King County House Price Sales Final Project

By Lahari Kuchibhotla, Stuti Sanghavi, Tanmayi Varansi and Sanjana Ramakandth

Abstract

This report examines the causal effect house size along with a variety of amenities has on house price evidenced by data collected for house sales in King County, USA between 2014 and 2015. After utilizing multivariable regression for the analysis, the conclusion drawn was that as house size increases and amenities such as having a waterfront, a nicer view, renovated home or better construction all play a role in increasing the price of a house.

Introduction

The Housing Sales dataset from Kaggle is about homes sold in King County, Seattle. The observations were made between May 2014 and May 2015. The data contains around 21K rows and 21 variables about various features of the home as well as the price the home was sold at and the location.

Data Cleaning and Manipulation

```
#Reading the csv file
data <- read.csv("kc_house_data.csv")
head(data)
```

	id	date	price	bedroo…	bathroo…	sqft_living	sqft_lot	floors	w
	<dbl>	<fctr>	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>	
1	7129300520	20141013T000000	221900	3	1.00	1180	5650	1	
2	6414100192	20141209T000000	538000	3	2.25	2570	7242	2	
3	5631500400	20150225T000000	180000	2	1.00	770	10000	1	
4	2487200875	20141209T000000	604000	4	3.00	1960	5000	1	
5	1954400510	20150218T000000	510000	3	2.00	1680	8080	1	
6	7237550310	20140512T000000	1225000	4	4.50	5420	101930	1	

6 rows | 1-10 of 22 columns

Basic Descriptive Statistics

```
#Getting descriptive statistics for data
res <- stat.desc(data)
print(res)
```

```
##                        id date        price     bedrooms    bathrooms
## nbr.val      2.161300e+04   NA 2.161300e+04 2.161300e+04 2.161300e+04
## nbr.null     0.000000e+00   NA 0.000000e+00 1.300000e+01 1.000000e+01
## nbr.na       0.000000e+00   NA 0.000000e+00 0.000000e+00 0.000000e+00
## min          1.000102e+06   NA 7.500000e+04 0.000000e+00 0.000000e+00
## max          9.900000e+09   NA 7.700000e+06 3.300000e+01 8.000000e+00
## range        9.899000e+09   NA 7.625000e+06 3.300000e+01 8.000000e+00
## sum          9.899406e+13   NA 1.167293e+10 7.285400e+04 4.570625e+04
## median       3.904930e+09   NA 4.500000e+05 3.000000e+00 2.250000e+00
## mean         4.580302e+09   NA 5.400881e+05 3.370842e+00 2.114757e+00
## SE.mean      1.956666e+07   NA 2.497233e+03 6.326366e-03 5.238720e-03
## CI.mean.0.95 3.835210e+07   NA 4.894760e+03 1.240014e-02 1.026828e-02
## var          8.274629e+18   NA 1.347824e+11 8.650150e-01 5.931513e-01
## std.dev      2.876566e+09   NA 3.671272e+05 9.300618e-01 7.701632e-01
## coef.var     6.280297e-01   NA 6.797542e-01 2.759138e-01 3.641851e-01
##               sqft_living     sqft_lot       floors   waterfront         view
## nbr.val      2.161300e+04 2.161300e+04 2.161300e+04 2.161300e+04 2.161300e+04
## nbr.null     0.000000e+00 0.000000e+00 0.000000e+00 2.145000e+04 1.948900e+04
## nbr.na       0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min          2.900000e+02 5.200000e+02 1.000000e+00 0.000000e+00 0.000000e+00
## max          1.354000e+04 1.651359e+06 3.500000e+00 1.000000e+00 4.000000e+00
## range        1.325000e+04 1.650839e+06 2.500000e+00 1.000000e+00 4.000000e+00
## sum          4.495287e+07 3.265069e+08 3.229650e+04 1.630000e+02 5.064000e+03
## median       1.910000e+03 7.618000e+03 1.500000e+00 0.000000e+00 0.000000e+00
## mean         2.079900e+03 1.510697e+04 1.494309e+00 7.541757e-03 2.343034e-01
## SE.mean      6.247319e+00 2.817461e+02 3.673054e-03 5.884979e-04 5.212562e-03
## CI.mean.0.95 1.224521e+01 5.522432e+02 7.199457e-03 1.153499e-03 1.021701e-02
## var          8.435337e+05 1.715659e+09 2.915880e-01 7.485226e-03 5.872426e-01
## std.dev      9.184409e+02 4.142051e+04 5.399889e-01 8.651720e-02 7.663176e-01
## coef.var     4.415794e-01 2.741815e+00 3.613636e-01 1.147176e+01 3.270620e+00
##                 condition        grade   sqft_above sqft_basement     yr_built
## nbr.val      2.161300e+04 2.161300e+04 2.161300e+04  2.161300e+04 2.161300e+04
## nbr.null     0.000000e+00 0.000000e+00 0.000000e+00  1.312600e+04 0.000000e+00
## nbr.na       0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00 0.000000e+00
## min          1.000000e+00 1.000000e+00 2.900000e+02  0.000000e+00 1.900000e+03
## max          5.000000e+00 1.300000e+01 9.410000e+03  4.820000e+03 2.015000e+03
## range        4.000000e+00 1.200000e+01 9.120000e+03  4.820000e+03 1.150000e+02
## sum          7.368800e+04 1.654880e+05 3.865249e+07  6.300385e+06 4.259933e+07
## median       3.000000e+00 7.000000e+00 1.560000e+03  0.000000e+00 1.975000e+03
## mean         3.409430e+00 7.656873e+00 1.788391e+03  2.915090e+02 1.971005e+03
## SE.mean      4.426414e-03 7.995578e-03 5.632751e+00  3.010436e+00 1.998006e-01
## CI.mean.0.95 8.676097e-03 1.567192e-02 1.104061e+01  5.900677e+00 3.916240e-01
## var          4.234665e-01 1.381703e+00 6.857347e+05  1.958727e+05 8.627973e+02
## std.dev      6.507430e-01 1.175459e+00 8.280910e+02  4.425750e+02 2.937341e+01
## coef.var     1.908657e-01 1.535168e-01 4.630370e-01  1.518221e+00 1.490276e-02
##              yr_renovated      zipcode          lat          long sqft_living15
## nbr.val      2.161300e+04 2.161300e+04 2.161300e+04  2.161300e+04  2.161300e+04
## nbr.null     2.069900e+04 0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00
## nbr.na       0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00
## min          0.000000e+00 9.800100e+04 4.715590e+01 -1.225190e+02  3.990000e+02
## max          2.015000e+03 9.819900e+04 4.777760e+01 -1.213150e+02  6.210000e+03
## range        2.015000e+03 1.980000e+02 6.217000e-01  1.204000e+00  5.811000e+03
## sum          1.824186e+06 2.119759e+09 1.027915e+06 -2.641409e+06  4.293536e+07
## median       0.000000e+00 9.806500e+04 4.757180e+01 -1.222300e+02  1.840000e+03
## mean         8.440226e+01 9.807794e+04 4.756005e+01 -1.222139e+02  1.986552e+03
## SE.mean      2.732259e+00 3.639461e-01 9.425230e-04  9.579273e-04  4.662094e+00
## CI.mean.0.95 5.355429e+00 7.133612e-01 1.847415e-03  1.877608e-03  9.138049e+00
## var          1.613462e+05 2.862788e+03 1.919990e-02  1.983262e-02  4.697612e+05
## std.dev      4.016792e+02 5.350503e+01 1.385637e-01  1.408283e-01  6.853913e+02
## coef.var     4.759105e+00 5.455358e-04 2.913447e-03 -1.152310e-03  3.450155e-01
##                sqft_lot15
## nbr.val      2.161300e+04
## nbr.null     0.000000e+00
## nbr.na       0.000000e+00
## min          6.510000e+02
## max          8.712000e+05
## range        8.705490e+05
## sum          2.759646e+08
## median       7.620000e+03
## mean         1.276846e+04
## SE.mean      1.857255e+02
## CI.mean.0.95 3.640357e+02
## var          7.455182e+08
## std.dev      2.730418e+04
## coef.var     2.138409e+00
```

