EDA on house prices

Sanjana Suresh

18/Sep/2023

Sanjana Suresh

Dataset: Exploratory Data Analysis on factors affecting housing prices

About:

Size of dataset: 528 KB

Features:

The dataset consists of the attributes such as

1. Number of bedrooms, bathrooms, floors

2. Condition - Rated in a range from 1 to 5

2. Square feet of the house

3. Year built

4. Location

5. Price of the property, etc..

All of the attributes are real in this data

Step 1: Load & View the Data

```
#Loading house prices csv file
options(repos = "https://cran.rstudio.com/")
housing_csv<-read.csv("/Users/sanjana/Desktop/MS/UTA/PS/Dataset/housePricing.csv")


#Displaying first 6 rows of the dataset
head(housing_csv)
```

```
##                 date   price bedrooms bathrooms sqft_living sqft_lot floors
## 1 2014-05-02 00:00:00  313000        3      1.50        1340     7912    1.5
## 2 2014-05-02 00:00:00 2384000        5      2.50        3650     9050    2.0
```

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

```
#Getting the dimensions of the dataset
dim(housing_csv)
```

```
## [1] 4600   18
```

This shows that there are 4600 rows and 18 columns in the given dataset

Step 2: Summarize the Data

Lets first start our analysis by getting a summary using describe()

describe() - It is a part of the Hmisc library. Provides critical statistical information about the dataset such as missing, distinct values,mean, lowest and highest values,etc..

```
#Importing the Hmisc library that contains the describe()
library(Hmisc)
#Applying describe() to the dataset
print(describe(housing_csv))
```

```
## housing_csv
##
##  18  Variables      4600  Observations
## ---------------------------------------------------------------------------
## date
##        n  missing distinct
##     4600        0       70
##
## lowest : 2014-05-02 00:00:00 2014-05-03 00:00:00 2014-05-04 00:00:00 2014-05-05 00:00:00 2014-05-06 0
## highest: 2014-07-06 00:00:00 2014-07-07 00:00:00 2014-07-08 00:00:00 2014-07-09 00:00:00 2014-07-10 0
## ---------------------------------------------------------------------------
## price
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0     1741        1   551963   356066   200000   239950
```

```
##      .25      .50      .75      .90      .95
##   322875   460943   654962   900000  1184050
##
## lowest :        0     7800    80000    83000    83300
## highest:  4489000  4668000  7062500 12899000 26590000
## ---------------------------------------------------------------------------
## bedrooms
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0       10    0.875    3.401   0.9486        2        2
##      .25      .50      .75      .90      .95
##        3        3        4        4        5
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency      2    38   566  2032  1531   353    61    14     2     1
## Proportion 0.000 0.008 0.123 0.442 0.333 0.077 0.013 0.003 0.000 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## bathrooms
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0       26    0.974    2.161   0.8521     1.00     1.00
##      .25      .50      .75      .90      .95
##     1.75     2.25     2.50     3.00     3.50
##
## lowest : 0    0.75 1    1.25 1.5 , highest: 5.75 6.25 6.5  6.75 8
## ---------------------------------------------------------------------------
## sqft_living
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0      566        1     2139     1015      950     1110
##      .25      .50      .75      .90      .95
##     1460     1980     2620     3340     3870
##
## lowest :   370   380   420   430   490, highest:  8020  8670  9640 10040 13540
## ---------------------------------------------------------------------------
## sqft_lot
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0     3113        1    14853    17219     1691     3300
