Module: ***ADVANCED METHODS FOR MANAGERS (MBA60)***
Academic Year: ***2021-22***

1st Written Assignment (WA1)

Model Answers

- Model answers are provided below for all WA1 questions.
- Marking is based on the outcomes and interpretation.
- R s/w package was utilized for the generation of all answers below. Source code is also provided as a reference for all questions.
- Students are free to use any type of software to work with.

Question 1: (5%)

Create the following new variables within the given dataset:

- **month**: From the original variable Month replace month numbers with labels as follows: 5=May, 6=June, 7=July, 8=August, 9=September.
- **TempCelsius**: Transform daily temperatures from Fahrenheit to Celsius using the appropriate conversion formula.
- **HotTemp**: A dichotomous variable with the value 1 if TempCelsius > 25 and 0 otherwise.
- **OzAlertLevel**: An ordinal variable with value 2 if Ozone > 65 & TempCelsius > 25, 1 if Ozone > 50 & Ozone ≤ 65 & TempCelsius > 25 and 0 in all other cases.

Answers

We create new variables using different techniques for each case.

a) **month**: Code numerical values to characters. e.g.

Month	MonthL
5	May
5	May
5	May
5	May

b) **TempCelsius:** Convert from one scale (Fahrenheit) to another (Celsius). The conversion formula from Fahrenheit to Celsius is **C=(F-32)*(5/9).** e. g.

Temp	TempCelsius
67	19,44
72	22,22
74	23,33
62	16,67
56	13,33
66	18,89
65	18,33

c) **HotTemp:** Create a dichotomous variable. e.g.

TempCelsius	HotTemp
24,44	0
25,56	1
23,33	0
19,44	0
28,89	1
29,44	1
26,11	1

d) **OzAlertLevel:** Create group data from numerical data. e.g.

TempCelsius	Ozone	OzAlertLevel
27,22	32	0
28,89	59	1
28,33	64	1
28,33	40	0
31,11	77	2
33,33	97	2
33,33	97	2
31,67	85	2
27,78	59	1
22,78	10	0

Indicative R code

```
install.packages("readxl")
install.packages("writexl")

library("readxl")
library("writexl")

data <- read_excel("C:\\EAP\\MBA60\\MBA60_2022_WA1_DATA.xlsx")

#Question 1
    data$month=c(rep("May",31),rep("June",30),rep("July",31),
             rep("August",31),rep("September",30)) # create month variable

    data$TempCelsius <- (data$Temp - 32) * 5/9 # convert F to C, create TempCelsius
var

    data$TempHot <- ifelse(data$TempCelsius>25, 1, 0) # create TempHot variable
```

```
    data$OzAlertLevel<-with(data,
                        ifelse(Ozone <=65 & Ozone>50 & TempCelsius>25, 1,
                        ifelse(Ozone >65 & TempCelsius>25, 2,
                                0))) # create OzAlertLevel variable

write_xlsx(data, "C:\\EAP\\MBA60\\MBA60_2022_WA1_DATA_NEW.xlsx")
```

Question 2: (25%)

2.1 Calculate the mean and the standard deviation of the temperature (in Celsius) for each month. Compare and comment on your results.

2.2 Calculate the proportion of the hot days for each month (hot days are defined as those with temperature >25 Celsius degrees). Comment on your findings. Are these in line with the ones from question 2.1?

2.3 Calculate the median, the 1st and the 3rd qua.rtile of the variable **Wind** for each month. Interpret your findings to describe the distribution of the Wind data in August.

2.4 Compare the variability of **Ozone**, **TempCelsius**, **Solar.R** and **Wind** for the entire period, using the appropriate measure(s). Which variable shows the largest variability? Comment on your results.

2.5 Construct the frequency distribution and the percentage distribution of the variable **Solar.R** for the entire period, by choosing the appropriate number of classes (justify your choice). Calculate the mean twice using the raw and the grouped data. Compare and comment on the results. Which one will you use and why?

