# Business Intelligence Assignment

1. 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 | 45 | 62 |
| 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 | 1.03 | 1.33 |

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 a page or two) which summarises and interprets the findings, using all of the statistics given in the table above.

**Solution:**

Based on the data provided, a total of 107 gym participants participated in the survey. 45 gym participants attended exercise class and 62 gym participants attended gym only. Superficially, the average weight loss by the participants is higher for participants who attended the gym-only workouts. However, the standard deviation of gym-only workouts participants is higher than the standard deviation of those who participated in exercise classes. This implies that the weight lost by the participants who participated in gym-only workouts is highly dispersed than those who participated in exercise classes.

In addition, the value of the mode of both group of participants, which indicates the highest occurring weight loss amongst the participants who participated in exercise classes and gym-only workouts are 1.5kg and 1.7kg respectively.

To determine which type of gym workouts is effective as a weight loss program, we assumed that the sample participants represent the participants population and that the weight loss distribution is normally distributed as shown in figure below.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Based on the probability distribution curve shown above, the probability of losing the mean weight loss for exercise class group is ~0.38 while the probability of losing the mean weight loss of 2.5kg with the gym-only workouts is ~0.3. As a result, the probability of losing the mean weight loss for the exercise class group is higher than the gym-only workouts. This implies that the exercise class method is more effective as a weight loss program the gym-only workouts.

2. Describe a way to deal with missing data values in data for processing it. [No more than 500 words]

**Solution:**

In data cleaning, there are three broad options for dealing with missing values. These options are (McKinney):

* Delete the rows with the missing values.
* Delete the columns with the missing values.
* Replace the missing values with some values such as zero, mean of the affected column, median of the affected column, interpolated value between the penultimate values and next available data in the affected columns.

Most common approach to dealing with missing values is to replace the missing values with the averages of the affected columns. As a result, the following steps are taken to address the missing values:

1. Find the columns with missing values.
2. Compute the average value of the columns with missing values neglecting the missing values.
3. Replace the missing values with the computed averages in step ii.

3. 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: (10)

1. region 5
2. region 3
3. regions 1 and 2 together
4. regions 4 and 7 together
5. regions 3, 6, 7, and 8 together

Diagram, venn diagram

Description automatically generated

**Solution:**

1. region 5: Is the event that the family experiences a mechanical problem but did not receive a traffic violation ticket and will not arrived at a campsite with no vacancies. In other words, the family experienced a mechanical problem only.
2. region 3: Is the event that the family will arrive at a campsite with no vacancies and received a traffic violation ticket but did not have a mechanical problem.
3. regions 1 and 2 together: The event that the family experiences a mechanical problem and arrived at the campsite with no vacancies.
4. regions 4 and 7 together: The event that the family received a traffic violation ticket but did not arrive at the campsite with no vacancies.
5. regions 3, 6, 7, and 8 together: The event that the family did not experience a mechanical problem during the trip.

4. According to Consumer Digest Magazine the probable location of an LCD TV in the home is as follows:

Adult bedroom: 0.15

Child bedroom: 0.07

Other bedroom: 0.12

Office or den: 0.38

Other rooms: 0.28

(a) What is the probability that an LCD TV is in a bedroom? (2)

(b) What is the probability that it is not in a bedroom? (2)

(c) Suppose a household is selected at random from households with an LCD TV; in what room would you expect to find an LCD TV? (2)

**Solution:**

1. Since the probability of having an LCD Tv in the office or den is higher than other bedroom as presented above, I would expect the office or den will be selected at random.

5. The probability that an iPhone will survive a shock test is 0.70. Find the probability that exactly 3 of the next 5 iPhones tested survive. These tests are independent.

**Solution:**

Let be the probability that the iPhone survives the shock test and be the probability that the iPhone will not survive the shock test, then

Since the possible outcome of the events is either a success or failure, the probability distribution of the random variable is a binomial probability distribution. As a result, for trials, the probability that the iphone survives (success) exactly times is given as:

6. 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.

