International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comprehensive Statistical Study on Health And Lifestyle of adults

Dr. Nileema Bhalerao

Associate Professor, Department of Statistics, Fergusson College (Autonomous) Pune

Abstract:
'Health is Wealth' is a concept which is deeply rooted in Indian culture. We grow up learning this through shlokas as well as stories told by our parents and grandparents.
Ayurveda is the first ancient Indian text that speaks about rest, movement, food and nutrition, meditation and emotional wellbeing, which originated more than 5000 years ago. It is considered as a guide for a joyful and healthy life.
Health is the biggest wealth for a human being in his/her entire lifetime. One can survive without excess money but cannot survive without good health.
Health is something that we cannot buy with money but we can take care of it and we can cure it when needed with the help of the money. Health refers to the physical and mental state of a human being. To stay healthy is not an option but a necessity to live a happy life. The basic laws of good health are related to the food we eat, the amount of physical exercise we do, our cleanliness, rest, and relaxation. A healthy person is normally more confident, self-assured, sociable, and energetic.
In recent times, we come across many people fighting with chronic diseases, physical health issues from infancy till death. The primary reason behind it must be changing environment, pollution along with lack of health literacy. The changes happening on a global scale may not be easily controlled by an individual, however maintaining an individual's health and leading a good lifestyle is in our hands.

INTRODUCTION

'Health is Wealth'
'Health is Wealth' is a concept which is deeply rooted in Indian culture. We grow up learning this through shlokas as well as stories told by our parents and grandparents.
Ayurveda is the first ancient Indian text that speaks about rest, movement, food and nutrition, meditation and emotional wellbeing, which originated more than5000 years ago. It is considered as a guide for a joyful and healthy life.
Health is the biggest wealth for a human being in his/her entire lifetime. One can survive without excess money but cannot survive without good health.
Health is something that we cannot buy with money but we can take care of it and we can cure it when needed with the help of the money. Health refers to thephysical and mental state of a human being. To stay healthy is not an option but a necessity to live a happy life. The basic laws of good health are related to the food we eat, the amount of physical exercise we do, our cleanliness, rest, and relaxation. A healthy person is normally more confident, self-assured, sociable, and energetic.
In recent times, we come across many people fighting with chronic diseases, physical health issues from infancy till death. The primary reason behind it must be changing environment, pollution along with lack of health literacy. The changes happening on a global scale may not be easily controlled by an individual, however maintaining an individual's health and leading a good lifestyle is in our hands.

MOTIVATION AND OBJECTIVES

It is well-known that the lifestyle of an individual has a very important role to play indeciding their health. In the earlier days, all people had a similar routine- sleep schedule, eating habits, work type as well as the

habitat. There was no extreme difference between the lifestyle of the rich and the poor, rural and urban person as the resources were limited. The major difference in lifestyle was observed only in different geographical regions. Yet these differences were veryorganic in nature and actually helped in maintaining good health of people in that region.

Nowadays, globalization, advancement in science and technology, population growth has played a key role for changes in lifestyle of an individual. We come across a vast variety ofdifferent resources, types of work (including shifts), eating habits, sleeping schedule etc, resulting in a unique lifestyle of every individual. These changes in lifestyle are not necessarily helping in maintaining one's health.

The study is an elaborate overview of different aspects of lifestyle and their relations with physical and mental health. We understand that health is a concerning topic for a lot of peopleat present. Many people are very conscious about maintaining good health, while others are not too worried. Their concern, or lack thereof, of the same influences a lot of their lifestyle, thereby impacting their health quality.

Through this work, I wanted to quantify the impact of lifestyle on health. The main idea behind the project is to find the lifestyle factors that are most responsible for shaping health -both physical and mental, of a person. Also, I wanted to find out interesting trends in health quality among different age groups, professions, income groups, streams of study and over different lifestyles.

I worked on this project with the following 5 key objectives in mind.

1. To find how health quality varies over age, profession, stream of study and several otherfactors.
2. To check if adults follow healthy lifestyle habits.
3. To check the significance of different lifestyle factors in maintaining good physical andmental health.
4. To understand trends and patterns in a specific lifestyle factor and its interrelationship withIndian socio-cultural environment.
5. To spread awareness among our generation about the most important lifestyle factorswhich affect one's health, as we are the future of our country.

METHODOLOGY AND DATA

In this project, I was mainly interested in studying how lifestyle factors affect health quality of adults. The type of data and the variables I was interested in studying were notreadily available in the form of a secondary dataset and hence, our team decided to collect primary data and analyze it with our specific goals in mind.

Our team created a google form with 41 questions related to the health and lifestyle of individualsand collected responses from adults between the ages of 18 and 60. The data collection wasdone both online and on-ground. On-ground data was collected in Pune whereas the online collection extended to a few states of India. In total, we received 591 responses and after cleaning the dataset, we worked with 579 responses.

One interesting aspect of our data was that, with the help of the primary data collected, we defined a new variable of our own called Physical Health Score. Based on the responses ofdifferent lifestyle factors given by the respondents, we gave each of them a physical healthscore, to signify their physical health quality as a function of different lifestyle factors. Themental health score was assigned to each respondent based on a standard questionnaire forgeneral mental health status. The project deals with exploiting both these measures to see how they interact with each other and all the lifestyle factors.

Note: As a statistician, good data is what we desire the most. Our project is based on primarydata mainly, hence getting the data as accurate as possible was a huge task for us. During theprocess of collection of primary data, these are some of the many things we became aware ofand it was a huge learning experience for us:

1. For primary data collection, the most attention should be paid towards preparing the questionnaire. Each question needed to be clear, to the point and easy to understand. This helps in getting the most

accurate and honest answers from respondents. Also, itbecomes easier to clean the data, as there are very few absurd entries.

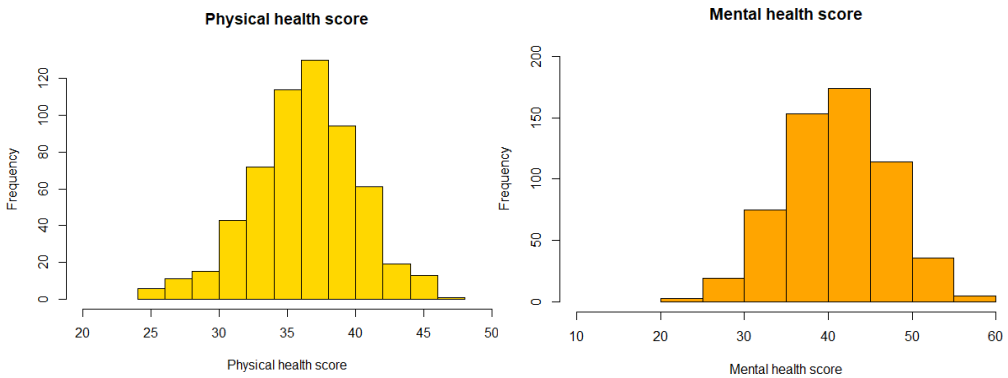
2. Preparing the questionnaire in 2 languages (English- as official language and Marathi- as local language) helped in the data collection. More people were comfortable to fill the questionnaire as there were language options. People from all socio-economic backgrounds could fill the questionnaire, resulting in good diversity among the respondents and more number of responses.
3. We realised that on ground data collection is very tough. The investment of time, efforts and making strangers interested in the study, so that they will be willing to fillthe questionnaire is hard. Despite these challenges, we could collect data from those who did not have smart phones, could not read or write only due to on ground data collection.

Software Used:

R-Studio, MS-Excel

EXPLORATORY DATA ANALYSIS

1. Histogram



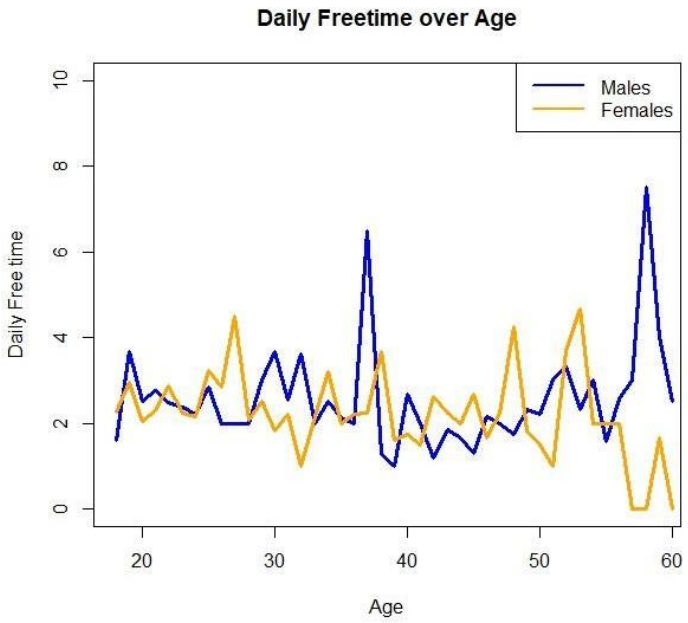
Interpretation: Looking at the bell-shaped curve, physical and mental health scores maypotentially follow normal distribution.

36-37 is the modal class of physical health score for entire population. We have decided anyvalue of or above 35 as good health. Most people have physical health score in the neighbourhood of the threshold value - a satisfactory observation.

40-45 is the modal class for mental health score. Threshold value is 42. Majority people havemental health score in the neighbourhood of the deciding value – again, a satisfactory observation.

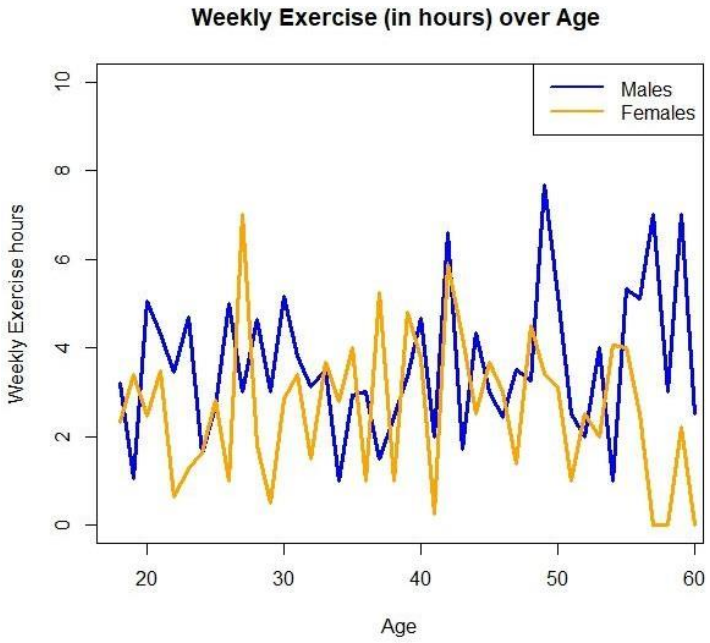
Interpretation: Proportion of people engaging in exercise is less than the proportion of peoplenot doing any exercise. People who do exercise have a higher mean health score than those who do not.

