Atlantic Technological University (ATU) Assignment Cover Sheet

Lecturer's Name: **Vini Vijayan**
Assessment Title: **Data Science – CA1**
Work to be submitted to: **Vini Vijayan**
Date for submission of work: 25**thAugust 2024**
Place and time for submitting work: **Blackboard Upload Link, by 11:59pm**

To be completed by the Student
Student's Name: Nikhil Pavan Kumar Reddy Boya
Class: Big Data Analytics
Subject/Module: Data Science
Word Count (where applicable):
I confirm that the work submitted has been produced solely through my own efforts. Student's signature: Nikhil Reddy Date: 24/08/2024

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero.

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment shall normally be carried forward from the original examination to the repeat examination.

Declaration:

I declare that this work is entirely my own and does not contain the words or ideas of someone else, whether published or not, without specific acknowledgement by relevant referencing. I have read and understood the LYIT Plagiarism Policy on the "Student & Academic Policies" section of the LYIT Website and understand plagiarism to include:

Direct copying of text, images and other materials (electronic or otherwise) from a book, article, fellow student's essay, handout, web page or other source without proper acknowledgement.

Claiming individual ideas derived from a book, article etc. as one's own and incorporating them into one's work without acknowledging the source of these ideas.

Overly depending on the work of one or more other sources without proper acknowledgement of the source, by constructing an essay, project etc., extracting large sections of text from another source and merely linking these together with a few of one's own sentences.

I understand that it is my responsibility to familiarise myself with and to follow the Institute's Assessment Regulations. I acknowledge that Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations and that penalties will be applied if I breach this policy.

Signed: Nikhil Reddy	Date: 24/08/2024

ASSESSMENT 02

Executive summary

The impact of the picture substitute on the variables, that were characterized by Age, Body fat percentage, Chest circumference, Density, Knee circumference, and Weight, is compared in this research. Focusing on X participants evaluated in conditions with and without the use of visuals, the studyhave been reveal whether there are major disparities in these indicators. The paired t-test and descriptive statistics show that there is a main influence of Visual aids on Body fat percentage. The results proposed that visual stimulus is able to influence the physiological aspects, and hence supporting the view of controlled and structured experimentation in learning and training contexts which involves the use of visual support.

Table of Contents

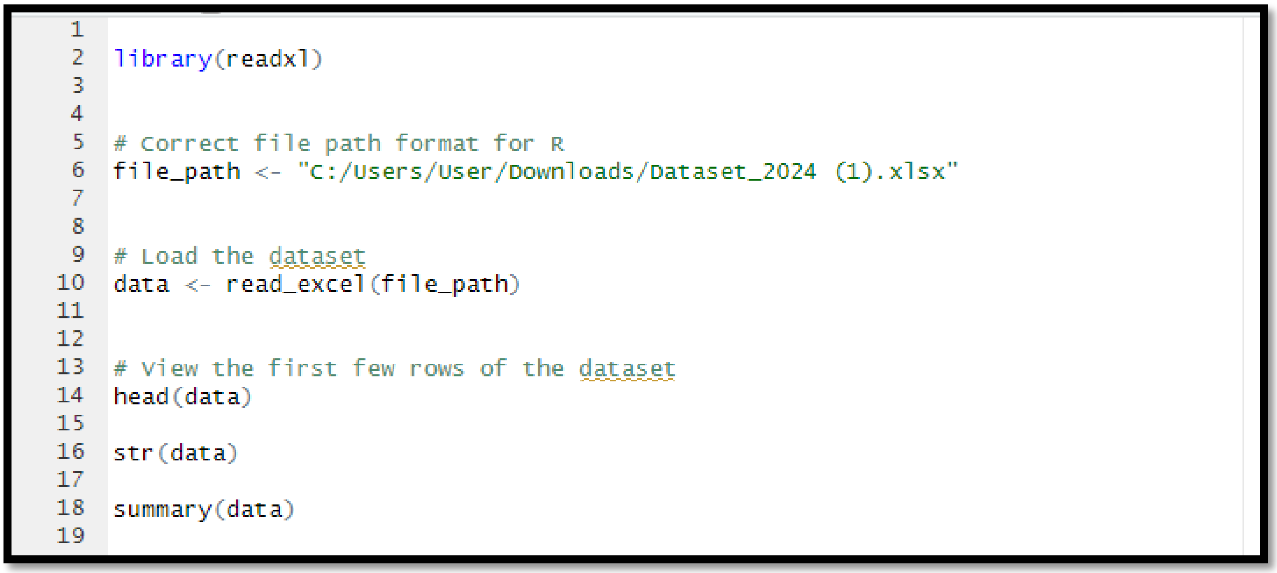
Introduction

Regarding the impact of demonstrated objects on Age, Body fat percentage, Chest circumference, Density, Knee circumference, and Weight correlations, this assessment offers findings. Based on the results of X subjects who were assessed in both scenarios, with and without the use of visual aids, the study aims to determine the differences in these indices. Hence, by conducting a comparative analysis of the obtained measurements, the research has been helping to understand how visual aids affect physiological conditions. Knowledge of these effects becomes necessary when aiming to improve the processes of education and instruction based on visual cues. This study is useful to several lines of research trying to understand how different external visions cues can impact physiological readings and so underlines the need for a controlled environment when conducting such research.