##      .25      .50      .75      .90      .95
##     5001     7683    11001    24302    43560
##
## lowest :     638     681     704     746     747
## highest:  423838  435600  478288  641203 1074218
## ---------------------------------------------------------------------------
## floors
##        n  missing distinct     Info     Mean      Gmd
##     4600        0        6    0.832    1.512   0.5583
##
## Value        1.0   1.5   2.0   2.5   3.0   3.5
## Frequency   2174   444  1811    41   128     2
## Proportion 0.473 0.097 0.394 0.009 0.028 0.000
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------
## waterfront
```

```
##        n  missing distinct     Info      Sum     Mean      Gmd
##     4600        0        2    0.021       33 0.007174  0.01425
##
## ----------------------------------------------------------------------------
## view
##        n  missing distinct     Info     Mean      Gmd
##     4600        0        5    0.271   0.2407   0.4432
##
## Value          0     1     2     3     4
## Frequency   4140    69   205   116    70
## Proportion 0.900 0.015 0.045 0.025 0.015
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## condition
##        n  missing distinct     Info     Mean      Gmd
##     4600        0        5    0.735    3.452   0.6549
##
## Value          1     2     3     4     5
## Frequency      6    32  2875  1252   435
## Proportion 0.001 0.007 0.625 0.272 0.095
##
## For the frequency table, variable is rounded to the nearest 0
## ----------------------------------------------------------------------------
## sqft_above
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0      511        1     1827    912.2      860      970
##      .25      .50      .75      .90      .95
##     1190     1590     2300     3030     3440
##
## lowest :  370  380  420  430  490, highest: 6640 7320 7680 8020 9410
## ----------------------------------------------------------------------------
## sqft_basement
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0      207    0.787    312.1    445.8        0        0
##      .25      .50      .75      .90      .95
##        0        0      610     1000     1210
##
## lowest :    0   20   50   60   65, highest: 2550 2730 2850 4130 4820
## ----------------------------------------------------------------------------
## yr_built
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0      115        1     1971    33.74     1913     1925
##      .25      .50      .75      .90      .95
##     1951     1976     1997     2006     2009
##
## lowest : 1900 1901 1902 1903 1904, highest: 2010 2011 2012 2013 2014
## ----------------------------------------------------------------------------
## yr_renovated
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4600        0       60     0.79    808.6    964.9        0        0
##      .25      .50      .75      .90      .95
##        0        0     1999     2006     2011
##
```

```
## lowest :    0 1912 1913 1923 1934, highest: 2010 2011 2012 2013 2014
## ---------------------------------------------------------------------------
## street
##        n  missing distinct
##     4600        0     4525
##
## lowest : 1 View Ln NE        10 W Etruria St     100 20th Ave E      100 24th Ave E      100 Mt Si Pl
## highest: Shangri-La Way NW   Sunrise Loop Trail  Tolt Pipeline Trail Trossachs Blvd SE   Valley View
## ---------------------------------------------------------------------------
## city
##        n  missing distinct
##     4600        0       44
##
## lowest : Algona             Auburn             Beaux Arts Village Bellevue           Black Diamond
## highest: Snoqualmie Pass    Tukwila            Vashon             Woodinville        Yarrow Point
## ---------------------------------------------------------------------------
## statezip
##        n  missing distinct
##     4600        0       77
##
## lowest : WA 98001 WA 98002 WA 98003 WA 98004 WA 98005
## highest: WA 98188 WA 98198 WA 98199 WA 98288 WA 98354
## ---------------------------------------------------------------------------
## country
##        n  missing distinct    value
##     4600        0        1      USA
##
## Value       USA
## Frequency  4600
## Proportion    1
## ---------------------------------------------------------------------------
```

