9/5/2023

# ST 3010 CASE STUDY

Analyzing Health Data for a Community Health Program

**Lasani Balasuriya**
S15598

# Table of Contents

# Objectives of the Study

- Assess the changes in health risk factors (obesity and hypertension) before and after the program.
- Determine if there are any significant differences between various demographic groups in terms of health outcomes.
- Identify correlations between health risk factors and health outcomes.

# Data Description

Note that there are five categorical variables as well as eight quantitative variables as below.

- ID: Unique identifier for each participant.
- Age: Age of the participant.
- Gender: Gender of the participant (Male/Female).
- Ethnicity: Ethnic background of the participant.
- Pre-Weight (kg): Weight of the participant before the program.
- Pre-Height (cm): Height of the participant before the program.
- Pre-BMI: Body Mass Index before the program.
- Pre-Hypertension: Whether the participant had hypertension before the program (Yes/No).
- Pre-Obesity: Whether the participant was obese before the program (Yes/No).
- Post-Weight (kg): Weight of the participant after the program.
- Post-Height (cm): Height of the participant after the program.
- Post-BMI: Body Mass Index after the program.
- Post-Hypertension: Whether the participant had hypertension after the program (Yes/No).
- Post-Obesity: Whether the participant was obese after the program (Yes/No).

## Checking for Missing Values (Data Cleansing)

```
> #Checking for Missing Values
> colSums(is.na(Case_Study))
                ID                Age             Gender 
                 0                  0                  0 
         Ethnicity         Pre_Weight         Pre_Height 
                 0                  0                  0 
           Pre_BMI  Pre_Hypertension        Pre_Obesity 
                 0                  0                  0 
       Post_Weight        Post_Height           Post_BMI 
                 0                  0                  0 
Post_Hypertension       Post_Obesity 
                 0                  0 
```

This shows that there are no missing values for any of the variables.

## Checking for any Duplicate Values

```
> #Checking for any Duplicate Values
> sum(duplicated(Case_Study))
[1] 0
```

Shows that there are none.

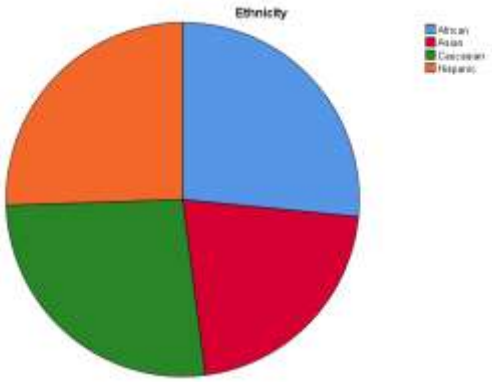
# Data Analysis and Interpretation

## Descriptive Analysis

Note that there are no missing values in the data set. Let's analyze both categorical and numerical data variables.

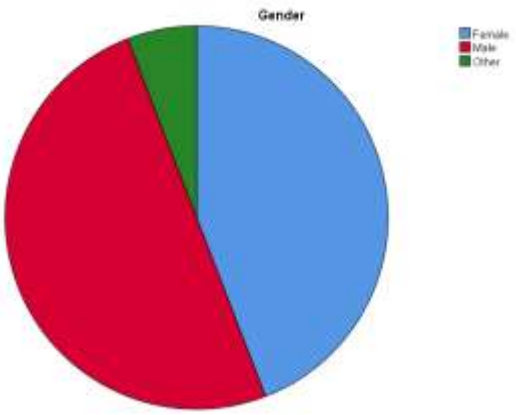
### Categorical Variables

1. Gender
2. Ethnicity
3. Pre and Post hypertension
4. Pre and Post Obesity




Gender

Contains 3 categories. Most of them are males (145), while females (128) and other (17) categories take second and third most in the respective order.




Ethnicity

Contains 4 categories which are African, Asian, Caucasian, and Hispanic. There are respectively 77, 62, 77, 74 respondents from each.

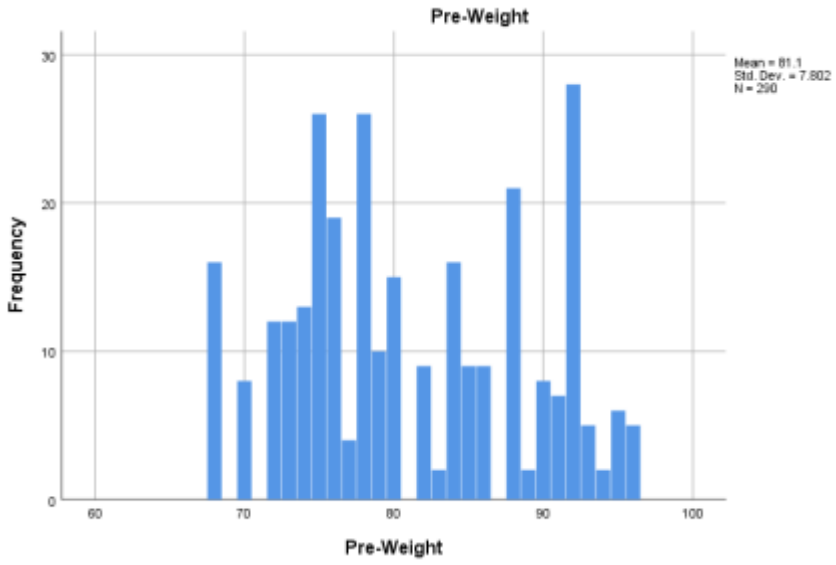
Pre and Post hypertension & Pre and Post Obesity

These contain two categories each. Analysis is done below.

