MOSCOW REAL ESTATE ANALYSIS

Abhishikt Emmanuel Prakash

```
# Guided by
# Christ (Deemed to be University) Bengaluru
# Department of Statistics:
# Ms. Kavitha R, Mr. Dibu A
# Bangalore - 29

# The secondary data set for this statistical study has been obtained from
Kaggel.
# This data set includes information about one room apartment real estate in
Moscow
# The aim of this statistical analysis is to make use of various R
functionalities
# such as plot, bar plot, histogram, skewness, kurtosis, standard deviation
and
# the summary function to study the real estate market in Moscow.

# This data set includes the following variables:
# metro (nominal) : The nearest metro station to to the apartment
# price (numerical) : The rent price for the apartment
# way (nominal) : Mode of transportation to reach metro station (on foot or
by public transport)
# views (numerical) : The number of views for each apartment
# provider (nominal) : A person or agency who is renting apartment
# fee_percent (numerical) : Fee percent of an agency or realtor
# storey (numerical) : The storey, where the apartment located
# minutes (numerical) : Time to reach nearest metro station
# storeys (numerical) : The total number of storeys in a building
# living_area (numerical) : Square foot of the living area
# kitchen_area (numerical) : Square foot of the kitchen area
# total_area (numerical) : Total square of each apartment
# total_area_description (ordinal) : The size of apartment is Large, Medium
or small
# distance_description (ordinal) : The distance from the metro station is far
or near
library(readxl)# import read excel library
Data_temp <-
read_excel("/Users/abhishikt_mac/Downloads/real_estate_moscow.xlsx")# import
excel data set
data <- data.frame(Data_temp)# create data frame
library(ggplot2)# import ggplot2 library
library(moments) # import the moments library
attach(data)
# Basic operations on all the columns in the data set------------------------
---
```

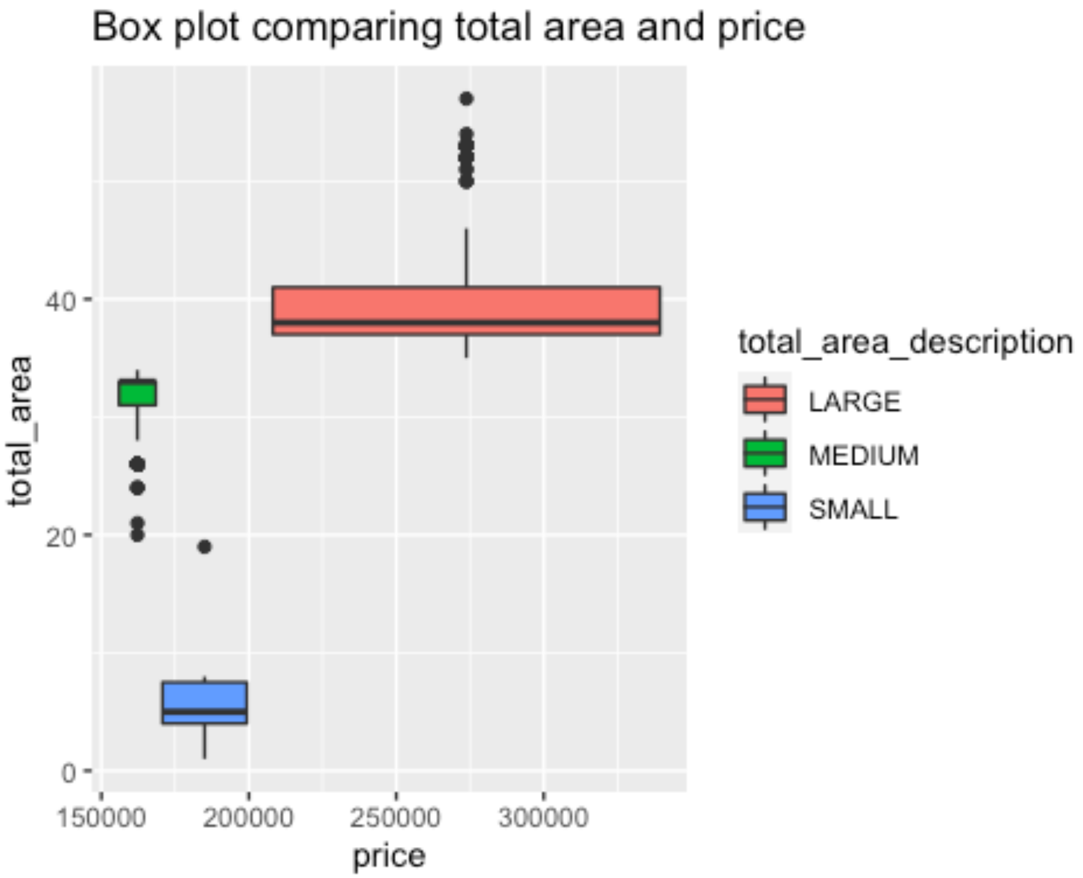
```
#summary of the data set i.e for each variable
summary(data)
```

```
