Analysing health parameters on a young adult

David Ortego Casado

Table of contents

## 1 Objective

The goal of this project is to do a comparison between some health parameters (systolic pressure, diastolic pressure and weight) of one young adult that exercises four times a week versus the average parameters of a healthy young adult to practise how to program in R as well as to practise some statistical concepts.

## 2 Material and methods

This project was the result of considering how I could improve my R programming skils as well as practising some basic statistics, the solution was to practise with a dataset, so I created my own dataset by measuring some health parameters that could easily be measured at home, I took a total of 16 observations of my blood pressure , my heart rate (although this variable was not used in the analysis) and my weight.

After having all the measures, the data was processed and later analysed. After the analysis was done , a report was made with Quarto and the results of the project were shared in Github.

The reference values of blood pressure were obtained from the Center for Disease Control and Prevention (1) and the reference weight was calculated using Calculators.org. The analysis was done in R and RStudio.

## 3 Results

#Loading our dataset  
library(readxl)  
library(tidyverse)  
#Selecting the Excel file that contains our data-----------------------------------  
  
file.choose()

[1] "D:\\OneDrive - UAM\\Escritorio\\Data analysis\\R documents\\Proyecto analisis estadistico mis parametros de salud\\Health\_analysis\\Quarto\_refort\_definitive\\Quarto report.qmd"

# We will call raw\_data the file location that we got from file.choose(), we add everything we see in the console after using the function, this path will be different than the path I use.   
  
raw\_data <- "D:\\OneDrive - UAM\\Escritorio\\Data analysis\\R documents\\Proyecto analisis estadistico mis parametros de salud\\Health\_analysis\\Quarto\_refort\_definitive\\Health\_parameters.xlsx"  
  
first\_data <- read\_excel(raw\_data)  
first\_data

# A tibble: 16 × 5  
 time sp dp pulse weight  
 <dttm> <dbl> <dbl> <dbl> <dbl>  
 1 1899-12-31 09:30:00 104 52 52 72.6  
 2 1899-12-31 10:30:00 109 52 58 72.6  
 3 1899-12-31 11:30:00 100 52 50 72.9  
 4 1899-12-31 12:30:00 98 54 47 72.7  
 5 1899-12-31 13:30:00 103 57 44 72.2  
 6 1899-12-31 14:30:00 102 60 45 72.1  
 7 1899-12-31 15:30:00 112 54 44 72.8  
 8 1899-12-31 16:30:00 111 53 47 72.9  
 9 1899-12-31 17:30:00 120 57 49 73.1  
10 1899-12-31 18:30:00 109 55 46 72.6  
11 1899-12-31 19:30:00 111 58 46 73.1  
12 1899-12-31 20:30:00 116 60 49 72.7  
13 1899-12-31 21:30:00 112 63 44 72.4  
14 1899-12-31 22:30:00 114 59 49 73.5  
15 1899-12-31 23:30:00 111 55 49 73.2  
16 1900-01-01 00:30:00 121 66 51 73.2

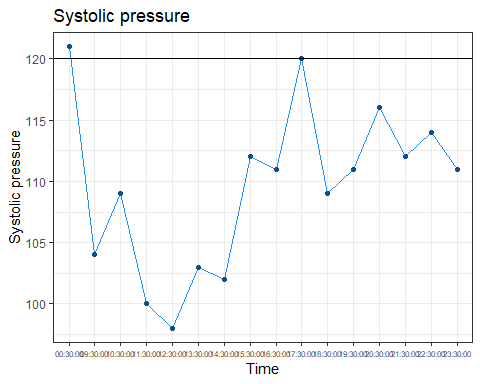
# Removing a date from the variable momento\_del\_dia--------------------------------  
  
second\_data <- separate(first\_data, time, into =c("day", "time"),sep=" ")  
  
second\_data

# A tibble: 16 × 6  
 day time sp dp pulse weight  
 <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
 1 1899-12-31 09:30:00 104 52 52 72.6  
 2 1899-12-31 10:30:00 109 52 58 72.6  
 3 1899-12-31 11:30:00 100 52 50 72.9  
 4 1899-12-31 12:30:00 98 54 47 72.7  
 5 1899-12-31 13:30:00 103 57 44 72.2  
 6 1899-12-31 14:30:00 102 60 45 72.1  
 7 1899-12-31 15:30:00 112 54 44 72.8  
 8 1899-12-31 16:30:00 111 53 47 72.9  
 9 1899-12-31 17:30:00 120 57 49 73.1  
10 1899-12-31 18:30:00 109 55 46 72.6  
11 1899-12-31 19:30:00 111 58 46 73.1  
12 1899-12-31 20:30:00 116 60 49 72.7  
13 1899-12-31 21:30:00 112 63 44 72.4  
14 1899-12-31 22:30:00 114 59 49 73.5  
15 1899-12-31 23:30:00 111 55 49 73.2  
16 1900-01-01 00:30:00 121 66 51 73.2

