The Mental Health in Tech Survey

Rahela Jawadi

Supervisor: Dr. Ahmad Al-Janad

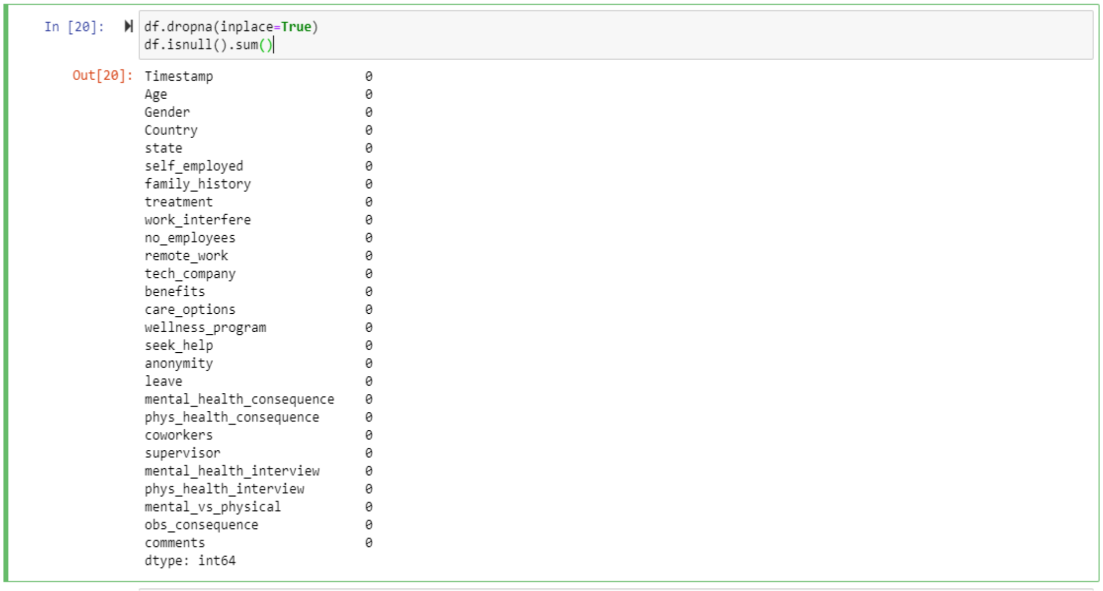
American University *of* Afghanistan

Division *of* Science Technology and Mathematics

Dec 18, 2020

1. Introduction

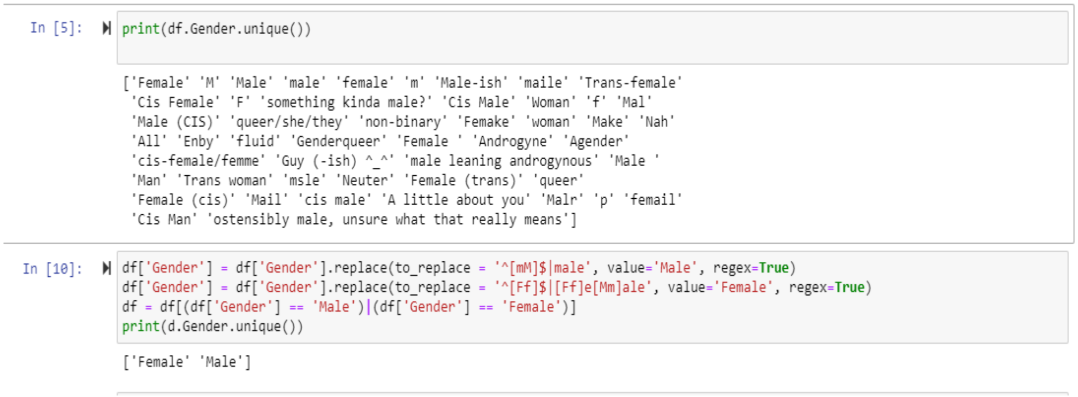
The present report analyzes and measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. The dataset that will be analyzed in this report is the Mental Health in Tech Survey dataset that was conducted in 2014, and taken from Kaggle website. Ignoring the null values, the dataset contains 1259 observations and 27 variables that among them age, gender, country, and family history are bold variables that will be used mostly for analyzing the dataset. The variables are divided into qualitative and quantitative variables. In order to find out what are the strongest predictors of mental health illness or certain attitudes towards mental health in the workplace, I have divided the variables into two different categories: univariates analysis and bivariate analysis. Before starting to analyze and visualize the dataset, is the data clearing phase and for that I found out the null values, outliers and columns that contained variations in their capitalization and the use of abbreviations and replaced or dropped those columns. After checking the null values, I have dropped the comment and state column since I don't need them in my visualization and analysis since I believe they are not a prefect indicator for representing the dataset. Also, those two columns contained more than 500 null values that replacing them with another value was not a good idea.



```
In [20]:  df.dropna(inplace=True)
          df.isnull().sum()

Out[20]:  Timestamp                    0
          Age                          0
          Gender                       0
          Country                      0
          state                        0
          self_employed                0
          family_history               0
          treatment                    0
          work_interfere               0
          no_employees                 0
          remote_work                  0
          tech_company                 0
          benefits                     0
          care_options                 0
          wellness_program             0
          seek_help                    0
          anonymity                    0
          leave                        0
          mental_health_consequence    0
          phys_health_consequence      0
          coworkers                    0
          supervisor                   0
          mental_health_interview      0
          phys_health_interview        0
          mental_vs_physical           0
          obs_consequence              0
          comments                     0
          dtype: int64
```

