Assignment 1 - Medical Cost
Statistical Inference and Modelling - SIM
1st Semester 2022

Ander Barrio Campos, Odysseas Kyparissis

2022-11-19

Contents

1 Explanatory Data Analysis - EDA

1.1 Loading Insurance Data

In this part of the report, setting up the working environment and loading of the data into R are taking place. Additionally, a first look at the summary of the raw insurance data set is taken.

```
df <- read.csv("insurance.csv")
summary(df)
```

```
##       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
```

1.2 Data Types

To begin with, the types of the raw variables contained into the data set are being checked.

```
typeof(df$age)
```

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

```
typeof(df$sex)
```

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

```
typeof(df$bmi)
```

```
## [1] "double"
```

```
typeof(df$children)
```

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

```
typeof(df$smoker)
```

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

```
typeof(df$region)
```

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

```
typeof(df$charges)
```

```
## [1] "double"
```

It is clear, that the data set consists of 4 numerical variables and 3 categorical ones. The numeric variables are the following: age, BMI, children and charges, while the categorical ones are: sex, smoker and region. In the following sections, categorical variables will be transformed into labeled factors, as well as, new derived factors will be produced from the numerical values in order to see their performance on the regression modelling process.

1.3 Checking for Missing Data

To continue with, a check for missing data is conducted on the raw data set. Considering the summary of the data set presented before, there are no NA values in the variables of the data set. The same conclusion is derived when a check is completed for each individual variable.

1.4 Checking for Duplicates

By checking if there are duplicate rows inside the raw data set, the result indicates that a single occurrence of a duplicate exists and its index is equal to 582 as it is shown below.

```
dupli <- duplicated(df)
dupli_ind <- which(dupli)
dupli_ind; length(dupli_ind)
```

```
## [1] 582
```

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

With the following command, a closer look can be taken into the values of the duplicate row.

```
df [dupli_ind,]
```

```
##     age  sex   bmi children smoker    region  charges
## 582  19 male 30.59        0     no northwest 1639.563
```

To continue with the explanatory data analysis, the duplicate row is removed from the raw data set. Finally, a check for the existence of duplicates is taking place which results in a FALSE statement.

```
df<- df[-dupli_ind, ]
any(duplicated(df))
```

```
## [1] FALSE
```

1.5 Creating Factors for Qualitative Variables

In this subsection of EDA, all qualitative variables are transformed into labeled factors. The qualitative variables, as mentioned before, are sex, smoker and region. First of all, the unique values of these 3 variables are presented below:

```
unique(df$sex)
```

```
## [1] "female" "male"
```

```
unique(df$smoker)
```

```
## [1] "yes" "no"
```

```r
unique(df$region)
```

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

The next step includes the creation of the labeled factors based on the unique values of the categorical variables. Following the practice below, in case a categorical variable includes NA values, they will be transformed into zeros, which is an incorrect approach. In this case, once missing values check indicated that there are no missing data, proceeding with this practice does not result in erroneous data.

1.5.1 Sex to Labeled Factor

```r
# 1 - Initialize a variable with all values setted at 0
df$f.sex<-0
# 2 - Change its value for the cases where type of sex equals "male"
df$f.sex[df$sex=="male"]<-1
# 3 - Convering f.sex to labeled factor
df$f.sex<-factor(df$f.sex, labels=c("F","M"))
```

1.5.2 Smoker to Labeled Factor

```r
# 1 - Initialize a variable with all values setted at 0
df$f.smok<-0
# 2 - Change its value for the cases where type of smoker equals "yes"
df$f.smok[df$smoker=="yes"]<-1
# 3 - Convering f.smok to labeled factor
df$f.smok<-factor(df$f.smok, labels=c("No","Yes"))
```

1.5.3 Region to Labeled Factor

```r
# 1 - Initialize a variable with all values setted at 0
df$f.reg<-0
# 2 - Change its value for the cases where type of region equals:
# "southwest", "southeast", "northwest"
df$f.reg[df$region=="southwest"]<-3
df$f.reg[df$region=="southeast"]<-2
df$f.reg[df$region=="northwest"]<-1
# 3 - Convering f.reg to labeled factor
df$f.reg<-factor(df$f.reg,labels=c("NE","NW","SE","SW"))
```

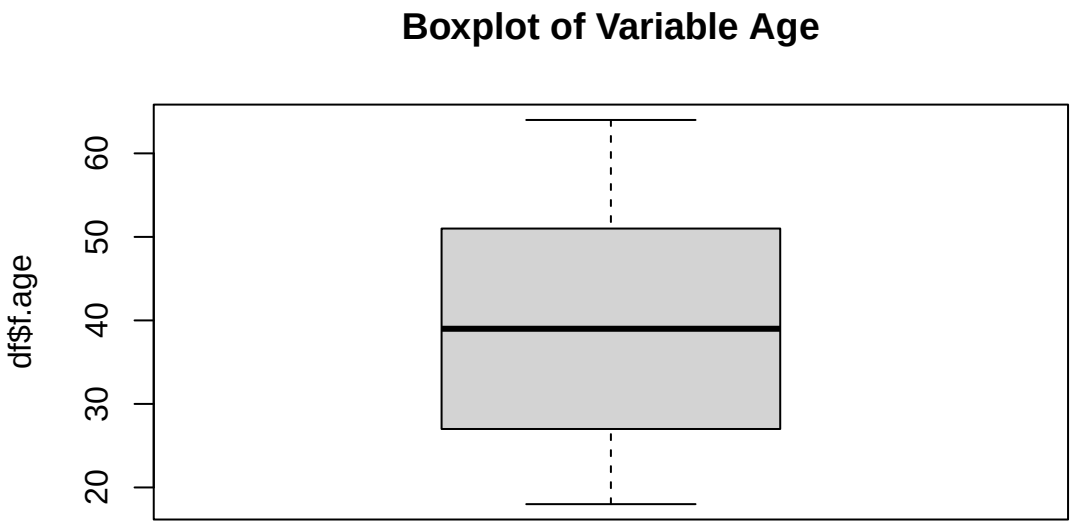
1.6 Creating Factors for Numerical Variables

This step is created in order to extract factors from numerical variables. This approach's goal is to check if some variables are more descriptive as a factor rather than as a numeric feature while training linear models. From the numerical variables of the data set, only age and BMI will be converted to factors, not the target variable (charges). Firstly, the discretization of the variable's values is taking place followed by the assigning of a label for each divided group.

1.6.1 Age to Labeled Factor

```
# 1 - Copying column age to a new column named f.age
df$f.age<-df$age

# 2 - Checking distribution of sample for variable age
# in order to decide how to discretize the values
Boxplot(df$f.age, main= "Boxplot of Variable Age")
```




```
# Once age distribution is almost equally devided
# into its range values, it is decided to discretize
# by using equal intervals
df$f.age<-discretize(df$f.age, method = "interval", breaks = 3,
                     labels = c("Young","Medium", "Old"))
```

The result of the discretization for age is calculated by separating the values of the variable into 3 equally-interval groups with labels: “Young”, “Medium” and “Old” respectively. The interval is equal to 15 years. Thus, group “Young” contain people in ages [18,33], group “Medium” contain individuals with ages [34,48] and finally “Old” group consist of people with ages [49,64].

