Question 1 – 10 marks

Six months ago, a local gym set up a research programme to find out if gym members who attended exercise classes were more likely to lose weight than those who exercised alone. A census of all participants was conducted. These were the results they recorded:

	Exercise class	Gym-only workouts
Participants	46	63
Mean weight loss over 6 month	1.8 kgs	2.5 kgs
Mode weight loss over 6 months	1.5 kgs	1.7 kgs
Standard deviation	1.04	1.34

The staff at the gym want to know which type of exercise – gym only workouts or attending exercise classes – is most effective in helping individuals lose weight. Prepare a short report (not more than 1100 words) which summarises and interprets the findings, using all of the statistics given in the table above.

As Busch (2008) explains, 'businesses must collect and assess knowledge to decide on the kinds of products and services to deliver'. Comparing the impact of an exercise class and the gym-only work outs through statistics can help produce meaningful and accurate statements in relation to weight loss which could in turn lead to advertisements to the number of people joining the gym or exercise classes. It should be worth noting that this is purely based on weight loss. Other possible areas that can affect the data are not considered such as BMI or diet.

Initially, we will compare the participant numbers. From the research programme, 17 more people took part in gym-only workouts compared to exercise classes. As the sample size is quite small, the difference between the two groups can affect the mean, mode and standard deviation of the data. As Indicative Team (n.d.) explains, 'the larger volume of data is beneficial to organisations (because of) the more insight they can extract'. From the given data, it is unclear whether the sample is indicative of the whole exercise class or just a small part. This should be worth noting throughout the interpretations detailed below.

Comparing the mean weight, the gym-only workouts show greater weight loss compared to those taking part in exercises classes. Gym only workouts exceed exercise classes with 2.5kg and 1.8kg of mean weight lost respectively. This difference is quite significant at 0.7kg. Bhandari (2022) states, 'the mode tells you the most popular category' so, in this context, will be an important factor in showing which method of exercise produces the better results. From the given results, the mode indicates that the gym-only workouts led to higher weight loss, albeit by a smaller amount. This compared 1.7kg and 1.5kg, a smaller difference of 0.2kg.

It is also important to note the standard deviation between the two groups, 1.04 for the exercise class and 1.34 for the gym-only workouts. This indicates a greater variation between individuals in the gym-only workouts. As the distribution of the above data is different, in that the standard deviation, mean and mode are different, we can standardise by finding the z-value. I have taken the value of the most weight lost (mode) to determine the z-value:

Exercise class: (1.5 – 1.8) / 1.04 = -0.28846

Gym-only: (1.7 – 2.5) / 1.34 = -0.59701

From these values, in an exercise class, the weight lost by the most people is 0.29 standard deviations below the class average. However, the gym-only workouts are 0.60 standard deviations below the class average. Therefore, relatively speaking, those in the exercise class performed better than those in the gym only class. Again, it is worth reiterating that this was taken from the mode value rather than any raw data.

It should be worth noting that the above dataset only explores exercise classes and gym-only workouts without considering other external factors. Diet, for example, can play a huge role in weight loss. Combining this with either of the above may result in higher weight loss and affect the above data. It also isn't clear how long this programme lasted making any meaningful statements in relation to weight loss less clear.

Furthermore, joining an exercise class can impact on an individual's life outside of weight loss. As Cherry (2017) explains, taking part in exercise classes can lead to 'significantly lower stress levels and increased physical, mental and emotional quality of life' comparing this with the non (grouped) exercise group who 'didn't show a significant change'. Therefore, while the above figures may indicate gym only workouts have a higher impact on weight loss, taking part in an exercise class may lead to an improvement in a person's well-being.

To improve the overall findings, results of individual raw data would provide for accurate and reliable conclusions being drawn. An accurate median could be drawn allowing for certain plots to be presented to further demonstrate the greater weight loss between either of the exercise groups. As Knaflic (2015) states, 'effective data visualisation can mean the difference between success and failure' when it comes to communicating findings. Having this would have allowed the findings to be clearly displayed to the relevant audiences.

Overall, if a person's main target is to lose weight, the gym-only workouts provide improved weight loss over a period of 6 months. The average (mean) weight loss is 0.7kg greater than if an individual participated in exercise classes. Furthermore, when comparing the amount of weight most people lost, albeit a smaller difference, gym-only workouts the weight lost in terms of frequency was 1.7kg compared to exercises classes at 1.5kg – a difference of 0.2kg. If weight loss is the only target, gym only workouts provide stronger results. However, as mentioned above in reference to Cherry (2017), joining an exercise class can show further personal changes outside of weight loss. Having access to the raw data would also allow for representative plots to be drawn to further demonstrate the above conclusions.

Question 2 – 5 marks

Describe a way to deal with missing data values in data for processing it.

As Columbia University, Department for Statistics (n.d.) states, when dealing with 'how to handle missing data, it is helpful to know why they are missing'. Humprhies (n.d.) continues saying, this can be because 'certain groups are more likely to have missing values'. Ultimately, the type of missing data needs to be specified. Graham (2009) amongst other researchers quantify these as 'Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR)'. There are a range of methods that can be completed for each type of missing data. I believe that the multiple imputation method produces the most reliable results.

As Black (n.d.) states, multiple imputation 'is the practice of computing multiple different imputed dataset (allowing) us to complete multiple estimates… and combine them all'. Kang (2013) explains that this method 'begins with a prediction of the missing data using the existing data from other variables… the missing values are then replaced with the predicted values.' This creates a new dataset. As opposed to single imputation, the multiple imputation method produces multiple datasets, each to be analysed using the chosen statistical analysis. These results can then be combined creating the most accurate reflection of the missing data as possible. Kang (2013) believes 'multiple imputation… reflects the uncertainty associated with the estimation of the missing data'. As Acock (2005) states, 'multiple imputation allows for unbiased standard errors' reflecting the above point further. This is important to consider against other methods such as simple mean imputation as this doesn't account for variance. However, as Black (n.d.) explains, this approach 'makes the inference stage slightly more laborious' – with larger datasets this approach may take more time, and ultimately may not be the most cost effective.

Whilst the above method is the most preferred, Humphries (n.d.) states that it is most appropriate to 'try a few methods. Often if the result (has) similar estimates… the author can put as a footnote to support (the) method'. Therefore, exploring whether there are similarities in how missing data is inputted over a range of methods can allow for a more reliable method of dealing with missing data values. Graham (2009) surmises that we should 'try move away from the fear of missing data' as these situations will always occur – having an appropriate methodology towards missing data can make all the difference when processing and analysing data.

Question 3 – 10 marks

Suppose that a family is leaving on a summer vacation in their camper and that M is the event that they will experience mechanical problems, T is the event that they will receive a ticket for committing a traffic violation, and V is the event that they will arrive at a campsite with no vacancies. Referring to the Venn diagram of this situation in the Figure below, state in words the events represented by the following regions:

(a) region 5

- [M – (V ∪ T)]
- The family will experience mechanical problems but will not receive a ticket for committing a traffic violation and they will not arrive at a campsite which has no vacancies.

(b) region 3

- T ∩ V ∩ M'
- The family will commit a traffic violation and will arrive at a campsite with no vacancies. However, they will not experience mechanical problems.

(c) regions 1 and 2 together

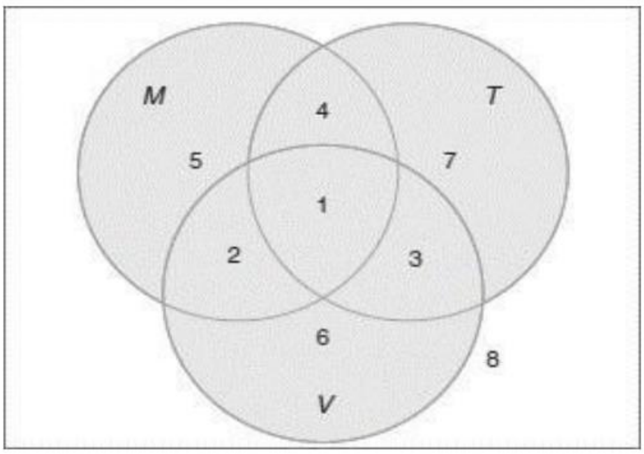
- M ∩ V
- The family will experience mechanical problems and they will arrive at a campsite with no vacancies.

(d) regions 4 and 7 together

- V' ∩ T
- During the trip, the family will receive a ticket for a traffic violation, but they will not arrive at a campsite which has no vacancies.

(e) regions 3, 6, 7, and 8 together

- ((T ∪ V) ∩ M') ∪ (T' ∪ M' ∪ V'))
- The family will receive a ticket violation or will arrive at a campsite with no vacancies. However, they won't experience mechanical problems. Furthermore, they may not experience any of the aforementioned issues (number 8).



Question 4 – 7 marks

British Airways is considering two different suppliers A and B for a critical chip component used in their planes. Each supplier has a different defect rate. The defect rate for supplier A and B are 10 out of 1000 and 8 out of 1200 respectively. Discuss how statistical analysis can be used to make a comprehensive and informed decision.

Using statistical analysis is fundamental in current business practices not only to find the edge over competitors but to make informed decisions which can save time, resources, and money. With British Airways arguably one the most well-known and successful airliners worldwide - as Calder (2023) reports making '£50 of profit a second in the first 9 months of 2023' - such decisions can be fundamental to the operations of the business. Sankar (2020) states, 'the data that is available presently is unlike any ever seen before' with Wamba et al (2015) taking this further explaining that 'big data has the potential to transform the entire business process'. Therefore, using statistical analysis can support British Airways in making informed decisions to maximise business' results.

One such way to use statistical analysis would be with relative frequency. As Frost (2021) describes, 'relative frequency indicates how often a specific kind of event occurs within the total number of observations.'

Relative frequency is calculated with the following simple formula:

RF = Event Count / Number of observations.

Using this, we can calculate the percentage that there is a defect rate of the chip component.

<u>Input:</u>

```
A_defect <- 10
B_defect <- 8

A_trials <- 1000
B_trials <- 1200

A_relativeF <- A_defect / A_trials
B_relativeF <- B_defect / B_trials

A_relativeF * 100 #calculates the percentage
B_relativeF * 100 #calcualtes the percentage
```

<u>Output:</u>

```
> A_relativeF * 100 #calculates the percentage
[1] 1
> B_relativeF * 100 #calcualtes the percentage
[1] 0.6666667
```

The output shows the value as the percentage themselves. As seen from above, supplier A has a defect rate of 1% whilst supplier B has a defect rate of 0.67% (2dp). This indicates that supplier B would be the preferred choice of supplier due to the lower defect rate.

Whilst not available, other data and information should be considered when making these decisions. Supplier reputation, costs and lead-time of the computer chip should all be factored into any business decision, alongside the above statistical analysis.

Question 5 – 6 marks

The probability that an iPhone will survive a shock test is 0.69. Find the probability that exactly 3 of the next 5 iPhones tested survive. These tests are independent. Also, describe which probability distribution is used to answer this question, why it is selected? Can this problem be solved without using the probability distribution, and how – discuss it?

The probability distribution that would be used is the binomial probability distribution. As Bourne (2018) explains, there needs to be 'repeated trials', 'an outcome that may be classified as a success or a failure' and 'the probability of success'. As these are prerequisites, it matches the binomial probability distribution.

This could also be denoted as below:

n = 5 (number of tests)

X = 3 (number of iPhones that will pass the test)

p = P(iPhone survives) = 0.69

q = P(iPhone doesn't survive) = 0.31

$P(X = 3) = {}^{5}C_{3}\,(0.69)^{3}\,(0.31)^{5-3}$

= {5! / (5-3)! 3!} (0.328509)(0.0961)

= {(5 x 4 x 3 x 2 x 1) / 12} (0.0315697149)

= 0.315697149

This shows that the probability of 3 iPhones out of 5 surviving the shock test would be 31.57% (to 2dp).

This could also be performed using the dbinom function in R:

Input:

```
n <- 5 #number of trials
k <- 3 #number of successes
p <- 0.69 #probability of success

probability <- dbinom(k, size = n, prob = p)

print(probability) #prints the resulting probability
```

Output:

```
> n <- 5
> k <- 3
> p <- 0.69
>
> probability <- dbinom(k, size = n, prob = p)
>
> print(probability)
[1] 0.3156971
```

There is an alternative, albeit more time-consuming, method of completing the above problem using the product rule. This is because we are finding 3 lots of the iPhone succeeding (3 x 0.69) and 2 lots of the iPhone not succeeding in the shock test (2 x 0.31). To show this, I have created a table showing the collation of results for when this instance occurs.

