Exploratory Data Analysis on the Mental Health Survey Data Set

Introduction:

This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorder within companies. This survey was conducted world wide.

SOURCE:
Name: Mental Health in Tech Survey
Dataset: (survey.csv) [I renamed the file to mentalhealth.csv]
Owner: Stephen Myers
Shared: With Everyone
Original Dataset can be found HERE

Dataset Summary:

DATA IN COLUMNS:

➤ Timestamp: `This is the exact date and time the person took the survey`
Data: The format is dd/mm/yyyy and time. I will not need the time value so I can drop it. I will also need to pass this into a column name as it does not have a name

➤ Age: `What is your age?`
Data: There format is intergers. Need to find out age ranges.

➤ Gender: `What is your gender?`
Data: There might be inconsistencies here. Will need to format

➤ Country: `What country are your from?`
Data: All countries world wide, unique says there are 48 countries, check for duplicates

➤ state: `If you live in the United States, which state or territory do you live in?`
Data: This is dependable on if they are from USA only. If missing values in country but have values in state, then can perform imputatuion. Otherwise will not need this column

➤ self_employed: `Are you self Employed?`
Data: This data is ordinal as can only say yes or no or other.

➤ family_history: `Do you have a family history of mental illness?`
Data: This data is ordinal as can only say yes or no or other. I think I might drop this column as it holds no real relevance for the exploration I want to scope

➤ Treatment: `Have you sought treatment for a mental health condition?`
Data: This data is ordinal as can only say yes or no or other.

➤ work_interfere: `If you have a mental health condition, do you feel that it interferes with your work?`
Data: This data is ordinal as can only say yes or no or other. Will also need to check data for keywords

➤ no_employees: `How many employees does your company or organization have?`
Data: The format is inconsistant, it needs to be nominal, need to therfore clean up data here or drop column

➤ remote_work: `Do you work remotely (outside of an office) at least 50% of the time?`
Data: This data is ordinal as can only say yes or no or other.

➤ tech_company: `Is your employer primarily a tech company/organization?`
Data: This data is ordinal as can only say yes or no or other.

➤ benefits: `Does your employer provide mental health benefits?`
Data: This data is ordinal as can only say yes or no or other.

➤ care_options: `Do you know the options for mental health care your employer provides?`
Data: This data is ordinal as can only say yes or no or other. If person days that they dont know in 'benefits' column then more than likely wont have a values input in this column either, as this column id dependable on 'benefits' column. Can also help if have any missing values in 'benefit' column but have it filled out here, can perform imputation.

➤ wellness_program: `Has your employer ever discussed mental health as part of an employee wellness program?`
Data: This data is ordinal as can only say yes or no or other.

➤ seek_help: `Does your employer provide resources to learn more about mental health issues and how to seek help?`
Data: This data is ordinal as can only say yes or no or other.

➤ anonymity: `Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?`
Data: This data is ordinal as can only say yes or no or other.I dont think I will need this column.

➤ leave: `How easy is it for you to take medical leave for a mental health condition?`
Data: This data is also ordinal.

➤ mental_health_consequence: `Do you think that discussing a mental health issue with your employer would have negative consequences?`
Data: This data is ordinal as can only say yes or no or other.

➤ phys_health_consequence: `Do you think that discussing a physical health issue with your employer would have negative consequences?`
Data: This data is ordinal as can only say yes or no or other.

➤ coworkers: `Would you be willing to discuss a mental health issue with your coworkers?`
Data: This data is ordinal as can only say yes or no or other.

➤ supervisor: `Would you be willing to discuss a mental health issue with your direct supervisor(s)?`
Data: This data is ordinal as can only say yes or no or other.

➤ mental_health_interview: `Would you bring up a mental health issue with a potential employer in an interview?`
Data: This data is ordinal as can only say yes or no or other.

➤ phys_health_interview: `Would you bring up a physical health issue with a potential employer in an interview?`
Data: This data is ordinal as can only say yes or no or other.

➤ mental_vs_physical: `Do you feel that your employer takes mental health as seriously as physical health?`
Data: This data is ordinal as can only say yes or no or other. I do not think I wil need this column, as I can explore with more accurate data from my exploration based on values in columns 'mental_health_interview' and 'Physical_health_interview'

➤ obs_consequence: `Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?`
Data: This data is ordinal as can only say yes or no or other. I do not think that I will need this column as it will be irrelevant for the exploration I want to scope into.

➤ comments: `Any additional notes or comments?`
Data: The data is random and is personalised, cant get much data from this. Could form a wordcloud visualisation in order to create a visual sense of people's emotions

Data Preperation & Cleaning

Missing values and Outliers.

