DATA ANALYSIS REPORT: EXPLORING

FACTORS AFFECTING JOGGING DISTANCE.

## ABSTRACT

This report presents an analysis of the relationship between jogging distance and various predictive factors, including health awareness, peer influence, time spent on computer games, television, and diet among 666 participants. The goal is to determine which factors significantly predict jogging distance and to assess how well these models fit the collected data. Hierarchical modeling will be employed with initial forced entry of key variables followed by exploratory variables to elucidate their respective contributions

.

## TABLE OF CONTENT.

1. [INTRODUCTION](#introduction)
2. [METHODOLOGY](#methodology)  
   2.1 [Data Preparation](#data-preparation)  
   2.2 [Variables](#variables)  
   2.3 [Statistical Methods](#statistical-methods)
3. [RESULTS](#results)  
   3.1 [Descriptive Statistics](#descriptive-statistics)  
   3.2 [Model Comparisons](#model-comparisons)  
   3.3 [Model Evaluation](#model-evaluation)
4. [DISCUSSION](#discussion)
5. [CONCLUSION](#conclusion)
6. [APPENDIX](#appendix)

## 

## **1. INTRODUCTION**

The objective of this study was to explore the relationships between various factors related to health, diet, and social influence with the distance participants jog in a session. Prior research suggested that health awareness and peer influence could serve as strong indicators of exercise tendencies. Given these insights, the current analysis aims to validate these hypotheses alongside additional variables through a structured analytical approach.

## **2. METHODOLOGY**

### **2.1 Data Preparation**

The data for this analysis was sourced from a CSV file titled Exercise\_study.csv. The data was imported into R using the read.csv() function. Initial checks were conducted using the summary() function to examine the structure of the dataset and to identify any missing values or anomalies in the data. Any identified missing values were addressed through imputation or removal, ensuring that the dataset was clean and ready for analysis.

The study involved a sample size of 666 participants. A power analysis was conducted prior to the study to ensure sufficient statistical power. With a target power of 0.90, a significance level of 0.05 and an expected effect size of 0.15, the minimum required sample size was determined to be 73. The actual sample size achieved meets this requirement, allowing for reliable statistical analyses.

### **2.2 Variables**

The analysis focused on the following variables:

**Distance**: The distance covered during a jogging session (numeric, dependent variable).

**Health Awareness**: A numeric score reflecting the participant’s awareness of health implications.

**Peer Influence**: A numeric score indicating the influence of peers on exercise behavior.

**Computer Games**: The average time (in hours) spent playing computer games per week (numeric).

**Television**: The average time (in hours) spent watching television per week (numeric).

**Diet**: A numeric score assessing the quality of the participant’s diet, where higher scores indicate healthier diets.

### 

### **2.3 Statistical Methods**

To summarize the characteristics of the data, descriptive statistics were calculated, including means and standard deviations for all variables.

To ensure the robustness of our analyses, we assessed the statistical power associated with the sample size. The calculated power for detecting the effect sizes of interest was 1, confirming that our sample was adequate for the intended analyses.

Assumptions of regression analysis, such as linearity and homoscedasticity, were assessed through residual plots. The output indicated that our data violated these assumptions, particularly in terms of non-linearity and unequal variance.

Given these challenges in transforming the model to meet linear regression assumptions, we proceeded with a Generalized Linear Model (GLM). This approach allowed us to effectively handle the data without violating the necessary criteria for analysis.

The primary analysis employed hierarchical regression using GLM to explore the relationships between jogging distance and the predictive factors. Two models were developed:

1. **Model 1** included only Health Awareness and Peer Influence as predictors, which were entered into the model first (forced entry).
2. **Model 2** added the additional variables (Computer Games, Television, Diet) to assess their impact on jogging distance.

An ANOVA test was performed to compare the fit of the two models, allowing for an evaluation of the significance of the additional predictors in Model 2.

