Mental Health in University and Tech Industry

Mental Health Maniacs

Amulya Rayasam, Ally De Vera, Ann Biju, Yuya Cho

Table of Contents

Data

Background

As mental health becomes an increasingly important topic of discussion in our society, we conducted research on the correlation between mental health and individuals pursuing a STEM field of study, as well as those working in the tech industry. To gain a comprehensive understanding of mental health within these fields, we examined two datasets. The first dataset pertains to mental health within academia, while the second dataset focuses on mental health in the tech industry.

In this report, we will first discuss the findings from the dataset regarding mental health in academia. Subsequently, we will delve into the findings from the dataset related to mental health within the tech industry. By analyzing these two datasets, we aim to identify common conclusions and trends concerning mental health in both academia and the tech industry.

MENTAL HEALTH IN ACADEMIA

Data Description

Student Mental Health

The first dataset was obtained from Kaggle, and it was generated through a survey administered by Google Forms, which targeted university students. The primary objective of the survey was to assess the prevalence of different mental health issues among students across various academic disciplines. The dataset comprises 101 records and 11 categorical variables, which were obtained using yes/no questions. To facilitate analysis, the yes/no responses were transformed into binary data. A logistic regression model was then constructed to explore the associations between the variables and to forecast possible outcomes.

Cleaning our Data

After examining the initial dataset, it was determined that certain variables, including: Timestamp, Gender, Age, Current Year of Study, and Marital Status, were not significant predictors of an individual's mental health. Therefore, these variables were removed from the dataset to simplify the analysis.

To further focus on the mental health of STEM students, only individuals enrolled in STEM courses were included in the analysis. The STEM programs included in the analysis were Engineering, BIT, BCS, and Biotechnology.

After these cleaning procedures were performed, the dataset was reduced to 46 observations with 8 variables.

Analysis

To ensure accurate analysis of the relationship between GPA and mental health issues such as depression, anxiety, and panic attacks in our dataset, we created a new variable by categorizing students into three groups based on their GPA scores. This categorization was necessary to make the logistic regression model non-zero and improve its effectiveness in analyzing the data. Specifically, we divided the students into low GPA group (GPA < 3), median GPA group (GPA = 3), and high GPA group (GPA > 3). By categorizing the students into these groups, we aim to provide a more comprehensive understanding of the relationship between GPA and mental health issues, allowing us to better identify potential risk factors and inform targeted interventions to support student well-being.

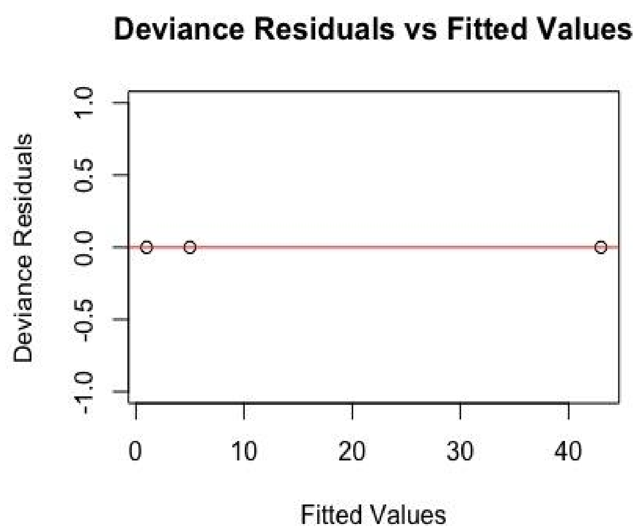
	gpa_group	depression	anxiety	panic_attack	total_issues
1	low	0	0	1	1
2	medium	2	1	2	5
3	high	15	17	11	43

The glm() model with poisson family is used to analyze count data, and the output indicates that the intercept (representing gpa_grouplow) is not statistically significant (p-value = 1), which means there is no significant difference in the log-odds of the response variable between the low and the reference group. The coefficient for gpa_groupmedium has a p-value of 0.142, which is greater than the significance level of 0.05, indicating that there is no statistically significant difference in the log-odds of the response variable between the medium and the reference group. On the other hand, the coefficient for gpa_grouphigh has a p-value of 0.000201, which is less than 0.001, indicating a statistically significant difference in the log-odds of the response variable between the high and the reference group. Therefore, it can be concluded that students with high GPAs have a significantly higher likelihood of experiencing mental health issues compared to students with low and medium GPAs.

```
> summary(model)

Call:
glm(formula = total_issues ~ gpa_group, family = poisson(), data = gpa_issues)

Deviance Residuals: 
[1]  0  0  0

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     4.676e-11  1.000e+00   0.000 1.000000    
gpa_groupmedium 1.609e+00  1.095e+00   1.469 0.141774    
gpa_grouphigh   3.761e+00  1.012e+00   3.718 0.000201 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  6.5823e+01  on 2  degrees of freedom
Residual deviance: -6.6614e-16  on 0  degrees of freedom
AIC: 17.084

Number of Fisher Scoring iterations: 3
```

Variable Selection

In our analysis, we performed residual analysis using deviance because it is a commonly used measure of goodness of fit that can identify potential outliers or influential data points that may affect the accuracy of a statistical model. Our analysis revealed no outliers in the deviance residual plot, indicating that the current model fits well and requires no data transformation or adjustment. This finding supports the appropriateness of the current model for the data at hand and increases our confidence in the results.



After confirming that our data does not require any transformation, we proceeded with variable selection using stepwise regression analysis, specifically stepAIC. The output of our analysis showed that removing gpa_group from the model increased the AIC value from 17.084 to 78.907. This indicates that the current model with gpa_group as a predictor variable is the most suitable one for our data. Therefore, we will retain gpa_group in our final model and use it for further analysis and interpretation of the results.

```
> stepAIC(model, direction = "both")
Start:  AIC=17.08
total_issues ~ gpa_group

            Df Deviance    AIC
<none>            0.000 17.084
- gpa_group  2   65.823 78.907

Call:  glm(formula = total_issues ~ gpa_group, family = poisson(), data = gpa_issues)

Coefficients:
    (Intercept)  gpa_groupmedium    gpa_grouphigh
      4.676e-11        1.609e+00        3.761e+00

Degrees of Freedom: 2 Total (i.e. Null);  0 Residual
Null Deviance:      65.82
Residual Deviance: -6.661e-16   AIC: 17.08
```

Since we are using the Poisson regression model, we need to exponentiate the model coefficients to interpret the results. Our analysis revealed that students in the high GPA group had a 43-fold increase in mental health issues compared to those in the low GPA group. In other words, students in the high GPA group were found to have a significantly greater number of mental health issues than those in the low GPA group. Additionally, the exponentiated coefficient for the medium GPA group was 5, which suggests that students in this group had a moderately higher number of mental health issues than those in the low GPA group. Overall, these findings suggest a significant relationship between GPA and mental health issues among students, with high GPA being a strong predictor of greater mental health issues.

```
> exp(coef(stepwise_model))
  (Intercept) gpa_groupmedium    gpa_grouphigh
            1               5               43
```

