

Northumbria University
Department of Computer and Information Sciences

PE7050 Statistics & Business Intelligence
Final Assessment Submission

Student: Scott Cumming
Student Number: 21056374
Date: 25th March 2023

Contents

1. Introduction	2
2. Answer to Question 1	2
3. Answer to Question 2	3
4. Answer to Question 3	4
5. Answer to Question 4	5
6. Answer to Question 5	6
7. Answer to Question 6	7
8. Answer to Question 7	8
9. Answer to Question 8	11
10. Answer to Question 9	13
 Appendix A – Bibliography	 19
Annex A – Question 10 Report	21

1. Introduction

This document contains the author's submission for the PE7050 Statistics and Business Intelligence module final assessment. The answers to questions 1 to 9 are provided below in the body of the document. The report containing the response to question 10 can be found in Annex A.

2. Answer to Question 1

On face value, the study appears to show that gym-only workouts are more effective than exercise classes at helping individuals to lose weight. The mean weight loss achieved by the gym-only group was 2.3kg over 6 months, compared to 1.6kg for those attending classes. However, the mode weight loss for those undertaking gym-only workouts was 1.5kg, which suggests the mean may be higher due to the presence of extreme values. This does not appear to be the case in the exercise class group as their mode (1.3kg) was closer to the mean. Therefore, it is important to scrutinise the data that has been provided to determine what conclusions can be drawn from the study.

The standard deviation shows that the typical weight loss for the exercise class group was $1.6\text{kg} \pm 1.03\text{kg}$ (between 0.57kg and 2.63kg), compared to $2.3\text{kg} \pm 1.33\text{kg}$ (between 0.97kg and 3.63kg) for the gym-only group. This reveals that the latter values are spread further from the mean and there is more variability in this dataset, which supports the argument that there are extreme values in the gym-only group.

The statistics given permitted the calculation of Pearson's first coefficient of skewness, using the following formula:

$$Sk1 = \frac{X - Mo}{s}$$

The results for each group are shown in the table below:

Exercise Class	$Sk1 = \frac{(1.6 - 1.3)}{1.03} = 0.291262$
Gym-Only Workouts	$Sk1 = \frac{(2.3 - 1.5)}{1.33} = 0.601504$

Table 1 - Pearson's First Coefficient of Skewness for Exercise Class & Gym-Only Groups

A value of 0.29 shows the exercise class group dataset is nearly symmetrical with a slightly positive skew. The gym-only workout group's dataset is more positively skewed as its coefficient is greater; at 0.60 it is considered to have a moderate skew

(Menon, 2023). These results provide yet more evidence that there are extreme values in the gym-only workout group. They also show that the median for each group – which was not provided – will be less than the mean but more than the mode (Kovchegov, 2022).

The gym staff want to find out which form of exercise is better at “helping individuals lose weight”, implying they may want inferences to be made about the population of gym members in general, not just their own members. Consequently, it is important to analyse how closely the samples included in the study represent this wider population.

Convenience sampling was employed in this study as participants were selected purely from the gym membership. Urdan (2017, p.4) states that this form of sampling is considered acceptable when the samples do not differ from the wider population “in ways that influence the outcome of the study”. The samples in this study do not necessarily represent the wider population of gym members as they are drawn exclusively from one club, which might attract a narrow demographic. For example, less affluent individuals are unlikely to join a health and fitness club which charges expensive fees, meaning a sample drawn solely from its membership cannot be considered representative of all gym members. Similarly, the demographics of the area where a club is located and recruits its members can vary from that of another club located in a different area.

As a result, the findings of the study should not be used to make inferences about the wider population of gym members, unless the staff are confident the samples selected are representative of this population. If this is not the case, further research is required.

Nevertheless, conclusions can be drawn about the members at this specific gym because a census of the whole population was conducted. The findings suggest that gym-only workouts are slightly more effective than exercise classes at helping individuals who are members at this gym to lose weight. Although the gym-only workout group contained extreme values which increased the mean, these values do not affect the mode. The mode for this group was 0.2kg higher than the exercise class group, indicating that weight loss was generally higher when considered alongside the mean and standard deviation. When the mode is based on a dataset with a small number of values, it may be unreliable (Hayes, 2022). However, the mode for each group was based on a population of 62 and 45 in this study, therefore its reliability here is not a concern.

3. Answer to Question 2

The main approaches to dealing with missing data involve the use of deletion or procedures based on weighting, imputation or modelling (Carver, 2017). Each approach “has some flaws and the potential to introduce some bias” into an analysis (Dodd, 2022, p.55), therefore the one selected will depend on the aims of the study, the size of the dataset, the data type and the reason why the values are missing.

The question asks for a description of one of these approaches, therefore this answer will explore the use of imputation. This approach replaces missing data with estimated values and can be done using a variety of methods, including last value carried forward, mean substitution, hot-deck imputation, cold-deck imputation and multiple imputation (Bennett, 2001), amongst others.

The last value carried forward method is only applicable in longitudinal studies and uses the value of the variable when it was last observed. Mean substitution involves replacing the missing value with the mean and hot/cold-deck imputation takes a value from an estimated distribution of the current data. All of these methods are considered a form of single imputation and although simple and easy to implement, they generally underestimate the variance and ignore uncertainty (Takahashi & Ito, 2012), providing replacement values that are often unrealistic (Salgado *et al.*, 2016). This is a concern as the variance is used regularly to calculate other statistics (Urdan, 2017), meaning the use of single imputation could lead to biased or flawed analyses.

Multiple imputation attempts to reduce the uncertainty inherent in these methods and reflect the increased variation caused by missing values (Bennett, 2001; Sainani, 2015). For each missing data point, a value is generated M times to produce M complete datasets using a model that employs random variation. Each dataset is analysed individually then the results are combined and used to calculate one, single value. This value is considered a plausible estimate of the missing data point as it is derived from the “distributions of and relationships among observed variables in the data set” (Li, Stuart & Allison, 2015), therefore it is understandable why multiple imputation is favoured by many (Nguyen, Carlin & Lee, 2017).

Despite the advantages over other approaches, adopting multiple imputation isn't always appropriate. Liu *et al.* (2005) note it can be used only with data that is Missing At Random (MAR), not Missing Completely At Random (MCAR) or Not Missing At Random (NMAR). Similarly, the approach often assumes the data being imputed follows a normal distribution, so including variables that aren't distributed this way can introduce bias (Sterne *et al.*, 2009). Moreover, some choose not to use multiple imputation as it incurs higher costs than other approaches and involves more complexity (Stats NZ, 2022). However, this latter issue can be overcome using modern statistical software with integrated algorithms which automate much of the process and make multiple imputation more accessible (Sainani, 2015).

(Word count – 477)

4. Answer to Question 3

(a) The event represented by region 5 is simply that the family will experience mechanical problems. In notation form, this is equivalent to:

$$M - (T \cup V) = M \cap T' \cap V' = \{\text{mechanical problems}\}$$

(b) Region 3 are the events the family will receive a ticket for committing a traffic violation and that they will arrive at a campsite with no vacancies. These events can be represented by the notation:

$$T \cap V \cap M' = \{\text{traffic violation, no vacancies}\}$$

(c) The events represented by regions 1 and 2 includes all 3 sets and involves the family experiencing mechanical problems, receiving a ticket for committing a traffic violation and arriving at a campsite with no vacancies. In notation form, this is:

$$M \cap V = \{\text{mechanical problems, traffic violation, no vacancies}\}$$

(d) Regions 4 and 7 are the events the family will experience mechanical problems and that they will receive a ticket for committing a traffic violation. This is represented by the notation:

$$V' \cap T = \{\text{mechanical problems, traffic violation}\}$$

(e) The events represented by regions 3, 6 and 7 are the family will receive a ticket for committing a traffic violation and that they will arrive at a campsite with no vacancies. Region 8 does not represent an event or a set. In notation form, regions 3, 6 and 7 are represented by:

$$(V \cup T) \cap M' = \{\text{traffic violation, no vacancies}\}$$

5. Answer to Question 4

(a) The probability of an event is $0 \leq P(E) \leq 1$ and the probabilities given for each location in this scenario add up to 1 ($0.15 + 0.07 + 0.12 + 0.38 + 0.28$). Three of the location types given are bedrooms (adult, child and other bedroom) and the sum of their probabilities is 0.34 ($0.15 + 0.07 + 0.12$). Consequently, the probability that an LCD TV is in a bedroom is 0.34.

(b) The probability that an LCD TV is not in a bedroom can be calculated in two ways, by subtracting the figure obtained in the answer to (a) from one ($1 - 0.34$) or by calculating the sum of the other locations ($0.38 + 0.28$). Each approach provides the answer 0.66.

(c) If a household is selected at random from those with an LCD TV, the device is most likely to be found in an office or den. The probability of this is 0.38, compared to 0.34 for a bedroom (adult, child and other combined) and 0.28 for any other room. However, it's still important to note that there is a 0.62 probability that it will not be in an office or den, despite this location being the most likely out of all the locations.

6. Answer to Question 5

A1, A2, A3, A4 and A5 will represent each Apple iPhone subjected to the shock test. All possible variations of the outcome are denoted by the events E1, E2, ...E32. It is noted that the probability an iPhone will survive the test is 0.70, therefore the probability it will not survive is 0.30 ($1 - 0.70$). In other words:

$$P(S) = \frac{7}{10} \quad \text{and} \quad P(NS) = \frac{3}{10}$$

The outcomes where exactly 3 out of the next 5 iPhones survive are highlighted in yellow in the table below.

	A1	A2	A3	A4	A5
E1	0	0	0	0	0
E2	0	0	0	0	1
E3	0	0	0	1	0
E4	0	0	0	1	1
E5	0	0	1	0	0
E6	0	0	1	0	1
E7	0	0	1	1	0
E8	0	0	1	1	1
E9	0	1	0	0	0
E10	0	1	0	0	1
E11	0	1	0	1	0
E12	0	1	0	1	1
E13	0	1	1	0	0
E14	0	1	1	0	1
E15	0	1	1	1	0
E16	0	1	1	1	1
E17	1	0	0	0	0
E18	1	0	0	0	1
E19	1	0	0	1	0
E20	1	0	0	1	1
E21	1	0	1	0	0
E22	1	0	1	0	1
E23	1	0	1	1	0
E24	1	0	1	1	1
E25	1	1	0	0	0
E26	1	1	0	0	1
E27	1	1	0	1	0
E28	1	1	0	1	1
E29	1	1	1	0	0
E30	1	1	1	0	1
E31	1	1	1	1	0
E32	1	1	1	1	1

Table 2 – All possible outcomes from shock testing the next 5 iPhones

There are 10 total possibilities or combinations, which could also have been calculated using the formula below:

$$C(n, r) = \frac{n!}{r!(n-r)!} = \frac{5!}{3!(5-3)!} = \frac{120}{6(2)!} = \frac{120}{12} = 10$$

The probability of the event occurring and one of these outcomes happening is:

$$P(E8) + P(E12) + P(E14) + P(E15) + P(E20) + P(E22) + P(E23) + P(E26) + P(E27) + P(E29)$$

$$= (0.3 \times 0.3 \times 0.7 \times 0.7 \times 0.7) + (0.3 \times 0.7 \times 0.3 \times 0.7 \times 0.7) + (0.3 \times 0.7 \times 0.7 \times 0.3 \times 0.7) + (0.3 \times 0.7 \times 0.7 \times 0.7 \times 0.3) + (0.7 \times 0.3 \times 0.3 \times 0.7 \times 0.7) + (0.7 \times 0.3 \times 0.7 \times 0.3 \times 0.7) + (0.7 \times 0.3 \times 0.7 \times 0.7 \times 0.3) + (0.7 \times 0.7 \times 0.3 \times 0.3 \times 0.7) + (0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.3) + (0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.3)$$

$$= 10 \times (0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.3)$$

$$= 10 \times 0.03087$$

$$= 0.3087$$

Therefore the probability of exactly 3 out of the next 5 iPhones surviving a test is **0.3087**.

