Final-Report.r

muj_m

2022-04-06

```
#####Initial report####

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(ggplot2)
library(relaimpo)
```

```
## Warning: package 'relaimpo' was built under R version 4.1.2
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: boot
```

```
## Loading required package: survey
```

```
## Warning: package 'survey' was built under R version 4.1.2
```

```
## Loading required package: grid
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:boot':
##
##     aml
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
## Loading required package: mitools
```

```
## Warning: package 'mitools' was built under R version 4.1.2
```

```
## This is the global version of package relaimpo.
```

```
## If you are a non-US user, a version with the interesting additional metric pmvd is availab
le
```

```
## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.2
```

```
## corrplot 0.92 loaded
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v purrr   0.3.4
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x MASS::select()  masks dplyr::select()
## x tidyr::unpack() masks Matrix::unpack()
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

```
## Loading required package: lattice
```

```
## 
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
## 
##     melanoma
```

```
## 
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
## 
##     lift
```

```
## The following object is masked from 'package:survival':
## 
##     cluster
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2
```

```
## 
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:boot':
## 
##     logit
```

```
## The following objects are masked from 'package:ggplot2':
## 
##     %+%, alpha
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.1.3
```

```
setwd("C:\\Users\\muj_m\\Desktop\\aly_6015\\Major project\\Insurence_Cost")
insurance <- read.csv("insurance.csv")
class(insurance$region)
```

```
## [1] "character"
```

```
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

```
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
dim(insurance)
```

```
## [1] 1338    7
```

```
summary(insurance$charges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```
summary(insurance$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   27.00   39.00   39.21   51.00   64.00
```

```
summary(insurance$sex)
```

```
##    Length     Class      Mode
##      1338 character character
```

```
summary(insurance$smoker)
```

```
##    Length     Class      Mode
##      1338 character character
```

```
summary(insurance$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.96   26.30   30.40   30.66   34.69   53.13
```

```
summary(insurance$children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.095   2.000   5.000
```

```
class(insurance$age)
```

```
## [1] "integer"
```

```
typeof(insurance$age)
```

```
## [1] "integer"
```

```
#checking if we have any empty values
sum(is.na(insurance))
```

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

```
summary(insurance)
```

```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000  
##  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  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770  
```

