Data Warehouse

Project Report

Realised by: Elmehdi TERRAF

Professor: Maryem RHANOUI

Table des matières

Introduction

Oftentimes, we discuss mental health issues in the workplace in the context of social responsibility and caring for our fellow human beings. However, making mental health services a priority within a company is also about profit. According to the Centers for Disease Control and Prevention (CDC), depression is responsible for 200 million lost workdays each year, at a cost to employers of $17 to $44 billion.

Mental health affects your emotional, psychological, and social well-being. It affects how we think, feel and act. It also helps determine how we deal with stress, relationships with others and the choices we make. In the workplace, communication and inclusion are key skills for successful teams or employees. The impact of mental health on an organization can result in increased days away from work and decreased productivity and engagement. In the United States, approximately 70% of adults with depression are in the workforce. It is estimated that employees with depression will miss 35 million workdays per year due to mental illness. It is estimated that workers with unresolved depression experience a 35% decrease in productivity, costing employers $105 billion each year.

Basic computer science students at highly competitive institutions, experience an enormous amount of stress. Classes are fast-paced and extremely rigorous, often requiring algorithmic thinking skills as well as strong programming abilities. In large classes, students can feel like a nameless face, especially if they belong to historically underrepresented or marginalized groups. And outside of school, there is often pressured to get a job at a large IT company, whose hiring processes are often extremely competitive and selective.

Some level of stress is good and part of the learning process, but now more than ever, we as computer science students are seeing the stress we experience turn into distress. Students often feel like they are unable to cope with their workload and self-imposed pressures, in addition to the pressures we feel from society, our peers, and our familie

Data Sources

Open Sourcing Mental Illness (OSMI) is a nonprofit corporation dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech and open-source communities. Each year, OSMI conducts its Mental Health in Tech survey, aimed at measuring attitudes toward mental health in the tech workplace and examining the frequency of mental health disorders among tech workers.

We chose to analyze the OSMI Mental Health in Tech survey from 2014,2016,2017,2018,2019,2020 (full data sets available on the OSMI website) as the survey questions and number of respondents changed from year to year, and these surveys had the greatest number of respondents overall (1467 participants).

Research Questions

The key research area of this project is to understand the provisions at tech workplaces to address mental health issues in employees:

- What are the provisions at tech workplaces to address mental health issues in employees?
- Do employees feel that the tech industry supports mental health and wellbeing?
- How many respondents have experienced mental health problems or are currently being diagnosed?
- Are employers providing healthcare benefits for mental health issues?
- Are employees aware of their healthcare benefits and resources?
- How do employers weigh mental health issues at the same level as to other medical issues?
- Are there any leave policies for those suffering from mental health issues?
- How do the tech industries respond to the issue of mental health and how do they deal with it?
- Prediction on whether the person should go for the treatment or not?

Key Performance Indicators (KPIs)

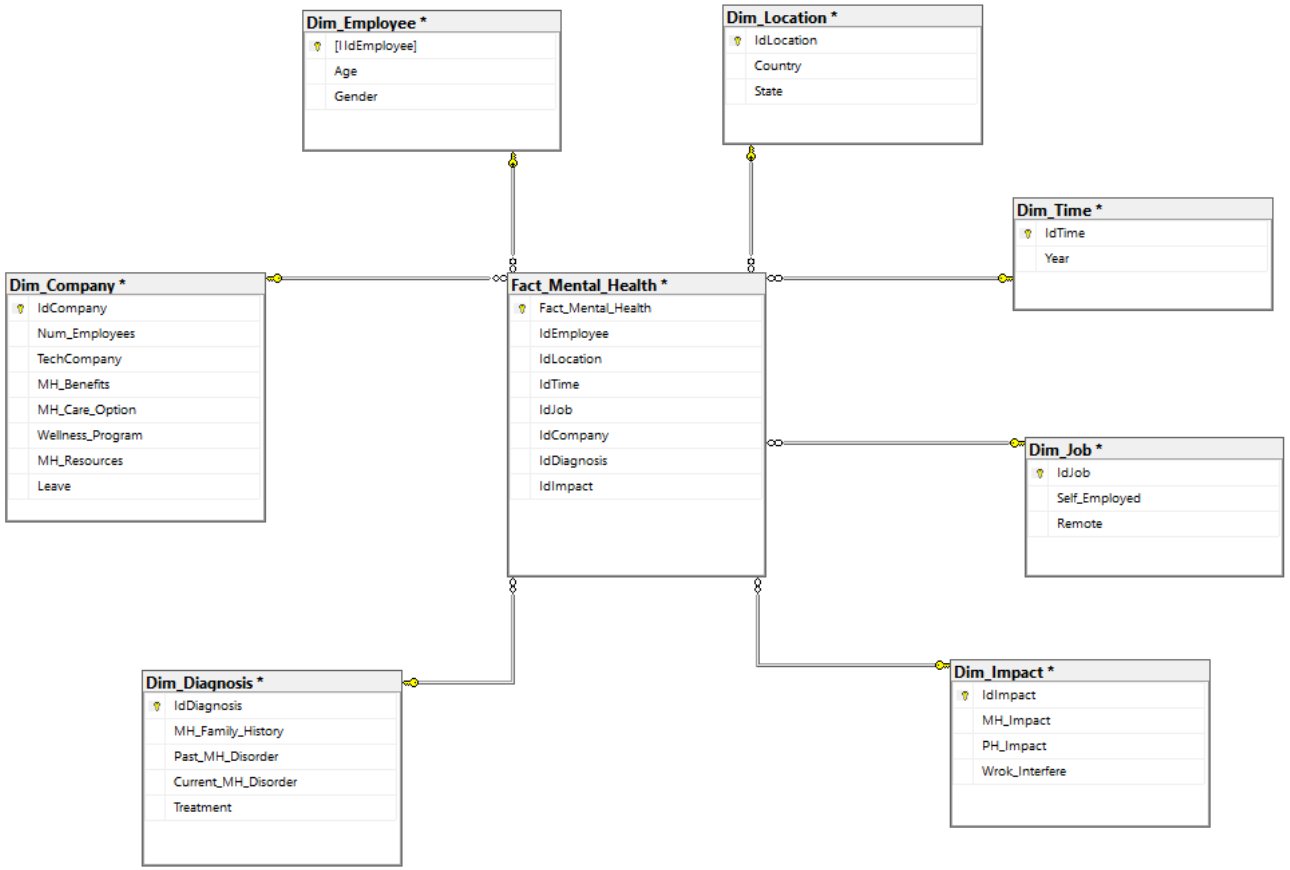
Index	Indicator
I1	Support for mental health in tech
I2	Mental health issues experienced
I3	Medical healthcare coverage & awareness
I4	Importance for mental & physical health
I5	Would mental health problem affect my career?
I6	Currently have a mental health problem
I7	Are companies taking seriously mental health?
I8	Leave policy

Dimensional Matrix

Features / Indicators	MH_Care_ Option	Current_MH_ Disorder	Past_MH_ Disorder	Treatment	MH_Benefits	Wellness_ Program	MH_Resources	MH_Impact	PH_Impact	leave
I1					X	X	X			
I2			X							
I3	X				X	X	X			
I4								X	X	

I5								X		
I6			X	X						
I7	X						X			
I8										X

Data Warehouse Design



Data Preparation & Performing Feature Engineering

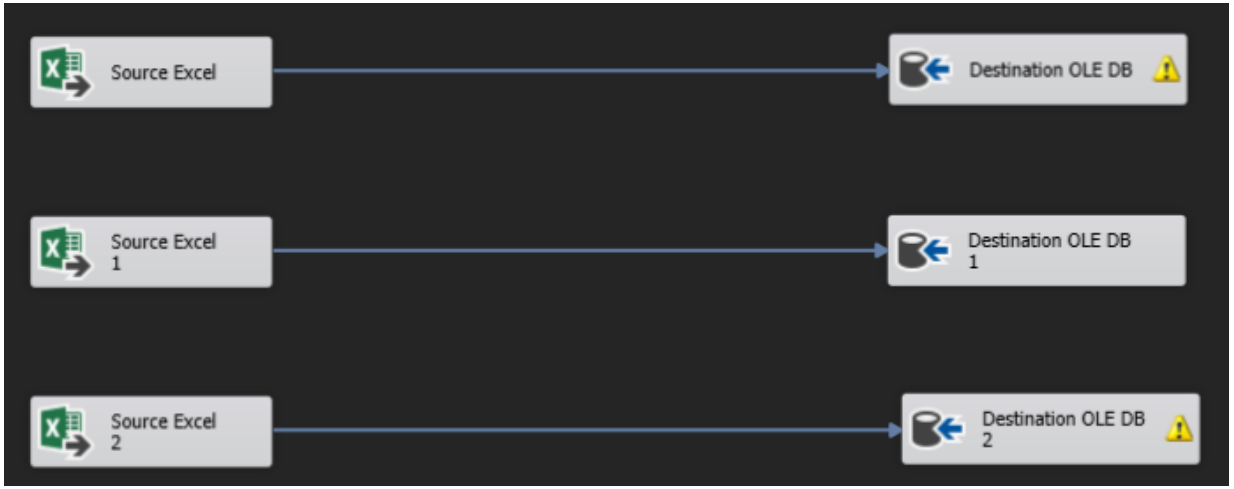
We used Python with the pandas library to clean and prepare our data for analysis.