Further Analysis

Additionally, we discovered that getting treatment does have an effect on mental health - more specifically depression, anxiety, and panic attacks - as seen in university students. This can be confirmed with the results of the stepAIC() model used on the student mental dataset, which had a p-value of 0.995 and a low AIC score of 22.55. Both of these values point directly to the fact that getting treatment has a strong correlation with the students' mental health of students - or more specifically depression rates. Below are our results.

```
> step.mod <- stepAIC(model, direction = "both")
Start:  AIC=24.71
treatment ~ depression + panic.attack + anxiety

               Df Deviance    AIC
- anxiety       1   16.864 22.864
- panic.attack  1   18.532 24.532
<none>              16.709 24.709
- depression    1   23.307 29.307

Step:  AIC=22.86
treatment ~ depression + panic.attack

               Df Deviance    AIC
- panic.attack  1   18.550 22.550
<none>              16.864 22.864
+ anxiety       1   16.709 24.709
- depression    1   23.448 27.448

Step:  AIC=22.55
treatment ~ depression

               Df Deviance    AIC
<none>              18.550 22.550
+ panic.attack  1   16.864 22.864
+ anxiety       1   18.532 24.532
- depression    1   27.180 29.180
```

```
> summary(step.mod)

Call:
glm(formula = treatment ~ depression, family = "binomial", data = clean.students)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.73248  -0.73248  -0.00005  -0.00005   1.70113

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -20.57    3292.45  -0.006    0.995
depression     19.39    3292.45   0.006    0.995

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.18  on 45  degrees of freedom
Residual deviance: 18.55  on 44  degrees of freedom
AIC: 22.55

Number of Fisher Scoring iterations: 19
```

Reflection

Cleaning, and analyzing this dataset gave us much insight into the mental state, gpa, and other relevant factors of an average university student. After cleaning the dataset, we were quick to realize that conventional linear regression could not be used to analyze linear trends and make models from the categorical data that we were dealing with. This proved to be a great learning opportunity for us, since we did not have much prior experience working with categorical data and hence logistic regression techniques. So this project gave us just the chance to explore an area we had yet to get our hands on!

MENTAL HEALTH IN THE TECH INDUSTRY

Data Description

Mental Health in the Tech Industry

The second dataset, also sourced from Kaggle, looks at the presence of mental health issues of individuals working in tech. It also looks at how various aspects of their work affect the presence of these mental health conditions or vice versa. Initially, this data set contained 1259 entries with 27 categorical/logical variables. Each data instance measures attitudes towards mental health in the workplace, perceived consequences of mental health in the workplace, as well as how individuals identify themselves within the scope of mental health. As a result of our data being categorical, we chose to use a logistic regression model.

Cleaning our Data

Of the 27 variables contained within our initial data set, we deemed certain variables within this dataset insignificant predictors of the existence of mental health within an individual. These removed variables are as follows:

- Timestamp
- Self_employed: whether the individual is self-employed or not
- work_interfere: Does your mental health condition affect your work?
- no_employees: Number of employees in company/organization
- benefits: Does your employer provide mental health benefits?
- care_options: Do you know the options for mental health care your employer provides?
- wellness_program: Did your employer include mental health in an employee wellness program?
- seek_help: Does your employer offer resources to educate about mental health concerns and how to get help?
- anonymity: Is your identity kept confidential if you use mental health or substance abuse treatment resources?
- leave: Is it easy for you to take medical leave for a mental health issue?
- mental_health_consequence: Do you believe that disclosing a mental health issue to your employer would result in negative outcomes?
- phys_health_consequence: Do you believe that discussing a physical health issue with your employer would have negative repercussions?
- coworkers: Are you comfortable discussing a mental health issue with your coworkers?
- supervisor: Are you comfortable discussing a mental health issue with your supervisor?

- mental_health_interview: Would you mention a mental health issue to a potential employer in interview?
- phys_health_interview: Would you mention a physical health issue to a potential employer in interview?
- mental_vs_physical: Do you think your employer regards mental health as important as physical health?
- obs_consequence: Have you witnessed or been aware of any adverse effects for coworkers with mental health conditions in your workplace?

We chose to remove these variables when creating our initial logistic regression model since they concerned attitudes towards mental health within the workplace, and therefore would not be accurate predictors of the existence of mental health within an individual working in tech. However, in our later analysis, we look at work_interfere a variable not initially considered for our predictive model.
The variables we chose to keep when creating our initial logistic regression model were the following:

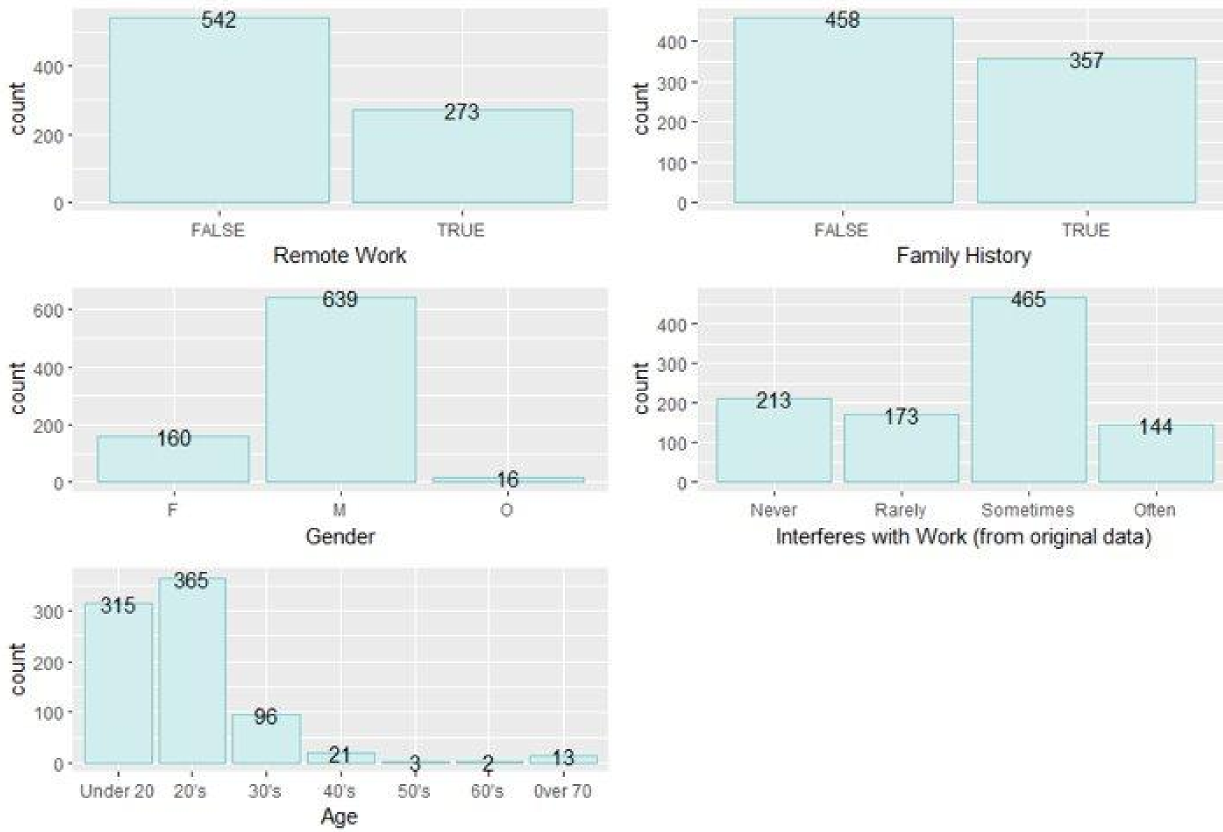
- Age
- Gender
- Country
- Family_history: family history of mental illness
- treatment: Receive treatment for a mental health issue
- remote_work: Work remotely at least 50% of the time
- Tech_company: whether the individual works in a tech company or not

Upon identifying these variables we considered valid predictors of mental health within an individual, we further subsetted our data by only looking at individuals that identified as working within a tech company. Once this data was subsetted, as well as NA values were removed, we were left with 815 instances of 7 variables. We identified treatment as our response variable, since we determined that an individual would not choose to seek treatment unless they identified as having a mental health condition. Later, in order to transform our model, we performed binary encoding on the relevant variables.