Also, by checking the unique values of the gender columns I came to know that the column contains a lot of variations.



```
In [5]:  print(df.Gender.unique())

         ['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']
```

```
In [10]: df['Gender'] = df['Gender'].replace(to_replace = '^[mM]$|male', value='Male', regex=True)
         df['Gender'] = df['Gender'].replace(to_replace = '^[Ff]$|[Ff]e[Mm]ale', value='Female', regex=True)
         df = df[(df['Gender'] == 'Male')|(df['Gender'] == 'Female')]
         print(d.Gender.unique())

         ['Female' 'Male']
```

As shown above, I have replaced all other values for instance m or f abbreviations to Male and Female that can be used easily on our visualization later on. Figure1 below, shows an overview of some of the independent variables and the dependent variable from the dataset after cleaning:

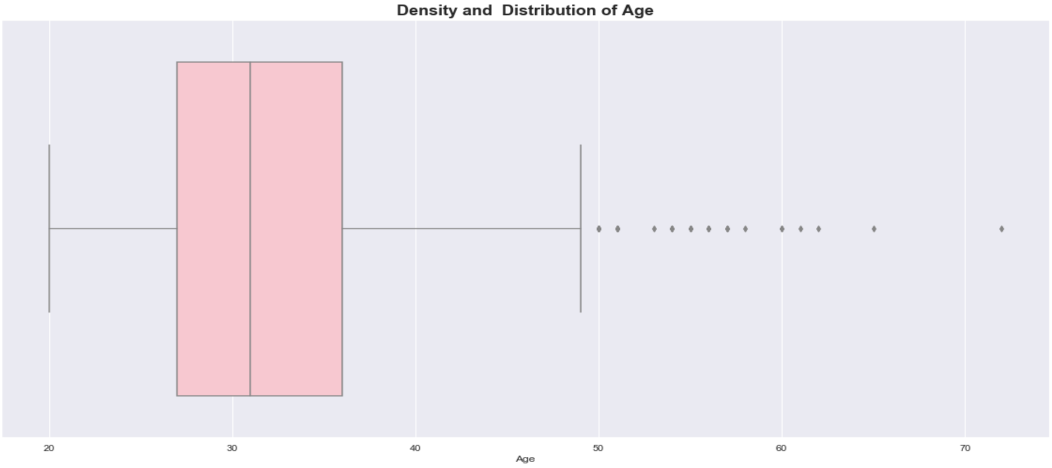
Table 1: An overview of some Independent variables and the Dependent variable