2. Line graph



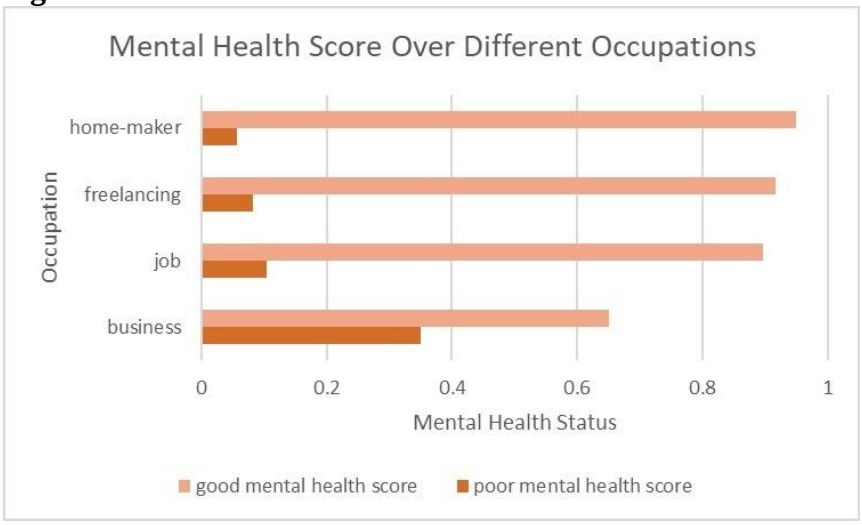
Interpretation: Age 25-30: Females have more free time than menAge 30-40: Males have more free time than females

This represents Indian social and cultural norms where men focus more on their career -resulting in having less free time during the ages 25-30. Whereas, women focus more ontaking care of the family and children - resulting in less free time during the ages 30-40

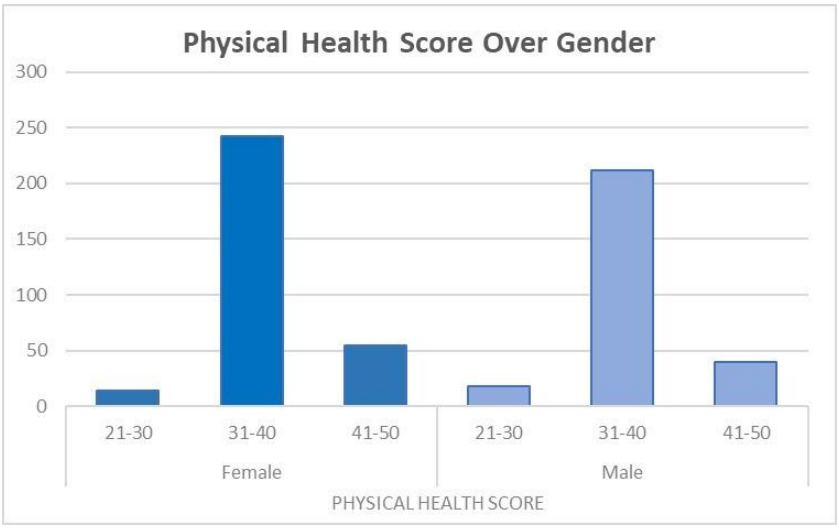


Interpretation: Women tend to engage in more exercise during the ages 25-30. This maybe to prevent the decline in physical health after age 30-35 and to have good health for future pregnancy. Men have consistent exercise hours during ages 20-35. After the age of 50, We see a sudden spike in exercise hours. This may indicate that men tend to devote more time toexercise after retirement.

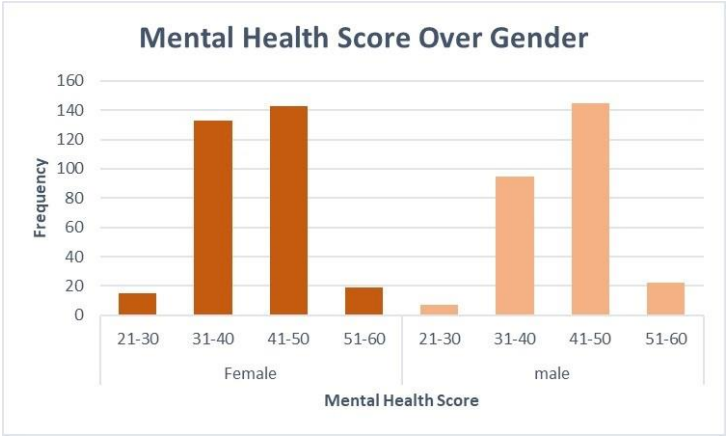
3. Multiple bar diagram



Interpretation: The proportion of people having good mental health status is highest among homemakers. Whereas it is lowest among business professionals. This may be dueto the extra/odd work hours, lack of security (in terms of fixed regular income) and a lot of responsibility/ups and downs of the business faced by an individual.

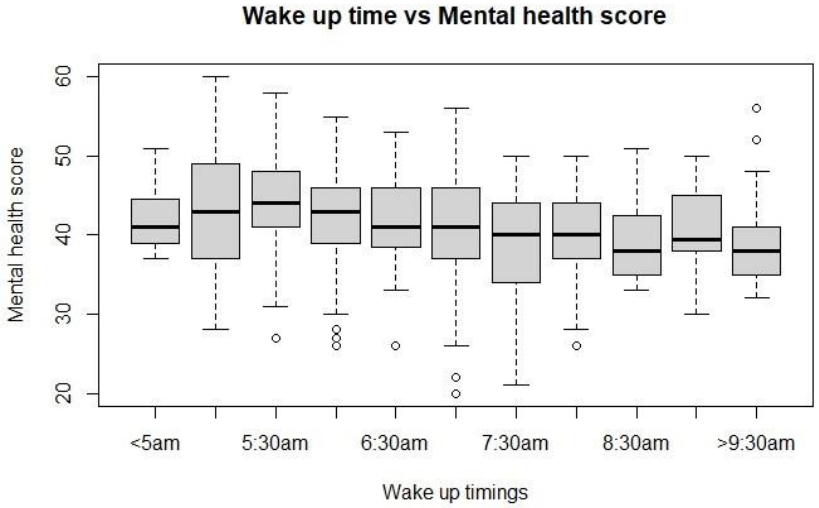


Interpretation: There is not a huge difference in the distribution of physical health scoresamong the two genders – almost similar proportions of men and women have satisfactory-good physical health score.

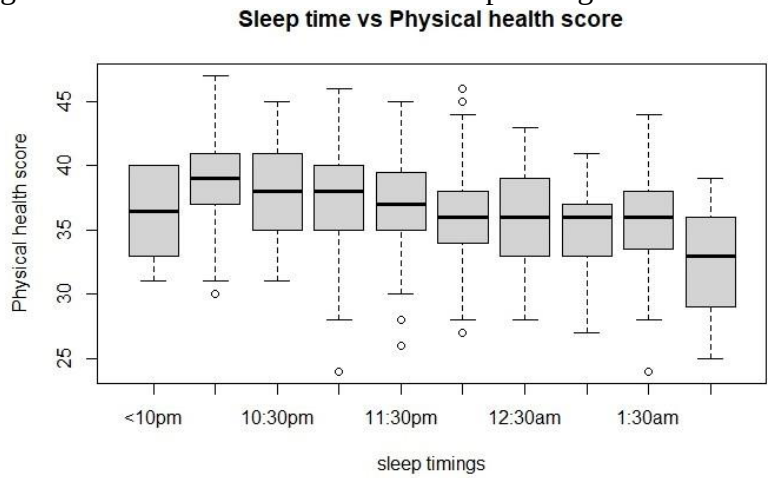


Interpretation: More women lie in the satisfactory-good mental health range than men.Maybe women are more open to talking about such mental health issues than men, andhence, resulting in better mental health.

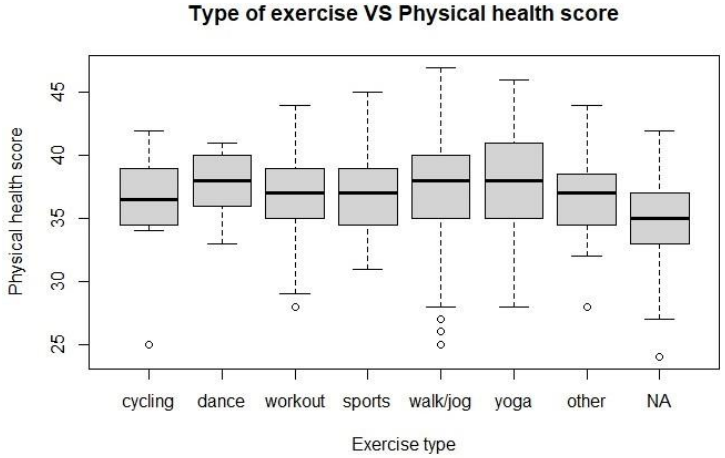
4. Boxplot



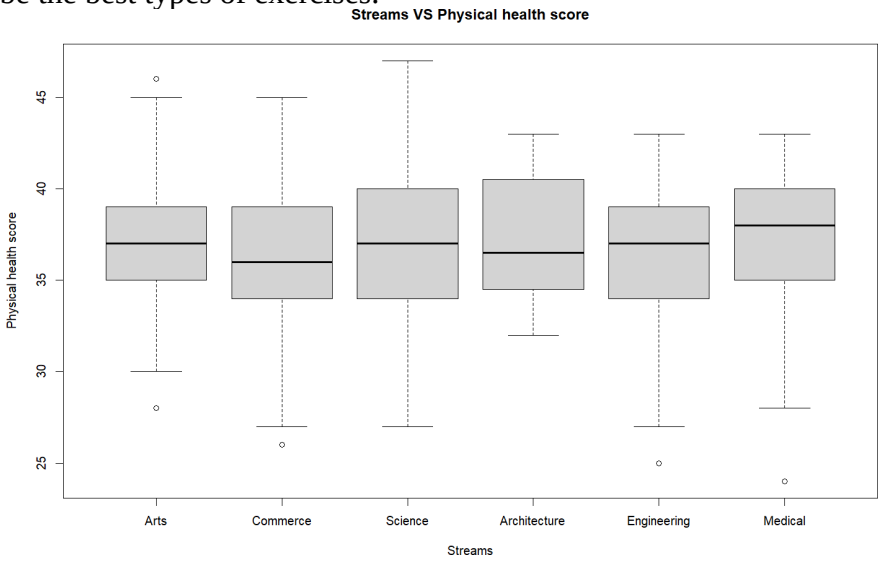
Interpretation: Mental health score is decreasing with increasing wake up time. This could imply that waking up early has a good effect on our health. Best wakeup timing is 5-5:30 am.



