DSC_424_HW1_SanketPatil

Sanket Praveen Patil

2024-01-27

Question 2:

Define the matrices and vectors

```
Z <- matrix(c(1, 1, 1, 1,
              -1, 1, 0, 3), nrow=4, byrow=FALSE)

Y <- matrix(c(0, 8, 0, 6), nrow=4, byrow=TRUE)

M <- matrix(c(2, 11, 0,
              1, 3, 40,
              4, 28, 73), nrow=3, byrow=FALSE)

N <- matrix(c(-4, 7, 9,
              -3, 2, 7,
              0, 1, -8), nrow=3, byrow=FALSE)

v <- matrix(c(-3, 39, 15), nrow=3)

w <- matrix(c(0, 10, 29), nrow=3)
```

a. v · w (dot product)

```
Question_a <- sum(v * w)
Question_a

## [1] 825
```

b. -3 * w

```
Question_b <- -3 * w
Question_b

##      [,1]
## [1,]    0
## [2,]  -30
## [3,]  -87
```

c. M * v

```
Question_c <- M %*% v
Question_c

##      [,1]
## [1,]   93
## [2,]  504
## [3,] 2655
```

d. M + N

```
Question_d <- M + N
Question_d

##      [,1] [,2] [,3]
## [1,]   -2   -2    4
## [2,]   18    5   29
## [3,]    9   47   65
```

e. M - N

```
Question_e <- M - N
Question_e

##      [,1] [,2] [,3]
## [1,]    6    4    4
## [2,]    4    1   27
## [3,]   -9   33   81
```

f. Z’Z

```
Question_f <- t(Z) %*% Z
Question_f

##      [,1] [,2]
## [1,]    4    3
## [2,]    3   11
```

g. (Z’Z)^-1 (inverse of Z’Z)

```
Question_g <- solve(Question_f)
Question_g

##             [,1]        [,2]
## [1,]  0.31428571 -0.08571429
## [2,] -0.08571429  0.11428571
```

h. Z'Y (matrix multiplication of Z transpose and Y)

```r
Question_h <- t(Z) %*% Y
Question_h

##      [,1]
## [1,]   14
## [2,]   26
```

i. β = (Z'Z)^-1 Z'Y

```r
Question_i <- Question_g %*% Question_h
Question_i

##          [,1]
## [1,] 2.171429
## [2,] 1.771429
```

j. det(Z'Z) (determinant of Z'Z)

```r
Question_j <- det(Question_f)
Question_j

## [1] 35
```

Question 3 - Ridge Regression

Author: Gary C. McDonald

Summary:

In his paper, McDonald (2009) talks about a method called ridge regression, which helps solve a common problem in statistics called collinearity. Collinearity happens when the variables in a regression model are too closely related, making it hard for the model to give accurate predictions. Ridge regression introduces a special parameter that helps balance things out, giving us better estimates even when the variables are strongly correlated. McDonald explains how this parameter works and why it's helpful in making our predictions more reliable.

He also talks about why ridge regression is useful. Sometimes, in real-life situations, we can't avoid having variables that are closely related. Ridge regression lets us handle these situations without throwing out important information. McDonald shows us how we can choose the right value for this special parameter to get the best results in our predictions. He uses examples to illustrate how ridge regression can be a powerful tool for statisticians when dealing with tricky data problems.

Overall, McDonald's paper helps us understand how ridge regression works and why it's important. By using this method, statisticians can improve their models and make more accurate predictions, even when faced with challenging data. McDonald's explanations make the complex topic of regression more accessible, showing us how statistical techniques can be applied to solve real-world problems.

Reference:

McDonald, G. C. (2009). Ridge regression. WIREs Computational Statistics, 1, 93–100. https://doi.org/10.1002/wics.14

Question 4 - Data Ethics or Data Integrity

Author: B.Y. Anom

Summary:

The article "Ethics of Big Data and Artificial Intelligence in Medicine" by B.Y. Anom, published in 2020, talks about how big data and artificial intelligence (AI) are changing healthcare. It explains how these technologies make healthcare easier and more efficient for both patients and doctors. Basically, big data means analyzing large amounts of information to find patterns and make predictions. AI is like teaching computers to think and learn like humans do.

The article also looks at the ethical issues that come with using big data and AI in healthcare. It says it's really important to protect patients' privacy when using these technologies. Sometimes, using big data and AI can go against traditional rules about how doctors should act. So, the article suggests ways to deal with these problems and says everyone involved in healthcare – doctors, people who study ethics, and governments – should work together to make rules that keep patients safe.

It's divided into three parts: first, it explains what big data and AI are and how they're used in healthcare. Then, it talks about the ethical problems these technologies bring up. Finally, it looks at how these technologies fit with the basic rules of medical ethics. Throughout the article, Anom gives examples of how big data and AI have changed healthcare and why it's important to think about ethics when using them.

In terms of being ethical with data, the article says it's really important to handle patients' information carefully. It warns that using big data and AI without thinking about ethics could lead to problems like invading patients' privacy, unfair decisions made by computer programs, and people not trusting the healthcare system anymore. So, the article suggests following clear rules to make sure patients' information is used safely and fairly. It says that by thinking about ethics, we can make sure big data and AI help patients without causing harm.

References:

Anom, B.Y. (2020). Ethics of Big Data and artificial intelligence in medicine. Journal of Ethics, Medicine and Public Health, 15, 100568. https://doi.org/10.1016/j.jemep.2020.100568

Question 5 - Indian housing data

Load necessary libraries

```
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 
2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── 
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all 
conflicts to become errors

library(dplyr)
library(modeest)
library(caret)

## Loading required package: lattice
## 
## Attaching package: 'caret'
## 
## The following object is masked from 'package:purrr':
## 
##     lift

library(fastDummies)

## Thank you for using fastDummies!
## To acknowledge our work, please cite the package:
## Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy 
(Binary) Columns and Rows from Categorical Variables. Version 1.7.1. URL: 
https://github.com/jacobkap/fastDummies, 
https://jacobkap.github.io/fastDummies/.
```

Read the CSV file

```
df <- 
read.csv("D:/Assignments_Depaul/DSC_424_Advance_Data_Analysis/HW1/indian_hous
ing_data.csv", header = TRUE)
head(df)

##   exactPrice sqftPrice securityDeposit          propertyType    postedOn
## 1     240000       171               9 Multistorey Apartment Jun 20, '23
```

```
## 2      12000        12          12000 Multistorey Apartment Jun 19, '23
## 3      17000         7              9     Residential House Jun 21, '23
## 4       5000         9              9     Residential House Jun 23, '23
## 5      12000         9          24000 Multistorey Apartment Jun 24, '23
## 6      18000        16              9 Multistorey Apartment Jun 24, '23
##   noOfLifts maintenanceChargesFrequency maintenanceCharges
## 1         9                           9                  9
## 2         1                     Monthly               1500
## 3         9                           9                  9
## 4         9                           9                  9
## 5         1                     Monthly                500
## 6         9                           9                  9
##                 locality     furnishing flrNum firstMonthCharges facing
## 1                Danapur Semi-Furnished      4                 9      9
## 2                      9 Semi-Furnished      4             25500      9
## 3 Phase 1 Ashiana Nagar Semi-Furnished Ground                 9      9
## 4                Kumhrar      Furnished      9                 9      9
## 5                Kumhrar    Unfurnished      1             36500   East
## 6             Lalji Tola    Unfurnished      1                 9  North
##   totalFlrNum  city carpetAreaUnit carpetArea brokerage bedrooms bathrooms
## 1           6 Patna              9          9         9        3         2
## 2           5 Patna          Sq-ft        900         9        2         2
## 3           2 Patna          Sq-ft       1300         9        3         3
## 4           3 Patna          Sq-ft        120         9        1         1
## 5           5 Patna          Sq-ft       1200         9        2         2
## 6           4 Patna          Sq-ft       1040         9        2         2
##   balconies Water_Storage Waste_Disposal Visitor_Parking Vaastu_Compliant
## 1         9             1              0               1                1
## 2         2             9              9               9                9
## 3         3             9              9               9                9
## 4         9             9              9               9                9
## 5         3             9              9               9                9
## 6         2             9              9               9                9
##                                                                                                         
URLs
## 1          https://www.magicbricks.com/propertyDetails/3-BHK-1407-Sq-ft-
Multistorey-Apartment-FOR-Rent-Danapur-in-Patna&id=4d423636393433373437
## 2                   https://www.magicbricks.com/propertyDetails/2-BHK-
980-Sq-ft-Multistorey-Apartment-FOR-Rent-in-Patna&id=4d423637363030303937
## 3 https://www.magicbricks.com/propertyDetails/3-BHK-2500-Sq-ft-
Residential-House-FOR-Rent-Phase-1-Ashiana-Nagar-in-
Patna&id=4d423637363333393731
## 4                 https://www.magicbricks.com/propertyDetails/1-BHK-120-Sq-
ft-Residential-House-FOR-Rent-Kumhrar-in-Patna&id=4d423637363638313337
## 5          https://www.magicbricks.com/propertyDetails/2-BHK-1200-Sq-ft-
Multistorey-Apartment-FOR-Rent-Kumhrar-in-Patna&id=4d423637363739323233
## 6       https://www.magicbricks.com/propertyDetails/2-BHK-1100-Sq-ft-
Multistorey-Apartment-FOR-Rent-Lalji-Tola-in-Patna&id=4d423631393339333635
##   Swimming_Pool Skydeck Service_Or_Goods_Lift Security
## 1             1       0                     0        1
```

```
## 2             9       9                     9        9
## 3             9       9                     9        9
## 4             9       9                     9        9
## 5             9       9                     9        9
## 6             9       9                     9        9
##   Retail_Boulevard___Retail_Shops__ Reserved_Parking
Rentable_Community_Space
## 1                                 0                1
0
## 2                                 9                9
9
## 3                                 9                9
9
## 4                                 9                9
9
## 5                                 9                9
9
## 6                                 9                9
9
##   RentOrSale Recreational_Pool Rain_Water_Harvesting RO_Water_System
## 1       Rent                 0                     1               0
## 2       Rent                 9                     9               9
## 3       Rent                 9                     9               9
## 4       Rent                 9                     9               9
## 5       Rent                 9                     9               9
## 6       Rent                 9                     9               9
##   Private_Terrace_Or_Garden Private_Garden Power_Back_Up Piped_Gas Park
## 1                         0              0             1         1    1
## 2                         9              9             9         9    9
## 3                         9              9             9         9    9
## 4                         9              9             9         9    9
## 5                         9              9             9         9    9
## 6                         9              9             9         9    9
##   Outdoor_Tennis_Courts Multipurpose_Hall Multipurpose_Courts
## 1                     1                 0                   0
## 2                     9                 9                   9
## 3                     9                 9                   9
## 4                     9                 9                   9
## 5                     9                 9                   9
## 6                     9                 9                   9
##   Mini_Cinema_Theatre Meditation_Area Maintenance_Staff     Long Lift
## 1                   0               0                 1 85.05633    0
## 2                   9               9                 9  9.00000    9
## 3                   9               9                 9 85.07996    9
## 4                   9               9                 9 85.18501    9
## 5                   9               9                 9 85.18501    9
## 6                   9               9                 9 85.14404    9
##   Library_And_Business_Centre Library Laundry_Service      Lat
## 1                           0       0               0 25.60590
## 2                           9       9               9  9.00000
```

```
## 3                          9       9              9 25.62143
## 4                          9       9              9 25.59309
## 5                          9       9              9 25.59309
## 6                          9       9              9 25.60508
##   Kids_Play_Pool_With_Water_Slides Kids_Play_Area Kids_Club
## 1                                0              0         0
## 2                                9              9         9
## 3                                9              9         9
## 4                                9              9         9
## 5                                9              9         9
## 6                                9              9         9
##   Jogging_and_Strolling_Track Internet_Or_Wi_Fi_Connectivity
Intercom_Facility
## 1                           1                              1
1
## 2                           9                              9
9
## 3                           9                              9
9
## 4                           9                              9
9
## 5                           9                              9
9
## 6                           9                              9
9
##   Indoor_Squash__And__Badminton_Courts Indoor_Games_Room
## 1                                    0                 0
## 2                                    9                 9
## 3                                    9                 9
## 4                                    9                 9
## 5                                    9                 9
## 6                                    9                 9
##   Health_club_with_Steam__Or__Jaccuzi Gymnasium Guest_Accommodation
## 1                                   0         1                   0
## 2                                   9         9                   9
## 3                                   9         9                   9
## 4                                   9         9                   9
## 5                                   9         9                   9
## 6                                   9         9                   9
##   Grand_Entrance_lobby Golf_Course Flower_Gardens Fire_Fighting_Equipment
## 1                    0           0              0                       0
## 2                    9           9              9                       9
## 3                    9           9              9                       9
## 4                    9           9              9                       9
## 5                    9           9              9                       9
## 6                    9           9              9                       9
##   Event_Space__And__Amphitheatre Earth_quake_resistant
Early_Learning_Centre
## 1                              0                     0
0
```

```
## 2                              9                      9
9
## 3                              9                      9
9
## 4                              9                      9
9
## 5                              9                      9
9
## 6                              9                      9
9
##   Dance_Studio DTH_Television_Facility Cycling__And__Jogging_Track
## 1            0                       1                           0
## 2            9                       9                           9
## 3            9                       9                           9
## 4            9                       9                           9
## 5            9                       9                           9
## 6            9                       9                           9
##   Cricket_net_practice Conference_Room Concierge_Services
## 1                    0               1                  0
## 2                    9               9                  9
## 3                    9               9                  9
## 4                    9               9                  9
## 5                    9               9                  9
## 6                    9               9                  9
##   Coffee_Lounge__And__Restaurants Club_House Canopy_Walk
## 1                               0          1           0
## 2                               9          9           9
## 3                               9          9           9
## 4                               9          9           9
## 5                               9          9           9
## 6                               9          9           9
##   Cafeteria_Or_Food_Court CCTV_Camera Barbeque_Pit Bar_Or_Lounge
Banquet_Hall
## 1                       1           0            0             1
1
## 2                       9           9            9             9
9
## 3                       9           9            9             9
9
## 4                       9           9            9             9
9
## 5                       9           9            9             9
9
## 6                       9           9            9             9
9
##   Bank__And__ATM Arts__And__Craft_Studio Air_Conditioned Activity_Deck4
## 1              0                       0               0              0
## 2              9                       9               9              9
## 3              9                       9               9              9
## 4              9                       9               9              9
```

```
## 5             9                        9               9              9
## 6             9                        9               9              9
##   AEROBICS_ROOM
## 1            0
## 2            9
## 3            9
## 4            9
## 5            9
## 6            9
```

