Final Project-Team Four

Isabella Oakes, Jeffrey Burnett, and Chris Robinson

12/12/2020

```
library(ggplot2)
house <- read.csv('C:/Users/Ladybug/Desktop/house_sales.csv')
print(summary(house))
```

```
##        id                 date               price            bedrooms
##  Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000
##  1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
##  Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
##  Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.373
##  3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000
##                                                           NA's   :1134
##    bathrooms      sqft_living       sqft_lot           floors
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.500   1st Qu.: 1430   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1920   Median :   7620   Median :1.500
##  Mean   :2.114   Mean   : 2081   Mean   :  15180   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10708   3rd Qu.:2.000
##  Max.   :8.000   Max.   :12050   Max.   :1651359   Max.   :3.500
##  NA's   :1068    NA's   :1110    NA's   :1044
##    waterfront            view          condition         grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
##  Median :1560   Median :   0.0   Median :1975   Median :   0.0
##  Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
##
##     zipcode           lat             long        sqft_living15
##  Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  Median :98065   Median :47.57   Median :-122.2   Median :1840
##  Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
```

```
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
##
##    sqft_lot15
##  Min.   :   651
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12768
##  3rd Qu.: 10083
##  Max.   :871200
##
```

```
house_subset <- subset(house, select =-c(1, 2, 6, 7, 16, 18, 19), bedrooms < 30)
house_subset$bedrooms[is.na(house_subset$bedrooms)] <- 3
house_subset$bathrooms[is.na(house_subset$bathrooms)] <- 2.25
```

Before looking at correlation, we removed the housing ID, date, sqft_living, sqft_lot, lat, and long. The housing ID would not be useful for correlation, so it was removed. The date field was within a year, so there was not significant change in housing prices within the year. The sqft_living and sqft_lot both have missing values and we chose to use sqft_living15 and sqft_lot15 over those columns, as they had no missing values. We also removed the entry for 33 bedrooms, as it was an outlier. We also replaced NA values for bedrooms with the median number of bedrooms (3) and NA values for bathrooms with the median as well (2.25). We chose to not omit them as there were over 1,000 NA values for each. We chose to remove lat and long as we weren't going to be using them to do mapping. We also removed yr_renovated, as it did not have significant information and most homes were not renovated.

The values we are working with are numerical, with zipcode being a nominal variable and view, condition, and grade being categorical variables.

```
print(summary(house_subset))
```

```
##      price            bedrooms        bathrooms        floors
##  Min.   :  75000   Min.   : 0.000   Min.   :0.00   Min.   :1.000
##  1st Qu.: 322050   1st Qu.: 3.000   1st Qu.:1.75   1st Qu.:1.000
##  Median : 450000   Median : 3.000   Median :2.25   Median :1.500
##  Mean   : 540397   Mean   : 3.371   Mean   :2.12   Mean   :1.494
##  3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.50   3rd Qu.:2.000
##  Max.   :7700000   Max.   :10.000   Max.   :8.00   Max.   :3.500
##    waterfront            view          condition         grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 3.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007374   Mean   :0.2341   Mean   :3.411   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##    sqft_above   sqft_basement       yr_built       zipcode      sqft_living15
##  Min.   : 370   Min.   :   0.0   Min.   :1900   Min.   :98001   Min.   : 399
##  1st Qu.:1200   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:98033   1st Qu.:1490
##  Median :1560   Median :   0.0   Median :1975   Median :98065   Median :1840
##  Mean   :1791   Mean   : 290.8   Mean   :1971   Mean   :98078   Mean   :1987
##  3rd Qu.:2219   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:98118   3rd Qu.:2360
##  Max.   :8860   Max.   :4820.0   Max.   :2015   Max.   :98199   Max.   :6210
##    sqft_lot15
```

```
##  Min.   :   659
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12740
##  3rd Qu.: 10080
##  Max.   :871200
```

```
boxplot(house_subset$bedrooms, house_subset$bathrooms, house_subset$floors,
        main = "House features",
        xlab = "Features",
        ylab = "Amount",
        names = c("Bedrooms", "Bathrooms", "Floors"))
```

Looking at the average housing features, the median number of bedrooms is 3, with a range of 0-10. Most homes have between 3-4 bedrooms. The median number of bathrooms is 2.25, with most homes having between 1.75-2.5 bathrooms. The median number of floors was 1.5, showing a positive skew in number of floors so a higher amount of homes had just one floor.

```
boxplot(house_subset$sqft_living15, house_subset$sqft_above, house_subset$sqft_basement,
        main = "House size",
        xlab = "Area of home",
        ylab = "Size (sq ft)",
        names = c("Total Living Square Feet", "Square Feet Main", "Square Feet of Basement"))
```

