Assessing comfort level of employees in discussing mental health issues within the Tech Industry

INFX 573 – Data Science I: Theoretical Foundations

Winter 2018

Team Members:

Harkar Talwar, Manisha Vyas, Neha Palsokar, Sayali Chakradeo, Shreya Sabharwal

Abstract

Mental health is often considered a social stigma that restricts people suffering from such disorders to openly discuss these issues with their supervisors at workplace. Open Source Mental Illness, a non-profit organization conducted a mental health survey in 2016 to raise awareness and support mental health wellness. This research is an attempt to analyze comfort level of employees in discussing mental health disorders with their supervisors in tech industry based on the survey questions answered by the respondents.

Based on preliminary analysis, few key features were identified which linked to comfort level evaluation. It led to formulation of four research question and specific hypothesis for testing correlation of age, organization size, protection of anonymity, mental health coverage and awareness of mental health benefits amongst the employees. Exploratory data analysis was performed to understand these correlations by plotting specific features and determining statistical significance using multinomial regression model.

Results from exploratory data analysis and backward elimination method were utilized to identify two models (full and reduced). Multinomial logistic regression and random forest prediction algorithms followed by k-fold cross validation method was applied to determine accuracy of these models. Though the prediction accuracy for both the algorithms was low, multinomial logistic regression on reduced model fitted the data best.

As per the findings, recommendations include developing firm policies which are age agnostic and focus on educating supervisors and coworkers for handling employees with mental health issues, in an effective manner. They should include terms on anonymity protection and information regarding mental health benefits provided by the employers.

Table of Contents

1. Introduction

The statistics regarding mental health issues at workplace are alarming. As reported by US Department of Health and Human Services, 1 in 5 adults in the US are living with mental health illness (Kasbergen, 2017). Approximately 84% work through lunch hours and 33% work a 50+ hours a week and yet only 11% of employees discuss a recent mental health problem with their line managers (*Mental health in the workspace Infographic*, 2017).

The stigma attached to having a psychiatric disorder is such that employees may be reluctant to seek treatment, especially in the current economic climate out of fear of negative consequences, change in perceptions of co-workers and jeopardizing their jobs. Additionally, managers may want to help but aren't sure how to do so. As a result, mental health disorders often go unrecognized and untreated not only damaging an individual's health, but also reducing productivity at work. Discussion and adequate treatment can alleviate such issues and improve employee job performance.

Open Sourcing Mental Illness (OSMI) is a non-profit dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech and open source communities. They create opportunities for public, open dialogue about mental health and provide people the opportunity to see they are not alone in dealing with mental health challenges. We are using the survey data available from OSMI for our research.

The broader research problem in our study focuses on identifying comfort level of employees when discussing mental health issues with their employers based on certain factors such as anonymity protection, organization size, age and benefits coverage. The analysis would be used in raising awareness and improving conditions for those with mental health disorders in the IT workplace. Here are the set of research question that we aim to explore through our analysis:

1. Is younger generation more likely to discuss mood and anxiety disorders with the employer as compared to the older generation?
2. What percentage of employees reach out for help to the employer, given that the employer is providing benefits to those who are suffering from mental illness?
3. Are employees in larger companies (>1000 employees) more comfortable in discussing mental health issues with employers, compared to those in smaller companies (<=1000 employees)?
4. What percentage of employees reach out for help to the employer, given that the employer is providing benefits to those who are suffering from mental illness?

2. Methods

Open Sourcing Mental Illness (OSMI) conducted a survey in 2016 to gauge perceptions and attitudes of workers in the IT industry towards mental health issues. The survey dataset has 1433 responses and 63 questions and the respondents having voluntarily taken the survey online (OSMI (n.d.).

The following section details out the steps taken to perform feature identification, dataset cleaning, preliminary analysis, exploratory data analysis and the full and reduced model developed through various algorithms (Multinomial regression and Random Forest).

2.1. Identification of relevant features

As part of this step, the data set was analyzed to identify features that were pertinent to the research questions and the broader theme of the study. In total 15 features were identified out of which some of the key features directly related to the research questions are described below, a complete list of these features as well as their description, data type and possible values are listed in (Appendix. I):

Feature Name	Associated Survey Question
age	Age of the Respondent
discuss.supervisor	Would the respondent feel comfortable discussing a mental health disorder with his/her direct supervisor(s)?
organization.size.large	Number of employees in the organization
prev.anonymity.protected	Was anonymity protected if the respondent chose to take advantage of mental health or substance abuse treatment resources with previous employers?
mental.health.coverage	Does your employer provide mental health benefits as part of healthcare coverage?
mental.health.options	Do you know the options for mental health care available under your employer-provided coverage?

Table 1. Specific Features from the dataset

2.2. Cleaning of the dataset for the identified features

Based on identified relevant features, we performed initial analysis (Appendix.II) in R to understand cleaning requirements for each variable as some of responses were in free form text:

- **Updating column headers:** As the features in their raw form had longer names (in the form of survey questions), the column headers were renamed to specific keywords as defined in the Step 1. A vector list with new names was created and was used to change the column names in the data frame.
- **Cleaning the 'age' feature:** To clean the age variable, some unrealistic values for age such as 3, 323 were removed. This was achieved by filtering the data based on realistic lower and upper bounds for age that the team agreed upon. Since, such records formed a very small percentage of the total records from the dataset (less than 0.3%), it would not impact the analysis.
- **Cleaning the 'gender' feature:** The gender variable also had issues such as missing or inappropriate values. The following four broad categories for the gender variable were defined and were assigned to the gender column based on individual responses.
  M: male, F: female; TG: transgender; GQ: genderqueer; Refused: not answered.

2.3. Exploratory data Analysis (EDA)

As part of the exploratory data analysis we analyzed key features individually as well as in conjunction with the response feature (comfort level with supervisor). These analyses have been explained in detail below:

Comfort Level in discussion (about mental health issues) with the supervisor

We identified the 'discuss.supervisor' feature for understanding comfort level of employees for discussing mental health issues with their employers. It will be treated as response feature and the correlation of other features will be identified with respect to this feature.

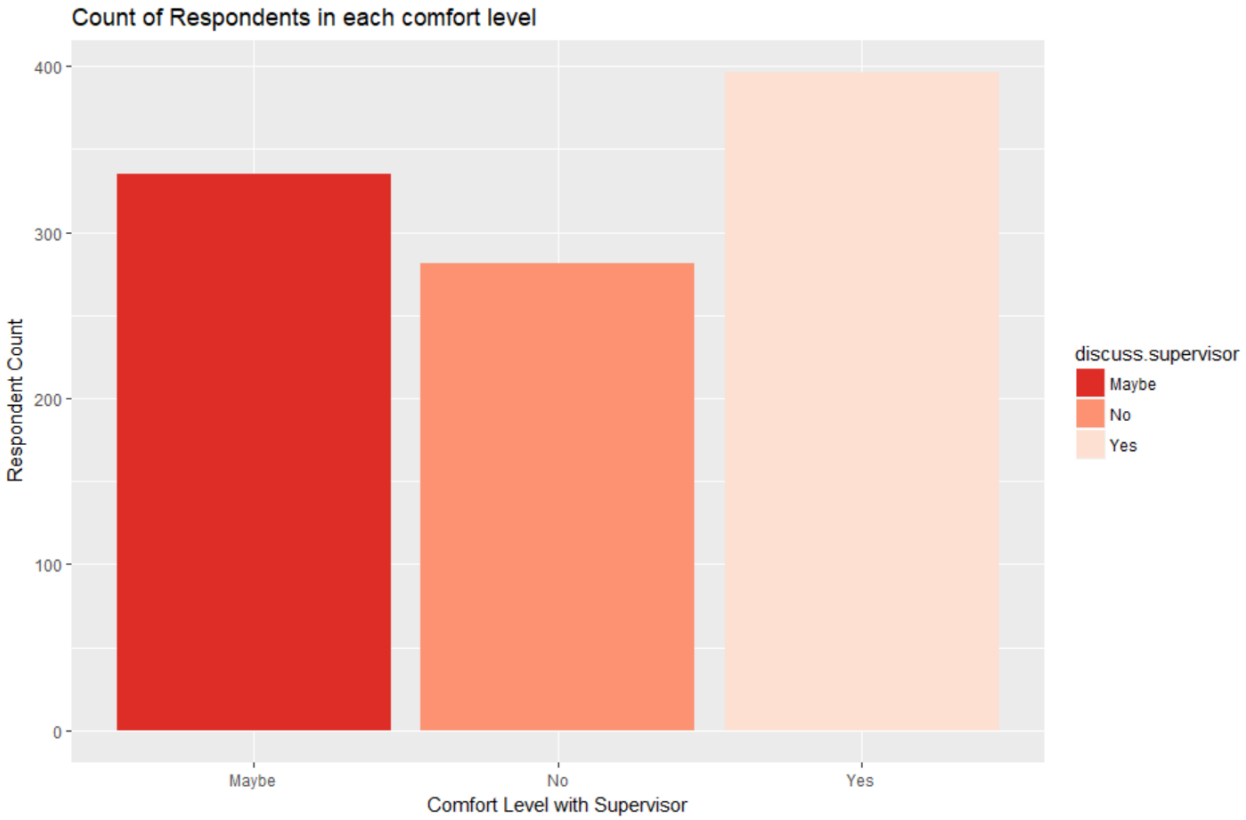


Figure 1. Distribution of the outcome variable

As observed, a fairly good number of respondents were not so sure about whether they feel comfortable discussing these concerns. It showed a good number of respondents who felt uncomfortable as well which outlines the significance of why we need better practices in the tech industry work cultures to increase the comfort level of these employees.

Effect of age on comfort level of the employees

Expectation: Analysis focused on understanding effect of age on comfort level of employees in discussing mental health concerns with their employers. Below is the hypothesis created for analyzing this effect:

(H0): Age Group has no effect on comfort level of employees in discussing mental health issues with employers.
(HA): Age Group influences comfort level of employees in discussing mental health issues with employers.

Visualizations:

Comfort level of the employees with Age Group: The 1012 respondents (after cleaning) were segregated into two categories, namely, 'Young' and 'Senior'. We have 656 respondents in 'Young' category and 356 in 'Senior' Category.

Frequency Distribution plot for Age Groups and Comfort level for discussion with employers: Based on our analysis of the plot, there is no significant correlation between age groups and comfort level.