1.6.2 BMI to Labeled Factor

For BMI, the discretization of the numerical value will be completed by using the labels “Low”, “Normal” and “High”. The values for creating the groups in this step are selected base on the Adult Body Mass Index values from healthcare bibliography.

```
# 1 - Copying column bmi to a new column named f.bmi
df$f.bmi<-df$bmi
# 2- Discretizing directly based on normal BMI values of Bibliography
df$f.bmi<-discretize(df$f.bmi, method = "fixed", breaks = c(-1,18.5,24.9,1000),
                     labels = c("Low","Normal", "High"))
```

1.7 Factor Conversion Check

After checking both manually and by executing commands on the terminal, the conversion of the categorical and numerical variables to factors has been completed correctly. In addition, while the categorical variables sex, region and smoker have been transformed into labeled factors, their old versions of type "chr" are discarded from the data frame. Below is presented the new structure of the data frame.

```
#We will only continue with those factor so we delete the previous variables
df$sex <- NULL #delete sex
df$region <- NULL #delete region
df$smoker <- NULL #delete smoker
str(df)
```

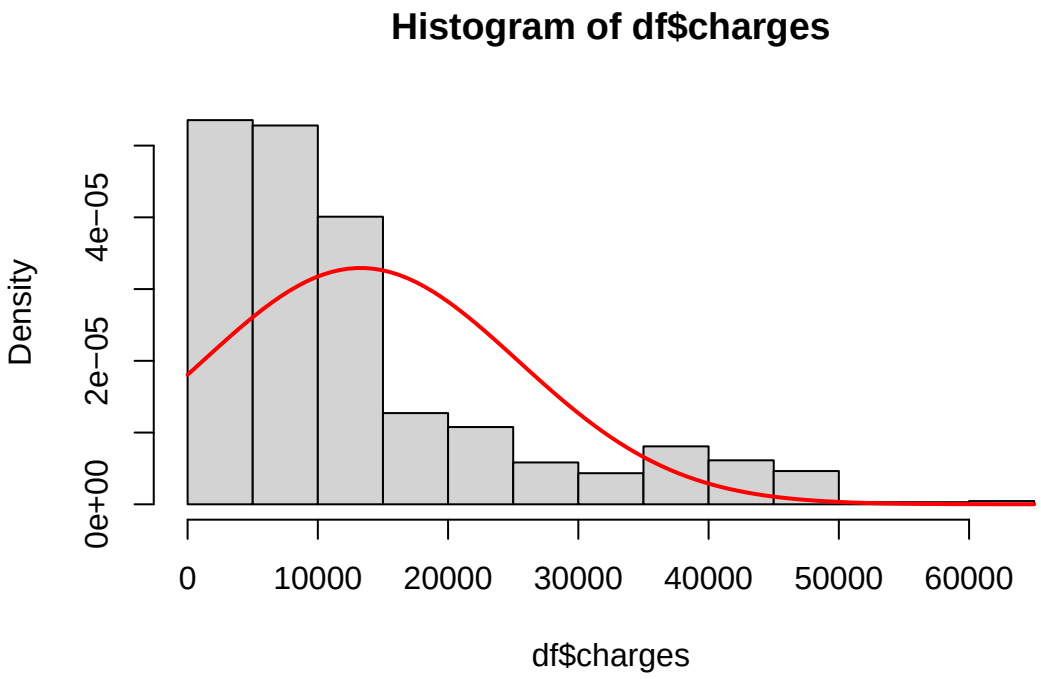
```
## 'data.frame':    1337 obs. of  9 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
##  $ f.sex   : Factor w/ 2 levels "F","M": 1 2 2 2 2 1 1 1 2 1 ...
##  $ f.smok  : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ f.reg   : Factor w/ 4 levels "NE","NW","SE",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ f.age   : Factor w/ 3 levels "Young","Medium",..: 1 1 1 1 1 1 2 2 2 3 ...
##   ..- attr(*, "discretized:breaks")= num [1:4] 18 33.3 48.7 64
##   ..- attr(*, "discretized:method")= chr "interval"
##  $ f.bmi   : Factor w/ 3 levels "Low","Normal",..: 3 3 3 2 3 3 3 3 3 3 ...
##   ..- attr(*, "discretized:breaks")= num [1:4] -1 18.5 24.9 1000
##   ..- attr(*, "discretized:method")= chr "fixed"
```

1.8 Normal Distribution Test for Target Variable (charges)

By taking a look at the histogram of the target variable and the density curve that describe a normal distribution with mean and standard deviation equal to the respective values of the data set, one can understand that the target variable does not follow a normal distribution. In order to be precise, by running the Shapiro test, the result indicate a value less than 0.05. Thus, the null hypothesis can be rejected and conclude that the target variable does not follow a normal distribution.

```
## [1] 13279.12
```

```
## [1] 12110.36
```

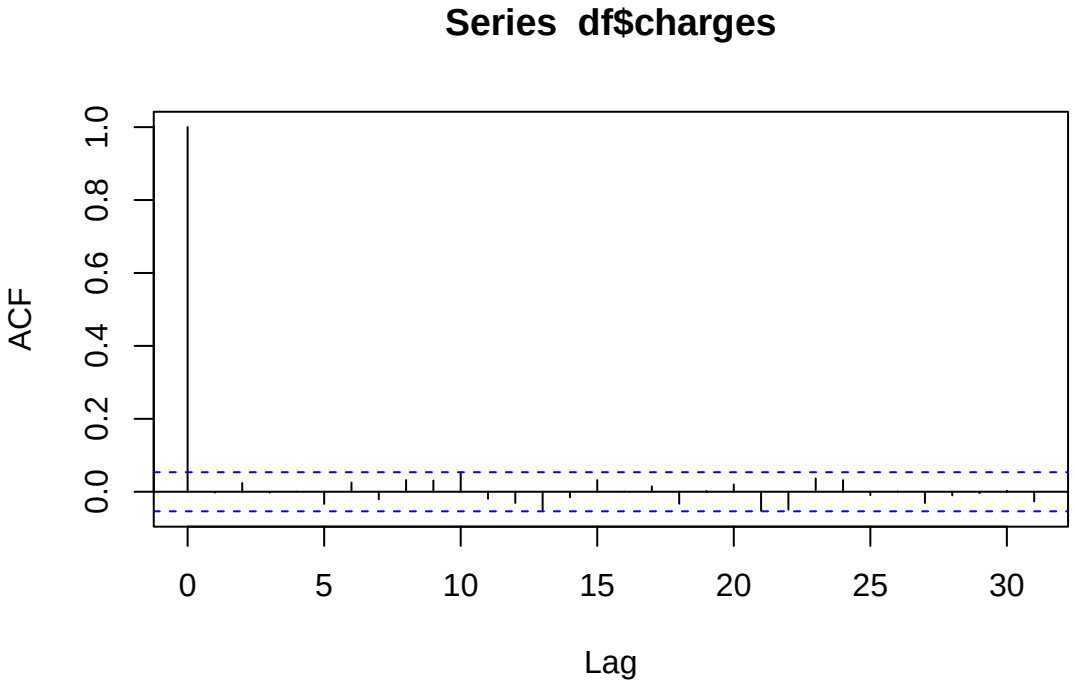



```
shapiro.test(df$charges)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$charges
## W = 0.81476, p-value < 2.2e-16
```

1.9 Serial Correlation

In order to address the serial correlation for the target variable two different approaches were followed. Firstly, the autocorrelation function was used which produces the ACF graph shown below.