7. Answer to Question 6

In this scenario, the **null hypothesis** is that 63% of the homes sampled in the study will be found to have 3-bedrooms. In other words, the hypothesis proposes the test statistic will equal 0.63, or:

$$H_0: p = 0.63$$

The real estate agent is claiming that exactly 63% of all private residences being built today are 3-bedroom homes, therefore the **alternative hypothesis** is this figure is not 63%, or the test statistic is not 0.63:

$$H_1: p \neq 0.63$$

The alternative hypothesis implies the test statistic is greater or less than 0.63, meaning the critical region occupies both tails of the \hat{p} distribution of the test statistic.

As a result, a two-tailed test will be required to determine whether the final result is statistically significant.

8. Answer to Question 7

Firstly, it is important to check which correlation statistic should be used with the data provided. There is no nominal or ordinal data and the scatterplot in Figure 1 below appears to show a linear relationship between both variables and no significant outliers. Consequently, Pearson's correlation coefficient is preferred over Spearman's coefficient.

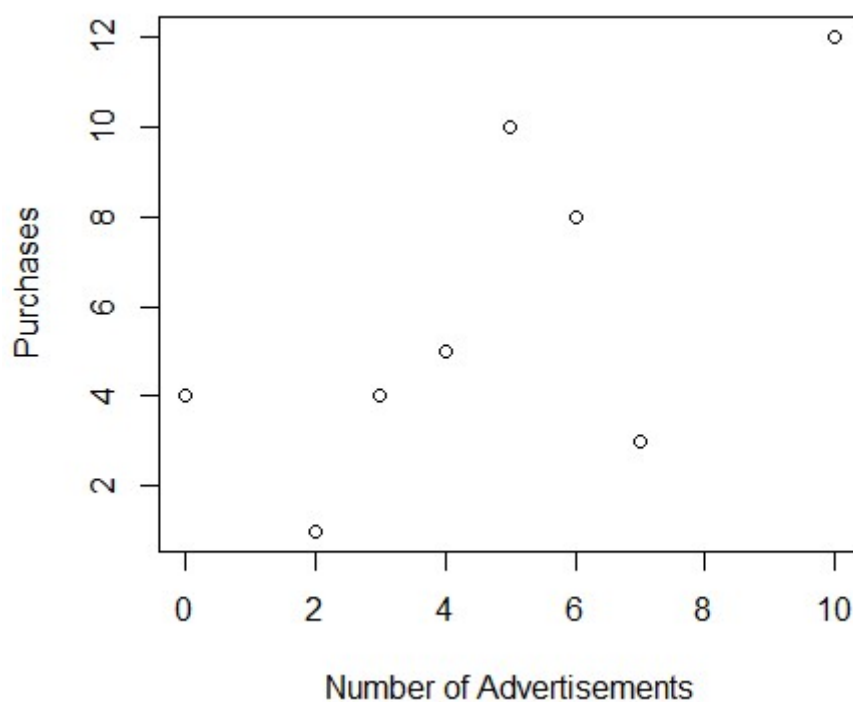


Figure 1: Ad count versus purchases

The `cor` function in R calculates the Pearson coefficient as 0.6790033, as shown in the screenshot in Image 1 below.

```

PE7050 Assessment - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins

Console Terminal Background Jobs
R 4.2.2 ~ /PE7050 Assessment/
> num_ads <- c(10, 7, 6, 5, 4, 3, 2, 0)
> purchases <- c(12, 3, 8, 10, 5, 4, 1, 4)
> q7 <- data.frame(num_ads, purchases)
> print(q7)
  num_ads purchases
1      10        12
2       7         3
3       6         8
4       5        10
5       4         5
6       3         4
7       2         1
8       0         4
> plot(num_ads, purchases, xlab="Number of Advertisements", ylab="Purchases", main="Figure 1: Ad Count vs Purchases")
> cor(num_ads, purchases)
[1] 0.6790033
>

```

Image 1: Using cor function in R to get Pearson's correlation coefficient

As there are relatively few variables in the dataset, the coefficient can also be calculated by hand using the values in the table below:

x	y	$x * y$	x^2	y^2
10	12	120	100	144
7	3	21	47	9
6	8	48	36	64
5	10	50	25	100
4	5	20	16	25
3	4	12	9	16
2	1	2	4	1
0	4	0	0	16
37	47	273	237	375

Table 3 – Values to be used in Pearson's correlation coefficient formula

$$r = \frac{\left(\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \right)}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)} \sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)}}$$

$$r = \frac{\left(273 - \frac{(37)(47)}{8} \right)}{\sqrt{\left(237 - \frac{(37)^2}{8} \right)} \sqrt{\left(375 - \frac{(47)^2}{8} \right)}}$$

$$r = \frac{(273 - 217.375)}{\sqrt{(237 - 171.125)} \sqrt{(375 - 276.125)}}$$

$$r = \frac{55.625}{8.1163415404 * 9.9435909007}$$

$$r = \frac{55.625}{80.7055798881}$$

$$r = 0.6892336326 = 0.69$$

As can be seen, there is a small difference of 0.01 between the calculation above and the figure given in R. This is because R uses a different formula to calculate the coefficient, as shown below:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

However, the difference in the calculations in this instance is only slight. Both statistics indicate there is a moderate, positive correlation between the number of advertisements and the number of products purchased. In other words, greater quantities of the product seem to be purchased when it is advertised more, but not in all cases.

The *cor.test* function in R can be used to explore the correlation further, as shown in screenshot below:

```
[workspace loaded from ~/PE7050 Assessment/.RData]
> cor.test(num_ads, purchases)

Pearson's product-moment correlation

data: num_ads and purchases
t = 2.2655, df = 6, p-value = 0.06406
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04922043  0.93588054
sample estimates:
cor
0.6790033
>
```

Image 2: Using *cor.test* function in R

The p-value for the dataset is 0.064, which means the correlation would not be considered statistically significant if the alpha level (α) was set at 0.05.

9. Answer to Question 8

(a) The *lm*, *plot* and *abline* functions in R were used to estimate the linear regression line and obtain the coefficients, as shown in the screenshot and figure below:

```
> region <- c(1,2,3,4,5,6)
> expenditure <- c(1.5, 4.5, 8.0, 4.0, 2.0, 4.0)
> sales <- c(2.0, 3.0, 4.5, 2.5, 2.0, 5.0)
> df <- data.frame(region, expenditure, sales)
> print(df)
  region expenditure sales
1      1          1.5    2.0
2      2          4.5    3.0
3      3          8.0    4.5
4      4          4.0    2.5
5      5          2.0    2.0
6      6          4.0    5.0
> model <- lm(sales~expenditure, data=df)
> model

Call:
lm(formula = sales ~ expenditure, data = df)

Coefficients:
(Intercept)  expenditure
      1.5818         0.3962

> cor(df$expenditure, df$sales)^2
[1] 0.4992453
> plot(df$expenditure, df$sales, xlab="Ad Expenditure", ylab="Sales", main="Figure 2: Linear Regression - Expenditure vs Sales")
> abline(model, col='blue')
```

Image 3: Using *lm*, *plot* and *abline* functions in R to estimate linear regression between advertising expenditure and sales

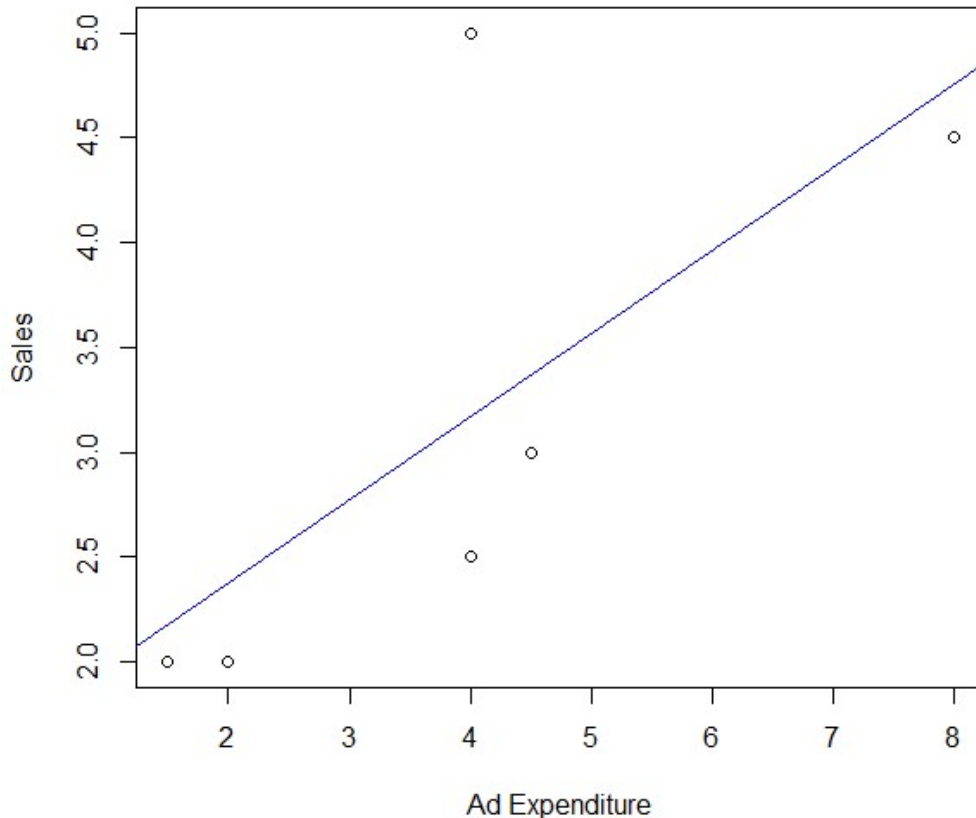


Figure 2: Linear regression – ad expenditure versus sales

As shown in Image 3, the intercept coefficient b_0 is 1.5818 and the slope b_1 is 0.3962. Therefore, the predicted value of sales (\hat{y}) can be obtained using the following equation:

$$\hat{y} = 1.5818 + 0.3962x$$

The coefficients b_0 and b_1 can also be obtained manually, using the table and equations below:

x	y	xy	x^2
1.5	2.0	3.00	2.25
4.5	3.0	13.50	20.25
8.0	4.5	36.00	64.00
4.0	2.5	10.00	16.00
2.0	2.0	4.00	4.00
4.0	5.0	20.00	16.00
24	19	86.5	122.5

Table 4 – Values used to calculate b_0 and b_1

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \frac{\sum y - b_1 \sum x}{n}$$

$$b_1 = \frac{6(86.5) - (24)(19)}{6(122.5) - (24)^2}$$

$$b_0 = \frac{19 - (0.3962264151)(24)}{6}$$

$$b_1 = \frac{519 - 456}{735 - 576}$$

$$b_0 = \frac{9.4912}{6}$$

$$b_1 = \frac{519 - 456}{735 - 576}$$

$$b_0 = 1.5817610063 = 1.5818$$

$$b_1 = \frac{63}{159}$$

$$b_1 = 0.3962264151 = 0.3962$$

The coefficient of determination (r^2) is also shown in Image 3 and is 0.4992. This means that 49.9% of the variation in sales is explained by the variation in advertising expenditure.