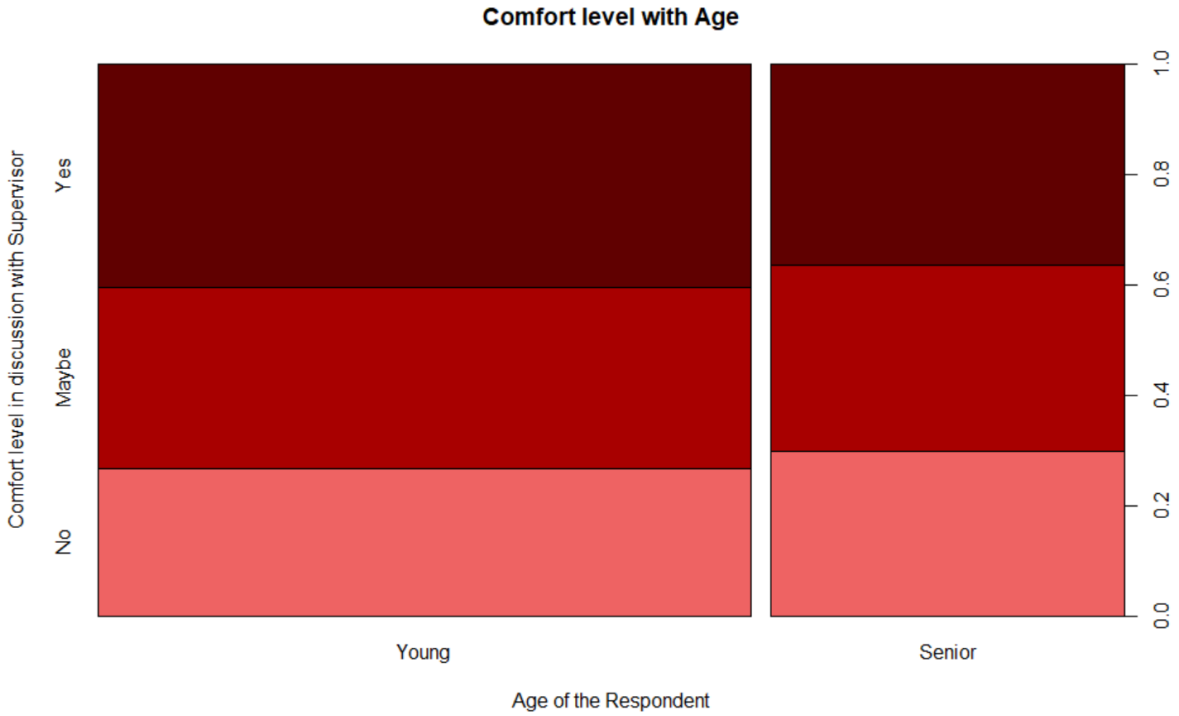


Figure 2. Distribution of comfort with age group

Boxplot for age and comfort level in discussion with employer: To further analyze the correlation between these two features, we plotted them on a box plot and as it can be seen from the plot below, there is no

Figure 3. Box plot for comfort with age

significant variation in comfort level with change in age. For this analysis, we kept the age as continuous variable to ensure that the grouping that we have selected has no effect on how this correlation is interpreted.

Statistical Significance:

Multinomial Logistic Regression: In the multinomial regression, the log odds of feeling comfortable vs not feeling comfortable decreases by 0.214 when we move younger age group to senior age group. Since, the p value was higher (0.188) than the alpha value, the results were not statistically significant and hence, we inferred that age has no effect on the comfort level.

```
> mod <- multinom(discuss.supervisor~age, family="multinom", data=dat)
# weights:  9 (4 variable)
initial  value 1111.795636
final  value 1101.059046
converged
> summary(mod)
Call:
multinom(formula = discuss.supervisor ~ age, data = dat, family = "multinom")

Coefficients:
      (Intercept)   ageSenior
Maybe   0.2058517 -0.08180029
Yes     0.4187112 -0.21461616

Std. Errors:
      (Intercept) ageSenior
Maybe  0.10181087 0.1677282
Yes    0.09733285 0.1630949

Residual Deviance: 2202.118
AIC: 2210.118
> z <- summary(mod)$coefficients/summary(mod)$standard.errors
> z
      (Intercept)  ageSenior
Maybe    2.021903 -0.4876955
Yes      4.301849 -1.3158974
> # 2-tailed z test
>  p <- (1 - pnorm(abs(z), 0, 1)) * 2
> p
       (Intercept) ageSenior
Maybe 0.0431863675 0.6257655
Yes   0.0000169379 0.1882085
```

Figure 4. Summary of comfort versus age single regression model

Effect of benefits coverage on comfort level of the employees

Expectation: Analysis focused on understanding effect of benefits on comfort level of employees. Below is the hypothesis created for analyzing this effect:

(H0): Providing mental health benefits has no correlation on comfort level of employees in discussing mental health issues with their employers.

(HA): Providing mental health benefits has a correlation on comfort level of employees in discussing mental health issues with their employers.

Visualization:

Frequency plot for mental health benefits provided versus comfort level: Higher number of respondents comfortable discussing mental health issues when benefits provided versus when not provided.

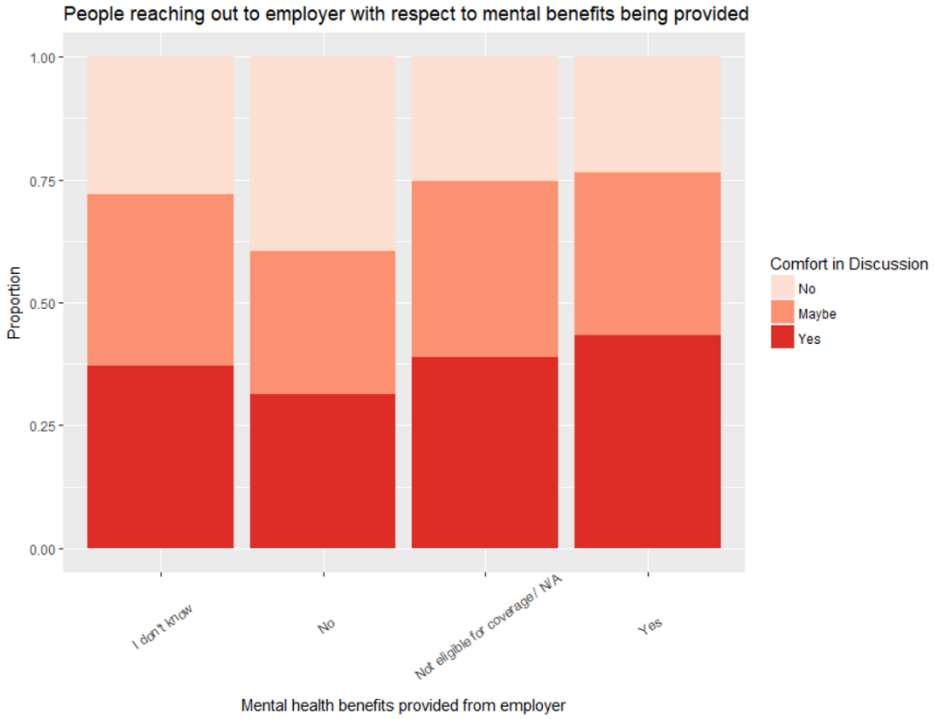


Figure 5. Variation in comfort levels with mental health benefits

Statistical Significance:

Multinomial Logistic Regression: Log odds of feeling comfortable versus not feeling comfortable increase by 0.844 when we move from benefits not provided to benefits provided.

Since, the p value is small (6.5e-05), the results are statistically significant, and we reject the null hypothesis.

```
> summary(mod)
Call:
multinom(formula = discuss.supervisor ~ mental.health.coverage,
    data = company.size)

Coefficients:
      (Intercept) mental.health.coverageI don't know
Maybe  -0.3014720                          0.5271422
Yes    -0.2300183                          0.5145273
      mental.health.coverageNot eligible for coverage / N/A mental.health.coverageYes
Maybe                                          0.6463129                   0.6455692
Yes                                            0.6548999                   0.8442404

Std. Errors:
      (Intercept) mental.health.coverageI don't know
Maybe   0.1794912                          0.2344703
Yes     0.1758979                          0.2304823
      mental.health.coverageNot eligible for coverage / N/A mental.health.coverageYes
Maybe                                          0.3642900                   0.2178871
Yes                                            0.3580854                   0.2114227

Residual Deviance: 2185.561
AIC: 2201.561
> z <- summary(mod)$coefficients/summary(mod)$standard.errors
> p <- (1 - pnorm(abs(z), 0, 1))*2
> p
      (Intercept) mental.health.coverageI don't know
Maybe  0.09303676                         0.02456177
Yes    0.19098141                         0.02558892
      mental.health.coverageNot eligible for coverage / N/A mental.health.coverageYes
Maybe                                         0.07603484                  3.047942e-03
Yes                                           0.06741562                  6.520396e-05
```

Figure 6. Summary of comfort versus benefits single regression model

Effect of company size on comfort level of the employees

Expectation: Analysis focused on understanding effect of company size on comfort level of employees. Below is the hypothesis created for analyzing this effect:

(H0): Company size has no correlation with the comfort level of employees in discussing mental issues with the employers.
(HA): Company size has a correlation with the comfort level of employees in discussing mental issues with the employers.

Visualization:

Frequency Distribution plot for Company Size and Comfort level for discussion with employers: Based on our analysis of the plot, there seems to be some correlation between the company size and comfort level. We had 795 respondents in the smaller company size and 217 in the larger company size.

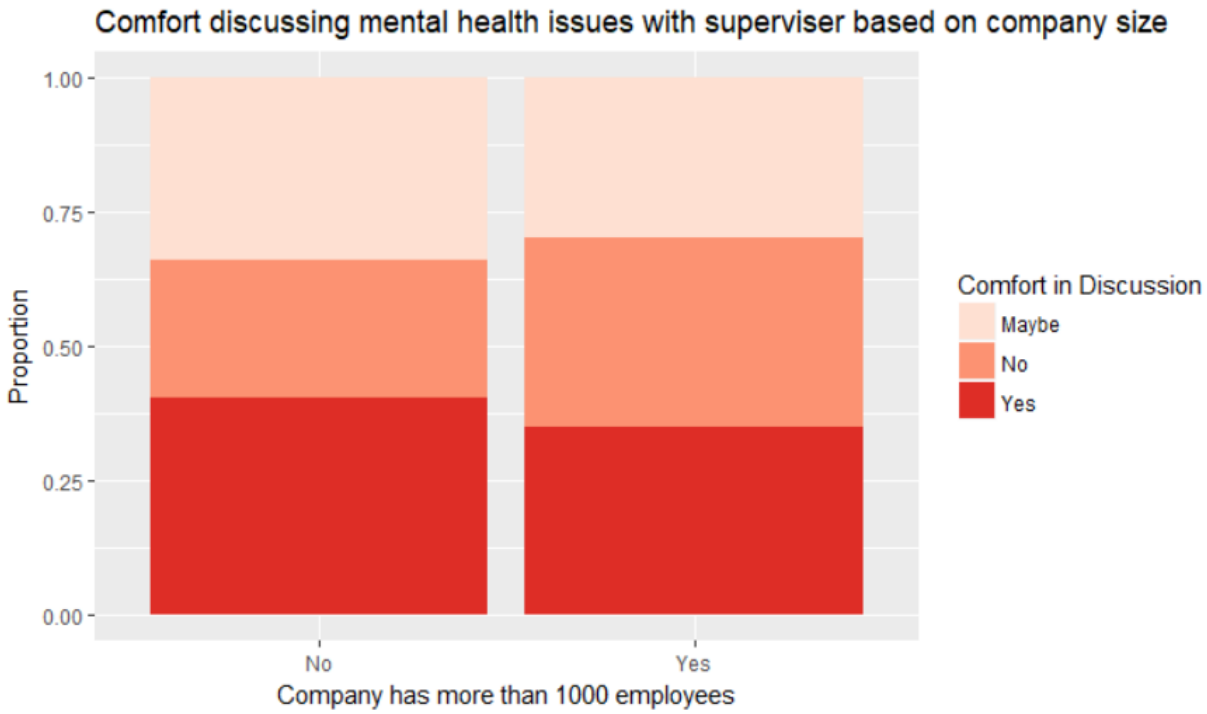


Figure 7. Variation in comfort with company size

Statistical Significance:

Multinomial Logistic Regression: From coefficients computed by multinomial regression, we found that log odds of respondents feeling comfortable versus not feeling comfortable decrease by 0.44 when we move from smaller organizations to larger organizations. Since, p-value (0.016) is smaller than our significance level of (0.05), so we reject null hypothesis that there is no association between company size and comfort in discussion with supervisors.

```
> summary(mod)
Call:
multinom(formula = discuss.supervisor ~ organization.size.large,
    data = company.size)

Coefficients:
      (Intercept) organization.size.largeYes
Maybe   0.2754097                 -0.4317561
Yes     0.4453145                 -0.4453203

Std. Errors:
      (Intercept) organization.size.largeYes
Maybe  0.09263782                 0.1926765
Yes    0.08945973                 0.1852535

Residual Deviance: 2196.919
AIC: 2204.919
> z <- summary(mod)$coefficients/summary(mod)$standard.errors
> p <- (1 - pnorm(abs(z), 0, 1))*2
> p
      (Intercept) organization.size.largeYes
Maybe 2.949309e-03                0.02503678
Yes   6.430408e-07                0.01622374
```

Figure 8. Summary of comfort versus company size single regression model

Effect of anonymity protection on comfort level of the employees

Expectation: To assess correlation between an employee's anonymity being protected at previous workplace and their comfort level discussing their mental health problems, if any, with current employer.

