IMPACT OF A WORKPLACE WELLNESS PROGRAM ON EMPLOYEE HEALTH

Case Study

ST 3010

Tharindu Darshana
s16341

Table of Contents

1. Introduction

Employee welfare has become a major concern for businesses that want to raise productivity, lower absenteeism and boost overall job satisfaction. Workplace wellness programs have emerged as an approach to the promotion of the health and well-being of employees. Such initiatives usually include a variety of activities and resources that encourage healthier lifestyles such as exercise sessions, mental health support and nutritional counseling. Wellness programs aim at physical and mental health to establish a more supportive and health-conscious work environment.

This case study examines the impact of a nine-month workplace wellness programme undertaken by a large corporation. The primary goals were to reduce stress, promote physical fitness, and improve mental health among workers. Regular workout sessions, mental health support workshops, and personalized nutrition advice were some of the features incorporated into this kind of plan.

The research collected data on several categories related to people's lifestyles such as demography information before and after the program to evaluate how effective this initiative was. The purpose of this study is threefold; firstly, it seeks to determine whether stress levels have changed as a result of the wellness program. Secondly, it aims to establish whether there are any documented changes in these employees' physical well-being in terms of fitness levels or body mass index (BMI).

2. Methodology

The data for this research has been taken from a corporate wellness program that has run over a period of nine months. This dataset consists of 200 employees across different departments, with explicit data regarding their demographics, job roles, and health metrics at the start and end of the program. Each of these employees is uniquely identified by an Employee ID. Key variables in the dataset include age, gender, and department, all denoting a highly diverse workforce. The health-related metrics were measured and recorded at baseline and post, so they gave a comprehensive look at the changes in the health status of the employees over the period of intervention.

The following variables were included in the dataset:

- **Demographic and Job Information:**
  - `Employee_ID`: Unique identifier for each employee.
  - `Age`: Age of the employee, ranging from 22 to 65 years.
  - `Gender`: Gender of the employee, categorized as "Male" or "Female."
  - `Department`: Department in which the employee works, categorized as "Sales," "Engineering," "HR," or "Finance."
  - `Job_Role`: Specific job role within the department, classified according to the department.
- **Pre-Program Health Metrics:**
  - `Pre_Stress_Level`: Employee's stress level before the wellness program, measured on a scale from 4 (low) to 10 (high).

- Pre_Exercise_Frequency: Frequency of exercise per week before the program, ranging from 0 to 5 times.
- Pre_Smoking_Status: Employee's smoking status before the program, categorized as "Smoker" or "Non-Smoker."
- Pre_BMI: Body Mass Index before the program, ranging from 18.5 to 35.0.
- Pre_Mental_Health_Score: Employee's mental health score before the program, measured on a scale from 4 (low) to 10 (high).

- **Post-Program Health Metrics:**
  - Post_Stress_Level: Employee's stress level after the wellness program, adjusted based on the program's impact.
  - Post_Exercise_Frequency: Frequency of exercise per week after the program, reflecting changes due to the program.
  - Post_Smoking_Status: Employee's smoking status after the program, indicating whether any employees quit smoking.
  - Post_BMI: Body Mass Index after the program, reflecting weight changes due to the the program.
  - Post_Mental_Health_Score: Employee's mental health score after the program, showing the program's impact on mental health.

To assess the effectiveness of the wellness program and investigate potential variations in outcomes, several statistical analyses are to be conducted using R.

- To access the impact of wellness program:
  - Paired t-tests can be used to evaluate the changes in health metrics like 'Stress_Level', 'Exercise_Frequency' and 'Mental_Health_Score' before and after the wellness program. This test is appropriate for comparing two related samples or measurements, in this case, the pre and post-program data for each employee. Wilcoxon signed-rank test can be used as a corresponding non-parametric test to the paired t-test.

- Investigate variations by Job Role or Department:
  - Since there are more than two categories for job roles and departments, we can use ANOVA (Analysis of Variance) to compare mean changes in health metrics across different departments and job roles. As a corresponding non-parametric test we can use the Kruskal-Wallis test.

- Investigate Relationships Between Lifestyle Factors and Health Outcomes:
  - To analyze the relationship between smoking status and health outcomes such as stress level, BMI value and mental health score, we can use Wilcoxon's rank sum test for two individual samples. We can measure the correlation between exercise frequency and health outcomes using Pearson's correlation as a parametric test and Kendall's correlation as a non-parametric test.

3. Analysis

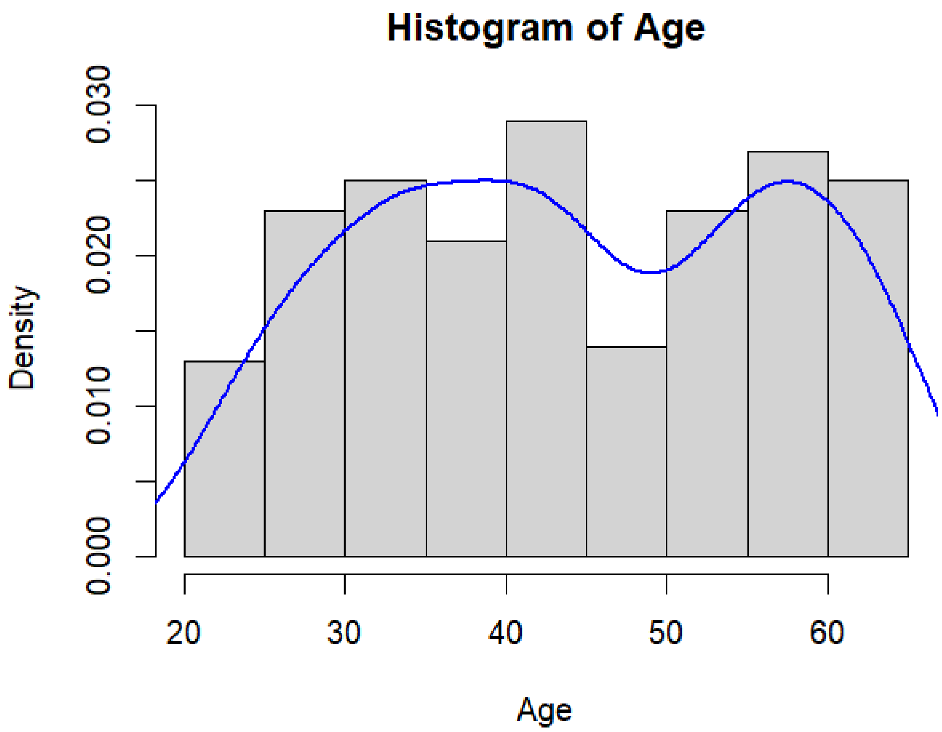
This section aims to evaluate the effect the workplace wellness program has on the health outcomes of employees. The analysis is divided into two major parts: Descriptive Analysis and Advanced Analysis. Descriptive Analysis is intended to provide a comprehensive overview of the dataset, summarizing key demographic and health-related variables. The Advanced Analysis involves the conduction of more advanced statistical tests to consider the effectiveness of the wellness program and exploration of the relationship between lifestyle factors and changes in health outcomes. Together, the analyses provide a thorough evaluation of the data, insights into the success of the wellness program, and identification of drivers behind positive health changes among employees.

3.1 Descriptive Analysis

3.1.1 Demographic Variables

Age

```
> summary(Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   22.0    34.0    43.0    44.4    56.0    65.0 
```



The participants' ages range from about 22 to 65. The mean and median ages for this group are 43 and 44, respectively. By observing the histogram we can see the majority of the employees are in the range of age 30-45 and 50-60

Gender Distribution

```
> table(Gender)
Gender
Female   Male
    88    112
```

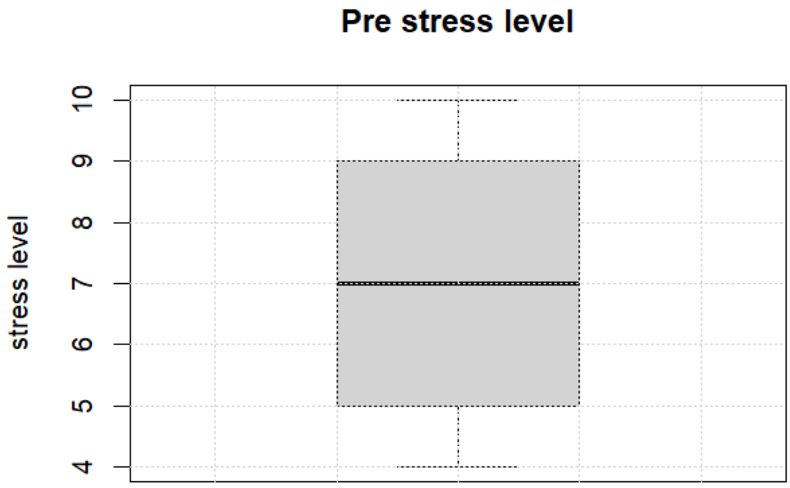
Department Distribution

```
> table(Department)
Department
Engineering     Finance          HR        Sales
         54          51          40           55
```

This table shows the number of employees in each department. This also seems to be fairly evenly distributed across departments.

