UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2019/2020

COURSE NAME	:	DATA VISUALIZATION
COURSE CODE	:	BWB 31503
PROGRAMME	:	BWQ
EXAMINATION DATE	:	DECEMBER 2019 / JANUARY 2020
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **NINE (9)** PAGES

Q1 (a) (i) Explain the box plot and state **TWO (2)** uses of it. Based on **Figure Q1(a)**, conclude the contents in the box plot.

(6 marks)

(ii) Define the inferential statistics appropriately.

(2 marks)

(b) (i) Explain the process of data analysis.

(2 marks)

(ii) List **FOUR (4)** types of data in statistics. Explain each of them accordingly.

(8 marks)

(c) Determine if following statements are **TRUE** or **FALSE**.

(i) Descriptive statistics is a statistical and graphic techniques to present information about the dataset being studied.

(1 mark)

(ii) Computing descriptive statistics is a preliminary step before proceed to an inferential statistical analysis.

(1 mark)

(iii) A common practice is to begin an analysis by examining graphical displays of a dataset and to compute some basic inferential statistics to obtain some sense of the data behaviour to be analysed.

(1 mark)

(d) Data analytic process is extremely important for the Risk Manager to improve decision making and increase accountability. Identify **FOUR (4)** challenges in the data analytic process.

(4 marks)

Q2 (a) A simple statistical method to produce and visualize the summary of low-dimensional datasets, but as the dimensionality increase, some problems arise. Determine **ONE (1)** problem mentioned. Explain the suitable statistical methods appropriately.

(3 marks)

(b) Explain univariate analysis and multivariate analysis.

(4 marks)

(c) (i) Define the textual data visualization and explain **ONE (1)** example of graphical display to visualize it.

(4 marks)

(ii) A researcher process a speech titled "I have a dream speech" from "Martin Luther King". Based on **Figure Q2(c)(i)** and **Figure Q2(c)(ii)**, examine the best conclusion could be made from both figures.

(2 marks)

(d) (i) Animation visualization is a good technique to attract an interest of the audience. Determine **THREE (3)** reasons of using this technique.

(3 marks)

(ii) Define the temporal data visualization.

(2 marks)

(e) Explain the network data visualization. Determine **TWO (2)** concerns before designing a network visualization. Based on the given *R-codes*, sketch the appropriate network data visualization that will be produce.

```
> library(igraph)
> netdata <- graph(edges=c(1,2, 2,3, 3, 1), n=10)
> plot(netdata)
```

(7 marks)

Q3 (a) (i) Explain the logistic regression.

(2 marks)

(ii) Identify **THREE (3)** assumptions of the logistic regression.

(3 marks)

(iii) Explain **TWO (2)** situations, the response and explanatory variables before adopting binary logistic regression:

(6 marks)

(iv) Explain the overfitting problem in logistic regression accordingly.

(3 marks)

(b) A study an outlier based on `cars` dataset in R has been runned with the appropriate plots and data analyses as shown in **Table Q3(b)** and **Figure Q3(b)**.

(i) Define the outlier appropriately.

(2 marks)

(ii) Based on **Figure Q3(b)**, interpret both plots (i) and (ii) respectively.

(2 marks)

(iii) Assume that we use **Figure Q3(b)**, plot (i) to train the model, discuss the result of prediction may obtain.

(2 marks)

(iv) Time series is any metric measured over regular time intervals that forms a time series. Determine **THREE (3)** conditions if a time series is said to be stationary. Then, explain the used of lags in the time series.

(5 marks)

Q4 (a) (i) Categorical data is the statistical data type consists of categorical variables that has been converted into grouped data. It's derived from observations made of qualitative data that are summarised as counts or cross tabulations. Purely categorical data are summarised in the form of a contingency table. State **THREE (3)** techniques to analyse categorical data.

