2021

Laboratory II – Group Assignment

SCS2211

GROUP 33

Abstract

The statistical analysis of the dataset “**Medical Cost Personal Datasets**” using the R programming language is given in this report. The report comprises observations of the dataset, justification of hypotheses, and the analysis is done by making use of plots, charts, multivariate data plots, least squares regression line, residual plot and clustering in the R programming language to graphically represent data.

Content

Introduction

The goal of this task is to analyse the selected dataset in a statistical manner. The major tool for this statistical analysis is the R programming language. By using R, this analysis can be carried out easily due to the number of packages that are included in R. Also, R programming language can be also used to represent data graphically with minimum effort.

This task is focused on three main areas. Which are,

1. Analyse dataset and plot data
2. Building and justifying hypothesis
3. Plotting multivariate data and depicting the strongest relationship
4. Identify natural grouping in the dataset

This selected dataset is about insurance beneficiaries named “Medical Cost Personal Datasets” which includes the following content. [1]

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

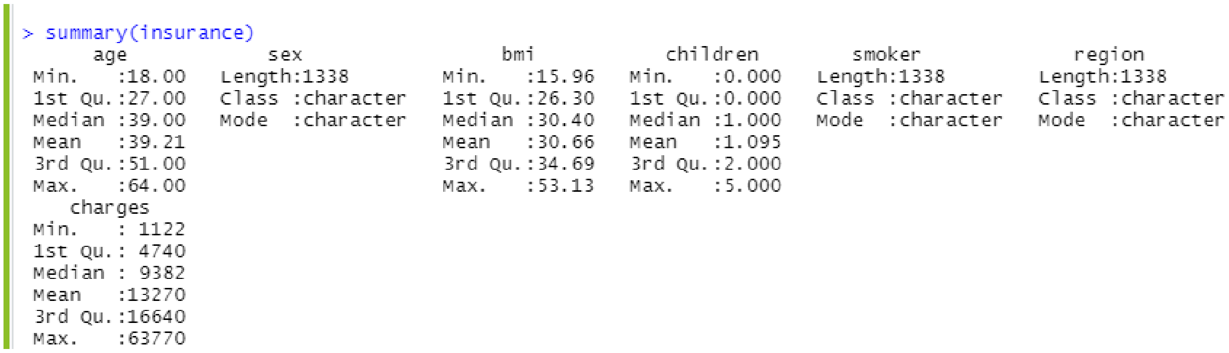
Experiment Procedure

Observations Of The Dataset

Considering the selected dataset, following facts can be concluded.

- This dataset consists of both quantitative (age of beneficiary, BMI, number of children and medical charges) and qualitative data (gender, smoking habits and region).
- Consists of 1338 records about various insurance beneficiaries.
- There are 7 different fields that are taken into consideration when collecting data about each beneficiary such as age, gender, BMI, number of children, whether a smoker or not, region in US and medical charges.

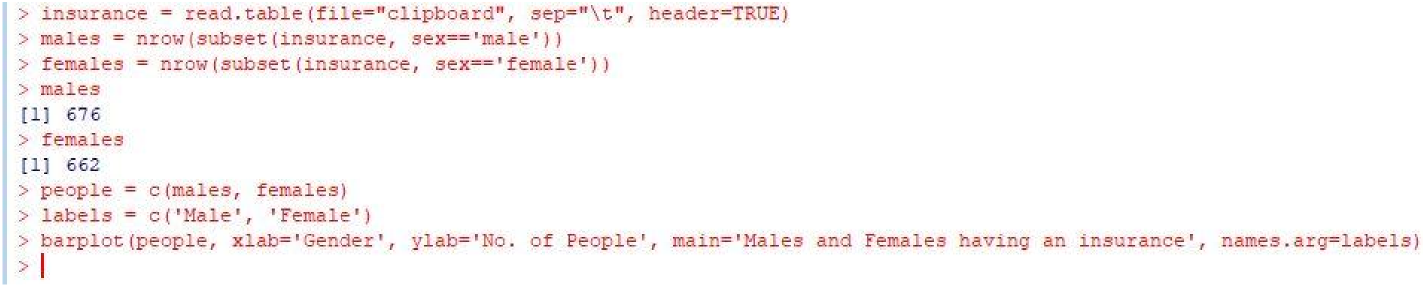
A summary of the dataset can be obtained from R as follows.



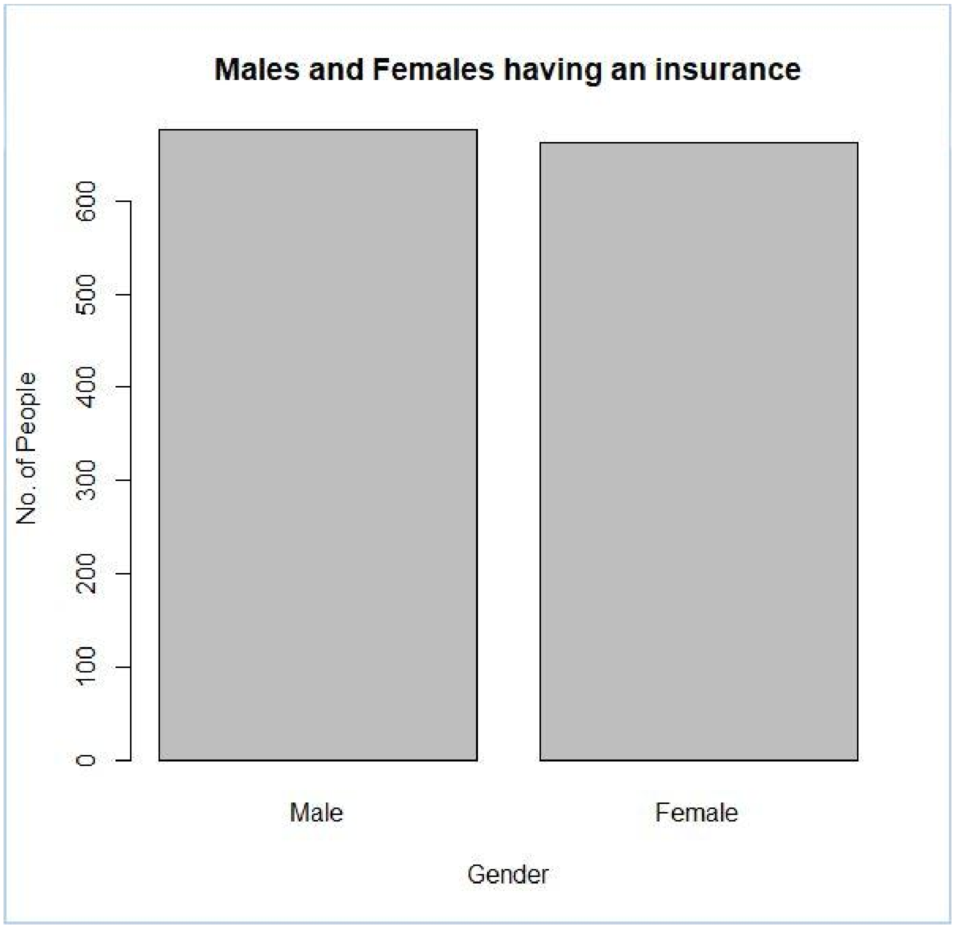
```
> summary(insurance)
      age            sex                 bmi           children        smoker             region         
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338       
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character  
 Mean   :39.21                      Mean   :30.66   Mean   :1.095                                        
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000                                        
 Max.   :64.00                      Max.   :53.13   Max.   :5.000                                        
    charges     
 Min.   : 1122  
 1st Qu.: 4740  
 Median : 9382  
 Mean   :13270  
 3rd Qu.:16640  
 Max.   :63770  
```