Check basic information

```
str(df)

## 'data.frame':    27900 obs. of  91 variables:
##  $ exactPrice                       : num  240000 12000 17000 5000
12000 18000 8500 10000 11000 7000 ...
##  $ sqftPrice                        : int  171 12 7 9 9 16 7 8 9 12 ...
##  $ securityDeposit                  : int  9 12000 9 9 24000 9 9 20000
11000 7000 ...
##  $ propertyType                     : chr  "Multistorey Apartment"
"Multistorey Apartment" "Residential House" "Residential House" ...
##  $ postedOn                         : chr  "Jun 20, '23" "Jun 19, '23"
"Jun 21, '23" "Jun 23, '23" ...
##  $ noOfLifts                        : chr  "9" "1" "9" "9" ...
##  $ maintenanceChargesFrequency      : chr  "9" "Monthly" "9" "9" ...
##  $ maintenanceCharges               : num  9 1500 9 9 500 9 500 2000 9
9 ...
##  $ locality                         : chr  "Danapur" "9" "Phase 1
Ashiana Nagar" "Kumhrar" ...
##  $ furnishing                       : chr  "Semi-Furnished" "Semi-
Furnished" "Semi-Furnished" "Furnished" ...
##  $ flrNum                           : chr  "4" "4" "Ground" "9" ...
##  $ firstMonthCharges                : num  9 25500 9 9 36500 9 9000
32000 22000 14000 ...
##  $ facing                           : chr  "9" "9" "9" "9" ...
##  $ totalFlrNum                      : int  6 5 2 3 5 4 3 5 2 2 ...
##  $ city                             : chr  "Patna" "Patna" "Patna"
"Patna" ...
##  $ carpetAreaUnit                   : chr  "9" "Sq-ft" "Sq-ft" "Sq-ft"
...
##  $ carpetArea                       : int  9 900 1300 120 1200 1040
1000 930 1000 500 ...
##  $ brokerage                        : chr  "9" "9" "9" "9" ...
##  $ bedrooms                         : int  3 2 3 1 2 2 2 2 3 2 ...
##  $ bathrooms                        : int  2 2 3 1 2 2 1 2 1 1 ...
##  $ balconies                        : int  9 2 3 9 3 2 9 3 9 1 ...
##  $ Water_Storage                    : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Waste_Disposal                   : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Visitor_Parking                  : int  1 9 9 9 9 9 9 9 9 9 ...
```

```
##  $ Vaastu_Compliant                      : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ URLs                                  : chr
"https://www.magicbricks.com/propertyDetails/3-BHK-1407-Sq-ft-Multistorey-Apartment-FOR-Rent-Danapur-in-Patna&id"| __truncated__
"https://www.magicbricks.com/propertyDetails/2-BHK-980-Sq-ft-Multistorey-Apartment-FOR-Rent-in-Patna&id=4d423637363030303937"
"https://www.magicbricks.com/propertyDetails/3-BHK-2500-Sq-ft-Residential-House-FOR-Rent-Phase-1-Ashiana-Nagar-i"| __truncated__
"https://www.magicbricks.com/propertyDetails/1-BHK-120-Sq-ft-Residential-House-FOR-Rent-Kumhrar-in-Patna&id=4d42"| __truncated__ ...
##  $ Swimming_Pool                         : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Skydeck                               : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Service_Or_Goods_Lift                 : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Security                              : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Retail_Boulevard___Retail_Shops__     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Reserved_Parking                      : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Rentable_Community_Space              : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ RentOrSale                            : chr  "Rent" "Rent" "Rent" "Rent"
...
##  $ Recreational_Pool                     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Rain_Water_Harvesting                 : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ RO_Water_System                       : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Private_Terrace_Or_Garden             : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Private_Garden                        : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Power_Back_Up                         : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Piped_Gas                             : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Park                                  : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Outdoor_Tennis_Courts                 : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Multipurpose_Hall                     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Multipurpose_Courts                   : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Mini_Cinema_Theatre                   : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Meditation_Area                       : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Maintenance_Staff                     : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Long                                  : num  85.1 9 85.1 85.2 85.2 ...
##  $ Lift                                  : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Library_And_Business_Centre           : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Library                               : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Laundry_Service                       : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Lat                                   : num  25.6 9 25.6 25.6 25.6 ...
##  $ Kids_Play_Pool_With_Water_Slides      : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Kids_Play_Area                        : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Kids_Club                             : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Jogging_and_Strolling_Track           : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Internet_Or_Wi_Fi_Connectivity        : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Intercom_Facility                     : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Indoor_Squash__And__Badminton_Courts: int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Indoor_Games_Room                     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Health_club_with_Steam__Or__Jaccuzi : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Gymnasium                             : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Guest_Accommodation                   : int  0 9 9 9 9 9 9 9 9 9 ...
```

```
##  $ Grand_Entrance_lobby                : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Golf_Course                         : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Flower_Gardens                      : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Fire_Fighting_Equipment             : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Event_Space__And__Amphitheatre      : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Earth_quake_resistant               : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Early_Learning_Centre               : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Dance_Studio                        : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ DTH_Television_Facility             : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Cycling__And__Jogging_Track         : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Cricket_net_practice                : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Conference_Room                     : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Concierge_Services                  : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Coffee_Lounge__And__Restaurants     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Club_House                          : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Canopy_Walk                         : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Cafeteria_Or_Food_Court             : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ CCTV_Camera                         : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Barbeque_Pit                        : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Bar_Or_Lounge                       : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Banquet_Hall                        : int  1 9 9 9 9 9 9 9 9 9 ...
##  $ Bank__And__ATM                      : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Arts__And__Craft_Studio             : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Air_Conditioned                     : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ Activity_Deck4                      : int  0 9 9 9 9 9 9 9 9 9 ...
##  $ AEROBICS_ROOM                       : int  0 9 9 9 9 9 9 9 9 9 ...
```

Check summary of the variables

```
summary(df)
```

```
##    exactPrice          sqftPrice          securityDeposit    propertyType
##  Min.   :9.000e+00   Min.   :        0   Min.   :      1   Length:27900
##  1st Qu.:1.300e+04   1st Qu.:       11   1st Qu.:      9   Class
:character
##  Median :3.000e+04   Median :       21   Median :      9   Mode
:character
##  Mean   :5.428e+06   Mean   :    42933   Mean   :  24079
##  3rd Qu.:5.270e+06   3rd Qu.:     3864   3rd Qu.:  14000
##  Max.   :3.250e+09   Max.   :200000000   Max.   :5000000
##    postedOn           noOfLifts          maintenanceChargesFrequency
##  Length:27900       Length:27900       Length:27900
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  maintenanceCharges    locality           furnishing            flrNum
##  Min.   :0.000e+00   Length:27900       Length:27900       Length:27900
##  1st Qu.:9.000e+00   Class :character   Class :character   Class
```

```
:character
##  Median :9.000e+00   Mode  :character   Mode  :character   Mode
:character
##  Mean   :2.902e+05
##  3rd Qu.:9.000e+00
##  Max.   :8.076e+09
##  firstMonthCharges      facing           totalFlrNum         city
##  Min.   :9.000e+00   Length:27900       Min.   :  1.000   Length:27900
##  1st Qu.:9.000e+00   Class :character   1st Qu.:  2.000   Class :character
##  Median :9.000e+00   Mode  :character   Median :  4.000   Mode  :character
##  Mean   :3.328e+05                      Mean   :  5.666
##  3rd Qu.:3.000e+04                      3rd Qu.:  7.000
##  Max.   :8.077e+09                      Max.   :200.000
##  carpetAreaUnit       carpetArea     brokerage           bedrooms
##  Length:27900       Min.   :    1   Length:27900       Min.   : 1.000
##  Class :character   1st Qu.:    9   Class :character   1st Qu.: 2.000
##  Mode  :character   Median :  125   Mode  :character   Median : 2.000
##                     Mean   :  610                      Mean   : 2.673
##                     3rd Qu.: 1050                      3rd Qu.: 3.000
##                     Max.   :13000                      Max.   :10.000
##    bathrooms        balconies      Water_Storage   Waste_Disposal
##  Min.   : 1.000   Min.   : 1.000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:9.000   1st Qu.:9.000
##  Median : 2.000   Median : 3.000   Median :9.000   Median :9.000
##  Mean   : 2.483   Mean   : 4.677   Mean   :7.198   Mean   :7.184
##  3rd Qu.: 3.000   3rd Qu.: 9.000   3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :10.000   Max.   :10.000   Max.   :9.000   Max.   :9.000
##  Visitor_Parking Vaastu_Compliant     URLs           Swimming_Pool
##  Min.   :0.00    Min.   :0.000    Length:27900       Min.   :0.000
##  1st Qu.:9.00    1st Qu.:9.000    Class :character   1st Qu.:9.000
##  Median :9.00    Median :9.000    Mode  :character   Median :9.000
##  Mean   :7.21    Mean   :7.191                       Mean   :7.253
##  3rd Qu.:9.00    3rd Qu.:9.000                       3rd Qu.:9.000
##  Max.   :9.00    Max.   :9.000                       Max.   :9.000
##     Skydeck       Service_Or_Goods_Lift    Security
##  Min.   :0.000   Min.   :0.000         Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000         1st Qu.:9.000
##  Median :9.000   Median :9.000         Median :9.000
##  Mean   :7.122   Mean   :7.151         Mean   :7.287
##  3rd Qu.:9.000   3rd Qu.:9.000         3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000         Max.   :9.000
##  Retail_Boulevard___Retail_Shops__ Reserved_Parking
Rentable_Community_Space
##  Min.   :0.00                      Min.   :0.000    Min.   :0.00
##  1st Qu.:9.00                      1st Qu.:9.000    1st Qu.:9.00
##  Median :9.00                      Median :9.000    Median :9.00
##  Mean   :7.13                      Mean   :7.245    Mean   :7.13
##  3rd Qu.:9.00                      3rd Qu.:9.000    3rd Qu.:9.00
##  Max.   :9.00                      Max.   :9.000    Max.   :9.00
##   RentOrSale        Recreational_Pool Rain_Water_Harvesting
```

```
RO_Water_System
##  Length:27900       Min.   :0.000     Min.   :0.000         Min.   :0.000
##  Class :character   1st Qu.:9.000     1st Qu.:9.000         1st Qu.:9.000
##  Mode  :character   Median :9.000     Median :9.000         Median :9.000
##                     Mean   :7.128     Mean   :7.225         Mean   :7.146
##                     3rd Qu.:9.000     3rd Qu.:9.000         3rd Qu.:9.000
##                     Max.   :9.000     Max.   :9.000         Max.   :9.000
##  Private_Terrace_Or_Garden Private_Garden  Power_Back_Up     Piped_Gas
##  Min.   :0.000             Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000             1st Qu.:9.000   1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000             Median :9.000   Median :9.000   Median :9.000
##  Mean   :7.151             Mean   :7.116   Mean   :7.258   Mean   :7.155
##  3rd Qu.:9.000             3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000             Max.   :9.000   Max.   :9.000   Max.   :9.000
##       Park      Outdoor_Tennis_Courts Multipurpose_Hall
Multipurpose_Courts
##  Min.   :0.00   Min.   :0.000         Min.   :0.000     Min.   :0.000
##  1st Qu.:9.00   1st Qu.:9.000         1st Qu.:9.000     1st Qu.:9.000
##  Median :9.00   Median :9.000         Median :9.000     Median :9.000
##  Mean   :7.23   Mean   :7.166         Mean   :7.147     Mean   :7.151
##  3rd Qu.:9.00   3rd Qu.:9.000         3rd Qu.:9.000     3rd Qu.:9.000
##  Max.   :9.00   Max.   :9.000         Max.   :9.000     Max.   :9.000
##  Mini_Cinema_Theatre Meditation_Area Maintenance_Staff      Long
##  Min.   :0.000       Min.   :0.00    Min.   :0.000     Min.   : 0.00
##  1st Qu.:9.000       1st Qu.:9.00    1st Qu.:9.000     1st Qu.:75.69
##  Median :9.000       Median :9.00    Median :9.000     Median :77.44
##  Mean   :7.128       Mean   :7.17    Mean   :7.183     Mean   :72.86
##  3rd Qu.:9.000       3rd Qu.:9.00    3rd Qu.:9.000     3rd Qu.:80.85
##  Max.   :9.000       Max.   :9.00    Max.   :9.000     Max.   :91.29
##       Lift       Library_And_Business_Centre    Library
Laundry_Service
##  Min.   :0.000   Min.   :0.000              Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000              1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000   Median :9.000              Median :9.000   Median :9.000
##  Mean   :7.258   Mean   :7.139              Mean   :7.121   Mean   :7.145
##  3rd Qu.:9.000   3rd Qu.:9.000              3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000              Max.   :9.000   Max.   :9.000
##       Lat        Kids_Play_Pool_With_Water_Slides Kids_Play_Area
##  Min.   : 0.00   Min.   :0.000                    Min.   :0.000
##  1st Qu.:17.30   1st Qu.:9.000                    1st Qu.:9.000
##  Median :23.19   Median :9.000                    Median :9.000
##  Mean   :22.24   Mean   :7.139                    Mean   :7.224
##  3rd Qu.:26.91   3rd Qu.:9.000                    3rd Qu.:9.000
##  Max.   :85.06   Max.   :9.000                    Max.   :9.000
##    Kids_Club     Jogging_and_Strolling_Track
Internet_Or_Wi_Fi_Connectivity
##  Min.   :0.000   Min.   :0.000              Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000              1st Qu.:9.000
##  Median :9.000   Median :9.000              Median :9.000
##  Mean   :7.141   Mean   :7.186              Mean   :7.161
```

```
##  3rd Qu.:9.000   3rd Qu.:9.000               3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000               Max.   :9.000
##  Intercom_Facility Indoor_Squash__And__Badminton_Courts Indoor_Games_Room
##  Min.   :0.000     Min.   :0.000                        Min.   :0.000
##  1st Qu.:9.000     1st Qu.:9.000                        1st Qu.:9.000
##  Median :9.000     Median :9.000                        Median :9.000
##  Mean   :7.201     Mean   :7.167                        Mean   :7.205
##  3rd Qu.:9.000     3rd Qu.:9.000                        3rd Qu.:9.000
##  Max.   :9.000     Max.   :9.000                        Max.   :9.000
##  Health_club_with_Steam__Or__Jaccuzi   Gymnasium     Guest_Accommodation
##  Min.   :0.000                       Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000                       1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000                       Median :9.000   Median :9.000
##  Mean   :7.129                       Mean   :7.269   Mean   :7.128
##  3rd Qu.:9.000                       3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000                       Max.   :9.000   Max.   :9.000
##  Grand_Entrance_lobby  Golf_Course    Flower_Gardens 
Fire_Fighting_Equipment
##  Min.   :0.000        Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000        1st Qu.:9.000   1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000        Median :9.000   Median :9.000   Median :9.000
##  Mean   :7.129        Mean   :7.122   Mean   :7.176   Mean   :7.217
##  3rd Qu.:9.000        3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000        Max.   :9.000   Max.   :9.000   Max.   :9.000
##  Event_Space__And__Amphitheatre Earth_quake_resistant 
Early_Learning_Centre
##  Min.   :0.000                  Min.   :0.000         Min.   :0.000
##  1st Qu.:9.000                  1st Qu.:9.000         1st Qu.:9.000
##  Median :9.000                  Median :9.000         Median :9.000
##  Mean   :7.144                  Mean   :7.156         Mean   :7.126
##  3rd Qu.:9.000                  3rd Qu.:9.000         3rd Qu.:9.000
##  Max.   :9.000                  Max.   :9.000         Max.   :9.000
##   Dance_Studio   DTH_Television_Facility Cycling__And__Jogging_Track
##  Min.   :0.000   Min.   :0.000           Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000           1st Qu.:9.000
##  Median :9.000   Median :9.000           Median :9.000
##  Mean   :7.121   Mean   :7.157           Mean   :7.165
##  3rd Qu.:9.000   3rd Qu.:9.000           3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000           Max.   :9.000
##  Cricket_net_practice Conference_Room Concierge_Services
##  Min.   :0.000        Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000        1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000        Median :9.000   Median :9.000
##  Mean   :7.121        Mean   :7.144   Mean   :7.126
##  3rd Qu.:9.000        3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000        Max.   :9.000   Max.   :9.000
##  Coffee_Lounge__And__Restaurants   Club_House     Canopy_Walk
##  Min.   :0.000                   Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000                   1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000                   Median :9.000   Median :9.000
```

```
##  Mean   :7.127                          Mean   :7.243   Mean   :7.123
##  3rd Qu.:9.000                          3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000                          Max.   :9.000   Max.   :9.000
##  Cafeteria_Or_Food_Court  CCTV_Camera     Barbeque_Pit    Bar_Or_Lounge
##  Min.   :0.000           Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000           1st Qu.:9.000   1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000           Median :9.000   Median :9.000   Median :9.000
##  Mean   :7.153           Mean   :7.155   Mean   :7.123   Mean   :7.134
##  3rd Qu.:9.000           3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000           Max.   :9.000   Max.   :9.000   Max.   :9.000
##   Banquet_Hall   Bank__And__ATM  Arts__And__Craft_Studio Air_Conditioned
##  Min.   :0.000   Min.   :0.000   Min.   :0.000           Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000   1st Qu.:9.000           1st Qu.:9.000
##  Median :9.000   Median :9.000   Median :9.000           Median :9.000
##  Mean   :7.176   Mean   :7.135   Mean   :7.122           Mean   :7.142
##  3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:9.000           3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000           Max.   :9.000
##  Activity_Deck4  AEROBICS_ROOM
##  Min.   :0.000   Min.   :0.000
##  1st Qu.:9.000   1st Qu.:9.000
##  Median :9.000   Median :9.000
##  Mean   :7.127   Mean   :7.147
##  3rd Qu.:9.000   3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000
```