3.1.2 Pre-Program Health Metrics

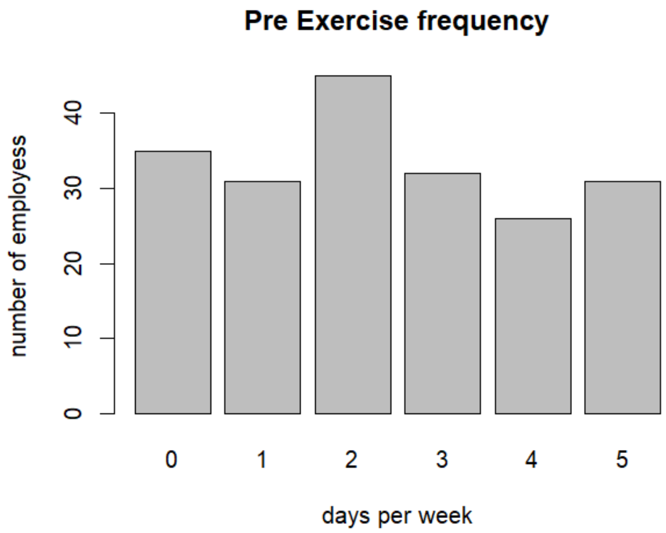
Stress Level




```
> summary(Pre_Stress_Level)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.000   5.000   7.000   7.075   9.000  10.000 
```

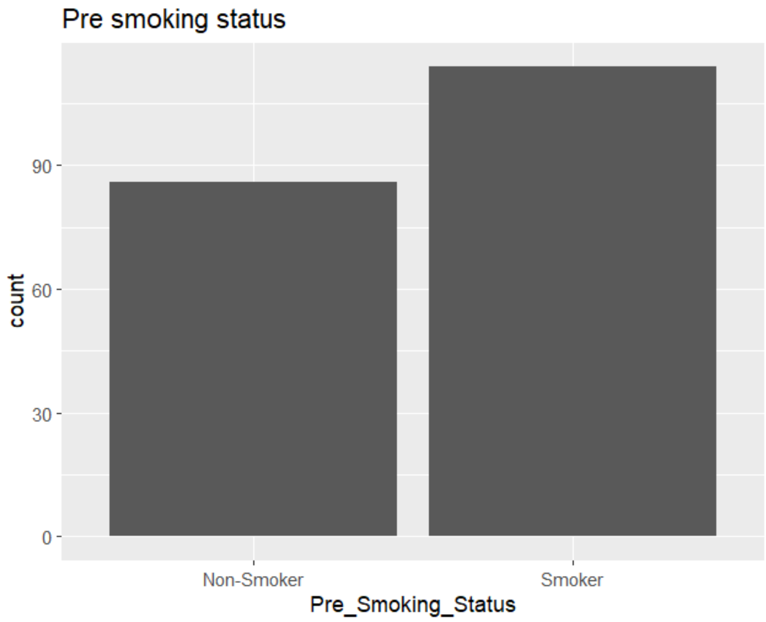
The median and mean stress levels of employees before the wellness program are around 7, ranging from 4 to 10.

Exercise Frequency



This plot shows the number of times employees exercise per week. The majority of employees exercise twice a week, while other frequencies are approximately uniformly distributed.

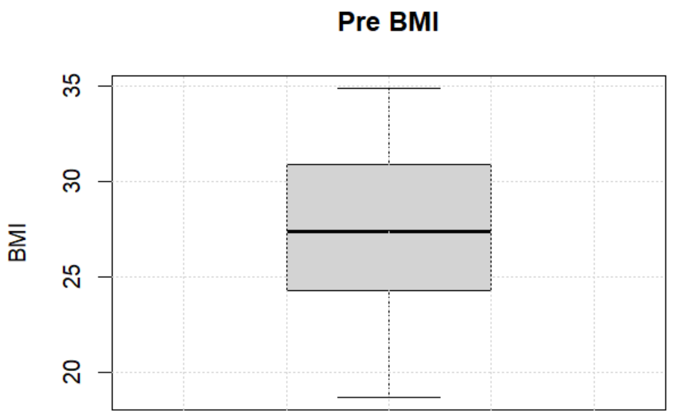
Smoking Status

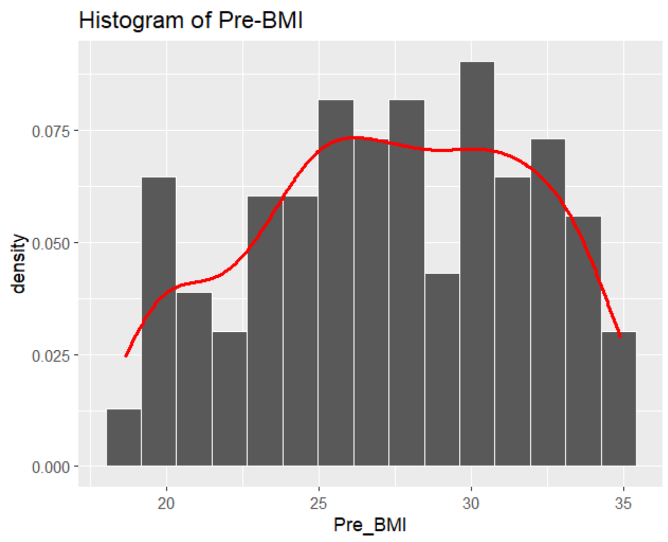



```
> table(Pre_Smoking_Status)
Pre_Smoking_Status
Non-Smoker     Smoker
        86        114
```

Here we can observe number of smokers is higher than non-smokers before the program.

BMI

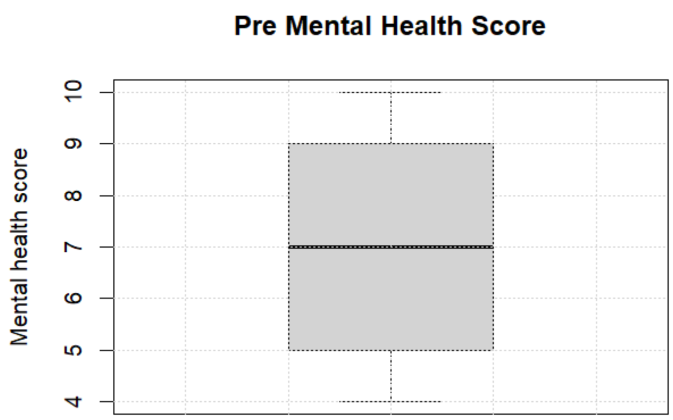





```
> summary(Pre_BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.65   24.29   27.39   27.28   30.90   34.91 
```

We can see the median and mean BMI of employees before the workout around 27 and it ranges from 18.65 to 34.91. By observing the histogram we can say that the distribution of BMI is approximately normally distributed.

Mental Health Score




```
> summary(Pre_Mental_Health_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.000   5.000   7.000   7.025   9.000  10.000 
```

The median and mean mental health scores of employees before the wellness program are around 7, ranging from 4 to 10.

