Project

GROUP_5||Economic Statistics

2024-01-25

```
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(psych)

## Warning: package 'psych' was built under R version 4.3.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

library(readxl)

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

STATISTICAL COMPUTING PROJECT

SECTION 1 USING R

Q1. reading and understanding the dataset.

A) Reading and Understanding the Dataset

```
#i.  Read or import NHANES dataset in R.
NHANES<-read.csv("C:/Users/USER/Desktop/R_Programming/NHANES.csv")
```

```r
#ii. How many variables in the dataset? What are the names of the variables
in the dataset?
length(NHANES) # there are 76 variables in dataset
```

```
## [1] 76
```

```r
names(NHANES) # Here are names of 76 variables in our dataset
```

```
##  [1] "ID"                "SurveyYr"          "Gender"            "Age"
##  [5] "AgeDecade"         "AgeMonths"         "Race1"             "Race3"
##  [9] "Education"         "MaritalStatus"     "HHIncome"
"HHIncomeMid"
## [13] "Poverty"           "HomeRooms"         "HomeOwn"           "Work"
## [17] "Weight"            "Length"            "HeadCirc"          "Height"
## [21] "BMI"               "BMICatUnder20yrs" "BMI_WHO"           "Pulse"
## [25] "BPSysAve"          "BPDiaAve"          "BPSys1"            "BPDia1"
## [29] "BPSys2"            "BPDia2"            "BPSys3"            "BPDia3"
## [33] "Testosterone"      "DirectChol"        "TotChol"           "UrineVol1"
## [37] "UrineFlow1"        "UrineVol2"         "UrineFlow2"        "Diabetes"
## [41] "DiabetesAge"       "HealthGen"         "DaysPhysHlthBad"
"DaysMentHlthBad"
## [45] "LittleInterest"    "Depressed"         "nPregnancies"      "nBabies"
## [49] "Age1stBaby"        "SleepHrsNight"     "SleepTrouble"      "PhysActive"
## [53] "PhysActiveDays"    "TVHrsDay"          "CompHrsDay"
"TVHrsDayChild"
## [57] "CompHrsDayChild"   "Alcohol12PlusYr"   "AlcoholDay"
"AlcoholYear"
## [61] "SmokeNow"          "Smoke100"          "Smoke100n"         "SmokeAge"
## [65] "Marijuana"         "AgeFirstMarij"     "RegularMarij"
"AgeRegMarij"
## [69] "HardDrugs"         "SexEver"           "SexAge"
"SexNumPartnLife"
## [73] "SexNumPartYear"    "SameSex"           "SexOrientation"
"PregnantNow"
```

```r
#iii. Select following column number 3, 4, 1, 7,9,10,12,14,15,16,35,40,46
Selected_var<- NHANES%>%select(c(3,4,1,7,10,12,14,15,35,40,46))
head(Selected_var) # Here we get a new data set of only 11 variables from our
NHANES data set and I called head to minimize the space.
```

```
##   Gender Age    ID Race1 MaritalStatus HHIncomeMid HomeRooms HomeOwn
TotChol
## 1   male  34 51624 White       Married       30000         6     Own
3.49
## 2   male  34 51624 White       Married       30000         6     Own
3.49
## 3   male  34 51624 White       Married       30000         6     Own
3.49
## 4   male   4 51625 Other          <NA>       22500         9     Own
NA
## 5 female  49 51630 White   LivePartner       40000         5    Rent
```

```
6.70
## 6   male   9 51638 White          <NA>       87500         6    Rent
4.86
##   Diabetes Depressed
## 1       No   Several
## 2       No   Several
## 3       No   Several
## 4       No      <NA>
## 5       No   Several
## 6       No      <NA>

#iv.What data types are associated with each variable selected?
Data_type<- sapply(Selected_var, typeof)%>%as.data.frame()
Data_type

##                       .
## Gender        character
## Age             integer
## ID              integer
## Race1         character
## MaritalStatus character
## HHIncomeMid     integer
## HomeRooms       integer
## HomeOwn       character
## TotChol          double
## Diabetes      character
## Depressed     character

#v. Provide the numerical and categorical variables in the dataset.
num_vars<-sapply(Selected_var, is.numeric) # Those are numeric.
num_vars<-as.data.frame(num_vars)|> filter(num_vars==TRUE)
num_vars

##             num_vars
## Age             TRUE
## ID              TRUE
## HHIncomeMid     TRUE
## HomeRooms       TRUE
## TotChol         TRUE

chr_vars<-sapply(Selected_var,is.character)
chr_vars<- as.data.frame(chr_vars)%>%filter(chr_vars==TRUE) # Those are
character variables
chr_vars

##               chr_vars
## Gender            TRUE
## Race1             TRUE
## MaritalStatus     TRUE
## HomeOwn           TRUE
```

