Project 2

Julia, Raymond, Neil, Jasmine (Group 13)

2022-08-10

Section 1

In this report two models were designed from a dataset on over 21,000 houses collected in King County, Washington in 2014 and 2015. This dataset was used to make two different models. The first model predicts the square footage of houses in the dataset. The second model predicts whether a house is old or new. For the purpose of this study "old" houses were determined to be houses that were built or last renovated at least 50 years ago, while houses that were built or last renovated within the last 50 years are considered new. Instead of arbitrarily creating these categories, they were created this way because the National Register of Historic Places categorizes a house as old if it is around 50 years old or older(1).

The first model made in this analysis attempted to predict the square footage of a house given several different factors. While attempting to find the factors that would be the most important in predicting the square footage inside of a house, it was found that many factors were contenders. After conducting many optimization methods, it was found that the best predictors of the square footage inside of a house are the price, the number of bedrooms, the quality grade of the house, the year built, the square footage of the neighboring houses, the square footage of the lot, and the houses distance to downtown. The results of this model showed that it was able to predict square footage inside of a house pretty accurately.

This model can be used for many purposes, but the main insight gained from the results are how different factors related to the square footage of a house. Specifically, houses with more square footage are more expensive, have more bedrooms, have higher quality grades, are farther from downtown, have larger lots, and their neighbors also have bigger houses. Additionally, houses with less indoor square footage tend to be older. This is a realistic way to view the model, because if buyers want specific features in a house they can use this model to decide what square footage they should look for in a house. For example, if a buyer wants an older house very close to downtown they should look at a smaller house.

Next, the data was examined in an attempt to find which characteristics of homes would be most useful in predicting whether a home is new or old. In an initial investigation, using visualizations, many factors were found that were expected to be good predictors of the age of a home. With the help of these visualizations it was found that the number of bathrooms, the quality grade of the house, and the square footage of the house that was above ground were related to whether a house is new or old. As a result, the model using these factors to predict the age of a house was pretty accurate, but it could still be improved.

In order to improve this model, more factors were added that appeared to be less related to age of home than the original three factors, but still appeared to be related to whether a home was new or old. These factors are the number of bedrooms, whether the house is waterfront, the number of floors in the house, the price of the house, the conditions of the house, and whether or not the house has a basement. The original three factors of number of bathrooms, square footage above ground, and quality grade of the house were still included. This model was found to be even better than the previously created model with only the three intial factors.

As a result, this model was kept as our final model. This model was able to accurately predict whether a house is new or old 82 percent of the time. As a result it was concluded that there is a difference between new and old houses as they can be differentiated between most of the time. Moreover, this model shows that

if a home buyer wants a house with more bathrooms, better quality grades, more floors, or a home that is waterfront, they should focus on looking at newer homes. While, if a buyer wants a home with a basement, more bedrooms, or better conditions, they should focus on older homes.

Section 2

Data & Variable Description

The data used for this analysis includes information about houses and their sales in King County, from May 2014 to May 2015. There is information on over 21,000 houses and includes many important variables that are included when making decisions about house sales. This includes categorical variables such as whether a house has a view or not or has water on their property, and also quantitative variables such as square footage or price of the house. The table below describes what each variable in the dataset means and the type of variable that it is.

Variable	Description	Data Type
id	Unique ID for each home sold	float
date	Date of the home sale	string
price	Price of each home sold	float
bedrooms	Number of bedrooms	integer
bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower	float
sqft_living	Square footage of the apartments interior living space	integer
sqft_lot	Square footage of the land space	integer
floors	Number of floors	float
waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not	integer
view	An index from 0 to 4 of how good the view of the property was	integer
condition	An index from 1 to 5 on the condition of the apartment,	integer
grade	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 have a high quality level of construction and design.	integer
sqft_above	The square footage of the interior housing space that is above ground level	integer
sqft_basement	The square footage of the interior housing space that is below ground level	integer
yr_built	The year the house was initially built	integer
yr_renovated	The year of the house's last renovation	integer
zipcode	What zipcode area the house is in	integer
lat	Lattitude	float
long	Longitude	float
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors	integer
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors	integer

The id variable gives each house sold a unique identifier number. The date variable has the date in string

format for the day that the house was sold. Price is an integer value of the price that the house was sold for, in US dollars. The prices in this dataset ranged from $75,000 to $7.7 million dollars. The bedroom variable represents the number of bedrooms in the house. This ranges from 0 to 33 bedrooms. The bathroom variable that ranges from 0 to 8, and includes 0.5 values for bathrooms that do not include a shower/bathtub. The sqft_living variable is a quantitative variable that includes the square footage of the house and sqft_lot is the square footage of the total area of the property that includes the house. For this analysis, the model that will be built will be used to predict square footage of the house. Waterfront and view are binary variables that are one for if there is a body of water or view on the property and 0 if there is not. Condition is a range from 1-5 that ranks the condition of the apartment. Grade is an index from 1-11 that rates the quality of construction and design of a house. Sqft_above and sqft_basement is above and below ground level square footage. Year built and year renovated are also variables included in the dataset, and the logistic model for this analysis will be used to predict whether a house has been renovated/built in the past five years (2010 and recent). Zipcode, lat and long have to do with location of the house. Sqft_living15 and sqft_lot15 are square footage of the 15 surrounding houses, and are interesting variables in a dataset as houses that are close to each other are usually similar in characteristics.

Before model building, exploratory analysis must be conducted to visualize variable interactions with each other, and with the response variable. We created univariate, bivariate and multivariate visualizations and analyzed those to give us a better idea of what variables could be included in the model.

We also created new variables from existing variables for further analysis with factors that we believed could also be added into a model for square footage of a home. First, we created the variable Year that combines the year the house was built and the year the house was last renovated so that it contains the maximum of the two. Additionally, we created our logistic regression response variable called AgeOfHome that categorizes a house as new or old. A house is categorized if it was built or last renovated over 50 years ago. A house is considered new if it was built or renovated in the last 50 years.

The model that is being created for linear regression has square footage of a house as the response variable. The following graphs and plots have square footage as the response variable and variables that we thought based off of common knowledge could be included in a model used to predict the square footage of the home.

Questions of Interest

1. Can we predict the square footage of the inside of a house based on its other features? Our motivation for this question is that we can see which factors influence square footage the most. For example does the condition of a house go down as its square footage goes up? These are important questions for house buyers so that they can understand what other factors diminish or increase with square footage.

2. Can we predict whether a house is old or new? Our motivation for this question is to determine whether there is a significant different between old and new houses. If we are able to predict with good accuracy whether a house is old or new, there must be significant differences between these homes. Home buyers should know whether they want a old house or a new house, if the features of these houses are significantly different.

Create New Variables

First, we will create the variable Year that combines the year the house was built and the year the house was last renovated so that it contains the maximum of the two. Additionally, we will create our logistic regression response variable called AgeOfHome that categorizes a house as new or old. A house is categorized if it was built or last renovated over 50 years ago. A house is considered new if it was built or renovated in the last 50 years.

In order to decide if we can use this new variable as the response in the logistic regression, we have to check the sample sizes of the two categories.

```
##
##   Old   New
##  7872 13741
```

```
##
##       Old       New
## 0.3642252 0.6357748
```

The data has 7872 observations that are categorized as old, and it has 13741 observations that are categorized as new. Additionally, 54 percent of the data is categorized as new while 36 percent is categorized as old.

Additionally, we created a variable called basement that contains a 0 if the house does not have a basement and 1 if the house does have a basement.

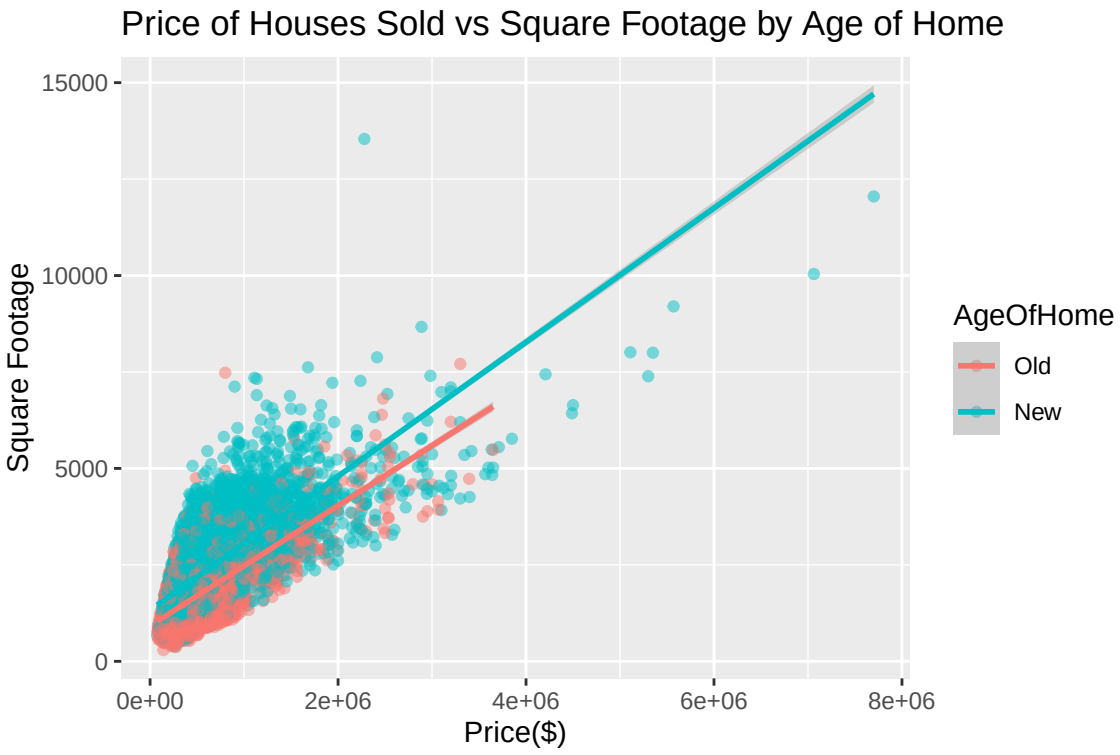
In addition, new feature is created called `distdt` from `long` and `lat` which measures the distance between the house in the observation and downtown Seattle in degrees and assuming a small angle approximation. This feature might give us more insight than the two dimensional world coordinates provided. With `distdt` created we drop `long` and `lat` from our data.

Split into a training and testing set

Next, we split the data into a training set and testing set. The training set contains 80% of the data while the testing set contians 20% of the data.

