Part 2: Exercise 7

Load Dataset

For the purpose of the exercise we will use the package *psych*. Psych is a package developed for personality, psychometric and psychology research. It provides useful functions for such analysis and it is a core part of International Cognitive Ability Resource (ICAR) project[1].

Dataset consists of 1338 records and 7 features. The column *charges* is the dependent variable, the other 6 will be used to analyze their impact to total costs.

```
require(psych)
```

```
## Loading required package: psych
```

```
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
require(randomForest)
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:psych':
##
##     outlier
```

```
require(randomForestExplainer)
```

```
## Loading required package: randomForestExplainer
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
df <- read.csv(file = 'insurance.csv')
nrow(df)
```

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

```
ncol(df)
```

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

```
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  
```

Describe dataset

Now we will use the *describe* method provided by psych package. It let us for a more in depth overview of the data by presenting the most frequently used descriptive statistics for psychometric and psychology research. Note the symbol * indicates that the variable is categorical.

```
describe(df)
```

```
##          vars    n     mean       sd  median  trimmed     mad     min      max
## age         1 1338    39.21    14.05   39.00    39.01   17.79   18.00    64.00
## sex*        2 1338     1.51     0.50    2.00     1.51    0.00    1.00     2.00
## bmi         3 1338    30.66     6.10   30.40    30.50    6.20   15.96    53.13
## children    4 1338     1.09     1.21    1.00     0.94    1.48    0.00     5.00
## smoker*     5 1338     1.20     0.40    1.00     1.13    0.00    1.00     2.00
## region*     6 1338     2.52     1.10    3.00     2.52    1.48    1.00     4.00
## charges     7 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87 63770.43
##             range  skew kurtosis     se
## age         46.00  0.06    -1.25   0.38
## sex*         1.00 -0.02    -2.00   0.01
## bmi         37.17  0.28    -0.06   0.17
## children     5.00  0.94     0.19   0.03
## smoker*      1.00  1.46     0.14   0.01
## region*      3.00 -0.04    -1.33   0.03
## charges  62648.55  1.51     1.59 331.07
```

We can see about the *mean*, *standard deviation*, *median*, *trimmed*, *mean absolute deviation*, *min*, *max*, *range*, *skew*, *kurtosis* and *standard error*. Before proceeding to model construction that it could explain/predict the dependent variable(*charges*) we need to define skewness and kyrtosis.

Skewness

Skewness is described as a measure of data symmetry. A perfectly symmetrical data will have a skewness of 0 which might indicate a Normal distribution as the value of skewness for the latter is also 0.

Skewness is defined as:

$$a_3 = \sum \frac{(X_i - \bar{X})^3}{ns} \tag{1}$$

where:

- n is the sample size
- X_i is the i^{th} X value
- $\bar{X}$ is the average
- s is the sample standard deviation

The exponent *3* is referred to the third standardized central moment for the probability model.

Usually, we interpret the value (rule of thumb) as:

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed
- If the skewness is less than -1 or greater than 1, the data are highly skewed

Kyrtosis

Kurtosis is a measure of whether a distribution is narrowly concentrated to the middle; most of the responses are in the center. In other words is a measure of peakedness or flatness of data points.

Kurtosis is defined as:

$$a_4 = \sum \frac{(X_i - \bar{X})^4}{ns} \tag{2}$$

where:

- n is the sample size
- X_i is the i^{th} X value
- $\bar{X}$ is the average
- s is the sample standard deviation

The exponent *4* is referred to the fourth standardized central moment for the probability model.

Analysing the numerical variables of the dataset and the output of *psych.describe* we can see that the variable *bmi* with *skew = 0.28* and *kyrtosis = -0.06* is distributed fairly normally.