The median for total living square feet is 1840sq ft, which is similar to the square feet of the main median of 1460. All three of the housing square feet measures show positive skew, with many outliers for larger than average homes. The basement square feet has a median of 0, which shows that less than half of the homes have a basement.

```
cor(house_subset)
```

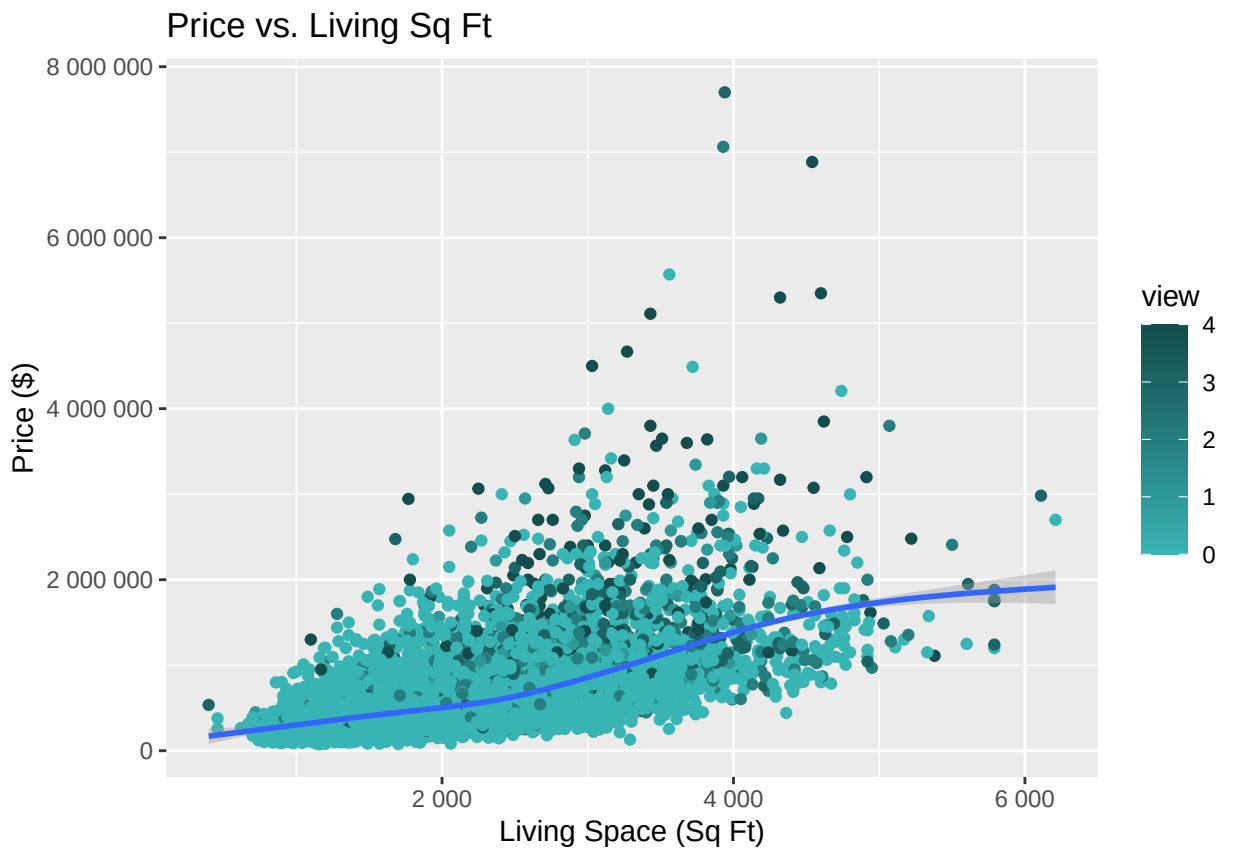
```
##                    price     bedrooms   bathrooms      floors   waterfront
## price          1.00000000  0.318380431  0.51587618  0.25837665  0.263563995
## bedrooms       0.31838043  1.000000000  0.51672443  0.18452553 -0.004452056
## bathrooms      0.51587618  0.516724431  1.00000000  0.48978484  0.067073787
## floors         0.25837665  0.184525531  0.48978484  1.00000000  0.022711927
## waterfront     0.26356400 -0.004452056  0.06707379  0.02271193  1.000000000
## view           0.39793619  0.085189403  0.18181108  0.02907906  0.398161452
## condition      0.03668597  0.025007679 -0.12189206 -0.26155725  0.014882031
## grade          0.66861029  0.369395077  0.65141700  0.46081399  0.082386778
## sqft_above     0.60775789  0.492126360  0.66887810  0.52676405  0.073105464
## sqft_basement  0.32329434  0.308893543  0.27412445 -0.24589323  0.079240717
## yr_built       0.05382920  0.161564916  0.49562788  0.48928857 -0.026432058
## zipcode       -0.05209556 -0.159052644 -0.20044079 -0.06347000  0.030055924
## sqft_living15  0.58669810  0.404481665  0.55511161  0.28347757  0.086635367
## sqft_lot15     0.08049890  0.026444512  0.08209564 -0.01028842  0.032948079
##                      view    condition       grade  sqft_above sqft_basement
## price          0.39793619  0.036685973  0.66861029  0.60775789    0.32329434
## bedrooms       0.08518940  0.025007679  0.36939508  0.49212636    0.30889354
## bathrooms      0.18181108 -0.121892056  0.65141700  0.66887810    0.27412445
```

