Univariate Plots Section

This report explores the King county house prices dataset which is available on Kaggle website. In this analysis we will try to find out on which factors is the house price dependend on.

```
##           id            date   price bedrooms bathrooms sqft_living
## 1 7129300520 20141013T000000  221900        3      1.00        1180
## 2 6414100192 20141209T000000  538000        3      2.25        2570
## 3 5631500400 20150225T000000  180000        2      1.00         770
## 4 2487200875 20141209T000000  604000        4      3.00        1960
## 5 1954400510 20150218T000000  510000        3      2.00        1680
## 6 7237550310 20140512T000000 1225000        4      4.50        5420
##   sqft_lot floors waterfront view condition grade sqft_above sqft_basement
## 1     5650      1          0    0         3     7       1180             0
## 2     7242      2          0    0         3     7       2170           400
## 3    10000      1          0    0         3     6        770             0
## 4     5000      1          0    0         5     7       1050           910
## 5     8080      1          0    0         3     8       1680             0
## 6   101930      1          0    0         3    11       3890          1530
##   yr_built yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1     1955            0   98178 47.5112 -122.257          1340       5650
## 2     1951         1991   98125 47.7210 -122.319          1690       7639
## 3     1933            0   98028 47.7379 -122.233          2720       8062
## 4     1965            0   98136 47.5208 -122.393          1360       5000
## 5     1987            0   98074 47.6168 -122.045          1800       7503
## 6     2001            0   98053 47.6561 -122.005          4760     101930
```

```
## [1] 21613    21
```

Our dataset consists of 21 variables, with almost 21613 observations.

```
## 'data.frame':    21613 obs. of  21 variables:
##  $ id           : num  7129300520 6414100192 5631500400 2487200875 1954400510 ...
##  $ date         : Factor w/ 372 levels "20140502T000000",..: 165 221 291 221 284 11 57 25
##  $ price        : num  221900 538000 180000 604000 510000 ...
##  $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
##  $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
```

```
##  $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
##  $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

```
##        id                           date           price
##  Min.   :   1000102   20140623T000000:  142   Min.   :  75000
##  1st Qu.:2123049194   20140625T000000:  131   1st Qu.: 321950
##  Median :3904930410   20140626T000000:  131   Median : 450000
##  Mean   :4580301521   20140708T000000:  127   Mean   : 540088
##  3rd Qu.:7308900445   20150427T000000:  126   3rd Qu.: 645000
##  Max.   :9900000190   20150325T000000:  123   Max.   :7700000
##                       (Other)        :20833
##     bedrooms        bathrooms      sqft_living       sqft_lot
##  Min.   : 0.000   Min.   :0.000   Min.   :  290   Min.   :    520
##  1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040
##  Median : 3.000   Median :2.250   Median : 1910   Median :   7618
##  Mean   : 3.371   Mean   :2.115   Mean   : 2080   Mean   :  15107
##  3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688
##  Max.   :33.000   Max.   :8.000   Max.   :13540   Max.   :1651359
##
##      floors        waterfront            view          condition
##  Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000
##  Median :1.500   Median :0.000000   Median :0.0000   Median :3.000
##  Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409
##  3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000
##  Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000
##
##      grade          sqft_above   sqft_basement       yr_built
##  Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900
##  1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951
##  Median : 7.000   Median :1560   Median :   0.0   Median :1975
##  Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971
##  3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997
##  Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015
```

```
##
##   yr_renovated       zipcode           lat             long
##  Min.   :   0.0   Min.   :98001   Min.   :47.16   Min.   :-122.5
##  1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
##  Median :   0.0   Median :98065   Median :47.57   Median :-122.2
##  Mean   :  84.4   Mean   :98078   Mean   :47.56   Mean   :-122.2
##  3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :2015.0   Max.   :98199   Max.   :47.78   Max.   :-121.3
##
##  sqft_living15    sqft_lot15
##  Min.   : 399   Min.   :   651
##  1st Qu.:1490   1st Qu.:  5100
##  Median :1840   Median :  7620
##  Mean   :1987   Mean   : 12768
##  3rd Qu.:2360   3rd Qu.: 10083
##  Max.   :6210   Max.   :871200
##
```

```
##  [1] "id"            "date"          "price"         "bedrooms"
##  [5] "bathrooms"     "sqft_living"   "sqft_lot"      "floors"
##  [9] "waterfront"    "view"          "condition"     "grade"
## [13] "sqft_above"    "sqft_basement" "yr_built"      "yr_renovated"
## [17] "zipcode"       "lat"           "long"          "sqft_living15"
## [21] "sqft_lot15"
```

The metatdata description of the variables is not provided in Kaggle. Although some of them are easy to understand there are a few which is ambigious.
The dataset contains the following 21 variables:

id : The id of the house
date : The date when this information was taken.(dates between May 2014 and May 2015.)
price : Price of the house
bedrooms : No of bedrooms
bathrooms : No of bathrooms
sqft_living : Living area in square feet
sqft_lot : Lot area in square feet
floors : No of floors
waterfront : House has a waterfront or not view : Views of the houses
condition : House condition ranging from 1 to 5
grade : House grade ranging from 1 to 13
sqft_above: Living area excluding the basement sqft_basement : Basement area
yr_built : The year the house was build yr_renovated : The year the house was renovated
zipcode : Zipcode of the house lat : Latitude long : Longitude
sqft_living15 : The average house square footage of the 15 closest houses
sqft_lot15 : The average lot square footage of the 15 closest houses

The columns id,date are redundant in this analysis so it is better to remove them.

```
##     price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
## 1  221900        3      1.00        1180     5650      1          0    0
## 2  538000        3      2.25        2570     7242      2          0    0
## 3  180000        2      1.00         770    10000      1          0    0
## 4  604000        4      3.00        1960     5000      1          0    0
## 5  510000        3      2.00        1680     8080      1          0    0
## 6 1225000        4      4.50        5420   101930      1          0    0
##   condition grade sqft_above sqft_basement yr_built yr_renovated zipcode
## 1         3     7       1180             0     1955            0   98178
## 2         3     7       2170           400     1951         1991   98125
## 3         3     6        770             0     1933            0   98028
## 4         5     7       1050           910     1965            0   98136
## 5         3     8       1680             0     1987            0   98074
## 6         3    11       3890          1530     2001            0   98053
##       lat     long sqft_living15 sqft_lot15
## 1 47.5112 -122.257          1340       5650
## 2 47.7210 -122.319          1690       7639
## 3 47.7379 -122.233          2720       8062
## 4 47.5208 -122.393          1360       5000
## 5 47.6168 -122.045          1800       7503
## 6 47.6561 -122.005          4760     101930
```

Preparing the dataset

Adding and modifying columns to be used later.

Creating a categorical variable for price and assigning them to “cheap”,“moderate”,“high”,“expensive”,“max”. Creating categorical variable for the year_built column. Created 5 categories for this.

Converting the varibles waterfront,condition,floors,grade,bedrooms into factor variables.

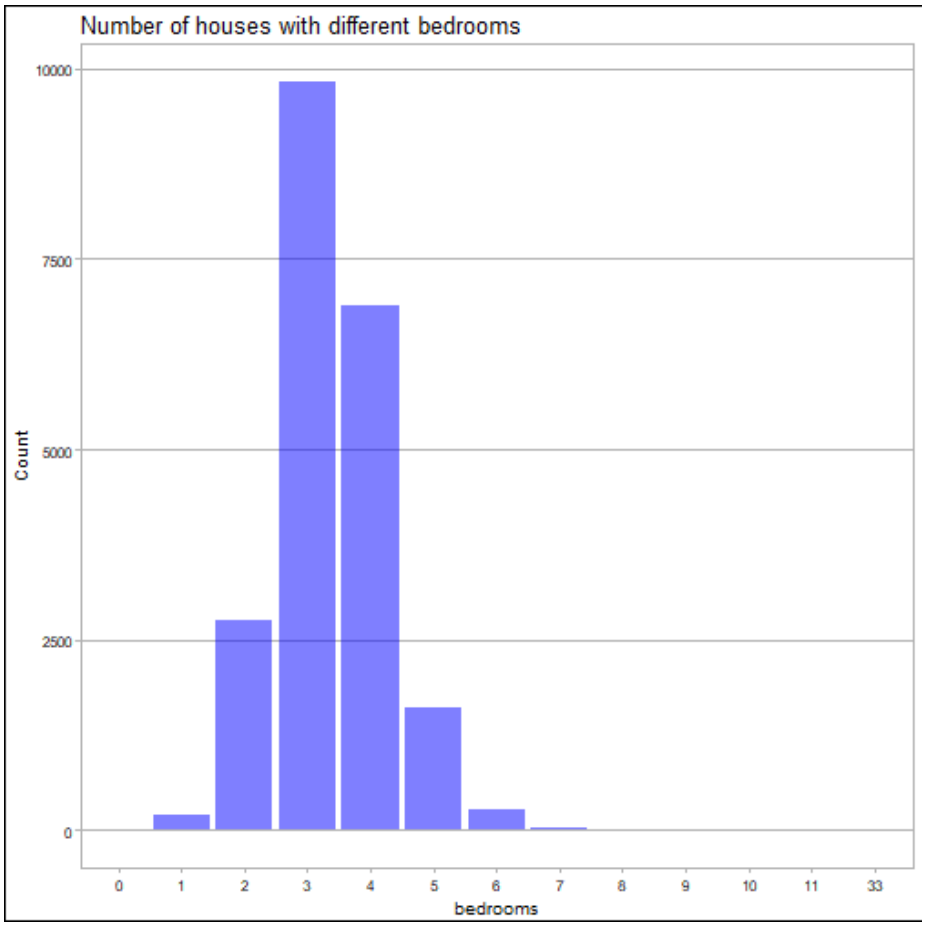
```
## 'data.frame':    21613 obs. of  21 variables:
##  $ price        : num  221900 538000 180000 604000 510000 ...
##  $ bedrooms     : Factor w/ 13 levels "0","1","2","3",..: 4 4 3 5 4 5 4 4 4 4 ...
##  $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
##  $ floors       : Factor w/ 6 levels "1","1.5","2",..: 1 3 1 1 1 1 3 1 1 3 ...
##  $ waterfront   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition    : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : Factor w/ 12 levels "1","3","4","5",..: 6 6 5 6 7 10 6 6 6 6 ...
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
```

