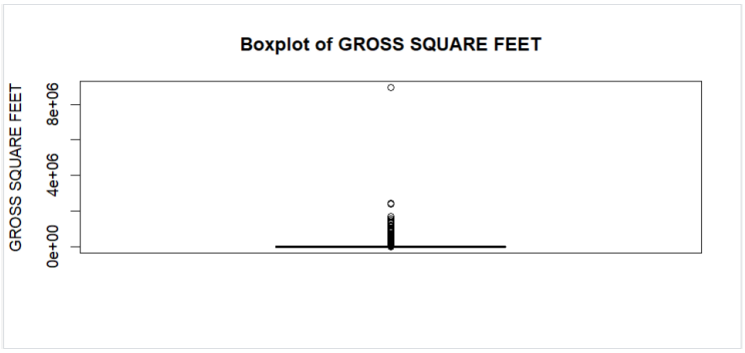
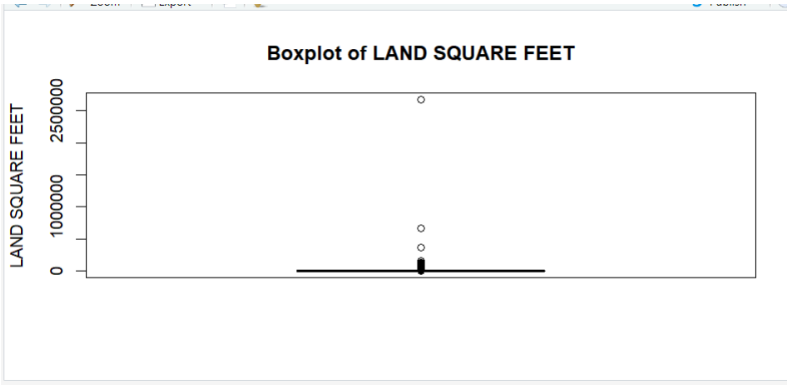
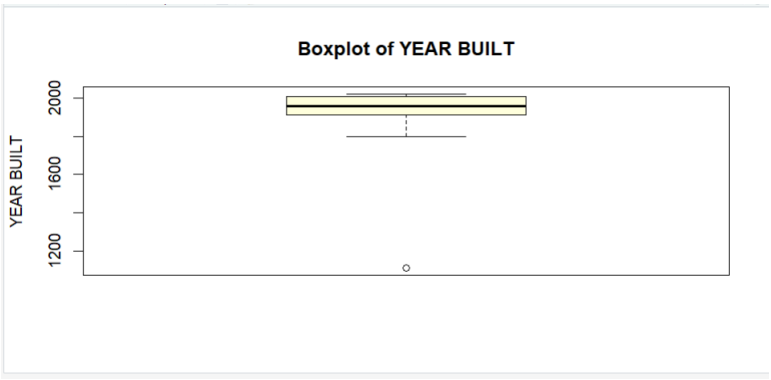
NOTE: Written components can be found in the code as comments tagged by "NOTE."