**Solution:**

Null Hypothesis : The probability of new private residence homes with 3 bedrooms is 0.64

Alternate Hypothesis : The probability of new private residence homes with 3-bedrooms is less than 0.64

Based on the hypothesis statement, the null hypothesis is rejected if the probability of the new private residence homes computed from the samples is . However, if the probability of new private residence homes computed is , then the test failed to reject the null hypothesis . This implies that the critical value of the hypothesis testing is .

7. 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.

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

**Solution:**

R-code:

A picture containing chart

Description automatically generated

The correlation between adverts and purchases is 0.6790. There is a positive correlation between the number of advertisements and number of purchases made as shown above. This implies that there is a relationship between number of advertisements and number of purchases made in the company. The R-code used to compute the correlation is provided in the appendix.

8. 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 | 5.0 | 4.0 |
| 3 | 8.0 | 4.5 |
| 4 | 2.0 | 2.0 |
| 5 | 4.0 | 2.5 |
| 6 | 4.5 | 3.0 |

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

(b) Estimate the expected sales for a region where 6.3 thousand pounds are beingspent on advertising.

**Solution:**

1. R-code:

Graphical user interface, text

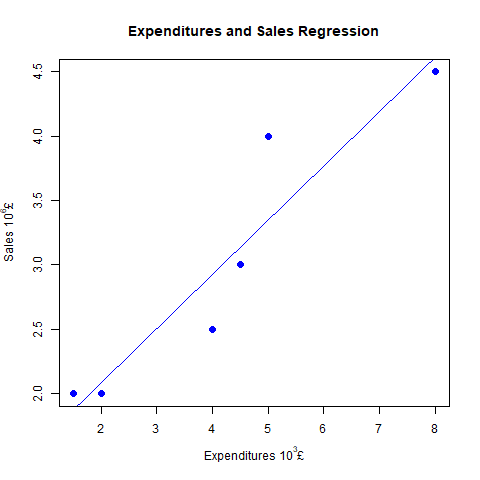
Description automatically generated

To compute the relationship between the advertising expenditure and sales, the linear regression model from the r-packages is used with the data provided as shown in the appendix.

The summary of the linear regression model is given as Text, letter

Description automatically generated

The plot of the linear regression is shown below



The linear relationship between the sales and expenditure of the marketing is given as:

Where and .

(b) Estimate the expected sales for a region where 6.3 thousand pounds are being spent on advertising.

R-Code:

Text, letter

Description automatically generated

Expected sales with expenditure of £6300 is

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

(a) Find the number of clusters for the given data (4)

(b) Cluster the given data and comment on each cluster of the data (8)

