Table of Contents

[1.0 INTRODUCTION 2](#_Toc128788309)

[2.0 ANALYSIS BACKGROUND 2](#_Toc128788310)

[2.1 PURPOSE 2](#_Toc128788311)

[2.2 OBJECTIVES 2](#_Toc128788312)

[3.0 SUITABLE DATASETS 3](#_Toc128788313)

[3.1 DATASET 3](#_Toc128788314)

[3.2 STRUCTURE OF DATASET 3](#_Toc128788315)

[4.0 EXPLORATORY DATA ANALYSIS 4](#_Toc128788316)

[4.1 UNIVARIATE ANALYSIS 4](#_Toc128788317)

[4.1.1 CATEGORICAL VARIABLE 4](#_Toc128788318)

[4.1.2 NUMERICAL VARIABLE 5](#_Toc128788319)

[4.2 BIVARIATE ANALYSIS 7](#_Toc128788320)

[4.2.1 CATEGORICAL VS NUMERICAL 7](#_Toc128788321)

[4.2.2 NUMERICAL VS NUMERICAL 8](#_Toc128788322)

[4.4 CLEANING TO TREAT QUANTITIES OF ZEROS IN THE DATASET 9](#_Toc128788323)

[5.0 DATA VISUALIZATION & PRESENTATION 12](#_Toc128788324)

[5.1 FIRST VISUALIZATION (Usage vs Gender) 12](#_Toc128788325)

[5.2 SECOND VISUALIZATION (Fitness VS Expected to run (Miles)) 14](#_Toc128788326)

[6.0 REFERENCE 16](#_Toc128788327)

[7.0 APPENDICES 17](#_Toc128788328)

# INTRODUCTION

About Assignment The world is becoming more and more data-driven, with endless amounts of data available to work with. Data analysis is widely used in almost every aspect of our life, and it has been used by small businesses, retail companies, in medicine, and even in the world of sports. Due to the importance and valuable data nowadays, data analysis has become an important field. This assignment is required to complete a data analysis project by using R programming to analyse a dataset that has been choose and produce a report based on the finding.

# 2.0 ANALYSIS BACKGROUND

## 2.1 PURPOSE

The major goal of this analysis is to investigate the correlation of treadmill usage and how far the expected run(miles) based on fitness.

## 2.2 OBJECTIVES

Objectives that can be achieved during this analysis:

1. To investigate how many times treadmill product is being used for every week to train user fitness.
2. To investigate the fitness based on expected run (Miles).

2.3 TARGET AUDIENCE

The following are some of target populations for fitness studies:

1. **People who want to train their fitness or cardiovascular:**

Fitness studies can help in determining the suitable usage of fitness product, which can help person’s to properly train his fitness to the maximum extent.

1. **Athletes:**

Fitness studies can assist athletes in getting the best possible fitness order to enhance performance and lower their risk of injury.

# 3.0 SUITABLE DATASETS

## 3.1 DATASET

**CARDIO GOOD FITNESS DATASET**

This dataset has been obtained from **“Kaggle” (**[**https://www.kaggle.com/**](https://www.kaggle.com/)**)** an online community platform for data scientists and machine learning enthusiasts. The specific for the dataset is:

**(**[**https://www.kaggle.com/datasets/vsridevi/cardio-good-fitness**](https://www.kaggle.com/datasets/vsridevi/cardio-good-fitness)**)**

The dataset includes details on a group of test volunteers' fitness customer. There is nine variables in this dataset which is “**Product**” for the model no. of the treadmill, “**Age**” in number of years of the customer, “**Gender**” of the customer, “**Education**” , “**Marital Status**” of the customer, “**Usage**” is for average times the customer wants to use the treadmill every week, “**Fitness**” is for self rated fitness score of the customer the score will be (5 - very fit, 1 - very unfit), “**Income**” of the customer and “**Miles**” for expected to run.

## 3.2 STRUCTURE OF DATASET

Text

Description automatically generated

*Figure 1 shows a screenshot of command str() for cardio good fitness dataset.*

|  |  |  |
| --- | --- | --- |
| **USED VARIABLE IN THIS ANALYSIS** | | |
| **NUM** | **VARIABLE NAME** | **DATA TYPE** |
| 1. | Product | chr |
| 2. | Gender | chr |
| 3. | MaritalStatus | chr |
| 4. | Usage | int |
| 5. | Fitness | int |
| 6. | Miles | int |

*Table 1 shows the variable used in the dataset.*

# 4.0 EXPLORATORY DATA ANALYSIS

## 4.1 UNIVARIATE ANALYSIS

A statistical technique called univariate analysis is used to analyse and understanding the properties of a single variable or feature in a dataset. It includes using multiple statistical methods, including descriptive statistics, histograms, box plots, and probability density functions, to examine the distribution, central tendency, and variability of a single variable.