third\_data <- second\_data %>%   
 select(time, sp, dp, pulse, weight)  
third\_data

# A tibble: 16 × 5  
 time sp dp pulse weight  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 09:30:00 104 52 52 72.6  
 2 10:30:00 109 52 58 72.6  
 3 11:30:00 100 52 50 72.9  
 4 12:30:00 98 54 47 72.7  
 5 13:30:00 103 57 44 72.2  
 6 14:30:00 102 60 45 72.1  
 7 15:30:00 112 54 44 72.8  
 8 16:30:00 111 53 47 72.9  
 9 17:30:00 120 57 49 73.1  
10 18:30:00 109 55 46 72.6  
11 19:30:00 111 58 46 73.1  
12 20:30:00 116 60 49 72.7  
13 21:30:00 112 63 44 72.4  
14 22:30:00 114 59 49 73.5  
15 23:30:00 111 55 49 73.2  
16 00:30:00 121 66 51 73.2

# Plotting Systolic pressure-------------------------------------------------------   
ps\_graph <- third\_data %>%   
 ggplot(aes(x= time,  
 y=sp))  
  
ps\_graph+  
 geom\_line(group= "sp",  
 color= "dodgerblue")+  
 geom\_point(color="dodgerblue4")+  
 geom\_hline(yintercept = 120)+  
 labs(title = "Systolic pressure",  
 x= "Time",  
 y= "Systolic pressure")+  
 theme\_bw()+  
 theme(axis.text.x= element\_text( size= "6"))



# Statistical inference of systolic pressure (ps) ---------------------------------  
  
  
mean\_sp <- mean(third\_data$ps)  
mean\_sp

[1] NA

standard\_deviation\_sp <- sd(third\_data$sp)  
standard\_deviation\_sp

[1] 6.712861

# The population value for sistolic pressure for a young adult is 120  
  
population\_sp\_mean <- 120  
  
# Since we have a small sample size, only 16 observations and we dont know the population standard deviation we are on a t-student distribution.  
  
# First hypothesis-----------------------------------------------------------------  
# H0; mean\_sp = population\_ps\_mean  
# H1; mean\_sp is different than population\_ps\_mean  
  
# 1. Calculating 95% Confidence interval (CI) -------------------------------------  
#   
  
# n= 16  
  
# mean\_sp= 109.5625  
  
# standard\_deviation\_sp= 6.712861  
  
# degrees of freedom = n-1 = 16-1=15  
  
#since its a 95%CI we do:  
  
alpha <- 0.05  
half\_alpha <- alpha/2  
percentile <- 1-half\_alpha  
  
# Now lets look on the t-student distribution table for the value of n-1=16-1=15 degrees of freedom where 97.5% of the values are inferior to it:   
  
t\_student\_sp <- 2.131  
  
# The CI is:   
  
lower\_CI <- mean\_sp-t\_student\_sp\*(standard\_deviation\_sp/sqrt(16))  
lower\_CI

[1] NA

upper\_CI <- mean\_sp+ t\_student\_sp\*(standard\_deviation\_sp/sqrt(16))  
upper\_CI

[1] NA

CI\_ps <- c(lower\_CI,upper\_CI)  
CI\_ps

[1] NA NA

# Since the CI does not inclue the H0 value, we can reject H0 and accept H1, the systolic pressure of this young adult that works out 4 times a week is #different than the reference value of a healthy young adult.  
  
# 2. Method of the rejection area -------------------------------------------------  
# we have to calculate the pivot value:   
  
pivot <- (mean\_sp-population\_sp\_mean)/(standard\_deviation\_sp/sqrt(16))  
pivot

[1] NA

# Delimiting rejection zone:   
  
#Since we did a bilateral significance test, we will establish two regions , one for values much higher than 120 and one for values much lower than 120.  
  
