21125052 - PHAM VO QUYNH NHU - DATA 9

SUMMER 2023

Table of Contents

Activity 01

I. Introduction

1. Setting environment for data

```
data<-read.csv('smoke.csv', header = FALSE)
header_names <- c("educ", "cigpric", "white", "age", "income", "cigs",
"restaurn", "lincome", "agesq", "lcigpric")
colnames(data) <- header_names
attach(data)
str(data)
```

```
## 'data.frame':    807 obs. of  10 variables:
##  $ educ    : num  16 16 12 13.5 10 6 12 15 12 12 ...
##  $ cigpric : num  60.5 57.9 57.7 57.9 58.3 ...
##  $ white   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  46 40 58 30 17 86 35 48 48 31 ...
##  $ income  : int  20000 30000 30000 20000 20000 6500 20000 30000 20000
20000 ...
##  $ cigs    : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ restaurn: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lincome : num  9.9 10.3 10.3 9.9 9.9 ...
##  $ agesq   : int  2116 1600 3364 900 289 7396 1225 2304 2304 961 ...
##  $ lcigpric: num  4.1 4.06 4.05 4.06 4.07 ...
```

2. Data size

```
dim(data)
```

```
## [1] 807  10
```

```
sum(is.na(data))
```

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

The result shows that this data has 807 rows (observations) and 10 columns (variables). And the second row show that no missing values are present in the data set.

3. Brief summary of the data

```
##       educ          cigpric          white             age
##  Min.   : 6.00   Min.   :44.00   Min.   :0.0000   Min.   :17.00
##  1st Qu.:10.00   1st Qu.:58.14   1st Qu.:1.0000   1st Qu.:28.00
##  Median :12.00   Median :61.05   Median :1.0000   Median :38.00
##  Mean   :12.47   Mean   :60.30   Mean   :0.8786   Mean   :41.24
##  3rd Qu.:13.50   3rd Qu.:63.18   3rd Qu.:1.0000   3rd Qu.:54.00
##  Max.   :18.00   Max.   :70.13   Max.   :1.0000   Max.   :88.00
##      income           cigs          restaurn         lincome
##  Min.   :  500   Min.   : 0.000   Min.   :0.0000   Min.   : 6.215
##  1st Qu.:12500   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 9.433
##  Median :20000   Median : 0.000   Median :0.0000   Median : 9.903
##  Mean   :19305   Mean   : 8.686   Mean   :0.2466   Mean   : 9.687
```

```
##  3rd Qu.:30000   3rd Qu.:20.000   3rd Qu.:0.0000   3rd Qu.:10.309
##  Max.   :30000   Max.   :80.000   Max.   :1.0000   Max.   :10.309
##      agesq         lcigpric
##  Min.   : 289   Min.   :3.784
##  1st Qu.: 784   1st Qu.:4.063
##  Median :1444   Median :4.112
##  Mean   :1990   Mean   :4.096
##  3rd Qu.:2916   3rd Qu.:4.146
##  Max.   :7744   Max.   :4.250
```

The observations include the following factors:

- educ: Number of years of education
- cigpric: Price of cigarettes in the state (in cents per pack)
- white: 1 if the individual is white, 0 otherwise
- age: Age of the individual in years
- income: Annual income in dollars
- cigs: Number of cigarettes smoked per day
- restaurn: 1 if state smoking restrictions
- lincome: Log of income.
- agesq: Age squared (age * age).
- lcigpric: Log of cigarette price.

→ Via observing the provided data, it is evident that the column set comprises 8 quantitative variables and 2 qualitative variables.

II. Descriptive Statistics

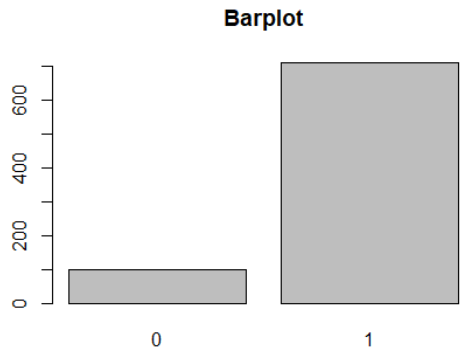
0. Prepared functions

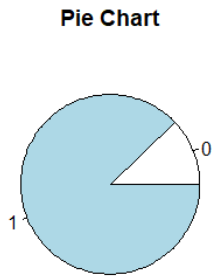
```
mode <- function( x, na.rm = FALSE) {
  if(na.rm){ x = x[!is.na(x)]  }
  val <- unique(x)
  return(val[which.max(tabulate(match(x, val)))])
}

outliers <- function(x) {
# 1st and 3rd quantiles
  q75 = quantile(x, 0.75)
  q25 = quantile(x, 0.25)
  IQR = q75-q25
# lower bound
  lower_bound = q25 - 1.5 * IQR
# upper bound
  upper_bound = q75 + 1.5 * IQR
# outliers
  outlier_ind <- which(x < lower_bound | x > upper_bound)
  if (length(outlier_ind) == 0) return (0)
```

```
  return(outlier_ind)
}
```

1. Number of whites

```
## white
##   0   1
##  98 709
```

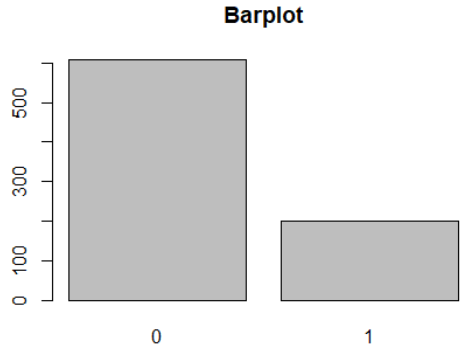


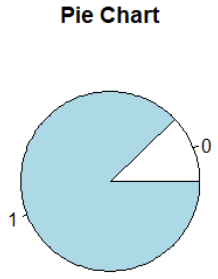


→ The major of observers are white.

2. Number of restaurant having smoking restrictions

```
## restaurn
##   0   1
## 608 199
```

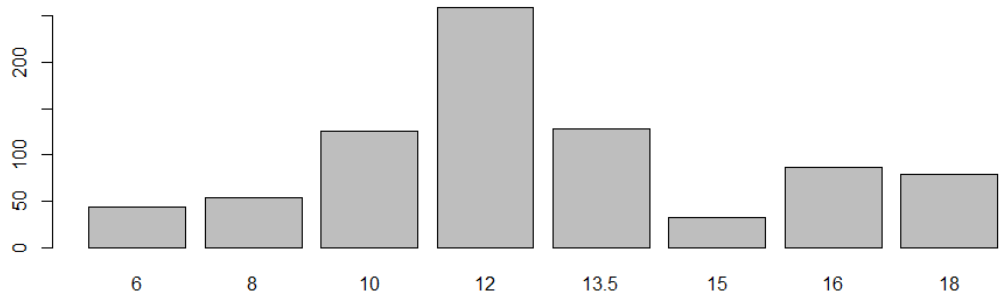




→ The majority of states have smoking restrictions in their restaurants.

3. Number of years of education

```
## educ
##    6    8   10   12 13.5   15   16   18
##   43   54  126  259  128   32   86   79
```

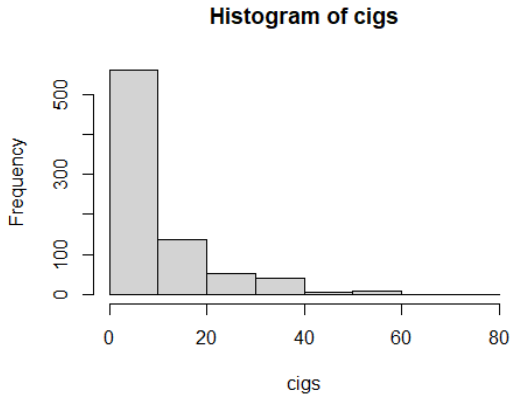


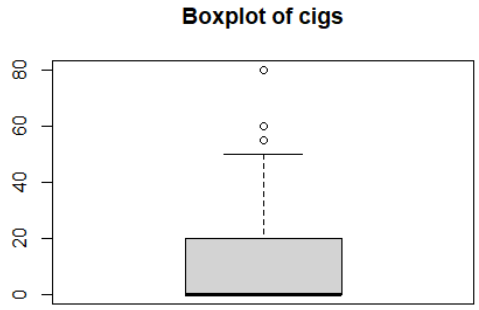

```
## Mode of years of education:  12

## Summary of years of education:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   10.00   12.00   12.47   13.50   18.00
```

→ Based on the data, it appears that the majority of observers have completed 12 years of education.

4. Number of cigarettes smoked per day

```
## cigs
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  18
19  20
## 497   7   5   5   2   7   3   2   3   2  28   2   4   2   1  23   1   3
1 101
##  25  28  30  33  35  40  50  55  60  80
##   7   3  42   1   2  37   6   1   8   1
```

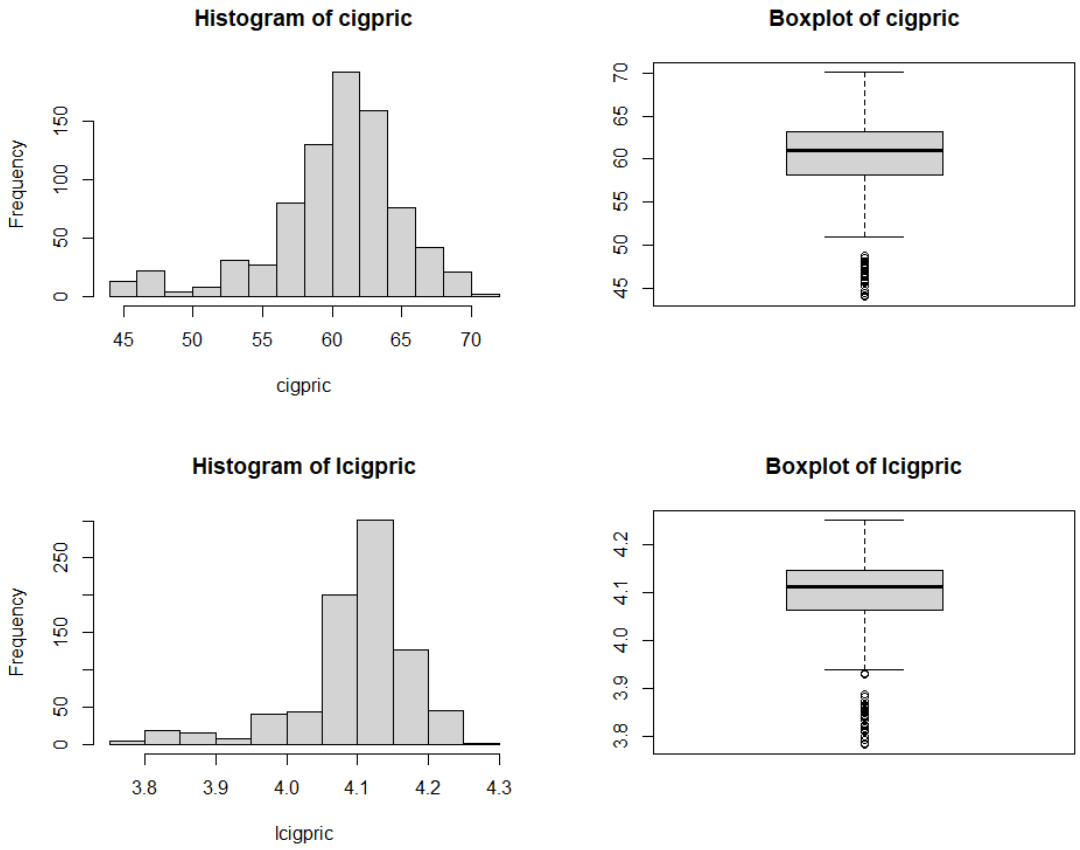





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   8.686  20.000  80.000
```

- The histogram indicates that the number of cigarettes smoked per day is concentrated mainly in the range of 0 to 40 cigarettes.
- This box-plot shows that there are still exist outliers but not too much (3 outliers).

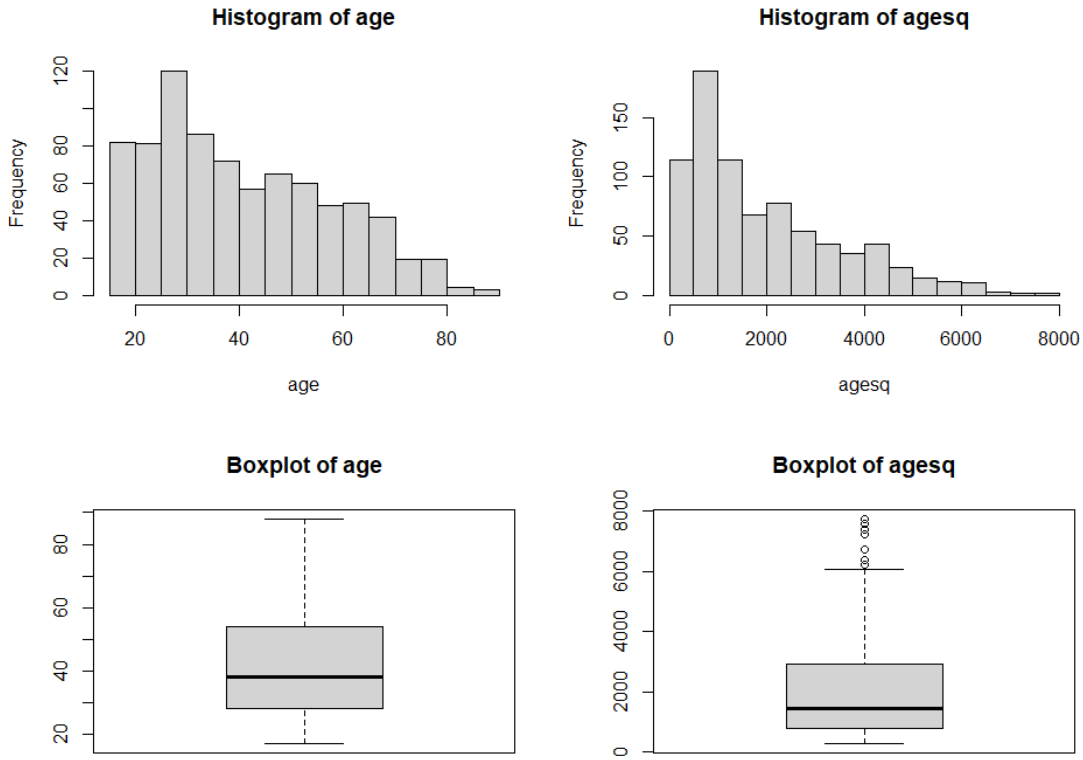
5. Price of cigarettes in the state (in cents per pack)