[Hint: Read this paper – Anwar, Naveed, Oakes, Michael, Wermter, Stefan and Heinrich,Stefan (2010) Clustering audiology data. In: 19th Annual Machine Learning Conference ofBelgium and TheNetherlands, 27-28 May 2010, Leuven. Web link:https://dtai.cs.kuleuven.be/events/Benelearn2010/submissions/benelearn2010\_submission\_7.pdf ]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Freq250 | Freq500 | Freq1K | Freq2K | Freq4K | Freq8K |
| 50 | 50 | 55 | 50 | 50 | 75 |
| 30 | 30 | 25 | 55 | 70 | 130 |
| 35 | 50 | 45 | 60 | 80 | 130 |
| 55 | 60 | 70 | 60 | 95 | 90 |
| 40 | 30 | 30 | 30 | 70 | 85 |
| 30 | 30 | 40 | 35 | 55 | 65 |
| 35 | 25 | 35 | 35 | 55 | 75 |
| 35 | 35 | 50 | 50 | 80 | 100 |
| 65 | 65 | 65 | 65 | 70 | 70 |
| 40 | 40 | 35 | 50 | 60 | 65 |
| 20 | 20 | 40 | 50 | 85 | 130 |
| 35 | 35 | 35 | 55 | 60 | 75 |
| 25 | 20 | 20 | 65 | 70 | 55 |
| 50 | 45 | 40 | 25 | 30 | 70 |
| 30 | 25 | 30 | 35 | 50 | 75 |
| 70 | 70 | 70 | 60 | 85 | 85 |
| 55 | 55 | 55 | 55 | 65 | 130 |
| 40 | 35 | 45 | 45 | 55 | 65 |
| 45 | 40 | 40 | 50 | 65 | 85 |
| 60 | 75 | 65 | 65 | 75 | 130 |
| 20 | 10 | 15 | 55 | 60 | 70 |
| 35 | 50 | 65 | 65 | 70 | 95 |
| 65 | 65 | 65 | 65 | 70 | 75 |
| 60 | 55 | 30 | 40 | 65 | 70 |
| 30 | 30 | 40 | 30 | 70 | 55 |
| 55 | 45 | 50 | 40 | 35 | 35 |
| 45 | 35 | 40 | 55 | 80 | 130 |
| 15 | 15 | 10 | 55 | 65 | 70 |
| 50 | 55 | 60 | 60 | 70 | 85 |
| 35 | 35 | 50 | 55 | 70 | 85 |
| 65 | 60 | 55 | 55 | 60 | 65 |
| 60 | 50 | 35 | 50 | 60 | 95 |
| 20 | 20 | 10 | 30 | 75 | 75 |
| 20 | 30 | 30 | 45 | 65 | 80 |
| 35 | 30 | 30 | 60 | 60 | 65 |
| 30 | 45 | 50 | 50 | 50 | 80 |
| 50 | 65 | 55 | 50 | 50 | 80 |
| 50 | 60 | 60 | 60 | 70 | 70 |
| 50 | 55 | 65 | 70 | 75 | 130 |
| 15 | 15 | 25 | 60 | 70 | 55 |
| 75 | 75 | 75 | 70 | 75 | 70 |
| 20 | 15 | 15 | 55 | 65 | 75 |
| 70 | 65 | 75 | 70 | 65 | 80 |
| 20 | 20 | 40 | 45 | 50 | 60 |
| 30 | 25 | 20 | 30 | 45 | 55 |
| 30 | 30 | 40 | 55 | 75 | 95 |
| 45 | 45 | 45 | 65 | 80 | 70 |
| 30 | 20 | 25 | 35 | 70 | 80 |
| 15 | 15 | 25 | 55 | 90 | 95 |
| 80 | 60 | 45 | 45 | 65 | 130 |

**Solution:**

1. Find the number of clusters for the given data (4)

Table

Description automatically generatedText

Description automatically generated

Graphical user interface, text

Description automatically generated

To determine the number of clusters for the given datasets of the audiograph, there are three methods available such as the Elbow method, Silhouette method, and the Gap statistic. However, in the solution provided below, the Silhouette method was used to determine the optimum cluster size required for the datasets.

The Silhouette method computes the quality of the clusters by determining how closely knitted the points are within the clusters. In this regard, the optimized cluster size has the highest average Silhouette width in relation to other cluster size.

Computing the average Silhouette width for different cluster size as shown below, the optimized cluster size was found to be 2.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

1. Cluster the given data and comment on each cluster of the data

Text

Description automatically generatedText

Description automatically generated

For optimized clustering, a cluster size of 2 with maximum average Silhouette value of 0.34 as shown in the figure above. Using the cluster size of 2, the mean audiogram of each cluster at the frequencies are shown in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Clusters | Freq250 | Freq500 | Freq1K | Freq2K | Freq4K | Freq8K |
| 1 | 31.03448 | 28.62069 | 31.72414 | 45.68966 | 62.24138 | 75.00000 |
| 2 | 55.47619 | 57.14286 | 57.85714 | 59.28571 | 70.95238 | 94.52381 |

The silhouette widths for cluster size of 2 and 5 are shown in the figure below

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

The cluster plot of the audiograms provided is shown in the figure below.