```
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
##  $ price_cat    : Factor w/ 5 levels "Cheap","Moderate",..: 2 3 1 3 3 4 2 2 2 2 ...
##  $ yr_builtR    : chr  "1951-1975" "1951-1975" "1926-1950" "1951-1975" ...
```

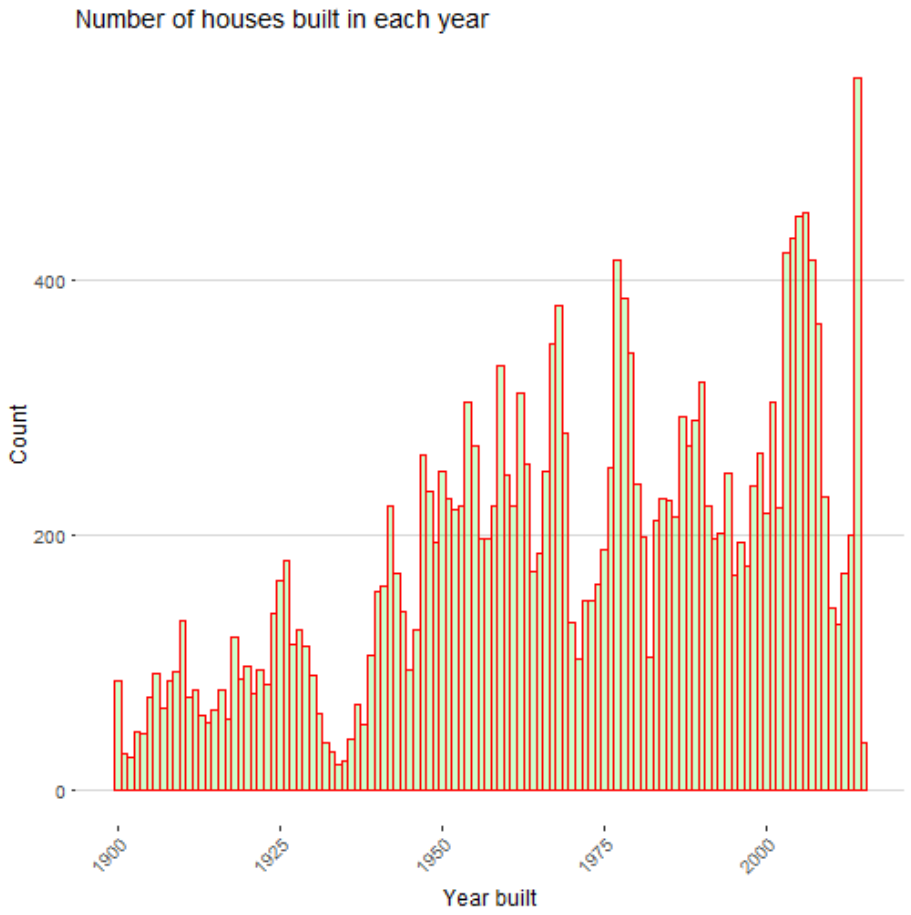
```
##     price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
## 1  221900        3      1.00        1180     5650      1          0    0
## 2  538000        3      2.25        2570     7242      2          0    0
## 3  180000        2      1.00         770    10000      1          0    0
## 4  604000        4      3.00        1960     5000      1          0    0
## 5  510000        3      2.00        1680     8080      1          0    0
## 6 1225000        4      4.50        5420   101930      1          0    0
##   condition grade sqft_above sqft_basement yr_built yr_renovated zipcode
## 1         3     7       1180             0     1955            0   98178
## 2         3     7       2170           400     1951         1991   98125
## 3         3     6        770             0     1933            0   98028
## 4         5     7       1050           910     1965            0   98136
## 5         3     8       1680             0     1987            0   98074
## 6         3    11       3890          1530     2001            0   98053
##       lat     long sqft_living15 sqft_lot15 price_cat yr_builtR
## 1 47.5112 -122.257          1340       5650  Moderate 1951-1975
## 2 47.7210 -122.319          1690       7639      High 1951-1975
## 3 47.7379 -122.233          2720       8062     Cheap 1926-1950
## 4 47.5208 -122.393          1360       5000      High 1951-1975
## 5 47.6168 -122.045          1800       7503      High 1976-2000
## 6 47.6561 -122.005          4760     101930 Expensive 2001-2015
```

Libraries used:

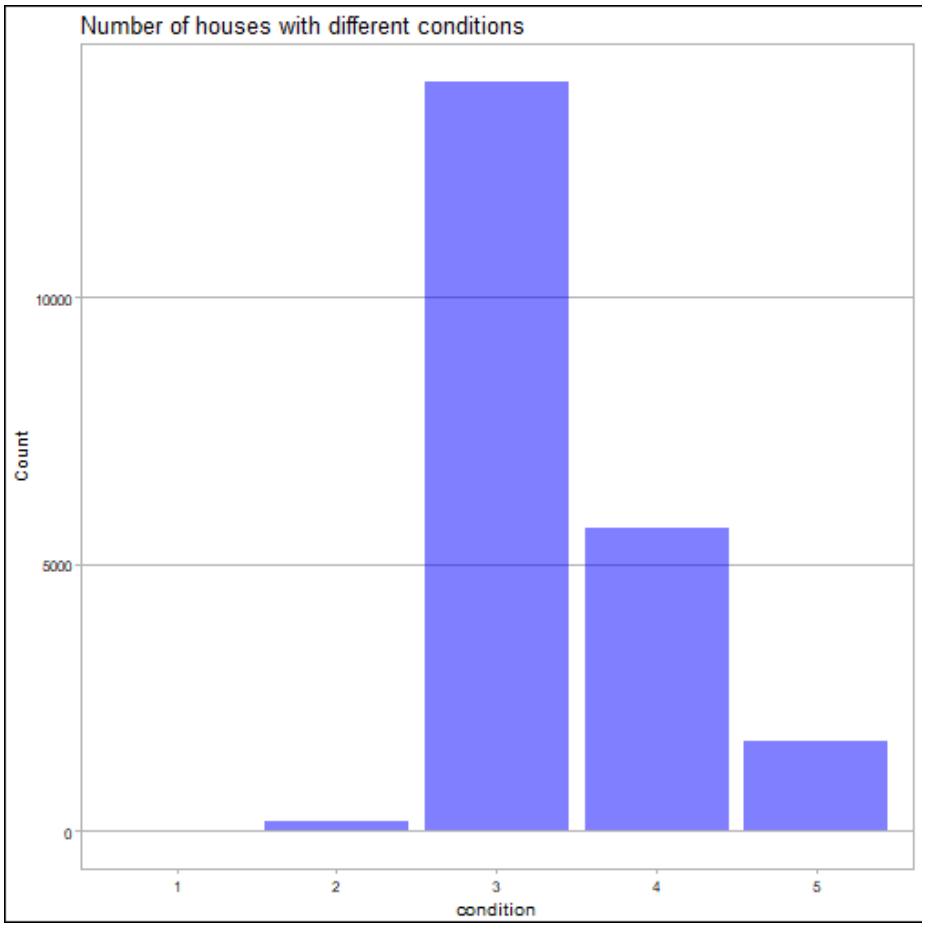
Univariate Plots Section



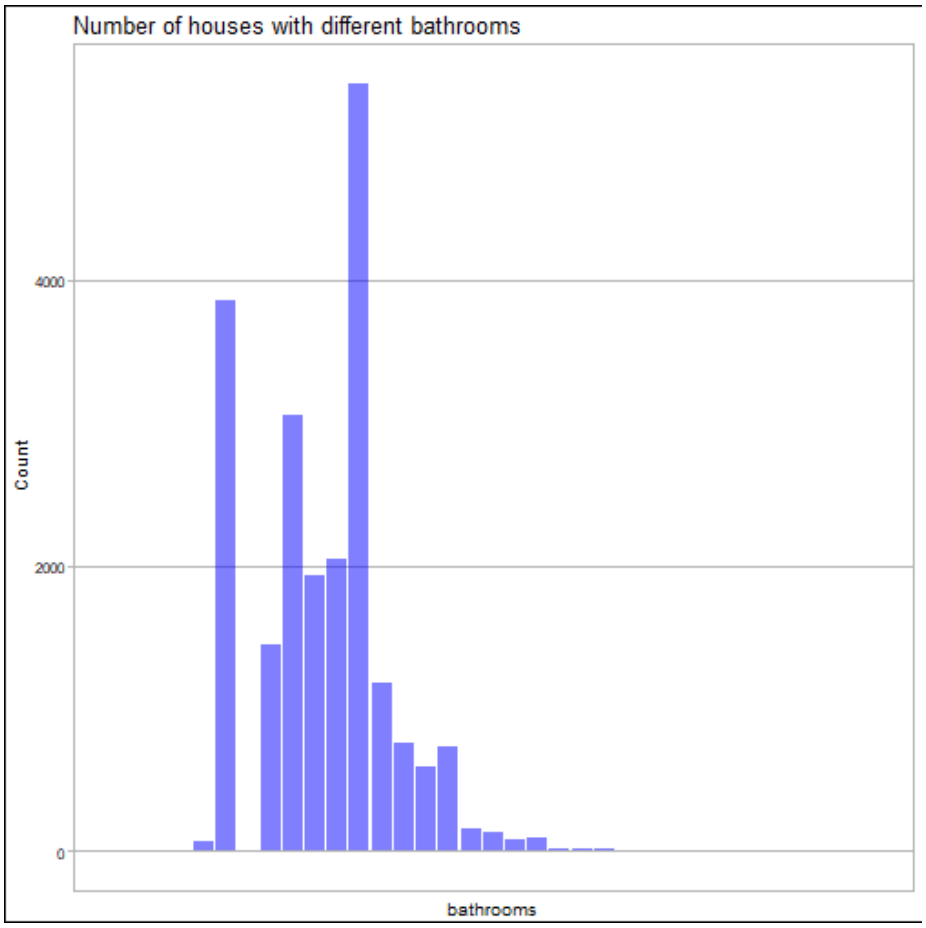
Out of the houses in King county a major portion of them are of 3 bedrooms which is ideal for small families.



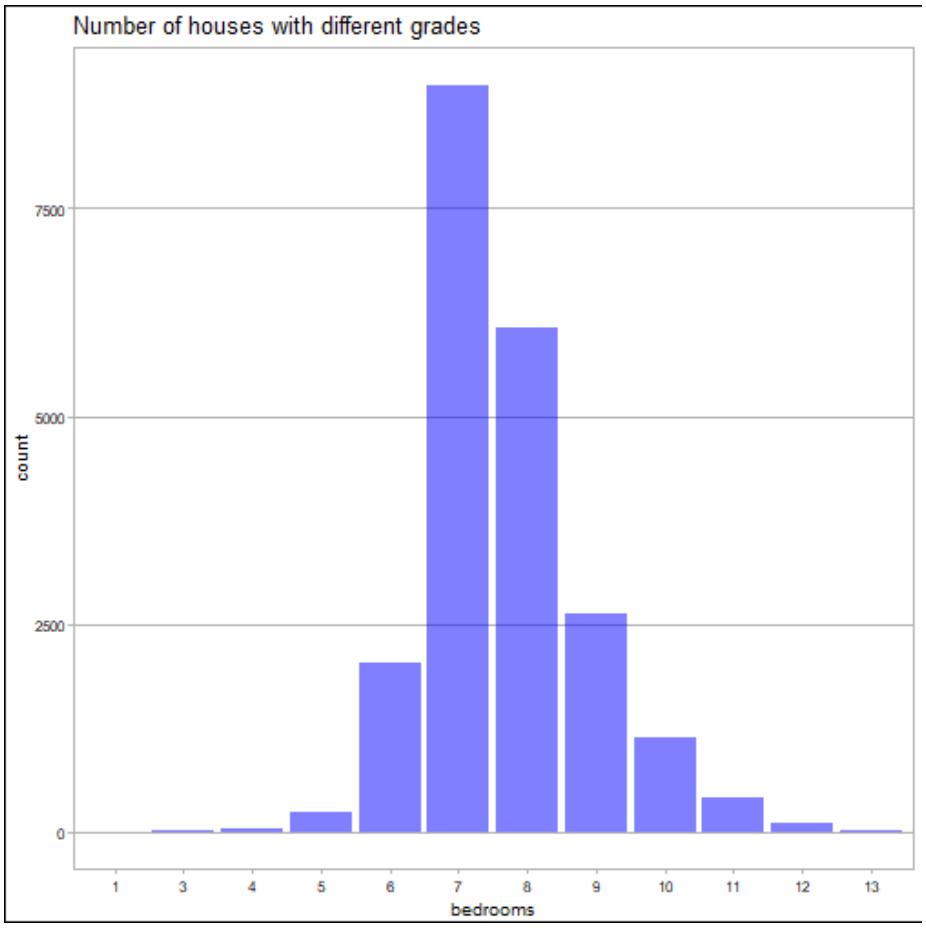
There is a gradual increase in the number of houses built across the years with the maximum number of houses built in 2014.



Most of the houses have a condition of 3 in King County.



The most common bathrooms are 1 full bath and 2 full bath and 1 powder room in the houses in king county.



The most common grade is 7 in King county with grades 1,2,3,13 being nonexistent

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     290    1427    1910    2080    2550   13540
```