Mean Absolute Devation(MAD)

The mean absolute deviation is the average distance between each data point and the mean. It gives us an idea about the *variability* in a dataset. It is defined as:

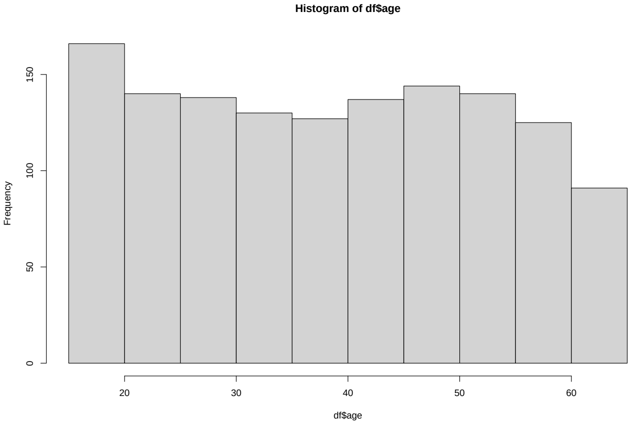
$$mad = \frac{\sum x_i - \bar{x}}{n} \tag{3}$$

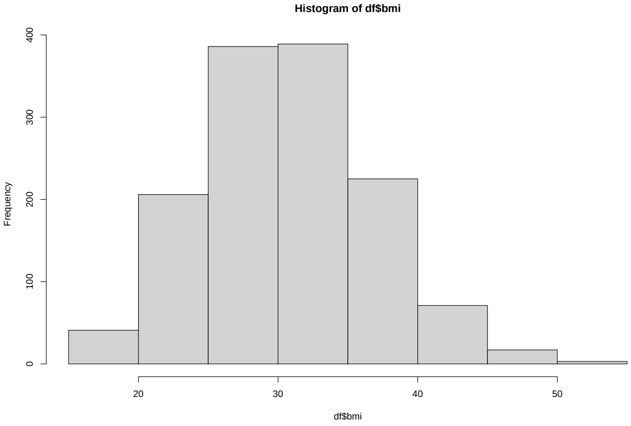
We observe that *age* has some variability *mad* which could also be conlcuded from standard devation.

Histograms

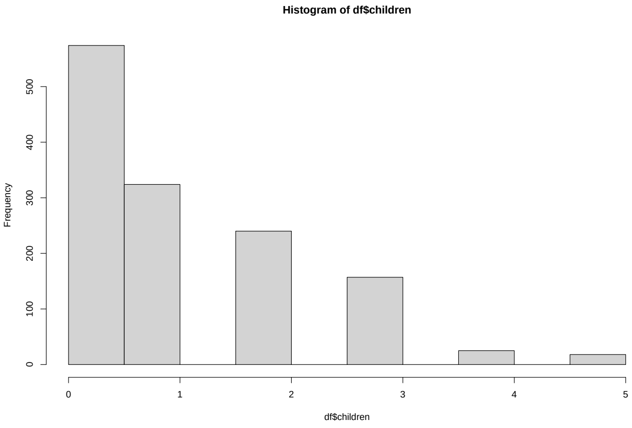
We will plot the histograms to get a better visual understanding of numerical variables of the dataset.

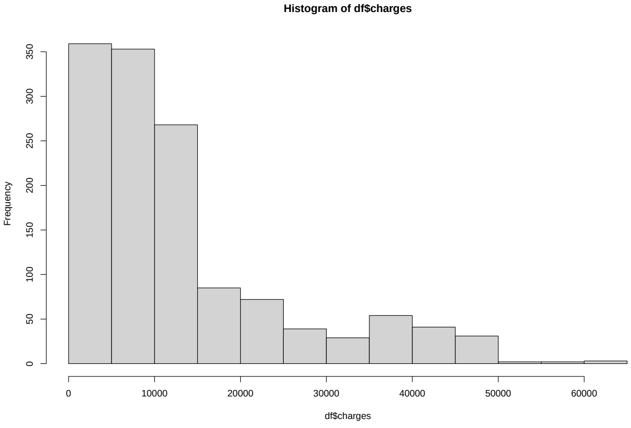
```
hist(df$age)
hist(df$bmi)
```





```
hist(df$children)
hist(df$charges)
```





Variables *children* and *charges* are skewed, *bmi* as hinted above looks to follow a normal distribution and *age* is a uniform distribution.

Model

We will test two models, *multivariate linear regression* and *decision tree*. For the first case we need to transform categorical variables into numerical. We use the command *str(structure)* to see the datatypes the set.

Multivariate Linear Regression

```
str(df)
```

```
## ’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 ...
```

There are 3 features which are categorical. In the snippet above we see the *sex*, *smoker* and *region* have a structure of chr. We need to convert them into *Factors* in order to fit a linear regression model. We call the *as.factor* method.

