```
#STUDENT NAME - SHRADDHA GANESH
#TAKE-HOME MID-TERM ASSIGNMENT
#DATE OF SUBMISSION - 02-09-2023
#PROF - SUCHIKA CHOPRA

library (wooldridge)

#QUESTION 1

data("bwght")
?bwght
view(bwght)

#1.
sum(bwght$male==0) #No of mothers.Ans-665
round(proportions(table(bwght$cigs==0)),2) #ANS - TURE VALUE = 85%.

#2.
round(mean(bwght$cigs),2) #Average cigarettes smoked per day.Ans-2.09.

#3.
round(mean(bwght$cigs>0),2) #Average cigarettes smoked among pregnant woman who
#smoke. Ans-0.15

#4.
round(mean(bwght$fatheduc,na.rm = T),2) #Mean of fatheduc.Ans-13.19.
summary(bwght$fatheduc=="NA's") #ANS= FALSE VALUE - (1192 observations)
#Taken to compute the average.Total observations (1388) - No of NA's (196).

#Reasoning - By using "na.rm=T", R removes the missing values and calculates the
#mean value with the remaining observations. Therefore, the number of
#observations taken to calculate the mean in this questions is 1192.
#Total observations (1388) - No of NA's (196). Without using "na.rm=F", R
#calculates the mean value with the missing values. In this case, the mean value
#comes out as "NA" due to the missing values included in the calculations.

#5.
round(mean(bwght$faminc),2)#Ans-29.03
round(sd(bwght$faminc),2)#Ans-18.74

#6.
library(tidyverse)
library(ggplot2)
library(ggthemes)

ggplot(bwght,aes(x=cigs,y=bwght))+
  geom_point()+
  geom_smooth(method="lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Effects of smoking on birth weight",
       x = "No of Cigarettes smoked per day",
       y = "Birth Weight (ounces)")+
  theme(plot.title = element_text(hjust=0.5))

#There is a negative linear correlation between smoking
#and birth weight. Therefore, as the number of cigarettes smoked
#per day increases, the birth weight decreases.

#7.
round(cor(x=bwght$lfaminc[bwght$cigs>0],y=bwght$bwght[bwght$cigs>0]),2)
#Ans-0.11
round(cor(x=bwght$lfaminc[bwght$cigs==0],y=bwght$bwght[bwght$cigs==0]),2)
```

```
#Ans-0.08

#For both, mother who smoke and for those who don't, there is a weak positive
#correlation between birth weight and log of family income. Therefore, as
#family income increases, body weight increases too. However, while both of
#them have a weak positive correlation, non-smoking mothers have a weaker
#positive correlation than smoking mothers do.

#QUESTION 2

?meap01
view(meap01)

#1.
min(meap01$read4)#Ans-0
max(meap01$read4)#Ans-100

#No, the difference/range of this variable does not make sense.Because it is
#calculated using only two values that are extreme. Therefore, it does
#represent the data set accurately.

#2.
sum(meap01$read4==100)#Number of schools with perfect pass rate on the reading
#test. Ans-6
proportions(table(meap01$read4==100))#Ans-True value - which is close to 0%.
sum(meap01$read4==50)#Number of schools with exactly 50% pass rate on the
#reading test.

#3.
mean1 <- mean(meap01$math4)#average pass rate in math.
mean2 <- mean(meap01$read4)#average pass rate in reading.
mean1 #Ans-71.909
mean2 #Ans-60.06188
#The reading test is harder to pass because it has a relatively lower mean pass
#rate than math.

plot1 <- ggplot(meap01,aes(x=math4,y=mean1))+
  geom_boxplot()+
  labs(title="Mean Math Pass Rate",
       x = "Mean Value",
       y ="Math")+
  coord_flip()+
  theme(plot.title = element_text(hjust=0.5))
plot2 <- ggplot(meap01,aes(x=read4,y=mean2))+
  geom_boxplot()+
  labs(title="Mean Reading Pass Rate",
       x = "Mean Value",
       y = "Reading")+
  coord_flip()+
  theme(plot.title = element_text(hjust=0.5))

library(patchwork)
plot1 | plot2

#4.
mean(meap01$exppp)#Ans-5194.865
sd(meap01$exppp)#Ans-1091.89
#Yes, there is a wide variation in per pupil spending. The standard deviation is
#approximately 21% of the mean value (coefficient of variance) which is quite
#large. Most of the data points fall within the interval 4102.975
#(mean - standard deviation) - 6285.755 (mean + standard deviation).

#5.

#a.
```

```r
ggplot(meap01,aes(x=math4, y=read4))+
  geom_point()+
  geom_smooth(method = "lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Correlation between Math and Readings",
       x = "Math Scores",
       y = "Reading Scores")+
  theme(plot.title = element_text(hjust=0.5))

#Yes, the students who perform well on math tend to perform well
#on reading too. From the scatter plot, we can see that there is a
#positive linear relationship between math scores and reading scores.