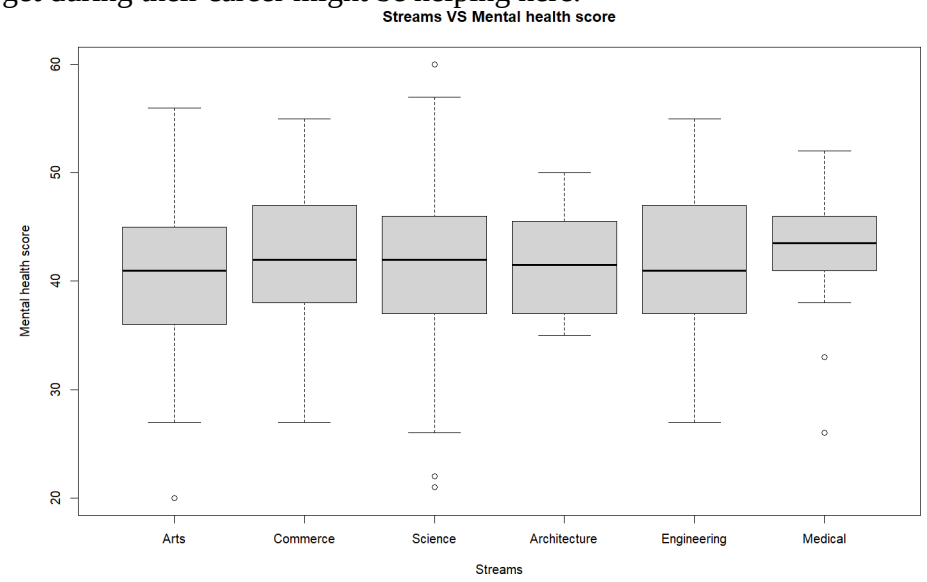
Interpretation: Physical health score is decreasing with increasing sleep time. This couldmean that sleeping early has a good effect on our health. Best sleep timing is 10 pm.



Interpretation: The physical health of people not doing any exercise is surely affectedadversely. Dance and yoga seem to be the best types of exercises.



Interpretation: Stream wise physical health score is almost the same. Medical professionalsare on the higher side of it. Even though they are very busy and have a hectic schedule, awareness about physical health that they get during their career might be helping here.



Interpretation: Stream wise mental health score is almost the same. Medical professionals areon the higher side of it. Even though they are very busy and have a hectic schedule, awareness about mental health that they get during their career might be helping here.

REGRESSION ANALYSIS: THEORY

Multiple Logistic Regression: Introduction

Multiple logistic regression is a statistical tool used to model the relationship between abinary response variable and multiple predictor variables. When the predictors are quantitative, the logistic regression model takes the form:

$Y = \prod(X) + \varepsilon$

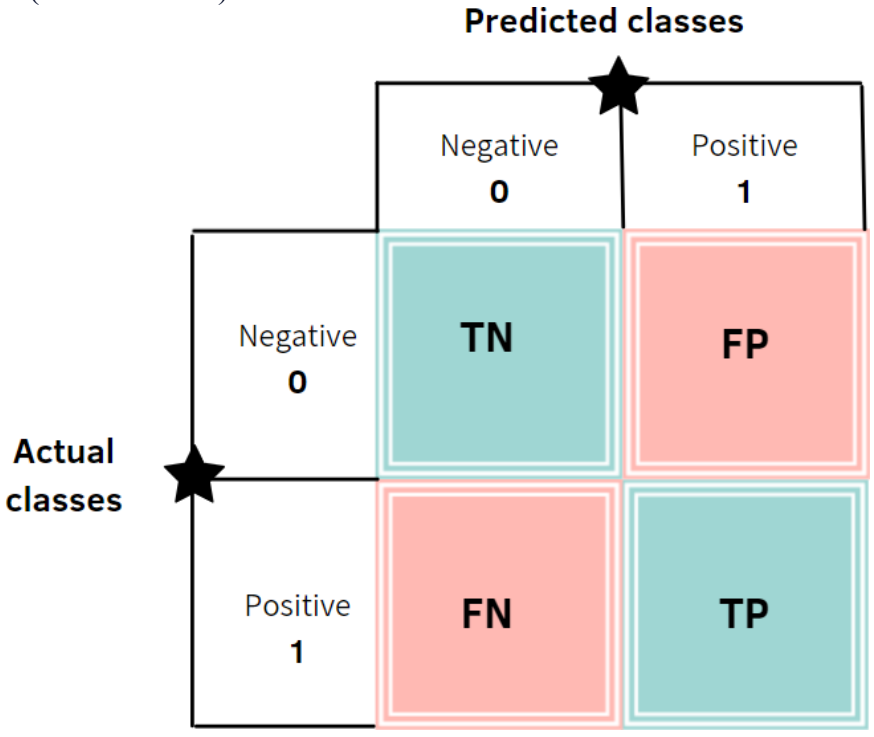
where, $\prod(X) = \frac{(e\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta kXk)}{1+(e\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta kXk)}$

and, $\varepsilon \sim B(\prod(X))$

Note that β_0, β_1, β_2 are the regression coefficients.
Logistic regression belongs to a family, named Generalized Linear Model (GLM), developedfor extending the linear regression model to other situations. Other synonyms are binary logistic regression, binomial logistic regression and logit model.
The logistic regression model can be fit using maximum likelihood estimation to estimate theregression coefficients and the intercept term.
Logistic regression does not return directly the class of observations. It allows us to estimate the probability (p) of class membership. The probability will range between 0 and 1. We needto decide the threshold probability at which the category flips from one to the other. by default, this is set to p=0.5, but in reality, it should be settled based on the analysis purpose.
The regressor variables and response variables are listed below.

Variable Definition	**Notation**
Physical health score, mental health score	Y
Water intake	X1
Sleep hours	X2
Study/work hours	X3
Screen time	X4
Cigarette	X5
Exercise	X6
Free time	X7
Junk Food consumption	X8

This data is now segregated into 80% training and 20% test data-set. A logistic regression model is fitted on the training data set, and using it we can proceed to predict the values ofthe test data set.
Confusion matrix: A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the truevalues are known.



The confusion matrix has basic 4 combinations
1. True positive rate
2. True negative rate
3. False positive rate
4. False negative rate

Terminologies in confusion matrix are as follows.
Accuracy: Overall, how often the classifier is correct? $(TP + TN)/TOTAL$
Mis-classification Rate: Overall, how often is it wrong? $(FP + FN)/TOTAL$
Matthew's Correlation Coefficient: $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FN)(FP+TN)}}$

It is the correlation between actual condition (actual class of dependent variable) andpredicted condition (predicted class of response variable).

ROC Curve: An ROC curve (receiver operating characteristic curve) is a graph showing theperformance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate and False Positive Rate

AUC: AUC stands for "Area under the ROC Curve. AUC provides an aggregate measure ofperformance across all possible classification thresholds

REGRESSION ANALYSIS: ANALYSIS OF DATA

1) Fitting Multiple Logistic Regression on Physical Health Status Based on LifestyleFactors

Predicted Variable	Estimate	Standard Error	Z- value	P-value
(Intercept)	0.26403	0.89622	0.295	0.768298
Water Intake	0.31688	0.11195	2.831	0.004646
Sleep Hours	0.20457	0.10541	1.941	0.052289
Study/Work hours	-0.04392	0.03729	-1.178	0.238835
Screen time	-0.15196	0.03791	-4.008	6.12e-05
Weekly smoking	-0.03011	0.01461	-2.061	0.039283
Weekly exercise hours	0.15446	0.04617	3.345	0.000822
Daily Free time	-0.10356	0.07481	-1.384	0.166261
Weekly Junk food consumption	-0.24215	0.05409	-4.477	7.59e-06

Null deviance: 526.2
Residual deviance: 440.87
AIC: 458.87
Here, we checked the significance of regressors. As $(Null\ Deviance - Residual\ Deviance) > \chi 2$ (8,0.05), we may conclude that regression model is significantat 5% level of significance. Further, we get to know that following regressors are significant:
Water intake, Sleep hours, Screen time, Cigarette, Exercise hours, Junk food

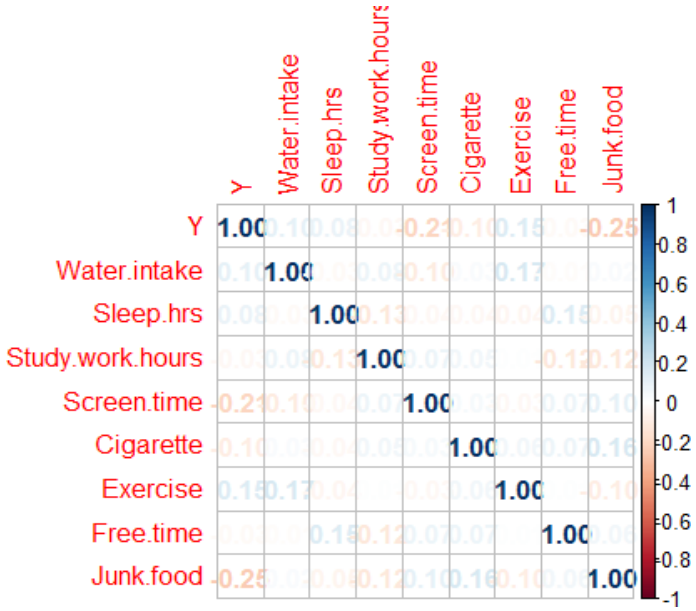
A) Checking model assumptions

A.1) To check if residuals are random:
Run Test is used to check the randomness of residuals.Null Hypothesis: Residuals are random

P-value of test 0.4571, which is greater than α (0.05).Hence, we can say that residuals are random.