```
## Diabetes          TRUE
## Depressed         TRUE
```

```
#vi. Provide summary statistics for numerical variables (mean, median, standard deviation)
# since we have 5 numeric variables we'll look summary for each.
for_age<- describe(Selected_var$Age)
for_age<- as.data.frame(for_age)%>%select(mean,median,sd)
for_HHY<- describe(Selected_var$HHIncomeMid)%>%select(mean,median,sd)
for_HHY<- as.data.frame(for_HHY)
for_room<- describe(Selected_var$HomeRooms)%>%select(mean,median,sd)
for_room<- as.data.frame(for_room)
for_TotChol<- describe(Selected_var$TotChol)%>%select(mean,median,sd)
for_TotChol<- as.data.frame(for_TotChol)
sum_stat<- rbind(for_age,for_HHY,for_room,for_TotChol)# Here is summary statistics where x1,x11,x12,and x13 stands for statistics for age, HHincomeMid, Rooms, and total cholesterol respectively.
rownames(sum_stat)<- c("for_age","for_HHincome","for_rooms","for_TotChol")
sum_stat
```

```
##                       mean   median           sd
## for_age          36.742100    36.00    22.397566
## for_HHincome 57206.170421 50000.00 33020.276584
## for_rooms        6.248918     6.00     2.277538
## for_TotChol      4.879220     4.78     1.075583
```

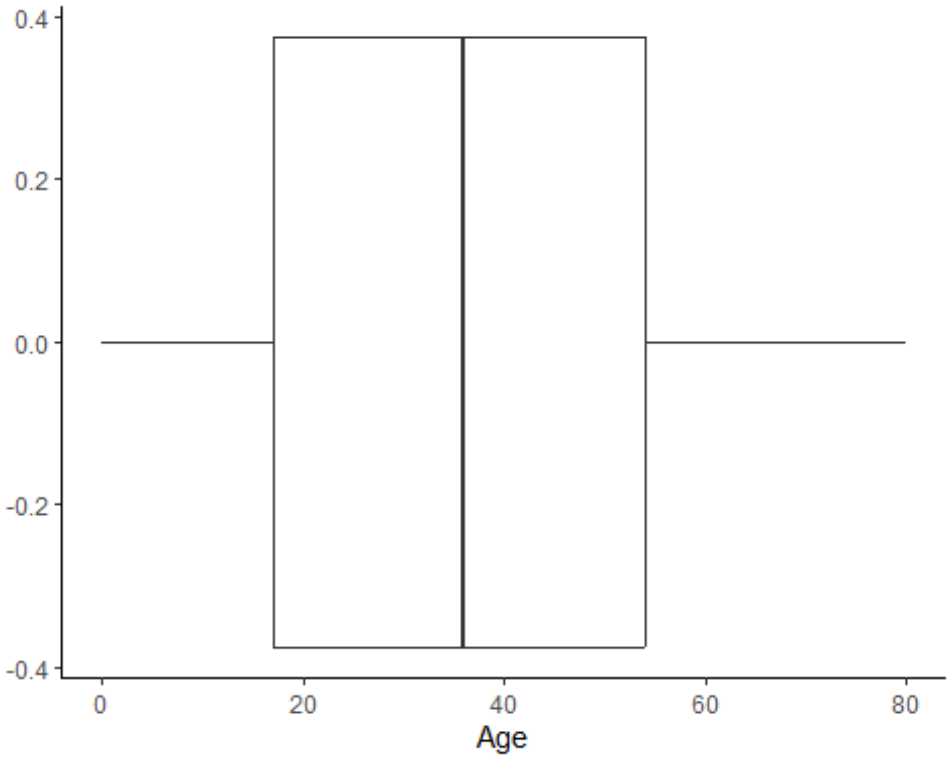
```
#vii.Are there any missing values and duplicates in the dataset, and if so, how many?
sum(is.na(NHANES)) # missing values are 277,677
```

