Final Report: Statistical Analysis of Housing Prices in King County, USA Between May 2014 and May 2015

Submitted to:

Dr. Weihong Grace Guo

Prepared by:

Kyle Aitken (RUID: 171002178)

December 10th, 2020

Kyle Aitken
kyle.aitken@rutgers.edu

1. Data Exploration and Visualization

(a) The variables `id` , `view`, `condition`, and `grade` are qualitative. The `id` variable is nominal and has no numerical meaning, but the others are ordinal and their values will be studied in this analysis. All other variables are quantitative.

(b) Summary statistics for the `price` variable and the corresponding box-and-whisker plot are displayed in Table 1.1 and Figure 1.1, respectively. The `price` variable ranges from $107,000 to $1.445M with an average of $495,830 and a median of $432,500. The outliers (defined as any value less than Q1 or greater that Q3 by at least 1.5*IQR where IQR=Q3-Q1) are plotted as separate points on the box-and-whisker plot.

Histograms with 25, 50, 75, and 100 thousand dollar bin sizes are displayed in Figure 1.2 and indicate that the `price` variable is skewed to the right with a peak between $250,000 and $500,000.

Table 1.1 Summary statistics for the `price` variable, measured in $1000 units.

Statistic	Value ($1000)
Mean	495.83
Standard Deviation	249.02
Minimum	107.00
25th Percentile (Q1)	309.75
50th Percentile (Q2/Median)	432.50
75th Percentile (Q3)	627.50
Maximum	1445.00

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.1 Box-and-whisker plot for the `price` variable with outliers denoted by separate points.

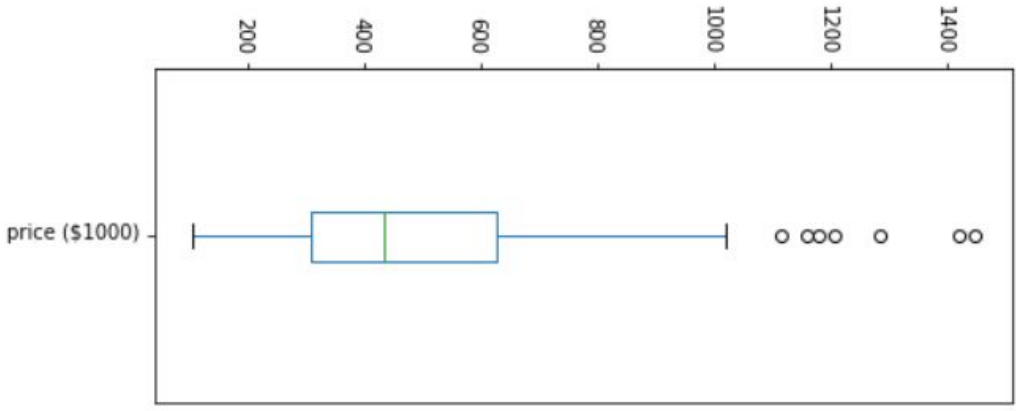
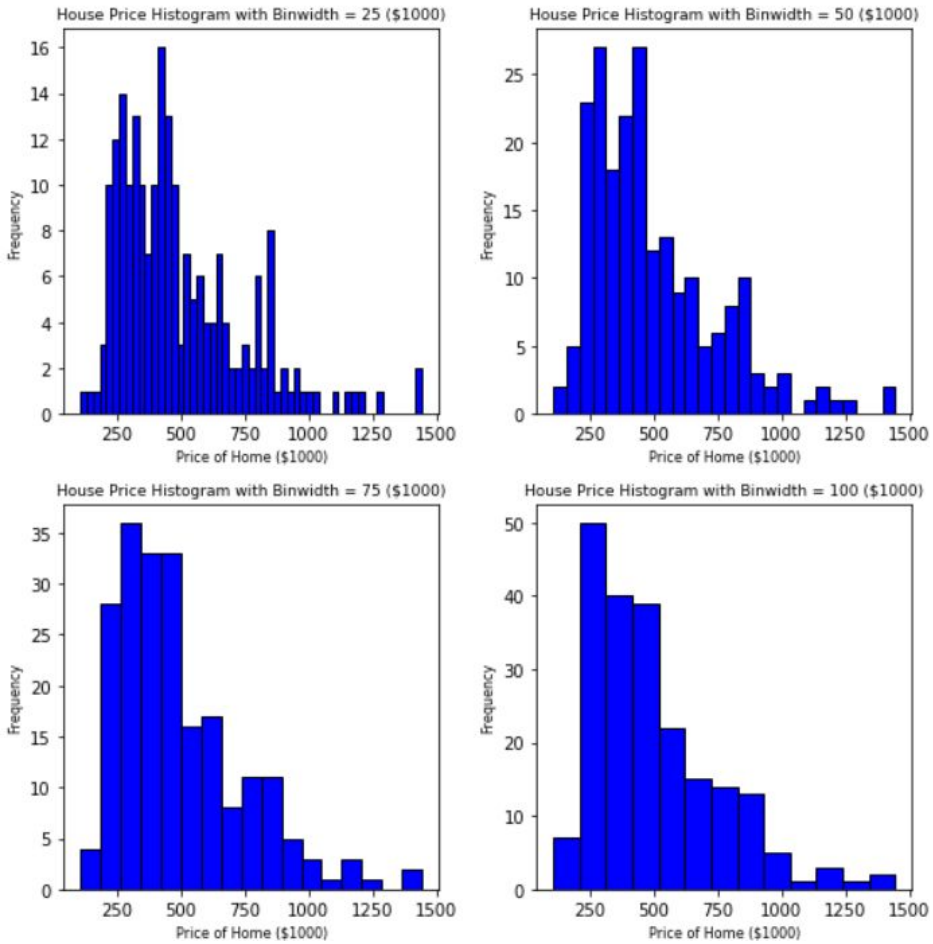


Figure 1.2 Histograms for the price variable with 25, 50, 75, and 100 thousand dollars bin sizes.



Kyle Aitken
kyle.aitken@rutgers.edu

(c) Summary statistics for the remaining variables (other than `id`) are provided in Table 1.2. It should be noted that unrenovated houses (`yr_renovated` = 0) were excluded from the summary statistics' calculations for `yr_renovated` to avoid being misleading. Additionally, the means and standard deviations for categorical variables (`view`, `condition`, and `grade`) should be viewed with skepticism, as their values may be deceptive.

Figures 1.3, 1.4, and 1.5 contain histograms and frequency bar charts for the remaining variables (other than `id`). Frequency bar charts were the chosen method of display for discrete variables and/or variables that mapped to a relatively small set of values. Histograms were chosen for continuous variables and/or variables that mapped to values over a relatively long range. The histogram for the `yr_renovated` variable does not include homes that were never renovated, similar to it's summary statistics in Table 1.2. The shapes of the remaining variables' distributions are summarized below:

- `bedrooms`: approximately normal distribution, skewed slightly to the right with a peak at 3 and a mean around 3.4 bedrooms.
- `bathrooms`: skewed to the right with three clear peaks at 1, 1.75, and 2.5 bathrooms. The mean is around 2.1 bathrooms.
- `sqft_living`: skewed to the right with a clear peak between 1,400 and 1,600 sqft. The mean is around 2046 sqft.
- `sqft_above`: skewed to the right with a clear peak between 1,00 and 1,200 sqft. The mean is around 1764 sqft.
- `floors`: skewed to the right with 2 clear peaks at 1 and 2 floors. The mean is around 1.5 floors.
- `yr_built`: skewed to the left with several peaks, most notably in the late 2000s. The mean is around 1971.
- `yr_renovated`: approximately normal distribution with a peak between 1985 and 1990. The mean is around 1990.
- `sqtf_living15`: skewed to the right with several peaks, most notably between 1,400 and 1,600 sqft. The mean is around 1974 sqft.
- `view`: skewed to the right with an obvious peak at 0 and a median of 0.

Kyle Aitken
kyle.aitken@rutgers.edu

- `condition`: skewed slightly to the right with an obvious peak at 3 and a median of 3.
- `grade`: skewed slightly to the right with an obvious peak at 7 and a median of 7.
- `latitude`: skewed to the left with several obvious peaks and a mean around 47.5 degrees.
- `longitude`: skewed to the right with several obvious peaks and a mean around -122.2 degrees. There are several outliers between -121.8 and -121.7 degrees.

Figure 1.6 contains a scatter plot of the houses' `latitude` and `longitude` values and puts the location of these homes in context. There is a large concentration of houses in the top left portion of the graph, which corresponds to the Seattle metropolitan area.

Table 1.2 Summary statistics for other variables (except `id`).

	count	mean	std	min	25%	50%	75%	max
bedrooms	212.0	3.424528	0.953399	1.0000	3.000000	3.00000	4.00000	6.0000
bathrooms	212.0	2.086085	0.750960	1.0000	1.687500	2.25000	2.50000	5.0000
sqft_living	212.0	2045.905660	828.253093	600.0000	1465.000000	1875.00000	2502.50000	4920.0000
floors	212.0	1.544811	0.554278	1.0000	1.000000	1.50000	2.00000	3.0000
view	212.0	0.207547	0.718305	0.0000	0.000000	0.00000	0.00000	4.0000
condition	212.0	3.330189	0.595746	2.0000	3.000000	3.00000	4.00000	5.0000
grade	212.0	7.646226	1.119553	5.0000	7.000000	7.00000	8.00000	11.0000
sqft_above	212.0	1764.349057	779.061521	600.0000	1175.000000	1550.00000	2260.00000	4130.0000
yr_built	212.0	1971.136792	29.787640	1901.0000	1948.000000	1976.50000	1997.00000	2015.0000
yr_renovated	8.0	1989.500000	15.099669	1964.0000	1985.250000	1988.50000	1995.50000	2012.0000
latitude	212.0	47.548950	0.141132	47.1938	47.453575	47.55465	47.67255	47.7761
longitude	212.0	-122.216368	0.146630	-122.4740	-122.337250	-122.23400	-122.11450	-121.7160
sqft_living15	212.0	1974.735849	687.977292	720.0000	1467.500000	1815.00000	2300.00000	3950.0000

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.3 Distributions of the houses' quantitative variables (excluding variables describing location).