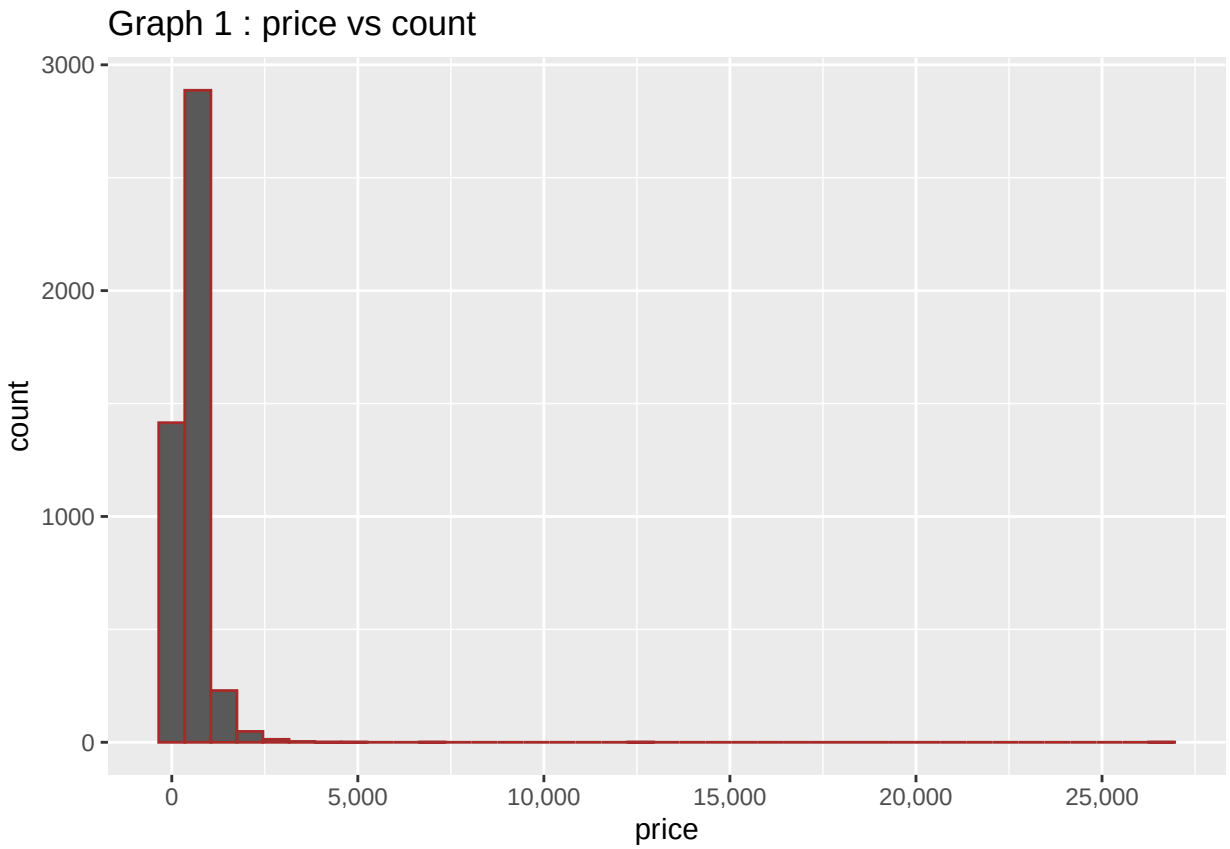
From the above we can understand that there are no missing values and hence we can move to the next step in the data wrangling process