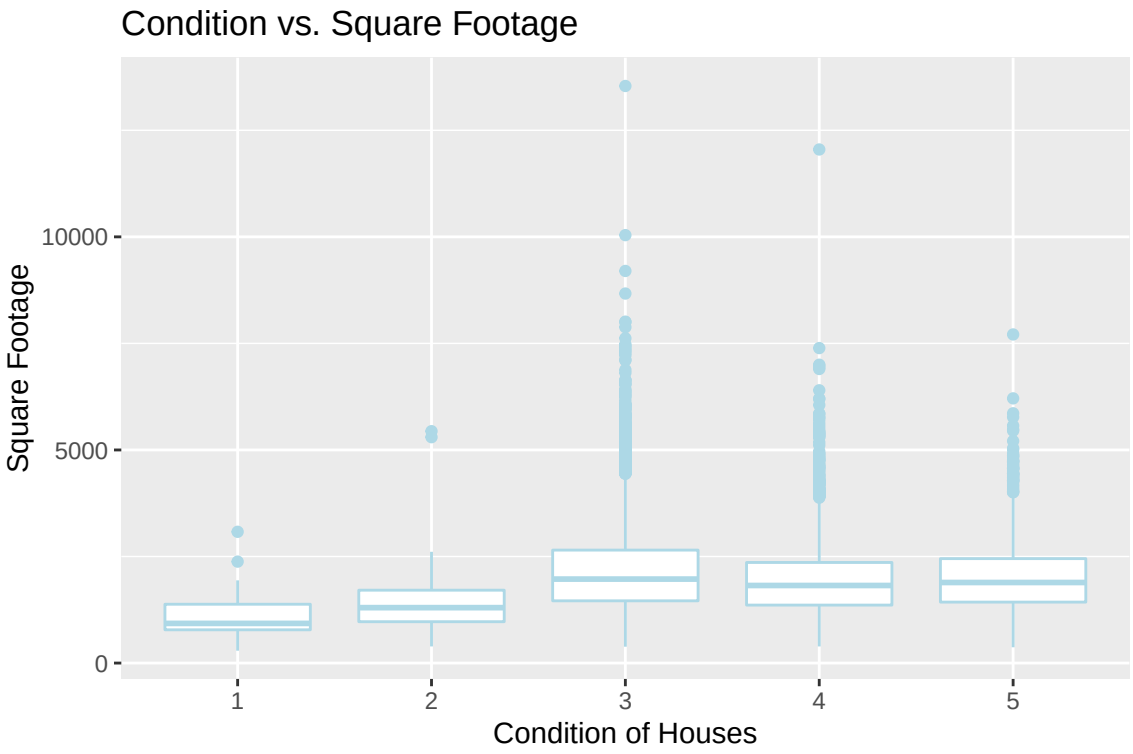
Plots for Square Footage

```
## `geom_smooth()` using formula 'y ~ x'
```

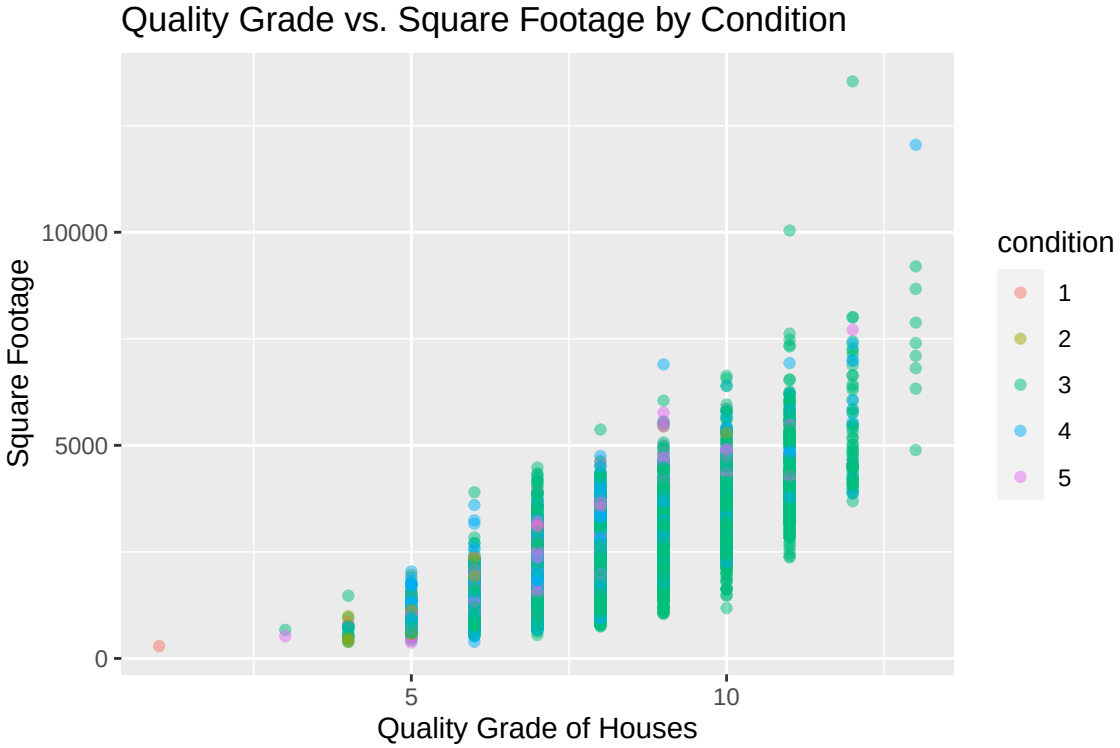


From the above plot, we are able to see that there is a linear relationship between price and square footage. In addition, we can see that old and new homes have different regression lines, so they can be treated differently.

The above graph shows that as the number of bedr0oms increases, the square footage of the house increases. Condition so not show any strong trend in relation to this data.



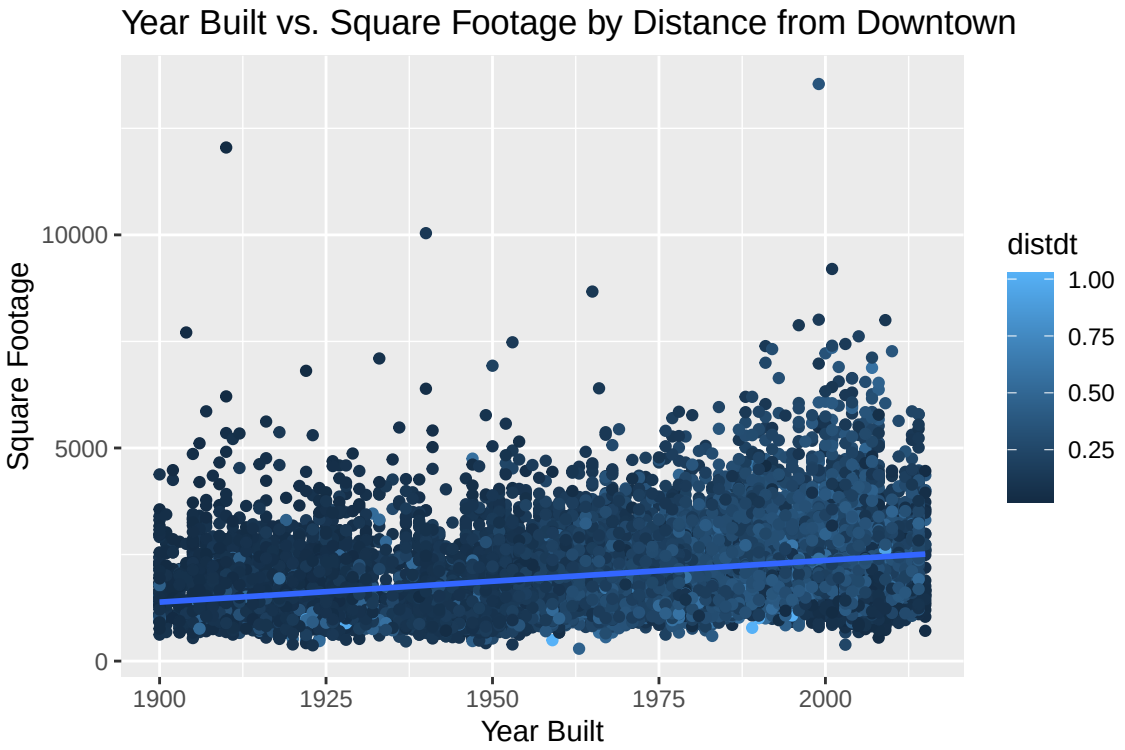
The above graph shows the relationship between condition of houses and square footage. Houses with a condition of 3 tend to have the most square footage which houses with a condition of 1 and 2 tend to have the smallest square footage.



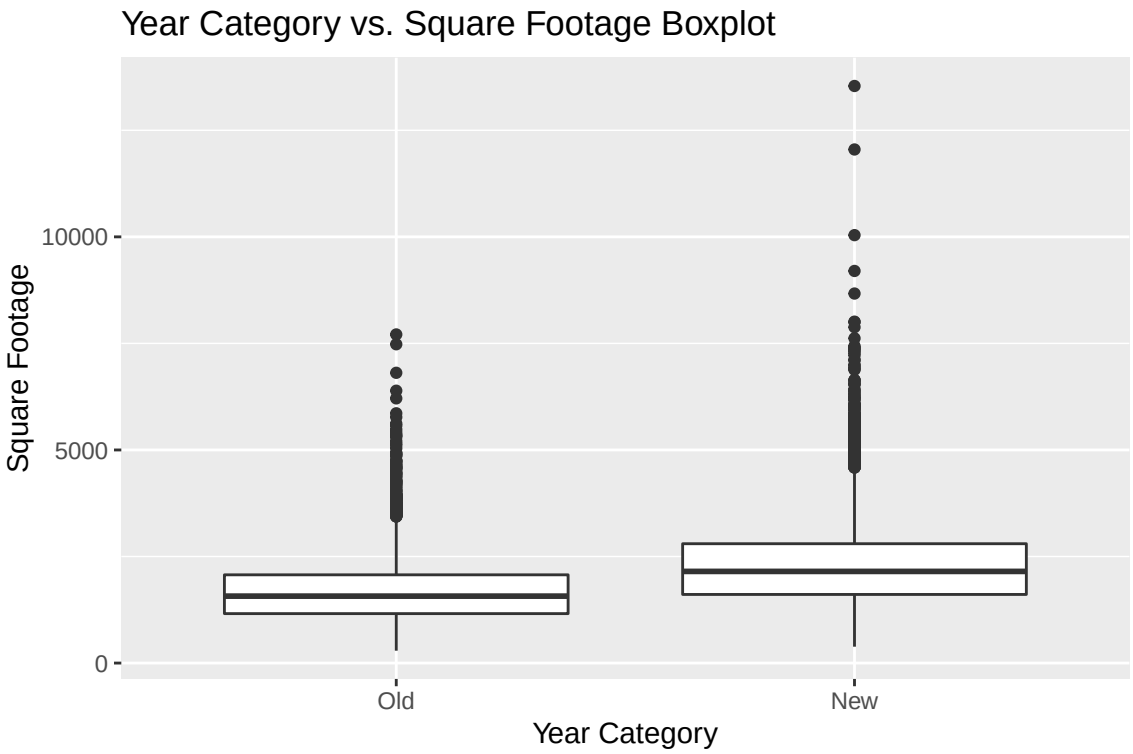
In this plot, we can see that as the quality grade of houses increases so does the square footage. Additionally,

the condition appears to have no obvious trend with these two variables.

```
## `geom_smooth()` using formula 'y ~ x'
```

