STUDENT: TEJASHRI DENDI **UIN**: 670653291

Programming Project (100 points)

The goal of this project is to evaluate the understanding of the knowledge we learned from class, and ability to apply the knowledge into real industry areas. It requires students to download a dataset and analyze it using the techniques in our class. Students must explore the relationship between one variable (for example, the score variable) with other variables in the dataset. Finally, write a report to answer the following questions and describe the steps and results. The program language must be R or Python and the dataset must be from kaggle.com.

The detailed steps are required as follows:

1. Select a dataset in the website: kaggle.com and download it.

   The dataset is collected from kaggle.

2. Describe the data set using a paragraph including its source, sample numbers, and variables and their meaning, etc. (10 points)

Exploration of real estate data

Can we predict housing prices based on the features?

How are housing price and location attributes correlated?

What is the overall picture of the USA housing prices w.r.t. locations?

Do house attributes (bedroom, bathroom count) strongly correlate with the price? Are there any hidden patterns?

Dataset description:

Downloaded the dataset realtor.csv from kaggle. This dataset is from Kaggle and contains both categorical/discrete (nominal and ordinal) and numeric (continuous) variables scraped from www.realtor.com real estate website. The data has over 900K observations (houses) and 12 columns (various attributes of houses). The goal is to explore the price variable and find an association between house attributes and its price.

The dataset has 1 CSV file with 10 columns -

realtor-data.csv (2,226,382 entries)

brokered by (categorically encoded agency/broker)

status (Housing status - a. ready for sale or b. ready to build)

price (Housing price, it is either the current listing price or recently sold price if the house is sold recently)

bed (# of beds)

bath (# of bathrooms)

acre_lot (Property / Land size in acres)

street (categorically encoded street address)

city (city name)

state (state name)

zip_code (postal code of the area)

house_size (house area/size/living space in square feet)

prev_sold_date (Previously sold date)

NB:

brokered by and street addresses were categorically encoded due to privacy policy

acre_lot means the total land area, and house_size denotes the living space/building area

3) Explore the overall summary of the dataset using codes and show the results (10 points)

Summary of the dataset:

```{r}
df=read.csv("C:/Users/Swathi/Downloads/realtor-1.csv")
str(df)
summary(df)
```

```
    status             price                bed              bath        
 Length:923159      Min.   :        0   Min.   :  1.00   Min.   :  1.00  
 Class :character   1st Qu.:   269000   1st Qu.:  2.00   1st Qu.:  1.00  
 Mode  :character   Median :   475000   Median :  3.00   Median :  2.00  
                    Mean   :   884123   Mean   :  3.33   Mean   :  2.49  
                    3rd Qu.:   839900   3rd Qu.:  4.00   3rd Qu.:  3.00  
                    Max.   :875000000   Max.   :123.00   Max.   :198.00  
                    NA's   :71          NA's   :131703   NA's   :115192  
    state              zip_code       house_size        sold_date        
 Length:923159      Min.   :  601   Min.   :    100   Length:923159      
 Class :character   1st Qu.: 2919   1st Qu.:   1130   Class :character   
 Mode  :character   Median : 7004   Median :   1651   Mode  :character   
                    Mean   : 6590   Mean   :   2142                      
                    3rd Qu.:10001   3rd Qu.:   2499                      
                    Max.   :99999   Max.   :1450112                      
                    NA's   :205     NA's   :297843                       
```

```
    acre_lot          full_address           street              city
 Min.    :     0.00   Length:923159      Length:923159      Length:923159
 1st Qu.:     0.11   Class :character   Class :character   Class :character
 Median :     0.29   Mode  :character   Mode  :character   Mode  :character
 Mean   :    17.08
 3rd Qu.:     1.15
 Max.   :100000.00
 NA's   :273623
```

Structure of each variable:

```
'data.frame':   923159 obs. of  12 variables:
 $ status      : chr  "for_sale" "for_sale" "for_sale" "for_sale" ...
 $ price       : num  105000 80000 67000 145000 65000 179000 50000 71600 100000 300000 ...
 $ bed         : num  3 4 2 4 6 4 3 3 2 5 ...
 $ bath        : num  2 2 1 2 2 3 1 2 1 3 ...
 $ acre_lot    : num  0.12 0.08 0.15 0.1 0.05 0.46 0.2 0.08 0.09 7.46 ...
 $ full_address: chr  "Sector Yahuecas Titulo # V84, Adjuntas, PR, 00601" "Km 78 9 Carr # 13
El Paraso Calle De Oro R-5 Ponce, Ponce, PR, 00731" ...
 $ street      : chr  "Sector Yahuecas Titulo # V84" "Km 78 9 Carr # 135" "556G 556-G 16 St'
 $ city        : chr  "Adjuntas" "Adjuntas" "Juana Diaz" "Ponce" ...
 $ state       : chr  "Puerto Rico" "Puerto Rico" "Puerto Rico" "Puerto Rico" ...
 $ zip_code    : num  601 601 795 731 680 612 639 731 730 670 ...
 $ house_size  : num  920 1527 748 1800 NA ...
 $ sold_date   : chr  "" "" "" "" ...
```

Execution:

```{r}
df=read.csv("C:/Users/Swathi/Downloads/realtor-1.csv")
str(df)
summary(df)
```

```
'data.frame':   923159 obs. of  12 variables:
 $ status      : chr  "for_sale" "for_sale" "for_sale" "for_sale" ...
 $ price       : num  105000 80000 67000 145000 65000 179000 50000 71600 1
 $ bed         : num  3 4 2 4 6 4 3 3 2 5 ...
 $ bath        : num  2 2 1 2 2 3 1 2 1 3 ...
 $ acre_lot    : num  0.12 0.08 0.15 0.1 0.05 0.46 0.2 0.08 0.09 7.46 ...
 $ full_address: chr  "Sector Yahuecas Titulo # V84, Adjuntas, PR, 00601"
El Paraso Calle De Oro R-5 Ponce, Ponce, PR, 00731" ...
 $ street      : chr  "Sector Yahuecas Titulo # V84" "Km 78 9 Carr # 135"
 $ city        : chr  "Adjuntas" "Adjuntas" "Juana Diaz" "Ponce" ...
 $ state       : chr  "Puerto Rico" "Puerto Rico" "Puerto Rico" "Puerto Ri
 $ zip_code    : num  601 601 795 731 680 612 639 731 730 670 ...
 $ house_size  : num  920 1527 748 1800 NA ...
 $ sold_date   : chr  "" "" "" "" ...
   status               price                bed              bath
Length:923159      Min.   :        0   Min.   :  1.00   Min.   :  1.00
Class :character   1st Qu.:   269000   1st Qu.:  2.00   1st Qu.:  1.00
Mode  :character   Median :   475000   Median :  3.00   Median :  2.00
                   Mean   :   884123   Mean   :  3.33   Mean   :  2.49
                   3rd Qu.:   839900   3rd Qu.:  4.00   3rd Qu.:  3.00
                   Max.   :875000000   Max.   :123.00   Max.   :198.00
                   NA's   :71          NA's   :131703   NA's   :115192
   state              zip_code       house_size       sold_date
Length:923159      Min.   :  601   Min.   :    100   Length:923159
Class :character   1st Qu.: 2919   1st Qu.:   1130   Class :character
Mode  :character   Median : 7004   Median :   1651   Mode  :character
                   Mean   : 6590   Mean   :   2142
                   3rd Qu.:10001   3rd Qu.:   2499
                   Max.   :99999   Max.   :1450112
                   NA's   :205     NA's   :297843
```

Analysis from the above is as follows:

```
# bed           : numerical  discrete #3 4 2 4 6 4 3 3 2 5 ...
# bath          : numerical discrete #2 2 1 2 2 3 1 2 1 3 ...
# acre_lot      : numerical continous #0.12 0.08 0.15 0.1 0.05 0.46 0.2
# city          : categorical nominal
# state         : categorical nominal
# zip_code      : numerical  discrete #601 601 795 731 680 612 639 731
# house_size    : numerical  discrete #920 1527 748 1800 NA ...
# prev_sold_date: categorical ordinal
# price         : numerical, discrete  #105000 80000 67000 145000 65000
```

3. Clean the data by removing the row or column with invalid or missing values and show the results. (10 points)

   Cleaning data:

   a) Checking for duplicated rows:

```r
# Check for duplicate observations
has_duplicates <- duplicated(df)



# Remove duplicate observations
unique_df <- df[!has_duplicates, ]

# Print the dimensions of the dataset before and after removing duplicates
cat("Original dataset dimensions:", dim(df), "\n")
cat("Dataset dimensions after removing duplicates:", dim(unique_df), "\n")
```

```
Original dataset dimensions: 923159 12
Dataset dimensions after removing duplicates: 113789 12
```

b) Checking missing values in variables of the dataset:

```r
# Check for missing values in each column in original dataset


missing_values_df <- colSums(is.na(df))
cat("Columns with missing values in  original dataset:", missing_values_df, "\n")


# Print columns with missing values
columns_with_missing <- names(missing_values_df[missing_values_df > 0])
cat("Columns with missing values:", columns_with_missing, "\n")

# Print the number of missing values for each column
cat("Number of missing values for each column:\n")
print(missing_values_df[missing_values_df > 0])

missing_val_uniqdf <- colSums(is.na(unique_df))
cat("Columns with missing values in distinct dataset:", missing_val_uniqdf, "\n")

# Print columns with missing values
uniqdf_missing_cols <- names(missing_values_df[missing_val_uniqdf > 0])
cat("Columns with missing values:", uniqdf_missing_cols, "\n")



```

```
Columns with missing values in  original dataset: 0 71 131703 115192 273623 0 0 0 0
Columns with missing values: price bed bath acre_lot zip_code house_size
Number of missing values for each column:
     price        bed       bath   acre_lot   zip_code house_size
        71     131703     115192     273623        205     297843
Columns with missing values in distinct dataset: 0 18 17516 16297 31123 0 0 0 0 33 3
Columns with missing values: price bed bath acre_lot zip_code house_size
```

It is cleared from the above that there are 5 columns with missing values namely Price, bed, bath, acre_lot, zip_code and house_size.