```
## [1] 277677
```

```
sum(duplicated(NHANES)) # duplicated observations are 2,168
```

```
## [1] 2168
```

```
#viii.  Visualize outliers in Age using boxplot. (remember to set limit if your limits are too large)
Box_plot<- ggplot(data = Selected_var)+ geom_boxplot(mapping = aes(Age), outlier.shape = 4, outlier.colour = "blue", outlier.size = 3)+
theme_classic()
Box_plot # There are no outliers in age variable
```



B) Data Cleaning, manipulation and Exploratory Data Analysis

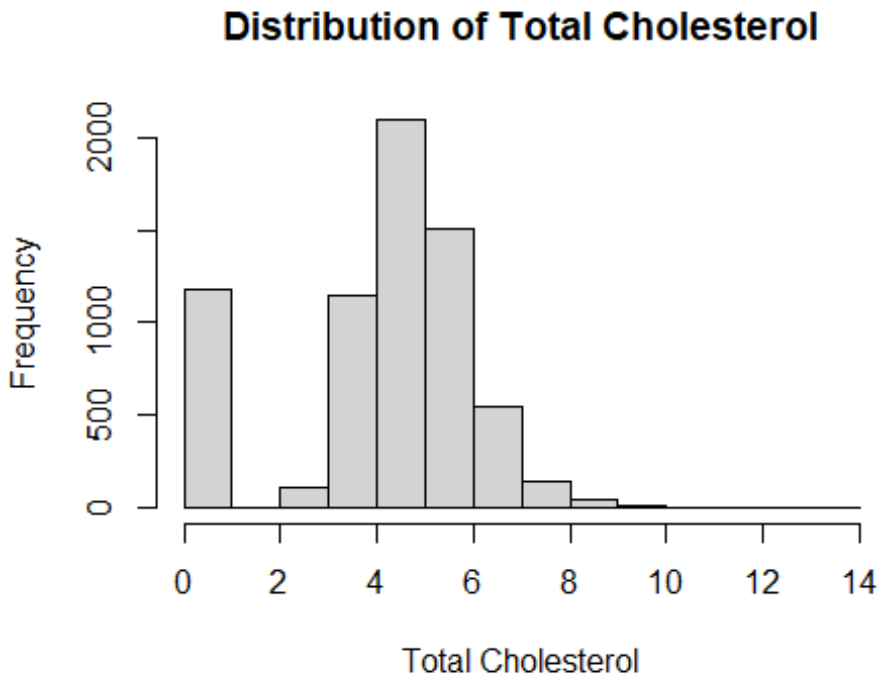
```
#i. Fill the missing values in the ' TotChol' column by zero (0) and check
again if there is no missing in that column.
Selected_var$TotChol[is.na(Selected_var$TotChol)]<-0
sum(is.na(Selected_var$TotChol)) # Now number of NA is zero because they've
been replaced by zero.
```

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

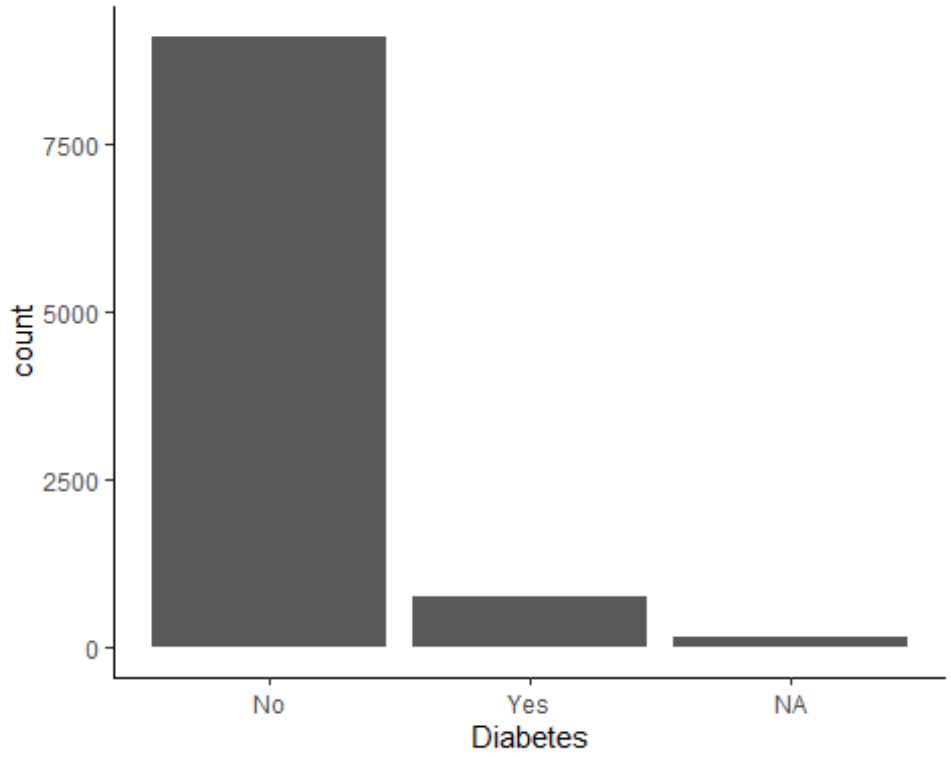
```
#ii. Remove all duplicated identified if any and check again if no duplicates
Selected_var<- Selected_var[!duplicated(Selected_var),]
sum(duplicated(Selected_var))# before there were 2,168 duplicates but now
they are all removed which has reduced nrow as well.
```

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

```
#iii.  Explore the distribution of ' TotChol' among patients. What patterns
or variations do you observe, and how might this information be relevant to
health analysis?
hist(Selected_var$TotChol, xlab = "Total Cholesterol", main = "Distribution
of Total Cholesterol") # the pattern is that TotChol between 3&6 is very
high, as well as at zero which is alarming that there are many people with
high risk of health issues like being venerable to desease
```




```r
#iv. Visualize and describe the distribution of diabetes across patients using bar plot. What insights can you gather from the diabetes distribution?
ggplot(data = NHANES)+geom_bar(mapping = aes(Diabetes))+theme_classic()  # This bar graph reveals that many of patients are diabetes negative but since they are NAs we can't be sure on exactly number of patients with diabetes.
```




```
#v.  Using the 'TotChol' column in the dataset, create a new categorical
variable named 'TotChol_group' to categorize patients into the following
cholesterol groups a. Group1: 0-1, b. Group2: 1-5, c. Group3: 5 and above
is.na(Selected_var$TotChol)
```

```
FALSE
```

```
 # since cut can not work in presence of NAs, I'll remove missing values.
is.na(Selected_var$TotChol)<- mean(Selected_var$TotChol)
Selected_var <- Selected_var[complete.cases(Selected_var$TotChol), ]
sum(is.na(Selected_var$TotChol))
```

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