### 4.1.1 CATEGORICAL VARIABLE

A categorical variable is a type of variable that represents data in the form of categories or groups. Categorical variables are qualitative variables and do not have numerical values. Instead, they represent characteristics such as gender, race, colour, country of origin, education level, marital status, etc.

In the cardio good fitness dataset ***“Gender”*** is an example of a categorical variable.

Chart, bar chart

Description automatically generated

*Figure 2 shows a bar chart about gender in cardio good fitness dataset*

Based on *Figure 2,* a bar chart is used for this specific data as the data is categorical which is gender. Product consists of two element, **Female and Male.** As observed above, we can easily identify that Femail is the red bar, and the blue bar represents the Male. Both are over 50 & specifically, the count of Female is 75, and Male is 105. A total amount of 180 customers. We also can see that Male customer is greater than Female customer.

### 4.1.2 NUMERICAL VARIABLE

A numerical variable is a type of variable that represents data with numerical values. Numerical variables can be further divided into two subtypes: discrete and continuous variables. Numerical variables are often analysed using various statistical methods, including descriptive statistics, correlation analysis, regression analysis, and hypothesis testing. These statistical methods are used to describe the distribution, central tendency, and variability of the data, as well as to test hypotheses about relationships between different numerical variables.

In the cardio good fitness dataset ***“Fitness”***. an example of a numerical variable.

Chart, line chart, histogram

Description automatically generated

*Figure 3 shows a histogram about the variable Fitness.*

Based on *Figure 3,* shows a histogram that is used specifically for numerical data which on this case is customer Fitness. The x-axis consists of 4 items which is 0, 2, 4 & 6. As observed, the highest count would be 3 at the middle between 2 and 4 the amount is nearly 100 counts. Beside that, the lowest would be 1 there is between 0 and 2 gathering a total count below 10. The diagram tells us that the Fitness is slightly skewed to the right. To confirm that the diagram is skewed to the right we will use box plot.

Graphical user interface, chart

Description automatically generated

*Figure 4 shows a box plot about the variable Fitness.*

*Text

Description automatically generated with medium confidence*

*Figure 5 shows a summary about the variable Fitness.*

Based on Figure 4, we can see that the box is stretch more to the right this show that it is true the diagram on Figure 3 is skewed to the right. As a result, we can see on Figure 5, the mean is greater than the median. That is way the box plot is stretch more to the right and make it a right skewed.

## 4.2 BIVARIATE ANALYSIS

Bivariate analysis can be used to examine the relationship between two continuous variables, two categorical variables, or one continuous and one categorical variable. In each case, different statistical techniques can be used to explore the relationship between the two variables using various statistical techniques such as scatter plots, correlation analysis, and regression analysis.

### 4.2.1 CATEGORICAL VS NUMERICAL

Categorical vs continuous bivariate analysis is an analysis consists of two variable which is categorical data & numerical data.