The above table shows the descriptive statistics for the entire dataset. Majority of the outliers were linked to having a nice view or a waterfront so they were not removed from the model. The one outlier that was removed was a house that had 33 bathrooms and only 1.75 bathrooms. This does not make sense as the house was only 1600 sqft.

Data Cleaning

Before running our models we cleaned the dataset.

- As price was on a much larger scale than the other variables, we scaled it down by 1000 so that when the variables were plotted together it would be easier to analyze.
- For easier interpretation of results, we added another column called **age_of_house** where we subtracted the year build from the current year.
- For easier interpretation of results, we converted the yr_renovated column into a binary column and named it **renovated_factor**.
- As discussed above, removed the outlier that had 33 bathrooms and only 1.75 bathrooms.

```
# Scaling down the price variable for better readability
data$price = data$price/1000

# Adding a new column "age_of_house"
data$age_of_house <- as.integer(format(Sys.Date(), "%Y")) - data$yr_built
data <- data[,-c(1,2,15)]

#Creating a binary column for yr_rennovated where:
# 1 = House is renovated
# 0 = House is not renovated
data$renovated_factor=ifelse(data$yr_renovated != 0, 1, 0)

#Removing the outlier
data <- subset(data, bedrooms <30)

head(data)
```

	price	bedroo...	bathrooms	sqft_living	sqft_lot	floors	waterfront	vi...	condition
	<dbl>	<int>	<dbl>	<int>	<int>	<dbl>	<int>	<int>	<int>
1	221.9	3	1.00	1180	5650	1	0	0	3
2	538.0	3	2.25	2570	7242	2	0	0	3
3	180.0	2	1.00	770	10000	1	0	0	3
4	604.0	4	3.00	1960	5000	1	0	0	5
5	510.0	3	2.00	1680	8080	1	0	0	3
6	1225.0	4	4.50	5420	101930	1	0	0	3

6 rows | 1-10 of 21 columns

Variables used in our analysis:

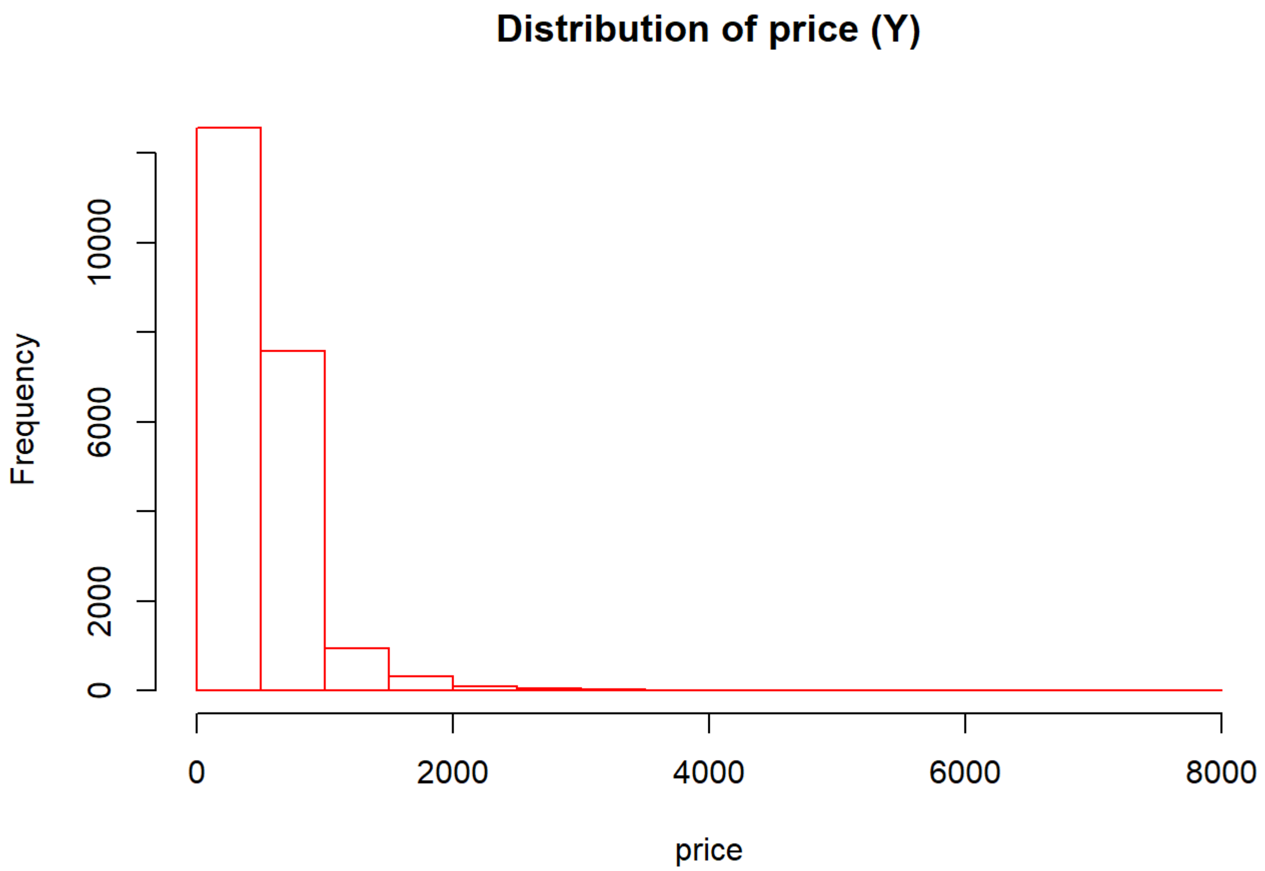
1. **price** - Price of each home sold (outcome variable)
2. **bedrooms** - Number of bedrooms
3. **bathrooms** - Number of bathrooms, where .5 accounts for a room with a toilet but no shower
4. **sqft_living** - Square footage of interior living space of the house
5. **sqft_lot** - Square footage of the land space
6. **floors** - Number of floors
7. **waterfront** - A dummy variable for whether the apartment was overlooking the waterfront or not
8. **view** - An index from 0 to 4 of how good the view of the property was
9. **condition** - An index from 1 to 5 on the condition of the apartment,
10. **grade** - An index from 1 to 13, where 1-6 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.

