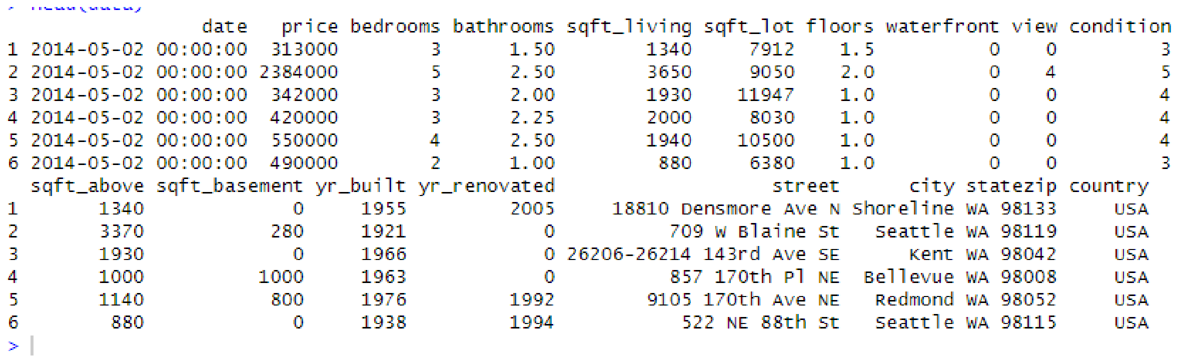
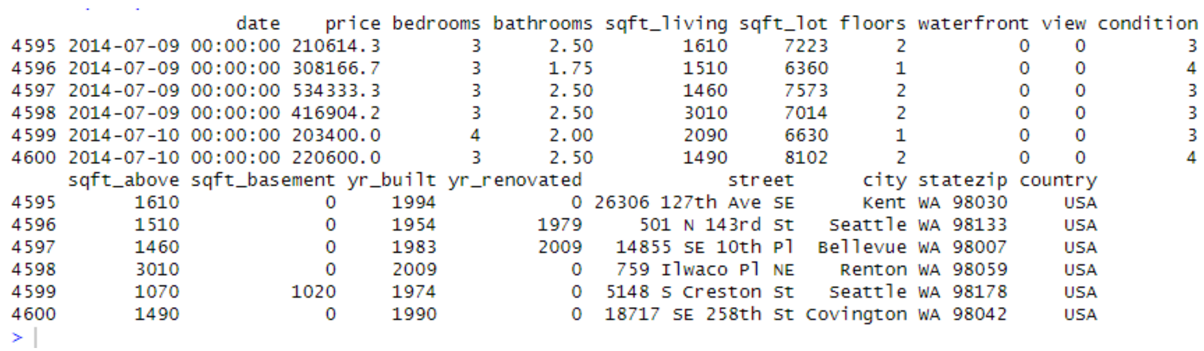
House price prediction using R programming

```
library(ggplot2)

data <- read.csv(file = 'D:/data.csv')

head(data)
```



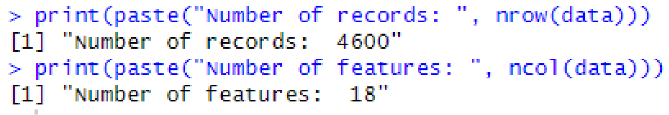
```
                 date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
1 2014-05-02 00:00:00  313000        3      1.50        1340     7912    1.5          0    0         3
2 2014-05-02 00:00:00 2384000        5      2.50        3650     9050    2.0          0    4         5
3 2014-05-02 00:00:00  342000        3      2.00        1930    11947    1.0          0    0         4
4 2014-05-02 00:00:00  420000        3      2.25        2000     8030    1.0          0    0         4
5 2014-05-02 00:00:00  550000        4      2.50        1940    10500    1.0          0    0         4
6 2014-05-02 00:00:00  490000        2      1.00         880     6380    1.0          0    0         3
  sqft_above sqft_basement yr_built yr_renovated                    street      city statezip country
1       1340             0     1955         2005      18810 Densmore Ave N Shoreline WA 98133     USA
2       3370           280     1921            0           709 W Blaine St   Seattle WA 98119     USA
3       1930             0     1966            0 26206-26214 143rd Ave SE      Kent WA 98042     USA
4       1000          1000     1963            0           857 170th Pl NE  Bellevue WA 98008     USA
5       1140           800     1976         1992         9105 170th Ave NE   Redmond WA 98052     USA
6        880             0     1938         1994            522 NE 88th St   Seattle WA 98115     USA
> |
```

```
tail(data)
```