(H0): If anonymity was previously protected at workplace, it has no effect on comfort level for discussion with supervisors at current workplace.
(HA): If anonymity was not previously protected at workplace, it effects on comfort level of discussion with supervisors at current workplace.

Visualization:

Frequency plot for comfort vs previous anonymity protection: Large proportion of employees express not being comfortable when anonymity wasn't protected previously.

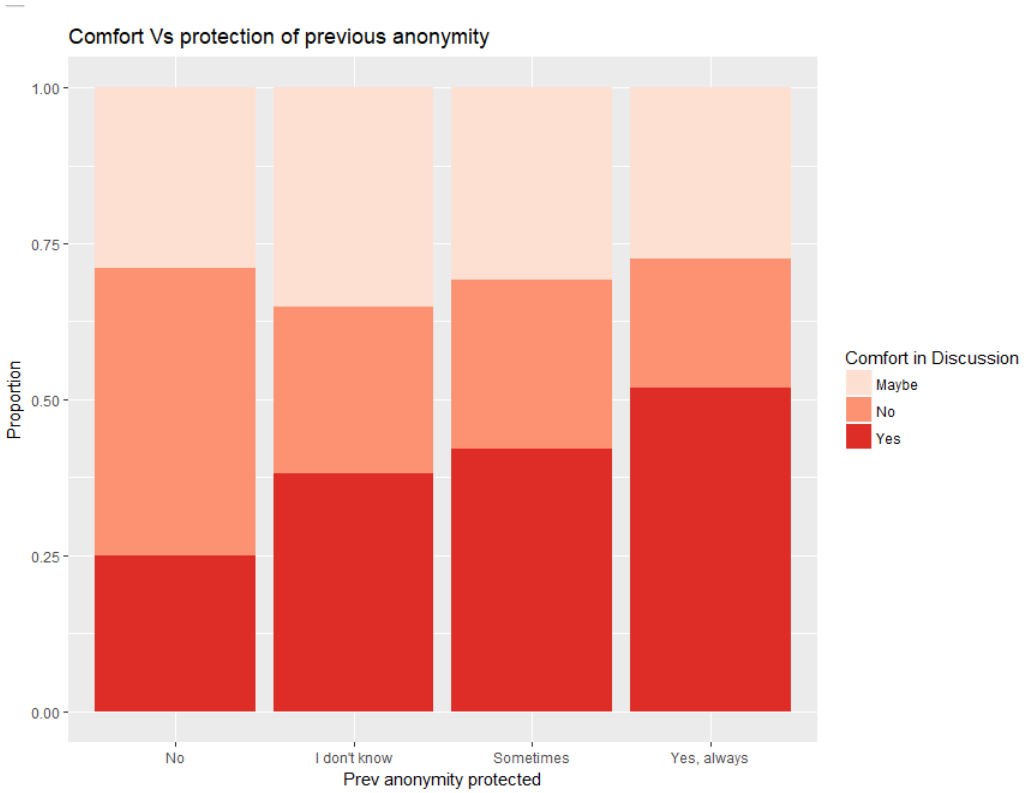


Figure 9. Variation in comfort with the protection of anonymity

Statistical Significance:

Multinomial Logistic Regression: Log odds of feeling comfortable versus not feeling comfortable increases by 1.5 when previously anonymity is protected versus not protected.

```
> summary(mod)
Call:
multinom(formula = discuss.supervisor ~ prev.anonymity.protected,
    data = company.size)

Coefficients:
      (Intercept) prev.anonymity.protectedI don't know prev.anonymity.protectedSometimes
Maybe  -0.4613069                            0.7433614                          0.5904953
Yes    -0.6097688                            0.9731592                          1.0491105
      prev.anonymity.protectedYes, always
Maybe                           0.7400299
Yes                             1.5260640

Std. Errors:
      (Intercept) prev.anonymity.protectedI don't know prev.anonymity.protectedSometimes
Maybe   0.2371100                            0.2570533                          0.3478599
Yes     0.2484752                            0.2669535                          0.3441576
      prev.anonymity.protectedYes, always
Maybe                           0.3449094
Yes                             0.3342755

Residual Deviance: 2176.666
AIC: 2192.666
> z <- summary(mod)$coefficients/summary(mod)$standard.errors
> p <- (1 - pnorm(abs(z), 0, 1))*2
> p
      (Intercept) prev.anonymity.protectedI don't know prev.anonymity.protectedSometimes
Maybe  0.05171004                           0.003829720                        0.089600469
Yes    0.01412599                           0.000266949                        0.002301068
      prev.anonymity.protectedYes, always
Maybe                        3.190669e-02
Yes                          4.988074e-06
```

Figure 10. Summary of comfort versus anonymity protection single regression model

Since, the p value is smaller(4.9e-06), the results are statistically significant, and we can reject the null hypothesis.

2.4. Models

A full model was developed based on initial research questions formulated and using backward elimination process, a reduced model was identified with the statistically significant features. In addition to features previously identified through single regressions, another feature (awareness of mental health options) was identified which was found to be positively correlated with comfort level.

Multinomial Logistic Regression Method

As the response feature (discuss.supervisor) is a factor with 3 levels 'Yes', 'No' and 'Maybe', we performed multinomial logistic regression analysis, with the reference level set at 'No' using 'nnet' package in R.

Full Model: This model included 'age', 'organization.size.large', 'prev.anonymity.protected', 'mental.health.coverage' and 'mental.health.options' as predictors and 'discuss.supervisor' as the response variable. Below summary indicates that age is not statistically significant. Additionally, organization size, anonymity protection, mental health coverage and awareness of mental health options are significant and their correlations to comfort and agree with the single multinomial regression performed as part of EDA.

Coefficients:

```
Call:
multinom(formula = discuss.supervisor ~ age + organization.size.large +
    prev.anonymity.protected + mental.health.coverage + mental.health.options,
    data = mdat4)

Coefficients:
      (Intercept)   ageSenior organization.size.largeYes prev.anonymity.protectedI don't know
Maybe  -0.9357292 -0.02889494                -0.5870206                          0.6726971
Yes    -1.3554874 -0.19754389                -0.5511425                          0.8706245
      prev.anonymity.protectedSometimes prev.anonymity.protectedYes, always
Maybe                         0.5067449                           0.6253331
Yes                           0.9406522                           1.3336427
      mental.health.coverageI don't know mental.health.coverageYes
Maybe                           0.4410408                  0.7319060
Yes                             0.7261981                  0.9466876
      mental.health.coverageNot eligible for coverage / N/A mental.health.optionsI am not sure
Maybe                                             0.5679709                          0.4901737
Yes                                               0.3827016                          0.3957778
      mental.health.optionsN/A mental.health.optionsYes
Maybe                0.3284048                0.06206543
Yes                  0.9788873                0.54424254
```

Figure 11. Summary of full multiple regression model output with all features of interest

Responses	Age Senior	Org Size Large	Anonymity protected Yes	Mental health coverage Yes	Mental health options Yes
Maybe	-0.02889494	-0.5870206	0.6253331	0.7319060	0.06206543
Yes	-0.19754389	-0.5511425	1.3336427	0.9466876	0.54424254

Table 2. Coefficients for relevant features from the full multiple regression model

P-values:

```
> z <- summary(mod1)$coefficients/summary(mod1)$standard.errors
> p <- (1 - pnorm(abs(z), 0, 1))*2
> p
       (Intercept) ageSenior organization.size.largeYes prev.anonymity.protectedI don't know
Maybe 2.501078e-03 0.8689589                0.004136788                          0.010348271
Yes   3.650144e-05 0.2514821                0.005614765                          0.001451165
      prev.anonymity.protectedSometimes prev.anonymity.protectedYes, always
Maybe                       0.152565140                         0.0763080434
Yes                         0.007664049                         0.0001060427
      mental.health.coverageI don't know mental.health.coverageYes
Maybe                         0.090759849                0.0055060816
Yes                           0.006437211                0.0004167449
      mental.health.coverageNot eligible for coverage / N/A mental.health.optionsI am not sure
Maybe                                           0.1418786                          0.01851807
Yes                                             0.3185962                          0.06104235
      mental.health.optionsN/A mental.health.optionsYes
Maybe               0.33058303               0.80500894
Yes                 0.00293408               0.02466187
```

Figure 12. p-value output for the full multiple regression model

Responses	Age Senior	Org Size Large	Anonymity protected Yes	Mental health coverage Yes	Mental health options Yes
Maybe	0.8689589	0.004136788	0.0763080434	0.0055060816	0.80500894
Yes	0.2514821	0.005614765	0.0001060427	0.0004167449	0.02466187

Table 3. Summary of p-values for the full multiple regression model

Reduced model: This model was built excluding 'age' features as it was found to be statistically insignificant in the full model. The statistical significance and correlation of other relevant features with comfort outcome remained the similar, as shown in the summary below.

Coefficients:

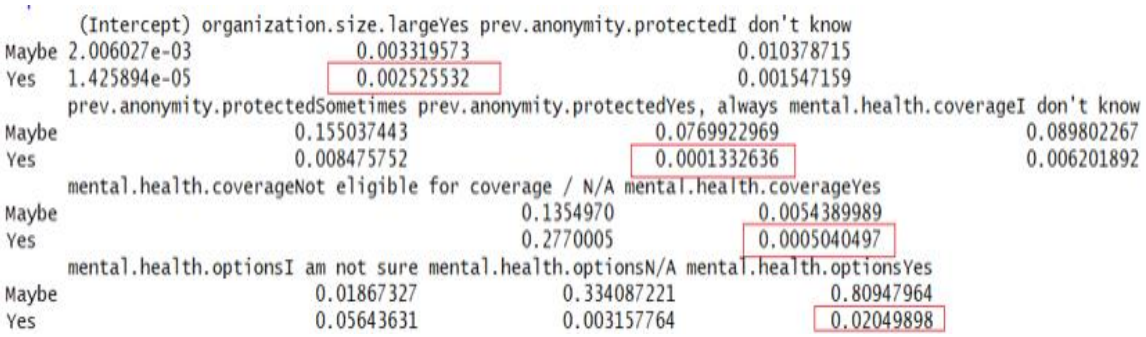
```
Call:
multinom(formula = discuss.supervisor ~ organization.size.large +
    prev.anonymity.protected + mental.health.coverage + mental.health.options,
    data = mdat4)

Coefficients:
      (Intercept) organization.size.largeYes prev.anonymity.protectedI don't know
Maybe  -0.9442307                 -0.5920654                              0.6721263
Yes    -1.4093187                 -0.5916918                              0.8651997
      prev.anonymity.protectedSometimes prev.anonymity.protectedYes, always
Maybe                         0.5033554                           0.6226737
Yes                           0.9273311                           1.3112163
      mental.health.coverageI don't know
Maybe                          0.4420993
Yes                            0.7289727
      mental.health.coverageNot eligible for coverage / N/A mental.health.coverageYes
Maybe                                             0.5750582                   0.7323285
Yes                                               0.4157656                   0.9314683
      mental.health.optionsI am not sure mental.health.optionsN/A
Maybe                          0.4894155                0.3255853
Yes                            0.4027869                0.9701761
      mental.health.optionsYes
Maybe                 0.0605879
Yes                   0.5605082
```

