Mental Health Analysis in Tech Industry

Author: Saelee, KC
Date: 05/11/2025

Background:

Mental health issues in the tech industry are often overlooked due to stigma and high performance expectations. This dataset contains responses from a 2014 survey of tech employees about mental health awareness, support, and treatment in their workplaces.

Project Goal:

The goal is to analyze trends and patterns in mental health disclosures, support availability, and openness among employees in the tech industry to identify areas for policy improvement.

Objective:

Our objective is to determine what factors (age, gender, company size, awareness of benefits, etc.) are associated with whether an employee seeks mental health treatment

Analysis:

First, let's load our data file
Then we'll look at a snapshot view of a few variables (columns) and observations (rows)

Obs	Age	Gender	Country	treatment	no_employees
1	37	Female	United States	Yes	6-25
2	44	M	United States	No	More than 1000
3	32	Male	Canada	No	6-25
4	31	Male	United Kingdom	Yes	26-100
5	31	Male	United States	No	100-500
6	33	Male	United States	No	6-25
7	35	Female	United States	Yes	1-5
8	39	M	Canada	No	1-5
9	42	Female	United States	Yes	100-500
10	23	Male	Canada	No	26-100

We'll look into the data structure as well (variables, rows, datatypes, etc.) and use thisinformation to identify where data preprocessing and cleaning is needed.

The CONTENTS Procedure

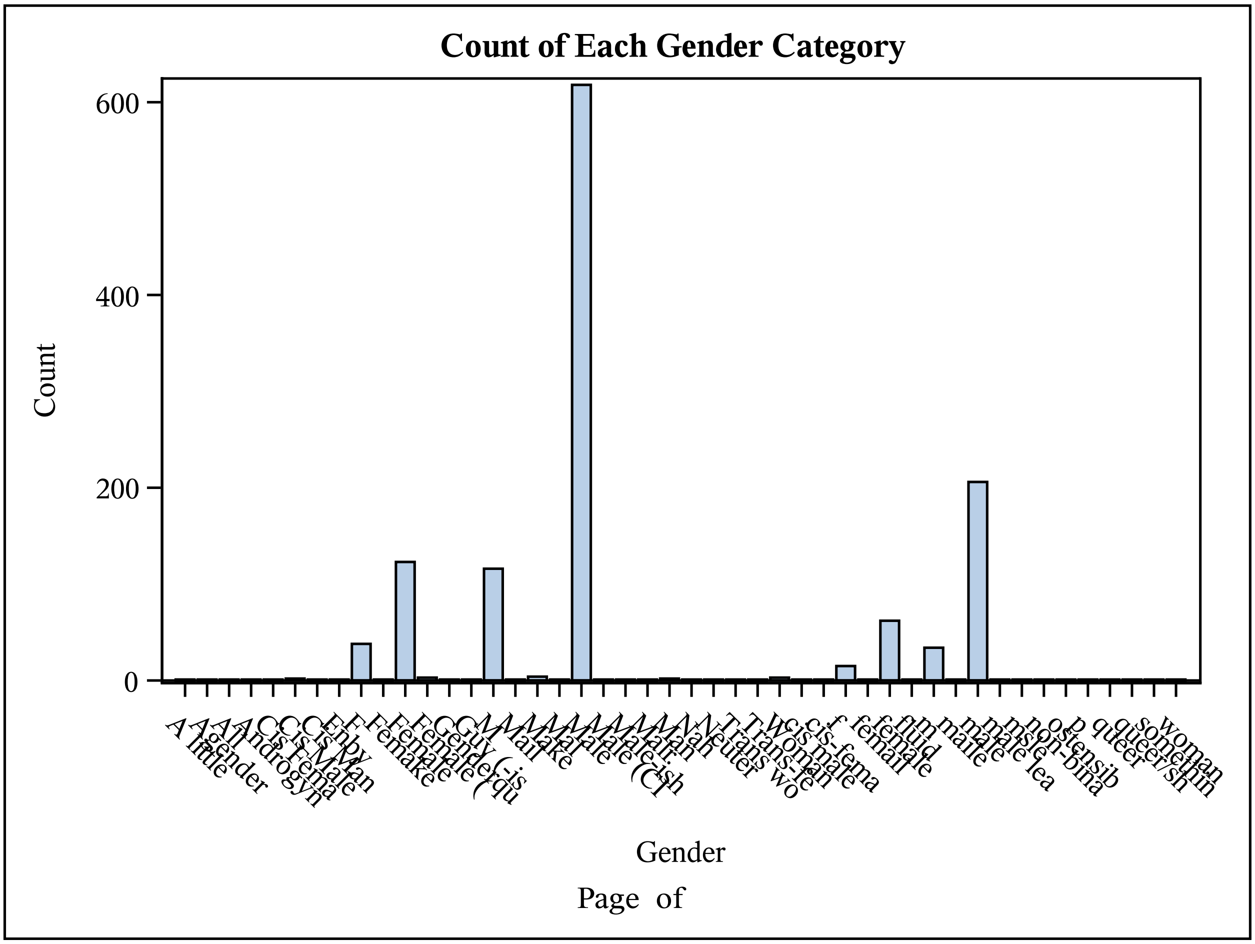
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
2	Age	Num	8	BEST12.	BEST32.
4	Country	Char	16	$16.	$16.
3	Gender	Char	8	$8.	$8.
1	Timestamp	Num	8	DATETIME.	ANYDTDTM40.
17	anonymity	Char	12	$12.	$12.
13	benefits	Char	12	$12.	$12.
14	care_options	Char	10	$10.	$10.
27	comments	Char	263	$263.	$263.
21	coworkers	Char	14	$14.	$14.
7	family_history	Char	5	$5.	$5.
18	leave	Char	20	$20.	$20.
19	mental_health_consequence	Char	7	$7.	$7.
23	mental_health_interview	Char	7	$7.	$7.
25	mental_vs_physical	Char	12	$12.	$12.
10	no_employees	Char	16	$16.	$16.
26	obs_consequence	Char	5	$5.	$5.
20	phys_health_consequence	Char	7	$7.	$7.
24	phys_health_interview	Char	7	$7.	$7.
11	remote_work	Char	5	$5.	$5.
16	seek_help	Char	12	$12.	$12.
6	self_employed	Char	5	$5.	$5.
5	state	Char	4	$4.	$4.
22	supervisor	Char	14	$14.	$14.
12	tech_company	Char	5	$5.	$5.
8	treatment	Char	5	$5.	$5.
15	wellness_program	Char	12	$12.	$12.
9	work_interfere	Char	11	$11.	$11.

