Social Media's Impact on Mental Health Analysis
STA 9750
Section: UWA

Group 2:
Mohammed Fayyaad Buharie (mohammed.buharie@baruchmail.cuny.edu)
Vrinda Arora (vrinda.arora@baruchmail.cuny.edu)
Margie Jurado Espinoza (margie.juradoespinoza@baruchmail.cuny.edu)
Phuong Trang Tran (phuongtrang.tran@baruchmail.cuny.edu)

1. Real Problem Scenario

Our approach to finding a meaningful topic that presented a real life problem came down to determining what aspects of our daily life contributed towards our overall well-being. Immediately, social media came to mind in this very technologically savvy day and age where we have access to everything online in our pockets. Social media allows you to constantly be connected with all your loved ones and even strangers online 24/7. We wanted to explore the impact of this ease of access on mental well-being. Whether that is due to the ability to see live world events and catastrophes unfolding, comparing yourself to others, the realm of influencers, online bullying, etc. Therefore, we found two datasets to aid with our statistical analysis.

The first dataset had unique responses from 791 adults via survey collected in 2021. The responses were collected from all individuals across the country Bangladesh regardless of sociodemographic background. The survey determined that Facebook is the most popular social media in Bangladesh with 669 Facebook users and 122 non-Facebook-users aged between 15 to 40 years in this data set.The survey related to the four dimensions of psychological distress including depression, anxiety, loneliness, and sleep disturbances. This dataset first started off by asking questions about how long the individuals spent on social media, which platforms they used, purpose of using, and how many friends/groups they had. Then the survey started asking questions such as if seeing a post emotionally influences them, if they feel isolated/lonely, feel depressed, have trouble falling asleep, worrying, etc. We really liked how this survey first gathered logistical data on the usage of social media and then dived into the impacts it had on their mental health. It was a very holistic dataset and covered many sides of mental well-being including factors such as sleep and appetite. Overall, this gave us a very good view to conduct our analysis.

The second dataset had 481 responses from various groups via survey to discuss the correlation between amount of time spent on social media and general mental well-being. Around 284 of the users fall in the age range of 21-28 years old, fairly evenly split among males and females and 61% are university students. After conducting the survey, machine learning techniques were used to create a predictive model to determine whether the individual should seek professional help. We used this dataset to obtain more quantitative data to aid with our regression analysis and overall to have a more solid analysis. This dataset starts off with demographic data, how much social media is used and which platform. It then asks users to rate on a scale from 1-5 on questions such as whether you get distracted by social media, if you compare yourself to others, purpose, and seeking validation. These impactful questions truly help to paint the picture on the impacts on well-being.

Both of these datasets in conjunction will help us to determine whether social media has an impact on mental health and we will conduct our analysis using RStudio for data analysis such as association and regression analyses.

2. Data Cleaning

In order to use these datasets, we conducted a variety of data cleaning techniques to begin with so that we had standardized data we could work with to conduct our statistical analyses.

To start off, we first imported the csv files into R. Then we created data frames for the files. The first thing we did when examining the data was realize how many extra columns the first dataset had that were not very relevant to our analysis. These were columns like "which type of internet connection do you use?" or super specific questions such as "what time do you wake up in the morning?" Since it was a lengthy dataset, we removed around 37 of the 76 columns. We used this function to help us quickly remove:

socialmediadata[,56:69] <- NULL

Next, we wanted to rename the columns because some of them had very lengthy names that were survey questions such as "Do you think, your mental wellbeing would be better if you do not use social media?" We thought it would make our analysis easier to shorten these so we could call on these columns when running the association or regression commands. We used the following to rename some of these columns:

colnames(socialmediadata)[18:24]<-c("Gender","Education","Profession","Monthly income","Area of residence","Living with","Smoking habit")