##       S.No          metro               price              way
##  Min.   :   0.0   Length:1446        Min.   : 14000   Length:1446
##  1st Qu.: 361.2   Class :character   1st Qu.: 29000   Class :character
##  Median : 722.5   Mode  :character   Median : 38000   Mode  :character
##  Mean   : 722.5                      Mean   : 43771
##  3rd Qu.:1083.8                      3rd Qu.: 45000
##  Max.   :1445.0                      Max.   :500000
##      views          provider          fee_percent          storey
##  Min.   :   4.0   Length:1446        Min.   :  0.00   Min.   :  1.00
##  1st Qu.:  38.0   Class :character   1st Qu.:  0.00   1st Qu.:  4.00
##  Median : 103.0   Mode  :character   Median : 50.00   Median :  6.00
##  Mean   : 417.9                      Mean   : 37.95   Mean   :  7.09
##  3rd Qu.: 414.0                      3rd Qu.: 50.00   3rd Qu.:  9.00
##  Max.   :5174.0                      Max.   :100.00   Max.   :613.00
##    minutes          storeys          living_area     kitchen_area
##  Min.   : 0.000   Min.   :    1.00   Min.   : 6.00   Min.   : 3.00
##  1st Qu.: 5.000   1st Qu.:    9.00   1st Qu.:18.00   1st Qu.: 7.00
##  Median : 7.000   Median :   12.00   Median :20.00   Median :10.00
##  Mean   : 8.754   Mean   :   22.55   Mean   :20.59   Mean   :11.37
##  3rd Qu.:12.000   3rd Qu.:   16.00   3rd Qu.:21.00   3rd Qu.:10.00
##  Max.   :47.000   Max.   :13217.00   Max.   :37.00   Max.   :37.00
##   total_area    total_area_description distance_description
##  Min.   : 1.00   Length:1446            Length:1446
##  1st Qu.:34.00   Class :character       Class :character
##  Median :37.00   Mode  :character       Mode  :character
##  Mean   :37.27
##  3rd Qu.:40.00
##  Max.   :57.00
```

```
# Basic structure of the data set
str(data)
```

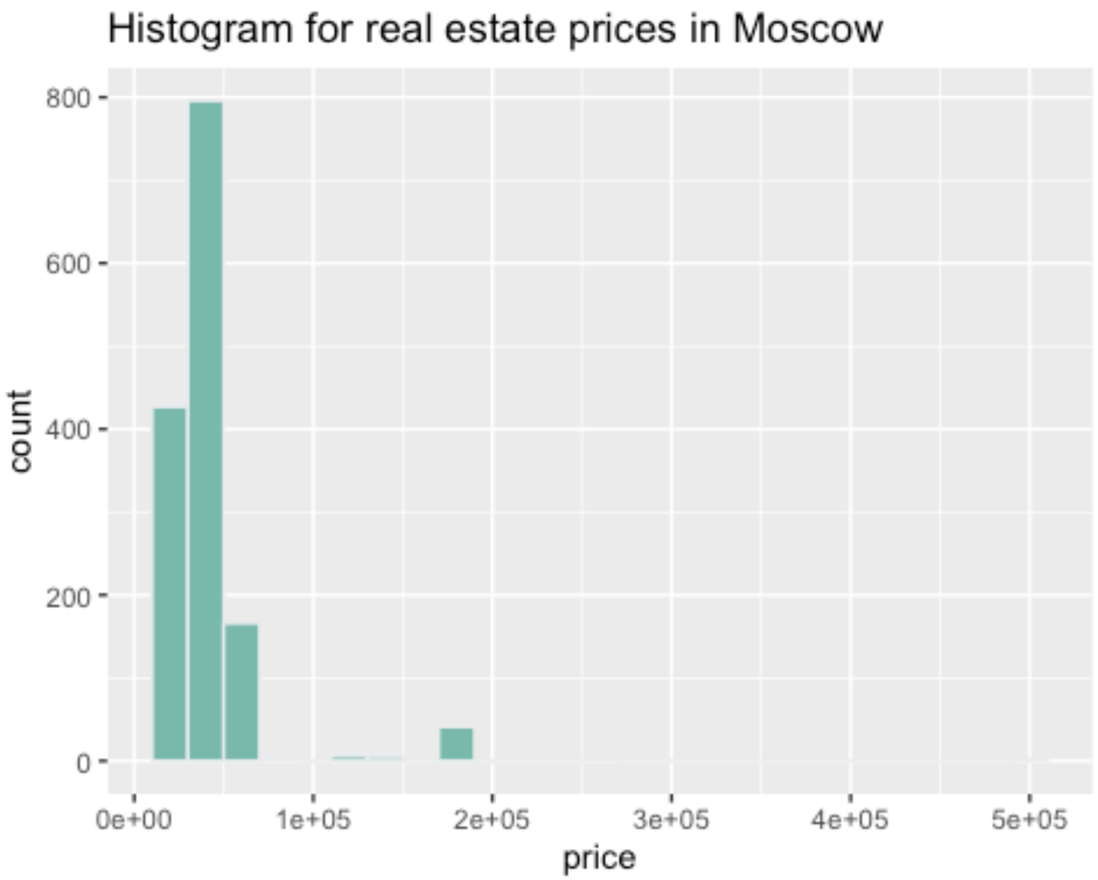
```
## 'data.frame':    1446 obs. of  15 variables:
##  $ S.No                  : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ metro                 : chr  "Planernaia" "VDNKh" "Alekseevskaia"
"Sviblovo" ...
##  $ price                 : num  45000 50000 50000 38000 55999 ...
##  $ way                   : chr  "walk" "walk" "walk" "walk" ...
##  $ views                 : num  513 389 483 414 360 ...
##  $ provider              : chr  "realtor" "realtor" "realtor" "realtor"
...