Figure 13. Summary of reduced multiple regression model coefficients

Responses	Org Size Large	Anonymity protected Yes	Mental health coverage Yes	Mental health options Yes
Maybe	-0.5920654	0.6226737	0.7323285	0.0605879
Yes	-0.5916918	1.3112163	0.9314683	0.5605082

Table 4. Summary of coefficients for the reduced multiple regression model

P-values:



```
      (Intercept) organization.size.largeYes prev.anonymity.protectedI don't know
Maybe 2.006027e-03               0.003319573                          0.010378715
Yes   1.425894e-05               0.002525532                          0.001547159
      prev.anonymity.protectedSometimes prev.anonymity.protectedYes, always mental.health.coverageI don't know
Maybe                       0.155037443                          0.0769922969                      0.089802267
Yes                         0.008475752                          0.0001332636                      0.006201892
      mental.health.coverageNot eligible for coverage / N/A mental.health.coverageYes
Maybe                                             0.1354970                0.0054389989
Yes                                               0.2770005                0.0005040497
      mental.health.optionsI am not sure mental.health.optionsN/A mental.health.optionsYes
Maybe                         0.01867327                0.334087221               0.80947964
Yes                           0.05643631                0.003157764               0.02049898
```

Figure 14. p-value output for the reduced multiple regression model

Responses	Org Size Large	Anonymity protected Yes	Mental health coverage Yes	Mental health options Yes
Maybe	0.003319573	0.0769922969	0.0054389989	0.80947964
Yes	0.002525532	0.0001332636	0.0005040497	0.02049898

Table 5. Summary of p-values for the reduced multiple regression model

Random Forest Method

Similar approach was followed for creation of models for random forest algorithm to estimate significance of each feature. Dataset was divided in 70:30 ratio for training and testing sets. Model consisted of 100 trees initially and five predictors. Variable importance was plotted to find top predictors.

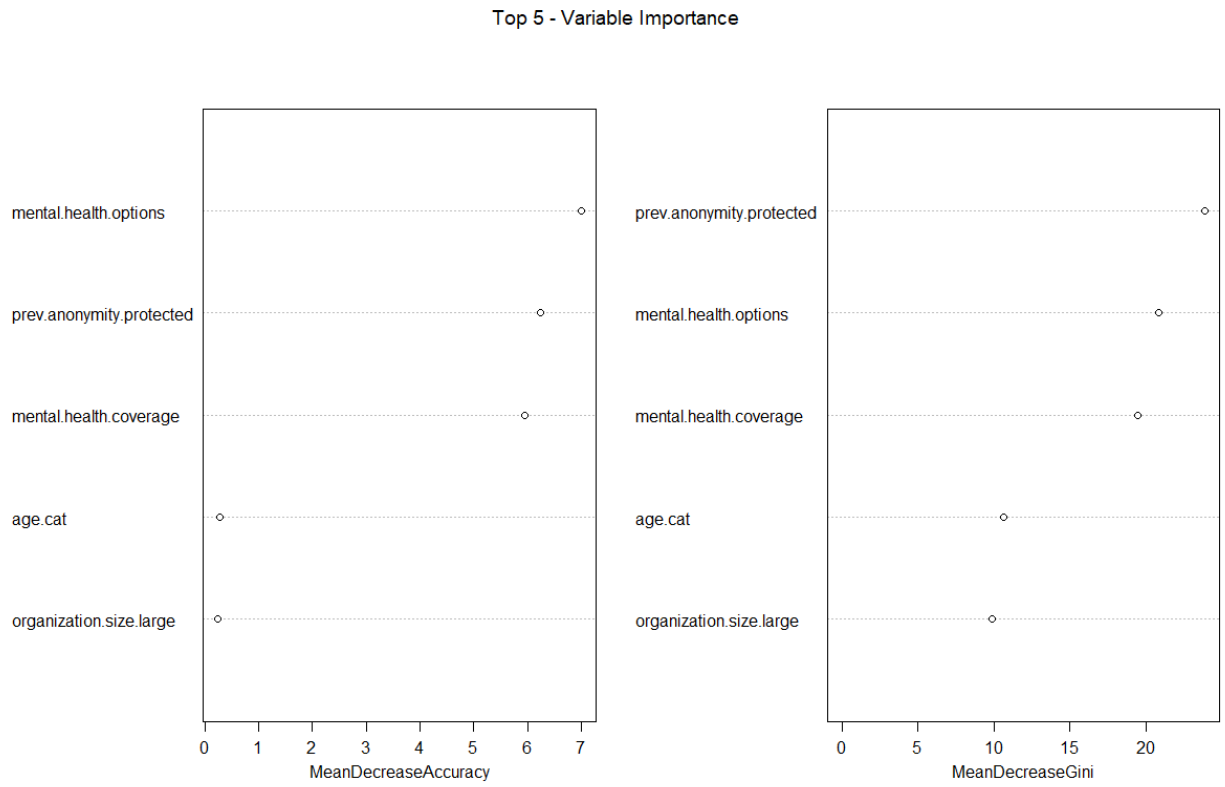


Figure 15. Variable importance for Random Forest method

'MeanDecreaseGini' in the 'measures importance of features over all splits done in the forest.

Full Model: It was built with all five predictors - age, protection of anonymity, organization size, mental health benefits and awareness of mental health benefits gave the below confusion matrix.

```
Call:
 randomForest(formula = discuss.supervisor ~ mental.health.coverage +      mental.health.options + prev.anonymity
.protected + age.cat +      organization.size.large, data = train, importance = TRUE,      ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 56.36%
Confusion matrix:
      Maybe No Yes class.error
Maybe   116 25  93   0.5042735
No       64 46  87   0.7664975
Yes      85 45 147   0.4693141
```

Figure 16. Summary of output for the full random forest model

Reduced model: Reduced model with top 3 predictors as identified by the variable importance plot were used to build another random forest model with 420 trees.

```
Call:
 randomForest(formula = discuss.supervisor ~ mental.health.coverage +      prev.anonymity.protected + mental.heal
th.options, data = train,      ntree = 420)
               Type of random forest: classification
                     Number of trees: 420
No. of variables tried at each split: 1

        OOB estimate of  error rate: 57.91%
Confusion matrix:
      Maybe No Yes class.error
Maybe    88 42 104   0.6239316
No       53 50  94   0.7461929
Yes      77 40 160   0.4223827
```

Figure 17. Summary of output for the reduced random forest method

3. Results

3.1. Exploratory Data Analysis Results

Except for age and organization size, all the other predictors have a positive correlation with the response variable('discuss.supervisor'). All the results were statistically significant except age variable, hence indicating that except for age all other predictors influencing the comfort level. Previous anonymity protection had the most effect on the comfort level with highest coefficient for slope in the single multinomial regression.

3.2. Prediction Model Results

Multinomial Logistic Regression: K-fold validation was performed to determine the accuracy for both the model identified in multinomial logistic regression. Reduced model performed better with 50% accuracy as compared to the full model with 49% accuracy.

Full Model Results: Accuracy for the full model was calculated to be 49.01% with 95% Confidence Interval of (41.93% to 56.12%). Below are the confusion matrix and summary details.

Confusion Matrix:

Prediction	No	Maybe	Yes
No	20	8	9
Maybe	20	28	19
Yes	16	31	51

Table 6. Confusion matrix for the full multiple regression model

```
Statistics by Class:

                     Class: No Class: Maybe Class: Yes
Sensitivity            0.23214       0.29851     0.6076
Specificity            0.87671       0.73333     0.4553
Pos Pred Value         0.41935       0.35714     0.4174
Neg Pred Value         0.74854       0.67808     0.6437
Prevalence             0.27723       0.33168     0.3911
Detection Rate         0.06436       0.09901     0.2376
Detection Prevalence   0.15347       0.27723     0.5693
Balanced Accuracy      0.55443       0.51592     0.5314


Residual Deviance: 2129.852
AIC: 2177.852
```

Figure 18. Summary statistics for the full multiple regression model

Reduced Model results: Accuracy for the reduced model was calculated to be 50% with 95% Confidence Interval of (42.9% to 57.1%). Below are the confusion matrix and summary details.

Confusion Matrix:

Prediction	No	Maybe	Yes
No	21	8	9
Maybe	20	31	21
Yes	15	28	49

Table 7. Confusion matrix for the reduced multiple regression model

```
Statistics by Class:

                     Class: No Class: Maybe Class: Yes
Sensitivity            0.23214       0.3284     0.5949
Specificity            0.87671       0.7037     0.4959
Pos Pred Value         0.41935       0.3548     0.4312
Neg Pred Value         0.74854       0.6786     0.6559
Prevalence             0.27723       0.3317     0.3911
Detection Rate         0.06436       0.1089     0.2327
Detection Prevalence   0.15347       0.3069     0.5396
Balanced Accuracy      0.55443       0.5160     0.5454

Residual Deviance: 2131.514
AIC: 2175.514
```

Figure 19. Summary statistics for the reduced multiple regression model

Random Forest Prediction Results: K-fold validation was performed to determine the accuracy for both the model identified in random forest. Reduced model performed better with 41.12% accuracy as compared to the full model with 39.14% accuracy.

Full Model Results: Accuracy for the full model was calculated to be 39.14% with 95% Confidence Interval of (33.62% to 44.88%). Below are the confusion matrix and summary details.

Prediction	Maybe	No	Yes
Maybe	40	25	47
No	18	24	17
Yes	43	35	55

Table 8. Confusion matrix for the full random forest model

```
Statistics by Class:

                     Class: Maybe Class: No Class: Yes
Sensitivity                0.3960   0.28571     0.4622
Specificity                0.6453   0.84091     0.5784
Pos Pred Value             0.3571   0.40678     0.4135
Neg Pred Value             0.6823   0.75510     0.6257
Prevalence                 0.3322   0.27632     0.3914
Detection Rate             0.1316   0.07895     0.1809
Detection Prevalence       0.3684   0.19408     0.4375
Balanced Accuracy          0.5207   0.56331     0.5203
```

Figure 20. Summary statistics for the full random forest model

Reduced Model Results: Accuracy for the reduced model was calculated to be 41.45% with 95% Confidence Interval of (35.85% to 47.21%). Below are the confusion matrix and summary details.

Prediction	No	Maybe	Yes
Maybe	32	21	33
Yes	15	25	17
Yes	54	38	69

Table 9. Confusion matrix for the reduced random forest model

```
Statistics by Class:

                     Class: Maybe Class: No Class: Yes
Sensitivity                0.3168   0.29762     0.5798
Specificity                0.7340   0.85455     0.5027
Pos Pred Value             0.3721   0.43860     0.4286
Neg Pred Value             0.6835   0.76113     0.6503
Prevalence                 0.3322   0.27632     0.3914
Detection Rate             0.1053   0.08224     0.2270
Detection Prevalence       0.2829   0.18750     0.5296
Balanced Accuracy          0.5254   0.57608     0.5413
```

Figure 21. Summary statistics for the reduced random forest model

3.3. Comparing Results from the two Models (Logistic and Radom Forest)

Reduced multinomial regression model had the highest accuracy (50%) and the predictors with the strongest significance include 'organization size', 'anonymity protection', 'mental Health Benefits', and 'mental health options'.