A picture containing bar chart

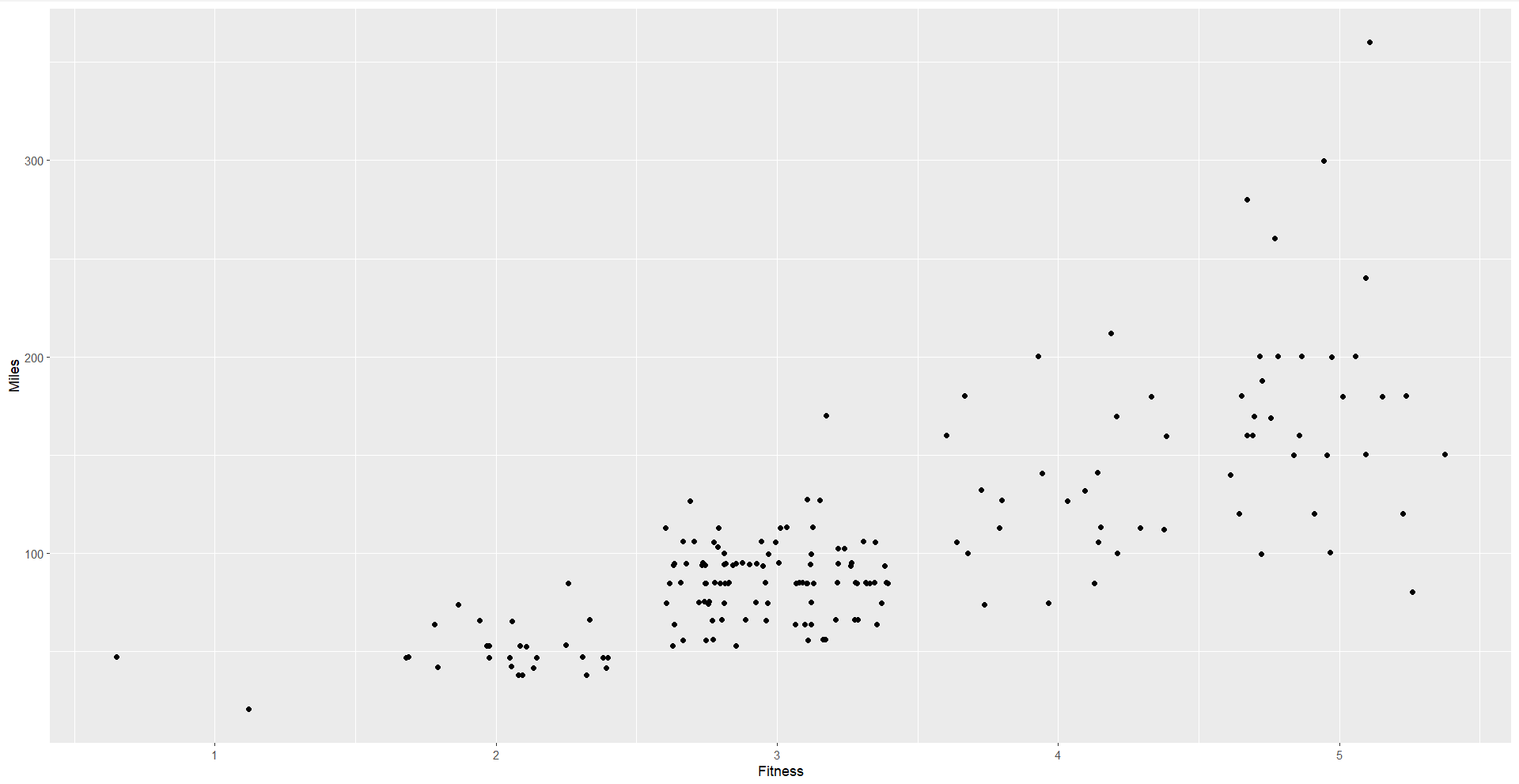
Description automatically generated

*Figure 6 shows a histogram about the variable Product and Usage*

Based on *Figure 6,* shows a bar plot that is used specifically for numerical data and categorical data which on this case is **Usage** and **Product** respectively. The x-axis consists of 3 items which is TM195, TM498 and TM798, representing Product. Y-axis on the other hand represent the total of usage by customer. As observed, the highest total usage would be TM195. Beside that, the lowest total usage would be TM498.

### 4.2.2 NUMERICAL VS NUMERICAL

Numerical vs Numerical bivariate analysis is an analysis consists of two numerical data.



*Figure 7 shows a scatterplot of Fitness and Miles (expected to run)*

Based on *Figure 7,* shows a scatter plot that is used for numerical data and numerical data which on this case is **Fitness** and **Miles (expected to run)** respectively. The x-axis consists of 5 items which is 1, 2, 3, 4 and 5 representing Fitness. Y-axis on the other hand represent Miles (expected to run) consists of 3 items starting from 100, 200, and 300. As observed, the higher the fitness, the farther a customer is expected to run (Miles).

## 4.4 CLEANING TO TREAT QUANTITIES OF ZEROS IN THE DATASET

A picture containing table

Description automatically generated

*Figure 8 show a data status of cardio good fitness dataset.*

In these cases, based on Figure 8 there is no quantities of zeros in the cardio good fitness dataset. Therefore, another dataset will be used to perform the step to treat quantities of zero in dataset.

