BDAT 1005
Math For Data Analytics

Image from https://seleritysas.com/blog/2019/12/06/preparing-for-the-future-of-data-analytics/ downloaded Jan 2023

Dataset Exploration – Part 4

PROJECT – REPORT

Document Summary

Document Item	Status
Document Title	Project Report on Dataset Exploration Part 4
Date Last Modified	4-April-2023
Status	Final
Document Description	This document provides a detailed analysis of Covid 19 survey student response dataset – Dataset Overview, Data Dictionary, Research Questions, Univariate Analysis, Coding, Data-cleaning, Hypothesis Testing, Inferential Techniques (interpolation & extrapolation), Tracking status.

Created By

Name & SID
Ruchika Gupta
SID: 200559617

Acknowledgement

I'm overwhelmed with gratitude and humility for everyone who has helped me translate these concepts into something concrete that is much above the level of simple.

I want to express my sincere gratitude to my professor Dr. Ehsan Pourjavad for giving me the chance to complete this excellent project on the subject of "Data Exploration". This project also enabled me to conduct extensive research and learn about a variety of new topics. I truly appreciate them.

Any effort, no matter the degree, cannot be successfully performed without my daughter's support and my husband's guidance.

I want to thank my daughter for letting me spend hours gathering data and a variety of facts. Despite his busy schedule, my spouse gave me various suggestions for how to make this endeavor stand out.

Thanking You

Ruchika Gupta

BDAT 1005-01

Table of Contents

Introduction

COVID-19 and Its Impact on Students

Due to my expertise in education, I am constantly worried about how well students are learning, which is what motivates me to do this analysis.

The COVID-19 epidemic has had a significant impact on students' lives. Their way of life has drastically changed because of switching from traditional classroom instruction to online education during the lockdown.

All students, regardless of age group, have been impacted by this quick evolution on such a huge scale. Students' academic performance, social lives, and mental health would all be significantly impacted by the disease's spread, travel bans, and the shutdown of educational facilities across the nation. The COVID-19 pandemic has had a greater negative impact on kids from less fortunate families.

I will investigate how the COVID-19 pandemic might affect the lives of students in this analysis. I'll try to determine whether there is a significant disparity between ambitions and the actual adoption of these online education rules at the local level. Additionally, I'll try to evaluate the mental health of pupils in various age groups using a variety of criteria, such as sleeping patterns, a daily exercise schedule, and social support. In addition, I will examine several coping techniques employed by pupils to handle the current circumstance.

Dataset Overview

Covid19 and its impact on students by Kaggle

(https://www.kaggle.com/kunal28chaturvedi/covid19-and-its-impact-on-students)

1182 students from various age groups and educational institutions in Delhi National Capital Region serve as the sample size for a cross-sectional survey that is being undertaken for this analysis.

This dataset, which is made up of 19 distinct variables, shows students':

- General demographics, such as age and place of residence.
- Details on the daily schedule of online learning in Indian educational institutions after the changeover from offline learning, including average daily online study time (hours), the average daily online study time (hours), and the average daily self-study time (hours).
- Evaluation of the online learning experience to determine the degree of student satisfaction.
- Evaluation of health because of lifestyle changes: average daily sleep time (hours), change in weight, average daily exercise time (hours), number of meals per day, and other factors including family cohesion and stress management techniques used during the epidemic.

Data Dictionary

This document will list the names, attributes and description of the various variables used in this dataset.

S.No.	Variables	Categorical / Numerical	Type	Missing/ Invalid Values	Description
1	**ID**	Categorical	Nominal Polytomous	0	Unique alpha numeric word represents the identity of a student
2	**Region of residence**	Categorical	Nominal Dichotomous	0	represents region where student resides
3	**Age of Subject**	Numerical	Ratio Discrete	0	It shows the age of the student
4	**Time spent on Online Class**	Numerical	Ratio Continuous	0	It shows number of **hours** a student spent on Online classes
5	**Rating of Online Class experience**	Categorical	Ordinal Polytomous	24	represents the level of satisfaction a student had with Online Classes
6	**Medium for online class**	Categorical	Nominal Polytomous	51	shows the device used by a student to attend online classes
7	**Time spent on self-study**	Numerical	Ratio Continuous	0	It displays number of **hours** a student spent on Self learning during Covid 19
8	**Time spent on fitness**	Numerical	Ratio Continuous	0	It shows number of **hours** a student gave for his/her fitness during pandemic
9	**Time spent on sleep**	Numerical	Ratio Continuous	0	It displays number of **hours** a student took the sleep during pandemic
10	**Time spent on social media**	Numerical	Ratio Continuous	0	It shows number of **hours** a student spent on Social media platforms during epidemic era
11	**Preferred social media platform**	Categorical	Nominal Polytomous	0	It shows, which social media platform liked by a student most
12	**Time spent on TV**	Numerical	Ratio Continuous	0	It displays number of **hours** a student spent in watching TV throughout Covid 19 era
13	**Number of meals per day**	Numerical	Ratio Discrete	0	It shows how many times a student took his/her meal
14	**Change in your weight**	Categorical	Ordinal Polytomous	0	represents change in the weight of a student during pandemic
15	**Health issue during lockdown**	Categorical	Nominal Dichotomous	0	shows whether a student faced any health issue or not during lockdown

16	**Stress busters**	Categorical	Nominal Polytomous	0	Represents activities chosen by students to release their stress
17	**Time utilized**	Categorical	Nominal Dichotomous	0	shows whether that time was actually utilized by a student or not
18	**Do you find yourself more connected with your family, close friends , relatives ?**	Categorical	Nominal Dichotomous	0	displays if student felt more connected to their family and pals
19	**What you miss the most**	Categorical	Nominal Polytomous	1	shows what students missed the most during pandemic

Research Questions

If we do not understand what conclusions to draw from the data, the analysis will be ineffective. By interpreting a dataset incorrectly, we run the risk of making a poor choice regarding a crucial matter. Therefore, it is critical for the analyst to create solid (F.I.N.E.R) research questions.

For this analysis my research questions would be the following:

Question 1

What kind of learning environment do pupils currently have to deal with due to the epidemic?

Supporting Question:

- How pandemic affected the level of satisfaction among students?

- What are the psychological effects on students due to Covid 19 outbreak?

Question 2

In this epidemic era, how are students' online class situations?

Supporting Questions:

- What are the main worries about online learning?
- What is the level of availability of digital infrastructure and skill set for online learning during epidemic era?

Question 3

How are each school level's students behaving in terms of their health?

Supporting Question:

- How can the stress levels of children at each level of school be measured during the pandemic?

F.I.N.E.R Framework

Feasible: The above research questions are Feasible enough as their scope leads to an attainable objective using adequate pool of sampling.

Interesting: The above research questions are for sure Interesting as they serve the immediate needs of education domain.

Novel: The above research questions are Novel as they will make meaningful contribution to the current situation of education domain.

Ethical: The above research questions are Ethical as there is no risk to participate under this analysis.

Relevance: The above research questions are Relevant as they guide future research endeavors.

Dependent / Independent Variables

My analysis will be based on the following factors, which are based on my research questions:

Independent Variable:

1. Student’s level of School

Dependent Variables:

1. Time spent on Online Class
2. Time spent on Self Study
3. Time spent on sleep
4. time spent on social media
5. Rating on Online Class
6. Student's self study vs social media orientation
7. Is there enough time for the student to study?
8. Does the student have a regular sleeping pattern?

9. Do students spend more time on social media and TV than on independent study and online classes?

Assumptions

1. Dummy variable: Level of School

Created a dummy variable “Level of School” classified by the age of students as per the need of research questions.

Classified students by age

0 - 11 (Elementary School Students)

12-14 (Junior High School Students)

15-17 (Senior High School Students)

18 > (College Students)

2. Removed Following Columns as they don’t have any relevance to the research questions

- Do you find yourself more connected with your family, close friends, relatives?

- What you miss the most

Coding

1. Rating of online class experience

Rating	Code
Average	3
Excellent	5
Good	4
Poor	2
Very poor	1

2. Medium for online class

Device	Code
Any Gadget	1
Laptop/Desktop	2
Smartphone	3
Smartphone or Laptop/Desktop	4
Tablet	5

3. Change in Weight

Weight	Code
Decreased	1
Increased	3

Remain Constant	2

4. Preferred social media platform

Social Media	Code
Elyment	1
Facebook	2
Instagram	3
Linkedin	4
None	5
Omegle	6
Quora	7
Reddit	8
Snapchat	9
Talklife	10
Telegram	11
Twitter	12
Whatsapp	13
Youtube	14

5. Health issues / Time Utilized

Health/Time Utilized	Code
No	0

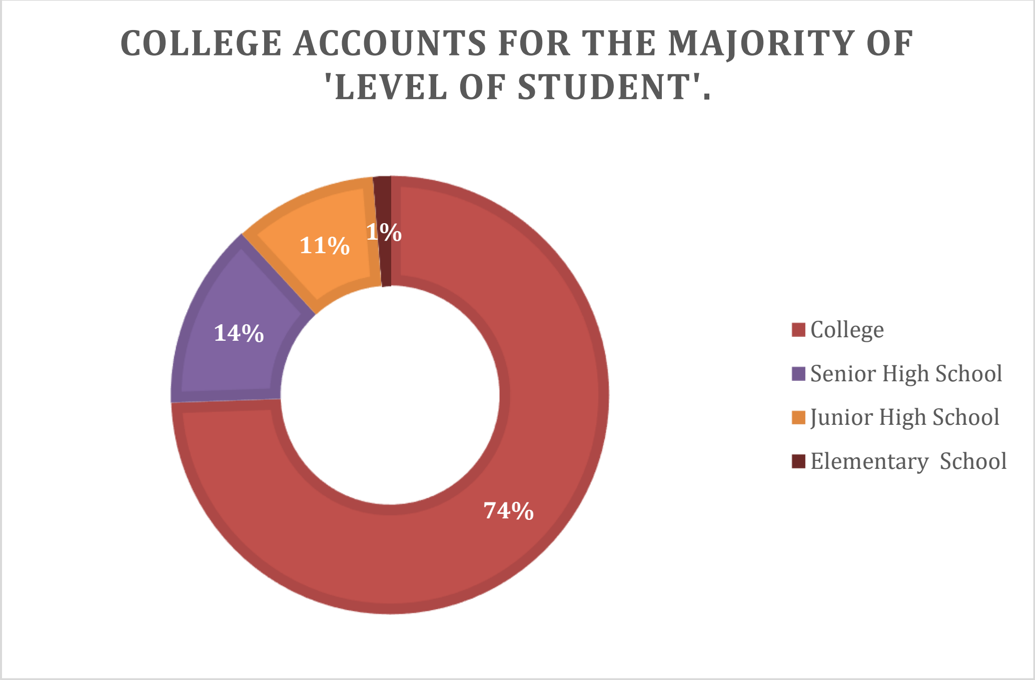
Yes	1

Univariate Analysis

a) Categorical (Qualitative) Variables

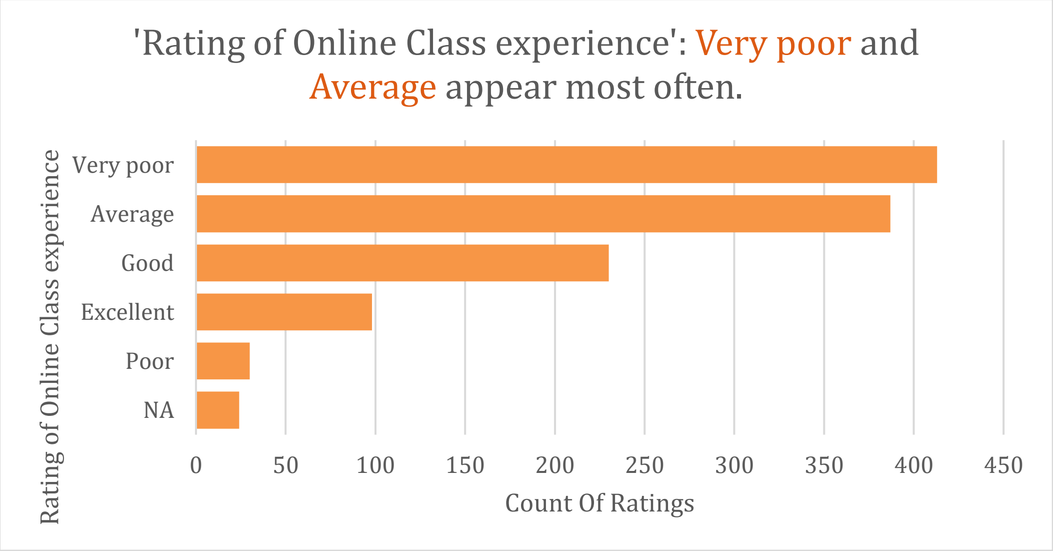
1. Level of School

Level of Student	**Count of Level of Student**
College	880
Senior High School	162
Junior High School	125
Elementary School	15
Grand Total	**1182**


COLLEGE ACCOUNTS FOR THE MAJORITY OF 'LEVEL OF STUDENT'.