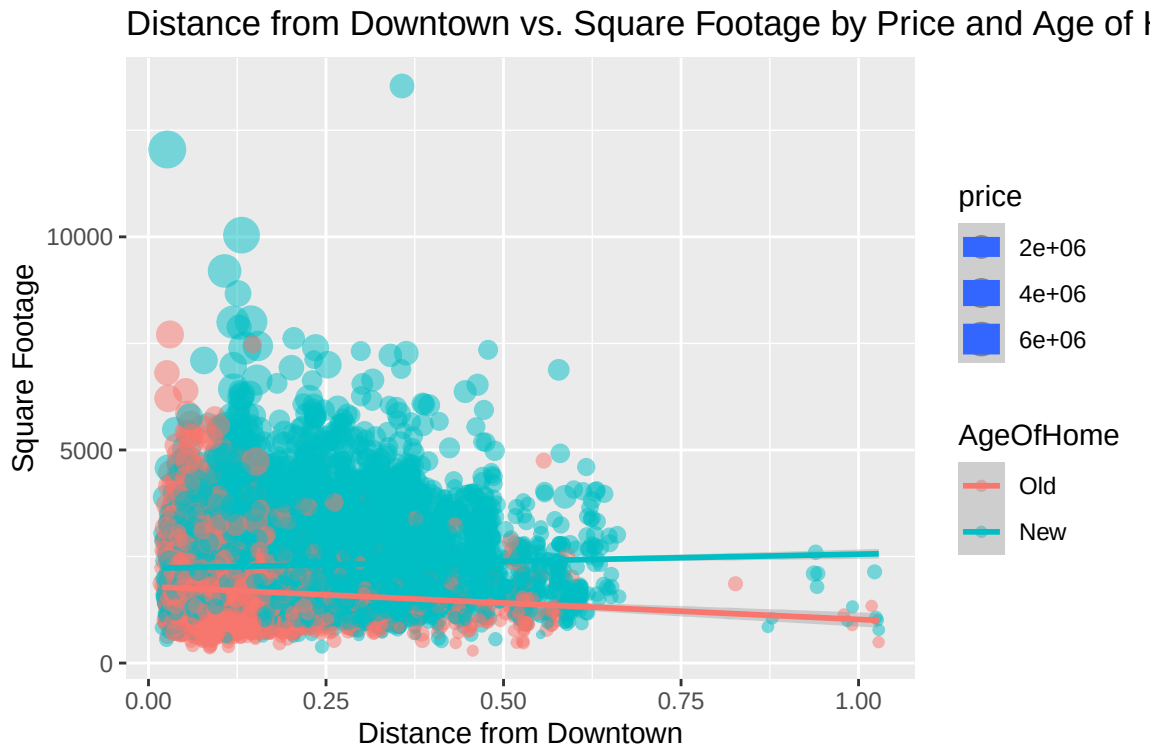


The year built and the square footage do not appear to have a strong linear relationship. Although in this graph we are able to see that in more recent years houses built near downtown are much smaller than houses that are not built near downtown Seattle.

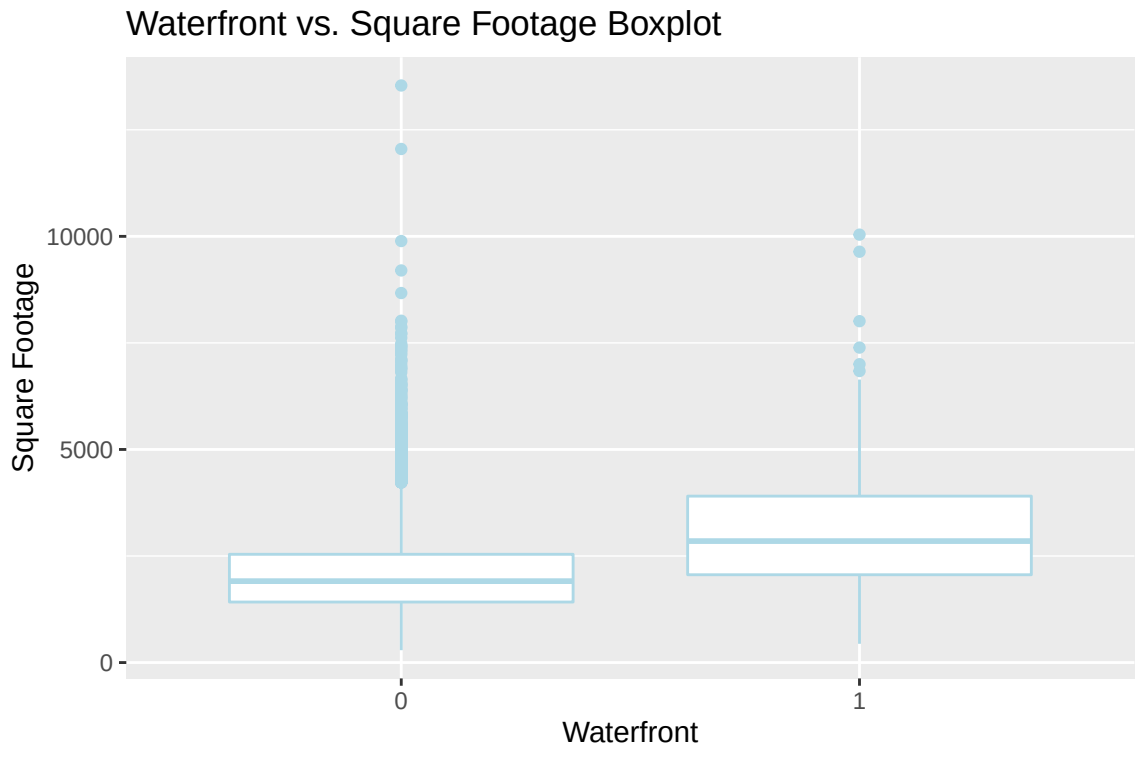


From this plot, we see that houses built within the last 50 years are a little bit bigger than houses built 50 or more years ago. This implies that square footage could be a good predictor of whether a house is new or old.

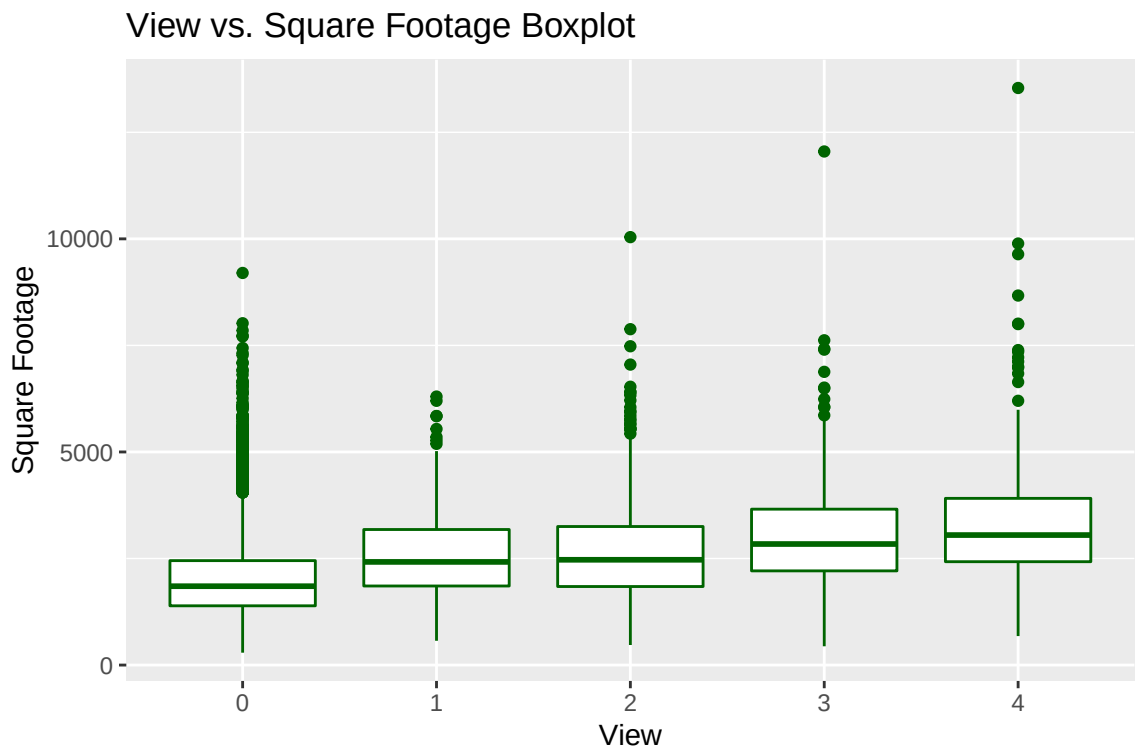
```
## `geom_smooth()` using formula 'y ~ x'
```



As the distance from downtown increases, the square footage of old homes tends to decrease while the square footage of new homes tends to increase. Although, these are not necessarily strong linear relationships.