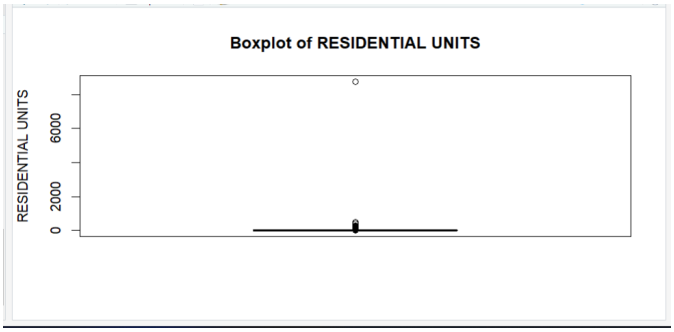
Problem 1b

Initial Boxplots

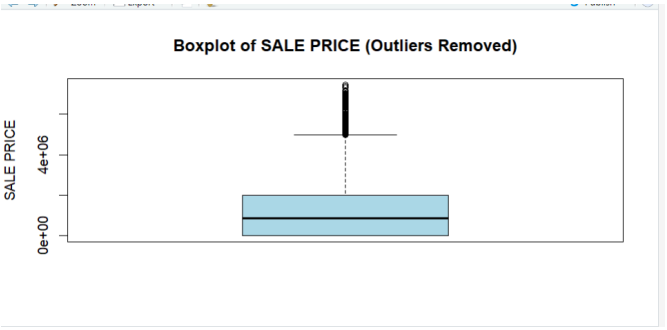


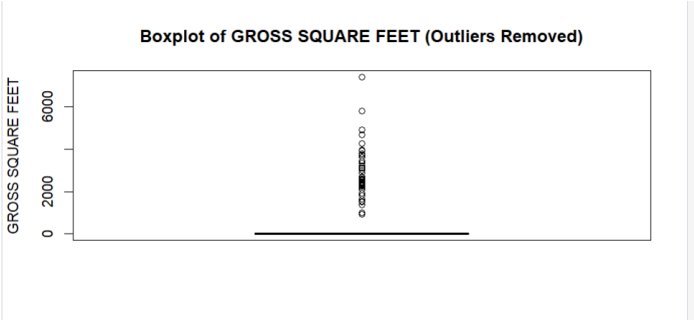


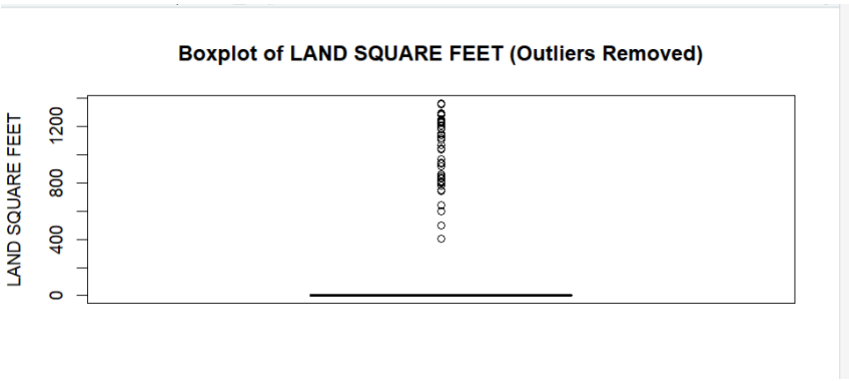


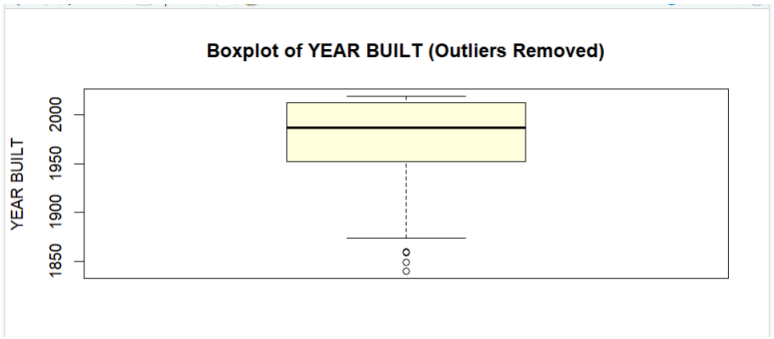


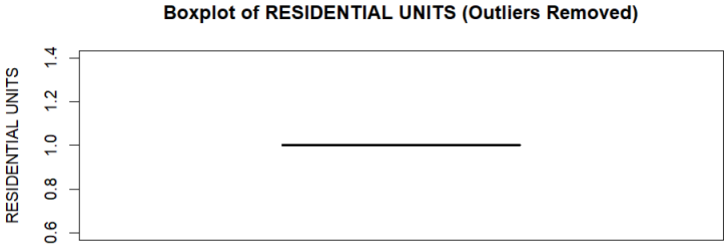
Boxplots (No Outliers)