Chart

Description automatically generated

10. A scenario (30)

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.

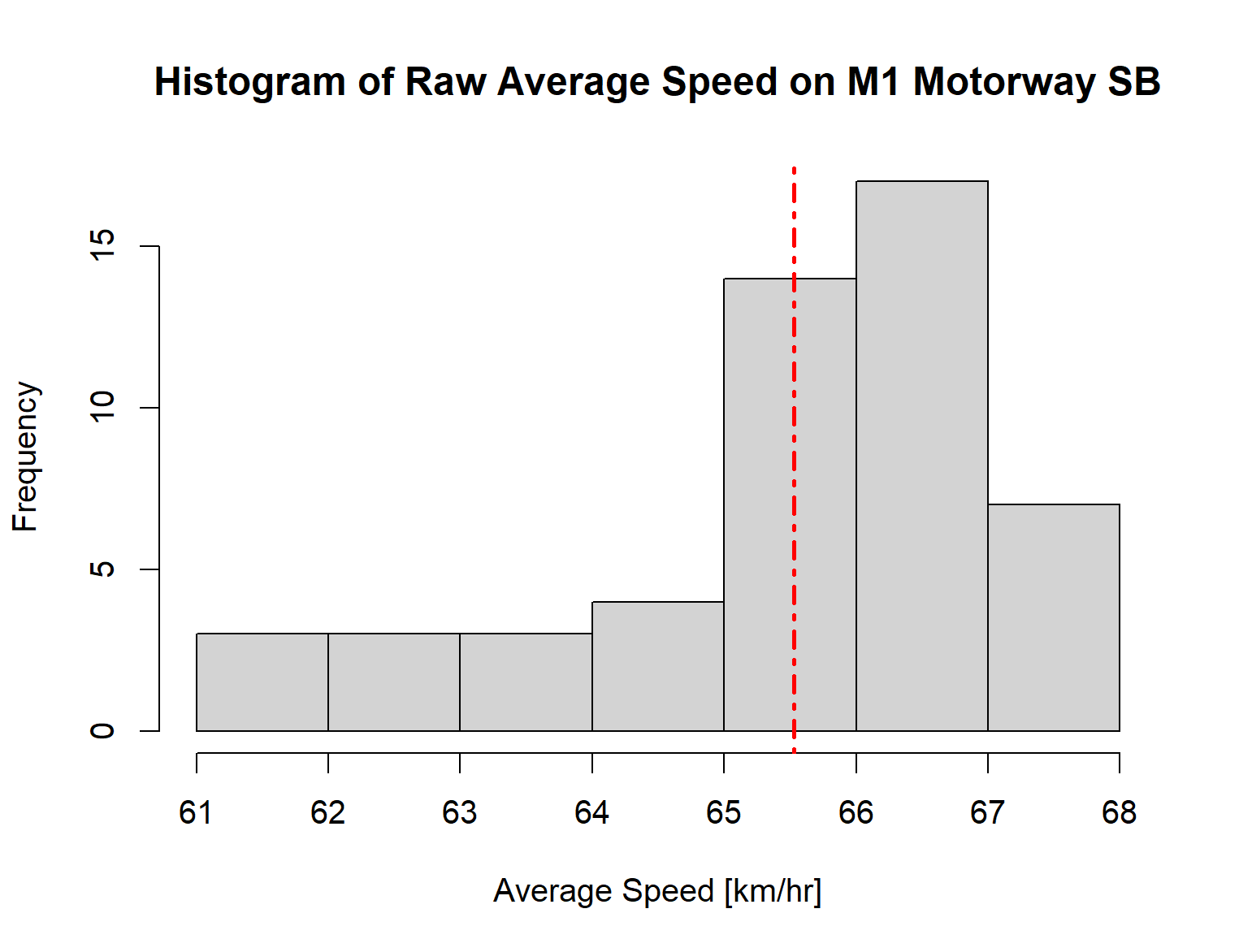