11. **sqft_above** - The square footage of the interior housing space that is above ground level
12. **sqft_basement** - The square footage of the interior housing space that is below ground level
13. **age_of_house** - Tells us how old the house is
14. **yr_renovated** - The year of the house’s last renovation
15. **zipcode** - What zipcode area the house is in

Exploring the data

Basic Distribution plot

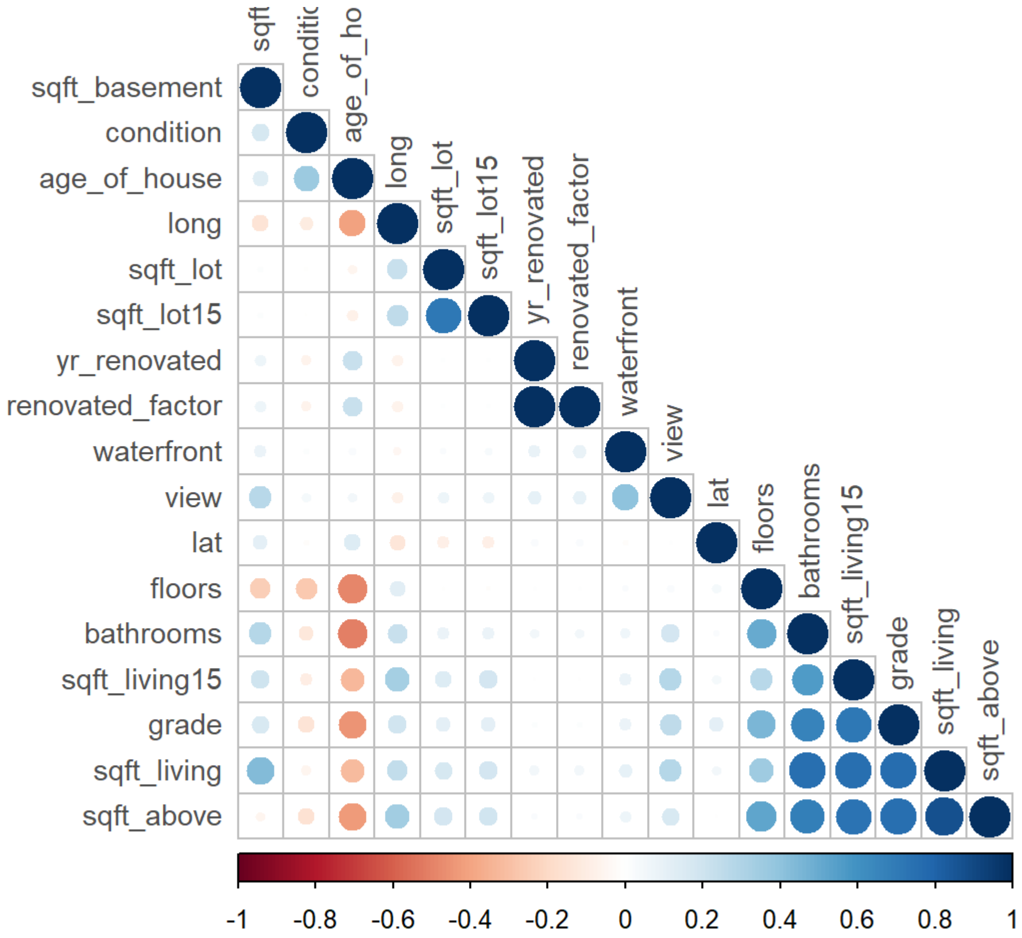
```
#Distribution of the outcome variable (price)

# We can see from the graph that the data is skewed.
hist(data$price,border="red", xlab="price", main = "Distribution of price (Y)")
```



Correlations between the variables

Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors.

```
# Correlation plot
data[, 3:20] %>%
  dplyr::select(-zipcode) %>% cor() %>%
  corrplot::corrplot(type = "lower", order = "hclust", tl.col = "grey30", tl.cex = 0.9)
```



Building our base specification model

Regressing house price on sqft_living

```
#Base specification model between our dependent variable and variable of interest.
sqft_living <- lm(price ~ sqft_living, data = data)

summary(sqft_living)
```

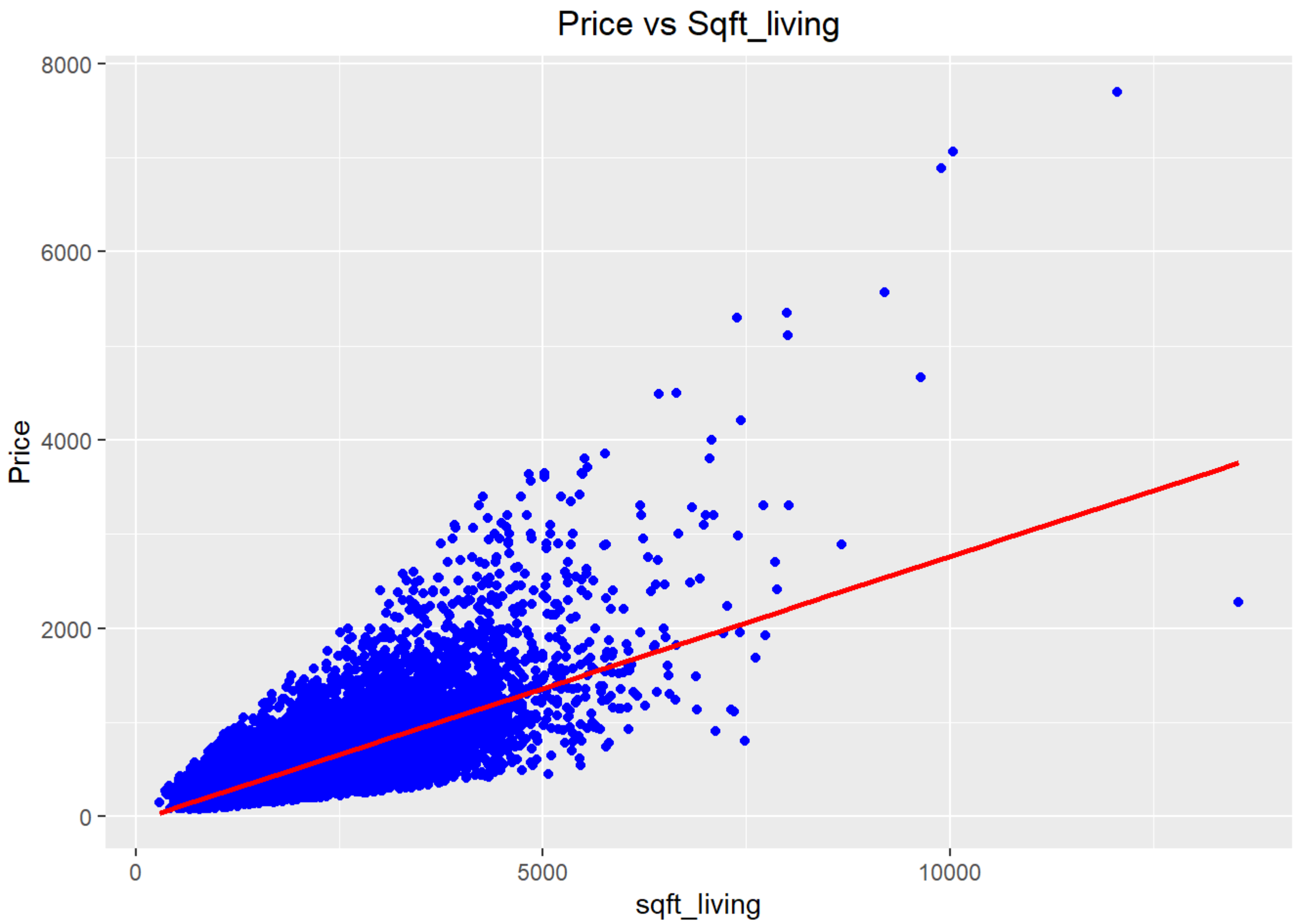
```
##
## Call:
## lm(formula = price ~ sqft_living, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1476.1  -147.5   -24.1   106.2  4362.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.603353   4.402789  -9.904   <2e-16 ***
## sqft_living   0.280629   0.001936 144.922   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 261.5 on 21610 degrees of freedom
## Multiple R-squared:  0.4929, Adjusted R-squared:  0.4928
## F-statistic: 2.1e+04 on 1 and 21610 DF,  p-value: < 2.2e-16
```