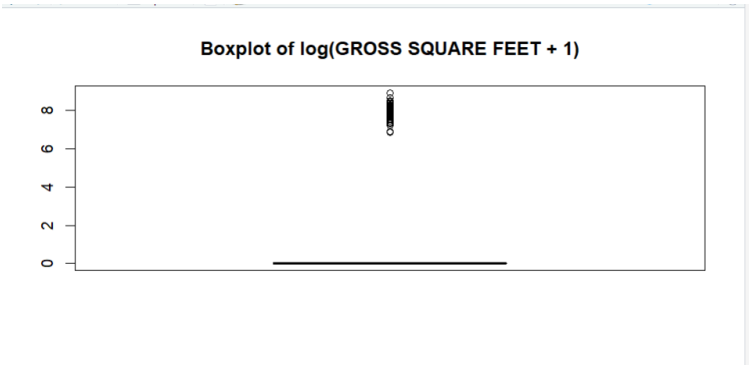


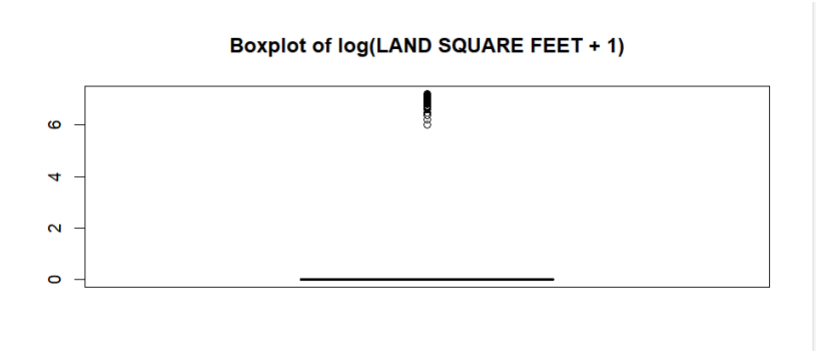


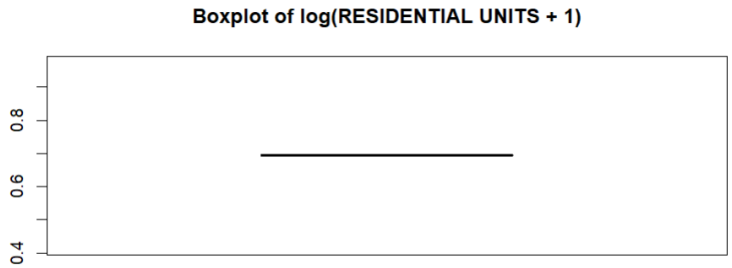




Boxplots (No Outliers + Logistically Transformed Features)



