QMM 1001 Case Study 1 [20%]

SOWMYA GANESAN A00295227

## Introduction

From 8th January 2024 to 24th February 2024 I gathered personal data to analyze my daily activities. I recorded each day for the dates mentioned above for the following task variables:- Hours spent on Zoom where I attend my college online classes as well as in-person classes, Time invested in studying, sleeping hours, number of times I have filled my water bottles, whether I have listened to music or not on that particular day, How did the overall day go, Happiness on the scale of 1 to 5, number of steps I walked.

The First five variables are quantitative. As a student, most time invested is in Zoom classes and studies which involves completing assignments that take up a major part of the day. It will be fascinating to know whether my sleep schedule and amount of water consumption have any relation with my hectic work. There are days when I have sleepless nights as I have a lot of assignments to complete that sometimes can impact my water intake. Another Quantitative variable is whether I listened to music or not during the day as I love listening to music for relaxing on a busy day. The last variable is several steps I walked in a day, we can analyze if the physical activities have been less on any day and if they did then what led to this.

I also collected data on one nominal and one ordinal categorical variable. The nominal variable is how my day went whether it was busy, happy, sad, or average. This is a nominal variable because it is based on categories that don’t have intrinsic order to them. I was interested in this data because I could get to know how the overall day went on the particular date as well as what were the circumstances contributing this to happening. The ordinal variable is the happiness factor based on a scale of 1 to 5 how much happy I was on that day as well as I was interested in this factor as I wanted to know did the overall day has any connection to the happiness level.

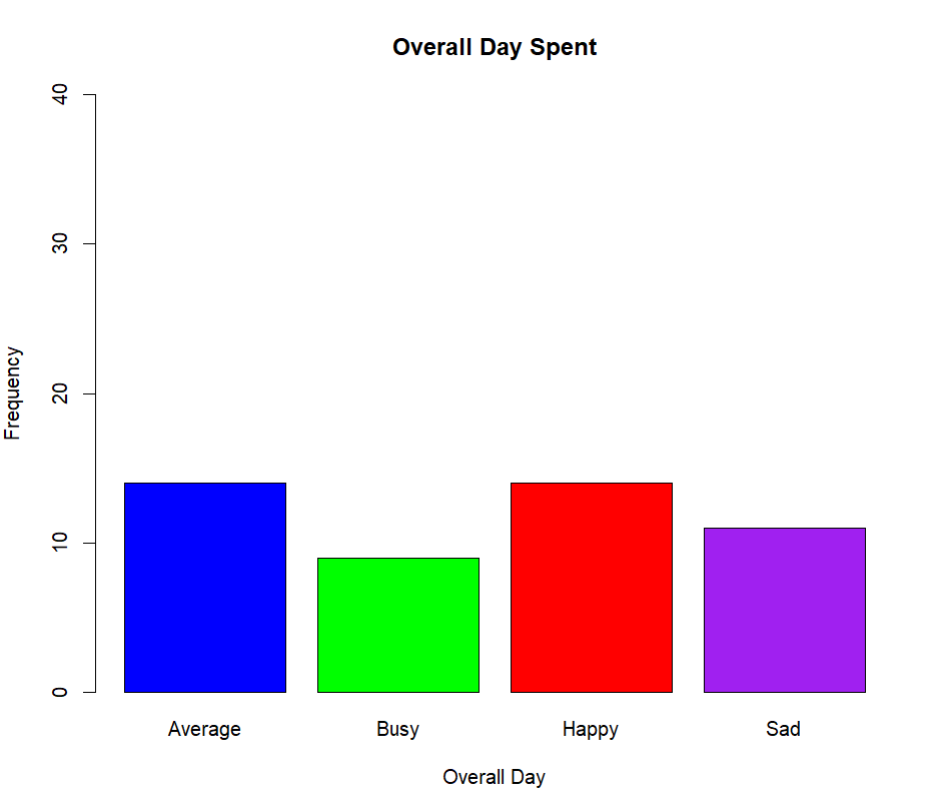
I hope after this analysis I can get better track of my daily activities as better understand what I do in a day, also establish whether any of these variables are correlated with one another.

**Categorical Variables**

1. **Overall Day Spent (Nominal Variable) –**

The first categorical variable is nominal that is based on how the overall day spent has different categories like ‘Busy’,’ Average’,’ Sad’, and ‘Happy’ with no inherent ranking . From Table-1 we can see Average and Happy days percentages are similar with 29.17% as well as have the highest frequency out of all the categories. Lowest day percentage is 18.75% i.e. busy days were very less during the data collection period.

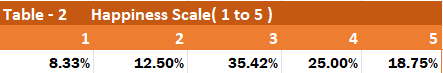
In Figure -1 We can see from the bar graph that happy and average days are similar this can be due to the fact that in our college, given the numerous events happening concurrently with our assignments, we find ourselves in a situation where we can get some time off from work . The Sad days percentage is 22.92% that can be caused with various factors like lack of sleep,fatigue, untimely food consumption etc.

