Mental Health Analysis in Tech Industries

Pallavi Chandrashekar (A20427289)
Priyanka Mutha (A20450968)
Mayank Phadke (A20456460)

CSP 571 - Data Preparation and Analysis

Prof. Adam McElhinney

May 3rd, 2020

Introduction:

Mental health problems are one of the biggest hurdles faced by the general population in this day and age. In 2016, out of the adult US population, about 18.3 % (44.7 million) people reported having issues with their mental health, reporting having symptoms synonymous with stress such frequent headaches, elevated heartbeat, or feeling overwhelmed or anxious over an irregular period of time. Treating mental health issues along with physical health ones can amp up the cost by 2 to 3 times

Around 63% of the US population is part of the labor force. The workplace can be a common and important location to work upon, to improve the well-being of the adults. Conducting wellness programs can help identify possible causes and triggers for these issues, mainly in context to the workplace environment. Support systems can be created for the employees. By addressing and working on these issues, workplaces can reduce the health care costs for their employees and create a more fulfilling environment.

In a study conducted by British Interactive Media Association (BIMA) in the UK, it was found that in the technology industry, issues such as depression, anxiety, and a range of other issues have blown up exponentially. They have increased to such an extent that a worker in this field is just about as stressed as a healthcare employee and up to 5 times more depressed than the average worker. The companies are growing at an incredible phase, and so do the risks involved, and the average employee is finding it difficult to keep up. The workers find themselves in a continuous routine of high-risk critical decision making, and lack of work-life balance. The employer needs to put in a system to support and help their employees as they have a duty of care towards them.

The workplace can be an optimal environment for promoting discussion and treatment of such issues because:

- Communication has already developed between the supervisors and among the employees.
- Social support systems (friends) are already in place, who all have similar experiences at the workplace.
- Common policies will come from a central body, equally available for everyone.
- Incentives can be provided to encourage healthy coping mechanisms and behaviors.
- Employers can gather data easily to track progress and suggest changes as appropriate.

Problem Statement:

Mental health issues in the technology industry is a widespread problem, with increased cases of anxiety, depression, and stress that is spread among employees in a wide range of sectors.

Due to the high stakes nature of work, as well the stigma surrounding the discussion of such issues at the workplace, an employee may feel discouraged to seek help and treatment for their ailments.

Our main goal is to find out some of the main factors in the workplace that affect the mental health of tech employees and may affect their decision to reach out for help. Another part would be to predict, based on the main factors found, the probability of an employee to seek treatment. The ultimate goal is to identify the main causes and triggers to stress-induced mental health issues and to provide suggestions to reduce these issues based on the results.

Some questions that aid this analysis can be:

1. How does the size of a company relate to an employer formally discussing mental health?
2. How does the age of an employee relate to their comfort in discussing mental health issues with their peers?
3. Which parts of the world are most affected?
4. Is the company’s attitude towards the health of an employee the main factor?
5. Does the employer provide mental health benefits?
6. Are the employees comfortable talking about mental health with their coworkers?
7. What kind of job has more mental health issues?
8. Does the employer provide help for mental health issues?

Literature Review:

1. **Promoting mental health at the workplace: the prevention side of stress management. Occupational Medicine. (Elkin AJ, Rosch PJ).**

   Stress in the workplace is a growing problem that is slowly gaining recognition in the American industry. Surely, companies are starting to provide some stress management and wellness programs. This being a serious matter requires a lot of guidance and direction which is not readily available. This study found that job stress may be linked to adverse health behaviors which can lead to further major health issues. They found that workplace stress has to increase BMI and smoking in Australian workers, which could further cardiovascular diseases.

   This study may help companies to better understand the sources and trigger points of stress at the workplace. It can also help employers to assess, implement, and evaluate their stress management and wellness programs.

2. An Introduction to Logistic Regression Analysis and Reporting.

In this paper, we demonstrate that logistic regression is often a robust analytical technique to be used when the outcome variable is dichotomous. The effectiveness of the logistic model was shown to be supported by (a) significance tests of the model against the null model, (b) the important test of every predictor. The trend is obvious within the better education journals. Such popularity is often attributed to researchers' easy accessibility to classy statistical software that performs comprehensive analyses of this method. it's anticipated that the appliance of the logistic regression technique is probably going to extend. This potential expanded usage demands that researchers, editors, and readers be coached in what to expect from a commentary that uses the logistic regression technique. What tables, charts, or figures should be included? What assumptions should be verified? it's hoped that this text has answered these questions with an illustration of logistic regression applied to a knowledge set and with guidelines and proposals offered on a preferred pattern of application of logistic methods

3. Random Forests and Decision Trees

In this paper, bagging is explained. Use Bootstrap sampling to select n samples from the sample set on all attributes, establish a classifier for these n samples (CART or SVM or ...). Repeat the above two steps m times, i.e. build m classifiers (CART or SVM or ...). Put the data on these m classifiers and run, finally vote to see which category is the bottom. Fit many large trees to bootstrap resampled versions of the training data, and classify by majority vote. The following figure is the selection strategy of Bagging. Each time sampling n times from N data to get a bag of n data, a total of B times are selected to get B bags, which is B bootstrap samples.

Random forest using bagging - Use Bootstrap sampling to select n samples from the sample set and pre-create CART. On each node of the tree, randomly select k attributes from all attributes, and select an optimal segmentation attribute as a node (this is the biggest difference from Bagging). Repeat the above two steps m times, ie build m CART. These m CARTs form Random Forest. Random here means - Randomly selected subsamples in Bootstrap. The random subspace algorithm randomly selects k attributes from the attribute set. When each tree node splits, from the random k attributes, the optimal one is selected.

Data Acquisition:

The Primary Dataset:

The primary dataset is from OSMI (Open Source Mental Illness), a non-profit organization which is dedicated to raising awareness about mental health in tech communities. This dataset was published by Kaggle. This dataset contains responses of the people who are working in tech and non-tech companies for the survey conducted in the year 2014 and 2016. The 2014 survey results will be the main dataset, while the 2016 results will be used to validate the observed trends and findings.

2014 (survey.csv) dataset has 1259 responses, and contained the following attributes:

- **Age**
- **Gender**
- **Country**
- **state**: If you live in the United States, which state or territory do you live in?
- **self_employed**: Are you self-employed?
- **family_history**: Do you have a family history of mental illness?
- **treatment**: Have you sought treatment for a mental health condition?
- **work_interfere**: If you have a mental health condition, do you feel that it interferes with your work?
- **no_employees**: How many employees does your company or organization have?
- **remote_work**: Do you work remotely at least 50% of the time?
- **tech_company**: Is your employer primarily a tech company/organization?
- **benefits**: Does your employer provide mental health benefits?
- **care_options**: Do you know the options for mental health care your employer provides?
- **wellness_program**: Has your employer ever discussed mental health as part of an employee wellness program?
- **seek_help**: Does your employer provide resources to learn more about mental health issues and how to seek help?
- **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
- **leave**: How easy is it for you to take medical leave for a mental health condition?
- **mental_health_consequence**: Do you think that discussing a mental health issue with your employer would have negative consequences?
- **phys_health_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
- **coworkers**: Would you be willing to discuss a mental health issue with your coworkers?

- **supervisor**: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
- **mental_health_interview**: Would you bring up a mental health issue with a potential employer in an interview?
- **phys_health_interview**: Would you bring up a physical health issue with a potential employer in an interview?
- **mental_vs_physical**: Do you feel that your employer takes mental health as seriously as physical health?
- **obs_consequence**: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- **comments**: Any additional notes or comments

The Secondary Dataset:

The secondary dataset is from the Uniform Crime Reporting (UCR). It is a program that compiles official data on crime in the United States, published by the Federal Bureau of Investigation (FBI). This dataset describes the crime statistics in the USA. It will be used to analyze whether leveraging the crime rates in the various states affects the prediction accuracy based on the year 2014. This dataset was published by www.data.world.

This dataset contains 25 tables.

- **14table-1a:** contains the crime rate comparisons of 2014 and 2005,2010 and 2013.
- **14table-1_2:** contains crime in the United States by volume and rate per 100000 inhabitants 1995-2014.
- **14table2:** contains crime in the United States by community type
- **14table3:** contains crime in the United States offense and population distribution by region
- **14table4:** contains crime in the United States by region geographic division and state 2013-2014
- **14table5:** contains crime in the United States by state 2014
- **14table6:** contains Crime in the United States by Metropolitan Statistical Area 2014
- **14table7:** contains offense analysis the United States 2010-2014
- **14table8:** contains city crime rates in 2014
- **14table9:** contains universities in United States crime rates in 2014
- **14table10:** contains city crime rates with County information
- **14table11:** contains offenses known to law enforcement by state by state-tribal and other agencies in 2014
- **14table12:** contains crime trends by population group 2013-2014
- **14table13:** contains crime trends by suburban and non suburban cities by population Group 2013-2014

- **14table14:** contains crime trends by Metropolitan and Nonmetropolitan Counties by Population Group 2013-2014
- **14table15:** contains crime trends additional information about selected offenses by population Group 2013-2014
- **14table16:** contains crime rates per 100000 Inhabitants
- **14table17:** contains crime rates per 100000 Inhabitants by with suburban and non-suburban information
- **14table18:** contains crime rates per 100000 Inhabitants by Metropolitan and Nonmetropolitan Counties by population.
- **14table19:** contains crime rate per 100000 inhabitants additional information about selected offenses by population group in 2014
- **14table20:** contains murder by State types of weapons.
- **14table21:** contains robbery by State types of weapons.
- **14table22:** contains aggravated assault by State types of weapons
- **14table23:** contains offense analysis number and percent change 2013-2014
- **14table24:** contains property stolen and recovered by type and value

We will be using **14table5**, which contains the following attributes:

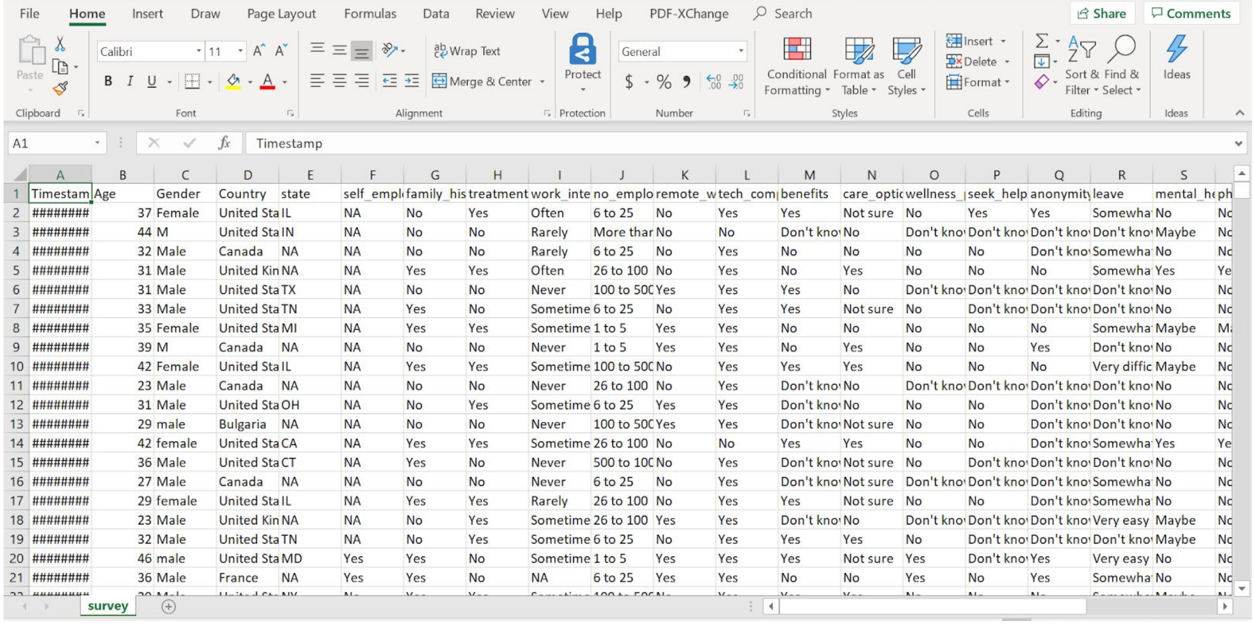
- **Timestamp**
- **State**
- **County**
- **Violent.crime:** No. of violent crimes committed.
- **Murder.and.nonnegligent.manslaughter:** No. of murders committed
- **Rape..revised.definition.1:** No. of rapes committed according to legacy definition one.
- **Rape..legacy.definition.2:** No. of rapes committed according to legacy definition two.
- **Robbery:** No. of robberies committed.
- **Aggravated.assault:** No. of aggravated assaults committed.
- **Property.crime:** No. of property crimes committed.
- **Burglary:** No. of burglaries committed.
- **Larceny..theft:** No. of larcenies committed.
- **Motor.vehicle.theft:** No. of motor thefts committed.
- **Arson3:** No. of arsons committed.