##  $ fee_percent           : num  50 50 50 50 99 40 40 50 50 50 ...
##  $ storey                : num  7 16 5 3 6 2 10 8 6 9 ...
##  $ minutes               : num  10 10 3 15 7 15 5 10 15 5 ...
##  $ storeys               : num  12 16 12 5 17 5 17 9 9 12 ...
##  $ living_area           : num  19 18 19 37 21 17 18 18 17 20 ...
##  $ kitchen_area          : num  8 8 5 37 10 7 11 7 6 10 ...
##  $ total_area            : num  38 41 33 37 40 31 41 33 34 35 ...
##  $ total_area_description: chr  "LARGE" "LARGE" "MEDIUM" "LARGE" ...
##  $ distance_description  : chr  "FAR" "FAR" "NEAR" "FAR" ...

# Get the top six rows of the data set
head(data)

##   S.No         metro price  way views provider fee_percent storey minutes
## 1    0    Planernaia 45000 walk   513  realtor          50      7      10
## 2    1         VDNKh 50000 walk   389  realtor          50     16      10
## 3    2 Alekseevskaia 50000 walk   483  realtor          50      5       3
## 4    3      Sviblovo 38000 walk   414  realtor          50      3      15
## 5    4      Rimskaia 55999 walk   360  realtor          99      6       7
## 6    5        Perovo 29000 walk  5174  realtor          40      2      15
##   storeys living_area kitchen_area total_area total_area_description
## 1      12          19            8         38                  LARGE
## 2      16          18            8         41                  LARGE
## 3      12          19            5         33                 MEDIUM
## 4       5          37           37         37                  LARGE
## 5      17          21           10         40                  LARGE
## 6       5          17            7         31                 MEDIUM
##   distance_description
## 1                  FAR
## 2                  FAR
## 3                 NEAR
## 4                  FAR
## 5                 NEAR
## 6                  FAR

#shape of the data set
dim(data)

## [1] 1446   15

# Metro stations--------------------------------------------------------------
----
# Summary of the variable metro i.e all the metro stations with the number of
real estate options near them
head(summary(factor(metro)))

##  Planernaia  Medvedkovo       VDNKh Rasskazovka    Altufevo  Nekrasovka
##         126          83          83          48          43          43

# Prices for each listed real estate------------------------------------------
----
# summary
summary(price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14000   29000   38000   43771   45000  500000
```

```
#box plot comparing total area and price
boxplot<-ggplot(data,aes(y=total_area, x = price, fill =
total_area_description))+geom_boxplot() + ggtitle("Box plot comparing total
area and price")
boxplot
```

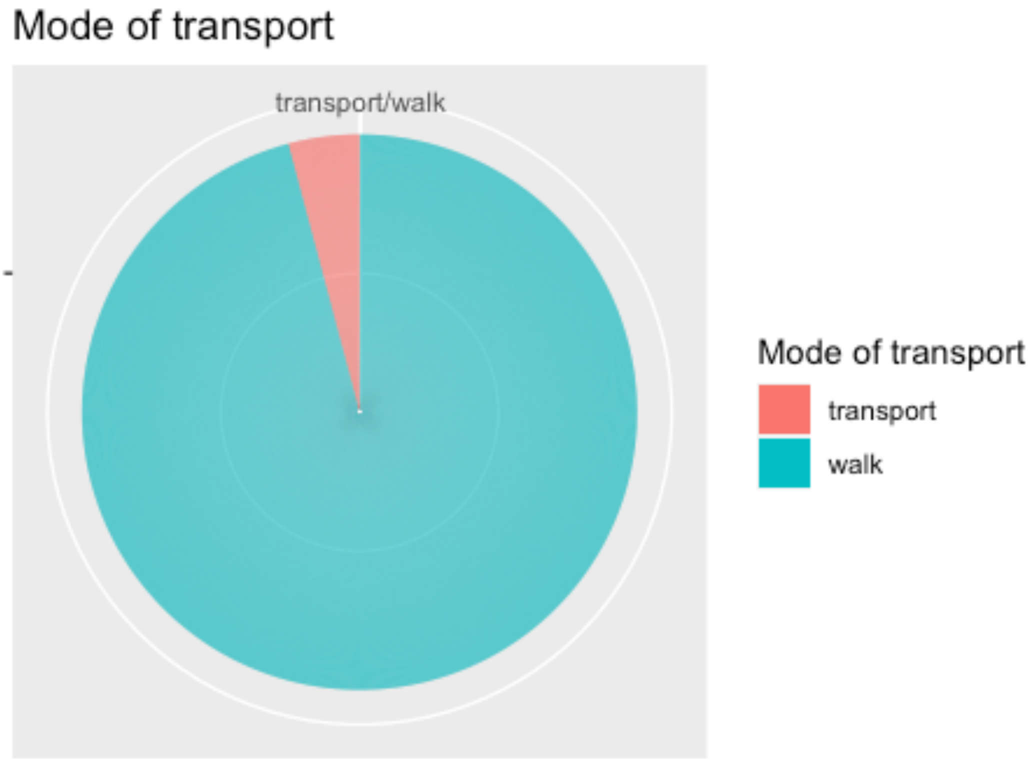


```
# Inference: Most of the real estate area for large properties is between
35sqft to 40sqft mostly below 40sqft which costs anywhere between 200K to
350K, while for medium properties size range don't vary much and are mostly
around 33sqft which costs around 150K to 175K, also the size for small
properties is less than 10sqft which costs around 175K to 200K.