Event	P1	P2	P3	P4	P5
1	0	0	0	0	0

2	0	0	0	0	1
3	0	0	0	1	0
4	0	0	0	1	1
5	0	0	1	0	0
6	0	0	1	0	1
7	0	0	1	1	0
8	0	0	1	1	1
9	0	1	0	0	0
10	0	1	0	0	1
11	0	1	0	1	0
12	0	1	0	1	1
13	0	1	1	0	0
14	0	1	1	0	1
15	0	1	1	1	0
16	0	1	1	1	1
17	1	0	0	0	0
18	1	0	0	0	1
19	1	0	0	1	0
20	1	0	0	1	1
21	1	0	1	0	0
22	1	0	1	0	1
23	1	0	1	1	0
24	1	0	1	1	1
25	1	1	0	0	0
26	1	1	0	0	1
27	1	1	0	1	0
28	1	1	0	1	1
29	1	1	1	0	0
30	1	1	1	0	1
31	1	1	1	1	0
32	1	1	1	1	1

There are 10 events where exactly 3 of the next 5 iPhones will survive the shock test. These have been highlighted above. This can be summarised as:

10 x (0.69 x 0.69 x 0.69 x 0.31 x 0.31)

10 x (0.0315697149) = 0.315697149

As a percentage, this matches to the earlier figure of a 31.57% (to 2dp) chance that exactly 3 out of 5 iPhones will successfully complete the shock test.

Question 6 – 6 marks

A real estate agent claims that 64% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the location of the critical region.

Assume α = 0.05.

What does α = 0.05 show here?

As Walpole et al (2016) states, ‘a statistical hypothesis is an assertion or conjecture concerning one or more populations’. In this case, the null hypothesis is that 64% of all private residencies being built today are 3-bedroom homes. The ‘*α = 0.05’* statement is the level of significance. The alternative hypothesis is the percentage of private residencies built is not 64%. This could be written as:

$H_0 = 0.64$

$H_1 \neq 0.64$

The alternative hypothesis indicates a two-tailed test as H_0 can be rejected if either above or below the hypothesis value – in this case 0.64 (64%) – at a significant rate. As Imai (n.d.) explains, 'we calculate the critical percentage for the α significance level by qnorm(1 − α/2) for a two-sided test'. In this case, this would be:

1 – 0.05 / 2

1 – 0.025 = 0.975

Input:

```
qnorm(0.975)
```

Output:

```
> qnorm(0.975)
[1] 1.959964
```

This result tells us that the critical region would be either 1.96%(2dp) above or below the null hypothesis statistics. We can then use this to determine the critical value is between these percentages at a confidence level of 5%:

```
> 64 + 1.96
[1] 65.96
> 64 - 1.96
[1] 62.04
```

Question 7 – 6 marks

The following data is taken from a company about its advertisements and purchases of the product. Calculate coefficient of correlation to measure the strength and direction of relationship between the number of advertisements and purchases made, and comment on it.

Does it imply causation or not? And for both cases (implying causation or not) discuss why?

Number of advertisements	0	2	3	4	5	6	7	10
Purchases	4	1	4	5	10	8	3	12

As Gunner (2022) states, 'finding the positive or negative correlation between two variables is an important way to study cause and effect.' Despite this, as Maths Tutor (2017) explains, 'correlation does not imply a causation' and it is important that the results of any correlative tests are analysed in the context of the data itself. To achieve this, I have used the 'cor' function to find the correlation coefficient between the number of advertisements and the purchases made.

Input:

```
advert_no <- c(0, 2, 3, 4, 5, 6, 7, 10)
purchases <- c(4, 1, 4, 5, 10, 8, 3, 12)

correlation_coefficient <- cor(advert_no, purchases)
correlation_coefficient
```

Output:

```
[1] 0.6790033
```

The above result shows a moderate positive correlation between the number of advertisements and purchases made. This indicates there is a relationship between the number of advertisements and purchases – i.e. The greater the number of advertisements, the more purchases are made. It is worth noting, however, that there are inconsistencies in the above data. There are missing values for 1, 8 or 9 advertisements being purchased. Having these values within the above dataset could have resulted in a more accurate picture of the correlation between the two variables.

In my opinion, the result of the above dataset does show causation but other factors do need to be considered. When examining how the purchases made changes as the number of advertisements increases, there are certain values which show that other factors may contribute to the purchases made. For example, when comparing 0 advertisements and 3 advertisements, this has had no impact on the number of purchases made. When the company took out 7 advertisements, the number of purchases made decreases compared with smaller values. As the Australian Bureau of Statistics (n.d.) states, 'relationships can be due to other factors'. In this context, this could range from special offers, time of year amongst others. Ultimately, I believe there is some causation between the number of advertisements and the purchases made but more analysis with other factors needs to be completed.

Question 8 – 8 marks

A famous company selling household appliances wants to determine the relationship between advertising expenditures and sales. The following data was taken from 6 major sales regions. The expenditure is in thousands of pounds and sales are in millions of pounds.

Region	Expenditure, x	Sales, y
1	1.5	2.0
2	2.0	2.0
3	4.0	2.5
4	4.0	5.0
5	4.5	3.5
6	8.0	4.5

(a) *Estimate the linear regression line to provide a chart and summary statistics together with the coefficients and discuss them.*

As Penn State (n.d.) explains, 'simple linear regression is a statistical method that allows us to summarise and study relationships between two variables'. From here, we can begin to explore the covariance between these variables. I first created a list of both expenditure and sales and used the 'lm' function to store this in the 'model' variable.

<u>Input:</u>

```
exp <- c(1.5,2,4,4,4.5,8)
sales <- c(2,2,2.5,5,3.5,4.5)

model <- lm(sales ~ exp)
model
```

<u>Output:</u>

```
> model <- lm(sales ~ exp)
> model

Call:
lm(formula = sales ~ exp)

Coefficients:
(Intercept)          exp
     1.6274       0.4057
```

From this, the linear regression model is: sales = 1.6247 + 0.4057 * expenditure

The expenditure value would be in thousands and the sales value would be in millions.

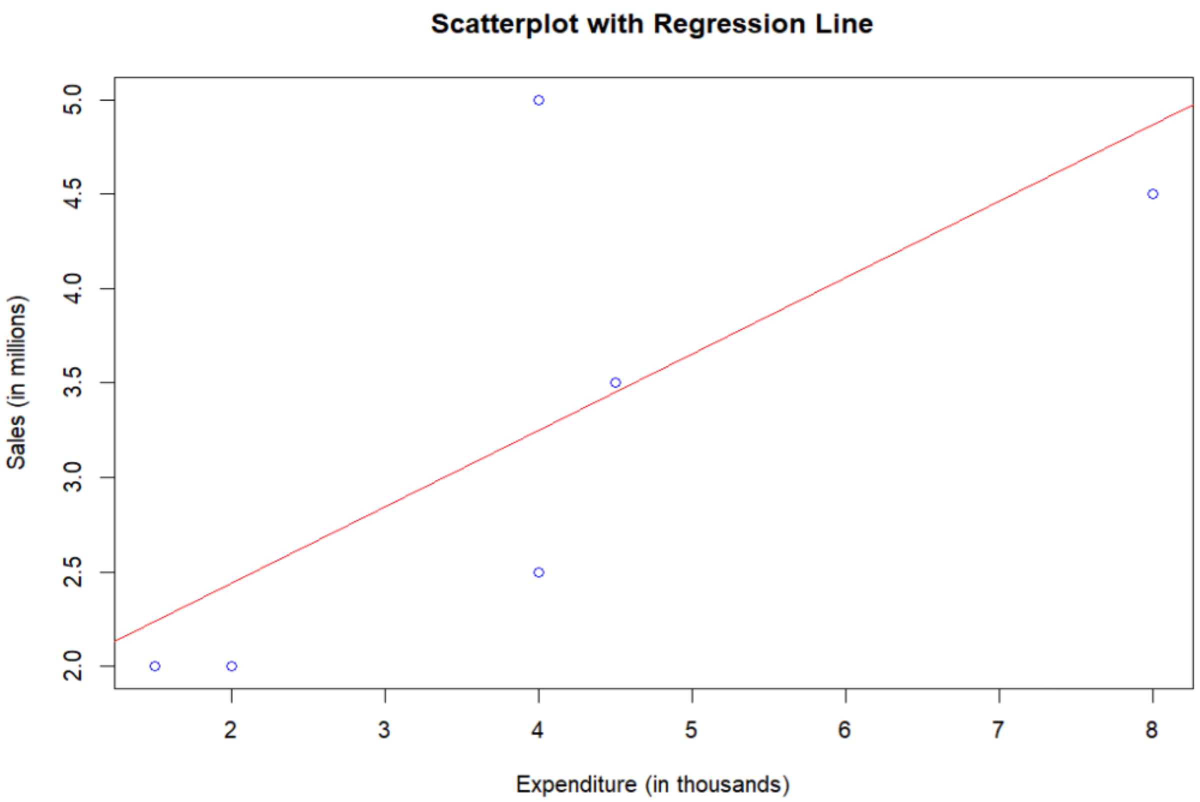
We can plot the chart showing the linear regression and values given.

Input:

```
plot(exp,
     sales,
     main = "Scatterplot with Regression Line",
     xlab = "Expenditure (in thousands)",
     ylab = "Sales (in millions)",
     col = 'blue')

abline(model, col = 'red')
```

Output:



Nguyen (2017) describes the notion of covariance as the calculation that 'shows you the direction of the relationship. If one variable increases and the other variable tends to also increase, the covariance would be positive.' This can be seen from the above chart produced; the greater the expenditure, the greater sales. However, as Pennsylvania State University (2022) state, 'association is not causation'. We can explore the summary statistics, notably r^2 to explore the association in more detail and determine the strength of the relationship between expenditure and sales.

Input:

```
cor(sales, exp)^2
```

Output:

```
> cor(sales, exp)^2
[1] 0.5206984
```

As seen from this value, the correlation, while positive, isn't necessarily strong. I would conclude that there isn't a strong enough correlation to indicate that a higher expenditure causes higher sales. Other factors, combined with higher expenditure, may result in greater sales.

(b) *Estimate the expected sales for a region where 6.2 to 6.8 thousand pounds are being spent on advertising.*

To find the expected sales of the given amounts using the linear regression, I created a list of the expenditure values. I then used the model created for the above part of the question to find the predicted sales. Again, the expenditure would be in thousands and the sales in millions.

Input:

```
mydf <- data.frame(exp = c(6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8))
predict(model, newdata = mydf)
```

Output:

```
       1        2        3        4        5        6        7 
4.142453 4.183019 4.223585 4.264151 4.304717 4.345283 4.385849 
```

Tabular format of the above:

Projected Expenditure (thousands)	Projected Sales (millions to 2dp)
6.2	4.14
6.3	4.18
6.4	4.22
6.5	4.26
6.6	4.30
6.7	4.35
6.8	4.39

Question 9 – 12 marks

Below are given some hearing frequencies (audiograms), you are required to: (12)

(a) *Find the number of clusters for the given data.*

Initially, I gave the original data a title 'Frequency_Data'. When clustering the above data, I then ensured it scaled the data so that the means for each frequency test were at 0.

Input:

```
df <- scale(Frequency_Data)
summary(df)
```

Output:

```
    Freq250             Freq500             Freq1K              Freq2K              Freq4K       
 Min.   :-1.5320   Min.   :-1.7078   Min.   :-1.8970   Min.   :-2.2325   Min.   :-2.7949  
 1st Qu.:-0.6582   1st Qu.:-0.8009   1st Qu.:-0.7367   1st Qu.:-0.5412   1st Qu.:-0.4593  
 Median :-0.2214   Median :-0.1730   Median :-0.1566   Median : 0.3044   Median : 0.1246  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.7252   3rd Qu.: 0.8037   3rd Qu.: 0.7135   3rd Qu.: 0.7273   3rd Qu.: 0.6111  
 Max.   : 2.2543   Max.   : 1.9198   Max.   : 1.8738   Max.   : 1.5729   Max.   : 2.2655  
     Freq8K       
 Min.   :-2.0150  
 1st Qu.:-0.5518  
 Median :-0.3428  
 Mean   : 0.0000  
 3rd Qu.: 0.4410  
 Max.   : 1.9565  
```

I then used a scale of 1 to 15 to find the optimal number of clusters.