(3 marks)

(ii) A study by a group of student of BWB 31503 Data Visualization aims to determine which smartphone brands are the most popular in 2019. The data consist of variable brand in 2019 together with other variables. Since this is a categorical variable, a simple frequency table regarding their frequencies has been produced. Explain any features that appear in **Table Q4(a)**.

(3 marks)

(iii) Before any algorithm is applied to the data, it should be structured in the real world. Most of the data are unstructured. State **ONE (1)** reason the data untidy before any further algorithm applied.

(1 mark)

(b) (i) List **TWO (2)** techniques of checking data normality and for each of techniques, explain **TWO (2)** uses of these methods.

(10 marks)

(ii) Based on **Figure Q4(b)(ii)**, interpret any results that could be obtained.

(2 marks)

(iii) In some cases, the true relationship between the outcome and a predictor variable might not be linear. There are different solutions extending the linear regression model to capture these non-linear effects. Identify **THREE (3)** solutions to capture these non-linear effects.

(3 marks)

(iv) Based on **Figure Q4(b)(iv)**, identify the problem of the plot.

(3 marks)

- END OF QUESTIONS –

FINAL EXAMINATION

SEMESTER / SESSION : SEM I / 2019/2020
PROGRAMME CODE : BWQ
COURSE NAME : DATA VISUALIZATION
COURSE CODE : BWB 31503

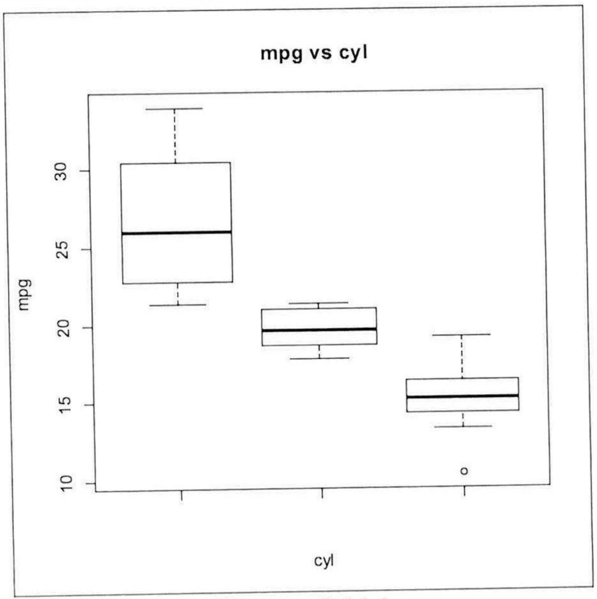


Figure Q1(a)

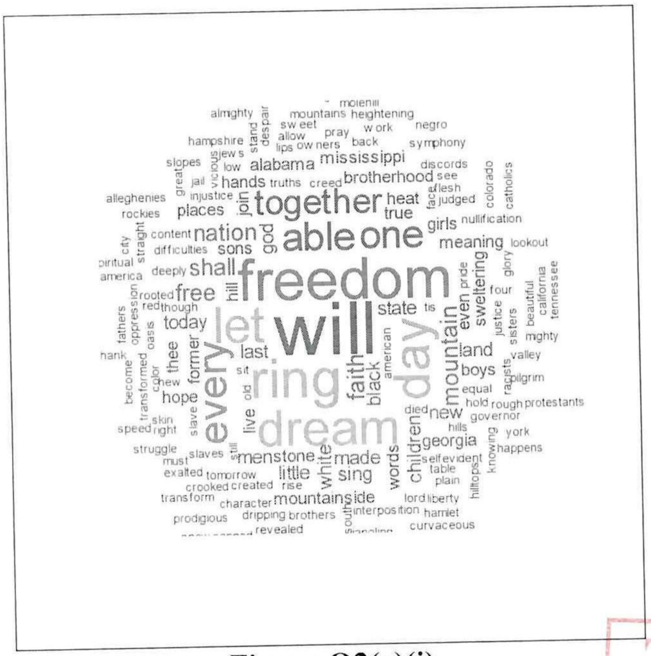


Figure Q2(c)(i)