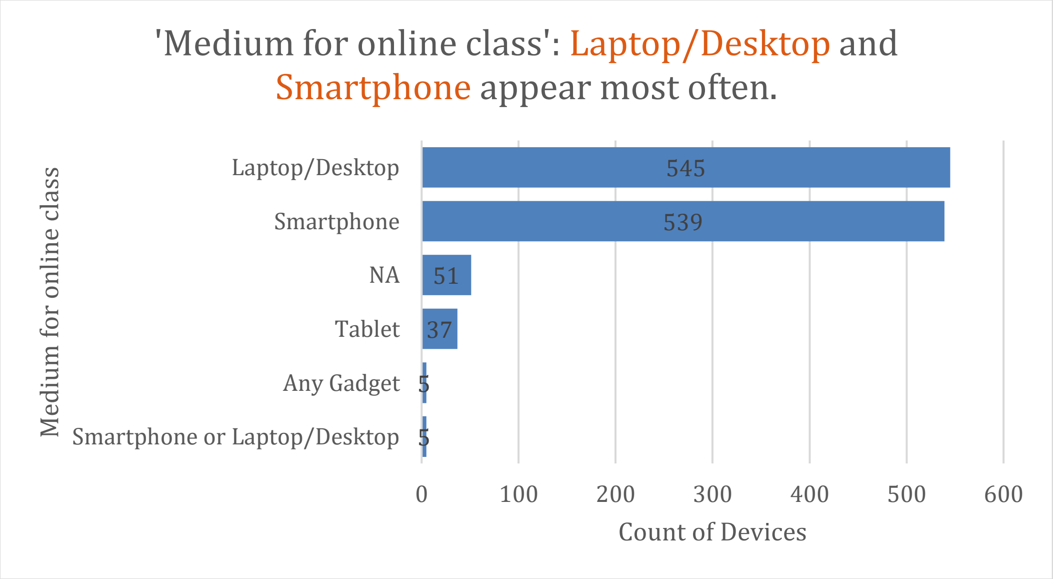
2. Rating of online class experience

Rating of Online Class experience	Count of Rating of Online Class experience
Very poor	413
Average	387
Good	230
Excellent	98
Poor	30
NA	24
Grand Total	**1182**



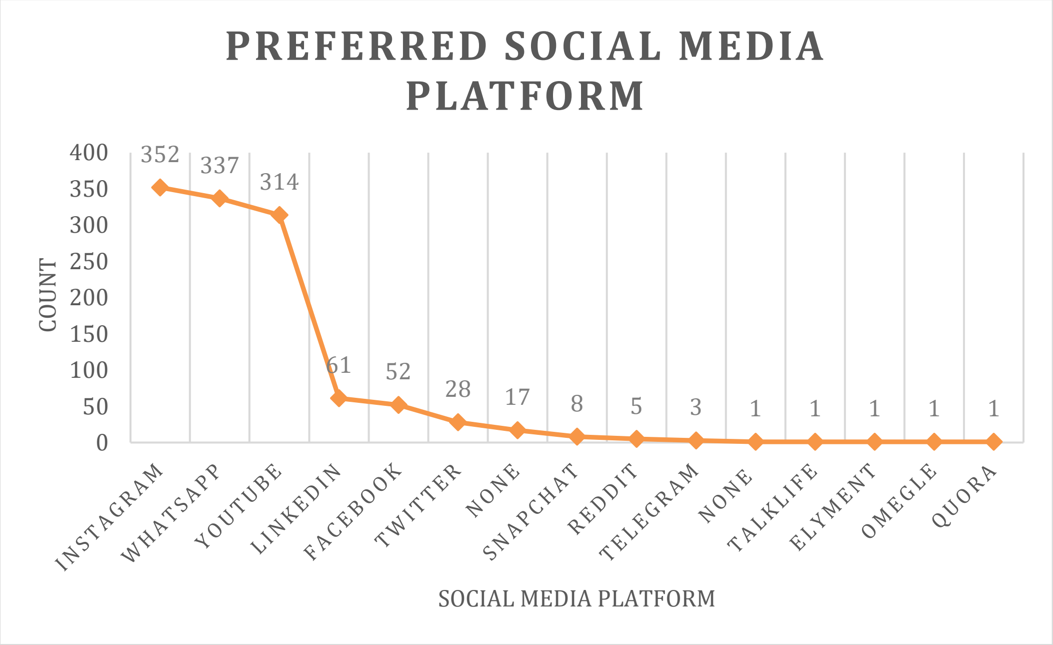
3. Medium for online class

Medium for online class	Count of Medium for online class
Laptop/Desktop	545
Smartphone	539
NA	51
Tablet	37
Any Gadget	5
Smartphone or Laptop/Desktop	5
Grand Total	**1182**



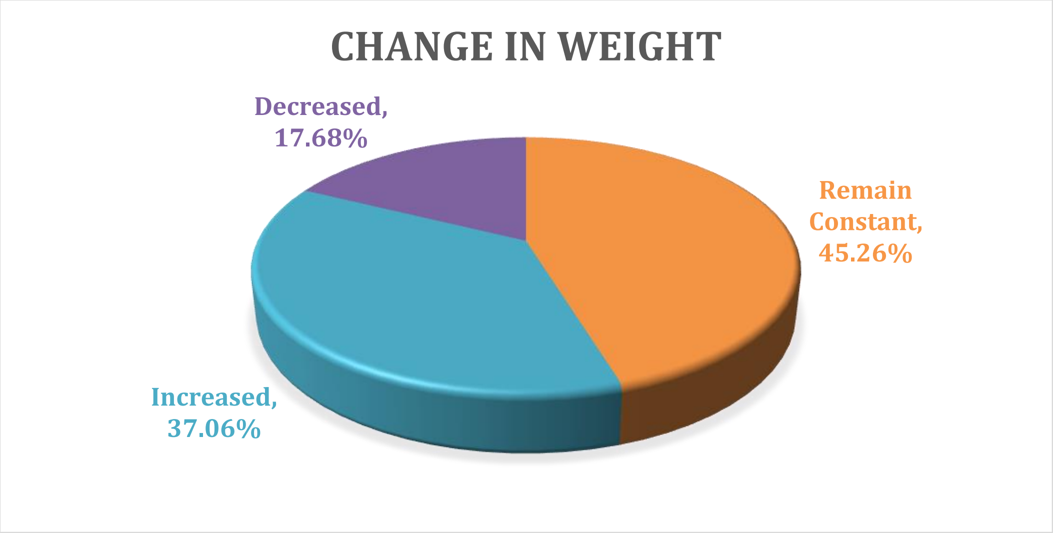
4. Preferred social media platform

Preferred social media platform	Count of Preferred social media platform
Instagram	352
Whatsapp	337
Youtube	314
Linkedin	61
Facebook	52
Twitter	28
None	17
Snapchat	8
Reddit	5
Telegram	3
None	1
Talklife	1
Elyment	1
Omegle	1
Quora	1
Grand Total	**1182**



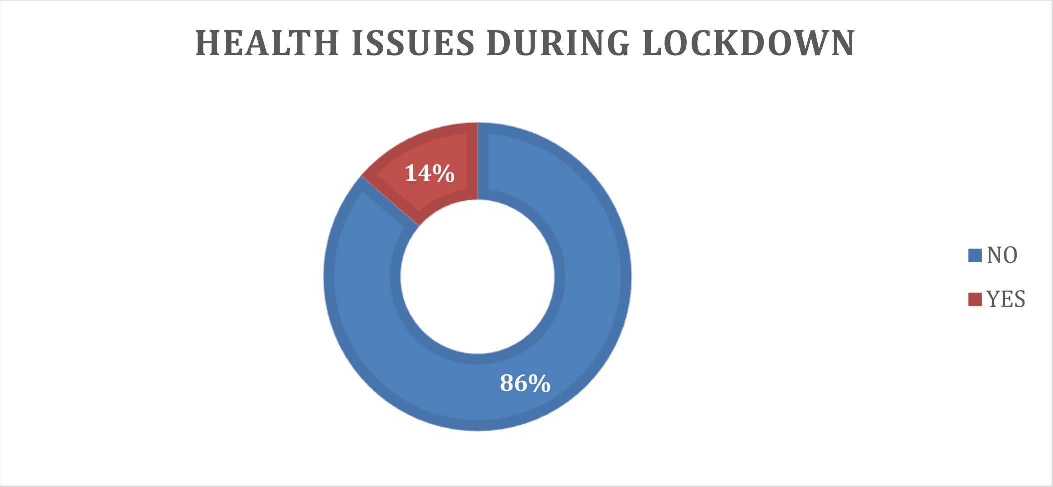
5. Change in Weight

Change in your weight	Count of Change in your weight
Remain Constant	45.26%
Increased	37.06%
Decreased	17.68%
Grand Total	**100.00%**



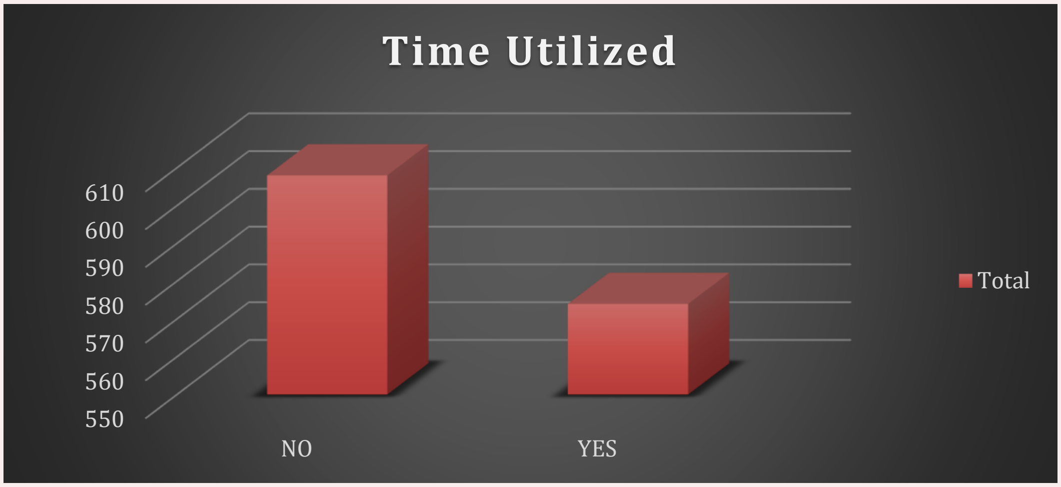
6. Health issues during Lockdown

Health issue during lockdown	Count of Health issue during lockdown
NO	1021
YES	161
Grand Total	**1182**



7. Time Utilized

Time utilized	Count of Time utilized
NO	608
YES	574
Grand Total	**1182**

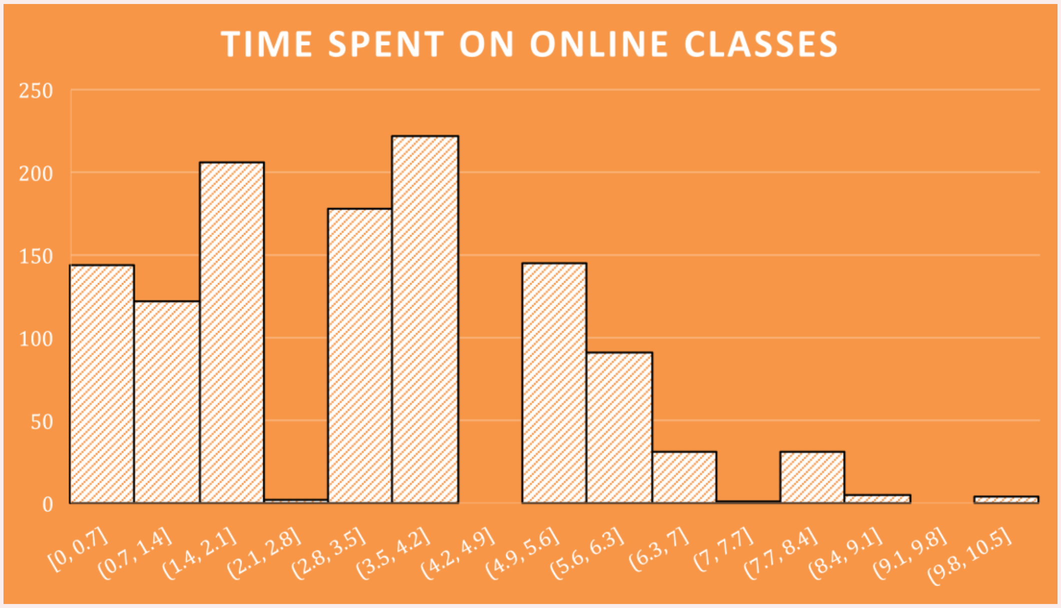


b) Numerical Variables

1. Time spent on online classes

Time spent on Online Classes	
Mean	3.208840948
Standard Error	0.061132705
Median	3
Mode	4
Standard Deviation	2.101756274
Sample Variance	4.417379434
Kurtosis	-0.281460036
Skewness	0.366112449
Range	10
Minimum	0
Maximum	10
Sum	3792.85
Count	1182