```
## Total outlier of cigpric:  39   0.04832714

## Summary of cigpric:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   58.14   61.05   60.30   63.18   70.13
```

- The histogram indicates that price of cigarettes in states is concentrated mainly in the range of 55 cents to 65 cents.
- This box-plot shows that there are still exist outliers (39 outliers).

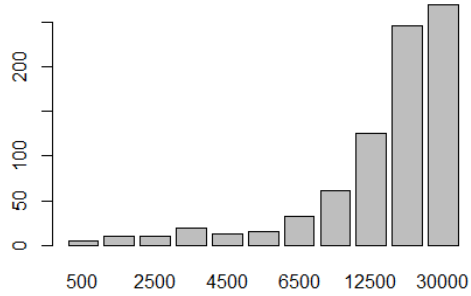
6. Age

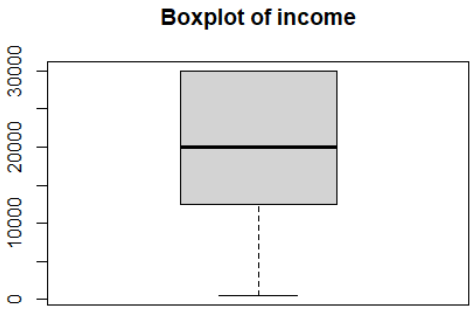


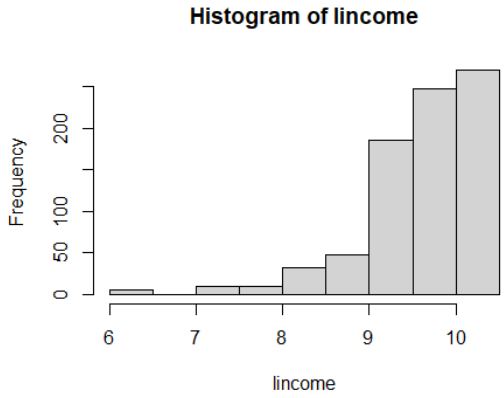

```
## Mode of age:  30

## Summary of age

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   38.00   41.24   54.00   88.00
```

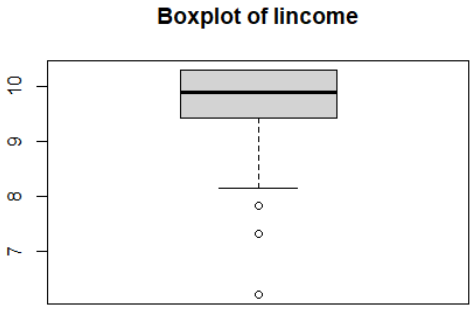
The histogram indicates that the age of observers is concentrated mainly in the range of 20 years to 40 years. Moreover, it seems to be skewed right.

7. Income










```
## Mode of age:  30000

## Summary of income:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     500   12500   20000   19305   30000   30000
```

The bar chart indicates that income of observers is concentrated mainly in the range of $6500 to $20000 and the histogram of lincome seems to be skewed left.

III. Inference Statistics

1. Testing for price of cigarettes in states

1.1 General

Utilizing the cigpric summary provided, where the average cigpric within the dataset stands at $60.30, our objective is to ascertain whether the actual mean corresponds to $60.

Hypothesis:

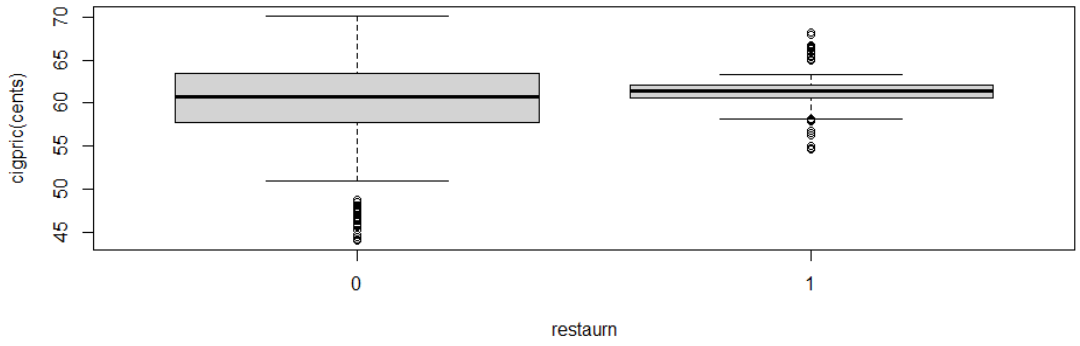
$$H_0: \mu_0 = 60$$

$$H_1: \mu_0 \neq 60$$

```
##
##  One Sample t-test
##
## data:  cigpric
## t = 1.801, df = 806, p-value = 0.07208
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  59.97299 60.62783
## sample estimates:
## mean of x
##  60.30041
```

The obtained p-value is 0.07208, which is greater than the predetermined significance level of α=0.05 for a confidence level of 95%. Consequently, we do not have sufficient evidence to reject the null hypothesis. Therefore, we conclude that the mean value of the price of cigarettes is $60.

1.2 Testing price between two categories



From the view of the data context, we have the reason to assume that the price of cigarettes at state having smoking restriction in restaurants is larger than the price of cigarettes at state not having smoking restriction in restaurants.

Hypothesis:

μ_1: means of prices of cigarettes at state having smoking restriction in restaurants

μ_2: means of prices of cigarettes at state not having smoking restriction in restaurants

$$H_0: \mu_{10} = \mu_{20}$$

$$H_1: \mu_{10} < \mu_{20}$$

```
##
##  Welch Two Sample t-test
##
```

```
## data:  cigpric[restaurn == 1] and cigpric[restaurn == 0]
## t = 5.6466, df = 728.78, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.974858
## sample estimates:
## mean of x mean of y
##  61.45231  59.92339
```

The obtained p-value is 1, which is undoubtly accept the null hypothesis. Therefore, the mean price of cigarettes at state having smoking restriction in restaurants is larger than the price of cigarettes at state not having smoking restriction in restaurants.

2. Testing for years of education

Utilizing the educ summary provided, where the average educ within the dataset stands at 12.47 years, our objective is to ascertain whether the actual mean corresponds to 13 years.

Hypothesis:

$$H_0: \mu_0 = 13$$

$$H_1: \mu_0 \neq 13$$

```
##
##  One Sample t-test
##
## data:  educ
## t = -4.9167, df = 806, p-value = 1.067e-06
## alternative hypothesis: true mean is not equal to 13
## 95 percent confidence interval:
##  12.25964 12.68212
## sample estimates:
## mean of x
##  12.47088
```

The obtained p-value is 0, which is smaller than the predetermined significance level of α=0.05 for a confidence level of 95%. Consequently, we reject the null hypothesis. Therefore, we conclude that the mean value of the years of education is not equal to 13 years.

3. Testing for number of whites

```
## [1] 0.1214374
```

The result show that 12,14% data are not of white ethnicity. So for now we will test whether this proportion is larger than 10% or not.

Hypothesis:

$$H_0: \rho_0 = 0.10$$

$$H_1: \rho_0 > 0.10$$

```
##
##  1-sample proportions test with continuity correction
##
## data:  table(white), null probability 0.1
## X-squared = 3.886, df = 1, p-value = 0.02435
## alternative hypothesis: true p is greater than 0.1
## 95 percent confidence interval:
##  0.1032028 1.0000000
## sample estimates:
##         p
## 0.1214374
```

The obtained p-value is p-value = 0.02435 which is smaller than the predetermined significance level of α=0.05 for a confidence level of 95%. Consequently, we reject the null hypothesis. The proportion of not of white ethnicity is larger than 10%

4. Testing for restaurants having smoking restrictions

4.1 General

```
## [1] 0.7534077
```

The result show that 75,34% data are restaurants not having smoking restrictions. So for now we will test whether this proportion is larger than 75% or not.

Hypothesis:

$$H_0: \rho_0 = 0.75$$

$$H_1: \rho_0 < 0.75$$

```
##
##  1-sample proportions test with continuity correction
##
## data:  table(restaurn), null probability 0.75
## X-squared = 0.033457, df = 1, p-value = 0.5726
## alternative hypothesis: true p is less than 0.75
## 95 percent confidence interval:
##  0.0000000 0.7780873
## sample estimates:
##         p
## 0.7534077
```

The obtained p-value is p-value = 0.5726 which is greater than the predetermined significance level of α=0.05 for a confidence level of 95%. Consequently, we accept the null hypothesis. The proportion of restaurants not having smoking restrictions is larger than 75%.

*4.2 Testing relationship between smokers and restaurant having smoking restrictions***

```
##          smoker
## restaurn   0   1
##        0 359 249
##        1 138  61

## Smoking restrictions and non-smoker:  0.1710037

## Smoking restrictions and smoker:  0.0755886
```

Hypothesis:

ρ_1: proportion of smoking restrictions in states' restaurant and non-smoker

ρ_2: proportion of smoking restrictions in states' restaurant and smoker

$$H_0: \rho_{10} = \rho_{20}$$

$$H_1: \rho_{10} > \rho_{20}$$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(nrow(data[cigs == 0 & restaurn == 1, ]), nrow(data[cigs > 0 &
restaurn == 1, ])) out of c(length(cigs), length((restaurn)))
## X-squared = 33.107, df = 1, p-value = 4.361e-09
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.06753894 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.1710037 0.0755886
```

The obtained p-value is p-value = 4.361e-09 which is smaller than the predetermined significance level of α=0.05 for a confidence level of 95%. Consequently, we reject the null hypothesis. It means that with states having smoking restrictions in restaurants, proportion of non-smokers is higher than smokers.

III. Linear regression model

According to the data, the number of cigarettes smoked per day is the attribute that is most commonly estimated by others. As a result, we are currently working on developing a model that focuses on the cigs attribute.

1. Pre-process

Freshing data:

In the above section, it has show that there is no missing values in the data so that we can skip this stage.