```
## floors         0.02907906 -0.261557253  0.46081399  0.52676405   -0.24589323
## waterfront     0.39816145  0.014882031  0.08238678  0.07310546    0.07924072
## view           1.00000000  0.045811905  0.25111449  0.16789439    0.27841878
## condition      0.04581191  1.000000000 -0.14530068 -0.15704789    0.17563716
## grade          0.25111449 -0.145300676  1.00000000  0.75844868    0.16784187
## sqft_above     0.16789439 -0.157047894  0.75844868  1.00000000   -0.05266817
## sqft_basement  0.27841878  0.175637161  0.16784187 -0.05266817    1.00000000
## yr_built      -0.05469142 -0.361409396  0.44693073  0.42371232   -0.13492177
## zipcode        0.08475903  0.004327130 -0.18524054 -0.26096111    0.07664762
## sqft_living15  0.28294263 -0.090515070  0.71536412  0.73221499    0.20077923
## sqft_lot15     0.07267963 -0.001916139  0.11656327  0.18917409    0.01667778
##                  yr_built     zipcode sqft_living15   sqft_lot15
## price          0.05382920 -0.05209556    0.58669810  0.080498899
## bedrooms       0.16156492 -0.15905264    0.40448167  0.026444512
## bathrooms      0.49562788 -0.20044079    0.55511161  0.082095643
## floors         0.48928857 -0.06347000    0.28347757 -0.010288422
## waterfront    -0.02643206  0.03005592    0.08663537  0.032948079
## view          -0.05469142  0.08475903    0.28294263  0.072679628
## condition     -0.36140940  0.00432713   -0.09051507 -0.001916139
## grade          0.44693073 -0.18524054    0.71536412  0.116563268
## sqft_above     0.42371232 -0.26096111    0.73221499  0.189174093
## sqft_basement -0.13492177  0.07664762    0.20077923  0.016677779
## yr_built       1.00000000 -0.34810204    0.32459737  0.070452247
## zipcode       -0.34810204  1.00000000   -0.27692453 -0.146407131
## sqft_living15  0.32459737 -0.27692453    1.00000000  0.181418465
## sqft_lot15     0.07045225 -0.14640713    0.18141846  1.000000000
```