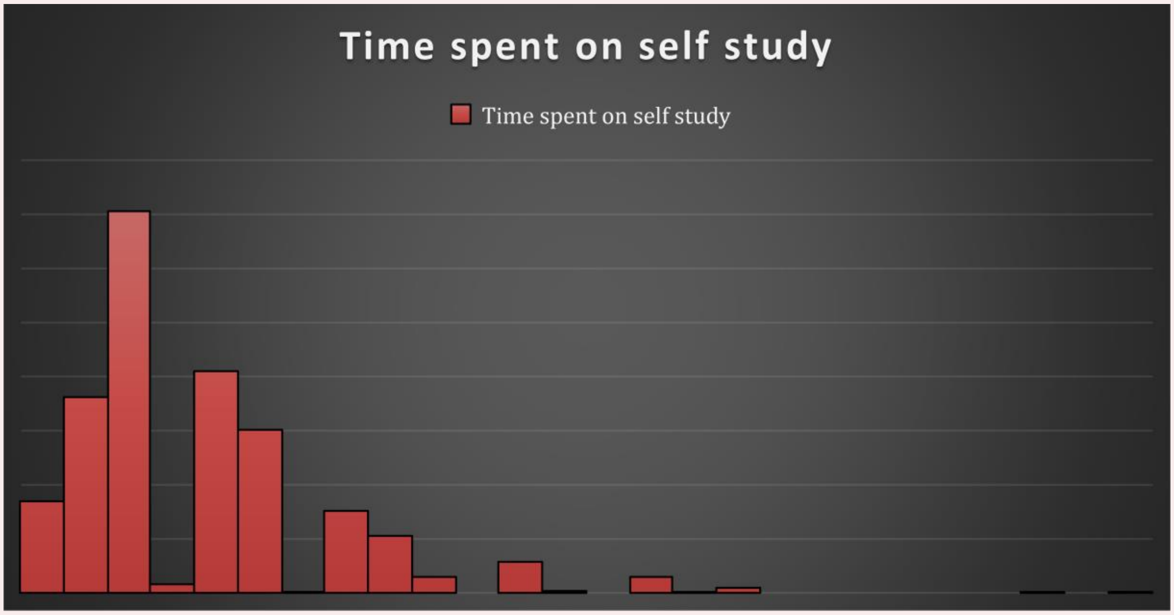
- **High Variance**
- **Positive Skewness**
- **Low Kurtosis**



2. Time spent on self-study

Time Spent on Self study	
Mean	2.911590525
Standard Error	0.062262247
Median	2
Mode	2
Standard Deviation	2.140590185
Sample Variance	4.582126342
Kurtosis	5.451284331
Skewness	1.732713865
Range	18
Minimum	0
Maximum	18
Sum	3441.5
Count	1182

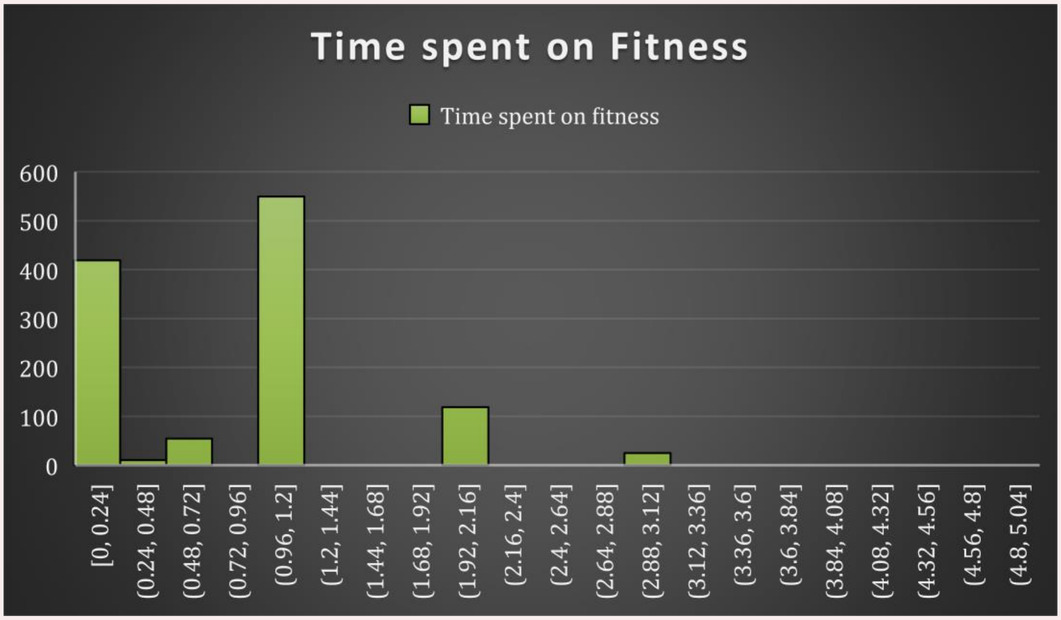
- **High Variance**
- **Positive Skewness**
- **High Kurtosis**



3. Time spent on fitness

Time Spent on fitness	
Mean	0.765820643
Standard Error	0.021071748
Median	1
Mode	1
Standard Deviation	0.724451472
Sample Variance	0.524829936
Kurtosis	1.582343265
Skewness	0.968172909
Range	5
Minimum	0
Maximum	5
Sum	905.2
Count	1182

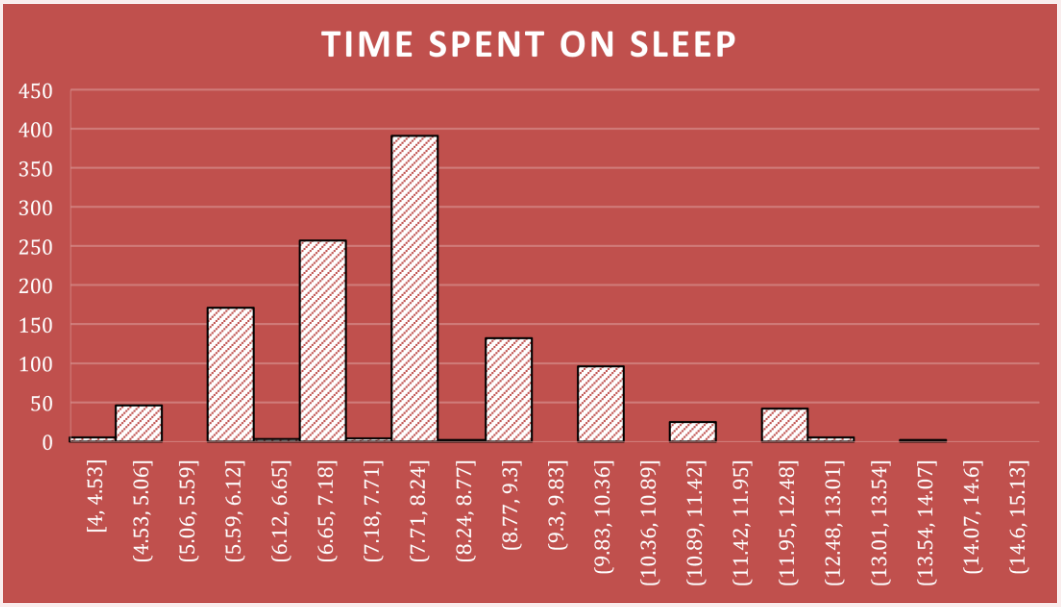
- **Low Variance**
- **Positive Skewness**
- **Low Kurtosis**



4. Time spent on sleep

Time Spent on Sleep	
Mean	7.871235195
Standard Error	0.046996846
Median	8
Mode	8
Standard Deviation	1.61576222
Sample Variance	2.61068755
Kurtosis	1.007620845
Skewness	0.735499394
Range	11
Minimum	4
Maximum	15
Sum	9303.8
Count	1182

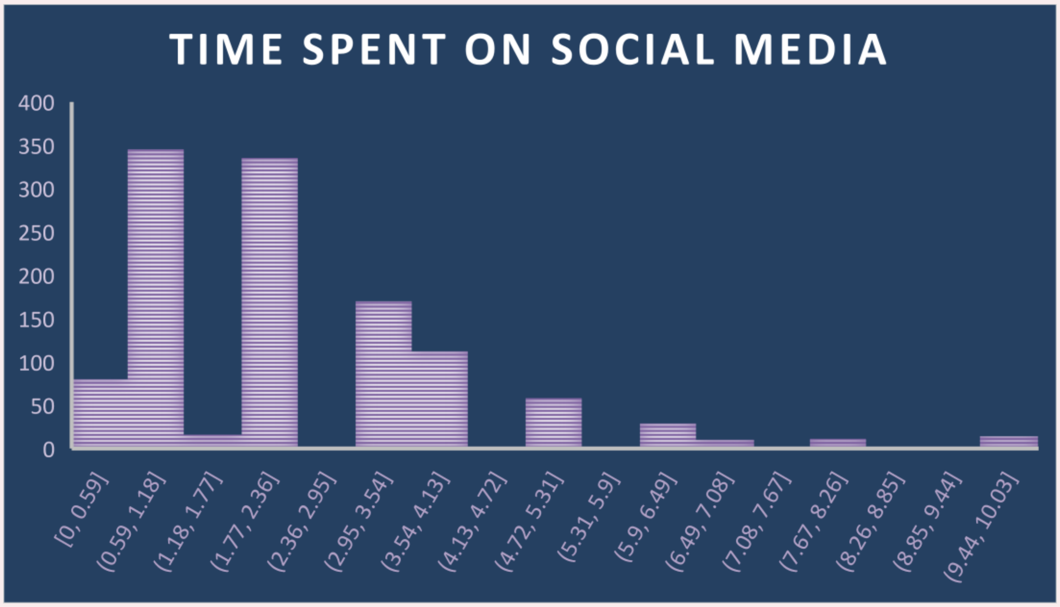
- **Low Variance**
- **Positive Skewness**
- **Low Kurtosis**



5. Time spent on social media

Time Spent on Social media	
Mean	2.365693739
Standard Error	0.051405599
Median	2
Mode	1
Standard Deviation	1.767336142
Sample Variance	3.123477037
Kurtosis	3.86399237
Skewness	1.699306818
Range	10
Minimum	0
Maximum	10
Sum	2796.25
Count	1182

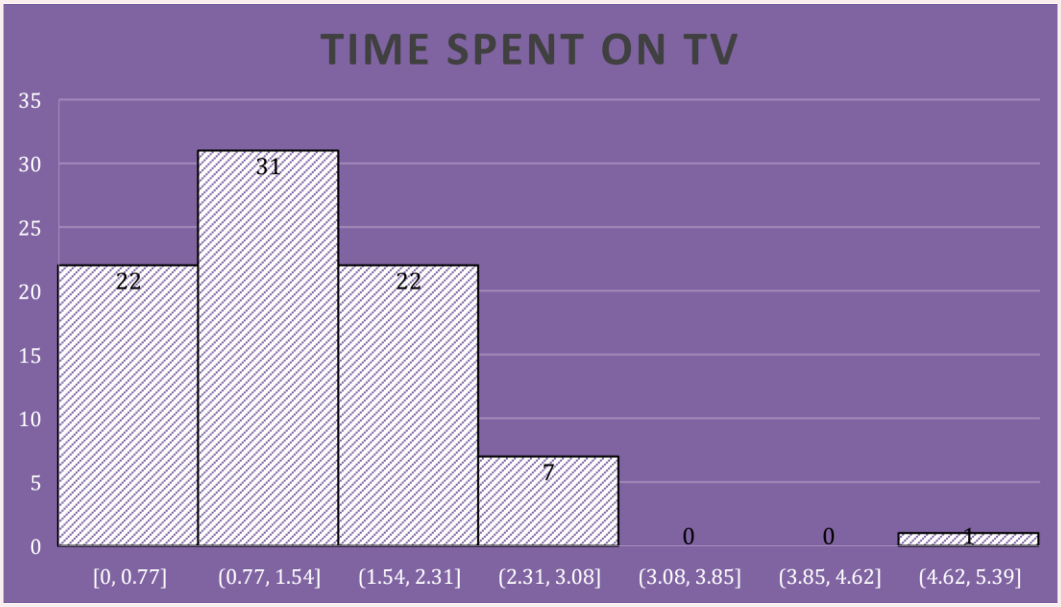
- **High Variance**
- **Positive Skewness**
- **Low Kurtosis**



6. Time spent on TV

Time Spent on TV	
Mean	1.021573604
Standard Error	0.036766156
Median	1
Mode	0
Standard Deviation	1.26402877
Sample Variance	1.597768733
Kurtosis	16.05917791
Skewness	2.690825489
Range	15
Minimum	0
Maximum	15
Sum	1207.5
Count	1182

- **Low Variance**
- **Positive Skewness**
- **High Kurtosis**



Data Cleaning

a) Handling of Missing Values or Invalid Values

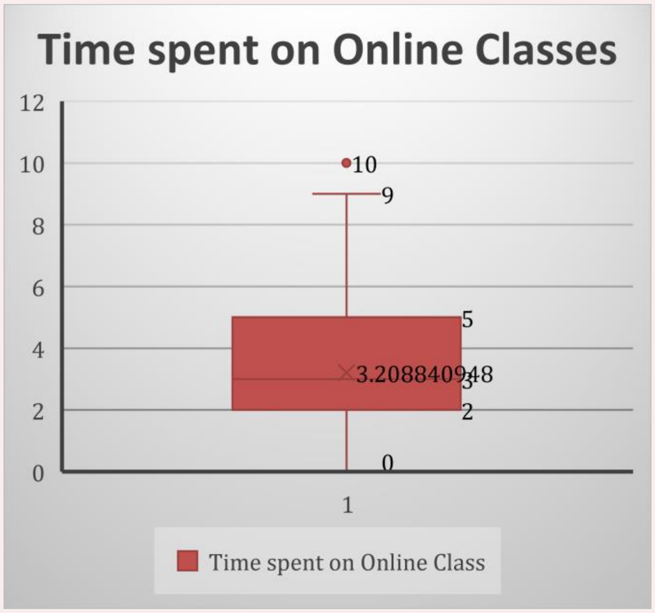
Variable	Missing Value
Rating of Online Class Experience	24
Medium of Online Class	51

