MILESTONE 2

MENTAL HEALTH PREDICTION

INTRODUCTION:

Mental health is a critical issue in society that has been gaining increased attention in recent years. Mental health conditions, such as depression, anxiety, and bipolar disorder, are common and can have significant impacts on individuals, families, and communities. It is important to understand the prevalence of mental health conditions in different populations and to identify the factors that contribute to the development of these conditions. This understanding can help in the development of effective prevention and intervention strategies.

In this report, we analyze a dataset that measures attitudes towards mental health and the frequency of mental health disorders in the tech workplace. The dataset was collected through a survey conducted in 2014, with additional data collected in 2016. Our goal is to develop a predictive model that can help identify individuals who may be at risk for mental health conditions based on demographic and workplace factors. We will also explore how attitudes towards mental health and the frequency of mental health disorders vary by geographic location and identify the strongest predictors of mental health illness or certain attitudes towards mental health in the workplace.

BACKGROUND:

Mental health conditions are a significant public health concern. According to the World Health Organization (WHO), an estimated 264 million people worldwide suffer from depression, and around 450 million people suffer from some form of mental or neurological disorder. Mental health conditions can have a significant impact on an individual's physical, emotional, and social well-being, as well as their ability to work and participate in daily activities.

The tech industry is known for its fast-paced and high-stress work environments, which can have negative impacts on mental health. According to a 2019 study by Blind, a workplace social network, tech workers are more likely to report mental health issues than workers in other industries. The study found that 57% of tech workers reported experiencing symptoms of burnout, compared to 43% of non-tech workers.

Despite the high prevalence of mental health conditions in the tech industry, there is still a stigma attached to mental illness, which can prevent individuals from seeking help. According to the same Blind study, only 44% of tech workers reported feeling comfortable discussing mental health with their managers or colleagues. This highlights the need for increased awareness and education about mental health issues in the workplace.

To address this issue, organizations and policymakers need to have a better understanding of the prevalence and factors that contribute to mental health conditions in the tech workplace. This report aims to contribute to this understanding by analyzing a dataset that measures attitudes toward mental health and the frequency of mental health disorders in the tech workplace. By identifying the strongest predictors of mental health illness and attitudes towards mental health, we can develop targeted interventions to support mental health in the tech workplace.

DATA SET DESCRIPTION:

Quick info on the Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 27 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Timestamp                  1259 non-null   object
 1   Age                        1259 non-null   int64
 2   Gender                     1259 non-null   object
 3   Country                    1259 non-null   object
 4   state                      744 non-null    object
 5   self_employed              1241 non-null   object
 6   family_history             1259 non-null   object
 7   treatment                  1259 non-null   object
 8   work_interfere             995 non-null    object
 9   no_employees               1259 non-null   object
 10  remote_work                1259 non-null   object
 11  tech_company               1259 non-null   object
 12  benefits                   1259 non-null   object
 13  care_options               1259 non-null   object
 14  wellness_program           1259 non-null   object
 15  seek_help                  1259 non-null   object
 16  anonymity                  1259 non-null   object
 17  leave                      1259 non-null   object
 18  mental_health_consequence  1259 non-null   object
 19  phys_health_consequence    1259 non-null   object
 20  coworkers                  1259 non-null   object
 21  supervisor                 1259 non-null   object
 22  mental_health_interview    1259 non-null   object
 23  phys_health_interview      1259 non-null   object
 24  mental_vs_physical         1259 non-null   object
 25  obs_consequence            1259 non-null   object
 26  comments                   164 non-null    object
dtypes: int64(1), object(26)
memory usage: 265.7+ KB
```

- There are a total of **26** columns in the dataset.
- We see that except the age column, all the columns are of object datatype.
- Comment column seems to contain the most number **(70%)** of null values, which makes sense because it was an optional text box so it's reasonable to expect that many (most) respondents would leave it blank.
- We will be dropping the timestamp column because it contains the date, month, year, and time the respondent took this questionnaire, which is irrelevant to us.
- The state column also contains a lot of null values. We'll dig deeper into that.

Unique Values:

```
United States            751
United Kingdom           185
Canada                    72
Germany                   45
Ireland                   27
Netherlands               27
Australia                 21
France                    13
India                     10
New Zealand                8
Switzerland                7
Italy                      7
Sweden                     7
Poland                     7
Brazil                     6
South Africa               6
Belgium                    6
Israel                     5
Bulgaria                   4
Singapore                  4
Mexico                     3
Russia                     3
Finland                    3
Austria                    3
Finland                    3
Austria                    3
Colombia                   2
Denmark                    2
Greece                     2
Portugal                   2
Croatia                    2
Latvia                     1
Norway                     1
Uruguay                    1
Costa Rica                 1
Georgia                    1
Spain                      1
Zimbabwe                   1
Czech Republic             1
Slovenia                   1
Romania                    1
Japan                      1
Bahamas, The               1
Thailand                   1
Bosnia and Herzegovina     1
Philippines                1
Hungary                    1
China                      1
Moldova                    1
Nigeria                    1
Name: Country, dtype: int64
```

Important Inferences:

- It will be really misleading to conclude that a certain country faces more problems with the mental health of employees because around **60%** of the people belong to The US.
- Moreover there are a lot of countries that have only one respondent
- The country column thus becomes pointless. We will be dropping this
- A quick look at the states suggests that it is applicable for the ones only in The US, so we'll drop it as well.

Age Groups:

```
The dataset contains different age groups including:

[         37          44          32          31          33          35
          39          42          23          29          36          27
          46          41          34          30          40          38
          50          24          18          28          26          22
          19          25          45          21         -29          43
          56          60          54         329          55 99999999999
          48          20          57          58          47          62
          51          65          49       -1726           5          53
          61           8          11          -1          72]


The different gender notations used in our dataset are:

['Female' 'M' 'Male' 'male' 'female' 'm' 'Male-ish' 'maile' 'Trans-female'
 'Cis Female' 'F' 'something kinda male?' 'Cis Male' 'Woman' 'f' 'Mal'
 'Male (CIS)' 'queer/she/they' 'non-binary' 'Femake' 'woman' 'Make' 'Nah'
 'All' 'Enby' 'fluid' 'Genderqueer' 'Female ' 'Androgyne' 'Agender'
 'cis-female/femme' 'Guy (-ish) ^_^' 'male leaning androgynous' 'Male '
 'Man' 'Trans woman' 'msle' 'Neuter' 'Female (trans)' 'queer'
 'Female (cis)' 'Mail' 'cis male' 'A little about you' 'Malr' 'p' 'femail'
 'Cis Man' 'ostensibly male, unsure what that really means']
```

Important Inferences:

How can age be negative? And age **below 15 years**? Are they even legally allowed to work?

- Regarding gender, people have described themselves as male and female in such different ways!
- Let's get back to our work and correct these responses. While this may not be the best way, we will be using this approach for the gender column: We will be renaming and combining all the categories that mean the same into one.

- Male, or cis Male, means born as male and decides to be male.
- Female, or cis Female, means born as female and decide to be female.
- Other, is a word that describes sexual and gender identities other than straight and cisgender. Lesbian, gay, bisexual, and transgender people may all identify with the word other.

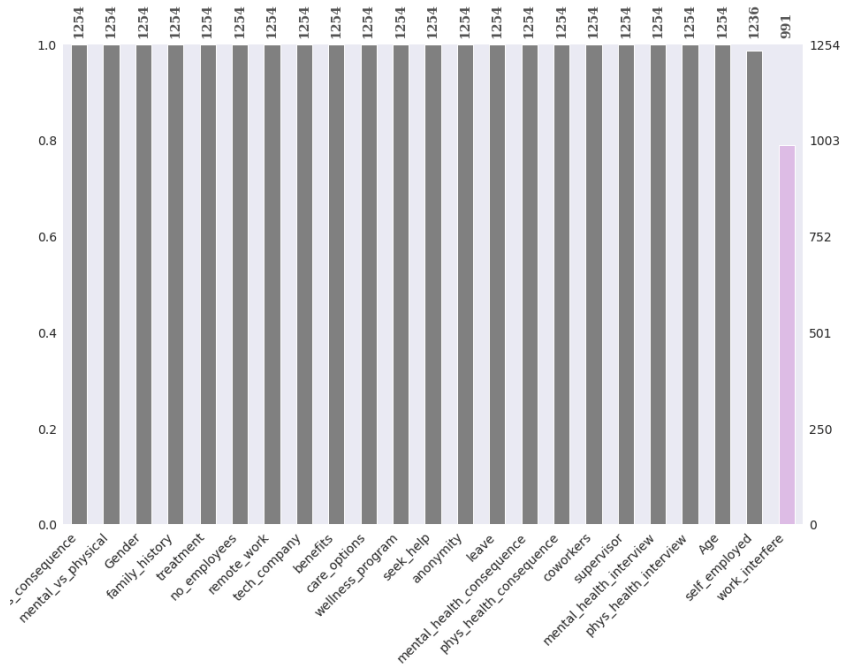
After Cleaning the Ages column & Sex Column:

```
array([37, 44, 32, 31, 33, 35, 39, 42, 23, 29, 36, 27, 46, 41, 34, 30, 40,
       38, 50, 24, 18, 28, 26, 22, 19, 25, 45, 21, 43, 56, 60, 54, 55, 48,
       20, 57, 58, 47, 62, 51, 65, 49,  5, 53, 61,  8, 11, 72])
```

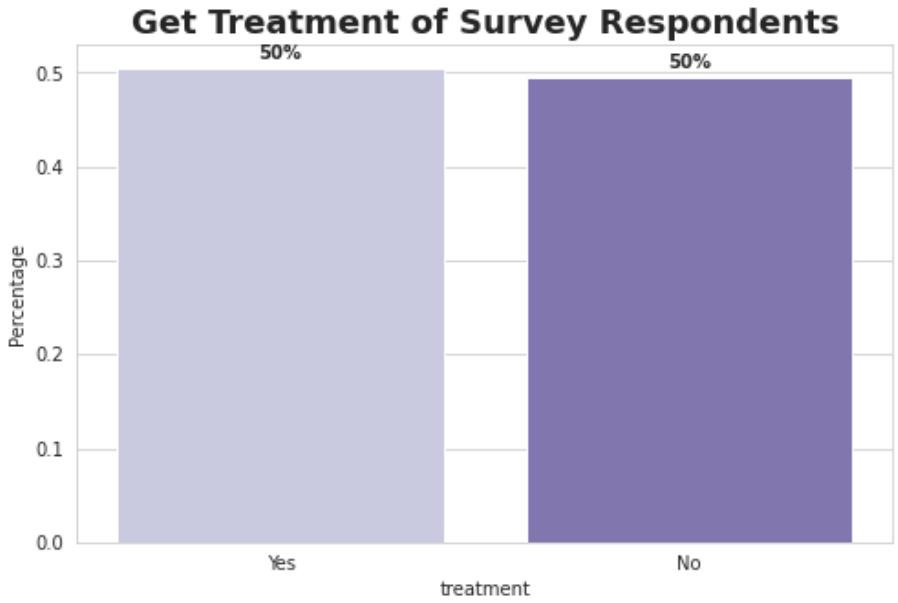
```
Male      988
Female    247
Other      19
Name: Gender, dtype: int64
```

Missing Values:

There's only one column which is **'work_interfere'** remaining that contains null values. For now we will proceed without any imputation.
Actually, there's another column, **'self_employed'** which contains around 18 null values which we failed to notice at first.



Exploratory data analysis:-

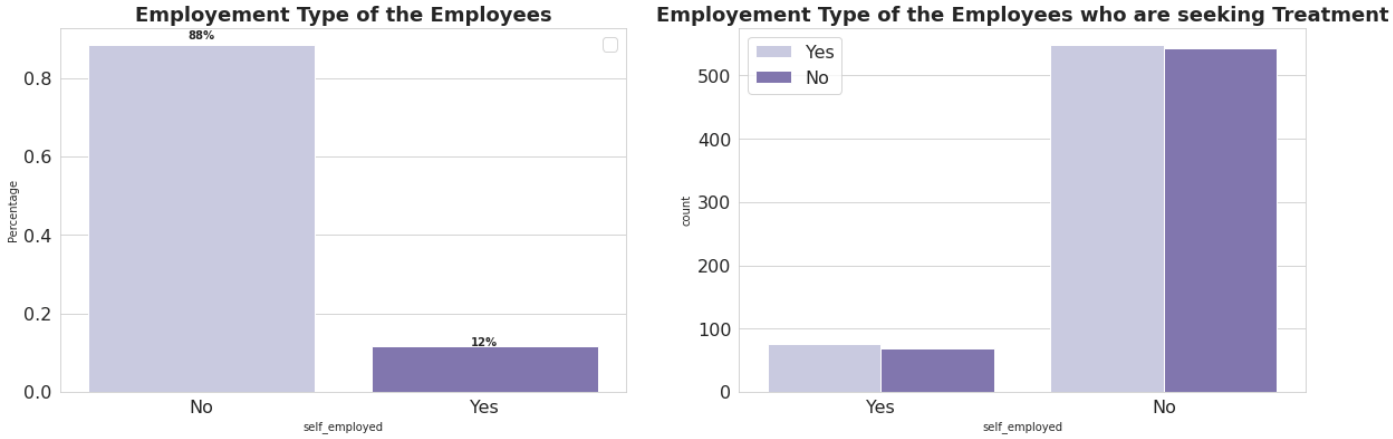


This is the respondent's result of the question, 'Have you sought treatment for a mental health condition?'

This is our target variable. Looking at the first graph, we see that the percentage of respondents who want to get treatment is exactly 50%. Workplaces that promote mental health and support people with mental disorders are more likely to have increased productivity, reduce absenteeism, and benefit from associated economic gains. If employees enjoy good mental health, employees can:

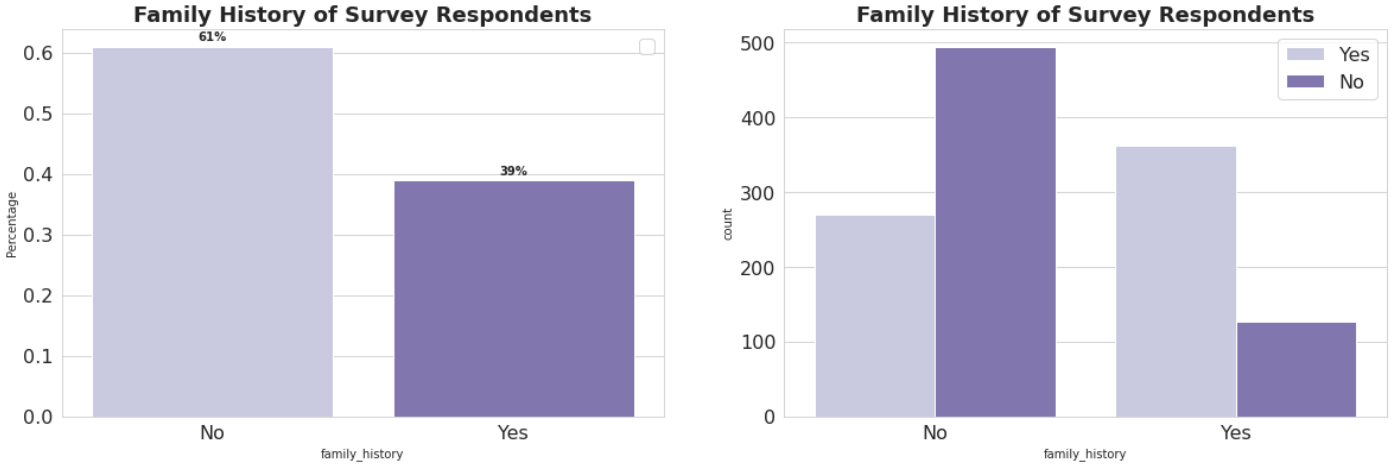
- Be more productive

- Take active participation in employee engagement activities and make better relations; both at workplace and personal life.
- Be more joyous and make people around them happy.



This is the respondent's answer to the question, 'Are you self-employed?'.

We see that the number of people who are self-employed are around 10%. Most of the people who responded to the survey belonged to working class. We also see that though there is a vast difference between people who are self employed or not, the number of people who seek treatment in both the categories is more or less similar.

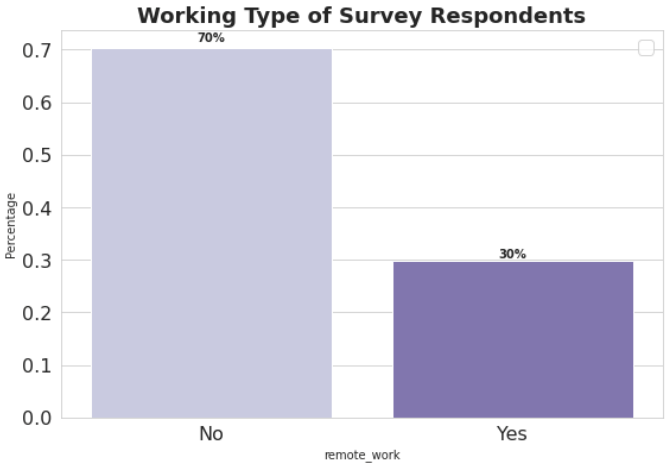


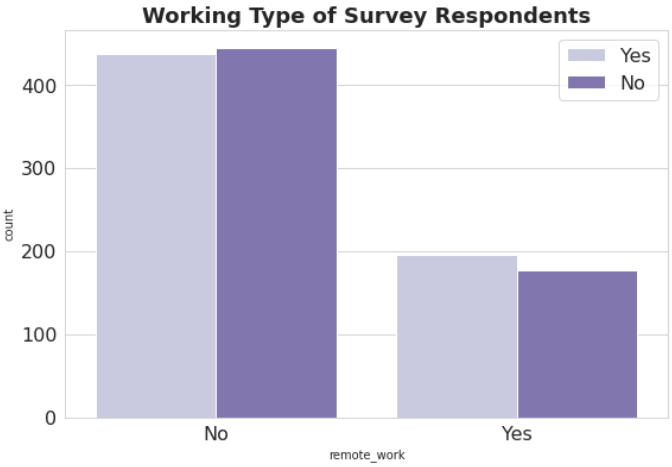
'Do you have a family history of mental illness?'.

From close to 40% of the respondents who say that they have a family history of mental illness, the plot shows that they significantly want to get treatment rather than without a family history. This is acceptable, remember the fact that people with a family history pay more attention to mental illness. Family history is a significant risk factor for many mental health disorders.

'If you have a mental health condition, do you feel that it interferes with your work?'.

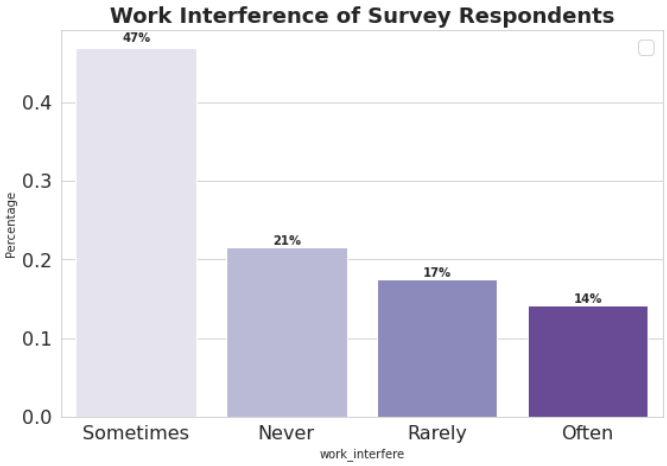
- On seeing the first graph we conclude that around 48% of people say that sometimes work interefers with their mental health. Now **'Sometimes'** is a really vague response to a question, and more often than not these are the people who actually face a condition but are too shy/reluctant to choose the extreme category.