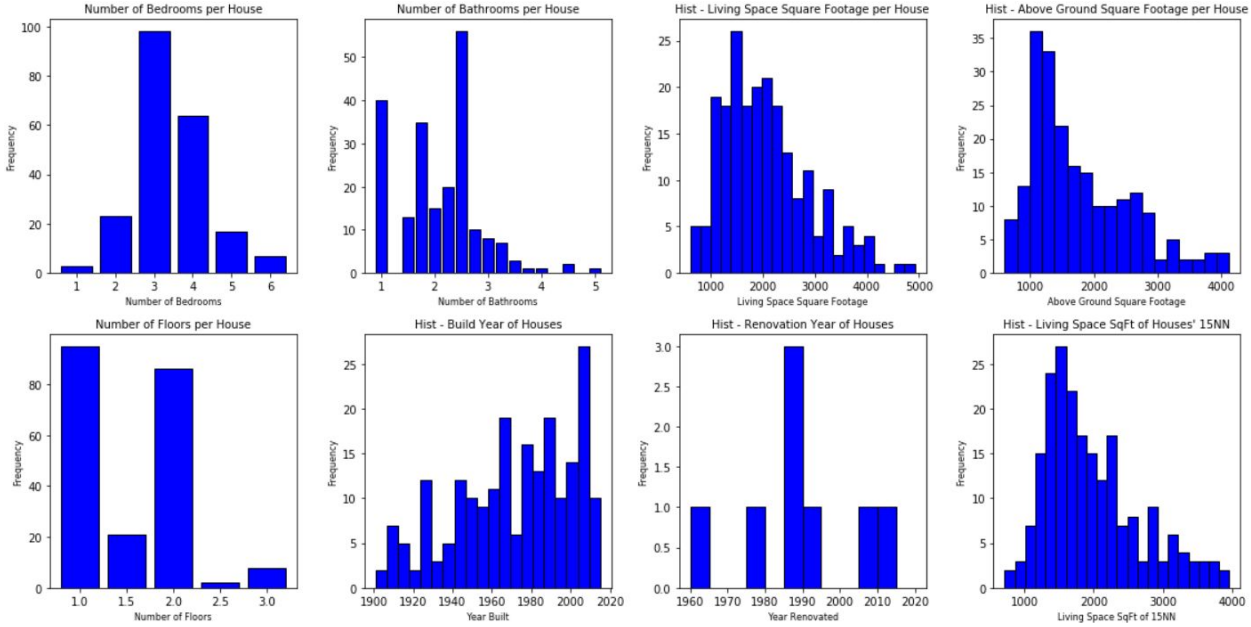
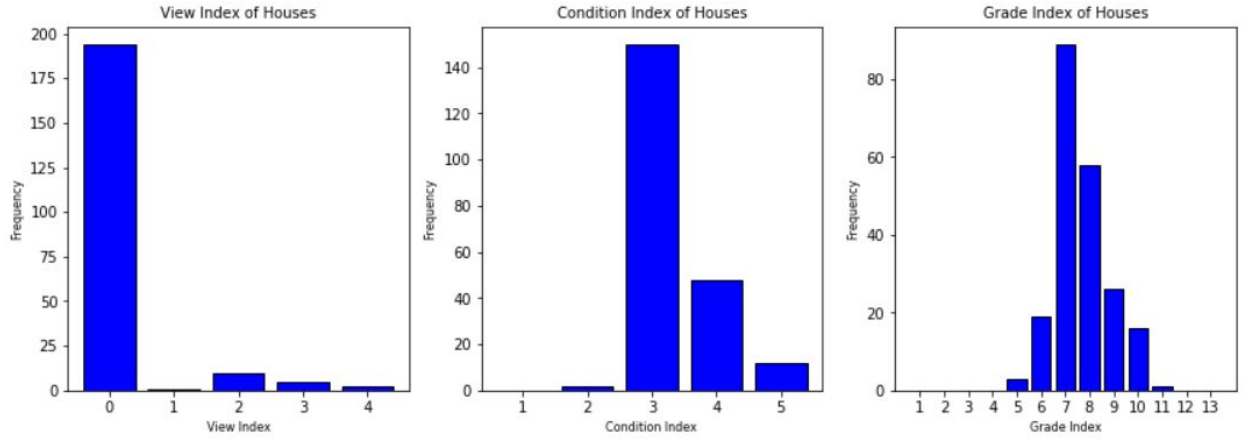
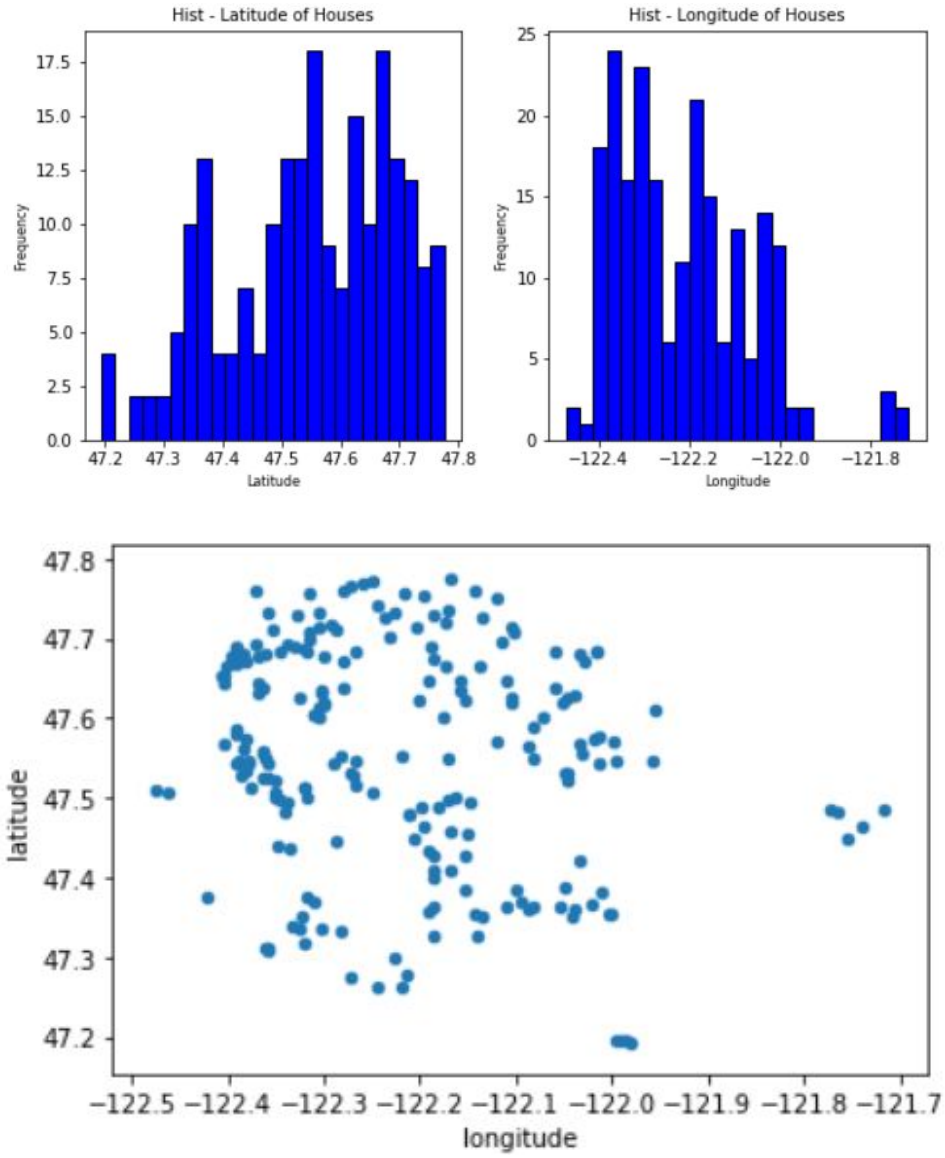


Figure 1.4 Distributions of the houses' categorical variables.



Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.5 (a) Distributions of the houses' location variables
(b) Scatterplot of `latitude` vs. `longitude`.



(d) The scatterplot matrix in Figure 1.6 shows the pairwise relationship between each of the 14 variables. A downloadable PDF of the scatterplot matrix is available here for closer inspection.