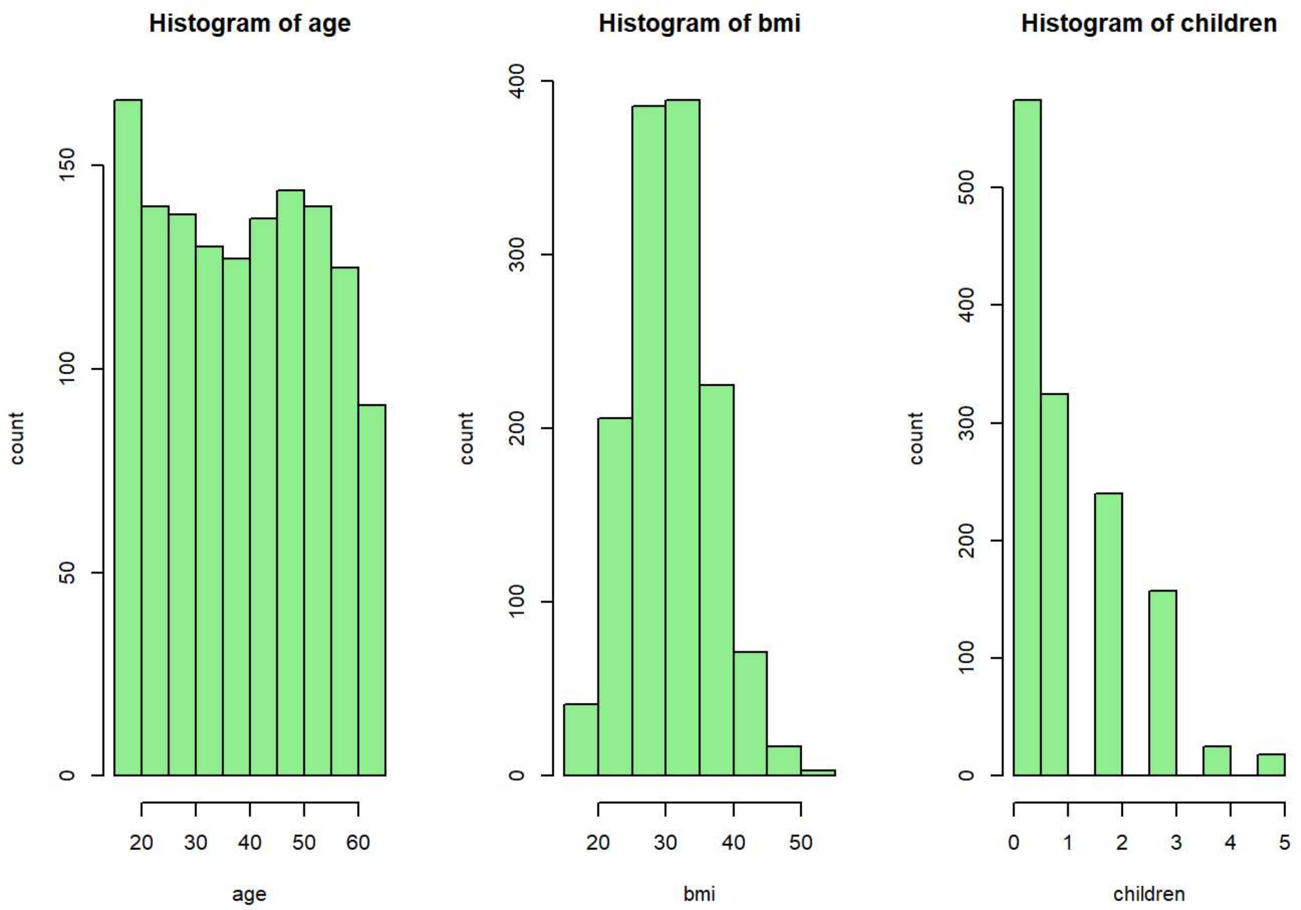
```
#checking the variable datatype we have
sapply(insurance,class)
```

```
##         age         sex         bmi    children      smoker      region 
##   "integer" "character"   "numeric"   "integer" "character" "character" 
##     charges 
##   "numeric" 
```

```
#converting sex and smoker into factor
insurance$sex <- as.factor(insurance$sex)
insurance$smoker <- as.factor(insurance$smoker)

#regions within our dataset
unique(insurance$region)
```

```
## [1] "southwest" "southeast" "northwest" "northeast"
```

```
dt <- data.frame(table(insurance$region))

#Ages within our dataset
insurance1 <- insurance[order(insurance$age,decreasing = FALSE),]
ages <- unique(insurance1$age)

#histogram of age, bmi & childs
par(mfrow = c(1,3))
hist(insurance$age, main = "Histogram of age", col = "lightgreen", xlab = "age", ylab= "coun
t")
hist(insurance$bmi, main = "Histogram of bmi", col = "lightgreen", xlab = "bmi", ylab= "coun
t")
hist(insurance$children, main = "Histogram of children", col = "lightgreen", xlab = "childre
n", ylab= "count")
```



```
#knowing which variable in important
model <- lm(charges ~ age + sex + bmi + children + smoker + region, data = insurance)
summary(model)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
relative_importance <- calc.relimp(model, type = "lmg", rela = TRUE)
sort(relative_importance$lmg, decreasing = TRUE)
```

```
##      smoker         age         bmi    children      region         sex
## 0.827775028 0.118343443 0.042750878 0.004552573 0.004459426 0.002118653
```

```
## Summarize medical expenses
summary(insurance$charges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```
## Correlation matrix
cor(insurance[c("age", "bmi", "children", "charges")])
```

```
##                age       bmi   children    charges
## age      1.0000000 0.1092719 0.04246900 0.29900819
## bmi      0.1092719 1.0000000 0.01275890 0.19834097
## children 0.0424690 0.0127589 1.00000000 0.06799823
## charges  0.2990082 0.1983410 0.06799823 1.00000000
```

```
model <- lm(charges ~ sex, data = insurance)
summary(model)$coef
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 12569.579   470.0717 26.739706 1.626108e-126
## sexmale      1387.172   661.3309  2.097547  3.613272e-02
```

```
model
```

```
## 
## Call:
## lm(formula = charges ~ sex, data = insurance)
## 
## Coefficients:
## (Intercept)      sexmale  
##       12570         1387
```

```
#Interpretations

# 1. Does smoking affect the insurance price?



boxplot(insurance$charges ~ insurance$smoker,main="Box plot of smoking records in terms of In
surance Charges",
        xlab = "Smoking Record", ylab = "Insurance Charges")




ggplot(insurance, aes(x=charges, y=bmi, color = smoker)) +
  geom_point(size=2, shape=23) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
# 2.Percentage of males vs females who have opted for insurance?



insurance %>% ggplot(aes(x = '', y = ..count.., fill = insurance$sex)) +
  geom_bar() + coord_polar('y', start = 0)
```

```
## Warning: Use of `insurance$sex` is discouraged. Use `sex` instead.
```

```
Total_count <- table(insurance$sex)
Total_count
```

```
## 
## female   male 
##    662    676
```

```
# 3. Can a dummy variable or subset be created within this data set?



MALE <- subset(insurance, insurance$sex == "male")
headtail(MALE)
```

```
## Warning: headtail is deprecated. Please use the headTail function
```