Furthermore, we transformed some variables from numerical values to factor variables. For example, column 9 asked "how many friends do you have on social media?" and the survey responses were categorical with fixed responses such as "less than 500" or "500-2000", etc. We transformed columns like this to factor variables so that we could easily create tables and plots to view the frequencies of the responses. We used the following query to create into factors:

socialmediadata[1:36]<- lapply(socialmediadata[1:36],factor)
lapply(socialmediadata,class)
summary(socialmediadata$Profession)

Lastly, some additional techniques we used were reformatting and filling in missing values. For example, there was a timestamp column in the first dataset that had the MM/DD/YYYY as well as the hours, minutes and seconds. We wanted to remove the time aspect so we could just have a clear date of when the survey record was submitted. Therefore, we altered the date format to accommodate for that. Additionally, one column had missing values so we used the complete cases function to help with that.

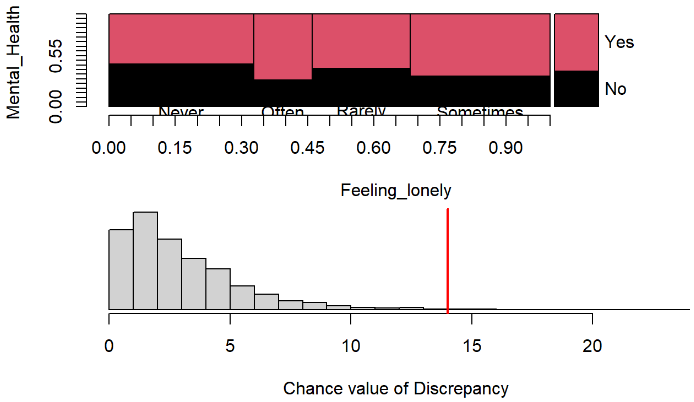
These data cleaning techniques really helped set us up better for our analyses and we also learned a lot on how to use a real life dataset. We were not aware of all the challenges we would have when beginning to clean all of this and truly examining every aspect of the data so this exercise was helpful for our growth in this class as well.

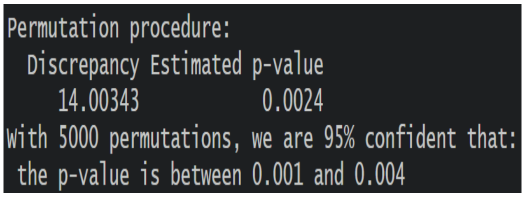
3. Association Analysis

For the association analysis, our dataset consisted of 791 unique responses from a survey conducted in Bangladesh. Our goal is to find if there is an association between the use of social media and the effect on an individual's mental health.

We tested our association with "Mental_Health" as our y-variable to five different x-variables. The y-variable is based off of the following question: Do you think your mental wellbeing would be better if you do not use social media? We found 2 variables that described a strong association with the predictor variable as discussed below:

3.1. Associating Mental Health with the Feeling of Loneliness:





```
Permutation procedure:
  Discrepancy Estimated p-value
     14.00343           0.0024
With 5000 permutations, we are 95% confident that:
 the p-value is between 0.001 and 0.004
```

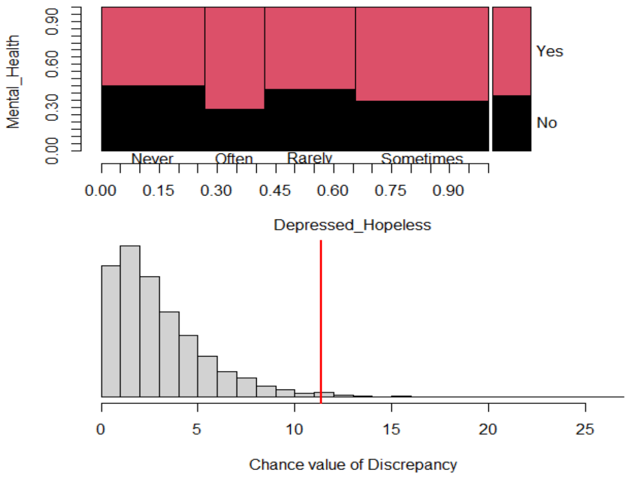
Figure 3.1: Is there an association with mental health and the feeling of loneliness?