(b) If 6.3 thousand pounds was spent on advertising, the predicted sales using the linear regression equation in (a) is 4.07786 million pounds, or £4,077,860. This is calculated as follows:

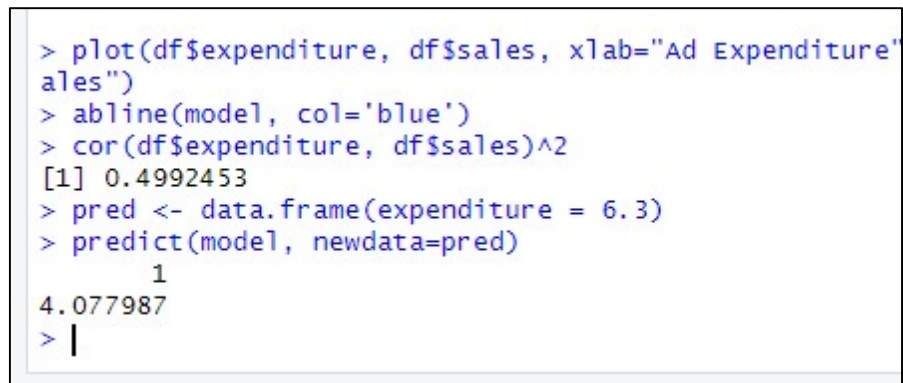
$$\hat{y} = 1.5818 + 0.3962x$$

$$\hat{y} = 1.5818 + 0.3962 * 6.3$$

$$\hat{y} = 1.5818 + 2.49606$$

$$\hat{y} = 4.07786$$

It is also shown calculated in R in the following screenshot:



```
> plot(df$expenditure, df$sales, xlab="Ad Expenditure"
ales")
> abline(model, col='blue')
> cor(df$expenditure, df$sales)^2
[1] 0.4992453
> pred <- data.frame(expenditure = 6.3)
> predict(model, newdata=pred)
      1
4.077987
> |
```

Image 4: Predicting sales in R if advertising expenditure is 6.3

The value of 4.077987 is slightly different from the figure calculated manually because R uses the exact values of b_0 and b_1 to calculate \hat{y} , whereas the coefficients are rounded to 4 decimal places in the equation above.

10. Answer to Question 9

(a) The simplest way to find the number of clusters is by using the general rule, as follows:

$$k = \frac{\sqrt{n}}{2}$$

$$k = \frac{\sqrt{50}}{2}$$

$$k = 3.54$$

An alternative approach is to use the elbow method, as shown in the screenshot and figure below:

```
> library(readxl)
> Q9_Data <- read_excel("~/MSC/Modules/Statistics & Business Intelligence/Assessment/Q9 Data.xlsx")
> view(Q9_Data)
> df <- scale(Q9_Data)
> library(factoextra)
Loading required package: ggplot2
welcome! want to learn more? see two factoextra-related books at https://goo.gl/ve3wBa
> fviz_nbclust(df, kmeans, method = 'wss')
> |
```

Image 5: Using R and the elbow method to find k

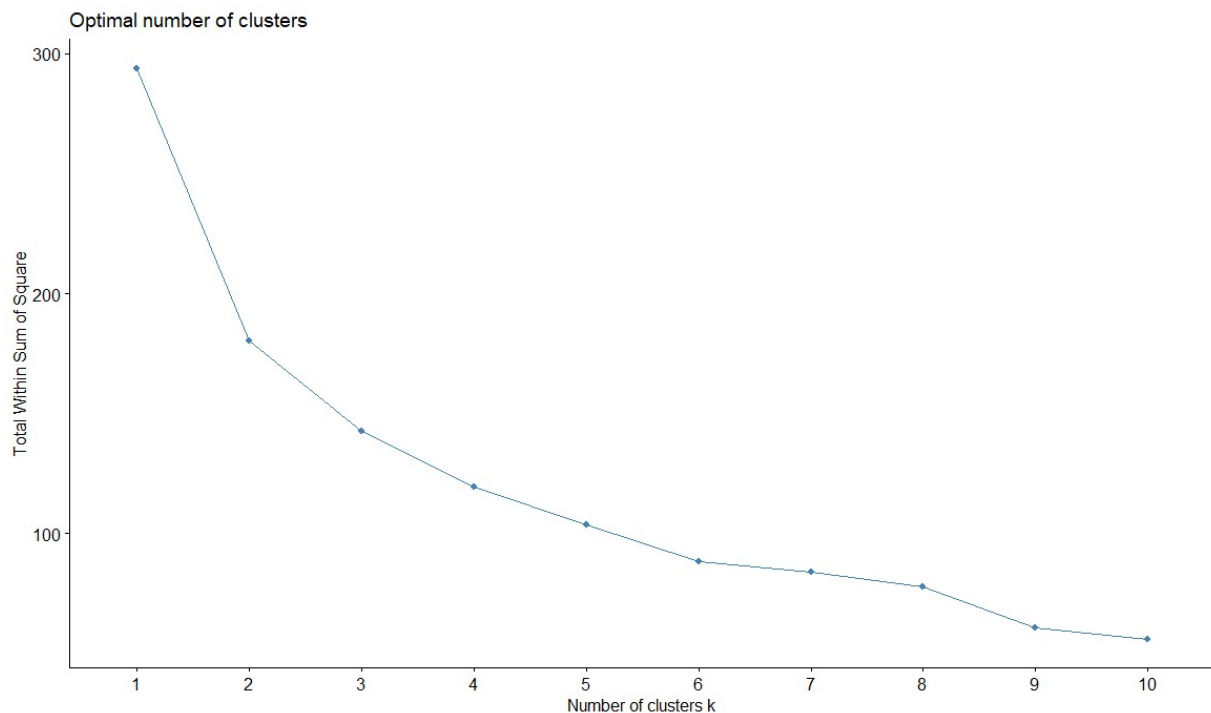


Figure 3: k versus total wss

As can be seen in Figure 3 above, the elbow method indicates the optimal number of clusters is between 2 and 4.

Calculating the average silhouette scores can also be used to find k. In this scenario, k is found in R using the 'fviz_nbclust' function shown in Image 5 and setting the method as 'silhouette'. The optimal number of clusters using this method is 2, as shown in Figure 4 below.

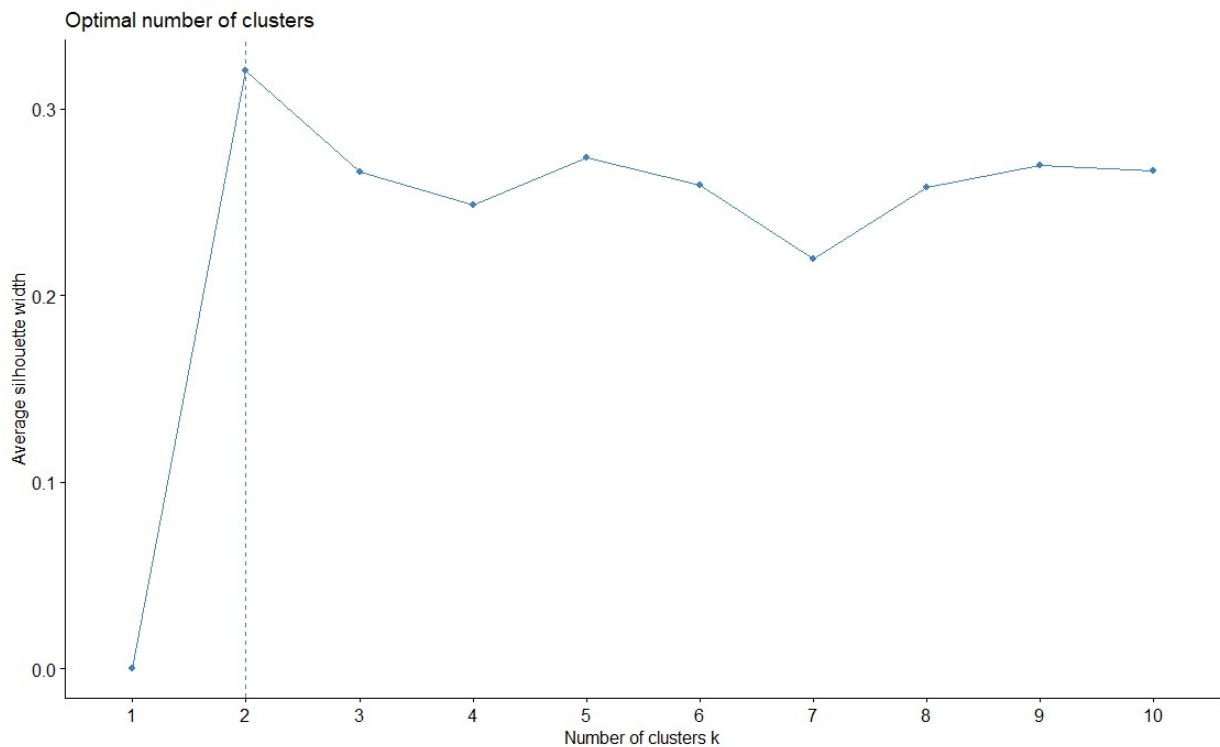


Figure 4: k versus average silhouette width

The three methods used above indicate k could be 2, 3 or 4. In the answer to the second part of this question, k will be set at 3.

(b) The data can be clustered using the 'kmeans' function in R, as shown in the screenshot below:

```
> set.seed(1)
> result <- kmeans(df, 3, nstart = 25)
> result
K-means clustering with 3 clusters of sizes 18, 16, 16

Cluster means:
      Freq250  Freq500  Freq1K  Freq2K  Freq4K  Freq8K
1 -0.04336499 -0.1265009 -0.2049720 -0.84658710 -0.9134670 -0.5402089
2  0.96187748  1.1001395  1.0761029  0.88581574  0.5624829  0.4541029
3 -0.91309187 -0.9578260 -0.8455094  0.06659474  0.4651675  0.1536321

Clustering vector:
[1] 2 2 3 2 3 2 1 1 3 2 3 3 1 3 2 2 1 1 1 3 2 1 3 1 3 1 1 3 3 2 3 2 1 3 3 3 1 1 2 1 2 2 1 1 2 3 2 2 1 1

within cluster sum of squares by cluster:
[1] 48.32203 47.09258 47.35401
(between_SS / total_SS = 51.4 %)

Available components:
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
>
```

Image 6: Using the kmeans function in R to cluster the data

A seed was set at 1 to make the results reproducible and the 'nstart' parameter at 25 as this value is reported to provide a more consistent result (Datanovia, 2023).