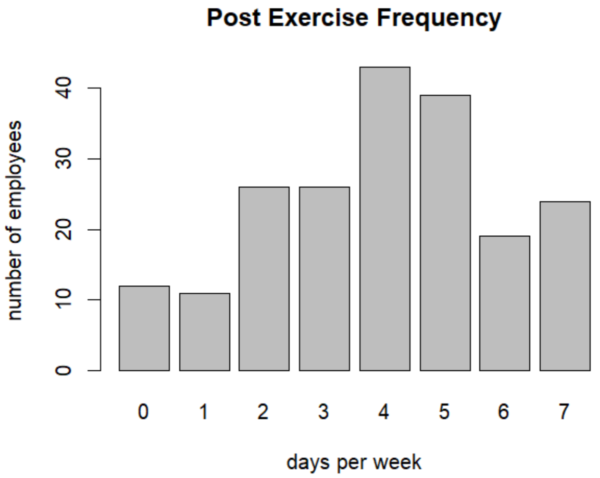
3.1.3 Post-Program Health Metrics

<u>Stress Level</u>


```
> summary(Post_Stress_Level)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   5.000   5.415   7.000  10.000
```

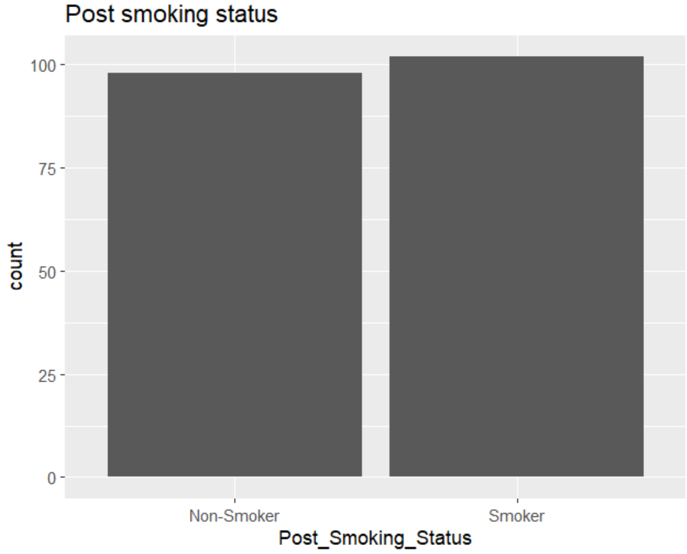
The median stress level of employees after the wellness program is 5, ranging from 1 to 10. In the boxplot, the median is closer to the first quartile than the third quartile, indicating a skew towards lower stress levels which might be an indicator of the program's effectiveness on stress level.

<u>Exercise Frequency</u>



We can observe that the majority of employees exercise 4 or 5 times per week after the program. While there are less number of employees exercise 0 or 1 times per week.

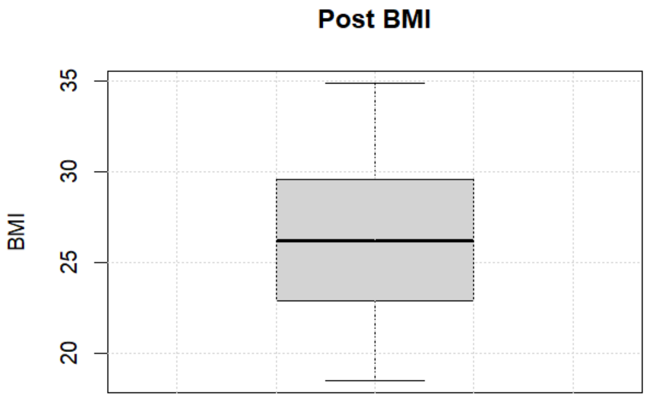
Smoking Status

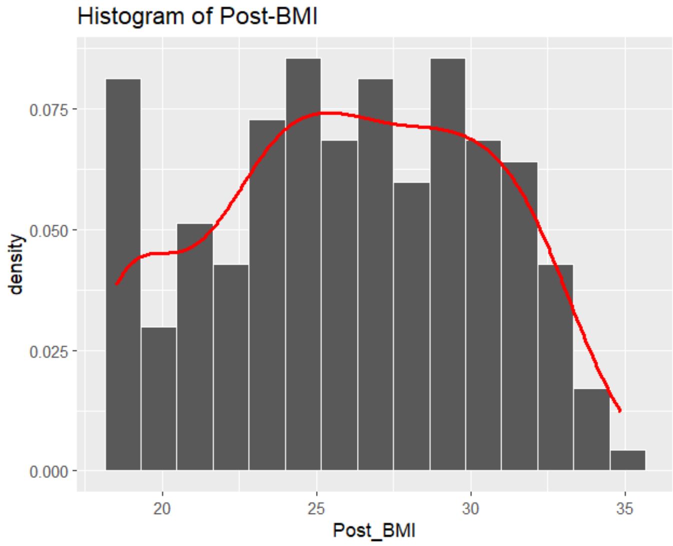



```
> table(Post_Smoking_Status)
Post_Smoking_Status
Non-Smoker     Smoker
        98        102
```

The number of smokers and non-smokers is almost equal after the program.

BMI

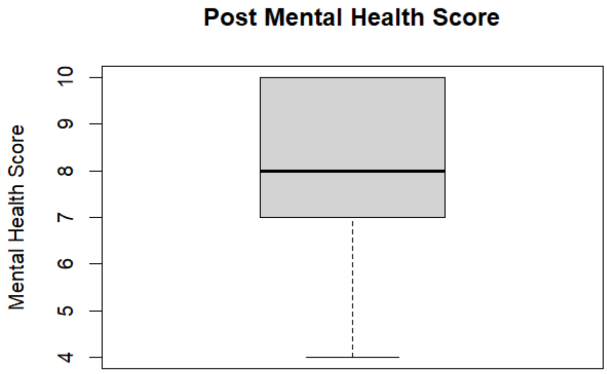





```
> summary(Post_BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.50   22.92   26.18   26.07   29.54   34.88
```

We can see the median and mean BMI of employees before the workout around 26 and it ranges from 18.5 to 34.88. By observing the histogram, we can see that the number of employees with very high BMI values is extremely low.

Mental Health Score




```
> summary(Post_Mental_Health_Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.000   7.000   8.000   8.105  10.000  10.000 
```

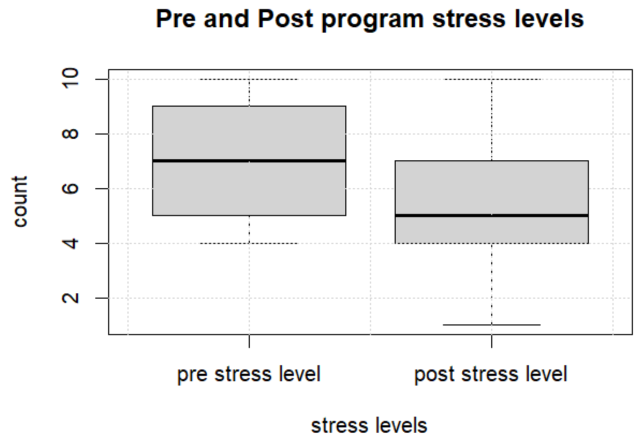
The median and mean mental health scores of employees before the wellness program are around 8, ranging from 4 to 10.

3.2 Advanced Analysis

This section aims to examine how the wellness program affects various health outcomes. This analysis goes beyond simple descriptive statistics to use more advanced methods of statistics. We will use statistical tests such as t-tests to investigate the significance of the impact of the wellness program on stress levels and physical fitness and explore variations across different job roles or departments.