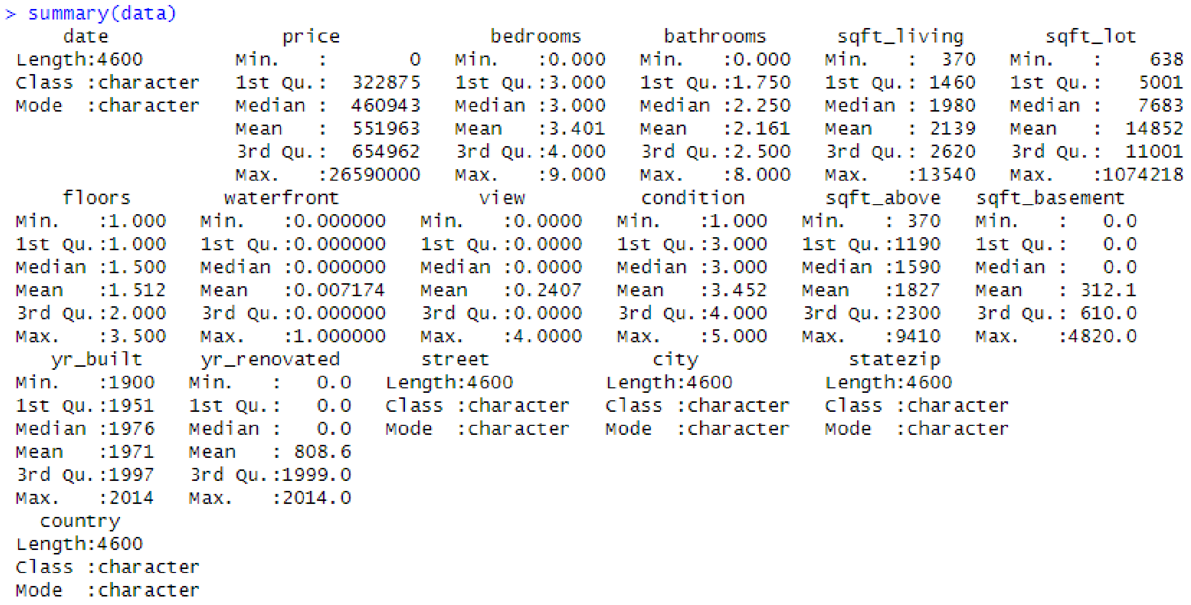
```
                    date    price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition
4595 2014-07-09 00:00:00 210614.3        3      2.50        1610     7223      2          0    0         3
4596 2014-07-09 00:00:00 308166.7        3      1.75        1510     6360      1          0    0         4
4597 2014-07-09 00:00:00 534333.3        3      2.50        1460     7573      2          0    0         3
4598 2014-07-09 00:00:00 416904.2        3      2.50        3010     7014      2          0    0         3
4599 2014-07-10 00:00:00 203400.0        4      2.00        2090     6630      1          0    0         3
4600 2014-07-10 00:00:00 220600.0        3      2.50        1490     8102      2          0    0         4
     sqft_above sqft_basement yr_built yr_renovated               street      city statezip country
4595       1610             0     1994            0 26306 127th Ave SE      Kent WA 98030     USA
4596       1510             0     1954         1979      501 N 143rd St   Seattle WA 98133     USA
4597       1460             0     1983         2009    14855 SE 10th Pl  Bellevue WA 98007     USA
4598       3010             0     2009            0    759 Ilwaco Pl NE    Renton WA 98059     USA
4599       1070          1020     1974            0    5148 S Creston St   Seattle WA 98178     USA
4600       1490             0     1990            0   18717 SE 258th St Covington WA 98042     USA
> |
```

```
print(paste("Number of records: ", nrow(data)))

print(paste("Number of features: ", ncol(data)))
```