Removing outliers:

Calculate the total outliers row :

```
## [1] 332
```

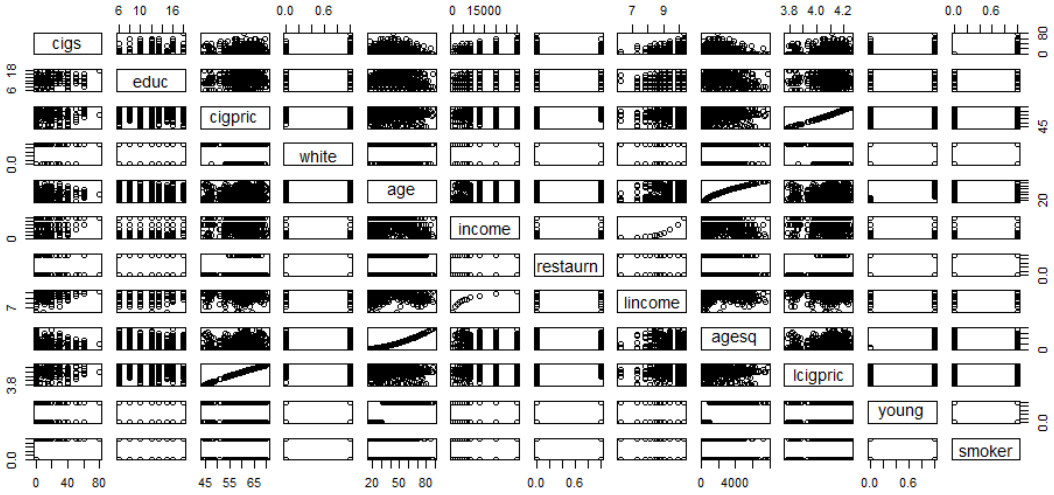
```
## [1] 0.4114002
```

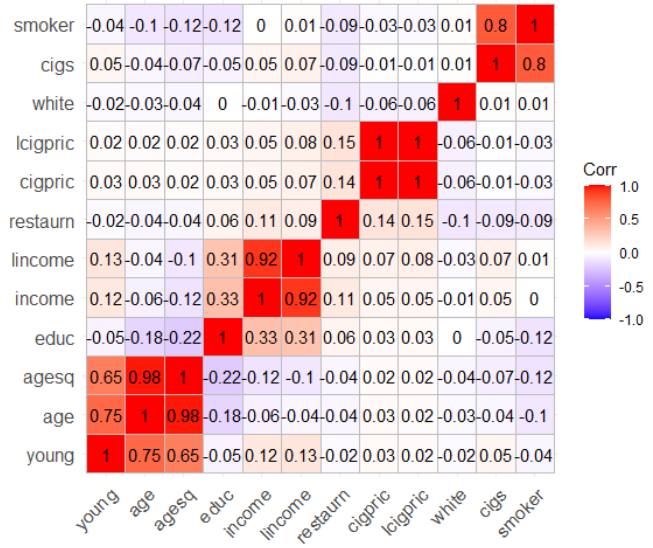
Observation: Since the number of outliers row is too much 332/807 observations (41.14%) so that we should not remove all of them.

Add categories for age and cigs:

```
#smoker =1 if cigs>0
smoker=1:length(cigs)
for (i in 1:length(cigs))
{if (cigs[i]<1)
{smoker[i]=0}
  else
  {smoker[i]=1}
}

#ageGroup=1 if age>30
ageGroup=1:length(age)
for (i in 1:length(age))
{if (age[i]<=30)
{ageGroup[i]=0}
  else
  {ageGroup[i]=1}
}

data$young<-ageGroup
data$smoker<-smoker
```

2. Checking correlation between variables






```
##       educ    cigpric      white        age     income    restaurn
lincome
##   1.224826 224.750335   1.023724  75.017783   6.711976   1.072329
6.654113
##      agesq   lcigpric      young     smoker
##  59.585589 225.349349   4.173298   1.068535
```

Conclusion: The Variance Inflation Factor (VIF) is a measure of the strength of the correlation between independent variables in a multiple regression model. A VIF value of less than 5 indicates that there is no strong correlation between the variables. In this case, the VIF values for cigpric, lcigpric, income, lincome age and agesq are strongly correlated to each other.

3. Building models

Starting the full model with all the variables except cigpric, age, income:

```
##
## Call:
## lm(formula = data$cigs ~ (. - age - cigpric - income), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.115  -2.153  -0.544   1.564  54.705
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.276e+01  1.464e+01  -0.871 0.383777
## educ         1.743e-01  1.016e-01   1.715 0.086796 .
## white        1.507e-01  8.789e-01   0.171 0.863915
## restaurn    -5.899e-01  6.768e-01  -0.872 0.383694
## lincome      5.920e-01  4.366e-01   1.356 0.175447
## agesq       -3.261e-04  2.533e-04  -1.287 0.198294
## lcigpric     8.484e-01  3.488e+00   0.243 0.807907
## young        3.089e+00  8.186e-01   3.773 0.000173 ***
## smoker       2.269e+01  6.011e-01  37.743  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.091 on 798 degrees of freedom
## Multiple R-squared:  0.6558, Adjusted R-squared:  0.6523
## F-statistic:   190 on 8 and 798 DF,  p-value: < 2.2e-16

##     educ    white restaurn  lincome    agesq lcigpric    young   smoker
## 1.188510 1.016029 1.049184 1.191966 1.964577 1.030105 1.881114 1.053904
```

The p-value for the cigpric and white variable is large enough to warrant further investigation into whether it can be eliminated from the analysis

```
##
## Call:
## lm(formula = data$cigs ~ (. - age - lcigpric - income - white -
##     cigpric), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.133  -2.133  -0.516   1.476  54.734
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.1943965  4.0653393  -2.262 0.023986 *
## educ         0.1744909  0.1015018   1.719 0.085985 .
## restaurn    -0.5786109  0.6660937  -0.869 0.385292
## lincome      0.5962013  0.4350923   1.370 0.170982
```

```
## agesq       -0.0003265  0.0002527  -1.292 0.196566
## young        3.0909894  0.8175168   3.781 0.000168 ***
## smoker      22.6864997  0.6003543  37.789  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.081 on 800 degrees of freedom
## Multiple R-squared:  0.6557, Adjusted R-squared:  0.6531
## F-statistic:   254 on 6 and 800 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: data$cigs ~ ((educ + cigpric + white + age + income + restaurn +
##     lincome + agesq + lcigpric + young + smoker) - age - cigpric -
##     income)
## Model 2: data$cigs ~ ((educ + cigpric + white + age + income + restaurn +
##     lincome + agesq + lcigpric + young + smoker) - age - lcigpric -
##     income - white - cigpric)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    798 52240
## 2    800 52245 -2   -5.5633 0.0425 0.9584
```

The summary of the new reduced model indicates that subtracting additional variables only results in a 0.08% increase in the adjusted R-squared value. This suggests that the cost of excepting the extra variables may not be worth the small improvement in the model's explanatory power. Additionally, the p-value in the anova table (0.9584) also support that comment. But seeking through the summary, the restaurn and agesq have proportions can be eliminate so further testing is needed.

```
##
## Call:
## lm(formula = data$cigs ~ (. - age - agesq - cigpric - income -
##     white - lcigpric - restaurn), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.219  -2.155  -0.674   1.328  54.887
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.58384    3.88768  -2.722  0.00662 **
## educ          0.19863    0.09953   1.996  0.04630 *
## lincome       0.66666    0.42691   1.562  0.11877
## young         2.39864    0.60583   3.959 8.18e-05 ***
## smoker       22.84987    0.59103  38.661  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.083 on 802 degrees of freedom
```

```
## Multiple R-squared:  0.6547, Adjusted R-squared:  0.653
## F-statistic: 380.2 on 4 and 802 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: data$cigs ~ ((educ + cigpric + white + age + income + restaurn +
##     lincome + agesq + lcigpric + young + smoker) - age - cigpric -
##     income)
## Model 2: data$cigs ~ ((educ + cigpric + white + age + income + restaurn +
##     lincome + agesq + lcigpric + young + smoker) - age - agesq -
##     cigpric - income - white - lcigpric - restaurn)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    798 52240
## 2    802 52400 -4   -160.56 0.6132 0.6532
```

The summary of the new reduced model indicates that including additional variables only results in a 0.07% increase in the adjusted R-squared value. This suggests that the cost of including the extra variables may not be worth the small improvement in the model's explanatory power. Additionally, the p-value in the anova table (0.6532) also support that comment.

Beside the original ways, there is another way to choose the good model based on the AIC :

```
## Start:  AIC=3383.41
## data$cigs ~ ((educ + cigpric + white + age + income + restaurn +
##     lincome + agesq + lcigpric + young + smoker) - age - cigpric -
##     income)
##
##            Df Sum of Sq    RSS    AIC
## - white     1         2  52241 3381.4
## - lcigpric  1         4  52243 3381.5
## - restaurn  1        50  52289 3382.2
## - agesq     1       109  52348 3383.1
## - lincome   1       120  52360 3383.3
## <none>                   52240 3383.4
## - educ      1       192  52432 3384.4
## - young     1       932  53172 3395.7
## - smoker    1     93256 145496 4208.0
##
## Step:  AIC=3381.44
## data$cigs ~ educ + restaurn + lincome + agesq + lcigpric + young +
##     smoker
##
##            Df Sum of Sq    RSS    AIC
## - lcigpric  1         4  52245 3379.5
## - restaurn  1        52  52294 3380.2
## - agesq     1       110  52351 3381.1
## - lincome   1       120  52361 3381.3
## <none>                   52241 3381.4
## - educ      1       192  52434 3382.4
```

```
## + white     1         2  52240 3383.4
## - young     1       934  53175 3393.7
## - smoker    1     93258 145499 4206.0
##
## Step:  AIC=3379.49
## data$cigs ~ educ + restaurn + lincome + agesq + young + smoker
##
##            Df Sum of Sq    RSS    AIC
## - restaurn  1        49  52294 3378.3
## - agesq     1       109  52354 3379.2
## - lincome   1       123  52368 3379.4
## <none>                   52245 3379.5
## - educ      1       193  52438 3380.5
## + lcigpric  1         4  52241 3381.4
## + white     1         2  52243 3381.5
## - young     1       934  53179 3391.8
## - smoker    1     93256 145501 4204.1
##
## Step:  AIC=3378.26
## data$cigs ~ educ + lincome + agesq + young + smoker
##
##            Df Sum of Sq    RSS    AIC
## - agesq     1       106  52400 3377.9
## - lincome   1       112  52406 3378.0
## <none>                   52294 3378.3
## - educ      1       190  52485 3379.2
## + restaurn  1        49  52245 3379.5
## + white     1         4  52290 3380.2
## + lcigpric  1         1  52294 3380.2
## - young     1       937  53232 3390.6
## - smoker    1     94466 146761 4209.0
##
## Step:  AIC=3377.89
## data$cigs ~ educ + lincome + young + smoker
##
##            Df Sum of Sq    RSS    AIC
## <none>                   52400 3377.9
## + agesq     1       106  52294 3378.3
## - lincome   1       159  52559 3378.3
## + restaurn  1        46  52354 3379.2
## + white     1         6  52394 3379.8
## + lcigpric  1         0  52400 3379.9
## - educ      1       260  52660 3379.9
## - young     1      1024  53424 3391.5
## - smoker    1     97657 150057 4224.9

##
## Call:
## lm(formula = data$cigs ~ educ + lincome + young + smoker, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.219  -2.155  -0.674   1.328  54.887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.58384    3.88768  -2.722  0.00662 **
## educ          0.19863    0.09953   1.996  0.04630 *
## lincome       0.66666    0.42691   1.562  0.11877
## young         2.39864    0.60583   3.959 8.18e-05 ***
## smoker       22.84987    0.59103  38.661  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.083 on 802 degrees of freedom
## Multiple R-squared:  0.6547, Adjusted R-squared:  0.653
## F-statistic: 380.2 on 4 and 802 DF,  p-value: < 2.2e-16
```

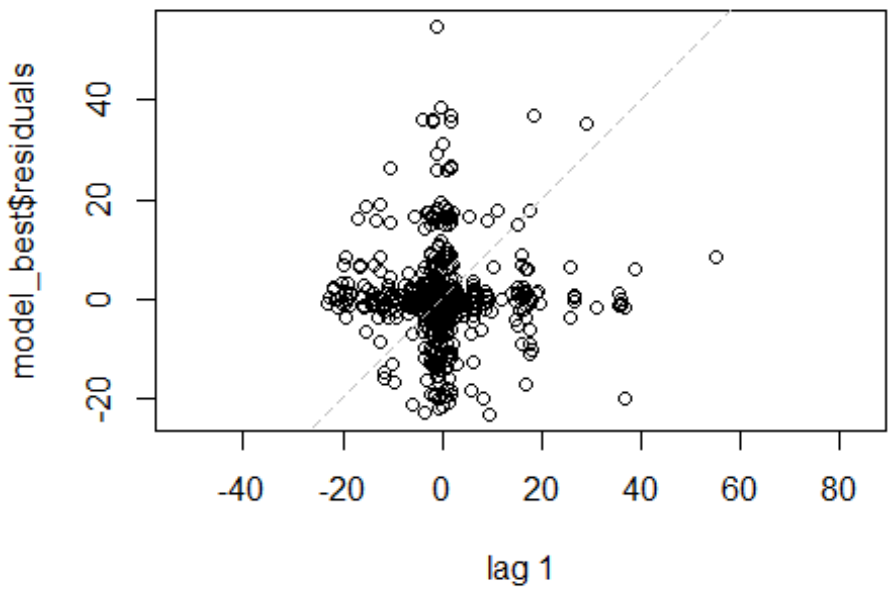
→ The result model is the same as the model with except age, agesq, cigpric, lcigpric, income, white and restaurn found before.

4. Checking the model

Independent check using Durbin-Watson Test:

H_0 : linear regression residuals of time series data are uncorrelated

H_1 : autocorrelation exists.




```
##  lag Autocorrelation D-W Statistic p-value
##    1    0.0004601491      1.998794   0.948
##  Alternative hypothesis: rho != 0
```

Conclusion: The result show that the p-value is 0.954, which means autocorrelation does not exist.