Types Of Plots To Analyze Data

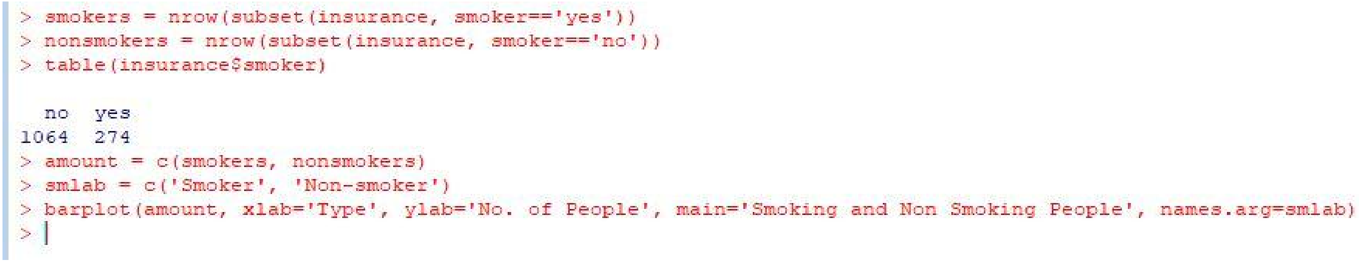
Bar graph to represent number of male beneficiaries and female beneficiaries



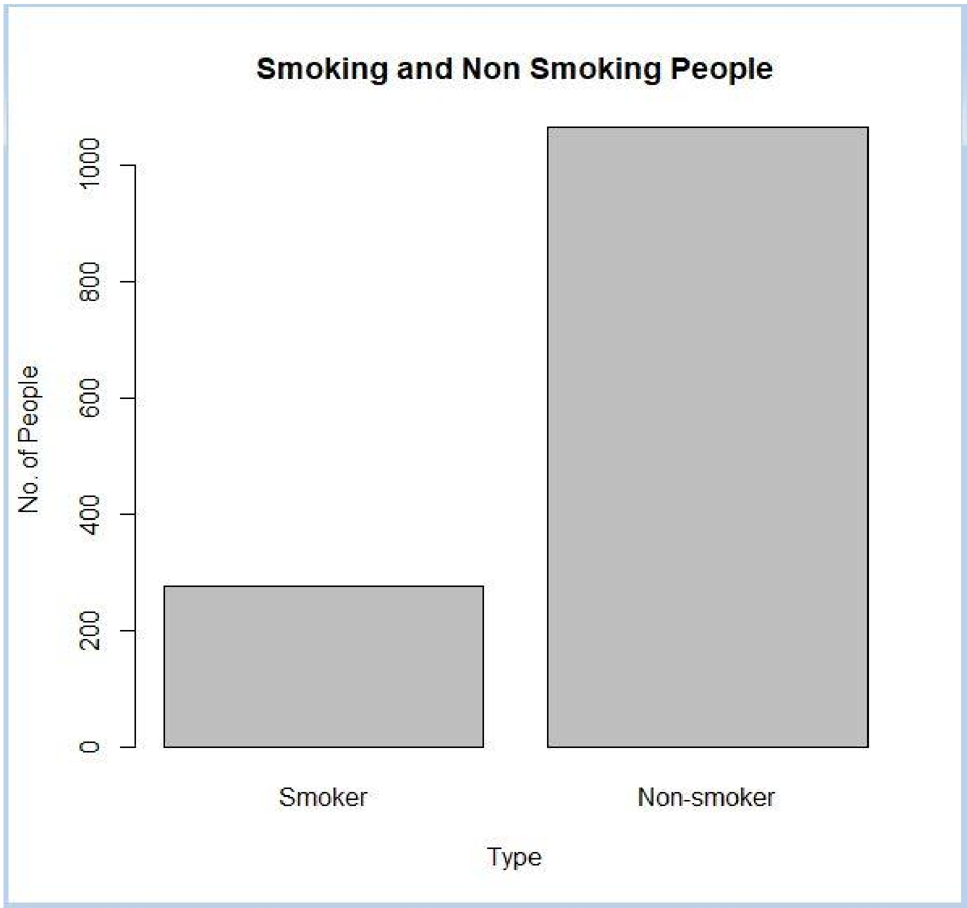
```
> insurance = read.table(file="clipboard", sep="\t", header=TRUE)
> males = nrow(subset(insurance, sex=='male'))
> females = nrow(subset(insurance, sex=='female'))
> males
[1] 676
> females
[1] 662
> people = c(males, females)
> labels = c('Male', 'Female')
> barplot(people, xlab='Gender', ylab='No. of People', main='Males and Females having an insurance', names.arg=labels)
> 
```



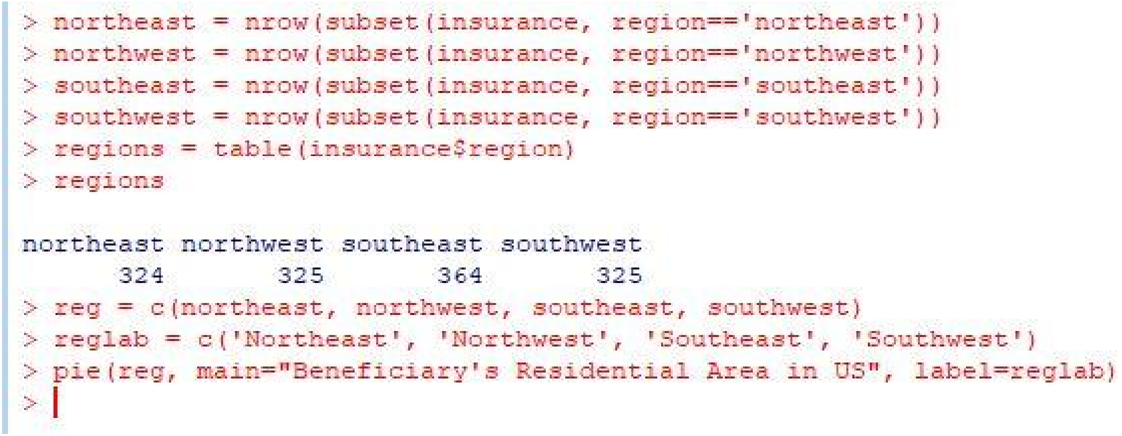
Bar graph to represent number of beneficiaries who are smoking and not smoking



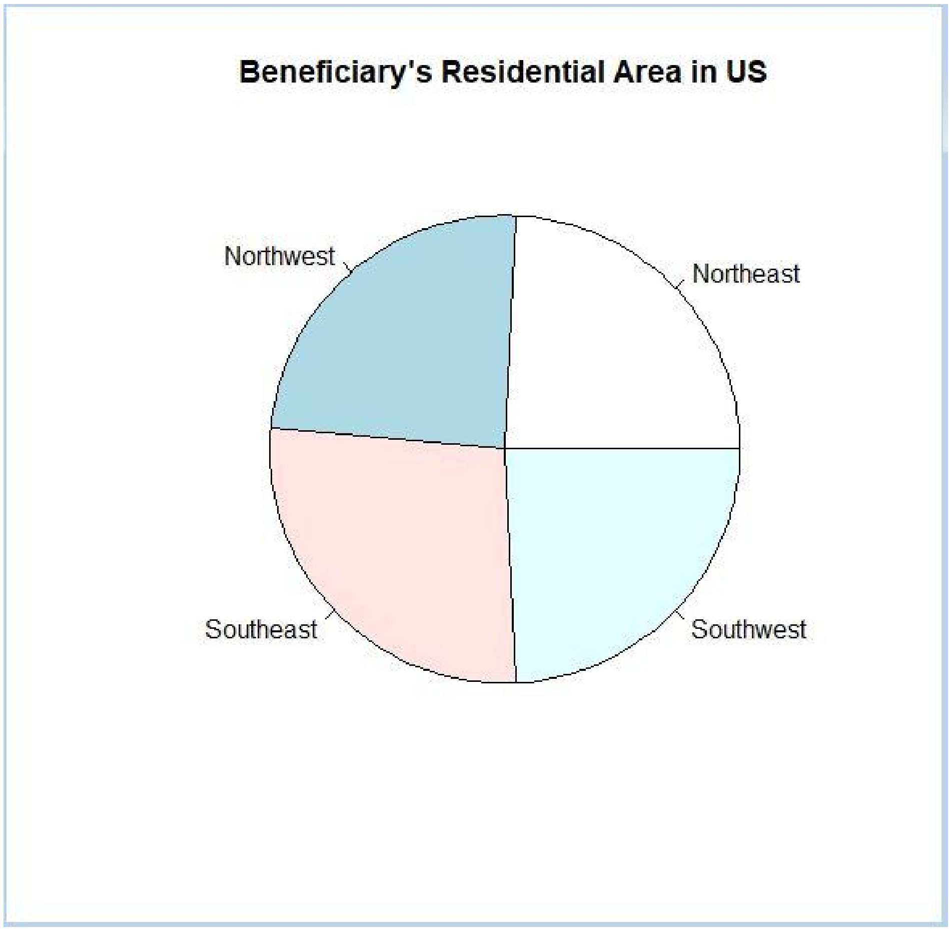
```
> smokers = nrow(subset(insurance, smoker=='yes'))
> nonsmokers = nrow(subset(insurance, smoker=='no'))
> table(insurance$smoker)

  no  yes 
1064  274 
> amount = c(smokers, nonsmokers)
> smlab = c('Smoker', 'Non-smoker')
> barplot(amount, xlab='Type', ylab='No. of People', main='Smoking and Non Smoking People', names.arg=smlab)
> |
```