- Coming to our second graph, we see that the people who chose **'Sometimes'** had the highest number of people who actually had a mental condition. Similar pattern was shown for the people who belonged to the **'Often category'*.
- But what is more surprising to know is that even for people whose mental health **'Never'** has interfered at work, there is a little group that still wants to get treatment before it become job stress. It can be triggered by a variety of reasons like the requirements of the job do not match the capabilities, resources, or needs of the worker.

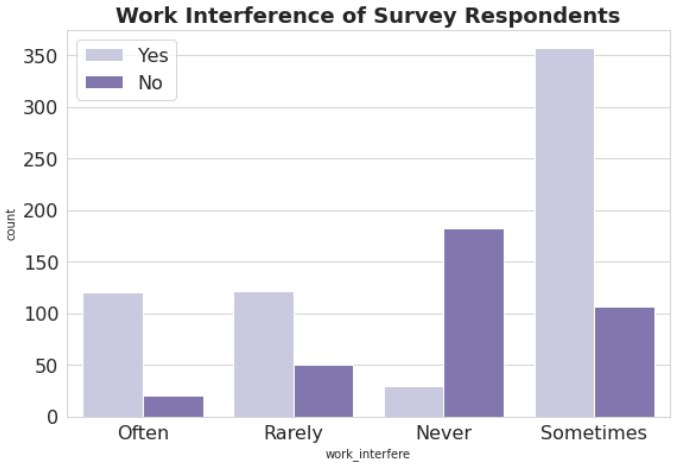


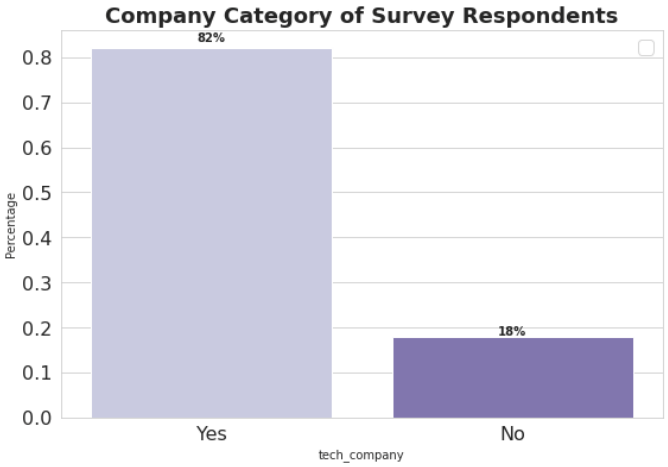


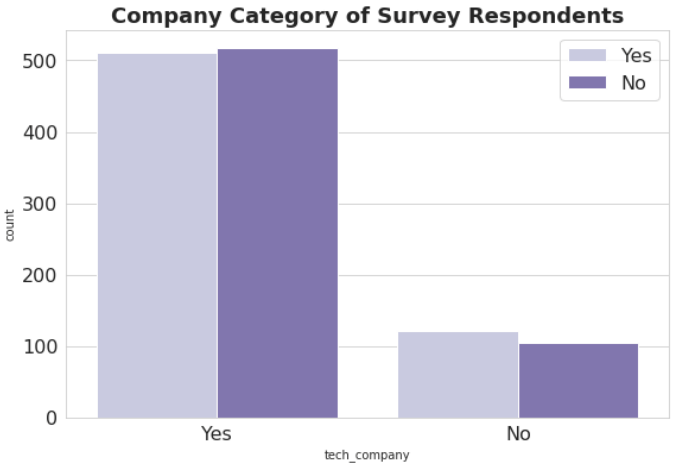
'Do you work remotely (outside of an office) at least 50% of the time?'.

Around 70% of respondents don't work remotely, which means the biggest factor of mental health disorder came up triggered on the workplace. On the other side, it has slightly different between an employee that want to get treatment and don't want to get a treatment. The number of people who seek treatment in both the categories is more or less similar and it does not affect our target variable.



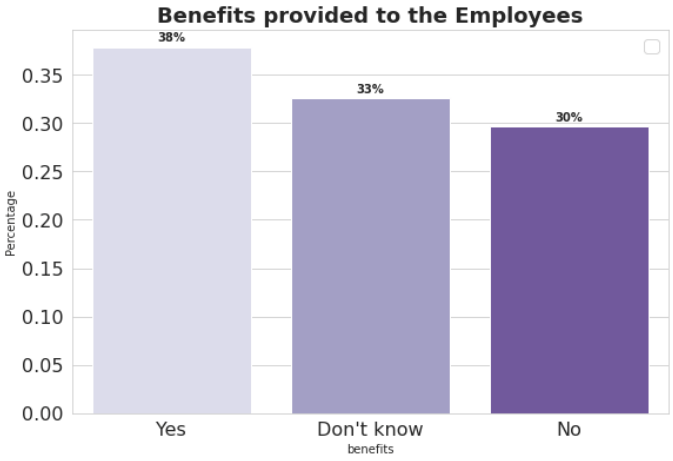


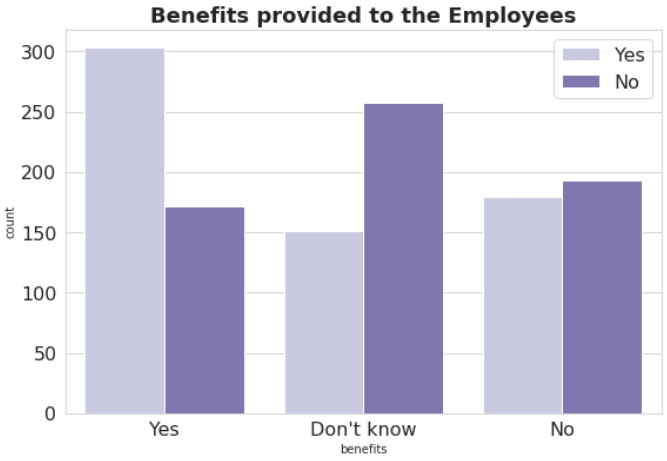




'Is your employer primarily a tech company/organization?'.

- Although the survey was specifically designed to be conducted in the tech field, there are close to 18% of the companies belong to the non-tech field. However, looking at the second graph, one may conclude that whether a person belongs to the tech field or not, mental health still becomes a big problem.
- However, on a deeper look we find that the number of employees in the tech sector who want to get treatment is slightly lower than the ones who don't. But in the non-tech field, the situation gets reversed.





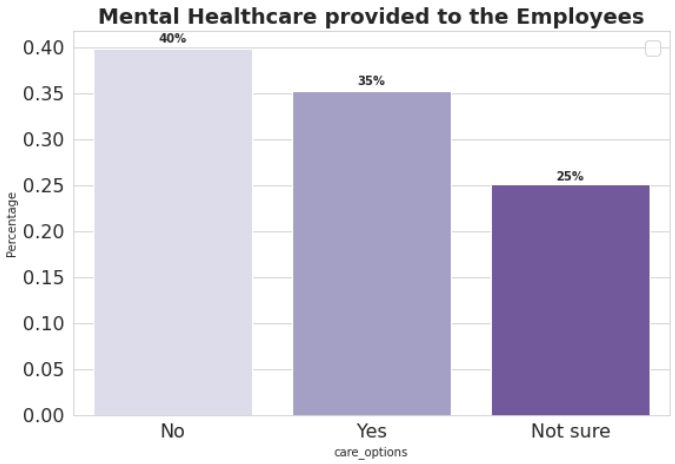
This was the respondent's answer to the question

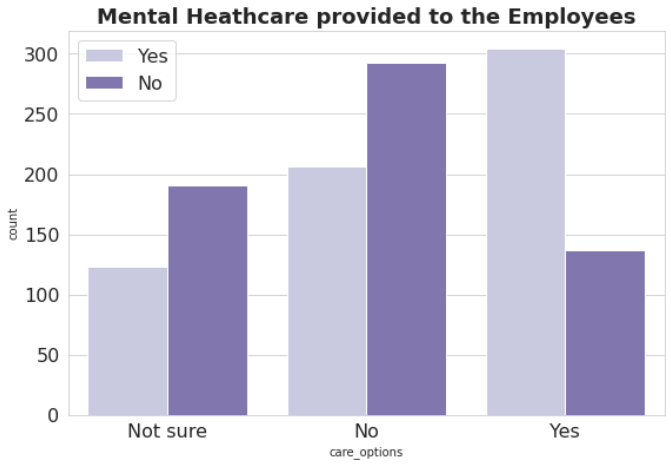
'Does your employer provide mental health benefits?'.

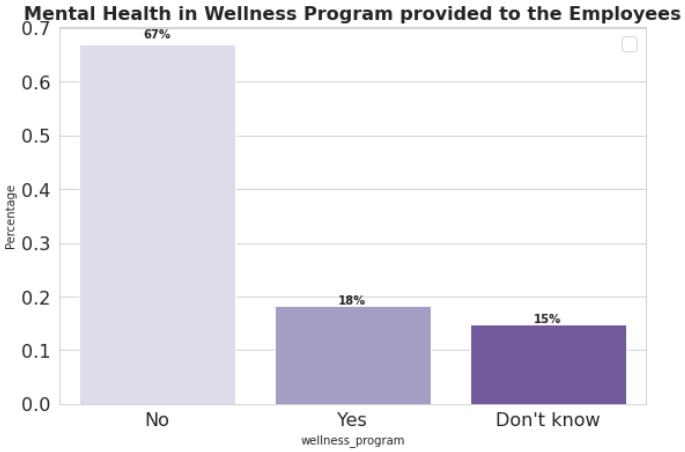
- We see that around 38% of the respondents said that their employer provided them mental health benefits, whereas a significant number (32%) of them didn't even know whether they were provided this benefit.
- Coming to the second graph, we see that for the people who **YES** said to mental health benefits, around 63% of them said that they were seeking medical help.
- Surprisingly, the people who said **NO** for the mental health benefits provided by the company, close to 45% of them who want to seek mental health treatment.

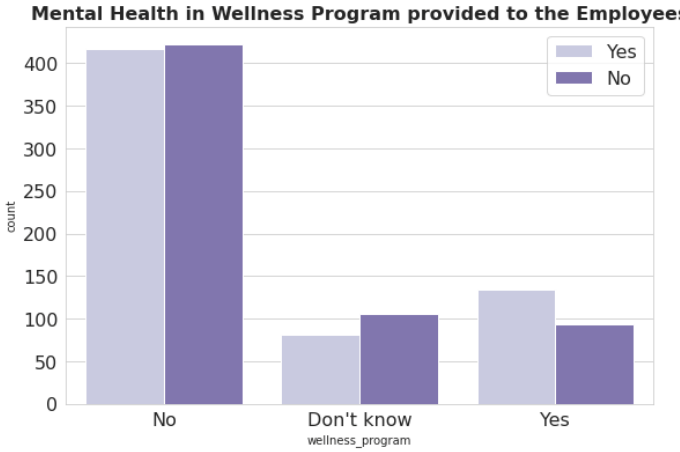
'Do you know the options for mental health care your employer provides?'.

