between the boroughs made it harder for the models to generalize.

c. Overall, my observations indicate clear trends in the data, particularly that both Gross Square Feet and Land Square Feet strongly affect sale price. However, these trends appear to be specific to each borough, as models that performed very well for Manhattan were not useful for predicting sales in Queens.