After creating the Data frame mental_raw_df, I printed the first few rows and columns in order to read the data and gain an overview:

	Year	Age	Gender	Country	treatment	work_interfere	remote_work	tech_company	benefits	care_options	wellness_program	seek_help	leave
0	2014	37	Female	United States	Yes	Yes	No	Yes	Yes	Not sure	No	Yes	Somewhat easy
1	2014	44	Male	United States	No	Yes	No	No	Don't know	No	Don't know	Don't know	Don't know
2	2014	32	Male	Canada	No	Yes	No	Yes	No	No	No	No	Somewhat difficult
3	2014	31	Male	United Kingdom	Yes	Yes	No	Yes	No	Yes	No	No	Somewhat difficult
4	2014	31	Male	United States	No	No	Yes	Yes	Yes	No	Don't know	Don't know	Don't know
5	2014	33	Male	United States	No	Yes	No	Yes	Yes	Not sure	No	Don't know	Don't know
6	2014	35	Female	United States	Yes	Yes	Yes	Yes	No	No	No	No	Somewhat difficult

```
Columns:['Age', 'Gender', 'Country', 'state', 'self_employed', 'family_
history', 'treatment', 'work_interfere', 'no_employees', 'remote_work',
'tech_company', 'benefits', 'care_options', 'wellness_program', 'seek_h
elp', 'anonymity', 'leave', 'mental_health_consequence', 'phys_health_c
onsequence', 'coworkers', 'supervisor', 'mental_health_interview', 'phy
s_health_interview', 'mental_vs_physical', 'obs_consequence', 'comments
']
```

I then used a function to check for all the missing values and possible outliers. What I discovered was the following:

UNIQUE VALUES:

In the columns Gender there were too many unique values (total = 49)
This was to be expected for the Age, Country, and other columns. I therefore needed to change the values for some of these as working with too many unique values would create a challenge for accurate analysis.

MISSING VALUES:

Missing data was found in columns *'state'; 'self_employed'; 'work_interfere' and 'comments*'. Now without having to compromising on dropping columns and rows for the missing data, I decided to explore other options as to not jeopardize the data.

I am going to use some logic for column *work_interfere*. If there is missing data, I can assume that the answer is no. If their mental health was interfering with their work, they would surely have said yes. Therefore, I am going to perform imputation on this column. For all the NA values I am going to fill with 'No'.

I am going to ignore missing values for 'comments', simply going to fill the missing values with an empty string. Not all people within surveys leave comments but

usually the people that do are the people who would really like to express their opinions/emotions.

Since there is no missing data in 'country' I am going to drop 'state' as I don't need it as it is dependable on Country being USA. I am also going to focus my exploration by country not by state.

IRRELEVANT COLUMNS:

Dropping columns:
'self_employed','no_employees','family_history','state','anonymity','mental_vs_physical','obs_consequence

IMPUTATION:

Performed imputation to place values inside of missing values as well as simplifying the dataset.

Work interfere:

- There are too many unique values and I want to simplify this by only have No or Yes as a value.
- Replacing values 'Often','Rarely','Sometimes' to value "Yes" (These values logically are all yes)
- Replacing values 'Never' and missing values (Nan) to value "No" (These values logically are all no)

Comments:

- Replaced the blanks with 0 values then converted it to empty strings. I wanted to keep the comments format in the string format.

Gender:

- Replaced all the values and simplified it to 3 types: Male, Female and Other

Timestamps:

- Converted the timestamps into date format and removed the time from the column.
- Then changed the data to only show the year
- Changed the column name to Year

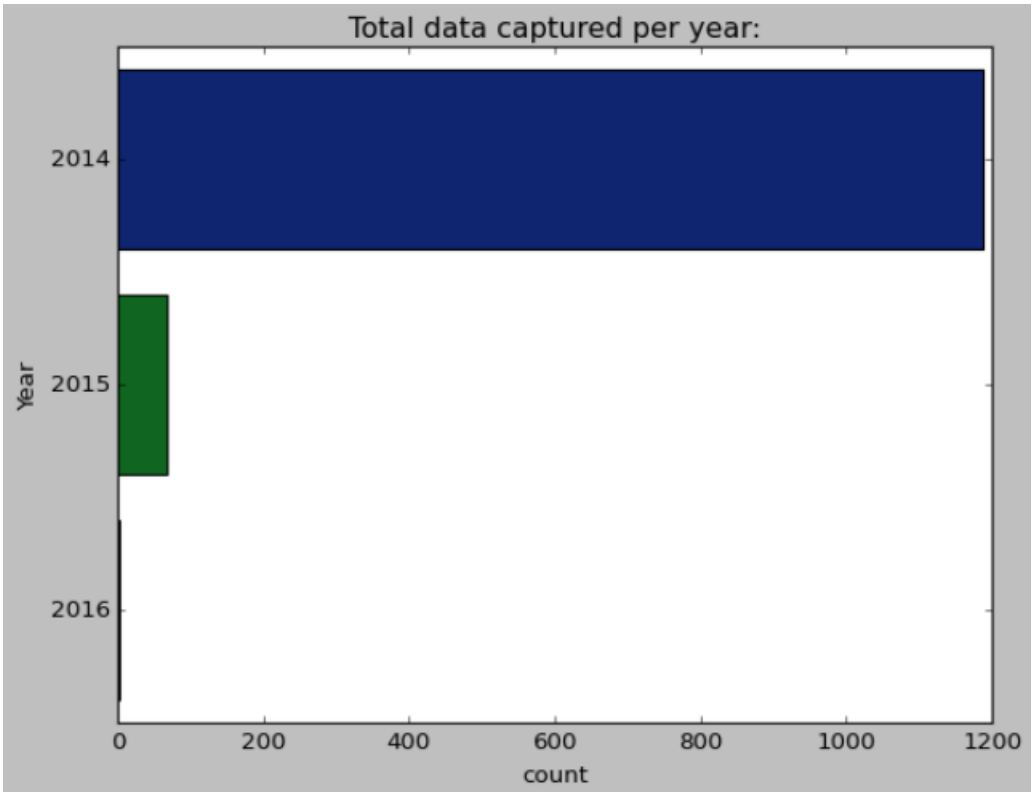
Benefits:

- Using 'loc' function to search if there is values in 'benefits' stating 'Don't Know' and if values in 'care_options' in same rows say 'Yes'. Since there is values in care_options stating Yes, and the same rows in in benefits is stating 'don't know'- I can therefore through logic decide to change the values in 'benefits' since this column is dependable to 'care-options.