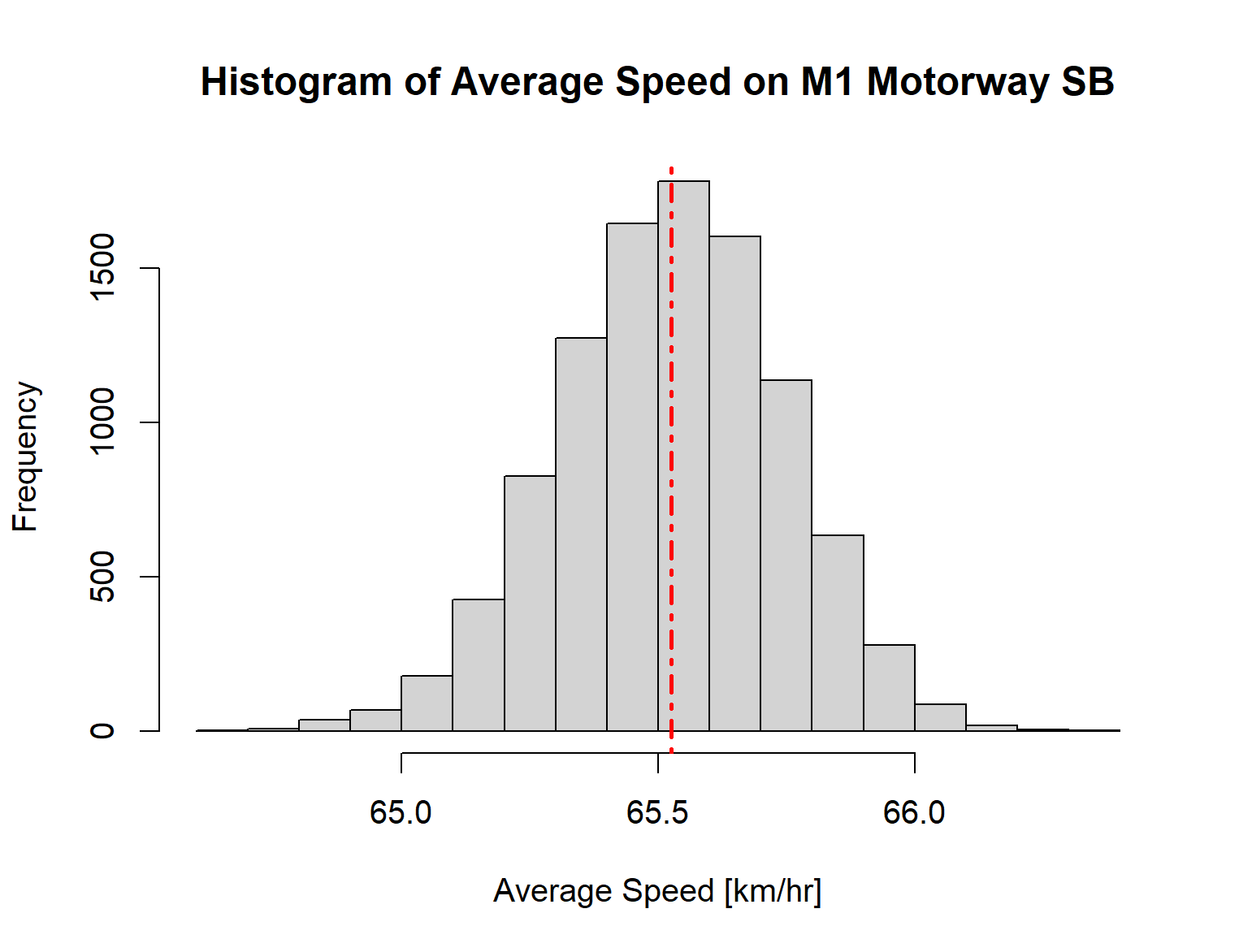
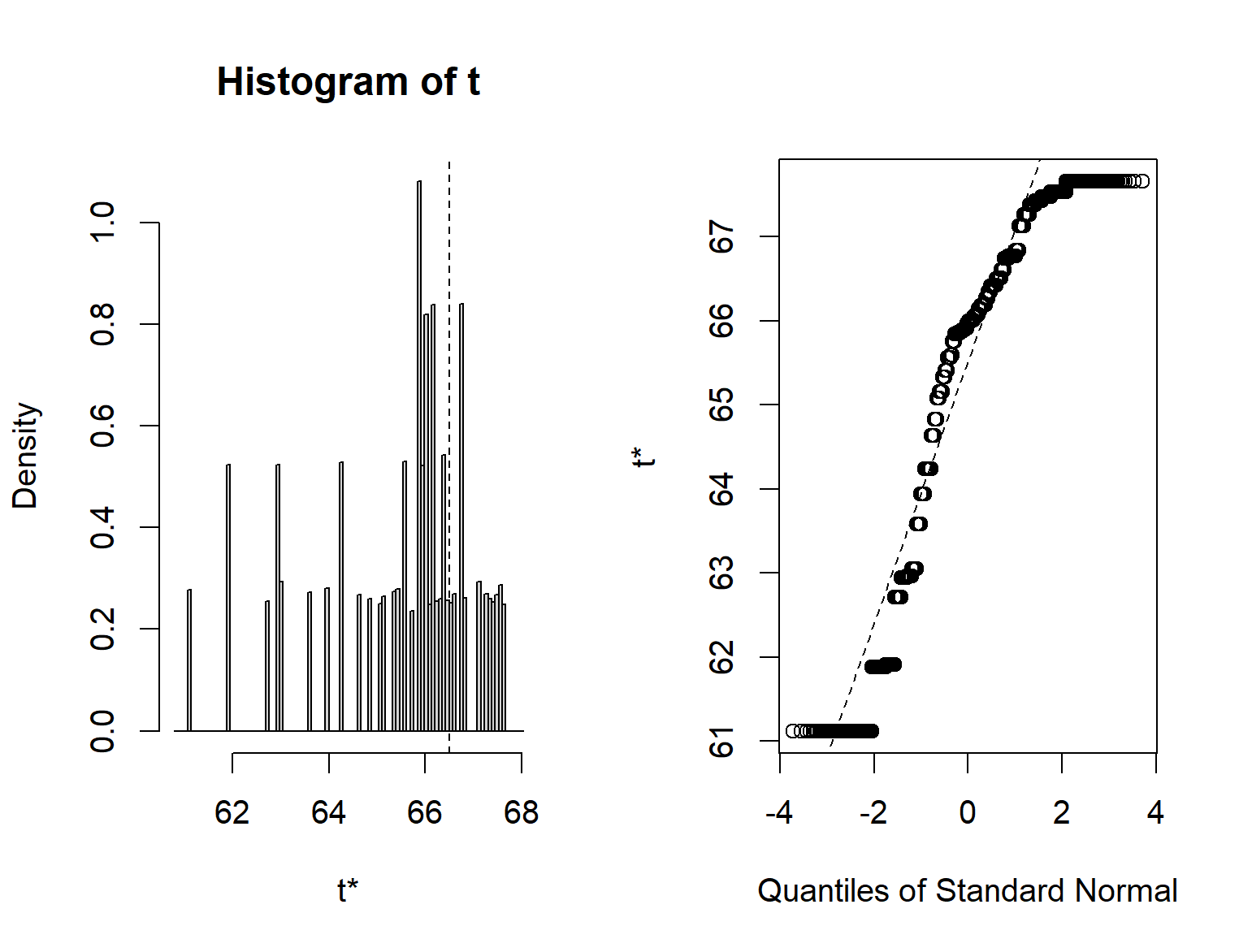
**Solution:**