4. Decide the variables you want to analyze and explain the reason in words. (5 points)

   house_price and the variables - house_size, bed, and

   Bath, acre_lot and state are used here to explore and analyze as these are the main attributes of house may lead to variation in price.

   Removed all houses with price less than or equal to 50K:

```r
# Create a new dataset without houses with price <= 50K
no50kp_uniqdf <- unique_df[unique_df$price > 50000, ]

# Print dimensions of the dataset before and after removing houses
cat("Original dataset dimensions:", dim(unique_df), "\n")
cat("Dataset dimensions after removing houses with price <= 50K:", dim(no50kp_uniqdf), "\n")
```

```
Original dataset dimensions: 113789 12
Dataset dimensions after removing houses with price <= 50K: 110480 12
```

   Removal of outliers:

   The price variable appears to have some extreme values. So, removed the outliers in the "price" variable using the IQR method.

```r
# Calculate the lower and upper bounds for outliers using the IQR method
lower_bound_val <- quantile(no50kp_uniqdf$price, 0.25, na.rm = TRUE) - 1.5 * IQR(no50kp_uniqdf$price, na.rm = TRUE)
upper_bound_val <- quantile(no50kp_uniqdf$price, 0.75, na.rm = TRUE) + 1.5 * IQR(no50kp_uniqdf$price, na.rm = TRUE)

# Remove outliers from the dataset
outliers_removed_df <- no50kp_uniqdf[no50kp_uniqdf$price >= lower_bound_val & no50kp_uniqdf$price <= upper_bound_val, ]

# Print dimensions of the dataset before and after removing outliers
cat("Original dataset dimensions:", dim(no50kp_uniqdf), "\n")
cat("Dataset dimensions after removing outliers:", dim(outliers_removed_df), "\n")
```

```
Original dataset dimensions: 110480 12
Dataset dimensions after removing outliers: 98804 12
```

   percentage of the observations are missing for the price variable:

```{r}
# Calculate the percentage of missing values for the 'price' variable
missing_percentage_uniqdfprice <- sum(is.na(unique_df$price)) / length(unique_df$price) * 100

# Print the result
cat("Percentage of missing observations for the 'price' variable of distinct dataset:", missing_percentage_uniqdfprice, "%\n")
# Calculate the pe                                                                    variable
missing_percentage_                                                                  df$price) * 100

# Print the result
cat("Percentage of missing observations for the 'price' variable of original dataset:", missing_percentage_dfprice, "%\n")
```

cat(..., file = "", sep = " ", fill = FALSE, labels = NULL, append = FALSE)

```
Percentage of missing observations for the 'price' variable of distinct dataset: 0.01581875 %
Percentage of missing observations for the 'price' variable of original dataset: 0.007690983 %
```

percentage of the observations are missing for the price variable after removing outliers:

```{r}
# Calculate the percentage of missing values for the 'price' variable
missing_percentage_outrmdfprice <- sum(is.na(outliers_removed_df$price)) / length(outliers_removed_df$price) * 100

# Print the result
cat("Percentage of missing observations for the 'price' variable fromoutliers removed dataframe:", missing_percentage_outrmdfprice, "%\n")
```

```
Percentage of missing observations for the 'price' variable fromoutliers removed dataframe: 0.01821789 %
```

Conversion of the “state” attribute to factor and removed states with only one observation from the #data:

```{r}
# Convert 'state' to factor
outliers_removed_df$state <- as.factor(outliers_removed_df$state)

# Take a summary to see the count of observations for each state
summary_state <- table(outliers_removed_df$state)
print(summary_state)

# Remove states with only one observation
states_to_remove <- names(summary_state[summary_state == 1])
after_fewstates_removed <- outliers_removed_df[!(outliers_removed_df$state %in% states_to_remove), ]

# Print the dimensions of the dataset before and after removing states
cat("Original dataset dimensions:", dim(df), "\n")
cat("Outliers removed dataset dimensions:", dim(outliers_removed_df), "\n")
cat("Dataset dimensions after removing states with only one observation:", dim(after_fewstates_removed), "\n")
```

```
   Connecticut       Delaware        Georgia          Maine  Massachusetts  New Hampshire     New Jersey
         12674           1262              5           4012           8673           3234          30363
       Vermont Virgin Islands       Virginia  West Virginia        Wyoming
          2206            606              7              1              1
Original dataset dimensions: 923159 12
Outliers removed dataset dimensions: 98804 14
Dataset dimensions after removing states with only one observation: 98802 14
```

5. Draw at least two different type of basic graphs to describe the distributions of the variables. (20 points)

   histogram and boxplot of the price:

   Code of histogram:

   ggplot(outliers_removed_df, aes(x = price)) +

   geom_histogram(binwidth = 5000, fill = "skyblue", color = "black", aes(y = ..density..), alpha = 0.7) +

   geom_density(color = "red") +

   labs(title = "Histogram of Price", x = "Price", y = "Density")

   Code of Boxplot:

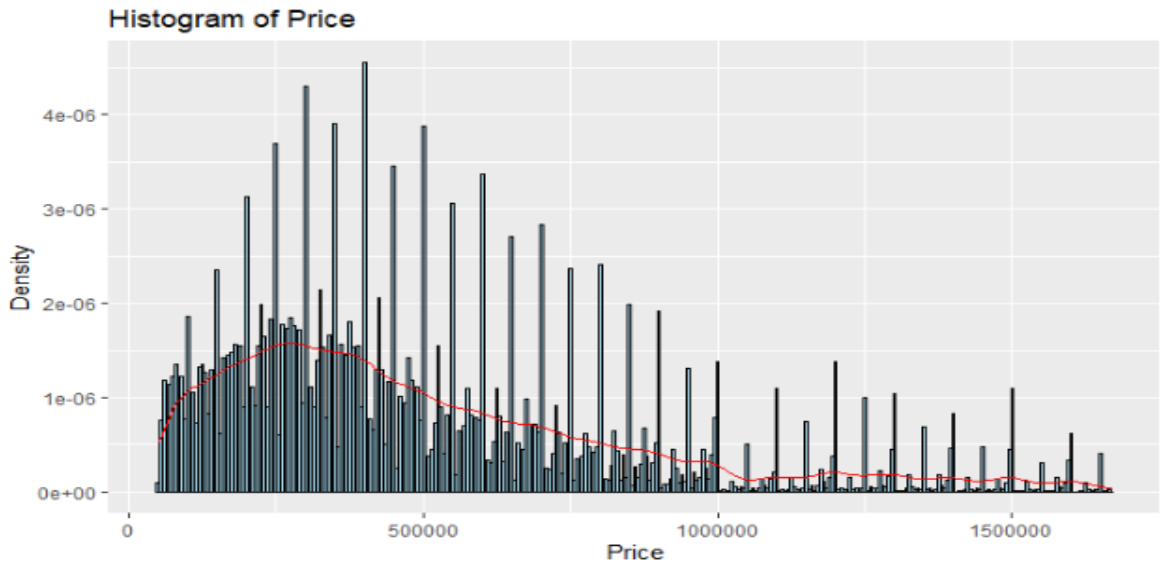
   ggplot(outliers_removed_df, aes(x = 1, y = price)) +

   geom_boxplot(fill = "lightgreen", color = "black") +

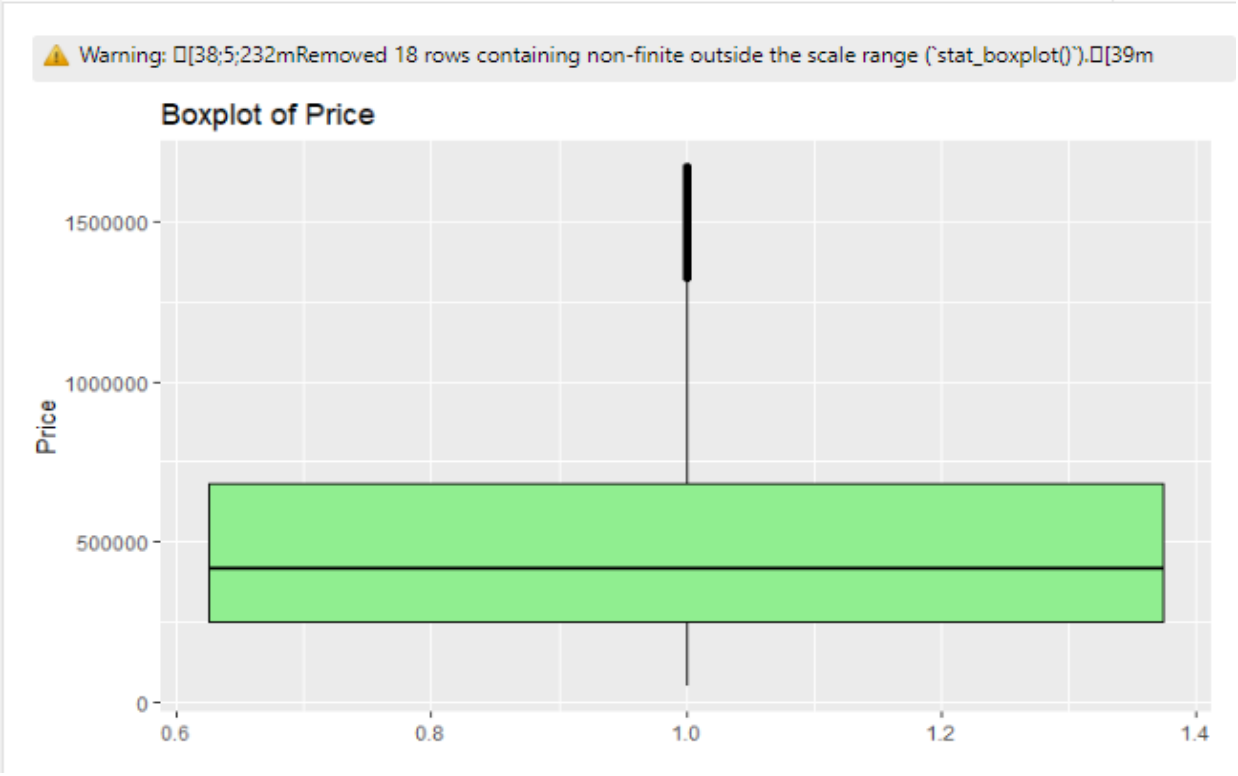
   labs(title = "Boxplot of Price", x = "", y = "Price")