Answers

Q 2.1

Mean of the temperature (in Celsius) for each month.

```
  Month TempCelsius
1     5    18.63799
2     6    26.16667
3     7    28.83513
4     8    28.87097
5     9    24.94444
```

Standard deviation of the temperature (in Celsius) for each month.

```
  Month TempCelsius
1     5    3.808261
2     6    3.665883
3     7    2.397507
4     8    3.658475
5     9    4.642040
```

Interpretation

As expected, July and August have the largest mean temperatures, both around 29 ^{0}C. However, the temperature in August shows a greater variability than in July. Furthermore, September seems to be the month with the greatest variability of all, having both warm and moderate cold days. Finally, weather in NY seems to be quite cold in May!

Q 2.2

Proportion of the hot days for each month (with temperature >25 Celsius degrees).

```
  Month    TempHot
1     5 0.06451613
2     6 0.53333333
3     7 0.93548387
4     8 0.83870968
5     9 0.40000000
```

Interpretation

July is the month with the largest proportion of days with temperature above 25 ^{0}C (0.94) followed by August (0.84). May is the coldest month, having only two days with max temperature above 25 ^{0}C (0.0645 x 31=2). These results are in line with the ones from question 2.1.

Q 2.3

Median, 1st and 3rd quartile of the variable Wind for each month.

```
  Month Wind.1st Qu. Wind.Median Wind.3rd Qu
1     5       8.90       11.50        14.05
2     6       8.00        9.70        11.50
3     7       6.90        8.60        10.90
4     8       6.60        8.60        11.20
5     9       7.55       10.30        12.32
```

Interpretation

August seems to be the less windy month in NY (no meltemia). 50% of the days in August have daily average wind less than 8.6 miles per hour and 25% of them have daily average wind less than 6.6. Furthermore, only 25% of the days in August have daily average wind above 11.2 miles per hour. Finally, the daily average wind in August in NY varies from a minimum of 2.3 to a maximum of 15.5.

The windiest month in NY is May. For exercise, you may try a similar interpretation.

Q 2.4

Variability of Ozone, TempCelsius, Solar.R and Win

In order to compare the variability between variables measured in different scales we calculate the coefficient of variation as the ratio of the standard deviation by the mean. Results for each variable are as follows.

```
Ozone:          0.7258156
TempCelsius:    0.2062943
Solar.R:        0.4742361
Wind:           0.3538032
```

Interpretation

Ozone, for the entire period, has the largest variability among the variables in our study. On the other hand, temperature seems to have the smallest variability.

Q 2.5

Frequency distribution and percentage distribution of the variable Solar.R.

In order to construct the frequency distribution, we have a breaking scheme of 7 classes with width=50.

Bin	Freq	PerCent Freq
[0,50)	17	0.11111111
[50,100)	17	0.11111111
[100,150)	17	0.11111111
[150,200)	27	0.17647059
[200,250)	28	0.18300654
[250,300)	38	0.24836601
[300,350)	9	0.05882353

A variety of rules can be applied for selection of the number of bins.
e.g.
Sturges rule k = 1 + 3.322 log n, where
k = the number of bins
n = the number of observations in the data set.

Mean for the raw data.

```
The mean is:  185.5294
```

Mean for the grouped data.

```
The mean is:  184.4771
```

Interpretation

One fourth of the observations (25%) are placed between 250 and 300, the mode interval of the distribution. Furthermore the distribution is skewed to the left.

The grouped mean (184.5) is slightly smaller than the raw data mean (185.5). If the raw data is available, then we better use the raw data mean, a measure that utilizes all the available information from the data.

Indicative R code

```
library("readxl")

data <- read_excel("C:\\EAP\\MBA60\\MBA60_2022_WA1_DATA_NEW.xlsx")

#Question 2.1
    aggregate(data["TempCelsius"],by=list(Month=data$Month),mean)
    aggregate(data["TempCelsius"], by=list(Month=data$Month),sd)

#Question 2.2
    as.numeric(data$TempHot)
    aggregate(data["TempHot"],by=list(month=data$Month),mean)

#Question 2.3
round(aggregate(data["Wind"],
                by=list(Month=data$Month),summary),2)

#Question 2.4
    CV=function(x){sd(x)/mean(x)} # create function for CV calc

    CV(data$Ozone)
    CV(data$TempCelsius)
    CV(data$Solar.R)
    CV(data$Wind)
```

```
#Question 2.5
    breaks = seq(0, 350, by=50)
    Solar.cut = cut(data$Solar.R, breaks, right=FALSE)
    cbind(table(Solar.cut),prop.table(table(Solar.cut)))

    install.packages("gds");library(gds)
    Solar.freq = cbind(table(Solar.cut))
    gds(seq(0,300,by=50), seq(50,350,by=50), Solar.freq)$mean
    mean(data$Solar.R)
```

Question 3: (25%)

3.1 Plot the histogram and the boxplot of the **Ozone** variable for the entire period using the appropriate labels and titles. Interpret your data visualization findings describing the main characteristics of the distribution.