Stability check using Breusch-Pagan Test:

H_0: Homoscedasticity is present (the residuals are distributed with equal variance)

H_1: Heteroscedasticity is present (the residuals are not distributed with equal variance)

```
##
##  studentized Breusch-Pagan test
##
## data:  model_best
## BP = 133.99, df = 4, p-value < 2.2e-16
```

Conclusion: The result show that it reject null hypothesis at any significant level. → The residuals are not distributed with equal variance.

Normality check using Shapiro-Wilk normality test: H_0 : the sample has been generated from a normal distribution

H_1 : the sample has not been generated from a normal distribution

```
##
##  Shapiro-Wilk normality test
##
## data:  model_best$residuals
## W = 0.80349, p-value < 2.2e-16
```

Conclusion: The p-value is too small to accept null hypothesis at any significant level.

5. Analysis and interpretation of the model results

```
##
## Call:
## lm(formula = data$cigs ~ educ + lincome + young + smoker, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.219  -2.155  -0.674   1.328  54.887
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.58384    3.88768  -2.722  0.00662 **
## educ          0.19863    0.09953   1.996  0.04630 *
## lincome       0.66666    0.42691   1.562  0.11877
## young         2.39864    0.60583   3.959 8.18e-05 ***
## smoker       22.84987    0.59103  38.661  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.083 on 802 degrees of freedom
## Multiple R-squared:  0.6547, Adjusted R-squared:  0.653
## F-statistic: 380.2 on 4 and 802 DF,  p-value: < 2.2e-16

##     educ  lincome    young   smoker
## 1.142144 1.141964 1.032241 1.020722
```

All the coefficients of the function model: $\widehat{\beta_0}$=-10.58384, $\widehat{\beta_1}$=0.19863, $\widehat{\beta_2}$= 0.66666, $\widehat{\beta_3}$=2.39864, $\widehat{\beta_4}$=22.84987

Explain the meaning of coefficients

- $\widehat{\beta_0}$ = -10.58384: When years of education, log of income, young and smoker are 0, and the estimate value of number of cigarettes smoked per day is -10.58384
- $\widehat{\beta_1}$ = 0.19863: When number years of education increases by 1 and other variables remain the same, the estimate number of cigarettes smoked per day increases by 0.19863.
- $\widehat{\beta_2}$ = 0.66666: When log of income increases by 1 and other variables remain the same, the estimate number of cigarettes smoked per day increases by 0.66666.
- $\widehat{\beta_3}$ = 2.39864: When the young=1 and other variables remain the same, the estimate number of cigarettes smoked per day will increase by 2.39864.

- $\widehat{\beta_4}$ = 22.84987: When the smoker=1 and other variables remain the same, the estimate number of cigarettes smoked per day will increase by 22.84987

From the result, the fitted least squares regression model for number of cigarettes smoked per day is:

$\hat{Y} = -10.58384 + 0.19863 \times educ + 0.66666 \times lincome + 2.39864 \times young + 22.84987 \times smoker$

Meaning of adjusted R-squared : The adjusted R-squared value is 0.6547. Therefore, the independent variables explain 65.47% of the variance in the dependent variable. The remaining 34.53% is explained by external variables and random error.

Model Utility Test

```
## Analysis of Variance Table
##
## Response: data$cigs
##            Df Sum Sq Mean Sq   F value    Pr(>F)
## educ        1    360     360    5.5071   0.01918 *
## lincome     1   1112    1112   17.0149 4.096e-05 ***
## young       1    225     225    3.4405   0.06398 .
## smoker      1  97657   97657 1494.6755 < 2.2e-16 ***
## Residuals 802  52400      65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

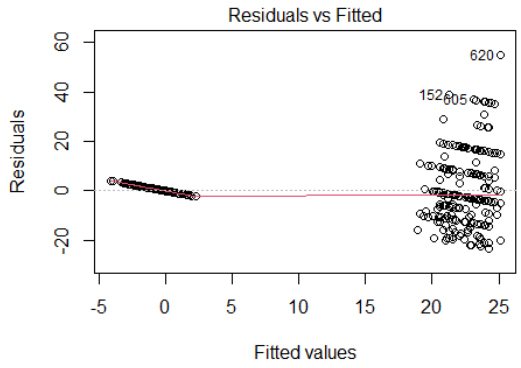
For all the variables, with the confidence level 90% (α = 0.1), p-value < 0.1

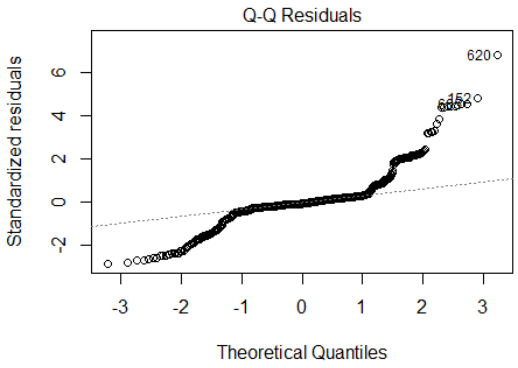
$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

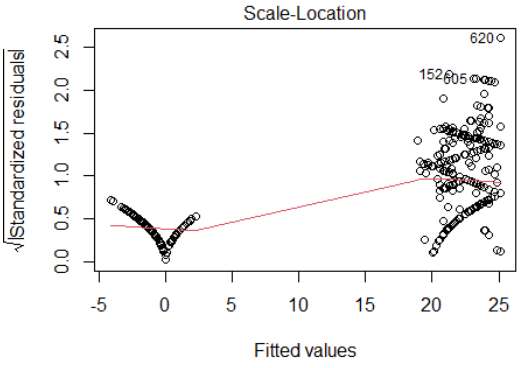
$$H_1: at\ least\ one\ \beta_i \neq 0 (i = 1,2,3,4)$$

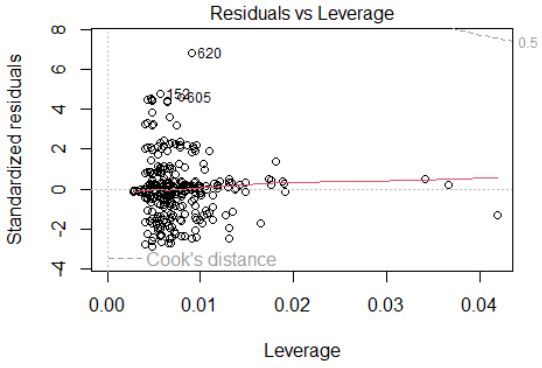
Since p-value < α, the null hypothesis is rejected. We can conclude that the variables that we are testing is useful in the regression model.

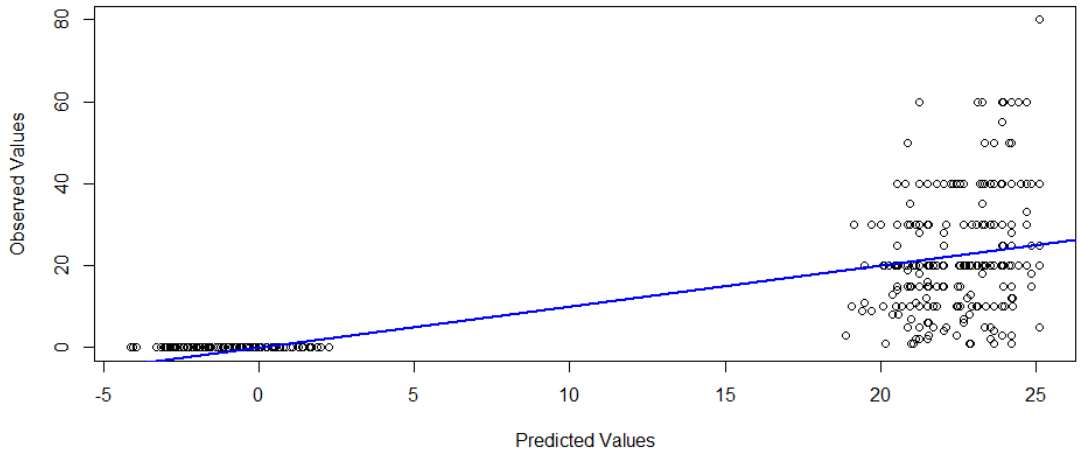
Therefore, all the variables are useful in the regression model.











Conclusion: The number of cigarettes smoked per day increases with years of education, income, and whether a person is a smoker or under 31 years old. Price of cigarettes, living in states having smoking restrictions in restaurant, being whites have small impacts on cigarettes' value.

Activity 02

I. Introduction

The data consists of 400 observations about graduate admission chances, each record has 9 description columns which are showed below :

- Serial.No: Identification number
- GRE.Score: Graduate Record Examination)
- TOEFL.Score: Test of English as a Foreign Language
- University.Rating: University ranking
- SOP: Statement of Purpose
- LOR: Letter of Recommendation
- CGPA: Cumulative Grade Point Average
- Research: 1 if the applicant has research experience, 0 otherwise
- Chance.of.Admit: Likelihood of admission

Conclusion: This dataset offers a robust foundation for investigating the intricate interplay of these attributes in shaping the outcomes of graduate admissions. By exploring the relationships embedded within these columns, researchers and analysts can unveil valuable insights into the factors contributing to successful admissions.

1. Setting environment for data

```
dataGA<-read.csv('US_graduate_schools_admission_parameters_dataset.csv',
header = TRUE)
str(dataGA)
```

```
## 'data.frame':    400 obs. of  9 variables:
##  $ Serial.No.       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ GRE.Score        : int  337 324 316 322 314 330 321 308 302 323 ...
##  $ TOEFL.Score      : int  118 107 104 110 103 115 109 101 102 108 ...
##  $ University.Rating: int  4 4 3 3 2 5 3 2 1 3 ...
##  $ SOP              : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
##  $ LOR              : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
##  $ CGPA             : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
##  $ Research         : int  1 1 1 1 0 1 1 0 0 0 ...
##  $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45
...
```

2. Data size

```
dim(dataGA)
```

```
## [1] 400   9
```

```
sum(is.na(dataGA))
```

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

The result shows that this data has 400 rows (observations) and 9 columns (variables).And the second row show that no missing values are present in the data set. ### 3. Brief summary of the data

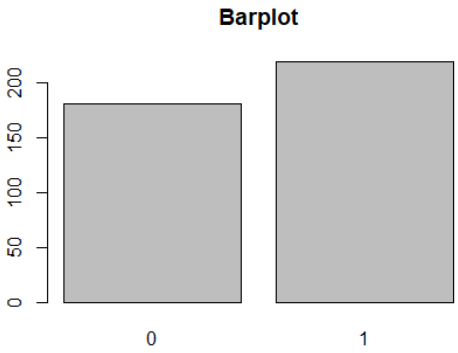
```
##    Serial.No.      GRE.Score      TOEFL.Score    University.Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
##  1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :200.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087
##  3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP             LOR             CGPA           Research
##  Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000
##  1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000
##  Median :3.5   Median :3.500   Median :8.610   Median :1.0000
##  Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475
##  3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000
##  Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000
##  Chance.of.Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

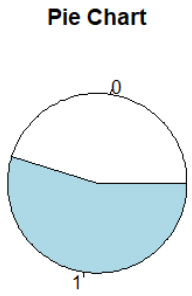
- GRE.Score and TOEFL.Score: The mean and median for both scores are close, indicating a balanced distribution of academic readiness. Notably, the relatively high maximum values (340 for GRE, 120 for TOEFL) showcase exceptional performance.
- University.Rating, SOP, and LOR: The median values align closely with the means, suggesting symmetrical distributions for university ratings, SOP, and LOR. The quartiles reveal a consistent evaluation trend across institutions and recommenders.
- CGPA: The slightly higher mean compared to the median in CGPA indicates a positively skewed distribution. The broader spread towards higher values underscores a notable concentration of strong academic performances.
- Research: The binary distribution of research experience is evident, with approximately half of applicants having research backgrounds. This highlights the substantial representation of candidates with practical engagement.
- Chance.of.Admit: The symmetrical distribution of admission probabilities, with mean and median closely aligned, implies a competitive range. Quartiles further emphasize the diverse probabilities among applicants.