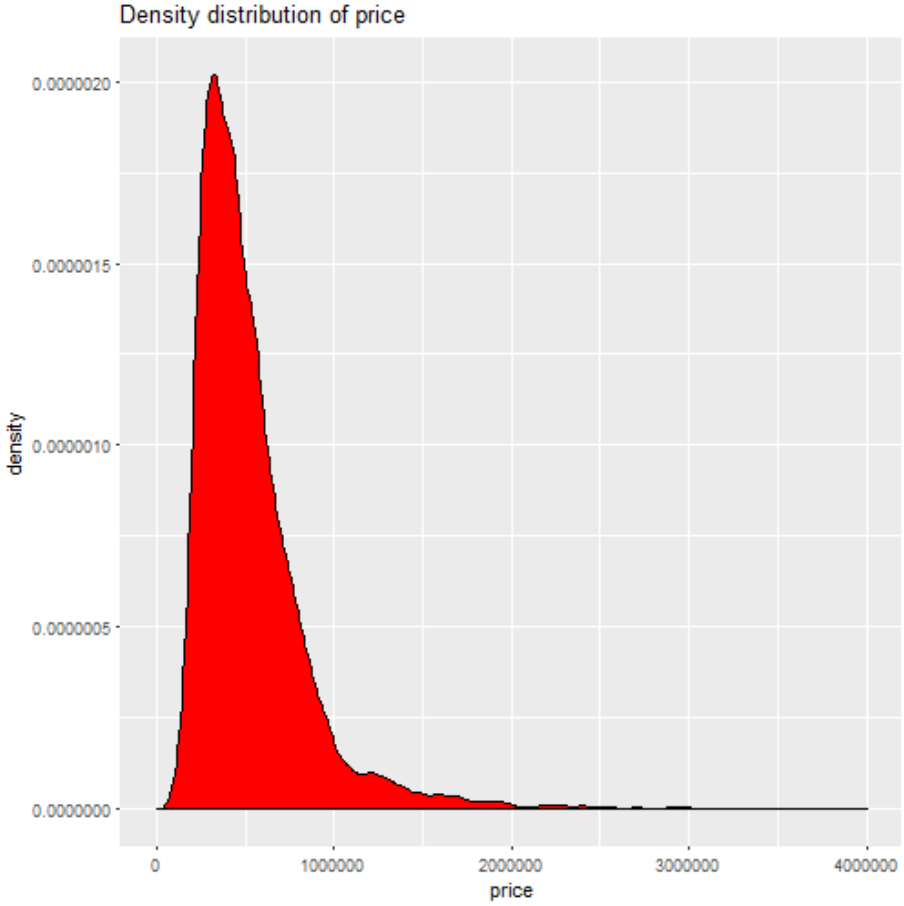
The median sqft_living is 1910 sq feet with the mean being 2080 sq feet.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     520    5040    7618   15107   10688 1651359
```

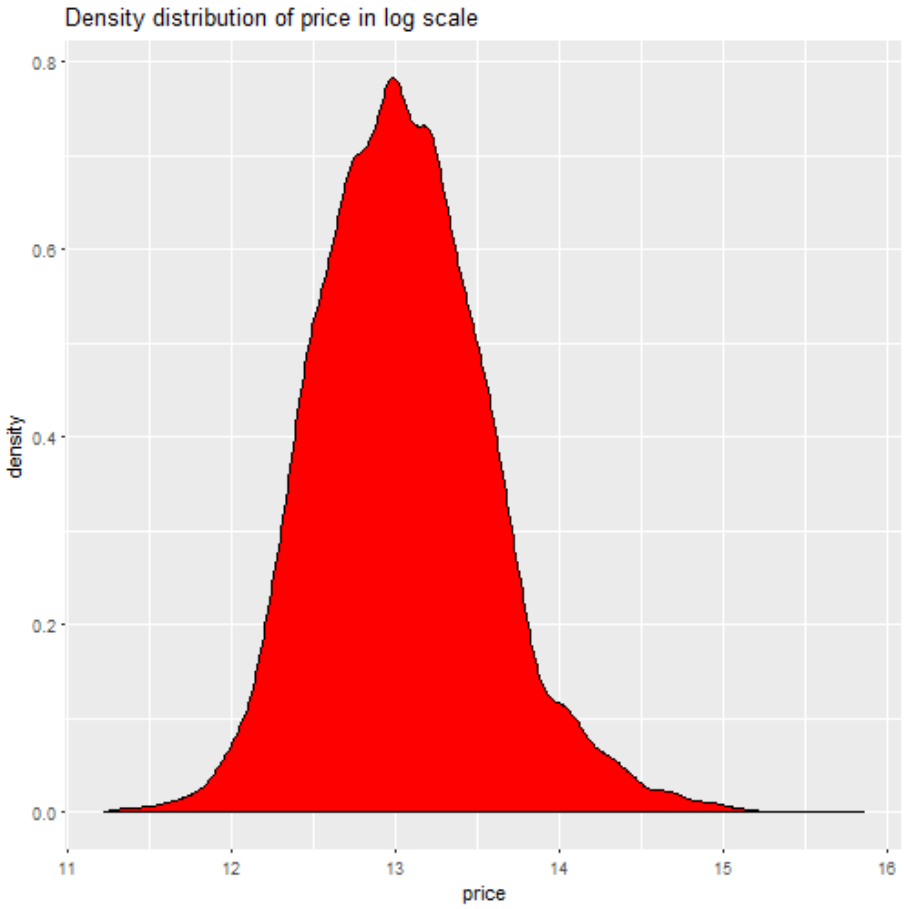
The median sqft_lot is 7618 sq feet with the mean being 15107 sq feet. The max lot size is quite huge compared to the the other houses and may be considered as an outlier.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   75000  321950  450000  540088  645000 7700000
```

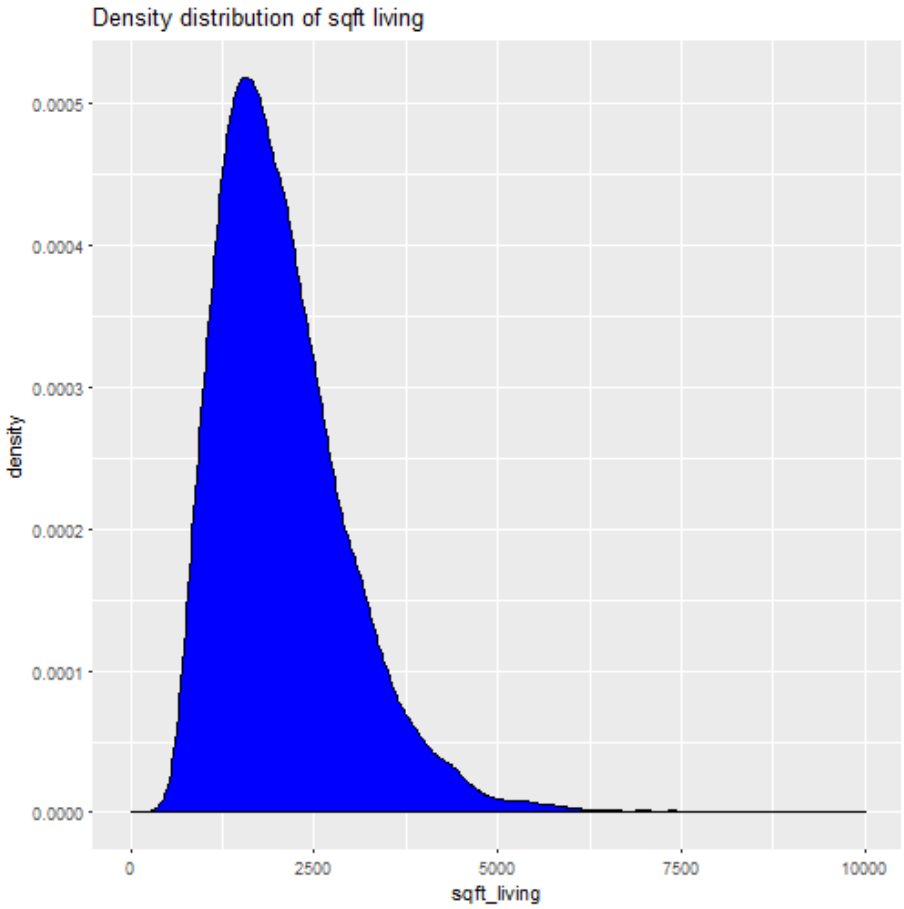
The median price of the houses in king county is around 45000.



The density distribution gives a plot that is right skewed.



When the price is change to log of price the distribution changes to normal distrbution from a skewed distribution.



Density distribution for sqft_living of the houses and it is right skewed.

```
##    zipcode count         city
## 43   98103   602      Seattle
## 24   98038   590 Maple Valley
## 50   98115   583      Seattle
## 29   98052   574      Redmond
## 52   98117   553      Seattle
```

These 5 zipcodes have the highest number of houses and the top 3 cities are Seattle, Maple valley and Redmond.

Univariate Analysis

What is the structure of your dataset?

There are 21613 houses in the dataset with 21 features. The variables condition, grade, and view, waterfront are ordered factor variables. From the analysis done it seems that the most common condition is 3. The median price of the houses is 45000,the most common bedroom is 3,the most common grade is 7. As it is not mentioned in the metatdata which of the values of the factor variables is better it has to be found out later in bivariate plots.

What is/are the main feature(s) of interest in your dataset?

The main features in the data set are price and sqft_living. I'd like to determine which features are best for predicting the price of a houses and which features have a strong correlation with price. I suspect sqft_living,sqft_lot, bedrooms and grade can help to build a predictive model to price of houses.

**What other features in the dataset do you think will help support your **

investigation into your feature(s) of interest?

I would like to explore condition, sqft_above, yr_build,yr_renovated and zipcode features of the dataset and find insights from the data.

Did you create any new variables from existing variables in the dataset?

Yes I created price_cat,yr_buildR to have caegorical variables for conducting furthur plots.

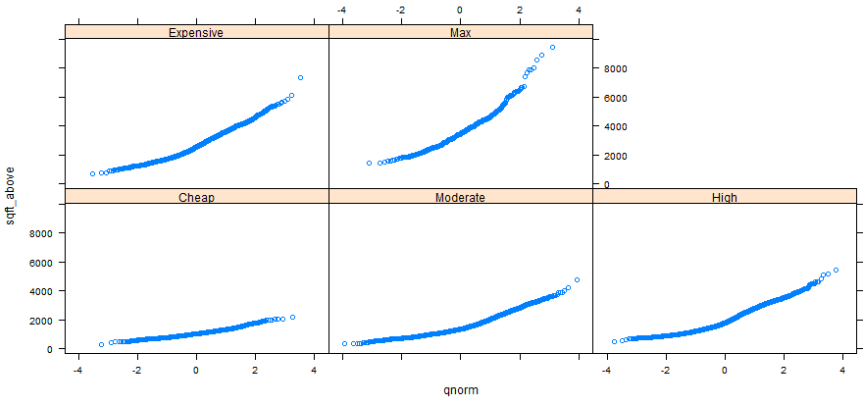
**Of the features you investigated, were there any unusual distributions? **

###Did you perform any operations on the data to tidy, adjust, or change the

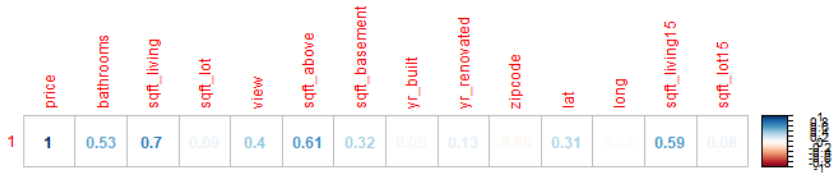
form of the data? If so, why did you do this?

I will convert waterfront,condition,floors,grade into factor variables for bivariate and multivariate analysis.