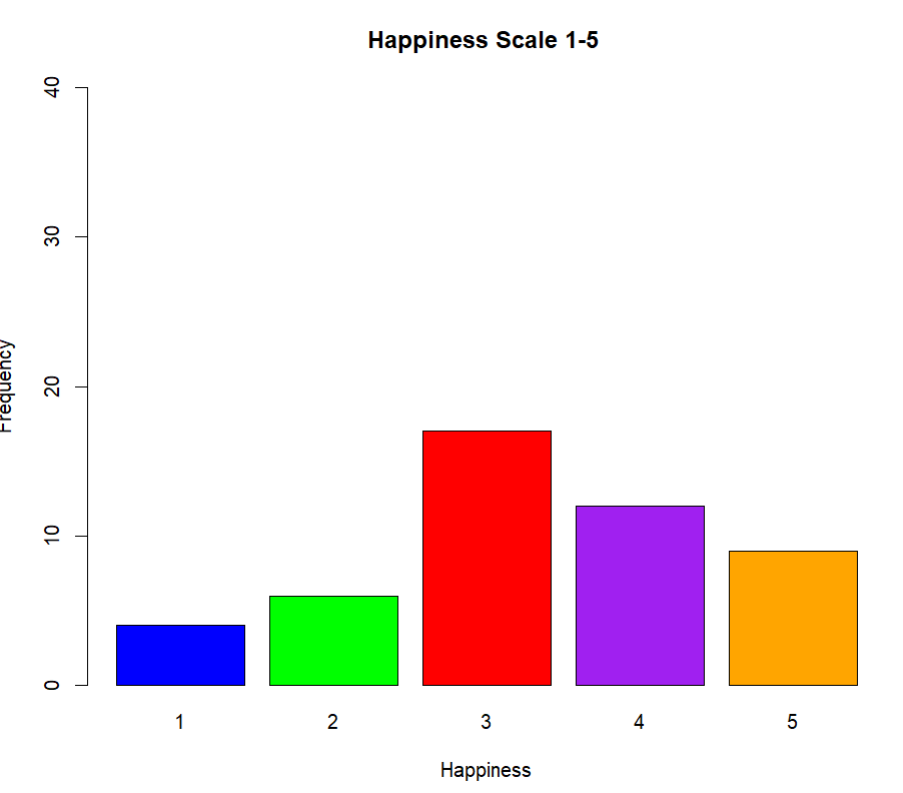
 



1. **Happiness Scale (1 to 5 ) ( Ordinal Variable) -**

**The second categorical variable is ordinal shows what is the happiness level on a specific day from a scale of 1 to 5 where** 1 represents the lowest level of happiness and 5 represents the highest with a clear order or ranking. From Table -2 we can see that from 1 to 5 , 3 has the highest percentage and 1 has the lowest percent with 8.33%

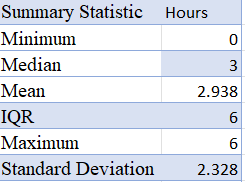
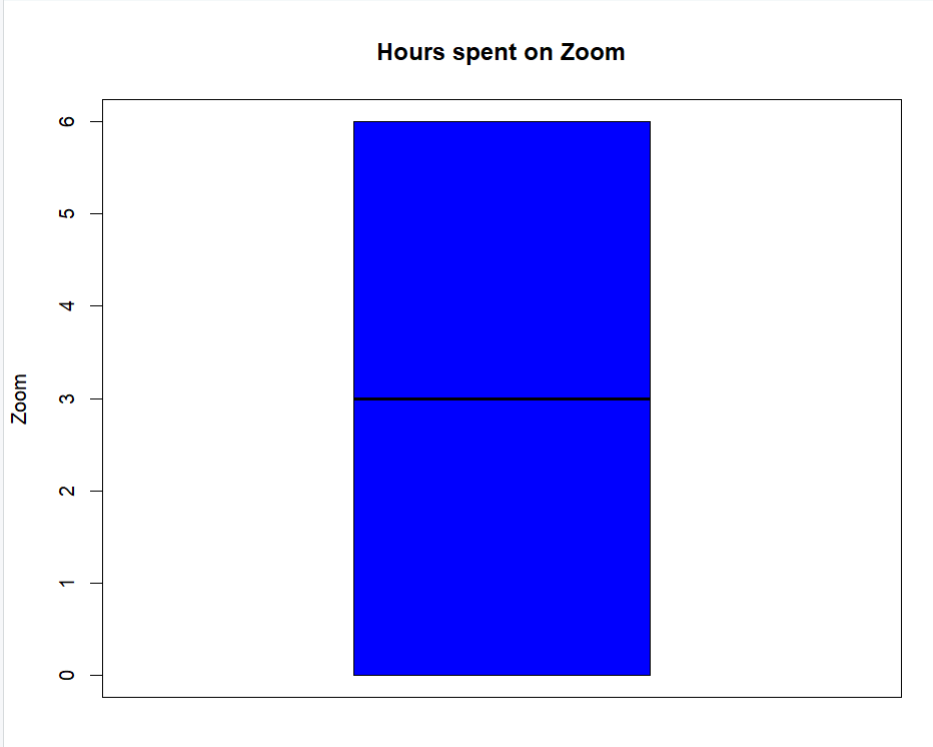




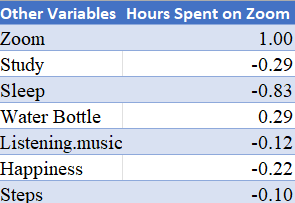
**Quantitative Variables**

1. **Zoom ( Quantitative Variable 1 )**

The first quantitative variable in my personalized data is Zoom which indicates the number of hours I spent online as well as offline college classes. The below histogram illustrates this variable wherein it shows a roughly symmetric distribution. The mode (highest peak ) was between 2 and 3 hours.The peak with the second highest frequency was for the days in which I spent 1-2 hours in class. The high frequency of days with 1 or 0 hours of class is likely due to the weekends when I did not attend class. The boxplot also shows that 50% of my data with the median being centered between the quartiles. Because this data is symmetric, the mean and the standard deviation would be the best measures of centrality and dispersion. It can be seen in Table, that the average amount of studying per day was 2.938 hours. The standard deviation was 2.328. The maximum Hours spent studying is 6 and the least is 0 which can be weekends. To determine if there were any outliers in the data the z-scores were calculated, and any data that received a z-score of +/- 3 would be considered an outlier. It was determined that there were no outliers in this data set. The boxplot in Figure 7 reiterates this notion.