Before we go to the data analysis step, we need to check for missing rows in the data and return rows with at least 1 missing data. Depending on the what values are missing, we need to handlethem accordingly.

Since the only missing value is only one record and it is in the comments column, we do not need to remove the recordor change it in any way and can proceed to the next data cleaning step.

All rows should be unique so we'll also remove duplicated rows and display the results

Now we'll check the 'Gender' column and see whether we need to do any data cleaning there.



We notice from the data that there are many different values in the 'Gender' column.For example, there are values such as 'm', 'woman' and 'transgender'. Since the other variables represent such as small size compared to 'Male' and 'Female' we will do our best in classifying the gender column so all words starting with the letter 'm' is 'Male' amd letter 'f' is 'female'. For the sake of simpifying the analysis, everything else will be classfied as 'Other'. However, we should be respectful of cognizant of people's view on gender and not marginalize it.

Next, we'll look into the 'Age' column. The table is showing that there are negative ages and ages as high as 9 million! Since age can only start at 0 years old and no human in recorded history has live past age 130, we need to filter out appropriate ages

Count of Each Gender Category

The MEANS Procedure

Analysis Variable : Age	
Minimum	**Maximum**
-1726.00	99999999999

Here we filter out odd ages such as negative numbers and very high unlrealistic numbers