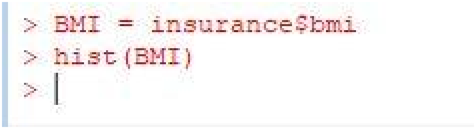
Pie chart to represent how beneficiaries scatter in four regions in the US



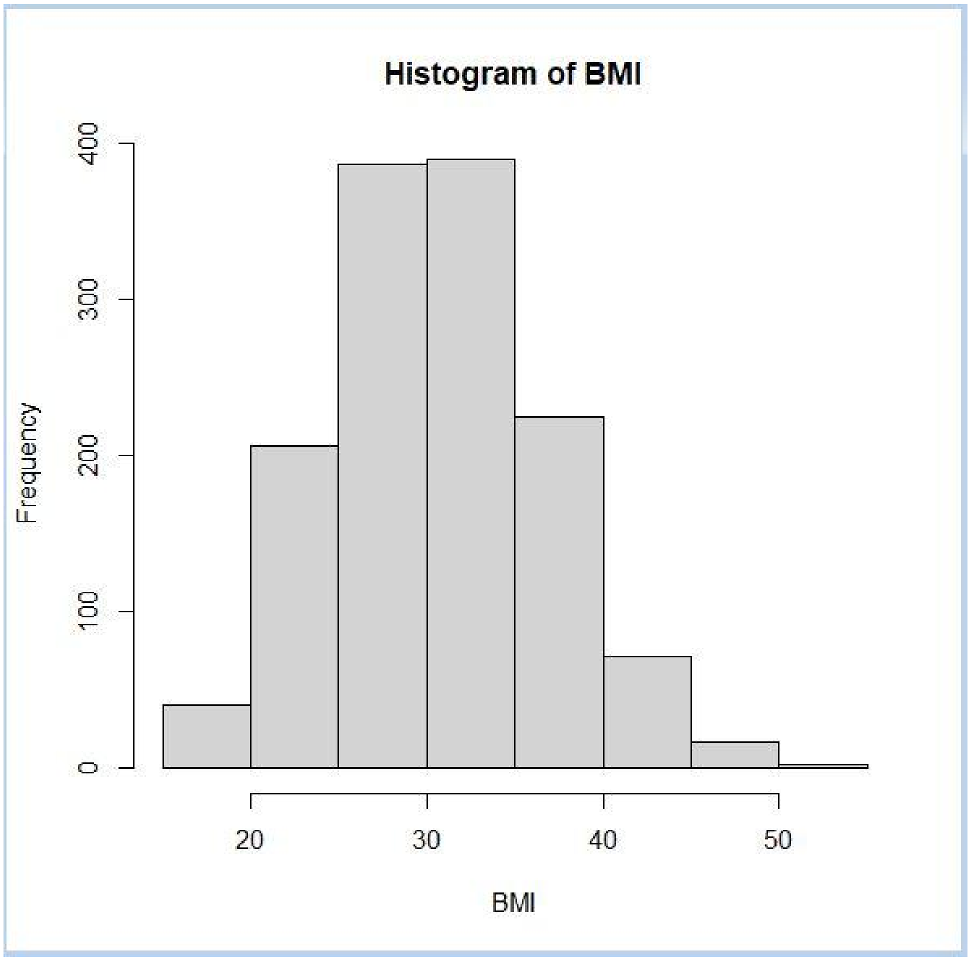
```
> northeast = nrow(subset(insurance, region=='northeast'))
> northwest = nrow(subset(insurance, region=='northwest'))
> southeast = nrow(subset(insurance, region=='southeast'))
> southwest = nrow(subset(insurance, region=='southwest'))
> regions = table(insurance$region)
> regions

northeast northwest southeast southwest 
      324       325       364       325 
> reg = c(northeast, northwest, southeast, southwest)
> reglab = c('Northeast', 'Northwest', 'Southeast', 'Southwest')
> pie(reg, main="Beneficiary's Residential Area in US", label=reglab)
> 
```



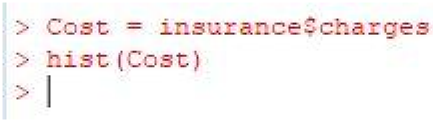
Histogram to represent data on BMI of beneficiaries



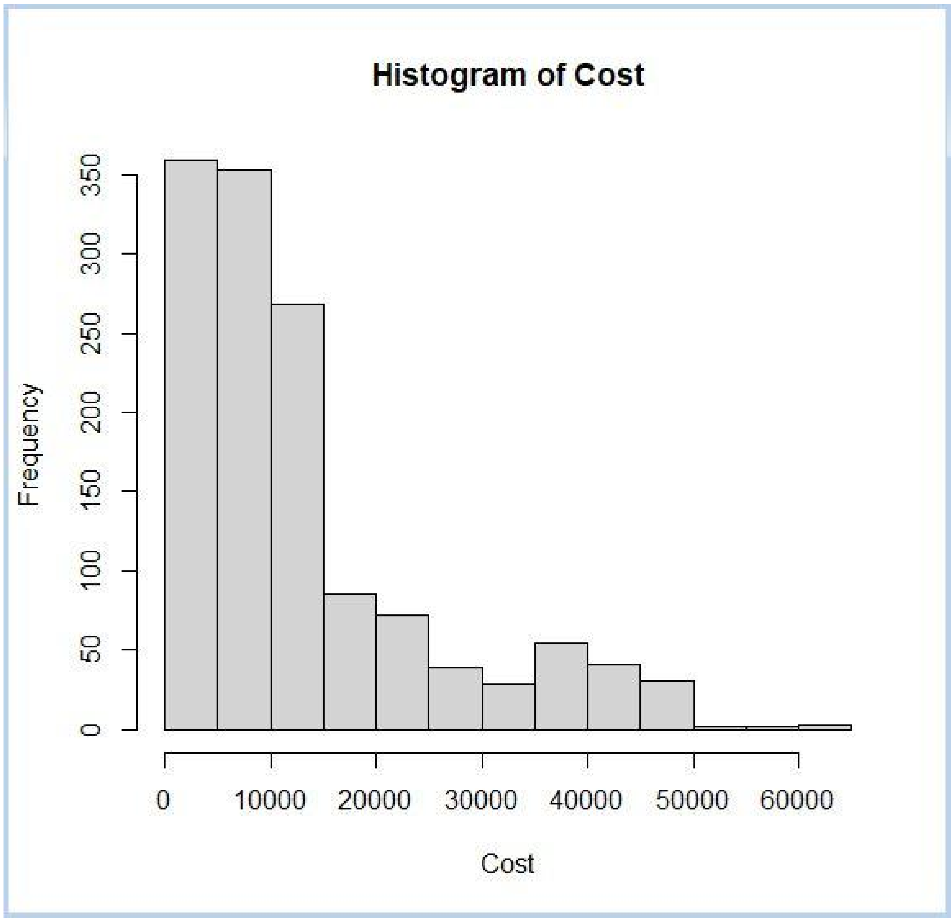
```
> BMI = insurance$bmi
> hist(BMI)
> |
```



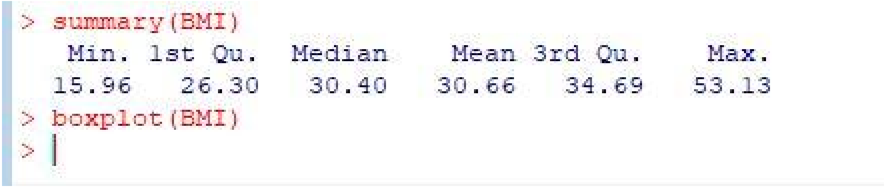
Histogram to represent data on individual medical charges of beneficiaries