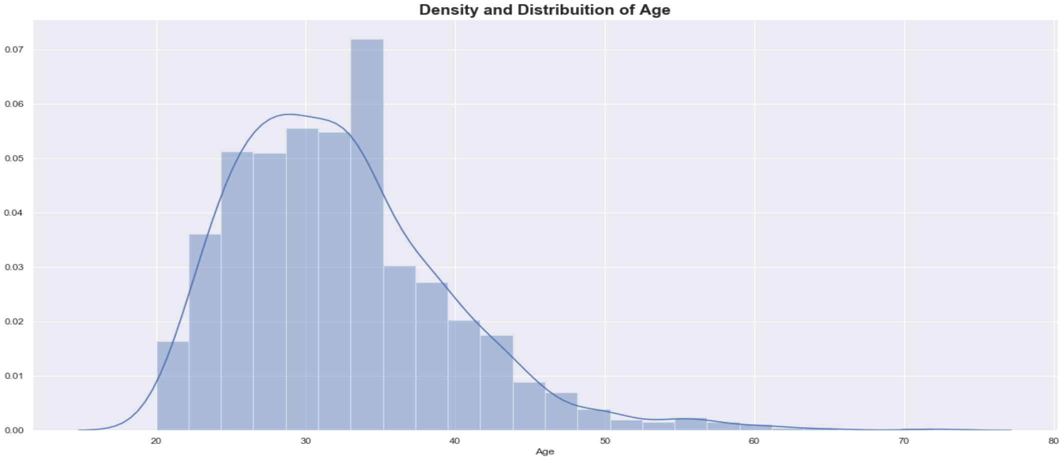
Age	Gender	Country	State	Work Interfere	Family History	Treatment
37	Female	United States	IL	Often	No	Yes
44	Male	United States	IN	Rarely	No	No
32	Male	Canada	NA	Rarely	No	No
31	Male	United Kingdom	NA	Often	Yes	Yes
31	Male	United States	TX	Never	No	No

The reason for selecting this dataset is that we are living in technology age that is controlling all aspects of our life and for sure living in this age is an indicator of more job demands in the tech marketplace. Also, another motive for selecting this dataset for me is the fact that I am a computer science student that will work in tech environment in near future. So, making sure whether this workplace is a secure and healthy workplace is of crucial importance at least for those who are living in tech environment. The analysis of the current dataset will reveal how working in tech company will affect the mental health of the employees and what other factors have a direct or indirect relationship on the employees' mental health in the workplace. In order to find out those factors that affect their mental health and working on them to remove or at least decrease those factors.

2. Discussion and Results

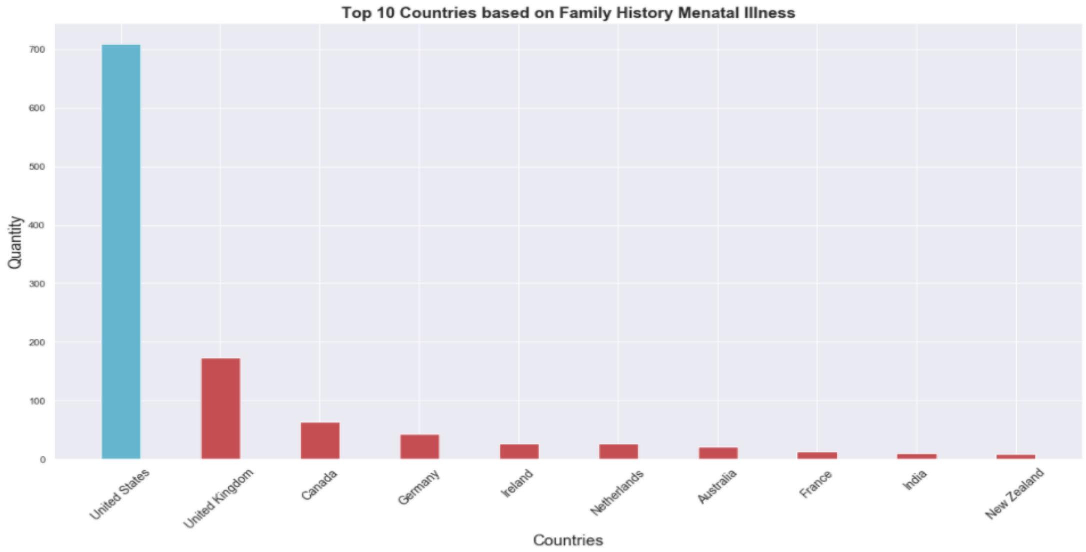
For the analysis of our univariate variables, I have used box plot and histogram to show the density of data as well as the outliers. The below box plot show that 50% of the participants age between 25 to 35 years old.



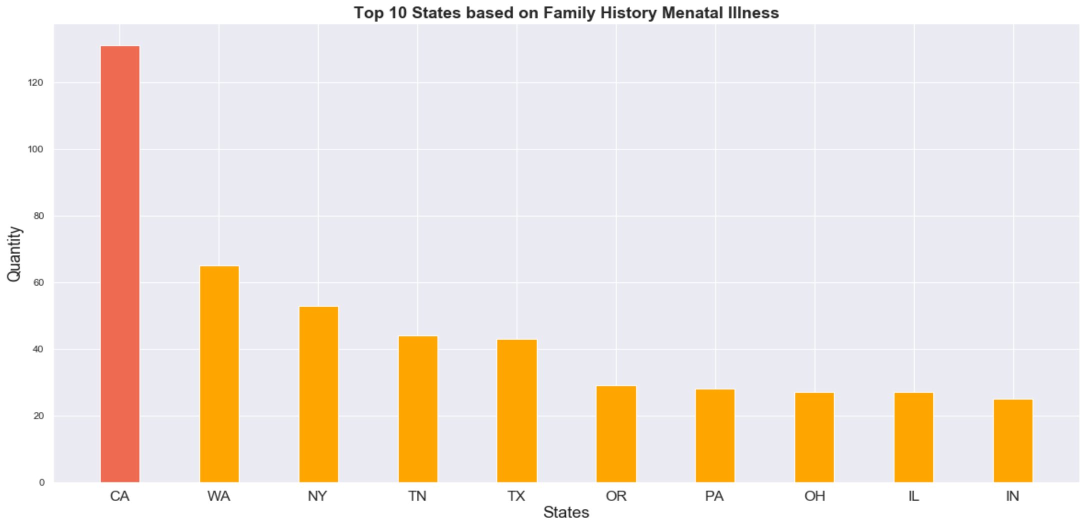


Both box plot and histogram above indicates the same thing which is more condensed data between age 25-35 and only a small range above 50 years old.

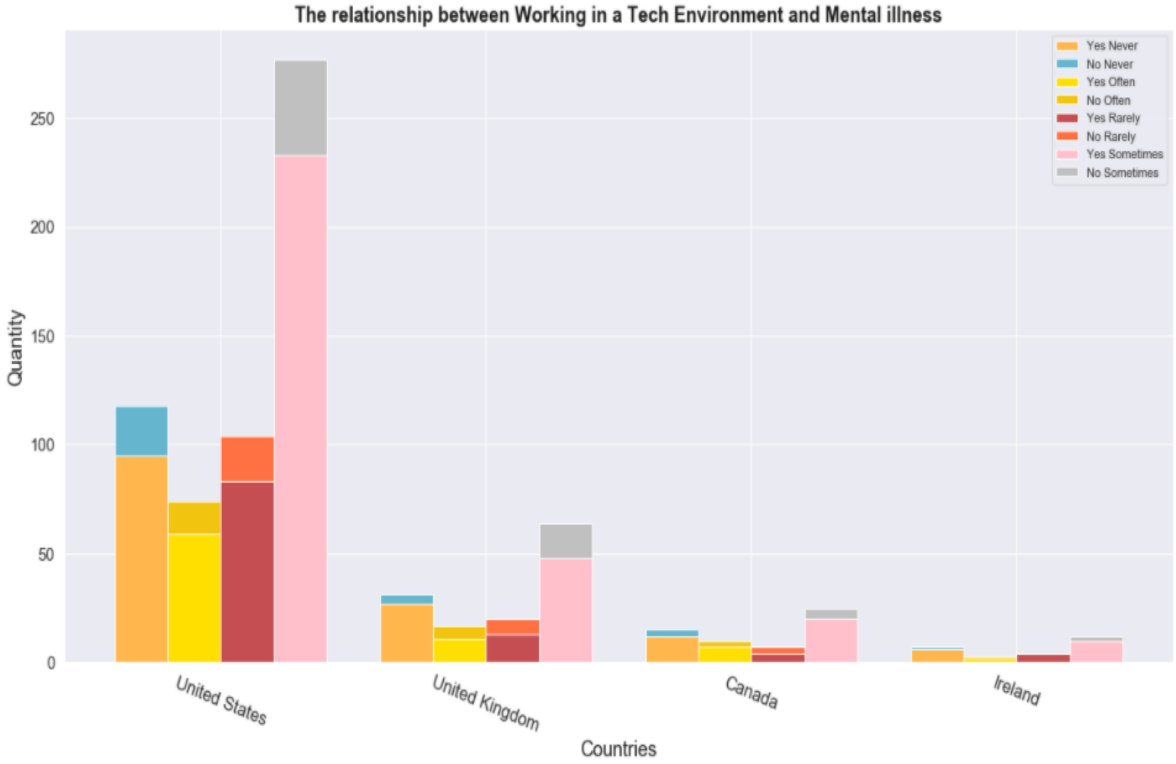
After cleaning the data and visualizing the only numerical variable which was age, now it is time to analyze and evaluate the categorical variables. The first variable among all categorical variables that I like to work on it is country that I like to find out which country among all has the highest record of family history of mental illness.