3.2.1 Impact of Wellness Program on Stress Level

Here we will use visual inspection followed by statistical tests to confirm the significance of the results. Since stress level is a discrete ordinary variable, we can't perform parametric tests on the stress level data as parametric tests can only be used with continuous data. Therefore we will use a non-parametric test as it can handle discrete ordinal data. Since we access the pre and post-stress levels, Wilcoxon's sign-ranked test for matched pairs is the most appropriate test for this.




```
> wilcox.test(Post_Stress_Level,Pre_Stress_Level,mu = 0,alternative = 'less',paired = T,conf.level = 0.95)

	Wilcoxon signed rank test with continuity correction

data:  Post_Stress_Level and Pre_Stress_Level
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

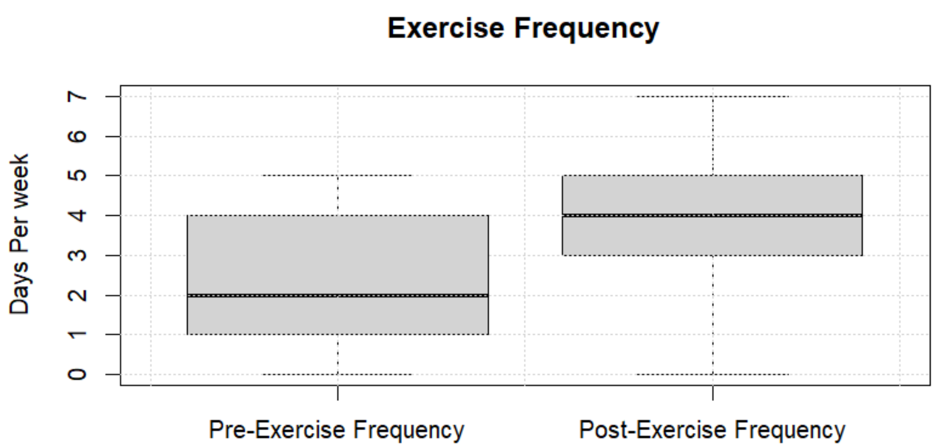
In the boxplot, we can observe a clear reduction in the median stress level after the program. This can be confirmed using Wilcoxon's sign rank test for match pairs. Here the hypotheses are,

H_0: The median of the distribution of differences between post-stress level and pre-stress level equal to zero

H_1: The median of the distribution of differences between post-stress level and pre-stress level is less than zero

Since the p-value is less than 0.05, we reject the null hypothesis. Hence, at 5% significance level, we can conclude that the median stress level after the program is less than before. Therefore we can say the wellness program reduced the stress level of employees.

3.2.2 Impact of Wellness Program on Physical Fitness



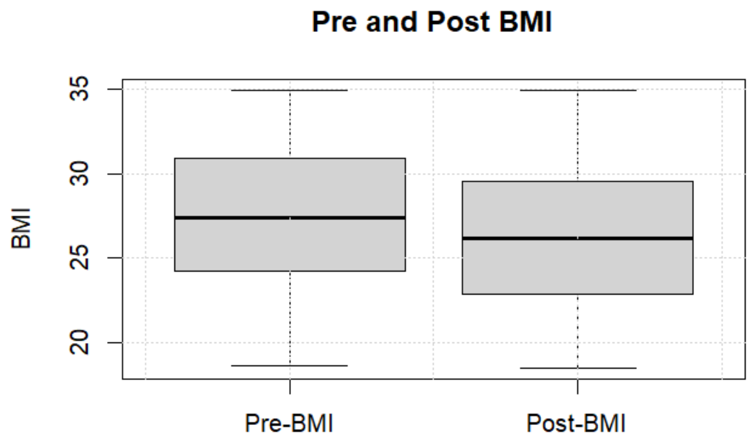
```
> wilcox.test(Post_Exercise_Frequency,Pre_Exercise_Frequency,alternative = 'greater',paired = T)

	Wilcoxon signed rank test with continuity correction

data:  Post_Exercise_Frequency and Pre_Exercise_Frequency
V = 11476, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

It can be observed that the median exercise frequency of employees after the program is higher than before. Here also we use Wilcoxon's sign rank test for matched pairs and its hypotheses as follows.

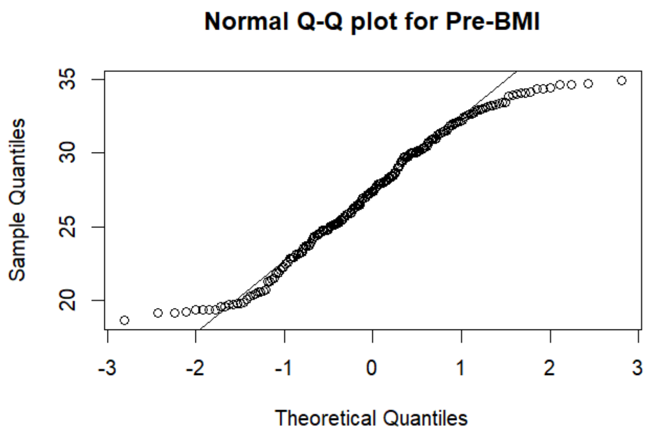
H_0: The median of the distribution of differences between post and pre-exercise frequencies equal to zero

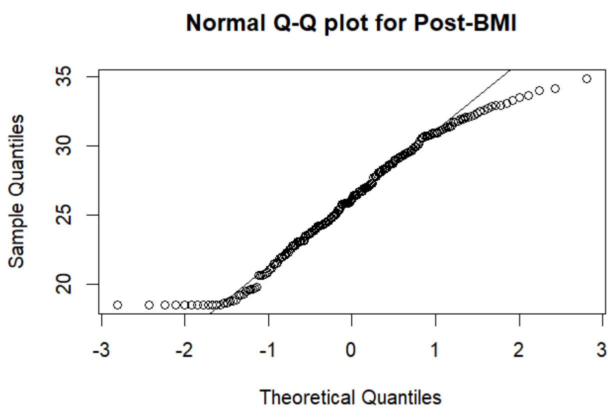
H_1: The median of the distribution of differences between post and pre-exercise frequencies is greater than zero

Since the p-value is less than 0.05, we reject the null hypothesis. Hence at 5% significance level, we can conclude that the exercise frequency of employees after the program is higher than before.



According to the boxplot visual, the BMI value of employees reduced after the program. For statistically test this result, we can consider using a parametric test since this is a continuous variable. Before that, we have to check for the normality assumption of this variable.





```
> shapiro.test(Pre_BMI)

	Shapiro-Wilk normality test

data:  Pre_BMI
W = 0.96531, p-value = 7.772e-05
```

```
> shapiro.test(Post_BMI)

	Shapiro-Wilk normality test

data:  Post_BMI
W = 0.96643, p-value = 0.0001048
```

Hypotheses of the Shapiro-Wilk test,

H_0: Data is normally distributed

H_1: Data is not normally distributed

Since the p-value for both pre and post BMI data is less than 0.05, we reject the null hypothesis. Hence the data is not normal. Therefore we have to use a non-parametric test. Since the pre and post-test data were measured from the same individual, as before we will use Wilcoxon's sign rank test for matched pairs.

H_0: The median of the distribution of differences between post and pre-BMI equal to zero

H_1: The median of the distribution of differences between post and pre-BMI is less than zero

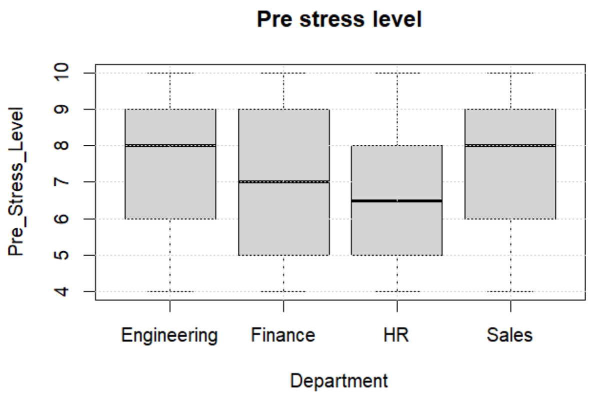
```
> wilcox.test(Post_BMI,Pre_BMI,alternative = 'less',paired = T)

	Wilcoxon signed rank test with continuity correction

data:  Post_BMI and Pre_BMI
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

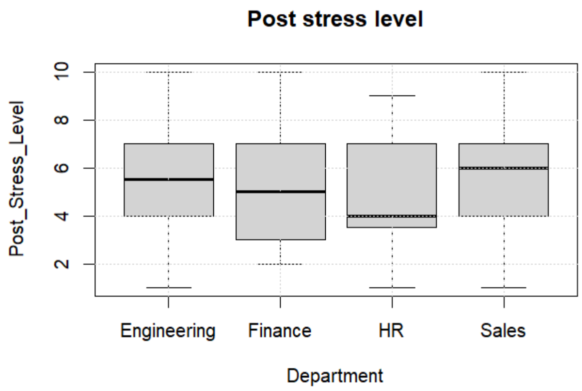
Since the p-value is less than 0.05, we reject the null hypothesis. Thereby we can say that the median BMI level of employees reduced after the program.

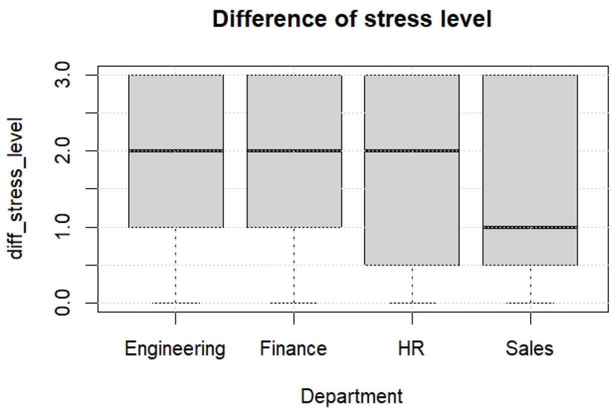
Given the statistical evidence showing that after the program, the exercise frequency of employees has increased and the BMI level of employees has reduced, we can conclude that the wellness program improves the physical fitness of employees.

3.2.3 Health Improvements Across Different Job Roles and Departments

Since we have confirmed that the wellness program improves employees' mental and physical health, we are now interested in finding out whether employees belonging to different departments experienced varying levels of health improvements.







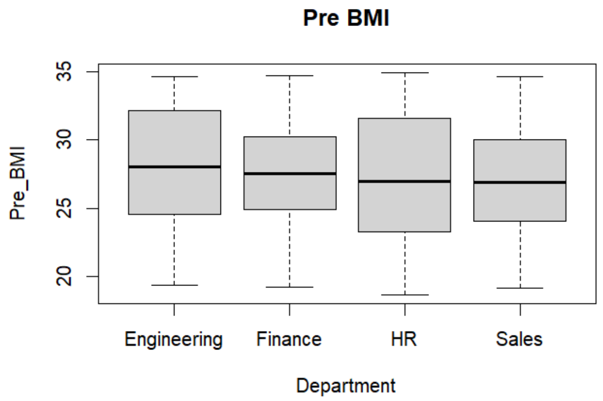

