STUDENTS ON GPA, & NUMBER OF ASSIGNMENTS

Final Individual Project

Division *of* Science Technology and Mathematics

ITC 255 - Statistical Data Analysis

Supervisor: Dr. Asadullah Jawid

Murtaza Malik Aqbal

American University of Afghanistan

December 2022

The individual project is a part of the assignment to pass the ITC 255 (Statistical Data science)

This sample is analyzed through R language.

0: Introduction:

In my dataset, there are two variables, both quantitative (numeric).

- The first quantitative variable in my dataset is the variable *Student GPA*. This variable gives information on student GPA.
- The second quantitative variable in my dataset is the variable *Number of assignments*. This variable gives information about the number of assignments students receive weekly.

I will conduct different numerical and graphical analysis on the variables in my dataset, namely finding their FDT, linear model of regression analysis, and visually representing the results using tools from the ggplot2 package.

1: Linear model and graphical representation

```
setwd("C:/Users/Mortaza/Documents/R")
library("dplyr")
library("ggplot2")
library("plotly")
```

Generating values for each variable using rnorm function:

```
gpa=rnorm(10000,2.8,1.1)
write.csv(m,file='C:/Users/Mortaza/Documents/R/gpa.csv')
assignments=rnorm(10000,3)
write.csv(n,file='C:/Users/ Mortaza /Documents/R/assignments.csv')
```

Importing the CSV file of the generated variables into R:

```
myData=read.csv(file='C:/Users/ Mortaza /Documents/R/M_dataset.csv')
View(myData)
```

Graphing Variable Y (Student's GPA) + FDT:

⇒ I made intervals to make it possible & easier to graphically represent the results.

```
gpa=c()
for (k in 1:length(myData$Y..GPA.)) {
```

```
  if(myData$Y..GPA.[k]<1.5){
    gpa[k]="Low GPA"
  } else if (myData$Y..GPA.[k] >=1.5 & myData$Y..GPA.[k]<=3) {
    gpa[k]="Average GPA"
  } else {
    gpa[k]="High GPA"
  }
}
gpa=as.data.frame(gpa)
```

Making the frequency distribution table (FDT) for Variable Y:

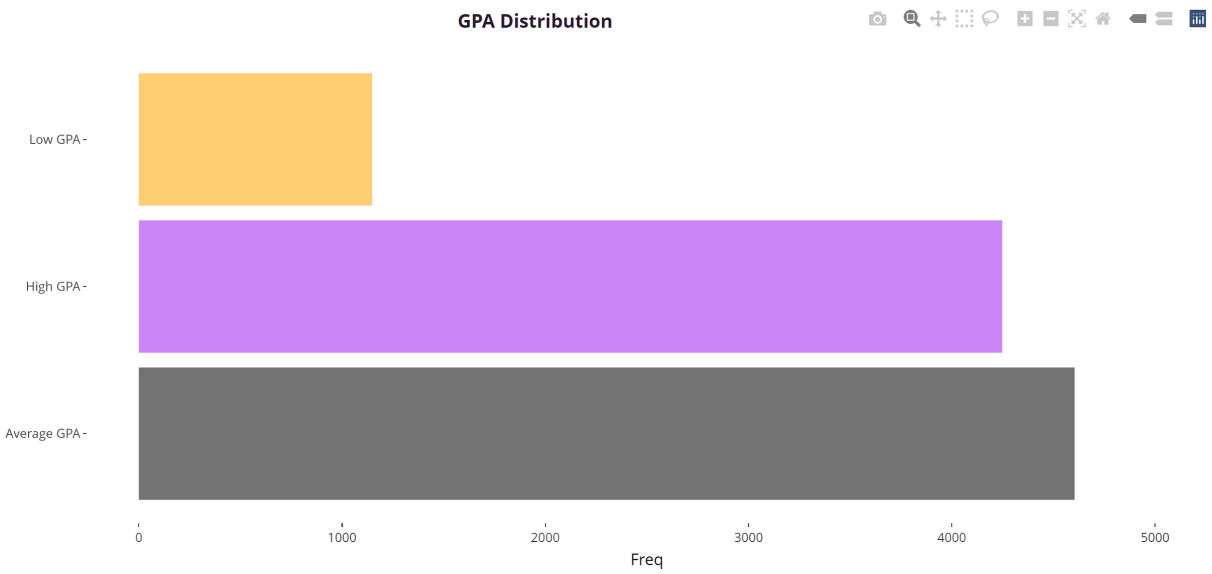
```
fdtFunc=function(x){
  absFreq=table(x)
  relFreq=prop.table(absFreq)
  cumFreq=cumsum(relFreq)
  fdtx=cbind(absFreq, relFreq, cumFreq)
  return(fdtx)
}
fdt=as.data.frame(fdtFunc(gpa))
View(fdt)

names_bar=c("Average GPA","High GPA","Low GPA")
```

Interpretation: I have transformed my numerical variable Y into categorical with 3 categories to make it possible to visually represent it in a bar chart.