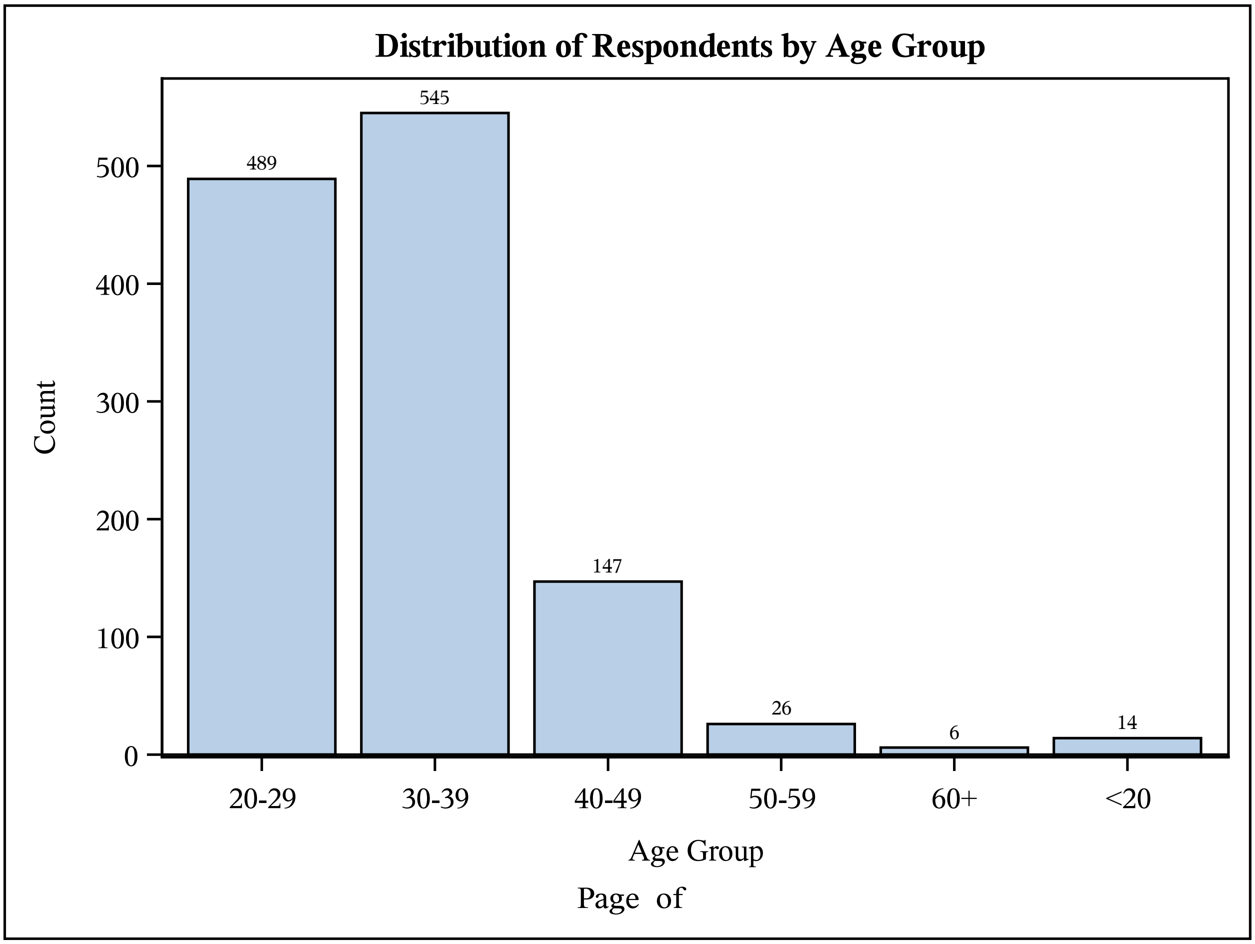
Count of Each Gender Category

The MEANS Procedure

Analysis Variable : Age	
Minimum	**Maximum**
18.0000000	72.0000000

There are many ages and it is a numeric variable. To answer the question of whether there aresignificant differences in ages, one approach to answer this question is to bin the ages into categorical groups

Now we visualize the Age groups to see how it looks



Now that are dataset is cleaned, we are ready to proceed with the analysis

We will test to see if there are statistically significant differences between treatment and other catergorial variables such as Age Group, Gender, Number of employeesin the company and Benefits (whether company provides benefits)

Age Group vs Treatment

Distribution of Respondents by Age Group

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Age_Group by treatment			
	treatment		
Age_Group	No	Yes	Total
20-29	262 21.35 53.58 42.74	227 18.50 46.42 36.97	489 39.85
30-39	268 21.84 49.17 43.72	277 22.58 50.83 45.11	545 44.42
40-49	61 4.97 41.50 9.95	86 7.01 58.50 14.01	147 11.98
50-59	11 0.90 42.31 1.79	15 1.22 57.69 2.44	26 2.12
60+	3 0.24 50.00 0.49	3 0.24 50.00 0.49	6 0.49
<20	8 0.65 57.14 1.31	6 0.49 42.86 0.98	14 1.14
Total	613 49.96	614 50.04	1227 100.00

Distribution of Respondents by Age Group

The FREQ Procedure

Statistics for Table of Age_Group by treatment

Statistic	DF	Value	Prob
Chi-Square	5	7.8057	0.1673
Likelihood Ratio Chi-Square	5	7.8320	0.1657
Mantel-Haenszel Chi-Square	1	3.6921	0.0547
Phi Coefficient		0.0798	
Contingency Coefficient		0.0795	
Cramer's V		0.0798	