Based on the graph above, United States has the highest record of family history of mental illness followed by followed by United Kingdom and the rest. After finding out the highest country which was US, I am eager to find out the ten top states in US and list them in ascending order based on their family history of mental illness.



By observing the graph above, California is state that has the highest number of family history of mental illness followed by WA, NY and the rest. Now, it is time to answer the most interesting question which is how working in a tech environment affects the mental illness of individuals and to what extend they are related to each other. In this case I have grouped them based on country and add the no values on the same stack bar chart for better visiualization and extract the four top countries and the graph is represented below:



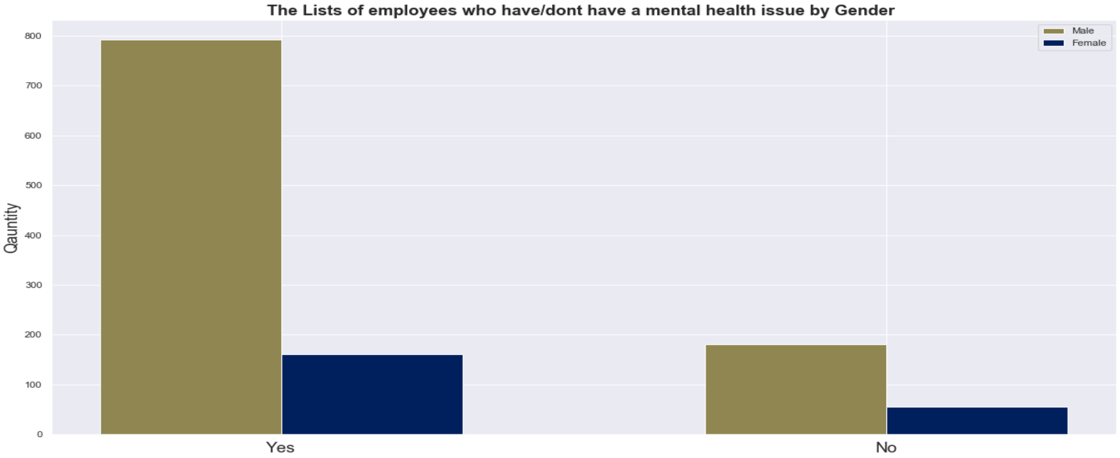
From the four categories of answers which were Sometimes, Never, Often, and Rarely it can be concluded that the number of replies for sometimes is more than other categories in all four countries. This indicates that there is a correlation between working in tech environment or industry and mental health issue, it means if they are working in that environment the chance of being affected by mental illness id more than those who are not in tech environment. Also, from the graph we can infer that among other countries in this survey US in the number one country.

After answering that question, I like to find out which gender is more affected by working in this environment. In order to answer this question, first I extracted the male and female values and stored their values in separate variables then using the count_values function I have counted the number of each value whether Yes or No in tech companies' columns. The below chart shows the calculations:

Table 2: Categories of Answers for whether Working in a Tech Environment Affect Their Mental Health