```
> kruskal.test(diff_stress_level~Department)

        Kruskal-Wallis rank sum test

data:  diff_stress_level by Department
Kruskal-Wallis chi-squared = 0.71936, df = 3, p-value = 0.8686
```

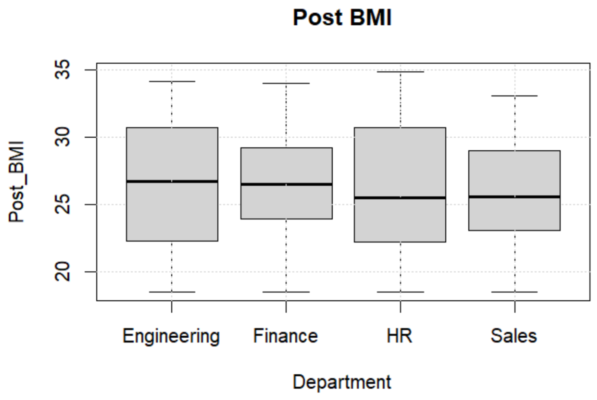
By observing pre and post department-wise stress levels we can see the median stress level of each department reduced after the program. And also we can see the range of stress levels has increased. The minimum stress level becomes 2. However, there is no significant difference in the median of the difference of pre and post-stress levels among the departments, as indicated by the third boxplot. To confirm that we can use a statistical test. Since stress level is a discrete variable we have to use a non-parametric test. As we have more than two groups to compare we can use Kruskal-Walli's test. The hypotheses are.

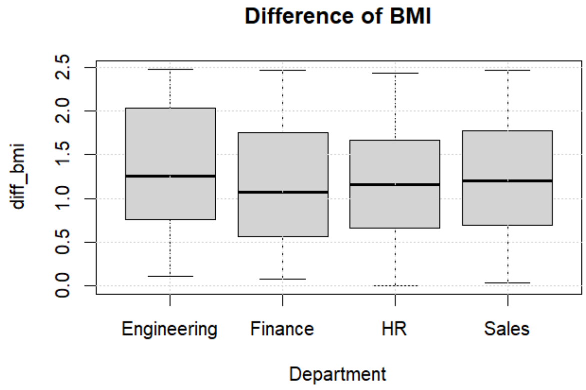
H0: There's no difference in location for stress level reduction between the departments

H1: There's a difference in location for stress level reduction in at least one or more departments

Since the p-value is greater than 0.05, we do not reject the null hypothesis. Therefore we do not have enough evidence to conclude that reduction in stress levels varies across the departments.







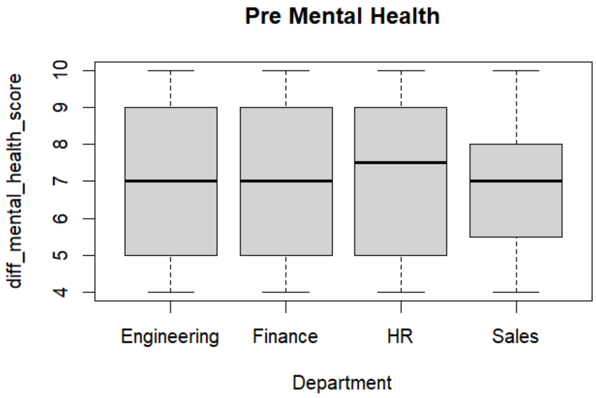

```
> kruskal.test(diff_bmi~Department)

        Kruskal-Wallis rank sum test

data:  diff_bmi by Department
Kruskal-Wallis chi-squared = 2.723, df = 3, p-value = 0.4363
```

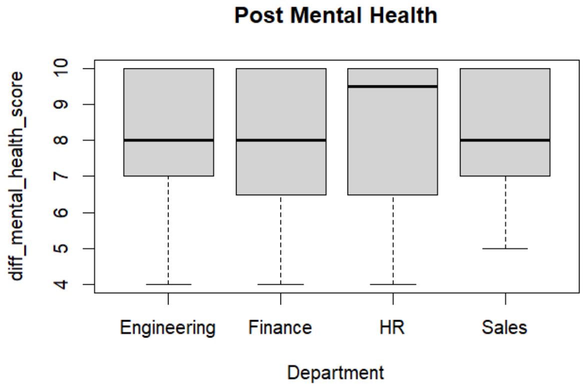
Here we can see a slight difference in the median of the difference of BMI levels before and after the program among the departments. We can use Kruskal-Walli's test to assess the significance of this.

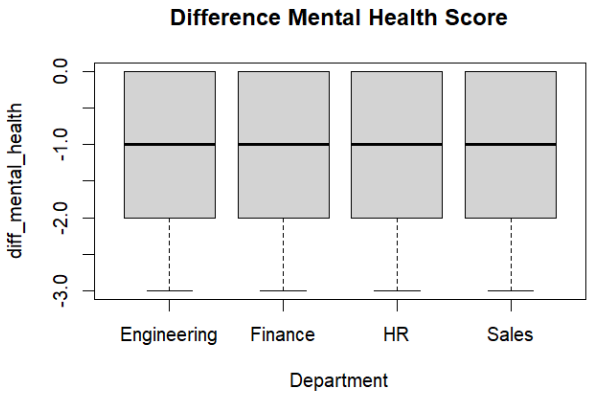
H0: There's no difference in location for BMI reduction between the departments

H1: There's a difference in location for BMI reduction in at least one or more departments

Since the p-value is greater than 0.05, we do not reject the null hypothesis. Even though we visually inspect slight differences in location for four departments, this is not statistically significant at the 5% level of significance.








```
> kruskal.test(diff_mental_health~Department)

        Kruskal-Wallis rank sum test

data:  diff_mental_health by Department
Kruskal-Wallis chi-squared = 4.939, df = 3, p-value = 0.1763
```

Here we can see from the third boxplot there's no difference in the median of the differences of mental health scores. We can confirm that from Kruskal-walli's test by not rejecting the null hypothesis.

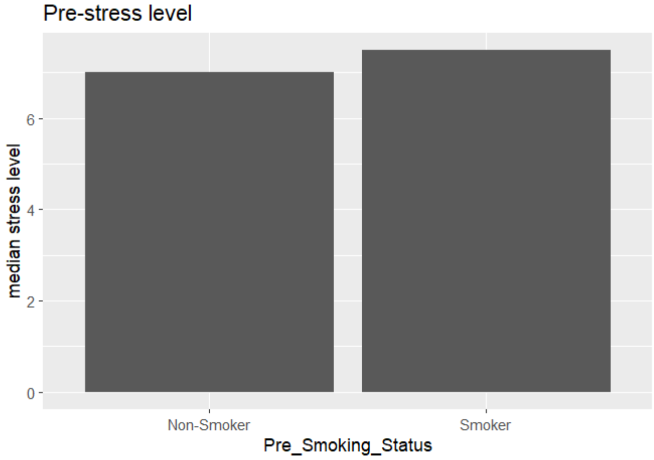
H0: There's no difference in location for mental health score difference between the departments

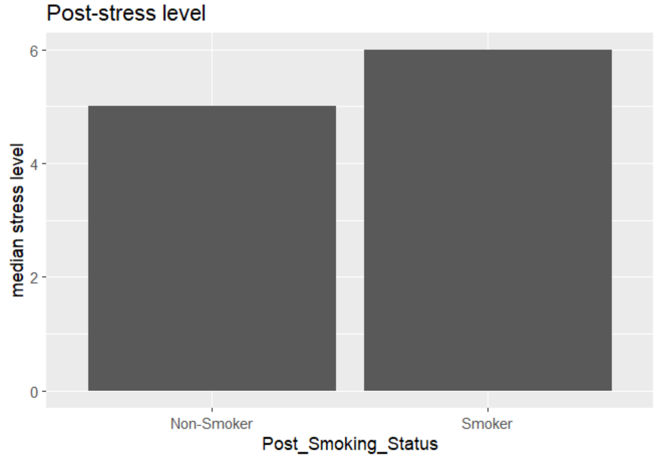
H1: There's a difference in location for health score difference in at least one or more departments

Since the p-value is greater than 0.05, we do not reject null hypothesis. Therefore we can conclude that there's no variation in mental health improvement across different departments.

3.2.4 Relationship between lifestyle factors and health outcomes

Smoking status with stress level