Since missing values are less than 25% and these are categorical variables, so these values are **replaced by mode (most frequent value).**

b) Handling of Outliers

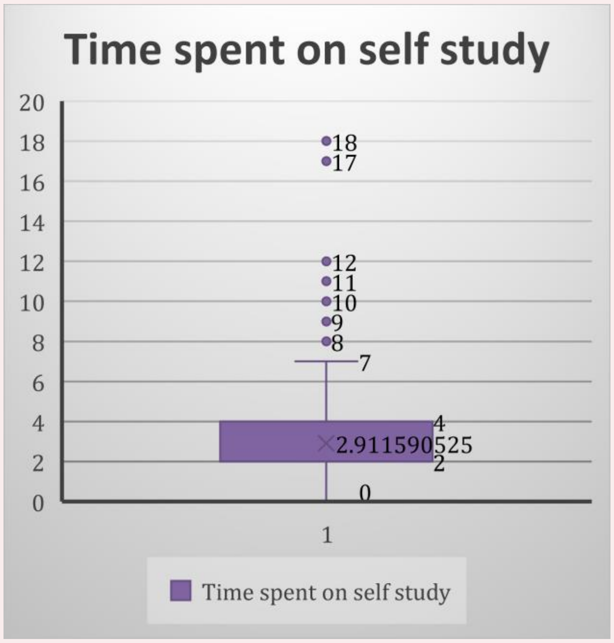
1. Time spent on online classes

IQR	3
3Q + 1.5IQR	9.5
Outlier	**Positive**



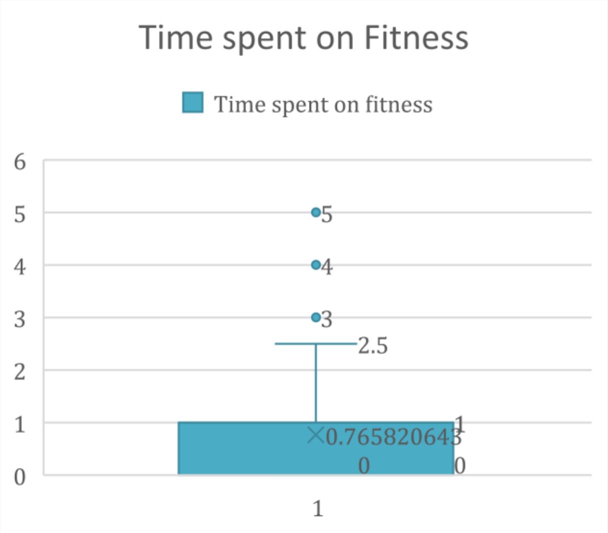
2. Time spent on self-study

IQR	2
3Q + 1.5IQR	7
Outlier	**Positive**



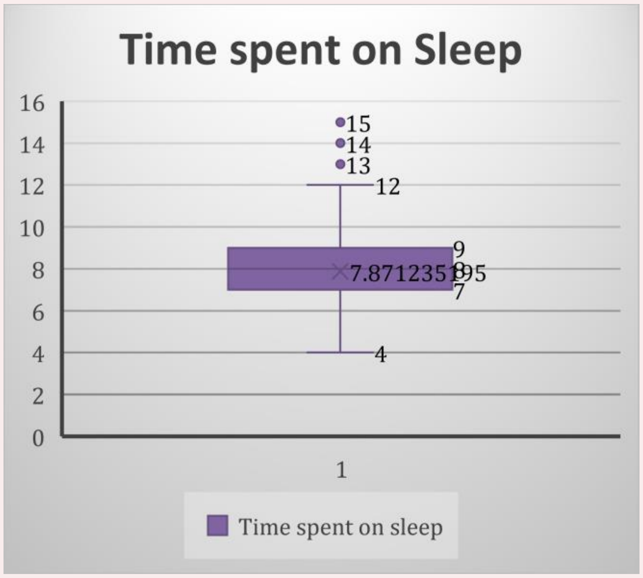
3. Time spent on fitness

IQR	1
3Q + 1.5IQR	2.5
Outlier	**Positive**



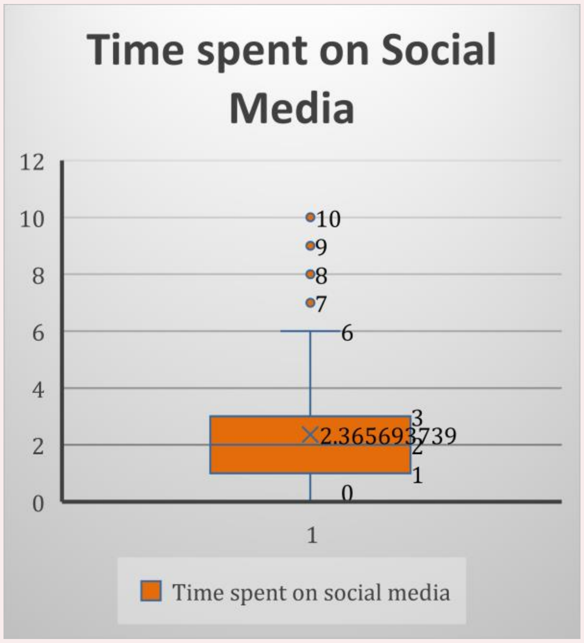
4. Time spent on sleep

IQR	2
3Q + 1.5IQR	12.5
1Q - 1.5IQR	3.5
Outlier	**Positive and negative**



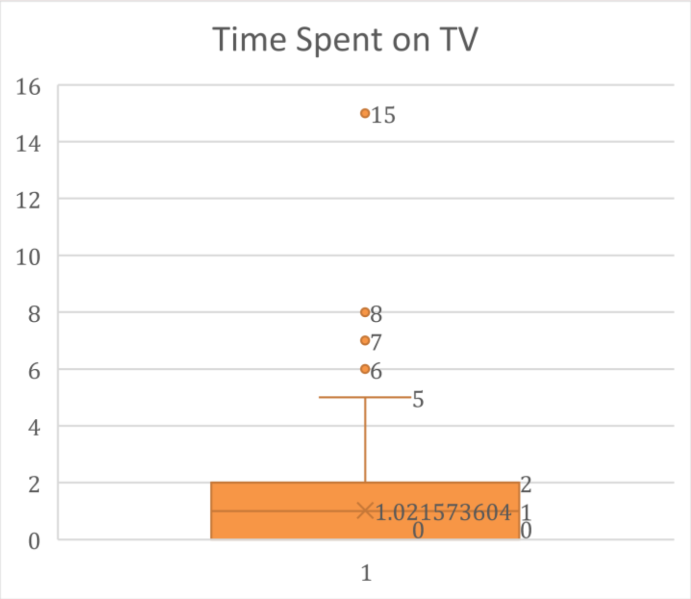
5. Time spent on social media

IQR	2
3Q + 1.5IQR	6.5
Outlier	**Positive**

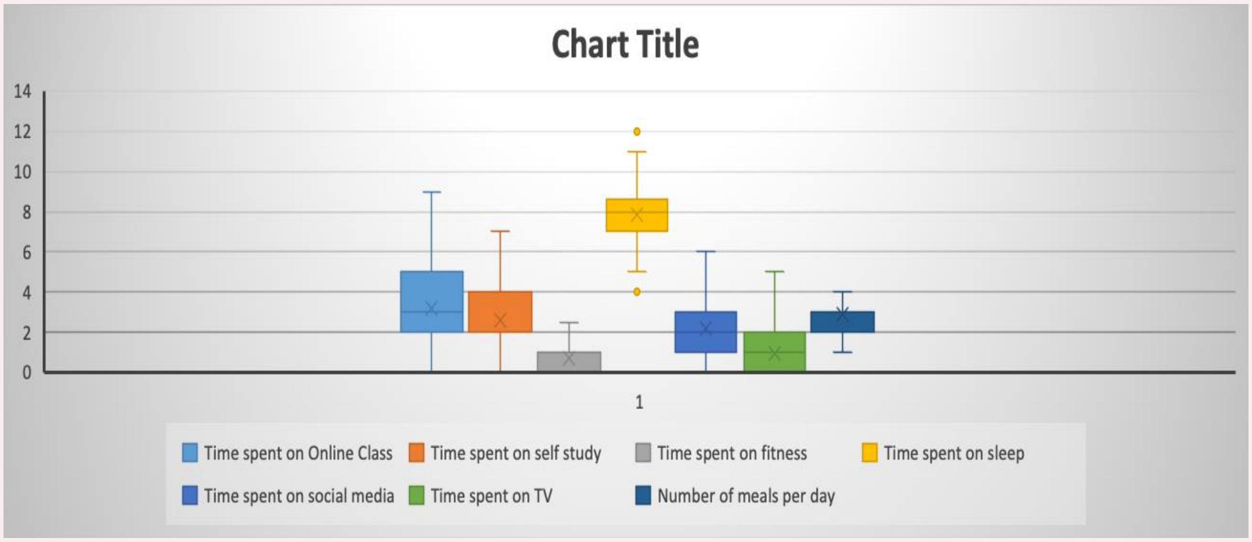


6. Time spent TV

IQR	2
3Q + 1.5IQR	5.5
Outlier	**Positive**



Data After Cleaning



Hypothesis Development

Research Question:

In this epidemic era, how are students' online class situations?

1. Hypothesis:

Time spent by a student on online classes during lockdown has a significant impact on the actual time utilization of the student.

Null Hypothesis (H_0):

There is no significant difference among time spent on online classes and actual time utilised by the student.

Research Question:

How are each school level's students behaving in terms of their learning?

2. Hypothesis:

Time spent by a student on social media during lockdown will depend on the school/college level of the student.

Null Hypothesis (H_0):

Level of a student have no impact on time spent by him/her on social media platforms.

---

3. Hypothesis:

Time spent by a student on his/her fitness during lockdown will be a factor for the occurrence of health issues.

Null Hypothesis (H_0):

There is no significant impact on health due to time spent on fitness and mean difference among groups of health issues occurred by chance.

Research Question:

How pandemic affected the level of satisfaction among students?

4. Hypothesis:

Actual time utilised by a student during pandemic is effected by the occurrence of health issues with the student.

Null Hypothesis (H_0):

There is no significant difference among actual time utilisation and occurrence of health issues.

---

5. Hypothesis:

Region of residence of a student is one the cause for the occurrence of health issues in student.

Null Hypothesis (H_0):

Region of residence has no impact on the occurrence of health issues.

Hypothesis Testing

1. Hypothesis:

Time spent by a student on online classes during lockdown has a significant impact on the actual time utilization of the student.

Null Hypothesis (H_0):

There is no significant difference among time spent on online classes and actual time utilised by the student.

This hypothesis consist one categorical (dichotomous) variable and one quantitative variable, also both groups have different number of observations and variance, therefore we will perform t-Test assuming unequal variance on it.

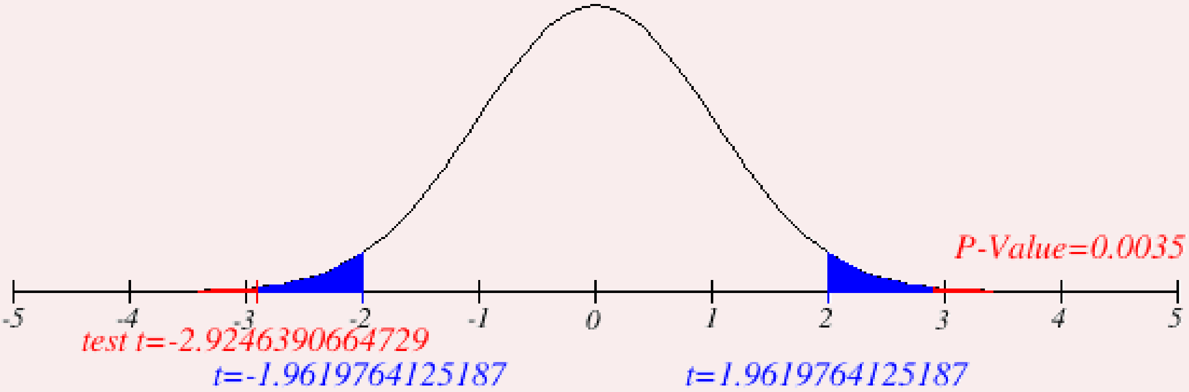
t- Test:

t-Test: Two-Sample Assuming Unequal Variances

	No	*Yes*
Mean	3.015460526	3.36489547
Variance	4.553096657	3.895593619
Observations	608	574
Hypothesized Mean Difference	0	
df	1180	
t Stat	-2.924639066	
P(T<=t) one-tail	0.001757236	
t Critical one-tail	1.646145977	
P(T<=t) two-tail	0.003514473	
t Critical two-tail	1.961976413	

Since p-value is less than alpha, i.e. p-value < 0.05, we will reject the null hypothesis.

Which means a student's actual time utilisation will get effected by his/her time spent on online classes.



2. Hypothesis:

Time spent by a student on social media during lockdown will depend on the school/college level of the student.

Null Hypothesis (H_0):

Level of a student have no impact on time spent by him/her on social Media platform.

Since this hypothesis consist one categorical (Polytomous) variable (Independent)and one quantitative variable(Dependent), we will perform ANOVA Test on it.

Note: Level of student is self-categorised variable using age.

