**PROJECT ONE**

**Remove all variables from the environment.**

rm(list=ls())

**#1.2 Set the working directory according to where you want your files to be saved.**

setwd("C:/PROJECT1")

getwd()

**The purpose of this question is to generate 200 observations, each with the first 6 numbers of a person’s ID (i.e. YYMMDD, e.g. 980331). The person who needs the dataset is only interested in people whose birthdays are in months that contain 31 days, i.e. Jan, Mar, May, Jul, Aug, Oct, and Dec.**

**For the year:**

**Create a variable year\_1 that contains all the years from 1926 to 2025. (1) Using year\_1, extract only the 3rd and 4th value. Name this year\_2. (1) Change the numbers in year\_2 to characters and save in year.**

year\_1 <- 1926:2025

year\_2 <- year\_1[3:4]

year <- as.character(year\_2)

**For the month:**

**Create a vector month, containing strings “01”, “03”, “05”, etc for the months containing 31 days.**

month <- c("01", "03", "05", "07", "08", "10", "12")

**For the day:**

**Create a vector day\_1, containing strings “01”,”02”, …”09” . (1) Create a vector day\_2, that contains the values from 10 to 31, saving them as characters. (1) Create a vector day by combining day\_1 and day\_2.**

day\_1 <- sprintf("%02d", 1:9)

day\_2 <- as.character(10:31)

day <- c(day\_1, day\_2)

**Generate 200 observations. The first six numbers of a person’s ID should be saved in a vector id. The first person’s id should be in position 1 in the vector, etc.**

**Use the following method: Create yr\_index, which should be a random integer number between (inclusive) 1 and 100.**

**Create mnth\_index, which should be a random integer number between (inclusive) 1 and 7. Create day\_index, which should be a random integer number between (inclusive) 1 and 31.**

**Use these indices to select the appropriate year, month and day for this person. e.g. if yr\_index=1, mnth\_index=3 and day\_index=12, then the person’s birthdate would be 1926-05-12 and the first 6 characters of his ID would be 260512. Use a loop to ensure that 200 observations are created**

set.seed(123)

id <- character(200)

for(i in 1:200)

{

yr\_index <- sample(1:100,1)

mnth\_index <- sample(1:7,1)

day\_index <- sample(1:31,1)

years <- substr(as.character(year\_1[yr\_index]),3,4)

months <- month[mnth\_index]

days <- day[day\_index]

id[i] <- paste0(years,months,days)

}

**#This is just to print out the first six id strings**

head(id)

person\_id <- print(id)

**#This prints out "260512"**

y <- paste0(substr(year[1],3,4),month[3],day[12])

y

**Create a vector gender, that consists of 80 males and 120 females. (Use “male” and “female”). USE ONLY ONE LINE OF CODE!.Create a vector educ, that consists of 50 grade 12s, 20 graduates, 10 post graduates, followed by another 70 grade 12s, 30 graduates, and 20 post graduates. (Use “gr12”, “grad” and “postgrad”).**

gender <- c(rep("males",80),rep("females",120))

gender

educ <- c(rep("gr12", 50), rep("grad", 20), rep("postgrad", 10), rep("gr12", 70), rep("grad", 30), rep("postgrad", 20))

educ

**Create a vector salary\_1 using the random normal variable generator. Round salary\_1 to integer numbers and save the integer numbers in the vector salary. Set the seed to 327.**

set.seed(327)

salary\_1 <- c(

rnorm(50, mean = 8000, sd = 500),

rnorm(20, mean = 15000, sd = 1000),

rnorm(10, mean = 30000, sd = 2000),

rnorm(70, mean = 8000, sd = 500),

rnorm(30, mean = 18000, sd = 1000),

rnorm(20, mean = 33000, sd = 2000)

)

salary <- round(salary\_1)

salary

**Create a dataframe mydata from the vectors created above. It should contain id, gender, educ, and salary.**

mydata <- data.frame(person\_id=id, gender = gender, educ = educ, salary = salary)

mydata

**#1.6**

setwd("C:/Assignment1")

getwd()

data1 <- dget("mydata\_given.txt")