#This two regions must cover 5% of the values which is the significance we are considering, the error in case we are wrong when rejecting H0:  
  
# Since its bilateral we will have halph\_alpha on each side of the graph, so we will reject values that are below the quartile region of 2.5 and above 97.5, we calculate the 97.5 percentile looking at the normal table and we get a value of 1.96 on the right and - 1.96 on the left. So values that are above 1.96 or below -1.96 will fall in the rejection region, allowing us to reject H0 and accept H1.   
  
# Our pivot value is : -6.21 which falls within the rejection region, therefore we reject H0 and we accept H1.   
  
# 3. Calculating the p-value for this test-----------------------------------------  
  
# We are considering both sides of the distribution since H1 is its different than 120, which could mean either its smaller or higher than 120:  
  
# First we have to calculate the pivot value:  
  
pivot\_sp <- (mean\_sp-population\_sp\_mean)/(standard\_deviation\_sp/sqrt(16))   
pivot\_sp

[1] NA

# Asumming Ho , the probability of finding lower values of our pivot is:   
  
probability\_pivot <- 1-0.9989  
probability\_pivot

[1] 0.0011

# Since we are considering both sides:  
  
p\_value\_sp <- probability\_pivot\*2  
p\_value\_sp

[1] 0.0022

#Second hypothesis-----------------------------------------------------------------  
  
# Now we know that the value is different than the reference value, we want to see if its smaller or higher, looking at the graph the data suggests that it is smaller, so we will do the next hypothesis:  
  
# H0: mean\_sp = ps\_mean\_pop; 109.5625= 120  
# H1: mean\_sp < ps\_mean\_pop; 109.5625 < 120  
  
# Here we are in a one tail hypothesis test, before we were looking at values above and below the mean, if H0 is false, it meant it was either higher or lower than the population mean.   
  
# Now if H1 is true, it means that ps\_mean\_david is smaller than the population mean.  
  
# 1. Calculating 95% Confidence interval (CI) -------------------------------------   
# Looking at the CI calculated before, since it doesnt include H0 , we can already reject H0 and say that the systolic pressure of this young adult that exercises 4 times a week is lower than the reference systolic pressure.  
  
  
# 2. Method of the rejection area -------------------------------------------------   
# Now to see the rejection area, since we are in a one tail, the rejection area covers all values below percentile 5  
  
# P5=P95 looking at the normal table, p95= 1.64 therefore p5=-1.64  
  
# Since our pivot = -6.21, it falls within the rejection area , therefore we reject H0 and we accept H1.  
  
# With the p-value method----------------------------------------------------------  
  
# Here we dont multiply it by 2 because we are only looking at one side of the distribution:   
  
probability\_pivot <- 1-0.9989  
probability\_pivot

[1] 0.0011

# Now we will do the same for the diastolic pressure and the weigh-----------------  
  
#Diastolic pressure comparison-----------------------------------------------------   
#This time , to clean the data and remove the date column that was imported from Excel and that is wrong and not needed, we will use a different function than the one used when calculating the systolic pressure, the result is the same:  
  
first\_data

# A tibble: 16 × 5  
 time sp dp pulse weight  
 <dttm> <dbl> <dbl> <dbl> <dbl>  
 1 1899-12-31 09:30:00 104 52 52 72.6  
 2 1899-12-31 10:30:00 109 52 58 72.6  
 3 1899-12-31 11:30:00 100 52 50 72.9  
 4 1899-12-31 12:30:00 98 54 47 72.7  
 5 1899-12-31 13:30:00 103 57 44 72.2  
 6 1899-12-31 14:30:00 102 60 45 72.1  
 7 1899-12-31 15:30:00 112 54 44 72.8  
 8 1899-12-31 16:30:00 111 53 47 72.9  
 9 1899-12-31 17:30:00 120 57 49 73.1  
10 1899-12-31 18:30:00 109 55 46 72.6  
11 1899-12-31 19:30:00 111 58 46 73.1  
12 1899-12-31 20:30:00 116 60 49 72.7  
13 1899-12-31 21:30:00 112 63 44 72.4  
14 1899-12-31 22:30:00 114 59 49 73.5  
15 1899-12-31 23:30:00 111 55 49 73.2  
16 1900-01-01 00:30:00 121 66 51 73.2

pd\_analysis <- first\_data %>%   
 separate\_wider\_delim(delim = " ",  
 cols = time,  
 names = c("day", "time")) %>%   
 select(time,dp)  
  
