Introduction

I have collected personalized data about myself on how I spend my days. I recorded data from January 14th, 2022, to February 15th, 2022, on these variables: number of hours I spent on zoom class, number of hours I spent studying, number of hours I spent sleeping, number of times I left the house, number of emails I received on that day, did I watch news that day, which fruit I ate that day and how did my dinner taste that day.

The first five variables listed above are quantitative. I am currently a student, so a lot of my time is spent on online zoom classes and studying. The number of times I left the house is usually for my job. So, it will be interesting to see how my class hours, study hours, the number of times I leave the house, and my sleep correlate with each other. The last quantitative variable is the number of emails I receive every day. Since every information is now shared online because of the pandemic, I had to keep checking my email to keep up with my assignments, class schedules, and job schedule so I tracked the numbers of emails I receive every day.

The last three variables that I collected are categorical. Did I watch the news that day and which fruit did I eat that day? Both these variables are nominal. I was interested in tracking whether I watch the news or not and which fruit I consume the most. I also recorded how my dinner tastes each day on a scale of 1 to 5, ‘1’ tasting very bad and ‘5’ tasting very good. Since I started living on my own and I am new to cooking I wanted to track if my cooking has improved. This is an ordinal categorical variable.

Since I am an international student and I am in a new environment than what I am used to, I plan what I am going to do each day. And I hope that this analysis will help me better understand what I spend my day on most, how the variables I have gathered are correlated with each other, and how can I better allocate my time to get more out of each day.

Data Analysis

Categorical Variables:

Frequency of the days I watched News:

This is the first nominal categorical variable that I recorded. It shows the frequency of times I watched the news in the 33 days. This data does not include the days I read the news but only the days I watched the news in a video format. The response 'Yes' means I watched the news on those days while 'No' means I didn't watch any news on those days. As Table 1 shows, out of 33 days, I only watched the news on 2 occasions. So, 93.93% of the total days I did not engage in any news watching activity. As Figure 1 shows, I would rather engage in some other activity than watch the news. With this analysis, I concluded that it is more convenient and faster to just read the news rather than watch it.

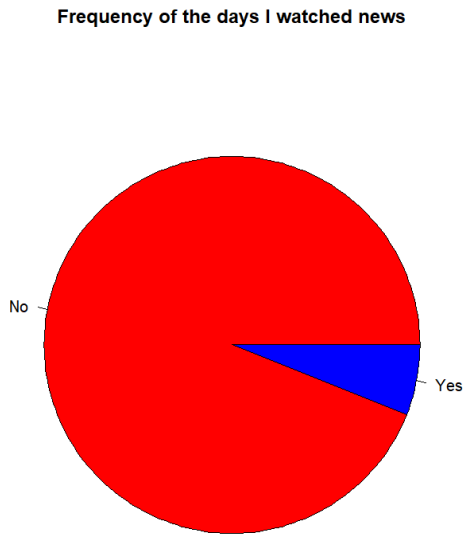


Figure 1: Frequency of the days I watched News

Table 1: Frequency of the days I watched News		
	Yes	No
Number of days	2	31
Percentages	6.06%	93.93%

Frequency of the fruits I ate:

This is the second nominal categorical variable that I recorded. I recorded this data to find out what types of fruits I was consuming each day and which fruit I prefer to others. As Table 2 shows, I ate 5 types of fruits in the 33 days. I ate 4 fruits, bananas, grapes, lemon, and oranges each for 3 days out of the total, but I ate apples for 21 days which means I ate an apple 63.63% of the time. Figure 2 shows the frequency of the fruits that I ate each day. And is clear from the bar chart that I prefer to eat apples to other fruits.