Age:

- Converted the inconsistent inputs (*eg—age = 1726*) to the average age group as a replacement.

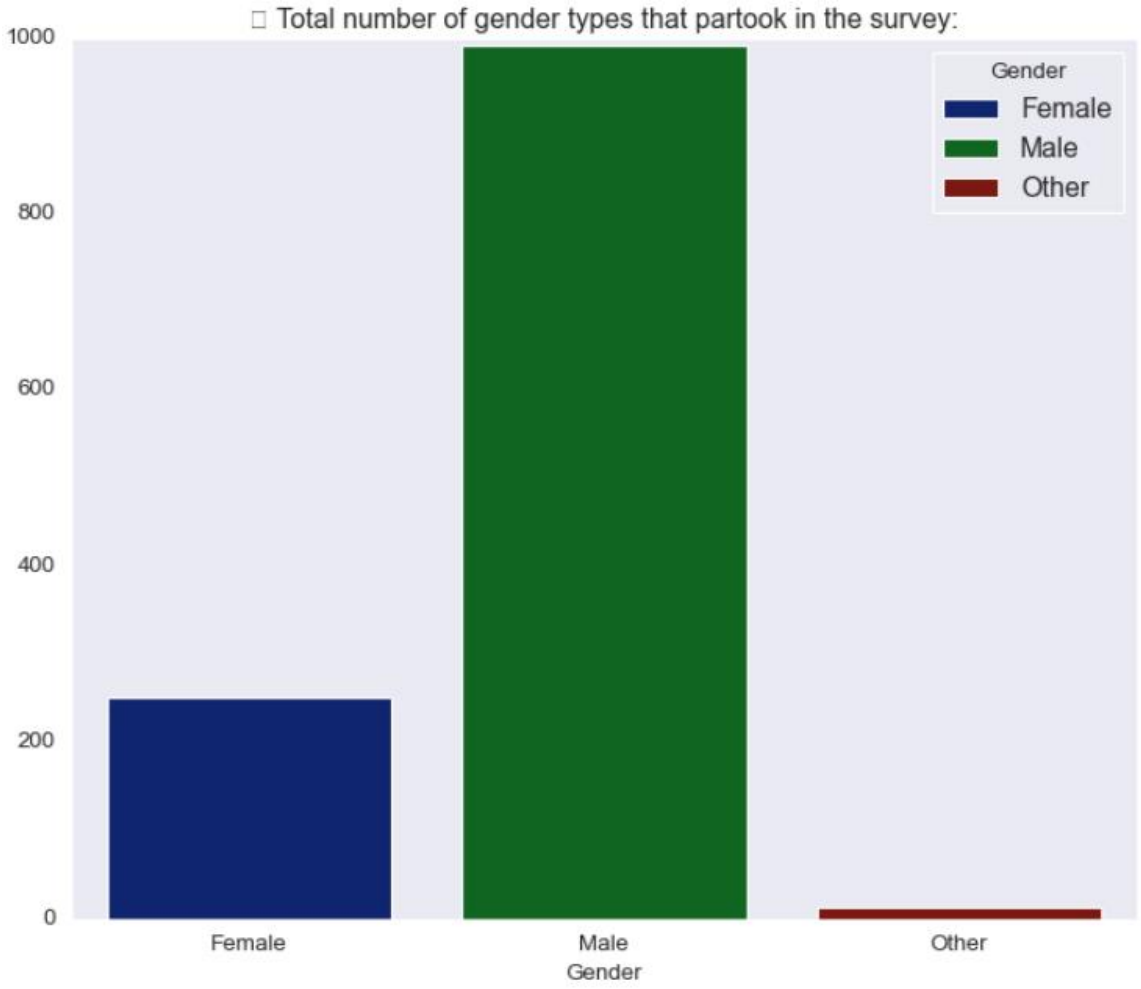
Visualisations:



Data Observations:

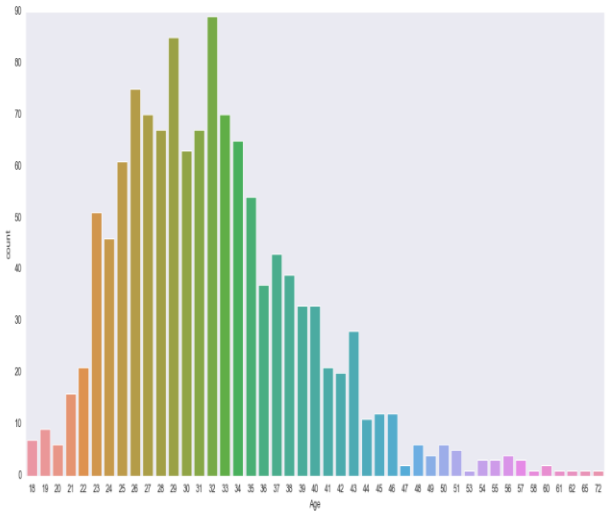
`The year 2014 had the most data captured. Could we safely assume that 2014 was a particular bad year for people that caused escalations for mental health issues? Or was it just the way the survey was sent, perhaps more interest in 2014 for people to partake in the survey.