Since this graph is more or less similar to the benefits one, we won't be discussing it in more detail.



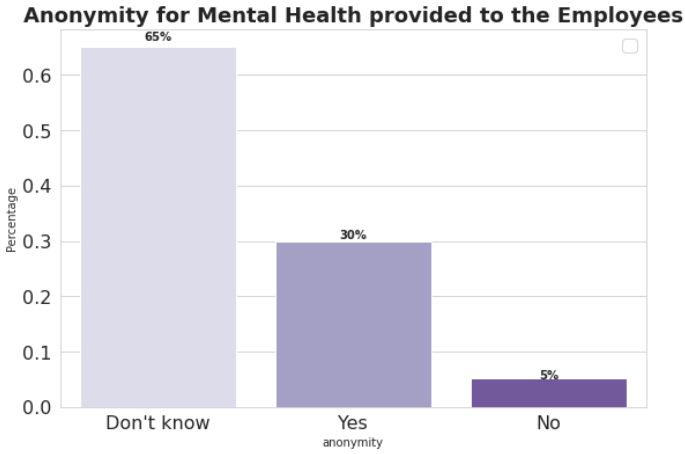


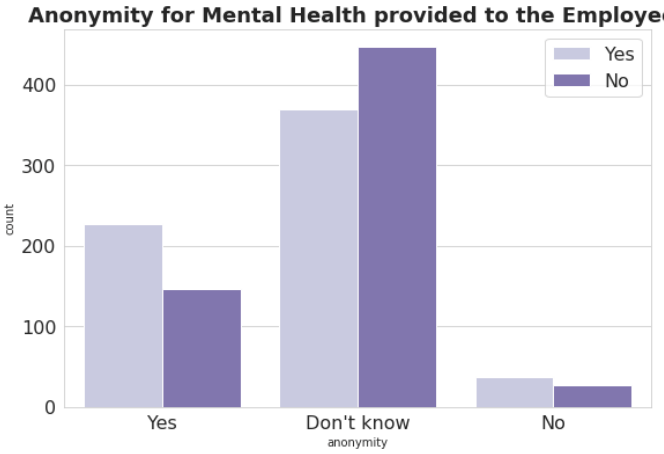




'Has your employer ever discussed mental health as part of an employee wellness program?'.

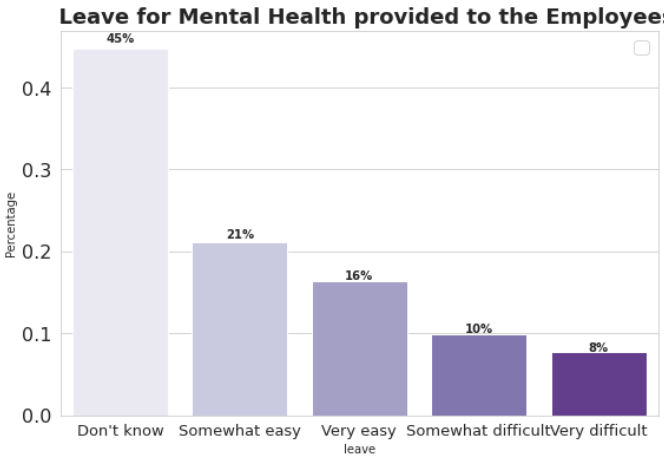
- About 19% of the respondents say **YES** about becoming a part of the employee wellness program and out of those 60% of employees want to get treatment.
- One shocking revelation is that more than 65% of respondents say that there aren't any wellness programs provided by their company. But close to half of those respondents want to get treatment, which means the company needs to fulfill its duty and provide it soon.

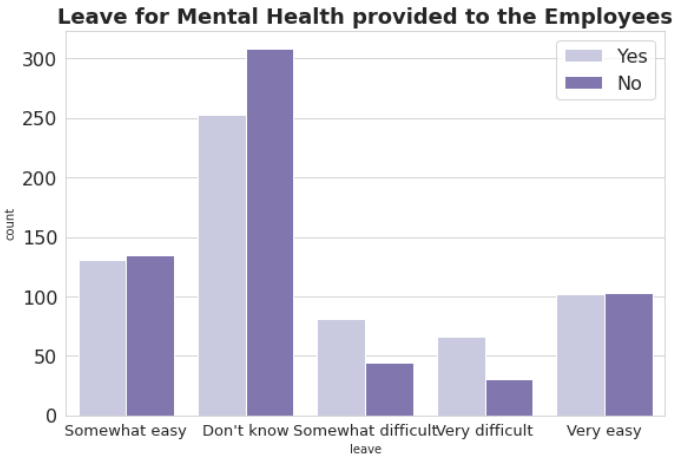




'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?'.

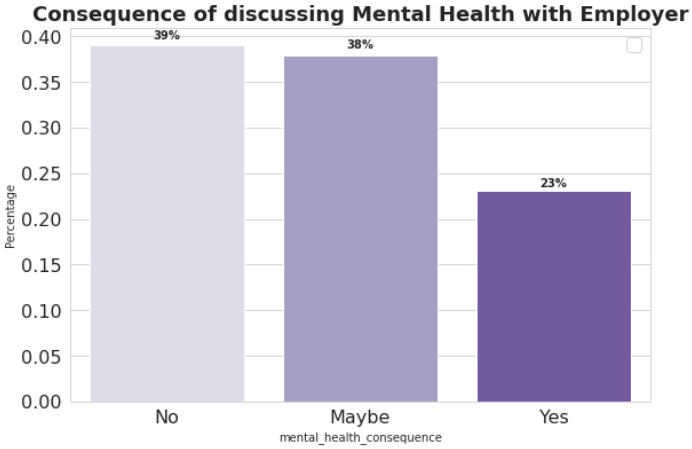
- Around 65% of the people were not aware whether anonymity was provided to them and 30% said yes to the provision of anonymity by the company.
- Looking at the second graph, we see that out of the people who answered yes to the provision of anonymity, around 60% of them were seeking help regarding their mental condition. Possible reasoning for this may be that the employee feels that the company has protected his/her privacy and can be trusted with knowing the mental health condition of it's workers. The most basic reason behind hiding this from the fellow workers can be the social stigma attached to mental health.

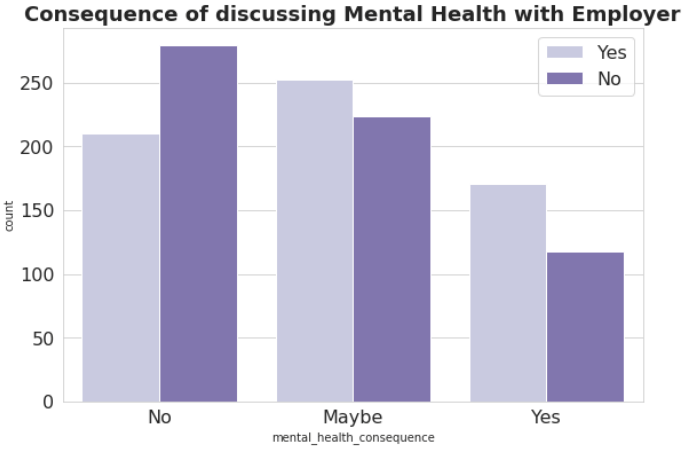




'How easy is it for you to take medical leave for a mental health condition?'

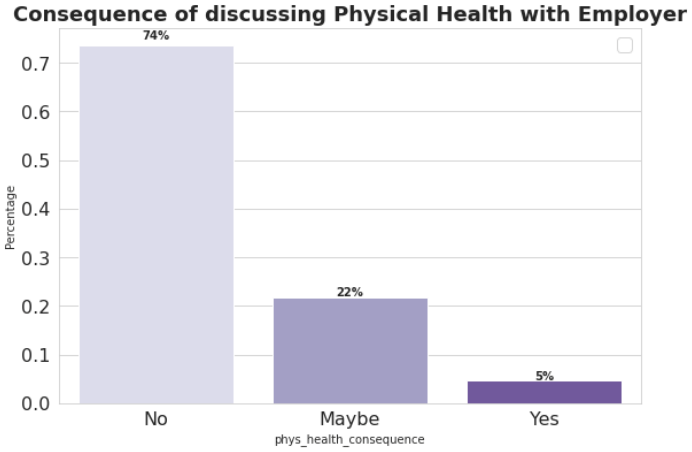
- While close to 50% of the people answered that they did not know about it, suprisingly around 45% of those people sought help for their condition.
- A small percent of people (around 8%) said that it was very difficult for them to get leave for mental health and out of those, 75% of them sought for help.
- Employees who said it was 'somewhat easy' or 'very easy' to get leave had almost 50% people seeking medical help.

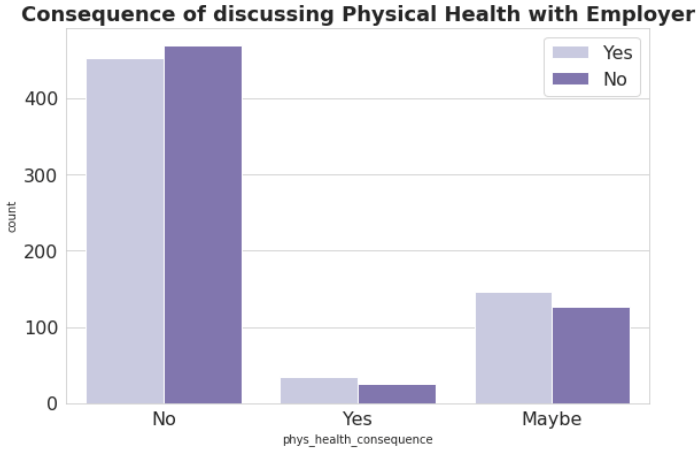




'Do you think that discussing a mental health issue with your employer would have negative consequences?'.

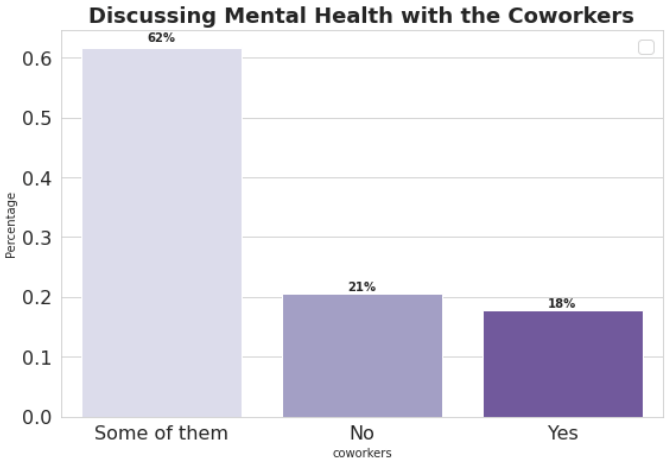
- Around same number of people (around 40% each) answered **Maybe** as well as **No** for the negative impact of discussing mental health consequences with the employer and about 23% said **Yes** to it.
- 23% is a significant number who feel that discussing their mental health might create a negative impact on their employer. This may be because of the stigma, decreased productivity, impact on promotions or any other preconcieved notion.