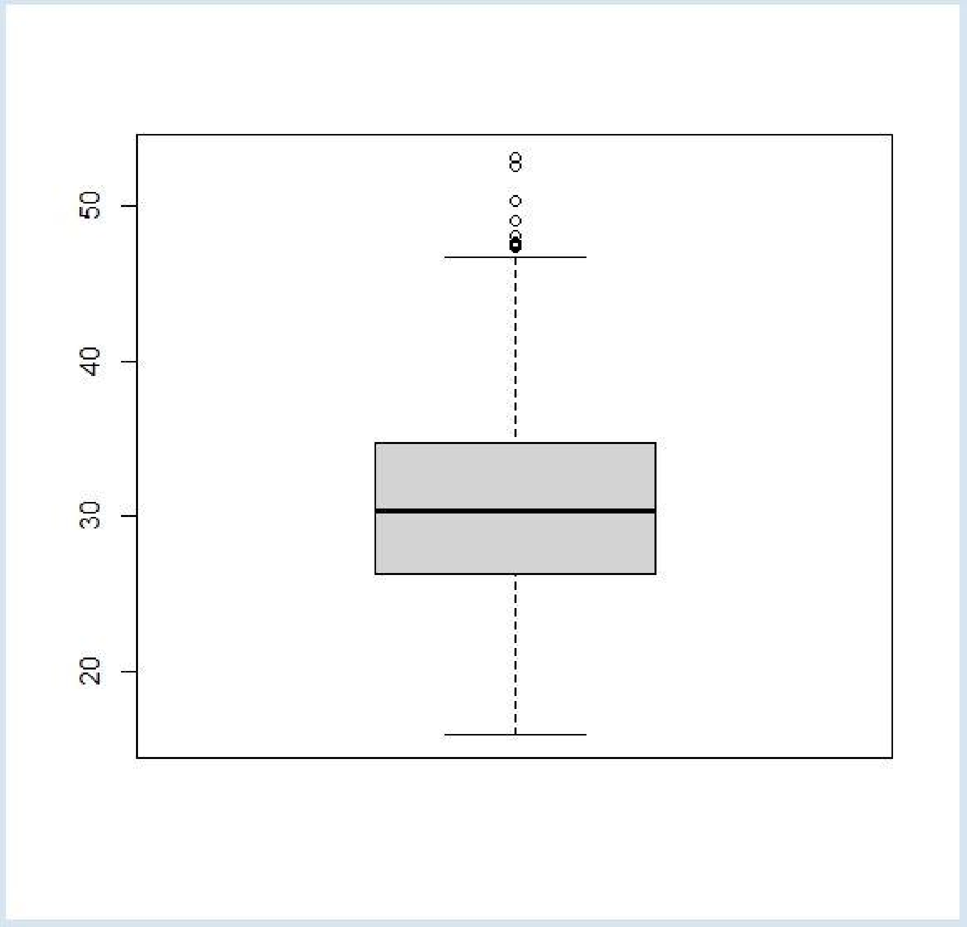
```
> Cost = insurance$charges
> hist(Cost)
> |
```



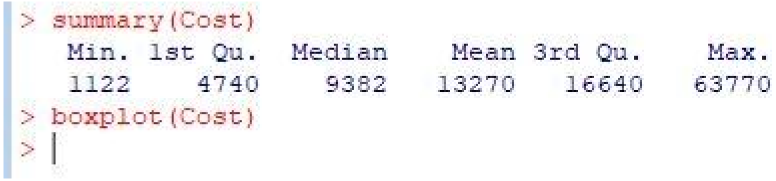
Boxplot to represent the summary of BMI of beneficiaries



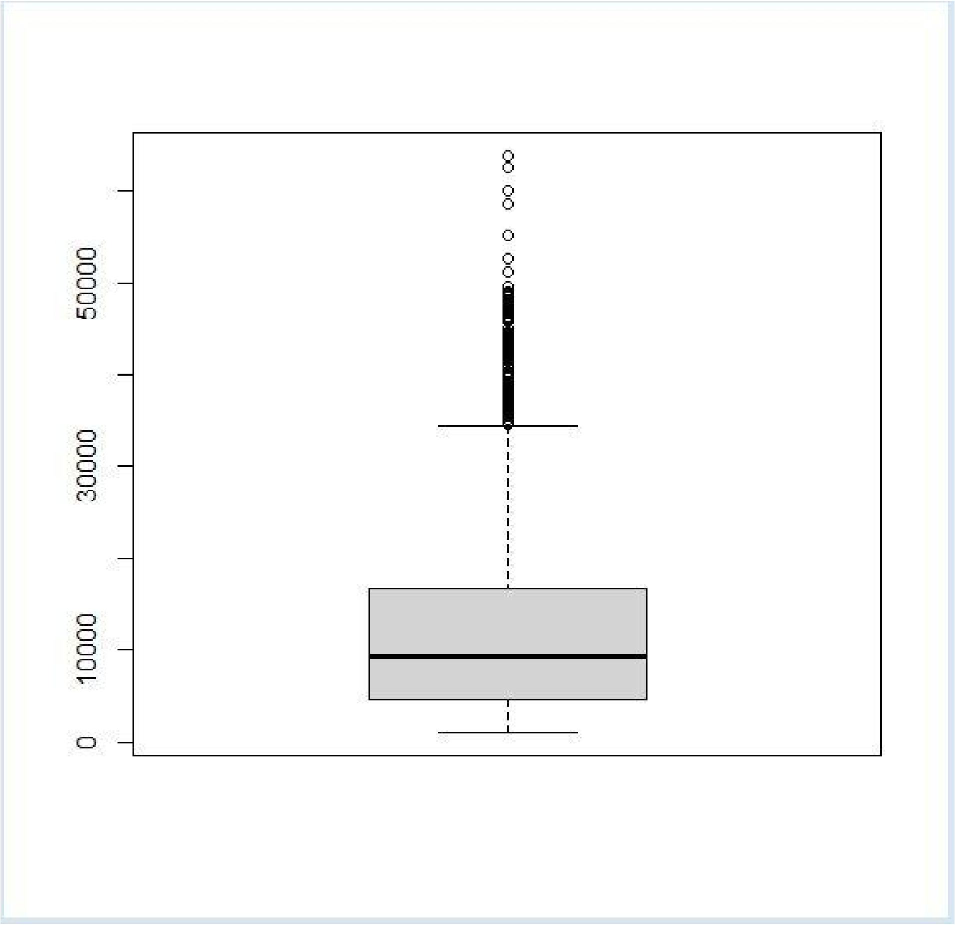
```
> summary(BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.96   26.30   30.40   30.66   34.69   53.13 
> boxplot(BMI)
> |
```



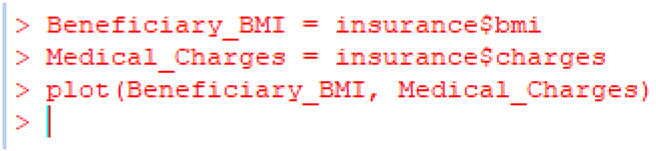
Boxplot to represent the summary of individual medical charges of beneficiaries



