***HOUSING REAL ESTATE INVESTMENTS:***

***IN TECH-EMERGING CITIES IN INDIA***

*-TEJASWINI KESIRAJU*

ABOUT THE DATA AND THE DATASET-

* The dataset is a real dataset containing the data from Makaan.com
* The dataset contains information only regarding tech-emerging cities in India (that is- Bangalore, Hyderabad, Chennai, Delhi, Mumbai, Kolkata and Lucknow)
* The data is already cleaned and doesn’t contain any null or inconsistent values.

Through this project, dataset is visualized and understand how the data can be processed using R programming language.

RESEARCH METHODOLOGY:

Data and information are the two most important pillars for any analysis, it is important that they are collected from verified sources for carrying out the analysis efficiently. In this project only secondary data is used.

Analysis is in entirely in R Studio, an IDE for R. R is a programming language which is primarily used for statistical analysis and data visualization. This analysis is limited to tech-emerging cities in India and primarily revolves around the objectives. Hence, relative datasets are taken out, modified and are analyzed for better understanding of the data.

METHODS:

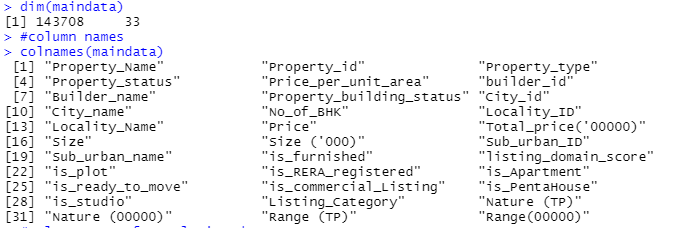
EXPLORATORY DATA ANALUSIS (EDA)- is used extensively to critically analyze the data to obtain the desired outputs. Data visualization tools (plots, charts) are used to better visualize the data.

* TREE MAP: Tree maps are used to present hierarchical data into a visualized from; file systems, structure of geographical arears, organizational structure etc. are only few examples. The large amount of different visualization approaches for trees speak to both importance of data type and its ability to spark the interest of the visualization.
* BAR PLOT: Bar plot (bar chat or graph) is used to represent values as bars in a graph. A bar plot is basically used for aggregating the categorical data.
* BOX PLOT: Box plot can be a good way to summarize the shape of a distribution, showing its Median, Inter-Quartile Range, Quartiles, Skewness and also possible outliers.
* CORRELATION ANALYSIS: Correlation analysis is a statistical technique which aims to establish whether a pair of variables are related or not. Correlation is how well data can be filled to a line ‘’scatter plot’’ is used to represent the data.
* SIMPLE REGRESSION: Simple regression is a statistical technique of estimating or predicting the unknown value of a dependent variables from a known value of an independent variable. X is usually the independent variable and Y is the dependent variable. Only two variables are studies in simple regression.
* MULTIPLE REGRESSION: When more than two variables are studied and their relationships are simultaneously worked out it is a case of multiple regression.

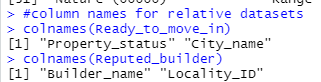
DATA SOURCE:

Data is collected from “MAKAAN.COM”. Link used for Downloading the data is – <https://iiitaphyd-mysharepoint.com>

Dataset is named as ‘maindata’ and it contains total of 143708 rows and 33 columns.



For better understanding of the data two other relative datasets are drawn from the above data.



ANALYSIS:

**OBJECTIVE:** TO ANALYSE THE DATA TO KNOW NUMBER OF READY TO MOVE PROPERTIES FOR LEASING IN TECH-EMERGING CITIES.

CODE:

#assigning table to a variable

RTMI <- table(maindata$is\_ready\_to\_move)

#barplot for no.of ready to move in properties

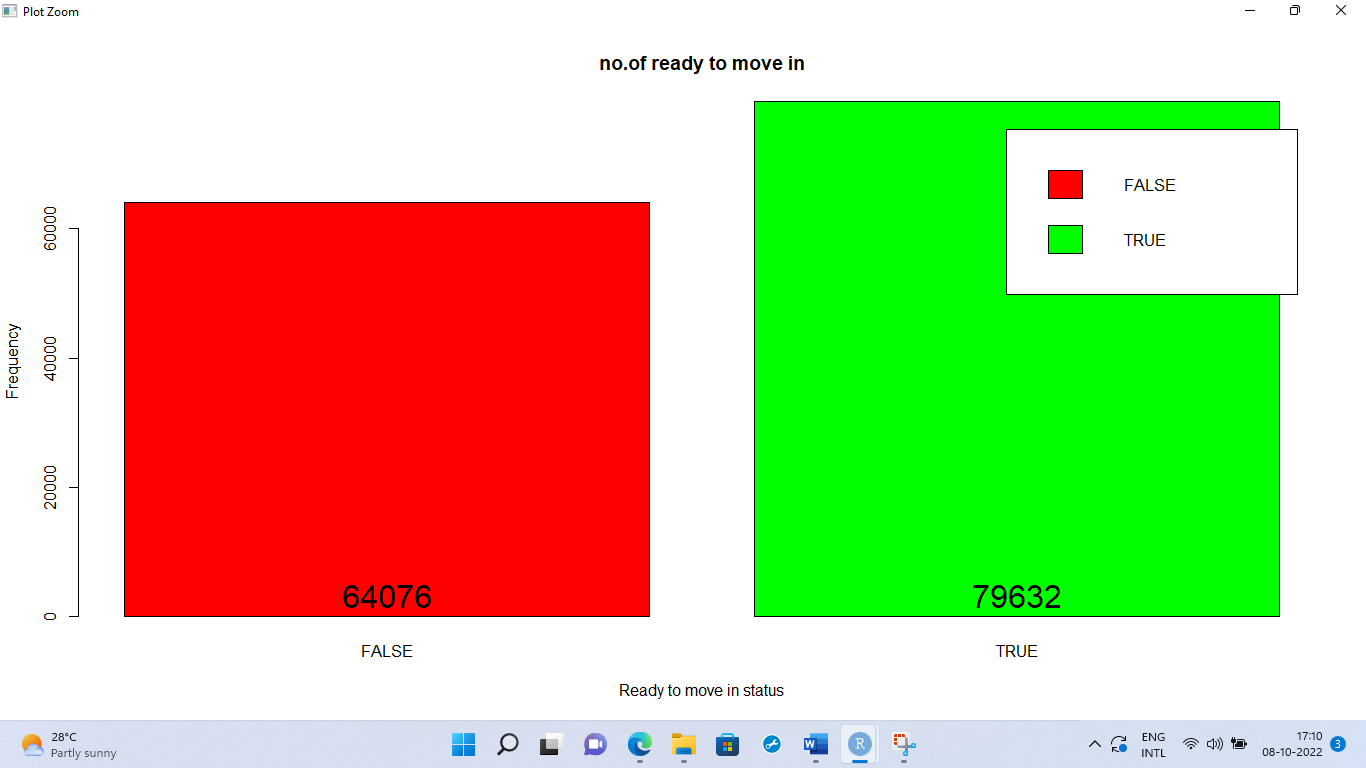
barplot(RTMI, main = 'no.of ready to move in', xlab = 'Ready to move in status', ylab = 'Frequency', col = c('RED','GREEN'),legend.text = TRUE )

show<- barplot(RTMI, main = 'no.of ready to move in', xlab = 'Ready to move in status', ylab = 'Frequency', col = c('RED','GREEN'),legend.text = TRUE )

#for numbers in each bar

text(show, 0, RTMI,cex = 2, pos = 3)

OUTPUT:



CODE:

#importing ready to move in excel file

#table() for property status and city name

table(Ready\_to\_move\_in$Property\_status, Ready\_to\_move\_in$City\_name)

#assigning Y variable to a variable

y <- table(Ready\_to\_move\_in$City\_name)

#assinging different colors to different cities

colours <- c("Blue","Yellow","Red","Orange","Green","pink","Brown","Grey")

barplot(y, xlab = 'City names', ylab = 'no. od ready to move in properties', main = "Number of ready to move in properties in each city", col = colours)