Bivariate Plots Section



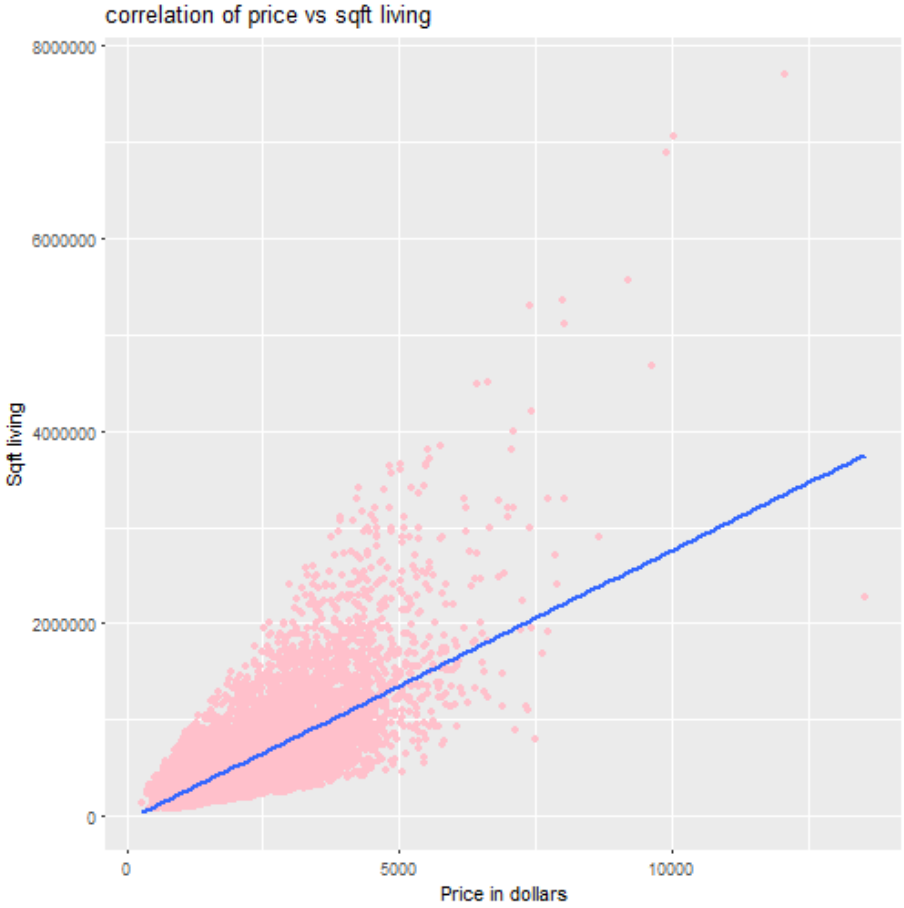
By this Q-Q plots for each price category we see the variable sqft_above has a right skewed distribution.



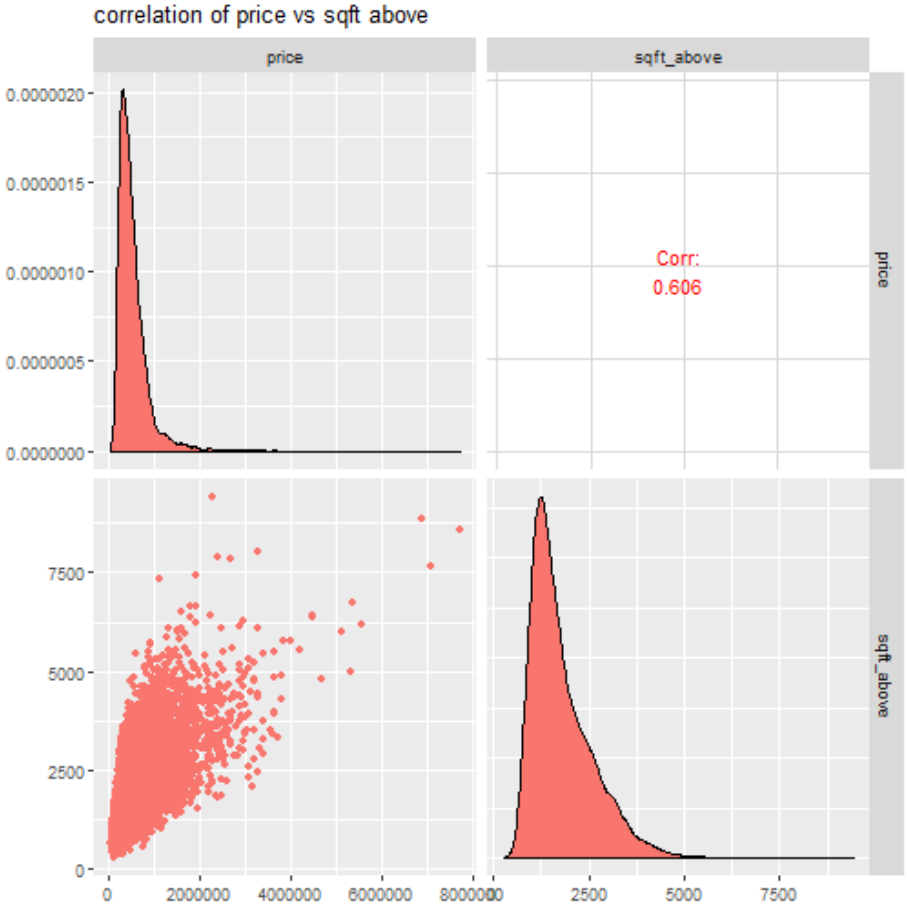
In this correlation matrix of the price with other variables, the most dominant correlations price has is with sqft_living(0.7), grade(0.67), sqft_above(0.61), sqft_living15(0.59) and bathrooms(0.53). This was different from what I had expected. I was hoping that bedrooms,zip_code,condition and sqft_lot should have a higher correlation with price. Zipcode has a negative correlation(-0.05) which is unusual. The average sqft living of 15 nearby houses (sqft_living15) is quite relevant as in neighbourhood which has higher sqft living size, most of the houses will be of similar size and also price.

So now I will focus on these variables with higher correlation.

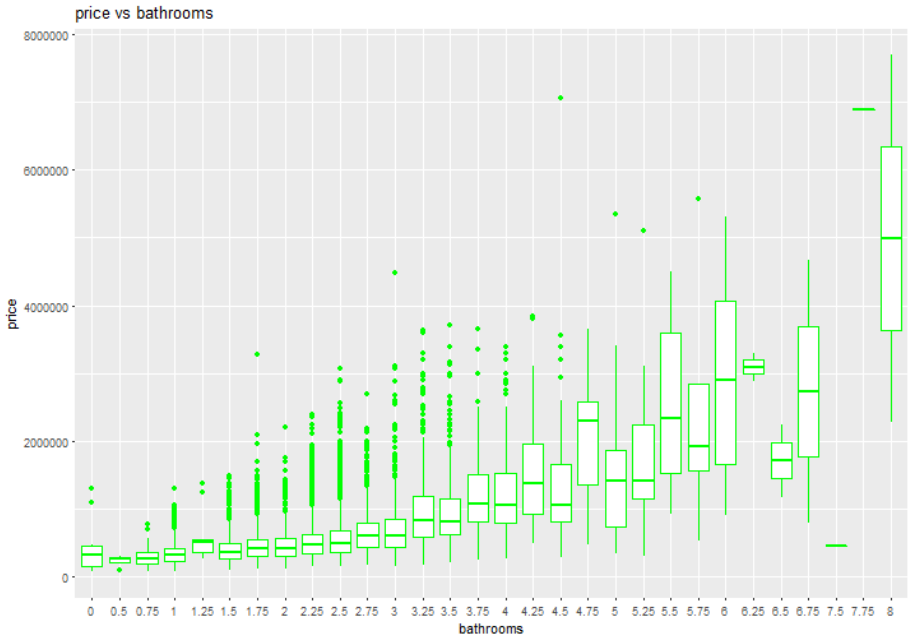
Now looking into the plots for those variables with higher correlation value.



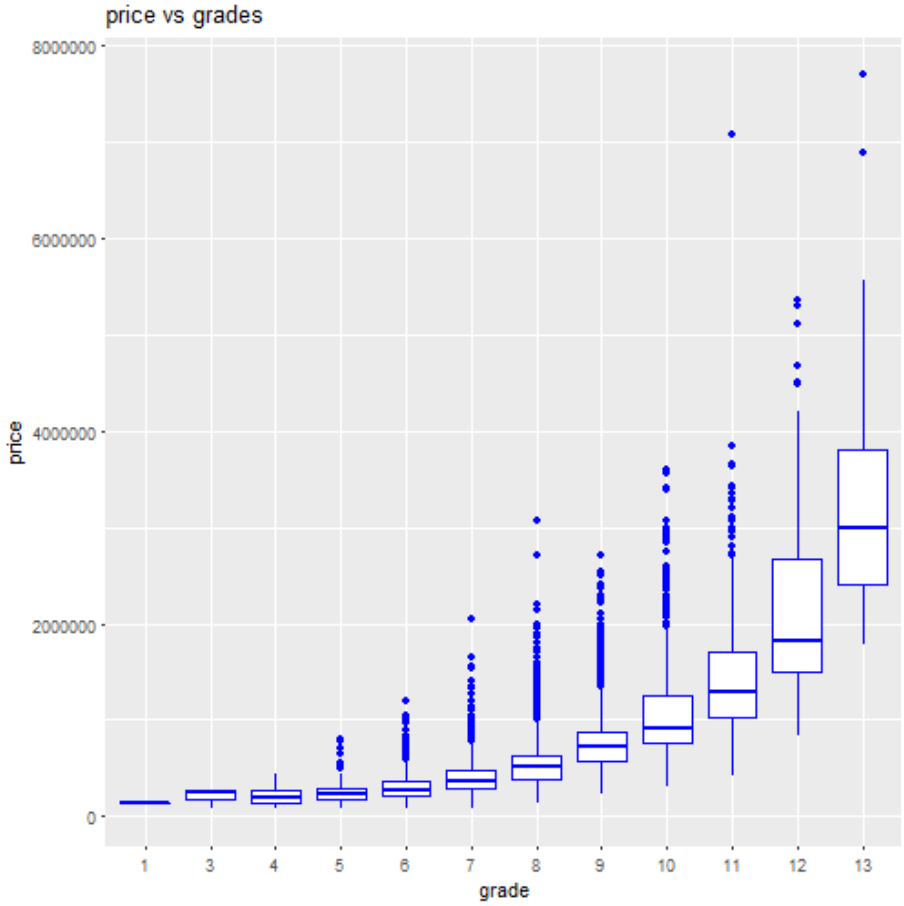
Sqft_living has the highest correlation with price which is obvious and in the price vs sqft_living we see a linear relationship between the two. The regression line shows that the relationship is linear.



This plots made using ggpairs gives a lot of information at one time. Similarly sqft_above has more or less a linear with price. The density function of sqft_above is also right skewed. This correlation is the third largest among the variables selected.



As bathrooms have a strong correlation with price we plotted this to further understand the relationship. As the umber of bathroom increase the spread increase of the boxplot indicating that the price difference is increasing. Also as the number of bathrooms increase the first quantile value price also goes up giving a positive correlation with price.