3.2 Plot the boxplots of the **Ozone** variable for each month, side by side, in one graph. Add a horizontal line to the graph at the value of the Grand Median (the Ozone median of the entire period). Discuss your findings as derived by the graph.

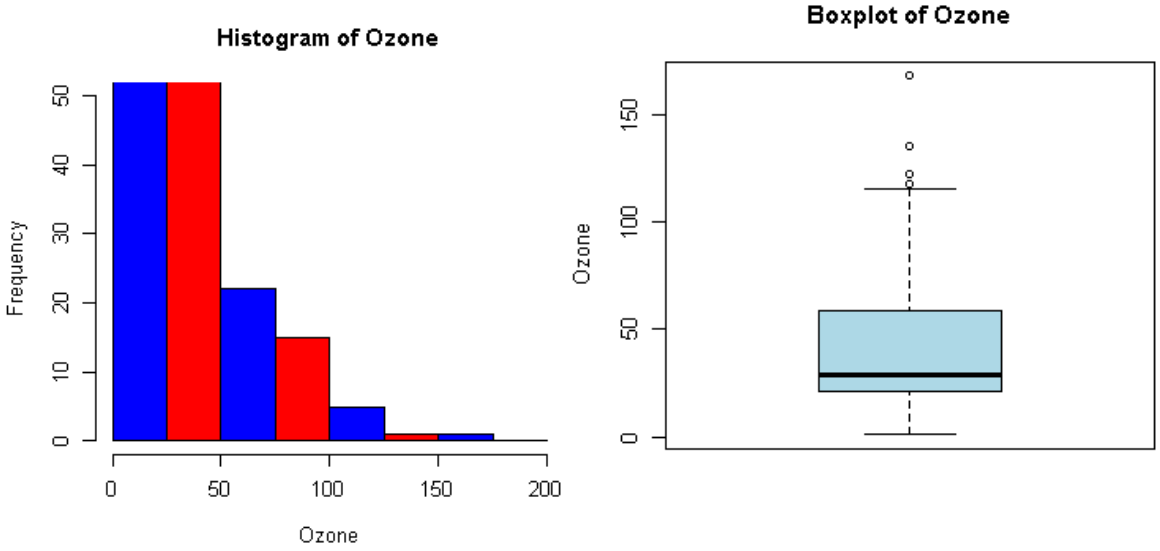
3.3 Construct scatter plots of the variable pairs (**Temp**, **Wind**), (**Temp**, **Ozone**), (**Ozone**, **Wind**). Detect and describe the possible relationship of the variables in each pair. Comment on your findings.

3.4 Create a clustered bar chart for the variable **OzAlertLevel** per month. Comment on the results.

Answers

Q 3.1

Histogram and boxplot of the Ozone variable.



Interpretation

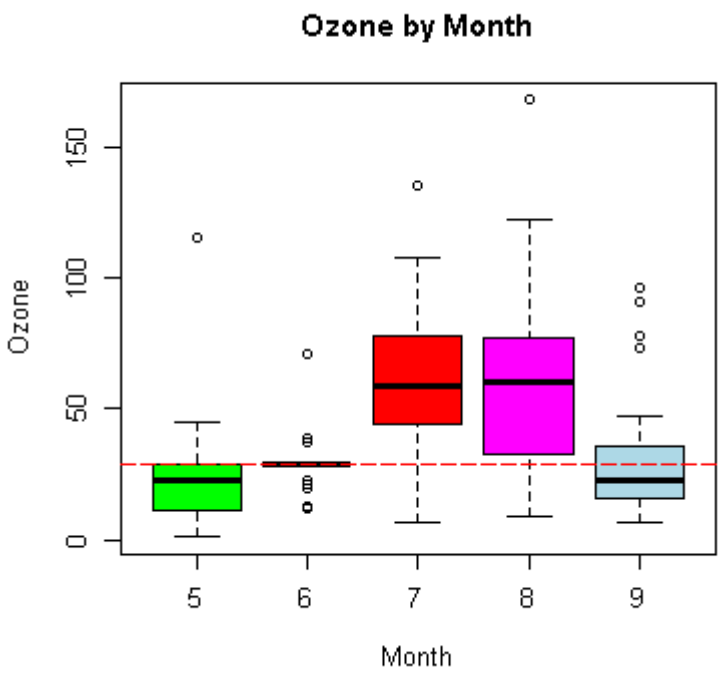
The boxplot and the histogram clearly indicate that