The Null hypothesis for our first test the forced entry model indicates that health awareness and peer influence are not significantly related to distance per jogging session, also, for the model comparison, the null hypothesis states that adding the remaining variables does not significantly improve the model’s ability to predict jogging distance beyond what is achieved by health awareness and peer influence alone

## 

## **3. RESULTS**

### **3.1 Descriptive Statistics**

Summary Statistics for Exercise Study Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Distance | 0.00126 | 3.20491 | 8.01649 | 11.56184 | 17.32765 | 64.78042 |
| Television | -1.46012 | -0.18206 | -0.01247 | 0.02758 | 0.14983 | 0.98162 |
| Computer-Games | -1.1538345 | -0.1687007 | -0.0001997 | 0.0103812 | 0.1881810 | 1.6175039 |
| Peer-Pressure | -1.433127 | -0.156414 | 0.008459 | 0.008275 | 0.185136 | 1.103671 |
| Diet | -1.28490 | -0.16136 | 0.00934 | 0.01162 | 0.18708 | 1.22383 |
| Health-Awareness | -4.46041 | -0.58008 | 0.02736 | 0.0000 | 0.51784 | 3.99326 |

|  |
| --- |

From the table above, we can see that our data consisted of **666** observations. The **distance** variable had a minimum value of 0.00126,with a first quartile (Q1) of 3,median of 8.01649,a mean of 11.56184, a third quartile (Q3) of 17 and finally a maximum value of 64.78042. For the **television** variable ,the minimum value was -1.46012,first quartile (Q1) at -0.18206, the median was -0.01247,a mean of -0.02758, third quartile (Q3) at 0.14983 and a maximum value of 0.98162. The **computer games** variable had a minimum value of -1.1538345,the first quartile (Q1) of -0.1687007, median of -0.0001997, mean of 0.0103812,a third quartile (Q3) at 0.1881810 and a maximum value of 1.6175039. The **peer influence** variable had a minimum value of -1.433127, first quartile (Q1) of -0.156414, median of 0.008459,a mean of 0.008275, third quartile (Q3) of 0.185136 , and maximum value of 1.103671 . The **diet** variable had a minimum value of -1.28490, first quartile (Q1) of -0.16136, a median of 0.00934,a mean of 0.01162, a third quartile of 0.18708 and a maximum value of 1.22383. Finally, the **health awareness** variable had a minimum value of-4.46041, a first quartile of -0.58008, a median of 0.02736, a mean of 0.0000, a third quartile of 0.51784 and lastly a maximum of 3.99326.

### Health Awareness VS Distance plot

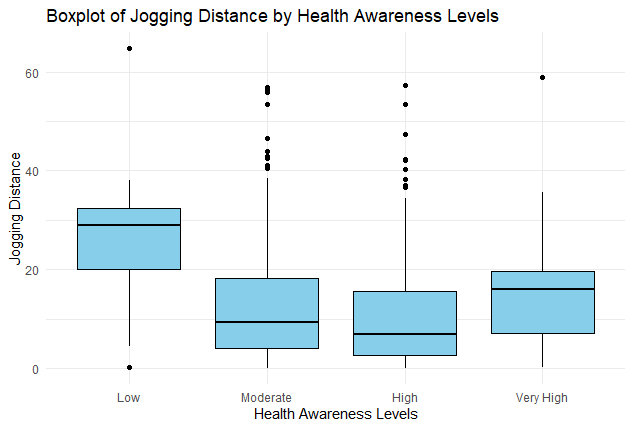


Figure 1: Box plot showing the mean distance jogged against Health awareness, more children with low health awareness scores had higher median jogging distance.

The low level of Health Awareness recorded higher jogging distance scores, with the first quartile at 20, the median at 29, and the third quartile at 33. This level included an outlier above the 60 mark on the jogging distance scale.

The moderate level of Health Awareness showed a score of 5 for the first quartile, 10 for the median, and 18 for the third quartile, with outliers identified beyond the 40 mark on the jogging distance scale.