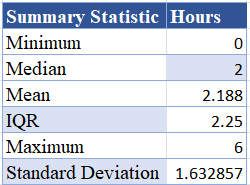
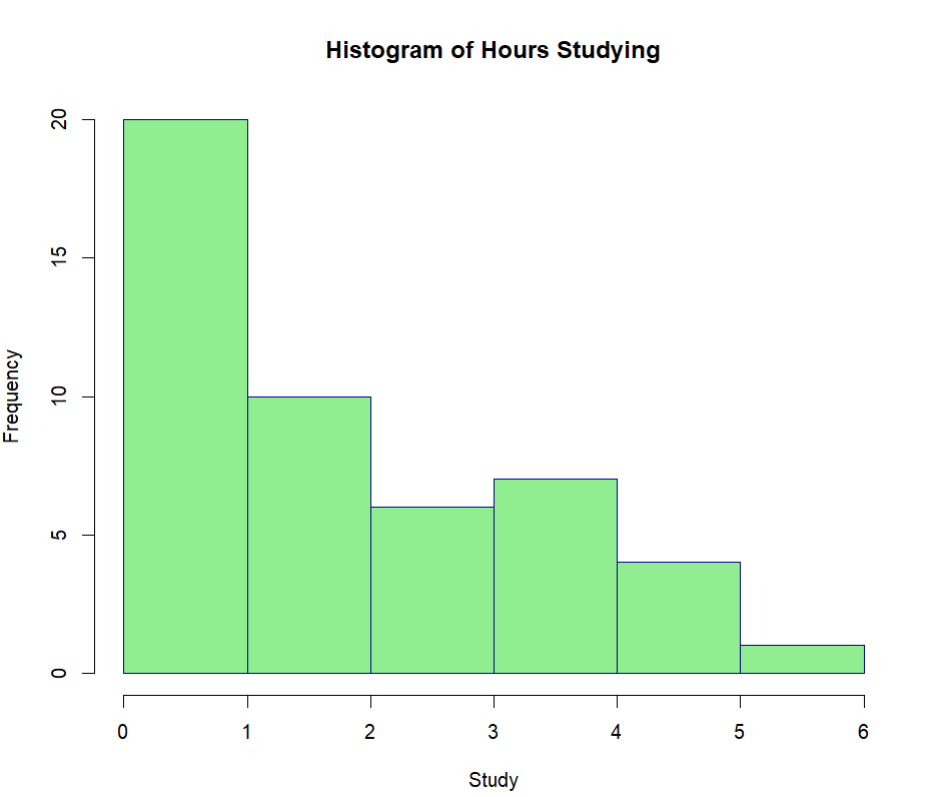
 

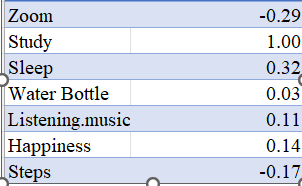
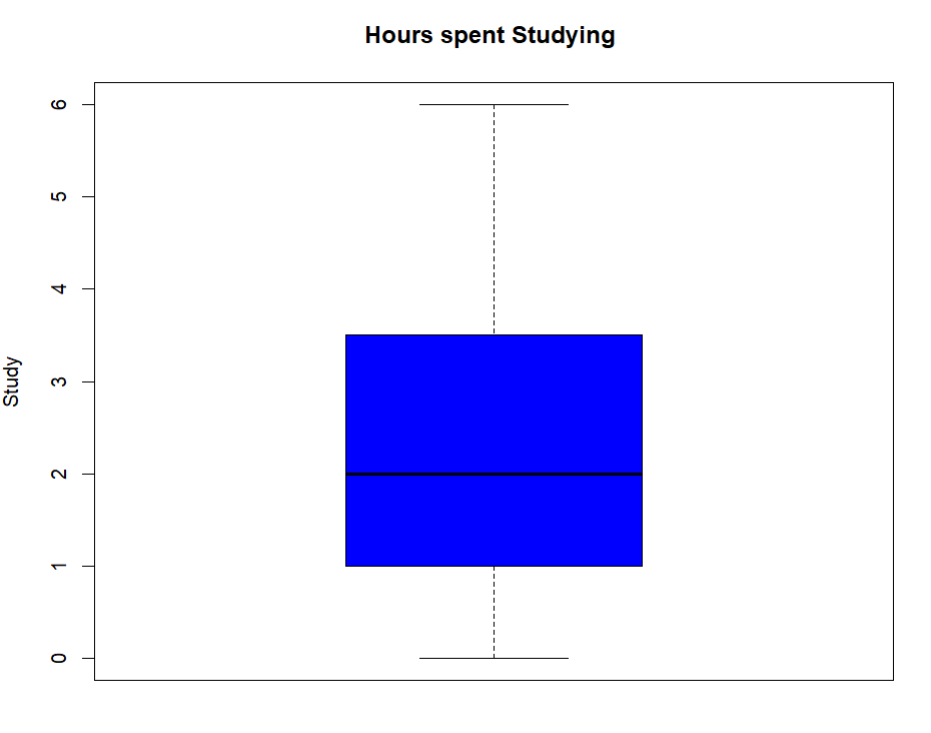
Below table is the Correation coefficient table that shows the correlation between hours spent on Zoom and the other quantitative variables in my data set. It can be seen, that the highest correlation that existed was with a number of water bottles filled with a positive correlation coefficient of 0.29. This can mean that increased hours spent on Zoom can increase the number of water bottles filled during the day .