First, we renamed the columns (the questions of the survey) to significant names, then we handled the nan values we used an interesting sklearn library “SimpleImputer” that replace missing values using a descriptive statistic, in our case we used the most-frequent strategy. Also, we dealt with each column and its unique values, for example the column “gender” has different inputs that’s normally should be either male or female…

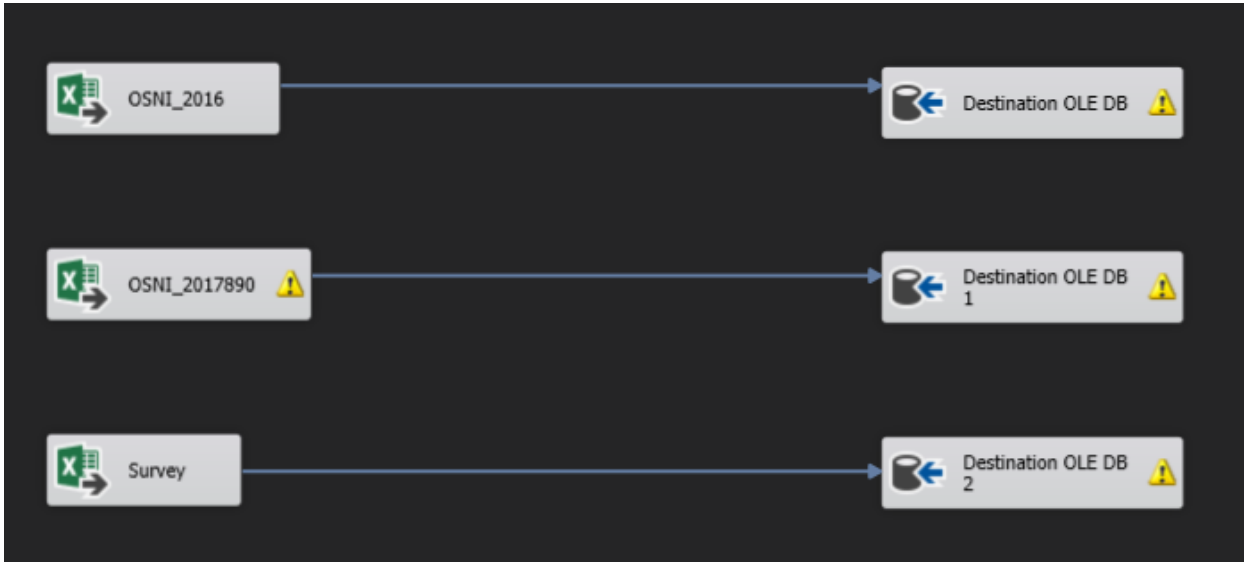
ETL Process

The process of extracting data from various sources, transforming it, and loading it onto a system that end users can access is known as "extract, transform, load." Using SQL Server Integration Services (SSIS) in Visual Studio, we will perform the ETL process for this project. We will extract the data from various flat excel and csv sources and load it into the Project Datawarehouse that was previously created on SQL Server. We must fill out all of the dimensions before moving on to fill out the fact table.