# Now we know the minimum, maximum price for an apartment and also the mean
price, median price, 1st and 3rd quartile prices
sd(price)

## [1] 33232.15

# Such a large standard deviation indicates that the prices of real estate
vary highly form the mean
var(price)

## [1] 1104375895
```

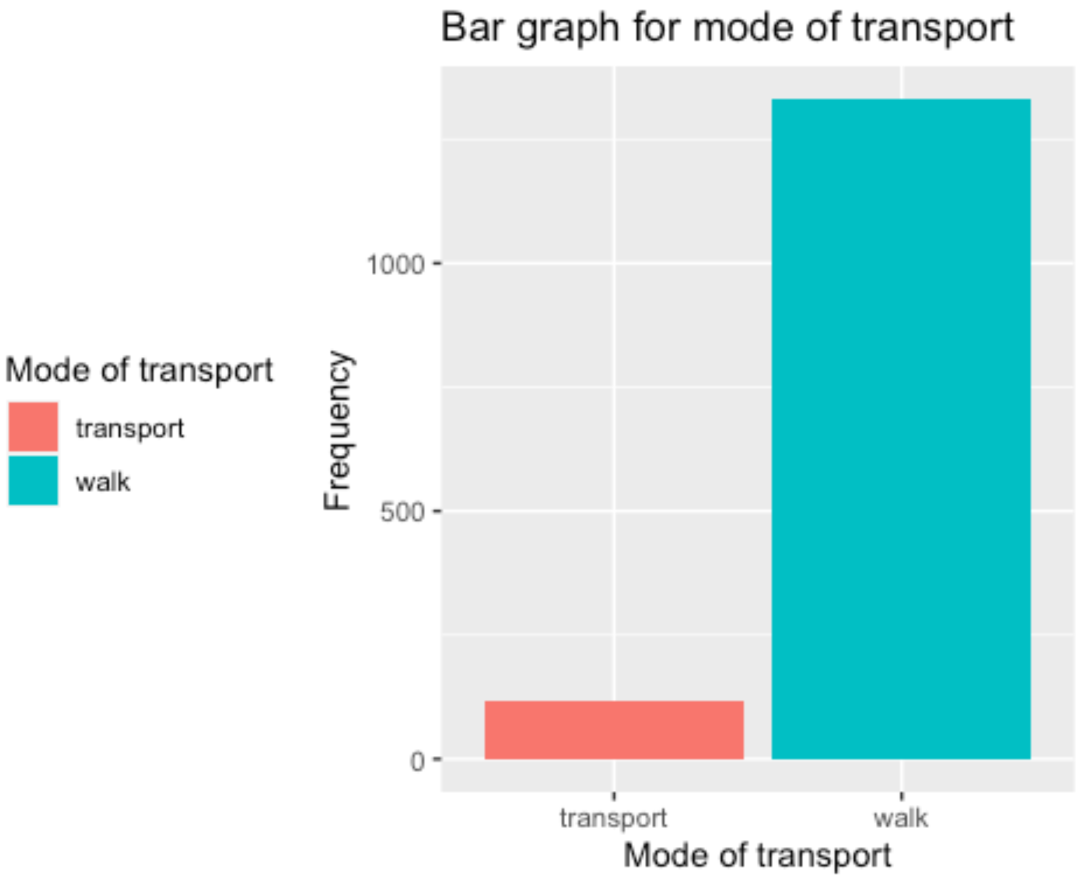
```
# Similarly a high varience is indicating that the prices of real estate in
Moscow vary highly from the mean and also there is a wast difference within
the prices.
#kurtosis
kurtosis(price)

## [1] 60.14116

# A positive kurtosis value represents that the data is highly peaked
#skewness
skewness(price)

## [1] 6.044986

# Since the value of skewness is 6.04 the data is highly skewed
# range of real estate prices in Moscow
range(price)

## [1]  14000 500000

# Histogram to visually represent the price distribution
ggplot(data, aes(x=price)) + geom_histogram( binwidth=20000, fill="#69b3a2",
color="#e9ecef", alpha=0.9) + ggtitle("Histogram for real estate prices in
Moscow")
```



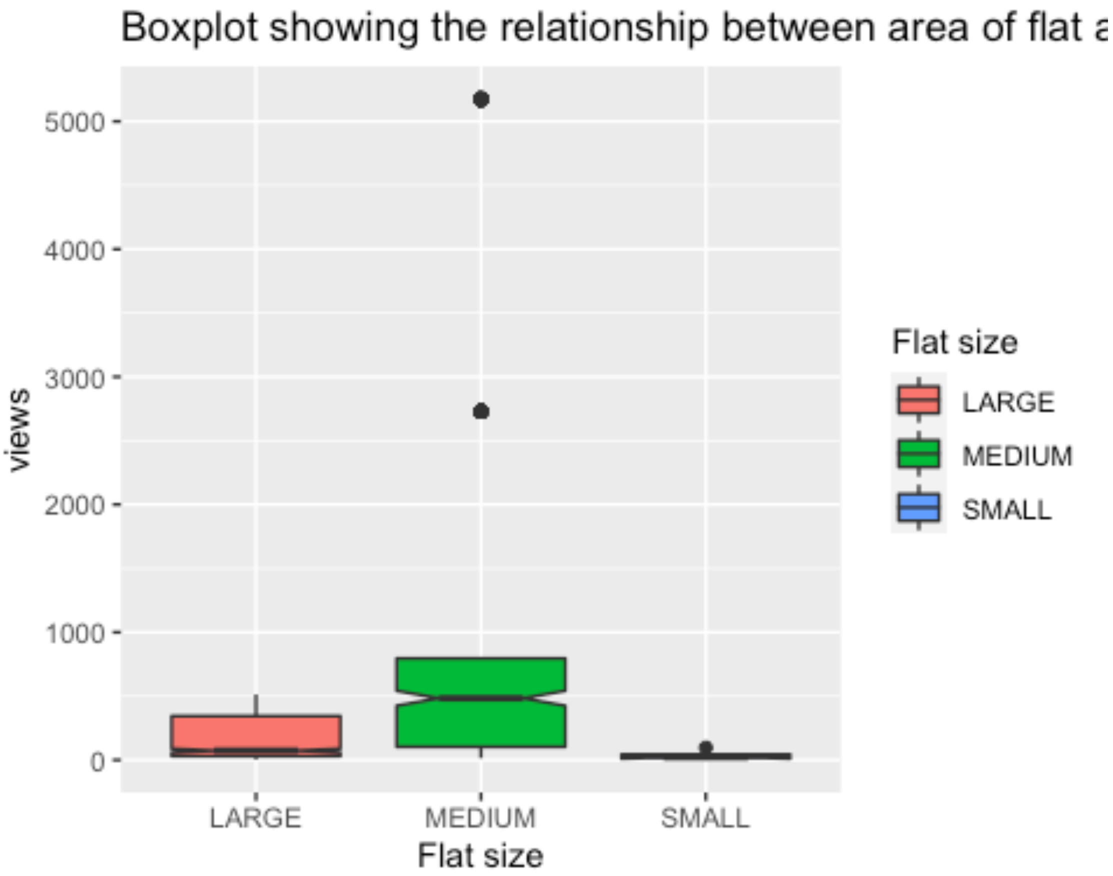
```
# Mode of transportation-----------------------------------------------------
---
summary(way)