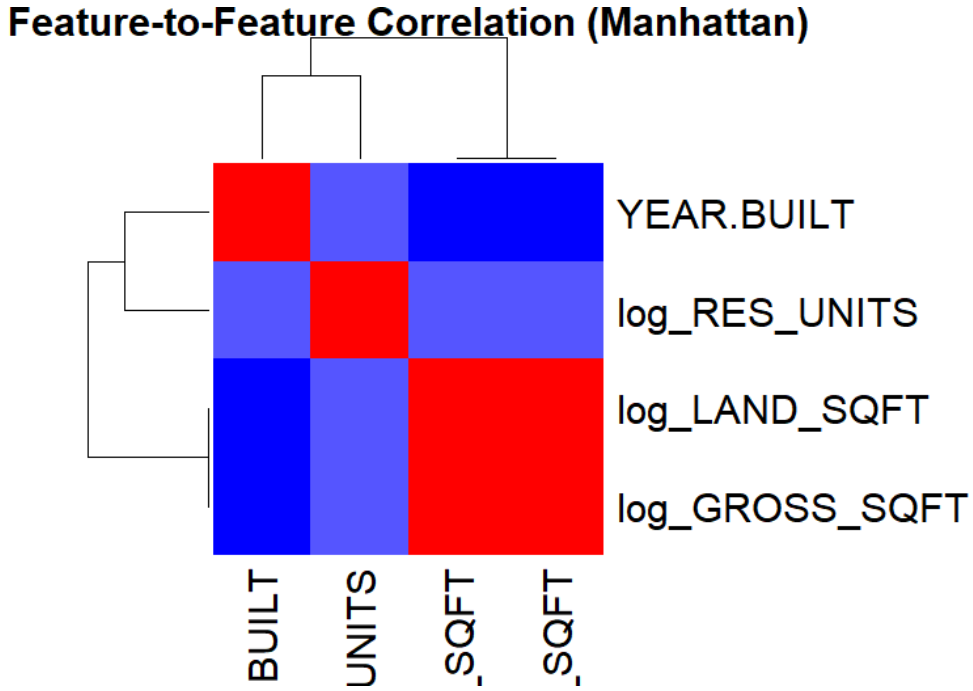




Logistically Transformed Variables' Summary Stats

```
> # Summary stats of Log Transformed Features
> summary(eda_data_clean$log_GROSS_SQFT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.09732 0.00000 8.91099
> summary(eda_data_clean$log_LAND_SQFT)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.08566 0.00000 7.22110
> summary(eda_data_clean$log_RES_UNITS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6931  0.6931  0.6931  0.6931  0.6931  0.6931
```

Feature-to-Feature Correlation Matrix (Heatmap)



Feature-to-Target Correlation Output

```
> print(cor_target_df)
                     Feature Correlation_with_SALE_PRICE
log_GROSS_SQFT log_GROSS_SQFT                0.104921593
log_LAND_SQFT   log_LAND_SQFT                0.103155408
YEAR.BUILT         YEAR.BUILT                0.009314719
log_RES_UNITS   log_RES_UNITS                         NA
```

Problem 1c

First Multiple Linear Regression Model Summary output:

```
> summary(mlr_model) # Print summary stats

Call:
lm(formula = SALE_PRICE ~ ., data = df_model)

Residuals:
     Min       1Q   Median       3Q      Max
-3563717 -1306581  -452612   658592  6140692

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1296988.7  1238503.2  -1.047   0.2951
YEAR.BUILT         1327.0      625.8   2.121   0.0340 *
log_GROSS_SQFT  1322047.0   534500.8   2.473   0.0134 *
log_LAND_SQFT  -1274235.7   607671.3  -2.097   0.0361 *
log_RES_UNITS          NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1566000 on 4352 degrees of freedom
Multiple R-squared:  0.01305,   Adjusted R-squared:  0.01237
F-statistic: 19.18 on 3 and 4352 DF,  p-value: 2.374e-12
```

Second MLR Model Summary Output:

```
> summary(mlr_model_2)

Call:
lm(formula = SALE_PRICE ~ ., data = df_model_2_final)

Residuals:
      Min        1Q    Median        3Q       Max
-60401156  -3473852  -1492685    433292 792860643

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.878e+09  1.977e+09  -1.961  0.04992 *
BLOCK              4.227e+03  1.434e+03   2.947  0.00322 **
LOT                1.050e+02  3.246e+02   0.323  0.74644
COMMERCIAL.UNITS  -7.471e+03  2.110e+04  -0.354  0.72332
TOTAL.UNITS        6.735e+03  2.069e+03   3.256  0.00114 **
YEAR.BUILT         2.478e+04  5.622e+03   4.408 1.06e-05 ***
Latitude          -4.882e+07  2.475e+07  -1.972  0.04862 *
Longitude         -7.868e+07  2.529e+07  -3.111  0.00187 **
Community.Board   -1.235e+04  5.386e+04  -0.229  0.81859
Council.District  -2.667e+04  2.294e+05  -0.116  0.90744
Census.Tract      -3.565e+01  3.782e+01  -0.943  0.34592
BIN               -9.932e-01  1.865e+00  -0.532  0.59440
BBL               -1.473e-03  2.448e-03  -0.602  0.54751
log_RESIDENTIAL_UNITS -1.826e+06  2.741e+05  -6.661 2.91e-11 ***
log_GROSS_SQFT     2.307e+06  2.811e+05   8.205 2.69e-16 ***
log_LAND_SQFT     -1.507e+06  3.260e+05  -4.622 3.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17840000 on 7472 degrees of freedom
Multiple R-squared:  0.04391,   Adjusted R-squared:  0.04199
F-statistic: 22.87 on 15 and 7472 DF,  p-value: < 2.2e-16
```