For each state, the number of crimes committed will be totaled, and all the states will be ranked according to that total number.

In both datasets, “state” is the variable is the common variable. Using this variable we will try to predict the changes in which the employee is having a mental health issue because of the environment he or she was brought up.

Data Cleaning:

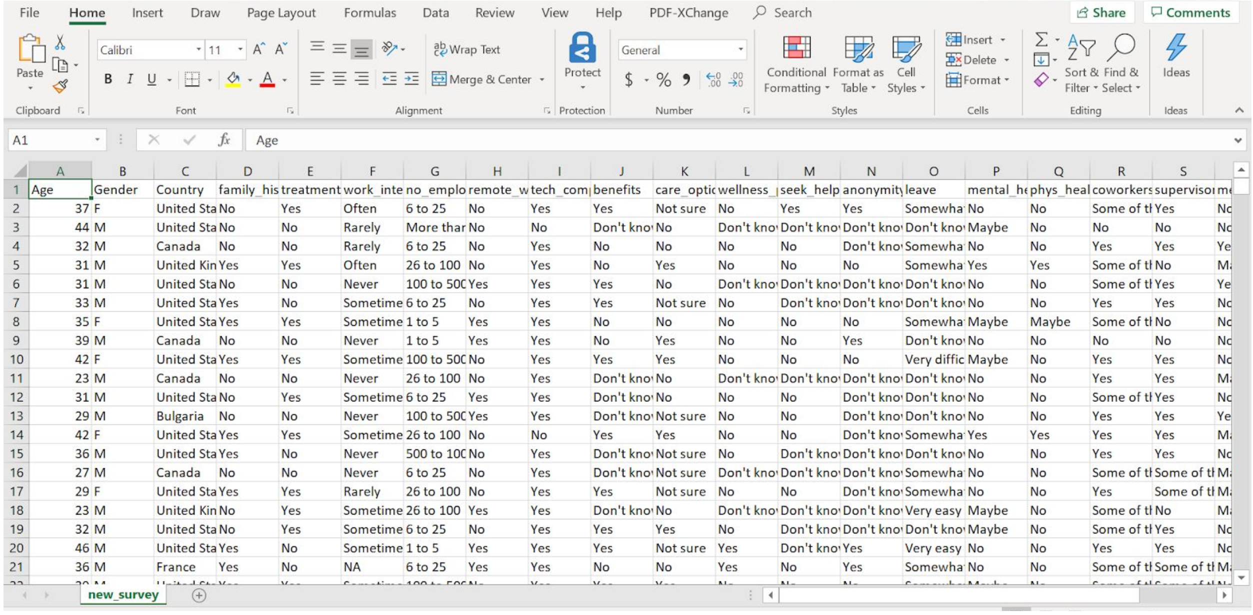
For 2014 dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Timestam	Age	Gender	Country	state	self_empl	family_his	treatment	work_inte	no_emplo	remote_w	tech_com	benefits	care_optic	wellness_	seek_help	anonymity	leave	mental_he	ph
2	########	37	Female	United Sta	IL	NA	No	Yes	Often	6 to 25	No	Yes	Yes	Not sure	No	Yes	Yes	Somewha	No	Nc
3	########	44	M	United Sta	IN	NA	No	No	Rarely	More thar	No	No	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	Maybe	Nc
4	########	32	Male	Canada	NA	NA	No	No	Rarely	6 to 25	No	Yes	No	No	No	No	Don't kno	Somewha	No	Nc
5	########	31	Male	United Kin	NA	NA	Yes	Yes	Often	26 to 100	No	Yes	No	Yes	No	No	No	Somewha	Yes	Ye
6	########	31	Male	United Sta	TX	NA	No	No	Never	100 to 50C	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Don't kno	No	Nc
7	########	33	Male	United Sta	TN	NA	Yes	No	Sometime	6 to 25	No	Yes	Yes	Not sure	No	Don't kno	Don't kno	Don't kno	No	Nc
8	########	35	Female	United Sta	MI	NA	Yes	Yes	Sometime	1 to 5	Yes	Yes	No	No	No	No	No	Somewha	Maybe	M
9	########	39	M	Canada	NA	NA	No	No	Never	1 to 5	Yes	Yes	No	Yes	No	No	Yes	Don't kno	No	Nc
10	########	42	Female	United Sta	IL	NA	Yes	Yes	Sometime	100 to 50C	No	Yes	Yes	Yes	No	No	No	Very diffic	Maybe	Nc
11	########	23	Male	Canada	NA	NA	No	No	Never	26 to 100	No	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	No	Nc
12	########	31	Male	United Sta	OH	NA	No	Yes	Sometime	6 to 25	Yes	Yes	Don't kno	No	No	No	Don't kno	Don't kno	No	Nc
13	########	29	male	Bulgaria	NA	NA	No	No	Never	100 to 50C	Yes	Yes	Don't kno	Not sure	No	No	Don't kno	Don't kno	No	Nc
14	########	42	female	United Sta	CA	NA	Yes	Yes	Sometime	26 to 100	No	No	Yes	Yes	No	No	Don't kno	Somewha	Yes	Ye
15	########	36	Male	United Sta	CT	NA	Yes	No	Never	500 to 10C	No	Yes	Don't kno	Not sure	No	Don't kno	Don't kno	Don't kno	No	Nc
16	########	27	Male	Canada	NA	NA	No	No	Never	6 to 25	No	Yes	Don't kno	Not sure	Don't kno	Don't kno	Don't kno	Somewha	No	Nc
17	########	29	female	United Sta	IL	NA	Yes	Yes	Rarely	26 to 100	No	Yes	Yes	Not sure	No	No	Don't kno	Somewha	No	Nc
18	########	23	Male	United Kin	NA	NA	No	Yes	Sometime	26 to 100	Yes	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Very easy	Maybe	Nc
19	########	32	Male	United Sta	TN	NA	No	Yes	Sometime	6 to 25	No	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Maybe	Nc
20	########	46	male	United Sta	MD	Yes	Yes	No	Sometime	1 to 5	Yes	Yes	Yes	Not sure	Yes	Don't kno	Yes	Very easy	No	Nc
21	########	36	Male	France	NA	Yes	Yes	No	NA	6 to 25	Yes	Yes	No	No	Yes	No	Yes	Somewha	No	Nc

1. Gender missing spelling: Most of the gender column was miss-spelled, like maile, make, male, Make, cis male, etc for male which we replaced with 'M', Femake, Woman, femail, woman, Female (cis), etc for female which we replaced with 'F' and genderqueer, fluid,enby, etc for others which we replaced with 'O'.
2. Converted tech_company column from binary (1 / 0) to Yes / No.
3. Age correction: few values for Age was less than 18 and greater than 65. We assumed that the ideal age to work in any industry is greater than 17 years and less than 65 years. We replaced all the outliers with the median value of age.
4. Timestamp, comments,self_employed: These columns were not providing much information so we replaced it with NULL.
5. Country: We replaced the Country values by the continent like South Africa, Zimbabwe, Nigeria as Africa.
6. State: Column had empty values and NA's we can't use a column with NA's or empty values for building models so replaced values with NULL.

After cleaning the data results are stored to a new CSV file called new_survey.csv:

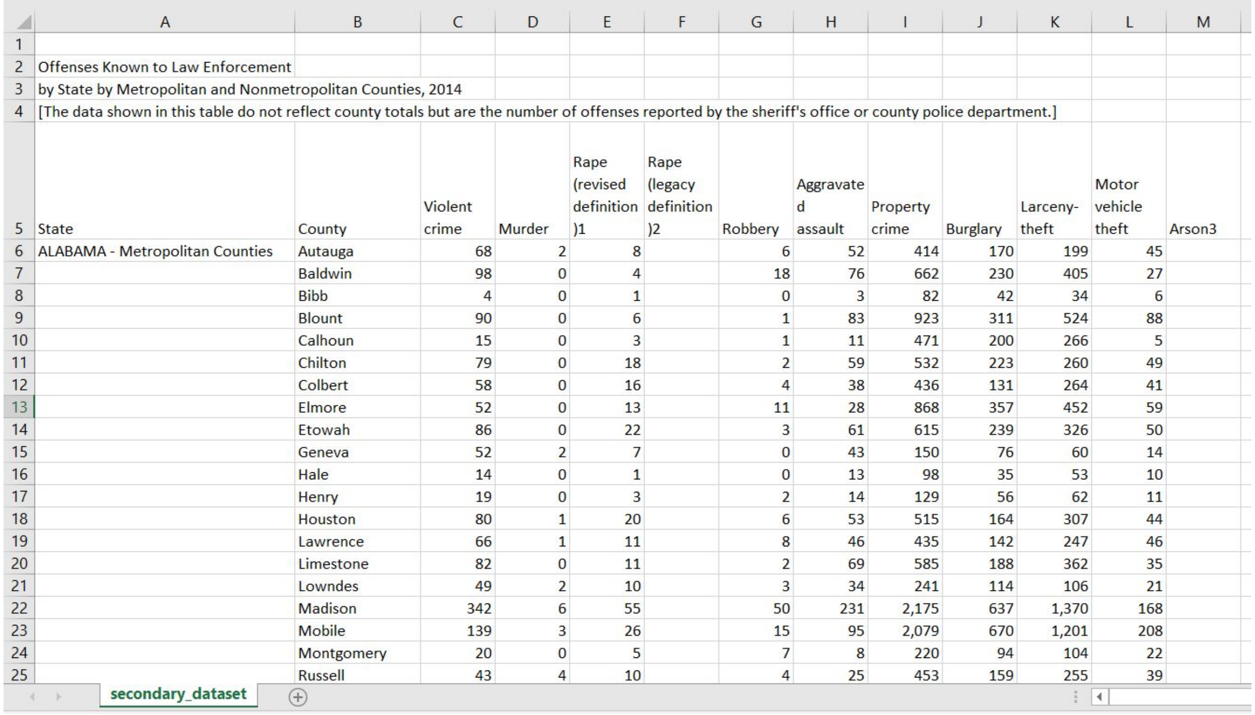


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Age	Gender	Country	family_his	treatment	work_inte	no_emplo	remote_w	tech_com	benefits	care_optic	wellness_	seek_help	anonymity	leave	mental_he	phys_heal	coworkers	supervisor	me
2	37	F	United Sta	No	Yes	Often	6 to 25	No	Yes	Yes	Not sure	No	Yes	Yes	Somewha	No	No	Some of tl	Yes	Nc
3	44	M	United Sta	No	No	Rarely	More thar	No	No	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	Maybe	No	No	No	Nc
4	32	M	Canada	No	No	Rarely	6 to 25	No	Yes	No	No	No	No	Don't kno	Somewha	No	No	Yes	Yes	Ye
5	31	M	United Kin	Yes	Yes	Often	26 to 100	No	Yes	No	Yes	No	No	No	Somewha	Yes	Yes	Some of tl	No	Ma
6	31	M	United Sta	No	No	Never	100 to 50C	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No	Some of tl	Yes	Ye
7	33	M	United Sta	Yes	No	Sometime	6 to 25	No	Yes	Yes	Not sure	No	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	Nc
8	35	F	United Sta	Yes	Yes	Sometime	1 to 5	Yes	Yes	No	No	No	No	No	Somewha	Maybe	Maybe	Some of tl	No	Nc
9	39	M	Canada	No	No	Never	1 to 5	Yes	Yes	No	Yes	No	No	Yes	Don't kno	No	No	No	No	Nc
10	42	F	United Sta	Yes	Yes	Sometime	100 to 50C	No	Yes	Yes	Yes	No	No	No	Very diffic	Maybe	No	Yes	Yes	Nc
11	23	M	Canada	No	No	Never	26 to 100	No	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	Ma
12	31	M	United Sta	No	Yes	Sometime	6 to 25	Yes	Yes	Don't kno	No	No	No	Don't kno	Don't kno	No	No	Some of tl	Yes	Nc
13	29	M	Bulgaria	No	No	Never	100 to 50C	Yes	Yes	Don't kno	Not sure	No	No	Don't kno	Don't kno	No	No	Yes	Yes	Ye
14	42	F	United Sta	Yes	Yes	Sometime	26 to 100	No	No	Yes	Yes	No	No	Don't kno	Somewha	Yes	Yes	Yes	Yes	Ma
15	36	M	United Sta	Yes	No	Never	500 to 10C	No	Yes	Don't kno	Not sure	No	Don't kno	Don't kno	Don't kno	No	No	Yes	Yes	Nc
16	27	M	Canada	No	No	Never	6 to 25	No	Yes	Don't kno	Not sure	Don't kno	Don't kno	Don't kno	Somewha	No	No	Some of tl	Some of tl	Ma
17	29	F	United Sta	Yes	Yes	Rarely	26 to 100	No	Yes	Yes	Not sure	No	No	Don't kno	Somewha	No	No	Yes	Some of tl	Ma
18	23	M	United Kin	No	Yes	Sometime	26 to 100	Yes	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Very easy	Maybe	No	Some of tl	No	Ma
19	32	M	United Sta	No	Yes	Sometime	6 to 25	No	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Maybe	No	Some of tl	Yes	Nc
20	46	M	United Sta	Yes	No	Sometime	1 to 5	Yes	Yes	Yes	Not sure	Yes	Don't kno	Yes	Very easy	No	No	Yes	Yes	Nc
21	36	M	France	Yes	No	NA	6 to 25	Yes	Yes	No	No	Yes	No	Yes	Somewha	No	No	Some of tl	Some of tl	Ma

<u>For 2016 dataset:</u>

1. Renamed all the columns to match the column names of the 2014 dataset.
2. Gender missing spelling: Most of the gender column was miss-spelled, like maile, make, male, Make, cis male, etc for male which we replaced with ‘M’, Femake, Woman, femail, woman, Female (cis), etc for female which we replaced with ‘F’ and genderqueer, fluid,enby, etc for others which we replaced with ‘O’.
3. Age correction: few values for Age was less than 18 and greater than 65. We assumed that the ideal age to work in any industry is greater than 17 years and less than 65 years. We replaced all the outliers with the median value of age.
4. Many columns were new and were not present in the dataset from 2014. Removed these columns so that they won't affect any data analysis.
5. Replaced the Country values by the continent like South Africa, Zimbabwe, Nigeria as Africa.
6. State: Column had empty values and NA’s we can’t use a column with NA’s or empty values for building models so replaced these values with NULL.
7. One of the questions from the 2014 survey was divided into two parts in the 2016 survey. So two columns from the 2016 survey (work interference while being treated / work interference while not being treated) had to be merged into one column (work interference) to match the 2014 data.

For crime statistics dataset:



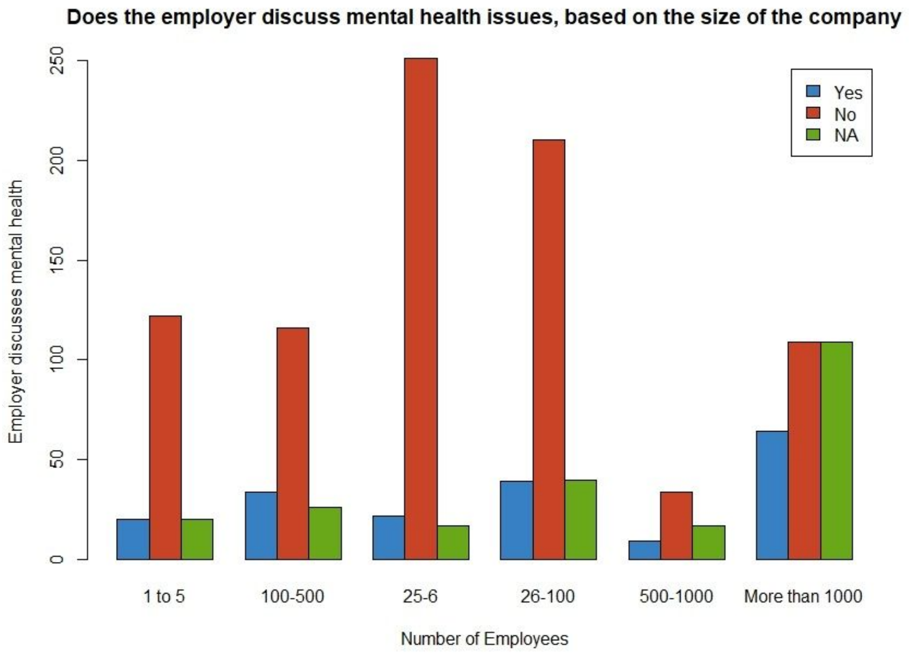
	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2	Offenses Known to Law Enforcement												
3	by State by Metropolitan and Nonmetropolitan Counties, 2014												
4	[The data shown in this table do not reflect county totals but are the number of offenses reported by the sheriff's office or county police department.]												
5	State	County	Violent crime	Murder	Rape (revised definition)1	Rape (legacy definition)2	Robbery	Aggravate d assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson3
6	ALABAMA - Metropolitan Counties	Autauga	68	2	8		6	52	414	170	199	45	
7		Baldwin	98	0	4		18	76	662	230	405	27	
8		Bibb	4	0	1		0	3	82	42	34	6	
9		Blount	90	0	6		1	83	923	311	524	88	
10		Calhoun	15	0	3		1	11	471	200	266	5	
11		Chilton	79	0	18		2	59	532	223	260	49	
12		Colbert	58	0	16		4	38	436	131	264	41	
13		Elmore	52	0	13		11	28	868	357	452	59	
14		Etowah	86	0	22		3	61	615	239	326	50	
15		Geneva	52	2	7		0	43	150	76	60	14	
16		Hale	14	0	1		0	13	98	35	53	10	
17		Henry	19	0	3		2	14	129	56	62	11	
18		Houston	80	1	20		6	53	515	164	307	44	
19		Lawrence	66	1	11		8	46	435	142	247	46	
20		Limestone	82	0	11		2	69	585	188	362	35	
21		Lowndes	49	2	10		3	34	241	114	106	21	
22		Madison	342	6	55		50	231	2,175	637	1,370	168	
23		Mobile	139	3	26		15	95	2,079	670	1,201	208	
24		Montgomery	20	0	5		7	8	220	94	104	22	
25		Russell	43	4	10		4	25	453	159	255	39	

1. State names missing: A lot of rows had their state names missing. The state value would remain empty until the counties of another state began. The name of the respective states was assigned to each county.
2. State name format: The state names were changed to their abbreviations to facilitate easy merging with the primary dataset for leveraging.
3. NA: Columns like Rape(legacy definition 2) and Arson mostly consisted of missing values. Only a minuscule amount of counties reported any crimes in that category. To replace the NA's with their mean or median would skew the data, so the NA's were replaced by zero.

Exploratory Data Analysis:

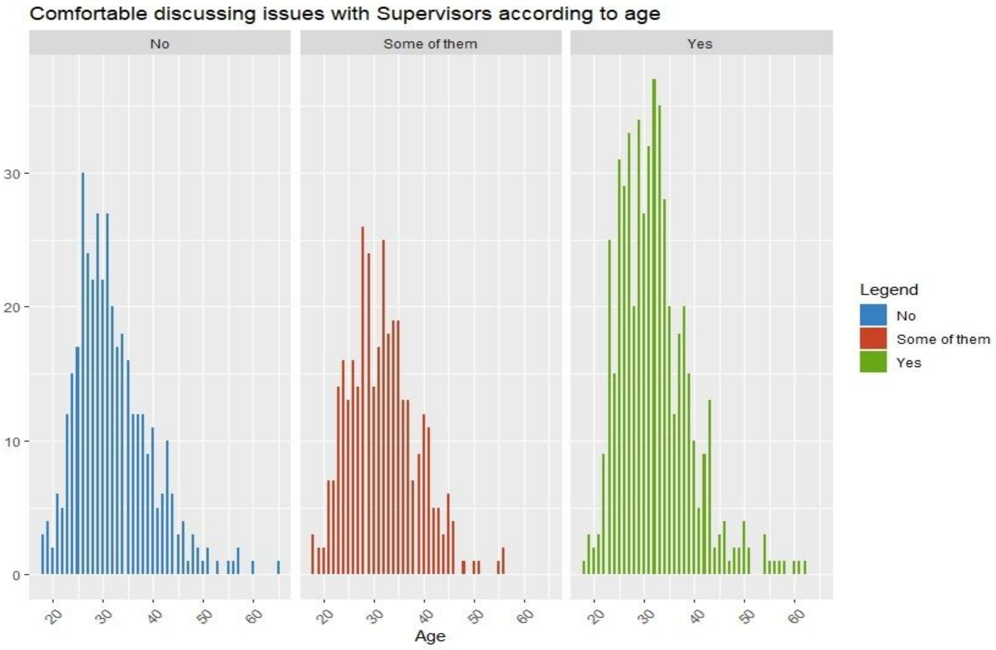
Initial exploration:

Does the size of the company affect their willingness to discuss mental health issues:



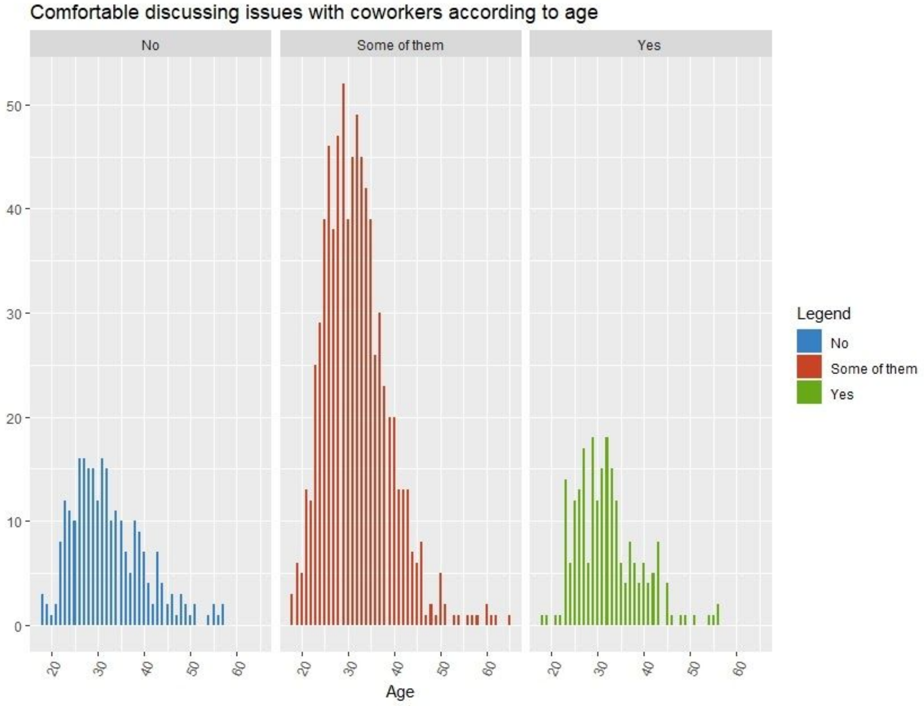
It can be observed that size does not affect the employer discussing issues. In each of the company size categories, the number of companies not discussing issues is way higher than the ones discussing.

At what age are employees more comfortable discussing issues with their supervisors?



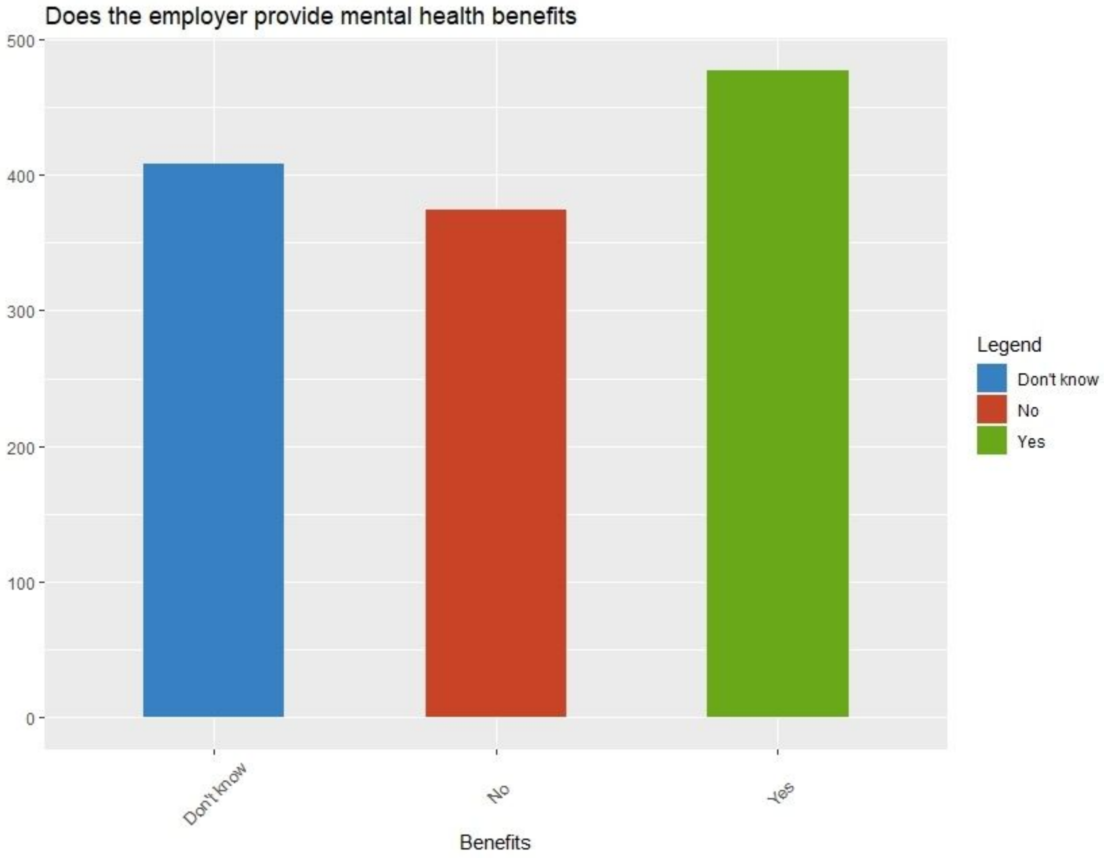
As observed, people in the 25-40 age group are more comfortable. This is probably due to knowledge and open-mindedness towards the issue.

At what age are employees more comfortable discussing issues with their coworkers?



Similarly, people in the 20-40 age group are comfortable with discussing their issues with some of their coworkers, if not all of them.

Does the employer provide health benefits for mental health issues:

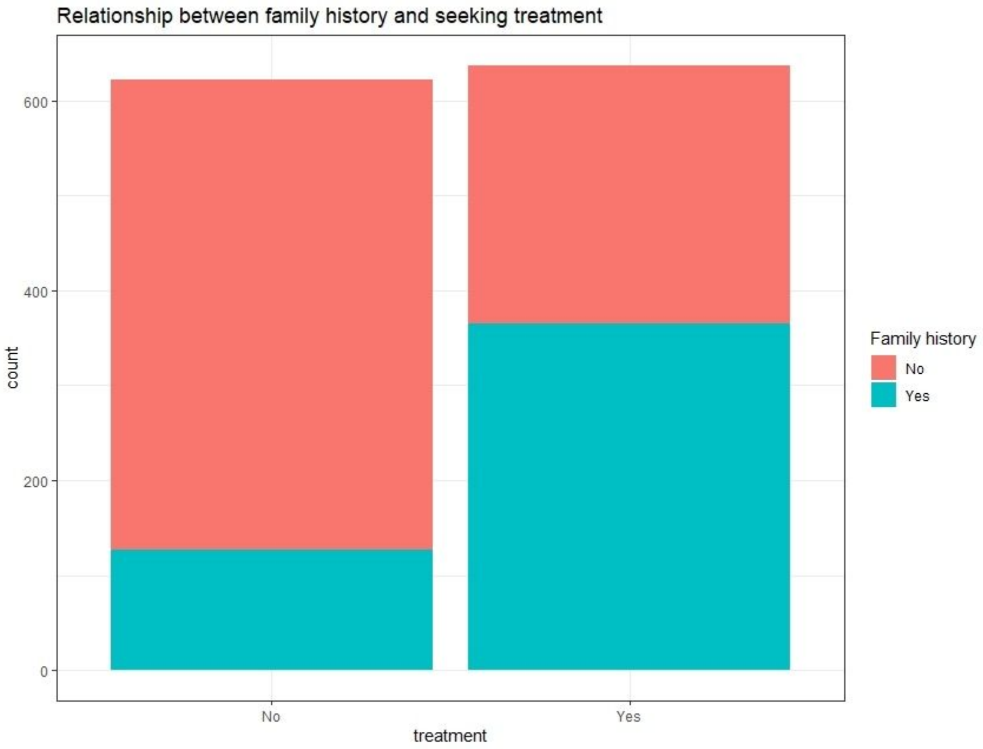


It can be observed that there is not a lot of difference between the number of companies not providing benefits and the ones providing it. But, there are a large number of employees that are not aware of whether they are provided benefits or not.

<u>Investigating the relationship with “treatment”:</u>

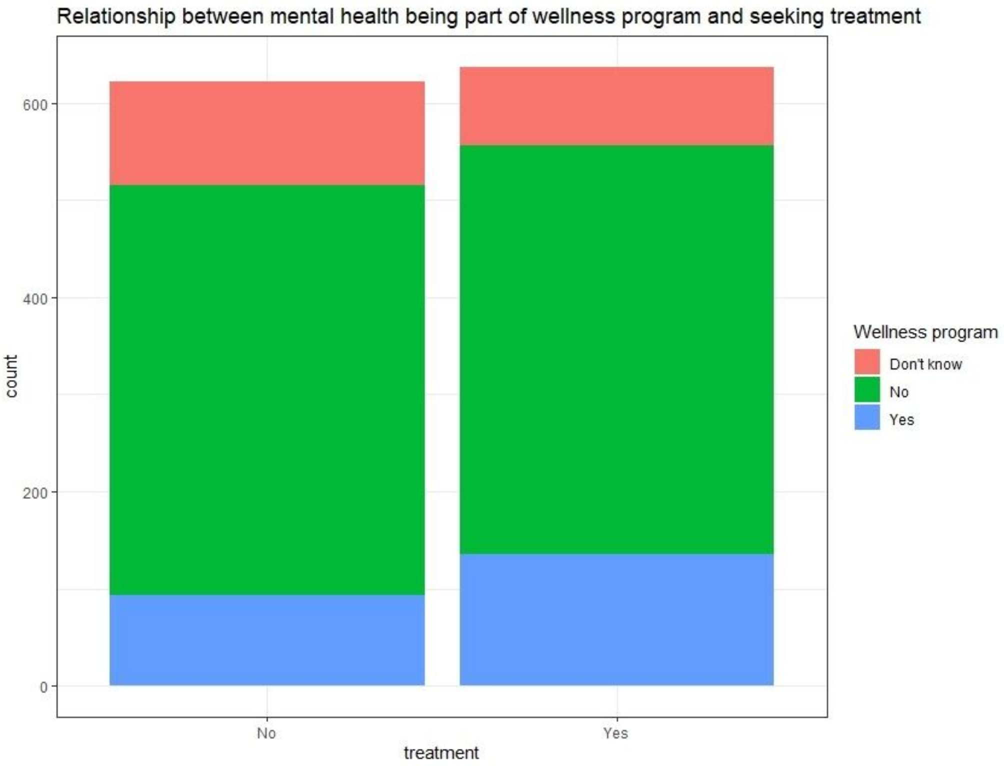
The aim here is to predict the probability of whether an employee will get treatment or not, data needs to be analyzed and explored, to see which factors or attributes can have a large influence on this decision. Based on some limited domain knowledge, the factor “treatment’ is compared against various variables (all categorical).