II. Descriptive Statistics

1. Research experience

```
##
##   0   1
## 181 219
```

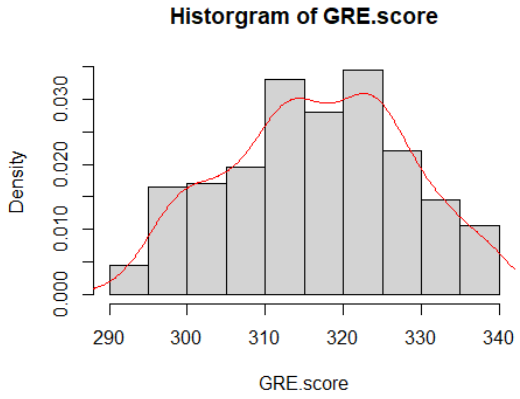


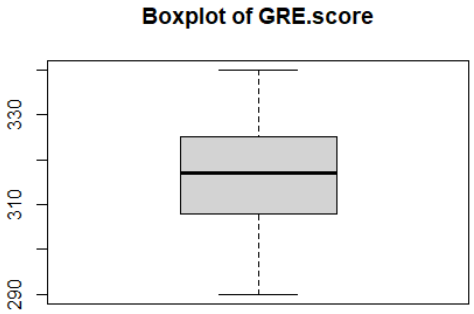


→ Over half of observers have experience in research.

2. Graduate Record Examination score

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   290.0   308.0   317.0   316.8   325.0   340.0
```

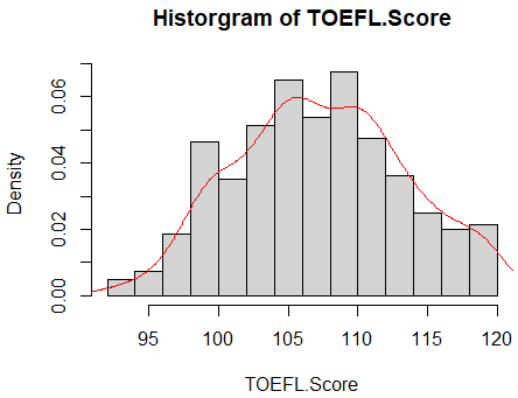


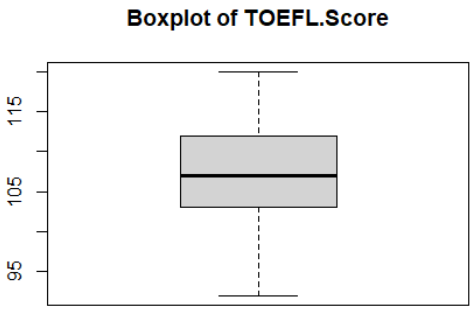


→ The majority of applicants exhibit GRE scores within the range of 310 to 320, indicating a concentration around this interval. The absence of outliers, as depicted by the boxplot, suggests a consistent distribution of scores without extreme deviations.

3. Test of English as a Foreign Language

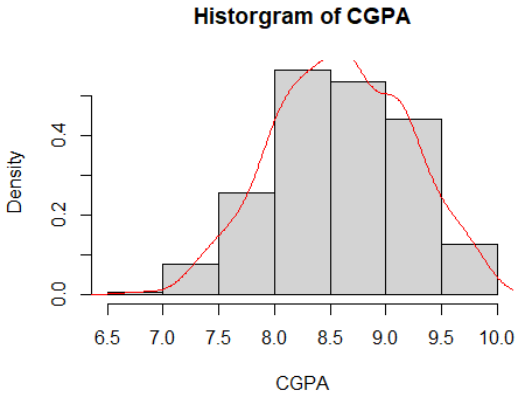
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    92.0   103.0   107.0   107.4   112.0   120.0
```

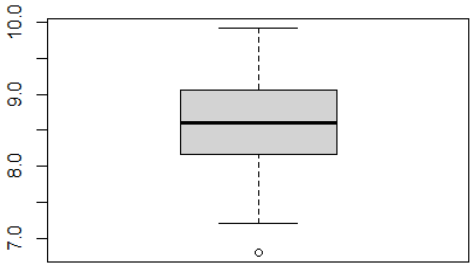


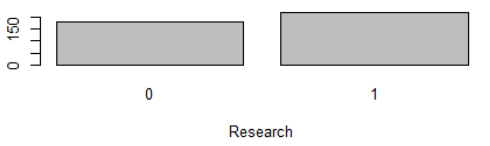


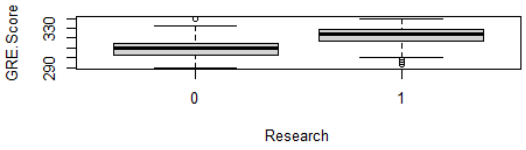
→ Most applicants have TOEFL scores between 105 and 110, and the boxplot indicates a consistent distribution without outliers. ### 4. Cumulative Grade Point Average

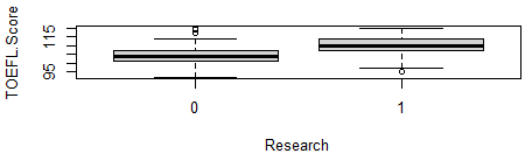
```
## Length  Class   Mode
##      0   NULL   NULL
```

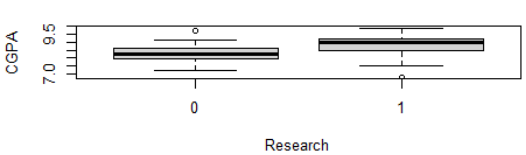








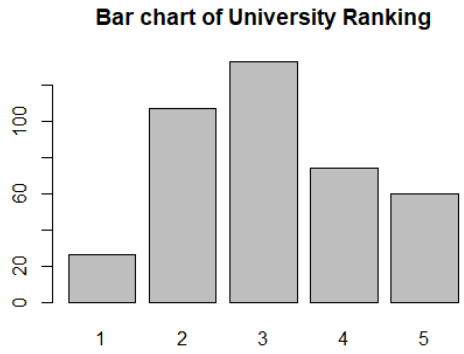


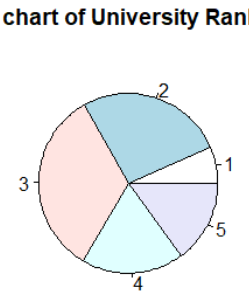


- CGPA for most applicants falls between 8.25 and 9.0, with one outlier.
- Boxplots indicate that candidates with research experience tend to achieve higher GRE, TOEFL, and CGPA scores compared to those without research experience.

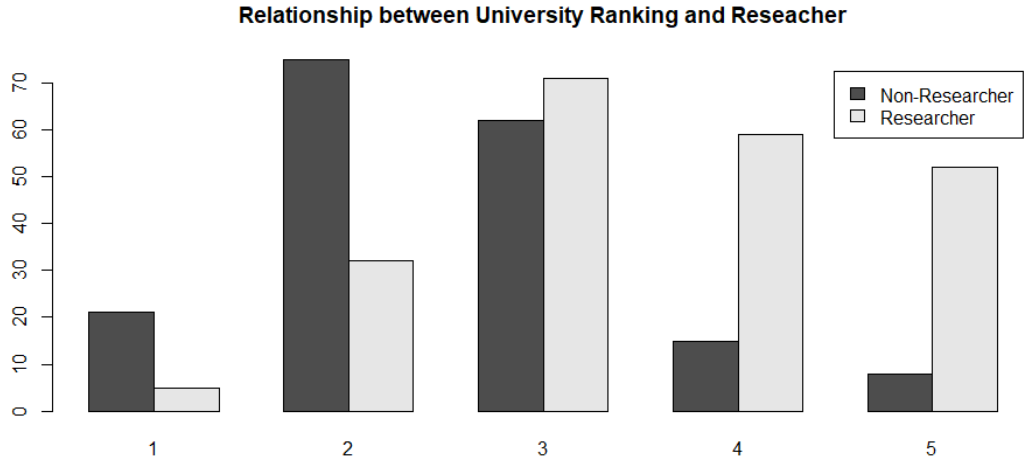
5. University Ranking

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.087   4.000   5.000
```





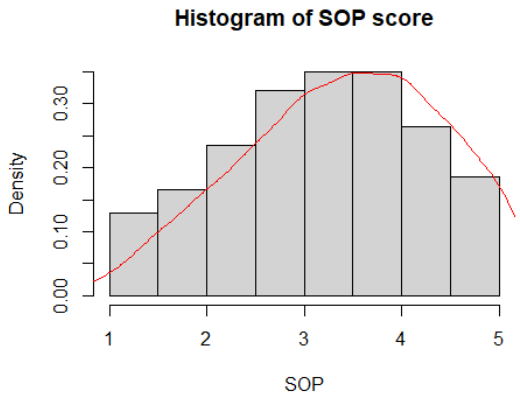

```
##
##      1  2  3  4  5
##   0 21 75 62 15  8
##   1  5 32 71 59 52
```

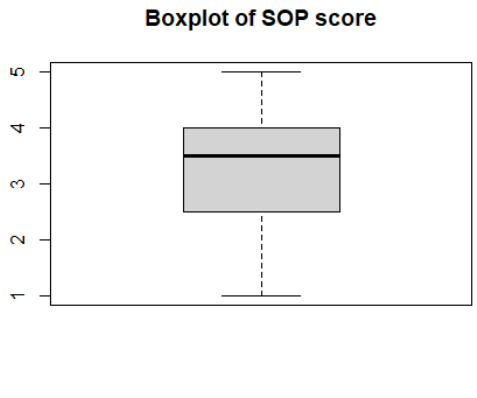


The majority of universities are ranked between 2 and 4. Interestingly, within the range of rankings 3 to 5, a higher number of applicants possess research experience compared to those without. Conversely, for universities ranked 1 to 2, there is a lower number of applicants with research experience in comparison to those without.

6. Statement of Purpose

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     2.5     3.5     3.4     4.0     5.0
```

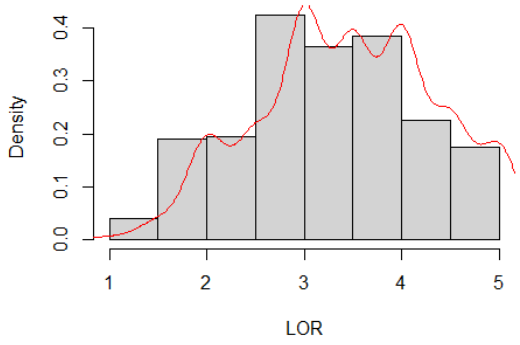


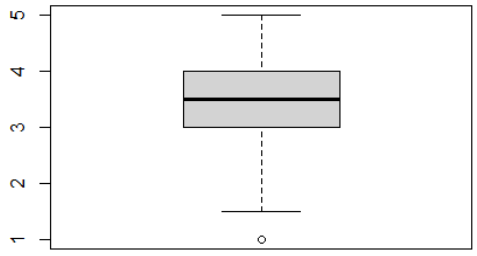


→ The majority of applicants exhibit Statement of Purpose scores ranging from 2 to 4.5, without any outliers.

7. Letter of Recommendation

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.500   3.453   4.000   5.000
```

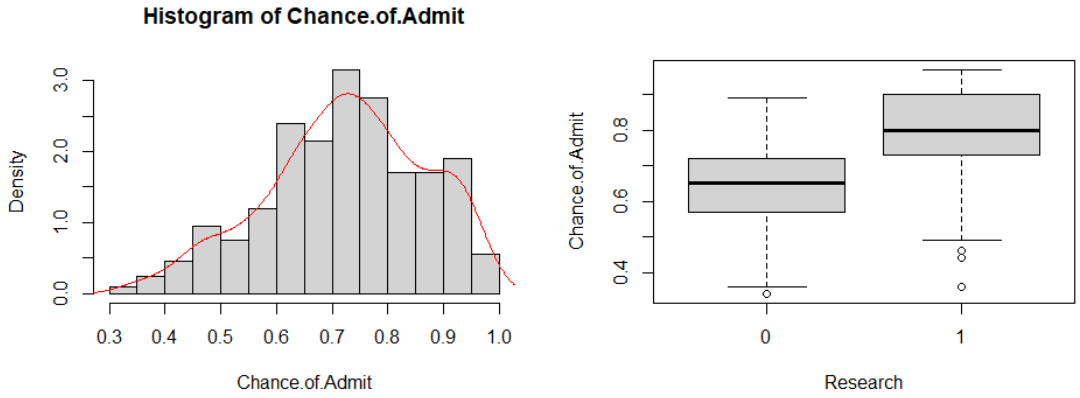




→ Most applicants' Letter of Recommendation scores fall within the range of 2.5 to 4.

8. Chance of admission

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3400  0.6400  0.7300  0.7244  0.8300  0.9700
```



→ Notably, a substantial portion of admissions probabilities cluster between 0.6 to 0.8. Intriguingly, applicants with research experience consistently demonstrate a higher proportion of favorable admission chances compared to those without such experience. This observation underscores the potential positive impact of research engagement on admission outcomes.