Third Model Summary Output:

```
> summary(mlr_model_3)

Call:
lm(formula = SALE_PRICE ~ ., data = df_model_3)

Residuals:
       Min         1Q     Median         3Q        Max
-123074147   -3649446   -2212538    -175373  781267749

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.892e+09  1.952e+09  -3.018 0.002551 **
BLOCK              3.437e+03  1.429e+03   2.404 0.016233 *
LOT               -1.382e+03  2.691e+02  -5.137 2.86e-07 ***
RESIDENTIAL.UNITS  1.884e+04  4.568e+03   4.124 3.77e-05 ***
COMMERCIAL.UNITS  -2.562e+03  2.051e+04  -0.125 0.900589
TOTAL.UNITS               NA         NA      NA       NA
LAND.SQUARE.FEET  -2.351e+02  1.984e+01 -11.850  < 2e-16 ***
GROSS.SQUARE.FEET  5.708e+01  3.510e+00  16.261  < 2e-16 ***
YEAR.BUILT        -3.013e+03  5.261e+03  -0.573 0.566871
Latitude          -2.295e+07  2.464e+07  -0.931 0.351715
Longitude         -9.241e+07  2.505e+07  -3.689 0.000227 ***
Community.Board    5.988e+03  5.347e+04   0.112 0.910834
Council.District  -1.273e+04  2.279e+05  -0.056 0.955448
Census.Tract       5.564e+00  3.737e+01   0.149 0.881632
BIN               -1.626e+00  1.854e+00  -0.877 0.380432
BBL               -4.087e-04  2.432e-03  -0.168 0.866529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17770000 on 7473 degrees of freedom
Multiple R-squared:  0.0513,    Adjusted R-squared:  0.04952
F-statistic: 28.86 on 14 and 7473 DF,  p-value: < 2.2e-16
```

Problem 1d

kNN Confusion Matrix Overall Stats:

```
Overall Statistics

               Accuracy : 0.1893
                 95% CI : (0.1733, 0.2062)
    No Information Rate : 0.1056
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1431

 Mcnemar's Test P-Value : NA
```

Random Forest Confusion Matrix Overall Stats:

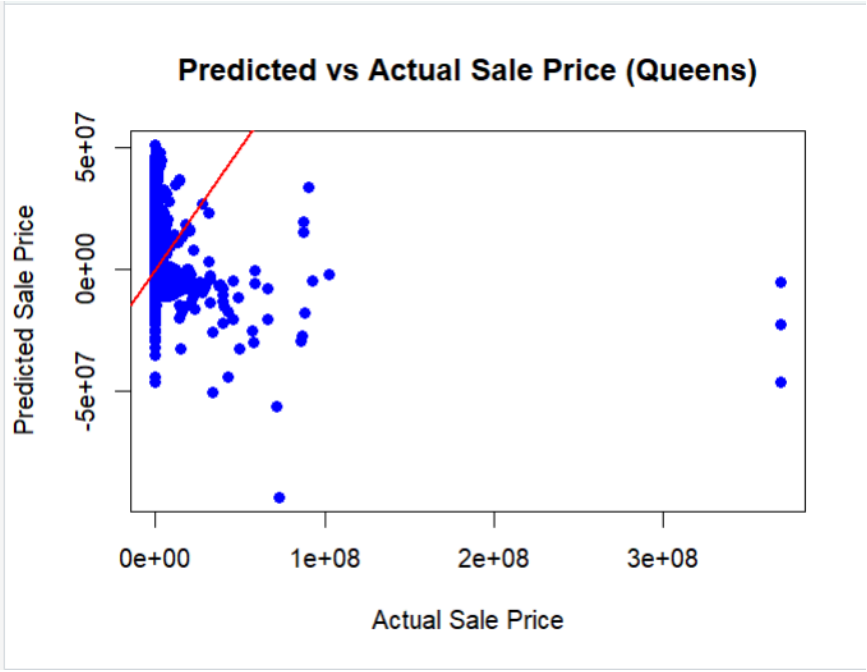
```
Overall Statistics

               Accuracy : 0.235
                 95% CI : (0.2176, 0.2532)
    No Information Rate : 0.1056
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1699

 Mcnemar's Test P-Value : NA
```