Analysis

Preliminary Analysis

We created a series of graphs in order to have a visualization of the distribution of our different predictors for our logistic regression model, looking at Country (not included below) as well.



We created this visualization in order to measure the different responses and compare between various responses. In order to create a condensed visualization of Age, we divided age into the intervals shown above. Some notable takeaways from these visualizations are the number of male responses as opposed to female and other, as well as the average age of individuals whose responses were recorded.
After we had this visualization, we proceeded to perform variable selection.

Variable Selection

We fit the largest possible logistic regression model to the data using the variables we initially identified as being valid predictors of mental health. Once this model was created, our summary of the model was as follows:

```
Call:
glm(formula = treatment ~ Age + Gender + Country + family_history +
    remote_work, data = mh_valid_tech)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0191  -0.4577   0.1658   0.3517   0.7947

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.827e-01  1.689e-01   3.450  0.00059 ***
Age                         2.143e-03  1.266e-03   1.693  0.09086 .
GenderF                     4.391e-02  1.252e-01   0.351  0.72595
GenderM                    -1.279e-01  1.215e-01  -1.052  0.29308
CountryAustria             -5.062e-01  4.601e-01  -1.100  0.27156
CountryBahamas, The         6.372e-02  4.763e-01   0.134  0.89362
CountryBelgium             -5.105e-01  4.601e-01  -1.110  0.26749
CountryBosnia and Herzegovina -8.449e-01  4.605e-01  -1.835  0.06692 .
CountryBrazil              -4.255e-01  2.494e-01  -1.706  0.08841 .
CountryBulgaria             7.739e-02  2.817e-01   0.275  0.78363
CountryCanada              -1.766e-01  1.288e-01  -1.371  0.17074
CountryChina               -5.556e-01  4.603e-01  -1.207  0.22780
CountryColombia            -5.127e-01  3.351e-01  -1.530  0.12641
CountryCroatia              2.880e-01  3.351e-01   0.859  0.39034
CountryCzech Republic      -8.469e-01  4.603e-01  -1.840  0.06615 .
CountryDenmark              1.488e-01  3.354e-01   0.444  0.65743
CountryFinland             -2.980e-02  3.351e-01  -0.089  0.92916
CountryFrance              -2.982e-02  2.498e-01  -0.119  0.90499
CountryGermany             -8.144e-02  1.400e-01  -0.582  0.56101
CountryHungary             -1.021e+00  4.615e-01  -2.212  0.02727 *
CountryIndia                1.210e-01  2.499e-01   0.484  0.62835
CountryIreland             -1.081e-01  1.468e-01  -0.737  0.46164
CountryIsrael              -7.571e-01  3.351e-01  -2.260  0.02413 *
CountryItaly               -4.400e-01  2.499e-01  -1.761  0.07867 .
```

```
CountryJapan                1.037e-01  4.609e-01   0.225  0.82201
CountryMexico              -1.873e-01  3.344e-01  -0.560  0.57557
CountryMoldova              4.895e-01  4.601e-01   1.064  0.28769
CountryNetherlands         -2.354e-01  1.514e-01  -1.555  0.12040
CountryNew Zealand         -7.784e-02  2.022e-01  -0.385  0.70033
CountryPhilippines         -5.212e-01  4.601e-01  -1.133  0.25757
CountryPoland              -7.356e-02  2.022e-01  -0.364  0.71605
CountryPortugal            -5.127e-01  4.601e-01  -1.114  0.26548
CountryRussia              -5.127e-01  3.351e-01  -1.530  0.12641
CountrySingapore           -3.661e-01  2.494e-01  -1.468  0.14255
CountrySlovenia             5.045e-01  4.602e-01   1.096  0.27337
CountrySouth Africa        -6.046e-02  2.818e-01  -0.215  0.83016
CountrySweden              -3.116e-01  2.292e-01  -1.360  0.17432
CountrySwitzerland         -8.372e-02  2.290e-01  -0.366  0.71475
CountryThailand            -5.405e-01  4.602e-01  -1.174  0.24059
CountryUnited Kingdom      -2.750e-02  1.199e-01  -0.229  0.81869
CountryUnited States       -5.986e-02  1.136e-01  -0.527  0.59826
CountryZimbabwe            -2.143e+08  1.266e+08  -1.693  0.09086 .
family_historyTRUE          3.214e-01  3.265e-02   9.844  < 2e-16 ***
remote_workTRUE             1.508e-02  3.446e-02   0.438  0.66182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1987842)

    Null deviance: 190.86  on 814  degrees of freedom
Residual deviance: 153.26  on 771  degrees of freedom
AIC: 1041

Number of Fisher Scoring iterations: 2
```

We can see we have an extremely low p-value of 0.00059, as well as an AIC score of 1041. We chose to perform stepwise AIC variable selection in order to find the best fit model to our data. Below is a summary of this process:

```
> transformed <- stepAIC(large_model, direction = "both")
Start:  AIC=1040.98
treatment ~ Age + Gender + Country + family_history + remote_work

                 Df Deviance    AIC
- Country        38   163.41 1017.2
- remote_work     1   153.30 1039.2
<none>                153.26 1041.0
- Age             1   153.83 1042.0
- Gender          2   156.92 1056.2
- family_history  1   172.53 1135.5

Step:  AIC=1017.23
treatment ~ Age + Gender + family_history + remote_work

                 Df Deviance    AIC
- Age             1   163.42 1015.3
- remote_work     1   163.57 1016.0
<none>                163.41 1017.2
- Gender          2   167.21 1032.0
+ Country        38   153.26 1041.0
- family_history  1   183.70 1110.6

Step:  AIC=1015.28
treatment ~ Gender + family_history + remote_work
```

```
Step:  AIC=1015.28
treatment ~ Gender + family_history + remote_work

                 Df Deviance    AIC
- remote_work     1   163.58 1014.1
<none>                163.42 1015.3
+ Age             1   163.41 1017.2
- Gender          2   167.24 1030.1
+ Country        38   153.83 1042.0
- family_history  1   183.74 1108.8

Step:  AIC=1014.07
treatment ~ Gender + family_history

                 Df Deviance    AIC
<none>                163.58 1014.1
+ remote_work     1   163.42 1015.3
+ Age             1   163.57 1016.0
- Gender          2   167.40 1028.9
+ Country        38   153.91 1040.4
- family_history  1   183.95 1107.7
```

```
> summary(transformed)

Call:
glm(formula = treatment ~ Gender + family_history, data = mh_valid_tech)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9410  -0.4484   0.2294   0.3812   0.5516

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.59099    0.11442   5.165 3.03e-07 ***
GenderF             0.02777    0.11783   0.236    0.814
GenderM            -0.14261    0.11405  -1.250    0.212
family_historyTRUE  0.32220    0.03206  10.050  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2017017)

    Null deviance: 190.86  on 814  degrees of freedom
Residual deviance: 163.58  on 811  degrees of freedom
AIC: 1014.1

Number of Fisher Scoring iterations: 2
```

After we performed this model selection, we looked at the summary of the model, shown to the left. We can see that our AIC score reduced to 1014, indicating that our model selection was successful. The variable selection removed age, country, and remote work as significant predictors of mental illness within an individual.