From the above results, we can see that sqft_living is positively correlated with the the price of the house. i.e. If the sqft_living increases by 1 foot, the price increases by $280.

Plotting price vs sqft_living

```
# Plot 1: plotting price vs sqft_living
ggplot(data, aes(x=sqft_living, y=price)) + geom_point(col="blue") +
labs(title = "Price vs Sqft_living", x = "sqft_living", y = "Price") +
stat_smooth(method = "lm", col = "red", se=FALSE) + theme(plot.title = element_text(hjust = 0.5
))
```

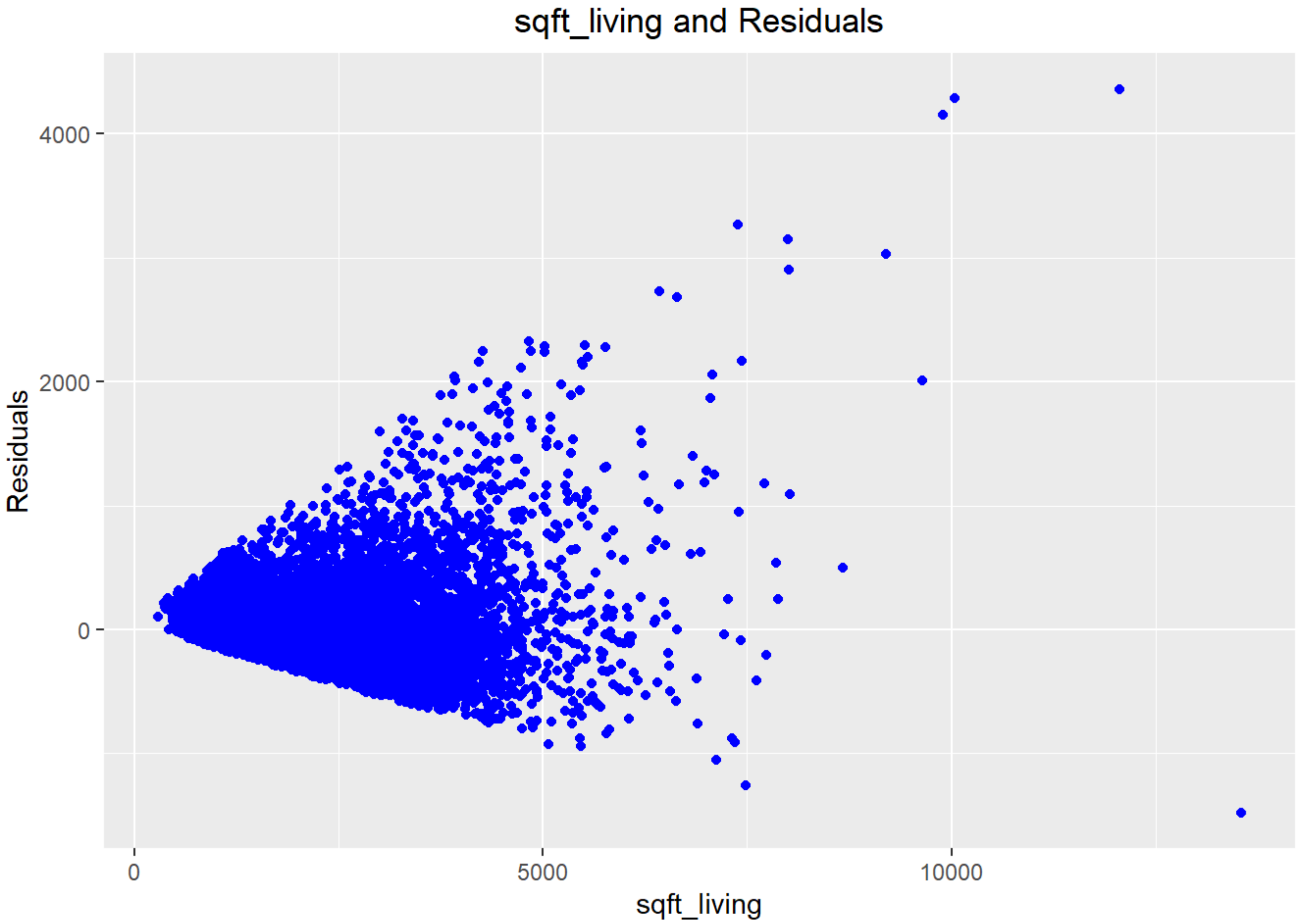
```
## `geom_smooth()` using formula 'y ~ x'
```



Plotting the sqft_living and residuals

Here, we notice that the variance of the residuals increases with sqft_living

```
#Plot 2: plotting the residuals
df_resid = data

df_resid$resid<-resid(sqft_living)

ggplot(df_resid, aes(x=sqft_living, y=resid)) + geom_point(col="blue") +
labs(title = "sqft_living and Residuals", x = "sqft_living", y = "Residuals")  + theme(plot.titl
e = element_text(hjust = 0.5))
```



From the Plot 1 and Plot 2, we see that the errors are heteroskadastic. Therefore we need to correct for standard errors.

