STAT 6021: Project 2 - Report

Group 10:
Hana Nur (hrn4ch), Fadumo Hussein (fmh7pv), Brendan Puglisi (btp6ht), Matt Rubin(jsd8ar)

7/14/2023

Section 1: Summary

The journey of homeownership is full of emotions. Excitement, nervousness, and even fatigue are emotions that potential homeowners can feel while searching for their forever home. But how do we choose the best house possible? Searching for a home takes a lot of time and consideration, and we want to ensure homeowners make the best decision. Using the information gathered in King County, Washington from May 2014 - 2015, we wanted to assist homeowners on their journey to homeownership. We built two models to help homeowners make the best decision possible. The first model is a linear regression that explores whether we can build a model to reliably predict the prices of dwellings across King County. From our final model for this question, we discovered that grade was the most influential factor in housing price. As a result, in our second question, we wanted to use a classification model to ask whether we can assemble a model to determine the likelihood of a high-grade home, which is defined by its quality level of construction and design with the other variables in the data set.

When building our linear model, we determined that this formula was the best at predicting the price of homes.

$$\hat{y} = 21.57 + 0.0001617x_{sqft-living} + 0.2252x_{grade} - 0.005502x_{yr-built} + 0.2764x_{waterfront} + 0.05068x_{view} + 0.08522x_{bathrooms} - 0.02395x_{bedrooms} - 0.000000264x_{sqft_lot15} + 0.08644x_{floors} + 0.04866I_{condition} + 0.01858I_{yr-renovated}$$

We are confident in this model, given how close our predicted prices are to the actual prices in the test data. When we built the classification model, we were interested in predicting the likelihood of homes having a high grade, since grade was important in creating the linear model. Here is the formula for the logistical model:

$$log(\pi hat/1-\pi hat) = -53.865248 + 0.573404(bedrooms) + 1.249434(floors) + 1.594117(waterfront) + 0.993799(I1) + 0.025331(yrbuilt)$$

We are confident in this model, given the error rate in the classification of a home grade is only 25%. The journey of real estate is a complex one. However, with these models, we hope to ensure the residents of King County, Washington that their experience will be full of excitement and relief.

Section 2: Description of Data

The data set that our group worked on contains house sale prices for King County, Washington, which includes Seattle. It comprises homes sold between May 2014 and May 2015. Before building our model, we looked at the variables given to us. We decided to remove four variables from our model. These variables are ID, Date, Longitude (long), and Latitude (lat). Our group believed these variables were not meaningful to the question at hand. Also, we mutated the variable condition to be a conditional variable with a factor of either ‘good’ or ‘bad’ based on whether the apartment condition earns below or above a rating of 3. We also mutated the yr_renovated variable to be categorical, where ‘no’ meant the house was never renovated and ‘yes’ represented that a house has undergone a renovation. For the logistic regression, we needed to cut grade for it to be a binary categorical variable. Every grade between 1-7 was deemed as ‘low’ and between 8-13 as ‘high’. Since we used all the variables in our initial model for logistic regression, we also mutated the zipcode variable to be categorical, splitting it between the top 20 ‘wealthy’ zip codes and the rest being ‘not_wealthy’.

Variable Name	Class	Description
price	num	Price of each home sold
bedrooms	int	Number of bedrooms
bathrooms	num	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	int	Square footage of the apartments interior living space
sqft_lot	int	Square footage of the land space
floors	num	Number of floors
waterfront	int	A dummy variable for whether the apartment was overlooking the waterfront or not
view	int	An index from 0 to 4 of how good the view of the property was
Linear regression: grade	int	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 has a high-quality level of construction and design. Only used in linear regression.
Logistic regression: grade - **Mutated variable**	Factor w/ 2 levels: "low" and "high"	Taking the grade variable used in linear regression and creating a factor from x<7.1 being low and x>7.1 being high. Only used in logistic regression.
sqft_above	int	The square footage of the interior housing space that is above ground level
sqft_basement	int	The square footage of the interior housing space that is below ground level

yr_built	int	The year the house was initially built
yr_renovated - **Mutated variable**	Factor w/ 2 levels: "no" and "yes"	Originally the year of the house's last renovation, we altered this variable to be a conditional variable dictating if the home was renovated or not using factors "no" and "yes".
Linear regression: zipcode	int	What zip code area the house is in
Logistic regression: zipcode - **Mutated variable**	Factor w/ 2 levels: "Wealthy" and "Not_Wealthy"	Taking the zipcode variable and splitting it between the top 20 wealthiest zip codes.
sqft_living15	int	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	int	The square footage of the land lots of the nearest 15 neighbors
Condition - **Mutated variable**	Factor w/ 2 levels: "poor" and "good"	Originally an index from 1 to 5, we mutated the condition to be a conditional variable with a factor from x<3.1 being poor and x>3.1 being good.

Section 3: Research Questions

For our linear regression model, our question of interest had two parts. First, we wanted to investigate the following question: can we build a housing model that could reliably predict the prices of dwellings across King County? Perhaps the top aspect of considering the use of a model is its efficacy. A model with a high error rate and low R^2 value can prove too weak to allow an accurate interpretation of the data. Our approach wished to find statistically significant patterns among county housing that can be used to predict price, our linear response variable, for a booming market. We had to determine whether our predictors were sufficiently correlated with prices that it was feasible for our model to use them as a predictive tool. The second area we

wished to explore as part of our linear analysis was which predictors had the greatest impact on prices throughout the county between 2014 and 2015. While many factors may have some influence over price, we sought to determine which ones have the strongest relationships with price so that we can weigh our model accordingly. Furthermore, we were curious whether links between housing aspects and prices that may not be inherently intuitive would emerge.

For our logistic regression model, we sought to study how the grade of a housing unit in King County can be influenced by the range of variables found within our model. Grade is arguably the paramount factor for individuals seeking to buy or sell real estate due to its profound impact on quality of life. Houses with poor grades may prove risky as their owners may need personal expenses to repair failing elements of the structure. Grade can be turned into a binary variable by separating housing units into low and high grades based on whether they were above or below the median. We feel that using grade as the response variable for our logistic regression will allow us to closely study which predictors tend to correlate with each level, which could help consumers make more informed choices about purchasing housing within the county.