Sample Size = 1227

Gender vs Treatment

Distribution of Respondents by Age Group

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Gender by treatment			
	treatment		
Gender	No	Yes	Total
Female	75 6.11 30.74 12.23	169 13.77 69.26 27.52	244 19.89
Male	538 43.85 54.73 87.77	445 36.27 45.27 72.48	983 80.11
Total	613 49.96	614 50.04	1227 100.00

Distribution of Respondents by Age Group

The FREQ Procedure

Statistics for Table of Gender by treatment

Statistic	DF	Value	Prob
Chi-Square	1	45.0109	<.0001
Likelihood Ratio Chi-Square	1	45.9776	<.0001
Continuity Adj. Chi-Square	1	44.0563	<.0001
Mantel-Haenszel Chi-Square	1	44.9742	<.0001
Phi Coefficient		-0.1915	
Contingency Coefficient		0.1881	
Cramer's V		-0.1915	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	75
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 1227

Number of employees vs Treatment

Distribution of Respondents by Age Group

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of no_employees by treatment

no_employees	treatment: No	treatment: Yes	Total
1-5	69 5.62 44.81 11.26	85 6.93 55.19 13.84	154 12.55
100-500	80 6.52 46.51 13.05	92 7.50 53.49 14.98	172 14.02
26-100	138 11.25 48.76 22.51	145 11.82 51.24 23.62	283 23.06
500-1000	32 2.61 54.24 5.22	27 2.20 45.76 4.40	59 4.81
6-25	162 13.20 56.84 26.43	123 10.02 43.16 20.03	285 23.23
More than 1000	132 10.76 48.18 21.53	142 11.57 51.82 23.13	274 22.33
Total	613 49.96	614 50.04	1227 100.00

Distribution of Respondents by Age Group

The FREQ Procedure

Statistics for Table of no_employees by treatment

Statistic	DF	Value	Prob
Chi-Square	5	8.7974	0.1174
Likelihood Ratio Chi-Square	5	8.8185	0.1165
Mantel-Haenszel Chi-Square	1	2.5933	0.1073
Phi Coefficient		0.0847	
Contingency Coefficient		0.0844	
Cramer's V		0.0847	

Sample Size = 1227

Benefits vs Treatment

Distribution of Respondents by Age Group

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of benefits by treatment			
	treatment		
benefits	No	Yes	Total
Don't know	255 20.78 63.43 41.60	147 11.98 36.57 23.94	402 32.76
No	191 15.57 52.47 31.16	173 14.10 47.53 28.18	364 29.67
Yes	167 13.61 36.23 27.24	294 23.96 63.77 47.88	461 37.57
Total	613 49.96	614 50.04	1227 100.00

Statistics for Table of benefits by treatment

Statistic	DF	Value	Prob
Chi-Square	2	64.8912	<.0001
Likelihood Ratio Chi-Square	2	65.7076	<.0001
Mantel-Haenszel Chi-Square	1	64.1244	<.0001
Phi Coefficient		0.2300	
Contingency Coefficient		0.2241	
Cramer's V		0.2300	

Sample Size = 1227

We can see from the tests that Age group and Number of employees were not stasttically significant at seeking treatmentbut gender and awarenss of benefits are ($p < 0.05$)

Next, we'll run the data in a logistic regression model and see which variables arethe most important predictors. Since we have many variables that could be correlated, we'll use a stepwise approach to select for variables in the model to include until the fit is appropriate

Distribution of Respondents by Age Group

The LOGISTIC Procedure

Model Information	
Data Set	WORK.MHCLEAN
Response Variable	treatment
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1227
Number of Observations Used	1227

Response Profile		
Ordered Value	treatment	Total Frequency
1	Yes	614
2	No	613

Probability modeled is treatment='Yes'.

Stepwise Selection Procedure

Class Level Information						
Class	Value	Design Variables				
Age_Group	20-29	1	0	0	0	0
	30-39	0	1	0	0	0
	40-49	0	0	1	0	0
	50-59	0	0	0	1	0
	60+	0	0	0	0	1
	<20	0	0	0	0	0
Gender	Female	1				
	Male	0				
benefits	Don't know	1	0			
	No	0	0			
	Yes	0	1			
no_employees	1-5	1	0	0	0	0
	100-500	0	1	0	0	0
	26-100	0	0	1	0	0

Distribution of Respondents by Age Group

The LOGISTIC Procedure

Class Level Information						
Class	**Value**	**Design Variables**				
	500-1000	0	0	0	1	0
	6-25	0	0	0	0	0
	More than 1000	0	0	0	0	1
family_history	**No**	0				
	Yes	1				
remote_work	**No**	0				
	Yes	1				

Step 0. Intercept entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L	=	1700.982

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
250.4266	15	<.0001

Step 1. Effect family_history entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1702.982	1519.954
SC	1708.095	1530.179
-2 Log L	1700.982	1515.954

Distribution of Respondents by Age Group

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	185.0283	1	<.0001
Score	179.2505	1	<.0001
Wald	167.0684	1	<.0001

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
83.9062	14	<.0001

Note: No effects for the model in Step 1 are removed.