The major dependent variable when looking at the data would be the housing price. Different features influence how much a house will sell for. One of the main independent variables is the square footage of the home. The scatterplot shows the price of the home vs. the number of square feet, with the view of the home shown by the color of the points.

```
ggplot(house_subset, aes(x=sqft_living15, y=price, color=view))+
  geom_point()+
  geom_smooth()+
  labs(title="Price vs. Living Sq Ft", x="Living Space (Sq Ft)", y="Price ($)")+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```

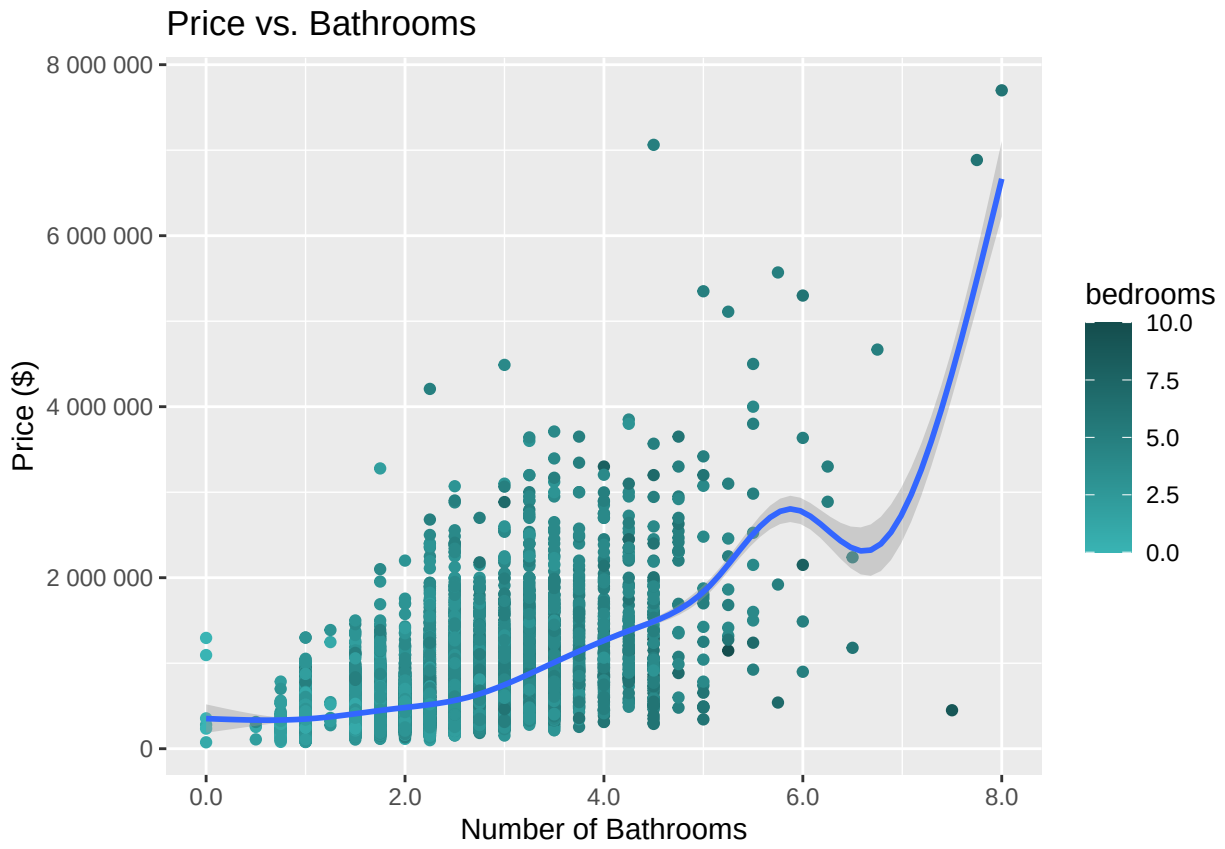
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Another relevant independent variable is the number of bathrooms. While number of bedrooms also has correlation to the price, bathrooms has a higher correlation. The scatterplot shows the number of price of the home vs. number of bathrooms, with the number of bedrooms shown in by color of the points.