Dataset that will be used is Sleep Efficiency that has been obtained from “Kaggle”(<https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency> )

1. Know the profile of your dataset.

Table

Description automatically generated

*Figure 9 shows a data structure of dataset “sleep\_data”.*

*Table

Description automatically generated*

*Figure 10 show the ordering data by percentage of zeros*

Based on *Figure 9 and 10,* knowing the structure of the dataset will help identify the quantities of zero in a dataset. By observing the variable of “*q\_zeros”* and

”*p\_zeros”* which means quantity zeros values and percentage of zeros values. As shown in *Figure 9 and 10,* there is a zeros value on the variable *“Alcohol.consumption”,* “*Caffeine.consumption”*, *“Exercise.frequency”*, and *“Awakenings”*,

1. Removing variables with high number of NA/zeros.



*Figure 11 shows an executed command of removing variable with high number of percentages zeros.*

Based on *Figure 11,* executing a command of *removing variable with high number of percentage zeros* will remove “Alcohol.consumption” variable because the percentages of zero values in more then 50%.

1. Keeping all columns except the ones present in 'vars\_to\_remove' vector.

CODING IN R STUDIO

sleep\_data\_2=select(sleep, - one\_of(vars\_to\_remove))

Graphical user interface, text, application

Description automatically generatedFigure 12 show the variable that have percentage zero value more then 50% is completely remove.

Based on Figure 12, after executing the command, variable that have high percentage zero values has been remove and will not be in the new dataset that have been sign to new variable which is “**sleep\_data\_2**”.

# 5.0 DATA VISUALIZATION & PRESENTATION

## 5.1 FIRST VISUALIZATION (Usage vs Gender)

CODING IN R STUDIO

ggplot(cardio, aes(x=Usage, fill=MaritalStatus))+

geom\_histogram(bins=5, colour="#1380A1") +

labs(title="Marital Status Rate by Gender", y="Number of User", subtitle = "Distribution by usage, gender and product")+

theme\_bw() +

facet\_grid(Gender~Product, scales="free")

Chart

Description automatically generated

*Figure 13 shows a data visualization of Number of User by Marital Status*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Product** | **Gender** | **Marital Status** | | **Total** | **Total Overall** |
| **Partnered** | **Single** |
| TM195 | Female | 27 | 13 | 40 | 80 |
| Male | 21 | 19 | 40 |
| TM498 | Female | 15 | 14 | 29 | 60 |
| Male | 21 | 10 | 31 |
| TM798 | Female | 4 | 3 | 7 | 40 |
| *Male* | *19* | *14* | *33* |

*Table 2 shows the count for each product based on gander and marital status.*

As shown in *Figure 13 and Table 2,* Product TM195 has the highest number of users, most of the user are Partnered. Mostly user of TM195 often to use the treadmill only 3 times every week. In other view, product TM798 has the lowest number of users, most of the users are male and they chose to use the treadmill 4 time every week.