a) the distribution of Ozone for the whole period is right skewed

b) there are outliers into the right tail of the distribution (large values of ozone)

Q 3.2

Boxplots of the Ozone variable for each month.

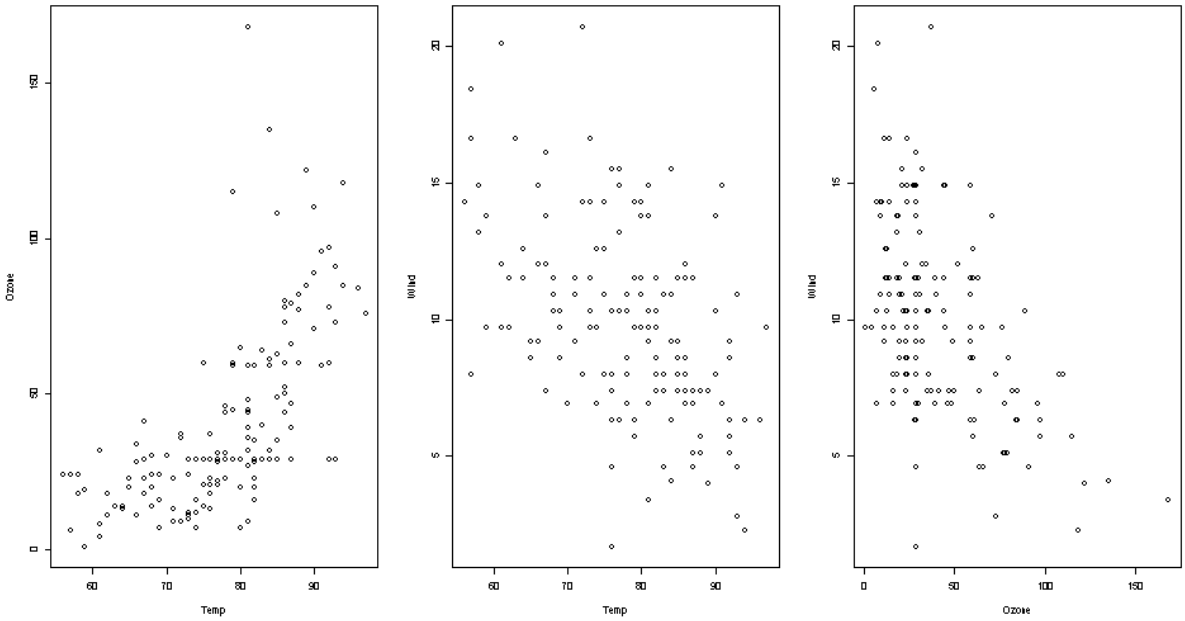


Interpretation

The median Ozone in the warm months, July and August, is way above the grand median. Even the 1st quartiles of these months are above the grand median. Therefore, at least the 75% of July and August days had Ozone level in the overall top 50%. On the other hand, in May and September median Ozone is below the grand median. June seems to be on the grand median with a considerable number of outliers in both tails. The September box plot reveals several outliers in the right tail of the ozone distribution.

Q 3.3

Scatter plots of the variable pairs (Temp, Wind), (Temp, Ozone), (Ozone, Wind).