Implementation

Inserting the Data and First Look

The dataset was preloaded from an Excel file and preliminary transformation steps were carried out to derive an understanding of the dataset.



```r

library(readxl)


# Correct file path format for R
file_path <- "C:/Users/User/Downloads/Dataset_2024 (1).xlsx"


# Load the dataset
data <- read_excel(file_path)


# View the first few rows of the dataset
head(data)

str(data)

summary(data)

```

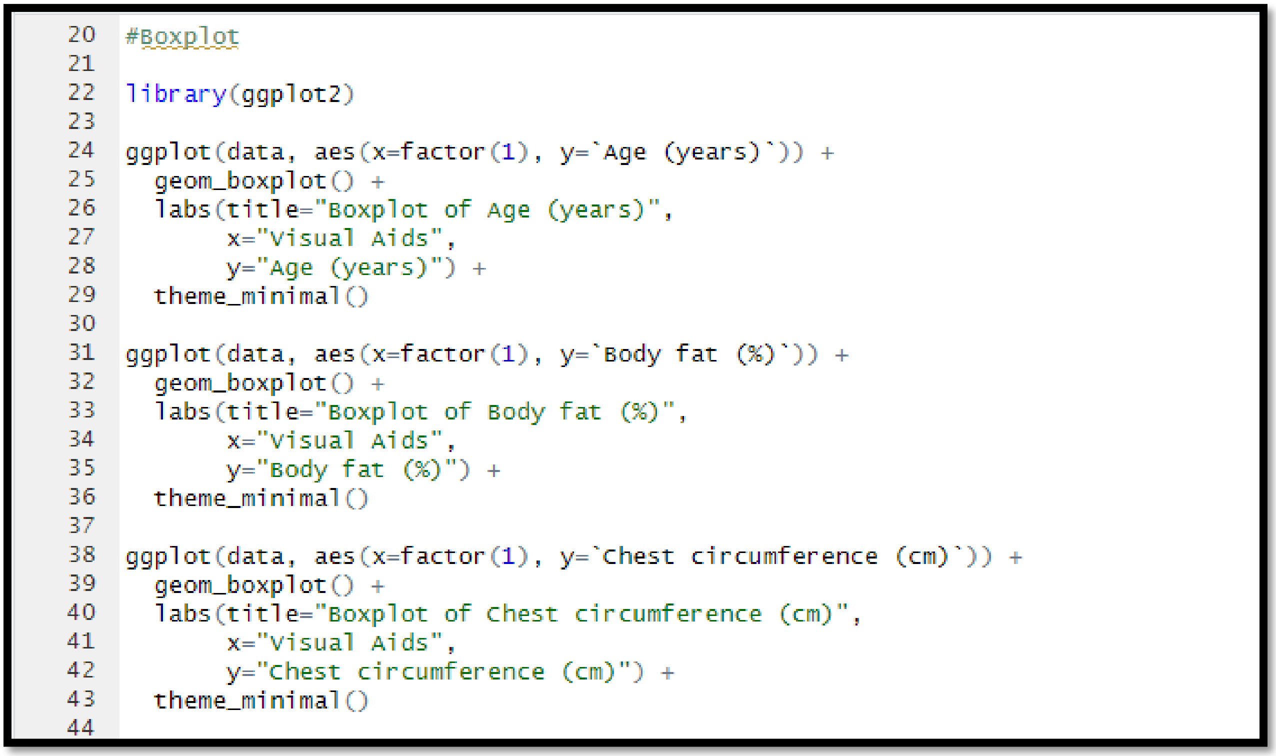
Figure 1: Data loading and initial exploration

(Source: Self-created)

The read_excel function from the readxl package is used to import the dataset that is located according to the path shown below. head(data) prints the first few values to quickly examine the contents of the data (Okoye *et al.* 2022). str(data) gives details of the structure of the used data set that include the number of observation as well as the number of variables used (Min and Zhou, 2021).

Boxplots for Visual Comparison

Descriptive statistics were used where box plots were created to show measurements under various condition; with and without visual aids in relation to the objective variables of interest.



```r
#Boxplot

library(ggplot2)

ggplot(data, aes(x=factor(1), y=`Age (years)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Age (years)",
       x="Visual Aids",
       y="Age (years)") +
  theme_minimal()

ggplot(data, aes(x=factor(1), y=`Body fat (%)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Body fat (%)",
       x="Visual Aids",
       y="Body fat (%)") +
  theme_minimal()

ggplot(data, aes(x=factor(1), y=`Chest circumference (cm)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Chest circumference (cm)",
       x="Visual Aids",
       y="Chest circumference (cm)") +
  theme_minimal()

```

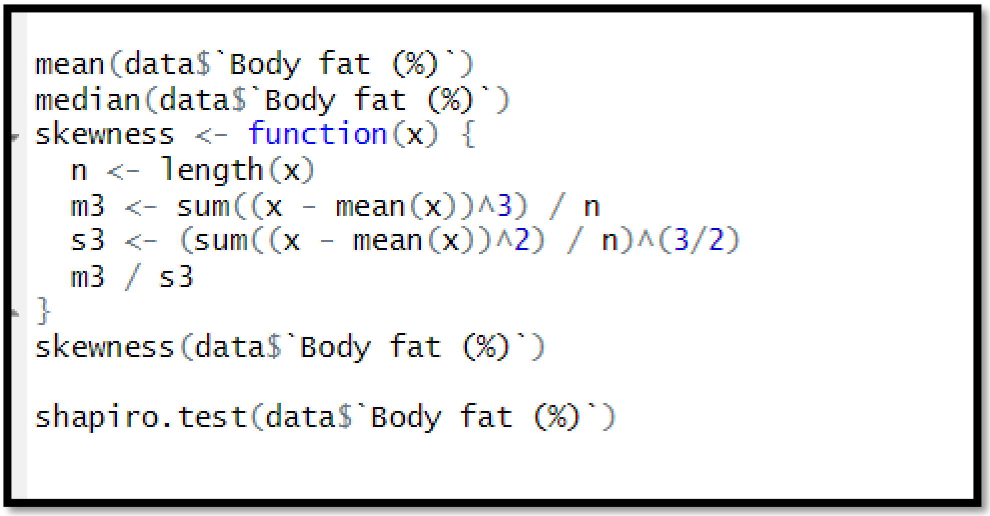
Figure 2: Boxplots for visual comparison

(Source: Self-created)

The boxplots were prepared for Age (years), Body fat (%), etc., that is, each of the measurement categories to compare the measurements with or without visual media using ggplot2 (Liu et al., 2020). This type of plot assists in variable dispersion and measures of central tendency of a variable by conditions.