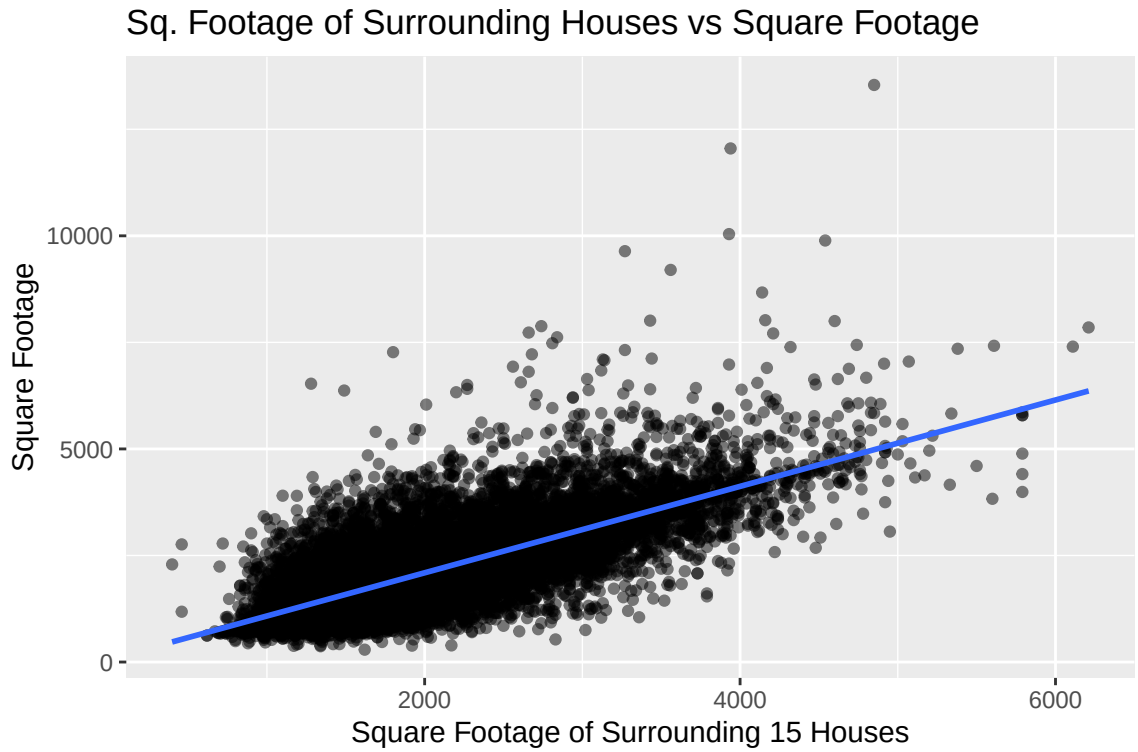


The box plot shows that waterfront homes tend to have a higher square footage. Waterfront could be a good predictor of square footage.



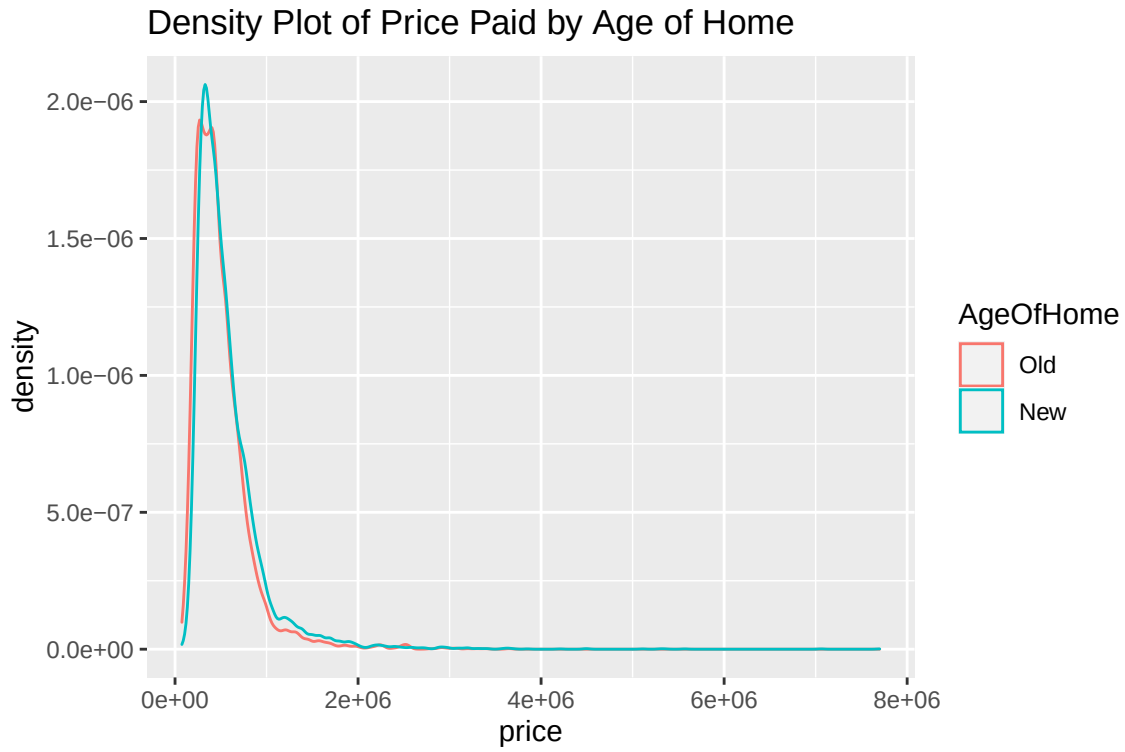
The boxplot above implies that as the view a home has gets better, the square footage of the home tends to increase. View could be a good predictor for our linear regression model.

```
## `geom_smooth()` using formula 'y ~ x'
```



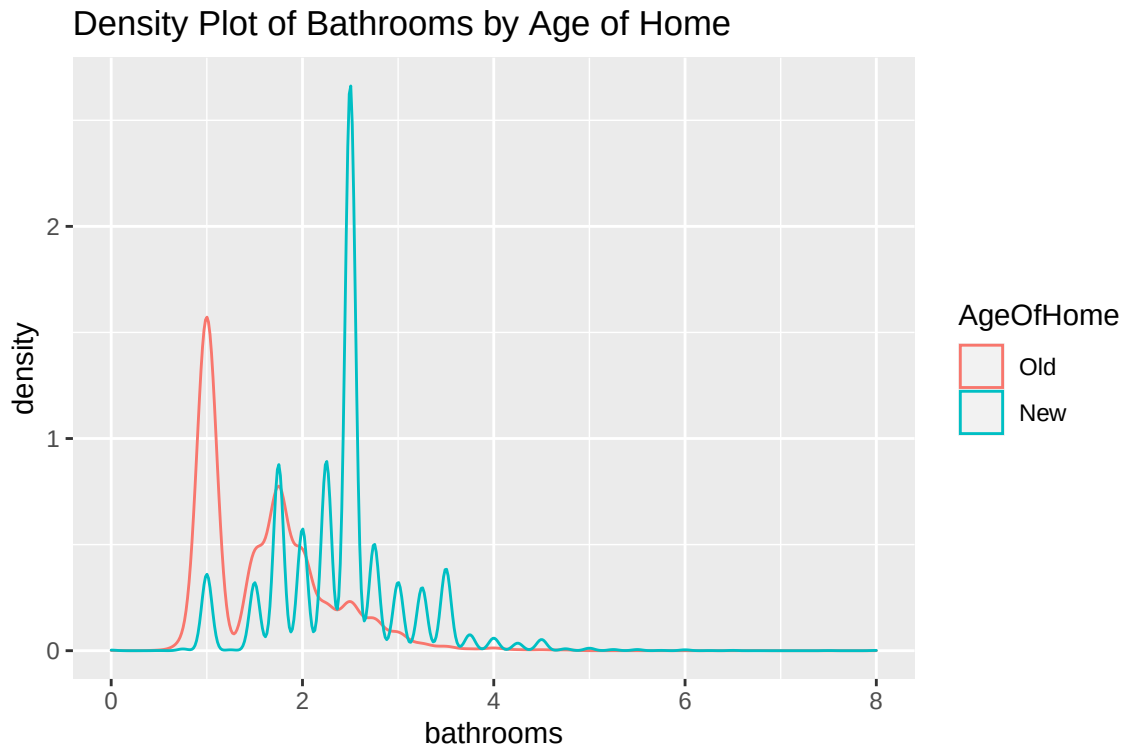
The above graph shows that as the square footage of surrounding houses increases, the square footage of the house should increase. This makes sense because neighborhoods tend to have similar sizes of homes.

#House Age Plots (Logistic Regression)

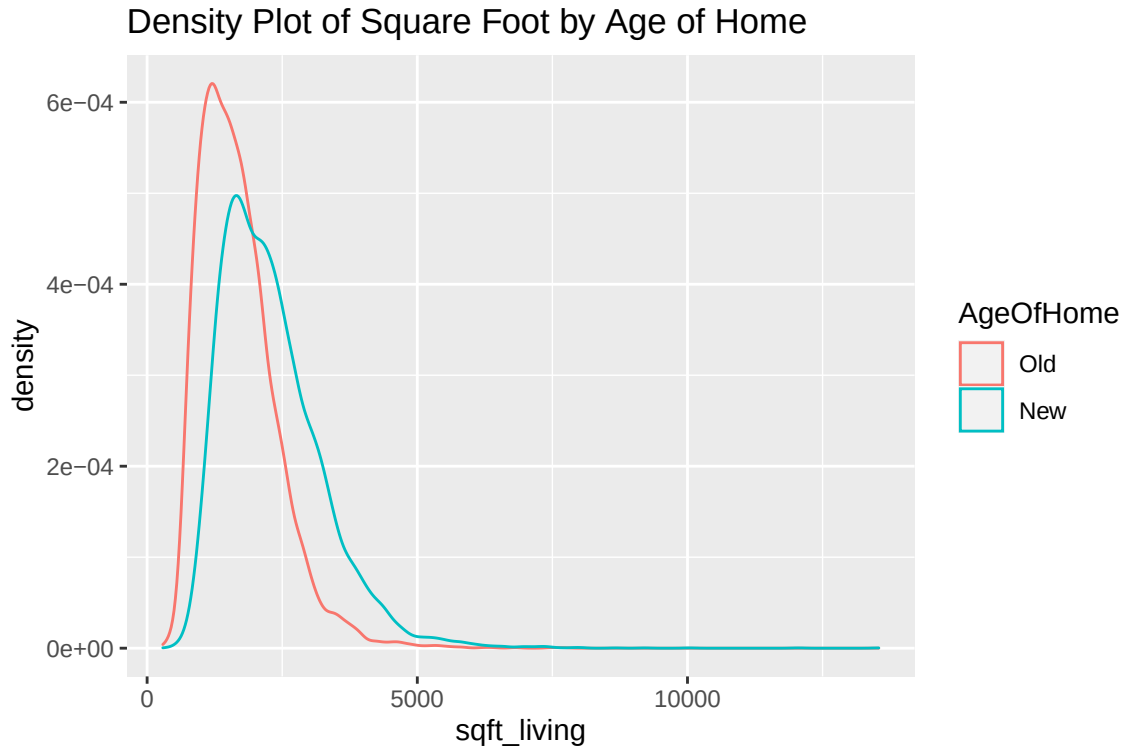


The density of price paid for homes does not appear different based on whether a home is old or new. This is not expected to be a significant decider on whether a home is old or new.