Input:

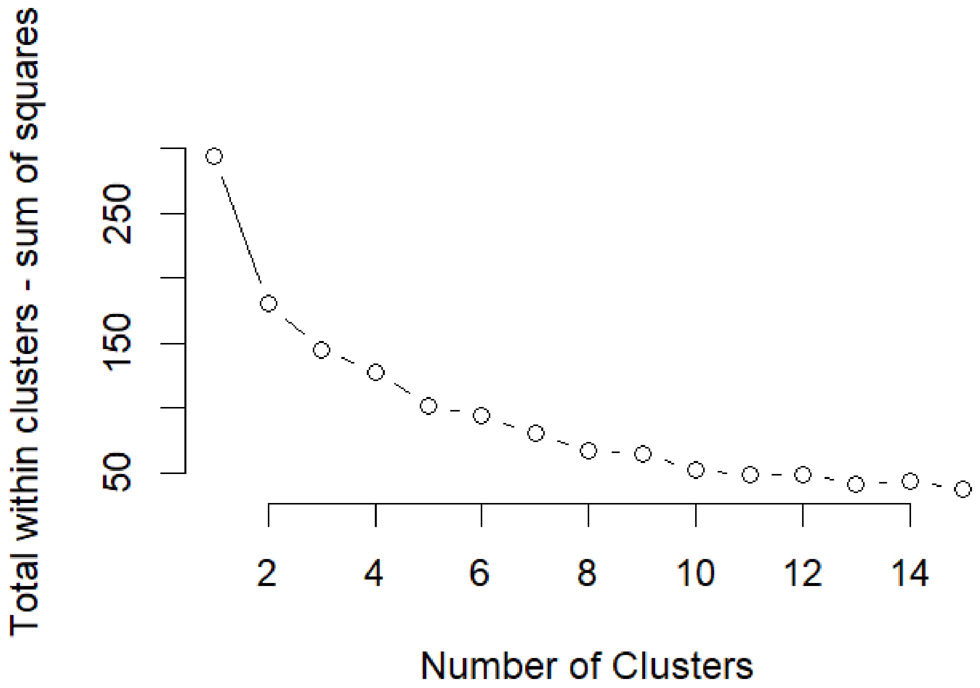
```
wss <- function(k) {
  kmeans(df, k)$tot.withinss
}

k <- 1:15

wssvalue <- map_dbl(k, wss)

plot(k,
     wssvalue,
     type = "b",
     frame = "FALSE",
     xlab = "Number of Clusters",
     ylab = "Total within clusters - sum of squares")
```

Output:



After finding the ‘elbow’, I determined that the optimal number of clusters could be a value between 2 and 4 as the plot then becomes inconsistent. For this, I chose 3 clusters.

(b) Cluster the given data and comment on each cluster of the data.

I first plotted the k-means clustering using the below script.

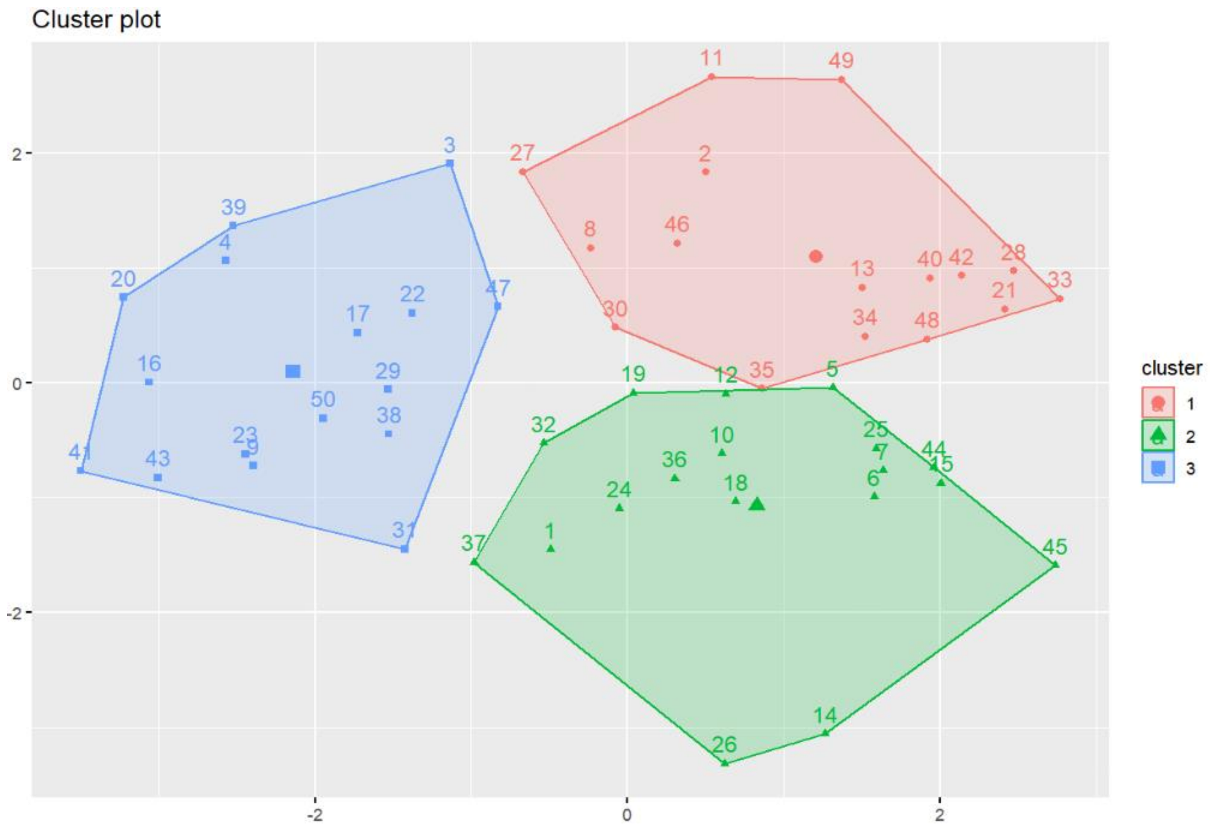
Input:

```
result <- kmeans(df, 3)

print(result)

fviz_cluster(result, data = df)
```

Output:



I then combined this by creating a self-organising map (SOM) to help further form my analysis of each cluster.

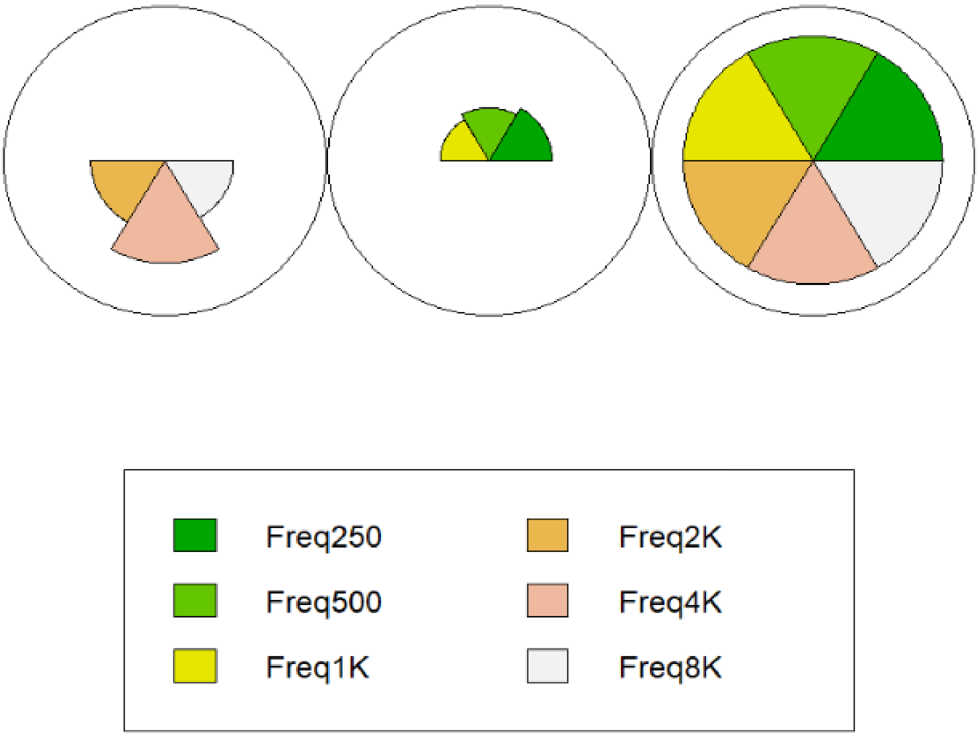
Input:

```
library(kohonen)


g <- somgrid(xdim = 3, ydim = 1, topo = "rectangular")
map <- som(df,
           grid = g,
           alpha = c(0.05, 0.01),
           radius = 1)

plot(map)
```

Output:



When exploring the k-means clustering plots, cluster 1, in general, shows that people had good hearing across the range of frequencies. The clusteroid indicates that people on average performed well across all tests. When analysing using the SOM, comparatively, these people had better hearing at the lower frequencies compared with the higher frequencies. Notably, this group had a large percentage of people with poorer hearing at frequency 4k from the SOM.

In cluster 2, the individuals had improved hearing at a higher frequency but the difference between cluster 1 is the performance in the lower frequencies with this cluster in the bottom right hand quadrant. This is justified in the SOM with the segments of the nodes showing greater values at the three frequencies under 2k. It is worth noting individuals 26 and 14 will have brought this average down. Further clusters could have affected the averages.

In cluster 3, the SOM isn't helpful in determining the hearing success of the individuals within node as each of the segments in the node show they are the same. The k-means clustering plot shows that the individuals within this cluster performed variably. Above the average, there are some who's audiograms scored highly at certain frequencies whilst others in the bottom left quadrant showed limited hearing at all frequencies.

Whilst the above analysis may be useful in one respect, having further information about the patients could provide further analysis, such as age, occupation and other medical conditions. This would mean further relationships between the audiogram results could be explored.

After comparing the clustering plotting and the self-organising map, adding further nodes and/or clusters may have resulted in more patterns to be obtained. Combining this with the above variables could allow for more accurate and reliable analysis.

Question 10 - A scenario – 30 marks

You are a director of a major manufacturing organisation and collecting various pieces of information for your potential clients, such as on one of your major clients who is based in London, will require delivery lorries to travel the length of the M1. You will investigate the speed on this road using the data available at https://www.trafficengland.com/traffic-report.

You should only use the source specified. You will need to adopt a sampling approach and credit will be given for schemes which show you have considered how to apply the principles of sampling to obtain the best results with the smallest possible dataset.

Logistics is a crucial component of any successful manufacturing organisation. Being able to transport goods efficiently and quickly increases productivity, potential and ultimately profit. By understanding and identifying the relationships between the speed, time it takes for lorries to travel the M1 and other factors, the success of the company can increase as there may be potential competitive edges which can be explored. Throughout this report, I will be considering rules and regulations surrounding haulage vehicles, sampling speed data from National Highways and presenting my analysis through the use of visualisations, machine learning and my explanations. Where possible, factors potentially affecting the data will be explored and discussed for a holistic viewpoint. Hodeghetta and Nayak (2023) explain that 'having only programming skills along with statistical or mathematical knowledge can sometimes lead to proposing impractical suggestions.' By combing these, the aim is to provide a clear and concise analysis of the speed of the length of the M1 and factors which can affect the average speed or how the given data's influence can be maximised.

The M1 interchange is one of the major motorways connecting London the north. Turning into the A1(m) junction, which leads towards Newcastle and Scotland, the M1, as reported by Jones et al (n.d.), 'carries between 130,000 and 140,000 vehicles every day'. As it's such an important route, haulage and logistic companies make up many of these vehicles. Before any analysis can be completed, the laws and regulations surrounding the speed that lorries can travel on the motorways, as well as the rules for drivers, need to be explored. In the first instance, the weight of a lorry can play a crucial part in the speed it can travel. For example, as the Department for Transport (n.d.) outline, 'Goods vehicles (more than 7.5 tonnes maximum laden weight) in England and Wales have a speed limit of 60mph. This is an important consideration, especially when a lot of the data collected exceeds this 60mph speed. Appropriate functions will be initiated to show this in further detail.

When collecting the data from National Highways, I wanted the same amount of data points for the days of the week. I started collecting data on Saturday 4th of November and finished on Friday 1st December. This would mean that each day of the week would have four datasets for both north and south. The raw data is found in the Appendix A. During data collection, it was observed at the same time, 12pm, each day.