Step 2. Effect benefits entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1702.982	1476.869
SC	1708.095	1497.319
-2 Log L	1700.982	1468.869

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	232.1132	3	<.0001
Score	219.2755	3	<.0001
Wald	193.2852	3	<.0001

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
38.4757	12	0.0001

Note: No effects for the model in Step 2 are removed.

Distribution of Respondents by Age Group

The LOGISTIC Procedure

Step 3. Effect Gender entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1702.982	1455.193
SC	1708.095	1480.754
-2 Log L	1700.982	1445.193

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	255.7897	4	<.0001
Score	238.0754	4	<.0001
Wald	204.5360	4	<.0001

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
15.1968	11	0.1737

Note: No effects for the model in Step 3 are removed.

Note: No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection							
	Effect						
Step	**Entered**	**Removed**	**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**
1	**family_history**		1	1	179.2505		<.0001
2	**benefits**		2	2	46.9797		<.0001
3	**Gender**		1	3	23.4620		<.0001

Distribution of Respondents by Age Group

The LOGISTIC Procedure

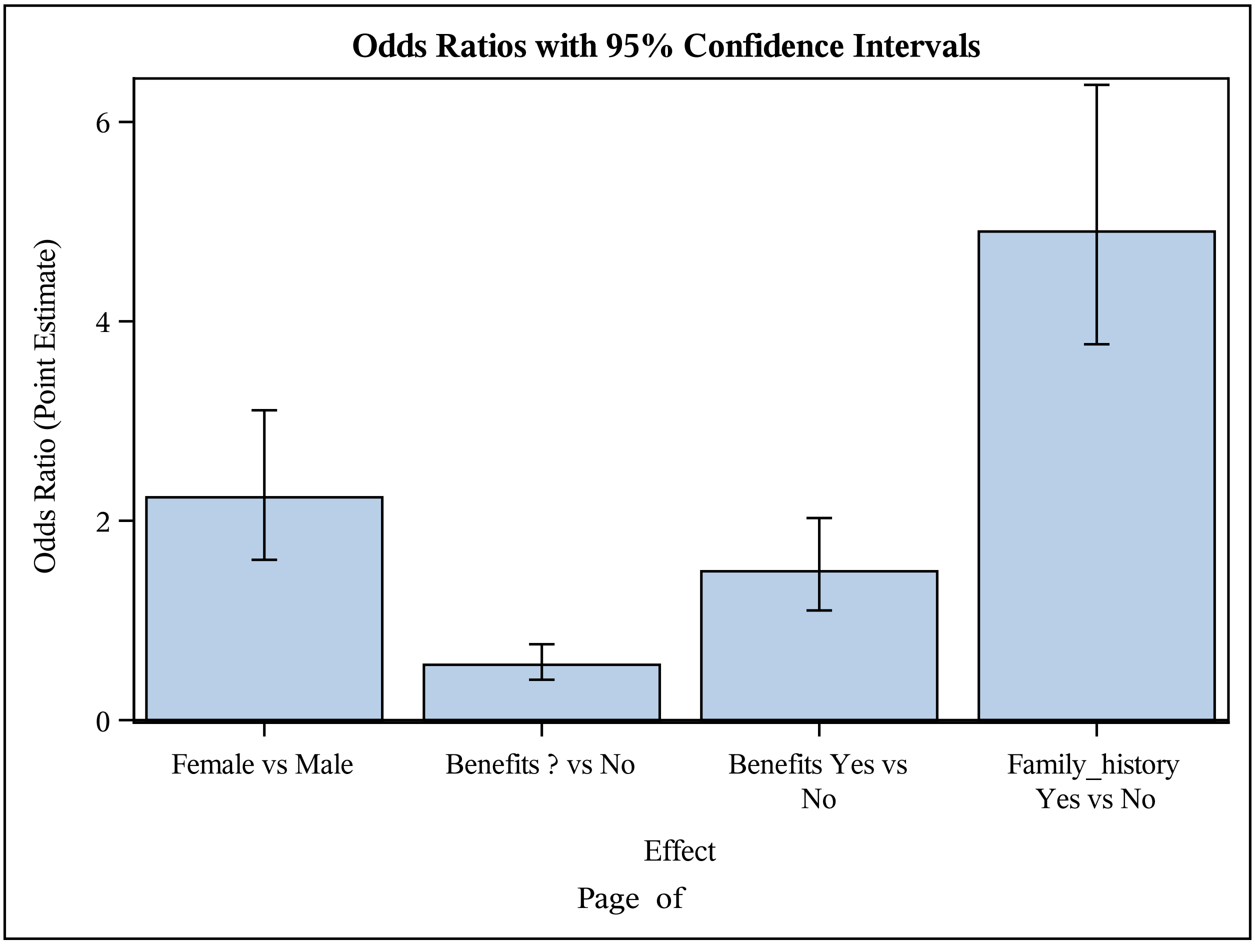
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	22.8791	<.0001
benefits	2	41.4336	<.0001
family_history	1	140.9109	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.7040	0.1215	33.5534	<.0001
Gender	Female	1	0.8043	0.1681	22.8791	<.0001
benefits	Don't know	1	-0.5889	0.1609	13.3964	0.0003
benefits	Yes	1	0.4009	0.1561	6.5984	0.0102
family_history	Yes	1	1.5895	0.1339	140.9109	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gender Female vs Male	2.235	1.608	3.108
benefits Don't know vs No	0.555	0.405	0.761
benefits Yes vs No	1.493	1.100	2.027
family_history Yes vs No	4.901	3.770	6.372