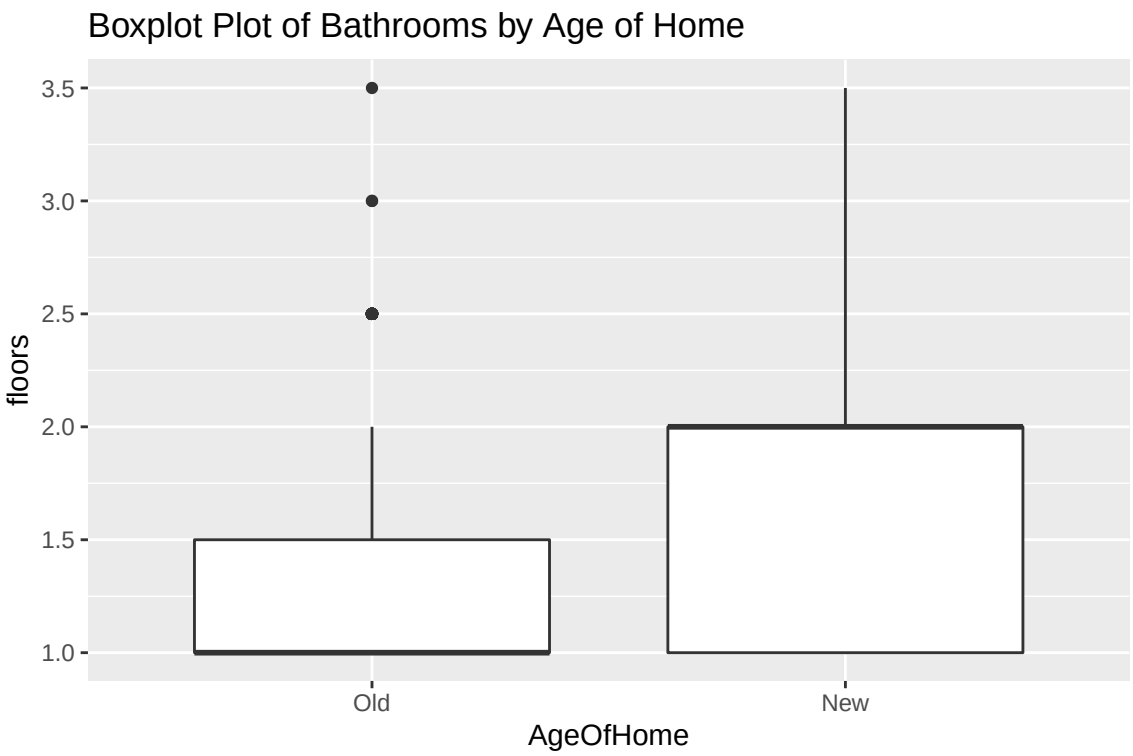
```
##density plots
ggplot(train,aes(x=bathrooms, color=AgeOfHome))+
  geom_density()+
  labs(title="Density Plot of Bathrooms by Age of Home")
```



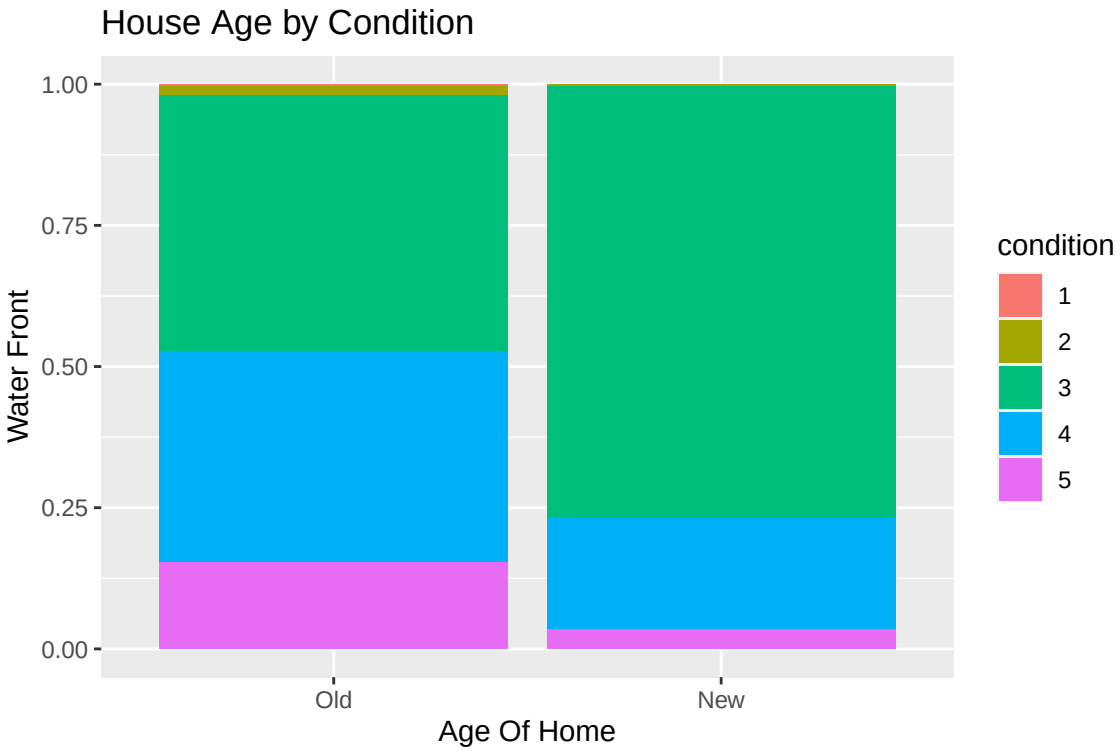
The density of bathrooms is very different based on whether a home is old or new. This variable should be significant in differentiating between old and new homes.



The density of square foot living is slightly different for old and new homes. This variable is questionable in whether it will be a significant predictor of whether a home is old or new.

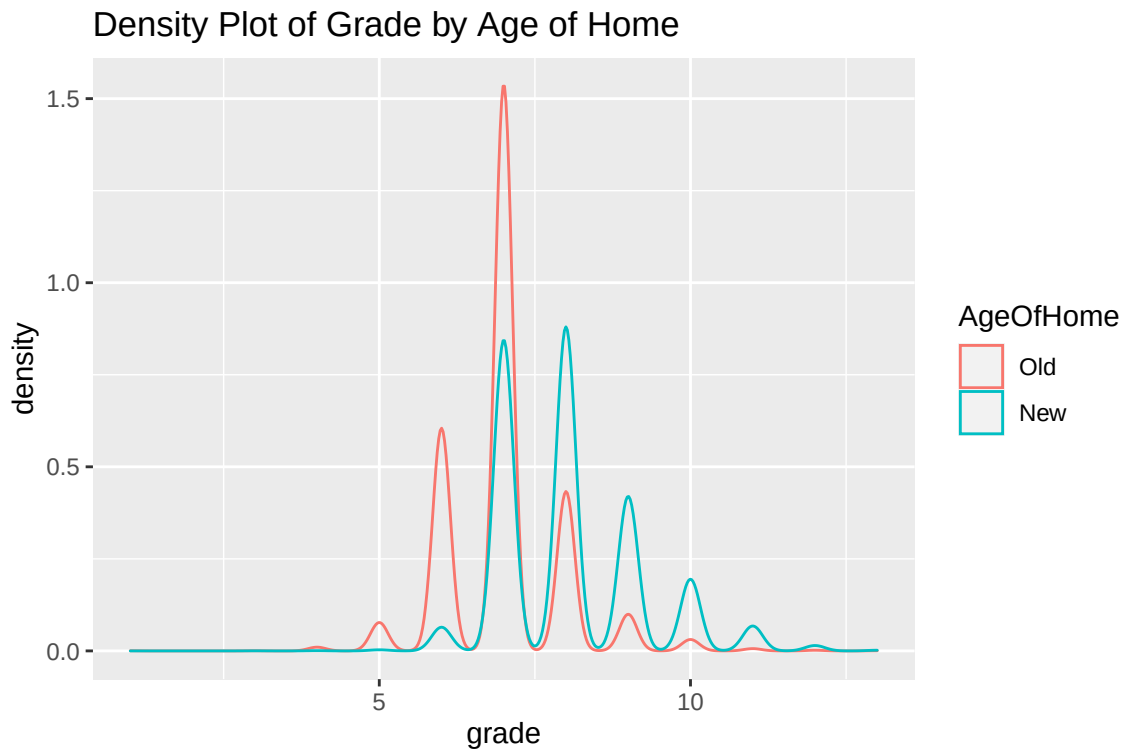


The above graph shows that new homes are more likely to have houses with more floors than old homes. Old homes are more likely to be first floor homes. Floors should be a significant predictor in predicting age of home.



From the above plot, we can see that new houses are much more likely too have a condition of 3, while old

houses are more likely to have a condition of 5. This shows that condition will likely be a factor in predicting the age of a home.



The density plot above shows that old and new homes have signficantly different densities based on grade, so this could be a good predictor of whether a home is old or new.

Section 3

3.0 Predicting `sqft_living`

We are interested in exploring what predictors, if any, may exist for the square footage of the interior living space of the homes in our data. Before doing anything to our data, we randomly split it into two sets. The *training* set contains 80% percent of the original data and the *test* set contains 20% of the original data. For reproducibility purposes of this paper, we included `set.seed(1)` in our `r` script before doing our random split. Now that we have a training set and a validation set, we are ready to continue.

3.1 Data Manipulation and Preparation

We start with our training data and decide to drop certain columns that we know will impede our model. `id` is dropped because it only serves as an identifier the same way the index does, it has no other relation to the data. A new feature is created called `distdt` from `long` and `lat` which measures the distance between the house in the observation and downtown Seattle in degrees and assuming a small angle approximation. This feature might give us more insight than the two dimensional world coordinates provided. With `distdt` created we drop `long` and `lat` from our data. Since our response variable `sqft_living` is a linear combination (the sum) of `sqft_above` and `sqft_basement` those two features are also dropped. `zipcode` and `date` are also dropped from our data. In the end we dropped seven features from our data (`date`, `id`, `lat`, `long`, `zipcode`, `sqft_above`, and `sqft_basement`) and created one (`distdt`).