Many of the scatterplots shown in Figure 1.6 display some form of relationship between the variables. The variables `price`, `bedrooms`, `bathrooms`, `sqft_living`, `grade`, `sqft_above`, and `sqft_living15` all appear to have a

Kyle Aitken
kyle.aitken@rutgers.edu

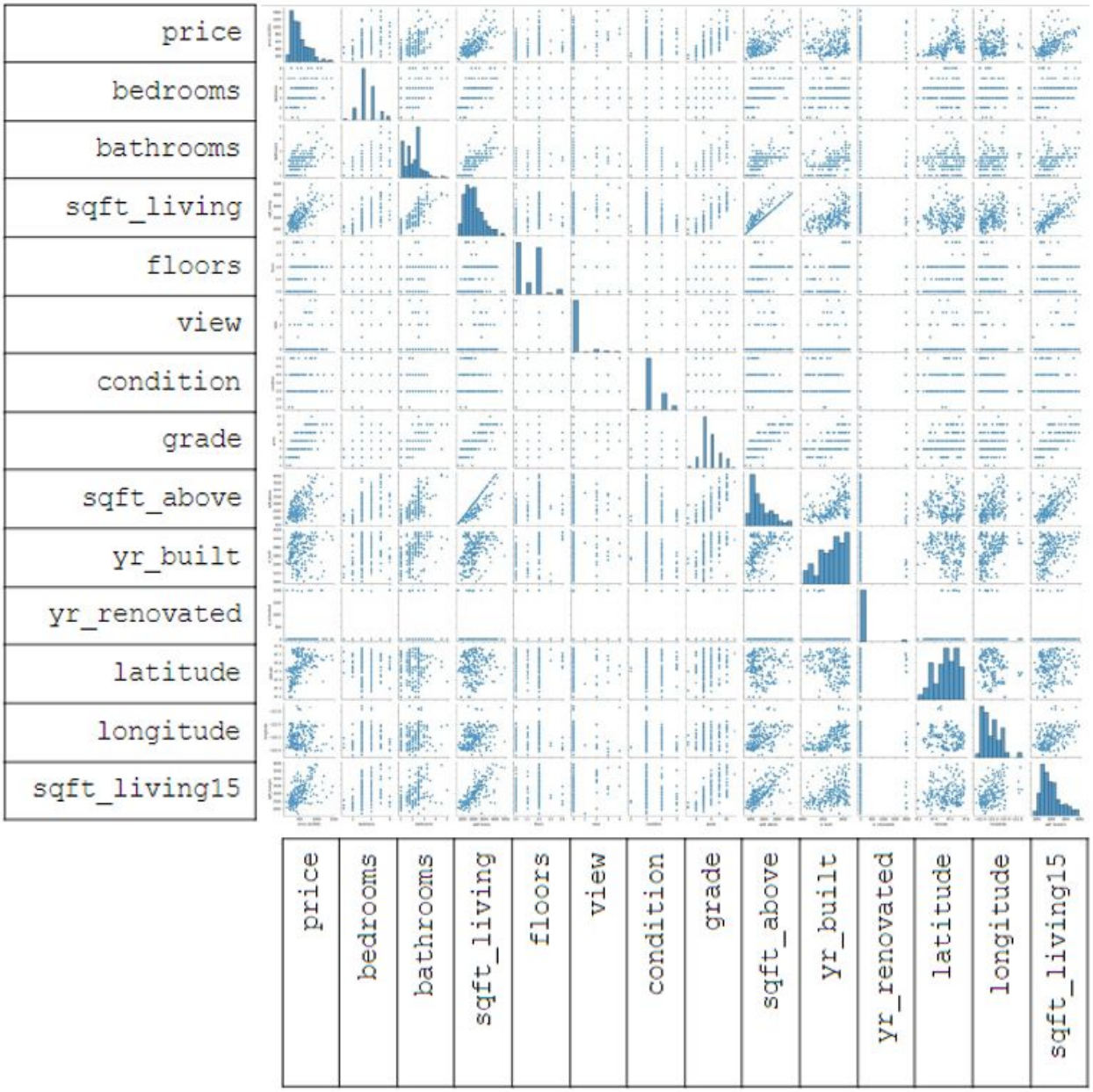
direct relationship - as one of these variables increases the other variables tend to increase as well, exhibiting a pairwise positive correlation between them. A scatterplot matrix for these seven variables is displayed in Figure 1.7.

The relationships between `yr_built` and the rest of the variables is not quite as clear as it was for the variables previously mentioned. The variable `yr_built` has a direct relationship with `grade` and appears to have a slightly direct relationship with `bedrooms`, `bathrooms`, and `sqft_living`. Not much can be said about the relationship between homes with high `yr_built` values and the other variables since newer houses take on a large range of values for nearly all the variables. However, some interesting things can be said about very old homes (pre-1940). Very old houses' `sqft_living` and `bathrooms` values very rarely exceed the mean, despite the fact that some of the most expensive houses in the data set are pre-1940. Additionally, very old houses almost always have high `latitude` values above the mean and low `longitude` values below the mean, which means very old homes were almost exclusively constructed in or around Seattle. Close-ups of the `yr_built` vs. `latitude` and `yr_built` vs. `longitude` scatterplots are shown in Figure 1.8.

The variables `floors`, `view`, `condition`, `yr_renovated`, `latitude`, and `longitude` do not have an obvious relationship with each other or with any of the other variables.

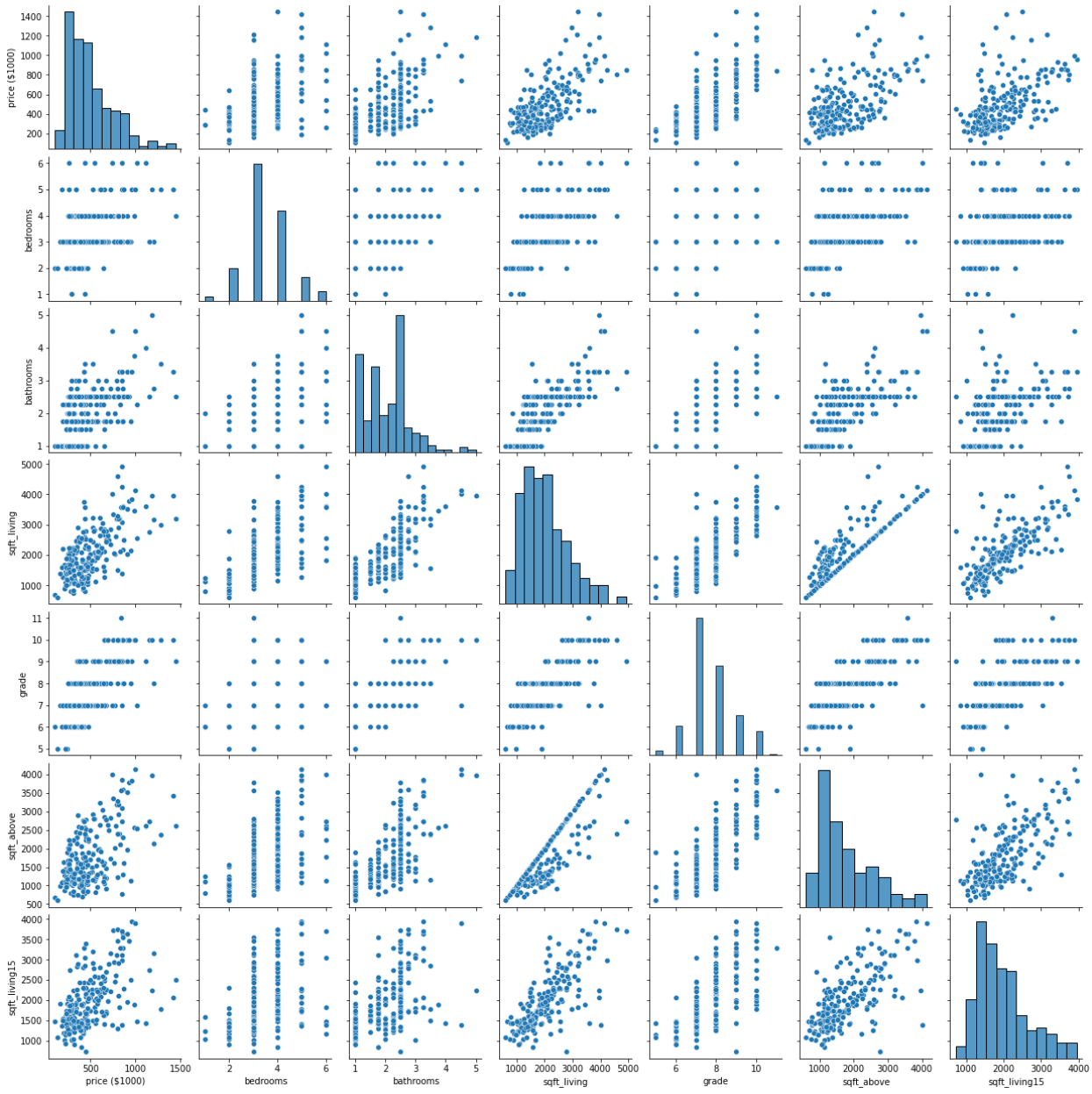
Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.6 Scatterplot matrix for all the variables in the dataset (except `id`).



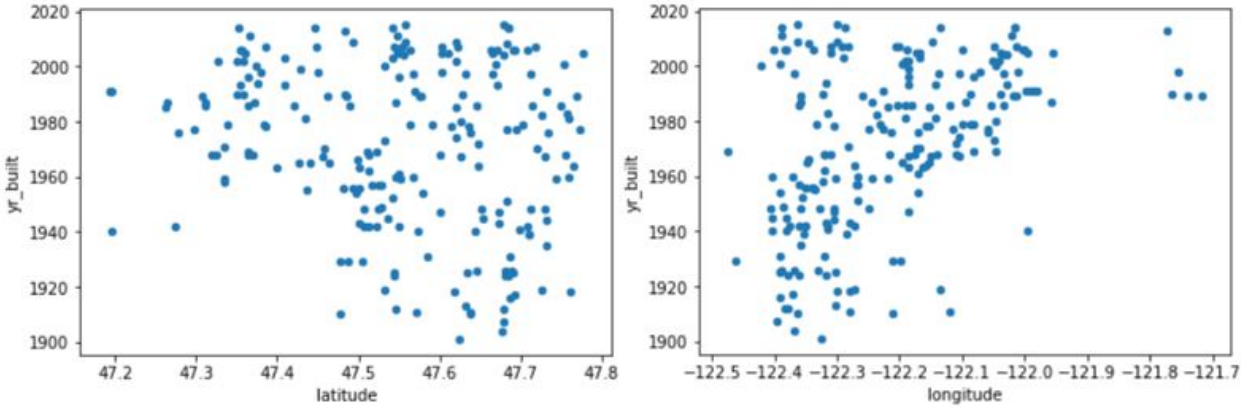
Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.7 Scatterplot matrix for `price`, `bedrooms`, `bathrooms`, `sqft_living`, `grade`, `sqft_above` and `sqft_living15`.