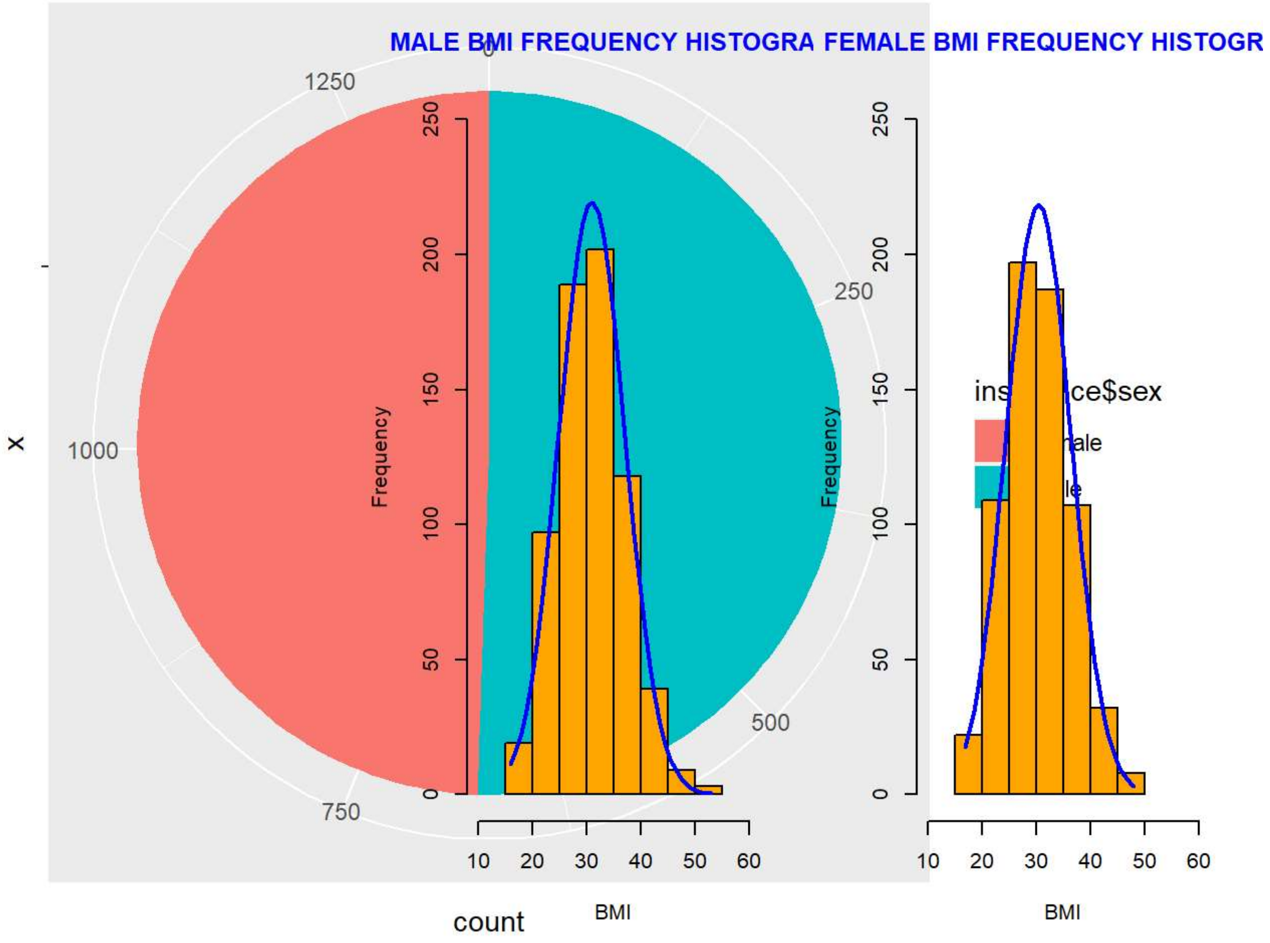
```
##      age  sex   bmi children smoker    region  charges
## 2     18 male 33.77        1     no southeast  1725.55
## 3     28 male    33        3     no southeast  4449.46
## 4     33 male  22.7        0     no northwest 21984.47
## 5     32 male 28.88        0     no northwest  3866.86
## ...  ... <NA>   ...      ...   <NA>      <NA>      ...
## 1326  61 male 33.53        0     no northeast 13143.34
## 1328  51 male 30.03        1     no southeast   9377.9
## 1330  52 male  38.6        2     no southwest 10325.21
## 1334  50 male 30.97        3     no northwest 10600.55
```

```
FEMALE <- subset(insurance, insurance$sex == "female")
headtail(FEMALE)
```

```
## Warning: headtail is deprecated. Please use the headTail function
```

```
##      age    sex   bmi children smoker    region  charges
## 1     19 female  27.9        0    yes southwest 16884.92
## 6     31 female 25.74        0     no southeast  3756.62
## 7     46 female 33.44        1     no southeast  8240.59
## 8     37 female 27.74        3     no northwest  7281.51
## ...  ...   <NA>   ...      ...   <NA>      <NA>      ...
## 1335  18 female 31.92        0     no northeast  2205.98
## 1336  18 female 36.85        0     no southeast  1629.83
## 1337  21 female  25.8        0     no southwest  2007.94
## 1338  61 female 29.07        0    yes northwest 29141.36
```

```
histo1 <-hist(MALE$bmi, ylab = "Frequency", xlab = "BMI", main = "MALE BMI FREQUENCY HISTOGRAM",
              xlim = c(10,60), ylim = c(0,250), col = "orange", col.main ="blue")
xfit<-seq(min(MALE$bmi),max(MALE$bmi),length=40)
yfit<-dnorm(xfit,mean=mean(MALE$bmi),sd=sd(MALE$bmi))
yfit <- yfit*diff(histo1$mids[1:2])*length(MALE$bmi)
lines(xfit, yfit, col="blue", lwd=2)




histo2 <-hist(FEMALE$bmi, ylab = "Frequency", xlab = "BMI", main = "FEMALE BMI FREQUENCY HISTOGRAM",
              xlim = c(10,60), ylim = c(0,250), col = "orange", col.main ="blue")
xfit<-seq(min(FEMALE$bmi),max(FEMALE$bmi),length=40)
yfit<-dnorm(xfit,mean=mean(FEMALE$bmi),sd=sd(FEMALE$bmi))
yfit <- yfit*diff(histo2$mids[1:2])*length(FEMALE$bmi)
lines(xfit, yfit, col="blue", lwd=2)
```



```
summary(MALE)
```

```
##       age             sex           bmi            children      smoker
##  Min.   :18.00   female:  0   Min.   :15.96   Min.   :0.000   no :517
##  1st Qu.:26.00   male  :676   1st Qu.:26.41   1st Qu.:0.000   yes:159
##  Median :39.00                Median :30.69   Median :1.000
##  Mean   :38.92                Mean   :30.94   Mean   :1.115
##  3rd Qu.:51.00                3rd Qu.:34.99   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##     region             charges
##  Length:676         Min.   : 1122
##  Class :character   1st Qu.: 4619
##  Mode  :character   Median : 9370
##                     Mean   :13957
##                     3rd Qu.:18990
##                     Max.   :62593
```