```
Selected_var$TotChol_group <- cut(Selected_var$TotChol, breaks = c(0, 1, 5,
Inf), labels = c("Group1", "Group2", "Group3"))
Selected_var%>%head(10)%>% select(ID,TotChol,TotChol_group) # Now we have new
variable containing groups of TotChol.
```

```
##       ID TotChol TotChol_group
## 1  51624    3.49        Group2
## 4  51625    0.00          <NA>
## 6  51638    4.86        Group2
## 7  51646    4.09        Group2
## 8  51647    5.82        Group3
## 11 51654    4.99        Group2
## 12 51656    4.24        Group2
## 13 51657    6.41        Group3
```

```
## 14 51659    0.00          <NA>
## 15 51666    4.78        Group2
```

```
#vii. Explore the distribution between gender and key health metrics such as
diabetes and cholesterol groups ('TotChol_group') using a cross-tabulation
(table). How do health metrics differ between genders, and what implications
might this have?
cross_table <- xtabs(~ Gender + TotChol_group + Diabetes, data =
Selected_var)
cross_table #Those cross-table indicate no significant difference between
gender as shown clearly by cross-tabe of proportions below
```

```
## , , Diabetes = No
##
##         TotChol_group
## Gender   Group1 Group2 Group3
##   female      0   1479   1108
##   male        0   1547    953
##
## , , Diabetes = Yes
##
##         TotChol_group
## Gender   Group1 Group2 Group3
##   female      0    143     99
##   male        0    179     85
```

```
proportion<- prop.table(cross_table, margin = 2)
round(proportion,  digits = 2)
```

```
## , , Diabetes = No
##
##         TotChol_group
## Gender   Group1 Group2 Group3
##   female          0.44   0.49
##   male            0.46   0.42
##
## , , Diabetes = Yes
##
##         TotChol_group
## Gender   Group1 Group2 Group3
##   female          0.04   0.04
##   male            0.05   0.04
```

C) Regression Modeling

```
#i. Using the same dataset, fit the multiple linear regression model and show
the variables that are significant and provide the reason of your response.
Consider the dependent variable being "TotChlor" and independent variables
are: Gender, Depressed,and Diabetes.
model <- lm(TotChol ~ Gender + Depressed + Diabetes, data = Selected_var)
summ_model<-summary(model)
```

```
summ_model# Since all p-values are less than 0.05 those variables; gender,
diabetes,an Depressed are significant.

##
## Call:
## lm(formula = TotChol ~ Gender + Depressed + Diabetes, data = Selected_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0982 -0.6081  0.0900  0.8184  8.8438
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.09825    0.09048  56.345  < 2e-16 ***
## Gendermale       -0.12188    0.04550  -2.678  0.00742 **
## DepressedNone    -0.21829    0.09173  -2.380  0.01737 *
## DepressedSeveral -0.17013    0.10461  -1.626  0.10394
## DiabetesYes      -0.35648    0.07108  -5.015 5.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.481 on 4265 degrees of freedom
##   (2508 observations deleted due to missingness)
## Multiple R-squared:  0.008833,   Adjusted R-squared:  0.007903
## F-statistic: 9.502 on 4 and 4265 DF,  p-value: 1.205e-07

#ii. Calculate and interpret R-squared of your regression model. How well
does the model fit the data? Can we use the model in prediction?
R_squared<- summary(model) #In this summary the multiple R-squared is 0.0132
which shows the severe poor of this model to fit of the data. From above
summary of model.
```

Q2.QUESTION 2: HOUSING ANALYSIS

a) Reading and Understanding the Dataset

```
#i. Read the dataset "Housing dataset" given to you.
Housing<-read_xlsx("C:/Users/USER/Desktop/R_Programming/Housing
dataset.xlsx")

#ii. Display the variables or columns in the dataset. How many columns and
rows
nrow(Housing) #There are 545 rows or observations in dataset

## [1] 545

length(Housing) #There are 13 variables or column in dataset.

## [1] 13

columns<-as.vector(colnames(Housing))
columns #Here is name of all 13 variables we have.
```

```
##  [1] "price"            "area"             "bedrooms"         "bathrooms"
##  [5] "stories"          "mainroad"         "guestroom"        "basement"
##  [9] "hotwaterheating"  "airconditioning"  "parking"          "prefarea"
## [13] "furnishingstatus"
```

```
#iii. Display the last few rows to provide an overview of the data.
tail(Housing) #By default r provided 6 last rows of dataset
```

```
## # A tibble: 6 × 13
##     price  area bedrooms bathrooms stories mainroad guestroom basement
##     <dbl> <dbl>    <dbl>     <dbl>   <dbl> <chr>    <chr>     <chr>
## 1 1855000  2990        2         1       1 no       no        no
## 2 1820000  3000        2         1       1 yes      no        yes
## 3 1767150  2400        3         1       1 no       no        no
## 4 1750000  3620        2         1       1 yes      no        no
## 5 1750000  2910        3         1       1 no       no        no
## 6 1750000  3850        3         1       2 yes      no        no
## # ℹ 5 more variables: hotwaterheating <chr>, airconditioning <chr>,
## #   parking <dbl>, prefarea <chr>, furnishingstatus <chr>
```