Variance Inflation Factor and Influence Analysis

Variance Inflation Factor

We also took a look at the variance inflation factor (VIF) within our best fit model in order to determine whether multicollinearity existed within our model. The result of our analysis is shown below:

```
> vif(transformed)
                   GVIF Df GVIF^(1/(2*Df))
Gender          1.02226  2        1.005519
family_history 1.02226  1        1.011069
```

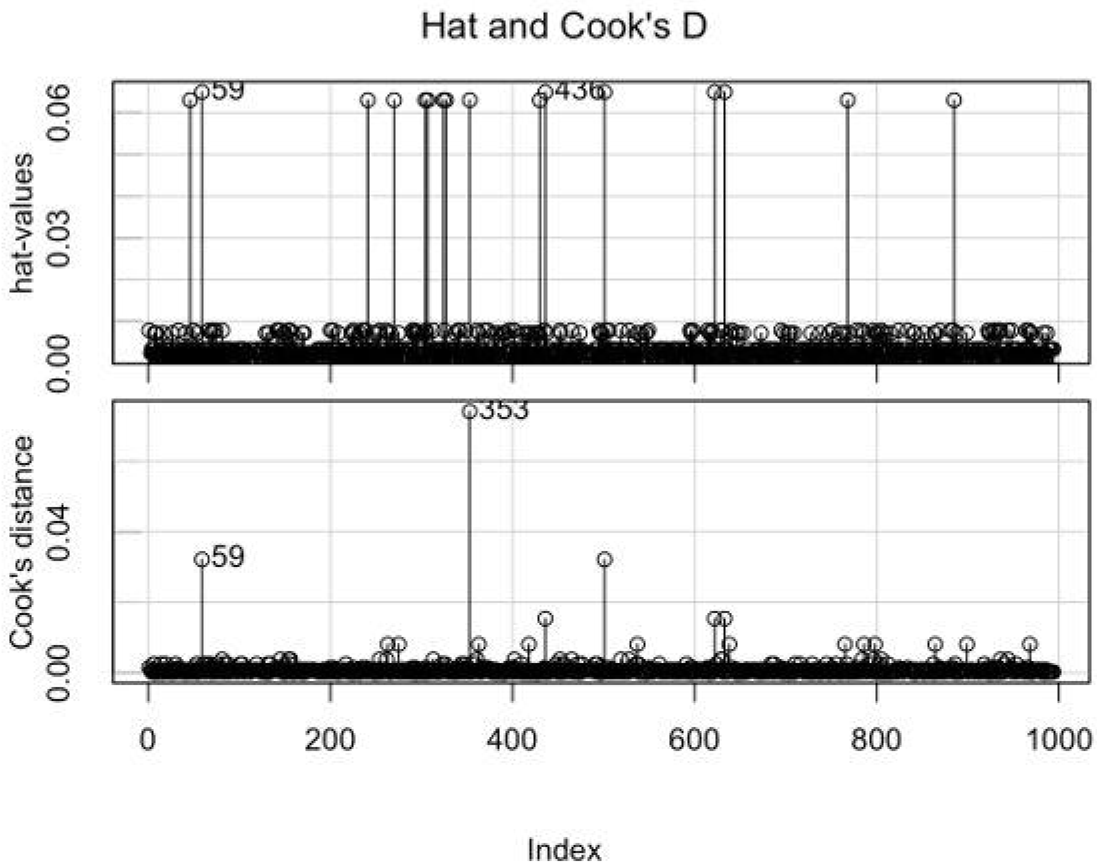
Since neither of the values are above 10, we can safely conclude that multicollinearity does not exist within our model.

Influence Analysis

We performed influence analysis of our best fit model in order to determine if there were any outliers within the data. As shown below, we identified 16 potentially influential data points.

```
> summary(influence.measures(transformed))
Potentially influential observations of
	 glm(formula = treatment ~ Gender + family_history, data = mh_valid_tech) :

    dfb.1_ dfb.GndF dfb.GndM dfb.f_TR dffit   cov.r   cook.d hat    
46   0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
59  -0.36   0.34     0.35     0.07    -0.36_*  1.06_*  0.03   0.06_*
241  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
270  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
304  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
307  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
324  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
327  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
353 -0.52   0.52     0.53    -0.05    -0.55_*  1.05_*  0.07   0.06_*
430  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
436  0.25  -0.23    -0.24    -0.05     0.25_*  1.07_*  0.02   0.06_*
501 -0.36   0.34     0.35     0.07    -0.36_*  1.06_*  0.03   0.06_*
622  0.25  -0.23    -0.24    -0.05     0.25_*  1.07_*  0.02   0.06_*
633  0.25  -0.23    -0.24    -0.05     0.25_*  1.07_*  0.02   0.06_*
768  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
885  0.05  -0.05    -0.05     0.00     0.05    1.07_*  0.00   0.06_*
```

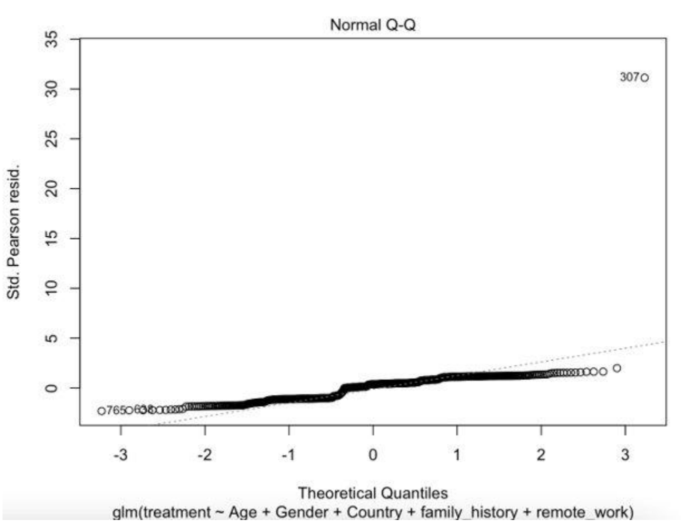
Shown below is also a graph of the hat values and Cook's D values illustrating the outliers identified in our model.



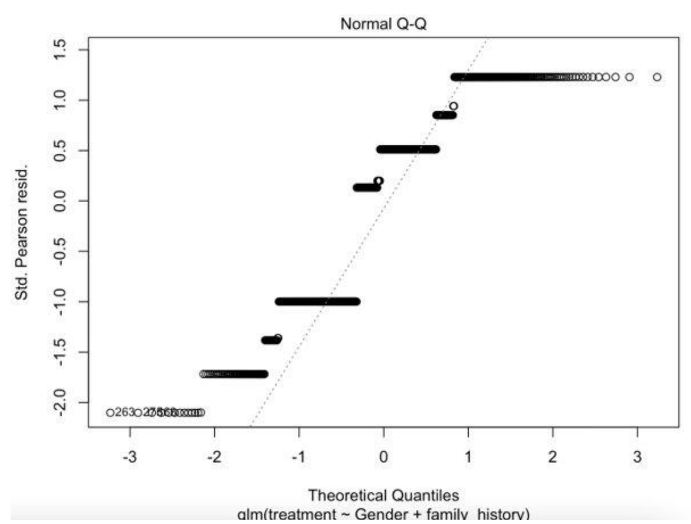
QQ Plots

We finally also analyzed the QQ plots, or quantile-quantile plots, in order to compare the distribution of residuals before and after finding the best fit model. The plots are shown below:

Before Model Transformation



After Model Transformation

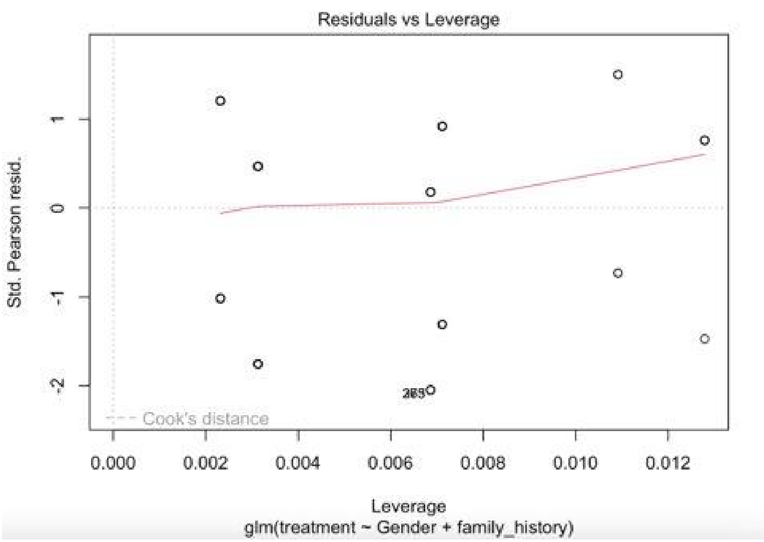


We can see that in the QQ plot before performing selection that our distribution of residuals mostly follows a normal distribution, since the residuals generally follow the line shown within the graph. However, upon analysis of the QQ plot after performing model selection, we can see that the residual distribution violates the normal assumption, since it doesn't follow the line shown in the plot.

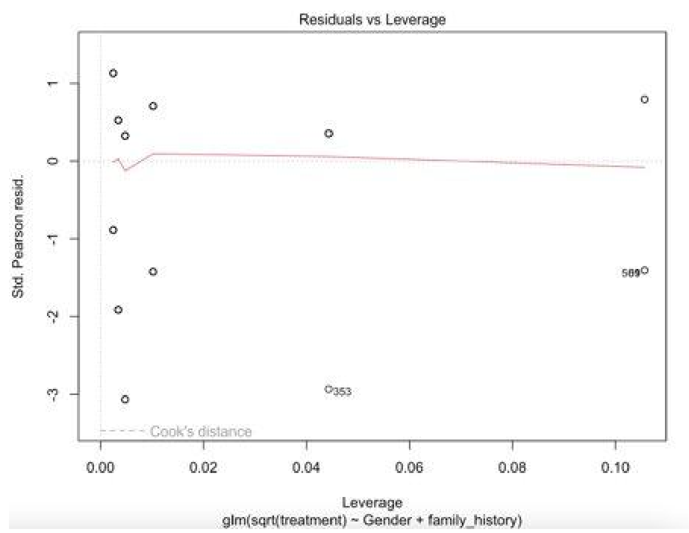
Model Transformation

Due to the results of our influence analysis as well as our QQ plot analysis, we decided to transform the model to see if it would have any effect on the distribution of our residuals. We chose to use a square root transformation, or taking the square root of our response variable. The results of this transformation are shown below:

Before Model Transformation

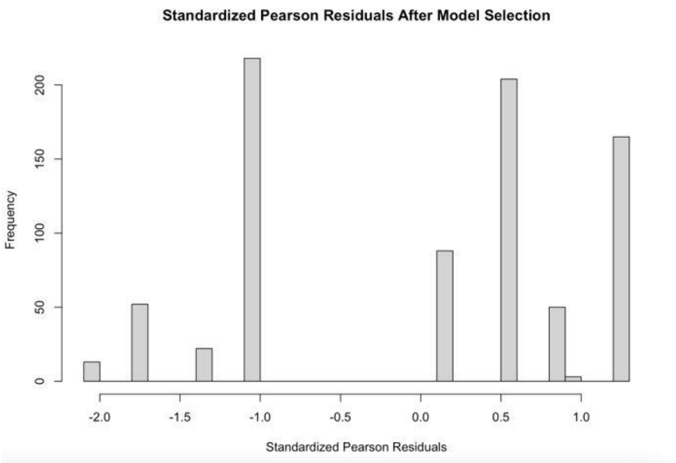


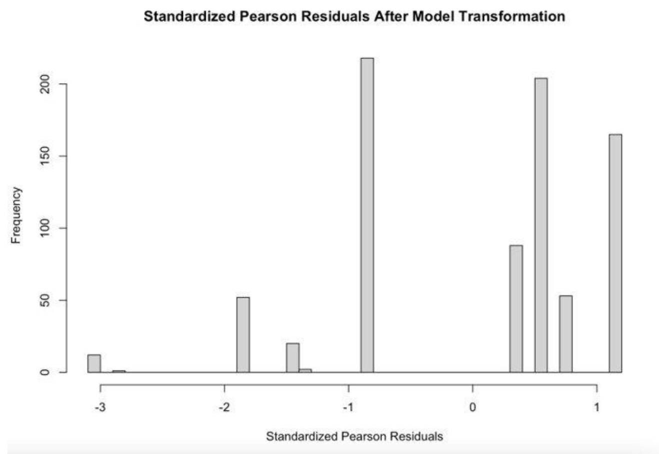
After Model Transformation



We see that before transforming our model, the data points are evenly scattered throughout the residuals vs leverage plot. This scattering indicates that the model is a good fit for the data. After transforming our model, however, we see that the data points skew towards the left and are not evenly scattered throughout the residuals vs leverage plot. This indicates that the model transformation did not provide a better fit for our data.

The distribution of residuals can also be shown as below:

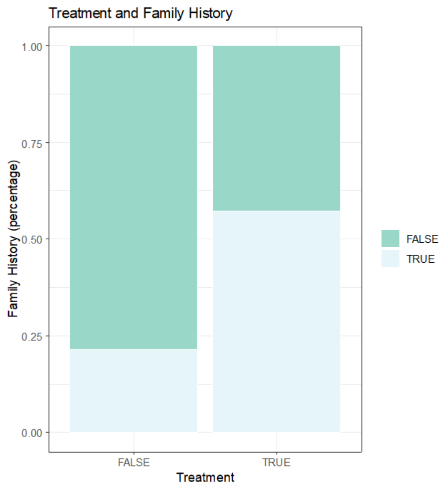




In order to create this histogram of residuals within the models, we used the Pearson method. We can see that initially, the distribution of residuals of our best fit model don't follow a normal distribution, and after performing the transformation, this distribution skews farther to the right. This also indicates that the model transformation did not provide a better fit for our data. Overall, we can conclude that the best fit model we found using the stepwise AIC variable selection process provided the best model for the data.

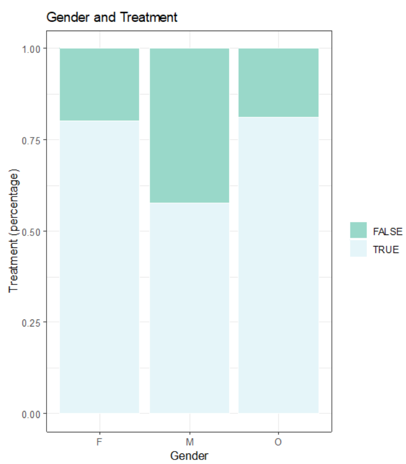
Further Analysis

In this section, we analyze variables that we initially removed for our predictive model as well as further analyze variables initially considered for our model. We look at perceived interference with work among genders, the distribution of genders that identified as having a mental health condition, and prevalence of family history of a mental health condition within individuals.