Step 3: Visualize the Data using ggplot

1. Histogram depicting price distribution

```
library(ggplot2)
library(scales)
#Setting up the aesthetic of the plot by mapping price along x axis
ggplot(data=housing_csv, aes(x=price)) +

#Adding a histogram layer to the ggplot object and setting interval width a 1000
geom_histogram(color="brown",binwidth=700000) +
scale_x_continuous(breaks = seq(0, max(housing_csv$price), by = 5000000),labels = comma_format(scale = 1
ggtitle("Graph 1 : price vs count")
```



Here,

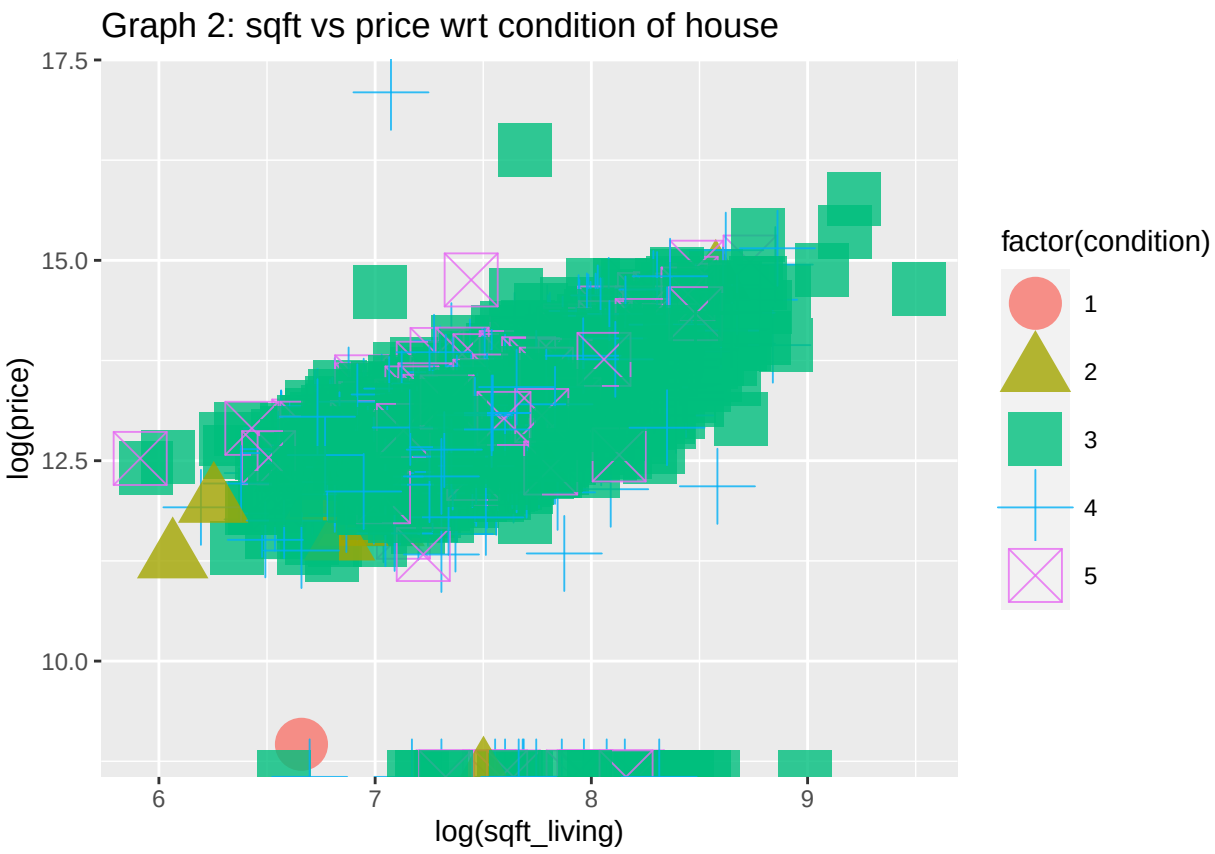
- scale_x_continuous is used to avoid exponential values and get a scaled range.
- breaks argument creates breaks at intervals of 5000000 starting from 0 up to the maximum value of the "price" variable
- labels argument is used to format labels with commas and scale values down by a factor of 1000 to make them more readable.

From the histogram above, we understand the following:

1. Maximum number of houses have an average price of around $600,000
2. Most houses are within $5000000
3. Count and price are inversely proportional, i.e the number of houses keeps decreasing as the price of the house gets expensive
4. Scatter plot of impact of sq feet on price based on condition

```
#Using the log function so as to distribute the data and compress due to high volume

p <- ggplot(housing_csv, aes(log(sqft_living), log(price), shape = factor(condition)))
p +
  geom_point(aes(colour = factor(condition)), size =9,alpha=0.8) +
   scale_x_continuous(labels = comma) +
   scale_y_continuous(labels = comma) +
ggtitle("Graph 2: sqft vs price wrt condition of house")
```