When completing the sampling method, I decided on using systematic random sampling. As Elfil and Negida (2017) explain, in this sampling method 'the researcher can start randomly and then systematically chooses next patients (in this case dates) using a fixed interval.' When completing this sampling, I used this sampling method every other day so that there would be two datasets for each day of the week, a total of 14 days being used for northbound and southbound travel. I used this approach opposed to simple random sampling as this may have resulted in certain days being repeated multiple times and other days not being represented at all. As Rumsey (2011) states, 'the quality of the data is extremely critical' and is the reason why this approach was taken. Northbound and Southbound data will also be combined so that the whole journey can be analysed too. To achieve, this I used the sequencing function to save the systematic randomly sampled data in a new data frame. The code to achieve this can be found in Appendix B.

Before completing any analysis on the data, I also transformed the data so that any speeds exceeding 60mph were changed to 60mph. As HGVs cannot travel above 60mph, any speed collected which indicates this would not be useful to the logistical company. I used a for loop to go through each column and transform any data which indicated speeds above 60mph to precisely 60mph. All other speeds would remain unchanged. This transformed data can be seen in Appendix C.

My first hypothesis I wanted to explore was the following: 'There is no difference in average speed on the M1 when travelling on different days of the week.'

This could be denoted as:

H_0 = There is no difference in average speed on the M1 on different days of the week

H_1 = There are different average speeds on the M1 on different days of the week.

This is a two-tailed test as the average speed can be higher or lower than on the different days. The significance level will be set at 0.05.

I began by analysing how the speed of the motorways changed from day to day on northbound, southbound, and combined routes. Separate dataframes of north and southbound average speeds, as well as this combined, can be seen in Appendix D per date. I used the mean aggregate function, after adapting the data so that it would be of long type. As these were still in dates, I converted them to the corresponding days of the week and grouped them together in a plot to show how the average speed was affected daily for both north and south routes. The code for this can be found in Appendix E. Throughout, it is important to consider how, as Murray (2019) states, visualisation can turn from 'unassuming visualisations into an emotion-filled data story' leading to false conclusions being drawn.

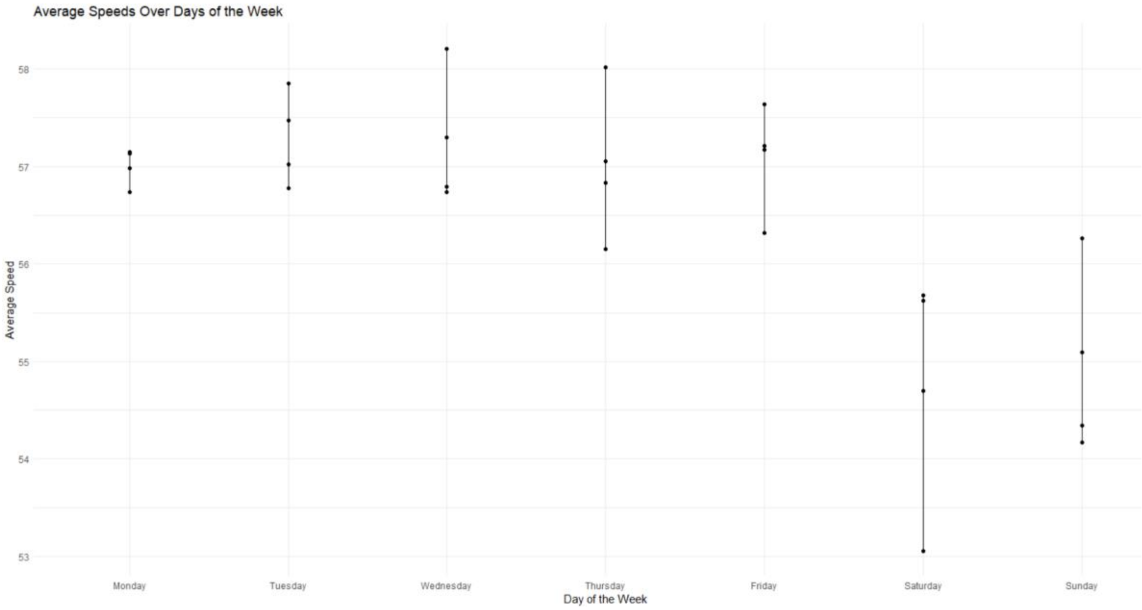


Figure 1 - Combined North and South Plot of Average Speeds on M1

As seen from this plot, there is a clear drop in average speed for the duration of the M1 on a Saturday and Sunday meaning the weekend is not an optimal time to be completing distance on these days. From the plot, Tuesday, Wednesday and Thursday allowed for the optimal speeds albeit in a range of 1mph.

From this, I wanted to see if there was a difference in the speeds northbound and southbound on the motorways.

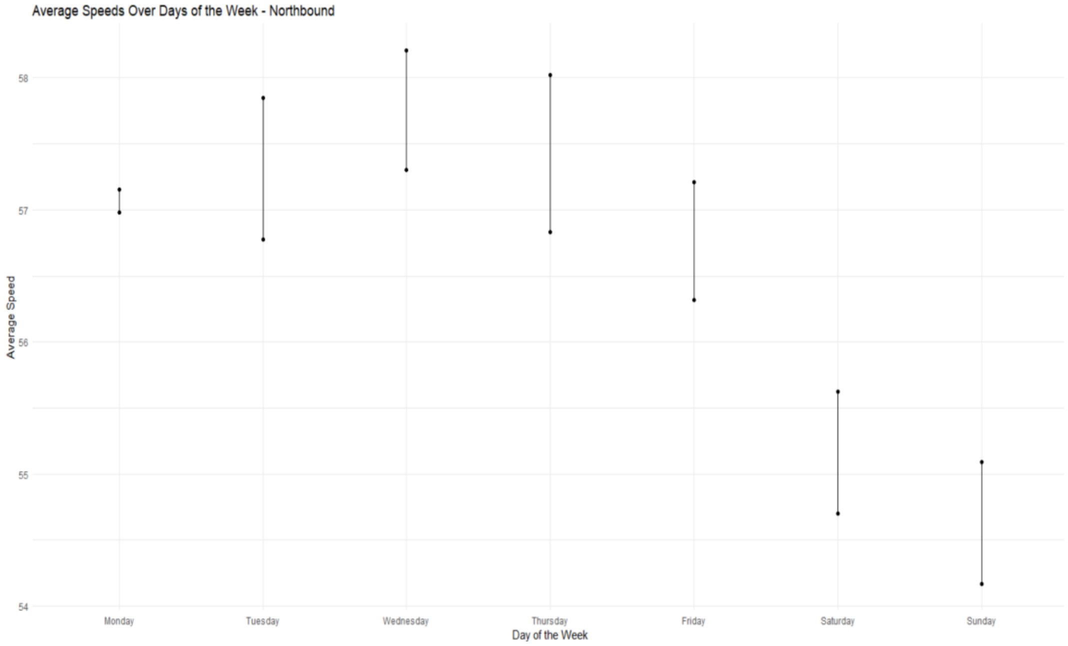


Figure 2 - Northbound average speed plot

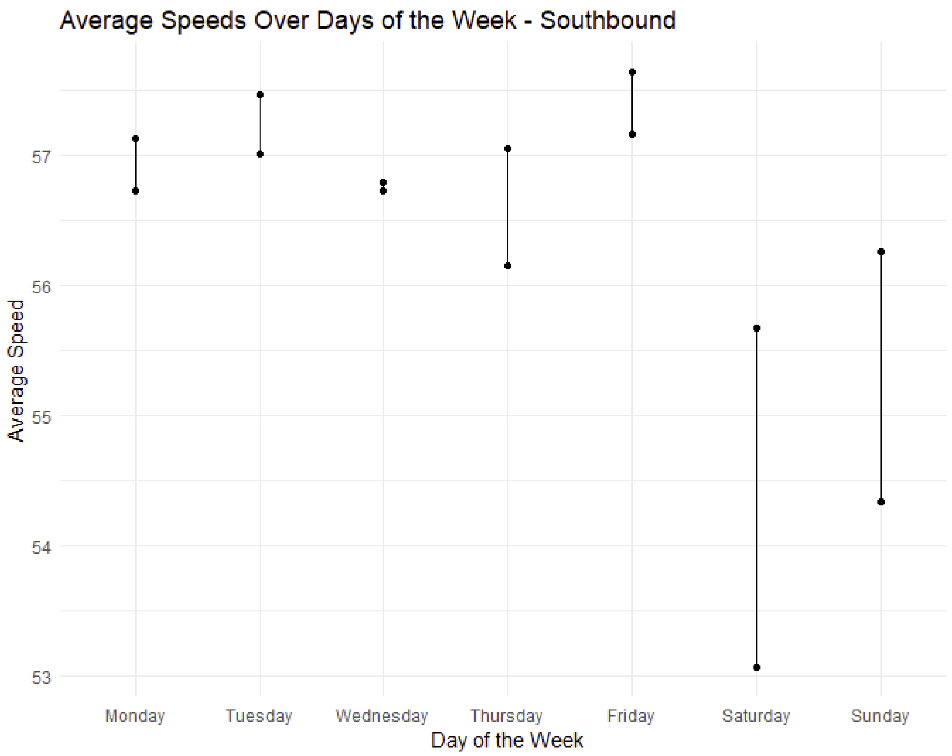


Figure 3 - Southbound Average Speed Plot

Similar visuals can be seen from these plots; the weekend sees a drop in average speed on the length of the M1 reiterating the points above. However, on the Sunday southbound plot, one of the average speeds exceeds an average Thursday speed. Despite this, Moto-way (n.d.) have commented 'Motorway traffic is considered to be lighter on certain days of the week, namely Tuesday, Wednesday and Thursday' which is supported to an extent in Figure 1. It is important to reiterate that this data was taken at 12pm daily and rush hour traffic isn't taken into account.

To support these visualisations, I then used an ANOVA test, which will be used throughout the report. As Gaur and Gaur (2009) state, ANOVA is used to compare the means of more than two populations, in this case the average speeds on the different days of the week. The ANOVA test will have a level of significance of 0.05. The null hypothesis will be rejected if the test statistic (p-value) is below 0.05.

```
> anova_result <- aov(Average_Speed ~ DayOfWeek, data = average_speeds_combined)
>
>
> summary(anova_result)
            Df Sum Sq Mean Sq F value   Pr(>F)    
DayOfWeek    6  29.60   4.933    8.55 8.95e-05 ***
Residuals   21  12.12   0.577
```

Figure 4 - ANOVA result for combined northbound and southbound data

The result of the ANOVA test indicates that the test statistic is below the confidence level of 0.05. Therefore, we can reject the null hypothesis and determine that there are different average speeds on the M1 on different days of the week when combining north and southbound data. As seen from above, weekend travel affects this. As there appeared to be some differences in the nature of the fall when travelling north and south from the plots in Figure 2 and Figure 3, I used the ANOVA test on the northbound and southbound data.

```
> anova_result <- aov(Average_Speed ~ DayOfWeek, data = average_speeds_north)
>
>
> summary(anova_result)
            Df Sum Sq Mean Sq F value Pr(>F)  
DayOfWeek    6 17.415  2.9025    6.87 0.0114 *
Residuals    7  2.957  0.4225


> anova_result <- aov(Average_Speed ~ DayOfWeek, data = average_speeds_south)
>
>
> summary(anova_result)
            Df Sum Sq Mean Sq F value Pr(>F)  
DayOfWeek    6 15.032  2.5053   2.925 0.0931 .
Residuals    7  5.995  0.8564
```

Figure 5 - ANOVA results for separate northbound and southbound travel

From these results, we can see the differences more clearly, supporting the earlier judgements as the test statistic is 0.01, rejecting the null hypothesis. The northbound data shows that there are different average speeds on the motorway depending on the day of travel. However, the southbound has a test-statistic result of 0.09 indicating that the null hypothesis can be accepted. As seen from Figure 3, the overlap between Thursday and Sunday can likely account to this.

To further the reliability and accuracy of the above, it would be pertinent in using a larger amount of data to further strengthen the notion that weekend travel is generally slower; this data only takes in two data points for the separate directions. Also, further hypothesis tests could be conducted on specific days of the week.

As there seems to be a form of relationship between days of the week in terms of average speed, in that some of the data points were closer together in the above plots, I then wanted to explore whether the speed travelling north corresponds to the speed travelling south using linear regression.

My hypothesis again would be two-tailed with a confidence level of 0.05.

H_0 = There is no difference in average speed on the length of the M1 whether travelling north or south.

H_1 = There is a difference in average speed on the length of the M1 whether travelling north or south.