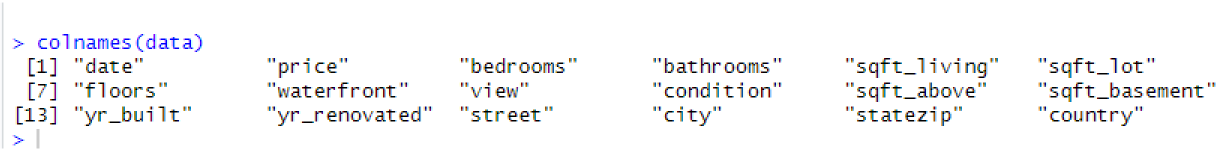
```
> print(paste("Number of records: ", nrow(data)))
[1] "Number of records:  4600"
> print(paste("Number of features: ", ncol(data)))
[1] "Number of features:  18"
```

summary(data)



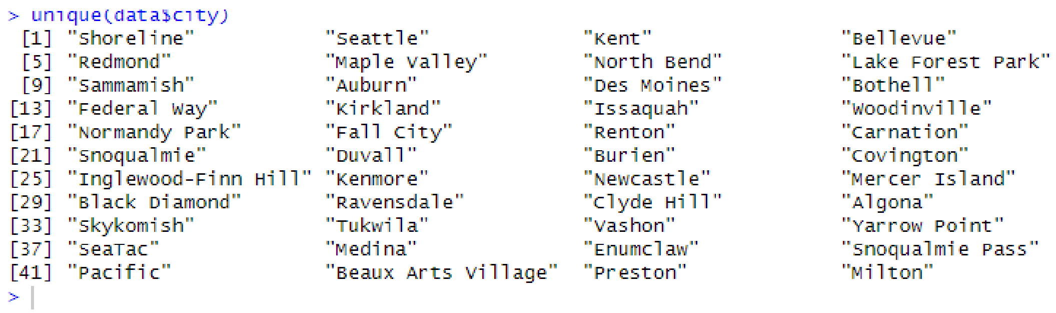
```
> summary(data)
     date               price             bedrooms       bathrooms      sqft_living       sqft_lot      
 Length:4600        Min.   :       0   Min.   :0.000   Min.   :0.000   Min.   :  370   Min.   :    638  
 Class :character   1st Qu.:  322875   1st Qu.:3.000   1st Qu.:1.750   1st Qu.: 1460   1st Qu.:   5001  
 Mode  :character   Median :  460943   Median :3.000   Median :2.250   Median : 1980   Median :   7683  
                    Mean   :  551963   Mean   :3.401   Mean   :2.161   Mean   : 2139   Mean   :  14852  
                    3rd Qu.:  654962   3rd Qu.:4.000   3rd Qu.:2.500   3rd Qu.: 2620   3rd Qu.:  11001  
                    Max.   :26590000   Max.   :9.000   Max.   :8.000   Max.   :13540   Max.   :1074218  
     floors        waterfront            view          condition       sqft_above   sqft_basement   
 Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 370   Min.   :   0.0  
 1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:1190   1st Qu.:   0.0  
 Median :1.500   Median :0.000000   Median :0.0000   Median :3.000   Median :1590   Median :   0.0  
 Mean   :1.512   Mean   :0.007174   Mean   :0.2407   Mean   :3.452   Mean   :1827   Mean   : 312.1  
 3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:2300   3rd Qu.: 610.0  
 Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :9410   Max.   :4820.0  
    yr_built     yr_renovated       street              city             statezip        
 Min.   :1900   Min.   :   0.0   Length:4600        Length:4600        Length:4600       
 1st Qu.:1951   1st Qu.:   0.0   Class :character   Class :character   Class :character  
 Median :1976   Median :   0.0   Mode  :character   Mode  :character   Mode  :character  
 Mean   :1971   Mean   : 808.6                                                           
 3rd Qu.:1997   3rd Qu.:1999.0                                                           
 Max.   :2014   Max.   :2014.0                                                           
   country         
 Length:4600       
 Class :character  
 Mode  :character  
```

colnames(data)