```
ggplot(house_subset, aes(x=bathrooms, y=price, color= bedrooms))+
  geom_point()+
  geom_smooth()+
  labs(title="Price vs. Bathrooms", x="Number of Bathrooms", y="Price ($)")+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```

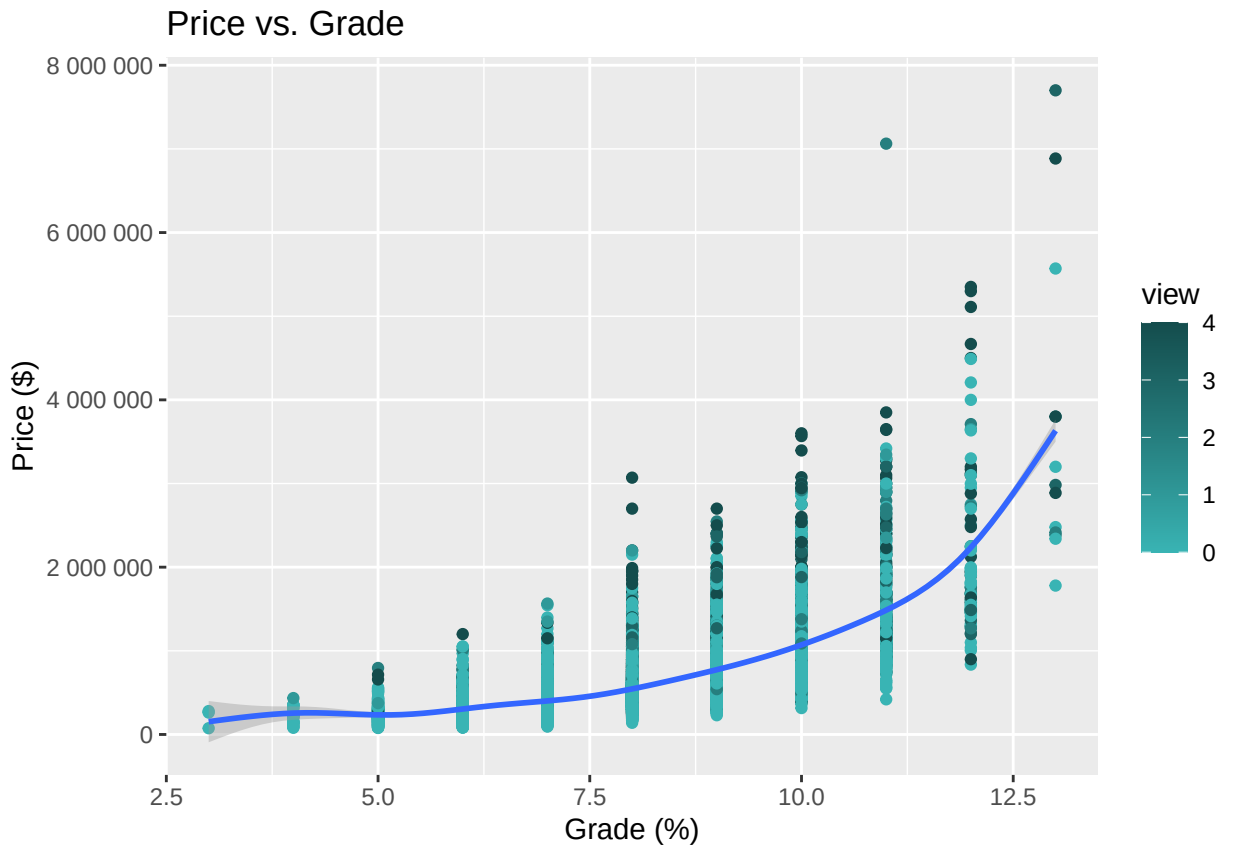
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The grade of the home is an important factor when looking at price, with the higher the grade the higher the price. The scatterplot shows the price vs. the grade of the home, colored by the view that the home has.

```
ggplot(house_subset, aes(x=grade, y=price, color=view))+
  geom_point()+
  geom_smooth()+
  labs(title="Price vs. Grade", x="Grade (%)", y="Price ($)")+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
linearMod <- lm(price ~ sqft_living15, data=house_subset)
print(linearMod)
```

```
##
## Call:
## lm(formula = price ~ sqft_living15, data = house_subset)
##
## Coefficients:
##   (Intercept)  sqft_living15
##        -85466            315
```

The linear model for price vs. square feet would be f(x)=(-85466)+315x where f(x) is the price and x is the square footage of the home.

```
ggplot(house_subset, aes(x=sqft_living15, y=price, color=view))+
  geom_point()+
  labs(title="Price vs. Living Sq Ft", x="Living Space (Sq Ft)", y="Price ($)")+
  geom_abline(aes(intercept=-85466, slope=315))+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```


```
cor.test(house_subset$sqft_living15, house_subset$price, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  house_subset$sqft_living15 and house_subset$price
## t = 103.67, df = 20476, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5776434 0.5956084
## sample estimates:
##       cor
## 0.5866981
```

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = price ~ sqft_living15, data = house_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -846623 -162219  -43529   95228 6544313
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -85465.746   6386.487  -13.38   <2e-16 ***
## sqft_living15    315.013      3.039  103.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 298200 on 20476 degrees of freedom
## Multiple R-squared:  0.3442, Adjusted R-squared:  0.3442
## F-statistic: 1.075e+04 on 1 and 20476 DF,  p-value: < 2.2e-16
```

The linear regression model for price vs. living space has an adjusted R-squared value of 0.3442, which is not a high r-squared value, meaning that much of the variation is not explained by the model. The p-value is very low, which shows that there is significant correlation, so the square footage does have an effect on price.

```
linearMod2 <- lm(price ~ bathrooms, data=house_subset)
print(linearMod2)
```