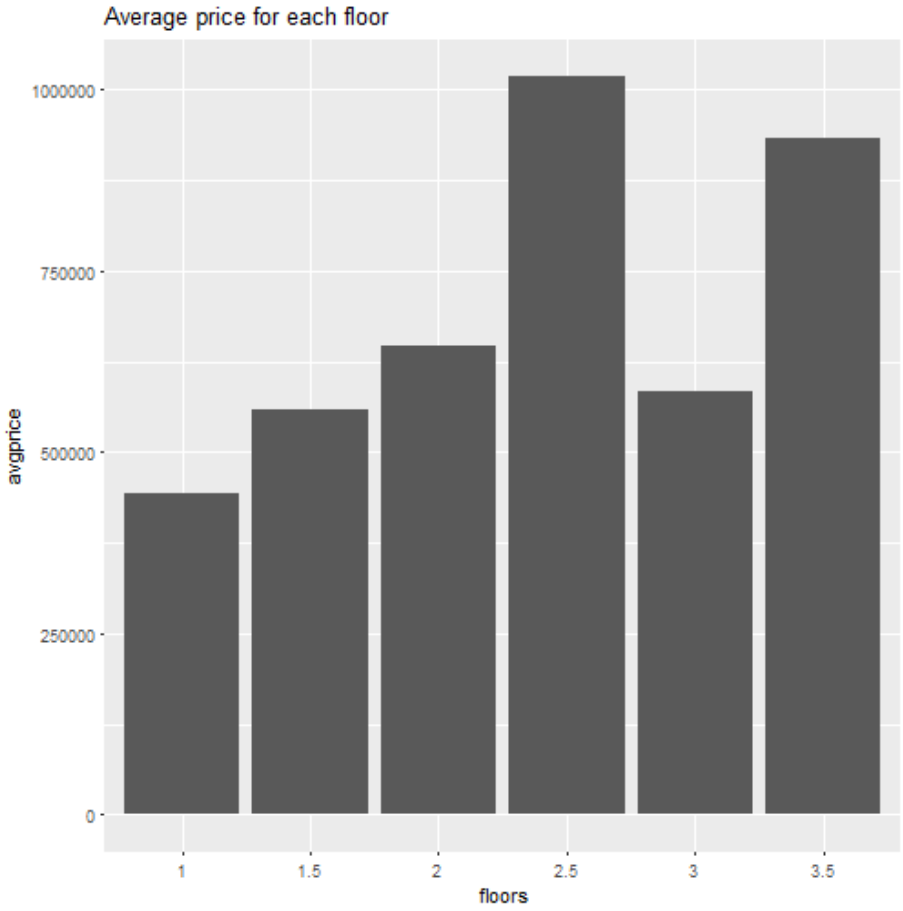


Similarly this plot shows a gradual positive correlation of price with increasing grade.Hence we can figure that houses with grade 13 is the best as it is the costliest. There are a lot of outliers in the grades ranging from 7 to 11

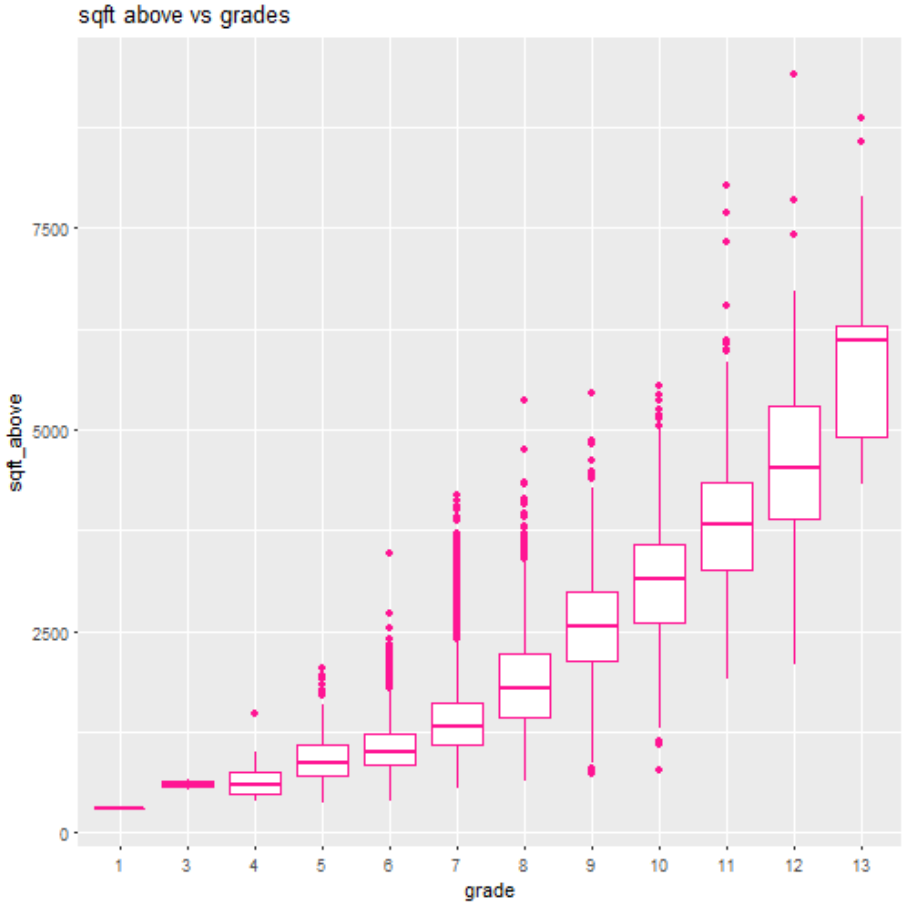
```
## # A tibble: 12 x 3
##     grade  avgPrice count
##    <fctr>     <dbl> <int>
##  1      1  142000.0     1
##  2      3  205666.7     3
##  3      4  214381.0    29
##  4      5  248524.0   242
##  5      6  301919.6  2038
##  6      7  402590.3  8981
##  7      8  542852.8  6068
##  8      9  773513.2  2615
##  9     10 1071771.1  1134
## 10     11 1482857.7   398
```

```
## 11     12 2191222.0    90
## 12     13 3058181.8    11
```

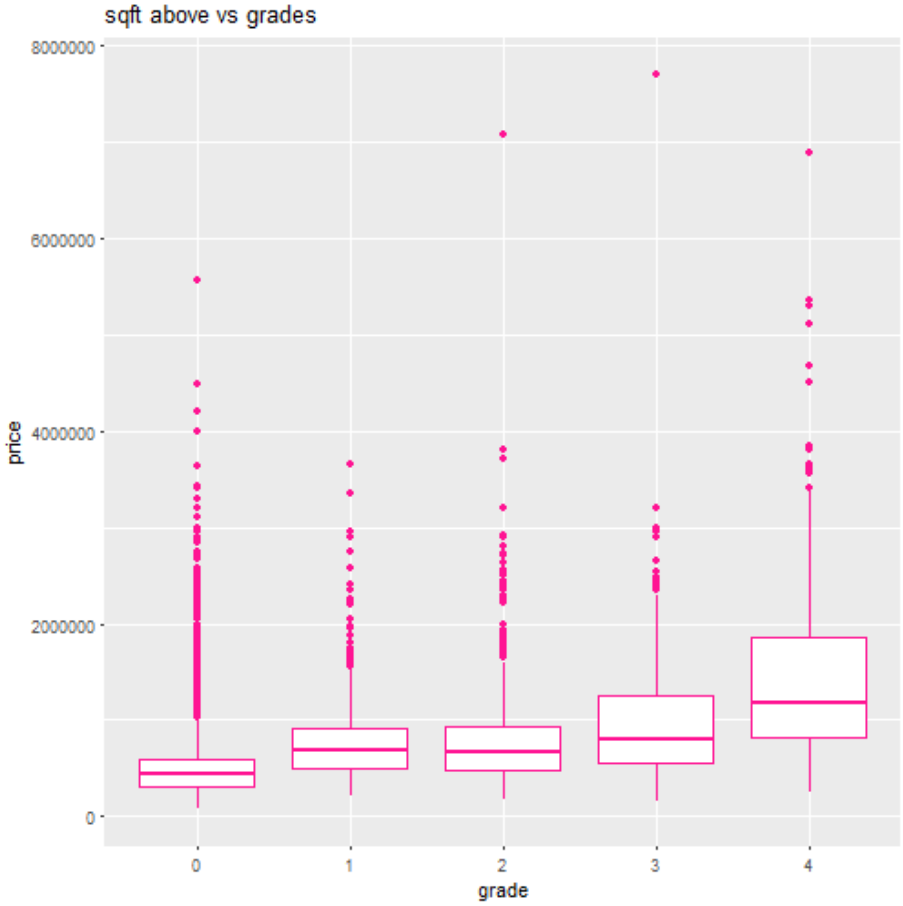
The above is also proved by this summary table.



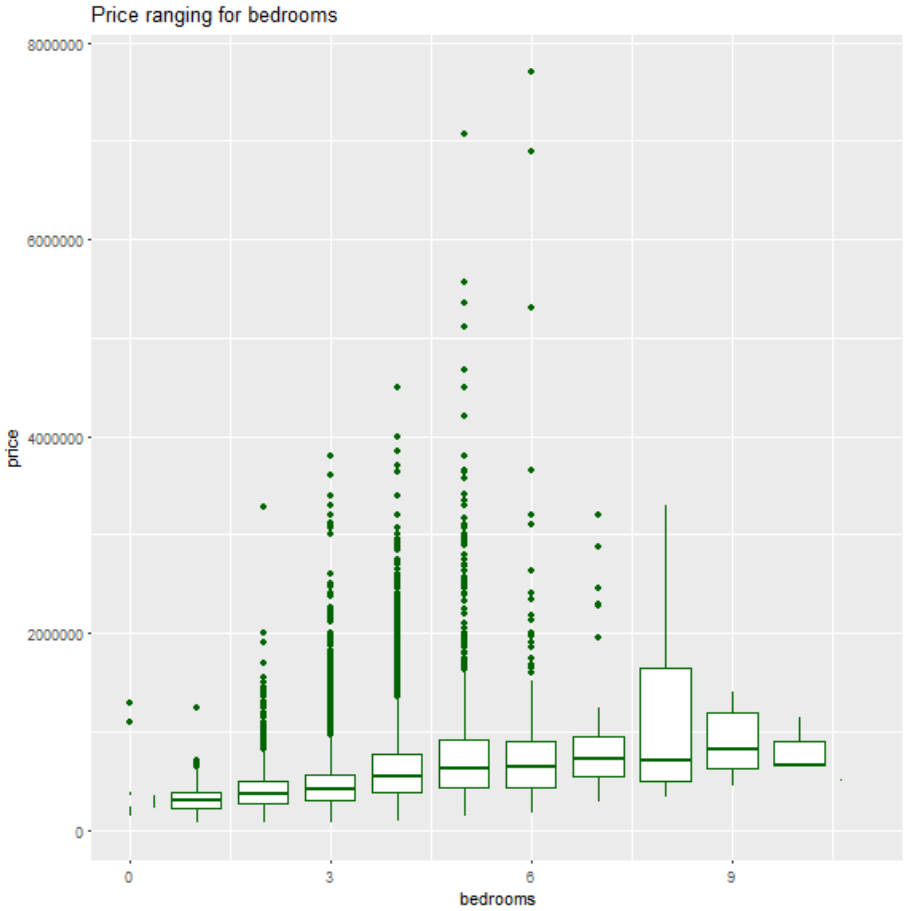
The floor that has the highest average price is 2.5 followed by 3.5.



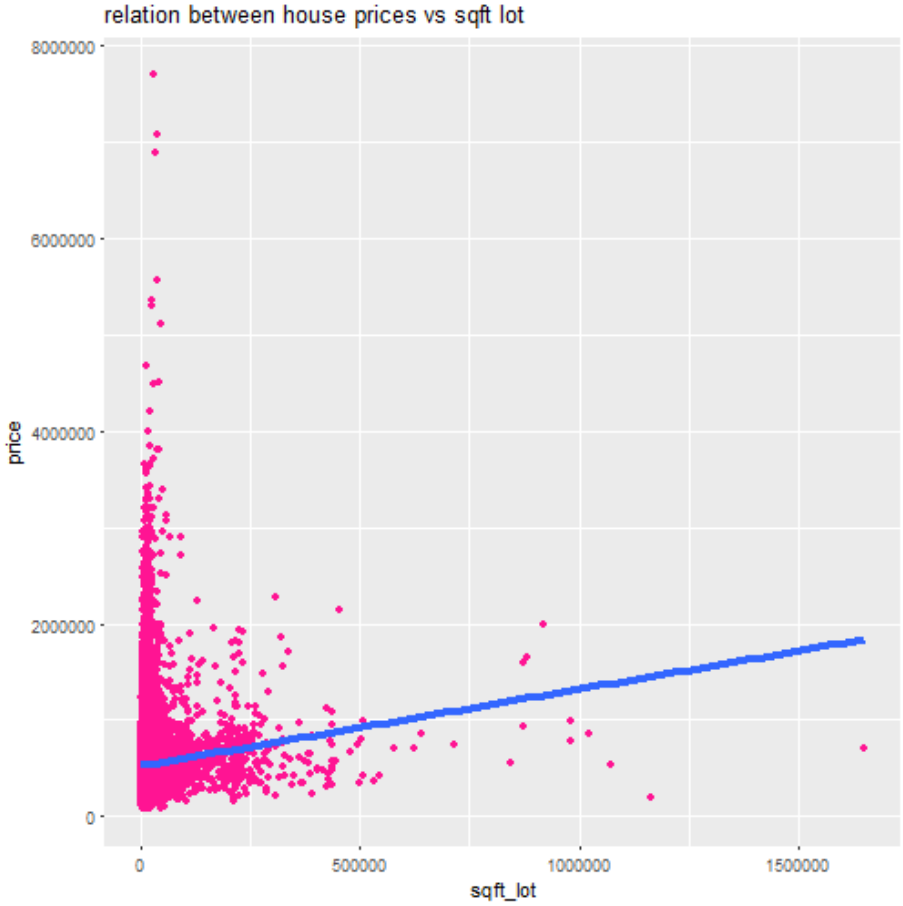
Now looking into the correlation between grades and sqft above, I see there is a positive correlation.