```
#iv. Examine the data types associated with each variable.
data_type<- sapply(Housing, typeof)%>%as.data.frame()
data_type #This table contains all 13 variable and type of data they are
holding.
```

```
##                          .
## price                double
## area                 double
## bedrooms             double
## bathrooms            double
## stories              double
## mainroad          character
## guestroom         character
## basement          character
## hotwaterheating   character
## airconditioning   character
## parking              double
## prefarea          character
## furnishingstatus  character
```

```
#v.Generate summary statistics for numerical variables, including mean,
median, and standard deviation.
#Firstly, we need to know which variable are numeric
numerics<-sapply(Housing,  is.numeric)
numerics<-as.data.frame(numerics)|> filter(numerics==TRUE)
numerics #We have price,area,bedrooms,bathrooms,stories,and parking as
numeric variables.
```

```
##           numerics
## price         TRUE
## area          TRUE
```

```
## bedrooms      TRUE
## bathrooms     TRUE
## stories       TRUE
## parking       TRUE
```

```
#To get their summary statistics I'll use describe.
price<- describe(Housing$price)%>%select(mean,median,sd)
area<- describe(Housing$area)%>%select(mean,median,sd)
bedrooms<- describe(Housing$bedrooms)%>%select(mean,median,sd)
bathrooms<- describe(Housing$bathrooms)%>%select(mean,median,sd)
stories<- describe(Housing$stories)%>%select(mean,median,sd)
#To present them clearly I'll bind them
All_summ_stat<- rbind(price,area,bedrooms,bathrooms,stories)
rownames(All_summ_stat)<- c("price","area","bedrooms","bathrooms","stories")
All_summ_stat #Combined data frame showing summary statistics for numeric
variables we have in our dataset.
```

```
##                mean  median         sd
## price     4766729.25 4340000 1870439.62
## area         5150.54    4600    2170.14
## bedrooms        2.97       3       0.74
## bathrooms       1.29       1       0.50
## stories         1.81       2       0.87
```

```
#vi.  Analyze the dataset for missing values and duplicates. Quantify the
number of missing values and duplicates.
sum(is.na(Housing)) #There is no missing value.
```

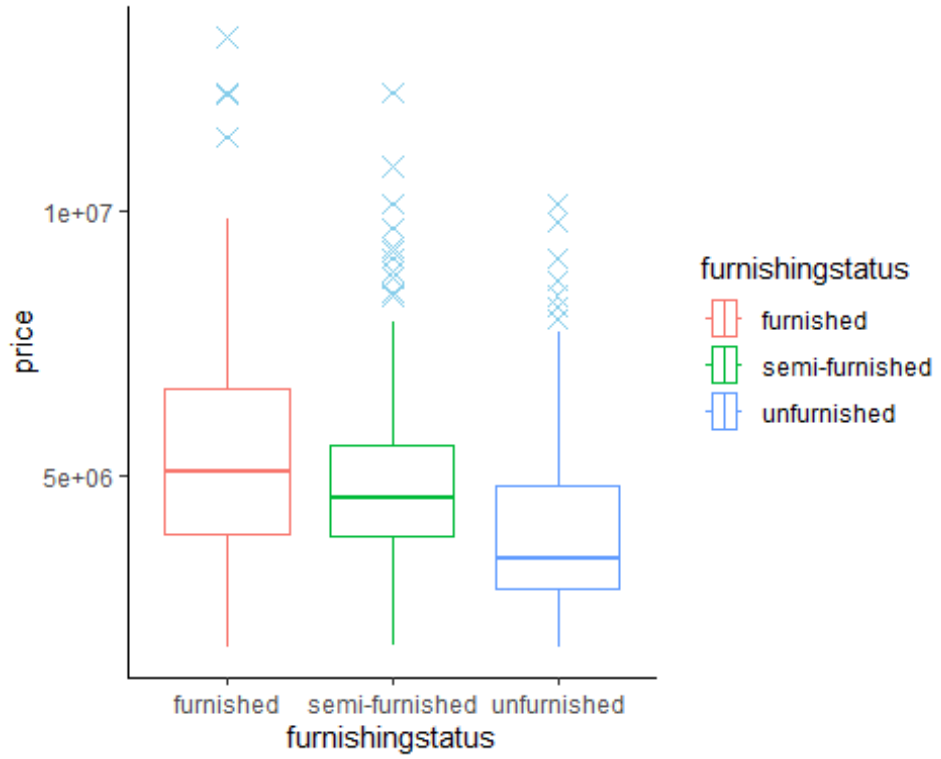
```
## [1] 0
```