ANOVA Testing:

Anova: SingleFactor

SUMMARY

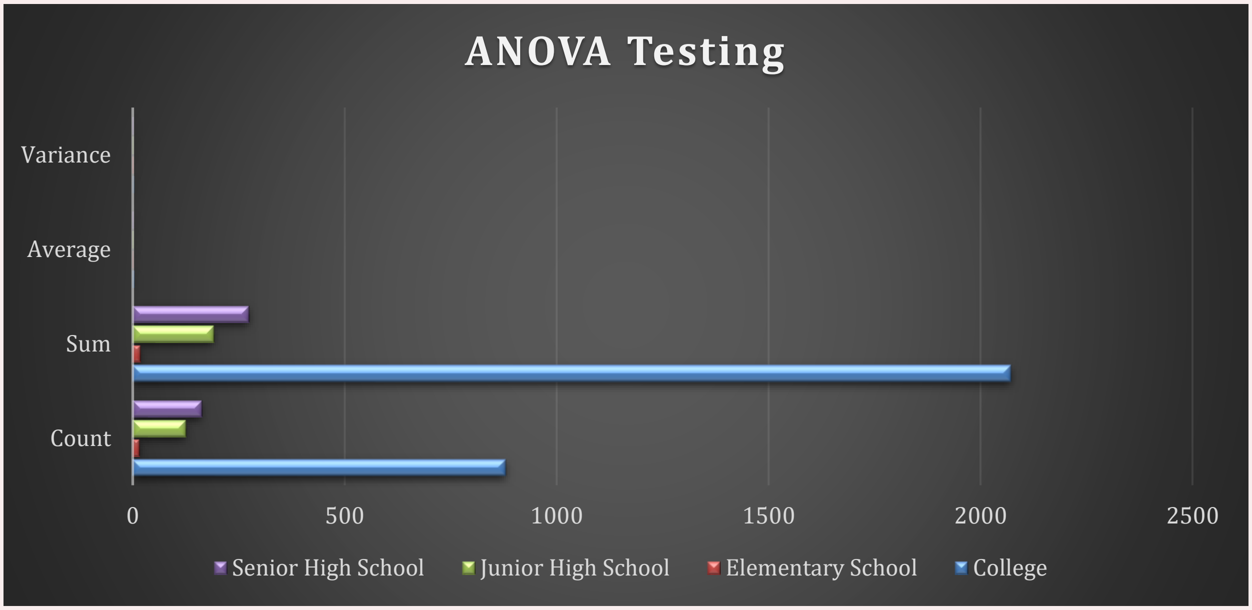
Groups	*Count*	*Sum*	*Average*	*Variance*
College	880	2071.9	2.35443182	1.85355827
Elementary School	15	17.5	1.16666667	0.70238095
Junior High School	125	191.05	1.5284	1.18469097
Senior High School	162	273.8	1.69012346	1.59356644

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	133.698693	3	44.566231	25.7023466	4.16024E-16	2.61245726
Within Groups	2042.57693	1178	1.73393627			
Total	2176.27562	1181				

Since p-value is less than alpha, i.e. p-value < 0.05, we will reject the null hypothesis.

Which means time spent by a student on social media during lockdown will be impacted by the school/college level of the student.



3. Hypothesis:

Time spent by a student on his/her fitness during lockdown will be a factor for the occurrence of health issues.

Null Hypothesis (H_0):

There is no significant impact on health due to time spent on fitness and mean difference among groups of health issues occurred by chance.

This hypothesis consist one categorical (dichotomous) variable and one quantitative variable, also both groups have different number of observations and variance, therefore we will perform t-Test assuming unequal variance on it.

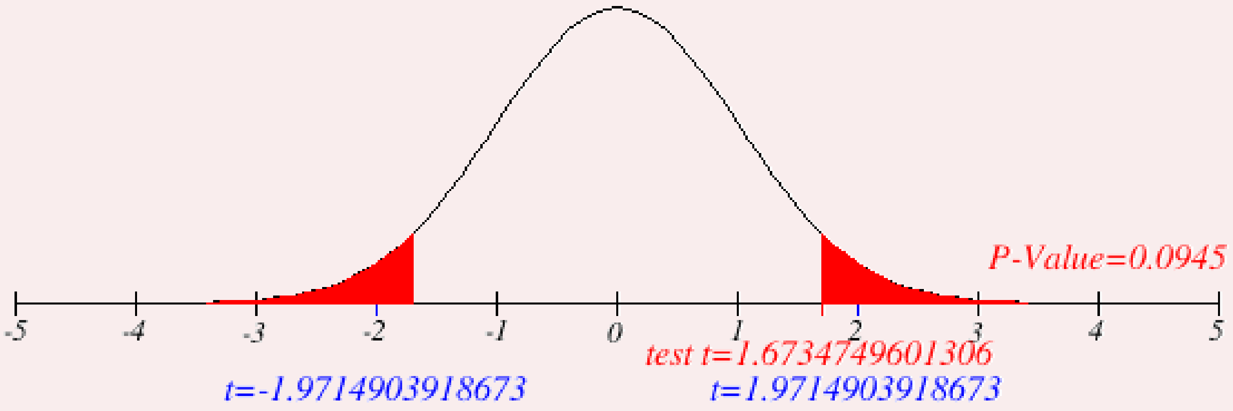
t- Test:

t-Test: Two-Sample Assuming Unequal Variances

	No	*Yes*
Mean	0.730264447	0.637267081
Variance	0.386646891	0.436227484
Observations	1021	161
Hypothesized Mean Difference	0	
df	207	
t Stat	1.67347496	
P(T<=t) one-tail	0.047872137	
t Critical one-tail	1.652248086	
P(T<=t) two-tail	0.095744273	
t Critical two-tail	1.971490392	

Since p-value is more than alpha, i.e. p-value > 0.05, we cannot reject the null hypothesis.

Which means there is no significant difference among health issues and time spent on fitness, i.e. mean differences occurred by chance.



4. Hypothesis:

Actual time utilised by a student during pandemic is effected by the occurrence of health issues with the student.

Null Hypothesis (H_0):

There is no significant difference among actual time utilisation and occurrence of health issues.

Since this hypothesis consist both categorical (dichotomous) variables, we will perform Chi-square Test on it.

Chi-square Testing:

Actual values

Count of ID	Time Utilized		
Health Issues	**YES**	**NO**	**Grand Total**
YES	59	102	161
NO	515	506	1021
Grand Total	**574**	**608**	**1182**

Expected values

Row Labels	Yes	No	Grand Total
Yes	78.18443316	82.81556684	161
No	495.8155668	525.1844332	1021
Grand Total	**574**	**608**	**1182**

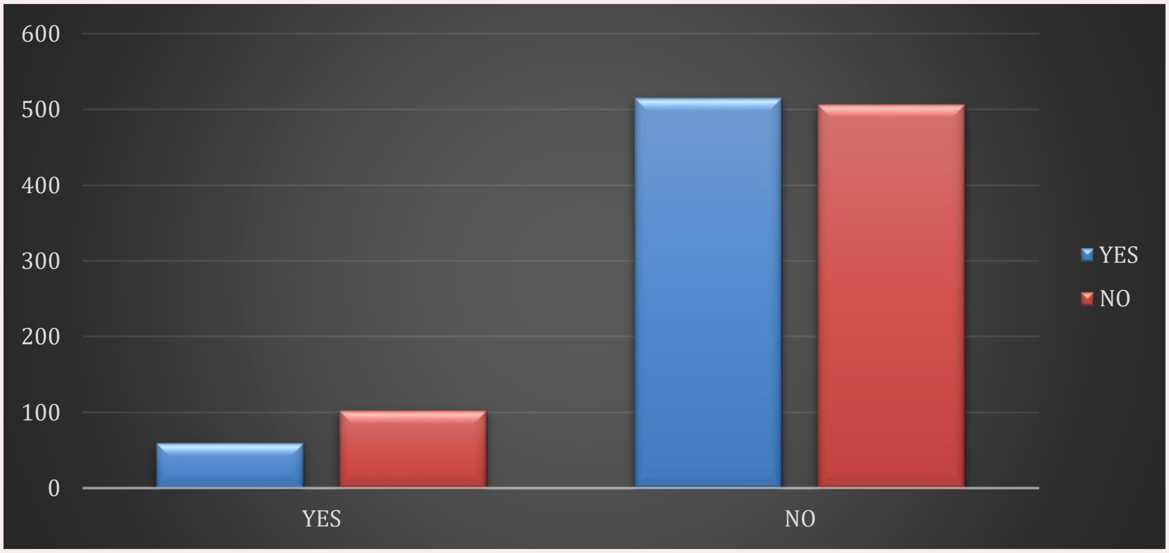
Chi-square

Row Labels	Yes	No	Grand Total
Yes	4.70736259	4.444121919	
No	0.742297137	0.7007871	
Grand Total			**10.59456875**

p-value for Chi-square	**0.001134204**

Since p-value is less than alpha, i.e. p-value < 0.05, we will reject the null hypothesis.

Which means if a student face a health issue during pandemic his/her actual time utilised will be effected.



5. Hypothesis:

Region of residence of a student is one the cause for the occurrence of health issues in student.

Null Hypothesis (H_0):

Region of residence has no impact on the occurrence of health issues.

Since this hypothesis consist both categorical (dichotomous) variables, we will perform Chi-square Test on it.

Odds Ratio Testing:

Count of ID	**Health Issues**		
Region of Residence	**NO**	**YES**	**Grand Total**
Delhi-NCR	635	86	721
Outside Delhi-NCR	386	75	461
Grand Total	**1021**	**161**	**1182**

Odds Ratio	(a*d)/(b*c)		**1.4346608**	*0.697028871*	*Inverted*

Confidence Interval of OR

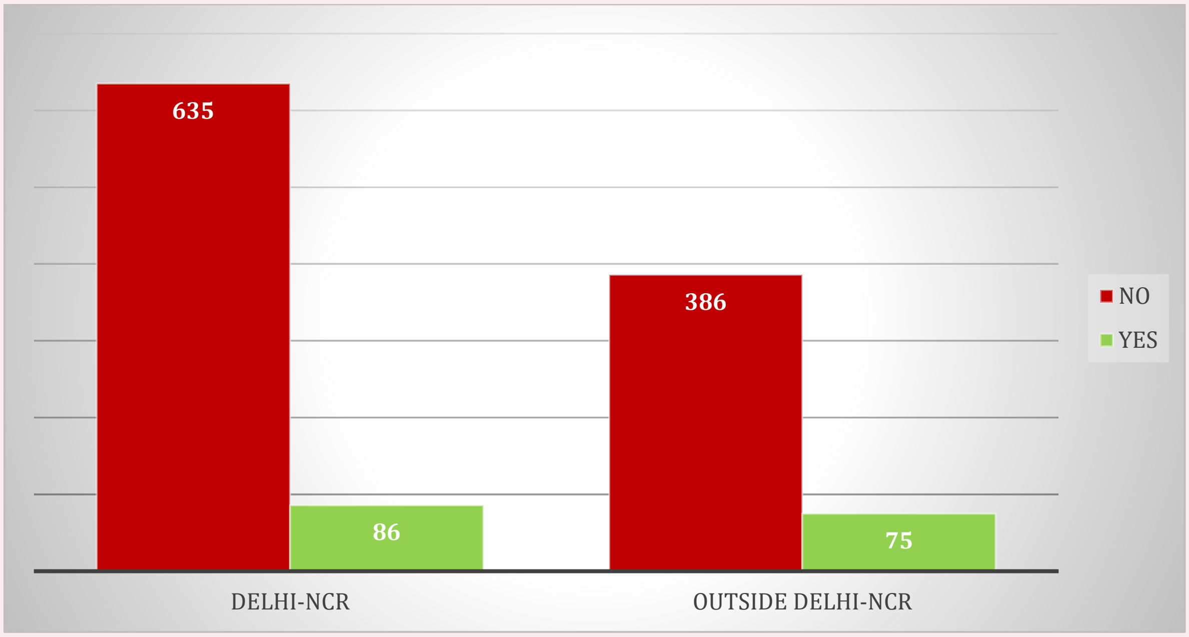
exp(ln(OR±($Z_{\alpha/2}$)*SE(ln(OR))) WHERE SE(ln(OR) is sqrt(1/a+1/b+1/c+1/d))

ln(OR) =	0.360928447	**SE(ln(OR)) =**	0.170665512
		for CL 95%, $Z_{\alpha/2}$ is	
$Z_{\alpha/2}$ =	1.96	*1.96 SD*	
Lower bound:	**1.026776253**	**Upper bound:**	**2.004576568**

Odds ratio = 1.43

Confidence Interval with 95% confidence: 1.02 – 2.00

Which means a student belongs to Delhi-NCR is 1.43 times more likely to have a health issue during lockdown.



Inferential Techniques

Inferential statistics allow us to make predictions based on observed data.

- Generalizing about a population based on sample data (interpolation)
- Predicting future results from historical data (extrapolation)

Here, I am using Multivariate Regression analysis on my dataset.

1. Linear Regression Analysis (Predictive)

In this analysis, I will find how different factors are affecting a student's time spent on self-study and later I will, make the prediction for this attribute.