A.2) Checking multicollinearity in data

Correlation Plot:



To check multicollinearity in the model, VIF score is used:

Predicted Variable	VIF Score
Daily Wate Intake	1.060676
Sleep hours	1.036477
Study/Work hours	1.085903
Daily screen time	1.015260
Cigarettes per week	1.064700
Exercise hours per week	1.075898
Daily free time	1.048559
Junk food consumption per week	1.047877

Interpretation:

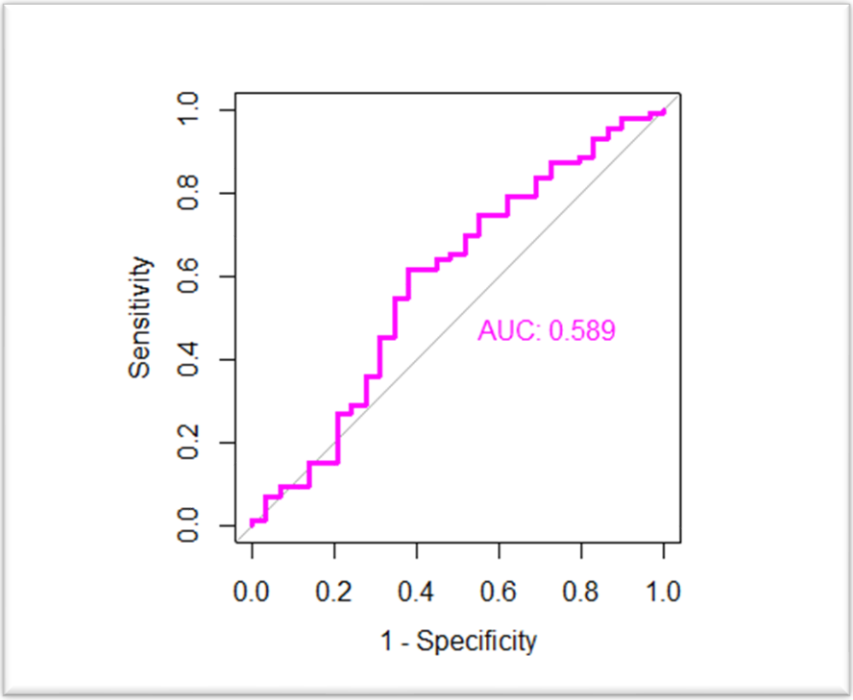
The VIF (variance inflation factor) score is a measure of multicollinearity between variables in a regression model. A VIF score of 1 indicates that there is no multicollinearity between the variable and the other variables in the model. Typically, a VIF score greater than 5 or 10 is considered a sign of high multicollinearity. As all the VIF scores in this table are well below5, it suggests that there is little or no multicollinearity between these variables.

Confusion Matrix

Reference Prediction	0	1
0	18	33

1	11	53

Metrics	Values
Accuracy	0.6174
95% CI	(0.5221,0.7065)
Sensitivity	0.6207
Specificity	0.6163
AUC	0.589
ROC curve:	



Point obtained (cut-off point): 0.7778832

Here we count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under the Curve**) ROC (**Receiver Operating Characteristics**) curve. It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (**Area Under the Receiver Operating Characteristics**). AUC is an effective way to summarise the overall diagnostic accuracy of the test. It takes value from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and value of 1 reflectsa perfectly accurate test. A value of 0.5 for AUC indicates that the ROC curve will fall on the diagonal (i.e. 45 degree line) and hence suggest that the diagnostic test has no discriminatory ability.

- AUC value obtained by us is 0.589 which means that our model has very slight discriminatoryability
- Value of Matthew's Correlation Coefficient = 0.2071 which implies that our prediction modelis average random method of classification
- From the confusion matrix it is clear that accuracy of model is 61.74%

2) Fitting Multiple Logistic Regression on Mental Health Status Based on LifestyleFactors

Predicted Variable	Estimate	Standard Error	Z- value	P-value
(Intercept)	-1.96263	0.81133	-2.419	0.015563
Water Intake	0.07998	0.07786	1.027	0.304327
Sleep Hours	0.15108	0.09348	1.616	0.106035
Study/Work hours	0.13229	0.03261	4.057	4.96e-05
Screen time	-0.09971	0.0407	-2.926	0.003429
Weekly smoking	0.01175	0.01296	0.907	0.364566
Weekly exercise hours	0.11705	0.03346	3.498	0.000469
Daily Free time	-0.02893	0.06459	-0.448	0.654223
Weekly Junk food consumption	-0.07588	0.04796	-1.582	0.113615

Null deviance: 641.75
Residual deviance: 590.64
AIC: 608.64

Here we checked the significance of regressors. As $(Null\ Deviance - Residual\ Deviance) > \chi^2_{8,0.05}$ hence we may conclude that regression model is significantat 5% level of significance. Further. we get to know that following regressors are significant:

Study/work hour, Screen time, Exercise hours

A) Checking model assumptions

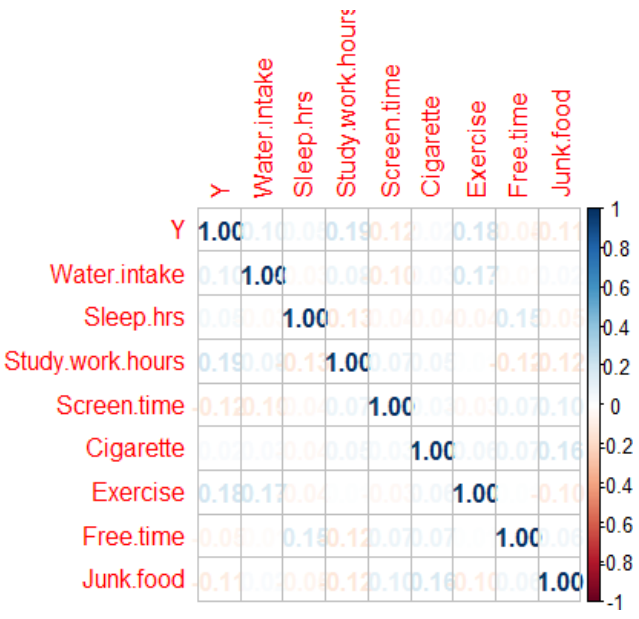
A.1) To check if residuals are random:

Run Test is used to check randomness of residuals.Null Hypothesis: Residuals are random.

P-value of test is 0.2259, which is greater than α (0.05).Hence, we can say that residuals are random.

A.2) Checking multicollinearity in data

Correlation Plot:



To check multicollinearity in the model, VIF score is used:

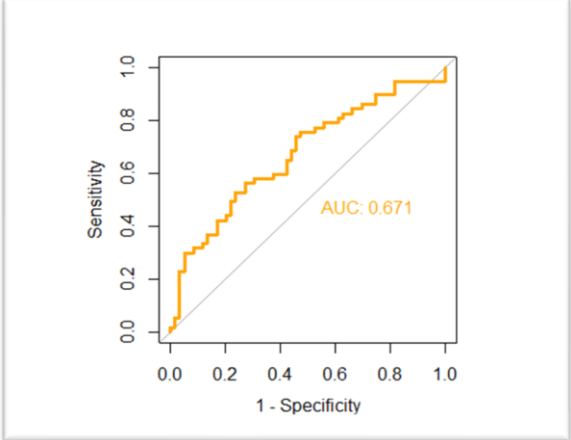
Predicted Variable	VIF Score
Daily Wate Intake	1.036188
Sleep hours	1.062537
Study/Work hours	1.076740
Daily screen time	1.039964
Cigarettes per week	1.045480
Exercise hours per week	1.049371
Daily free time	1.049502
Junk food consumption per week	1.073252

Interpretation:
The VIF (variance inflation factor) score is a measure of multicollinearity between variables in a regression model. A VIF score of 1 indicates that there is no multicollinearity between the variable and the other variables in the model. Typically, a VIF score greater than 5 or 10 is considered a sign of high multicollinearity. As all the VIF scores in this table are well below 5, it suggests that there is little or no multicollinearity between these variables.

Confusion Matrix

Reference Prediction	0	1
0	43	25
1	16	32

Metrics	Values
Accuracy	0.6466
95% CI	(0.5524,0.7331)
Sensitivity	0.7288
Specificity	0.5614
AUC	0.671
ROC curve:	



Point obtained (cut-off point): 0.495903

- AUC value obtained by us is 0.671 which means that our model has a discriminatoryability
- Value of Matthew's Correlation Coefficient = 0.2945, which implies that ourprediction model is average random method of classification
- From the confusion matrix it is clear that accuracy of model is 64.66%

Shapiro Wilk test of normality

1. Introduction: The Shapiro-Wilk test is a statistical test used to determine if a set of data follows a normal distribution. It is a commonly used test in statistics to assess the normality assumption required for many parametric tests. The Shapiro-Wilk test works by calculating atest statistic (W) that compares the observed distribution of data to the expected normal distribution.
2. Hypothesis: H_0: Data is normally distributed against H_1: Data is not normally distributed
3. Test Statistic:

$$W = \frac{(\sum_i a_i x_i)^2}{\sum (x - \bar{x})^2}$$

where,

x_i: i^{th} ordered observation from the sample
$\bar{x}$: sample mean
a_i : coefficients used to calculate the expected normal scores, which depend on the sample size and distribution

i) Decision Rule and Interpretation: The decision rule for the Shapiro-Wilk test is to reject the null hypothesis of normality if the test statistic W is less than the critical value at a chosen significance level. The critical values for W depend on the sample size, but tables of critical values are available for different sample sizes and significance levels.

Note: In the Shapiro-Wilk test, the null hypothesis is that the data are normally distributed. The test statistic, W, ranges between 0 and 1, where a value of 1 indicates perfect normality, and smaller values indicate departures from normality.

If the p-value from the Shapiro-Wilk test is less than the significance level (e.g., 0.05), we reject the null hypothesis of normality. However, even if the p-value is greater than the significance level, we should also examine the W statistic to assess the degree of normality.

If the W statistic is close to 1 (e.g., 0.95 or higher), we can conclude that the data are approximately normally distributed.

