Coding Assignment 1

Team 8

Due: 2021-09-29 23:59

Table of Contents

# Put any packages you want here  
  
library(readxl)  
library(gt)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(gtsummary)  
library(plotly)

##   
## Attaching package: 'plotly'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following object is masked from 'package:graphics':  
##   
## layout

library(readxl)  
library(plotly)  
library(corrplot)

## corrplot 0.92 loaded

A Florida health insurance company wants to predict annual claims for individual clients. The company pulls a random sample of 50 customers. The owner wishes to charge an actuarially fair premium to ensure a normal rate of return. The owner collects all of their current customer’s health care expenses from the last year and compares them with what is known about each customer’s plan.

The data on the 50 customers in the sample is as follows:

* Charges: Total medical expenses for a particular insurance plan (in dollars)
* Age: Age of the primary beneficiary
* BMI: Primary beneficiary’s body mass index (kg/m2)
* Female: Primary beneficiary’s birth sex (0 = Male, 1 = Female)
* Children: Number of children covered by health insurance plan (includes other dependents as well)
* Smoker: Indicator if primary beneficiary is a smoker (0 = non-smoker, 1 = smoker)
* Cities: Dummy variables for each city with the default being Sanford

Answer the following questions using complete sentences and attach all output, plots, etc. within this report.

**For this assignment, ignore the categorical variables (gender, smoker, cities)**

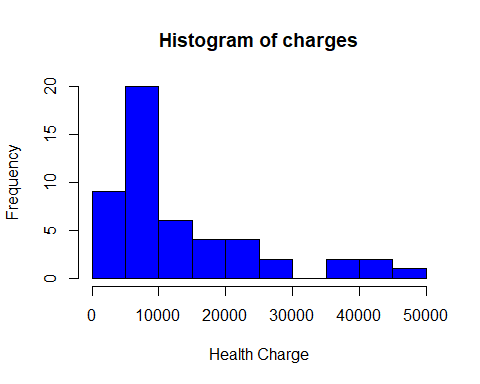
# Question 1

Perform univariate analyses on the quantitative variables (center, shape, spread). Include descriptive statistics, and histograms. Be sure to use terms discussed in class such as bimodal, skewed left, etc.

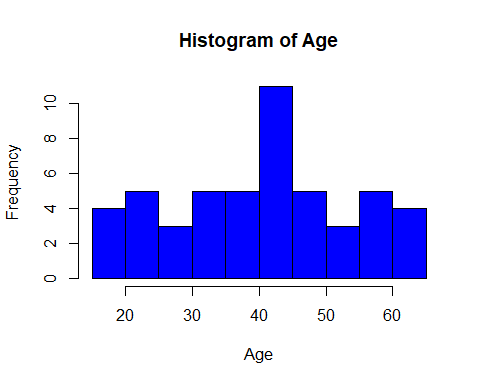
# Center   
  
summary(Health1)

## Charges Age BMI Children   
## Min. : 1640 Min. :19.00 Min. :18.34 Min. :0.0   
## 1st Qu.: 6241 1st Qu.:31.25 1st Qu.:28.33 1st Qu.:0.0   
## Median : 9095 Median :41.00 Median :31.11 Median :1.0   
## Mean :13251 Mean :40.82 Mean :32.55 Mean :1.1   
## 3rd Qu.:17544 3rd Qu.:48.00 3rd Qu.:36.51 3rd Qu.:2.0   
## Max. :47404 Max. :63.00 Max. :49.06 Max. :5.0

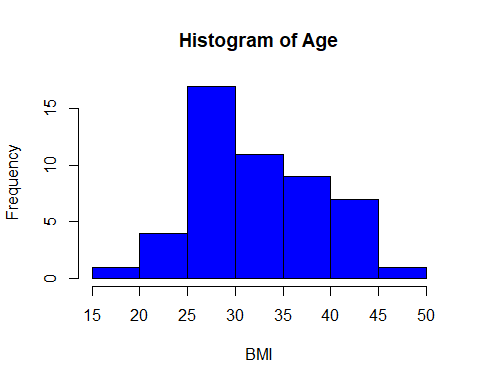
#shape  
  
  
  
  
hist(Health1$Charges , xlab = "Health Charge", ylab = "Frequency",main = "Histogram of charges", col = "blue")