Visualizing variable Y graphically:

```
g1=ggplot(data = fdt, aes(x = fdt$absFreq, y = names_bar, fill = names_bar, alpha = 0.5)) +
  geom_bar(stat = "identity")+
  scale_fill_manual(values=c('black', 'purple', 'orange'))+
  labs(x = "Freq", y = "") +
  theme(panel.background = element_blank(),
        legend.position = "none")+
  ggtitle("GPA Distribution")+
  theme(plot.title = element_text(colour = "#301934",
                                  size = 13,
                                  face = "bold",
                                  hjust = .4))+
  xlim(0,5000)
g1
g2=plotly::ggplotly(g1)
htmlwidgets::saveWidget(g2,
                        file = "Mortaza's Graph 1.html")
```



Interpretation: The majority of students (4,631 from 10,000) have a GPA between 1.5 to 3 (Average GPA=1.5-3). The minority being less than 15% of students (1,211) having a GPA below 1.5.

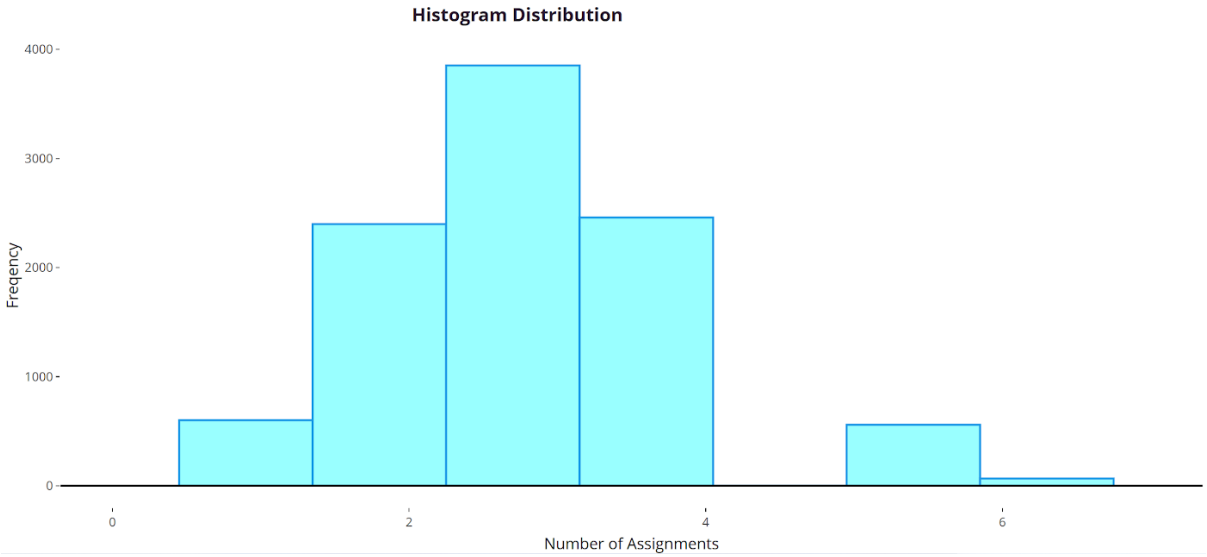
Making the frequency distribution table (FDT) for Variable X:

```
fdt2=as.data.frame(fdtFunc(myData$X....of.assignments.))
View(fdt2)
```

Visualizing variable X graphically:

```
g3=ggplot(data = myData, aes(x = myData$X....of.assignments.)) +
  geom_histogram(fill='#99FFFF',colour=4, binwidth=0.9)+
  theme(panel.background = element_blank())+
  labs(x = "Number of Assignments", y = "Freqency")+
  xlim(0,7)+
  ggtitle("Histogram Distribution")+
  theme(plot.title = element_text(colour = "#301934",
                                  size = 13,
                                  face = "bold",
                                  hjust = .4))+
  geom_hline(yintercept = 0)

g4=plotly::ggplotly(g3)
htmlwidgets::saveWidget(g4,
                        file = "Mortaza's Graph 2.html")
```



Interpretation: On weekly basis, most students (3,879) recieve 3 assignments. Less than 1% of students receive six assignments per week. Over 90% of all assignments per week are between 2 to 4.

2: Regression Analysis

GPA will Y=myData$Y..GPA.
X=myData$X....of.assignments.
regression_analysis=lm(Y~X)
summary(regression_analysis)
#2.834720 -- -0.008832 coefficient – p-value= 0.4
ahat= 2.83
bhat= -0.009
#Y=a+b*X --> Y=2.8-0.009*X

Interpretation: b has a negative sign, therefore, there is an indirect relationship between X (Student's GPA) & Y (Number of weekly assignments), meaning as the number of weekly assignments increases, student GPA decreases. Also, P-value is above 0.05 (Alpha), meaning there is weak association between variables X & Y.

Using Hypothesis testing to check the validity of claims under different scenarios for the entire population (Two quantitative variables)

Y= Student GPA
X= Weekly number of assignments
Alpha= 0.05

p-value= 0.4

Scenario 1:

H0= b=0 (there is no effect of X on Y)
H1= b>1 (there is a direct effect of X on Y)

p-value= 0.4>0.05

Interpretation: Since the P-value of X is greater than 0.05, meaning there is a high chance of making a type 1 error by rejecting H0, therefore we we **do not** reject H0 in favor of H1. However, this does not indicate that we should reject H1. Since our b value is a very small negative, there is a really weak indirect relationship between the given variables.

What will be the GPA is there be 10 assignments per week?
-> Y=2.8-0.009*10=2.71
Interpretation: GPA will be 2.71 if there are 10 assignments per week.

Conclusion: All my numerical and graphical analysis resulted in statistics, such collecting data on variables, managing them, presenting them graphically, to analysisng and interpreting them. the information in my dataset is a sample collected from a population on their GPA and weekly number of assignments. All of the data above, have been analysied using the R coding language.