```
> wilcox.test(Post_Stress_Level~Post_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Post_Stress_Level by Post_Smoking_Status
W = 4461.5, p-value = 0.1867
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Pre_Stress_Level~Pre_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Pre_Stress_Level by Pre_Smoking_Status
W = 4669.5, p-value = 0.5625
alternative hypothesis: true location shift is not equal to 0
```

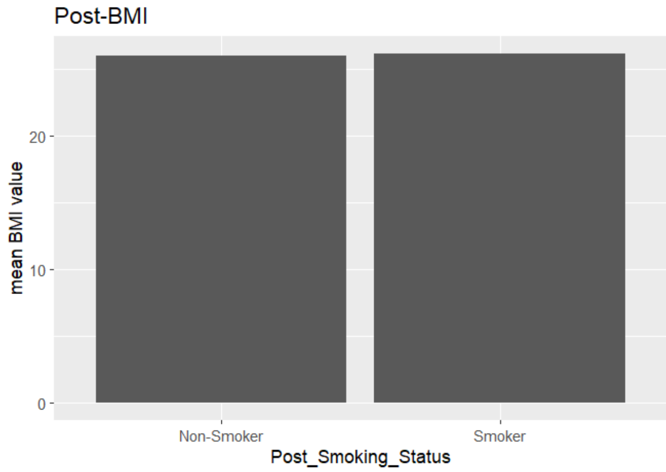
From the bar charts, we can see that the median stress level of non-smokers is less than that of smokers both before and after the program. However, the gap between the median stress level of non-smokers and smokers is greater after the program. To test that we can use wilcoxon's rank sum test.

H0: There's a difference in the location of stress levels between non-smokers and smokers

H1: There's no difference in the location of stress levels between non-smokers and smokers

Since the p-value is greater than 0.05 in both tests, we do not reject the null hypothesis. Even though we visually inspect a difference in stress levels between smokers and non-smokers, this is not statistically significant at the 5% significance level. It appears that smoking status did not significantly affect stress levels either before or after the program.

Smoking Status with BMI value




```
> wilcox.test(Pre_BMI~Pre_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Pre_BMI by Pre_Smoking_Status
W = 4721, p-value = 0.656
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Post_BMI~Post_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Post_BMI by Post_Smoking_Status
W = 4928, p-value = 0.8651
alternative hypothesis: true location shift is not equal to 0
```

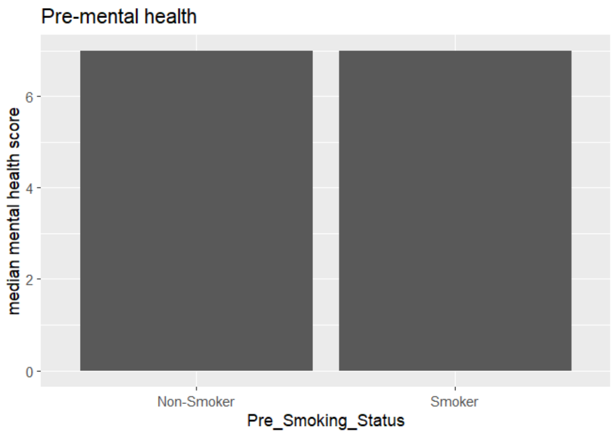
Here we can't see any difference in mean BMI value between smokers and non-smokers. The wilcoxon's rank sum test confirms that by not rejecting the null hypothesis.

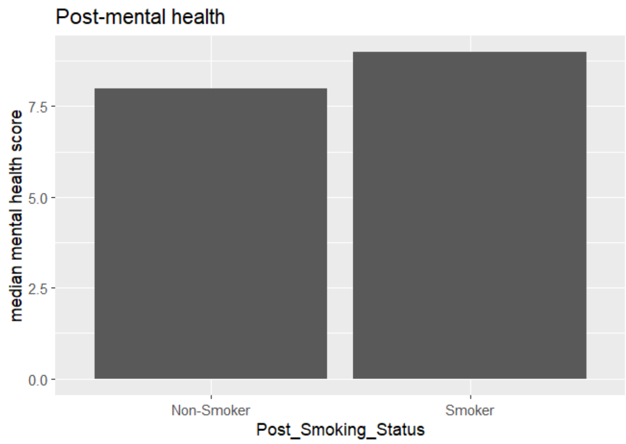
H0: There's a difference in the location of stress levels between non-smokers and smokers

H1: There's no difference in the location of stress levels between non-smokers and smokers

We do not reject the null hypothesis since the p-value is greater than 0.05 in both tests. Therefore we can conclude that smoking status did not affect the BMI value.

Smoking Status with Mental Health Score





```
> wilcox.test(Pre_Mental_Health_Score~Pre_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Pre_Mental_Health_Score by Pre_Smoking_Status
W = 4809, p-value = 0.8176
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Post_Mental_Health_Score~Post_Smoking_Status)

        Wilcoxon rank sum test with continuity correction

data:  Post_Mental_Health_Score by Post_Smoking_Status
W = 4822.5, p-value = 0.6588
alternative hypothesis: true location shift is not equal to 0
```

Here we can see there's no difference in median mental health score before the program between smokers and non-smokers. But we can see a slight improvement in the median mental health score among smokers compared to non-smokers after the program.

H0: There's a difference in the location of mental health scores between non-smokers and smokers

H1: There's no difference in the location of mental health scores between non-smokers and smokers

Since the p-values greater than 0.05, we do not reject the null hypothesis, suggesting that smoking status does not affect the mental health of employees.

We have to test the correlation between these variables to find out the relationship between exercise frequency and health outcomes. Since these are discrete variables we have to use a non-parametric test as normality assumptions are violated. Hence we will use the Spearman correlation coefficient to measure the relationship between these variables.

<u>Exercise frequency with stress level</u>

```
> cor.test(Pre_Exercise_Frequency,Pre_Stress_Level,method = 'kendall')

        Kendall's rank correlation tau

data:  Pre_Exercise_Frequency and Pre_Stress_Level
z = 0.16515, p-value = 0.8688
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau
0.009049043
```

```
> cor.test(Post_Exercise_Frequency,Post_Stress_Level,method = 'kendall')

        Kendall's rank correlation tau

data:  Post_Exercise_Frequency and Post_Stress_Level
z = 0.85704, p-value = 0.3914
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
0.04585555
```

H0: There’s no relationship between exercise frequency and stress level

H1: There’s a relationship between exercise frequency and stress level

Since the p-value is greater than 0.05 in both tests, we do not reject the null hypothesis. Therefore we can say that there’s no relationship between the exercise frequency of employees and their stress level.

Exercise frequency with BMI

```
> cor.test(Pre_Exercise_Frequency,Pre_BMI,method = 'kendall')

        Kendall's rank correlation tau

data:  Pre_Exercise_Frequency and Pre_BMI
z = -0.0085902, p-value = 0.9931
alternative hypothesis: true tau is not equal to 0
sample estimates:
          tau
-0.0004406369
```

```
> cor.test(Post_Exercise_Frequency,Post_BMI,method = 'kendall')