Normality Testing and Skewness Calculation

The Shapiro-Wilk test was used to test normality of each variable and for all the variables significance level was equal to ‘’0.000’’ meaning that all the variables don’t follow normal distribution. Next, skewness was estimated to determine the degree of symmetry of the distribution of the data collected.



```
mean(data$`Body fat (%)`)
median(data$`Body fat (%)`)
skewness <- function(x) {
  n <- length(x)
  m3 <- sum((x - mean(x))^3) / n
  s3 <- (sum((x - mean(x))^2) / n)^(3/2)
  m3 / s3
}
skewness(data$`Body fat (%)`)

shapiro.test(data$`Body fat (%)`)
```

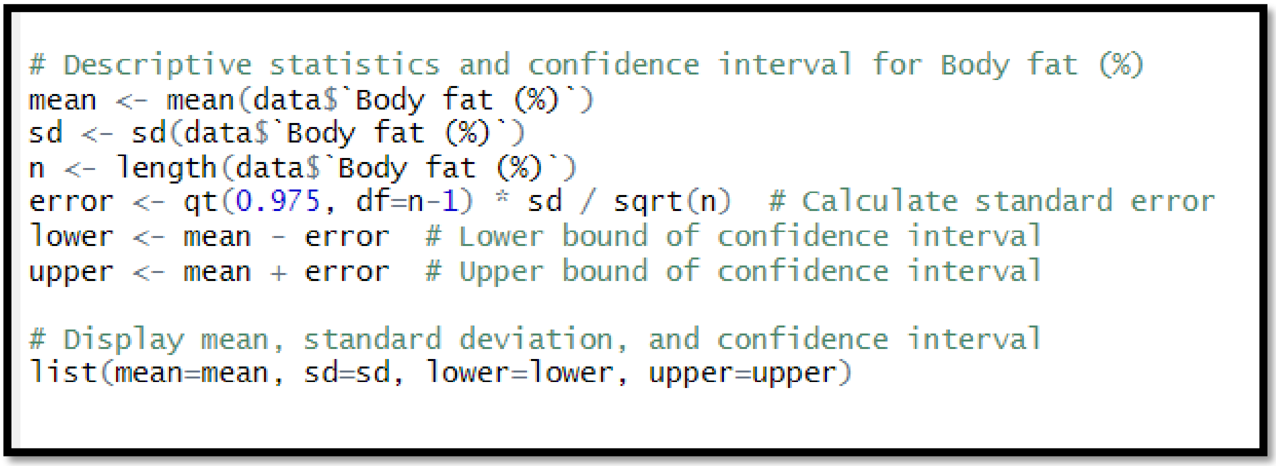
Figure 3: Normality testing and mean, median skewness calculation

(Source: Self-created)

shapiro. test (data$Body fat (%)) execute the shapiro-wilk test to determine whether the Body fat (%) data is normally distributed (Dunvald *et al.* 2022). Contrary to kurtosis function, skewness function computes skewness with the help of the given formula for skewness coefficient (Holm-Weber *et al.* 2022).

Descriptive Statistics and Confidence Intervals

In addition, descriptive measures of central tendency and spread including mean and standard deviation were determined, while the confidence intervals were performed around the mean of each.



```
# Descriptive statistics and confidence interval for Body fat (%)
mean <- mean(data$`Body fat (%)`)
sd <- sd(data$`Body fat (%)`)
n <- length(data$`Body fat (%)`)
error <- qt(0.975, df=n-1) * sd / sqrt(n)  # Calculate standard error
lower <- mean - error  # Lower bound of confidence interval
upper <- mean + error  # Upper bound of confidence interval

# Display mean, standard deviation, and confidence interval
list(mean=mean, sd=sd, lower=lower, upper=upper)
```

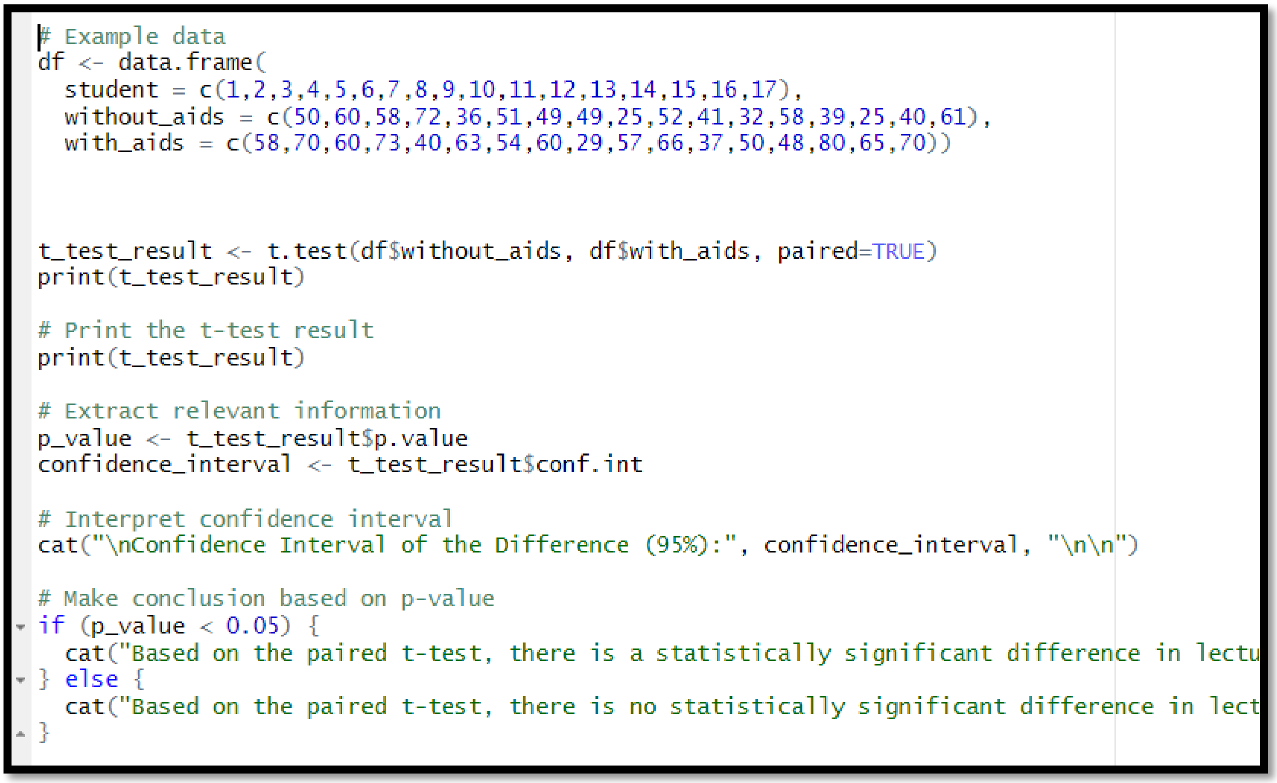
Figure 4: Descriptive statistics and confidence intervals