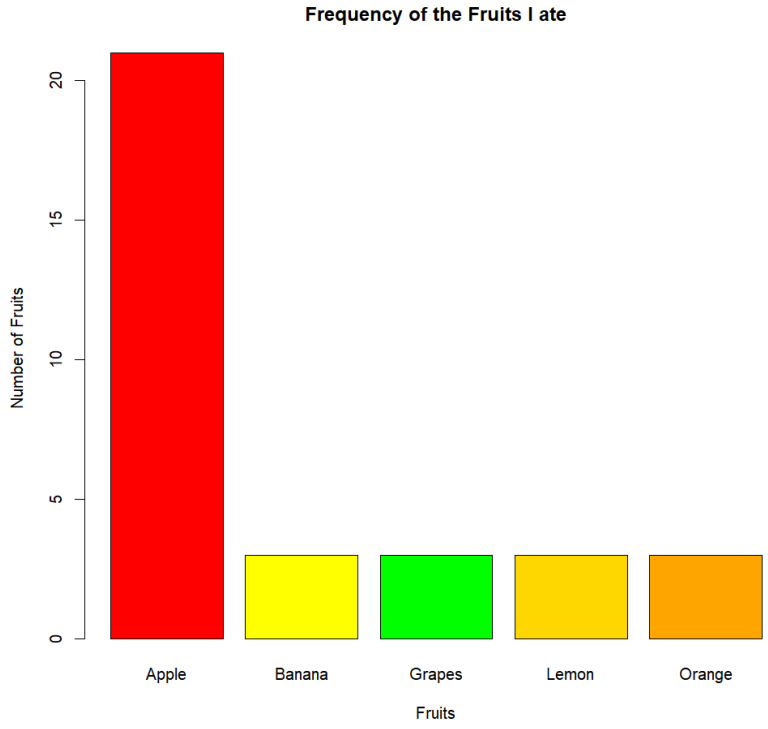


Figure 2: Frequency of the fruits I ate

Table 2: Frequency of the fruits I ate					
	Apple	**Banana**	**Grapes**	**Lemon**	**Orange**
Number of fruits	21	3	3	3	3
Percentages	63.63%	9.09%	9.09%	9.09%	9.09%

Frequency of dinner satisfaction:

The last categorical variable that I collected is ordinal. It measures my satisfaction with how the dinner I cooked tastes. Since I live alone some of my time is spent cooking dinner and I measured this variable because I also wanted to measure my improvement in cooking skills. It measures my satisfaction from '1' to '3' where '1' is 'OK', '2' is 'Satisfied and '3' is 'Very Satisfied'. Table 3 and Figure 3, show three variables with frequencies '24.24%', '39.39%' and '36.36%' for '1', '2' and '3' respectively. The highest value is recorded as '13' (39.39%) for '2' out of 33 days and the lowest value is '8' (24.24%) for '1' out of 33 days. With this data analysis, I can conclude that I seem to be improving my cooking skills.

Table 3: Frequency of dinner satisfaction			
	1	**2**	**3**
Number of frequencies	8	13	12
Percentages	24.24%	39.39%	36.36%

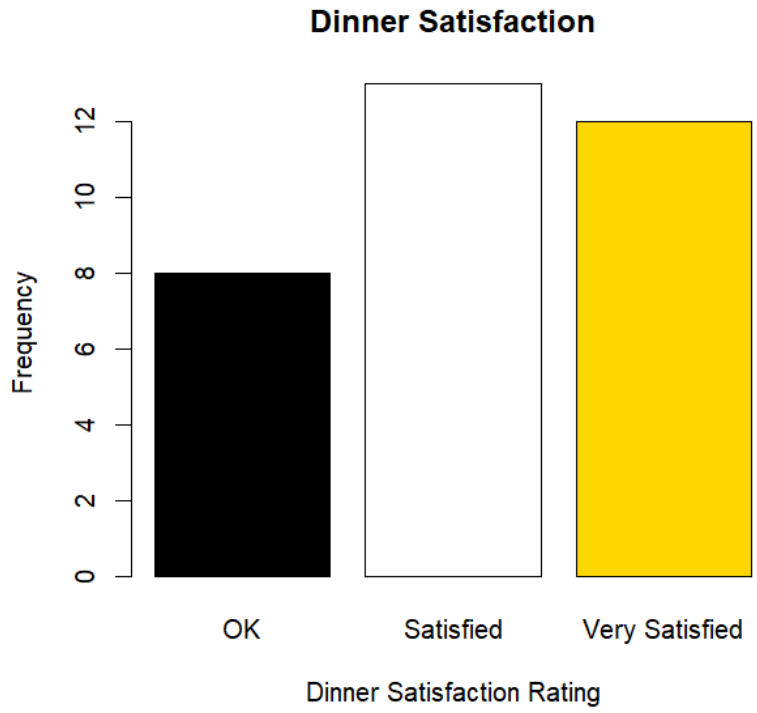


Figure 3: Dinner Satisfaction bar plot

Contingency Table:

I compared categorical variables of me consuming fruit and my dinner satisfaction to each other using PivotTable and the result for that is shown in Table 4. It is very unrealistic that these variables might be influenced by each other. But let's analyze these data and maybe some insight might appear that shows that these two variables influence each other.