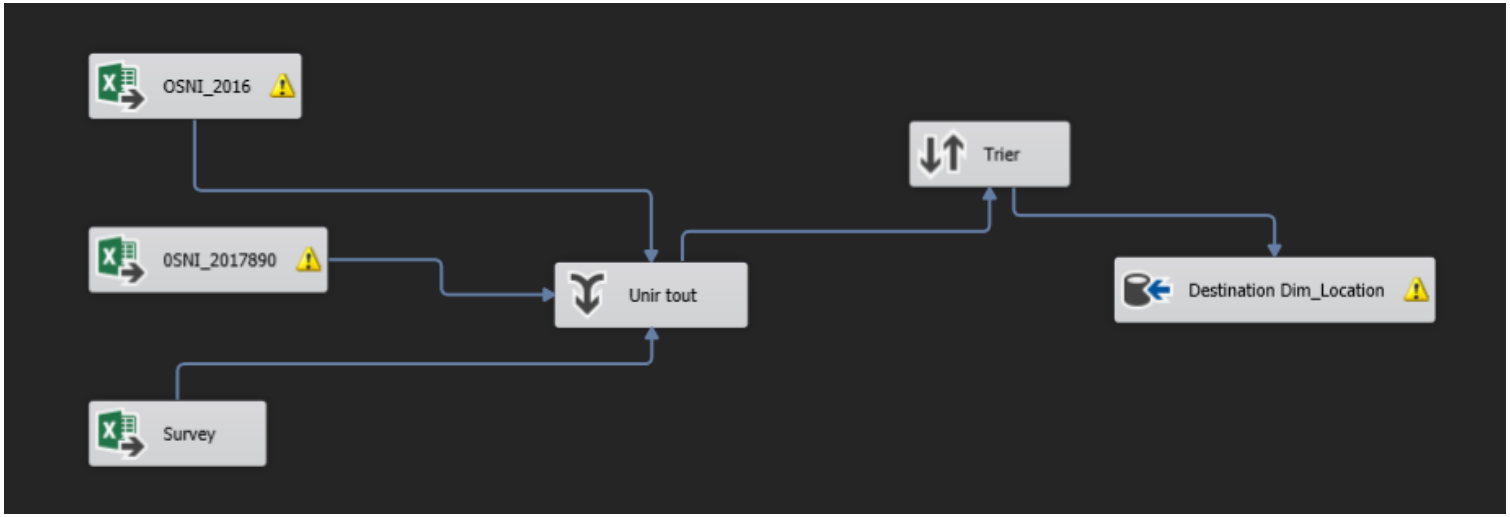
Dim_Company



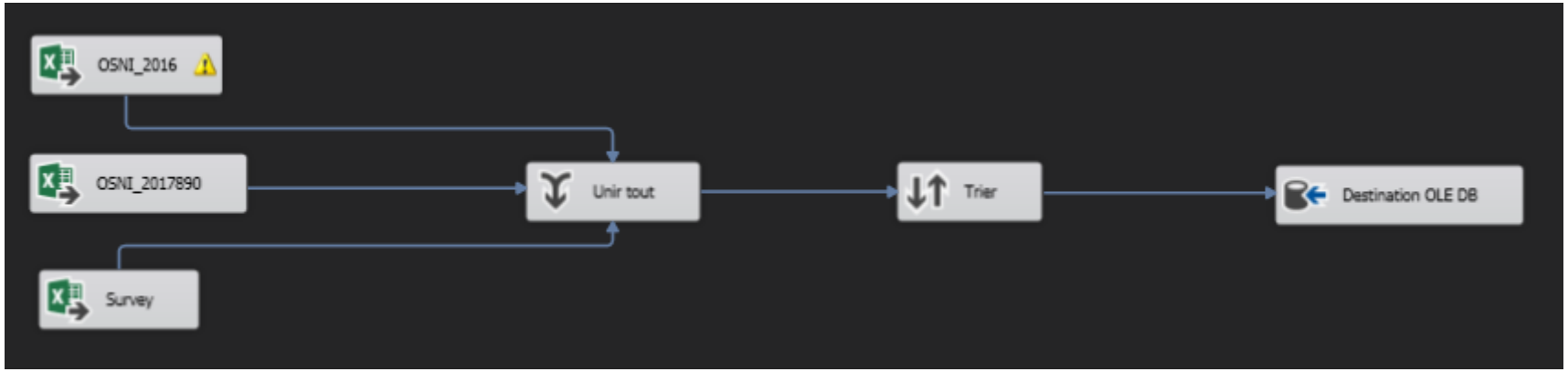
Dim_Employee



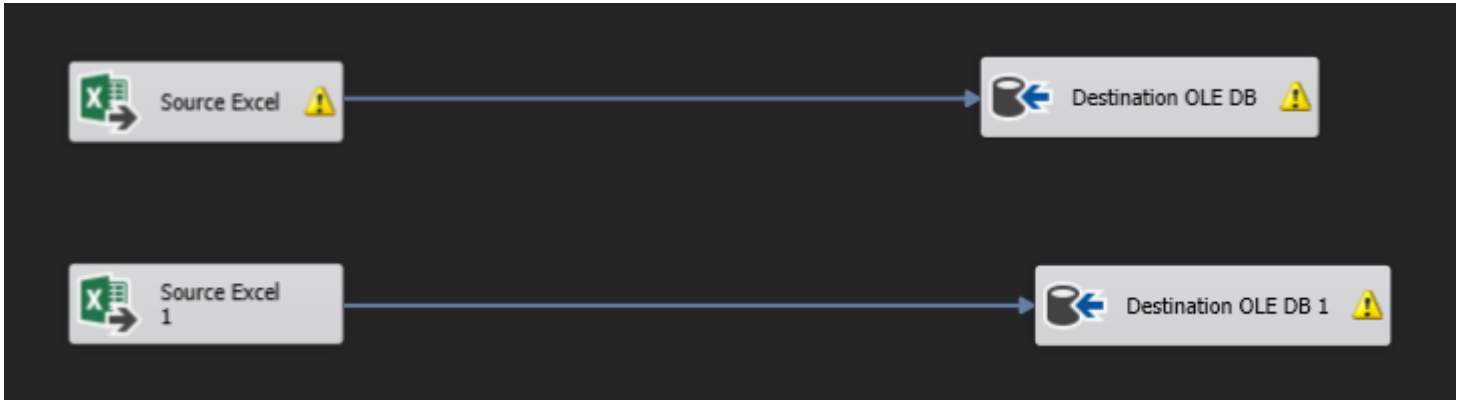
Dim_Location



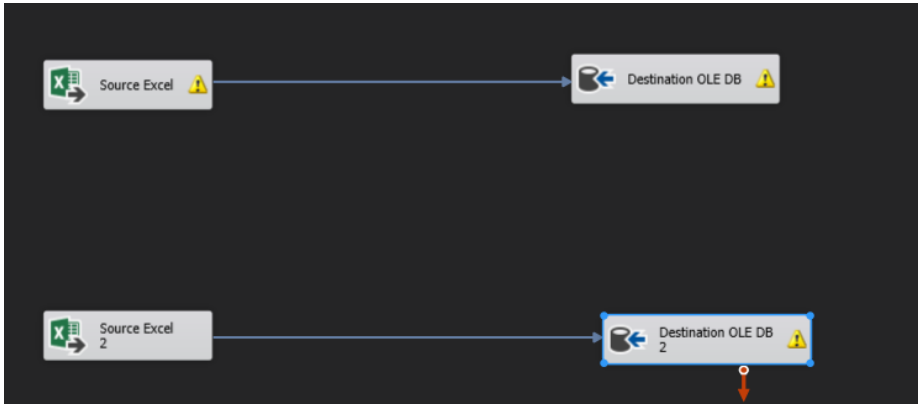
Dim_Time



Dim_Job

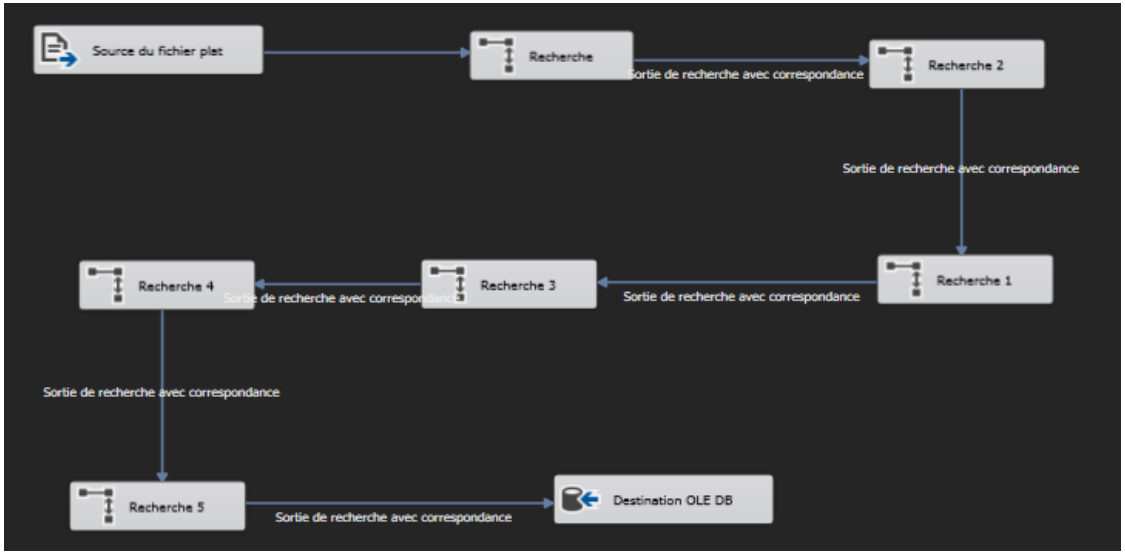


Dim_Impact



Dim_Diagnosis

Mental Health Fact Table

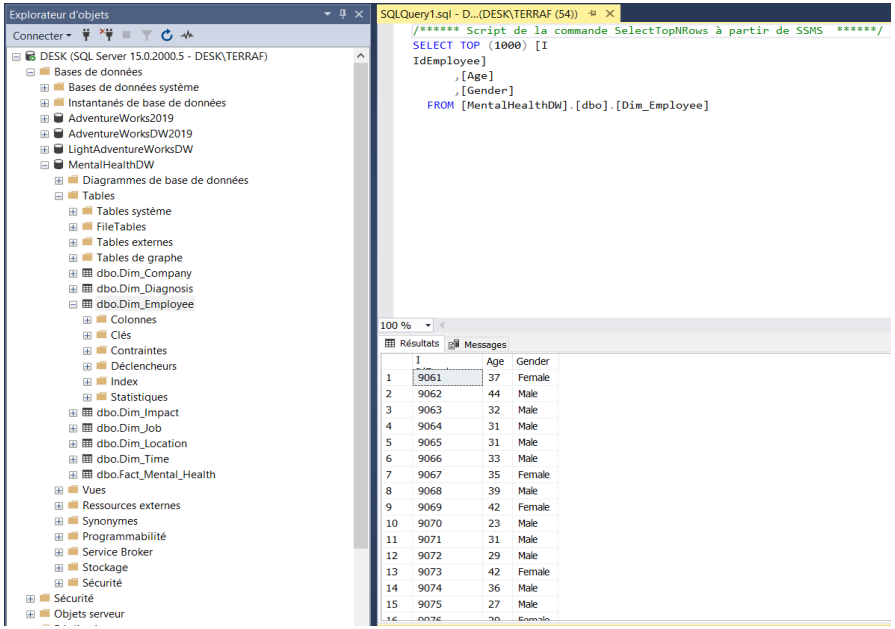


For the fact table, the lookup was used as many times as the number of foreign keys (ID) that it has.

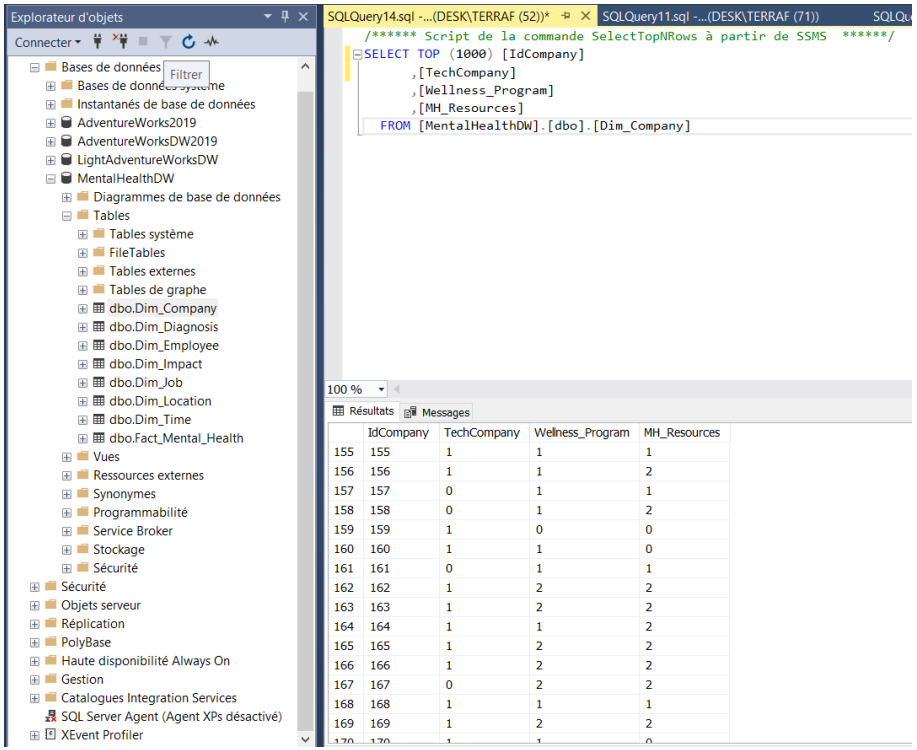
The whole ETL Process

SQL Server Tables

Dim_Employee

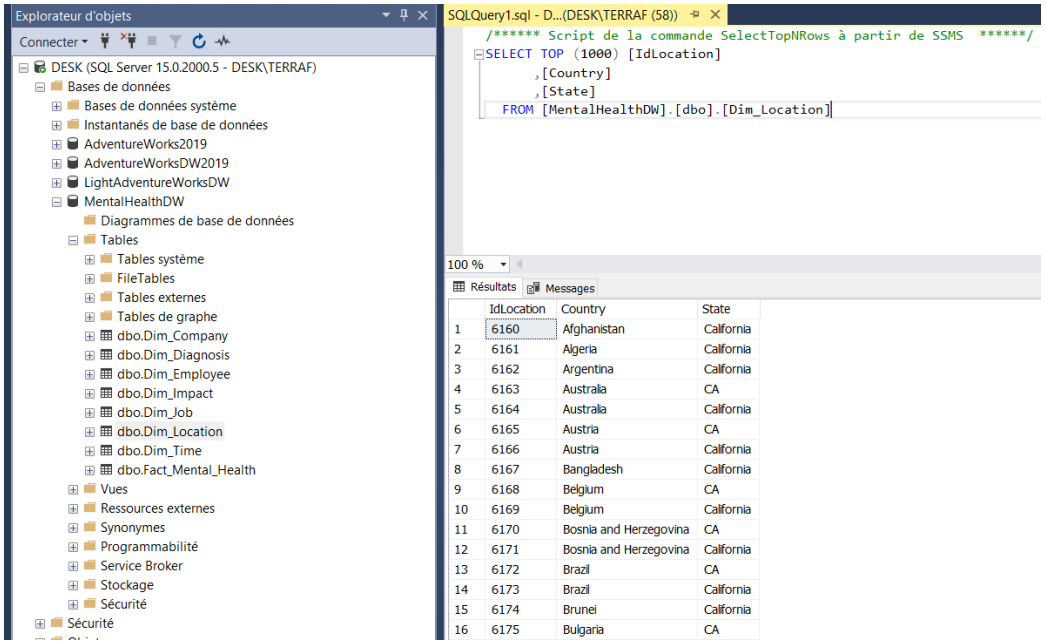


Dim_Compacy



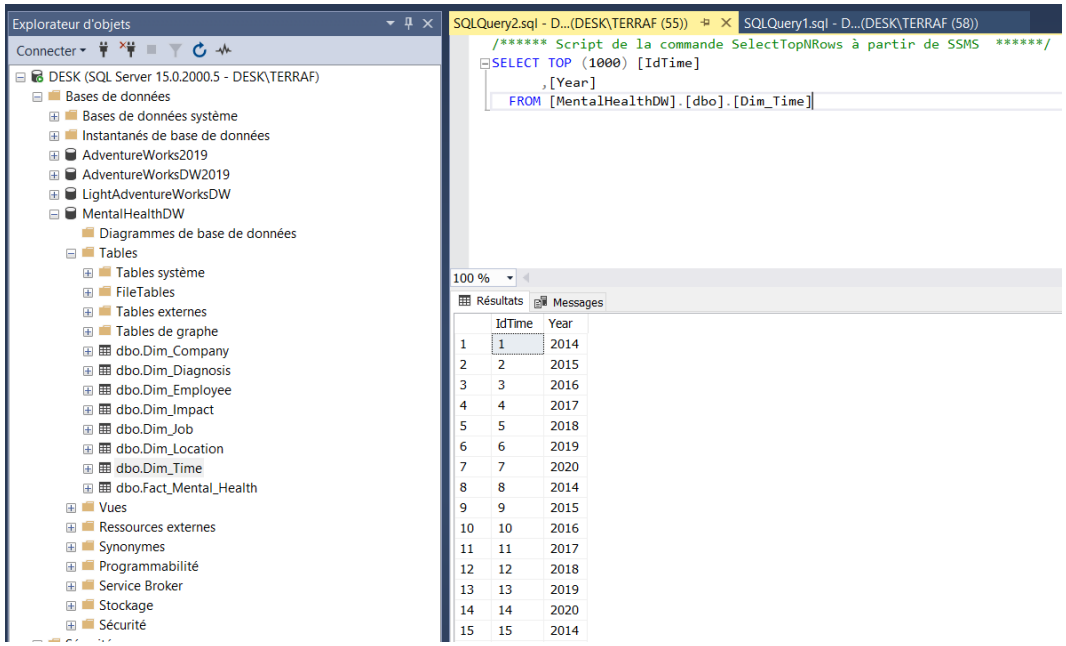
	IdCompany	TechCompany	Wellness_Program	MH_Resources
155	155	1	1	1
156	156	1	1	2
157	157	0	1	1
158	158	0	1	2
159	159	1	0	0
160	160	1	1	0
161	161	0	1	1
162	162	1	2	2
163	163	1	2	2
164	164	1	1	2
165	165	1	2	2
166	166	1	2	2
167	167	0	2	2
168	168	1	1	1
169	169	1	2	2