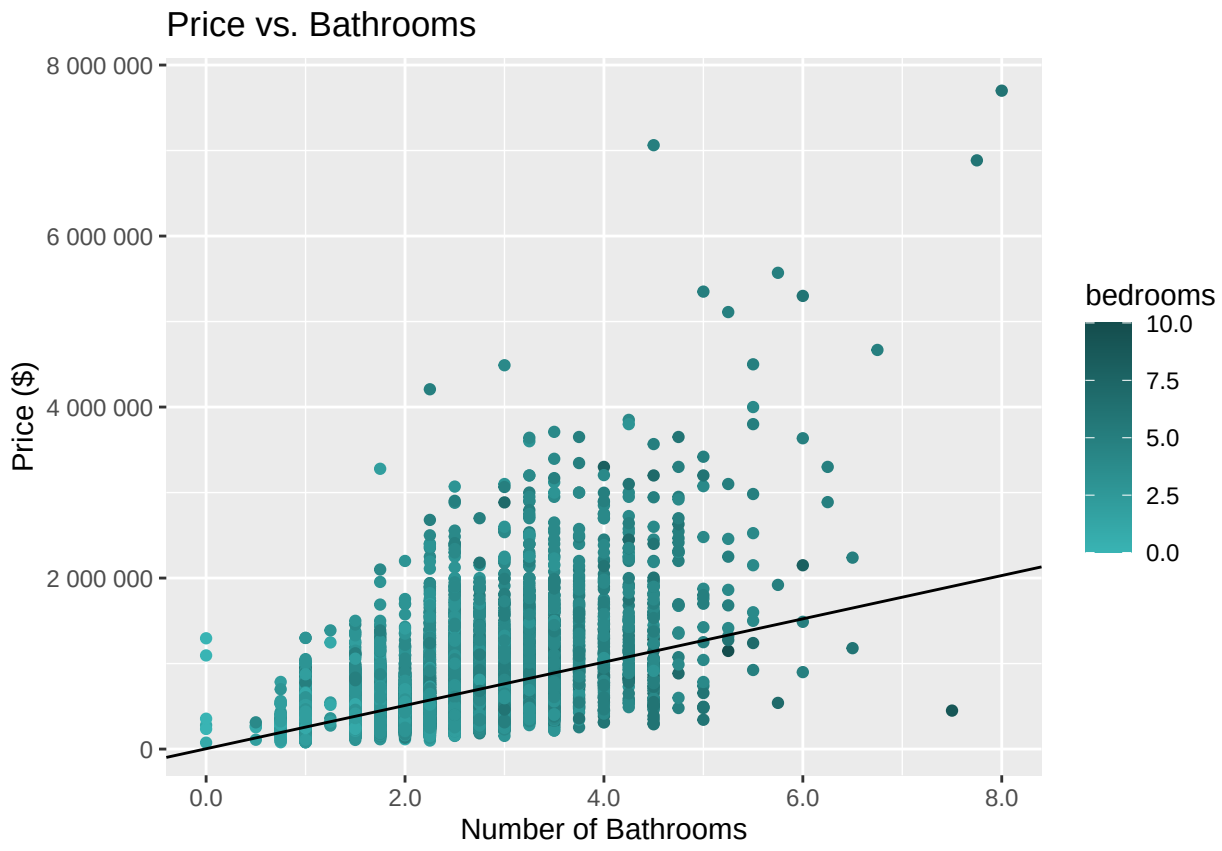
```
##
## Call:
## lm(formula = price ~ bathrooms, data = house_subset)
##
## Coefficients:
## (Intercept)    bathrooms
##        3551       253236
```

```
summary(linearMod2)
```

```
##
## Call:
## lm(formula = price ~ bathrooms, data = house_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1452818  -186801   -42750   113213  5919389
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3551       6608   0.537    0.591
## bathrooms     253236       2939  86.170   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 315400 on 20476 degrees of freedom
## Multiple R-squared:  0.2661, Adjusted R-squared:  0.2661
## F-statistic:  7425 on 1 and 20476 DF,  p-value: < 2.2e-16
```

The linear model for price vs. number of bathrooms would be f(x)=3551+253236x where f(x) is the price and x is the number of bathrooms in the home. The adjusted R-squared value is 0.2661, which is not a very high value, but the p-value shows that number of bathrooms does have a statistically significant correlation to the price.

```
ggplot(house_subset, aes(x=bathrooms, y=price, color= bedrooms))+
  geom_point()+
  geom_abline(aes(intercept=3551, slope=253236))+
  labs(title="Price vs. Bathrooms", x="Number of Bathrooms", y="Price ($)")+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```