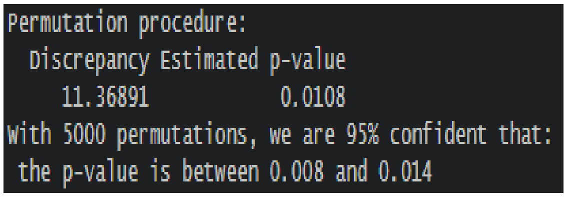
In our test, **Figure 3.1** displays the association of mental health with the feeling of loneliness.
We can correctly conclude that there is a significant association between the two variables as the data provides the estimated p-value at 0.0024. This may suggest that users of social media feel lonely as they use social media as they may not have a lot of friends, or are introverted in nature. These feelings of anxiety and loneliness play a significant role in the mental well-being of a person.

We ran the data using the following code*:*

associate(Mental_Health~Feeling_lonely,data=socialmediadata,permutations = 5000)

3.2. Associating Mental_Health with Feeling Depressed





```
Permutation procedure:
  Discrepancy Estimated p-value
     11.36891           0.0108
With 5000 permutations, we are 95% confident that:
 the p-value is between 0.008 and 0.014
```

Figure 3.2. Does Social Media Usage contribute to feelings of depression

This figure represents the use of social media and its association with feelings of depression and hopelessness. Similarly to Figure 3.1., we can see that there is a strong association between these two variables. In that the p-value is 0.0108 which is less than the alpha at 0.05.This suggests more individuals are likely to feel depressed through using social media and there may be a number of factors that contribute to it.

We ran the data using the following line of code:

associate(Mental_Health~Depressed_Hopeless,data=socialmediadata,permutations = 5000)

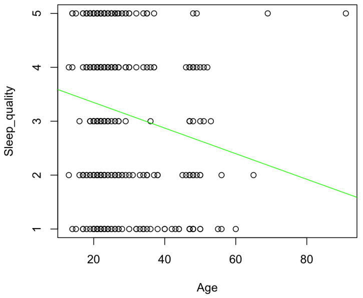
4. Regression Analysis

Our regression analysis is based on the second dataset, which had 481 responses from various groups due to the survey that discussed the correlation between social media use and general mental well-being. The dataset used machine learning techniques to create a predictive model to determine if the individual should seek professional help. Based on this dataset, we created four different models.

4.1. Regression Analysis on Sleep quality vs Age (Model 1)

4.1.1. Linear Regression Model

m1 <- lm(Sleep_quality~Age, data = smmh)
Our first model we created a linear regression analysis on the variable Sleep_quality and age. The scatter plot of the data points for Sleep_quality, the dependent variable, and Age, the independent variable. A scatter plot identifies if there are possible relationships between two quantitative variables. This plot shows that most of the people that got surveys are in their twenties and thirties. Mostly young people are using social media, so it does not show any significant correlation between sleep_quality and age. Sleep quality seems to decrease, the older the individual gets, which suggests a negative relationship.