Image 6 shows that 3 clusters were produced of sizes 18, 16 and 16 respectively. The dataset contains the Air Conduction (AC) threshold values which represent the faintest sound that the individual can hear in decibels (dB) at the following

frequencies: 250Hz, 500Hz, 1000Hz, 2000Hz, 4000Hz and 8000Hz (Anwar et al., 2010). The mean AC values for all frequencies in each cluster can be calculated in R, as shown in the screenshot below:

```
> aggregate(Q9_Data, by=list(cluster=result$cluster), mean)
  cluster Freq250 Freq500 Freq1K Freq2K Freq4K Freq8K
1       1  40.55556 38.33333 39.16667 41.38889 54.16667 70.27778
2       2  57.81250 60.31250 61.25000 61.87500 73.12500 94.06250
3       3  25.62500 23.43750 28.12500 52.18750 71.87500 86.87500
> |
```

Image 7: Finding the mean AC values in each cluster

This data can be plotted on a line chart to give a visual representation of the hearing ability of the three clusters. However, the data frame shown in Image 7 must be reformatted first using the following code:

```
> df2 <- aggregate(Q9_Data, by=list(cluster=result$cluster), mean)
> df3 <- data.frame(Freq250=unlist(df2, use.names=FALSE))
> df3 <- df3[-(1:3),]
> Cluster <- c("1","2","3","1","2","3","1","2","3","1","2","3","1","2","3",
,"1","2","3")
> df4 <- data.frame(Cluster,df3)
> Freq_Level <- c("250","250","250","500","500","500","1000","1000","1000",
,"2000","2000","2000","4000","4000","4000","8000","8000","8000")
> df5 <- data.frame(df4,Freq_Level)
> colnames(df5)[2] <- "Obs_Mean_Freq"
> df6 <- df5[,c(1,3,2)]
```

When executed, this code produces the following data frame:

	Cluster	Freq_Level	Obs_Mean_Freq
1	1	250	40.55556
2	2	250	57.81250
3	3	250	25.62500
4	1	500	38.33333
5	2	500	60.31250
6	3	500	23.43750
7	1	1000	39.16667
8	2	1000	61.25000
9	3	1000	28.12500
10	1	2000	41.38889
11	2	2000	61.87500
12	3	2000	52.18750
13	1	4000	54.16667
14	2	4000	73.12500
15	3	4000	71.87500
16	1	8000	70.27778
17	2	8000	94.06250
18	3	8000	86.87500

Image 8: Mean AC values in each cluster in reformatted data frame

The British Society of Audiology (2023) classifies hearing loss into the following categories based on the range of AC values that can be heard:

- **Normal hearing** – Less than 20dB.
- **Mild hearing loss** – 20 to 40dB.
- **Moderate hearing loss** – 41 to 70dB.
- **Severe hearing loss** – 71 to 95dB.
- **Profound hearing loss** – Over 95dB.

These ranges can be plotted on a line chart along with the AC values in Image 8 using the following code:

```
> ggplot(df6, aes(x = Freq_Level, y = Obs_Mean_Freq, shape = Cluster))+  
+ ylim(0,110)+  
+ geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=20, ymax=40), fill="grey", alpha  
=0.025)+  
+ geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=40, ymax=70), fill="grey", alpha  
=0.050)+  
+ geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=70, ymax=95), fill="grey", alpha  
=0.075)+  
+ geom_rect(aes(xmin=-Inf, xmax=Inf, ymin=95, ymax=Inf), fill="grey", alph  
a=0.1)+  
+ geom_line() +  
+ geom_point(size = 2)+  
+ xlab("Hearing Frequency (Hz)")+  
+ ylab("Mean Air Conduction Thresholds (dB)")+  
+ annotate("text", x = 8000, y = 10, label = "Normal Hearing", hjust = 1,  
fontface =2)+  
+ annotate("text", x = 8000, y = 30, label = "Mild Hearing Loss", hjust =  
1, fontface =2)+  
+ annotate("text", x = 8000, y = 55, label = "Moderate Hearing Loss", hjus  
t = 1, fontface =2)+  
+ annotate("text", x = 250, y = 83, label = "Severe Hearing Loss", hjust =  
0, fontface =2)+  
+ annotate("text", x = 250, y = 105, label = "Profound Hearing Loss", hjus  
t = 0, fontface =2)
```

When executed, the code produces the line chart on the following page in Figure 5. This chart shows the hearing ability of the participants in each cluster deteriorates as the hearing frequency increases. It also shows that each cluster suffers from hearing loss, although the extent of this loss varies between the clusters.

Cluster 1 contains individuals who suffer from moderate hearing loss at all frequencies. The loss suffered by those in cluster 2 is greater as it ranges from moderate to severe, bordering on profound at the higher frequencies. The individuals in cluster 3 have the best hearing at the lower frequencies, where the loss is mild. However, hearing ability in this cluster deteriorates more rapidly as the frequencies increase and is classed as severe at 8000Hz.

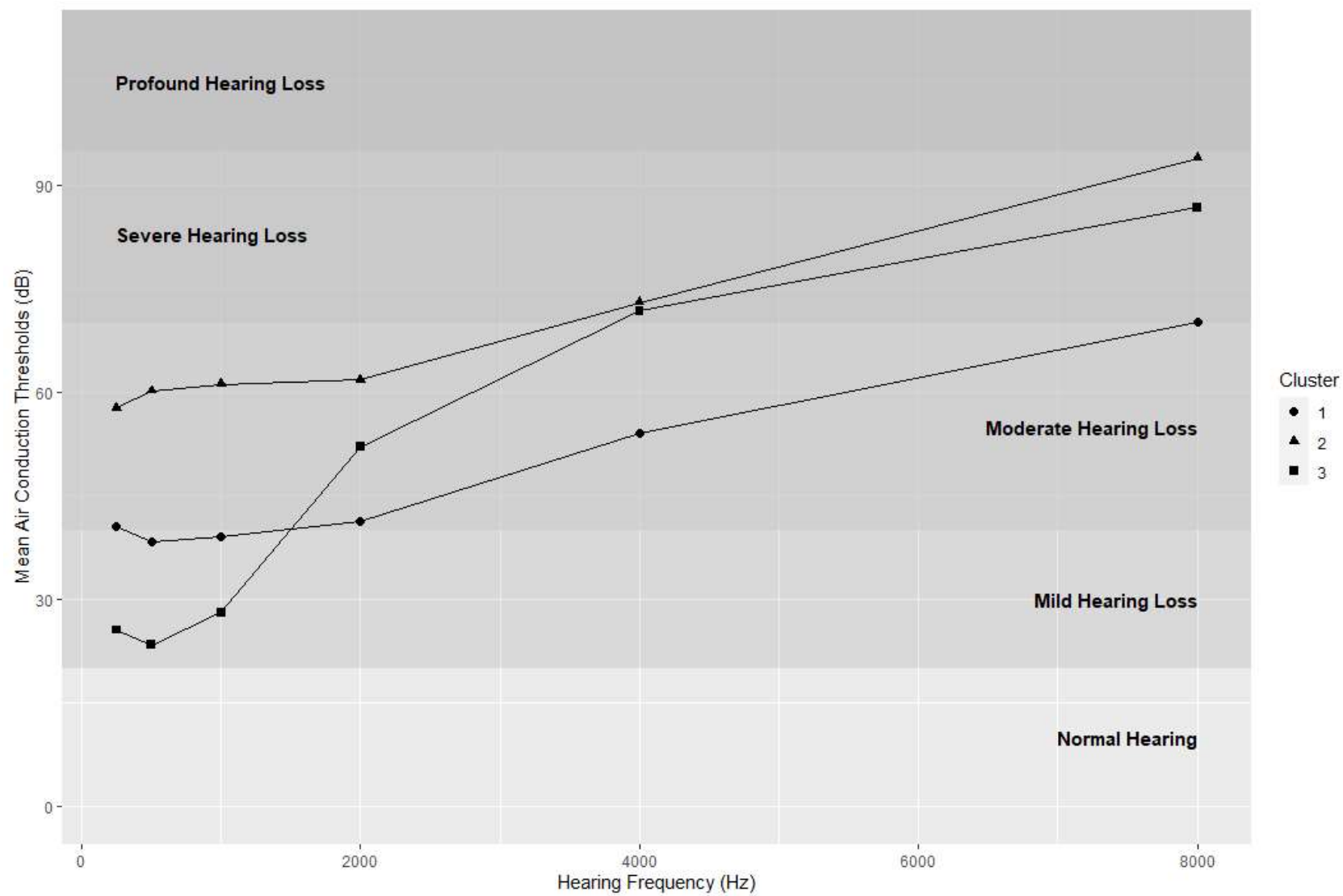


Figure 5: Mean AC values for each cluster at the frequencies 250, 500, 1000, 2000, 4000 and 8000 Hz

Appendix A – Bibliography

- Anwar, M.N. *et al.* (2010) 'Clustering Audiology Data', *19th Machine Learning Conference of Belgium and The Netherlands*, Leuven, Belgium, 27-28 May 2010. Available at: https://dtai.cs.kuleuven.be/events/Benelearn2010/submissions/benelearn2010_submission_7.pdf (Accessed: 14th March 2023).
- Bennett, D. A. (2001) 'How Can I deal with Missing Data in My Study', *Australian and New Zealand Journal of Public Health*, 25(5), pp. 387-479. Available at: <https://doi.org/10.1111/j.1467-842X.2001.tb00294.x> (Accessed: 5th February 2023).
- British Society of Audiology (2023) *FAQs: A useful set of Audiology and Hearing Loss Frequently Asked Questions*. Available at: <https://www.thebsa.org.uk/public-engagement/faqs/> (Accessed: 14th March 2023).
- Carver, R. (2017) *Preparing Data for Analysis with JMP*. Cary, NC: SAS Institute.
- Datanovia (2023) *Partitional Clustering in R: The Essentials*. Available at: <https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorithm-and-practical-examples/> (Accessed: 13th March 2023).
- Dodd, C. (2022) *CompTIA Data+: DAO-001 Certification Guide*. Birmingham: Packt Publishing.
- Gutierrez, D. (2018) *Unsupervised Learning: Evaluating Clusters*. Available at: <https://opendatascience.com/unsupervised-learning-evaluating-clusters/> (Accessed: 14th March 2023).
- Hayes, A. (2022) *Mode: What It Is in Statistics and How to Calculate It*. Available at: <https://www.investopedia.com/terms/m/mode.asp> (Accessed: 22nd January 2023).
- Kovchegov, Y. (2022) 'A New Life of Pearson's Skewness', *Journal of Theoretical Probability*, 35, pp. 2896–2915. Available at: <https://doi.org/10.1007/s10959-021-01149-7> (Accessed: 5th February 2023).
- Li, P., Stuart, E.A. and Allison, D.B. (2015) 'Multiple Imputation: A Flexible Tool for Handling Missing Data', *Journal of the American Medical Association*, 314(18), pp. 1966–1967. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4638176/> (Accessed: 5th February 2023).
- Liu, P. *et al.* (2005) 'An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset', *Advanced Data Mining and Applications*, Wuhan, China, 22-24 July. New York: Springer. pp. 583-590. Available at: https://doi.org/10.1007/11527503_69 (Accessed: 5th February 2023).
- Menon, K. (2023) *The Complete Guide to Skewness and Kurtosis*. Available at: <https://www.simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis> (Accessed: 22nd January 2023).