1. **Study ( Quantitative Variable 2 )**

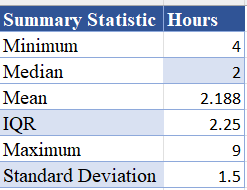
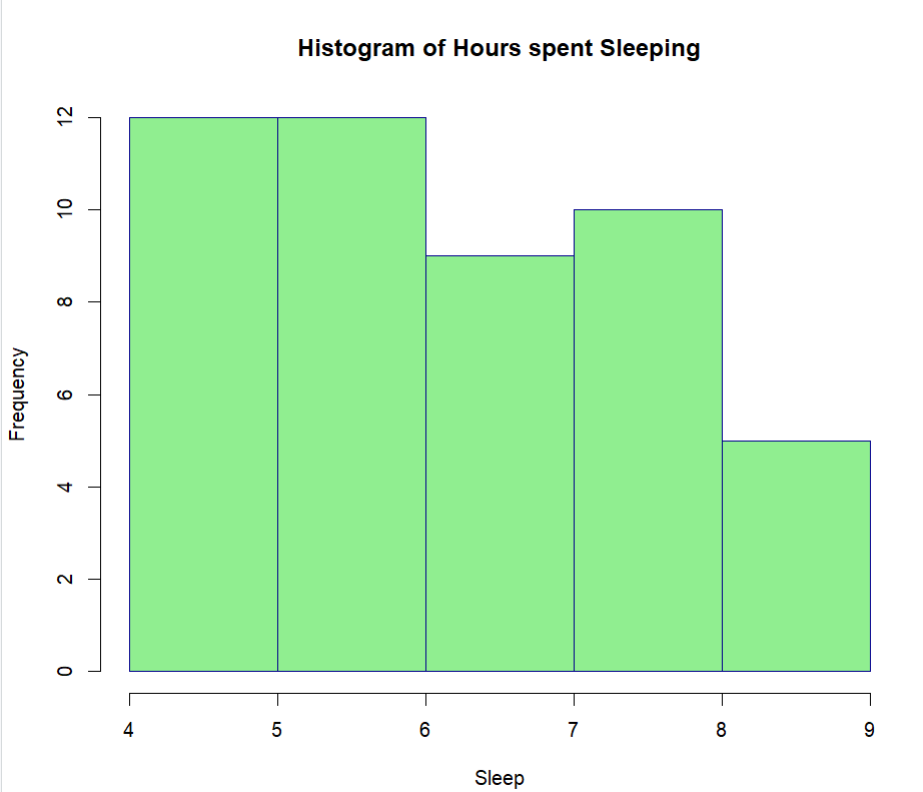
The Second quantitative variable in my personalized data is Study which indicates the number of hours I spent studying in a particular day apart from college. The below histogram illustrates this variable wherein it shows an asymmetric distribution right-skewed. The mode (highest peak ) was between 0 and 1 hours. The peak with the second highest frequency was for the days in which I spent 3-4 hours studying. The high frequency of days with 1 or 0 hours of study is likely due to the greater number of classes on that day wherein most of the time is spent in classes. The boxplot also shows that 50% of my data with the median being centered between the quartiles. Because the histogram is asymmetric, the median and interquartile range (IQR) are the best measures to use for centrality and dispersion. The summary statistics table for this variable shows a table that the median was 2, which indicates that the middle value in my data was 2 hours in studies. The IQR for the data was 2.25, Since the IQR was used to determine dispersion, the presence of outliers was also determined this way. Generally, outliers are values that are 1.5 times the IQR. Utilizing this method and looking at the boxplot, it can be determined that there are no outliers in this data. The correlation Table shows the correlation between the hours spent in studying and the other quantitative variables collected. It can be seen from the table that the strongest correlation that existed for hours spent studying was with the hours I spent sleeping. These two variables had a correlation coefficient of 0.32. Although it is the strongest correlation that exists, it is a very weak, positive relationship and it is very unlikely that these variables are related to one another.

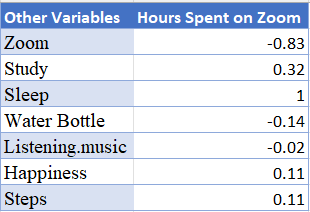
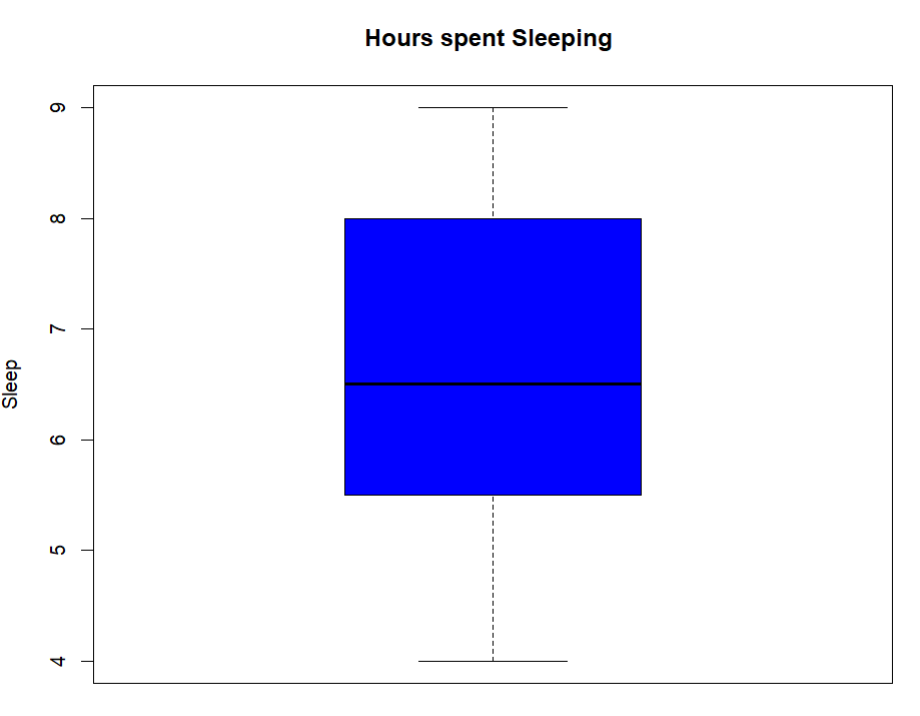
 