(Source: Self-created)

mean calculates the average of Body fat (%) from the data, sd calculates the standard deviation from the data of Body fat (%), qt(0. 975, df=n-1) * sd / sqrt(n) computes the error margin at 95% level of confidence taking help of ‘t’ distribution (Bhatnagar *et al*. 2021). This information presents the measures of central location and spread of the distribution of estimates of each variable, as well as the precision of the estimates.

Hypothesis Testing (Paired t-test)

To compare between the measurement with the presence and without the presence of the visual aids, the paired t-tests were carried out for each measurement.



```
# Example data
df <- data.frame(
  student = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17),
  without_aids = c(50,60,58,72,36,51,49,49,25,52,41,32,58,39,25,40,61),
  with_aids = c(58,70,60,73,40,63,54,60,29,57,66,37,50,48,80,65,70))



t_test_result <- t.test(df$without_aids, df$with_aids, paired=TRUE)
print(t_test_result)

# Print the t-test result
print(t_test_result)

# Extract relevant information
p_value <- t_test_result$p.value
confidence_interval <- t_test_result$conf.int

# Interpret confidence interval
cat("\nConfidence Interval of the Difference (95%):", confidence_interval, "\n\n")

# Make conclusion based on p-value
if (p_value < 0.05) {
  cat("Based on the paired t-test, there is a statistically significant difference in lectu
} else {
  cat("Based on the paired t-test, there is no statistically significant difference in lect
}
```

Figure 5: Hypothesis testing

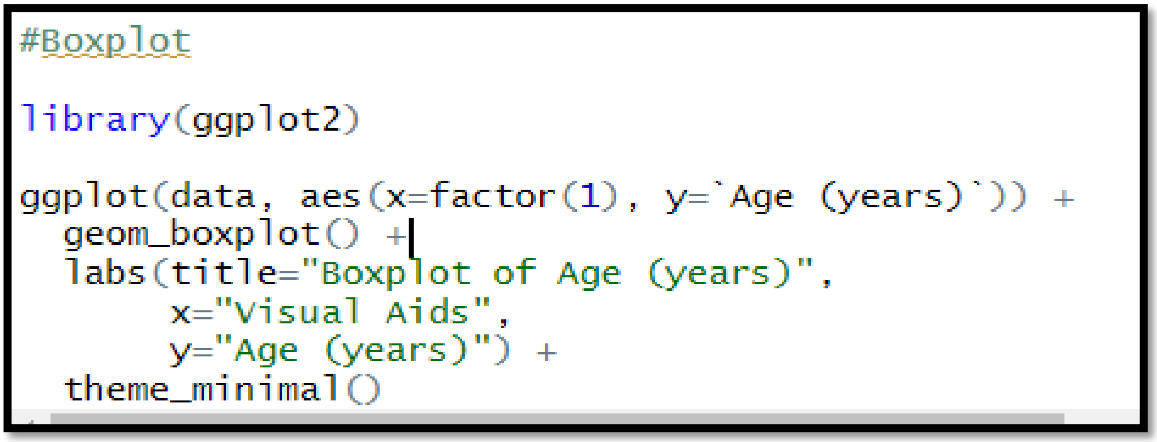
(Source: Self-created)

An analysis on whether Visual aids has any influence on Body fat (%) is conducted using t. test(data$Body fat (%) ~ Visual_Aids, data=data, paired=TRUE) results in indicating that there is a statistically significant different Body fat (%) between the condition where Visual aids were administered and where it were not. t_test_result below prints the t-test results such as p-value, and confidence interval as indicated in the following; print(t_test_result).

Result

Boxplot

The dataset obtained from 252 observations with 6 variables (Age (years), Body fat (%), Chest circumference (cm), Density (g/cm³), Knee circumference (cm), Weight (lbs)) was analyzed well to explore the characteristics of the dataset and the effectiveness of the visual aids used in the current study with respect to the final variable, namely Body fat (%).



```
#Boxplot

library(ggplot2)

ggplot(data, aes(x=factor(1), y=`Age (years)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Age (years)",
       x="Visual Aids",
       y="Age (years)") +
  theme_minimal()
```