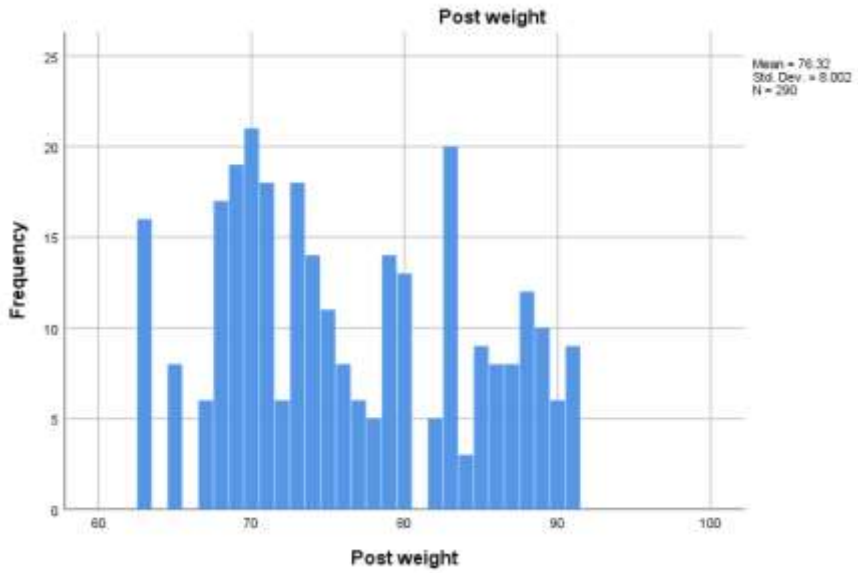
## Quantitative Variables

1. Pre-Weight & Post-Weight
2. Pre-Height & Post-Height
3. Pre-BMI & Post-BMI

Pre-Weight & Post-Weight

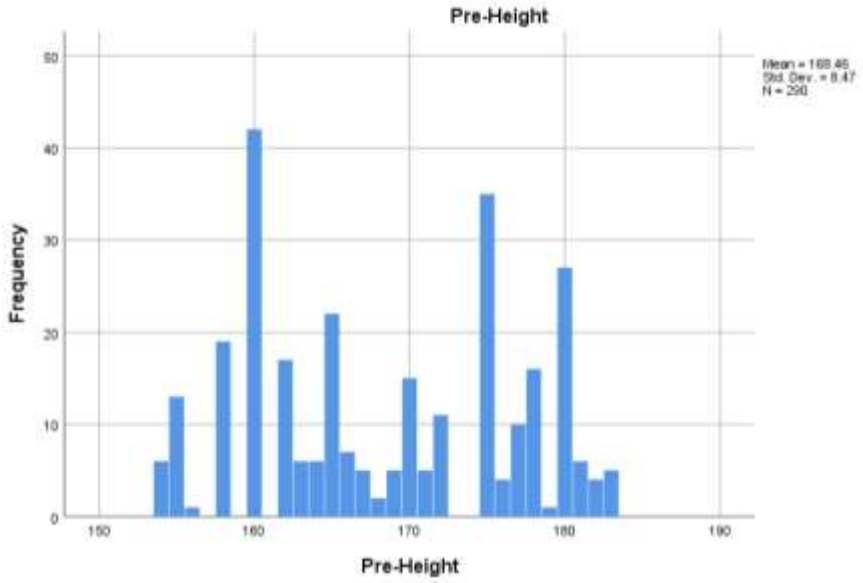



Pre-Weight data seems not to be normally distributed. Most number of respondents seems to be in the range of 70-80kg weights.




Post-Weight data too seems not to be normally distributed. Most number of respondents seems to be shifting from the pre-weight range to less than those weights.

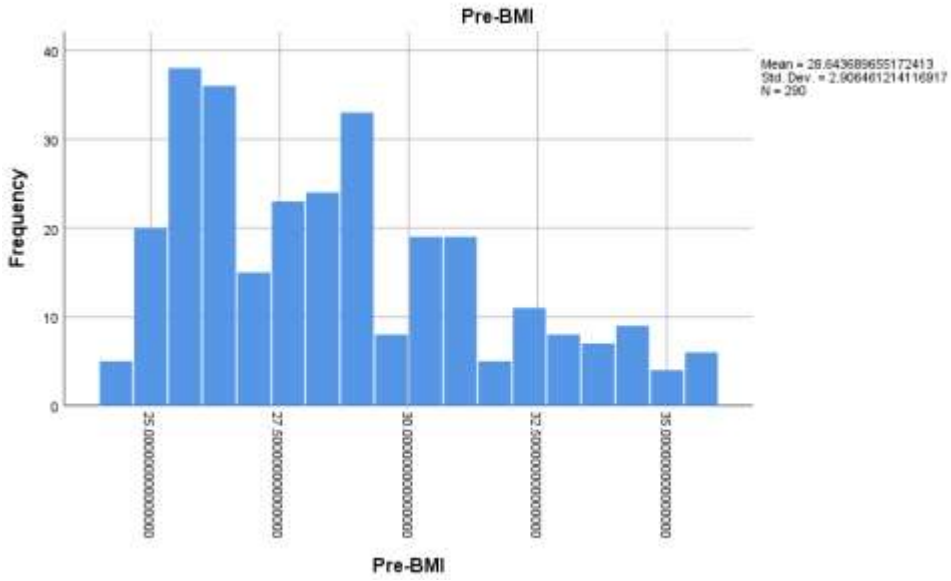
Pre-Height & Post-Height



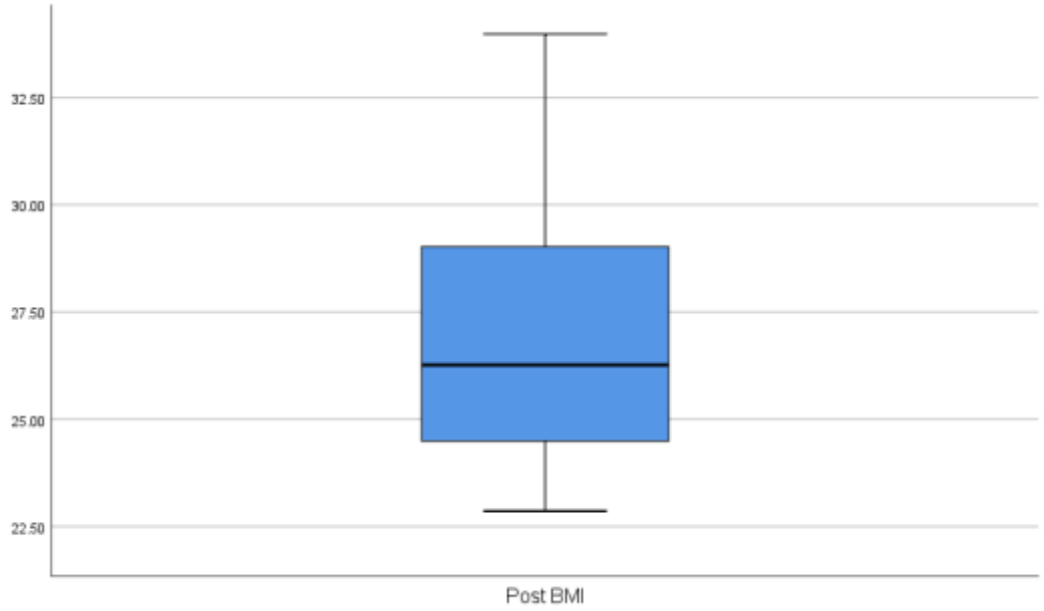

Pre-Height data seems not to be normally distributed. Most number of respondents seems to be in the heights 160 and 175.

Note that the Post-Height data too are the same as Pre-Height data.

Pre-BMI & Post-BMI




Pre-BMI data too seems not to be normally distributed but is skewed to the right.



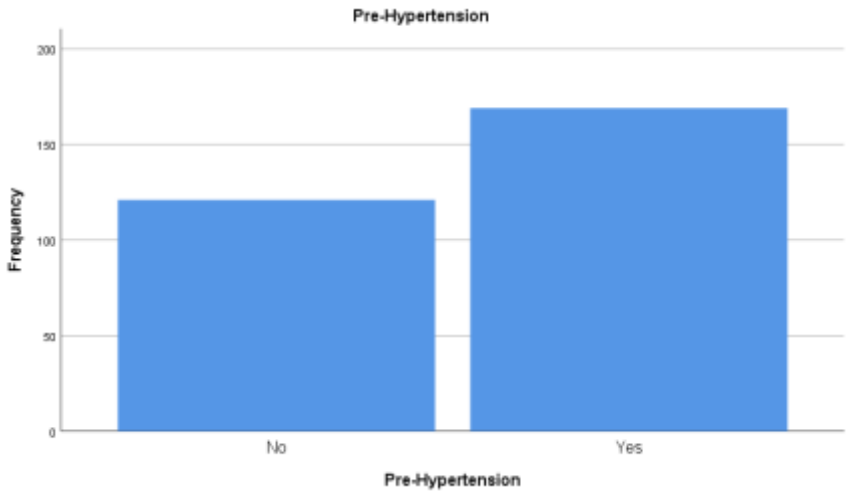

Box plot of Post-BMI variable shows that it is distributed with 26.265 median, skewed to the right, minimum BMI value being 22.86 and the maximum being 33.98.