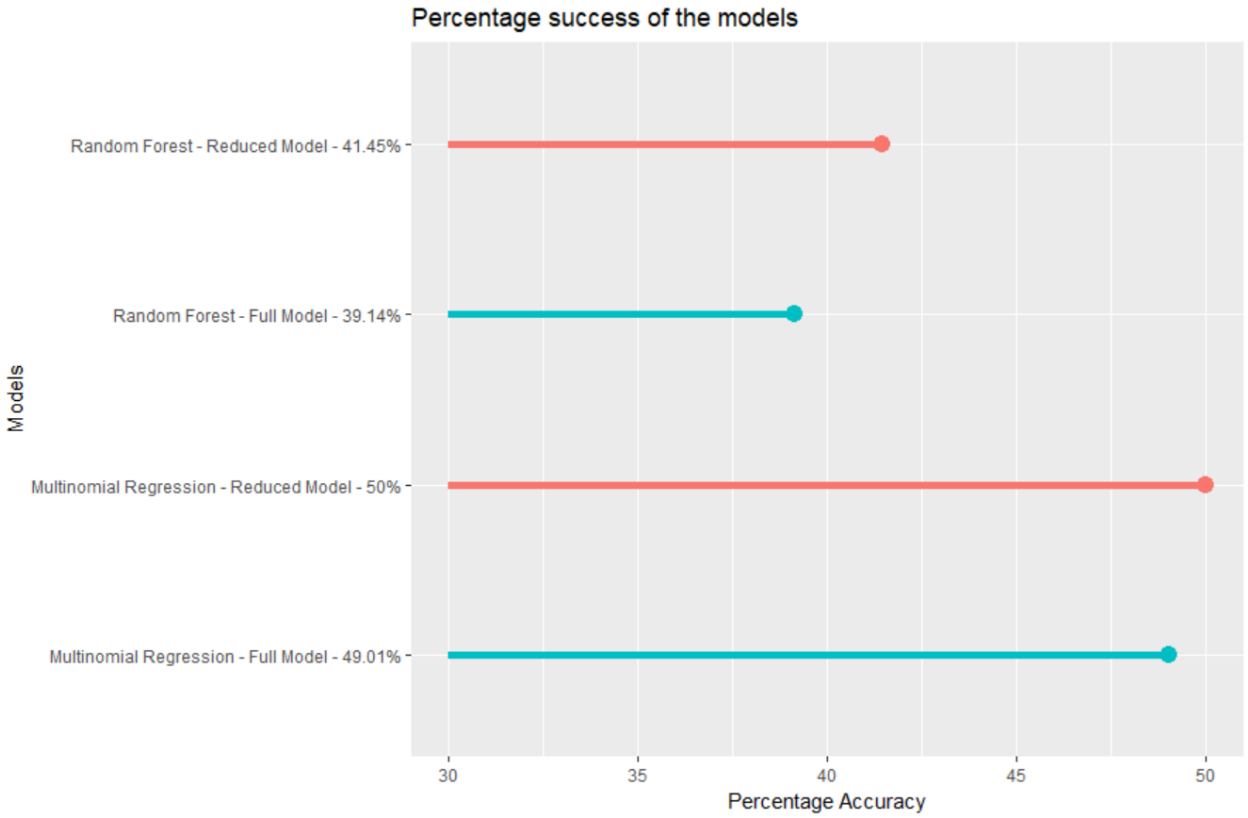


Figure 21. Comparison of accuracy of different models

4. Discussion

Research was performed to identify correlation between response and predictor variables and doesn't imply any causality. The dataset was cleansed for missing and incorrect values which led to reduction from 1433 responses to 1012 responses. As the research questions were identified based on our broader research problem, only key features were analyzed for EDA, there could be a chance of effect or association of other features in the dataset with the response variable which can be included in the future scope.

The highest accuracy obtained through the reduced multinomial logistic regression was 50% which is quite low, hence, additional feature engineering or algorithms can be applied to improve the accuracy and define a better model. Since, the responses are majorly from California (United States), results of this study could not be applied to everywhere as it fails to represent samples from all the other regions.

From our results we see that organization size, mental health benefits for employees, and protection of anonymity of employees suffering from mental health issues are all statistically significant factors that influence a tech worker's comfort in discussing mental wellbeing with employers. Based on our findings, we propose building age agnostic policies for mental health initiatives at tech organizations to assist the

employees facing such issues. It is key to train and increase awareness among the employers around how to support employees suffering from mental health issues. This would develop an environment which fosters assistance and creates an open environment that encourages employees to share and discuss such issues with their supervisors. As observed in the dataset responses, protecting anonymity of the employees suffering from mental health issues motivates them to share their concerns more freely. Hence, ensuring organization wide policies which support anonymity protection would go a long way in helping employees with these concerns. Additionally, ensuring employee awareness around mental health benefits provided by the employer and effective implementation of these benefits to support employees is needed.

References

Mental health in the workspace Infographic (2017). Retrieved from "https://elearninginfographics.com/mental-health-in-the-workplace-infographic/"

OSMI (n.d.). OSMI Mental Health in Tech Survey 2016: Data on prevalence and attitudes towards mental health among tech workers. Retrieved from https://www.kaggle.com/osmi/mental-health-in-tech-2016/feed

Nguyen, J. (2016). Are You More Than Okay: The State of Mental Health in Tech in 2016. Retrieved from "https://modelviewculture.com/pieces/are-you-more-than-okay-the-state-of-mental-health-in-tech-in-2016"

Kasbergen, N. (2017). Mental Health in Tech at NationJS 2016. Retrieved from "https://npr.codes/mental-health-in-tech-at-nationjs-2016-9ea6ce99f6f0"

Workplace stress and mental health in the workplace: infographic (2017). Retrieved from "https://www.thepeoplespace.com/brand/articles/workplace-stress-and-mental-health-workplace-infographic"

APPENDIX

I. Details of Features Identified from the dataset:

Sr. No.	Feature Name	Actual Survey Question	Variable Type
1	age	What is your age?	Numerical
2	gender	What is your gender? (Female, Male, Transgender, Genderqueer, Refused)	Categorical
3	work.country	What country do you work in? (List of all the countries)	Categorical
4	work.us.state	What US state or territory do you work in? (List of all the US states)	Categorical
5	leave.sanction	If a mental health issue prompted you to request a medical leave from work, asking for that leave would be: (I don't know, Neither easy nor difficult, Somewhat difficult, Somewhat easy, Very easy, Very difficult)	Categorical
6	discuss.coworker	Would you feel comfortable discussing a mental health disorder with your coworkers? (Maybe, Yes, No)	Categorical
7	discuss.supervisor	Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)? (Maybe, Yes, No)	Categorical
8	Company.size	How many employees does your company or organization have? (1-5, 6-25, 26-100, 100-500, 500-1000, More than 1000)	Categorical
9	mental.health.options	Do you know the options for mental health care available under your employer-provided coverage? (No, Yes, N/A, I am not sure)	Categorical
10	current.mental.disorder	Do you currently have a mental health disorder?	Categorical

		(Yes, No, Maybe)	
12	company.resources	Does your employer offer resources to learn more about mental health concerns and options for seeking help? (Yes, No, I don't know)	Categorical
14	mental.health.coverage	Does your employer provide mental health benefits as part of healthcare coverage? (Yes, No, I don't know, Not eligible for coverage/N/A)	Categorical
15	prev.anonymity.protected	Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers? (I don't know, No, Sometimes, Always)	Categorical

Table 10. Features of interest identified from the dataset

II. Preliminary Analysis:

Analyzing the top ten mental health disorders in the tech industry: Anxiety and mood disorders are top mental health issues faced by the employees in the tech industry.

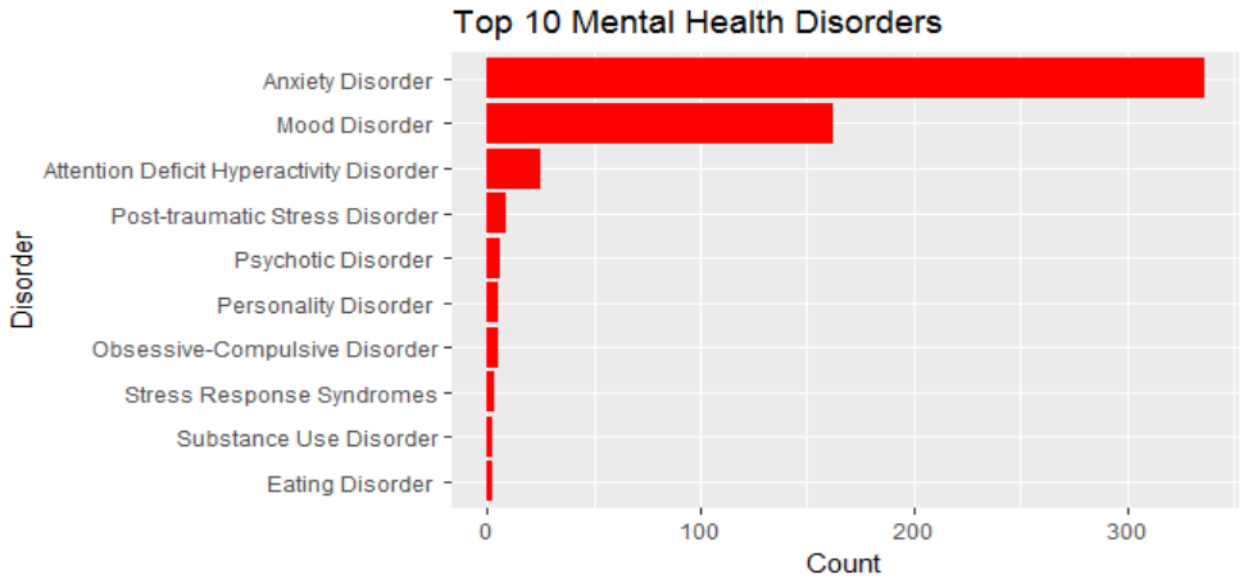


Figure 23. Distribution of respondents by disorders

Percentage of employees currently suffering from mental health issues: 40% of the respondents are currently experiencing mental health issues.

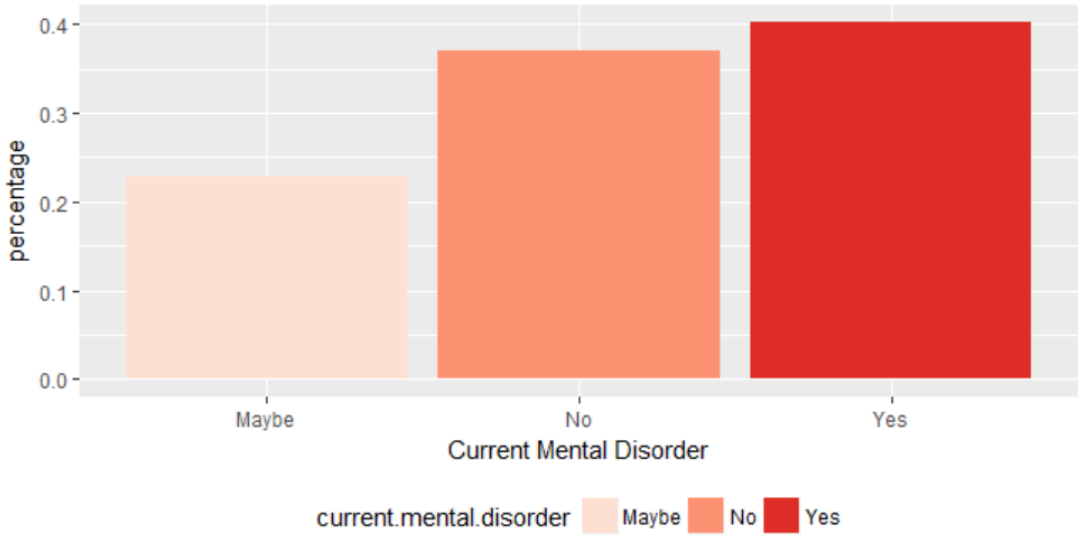


Figure 24. Distribution of respondents when asked if suffering from a mental health disorder

Mental health benefit awareness of the employees: Only **46%** of respondents are aware about mental health benefits provided by the employer.