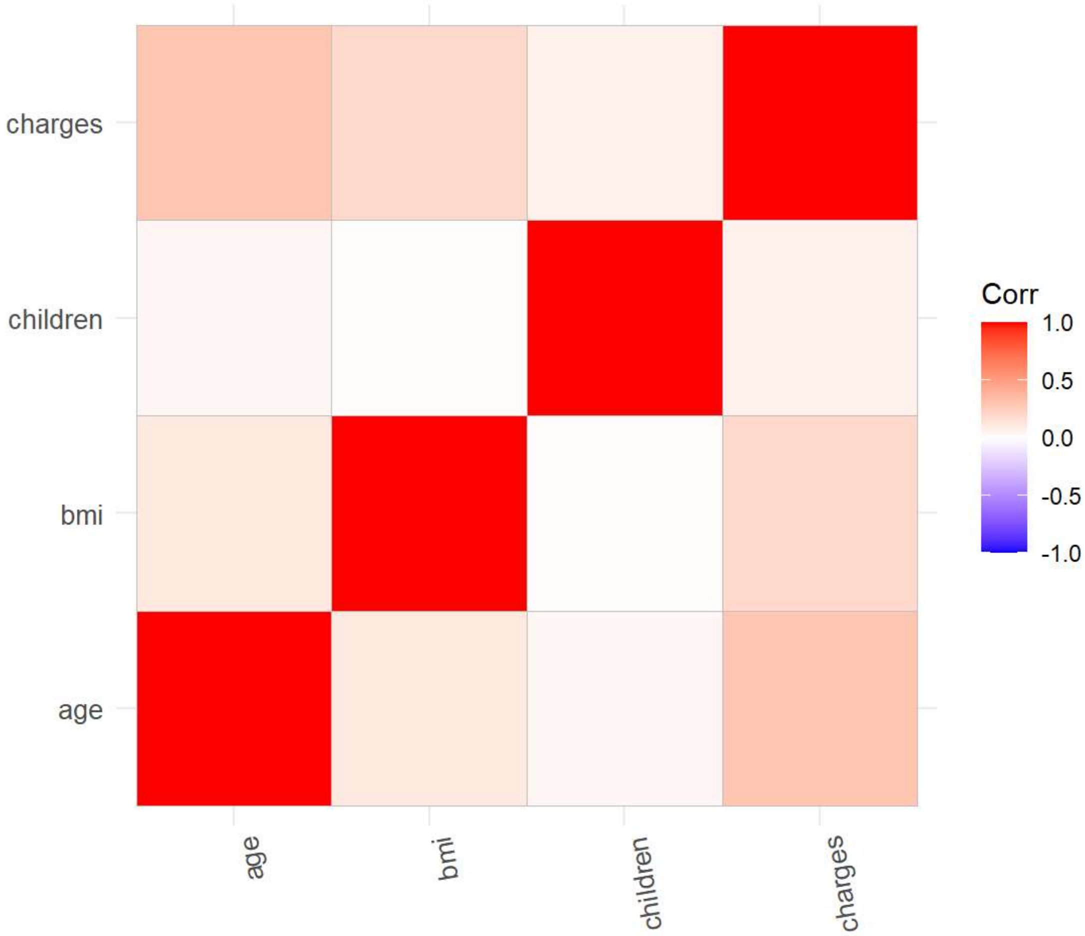
```
summary(FEMALE)
```

```
##       age             sex           bmi            children      smoker
##  Min.   :18.00   female:662   Min.   :16.82   Min.   :0.000   no :547
##  1st Qu.:27.00   male  :  0   1st Qu.:26.12   1st Qu.:0.000   yes:115
##  Median :40.00                Median :30.11   Median :1.000
##  Mean   :39.50                Mean   :30.38   Mean   :1.074
##  3rd Qu.:51.75                3rd Qu.:34.31   3rd Qu.:2.000
##  Max.   :64.00                Max.   :48.07   Max.   :5.000
##     region             charges
##  Length:662         Min.   : 1608
##  Class :character   1st Qu.: 4885
##  Mode  :character   Median : 9413
##                     Mean   :12570
##                     3rd Qu.:14455
##                     Max.   :63770
```

```
# 4. Are there any variables that influence price?

Correlation_matrix <-cor(insurance[sapply(insurance,is.numeric)])
Correlation_matrix
```

```
##                age       bmi   children    charges
## age      1.0000000 0.1092719 0.04246900 0.29900819
## bmi      0.1092719 1.0000000 0.01275890 0.19834097
## children 0.0424690 0.0127589 1.00000000 0.06799823
## charges  0.2990082 0.1983410 0.06799823 1.00000000
```