pd\_analysis

# A tibble: 16 × 2  
 time dp  
 <chr> <dbl>  
 1 09:30:00 52  
 2 10:30:00 52  
 3 11:30:00 52  
 4 12:30:00 54  
 5 13:30:00 57  
 6 14:30:00 60  
 7 15:30:00 54  
 8 16:30:00 53  
 9 17:30:00 57  
10 18:30:00 55  
11 19:30:00 58  
12 20:30:00 60  
13 21:30:00 63  
14 22:30:00 59  
15 23:30:00 55  
16 00:30:00 66

dp\_mean <- mean(pd\_analysis$dp)  
dp\_mean

[1] 56.6875

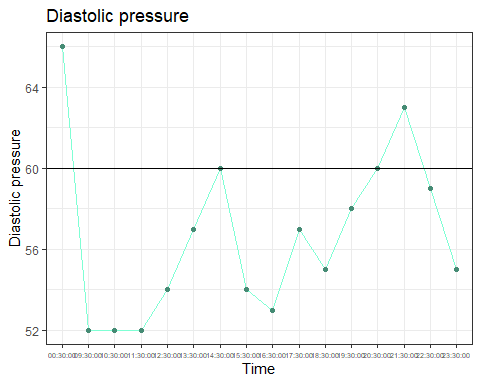
dp\_reference <- 80  
dp\_reference

[1] 80

standard\_deviation\_dp <- sd(pd\_analysis$pd)  
standard\_deviation\_dp

[1] NA

# Plotting the diastolic pressure-------------------------------------------------  
  
pd\_graph <-   
third\_data %>%   
 ggplot(aes(x= time,  
 y= dp))  
pd\_graph+  
 geom\_line(group= "dp",  
 color= "aquamarine")+  
 geom\_point(color="aquamarine4")+  
 geom\_hline(yintercept = 60)+  
 labs(title = "Diastolic pressure",  
 x= "Time",  
 y= "Diastolic pressure")+  
 theme\_bw()+  
 theme(axis.text.x= element\_text( size= "5"))



# Statistical inference of diastolic pressure (pd) --------------------------------  
  
# First hypothesis-----------------------------------------------------------------  
# H0; dp\_mean = pd\_reference  
# H1; dp\_mean is different than pd\_reference  
  
# 1. Calculating the 95% CI -------------------------------------------------------  
  
# Formula:   
  
# Lower CI: dp\_mean-t(n-1, 1-alpha/2)\*(standar\_deviation\_dp/sqrt(16)  
# Upper CI dp\_mean+t(n-1,1-alpha/2)\*(standar\_deviation\_dp/sqrt(16)  
  
t\_15 <- 2.131   
  
t\_15

[1] 2.131

CI\_dp\_15 <- c(dp\_mean-t\_15\*(standard\_deviation\_dp/sqrt(16)), dp\_mean+t\_15\*(standard\_deviation\_dp/sqrt(16)))  
  
CI\_dp\_15

[1] NA NA

# 2. Rejection area method --------------------------------------------------------  
  
# Calculating the pivot  
  
dp\_pivot <- (dp\_mean-dp\_reference)/(standard\_deviation\_dp/sqrt(16))  
  
dp\_pivot

[1] NA

alpha

[1] 0.05

halph\_alpha <- alpha/2  
  
halph\_alpha

[1] 0.025

1-halph\_alpha

[1] 0.975

# Checking normal distribution table  
  
p97.5 <- 1.96  
p2.5 <- -1.96  
  
# Since our pd\_pivot -22.59 , it falls in the rejection area, therefore we reject H0 that says that this young adult that worksout 4 times a week has the same diastolic pressure as a healthy young adult and we accept H1 that the diastolic pressure of this young adult that works out 4 times a week is different than the mean reference diastolic pressure of a healthy young adult.  
  