Checking the class of the columns

```
column_types <- sapply(df, class)
print(column_types)
```

```
##                              exactPrice                                sqftPrice
##                               "numeric"                                "integer"
##                         securityDeposit                             propertyType
##                               "integer"                              "character"
##                                postedOn                                noOfLifts
##                             "character"                              "character"
##              maintenanceChargesFrequency                       maintenanceCharges
##                             "character"                                "numeric"
##                                locality                                furnishing
##                             "character"                              "character"
##                                  flrNum                        firstMonthCharges
##                             "character"                                "numeric"
##                                  facing                              totalFlrNum
##                             "character"                                "integer"
##                                    city                           carpetAreaUnit
##                             "character"                              "character"
##                              carpetArea                                brokerage
##                               "integer"                              "character"
##                                bedrooms                                bathrooms
##                               "integer"                                "integer"
```

```
##                             balconies                         Water_Storage
##                             "integer"                             "integer"
##                        Waste_Disposal                       Visitor_Parking
##                             "integer"                             "integer"
##                      Vaastu_Compliant                                  URLs
##                             "integer"                           "character"
##                         Swimming_Pool                               Skydeck
##                             "integer"                             "integer"
##                Service_Or_Goods_Lift                              Security
##                             "integer"                             "integer"
##     Retail_Boulevard___Retail_Shops__                      Reserved_Parking
##                             "integer"                             "integer"
##              Rentable_Community_Space                            RentOrSale
##                             "integer"                           "character"
##                     Recreational_Pool                  Rain_Water_Harvesting
##                             "integer"                             "integer"
##                       RO_Water_System             Private_Terrace_Or_Garden
##                             "integer"                             "integer"
##                        Private_Garden                         Power_Back_Up
##                             "integer"                             "integer"
##                             Piped_Gas                                  Park
##                             "integer"                             "integer"
##                 Outdoor_Tennis_Courts                      Multipurpose_Hall
##                             "integer"                             "integer"
##                   Multipurpose_Courts                     Mini_Cinema_Theatre
##                             "integer"                             "integer"
##                       Meditation_Area                     Maintenance_Staff
##                             "integer"                             "integer"
##                                  Long                                  Lift
##                             "numeric"                             "integer"
##           Library_And_Business_Centre                               Library
##                             "integer"                             "integer"
##                       Laundry_Service                                   Lat
##                             "integer"                             "numeric"
##      Kids_Play_Pool_With_Water_Slides                        Kids_Play_Area
##                             "integer"                             "integer"
##                             Kids_Club            Jogging_and_Strolling_Track
##                             "integer"                             "integer"
##        Internet_Or_Wi_Fi_Connectivity                      Intercom_Facility
##                             "integer"                             "integer"
## Indoor_Squash__And__Badminton_Courts                      Indoor_Games_Room
##                             "integer"                             "integer"
##   Health_club_with_Steam__Or__Jaccuzi                              Gymnasium
##                             "integer"                             "integer"
##                   Guest_Accommodation                   Grand_Entrance_lobby
##                             "integer"                             "integer"
##                           Golf_Course                        Flower_Gardens
##                             "integer"                             "integer"
##               Fire_Fighting_Equipment          Event_Space__And__Amphitheatre
##                             "integer"                             "integer"
```

```
##                Earth_quake_resistant                 Early_Learning_Centre
##                            "integer"                             "integer"
##                         Dance_Studio               DTH_Television_Facility
##                            "integer"                             "integer"
##          Cycling__And__Jogging_Track                Cricket_net_practice
##                            "integer"                             "integer"
##                      Conference_Room                    Concierge_Services
##                            "integer"                             "integer"
##       Coffee_Lounge__And__Restaurants                          Club_House
##                            "integer"                             "integer"
##                          Canopy_Walk               Cafeteria_Or_Food_Court
##                            "integer"                             "integer"
##                          CCTV_Camera                         Barbeque_Pit
##                            "integer"                             "integer"
##                        Bar_Or_Lounge                          Banquet_Hall
##                            "integer"                             "integer"
##                       Bank__And__ATM               Arts__And__Craft_Studio
##                            "integer"                             "integer"
##                      Air_Conditioned                       Activity_Deck4
##                            "integer"                             "integer"
##                        AEROBICS_ROOM
##                            "integer"
```

Count the number of categorical and numerical variables

```
num_categorical <- sum(column_types == "factor" | column_types ==
"character")
num_numerical <- sum(column_types == "numeric" | column_types == "integer")
```

Print the results

```
cat("Number of Categorical Variables:", num_categorical, "\n")

## Number of Categorical Variables: 13

cat("Number of Numerical Variables:", num_numerical, "\n")

## Number of Numerical Variables: 78
```

Checking the columns with unique counts in the dataframe

```
unique_counts <- sapply(df, function(x) length(unique(x)))
```

Display the number of unique values for each column

```
print(unique_counts)

##                           exactPrice                             sqftPrice
##                                 2018                                  4916
```

```
##                securityDeposit                   propertyType
##                            314                              7
##                       postedOn                      noOfLifts
##                            174                             11
##   maintenanceChargesFrequency              maintenanceCharges
##                              6                            257
##                       locality                     furnishing
##                           3831                              4
##                         flrNum               firstMonthCharges
##                             62                           1947
##                         facing                    totalFlrNum
##                              9                             81
##                           city                  carpetAreaUnit
##                             20                             10
##                     carpetArea                      brokerage
##                           1596                             19
##                       bedrooms                      bathrooms
##                             10                             10
##                      balconies                  Water_Storage
##                             10                              3
##                 Waste_Disposal                Visitor_Parking
##                              3                              3
##               Vaastu_Compliant                           URLs
##                              3                          27870
##                  Swimming_Pool                        Skydeck
##                              3                              3
##           Service_Or_Goods_Lift                       Security
##                              3                              3
##  Retail_Boulevard___Retail_Shops__             Reserved_Parking
##                              3                              3
##       Rentable_Community_Space                     RentOrSale
##                              3                              3
##              Recreational_Pool           Rain_Water_Harvesting
##                              3                              3
##                RO_Water_System      Private_Terrace_Or_Garden
##                              3                              3
##                 Private_Garden                   Power_Back_Up
##                              3                              3
##                      Piped_Gas                           Park
##                              3                              3
##          Outdoor_Tennis_Courts               Multipurpose_Hall
##                              3                              3
##            Multipurpose_Courts              Mini_Cinema_Theatre
##                              3                              3
##                Meditation_Area               Maintenance_Staff
##                              3                              3
##                           Long                           Lift
##                           7059                              3
##     Library_And_Business_Centre                        Library
##                              3                              3
```

```
##                     Laundry_Service                                 Lat
##                                   3                                7099
##     Kids_Play_Pool_With_Water_Slides                      Kids_Play_Area
##                                   3                                   3
##                           Kids_Club          Jogging_and_Strolling_Track
##                                   3                                   3
##       Internet_Or_Wi_Fi_Connectivity                  Intercom_Facility
##                                   3                                   3
## Indoor_Squash__And__Badminton_Courts                 Indoor_Games_Room
##                                   3                                   3
##  Health_club_with_Steam__Or__Jaccuzi                         Gymnasium
##                                   3                                   3
##                 Guest_Accommodation               Grand_Entrance_lobby
##                                   3                                   3
##                         Golf_Course                     Flower_Gardens
##                                   3                                   3
##             Fire_Fighting_Equipment     Event_Space__And__Amphitheatre
##                                   3                                   3
##               Earth_quake_resistant               Early_Learning_Centre
##                                   3                                   3
##                        Dance_Studio             DTH_Television_Facility
##                                   3                                   3
##          Cycling__And__Jogging_Track              Cricket_net_practice
##                                   3                                   3
##                     Conference_Room                 Concierge_Services
##                                   3                                   3
##       Coffee_Lounge__And__Restaurants                        Club_House
##                                   3                                   3
##                         Canopy_Walk            Cafeteria_Or_Food_Court
##                                   3                                   3
##                         CCTV_Camera                       Barbeque_Pit
##                                   3                                   3
##                       Bar_Or_Lounge                       Banquet_Hall
##                                   3                                   3
##                      Bank__And__ATM           Arts__And__Craft_Studio
##                                   3                                   3
##                     Air_Conditioned                     Activity_Deck4
##                                   3                                   3
##                       AEROBICS_ROOM
##                                   3
```

Delete column with Unique values

```
df <- df[, !colnames(df) %in% "URLs"]
```

Delete column with dates

```
df <- df[, !colnames(df) %in% "postedOn"]
```

Remove locality as it will not help and will cause issue in creating dummy variables

```
df <- df[, !colnames(df) %in% "locality"]
```

Checking the % of rows with value = 9. We are treating them as NA as given in the problem statement

```
percentage_rows_exact_Value_9 <- colMeans(df == 9, na.rm = TRUE) * 100
print(percentage_rows_exact_Value_9)
```

```
##                     exactPrice                          sqftPrice
##                     4.32258065                        13.18637993
##                securityDeposit                       propertyType
##                    69.72043011                         0.04301075
##                      noOfLifts         maintenanceChargesFrequency
##                    82.34408602                        76.36559140
##             maintenanceCharges                         furnishing
##                    83.82078853                         2.43369176
##                         flrNum                 firstMonthCharges
##                    24.03584229                        68.19354839
##                         facing                        totalFlrNum
##                    52.44802867                         7.99641577
##                           city                      carpetAreaUnit
##                     0.04301075                        45.86021505
##                     carpetArea                          brokerage
##                    46.33333333                        79.26523297
##                       bedrooms                          bathrooms
##                     2.11469534                         2.61290323
##                      balconies                      Water_Storage
##                    38.83154122                        79.06810036
##                 Waste_Disposal                    Visitor_Parking
##                    79.06810036                        79.06810036
##                Vaastu_Compliant                      Swimming_Pool
##                    79.06810036                        79.06810036
##                        Skydeck                Service_Or_Goods_Lift
##                    79.06810036                        79.06810036
##                       Security      Retail_Boulevard___Retail_Shops__
##                    79.06810036                        79.06810036
##               Reserved_Parking           Rentable_Community_Space
##                    79.06810036                        79.06810036
##                     RentOrSale                   Recreational_Pool
##                     0.11111111                        79.06810036
##           Rain_Water_Harvesting                    RO_Water_System
##                    79.06810036                        79.06810036
##         Private_Terrace_Or_Garden                    Private_Garden
##                    79.06810036                        79.06810036
##                   Power_Back_Up                          Piped_Gas
##                    79.06810036                        79.06810036
```

```
##                                Park                Outdoor_Tennis_Courts
##                         79.06810036                          79.06810036
##                  Multipurpose_Hall                  Multipurpose_Courts
##                         79.06810036                          79.06810036
##                 Mini_Cinema_Theatre                      Meditation_Area
##                         79.06810036                          79.06810036
##                   Maintenance_Staff                                 Long
##                         79.06810036                           7.77419355
##                                Lift         Library_And_Business_Centre
##                         79.06810036                          79.06810036
##                             Library                      Laundry_Service
##                         79.06810036                          79.06810036
##                                 Lat     Kids_Play_Pool_With_Water_Slides
##                          7.77419355                          79.06810036
##                      Kids_Play_Area                            Kids_Club
##                         79.06810036                          79.06810036
##          Jogging_and_Strolling_Track        Internet_Or_Wi_Fi_Connectivity
##                         79.06810036                          79.06810036
##                   Intercom_Facility Indoor_Squash__And__Badminton_Courts
##                         79.06810036                          79.06810036
##                   Indoor_Games_Room  Health_club_with_Steam__Or__Jaccuzi
##                         79.06810036                          79.06810036
##                           Gymnasium                  Guest_Accommodation
##                         79.06810036                          79.06810036
##                Grand_Entrance_lobby                          Golf_Course
##                         79.06810036                          79.06810036
##                      Flower_Gardens               Fire_Fighting_Equipment
##                         79.06810036                          79.06810036
##      Event_Space__And__Amphitheatre                Earth_quake_resistant
##                         79.06810036                          79.06810036
##               Early_Learning_Centre                         Dance_Studio
##                         79.06810036                          79.06810036
##             DTH_Television_Facility          Cycling__And__Jogging_Track
##                         79.06810036                          79.06810036
##                Cricket_net_practice                      Conference_Room
##                         79.06810036                          79.06810036
##                  Concierge_Services        Coffee_Lounge__And__Restaurants
##                         79.06810036                          79.06810036
##                          Club_House                          Canopy_Walk
##                         79.06810036                          79.06810036
##             Cafeteria_Or_Food_Court                          CCTV_Camera
##                         79.06810036                          79.06810036
##                        Barbeque_Pit                         Bar_Or_Lounge
##                         79.06810036                          79.06810036
##                        Banquet_Hall                        Bank__And__ATM
##                         79.06810036                          79.06810036
##             Arts__And__Craft_Studio                      Air_Conditioned
##                         79.06810036                          79.06810036
##                       Activity_Deck4                        AEROBICS_ROOM
##                         79.06810036                          79.06810036
```

Find columns where percentage is greater than 60%

```
columns_greater_than_70_percent <-
names(percentage_rows_exact_Value_9[percentage_rows_exact_Value_9 > 70])
```