```
sum(duplicated(Housing)) #There is no even duplicates as well.
```

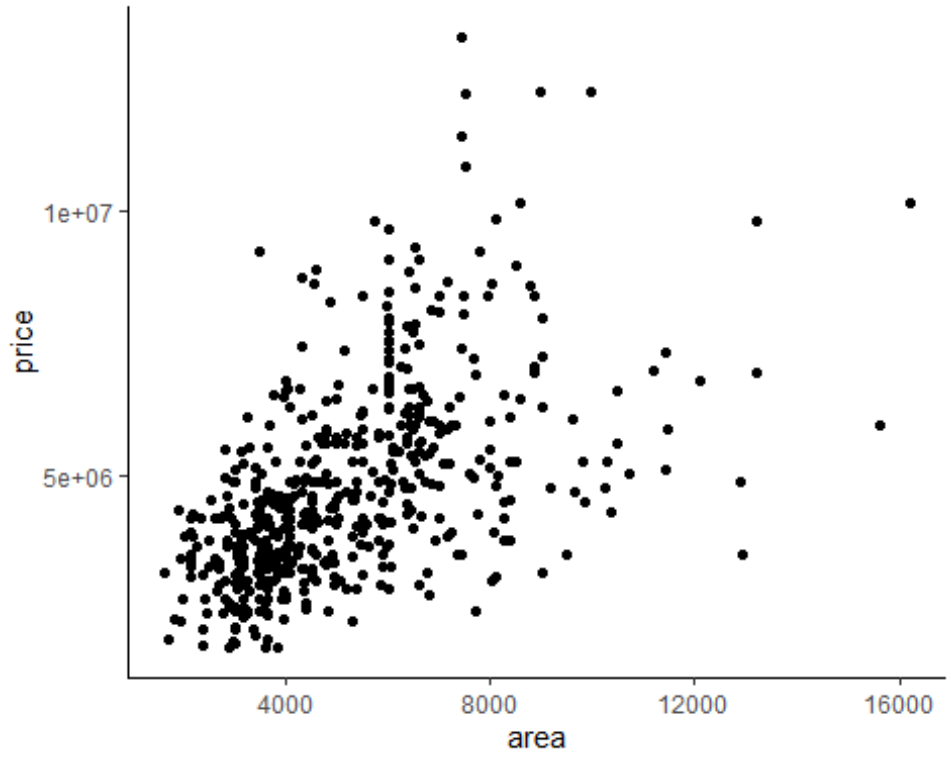
```
## [1] 0
```

b) Exploratory Data Analysis

```
#i. Explore the distribution of house prices and furnishing status using
boxplot and ggplotfunction. What patterns or variations do you observe, and
how might this information be relevant for buyers or investors?
box_plot<- ggplot(data = Housing)+geom_boxplot(mapping =
aes(price,furnishingstatus, colour= furnishingstatus), outlier.colour =
"skyblue", outlier.shape = 4,outlier.size = 3)+ theme_classic()+coord_flip()
box_plot #This box plots uncover that there are presence of outliers in high
prices for all furnishing status this inform investors that mostly houses
become expensive that normally expected.
```




```
#ii. Describe the distribution of area and price using scatter plot. What
insights can you gather from this analysis?
scatter<- ggplot(data = Housing)+geom_point(mapping =
aes(area,price))+theme_classic()+geom_abline()
scatter #This scatter depict the positive linear relation between area and
prices, however are unusual points which can be analyzed with further
analysis.
```




```
#iii. Create a table showing the number of bedrooms and bathrooms. What can
you say about the results?
bedroom_bathroom_table<-table(Housing$bedrooms,Housing$bathrooms)
bedroom_bathroom_table #It is impossible to have more bathrooms than
bedrooms, and majority of houses have two or three bedrooms with one
bathrooms.
```

```
##
##       1   2   3   4
##   1   2   0   0   0
##   2 128   8   0   0
##   3 224  72   4   0
##   4  42  48   4   1
##   5   4   4   2   0
##   6   1   1   0   0
```

c) Statistical testing and modelling Given below hypotheses: Null Hypothesis (H0): There is no significant difference in house prices with different parking numbers, Alternative Hypothesis (H1): There is a significant difference in house prices with different parking numbers,

```
#i. Check the normality of house prices with each parking number.
prices<- Housing$price[Housing$parking==Housing$parking]
shapiro_test<- shapiro.test(prices)
print(shapiro_test) #Those results indicate that prices for each number of
parking are not normally distributed.
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  prices
## W = 0.92163, p-value = 3.155e-16

#ii.For each parking number category, assess the normality of house prices
parking_number<- Housing$parking[Housing$price==Housing$price]
shpr_test<- shapiro.test(parking_number)
print(shpr_test) #This indicate extremely non-normality of prices in each
category of parking number.

##
##  Shapiro-Wilk normality test
##
## data:  parking_number
## W = 0.74844, p-value < 2.2e-16

#iii. What is the conclusion based on your p-values after running the anova
test?
#Since above shapiro-wilk test showed non-normality, we can not use
parametric test like ANOVA we can instead use some non-parametric methods.