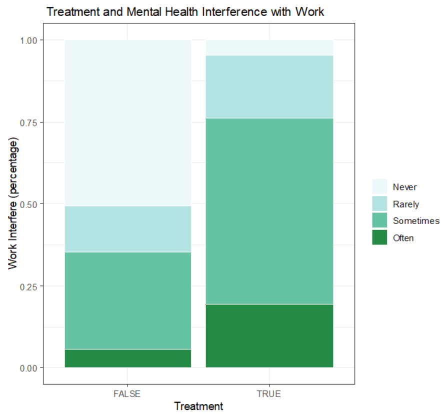

```
Call:
glm(formula = treatment ~ family_history, family = binomial,
    data = mental_health_transf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.846   -1.137   0.634    1.218    1.218

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.09614    0.09356  -1.028     0.304
family_historyTRUE  1.59851    0.16602   9.628    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1077.71  on 814  degrees of freedom
Residual deviance:  972.67  on 813  degrees of freedom
AIC: 976.67

Number of Fisher Scoring iterations: 4
```

From our models looking at how family history influences mental health, we saw a positive correlation between family history being true and the presence of a mental health disorder.




```
Call:
glm(formula = treatment ~ Gender, family = binomial, data = mental_health_transf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8297  -1.3126   0.6681   1.0480   1.0480

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.38629    0.19764   7.014 2.31e-12 ***
GenderM     -1.07392    0.21325  -5.036 4.76e-07 ***
GenderO      0.08004    0.67031   0.119    0.905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1077.7  on 814  degrees of freedom
Residual deviance: 1046.0  on 812  degrees of freedom
AIC: 1052

Number of Fisher Scoring iterations: 4
```

From our models looking at how gender influences mental health, we saw a positive correlation between being female/ other and the presence of a mental health disorder. About 60% of males still had some kind of mental health disorder present, but both females and other genders sat at about 80%.




```
Call:
glm(formula = treatment ~ work_interfere, family = binomial,
    data = mental_health)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9623  -0.5510   0.7232   0.7232   1.9800

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.8083     0.1970   -9.18   <2e-16 ***
work_interfereRarely      2.6805     0.2581   10.39   <2e-16 ***
work_interfereSometimes   3.0160     0.2257   13.36   <2e-16 ***
work_interfereOften       3.5760     0.3075   11.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1304.6  on 994  degrees of freedom
Residual deviance: 1004.3  on 991  degrees of freedom
AIC: 1012.3

Number of Fisher Scoring iterations: 4
```

While work interference wasn’t ultimately one of our predictors, we did still feel it was important to look into. Most prominently here we can conclude there is a high association between needing treatment and mental health interfering with work. We can also see that almost all employees who have a mental health disorder feel that it in some way interferes with work.

Reflection

This project allowed us to put all of our skills we learned this semester into use. We got to take a deep dive into what to look for in our mental health and our futures in the tech industry. Our analysis posed a lot of road blocks when it came to how we wanted to proceed with our categorical data. Ideally we would have been able to look at a wider variety of information, take a look at if there was a correlation between mental health and hours worked, vacation days taken, etc.

Appendix

Roles

<u>Mental Health in Tech</u>
Amulya: Performed model fitting, variable selection, and transformation of the mental health in tech dataset

Ally: Cleaned data and created informational graphs for the health and tech dataset

<u>Mental Health in University</u>
Ann: Cleaned the student mental health dataset, analyzed whether getting treatment affected mental health variables such as depression, anxiety, and panic attacks in students majoring in technology-related majors, fit a stepAIC model to the data.

Cindy: Cleaned the dataset on student mental health, performed transformations to group GPAs into low, median, and high categories, and conducted an analysis to determine if high GPA levels have an impact on mental health issues, including depression, anxiety, and panic attacks, and conducted variable selection on the dataset.

Sources

Islam, MD Shariful. "Student Mental Health." Kaggle, 17 Feb. 2023, www.kaggle.com/datasets/shariful07/student-mental-health.

Open Sourcing Mental Illness, LTD. "Mental Health in Tech Survey." Kaggle, 3 Nov. 2016, www.kaggle.com/datasets/osmi/mental-health-in-tech-survey.

R Code

<u>Mental Health in Academia:</u>

A - Defining Libraries:

```
library(dplyr)
```

```
library(MASS)
student <- read.csv("student.csv")
```

B - Cleaning our Data:

```
students <- student[, c(-1, -3)]
students <- students[, -c(3, 5)]
students <- students[which(student$What.is.your.course. == "Engineering" |
                     student$What.is.your.course. == "BIT" |
                     student$What.is.your.course. == "BCS" |
                     student$What.is.your.course. == "Biotechnology"), ]
#Change Column names
colnames(students)[1] = "gender"
colnames(students)[2] = "major"
colnames(students)[3] = "year"
colnames(students)[4] = "gpa"
colnames(students)[6] = "depression"
colnames(students)[7] = "anxiety"
colnames(students)[8] = "panic_attack"
colnames(students)[9] = "treatment"

#Transform yes/no to binary data
students$depression <- ifelse(students$depression == "Yes", 1, 0)
students$anxiety <- ifelse(students$anxiety == "Yes", 1, 0)
students$panic_attack <- ifelse(students$panic_attack == "Yes", 1, 0)
students$treatment <- ifelse(students$treatment == "Yes", 1, 0)

# Categorize GPA to 1 to 5 levels
for (i in seq_along(students$gpa)) {
  gpa_str <- students$gpa[i]
  if (gpa_str == "0 - 1.99") {
    students$gpa[i] <- 1
  } else if (gpa_str == "2.00 - 2.49") {
    students$gpa[i] <- 2
  } else if (gpa_str == "2.50 - 2.99") {
    students$gpa[i] <- 3
  } else if (gpa_str == "3.00 - 3.49") {
    students$gpa[i] <- 4
  } else if (gpa_str == "3.50 - 4.00") {
    students$gpa[i] <- 5
  }
}

# Categorize GPA to low, median, and high and aggregate with total issues
students$gpa_group <- ifelse(students$gpa < 3, "low", ifelse(students$gpa > 3,
"high", "medium"))
gpa_issues <- aggregate(cbind(depression, anxiety, panic_attack) ~ gpa_group, data =
students, FUN = sum)
gpa_issues$gpa_group <- factor(gpa_issues$gpa_group, levels = c("low", "medium",
"high"))
gpa_issues<- arrange(gpa_issues, gpa_group)
gpa_issues$total_issues <- gpa_issues$depression + gpa_issues$anxiety +
gpa_issues$panic_attack
```

C - Performing Model Analysis

```
#Model Fitting
model <- glm(total_issues ~ gpa_group, data = gpa_issues, family = poisson())
summary(model)

#Residual Anaysis
residuals_deviance <- residuals(model, type = "deviance")
plot(fitted(model), residuals_deviance, main = "Deviance Residuals vs Fitted Values",
xlab = "Fitted Values", ylab = "Deviance Residuals")
abline(0, 0, col = "red")

#Variable stepwise
stepwise_model <- stepAIC(model, direction = "both")
stepwise_model
summary(stepwise_model)

#Conclude the results
exp(coef(stepwise_model))
```

D - Plots of our Further Analysis:

```
sapply(clean.students[, 4:7], sd)
xtabs(~treatment + depression + panic.attack + anxiety, data = clean.students)

model <- glm(treatment ~ depression + panic.attack + anxiety, data = clean.students,
             family = "binomial")
summary(model)

# residual analysis
r.dev <- residuals(model, type = "deviance")
fm <- fitted(model)

o.plot <- plot(fm, r.dev, col = c("blue", "red"), pch = c(22, 17),
               main = "Deviant Residual vs Fitted Residual Values",
     xlab = "Fitted", ylab = "Deviant")
legend("topleft", pch = c(22, 17), c("Fitted", "Deviant"), col = c("blue", "red"))
abline(lm(fm ~ r.dev), col = "black")

# transforming the data
r.dev <- log10(r.dev)

new.plot <- plot(fm, r.dev, col = c("blue", "red"), pch = c(22, 17),
                 main = "Deviant Residual vs Fitted Residual Values",
                 xlab = "Fitted", ylab = "Deviant")
legend("topleft", pch = c(22, 17), c("Fitted", "Deviant"), col = c("blue", "red"))
abline(lm(fm ~ r.dev), col = "black")

# model selection
step.mod <- stepAIC(model, direction = "both")
summary(step.mod)
```