1. **Sleep ( Quantitative Variable 3)**

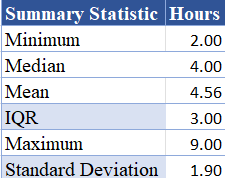
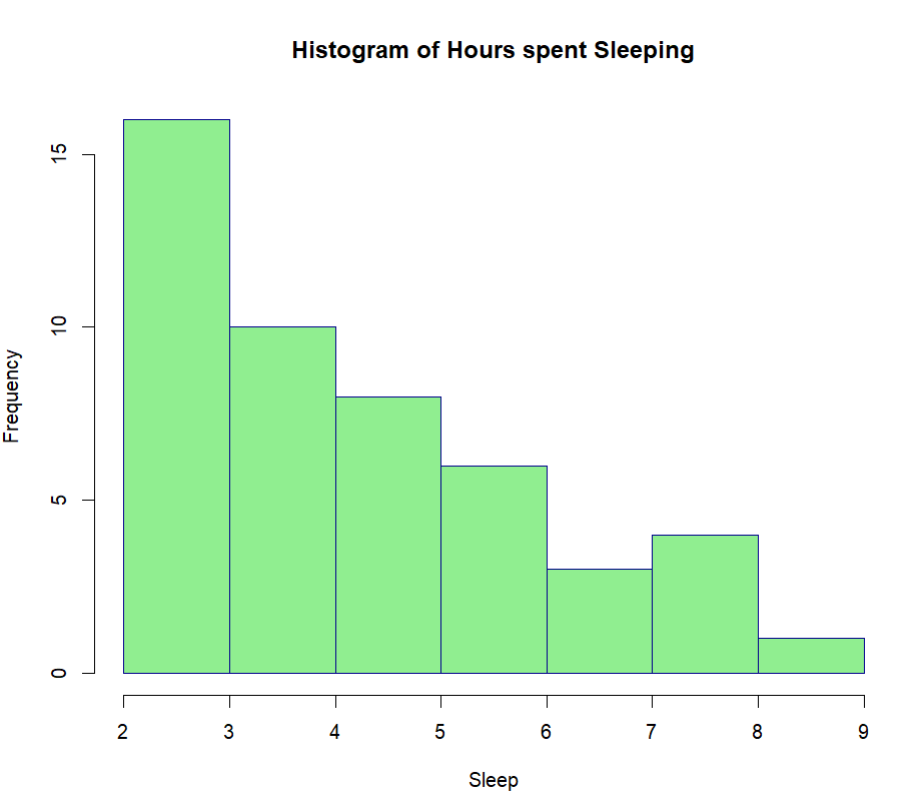
The third quantitative variable in my dataset was the number of hours spent sleeping. It can be seen in the histogram below that the data collected for this variable was symmetric. There are two highest highest peak with the same data representing the mode being between 4 and 5 hours The box plot in reiterates that the data is symmetric, showing that the median is slightly downwards between the quartiles and the ‘fences’ . Because this data is symmetric, the mean and the standard deviation would be the best measures of centrality and dispersion. It can be seen in summary statistics table , the average amount of sleeping per day was 6.5 hours. The standard deviation was 1.5 . To determine if there were any outliers in the data the z-scores were calculated, and any data that received a z-score of +/- 3 would be considered an outlier. It was determined that were no outliers in this data set. The boxplot reiterates this notion. Correlation coefficient table shows the correlation between hours spent sleeping and the other quantitative variables in my data set. It can be seen, the highest correlation that existed was with hours spent on study with a correlation coefficient of 0.320. This can imply that if I spend more number of hours sleeping I can concentrate more on my studies as with good mental health concentration power increases.