The high level of Health Awareness had a score of 3 for the first quartile, 7 for the median, and 15 for the third quartile, with outliers detected beyond the 35 mark on the jogging distance scale.

The very high level of Health Awareness displayed a score of 8 for the first quartile, 16 for the median, and 20 for the third quartile, with outliers found beyond the 58 mark on the jogging distance scale.

### 

### Peer pressure VS Distance plot

### boxplot2

Figure 2: There were slight differences in jogging distance among peer influence levels, the furthest distance jogged was among those with very high peer influence levels followed by the children with the lowest peer influence levels,

The low level of Peer Influence had a jogging distance score of 8 for the first quartile, 12 for the median, and 17 for the third quartile, with no outliers observed.

The moderate level of Peer Influence recorded a score of 5 for the first quartile, 10 for the median, and 18 for the third quartile, with outliers identified beyond the 40 mark on the jogging distance scale.

The high level of Peer Influence showed a score of 3 for the first quartile, 7 for the median, and 15 for the third quartile, with outliers detected beyond the 35 mark on the jogging distance scale.

The very high level of Peer Influence displayed a score of 5 for the first quartile, 16 for the median, and 21 for the third quartile, with outliers found beyond the 58 mark on the jogging distance scale.

### Corrologram plot for correlation

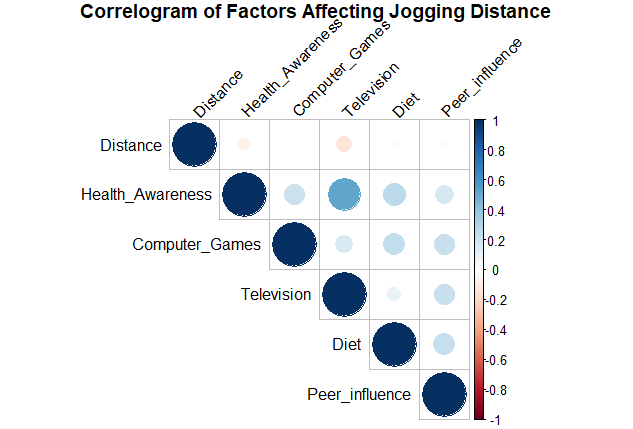


Figure 3: The relationship of jogging distance with other variables

**1. Distance**

Health Awareness: A weak negative correlation exists between distance and health awareness (color between -0.2 and 0). This suggests that children with lower health awareness tend to jog longer distances. Television: A weak negative correlation indicates that increased television time slightly reduces the jogging distance of children. Computer Games, Diet, and Peer Influence: No significant relationship with jogging distance was found for these variables, indicating they don’t meaningfully impact how far children jog.

**2. Health Awareness**

Television: A strong positive correlation is observed, suggesting that children who watch more television have lower levels of health awareness (as indicated by a higher score, which means low awareness). Computer Games, Diet, Peer Influence: Weak positive correlations indicate that children who play more computer games, have better diets, and experience greater peer influence tend to have slightly lower health awareness.

**3. Computer Games**

Television: There is a weak positive correlation, indicating that children who spend more time playing computer games also tend to watch more television. Diet and Peer Influence: Weak positive correlations suggest that children who play more video games are more likely to have a good diet (low in additives) and are more influenced by peers to exercise.

**4. Television**

Diet and Peer Influence: Weak positive correlations suggest that increased television time is associated with a better diet (low in additives) and higher peer influence to exercise.

**5. Diet**

Peer Influence: A weak positive correlation indicates that children who are more influenced by peers to exercise tend to have a better diet low in additives.