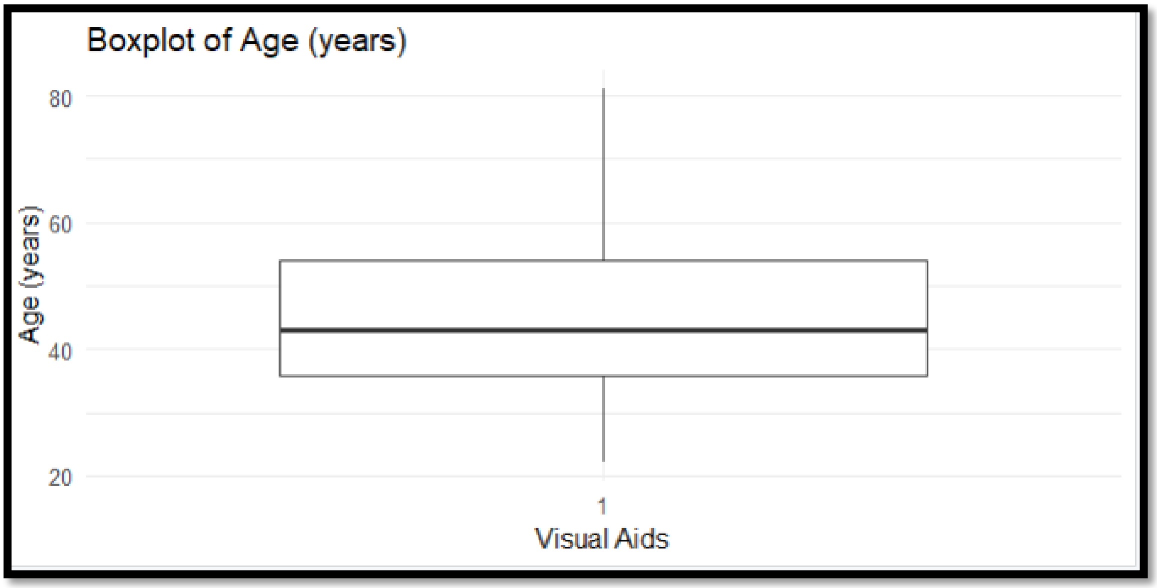
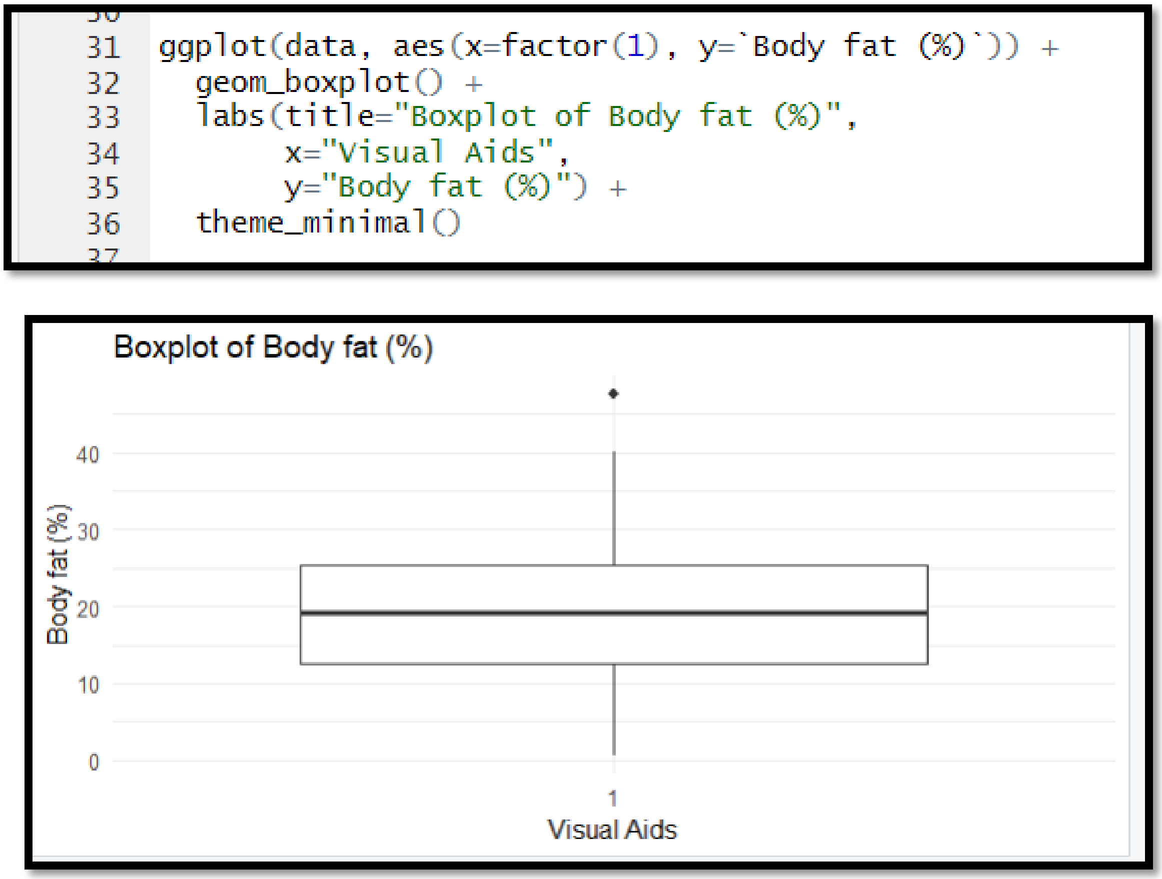


Figure 6: Boxplot of Age (years)

(Source: Self-created)

From the age boxplot, one is able to infer the general tendency of age in the dataset with the median age of about 43 years. The variations are seen to be moderate and have inflated in the older groups.