Further, statistical analysis is needed to determine the statistical significance of the above changes and to assess the program's effectiveness in health goals.

# Objective 1: Assess the changes in health risk factors (obesity and hypertension) before and after the program

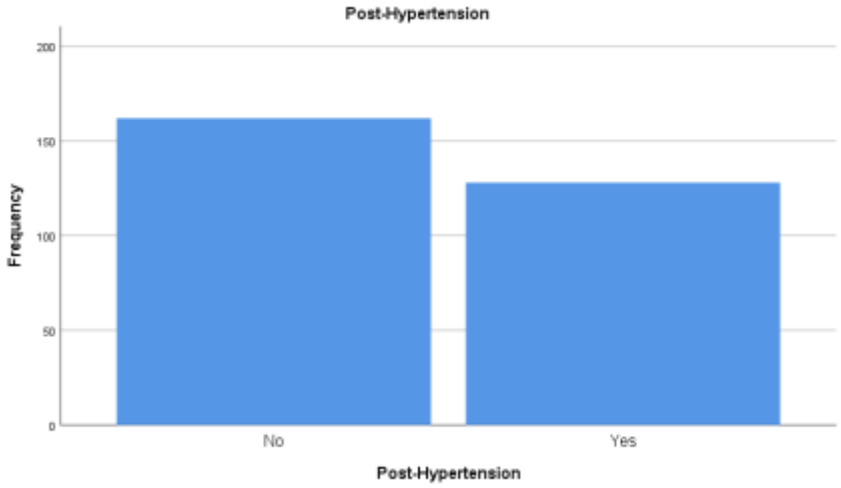
### Hypertension

## 1. Pre-Hypertension




This bar chart shows that out of 290, 169 respondents who suffered from hypertension before has participated to the program.

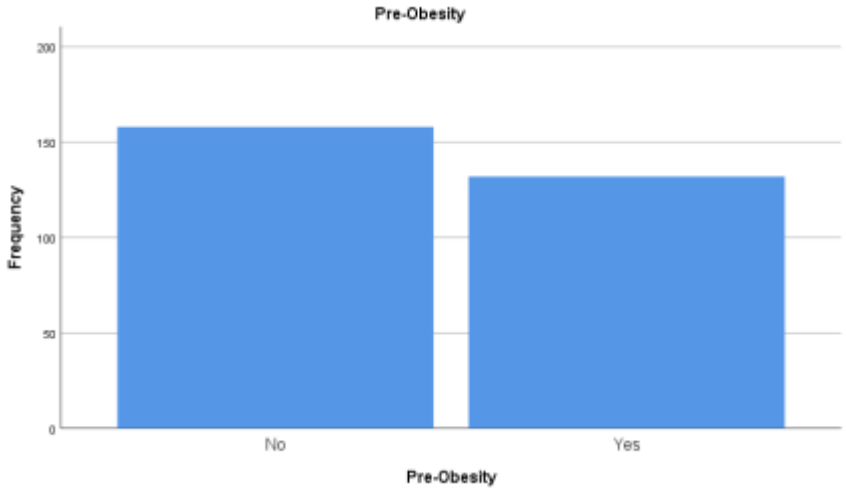
## 2. Post-Hypertension




This bar chart shows that out of 290, 128 respondents still have hypertension after the program. That is, 41 of respondents have recovered from the disease.

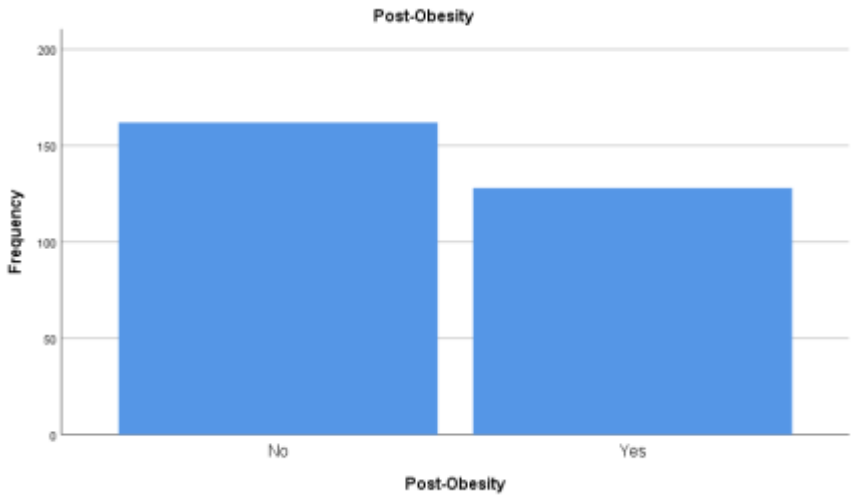
## Obesity

# 1. Pre-Obesity




This bar chart shows that out of 290, 132 respondents who suffered from obesity before has participated to the program.

# 2. Post-Obesity




This bar chart shows that out of 290, 128 respondents still have hypertension after the program. That is, 4 of respondents have recovered from the disease.

## Checking for Normality Assumption by Shapiro-Wilk Normality Test

Now to test for the hypothesis, checking for normality before selecting whether to do a parametric or non-parametric test type is necessary of the variables; hypertension and obesity.

```
> #Checking for normality
> #Shapiro-Wilk normality test
> shapiro.test(Case_Study$Post_Weight)

        Shapiro-Wilk normality test

data:  Case_Study$Post_Weight
W = 0.94875, p-value = 1.586e-08

> shapiro.test(Case_Study$Post_Height)

        Shapiro-Wilk normality test

data:  Case_Study$Post_Height
W = 0.93219, p-value = 3.047e-10

> shapiro.test(Case_Study$Post_BMI)

        Shapiro-Wilk normality test

data:  Case_Study$Post_BMI
W = 0.93482, p-value = 5.464e-10
```

Here, all p values being less than 0.05 implies that normality assumption violated. Another assumption being that sample size should be larger with respect to the target population is also not valid for this case. Therefore, non-parametric tests must be used.

## Wilcoxon Signed Rank Test for Matched Pairs

The most suitable test given the data follows non-normal distribution and there are pre and post data both available, we can use Wilcoxon Signed Rank Test for Matched Pairs.

Since the respondents with a higher BMI increases the risk of having obesity and hypertension, we analyze the BMI decrease or increase so that we can come into a conclusion easily.

Hypothesis;

> $H_0$ : There is no difference of the risk factors in health before and after the health program. ($M_{x-y} = 0$)

$H_1$ : Risk factors are reduced after the health program. ($M_{x-y} > 0$)

```
> #Wilcoxons signed rank test for matched pairs
> wilcox.test(Pre_BMI,Post_BMI, mu=0, alternative = "greater", paired =
T, exact = F, conf.int = T, conf.level = 0.95)

        Wilcoxon signed rank test with continuity correction

data:  Pre_BMI and Post_BMI
V = 42195, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 1.654976      Inf
sample estimates:
(pseudo)median
      1.690086
```

Decision rule is if p-value < 0.05, then reject $H_0$.

There is evidence to conclude that respondents' risk factors on health (Obesity, Hypertension) are reduced after the program.