- It is nice to know that out of the people who answered No, there were only around 40% of the people who actually sought after help, whereas in both the other categories, it is more than 50%.

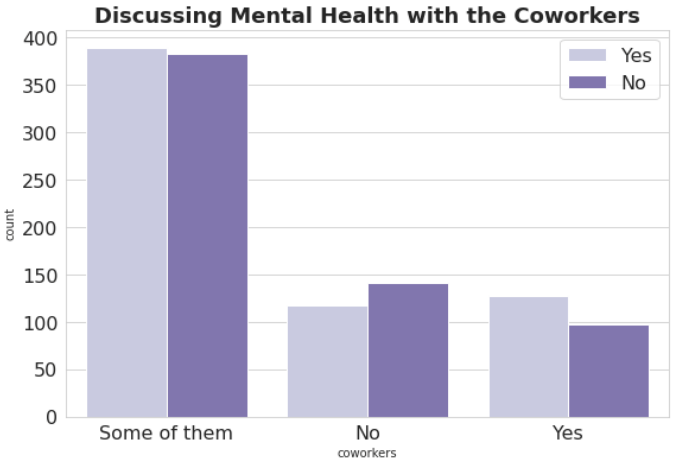




'Do you think that discussing a physical health issue with your employer would have negative consequences?'

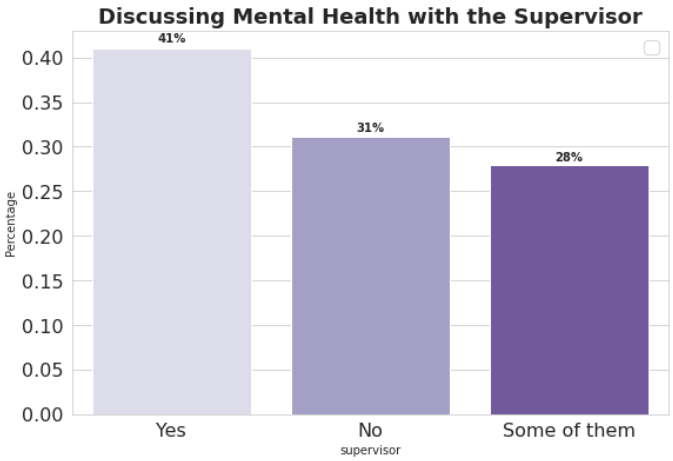
- There is a difference between the responses to the same question regarding mental and physical health. More than 70% of the employees believe that their physical health does not create a negative impact on their employer and only 5% of them believe that it does.
- While it may be incorrect for us to draw any conclusions about whether they seek mental help on the basis of their physical condition because it is more or less the same for all three categories, we must keep in mind about **how different mental and physical health is treated as a whole.**

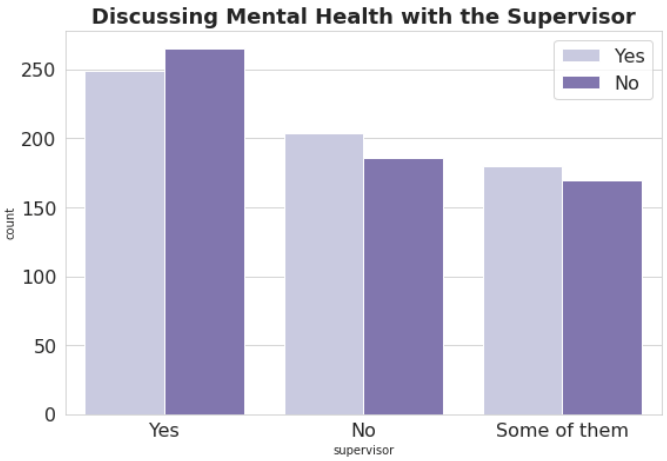




'Would you be willing to discuss a mental health issue with your coworkers?'

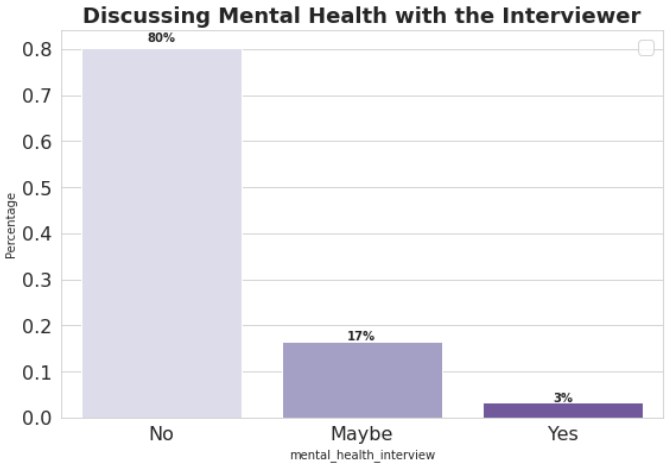
- Around 62% of the employees said that they might be comfortable discussing some type of mental problems with their coworkers, and out of them around 50% actually sought medical help.
- 20% of the employees believed that discussing mental health with their coworkers wasn't a good option for them.

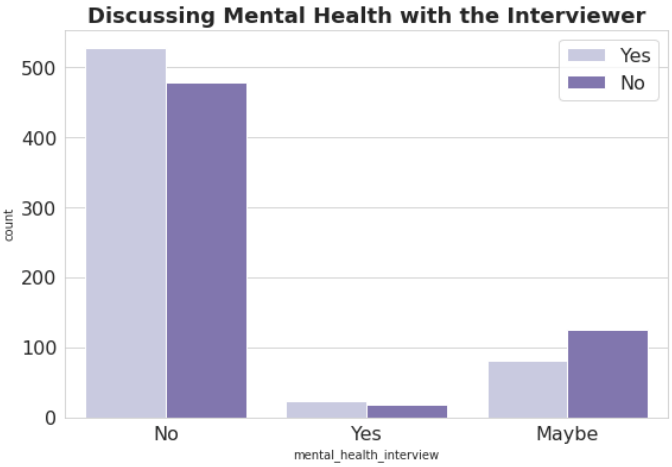




'Would you be willing to discuss a mental health issue with your direct supervisor(s)?'.

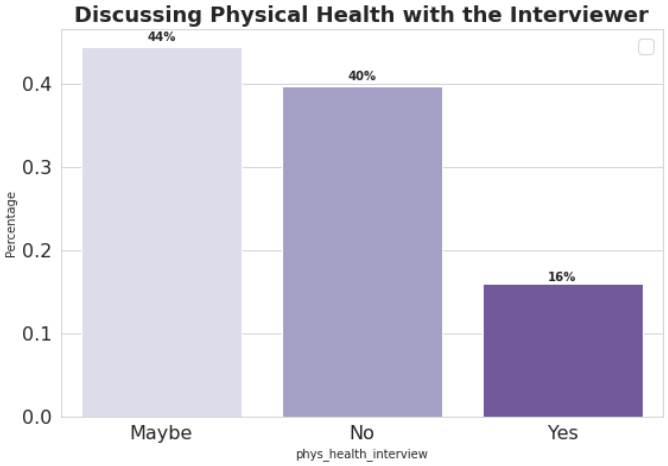
- This graph is quite different from the one of the coworker. Here, around 40% of the workers believe in sharing their mental health with their supervisors. This may have something to do with their performance etc.
- Looking at the second graph, employees who actually sought for help regarding their mental health was more or less similar for all the three categories.

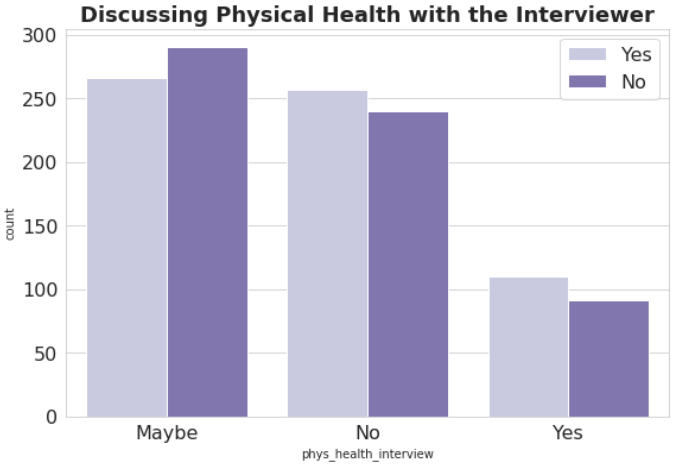




'Do you think that discussing a mental health issue with your employer would have negative consequences?'.

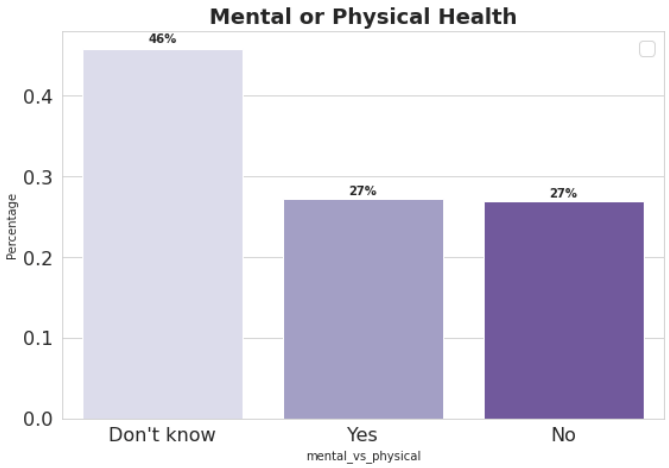
- As our intution might suggest us, 80% of the respondents believe that it is a good option to discuss your mental health with the future employer. This is actually a good thing! This might not have been the case 15 years ago.
- While around 15% of the candidates seem confused about whether they should be discussing their mental conditions with the future employer or not, less than 5% think that it may not be a good option discussing it.

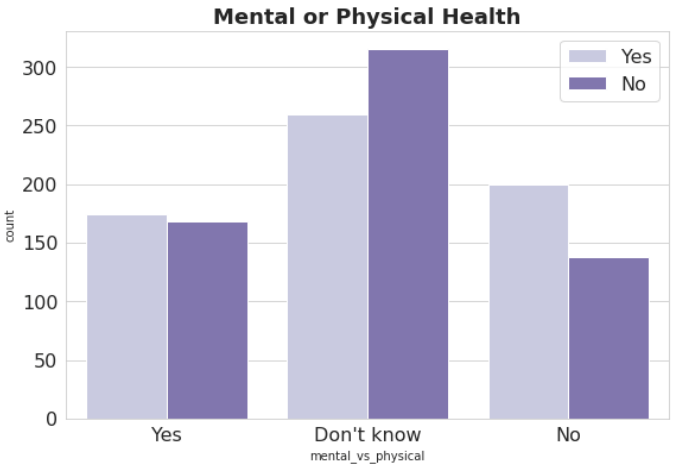




'Would you bring up a physical health issue with a potential employer in an interview?'.

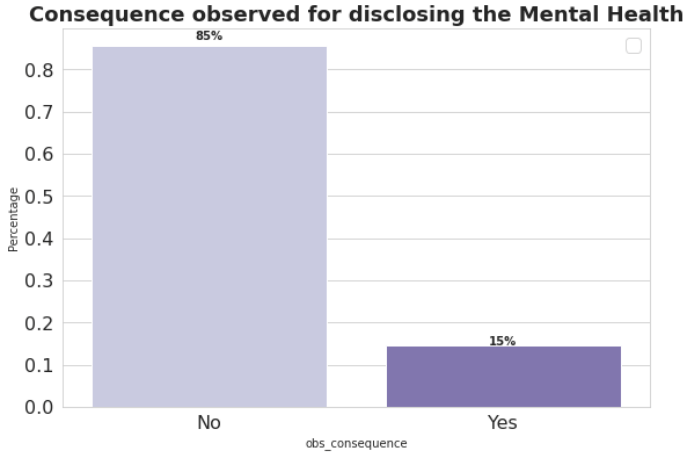