Dim_Location

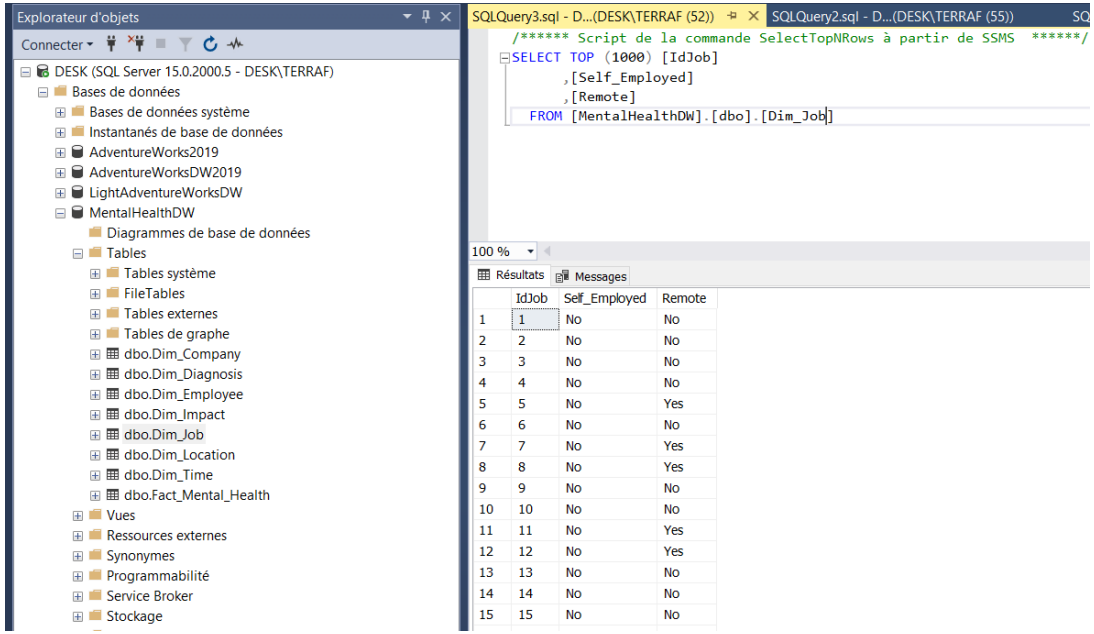


	IdLocation	Country	State
1	6160	Afghanistan	California
2	6161	Algeria	California
3	6162	Argentina	California
4	6163	Australa	CA
5	6164	Australia	California
6	6165	Austria	CA
7	6166	Austria	California
8	6167	Bangladesh	California
9	6168	Belgium	CA
10	6169	Belgium	California
11	6170	Bosnia and Herzegovina	CA
12	6171	Bosnia and Herzegovina	California
13	6172	Brazil	CA
14	6173	Brazil	California
15	6174	Brunei	California
16	6175	Bulgaria	CA