**#How to read csv file into R**

**# data1 <- read.csv("mydata.csv",header= TRUE)**

**# data1**

**#QUESTION2**

**Attach the data1 dataset. Calculate following summary statistics:**

**• The mean salary • The standard deviation of salary • The 85th percentile of salary • Use one line of code to calculate all the following statistics of salary at once: Minimum, 1st quartile, median, mean, 3rd quartile, and maximum.**

attach(data1)

mean\_sal <- mean(salary)

sd\_sal <- sd(salary)

Q\_085 <- quantile(salary,0.85)

SM <- summary(salary)

**Compare the median salary for the different genders. (1) 2.4 Create 3 vectors: Salary\_gr12: salary of people with grade 12 Salary\_grad: salary of people with a degree Salary\_postgrad: salary of people with a higher degree.**

f <- tapply(salary,gender, median)

Salary\_gr12 <- salary[educ == "gr12"]

Salary\_grad <- salary[educ == "grad"]

Salary\_postgrad <- salary[educ == "postgrad"]

**Calculate the mean salary for each of the three educational groups. Round the answers so they are integer values.**

mu\_gr12 <- round(mean(Salary\_gr12))

mu\_grad <- round(mean(Salary\_grad))

mu\_postgrad <- round(mean(Salary\_postgrad))

**#GRAPHS**

**Create a barplot of educ. The graduates should be indicated with light blue, the postgraduates with light green, and the grade 12s with light yellow. Once you have created the graph, ensure that the colours are correctly associated! Add a title, horizontal label and vertical label.**

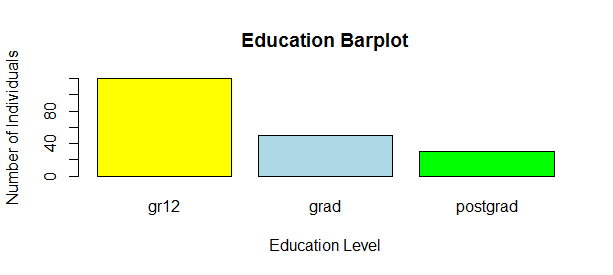
attach(data1)

barplot(table(educ),col=c("Yellow","lightblue","green"),

xlab = "Education Level",

ylab = "Number of Individuals",

main="Education Barplot")



**Create a histogram of salary. Add a horizontal label “Salary”. The colour should be blue. Add a red kernel density overlay and a green normal overlay. State whether the distribution is normal or not.**

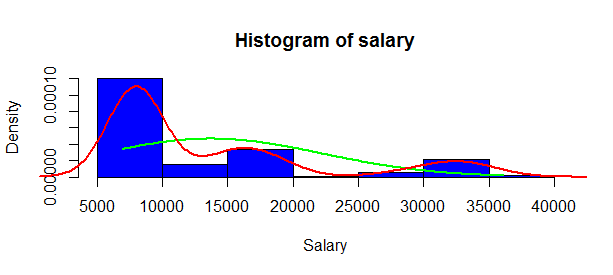
hist(salary,freq=FALSE, col="blue",xlab="Salary")

x.grid <- seq(min(salary),max(salary),length=200)

lines (x.grid,dnorm(x.grid,mean(salary),sd(salary)),col="green",lwd=2)

den <- density(salary)

lines(den,col="red",lwd=2)



**The distribution of the following graph above is not normal.**

**Create a 2x2 matrix with the following plots:**

**• A horizontal barchart of gender. • A boxplot of salary by gender. • A stacked barplot indicating the number of males and females per educational group. • A boxplot of salary by educ. Now, switch back to one graph per page.**