Section 4: Visualizations for Linear Regression

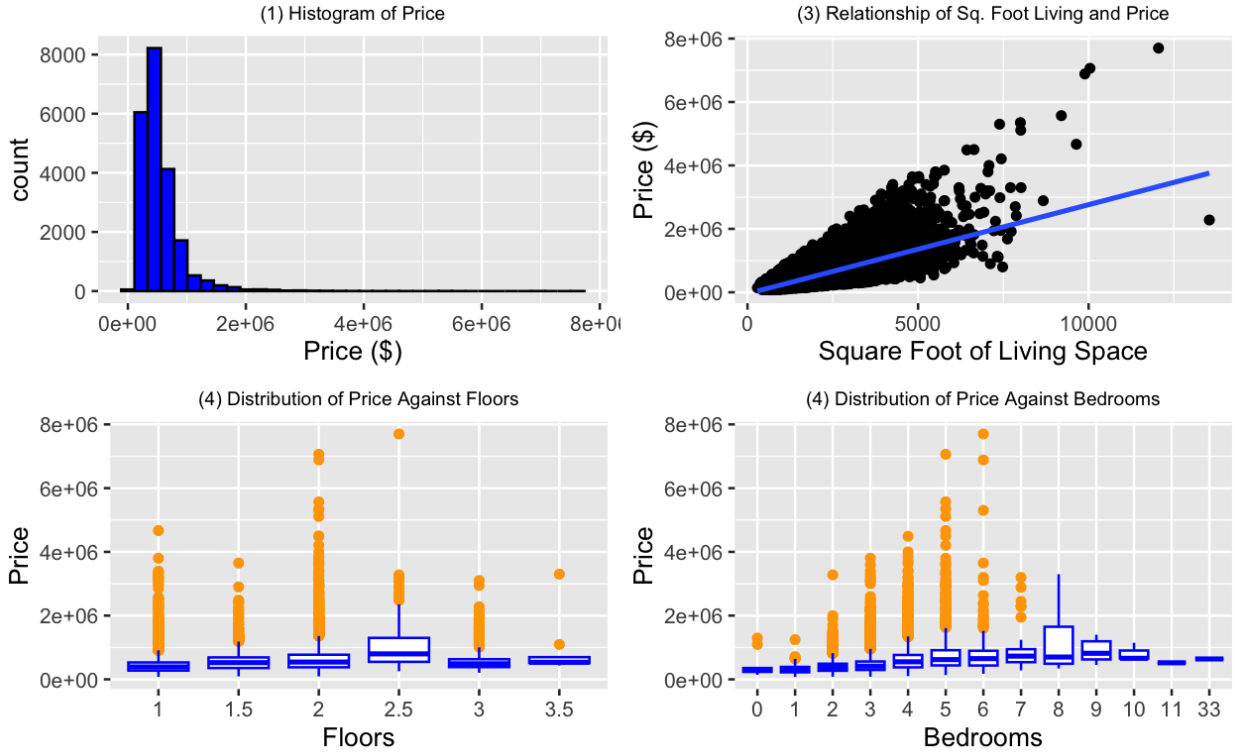


Figure 1

The graph in the top left of Figure 1 is a histogram of the price variable. This graph shows the distribution of houses from $75,000 to $7,700,000. This right-skewed graph shows us that the average ($540,088.1) is greater than the median ($450,000) price. The majority of homes are between $75,000 and $20,000,000.

The figure on the top right is a simple scatter plot that helps to further explore the relationship between the price of a house and the square foot of living space of a house. There is a strong positive correlation between both variables. We hope to analyze this relationship in our project.

The figure in the bottom left is a box plot of the distribution of price against the number of floors of a house. As the number of stories increases, there is an increase in price. However, this model shows that several homes are outliers for each number of floors. Houses with two

floors seem to have the most. The general expectation of homes is that if the number of floors increases, then the price should increase as well. However, homes above $4,000,000 primarily are 2 and 2.5 story homes. This raises the question of what other factors are going into the cost of a house given the number of floors a home has.

The figure on the bottom right is a box plot showing the distribution of price against the number of bedrooms in a home. As the number of bathrooms increases, the price also increases overall. There is an increase in spread as the number of bathrooms increases as well. This generally is a consistent idea that the more bedrooms you have in a home, the more expensive it is. However, there are outliers in several box plots in the model, raising the question of what other factors are going into the price.

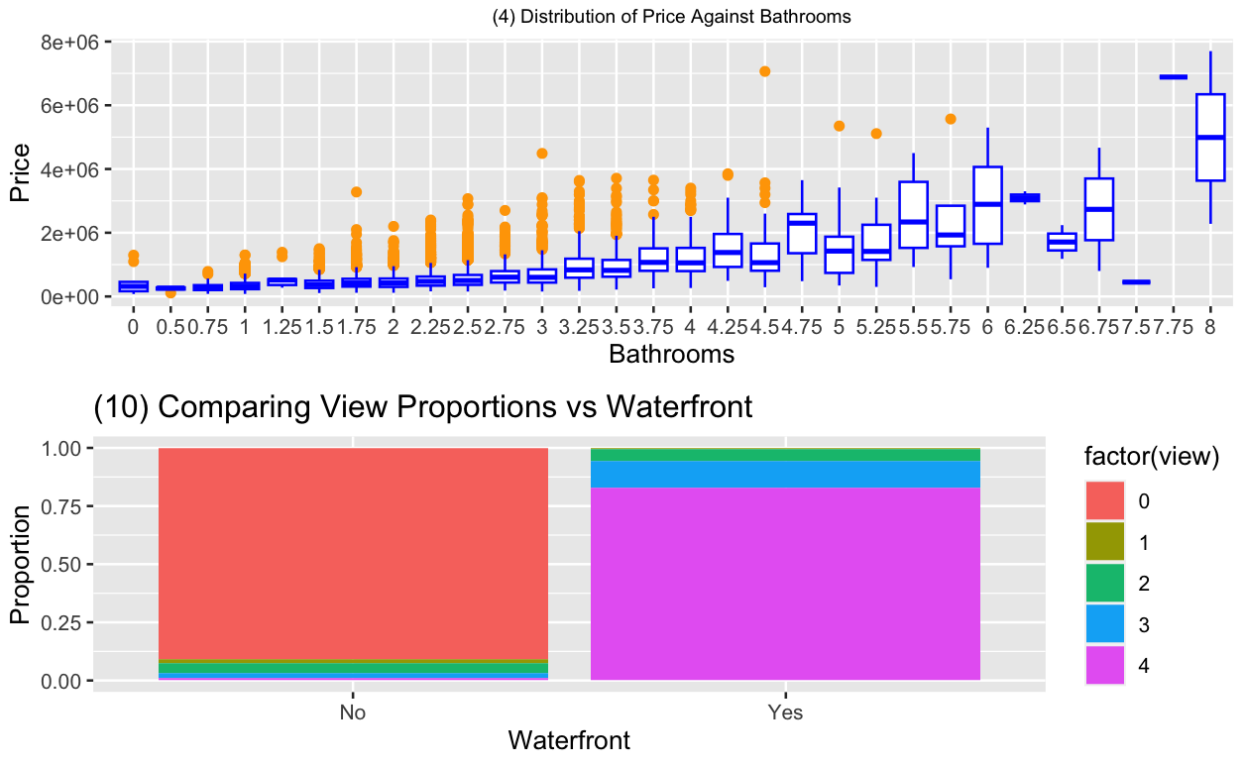
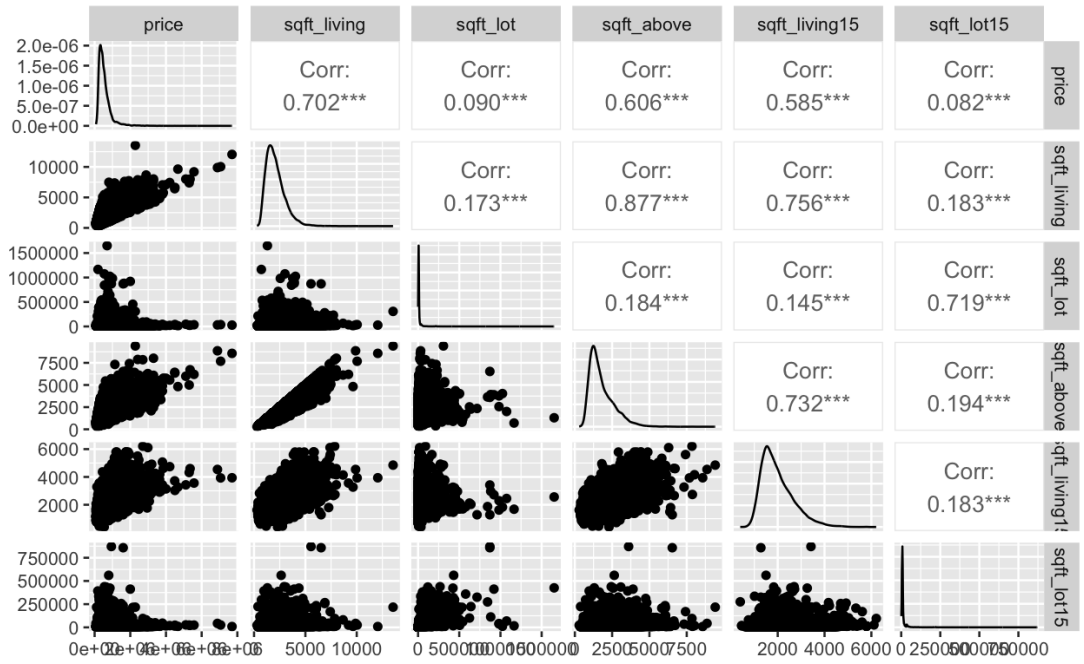


Figure 2

The top graph in Figure 2 is a box plot that shows the distribution of price against the number of bathrooms in a home. As the number of bathrooms increases, the price does as well.