Mental Health in Tech:

A - Defining Libraries:

```
library(readr)
library(dplyr)
library(tidyverse)
library(aod)
library(tibble)
library(MASS)
library(ggplot2)
library(reshape2)
library(grid)
library(gridExtra)
library(car)
library(cowplot)
library(reshape2)
mental_health <- read_csv("mental_health_in_tech.csv")
#View(mental_health)
```

B - Cleaning our Data:

```
#CLEANING _______________________________________________________________________
#removing unused data
mental_health <- subset(mental_health, select = -c(Timestamp, state, self_employed,
                                                   no_employees, comments))

#omit NA's
mental_health <- na.omit(mental_health)


#dim(mental_health)

#GENDER ________________________________________________________________________
#Male
index_M <- which(mental_health$Gender == "Male")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "male")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "m")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Cis Man")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "maile")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Cis Male")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Mal")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Male (CIS)")
```

```
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Make")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Man")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "msle")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Mail")
mental_health$Gender[index_M] <- "M"

index_M <- which(mental_health$Gender == "cis male")
mental_health$Gender[index_M] <- "M"
index_M <- which(mental_health$Gender == "Malr")
mental_health$Gender[index_M] <- "M"


#Female
index_F <- which(mental_health$Gender == "Female")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "female")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "f")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "Cis Female")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "Woman")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "woman")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "Femake")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "cis-female/femme")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "Female (cis)")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "femail")
mental_health$Gender[index_F] <- "F"
index_F <- which(mental_health$Gender == "Woman")
mental_health$Gender[index_F] <- "F"


#Other
index_O <- which(mental_health$Gender == "Male-ish")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "something kinda male?")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Trans-female")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "queer/she/they")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "non-binary")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Nah")
```

```
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "All")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Enby")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "fluid")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Genderqueer")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Androgyne")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Agender")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "male leaning androgynous")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Guy (-ish) ^_^")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Female (trans)")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Neuter")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "queer")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "A little about you")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "p")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "ostensibly male, unsure what that really
means")
mental_health$Gender[index_O] <- "O"
index_O <- which(mental_health$Gender == "Trans woman")
mental_health$Gender[index_O] <- "O"

mental_health$Gender <- factor(mental_health$Gender)




#FAMILY HISTORY________________________________________________________________

index_fh_true <- which(mental_health$family_history =="Yes")
mental_health$family_history[index_fh_true] <- TRUE

index_fh_false <- which(mental_health$family_history =="No")
mental_health$family_history[index_fh_false] <- FALSE

mental_health$family_history <- as.logical(mental_health$family_history)


#TREATMENT_____________________________________________________________________
index_treat_true <- which(mental_health$treatment =="Yes")
mental_health$treatment[index_treat_true] <- TRUE
```

```r
index_treat_false <- which(mental_health$treatment =="No")
mental_health$treatment[index_treat_false] <- FALSE

mental_health$treatment <- as.logical(mental_health$treatment)



#WORK_INTERFERE_________________________________________________________________

mental_health$work_interfere <- factor(mental_health$work_interfere,
                                       levels = c("Never", "Rarely", "Sometimes",
                                                  "Often"))


#REMOTE WORK____________________________________________________________________
rm_false <- which(mental_health$remote_work == "No")
mental_health$remote_work[rm_false] <- FALSE
rm_true <- which(mental_health$remote_work == "Yes")
mental_health$remote_work[rm_true] <- TRUE

mental_health$remote_work <- as.logical(mental_health$remote_work)

#TECH COMPANY___________________________________________________________________
tc_false <- which(mental_health$tech_company == "No")
mental_health$tech_company[tc_false] <- FALSE
tc_true <- which(mental_health$tech_company == "Yes")
mental_health$tech_company[tc_true] <- TRUE

mental_health$tech_company <- as.logical(mental_health$tech_company)


#BENEFITS_______________________________________________________________________
mental_health$benefits <- factor(mental_health$benefits,
                                 levels = c("No", "Don't know", "Yes"))

#CARE_OPTIONS___________________________________________________________________
index_NS <- which(mental_health$care_options == "Not sure")
mental_health$care_options[index_NS] <- "Don't know"


mental_health$care_options <- factor(mental_health$care_options,
                                     levels = c("No", "Don't know", "Yes"))


#WELLNESS_PROGRAM_______________________________________________________________
mental_health$wellness_program <- factor(mental_health$wellness_program,
                                         levels = c("No", "Don't know", "Yes"))


#SEEK_HELP______________________________________________________________________
mental_health$seek_help <- factor(mental_health$seek_help,
                                  levels = c("No", "Don't know", "Yes"))
```

```
#ANONIMITY________________________________________________________________________
mental_health$anonymity <- factor(mental_health$anonymity,
                                  levels = c("No", "Don't know", "Yes"))

#LEAVE____________________________________________________________________________
mental_health$leave <- factor(mental_health$leave,
                              levels = c("Don't know", "Very difficult", "Somewhat
difficult",
                                         "Somewhat easy", "Very easy"))

#MENTAL_HEALTH_DATA_______________________________________________________________
mental_health$mental_health_consequence <-
factor(mental_health$mental_health_consequence,
                                                  levels = c("No", "Maybe", "Yes"))

#PHYS_HEALTH_CONS_______________________________________________________________
mental_health$phys_health_consequence <- factor(mental_health$phys_health_consequence,
                                                levels = c("No", "Maybe", "Yes"))

#COWORKERS_____________________________________________________________
mental_health$coworkers <- factor(mental_health$coworkers,
                                  levels = c("No", "Some of them", "Yes"))

#COWORKERS_____________________________________________________________
mental_health$supervisor <- factor(mental_health$supervisor,
                                   levels = c("No", "Some of them", "Yes"))


#MENTAL_HEALTH_INTERVIEW_____________________________________________________________
mental_health$mental_health_interview <- factor(mental_health$mental_health_interview,
                                                levels = c("No", "Maybe", "Yes"))

#PHYS_HEALTH_INTERVIEW_____________________________________________________________
mental_health$phys_health_interview <- factor(mental_health$phys_health_interview,
                                              levels = c("No", "Maybe", "Yes"))

#MENTAL_HEALTH_INTERVIEW_____________________________________________________________
mental_health$mental_vs_physical<- factor(mental_health$mental_vs_physical,
                                          levels = c("No", "Don't know", "Yes"))

#OBSERVED NEGATIVE CONSEQUENCES
unique(mental_health$obs_consequence)
oc_false <- which(mental_health$obs_consequence == "No")
mental_health$obs_consequence[oc_false] <- FALSE
oc_true <- which(mental_health$obs_consequence == "Yes")
mental_health$obs_consequence[oc_true] <- TRUE
mental_health$obs_consequence <- as.logical(mental_health$obs_consequence)
```