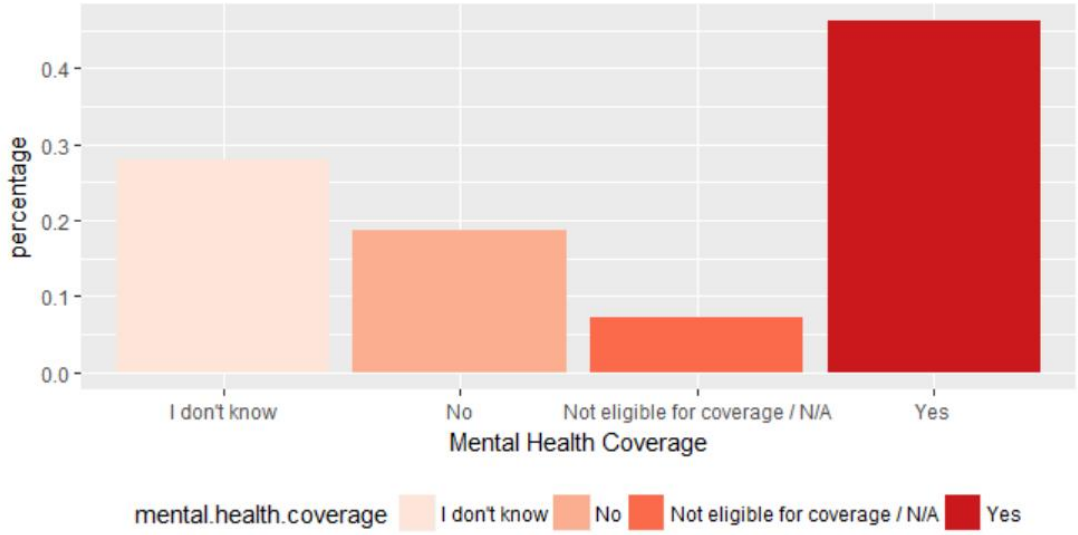


Figure 25. Distribution of respondents by provision of mental health coverage from employer

Anonymity protection: 69% of respondents are unsure about discussing mental health issues if their anonymity was not protected at previous workplace.

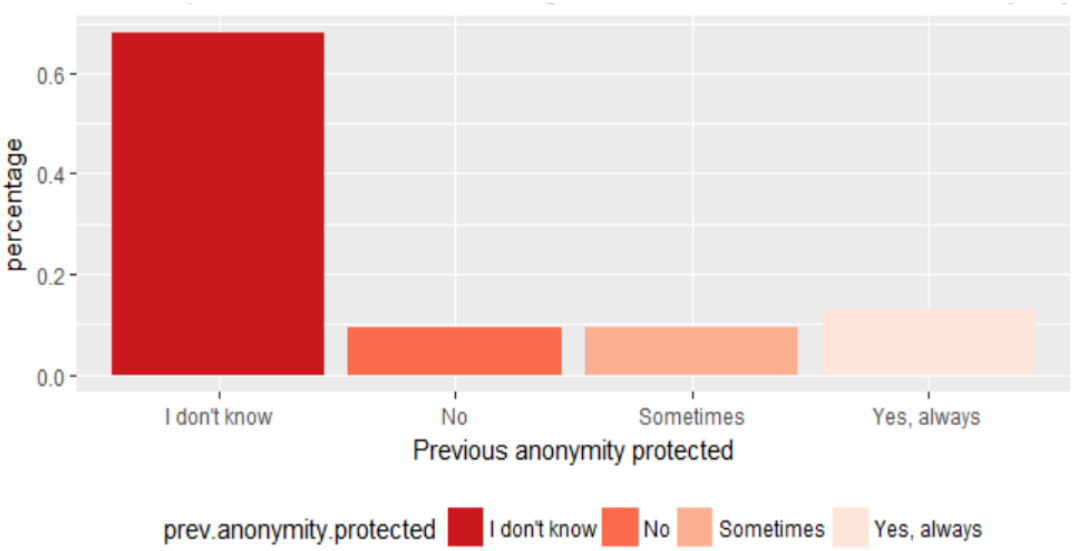


Figure 26. Distribution of respondents by protection of anonymity

Identification of geospatial aspects of the data: Since the survey was open to participants from all parts of the world, the distribution of respondents by country was explored, and the highest density country was further analyzed, to study region-wise density of respondents.

The highest number of respondents in the US were from **California** due to the location of the world's largest tech corporations in the **Silicon Valley**.

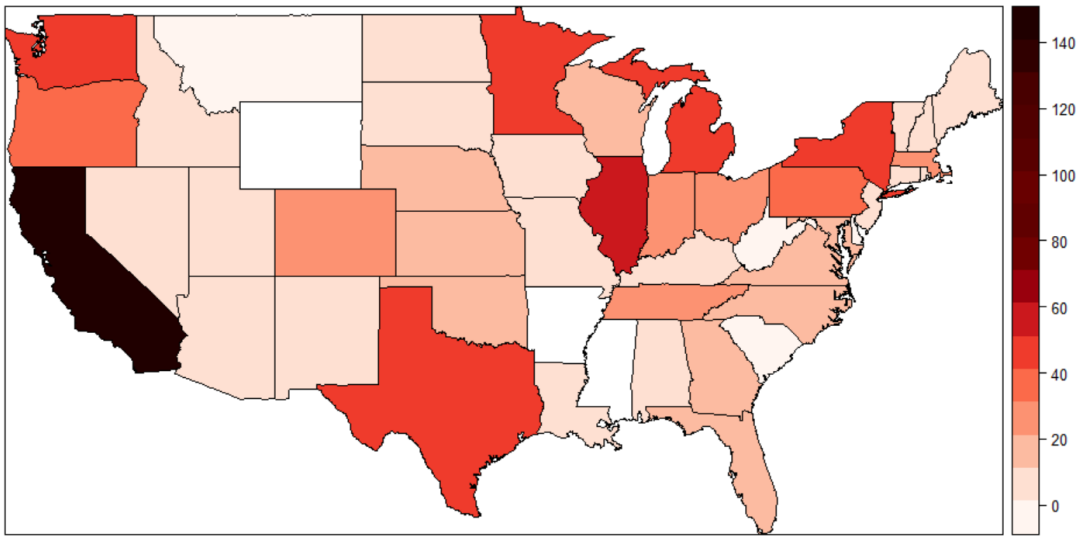


Figure 27. Distribution of respondents across the USA

III. R Code

Data Cleaning Code:

```
###### PACKAGES & LIBRARIES ######

library('dplyr')
library('ggplot2')
library('RColorBrewer')

###### READING THE DATASET ######

mentalHealth <- read.csv("mental-heath-in-tech-2016_20161114.csv")              # Reading the dataset

head(mentalHealth)                                                  # Glimpse of the dataset
colnames(mentalHealth)                                              # Columns in the dataset
dim(mentalHealth)                                                   # Dimensions of the dataset


###### DATASET PREPARATION ######

# RENAMING COLUMNS & CREATING A DATAFRAME

df <- data.frame(mentalHealth)

mdat <- rename(df,
        work.country = What.country.do.you.work.in.,
        work.us.state = What.US.state.or.territory.do.you.work.in.,
        age = What.is.your.age.,
        gender = What.is.your.gender.,
        discuss.supervisor =
Would.you.feel.comfortable.discussing.a.mental.health.disorder.with.your.direct.supervisor.s..,
        discuss.coworker =
Would.you.feel.comfortable.discussing.a.mental.health.disorder.with.your.coworkers.,
        leave.sanction =
If.a.mental.health.issue.prompted.you.to.request.a.medical.leave.from.work..asking.for.that.leave.would.
be.,
        employee.count= How.many.employees.does.your.company.or.organization.have.,
        mental.health.options =
Do.you.know.the.options.for.mental.health.care.available.under.your.employer.provided.coverage.,
        current.mental.disorder = Do.you.currently.have.a.mental.health.disorder.,
        self.employed = Are.you.self.employed.,
        company.resources =
Does.your.employer.offer.resources.to.learn.more.about.mental.health.concerns.and.options.for.seeking
.help.,
        mental.health.coverage =
Does.your.employer.provide.mental.health.benefits.as.part.of.healthcare.coverage.,
```

```
        prev.anonymity.protected =
Was.your.anonymity.protected.if.you.chose.to.take.advantage.of.mental.health.or.substance.abuse.treatment.resources.with.previous.employers.)

mdat %>%
select(age,gender,work.country,work.us.state,leave.sanction,discuss.supervisor,discuss.coworker,employee.count,mental.health.options,current.mental.disorder,self.employed,company.resources,mental.health.coverage,prev.anonymity.protected) %>% str()


# CLEANING GENDER COLUMN #

mdat$gender <- as.character(mdat$gender)                    # Converting gender column to character type

table(is.na(mdat$gender))                                   # Checking for null values


# Cleaning Male values
mdat[mdat$gender == "Male", "gender"] <- "M"
mdat[mdat$gender == "male", "gender"] <- "M"
mdat[mdat$gender == "MALE", "gender"] <- "M"
mdat[mdat$gender == "Man", "gender"] <- "M"
mdat[mdat$gender == "man", "gender"] <- "M"
mdat[mdat$gender == "m", "gender"] <- "M"
mdat[mdat$gender == "man ", "gender"] <- "M"
mdat[mdat$gender == "Dude", "gender"] <- "M"
mdat[mdat$gender == "mail", "gender"] <- "M"
mdat[mdat$gender == "M|", "gender"] <- "M"
mdat[mdat$gender == "Cis male", "gender"] <- "M"
mdat[mdat$gender == "Male (cis)", "gender"] <- "M"
mdat[mdat$gender == "Cis Male", "gender"] <- "M"
mdat[mdat$gender == "cis male", "gender"] <- "M"
mdat[mdat$gender == "cisdude", "gender"] <- "M"
mdat[mdat$gender == "cis man", "gender"] <- "M"
mdat[mdat$gender == "Male.", "gender"] <- "M"
mdat[mdat$gender == "Male ", "gender"] <- "M"
mdat[mdat$gender == "male ", "gender"] <- "M"
mdat[mdat$gender == "Malr", "gender"] <- "M"
mdat[841,"gender"] <- "M"

# Cleaning Female values
mdat[mdat$gender == "Female", "gender"] <- "F"
mdat[mdat$gender == "Female ", "gender"] <- "F"
mdat[mdat$gender == " Female", "gender"] <- "F"
mdat[mdat$gender == "female", "gender"] <- "F"
mdat[mdat$gender == "female ", "gender"] <- "F"
mdat[mdat$gender == "Woman", "gender"] <- "F"
mdat[mdat$gender == "woman", "gender"] <- "F"
```

```
mdat[mdat$gender == "f", "gender"] <- "F"
mdat[mdat$gender == "Cis female", "gender"] <- "F"
mdat[mdat$gender == "Cis female ", "gender"] <- "F"
mdat[mdat$gender == "Cisgender Female", "gender"] <- "F"
mdat[mdat$gender == "Cis-woman", "gender"] <- "F"
mdat[mdat$gender == "fem", "gender"] <- "F"
mdat[1091, "gender"] <- "F"
mdat[17, "gender"] <- "F"

# Cleaning Gender Queer Values
mdat[!is.na(mdat$gender)&mdat$gender == "Agender", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Androgynous", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Bigender", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Female or Multi-Gender Femme", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "female-bodied; no feelings about gender", "gender"] <-
"GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Fluid", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "fm", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "GenderFluid", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "GenderFluid (born female)", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Genderflux demi-girl", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "genderqueer", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Genderqueer", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "fm", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "genderqueer woman", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "human", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Human", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Unicorn", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Male/genderqueer", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "nb masculine", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "non-binary", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "Nonbinary", "gender"] <- "GQ"
mdat[!is.na(mdat$gender)&mdat$gender == "AFAB", "gender"] <- "GQ"

# Cleaning Transgender values
mdat[!is.na(mdat$gender)&mdat$gender == "Male (trans, FtM)", "gender"] <- "TG"
mdat[!is.na(mdat$gender)&mdat$gender == "Transgender woman", "gender"] <- "TG"


# see what's left
index <- which(mdat$gender != "M" & mdat$gender != "F" & mdat$gender != "GQ" & mdat$gender !=
"TG")

mdat[index, "gender"]

# create vector of final gender values to fill in based on index
last.genders <- c("F", "TG", "GQ", "GQ", "F", "GQ", "GQ", "GQ", "M", "Refused", "GQ", "GQ", "GQ", "TG",
"GQ", NA)
```

```
# fill in remaining values
mdat[index, "gender"] <- last.genders

# check gender
table(data$gender)

# convert gender back to factor
mdat$gender <- as.factor(mdat$gender)

# Plotting gender with frequency
ggplot(mdat, aes(x = gender, fill=gender)) +
  geom_bar() +
  scale_fill_brewer(palette = "Dark2", name="Gender",
                    breaks=c("M", "F", "TG", "GQ", "Refused", "NA"),
                    labels=c("Male", "Female", "Transgender", "Gender Queer", "Refused", "NA")) +
  ggtitle('Frequency distribution by gender') +
  xlab('Gender') +
  ylab('Count')


# CLEANING AGE COLUM #

summary(mdat$age)
# Filter records with age 3 and 323
mdat1 <- mdat %>% filter(age != 323 & age !=3)
summary(mdat1$age)
# Max age is still 99
mdat1 <- mdat %>% filter(age > 15 & age <= 80)
summary(mdat1$age)
dim(mdat1)

mdat1$age <-cut(mdat1$age, breaks = c(14,35,78), labels = c('Young','Senior'))

write.csv(mdat1, 'MentalHealthCleanedDataset.csv')
```