There also seems to be more spread as the number of bathrooms increases. This generally is a consistent idea that the more bathrooms you have in a home, the more expensive it is. However, there are outliers in several box plots in the model. This raises the question of what other factors go into the cost of a house given the number of bathrooms a home has.

The bottom graph in Figure 2 is a bar graph comparing the view of homes with and without a waterfront. Homes with a waterfront have higher-rated views than homes without one. Homes' views without a waterfront primarily were rated as zero, and homes with a waterfront view were rated as four. This graph generally makes sense as homes with a waterfront tend to have higher-rated views than homes without one**.**



The off-diagonal entries of the output give us the scatterplot and correlation between the corresponding pair of quantitative variables. Variables: sqft_living and price, sqft_above and price, sqft_living_15 and price, sqft_living_15 and sqft_living, sqft_lot15, and sqft_lot,

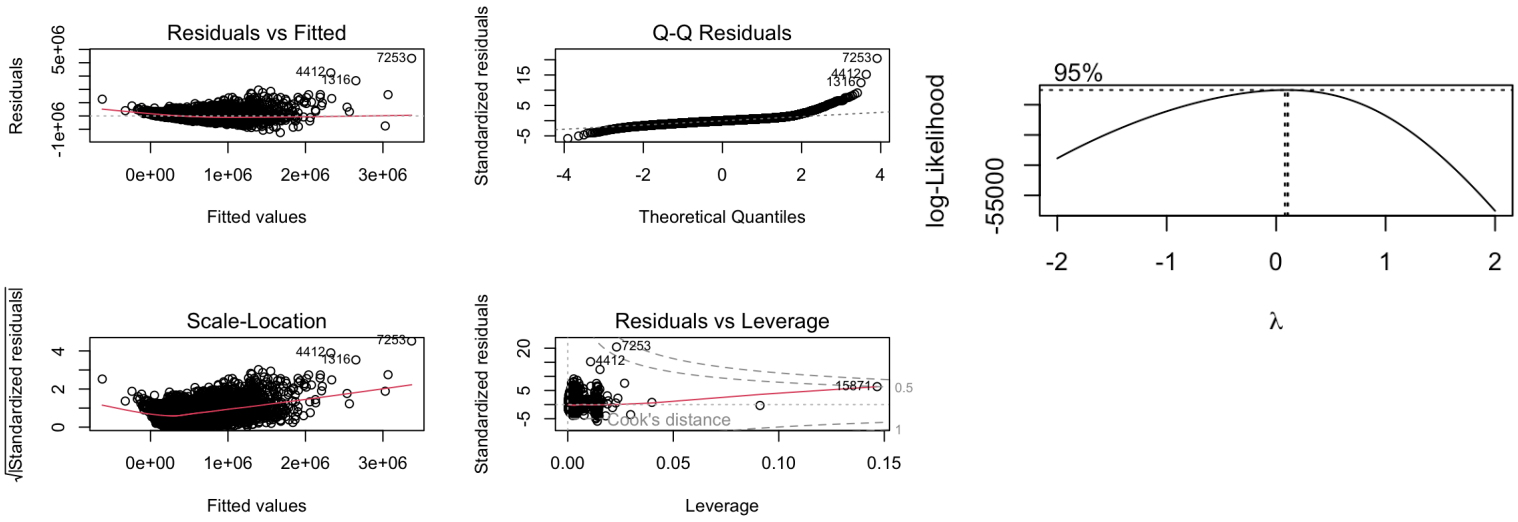
sqft_living_15 and sqft_above all have high correlations with sqft_above and sqft_living having the highest (0.877). Most of the distribution is right-skewed, similar to the price histogram in the first figure.