Kyle Aitken
kyle.aitken@rutgers.edu

Figure 1.8 Scatterplots for `yr_built` vs. `latitude` and `longitude`.



(e) The variables `bedrooms`, `bathrooms`, `sqtf_living`, `grade`, `sqft_above`, and `sqft_living15` all appear to have a direct relationship with the `price` variable. These variables would most likely be useful in predicting the price of a house.

2. Statistical Inference

(a) **95% Confidence Interval for Mean Sale Price of King County Houses:**

Known Values: alpha = .05, x_bar = $495,830, s = $249,020, N = 212

Critical Value: the critical value for alpha/2 = .025 with 211 degrees of freedom is:

t* = 1.97127

Calculation:

95%CI = ((x_bar - t*(s)/sqrt(212)), (x_bar + t*(s)/sqrt(212)))

95%CI = ($462,115.83, $529,544.17)

From these calculations, it can be said with 95% certainty that the average selling price of a King County, WA house between May 2014 and May 2015 is between $462,115.83 and $529,544.17.

Kyle Aitken
kyle.aitken@rutgers.edu

95% Confidence Interval for Mean Square Footage of King County Houses:

Known Values: alpha = .05, x_bar = 2045.91 sq. ft., s = 828.25 sq. ft., N = 212

Critical Value: the critical value for alpha/2 = .025 with 211 degrees of freedom is:

t* = 1.97127

Calculation:

95%CI = ((x_bar - t*(s)/sqrt(212)), (x_bar + t*(s)/sqrt(212)))

95%CI = (1933.78 sq. ft., 2158.04 sq.ft.)

From these calculations, it can be said with 95% certainty that the average square footage of a King County, WA house is between 1933.78 sq. ft. and 2158.04 sq.ft.

(b) **Testing Hypothesis:**

Known Values: n_y = 87, n_n = 125, x_bary = $549,862, x_barn = $458,224,
s_y = $274,445, s_n = $223,172

i. Parameter of Interest: difference of means, mu_y - mu_n

ii. Null Hypothesis: H_0: mu_y - mu_n = 0

iii. Alternative Hypothesis: H_1: mu_y - mu_n > 0

iv. Test Statistic: t_0' = (x_bary - x_barn - d_0) / sqrt(s_y^2/n_y + s_n^2/n_n)

v. Reject Null Hypothesis if: t_0' > t_alpha,v = t_.05,159 = 1.6545

v = (s_y^2/n_y + s_n^2/n_n)^2 / ((s_y^2/n_y)^2/(n_y-1) +
(s_n^2/n_n)^2/(n_n-1))

v = (274445^2/87 + 223172^2/125)^2 /
((274445^2/87)^2/86 + (223172^2/125)^2/124) = 159 (round down)

vi. Computations: Under null hypothesis, mu_y - mu_x = d_0 = 0.

t_0' = (x_bary - x_barn - d_0) / sqrt(s_y^2/n_y + s_n^2/n_n)

t_0' = (549862 - 458224) / sqrt(274445^2/87 + 223172^2/125)

t_0' = 2.5773

vii. Conclusion: Since t_0' = 2.5773 > 1.6545 = t_.05,159, we reject the null hypothesis, H_0: mu_y - mu_n = 0. We accept H_1 and agree with the claim that mu_y - mu_n > 0. In the context of this problem, we have strong evidence to support that the average selling price of houses with basements is higher than the average selling price of houses without basements in King County, WA between May 2014 and May 2015.

Kyle Aitken
kyle.aitken@rutgers.edu

P-Value:

P-Value = P(T_159 > 2.5773). Since t_.01,inf = 2.326 and t_.0025,inf = 2.807, the value P(T_159 > 2.5773) should be between .0025 and .01. So, .0025 <P-Value< .01. Since P-Value < .01 < .05 = alpha, we reject the null hypothesis and accept the alternative hypothesis, as stated above in the Conclusion section.

(c) **Two-Sided Confidence Interval for the Difference of Means:**

The general equation for a two-sided confidence interval for this problem is:

$$(\bar{x}_y - \bar{x}_n) - t_{\alpha,\nu}\sqrt{s_y^2/n_y + s_n^2/n_n} \leq \mu_y - \mu_n \leq (\bar{x}_y - \bar{x}_n) + t_{\alpha,\nu}\sqrt{s_y^2/n_y + s_n^2/n_n}$$

The calculations eventually lead to the two-sided confidence interval:

$$\$21{,}415.88 < \mu_y - \mu_n < \$161{,}860.12.$$

From this confidence interval, we can conclude, with 95% certainty, that the difference in means, mu_y - mu_n, is between $21,415.88 and $161,860.12. In the context of this problem, this means we can conclude, with 95% certainty, that the average selling price of a house with a basement is between $21,415.88 and $161,860.12 more than the average selling price of a house without a basement in King County, WA.

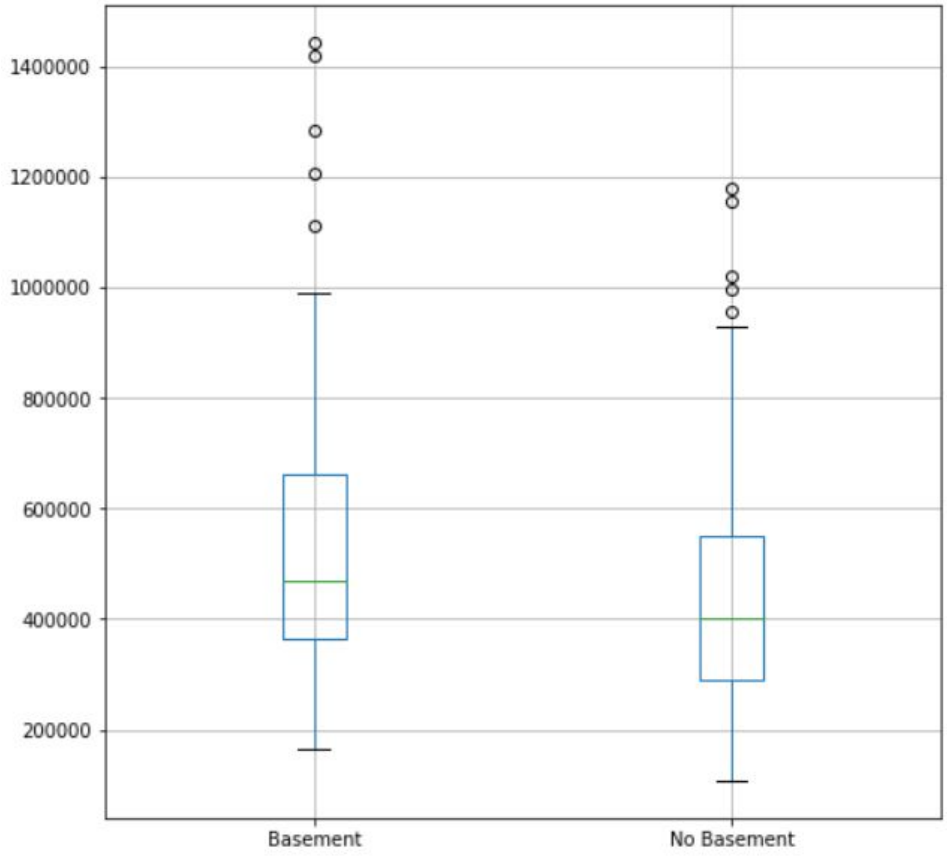
(d) Figure 2.1 shows a side-by-side boxplot comparing the price distributions of houses with basements (left) and houses without basements (right).

Houses with basements have a median price of around $470,000 and an interquartile range of about $300,000, with 5 outliers exceeding Q3+1.5*IQR. Houses without basements have a median price of around $400,000 and an interquartile range of about $260,000, with 5 outliers exceeding Q3+1.5*IQR.

Despite the fact that the center of the price distribution for houses with basements is slightly greater, these two distributions are very similar. The quartiles are relatively similar in length and each boxplot has similarly sized ‘whiskers’. In addition, both distributions are slightly skewed to the right with a few outliers lying above their upper whisker.

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 2.1 Side-by-side boxplot comparing price distributions for houses with basements and houses without basements.



3. Simple Linear Regression

(a) Figure 3.1 is a scatterplot of `price` vs. `sqft_living` with the calculated least squares regression line overlaid on top in red. The equation of the least squares regression line is:

`price($1000)` = 80.7945 + .2029*(`sqft_living`)

There is a relatively strong positive relationship between `price` and `sqft_living`, as these two variables have a correlation of .6747. According to the SLR model, the price of the house increases .2029($1000) = $202.90 for each additional square foot of living space. The predicted selling price of a 2,000 sq. ft. house is 486.51769($1000) = $486,517.69.

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 3.1 Least squares regression line overlaid on scatterplot of `price` vs. `sqft_living`.