# 3. Calculating the p-value-------------------------------------------------------  
  
# Now we want to see the probability of getting a pivot of -22.59 if H0 is true:   
# The probability of getting a value less than or equal to -22.59 is the same as  
# the probability of getting a value equal or higher to 22.59 therefore:   
  
# p(Z>22.59)= 1-P(Z<22.59):  
  
1-0.998

[1] 0.002

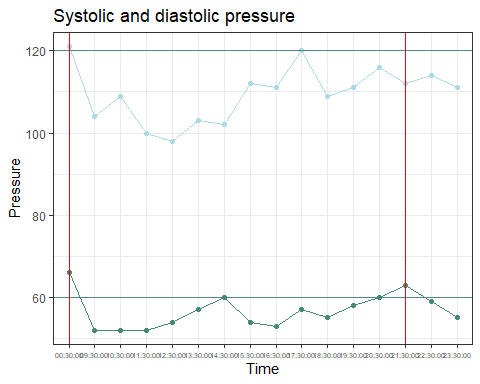
# Since its a 2 sided we multiply that by 2:  
  
pval\_dp <- (1-0.998)\*2  
pval\_dp

[1] 0.004

# Since its smaller than our alhpa (0.05) we reject H0 and accept H1 that its different.   
  
#Second hypothesis-----------------------------------------------------------------  
  
# Now, observing the data, it seems like the value is smaller than 60, we want to make the hypothesis that it is less that the average value for that population:   
  
# H0: dp\_mean = pd\_reference  
# H1: dp\_mean < pd\_reference  
#  
# 1. Calculating the 95% CI ----------------------------------------------------  
  
# With the CI we have already demonstrated that it is below, since it doesnt include H0.  
  
# 2. Rejection are method----------------------------------------------------------   
# Since out pivot is -22.59 it falls within the rejection region we calculated before, when using the rejection method for the systolic pressure, therefore we still reject H0 and we accept H1, the diastolic pressure of this young adult that works out 4 times a week is less than the average pressure value of a healthy young adult.  
  
# 3. Calculating the p-value-------------------------------------------------------  
  
# p(Z<-22.59)=P(z> 22.59)= 1-p(Z<22.59)= pval\_dp  
# Since its a one tail, the pval is the same as pval\_dp :  
  
pval\_one\_side\_dp <- pval\_dp  
pval\_one\_side\_dp

[1] 0.004

# Since its < alpha, which is 0.05 , the probability of having a diastolic presure of 56.6875 if H0 is true is 0.4 % , since it is so rare, H0 has to be false, and H1 is true, the probability of being wrong if we accept H1 is 0.4% , since it is #so low the chances of rejecting H0 and being wrong are extremely low, therefore we can say that H1 is true with very low probability of being wrong.   
  
# Plotting systolic and diastolic pressure together--------------------------------  
  
third\_data %>%   
 ggplot(aes(x= time))+  
 geom\_point(aes(y= sp), color= "lightblue")+  
 geom\_point(aes(y= dp), color= "aquamarine4")+  
 geom\_line(aes(y= sp), group= "sp", color= "lightblue")+  
 geom\_line(aes(y= dp), group= "dp", color= "aquamarine4")+  
 geom\_hline(yintercept = 120, color= "cadetblue4")+  
 geom\_hline(yintercept = 60, color= "cadetblue4")+  
 geom\_vline(xintercept = c("00:30:00", "21:30:00"),  
 color= "red")+  
 labs(title= "Systolic and diastolic pressure",  
 x= "Time",  
 y= "Pressure",  
 )+  
 theme\_bw()+  
 theme(axis.text.x= element\_text( size= "5"))



# Weight comparison----------------------------------------------------------------   
# To finish the main analysis, we will see if the weight of this person that works out 4 times a week is the same as the reference ideal body weight of a person that has a height of 182 cm:   
  
third\_data

# A tibble: 16 × 5  
 time sp dp pulse weight  
 <chr> <dbl> <dbl> <dbl> <dbl>  
 1 09:30:00 104 52 52 72.6  
 2 10:30:00 109 52 58 72.6  
 3 11:30:00 100 52 50 72.9  
 4 12:30:00 98 54 47 72.7  
 5 13:30:00 103 57 44 72.2  
 6 14:30:00 102 60 45 72.1  
 7 15:30:00 112 54 44 72.8  
 8 16:30:00 111 53 47 72.9  
 9 17:30:00 120 57 49 73.1  
10 18:30:00 109 55 46 72.6  
11 19:30:00 111 58 46 73.1  
12 20:30:00 116 60 49 72.7  
13 21:30:00 112 63 44 72.4  
14 22:30:00 114 59 49 73.5  
15 23:30:00 111 55 49 73.2  
16 00:30:00 121 66 51 73.2