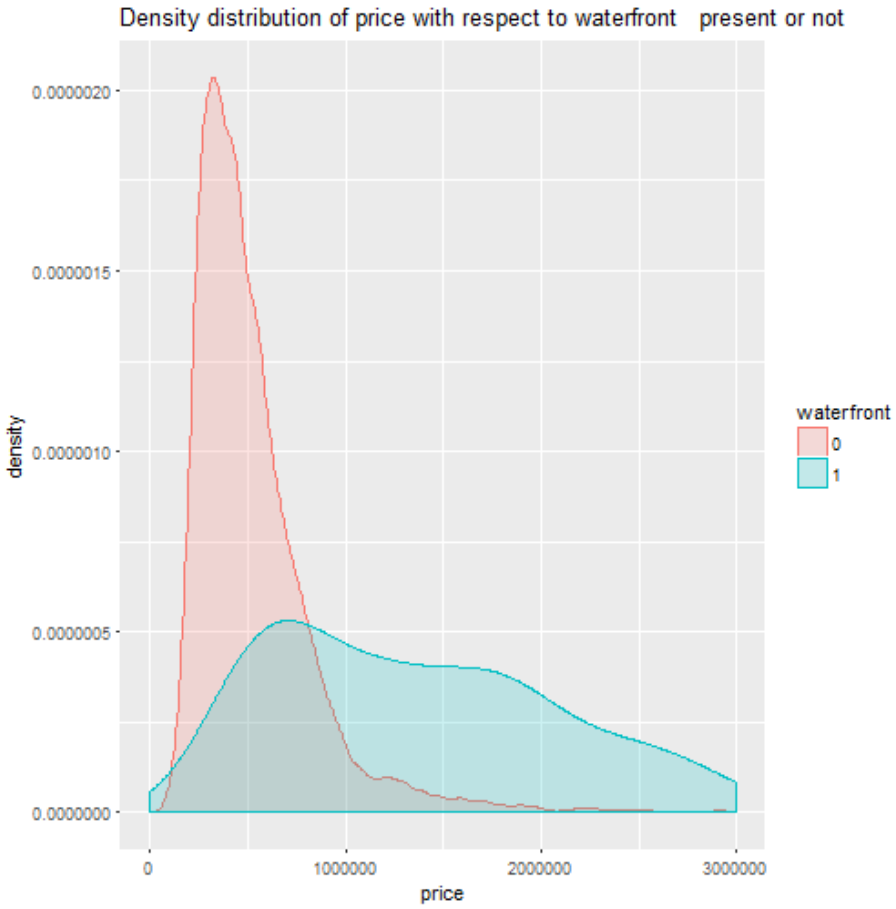
From this plot I figured that view 4 is the best among the four.



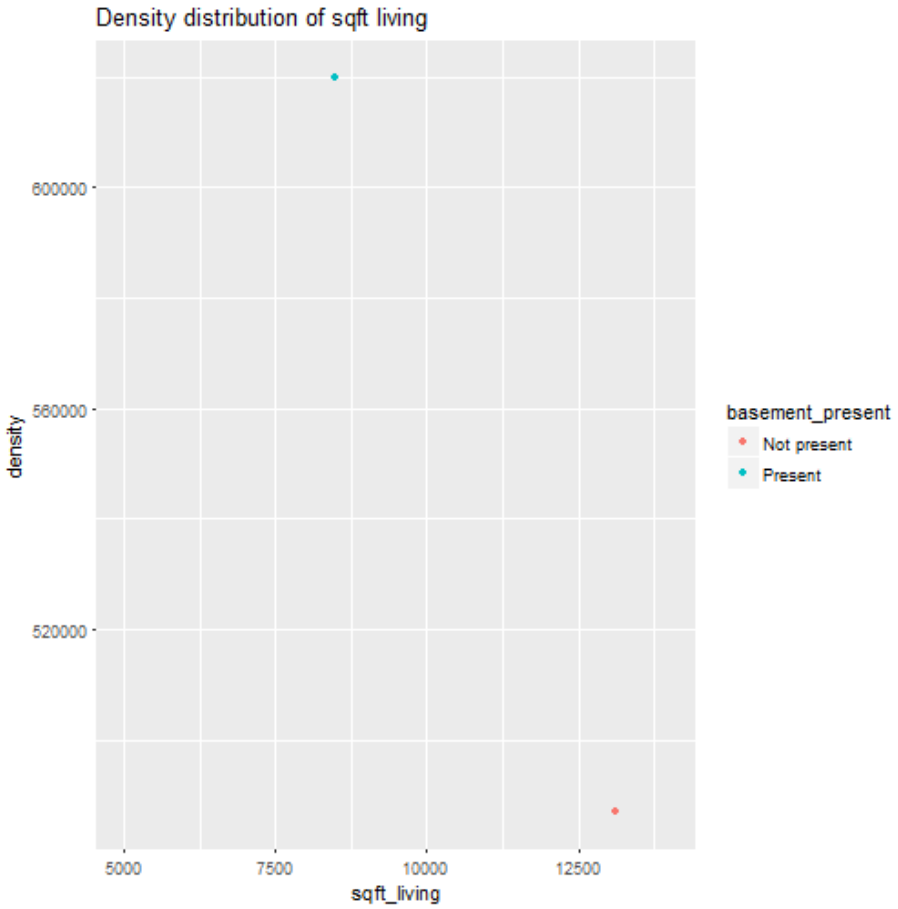
There are a lot of outliers in this boxplot for bedrooms 3,4,5,6 which indicate that price and bedrooms are not that correlated as I had assumed before.



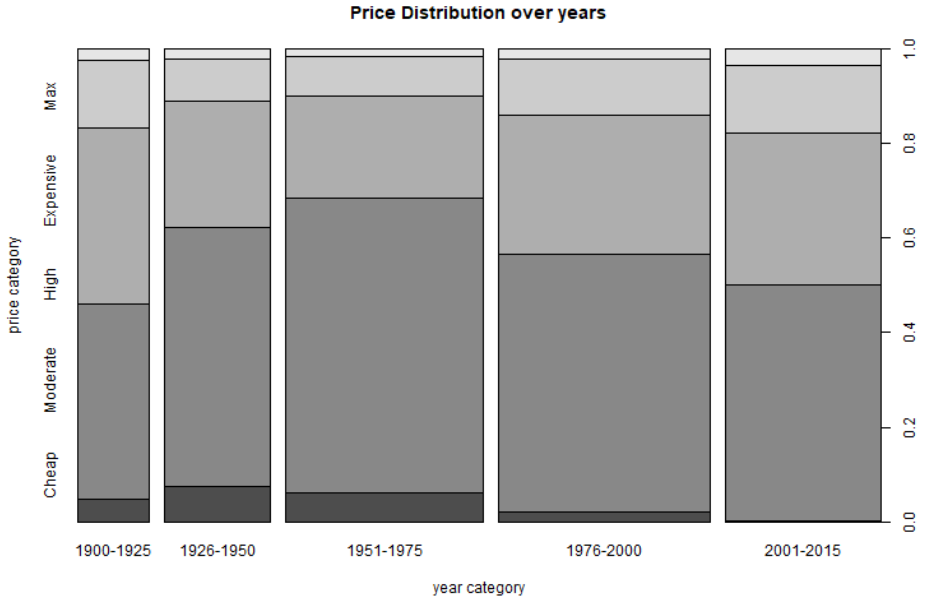
Relationship of price vs sqft lot is clearly not linear and hence sqft lot clearly does not affect price of houses in king county.



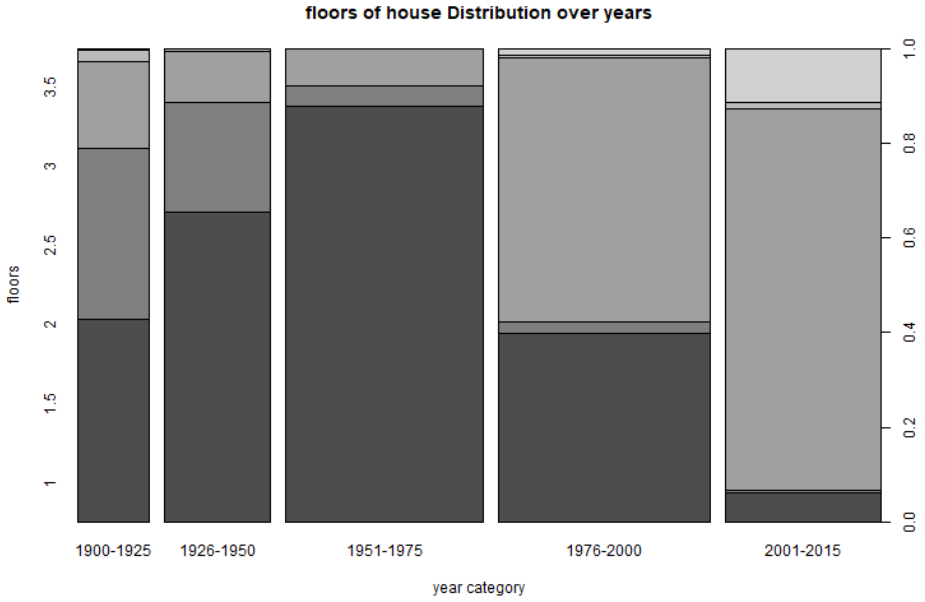
In this density plot the distribution of the price is studied with respect to waterfront being present or not. In the case where waterfront is not present the prices are lower with the spread being less compared to the normal distribution with waterfront where the price range is larger.



In the neighbourhood, the number of houses having basement is less than the number of housing not having basement. By comparing the average prices we see that it is higher for the houses with that have basement. It means most people live in houses that does not have basement as the prices is comparitively much higher.



In the time range of 1900-1925 houses were more expensive compared to 1926-2000. In 2000 there is again increase in the price range.



In the early 1900s(1900-1950) mostly houses were built with 1,1.5 floors. But,

gradually through the years in the 2000s this has changed and mostly houses are built with 2 floors and 3.5 floors is also on the high.

By plotting condition vs the year category I found that most of the houses built in he 2000s are of condition 3 compared to the previous years where other condition houses were also built. I can assume that condition 3 signifies a mid condition that can be afforable by the middle working class.

#Bivariate Analysis

###Talk about some of the relationships you observed in this part of the \ ###investigation. How did the feature(s) of interest vary with other features in \ ###the dataset?

Yes there were some relationships of interest.

Houses built in 2000s are mostly of condition 3.

Houses built in 2000s are mostly of 2 and 3.5 floors compared to 1900-1950

which were 1 and 1.5 floors

The feature of interest was price and it has a positive correlation with sqft_living,grade,sqft_above. Contrary to my belief, number of bedrooms,

sqft-lot and zipcode does not have a strong correlation with price.