I plotted the average speeds of the northbound M1 along with the southbound M1. I then used the 'cor' function to analyse the strength of correlation between the two. The code for this can be found in Appendix F. Rongpeng (2020) states that the 'regression model studies the direction of the correlation and the strength of the correlation'. This may be useful in determining whether average north speed can be predicted from average south speed, considering that HGVs are limited to 60mph. It would be expected that there would be greater variance if no limit was placed on speed.

```
[1] 0.5401903
```

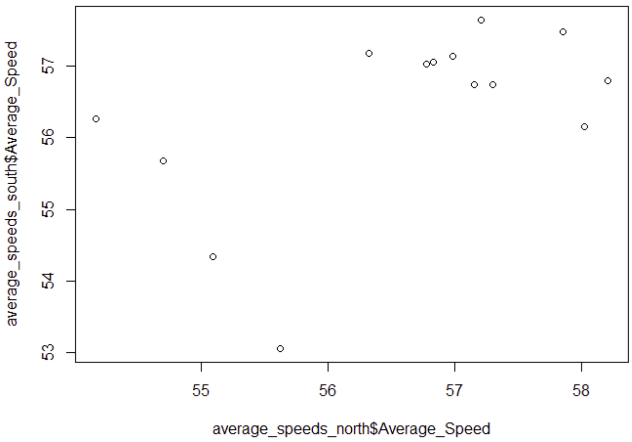


Figure 6 - Correlation value and plot of average north speed and average south speed

The result of the correlation indicates that in general there is a positive relationship between the northbound and southbound M1, albeit moderate in strength at 0.54 (2dp). This somewhat indicates that there isn't a difference in average speed travelling north or south from the data given. I then used linear regression to plot the relationship.

```
> model <- lm(average_speeds_north$Average_Speed ~ average_speeds_south$Average_Speed)
> model

Call:
lm(formula = average_speeds_north$Average_Speed ~ average_speeds_south$Average_Speed)

Coefficients:
                       (Intercept)  average_speeds_south$Average_Speed
                           26.6121                              0.5317
```

The result shows that the formula is as follows:

Northbound Speed = 26.6121 + (0.5317 x Southbound Speed)

This is then plotted below. The code can be found in Appendix G.

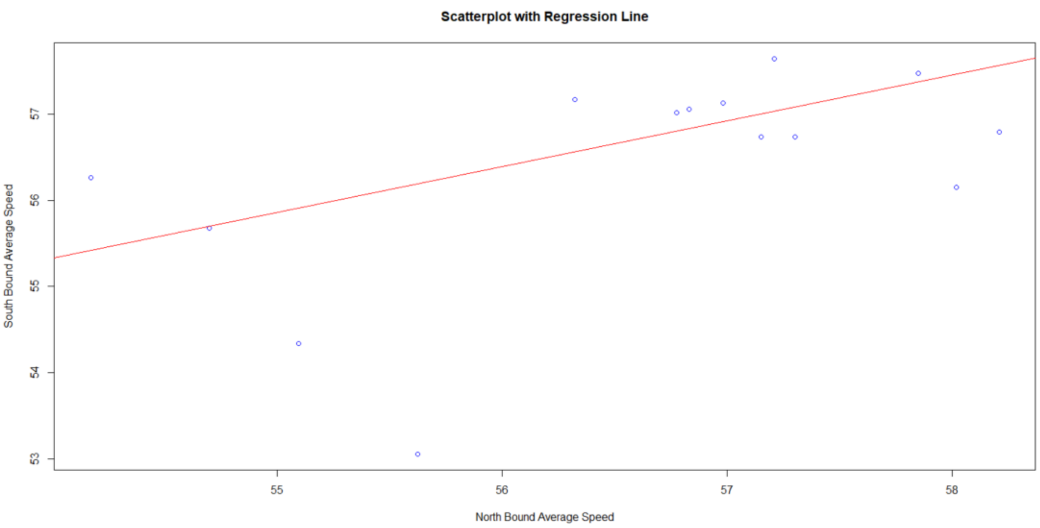


Figure 7 - plot with linear regression added

As can be seen there is a moderately positive relationship between the two variables. Different explanations can accommodate this, such as how slower speeds tend to be applied to both sides of the motorway such as if there has been an accident. The ANOVA test is then conducted below:

```
> anova_result <- aov(average_speeds_north$Average_Speed ~ average_speeds_south$Average_Speed)
>
> summary(anova_result)
                                   Df Sum Sq Mean Sq F value Pr(>F)
average_speeds_south$Average_Speed  1  5.945   5.945   4.944 0.0461 *
Residuals                          12 14.428   1.202
```

The result shows that the test statistic is 0.046 and below the 0.05 required to accept the null hypothesis. This shows that there is a difference between the speeds travelling south and north. It is again worth reiterating that the speeds travelling north and south were capped at 60mph.

After observing the speed that the lorries would be able to travel on the M1 as a whole, I then wanted to look more closely at the individual junctions and whether there were any significant speed differences from one junction to another. The hypothesis statement would be as follows.

H_0 = There is no difference in average speed from junction to junction along the M1.

H_1 = There is a difference in average speed from junction to junction along the M1.

Again, this is a two-tailed test as the difference (if any) can be higher or lower that the average.

I began this section by finding the average speed at each junction, using row means and plotting this. The code for this can be found in Appendix H.

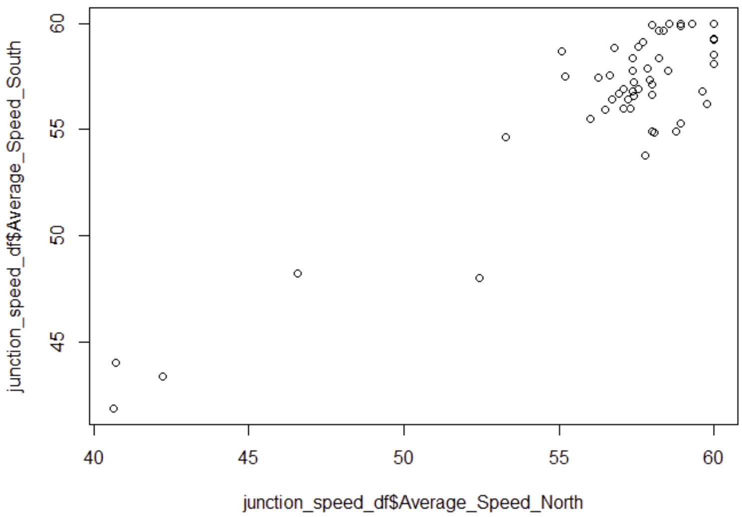


Figure 8 - Plot of Average Speeds North and South at different junctions as a Scatter

As seen from the plot, there are certain junctions which show a smaller speed compared with others. To explore this further, I will use cluster to identify whether there is a relationship between these junctions in terms of location or other factors. As Long et al (2010) suggests, this can lead us to ‘discover hidden groups’.

To cluster, I first scaled the dataframe made from Appendix H into a new dataframe and used kmeans and the ‘elbow’ to determine the optimal number of clusters. The code for this can be found in Appendix I. As Verma (2023) explains, ‘the elbow is the point where the rate in WSS sharply changes.’

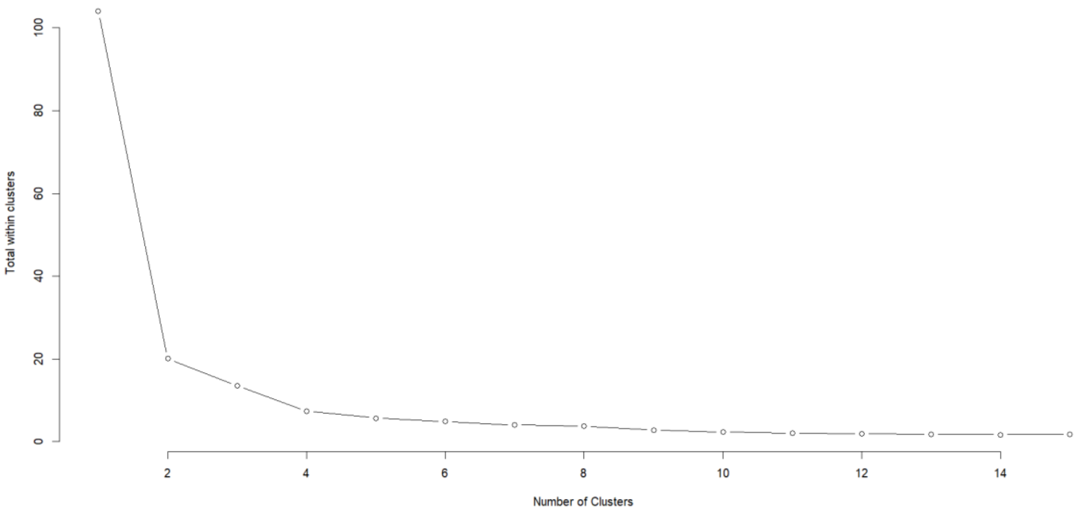
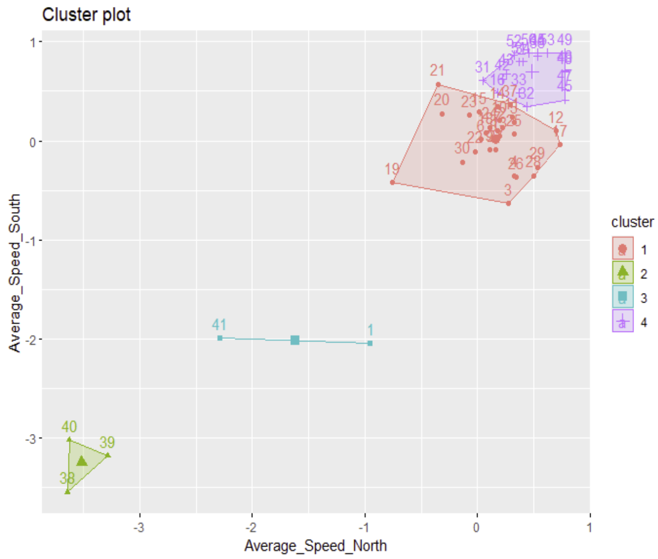


Figure 9 - Plotting of cluster points

From here, I determined that 4 would be the optimal number of clusters and plotted the clusters. The code can be found in Appendix J. Other options could have been 2 or 3.




```
   Junction Average_Speed_North Average_Speed_South Cluster
38       33           -3.641937           -3.545971       2
39       34           -3.283007           -3.179589       2
40       35           -3.625622           -3.022569       2
>
> df_cluster_3 <- df_clustered[df_clustered$Cluster==3,]
> df_cluster_3
   Junction Average_Speed_North Average_Speed_South Cluster
1         2          -0.9499635            -2.04555       3
41      35A          -2.2877929            -1.99321       3
```

Figure 10 - Cluster Plot and Accompanying Data regarding points in Clusters 2 and 3

As can be seen from the plot and accompanying results, there is a drop in speed from junctions 33 to 35A. After examining projects on National Highways (2023), they say they are ‘developing a programme to create additional emergency areas … (so) the left-hand will be closed throughout construction. Lanes two, three and four will remain open with a 50mph speed limit in place.’ This explains the drop in speed between these junctions on both the north and southbound carriageways. This is reported to end in Winter 2024. After examining this area of the M1 more closely, alternative routes may be quicker during this time, for example joining the M18 at junction 32 and travelling north up the A1. Further analysis on these two major roads could lead to lorries travelling this route to increase the average speed and decrease the time taken to travel the equivalent of the M1 through this period of roadworks. In addition, there is a second cluster which shows the average speed is slower than clusters 1 and 4 – cluster 3. Point 41 relates to the above-mentioned roadworks but shows a quicker speed as this is the end of the roadworks at least in one direction; vehicles naturally start speeding up once the roadworks have finished. The other notable junction within cluster 3 is point 1 (junction 2), near where the M1 starts and ends in London. This may be due to the large volume of traffic either leaving or joining in London.

I then conducted the final ANOVA result:

```
> anova_result <- aov(junction_speed_df$Average_Speed_North ~ junction_speed_df$Average_Speed_South)
>
> summary(anova_result)
                                      Df Sum Sq Mean Sq F value Pr(>F)    
junction_speed_df$Average_Speed_South  1  818.1   818.1   233.5 <2e-16 ***
Residuals                             51  178.6     3.5                   
```

From here, we can see that the test statistic is below 0.05. This supports our analysis and rejects the null hypothesis. Therefore, we can confidently say that the average speed at different junctions is not the same due to the above-mentioned reasons.