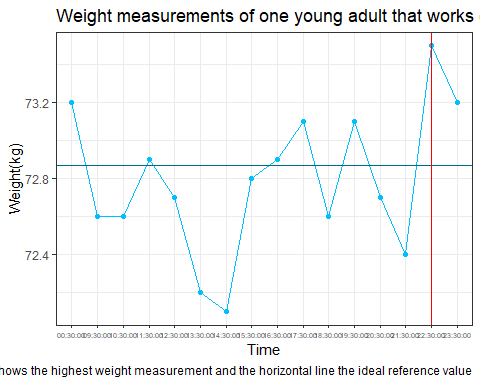
weight\_mean <- mean(third\_data$weight)  
weight\_mean

[1] 72.7875

weight\_reference <- 72.87  
weight\_reference

[1] 72.87

# Plotting weight ----------------------------------------------------------------  
  
third\_data %>%   
ggplot(aes(x= time))+  
 geom\_point(aes(y= weight),  
 color= "deepskyblue")+  
 geom\_line(aes(y= weight),group= "weight", color= "deepskyblue1")+  
 geom\_hline( yintercept= 72.87,  
 color= "deepskyblue4")+  
 geom\_vline(xintercept="22:30:00",  
 color= "red")+  
 labs( title = "Weight measurements of one young adult that works out 4 times a week",  
 caption = "The vertical line shows the highest weight measurement and the horizontal line the ideal reference value",  
 x= "Time",  
 y= "Weight(kg)")+  
 theme\_bw()+  
 theme(axis.text.x= element\_text( size= "5"))



# Statistical inference of weight -------------------------------------------------  
  
#1. calculating the 95% CI, alpha= 0.05--------------------------------------------  
  
# We dont have the population sd , therefore we are on a t-student  
  
sd\_weight <- sd(third\_data$weight)  
sd\_weight

[1] 0.3792537

t\_value\_15\_0.975 <- 2.131  
  
error\_weight <- 2.131\*(sd\_weight/sqrt(16))  
  
CI\_weight <- c(weight\_mean-error\_weight, weight\_mean+error\_weight)  
  
CI\_weight

[1] 72.58545 72.98955

# Since the confidence interval includes the reference value, with the information we have we cannot reject H0.   
  
#2. Rejection area method ---------------------------------------------------------  
  
# H0: weight\_mean = Reference\_weight  
# H1: weight\_mean is different than Reference\_weight  
# alpha= 0.05  
# Calculating the pivot:   
  
pivot\_weight <- (weight\_mean-weight\_reference)/(sd\_weight/sqrt(16))  
pivot\_weight

[1] -0.87013

# The t-student value is (tn-1,1-alpha/2)= t15,1-0.05/2)=t15, 0.975= ± 2.131  
  
# Since our pivot or t = -0.87013, it does not fall within the rejection area, therefore we fail to reject H0, there is not enough evidence that shows a difference between the weight of a person that is 182 cm and works out 4 times a week and the reference ideal value of a person of that height.   
  
#3. calculating the p-value--------------------------------------------------------   
# On a normal curve, the probability of getting a value that is equal or less than -0.87 = the probability of getting a value higher than 0.87 and the probability of getting a value higher than 0.87 = 1-probability of getting a value smaller than 0.87 therefore:   
  
# The probability of getting a value less than 0.87= 0.8078:  
one\_minus\_probability\_higher\_than\_0.87 <- 1-0.8078  
  
one\_minus\_probability\_higher\_than\_0.87

[1] 0.1922

# 0.1922= p(Z< -0.87)= p(Z>0.87)  
  
# Since we are in a two tail:   
  
p\_value\_weight <- 2\*one\_minus\_probability\_higher\_than\_0.87  
p\_value\_weight

[1] 0.3844

# Since the p-value is higher than alpha , we fail to reject H0, again there is not enough evidence that shows a difference between the weight of a person that is 182 cm and works out 4 times a week and the reference ideal value of a person of that height.

## 4 Conclusions

After analysing the systolic , diastolic and weight measurements from an individual that works out 4 times a week and comparing the results with the reference values, it was found that both the systolic and diastolic pressure were lower than the average parameters of 120 /80 mmHg of a healthy individual, however, there were no statistically significant differences between the weight and its reference value.

## 5 Bibliography

1. *High Blood Pressure Symptoms, Causes, and Problems | cdc.gov*. (2023, August 29). Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/about.htm
2. *How much should I weigh? Ideal body weight calculator for women & men*. (n.d.). https://www.calculators.org/health/ideal-weight.php