To do that, we use CSE function to calculate heteroskedastic-robust standard errors.

```
# CSE function is used to correct the standard errors
cse=function(reg) {

        rob=sqrt(diag(vcovHC(reg, type="HC2")))

        return(rob)
}
```

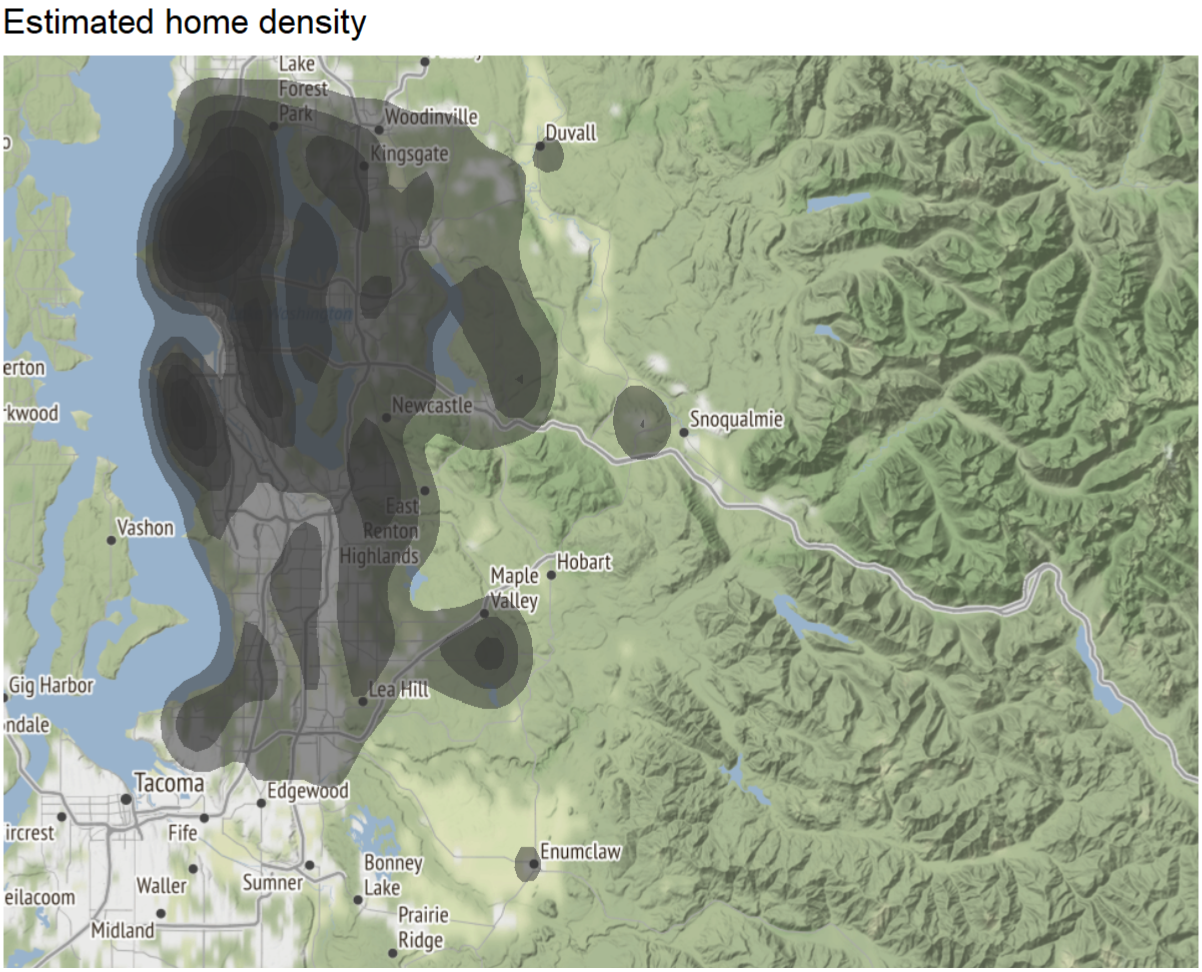
Looking at Potential Control Variables

After researching the relevant literature , we notice that our dependent variable (house price) is broadly dependent on different feature categories as follows:

1. **Size of the house :** We know that size of the house plays an important factor in determining whether or not the house prices increase and we expect that as the size of the house increases, the house price should increase. The size of the house is captured by various variables in the dataset such as bedrooms, bathrooms, sqft_living, floors, sqft_lot, sqft_above, sqft_basement.

2. **Location:** The Location of the house definitely plays an important role in determining the house price. We expect that houses having better location and safer neighbourhoods have higher prices as compared to unsafe locations and similarly, houses near water bodies are more expensive as compared to others. The variables in the dataset indicative of the location of the house include, waterfront, zipcode, lat and long. This is clearly indicated by the graph below where we see that the prices are higher for properties which are located near a waterfront / properties that may be in a safe neighbourhood as compared to others.

```
#Estimating home density and price
map <- qmplot(long, lat, data = data, geom = "blank",
       maptype = "toner-lite") +
  stat_density_2d(aes(fill = price),
                  geom = "polygon", alpha = .5, color = NA) +
  scale_fill_gradient2(low = "steelblue", mid = "seagreen4",
                        high = "indianred", midpoint = 7) +
  labs(title = "Estimated home density")
map
```



3. **View :** If a particular house has a view or not. We expect that houses with better views have higher prices. A house having a city view, ocean view, mountain view etc have higher prices as compared to houses which are more inland. The variable capturing that in the dataset is view.

4. **Conditions:** If the house is in a better condition, it would require lesser maintainence work and the house prices would generally be higher for prices with better condition. If the house has better construction, meaning it has better plumbing, roofing or newer appliances the house would be more expensive. Similarly

if the house had a lower Grade the house would be less expensive.The variables which talk about the conditions of the home in our dataset are, condition and grade

5. **Age of the house:** We would expect that if the age of the house increases, i.e. if its an old house, the price of the house should decrease. Similarly, if it’s a new construction / newly renovated home, the house price should be higher. The variables indicative of the age of the house in our dataset are, age_of_house and yr_renovated variable.

Other factors such as economic growth, supply and demand and interest rates all play a role in house prices. However our dataset does not include these factors so we will use the above mentioned variables to see which of these variables help us explain the effect of sqft_living on house price.

Research Question/Hypothesis:

- **Research Question:** What is the causal effect of sqft_living on house price?
- **Null Hypothesis:** There is no difference in house price when sqft_living increases and amenities such as having view, waterfront, renovation and grade are added.
- **Alternative Hypothesis:** Sqft_living, having a waterfront, view, Renovation and/or grade impact the price of the house.
- **Variable of Interest:** Sqft_living
- **Dependent Variable:** House Price
- **Control Variables:** Waterfront, View, Renovation and Grade

Nested Method Analysis

For the variables that fall under each of the categories mentioned above, we used a nested approach to determine whether they should be included in the final model or not. In the nested approach we run two variable regression models, wherein we see how adding one variable at a time to the base model affects the newly created model and how that compares to the base specification. For all the variables falling under one category, we plan to shortlist and use only one of the variables as control variables. The variable selected from each category would be the one that has the most effect or impact on house price when added in conjunction with the variable of interest. We plan to use only one variable from each category and not more than that because we might face an issue of multicollinearity. However we also plan to select atleast one variable from each category of size, location, view, condition and age to capture the highest impact of these features along with sqft_living on house price.