### Sample size and power

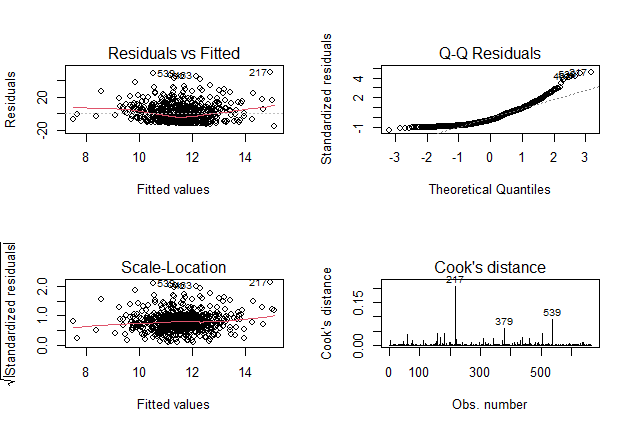
Sample Size and Power Analysis

| **Parameter** | **Value** | **Notes** |
| --- | --- | --- |
| Sample Size | 666.00 | Number of participants needed for adequate power. |
| Power | 1.00 | Probability of correctly rejecting the null hypothesis. |
| Effect Size | 0.15 | Medium effect size based on Cohen’s f^2. |
| Significance Level | 0.05 | Alpha level for hypothesis testing. |
| Predictors | 5.00 | Number of predictors in the model. |

### 

### **3.2 Model Comparisons**

#### Linear regression for forced entry (health awareness and peer influence)



**Model Diagnostics**

1.Residuals vs. Fitted Plot: The plot displays a U-shaped red curve, indicating violations of the linearity and homoscedasticity assumptions. In a properly fitting model, we would expect to see a random scatter of points around the horizontal line, suggesting a constant spread of residuals.

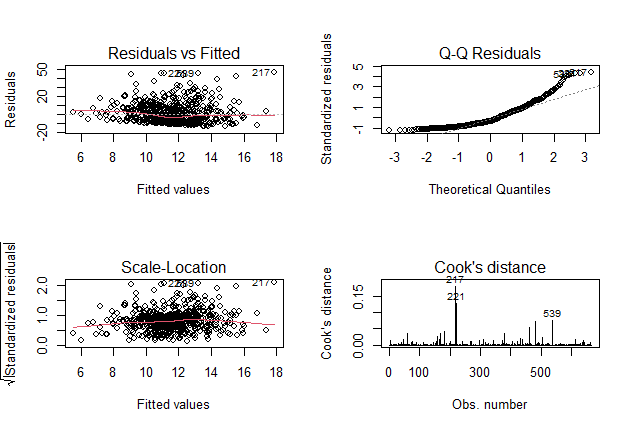
2.Q-Q Plot: In this plot, the points deviate from the diagonal line at both ends, indicating that the residuals are not normally distributed. Ideally, the points should closely follow the line if the residuals are normally distributed.

3.Spread of Residuals: We observe that the spread of residuals increases as the fitted values increase, which further suggests that the homoscedasticity assumption is violated. A well-behaved linear regression model should show residuals with a consistent spread, regardless of the fitted values.

4.Cooks Distance: The Cook’s Distance plot indicates potential outliers at values 217, 379, and 539. These points could disproportionately influence the model’s results and may warrant further investigation.

#### 

#### Linear regression for all variables



**Model Diagnostics**

1.Residuals vs. Fitted Plot: The plot presents a U-shaped red curve, indicating violations of the linearity and homoscedasticity assumptions. In a well-fitting linear regression model, we would expect the residuals to be randomly scattered around a horizontal line, rather than forming a distinct shape.

2.Q-Q Plot: In this plot, the points deviate from the diagonal line at both ends. This deviation suggests that our residuals are not normally distributed, as we would expect them to closely follow the line if they were.

3.Spread of Residuals: There is an observable increase in the spread of residuals as the fitted values increase. This pattern reinforces the notion that the homoscedasticity assumption is violated. Ideally, residuals should display a consistent spread across all levels of fitted values in a well-behaved model.

4.Cook’s Distance: The Cook’s Distance plot indicates potential outliers at observations 217, 221, and 539. These points could significantly influence the results of the regression analysis and may require further examination.