Dependent variable : Time spent on self-study (y) – Continuous in nature

Independent variables:

- Time spent on online classes (x_1)
- Time spent on social media (x_2)
- Time spent on TV(x_3)

Regression Statistics:

Regression Statistics	
Multiple R	0.22584656
R Square	0.05100667
Adjusted R Square	0.04858988
Standard Error	1.51234044
Observations	1182

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	144.813227	48.2710756	21.105121	2.5486E-13
Residual	1178	2694.2905	2.2871736		
Total	1181	2839.10373			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.59843306	0.11938577	21.7650143	1.504E-88	2.36420058	2.83266554	2.36420058	2.83266554
Time spent on online study	0.11766154	0.02154168	5.46204144	5.7381E-08	0.0753972	0.15992588	0.0753972	0.15992588
Time spent on social media	-0.1356919	0.03282497	-4.1338022	3.8208E-05	-0.2000939	-0.07129	-0.2000939	-0.07129
Time spent on TV	-0.1115714	0.04200917	-2.655882	0.00801691	-0.1939926	-0.0291503	-0.1939926	-0.0291503

From the above statistics, we can see:

p-value is less than alpha, i.e. p-value < 0.05, we can say that variables Time spent on online classes, Time spent on social media, Time spent on TV has a significant impact on **Time spent on self-study.**

Also,

Intercept (b_0) = 2.5984331

Coeff. of time spent on online classes (b_1) = 0.11766154

Coeff. of time spent on social media (b_2) = -0.135691949

Coeff. of time spent on TV (b_3) = -0.111571404

Regression linear equation is:

$$\mathbf{y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 +\ldots + b_n*x_n}$$

So, in this case regression linear equation becomes:

$$\mathbf{y = 2.5984331 + 0.11766154*(x_1) - 0.135691949*(x_2) - 0.111571404*(x_3)}$$

Prediction

Prediction for Time spent on self-study if:

Time spent on online class = 4 hrs

Time spent on social media = 2 hrs

Time spent on TV = 1 hr

Time spent on self-study

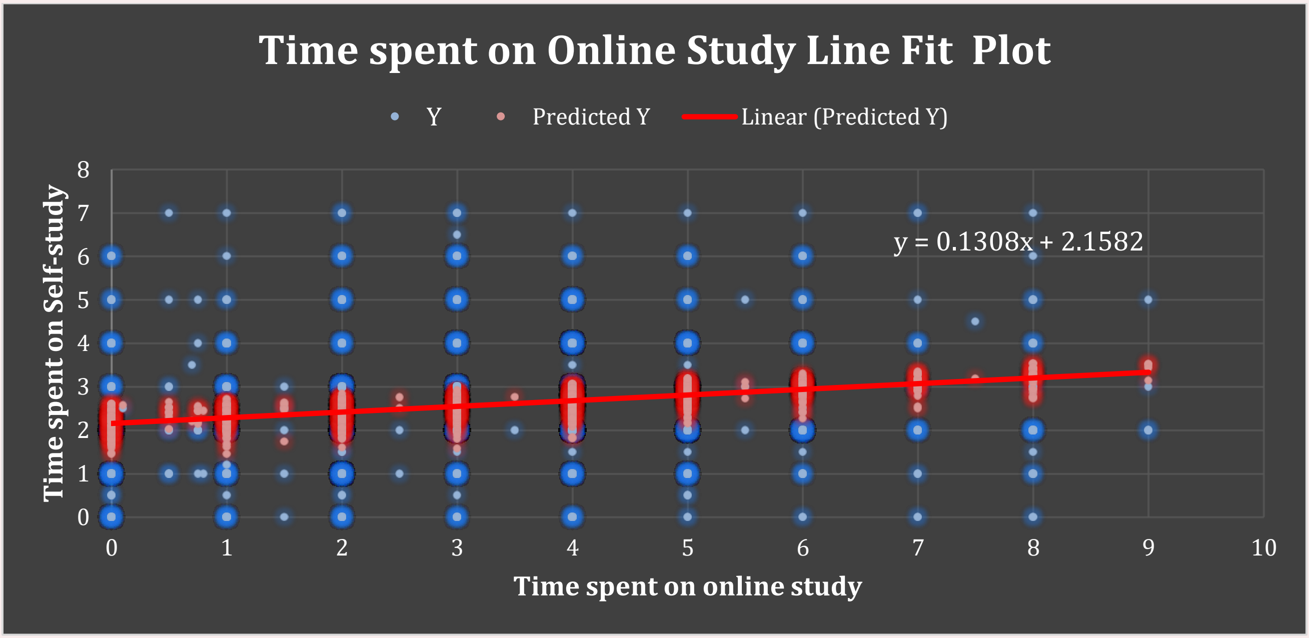
= 2.5984331 + 0.11766154*(4) - 0.135691949*(2) - 0.111571404*(1)

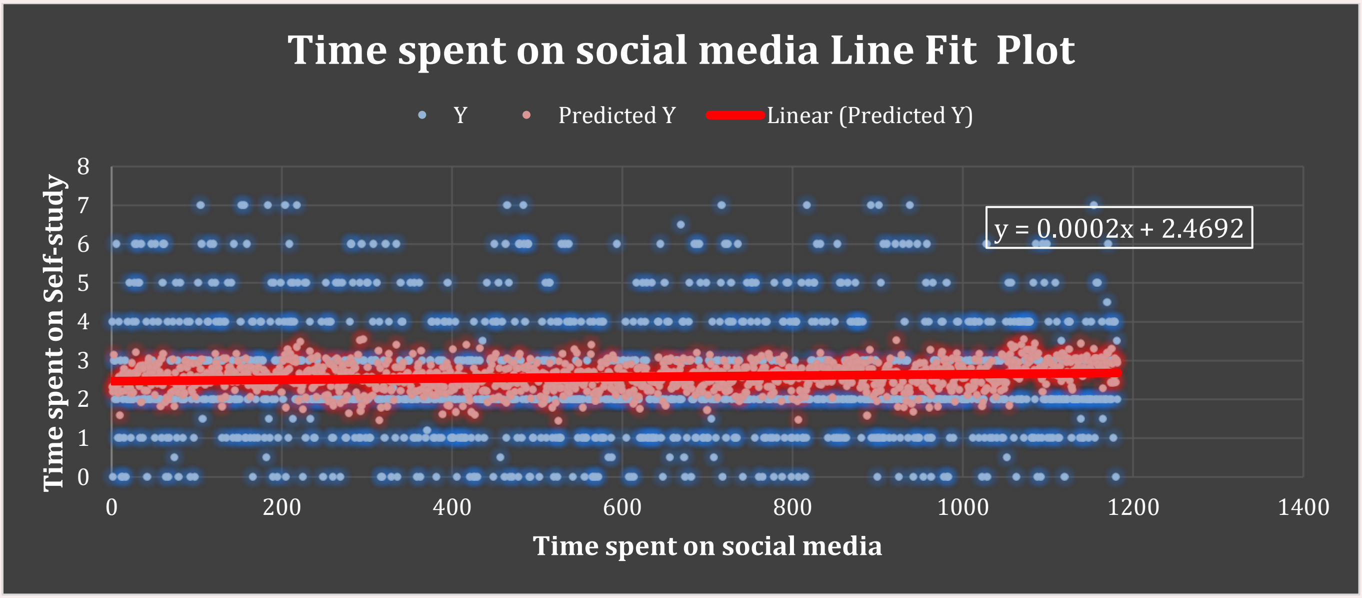
= 2.450800878

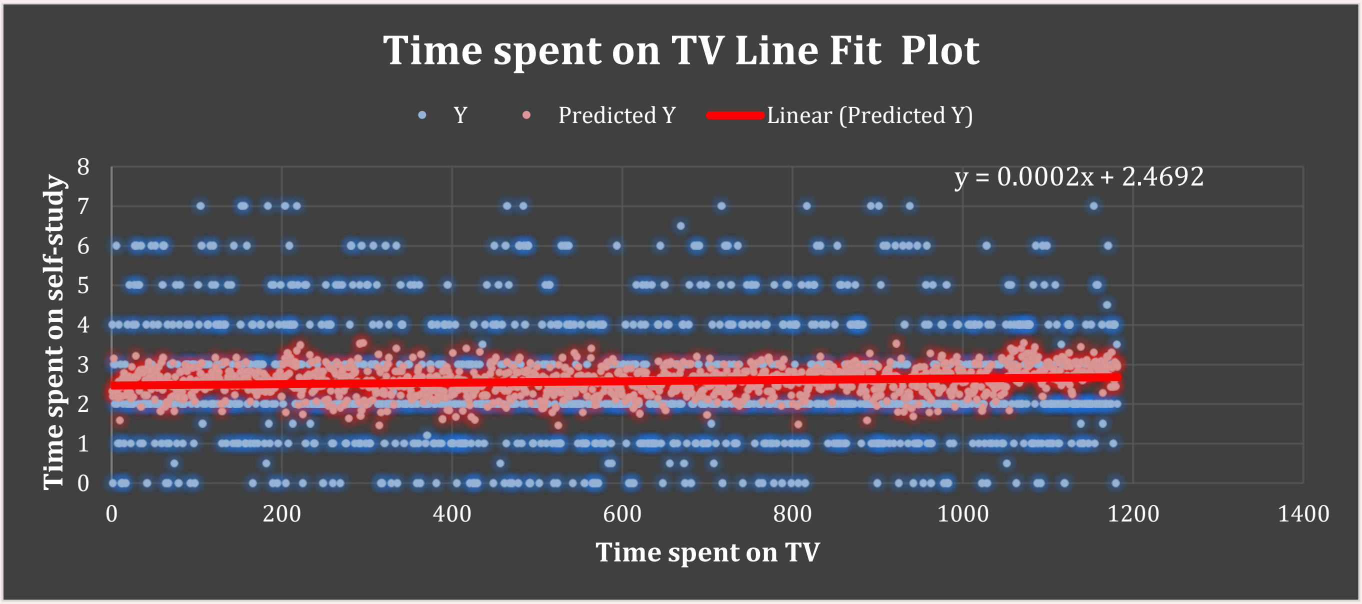
= **2.5 Hours**

Assumption of Linear Regression Model:

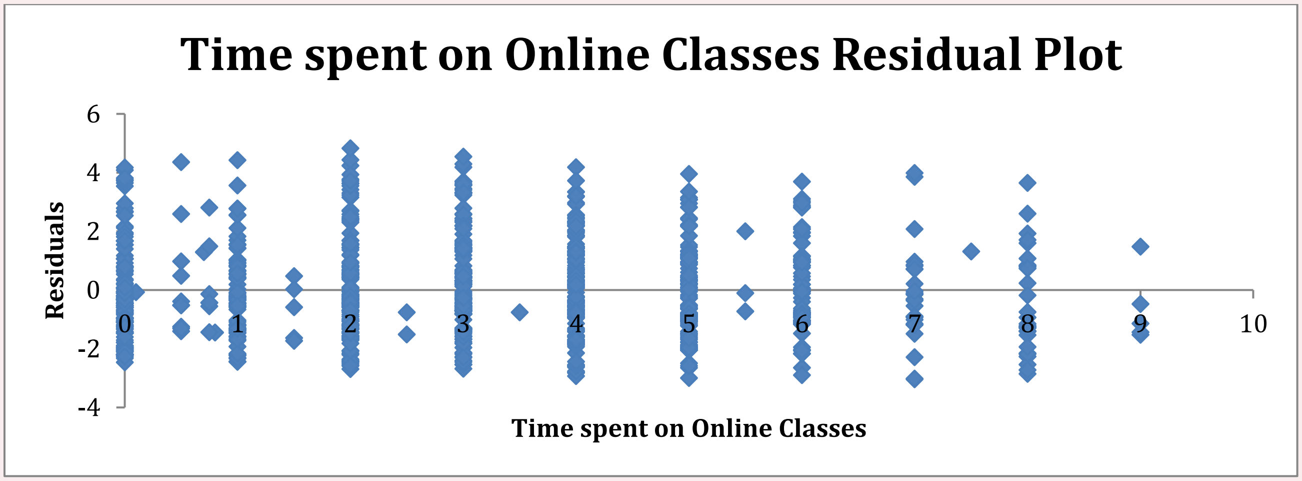
- **Linearity:** There is a linear relationship between dependent variable, Time spent on self-study and independent variables, time spent on online classes, time spent on social media, time spent on TV.