Analysis of Top 10 disorders:

```
require(dplyr)
require(ggplot2)
require(tidyr)
dat <- read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)
head(names(dat))
names(dat) = tolower(names(dat))
dat2 = rename(dat,
              mdis_now = current.mental.disorder,
              mdis_past = have.you.had.a.mental.health.disorder.in.the.past.,
```

```
        mdis_diagnosed =
have.you.been.diagnosed.with.a.mental.health.condition.by.a.medical.professional.,
        sex = gender,
        country = what.country.do.you.live.in.,
        mdis_cat = if.yes..what.condition.s..have.you.been.diagnosed.with.)


# Save all mdis variables in a new df
dat2 %>% select(contains('mdis'), age, sex, country) %>% str()
disorder_vars = dat2 %>% select(contains('mdis'))

names(disorder_vars)

tmp = dat2 %>% separate(mdis_cat,
                        sep = '\\|',
                        c('mdis_1', 'mdis_2', 'mdis_3', 'mdis_4', 'mdis_5',
                          'mdis_6', 'mdis_7', 'mdis_8', 'mdis_9'),
                        fill = 'right')
# Turn those new variables into factors
tmp2 = tmp %>% select(matches('mdis_[1-9]')) %>% mutate_all(.funs = 'as.factor')
tmp2 %>% select(matches('mdis_[1-9]')) %>% str()

# add these new variables to the disorder_vars df
disorder_vars = cbind(disorder_vars, tmp2)

# Order the factor levels of mdis_1 by frequency
tmp = table(disorder_vars$mdis_1)
disorder_vars$mdis_1 = factor(disorder_vars$mdis_1,
                              levels = names(tmp[order(tmp, decreasing = TRUE)]))

# shorten the factor levels (to plot below)
# split at first '(' and only take the first par
#levels(disorder_vars$mdis_1) = substr(levels(disorder_vars$mdis_1), 1, 15)
#strsplit(levels(disorder_vars$mdis_1), split = '\\(')

levels(disorder_vars$mdis_1) = sapply(strsplit(levels(disorder_vars$mdis_1), split = "\\("), `[`, 1)

subdat = subset(disorder_vars, mdis_1 != "I haven\'t been formally diagnosed, so I felt uncomfortable
answering, but Social Anxiety and Depression.")
subdat = subset(subdat, !is.na(mdis_1))


colnames(subdat)[13] <- 'mdis_1'
summary(subdat$mdis_1)
plot.mdis_1 <- subdat %>%
  group_by(mdis_1) %>%
  summarize(count = n()) %>%
  top_n(n=10, wt=count) %>%
```

```
  arrange(desc(count))

summary(plot.mdis_1)
ggplot(plot.mdis_1, aes(x = reorder(mdis_1, count), y= count, fill=mdis_1 )) +
  geom_bar(stat = 'identity') +
  scale_fill_manual(values = c("red","red","red","red","red","red","red","red","red","red")) +
  ggtitle('Top 10 Mental Health Disorders') +
  xlab('Disorder') +
  ylab('Count') +
  theme(legend.position="none") +
  coord_flip()
```

Geographical Distribution of Respondents:

```
install.packages('maptools')
install.packages('sp')
install.packages('rworldmap')
library(maps)
library(maptools)
library(sp)
library(rworldmap)

####### Analyzing US Specific reponse distribution ######

df1 <- read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)

t <- table(df1$work.us.state)
w <- as.data.frame(t)
w <- data.frame(w[-c(1),])
colnames(w) <- c("state","freq")
levels(w$state) <- tolower(levels(w$state))
w

mapUSA <- map('state',  fill = TRUE,  plot = FALSE)
nms <- sapply(strsplit(mapUSA$names,  ':'),  function(x)x[1])
USApolygons <- map2SpatialPolygons(mapUSA,  IDs = nms,  CRS('+proj=longlat'))

as.character(w$state)
idx <- match(unique(nms),w$state)
dat2 <- data.frame(value = w$freq[idx], state = unique(nms))
row.names(dat2) <- unique(nms)

USAsp <- SpatialPolygonsDataFrame(USApolygons,  data = dat2)

spplot(USAsp['value'],
col.regions=c(brewer.pal(8,"Reds"),"#700000","#600000","#680000","#580000","#500000","#480000","#
300000","#200000"))
```

```
library(ggplot2)
library(magrittr)
library(dplyr)
library(plotly)
library(RColorBrewer)
library(gridExtra)
```

Preliminary Analysis:

```
data <- read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)

#mental health coverage
No_cov<-subset(data, data$mental.health.coverage == "No"|data$mental.health.coverage ==
"Yes"|data$mental.health.coverage == "Not eligible for coverage / N/A"|data$mental.health.coverage ==
"I don't know")
ggplot(No_cov, aes(x = mental.health.coverage,y=percentage, fill=mental.health.coverage)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  #scale_fill_brewer(palette = "Reds")+
  #scale_colour_manual(values = rev(palette("Reds")))+
  scale_fill_brewer(palette="Reds")+
  #ggtitle("% of respondents unsure discussing mental health issues if their anonymity was not protected
at previous workplace")+
  xlab("Mental Health Coverage")+
  theme(legend.position= "bottom" )



#Anonymity protected
anom.data<-subset(data, data$prev.anonymity.protected=='I don\'t know' |
data$prev.anonymity.protected=='No' | data$prev.anonymity.protected== 'Sometimes' |
data$prev.anonymity.protected=='Yes, always')
ggplot(anom.data, aes(x = prev.anonymity.protected,y=percentage, fill=prev.anonymity.protected)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  #scale_fill_brewer(palette = "Reds")+
  #scale_colour_manual(values = rev(palette("Reds")))+
  scale_fill_brewer(palette="Reds", direction=-1)+
  ggtitle("% of respondents unsure discussing mental health issues if their anonymity was not protected at
previous workplace")+
  xlab("Previous anonymity protected")+
  theme(legend.position= "bottom" )

#current mental disorder
ggplot(data, aes(x = current.mental.disorder,y=percentage, fill=current.mental.disorder)) +
  geom_bar(aes(y = (..count..)/sum(..count..)))+
  #scale_fill_brewer(palette = "Reds")+
  #scale_colour_manual(values = rev(palette("Reds")))+
  scale_fill_brewer(palette="Reds")+
```

```
  #ggtitle("% of respondents unsure discussing mental health issues if their anonymity was not protected at previous workplace")+
  xlab("Current Mental Disorder")+
  theme(legend.position= "bottom")
```

Question 1: Analysis

```
library('dplyr')
library('ggplot2')
library('RColorBrewer')
library('nnet')

dat <- read.csv('MentalHealthCleanedDataset.csv')

dim(dat)

# Plot of comfort level with Supervisor
ggplot(dat, aes(x=discuss.supervisor, fill=discuss.supervisor, group=discuss.supervisor)) +
geom_bar(position = 'dodge') + scale_fill_brewer(palette = "Reds", direction = -1) + labs(x="Comfort Level with Supervisor",y="Respondent Count",title="Count of Respondents in each comfort level")

# Plot of age groups and respondet count

ggplot(dat, aes(x=age, fill=age, group=age)) + geom_bar(position = 'dodge') + scale_fill_brewer(palette = "Reds", direction = -1) + labs(x="Age Group",y="Respondent Count",title="Count of Respondents in each Age group")

# chi-square for age

t <- with(dat, table(discuss.supervisor, age))

chisq.test(t)

## Multinom for age and comfort variable

dat$discuss.supervisor <- relevel(dat$discuss.supervisor, ref="No")
dat$age <- relevel(dat$age, ref="Young")

mod <- multinom(discuss.supervisor~age, family="multinom", data=dat)
summary(mod)
z <- summary(mod)$coefficients/summary(mod)$standard.errors
z
# 2-tailed z test
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

Question 2: Analysis

```
library(ggplot2)
library(RColorBrewer)
library(dplyr)

data<-read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)

dim(data)
# DISCARDING RECORDS FOR SELF-EMPLOYED PEOPLE
survey.data <- data %>% filter(self.employed==0)
#data <- data[data$self.employed==0,]
dim(survey.data)

#3 a. People reaching out to employer with respect to mental benefits being provided

survey.data$mental.health.comfort.supervisor2 <- relevel(survey.data$discuss.supervisor, ref = "No")

discuss.supervisor.group <- survey.data %>% group_by(mental.health.coverage,
mental.health.comfort.supervisor2)
discuss.supervisor.group.plot <- discuss.supervisor.group %>% summarise(count = n()) %>%
mutate(frequency = count/sum(count))

ggplot(discuss.supervisor.group.plot,
      aes(x = mental.health.coverage, y = frequency, fill = mental.health.comfort.supervisor2)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Reds") +
  ggtitle('People reaching out to employer with respect to mental benefits being provided') +
  xlab('Mental health benefits provided from employer') +
  ylab('Proportion') +
  theme(axis.text.x = element_text(angle=35,vjust=0.6))+
  guides(fill=guide_legend(title='Comfort in Discussion'))

# Test of statistical significance
survey.data$mental.health.comfort.supervisor2 <- relevel(survey.data$discuss.supervisor, ref = "No")
survey.data$mental.health.coverage<- relevel(survey.data$mental.health.coverage,ref="No")
mod <- multinom(mental.health.comfort.supervisor2 ~ mental.health.coverage + mental.health.options,
data = survey.data)
summary(mod)

z <- summary(mod)$coefficients/summary(mod)$standard.errors
z

p <- (1 - pnorm(abs(z), 0, 1))*2
p
```

Question 3: Analysis

```
library(dplyr)
```

```
library(ggplot2)
library(RColorBrewer)
library(nnet)