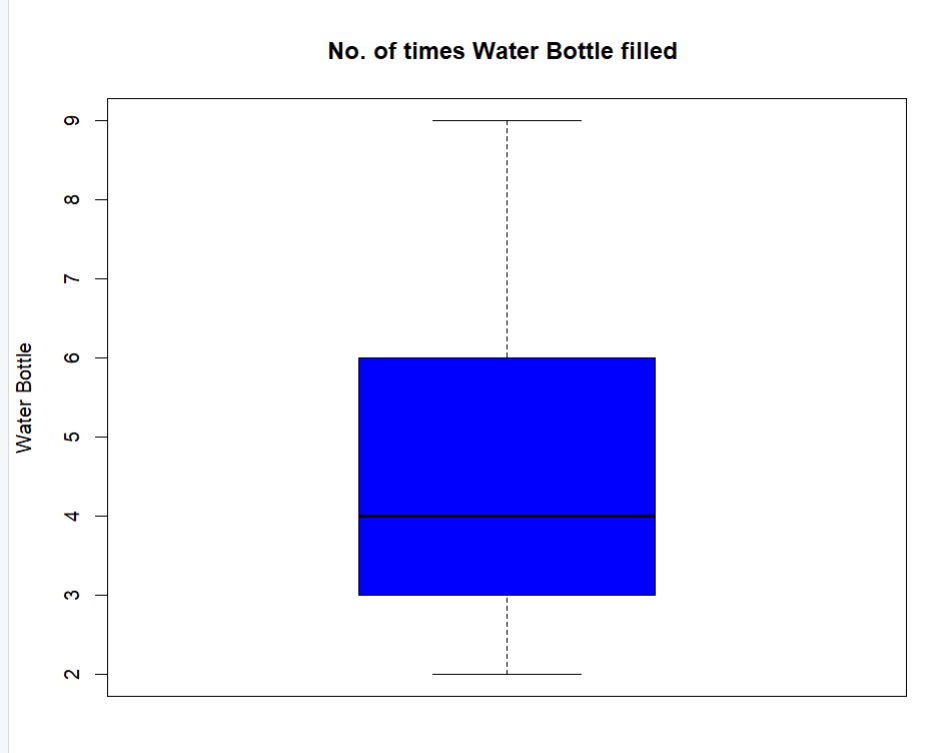
 

1. **Water Bottle ( Quantitative Variable 4)**

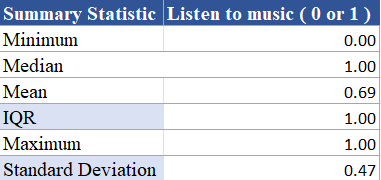
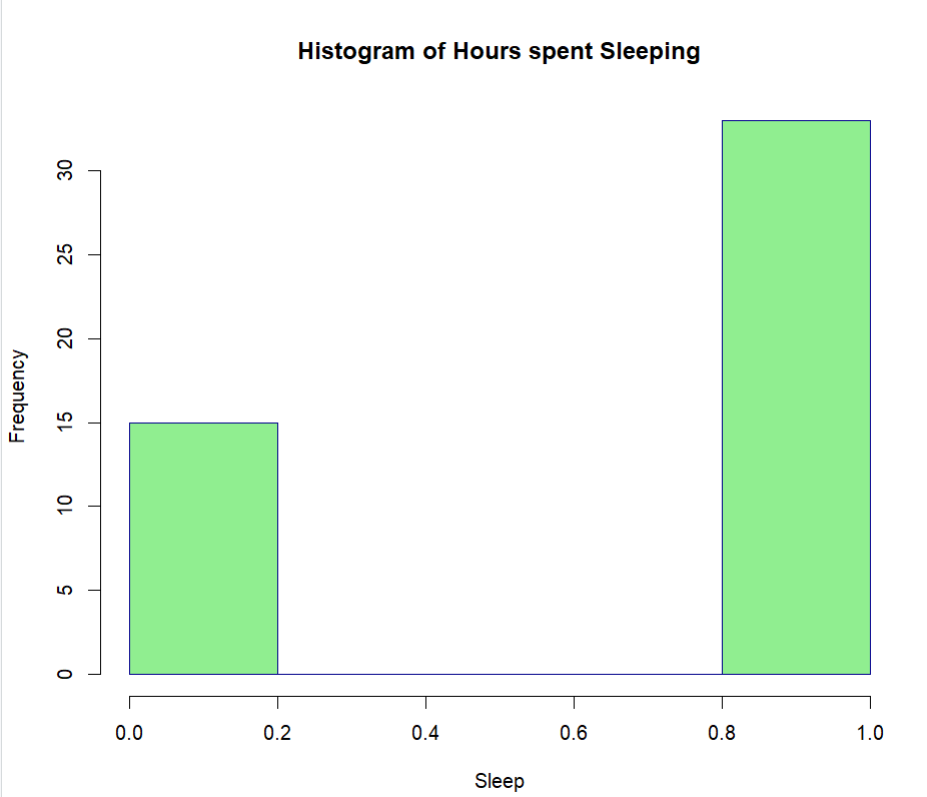
The Fourth quantitative variable in my personalized data is No. of Water Bottles filled which indicates the number of water bottles I filled in a particular day. The below histogram illustrates this variable wherein it shows an asymmetric distribution right-skewed. The mode (highest peak ) was between 2 and 3. The peak with the second highest frequency was for 3 and 4 bottles . The boxplot also shows that 50% of my data with the median being downwards towards the quartiles. Because the histogram is asymmetric, the median and interquartile range (IQR) are the best measures to use for centrality and dispersion. The summary statistics table for this variable shows a table that the median was 4, which indicates that the middle value in my data was 4 bottles in a day . The IQR for the data was 3.00, Since the IQR was used to determine dispersion, the presence of outliers was also determined this way. Generally, outliers are values that are 1.5 times the IQR. Utilizing this method and looking at the boxplot, it can be determined that there are no outliers in this data. The correlation Table shows the correlation between the hours spent in studying and the other quantitative variables collected. It can be seen from the table that the strongest correlation that existed for Water bottles was with the hours I spent studying. These two variables had a correlation coefficient of 0.03. With the increase in studying hours, our body gets tired which may lead to exertion. So a high quantity of water intake is required.