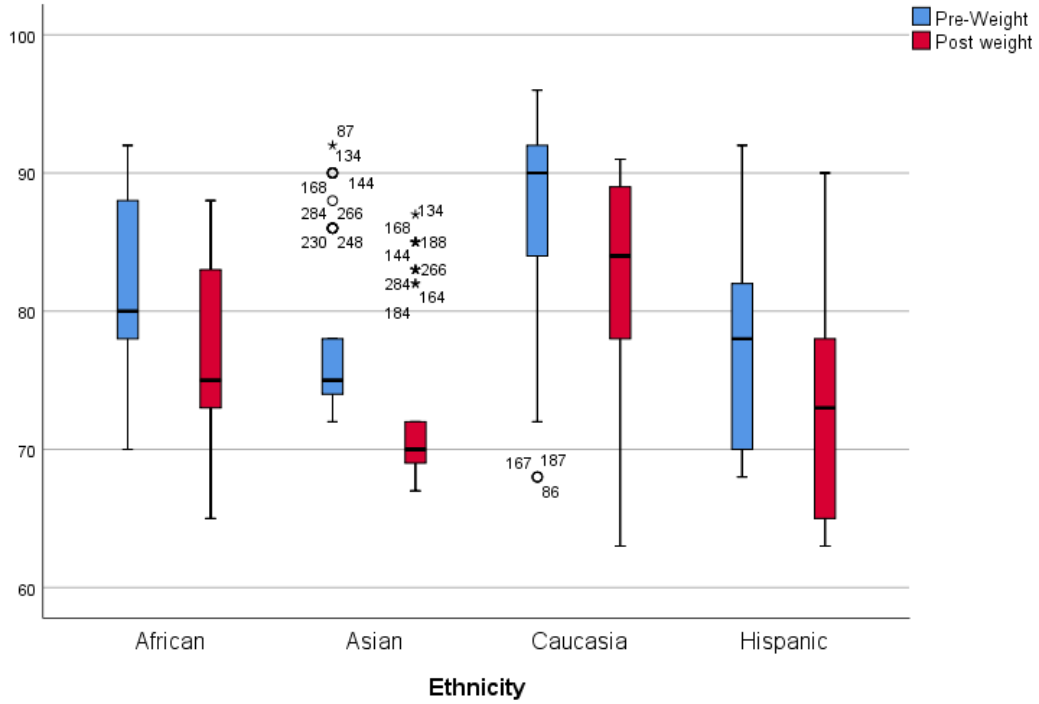
# Objective 2 : Determine if there are any significant differences between various demographic groups in terms of health outcomes

Now we may draw boxplots to do a descriptive analysis and Kruskal Wallis test to do a statistical analysis.

1. Ethnicity with Pre-weight and Post-weight
2. Ethnicity with Pre-height and Post-height
3. Ethnicity with Pre-BMI and Post-BMI
4. Gender with Pre-weight and Post-weight
5. Gender with Pre-height and Post-height
6. Gender with Pre-BMI and Post-BMI
7. Age with Pre-weight and Post-weight
8. Age with Pre-height and Post-height
9. Age with Pre-BMI and Post-BMI

# Ethnicity with Pre-weight and Post-weight

Step 1: Boxplot




Overall results show that post-weights are lower than pre-weights for each of the four ethnicities. So, the reduction of BMI values implies that the respondents have taken benefits from the program in which mostly they have reduced weights. This has reduced the risk of obesity.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-weights are the same for the 4 ethnicities.

$H_1$ : At least one of median post-weights are not the same for an ethnicity group.

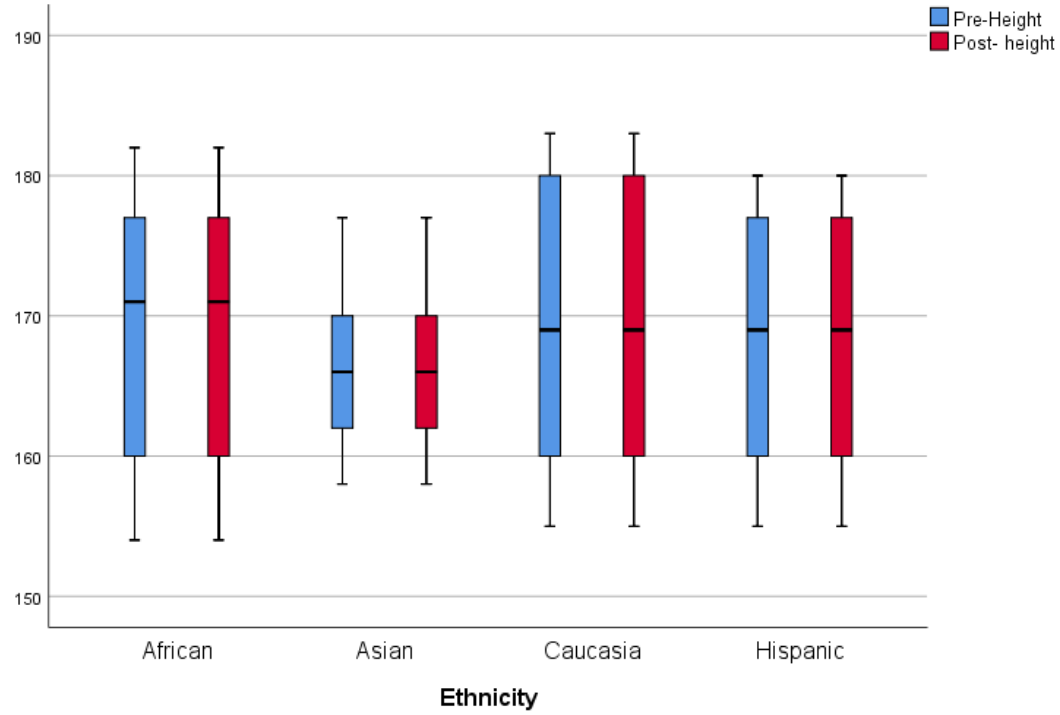
Decision rule is if p-value < 0.05, then reject $H_0$.

```
> #Kruskal Wallis Rank Sum Test
> kruskal.test(Post_Weight ~ Ethnicity, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Weight by Ethnicity
Kruskal-Wallis chi-squared = 67.277, df = 3, p-value =
1.634e-14
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-weights are not the same for the 4 ethnicities.

## Ethnicity with Pre-Height and Post-Height

Step 1: Boxplot




Overall results show that there is no difference between post-and pre-heights for each of the four ethnicities. So, it shows that the respondents have not taken benefits from the program.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-heights are the same for the 4 ethnicities.

$H_1$ : At least one of median post-heights are not the same for any ethnicity group.

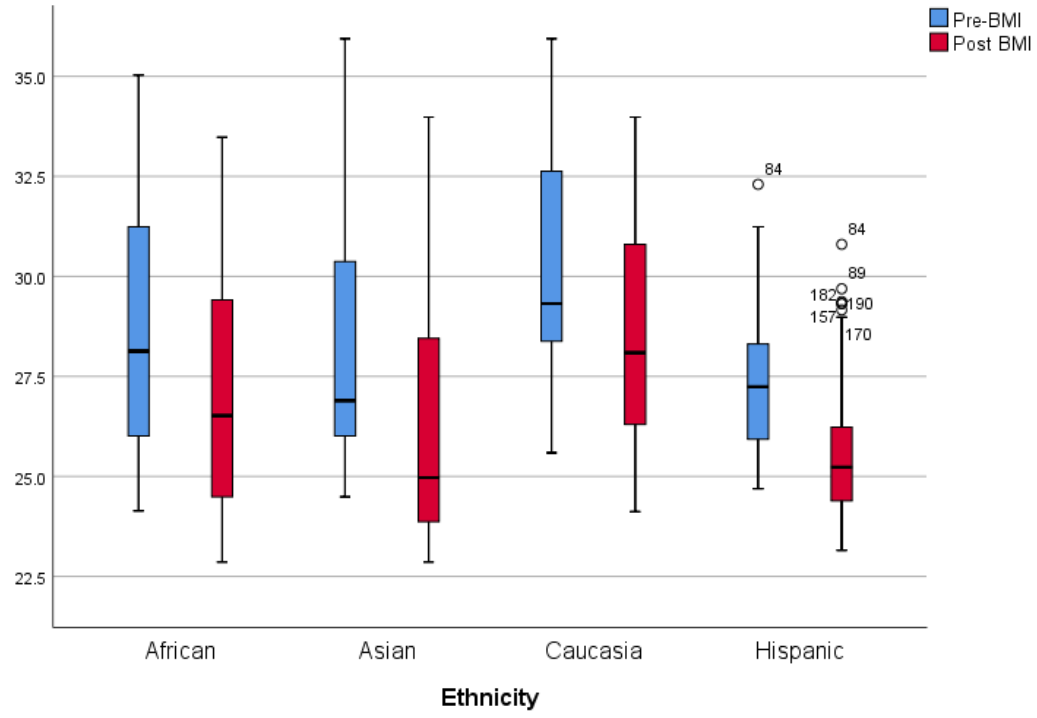
Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_Height ~ Ethnicity, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Height by Ethnicity
Kruskal-Wallis chi-squared = 6.8226, df = 3, p-value =
0.07777
```