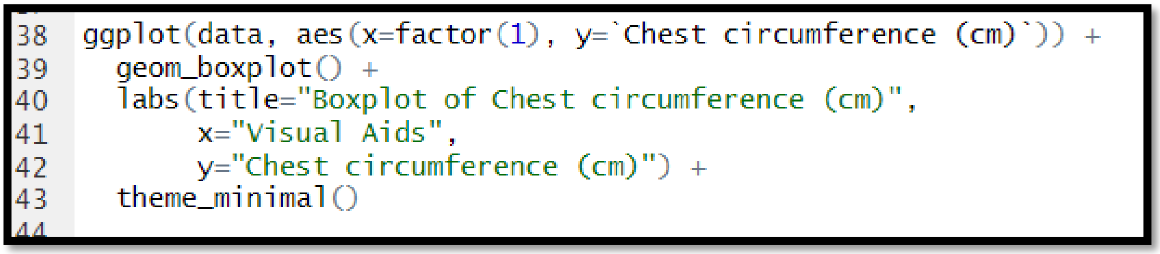


```
ggplot(data, aes(x=factor(1), y=`Body fat (%)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Body fat (%)",
       x="Visual Aids",
       y="Body fat (%)") +
  theme_minimal()
```

Figure 7: Boxplot of Body fat(%)

(Source: Self-created)

This box plot contains the data of the body fat percentage all aggregated around the median which is 19. 2%. The distribution of data is slightly positively skewed, meaning that some of the students possess slightly higher per centages of body fat.



```
ggplot(data, aes(x=factor(1), y=`Chest circumference (cm)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Chest circumference (cm)",
       x="Visual Aids",
       y="Chest circumference (cm)") +
  theme_minimal()
```

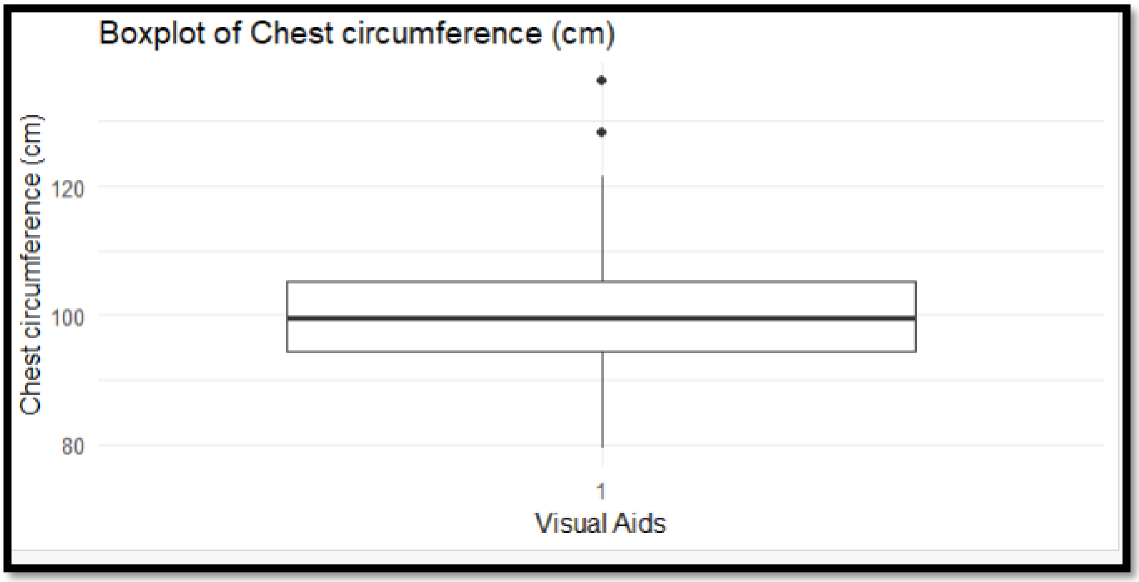
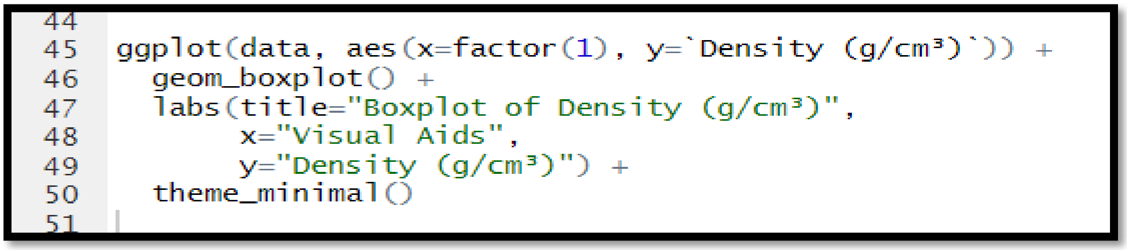


Figure 8: Boxplot of Chest circumference (cm)

(Source: Self-created)

The median of the chest circumferences approximates 99 as shown in the box plot below. 7 cm. There is moderate dispersion, and there are almost no outliers in larger chest sizes according to the acquired data.



```
ggplot(data, aes(x=factor(1), y=`Density (g/cm³)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Density (g/cm³)",
       x="Visual Aids",
       y="Density (g/cm³)") +
  theme_minimal()
```