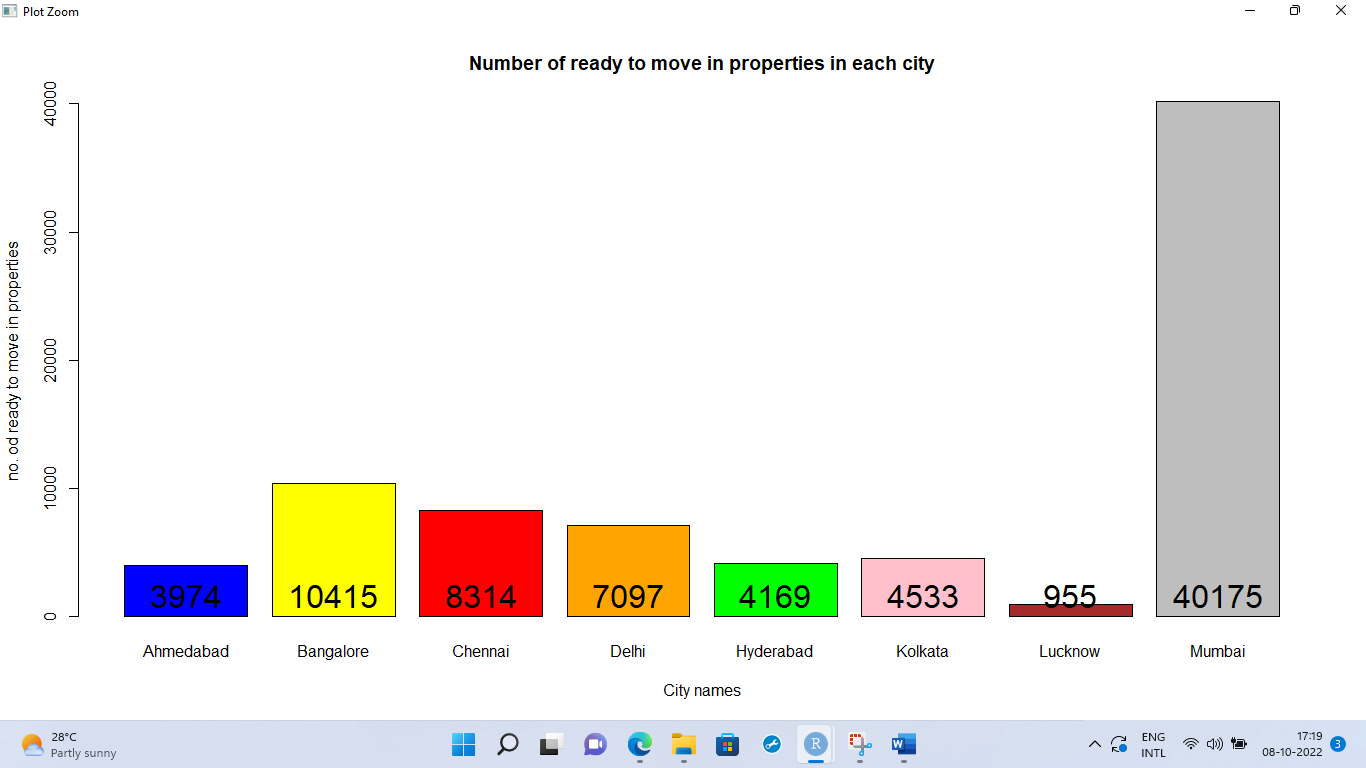
#assinging plot to a variable

final <- barplot(y, xlab = 'City names', ylab = 'no. od ready to move in properties', main = "Number of ready to move in properties in each city", col = colours)

#numbers against each bar

text(final, 0, y,cex = 2, pos = 3)

OUTPUT:



**OBJECTIVE:** TO PROVIDE BUDGET INSIGHTS AND DIVIDING THE BUFGET RANGE INTO THREE DIFFERENT GROUPS (HIGH, MID, LOW)

CODE:

#assigning table to a variable

LMH <- table(maindata$`Nature (00000)`)

#defining colours for bar plot

colours = c('Red','Blue','Green')

#barplot()

barplot(LMH, main = 'Budget range', xlab='Price nature', ylab='frequency',col = colours,legend.text = TRUE)

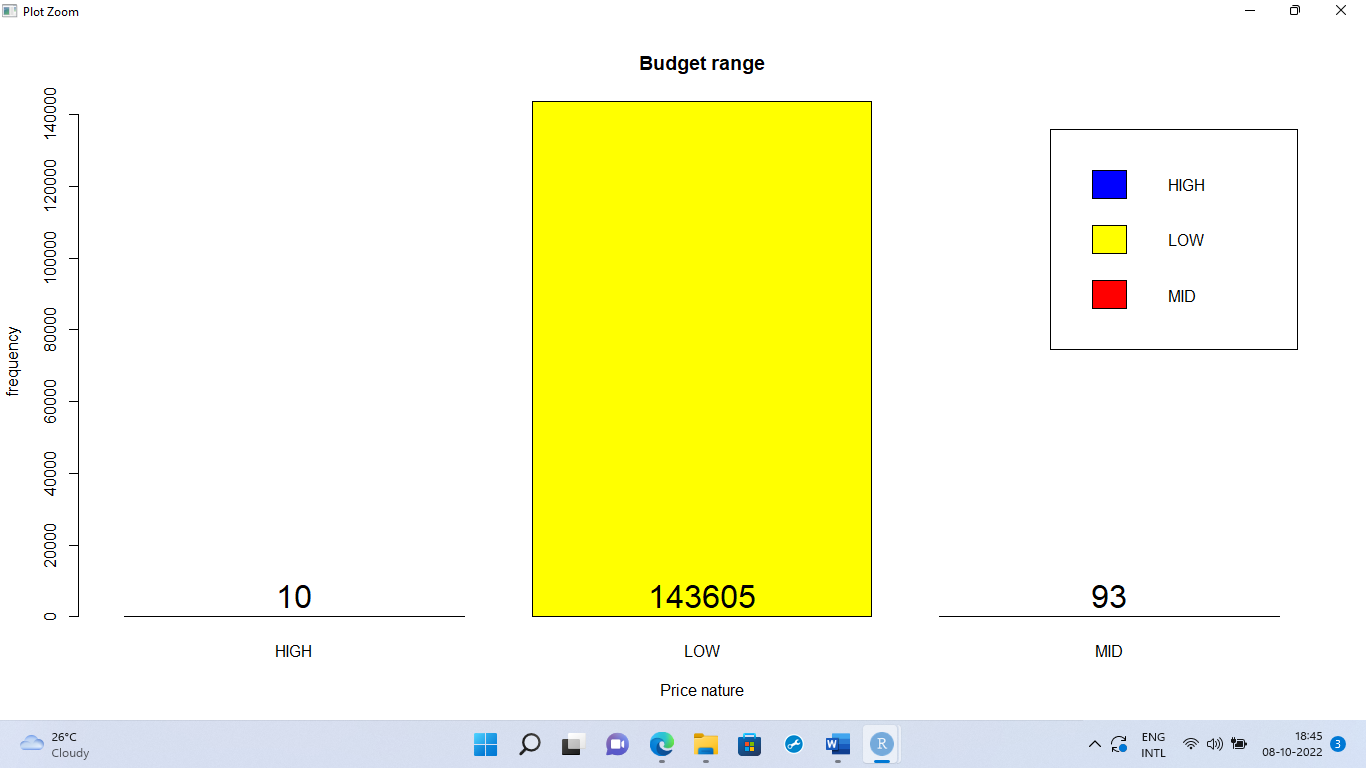
#assigning to a variable PR

PR <- barplot(LMH, main = 'Budget range', xlab='Price nature', ylab='frequency',col = colours,legend.text = TRUE)

#numbers gainst the bars

text(PR, 0, LMH,cex = 2, pos = 3)

OUTPUT:



CODE:

#summary()

summary(maindata$Price\_per\_unit\_area)

#plot one-price per unit

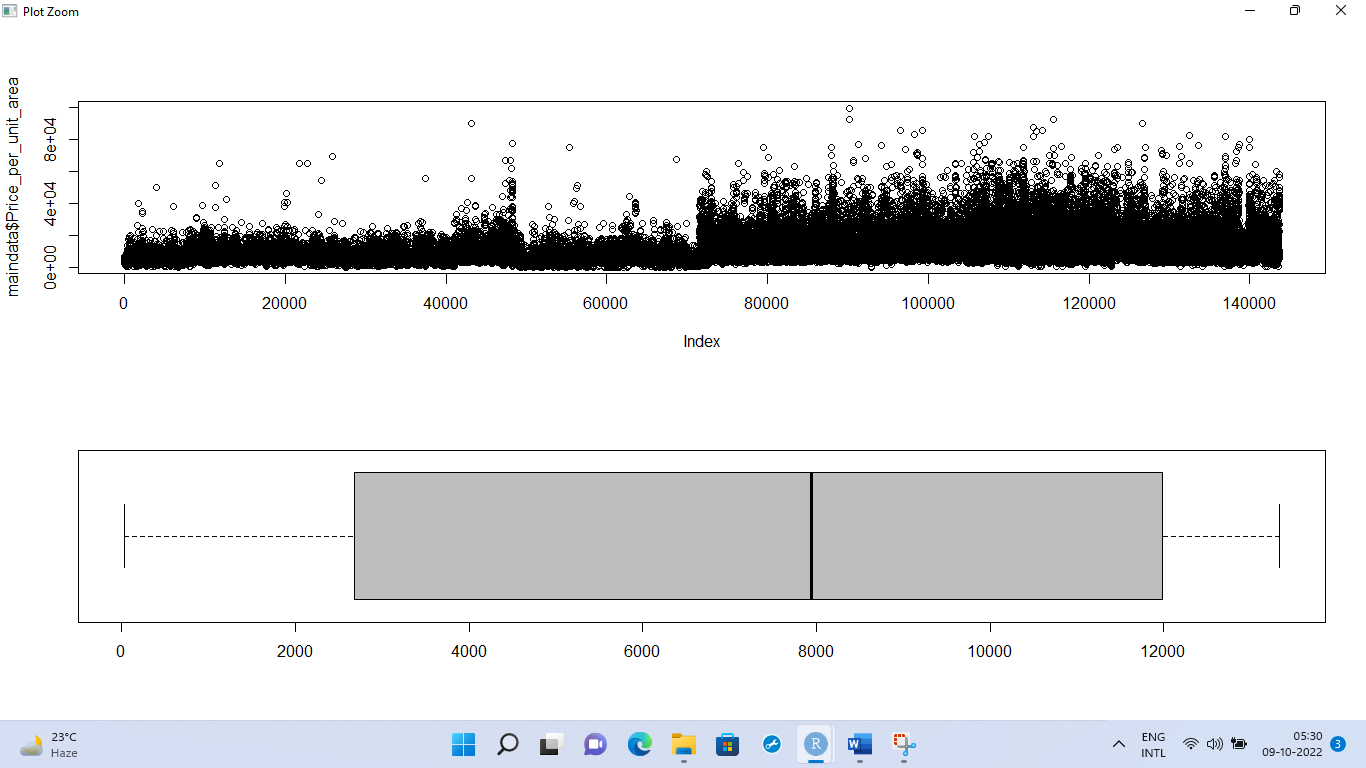
par(mfcol = c(2,1))

plot(maindata$Price\_per\_unit\_area)

boxplot(summary(maindata$Price\_per\_unit\_area), varwidth = TRUE, notch = FALSE, col = 'GREY', horizontal = TRUE, outline = FALSE)

OUTPUT:





CODE:

#summary()

summary(maindata$`Total\_price('00000)`)

#plot two- Price

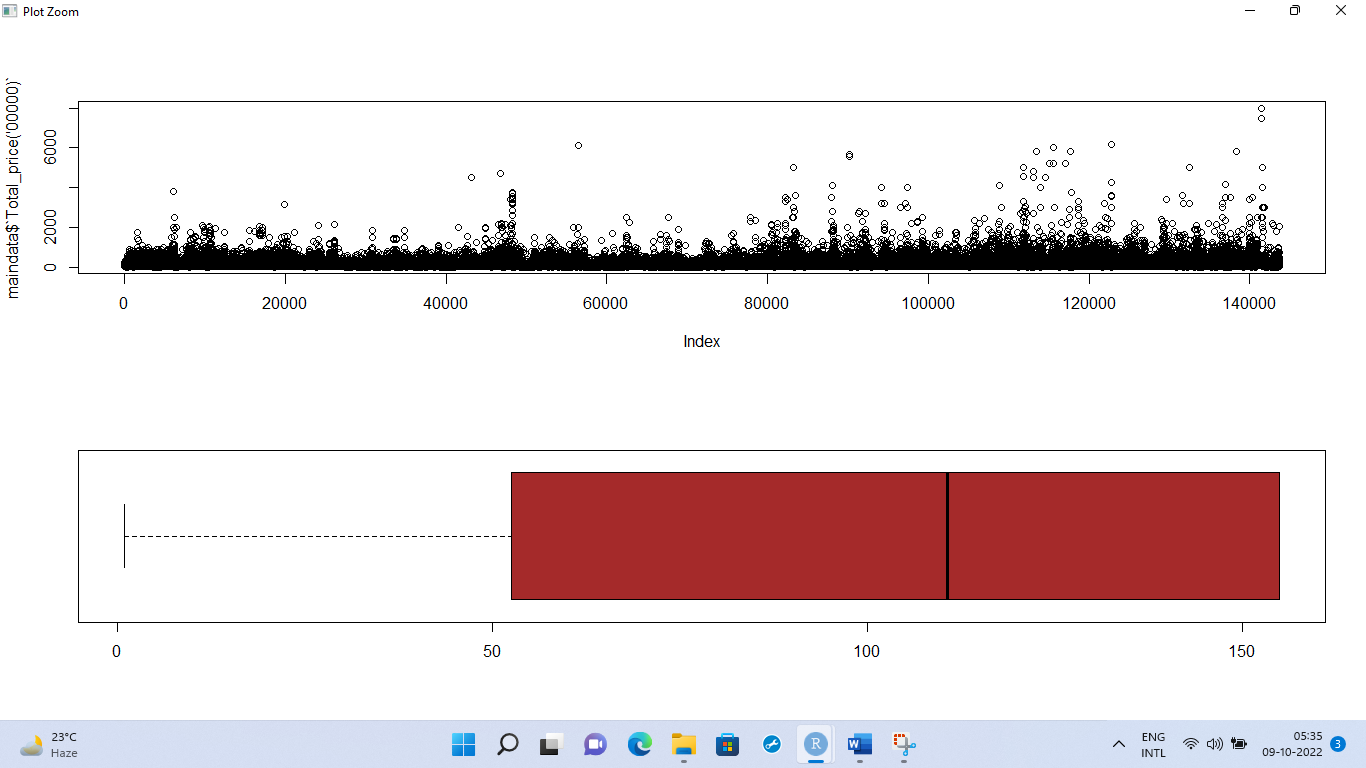
par(mfcol = c(2,1))

plot(maindata$`Total\_price('00000)`)

boxplot(summary(maindata$`Total\_price('00000)`), varwidth = TRUE, notch = FALSE, col = 'BROWN', horizontal = TRUE, outline = FALSE)

OUTPUT:





CODE:

#setting one image per frame

par(mfcol = c(1,1))