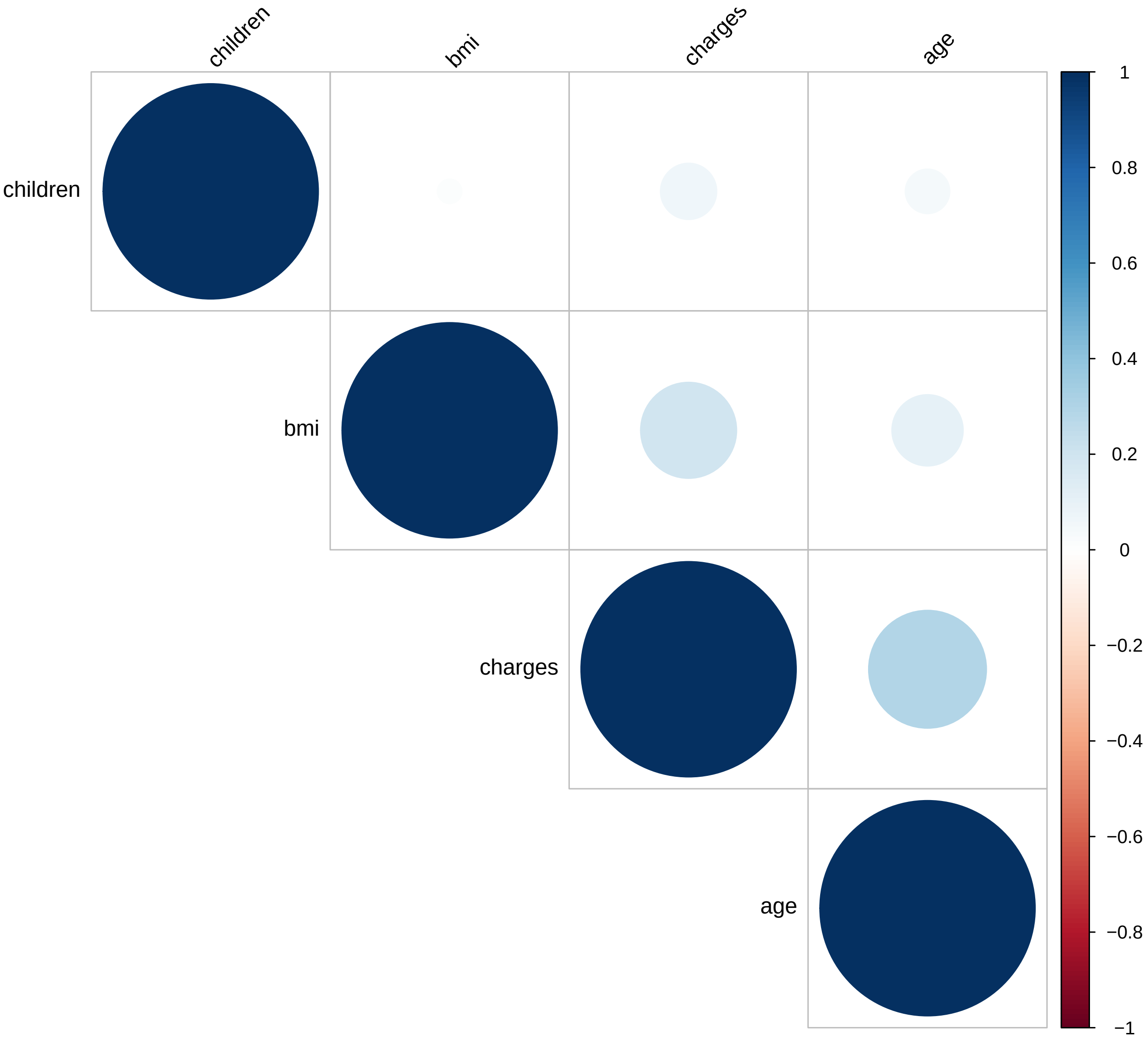
```
df$sex <- as.factor(df$sex)
df$smoker <- as.factor(df$smoker)
df$region <- as.factor(df$region)
str(df)
```

```
## ’data.frame’:    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Correlation matrix

Dependent variable seems to have a fairly strong correlation with the *age*.