   Execution and result:

```{r}
#install.packages("ggplot2")
# Load necessary libraries
library(ggplot2)

# Draw a histogram
ggplot(outliers_removed_df, aes(x = price)) +
geom_histogram(binwidth = 5000, fill = "skyblue", color = "black", aes(y = ..density..), alpha = 0.7) +
geom_density(color = "red") +
labs(title = "Histogram of Price", x = "Price", y = "Density")


```




```{r}
# Draw a boxplot
ggplot(outliers_removed_df, aes(x = 1, y = price)) +
geom_boxplot(fill = "lightgreen", color = "black") +
labs(title = "Boxplot of Price", x = "", y = "Price")
```



From the above, the histogram is skewed to the right.

Calculation of skewness of price variable:

```r
# calculate skewness in r
install.packages("moments")
library(moments)

# Calculate skewness
skewness_val <- skewness(outliers_removed_df$price, na.rm = TRUE)
cat("Skewness of Price variable:", skewness_val, "\n")
```

```
Skewness of Price variable: 1.1601
```

As the skewness value is positive, the distribution of price variable is positively skewed.

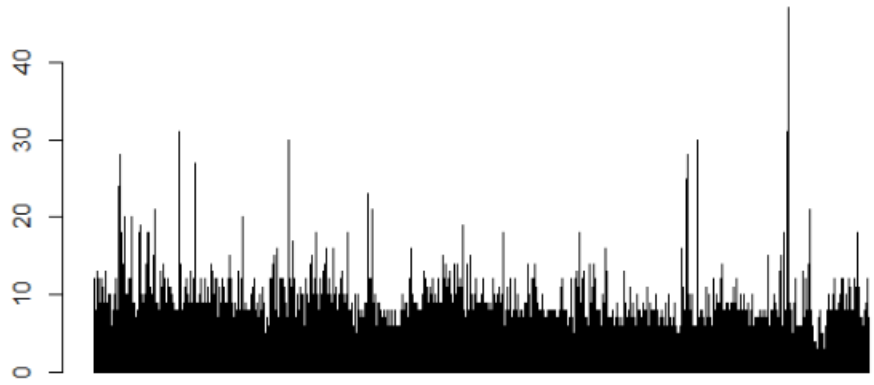
Graphs of 'bed' variable:

barplot(outliers_removed_df$bed)

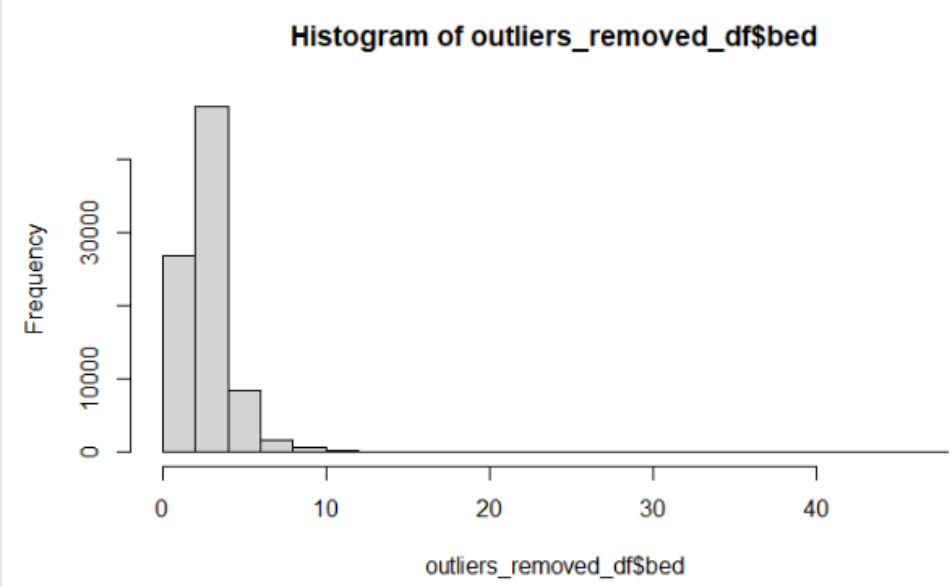
hist(outliers_removed_df$bed)

plot(outliers_removed_df$bed)

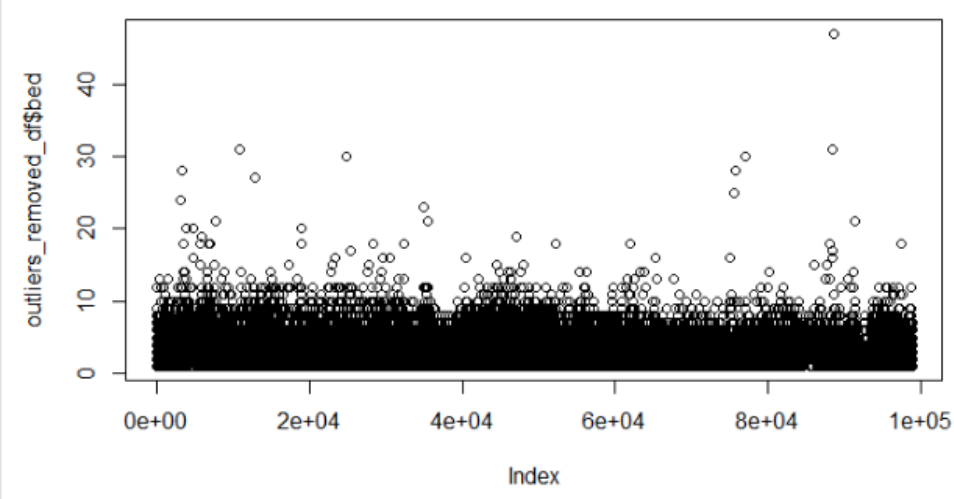
```r
barplot(outliers_removed_df$bed)
hist(outliers_removed_df$bed)
plot(outliers_removed_df$bed)S
```



Histogram of 'bed':



Scatterplot of 'bed':

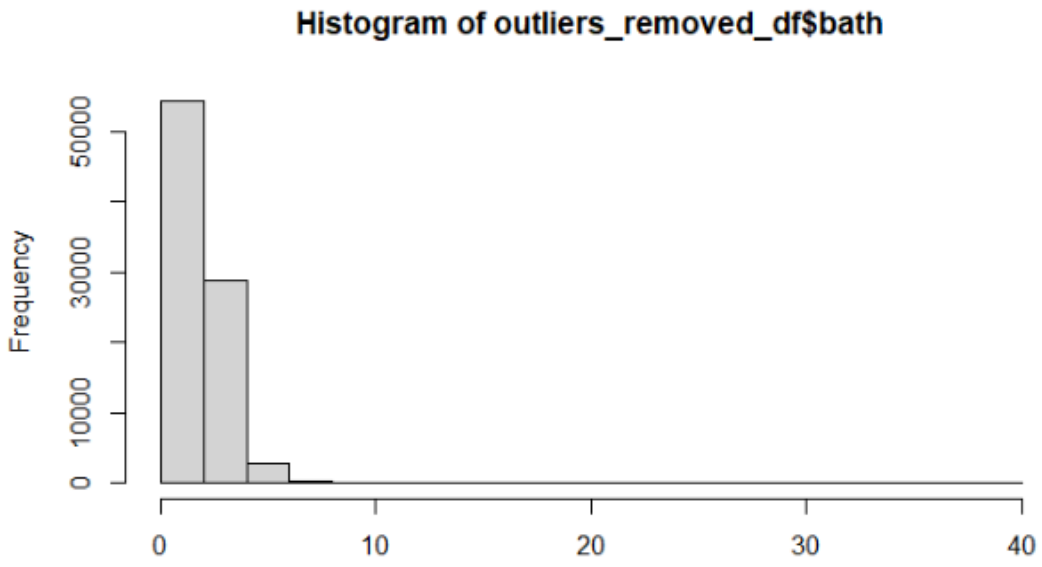


<u>Graphs of 'bath' variable:</u>

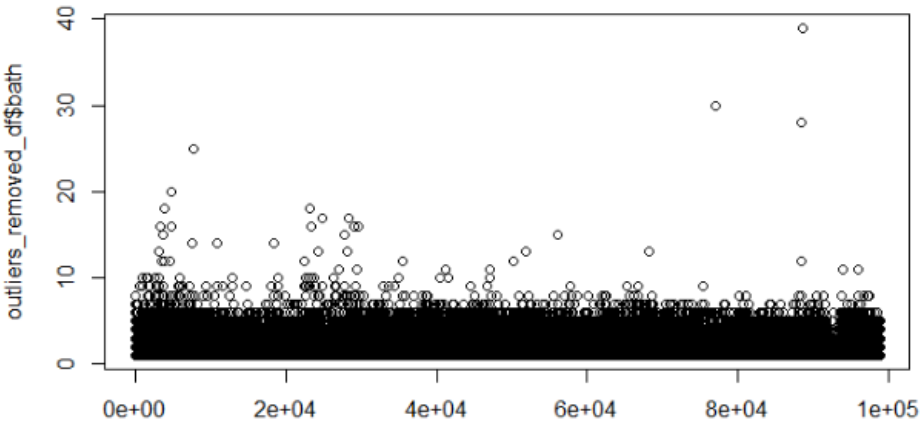
hist(outliers_removed_df$bath)

plot(outliers_removed_df$bath)

```r
barplot(outliers_removed_df$bath)
hist(outliers_removed_df$bath)
plot(outliers_removed_df$bath)
```