```
linearMod3 <- lm(price ~ grade, data=house_subset)
print(linearMod3)
```

```
##
## Call:
## lm(formula = price ~ grade, data = house_subset)
##
## Coefficients:
## (Intercept)        grade
##    -1062438       209321
```

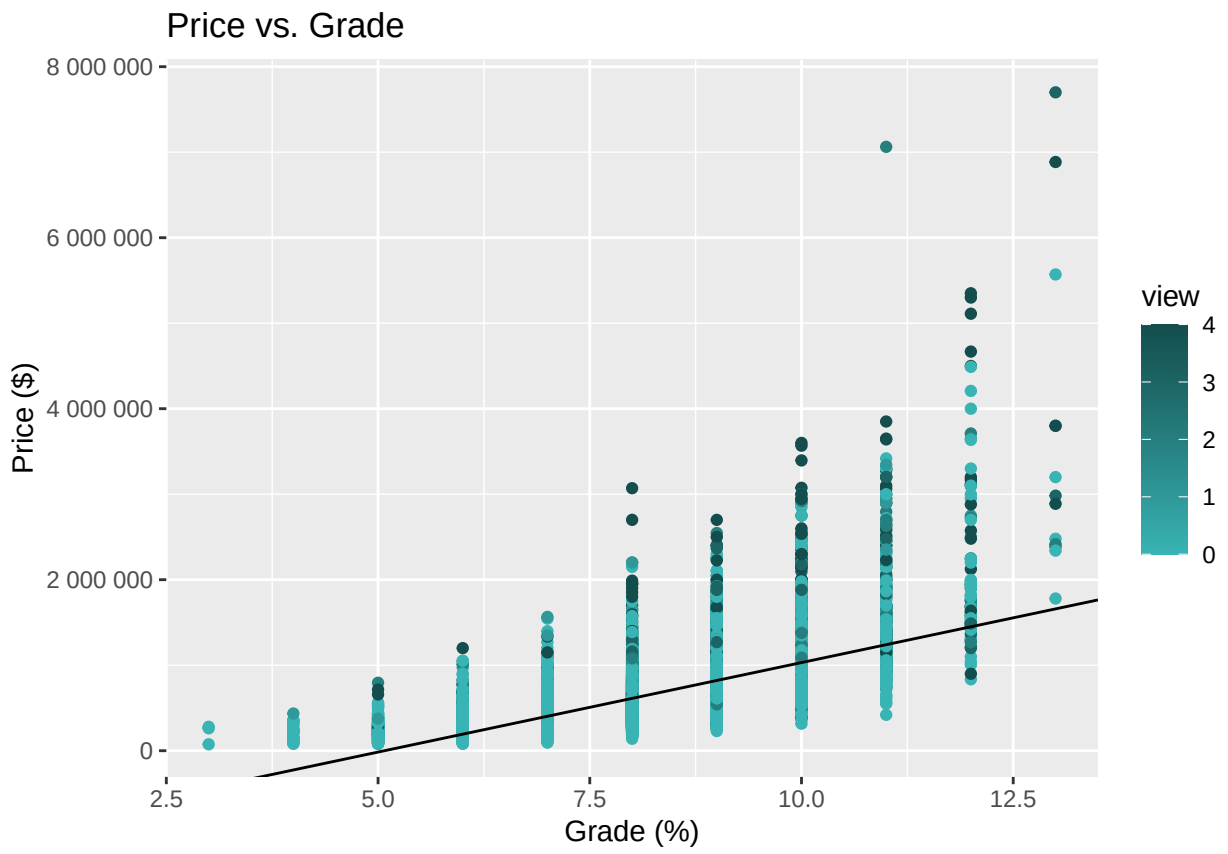
```
summary(linearMod3)
```

```
##
## Call:
## lm(formula = price ~ grade, data = house_subset)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -820098 -152133  -36455   97867 6041260
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1062438      12604   -84.3   <2e-16 ***
## grade         209321       1627   128.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273800 on 20476 degrees of freedom
## Multiple R-squared:  0.447,  Adjusted R-squared:  0.447
## F-statistic: 1.655e+04 on 1 and 20476 DF,  p-value: < 2.2e-16
```

The linear model for price vs. grade of the lot is f(x)=(-1062438)+209321x where f(x) is the price and x is the grade of the home. The adjusted R-squared value is 0.447, which is the highest R-squared value out of the variables, with a similar p-value compared to the other variables, showing that the grade of the home does have a statistically significant correlation to the price.

```
ggplot(house_subset, aes(x=grade, y=price, color=view))+
  geom_point()+
  labs(title="Price vs. Grade", x="Grade (%)", y="Price ($)")+
  geom_abline(aes(intercept=-1062438, slope=209321))+
  scale_x_continuous(labels=scales::number)+
  scale_y_continuous(labels=scales::number)+
  scale_colour_gradient(low = "#39B4B4", high="#144D4D")
```




```
fit <- lm(price ~ sqft_living15 + bathrooms + grade, data=house_subset)
summary(fit)
```