#boxplot by price and city

boxplot(maindata$`Total\_price('00000)`~maindata$City\_name,

main = 'price-city boxplot',

xlab = 'City\_names',

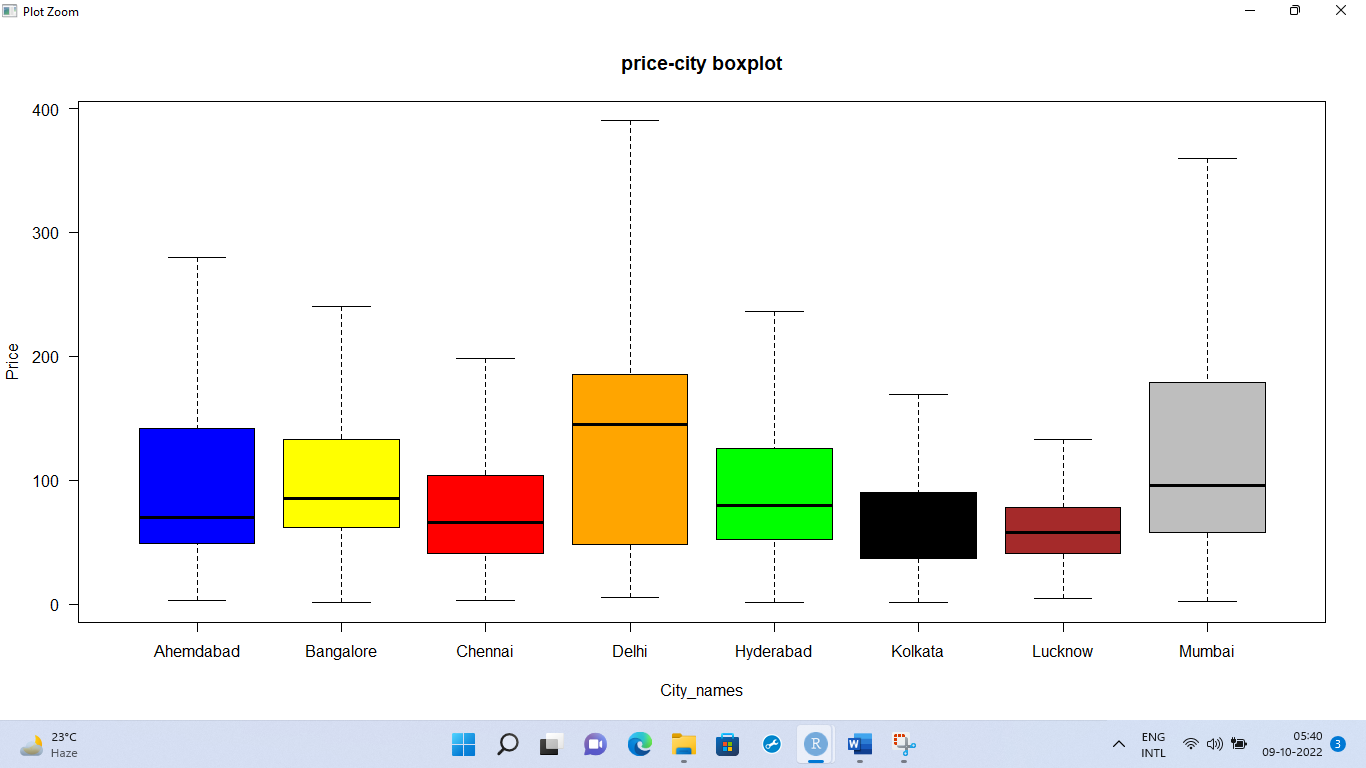
ylab = 'Price', outline = FALSE,

las = 1, col = colours,

names = c('Ahemdabad','Bangalore','Chennai','Delhi','Hyderabad','Kolkata','Lucknow','Mumbai')

)

OUTPUT:



**OBJECTIVE:** TO ANALYSE THE DATASET TO KNOW THE POTENTIAL LOCALITIES FOR INVESTORS WHO WANT TO MAKE ALL THE INVESTMENT AT ONE PLACE AND FOR OTHERS IN ANOTHER PLACE.

CODE:

#treemap

install.packages('treemap')

library(treemap)

#treemap

treemap(maindata,

index=c("City\_name","Locality\_Name","Sub\_urban\_name"),

vSize = "Price",

type = "index",

title = 'To make all investments in one place')

#TABLE()

table(maindata$City\_name)

OUTPUTS:



**OBJECTIVE**: TAKING THE LOCALITY OF BUILDER NAME ‘’REPUTED BUILDER’, TO KNOW THE COUNT OF NUMBER OF PROPERTIES IN THE SAME LOCALITY

CODE:

#importing the reputed builder excel file

#table()

table(Reputed\_builder$Locality\_ID)

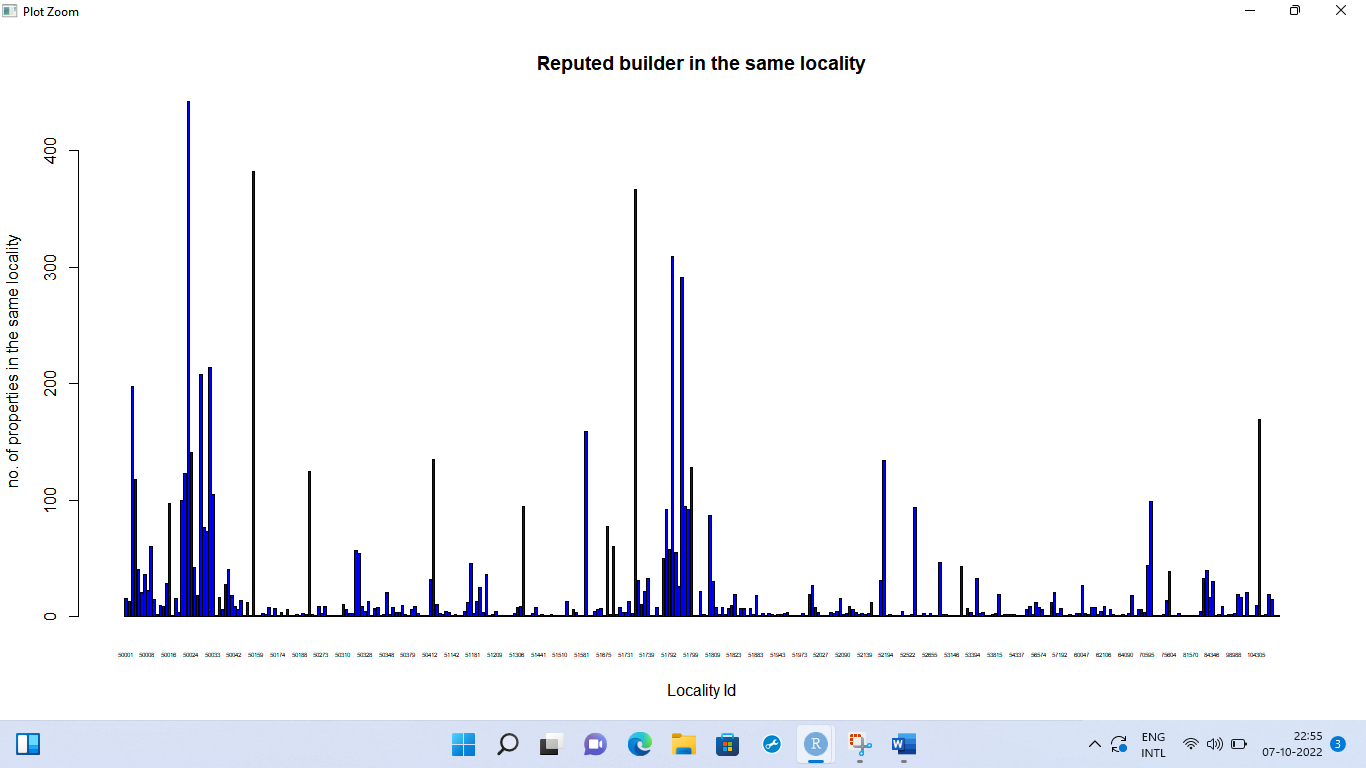
#assigning

x <- table(Reputed\_builder$Locality\_ID)