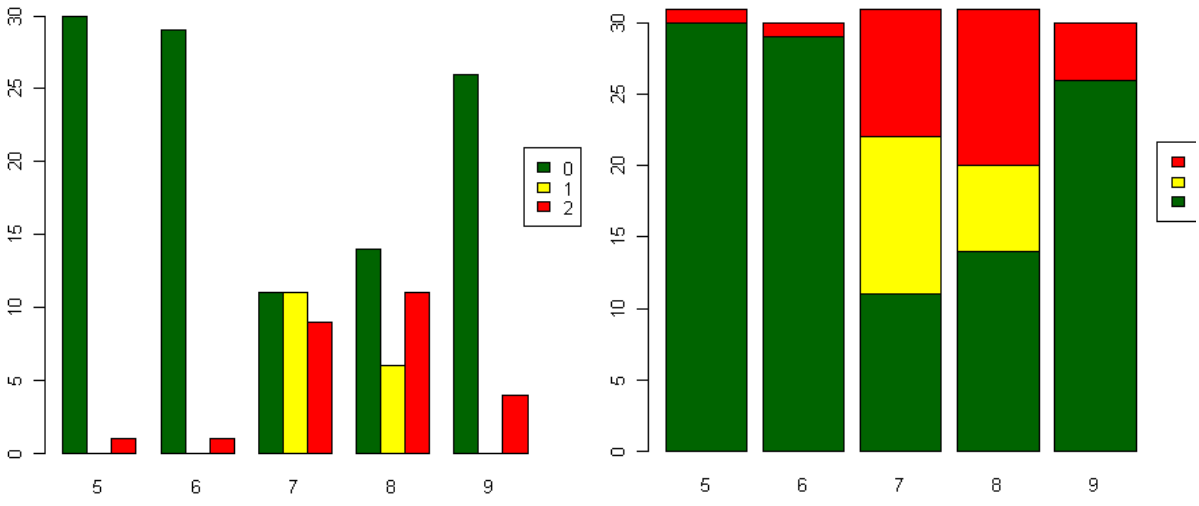


Interpretation

Ozone and Temperature shows a positive relation, meaning that large values for temperature are related with large values for ozone. On the contrary, both temperature and ozone are going down during the windy days of the period under study.

Q 3.4

Clustered bar chart for the variable OzAlertLevel per month.



Interpretation

July and August are the months with a considerable number of days with Ozone Alert at level 1 & 2. For the rest of the months this number is very small.

Indicative R code

```
#Question 3.1
    hist (data$Ozone,
      breaks=seq(from=0,to=200,by=25),freq=TRUE,col=c("blue","red"),border="black",
      main="Histogram of Ozone", xlab="Ozone",ylab="Frequency",
      xlim=c(0,200), ylim=c(0,50),
      axes=TRUE, labels=FALSE)

    boxplot(data$Ozone,col="lightblue",
      main="Boxplot of Ozone", xlab="", ylab="Ozone")

#Question 3.2
    boxplot(Ozone ~ Month, data = data,
      col=c("green","orange","red","magenta", "lightblue"),
      main="Ozone by Month")
    abline(h=median(data$Ozone), col="red",lty=5)

#Question 3.3
    par(mfrow=c(1,3))
    plot(data$Temp, data$Ozone, xlab="Temp", ylab="Ozone")
    plot(data$Temp,data$Wind, xlab="Temp", ylab="Wind")
    plot(data$Ozone,data$Wind, xlab="Ozone", ylab="Wind")

#Question 3.4
    m<-table(data$OzAlertLevel,data$Month)
    barplot(m, col=c("darkgreen","yellow","red"),
      legend = rownames(m), beside=T,
      args.legend = list(x = "topright",inset = c(-0.07, 0.3)))

    barplot(m, col=c("darkgreen","yellow","red"),
      legend = rownames(m), beside=F,
      args.legend = list(x = "topright",inset = c(-0.12, 0.3)))
```

Question 4: (15%)

When the Environmental Protection Agency (EPA) in New York detects Ozone level above 65, they issue the first warning (yellow flag).

4.1 Calculate the probability of at least one yellow flag within a week (7 days) between May and September.
4.2 How many yellow flags are expected during the same week?

For answering the above questions assume that the probability of a yellow flag in any given day is estimated from the percentage of yellow flags in the given data set. (Percentage of days with Ozone > 65).

Answers

Probability of a yellow flag in any given day = $\frac{Number\ of\ days\ with\ Yellow\ flag\ (Ozone>65)}{Total\ number\ of\ observations} = \frac{26}{153} = 0.1699346$

Q 4.1

The number of yellow flagged days in a week follows Binomial distribution with n=7 and p=0.1699.