III. Linear regression model

1. Pre-process

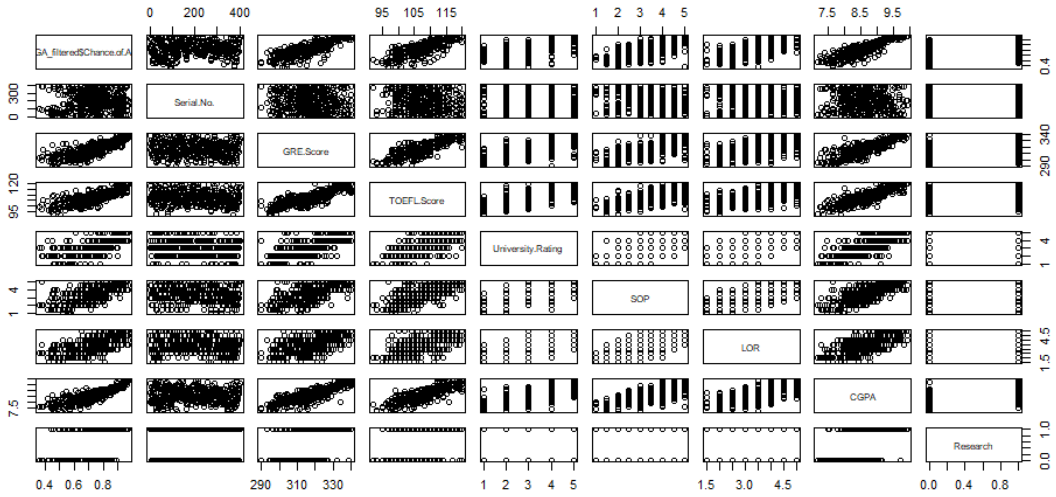
Checking outliers

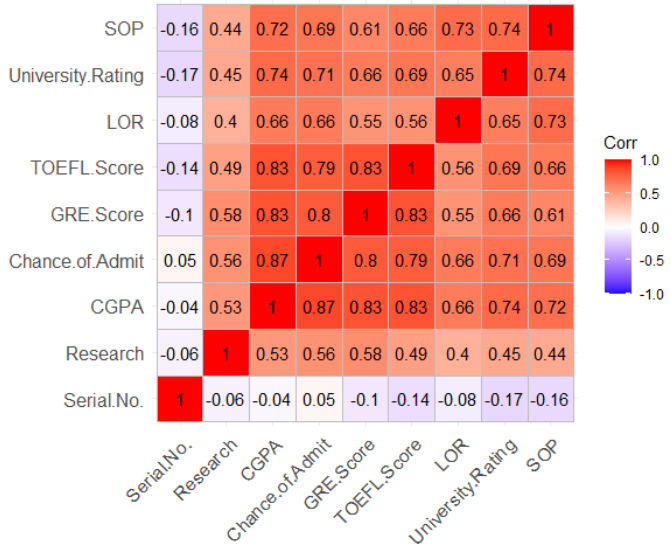
```
result_vector <- unlist(apply(dataGA[,c(1,2,3,4,5,6,7,8,9)],2,outliers))
length(unique(result_vector[result_vector != 0]))

## [1] 4

length(unique(result_vector[result_vector != 0])) / nrow(dataGA)

## [1] 0.01
```

Observation: Since the number of outliers row is not too much, 4 observations (1%) so that they could be remove.

```
dataGA_filtered <- dataGA[-unique(result_vector[result_vector != 0]),]
dim(dataGA_filtered)

## [1] 396   9
```

Checking multicollinearity:






```
##          Serial.No.           GRE.Score         TOEFL.Score University.Rating
##            1.084580            4.513560            4.251884            2.916775
##                 SOP                 LOR                CGPA            Research
##            3.149879            2.379183            5.359569            1.547918
```

Conclusion: The Variance Inflation Factor (VIF) is a measure of the strength of the correlation between independent variables in a multiple regression model. A VIF value of less than 10 indicates that there is no strong correlation between the variables. In this case, the VIF values for all variables are less than 10, which means that there are no variables that are strongly correlated with each other.

Data splitting for training and validation process

```
## Dimensions of training_data: 317 9

## Dimensions of validation_data: 79 9
```

The training dataset will contain a random selection of 80% of the original data, and the validation dataset will contain the remaining 20% of the data.

```
##         GRE.Score       TOEFL.Score University.Rating               SOP
##          4.869778          4.590635          2.967035          3.005907
##              LOR              CGPA          Research
##          2.354207          5.476878          1.591776

##         GRE.Score       TOEFL.Score University.Rating               SOP
##          3.442388          3.184291          2.804954          4.289578
##              LOR              CGPA          Research
##          2.720861          4.270238          1.664103
```

2. Building model

At first, we examine the linear model with chance of admission be the dependent variable.

```
##
## Call:
## lm(formula = Chance.of.Admit ~ (. - Serial.No.), data = training_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.233360 -0.021373  0.008385  0.034445  0.151535
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.2027921  0.1301184  -9.244  < 2e-16 ***
## GRE.Score          0.0016774  0.0006259   2.680  0.00776 **
## TOEFL.Score        0.0028609  0.0011541   2.479  0.01371 *
## University.Rating  0.0038583  0.0049938   0.773  0.44033
## SOP                0.0005385  0.0056141   0.096  0.92364
## LOR                0.0198654  0.0056972   3.487  0.00056 ***
## CGPA               0.1158878  0.0131887   8.787  < 2e-16 ***
## Research           0.0245160  0.0084050   2.917  0.00379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05906 on 309 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.8164
## F-statistic: 201.7 on 7 and 309 DF,  p-value: < 2.2e-16
```

Choosing the best model via the AIC criterion :

```
## Start:  AIC=-1785.85
## Chance.of.Admit ~ ((Serial.No. + GRE.Score + TOEFL.Score +
University.Rating +
##     SOP + LOR + CGPA + Research) - Serial.No.)
##
##                     Df Sum of Sq    RSS     AIC
## - SOP                1  0.000032 1.0777 -1787.8
```

```
## - University.Rating  1  0.002082 1.0798 -1787.2
## <none>                           1.0777 -1785.8
## - TOEFL.Score        1  0.021433 1.0991 -1781.6
## - GRE.Score          1  0.025047 1.1027 -1780.6
## - Research           1  0.029673 1.1074 -1779.2
## - LOR                1  0.042404 1.1201 -1775.6
## - CGPA               1  0.269283 1.3470 -1717.2
##
## Step:  AIC=-1787.84
## Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating +
##     LOR + CGPA + Research
##
##                     Df Sum of Sq    RSS     AIC
## - University.Rating  1  0.002505 1.0802 -1789.1
## <none>                           1.0777 -1787.8
## + SOP                1  0.000032 1.0777 -1785.8
## - TOEFL.Score        1  0.021766 1.0995 -1783.5
## - GRE.Score          1  0.025015 1.1027 -1782.6
## - Research           1  0.029745 1.1075 -1781.2
## - LOR                1  0.050541 1.1283 -1775.3
## - CGPA               1  0.276015 1.3537 -1717.6
##
## Step:  AIC=-1789.11
## Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + CGPA + Research
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           1.0802 -1789.1
## + University.Rating  1  0.002505 1.0777 -1787.8
## + SOP                1  0.000455 1.0798 -1787.2
## - GRE.Score          1  0.025220 1.1055 -1783.8
## - TOEFL.Score        1  0.025614 1.1058 -1783.7
## - Research           1  0.029478 1.1097 -1782.6
## - LOR                1  0.065579 1.1458 -1772.4
## - CGPA               1  0.306180 1.3864 -1712.0

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR +
##     CGPA + Research, data = training_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.231838 -0.021396  0.007437  0.035194  0.150421
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2415291  0.1213790 -10.229  < 2e-16 ***
## GRE.Score    0.0016818  0.0006241   2.695  0.00743 **
## TOEFL.Score  0.0030571  0.0011258   2.716  0.00699 **
## LOR          0.0215602  0.0049619   4.345 1.89e-05 ***
```

```
## CGPA         0.1187133  0.0126441   9.389  < 2e-16 ***
## Research     0.0244236  0.0083838   2.913  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05894 on 311 degrees of freedom
## Multiple R-squared:   0.82,  Adjusted R-squared:  0.8171
## F-statistic: 283.4 on 5 and 311 DF,  p-value: < 2.2e-16
```

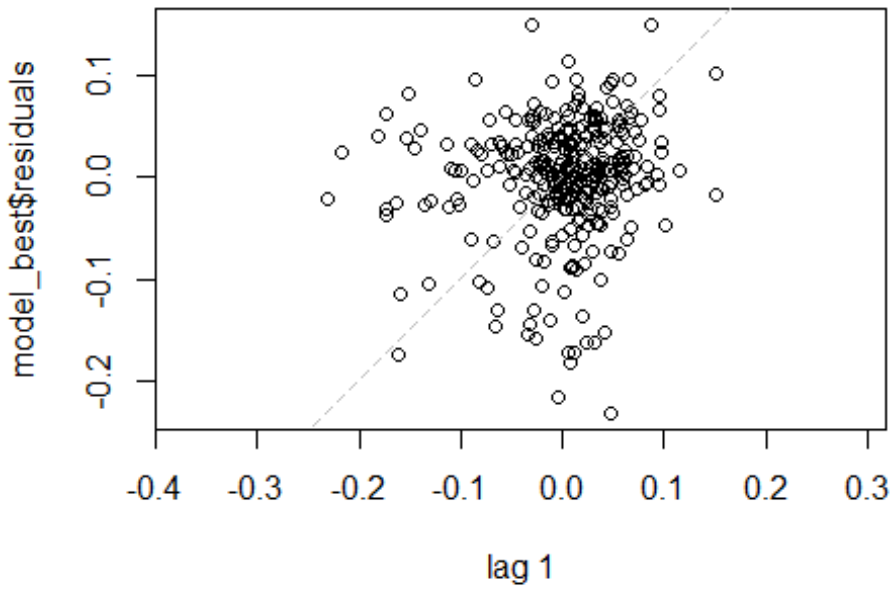
Observation: The p-value of all variables are significant small, that lead to that conclusion that non of them could be eliminated from this model.

3. Checking the model

Independent check using Durbin-Watson Test:

H_0 : linear regression residuals of time series data are uncorrelated

H_1 : autocorrelation exists.




```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1166149      1.766329    0.05
##  Alternative hypothesis: rho != 0
```

Conclusion: The p-value associated with the test is 0.034 → Reject null hypothesis, linear regression residuals of time series data are not uncorrelated

Stability check using Breusch-Pagan Test:

H_0: Homoscedasticity is present (the residuals are distributed with equal variance)

H_1: Heteroscedasticity is present (the residuals are not distributed with equal variance)

```
##
##  studentized Breusch-Pagan test
##
## data:  model_best
## BP = 17.875, df = 5, p-value = 0.003107
```

Conclusion: The p-value of the result is 0.003107, hence suggests that there is evidence to reject the null hypothesis, and the residuals are not distributed with equal variance.

Normality check using Shapiro-Wilk normality test:

H_0 : the sample has been generated from a normal distribution

H_1 : the sample has not been generated from a normal distribution

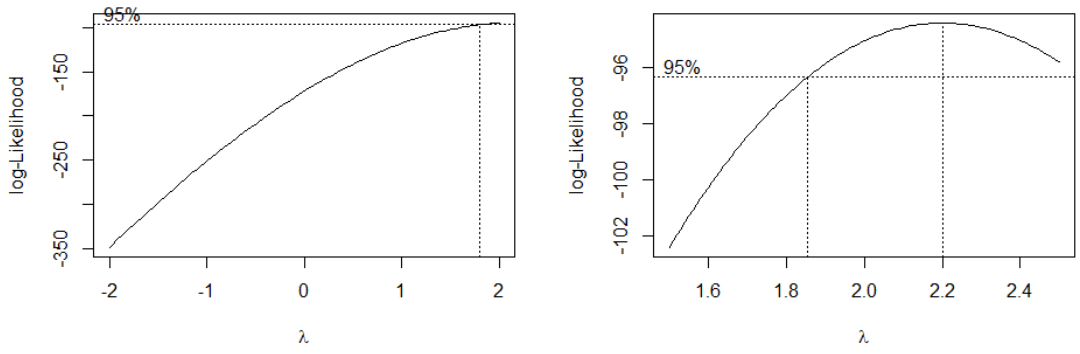
```
##
##  Shapiro-Wilk normality test
##
## data:  model_best$residuals
## W = 0.92538, p-value = 1.729e-11
```

Conclusion: The p-value is too small to accept null hypothesis at any significant level. Therefore, the data does not follow a normal distribution, and those normal assumptions we made above are likely not to be valid.

4. Transformation model

By the reason that our model did not pass through the normality checking, we are now transformation the model using Box-Cox.