```
df_num <- df[c(7,1,3,4)]
res <- cor(df_num)
corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```



Fit model

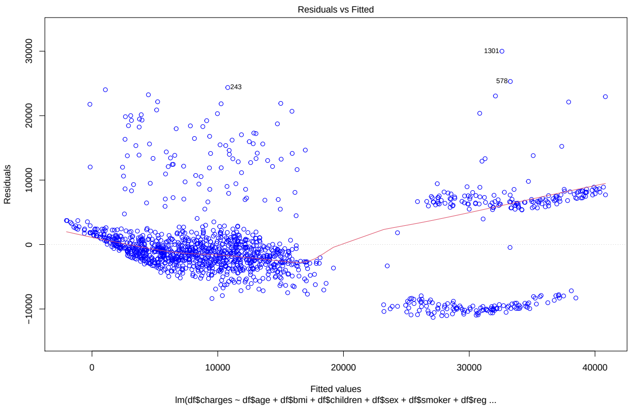
```
linear <- lm(df$charges~df$age + df$bmi + df$children + df$sex + df$smoker + df$region)
summary(linear)
```

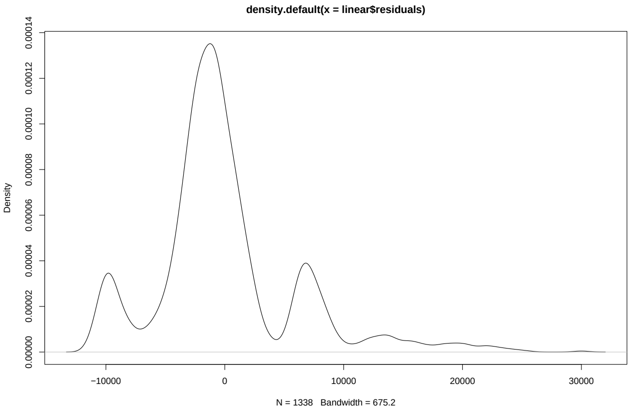
```
##
## Call:
## lm(formula = df$charges ~ df$age + df$bmi + df$children + df$sex +
##     df$smoker + df$region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
```

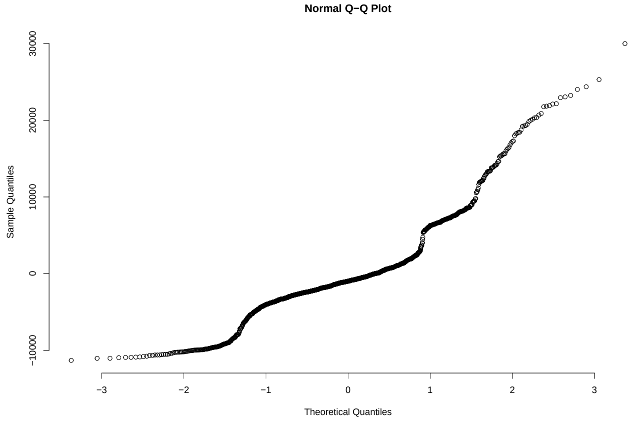
```
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -11938.5      987.8 -12.086  < 2e-16 ***
## df$age                  256.9       11.9  21.587  < 2e-16 ***
## df$bmi                  339.2       28.6  11.860  < 2e-16 ***
## df$children             475.5      137.8   3.451 0.000577 ***
## df$sexmale             -131.3      332.9  -0.394 0.693348    
## df$smokeryes          23848.5      413.1  57.723  < 2e-16 ***
## df$regionnorthwest     -353.0      476.3  -0.741 0.458769    
## df$regionsoutheast    -1035.0      478.7  -2.162 0.030782 *  
## df$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
```

Linear regression mke the assumptions of normality of residuals, homoscedasticity[2] (the variability in the response variable is the same at all levels of the explanatory variable) etc. Thus, we need to analyse the residuals of the model.

```
linear$predicted <- predict(linear)
linear$residuals <- residuals(linear)
plot(linear, which=1, col=c("blue"))
plot(density(linear$residuals))
qqnorm(linear$residuals, pch = 1, frame = FALSE)
```







Even though the multivariate regression model is able to explain the variance in *charges* achieving $R^2 = 0.7509$ and Adjusted $R^2 = 0.7509$ it fails to full fill the assumption of normality of residuals as is shown above.