From the graph, one can understand that all the vertical lines are inside the two horizontal blue lines except for the first one. The interpretation of this result is that there is no serial correlation for the target variable. Furthermore, for using statistical methods to address the same problem, the Durbin-Watson (DW) test was applied. The result is presented here:

```
dwtest(df$charges~1)
```

```
##
##  Durbin-Watson test
##
## data:  df$charges ~ 1
## DW = 2.0033, p-value = 0.5244
## alternative hypothesis: true autocorrelation is greater than 0
```

Once the resulting p-value is equal to 0.5 approximately, it means that the null hypothesis can not be rejected. The Durbin-Watson test has the null hypothesis that the autocorrelation of the disturbances is 0, thus serial correlation for the target variable is discarded.

1.10 Outliers Detection

In the following subsections both uni-variate and multivariate outliers will be detected and treated.

1.10.1 Univariate Outliers

To start with, in the following subsection the uni-variate outliers will be detected for the numerical variables: age, bmi, children and charges with the respective order. It is crucial to mention here, that only severe outliers were taken into account and not mild ones. Now, concerning variable age, as it was depicted before in the boxplot of the variable, outliers did not exist. The same result is derived after trying to detect outliers using the IQR method, which is implemented by function calcQ.

```
# 1 - AGE:
var_out<-calcQ(df$age)
llout_age<-which((df$age<var_out$souti)|(df$age>var_out$souts))
length(llout_age)
```

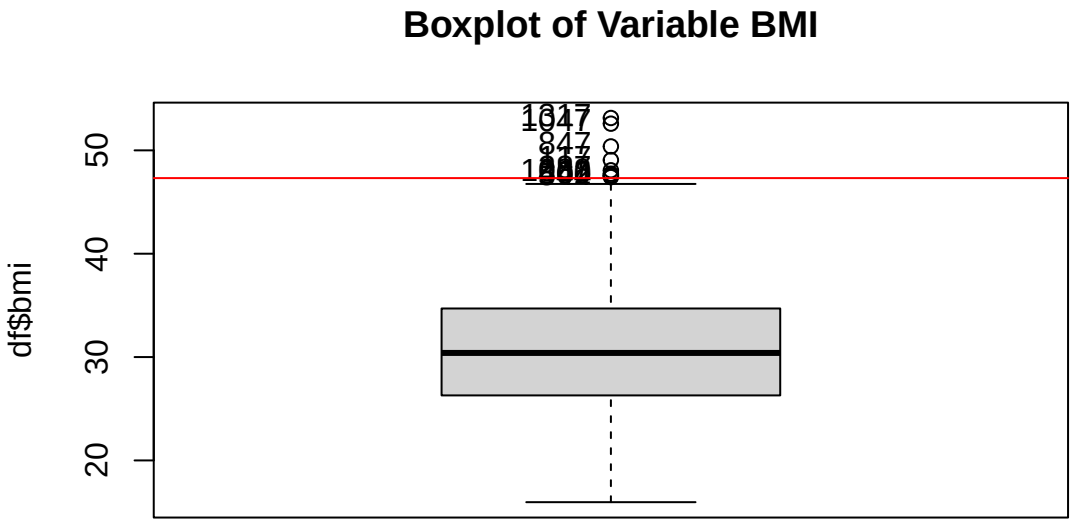
```
## [1] 0
```

This number indicates the number of indexes belonging to outlier observations for variable age, thus while it is zero it means that there are no severe outliers for variable age. Following by, the same approach is followed for variable BMI.

```
# 2 - BMI:
Boxplot(df$bmi, main = "Boxplot of Variable BMI")
```

```
## [1]  117  287  402  544  847  860 1047 1088 1317
```

```
var_out<-calcQ(df$bmi)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="red")
abline(h=var_out$mouts,col="red")
```



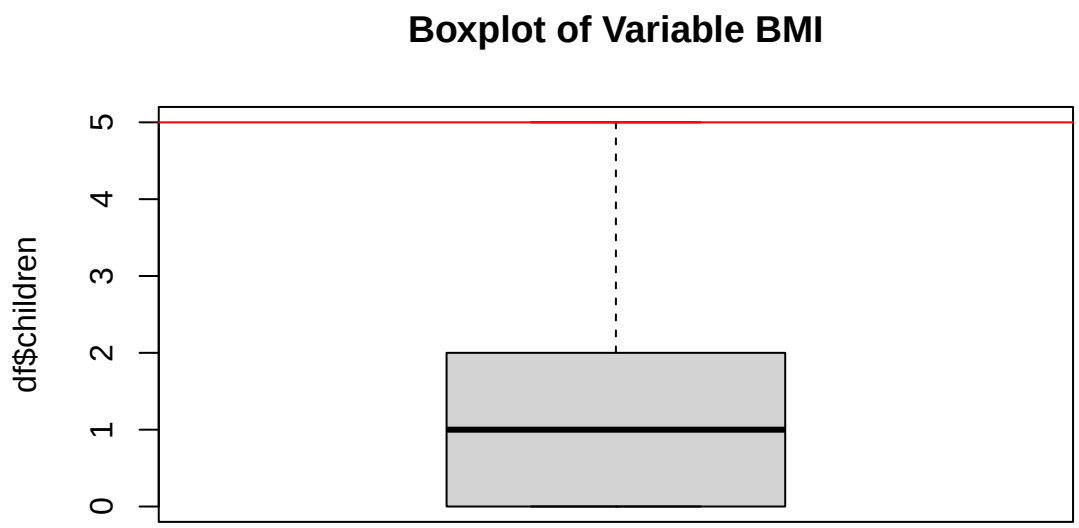

```
llout_bmi<-which((df$bmi<var_out$souti)|(df$bmi>var_out$souts))
length(llout_bmi)
```