FINAL EXAMINATION

SEMESTER / SESSION : SEM I / 2019/2020
COURSE NAME : DATA VISUALIZATION
PROGRAMME CODE : BWQ
COURSE CODE : BWB 31503

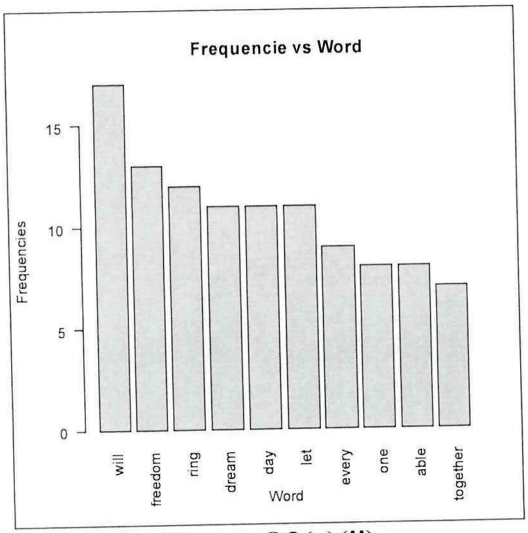
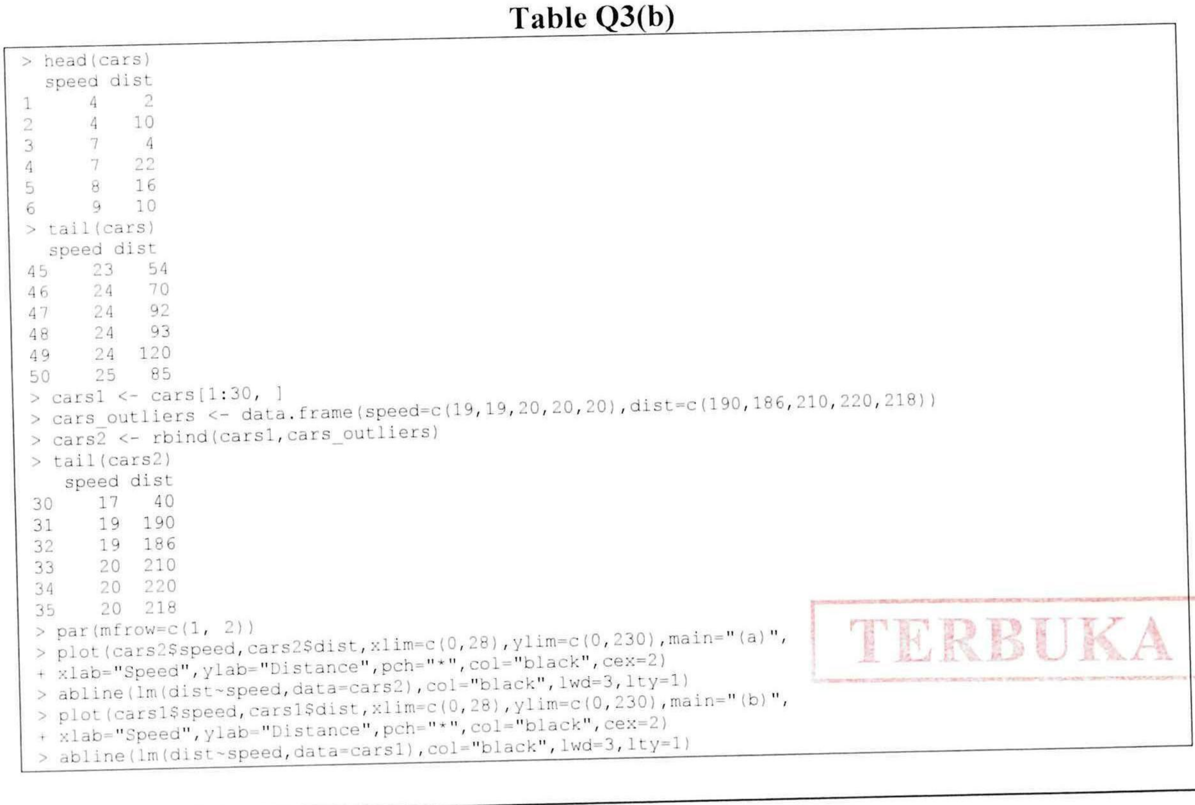