##    Length     Class      Mode 
##      1446 character character 

ggplot(data, aes(x = "", y =way,fill = as.factor(way))) +
  geom_col() +
  coord_polar(theta = "y")+ggtitle("Mode of transport")+
  labs(x="",y="",fill="Mode of transport")
```



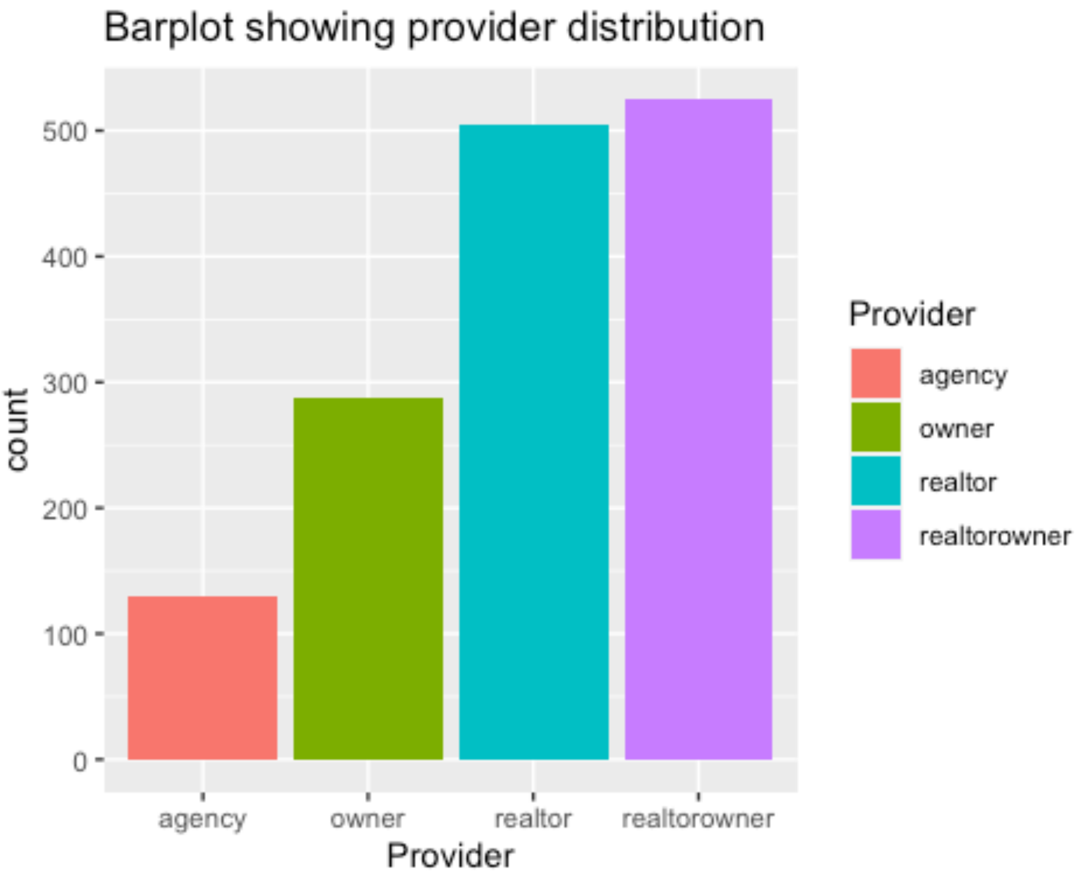
```
ggplot(data=data,aes(x=way,fill=as.factor(way)))+
  geom_bar(position='dodge')+ggtitle("Bar graph for mode of 
transport")+labs(x="Mode of transport",y="Frequency",fill="Mode of 
transport")+
  theme(legend.position='left')
```



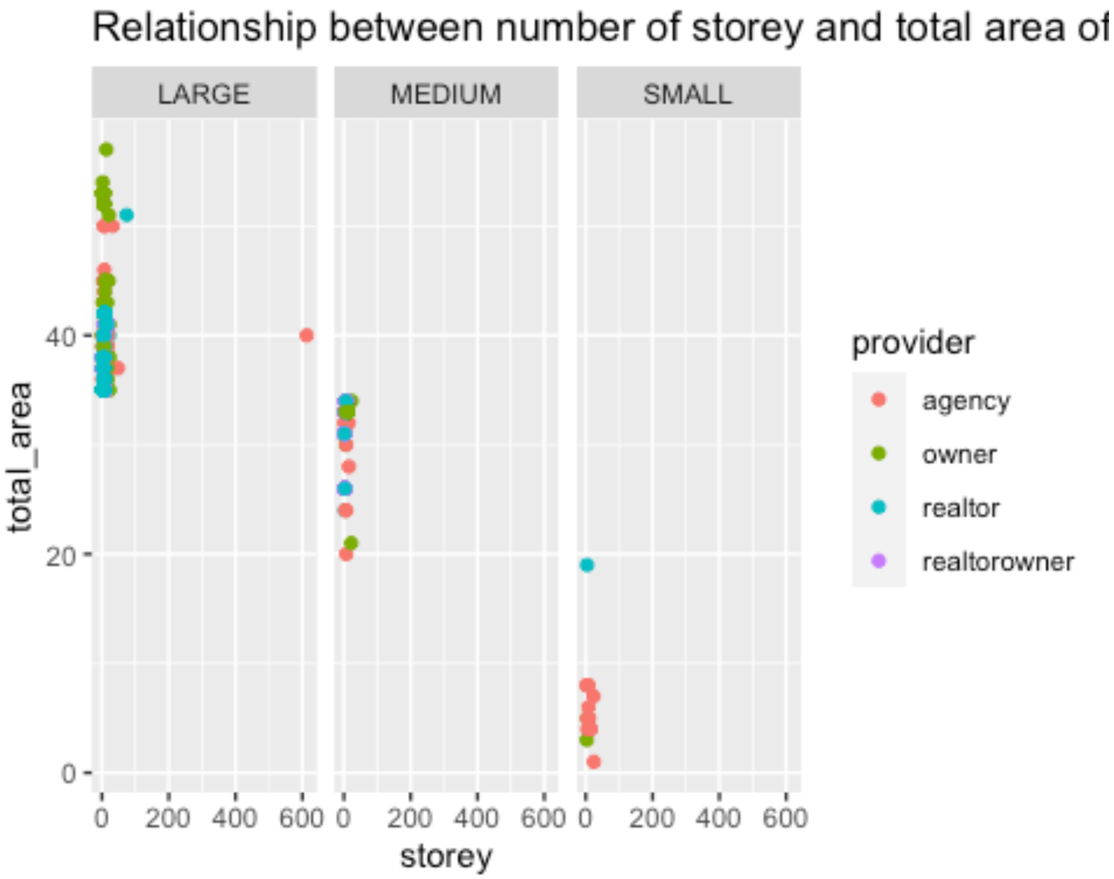
```
# Inference : Most of the apartments are at walking distance from metro
stations

# Number of views for the site-----------------------------------------------
---
summary( views)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    38.0   103.0   417.9   414.0  5174.0

# Inference : Median value fro views is 103

ggplot(data=data, aes(x=factor(total_area_description), y=views,
                      fill=factor(total_area_description))) +
  geom_boxplot(notch=T) +
  labs(title="Boxplot showing the relationship between area of flat and
number of views",
       fill = "Flat size", x="Flat size")