Here, p-value is greater than 0.05. Then it does not reject $H_0$. Therefore, there is no enough evidence at 5% significance level to conclude that respondents' median post-heights are not the same for the 4 ethnicities.

## Ethnicity with Pre-BMI and Post-BMI

Step 1: Boxplot



Overall results show that post-BMIs are lower than pre-BMIs for each of the four ethnicities. So, the reduction of BMI values implies that the respondents have taken benefits from the program. This has reduced the risk of obesity.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-BMIs are the same for the 4 ethnicities.

$H_1$ : At least one of median post-BMIs are not the same for an ethnicity group.

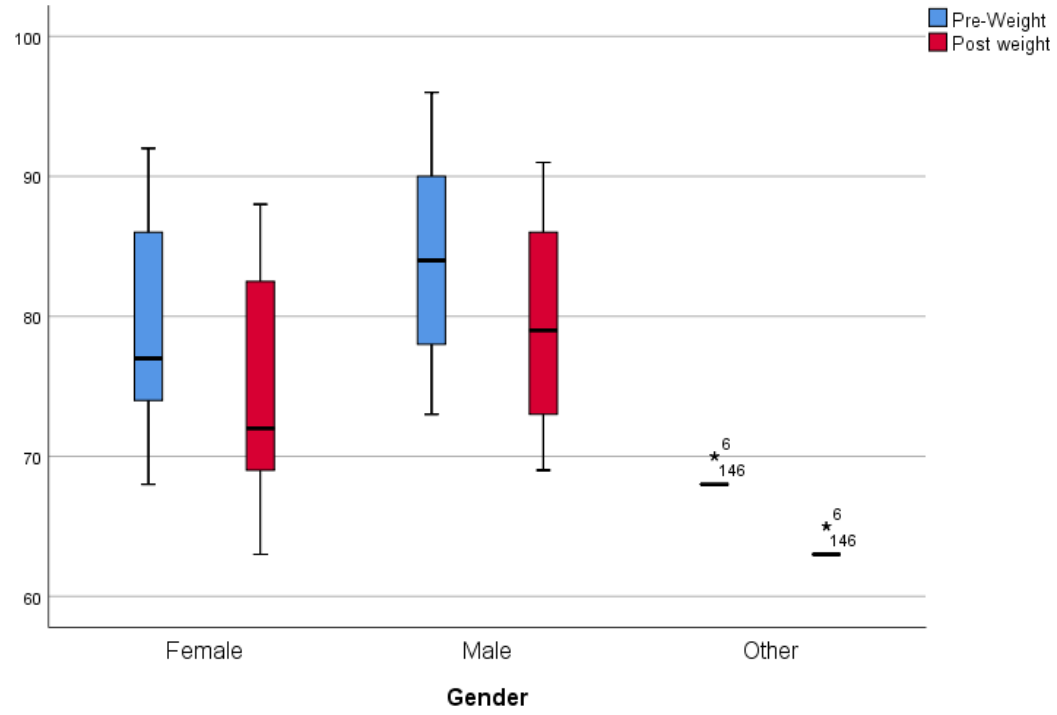
Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_BMI ~ Ethnicity, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_BMI by Ethnicity
Kruskal-Wallis chi-squared = 35.326, df = 3, p-value =
1.04e-07
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-BMIs are not the same for the 4 ethnicities.

# Gender with Pre-weight and Post-weight

Step 1: Boxplot




Here, there is a significance difference between male and females' median pre-weights and post-weights. That is, males have a higher overall pre-weight and a higher overall post-weight than the females, also where the post-weights being less than pre-weights for both gender groups.

Note that since there is a smaller number of response rate for the gender category, 'other' rather than males and females, we cannot use it for comparison. The boxplot doesn't show for the other category. Thus, we can disregard this.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-weights are the same for the 3 gender groups.

$H_1$ : At least one of median post-weights are not the same for any gender group.

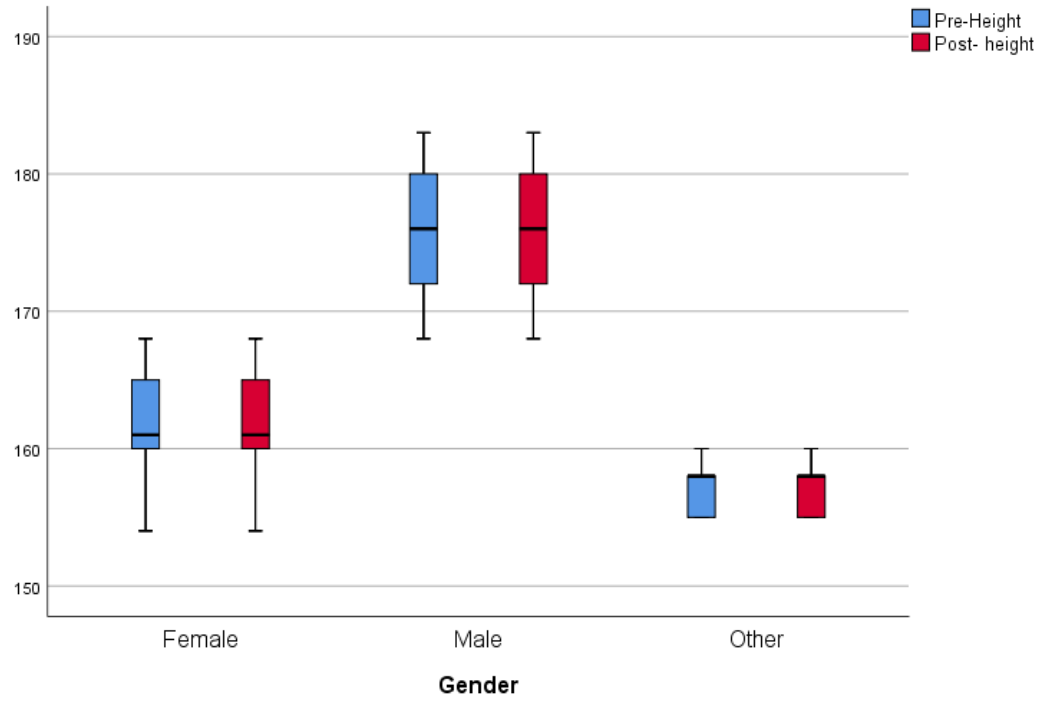
Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_Weight ~ Gender, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Weight by Gender
Kruskal-Wallis chi-squared = 80.913, df = 2, p-value <
2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-weights are not the same for the 3 gender groups.