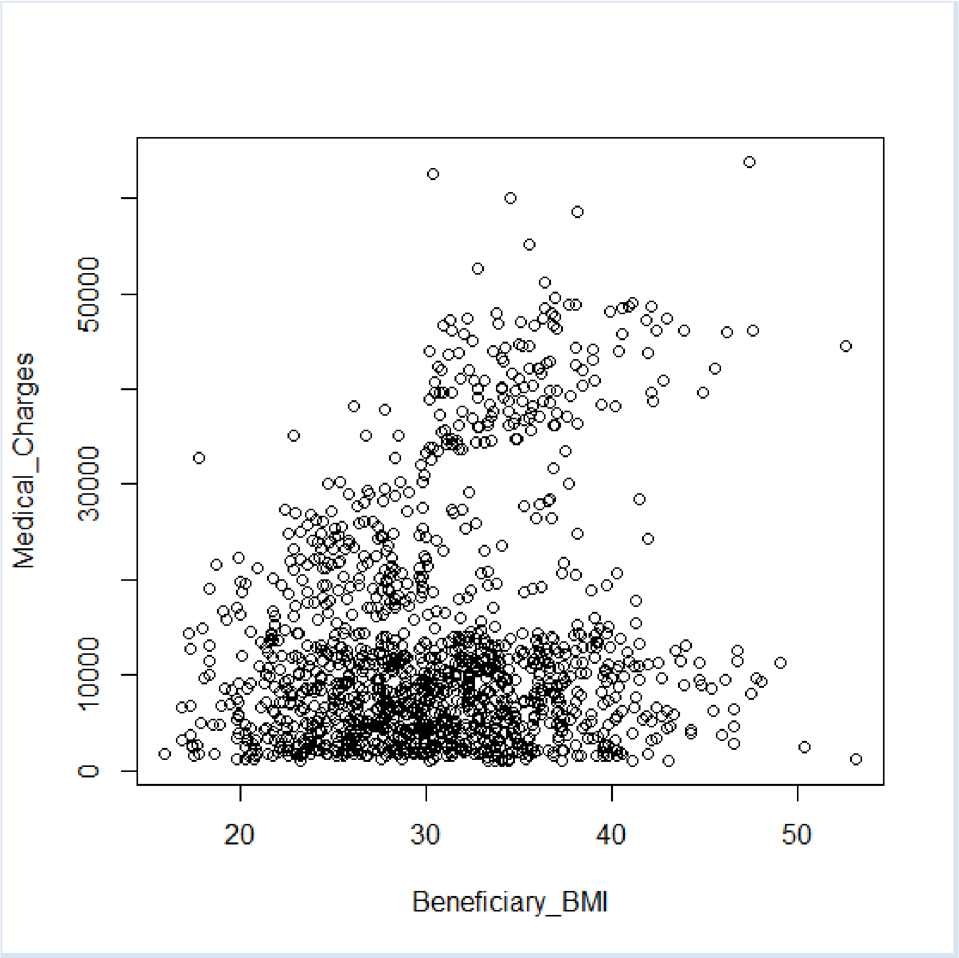
```
> summary(Cost)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1122    4740    9382   13270   16640   63770 
> boxplot(Cost)
> |
```



Scatter plot to represent the relationship between BMI and individual medical charges of beneficiaries



```
> Beneficiary_BMI = insurance$bmi
> Medical_Charges = insurance$charges
> plot(Beneficiary_BMI, Medical_Charges)
> 
```



Hypothesis

Two hypotheses are developed as follows from the dataset.

1. **Null hypothesis** - The difference between the mean of BMI of males (mu1) and the mean of BMI of females (mu2) is greater than or equal to 0, that is mu1 - mu2 >= 0.
   Alternative hypothesis - The difference between the mean of BMI of males (mu1) and the mean of BMI of females (mu2) is less than 0, that is mu1 - mu2 < 0.

2. **Null hypothesis** - The difference between the medical costs of males (mu3) and the medical costs of females (mu4) is equal to 0, that is mu3 - mu4 = 0.
   Alternative hypothesis - The difference between the medical costs of males (mu3) and the medical costs of females (mu4) is not equal to 0, that is mu3 - mu4 != 0.

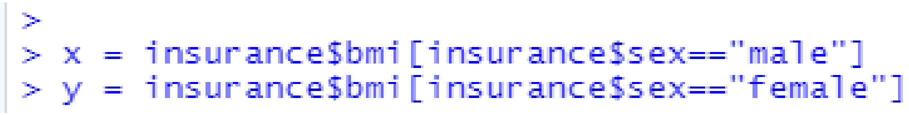
The types of tests that can be used are,

1. Z test - If the population standard deviation is known, then this test can be used.
2. T test - If the population standard deviation is unknown but the sample standard deviation is known, then this test can be used.

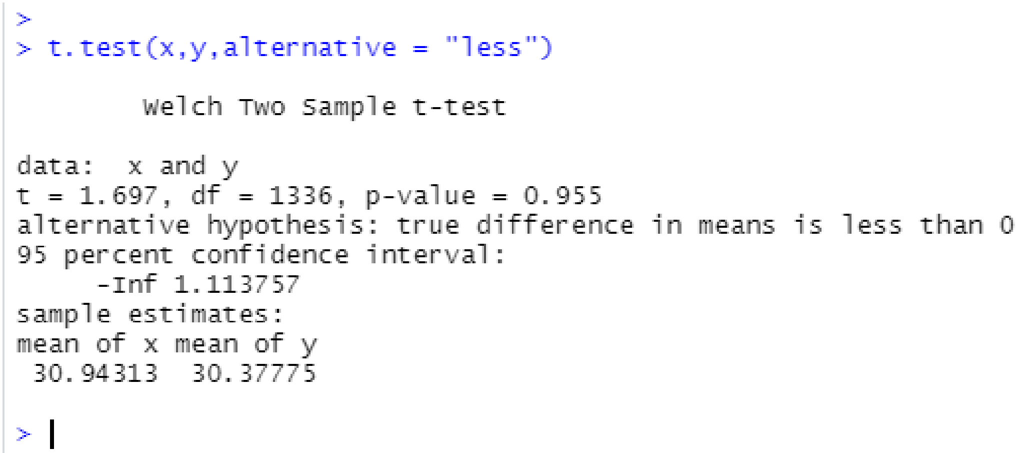
The T test is chosen to carry out the justification of both hypotheses.

Hypothesis Justification

1. **Null hypothesis** - The difference between the mean of BMI of males (mu1) and the mean of BMI of females (mu2) is greater than or equal to 0, that is mu1 - mu2 >= 0.
   Alternative hypothesis - The difference between the mean of BMI of males (mu1) and the mean of BMI of females (mu2) is less than 0, that is mu1 - mu2 < 0.



```
>
> x = insurance$bmi[insurance$sex=="male"]
> y = insurance$bmi[insurance$sex=="female"]
```

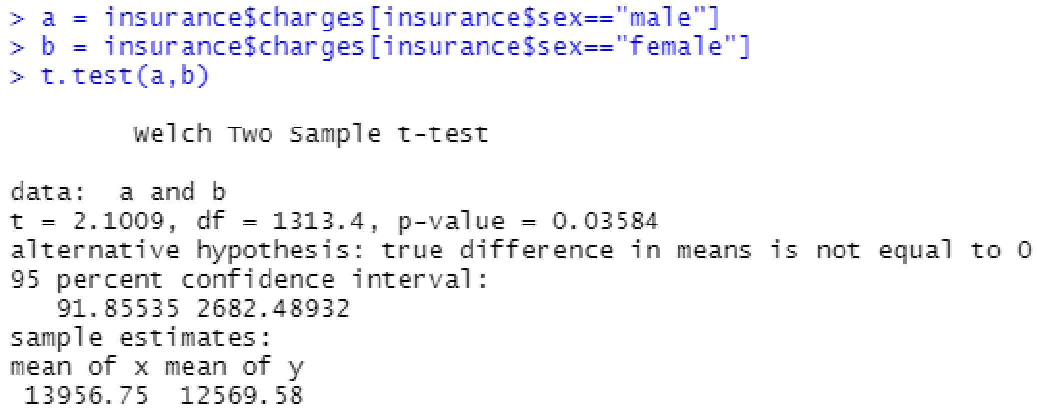


```
>
> t.test(x,y,alternative = "less")

        Welch Two Sample t-test

data:  x and y
t = 1.697, df = 1336, p-value = 0.955
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.113757
sample estimates:
mean of x mean of y
 30.94313  30.37775

> |
```

The significance level is 5% and the p-value is 0.955. The hypothesis can be justified by comparing the p-value with the significance level. Since the p-value is greater than the 0.05 (5%) significance level, it can be said that at 5% significance level, we do not reject the null hypothesis that mu1 - mu2 >= 0. Therefore, the null hypothesis which states that the difference between the mean of BMI of males (mu1) and the mean of BMI of females (mu2) is greater than or equal to 0 is accepted.