#iv. Fit a multiple linear regression model with the dependent variable
“price” and all predictors (area, parking, furnishingstatus, mainroad,
bedrooms, stories, bathrooms, guestroom, basement, hotwaterheating,
airconditioning, prefarea).
Model<- lm(price ~ area + parking + furnishingstatus + mainroad + bedrooms +
stories + guestroom + hotwaterheating + bathrooms + basement + prefarea ,
data = Housing)
Model #For obtaining model

##
## Call:
## lm(formula = price ~ area + parking + furnishingstatus + mainroad +
##     bedrooms + stories + guestroom + hotwaterheating + bathrooms +
##     basement + prefarea, data = Housing)
##
## Coefficients:
##                     (Intercept)                             area
##                       -42590.3                            268.8
##                         parking  furnishingstatussemi-furnished
##                        317597.0                         -131003.6
##    furnishingstatusunfurnished                       mainroadyes
##                        -488639.4                         407645.7
##                        bedrooms                           stories
##                        109078.0                         574082.7
##                     guestroomyes              hotwaterheatingyes
##                        374836.9                         588181.9
##                       bathrooms                       basementyes
##                       1026959.7                         384802.8
##                     prefareayes
##                        679006.6
```

```r
#v. What are the significant variables based on the p-value.
model_summary<-summary(Model)
model_summary
```

```
##
## Call:
## lm(formula = price ~ area + parking + furnishingstatus + mainroad +
##     bedrooms + stories + guestroom + hotwaterheating + bathrooms +
##     basement + prefarea, data = Housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3063929  -641651   -38225   547453  5414056
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -42590.30  279231.80  -0.153 0.878829
## area                               268.76      25.47  10.551  < 2e-16 ***
## parking                         317597.03   61647.07   5.152 3.64e-07 ***
## furnishingstatussemi-furnished -131003.58  122743.61  -1.067 0.286323
## furnishingstatusunfurnished    -488639.39  133049.02  -3.673 0.000264 ***
## mainroadyes                     407645.69  150363.89   2.711 0.006924 **
## bedrooms                        109077.96   76755.43   1.421 0.155870
## stories                         574082.74   65853.63   8.718  < 2e-16 ***
## guestroomyes                    374836.86  138910.02   2.698 0.007188 **
## hotwaterheatingyes              588181.88  233270.46   2.521 0.011978 *
## bathrooms                      1026959.73  109161.47   9.408  < 2e-16 ***
## basementyes                     384802.80  116513.63   3.303 0.001022 **
## prefareayes                     679006.58  122257.77   5.554 4.42e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129000 on 532 degrees of freedom
## Multiple R-squared:  0.6436, Adjusted R-squared:  0.6356
## F-statistic: 80.06 on 12 and 532 DF,  p-value: < 2.2e-16
```

```r
model_summary$coefficients[, "Pr(>|t|)"] <-
round(model_summary$coefficients[, "Pr(>|t|)"], 2)
model_summary
```

```
##
## Call:
## lm(formula = price ~ area + parking + furnishingstatus + mainroad +
##     bedrooms + stories + guestroom + hotwaterheating + bathrooms +
##     basement + prefarea, data = Housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3063929  -641651   -38225   547453  5414056
##
```

```
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -42590.30  279231.80  -0.153     0.88
## area                                268.76      25.47  10.551   <2e-16 ***
## parking                          317597.03   61647.07   5.152   <2e-16 ***
## furnishingstatussemi-furnished -131003.58  122743.61  -1.067     0.29
## furnishingstatusunfurnished    -488639.39  133049.02  -3.673   <2e-16 ***
## mainroadyes                      407645.69  150363.89   2.711     0.01 **
## bedrooms                         109077.96   76755.43   1.421     0.16
## stories                          574082.74   65853.63   8.718   <2e-16 ***
## guestroomyes                     374836.86  138910.02   2.698     0.01 **
## hotwaterheatingyes               588181.88  233270.46   2.521     0.01 **
## bathrooms                       1026959.73  109161.47   9.408   <2e-16 ***
## basementyes                      384802.80  116513.63   3.303   <2e-16 ***
## prefareayes                      679006.58  122257.77   5.554   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129000 on 532 degrees of freedom
## Multiple R-squared:  0.6436, Adjusted R-squared:  0.6356
## F-statistic: 80.06 on 12 and 532 DF,  p-value: < 2.2e-16

significant_variables <-
model_summary$coefficients[model_summary$coefficients[, "Pr(>|t|)"] < 0.05, ]
significant_variables #Base on their p-value which is less than 0.05 the
following variables are significant; area,parking,mainroad,stories,bathrooms,
and others.

##                                    Estimate   Std. Error   t value Pr(>|t|)
## area                                268.7612     25.47285 10.550890     0.00
## parking                          317597.0264  61647.06573  5.151860     0.00
## furnishingstatusunfurnished -488639.3906 133049.02426 -3.672627     0.00
## mainroadyes                      407645.6862 150363.89115  2.711061     0.01
## stories                          574082.7353  65853.63118  8.717556     0.00
## guestroomyes                     374836.8631 138910.02023  2.698415     0.01
## hotwaterheatingyes               588181.8825 233270.45860  2.521459     0.01
## bathrooms                       1026959.7310 109161.47426  9.407712     0.00
## basementyes                      384802.7955 116513.63494  3.302642     0.00
## prefareayes                      679006.5829 122257.77218  5.553893     0.00

#vi. Calculate and interpret the R-squared value to assess how well the
regression model explains the variability in house prices.
model_summary #In this model summary the values of R-squared in 0.6356 which
means that the variation in house price can be explained by variation in
other independent variables in model at 63.5%. using Adjusted R-squared.