As seen in Table 4, for days that I ate orange the chance of dinner satisfaction being '1', '2' or '3' are equal with 3% for each value. If we look at dinner satisfaction when I ate banana chances of me being 'satisfied' or 'vary satisfied' are 6.1% and 3%, with zero occurrences of '1'. With grape, the chances are I will be 'satisfied' with my dinner but not 'ok' or 'very satisfied. If we look at data for apple, the chances of dinner satisfaction are '18.2%', '21.2%' and '24.2%' for '3', '4' and '5' respectively. These values in the contingency table don't make sense in my analysis. So, I conclude that my satisfaction with my dinner is not influenced by me eating apples or any other fruits.

From the data collected, I would like to conclude that my dinner satisfaction and my eating fruit do not influence one another. Maybe if more data is collected for these variables some new insights may appear that might suggest that these variables depend on each other but for now, these variables seem to be independent of each other.

Table 4: Contingency Table comparing the fruits I ate and my satisfaction with dinner

		Fruits				
Dinner Satisfactions		**Apple**	**Banana**	**Grapes**	**Lemon**	**Orange**
	1	18.2%	0%	0%	3%	3%
	2	21.2%	6.1%	9.1%	0%	3%
	3	24.2%	3%	0%	6%	3%

Quantitative Variables Analysis:

Zoom Variable:

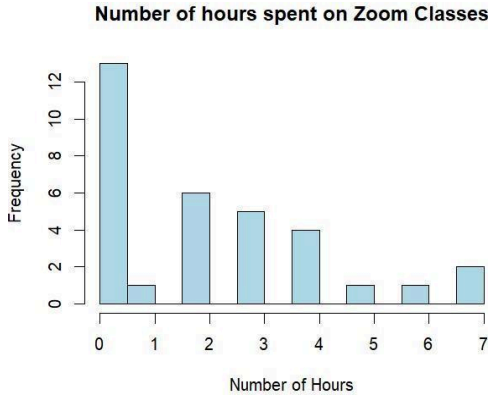
The first quantitative variable I created for this analysis is Zoom. This variable contains the total number of hours I spent on Zoom classes. This data only records the hours I attended the zoom classes in real-time and does not count the hours I watched the recordings of my missed classes.

Figure 4 shows the histogram for the Zoom variable. Since the mode was unclear between '0' and '1', I had to add more breaks to the histogram to identify the mode as '0'. The data is unimodal and is right-skewed. There seem to be no outliers when you look at Figure 4. The mode equals '0' maybe because there are zero zoom classes on weekends (Saturday and Sunday every week), and I might have missed all classes on certain days due to my work schedules.

Table 5 shows the statistics for the Zoom variable. It illustrates that the mean was 2.090909 and the median was 2. Since the mean is greater than the median it can be concluded that the data is right-skewed. The data also shows a standard deviation of 2.155859 and an IQR of 3. Since the data is asymmetric, median and interquartile ranges are the best measures for centrality and dispersion.

Since the data was asymmetric, I decided to use a boxplot to calculate outliers. And it can be seen in Figure 5 that there are no outliers in the data. This may be because there are no extreme values in the data and the data set is too small to contain any outliers. Also, the boxplot only contains a fence above the upper quartile of my data but no fence below the lower quartile.

Table 6 shows the correlation coefficient of Zoom with other quantitative variables. Here zoom has the strongest correlation with study with a correlation coefficient of 0.3933782. It is not a strong correlation, but it is the strongest compared to correlations with other variables.



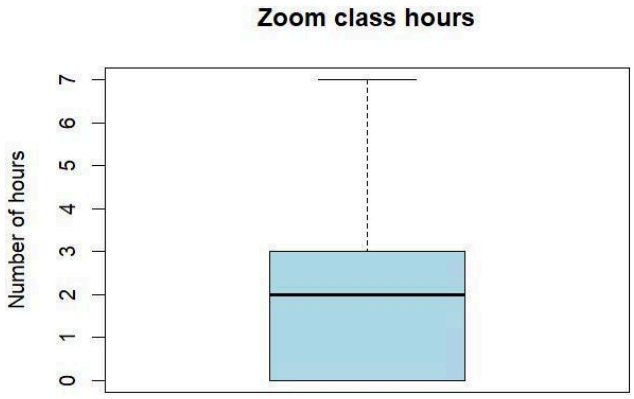


Figure 4: Histogram for Zoom

Figure 5: Boxplot for Zoom

Table 6: Correlation Coefficient of Zoom with other variables	
Other Variables	**Zoom**
Study	0.3933782
Sleep	0.1864882
House	-0.02483867
E-mail	0.285793
Table 5: Statistics for Zoom Variable	
Statistics	**Value**
Mean	2.090909
Median	2
Standard Deviation	2.155859
IQR	3

Study Variable:

The second quantitative variable I recorded for this data set was Study. This variable contains all the hours I spent studying in classes, doing assignments, and doing online courses. This variable always contained a higher value than the zoom variable because I added zoom classes hours to other hours I spend studying. I spend a lot of time learning online courses on Data Camp. The study variable also records the hours I spend on Data Camp.