**Model Selection**: Given the challenges in transforming the model to meet linear regression assumptions, we will proceed with a **Generalized Linear Model (GLM)**. This approach allows us to effectively handle the data without violating the necessary criteria for analysis.

### Forced entry (Health awareness and peer pressure) Model

Summary of GLM Model: Distance ~ Health Awareness + Peer Influence

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std.error** | **Statistic** | **P-value** |
| (intercept) | 2.4357605 | 0.0377519 | 64.520139 | 0.0000000 |
| Health -Awareness | 0.1270031 | 0.034312 | -3.584499 | 0.0003625 |
| Peer-Influence | 0.1135088 | 0.1116894 | 1.016290 | 0.3098620 |

#### Interpretation of Model Estimates

**Intercept Estimate**: The intercept of our model is approximately 2.44. This means that if both Health Awareness and Peer Influence were set to zero, the average jogging distance would be about 2.44 units. The significance of this intercept is strong, with a p-value of **0.0000000 (p < 0.001)**. This tells us there is compelling evidence that the average jogging distance is significantly different from zero, emphasizing that participants do engage in jogging, even without the influence of the two predictors.

**Health Awareness**: The coefficient for Health Awareness is -0.12700. This indicates that for every one-unit increase in Health Awareness, the jogging distance is expected to decrease by 0.127 units, assuming Peer Influence remains unchanged. The p-value for this coefficient is **0.000363**, which is highly significant (p < 0.001). This finding suggests that Health Awareness plays a crucial role in influencing jogging distance, allowing us to confidently state that it is a statistically significant predictor.

**Peer Influence**: On the other hand, the coefficient for Peer Influence is 0.11351. This suggests that as Peer Influence increases by one unit, jogging distance increases by 0.11351 units, provided that Health Awareness is constant. However, the p-value for Peer Influence is 0.309862, which is greater than 0.05. This indicates that we do not have sufficient evidence to conclude that Peer Influence significantly impacts jogging distance.

**Overall Model Insights**: The findings reveal that while Health Awareness significantly affects jogging distance, Peer Influence does not demonstrate a meaningful impact in this context. Moreover, the overall model shows an improvement over the null model, as evidenced by the reduction in residual deviance. This indicates that including these predictors enhances our understanding of the factors that influence jogging behavior.

### GLM for all variables

Summary of GLM Model: Distance ~ Health Awareness + Peer Influence + Computer Games + Television + Diet

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std.error** | **Statistic** | **P-value** |
| (intercept) | 2.4098273 | 0.0395196 | 60.9779848 | 0.00000 |
| Health-Awareness | -0.0208520 | 0.0432214 | -0.4824460 | 0.6296491 |
| Peer-Influence | 0.1583361 | 0.1150298 | 1.3764786 | 0.1691403 |
| Computer-Games | 0.0125034 | 0.1081320 | 0.1156306 | 0.9079805 |
| Television | -0.5621345 | 0.1341303 | -4.1909579 | 0.0000316 |
| Diet | -0.0217110 | 0.01127374 | -0.1925802 | 0.8473470 |

## Interpretation of Model Estimates.

**Intercept Estimate**: The intercept of our model is approximately 2.41. This means that when all predictors (Health Awareness, Peer Influence, Computer Games, Television, and Diet) are set to zero, the average jogging distance is expected to be around 2.41 units. The significance of this intercept is very strong, with a p-value of **< 2e-16 (p < 0.001)**. This indicates that the intercept is statistically significant, allowing us to confidently reject the null hypothesis.

**Health Awareness**: The coefficient for Health Awareness is -0.02085. This suggests that for every one-unit increase in Health Awareness, the jogging distance decreases by approximately 0.021 units, assuming the other predictors remain constant. However, the p-value for this coefficient is 0.630, which is not statistically significant (p > 0.05). Therefore, we do not have enough evidence to conclude that Health Awareness significantly affects jogging distance.