**Did you observe any interesting relationships between the other features **

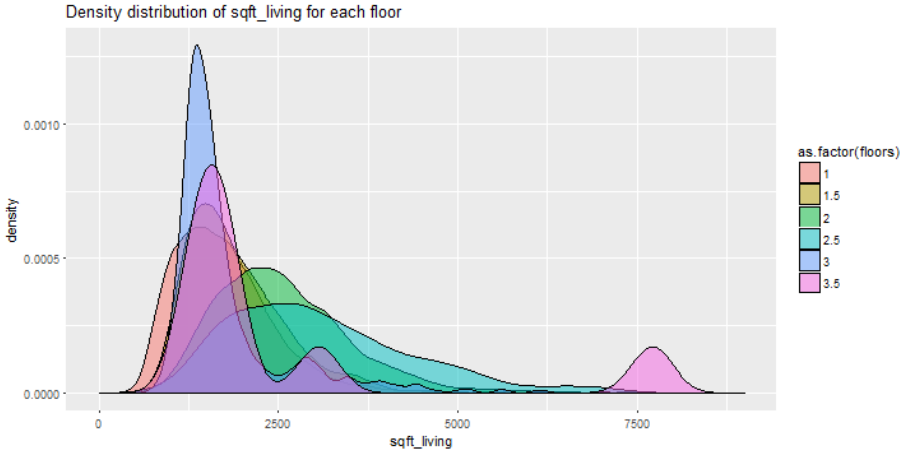
(not the main feature(s) of interest)?

In the early 1900s(1900-1950) mostly houses were built with 1,1.5 floors. But, gradually through the years in the 2000s this has changed and mostly houses are built with 2 floors and 3.5 floors is also on the high.

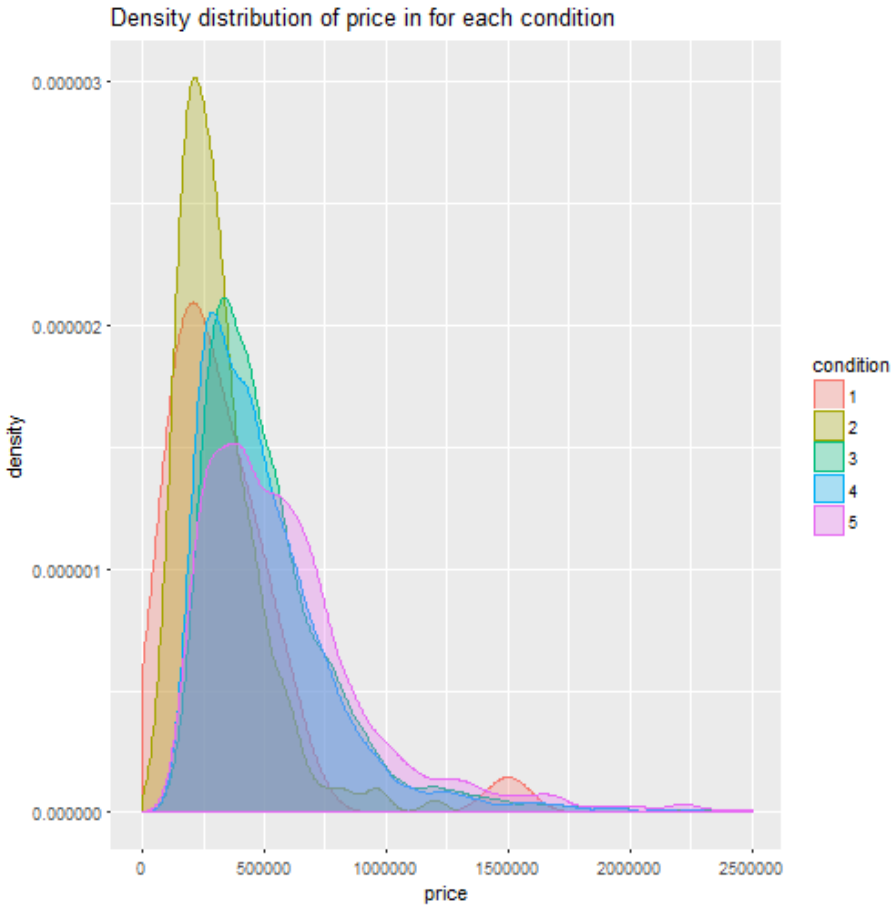
What was the strongest relationship you found?

The strongest relationship is of price vs sqft_living as it has a positive correlation of 0.7

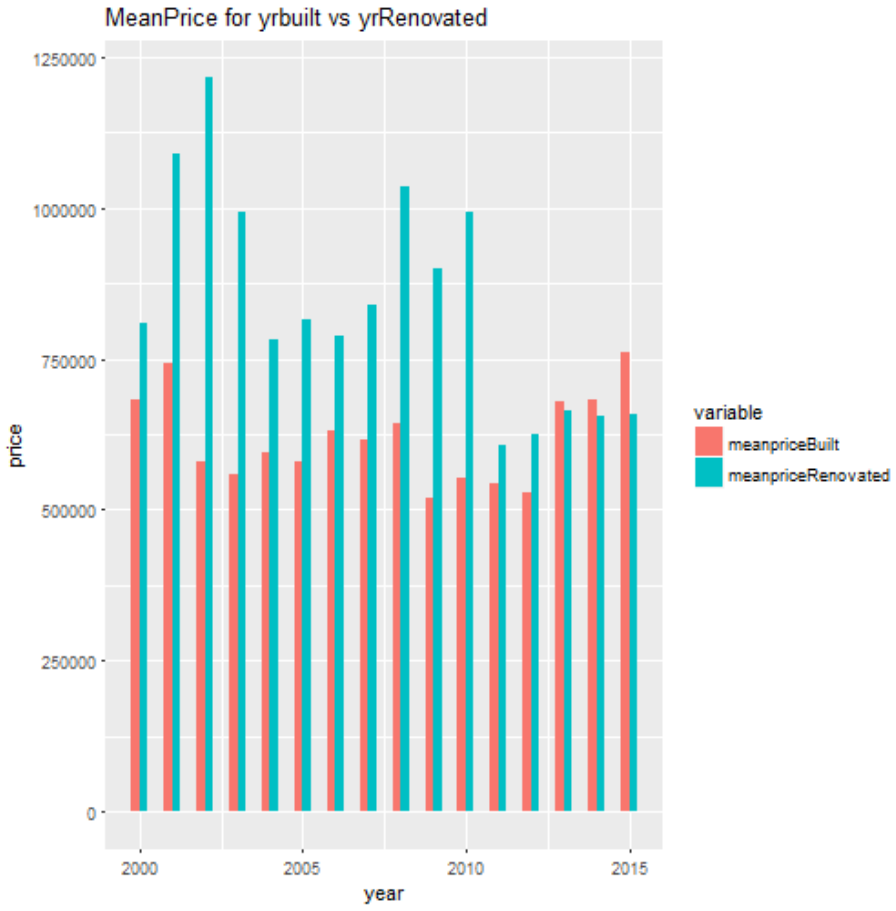
#Multivariate Plots



In this plot I tried to density distribution of sqft living for each floor and found that similar to price the height for 2.5 floor is maximum and variation of price is less as compared to the other floor factors.

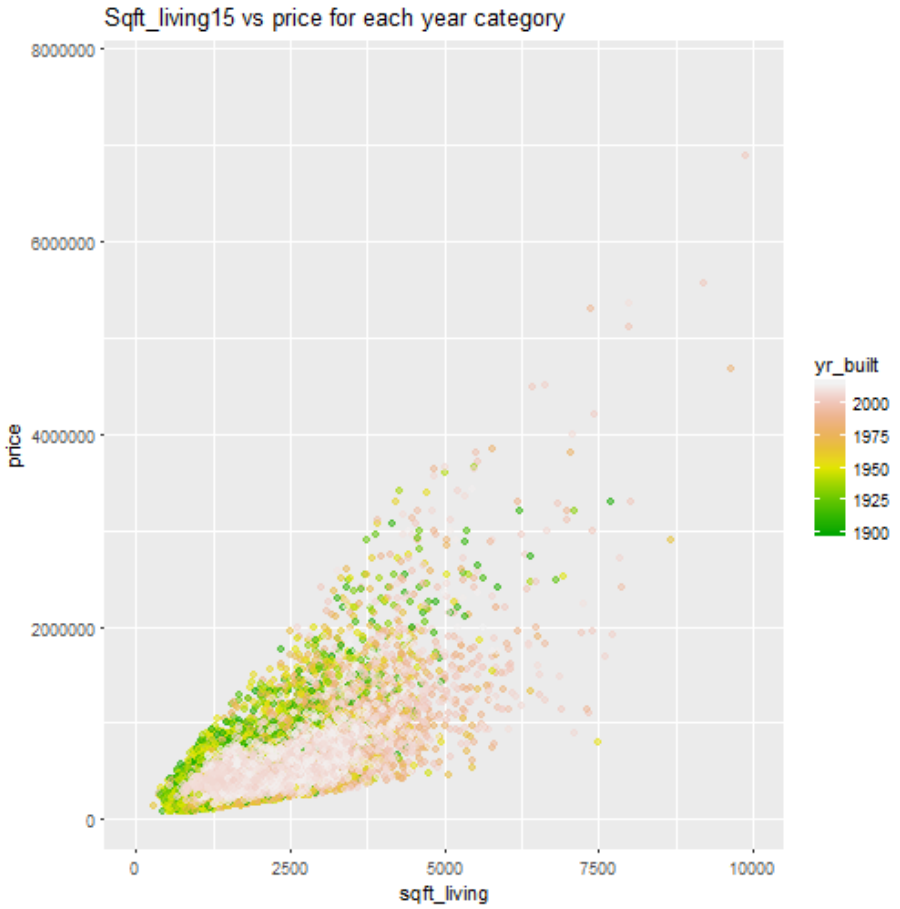


In this distribution the price is compared for each condition value. For the condition 2 the variation of price is less compared to the the other conditions.

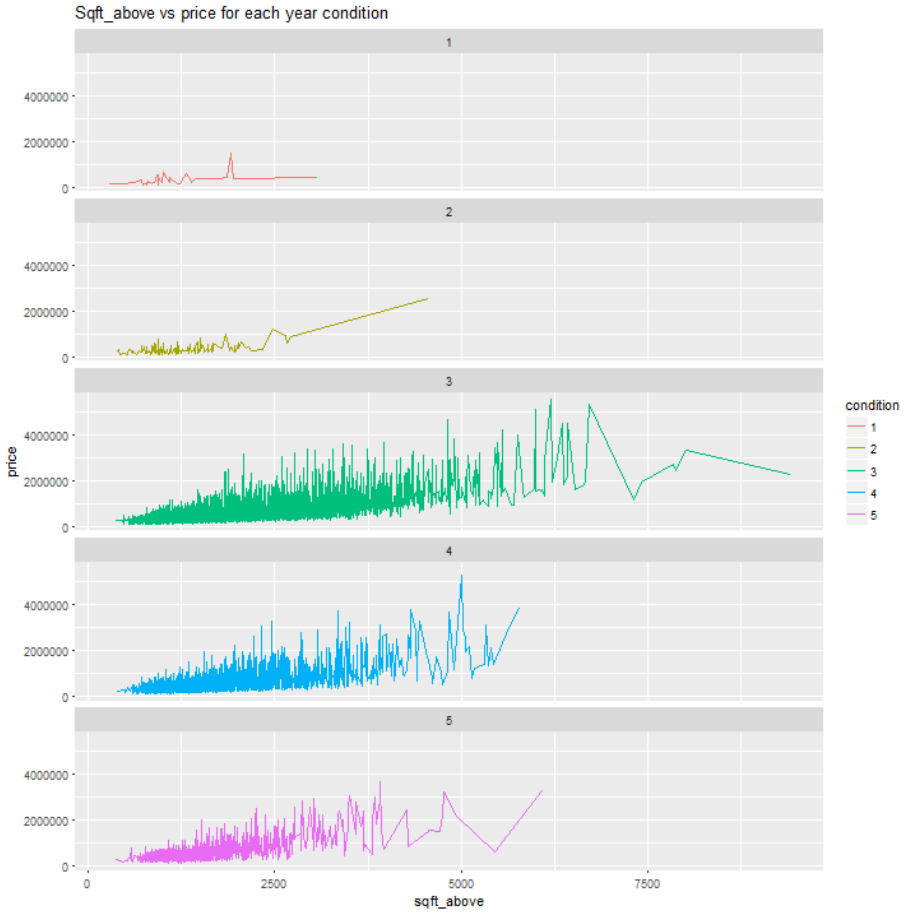


The mean prices of the houses that were built in 2000 to 2012 was way lower that those renovated in that period. But there is a gradual shift and the mean price built has increased after 2012.