Figure 6 illustrates the histogram for the Study variable. It can be seen in the histogram that the data is bimodal (i.e., it has two modes) having modes as ‘4’ and ‘6’. The histogram might look right-skewed, but it is symmetric, and it contains no outliers. The histogram shows that I spend 4 to 6 hours studying most of my days but there are rare occasions when I have outdone myself by studying from 8 to 9 hours in a day.

Table 7 shows the statistics for the Study variable. It illustrates that the mean was 5.212121 and the median was 5. These data lead me to believe that the variable is symmetric and to use standardized values for calculating the outliers. The data also shows a standard deviation of 1.634732and an IQR of 2.

Since the data is symmetric, I used standardized values (z-score) for calculating outliers. And the calculations show that there are no outliers for this data. I also plotted a boxplot to support my conclusions about zero outliers and to show that the data is symmetric. It can be concluded by analyzing the box plot that the data is symmetric since the median is almost equal to the mean. The data is symmetric, so mean and standard deviation are the best measures for centrality and dispersion.

Table 8 shows the correlation of the Study variable with other variables. The strongest correlation is between Study and Sleep variables with the correlation coefficient of 0.4754795. this correlation is moderate in strength according to the definition, but it is the strongest compared to all other variables. And form this analysis, it can be concluded that if I sleep for more hours, chances are I will also study for more hours on the same day. On the other hand, the weakest correlation is between study and E-mail because the number of emails I receive doesn't influence how many hours I will study that day.

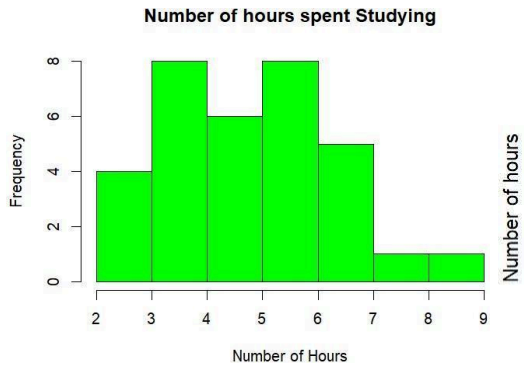


Figure 6: Histogram for Study

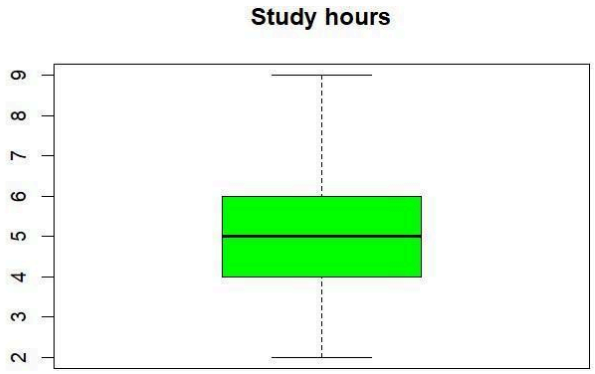


Figure 7: Boxplot for Study

Table 7: Statistics for Study Variable	
Statistics	**Value**
Mean	5.212121
Median	5
Standard Deviation	1.634732
IQR	2

Table 8: Correlation Coefficient of Study with other variables	
Other Variables	**Study**
Zoom	0.3933782
Sleep	0.4754795
House	-0.3867128
E-mail	-0.06767811

Sleep Variable:

The next variable that I stored was Sleep. It stores the number of hours I slept on that day. Figure 8 and Figure 9 show the histogram and box plot for the Sleep variable. And Table 9

shows the statistics for the Sleep variable (which is mean = 7.666667, median = 8, standard deviation = 0.7772816 and IQR = 1).

In this analysis, if you look at the histogram the data may seem symmetric. And Table 9 also shows mean and median are 7.666667 and 8 respectively. The mean and median seem too close that one may categorize this variable as symmetric. But if you look closely at the boxplot, it can be seen the median is equal to the 3rd quartile and not equal to the mean. Which makes it clear that the variable is asymmetric. This may have occurred due to a low IQR value which is '1'.