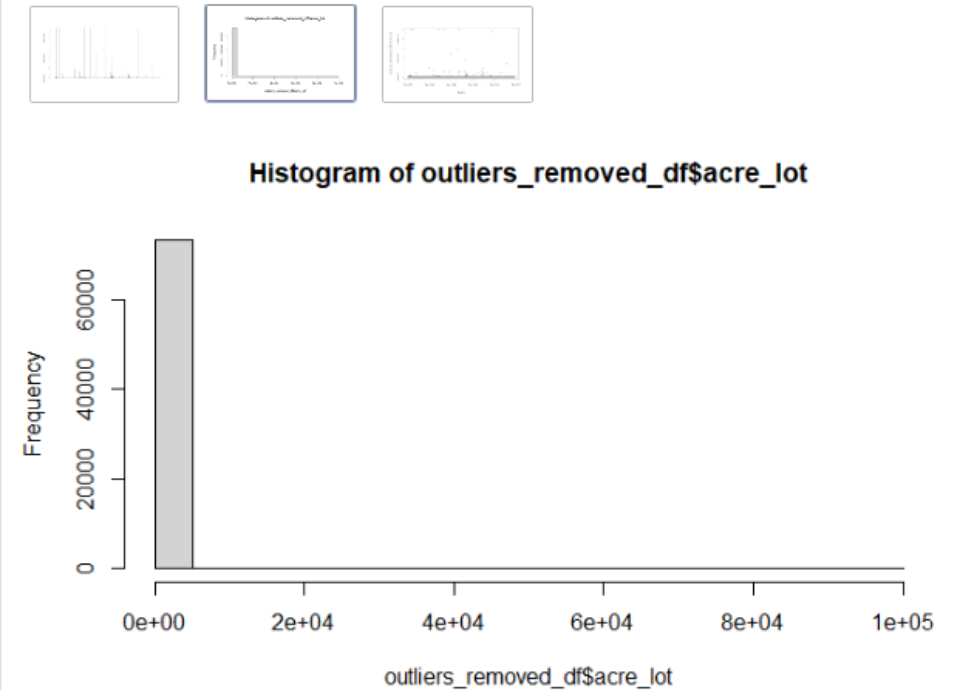


Scatterplot of 'bath' variable:

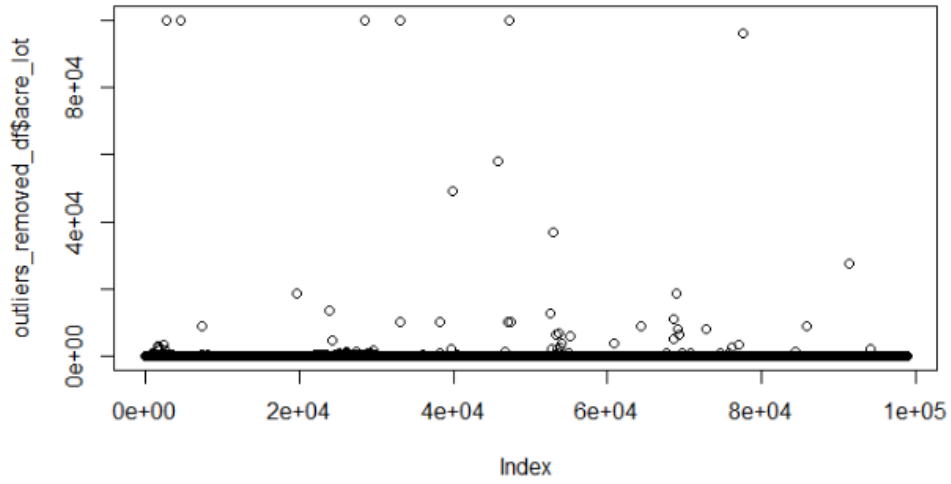


Graphs of 'acre_lot' variable:

```
hist(outliers_removed_df$acre_lot)
plot(outliers_removed_df$acre_lot)
```



Scatterplot of 'acre_lot' variable:



6. Add more explanation in the above graphs using at least four

Advanced techniques such as legend, text, graph title, grid line:

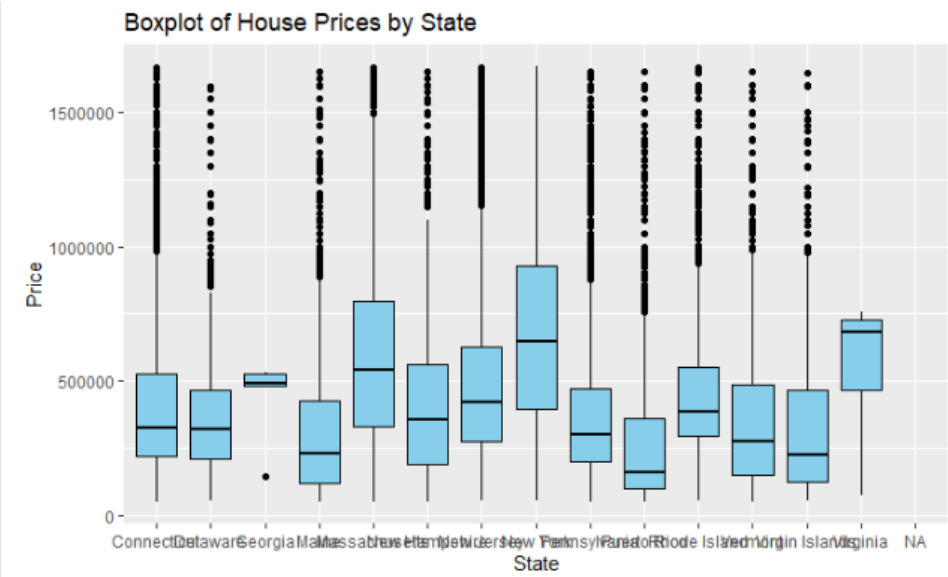
Used labs():

ggplot(after_fewstates_removed, aes(x = state, y = price)) +

geom_boxplot(fill = "skyblue", color = "black") +

labs(title = "Boxplot of House Prices by State", x = "State", y = "Price")

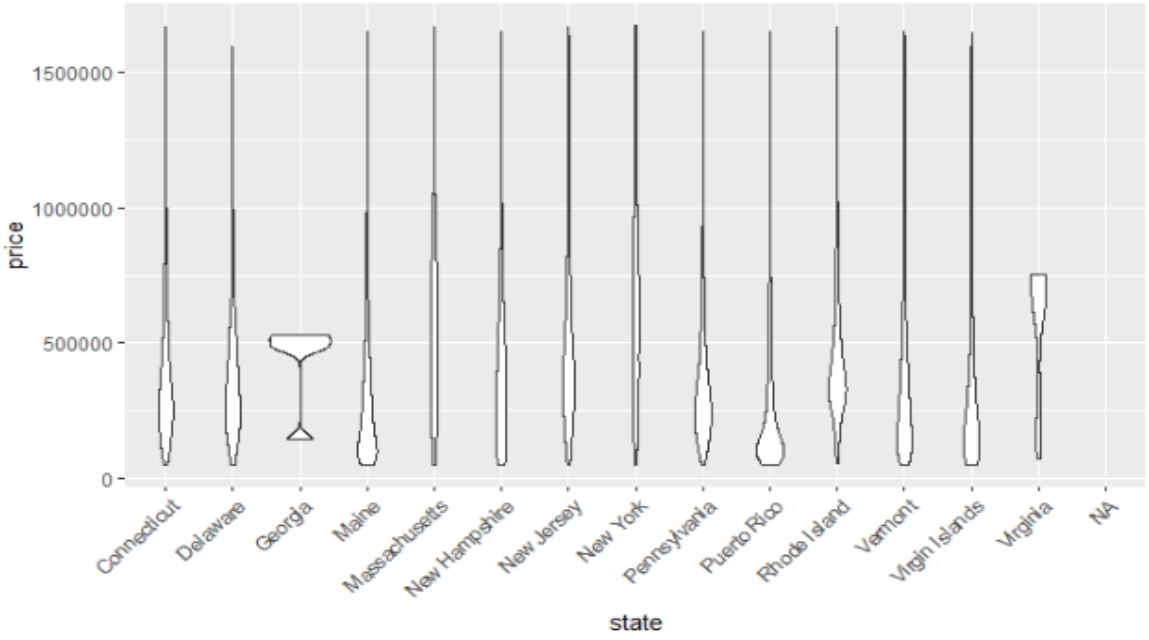
```r
# Load necessary libraries
library(ggplot2)
library(tidyr)

# Create a boxplot
ggplot(after_fewstates_removed, aes(x = state, y = price)) +
geom_boxplot(fill = "skyblue", color = "black") +
labs(title = "Boxplot of House Prices by State", x = "State", y = "Price")
```



Violin plot of state ~ price with **theme**():