Models - Part 1

Our Base specification model is when we regress Price with Sqft_living. The Alternative specifications are shown below.

```
#renovated_factor and sqft_living
sqft_renovated <- lm(price ~ sqft_living + as.factor(renovated_factor) , data=data)

#Bathroom and sqft_living
sqft_bathrooms <-  lm(price ~ sqft_living + bathrooms , data=data)

#Bedrooms and sqft_living
sqft_bedrooms <-  lm(price ~ sqft_living + bedrooms , data=data)

#Waterfront and sqft_living
sqft_waterfront <-  lm(price ~ sqft_living + as.factor(waterfront) , data=data)

#Sqft_above and sqft_living
sqft_sabove <- lm((price) ~ sqft_living + sqft_above, data=data)


stargazer(sqft_living,sqft_bathrooms,sqft_bedrooms,sqft_waterfront,sqft_sabove,sqft_renovated, s
e=list(cse(sqft_living),cse(sqft_bathrooms), cse(sqft_bedrooms),cse(sqft_waterfront),cse(sqft_sa
bove),cse(sqft_renovated)),title="Price vs renovated_factor, Bedrooms,Bathrooms,waterfront,squar
e_footage_above ", type="text", star.cutoffs=NA, df=FALSE, digits=3)
```

```
##
## Price vs renovatedBedrooms,Bathrooms,waterfront,square
## =============================================================================================
=
##                                                     Dependent variable:
##                              ----------------------------------------------------------------
-
##                                                 price                     (price)     price
##                                 (1)        (2)        (3)        (4)        (5)        (6)
## ---------------------------------------------------------------------------------------------
-
## sqft_living                    0.281      0.284      0.317      0.273      0.295      0.279
##                               (0.006)    (0.007)    (0.007)    (0.005)    (0.009)    (0.006)
##
## bathrooms                                 -5.162
##                                          (4.744)
##
## bedrooms                                            -62.062
##                                                     (3.542)
##
## as.factor(waterfront)1                                         829.988
##                                                                (59.188)
##
## sqft_above                                                                 -0.019
##                                                                           (0.007)
##
## as.factor(renovated_factor)1                                                         159.957
##                                                                                      (13.464)
##
## Constant                      -43.603    -39.482     90.034    -32.981    -40.885    -46.352
##                               (10.812)   (10.483)   (8.178)    (10.102)   (10.627)   (10.742)
##
## ---------------------------------------------------------------------------------------------
-
## Observations                   21,612     21,612     21,612     21,612     21,612     21,612
## R2                             0.493      0.493      0.508      0.531      0.493      0.501
## Adjusted R2                    0.493      0.493      0.508      0.531      0.493      0.500
## Residual Std. Error           261.454    261.447    257.482    251.515    261.353    259.477
## F Statistic                  21,002.300 10,502.790 11,164.190 12,218.800 10,518.180 10,827.65
0
## =============================================================================================
=
## Note:                                                                                       N
A
```

Correlations:

```
#Bathroom and sqft_living
paste("Correlation between bathrooms and Living Square footage: ", round(cor(data$bathrooms, dat
a$sqft_living), digits=2))
```

```
## [1] "Correlation between bathrooms and Living Square footage:  0.75"
```

```
#Bedrooms and sqft_living
paste("Correlation between Bedrooms and Living Square footage: ", round(cor(data$bedrooms, data
$sqft_living), digits=2))
```

```
## [1] "Correlation between Bedrooms and Living Square footage:  0.59"
```

```
#Waterfront and sqft_living
paste("Correlation between Waterfront and Living Square footage: ", round(cor(data$waterfront, d
ata$sqft_living), digits=2))
```

```
## [1] "Correlation between Waterfront and Living Square footage:  0.1"
```

```
#Sqft_above and sqft_living
paste("Correlation between sqft_above and Living Square footage: ", round(cor(data$sqft_above, d
ata$sqft_living), digits=2))
```

```
## [1] "Correlation between sqft_above and Living Square footage:  0.88"
```

```
#renovated_factor and sqft_living
paste("Correlation between renovated_factor and Living Square footage: ", round(cor(data$renovat
ed_factor, data$sqft_living), digits=2))
```

```
## [1] "Correlation between renovated_factor and Living Square footage:  0.06"
```

```
#Price vs bedroom and bathroom
#Bathroom and price
price_bathrooms <-  lm(price ~ bathrooms , data=data)

#Bedrooms and price
price_bedrooms <-  lm(price ~ bedrooms , data=data)

stargazer(price_bathrooms,price_bedrooms, se=list(cse(price_bathrooms), cse(price_bedrooms)),tit
le="Price vs Bedrooms and Bathrooms", type="text", star.cutoffs=NA, df=FALSE, digits=3)
```

```
##
## Price vs Bedrooms and Bathrooms
## ========================================
##                     Dependent variable:
##                     --------------------
##                            price
##                          (1)        (2)
## ----------------------------------------
## bathrooms            250.332
##                      (6.101)
##
## bedrooms                       127.548
##                                (3.608)
##
## Constant              10.688    110.316
##                      (11.796)  (11.075)
##
## ----------------------------------------
## Observations          21,612    21,612
## R2                    0.276      0.100
## Adjusted R2           0.276      0.099
## Residual Std. Error  312.443    348.399
## F Statistic         8,228.977  2,387.925
## ========================================
## Note:                                NA
```

- When the number of bedrooms or bathrooms increases the price intuitively should increase but we don't see that in the first regression table. We then regressed bedrooms and bathrooms individually with price and they had a positive relationship. But when we add these variables in conjunction with sqft_living, their coefficients are negative. This indicates that these models are suffering from imperfect multicollinearity and hence the variables bedrooms and bathrooms should not be included in the model.
- When waterfront is included, the adjusted R2 increases. The coefficient of sqft_living decreases. This means the model was suffering from upward omitted variable bias. Generally waterfront properties have higher sqft_living and higher prices which explains the upward omitted variable bias. Therefore, we include waterfront as one of our control variables.
- We see a similar problem as that of bedrooms and bathrooms, when sqft_above is added. The standard error increases and as the correlation between sqft_living and sqft_living is 0.87 it is suffering from imperfect multicollinearity as well. Hence we exclude the variable sqft_above from our final model.
- When we include renovated_factor in our model, the adjusted R2 increases and the positive coefficient makes sense intuitively. i.e. If the house is renovated, the price increase by $160000. Also, the coefficient of the sqft_living decreases, which means that the model was suffering from upward omitted variable bias and the variable renovated factor helps us correct for that and hence we decide to include the variable in our model.