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.3	Somers' D	0.505
Percent Discordant	19.7	Gamma	0.561
Percent Tied	10.0	Tau-a	0.253
Pairs	376382	c	0.753

Here is the odds ratio (point estimate) of each of the significant factor levels



Here are the results from the model:
1. Women have 2.24 times higher odds seeking treatment compared men.
2. Those who answered 'Don't know' have lower odds (by ~45%) compared to those who answered 'No'
3. Those who answered 'Yes' have 1.49 times higher odds of seeking treatment than those who answered 'No'
4. Those with a family history have 4.9 times higher odds of seeking treatment.
All effects were statsitically signifcant ($p < 0.05$).

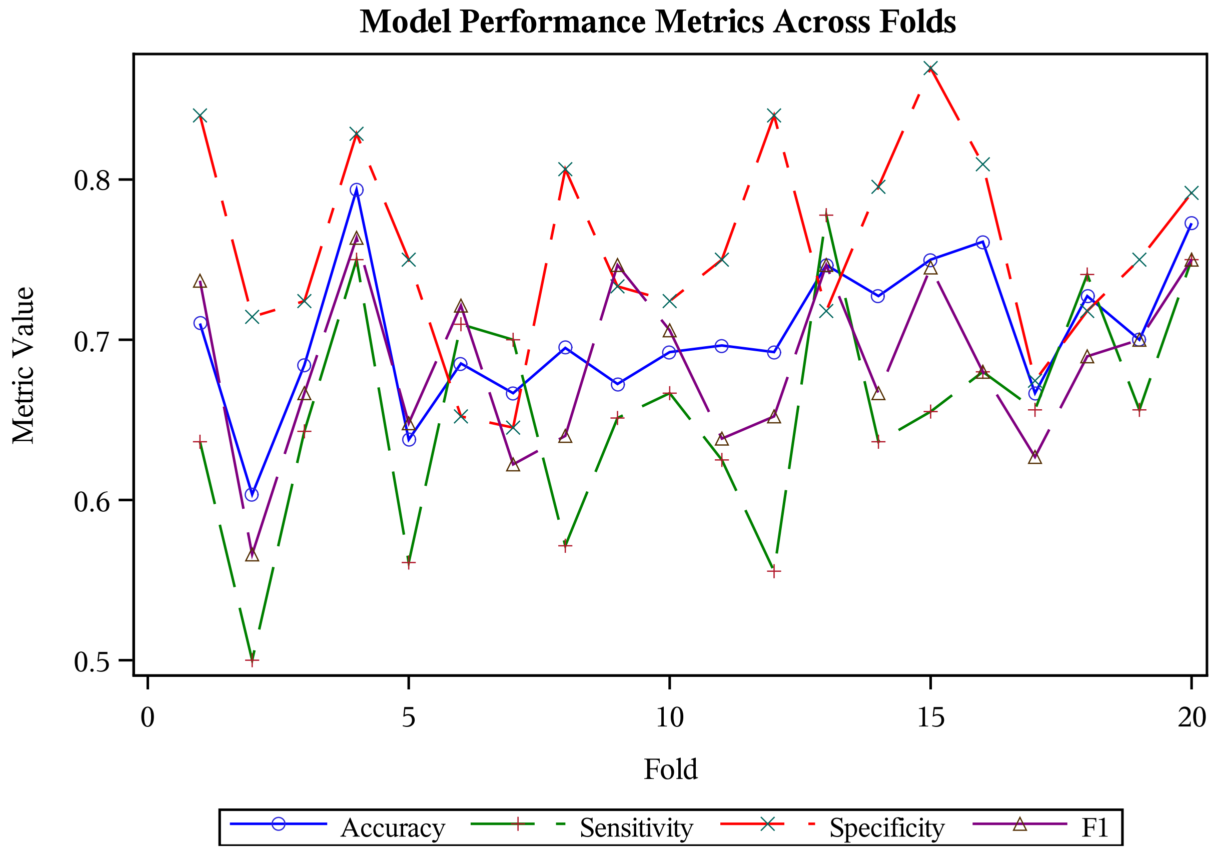
Finally, we will validate the dataset using cross validation to see how well the dataset generalizes on unseen data.To do this, we'll implement a macros doing 20 k-fold cross validation were k-1 folds will be used for training and the remianing1 fold is used for testing and validation. Metrics such as accuracy, specificity, sensitivity, precision, recall and f1 will be returned.The metrics will be represented as the average of the folds.