Section 5: Linear Regression

We utilized forward, backward, and stepwise selection to determine a starting point for our linear regression models. All selection methods resulted in identical models with adjusted R^2 of 0.6526 and mean squared error of 1.0999. We also evaluated the models with the best adjusted R^2, BIC, and CP. These "best subset" methods all resulted in the same model, which had a greater mean squared error, 1.27, than the previously identified models. We decided to start with the model determined by the step selection methods for our linear regression as it had the lowest mean squared error. Because the step selection methods do not consider interaction, we attempted to include interaction in the model. All interaction predictors resulted in t statistics with large p-values, indicating they are insignificant, so our model includes only additive conditional predictors.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.399e+06  1.915e+05  33.413  < 2e-16 ***
sqft_living             1.635e+02  6.381e+00  25.618  < 2e-16 ***
grade                   1.227e+05  3.119e+03  39.329  < 2e-16 ***
yr_built               -3.661e+03  1.001e+02 -36.572  < 2e-16 ***
waterfront              5.048e+05  2.580e+04  19.565  < 2e-16 ***
view                    3.946e+04  3.112e+03  12.681  < 2e-16 ***
bathrooms               4.812e+04  4.819e+03   9.985  < 2e-16 ***
bedrooms               -3.574e+04  2.756e+03 -12.969  < 2e-16 ***
sqft_lot15             -5.493e-01  7.550e-02  -7.275 3.69e-13 ***
floors                  3.442e+04  5.294e+03   6.501 8.35e-11 ***
condition_cgood         2.696e+04  4.853e+03   5.554 2.85e-08 ***
sqft_living15           2.789e+01  5.058e+00   5.514 3.58e-08 ***
sqft_above             -1.553e+01  6.312e+00  -2.460   0.0139 *
yr_renovatedrenovated   2.670e+04  1.087e+04   2.457   0.0140 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 214400 on 10792 degrees of freedom
Multiple R-squared:  0.653,     Adjusted R-squared:  0.6526
F-statistic:  1563 on 13 and 10792 DF,  p-value: < 2.2e-16
```



The first assumption is that errors have a mean of zero. A by-product of this statement is that the relationship between y and x, as expressed via $y \approx f(x)$, is correct, meaning our model should be linear to not violate this assumption. The scatterplot of the residuals vs. fitted is nonlinear but has an increasing variance. The data points are generally not evenly scattered on both sides of the overlaid regression line, meaning that this assumption is violated. A major consequence of violating this assumption is that predicted values will over- or underestimate the

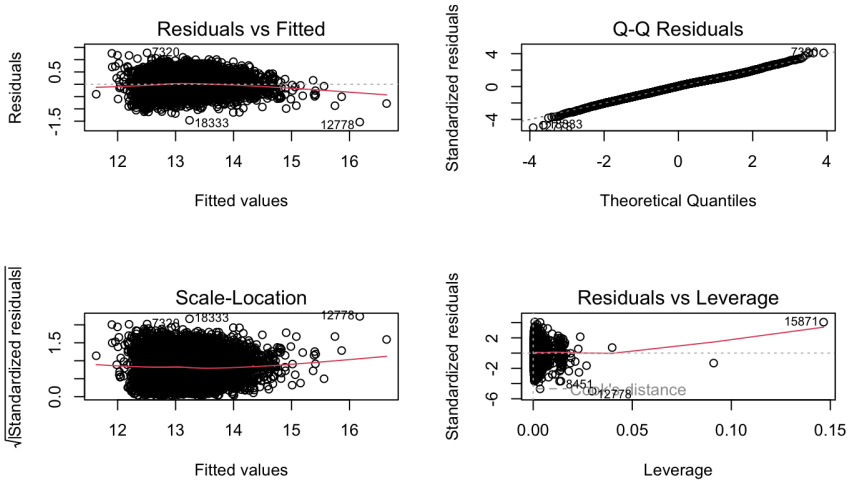
true values of the response variables. The residual plot shows that the model will overpredict the price of houses.

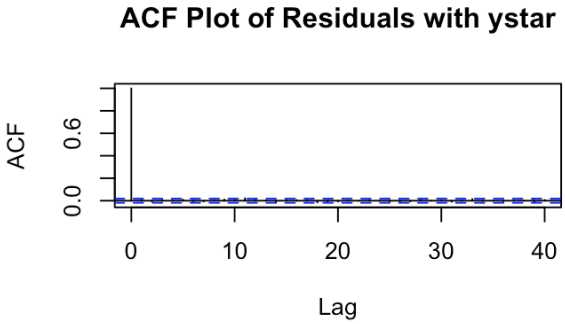
The second assumption is that errors have a constant variance. A by-product of this statement is that a scatterplot has the same vertical variation of data points around the regression equation and has the same magnitude everywhere. In the Scale-Location graph, the vertical variance increases as the response variable gets larger, meaning that variance is not constant. A major consequence of violating this assumption is statistical inference is not reliable. This raises issues as any hypothesis test, calculation of confidence and prediction intervals of the prices of houses are all unusable.

The fourth assumption is assessed using the normal probability plot (also called a QQ plot). If the residuals are normal, the residuals should fall along the 45-degree line. The residuals on both ends of the graph do not fall on the line.

The Residual vs. Leverage plot is used to identify influential outliers. Data points that lie in the contour lines with large Cook's distance are influential. Residual point 40 of our data points have Cook's distance greater than 0.5. However, none of the observed points are greater than 1, meaning we have no influential variables in the model.

The Boxcox plot shows a lambda very close to but not exactly zero. We chose to do a logarithmic transformation despite this so that we could interpret the results. Once we completed a logarithmic transformation, the residual plots demonstrated that the assumptions were met.





Observations are evenly distributed above and below the residual plot line, indicating assumption one is met. Additionally, variance is much more constant as there is no longer an increase in variance as the fitted values increase, indicating the model meets assumption two. The residuals closely follow the Q-Q plot line, showing assumption four is also met.

The ACF plot demonstrates the residuals are independent, as there is no significant ACF for values after lag 0. Therefore, the residuals are independent, and assumption three is met. Because all assumptions are met, we moved on to hypothesis testing to determine all necessary predictors and if we may reduce our model.

We began with an ANOVA F test to determine if our model is better than an intercept-only model. Our null hypothesis is that all coefficients are zero, meaning all predictors can be dropped, and our alternative hypothesis is that at least one coefficient is non-zero, meaning predictors cannot be dropped. The F statistic is 1563 with a p-value of nearly zero, as seen in the model summary. Therefore, we reject the null hypothesis, and conclude our model is more useful than an intercept-only model.

Due to the high number of predictors in the model summary, we ran a general linear F-test to see if we can drop the three predictors with the greatest p-values.

H_0: $\beta_{sqft-living15} = \beta_{yr-renovated} = \beta_{sqft-above} = 0$; H_a: *at least one coefficient in* H_0 *is not zero*

The null hypothesis is that all coefficients for the predictors we wish to drop are zero, meaning they can be dropped. The alternative hypothesis is that the coefficients of the predictors are non-zero, meaning they are significant and cannot be dropped.

The F statistic, which we determined to be 12.42197, is greater than the critical value, 2.605732, and the p-value, 4.167185e-08, is near zero. Therefore, we reject the null hypothesis and conclude that the data supports the alternative hypothesis, meaning we cannot drop these predictors and must use the full model.

The model predictors are significant according to both the general linear F-test results as well as the t-test results for individual predictors, but our contextual understanding of the predictors indicates that there may be some linear dependence between variables. For example, the sqft_above predictor indicates the number of square feet a home has above the basement floor. Depending on the presence of a basement, this may be equal to the square feet of the home interior, sqft_living. Due to this potential source of collinearity, we decided to investigate the model further to determine if any variables are linearly dependent on others.

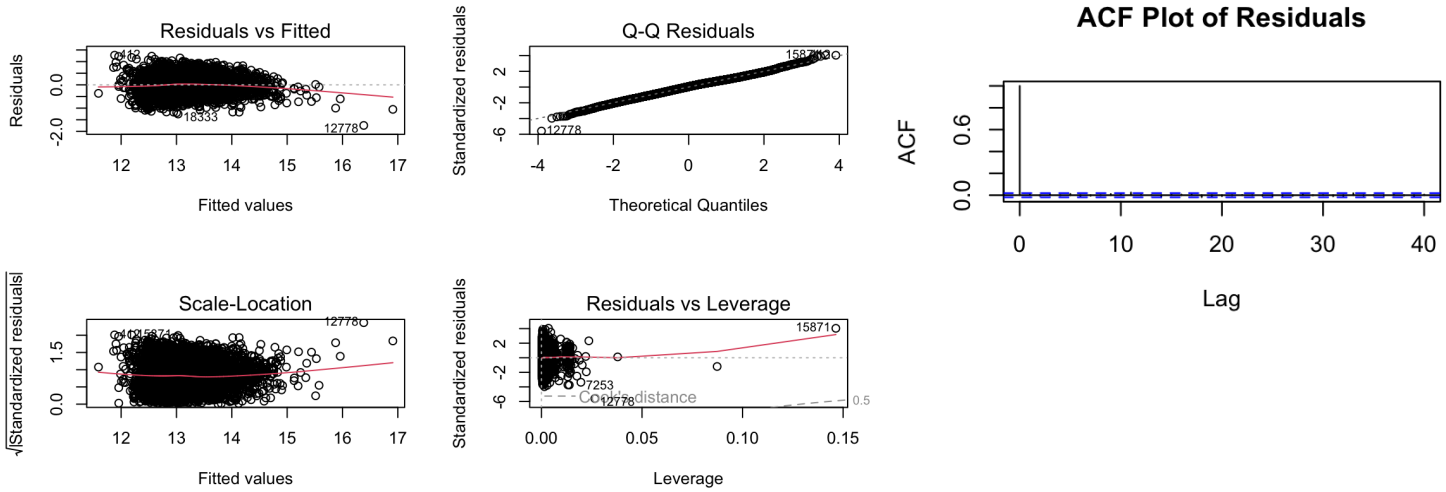
sqft_lot15	yr_renovatedrenovated	waterfront	condition_cgood	view
1.072672	1.149874	1.207119	1.252640	1.406005
bedrooms	floors	yr_built	sqft_living15	grade
1.594562	1.931336	2.010447	2.803321	3.173848
bathrooms	sqft_above	sqft_living		
3.298012	6.365248	8.186901		

We first compare the VIFs of the predictors shown in the figure above. Several are above 5, indicating the presence of some degree of multicollinearity in the model. Because sqft_living has the highest VIF, we keep this predictor and remove the predictors that are highly correlated

with sqft_living to resolve any multicollinearity. The predictors sqft_above and sqft_living15 are both highly correlated with sqft_living, so we remove both sqft_above and sqft_living15 from the model.

sqft_lot15	yr_renovatedrenovated	waterfront	condition_cgood	view	floors
1.058249	1.147694	1.205394	1.250055	1.350628	1.585717
bedrooms	yr_built	grade	bathrooms	sqft_living	
1.593921	1.981811	2.841952	3.197249	4.096389	

The resulting model has predictors with VIFs all below 5, and all predictors have t-test statistics with p-values indicating they are significant. This indicates that the multicollinearity among predictors has been removed, and the predictors are now linearly independent.



The residual plot demonstrates assumption one is met as there are an even number of observations above and below the regression line. Additionally, assumption two is also met as there is constant variance. We conclude that assumption four is met as the residuals closely follow the Q-Q plot line, indicating the errors follow a normal distribution. The ACF plot demonstrates assumption three is met, as the ACF for values after lag 0 are all below the critical value.

There are a number of observations with a DFBETAS and DFFITS values marginally greater than the suggested values, $\frac{2}{\sqrt{n}}$ and $2\sqrt{\frac{p}{n}}$ respectively. Similarly, several observations have externally studentized residuals that narrowly exceed the suggested value, $\frac{2p}{n}$. However, the Cook's distance for all observations is below 1, with the greatest Cook's distance being 0.23292 (seen as observation 15871 in the Residuals vs Leverage plot above). Therefore, we conclude that there are no influential observations in our final linear regression model.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.157e+01  2.795e-01  77.191  < 2e-16 ***
sqft_living           1.617e-04  6.649e-06  24.313  < 2e-16 ***
grade                 2.252e-01  4.348e-03  51.788  < 2e-16 ***
yr_built             -5.502e-03  1.464e-04 -37.569  < 2e-16 ***
waterfront            2.764e-01  3.799e-02   7.275 3.71e-13 ***
view                  5.068e-02  4.493e-03  11.279  < 2e-16 ***
bathrooms             8.522e-02  6.990e-03  12.191  < 2e-16 ***
bedrooms             -2.395e-02  4.059e-03  -5.899 3.77e-09 ***
sqft_lot15           -2.640e-07  1.105e-07  -2.390   0.0169 *
floors                8.644e-02  7.068e-03  12.231  < 2e-16 ***
condition_cgood       4.866e-02  7.142e-03   6.812 1.01e-11 ***
yr_renovatedrenovated 1.858e-02  1.599e-02   1.162   0.2452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3159 on 10794 degrees of freedom
Multiple R-squared:  0.6398,    Adjusted R-squared:  0.6394
F-statistic:  1743 on 11 and 10794 DF,  p-value: < 2.2e-16
```