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

The results are the same, there are no severe outliers for variable BMI as well, but in this case some mild ones appear but will not be treated. To continue with, same technique is used for variable children.

```
# 3 - CHILDREN:
Boxplot(df$children, main = "Boxplot of Variable BMI")
var_out<-calcQ(df$children)
```

```
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="red")
abline(h=var_out$mouts,col="red")
```




```
llout_children<-which((df$children<var_out$souti)|(df$children>var_out$souts))
length(llout_children)
```

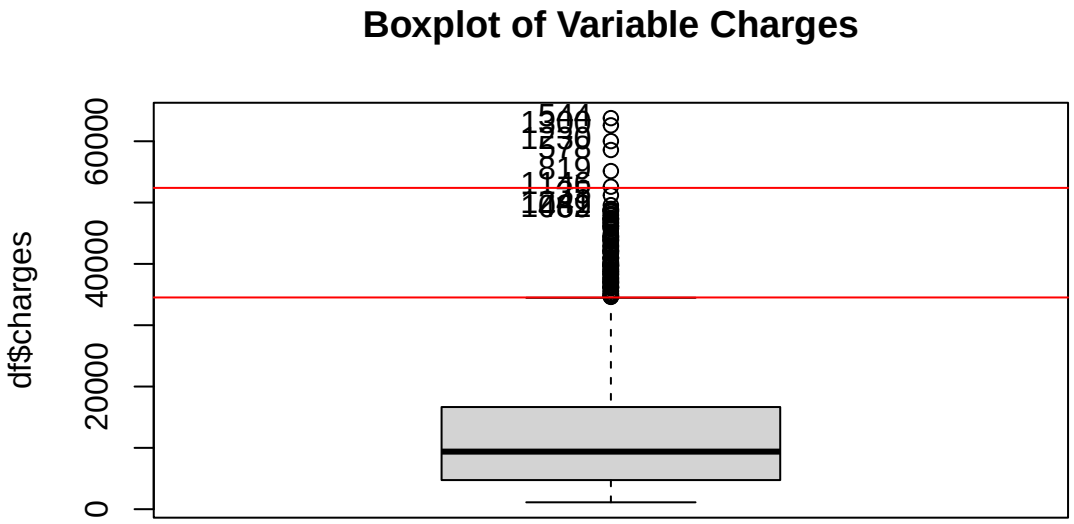
```
## [1] 0
```

And the results are again the same, there are not outliers for this variable as well. Finally, the outlier detection for the target variable is taking place.

```
# 4 - CHARGES:
Boxplot(df$charges, main = "Boxplot of Variable Charges")
```

```
##  [1]  544 1300 1230  578  819 1146   35 1241 1062  489
```

```
var_out<-calcQ(df$charges)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="red")
abline(h=var_out$mouts,col="red")
```




```
llout_charges<-which((df$charges<var_out$souti)|(df$charges>var_out$souts))
length(llout_charges)
```

```
## [1] 6
```

In this case, there are 6 extreme outliers for the target variable, which are presented below.

```
df[llout_charges,]
```

```
##      age    bmi children  charges f.sex f.smok f.reg  f.age f.bmi
## 544   54 47.410        0 63770.43     F    Yes    SE    Old  High
## 578   31 38.095        1 58571.07     F    Yes    NE  Young  High
## 820   33 35.530        0 55135.40     F    Yes    NW  Young  High
## 1147  60 32.800        0 52590.83     M    Yes    SW    Old  High
## 1231  52 34.485        3 60021.40     M    Yes    NW    Old  High
## 1301  45 30.360        0 62592.87     M    Yes    SE Medium  High
```

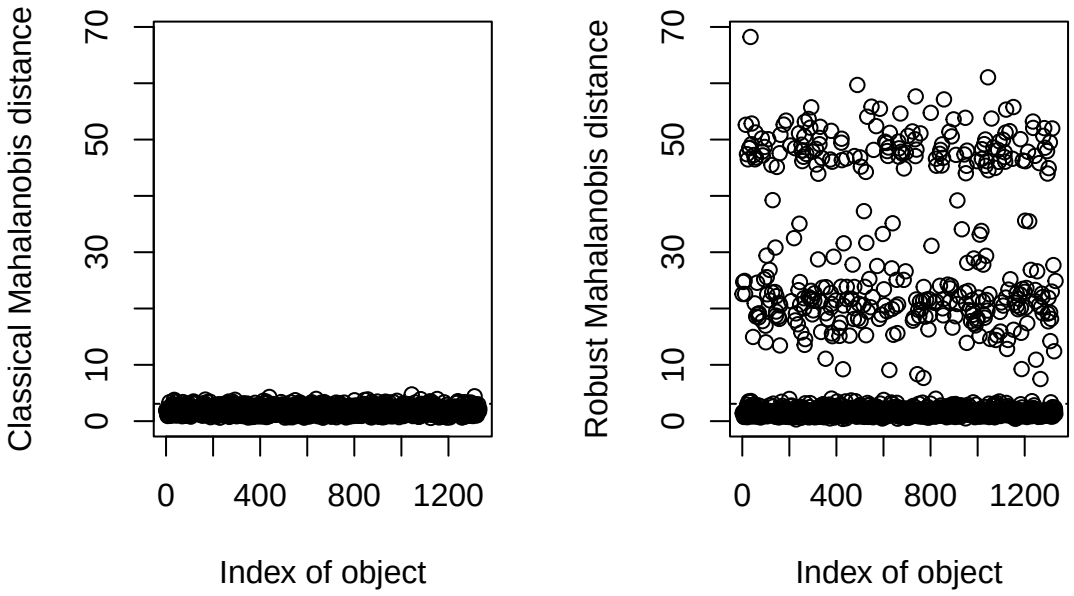
Due to the fact that the outliers are presented for the target variable, treating them would not be an ideal approach. By checking the values of the data set for those 6 cases, one conclusion that can be drawn is that all of them consern people who smoke and have a high BMI. The values of the other variables for those 6 cases are fluctuating. For this reason, it was decided to remove those observations from our further analysis.

```
df <- df[-llout_charges,]
```

1.10.2 Multivariate Outliers

In this subsection, the detection of multivariate outliers is taking place. As a first step, the calculation of the Mahalanobis distance with a confidence interval of 95% is completed. The Classical and Robust Mahalanobis distances for the numerical values of the dataset are presented in the following figure:

```
res.mout <- Moutlier( df[ , c(1:4)], quantile = 0.95 )
```