Since the variable was asymmetric, I used the 1.5 IQR rule to calculate the outliers but there seems to be no outlier in this variable. Also, median and interquartile ranges are the best measures for centrality and dispersion.

Table 10 shows the correlation coefficient of the Sleep variable with other variables. The sleep variable has the strongest correlation coefficient with the study variable with a value equal to 0.4754795. This might mean that if I study for more hours, chances are I will be tired and have a good sleep. And the weakest correlation is between the Sleep and the Zoom variables. It may be because I sleep at night and all my classes are during the daytime so, Zoom and Sleep variables are independent of each other.

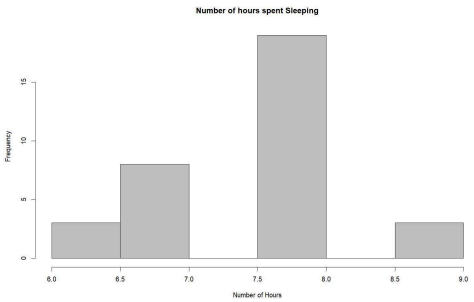


Figure 8: Histogram for Study

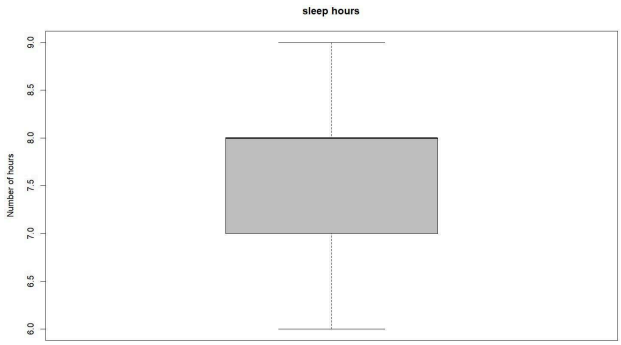


Figure 9: Boxplot for Study

Table 9: Statistics for Sleep Variable	
Statistics	**Value**
Mean	7.666667
Median	8
Standard Deviation	0.7772816
IQR	1

Table 10: Correlation Coefficient of Sleep with other variables	
Other Variables	**Sleep**
Zoom	0.1864882
Study	0.4754795

House	-0.2736553
E-mail	0.1894233

House Variable:

The next variable that I gathered was the House variable. This variable stores the number of times I went out of the house. Since I come from a hot weather city in Nepal where temperatures are between 5C to 10C during winter and above 40C during summer, I am used to hot weather. But here in Canada temperature is usually negative. So, it is a very different environment than what I am used to. Due to that fact, I do not go out of my house very often. All the times I went for an outing was for my job.

Figure 10 and Figure 11 show that the variable is unimodal with mode equals 1 and it is left-skewed. And there are no outliers for this variable which is calculated using the 1.5 IQR rule and boxplot. Table 11 shows the statistics of the variable, mean, median, standard deviation, and IQR as 0.6969697, 1, 0.6366341, and 1 respectively. Median and interquartile ranges are the best measures for centrality and dispersion because the data is skewed.

Table 12 illustrates the correlation coefficient of the House variable with other variables. Every variable is negatively correlated with the House variable. House variable has the highest correlation with study variable and second highest with sleep variable with a correlation coefficient of '-0.38' and '-0.27' respectively, which means that every time I went out of the house, I studied for fewer hours and, slept for fewer hours that night. Since the House variable has a very low correlation with Zoom and E-mail, I can conclude that these variables are not correlated.

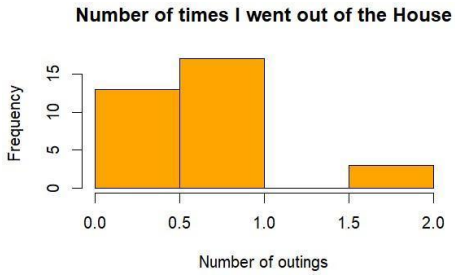


Figure 10: Histogram for House

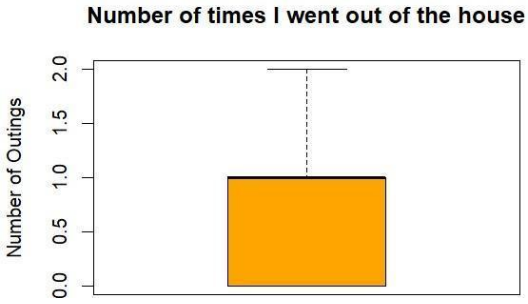


Figure 11: Boxplot for House

Table 11: Statistics for House Variable	
Statistics	**Value**
Mean	0.6969697
Median	1
Standard Deviation	0.6366341