C - Performing Model Analysis

```
#VARIABLE SELECTION ------------------------------------------------------------
#new data frame with only have valid predictors of mental health
mental_health_valid <- cbind(mental_health[1:5], mental_health[7:8])
mh_valid_tech <- subset(mental_health_valid,
                        mental_health_valid$tech_company == TRUE)
#creating logistic regression model
large_model <- glm(treatment ~ Age + Gender + Country + family_history +
                     remote_work, data = mh_valid_tech)
summary(large_model)
# Using stepwise variable selection for model
transformed <- stepAIC(large_model, direction = "both")
summary(transformed)
transformed <- glm(formula = treatment ~ Gender + family_history,
                   data = mh_valid_tech)

#visualizing probabilities using the transformed model
#VARIABLE SELECTION ------------------------------------------------------------

#VIF FOR MULTICOLLINEARITY -----------------------------------------------------
vif(large_model)
vif(transformed)
#VIF FOR MULTICOLLINEARITY -----------------------------------------------------

#INFLUENCE ANALYSIS ------------------------------------------------------------
summary(influence.measures(large_model))
summary(influence.measures(transformed))
infIndexPlot(transformed, vars = c("hat"), main = "Hat")
#INFLUENCE ANALYSIS ------------------------------------------------------------

#RESIDUAL ANALYSIS -------------------------------------------------------------
#on largest model
large_model_resid <- resid(large_model, type = "pearson")
lmr_standard <- large_model_resid / sd(large_model_resid)
hist(lmr_standard, main = "Standardized Pearson Residuals before Model Selection",
     xlab = "Standardized Pearson Residuals", breaks = 30)
#on best fit model
transformed_resid <- resid(transformed, type = "pearson")
trans_standard <- transformed_resid / sd(transformed_resid)
hist(trans_standard, main = "Standardized Pearson Residuals After Model Selection",
     xlab = "Standardized Pearson Residuals", breaks = 30)
#RESIDUAL ANALYSIS -------------------------------------------------------------

#to get QQ plots before and after model selection
plot(large_model)
plot(transformed)

#encoding appropriate data to perform transformations
mhvt2 <- mh_valid_tech
mhvt2$Gender <-as.numeric(mhvt2$Gender)
mhvt2$family_history <- as.numeric(mhvt2$family_history)
mhvt2$treatment <- as.numeric(mhvt2$treatment)
mhvt2$tech_company <- as.numeric(mhvt2$tech_company)
mhvt2$remote_work <- as.numeric(mhvt2$remote_work)
```

```
#apply square root transformation to best model
trans <- glm(formula = sqrt(treatment) ~ Gender + family_history,
                   data = mhvt2,
                   family = binomial(link = "logit"))
plot(trans)
trans_resid_clean <- resid(trans, type = "pearson")
trans_standard_clean <- trans_resid_clean / sd(trans_resid_clean)
hist(trans_standard_clean, main = "Standardized Pearson Residuals After Model
Transformation",
     xlab = "Standardized Pearson Residuals", breaks = 30)
```

D - Plots of our Further Analysis:

```
#VALIDATION ____________________________________________________________________
{
  mental_health_valid <- cbind(mental_health[1:5], mental_health[7:8])
  mental_health_transf <- subset(mental_health_valid,
                                 mental_health_valid$tech_company == TRUE)
}

#AGE AS FACTOR__________________________________________________________________
{mental_health_transf <- cbind(mental_health_transf, mental_health_transf$Age)

  colnames(mental_health_transf)[8] <- "Age_fact"

  index_10 <- which(mental_health_transf$Age_fact < 20)
  mental_health_transf$Age_fact[index_10] <- "Under 20"

  index_20 <- which(mental_health_transf$Age_fact < 30 & mental_health_transf$Age_fact
>= 20)
  mental_health_transf$Age_fact[index_20] <- "20's"

  index_30 <- which(mental_health_transf$Age_fact < 40  & mental_health_transf$Age_fact
>= 30)
  mental_health_transf$Age_fact[index_30] <- "30's"

  index_40 <- which(mental_health_transf$Age_fact < 50 & mental_health_transf$Age_fact
>= 40)
  mental_health_transf$Age_fact[index_40] <- "40's"

  index_50 <- which(mental_health_transf$Age_fact < 60 & mental_health_transf$Age_fact
>= 50)
  mental_health_transf$Age_fact[index_50] <- "50's"

  index_60 <- which(mental_health_transf$Age_fact < 70 & mental_health_transf$Age_fact
>= 60)
  mental_health_transf$Age_fact[index_60] <- "60's"

  index_70 <- which(mental_health_transf$Age >= 70)
  mental_health_transf$Age_fact[index_70] <- "Over 70"
```

```
  mental_health_transf$Age_fact <- as.factor(mental_health_transf$Age_fact)

  levels(mental_health_transf$Age_fact) <- c("Under 20", "20's", "30's", "40's",
"50's", "60's", "Over 70")

}

#VARIABLE COUNT PLOTS___________________________________________________________
{
  wi <- ggplot(data = mental_health) +
    geom_bar(mapping = aes(x = work_interfere) , fill = "lightcyan2", color =
"darkslategray3") + labs(x = "Interferes with Work (from original data)") +
    geom_text(mapping = aes(x = work_interfere, label=..count..), stat = 'count')


  fh <- ggplot(data = mental_health_transf) +
    geom_bar(mapping = aes(x = family_history) , fill = "lightcyan2", color =
"darkslategray3") + labs(x = "Family History") +
    geom_text(mapping = aes(x = family_history, label=..count..), stat = 'count')

  gender <- ggplot(data = mental_health_transf) +
    geom_bar(mapping = aes(x = Gender) , fill = "lightcyan2",  color = "darkslategray3"
) + labs(x = "Gender") +
    geom_text(mapping = aes(x = Gender, label=..count..), stat = 'count')

  rw <- ggplot(data = mental_health_transf) +
    geom_bar(mapping = aes(x = remote_work) , fill = "lightcyan2",  color =
"darkslategray3" ) + labs(x = "Remote Work") +
    geom_text(mapping = aes(x = remote_work, label=..count..), stat = 'count')

  age <- ggplot(data = mental_health_transf) +
    geom_bar(mapping = aes(x = as.factor(Age_fact)) , fill = "lightcyan2",  color =
"darkslategray3" ) + labs(x = "Age") +
    geom_text(mapping = aes(x = as.factor(Age_fact), label=..count..), stat = 'count')


  plot_grid(rw, fh, gender, wi, age, nrow = 3)

}

#WORK INTERFERE PLOTS___________________________________________________________
{
  lm_workInfft = glm(treatment~work_interfere, data = mental_health,
                     family = poisson())
  summary(lm_workInfft)

  ggplot(data = mental_health) +
    geom_bar(mapping = aes(x = treatment,  fill = work_interfere),
             position = "fill", color = "white") +
    labs(title = " Treatment and Mental Health Interference with Work",
         x = "Treatment",
         y = "Work Interfere (percentage)",
```

```
        fill = "") +
    scale_fill_brewer(palette = "BuGn", direction = 1) + theme_bw()
}

#FAMILY HISTORY PLOTS___________________________________________________________
{
  lm_fam = glm(treatment~family_history, data = mental_health_transf,
               family = binomial)
  summary(lm_fam)

  ggplot(data = mental_health_transf) +
    geom_bar(mapping = aes(x = treatment,  fill = family_history),
             position = "fill", color = 'white') +
    labs(title = "Treatment and Family History ",
         x = "Treatment",
         y = "Family History (percentage)",
         fill = "") +
    scale_fill_brewer(palette = "BuGn", direction = -1) + theme_bw()
}

#GENDER PLOTS___________________________________________________________________
{
  lm_gen = glm(treatment~family_history, data = mental_health_transf,
               family = binomial)
  summary(lm_gen)

  ggplot(data = mental_health) +
    geom_bar(mapping = aes(x = Gender,  fill = treatment),
             position = "fill", color = 'white') +
    labs(title = "Mental Health by Gender",
         x = "Treatment",
         y = "Gender (percentage)",
         fill = "") +
    scale_fill_brewer(palette = "BuGn", direction = -1) + theme_bw()
}
```