Moreover, the hypothesis can also be justified by comparing the test statistic with the critical value (alpha) at 5% significance level. The test statistic is 1.697 and the critical value is 1.113757. This is a lower tail test. Therefore, if the test statistic is less than the critical value, the null hypothesis is rejected. In this scenario, the test statistic 1.697 is not less than the critical value 1.113757. Therefore, at 5% significance level, we do not reject the null hypothesis that mu1 - mu2 >= 0.

2. **Null hypothesis** - The difference between the medical costs of males (mu3) and the medical costs of females (mu4) is equal to 0, that is mu3 - mu4 = 0.
   Alternative hypothesis - The difference between the medical costs of males (mu3) and the medical costs of females (mu4) is not equal to 0, that is mu3 - mu4 != 0.



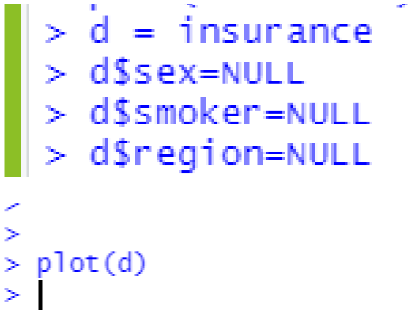
```
> a = insurance$charges[insurance$sex=="male"]
> b = insurance$charges[insurance$sex=="female"]
> t.test(a,b)

        Welch Two Sample t-test

data:  a and b
t = 2.1009, df = 1313.4, p-value = 0.03584
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   91.85535 2682.48932
sample estimates:
mean of x mean of y 
 13956.75  12569.58 
```

The significance level is 5% and the p-value is 0.03584. Since the p-value is less than the 0.05 (5%) significance level, it can be said that at 5% significance level, we reject the null hypothesis that mu3 - mu4 = 0. Therefore, the null hypothesis which states that the difference between the medical costs of males (mu3) and the medical costs of females (mu4) is equal to 0 is rejected.

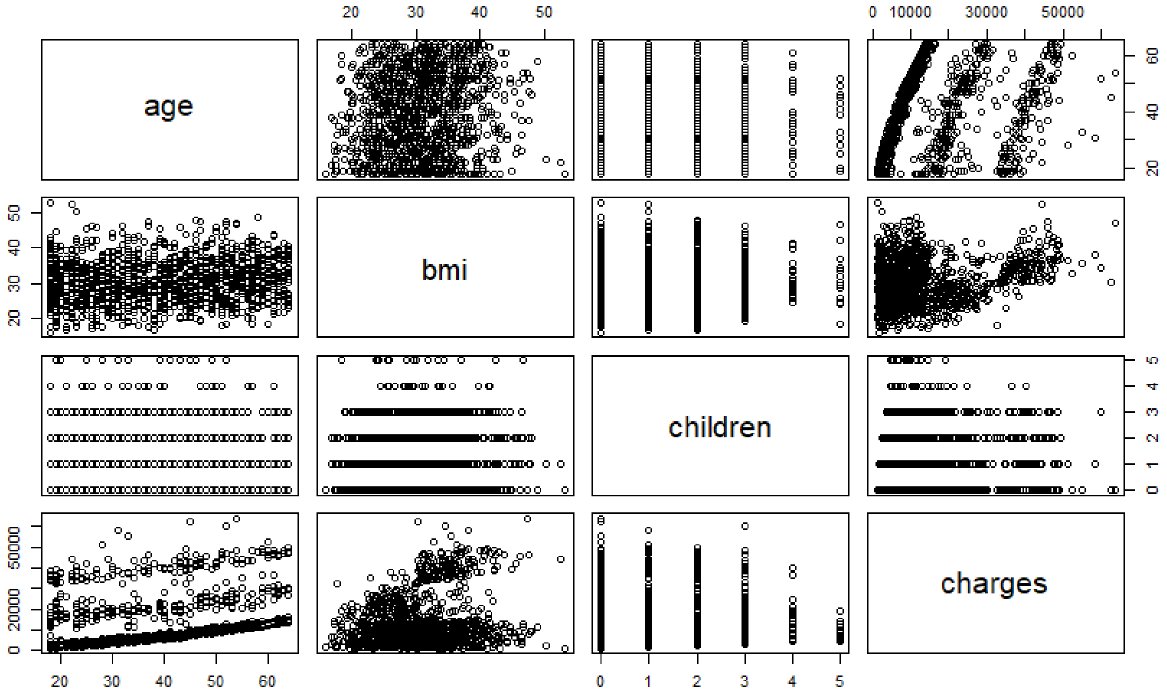
Moreover, the hypothesis can also be justified by comparing the test statistic with the critical values (alpha) at 5% significance level. The test statistic is 2.1009 and the critical values are 91.85535 and 2682.48932. This is a two tailed test. Therefore, if the test statistic is in the range of the critical values, the null hypothesis is not rejected. In this scenario, the test statistic 2.1009 is less than the critical value 91.85535 and it does not lie between the critical values. Therefore, at 5% significance level, we reject the null hypothesis that mu3 - mu4 = 0.

Plotting of Multivariate Data

In multivariate data, data arise from more than one response variable as shown in the following plot. Multivariate data analysis is used to analyze these kinds of datasets.



```
> d = insurance
> d$sex=NULL
> d$smoker=NULL
> d$region=NULL

>
> plot(d)
> |
```

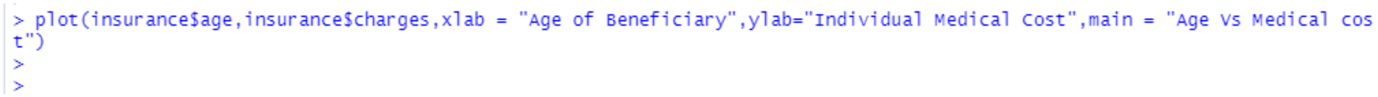


This multivariate plot is drawn between the quantitative data in the dataset to get a meaningful result. When examining the multivariate plot, it can be seen that the relationships age vs. charges and bmi vs. charges have a strong relationship while other relationships don’t.

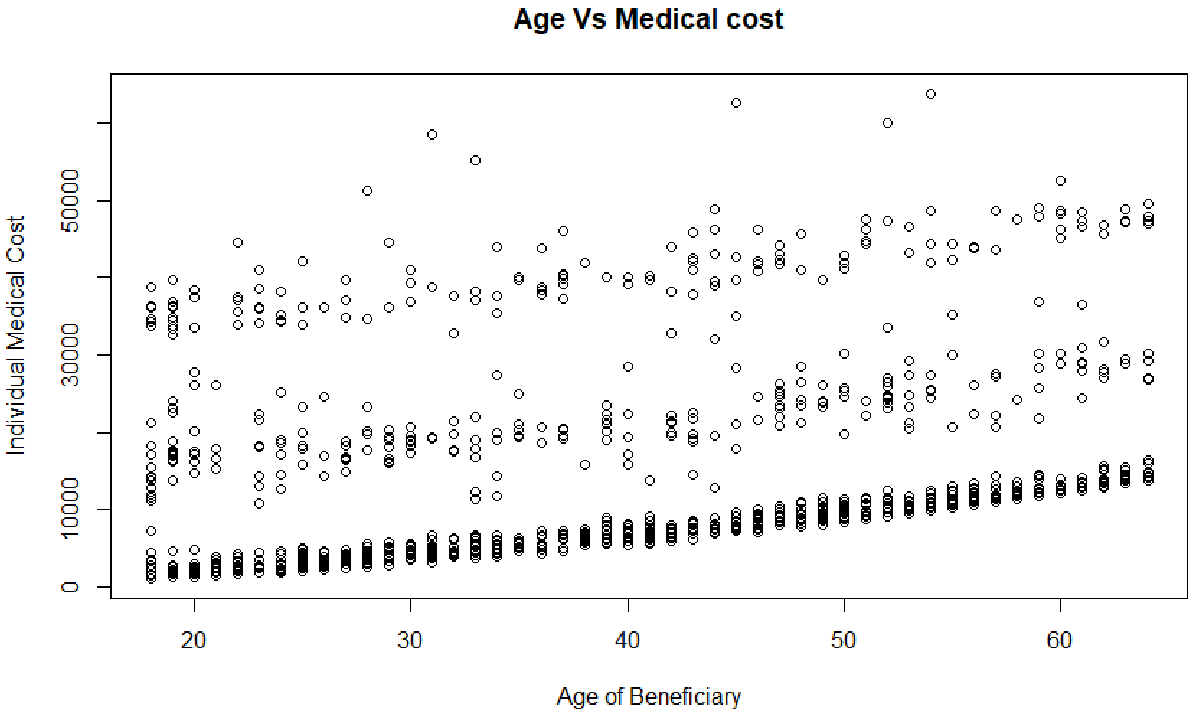
Description of the two variables that shows the strongest relationship

Considering age vs. charges and bmi vs. charges relationships, the strongest relationship we chose is the age vs. charges relationship since the data points lie in a compact cluster.