Log-likelihood function of λ:



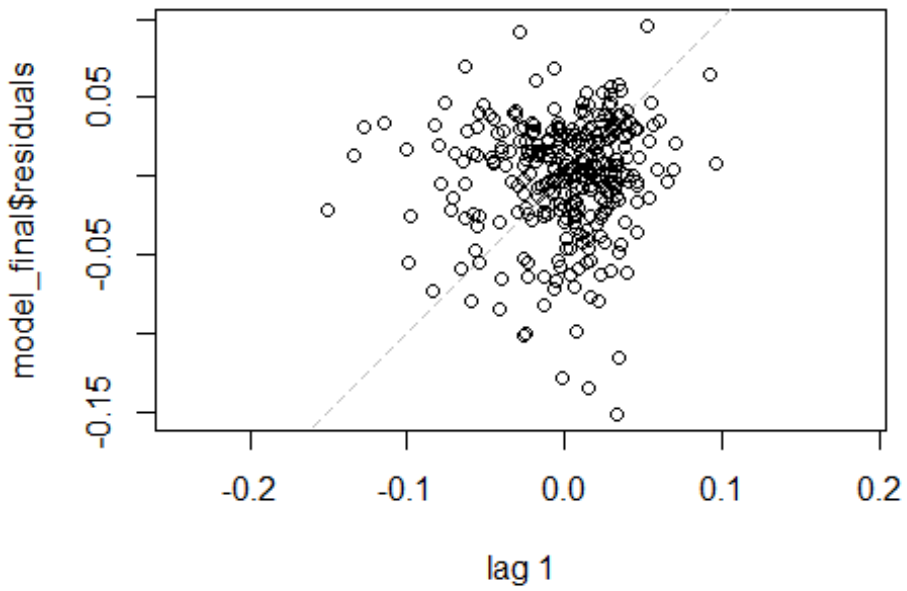
Observation: The confidence interval for the λ value that maximizes the log-likelihood is close the range [1.5,2.5]. Here we see that λ = 2.2 is both in the confidence interval, and is extremely close to the maximum. This suggests a transformation of the form:

$$\frac{y^{2.2}-1}{2.2}$$

Therefore, we create a new model after transformation:

```
model_final<-lm((Chance.of.Admit^2.2-1)/2.2 ~ GRE.Score + TOEFL.Score + LOR +
    CGPA + Research, data = training_data)
summary(model_final)

##
## Call:
## lm(formula = (Chance.of.Admit^2.2 - 1)/2.2 ~ GRE.Score + TOEFL.Score +
##     LOR + CGPA + Research, data = training_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.151119 -0.018462  0.006232  0.025838  0.095640
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.5597271  0.0754849 -20.663  < 2e-16 ***
## GRE.Score    0.0010577  0.0003882   2.725 0.006795 **
## TOEFL.Score  0.0023204  0.0007001   3.314 0.001027 **
## LOR          0.0139788  0.0030858   4.530 8.42e-06 ***
## CGPA         0.0813610  0.0078633  10.347  < 2e-16 ***
## Research     0.0180335  0.0052138   3.459 0.000618 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03665 on 311 degrees of freedom
## Multiple R-squared:  0.8473, Adjusted R-squared:  0.8449
## F-statistic: 345.2 on 5 and 311 DF,  p-value: < 2.2e-16
```

Checking for the new model:




```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.0713386      1.857278   0.214
##  Alternative hypothesis: rho != 0

##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  model_final
## LM test = 1.6698, df = 1, p-value = 0.1963

##
##  Shapiro-Wilk normality test
##
## data:  model_final$residuals
## W = 0.94854, p-value = 4.437e-09
```

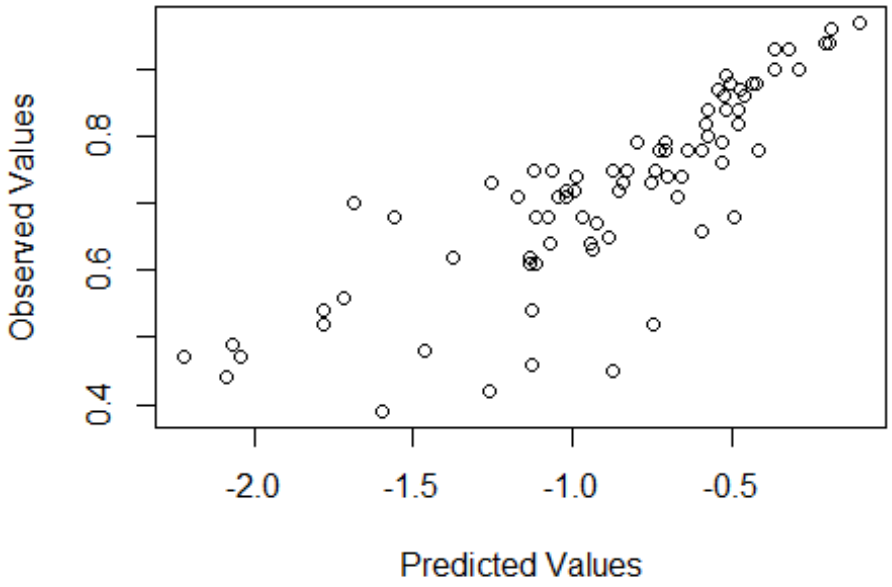
Observation: After applying the Box-Cox transformation to the data, the p-value for both the Breusch-Pagan test and the Durbin-Watson test has increased. This suggests that the transformation has helped stabilize the residual variance and has resulted in a diminished amount of compelling evidence against the null hypothesis of no autocorrelation.

Conclusion: Could use the model after transformation.

5. Double checking the model

Using model after transformation:

```
##          7         10         12         35         36         41
## -0.8729851 -0.8711698 -0.4795344 -0.1940170 -0.4255700 -1.1285685
```

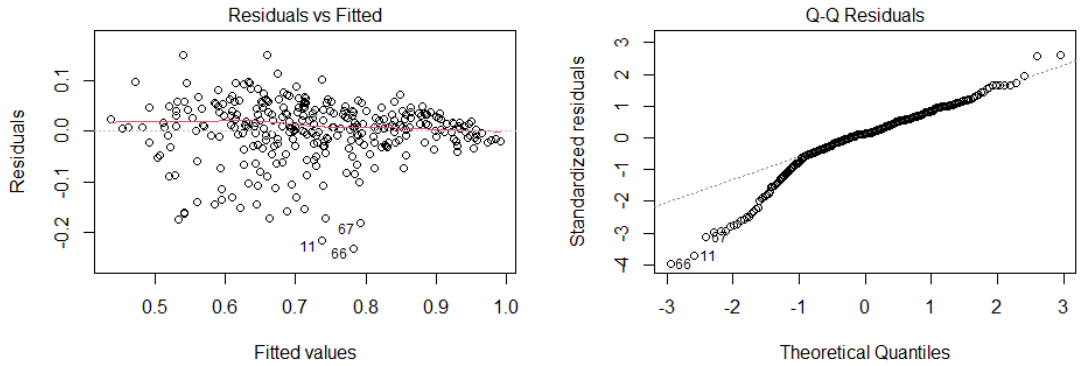


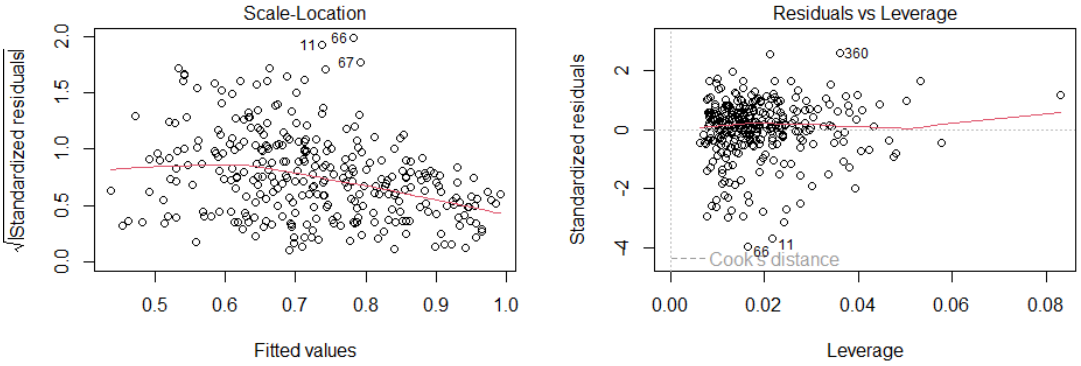

```
## [1] "Root mean square error:"

## [1] 1.649557
```

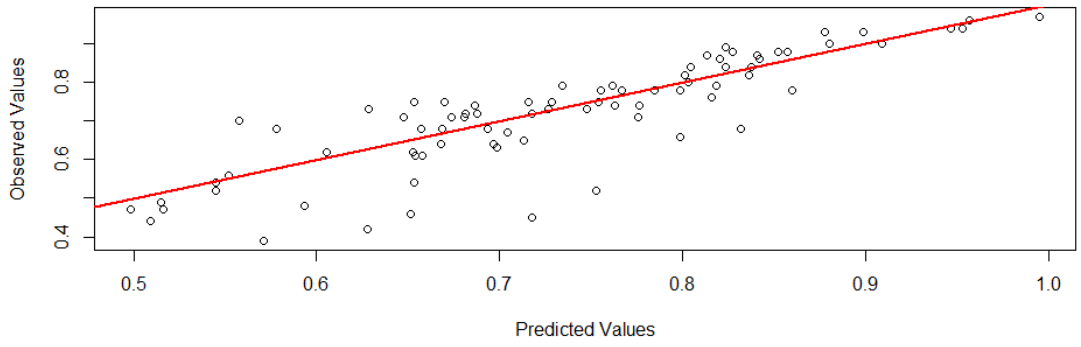
Observation: Despite the improved performance of the model after transformation compared to its pre-transformation state, it is noteworthy that the predicted values for "Chance of Admission" are consistently negative, whereas in reality, this variable is positive. This discrepancy indicates that while the transformed model may exhibit better statistical characteristics, it fails to provide meaningful and accurate predictions in the context of the problem. As a result, it is advisable to consider utilizing the model before the transformation, as it produces predictions that align with the logical expectations for the "Chance of Admission" variable.

Using model before transformation

```
##         7        10        12        35        36        41
## 0.7156696 0.7174777 0.8376253 0.9531604 0.8573183 0.6515603

## [1] "Root mean square error:"

## [1] 0.0753715
```







6. Analysis and interpretation of the model results

```
##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR +
##     CGPA + Research, data = training_data)
##
## Residuals:
##       Min         1Q    Median        3Q       Max
## -0.231838 -0.021396  0.007437  0.035194  0.150421
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2415291  0.1213790 -10.229  < 2e-16 ***
## GRE.Score    0.0016818  0.0006241   2.695  0.00743 **
## TOEFL.Score  0.0030571  0.0011258   2.716  0.00699 **
## LOR          0.0215602  0.0049619   4.345 1.89e-05 ***
## CGPA         0.1187133  0.0126441   9.389  < 2e-16 ***
## Research     0.0244236  0.0083838   2.913  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05894 on 311 degrees of freedom
## Multiple R-squared:   0.82,  Adjusted R-squared:  0.8171
## F-statistic: 283.4 on 5 and 311 DF,  p-value: < 2.2e-16

##   GRE.Score TOEFL.Score         LOR        CGPA    Research
##    4.861717    4.386228    1.793084    5.054612    1.590279
```

$Chance.of.Admit = -1.2415 + 0.0016818 \times GRE.Score + 0.0030571 \times TOEFL.Score + 0.0215602 \times LOR + 0.1187133 \times CGPA + 0.0244236 \times Research$

The linear regression model estimates the "Chance of Admission" for a student based on the following predictor variables:

- The intercept term of -1.2415 represents the estimated "Chance of Admission" when all predictor variables are zero.
- GRE Score: For every one-point increase in GRE score, the predicted chance of admission increases by approximately 0.0017.
- TOEFL Score: For every one-point increase in TOEFL score, the predicted chance of admission increases by approximately 0.0031.
- LOR (Letter of Recommendation): For every one-unit increase in the LOR rating, the predicted chance of admission increases by approximately 0.0216.
- CGPA (Cumulative Grade Point Average): For every one-unit increase in CGPA, the predicted chance of admission increases by approximately 0.1187.
- Research: If a student has research experience (Research = 1), the predicted chance of admission increases by approximately 0.0244.

The model's performance is indicated by the adjusted R-squared value of approximately 0.8171, which suggests that the predictor variables in the model explain about 81.71% of the variance in the "Chance of Admission."

Conclusion: The model indicates that variables such as GRE score, TOEFL score, letter of recommendation (LOR), CGPA, and research experience play crucial roles in influencing the chance of admission. When these factors rise, the predicted chance of admission rises as well. However, it appears that statement of purpose (SOP) and university ranking have limited impact on the chance of admission based on the model's analysis.