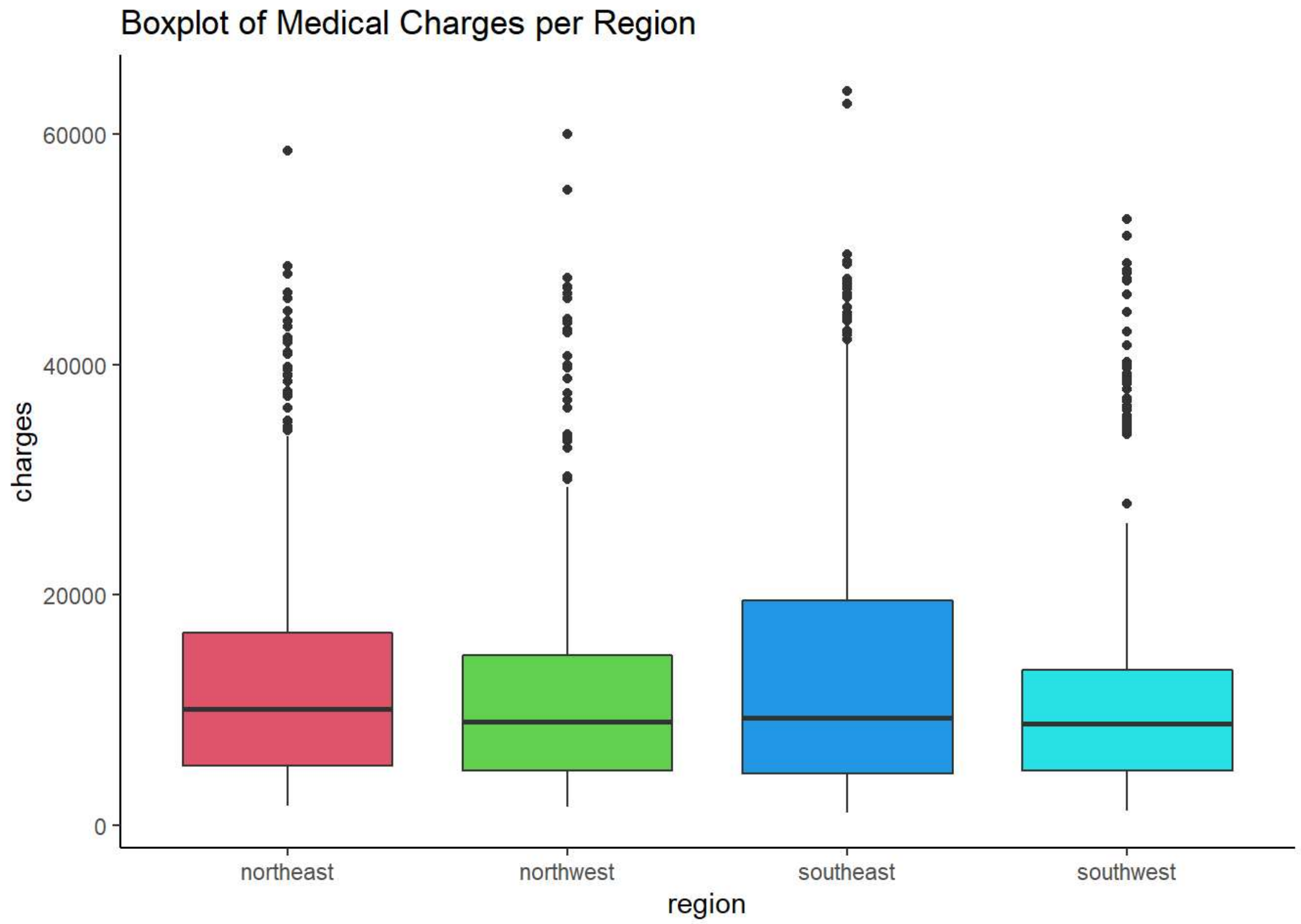
```
ggcorrplot(Correlation_matrix,method = "square",tl.cex = 10, tl.srt = 100)
```