Below are the results for the corss validation,

Cross-Validation Metrics by Fold

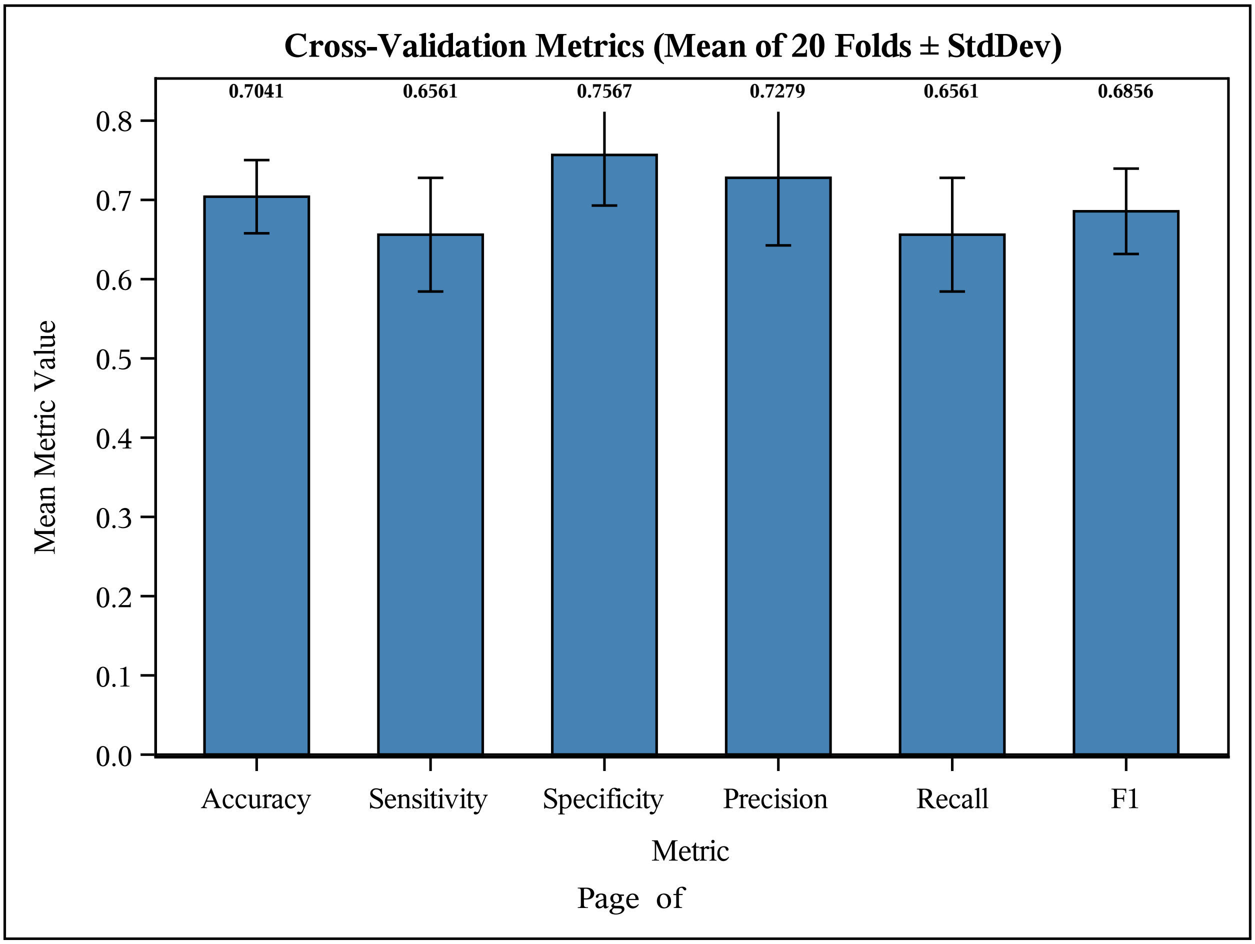
Obs	Fold	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
1	1	0.71014	0.63636	0.84000	0.87500	0.63636	0.73684
2	2	0.60345	0.50000	0.71429	0.65217	0.50000	0.56604
3	3	0.68421	0.64286	0.72414	0.69231	0.64286	0.66667
4	4	0.79365	0.75000	0.82857	0.77778	0.75000	0.76364
5	5	0.63768	0.56098	0.75000	0.76667	0.56098	0.64789
6	6	0.68519	0.70968	0.65217	0.73333	0.70968	0.72131
7	7	0.66667	0.70000	0.64516	0.56000	0.70000	0.62222
8	8	0.69492	0.57143	0.80645	0.72727	0.57143	0.64000
9	9	0.67241	0.65116	0.73333	0.87500	0.65116	0.74667
10	10	0.69231	0.66667	0.72414	0.75000	0.66667	0.70588
11	11	0.69643	0.62500	0.75000	0.65217	0.62500	0.63830
12	12	0.69231	0.55556	0.84000	0.78947	0.55556	0.65217
13	13	0.74667	0.77778	0.71795	0.71795	0.77778	0.74667
14	14	0.72727	0.63636	0.79545	0.70000	0.63636	0.66667
15	15	0.75000	0.65517	0.86957	0.86364	0.65517	0.74510
16	16	0.76119	0.68000	0.80952	0.68000	0.68000	0.68000
17	17	0.66667	0.65625	0.67442	0.60000	0.65625	0.62687
18	18	0.72727	0.74074	0.71795	0.64516	0.74074	0.68966
19	19	0.70000	0.65625	0.75000	0.75000	0.65625	0.70000
20	20	0.77273	0.75000	0.79167	0.75000	0.75000	0.75000