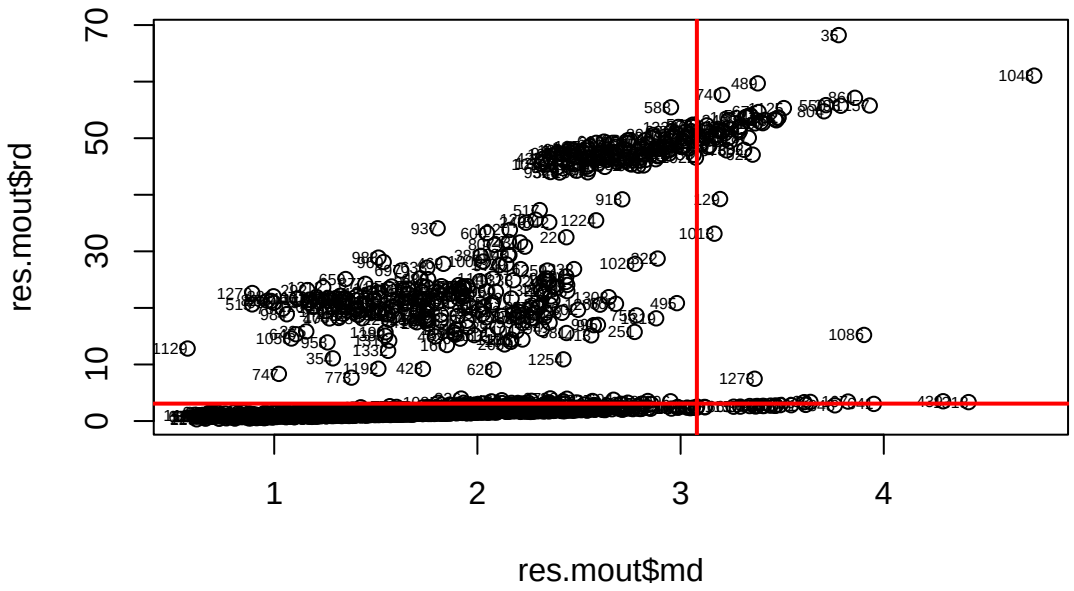


After calculating Mahalanobis distance at a 95% confidence interval, the cut off given is presented below.

```
res.mout$cutoff
```

```
## [1] 3.080216
```

Then, all the observations which have a classical and a robust distance bigger than this cut off are marked as multivariate outliers. After detecting them, a new factor is being created in the data set, indicating if an observation belongs to multivariate outliers or not. It can be seen in the final result that 53 observations are marked as multivariate outliers. Further analysis about them will be conducted in the following sections.




```
llmout <- which((res.mout$md > res.mout$cutoff)
                & (res.mout$rd > res.mout$cutoff))
#df[llmout,] # Observations of Multiple Outliers
df$mout <- 0
df$mout[llmout] <- 1
df$mout <- factor( df$mout, labels = c("MvOut.No","MvOut.Yes"))
summary(df["mout"])
```

```
##         mout
##  MvOut.No :1277
##  MvOut.Yes:  54
```

1.11 Preliminary Exploratory Analysis

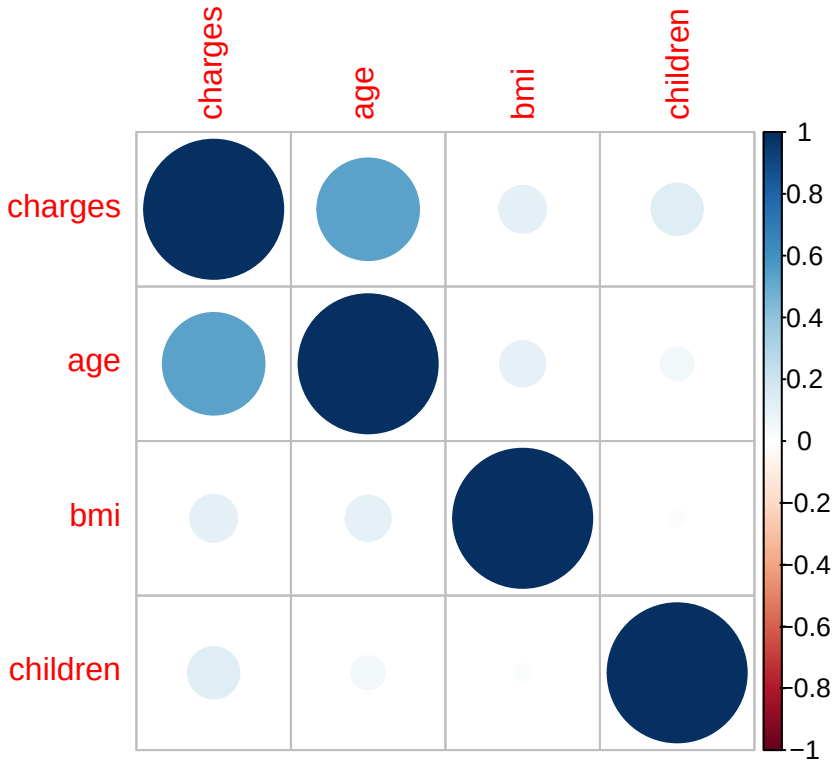
The goal of this chapter is to discover the relationships between the different variables of the data set. In order to do so, the following techniques were used: calculation of the Spearman correlation for the numerical variables, once the target variable does not follow a normal distribution, use of the library FactoMineR and Boxplots for presenting interactions between the target and categorical and numerical variables.

In the following graph, the correlation of the numerical values is presented. The only significant observation here is that age and charges have a slight strong correlation. By checking the correlation matrix the value between those two variables is equal to 0.53.

```
M <- cor(df[,c(4,1:3)],method="spearman");M #Non Parametric version
```

```
##             charges        age        bmi    children
## charges   1.0000000 0.53429693 0.11365380 0.13672344
## age       0.5342969 1.00000000 0.10664431 0.05742152
## bmi       0.1136538 0.10664431 1.00000000 0.01717882
## children 0.1367234 0.05742152 0.01717882 1.00000000
```

```
corrplot(M, method="circle")
```



Moreover, with the usage of the library FactoMineR and specifically the function condes, which calculate the dependencies of a continuous variable, it is able to check the dependencies of the target variable with the rest of the variables of the data set. At first, by looking at the dependency between the target variable and the numerical variables the following result is produced.

```
res.con<-condes(df[,1:9],4)
res.con$quanti
```

```
##          correlation      p.value
## age       0.30062715 3.329401e-29
## bmi       0.18862792 3.988446e-12
## children  0.07587579 5.613221e-03
```

It is clear again, that the bigger correlation (0.3 here) with the target variable exist for the age variable, with a great statistical significance due to the really small value of p-value presented. The conclutions here are the same with the ones produced by the correlation matrix.

Furthermore, the relationship of the target variable with the qualitative variables is conducted below.

```
res.con$quali
```

```
##                  R2       p.value
## f.smok 0.618076780 4.650154e-280
## f.age  0.077526782  5.361680e-24
## f.bmi  0.013067483  1.610172e-04
## f.sex  0.003654488  2.742457e-02
## f.reg  0.006613994  3.193769e-02
```

It is obvious, that the most important relationship (for the categorical variables) exist between charges and the factor "f.smok" with an R-square value approximately equal to 0.62 and an extremely low p-value. The rest of the categorical variables ("f.age", "f.bmi", "f.sex" and "f.reg") do not show a great relationship with the target variable since the R-squared values they achieve is really small (close to 0).