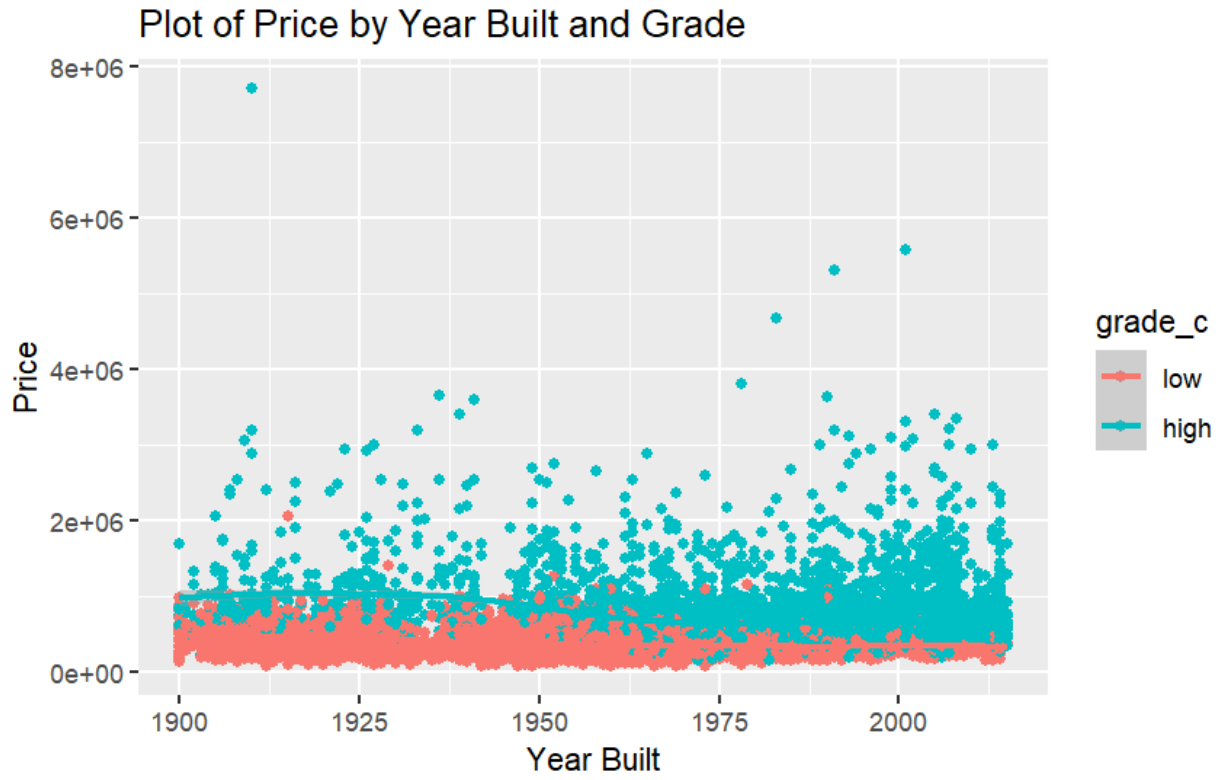
$$\hat{y} = 21.57 + 0.0001617x_{sqft-living} + 0.2252x_{grade} - 0.005502x_{yr-built} + 0.2764x_{waterfront} + 0.05068x_{view} + 0.08522x_{bathrooms} - 0.02395x_{bedrooms} - 0.000000264x_{sqft_lot15} + 0.08644x_{floors} + 0.04866I_{condition} + 0.01858I_{yr-renovated}$$

Our final model has a mean squared error of 1.10 and an adjusted R^2 of 0.6394. These quantities tell us our reduced model fits the data similarly to our original full model as determined by forward selection. Additionally, because multicollinearity has been removed, we can accurately interpret the coefficients to determine the impact of individual predictors on price.

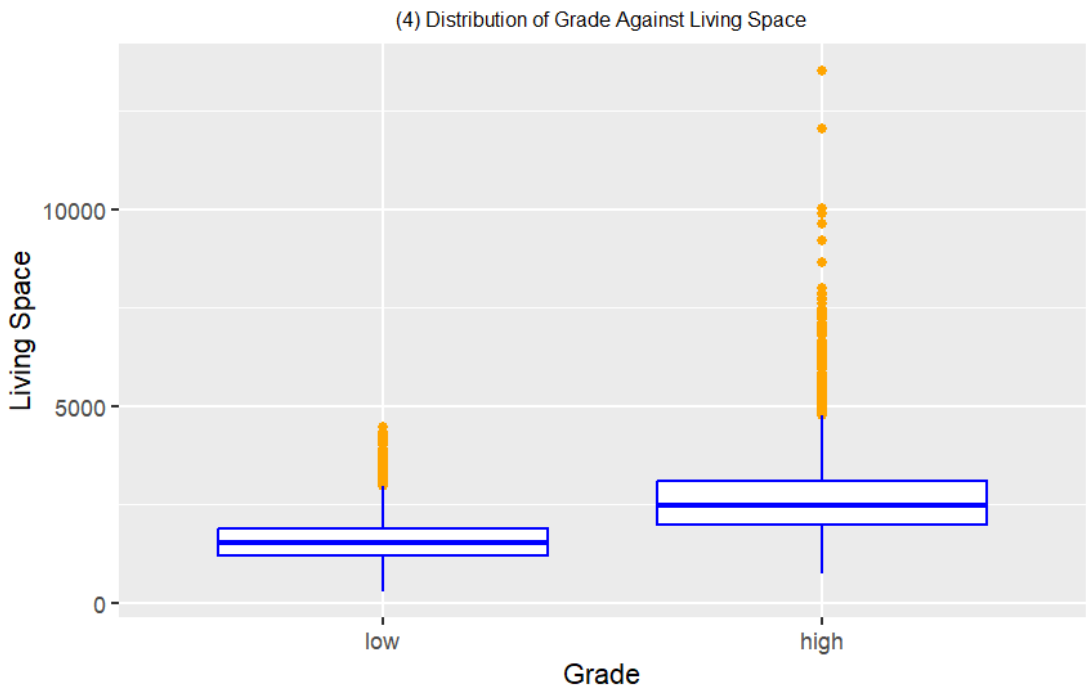
Based on the results from the test data, this model reliably predicts the prices of dwellings across King County.