## 5.2 SECOND VISUALIZATION (Fitness VS Expected to run (Miles))

CODING IN R STUDIO

ggplot(cardio, aes(Fitness, Miles)) +

geom\_jitter(aes(colour = Product)) +

geom\_smooth(method="lm", se=FALSE) +

labs(title="Fitness VS Miles",

y="Miles",

subtitle = "Distribution by fitness, miles and product") +

theme(plot.title = element\_text(size=20,

face="bold",

family="American Typewriter",

color="tomato",

hjust=0.5,

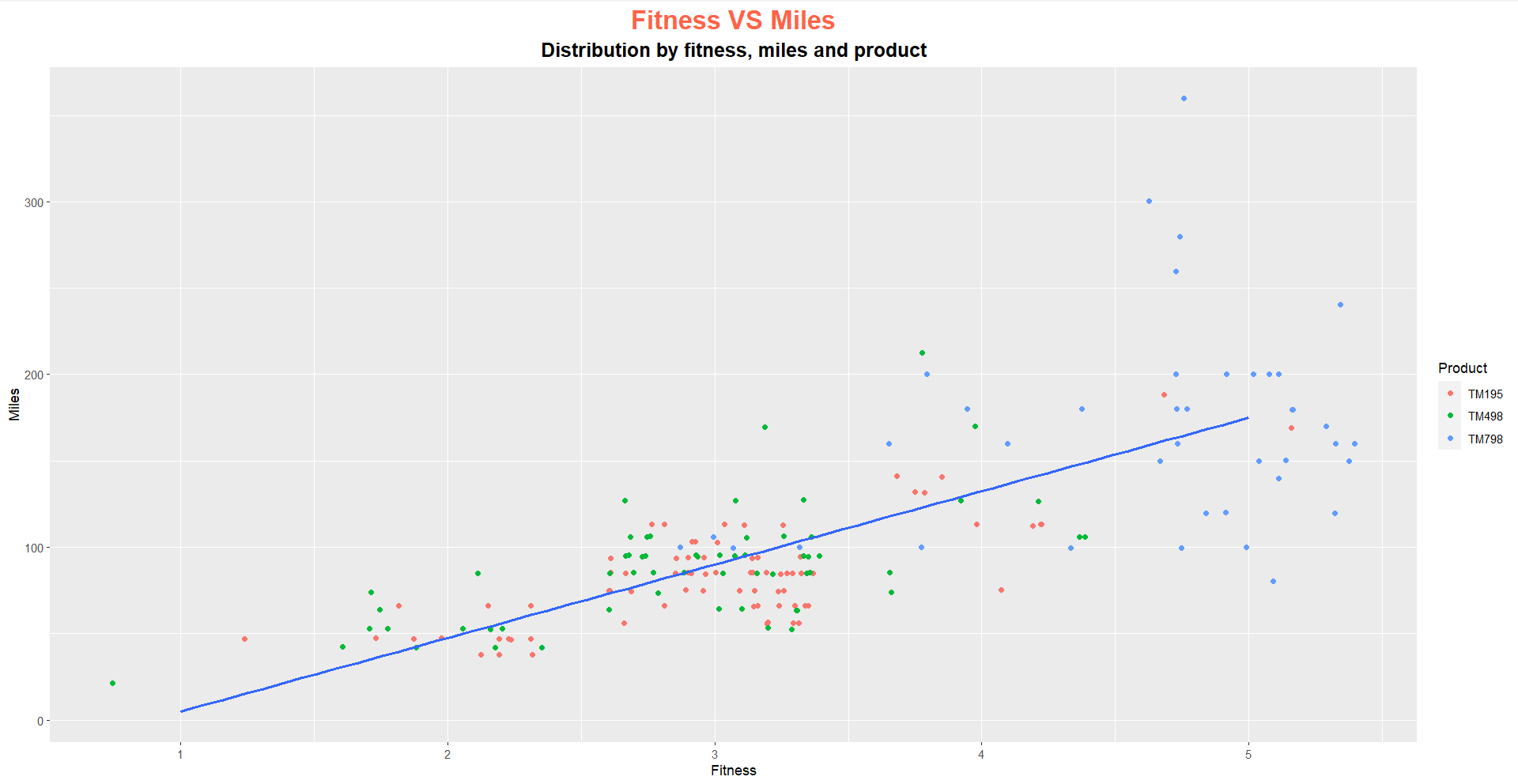
lineheight = 1.2),

plot.subtitle=element\_text(size=15,

family="American Typewriter",

face="bold",

hjust=0.5))



*Figure 14 shows a data visualization of Fitness VS Expected to run (Miles))*

As shown in *Figure 14*, most who has fitness score 5 will be expected to run farther other then any score fitness. Average persons that have fitness score 3 will be expected to run around 50 to 150 miles. Lastly, persons that has fitness score below 2 will be expected to run below 100 miles.

# 6.0 REFERENCE

1. **DATASET**

*Exploratory data analysis/Data visualization*. (2022, November 3). Kaggle. <https://www.kaggle.com/datasets/vsridevi/cardio-good-fitness>

1. **DATASET THAT BEING USED TO TREAT QUANTITIES OF ZEROS IN THE DATASET**

*Sleep Efficiency Dataset*. (2023, February 21). Kaggle. <https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency>

1. **SKEWED DISTRIBUTION**

*Skewed Distribution: Definition, Examples - Statistics How To*. (2022, February 23). Statistics How To. <https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/>

# 7.0 APPENDICES

A screenshot of a computer

Description automatically generated with medium confidence

*Appendix 1 shows the page to retrieve Cardio Good Fitness dataset.*

*A screenshot of a computer

Description automatically generated with medium confidence*

*Appendix 2 shows the page to retrieve Sleep Efficiency dataset.*

Graphical user interface, text, application

Description automatically generated

*Appendix 3 shows the page about Skewed Right information*