Nguyen, C.D., Carlin, J.B. and Lee, K.J. (2017) 'Model checking in multiple imputation: an overview and case study', *Emerging Themes in Epidemiology*, 14(8). Available at: <https://doi.org/10.1186/s12982-017-0062-6> (Accessed: 5th February 2023).

Sainani, K.L. (2015) 'Dealing With Missing Data', *American Academy of Physical Medicine and Rehabilitation*, 7(9), pp. 905-1019. Available at: <https://doi.org/10.1016/j.pmrj.2015.07.011> (Accessed: 5th February 2023).

Salgado, C.M. *et al.* (2016) 'Missing Data', in *Secondary Analysis of Electronic Health Records*. Cambridge, MA: Springer, pp. 143-161. Available at: https://doi.org/10.1007/978-3-319-43742-2_13 (Accessed: 5th February 2023).

Stats NZ (2022) *Exploring the benefits and costs of multiple imputation for the 2018 Census*. Available at: <https://www.stats.govt.nz/methods/exploring-the-benefits-and-costs-of-multiple-imputation-for-the-2018-census/> (Accessed: 5th February 2023).

Sterne, J.A.C. *et al.* (2009) 'Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls', *British Medical Journal*, 338(b2393). Available at: <https://doi.org/10.1136/bmj.b2393> (Accessed: 5th February 2023).

Takahashi, M. & Ito, T. (2012) 'Multiple Imputation of Turnover in Edinet Data: Toward the Improvement of Imputation for the Economic Census', *Conference of European Statisticians: Work Session on Statistical Data Editing*, Oslo, Norway, 24-26 September. Available at: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/35_Japan.pdf (Accessed: 5th February 2023).

Urdan, T.C. (2017) *Statistics in Plain English*. 4th edn. New York: Routledge.

Annex A - Question 10 Report

Contents

Executive Summary	23
Introduction	23
Data Collection and Preparation	24
Data Analysis	26
Bibliography	31
 Appendix A – Final Dataset	 33
Appendix B – R Studio Code and Calculations	37

Executive Summary

In 2014, the average speed of all vehicles using motorways in Great Britain was 61.86 miles per hour (mph). This study tested the hypothesis that the average speed on the M1 motorway was lower than this figure, using data collected in January 2023. It also tested whether the average speed on the northbound and southbound lanes of the M1 differed during this time period.

Stratified random sampling was used to test the hypotheses and data was collected from 144 sites on the M1 using the National Highways WebTRIS service. The findings showed the average speed was not lower than 61.86 mph, nor was there any difference between the lanes. On the contrary, there is a 0.95 probability that the mean speed of all vehicles using the M1 in January 2023 was between 62.18 and 63.02 mph. This suggests that the average speed in general may be higher than 61.86 mph, however further research using data collected over a longer timeframe is required before this result can be generalised.

Introduction

The M1 motorway is 193.6 miles long and stretches from Edgware in North London to the outskirts of Leeds (RAC, 2019). It is one of the few arterial routes that connects the North and South of England, therefore it is used frequently by logistics and delivery companies to move goods across the UK.

In 2014, the average speed of all types of vehicle using motorways in Great Britain was 61.86 mph (Statista, 2015)¹. Since then, various events have occurred which indicate the average speed on UK motorways may have fallen. For instance, speed limits are being trialled in various locations to assess whether lower speeds result in improved air quality (National Highways, 2023). Additionally, it has been reported that motorists are driving at slower speeds to increase fuel efficiency and save money in the current cost of living crisis (Ruff, 2022).

On the M1, recent changes in legislation are also likely to have affected the average vehicle speed. The M1 Motorway (Junction 2) (50 Miles Per Hour Speed Limit) Regulations 2020 and M1 Motorway (Junctions 13 to 16) (Variable Speed Limits) Regulations 2020 both lowered the speed limit on sections of this road. Consequently, this study investigated whether the average vehicle speed on the M1 is lower than 61.86 mph by testing the following hypotheses:

$$H_0: \mu = 61.86 \text{ mph}$$

$$H_1: \mu < 61.86 \text{ mph}$$

¹ This figure was calculated by taking the mean of the average speeds for each vehicle type reported by Statista.

The figures above refer to the average speed on both the northbound and southbound lanes of the M1. It is possible that this speed varies depending on the direction of travel. Therefore, the study also tested the following hypotheses:

$$H_0: \mu_n = \mu_s$$

$$H_1: \mu_n \neq \mu_s$$

Data Collection and Preparation

The National Highways WebTRIS service² collects data from sites located across the road network and makes it freely available online. In this study, Microsoft Excel and Power Query were used to obtain a list of all of these sites from the WebTRIS Application Programming Interface (API)³. The list was then filtered to display northbound and southbound sites on the M1 which are currently active and collecting data⁴. The data was then filtered further to remove sites classified as slip roads⁵, which left 599 northbound sites and 576 southbound sites available for sampling.

Stratified random sampling was used to select the sample as it ensured both northbound and southbound sites were adequately represented. The sites were divided into separate lists according to their direction of travel and each one was allocated a random number using the MS Excel RAND function (Microsoft, 2023a). In each list, the sites were then sorted in order from the smallest to largest value and the first 100 were sampled. This approach was taken to minimise the risk of selection bias and enable inferences to be made about the wider population (McCombes, 2022).

In total, 200 samples were selected. This number was chosen as bigger sample sizes result in smaller standard errors, which means statistically significant results are more likely (Urdan, 2017, p.68). Moreover, it quickly became apparent that a considerable number of sites, although active, had a low level of data availability⁶. Therefore, a larger sample size was selected in anticipation of many sites not being able to provide sufficient data.

Looking through the list of samples, data availability seemed to be highest for the month of January 2023. As a result, daily speed data was collected for this time period due to it being both current and reliable.

² This service can be accessed at <https://webtris.nationalhighways.co.uk/>.

³ This list was obtained using the URL <http://webtris.nationalhighways.co.uk/api/v1.0/sites>.

⁴ This list and most of the other data described in this report cannot be provided in an appendix as it is too voluminous. Therefore, it has been made available to the module tutors in a folder in the author's University OneDrive account called [PE7050 Q10 Assessment Data](#) (click link to access). The data is in spreadsheets, which also contain the formulas used to calculate imputation values.

⁵ Site references ending in J, K, L or M represent slip roads whereas those ending in A or B are assigned to the motorway.

⁶ The availability of data at each site is represented by a colour on the WebTRIS service, as explained in the Frequently Asked Questions (FAQs) section of the site (<https://webtris.nationalhighways.co.uk/Home/Faqs>).

Out of the 100 northbound sites, 31 had less than 76% of the data available for the month. This figure was 25 for the southbound sites. It was important to check whether any unseen, systemic factors caused lower data availability at these sites, such as a batch of faulty traffic sensors being used in a small number of areas. The locations of these sites were therefore plotted on a map, as shown in figure 1 below.

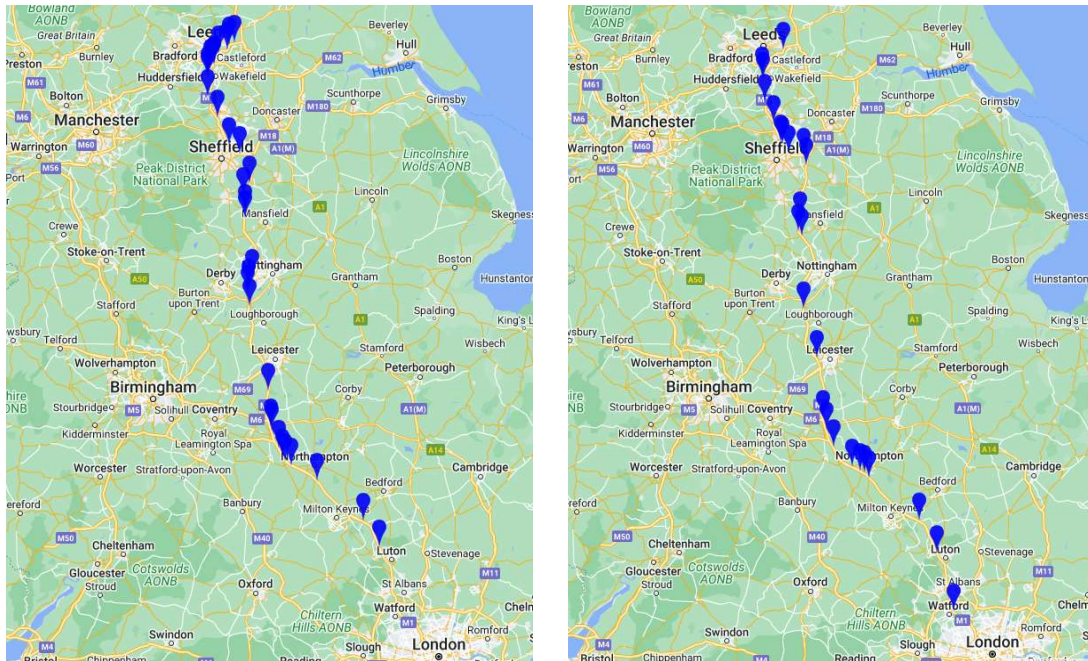


Figure 1 – Location of sites with less than 76% data availability on the northbound (l) and southbound (r) lanes, plotted using MI Map Tools: GeoPlotter (2023)

The maps above show the sites were spread across the length of the M1 and there did not appear to be any patterns present in the data. As such, the data was considered Missing Completely At Random (MCAR) and deleted without fear of introducing bias into the sample (Wiley & Wiley, 2019). This meant the final sample size was 144, comprising 69 northbound and 75 southbound sites. The list-wise deletion method for handling missing data was deemed acceptable in this instance as the size of the sample remained large and no significant loss of statistical power was anticipated (Liu, 2015).

The WebTRIS batch reporting tool was used to download the daily speed data in January 2023 for each site in the sample. This data was received in multiple csv files, therefore it was combined into one northbound and one southbound MS Excel spreadsheet using Power Query (Microsoft, 2023b).