3.2 Fitting a MLR

To look for a regression equation that represents the best model for `sqft_living` we regress all possible subsets of our predictors in our trimmed data using `regsubsets` from the `leaps` package in `r`. Once we regressed all possible subsets, we took a look at what models returned optimal values for important statistics in their predictive power and reliability in comparison to all other models in the finite set. In particular, we were interested in the models that provided the highest $R^2_{adjusted}$ (equation 3.1), the lowest Mallow's C_p (equation 3.2) and the lowest Bayesian Information Criterion, BIC (equation 3.3). As you can see, the models returned for each of these statistics are identical. The model that was returned satisfying the optimization of all three statistics of interest uses seven predictors including our new `distdt` predictor.

Model with Highest $R^2_{adjusted}$

$$\hat{y} = 2868.086 + 0.0006206373 \cdot price + 192.0078 \cdot bedrooms + 148.6072 \cdot grade - 2.227198 \cdot yr\_built$$

$$+0.3119265 \cdot sqft\_living15 + 0.001.741457 \cdot sqft\_lot15 + 620.3304 \cdot distdt \ (Equation 3.1)$$

Model with lowest Mallow's C_{p}

$$\hat{y} = 2868.086 + 0.0006206373 \cdot price + 192.0078 \cdot bedrooms + 148.6072 \cdot grade - 2.227198 \cdot yr\_built$$

$$+0.3119265 \cdot sqft\_living15 + 0.001.741457 \cdot sqft\_lot15 + 620.3304 \cdot distdt \ (Equation 3.2)$$

Model with lowest BIC

$$\hat{y} = 2868.086 + 0.0006206373 \cdot price + 192.0078 \cdot bedrooms + 148.6072 \cdot grade - 2.227198 \cdot yr\_built$$

$$+0.3119265 \cdot sqft\_living15 + 0.001.741457 \cdot sqft\_lot15 + 620.3304 \cdot distdt \ (Equation 3.3)$$

To further validate this model we use `r` to evaluate $R^2_{prediction}$ and R^2. We find that $R^2_{prediction} = 0.82013$ meaning that this model might be able to explain about 81.91% of the variability in the new observations. Additionally, we find that $R^2 = 0.8199$. The proximity of these two values allows us to conclude that that the model has good predictive ability.

3.3 Validating the Model

Before proceeding we must evaluate whether or not the model represented in **Equation 3.1** is useful in predicting `sqft_living`. We do this my performing the following hypothesis test

$H_0 : \beta_{price} = \beta_{bedrooms} = \beta_{grade} = \beta_{yr\_built} = \beta_{sqft\_living15} = \beta_{sqft\_lot15} = \beta_{distdt} = 0,$

H_a : at least one of the coefficients in H_0 is not zero.

Using the summary table of our model below, we find that the *ANOVA F* statistic is 9835 with p-value approaching 0. Given that the p-value is less than 0.05, we reject the null hypothesis. The results support the claim that the model represented in **Equation 3.1** is useful in predicting `sqft_living`. Additionally, we see that all of our features appear to be significant informing us that we can proceed without dropping any of them.

```
##
## Call:
## lm(formula = sqft_living ~ price + bedrooms + bathrooms + grade +
##     yr_built + sqft_living15 + sqft_lot15 + distdt, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5748.6  -235.8   -37.0   197.9  5768.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.868e+03  2.663e+02   10.77   <2e-16 ***
## price          6.206e-04  1.302e-05   47.66   <2e-16 ***
## bedrooms       1.920e+02  3.785e+00   50.73   <2e-16 ***
## bathrooms      3.374e+02  6.211e+00   54.33   <2e-16 ***
## grade          1.486e+02  4.620e+00   32.16   <2e-16 ***
## yr_built      -2.227e+00  1.417e-01  -15.71   <2e-16 ***
## sqft_living15  3.119e-01  6.790e-03   45.94   <2e-16 ***
## sqft_lot15     1.741e-03  1.137e-04   15.31   <2e-16 ***
## distdt         6.203e+02  3.013e+01   20.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 388.5 on 17281 degrees of freedom
## Multiple R-squared:  0.8199, Adjusted R-squared:  0.8198
## F-statistic:  9835 on 8 and 17281 DF,  p-value: < 2.2e-16
```

Next, we took a look at influential observations. We found that there are 1507 high leverage points within our train set. This is likely due to a right skew in the square foot data, because there are some houses that have a very high square footage. Additionally, there are 1224 points that are influential based on DFFITS, this is also likely due to the same problem.

Additionally, we took a look the cooks distance of our model. We found that one observation appeared to be influential. It was observation 15871. We took a closer look at this observation and found that it had 1.75 bathrooms, was 1600 square feet, but had 33 bedrooms. As a result, we are assuming that this data point was entered incorrectly, so will take this point out of the data.

```
##   15871
## 2.47809
```

```
##
## Call:
## lm(formula = sqft_living ~ price + bedrooms + bathrooms + grade +
##     yr_built + sqft_living15 + sqft_lot15 + distdt, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1906.4  -235.7   -36.7   196.5  5772.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.634e+03  2.649e+02   9.942   <2e-16 ***
## price          6.241e-04  1.293e-05  48.262   <2e-16 ***
## bedrooms       2.099e+02  3.930e+00  53.406   <2e-16 ***
```

```
## bathrooms      3.258e+02  6.213e+00  52.435   <2e-16 ***
## grade          1.487e+02  4.588e+00  32.416   <2e-16 ***
## yr_built      -2.122e+00  1.409e-01 -15.056   <2e-16 ***
## sqft_living15  3.068e-01  6.751e-03  45.449   <2e-16 ***
## sqft_lot15     1.771e-03  1.129e-04  15.680   <2e-16 ***
## distdt         6.120e+02  2.993e+01  20.448   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 385.8 on 17280 degrees of freedom
## Multiple R-squared:  0.8224, Adjusted R-squared:  0.8223
## F-statistic: 1e+04 on 8 and 17280 DF,  p-value: < 2.2e-16
```

After removing the outlier, we found the best model again on the training set by finding the model with the best bic, cp malle, and r squared adjusted, and all of the same predictors in the model remained, the only change was the value of the intercept.

$$\hat{y} = 2633.810 + 0.0006206373 \cdot price + 192.0078 \cdot bedrooms + 148.6072 \cdot grade - 2.227198 \cdot yr\_built$$

$$+0.3119265 \cdot sqft\_living15 + 0.001.741457 \cdot sqft\_lot15 + 620.3304 \cdot distdt \ (Equation 3.4)$$

3.4 Testing the Model for Assumptions for Linear Regression

Now that we have a potential model for predicting `sqft_living` we test if it meets the assumptions for linear regression. We test for the first two assumptions (i.e. if the residuals fall in a horizontal band around 0 and if they have similar variation across fits) by creating the residual plot found in **Figure 3.1**. Using **Figure 3.1** as reference, we are satisfied that the first two assumptions for linear regression are met and continue to check for the other assumptions.

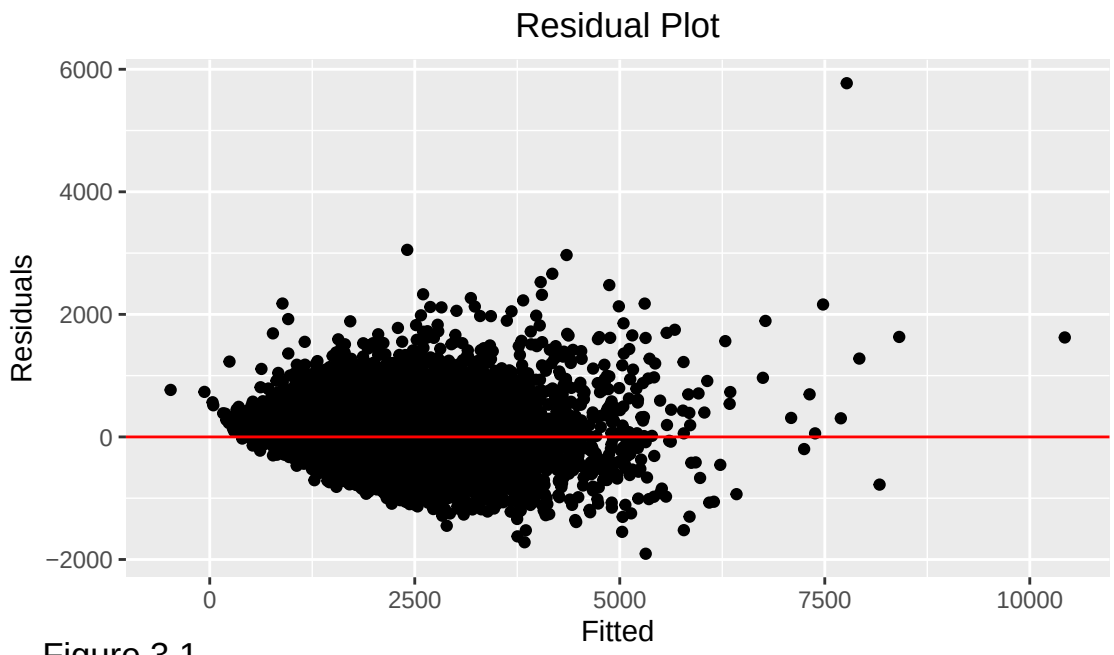


Figure 3.1

The third assumption for linear regression states that residuals are uncorrelated. To test this assumption we create an autocorrelation plot of the residuals in **Figure 3.2** and assess that the ACF never passes the dashed lines save for when *Lab = 0*. The results displayed in **Figure 3.2** indicate that assumption 3 for linear regression is met and we may continue.