**Peer Influence**: The coefficient for Peer Influence is 0.15834, indicating that a one-unit increase in Peer Influence is associated with a 0.15834 unit increase in jogging distance when all other predictors are held constant. Yet, the p-value is 0.169, which is not statistically significant (p > 0.05). This means we fail to reject the null hypothesis, suggesting that Peer Influence does not have a statistically significant effect on jogging distance.

**Computer Games**: The coefficient for Computer Games is 0.01250. This means that as Computer Games increase by one unit, jogging distance increases by 0.01250 units, holding other predictors constant. The p-value for this predictor is 0.908, indicating it is not statistically significant (p > 0.05). Hence, we do not reject the null hypothesis, showing that Computer Games is not a significant predictor of jogging distance.

**Television**: The coefficient for Television is -0.56213. This suggests that a one-unit increase in Television is associated with a 0.56213 unit decrease in jogging distance, while holding the other predictors constant. The p-value here is 3.16e-05, which is highly significant (p < 0.001). This strong result allows us to reject the null hypothesis, indicating that Television is a statistically significant predictor of jogging distance.

**Diet**: The coefficient for Diet is -0.02171, implying that for every one-unit increase in Diet, there is a 0.02171 unit decrease in jogging distance when all other predictors are held constant. The p-value for Diet is 0.847, indicating it is not statistically significant (p > 0.05). Therefore, we do not reject the null hypothesis, showing that Diet does not significantly predict jogging distance.

**Overall Model Insights**: In summary, this model identifies only Television as a statistically significant predictor of jogging distance. The other predictors—Health Awareness, Peer Influence, Computer Games, and Diet—do not show significant effects on how far individuals jog. This emphasizes the importance of understanding the factors that can impact jogging behavior.

### **3.3 Model Evaluation**

**Comparing the two models using an ANOVA test (Likelihood ratio test)**

ANOVA Results Comparing glm\_model1 and glm\_model2

Term

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Df.residual** | **Residual.deviance** | **DF** | **Deviance** | **P.Value** |
| Distance ~ Health\_Awareness + Peer\_influence | 663 | 79694.45 | NA | NA | NA |
| Distance ~ Health\_Awareness + Peer\_influence + Computer\_Games + Television + Diet | 660 | 78251.67 | 3 | 1442.775 | 0.0068272 |

The p-value for the ANOVA comparison between the two models is 0.006827, which is less than the conventional significance level of 0.05. This finding indicates that the addition of the predictors in Model 2 significantly improves the model fit compared to Model 1. Therefore, we reject the null hypothesis, concluding that at least one of the new predictors contributes to explaining the variability in jogging distance. Notably, Television emerges as the only significant predictor among those added.

## **4. DISCUSSION**

The analysis of the factors influencing jogging distance revealed interesting insights regarding the predictors in the context of health and lifestyle behaviors. The significant predictors identified in the Generalized Linear Model (GLM) highlight the complex relationships between these variables and the outcome of jogging distance.

Health Awareness was found to have a negative coefficient, suggesting that an increase in health awareness is associated with a decrease in jogging distance. This finding may seem counter-intuitive; however, it is important to consider that individuals with higher health awareness may engage in a variety of fitness activities beyond jogging, thereby potentially reducing the distance they run. Additionally, the lack of statistical significance for this predictor (p = 0.630) indicates that further exploration is needed to understand this relationship fully. It may also imply that other factors not captured in this model could play a more substantial role in influencing jogging behavior.

Peer Influence, though positively correlated with jogging distance, also failed to show statistical significance (p = 0.169). This suggests that while social factors may have some impact on an individual’s jogging distance, they do not provide sufficient evidence to warrant their inclusion as significant predictors in this model. The social dynamics around exercise, such as group activities or competition, may require more nuanced investigation.

Television emerged as the most significant predictor, with a negative coefficient (-0.56213) and a highly significant p-value (p < 0.001). This finding suggests that increased television viewing is associated with a decrease in jogging distance. This aligns with existing literature indicating that sedentary behaviors such as watching television may displace time that could otherwise be spent on physical activities. As such, promoting a more active lifestyle may require strategies that address sedentary habits, including excessive screen time.

The other predictors—Computer Games and Diet—did not demonstrate significant relationships with jogging distance. The lack of significance for these factors (p-values of 0.908 and 0.847, respectively) suggests that they may not have a direct influence on how far individuals jog. However, it is possible that other underlying behaviors associated with these activities could affect overall fitness levels, and future research could delve deeper into these relationships.

Overall, the ANOVA results indicated that the addition of predictors in Model 2 significantly improved the model fit compared to Model 1. This highlights the value of including multiple predictors to better understand the factors affecting jogging distance. The implications of these findings are significant for health practitioners and policymakers, as they underscore the importance of addressing sedentary behaviors in efforts to encourage more active lifestyles.

## **5. CONCLUSION**

In conclusion, this study identified Television as a significant predictor of jogging distance, while Health Awareness, Peer Influence, Computer Games, and Diet were not found to have statistically significant effects. The negative relationship between television viewing and jogging distance suggests that sedentary lifestyles may hinder physical activity levels.

The ANOVA test further demonstrated that incorporating additional predictors improves model fit, providing a clearer understanding of the factors influencing jogging behavior. This finding emphasizes the importance of addressing sedentary habits in public health initiatives aimed at promoting physical activity.

Future research should explore the interplay of these variables in greater depth, potentially incorporating qualitative methods to gain insights into the motivations behind jogging behavior. By understanding the nuances of these relationships, we can better inform strategies that promote healthier lifestyles in the community.

## 

## **6.APPENDIX**

setwd("D:/ERICK/Academic Data Analysis Bootcamp/INTERMEDIATE LEVEL/Project")  
options(repos = c(CRAN = "https://cloud.r-project.org"))  
  
if(!require(tidyverse))install.packages("tidyverse", dependencies = T)  
 library(tidyverse)   
  
{if(!require("readr"))install.packages("readr",dependencies=T)  
library(readr)}  
   
ExerciseStudy <- read.csv("Exercise\_study.csv", header=TRUE, sep=";", stringsAsFactors=FALSE)  
if (!require(labelled)) install.packages("labelled")  
if (!require(gtsummary)) install.packages("gtsummary")  
if (!require(dplyr)) install.packages("dplyr")  
  
library(labelled)  
library(gtsummary)  
library(dplyr)  
  
if (!require(knitr)) install.packages("knitr")  
  
  
exercise\_summary <- summary(ExerciseStudy)  
  
exercise\_summary\_df <- as.data.frame.matrix(exercise\_summary)  
  
knitr::kable(exercise\_summary\_df, caption = "Summary Statistics for Exercise Study Data")  
  
ExerciseStudy$Health\_Awareness\_Bin <- cut(ExerciseStudy$Health\_Awareness,  
 breaks = 4, # Create 4 bins  
 labels = c("Low", "Moderate", "High", "Very High"),  
 include.lowest = TRUE)  
  
ggplot(ExerciseStudy, aes(x = Health\_Awareness\_Bin, y = Distance)) +  
 geom\_boxplot(fill = "skyblue", color = "black") +  
 labs(title = "Boxplot of Jogging Distance by Health Awareness Levels",  
 x = "Health Awareness Levels",  
 y = "Jogging Distance") +  
 theme\_minimal()  
  