Models - Part 2

Next we regressed Price with view, condition, grade, age_of_house and yr_renovated.

```
#Condition and Living Square footage
sqft_condition <-  lm(price ~ sqft_living + condition , data=data)

#Age of the house and Living Square footage
sqft_age_of_house <-  lm(price ~ sqft_living + age_of_house , data=data)

#Grade and Living Square footage
sqft_grade <-  lm(price ~ sqft_living + grade , data=data)

#View and Living Square footage
sqft_view <-  lm(price ~ sqft_living + as.factor(view) , data=data)

#Floors and Living Square footage
sqft_floors <-  lm(price ~ sqft_living + floors , data=data)


stargazer(sqft_living, sqft_condition,sqft_age_of_house,sqft_grade,sqft_view, sqft_floors,  se=l
ist(cse(sqft_living),cse(sqft_condition), cse(sqft_age_of_house),cse(sqft_grade),cse(sqft_view),
cse(sqft_floors)),

                     title="Price vs Condition, Age of the House, Grade and View respectively ",
type="text", star.cutoffs=NA, df=FALSE, digits=3)
```

```
##
## Price vs Condition, Age of the House, Grade and View respectively
## =====================================================================================
##                                           Dependent variable:
##                     -----------------------------------------------------------------
##                                                  price
##                        (1)        (2)        (3)        (4)        (5)       (6)
## -------------------------------------------------------------------------------------
## sqft_living           0.281      0.282      0.305      0.184      0.256     0.279
##                      (0.006)    (0.006)    (0.006)    (0.008)    (0.005)   (0.006)
##
## condition                        43.908
##                                 (2.810)
##
## age_of_house                                2.354
##                                            (0.085)
##
## grade                                                 98.559
##                                                       (3.729)
##
## as.factor(view)1                                                 169.618
##                                                                 (20.421)
##
## as.factor(view)2                                                 127.664
##                                                                 (11.784)
##
## as.factor(view)3                                                 214.295
##                                                                 (19.534)
##
## as.factor(view)4                                                 620.884
##                                                                 (39.226)
##
## floors                                                                      6.475
##                                                                            (3.890)
##
## Constant             -43.603    -197.099   -208.701   -598.157   -14.469   -50.477
##                      (10.812)   (15.558)   (14.716)   (17.748)  (10.310)   (9.423)
##
## -------------------------------------------------------------------------------------
## Observations          21,612     21,612     21,612     21,612    21,612     21,612
## R2                    0.493      0.499      0.525      0.535      0.544     0.493
## Adjusted R2           0.493      0.499      0.525      0.534      0.544     0.493
## Residual Std. Error  261.454    259.900    253.111    250.493    247.900   261.440
## F Statistic         21,002.300 10,757.240 11,929.370 12,407.130 5,158.716 10,504.000
## =====================================================================================
## Note:                                                                              NA
```

Correlations:

```
#Condition and Living Square footage
paste("Condition and Living Square footage: ", round(cor(data$condition, data$sqft_living), digi
ts=2))
```

```
## [1] "Condition and Living Square footage:  -0.06"
```

```
# Age of the house and Living Square footage
paste("Year the age_of_house and Living Square footage: ", round(cor(data$age_of_house, data$sqf
t_living), digits=2))
```

```
## [1] "Year the age_of_house and Living Square footage:  -0.32"
```

```
# View and Living Square footage
paste("View and Living Square footage: ", round(cor(data$view, data$sqft_living), digits=2))
```

```
## [1] "View and Living Square footage:  0.28"
```

```
# Floors and Living Square footage
paste("Floors and Living Square footage: ", round(cor(data$floors, data$sqft_living), digits=2))
```

```
## [1] "Floors and Living Square footage:  0.35"
```

- The correlation between condition with sqft_living is very small (i.e -0.06). That means that the model was suffering from omitted variable bias which is not severe since the values are not significantly different from the base specification. Hence we decide to exclude this from our final model.
- Age_of_house has a downward bias as the coefficient of sqft_living increases and also it is negatively correlated with price. However even though the adjusted R2 of the model increases, it is counter intuitive because an older house should be cheaper and hence we decide to exclude age_of_house from our final model.
- The Floors variable when added to the model isn’t significant and the R2 does not increase. Therefore we will not add it in our model.
- The addition of the other two variables grade and view with sqft_living, indicates upward omitted variable bias has been removed since the coefficient of sqft_living decreases and overall model performance increases. Hence we decide to keep grade and view in our final model.

Final Models:

```
#Model 1
sqft_living <- lm(price ~ sqft_living, data=data)

#Model2
sqft_waterfront <-  lm(price ~ sqft_living + as.factor(waterfront), data=data)

#Model 3
sqft_waterfront_grade <-  lm(price ~ sqft_living + as.factor(waterfront) + grade , data=data)

#Model 4
sqft_waterfront_g_v <-  lm(price ~ sqft_living + as.factor(waterfront) + grade + as.factor(view)
, data=data)

#Model 5
sqft_waterfront_g_v_r <-  lm(price ~ sqft_living + as.factor(waterfront) + renovated_factor + gr
ade + as.factor(view), data=data)

#Model 6
reg6 <- lm(price ~ sqft_living + as.factor(waterfront) + grade + renovated_factor + as.factor(vi
ew) +floors*renovated_factor, data=data)


#stargazer
stargazer(sqft_living, sqft_waterfront, sqft_waterfront_grade, sqft_waterfront_g_v, sqft_waterfr
ont_g_v_r, reg6, se=list(cse(sqft_living),cse(sqft_waterfront),cse(sqft_waterfront_grade), cse(s
qft_waterfront_g_v), cse(sqft_waterfront_g_v_r), cse(reg6)),title="Price vs Square living footag
e with control variables ", type="text", star.cutoffs=NA, df=FALSE, digits=3)
```

```
##
## Price vs Square living footage with control variables
## =====================================================================================
##                                              Dependent variable:
##                         -------------------------------------------------------------
##                                                     price
##                            (1)        (2)        (3)       (4)       (5)       (6)
## -------------------------------------------------------------------------------------
## sqft_living               0.281      0.273      0.177     0.165     0.163     0.163
##                          (0.006)    (0.005)    (0.008)   (0.008)   (0.008)   (0.007)
##
## as.factor(waterfront)1              829.988    825.150   511.762   492.703   495.239
##                                     (59.188)  (59.683)  (75.956)  (75.686)  (75.330)
##
## renovated_factor                                                  129.925   -20.337
##                                                                   (12.057)  (41.723)
##
## grade                                          98.037    94.505    96.134    103.109
##                                                (3.499)   (3.493)   (3.415)   (3.404)
##
## as.factor(view)1                                         167.535   160.188   152.263
##                                                         (19.534)  (19.262)  (19.261)
##
## as.factor(view)2                                         111.457   107.162   100.936
##                                                         (11.075)  (10.980)  (11.060)
##
## as.factor(view)3                                         177.172   169.044   161.201
##                                                         (18.917)  (18.645)  (18.575)
##
## as.factor(view)4                                         383.768   374.421   363.556
##                                                         (45.861)  (45.799)  (45.756)
##
## floors                                                                      -34.632
##                                                                              (3.506)
##
## renovated_factor:floors                                                      100.213
##                                                                             (30.713)
##
## Constant                 -43.603    -32.981   -584.661  -548.597  -560.888  -561.926
##                          (10.812)   (10.102)  (16.812)  (16.286)  (16.033)  (16.263)
##
## -------------------------------------------------------------------------------------
## Observations              21,612     21,612    21,612    21,612    21,612    21,612
## R2                        0.493      0.531      0.572     0.591     0.596     0.598
## Adjusted R2               0.493      0.531      0.572     0.591     0.596     0.598
## Residual Std. Error      261.454    251.515    240.223   234.799   233.369   232.697
## F Statistic           21,002.300 12,218.800 9,623.096 4,461.779 3,985.364 3,219.449
## =====================================================================================
## Note:                                                                              NA
```

For each category of dependent variable price, we found one variable to control each aspect. For example, from size aspect, we just selected the variable of interest sqft_living in our model. We did not include any of the remaining variables from the size category in our final model. If we were to include more than one control variable to control for the same category of our dependent variable, multicollinearity becomes a problem which would throw off our coefficient estimates.To avoid that, we keep **regression 5 (sqft_waterfront_g_v_r)** as our final model. The variables included in the model from each category are:

Grade: because we feel grade or the quality of construction for the house would play a role in how low or high the price of the house would be. If the house has better construction, meaning it has better plumbing, roofing or newer appliances the house would be more expensive. Similarily if the house had a lower Grade the house would be less expensive. Overall,as grade increases it has a positive relationship with price.