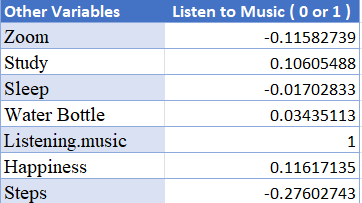
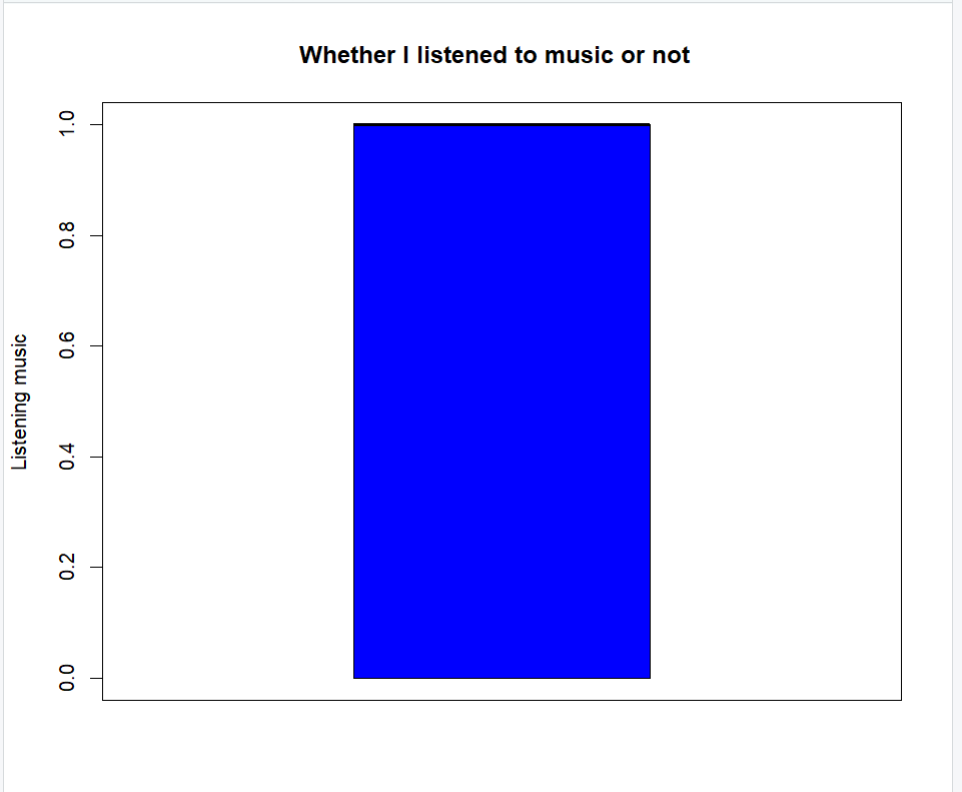
 

** **

1. **Listening Music ( Quantitative Variable 5)**

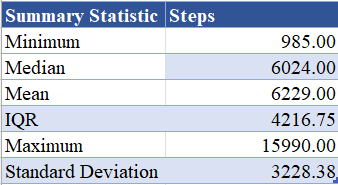
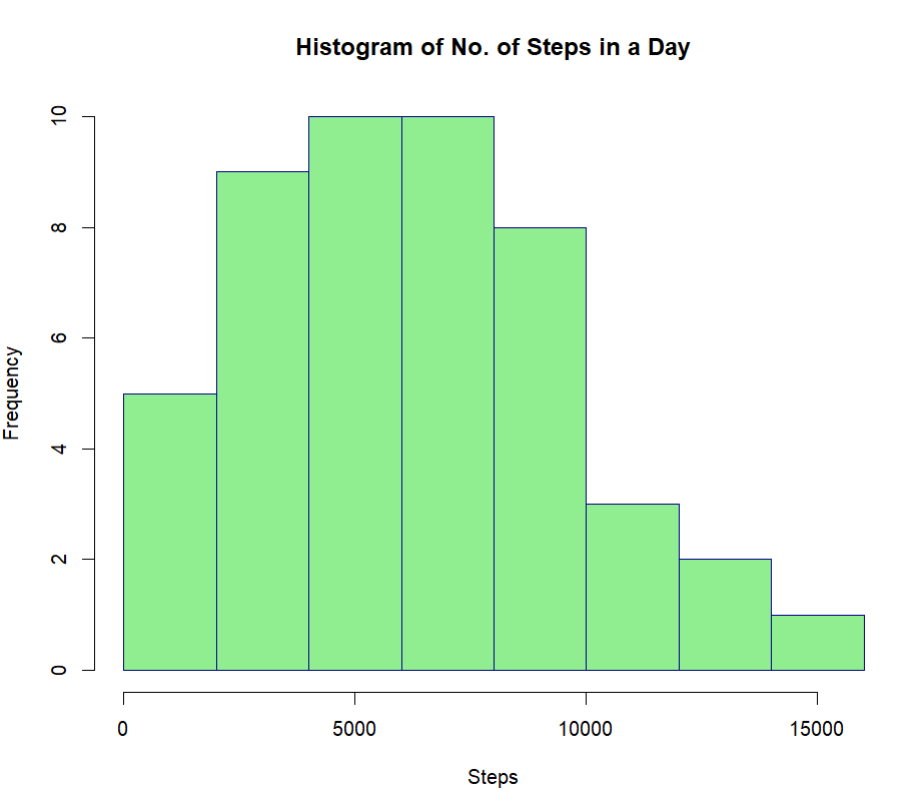
The first quantitative variable in my dataset was whether I listened to music in particular or not .The below histogram illustrates this variable. It can be seen in the graph that whether I listened to music showed an symmetric bimodal distribution . The mode (highest peak) was between 0.8 and 1.0. Because this data is symmetric, the mean and the standard deviation would be the best measures of centrality and dispersion. It can be seen in Table, that the average amount of listening music per day was 0.69 hours. The standard deviation was 0.47 hours . Below table is the Correation coefficient table that shows the correlation between listening music and the other quantitative variables in my data set. It can be seen, that the highest correlation that existed was with happiness with a positive correlation coefficient of 0.12. This can mean that if I listen to music on a particular day my happiness level increases.