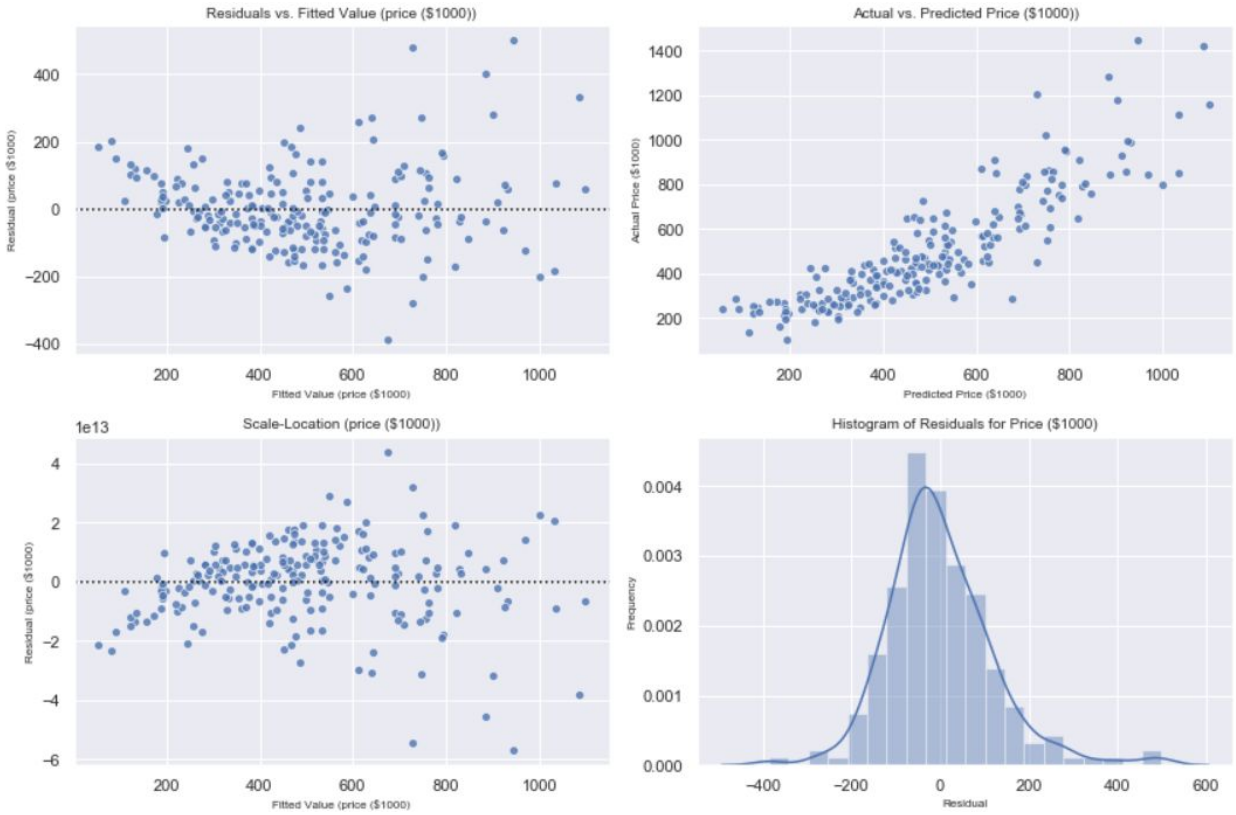
(b) Figure 3.2 shows diagnostic plots for the linear regression model in 3(a). The residuals in the residual plot and the scale-location plot tend to increase in magnitude as square footage increases. This is confirmed by the increase in sparsity of the points on the Actual vs. Fitted plot. Ideally, there would not be a clear pattern in the residual or scale-location plots and the locations of the data points would be close to the 45 degree line on the Actual vs. Fitted plot over the whole range of fitted values.

Additionally, the histogram for the residuals would ideally exhibit a normal distribution centered at 0, but this histogram for residuals is bimodal and skewed to the right.

These unideal characteristics of the diagnostic plots indicated that the simple linear regression model may need fine tuning and/or may not be the best fit for this data.

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 3.2 Diagnostics plot for the simple linear regression line in 3(a).



(c) Figure 3.3 shows the regression lines overlaid on scatterplots of `price` vs. `sqft_living` and residual vs. fitted value (predicted value of `price`) plots for each of the three models. Table 3.1 displays the root mean squared error, the mean absolute percentage error, R^2, and Adjusted R^2 of the three regression models. All of the R^2 and Adjusted R^2 values are relatively the same. Of the three models, the linear fit with a logarithmic transformation on the `price` variable is the best model for predicting `price` given `sqft_living`. This model predicted `price` with the lowest MAPE, meaning it predicted the response variable with a lower absolute percentage error than the other two models. Furthermore, the third model's residual plot exhibits no clear pattern, meaning the model is likely unbiased, unlike the other two residual plots where the magnitude of the residual increases with fitted value.

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 3.3 Least squares regression lines on scatterplots of `price` vs. `sqft_living` and Residual vs. Fitted Value plots of `price` for each of the three models.

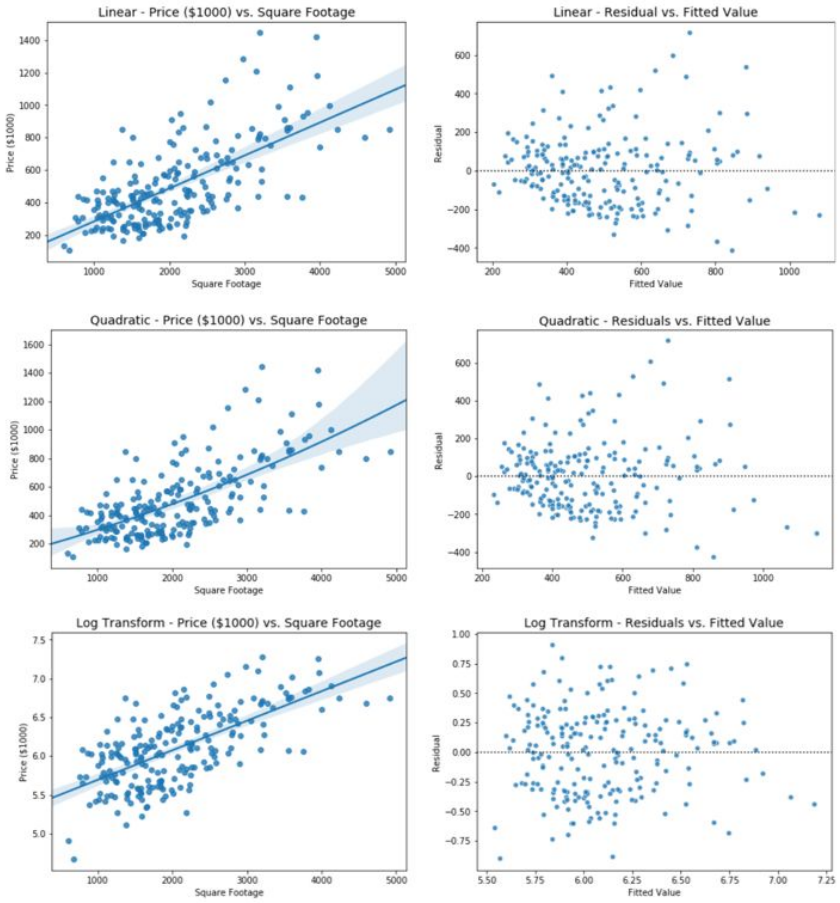


Table 3.1 Root mean squared error, mean absolute percentage error, R^2, and R^2_adj values for the three regression models.

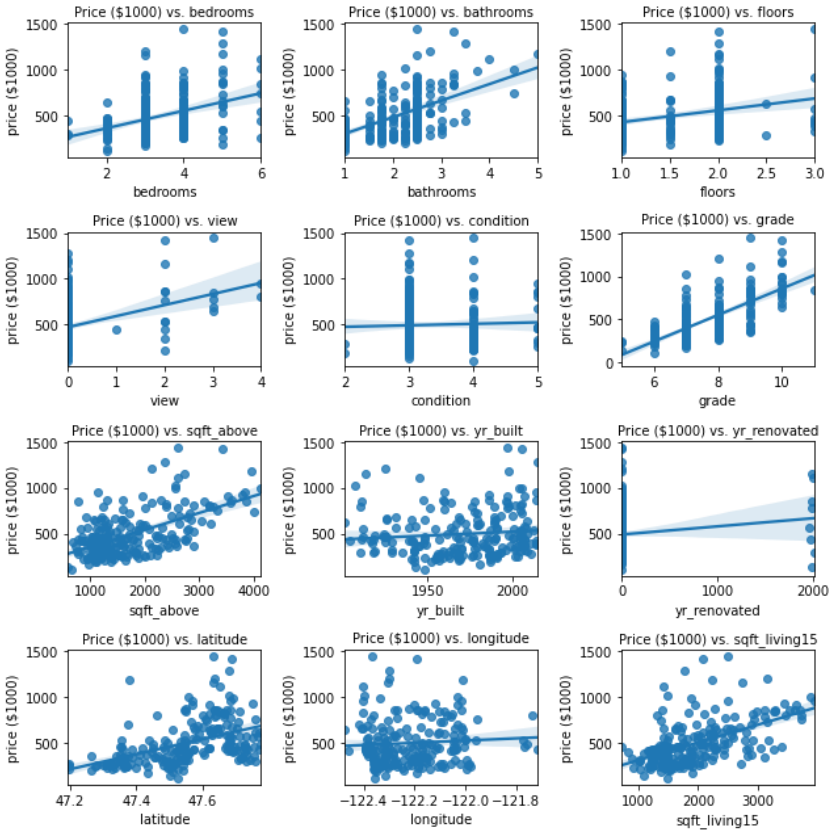
	RMSE	MAPE	R2	R2_adj
Linear Model	183.358588	.0332961	.455260	.4526666
Quadratic Fit Model	183.000241	.330846	.457387	.454803
Logarithmic Transform Model	.0357463	.049296	.436105	.433420

Kyle Aitken
kyle.aitken@rutgers.edu

(d) Figure 3.4 shows the individual scatterplots of `price` vs. each variable and the least squares line calculated from a simple linear regression. Table 3.2 shows the slope coefficients and their corresponding p-values for each variable's regression fit line.