In each spreadsheet, there were 96 data points per day for each site which represented 15-minute periods throughout the day. The average speed of all vehicles registered by the sensor during each period was provided, along with other

data⁷. In total, there were 205,344 and 223,200 data points for the northbound and southbound sites.

Speed data was missing in 6,666 (3.2%) of the northbound data points and there were 7,304 (3.3%) missing values amongst the southbound sites. To resolve this issue, the median speed of each 15-minute period was calculated using the data points which contained values. The data points with missing values were subsequently assigned the imputed median speed value of the corresponding timeframe.

Single imputation was selected to deal with this issue as it was easy to implement and the number of missing values in the sample was relatively low, meaning the impact on the standard deviation was likely to be negligible (Sainani, 2015). In addition, the median was chosen as the imputation value instead of the mean because it is less susceptible to the influence of outliers (Bonthu, 2022).

Finally, pivot tables were used to aggregate the data points to produce one average speed figure for each site. This figure represented the average or mean speed of all vehicles passing through the site during the entire month of January 2023. The final dataset produced can be found in Appendix A.

Data Analysis

R Studio was used to conduct data analysis on the whole dataset, including each individual strata. The code and calculations underpinning this analysis can be found in Appendix B.

Measures of central tendency and variability were calculated first. The statistics that were produced are shown below in table 1.

Direction	Min	Q1	Median	Mean	Mode	Q3	Max	IQR	SD
Northbound	56.20	61.73	62.88	62.73	62.99	64.06	69.80	2.33	2.54
Southbound	56.48	61.30	63.18	62.48	63.80	63.92	69.17	2.62	2.58
Both	56.20	61.39	63.00	62.60	63.79	63.93	69.80	2.54	2.55

Table 1 – Summary descriptive statistics for M1 speed data

The statistics for the whole dataset and both stratas are very similar. The data is largely clustered around the mean given the small standard deviation (SD) in each case. Furthermore, the median, mean and mode are largely the same, which indicates the data is normally distributed (Steinberg & Price, 2020).

⁷ This data included the total number of vehicles passing the sensor in the 15-minute window and their length classification.

Levels of skewness and kurtosis in the data were also calculated, as shown below in table 2.

Direction	Skewedness	Kurtosis
Northbound	0.03549786	3.459380
Southbound	-0.73629859	3.323185
Both	-0.37565453	3.462557

Table 2 – Level of skewedness and kurtosis in M1 speed data

The results confirm that the northbound strata closely follows a normal distribution as the skewness statistic (0.04) is negligible, meaning the data is essentially symmetrical (Gawali, 2023). The southbound strata has a slight, negative skew (-0.74) which affects the level of skewness in the whole dataset (-0.38), however this is relatively minor. The kurtosis values are close to 3 in all three cases, therefore the distributions are considered mesokurtic with medium-sized tails (Turney, 2022).

The distribution of the whole dataset and both stratas are represented visually in the box plots and histograms below. The slight, negative skew in the southbound strata can be seen in Figure 4 and the outline of the normal distribution is evident in each histogram.