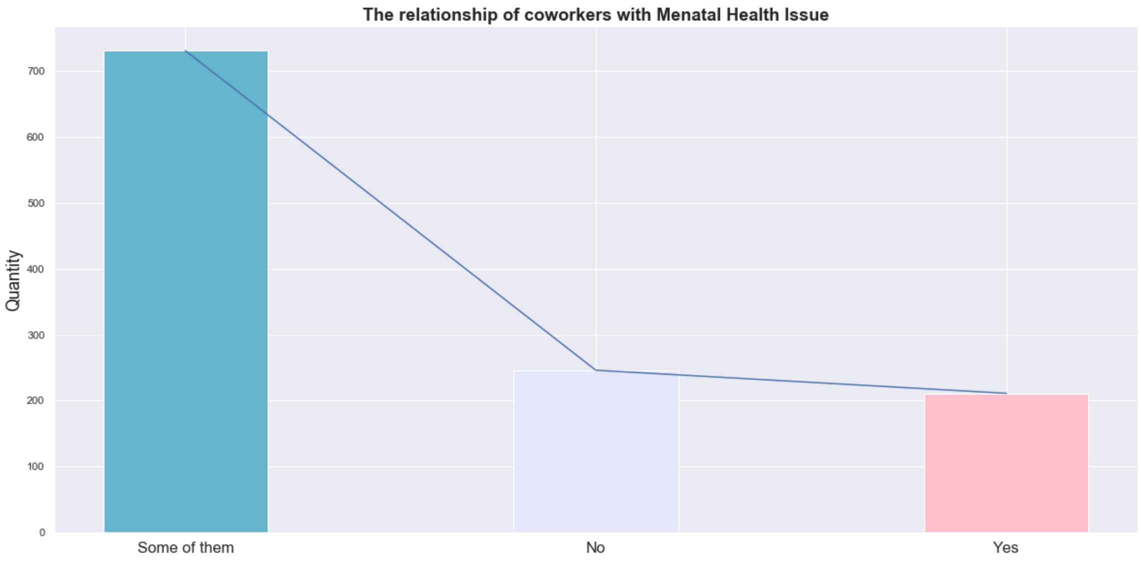
Quantity	Gender	Answers for working in Tech Environment Affecting their Mental Situation
792	Male	Yes
161	Male	No
180	Female	Yes
55	Female	No

As shown in figure 2 above, number of yes answer for both male and female category is more than no that the number of yes in total is 972 and the number no in total is 216. Also, the calculations above are an indication of the tech job environment that is dominated by men comparing to women as the number male participants in this survey working in tech industry is 4.5 times of women. However, this number is not exact 100%, still we can consider it as a sample. The below stack chart will show the calculations visually for better understanding.



Seeing the chart above we can infer that the number of yes answers are more than no in general and the male participants are more than female.

While reading the comments part, a comment that was written by a participant with index 162 pointed out, “I feel that my employer and colleagues have created my mental health issue”. Reading this comment motivates me to see and analyze what is other participants’ idea about their coworkers and can they really cause mental health issue for each other or not. So, I start by extracting the coworkers’ column, categorizing them, and finally counting the values for each category using the count values function. I was surprised that more that 731 of them which is almost 50% of them claimed that some of their coworkers cause mental health issue, and 246 of them said no while another 211 of them said yes. The bar graph below is used for better clarification.



As seen in the bar graph above, the bar with the cyan color has the highest value which belongs to some of them and followed by no and the pink one which represents the answers for yes values. I believe that the employee’s mental health was more effected by the coworkers’ factor than the previous factors such as leave, work interfere, and care option. Since more than 70% of them claimed that one factor for their mental health issue is their coworkers.

3.Conclusion

The Mental Health in Tech Survey dataset used in this report has different variables, qualitative and quantitative and the only quantitative variable in the dataset was age. The aim for this paper was to find out if in general working environment play a role on mental health and the specific goal was to find out how working in tech environment is different than other working environments. After applying the univariate and bivariate technique analysis we came to know that there is a US has the highest number of mental health illness and among other states of US, California had the highest number of mental health illness. Another finding was that there is a strong association between working environment and mental health problems, especially when it comes to working in tech companies. Among the many factors from this dataset, I have picked work interfere, care option, and coworkers as the bold variables that had a direct effect on the mental health.

The relationship of coworkers with Mental Health Issue	
Some of Them	731
No	246
Yes	211

The List of those who Believe that their Mental Health issue Interfere with their Work	
Sometimes	438
Never	203
Often	165
Rarely	132

Does their employer provides mental health care option for them	
No	479
Yes	415
Not Sure	249

Tables above are the numeric values for the graphs used in this report that clearly shows the employees view regarding different factors that play a crucial role on their mental health. The first table is the employee's mental health relationship with their coworkers in their companies that a large percentage of them pointed out that their coworkers in some cases all of them or at least some of them in some other cases play a crucial role on their mental health. The second table shows how they view their mental health to interfere with their work and still a huge number of them pointed out that their mental health often or at least sometimes interfere with their work. Finally, the third table shows the care option provided by their employers and unfortunately more than 50% of them pointed out that they are not provided with care option. In general, I would like to add the mental health illness is affected mainly by three factors of work interfere, care option, and mainly coworkers in the work environment and tech companies. The only shortcoming of the current report I believe was time limitation, since if I had more time I would analyze each and every variable on the dataset to find out which factors exactly play a crucial role on employee's mental health.