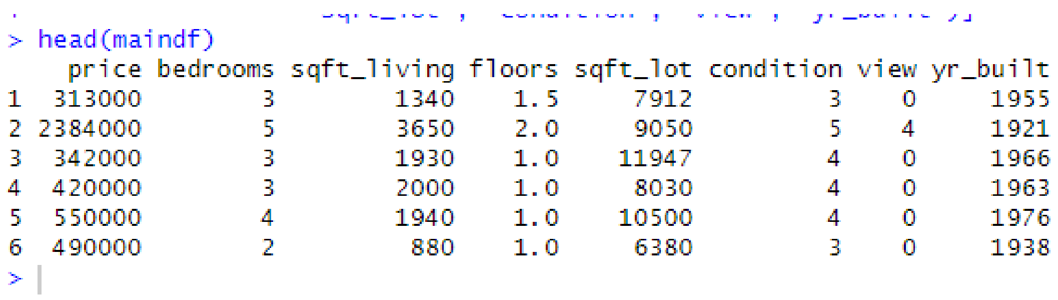
```
> colnames(data)
 [1] "date"          "price"         "bedrooms"      "bathrooms"     "sqft_living"   "sqft_lot"     
 [7] "floors"        "waterfront"    "view"          "condition"     "sqft_above"    "sqft_basement"
[13] "yr_built"      "yr_renovated"  "street"        "city"          "statezip"      "country"      
> 
```

unique(data$city)



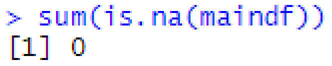
```
> unique(data$city)
 [1] "Shoreline"           "Seattle"             "Kent"                "Bellevue"           
 [5] "Redmond"             "Maple Valley"        "North Bend"          "Lake Forest Park"   
 [9] "Sammamish"           "Auburn"              "Des Moines"          "Bothell"            
[13] "Federal Way"         "Kirkland"            "Issaquah"            "Woodinville"        
[17] "Normandy Park"       "Fall City"           "Renton"              "Carnation"          
[21] "Snoqualmie"          "Duvall"              "Burien"              "Covington"          
[25] "Inglewood-Finn Hill" "Kenmore"             "Newcastle"           "Mercer Island"      
[29] "Black Diamond"       "Ravensdale"          "Clyde Hill"          "Algona"             
[33] "Skykomish"           "Tukwila"             "Vashon"              "Yarrow Point"       
[37] "SeaTac"              "Medina"              "Enumclaw"            "Snoqualmie Pass"    
[41] "Pacific"             "Beaux Arts Village"  "Preston"             "Milton"             
> 
```

```
maindf <- data[,c("price","bedrooms","sqft_living","floors",
                "sqft_lot", "condition", "view", "yr_built")]
head(maindf)
```