#barplot

barplot(x, xlab = 'Locality Id',ylab = "no. of properties in the same locality", main = "Reputed builder in the same locality", col = 'blue', cex.names = 0.4)

OUTPUT:



**OBJECTIVE:** TO KNOW ABOUT THE SIZE INSIGHTS

CODE:

#summary()

summary(maindata$`Size ('000)`)

#plot three- size

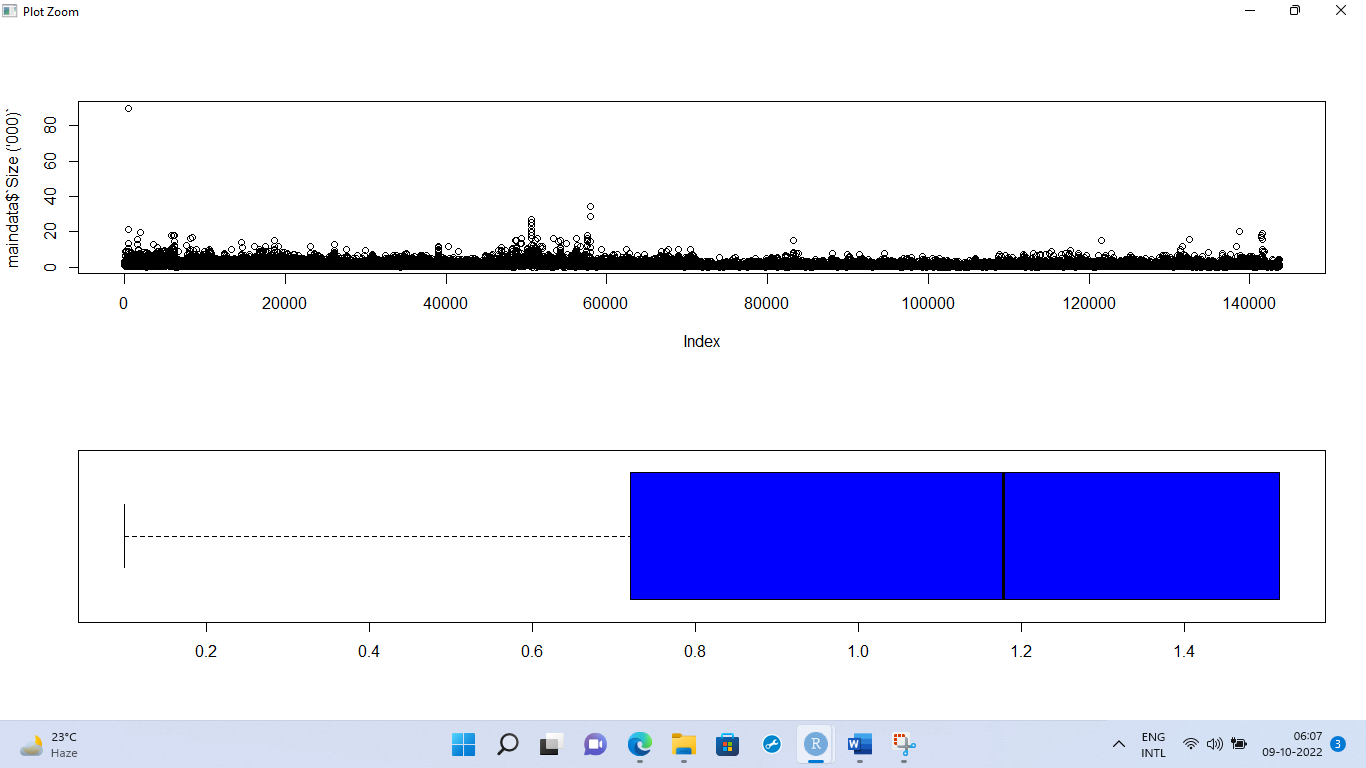
par(mfcol = c(2,1))

plot(maindata$`Size ('000)`)

boxplot(summary(maindata$`Size ('000)`), varwidth = TRUE, notch = FALSE, col = 'BLUE', horizontal = TRUE, outline = FALSE)

OUTPUT:





**OBJECTIVE:** TO KNOW THE CORRELATION BETWEEN PRICE AND THE SIZE

CODE:

#correlation

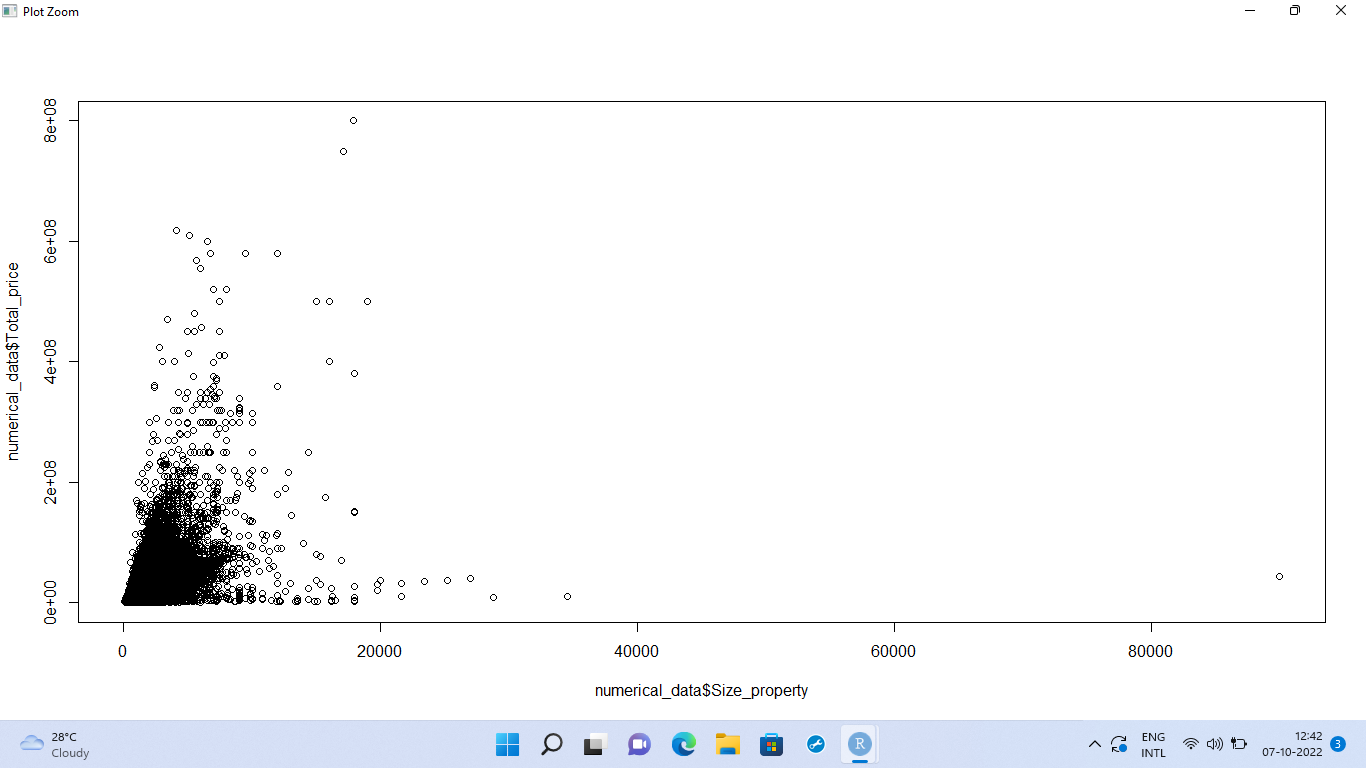
cor(maindata$`Size ('000)`, maindata$`Total\_price('00000)`)

#plotting the correlation

plot(maindata$`Size ('000)`, maindata$`Total\_price('00000)`,main = 'Plot for studying relation between price and size', xlab = 'Size' , ylab = 'Price')

OUTPUTS:





**OBJECTIVE:** TO FIND SUM OF DIFFERENT TYPES OF PROPERTIES IN EACH CITY AND ALSO OVERALL

CODE:

#assinging to a variable

type <- table(maindata$Property\_type)

#barplot()

barplot(type, main = 'Different types of properties',xlab = 'type of property', ylab = 'frequency',legend.text = TRUE,col = c('Pink','lightblue','brown','Grey','yellow'))

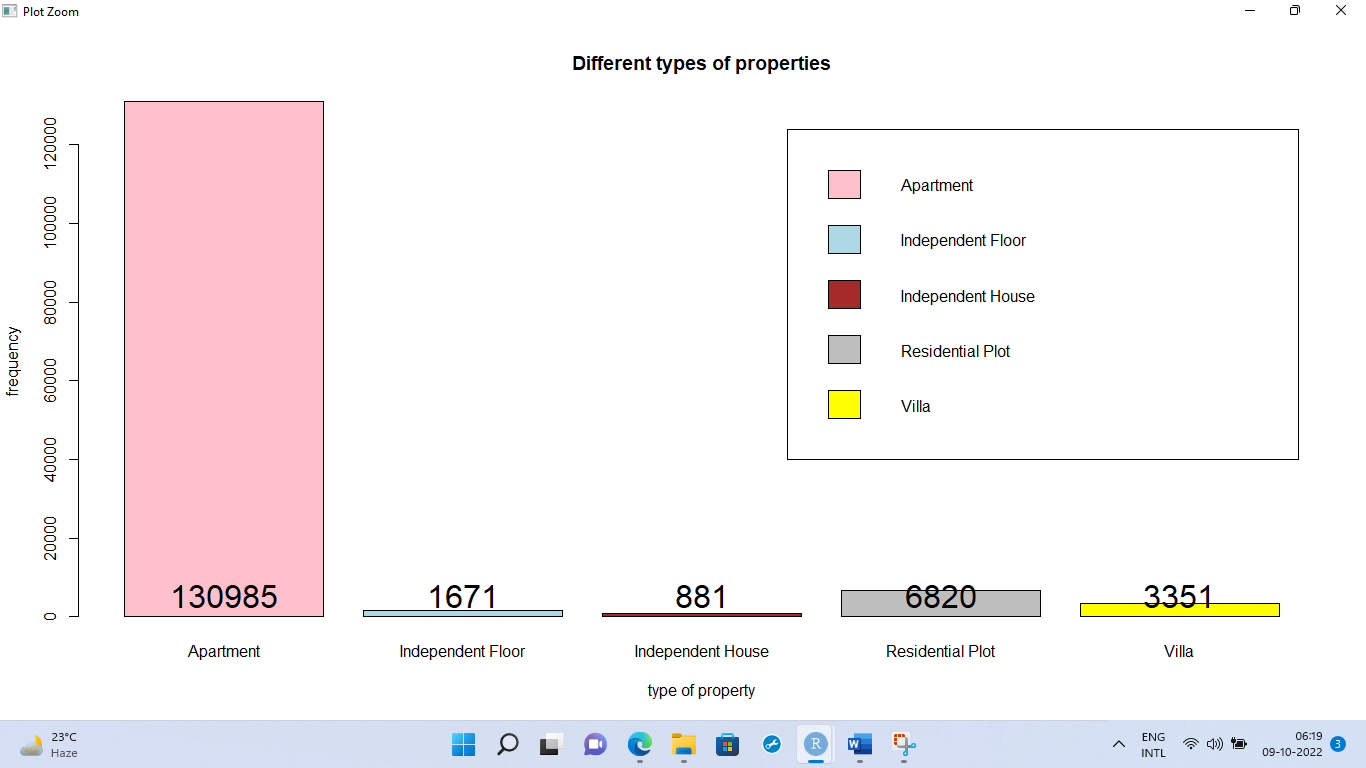
#assigning to a variable

DP <- barplot(type, main = 'Different types of properties',xlab = 'type of property', ylab = 'frequency',legend.text = TRUE,col = c('Pink','lightblue','brown','Grey','yellow'))

#for numbers against each bar

text(DP, 0, type, cex = 2, pos = 3)

OUTPUT:



CODE:

install.packages('ggplot2')

install.packages('ggpubr')

library(ggplot2)

library(ggpubr)

#barplot

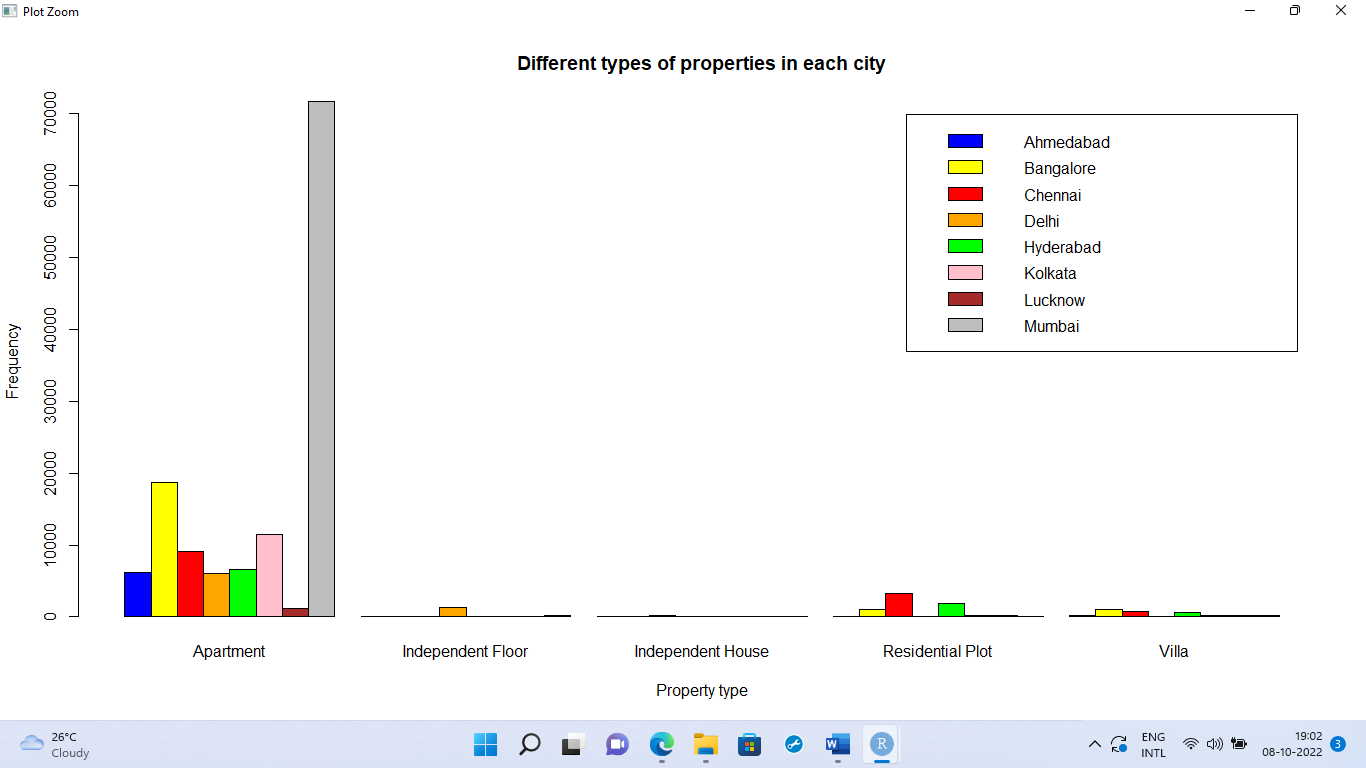
barplot(table(maindata$City\_name, maindata$Property\_type),main = 'Different types of properties in each city',

col = colours,beside = TRUE,

legend.text = TRUE, ylab = 'Frequency', xlab = 'Property type')

table(maindata$City\_name, maindata$Property\_type)

OUTPUT:



**OBJECTIVE:** TO KNOW THE NUMBER OF FURNISHED, SEMI-FURNISHED, UNFURNISHED HOUSES IN EACH CITY

CODE:

#table() city name and furniture

table(maindata$City\_name, maindata$is\_furnished)

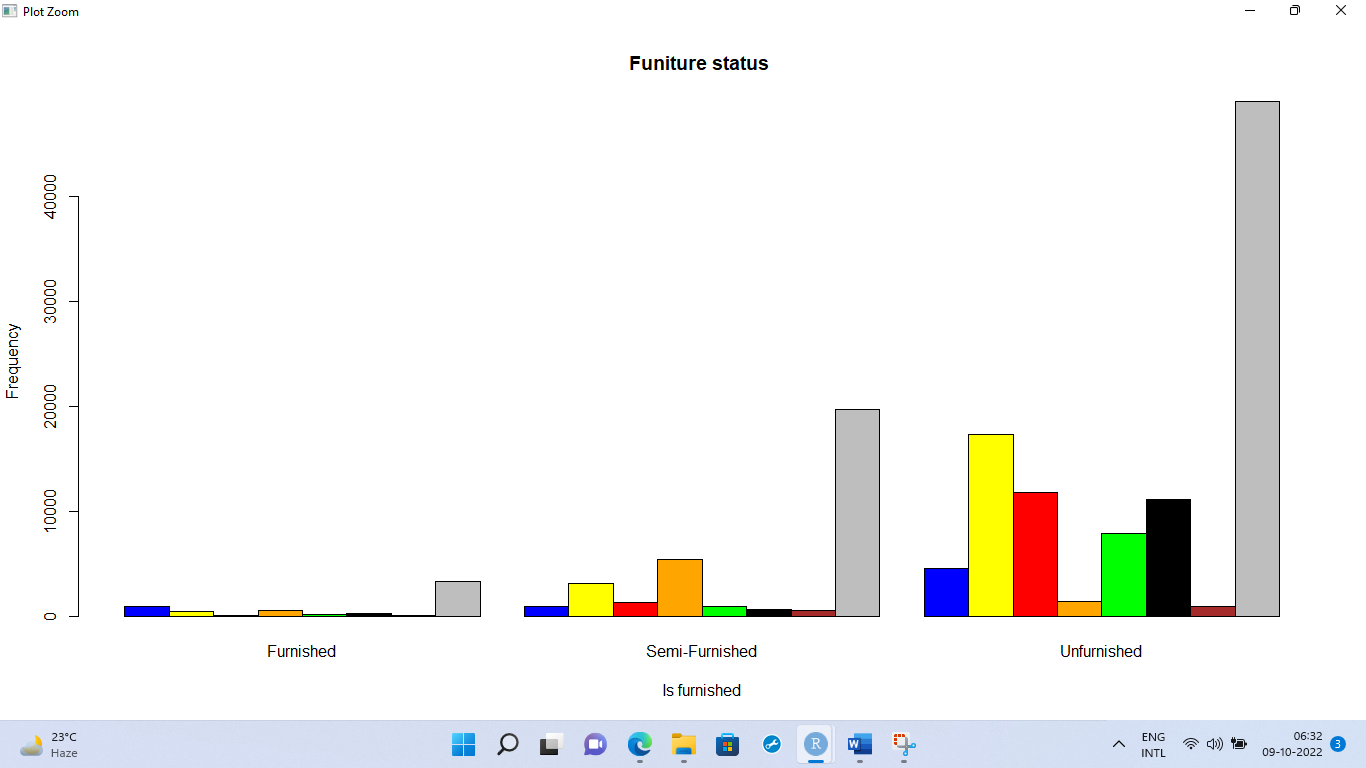
#barplot()

barplot(table(maindata$City\_name, maindata$is\_furnished),main = 'Funiture status ',

col = colours,beside = TRUE,

legend.text = FALSE, ylab = 'Frequency', xlab = 'Is furnished')

OUTPUT:



**OBJECTIVE:** TO KNOW OTHER INSIGHTS ABOUT LISTINGS,

CODE:

par(mfcol = c(2,2))

#is plot

table(maindata$is\_plot)

barplot(table(maindata$is\_plot), main = 'plots', ylab = 'Frequency', xlab = 'plot', col = c('RED', 'GREEN'), legend.text = FALSE)

#penthouse

table(maindata$is\_PentaHouse)

barplot(table(maindata$is\_PentaHouse), main = 'Penthouse', ylab = 'Frequency', xlab = 'Pent house ', col = c('RED', 'GREEN'), legend.text = FALSE)

#studio

table(maindata$is\_studio)

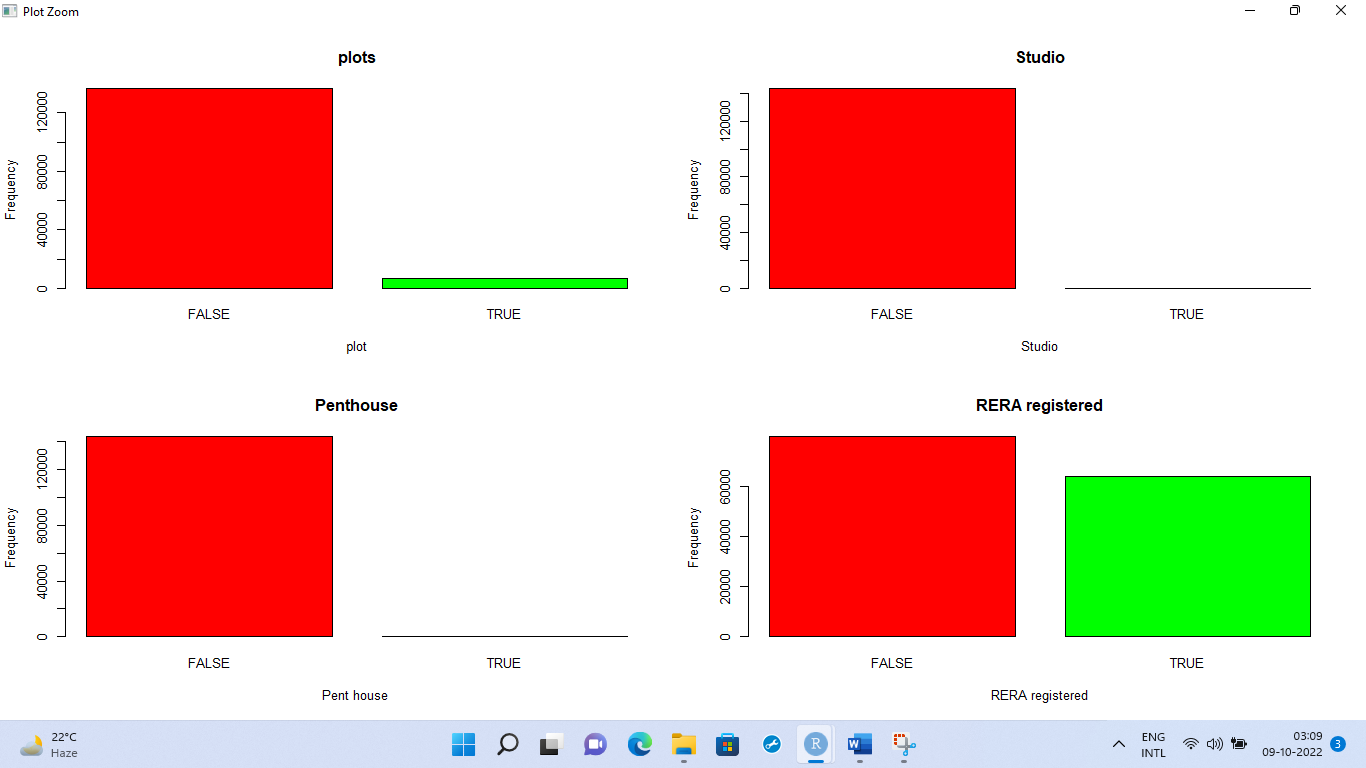
barplot(table(maindata$is\_PentaHouse), main = 'Studio', ylab = 'Frequency', xlab = 'Studio', col = c('RED', 'GREEN'), legend.text = FALSE)