IQR	1
Table 12: Correlation Coefficient of House with other variables	
Other Variables	**House**
Zoom	-0.02483867
Study	-0.3867128
Sleep	-0.2736553
E-mail	-0.03515062

E-mail variable:

The last variable that I saved was the E-mail variable. Since I had just arrived in Canada, I had to find a job and I applied to a lot of jobs online through email and had to keep checking my emails for any responses from employers, so I started to record the number of emails I receive every day. And I found a job after some time. But I kept recording my emails to see if there are any changes to the numbers. And, I had to keep using my email to keep up with my schedule for the day.

In Figure 12 it is apparent that the variable is right-skewed and is unimodal. And there are no outliers in this variable. The maximum number of emails I received every day is 20. These are the peak days when I was regularly applying to jobs online. But after I got a job the number of emails, I receive every day gradually dropped.

Table 13 shows the statistics for the E-mail variable where mean equals 9.575758, median equals 9, standard deviation equals 4.527902, and IQR equals 7. Since the variable is right-skewed, I calculated outliers using the 1.5 IQR rule and boxplot. Figure 13 shows the boxplot for the E-mail variable. As the boxplot shows no outliers were found for this variable. Also, median and interquartile ranges are the best measures for centrality and dispersion.

Table 14 shows the correlation coefficient of the E-mail variable with other variables. The E-mail variable has a very low correlation with all the other variables. The highest correlation it has is with Zoom with a correlation coefficient of 0.285793. From this analysis, it can be concluded that the E-mail variable is not influencing or is being influenced by any of the other variables since all the correlation coefficient is weak.

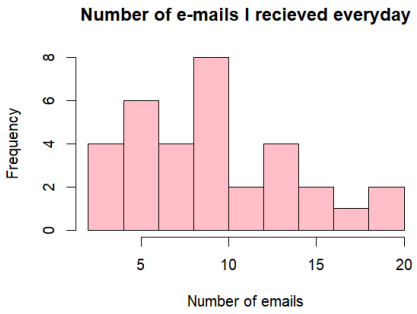


Figure 12: Histogram for Study

Figure 13: Boxplot for Study

Table 13: Statistics for E-mail Variable	
Statistics	**Value**
Mean	9.575758
Median	9
Standard Deviation	4.527902
IQR	7

Table 14: Correlation Coefficient of E-mail with other variables	
Other Variables	**E-mail**
Zoom	0.285793
Study	-0.06767811
Sleep	0.1894233
House	-0.03515062

Quantitative Variables with Highest Correlation:

From Table 15, it can be interpreted that the sleep variable and the study variable have the highest correlation among all the recorded quantitative data. The correlation coefficient equals 0.4754 which is a moderate and a positive relationship. This means that the day is sleep more, I will also study more or vice versa depending on which variable is dependent and independent.

Table 15: Correlation Coefficient between all quantitative variables					
	Zoom	**Study**	**Sleep**	**House**	**E-mail**
Zoom	1.0000	0.3933	0.1864	-0.0248	0.2857
Study	0.3933	1.0000	0.4754	-0.3867	-0.0676
Sleep	0.1864	0.4754	1.0000	-0.2736	0.1894

House	-0.0248	-0.3867	-0.2736	1.0000	-0.0351
E-mail	0.2857	-0.0676	0.1894	-0.0351	1.0000

4 conditions need to be met for regression. The conditions are:

1. **Quantitative Variable Condition:** This condition states that both variables are quantitative. This is true for this case because both variables are collected for this analysis are quantitative.
2. **Linearity Condition:** The second condition is that the relationship should be linear because regression only measures the strength of a linear relationship. Figure 14 checks the linearity condition for two quantitative variables. Since the red line in Figure 14 does not follow a horizontal line, the linearity condition is not met.
3. **Outlier Condition:** The third condition is the outlier condition. Figure 15 shows the Normal Q-Q plot which checks the outlier condition. Since the points do not follow the straight line in Figure 15 this condition is also not met.
4. **Equal Spread Condition:** The last condition is the equal spread condition. Figure 16 checks the equal spread condition by plotting a Scale-Location plot. Since the plot does not give a straight horizontal line (i.e., the red line in Figure 14) and the points are not evenly dispersed the equal spread condition is not met.