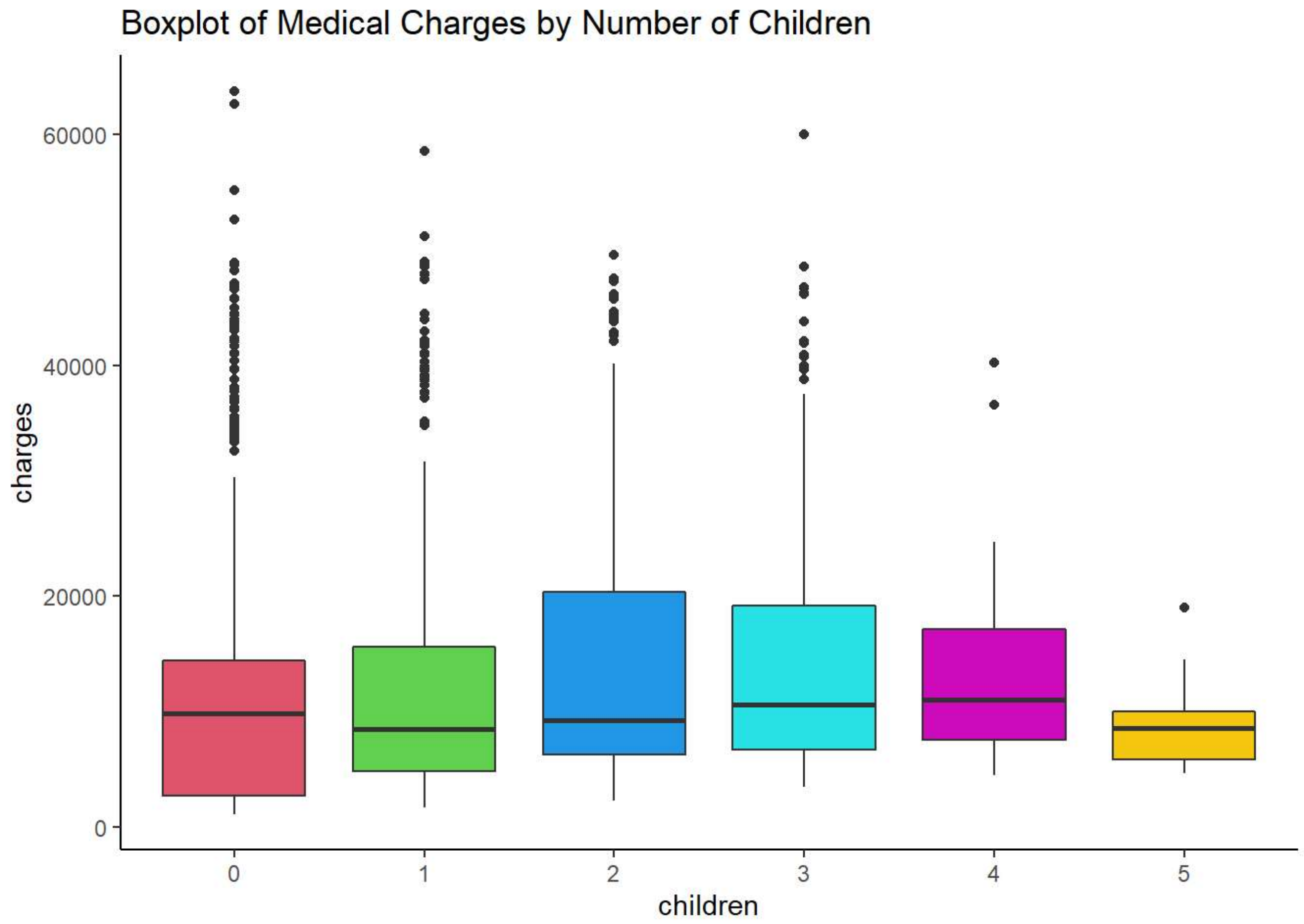
```
#Linear model for smoking and charges
lm(charges~smoker, data=insurance)
```

```
##
## Call:
## lm(formula = charges ~ smoker, data = insurance)
##
## Coefficients:
## (Intercept)    smokeryes
##        8434        23616
```

```
#region affecting price (region does not have much impact on the charges)
ggplot(data = insurance,aes(region,charges)) + geom_boxplot(fill = c(2:5)) +
  theme_classic() + ggtitle("Boxplot of Medical Charges per Region")
```



```
#children affecting price
ggplot(data = insurance,aes(as.factor(children),charges)) + geom_boxplot(fill = c(2:7)) +
  theme_classic() +  xlab("children") +
  ggtitle("Boxplot of Medical Charges by Number of Children")
```



```
##### draft report #####



#chi-square
table3 <- table("smoker" =insurance$smoker, "charges"=insurance$charges)
View(table3)

table4 <- table("sex" = insurance$sex,"smoker" = insurance$smoker)
View(table4)

#h0: gender and smoker are independent of one another
#h1: gender and smoker are dependent on one another
#critical value = 0.05
ch2 <- chisq.test(table4)
ch2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table4
## X-squared = 7.3929, df = 1, p-value = 0.006548
```

```
alpha <- 0.05
ifelse(ch2$p.value>alpha,"Fail to reject Null Hypothesis","Reject Null Hypothesis")
```

```
## [1] "Reject Null Hypothesis"
```

```
#h0: smoker affects the charges
#h1: smoker does not affects the charges
#critical value = 0.05
ch1 <- chisq.test(table3)
```

```
## Warning in chisq.test(table3): Chi-squared approximation may be incorrect
```

```
ch1
```

```
##
##  Pearson's Chi-squared test
##
## data:  table3
## X-squared = 1338, df = 1336, p-value = 0.4794
```

```
alpha <- 0.05
ifelse(ch1$p.value>alpha,"Fail to reject Null Hypothesis","Reject Null Hypothesis")
```

```
## [1] "Fail to reject Null Hypothesis"
```

```
#### method 1 Linear regression####
#splitting on 80%
n_train <- round(0.8 * nrow(insurance))
trainIndex <- sample(1:nrow(insurance), n_train)
train <- insurance[trainIndex, ]
test <- insurance[-trainIndex, ]

#linear model
model1<- lm(charges ~ age + sex + bmi + children + smoker + region, data=train)
summary(model1)
```

```
## 
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker + 
##     region, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11259.7  -2883.2   -861.1   1509.8  29898.8 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -12743.51    1097.01 -11.617  < 2e-16 ***
## age                266.54      13.21  20.172  < 2e-16 ***
## sexmale            -57.08     365.61  -0.156  0.87598    
## bmi                356.21      31.96  11.147  < 2e-16 ***
## children           383.82     149.18   2.573  0.01022 *  
## smokeryes        24062.79     454.20  52.978  < 2e-16 ***
## regionnorthwest   -611.15     524.09  -1.166  0.24383    
## regionsoutheast  -1377.09     523.61  -2.630  0.00866 ** 
## regionsouthwest  -1185.30     524.60  -2.259  0.02406 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5944 on 1061 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7571 
## F-statistic: 417.5 on 8 and 1061 DF,  p-value: < 2.2e-16
```