Appendix

1. Dataset

- Dataset for activity 02: US graduate school's admission parameters.
- Link: https://www.kaggle.com/datasets/tanmoyie/us-graduate-schools-admission-parameters

2. Source code

```
knitr::opts_chunk$set(echo = TRUE)
data<-read.csv('smoke.csv', header = FALSE)
header_names <- c("educ", "cigpric", "white", "age", "income", "cigs",
"restaurn", "lincome", "agesq", "lcigpric")
colnames(data) <- header_names
attach(data)
str(data)
dim(data)
sum(is.na(data))
summary(data)
mode <- function( x, na.rm = FALSE) {
  if(na.rm){ x = x[!is.na(x)]  }
  val <- unique(x)
  return(val[which.max(tabulate(match(x, val)))])
}

outliers <- function(x) {
# 1st and 3rd quantiles
  q75 = quantile(x, 0.75)
  q25 = quantile(x, 0.25)
  IQR = q75-q25
# lower bound
  lower_bound = q25 - 1.5 * IQR
# upper bound
  upper_bound = q75 + 1.5 * IQR
# outliers
  outlier_ind <- which(x < lower_bound | x > upper_bound)
  if (length(outlier_ind) == 0) return (0)
  return(outlier_ind)
}
par(mfrow=c(1,2))
whiteTab <- table(white)
whiteTab
barplot(whiteTab)
title(main = "Barplot")
pie(whiteTab, labels = c("0", "1"), col = c("white", "lightblue"))
title(main = "Pie Chart")
par(mfrow=c(1,2))
restaurnTab <- table(restaurn)
restaurnTab
```

```
barplot(restaurnTab)
title(main = "Barplot")
pie(whiteTab, labels = c("0", "1"), col = c("white", "lightblue"))
title(main = "Pie Chart")
educTab<-table(educ)
educTab
barplot(educTab)
cat("Mode of years of education: ", mode(educ) ,"\n")
cat("Summary of years of education: \n")
summary(educ)
cigTab<-table(cigs)
cigTab
par(mfrow=c(1,2))
hist(cigs)
boxplot(cigs)
title(main="Boxplot of cigs")
summary(cigs)
par(mfrow=c(1,2))
hist(cigpric)
boxplot(cigpric)
title(main="Boxplot of cigpric")

hist(lcigpric)
boxplot(lcigpric)
title(main="Boxplot of lcigpric")

cat("Total outlier of cigpric: ", length( boxplot.stats(cigpric)$out ), " ",
length( boxplot.stats(cigpric)$out) / length(cigpric), "\n")
cat("Summary of cigpric: \n")
summary(cigpric)
par(mfrow=c(1,2))
hist(age)
hist(agesq)

boxplot(age)
title(main="Boxplot of age")
boxplot(agesq)
title(main="Boxplot of agesq")
cat("Mode of age: ", mode(age), "\n")
cat("Summary of age \n")
summary(age)
par(mfrow=c(1,2))
barplot(table(income))
boxplot(income)
title(main="Boxplot of income")

hist(lincome)
boxplot(lincome)
title(main="Boxplot of lincome")
```

```r
cat("Mode of age: ", mode(income), "\n")
cat("Summary of income: \n")
summary(income)
t.test(x=cigpric, alternative = "two.sided", mu=60)
boxplot(cigpric~restaurn, ylab="cigpric(cents)")
t.test(cigpric[restaurn==1], cigpric[restaurn==0], alternative = "less")
t.test(x=educ, alternative = "two.sided", mu=13)

length(white[white==0])/length(white)
prop.test(table(white),p=0.10, alternative = "greater")
length(restaurn[restaurn==0])/length(restaurn)
prop.test(table(restaurn),p=0.75, alternative = "less")
#smoker and non-smoker
smoker=1:length(cigs)
for (i in 1:length(cigs))
{if (cigs[i]<1)
{smoker[i]=0}
  else
  {smoker[i]=1}
}

table(restaurn,smoker)
cat("Smoking restrictions and non-smoker: ", nrow(data[cigs == 0 & restaurn
== 1,])/nrow(data), "\n")
cat("Smoking restrictions and smoker: ", nrow(data[cigs > 0 & restaurn ==
1,])/nrow(data), "\n")


prop.test( x = c( nrow(data[cigs == 0 & restaurn == 1,]),
                  nrow(data[cigs > 0 & restaurn == 1,])), n = c(length(cigs),
length((restaurn))), alternative = "greater" )

result_vector <- unlist(apply(data[,c(1,2,3,4,6,7,8,9,10)],2,outliers))
length(unique(result_vector[result_vector != 0]))

length(unique(result_vector[result_vector != 0])) / nrow(data)
#smoker =1 if cigs>0
smoker=1:length(cigs)
for (i in 1:length(cigs))
{if (cigs[i]<1)
{smoker[i]=0}
  else
  {smoker[i]=1}
}

#ageGroup=1 if age>30
ageGroup=1:length(age)
for (i in 1:length(age))
{if (age[i]<=30)
```

```
{ageGroup[i]=0}
  else
  {ageGroup[i]=1}
}

data$young<-ageGroup
data$smoker<-smoker
pairs(cigs ~ ., data = data )
r <- cor(dplyr::select_if(data, is.numeric))
ggcorrplot::ggcorrplot(r,
           hc.order = TRUE,
           lab = TRUE)
car::vif(lm(cigs~., data = data))
model<-lm(data$cigs ~ (.-age-cigpric-income), data=data)
summary(model)
car::vif(model)
model_exceptCigpricAndWhite<-lm(data$cigs ~(.-age-lcigpric-income-white-
cigpric), data=data)
summary(model_exceptCigpricAndWhite)
anova(model, model_exceptCigpricAndWhite)
model_exceptCigpricAndWhiteAndResAndAge<-lm(data$cigs ~(.-age-agesq-cigpric-
income-white-lcigpric-restaurn), data=data)
summary(model_exceptCigpricAndWhiteAndResAndAge)
anova(model, model_exceptCigpricAndWhiteAndResAndAge)
model_best<-step(lm(data$cigs ~(.-age-cigpric-income), data=data),direction =
"both",k = 2)
summary(model_best)
lag.plot(model_best$residuals)
car::durbinWatsonTest(model_best)
lmtest::bptest(model_best)
shapiro.test(model_best$residuals)
summary(model_best)
car::vif(model_best)
anova(model_best)
par(mfrow=c(1,2))
plot(model_best)
plot(predict(model_best, newdata = data), cigs, xlab = "Predicted Values",
ylab = "Observed Values")
abline(a = 0, b = 1, col = "blue", lwd = 2)
dataGA<-read.csv('US_graduate_schools_admission_parameters_dataset.csv',
header = TRUE)
str(dataGA)
dim(dataGA)
sum(is.na(dataGA))
summary(dataGA)
par(mfrow=c(1,2))
REtable<-table(dataGA$Research)
REtable
barplot(REtable)
title(main = "Barplot")
```

```
pie(REtable,  labels = c("0", "1"), col = c("white", "lightblue"))
title(main = "Pie Chart")
par(mfrow=c(1,2))
summary(dataGA$GRE.Score)
hist(dataGA$GRE.Score,freq = FALSE, main="Historgram of GRE.score",
xlab="GRE.score")
lines(density(dataGA$GRE.Score), col = "red")
boxplot(dataGA$GRE.Score)
title(main="Boxplot of GRE.score")

par(mfrow=c(1,2))
summary(dataGA$TOEFL.Score)
hist(dataGA$TOEFL.Score,freq = FALSE, main="Historgram of TOEFL.Score",
xlab="TOEFL.Score")
lines(density(dataGA$TOEFL.Score), col = "red")
boxplot(dataGA$TOEFL.Score)
title(main="Boxplot of TOEFL.Score")
par(mfrow=c(1,2))
summary(dataGA$GPA)
hist(dataGA$CGPA,freq = FALSE, main="Historgram of CGPA", xlab="CGPA")
lines(density(dataGA$CGPA), col = "red")
boxplot(dataGA$CGPA)
title(main="Boxplot of CGPA")

par(mfrow=c(2,2))
barplot(table(dataGA$Research), xlab="Research")
boxplot(dataGA$GRE.Score~dataGA$Research, xlab="Research", ylab="GRE.Score")
boxplot(dataGA$TOEFL.Score~dataGA$Research, xlab="Research",
ylab="TOEFL.Score")
boxplot(dataGA$CGPA~dataGA$Research, xlab="Research", ylab="CGPA")

par(mfrow=c(1,2))
summary(dataGA$University.Rating)
rank<-table(dataGA$University.Rating)
barplot(rank, main="Bar chart of University Ranking")
pie(rank,main="Pie chart of University Ranking")
par(mfrow=c(1,1))
table(dataGA$Research,dataGA$University.Rating)

barplot(table(dataGA$Research,dataGA$University.Rating), beside=TRUE,legend =
c("Non-Researcher", "Researcher"),main="Relationship between University
Ranking and Reseacher")
par(mfrow=c(1,2))
summary(dataGA$SOP)
hist(dataGA$SOP, freq=FALSE, main="Histogram of SOP score", xlab="SOP")
lines(density(dataGA$SOP), col = "red")
boxplot(dataGA$SOP, main="Boxplot of SOP score")
par(mfrow=c(1,2))
```

```r
summary(dataGA$LOR)
hist(dataGA$LOR, freq=FALSE, main="Histogram of LOR score", xlab="LOR")
lines(density(dataGA$LOR), col = "red")
boxplot(dataGA$LOR, main="Boxplot of LOR score")
par(mfrow=c(1,2))
summary(dataGA$Chance.of.Admit)
hist(dataGA$Chance.of.Admit, freq=FALSE, main="Histogram of Chance.of.Admit",
xlab="Chance.of.Admit")
lines(density(dataGA$Chance.of.Admit), col = "red")
boxplot(dataGA$Chance.of.Admit~dataGA$Research, xlab="Research",
ylab="Chance.of.Admit")

result_vector <- unlist(apply(dataGA[,c(1,2,3,4,5,6,7,8,9)],2,outliers))
length(unique(result_vector[result_vector != 0]))
length(unique(result_vector[result_vector != 0])) / nrow(dataGA)
dataGA_filtered <- dataGA[-unique(result_vector[result_vector != 0]),]
dim(dataGA_filtered)

pairs(dataGA_filtered$Chance.of.Admit ~ ., data = dataGA_filtered )
r <- cor(dplyr::select_if(dataGA_filtered, is.numeric))
ggcorrplot::ggcorrplot(r, hc.order = TRUE, lab = TRUE)
car::vif( lm(dataGA_filtered$Chance.of.Admit ~ ., data = dataGA_filtered ) )

set.seed(190)  # Set a random seed for reproducibility

# Calculate the number of observations for training and validation
num_train <- round(0.8 * nrow(dataGA_filtered))  # Use 80% for training
num_val <- nrow(dataGA_filtered) - num_train

# Generate random indices for training and validation
train_indices <- sample(nrow(dataGA_filtered), num_train, replace = FALSE)
val_indices <- setdiff(1:nrow(dataGA_filtered), train_indices)

# Create training and validation datasets
training_data <- dataGA_filtered[train_indices, ]
validation_data <- dataGA_filtered[val_indices, ]

# Print the dimensions of training_data
cat("Dimensions of training_data:", dim(training_data), "\n")

# Print the dimensions of validation_data
cat("Dimensions of validation_data:", dim(validation_data), "\n")

car::vif( lm( Chance.of.Admit ~ (.-Serial.No.) , data = training_data ) )
car::vif( lm( Chance.of.Admit ~ (.-Serial.No.) , data = validation_data ) )

model<-lm( Chance.of.Admit ~ (.-Serial.No.) , data = training_data )
summary(model)
model_best<-step(model, direction='both', k=2)
```

```r
summary(model_best)
lag.plot(model_best$residuals)
car::durbinWatsonTest(model_best)
lmtest::bptest(model_best)
shapiro.test(model_best$residuals)
par(mfrow=c(1,2))
MASS::boxcox(model_best, plotit = TRUE)
MASS::boxcox(model_best, plotit = TRUE, lambda = seq( 1.5,2.5, by = 0.01 ) )

model_final<-lm((Chance.of.Admit^2.2-1)/2.2 ~ GRE.Score + TOEFL.Score + LOR +
    CGPA + Research, data = training_data)
summary(model_final)
lag.plot(model_final$residuals)
car::durbinWatsonTest(model_final)
lmtest::bgtest(model_final)
shapiro.test(model_final$residuals)
predict_Chance.of.Admit<-log(((predict(model_final, newdata =
validation_data))*2.2+1),base = 2.2)
head(predict_Chance.of.Admit)
plot(predict_Chance.of.Admit, validation_data$Chance.of.Admit,
     xlab = "Predicted Values",
     ylab = "Observed Values")

#RMSE
print("Root mean square error:")
sqrt(mean((predict_Chance.of.Admit - validation_data$Chance.of.Admit)^2))

predict_Chance.of.Admit <- predict(model_best, newdata = validation_data)
head(predict_Chance.of.Admit)
#RMSE
print("Root mean square error:")
sqrt(mean((predict_Chance.of.Admit - validation_data$Chance.of.Admit)^2))

par(mfrow=c(1,2))
plot(model_best)
par(mfrow=c(1,1))

plot(predict(model_best, newdata = validation_data),
     validation_data$Chance.of.Admit,
     xlab = "Predicted Values",
     ylab = "Observed Values")
abline(a = 0,
       b = 1,
       col = "red",
       lwd = 2)
```

```
summary(model_best)
car::vif(model_best)
```