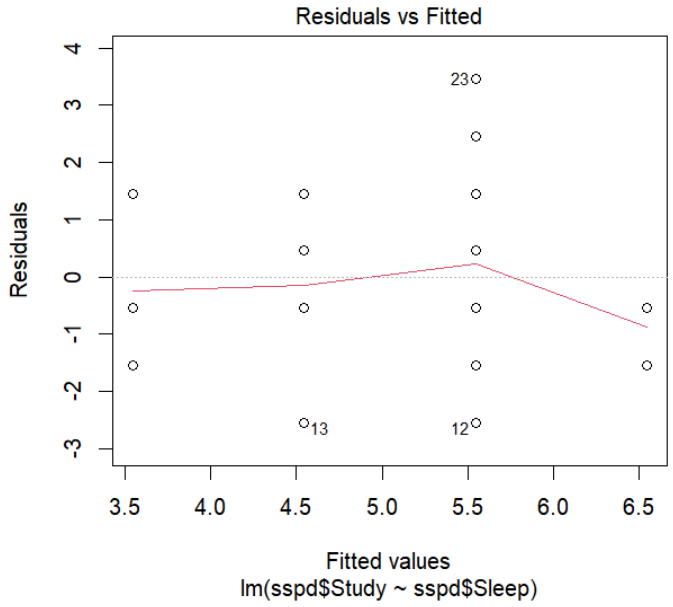


Figure 14: Residuals Plot Fitted

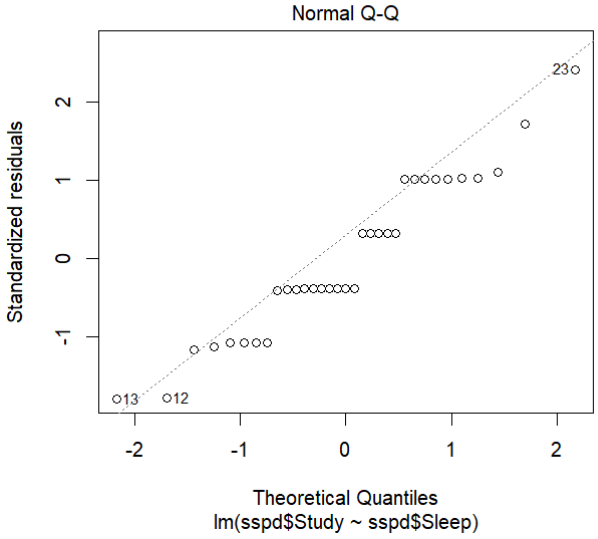


Figure 15: Normal Q-Q Plot

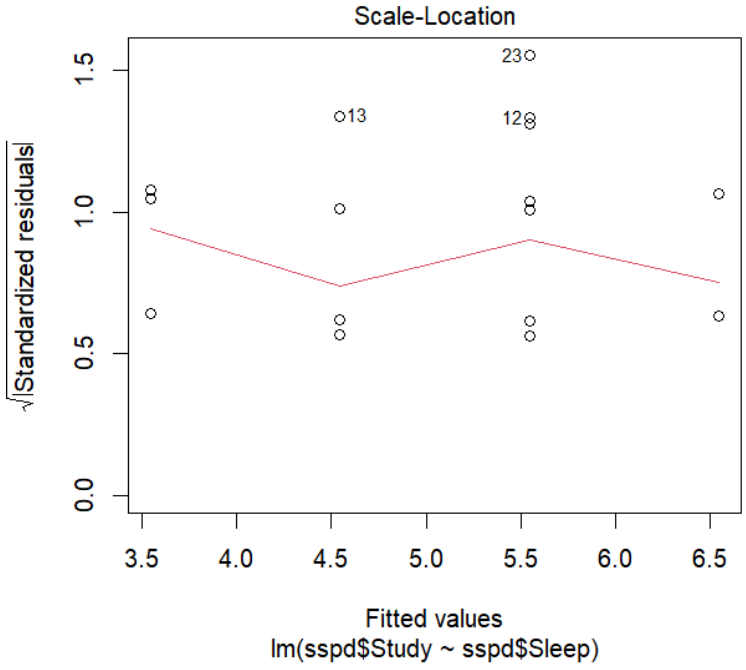


Figure 16: Scale-Location Plo

Equation of the regression line:

The equation of the regression line is,

Study = Sleep – 2.454545

The slope of the line is ‘1’, which means every time the sleep variable increases by 1-hour study variable will increase by 1-hour. The y-intercept value is -2.454545, which means that when the sleep variable equals zero the study variable equals -2.454545.