#Therefore, as x (math score) increases, y also increases (reading score).

cor(x = meap01$math4, y = meap01$read4,method="pearson")#Ans-0.8427281

#STRENGTH & DIRECTION - There is a VERY STONG POSITIVE correlation between
#math scores and reading scores. Strong because the value is almost close to 1,
#and positive because the value is positive and not negative. Therefore, as
#x increases, y also increases.

#b.

?scale_color_distiller
ggplot(meap01, aes(x=math4,y=read4,color=lunch))+
  geom_point()+
  geom_smooth(method ="lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Relationship between Math and Reasing scores",
       x = "Math Score", y = "Reading score")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_color_distiller(palette = 4)

#While there is a positive linear relationship between math and reading scores,
#there is a negative linear relationship between the availability of free/
#reduced lunch and examination scores. Students who struggle to
#get free/reduced lunch tend to score higher in math and reading. Whereas,
#students eligible for free/reduced lunch tend to score less
#in math and reading.

#QUESTION 3

?wage1
view(wage1)

#1.
mean(wage1$educ) #Average education level. Ans-12.56274
min(wage1$educ) #Min education level.Ans-0
max(wage1$educ) #Max education level.Ans-18

#2,
mean(wage1$wage)#Ans-5.896103
#The average per-hour wage within the sample seems to be
#on the lower side.

#3.
round(proportions(table(wage1$female==1)),2) #Ans - True value (48%).
round(proportions(table(wage1$female==0)),2) #Ans - True value (52%).

#4.
round(proportions(table(wage1$married==1)),2) #Ans - True value (61%).

#5.
```

```
#a.
ggplot(wage1,aes(x=educ,y=wage))+
  geom_point()+
  geom_smooth(method = "lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Correlation between Education level & Hourly-Wage level",
       x = "Education Level (Years)",
       y = "Hourly-Wage Level")+
  theme(plot.title = element_text(hjust=0.5))

round(cor(x = wage1$educ, y = wage1$wage,method="pearson"),2)#Ans-0.41

#ASSOCIATION - As the education level increases, the hourly wage increases.
#STRENGTH & DIRECTION - There is a moderate positive correlation between
#education level and hourly-wage level.

#b.
ggplot(wage1,aes(x=educ,y=wage))+
  geom_point()+
  geom_smooth(method = "lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Correlation between Education level & Hourly-Wage level",
       x = "Education Level (Years)",
       y = "Hourly-Wage Level")+
  theme(plot.title = element_text(hjust=0.5))+
  facet_wrap(~married)

#For both married and unmarried people, there is a positive linear relationship
#between education level and hourly=wage level.However, the strength of the
#positive relationship differ. The strength is computed below.

round(cor(x = wage1$educ[wage1$married==0], y = wage1$wage[wage1$married==0],
          method="pearson"),2)#Ans-0.44

#There is moderate positive correlation between education level and hourly-
#wage level for unmarried people. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is moderate.

round(cor(x = wage1$educ[wage1$married==1], y = wage1$wage[wage1$married==1],
          method="pearson"),2)#Ans-0.39

#There is weak positive correlation between education level and hourly-wage
#level for married people. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is weak.

#c.
ggplot(wage1,aes(x=educ,y=wage))+
  geom_point()+
  geom_smooth(method = "lm",
              formula = y~x,
              se = FALSE)+
  labs(title = "Correlation between Education level & Hourly-Wage level",
       x = "Education Level (Years)",
       y = "Hourly-Wage Level")+
  theme(plot.title = element_text(hjust=0.5))+
  facet_wrap(~married~female)

round(cor(x = wage1$educ[wage1$married==0 & wage1$female==0],
          y = wage1$wage[wage1$married==0 & wage1$female==0],
          method="pearson"),2)#Ans-0.36

#STRENGTH & DIRECTION (Unmarried Male education level on wage)-
```

```
#Weak positive correlation. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is weak.

round(cor(x = wage1$educ[wage1$married==0 & wage1$female==1],
          y = wage1$wage[wage1$married==0 & wage1$female==1],
          method="pearson"),2)#Ans-0.5

#STRENGTH & DIRECTION (Unmarried Female education level on wage)-
#Moderate positive correlation. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is moderate.

round(cor(x = wage1$educ[wage1$married==1 & wage1$female==1],
          y = wage1$wage[wage1$married==1 & wage1$female==1],
          method="pearson"),2)#Ans-0.38

#STRENGTH & DIRECTION (Married Female education level on wage)-
#Weak positive correlation. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is weak.

round(cor(x = wage1$educ[wage1$married==1 & wage1$female==0],
          y = wage1$wage[wage1$married==1 & wage1$female==0],
          method="pearson"),2)#0.41

#STRENGTH & DIRECTION (Married Male education level on wage)-
#Moderate positive correlation. Therefore, with increase in education level,
#hourly-wage increases but the strength of this relation is moderate.
```