Relationship between getting help for mental health issues and family history for the same:



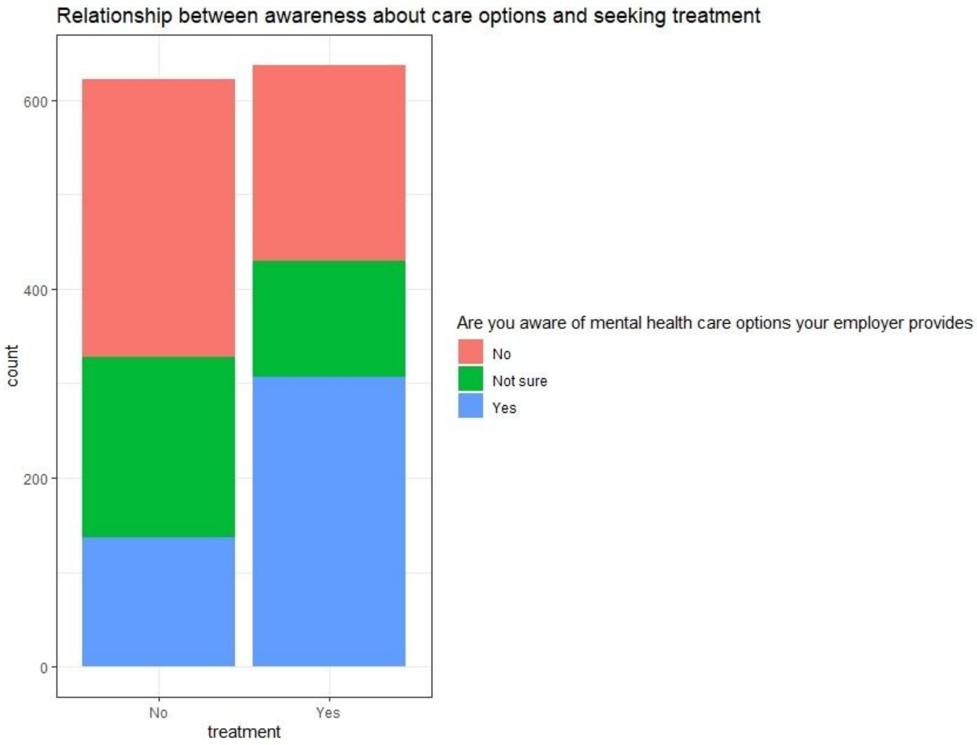
Over half the employees that are seeking treatment have a family history of mental health issues. This may be because after seeing a family member suffer due to their issues, they develop empathy and understand the importance of seeking help. The stigma is reduced. They may tend to become more accepting of the situation and more responsive to their own needs. This variable can have a major impact on their decision to seek help.

Relationship between getting help for mental health issues and mental health being part of the employers' wellness program:



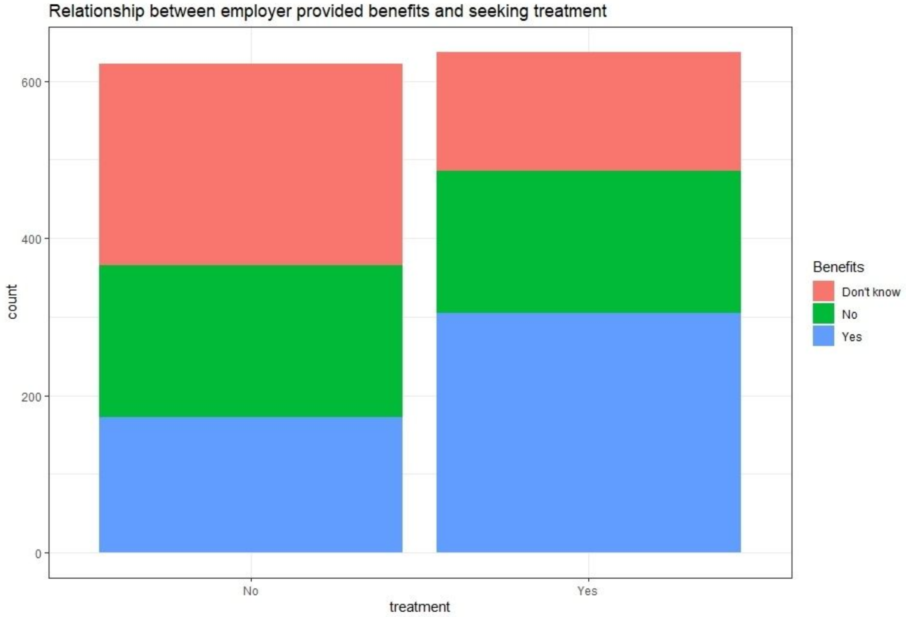
As observed above, a majority of employers don't even offer mental health care as a part of their wellness program. Not much difference is observed between the number of people seeking and no seeking help based on this variable. While this is slightly unexpected, we can assume that this variable may not have a lot of influence over an employee's decision to seek help.

Relationship between getting help for mental health issues and being aware that the employer provides care options:



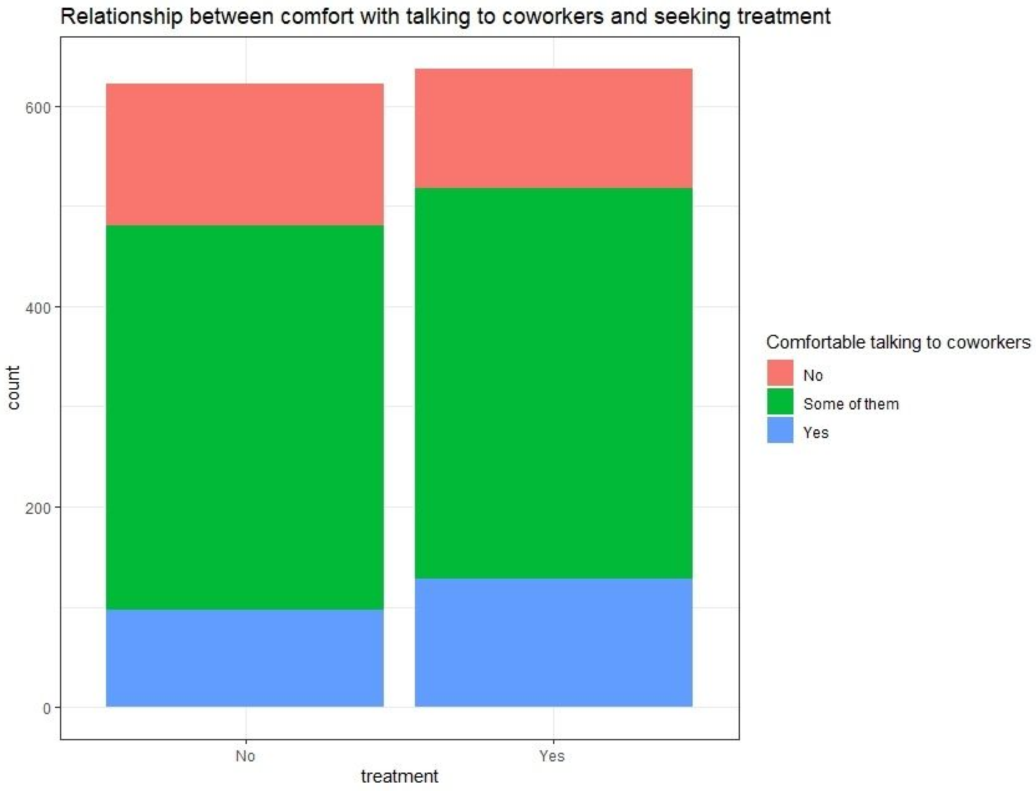
It could be assumed that just the simple knowledge about care options being available would alleviate some stress and encourage employees to seek treatment. A sense of security is formed, which may reduce the risk of developing mental health issues. It is observed that approximately half the people currently seeking treatment are aware of their options. This variable may be slightly influential.

Relationship between getting help for mental health issues and whether an employer provides benefits for mental health issues:



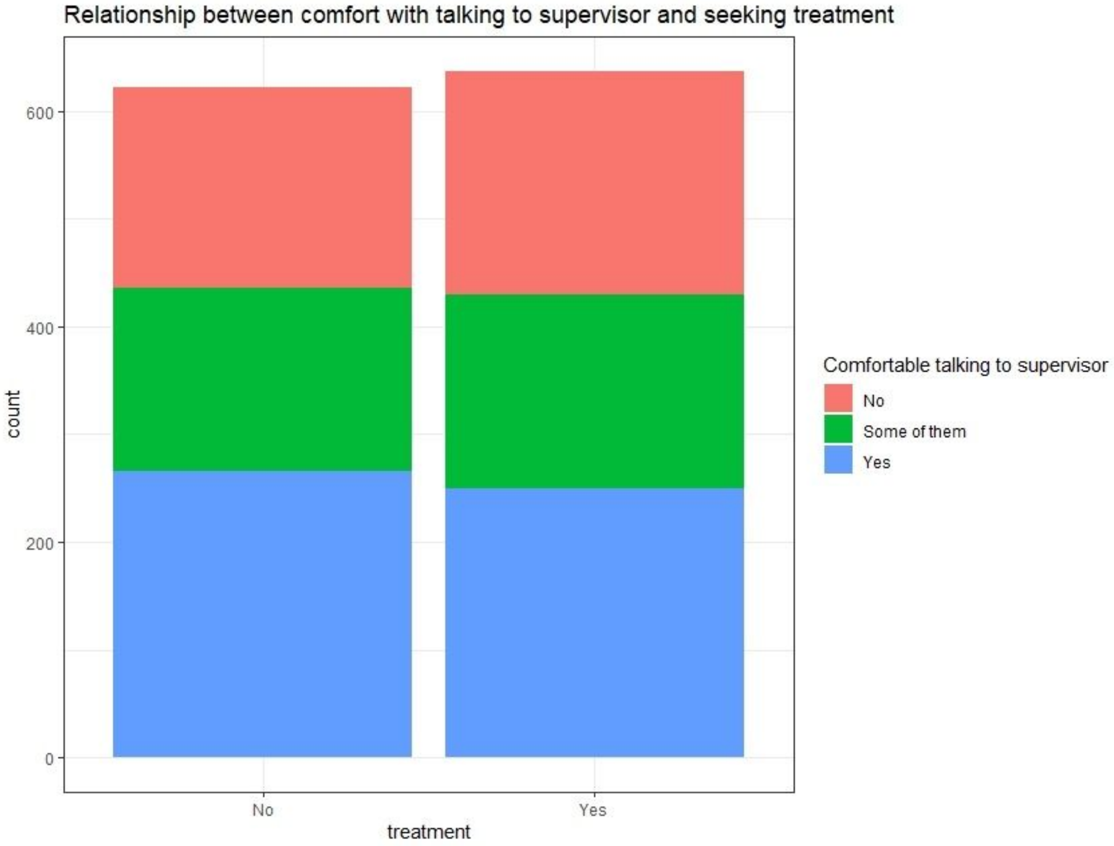
It is observed that there is a halfway split, but a concrete conclusion cannot be drawn since there are a lot of employees who are not aware of whether they are provided benefits or not.