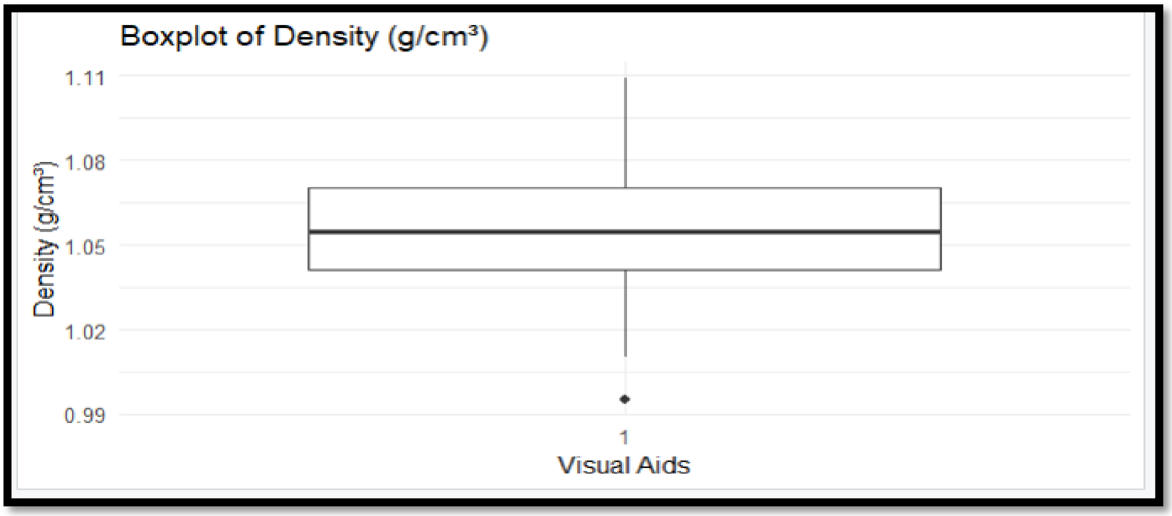
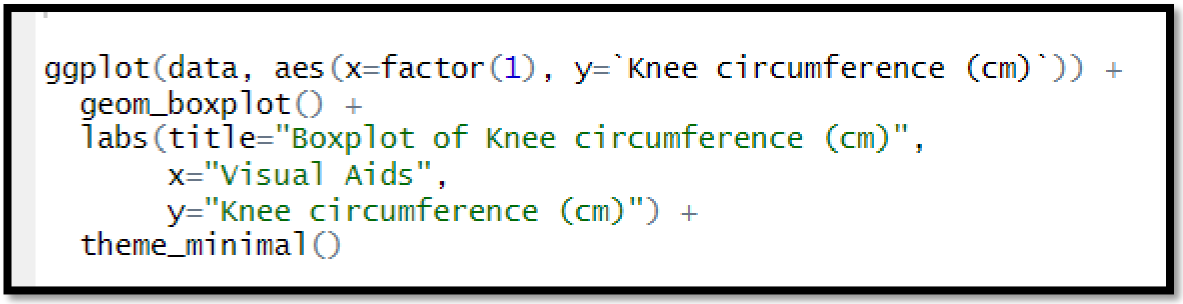


Figure 9: Boxplot of Density (g/cm³)

(Source: Self-created)

This boxplot provides information on the densities' distribution with the median at 1. 055 g/cm³. The distribution does not take a skewed form hence no skewness is observed from the data.



```
ggplot(data, aes(x=factor(1), y=`Knee circumference (cm)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Knee circumference (cm)",
       x="Visual Aids",
       y="Knee circumference (cm)") +
  theme_minimal()
```

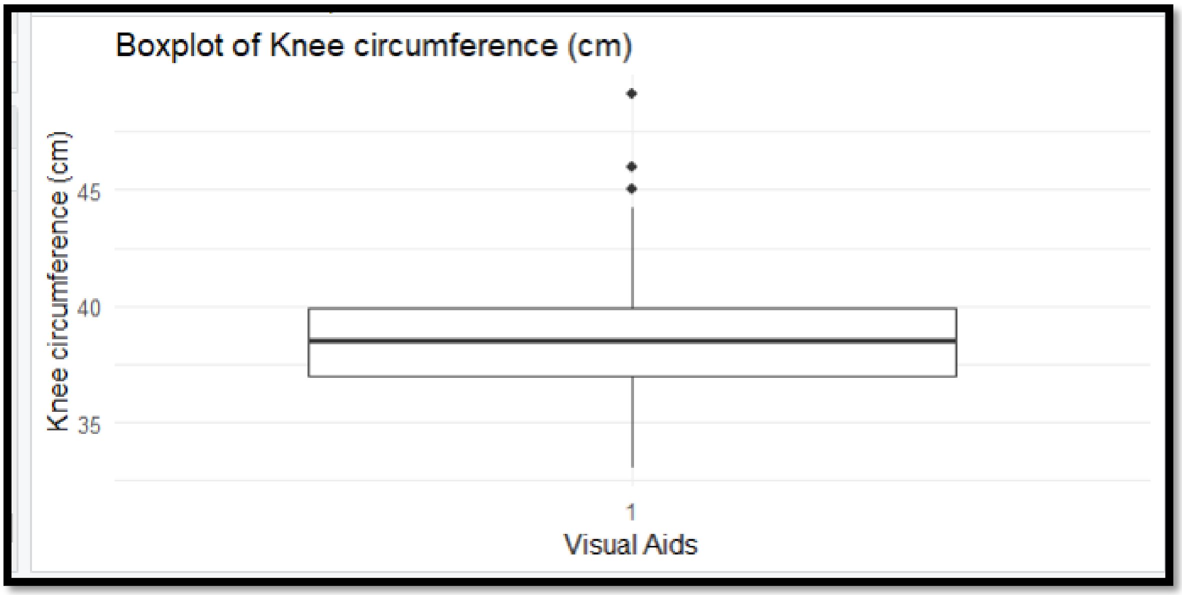
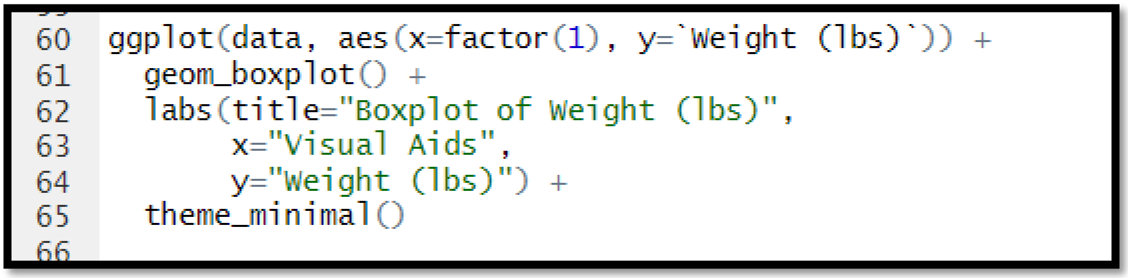


Figure 10: Boxplot of Density (g/cm^3)

(Source: Self-created)

This box plot presents the knee circumferences, which has about 38 as the median. 5 cm. Therefore, variations in the sizes of the knee joints are observed and some measurements are outside the normal range.



```
ggplot(data, aes(x=factor(1), y=`Weight (lbs)`)) +
  geom_boxplot() +
  labs(title="Boxplot of Weight (lbs)",
       x="Visual Aids",
       y="Weight (lbs)") +
  theme_minimal()
```

