**NAME: SANDEEP PATIL**

**Email: saashasandeep@gmail.com**

**DATA ANALYTICS WITH R, EXCEL and TABLEAU**

**Session 11 – Assignment – 11.1**

#1. Use the given link and locate the bank marketing dataset. Data Set Link

#Perform the below operations:

#a. Create a visual for representing missing values in the dataset.

getwd()

setwd("D:/BIG DATA/DATA ANALYTICS WITH R, EXCEL & TABLEAU/11 LINEAR MODELS")

getwd()

dm = read.csv("bank-additional.csv", sep = ";")

View(dm)

sum(is.na(dm))

dm$default[dm$default=="no"]<-NA

View(dm)

sum(is.na(dm))

library(Amelia)

missmap(dm)

#b. Show a distribution of clients based on a Job.

#\*\*\*\*NOTE\*\*\*\*

#since in dataset I'm unable to find variable clients therefore i am using

#another variable say age for

#showing you distribution of a age based on job

#Set a different color for each group

bank.additional = read.csv("bank-additional.csv", sep = ";")

library(ggplot2)

ggplot(bank.additional, aes(x=job, y=age, fill=job)) +

geom\_boxplot(alpha=0.3) +

theme(legend.position="none")+

ggtitle("Distribution of age based on a Job")

#OR

ggplot(bank.additional,aes(x="",y=bank.additional$age,fill=bank.additional$job))+geom\_bar(width=1,stat="identity")+coord\_polar("y",start=0)

#c. Check whether is there any relation between Job and Marital Status?

#we are using Chi-Square Test for checking relation as both job and marital status are categorical variables

#so first defining the null hypothesis

#Ho: There is no relation between job and marital status

#Ha: There is relation between job and marital status

chisq.test(bank.additional$job ,bank.additional$marital)

#now as we can see p value is nearly 0 or less which is henceforth less than 0.05

#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis

#which says that There is relation between job and marital status

#also another way to check relation and more example

#by

#Pearson correlation test

#i'm using variable job and education in this

#Correlation test between job and education variables:

newbank = bank.additional

newbank$job <-as.numeric(newbank$job)

newbank$education <-as.numeric(newbank$education)

result <- cor.test(newbank$job, newbank$education, method = "pearson")

result

#In the result above :

#t is the t-test statistic value (t = 8.7235),

#df is the degrees of freedom (df= 4117),

#p-value is the significance level of the t-test (p-value nearly 0).

#conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [0.1046068, 0.1645802]);

#sample estimates is the correlation coefficient (Cor.coeff = 0.1347169).

#The p-value of the test is nearly 0, which is less than the significance level alpha = 0.05.

#We can conclude that job and education are significantly correlated with a correlation coefficient of 0.1347169 and p-value of 2.210^{-16}

#d. Check whether is there any association between Job and Education?

#we are using Chi-Square Test for checking association as both job and education are categorical variables

#hence Chi-Square Test for checking association

#so first defining the null hypothesis

#Ho: There is no association between job and education

#Ha: There is association between job and education

chisq.test(bank.additional$job ,bank.additional$education)

#now as we can see p value is nearly 0 or less which is henceforth less than 0.05

#p value<0.05 hence we will reject the null hypo and accept the alternative hypothesis

#which says that There is association between job and education