The predictor with the greatest coefficient in our model is the grade. This indicates that a one-unit increase in the grade value has the highest percent increase in price when holding all other predictors constant. However, when comparing the R^2 of all possible one-predictor models, the model using sqft_living had the best R^2. So, in the presence of other predictors, grade has the greatest impact on the percent increase in price, but sqft_living is the predictor with the greatest linear relationship with the logarithmically transformed price. These two variables are the most highly correlated with price. Interestingly, increasing the number of bedrooms does not increase the percent price, and older homes have a greater effect on percent price than newer homes when holding all other predictors constant.

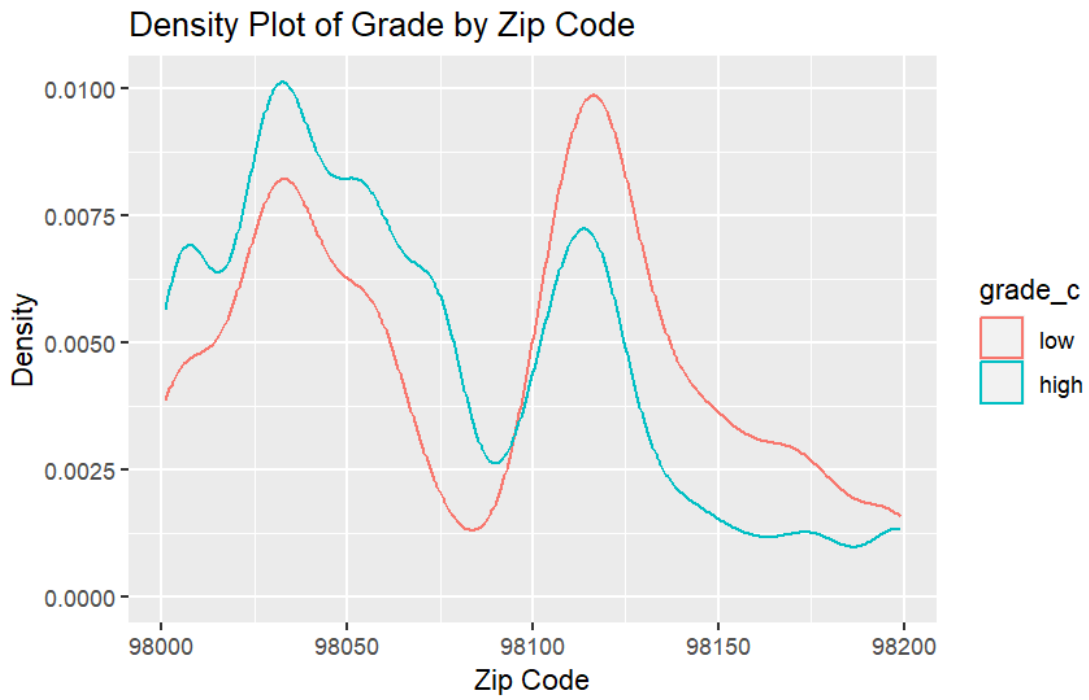
Section 6: Visualizations for Logistic Regression



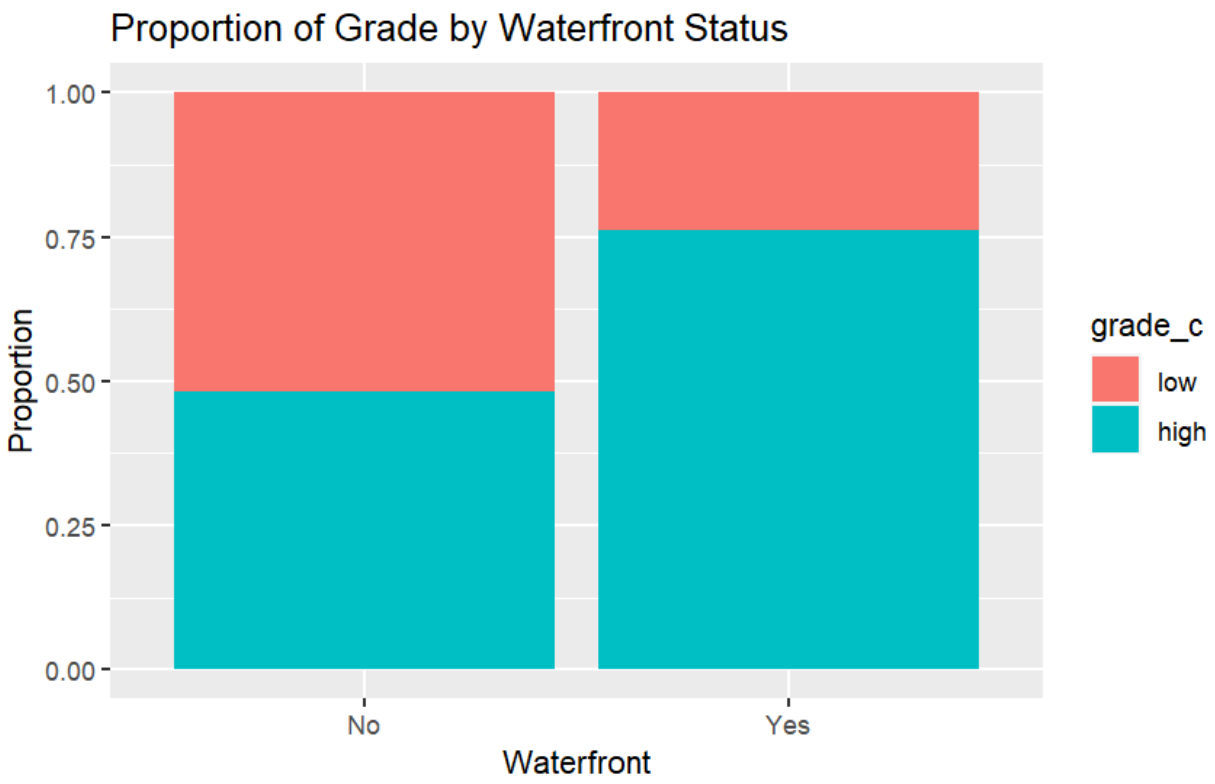
We found an interesting relationship to be present between price, grade, and the construction year of a residence in King County. As time passed, there was a clear increase in the proportion of high-grade houses thanks to growth in technology and building standards. It is also noteworthy that the low-end price of high-grade residences appears to decline significantly over this period. However, we eventually dropped price from our model due in part to the presence of several significant outliers visible on this figure.



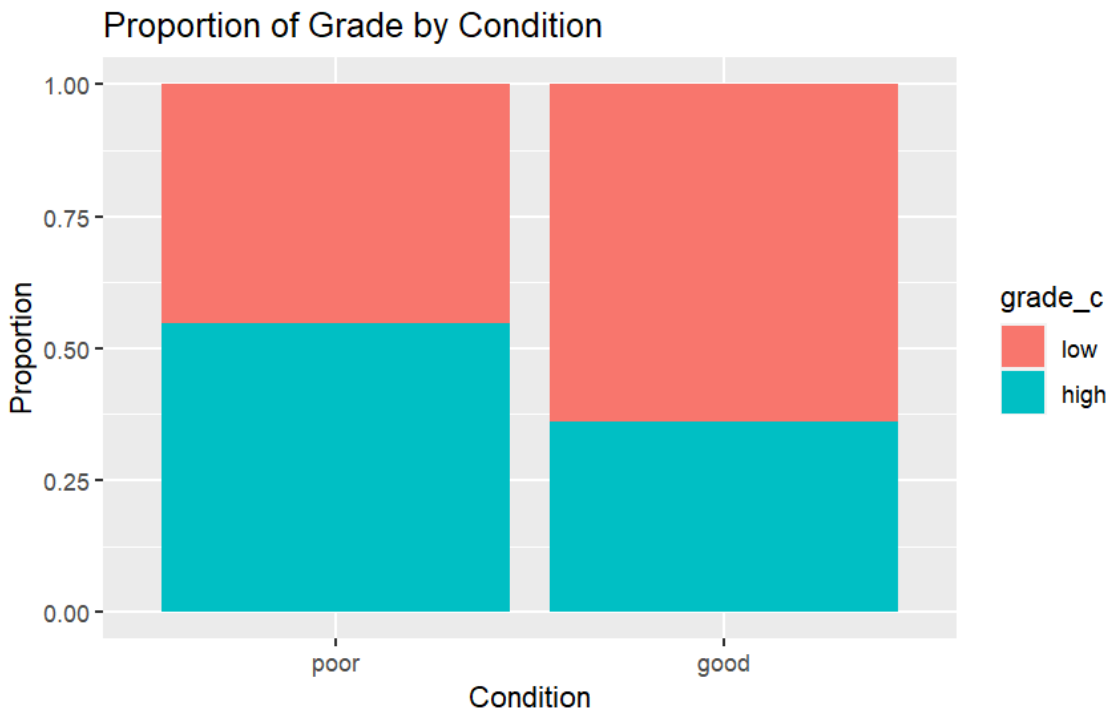
We found that houses with high grades generally have more living space than houses with low grades. It is interesting to note the much wider spread of outliers among the high-grade residences throughout the county. This likely results from the extreme size of the largest and most prestigious houses in the area. Additionally, the lower quartile square footage of high-grade houses is greater than the upper quartile square footage of low-grade residences. Living space, like price, was ultimately judged to not be critical to the logistic regression and was dropped from the model.