- While a majority of the people are still dubious about discussing their physical health condition with the future employer, however, close to 17% believe that there is no issue in discussing their physical health conditions.
- Around 50% of the people still remain confused about whether it is a good option to discuss their condition or not.

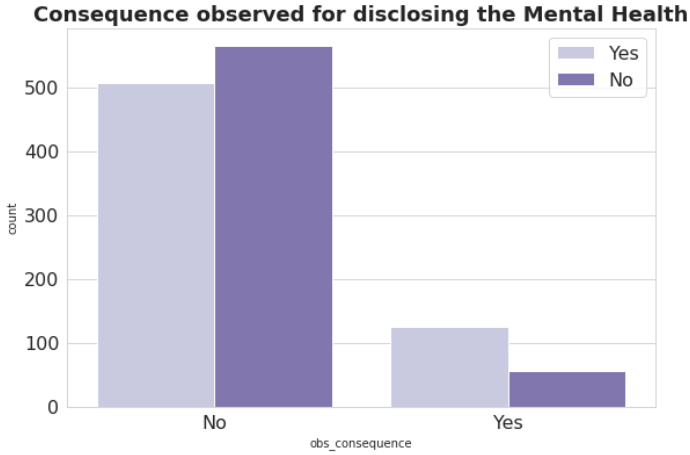




'Do you feel that your employer takes mental health as seriously as physical health?'

- While close to 50% people said that they didn't know, the number of people who answered **Yes** as well as **No** were completely equal.
- For the people who answered Yes as well as the ones who answered No, more than 505 of them sought after medical help for their mental health, whereas it was not the case for the one's belonging to the 'Don't know' category.

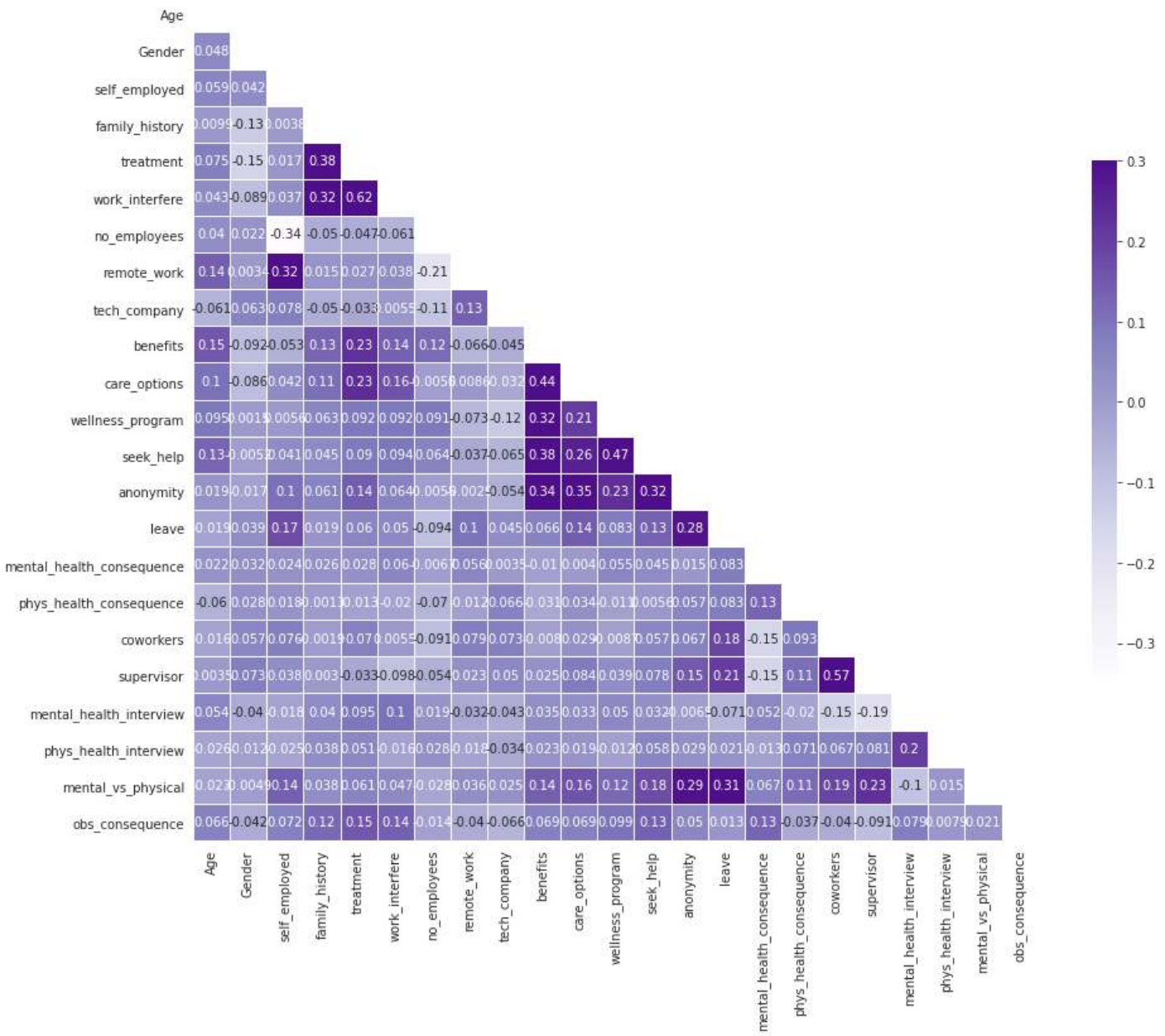




This was the respondent's answer to the question, **'Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?'**

- Majority (85%) of the people, answered **No** to this question. This is quite important to note that IT being an organized sector, follows strict guidelines for employee satisfaction, etc. Thus, we didn't come across any major issues regarding the employer behavior as such.

CORRELATION MATRIX:



MEASURABLE METRICS AND ML ALGORITHMS/ UTILITIES:

ML Algorithms planning to implement:

Logistic Regression is a type of statistical model used to predict a binary outcome, such as whether a customer will make a purchase or not. It works by fitting a logistic function to a set of input data, which allows it to estimate the probability of a certain outcome. Logistic regression can be used for both classification and regression problems, and is particularly useful when the data is linearly separable.

KNearestNeighbours (KNN) is a simple and effective classification algorithm that works by finding the K nearest data points to a given input point and assigning the input point the most common class label among its K nearest neighbors. KNN is a non-parametric algorithm, which means it does not make any assumptions about the distribution of the data. KNN is often used in image recognition, natural language processing, and recommendation systems.

Decision Tree Classifier is a popular machine learning algorithm that works by recursively splitting the data into subsets based on the most significant feature at each step, until the subsets are as pure as possible (i.e., all elements in a subset belong to the same class). Decision trees are particularly useful when the data has a hierarchical structure, or when the data contains a mix of categorical and numerical features.

Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of the model. Each decision tree in the random forest is trained on a random subset of the data and a random subset of the features. The final prediction is then made by aggregating the predictions of all the trees. Random forests are known for their ability to handle noisy data and avoid overfitting.

Gradient Boost Classifier is another ensemble learning algorithm that works by combining multiple weak models into a strong model. Unlike random forests, gradient boosting builds the models sequentially, with each new model fitting the residual errors of the previous model. Gradient boosting is often used in regression problems, as well as classification problems.

AdaBoost Classifier is a popular boosting algorithm that works by iteratively adjusting the weights of the misclassified data points to focus on the harder examples. In each iteration, a new weak model is trained on the weighted data, and the weights are adjusted again. The final prediction is then made by aggregating the predictions of all the weak models, weighted by their performance. AdaBoost is often used in face detection and text classification.

XGB Classifier (Extreme Gradient Boosting) is a gradient boosting algorithm that uses a tree-based model and a more advanced regularization technique than traditional gradient boosting. XGB is known for its speed and scalability, and is often used in data mining, text mining, and natural language processing. XGB also has a built-in feature selection mechanism, which can be useful when dealing with high-dimensional data.

Metrics planning to use:

Accuracy: This measures the proportion of correct predictions made by the model. It is a simple and intuitive metric, but can be misleading if the data is imbalanced or the cost of false positives/negatives is not equal.

Precision: This measures the proportion of true positives among all predicted positives. It is a useful metric when the goal is to minimize false positives (e.g., in spam filtering).

Recall: This measures the proportion of true positives among all actual positives. It is a useful metric when the goal is to minimize false negatives (e.g., in disease diagnosis).

F1 score: This is the harmonic mean of precision and recall, and provides a balanced measure of both metrics. It is often used when both false positives and false negatives are equally important.

ROC AUC: This measures the trade-off between true positive rate and false positive rate at different probability thresholds. It is a useful metric when the goal is to maximize the overall performance of the model, regardless of the specific threshold used.

Mean Squared Error (MSE): This measures the average squared difference between the predicted and actual values. It is often used in regression problems, where the goal is to minimize the difference between predicted and actual values.