##
## Call:
## lm(formula = price ~ area + parking + furnishingstatus + mainroad +
##     bedrooms + stories + guestroom + hotwaterheating + bathrooms +
##     basement + prefarea, data = Housing)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3063929  -641651   -38225   547453  5414056
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -42590.30  279231.80  -0.153     0.88
## area                               268.76      25.47  10.551   <2e-16 ***
## parking                         317597.03   61647.07   5.152   <2e-16 ***
## furnishingstatussemi-furnished -131003.58  122743.61  -1.067     0.29
## furnishingstatusunfurnished    -488639.39  133049.02  -3.673   <2e-16 ***
## mainroadyes                     407645.69  150363.89   2.711     0.01 **
## bedrooms                        109077.96   76755.43   1.421     0.16
## stories                         574082.74   65853.63   8.718   <2e-16 ***
## guestroomyes                    374836.86  138910.02   2.698     0.01 **
## hotwaterheatingyes              588181.88  233270.46   2.521     0.01 **
## bathrooms                      1026959.73  109161.47   9.408   <2e-16 ***
## basementyes                     384802.80  116513.63   3.303   <2e-16 ***
## prefareayes                     679006.58  122257.77   5.554   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129000 on 532 degrees of freedom
## Multiple R-squared:  0.6436, Adjusted R-squared:  0.6356
## F-statistic: 80.06 on 12 and 532 DF,  p-value: < 2.2e-16
```

```
#vii. Assess how the model's performance changes when certain predictors such as area and furnishing status are excluded. Discuss the sensitivity of the model to specific variables.
model2<- lm(price ~ parking + mainroad + bedrooms + stories + guestroom + hotwaterheating + bathrooms + basement + prefarea , data = Housing)
model2
```

```
##
## Call:
## lm(formula = price ~ parking + mainroad + bedrooms + stories +
##     guestroom + hotwaterheating + bathrooms + basement + prefarea,
##     data = Housing)
##
## Coefficients:
##       (Intercept)            parking         mainroadyes
bedrooms
##            187266             530013             797063
189248
##           stories       guestroomyes  hotwaterheatingyes
bathrooms
##            554309             545504             552156
1176992
```

```
##        basementyes         prefareayes
##             327837              912239
```

```
summary(model2) #Removing area and furnishing status the R-squared reduced dramatically, this means that now the ability of a model as a prediction tool reduced
```

```
##
## Call:
## lm(formula = price ~ parking + mainroad + bedrooms + stories +
##     guestroom + hotwaterheating + bathrooms + basement + prefarea,
##     data = Housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3280107  -735633   -38057   608875  5569504
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           187266     268839   0.697 0.486374
## parking               530013      65502   8.092 3.98e-15 ***
## mainroadyes           797063     163136   4.886 1.36e-06 ***
## bedrooms              189248      85104   2.224 0.026583 *
## stories               554309      73180   7.575 1.59e-13 ***
## guestroomyes          545504     153777   3.547 0.000423 ***
## hotwaterheatingyes    552156     259269   2.130 0.033655 *
## bathrooms            1176992     120707   9.751  < 2e-16 ***
## basementyes           327837     129044   2.541 0.011351 *
## prefareayes           912239     134120   6.802 2.77e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1258000 on 535 degrees of freedom
## Multiple R-squared:  0.5553, Adjusted R-squared:  0.5478
## F-statistic: 74.23 on 9 and 535 DF,  p-value: < 2.2e-16
```

```
#The sensitivity of those two variables excluded can be aligned with the decrease in R-square of 0.088.
```

```
getwd()
```

```
## [1] "C:/Users/USER/Desktop/R_Programming"
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

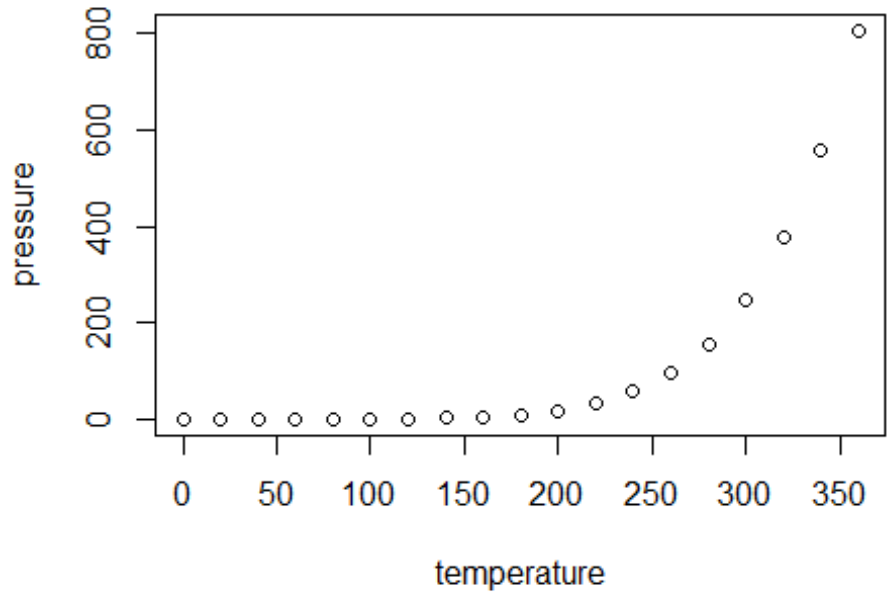
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.