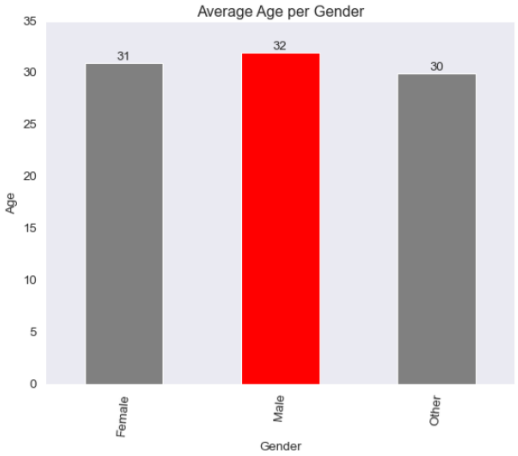
- There were a few historical events that happened in 2014 that might have led to this: (https://abcnews.go.com/International/biggest-news-stories-2014/story?id=27466867#:~:text=Major%20news%20events%2C%20including%20the,all%20corners%20of%20the%20globe.)`



Data Observations:

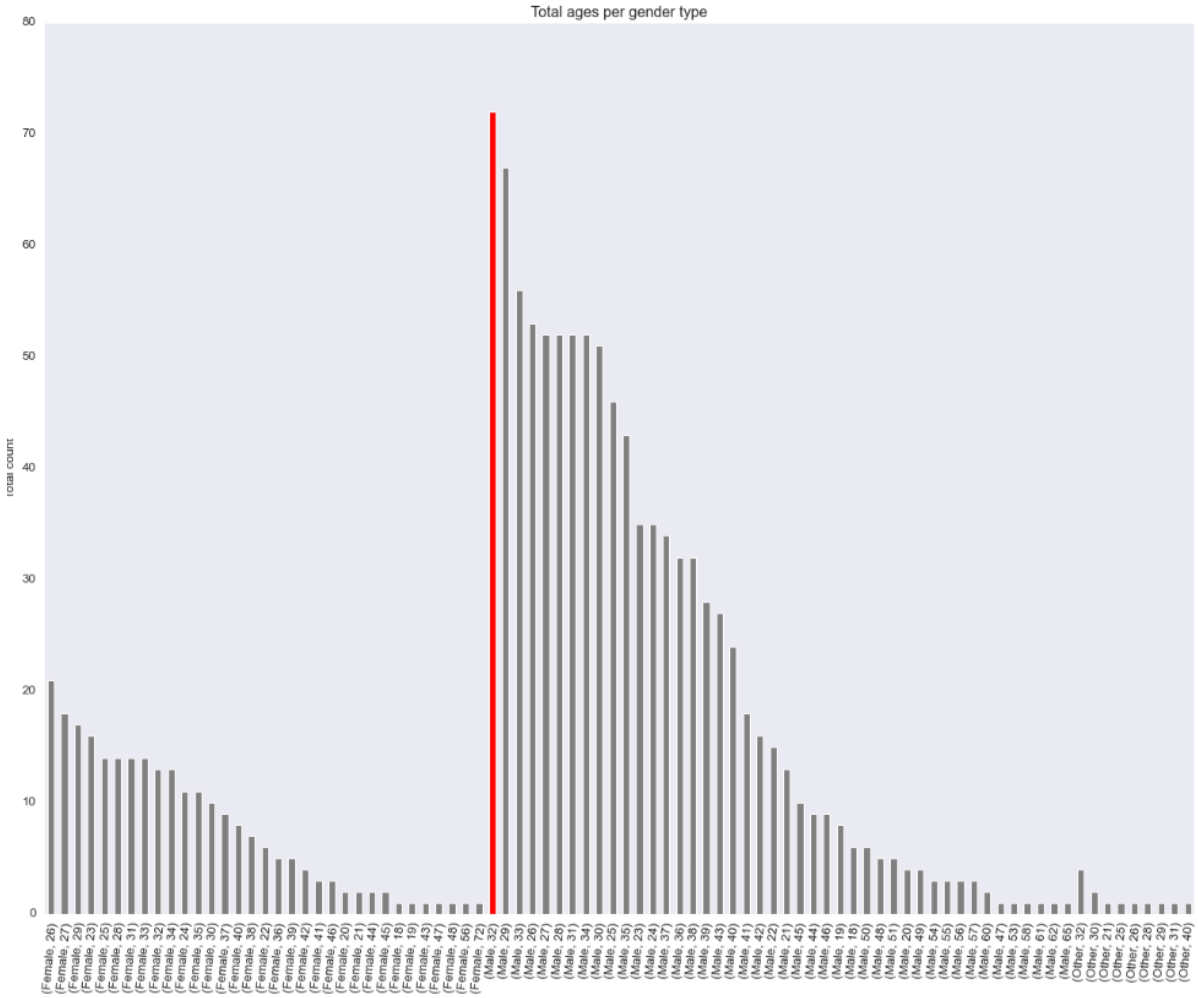
The majority of people that partook of the survey were men. This could also mean that the survey was specifically targeted for people that suffer with some form of mental illness. It is therefore assumed that men suffer more with mental illness than women. Now let's look at the age groups.