Figure Q2(c)(ii)

Table Q3(b)



```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
> tail(cars)
   speed dist
45    23   54
46    24   70
47    24   92
48    24   93
49    24  120
50    25   85
> cars1 <- cars[1:30, ]
> cars_outliers <- data.frame(speed=c(19,19,20,20,20),dist=c(190,186,210,220,218))
> cars2 <- rbind(cars1,cars_outliers)
> tail(cars2)
   speed dist
30    17   40
31    19  190
32    19  186
33    20  210
34    20  220
35    20  218
> par(mfrow=c(1, 2))
> plot(cars2$speed,cars2$dist,xlim=c(0,28),ylim=c(0,230),main="(a)",
+ xlab="Speed",ylab="Distance",pch="*",col="black",cex=2)
> abline(lm(dist~speed,data=cars2),col="black",lwd=3,lty=1)
> plot(cars1$speed,cars1$dist,xlim=c(0,28),ylim=c(0,230),main="(b)",
+ xlab="Speed",ylab="Distance",pch="*",col="black",cex=2)
> abline(lm(dist~speed,data=cars1),col="black",lwd=3,lty=1)
```

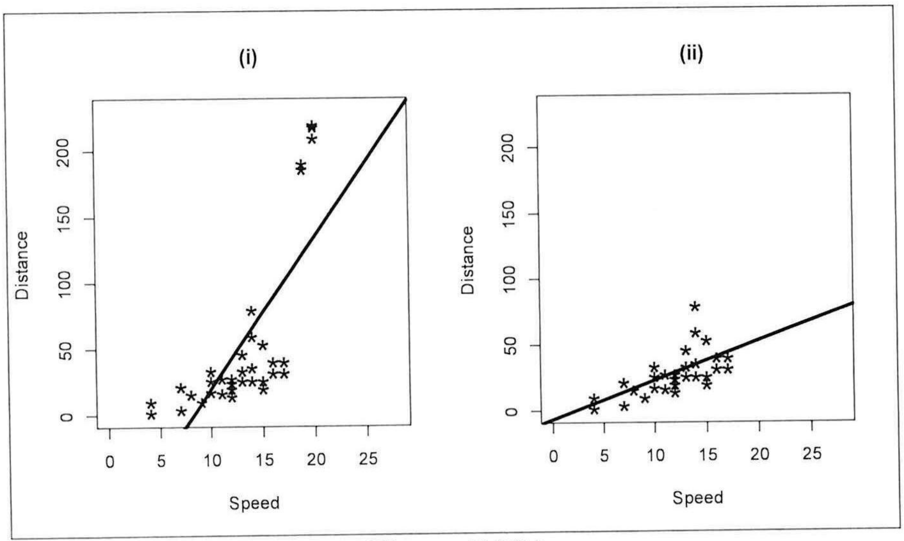
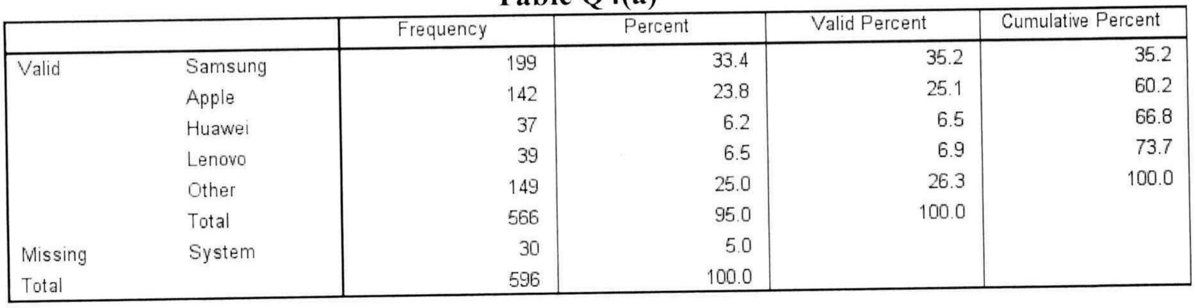