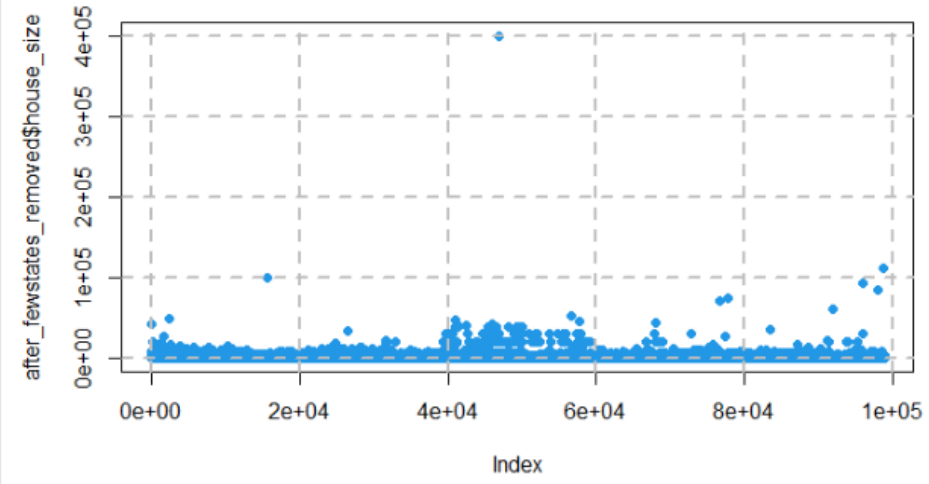
```r
#  plotting a violin plot - state and price
ggplot(data = after_fewstates_removed, aes(x = state, y = price)) +
  geom_violin() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: [38;5;232mRemoved 18 rows containing non-finite outside the scale range (`stat_ydensity()`).[39m
```



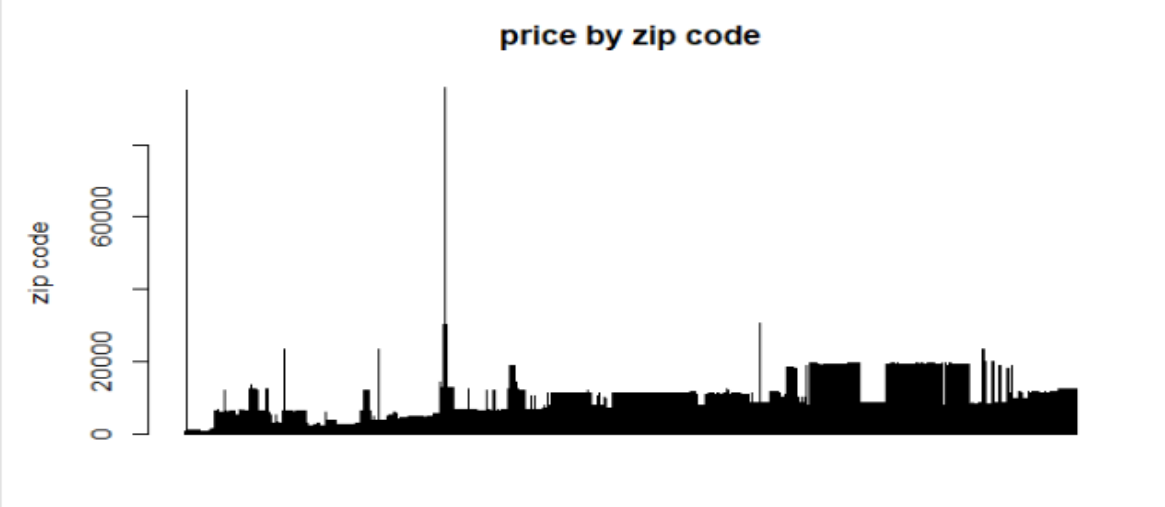
Distribution of ‘house size’ variable (used **grid**()):

```r
plot(after_fewstates_removed$house_size, pch = 19, col = 4)
grid(nx = NULL, ny = NULL,lty = 2, col = "gray", lwd = 2)
```



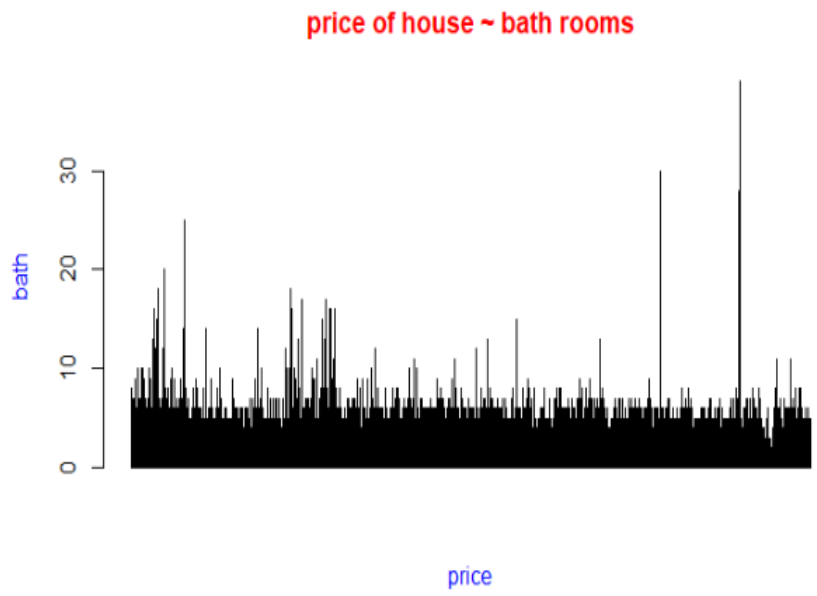
Title of graph using **main**(): Zip code ~ price

```r
barplot(after_fewstates_removed$zip_code, main="price by zip code", ylab = "zip code")
```



bath rooms ~ price (used **text**()):

```{r}
barplot(after_fewstates_removed$bath, main=" price of house ~ bath rooms ",xlab="price", ylab="bath", col.main="red",
col.lab="blue")


```

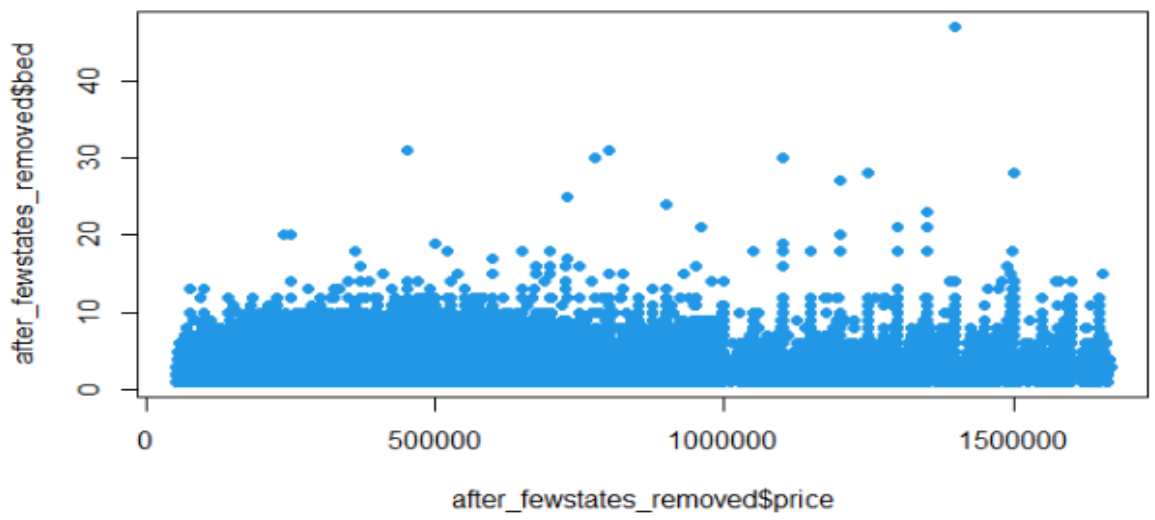


<u>The relationship between 'price' and other variables:</u>

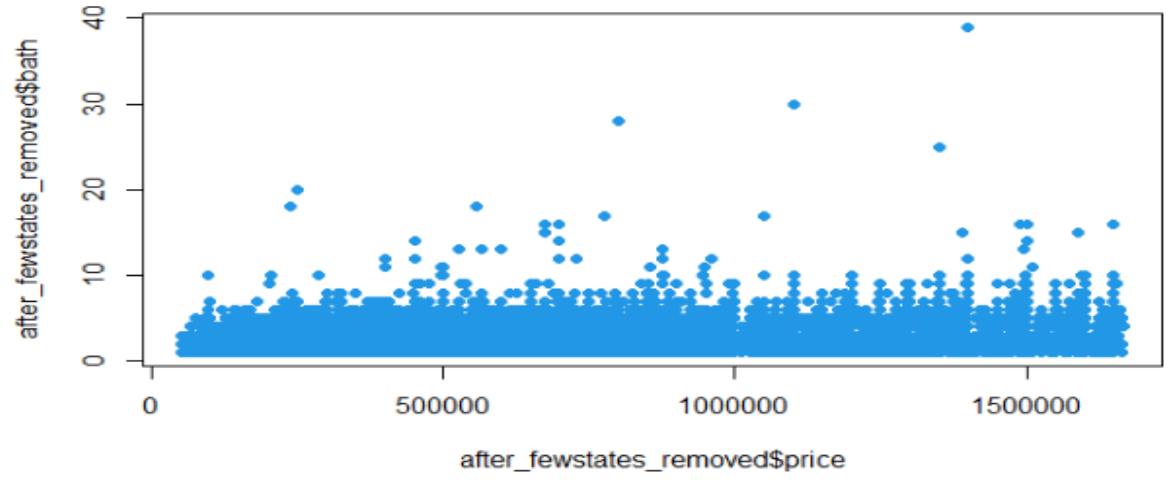
<u>Type1:</u>

Scatter plot - Price ~ bed rooms:

```{r}
plot(after_fewstates_removed$price, after_fewstates_removed$bed, pch = 19, col = 4)
```