The plots/table show that the model calculated a positive linear relationship between all of the variables and price. Furthermore, the p-values for each of the slope coefficients, except for `condition`, are less than .05. This indicates that the relationships calculated by the linear regression models between `price` and all other variables (except `condition`) are statistically significant. That is to say that there is strong evidence to support that there is a relationship between every variable (except `condition`) and `price`. Conversely, there is not enough evidence to reject the null hypothesis that `condition` has no relationship with `price`.

Figure 3.4 Linear regression fit lines on scatterplots of `price` vs. all other variables (except `id` and `sqft_living`).



Kyle Aitken
kyle.aitken@rutgers.edu

Table 3.2 Table showing each variable's slope coefficient and p-value after running a linear regression against `price`.

Variable	Slope Coefficient	P-Value
bedrooms	95.47	0.0000
bathrooms	180.63	0.0000
sqft_living	0.20	0.0000
floors	129.22	0.0000
view	120.45	0.0000
condition	16.26	0.1906
grade	154.17	0.0000
sqft_above	0.19	0.0000
yr_built	0.84	0.0000
yr_renovated	0.09	0.0435
latitude	820.36	0.0000
longitude	121.87	0.0000
sqft_living15	0.19	0.0000
bedrooms	95.47	0.0000

4. Multiple Linear Regression

(a) The multiple linear regression model has an R^2 value of .77, an R^2_adj of .75, an RMSE of 119.92, and a MAPE of .33. All of these statistics are improvements over the simple linear regression models from Section 3, so there is definitely a relationship between the predictors and the response variable.

Summary statistics for the MLR model's coefficients are shown in Figure 4.1. The p-values for all of the variables except `bedrooms`, `floors`, `sqft_above`, and

Kyle Aitken
kyle.aitken@rutgers.edu

`yr_renovated` are below alpha = .05. This means we have strong evidence to support that the variables with coefficient p-values below .05 have a statistically significant relationship with `price` and are likely useful additions to the regression model.

The coefficient for `latitude` tells us that, according to the model, the rate of change of the `price` variable's conditional mean with respect to `latitude` is around $656,000 per degree latitude. The `latitude` variable has a p-score or 0.0000, meaning the value of its coefficient is statistically significant and it is very likely that `price` and `latitude` have a direct relationship. Similarly, the coefficient for `grade` has a p-value of 0.0000, which means it is very likely `grade` and `price` have some sort of relationship. The value of the coefficient tells us that, with all other variables remaining the same, a `grade` index increase of 1 corresponds to around a $101,838 bump in house price.

Figure 4.1 Summary statistics for the MLR model used to predict the `price` variable.

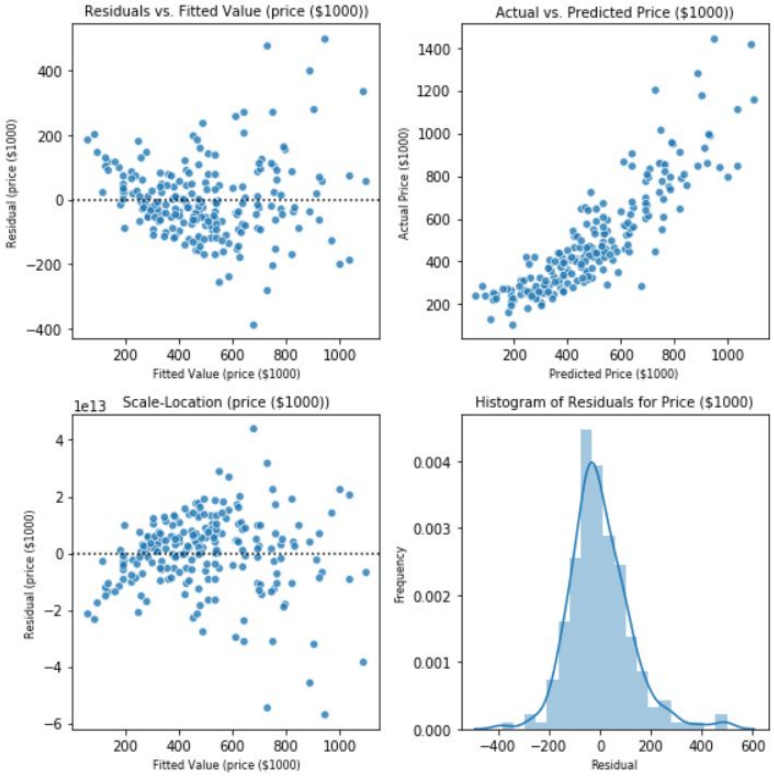
```
========================= SUMMARY =========================
Residuals:
     Min       1Q    Median        3Q       Max
 53.9314  334.849  472.1206  635.0334  1098.7871


Coefficients:
                  Estimate   Std. Error  t value   p value
_intercept    -19742.188213  9497.098454  -2.0788  0.038848
bedrooms          13.409008    11.676583   1.1484  0.252117
bathrooms         61.871120    20.515548   3.0158  0.002877
sqft_living        0.042339     0.018632   2.2724  0.024072
floors            17.739773    21.411955   0.8285  0.408324
view              62.513922    13.427477   4.6557  0.000006
condition         45.557019    15.433232   2.9519  0.003516
grade            101.837776    13.378191   7.6122  0.000000
sqft_above         0.027649     0.018266   1.5136  0.131612
yr_built          -1.813369     0.035941 -50.4535  0.000000
yr_renovated       0.029580     0.021745   1.3603  0.175190
latitude         656.648726    62.254458  10.5478  0.000000
longitude         70.540058    15.225487   4.6330  0.000006
sqft_living15     -0.036851     0.016190  -2.2762  0.023839
---
R-squared:  0.76700,    Adjusted R-squared:  0.75170
F-statistic: 50.14 on 13 features
```

Kyle Aitken
kyle.aitken@rutgers.edu

(b) Diagnostic plots for the multiple linear regression model are displayed in Figure 4.2. The residual plot and the scale-location plot exhibit slightly u-shaped and inverse u-shaped distributions, respectively. This indicates that there might be a better fit for this data with a non-linear model.

The points on the actual value vs. predicted value plot are slightly below the diagonal (especially in the middle part of the x-axis), suggesting that the MLR model tends to over estimate. The center of the histogram is slightly below 0, which also suggests that the model tends to over predict the response variable. The slight curve in the actual vs. predicted plot and the high concentration of negative residues in the middle of the fitted value range on the residue plot suggests that a transformation or new, non-linear model might be needed to improve accuracy. Additionally, there are some very extreme outliers when the Fitted Value is large (> 600 ($1000)), indicated by the points with large (in magnitude) residuals on the residual plots.

Figure 4.2 Diagnostic plots for the Multiple Linear Regression Model

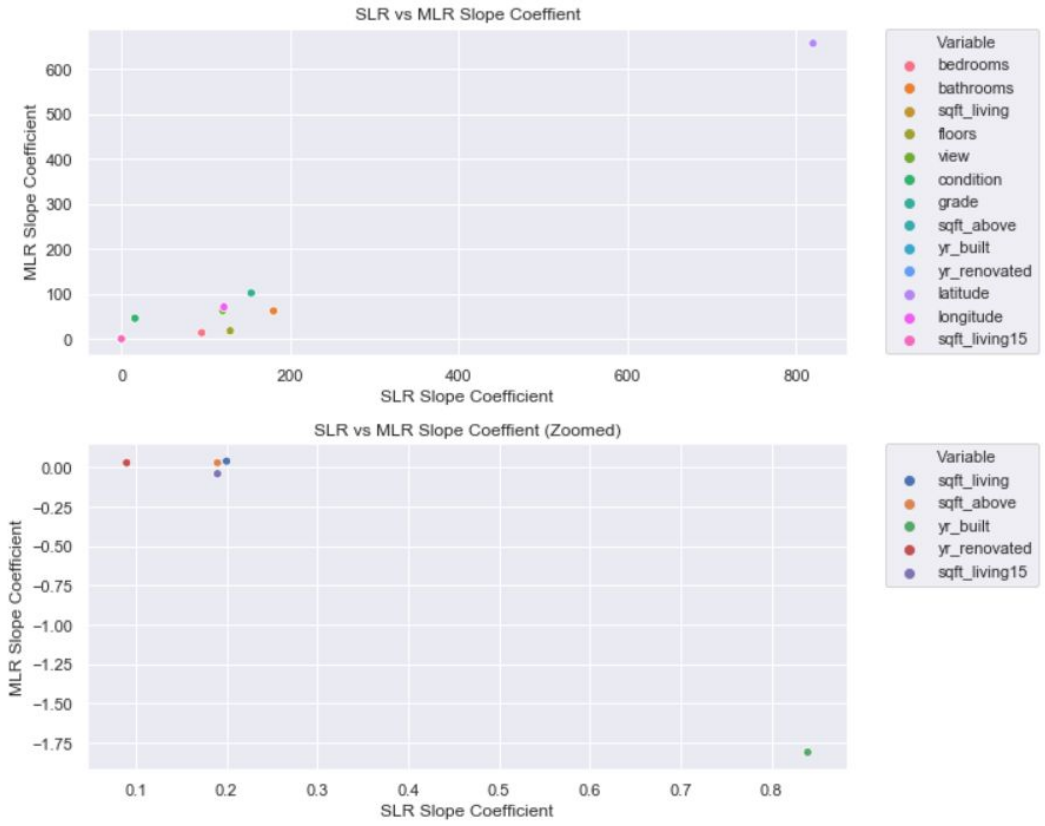


Kyle Aitken
kyle.aitken@rutgers.edu

(c) Figure 4.3 compares the univariate coefficients obtained in 3(d) and the multiple linear regression coefficients obtained in 4(a). The first graph plots SLR Slope Coefficient vs. MLR Slope Coefficient for all of the variables. The second plot is the same as the first, but zoomed to more clearly display the data points with an SLR Slope Coefficient lower than 1.