  
ExerciseStudy$Peer\_influence\_Bin <- cut(ExerciseStudy$Peer\_influence,  
 breaks = 4,   
 labels = c("Low", "Moderate", "High", "Very High"),  
 include.lowest = TRUE)  
  
ggplot(ExerciseStudy, aes(x = Peer\_influence\_Bin, y = Distance)) +  
 geom\_boxplot(fill = "skyblue", color = "black") +  
 labs(title = "Boxplot of Jogging Distance by Peer Influence Levels",  
 x = "Peer Influence Levels",  
 y = "Jogging Distance") +  
 theme\_minimal()  
  
  
install.packages("corrplot")  
library(corrplot)  
correlation\_data <- ExerciseStudy %>%  
 select(Distance, Health\_Awareness, Computer\_Games, Television, Diet, Peer\_influence)  
  
correlation\_matrix <- cor(correlation\_data, use = "complete.obs")  
  
corrplot(correlation\_matrix, method = "circle", type = "upper",  
 tl.col = "black", tl.srt = 45,   
 title = "Correlogram of Factors Affecting Jogging Distance",  
 mar = c(0, 0, 1, 0))  
  
  
if (!require(pwr)) install.packages("pwr")  
library(pwr)  
  
sample\_size <- pwr::pwr.f2.test(u = 1, v = NULL, f2 = 0.15, sig.level = 0.05, power = 0.9)  
  
print(sample\_size)  
  
  
power <- pwr::pwr.f2.test(u = 1, v = 666, f2 = 0.15, sig.level = 0.05, )  
  
print(power)  
  
  
if (!require(knitr)) install.packages("knitr")  
library(knitr)  
  
sample\_size\_table <- data.frame(  
 Parameter = c("Sample Size", "Power", "Effect Size", "Significance Level", "Predictors"),  
 Value = c(666, 1, 0.15, 0.05, 5),   
 Notes = c("Number of participants needed for adequate power.",   
 "Probability of correctly rejecting the null hypothesis.",   
 "Medium effect size based on Cohen's f^2.",   
 "Alpha level for hypothesis testing.",   
 "Number of predictors in the model.")  
)  
  
kable(sample\_size\_table, caption = "Sample Size and Power Analysis")  
  
  
missing\_values <- colSums(is.na(ExerciseStudy))  
print(missing\_values)  
  
  
  
model1<-lm(Distance~Health\_Awareness+Peer\_influence,data=ExerciseStudy)  
  
par(mfrow=c(2,2))  
  
plot(model1, which=1:4)  
  
  
  
model2<-lm(Distance~Health\_Awareness+Peer\_influence+ Computer\_Games + Television + Diet,data = ExerciseStudy)  
  
par(mfrow = c(2, 2))  
plot(model2, which = 1:4)  
  
  
  
if(!require(broom)) install.packages("broom")  
if(!require(kableExtra)) install.packages("kableExtra")  
  
library(broom)  
library(kableExtra)  
  
# Fit the GLM model  
glm\_model1 <- glm(Distance ~ Health\_Awareness + Peer\_influence, family = gaussian(link = "log"), data = ExerciseStudy)  
  
# Tidy the model output  
glm\_results <- tidy(glm\_model1)  
  
# Create a table of the results  
kable(glm\_results, caption = "Summary of GLM Model: Distance ~ Health Awareness + Peer Influence") %>%  
 kable\_styling(full\_width = FALSE)  
  
  
  
  
library(broom)  
library(kableExtra)  
  
  
glm\_model2 <- glm(Distance ~ Health\_Awareness + Peer\_influence + Computer\_Games + Television + Diet, family = gaussian(link = "log"), data = ExerciseStudy)  
  
  
glm\_results2 <- tidy(glm\_model2)  
  
kable(glm\_results2, caption = "Summary of GLM Model: Distance ~ Health Awareness + Peer Influence + Computer Games + Television + Diet") %>%  
 kable\_styling(full\_width = FALSE)  
  
  
  
  
library(broom)  
  
anova\_results <- anova(glm\_model1, glm\_model2, test = "Chisq") # out anova is different due to the term “test= Chiq”, the Chi square test Improves the model fit. It compares the likelihood ratio tests for the two GLM   
  
anova\_tidy <- broom::tidy(anova\_results)  
  
kable(anova\_tidy, caption = "ANOVA Results Comparing glm\_model1 and glm\_model2") %>%  
 kable\_styling(full\_width = FALSE)