## Gender with Pre-height and Post-height

Step 1: Boxplot




Here, there is a significance difference between males', females' and other categories 'median pre-heights and post-heights. That is, males have a higher overall pre-height and a higher overall post-height than the females and other gender category, also where there is no difference between post-heights and pre-heights for both gender groups.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-heights are the same for the 3 gender groups.

$H_1$ : At least one of median post-heights are not the same for any gender group.

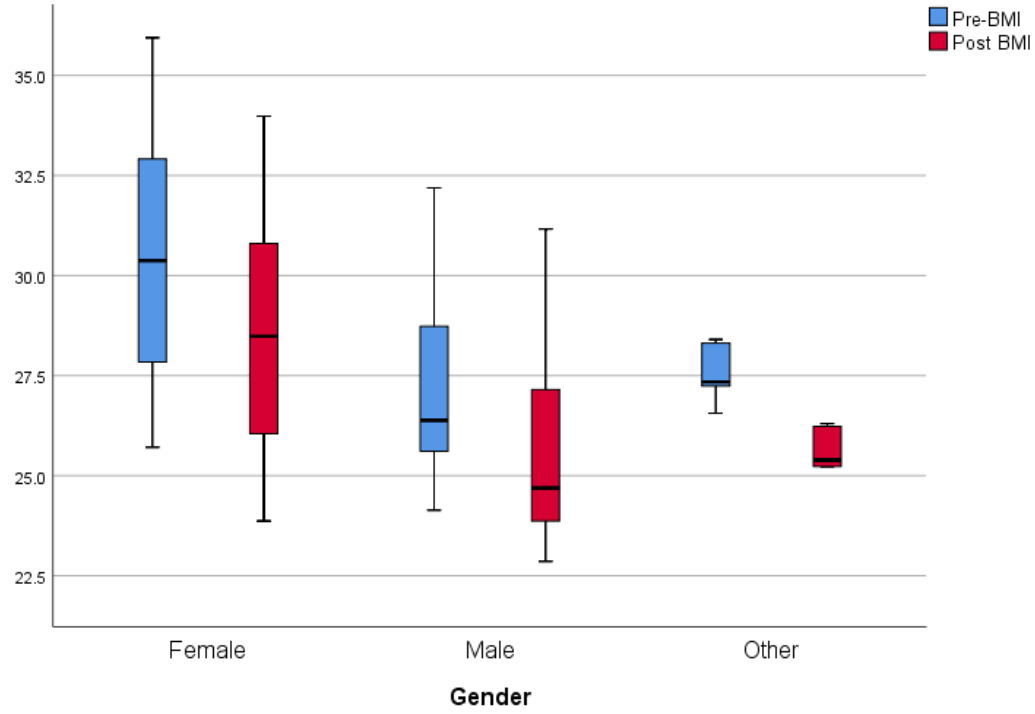
Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_Height ~ Gender, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Height by Gender
Kruskal-Wallis chi-squared = 223.34, df = 2, p-value <
2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-heights are not the same for the 3 gender groups.

# Gender with Pre-BMI and Post-BMI

Step 1: Boxplot




Here, there is a significance difference between males', females' and other categories 'median pre-BMIs and post-BMIs. That is, females have a higher overall pre-BMI and a higher overall post-BMI than the males and other gender category, also where the post-BMIs being less than pre-BMIs for all gender groups.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-BMIs are the same for the 3 gender groups.

$H_1$ : At least one of median post-BMIs are not the same for any gender group.

Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_BMI ~ Gender, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_BMI by Gender
Kruskal-Wallis chi-squared = 72.262, df = 2, p-value <
2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-BMIs are not the same for the 3 gender groups.

## Age

Since Age variable is numerical, it's better to divide them into three age categories so that the analysis is easier to interpret.

Less than and equals 36 → "Young"

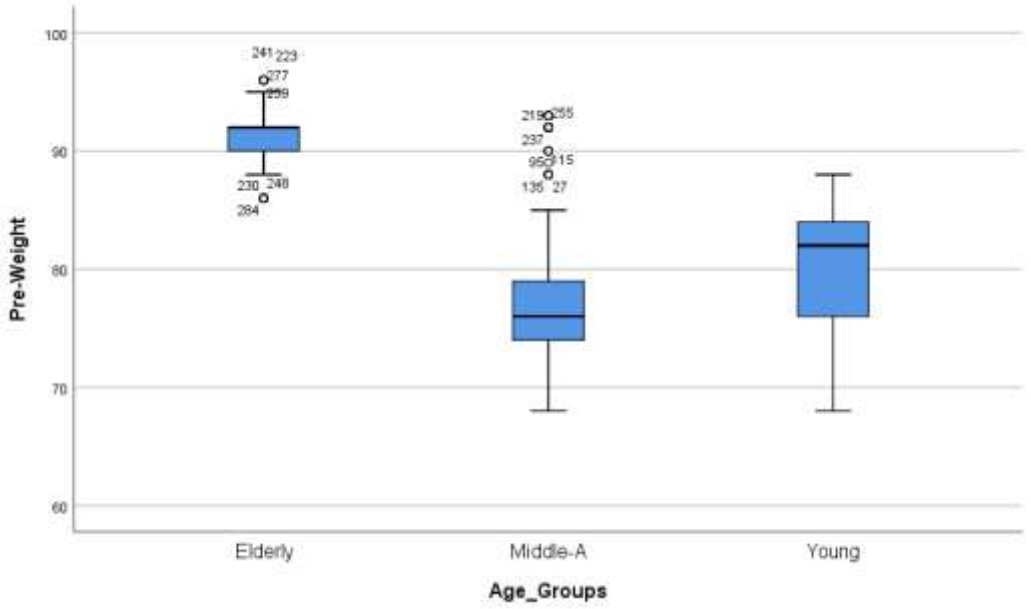
Between 36 and 52 → "Middle-aged"