#RERA Registered

table(maindata$is\_RERA\_registered)

barplot(table(maindata$is\_RERA\_registered), main = 'RERA registered', ylab = 'Frequency', xlab = 'RERA registered', col = c('RED', 'GREEN'), legend.text = FALSE)

OUTPUT:



CODE:

#Property status

table(maindata$Property\_status)

barplot(table(maindata$Property\_status), main = 'Property status', ylab = 'Frequency', xlab = 'Property status', col = c('RED', 'GREEN'), legend.text = FALSE)

#no of BHK

table(maindata$No\_of\_BHK)

barplot(table(maindata$No\_of\_BHK), main = 'no.of BHKs', ylab = 'Frequency', xlab = 'number', col = colours)

#city name

table(maindata$City\_name)

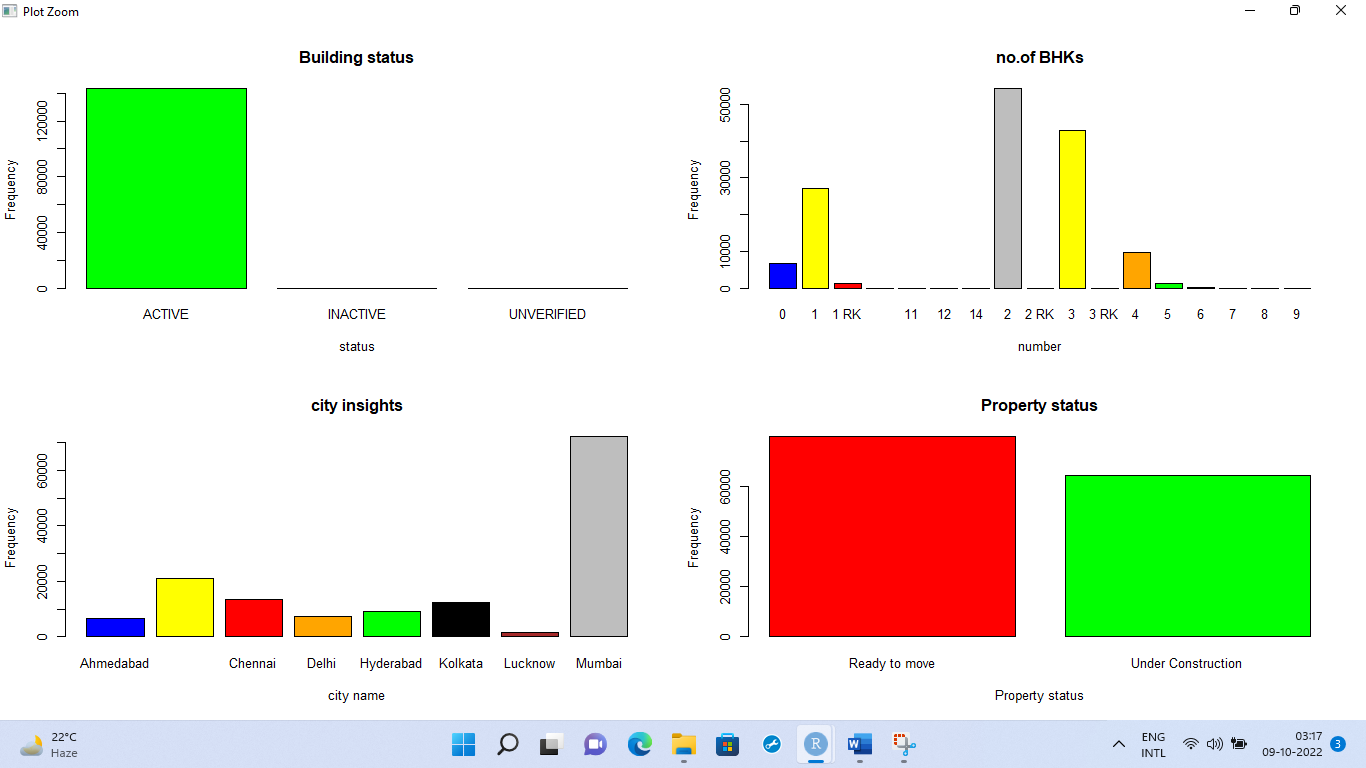
barplot(table(maindata$City\_name), main = 'city insights', ylab = 'Frequency', xlab = 'city name', col = colours)

#property building status

table(maindata$Property\_building\_status)

barplot(table(maindata$Property\_building\_status), main = 'Building status', ylab = 'Frequency', xlab = 'status', col = c('GREEN', 'RED', 'BLUE'))

OUTPUT:



CODE:

#tables by each city

table(maindata$Property\_building\_status, maindata$City\_name)

table(maindata$No\_of\_BHK, maindata$City\_name)

table(maindata$Property\_status, maindata$City\_name)

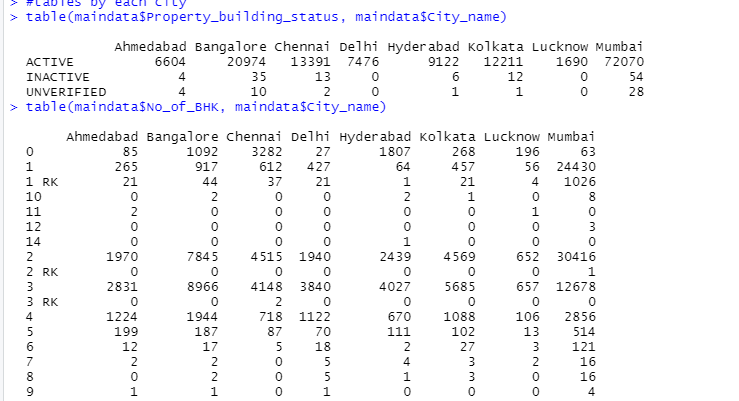
table(maindata$is\_RERA\_registered, maindata$City\_name)

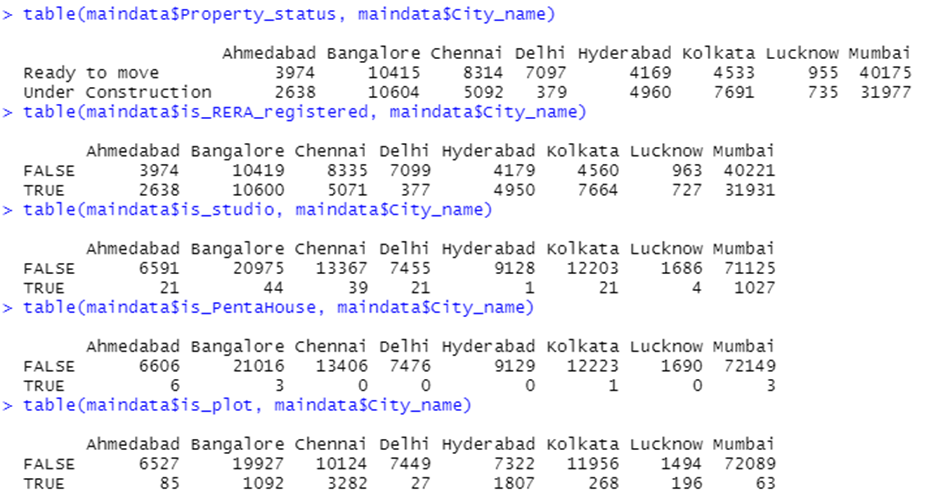
table(maindata$is\_studio, maindata$City\_name)

table(maindata$is\_PentaHouse, maindata$City\_name)

table(maindata$is\_plot, maindata$City\_name)

OUTPUTS:





**OBJECTIVE:** TO FIND THE RELATIONSHIP BETWEEN SIZE AND THE PRICE

CODE:

SR <- lm(maindata$`Total\_price('00000)` ~ maindata$`Size ('000)`)

summary(SR)

plot(SR)

OUTPUTS:

