FOSTER SCHOOL OF BUSINESS

UNIVERSITY of WASHINGTON

MSIS 510 - Introduction to Data Mining and Analytics (Group Project)

Descriptive and Predictive Analysis on Mental Health at Work Dataset

MSIS Purple Team 9

Ashima Dogra

Harkirat Kaur

Grace Zhou

Saswati Prusty

Jagan Reddy

Weiming Luo

Project Goal

As a Mental Health consulting firm, our goal is to identify factors which impact mental health treatment choices for working individuals. Based on factors (e.g., age, employment, company benefits), we predict whether an individual will seek treatment or not? How much does each factor correlate to treatment choice? This will help us design our services as per the needs of individuals and help our clients make better choices.

Description of Dataset

For our analysis, we picked the Mental Health in Tech Survey dataset from Kaggle. The data was collected in a global survey conducted by Open Sourcing Mental Illness.
Mental Health in Tech Survey | Kaggle

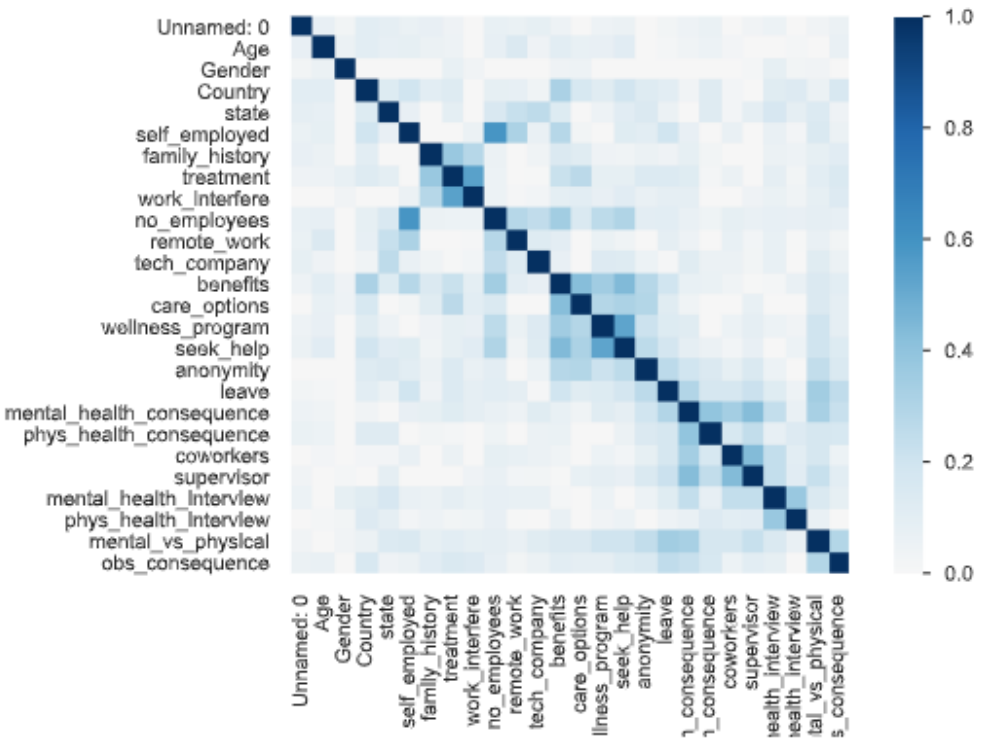
The original dataset has 1259 observations consisting of 27 variables. Below are the variables and their description:

Name	Type	Description
Timestamp	date_time	Time the survey was submitted
Age	int	Respondent age
Gender	factor	Respondent gender
Country	factor	Respondent country
state	factor	If you live in the United States, which state or territory do you live in?
self_employed	factor	Are you self-employed?
family_history	factor	Do you have a family history of mental illness?
treatment	factor	Have you sought treatment for a mental health condition?
work_interfere	factor	If you have a mental health condition, do you feel that it interferes with your work?
no_employees	range	How many employees does your company or organization have?
remote_work	factor	Do you work remotely (outside of an office) at least 50% of the time?
tech_company	factor	Is your employer primarily a tech company/organization?

benefits	factor	Does your employer provide mental health benefits?
care_options	factor	Do you know the options for mental health care your employer provides?
wellness_program	factor	Has your employer ever discussed mental health as part of an employee wellness program?
seek_help	factor	Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity	factor	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
leave	factor	How easy is it for you to take medical leave for a mental health condition?
mentalhealthconsequence	factor	Do you think that discussing a mental health issue with your employer would have negative consequences?
physhealthconsequence	factor	Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers	factor	Would you be willing to discuss a mental health issue with your coworkers?
Supervisor	factor	Would you be willing to discuss a mental health issue with your direct supervisor(s)?
physhealthinterview	factor	Would you bring up a physical health issue with a potential employer in an interview?
mentalhealthinterview	factor	Would you bring up a mental health issue with a potential employer in an interview?
mentalvsphysical	factor	Do you feel that your employer takes mental health as seriously as physical health?
obs_consequence	factor	Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
comments	strings	Any additional notes or comments

The chart below shows correlation between various parameters.

Correlation chart:



For our analysis, we are considering **TREATMENT** as target variable.

Data Preparation Details

Data cleaning -

1. Removed any outliers in age by using 16-100 as upper and lower age limits. Further reduced cardinality by binning age among different age groups.
2. Reformatted and categorized gender into three categories. Removed spelling errors, short forms (like F for female).
3. Deleted any missing values/rows in the self-employed column.
4. Binned the no. of employees within the organizations to distinct groups.
5. For logistic regression model, to make better predictions, country and state were removed. These are very high cardinality parameters.
6. Any NAs in comments were removed as well to process text.

Data Visualization and Exploratory Analysis

Focus for Exploratory Analysis:

The analysis focused on finding below information based on survey from different respondents:

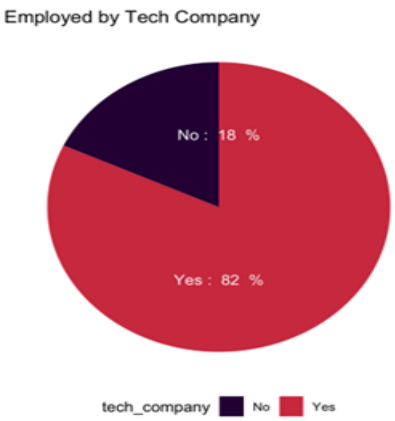
- Attitude of the employees towards issues related to mental health.

- Employer's readiness for employees regarding mental health issues.
- Differences in attitude between physical health vs. mental health treatment.

1. Observations based on Industry for the survey:

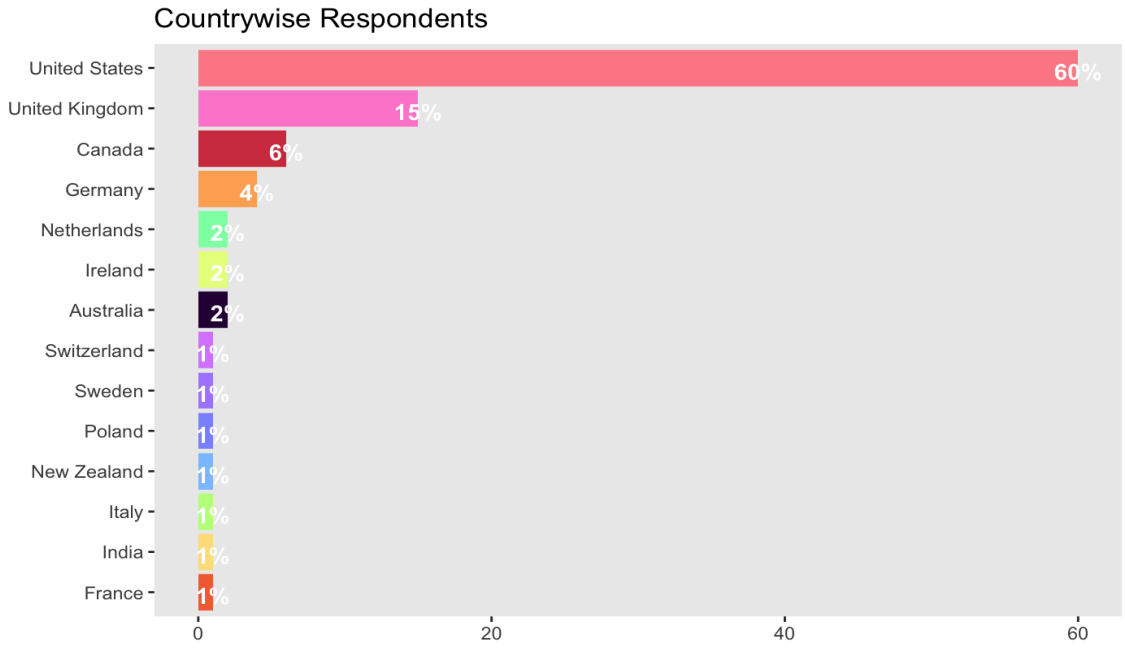
The data visualization has been done majorly for the employees working in Tech Industry during the year 2014.

The pie chart depicts that 82% of the respondents were from the Tech company while other were from non-tech companies.

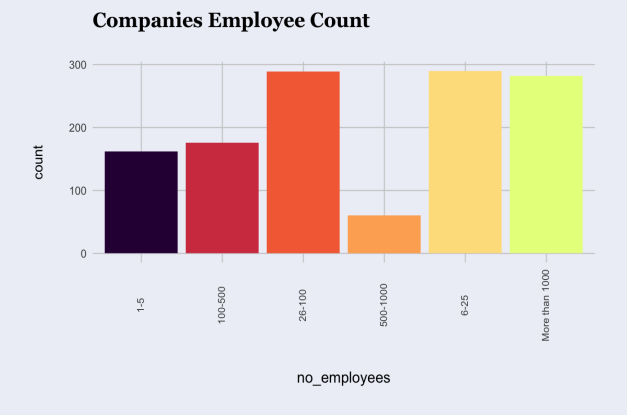


2. Observations based on Geographic locations -

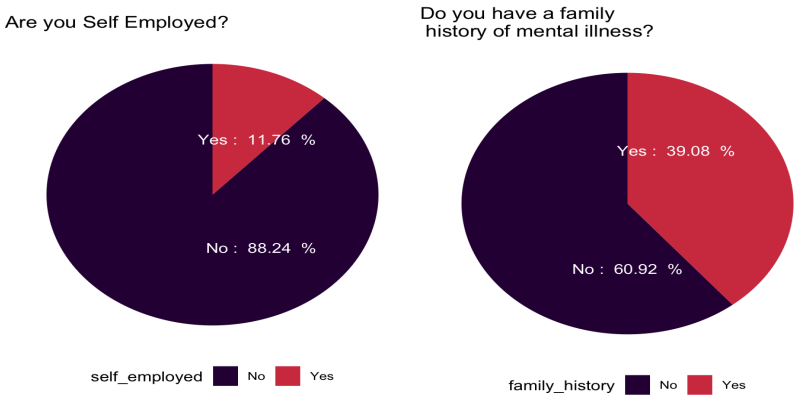
The visualization shows country wise respondents' information, the survey done has maximum number of respondents from United States, followed by United Kingdom, Canada with least respondents being from France.



3. **Observations based on Age Range and Company Size -**
   The company size on which the survey has been conducted is highlighted in the chart, it shows most companies having the number of employees in the range from '26-100', followed by '6-25' and 'more than 1000'.
   Age distribution of the employees is show from the plot below - age range is between 20-40 years for the maximum correspondents.



4. **Observations based on Self-Employed and having a family history -**
   Most of the respondents are working for a company. Also, many respondents do not have a family history of mental illness, however those who seek treatment are significantly having a family history of mental illness.

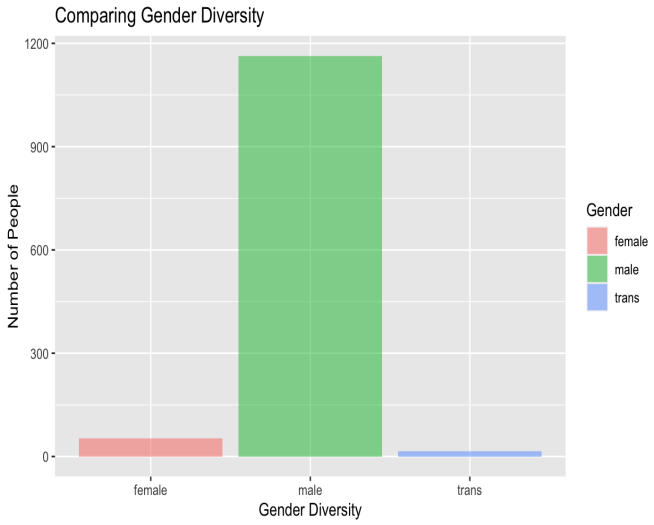


5. Observations Based on Gender Diversity

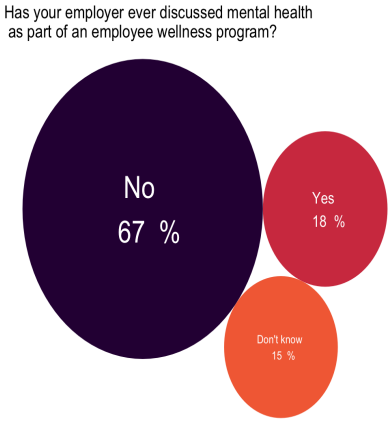
It can be seen from the Gender diversity bar graph that there are more Male respondents in the survey, followed by females and then Trans gender. It is possible that since there are a wider number of males working in the tech industry, there is such a discrepancy in gender diversity.

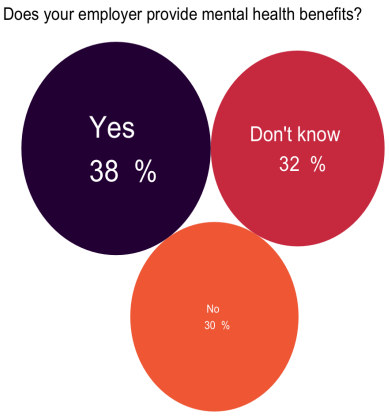
6. Employer readiness for Employee Mental Health Treatment -

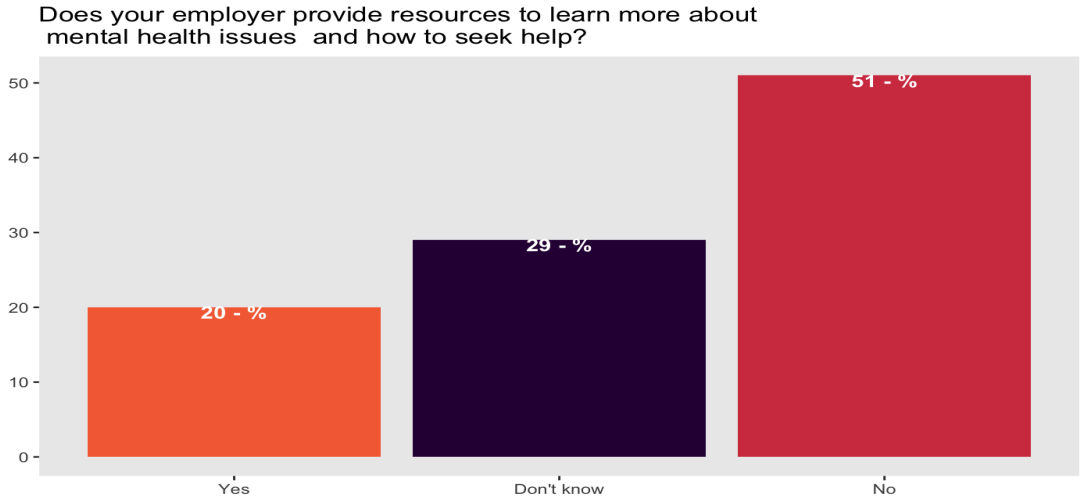
The below shows respondents responding to the Employer's readiness for mental health of employees. Respondents said 'Yes' (38%) when asked if Employer provides mental health benefits, while most of the respondents either said 'No' or 'Don't know' which shows mental health benefits options are still not explored by employees. It can also be seen that respondents confirmed saying 'No' when asked if the employer ever discussed mental health as part of the Employee wellness program.



This shows Employers are still not fully ready when it comes to employee mental health. The bar graph below shows 51% employees saying 'No' when asked there were enough resources provided by their employer for mental health issues.

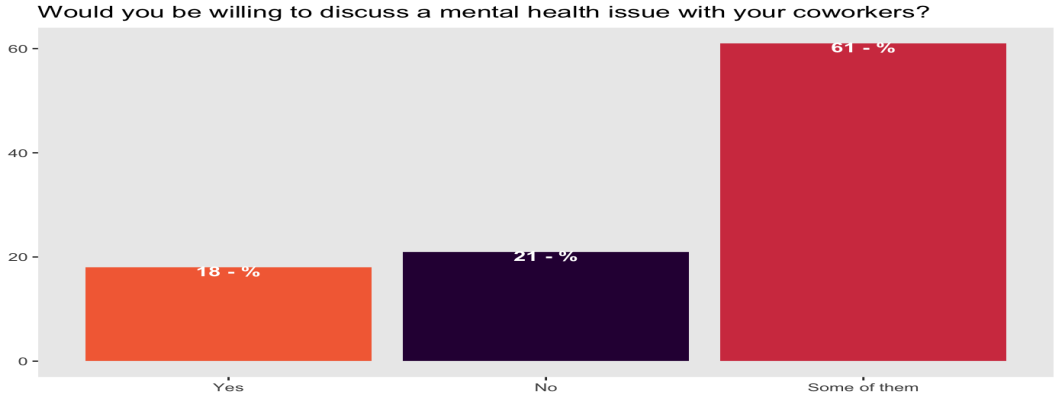


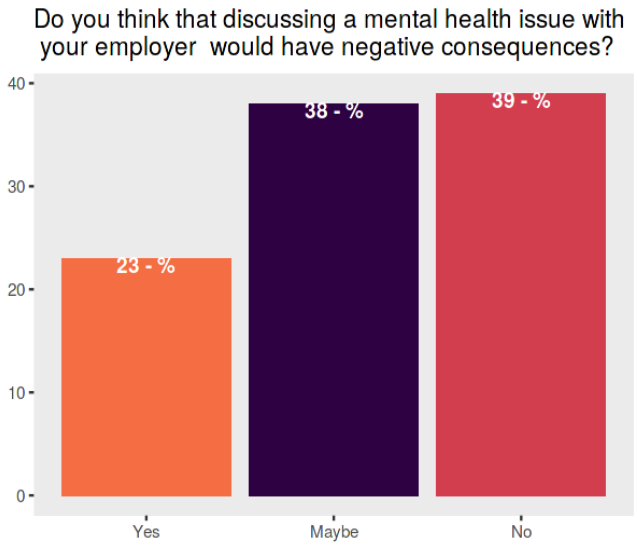


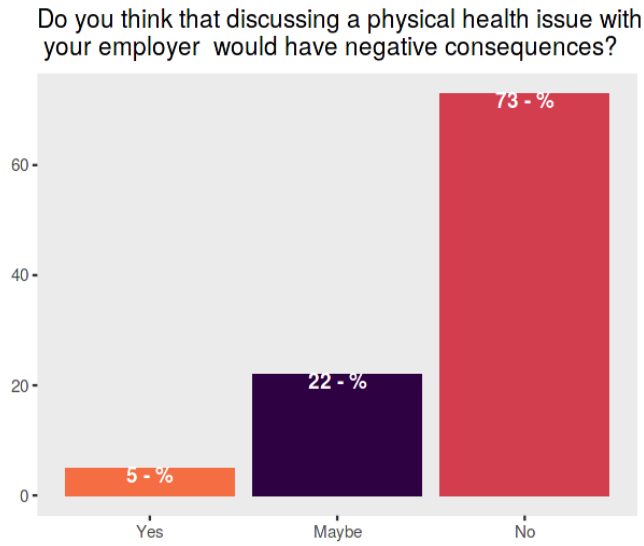


7. Difference in attitude towards mental health vs. physical health:

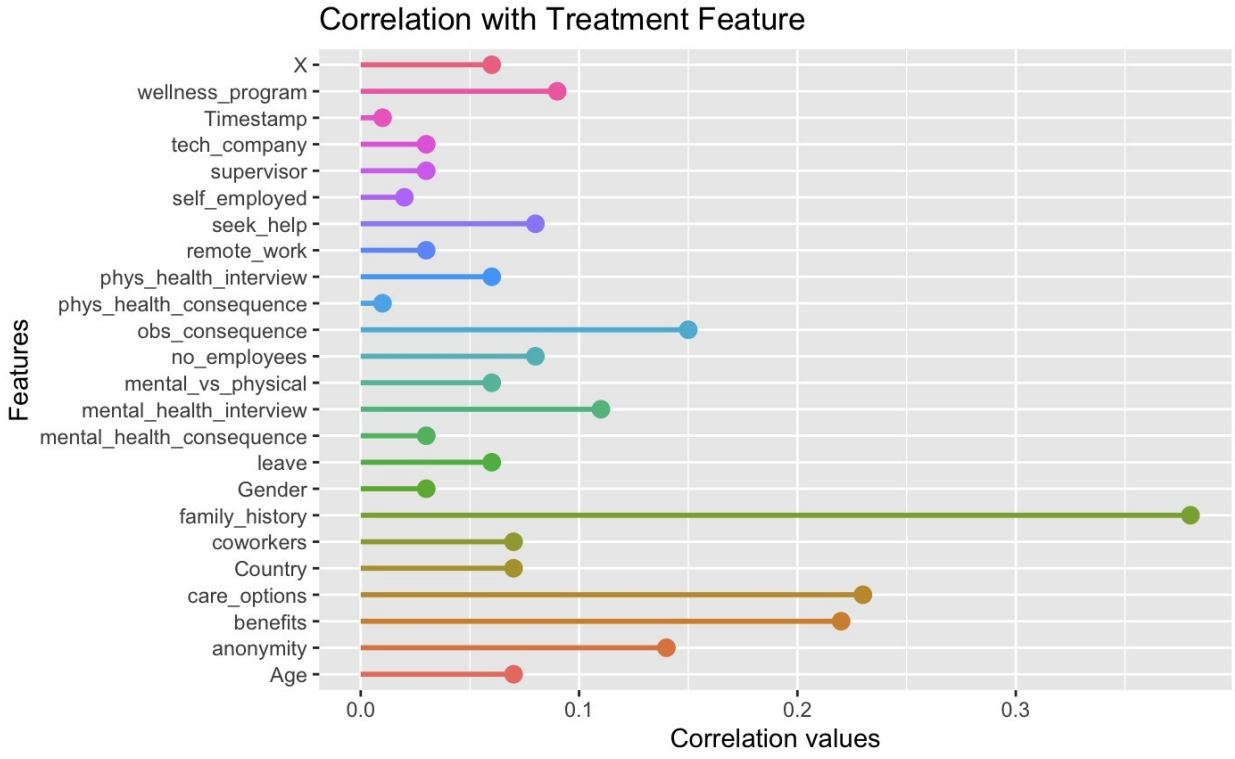
It can be seen clearly that employees are still hesitant about discussing mental health issues with employer and coworkers thinking that it will have negative consequence. 73% respondents are willing to discuss physical health issue over mental health issue.







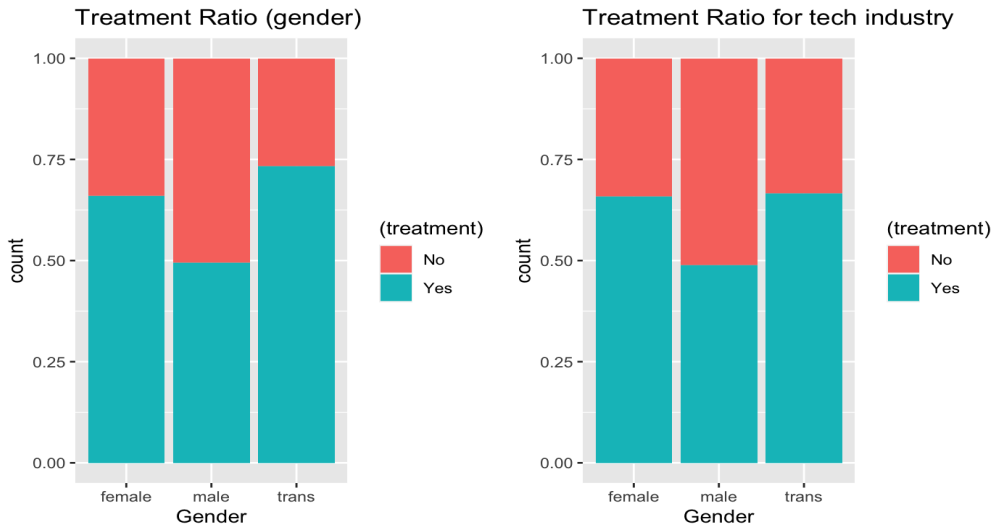
8. Correlation between different factors for treatment



9. Mental Health Treatment based on Gender

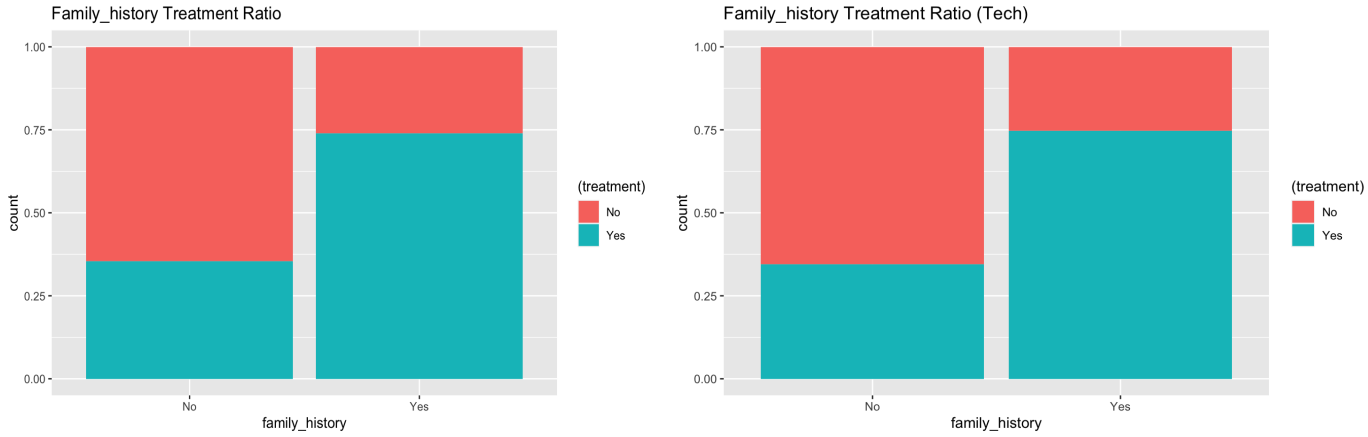
It is visible from the comparative graphs below; females are most likely to seek mental health treatment. Trans Gender are more likely to seek treatment compared to males. The

reason for such a difference could be a general view that males are less likely to call out the mental health issues compared to females and trans gender.



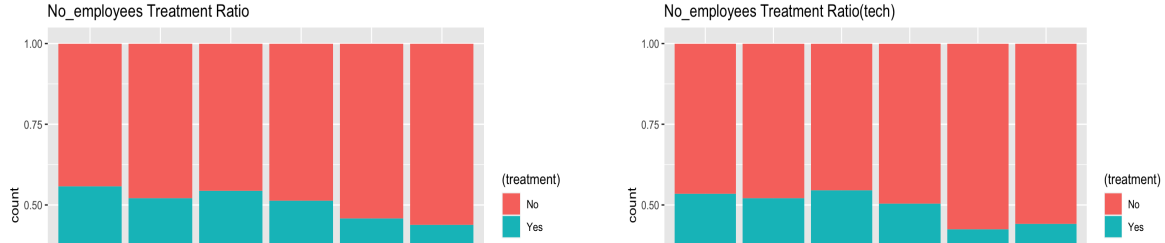
10. Mental Health Treatment based on Family history

It can be seen that respondents that have a family history of treatment are more likely to get treatment compared to those who do not have a family history of mental health.