Print or use the column names as needed

```
cat("Columns where more than 70% of values are 9:\n")

## Columns where more than 70% of values are 9:

print(columns_greater_than_70_percent)

##  [1] "noOfLifts"
##  [2] "maintenanceChargesFrequency"
##  [3] "maintenanceCharges"
##  [4] "brokerage"
##  [5] "Water_Storage"
##  [6] "Waste_Disposal"
##  [7] "Visitor_Parking"
##  [8] "Vaastu_Compliant"
##  [9] "Swimming_Pool"
## [10] "Skydeck"
## [11] "Service_Or_Goods_Lift"
## [12] "Security"
## [13] "Retail_Boulevard___Retail_Shops__"
## [14] "Reserved_Parking"
## [15] "Rentable_Community_Space"
## [16] "Recreational_Pool"
## [17] "Rain_Water_Harvesting"
## [18] "RO_Water_System"
## [19] "Private_Terrace_Or_Garden"
## [20] "Private_Garden"
## [21] "Power_Back_Up"
## [22] "Piped_Gas"
## [23] "Park"
## [24] "Outdoor_Tennis_Courts"
## [25] "Multipurpose_Hall"
## [26] "Multipurpose_Courts"
## [27] "Mini_Cinema_Theatre"
## [28] "Meditation_Area"
## [29] "Maintenance_Staff"
## [30] "Lift"
## [31] "Library_And_Business_Centre"
## [32] "Library"
## [33] "Laundry_Service"
## [34] "Kids_Play_Pool_With_Water_Slides"
## [35] "Kids_Play_Area"
## [36] "Kids_Club"
## [37] "Jogging_and_Strolling_Track"
## [38] "Internet_Or_Wi_Fi_Connectivity"
```

```
## [39] "Intercom_Facility"
## [40] "Indoor_Squash__And__Badminton_Courts"
## [41] "Indoor_Games_Room"
## [42] "Health_club_with_Steam__Or__Jaccuzi"
## [43] "Gymnasium"
## [44] "Guest_Accommodation"
## [45] "Grand_Entrance_lobby"
## [46] "Golf_Course"
## [47] "Flower_Gardens"
## [48] "Fire_Fighting_Equipment"
## [49] "Event_Space__And__Amphitheatre"
## [50] "Earth_quake_resistant"
## [51] "Early_Learning_Centre"
## [52] "Dance_Studio"
## [53] "DTH_Television_Facility"
## [54] "Cycling__And__Jogging_Track"
## [55] "Cricket_net_practice"
## [56] "Conference_Room"
## [57] "Concierge_Services"
## [58] "Coffee_Lounge__And__Restaurants"
## [59] "Club_House"
## [60] "Canopy_Walk"
## [61] "Cafeteria_Or_Food_Court"
## [62] "CCTV_Camera"
## [63] "Barbeque_Pit"
## [64] "Bar_Or_Lounge"
## [65] "Banquet_Hall"
## [66] "Bank__And__ATM"
## [67] "Arts__And__Craft_Studio"
## [68] "Air_Conditioned"
## [69] "Activity_Deck4"
## [70] "AEROBICS_ROOM"
```

Remove columns from df

```
df <- df[, !(names(df) %in% columns_greater_than_70_percent)]

dim(df)

## [1] 27900    18
```

Check for NA values column wise

```
na_percentages <- colMeans(is.na(df)) * 100
na_percentages

##        exactPrice         sqftPrice   securityDeposit      propertyType
##                 0                 0                 0                 0
##        furnishing            flrNum firstMonthCharges            facing
```

```
##                 0                 0                 0                 0
##       totalFlrNum              city     carpetAreaUnit        carpetArea
##                 0                 0                 0                 0
##          bedrooms         bathrooms         balconies        RentOrSale
##                 0                 0                 0                 0
##              Long               Lat
##                 0                 0
```

Calculate the percentage of rows affected by NA

```
percentage_na_rows <- mean(apply(df, 1, function(row) any(is.na(row)))) * 100
print(percentage_na_rows)

## [1] 0
```

Checking unique values

```
sapply(df, function(x) length(unique(x)))

##        exactPrice         sqftPrice   securityDeposit      propertyType
##              2018              4916               314                 7
##        furnishing            flrNum firstMonthCharges            facing
##                 4                62              1947                 9
##       totalFlrNum              city     carpetAreaUnit        carpetArea
##                81                20                10              1596
##          bedrooms         bathrooms         balconies        RentOrSale
##                10                10                10                 3
##              Long               Lat
##              7059              7099
```

—-—- Handeling carpet area problem as it has different units ———-

Checking unique values first

```
unique(df$carpetAreaUnit)

##  [1] "9"      "Sq-ft"  "Kanal"  "Marla"  "Sq-yrd" "Biswa1" "Sq-m"   "Rood"
##  [9] "Biswa2" "Acre"
```

Converting other units to sqft

```
convert_to_sqft_and_replace <- function(df) {
   convert <- tibble(
    unit = c("Sq_ft", "Kanal", "Marla", "Sq_yrd", "Biswa1", "Sq_m", "Rood",
"Biswa2", "Acre"),
    factor = c(1, 5445, 272.25, 9, 1350, 10.764, 10890, 2700, 43560)
  )
```

```r
  df$carpetArea <- mapply(function(area, unit) {
    if (unit %in% names(convert) && unit != "Sq-ft" && unit != "9") {
      return(area * convert[unit])
    } else {
      return(area)
    }
  }, df$carpetArea, df$carpetAreaUnit)

  # Return the modified data frame
  return(df)
}
```

Converting the other units to sqft

```r
df <- convert_to_sqft_and_replace(df)
```

Now we don't need carpetAreaUnit column as we have converted all values to sq-ft.

Delete column carpetAreaUnit

```r
df <- df[, !colnames(df) %in% "carpetAreaUnit"]
```

We have a column called flrNum which has some categorical values. We will convert them to numeric

```r
unique(df$flrNum)

##  [1] "4"              "Ground"         "9"              "1"
##  [5] "2"              "3"              "5"              "6"
##  [9] "7"              "13"             "8"              "14"
## [13] "10"             "12"             "Upper Basement" "Lower Basement"
## [17] "11"             "15"             "16"             "24"
## [21] "21"             "17"             "27"             "19"
## [25] "29"             "23"             "18"             "28"
## [29] "20"             "30"             "22"             "25"
## [33] "39"             "33"             "26"             "31"
## [37] "34"             "32"             "35"             "36"
## [41] "40"             "38"             "70"             "37"
## [45] "58"             "56"             "45"             "42"
## [49] "50"             "53"             "43"             "41"
## [53] "47"             "46"             "61"             "44"
## [57] "54"             "60"             "65"             "66"
## [61] "55"             "63"
```

Assigning Ground as 0, Lower basement as -2 and Upper Basement as -1

```
df <- df %>%
  mutate(flrNum = case_when(
    flrNum == "Ground" ~ 0,
    flrNum == "Upper Basement" ~ -1,
    flrNum == "Lower Basement" ~ -2,
    TRUE ~ as.numeric(flrNum)
  ))

## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `flrNum = case_when(...)`.
## Caused by warning:
## ! NAs introduced by coercion
```

Converting whole column to numeric

```
df$flrNum <- as.numeric(df$flrNum)
```

replacing flrNum=9 by 4 as 4 is mean of flrNum column

```
df <- df %>%
  mutate(flrNum = case_when(
    flrNum == 9 ~ 4, TRUE ~ as.numeric(flrNum)))
```

Converting some of the columns as factors as they will not provide any useful information with numeric as they have less number if unique values.

Hence we can treat those columns as factors to find non linear relations between them

```
df$bedrooms <- as.factor(df$bedrooms)
df$bathrooms <- as.factor(df$bathrooms)
df$balconies <- as.factor(df$balconies)
```

———————— Treating value = 9 as NA which is provided in the problem statement and replacing them ————-

Checking % of value=9 column wise

```
colMeans(df == 9, na.rm = TRUE) * 100
```

```
##        exactPrice          sqftPrice   securityDeposit       propertyType
##        4.32258065        13.18637993       69.72043011         0.04301075
##        furnishing             flrNum firstMonthCharges             facing
##        2.43369176         0.00000000       68.19354839        52.44802867
##       totalFlrNum               city         carpetArea           bedrooms
##        7.99641577         0.04301075       46.33333333         2.11469534
##         bathrooms          balconies        RentOrSale               Long
##        2.61290323        38.83154122        0.11111111         7.77419355
##               Lat
##        7.77419355
```

Creating a function to replace NA values

```r
replace_9_with_mean_or_mode <- function(df) {
  for (col in names(df)) {

    is_numeric <- is.numeric(df[[col]])
    is_character <- is.character(df[[col]])
    is_factor <- is.factor(df[[col]])

    # Replace 9 with mean for numeric columns
    if (is_numeric) {
      df[[col]][df[[col]] == 9] <- mean(df[[col]][df[[col]] != 9], na.rm =
TRUE)
    }
    # Replace 9 with mode for character and factor columns
    else if (is_character || is_factor) {
      mode_val <- as.character(names(sort(table(df[[col]][df[[col]] != 9]),
decreasing = TRUE)[1]))
      df[[col]][df[[col]] == 9] <- mode_val
    }
  }

  return(df)
}
```

Replace the values now

```r
df <- replace_9_with_mean_or_mode(df)
```

Checking if the values are replaced or not

```r
colMeans(df == 9, na.rm = TRUE) * 100
```

```
##        exactPrice          sqftPrice   securityDeposit       propertyType
##                 0                  0                 0                  0
##        furnishing             flrNum firstMonthCharges             facing
##                 0                  0                 0                  0
##       totalFlrNum               city         carpetArea           bedrooms
```

```
##                 0                 0                 0                 0
##         bathrooms         balconies        RentOrSale              Long
##                 0                 0                 0                 0
##               Lat
##                 0
```

Check for NA values generated during preprocessing

```r
na_percentages <- colMeans(is.na(df)) * 100
na_percentages
```

```
##        exactPrice         sqftPrice   securityDeposit      propertyType
##                 0                 0                 0                 0
##        furnishing            flrNum firstMonthCharges            facing
##                 0                 0                 0                 0
##       totalFlrNum              city        carpetArea          bedrooms
##                 0                 0                 0                 0
##         bathrooms         balconies        RentOrSale              Long
##                 0                 0                 0                 0
##               Lat
##                 0
```

Calculate the percentage of rows affected by NA

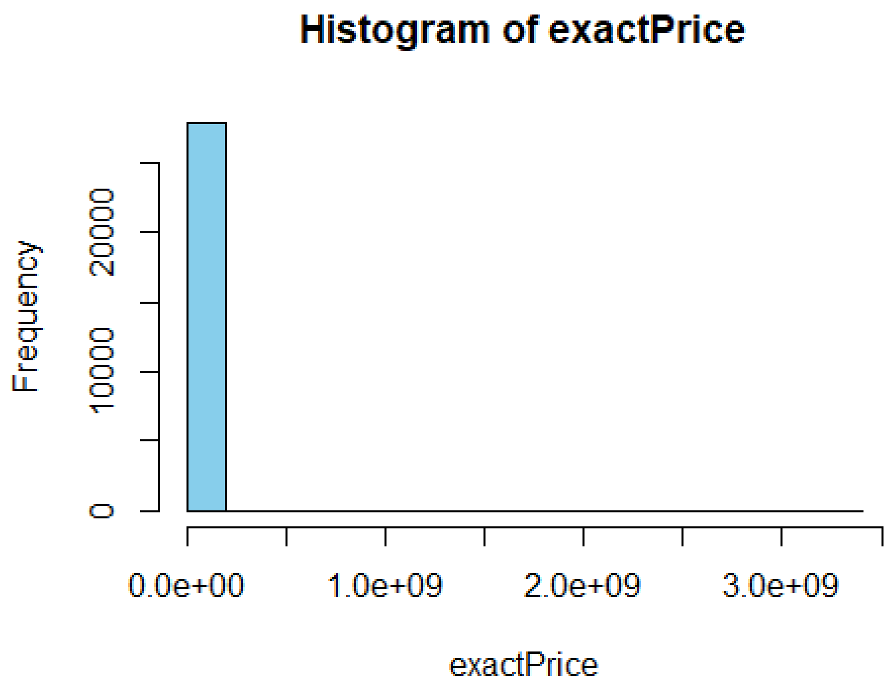
```r
percentage_na_rows <- mean(apply(df, 1, function(row) any(is.na(row)))) * 100
print(percentage_na_rows)
```

```
## [1] 0
```

———————————- Target variable Analysis ——————————————

Plotting histogram of target variable

```r
hist(df$exactPrice, main = "Histogram of exactPrice", xlab = "exactPrice",
col = "skyblue", border = "black")
```




```
exactPrice_skewness <- skewness(df$exactPrice)

cat("Skewness of exactPrice:", exactPrice_skewness, "\n")

## Skewness of exactPrice: 68.05725
```

Target variable is highly skewed hence we need handle this.

Finding outliers

Calculate Z-scores

```
z_scores <- scale(df$exactPrice)
```

Set a threshold (e.g., 3 or -3)

```
threshold <- 3
```

Identify outliers

```
outliers <- which(abs(z_scores) > threshold)
```

Print the indices of outliers

```r
cat("Indices of outliers in exactPrice:", outliers, "\n")
```

```
## Indices of outliers in exactPrice: 4196 12208 12890 12950 13059 13340
13386 13474 13487 13498 13777 13930 13957 14156 14315 14494 14496 14509 14846
15058 15816 16483 17017 17532 17693 18099 18133 18729 18866 19486 19782 19791
20059 20060 20063 20064 20070 20111 20113 20115 20155 20161 20208 20290 20294
20300 20305 20309 20468 20469 20606 20610 20614 20628 20753 20766 20769 20777
20880 20925 20928 20933 20971 20977 20978 21185 21224 21251 21252 21459 21476
21477 21549 21553 21611 21613 21618 21619 21655 21656 21669 21725 21761 21775
21867 21917 21960 22014 22103 22125 22165 22239 22297 22323 22332 22335 22342
22380 22382 22383 22389 22503 22546 22547 22602 22611 22616 22728 22839 22865
22888 23008 23010 23126 23127 23137 23314 23337 23431 23434 23436 23439 23609
23612 23614 23621 23622 23623 23719 23721 23802 23830 23855 23951 23961 24036
24083 24111 24130 24184 24194 24236 24264 24309 24367 24372 24414 24447 24515
24523 24599 24675 24747 24749 24763 24768 24772 24812 24813 24824 24874 24903
24913 24953 24954 25035 25045 25064 25095 25130 25133 25134 25287 25293 25300
25305 26061 26172 26360 27072
```

Print the values of outliers

```r
cat("Values of outliers in exactPrice:", df$exactPrice[outliers], "\n")
```