## notch went outside hinges. Try setting notch=FALSE.
```



```
# Inference : This shows medium size flats are in the highest demand and
experience a median view count of 500.

# Provider of the apartment-----------------------------------------------------
---
summary(factor( provider))

##       agency        owner      realtor realtorowner
##          129          287          505          525

ggplot(data=data, aes(x = factor(provider), fill = factor(provider),))
+geom_bar(position="dodge")+ labs(title = "Barplot showing provider
distribution", x = "Provider", fill = "Provider")
```



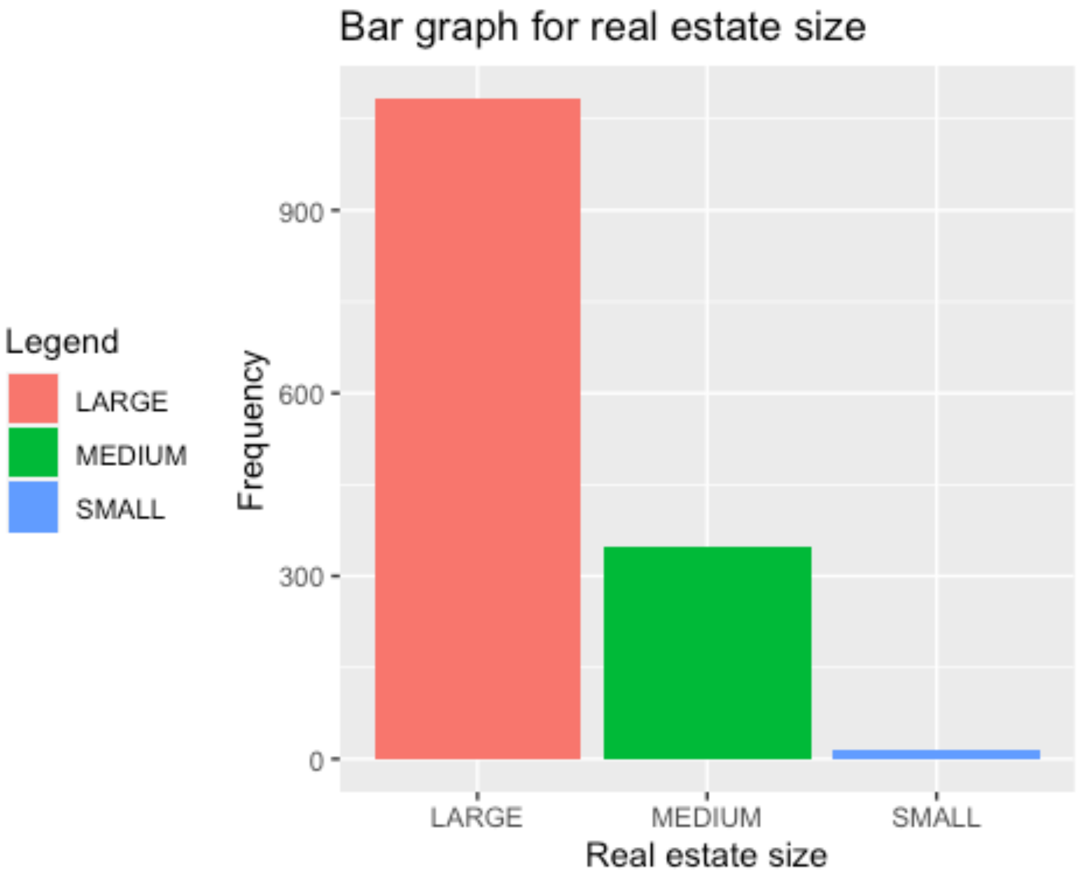
```
# Inference most of the apartments in Moscow are rented out by Reltor or
owners
summary(views)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    38.0   103.0   417.9   414.0  5174.0

ggplot(data=data, aes(x=storey, y = total_area ,col=provider)) +
  geom_point() +
  labs(title="Relationship between number of storey and total area of flat ")
+
  facet_grid(~total_area_description)
```



```
#Inference : Maximum data for large is concentrated towards total area
40sqft, while that for medium is around 30sqft and that of small properties
is less than 10sqft
# Fee percentage charged-----------------------------------------------------
---
summary( fee_percent)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00   50.00   37.95   50.00  100.00

# Inference : 50 percent is the average fee percentage charged by the broker
var( fee_percent)

## [1] 723.2521

# Inference: Fee percentage has a high variance i.e value of fee percentage
can vary a lot

# Number of stories----------------------------------------------------------
---
summary( storey)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00    6.00    7.09    9.00  613.00
```

```
# Inference: Number of stories mostly lie between 6 to 9 but some apartments
might go up to 613 stories

# Living area----------------------------------------------------------------
---
summary( living_area)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   18.00   20.00   20.59   21.00   37.00

# Inference : Average living in Moscow apartments is 20 square foot
range( living_area)

## [1]  6 37

# Inference : Living area ranges from 6 square foot to 37 square foot

# Kitchen area---------------------------------------------------------------
---
summary( kitchen_area)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    7.00   10.00   11.37   10.00   37.00

#Inference : The average kitchen area in Moscow apartments is 10 square foot
range( kitchen_area)

## [1]  3 37

#Inference : The kitchen area in Moscow apartments varies from 3 square foot
to 37 square foot

# Living area----------------------------------------------------------------
---
summary( total_area)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.00   37.00   37.27   40.00   57.00

range( total_area)

## [1]  1 57

# Inference : The average area of apartments in Moscow is 37 square foot
whereas the area might vary anywhere from 1 square foot to 57 square foot

# Area description-----------------------------------------------------------
---
summary(factor( total_area_description))

##  LARGE MEDIUM  SMALL
##   1083    348     15
```

```
ggplot(data=data,aes(x=total_area_description,fill=as.factor(total_area_descr
iption)))+
  geom_bar(position='dodge')+ggtitle("Bar graph for real estate size
")+labs(x="Real estate size",y="Frequency",fill="Legend")+
  theme(legend.position='left')
```



```
# Inference : This bar graph shows that most of the real estate in Moscow has
an area greater than 35 square foot and very few real estate have an area