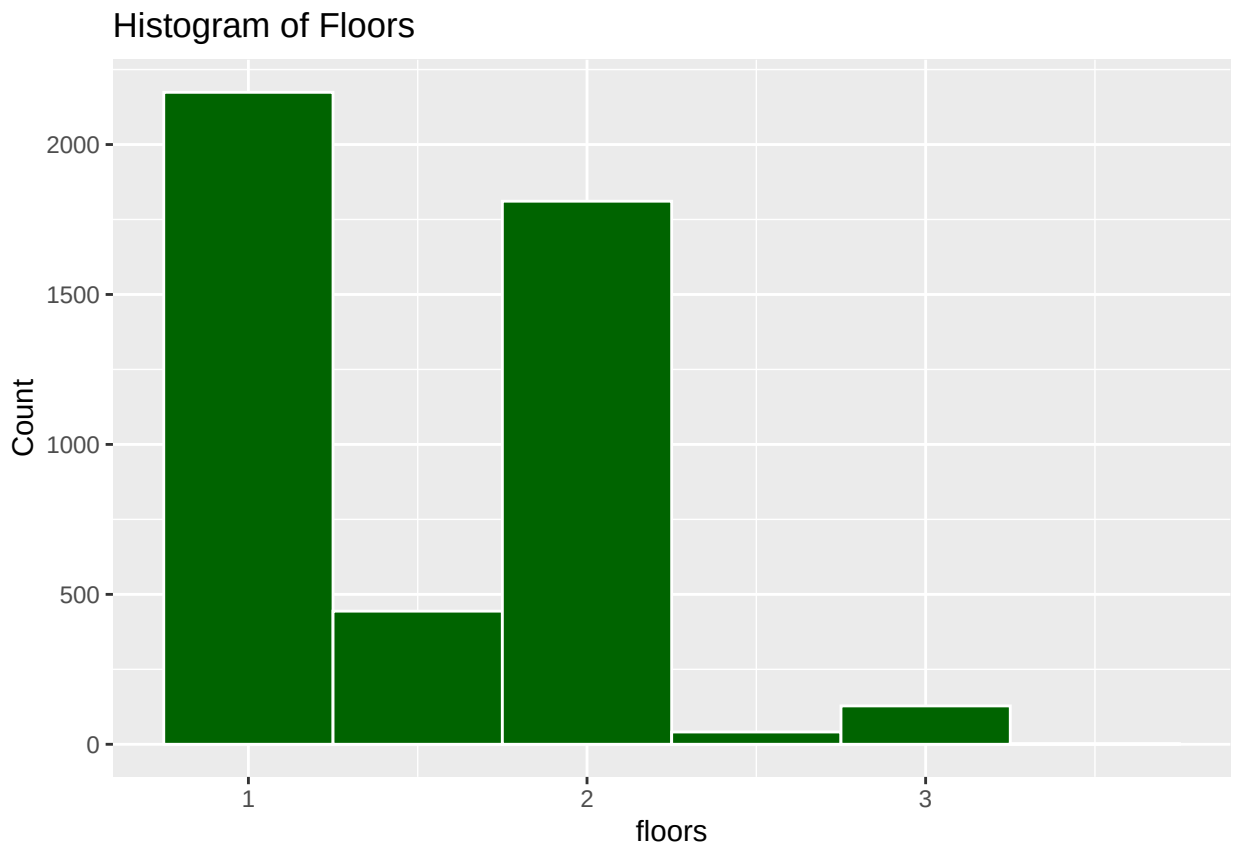


From the scatterplot we understand that:

- There are more number of houses with a mediocre condition of 3 and the price is also average.
- There is only one house that is of poor condition
- The price range of houses of very good quality range between 11.25 and 15. However, There is one house that is very expensive and square feet is also less.
- Strong, negative relationship: For an optimal solution, as the variable on the x-axis increases, the variable on the y-axis should decrease. The dots are packed tightly together, which indicates a strong relationship. The above depicts a strong relationship as desired.

3. Histogram

```
ggplot(data = housing_csv, aes(x = floors,color=floors)) +
  geom_histogram(colour="white",fill="darkgreen",binwidth = 0.5) +
  labs(title = "Histogram of Floors",
       x = "floors",
       y = "Count")
```



From the histogram, we can understand that there is more availability of 1 and 2 floor houses

4. CORRELATION MATRIX

- The correlation matrix is a very useful metric that helps to establish a relationship between two or more variables in a dataset
- The coefficient indicates both the strength of the relationship as well as the direction(Positive or Negative correlation)
- In our case, it helps to determine the factors that strongly and weakly impact the price of a house

```
#Loading the corrplot library
library(corrplot)
```