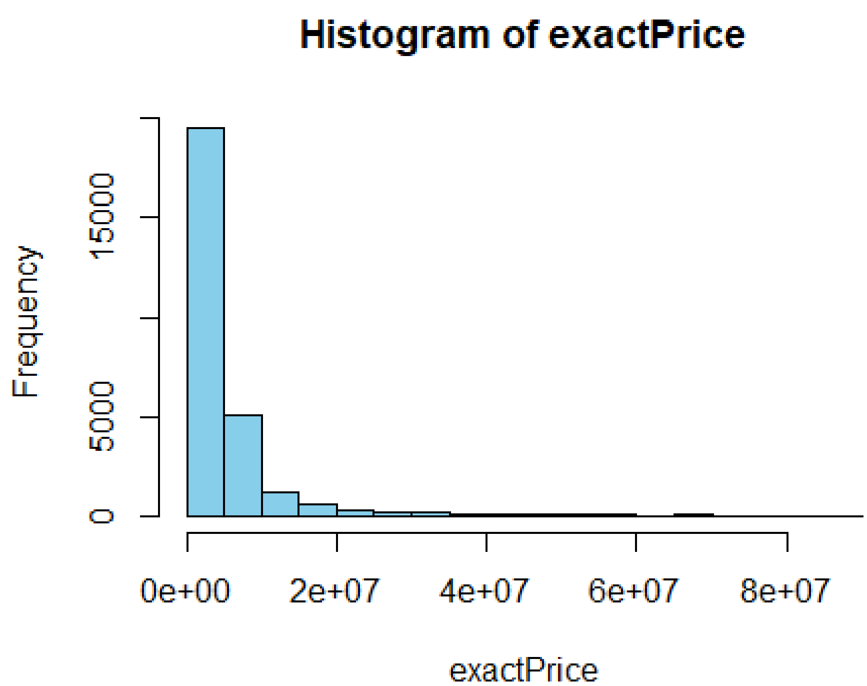
```
## Values of outliers in exactPrice: 1.1e+08 1.2e+08 1.1e+08 1.1e+08 9e+07
120300000 1.55e+08 1.05e+08 1.2e+08 1.4e+08 1.6e+08 9e+07 1.2e+08 9e+07 1e+08
1.4e+08 9e+07 1.65e+08 9.5e+07 1.4e+08 9.3e+07 1.6e+08 1.65e+08 1e+08 1.8e+08
2.45e+08 9.5e+07 1e+08 1.2e+08 9.5e+07 1.1e+08 1e+08 1.15e+08 1.8e+08 8.5e+08
1.55e+08 2.5e+08 7.8e+08 2.2e+08 1.05e+08 1.05e+08 4.5e+08 1.2e+08 9.5e+07
1.55e+08 4.6e+08 2.4e+08 2.4e+08 4.3e+08 2.4e+08 1.05e+08 97500000 5.7e+08
1.25e+08 9.5e+07 1.1e+08 4.8e+08 1.7e+08 1.02e+08 1.1e+08 2.1e+08 3.25e+09
1e+08 1.45e+08 97500000 1.15e+08 1.75e+08 142500000 1e+08 122500000 9.5e+07
167500000 1e+08 124800000 1.3e+08 3.3e+08 1.3e+08 9e+07 1e+08 1.55e+08
9.5e+07 2.15e+08 1.1e+08 1e+08 5e+08 1.05e+08 115500000 3.7e+08 1.2e+08
1.35e+08 1.5e+08 102500000 1e+08 1.2e+08 9e+07 1.85e+08 1.2e+08 1.05e+08
1.2e+08 5.25e+08 1e+08 1.7e+08 9.5e+07 1.1e+08 9.5e+07 182700000 109900000
107500000 1.3e+08 1.15e+08 97500000 2.05e+08 172500000 1.2e+08 1.25e+08
150050000 1.2e+08 1.05e+08 9.5e+07 1e+08 1.2e+08 1.05e+08 3.5e+08 1.15e+08
2.1e+08 1.1e+08 1.4e+08 1.3e+08 1.15e+08 9.5e+07 98600000 2e+08 1.1e+08
1.35e+08 3.6e+08 3.55e+08 2.5e+08 1.4e+08 151830782 1.1e+08 117500000 1.1e+08
108100000 1.25e+08 1.1e+08 111111111 225900000 1.05e+08 2e+08 1e+08 1.3e+08
1.75e+08 3.4e+08 162100000 2.1e+08 1.7e+08 7.5e+08 1.15e+08 1.5e+08 1e+08
94600000 1.35e+08 2.2e+08 9e+07 1.35e+08 1.1e+08 3.4e+08 1.05e+08 1.2e+08
1.3e+08 1.3e+08 139800000 122200000 1.26e+08 1.35e+08 148500000 1.2e+08 1e+08
2.8e+08 1.5e+08
```

Remove rows with outliers

```r
df <- df[-outliers, ]
```

Print information about removed rows

```
cat("Number of rows removed:", length(outliers), "\n")

## Number of rows removed: 180

hist(df$exactPrice, main = "Histogram of exactPrice", xlab = "exactPrice",
col = "skyblue", border = "black")
```




```
skewness(df$exactPrice)

## [1] 4.180399
```

Still data is highly skewed after removing outliers.

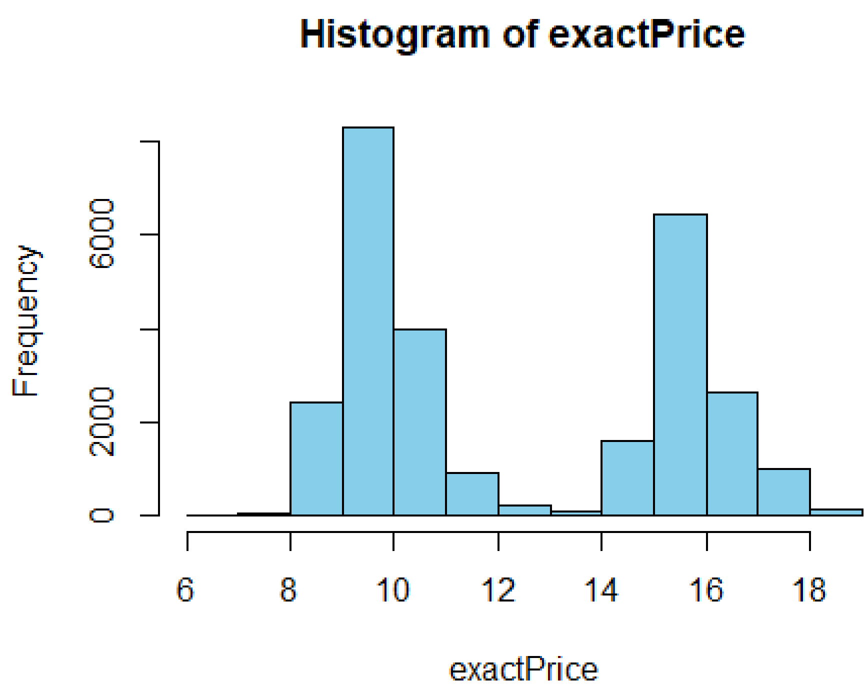
We will use log scaling to scale the data in target variable

```
df$exactPrice <- log(df$exactPrice)

skewness(df$exactPrice)

## [1] 0.2842498
```

Now the skewness is under range of -1 to 1.

Plotting the histogram of the data

```
hist(df$exactPrice, main = "Histogram of exactPrice", xlab = "exactPrice",
col = "skyblue", border = "black")
```



Question A :

- Checking the multicollinearity. We will check first the correlation values for each variable. If a variable is highly correlated, then we can say there is presence of multicollinearity hence we can simply remove those variables. We will set threshold as 0.8. Hence if a variable has correlation value greater than 0.8, we will remove the variable.
- Also we can apply a multiple linear model with the help of current columns and check the VIF values of the variables. If the vif value of the variables exceeds 10, we can say there is problem of multicollinearity due to that variable and we can simply remove the respective variable.
- We were having lot of issues in the data and we have cleaned and preprocessed the data. You can refer above steps of preprocessing where we did a lot of cleaning and preprocessing. Example removing the columns with unique values, removing the date column, removing the values with high NA%, converting the units of the column, removing other unwanted columns, removing outliers, scaling the target variable etc.

———————————— Checking Coorelation ————————————-

Find numeric columns in predictors

```
numeric_columns <- sapply(df, is.numeric)
```

Create a data frame with only numeric columns

```
numeric_predictors <- df[, numeric_columns]
```

Check for correlations between predictors

```
cor_matrix <- cor(numeric_predictors)
print(cor_matrix)

##                      exactPrice     sqftPrice securityDeposit       flrNum
## exactPrice         1.000000e+00  0.0780293500    0.0796377057  0.197169244
## sqftPrice          7.802935e-02  1.0000000000   -0.0033778434  0.008140996
## securityDeposit    7.963771e-02 -0.0033778434    1.0000000000  0.059483819
## flrNum             1.971692e-01  0.0081409961    0.0594838192  1.000000000
## firstMonthCharges -8.710237e-06 -0.0002866872   -0.0007531474  0.001691459
## totalFlrNum        2.120496e-01 -0.0025866414    0.0432074066  0.680816243
## carpetArea         9.379798e-02  0.0331048390    0.2447428611 -0.011463416
## Long              -8.292512e-02 -0.0132562207   -0.0118584337 -0.087016701
## Lat                1.547505e-01  0.0290671605   -0.1553667724 -0.074702093
##                   firstMonthCharges  totalFlrNum   carpetArea         Long
## exactPrice            -8.710237e-06  0.212049564  0.093797981 -0.082925116
## sqftPrice             -2.866872e-04 -0.002586641  0.033104839 -0.013256221
## securityDeposit       -7.531474e-04  0.043207407  0.244742861 -0.011858434
```

```
## flrNum                   1.691459e-03  0.680816243 -0.011463416 -0.087016701
## firstMonthCharges        1.000000e+00 -0.003537810 -0.001147977 -0.001334649
## totalFlrNum             -3.537810e-03  1.000000000 -0.046795891 -0.103870968
## carpetArea              -1.147977e-03 -0.046795891  1.000000000  0.060701773
## Long                    -1.334649e-03 -0.103870968  0.060701773  1.000000000
## Lat                      5.681558e-03 -0.068569175  0.094193321  0.085587589
##                              Lat
## exactPrice           0.154750499
## sqftPrice            0.029067161
## securityDeposit     -0.155366772
## flrNum              -0.074702093
## firstMonthCharges    0.005681558
## totalFlrNum         -0.068569175
## carpetArea           0.094193321
## Long                 0.085587589
## Lat                  1.000000000
```

```r
highly_correlated_vars <- findCorrelation(cor_matrix, cutoff = 0.8)
highly_correlated_vars
```

```
## integer(0)
```

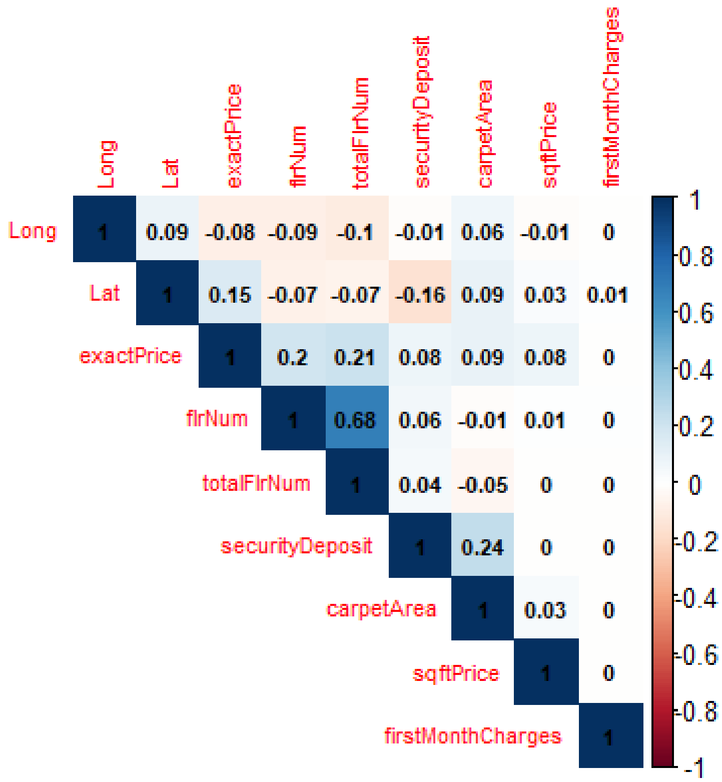
Creating a heatmap

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```r
corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
tl.cex = 0.7, addCoef.col = "black", number.cex = 0.7, number.digits = 2)
```



By checking correlation values, we can see there is not a single variable which is highly correlated with target variale.

We will fit a temporary model only on numeric data to check if there is multicollinearity

```
model_temp <- lm(exactPrice ~ ., data = numeric_predictors)
summary(model_temp)

##
## Call:
## lm(formula = exactPrice ~ ., data = numeric_predictors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.035  -2.553  -1.362   2.973   7.013
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.239e+01  2.769e-01   44.74   <2e-16 ***
## sqftPrice          1.416e-06  1.164e-07   12.17   <2e-16 ***
## securityDeposit    2.383e-06  1.816e-07   13.12   <2e-16 ***
## flrNum             8.710e-02  6.958e-03   12.52   <2e-16 ***
## firstMonthCharges -3.966e-11  3.608e-10   -0.11    0.912
## totalFlrNum        7.009e-02  3.634e-03   19.29   <2e-16 ***
## carpetArea         3.426e-04  2.998e-05   11.43   <2e-16 ***
```

```
## Long                -4.595e-02  3.440e-03  -13.36   <2e-16 ***
## Lat                  9.934e-02  3.177e-03   31.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.914 on 27711 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.1035
## F-statistic: 400.9 on 8 and 27711 DF,  p-value: < 2.2e-16
```

- We can see we got very low adjusted R-Squared with this numeric data. Now we will check vif values for these numeric columns.

Calculate VIF

```
vif_values <- car::vif(model_temp)
print(model_temp)
```

```
##
## Call:
## lm(formula = exactPrice ~ ., data = numeric_predictors)
##
## Coefficients:
##       (Intercept)           sqftPrice     securityDeposit              flrNum
##         1.239e+01           1.416e-06           2.383e-06           8.710e-02
## firstMonthCharges         totalFlrNum          carpetArea                Long
##        -3.966e-11           7.009e-02           3.426e-04          -4.595e-02
##               Lat
##         9.934e-02
```

- We can see all the variables have vif values under 10. Hence we can say there is no multicollinearity in the data.

—————————————- EDA ————————————————

Question B:

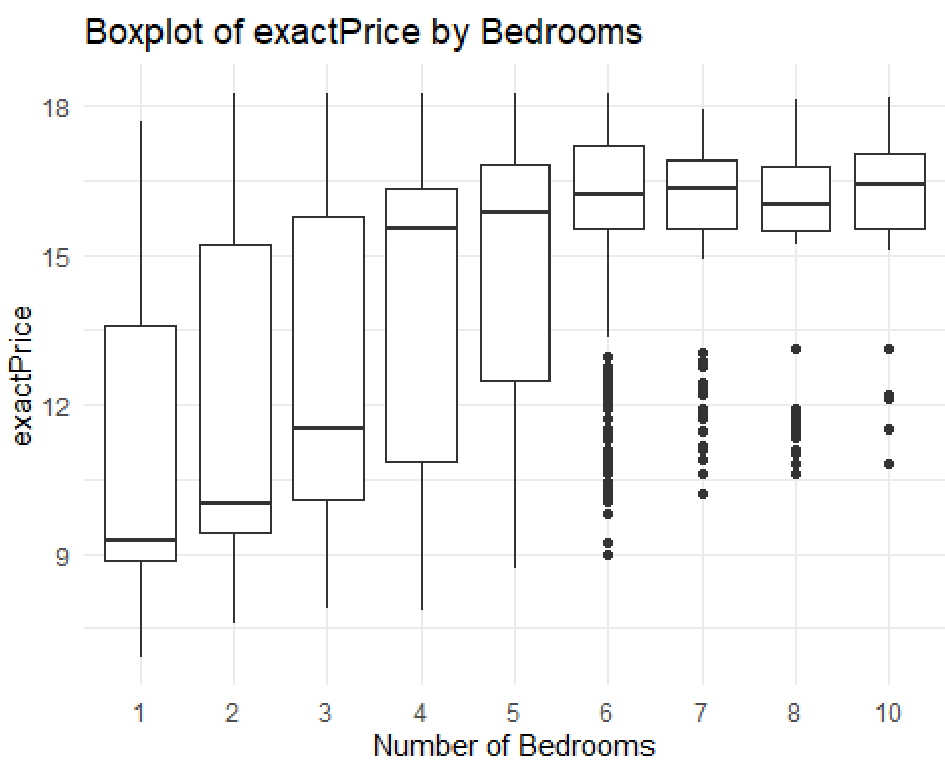
ii) Now we will see the exploratory data analysis for the variables.

Please refer below mentioned plots and description for the data insights