less than 20 square foot

# Probability of finding a large real estate in Moscow
paste("There is a" ,round(((table(
total_area_description)["LARGE"])/nrow(data))*100, 2), "% probability of
finding a LARGE appartment in Moscow")

## [1] "There is a 74.9 % probability of finding a LARGE appartment in
Moscow"

paste("There is a" ,round(((table(
total_area_description)["MEDIUM"])/nrow(data))*100, 2), "% probability of
finding a MEDIUM appartment in Moscow")

## [1] "There is a 24.07 % probability of finding a MEDIUM appartment in
Moscow"
```

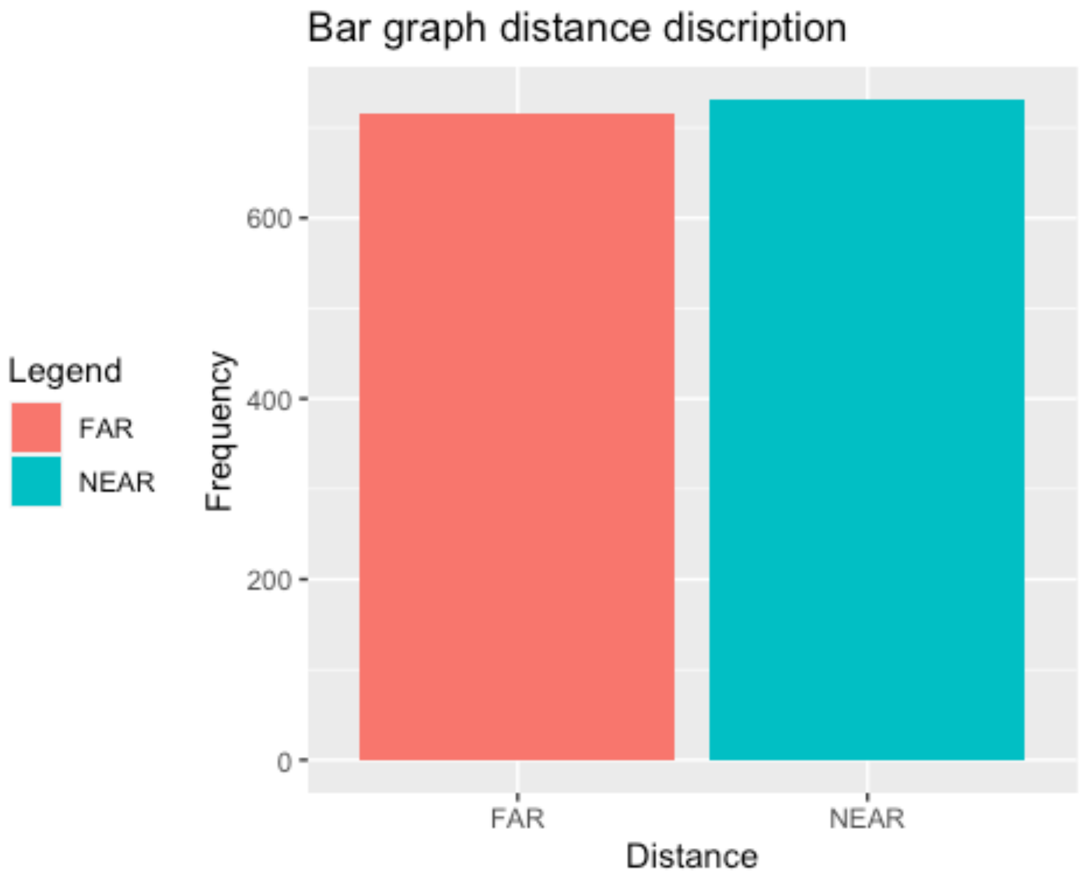
```
paste("There is a" ,round(((table(
total_area_description)["SMALL"])/nrow(data))*100, 2), "% probability of
finding a SMALL appartment in Moscow")
```

```
## [1] "There is a 1.04 % probability of finding a SMALL appartment in
Moscow"
```

```
# Distance from metro description---------------------------------------------
---
summary(factor( distance_description))
```

```
##  FAR NEAR
##  715  731
```

```
ggplot(data=data,aes(x=distance_description,fill=as.factor(distance_descripti
on)))+
  geom_bar(position='dodge')+ggtitle("Bar graph distance discription
")+labs(x="Distance",y="Frequency",fill="Legend")+
  theme(legend.position='left')
```



```
# Inference : Half of the real estate in Moscow lies such that travel time to
metro stations is less than 7 minutes whereas half of the real estate in
Moscow lies far from any metro station
```

```
# Correlation between price and total area-----------------------------------
---
cor( price,  total_area)

## [1] 0.2745654

# Inference: This shows that the prices of real estate and the area of
apartment are moderately correlated which means more the area more the prices

# Correlation between price and total area-----------------------------------
---
cor( price,  views)

## [1] -0.1208093

# Inference: This shows that the prices of real estate and the number of
views are negatively correlated which means higher the number of views lower
are the prices

# HO: Data follows normal distribution
# HA: Data does not follow normal distribution

shapiro.test(views)

##
##  Shapiro-Wilk normality test
##
## data:  views
## W = 0.41104, p-value < 2.2e-16

#Inference:
# Since the p-value is lesser than 0.05, we must accept HA and conclude that
the
# variable is not a part of the normal distribution

# Finding the Minimum and maximum of each variable using the lapply() and
# sapply() functions:

lapply(data, max)

## $S.No
## [1] 1445
##
## $metro
## [1] "Ziablikovo"
##
## $price
## [1] 5e+05
##
## $way
## [1] "walk"
```

```
## 
## $views
## [1] 5174
## 
## $provider
## [1] "realtorowner"
## 
## $fee_percent
## [1] 100
## 
## $storey
## [1] 613
## 
## $minutes
## [1] 47
## 
## $storeys
## [1] 13217
## 
## $living_area
## [1] 37
## 
## $kitchen_area
## [1] 37
## 
## $total_area
## [1] 57
## 
## $total_area_description
## [1] "SMALL"
## 
## $distance_description
## [1] "NEAR"
```

```
sapply(data, min)
```

```
##                    S.No                  metro                  price 
##                     "0"      "Akademicheskaia"                "14000" 
##                     way                  views               provider 
##             "transport"                    "4"               "agency" 
##             fee_percent                 storey                minutes 
##                     "0"                    "1"                    "0" 
##                 storeys            living_area           kitchen_area 
##                     "1"                    "6"                    "3" 
##              total_area total_area_description   distance_description 
##                     "1"                "LARGE"                  "FAR"
```