Figure Q3(b)

Table Q4(a)



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Samsung	199	33.4	35.2	35.2
	Apple	142	23.8	25.1	60.2
	Huawei	37	6.2	6.5	66.8
	Lenovo	39	6.5	6.9	73.7
	Other	149	25.0	26.3	100.0
	Total	566	95.0	100.0	
Missing	System	30	5.0		
Total		596	100.0		

FINAL EXAMINATION

SEMESTER / SESSION	: SEM I / 2019/2020	PROGRAMME CODE	: BWQ
COURSE NAME	: DATA VISUALIZATION	COURSE CODE	: BWB 31503

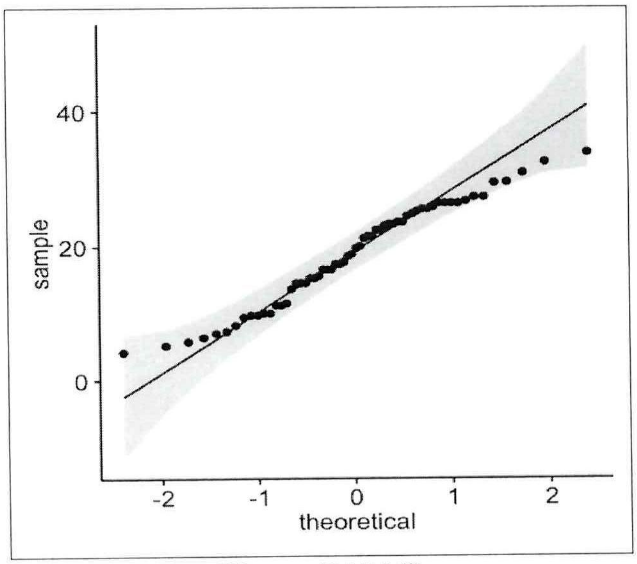


Figure Q4(b)(i)

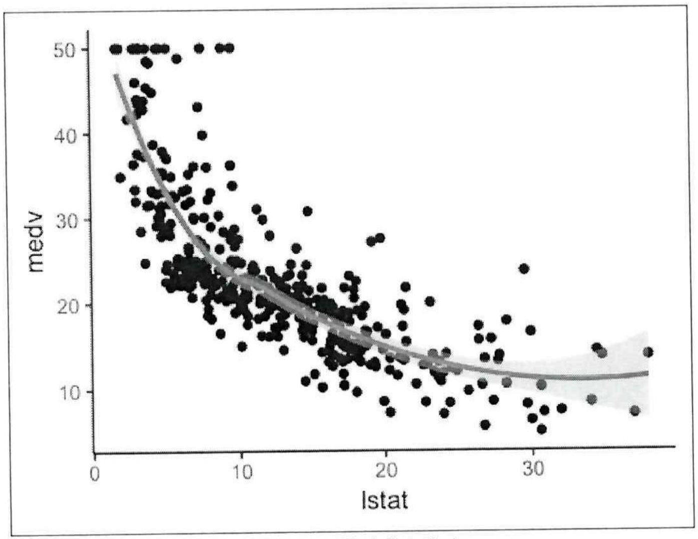


Figure Q4(b)(iv)