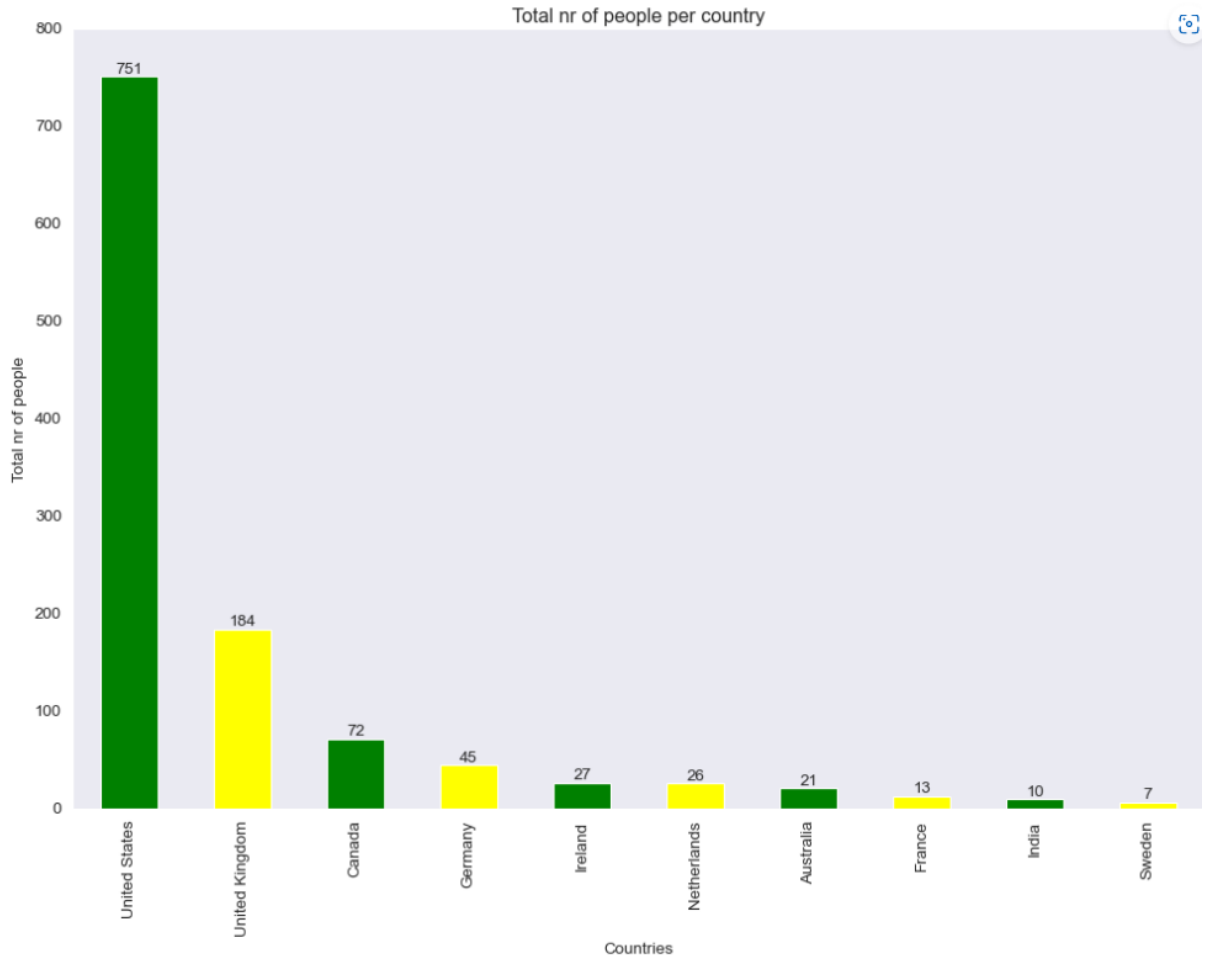




Data Observations:

Most people that partook of the survey were around the average ages of between 26-35. This seems to indicate that during this age gap, young adults seem to suffer more with mental illness. We also know that the average age is 32.