Before my concluding remarks, some considerations do need to be reiterated. For example, the data is taken from the month of November. A study over a long period of time may elicit further observations such as how the summer months affect road speed as well as the Christmas period. This could be achieved through a time series as Nielsen describes ‘to diagnose past behaviour as well as to predict future behaviour’ which could again provide competitive edges over competitors. Furthermore, when exploring hypothesis 1, there were only 2 data points per day of the week (for the sperate directions) which may not be representative; again, further data points may enhance the strength of argument. However, there should always be caution with this; as Sadkaoui (2018) states, ‘more data are not necessarily better data’.

In conclusion, the optimal days of the week to travel on are weekdays from the data given. From Figure 1, there is a pattern in that weekend travel leads to a decrease in average speed which affects the time taken for drivers to make the journey on the M1. Likewise, as seen from the linear model in Figure 7, certain predictions about the speed of the motorway northbound corresponds to the southbound carriageway. It is worth reiterating that the strength of correlation between the two was found to be 0.54 (2dp). Furthermore, currently roadworks are affecting the time it takes to travel between certain junctions, notably around Sheffield, as there is a speed limit 50mph. This has been identified by the junctions in cluster 2 from the plot in Figure 10. There is also a slow average speed when vehicles are either leaving or entering the M1 at junction 2. This is likely because of the high volumes of traffic leaving or entering London. Considerations for the clients are that alternative routes around Sheffield are at least explored. With the nearby M18 and A1, this could provide a quicker route, but similar analysis would need to be explored to identify whether this would be more efficient. Furthermore, with current technology in HGVs, there is potential for the linear model to be adapted and applied to certain routing algorithms. For example, if drivers reported their speed and location, alternative routes could be planned at an increased speed leading to an ultimately more efficient, profitable, and successful business operation.

Appendix for Question 10

Appendix A – Raw Data. I saved this data in two separate sheets. These are below. I had junction as the first column followed by speed data for the time period. The below are screenshots of PDFs saved from Excel.

Northbound

Junction	04/11/2023	05/11/2023	06/11/2023	07/11/2023	08/11/2023	09/11/2023	10/11/2023	11/11/2023	12/11/2023	13/11/2023	14/11/2023	15/11/2023	16/11/2023	17/11/2023	18/11/2023	19/11/2023	20/11/2023
2	45	63	45	57	47	61	54	43	46	43	57	50	52	44	45	60	41
4	63	62	68	52	69	60	55	56	48	69	69	70	59	60	35	56	54
5	69	56	66	56	61	69	62	58	50	61	59	63	55	53	50	61	65
6	57	58	56	55	60	60	68	67	55	65	55	53	55	54	40	70	63
6A	40	65	63	58	60	61	50	54	60	54	51	64	59	52	50	67	54
7	58	49	60	57	60	55	63	54	58	70	55	68	52	64	55	69	50
8	69	60	51	65	64	68	60	66	63	58	51	66	51	62	60	45	67
9	55	53	67	70	64	62	68	64	60	67	62	52	64	50	58	50	67
10	55	52	61	64	50	64	60	50	52	56	55	69	52	62	54	57	52
11	62	63	67	58	69	61	60	59	49	61	69	54	58	58	69	54	69
11A	53	61	55	62	58	61	60	60	56	59	56	70	64	63	68	59	58
12	61	60	68	59	70	64	50	60	60	69	69	63	62	67	54	58	54
13	56	61	57	67	56	64	56	45	60	67	56	53	60	56	40	52	55
14	53	63	54	67	50	58	54	40	69	68	55	64	57	59	52	63	70
15	60	66	53	50	69	54	61	45	57	58	51	69	52	67	60	49	66
15A	52	62	53	59	64	68	59	48	48	65	65	66	58	52	70	50	70
16	60	61	62	70	63	62	67	60	50	69	55	66	68	58	63	64	65
17	59	53	52	52	61	58	62	55	52	55	51	65	64	60	61	69	54
18	52	15	52	56	67	60	65	53	50	54	53	55	67	51	54	69	66
19	53	25	59	70	54	62	66	63	50	52	56	53	59	52	53	65	68
20	62	30	54	60	51	55	53	60	51	58	52	68	69	62	64	55	53
21	51	58	66	55	59	59	58	60	55	68	67	69	54	53	54	56	56
21A	52	56	68	51	58	64	54	58	56	52	51	62	54	63	51	58	69
22	50	59	63	66	65	66	52	57	60	67	63	58	66	50	60	59	56
23	51	63	70	54	51	67	62	66	51	64	66	59	59	53	56	60	63
23A	50	51	68	59	50	68	58	59	59	66	66	66	40	66	57	58	66
24	63	62	67	59	63	57	52	64	50	52	63	59	29	52	57	51	62
24A	57	60	55	69	55	70	67	66	57	69	56	62	56	69	60	62	60
25	63	58	64	54	54	67	59	68	52	62	56	56	50	57	60	65	56
26	50	51	54	65	67	50	67	59	69	56	65	67	59	50	51	57	50
27	59	39	60	56	68	55	63	62	68	57	64	56	55	61	61	55	59
28	66	50	60	65	68	58	67	59	66	70	56	64	61	67	59	52	70
29	55	50	60	64	64	64	60	63	60	66	56	69	57	58	55	50	64
29A	62	53	65	59	66	60	69	65	55	70	64	69	68	63	52	50	60
30	64	65	70	59	64	61	64	63	53	60	59	70	69	59	55	48	62
31	62	60	68	70	64	64	61	54	52	61	62	65	62	59	54	59	66
32	66	60	64	65	52	63	67	59	48	58	67	70	50	52	57	63	69
33	50	36	35	42	39	40	49	46	45	35	37	41	44	51	37	36	45
34	46	35	48	45	45	48	44	45	40	46	36	38	35	39	38	37	48
35	45	37	43	36	44	37	44	35	48	43	46	42	49	49	42	41	46
35A	50	50	48	40	50	52	45	51	40	43	46	44	42	44	48	42	40
36	63	62	58	55	63	67	64	60	35	66	66	56	68	62	63	40	66
37	68	57	63	56	70	59	63	63	30	69	60	58	68	67	55	51	63
38	65	63	65	67	68	63	60	66	52	67	65	65	66	66	65	55	60
39	58	64	60	67	61	65	61	62	51	65	62	62	63	61	70	69	65
40	63	65	68	60	65	61	61	61	50	67	66	65	66	68	65	61	63
41	45	62	67	69	63	69	62	60	50	60	65	64	60	68	61	66	65
42	50	66	68	65	64	65	64	68	52	69	63	67	69	65	68	65	65
43|44	65	68	69	66	69	65	65	65	29	69	70	69	66	70	66	68	65
45	70	67	70	70	67	65	68	40	68	66	70	69	68	65	68	70	66
46	61	62	65	69	70	68	69	35	67	65	67	68	68	69	69	69	70
47	60	69	70	69	70	65	65	32	69	70	68	68	66	65	65	66	68
48	58	67	69	65	65	68	66	50	67	68	67	66	65	71	68	65	68

21/11/2023	22/11/2023	23/11/2023	24/11/2023	25/11/2023	26/11/2023	27/11/2023	28/11/2023	29/11/2023	30/11/2023	01/12/2023
56	48	59	46	49	40	50	55	63	45	43
69	59	50	62	52	32	68	55	69	59	62
63	69	67	62	55	35	55	66	61	65	56
59	50	56	61	58	40	62	66	59	53	67
66	53	54	59	60	45	65	58	70	67	65
51	70	67	67	61	50	60	62	53	60	55
54	52	60	69	60	58	69	59	51	67	55
56	63	65	66	64	61	68	61	70	63	61
65	68	55	55	52	63	57	57	69	69	62
61	56	62	51	51	52	54	65	69	68	56
56	59	53	63	54	50	57	57	58	69	55
67	68	68	57	58	58	66	58	63	68	61
66	70	60	52	58	59	68	51	67	64	67
56	61	70	67	59	52	59	57	66	56	52
62	54	66	59	59	58	59	65	59	62	70
58	66	69	65	58	58	70	52	58	56	70
69	58	59	66	61	50	69	56	60	69	66
61	59	51	57	55	55	58	64	66	50	67
68	64	51	60	61	56	51	64	63	55	67
70	59	53	53	58	51	62	55	67	57	66
68	56	52	69	50	47	51	68	65	63	60
54	58	70	59	50	42	52	58	70	63	54
59	64	54	67	52	52	51	65	57	53	70
61	51	56	50	50	62	69	63	70	58	50
59	58	52	50	55	61	68	68	65	51	62
55	60	65	55	61	56	60	63	51	52	66
61	68	66	57	60	60	66	61	65	70	52
59	55	67	70	61	66	50	55	54	50	60
68	66	64	62	70	51	63	52	60	53	70
60	57	63	68	57	70	55	59	50	61	59
64	66	68	69	59	69	68	55	61	59	58
62	68	69	67	60	55	68	63	62	56	60
57	57	67	59	62	60	68	64	62	59	63
66	65	69	59	55	62	65	62	67	67	68
62	69	70	59	59	65	60	70	60	65	68
65	60	65	60	53	61	66	68	68	69	66
63	59	60	57	44	50	67	63	61	59	57
40	41	36	46	37	49	44	49	46	37	39
49	39	39	36	44	44	42	46	46	44	38
47	45	38	37	42	43	37	35	44	49	42
44	51	44	48	45	45	43	45	59	56	51
62	56	66	62	64	67	55	68	66	56	60
57	63	65	69	50	65	68	64	66	57	60
70	64	67	61	50	66	61	69	61	60	69
66	67	62	68	63	61	62	65	69	68	60
65	69	69	69	62	51	67	66	67	61	64
69	66	65	65	61	42	68	69	63	61	68
66	66	65	68	65	50	69	67	68	67	63
65	70	67	70	66	51	68	67	70	67	68
67	66	66	68	67	68	69	68	66	65	66
70	66	65	66	65	67	69	65	65	67	70
69	67	70	69	68	68	69	68	69	68	67
67	65	65	70	66	65	65	67	74	67	69

Southbound

Junction	04/11/2023	05/11/2023	06/11/2023	07/11/2023	08/11/2023	09/11/2023	10/11/2023	11/11/2023	12/11/2023	13/11/2023	14/11/2023	15/11/2023	16/11/2023	17/11/2023	18/11/2023	19/11/2023	20/11/2023
2	53	44	47	54	62	44	55	55	55	43	50	42	63	46	62	53	46
4	52	51	70	52	60	63	58	66	59	59	60	50	63	70	53	60	67
5	53	50	58	67	58	64	61	55	55	52	68	68	68	69	55	51	65
6	49	60	61	57	54	66	62	52	54	51	58	50	51	58	58	50	67
6A	59	59	66	70	67	69	62	50	61	64	63	62	63	64	55	47	54
7	55	64	68	51	69	62	69	59	51	63	59	55	62	55	51	45	61
8	65	52	65	56	57	51	59	60	64	62	70	63	60	69	63	51	54
9	59	54	59	61	51	69	58	60	66	60	55	69	58	59	54	41	51
10	55	58	65	59	69	50	55	63	53	70	60	65	53	65	50	40	61
11	50	52	64	54	60	59	54	55	51	66	55	70	66	50	55	67	64
11A	51	55	59	61	63	60	67	60	63	55	50	64	52	61	59	53	57
12	57	55	57	66	56	63	63	67	56	51	64	67	64	69	58	68	66
13	62	54	66	69	52	56	69	56	55	57	61	50	52	55	55	61	60
14	59	55	69	67	55	53	66	63	66	55	68	50	57	69	51	59	66
15	58	54	54	55	53	63	60	66	67	56	70	53	62	60	59	62	56
15A	61	60	52	64	54	66	65	59	54	67	65	61	54	61	61	67	55
16	68	53	65	68	60	51	55	57	54	58	70	61	54	63	49	42	63
17	53	59	69	54	68	54	53	53	51	63	55	53	56	67	62	50	60
18	68	53	53	50	57	51	63	50	52	64	66	58	51	68	61	50	50
19	54	60	67	61	50	57	62	60	55	50	55	59	50	51	55	69	64
20	52	59	66	56	68	64	57	58	51	67	70	59	57	64	58	56	63
21	62	55	57	56	50	66	60	67	59	55	69	70	70	53	57	50	60
21A	55	59	63	62	61	53	60	59	50	58	52	58	54	59	62	50	61
22	59	52	66	60	51	52	52	69	48	65	57	66	64	53	51	50	69
23	58	56	64	52	65	64	69	45	40	69	66	58	65	69	38	53	52
23A	50	59	69	56	66	54	50	40	39	53	58	53	52	54	40	58	63
24	61	58	52	58	66	59	61	50	44	50	57	54	50	70	50	50	66
24A	56	69	54	66	55	55	56	55	67	51	58	69	70	50	61	56	60
25	51	58	67	62	65	52	65	51	58	62	62	68	50	70	49	62	65
26	55	54	53	68	68	63	65	59	58	51	55	69	64	50	53	57	69
27	62	57	68	67	63	58	59	50	50	69	58	69	58	67	55	61	56
28	55	68	59	66	56	59	64	59	50	58	61	45	70	56	51	68	60
29	58	55	61	65	59	62	65	55	68	61	63	56	59	59	52	61	59
29A	50	59	60	60	62	64	67	59	67	68	63	60	68	64	62	59	60
30	55	60	60	61	66	68	65	66	61	67	60	68	69	61	64	66	70
31	53	67	61	65	63	59	60	65	61	61	68	63	65	67	61	62	59
32	51	61	67	50	56	56	67	68	67	61	56	62	70	58	57	61	65
33	41	40	37	36	49	35	48	39	42	41	40	48	39	44	37	43	47
34	42	49	44	47	49	49	43	39	37	41	42	48	43	41	43	36	40
35	47	46	46	46	47	36	47	44	41	49	46	44	37	45	37	42	44
35A	54	55	48	45	52	52	44	45	42	53	46	45	43	59	56	53	46
36	61	61	55	67	63	61	59	60	62	69	62	59	57	56	56	55	55
37	65	65	68	58	64	56	57	59	64	65	68	60	62	59	55	59	65
38	67	69	65	62	69	62	70	61	70	69	68	65	62	62	59	60	68
39	68	63	61	69	63	45	60	55	68	66	67	64	63	62	60	60	69
40	64	62	68	60	68	60	70	59	65	70	62	69	66	67	51	57	63
41	62	59	64	69	69	70	68	50	67	70	64	66	66	69	63	55	68
42	68	64	65	65	65	69	64	66	66	67	66	66	63	65	66	49	67
43|44	70	65	65	68	69	66	69	65	67	65	68	66	68	70	70	69	65
45	66	66	67	69	68	68	66	70	55	66	66	69	67	70	65	65	69
46	68	55	66	68	67	65	68	66	59	68	65	70	68	68	66	65	69
47	65	59	65	65	67	69	65	67	59	69	65	65	66	65	66	65	69
48	68	66	69	70	67	65	70	69	68	66	68	66	65	74	70	68	67