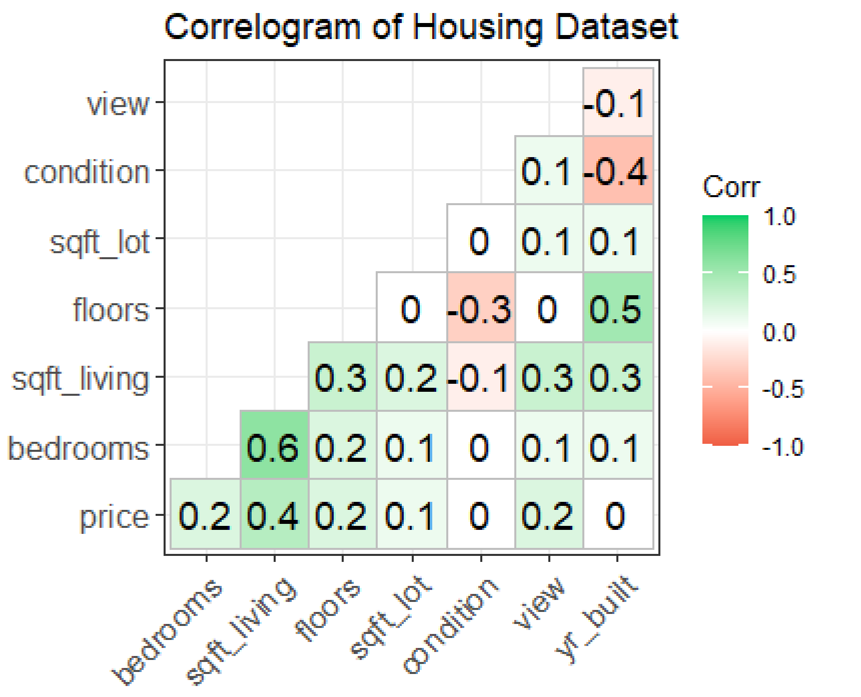
```
+                   "sqft_lot", "condition", "view", "yr_built")]
> head(maindf)
    price bedrooms sqft_living floors sqft_lot condition view yr_built
1  313000        3        1340    1.5     7912         3    0     1955
2 2384000        5        3650    2.0     9050         5    4     1921
3  342000        3        1930    1.0    11947         4    0     1966
4  420000        3        2000    1.0     8030         4    0     1963
5  550000        4        1940    1.0    10500         4    0     1976
6  490000        2         880    1.0     6380         3    0     1938
> 
```

```
sum(is.na(maindf))
```