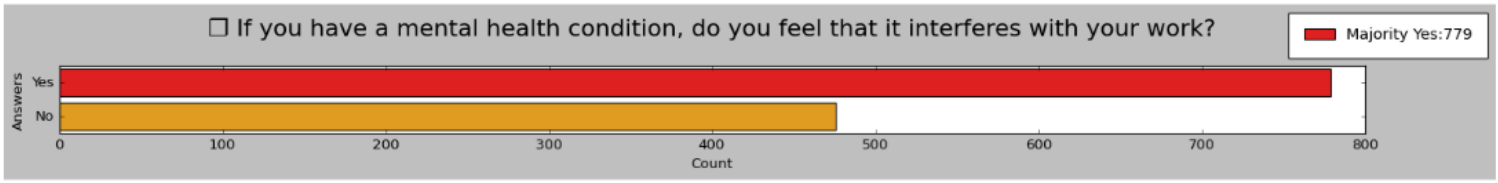

Total nr of people per country
Total nr of people
800
700
600
500
400
300
200
100
0
751
184
72
45
27
26
21
13
10
7
United States
United Kingdom
Canada
Germany
Ireland
Netherlands
Australia
France
India
Sweden
Countries

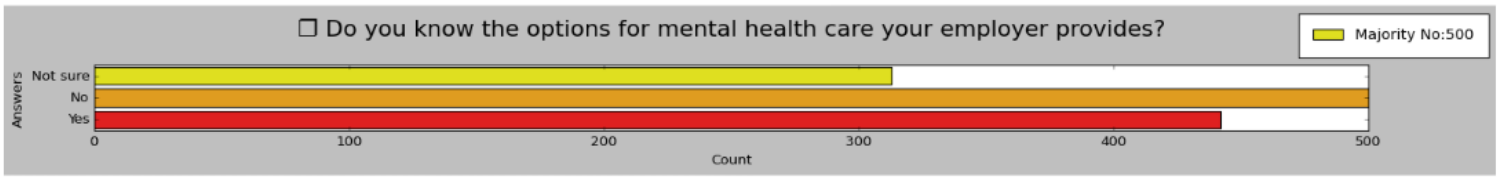
Data Observations:

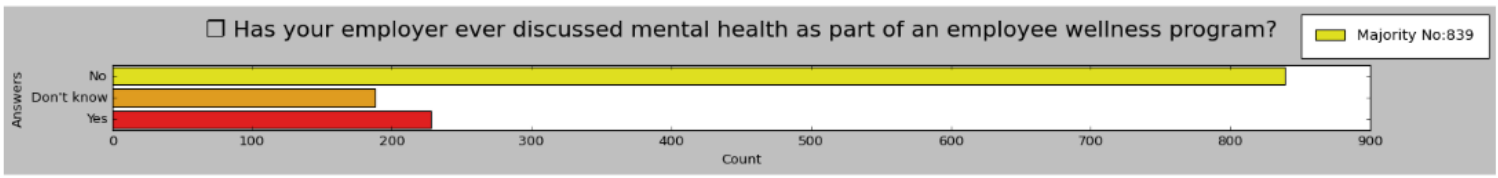
From both charts above I can see an overview of the countries and total of people in general as well as gender group. It is clear the **United States** and **United Kingdom** have the highest amounts. This could mean that more people face mental illness in those countries or this is more of a awareness as opposed to other countries.

The top 5 countries are all **first world countries**, this could be a factor as there is perhaps more stress in the workplace?

It is also interesting to see highest number of **males** are in US, UK, Canada, Germany, and Ireland. Also, the US and UK seem to have most of the **'other'** gender as they are countries that have embraced various gender types more openly than other countries.







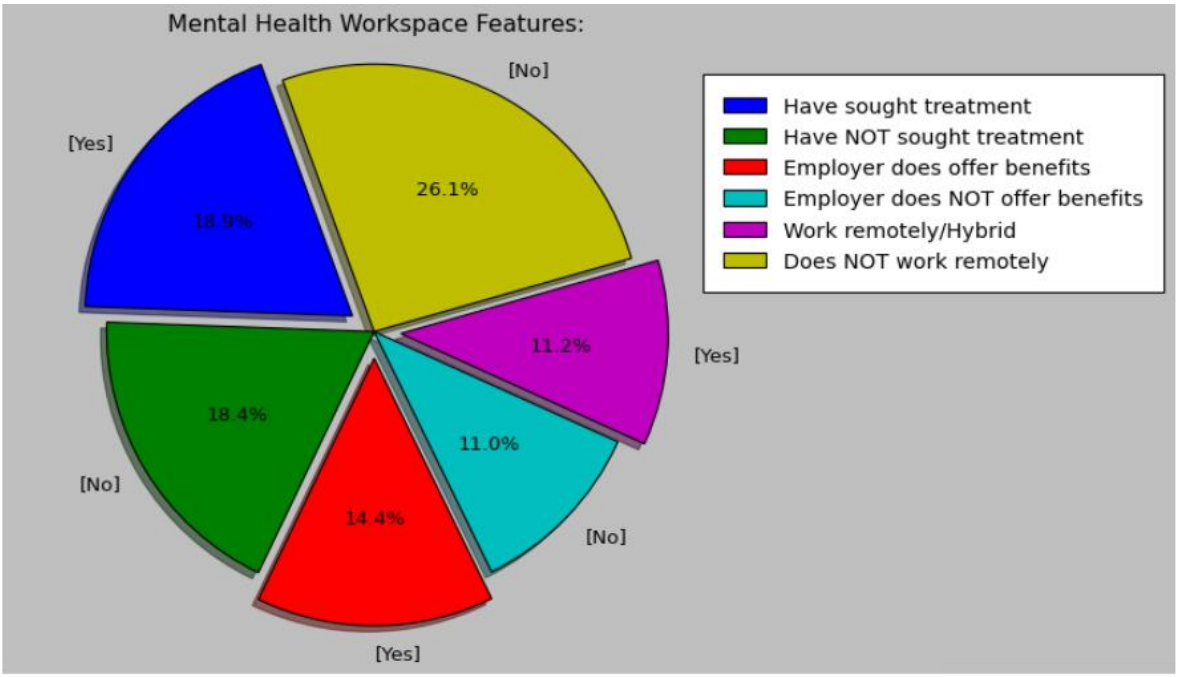
Data Observations:

From focusing on the survey relevant to the effects and benefits of mental health within the workplace.