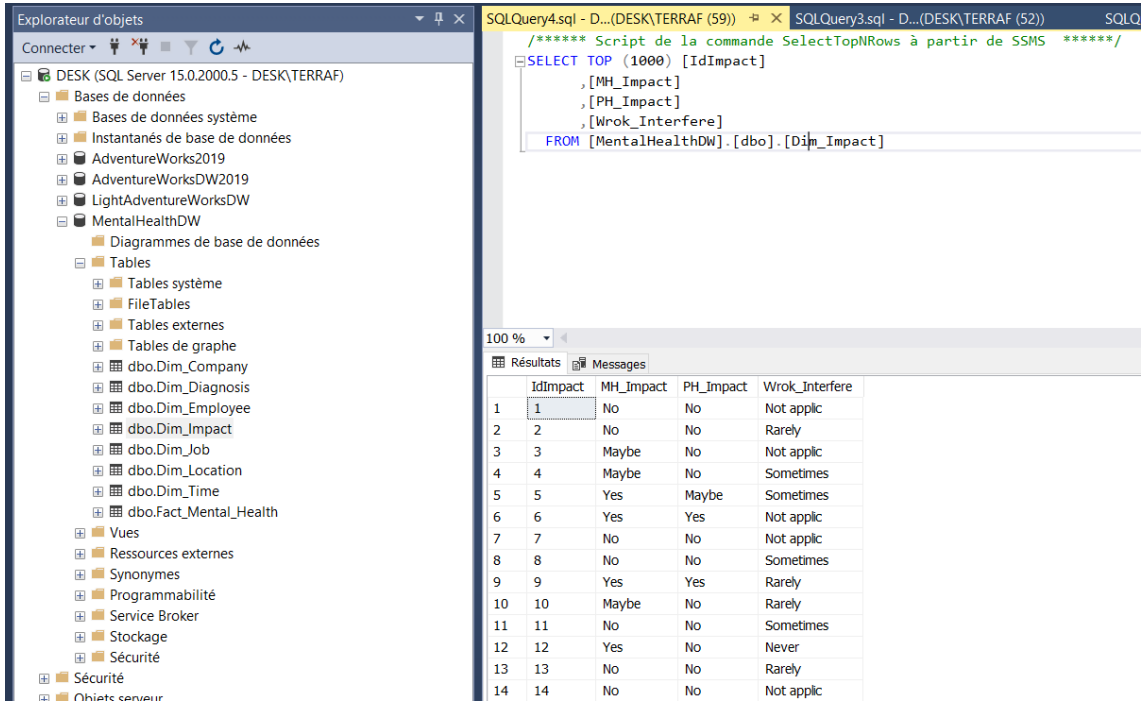
Dim_Time



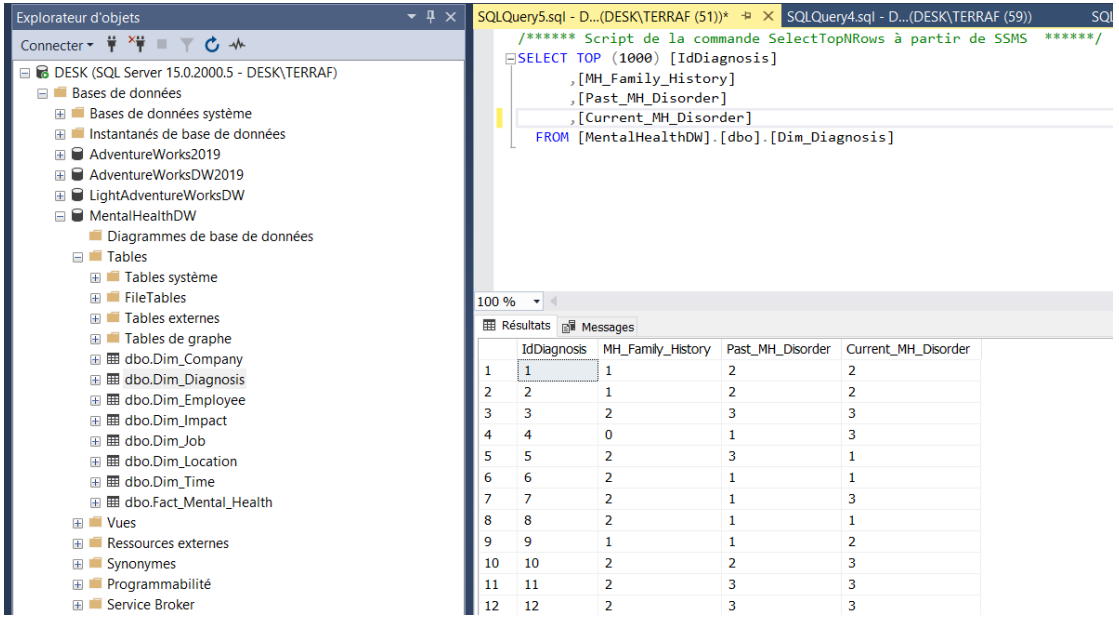
Dim_Job



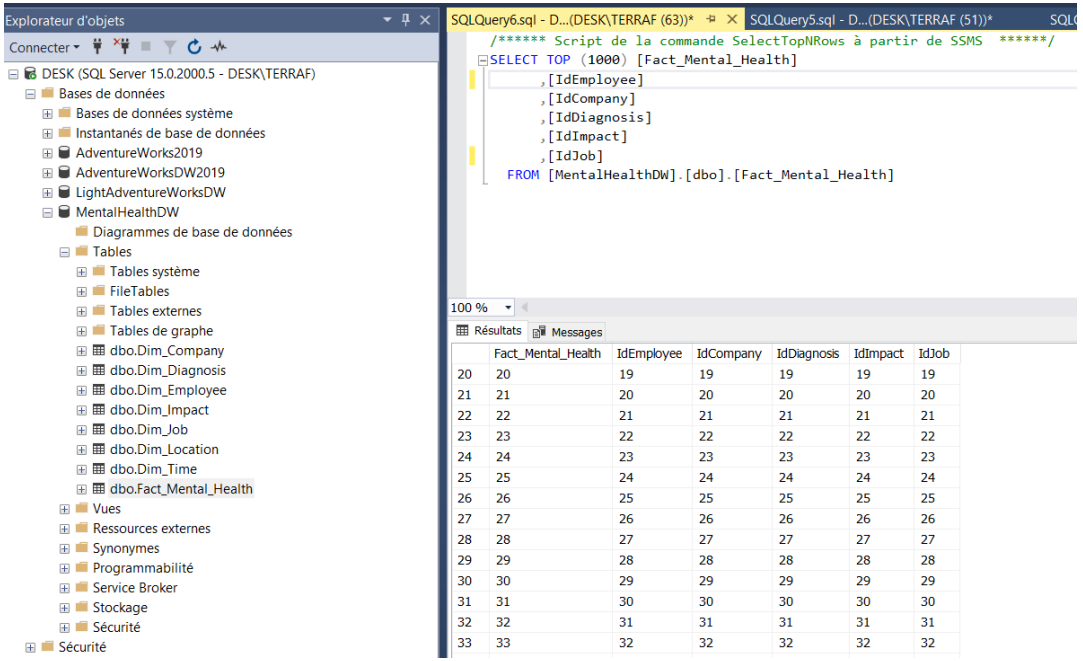
Dim_Impact



Dim_Diagnosis



Fact_Mental_Health



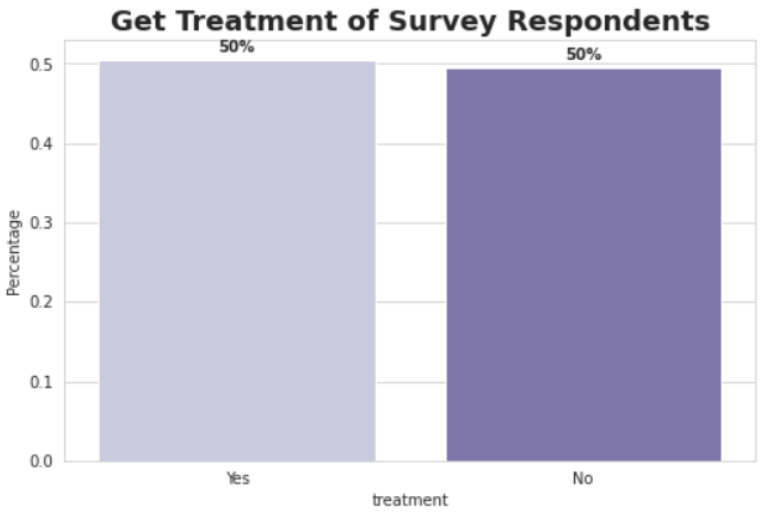
	Fact_Mental_Health	IdEmployee	IdCompany	IdDiagnosis	IdImpact	IdJob
20	20	19	19	19	19	19
21	21	20	20	20	20	20
22	22	21	21	21	21	21
23	23	22	22	22	22	22
24	24	23	23	23	23	23
25	25	24	24	24	24	24
26	26	25	25	25	25	25
27	27	26	26	26	26	26
28	28	27	27	27	27	27
29	29	28	28	28	28	28
30	30	29	29	29	29	29
31	31	30	30	30	30	30
32	32	31	31	31	31	31
33	33	32	32	32	32	32

Exploratory Data Analysis

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

This is our target variable.

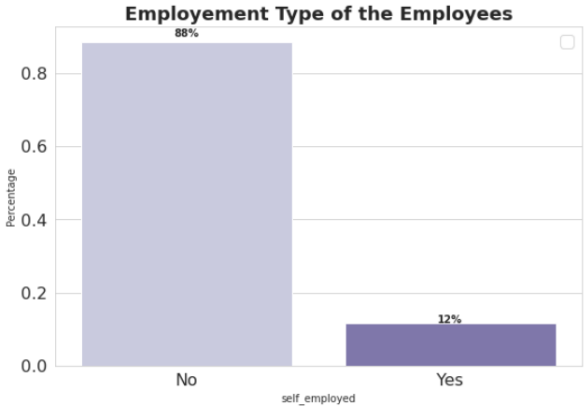
Looking at the first graph, we see that the percentage of respondents who want to get treatment is exactly 50%. Workplaces that promote mental health and support people with mental disorders are more likely to have increased productivity, reduce absenteeism, and benefit from associated economic gains.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

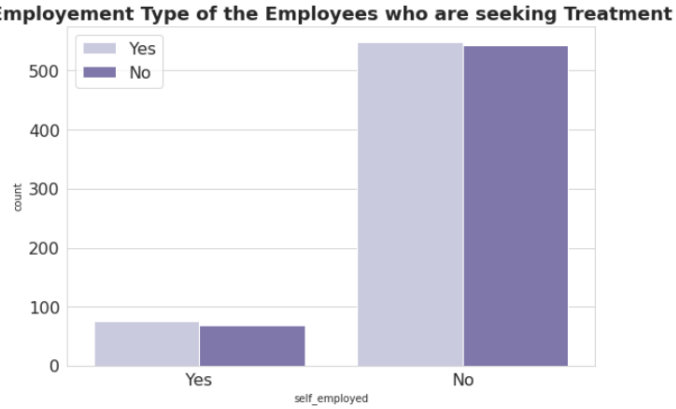


- Be more productive
- Take active participation in employee engagement activities and make better relations; both at workplace and personal life.
- Be more joyous and make people around them happy.

This is 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.

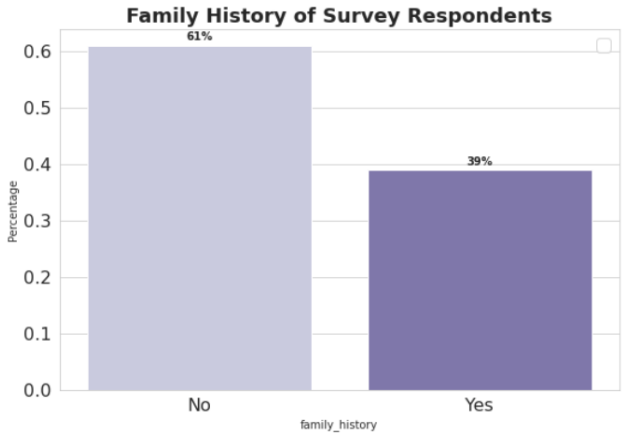


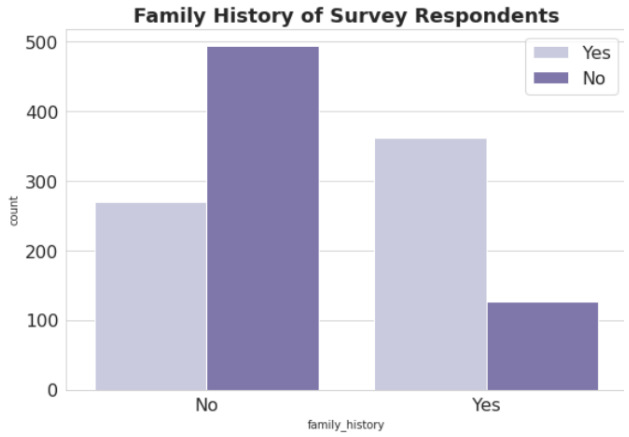


Thus, we may conclude that whether a person is self-employed or not, does not largely affect whether he may be seeking mental treatment or not.

This is the respondents answer to the question, 'Do you have a family history of mental illness?'.

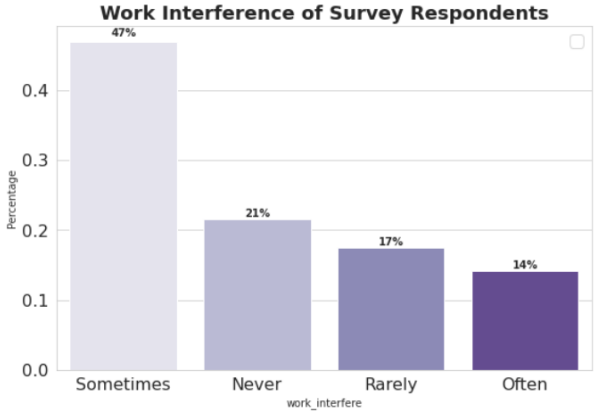
From close to 40% of the respondents who say that they have a family history of mental illness, the plot shows that they significantly want to get treatment rather than without a family history. This is acceptable, remember the fact that people with a family history pay more attention to mental illness.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

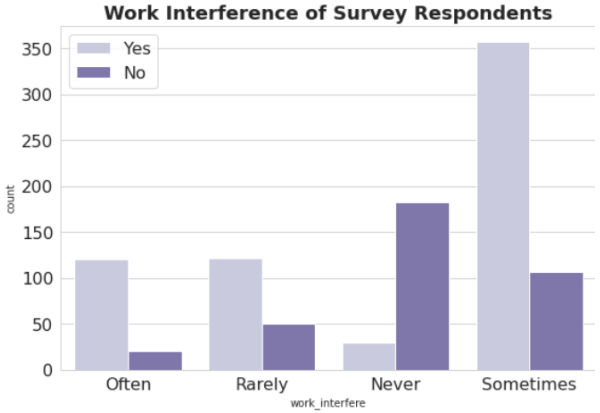




Thus, this is an important factor that has to be taken under consideration as it influences the behavior of the employees to a significant extent.

This was the respondent's answer to the question, 'If you have a mental health condition, do you feel that it interferes with your work?'.

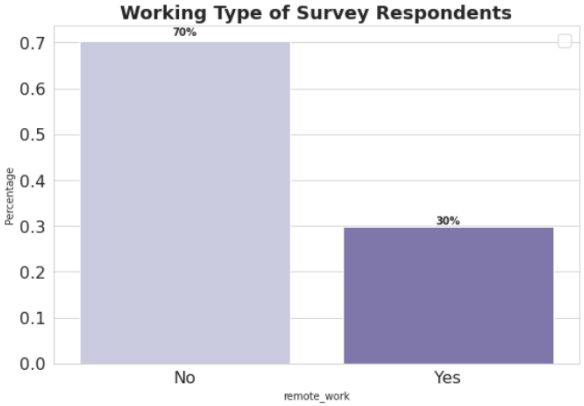


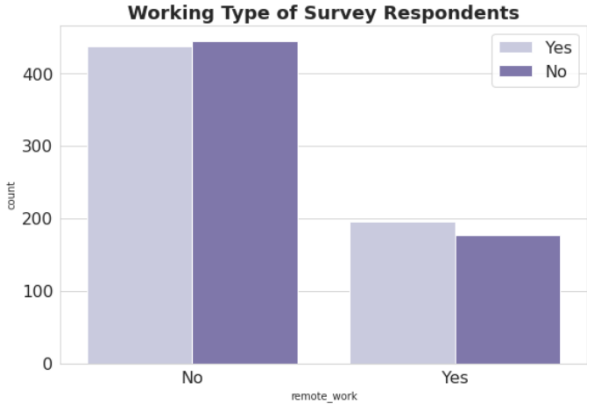


- On seeing the first graph we conclude that around 48% of people say that sometimes work interferes with their mental health. Now 'Sometimes' is a vague response to a question, and more often than not these are the people who actually face a condition but are too shy/reluctant to choose the extreme category.
- Coming to our second graph, we see that the people who chose 'Sometimes' had the highest number of people who 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 want to get treatment before it become a job stress. It can be triggered a variety of reasons like the requirements of the job do not match the capabilities, resources, or needs of the worker.