11. Mental Health Treatment based on Number of Employees

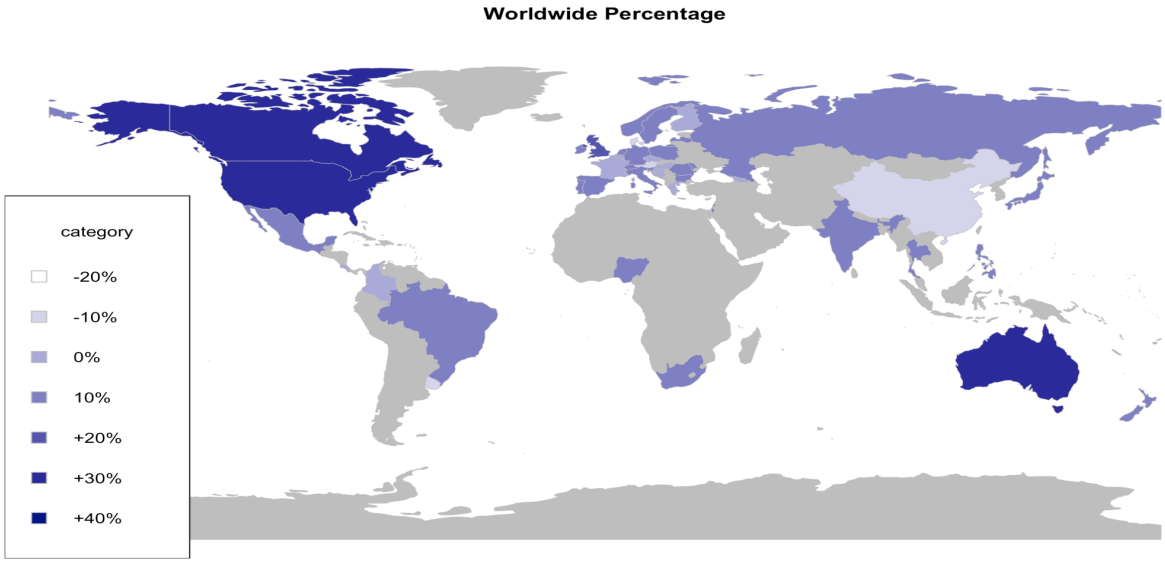
From the observations, it is visible that more the number of employees, they are less likely to seek mental health treatment. There could be various reasons, like employer's readiness towards providing benefits, more the employees there could less focus on such benefits.



12. Word Cloud from Survey comments

The word cloud brings out the essence of mental health survey comments from employees, showing the attitude of depression, anxiety when mental health is discussed in the workspace.

13.Country wise distribution of mental health treatment choice:



According to the map generated from the survey through R, in most countries, the percentage of respondents seek for mental health treatment is relatively equal or slightly higher than the number of respondents did not seek for mental health treatment. For some of the countries the percentage who have sought treatments are lower than the respondents who did, for example: China, Uruguay, Austria. People in France and Finland were surprisingly among the countries who are less likely to seek mental health treatment even though they are commonly considered to be part of the developed western culture.

In contrast, respondents in Australia, Canada, and the United States who seek treatments at least 40% more than the non-treatments.

Data Modeling and Methodology

Logistic Regression:

Based on the correlation derived during data cleaning, only highly correlated predictors are used for creating this model, which can predict the treatment better.

Initially we included more predictors including supervisor, coworker, leave, mentalhealth_consequences and we ran at a test to train ratio of 70:30. In this case our model accuracy was much lower at 63%. After that we decided to omit some less correlated and high cardinality parameters and used a test train split of 80:20.

Type of questions answered by this model:

Few of the questions that can be answered with this model are:

1. How does age group and gender impact mental health treatment choices?
2. How much impact does family history have on employee mental health conditions?
3. How does resources provided by employer's impact employees attitude towards seeking mental health treatment?

Results of model:

Fitting the logistic model to full data gives following results:

```
Call:
glm(formula = treatment ~ Age + Gender + no_employees + work_interfere +
    family_history + self_employed + benefits + care_options +
    mental_health_interview + seek_help + anonymity, family = "binomial",
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4039  -0.4551   0.3933   0.6693   2.6934

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.11639    1.06032  -0.110   0.9126
Age<20                       0.03639    0.67751   0.054   0.9572
Age>60                       2.07238    1.39249   1.488   0.1367
Age31-40                    -0.01010    0.21831  -0.046   0.9631
Age41-50                     0.34607    0.34703   0.997   0.3187
Age51-60                     1.53555    0.87884   1.747   0.0806 .
Gendermale                  -0.08137    0.44976  -0.181   0.8564
Gendertrans                 -0.06789    0.97471  -0.070   0.9445
no_employees> 1000           0.20376    0.41582   0.490   0.6241
no_employees100 - 500        0.59518    0.43425   1.371   0.1705
no_employees26 - 100         0.30967    0.38077   0.813   0.4161
no_employees500 - 1000       0.67876    0.63106   1.076   0.2821
no_employees6-25             0.13459    0.36083   0.373   0.7091
work_interfereOften          3.82608    0.40079   9.546  < 2e-16 ***
work_interfereRarely         2.58873    0.33179   7.802 6.08e-15 ***
work_interfereSometimes      3.19837    0.30335  10.543  < 2e-16 ***
family_historyNo            -1.10565    0.20618  -5.363 8.20e-08 ***
self_employedYes             0.12812    0.34532   0.371   0.7106
benefitsDon't know          -0.20225    0.28903  -0.700   0.4841
benefitsYes                  0.56620    0.31545   1.795   0.0727 .
care_optionsNot sure        -0.30032    0.26345  -1.140   0.2543
care_optionsYes              0.61681    0.27068   2.279   0.0227 *
mental_health_interviewMaybe -1.86445   0.82445  -2.261   0.0237 *
mental_health_interviewNo    -1.71342   0.78993  -2.169   0.0301 *
seek_helpDon't know          0.83376    0.32989   2.527   0.0115 *
seek_helpNo                  0.36528    0.30991   1.179   0.2385
anonymityDon't know         -0.49895    0.25529  -1.954   0.0506 .
anonymityNo                 -0.56642    0.47721  -1.187   0.2353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1008.65  on 776  degrees of freedom
Residual deviance:  671.99  on 749  degrees of freedom
  (223 observations deleted due to missingness)
AIC: 727.99

Number of Fisher Scoring iterations: 5
```

Based on this we can see that:

1. Work interference, family history and employers' mental health benefits/options have a great impact on the target variable.
2. P value is less than 0.05 which indicates usefulness of model in making correct predictions.