In general, the magnitude of the slope coefficients were smaller for the MLR model, with the exception being `condition`. Furthermore, while all the variables' slope coefficients were different in each model, the relative magnitude of the coefficients was more or less that same. That is, variables with lower slope coefficients in the SLR model generally had lower slope coefficients in the MLR model. It is also important to note that the coefficients for yr_built and sqft_living15 are negative for the MLR model but positive in the SLR model, potentially due to multicollinearity between these variables and other predictor variables used in the MLR model.

Figure 4.3 Plots comparing SLR and MLR Slope Coefficients for each variable.



Kyle Aitken
kyle.aitken@rutgers.edu

(d) The multiple linear regression model was run 78 addition times, each time including 1 of the 78 interactions (excluding interactions of degree 2). For each iteration, the p-value of the coefficient for the interaction, the Adjusted R^2, and the RMSE were kept track of. A list of 5 interactions that produced a p-value less than .05, an Adjusted R^2 greater than .77, and an RMSE less than 119 was created and can be seen in Table 4.1. These interactions improved the Adjusted R^2 and RMSE of the original multiple linear regression (Adjusted R^2 = .75170, RMSE = 119.91876) by a relatively significant amount.

The multiple linear regression was run yet again, this time using the original 13 variables and the 5 most relevant interactions as predictors. This model had an Adjusted R^2 of .78971, an RMSE of 108.96, and a MAPE of .1966. However, it changed the p-values of several of the variables. The summary statistics for this model can be seen in Figure 4.4.

Table 4.1 Table showing the adjusted R^2 and RMSE of the MLR model with added interactions.

Interaction	R2 Adj	RMSE
bathrooms*floors	0.777109	113.330013
bathrooms*grade	0.771990	114.623911
sqft_living*floors	0.783212	111.767691
floors*grade	0.774354	114.028288
floors*sqft_above	0.773127	114.337781

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 4.4 Summary statistics for the MLR model including the relevant interactions as predictors.

```
============================ SUMMARY ============================
Residuals:
     Min        1Q    Median        3Q        Max
 68.5709  344.4643  455.1361  627.5059  1196.2303


Coefficients:
                       Estimate   Std. Error  t value   p value
_intercept         -16555.051116  8768.767296  -1.8880  0.060404
bedrooms               22.359089    10.674987   2.0945  0.037406
bathrooms            -146.073616    78.305335  -1.8654  0.063508
sqft_living            -0.083764     0.042824  -1.9560  0.051782
floors               -323.443680   158.861036  -2.0360  0.042998
view                   70.268657    12.343528   5.6928  0.000000
condition              49.450083    14.128041   3.5001  0.000567
grade                  19.186530    30.997503   0.6190  0.536604
sqft_above              0.009748     0.061134   0.1595  0.873464
yr_built               -1.620029     0.034778 -46.5821  0.000000
yr_renovated            0.015421     0.019806   0.7786  0.437075
latitude              635.512438    56.969479  11.1553  0.000000
longitude              84.598757    14.309515   5.9121  0.000000
sqft_living15          -0.008886     0.015171  -0.5857  0.558683
bathrooms*floors       39.355377    31.497966   1.2495  0.212882
bathrooms*grade        19.421372     3.755958   5.1708  0.000001
sqft_living*floors      0.110651     0.025828   4.2841  0.000028
floors*grade           21.723540     9.699489   2.2397  0.026156
floors*sqft_above      -0.049141     0.029270  -1.6789  0.094660
---
R-squared:  0.80765,    Adjusted R-squared:  0.78971
F-statistic: 45.02 on 18 features
```

(e) Table 4.2 shows the Adjusted R^2 and MAPE of the multiple linear regression models (with interactions included) with several different transforms of the response variable. Squaring the response variable prior to running the MLR model proved to be less effective than using the actual value for the response. However, taking the natural log of the response variable and taking the square root of the response variable both improved the Adjusted R^2 and MAPE of the MLR model.

Kyle Aitken
kyle.aitken@rutgers.edu

Taking the natural log of the price variable prior to running the MLR model with the relevant interaction appears to be the best transformation, as is produces an Adjusted R^2 value of .7949 and a MAPE of .0258, both much better value than that of those produced by the model run in 4(d).

Table 4.2 Adjusted R^2 and MAPE Values for the MLR model (with relevant interactions) run with different transformations.

	log(`price`)	sqrt(`price`)	(`price`)^2
Adjusted R^2	.7949	.8061	.7247
MAPE	.0258	.0826	.7285

(f) The best MLR so far is the model that takes a logarithmic transform of the `price` variable and uses the 13 original variables and the 5 interactions as predictor variables. In an effort to simplify this model by reducing the number of predictors, I removed the variables that had slope coefficients greater than .05 in the original MLR model with the interactions included (summary shown in Figure 4.4). The variables removed were: `bathrooms`, `sqft_living`, `grade`, `sqft_above`, `sqft_living15`, `yr_renovated`, `bathrooms*floors`, and `floors*sqft_above`. This model has the following attributes:

- R^2: .7993
- Adjusted R^2: .7893
- RMSE: .2133
- MAPE: .0264

Excluding the variables mentioned above did not severely alter the statistics shown above and is much simpler, in that it takes only 10 predictor variables instead of the original 18. Therefore, the best MLR model for price is modeled by:

Kyle Aitken
kyle.aitken@rutgers.edu

ln(price) = -25.923677 + .036093*bedrooms - .590162*floors + .116462*view + .086284*condition - .003613*yr_built + 1.522096*latitude + .279368*longitude + .020220*bathroom*grade + .000026*sqft_living*floors + .079752*floors*grade

The summary for this model can be seen below in Figure 4.5.

Figure 4.5 Summary statistics for the best MLR model for predicting price.

```
========================== SUMMARY ==========================
Residuals:
    Min      1Q  Median      3Q     Max
-0.5756 -0.1274  0.0029  0.1138  0.9633


Coefficients:
                     Estimate  Std. Error  t value   p value
_intercept         -25.923677   15.658468  -1.6556  0.099296
bedrooms             0.036093    0.019815   1.8215  0.069946
floors              -0.590162    0.105602  -5.5885  0.000000
view                 0.116462    0.021313   5.4645  0.000000
condition            0.086284    0.027232   3.1685  0.001760
yr_built            -0.003613    0.000056 -64.8660  0.000000
latitude             1.522096    0.104901  14.5099  0.000000
longitude            0.279368    0.025925  10.7759  0.000000
bathrooms*grade      0.020220    0.003690   5.4792  0.000000
sqft_living*floors   0.000026    0.000006   4.4778  0.000012
floors*grade         0.079752    0.007252  10.9967  0.000000
---
R-squared:  0.79930,    Adjusted R-squared:  0.78932
F-statistic: 80.05 on 10 features
```

5. Classification

(a) A binary variable before1980 was created in the dataset. The variable takes on a value of 1 if the house was built prior to 1980 and takes on a value of 0 if the house was built in or after 1980.

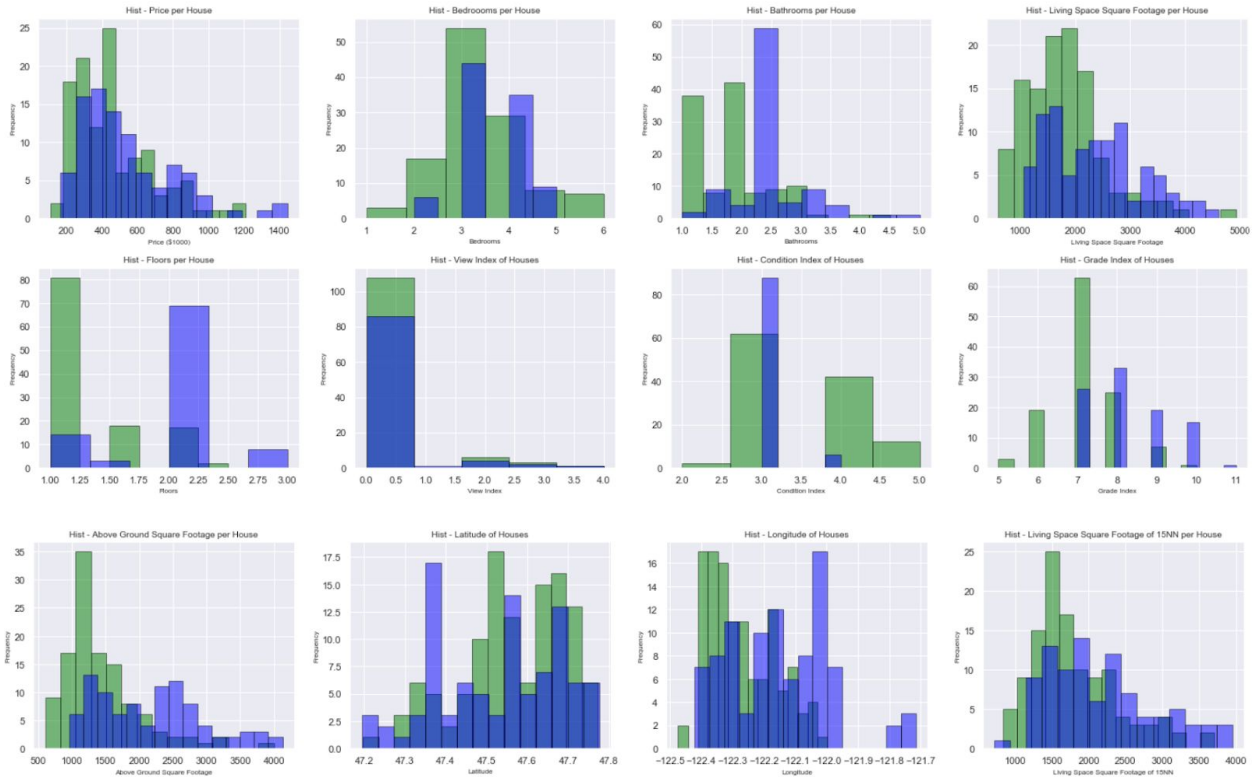
(b) Histograms for 12 of the predictor variables are plotted in Figure 5.1 (the variable yr_renovated equals 0 for every house built after 1980, so it was omitted from this figure and will not be considered in our classification analysis). The green bars represent

Kyle Aitken
kyle.aitken@rutgers.edu

houses that were built before 1980 (`before1980` = 1) and the blue bars represent houses that were built in or after 1980 (`before1980` = 0).