Plot the Metrics. This visualization shows the accuracy, sensitivity, specificity, etc.for each fold



Page of

Lastly, below is the overall average of the metrics from the cross validation.



The model produced the following:
1. Accuracy: The model had an overall 70% of the predictions that were correctly classfied. This is misleading however because there are some imbalances in the dataset classes
2. Sensitivty (Recall) is the true positive rate. 66% of people who sought treatment were correctly classified
3. Specificity is also known as the true negative rate. 76% of people who didn't seek treatment were correctly identified
4. Precision is the true predictive value. 73% of those flagged for needing treatment were correct.
5. F1 Score is the balance between Precision and Recall. So 69% were correctly idenfied as treatment seekers and with the same rate of false alarms

Results Review

1. Chi-square tests showed a significant association between benefits awareness and seeking treatment ($p < 0.05$).

2. Logistic regression found gender, awareness of benefits, and family history to be significant predictors ($p < 0.05$). Employees aware of benefits were 1.49 times more likely to seek treatment. Additionally, employees with a family hisotry of mental health issues were 4.9 times more likely to seek treatment

3. Age group, number of employees and remote work were dropped from the stepwise variable selection in the model meaning that they are likely not as highly associated with seeking treatment than factors including gender, awareness of benefits and family history

4. Using k-folds cross validation, the model does a decent job in predicting those that sought treatment at 66% tru postive rate, and those that did not seek treatment at 76% true negative rate

Conclusion

These results show that promoting awareness of benefits is key to increasing mental health treatment uptake. Employees who were aware of benefits provided by the company were 1.49 times likely to seek treatment. Additionally, those with family history of mental health issues seem to be more aware of the issues and are even more likely (4.9x) to seek treatment

Gender-specific approaches may also improve outreach effectiveness. Women were 2.2x more likely to seek treatment. This shows that societal norms play a large part in how men and women percieve and act around mental health. It seems that there could be more in the data that shows us why men seem to not seek treatment.

Number of Employees did not significantly predict treatment behavior in this sample nor did age group but there may be more insight to be gathered from the data especially if we look at each category more granularly

Recommendations/Next Steps

Encourage HR transparency on available mental health benefits.
Tailor mental health communication efforts to different demographic groups.