Naive Bayes Confusion Matrix Overall Stats:

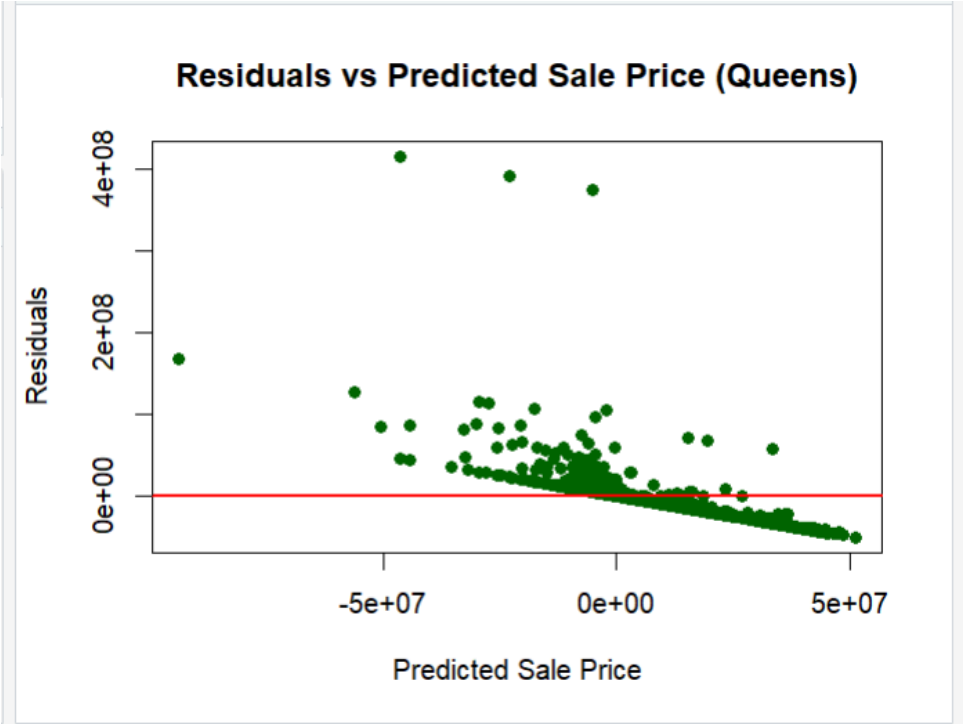
```
Overall Statistics

               Accuracy : 0.0363
                 95% CI : (0.0289, 0.0449)
    No Information Rate : 0.1056
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0067

 Mcnemar's Test P-Value : NA
```

Problem 2a

Prediction vs Actual



Residuals Plot



Problem 2b

k-NN Overall Stats

```
Overall Statistics

               Accuracy : 0
                 95% CI : (0, 1e-04)
    No Information Rate : 0.0851
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : NA
```

Random Forest Overall Stats

```
Overall Statistics

               Accuracy : 0
                 95% CI : (0, 1e-0
    No Information Rate : 0.0851
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : NA
```

Naive Bayes Overall Stats

```
Overall Statistics

               Accuracy : 0
                 95% CI : (0, 1e-04)
    No Information Rate : 0.0851
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : NA
```