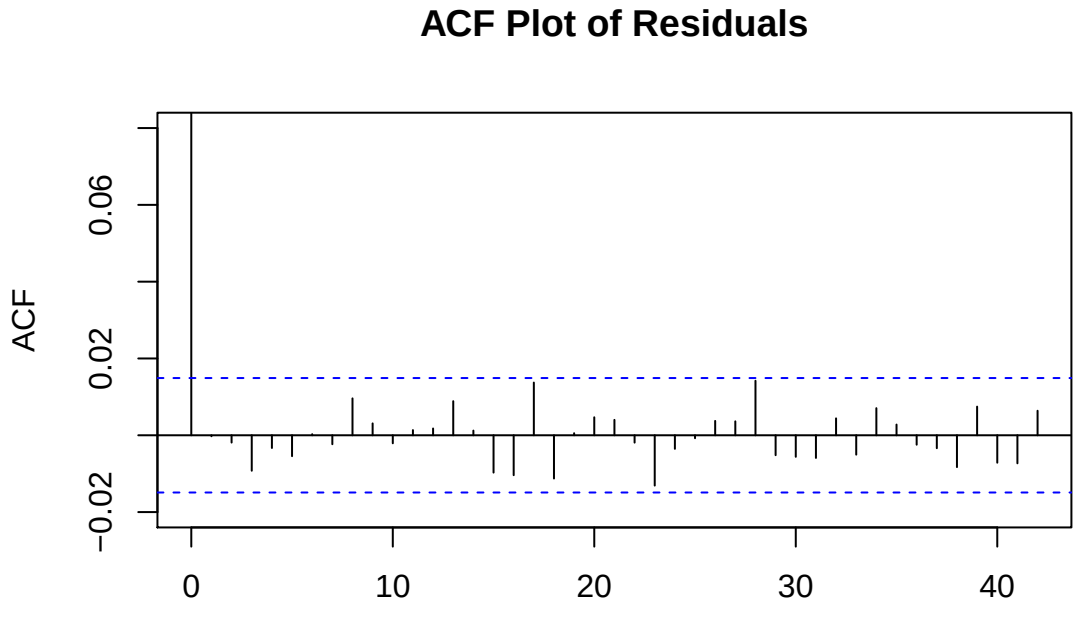


Figure 3.2

The final assumption for linear regression that we are testing for is that the error terms in our model follow a normal distribution. We test for this assumption by creating a QQ plot of the residuals in **Figure 3.3**. We expect the data in our plot to fall close to the line representing the expected value under normality. This is exactly what our data is doing and we can confidently say that assumption 4 for linear regression has been met.

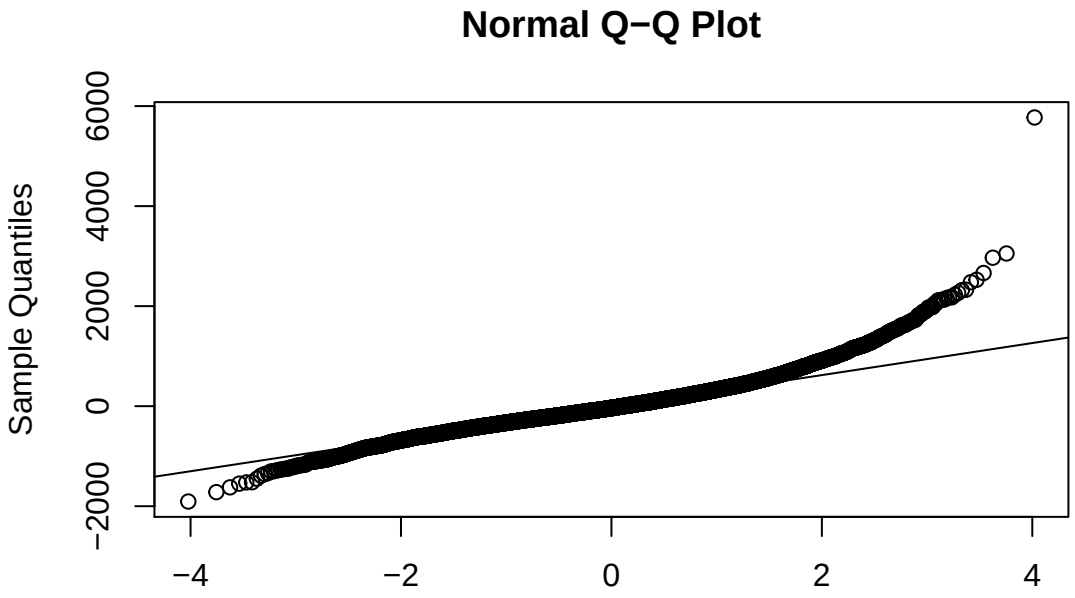


Figure 3.3

3.5 Additional Tests

In order to further validate our model, we check for a few more statistics to provide further insight to how robust our model is. We look at the MSE of our test set using the current model and at the variance inflation factor (VIF) for each predictor of our model. The MSE will evaluate the error for our validation (i.e. `test`) set. We calculate $MSE = 153130.3$ which is acceptable in this context. VIF evaluates how many times larger the variance for a predictor is that it would have been without collinearity. The VIF for each of our predictors can be found in the table below. All values in the table below are under 4, meaning that we can feel confident in the fact that the statistical significance of our independent features is high enough to keep them and not worry about futher manipulating the data.

```
##         price      bedrooms     bathrooms         grade      yr_built
##      2.571247      1.466329      2.650278      3.367826      1.983555
## sqft_living15    sqft_lot15        distdt
##      2.485425      1.138750      1.583064
```

Predicting `sqft_living` Conclusions

In order to both reduce the amount of time it would take to run all possible regression subsets, avoid including linearly dependent variables, and deal with irrelevant features we reduced the size of our data by removing the appropriate predictors in `Section 3.1`. In `Section 3.2` we fit all possible regressions on our

data and found that the model represented by `Equation 3.1` satisfies optimization of the three statistics we were interested in, $R^2_{adjusted}$, $Mallow's\, C_p$, and BIC. In sections `3.3 - 3.5` we ensured that our model did not have collinear features, that assumptions for linear regression were met, and that for our purposes it was robust for predicting `sqft_living`. What we end up with is `Equation 3.4` which serves as a reliable model to predict `sqft_living`. In `Figure 3.4` we finally plot our predicted values for the test set against the actual values for the test set. A perfect model free of even normal errors would follow a straight line along the plotted red line (y=x). Our model seems to follow this line closely with a few exceptions.

$$\hat{y} = 2634 + 0.0006206373 \cdot price + 192.0078 \cdot bedrooms + 148.6072 \cdot grade - 2.227198 \cdot yr\_built$$

$$+0.3119265 \cdot sqft\_living15 + 0.001.741457 \cdot sqft\_lot15 + 620.3304 \cdot distdt \; (Equation 3.5)$$

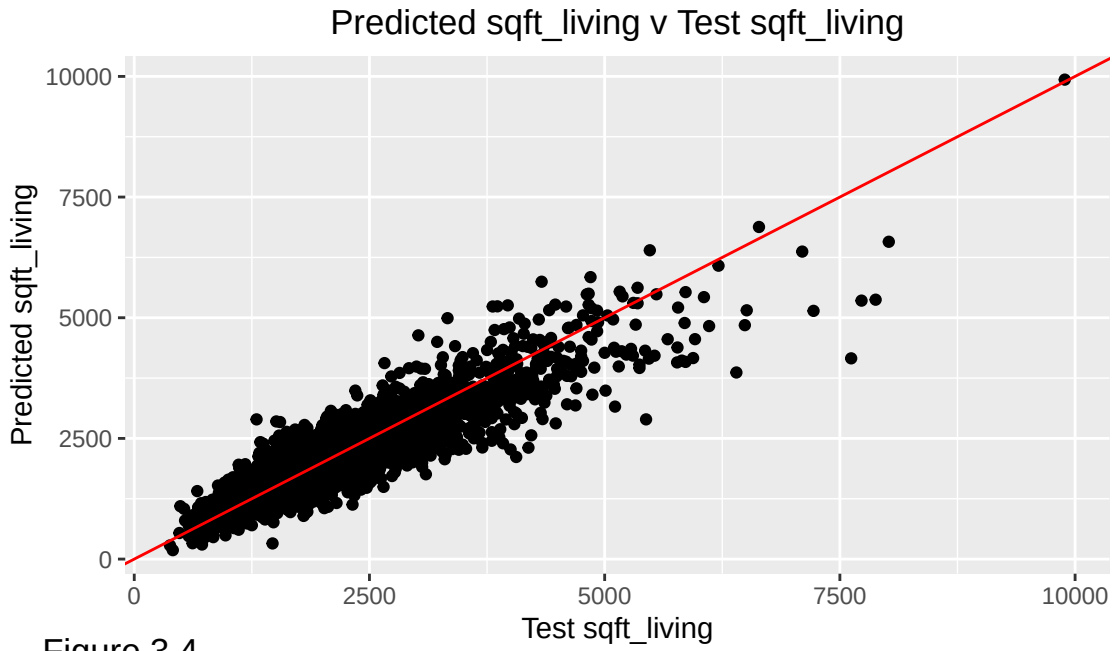


Figure 3.4

Section 4

Initial Model

Based on our EDA, we could immediately tell that bathrooms, grade, and the square footage of the interior housing above ground level would be the most significant in predicting whether a home is new or old.

```
##
## Call:
## glm(formula = AgeOfHome ~ bathrooms + grade + sqft_above, family = "binomial",
##     data = train)
##
## Deviance Residuals:
```

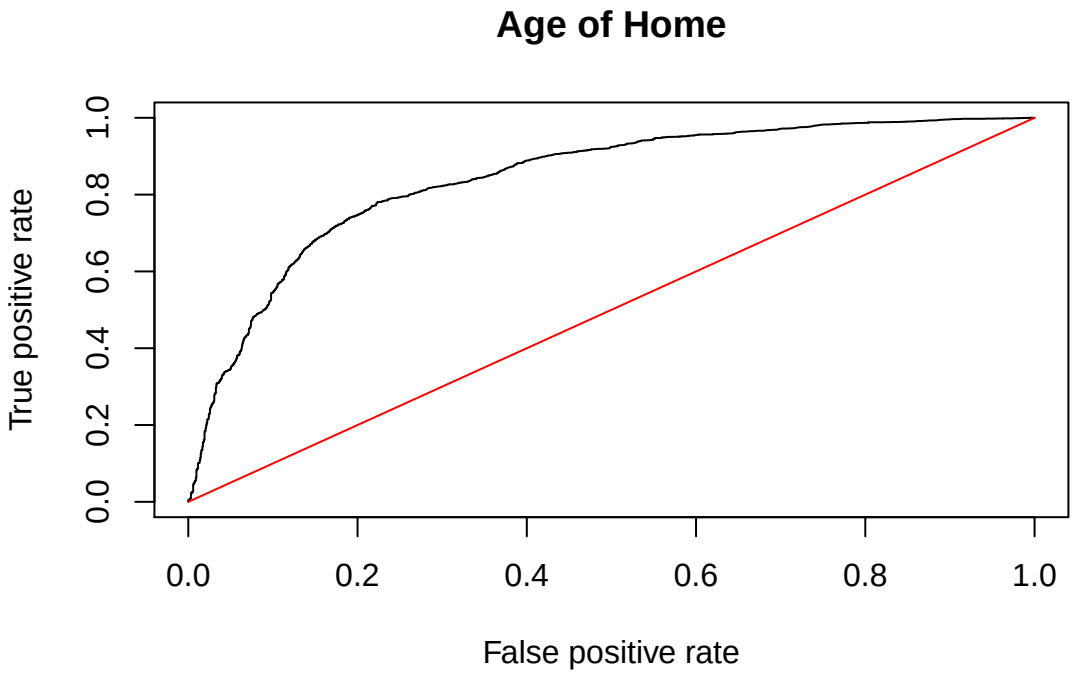
```
##     Min       1Q   Median       3Q      Max
## -4.3866  -0.7574   0.4138   0.7611   3.0036
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.147e+00  1.807e-01 -34.029  < 2e-16 ***
## bathrooms    1.319e+00  3.963e-02  33.287  < 2e-16 ***
## grade        4.791e-01  2.953e-02  16.226  < 2e-16 ***
## sqft_above   3.139e-04  4.427e-05   7.091 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22690  on 17289  degrees of freedom
## Residual deviance: 16944  on 17286  degrees of freedom
## AIC: 16952
##
## Number of Fisher Scoring iterations: 5
```

Hypothesis Tests

From the above model we were able to see that all three of our predictors passed the wald test.