21/11/2023	22/11/2023	23/11/2023	24/11/2023	25/11/2023	26/11/2023	27/11/2023	28/11/2023	29/11/2023	30/11/2023	01/12/2023
55	54	45	51	44	41	62	56	40	51	47
64	63	57	58	48	50	65	69	60	70	63
50	56	51	61	40	55	54	56	50	54	67
61	63	66	68	49	52	51	54	67	51	51
56	53	67	69	51	51	69	55	57	67	70
69	64	69	57	54	53	64	70	52	69	59
70	66	52	64	53	53	67	51	65	56	64
68	63	57	58	50	54	51	52	59	50	59
52	61	66	65	54	52	54	55	59	65	58
62	57	61	50	53	65	58	66	62	57	70
67	63	70	54	52	52	50	63	64	54	58
55	55	50	68	50	67	54	57	65	64	64
61	69	51	68	59	58	67	64	61	62	54
62	53	67	65	57	59	66	50	67	55	60
57	63	60	64	51	50	70	59	61	67	66
65	65	61	62	50	64	63	68	53	51	55
66	57	60	64	56	59	63	55	50	56	62
65	63	65	50	52	50	59	62	61	54	70
69	62	51	62	69	52	51	56	51	59	60
65	57	69	70	66	51	54	53	57	55	57
57	58	68	61	70	54	57	57	70	66	63
58	52	58	53	52	60	55	59	68	54	51
69	55	63	50	51	55	69	51	69	50	57
61	70	63	51	52	59	60	57	58	57	64
56	69	70	51	70	56	61	57	53	70	65
68	56	51	66	50	59	60	64	67	59	66
62	68	67	64	45	66	63	62	65	51	64
58	57	65	62	32	68	52	66	63	51	67
52	54	69	65	30	55	58	67	53	59	63
57	66	55	62	49	70	54	65	55	54	56
59	61	65	68	61	65	65	56	67	63	66
55	61	60	60	57	58	70	57	62	55	63
61	63	59	64	59	59	55	57	60	57	59
64	66	64	59	61	59	66	65	59	65	59
61	66	64	68	64	51	68	67	69	61	70
67	67	59	59	70	51	65	67	67	66	62
53	51	66	51	53	52	65	70	65	51	65
44	39	41	38	49	49	42	51	43	40	41
43	49	38	42	43	35	44	39	42	38	47
36	43	44	43	43	39	49	38	43	48	49
53	42	45	46	42	53	46	40	41	53	41
66	62	69	65	65	54	64	55	55	59	68
70	70	62	56	66	52	68	64	65	64	57
67	66	63	60	60	51	62	66	64	61	62
70	60	64	69	53	61	70	70	61	69	61
64	67	63	60	54	61	69	63	64	64	64
60	65	65	61	55	68	60	65	64	65	69
63	69	64	66	65	69	65	65	69	67	66
70	67	67	69	65	69	68	66	69	65	67
67	69	68	69	69	65	67	69	68	65	70
70	66	67	65	65	68	68	65	67	69	68
69	69	70	69	65	70	66	68	67	66	68
70	69	65	65	70	68	65	71	65	65	70

Appendix B – Northbound and Southbound data loaded into RStudio with code for sampling strategy. Output is the new dataframe of the sampled data.

```
> n_df <- traffic_data_north # stores north speeds in a new df
> s_df <- traffic_data_south # stores south speeds in a new df
>
> n_selected_columns <- seq(1,ncol(n_df), by = 2) #systematic random sampling on the columns
> s_selected_columns <- seq(1,ncol(s_df), by = 2)
>
> sample_n_df <- n_df[,n_selected_columns] #creates a new df with the sampling now taken place
> sample_s_df <- s_df[,s_selected_columns]
>
> print(sample_n_df) #prints the sampled north data
# A tibble: 53 × 15
   Junction `05/11/2023` `07/11/2023` `09/11/2023` `11/11/2023` `13/11/2023` `15/11/2023` `17/11/2023` `19/11/2023` `21/11/2023` `23/11/2023`
   <chr>           <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
 1 2                  63           57           61           43           43           50           44           60           56           59
 2 4                  62           52           60           56           69           70           60           56           69           50
 3 5                  56           56           69           58           61           63           53           61           63           67
 4 6                  58           55           60           67           65           53           54           70           59           56
 5 6A                 65           58           61           54           54           64           52           67           66           54
 6 7                  49           57           55           54           70           68           64           69           51           67
 7 8                  60           65           68           66           58           66           62           45           54           60
 8 9                  53           70           62           64           67           52           50           50           56           65
 9 10                 52           64           64           50           56           69           62           57           65           55
10 11                 63           58           61           59           61           54           58           54           61           62
# ℹ 43 more rows
# ℹ 4 more variables: `25/11/2023` <dbl>, `27/11/2023` <dbl>, `29/11/2023` <dbl>, `01/12/2023` <dbl>
# ℹ Use `print(n = ...)` to see more rows
> print(sample_s_df) #prints the sampled south data
# A tibble: 53 × 15
   Junction `05/11/2023` `07/11/2023` `09/11/2023` `11/11/2023` `13/11/2023` `15/11/2023` `17/11/2023` `19/11/2023` `21/11/2023` `23/11/2023`
   <chr>           <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
 1 2                  44           54           44           55           43           42           46           53           55           45
 2 4                  51           52           63           66           59           50           70           60           64           57
 3 5                  50           67           64           55           52           68           69           51           50           51
 4 6                  60           57           66           52           51           50           58           50           61           66
 5 6A                 59           70           69           50           64           62           64           47           56           67
 6 7                  64           51           62           59           63           55           55           45           69           69
 7 8                  52           56           51           60           62           63           69           51           70           52
 8 9                  54           61           69           60           60           69           59           41           68           57
 9 10                 58           59           50           63           70           65           65           40           52           66
10 11                 52           54           59           55           66           70           50           67           62           61
# ℹ 43 more rows
# ℹ 4 more variables: `25/11/2023` <dbl>, `27/11/2023` <dbl>, `29/11/2023` <dbl>, `01/12/2023` <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Appendix C - Northbound and Southbound transformed data to show maximum speed of 60mph (the maximum speed a HGV can travel).

```
> for (col in names(sample_n_df)[-1]) {
+   sample_n_df[[col]][sample_n_df[[col]] > 60] <- 60 #changes speeds above 60 to 60 (the maximum a HGV can travel)
+ }
>
> print(sample_n_df) #prints the new sample_df
# A tibble: 53 × 15
   Junction `05/11/2023` `07/11/2023` `09/11/2023` `11/11/2023` `13/11/2023` `15/11/2023` `17/11/2023` `19/11/2023` `21/11/2023` `23/11/2023`
   <chr>           <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
 1 2                  60           57           60           43           43           50           44           60           56           59
 2 4                  60           52           60           56           60           60           60           56           60           50
 3 5                  56           56           60           58           60           60           53           60           60           60
 4 6                  58           55           60           60           60           53           54           60           59           56
 5 6A                 60           58           60           54           54           60           52           60           60           54
 6 7                  49           57           55           54           60           60           60           60           51           60
 7 8                  60           60           60           60           58           60           60           45           54           60
 8 9                  53           60           60           60           60           52           50           50           56           60
 9 10                 52           60           60           50           56           60           60           57           60           55
10 11                 60           58           60           59           60           54           58           54           60           60
# ℹ 43 more rows
# ℹ 4 more variables: `25/11/2023` <dbl>, `27/11/2023` <dbl>, `29/11/2023` <dbl>, `01/12/2023` <dbl>
# ℹ Use `print(n = ...)` to see more rows
>
> for (col in names(sample_s_df)[-1]) {
+   sample_s_df[[col]][sample_s_df[[col]] > 60] <- 60 #changes speeds above 60 to 60 (the maximum a HGV can travel)
+ }
>
> print(sample_s_df) #prints the new sample_df
# A tibble: 53 × 15
   Junction `05/11/2023` `07/11/2023` `09/11/2023` `11/11/2023` `13/11/2023` `15/11/2023` `17/11/2023` `19/11/2023` `21/11/2023` `23/11/2023`
   <chr>           <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
 1 2                  44           54           44           55           43           42           46           53           55           45
 2 4                  51           52           60           60           59           50           60           60           60           57
 3 5                  50           60           60           55           52           60           60           51           50           51
 4 6                  60           57           60           52           51           50           58           50           60           60
 5 6A                 59           60           60           50           60           60           60           47           56           60
 6 7                  60           51           60           59           60           55           55           45           60           60
 7 8                  52           56           51           60           60           60           60           51           60           52
 8 9                  54           60           60           60           60           60           59           41           60           57
 9 10                 58           59           50           60           60           60           60           40           52           60
10 11                 52           54           59           55           60           60           50           60           60           60
# ℹ 43 more rows
# ℹ 4 more variables: `25/11/2023` <dbl>, `27/11/2023` <dbl>, `29/11/2023` <dbl>, `01/12/2023` <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Appendix D – average speeds north and south

```
> # Finds the mean of the speeds per date
> average_speeds_north <- aggregate(value ~ variable, data = date_df_north_long, FUN = mean)
> average_speeds_south <- aggregate(value ~ variable, data = date_df_south_long, FUN = mean)
>
> # Renames the columns to allow for easier access below
> colnames(average_speeds_north) <- c("Date", "Average_Speed")
> colnames(average_speeds_south) <- c("Date", "Average_Speed")
>
> # Converts to date format
> average_speeds_north$Date <- as.Date(average_speeds_north$Date, format="%d/%m/%Y")
> average_speeds_south$Date <- as.Date(average_speeds_south$Date, format="%d/%m/%Y")
>
> #Shows north, sound and combined to be analysed below
> average_speeds_north
         Date Average_Speed
1  2023-11-05      54.16981
2  2023-11-07      56.77358
3  2023-11-09      58.01887
4  2023-11-11      54.69811
5  2023-11-13      57.15094
6  2023-11-15      57.30189
7  2023-11-17      56.32075
8  2023-11-19      55.09434
9  2023-11-21      57.84906
10 2023-11-23      56.83019
11 2023-11-25      55.62264
12 2023-11-27      56.98113
13 2023-11-29      58.20755
14 2023-12-01      57.20755
> average_speeds_south
         Date Average_Speed
1  2023-11-05      56.26415
2  2023-11-07      57.01887
3  2023-11-09      56.15094
4  2023-11-11      55.67925
5  2023-11-13      56.73585
6  2023-11-15      56.73585
7  2023-11-17      57.16981
8  2023-11-19      54.33962
9  2023-11-21      57.47170
10 2023-11-23      57.05660
11 2023-11-25      53.05660
12 2023-11-27      57.13208
13 2023-11-29      56.79245
14 2023-12-01      57.64151
```

```
> average_speeds_combined
         Date Average_Speed
1  2023-11-05      54.16981
2  2023-11-07      56.77358
3  2023-11-09      58.01887
4  2023-11-11      54.69811
5  2023-11-13      57.15094
6  2023-11-15      57.30189
7  2023-11-17      56.32075
8  2023-11-19      55.09434
9  2023-11-21      57.84906
10 2023-11-23      56.83019
11 2023-11-25      55.62264
12 2023-11-27      56.98113
13 2023-11-29      58.20755
14 2023-12-01      57.20755
15 2023-11-05      56.26415
16 2023-11-07      57.01887
17 2023-11-09      56.15094
18 2023-11-11      55.67925
19 2023-11-13      56.73585
20 2023-11-15      56.73585
21 2023-11-17      57.16981
22 2023-11-19      54.33962
23 2023-11-21      57.47170
24 2023-11-23      57.05660
25 2023-11-25      53.05660
26 2023-11-27      57.13208
27 2023-11-29      56.79245
28 2023-12-01      57.64151
```