R-squared (R2): This measures the proportion of variance in the target variable that is explained by the model. It is a useful metric when the goal is to understand the overall performance of the model and how well it fits the data.

NEXT STEPS:

1. **Feature engineering:** After data cleaning, I'm planning to extract relevant features from the survey data that might be useful for predicting mental health outcomes. This might involve aggregating or transforming variables, creating new variables, or encoding categorical variables as numerical values.
2. **Split the data:** Split dataset into training and testing sets. Use the training set to train your model and the testing set to evaluate its performance.
3. **Choose a suitable model:** There are many different machine learning models that can be used for predicting mental health outcomes, such as logistic regression, decision trees, random forests, and support vector machines. Choose a model that is appropriate for our specific problem and data.
4. **Train the model:** Use the training set to fit the chosen model to the data. This involves choosing appropriate hyperparameters, such as the learning rate or the number of trees in a random forest.
5. **Evaluate the model:** Once the model is trained, using the testing set to evaluate its performance. Calculate metrics such as accuracy, precision, recall, F1 score, ROC AUC, mean squared error, and R-squared to assess the model's performance.
6. **Tune the model:** If the model's performance is not satisfactory, tuning the hyperparameters or trying a different model.

References

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
3. Eysenbach, G., & Langer, J. (2019). Towards rigorous and reproducible deep learning: The DAWNBench benchmark. arXiv preprint arXiv:1903.10520.
4. Holmes, A. K., & Spann, C. A. (2019). Using machine learning to predict mental health outcomes from clinical records. Journal of Mental Health, 28(6), 617-624.
5. IBM Knowledge Center. (n.d.). Evaluation metrics for classification models. Retrieved from https://www.ibm.com/docs/en/wmla/1.2.3?topic=metrics-evaluation-metrics-classification-models
6. Kaggle. (n.d.). Mental health. Retrieved from https://www.kaggle.com/search?q=mental+health
7. NIH National Institute of Mental Health. (2021). Data archive. Retrieved from https://www.nimh.nih.gov/research/research-funded-by-nimh/data-archive.shtml
8. OpenAI. (2021). AI safety for mental health. Retrieved from https://openai.com/research/ai-safety-for-mental-health/
9. Scikit-learn. (n.d.). Model evaluation: quantifying the quality of predictions. Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html