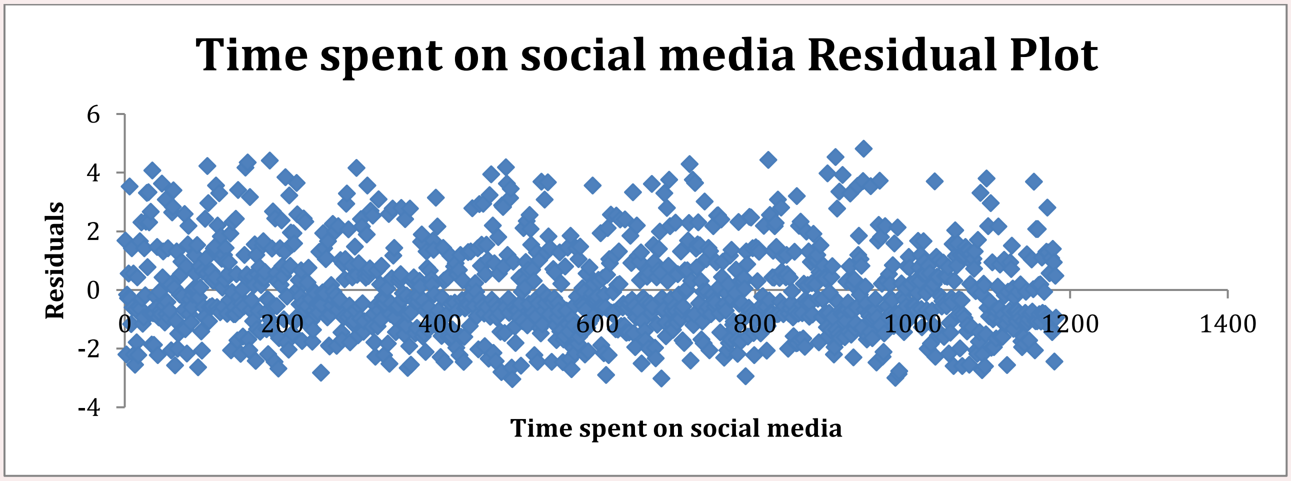


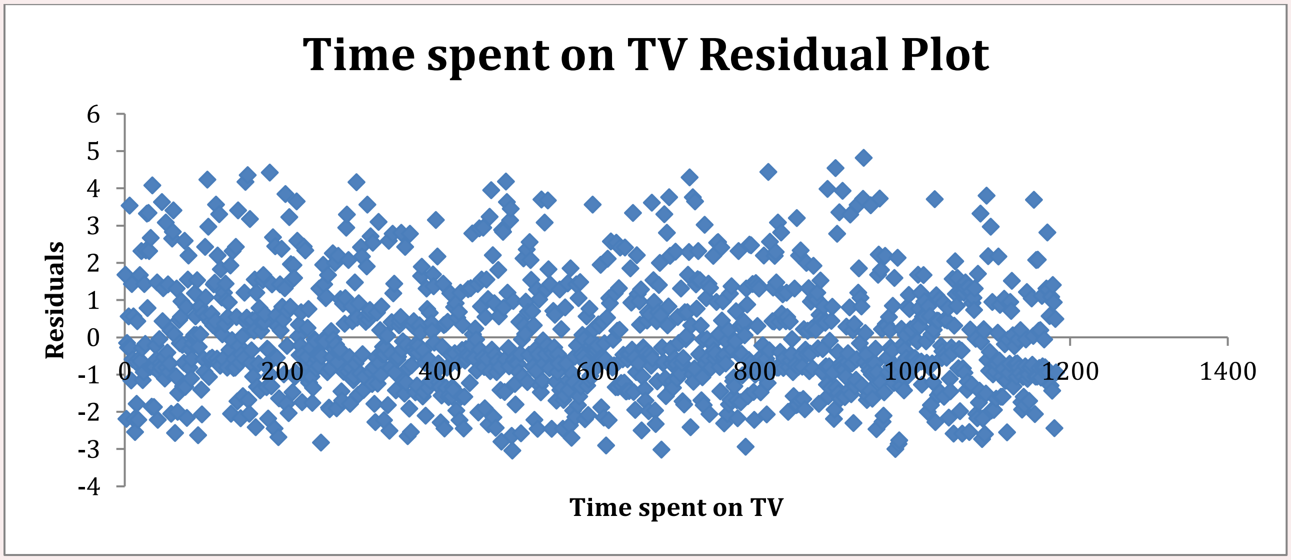




- **Homoscedasticity:** The variance of residual is constant across all values of Time spent on online classes, Time spent on social media, Time spent on TV.

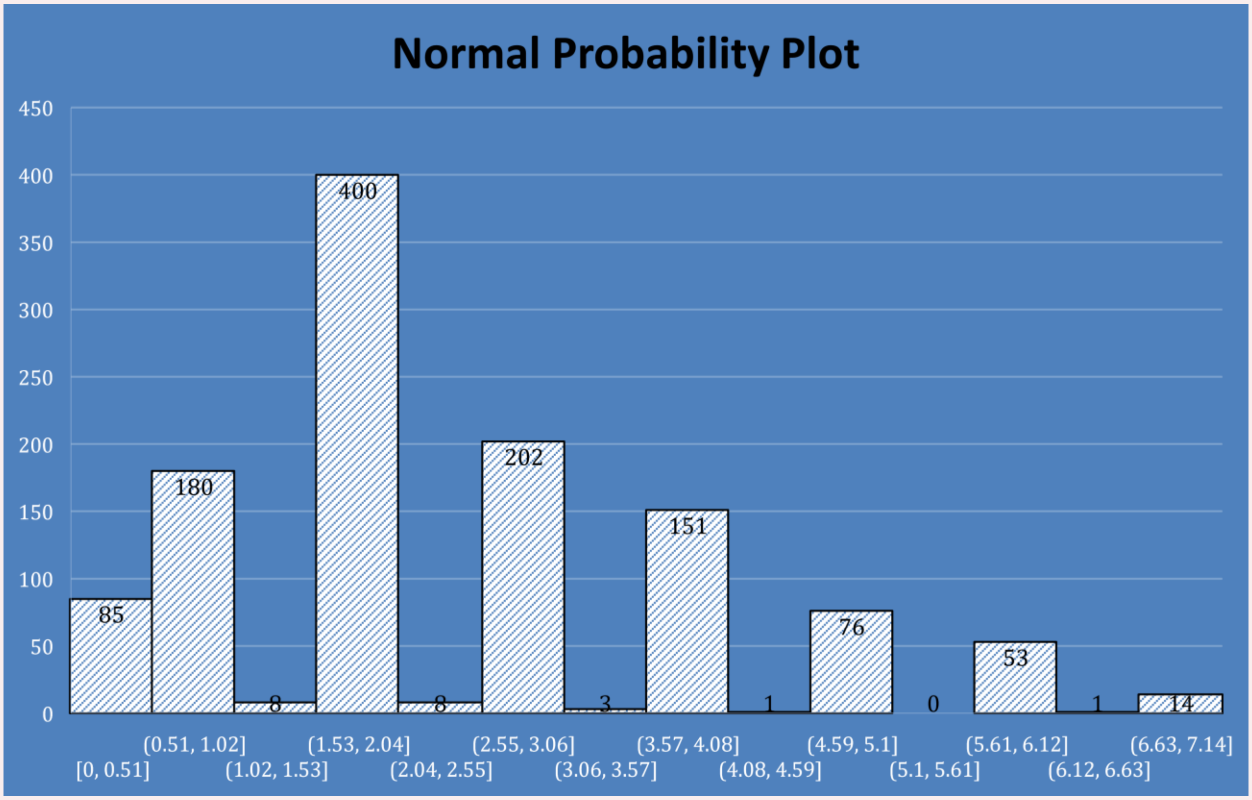

Time spent on Online Classes Residual Plot
Residuals
6
4
2
0
-2
-4
0
1
2
3
4
5
6
7
8
9
10
Time spent on Online Classes


Time spent on social media Residual Plot
Residuals
6
4
2
0
-2
-4
0
200
400
600
800
1000
1200
1400
Time spent on social media


Time spent on TV Residual Plot
Residuals
6
5
4
3
2
1
0
-1
-2
-3
-4
0
200
400
600
800
1000
1200
1400
Time spent on TV

- **Independence:** From the above charts we can see all the observations are independent of each other.

- **Normality:** The values of Time spent on self-study are normally distributed across all values of Time spent on online classes, on social media and on TV.



2. Logistic Regression Analysis

In this analysis, I will find how Change in weight impacted Health issues occurred in student's during lockdown.

Dependent variable : Health Issues (y) – Binary in nature

Independent variables: Change in weight (x)

Since we are supposed to use MS Excel as a tool for this course, so to get the results of logistic regression, we need to do the linear regression analysis on this first and then we can **tune the values of intercept and slope** accordingly.

Regression Statistics:

Regression Statistics	
Multiple R	0.07665731
R Square	0.00587634
Adjusted R Square	0.00503387
Standard Error	0.34229162
Observations	1182

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.81722444	0.81722444	6.97507377	0.00837445
Residual	1180	138.252996	0.11716356		
Total	1181	139.07022			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.2169945	0.03216775	6.74571577	2.3797E-11	0.15388213	0.28010686	0.15388213	0.28010686
Change in weight	-0.0368251	0.01394343	-2.6410365	0.00837445	-0.0641818	-0.0094684	-0.0641818	-0.0094684

From the above statistics, we can see:

p-value is less than alpha, i.e. p-value < 0.05, we can say that variable **Change in weight** has a significant impact on **Health Issues** occurred in students during Lockdown.

Also,

Intercept = b_0 = 0.2169945

Coeff. of Change in weight = b_1 = -0.0368251

Regression linear equation is:

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 +\dots + b_n*x_n$$

So, in this case regression linear equation becomes:

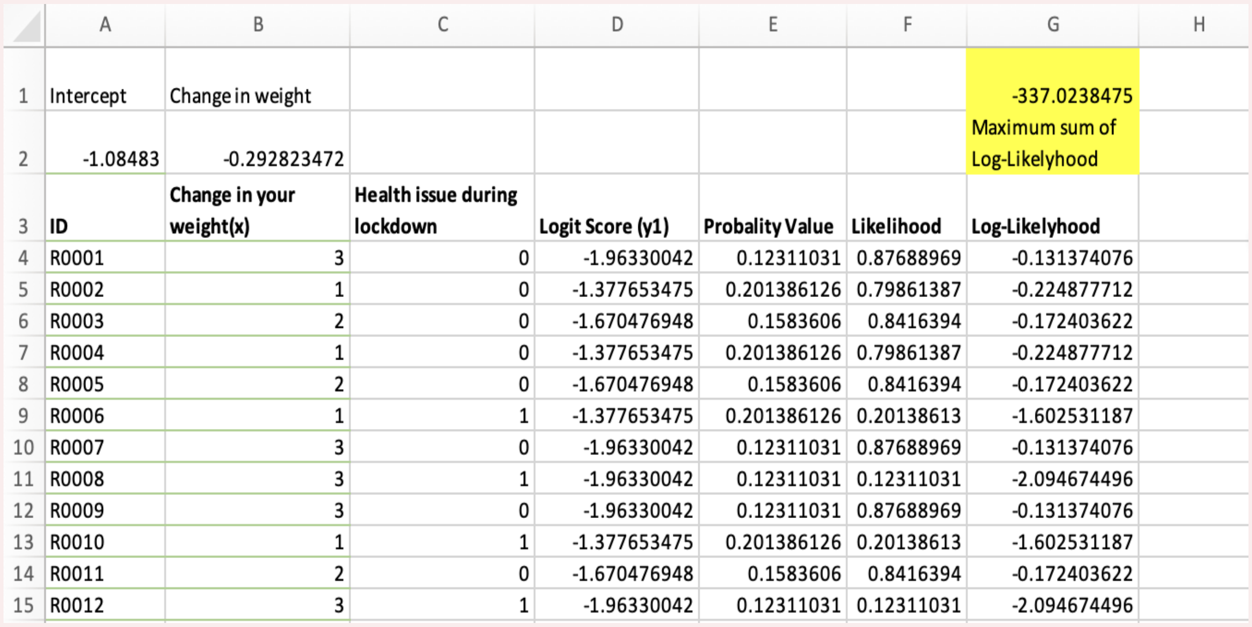
Regression score (y) = 0.2169945 – 0.0368251 * (x)

But we need to calculate Logit score. To get that score in MS Excel, I followed the steps stated below:

- We first calculate regression score for all the observations.
- Then calculate **Probability value** for all the observations as

$$\mathbf{P = e^y / (1 + e^y)}$$

- Then calculate **Likelihood value** as if student has health issue, then insert value of y, else (1 – y).
- Next, calculate **Log Likelihood** as Log of Likelihood value.
- Then, calculate **maximum sum of Log-Likelihood and tune the values of Intercept & coeff. of change in weight by using Solver tool under Data Analysis tab.**



	A	B	C	D	E	F	G	H
1	Intercept	Change in weight					-337.0238475	
2	-1.08483	-0.292823472					Maximum sum of Log-Likelyhood	
3	**ID**	**Change in your weight(x)**	**Health issue during lockdown**	**Logit Score (y1)**	**Probality Value**	**Likelihood**	**Log-Likelyhood**	
4	R0001	3	0	-1.96330042	0.12311031	0.87688969	-0.131374076	
5	R0002	1	0	-1.377653475	0.201386126	0.79861387	-0.224877712	
6	R0003	2	0	-1.670476948	0.1583606	0.8416394	-0.172403622	
7	R0004	1	0	-1.377653475	0.201386126	0.79861387	-0.224877712	
8	R0005	2	0	-1.670476948	0.1583606	0.8416394	-0.172403622	
9	R0006	1	1	-1.377653475	0.201386126	0.20138613	-1.602531187	
10	R0007	3	0	-1.96330042	0.12311031	0.87688969	-0.131374076	
11	R0008	3	1	-1.96330042	0.12311031	0.12311031	-2.094674496	
12	R0009	3	0	-1.96330042	0.12311031	0.87688969	-0.131374076	
13	R0010	1	1	-1.377653475	0.201386126	0.20138613	-1.602531187	
14	R0011	2	0	-1.670476948	0.1583606	0.8416394	-0.172403622	
15	R0012	3	1	-1.96330042	0.12311031	0.12311031	-2.094674496	

By doing so, we get maximum sum of Log-Likelihood as **-337.023**

After tunning we get the values of Intercept & Coeff. as

Intercept (b_0) = -1.08483

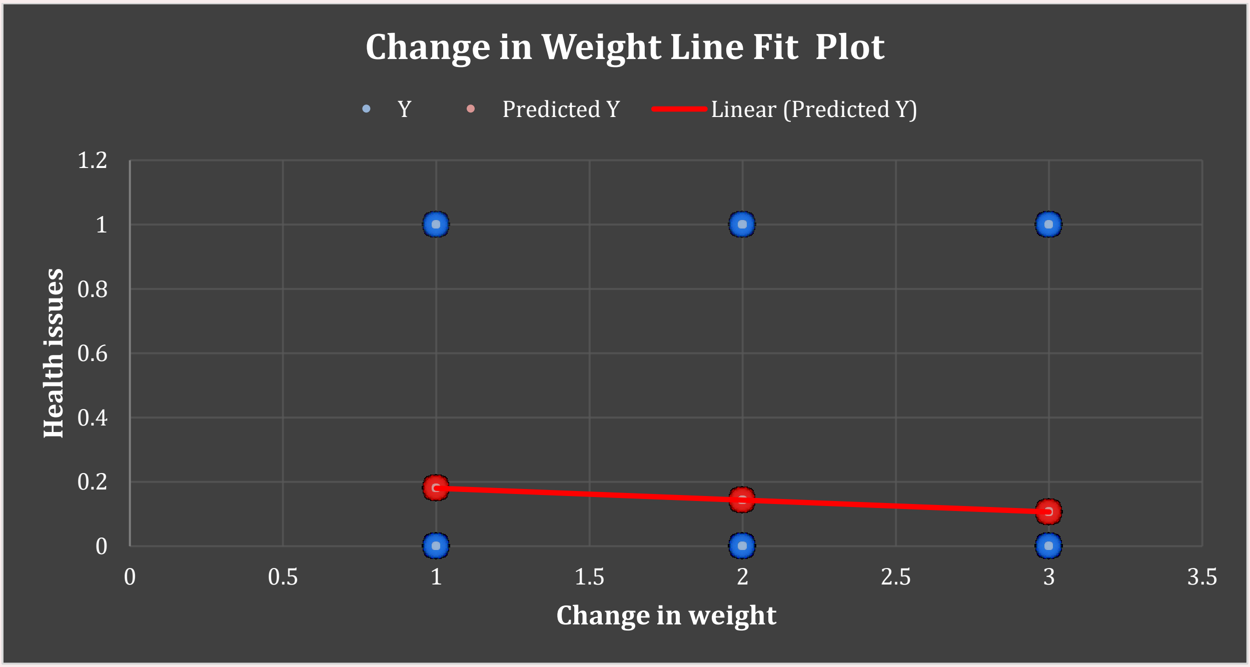
Coeff. of Change in weight (b_1) = -0.292823472