The delivery of lorries that travel the length of M1 to London depends on the traffic condition – speed on the M1 road at each junction of the motorway. The average speed on the motorway was investigated by collecting data of the motorway reported by the traffic regulatory body at <https://www.trafficengland.com/traffic-report> at each junction for a duration of 25hr at an interval of 30mins as shown in the data provided in github repo [problem10\_05-08-2023\_01\_19\_10](https://github.com/ebereokafor4/KC7021_Stats_Bus_Intelligence_Assignment/blob/main/Data/Problem10/problem10_05-08-2023_01_19_10.csv) (Okafor).

The collected data was cleaned for the motorway M1 was cleaned up for the average speed of motorway at each junction for the SouthBound and NorthBound direction. This data is available on github at this location [SouthBound](https://github.com/ebereokafor4/KC7021_Stats_Bus_Intelligence_Assignment/blob/main/Data/Problem10/upstream_problem10_05-08-2023_01_19_10.csv) and [NorthBound](https://github.com/ebereokafor4/KC7021_Stats_Bus_Intelligence_Assignment/blob/main/Data/Problem10/downstream_problem10_05-08-2023_01_19_10.csv) (Okafor). The cleanup was analyzed to estimate the parametric mean average speed using a sampling method called random sampling with replacement also known as Bootstrap Sampling. This approach was used because the data being used does not necessarily have to meet the assumptions of normally distributed data to compute the statistics parameter – mean (Bruce, Bruce and Gedeck).

This sampling approach repeatedly selected samples from the provided data for SouthBound and NorthBound traffic collected while reshuffling the data in the process. The analysis was conducted using R-programming and the script used for the analysis is available in the github repo at this location [problem10\_code](https://github.com/ebereokafor4/KC7021_Stats_Bus_Intelligence_Assignment/blob/main/Code/Problem10/problem10_code.R) (Okafor).

Based on the analysis on the SouthBound(Upstream) data, the average speed of the traffic travelling along the M1 motorway is computed as 65.53km/hr with a standard deviation of 1.564. With 95% confidence level, the average speed of the lorries travelling along M1 motorway in the SouthBound direction is within the interval of .



As shown, the bootstrap sampling method, generated a normally distributed data from a raw data which is not normally distributed.

Additionally, the analysis on the NorthBound (Downstream) data, the average speed of the traffic travelling along the M1 motorway is computed as 65.73km/hr with a standard deviation of 1.803. With 95% confidence level, the average speed of the lorries travelling along M1 motorway in the NorthBound direction is within the interval of .

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

As shown, the bootstrap sampling method, generated a normally distributed data from a raw data which is not normally distributed.

Appendix: R Code

1. Problem 7 Code

**R-code:**

adverts <- c(10, 7, 6, 5, 4, 3, 2, 0)

purchases <- c(12, 3, 8, 10, 5, 4, 1, 4)

correlation <- cor(adverts, purchases, method = "pearson")

print(correlation)

1. Problem 8 Code

# Problem 8: Linear Regression

appliance.data <- data.frame(

Region = c(1:6), Expenditure = c(1.5, 5.0, 8.0, 2.0, 4.0,4.5),

Sales = c(2.0,4.0,4.5,2.0,2.5,3.0), stringsAsFactors = FALSE )

appliance.data

relation <- lm(formula = Sales ~ Expenditure,

data = appliance.data)

# Give the chart file a name.

png(file = "problem8\_Sales\_Expenditure\_regression.png")

# Plot the chart.

plot(appliance.data$Expenditure,appliance.data$Sales,col = "blue",main = "Expenditures and Sales Regression",

abline(relation, col="blue"),cex = 1.3,pch = 16,xlab = expression(paste("Expenditures ", 10^3, "£")),

ylab = expression(paste("Sales ", 10^6, "£")) )