We chose the variable “age” as the explanatory variable and the variable “charges” as the response variable. When we observe the relationship, it is the case that when the age of a beneficiary increases, the individual medical charges also increase. The relationship between age and medical charges shows a positive association.

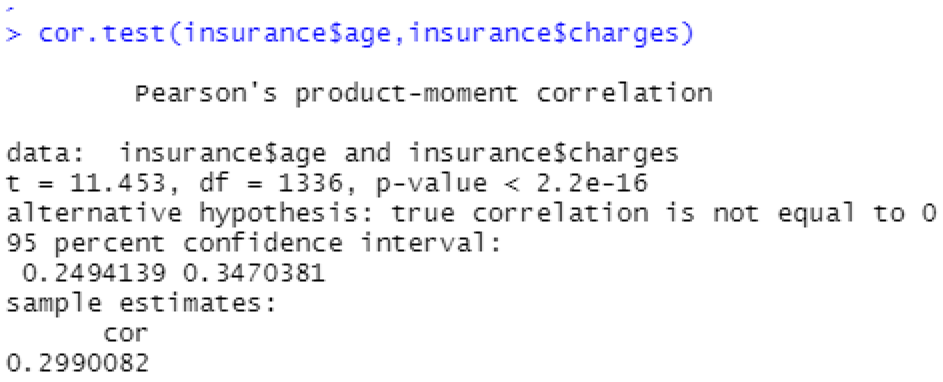


```
> plot(insurance$age,insurance$charges,xlab = "Age of Beneficiary",ylab="Individual Medical Cost",main = "Age Vs Medical cos
t")
>
>
```



Finding the correlation of selected variables

- Explanatory variable - Age of the beneficiary
- Response variable - Individual medical cost



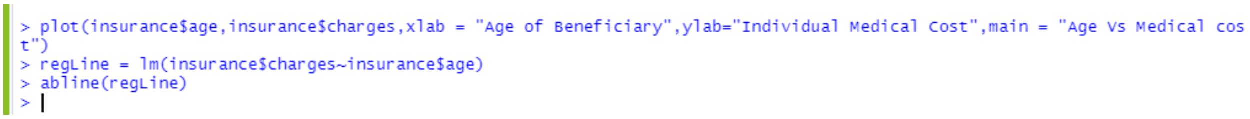
```
> cor.test(insurance$age,insurance$charges)

	Pearson's product-moment correlation

data:  insurance$age and insurance$charges
t = 11.453, df = 1336, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2494139 0.3470381
sample estimates:
      cor 
0.2990082
```

The correlation is used to measure the direction and the strength of the linear relationship between two quantitative variables. The more correlation is closer to +1 or -1, the stronger the linear relationship is. The correlation of the selected two variables “age” and “charges” is 0.2990082 which depicts that this linear relationship is moderately strong and it has a positive association since the value is positive.

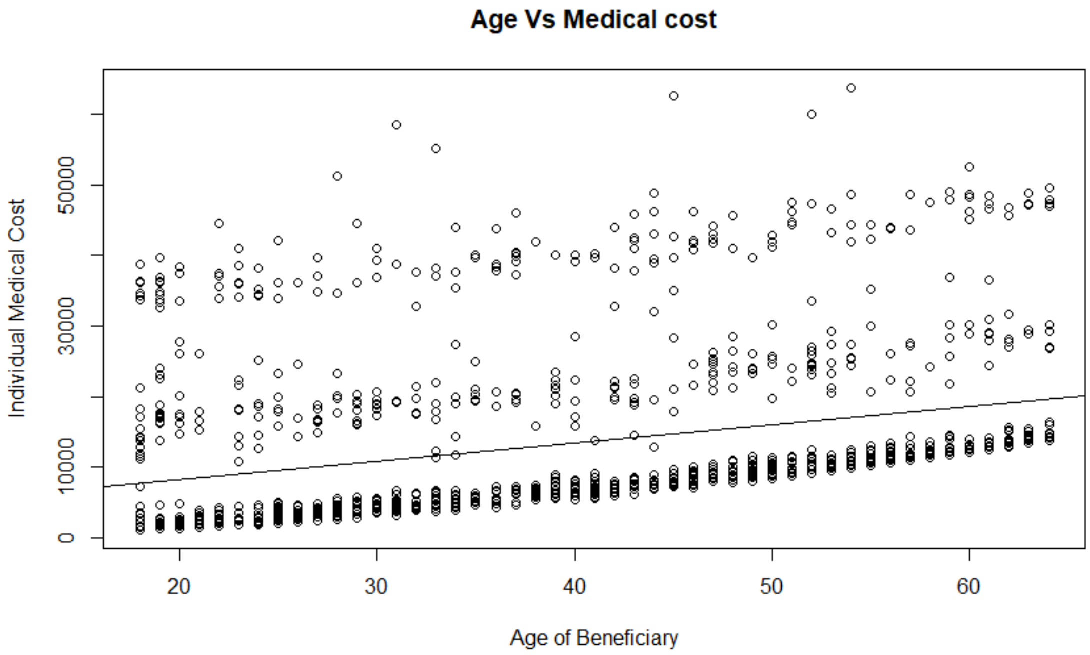
Least squares regression line

The least squares regression line is the straight line which goes through data points having the sum of squares of the vertical gap between the line and the data points as small as possible.

The least squares regression line of the linear relationship between “age” and “charges” is shown below.



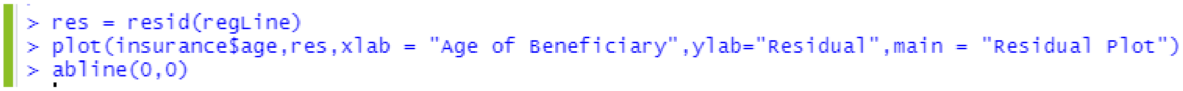
```
> plot(insurance$age,insurance$charges,xlab = "Age of Beneficiary",ylab="Individual Medical Cost",main = "Age Vs Medical cos
t")
> regLine = lm(insurance$charges~insurance$age)
> abline(regLine)
> |
```



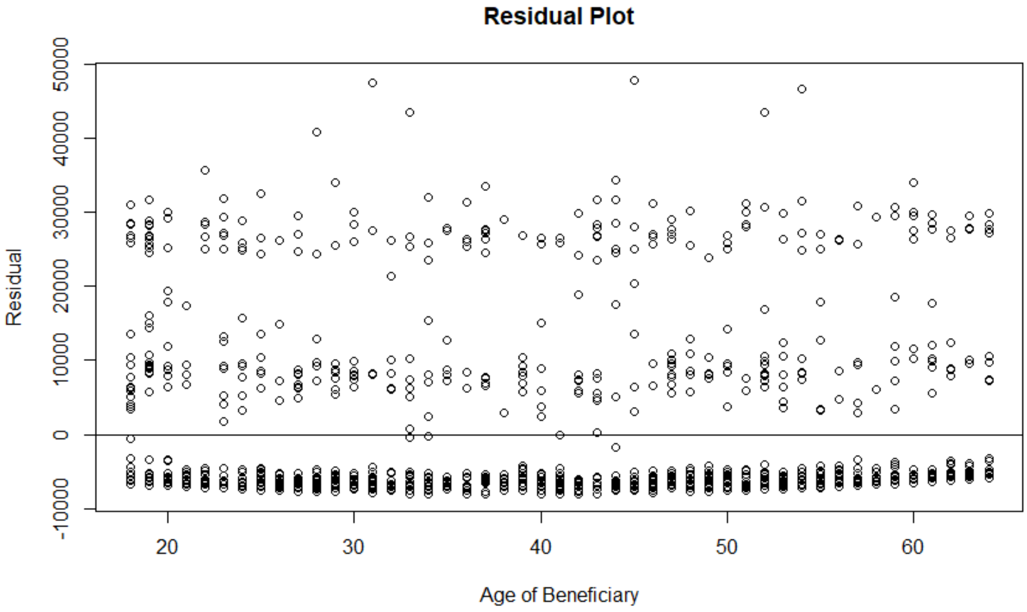
The residual plot

The residual is the difference between the observed value of the response variable and the predicted value taken by the regression line. It is basically the error in predicting the response variable using the regression line.