Now, equation becomes:

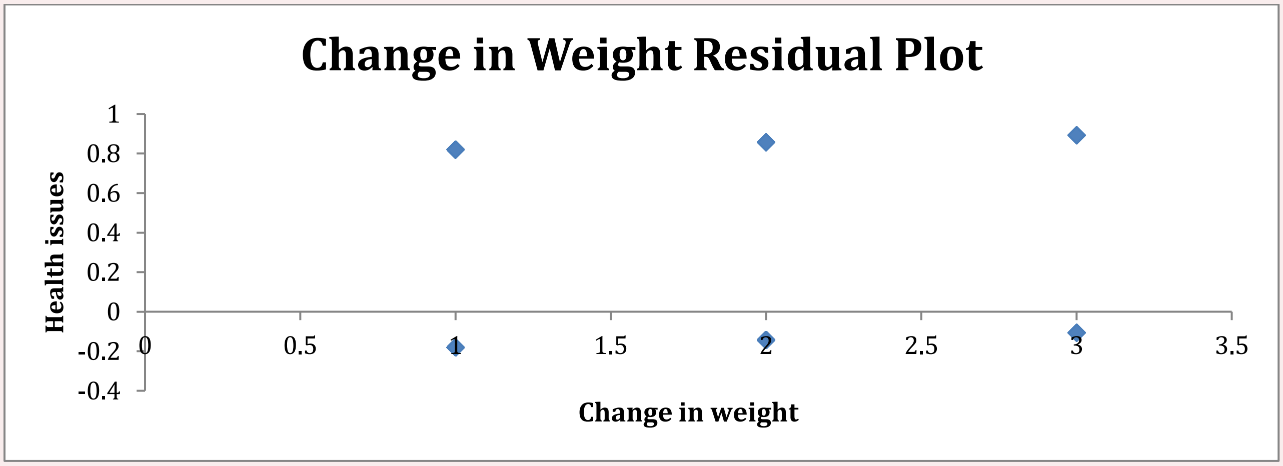
Logit score (y1) = -1.08483 - 0.292823472(x)

Assumption of Regression Model:

- **Linearity:** There is a linear relationship between dependent variable, Health Issues, and independent variable Change in weight.

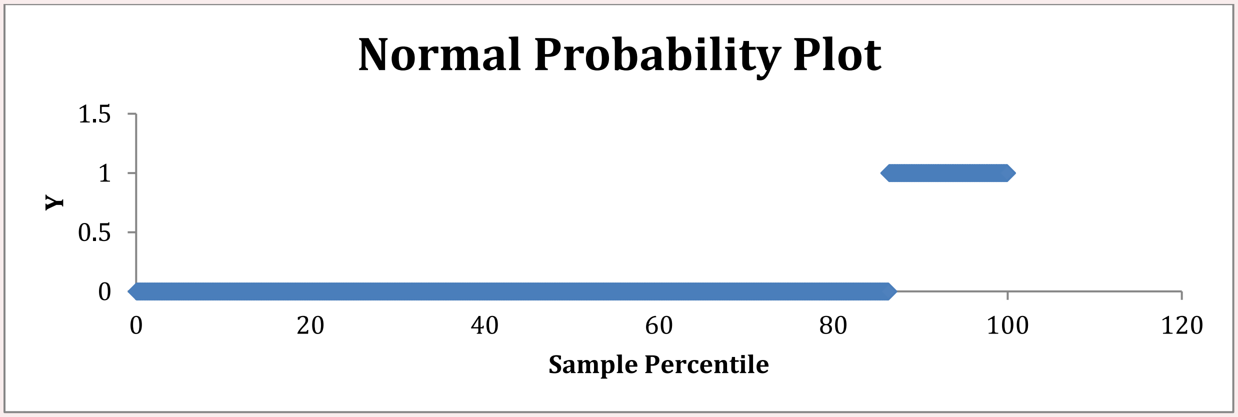


- **Homoscedasticity:** The variance of residual is constant across all values of Change in weight.



- **Independence:** From the above charts we can see all the observations are independent of each other.

- **Normality:** The values of variable Health Issues are normally distributed across all values of Change in weight.



Discussion

In this analysis, first research question was:

What kind of learning environment do pupils currently have to deal with due to the epidemic?

Assessment of Online Learning

Time spent on Online Classes

Groups	*Count*	*Sum*	*Average*	*Variance*
College	880	2454.3	2.78897727	3.77546994
Elementary School	15	48	3.2	2.02857143
Junior High School	125	506.75	4.054	3.12256452
Senior High School	162	755.8	4.6654321	4.41153056

By looking at the analysis, which states:

Elementary School - Mean: 3.2 hours

Junior High School - Mean: 4.1 hours

Senior High School - Mean: 4.6 hours

College - Mean: 2.8 hours

College students, who are supposedly enrolled in a higher level of education, spend the least amount of time in online classes—even less than kids in elementary school—so the relationship between level of education and time spent online is ambiguous.

We can observe that schools do not place a high priority on online learning, and students spend a wide range of amounts of time in online classes.

Meaning that there are different amounts of time that each student spends engaging in online learning depending on the level of education.

	Online Class experience					
Level of school	Average	Excellent	Good	Poor	Very poor	Grand Total
College	32.4%	4.4%	16.0%	3.0%	44.2%	100.0%
Elementary School	33.3%	33.3%	33.3%	0.0%	0.0%	100.0%
Junior High School	25.6%	32.0%	36.0%	0.0%	6.4%	100.0%
Senior High School	40.1%	8.6%	24.1%	2.5%	24.7%	100.0%

When the distribution of ratings for online learning was examined, the data was homogenous, indicating that the data had a significant trend value.

Online learning is of higher quality at lower grade levels.

College students spend the least amount of time learning online, and they rate online classes the worst.

Based on the amount of time spent in an online class and the rating given, it causes college students to have the worst online class quality.

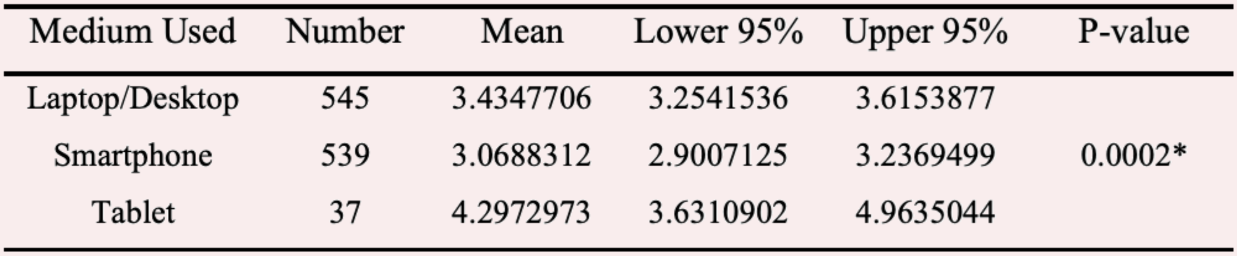
Based on the amount of time spent in the online class and the rating it received, elementary school students had the best online class quality.

In this analysis, second research question was:

In this epidemic era, how are students' online class situations?

	Device used to attend online classes					
Level of School	Any Gadget	Laptop/Desktop	Smartphone	Smartphone or Laptop/Desktop	Tablet	Grand Total
College	0.3%	57.6%	40.6%	0.0%	1.5%	100.0%
Elementary School	0.0%	40.0%	46.7%	6.7%	6.7%	100.0%
Junior High School	0.8%	22.4%	63.2%	1.6%	12.0%	100.0%
Senior High School	0.6%	34.0%	59.3%	1.2%	4.9%	100.0%

If we talk about the medium used for online study, the majority of students at College level used Laptop/Desktop and students at the lower level used Smartphone to study online. A very less percentage of students used Tablets to attend classes.



Medium Used	Number	Mean	Lower 95%	Upper 95%	P-value
Laptop/Desktop	545	3.4347706	3.2541536	3.6153877	
Smartphone	539	3.0688312	2.9007125	3.2369499	0.0002*
Tablet	37	4.2972973	3.6310902	4.9635044	

With regard to the time spent in online classes, there was a statistically significant difference between the various mediums used (P=0.0002). As shown above, 4.29 hours was the average time spent on online classes using tablets, 3.43 hours when using laptop/desktop, and 3.06 hours when using smartphones.

In this analysis, third research question was:

How are each school level's students behaving in terms of their health?

Average Time spent on sleep

College		*Elementary School*		*Junior High School*		*Senior High School*	
Mean	7.821454629	Mean	8.377777778	Mean	8.1463332	Mean	7.53298345
Standard Error	0.069120624	Standard Error	0.232006811	Standard Error	0.0954707	Standard Error	0.0120939
Median	7.761006289	Median	8.333333333	Median	8.2033898	Median	7.54166667
Standard Deviation	0.119720432	Standard Deviation	0.401847585	Standard Deviation	0.1653601	Standard Deviation	0.02094726
Minimum	7.744010417	Minimum	8	Minimum	7.96	Minimum	7.50909091
Maximum	7.959347181	Maximum	8.8	Maximum	8.2756098	Maximum	7.54819277
Confidence Level(95.0%)	0.29740204	Confidence Level(95.0%)	0.99824474	Confidence Level(95.0%)	0.4107773	Confidence Level(95.0%)	0.05203587

If we examined the average amount of time spent sleeping, all pupils slept for the recommended amount of time.

However, if we examined the minimum, maximum, errors, and distribution of the data, we found that a small number of students tended to spend more time sleeping. (bigger than 8 hours).

Change in Weight	Number of students
Decreased	17.7%
Increased	37.1%
Remain Constant	45.3%

Further if we analyse the change in body weight within this period, 37.1% reported an increase in weight, 17.7% reported a decrease in weight, and 45.3% reported no change in weight.

Alarmingly, 51.4% of respondents said they wasted time while the school was under lockdown. Additionally, their regular physical routines, social interactions, and sleeping patterns all had a big impact on their health.

Conclusion

Our findings in this study suggested that the Covid-19 outbreak has had a substantial effect on students' mental health, education, and everyday routines.

The Covid-19-related pauses highlight important issues and give a chance to better assess potential solutions in the educational field.

Project Tracking

Project		Deliverables				
TASK	DESCRIPTION	DELIVERABLE	% DONE	PRIORITY	DEADLINE	STATUS
Task 1	Collection of Dataset	Explored many data sources like Git Hub, Google Datasets, Kaggle and finally landed to https://www.kaggle.com/kunal28chaturvedi/covid19-and-its-impact-on-students and selected a dataset Impact of Covid 19 on students	100%	High	25th January 2023	Complete
Task 2	Dataset Overview	Dataset consisting of 1182 records with 19 separate variables	100%	Medium	26th January 2023	Complete
Task 3	Data Dictionary	Described each variable in detail along with its type, missing values, units, etc.	100%	Medium	29th January 2023	Complete
Task 4	Research Questions	Defined at least three Research questions along with the supportive questions to start the analysis	100%	High	1st February 2023	Complete
Task 5	Project Report	Created a detailed Project report of the analysis done on the initial stage	100%	High	5th February 2023	Complete
Task 6	Univariate Analysis	Univariate analysis of both categorical and numerical variables	100%	Medium	15th February 2023	Complete

Task 5	Data Cleaning	Cleaning of data by handling missing values and outliers	100%	High	20th February 2023	Complete
Task 7	Coding	Coding of categorical variables according to the research questions	100%	High	22nd February 2023	Complete
Task 8	Development of Hypothesis	Develop five hypothesis related to the research questions.	100%	High	2nd March 2023	Complete
Task 9	Hypothesis Testing	Perform OR, RR, Chi-square test, t-test, ANOVA and MANOVA accordingly.	100%	High	19th March 2023	Complete
Task 10	Inferential Techniques	Perform any 2 Inferential techniques in which at least one should be predictive.	100%	Medium	16th April 2023	Complete
Task 11	Discussion /Conclusion	Discuss the whole analysis with respect to the research questions	100%	High	16th April 2023	Complete

References

1. https://www.researchgate.net/publication/347935769_COVID-19_and_its_impact_on_education_social_life_and_mental_health_of_students_A_Survey
2. https://www.kaggle.com/kunal28chaturvedi/covid19-and-its-impact-on-students
3. https://seleritysas.com/blog/2019/12/06/preparing-for-the-future-of-data-analytics/
4. https://www.smartsheet.com/top-project-management-excel-templates