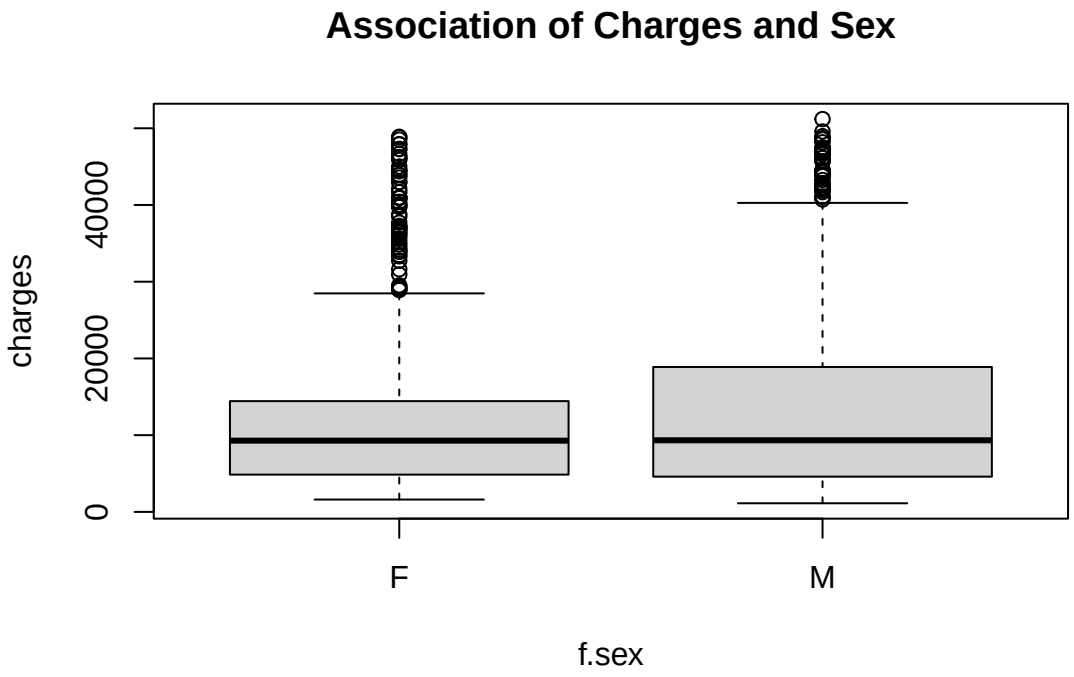
A closer look to the relationship of charges and the different categories of the categorical variables is presented below.

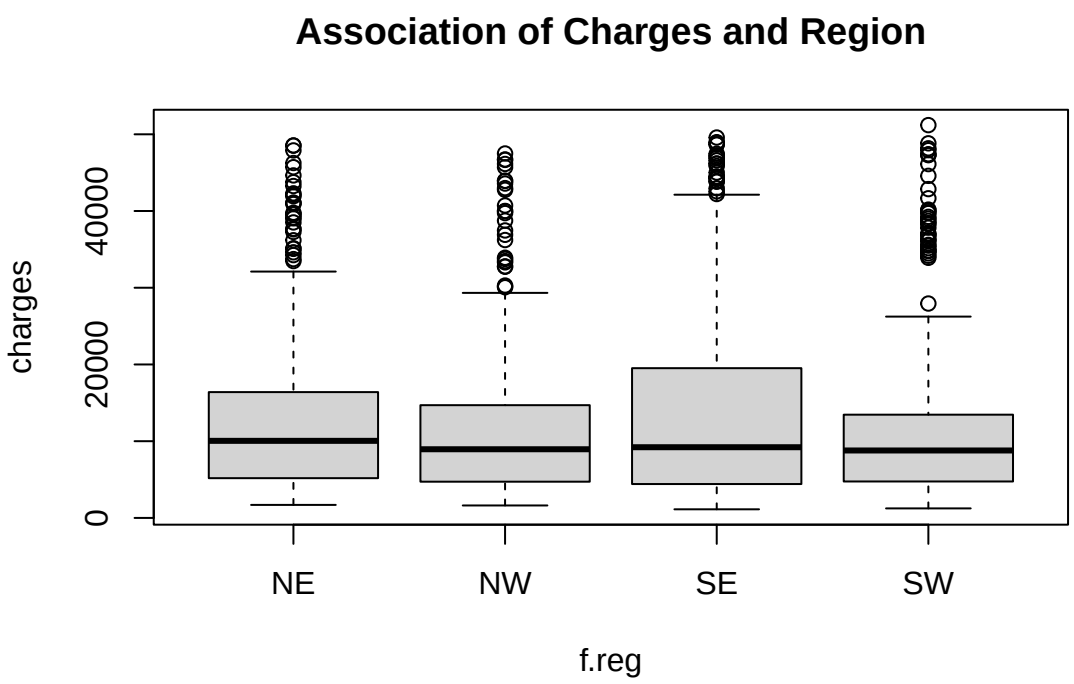
```
res.con$category
```

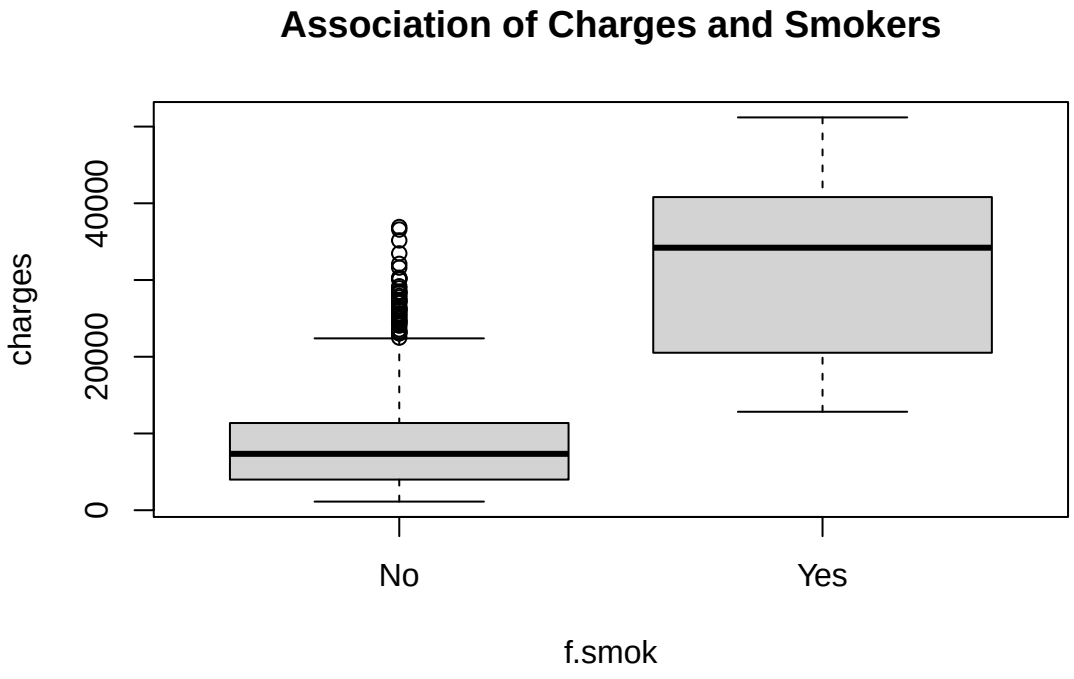
```
##                 Estimate       p.value
## f.smok=Yes    11505.5682 4.650154e-280
## f.age=Old      3842.5818  1.736319e-18
## f.bmi=High     2723.3239  3.435144e-05
## f.reg=SE       1435.8572  8.076659e-03
## f.sex=M         709.5889  2.742457e-02
## f.sex=F        -709.5889  2.742457e-02
## f.bmi=Normal   -598.0124  1.743093e-04
## f.age=Young   -3994.8994  2.095461e-20
## f.smok=No    -11505.5682 4.650154e-280
```