# load dataset
survey.data <- read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)
str(survey.data)

dim(survey.data)

# filter the missing responses
company.size <- survey.data %>% filter(employee.count != '')
summary(company.size$employee.count)
levels(company.size$organization.size.large)
company.size$organization.size.large <- rep("0", nrow(company.size))
company.size$organization.size.large[company.size$employee.count=="More than 1000"] <- "1"
company.size$organization.size.large <- as.factor(company.size$organization.size.large)
levels(company.size$organization.size.large) <- c("No","Yes")
table(company.size$organization.size.large)


# ordering the factors
levels(company.size$discuss.supervisor)
company.size$discuss.supervisor <- relevel(company.size$discuss.supervisor, ref = 'No')
company.size$discuss.supervisor <- factor(company.size$discuss.supervisor, levels = c('Maybe', 'No',
'Yes'))

# Frequency distribution of respondents overall by company size
ggplot(company.size, aes(x = organization.size.large, fill=organization.size.large)) +
  geom_bar() + scale_fill_brewer(palette = "Reds") +
  ggtitle('Frequency distribution of respondents by company size') +
  xlab('Organization has more than 1000 employees') +
  ylab('Respondent Count') +
  guides(fill=guide_legend(title='More than 1000 employees'))

# Relation between company size and discussing a mental health issue with a supervisor
summary(company.size$discuss.supervisor)
discuss.supervisor.group <- company.size %>% group_by(organization.size.large, discuss.supervisor)
discuss.supervisor.group.plot <- discuss.supervisor.group %>% summarise(count = n()) %>%
mutate(frequency = count/sum(count))
# plot the frequency distribution
ggplot(discuss.supervisor.group.plot,
       aes(x = organization.size.large, y = frequency, fill = discuss.supervisor)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Reds") +
  ggtitle('Comfort discussing mental health issues with superviser based on company size') +
  xlab('Company has more than 1000 employees') +
  ylab('Proportion') +
```

```
  guides(fill=guide_legend(title='Comfort in Discussion'))

chisq.test(table(company.size$discuss.supervisor, company.size$organization.size.large))

levels(company.size$mental.health.coverage)
company.size$mental.health.coverage <- relevel(company.size$mental.health.coverage, ref = 'No')

company.size$discuss.supervisor <- relevel(company.size$discuss.supervisor, ref = "No")
mod <- multinom(discuss.supervisor ~ organization.size.large, data = company.size)

summary(mod)

z <- summary(mod)$coefficients/summary(mod)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2
p
```

Question 4: Analysis

```
library(ggplot2)
library(magrittr)
library(dplyr)
library(plotly)
library(RColorBrewer)
library(nnet)
library(gridExtra)

data <- read.csv("MentalHealthCleanedDataset.csv", header = TRUE, stringsAsFactors = TRUE)

anom.data<-subset(data, data$prev.anonymity.protected=='I don\'t know' |
data$prev.anonymity.protected=='No' | data$prev.anonymity.protected== 'Sometimes' |
data$prev.anonymity.protected=='Yes, always')
anom.data<-subset(anom.data, anom.data$discuss.coworker=='Maybe'|
anom.data$discuss.coworker=='Yes' | anom.data$discuss.coworker=='No' )
anom.data<-subset(anom.data, anom.data$discuss.supervisor=='Maybe'|
anom.data$discuss.supervisor=='Yes' | anom.data$discuss.supervisor=='No' )

#Previous anonymity protected vs discusing with supervisor
anom.gropued<- anom.data %>% group_by(prev.anonymity.protected,discuss.supervisor)
forPlotting<-anom.gropued %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
gg <- ggplot(forPlotting, aes(x = prev.anonymity.protected, y = freq, fill = discuss.supervisor)) +
  geom_bar(stat = "identity",position = "stack") +
  ggtitle("Comfort Vs protection of previous anonymity") +
  xlab("Previous anonymity protected") +
  ylab("Proportion") +
  guides(fill=guide_legend(title="Comfort in Discussion")) +
```

```
  scale_fill_brewer(palette = "Reds")+
  theme(axis.text.x = element_text(angle=65,vjust=0.6))
ggplotly(gg)


anom.data$discuss.coworker <-factor(anom.data$discuss.coworker)
levels(anom.data$discuss.coworker)

#multinomial logistic regression for coworkers
anom.data$discuss.coworker2 <- relevel(anom.data$discuss.coworker, ref = 'No')
test <- multinom(discuss.coworker2 ~ prev.anonymity.protected, data = anom.data)
summary(test)

z <- summary(test)$coefficients/summary(test)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2
p

#multinomial logistic regression for supervisors
levels(anom.data$discuss.supervisor)
anom.data$discuss.supervisor2 <- relevel(anom.data$discuss.supervisor, ref = 'No')
levels(anom.data$discuss.supervisor2)
mod <- multinom(discuss.supervisor2 ~ prev.anonymity.protected, data = survey.data)
summary(mod)

z <- summary(mod)$coefficients/summary(mod)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2
p
```

Multiple Multinomial Regression:

```
library('caret')
library('dplyr')
library(AUC)
library('nnet')
library('RColorBrewer')
library('ggplot2')

##### Load the clean dataset #####

mdat1 <- read.csv('MentalHealthCleanedDataset.csv')   # Reading the dataset

#### Setting the baseline for different variables ####

mdat4$discuss.supervisor <- relevel(mdat1$discuss.supervisor, ref="No")

mdat4$prev.anonymity.protected <- relevel(mdat1$prev.anonymity.protected, ref="No")
```

```
mdat4$mental.health.coverage <- relevel(mdat1$mental.health.coverage, ref="No")

mdat4$mental.health.options <- relevel(mdat1$mental.health.options, ref="No")

levels(mdat4$discuss.supervisor)

#### Logistic regression model with all our predictors ####

mod1 <- multinom(discuss.supervisor~ age + organization.size.large + prev.anonymity.protected +
mental.health.coverage + mental.health.options, data=mdat4)

mostImportantVariables <- varImp(mod1)
mostImportantVariables$Variables <- row.names(mostImportantVariables)
mostImportantVariables <- mostImportantVariables[order(-mostImportantVariables$Overall),]
print(mostImportantVariables)
summary(mod1)

z <- summary(mod1)$coefficients/summary(mod1)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2
p

##### Model 2 ####

mod2 <- multinom(discuss.supervisor~ organization.size.large + prev.anonymity.protected +
mental.health.coverage + mental.health.options, data=mdat4)

summary(mod2)

z <- summary(mod2)$coefficients/summary(mod2)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1))*2
p

######################## K Fold Validation ###################

###### Model 1 ######

Train <- createDataPartition(mdat4$discuss.supervisor, p=0.8, list=FALSE)

training <- mdat4[Train, ]
testing <- mdat4[-Train, ]

ctrl <- trainControl(method = "repeatedcv", number = 6, savePredictions = TRUE)

mod_fit <- train(discuss.supervisor~age + organization.size.large + prev.anonymity.protected +
mental.health.coverage + mental.health.options,  data=training, method="multinom",
```

```
            trControl = ctrl, tuneLength = 5)


pred = predict(mod_fit, newdata=testing)
cm <- confusionMatrix(data=pred, testing$discuss.supervisor)
cm
mc <- cm$table
percent$value[1] <- sum(diag(mc)) / sum(mc) * 100
percent

##### Model 2 ####

mod_fit2 <- train(discuss.supervisor~organization.size.large + mental.health.coverage +
prev.anonymity.protected + mental.health.options,  data=training, method="multinom",
            trControl = ctrl, tuneLength = 5)


pred2 = predict(mod_fit2, newdata=testing)
cm2 <- confusionMatrix(data=pred2, testing$discuss.supervisor)
cm2
mc2 <- cm2$table
percent$value[2] <- sum(diag(mc2)) / sum(mc2) * 100
percent
```

Random Forest Classification:

```
library('caret')
library('dplyr')
library(AUC)
library(caTools)
library(randomForest)


##### Load the clean dataset #####

mdat4 <- read.csv('MentalHealthCleanedDataset.csv')   # Reading the dataset

mdat4 %>%
select(age,gender,work.country,work.us.state,leave.sanction,discuss.supervisor,discuss.coworker,employ
ee.count,mental.health.options,current.mental.disorder,self.employed,company.resources,mental.health
.coverage,prev.anonymity.protected,family.history) %>% str()

#######################################

mdat4$discuss.supervisor <- relevel(mdat4$discuss.supervisor, ref="No")

mdat4$prev.anonymity.protected <- relevel(mdat4$prev.anonymity.protected, ref="No")
```

```
mdat4$mental.health.coverage <- relevel(mdat4$mental.health.coverage, ref="No")

levels(mdat4$organization.size.large)

# RANDOM FOREST -starts

set.seed(123)
split = sample.split(mdat4$discuss.supervisor, SplitRatio = 0.7)
train = subset(mdat4, split==TRUE)
test = subset(mdat4, split==FALSE)

train$discuss.supervisor<-factor(train$discuss.supervisor)
test$discuss.supervisor<-factor(test$discuss.supervisor)


mod.forest2 <- randomForest(discuss.supervisor ~ mental.health.coverage +mental.health.options +
prev.anonymity.protected + age + organization.size.large, data=train, importance=TRUE)

tree.pred2 = predict(mod.forest2, newdata=test)
tab3 <- table(tree.pred2, test$discuss.supervisor)
(tab3[1,1]+tab3[2,2])/(tab3[1,1]+tab3[1,2]+tab3[2,1]+tab3[2,2])
# 60.09
mod.forest2

# Variable Importance
varImpPlot(mod.forest2,
        sort = T,
        n.var=5,
        main="Top 5 - Variable Importance")



mod.forest3 <- randomForest(discuss.supervisor ~ mental.health.coverage  + prev.anonymity.protected +
mental.health.options,data=train)

tree.pred3 = predict(mod.forest3, newdata=test)
tab4 <- table(tree.pred3, test$discuss.supervisor)
(tab4[1,1]+tab4[2,2])/(tab4[1,1]+tab4[1,2]+tab4[2,1]+tab4[2,2])
# 61.57
mod.forest3

# K-fold
ctrl <- trainControl(method = "repeatedcv", number = 10, savePredictions = TRUE)
seed<-7
mod_fit <- train(discuss.supervisor~ mental.health.coverage + mental.health.options +
prev.anonymity.protected,  data=train, method="rf",
            trControl = ctrl, tuneLength = 5)
```

```
pred = predict(mod_fit, newdata=test)
confusionMatrix(data=pred, test$discuss.supervisor)
#58.13%
# RANDOM FOREST - ends
```

Comparison of Models:

```
# Comparison of models
percent <- data.frame(methods=c("Multinomial Regression - Reduced Model", "Multinomial Regression -
Full Model","Random Forest - Reduced Model","Random Forest - Full Model"), value=c(0,0,0,0))
percent$value[1] <- 50.00
percent$value[2] <- 49.01
percent$value[3] <- 41.45
percent$value[4] <- 39.14
#plotting accurcy of all models
percent$methods <- paste(percent$methods, " - " , round(percent$value,digits = 2) , "%" , sep = "")
percent1 <- data.frame(methods = percent$methods, value = percent$value)
percent2 <- rbind(percent1,data.frame(value=30, methods=percent1$methods))
ggplot() +
  geom_point(data = percent, aes(x = value, y = methods, col=c("blue","red","blue","red")), size = 4) +
  geom_path(data = percent2, aes(x = value, y = methods,
col=c("blue","red","blue","red","blue","red","blue","red")), size = 2) +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 0, vjust = 0.5, hjust = 0.5)) +
  labs(
    x = "Percentage Accuracy",
    y = "Models",
    title = "Percentage success of the models"
  )
```