Decision Trees

In this section we will use a *Random Forest* tree with default hyper-parameters

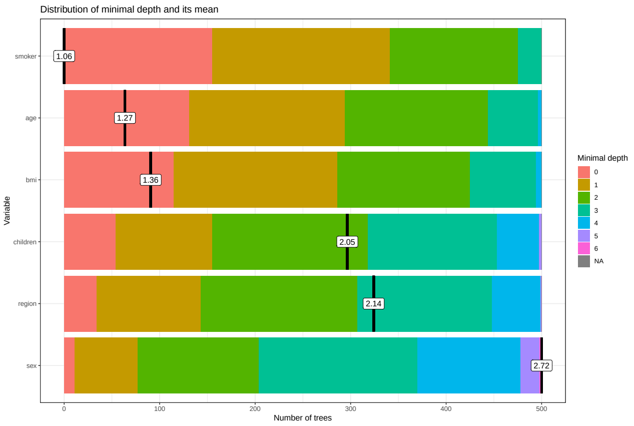
```
require(randomForest)
# Create a Random Forest model with default parameters
rf <- randomForest(charges ~ ., data = df, importance = TRUE)
rf
```

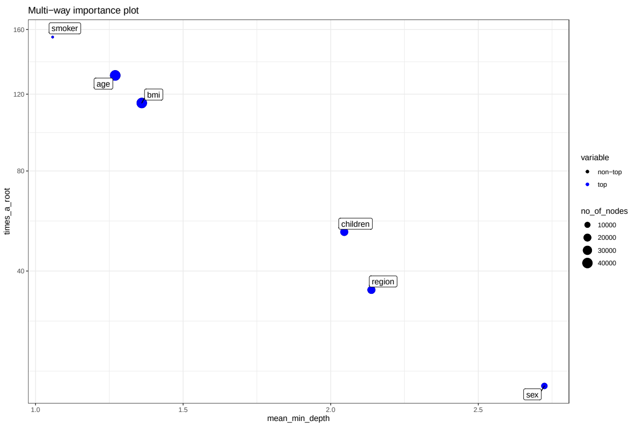
```
##
## Call:
##  randomForest(formula = charges ~ ., data = df, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 22091896
##                     % Var explained: 84.92
```

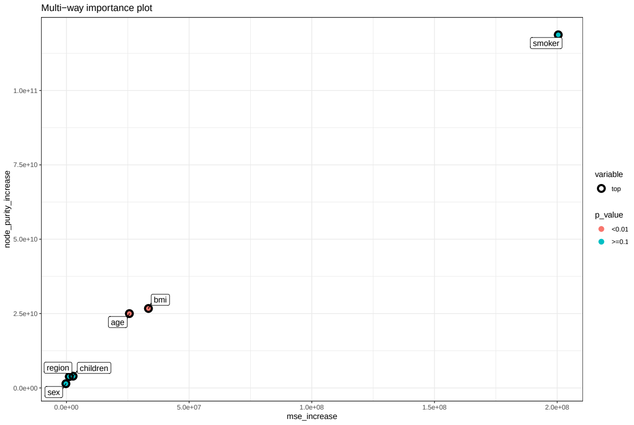
By default, number of trees is 500 and number of variables tried at each split is 2 in this case. 84% of the variance was explained.

```
min_depth_frame <- min_depth_distribution(rf)
plot_min_depth_distribution(min_depth_frame, mean_sample = "relevant_trees", k = 15)
importance_frame <- measure_importance(rf)
plot_multi_way_importance(importance_frame, size_measure = "no_of_nodes")
plot_multi_way_importance(importance_frame, x_measure = "mse_increase", y_measure = "node_purity_increa
```

```
## Warning: Using alpha for a discrete variable is not advised.
```







We see that feature *smoker* has the higher node_purity_increase and mean-squared-error increase(gini index for classification). We can concluded that the variance of *charges* can be explained mainly from the variable *smoker* as it seems to be the most important feature.

References

psych package [@https://personality-project.org/r/psych/].

homoscedasticity [@https://en.wikipedia.org/wiki/Homoscedasticity].