There is a strong correlation between sqft_living and sqft_living15. The grades from 1 to 6 has usually smaller values for sqft_living and sqft_living15 and it gradually increases for grades 7 to 13.

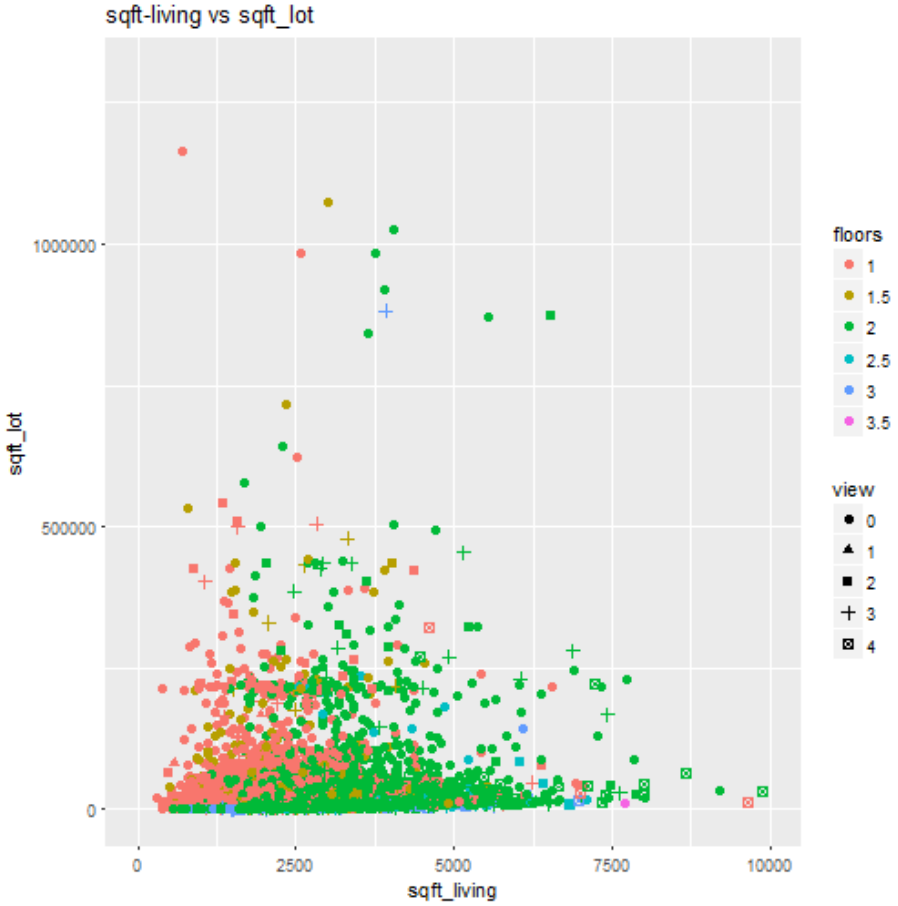


For the same sqft_living houses built in 1900 has higher price than the ones built in 2000s. This is very interesting observation. But houses made in larger sqft_living has increased in the 2000s.

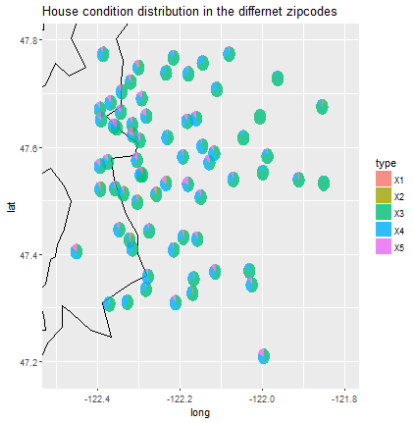


This plot helps to understand that houses with condition 1,2 has both less range of price and sqft above compared to the other 3.

This plot of bedrooms and floors has horizontal strips. The majority of houses with 3 floors have grade 8. For floors 2, the majority is 8 and 9. For floors 1 and 1.5 the most common grade is 6 and 7.



The plots shows that lot size is quite high compared to the sqft living. the majority of houses are with 1 and 2 floors and between 2400 to 7000 sqft_living. Most of the houses are of view 0.



In this plot I am trying to figure how the house condition is distributed in the different zipcodes. The most common condition is 3 and it has the majority in most zipcodes followed by 4 condition.

```
## Error in file(con, "rb"): cannot open the connection
```

This visualization predicts that the houses near the coastline are more expensive than the houses that are not.

In this plot I wanted o see how the houses were built over the years. In the Seattle region the houses were built near the coastline first. Then as time progressed people started moving towards the mainland. In the right side of the map we see most houses built after 1950s indicating newer settlemets than the left side. Also I have addded waterfront as size so we can identify houses with waterfront.

**Talk about some of the relationships you observed in this part of the **

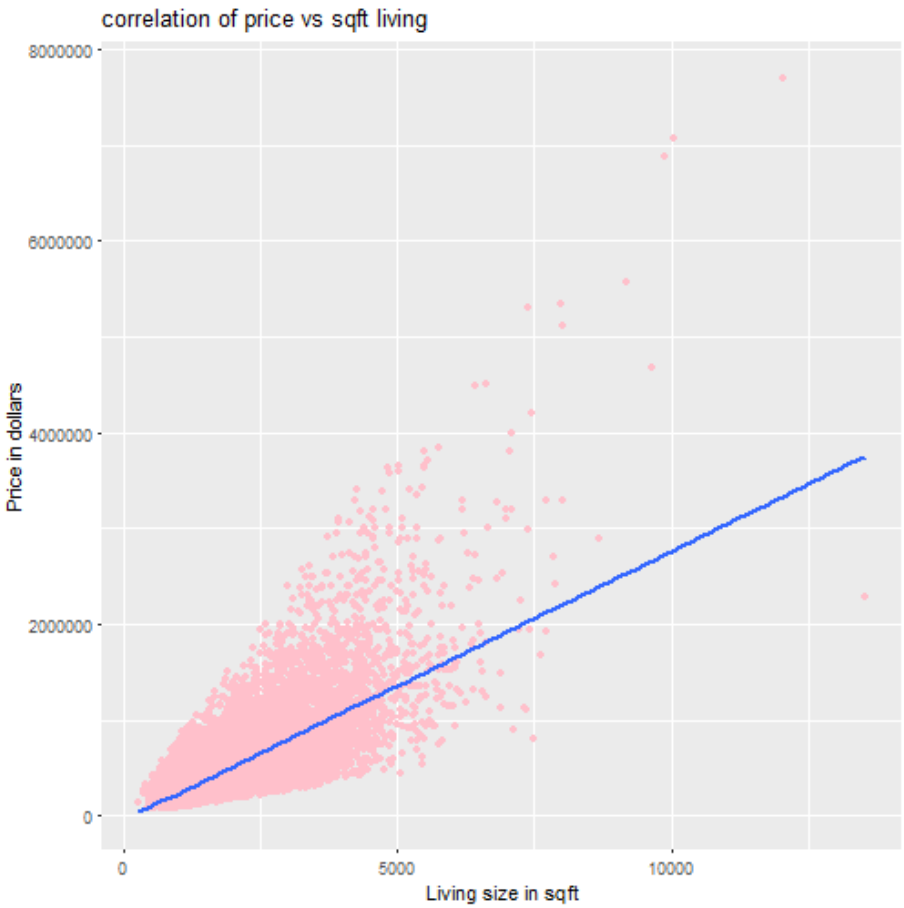
**investigation. Were there features that strengthened each other in terms of **

looking at your feature(s) of interest?

The costlier houses are built along the coastline and they are built in the 1900. The prices have remained high for those houses.

The mean prices of the houses that were built in 2000 to 2012 was way lower that those renovated in that period. But there is a gradual shift and the mean price built has increased after 2012.

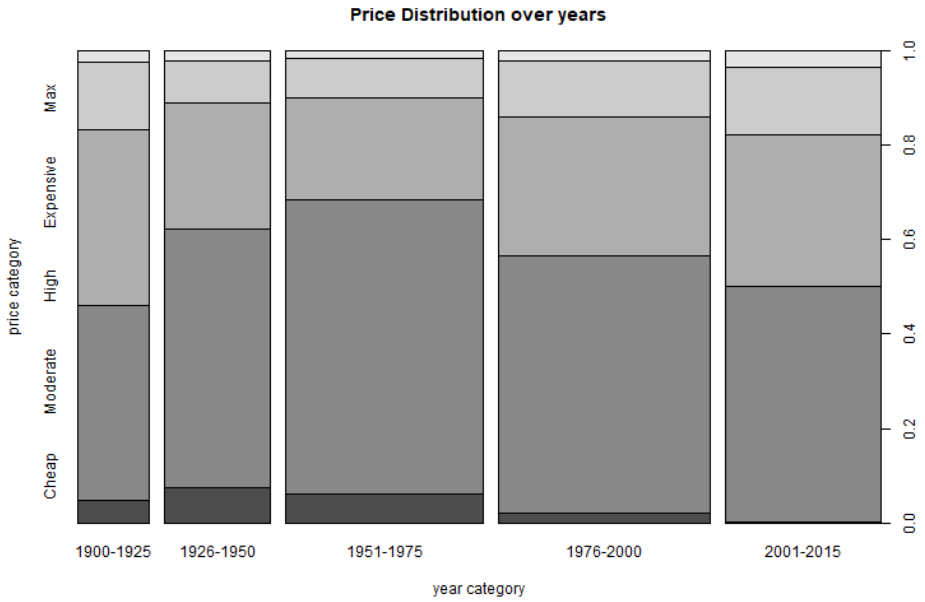
Plot One



Description One

I choose this as price has the as this gives us the strongest correlation with sqft_living(0.71).

Plot Two



Description Two

The construction of houses in the cheap category has gradually reduced and nowerdays cheap category is not built at all. It is edging towards the expensive side now in the 2000s. The moderate price category was maximun in the 1951-1975 range.

Plot Three

```
## Error in file(con, "rb"): cannot open the connection
```

Description Three

The houses near the coastline is the cosliest and also built earlier in time. This indicates that people first inhabited the coastline and then in the later years more towards the interior in King County.

Reflection

First for doing this project comming up with a tidy dataset with lots of features was a challenge. I did a lot of research and found this tidy dataset on Kaggle which suited my requirements. The King county house pricing dataset has 21613 observations and 21 features. I wanted to explore all the 21 features and started by plotting the individual variables in the data set. Then I looked for any interesting relationships present in the dataset mong the features. I had expected that price was positively correlated with a lot of variables. But I was wrong as value for number of bedrooms,zipcode, lot size was lot. The most important correlation is sqft of living. I found out the that condition 3 is the most prevelant among all houses. The houses were first built along the coastline and then people moved towards the mainland. I loved working with this housing prices data. It would have been great if the dataset contained day and month for the built year. I could have drawn a lot of insights and also use time series forcasting to predict for the next years to come. I would continue with price prediction later on for this dataset. I plan to do both of these things if possible with this dataset.