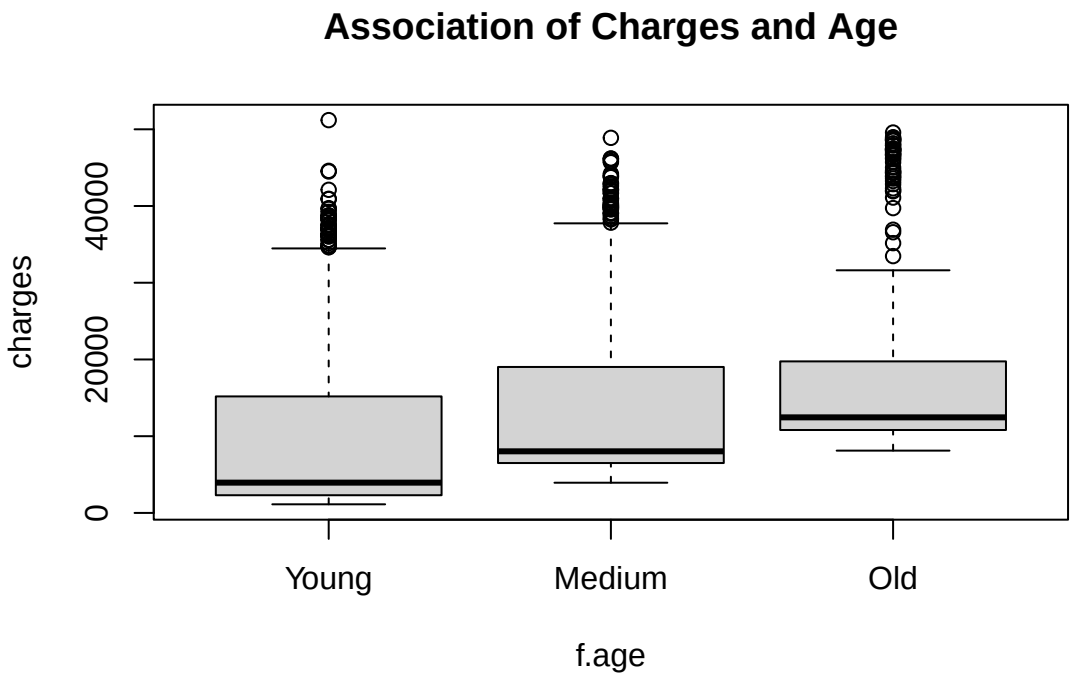
By interpreting the results, one can understand that the most influential categories that lead to the maximization of the values of the target variable is people who are smokers, following by old people and people with high BMI values. On the other hand, exactly the opposite is true for non smokers and young people.

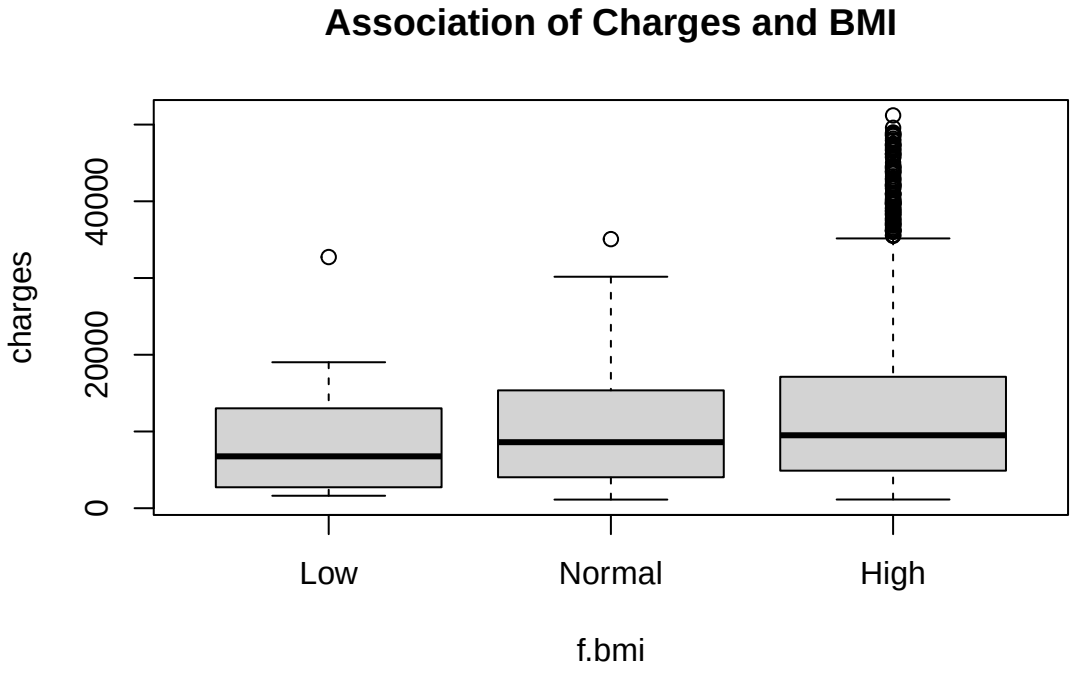
With the following graphs, a visual result of the above conclusions is presented.