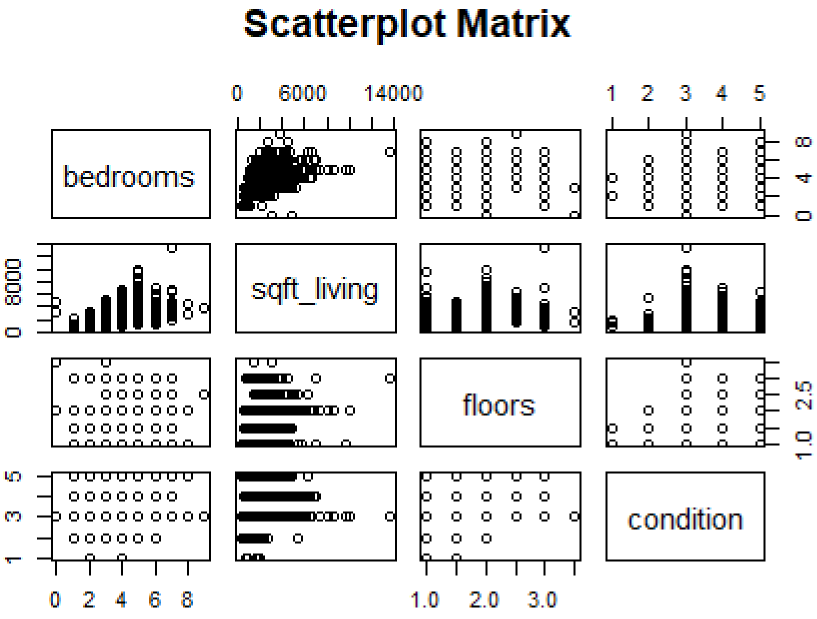


```
> sum(is.na(maindf))
[1] 0
```

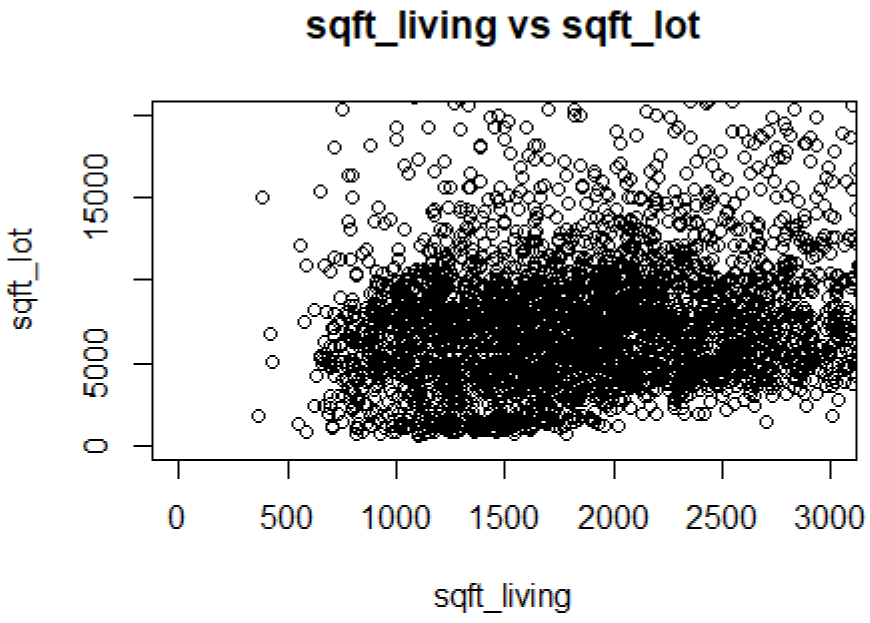
```
install.packages("ggcorrplot")
install.packages("Rcpp")
install.packages("stringi")


library(ggcorrplot)
corr <- round(cor(maindf), 1)


# Plot
ggcorrplot(corr,
         type = "lower",
         lab = TRUE,
         lab_size = 5,
         colors = c("tomato2", "white", "springgreen3"),
         title="Correlogram of Housing Dataset",
         ggtheme=theme_bw)
```



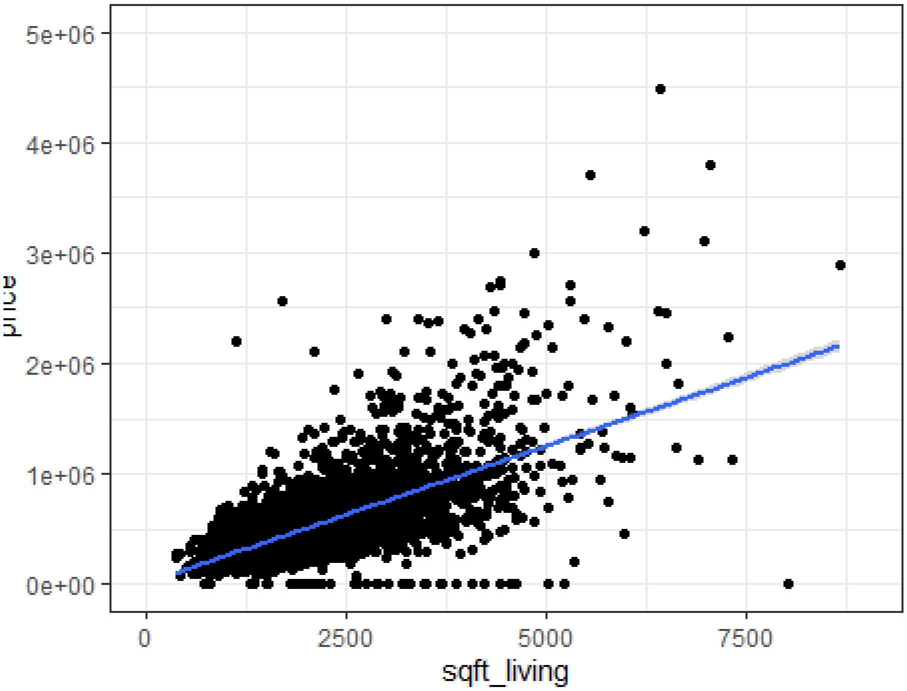
```
pairs(~bedrooms + sqft_living + floors + condition, data = maindf,
    main = "Scatterplot Matrix")
```