```
# Inference:
# The lapply and sapply functions are used in order to apply a given function
# and result in it's value being given out.
# Exploring the mean of various variables using the tapply() function:

tapply(price, total_area_description, mean)

##    LARGE   MEDIUM    SMALL
## 44463.18 40995.28 58166.67

# Inference:


tapply(total_area, distance_description , mean)

##      FAR     NEAR
## 36.17762 38.32969

# Inference:

# Analyzing the relationship between Provider and total area results
# using the Chi-square test


# HO: There is no association between BMI result and Diabetes results
# HA: There is an association between BMI result and Diabetes results
chisq.test(provider, total_area_description)

## Warning in chisq.test(provider, total_area_description): Chi-squared
## approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  provider and total_area_description
## X-squared = 260.49, df = 6, p-value < 2.2e-16

# Inference:
# Chi-square test is used to find the association between two categorical
# variables. Since the p-value is lesser than 0.05, we can conclude that
there
# is no association between Provider and total area.

# Finding linear regression between views and price of real estate
regression = lm(views~price)
regression

##
## Call:
## lm(formula = views ~ price)
##
## Coefficients:
## (Intercept)        price
##  566.938431    -0.003405

#Inference:
summary(regression)
```

```
## 
## Call:
## lm(formula = views ~ price)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -495.3 -381.9 -292.8   -7.7 4705.8 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.669e+02  4.045e+01  14.015  < 2e-16 ***
## price       -3.405e-03  7.362e-04  -4.625 4.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 930 on 1444 degrees of freedom
## Multiple R-squared:  0.01459,    Adjusted R-squared:  0.01391 
## F-statistic: 21.39 on 1 and 1444 DF,  p-value: 4.088e-06
```

```
# Finding linear regression between living area and price
regression = lm(living_area~price)
regression
```

```
## 
## Call:
## lm(formula = living_area ~ price)
## 
## Coefficients:
## (Intercept)        price  
##   2.054e+01    9.751e-07
```

```
#Inference:
summary(regression)
```

```
## 
## Call:
## lm(formula = living_area ~ price)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14.5762  -2.5811  -0.6016   0.4126  16.4345 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.054e+01  2.441e-01   84.17   <2e-16 ***
## price       9.751e-07  4.442e-06    0.22    0.826    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.611 on 1444 degrees of freedom
## Multiple R-squared:  3.338e-05,  Adjusted R-squared:  -0.0006591
## F-statistic: 0.0482 on 1 and 1444 DF,  p-value: 0.8263

# Conclusion-----------------------------------------------------------------
---
# With the help of this statistical study we are able to identify that in
Moscow most of the real estate prices range from 14000 to 38000, and also
most apartments are at a walking distance from the nearest metro station,
most of the sites experience a view rate upto 500, most of the sites are
rented out by either retailer or the owner and also most of the properties
are large. We also discovered that higher the number of views lower are the
prices and more the area of the apartment more is the price. This assignment
has been really helpful in understanding the R functionality and also the use
of various R libraries, such as ggplot 2, moments, readxl.
```