Checking normality of different variables in the data

Hypothesis: H_0: Given variable is normally distributed Against H_1: Given variable is not normally distributed

Decision: The test statistic, W, ranges between 0 and 1, where a value of 1 indicates perfect normality, and smaller values indicate departures from normality.

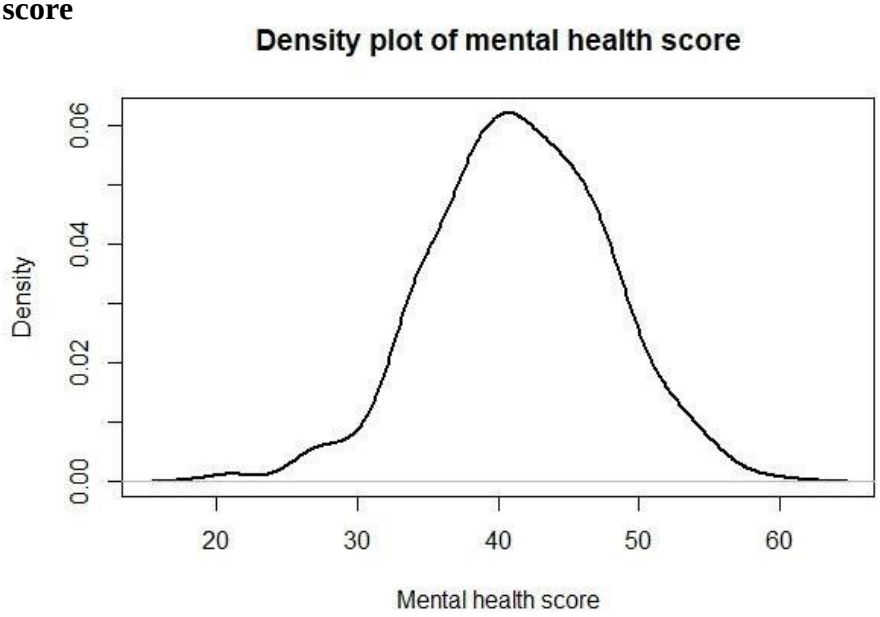
Variable	W-value	Decision
Mental Health Score	0.99517	Accept H_0
Physical Health Score	0.98802	Accept H_0
Daily Screen time (in hrs)	0.90795	Reject H_0
Daily Water Intake (in Ltr)	0.61622	Reject H_0
Sleep Duration	0.93226	Accept H_0

Hence from the above table we can conclude that following variables are normally distributed.
1) Mental Health Score
2) Physical Health Score

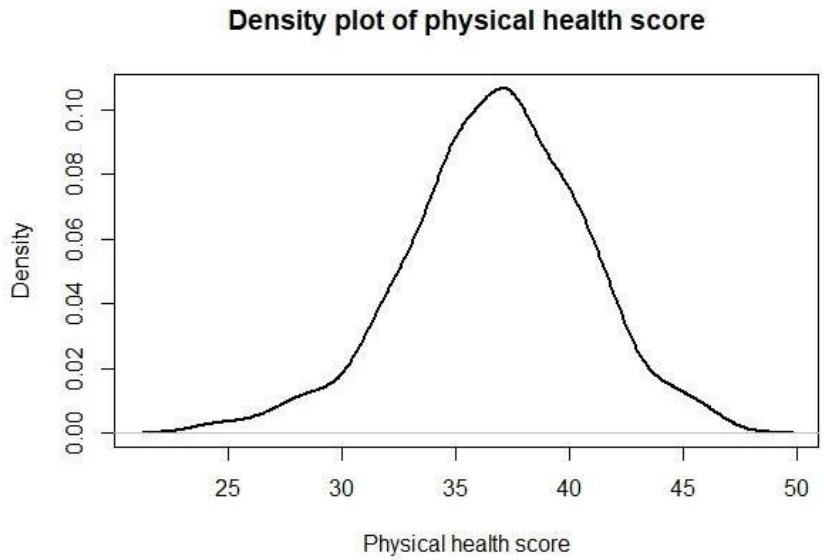
And variables which are not normally distributed are:
1) Daily Screentime (in hrs)
2) Daily Water Intake (in litre)

Density plots of normally distributed variables: Even from visual inspection alone, they appear to be normally distributed quite well for real-life data.

1) Mental health score



2) Physical health score



Analysis of Variance using Completely Randomised Design

1. *Introduction*: Completely Randomized Design is a statistical method used to analyze data from an experiment in which the treatments are randomly assigned to the experimental units without any specific blocking or grouping. In ANOVA for CRD, the main goal is to determine if there is a statistically significant difference among the means of different treatments.
2. *Assumptions:* 1. Residuals are normally distributed with mean zero and constant variance.
3. Effects of the treatments are additive in nature.
4. *Hypothesis*: H0: The means of all the groups are not significantly different against H1: Themeans of the groups are significantly different
5. *Test Statistic:*

where,

$$F = \frac{MS_{between}}{MS_{within}} \rightarrow F_{t-1,n-t}$$

$MS_{between}$: mean sum of squares between the groups, calculated as the sum of squares between the group means divided by the degrees of freedom between thegroups

MS_{within}: mean sum of squares within the groups, calculated as the sum of squareswithin each group divided by the degrees of freedom within the groups

t: number of treatments

n: total number of replications

i) Decision Rule and Interpretation: If the F-value is greater than the critical F-value, we reject the null hypothesis and conclude that there is a significant difference between the means of at least one pair of groups. If the F-value is less than or equal to the critical F-value,we fail to reject the null hypothesis and conclude that there is not enough evidence to suggestthat the means of the groups are different.

Post-hoc analysis using Method of Critical Difference

i) Introduction: Post hoc analysis is a statistical method used to determine which pairs of treatments in an experiment are significantly different from each other after a significant overall effect has been detected by an ANOVA or other statistical test. One commonly usedmethod for post hoc analysis is the method of critical difference. The method of critical difference involves calculating a critical value for the difference between two means that must be exceeded in order to conclude that the means are significantly different from each other.

ii) Assumptions: Independence, normality, and homogeneity of variance of residuals.

iii) Hypothesis: H0: The mean difference between any two groups is not significant againstH1: The mean difference between at least one pair of groups is significant

iv) Test Statistic:

$$CD = q(\alpha, df)\sqrt{\frac{MS_{error}}{r}}$$

where,

df: degrees of freedom

q(α,df): critical value from the Studentized range distribution with α level ofsignificance and df degrees of freedom

MSerror: mean sum of squares of the error term from the ANOVA tableR: number of replicates per treatment

v) Decision Rule and Interpretation: The decision rule is to reject the null hypothesis of nosignificant difference between two means if the absolute difference between their means isgreater than the critical difference value.

Comparing mental health scores over different income groups

ANOVA

Response Variable: Mental health scoreTreatments: Different income ranges

Hypothesis: H0: Average mental health score for different income ranges is same.
Against H1: Average mental health score for different income ranges is significantly different.ANOVA

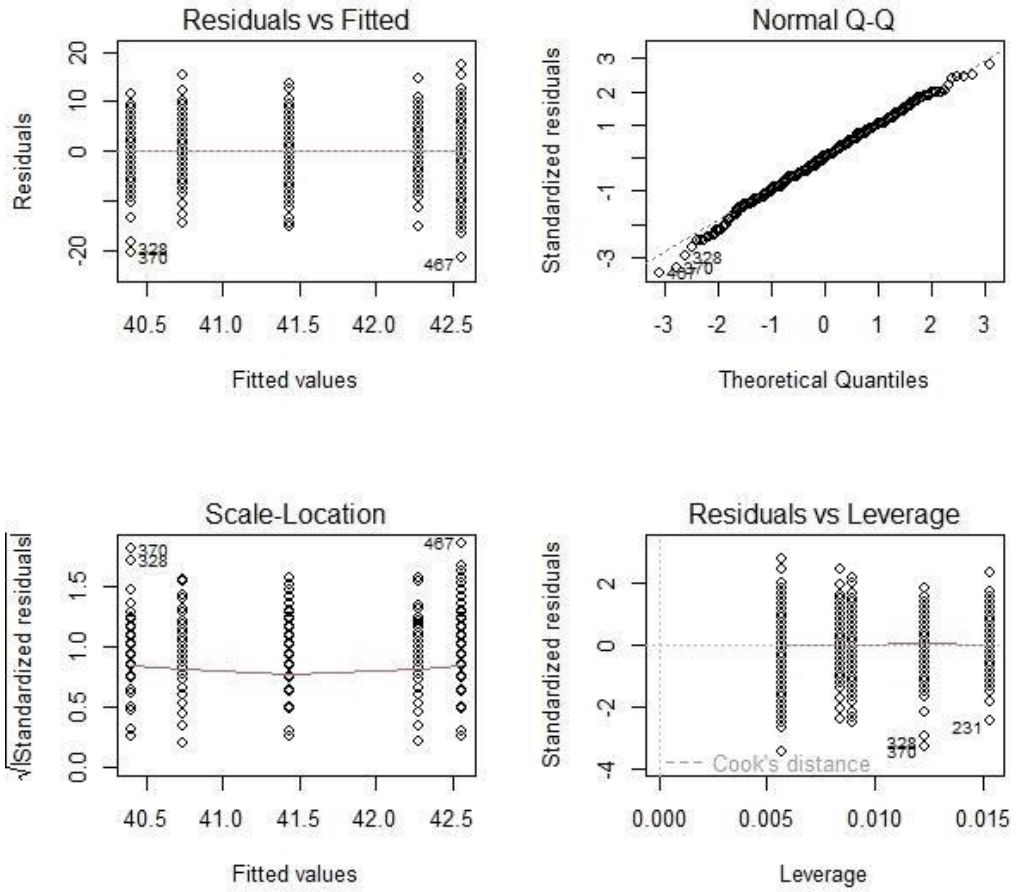
Table:

Variable	Df	SS	MSS	F-ratio	P-value
Treatments	4	402	100.40	2.577	0.0367
Residuals	547	21316	38.97		

Decision: As p-value is less than the level of significance (α = 0.05), we will reject the nullhypothesis.

Conclusion: Average Mental health score for different income ranges may be significantlydifferent.

Residual Analysis:



From the above residual plots, we can conclude that residuals are iid normal variates.

Post-hoc analysis

μ_i: $Average\ mental\ health\ score\ for\ i^{th}\ income\ group$ (for i=1,2,3,4,5)

Hypothesis: H0: $\mu_i = \mu_j\ (\forall\ i, j = 1,2,3,4,5), i \neq j$ against H1: $\mu_i \neq \mu_j\ (\forall\ i, j = 1,2,3,4,5), i \neq j$

Post-hoc table:

Following is the table of p-values for different treatment mean combinations.

	μ_1	μ_2	μ_3	μ_4
μ_2	0.401			
μ_3	0.111	0.387		
μ_4	0.702	0.256	0.071	
μ_5	0.014	0.136	0.753	0.010

Decision: The mean combinations (μ1, μ5) and (μ4, μ5) shows inequality as their P-value isless than level of significance (α = 0.05).

Conclusion: There is significant difference in average mental health score of income range

1. less than 2.5 lakh and more than 10 lakhs.
2. between 7.5 -10 lakh and more than 10 lakhs.

Further, by doing individual testing of means of above combinations using t-test for equality of two

population means, we found that mental health score of people with income more than10 lakh is more.

Mental Health score for income range less than 2.5 lakh and more than 10 lakh
Let $\mu 1$ be the mental health score for income range less than 2.5 lakh and
$\mu 2$ be the average mental health score for income range more than 10 lakh

Alternative hypothesis	P-value	Decision
$\mu 1 < \mu 2$	0.006272	Reject H_0

Conclusion: Average mental health of people in income range more than 10 lakh is betterthan that of people in income range less than 2.5 lakh.

Mental Health score for income range 7.5 to 10 lakh and more than 10 lakh
Let $\mu 1$ be the mental health score for income range 7.5 to 10 lakh and
$\mu 2$ be the average mental health score for income range more than 10 lakh

Alternative hypothesis	P-value	Decision
$\mu 1 < \mu 2$	0.008631	Reject H_0

Conclusion: Average mental health of people in income range more than 10 lakh is betterthan that of people in income range 7.5 to 10 lakh.

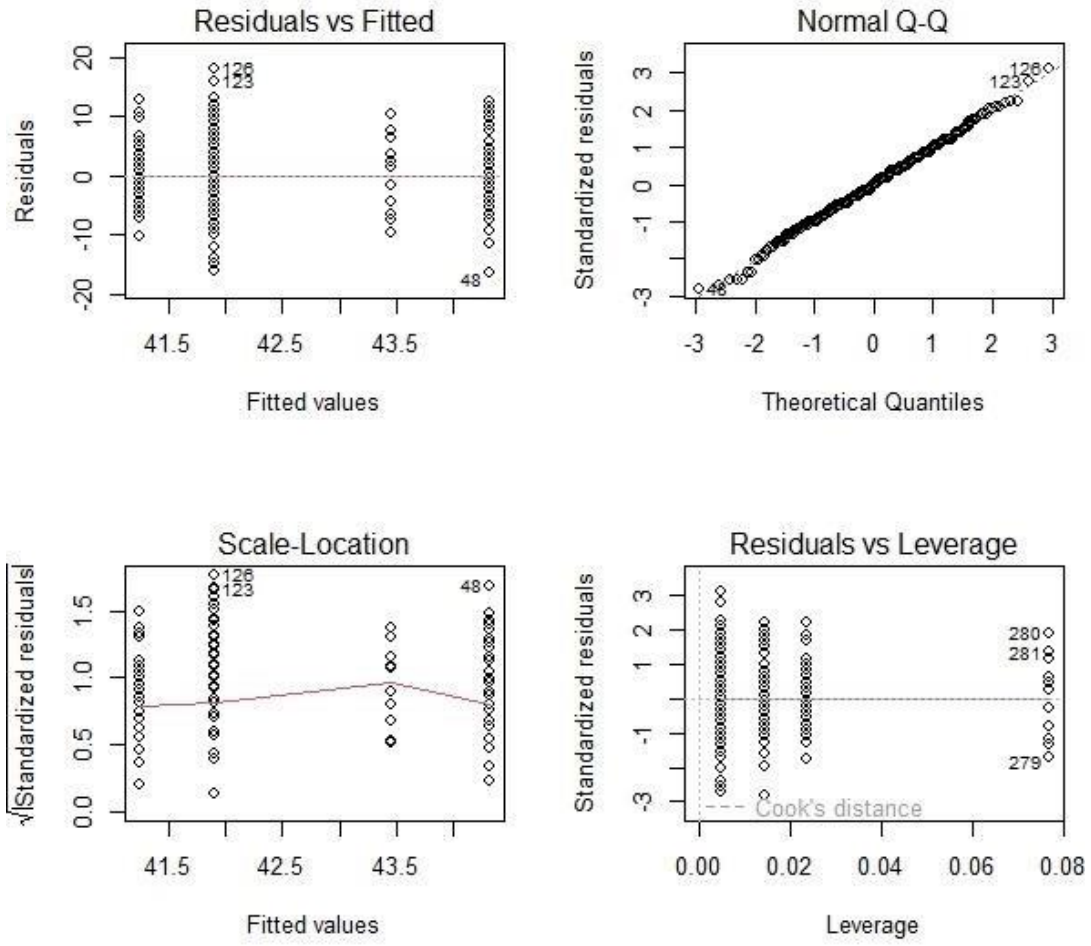
Comparison of mental health score over different occupations
ANOVA
Response Variable: Mental health scoreTreatments: Different occupations

Hypothesis: H_0: Average mental health score for different occupations is same.
Against H_1: Average mental health score for different occupations is significantly different. ANOVA Table:

Variable	**Df**	**SS**	**MSS**	**F-ratio**	**P-value**
Treatments	3	375	124.85	3.676	0.0125
Residuals	324	11003	33.96		

Decision: As p-value is less than the level of significance ($\alpha = 0.05$), we will reject the nullhypothesis.

Conclusion: Average Mental health score for different occupations may be significantlydifferent.
Residual Analysis:



From the above residual plots, we can conclude that residuals are iid normal variates.

Post-hoc analysis

μ_i: *Average mental health score for* i^{th} *income group* (for i=1,2,3,4,5)

Hypothesis: H0: $\mu_i = \mu_j$ $(\forall\ i, j = 1,2,3,4,5), i \neq j$ against H1: $\mu_i \neq \mu_j$ $(\forall\ i, j = 1,2,3,4,5), i \neq j$

Post-hoc table:

The following is the table of p-values for different treatment mean combinations.

	μ_1	μ_2	μ_3
μ_2	0.0032		
μ_3	0.6254	0.3488	
μ_4	0.0073	0.5045	0.2302

Decision: The mean combinations $(\mu 1, \mu 2)$ *and* $(\mu 4, \mu 1)$ shows inequality as their P-value isless than level of significance $(\alpha = 0.05)$.

Conclusion: There is significant difference in average mental health score of followingoccupation pairs:

i) Home-maker and business

ii) Business and job

Also, by doing the individual testing of means of above combinations using t-test for equalityof two population means, we found that mental health of people doing business is more.

Mental health score for people doing business and job

Let $\mu 1$ be the mental health score for people doing business and
$\mu 2$ be the average mental health score for people with a job

Alternative hypothesis	P-value	Decision
$\mu 1 > \mu 2$	0.001721	Reject H_0

Conclusion: Average mental health of people with business is better than that of people witha job.

Mental health score for people doing business and home-makers

Let $\mu 1$ be the mental health score for for people doing business and
$\mu 2$ be the average mental health score for Home-makers

Alternative hypothesis	P-value	Decision
$\mu 1 > \mu 2$	0.002567	Reject H_0

Conclusion: Average mental health of people with business is better than that of Home-makers.

Comparison of physical health score over different income ranges

ANOVA

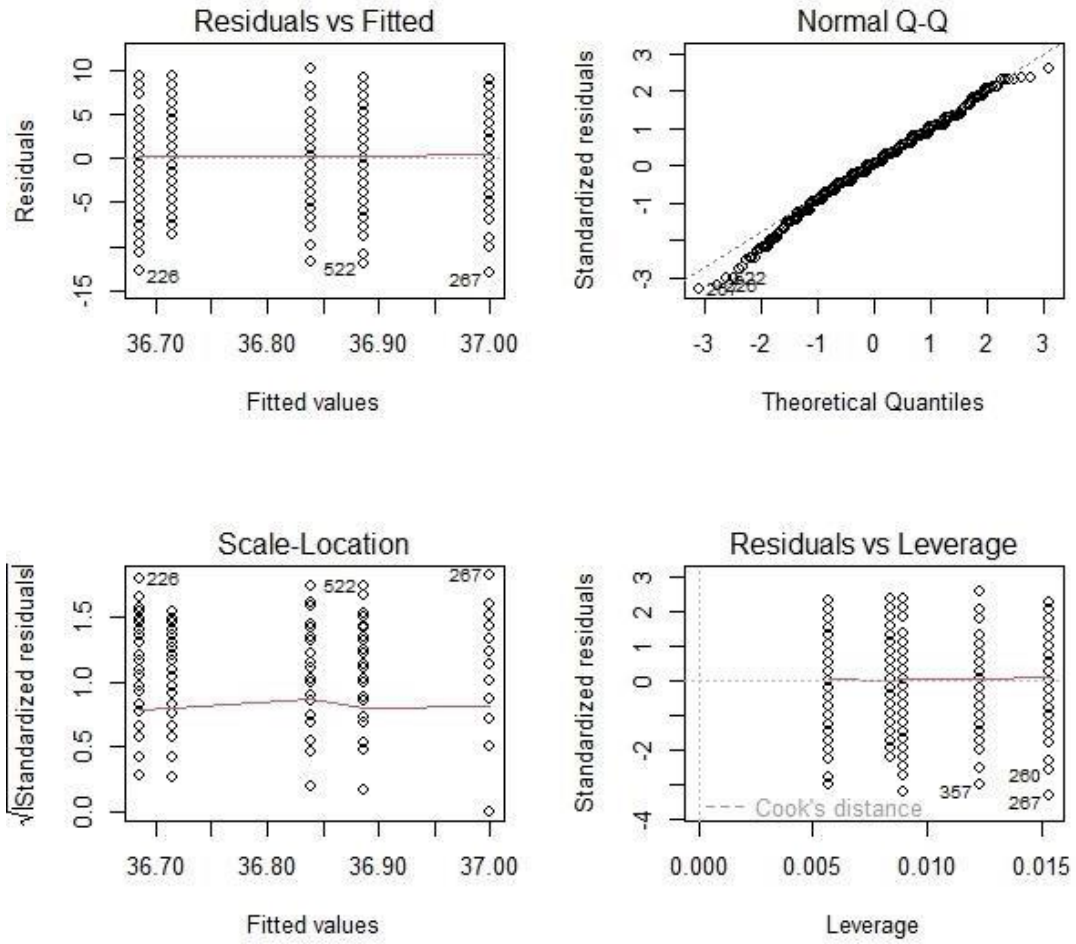
Response Variable: Physical health scoreTreatments: Different income ranges

Hypothesis: H_0: Average physical health score for different income ranges is same.
Against H_1: Average physical health score for different income ranges is significantlydifferent.
ANOVA Table:

Variable	**Df**	**SS**	**MSS**	**F-ratio**	**P-value**
Treatments	4	6	1.565	0.1	0.982
Residuals	547	8541	15.614		

Decision: As p-value is greater than the level of significance ($\alpha = 0.05$), we failed to rejectthe null hypothesis.