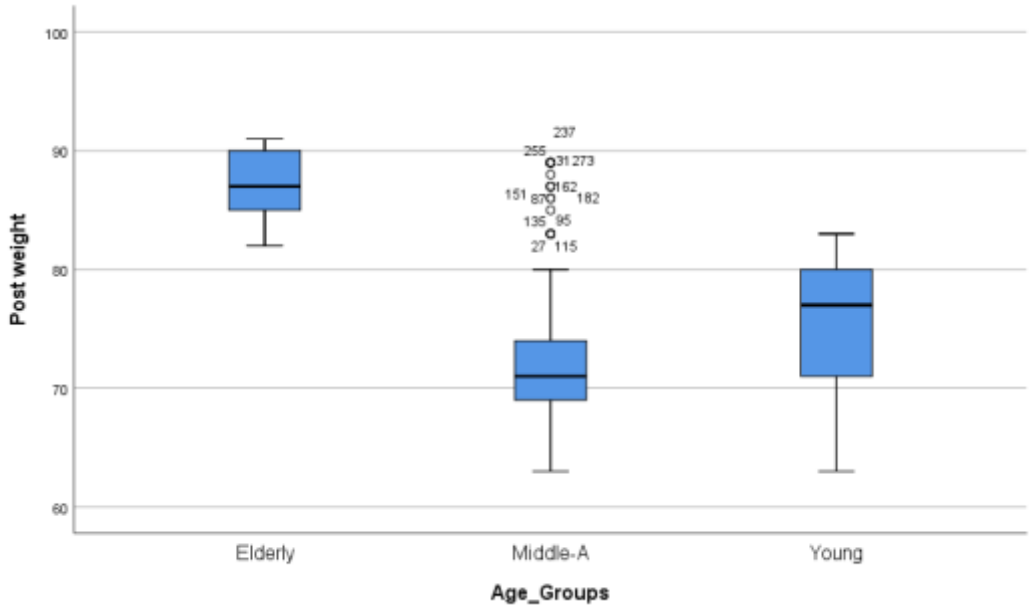
Greater than or equals 66 → "Elderly"

Then we will be applying Kruskal Wallis test for further analysis between age categories.

# Age with Pre-weight and Post-weight

Step 1: Boxplot




This shows that the elderly age group has higher pre-weights and post-weights.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-weights are the same for the 3 age groups.

$H_1$ : At least one of median post-weights are not the same for any age group.

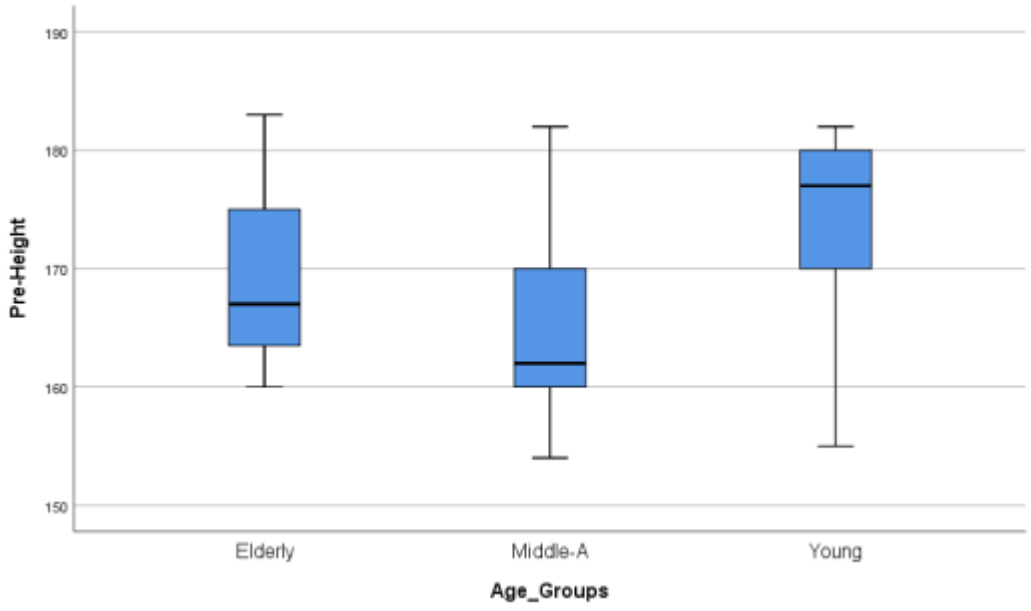
Decision rule is if p-value < 0.05, then reject $H_0$.

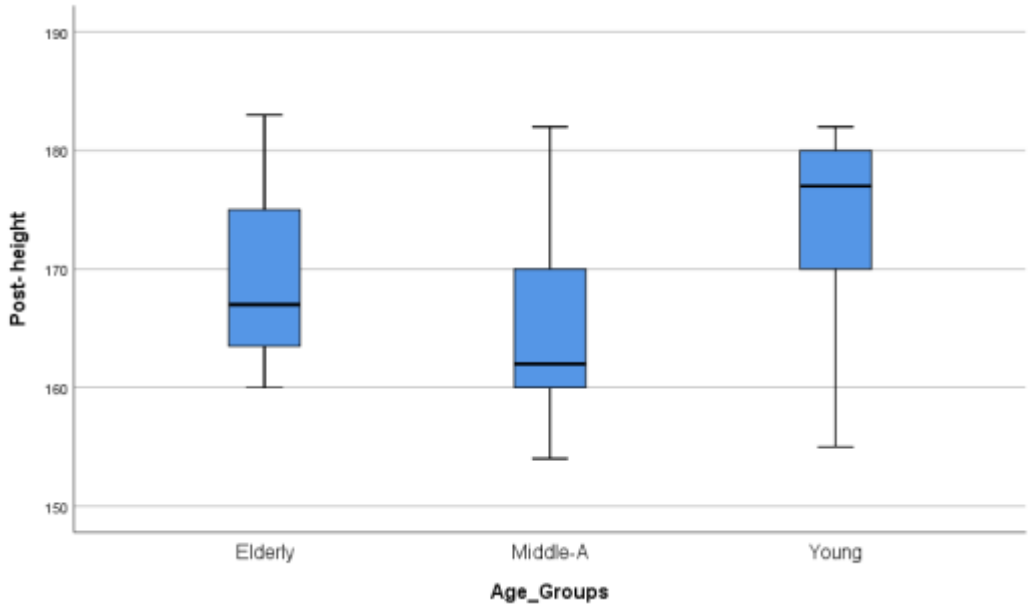
```
> kruskal.test(Post_Weight ~ age_categories, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Weight by age_categories
Kruskal-Wallis chi-squared = 117, df = 2, p-value < 2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-weights are not the same for the 3 age groups.

## Age with Pre-height and Post-height

Step 1: Boxplot






This shows that the younger age group has higher pre-heights and post-heights.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-heights are the same for the 3 age groups.

$H_1$ : At least one of median post-heights are not the same for any age group.

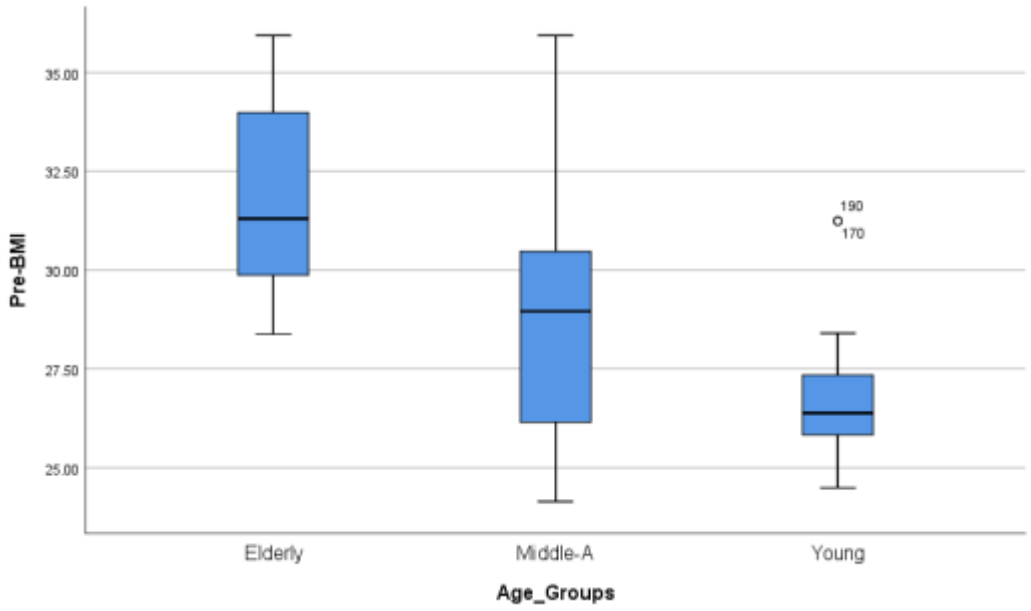
Decision rule is if p-value < 0.05, then reject $H_0$.

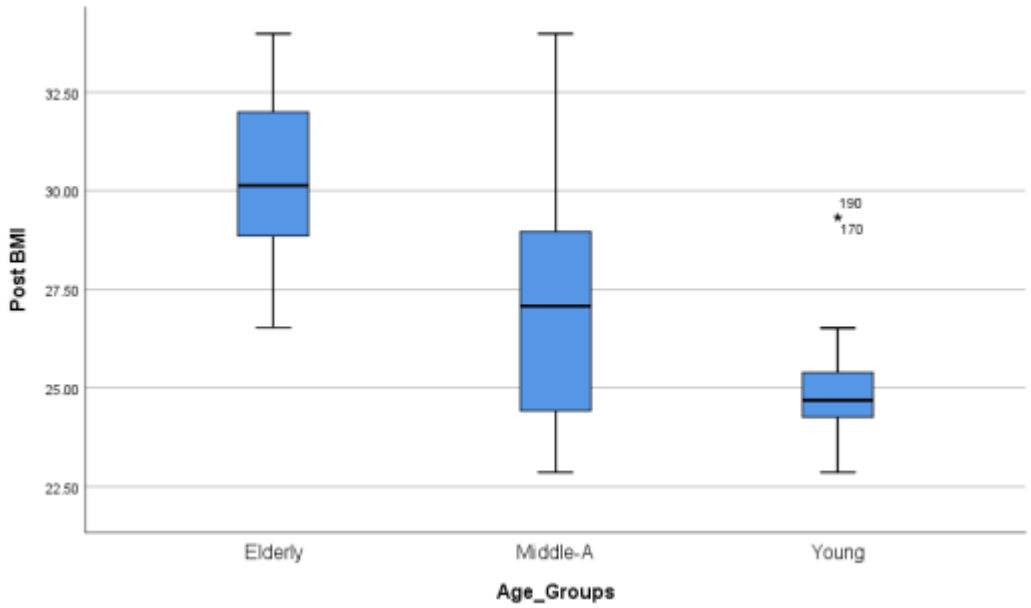
```
> kruskal.test(Post_Weight ~ age_categories, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_Weight by age_categories
Kruskal-Wallis chi-squared = 117, df = 2, p-value < 2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-heights are not the same for the 3 age groups.