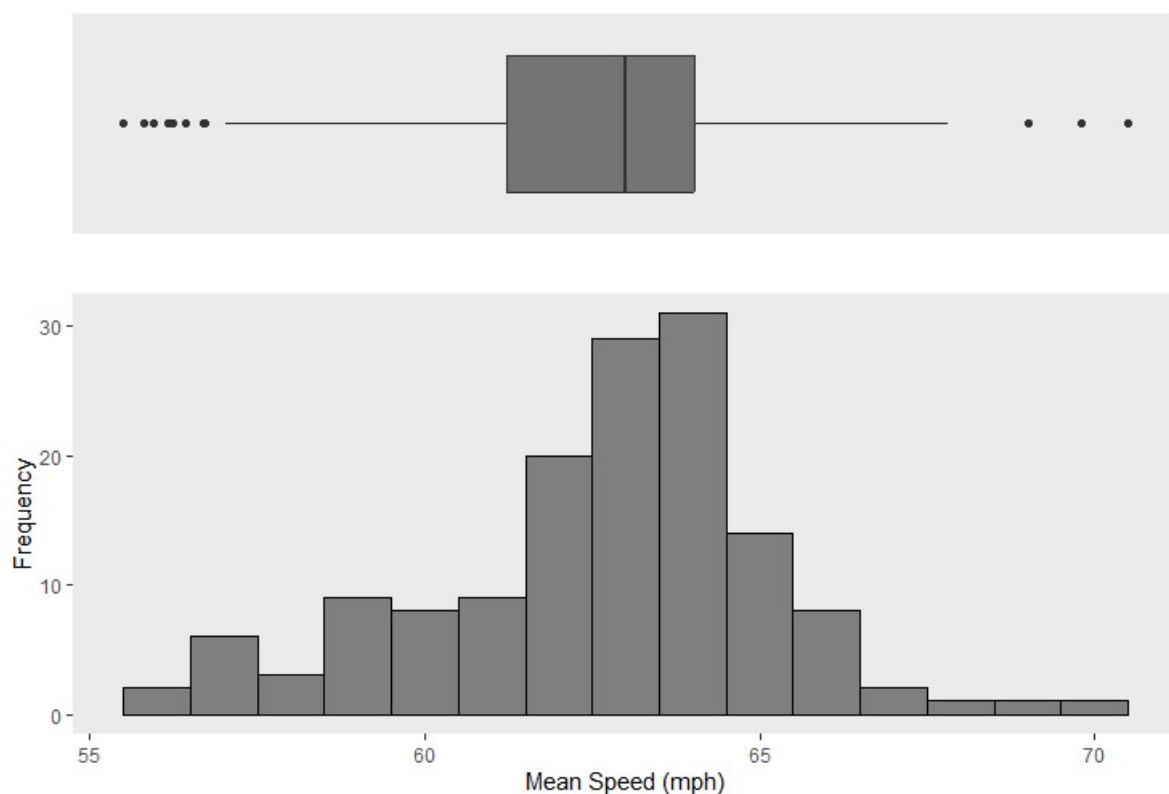


Figure 2 – Box plot and histogram for whole dataset

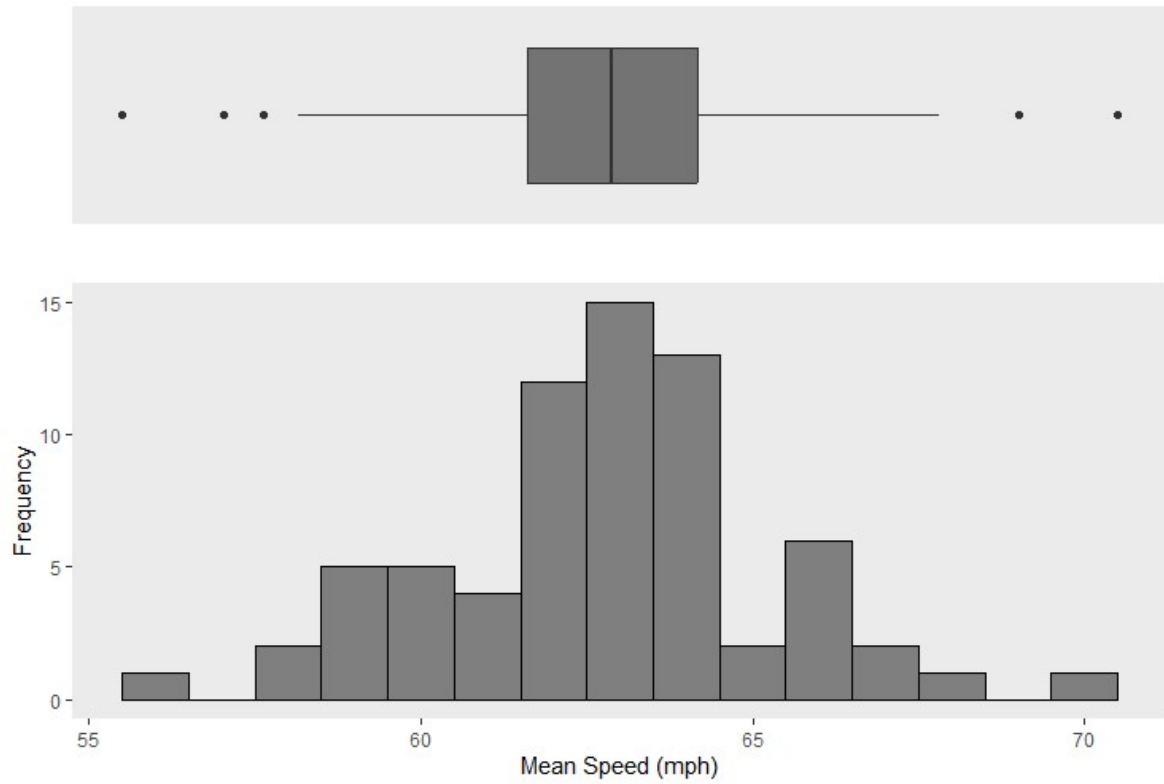


Figure 3 – Box plot and histogram for northbound strata

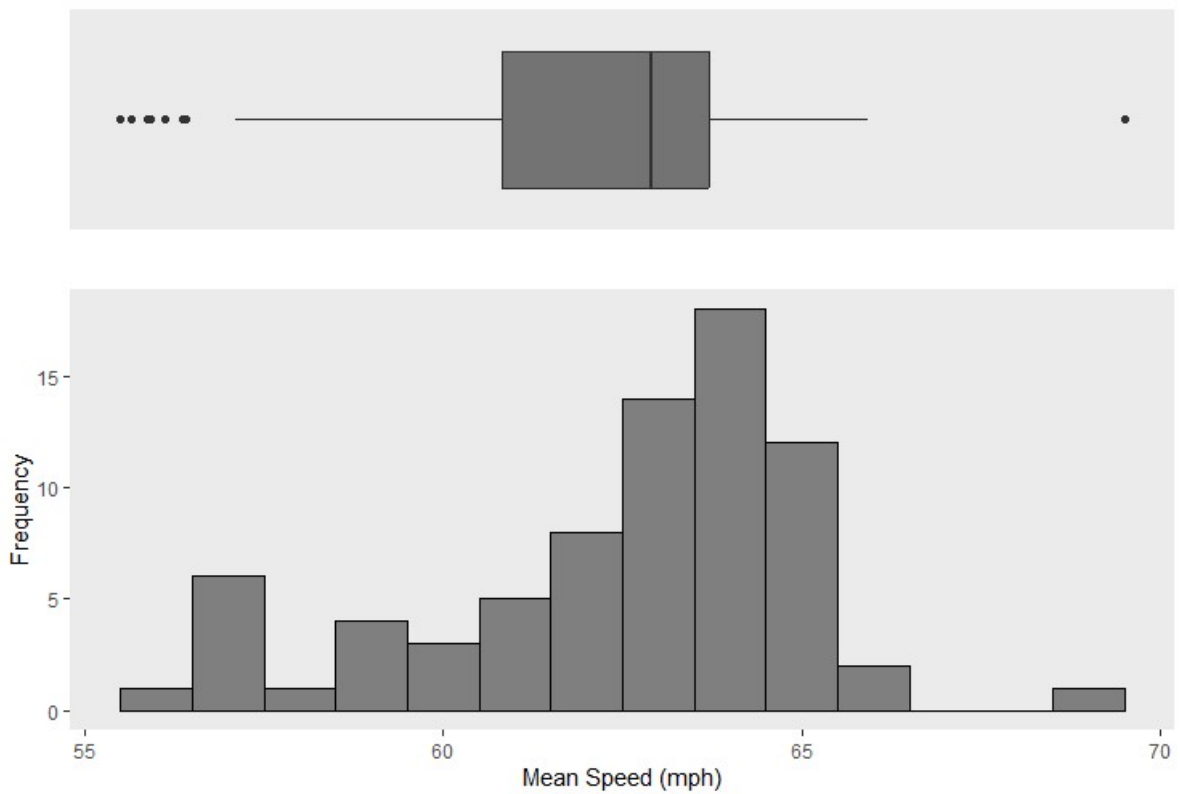


Figure 4 – Box plot and histogram for southbound strata

Hypothesis testing was conducted to determine whether the average vehicle speed on the M1 is lower than 61.86 mph. Although the population variance was unknown, a z-test was performed because the sample size was greater than 30 (Newcastle University, 2023). Moreover, as a parametric test it offered more statistical power than nonparametric methods (Minitab, 2015). The result of the hypothesis test is shown in the screenshot below.

```
> z.test(m1data$`Mean Speed (mph)` ,alternative='less',mu=61.86, sigma.x=overall_sd,conf.level=0.95)

One-sample z-Test

data: m1data$`Mean Speed (mph)`
z = 3.4573, p-value = 0.9997
alternative hypothesis: true mean is less than 61.86
95 percent confidence interval:
 NA 62.94591
sample estimates:
mean of x
62.59583
```

Image 1 – Z-test to determine whether the population mean speed is less than 61.86 mph

The p-value is 0.9997 and more than the alpha level (0.05), therefore the null hypothesis ($H_0: \mu = 61.86 \text{ mph}$) is retained. In fact, the findings of the study suggest that the true average speed on the M1 may be greater than 61.86 mph, given that the sample mean was 62.6 mph. To investigate this further, the standard error was calculated for the whole dataset and each strata, as shown below in Table 3.

Direction	SE
Northbound	0.3059248
Southbound	0.2975503
Both	0.2128329

Table 3 – Standard error values for mean speed

The results show that the mean speed of the random sample of 144 M1 sites was 62.6 mph \pm 0.21 (SE). For the 69 northbound sites, the mean speed was 62.73 mph \pm 0.31 (SE) and for the 75 southbound sites it was 62.48 mph \pm 0.30 (SE).

The SE for the whole dataset was subsequently used to construct a 95% confidence interval, in order to estimate the population mean. The calculations were done manually as follows:

$$\text{Lower limit} = \bar{x} - (1.96 \times SE) = 62.6 - (1.96 \times 0.2128329) = 62.18$$

$$\text{Upper limit} = \bar{x} + (1.96 \times SE) = 62.6 + (1.96 \times 0.2128329) = 63.02$$

The results show there is a 0.95 probability that the mean speed of all vehicles using the M1 in January 2023 was between 62.18 and 63.02 mph. This indicates that the average speed on the M1 in general may be greater than 61.86 mph, however more research is needed to determine whether the results of this study are generalisable. This is because traffic speed on the M1 is likely to vary across the year due to various factors, such as the onset of the school holidays. Therefore, any future study should collect data over a 12-month period at a minimum.

A second hypothesis test was conducted to determine whether the average speed on the northbound lane of the M1 differed from the southbound lane. The result of the z-test is shown in the screenshot below.

```
> z.test(x=nb_data$`Mean Speed (mph)`, y=sb_data$`Mean Speed (mph)`, mu=0, sigma.x=nb_sd, sigma.y=sb_sd)

      Two-sample z-Test

data:  nb_data$`Mean Speed (mph)` and sb_data$`Mean Speed (mph)`
z = 0.58927, p-value = 0.5557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5849608  1.0879173
sample estimates:
mean of x mean of y
 62.72681  62.47533
```

Image 2 – Z-test to determine whether the mean speed differs between each lane of the M1

The p-value (0.5557) is more than the alpha level (0.05), therefore the null hypothesis ($H_0: \mu_n = \mu_s$) is retained. In other words, the average speed on the northbound and southbound lanes of the M1 in January 2023 was not significantly different.

Bibliography

Bonthu, H. (2022) *Detecting and Treating Outliers | Treating the odd one out!* Available at: <https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/> (Accessed: 25th March 2023).

Gawali, S. (2023) *Skewness and Kurtosis: Quick Guide (Updated 2023)*. Available at: <https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/> (Accessed: 25th March 2023).

Liu, X. (2015) *Methods and Applications of Longitudinal Data Analysis*. Available at: <https://doi.org/10.1016/C2013-0-13082-6> (Accessed: 20th March 2023).

M1 Motorway (Junction 2) (50 Miles Per Hour Speed Limit) Regulations 2020 (SI 2022/1136). Available at: <https://www.legislation.gov.uk/uksi/2020/1136/made> (Accessed: 19th March 2023).

M1 Motorway (Junctions 13 to 16) (Variable Speed Limits) Regulations 2020 (SI 2022/956). Available at: <https://www.legislation.gov.uk/uksi/2020/956/made> (Accessed: 19th March 2023).

McCombes, S. (2022) *Sampling Methods | Types, Techniques, & Examples*. Available at: <https://www.scribbr.co.uk/research-methods/sampling/> (Accessed: 25th March 2023).

Microsoft (2023a) *RAND function*. Available at: <https://support.microsoft.com/en-us/office/rand-function-4cbfa695-8869-4788-8d90-021ea9f5be73> (Accessed: 25th March 2023).

Microsoft (2023b) *Import data from a folder with multiple files (Power Query)*. Available at: <https://support.microsoft.com/en-us/office/import-data-from-a-folder-with-multiple-files-power-query-94b8023c-2e66-4f6b-8c78-6a00041c90e4#:~:text=Combine%20and%20Transform%20Data%20To,select%20Combine%20%3E%20Combine%20and%20Load> (Accessed: 25th March 2023).

MI Map Tools: GeoPlotter (2023) *M1 sites with less than 76% data availability*. Available at: <https://mobisoftinfotech.com/tools/plot-multiple-points-on-map/> (Accessed: 20th March 2023).

Minitab (2015) *Choosing Between a Nonparametric Test and a Parametric Test*. Available at: <https://blog.minitab.com/en/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test> (Accessed: 25th March 2023).

National Highways (2023) *Air quality speed limit trials*. Available at: <https://nationalhighways.co.uk/our-work/environment/air-quality-and-noise/air-quality/air-quality-speed-limit-trials/> (Accessed: 25th March 2023).

Newcastle University (2023) *Z-tests and T-tests (Business)*. Available at: <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths->

<resources/business/hypothesis-tests/z-tests-and-t-tests.html> (Accessed: 25th March 2023).

RAC (2019) *M1 Traffic News*. Available at: <https://www.rac.co.uk/route-planner/traffic-news/m1/> (Accessed: 25th March 2023).

Ruff, M. (2022) *UK driving behaviours changing due to cost-of-living crisis, survey finds*. Available at: <https://garagewire.co.uk/news/uk-driving-behaviours-changing-due-to-cost-of-living-crisis-survey-finds/> (Accessed: 25th March 2023).

Sainani, K.L. (2015) 'Dealing With Missing Data', *American Academy of Physical Medicine and Rehabilitation*, 7(9), pp. 905-1019. Available at: <https://doi.org/10.1016/j.pmrj.2015.07.011> (Accessed: 21st March 2023).

Statista (2015) *Average speed on motorways in Great Britain in 2014, by vehicle type*. Available at: <https://www.statista.com/statistics/303425/average-speed-on-motorways-in-the-uk/> (Accessed: 25th March 2023).

Steinberg, W.J. & Price, M. (2020) *Statistics Alive!* 3rd edition. Thousand Oaks, California: Sage.

Turney, S. (2022) *What Is Kurtosis? | Definition, Examples & Formula*. Available at: https://www.scribbr.com/statistics/kurtosis/?_ga=2.151976506.247020522.1679693520-1138181901.1674235798 (Accessed: 25th March 2023).

Urdan, T.C. (2017) *Statistics in Plain English*. 4th edn. New York: Routledge.

Wiley, M. & Wiley, J.F. (2019) *Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization*. Berkeley, CA: Apress.

Appendix A – Final Dataset

The final dataset below was imported into RStudio under the file name **m1data**.

Site	Direction	Mean Speed (mph)
M1/2220A	Northbound	64.14
M1/2400A	Northbound	60.54
M1/2405A	Northbound	61.89
M1/2410A	Northbound	59.99
M1/2420A	Northbound	60.38
M1/2427A	Northbound	60.69
M1/2533A	Northbound	58.80
M1/2588A	Northbound	61.76
M1/2618A	Northbound	60.70
M1/2681A	Northbound	58.12
M1/2697A	Northbound	58.65
M1/2745A	Northbound	64.13
M1/2756A	Northbound	65.78
M1/2792A	Northbound	63.18
M1/2850A	Northbound	62.88
M1/2864A	Northbound	62.01
M1/2916A	Northbound	64.51
M1/2933A	Northbound	63.70
M1/2959A	Northbound	59.37
M1/3010A	Northbound	58.62
M1/3044A	Northbound	61.39
M1/3079A	Northbound	56.20
M1/3130A	Northbound	62.85
M1/3166A	Northbound	62.00
M1/3336A	Northbound	63.04
M1/3338A	Northbound	63.00
M1/3519A	Northbound	63.69
M1/3652A	Northbound	63.57
M1/3685A	Northbound	62.82
M1/3753A	Northbound	66.41
M1/3780A	Northbound	61.82
M1/3785A	Northbound	63.12
M1/3849A	Northbound	66.50
M1/3857A	Northbound	65.58
M1/3892A	Northbound	62.99
M1/3915A	Northbound	65.74
M1/3977A	Northbound	64.29
M1/4053A	Northbound	63.91
M1/4069A	Northbound	62.99
M1/4095A	Northbound	62.80