Additionally, we tested whether our model was a better predictor than an intercept only model. The null and alternative hypotheses are: H0 : B1 = B2 = B3 = 0 Ha : at least one of the coefficients in H0 is not zero We found that the pvalue was zero, so we rejected the null hypothesis and found that our three predictor model is useful.

Testing the model



From the above ROC curve, we can see that our model performs better than random chance. It also increases pretty quickly which implies that our true positive rate will be generally high when our false positive rate is low.

Additionally, the AUC we found is 0.84, which also tells us that the model performs better than random chance.

Next, we find the confusion matrix for our model.

```
##
##       FALSE TRUE
##   Old  1002  562
##   New   409 2350
```

From the confusion matrix, we can see that the false positive rate is 35 percent while the true positve rate is 84 percent.

Try to improve Model

In order to try and improve the model, we will add more predictors to our initial model. Based on our EDA, the predictors bedrooms, whether the house is waterfront, square feet of the living space, number of floors, price, conditions, whether the house has a basement could all be significant in predicting whether a home is old or new.

```
##
## Call:
## glm(formula = AgeOfHome ~ bathrooms + grade + sqft_above + bedrooms +
```

```
##     sqft_living + floors + price + condition + Basement + waterfront,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0309  -0.6268   0.2020   0.5498   4.7836
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.323e+00  2.706e-01 -27.063  < 2e-16 ***
## bathrooms    1.879e+00  5.251e-02  35.787  < 2e-16 ***
## grade        1.147e+00  3.950e-02  29.050  < 2e-16 ***
## sqft_above   4.710e-04  1.106e-04   4.258 2.06e-05 ***
## bedrooms    -3.545e-01  3.109e-02 -11.404  < 2e-16 ***
## sqft_living -1.097e-04  9.555e-05  -1.148    0.251
## floors       5.249e-01  5.909e-02   8.884  < 2e-16 ***
## price       -3.737e-06  1.087e-07 -34.377  < 2e-16 ***
## condition   -6.708e-01  3.191e-02 -21.021  < 2e-16 ***
## Basement1   -4.664e-01  7.928e-02  -5.882 4.04e-09 ***
## waterfront   1.882e+00  2.834e-01   6.642 3.09e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22690  on 17289  degrees of freedom
## Residual deviance: 13465  on 17279  degrees of freedom
## AIC: 13487
##
## Number of Fisher Scoring iterations: 6
```

Hypothesis Tests

Based on the Wald test, we are able to remove the predictor, square foot of living space from the model. All of the other predictors pass the Wald Test.

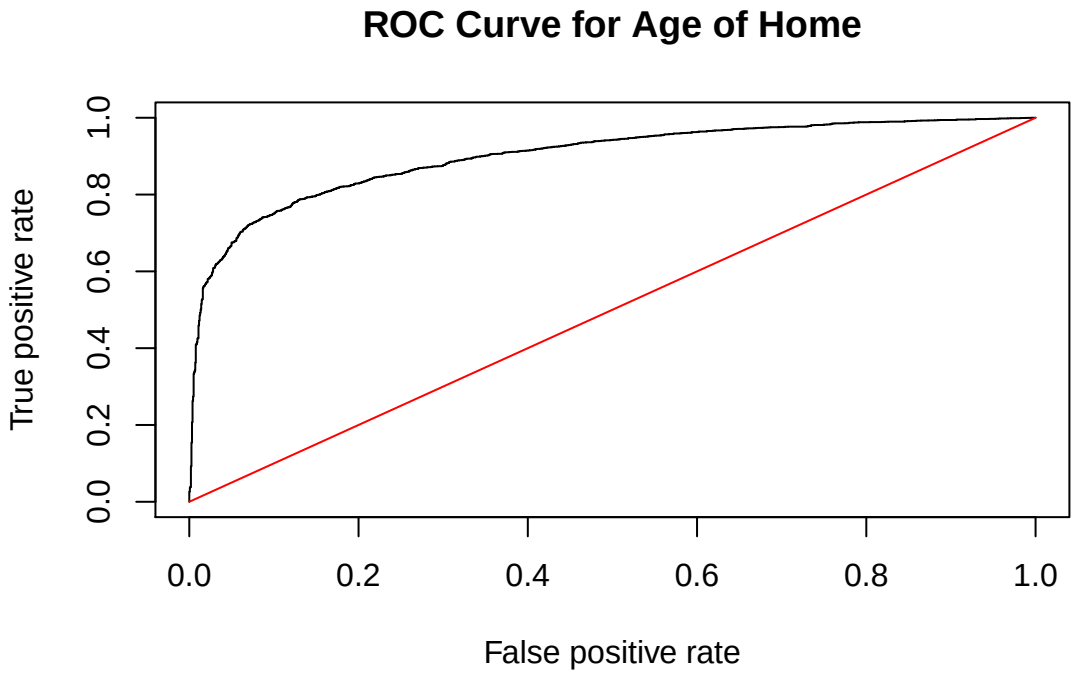
Our new model is below.

```
##
## Call:
## glm(formula = AgeOfHome ~ bathrooms + grade + sqft_above + bedrooms +
##     floors + price + condition + Basement + waterfront, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0591  -0.6268   0.2019   0.5500   4.7754
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.281e+00  2.680e-01 -27.172  < 2e-16 ***
## bathrooms    1.869e+00  5.178e-02  36.101  < 2e-16 ***
## grade        1.145e+00  3.943e-02  29.039  < 2e-16 ***
## sqft_above   3.707e-04  6.791e-05   5.458 4.80e-08 ***
```

```
## bedrooms    -3.626e-01  3.029e-02 -11.974  < 2e-16 ***
## floors       5.406e-01  5.752e-02   9.399  < 2e-16 ***
## price       -3.761e-06  1.068e-07 -35.229  < 2e-16 ***
## condition   -6.732e-01  3.184e-02 -21.143  < 2e-16 ***
## Basement1   -5.306e-01  5.621e-02  -9.440  < 2e-16 ***
## waterfront   1.872e+00  2.831e-01   6.612 3.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22690  on 17289  degrees of freedom
## Residual deviance: 13466  on 17280  degrees of freedom
## AIC: 13486
##
## Number of Fisher Scoring iterations: 6
```

Next, we test whether our 9 predictor model is more useful than our 3 predictor model. Our pvalue is zero so we find that the 9 predictor model is better than the 3 predictor model.

Testing the Model



From the above ROC curve, we can see that our model performs better than random chance. It also increases pretty quickly which implies that our true positive rate will be generally high when our false positive rate is low.

Additionally, the AUC we found is 0.90, which also tells us that the model performs better than random chance.

```
##
##       FALSE TRUE
##   Old  1167  397
##   New   394 2365
```

From the confusion matrix, we can see that the false positive rate is 25 percent while the true positve rate is 86 percent. Additionally, the accuracy of the model is 82 percent.

Final Model

```
##
## Call:
## glm(formula = AgeOfHome ~ bathrooms + grade + sqft_above + bedrooms +
##     floors + price + condition + Basement + waterfront, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0591  -0.6268   0.2019   0.5500   4.7754
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.281e+00  2.680e-01 -27.172  < 2e-16 ***
## bathrooms    1.869e+00  5.178e-02  36.101  < 2e-16 ***
## grade        1.145e+00  3.943e-02  29.039  < 2e-16 ***
## sqft_above   3.707e-04  6.791e-05   5.458 4.80e-08 ***
## bedrooms    -3.626e-01  3.029e-02 -11.974  < 2e-16 ***
## floors       5.406e-01  5.752e-02   9.399  < 2e-16 ***
## price       -3.761e-06  1.068e-07 -35.229  < 2e-16 ***
## condition   -6.732e-01  3.184e-02 -21.143  < 2e-16 ***
## Basement1   -5.306e-01  5.621e-02  -9.440  < 2e-16 ***
## waterfront   1.872e+00  2.831e-01   6.612 3.79e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22690  on 17289  degrees of freedom
## Residual deviance: 13466  on 17280  degrees of freedom
## AIC: 13486
##
## Number of Fisher Scoring iterations: 6
```

Our final model is the 9 predictor model. The predictors this model uses are the number of bathrooms, the house grade, the square foot of the home that is above ground, the number of bedrooms, the number of floors, the price of the home, the condition the home is in, whether the home has a basement, and whether the house is waterfront. Both our ROC curve and AUC show that the nine predictor model performs better than random guessing. Additionally the accuracy of this model is around 82 percent, while the false positive rate is around 25 percent, so this model is a pretty good at predicting whether a house is new or old. We have also concluded that this model is better than both a three predictor model and the intercept only model. The prediction equation for our model is:

$log\frac{\hat{\pi}}{1-\hat{\pi}} = -7.281 + 1.869 * (B1) + 1.145 * (B2) + 0.00037 * (B3) - 0.3626 * (B4) + 0.5406 * (B5) - 0.0000038 * (B6) - 0.6732 * (B7) + 1.872 * (B8) - 0.5306 * (I1)$

Cite Sources

1. “How to List A Property.” National Parks Service, U.S. Department of the Interior, https://www.nps.gov/subjects/nationalregister/how-to-list-a-property.htm.