# Save the file.

dev.off()

print(summary(relation))

Call:

lm(formula = Sales ~ Expenditure, data = appliance.data)

Residuals:

1 2 3 4 5 6

0.12195 0.64939 -0.11280 -0.08841 -0.42988 -0.14024

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.24695 0.36417 3.424 0.02668 \*

Expenditure 0.42073 0.07779 5.409 0.00566 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4067 on 4 degrees of freedom

Multiple R-squared: 0.8797, Adjusted R-squared: 0.8496

F-statistic: 29.25 on 1 and 4 DF, p-value: 0.00566

The plot of the linear regression is shown below

1. Problem 10

#install.packages('boot',dep=TRUE)

library(boot)

# Load dplyr

library('dplyr')

library(ggplot2)

mydata.SB <- read.csv("upstream\_problem10\_05-08-2023\_01\_19\_10.csv");

mydata.NB <- read.csv("downstream\_problem10\_05-08-2023\_01\_19\_10.csv");

mydata\_df.SB <- mydata.SB %>% select("X01.J1":"X53.J47");

mydata\_df.NB <- mydata.NB %>% select("X01.J1":"X53.J47");

head(mydata\_df.SB)

# Define the function that calculates the metric of interest - mean speed on the

# motorway M1 from J1 - J48

average\_speed <- function(data, i){

d2 <- data[i,]

return(rowMeans(d2))

}

# Set the seed for repeatability

set.seed(412)

# Using Random Sampling with Replacement

upstream.bootstrap <- boot(mydata\_df.SB,average\_speed,R=10000);

downstream.bootstrap <- boot(mydata\_df.NB,average\_speed,R=10000);

upstream.bootstrap

downstream.bootstrap

# Display the summary of the sampling of the upstream average speed

summary(upstream.bootstrap)

# Display the summary of the sampling of the upstream average speed

summary(downstream.bootstrap)

# Compute the statistics of the sampling process

print(range(upstream.bootstrap$t))

print(mean(upstream.bootstrap$t))

print(sd(upstream.bootstrap$t))

# Confidence Interval of the SouthBound average Speed

boot.ci(boot.out=upstream.bootstrap,type=c('norm','basic','perc','bca'))

# Compute the statistics of the sampling process

print(range(downstream.bootstrap$t))

print(mean(downstream.bootstrap$t))

print(sd(downstream.bootstrap$t))

# Confidence Interval of the NorthBound average Speed

boot.ci(boot.out=downstream.bootstrap,type=c('norm','basic','perc','bca'))

# Plot the SouthBound Average speed Histogram

plot(upstream.bootstrap)

hist(rowMeans(upstream.bootstrap$t), xlab='Average Speed [km/hr]',

main='Histogram of Average Speed on M1 Motorway SB')

abline(v=mean(upstream.bootstrap$t), lty=4, lwd=2, col='red' )

hist(rowMeans(upstream.bootstrap$data), xlab='Average Speed [km/hr]',

main='Histogram of Raw Average Speed on M1 Motorway SB')

abline(v=mean(rowMeans(upstream.bootstrap$data)), lty=4, lwd=2, col='red' )

# Plot the NorthBound Average speed Histogram

plot(downstream.bootstrap)

hist(rowMeans(downstream.bootstrap$t), xlab='Average Speed [km/hr]',

main='Histogram of Average Speed on M1 Motorway NB')

abline(v=mean(downstream.bootstrap$t), lty=4, lwd=2, col='red' )

hist(rowMeans(downstream.bootstrap$data), xlab='Average Speed [km/hr]',

main='Histogram of Raw Average Speed on M1 Motorway NB')

abline(v=mean(rowMeans(downstream.bootstrap$data)), lty=4, lwd=2, col='red' )