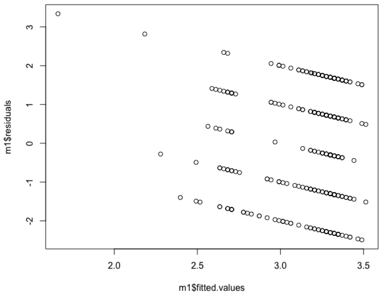


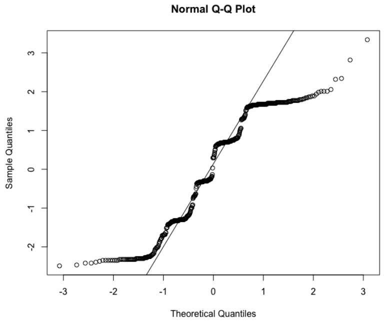
4.1.2. Correlation Analysis

Our correlation analysis we created by running the command:*associate(Sleep_quality ~ Age, data = smmh, permutations = 1000).* After conducting this code, the results were that the calculated Pearson’s correlation coefficient is approximately -0.1611 and the p-value associated is 0. While the Spearman’s rank correlation coefficient is -0.1366 and has an associated p-value of 0 as well.
It suggests a weak negative linear relationship between Sleep_quality and Age in both and the observed correlation is statistically significant. However, this model seems to have linear regression, so we suggest using Pearson’s correlation instead of the Spearman’s rank one. The t-value shows that our p-values are significant. The model 1 has a Residual Standard Error of 1.44 with a degree of freedom of 479. A lower RSE indicates a better model; in this summary the observed values deviate from the predicted values by 1.444 units. The F-statistic test shows the overall significance of the model and with the associated p-value of 0.0003884 determines that this model is statistically significant. Summary of m1 illustrates a statistically significant relationship between Sleep_quality and Age. If the coefficient of Age increases, the one of Sleep_quality decreases. Importantly, the R-squared value is low, which demonstrates that it has a small proportion of the variance in Sleep_quality.

4.1.3. Residual Analysis

For our Residual Analysis Plots we conducted a scatter plot of residuals against fitted values, residual against Age, a QQ plot and time series plot. Our first plot illustrates the residuals vs fitted values, there is a negative pattern. The plots are parallel to each other, which can be seen as homoscedasticity, where the variance is constant. Our Q-Q plot compares the quantities of the residuals to the quantiles of a standard normal distribution. Our graph suggests that our data is under-dispersed, which means the actual distribution of residuals has less variability in the tails compared to normal distribution. The time-series plot shows the data points are randomly distributed, and indicates no correlation between age and sleep-quality.





4.1.4. ANOVA Table

The ANOVA table suggests that the overall model is statistically significant, because the p-value is small with the F-statistic. Even the variable Age is individually significant to Sleep_quality by having a p-value of 0.0003884.

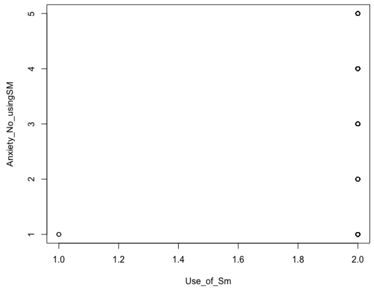
4.2. Regression Analysis on anxiety without social media vs use of social media (Model 2)

First, we converted the use of social media column into a factor in order to count if a person uses social media or not. We get 478 ‘Yes’ and 3 ‘No’.

4.2.1. Linear Regression Model

m2 <- lm(Anxiety_No_usingSm~Use_of_Sm, data = smmh)

The model2 is a linear regression model between the dependent variable “Anxiety_No_usingSm” and the independent variable “Use_of_Sm”. The plot does not show any kind of regression relation. The reason for not having any slope is because of the column of usage of social media. There are only two answer options, and people that do not use social media do not have any anxiety of using social media.

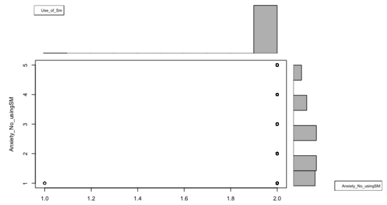


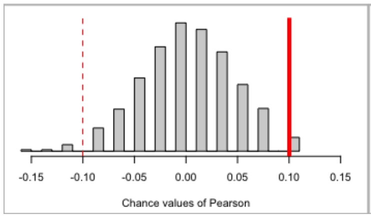
4.2.2. Correlation Analysis

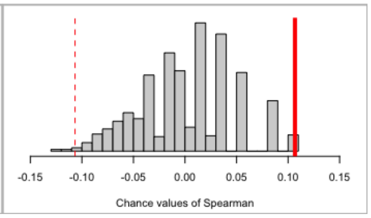
In order to run the correlation analysis we conducted this command: *associate(Anxiety_No_usingSm~Use_of_Sm, data= smmh, permutation = 1000)*