Relationship between getting help for mental health issues and whether an employee is comfortable with talking to their co-workers about their issues:



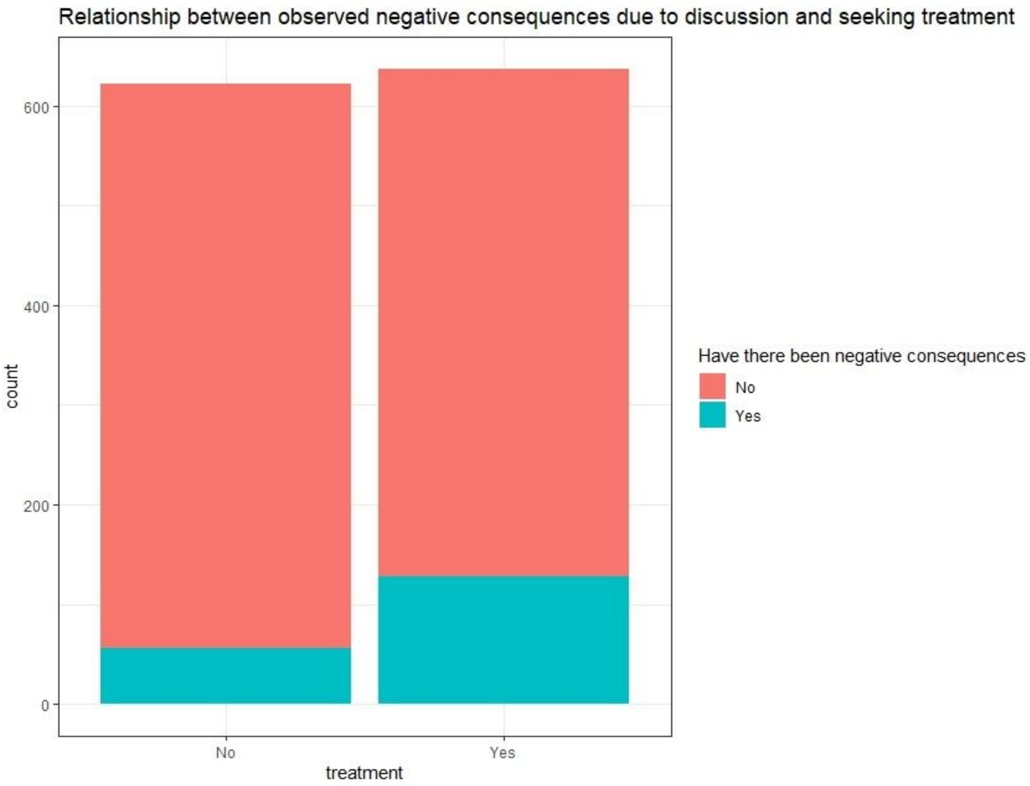
It is observed that a large fraction of people are comfortable with talking to at least someone about their problems. This creates a sense of security and kindness which may be beneficial towards reducing anxiety, loneliness, and stress. This variable may be a huge influence on the decision to seek help.

Relationship between getting help for mental health issues and whether an employee is comfortable with talking to their supervisor about their issues:



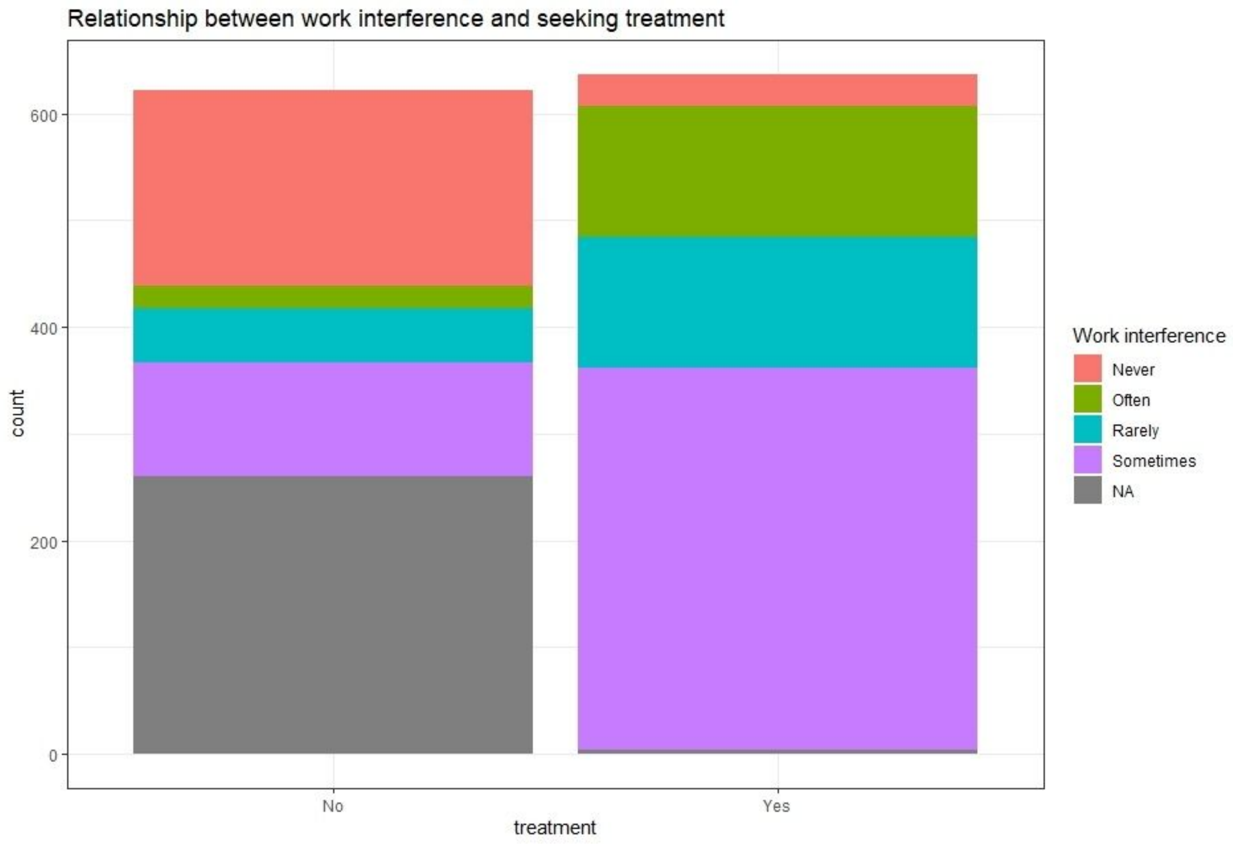
The trend here is very similar to that with “comfort with talking to a coworker”. A large fraction of people seeking help are comfortable discussing their issues with all or some of their supervisors. This again creates a sense of security and understanding. The employee feels that seeking help may not have a negative impact on their image in front of their superiors and feel supported, making them more responsive to seeking treatment. This variable may have a heavy influence on their decision.

Relationship between getting help for mental health issues and observed negative consequences due to discussing issues:



It can be safely assumed that experiencing or observing negative consequences to discussing mental health issues can be a great deterrent to employees wanting to seek help; it would discourage them from doing so. By looking at the distribution in the “Yes” and “No” graphs of seeking treatment, it can be observed that there aren't many cases where negative consequences are observed. While this is a good thing, due to a limited sample, we can’t say this is a global trend. For the sake of our prediction, it can be assumed that this variable will not have a lot of influence.

Relationship between getting help for mental health issues and whether mental health issues interfere with the employees' work:



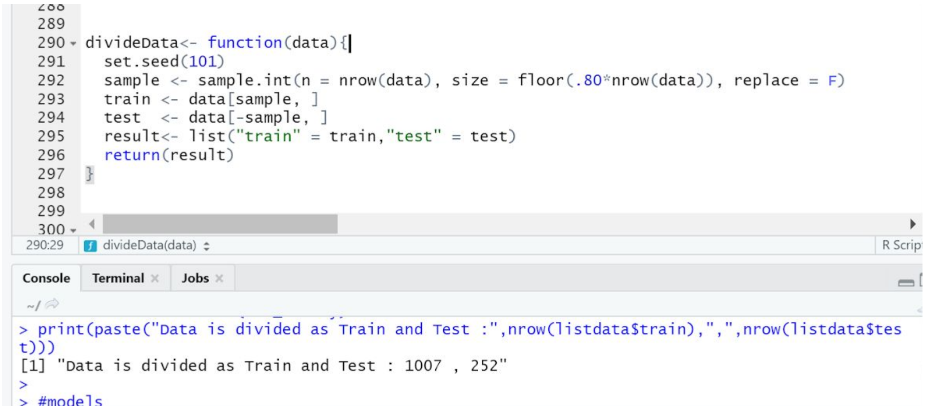
The above graph is quite conclusive. It can be observed that a clear majority (> 90%) of people who are currently seeking help for their mental health issues have experienced some sort of interference with their work. While it is obvious that such issues have a negative effect on the job, it does encourage people to seek help. This variable may have a great influence on the probability of seeking treatment.

Modeling:

The aim of our project is to predict whether an employee will seek help for their mental health issues depending on their work environment. Since the treatment variable is categorical we will be using the classification technique to build the models.

Based on the analysis we will be using a bootstrap aggregated tree (bagged) via random forest method and logistic regression method to build models.

Data is divided into two parts train data which is 80% and test data which is 20%.



```
divideData<- function(data){
  set.seed(101)
  sample <- sample.int(n = nrow(data), size = floor(.80*nrow(data)), replace = F)
  train <- data[sample, ]
  test  <- data[-sample, ]
  result<- list("train" = train,"test" = test)
  return(result)
}
```

```
> print(paste("Data is divided as Train and Test :",nrow(listdata$train),",",nrow(listdata$tes
t)))
[1] "Data is divided as Train and Test : 1007 , 252"
>
> #models
```

Random Forest :

Random forest is an algorithm that integrates multiple trees through the idea of integrated learning. Its basic unit is a decision tree, and its essence belongs to a large branch of machine learning, the integrated learning (Ensemble Learning) method. There are two keywords in the name of a random forest, one is "random" and the other is "forest". "Forest" we understand very well. One tree is called a tree, so hundreds or thousands of trees can be called a forest. This analogy is still very appropriate. In fact, this is also the embodiment of the main idea of random forest-integrated thought. We will talk about the meaning of "random" in the next section.

In fact, from an intuitive point of view, each decision tree is a classifier (assuming that it is now directed to a classification problem), then for an input sample, N trees will have N classification results. The random forest integrates all the classification voting results and specifies the category with the most votes as the final output. This is the simplest Bagging idea.

The main function of the **randomForest** package is classification and regression analysis, providing a total of 39 functions, the most commonly used is a random forest to achieve classification (Classification) and time series regression (Regression).

Logistic regression:

Logistic regression has many similarities with multiple linear regression. The biggest difference is that their dependent variables are different. The other basics are basically the same. Because of this, these two regressions can be attributed to the same family. That is the generalized linear model (generalized linear model). The model forms in this family are basically the same. The difference is that the dependent variables are different. If it is continuous, it is multiple linear regression. If it is a binomial distribution, it is logistic regression. If it is Negative binomial distribution is negative binomial regression, and so on. Just pay attention to distinguish their dependent variables.

The dependent variable of logistic regression can be dichotomous nonlinear difference equations, or it can be multi-classified, but the dichotomous classification is more commonly used and easier to explain. So the most commonly used in practice is the logistic regression of binary classification.