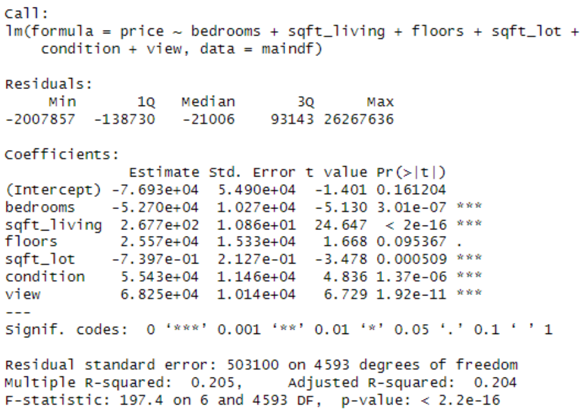
```
# Scatterplot

theme_set(theme_bw())

g <- ggplot(maindf, aes(bedrooms, floors))

g + geom_count(col="tomato3", show.legend=F) +

  labs(y="floors",

       x="bedrooms",

       title="Bedrooms vs Floors")


plot(x = maindf$sqft_living, y = maindf$sqft_lot,

     xlab = "sqft_living",

     ylab = "sqft_lot",

     xlim = c(0, 3000),

     ylim = c(0, 20000),

     main = "sqft_living vs sqft_lot"

)
```



```
ggplot(maindf,aes(y=price,x=sqft_living)) +
  geom_point() +
  xlim(0, 9000) +
  ylim(0, 5000000) +
  geom_smooth(formula = y ~ x,method="lm")
```



```
linearmodel = lm(price~bedrooms + sqft_living + floors + sqft_lot + condition + view + oldbuilt,
              data = maindf)
summary(linearmodel)
```



```
Call:
lm(formula = price ~ bedrooms + sqft_living + floors + sqft_lot +
    condition + view, data = maindf)

Residuals:
     Min       1Q   Median       3Q      Max
-2007857  -138730   -21006    93143 26267636

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.693e+04  5.490e+04  -1.401 0.161204
bedrooms    -5.270e+04  1.027e+04  -5.130 3.01e-07 ***
sqft_living  2.677e+02  1.086e+01  24.647  < 2e-16 ***
floors       2.557e+04  1.533e+04   1.668 0.095367 .
sqft_lot    -7.397e-01  2.127e-01  -3.478 0.000509 ***
condition    5.543e+04  1.146e+04   4.836 1.37e-06 ***
view         6.825e+04  1.014e+04   6.729 1.92e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 503100 on 4593 degrees of freedom
Multiple R-squared:  0.205,     Adjusted R-squared:  0.204
F-statistic: 197.4 on 6 and 4593 DF,  p-value: < 2.2e-16
```

Graph Description

The graphs generated by the provided code offer a variety of insights into the housing dataset. The correlation plot, or correlogram, visualizes the relationships between key features like `price`, `bedrooms`, `sqft_living`, `floors`, `sqft_lot`, `condition`, `view`, and `yr_built`, using color gradients to indicate the strength and direction of correlations. This plot helps quickly identify strong relationships, such as the positive correlation between `sqft_living` and `price`. The scatterplot matrix displays pairwise relationships between `bedrooms`, `sqft_living`, `floors`, and `condition`, helping to identify trends, correlations, and potential outliers across these variables. The bubble chart, which plots `bedrooms` against `floors`, uses bubble size to represent the frequency of each combination, offering insight into the most common configurations of bedrooms and floors in the dataset. The scatterplot of `sqft_living` versus `sqft_lot` explores the relationship between the living area and the lot size, with axis limits customized to highlight data within a certain range. Lastly, the scatterplot of `price` versus `sqft_living` is complemented with a linear regression line to demonstrate the general upward trend of house prices as the size of the living area increases. Together, these graphs provide a comprehensive view of the relationships and trends within the data, guiding further analysis or model development.

Conclusion

In this analysis, we uncovered some interesting patterns in the housing data. For example, we saw that the larger the living space (`sqft_living`), the higher the price, which makes sense—bigger homes usually cost more. We also explored how features like `bedrooms` and `floors` are connected, showing us common home configurations. The regression model reinforced that living area is a big factor in determining house prices. Overall, these findings give us a clearer picture of the housing market, setting us up for more in-depth predictions and insights moving forward.