This was the respondent's answer to the question, 'Do you work remotely (outside of an office) at least 50% of the time?'.

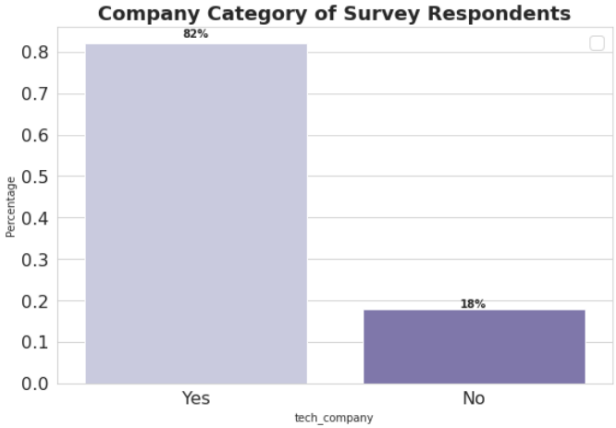
Around 70% of respondents don't work remotely, which means the biggest factor of mental health disorder came up triggered on the workplace. On the other side, it has slightly different between an employee that want to get treatment and don't want to get a treatment. The number of people who seek treatment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

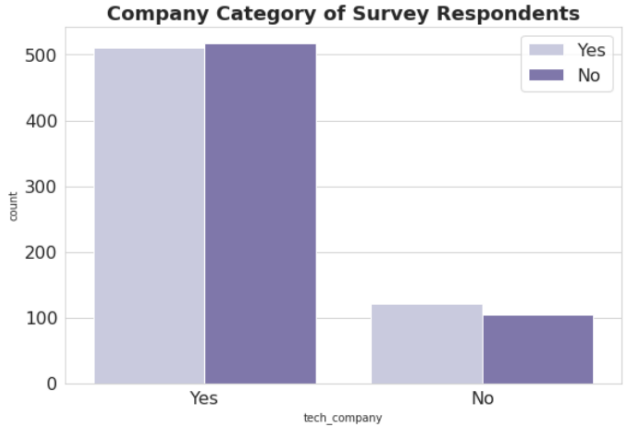




This is the respondents answer to the question, 'Is your employer primarily a tech company/organization?'.

- Although the survey was specifically designed to be conducted in the tech field, there are close to 18% of the companies belonging 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.

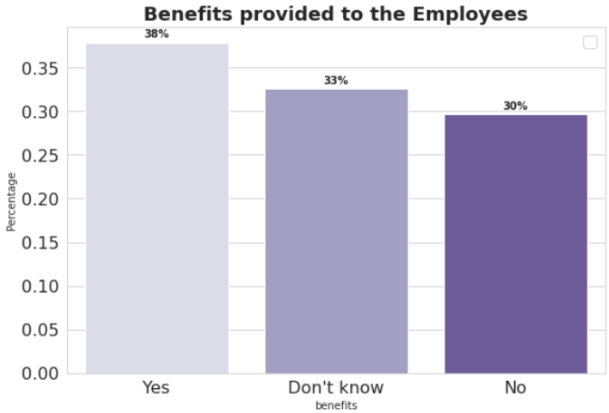


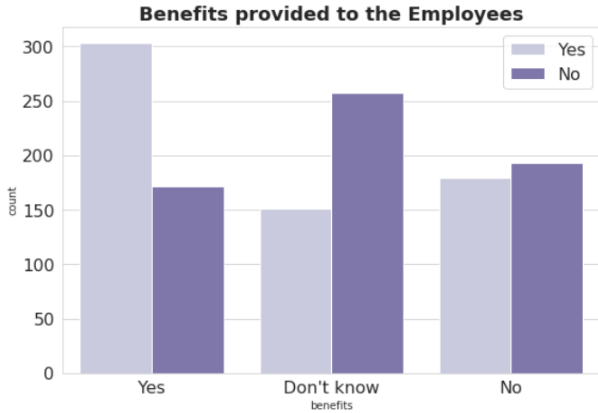


- However, on a deeper look we find that the number of employees in the tech sector who want to get treatment is slightly lower than the ones who don't. But in the non-tech field the situation gets reversed.

This was the respondent's answer to the question, 'Does your employer provide mental health benefits?'.

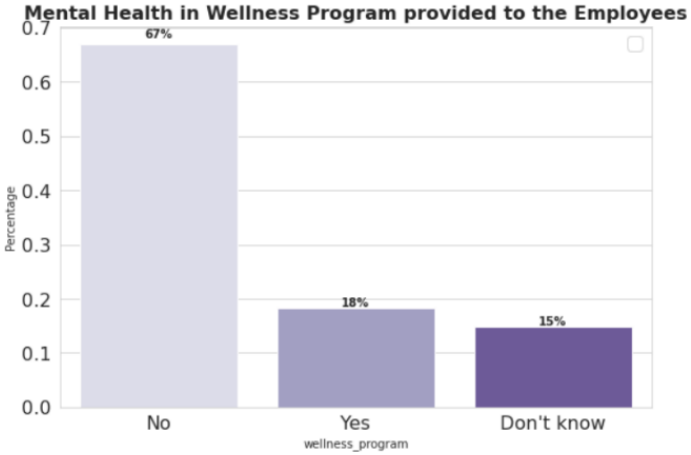
- We see that around 38% of the respondents said that their employer provided them mental health benefits, whereas a significant number (32%) of them didn't even know whether they were provided this benefit.
- Coming to the second graph, we see that for the people who YES said to mental health benefits, 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.

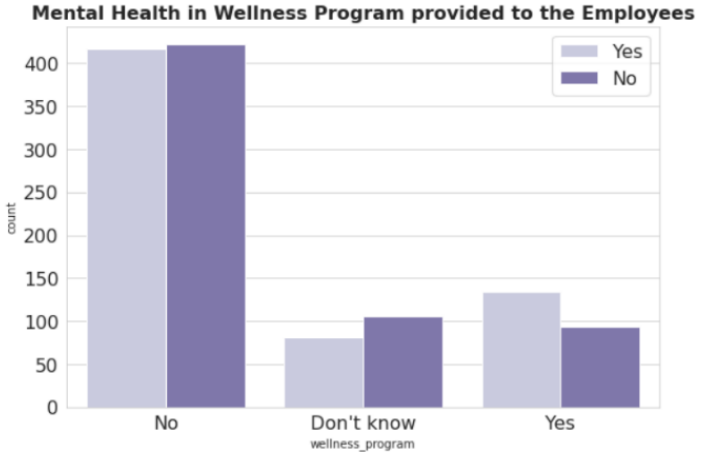




This is the respondents answer to the question, 'Has your employer ever discussed mental health as part of an employee wellness program?'.

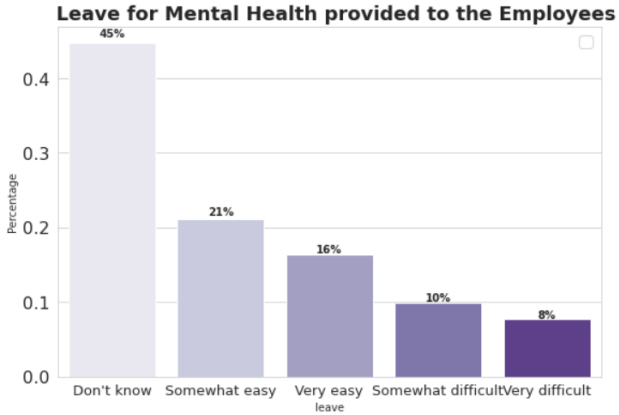
- About 19% of the respondents say YES about becoming a part of the employee wellness program and out of those 60% of employee want to get treatment.
- One shocking revelation is that more than 65% of respondents say that there aren't any wellness programs provided by their company. But close to half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

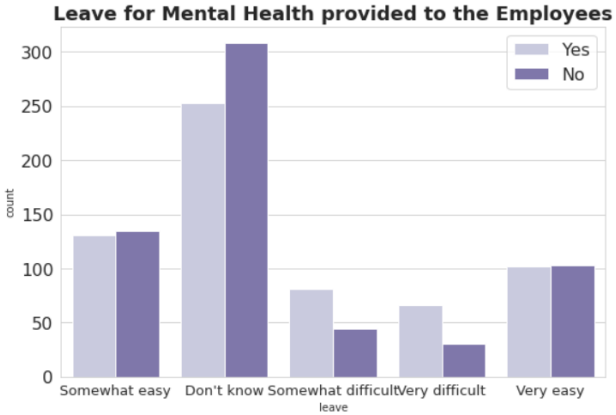




This is the respondent's answer to the question, 'How easy is it for you to take medical leave for a mental health condition?'