This chart displays how the various zip codes throughout King County relate to the grades of housing units. Zip codes correspond to geographic areas within the county, meaning that codes numerically close to one another are usually also geographically close. 98100 represents a clear dividing line for this relationship, with zip codes below that number tending towards higher grade housing and zip codes above that figure generally featuring lower grade residences.



Next, we plotted the proportional relationship between houses on the waterfront and their corresponding grades. Residences near the water tend to have higher grades than those further away from it by a large margin. Even so, the proportions of houses away from the waterfront with low and high grades are roughly even.



Our final visualization depicts a proportion plot between condition and grade, two of the most critical variables for the logistic regression question. Surprisingly, a higher proportion of poor-condition houses are classified as high-grade than the corresponding proportion of good-condition homes.

Section 7: Logistic Regression

In order to answer our logistic regression question of predicting house grade, we will use manual model-building and hypothesis testing to determine the optimal model. First, we check the correlations between the quantitative predictors to better understand their relationships and identify potential collinearity.

	price	sqft_living	sqft_lot	sqft_above	sqft_basement	sqft_living15	sqft_lot15	bedrooms	bathrooms	floors
price	1.000	0.700	0.081	0.601	0.338	0.585	0.072	0.299	0.530	0.257
sqft_living	0.700	1.000	0.167	0.874	0.454	0.754	0.177	0.563	0.755	0.347
sqft_lot	0.081	0.167	1.000	0.170	0.031	0.133	0.710	0.030	0.082	-0.007
sqft_above	0.601	0.874	0.170	1.000	-0.036	0.730	0.182	0.475	0.685	0.520
sqft_basement	0.338	0.454	0.031	-0.036	1.000	0.211	0.030	0.287	0.297	-0.240
sqft_living15	0.585	0.754	0.133	0.730	0.211	1.000	0.177	0.384	0.571	0.278
sqft_lot15	0.072	0.177	0.710	0.182	0.030	0.177	1.000	0.025	0.081	-0.009
bedrooms	0.299	0.563	0.030	0.475	0.287	0.384	0.025	1.000	0.505	0.170
bathrooms	0.530	0.755	0.082	0.685	0.297	0.571	0.081	0.505	1.000	0.495
floors	0.257	0.347	-0.007	0.520	-0.240	0.278	-0.009	0.170	0.495	1.000

Unsurprisingly, a high degree of correlation exists between square footage above ground and square footage of total living space. A modest relationship is also present between square footage of neighboring residences and square footage of living space. Price also correlates somewhat with square footage above ground and living square footage. Interestingly, the square footage of the lot does not have a significant relationship with any variables in our model other than the lot square footage of the neighbors. Bathrooms are highly correlated to living space, but bedrooms do not show a high correlation with any variable.

Now, we will make a logistic regression model with all predictors included and drop insignificant and multicollinear coefficients as needed. We choose this model-building technique as we attempted automatic model selection techniques such as forward selection, but since these were simply suggestions, we decided to build our own because both models had similar results and we were able to use our knowledge of real estate in order to keep the variables we viewed as most important for identifying a house's grade measured by its design and construction. From the visualizations, we see that all the predictors may influence the grade of a house. We use the glm() function to create a logistic regression with all of our variables.

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = grade ~ ., family = binomial, data = train)

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.809e+01  3.273e+00 -20.804  < 2e-16 ***
price               5.700e-06  2.421e-07  23.547  < 2e-16 ***
bedrooms           -3.370e-01  4.577e-02  -7.363 1.80e-13 ***
bathrooms           3.321e-01  7.168e-02   4.633 3.61e-06 ***
sqft_living         6.689e-04  1.026e-04   6.521 6.98e-11 ***
sqft_lot            6.852e-07  1.067e-06   0.642   0.5206
floors              5.387e-01  7.236e-02   7.445 9.73e-14 ***
waterfront         -1.273e+00  5.615e-01  -2.268   0.0233 *
view                1.244e-01  5.297e-02   2.348   0.0189 *
conditiongood      -8.431e-02  7.088e-02  -1.189   0.2342
sqft_above          4.113e-04  9.954e-05   4.132 3.59e-05 ***
sqft_basement              NA         NA      NA       NA
yr_built            3.074e-02  1.656e-03  18.559  < 2e-16 ***
yr_renovatedyes     3.369e-01  1.661e-01   2.029   0.0425 *
zipcodeNot_Wealthy  8.450e-04  7.882e-02   0.011   0.9914
sqft_living15       1.222e-03  8.808e-05  13.874  < 2e-16 ***
sqft_lot15         -7.573e-06  1.638e-06  -4.624 3.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14967.9  on 10805  degrees of freedom
Residual deviance:  7220.4  on 10790  degrees of freedom
AIC: 7252.4