Association of Charges and Region
charges
0
20000
40000
NE
NW
SE
SW
f.reg


Association of Charges and Smokers
charges
0
20000
40000
No
Yes
f.smok





Additionally, association tests on mean and variance are completed below between the target variable and the different explanatory variables, based on their type. It is important to mention here, that only non-parametric tests were completed due to the fact that charges do not follow a normal distribution. Firstly, Kruskal test is used for checking the mean of the target variable between the different groups of region, age and BMI multi-level factors.

```
kruskal.test(charges~f.age,data=df)
```

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  charges by f.age
## Kruskal-Wallis chi-squared = 333.83, df = 2, p-value < 2.2e-16
```

```
kruskal.test(charges~f.reg,data=df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  charges by f.reg
## Kruskal-Wallis chi-squared = 4.7239, df = 3, p-value = 0.1932
```

```
kruskal.test(charges~f.bmi,data=df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  charges by f.bmi
## Kruskal-Wallis chi-squared = 7.9869, df = 2, p-value = 0.01844
```

Kruskal test on mean for the target variable base on the different groups of age showed that the p-value is extremely small, so null hypothesis can be rejected, thus it can be said that the different groups of age have different mean of the variable charges. With the same approach, the same result is derived for categorical variables "f.reg" and "f.bmi", but here the p-values are not very small, in fact they are bigger than 0.01. This means that the difference of the means in the specific sample is not so significant.

To continue with, in order to test the mean difference between the different groups of binary factors the usage of Wilcox test is necessary. The test is completed one for the variable "f.sex" and one for "f.smoke".

```
wilcox.test(charges~f.sex,data=df, correct = TRUE, exact= FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  charges by f.sex
## W = 218631, p-value = 0.6904
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(charges~f.smok,data=df, correct = TRUE, exact= FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  charges by f.smok
## W = 7403, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

It is obvious, that for variable "f.smok" the p-value is extremely low (rejecting null hypothesis), thus the mean of the target variable for smokers and non-smokers have a significant difference, while for variable "f.sex" the p-value is large (0.69), meaning that mean values of charges do not have significant differences for men and women.

Moving on, the variance tests are taking place. For the same reason as for the mean tests, only

non-parametric tests were used. In the case of variance, Flinger test covers both multi-leveled factors and binary ones.

```
fligner.test(charges~f.age,data=df) # Non Parametric
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by f.age
## Fligner-Killeen:med chi-squared = 0.23836, df = 2, p-value = 0.8876
```

```
fligner.test(charges~f.reg,data=df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by f.reg
## Fligner-Killeen:med chi-squared = 18.996, df = 3, p-value = 0.0002739
```

```
fligner.test(charges~f.bmi,data=df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by f.bmi
## Fligner-Killeen:med chi-squared = 6.3264, df = 2, p-value = 0.04229
```

```
fligner.test(charges~f.sex,data=df)
```

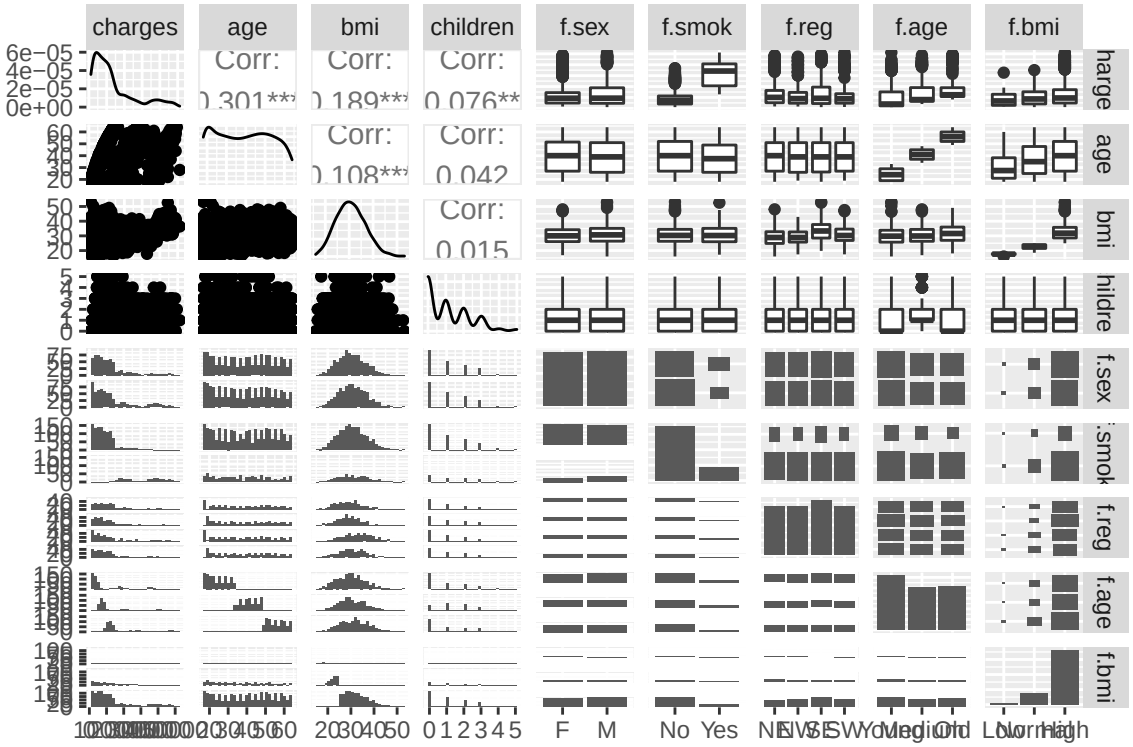
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by f.sex
## Fligner-Killeen:med chi-squared = 10.452, df = 1, p-value = 0.001225
```

```
fligner.test(charges~f.smok,data=df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  charges by f.smok
## Fligner-Killeen:med chi-squared = 223.81, df = 1, p-value < 2.2e-16
```

While Fligner test has as a null hypothesis the fact that variances in each of the groups defined by the levels of the factors are the same, the following results are concluded. For variable "f.age" we have a large p-value (0.89) indicating that the variance of charges in the different age groups is the same. For variable "f.reg" the p-value is less than 0.05 (0.0003) so here one can say that the variance of charges in different regions have a significant difference. The same is true for variable "f.sex" which achieve a p-value equal to 0.001. To continue with, for variable "f.bmi" the p-value of the test is slightly smaller than 0.05 (0.042), which means that null hypothesis is rejected, but due to the fact that the p-value is close to 0.05, it is concluded that the difference on the variance in the different BMI groups is not so significant. On the other hand variable "f.smok" achieves a p-value close to 0 (2.2e-16) showing that the difference of variance for variable charges is totally significant between smokers and non-smokers.

To conclude, all of the above relationships can be presented in total in the following pairplot.



2 Linear Regression Modelling

In this section of the report, the creation and comparison of multiple linear regression models is completed. The approach followed is presented here. As a first step, a linear model is trained addressing the target variable by using only numeric variables. Then, the addition of factors in the model is taking place, as well as the interactions between the different variables of the data set. Between each step, the quality of the models is being tested. The quality assurance of the models include the interpretation of several plots, like residuals plots, influential data plots, allEffects plots, etc.. Additionally, filtering of the influential data and several transformations on the explanatory variables are taking place in order to correct the final prediction quality of the model. The whole procedure is described in details in the next subsections.

2.1 Addressing Target Variable only with Numerical Variables