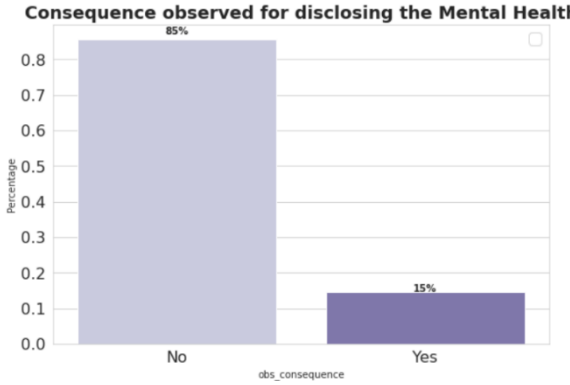
- While close to 50% of the people answered that they did not know about it, surprisingly around 45% of those people sought help for their condition.
- A small percent of people (around 8%) said that it was very difficult for them to get leave for mental health and out of those, 75% of them sought for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

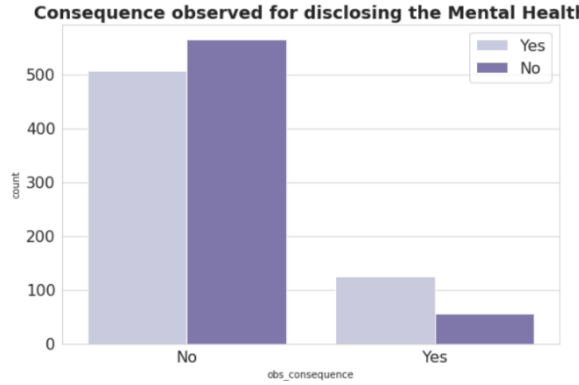




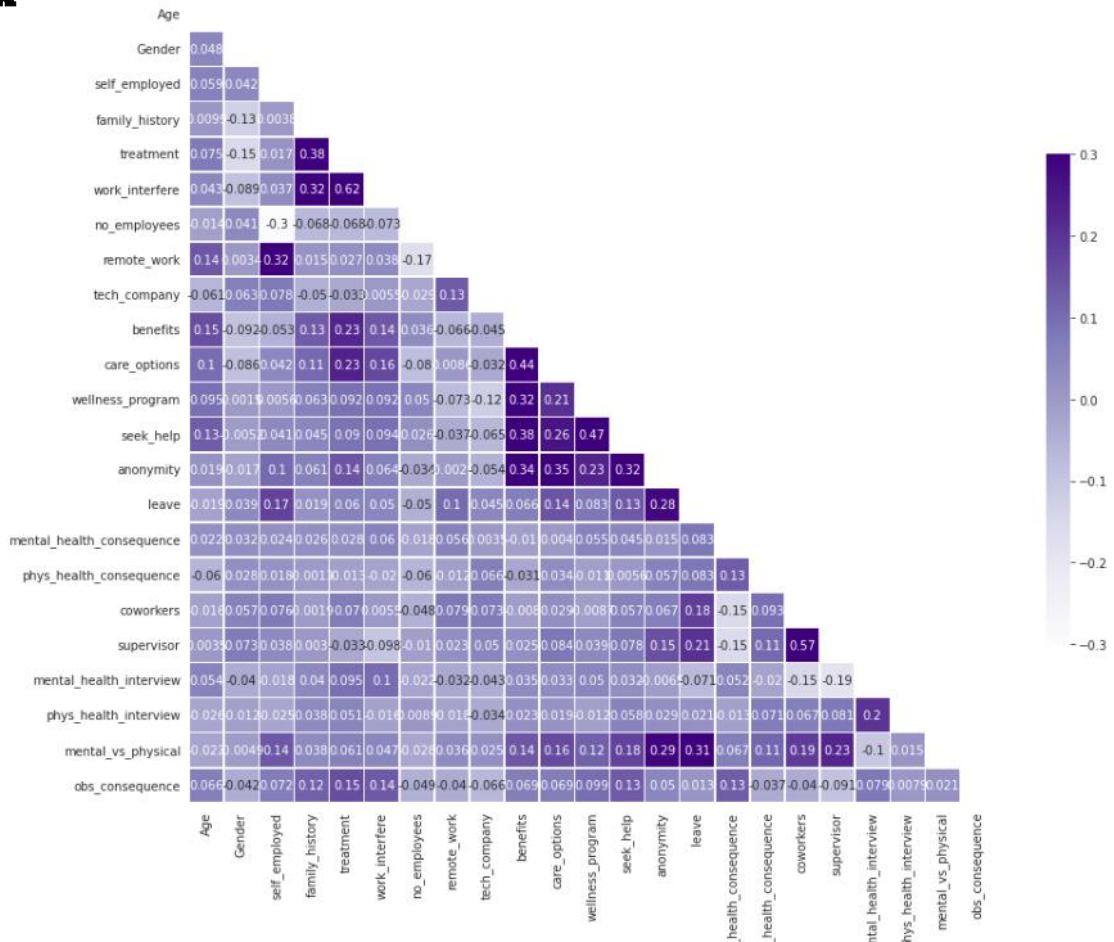
This was the respondent's answer to the question, 'Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?'

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 of employee satisfaction etc. Thus, we didn't come across any major issue regarding the employer behavior as such!



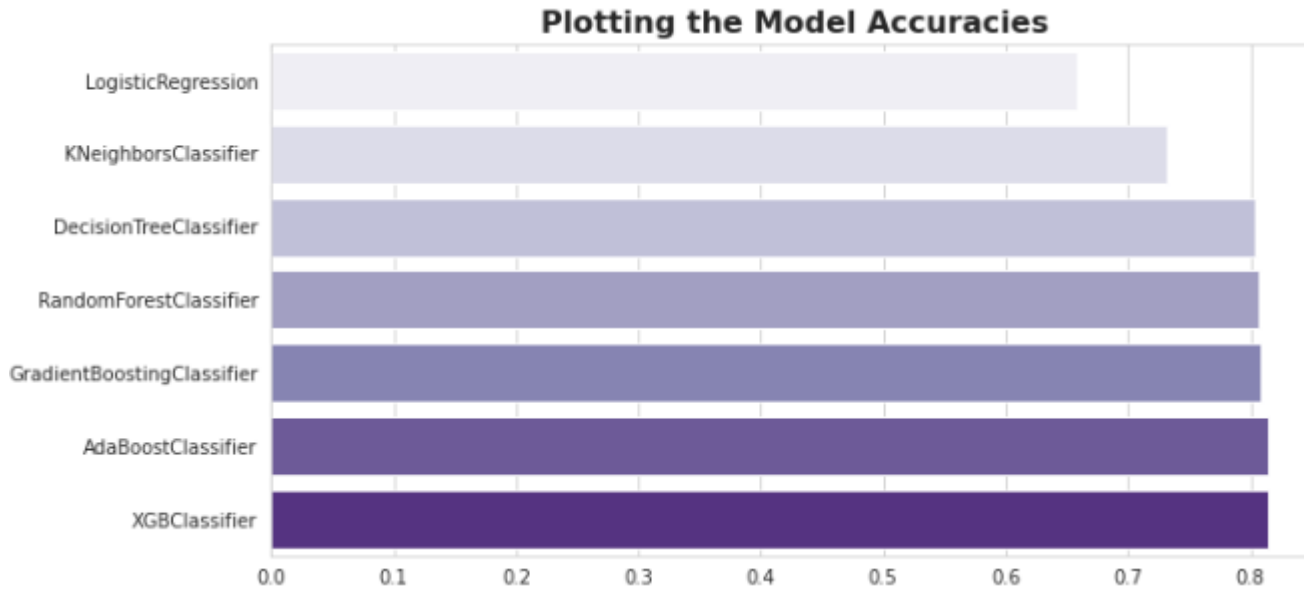


The correlation matrix



Predictive Analysis

The various machine learning techniques that were used for the data modeling are the following. The precision of each technique was different. After the preliminary training, the data models were also adjusted to improve their accuracy.



Logistic Regression

Logistic Regression is a predictive analysis method and is used to describe data and explain the relationship between dependent and independent variables. The independent variables in our model were the attributes in the survey data and the dependent variable was the person's mental state if he needed assistance or not.

K Neighbor Classifier

K neighbor classifier is used in pattern recognition and statistical estimation. It stores the labelled data and classifies the new data based on its previous data.

Decision Tree Classifier

Decision trees builds classification models in the form of a tree. The decision trees were used to find out the highest contributing factors so that due attention can be given to them.

Random Forest Classifier

Random forest is a supervised learning algorithm. It can be used for classification and regression problems. It builds multiple decision trees and merges them to get a stable and accurate prediction.

Gradient Boosting Classifier

Gradient boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects the error of its predecessor. Unlike Adaboost, the weights of the training instances are not changed, but each predictor is trained using the residual errors of its predecessor. using the residual errors of its predecessor as labels. It can be used for both regression and classification.