Number of Fisher Scoring iterations: 6
```

As shown by the model summary, a few predictors have highly insignificant coefficients: sqft_lot, condition, and zip code. This is not surprising; since we have so many variables in the model, there are bound to be several that are correlated, as seen in the correlation matrix above, resulting in insignificant results. For example, sqft_lot has a relatively high correlation to sqft_lot15, meaning we would only potentially need one of them in our model. Insignificant values could also signify that the predictor variable is not related to our target of grade. Either way, we chose to drop these three variables along with sqft_basement since it has NA's for its coefficients potentially due to multicollinearity. It is also important to note the warning of "fitted probabilities numerically 0 or 1 occurred" potentially indicating perfect separation in our data. However, since based on our question, we are simply using logistic regression for prediction, we can ignore this warning for now.

In likelihood ratio tests, the test statistic is the difference in the deviances of the two models. In this circumstance, the null hypothesis supports dropping the insignificant predictors, and the alternative hypothesis supports not dropping the four predictors.

H0: sqft_basement = sqft_lot = condition = zipcode = 0

Ha: at least one of the coefficients in the null hypothesis is not 0.

The ΔG2 test statistic is 1.813891, and is compared with a chi-squared distribution with 4 degrees of freedom since the degrees of freedom is equal to the number of terms we are dropping. The p-value is 0.7699399, meaning we fail to reject the null hypothesis. The data do not support using the full model with all the predictors, so we drop sqft_basement, sqft_lot, condition, and zip code.

Now we assess variance inflation factors (VIFs) in logistic regression to potentially identify multicollinearity before we drop any variables.

price	bedrooms	bathrooms	sqft_living	floors	waterfront	view	sqft_above
69.13506	20.18866	33.48673	96.15422	16.10329	26.14000	18.37760	72.56224
yr_built	yr_renovatedyes	sqft_living15	sqft_lot15				
22.46424	11.93195	38.22428	10.75162				

Due to the sheer number of variables in our model, it is not surprising that we have significant evidence of multicollinearity. Now we will attempt to manipulate and drop variables from our model to reduce multicollinearity and increase the significance of the variables present. Due to these extremely high levels of multicollinearity, the coefficients and p-values for these terms cannot necessarily be trusted, and as a result, we must reduce this multicollinearity in our model to avoid high variance with the estimated coefficients. Also, reducing the overall parameter count will help prevent overfitting.

Analyzing the variables from the previous model, we notice that price and sqft_living have massive VIF values. Using our background knowledge of real estate we recognize that both of these variables are highly correlated with several of the variables in the model as well as each other. For example, as sqft_above increases, sqft_living will also increase, and the corresponding price of a home increases. In light of this, we can either drop every variable related to price and sqft_living or simply drop price and sqft_living. Keeping our prediction variable of grade in mind, we decided to drop price and sqft_living as we think the other variables might give more insight into the actual design and structure of the house at hand, which are directly related to its grade.

Similar to the linear regression model, we will drop sqft_above and sqft_living15 due to high correlation with other variables that can explain these predictors. We also drop view for similar reasons since waterfront and floors can describe this variable. Similarly, we choose to keep bedrooms over bathrooms as bedrooms and floors will represent this variable since they are typically related. Also, we will drop sqft_lot15, as it is a characteristic of other houses that will not help us predict the grade of the particular house we are looking at.

Now we will build our final model to predict the grade of a house based on the predictors bedrooms, floors, waterfront, yr_renovated, and yr_built.

bedrooms	floors	waterfront	yr_renovatedyes	yr_built
7.496647	7.746026	7.258355	5.757404	8.838470

We see all of our VIFS for our model are between 5 and 10, meaning they show some signs of multicollinearity but are not necessarily identified as having a high level of it. Since our

question revolves around predictions, we can move forward with these predictors and resulting VIF levels since predictions can remain unbiased even in the presence of multicollinearity. Also, it is critical to note that 2 out of 5 of our predictors, waterfront and yr_renovated, are binary categorical variables, so VIFs do not apply to them. Let us take a look at the estimated coefficients for our reduced model:

```
Call:
glm(formula = grade ~ bedrooms + floors + waterfront + yr_renovated +
    yr_built, family = binomial, data = train)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -53.865248   1.915823 -28.116  < 2e-16 ***
bedrooms          0.573404   0.027870  20.574  < 2e-16 ***
floors            1.249434   0.049452  25.266  < 2e-16 ***
waterfront        1.594117   0.295103   5.402 6.59e-08 ***
yr_renovatedyes   0.993799   0.113398   8.764  < 2e-16 ***
yr_built          0.025331   0.000979  25.873  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14968  on 10805  degrees of freedom
Residual deviance: 11307  on 10800  degrees of freedom
AIC: 11319

Number of Fisher Scoring iterations: 4
```

So our logistic regression equation is:

$$log(\pi hat/1-\pi hat) = -53.865248 + 0.573404(bedrooms) + 1.249434(floors) + 1.594117(waterfront) + 0.993799(I1) + 0.025331(yrbuilt)$$

Where $I_1 = 1$ if the home was ever renovated.

Given that all these coefficients are positive, bedrooms, floors, yr_built, waterfront, and a renovated home are associated with higher likelihood of having a high grade.

- The odds of a high-grade house are multiplied by exp(0.573404) = 1.774296 for each additional bedroom when controlling for number of floors, waterfront, renovation, and year built.

- The odds of a high-grade house are multiplied by exp(1.249434) = 3.488368 for each additional floor when controlling for number of bedrooms, waterfront, renovation, and year built.
- The odds of a high-grade house are multiplied by exp(0.025331) = 1.025655 for each later year it was built when controlling for number of bedrooms, waterfront, renovation, and floors.
- The odds of a high-grade house having been renovated is exp(0.993799) = 2.701478 times the odds for a house that has not been renovated when controlling for number of bedrooms, waterfront, year built, and floors.

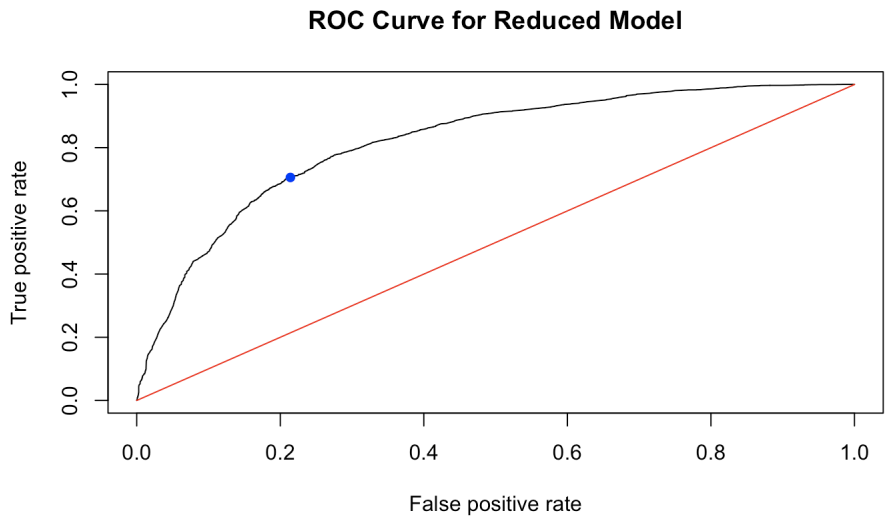
We see that all variables are highly significant, have small standard errors, and have the correct sign for their coefficients, proving our model is useful and lacks high multicollinearity. Now, we use the model to estimate the predicted probabilities of the test data and then use a threshold of 0.5 to classify the test data.

	FALSE	TRUE
low	4486	1222
high	1501	3598

Based on the confusion matrix above, the true negative (TN) is 4486, representing houses classified as low-grade by the model that truly were low-grade. The false positive (FP) is 1222, representing houses that were classified as high-grade but were truly low-grade. The false negative (FN) is 1501, representing houses that were really high-grade but were incorrectly classified as low-grade. The true positive (TP) given by the model is 3598, which is the number of houses that are high-grade and classified as high-grade correctly.

- The error rate of this model and threshold is 0.2519663, meaning 25% of the classifications were incorrect.
- The accuracy of this model and threshold is 0.7480337, meaning 75% of the classifications were correct.
- The False Positive Rate (FPR) is 0.2140855, meaning 21% of the true low grades were incorrectly classified as high-grade.
- The False Negative Rate (FNR) is 0.2943714, meaning 29% of the true high grades were incorrectly classified as low-grade.
- The sensitivity, also known as the true positive rate, is 0.7056286, meaning 71% of the true high grades were correctly classified as high-grade.
- The specificity, also known as the true negative rate, is 0.7859145, meaning 79% of the true low grades were correctly classified as low-grade.
- The precision of this model and threshold is 0.746473, meaning 74.6% of the houses classified as high-grade were truly high-grade.

In our question, neither of the two errors is more consequential than the other, as we simply want to focus on reducing the error rate. Therefore, we will stick with a threshold of 0.5 to minimize the error rate.



Our ROC curve is clearly above the diagonal line meaning our model does better than random guessing. Therefore, our model does in fact use information from the data to make its classification. The blue dot represents the sensitivity (TPR) and FPR of our logistic regression with a threshold of 0.5 from our confusion matrix above. Here, it is evident that TPR is not equal to FPR so it is not classified at random when dealing with grade observations of houses in King County, Washington. The area under the curve (AUC) of our model is 0.8191738. Since the AUC is larger than 0.5 and closer to 1, it further confirms that our model performs better than random guessing in classifying observations of house grade between low and high.

In conclusion, our model can effectively predict the likelihood of a high-grade house in King County, Washington with an accuracy of almost 75%. This is beneficial because grade is the most impactful variable in a house's price as seen in our linear regression model. As a result, we can use these two models in tandem to help aspiring homeowners make informed decisions when purchasing their house.