Conclusion: Average mental health score for different income ranges may be same.Residual Analysis:



From the above residual plots, we can conclude that residuals are approximately iid normalvariates.

Comparison of mean physical health scores over different occupations

ANOVA
Response Variable: Physical health scoreTreatments: Different Occupations

Hypothesis: H0: Average physical health score for different occupations is same.

Against H1: Average physical health score for different occupations is significantly different.

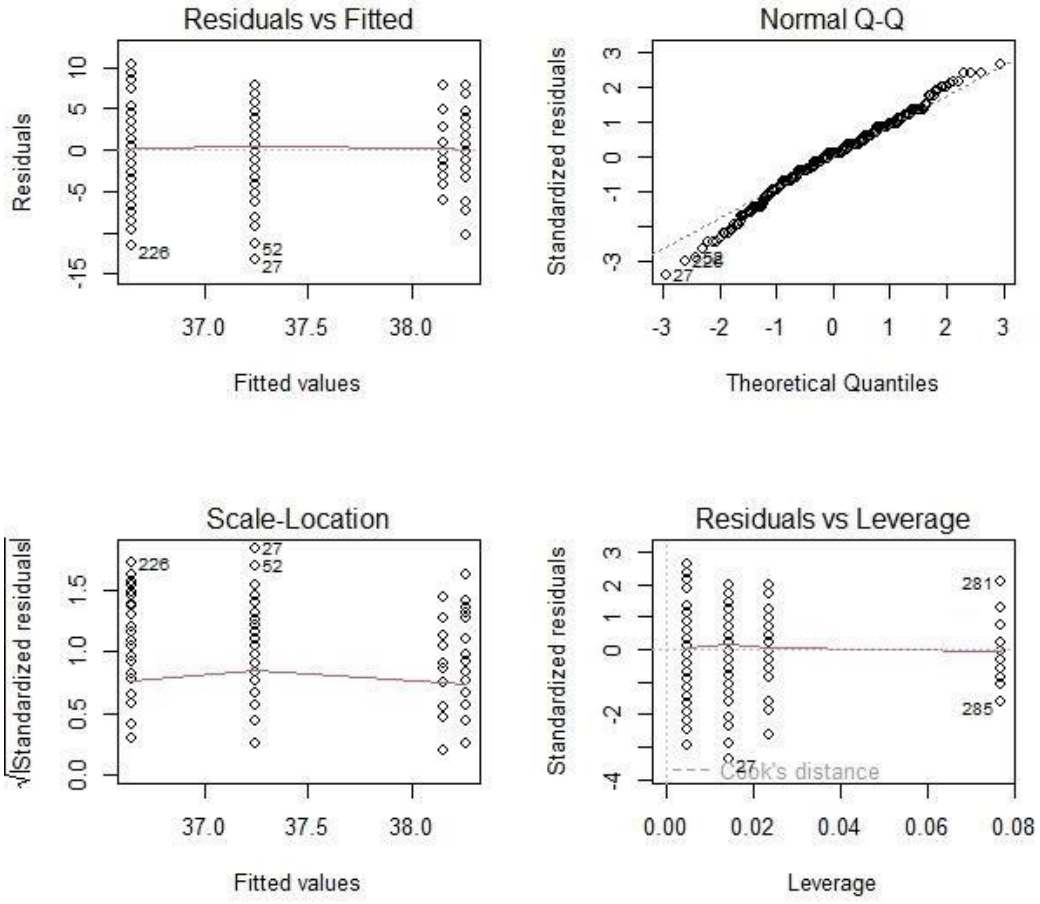
ANOVA Table:

Variable	Df	SS	MSS	F-ratio	P-value
Treatments	3	113	37.74	2.422	0.0658
Residuals	324	5047	15.58		

Decision: As p-value is greater than the level of significance ($\alpha = 0.05$), we fail to reject thenull hypothesis.

Conclusion: Average physical health score for different occupations may be same at the givenlevel of significance.

Residual Analysis:



From the above residual plots, we can conclude that residuals are approximately iid normalvariates.

Comparison of mental health score over different ranges of free time
ANOVA
Response Variable: Mental health score

Treatments: Different ranges of free time

Hypothesis: H_0: Average mental health score for different free time ranges is same.

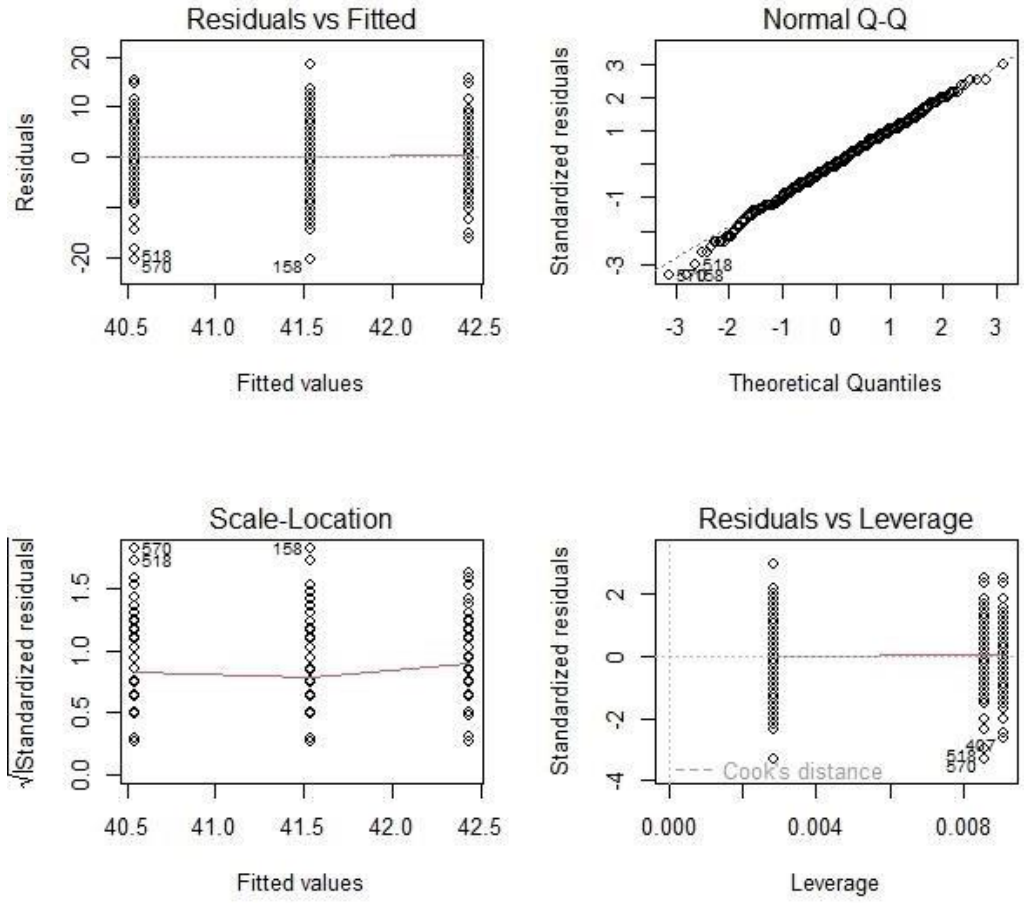
Against H_1: Average mental health score for different free time ranges is significantlydifferent.

ANOVA Table:

Variable	Df	SS	MSS	F-ratio	P-value
Treatments	2	205	102.57	2.665	0.0705
Residuals	576	22170	38.49		

Decision: As the p-value is greater than the level of significance ($\alpha = 0.05$), we fail to rejectthe null hypothesis.
Conclusion: Average mental health score for different free time ranges may be same at thegiven level of significance.

Residual Analysis:



From the above residual plots, we can conclude that residuals are approximately iid normalvariates.

Comparison of mental health score over streams of study
ANOVA
Response Variable: Mental health scoreTreatments: Different streams of study

Hypothesis: H_0: Average mental health score for different streams of study is same.

Against H_1: Average mental health score for different streams of study is significantlydifferent.

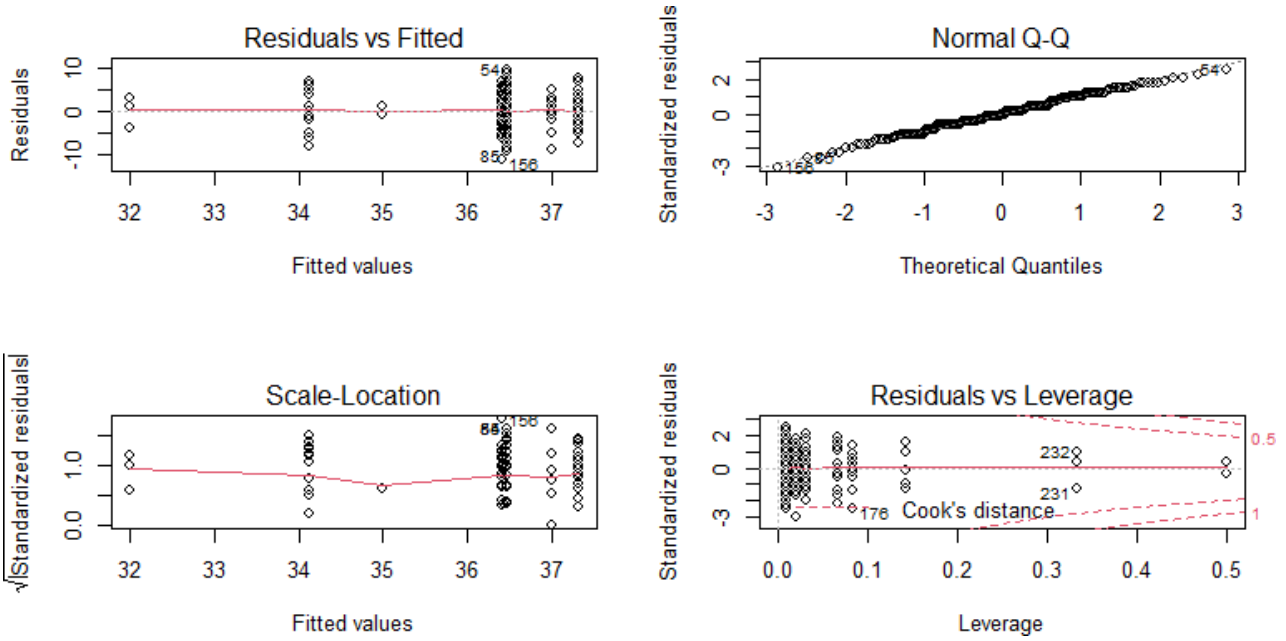
ANOVA Table:

Variable	Df	SS	MSS	F-ratio	P-value
Treatments	7	170.5	24.36	1.74	0.101
Residuals	225	3150	14		

Decision: As the p-value is greater than the level of significance ($\alpha = 0.05$), we fail to rejectthe null hypothesis.