Adaboost Classifier

AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

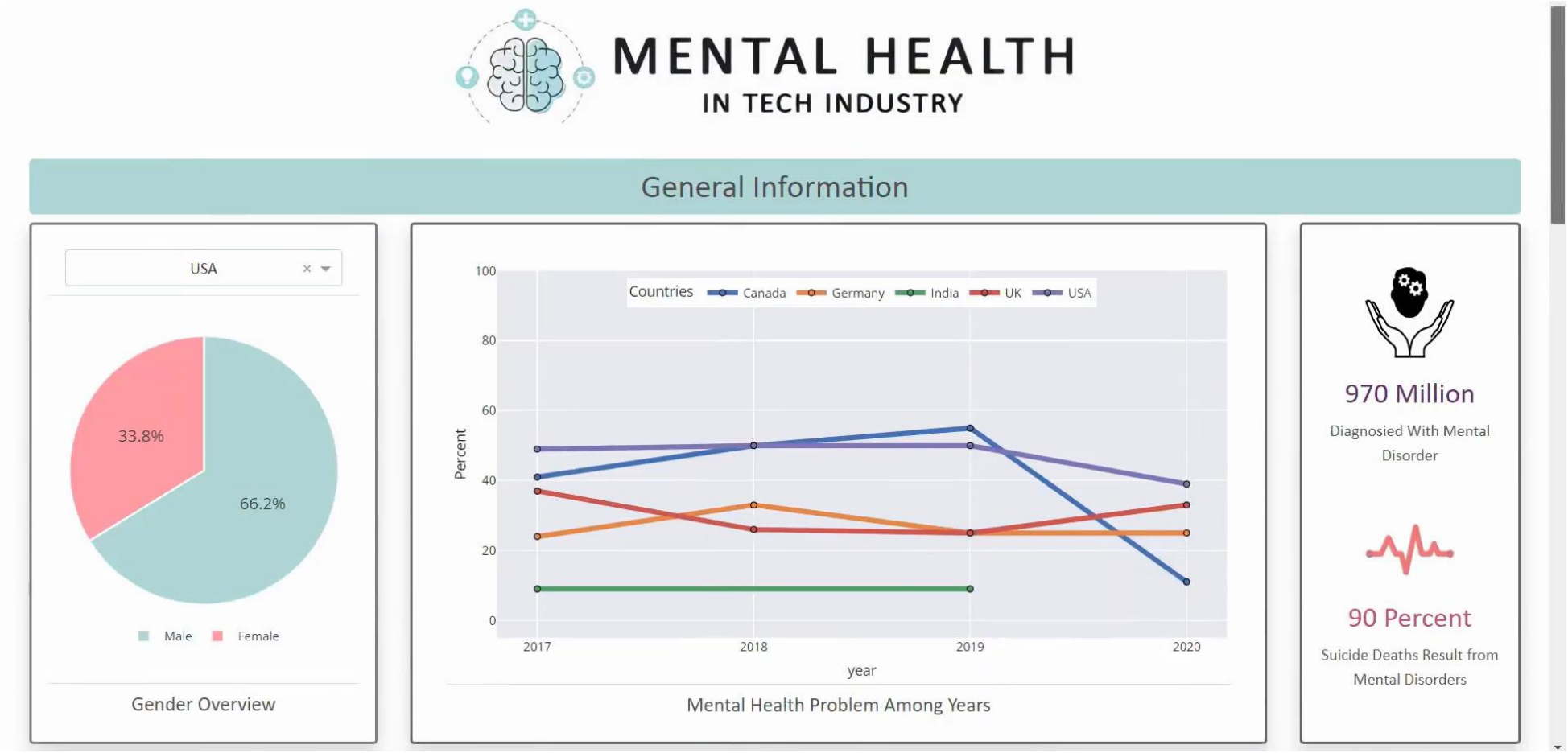
XGBoost Classifier

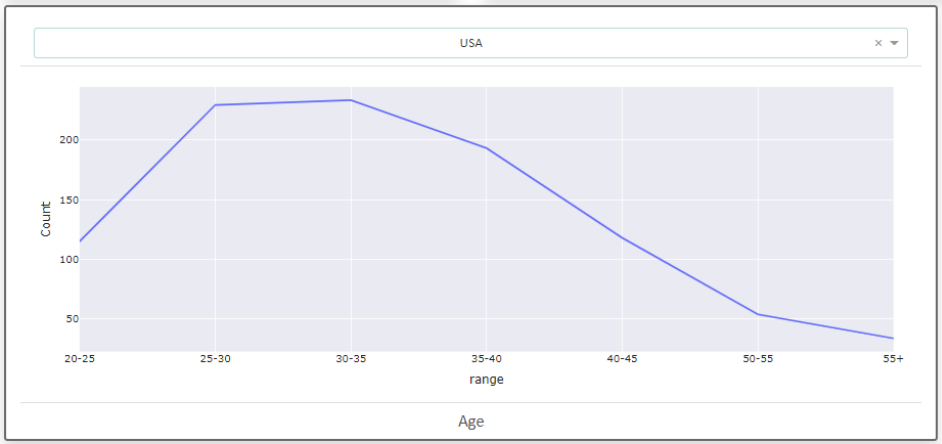
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework.

Reporting and Dashboard

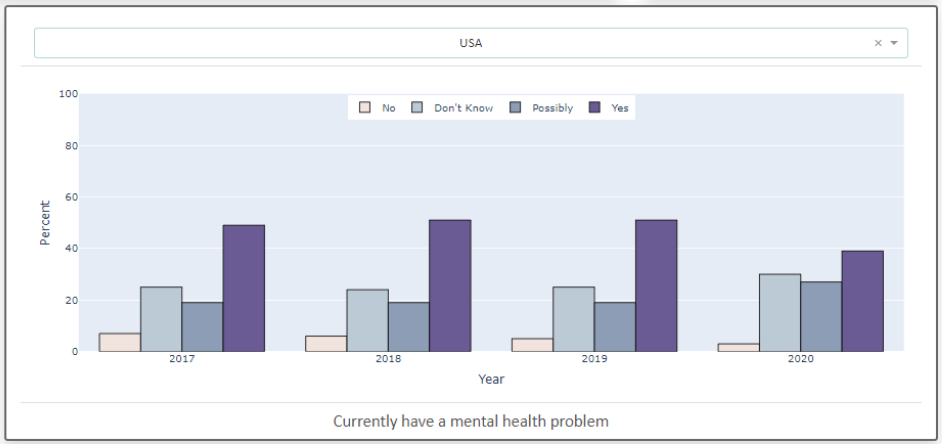
For the reporting and dashboarding, we used the Dash framework. With Dash, we created a variety of charts, graphs, and other visualizations that can be updated in real time based on user input or changes in the underlying data. This makes it easy to build dynamic dashboards that can provide a high-level overview of key metrics and trends, as well as more detailed reports on specific aspects of our data.

Additionally, Dash applications are built using Python, so we could use all the data analysis and manipulation tools available in the Python ecosystem to prepare and process our data before visualizing it. Overall, Dash is a powerful and flexible framework for building reporting and dashboard applications in Python.



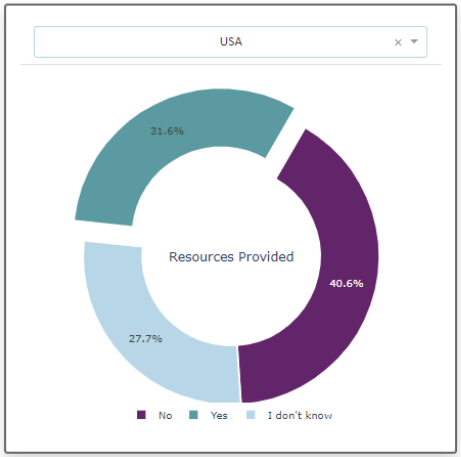


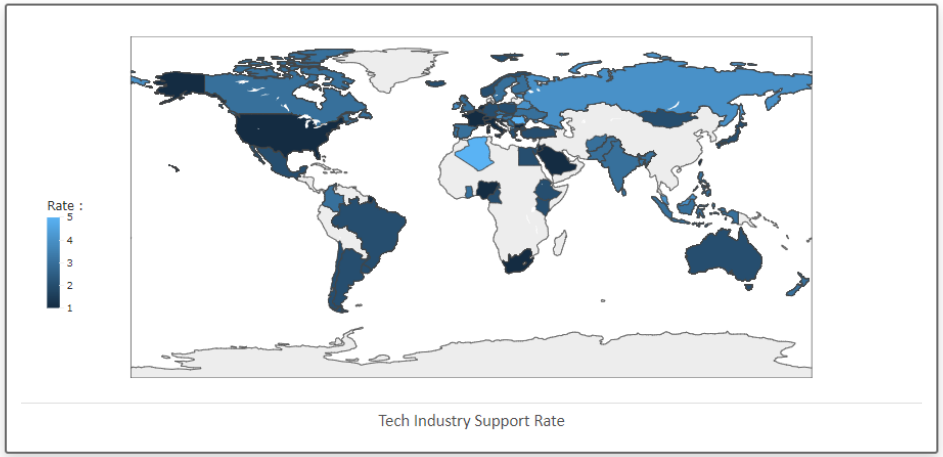
Age



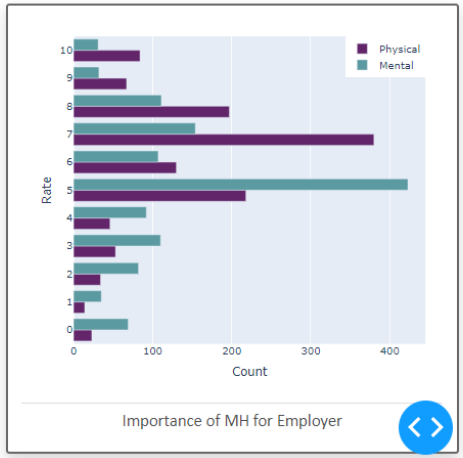
Currently have a mental health problem

Are Companies Taking Seriously Mental Health ?



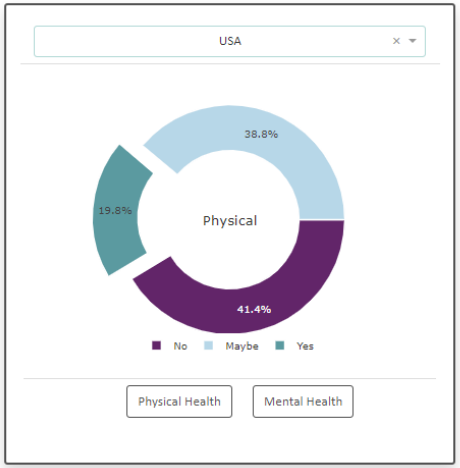


Tech Industry Support Rate

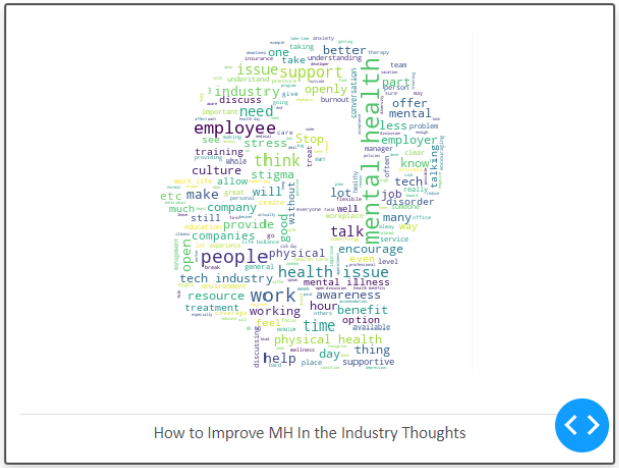


Importance of MH for Employer

Would Mental Health Problem Affect My Career ?



Admiting Physical/ Mental health issue in an interview



How to Improve MH In the Industry Thoughts