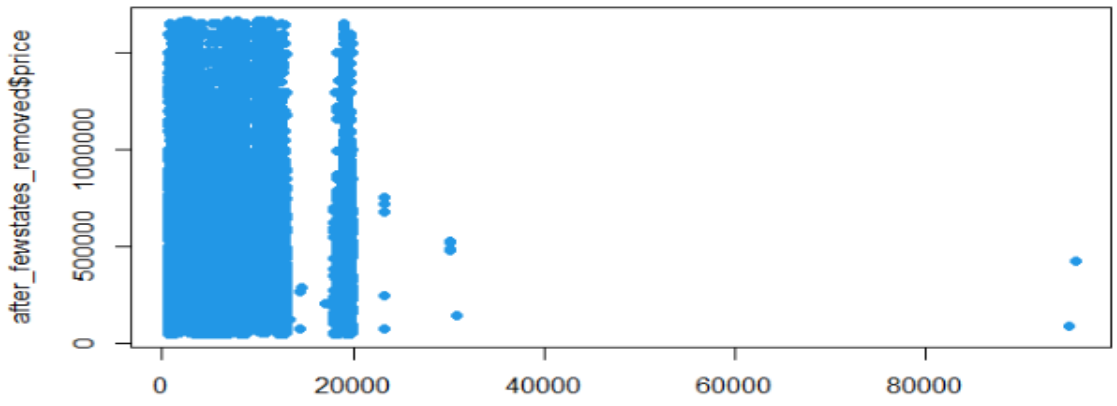



```{r}
plot(after_fewstates_removed$price, after_fewstates_removed$bath, pch = 19, col = 4)
```



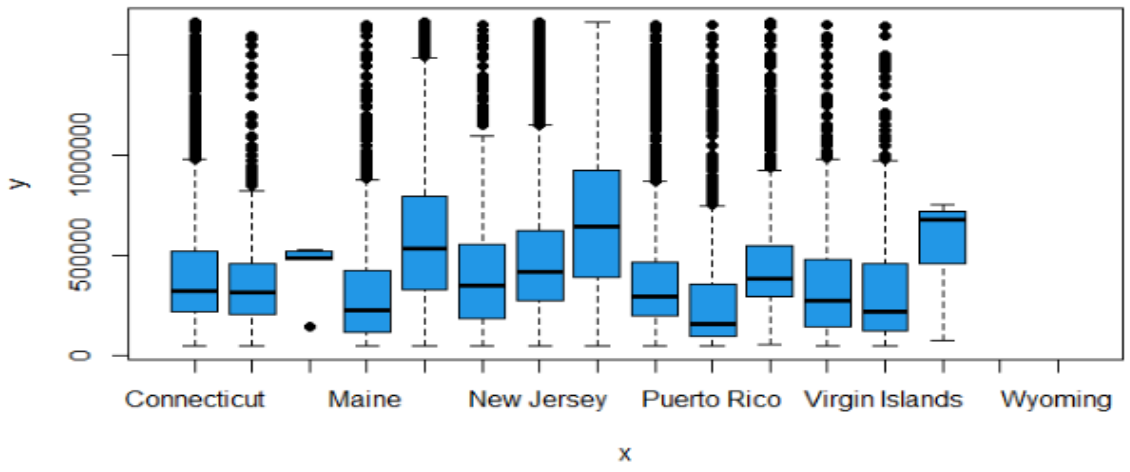
Zipcode and price:

```r
plot(after_fewstates_removed$zip_code, after_fewstates_removed$price, pch = 19, col = 4)
```



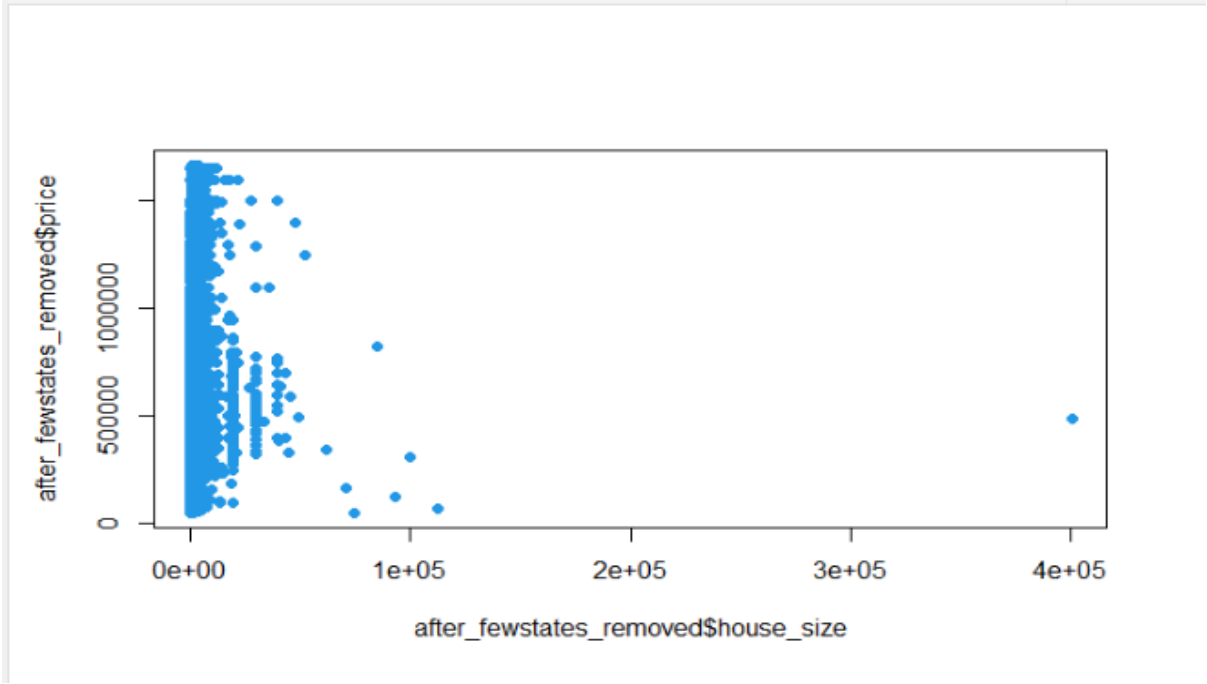
Boxplot of state ~ price:

```r
plot(after_fewstates_removed$state, after_fewstates_removed$price, pch = 19, col = 4)
```



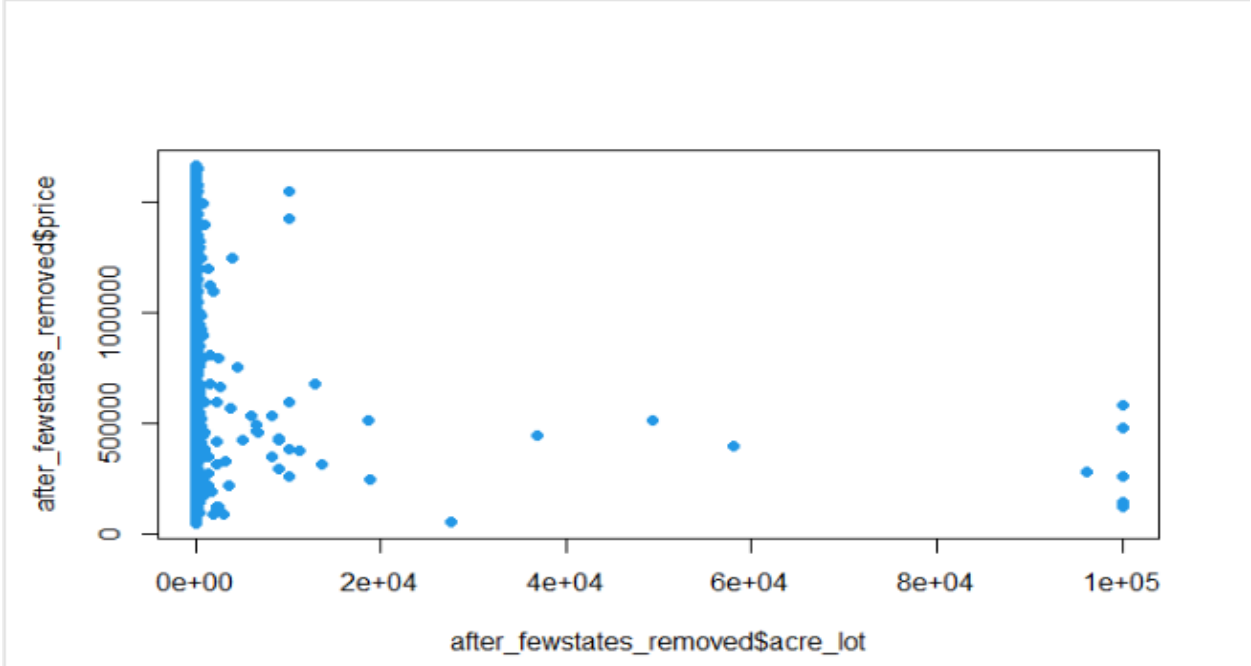
House_size vs price:

```r
plot(after_fewstates_removed$house_size, after_fewstates_removed$price, pch = 19, col = 4)
```