Conclusion: Average mental health score for different streams of study may be same at thegiven level of significance.

Residual Analysis:



From the above residual plots, we can conclude that residuals are approximately iid normalvariates.

A) Chi Square Test of Independence of Attributes

i) Introduction: The chi-square test of independence is a statistical test used to determine if there is a relationship between two categorical variables. It is commonly used to analyze datafrom a contingency table.

ii) Assumptions: Independence: The observations in each cell of the contingency table shouldbe independent of each other.

Sample size: The sample size should be large enough to ensure that the expected frequenciesare not too small.

Expected frequency: The expected frequency for each cell should be greater than 5.

iii) Hypothesis: H0: There is no association between the two categorical variables against H1:There is an association between the two categorical variables.

Test Statistic

$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(Oij-eij)^2}{eij}$ follows a chi square distribution with (r-1)(s-1) d.f under H_0

where

r: number of rows (in contingency table)

s: number of columns (in contingency table)

O_{ij}: observed frequency in each cell

e_{ij}: expected frequency in each cell

iv) Decision Rule and Interpretation: The decision rule is to reject the null hypothesis if the test statistic is greater than the critical value at a chosen significance level, or if the p-value isless than the chosen significance level.

A.1) Checking independence of physical and mental health

Here, Physical Health Score (PHS) greater than 35.5 is considered as good score. Similarly, Mental Health Score (MHS) greater than 42 is considered as good score.Hypothesis:

H_0: Physical health and mental health status are independent.Against H_1: Physical and mental health status are associated.Contingency Table:

	Good MHS	Bad MHS
Good PHS	213	207
Bad PHS	57	86

Decision:

χ^2 = 4.6101

$\chi^2_{1,0.05}$ = 3.841^{calc}

Hence, we reject H_0 at 5% level of significance.

Conclusion: Physical health score and mental health score are associated at 5% levelof significance.

A.2) Checking independence of exercising habit and self-confidence level

Exercise score (ES) as well as confidence score (CS) greater than or equal to 3 isconsidered to be good.

Hypothesis:

H_0: Exercising and self-confidence level are independent. Against H_1: Exercising and self-confidence level are associated.Contingency Table:

	Good CS	Bad CS
Good ES	120	35
Bad ES	284	140

Decision:

χ^2_{calc} = 5.3797

$\chi^2_{1,0.05}$ = 3.841

Hence reject H_0 at 5% level of significance

Conclusion: Exercise hours and confidence levels are associated.

A.3) Checking independence of exercising habit and obesity

Exercise score (ES) greater than or equal to 3 has been considered to be good.Hypothesis:

H_0: Exercising and obesity are independent. Against H_1: Exercising and obesity are associated. Contingency Table:

	Good ES	Bad ES
Obesity Present	17	25
Obesity absent	127	399

Decision:

χ^2_{calc} = 4.6527

$\chi^2_{1,0.05}$ = 3.841

Hence reject H0 at 5% level of significance Conclusion: Exercise hours and Obesity is associated.

A.4) Checking independence of junk food intake level and PCOD

Junk food frequency more than thrice a week is considered as high.This test is conducted among female participants.

Hypothesis:

H0: Junk food intake level and PCOD are independent. Against H1: Junk food intake level and PCOD are associated.Contingency Table:

	Low JFL	High JFL
PCOD Present	17	11
PCOD absent	192	87

Decision:

χ^2_{calc} = 0.44114

$\chi^2_{1,0.05}$ = 3.841

Hence, we accept H0 at 5% level of significance.

Conclusion: Junk food intake and presence of PCOD may be independent at 5% levelof significance.

Note: Above obtained result is contradictory to what doctors have proven till date. This may indicate that the junk food levels considered as “bad” in this test are not badenough to result into PCOD detection.

A.5) Checking independence of high BP and feeling worried

Hypothesis:

H0: High blood pressure and worriedness are independent. Against H1: High blood pressure and worriedness are associated.Contingency Table:

	Feeling worried	Not feeling worried
High BP	21	27
Normal BP	256	273

Decision:

χ^2_{calc} = 0.21684

$\chi^2_{1,0.05}$ = 3.841

Hence accept H0 at 5% level of significance

Conclusion: High blood pressure and worriedness may be independent at 5% level ofsignificance.

A.6) Checking independence of work hours and tiredness

Hypothesis:

H_0: Work hours and level of tiredness are independent. Against H_1: Work hours and level of tiredness are associatedContingency Table:

	0-3	**3.5-6**	**6.5-9**	**9.5-12**	**12.5-15**
1	0	1	8	9	0
2	3	10	42	26	7
3	6	11	32	15	2
4	5	17	48	39	0
5	2	5	22	10	1

Decision:

χ^2_{calc} = 21.901

$\chi^2_{16,0.05}$ = 26.296

Hence accept H_0 at 5% level of significance

Conclusion: Work hours and feeling of being tired is independent of each otherSimilar result found out for study hours.

A.7) Checking independence of beverage and sleep hours

Hypothesis:

H_0: Beverage choice and hours of sleep are independent. Against H_1: Beverage choice and hours of sleep are associatedContingency Table:

	3-4.9	5-6.9	7-8.9	9-10
Coffee	2	19	39	2
Tea	5	106	205	17
Milk	1	23	49	4
Fresh fruit juice	3	13	19	3

Decision:

χ^2 = 9.1345

χ^2_{cal} = 16.919

$\chi^2_{9,0.05}$

Hence accept H_0 at 5% level of significanceConclusion:

Beverage type and daily sleep hours are independent of each other.

A.8) Checking independence of beverage and sleep time

Hypothesis:

H_0: Beverage choice and time of sleep are independent Against H_1: Beverage choice and time of sleep are associatedContingency Table:

	coffee	tea	milk	Fresh fruit juice
Before 12 am	38	214	59	23
After 12 am	36	118	18	15

Decision:

χ^2_{calc} = 10.705

$\chi^2_{3,0.05}$ = 7.815

Hence accept H_0 at 5% level of significance Conclusion: Beverage type and sleep time is associated.

ODDS RATIO

The odds ratio is a statistical measure used to compare the odds of an event occurring in onegroup to the odds of the same event occurring in another group. It is commonly used in epidemiology and other fields to measure the association between two categorical variables,such as exposure to a risk factor and the occurrence of a disease.

		Event	
		Yes	No
Exposure	Yes	a	B
	No	c	d

$$Odds\ ratio = \frac{odds\ of\ the\ events\ in\ exposed\ group}{odds\ of\ the\ event\ in\ non\text{-}exposed\ group}$$

$$Odds\ ratio = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Interpretation of odds ratio:

1. OR >1 indicates increased occurrence of an event in the group of interest
2. OR <1 indicates decreased occurrence of an event in the group of interest
3. OR = 1 indicates equal occurrence of an event in both groups

Sr. No.	Attribute A	Categories of A	Attribute B	Categories of B	OddsRatio	Interpretation
1	Junk food intake	More junk food intake	Cholesterol	Has cholesterol	1.263562	Adults with higher junk food intake are 1.26 times more likely to develop cholesterol.
		Less junk food intake		Does not have cholesterol		
2	Insomnia	Has insomnia	Screen time	High screen time	1.392433	Adults with higher screen time are 1.39 times more
		Does not have		Low screen time		

		insomnia				likely to develop insomnia.
3	Diabetes	Has diabetes	Junk food intake	More junk food intake	1.456388	Adults with more junk food intake are 1.45 timesmore prone to developing diabetes.
		Does nothave diabetes		Less junk food intake		

CONCLUSIONS

1. Some of the results of our project were surprising, while the others weresatisfactorily close to expectation.
2. Even though we hear people complaining about not getting enough sleep,our population parameter for average sleep duration turned to be 7 hours.
3. As expected, hours of work and tiredness are associated, exercising habit and obesity are associated, exercising and self-confidence are associated and most importantly, physical and mental health status are associated.
4. Surprisingly, feeling of worriedness and high blood pressure may be independent, choice of beverage and hours of sleep may also be independent.
5. Water intake level of people with good physical health quality is significantly higher than that of people with lower physical health quality.
6. Daily water intake, hours of sleep, screen time, no. of cigarettes smoked, hours of exercise and junk food intake are the most significant factors for determining the physical health of a person.
7. Hours of study/work, screentime, hours of exercise turned out to be the most significant factors for determining the mental health of a person – all of which seem quite logical.

References

1. **Variables** **Source**
2. Daily WaterIntake https://rb.gy/o2j4t
3. Daily Sleep inHrs. https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html
4. Daily screentime Level recommended by an ophthalmologist
5. Daily freetime https://www.self.com/story/free-time-happiness
6. Smoking /Drinking Level recommended by a general physician
7. Daily Study /Work hrs. Study Time: https://rb.gy/jyuzmWork: https://rb.gy/2p35q
8. Weekly exercise hrs. https://rb.gy/wm7pj
9. Frequency ofjunk food eaten a week Level recommended by a general physician
10. Diabetes patients' data https://idf.org/