The residual plot is plotted taking residuals against the explanatory variable “age” as follows.



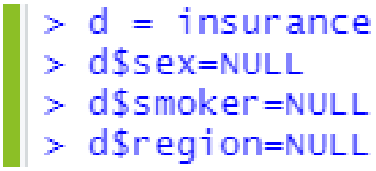
```
> res = resid(regLine)
> plot(insurance$age,res,xlab = "Age of Beneficiary",ylab="Residual",main = "Residual Plot")
> abline(0,0)
```



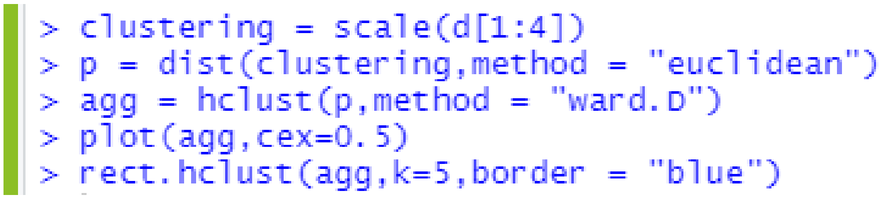
Identifying the Natural Grouping in the Dataset

Cluster analysis is used to identify natural groupings in multivariate data by grouping homogeneous data. There are mainly two types of methods for this namely hierarchical clustering and non hierarchical clustering.

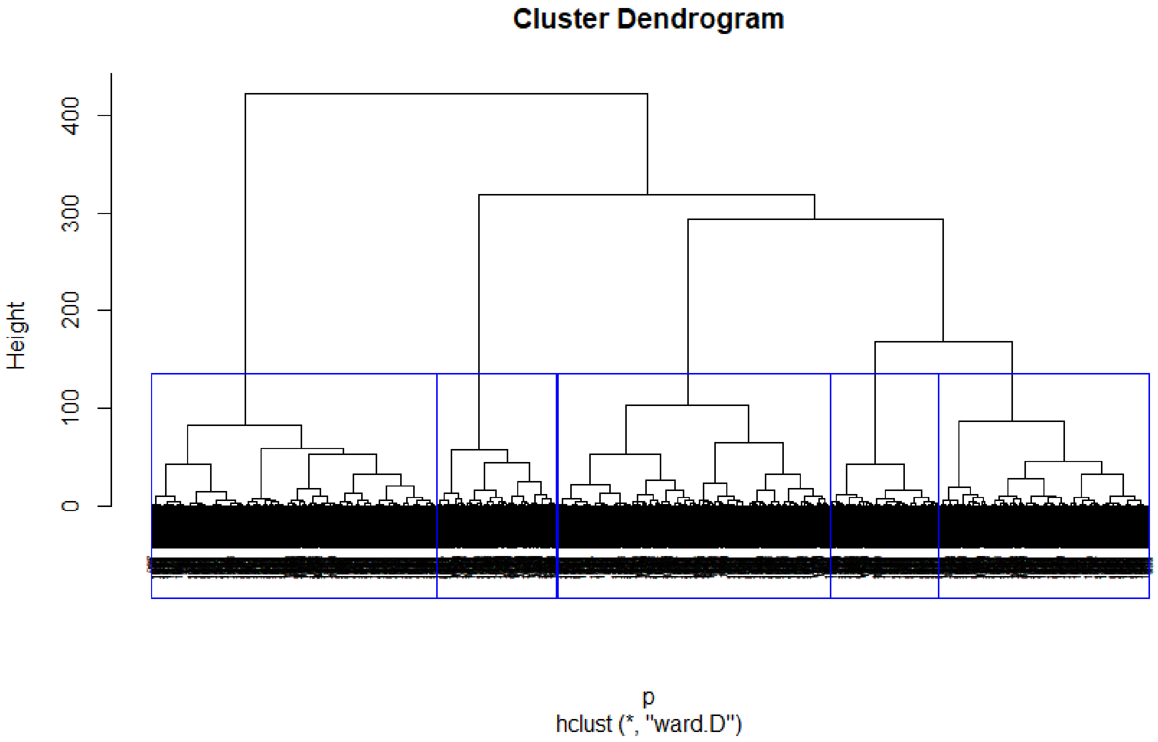
Using a hierarchical clustering method is suitable for this dataset since we are going to cluster it into 5 clusters. Agglomerative clustering which is a hierarchical clustering method is used in order to identify the natural grouping in the dataset.



```
> d = insurance
> d$sex=NULL
> d$smoker=NULL
> d$region=NULL
```



```
> clustering = scale(d[1:4])
> p = dist(clustering,method = "euclidean")
> agg = hclust(p,method = "ward.D")
> plot(agg,cex=0.5)
> rect.hclust(agg,k=5,border = "blue")
```



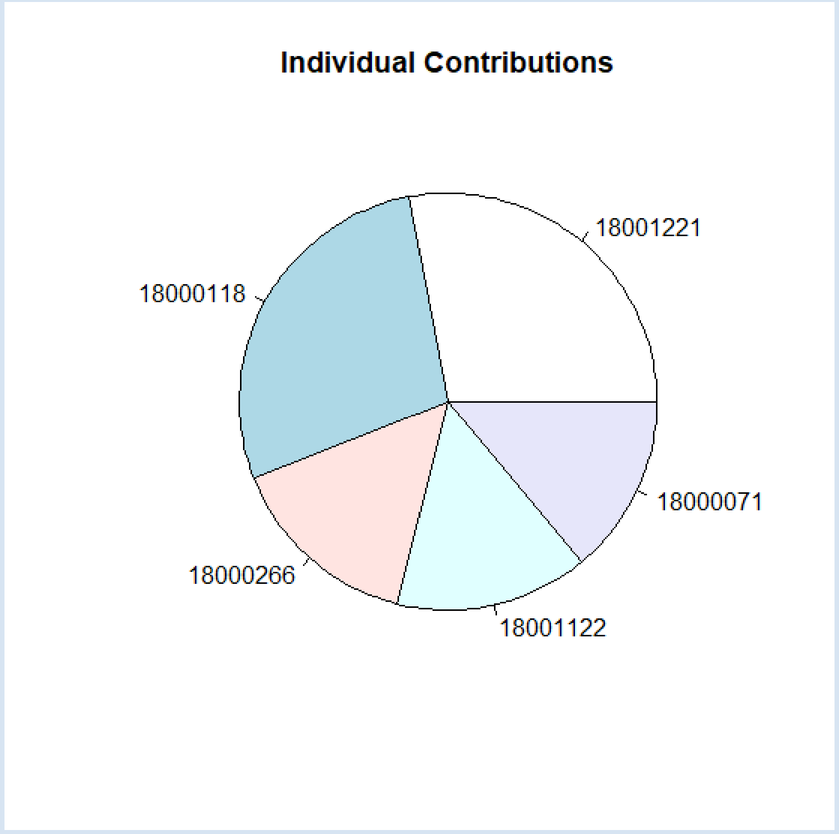
There are 5 clusters. Each of these 5 clusters contain homogenous data about beneficiaries in which age, BMI of the beneficiary, number of children that are covered by the insurance and individual medical costs are similar to one another within the same cluster and different from data in other clusters.

References

[1]"Medical Cost Personal Datasets", *Kaggle.com*, 2021. [Online]. Available: https://www.kaggle.com/mirichoi0218/insurance. [Accessed: 24- Mar- 2021].

Individual Contributions

1. M.U.I. Perera - 18001221
2. A.D.M.R. Amarasekara - 18000118
3. B.H.R. Cooray - 18000266
4. P.D. Niriella - 18001122
5. L.D. Abeywickrama - 18000071


Individual Contributions