```
library(ggplot2)
```

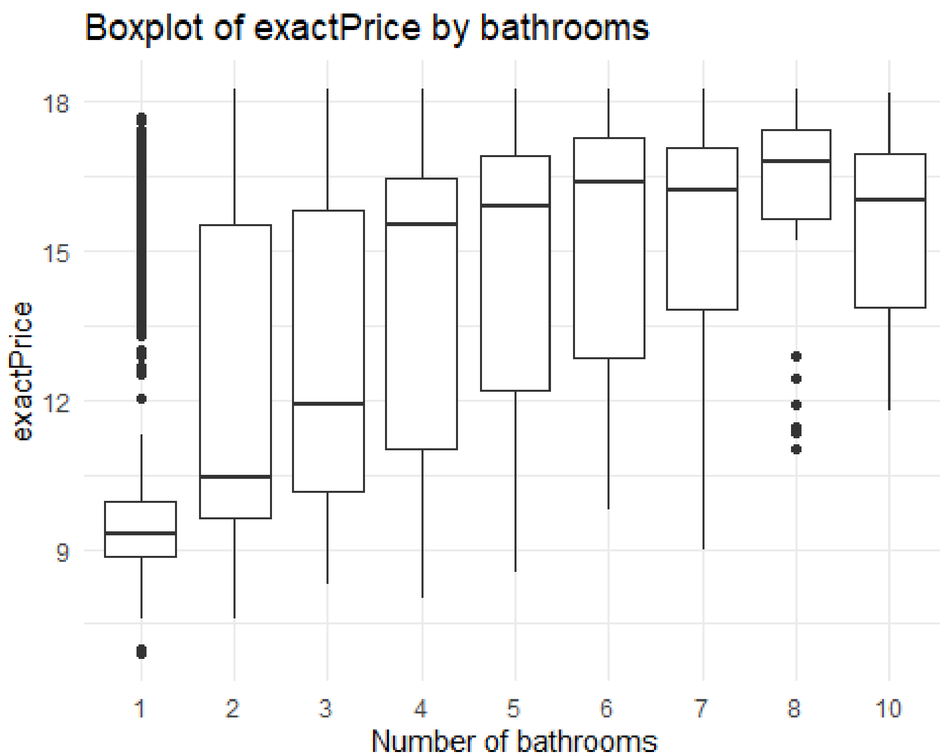
Here in the below mentioned plot, we can clearly see that as number of bedrooms increases, the price increases. We were able to say this on the basis of median of the boxplots as median increases as we increase the number of bedrooms.

```r
ggplot(df, aes(x = as.factor(bedrooms), y = exactPrice)) +
  geom_boxplot() +
  labs(title = "Boxplot of exactPrice by Bedrooms",
       x = "Number of Bedrooms",
       y = "exactPrice") +
  theme_minimal()
```



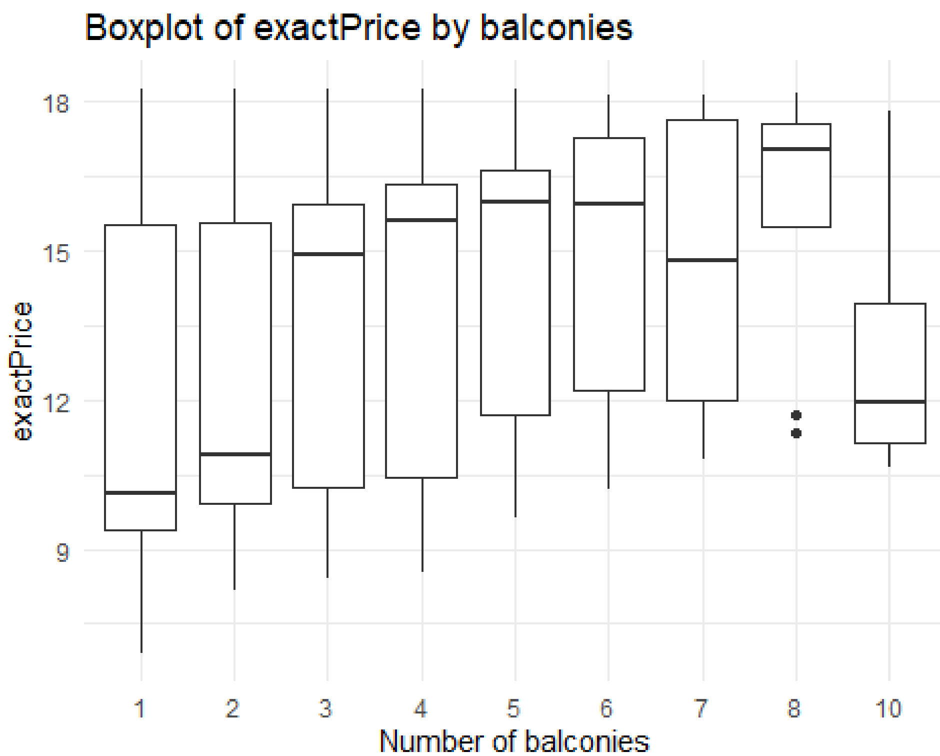
Same analysis goes for number of bathrooms, as we increase number of bathrooms, price of the house increases.

```r
ggplot(df, aes(x = as.factor(bathrooms), y = exactPrice)) +
  geom_boxplot() +
  labs(title = "Boxplot of exactPrice by bathrooms",
       x = "Number of bathrooms",
       y = "exactPrice") +
  theme_minimal()
```



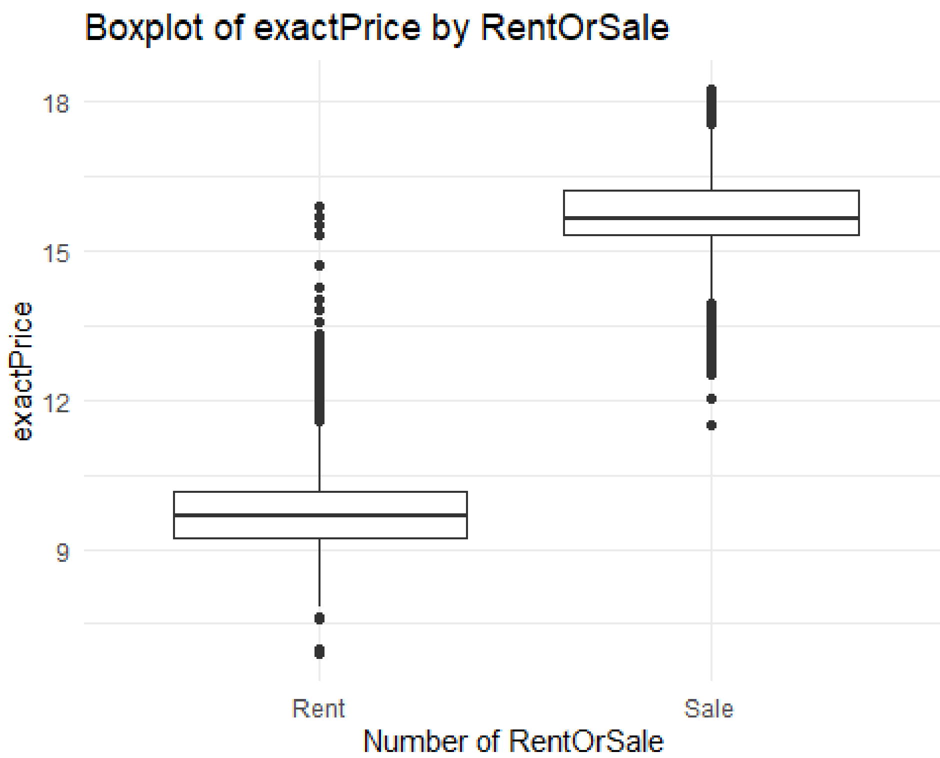
We can see here that prices range is high with number of balconies between 3 to 8. Price is low for number of balconies = 10 which is strange.

```
ggplot(df, aes(x = as.factor(balconies), y = exactPrice)) +
  geom_boxplot() +
  labs(title = "Boxplot of exactPrice by balconies",
       x = "Number of balconies",
       y = "exactPrice") +
  theme_minimal()
```



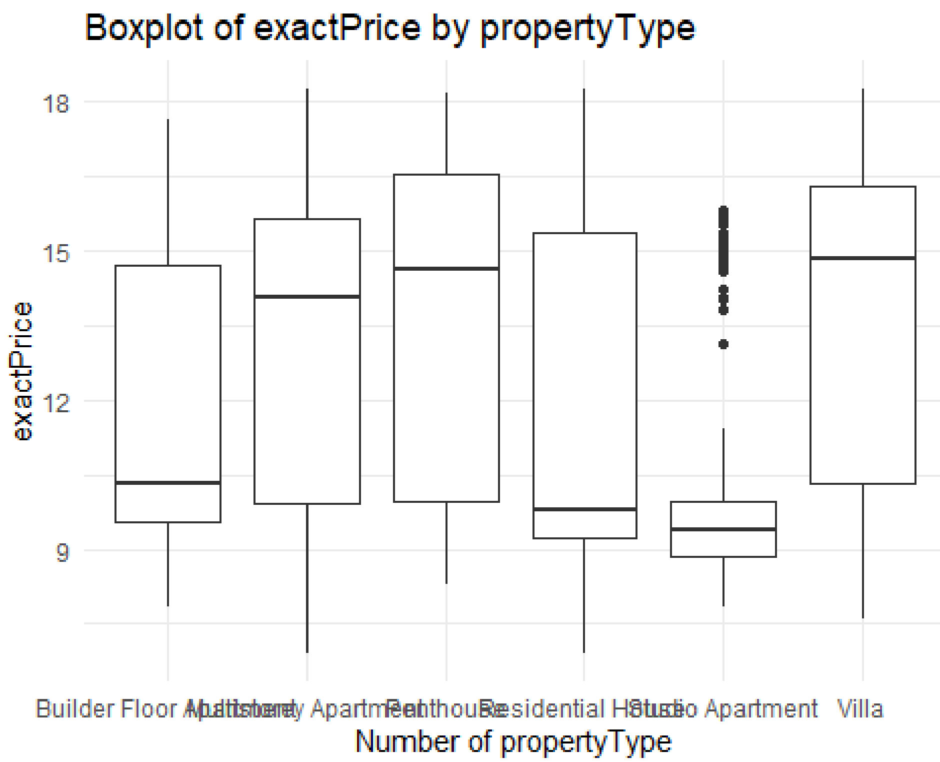
If a house is on sale, price is large as compared to house on rent which is obvious

```
ggplot(df, aes(x = as.factor(RentOrSale), y = exactPrice)) +
  geom_boxplot() +
  labs(title = "Boxplot of exactPrice by RentOrSale",
       x = "Number of RentOrSale",
       y = "exactPrice") +
  theme_minimal()
```



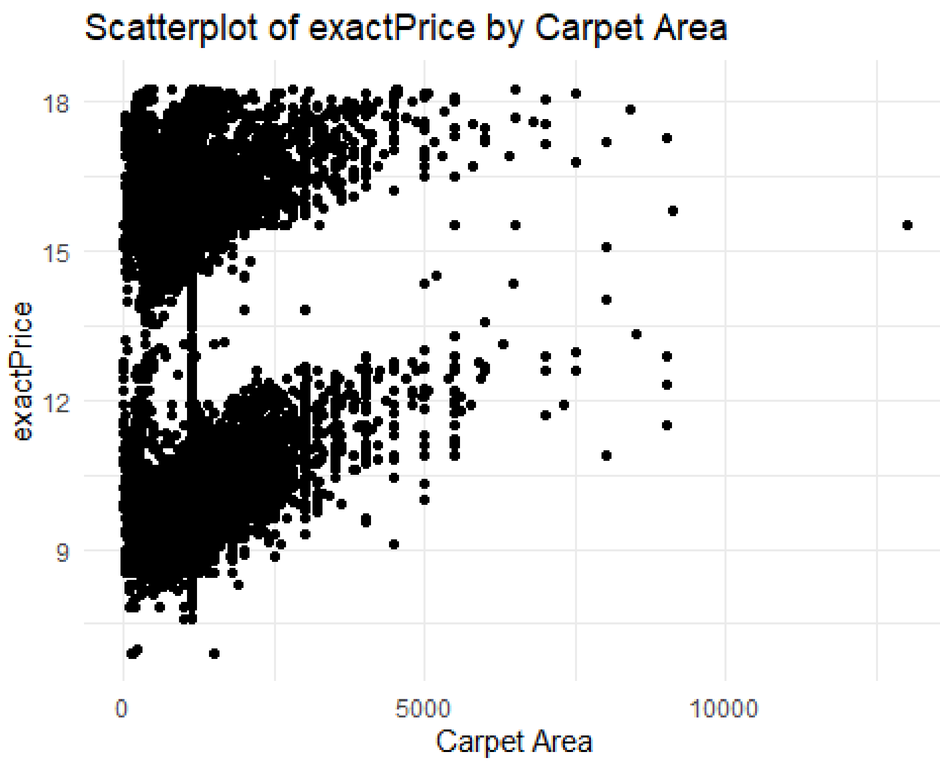
If a house is Multistory, Penthouse and Villa, the price is high so these variables must be significant in predicting the price.