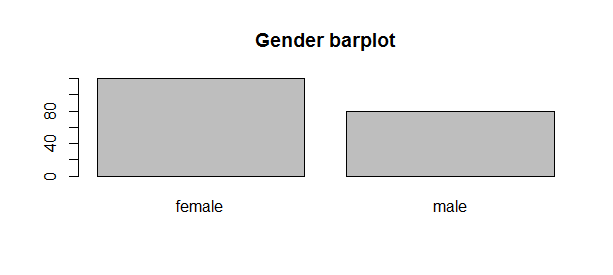
graphics.off()

plot.new()

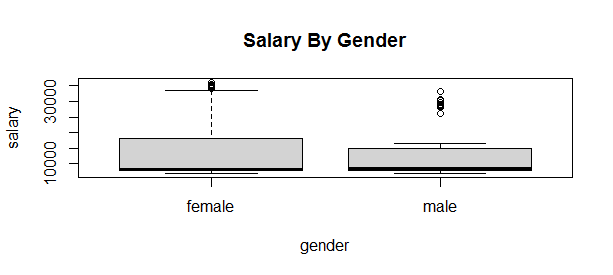
par(mfcol= c(2,2))

genders <- table(gender)

barplot(genders,main="Gender barplot")

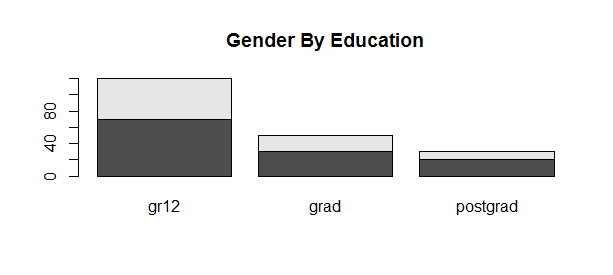


boxplot(salary ~ gender,main="Salary By Gender")

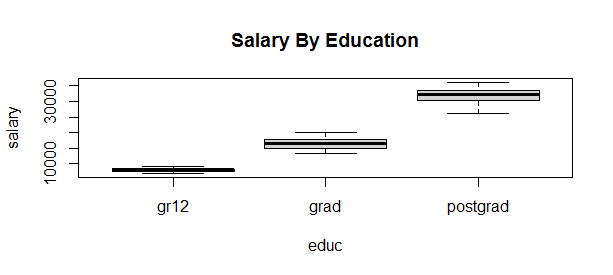


educ\_Gender <- table(gender,educ)

barplot(educ\_Gender,main="Gender By Education")



boxplot(salary ~ educ,main="Salary By Education")

****

**Use ggplot and create a histogram of salary with a kernel density overlay. Ensure that the histogram is light blue with blue lines and the density overlay is red with a line width of 1.0. Interpret the output.**

install.packages("ggplot2")

library(ggplot2)

df <- data.frame(salary = salary)

ggplot(df, aes(x = salary)) + geom\_histogram(aes(y = ..density..), fill = "lightblue" , color = "blue", bins = 20) +

geom\_density(color = "red", size = 1.0) + labs(title = "Salary Distribution with Kernel Density",x = "Salary",y = "Density")

**#The histogram shows that salary variable is not normally distributed, as the data is skewed to the right.**

**Use ggplot and create violin plot of salary versus educational group. Ensure that the colours differ according to gender. Add the title “Salary per education”. Interpret the output.**

library(ggplot2)

df <- data.frame(salary = salary, educ = educ,gender = gender)

ggplot(df, aes(x = educ, y = salary, fill = gender)) +

geom\_violin(trim = FALSE, scale = "width") +

labs(title = "Salary per Education",

x = "Education Level",

y = "Salary") +

scale\_fill\_manual(values = c("Male" = "blue", "Female" = "pink")) +

theme\_classic()

**#The gr12 group for both males and females do not show much variation, salaries are the same around the 8000 salary level.**

**#The graduate group, females have a wider spread around the 8000-20000 level,**

**#therefore females earn more than males at that salary level.**

**#The postgraduate group shows that there is a wider spread for male group around 30 000 level,**

**#so males are earning more in postgraduate roles.**