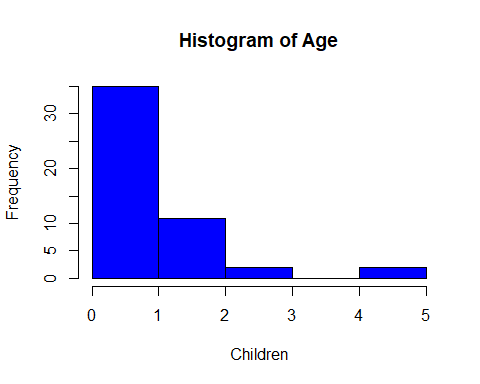
hist(Health1$Age, xlab = "Age", ylab = "Frequency",main = "Histogram of Age", col = "blue")



hist(Health1$BMI, xlab = "BMI", ylab = "Frequency",main = "Histogram of Age", col = "blue")



hist(Health1$Children, xlab = "Children", ylab = "Frequency",main = "Histogram of Age", col = "blue")



#Spread   
  
## variance  
  
var(Health1$Charges,na.rm = TRUE)

## [1] 129892851

var(Health1$Age,na.rm = TRUE)

## [1] 165.0078

var(Health1$BMI,na.rm = TRUE)

## [1] 39.56672

var(Health1$Children,na.rm = TRUE)

## [1] 1.397959

## standard deviation  
  
sd(Health1$Charges,na.rm = TRUE)

## [1] 11397.05

sd(Health1$Age,na.rm = TRUE)

## [1] 12.84553

sd(Health1$BMI,na.rm = TRUE)

## [1] 6.290208

sd(Health1$Children,na.rm = TRUE)

## [1] 1.182353

## IQR  
  
IQR(Health1$Charges,na.rm = TRUE)

## [1] 11303.03

IQR(Health1$Age,na.rm = TRUE)

## [1] 16.75

IQR(Health1$BMI,na.rm = TRUE)

## [1] 8.17625

IQR(Health1$Children,na.rm = TRUE)

## [1] 2

## range  
  
range(Health1$Charges,na.rm = TRUE)

## [1] 1639.563 47403.880

range(Health1$Age,na.rm = TRUE)

## [1] 19 63

range(Health1$BMI,na.rm = TRUE)

## [1] 18.335 49.060

range(Health1$Children,na.rm = TRUE)

## [1] 0 5

# Question 2

Perform bivariate analyses on the quantitative variables (direction, strength and form). Describe the linear association between all variables.

Bivariate Analysis of Charges & Age

This scatterplot shows a weak negative linear association between age of the primary beneficiary and total medical expenses for a particular insurance plan.

Bivariate Analysis of Charges & BMI

This scatterplot shows a weak negative linear association between age of the primary beneficiary and total medical expenses for a particular insurance plan.

Bivariate Analysis of Charges & Children

This scatterplot shows a weak negative linear association between age of the primary beneficiary and total medical expenses for a particular insurance plan

Bivariate Analysis of BMI & Age

This scatterplot shows a weak negative linear association between age of the primary beneficiary and primary beneficiary’s body mass index.

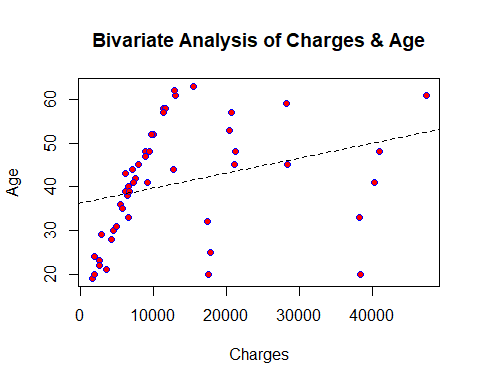
Bivariate Analysis of BMI & Children

This scatterplot shows a weak negative with no linear association between primary beneficiary and number of children covered by health insurance plan and primary beneficiary’s body mass index

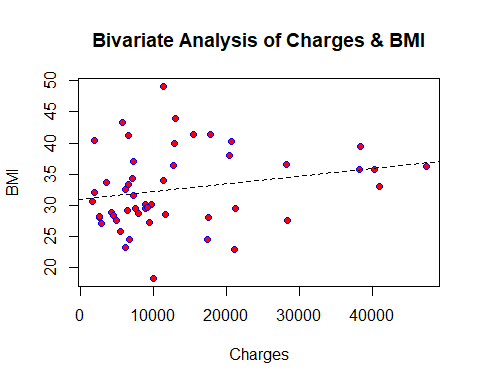
Bivariate Analysis of Age & Children

This scatterplot shows a weak negative with no linear association between age of the primary beneficiary and number of children covered by health insurance plan.