Figure 17 shows the Scatter Chart for the study variable and the sleep variable. It also includes the regression line for the variables. For this analysis, I chose the study variable as the dependent variable and the sleep variable as the independent variable.

The regression line equation predicts the dependent variable when you put the independent variable in the equation. For this example, I predicted that if I sleep for 12 hours a day, I will study for 9.54 hours the same day.

Study = 12 – 2.454545

=> Study = 9.545454

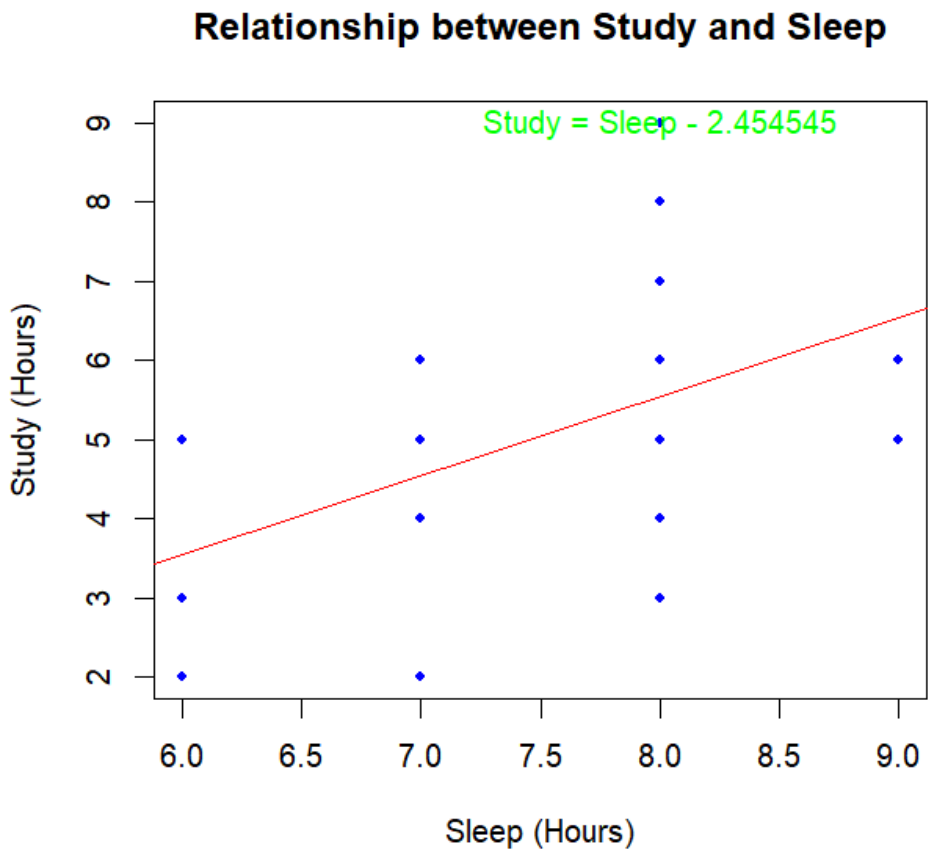


Figure 17: Scatter Chart for Study and Sleep

Conclusion

This analysis of my data set recorded from January 14th, 2022, to February 15th, 2022, has given me a lot of insights into how I spend my days and what changes I need to make to have a more productive day.

The categorical data showed that I do not spend my time watching the news. Since I only watched the news 2 days out of 33 days. It showed that I prefer to eat apples more than any other fruits. And finally, my dinner satisfaction data showed that I am improving my cooking skills and do not have to allocate more time to improve my cooking skills anymore. That time can be used to do more important things. The contingency table between 'fruits' and 'dinner satisfaction' showed that these variable does not influence each other.

The quantitative data showed the results as, I attend class 2.09 hours on average, study for 5.21 hours on average, sleep for 7.67 hours on average, leave house 0.69 times per day and receive 9.57 e-mails on average per day. The strongest correlation was between study and sleep with a correlation coefficient of 0.4754 which is a moderate relationship. Another moderate relationship was between the number of times I left the house and the number of hours I spent studying with a correlation coefficient of -0.3867 which is a negative relationship. Every other relationship was a weak relationship.

From this analysis, I can say that everything I do in a day seems to be going the way I plan it. But I guess I can go out of the house more often than I am doing right now. Maybe I will do that in the summer when the weather is warmer.