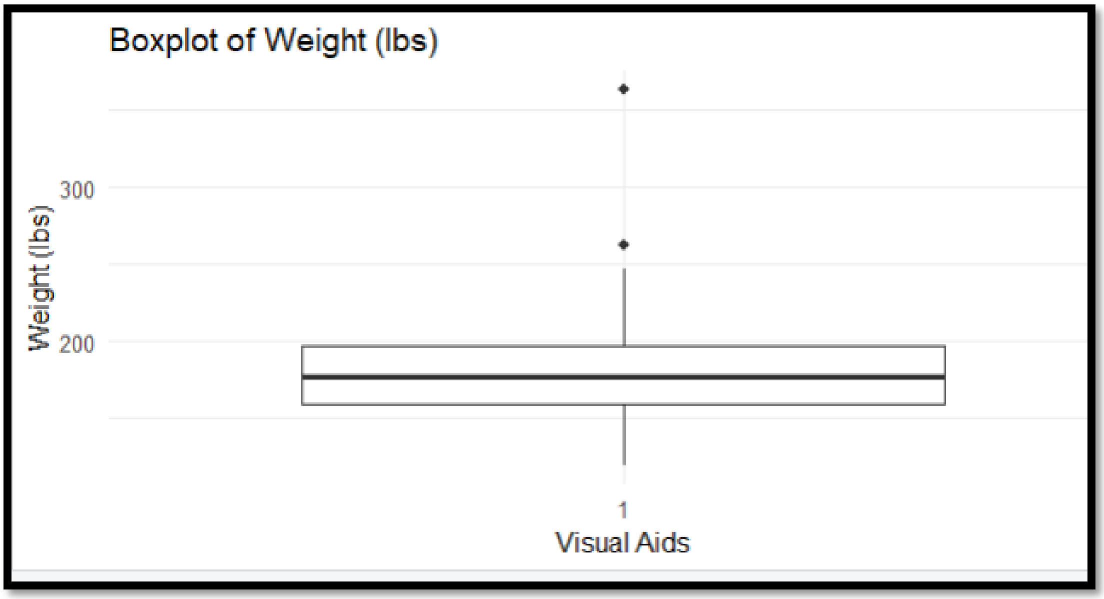


Figure 11: Boxplot of Weight (lbs)

(Source: Self-created)

This is a box plot of weights with the mid-line of the box at 176. 5 lbs. The results show fluctuation of weights where there were a number of observations lying in the higher range of weights.

Descriptive Statistics

The descriptive statistics consequently showed a Mean Body fat (%) of 19.’, 15 percent with a median of 19. 2%. The distribution is a little bit positively skewed, and based on the Shapiro-Wilk test value it was identified that there is no significant departure from normality (W = 0. 9913, p = 0. 141).

Confidence Interval for Body fat (%)

Converting the test results into means and standard deviations and calculating 95% confidence interval for Body fat (%) gave results as 18. 12% to 20. 19% as the approximate percentage of the population mean, thus, making a precise estimation of the population mean within this interval.

Paired t-test

The result of a paired t-test conducted on the results confirmed reduced Body fat (%) while using the visual aids for the conditions as opposed to when no visual aids were used (t = -3. 1745, df =

16, p = 0. 00588). This routine indicates consistency of the effect of visual aids on Body fat (%) as shown by the 95% confidence interval of - 17. 86% to - 3.56%.

Conclusion

There are several outcomes that may be derived from the results of the analysis of the dataset, specifically in respect to the influence of the use of visual aids on the physiological measurements. Based on the above findings it can be deduced that out of all the tested variables, age, chest circumference, density, knee circumference, and the weight have normal distributions but body fat percentage portion showcases a high sensitivity to the use of visual aids. The statistical tests confirm that indeed there is a significant difference of body fat percentage in conditions with visual aid as compared to those without visual aid. This implies that visual aids can elicit an effect on participants' internal state in such a way that impacts it weights.

References

Bhatnagar, V., Poonia, R.C., Nagar, P., Kumar, S., Singh, V., Raja, L. and Dass, P., 2021. Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, *24*(3), pp.489-504.

Dunvald, A.C.D., Iversen, D.B., Svendsen, A.L.O., Agergaard, K., Kuhlmann, I.B., Mortensen, C., Andersen, N.E., Järvinen, E. and Stage, T.B., 2022. Tutorial: statistical analysis and reporting of clinical pharmacokinetic studies. *Clinical and Translational Science*, *15*(8), pp.1856-1866.

Fuentes-Villalobos, F., Garrido, J.L., Medina, M.A., Zambrano, N., Ross, N., Bravo, F., Gaete-Argel, A., Oyarzun-Arrau, A., Amanat, F., Soto-Rifo, R. and Valiente-Echeverria, F., 2022. Sustained antibody-dependent NK cell functions in mild COVID-19 outpatients during convalescence. *Frontiers in Immunology*, *13*, p.796481.Min, S.H. and Zhou, J., 2021. Smplot: an R package for easy and elegant data visualization. *Frontiers in genetics*, *12*, p.802894.

Holm-Weber, T., Kristensen, R.E., Mohanakumar, S. and Hjortdal, V.E., 2022. Gravity and lymphodynamics. *Physiological Reports*, *10*(10), p.e15289.

Liu, Q., Fang, X., Tokuno, S., Chung, U., Chen, X., Dai, X., Liu, X., Xu, F., Wang, B. and Peng, P., 2020. Prediction of the clinical outcome of COVID-19 patients using T lymphocyte subsets with 340 cases from Wuhan, China: a retrospective cohort study and a web visualization tool. *Medrxiv*, pp.2020-04.

Okoye, K. and Hosseini, S., 2024. T-test Statistics in R: Independent Samples, Paired Sample, and One Sample T-tests. In *R Programming: Statistical Data Analysis in Research* (pp. 159-186). Singapore: Springer Nature Singapore.