We can clearly see that people who suffer with mental illness do struggle with their work, and it **does interfere** with their performance at work.

Even though the majority did say that their employer offers mental health benefits, the margins compared to those who said no are very close. Employers that offer benefits = **483**, don't know = **400**, No = **380**. That means roughly employers that don't offer benefits and those who answered know total roughly about **780**!

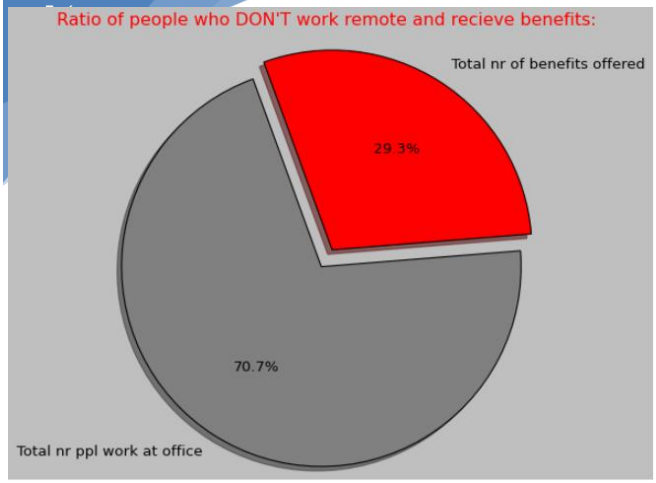
It is clear from the above that employers do not offer much benefits or any type of health wellness programs ect. This is strikingly alarming as the numbers suggests that mental health interferes with their work.

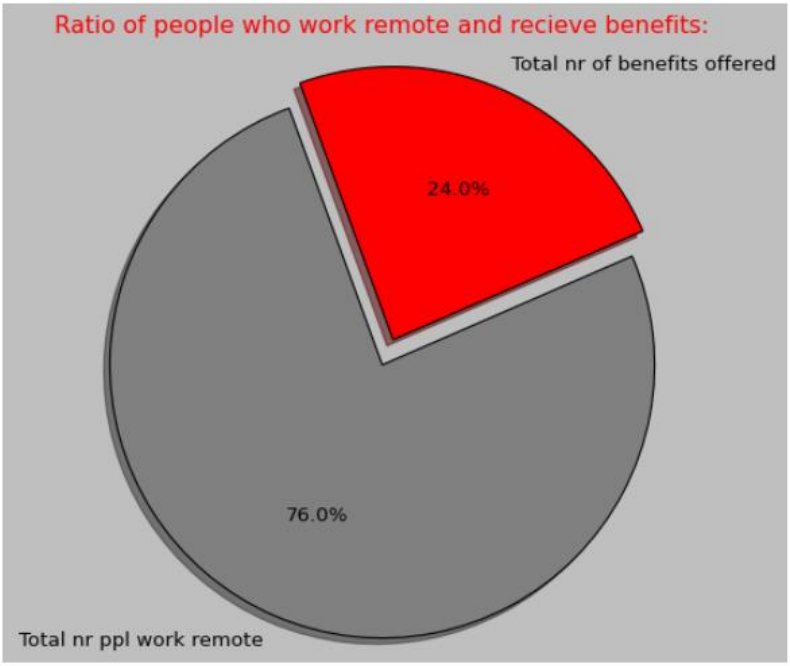


Data Observations:

From the data within the pie chart we are able to see the ratio's of the benefits, treatment and remote work sections. We can see that majority of the people **do not work remotely**. It is also interesting to see the people **who have sought for treatment** is higher than the people being offered benefits at work.

We might conclude that more people are seeking treatment outside of work, due to their being little or no benefits.