After running this code we got the association between Use_of_Sm and Anxiety_No_usingSm. The calculated Pearson’s correlation coefficient is 0.1002 and the associated p-value is reported to be 0.025. The Pearson’s Correlations indicate a weak positive linear relationship between the two variables and the observed correlation is statistically significant since it is lower than 0.05

significance level. While the Spearman's Rank Correlation reported a number of 0.1067 with a p-value of 0.023. Also, it suggests a weak positive monotonic relationship with a statistical significance of the p-value. In this model, the Spearman's rank correlation seems to be more efficient to use because of the data points.



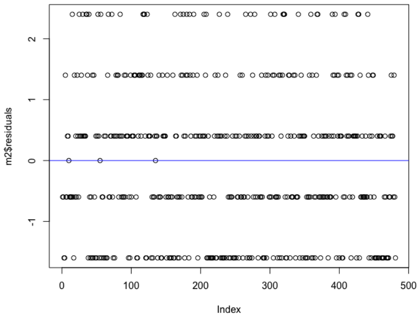




The summary of model 2 shows us the minimum and maximum residual, which are -1.5983 and 2.4017. In addition, the p-value associated with Use_of_Sm is 0.028, indicating that the variable is statistically significant at the 0.05 significance level. The RSE is 1.252 on 479 degrees of freedom. Also, the F-statistic shows the overall significance of the model and the associated p-value is 0.02799. Overall, the summary describes that there is a statistically significant relationship between Use_of_Sm and Anxiety_No_usingSM because of the F-test and small variance.

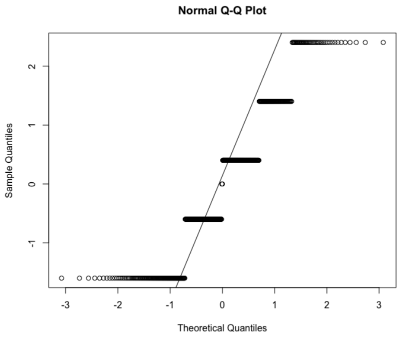
4.2.3. Residual Analysis

For our residual analysis we conducted plots for residuals against fitted values, residual against Use_of_Sm, a QQ plot for normality and time series plot again. However, the residuals vs fitted values and residual vs Use_of_Sm do not change. It looks similar to our linear regression plot because of the column Use_of_Sm of only offering two answer choices, either No = 1 and Yes = 2. However, the normal Q-Q plot seems to be interesting, because it does not show any sign of distribution. Also, the time-series plot seems to have any correlation, all the data points are randomly distributed. Besides the three data points on the horizontal line on 0. Those are the three data points of the people that do not use social media.



4.2.4. ANOVA Table

In the model 2 ANOVA table, it suggests that the variable Use_of_Sm is individually significant in predicting the other variable Anxiety_No_usingSm. According to the F-statistic value the overall model is statistically significant at the 0.05 significance level because of the F-test value of 4.8585 and the corresponding p-value 0.02799. Overall, Use_of_Sm has an impact on the prediction of the variable Anxiety_No_usingSm.



4.3. Regression on Concentration vs Comparison+Sleep Quality+Age (Model 3)

In our model 3, we look into the relationship between levels of concentration against comparison feelings, sleep quality, and age. Firstly, we convert column Use_of_Sm into a numerical column. The variables are how good people are focused against the issues of comparing to other people, sleep quality, and age.