Appendix E – Code for plotting the different visualisations to show the average speed on the different days of the week. These are combined, north and south.

```
# extracts so that the days of the week below can be applied
average_speeds_combined$DayOfWeek <- weekdays(average_speeds_combined$Date)
average_speeds_north$DayOfWeek <- weekdays(average_speeds_north$Date)
average_speeds_south$DayOfWeek <- weekdays(average_speeds_south$Date)

# Orders days of the week
ordered_days <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
average_speeds_combined$DayOfWeek <- factor(average_speeds_combined$DayOfWeek, levels = ordered_days)
average_speeds_north$DayOfWeek <- factor(average_speeds_north$DayOfWeek, levels = ordered_days)
average_speeds_south$DayOfWeek <- factor(average_speeds_south$DayOfWeek, levels = ordered_days)



# Plots the combined data
library(ggplot2)
ggplot(average_speeds_combined, aes(x = DayOfWeek, y = Average_Speed)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Speeds Over Days of the Week - combined",
       x = "Day of the Week",
       y = "Average Speed") +
  theme_minimal()

library(ggplot2)
ggplot(average_speeds_north, aes(x = DayOfWeek, y = Average_Speed)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Speeds Over Days of the Week - Northbound",
       x = "Day of the Week",
       y = "Average Speed") +
  theme_minimal()

library(ggplot2)
ggplot(average_speeds_south, aes(x = DayOfWeek, y = Average_Speed)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Speeds Over Days of the Week - Southbound",
       x = "Day of the Week",
       y = "Average Speed") +
  theme_minimal()
```

Appendix F – Code for plotting average north speeds and average south speeds along with correlation

```
> abline(model, col = 'red')
> plot(average_speeds_north$Average_Speed, average_speeds_south$Average_Speed)
> r <- cor(average_speeds_north$Average_Speed, average_speeds_south$Average_Speed)
> r
[1] 0.5401903
```

Appendix G – Code for plotting the above along with linear regression

```
> plot(average_speeds_north$Average_Speed,
+      average_speeds_south$Average_Speed,
+      main = "Scatterplot with Regression Line",
+      xlab = "North Bound Average Speed",
+      ylab = "South Bound Average Speed",
+      col = "blue")
>
> abline(model, col = 'red')
```

Appendix H – Code for plotting of junction speeds northbound and southbound

```
junction_speed_df <- data.frame( #creates a new dataframe with relevant data in
  Junction = sample_n_df$Junction,
  Average_Speed_North = sample_n_df$Average,
  Average_Speed_South = sample_s_df$Average
)

plot(junction_speed_df$Average_Speed_North,junction_speed_df$Average_Speed_South)
```

Appendix I – finding the optimal number of clusters

```
library('purrr')
wss <- function(k) {
  kmeans(scaled_df[, c("Average_Speed_North", "Average_Speed_South")], k)$tot.withinss
}

k <- 1:15

wssvalue <- map_dbl(k,wss)

plot(k,wssvalue,type="b",frame=FALSE,xlab = "Number of Clusters", ylab = "Total within clusters")
```

Appendix J – Plotting the clusters

```
fviz_nbclust(scaled_df[, c("Average_Speed_North", "Average_Speed_South")],kmeans,method='wss')
result <- kmeans(scaled_df[, c("Average_Speed_North", "Average_Speed_South")],4)

print(result)

fviz_cluster(result, data = scaled_df[, c("Average_Speed_North", "Average_Speed_South")])
```

Appendix K – Results of clustering

```
K-means clustering with 4 clusters of sizes 29, 3, 2, 19

Cluster means:
  Average_Speed_North Average_Speed_South
1           0.1588934          0.01677704
2          -3.5168556         -3.24937653
3          -1.6188782         -2.01938003
4           0.4831798          0.70001871

Clustering vector:
 [1] 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 1 2 2 2 3 4 4 4 4 4 4 4 4 4 4 4 4

Within cluster sum of squares by cluster:
[1] 4.6933656 0.2264411 0.8962634 1.5198247
 (between_SS / total_SS =  92.9 %)
```

Appendix L – identifying the junctions in clusters 2 and 3

```
clusters <- result$cluster

df_clustered <- cbind(scaled_df, Cluster = clusters)

df_cluster_2 <- df_clustered[df_clustered$Cluster==2,]
df_cluster_3 <- df_clustered[df_clustered$Cluster==3,]
```

Bilbliography for Question 10

Department for Transport (n.d.) Speed limits. Available at: https://www.gov.uk/speed-limits. (Accessed: 8th December 2023).

Elfil, Mohamed, and Ahmed Negida. (2017). *Sampling methods in Clinical Research; an Educational Review*. Emergency (Tehran, Iran) vol. 5,1 (2017): e52.

Hodeghatta, U., Nayak, U. (2023). *Practical Business Analytics Using R and Python: Solve Business Problems Using a Data-driven Approach*. New York. APress Media.

Jones, C., Morgan, J., Sprigge, R., Child, D., Hooper, I., Berry, P. (n.d.) *Stretching from London to the north, the M1 is Britain's first full-length motorway and possibly its most iconic.* Available at: https://www.roads.org.uk/motorway/m1#:~:text=Its%20original%20specification%20called%20for,it%20carries%20130%2C000%20to%20140%2C000. (Accessed: 5th December 2023).

Long, B, Zhang, Z, & Yu, PS 2010, *Relational Data Clustering : Models, Algorithms, and Applications*, CRC Press LLC, London. Available from: ProQuest Ebook Central. (Accessed: 17th December 2023).

Moto (n.d.) *When is the best time to travel on the motorway?* Available at: https://moto-way.com/2019/07/when-is-the-best-time-to-travel-on-the-motorway/. (Accessed: 10th December 2023).

Murray, E. (2019). *The importance of using color.* Available at: https://www.forbes.com/sites/evamurray/2019/03/22/the-importance-of-color-in-data-visualizations/?sh=4802ab8757ec. (Accessed 15th September 2023.)

National Highways (2023). M1 Junction 32 to 35a emergency area retrofit. Available at: https://nationalhighways.co.uk/our-roads/yorkshire-and-north-east/m1-junction-32-to-35a-emergency-area-retrofit/ (Accessed: 8th December 2023).

Nielsen, A. (2019) *Practical time series analysis : prediction with statistics and machine learning /*. First edition. Sebastopol, CA :: O'Reilly Media, Incorporated.

Rongpeng, L. (2020). *Essential Statistics for Non-STEM Data Analysts*. 1st Edition. Birmingham: Packt Publishing, Limited.

Rumsey, DJ 2011, *Statistics for Dummies*. Hoboken. John Wiley & Sons.

Verma, N. (2023). *Optimizing K-Means Clustering: A Guide to Using the Elbow Method for Determining the Number of Clusters*. Available at: https://blog.gopenai.com/optimizing-k-means-clustering-a-guide-to-using-the-elbow-method-for-determining-the-number-of-877c09b2c174. (Accessed: 29th December 2023).

Bibliography for Questions 1 – 9

Acock, A.C. (2005), *Working With Missing Values*. Journal of Marriage and Family, 67: 1012-1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Australian Bureau of Statistics (n.d.) *Statistical Language – Correlation and Causation* Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+correlation+and+causation (Accessed 15 December 2023).

Bhandari, P. (2022). Measures of Central Tendency. Available at: https://www.scribbr.co.uk/stats/measures-of-central-tendency/ (Accessed: 25th November 2023).

Black, B. (n.d.). Dealing with missing data. Available at: https://www.lancaster.ac.uk/~blackb/missingdata.html. (Accessed: 23rd December 2023).

Bourne, M. (2018). The Binomial Probability Distribution. Available at: https://www.intmath.com/counting-probability/11-probability-distributions-concepts.php (Accessed: 8th November 2023)

Busch, P 2008, Tacit Knowledge in Organizational Learning, IGI Global, Hershey. Available from: ProQuest Ebook Central. (Accessed: 2nd January 2024)

Calder, S. (2023). British Airways made £50 per second profit during first nine months of 2023. Available at: https://www.independent.co.uk/travel/news-and-advice/british-airways-profit-2023-iag-b2436997.html (Accessed: 30th November 2023).

Cherry, R. (2017). Study finds major benefits to taking workout classes vs. exercising alone. Available at: https://www.shape.com/fitness/trends/study-found-major-benefits-taking-class-vs-working-out-alone. Accessed (30th November 2023).

Columbia University, Department of Statistics (n.d.) Missing-data imputation. Available at: http://www.stat.columbia.edu/~gelman/arm/missing.pdf. (Accessed: 1st December 2023).

Frost, J. (2021). Relative Frequencies and Their Distributions. Available at: https://statisticsbyjim.com/basics/relative-frequency/. (Accessed: 29th November 2023)

Graham, J. (2009). Missing Data Analysis: Making It Work in the Real World. Available at: https://www.annualreviews.org/doi/pdf/10.1146/annurev.psych.58.110405.085530. (Accessed: 14th December 2023).

Gunner, J. (2022) Positive Correlation Examples in Real Life. Available at: https://www.yourdictionary.com/articles/examples-positive-correlation (Accessed 12 December 2023).

Humphries, M. (n.d.) Missing Data & How to Deal: An overview of missing data [online] (accessed 16 October 2023].

Imai, K. (n.d.). POL 345: Quantitative Analysis and Politics. Available at: https://imai.fas.harvard.edu/teaching/files/Handout8.pdf. (Accessed 2nd December 2023).

Indicative Team (n.d.). Volume of data. Available at: https://www.indicative.com/resource/volume-of-data/. (Accessed: 15th November 2023).

Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402

Knaflic, C. (2015). Storytelling with Data: A Data Visualisation Guide for Business Professionals. John Wiley and Sons.

Math Tutor (2017) Why Correlation does not Imply Causation in Statistics. Available at: https://www.mathtutordvd.com/public/Why-Correlation-does-not-Imply-Causation-in-Statistics.cfm (Accessed 12 December 2023).

Nguyen, J. (2017) *Regression Basics For Business Analysis* Available at: https://www.investopedia.com/articles/financial-theory/09/regression-analysis-basics-business.asp (accessed 30 November 2023).

Penn State (n.d.). Lesson 1: Simple Linear Regression. Available at: https://online.stat.psu.edu/stat501/book/export/html/639 (Accessed: 21st November 2023).

Sankar, S (2020). Big Data Consumer Analytics and the Transformation of Marketing. Available at: : https://www.researchgate.net/publication/348175860. (Accessed 17th December 2023).

The Pennsylvania State University (2022) '1.5 - The Coefficient of Determination, r^2'. Available at: https://online.stat.psu.edu/stat501/lesson/1/1.5 (Accessed 1 December 2023).

Walpole, R, Myers, R, Myers, S, & Ye, K 2016, Probability and Statistics for Engineers and Scientists, Global Edition, Pearson Education, Limited, Harlow. Available from: ProQuest Ebook Central.

Wamba, S., Akter, A., Edwards, A., Chopin E., Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics. Volume 165, Pages 234-246. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0925527314004253?via%3Dihub