```r
ggplot(df, aes(x = as.factor(propertyType), y = exactPrice)) +
  geom_boxplot() +
  labs(title = "Boxplot of exactPrice by propertyType",
       x = "Number of propertyType",
       y = "exactPrice") +
  theme_minimal()
```



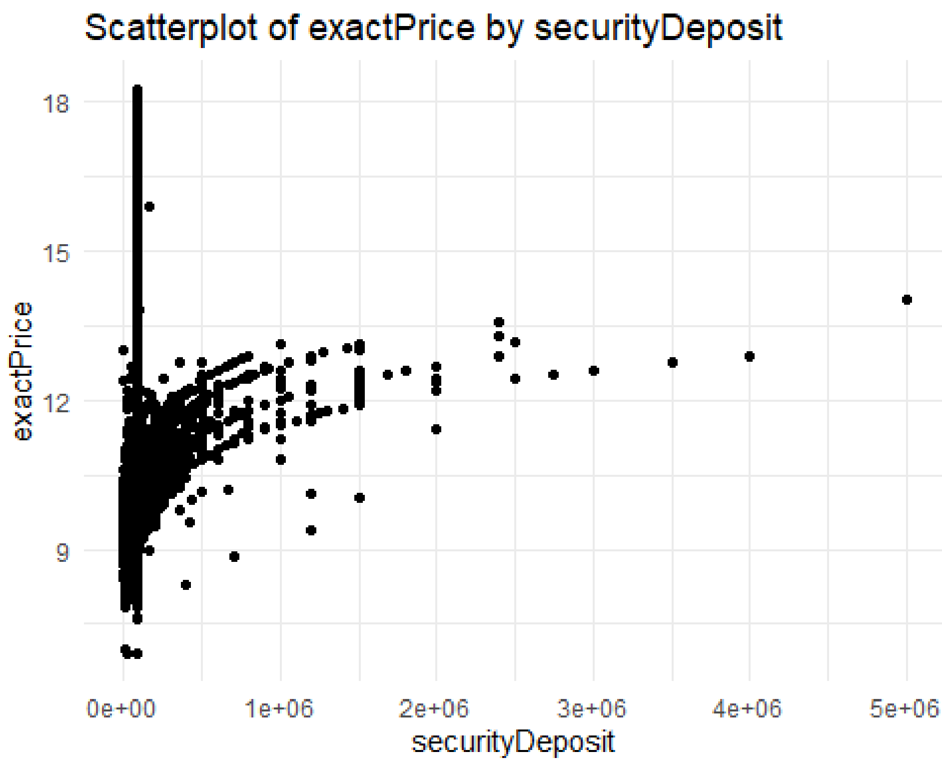
There are outliers present in the carpet are variable hence we are not able to capture any relation

```
ggplot(df, aes(x = carpetArea, y = exactPrice)) +
  geom_point() +
  labs(title = "Scatterplot of exactPrice by Carpet Area",
       x = "Carpet Area",
       y = "exactPrice") +
  theme_minimal()
```



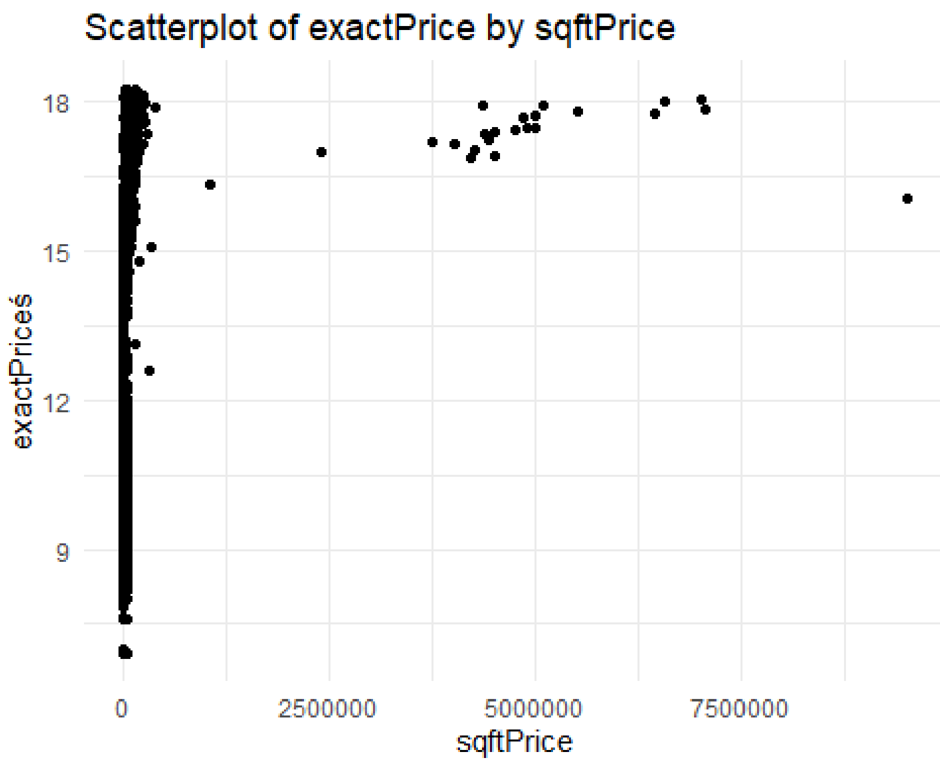
We can see there is slight positive correlation between these two variables

```
ggplot(df, aes(x = securityDeposit, y = exactPrice)) +
  geom_point() +
  labs(title = "Scatterplot of exactPrice by securityDeposit",
       x = "securityDeposit",
       y = "exactPrice") +
  theme_minimal()
```



We can see positive correlation between Sqft price and exact price

```
ggplot(df, aes(x = sqftPrice, y = exactPrice)) +  geom_point() +
  labs(title = "Scatterplot of exactPrice by sqftPrice",
       x = "sqftPrice",
       y = "exactPriceś") +
  theme_minimal()
```



—————————————— Combining Data ——————————————

Creating dummy variables

```
df_combined_dummies <- df %>% model.matrix(~ . - 1, data = .) %>%
as.data.frame()
dim(df_combined_dummies)

## [1] 27720    70
```

—————————————— Splitting Data ——————————————

Creating a train/test partition

```
set.seed(123)
splitIndex <- createDataPartition(df_combined_dummies$exactPrice, p = 0.8,
list = FALSE)
df_train <- df_combined_dummies[splitIndex, ]
df_test <- df_combined_dummies[-splitIndex, ]

dim(df_train)

## [1] 22178    70
```

```r
dim(df_test)

## [1] 5542   70
```

Apply linear regression

```r
Initial_model <- lm(exactPrice ~ ., data=df_train)
summary(Initial_model)

##
## Call:
## lm(formula = exactPrice ~ ., data = df_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.949 -0.324 -0.055  0.228  7.056
##
## Coefficients: (4 not defined because of singularities)
##                                       Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                          1.460e+01  6.777e-01  21.544  < 2e-
16
## sqftPrice                            2.441e-07  2.846e-08   8.576  < 2e-
16
## securityDeposit                      1.032e-06  4.748e-08  21.738  < 2e-
16
## `propertyTypeBuilder Floor Apartment` -1.594e-01  2.418e-02  -6.589 4.52e-
11
## `propertyTypeMultistorey Apartment`   -1.146e-01  2.057e-02  -5.574 2.53e-
08
## propertyTypePenthouse                -6.160e-02  6.671e-02  -0.923
0.355794
## `propertyTypeResidential House`      -1.847e-01  2.045e-02  -9.033  < 2e-
16
## `propertyTypeStudio Apartment`       -4.788e-01  5.227e-02  -9.160  < 2e-
16
## propertyTypeVilla                           NA         NA      NA
NA
## `furnishingSemi-Furnished`           -1.181e-01  1.275e-02  -9.262  < 2e-
16
## furnishingUnfurnished                -2.404e-01  1.302e-02 -18.469  < 2e-
16
## flrNum                               8.480e-03  1.843e-03   4.601 4.23e-
06
## firstMonthCharges                    -5.583e-11  8.272e-11  -0.675
0.499714
## facingNorth                          9.677e-02  1.877e-02   5.157 2.53e-
07
## `facingNorth - East`                 5.303e-03  1.723e-02   0.308
0.758267
```

```
## `facingNorth - West`                        1.688e-01  3.538e-02   4.772 1.84e-
06
## facingSouth                                 7.640e-02  3.388e-02   2.255
0.024165
## `facingSouth - East`                        1.172e-01  3.410e-02   3.437
0.000589
## `facingSouth -West`                         1.521e-01  4.575e-02   3.325
0.000886
## facingWest                                  3.879e-02  2.232e-02   1.738
0.082268
## totalFlrNum                                 6.129e-03  1.137e-03   5.389 7.16e-
08
## cityAgartala                               -5.909e+00  6.891e-01  -8.575  < 2e-
16
## cityBangalore                              -5.668e+00  6.678e-01  -8.488  < 2e-
16
## cityBhopal                                 -6.072e+00  6.674e-01  -9.098  < 2e-
16
## cityChandigarh                             -5.405e+00  6.676e-01  -8.097 5.92e-
16
## cityChennai                                -5.815e+00  6.678e-01  -8.708  < 2e-
16
## cityDehradun                               -5.562e+00  6.675e-01  -8.333  < 2e-
16
## cityGandhinagar                            -5.603e+00  6.675e-01  -8.393  < 2e-
16
## cityGangtok                                -5.548e+00  7.312e-01  -7.587 3.40e-
14
## cityGoa                                    -5.112e+00  6.676e-01  -7.658 1.97e-
14
## cityHyderabad                              -5.820e+00  6.676e-01  -8.718  < 2e-
16
## cityJaipur                                 -5.886e+00  6.674e-01  -8.820  < 2e-
16
## cityKolkata                                -5.823e+00  6.676e-01  -8.722  < 2e-
16
## cityLucknow                                -5.747e+00  6.675e-01  -8.611  < 2e-
16
## cityMumbai                                 -4.450e+00  6.680e-01  -6.661 2.78e-
11
## `cityNew-Delhi`                            -5.331e+00  6.678e-01  -7.983 1.49e-
15
## `cityNew Delhi`                            -5.429e+00  6.836e-01  -7.942 2.09e-
15
## cityPatna                                  -5.696e+00  6.674e-01  -8.535  < 2e-
16
## cityRaipur                                 -6.050e+00  6.675e-01  -9.064  < 2e-
16
## carpetArea                                  1.801e-04  8.675e-06  20.763  < 2e-
16
```

```
## bedrooms2                                  2.995e-01  1.873e-02  15.985  < 2e-
16
## bedrooms3                                  4.953e-01  2.252e-02  21.997  < 2e-
16
## bedrooms4                                  7.136e-01  3.023e-02  23.604  < 2e-
16
## bedrooms5                                  8.020e-01  4.400e-02  18.226  < 2e-
16
## bedrooms6                                  9.438e-01  5.594e-02  16.870  < 2e-
16
## bedrooms7                                  8.312e-01  8.259e-02  10.065  < 2e-
16
## bedrooms8                                  8.663e-01  1.048e-01   8.267  < 2e-
16
## bedrooms9                                         NA         NA      NA
NA
## bedrooms10                                 8.500e-01  1.382e-01   6.149 7.95e-
10
## bathrooms2                                 3.408e-01  1.699e-02  20.054  < 2e-
16
## bathrooms3                                 5.354e-01  2.253e-02  23.763  < 2e-
16
## bathrooms4                                 7.642e-01  3.062e-02  24.956  < 2e-
16
## bathrooms5                                 9.403e-01  4.425e-02  21.247  < 2e-
16
## bathrooms6                                 9.979e-01  6.336e-02  15.751  < 2e-
16
## bathrooms7                                 8.592e-01  9.576e-02   8.972  < 2e-
16
## bathrooms8                                 1.282e+00  1.417e-01   9.048  < 2e-
16
## bathrooms9                                        NA         NA      NA
NA
## bathrooms10                                1.288e+00  2.597e-01   4.959 7.15e-
07
## balconies2                                 2.168e-02  1.205e-02   1.799
0.072093
## balconies3                                 3.230e-02  1.687e-02   1.915
0.055550
## balconies4                                -9.041e-03  2.744e-02  -0.329
0.741837
## balconies5                                 7.836e-03  6.095e-02   0.129
0.897708
## balconies6                                 9.636e-02  1.175e-01   0.820
0.412017
## balconies7                                 3.685e-01  2.386e-01   1.545
0.122478
## balconies8                                 2.840e-01  2.374e-01   1.196
0.231556
```

```
## balconies9                                     NA         NA       NA
NA
## balconies10                             1.054e+00  3.863e-01   2.729
0.006353
## RentOrSaleSale                          5.614e+00  1.152e-02 487.395  < 2e-
16
## Long                                    3.316e-03  1.273e-03   2.605
0.009203
## Lat                                    -3.450e-03  2.198e-03  -1.569
0.116553
##
## (Intercept)                            ***
## sqftPrice                              ***
## securityDeposit                        ***
## `propertyTypeBuilder Floor Apartment`  ***
## `propertyTypeMultistorey Apartment`    ***
## propertyTypePenthouse
## `propertyTypeResidential House`        ***
## `propertyTypeStudio Apartment`         ***
## propertyTypeVilla
## `furnishingSemi-Furnished`             ***
## furnishingUnfurnished                  ***
## flrNum                                 ***
## firstMonthCharges
## facingNorth                            ***
## `facingNorth - East`
## `facingNorth - West`                   ***
## facingSouth                            *
## `facingSouth - East`                   ***
## `facingSouth -West`                    ***
## facingWest                             .
## totalFlrNum                            ***
## cityAgartala                           ***
## cityBangalore                          ***
## cityBhopal                             ***
## cityChandigarh                         ***
## cityChennai                            ***
## cityDehradun                           ***
## cityGandhinagar                        ***
## cityGangtok                            ***
## cityGoa                                ***
## cityHyderabad                          ***
## cityJaipur                             ***
## cityKolkata                            ***
## cityLucknow                            ***
## cityMumbai                             ***
## `cityNew-Delhi`                        ***
## `cityNew Delhi`                        ***
## cityPatna                              ***
## cityRaipur                             ***
```

```
## carpetArea                             ***
## bedrooms2                              ***
## bedrooms3                              ***
## bedrooms4                              ***
## bedrooms5                              ***
## bedrooms6                              ***
## bedrooms7                              ***
## bedrooms8                              ***
## bedrooms9
## bedrooms10                             ***
## bathrooms2                             ***
## bathrooms3                             ***
## bathrooms4                             ***
## bathrooms5                             ***
## bathrooms6                             ***
## bathrooms7                             ***
## bathrooms8                             ***
## bathrooms9
## bathrooms10                            ***
## balconies2                             .
## balconies3                             .
## balconies4
## balconies5
## balconies6
## balconies7
## balconies8
## balconies9
## balconies10                            **
## RentOrSaleSale                         ***
## Long                                   **
## Lat
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6671 on 22112 degrees of freedom
## Multiple R-squared:  0.9531, Adjusted R-squared:  0.953
## F-statistic:  6920 on 65 and 22112 DF,  p-value: < 2.2e-16
```

- We can see the Adjusted R-squared value is 0.9521. Also there are many insignificant variables which we can remove further with step function.

Making predictions on test data

```
predictions <- predict(Initial_model, newdata = df_test)
```

Calculate Mean Squared Error (MSE)

```
mse_initial <- mean((df_test$exactPrice - predictions)^2)
cat("Mean Squared Error (MSE):", mse_initial, "\n")

## Mean Squared Error (MSE): 0.4478998
```

Calculate Mean Absolute Error (MAE)

```
mae_initial <- mean(abs(df_test$exactPrice - predictions))
cat("Mean Absolute Error (MAE):", mae_initial, "\n")

## Mean Absolute Error (MAE): 0.3964109
```

We can see we got vary low MSE and MAE values.

Question B:

i) We will perform backward elimination method to select significant variables. Commented this line of code as it takes time to run the code.

backward_elimination <- step(Initial_model, direction = “backward”)

Creating the model based on the variables selected by backward elimination variable selection method.