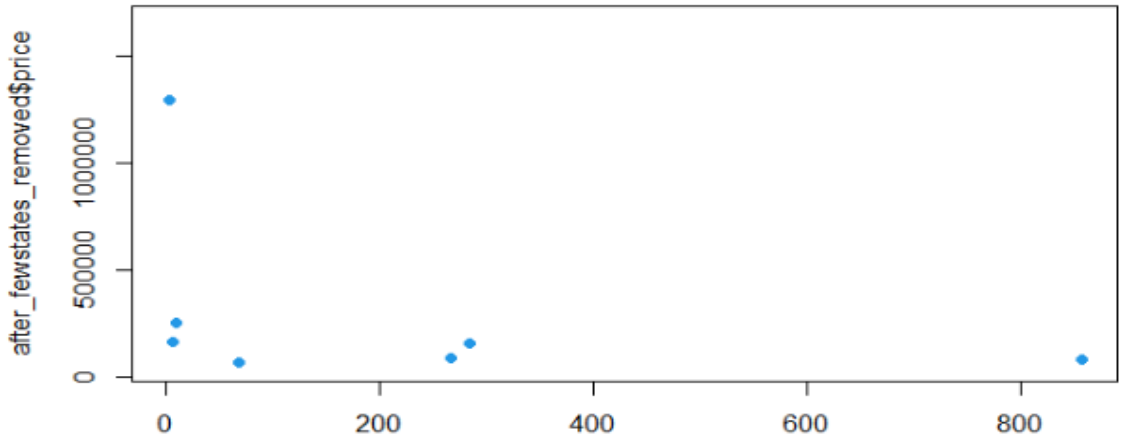
Acre_lot ~price:

```r
plot(after_fewstates_removed$acre_lot, after_fewstates_removed$price, pch = 19, col = 4)
```



Street ~price:

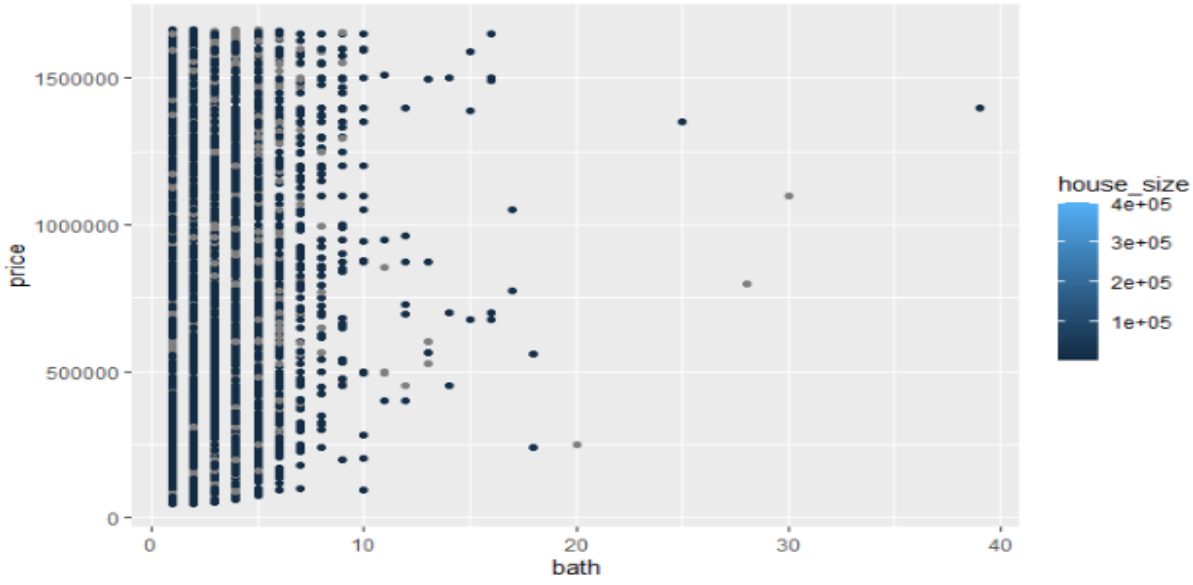
```
plot(after_fewstates_removed$street, after_fewstates_removed$price, pch = 19, col = 4)
```



Scatterplot of number of bathrooms, price, house_size:

```{r}
ggplot(after_fewstates_removed,aes(x=bath,y=price,color=house_size))+geom_point()
```

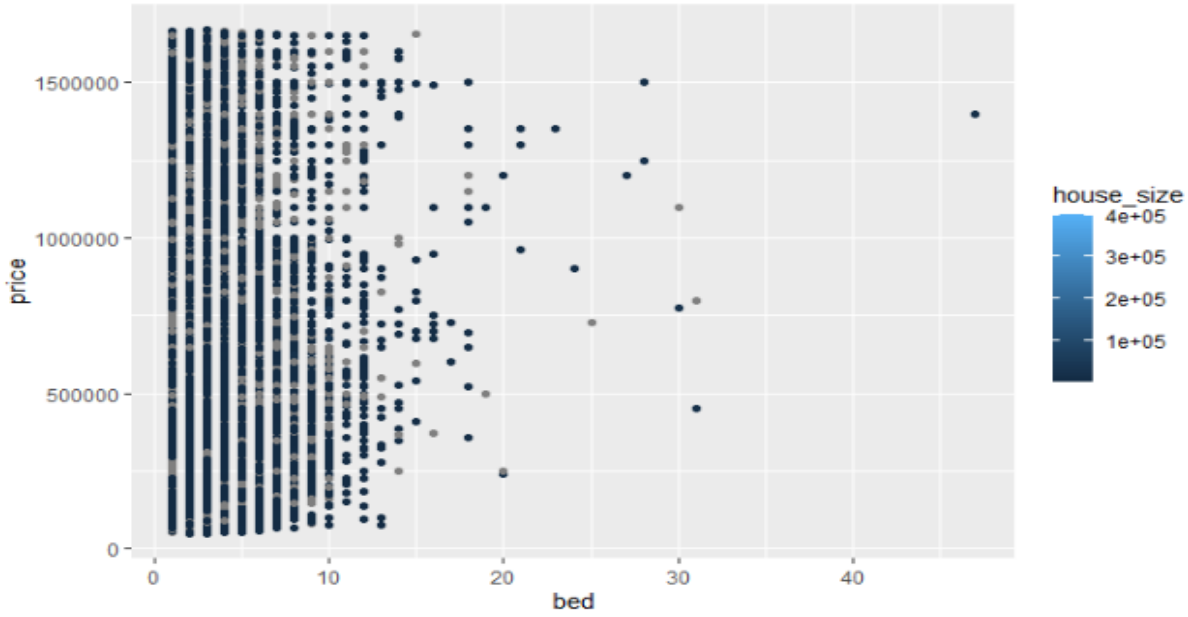
Warning: [38;5;232mRemoved 12635 rows containing missing values or values outside the scale range (`geom_point()`).[39m



Scatter plot of bed, price and house_size variables:

```{r}
ggplot(after_fewstates_removed,aes(x=bed,y=price,color=house_size))+geom_point()
```

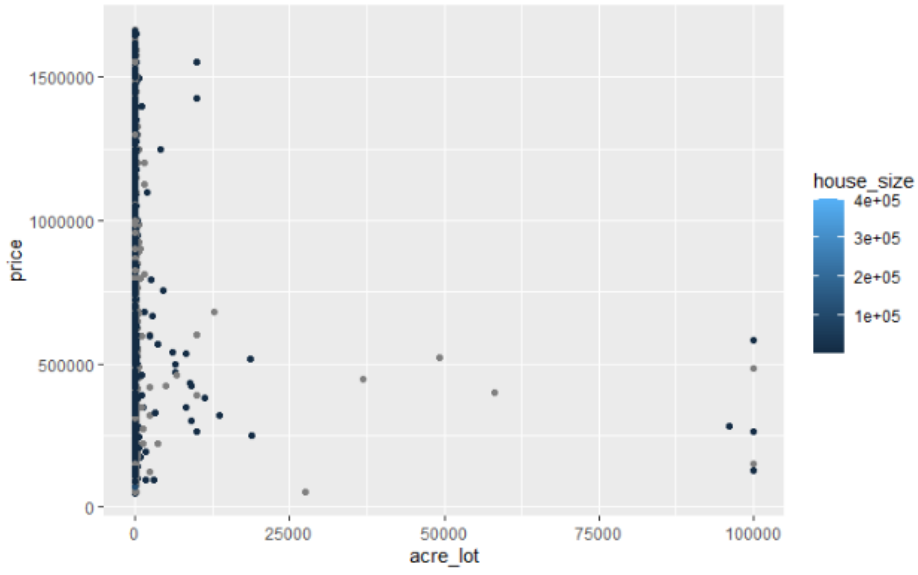
```
Warning: [38;5;232mRemoved 13764 rows containing missing values or values outside the scale range
(`geom_point()`).[39m
```



Scatter plot of acre_lot,price and house_size variables:

```{r}
ggplot(after_fewstates_removed,aes(x=acre_lot,y=price,color=house_size))+geom_point()
```

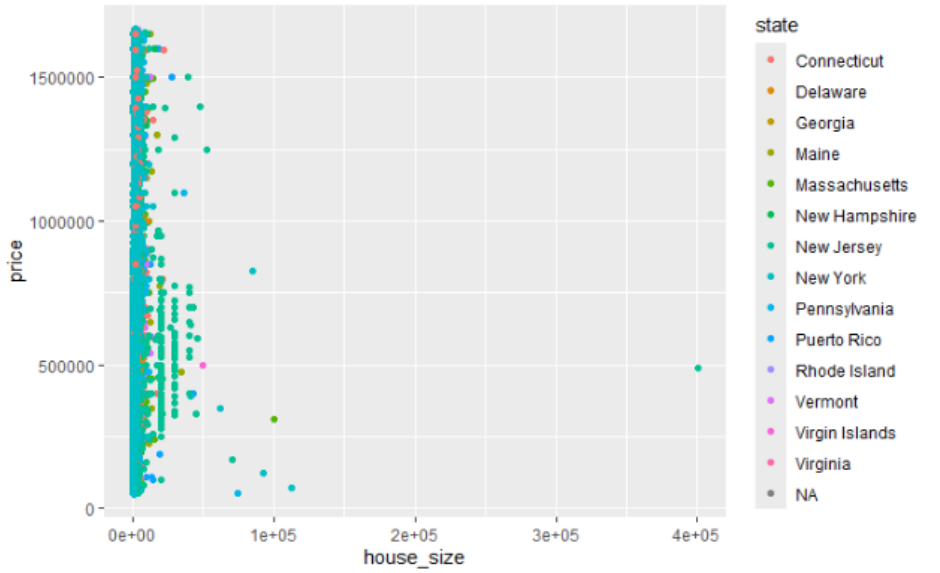
```
Warning: [38;5;232mRemoved 25305 rows containing missing values or values outside the scale range
(`geom_point()`).[39m
```