```
#getting statistics
r2 <- summary(model1)$r.squared
r2
```

```
## [1] 0.758898
```

```
#predicting data on test
pred <- predict(model1, newdata=test)

#rmse
rmse1 <- RMSE(test$charges,pred)
rmse1
```

```
## [1] 6540.783
```

```
#creating model without sex

model2 <- lm(charges ~ age + bmi + children + smoker + region, data=train)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11288.2  -2863.5   -857.5   1501.4  29873.6
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -12763.48    1089.03 -11.720  < 2e-16 ***
## age                266.61      13.20  20.196  < 2e-16 ***
## bmi                355.87      31.87  11.167  < 2e-16 ***
## children           383.25     149.07   2.571  0.01028 *
## smokeryes        24057.64     452.79  53.132  < 2e-16 ***
## regionnorthwest   -609.46     523.74  -1.164  0.24482
## regionsoutheast  -1376.60     523.36  -2.630  0.00865 **
## regionsouthwest  -1184.72     524.34  -2.259  0.02406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5941 on 1062 degrees of freedom
## Multiple R-squared:  0.7589, Adjusted R-squared:  0.7573
## F-statistic: 477.5 on 7 and 1062 DF,  p-value: < 2.2e-16
```

```
#getting statistics
r2_1 <- summary(model2)$r.squared
r2_1
```

```
## [1] 0.7588925
```

```
#predicting data on test
pred1 <- predict(model2, newdata=test)

#rmse
rmse2 <- RMSE(test$charges,pred1)
rmse2
```

```
## [1] 6541.604
```

```
#compare statistics for above linear and select which is best
#### higher r2 and lower rmse considers a good fit (going with model2)



#### method 2 stepwise selection through aic####

#1st model with only intercept
start <- lm(charges~1, data =insurance)

#2nd model with all predictor variables
all <- lm(charges~.,data = insurance)
formula(all)
```

```
## charges ~ age + sex + bmi + children + smoker + region
```

```
#performing stepwise to get best model with lower aic
step(start, direction = "both", scope = formula(all))
```

```
## Start:  AIC=25160.18
## charges ~ 1
##
##              Df  Sum of Sq        RSS   AIC
## + smoker      1 1.2152e+11 7.4554e+10 23868
## + age         1 1.7530e+10 1.7854e+11 25037
## + bmi         1 7.7134e+09 1.8836e+11 25109
## + children    1 9.0660e+08 1.9517e+11 25156
## + region      3 1.3008e+09 1.9477e+11 25157
## + sex         1 6.4359e+08 1.9543e+11 25158
## <none>                     1.9607e+11 25160
##
## Step:  AIC=23868.38
## charges ~ smoker
##
##              Df  Sum of Sq        RSS   AIC
## + age         1 1.9928e+10 5.4626e+10 23454
## + bmi         1 7.4856e+09 6.7069e+10 23729
## + children    1 7.5272e+08 7.3802e+10 23857
## <none>                     7.4554e+10 23868
## + sex         1 1.4213e+06 7.4553e+10 23870
## + region      3 1.0752e+08 7.4447e+10 23873
## - smoker      1 1.2152e+11 1.9607e+11 25160
##
## Step:  AIC=23454.24