For

X: number of days with yellow flag

n=7 trials

p=0,1699

and X∼Bin(n,p)

the probability that at least 1 out of 7 randomly selected days will be yellow flagged is

P(X>=1| n=7, p=0,1699) = P(X=1)+ P(X=2)+ P(X=3)+ P(X=4)+ P(X=5)+ P(X=6)+ P(X=7)= 0,389059044+0,239060377+0,081606956+0,016714678+0,002054093+0,000140239+0.000001 = 0.728411 = **72.84%**

Q 4.2

The expected number of yellow flags in a week is

$n * p = 7 * 0.1699 = \mathbf{1.189542}$

Indicative R code

```
#Question 4.1
    emerge1<-with(data, Ozone>65)
    as.numeric(emerge1)
    p<-mean(emerge1)
    p

    n=7
    pbinom(0,n,p,lower.tail=F) # P(X>=1)

#Question 4.2
    meanvalue<-mean(emerge1)*7;meanvalue #np
```

Question 5: (15%)

When the Environmental Protection Agency (EPA) in New York detects Ozone level above 65 and Temp above 90, they issue a second warning (red flag).

5.1 Calculate the probability of at least two red flags within a given month (31 days).

5.2 What is the probability of at least one but no more than three red flags during the same month?

For answering the above questions assume that the probability of a red flag in any given day is estimated from the percentage of red flags in the given data set. (Percentage of days with Ozone > 65 and Temp > 90). Use the Poisson approximation to the Binomial distribution.

Answers

Q 5.1

The percentage of red flagged days is 0.06536. In a given month with n = 31 days and p = 0.06536, the number of red flagged days follow approximately Poisson distribution with λ = np = 2,026144.

λ=2,026144

Y~P(λ)

P(Y>=2|λ=2,02614) = 1-P(X<=2) = 1- P(X=1)- P(X=0)= 1-0,26713-0,13184= 0,60103 = 60.10%

The probability of at least 2 red flags within a month is 60,1%.

Q 5.2

λ=2,026144

Y~P(λ)

P(0<Y<=3|λ=2,02614) = P(X=1)+ P(X=2)+ P(X=3) = 0,26713+ 0,27062+ 0,18277= 0,7261 = 72.61%

The probability of at least one but no more than three red flags during the same month is about 72%. (Sum up the probability of 1,2 and 3).

Indicative R code

```
#Question 5.1
    emerge2<-with(data, Ozone>65 & Temp > 90)
    as.numeric(emerge2)
    mean(emerge2)

    labmda<-mean(emerge2)*31
    ppois(1,labmda,lower.tail=F) # P(X>1)

#Question 5.2
    sum(dpois(1:3,labmda)) # P(1X>1)
```

Question 6: (15%)

The daily temperature between May and September follows approximately normal distribution. Using this fact and after estimating the **μ** and **σ** from the dataset as the mean and the standard deviation of the variable **TempCelsius,** perform the following calculations:

6.1 Calculate the probability that the temperature in a randomly selected day will be larger than 20 Celsius degrees.

6.2 Calculate the temperature that a day must have (at least) in order to be among the 10% of the warmest days of the season.

Answers

Q 6.1
For
X: daily temperature
X~N(μ= 25,4902, σ^2=27,6516) // μ and σ are calculated from the dataset

We have
P(X>20) = 1-P(X<=20) $=1-P\left(\frac{X-\mu}{\sigma}\leq\frac{20-25{,}49}{5{,}2584}\right)=1-P(Z\leq-1{,}044)=1-0{,}1492=$ 0,8508= **85,08%**

The probability that the temperature in a randomly selected day will be larger than 20 0C is about 85%.

Q 6.2
For
X: daily temperature
X~N(μ= 25,4902, σ^2=27,6516) // μ and σ are calculated from the dataset

We have
P(X<X_o) = 0,9 = Φ(1,28), so $Z=\frac{Xo-\mu}{\sigma}=\frac{Xo-25{,}49}{5{,}2584}=1{,}28 \Rightarrow Xo=\mathbf{32,22\ C}$

In order to be among the 10% of the warmest days of the season, a day must have temperature of at least 32.22 C.

Indicative R code

```
#Question 6.1
    mu<-mean(data$TempCelsius)
    sd<-sd(data$TempCelsius)

    pnorm(20,mu,sd,lower.tail=F)

#Question 6.2
    qnorm(.1,mu,sd,lower.tail=F)
```