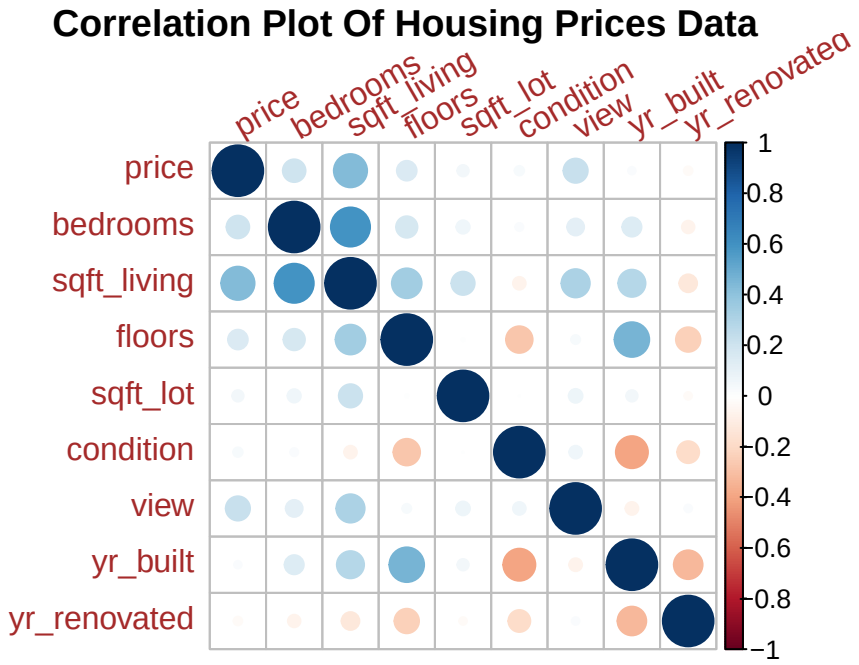
```
## corrplot 0.92 loaded
```

```
#Choosing the numeric columns for which correlation is to be calculated
num_housing <- housing_csv[,c("price","bedrooms","sqft_living","floors","sqft_lot", "condition", "view"
#Building the correlation matrix
cor_housing <- cor(num_housing)
cor_housing
```

```
##                    price    bedrooms sqft_living      floors     sqft_lot
## price          1.00000000  0.20033629  0.43041003  0.15146080  0.050451295
```

```
## bedrooms      0.20033629  1.00000000  0.59488406  0.17789490  0.068819355
## sqft_living   0.43041003  0.59488406  1.00000000  0.34485027  0.210538454
## floors        0.15146080  0.17789490  0.34485027  1.00000000  0.003749750
## sqft_lot      0.05045130  0.06881935  0.21053845  0.00374975  1.000000000
## condition     0.03491454  0.02507986 -0.06282598 -0.27501339  0.000558114
## view          0.22850417  0.11102800  0.31100944  0.03121095  0.073906741
## yr_built      0.02185683  0.14246104  0.28777522  0.46748066  0.050706346
## yr_renovated -0.02877365 -0.06108157 -0.12281688 -0.23399567 -0.022730309
##                 condition        view    yr_built yr_renovated
## price         0.034914537  0.22850417  0.02185683  -0.02877365
## bedrooms      0.025079856  0.11102800  0.14246104  -0.06108157
## sqft_living  -0.062825979  0.31100944  0.28777522  -0.12281688
## floors       -0.275013395  0.03121095  0.46748066  -0.23399567
## sqft_lot      0.000558114  0.07390674  0.05070635  -0.02273031
## condition     1.000000000  0.06307728 -0.39969823  -0.18681841
## view          0.063077281  1.00000000 -0.06446506   0.02296700
## yr_built     -0.399698234 -0.06446506  1.00000000  -0.32134228
## yr_renovated -0.186818414  0.02296700 -0.32134228   1.00000000
```

```
#Using the corrplot to develop a correlation plot to get a better visualisation of the relationship bet
corrplot(cor_housing, tl.col = "brown", tl.srt = 30, bg = "White",
         mar = c(3,3,3,3),title = "\n\n Correlation Plot Of Housing Prices Data",type = "full")
```



From the correlation plot, we derive the following:

- There is a negative relationship(-0.3213) of year built with the year renovated i.e properties that were built more recently tend to be less likely to have been renovated.

- The correlation between "price" and "view" is moderately positive (0.2285), indicating that there is a positive relationship between the view of the property and its price. In other words, properties with better views tend to have higher prices.
- "sqft_living" also has a moderate positive correlation with "price" (0.3110), suggesting that there is a positive relationship between the size of the living space and the price of the property. Larger living spaces tend to have higher prices.
- A coefficient of 0.46748066 between "year built" and the "number of floors" of a property suggests a moderate positive linear relationship between these two variables i.e recent properties have more number of floors.

Step 4: Identify Missing Values

```
sapply(housing_csv, function(x) sum(is.na(x)))
```

```
##          date         price      bedrooms     bathrooms   sqft_living
##             0             0             0             0             0
##      sqft_lot        floors    waterfront          view     condition
##             0             0             0             0             0
##    sqft_above sqft_basement      yr_built  yr_renovated        street
##             0             0             0             0             0
##          city      statezip       country
##             0             0             0
```

- The result indicates that there are no missing values in any of the factors impacting housing prices