Waterfront and View: Houses with better views are expected to be more expensive. The view of a house is also indicative of location. In our case, houses near the water are more expensive and have better views compared to houses more inland. The view variable accounts for two things and explains house price increases or decreases well which is why it is one of our control variables. Overall,as view increases it has a positive relationship with price.

And lastly, the **dummy variable Renovation** indicates if any remodelling was done. Renovations are known to increase the value of older homes and allow them to potentially be priced at a higher rate. If Renovation is 1 it causes an increase in price.

Each of these variables are included in the model and we see that, estimated beta hat of sqft_living is suffering from upward omitted variable bias and by adding these variables we are able to fix that bias. The coefficient estimate of sqft_living decreases as expected. And the model performance increases as we move from model 1 to 5.

Interpretation for each of the variable in the final model includes:

sqft_living

- Estimate : 0.163
- Interpretation: Holding everything else constant, the marginal effect of 1 foot increase in sqft_living increases the house price on an average by $163.

waterfront

- Estimate : 492.703
- Interpretation: Holding everything else constant, a house with a waterfront has house price higher by $492703 on an average as compared to houses without waterfront.

renovated_factor

- Estimate : 129.925
- Interpretation: Holding everything else constant, a house which is renovated increases the price on average by $129925 as compared to a house that is not renovated.

grade

- Estimate : 96.134
- Interpretation: Holding everything else constant, with each level increase in grade (i.e. from poor building construction and design to good building construction and design), the price increases by $96134 on an average.

view1

- Estimate : 167.535

- Interpretation: Holding everything else constant, if the house has view 1 on a scale of 0-4, the price increases by $167535 on an average.

view2

- Estimate : 107.162
- Interpretation: Holding everything else constant, if the house has view 2 on a scale of 0-4, the price increases by $107162 on an average.

view3

- Estimate : 169.044
- Interpretation: Holding everything else constant, if the house has view 3 on a scale of 0-4, the price increases by $169044 on an average.

view4

- Estimate : 374.421
- Interpretation: Holding everything else constant, if the house has view 4 on a scale of 0-4, the price increases by $374421 on an average.

```
#Calculating the correlation between control variables and variable of interest(sqft_living)

round(cor(data$sqft_living,data[c(7,8,10,20)]),2)
```

```
##      waterfront view grade renovated_factor
## [1,]        0.1 0.28  0.76             0.06
```

Another reason we have chosen to add these control variables along with our variable of interest is their correlation. A renovated home could mean anything from new appliances to expanding the actual sqft_living. With the potentially of meaningfully impacting the variable of interest we added the dummy variable to the model. Next the waterfront variable has a subtle relationship with living footage. Homes near the water are known to be more expensive for a multitude of reasons one of which being that the homes are expected to be larger. View is linked a little with waterfront as both of these variables take into consideration the location. As expected certain neighborhoods have larger homes and just in general more sqft_living. Lastly, grade is associated with construction which accounts for things like closet areas and stairways. Things such as enclosed porches are sometimes included in sqft_living which is associated with a higher grade value. Overall, these four control variables play a role in determining price and also have a relationship with our variable of interest.

We also validate our final model findings by running t and f-tests as follows:

Testing for significance of our control variables in model 5 using a T-test

- sqft_living : |20.38| > 1.96
- as.factor(waterfront)1 : |6.50| > 1.96
- renovated_factor : |10.78| > 1.96
- grade : |28.15| > 1.96
- as.factor(view)1 : |8.32| > 1.96

- as.factor(view)2 : |9.76| > 1.96
- as.factor(view)3 : |9.07| > 1.96
- as.factor(view)4 : |8.18| > 1.96

We can see that all the control variables in our final model are statistically significant at 5% significance level (i.e. |t-stat| > 1.96). Therefore, we can safely reject the null hypothesis that any of these variables do not have a coefficient of zero.

Performing F tests to determine the significance of the model

```
### Comparing model with waterfront, grade and view i.e. model 5 with model 1

unrestricted_model <- sqft_waterfront_g_v_r
restricted_model <- sqft_living
anova(restricted_model, unrestricted_model)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	21610	1477223932	NA	NA	NA	NA
2	21603	1176525231	7	300698700	788.7626	0

2 rows

Comparing the Restricted Model with the Unrestricted Model, we can see that the Unrestricted model is a better fit as its p value is less than 0.05 and the F value ie greater than the critical value. Therefore we reject the null hypothesis and state that at least one of the coefficients of the variables is non-zero

```
### Comparing model 6 with model 5(our final model)
unrestricted_model_1 <- reg6
restricted_model_1 <- sqft_waterfront_g_v_r
anova(restricted_model_1, unrestricted_model_1)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	21603	1176525231	NA	NA	NA	NA
2	21601	1169646389	2	6878842	63.51914	3.124555e-28

2 rows

Comparing model 6 and 5, from the above regression table we can see that the beta hat estimator is not changing. As the beta hat has converged we can say that we have gotten our final model. The addition of the interaction variable floors with renovated_factor improves the goodness of fit slightly. We will stick with regression 5 as our final modelbecause floors variable doesn’t add much value even though the R-square increases and also because we do not want to overfit the model. We believe that the control variables we have chosen take into consideration a variety of factors that impact house price as a whole.

Conclusion:

The Square Foot Living of the house has a positive impact on house prices. Furthermore, as new amenities are added to a house, the house price will increase.

Internal Validity Our estimators were unbiased and consistent because as the sample size increases the estimates are getting closer to the true value. Also the estimators are converging to the true value.

- Furthermore, we corrected for Heteroskedasticity by using the cse function which means that the first assumption of the least square i.e. Var(x|u) = 0 is met.
- The second assumption is met because our variable of interest and dependent variable are i.i.d which is assumed for cross-sectional observational data.
- The third assumption is met, as we determined which outliers could be kept and which needed to be removed.
- For the fourth assumption, we removed any control variables that resulted in Multicollinearity.
- As we have met the four assumptions for Least Squares we can say that our estimator is unbiased. E(hatβ0) =β0 and E(hatβ1) =β1.
- Also comparing our restricted model to our unrestricted model our R^2 and adjusted R^2 increased from 49% to 60%. This means our final model explains 11% more of the variation in Y than our base specification model.

External Validity As our data is only house sales for one year in King’s County, we think we would need a larger sample size to generalize our findings.