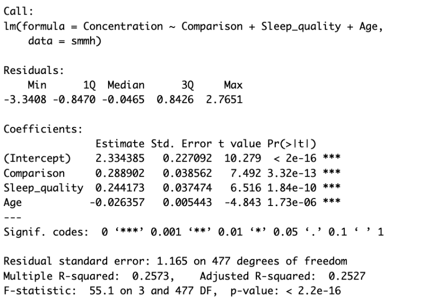
4.3.1. Linear Regression Model

The third model m3<- lm(Concentration~Comparison+Sleep_quality+Age,data=smmh). Our model is regressed between qualitative predictor like age and qualitative or indicator predictor like Sleep quality and Comparison levels.

We first demonstrate if the variables included in this model have a linear regression association with the model.

H0: B1=B2=B3 =0 or Ha: at least one of Bk is not Zero

The F value 55.1 and the p-value 2.2e-16 indicate that we should conclude our alternative hypothesis. Exists a linear regression relation between the variables included.

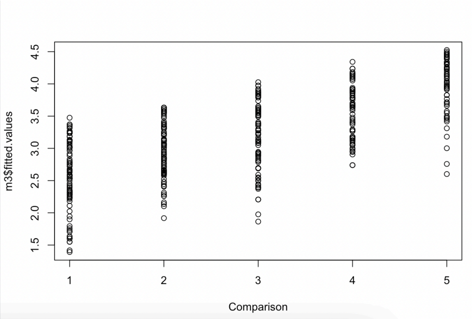


```
Call:
lm(formula = Concentration ~ Comparison + Sleep_quality + Age,
    data = smmh)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3408 -0.8470 -0.0465  0.8426  2.7651

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.334385   0.227092  10.279  < 2e-16 ***
Comparison     0.288902   0.038562   7.492 3.32e-13 ***
Sleep_quality  0.244173   0.037474   6.516 1.84e-10 ***
Age           -0.026357   0.005443  -4.843 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.165 on 477 degrees of freedom
Multiple R-squared:  0.2573,    Adjusted R-squared:  0.2527
F-statistic:  55.1 on 3 and 477 DF,  p-value: < 2.2e-16
```

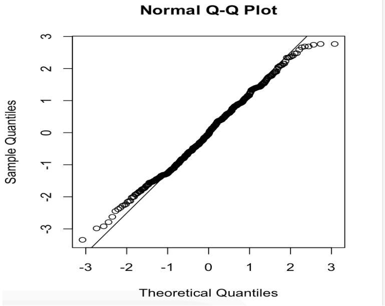


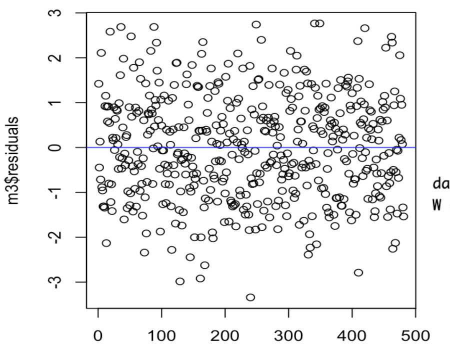
If we look at the t-statistics or p-values of each of the coefficients, all are small which indicate that they are statistically significant and contribute to explain the model.

Then, we performed a plot of fitted values against the predictor variables, to look for constant error variance or linearity violations. The coefficient of determination however is very small 25% of the variance in the response variable being studied is explained by the variance of the independent variables.

4.3.2. Correlation Analysis

The calculated Pearson’s correlation coefficient is 0.3546761 and the associated p-value is reported to be 1.051991e-15. The Pearson’s Correlations indicate a weak positive linear relationship between the three variables and the observed correlation is statistically significant since it is lower than 0.05 significance level.





```
Shapiro-Wilk normality test

data:  m3$residuals
W = 0.99452, p-value = 0.08337
```

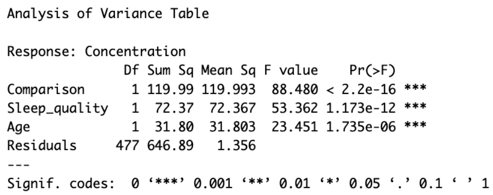
4.3.3. Normality Assumption and Collinearity:

The normal probability plot indicates that the residual does not show any significant departure from normality. The Shapiro test confirms the assumption of normality p-value of 0.08 concludes Ha: Normality. The points seem to follow a linear pattern. The time sequence plot doesn't show any specific pattern which may indicate no correlation. We will have to perform a formal test for no correlation.