Evaluation of model:

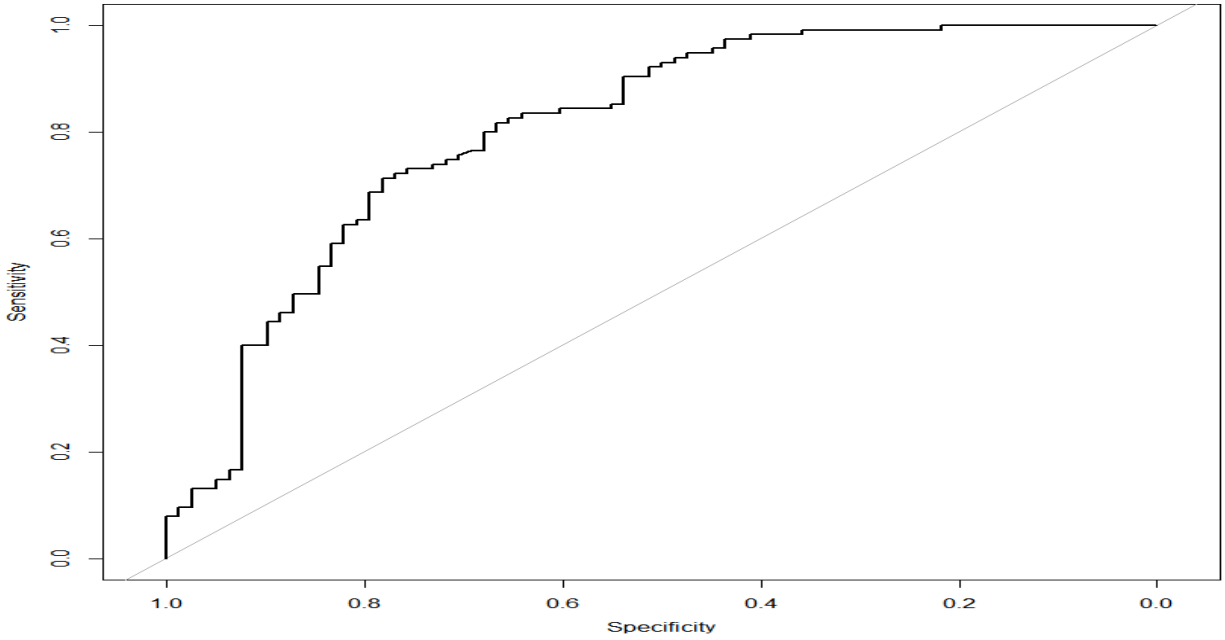
Confusion Matrix

p = 0.5		Actual class	
		0	1
Predicted	0	41	11
	1	37	104

Predicting Threshold

threshold	specificity	sensitivity
0.9	0.92	0.33
0.747	0.78	0.71
0.5	0.52	0.9

ROC curve :



Accuracy for this model is 75%.

Insights:

- Males are less likely to have sought treatment. This could be attributed to the fact that they do not need treatment or are reluctant to seek treatment. Although as per exploratory analysis, Trans are more likely to seek treatment but in comparison to females they are less likely. This could be due to low data availability because Trans gender are at higher targets and thus seek more treatment.

- Employees with Family_history are 5 times more likely to receive mental health treatment. Family history matters a lot. This could be attributed to the fact that they are more aware about mental health treatment, or they show more symptoms and thus need to seek treatment.
- In general, if employers provide resources, benefits, care options and employees anonymity this result in seeking more mental health care which is a positive sign.
- For employees whose mental health interferes with their work, they are more likely to seek treatment. This indicates a positive sign of employee's attitude towards mental health.

Classification Tree

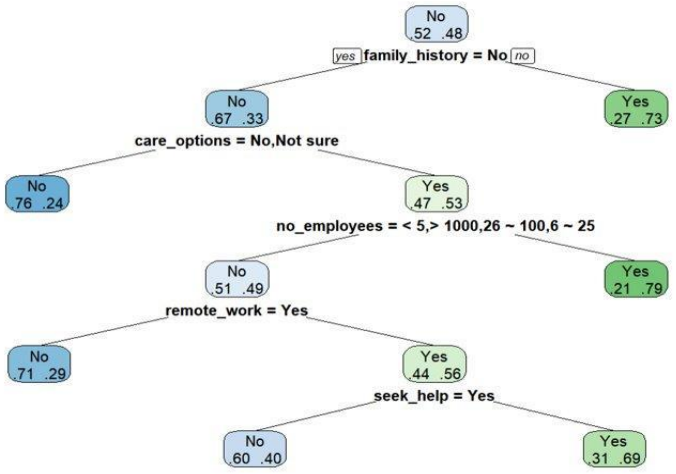
We used a classification tree to find the factors that can influence an employee's decision in seeking mental treatment. On our first attempt, the predictors include all variables except treatment and country. Prediction accuracy of the classification tree is 68.02%. In our second

attempt, we removed physhealthinterview, mentalhealthinterview and mentalvsphysical data, and the prediction accuracy increased to 70.54%.
First attempt:

Accuracy: 0.6802

Sensitivity: 0.6807

Specificity: 0.7031

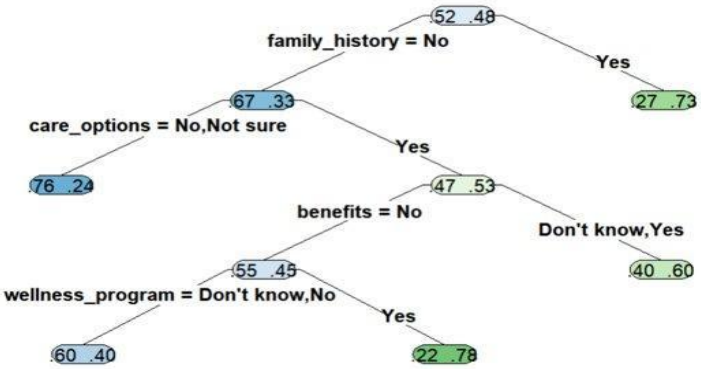


Second attempt:

Accuracy: 0.6923

Sensitivity: 0.6807

Specificity: 0.7031



Prediction Using Classification Tree
From the final classification tree, we found the following rules to predict whether an employee would seek treatment:

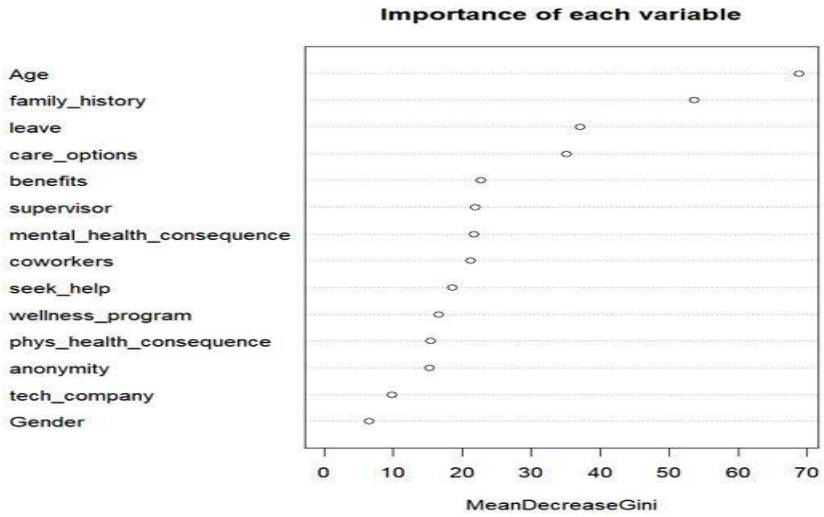
- **IF** family_history = Yes **THEN** treatment = Yes
- **IF** family_history = No **AND** care_options = Yes **AND** benefits = Don't know/Yes **THEN** treatment = Yes
- **IF** family_history = No **AND** care_options = Yes **AND** benefits = No **AND** wellness_program = Yes **THEN** treatment = Yes
- **IF** family_history = No **AND** care_options = No/Not sure **THEN** treatment = 0
- **IF** family_history = No **AND** care_options = Yes **AND** benefits = No **AND** wellness_program = Don't know/No **THEN** treatment = No

Insights:

- On one hand, the possible reason for employees with a family history is that they would be more aware of potential mental illness, thus seeking treatment actively.
- On the other hand, an employee is more likely to seek mental treatment if the company provides mental health benefits, or the employee knows the options for mental health care, or the employer discusses mental health as part of an employee wellness program.
- The advice for employers in the tech industry is to develop mental health programs and offer benefits. According to existing studies, participation in the mental health program for employees has increased awareness of mental health among themselves and individuals who they interact with daily. Moreover, they gained the ability to identify signals of someone who is potentially struggling and teach them the skills to seek for necessary resources.

Random Forest

We used Random Forest Model to confirm the findings in the classification tree model. Random forest model bagging so that it can average out the variances of the bootstrapped model. It can correct the decision trees' habit of overfitting.
Below is the graph of important of each variable in random forest model:



The prediction using the random forest model has an accuracy of 0.7126. The sensitivity is 0.6891 and the specificity is 0.7344.

Insights:

- Apart from family history, care options, benefits and wellness programs, there are some other factors that can contribute to an employee's action of seeking mental treatment.
- Age could be an important factor, and the detailed impact has been illustrated in the logistic regression model. How easy it is for employees to take medical leave for a mental health condition is also one main factor. If it is easier, employees might be more likely to seek treatment. In addition, the communication between employees and their coworkers/supervisors could influence the employees' decisions.
- In order to create a healthy and productive work environment that reduces the stigma associated with mental illness, companies should make it easier for employees to seek treatment and encourage communication between employees as well. Ultimately, it's beneficial for all parties which increases engagement and promotes an environment of inclusion and support.

Comparison between different models:

	Logistic Regression	CART	Random Forest
Accuracy	75%	69%	71%
Sensitivity	0.71	0.68	0.69
Specificity	0.78	0.70	0.73

Logistic regression is the best model to predict results for this dataset based on its high accuracy. It helps avoid any overfitting issues since it works on probability rather than a discrete yes or no as in random forest and CART.

Conclusion

Mental health treatment is as crucial as physical treatment. This datasets helps us assess the mental health treatment from both employees and employers' perspective. As a Mental Heath consulting firm, we can make use of the insights derived from this report to design better mental health treatment options for our clients. Other employers can also benefit from these insights to make healthy choices for their employees. Four priority areas are identified for focused attention to diminish the mental health treatment gap and to improve access to high-quality mental health services globally: diminishing pervasive stigma, improving mental health treatment resources and research capacity, implementing prevention programs to decrease the incidence of mental

disorders, and establishing sustainable scale up of public health systems to improve access to mental health treatment using evidence-based interventions.
If companies make mental health services more accessible and intervene in the workplace in ways that improve well-being, they will simultaneously make investments that will provide meaningful improvements in employee outcomes and consequently in company productivity.
As we did deeper into the data during our analysis, here are some improvement suggestions for the dataset:

1. Country - although it's a global data since multiple countries present but the final data majorly represents the US. If services are limited to customers only in USA, this data is good otherwise more data to be included for other countries.
2. High discrepancy in gender data - most of the survey takers were males.
3. Data is impartial for self-employed customers since not many data points are available.

Beyond implementation of a well-rounded system within companies, working with families and local social networks. Companies could have an open dialogue teams engage with people in crisis situations to provide support and stimulate dialogue regarding treatment options outside of the workplace.