M1/4113A	Northbound	63.30
M1/4136A	Northbound	68.45
M1/4162A	Northbound	67.11
M1/4169A	Northbound	62.83
M1/4172A	Northbound	63.79
M1/4209A	Northbound	65.29
M1/4279A	Northbound	61.73
M1/4294A	Northbound	67.37
M1/4315A	Northbound	64.21
M1/4411A	Northbound	63.79
M1/4484A	Northbound	63.21
M1/4497A	Northbound	59.81
M1/4546A	Northbound	62.11
M1/4561A	Northbound	60.47
M1/4577A	Northbound	57.58
M1/4589A	Northbound	58.60
M1/4612A	Northbound	62.43
M1/4634A	Northbound	62.74
M1/4657A	Northbound	61.90
M1/4745A	Northbound	64.12
M1/4767A	Northbound	62.26
M1/4777A	Northbound	61.98
M1/4792A	Northbound	65.61
M1/4843A	Northbound	64.06
M1/4856A	Northbound	63.55
M1/5011A	Northbound	63.09
M1/5021A	Northbound	60.09
M1/5073A	Northbound	62.32
M1/5191A	Northbound	69.80
M1/2220B	Southbound	63.46
M1/2331B	Southbound	63.53
M1/2338B	Southbound	62.85
M1/2354B	Southbound	61.07
M1/2358B	Southbound	57.27
M1/2424B	Southbound	60.83
M1/2467B	Southbound	59.38
M1/2482B	Southbound	57.93
M1/2502B	Southbound	57.05
M1/2563B	Southbound	56.61
M1/2579B	Southbound	56.48
M1/2595B	Southbound	58.72
M1/2607B	Southbound	59.04
M1/2681B	Southbound	62.08
M1/2726B	Southbound	59.39
M1/2850B	Southbound	62.27

M1/2901B	Southbound	62.56
M1/2945B	Southbound	60.25
M1/2969B	Southbound	57.31
M1/3017B	Southbound	56.81
M1/3058B	Southbound	56.86
M1/3219B	Southbound	64.78
M1/3248B	Southbound	63.72
M1/3273B	Southbound	63.80
M1/3323B	Southbound	64.84
M1/3799B	Southbound	63.69
M1/3863B	Southbound	63.92
M1/3892B	Southbound	63.15
M1/3898B2	Southbound	62.48
M1/3915B	Southbound	65.30
M1/3918B	Southbound	63.90
M1/3924B	Southbound	62.93
M1/3937B	Southbound	65.08
M1/3958B	Southbound	61.80
M1/3962B	Southbound	63.28
M1/3983B	Southbound	64.57
M1/3986B	Southbound	63.67
M1/4037B	Southbound	63.87
M1/4049B	Southbound	63.12
M1/4069B	Southbound	63.79
M1/4096B	Southbound	64.53
M1/4120B	Southbound	63.80
M1/4150B	Southbound	61.22
M1/4218B	Southbound	63.78
M1/4222B	Southbound	63.42
M1/4236B	Southbound	62.70
M1/4257B	Southbound	65.49
M1/4347B	Southbound	63.51
M1/4351B	Southbound	63.51
M1/4420B	Southbound	64.93
M1/4423B	Southbound	62.65
M1/4488B	Southbound	60.96
M1/4507B	Southbound	61.73
M1/4557B	Southbound	59.64
M1/4601B	Southbound	59.78
M1/4657B	Southbound	65.92
M1/4687B	Southbound	63.89
M1/4797B	Southbound	62.41
M1/4802B	Southbound	63.18
M1/4812B	Southbound	63.37
M1/4816B	Southbound	63.96

M1/4832B	Southbound	65.21
M1/4845B	Southbound	65.55
M1/4852B	Southbound	64.52
M1/4856B	Southbound	64.40
M1/4861B	Southbound	63.91
M1/4914B	Southbound	62.54
M1/4927B	Southbound	63.97
M1/4979B	Southbound	64.91
M1/5046B	Southbound	61.37
M1/5083B	Southbound	64.88
M1/5096B	Southbound	62.37
M1/5143B	Southbound	62.10
M1/5163B	Southbound	69.17
M1/5170B	Southbound	62.93

Appendix B – R Studio Code and Calculations

1. Loading dataset into R studio.

```
> library(readxl)
> mldata <- read_excel("~/MSc/Modules/Statistics & Business Intelligence/A
ssessment/Q10/PE7050 Q10 Assessment Data/mldata.xlsx")
> View(mldata)
```

2. Calculating descriptive statistics for the whole dataset and creating a function to find the mode.

```
> overall_summary <- summary(mldata$`Mean Speed (mph)`)
> find_mode <- function(x) {
+ ux <- na.omit(unique(x))
+ ux[which.max(tabulate(match(x, ux)))]
+ }
> overall_iqr <- IQR(mldata$`Mean Speed (mph)`)
> overall_mode <- find_mode(mldata$`Mean Speed (mph)`)
> overall_sd <- sd(mldata$`Mean Speed (mph)`)
```

3. Calculating descriptive statistics for the northbound strata only.

```
> nb_data <- c(subset(mldata, Direction == "Northbound"))
> nb_summary <- summary(nb_data$`Mean Speed (mph)`)
> nb_iqr <- IQR(nb_data$`Mean Speed (mph)`)
> nb_mode <- find_mode(nb_data$`Mean Speed (mph)`)
> nb_sd <- sd(nb_data$`Mean Speed (mph)`)
```

4. Calculating descriptive statistics for the southbound strata only.

```
> sb_data <- c(subset(mldata, Direction == "Southbound"))
> sb_summary <- summary(sb_data$`Mean Speed (mph)`)
> sb_iqr <- IQR(sb_data$`Mean Speed (mph)`)
> sb_mode <- find_mode(sb_data$`Mean Speed (mph)`)
> sb_sd <- sd(sb_data$`Mean Speed (mph)`)
```

5. Creating a data frame which contains all the statistics generated in steps 2 to 4 above.

```
> Direction <- c("Northbound", "Southbound", "Both")
> m1min <- c(nb_summary[1], sb_summary[1], overall_summary[1])
> m1q1 <- c(nb_summary[2], sb_summary[2], overall_summary[2])
> m1median <- c(nb_summary[3], sb_summary[3], overall_summary[3])
> m1mean <- c(nb_summary[4], sb_summary[4], overall_summary[4])
> m1mode <- c(nb_mode, sb_mode, overall_mode)
> m1q3 <- c(nb_summary[5], sb_summary[5], overall_summary[5])
> m1max <- c(nb_summary[6], sb_summary[6], overall_summary[6])
> m1iqr <- c(nb_iqr, sb_iqr, overall_iqr)
> m1sd <- c(nb_sd, sb_sd, overall_sd)
> ct_sd <- data.frame(Direction, m1min, m1q1, m1median, m1mean, m1mode, m1q3, m1max, m1iqr, m1sd)
> colnames(ct_sd)[2] <- "Min"
> colnames(ct_sd)[3] <- "Q1"
> colnames(ct_sd)[4] <- "Median"
> colnames(ct_sd)[5] <- "Mean"
> colnames(ct_sd)[6] <- "Mode"
> colnames(ct_sd)[7] <- "Q3"
> colnames(ct_sd)[8] <- "Max"
> colnames(ct_sd)[9] <- "IQR"
```

```
> colnames(ct_sd)[10] <- "SD"
> ct_sd[, -1] <- round(ct_sd[, -1], 2)
```

6. Creating a presentation table using the data frame in step 5.

```
> library(gridExtra)
> library(ggplot2)
> my_table <- tableGrob(ct_sd, rows=NULL)
> grid.arrange(my_table)
```

7. Calculated statistics for skewedness.

```
> library(moments)
> overall_skew <- skewness(mldata$`Mean Speed (mph)`)
> nb_skew <- skewness(nb_data$`Mean Speed (mph)`)
> sb_skew <- skewness(sb_data$`Mean Speed (mph)`)
> overall_kurt <- kurtosis(mldata$`Mean Speed (mph)`)
> nb_kurt <- kurtosis(nb_data$`Mean Speed (mph)`)
> sb_kurt <- kurtosis(sb_data$`Mean Speed (mph)`)
```

8. Create data frame then a presentation table using the statistics in step 7.

```
> Skewedness <- c(nb_skew, sb_skew, overall_skew)
> Kurtosis <- c(nb_kurt, sb_kurt, overall_kurt)
> shape <- data.frame(Direction, Skewedness, Kurtosis)
> my_table2 <- tableGrob(shape, rows=NULL)
> grid.arrange(my_table2)
```

9. Create aligned boxplot and histogram for the whole dataset.

```
> library(ggplot2)
> library(ggthemes)
> library(dplyr)
> library(egg)
> plt1 <- ggplot(mldata, aes(mldata$`Mean Speed (mph)`) +
+ geom_histogram(binwidth=1, alpha=0.9, fill="#737373", color="black")+
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
+ xlab("Mean Speed (mph)") + ylab("Frequency")
> plt2 <- ggplot(mldata, aes(mldata$`Mean Speed (mph)`,"")) + geom_boxplot(
+ fill = "#737373") +
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
+ axis.text.y = element_blank(), axis.ticks = element_blank(),
+ axis.text.x = element_blank()) +
+ xlab("") + ylab("")
> egg::ggarrange(plt2, plt1, heights = 1:2)
```

10. Create aligned boxplot and histogram for the northbound strata only.

```
> plt3 <- ggplot(subset(mldata, Direction %in% "Northbound"), aes(`Mean Speed (mph)`) +
+ geom_histogram(binwidth=1, alpha=0.9, fill="#737373", color="black")+
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
+ xlab("Mean Speed (mph)") + ylab("Frequency")
> plt4 <- ggplot(subset(mldata, Direction %in% "Northbound"), aes(`Mean Speed (mph)`,"")) + geom_boxplot(fill = "#737373") +
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
+ axis.text.y = element_blank(), axis.ticks = element_blank(),
```

```
+ axis.text.x = element_blank()+
+ xlab("") + ylab("")
> egg::ggarrange(plt4, plt3, heights = 1:2)
```

11. Create aligned boxplot and histogram for the southbound strata only.

```
> plt5 <- ggplot(subset(mldata, Direction %in% "Southbound"), aes(`Mean Speed (mph)`)) +
+ geom_histogram(binwidth=1, alpha=0.9, fill="#737373", color="black")+
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())+
+ xlab("Mean Speed (mph)") + ylab("Frequency")
> plt6 <- ggplot(subset(mldata, Direction %in% "Southbound"), aes(`Mean Speed (mph)`,"")) + geom_boxplot(fill = "#737373") +
+ theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()),
+ axis.text.y = element_blank(),axis.ticks = element_blank(),
+ axis.text.x = element_blank()+
+ xlab("") + ylab("")
> egg::ggarrange(plt6, plt5, heights = 1:2)
```

12. Calculate standard error for whole dataset and both stratas.

```
> library("plotrix")
> overall_se <- std.error(mldata$`Mean Speed (mph)` )
> nb_se <- std.error(nb_data$`Mean Speed (mph)` )
> sb_se <- std.error(sb_data$`Mean Speed (mph)` )
```

13. Create data frame then a presentation table using the statistics in step 7.

```
> SE <- c(nb_se, sb_se, overall_se)
> se_table <- data.frame(Direction, SE)
> my_table3 <- tableGrob(se_table, rows=NULL)
> grid.arrange(my_table3)
```

14. Conduct hypothesis testing (z tests).

```
> library(BSDA)
> z.test(mldata$`Mean Speed (mph)`,alternative='less',mu=61.86, sigma.x=overall_sd,conf.level=0.95)
> z.test(x=nb_data$`Mean Speed (mph)`, y=sb_data$`Mean Speed (mph)`, mu=0,
sigma.x=nb_sd, sigma.y=sb_sd)
```