```
##
## Call:
## lm(formula = price ~ sqft_living15 + bathrooms + grade, data = house_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -788779 -147126  -31580  100532 5876666
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.750e+05  1.339e+04  -65.32   <2e-16 ***
## sqft_living15  1.086e+02  3.937e+00   27.58   <2e-16 ***
## bathrooms      5.313e+04  3.315e+03   16.03   <2e-16 ***
## grade          1.419e+05  2.517e+03   56.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266100 on 20474 degrees of freedom
## Multiple R-squared:  0.4777, Adjusted R-squared:  0.4776
## F-statistic:  6241 on 3 and 20474 DF,  p-value: < 2.2e-16
```

Taking all three variables, the multiple regression equation for home price is f(x)=(-875000)+108.6x_1+53130x_2+141900x_3 where f(x) is price of the home, x_1 is the square feet of the home, x_2 is the number of bathrooms, and

x_3 is the grade of the home. The adjusted R-square value is 0.4776, which is higher than the individual linear regression models, showing that it is a better explanation for the variance.

Using ArcGIS, we plotted the homes based on location colored by sales price, which shows that the more expensive homes tend to be clustered around water, which likely contributes to the view rating. The grade of the home is very important when determining home price, as homes with better construction being more expensive as the features are more costly.

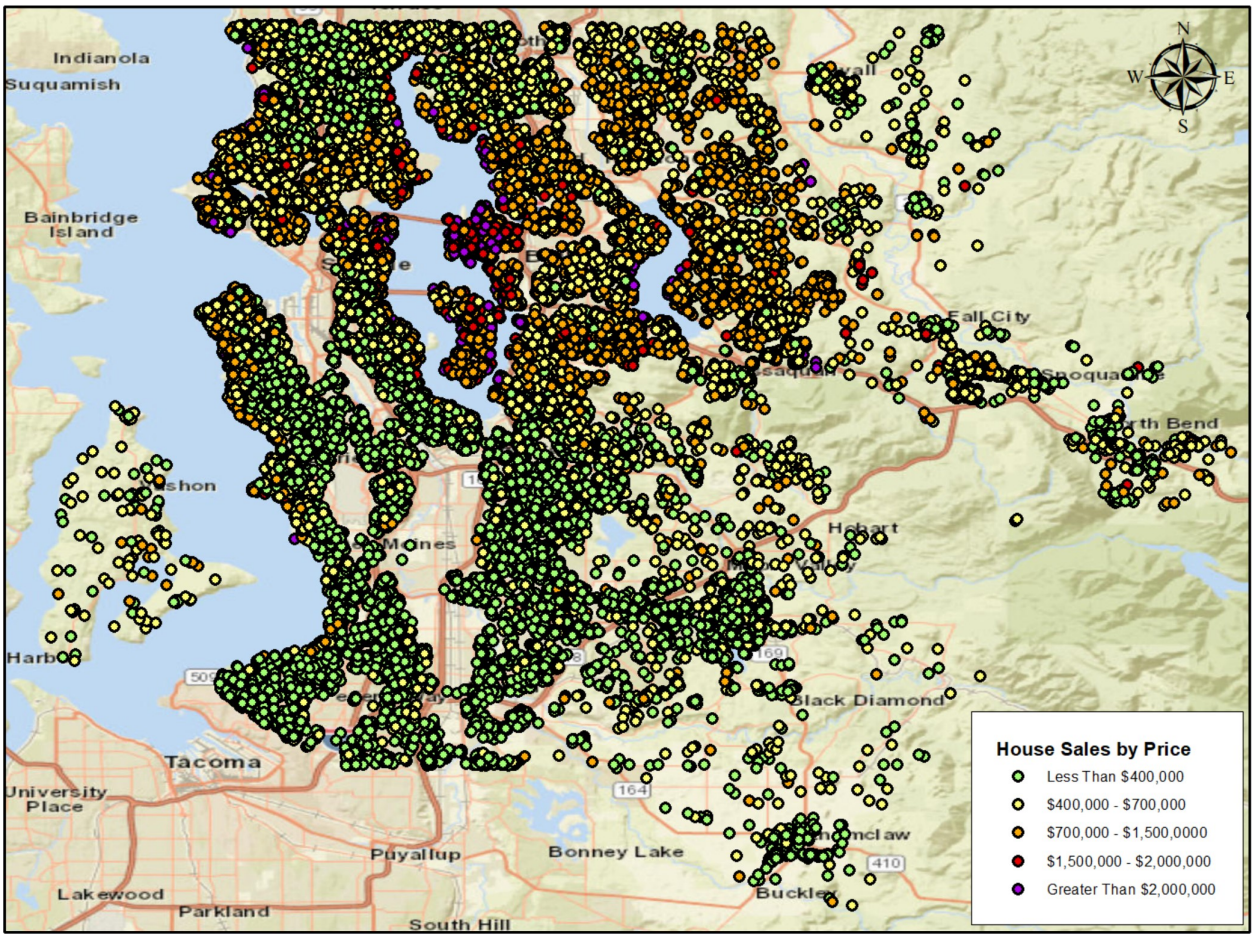


Figure 1: Price based on Location

When determining the sales price of a home, there are many factors that contribute to varying degrees. Location is a factor, and contributes to other variables such as view, whether a home is on a waterfront. The grade of the home is important because the materials that are used to build the home will cause a home to be a higher or lower price. The size of the home is also an important factor. Bathrooms can increase a home price, and has a higher correlation to price than bedrooms. All of these factors and more are taken into account when pricing a home, and are important considerations in buying or selling property.