```
backward_model <- lm(exactPrice ~ sqftPrice + securityDeposit +
`propertyTypeBuilder Floor Apartment` +
                       `propertyTypeMultistorey Apartment` +
`propertyTypeResidential House` +
                       `propertyTypeStudio Apartment` + `furnishingSemi-
Furnished` +
                       furnishingUnfurnished + flrNum + facingNorth +
`facingNorth - West` +
                       facingSouth + `facingSouth - East` + `facingSouth -
West` +
                       facingWest + totalFlrNum + cityAgartala +
cityBangalore +
                       cityBhopal + cityChandigarh + cityChennai +
cityDehradun +
                       cityGandhinagar + cityGangtok + cityGoa +
cityHyderabad +
```

```
                       cityJaipur + cityKolkata + cityLucknow + cityMumbai +
`cityNew-Delhi` +
                       `cityNew Delhi` + cityPatna + cityRaipur + carpetArea
+ bedrooms2 +
                       bedrooms3 + bedrooms4 + bedrooms5 + bedrooms6 +
bedrooms7 +
                       bedrooms8 + bedrooms10 + bathrooms2 + bathrooms3 +
bathrooms4 +
                       bathrooms5 + bathrooms6 + bathrooms7 + bathrooms8 +
bathrooms10 +
                       balconies2 + balconies3 + balconies4 + balconies6 +
balconies7 +
                       balconies8 + balconies10 + RentOrSaleSale + Long, data
= df_train)

summary(backward_model)

##
## Call:
## lm(formula = exactPrice ~ sqftPrice + securityDeposit +
`propertyTypeBuilder Floor Apartment` +
##     `propertyTypeMultistorey Apartment` + `propertyTypeResidential House`
+
##     `propertyTypeStudio Apartment` + `furnishingSemi-Furnished` +
##     furnishingUnfurnished + flrNum + facingNorth + `facingNorth - West` +
##     facingSouth + `facingSouth - East` + `facingSouth -West` +
##     facingWest + totalFlrNum + cityAgartala + cityBangalore +
##     cityBhopal + cityChandigarh + cityChennai + cityDehradun +
##     cityGandhinagar + cityGangtok + cityGoa + cityHyderabad +
##     cityJaipur + cityKolkata + cityLucknow + cityMumbai + `cityNew-Delhi`
+
##     `cityNew Delhi` + cityPatna + cityRaipur + carpetArea + bedrooms2 +
##     bedrooms3 + bedrooms4 + bedrooms5 + bedrooms6 + bedrooms7 +
##     bedrooms8 + bedrooms10 + bathrooms2 + bathrooms3 + bathrooms4 +
##     bathrooms5 + bathrooms6 + bathrooms7 + bathrooms8 + bathrooms10 +
##     balconies2 + balconies3 + balconies4 + balconies6 + balconies7 +
##     balconies8 + balconies10 + RentOrSaleSale + Long, data = df_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9432 -0.3246 -0.0554  0.2278  7.0549
##
## Coefficients:
##                                       Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                          1.450e+01  6.749e-01  21.484  < 2e-
16
## sqftPrice                            2.447e-07  2.845e-08   8.600  < 2e-
16
## securityDeposit                      1.035e-06  4.738e-08  21.838  < 2e-
```

```
16
## `propertyTypeBuilder Floor Apartment` -1.557e-01  2.375e-02  -6.559 5.54e-
11
## `propertyTypeMultistorey Apartment`   -1.105e-01  2.001e-02  -5.523 3.38e-
08
## `propertyTypeResidential House`       -1.812e-01  1.996e-02  -9.080  < 2e-
16
## `propertyTypeStudio Apartment`        -4.741e-01  5.207e-02  -9.106  < 2e-
16
## `furnishingSemi-Furnished`            -1.179e-01  1.275e-02  -9.248  < 2e-
16
## furnishingUnfurnished                 -2.401e-01  1.301e-02 -18.455  < 2e-
16
## flrNum                                 8.470e-03  1.843e-03   4.597 4.32e-
06
## facingNorth                            9.557e-02  1.861e-02   5.135 2.85e-
07
## `facingNorth - West`                   1.674e-01  3.521e-02   4.753 2.02e-
06
## facingSouth                            7.503e-02  3.384e-02   2.217
0.026614
## `facingSouth - East`                   1.166e-01  3.398e-02   3.430
0.000604
## `facingSouth -West`                    1.515e-01  4.566e-02   3.318
0.000908
## facingWest                             3.814e-02  2.226e-02   1.714
0.086605
## totalFlrNum                            6.113e-03  1.136e-03   5.383 7.38e-
08
## cityAgartala                          -5.912e+00  6.891e-01  -8.579  < 2e-
16
## cityBangalore                         -5.633e+00  6.674e-01  -8.440  < 2e-
16
## cityBhopal                            -6.070e+00  6.673e-01  -9.096  < 2e-
16
## cityChandigarh                        -5.427e+00  6.674e-01  -8.132 4.44e-
16
## cityChennai                           -5.782e+00  6.675e-01  -8.662  < 2e-
16
## cityDehradun                          -5.582e+00  6.673e-01  -8.365  < 2e-
16
## cityGandhinagar                       -5.600e+00  6.675e-01  -8.389  < 2e-
16
## cityGangtok                           -5.551e+00  7.312e-01  -7.592 3.28e-
14
## cityGoa                               -5.085e+00  6.674e-01  -7.620 2.64e-
14
## cityHyderabad                         -5.800e+00  6.675e-01  -8.690  < 2e-
16
## cityJaipur                            -5.896e+00  6.673e-01  -8.835  < 2e-
```

```
16
## cityKolkata                              -5.822e+00  6.676e-01  -8.721  < 2e-
16
## cityLucknow                              -5.758e+00  6.674e-01  -8.627  < 2e-
16
## cityMumbai                               -4.432e+00  6.679e-01  -6.637 3.28e-
11
## `cityNew-Delhi`                          -5.348e+00  6.677e-01  -8.010 1.20e-
15
## `cityNew Delhi`                          -5.427e+00  6.836e-01  -7.939 2.13e-
15
## cityPatna                                -5.704e+00  6.673e-01  -8.547  < 2e-
16
## cityRaipur                               -6.043e+00  6.675e-01  -9.053  < 2e-
16
## carpetArea                                1.798e-04  8.670e-06  20.742  < 2e-
16
## bedrooms2                                 2.993e-01  1.873e-02  15.977  < 2e-
16
## bedrooms3                                 4.954e-01  2.251e-02  22.005  < 2e-
16
## bedrooms4                                 7.143e-01  3.021e-02  23.645  < 2e-
16
## bedrooms5                                 8.031e-01  4.398e-02  18.260  < 2e-
16
## bedrooms6                                 9.443e-01  5.591e-02  16.891  < 2e-
16
## bedrooms7                                 8.321e-01  8.257e-02  10.077  < 2e-
16
## bedrooms8                                 8.665e-01  1.048e-01   8.269  < 2e-
16
## bedrooms10                                8.496e-01  1.382e-01   6.147 8.04e-
10
## bathrooms2                                3.412e-01  1.699e-02  20.082  < 2e-
16
## bathrooms3                                5.360e-01  2.252e-02  23.799  < 2e-
16
## bathrooms4                                7.653e-01  3.059e-02  25.016  < 2e-
16
## bathrooms5                                9.415e-01  4.410e-02  21.351  < 2e-
16
## bathrooms6                                9.993e-01  6.328e-02  15.792  < 2e-
16
## bathrooms7                                8.616e-01  9.571e-02   9.002  < 2e-
16
## bathrooms8                                1.286e+00  1.416e-01   9.080  < 2e-
16
## bathrooms10                               1.290e+00  2.595e-01   4.971 6.71e-
07
## balconies2                                2.154e-02  1.190e-02   1.810
```

```
0.070372
## balconies3                            3.185e-02  1.668e-02   1.909
0.056267
## balconies4                           -9.564e-03  2.716e-02  -0.352
0.724766
## balconies6                            9.506e-02  1.173e-01   0.810
0.417817
## balconies7                            3.683e-01  2.386e-01   1.544
0.122665
## balconies8                            2.918e-01  2.372e-01   1.230
0.218736
## balconies10                           1.054e+00  3.863e-01   2.727
0.006387
## RentOrSaleSale                        5.613e+00  1.151e-02 487.821  < 2e-
16
## Long                                  3.510e-03  1.267e-03   2.770
0.005616
##
## (Intercept)                          ***
## sqftPrice                            ***
## securityDeposit                      ***
## `propertyTypeBuilder Floor Apartment` ***
## `propertyTypeMultistorey Apartment`   ***
## `propertyTypeResidential House`       ***
## `propertyTypeStudio Apartment`        ***
## `furnishingSemi-Furnished`            ***
## furnishingUnfurnished                 ***
## flrNum                                ***
## facingNorth                           ***
## `facingNorth - West`                  ***
## facingSouth                           *
## `facingSouth - East`                  ***
## `facingSouth -West`                   ***
## facingWest                            .
## totalFlrNum                           ***
## cityAgartala                          ***
## cityBangalore                         ***
## cityBhopal                            ***
## cityChandigarh                        ***
## cityChennai                           ***
## cityDehradun                          ***
## cityGandhinagar                       ***
## cityGangtok                           ***
## cityGoa                               ***
## cityHyderabad                         ***
## cityJaipur                            ***
## cityKolkata                           ***
## cityLucknow                           ***
## cityMumbai                            ***
## `cityNew-Delhi`                       ***
```

```
## `cityNew Delhi`                        ***
## cityPatna                              ***
## cityRaipur                             ***
## carpetArea                             ***
## bedrooms2                              ***
## bedrooms3                              ***
## bedrooms4                              ***
## bedrooms5                              ***
## bedrooms6                              ***
## bedrooms7                              ***
## bedrooms8                              ***
## bedrooms10                             ***
## bathrooms2                             ***
## bathrooms3                             ***
## bathrooms4                             ***
## bathrooms5                             ***
## bathrooms6                             ***
## bathrooms7                             ***
## bathrooms8                             ***
## bathrooms10                            ***
## balconies2                             .
## balconies3                             .
## balconies4
## balconies6
## balconies7
## balconies8
## balconies10                            **
## RentOrSaleSale                         ***
## Long                                   **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.667 on 22117 degrees of freedom
## Multiple R-squared:  0.9531, Adjusted R-squared:  0.953
## F-statistic:  7497 on 60 and 22117 DF,  p-value: < 2.2e-16
```

Making predictions on test data

```
predictions_test <- predict(backward_model, newdata = df_test)
```

Calculate MSE

```
mse_backward <- mean((df_test$exactPrice - predictions_test)^2)
cat("Mean Squared Error (MSE):", mse_backward, "\n")

## Mean Squared Error (MSE): 0.4478945
```

Calculate MAE

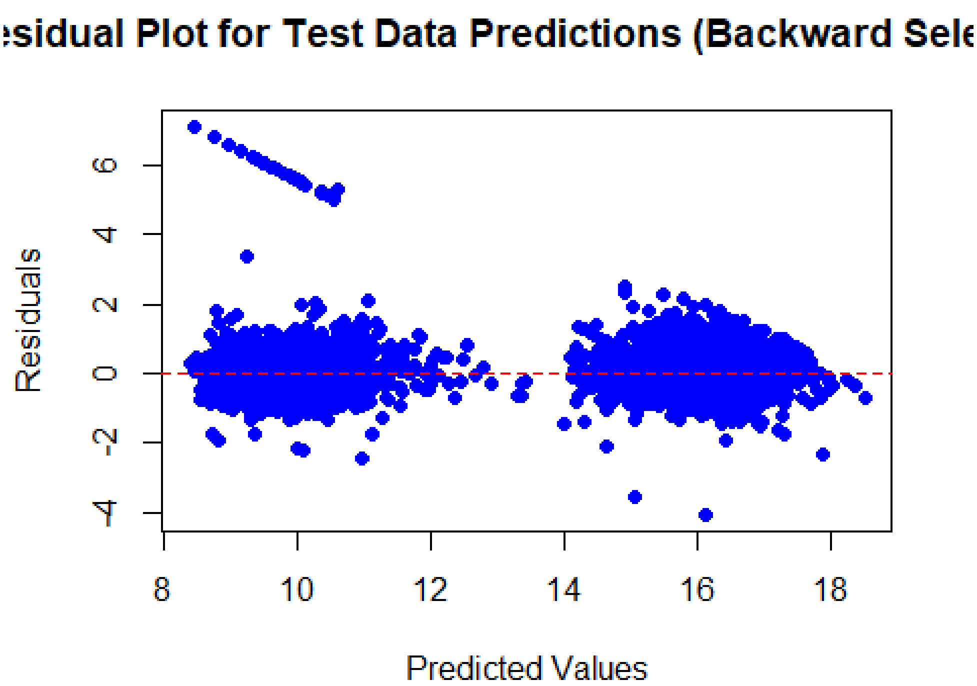
```r
mae_backward <- mean(abs(df_test$exactPrice - predictions_test))
cat("Mean Absolute Error (MAE):", mae_backward, "\n")

## Mean Absolute Error (MAE): 0.3964582
```

Calculate residuals

```r
residuals_backward <- df_test$exactPrice - predictions_test
```

Residual Plot

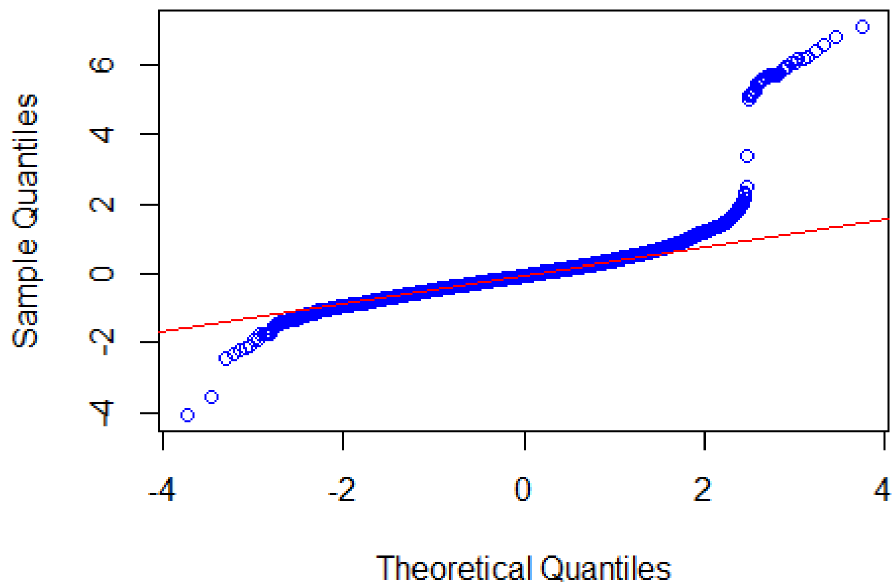
```r
plot(predictions_test, residuals_backward,
     xlab = "Predicted Values", ylab = "Residuals",
     main = "Residual Plot for Test Data Predictions (Backward Selection)",
     pch = 16, col = "blue")
abline(h = 0, col = "red", lty = 2)
```



QQ Plot

```r
qqnorm(residuals_backward, main = "QQ Plot for Test Data Predictions
(Backward Selection)", col = "blue")
qqline(residuals_backward, col = "red")
```

QQ Plot for Test Data Predictions (Backward Selecti



- We can see that variation of residuals with respect to predicted values is constant. Hence we can say the model is good. Also many points are following the line in the QQ plot.
- We can see all the significant variables are selected by backward elimination method. All the variables have p value less than 0.05 except some of the balconies variables.
- The large F-statistic and the very small p-value indicate that the regression model as a whole is highly significant, suggesting that the set of independent variables jointly have a significant effect on the dependent variable.
- The method chosen for variable selection is backward elimination. This method iteratively removes insignificant variables from the model until all remaining variables are statistically significant. Backward elimination starts with a full model including all variables and progressively removes variables based on their p-values until all remaining variables have p-values below a chosen threshold.
- We have selected backward elimination was employed to refine the initial model obtained through linear regression. By systematically removing variables with high p-values, the resulting model aims to improve interpretability, reduce overfitting, and enhance predictive accuracy by focusing on the most relevant predictors.
- The overall significance of the regression fit can be assessed based on several metrics:

- Adjusted R-squared: The adjusted R-squared value indicates the proportion of variance in the response variable that is explained by the model, adjusted for the number of predictors. In this case, the adjusted R-squared value is 0.9521, indicating that approximately 95.30% of the variance in the exactPrice variable is explained by the selected predictors.
- Significance of coefficients: The coefficients associated with each predictor variable provide insight into their impact on the response variable. In the summary output provided, most coefficients have extremely low p-values (indicated by '***'), suggesting that the corresponding predictors are statistically significant in predicting the exactPrice.
- Also what we saw in the EDA part, variables which we saw have linear or non-linear relations, those variables were selected by the backward elimination method and have signicant effect on target variable.