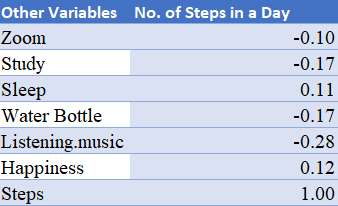
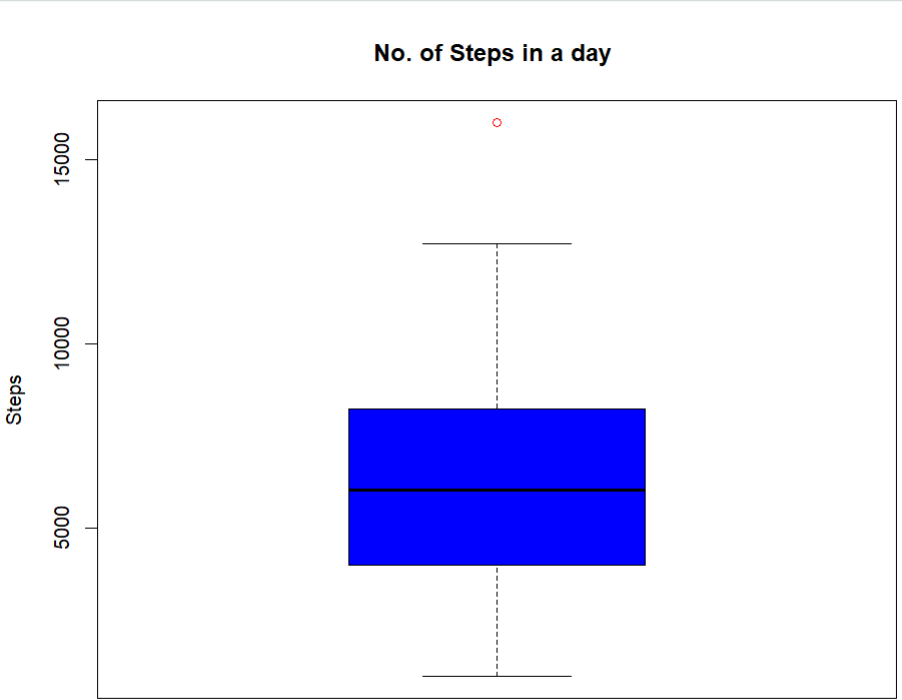
 

** **

1. **Steps ( Quantitative Variable 6)**

The Sixth quantitative variable that I collected was the number of steps I walked in each day. The histogram below illustrates the distribution of the data. The distribution appears to be symmetrical and bimodal around the mode of between 5000 and 10,000 steps . The median is evenly centered between the quartiles, which displays symmetry. The maximum and minimum values represented with the ‘fences’ are a relatively equal distance from the IQR. The summary statistics shows for the data collected for this variable. Since the data is symmetric, the mean and standard deviation are the best measures of centrality and dispersion. Due to the outliers that are present, it may however be more beneficial to use the median and IQR to represent the data set as they are not as easily affected by outliers. In this case, the mean and median are very close to one another, so either would be a good representation of centrality. On average, I walked 6229 steps in a day with a standard deviation of 3228.38 steps.  Both the standard deviation and IQR show little dispersion in the data set. Again, the z-scores were calculated, and 1 outlier was found. One outlier was a day that I received 2 hours of sleep and another outlier was a day in which I walked 15990 steps in a day . This outlier maybe because on that particular I was finding an apartment for myself which requires lot of walking in various places. The below table shows the correlation between the no. of steps I walked with the other quantitative variables. It shows No. of water bottles filled has the highest correlation coefficient with the steps walked.

**Quantitative Variables with Highest Correlation:**

It can be seen in Table 15 that the two quantitative variables in my data set with the highest correlation were the hours I spent Studying and the hours I spent sleeping ,The correlation coefficient was 0.32 , which represents a very weak, positive relationship between the variables. Potentially, this could suggest that If I spent more time in sleeping then I could concentrate more on studying which is true because if the mind is relaxed,stress-free one can focus on the work they are doing , The first condition states that both variables need to be quantitative, which is true in this case as the values collected for both variables were unique integers. The second condition of regression to be met is the linearity of the data, as correlation can only measure the strength of linear relationships.

Conclusion

### After analyzing the data collected during the period 8th January 2024 to 24th February 2024, I  can better understand how I spend my days.

Analysis of the categorical variables resulted in the following conclusions see Average and Happy days percentages are similar with 29.17% as well as have the highest frequency out of all the categories. Lowest day percentage is 18.75% i.e. busy days were very less during the data collection period. The second categorical variable is ordinal shows what is the happiness level on a specific day from a scale of 1 to 5 where 1 represents the lowest level of happiness and 5 represents the highest with a clear order or ranking. From Table -2 we can see that from 1 to 5 , 3 has the highest percentage and 1 has the lowest percent with 8.33%

### Analysis of the quantitative variables yielded the following results. For Zoom The mode (highest peak ) was between 2 and 3 hours.The peak with the second highest frequency was for the days in which I spent 1-2 hours in class. Study which indicates the number of hours I spent studying in a particular day apart from college. The below histogram illustrates this variable wherein it shows an asymmetric distribution right-skewed. The mode (highest peak ) was between 0 and 1 hours. The peak with the second highest frequency was for the days in which I spent 3-4 hours studying. For sleeping There are two highest highest peak with the same data representing the mode being between 4 and 5 hours.

For Water Bottle filling With the increase in studying hours, our body gets tired which may lead to exertion. So a high quantity of water intake is required.For Listening music the Correation coefficient table that shows the correlation between listening music and the other quantitative variables in my data set. It can be seen, that the highest correlation that existed was with happiness with a positive correlation coefficient of 0.12. This can mean that if I listen to music on a particular day my happiness level increases. For no. of steps walked It shows No. of water bottles filled has the highest correlation coefficient with the steps walked.

The above analysis shows the day to day activities which I do and what are the factors inter-related to each other as well what I can improve to have a better lifestyle .

### 