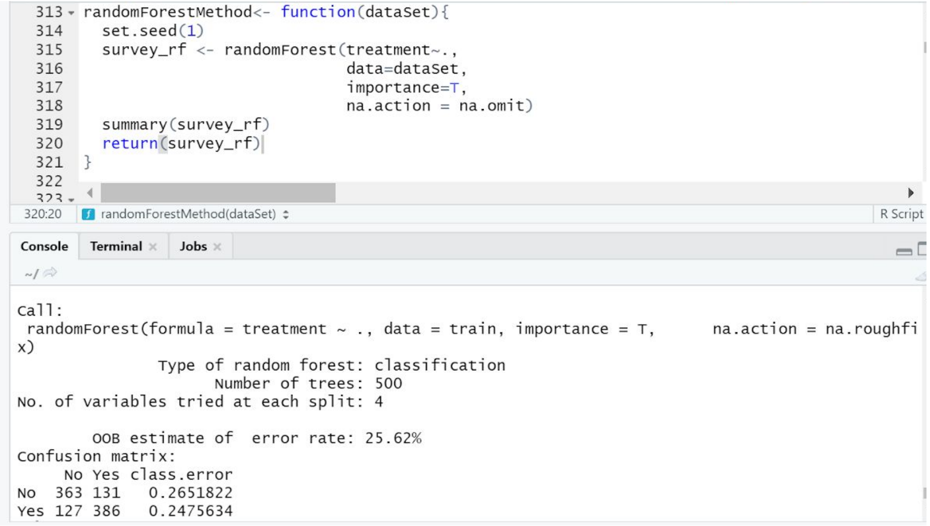
First, look for risk factors, as noted above to find a disease of the risk factors.

Second, Prediction. If you have established a logistic regression model, you can predict the probability of occurrence of a disease or a situation under different independent variables according to the model.

Third, Discrimination is actually similar to prediction. It is also based on the logistic model to determine the probability that a person belongs to a disease or a certain situation, that is, to see how likely this person is to belong to a disease.

Model Evaluation:

Random Forest

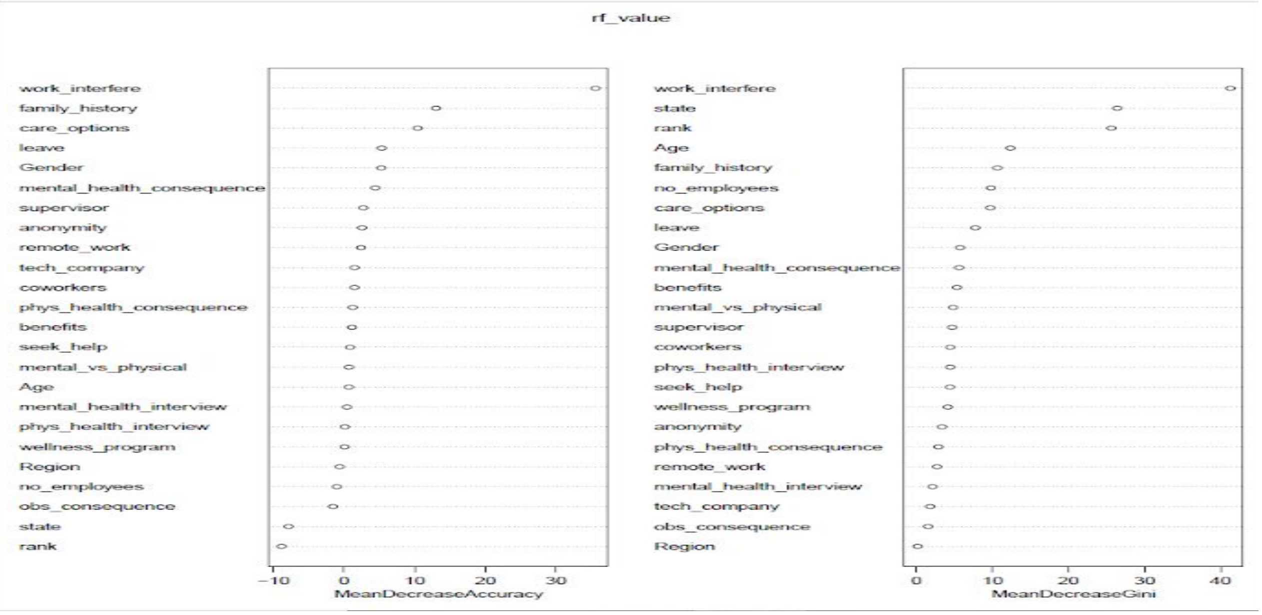


```
randomForestMethod<- function(dataSet){
  set.seed(1)
  survey_rf <- randomForest(treatment~.,
                             data=dataSet,
                             importance=T,
                             na.action = na.omit)
  summary(survey_rf)
  return(survey_rf)
}
```

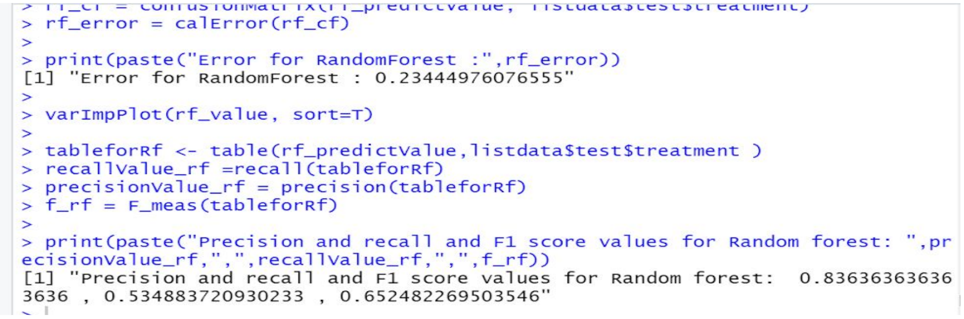
```
Call:
 randomForest(formula = treatment ~ ., data = train, importance = T,      na.action = na.roughfi
x)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 25.62%
Confusion matrix:
     No Yes class.error
No  363 131   0.2651822
Yes 127 386   0.2475634
```

By default, 500 trees are used. Four variable separation samples are obtained. The error evaluation matrix. The results are as follows: including analysis commands, optimization of the number of categorical variables selected 4, and data reclassification and error rate statistics. In this example, two variables are used for classification. The classification error rate is 25.62%. Visualization MeanDecreaseAccuracy, literally translated as average reduction accuracy, that is, without this feature, the degree of classification accuracy declines, which is equivalent to our commonly used concept of classification contribution.



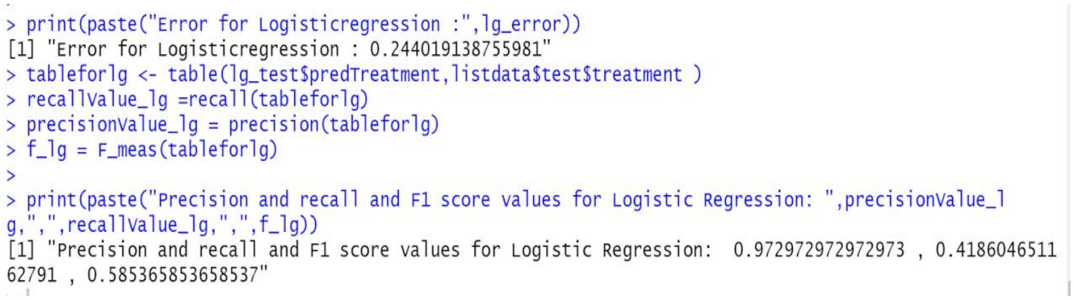
Error, Precision and recall and F1 score values for Random forest is



```
> rf_cf = confusionMatrix(rf_predictValue, listdata$test$treatment)
> rf_error = calError(rf_cf)
>
> print(paste("Error for RandomForest :",rf_error))
[1] "Error for RandomForest : 0.23444976076555"
>
> varImpPlot(rf_value, sort=T)
>
> tableforRf <- table(rf_predictValue,listdata$test$treatment )
> recallValue_rf =recall(tableforRf)
> precisionValue_rf = precision(tableforRf)
> f_rf = F_meas(tableforRf)
>
> print(paste("Precision and recall and F1 score values for Random forest: ",pr
ecisionValue_rf,",",recallValue_rf,",",f_rf))
[1] "Precision and recall and F1 score values for Random forest:  0.83636363636
3636 , 0.534883720930233 , 0.652482269503546"
```

Logistic regression:

Error, Precision and recall and F1 score values for Logistic regression:

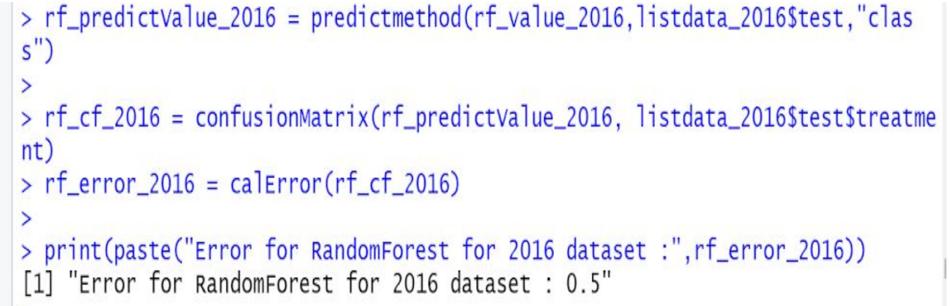


```
> print(paste("Error for Logisticregression :",lg_error))
[1] "Error for Logisticregression : 0.244019138755981"
> tableforlg <- table(lg_test$predTreatment,listdata$test$treatment )
> recallValue_lg =recall(tableforlg)
> precisionValue_lg = precision(tableforlg)
> f_lg = F_meas(tableforlg)
>
> print(paste("Precision and recall and F1 score values for Logistic Regression: ",precisionValue_l
g,",",recallValue_lg,",",f_lg))
[1] "Precision and recall and F1 score values for Logistic Regression:  0.972972972972973 , 0.4186046511
62791 , 0.585365853658537"
```

The random forest has less error value than logistic regression and not only error even recall metrics can be used to select the model. Recall metric can be best when there is a high cost associated with False Negative. So we will be using the random forest method.

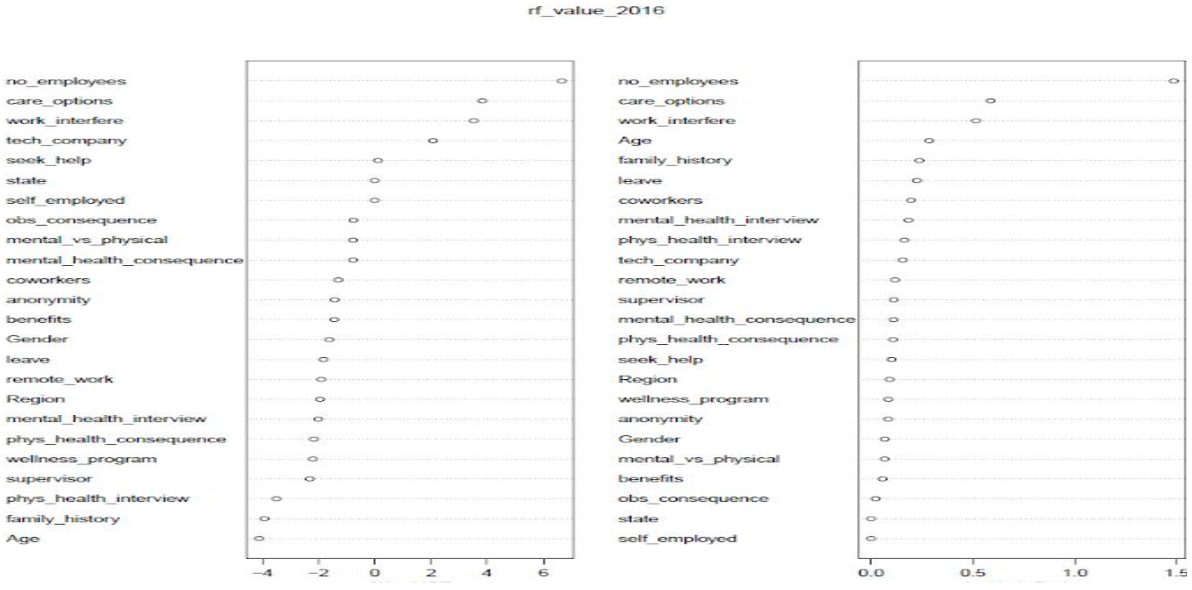
Random Forest for 2016:

Error for 2016 dataset:



```
> rf_predictValue_2016 = predictmethod(rf_value_2016,listdata_2016$test,"clas
s")
>
> rf_cf_2016 = confusionMatrix(rf_predictValue_2016, listdata_2016$test$treatme
nt)
> rf_error_2016 = calError(rf_cf_2016)
>
> print(paste("Error for RandomForest for 2016 dataset :",rf_error_2016))
[1] "Error for RandomForest for 2016 dataset : 0.5"
```

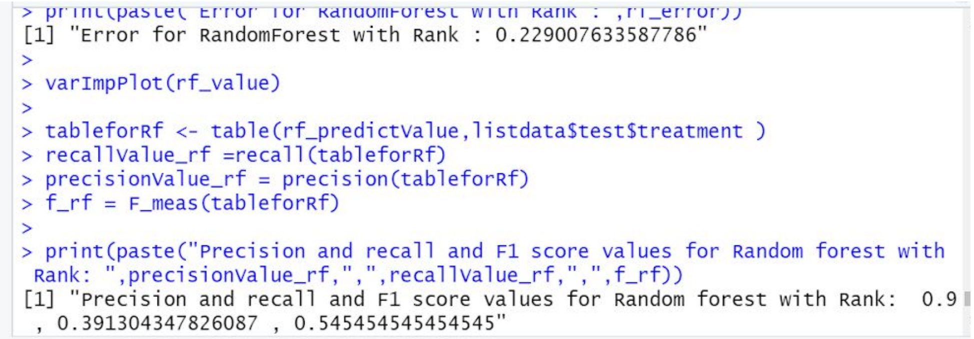
Variables used to form trees



Leveraging with Secondary dataset:

For the secondary dataset, we are using crime data in the United States. We have added crimes committed in each state in 2014 and ranked the states accordingly. After adding the crime ranks error reduces to 22.9% from 23.4%

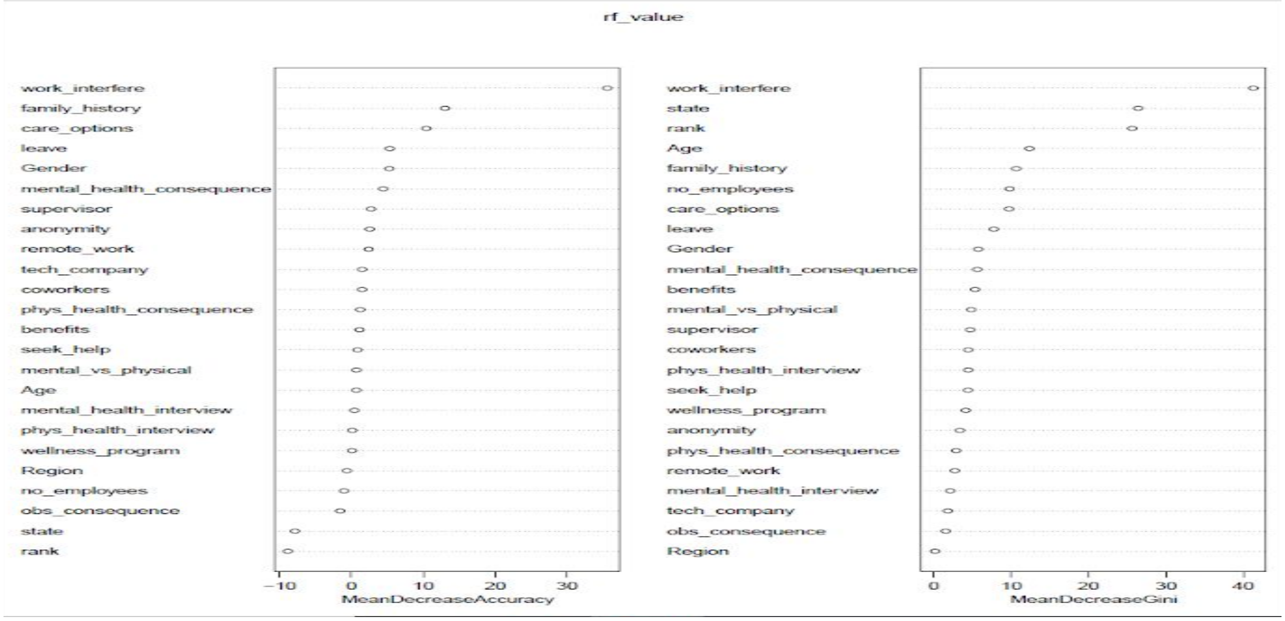
Error, Precision, Recall and F1 score values for Random forest with rank are:



```
> print(paste("Error for RandomForest with Rank : ",rf_error))
[1] "Error for RandomForest with Rank : 0.229007633587786"
>
> varImpPlot(rf_value)
>
> tableforRf <- table(rf_predictValue,listdata$test$treatment )
> recallValue_rf =recall(tableforRf)
> precisionValue_rf = precision(tableforRf)
> f_rf = F_meas(tableforRf)
>
> print(paste("Precision and recall and F1 score values for Random forest with
 Rank: ",precisionValue_rf,",",recallValue_rf,",",f_rf))
[1] "Precision and recall and F1 score values for Random forest with Rank:  0.9
 , 0.391304347826087 , 0.545454545454545"
```

Since the error is reduced after adding rank, we will be deploying the model with rank.

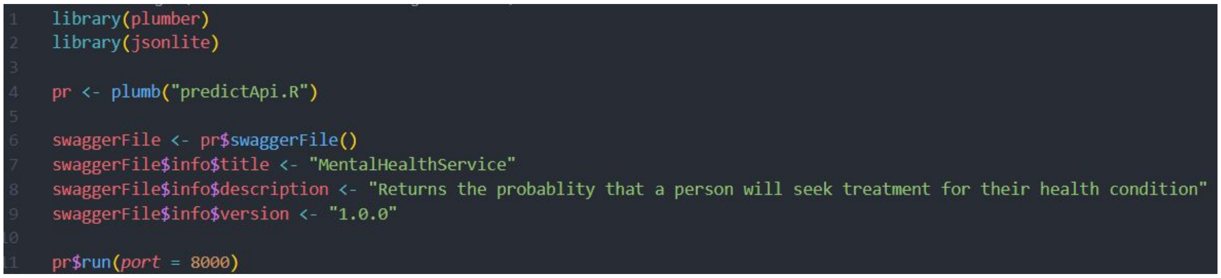
Variables used for the model are:



Deployment:

The plumber library is used for the deployment of the model. A simple API is created with only one endpoint which returns the probability of whether a person will seek treatment for their mental health condition. It is a POST request which requires data to predict the possibility.

Code for the server:



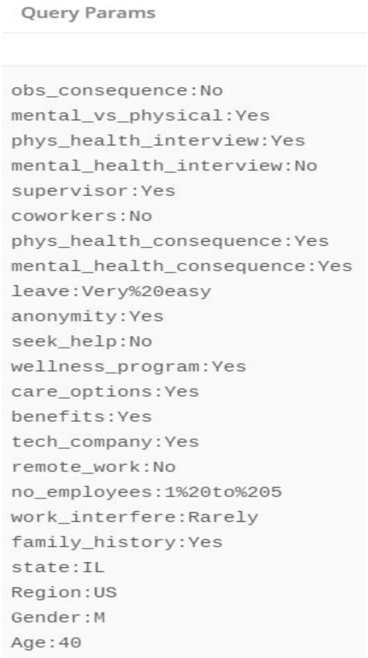
```
library(plumber)
library(jsonlite)

pr <- plumb("predictApi.R")

swaggerFile <- pr$swaggerFile()
swaggerFile$info$title <- "MentalHealthService"
swaggerFile$info$description <- "Returns the probablity that a person will seek treatment for their health condition"
swaggerFile$info$version <- "1.0.0"

pr$run(port = 8000)
```

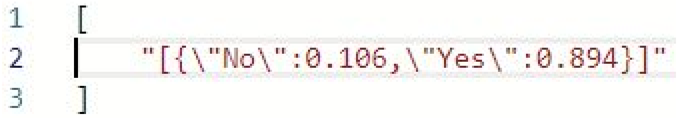
Parameters required by the server to make the prediction:



Query Params

```
obs_consequence:No
mental_vs_physical:Yes
phys_health_interview:Yes
mental_health_interview:No
supervisor:Yes
coworkers:No
phys_health_consequence:Yes
mental_health_consequence:Yes
leave:Very%20easy
anonymity:Yes
seek_help:No
wellness_program:Yes
care_options:Yes
benefits:Yes
tech_company:Yes
remote_work:No
no_employees:1%20to%205
work_interfere:Rarely
family_history:Yes
state:IL
Region:US
Gender:M
Age:40
```

The server generates the output and returns the response in JSON format. As seen here, the probability of “Yes” is 0.894 and the probability of “No” is 0.106 for the parameters which are specified above.



```
[
    "[{\"No\":0.106,\"Yes\":0.894}]"
]
```

Conclusion:

After all the modeling, it is observed that the main workplace factors that can affect a person's willingness to receive treatment for their mental health issues are:

- The issues' interference with their work efficiency.
- Knowledge about whether the employer provides mental health care options.
- Whether the employer provides benefits for mental health issues.
- The ease of taking leave for health-related reasons.
- Whether there are any negative consequences of discussing mental health issues with supervisors.

The poor mental health of an employee can have severe consequences. It affects their job productivity and performance, and their engagement with it. It breaks down the communication between coworkers, which will reduce their daily functioning.

What can employers do?

- Mental health self-assessment tools can be made available to all employees.
- Offer health insurance at low out-of-pocket charges, which includes counseling and medications, to all the employees.
- Sensitization of supervisors is required. Open discussion leads to breaking of the stigma surrounding these problems. Provide them with training that educates them about the issues, and about identifying early signs of stress and depression in their team members. Support from a figure of authority can be a huge boost.
- Educating employees is also necessary. Conduct seminars, distribute brochures talking about the signs and the importance of seeking treatment. Employees should be valued and feel the same.
- Designate quiet and safe spaces in the workplace where people can take a break and have much-needed discussions with someone they trust.

Citations:

- Elkin AJ, Rosch PJ. Promoting mental health at the workplace: the prevention side of stress management. Occup Med. 1990 Oct-Dec;5(4) 739-754. PMID: 2237702.

- Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786. https://www.researchgate.net/publication/242579096_An_Introduction_to_Logistic_Regression_Analysis_and_Reporting

- Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9. https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees

- Mental Health in the Workplace. (2019, April 10). Retrieved from https://www.cdc.gov/workplacehealthpromotion/tools-resources/workplace-health/mental-health/index.html

- Mental Health In Tech Survey. Retrieved from https://kaggle.com/osmi/mental-health-in-tech-survey

- Bagging trees retrieved from https://www2.isye.gatech.edu/~tzhao80/Lectures/Lecture_6.pdf