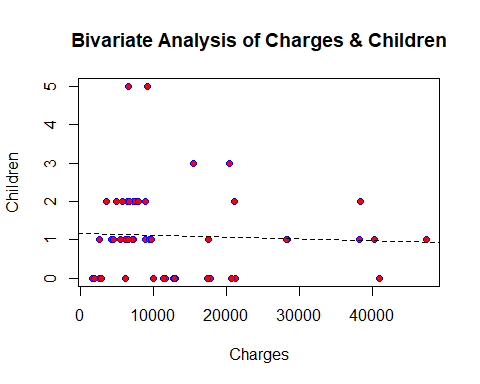
## [1] 0.3045961



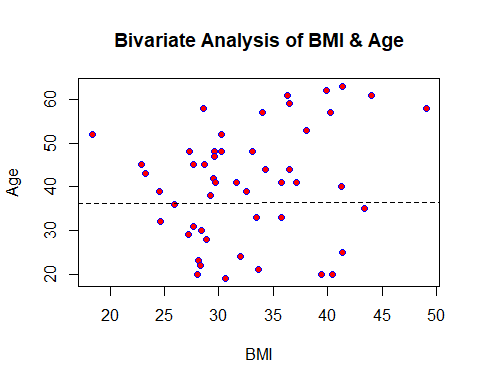
## [1] 0.2246895



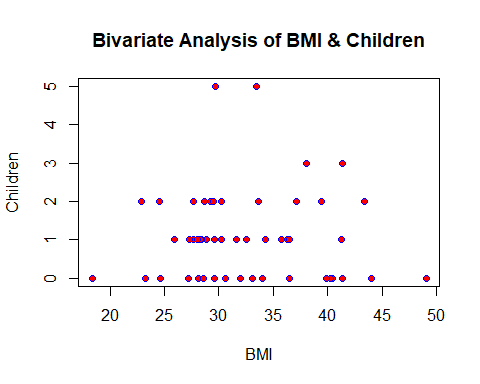
## [1] -0.04641818



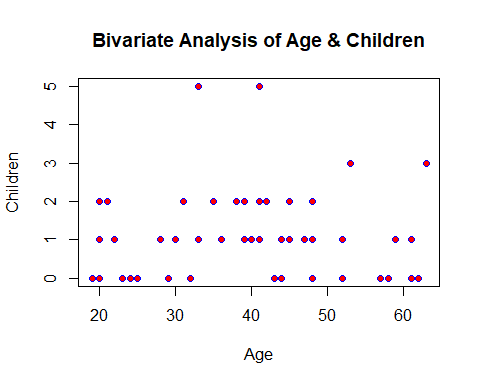
## [1] 0.2192622



## [1] -0.01766343



## [1] -0.04178928



# Question 3

Generate a regression equation in the following form:

ts\_model <- lm(Charges ~ Age + BMI + Children, data = Health1)  
summary(ts\_model)

##   
## Call:  
## lm(formula = Charges ~ Age + BMI + Children, data = Health1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11234 -6071 -4791 2692 28251   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5840.7 9105.8 -0.641 0.5244   
## Age 236.8 125.9 1.882 0.0662 .  
## BMI 300.0 256.9 1.168 0.2488   
## Children -311.7 1334.4 -0.234 0.8163   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11030 on 46 degrees of freedom  
## Multiple R-squared: 0.12, Adjusted R-squared: 0.06262   
## F-statistic: 2.091 on 3 and 46 DF, p-value: 0.1144

also write out the regression cleanly in this document.

Based on the data provided, the regression equation is as follows:

Charges = -5841 + 237Age + 300BMI - 312Children

# Question 4

An eager insurance representative comes back with a potential client. The client is 40, their BMI is 30, and they have one dependent. Using the regression equation above, predict the amount of medical expenses associated with this policy. (Provide a 95% confidence interval as well)

newPrediction <- data.frame(Age = 40 , BMI = 30 , Children = 1)  
predict(ts\_model,  
newdata = newPrediction,  
interval = "confidence",  
level = .95)

## fit lwr upr  
## 1 12321.89 8913.548 15730.23

The result indicates that the insurance representative can expect the value in medical expenses to be about 12,321.89 and that they can be 95 percent confident that the true value lies somewhere between 8,913,55 and 15,730.23.