        Kendall's rank correlation tau

data:  Post_Exercise_Frequency and Post_BMI
z = -0.23126, p-value = 0.8171
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau
-0.01175004
```

H0: There's no relationship between exercise frequency and BMI level

H1: There’s a relationship between exercise frequency and BMI level

We do not reject the null hypothesis since the p-values of both tests are greater than 0.05. Hence we can conclude there’s no relationship between exercise frequency and BMI values of employees.

Exercise frequency and mental health score

```
> cor.test(Pre_Exercise_Frequency,Pre_BMI,method = 'kendall')

        Kendall's rank correlation tau

data:  Pre_Exercise_Frequency and Pre_BMI
z = -0.0085902, p-value = 0.9931
alternative hypothesis: true tau is not equal to 0
sample estimates:
          tau
-0.0004406369
```

```
> cor.test(Post_Exercise_Frequency,Post_BMI,method = 'kendall')

        Kendall's rank correlation tau

data:  Post_Exercise_Frequency and Post_BMI
z = -0.23126, p-value = 0.8171
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau
-0.01175004
```

H0: There's no relationship between exercise frequency and mental health score

H1: There's a relationship between exercise frequency and mental health score

Since the p-values of both tests exceed 0.05, we do not reject the null hypothesis. Therefore we conclude that there's no relationship between exercise frequency and mental health score of employees.

4. Conclusion

The analysis of the wellness program data provided insightful findings regarding its impact on employee health and lifestyle habits. The key observations are as follows:

- We can observe a significant reduction in the overall stress level of employees, indicating that the wellness program had a positive impact in reducing stress. This suggests that the wellness program was effective in improving the mental well-being of employees.

- There was also a significant increase in exercise frequency after the wellness program, reflecting a shift toward healthier physical activity habits. This improvement further supports the program's success in encouraging healthier lifestyle choices among employees.

- Despite overall health improvements, no significant differences were found in the level of health improvement across different job roles or departments. This suggests that the wellness program was equally beneficial to employees regardless of their specific job functions.

- The analysis also indicated no significant relationship between lifestyle factors, such as smoking status or exercise frequency, and key health outcomes (BMI, stress levels, and mental health). This implies that while the wellness program improved health indicators broadly, individual lifestyle factors did not have a statistically significant influence on these outcomes within the dataset.

Overall, the wellness program demonstrated significant benefits in reducing stress and increasing exercise frequency, contributing positively to the health and well-being of the workforce. However, these improvements were not significantly influenced by employees' job roles, departments, or specific lifestyle factors. This analysis provides a foundation for further refinements in the wellness program to target individual well-being and optimize program outcomes for different groups within the corporation.

5. Dataset

	Employee_ID	Age	Gender	Department	Job_Role	Pre_Stress_Level	Pre_Exercise_Frequency	Pre_Smoking_Status	Pre_BMI	Pre_Mental_Health_Score	Post_Stress_Level	Post_Exercise_Frequency	Post_Smoking_Status	Pc
1	1	42	Male	Engineering	Engineer	8	4	Smoker	33.40772	10	6	4	Smoker	
2	2	28	Male	Sales	Manager	10	1	Smoker	29.71861	7	7	4	Smoker	
3	3	50	Female	HR	HR Manager	10	2	Non-Smoker	34.00515	6	7	4	Non-Smoker	
4	4	27	Female	Finance	Financial Analyst	10	5	Smoker	29.75548	6	9	7	Smoker	
5	5	30	Male	Sales	Manager	7	1	Non-Smoker	30.71989	8	5	4	Non-Smoker	
6	6	56	Female	Sales	Sales Executives	7	3	Smoker	19.32084	6	7	4	Smoker	
7	7	63	Male	HR	HR Manager	9	0	Smoker	31.64211	9	9	2	Smoker	
8	8	53	Male	Finance	Accountant	9	4	Smoker	29.12218	4	9	6	Smoker	
9	9	65	Male	HR	HR Manager	6	0	Smoker	19.55055	9	3	0	Smoker	
10	10	28	Male	Engineering	Senior Engineer	6	5	Smoker	30.03713	6	4	5	Smoker	
11	11	34	Male	Sales	Sales Executives	10	3	Smoker	24.83005	5	10	5	Non-Smoker	
12	12	55	Male	Finance	Financial Analyst	4	0	Smoker	29.98438	8	4	1	Smoker	
13	13	61	Female	Engineering	Technician	9	5	Smoker	28.29589	7	8	5	Smoker	
14	14	38	Female	Engineering	Engineer	6	3	Smoker	30.27416	5	3	6	Smoker	
15	15	64	Female	Sales	Sales Executives	5	2	Smoker	30.90044	7	2	3	Non-Smoker	
16	16	56	Female	HR	Recruiter	6	0	Smoker	18.64640	4	4	0	Non-Smoker	
17	17	63	Male	Engineering	Senior Engineer	5	4	Non-Smoker	31.26895	7	4	5	Non-Smoker	
18	18	60	Male	Finance	Financial Analyst	9	2	Smoker	25.85913	5	9	4	Smoker	
19	19	57	Female	Engineering	Engineer	5	0	Non-Smoker	28.66735	5	3	2	Non-Smoker	
20	20	41	Male	HR	Recruiter	4	5	Non-Smoker	30.14618	4	1	5	Non-Smoker	
21	21	41	Male	Sales	Manager	9	1	Non-Smoker	23.74910	4	9	3	Non-Smoker	
22	22	43	Female	HR	Recruiter	10	0	Smoker	31.49895	5	7	1	Smoker	
23	23	38	Female	Engineering	Senior Engineer	5	5	Non-Smoker	22.08104	10	4	7	Non-Smoker	
24	24	22	Female	Engineering	Technician	7	5	Non-Smoker	23.54142	9	5	7	Non-Smoker	
25	25	42	Male	Engineering	Engineer	6	3	Smoker	23.09419	7	5	6	Smoker	
26	26	41	Male	Sales	Sales Executives	10	4	Smoker	26.19295	4	9	4	Smoker	

Link to the dataset :-

https://drive.google.com/file/d/1QBbbH9ICXv0SZXWD8WydkUt_xrAD8xmv/view?usp=sharing

6. R code

```
#---------------------------
# Load necessary libraries
library(dplyr)

# Set seed for reproducibility
set.seed(16341)

# Number of observations
n <- 200

# Generate the dataset
data <- data.frame(
  Employee_ID = 1:n,
  Age = sample(22:65, n, replace = TRUE),
```

```
  Gender = sample(c("Male", "Female"), n, replace = TRUE),
  Department = sample(c("Sales", "Engineering", "HR", "Finance"), n, replace = TRUE)
)

# Assign Job_Role based on Department
data <- data %>%
  mutate(Job_Role = case_when(
    Department == "Sales" ~ sample(c("Manager", "Sales Executives"), n, replace = TRUE),
    Department == "Engineering" ~ sample(c("Engineer", "Senior Engineer", "Technician"), n,
replace = TRUE),
    Department == "HR" ~ sample(c("HR Manager", "Recruiter"), n, replace = TRUE),
    Department == "Finance" ~ sample(c("Financial Analyst", "Accountant"), n, replace = TRUE)
  ))

# Generate pre-program variables
data <- data %>%
  mutate(
    Pre_Stress_Level = sample(4:10, n, replace = TRUE),
    Pre_Exercise_Frequency = sample(0:5, n, replace = TRUE),
    Pre_Smoking_Status = sample(c("Smoker", "Non-Smoker"), n, replace = TRUE),
    Pre_BMI = runif(n, 18.5, 35.0),
    Pre_Mental_Health_Score = sample(4:10, n, replace = TRUE)
  )

# Generate post-program variables
data <- data %>%
  mutate(
    Post_Stress_Level = pmax(1, Pre_Stress_Level - sample(0:3, n, replace = TRUE)),
    Post_Exercise_Frequency = pmin(7, Pre_Exercise_Frequency + sample(0:3, n, replace = TRUE)),
    Post_Smoking_Status = ifelse(Pre_Smoking_Status == "Smoker" & runif(n) < 0.1, "Non-
Smoker", Pre_Smoking_Status),
    Post_BMI = pmax(18.5, Pre_BMI - runif(n, 0, 2.5)),
    Post_Mental_Health_Score = pmin(10, Pre_Mental_Health_Score + sample(0:3, n, replace =
TRUE)))
```

```
write.csv(data,"D:\\UOC\\3rd year\\1st semester\\ST\\ST 3010\\case study\\wellness.csv")