4.3.4. ANOVA Table

The ANOVA table indicates that the p-values for our 3 variables help to predict the model. Also, if we analyze the SSE closely, We observe that it is reduced with the addition of a new predictor. Which will indicate that the contribution of each variable contains new information for the model.

X1
X2|X1
X3|X2,X1



```
Analysis of Variance Table

Response: Concentration
              Df Sum Sq Mean Sq F value    Pr(>F)    
Comparison     1 119.99 119.993  88.480 < 2.2e-16 ***
Sleep_quality  1  72.37  72.367  53.362 1.173e-12 ***
Age            1  31.80  31.803  23.451 1.735e-06 ***
Residuals    477 646.89   1.356                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Alternative Techniques

Some alternative techniques that we applied include the use of ggplot2 to present a better visual of individuals whose mental health is affected while browsing social media.

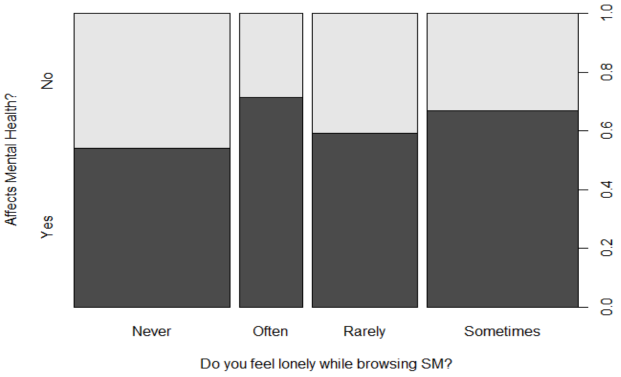


Figure 5.1.

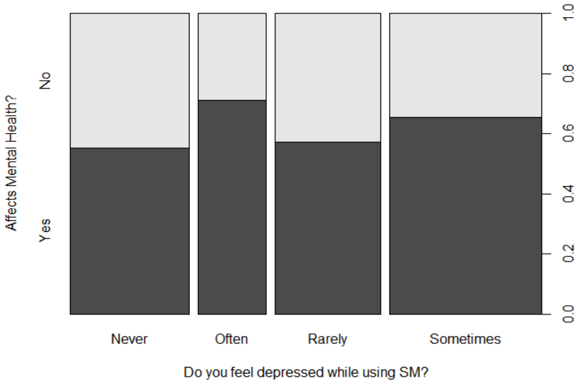


Figure 5.2.

Use of a new package called Dyplr was incorporated but scrapped later when we found it wasn't particularly useful for our analysis. The dplyr package helps typically with data manipulation and provides a set of verbs that help you solve the most common data manipulation challenges such as mutate, filter, summarize, and arrange.

6. Conclusion

Our study of the relationship between social media and mental health was successful. To sum it up, in our association analysis, we found that social media usage is tied to thoughts of feeling lonely & depressed. With our association analysis, we came to the conclusion that social media usage can contribute to feelings of loneliness due to a lack of friends or introverted nature. The second association analysis suggests that there is a potential negative impact of social media on a person's mental well-being. Individuals are more likely to feel depressed through using social media, but there are multiple other factors. In our second regression model we described the impact on people with social against people without and their anxiety level and could be related to our second association analysis. In our regression analysis, we found that the third model is explaining that if we try to explain sleeping quality only with the variable distraction level by user and feelings of comparison to others there is a significant relationship between them - the level of distraction for using social media and the emotions related to it can explain in some relation the predicted variable. These analyses are some examples of the multiple tests we generated. Both datasets have similar outcomes. Overall, we were able to utilize the statistical concepts taught in class to model a real-life scenario and dataset.