To find useful predictors for the `before1980` variable, we need to identify variables whose green and blue histogram distributions are noticeably different (different center, different peak locations, different skewness, etc). Based on these histograms, `bathrooms`, `sqft_living`, `floors`, `condition`, `grade`, `sqft_above`, `latitude`, and `longitude` appear to be the most useful in predicting `before1980`.

Figure 5.1 Histograms for the predictor variables (green bars represent houses built before 1980 and blur bars represent houses built in or after 1980).



(c) The variables that will be used in the Logistic Regression Classifier and k-NN Classifier are `bathrooms`, `sqft_living`, `floors`, `condition`, `grade`, `sqft_above`, `latitude`, and `longitude`. The data will be split into a training and test set, with

Kyle Aitken
kyle.aitken@rutgers.edu

70% (148 samples) included in the training set and 30% (64 samples) of the data included in the test set.

(d) Prior to running the logistic regression classifier, the predictor variables in the training and test data were normalized using the sklearn's `StandardScaler()` module.

After normalization, the logistic regression classifier was 90.625% accurate (9.375% test error rate), meaning 58 of the 64 houses in the test set were classified correct as being either built before 1980 or in/after 1980. The confusion matrix and classification report showing precision-recall are displayed in Figure 5.2.

Figure 5.2 Confusion matrix and classification report for the Logistic Regression Classifier.

```
The Confusion Matrix for the Logistic Regression Classifier is:

 [[29  3]
 [ 3 29]]

The Classification Report for the Logistic Regression Classifier is:

              precision    recall  f1-score   support

         0.0       0.91      0.91      0.91        32
         1.0       0.91      0.91      0.91        32

    accuracy                           0.91        64
   macro avg       0.91      0.91      0.91        64
weighted avg       0.91      0.91      0.91        64
```

(e) Prior to running the k-NN classifier, the predictor variables were normalized using sklearn's `StandardScaler()` module.

Before choosing the k-values, an Error Rate vs. K-Value graph was plotted for k-values ranging from 1 to 40. The resulting graph is shown in Figure 5.3. The k-values chosen for evaluation were 5, 12, and 20 because these were all local minimums on the Error Rate vs. K-Value graph. These k-values returned accuracy scores of 90.625% (9.375% error rate), 92.1875% (7.8125% error rate), and 89.0625% (10.9375% error rate), respectively. The optimal k for this dataset is k = 12. The confusion matrix and classification report for the k-NN Classifier (for k = 12) are shown in Figure 5.4.

Kyle Aitken
kyle.aitken@rutgers.edu

Figure 5.3 Error Rate vs. K-Value Plot for k-values ranging from 1 to 40.

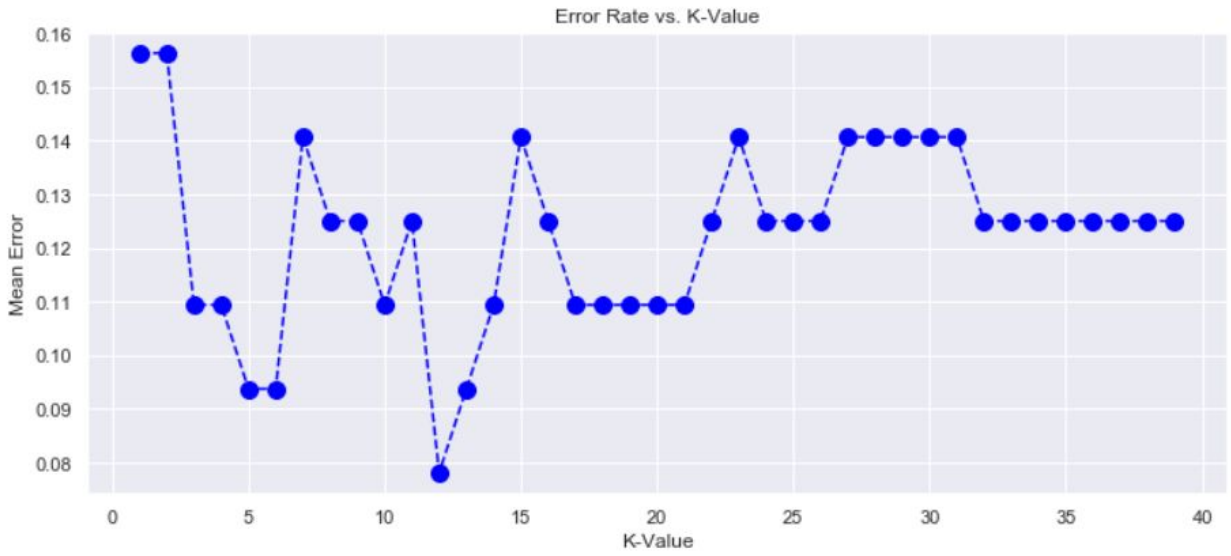


Figure 5.4 Confusion Matrix and Classification Report for the k-NN Classifier (k = 12).

```
The Confusion Matrix for the k-NN (k=12) Classifier is:

 [[30  2]
 [ 3 29]]

The Classification Report for the k-NN (k=12) Classifier is:

              precision    recall  f1-score   support

         0.0       0.91      0.94      0.92        32
         1.0       0.94      0.91      0.92        32

    accuracy                           0.92        64
   macro avg       0.92      0.92      0.92        64
weighted avg       0.92      0.92      0.92        64
```

(f) The precision and f1-score for the `before1980` = 1 response for the k-NN classifier with k = 12 are higher than they are for the logistic regression classifier. The results were largely the same, except the k-NN classifier correctly classified one more house as `before1980` = 0. Therefore, I would recommend the k-NN classifier with k = 12 for

Kyle Aitken
kyle.aitken@rutgers.edu

predicting whether or not a house was built before 1980 over the logistic regression classifier.

6. Extra Credit: Creative Exercise

In 5(f), it was determined that a k-NN classifier with k = 12 performed better as a classifier for the `before1980` variable than a logistic regression classifier. In Section 6, we are going to compare a k-NN classifier with k = 12 against linear discriminant analysis and quadratic discriminant analysis classifiers.

The training and test data for the discriminant analysis models are the same that were used for both the logistic regression classifier and the k-NN classifier. The predictor variables also underwent the same `StandardScaler()` normalization.

The results for the linear discriminant analysis are shown in Figure 6.1 and the results for the quadratic discriminant analysis are shown in Figure 6.2. Immediately, we can see that the LDA classifier has an accuracy of 94%, which is slightly higher than that of the k-NN (k = 12) classifier. Conversely, the QDA classifier has an accuracy of 66%, which is significantly lower than that of the k-NN (k = 12) classifier.

It is safe to say that the LDA and k-NN classifiers are much better predictors of the `before1980` variable than the QDA classifier. Furthermore, it is apparent that the discriminant analysis did not benefit from the added complexity of the QDA model, suggesting that there is an approximately linear decision boundary.

Without any context for the models and the purpose they serve, it is easy to say that the LDA classifier is better than the k-NN classifier, in that it has a higher accuracy and precision score for the `before1980` = 1 variable. However, if the goal of the model is to identify houses that potentially have asbestos, the recall of the model would be a more important statistic to look at than accuracy or precision. In this case, a false positive error is much less harmful than a false negative error because the purpose of the model is to alert as many homeowners/buyers as possible that their house was built before 1980 and potentially has asbestos. So, if this were the purpose of the model, although the LDA classifier has a higher accuracy, the two models have a very similar performance in terms of alerting homeowners/buyers that their homes may have asbestos (their recall scores are equivalent).

Kyle Aitken
kyle.aitken@rutgers.edu

Also, before jumping to conclusions about the relevance of the LDA and k-NN classifiers, it is important to remember that the models were tested on a very small set of houses (only 64 houses in the test dataset). The outputs of the two models only differed by one classification, and the accuracy of the two models may be very different when applied to a larger set of data.

However, in the context of this problem, the LDA classifier had better or equal accuracy, precision, and recall scores than the k-NN (k = 12) classifier and I would therefore recommend the LDA classifier for predicting the `before1980` variable.

Figure 6.1 Confusion matrix and summary statistics for the LDA classifier.

```
The Confusion Matrix for the LDA Classifier is:

 [[31  1]
 [ 3 29]]

The Classification Report for the LDA Classifier is:

              precision    recall  f1-score   support

         0.0       0.91      0.97      0.94        32
         1.0       0.97      0.91      0.94        32

    accuracy                           0.94        64
   macro avg       0.94      0.94      0.94        64
weighted avg       0.94      0.94      0.94        64
```

Figure 6.2 Confusion matrix and summary statistics for the QDA classifier.

```
The Confusion Matrix for the QDA Classifier is:

 [[27  5]
 [17 15]]

The Classification Report for the QDA Classifier is:

              precision    recall  f1-score   support

         0.0       0.61      0.84      0.71        32
         1.0       0.75      0.47      0.58        32

    accuracy                           0.66        64
   macro avg       0.68      0.66      0.64        64
weighted avg       0.68      0.66      0.64        64
```