# Age with Pre-BMI and Post-BMI

Step 1: Boxplot




This shows that the elderly age group has higher pre-BMIs and post-BMIs.

Step 2: Kruskal Wallis Rank Sum Test

Hypothesis;

$H_0$: Median post-heights are the same for the 3 age groups.

$H_1$ : At least one of median post-heights are not the same for any age group.

Decision rule is if p-value < 0.05, then reject $H_0$.

```
> kruskal.test(Post_BMI ~ age_categories, data = Case_Study)

        Kruskal-Wallis rank sum test

data:  Post_BMI by age_categories
Kruskal-Wallis chi-squared = 107.34, df = 2, p-value <
2.2e-16
```

Here, p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that respondents' median post-BMIs are not the same for the 3 age groups.

# Objective 3: Identify the correlations between health risk factors and health outcomes

Health Risk Factors:

- Pre-Weight
- Pre-Height
- Pre-BMI
- Pre-Hypertension
- Pre-Obesity

Health Outcomes:

- Post-Weight
- Post-Height
- Post-BMI
- Post-Hypertension
- Post-Obesity

Correlations means the relationship between two numerical variables. Therefore, above mentioned variables are used to see the relationship. Furthermore, Pearson correlation test is used.

Decision rule is if p-value < 0.05, then reject $H_0$.

```
> #Correlation tests
> cor.test(Pre_Weight,Post_Weight,method = "pearson")

        Pearson's product-moment correlation

data:  Pre_Weight and Post_Weight
t = 160.17, df = 288, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9929894 0.9955808
sample estimates:
      cor
0.9944335

> cor.test(Pre_Height,Post_Height,method = "pearson")

        Pearson's product-moment correlation

data:  Pre_Height and Post_Height
t = Inf, df = 288, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
cor
  1

> cor.test(Pre_BMI,Post_BMI,method = "pearson")

        Pearson's product-moment correlation

data:  Pre_BMI and Post_BMI
t = 143.74, df = 288, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9913146 0.9945234
sample estimates:
      cor
0.9931025
```

## Correlation between Pre-weight and Post-weight

$H_0$: There is no association between pre-weight and post-weight.

$H_1$: There is an association between pre-weight and post-weight.

From the above R codes, we see that the p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that there is an association between pre-

weights and post-weights. Also, since the correlation coefficient is 0.994335, there exists a strong positive relationship between respondents' pre-weights and post-weights.

## Correlation between Pre-weight and Post-weight

$H_0$: There is no association between pre-height and post-height.

$H_1$: There is an association between pre-height and post-height.

From the above R codes, we see that the p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that there is an association between pre-heights and post-heights. Also, since the correlation coefficient is 1, there exists a strong positive relationship between respondents' pre-heights and post-heights.

## Correlation between Pre-weight and Post-weight

$H_0$: There is no association between pre-BMI and post-BMI.

$H_1$: There is an association between pre-BMI and post-BMI.

From the above R codes, we see that the p-value is less than 0.05. Then it rejects $H_0$. Therefore, there is evidence at 5% significance level to conclude that there is an association between pre-BMIs and post-BMIs. Also, since the correlation coefficient is 0.9931025, there exists a strong positive relationship between respondents' pre-BMIs and post-BMIs.

# Discussions and Conclusions

According to the study we can conclude that, respondents' weights and BMI values get decreased after the health program is conducted which indicates that the health program had a good impact on reducing the risk factors of health, which are defined as obesity and hypertension. However, there seems to be no change of height along with the health program.

From the first objective we can conclude that the number of respondents who had obesity and hypertension decreased after the program. This gives a positive sign from the health program.

From the second objective, we detect that there are some significant differences between ethnicity groups, gender, and age groups itself in terms of pre and post values of weights and BMIs. (For each group, there were significant differences in pre and post values) However, there were no any significant difference in pre-height values and post-height values. Overall respondent heights have not changed, afterall.

Finally, from the third objective, we detect that there is a strong positive relationship between pre-weight and post-weight values as well as pre-BMI and post-BMI values.

In conclusion, this health program has had a positive impact on reducing hypertension, obesity and high BMI within the study population.