## charges ~ smoker + age
##
##              Df  Sum of Sq        RSS   AIC
## + bmi         1 5.1129e+09 4.9513e+10 23325
## + children    1 4.5928e+08 5.4167e+10 23445
## <none>                     5.4626e+10 23454
## + sex         1 2.2255e+06 5.4624e+10 23456
## + region      3 1.3843e+08 5.4488e+10 23457
## - age         1 1.9928e+10 7.4554e+10 23868
## - smoker      1 1.2392e+11 1.7854e+11 25037
##
## Step:  AIC=23324.76
## charges ~ smoker + age + bmi
##
##              Df  Sum of Sq        RSS   AIC
## + children    1 4.3477e+08 4.9078e+10 23315
## + region      3 2.3201e+08 4.9281e+10 23325
## <none>                     4.9513e+10 23325
## + sex         1 3.9429e+06 4.9509e+10 23327
## - bmi         1 5.1129e+09 5.4626e+10 23454
## - age         1 1.7556e+10 6.7069e+10 23729
## - smoker      1 1.2358e+11 1.7310e+11 24997
##
## Step:  AIC=23314.96
## charges ~ smoker + age + bmi + children
##
##              Df  Sum of Sq        RSS   AIC
## + region      3 2.3320e+08 4.8845e+10 23315
## <none>                     4.9078e+10 23315
## + sex         1 5.4861e+06 4.9073e+10 23317
## - children  1 4.3477e+08 4.9513e+10 23325
## - bmi       1 5.0884e+09 5.4167e+10 23445
## - age       1 1.7297e+10 6.6375e+10 23717
## - smoker    1 1.2345e+11 1.7253e+11 24995
## 
## Step:  AIC=23314.58
## charges ~ smoker + age + bmi + children + region
## 
##             Df  Sum of Sq        RSS   AIC
## <none>                    4.8845e+10 23315
## - region    3 2.3320e+08 4.9078e+10 23315
## + sex       1 5.7164e+06 4.8840e+10 23316
## - children  1 4.3596e+08 4.9281e+10 23325
## - bmi       1 5.1645e+09 5.4010e+10 23447
## - age       1 1.7151e+10 6.5996e+10 23715
## - smoker    1 1.2301e+11 1.7186e+11 24996
```

```
## 
## Call:
## lm(formula = charges ~ smoker + age + bmi + children + region, 
##     data = insurance)
## 
## Coefficients:
##     (Intercept)        smokeryes              age              bmi  
##        -11990.3          23836.3            257.0            338.7  
##        children  regionnorthwest  regionsoutheast  regionsouthwest  
##           474.6           -352.2          -1034.4           -959.4
```

```
#### appendix ####
#model comparison through anova and aic,bic

#h0: adding bmi, children and age does not improve the model
#h1: adding bmi, children and age does improve the model
#comparing models with anova
fit1 <- lm(formula = charges ~ smoker , data = insurance)
fit2 <- lm(formula = charges ~ smoker + bmi + children + age, data = insurance)
alpha = 0.05
ann <- anova(fit1,fit2)
p.value <- ann$`Pr(>F)`[2]
ifelse(p.value>alpha,"Fail to reject Null Hypothesis","Reject Null Hypothesis")
```

```
## [1] "Reject Null Hypothesis"
```

```
#using aic to compare
AIC(fit1,fit2)#low aic means best
```

```
##      df      AIC
## fit1  3 27667.46
## fit2  6 27114.04
```

```
BIC(fit1,fit2)#low bic means best
```

```
##      df      BIC
## fit1  3 27683.06
## fit2  6 27145.23
```

```
#fit2 is preferred.
```