Scatter plot of house_size, state, price variables:

```r
ggplot(after_fewstates_removed,aes(x=house_size,y=price,color=state))+geom_point()
```

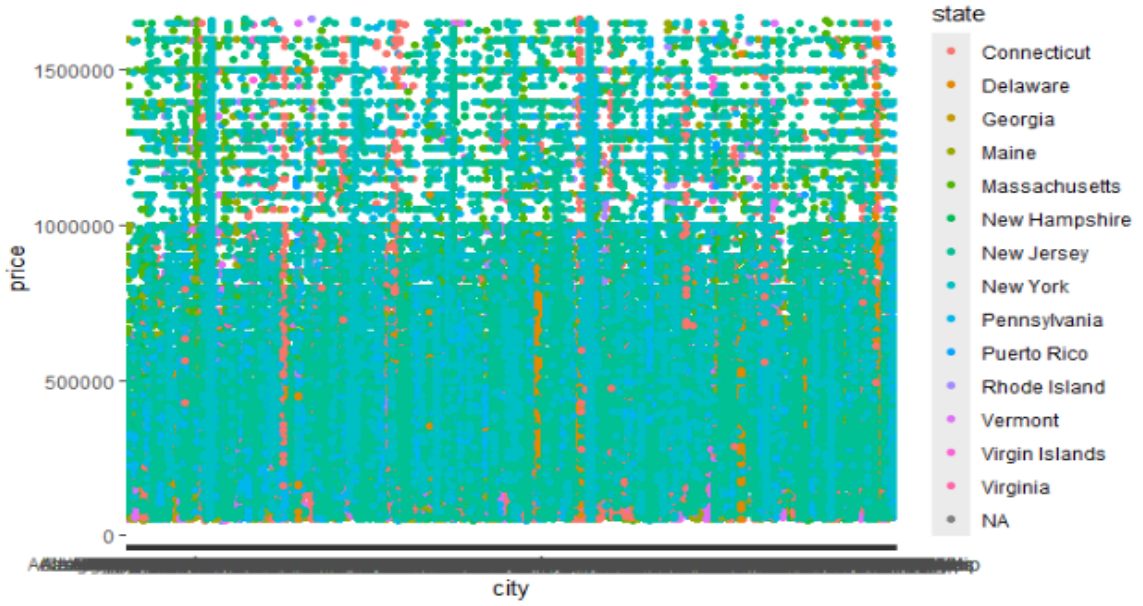
```
Warning: [38;5;232mRemoved 30640 rows containing missing values or values outside the scale range
(`geom_point()`).[39m
```



Scatterplot of city, state and price:

```{r}
ggplot(after_fewstates_removed,aes(x=city,y=price,color=state))+geom_point()
```

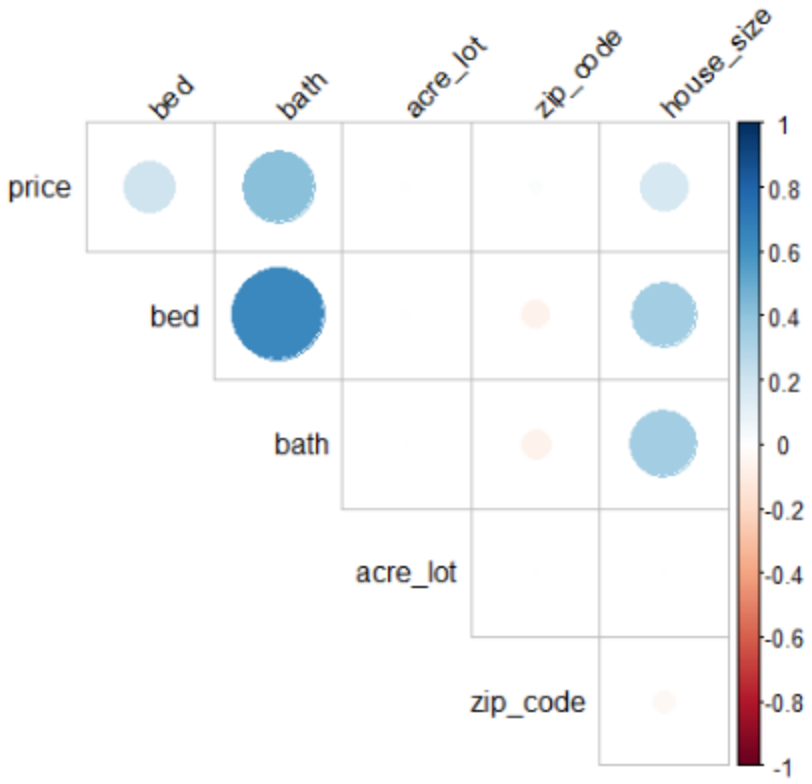
```
Warning: [38;5;232mRemoved 18 rows containing missing values or values outside the scale range
(`geom_point()`).[39m
```



Type2:

Correlation map between price and all numeric variables in dataset:

```{r}
# Load necessary library
library(dplyr)

# Select only numeric variables from the dataset
numeric_df <- after_fewstates_removed %>% select_if(is.numeric)

# Calculate correlation coefficients
correlation_matrix <- cor(numeric_df, use = "pairwise.complete.obs")

# Print the correlation matrix
print(correlation_matrix)
corrplot(correlation_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45, diag = FALSE)
```

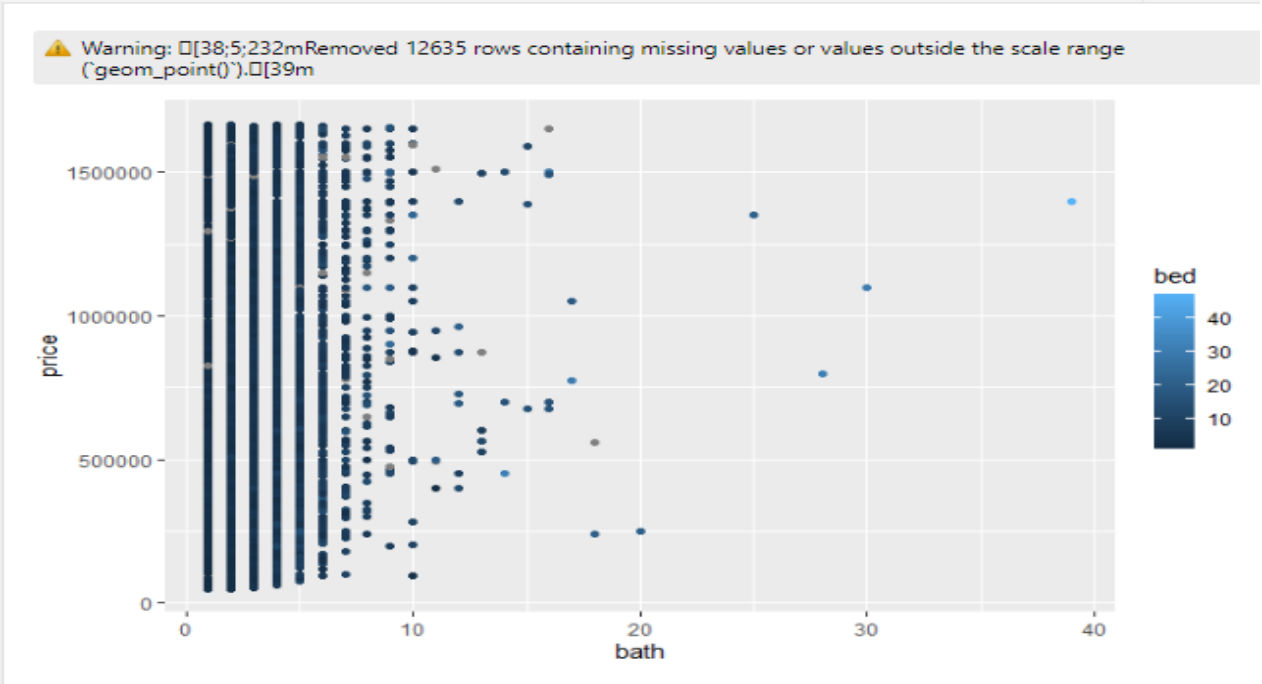


7. Discuss your observation results from the above graphs. (5 points)

   The number of bedrooms, bathrooms in a house and size of house impacts the price of a house slightly according to the data here when compared to other variables which are weakly correlated with the price of a house and the same is shown below:

```r
ggplot(after_fewstates_removed,aes(x=bath,y=price,color=bed))+geom_point()
```

```
Warning: [38;5;232mRemoved 12635 rows containing missing values or values outside the scale range
(`geom_point()`).[39m
```



Also from the above graphs, it is observed that New York, New jersey has more houses compared to all other states and priced in various ranges so that houses are available and affordable for all people. Apart from this, it is seen that houses of the same size are priced in wide range. Also, almost all the houses in all states are of the same size. There are more houses available in market which are valued under 1million dollars when compared to houses between 1 million and 1.5 million.

Price and zip code are weakly correlated with each other as per the visualization. But zip code and state/city/ street are multi-collinear, so it can be inferred that the location of house is associated with the price of house.

If the number of bathrooms and bedrooms is less than 10 in a house, then the price of such houses are available various prices range from the lowest to the highest. But if they are above 20 in number then they are available in price ranged above 5 lakh and became expensive.