#--------------------------------------------------------------------------------------------------------

wellness <- read.csv("D:/UOC/3rd year/1st semester/ST/ST 3010/case study/wellness.csv", stringsAsFactors=TRUE)

attach(wellness)


library(tidyverse)


summary(Age)

hist(Age,breaks = 15, probability = T)

lines(density(Age), col = 'blue', lwd = 2)


ggplot(data = wellness,aes(x = Age)) + geom_histogram(aes(y = ..density..),bins = 15,colour = 'white') +

  geom_density(colour = 'red', size = 1) + labs(title = 'Histogram of Age')



table(Gender)

barplot(table(Gender))

table(Department)

barplot(table(Department), ylab = 'count', main = 'Destribution of Departments')



boxplot(Pre_Stress_Level, ylab = 'stress level', main = 'Pre stress level', grid())

summary(Pre_Stress_Level)


barplot(table(Pre_Exercise_Frequency), xlab = 'days per week', ylab = 'number of employess', main = 'Pre Exercise frequency')

barplot(table(Pre_Smoking_Status))


ggplot(data = wellness, aes(x = Pre_Smoking_Status)) + geom_bar() + labs(title = 'Pre smoking status')

table(Pre_Smoking_Status)
```

```
boxplot(Pre_BMI, ylab = 'BMI', main = 'Pre BMI', grid())
summary(Pre_BMI)

boxplot(Pre_Mental_Health_Score, ylab = 'Mental health score', main = 'Pre Mental Health Score', grid())
summary(Pre_Mental_Health_Score)
hist(Pre_BMI)

mean_pre_bmi <- mean(Pre_BMI)
sd_pre_bmi <- sd

ggplot(data = wellness, aes(x = Pre_BMI)) + geom_histogram(aes(y = ..density..) ,bins = 15, colour = 'white') +
  labs(title = 'Histogram of Pre-BMI') + geom_density(colour = 'red', size = 1)

hist(Pre_BMI,breaks = 15, probability = T)
lines(density(Pre_BMI),col = 'blue', lwd = 2)

boxplot(Post_Stress_Level, ylab = 'stress level', main = 'Post stress level', grid())
summary(Post_Stress_Level)
barplot(table(Post_Stress_Level))

barplot(table(Post_Exercise_Frequency), xlab = 'days per week', ylab = 'number of employees', main = 'Post Exercise Frequency')
ggplot(data = wellness,aes(x = Post_Smoking_Status)) + geom_bar() + labs(title = 'Post smoking status')
table(Post_Smoking_Status)

ggplot(data = wellness, aes(x = Post_BMI)) + geom_histogram(aes(y = ..density..) ,bins = 15, colour = 'white') +
  labs(title = 'Histogram of Post-BMI') + geom_density(colour = 'red', size = 1)

boxplot(Post_BMI, ylab = 'BMI', main = 'Post BMI', grid())
summary(Post_BMI)
```

```
boxplot(Post_Mental_Health_Score, ylab = 'Mental Health Score', main = 'Post Mental Health Score')

summary(Post_Mental_Health_Score)


boxplot(Pre_Stress_Level, Post_Stress_Level,names = c('pre stress level','post stress level') , xlab = 'stress levels', ylab = 'count', main = 'Pre and Post program stress levels', grid())

library(BSDA)

wilcox.test(Post_Stress_Level,Pre_Stress_Level,mu = 0,alternative = 'less',paired = T,conf.level = 0.95)


#----------------------


boxplot(Pre_Stress_Level~Department)

boxplot(Post_Stress_Level~Department)

kruskal.test(Pre_Stress_Level~Department)


boxplot(Pre_Exercise_Frequency,Post_Exercise_Frequency, names = c('Pre-Exercise Frequency','Post-Exercise Frequency'), ylab = 'Days Per week', main = 'Exercise Frequency', grid())

wilcox.test(Post_Exercise_Frequency,Pre_Exercise_Frequency,alternative = 'greater',paired = T)


boxplot(Pre_BMI,Post_BMI, names = c('Pre-BMI','Post-BMI'), ylab = 'BMI', main = 'Pre and Post BMI', grid())

wilcox.test(Post_BMI,Pre_BMI,alternative = 'less',paired = T)


qqnorm(Pre_BMI,main = 'Normal Q-Q plot for Pre-BMI')

qqline(Pre_BMI)

shapiro.test(Pre_BMI)


qqnorm(Post_BMI, main = 'Normal Q-Q plot for Post-BMI')

qqline(Post_BMI)

shapiro.test(Post_BMI)

boxplot(Pre_Stress_Level~Department, main = 'Pre stress level', grid())

boxplot(Post_Stress_Level~Department, main = 'Post stress level',grid())
```

```
diff_stress_level <- Pre_Stress_Level - Post_Stress_Level
boxplot(diff_stress_level~Department, main = 'Difference of stress level',grid())
kruskal.test(diff_stress_level~Department)

boxplot(Pre_BMI~Department, main = 'Pre BMI')
boxplot(Post_BMI~Department, main = 'Post BMI')
diff_bmi <- Pre_BMI - Post_BMI
boxplot(diff_bmi~Department, main = 'Difference of BMI')
kruskal.test(diff_bmi~Department)

boxplot(Pre_Mental_Health_Score~Department, main = 'Pre Mental Health', ylab =
'diff_mental_health_score')
boxplot(Post_Mental_Health_Score~Department, main = 'Post Mental Health', ylab =
'diff_mental_health_score')
diff_mental_health <- Pre_Mental_Health_Score - Post_Mental_Health_Score
boxplot(diff_mental_health~Department, main = 'Difference Mental Health Score')

kruskal.test(diff_mental_health~Department)

cor.test(Post_Exercise_Frequency,Post_BMI,alternative = 'two.sided',method = 'kendall')

ggplot(data = wellness,aes(x = Pre_Smoking_Status,y = Pre_Stress_Level)) +
  geom_bar(stat = 'summary', fun ='median') + labs(title = 'Pre-stress level', y = 'median stress level')

ggplot(data = wellness,aes(x = Post_Smoking_Status,y = Post_Stress_Level)) +
  geom_bar(stat = 'summary', fun ='median') + labs(title = 'Post-stress level', y = 'median stress level')

wilcox.test(Post_Stress_Level~Post_Smoking_Status)
wilcox.test(Pre_Stress_Level~Pre_Smoking_Status)

ggplot(data = wellness,aes(x = Pre_Smoking_Status,y = Pre_BMI)) +
  geom_bar(stat = 'summary', fun ='mean') + labs(title = 'Pre-BMI', y = 'mean BMI value')

ggplot(data = wellness,aes(x = Post_Smoking_Status,y = Post_BMI)) +
```

```
  geom_bar(stat = 'summary', fun ='mean') + labs(title = 'Post-BMI', y = 'mean BMI value')

wilcox.test(Pre_BMI~Pre_Smoking_Status)
wilcox.test(Post_BMI~Post_Smoking_Status)

ggplot(data = wellness,aes(x = Pre_Smoking_Status,y = Pre_Mental_Health_Score)) +
  geom_bar(stat = 'summary', fun ='median') + labs(title = 'Pre-mental health', y = 'median mental health score')

ggplot(data = wellness,aes(x = Post_Smoking_Status,y = Post_Mental_Health_Score)) +
  geom_bar(stat = 'summary', fun ='median') + labs(title = 'Post-mental health', y = 'median mental health score')

wilcox.test(Pre_Mental_Health_Score~Pre_Smoking_Status)
wilcox.test(Post_Mental_Health_Score~Post_Smoking_Status)

cor.test(Post_Exercise_Frequency,Post_BMI, method = 'spearman')
cor.test(Post_Exercise_Frequency,Post_Stress_Level, method = 'spearman')
cor.test(Post_Exercise_Frequency,Post_Mental_Health_Score, method = 'spearman')

prop.test(table(Pre_Smoking_Status,Post_Smoking_Status),alternative = 'greater')
chisq.test(Pre_Smoking_Status,Post_Smoking_Status,correct = T)

cor.test(Pre_Exercise_Frequency,Pre_Stress_Level,method = 'kendall')
cor.test(Post_Exercise_Frequency,Post_Stress_Level,method = 'kendall')

cor.test(Pre_Exercise_Frequency,Pre_BMI,method = 'kendall')
cor.test(Post_Exercise_Frequency,Post_BMI,method = 'kendall')

cor.test(Pre_Exercise_Frequency,Pre_BMI,method = 'kendall')
cor.test(Post_Exercise_Frequency,Post_BMI,method = 'kendall')
```