Assignment 2

Dataset: Medical Insurance, Data source : datagov.com

Submitted by,

VARSHINI.S

Submitted to,

Dr.V.Bhuvaneshwari

**Assumption**:

* The dataset explains about the Medical Insurance.
* Insurance can be different for various age groups. Older people are at high risk of suffering from illness than the youngsters.
* For an Non smoker the charges are less and for a smoker the charges are more.
* Based on the region also charges differs for an individual.
* Bmi is the most important thing in insurance,that describes an individual how much to pay for insurance.
* We will be performing EDA to verify these assumptions.

library(lattice)  
library(MASS)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## 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(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.0  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ✔ readr 2.1.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::select() masks MASS::select()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(rmarkdown)  
  
df<-read.csv("expenses.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

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 ...

dim(df)

## [1] 1338 7

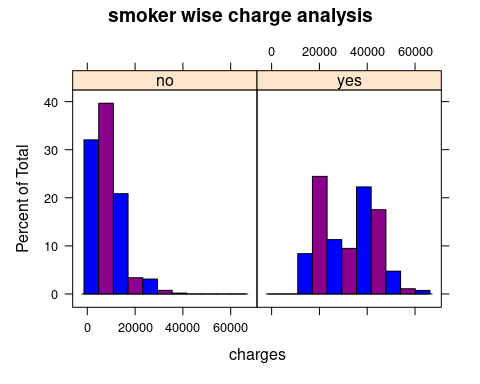
#character to categorical variable  
df$sex=as.factor(df$sex)  
df$smoker=as.factor(df$smoker)  
df$region=as.factor(df$region)  
str(df)

## '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 ...

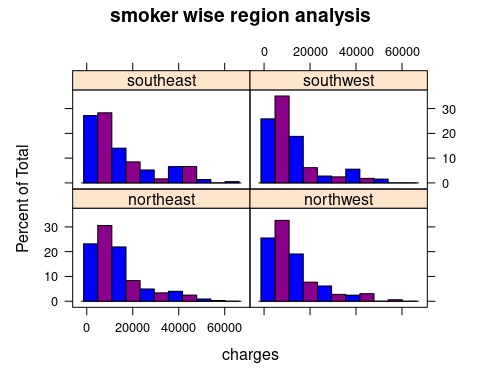
#histogram  
histogram(~age|smoker,data=df,col=c("blue","darkmagenta"),main="smoker wise age analysis")



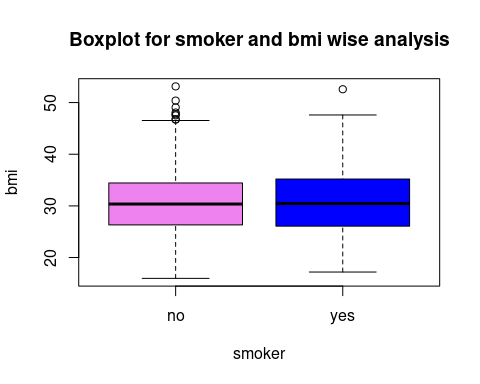
histogram(~charges|smoker,data=df,col=c("blue","darkmagenta"),main="smoker wise charge analysis")



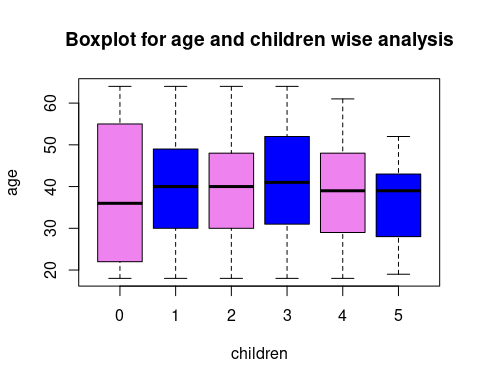
histogram(~charges|region,data=df,col=c("blue","darkmagenta"),main="smoker wise region analysis")



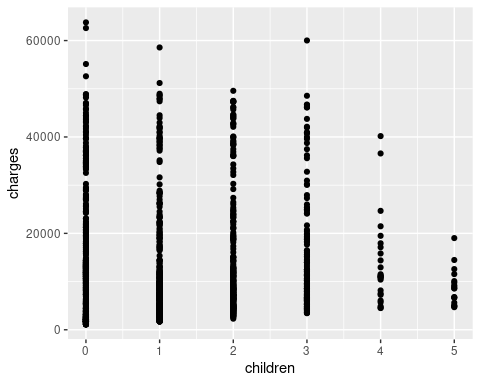
#boxplot  
boxplot(bmi~smoker,data=df,col=c("violet","blue"),main="Boxplot for smoker and bmi wise analysis")



p <-ggplot(df,aes(x=bmi,y=smoker))+  
 geom\_boxplot()  
  
boxplot(age~children,data=df,col=c("violet","blue"),main="Boxplot for age and children wise analysis")



p <-ggplot(df,aes(x=age,y=children))+  
 geom\_boxplot()  
  
#scatterplot  
  
ggplot(df, aes(x = children, y = charges)) +  
 geom\_point()



**Attribute Description:**

* The Medical Insurance dataset have 7variables and 1338 observations.
* The dataset is filled with numerical, integer and character data types.
* The Age group for the insurance starts 18 and ends in 64.
* There are two categories Male and Female.
* BMI is calculated for every individuals, If your BMI is 18.5 to 24.9, it falls within the Healthy Weight range. If your BMI is 25.0 to 29.9, it falls within the overweight range. If your BMI is 30.0 or higher, it falls within the obese range.
* There are both Smoker and Non Smoker.
* There are four different types of Region in the dataset
* Based on these attributes the charges has been calculated for an individual.

**INFERENCE**

Histogram:

* We can start our analysis with the smoker and age attributes .

18-22 age group people smoke more, 22-40 age people smoke often and 40-60 age people are the least smokers.

* For a non smoker the charges they pay are less, Maximum amount they pay is from 0- 20000 and averagely they pay about 20000-40000.
* But for a Smoker the chargers are high, they pay about 20000-60000.
* Moderately every region charges maximum at 0-20000 and it decreases after 20000.
* Southeast is the only region charges 60000

Boxplot:

* We are analyzing smoker and bmi. For both non smoker and smoker the median lies in the center , there is no skewness.
* For Children 0 and 3, there are maximum and minimum values ,the median lies near Q1 so it is positively skewed .
* For Children 1,2, and 4 ,there are maximum and minimum values the median lies in the center so it has no skewness.
* For Children 5 there are maximum and minimum values ,the median lies near Q3 so it is negatively skewed.

Scatterplot:

* X axis is plotted as the Children and Y axis is plotted as the Charges.
* NO CORRELATION because the points are scattered all over the plot it is difficult to conclude whether it is increasing or decreasing.

INSIGHTS

* Youngsters smoke more when compared to the old age people.
* Smoking is one of the major factors that affect your premium. Non Smokers pay less amount whereas Smokers pay large amount.
* Even with the regions the charges amount changes. Southeast region pays larger amount.
* BMI can also affect your premium, if your bmi falls within the range it does not affect your charges ,if it is higher it will surely affect your charges.
* The charges is maximized is you have more children and it is minimized if you have less children.