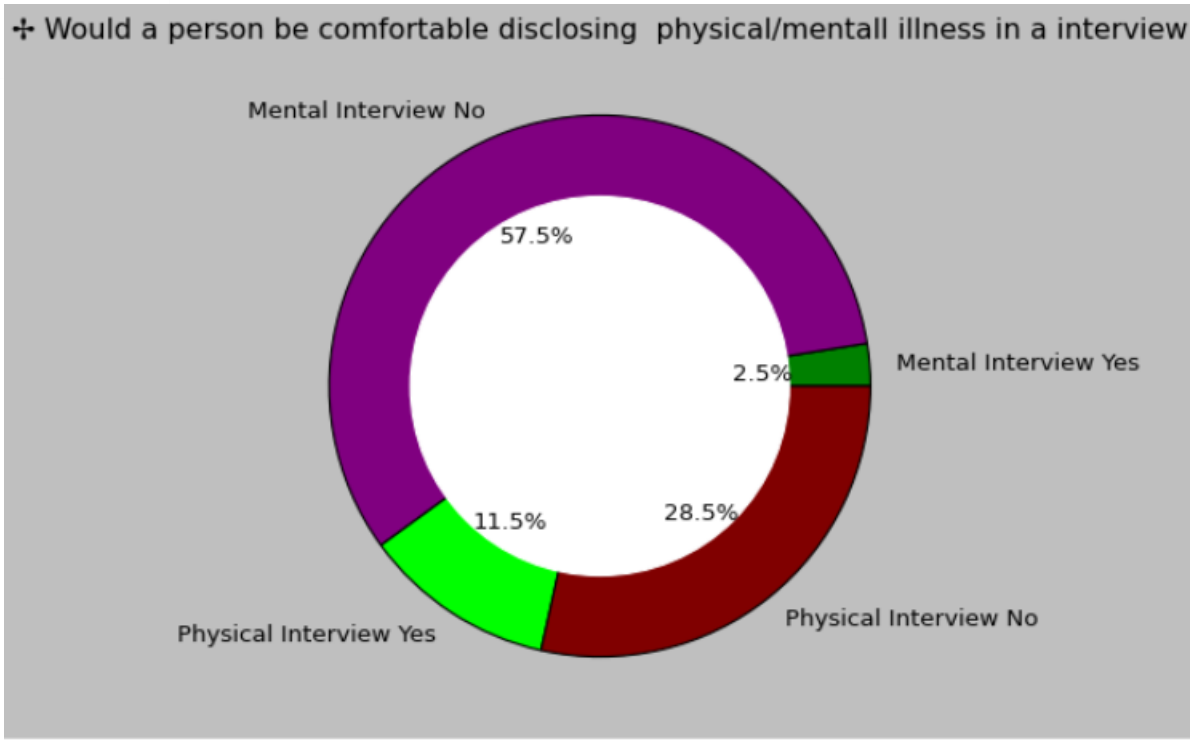




Data Observations:

From the data above we can conclude that **people that work remotely have a lower percentage of benefits compared to people that work in a office**.

This could mean that generally companies that have office a larger and have more money to spend on benefits for staff or we could say that the people who work remotely are less likely to need benefits. People who work at home might feel more comfortable or feel more in a 'safe' place which might assist their mental health interferences with work.

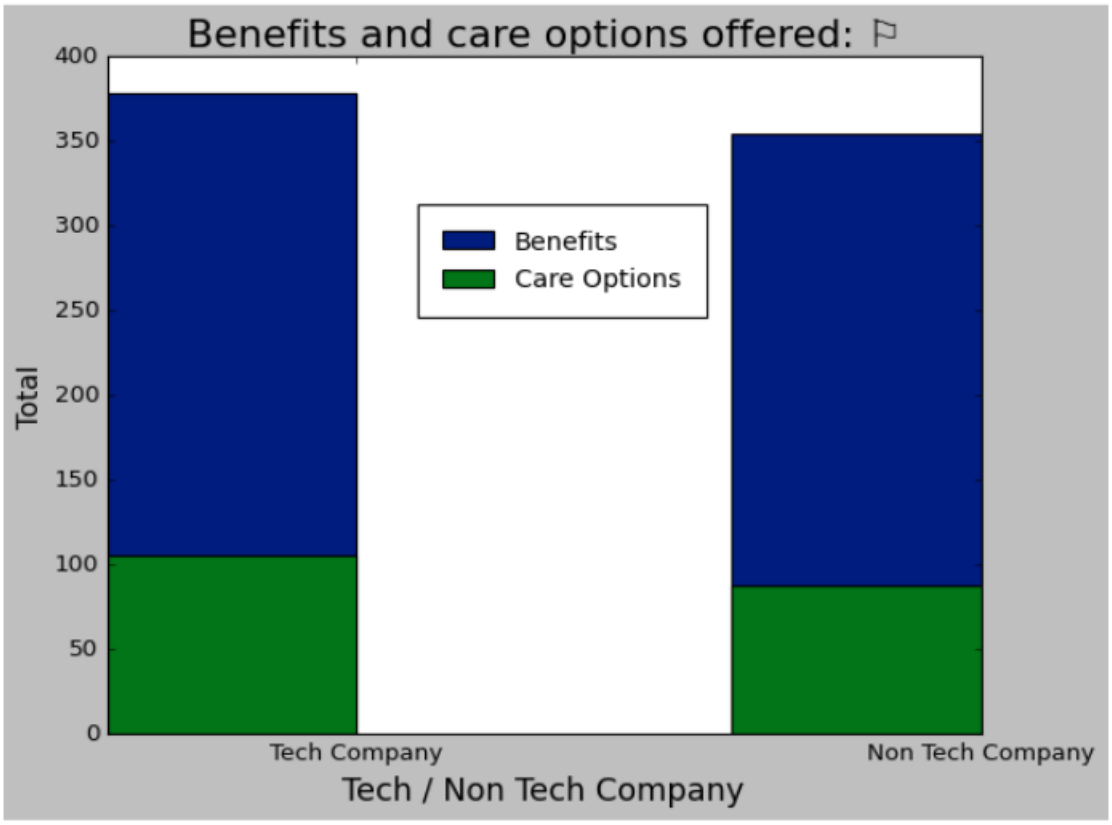


Data Observations:

From the data above it is clear to see that people would be more willing **not to share any illness** within an interview as the stigma might cause them to miss the opportunity.

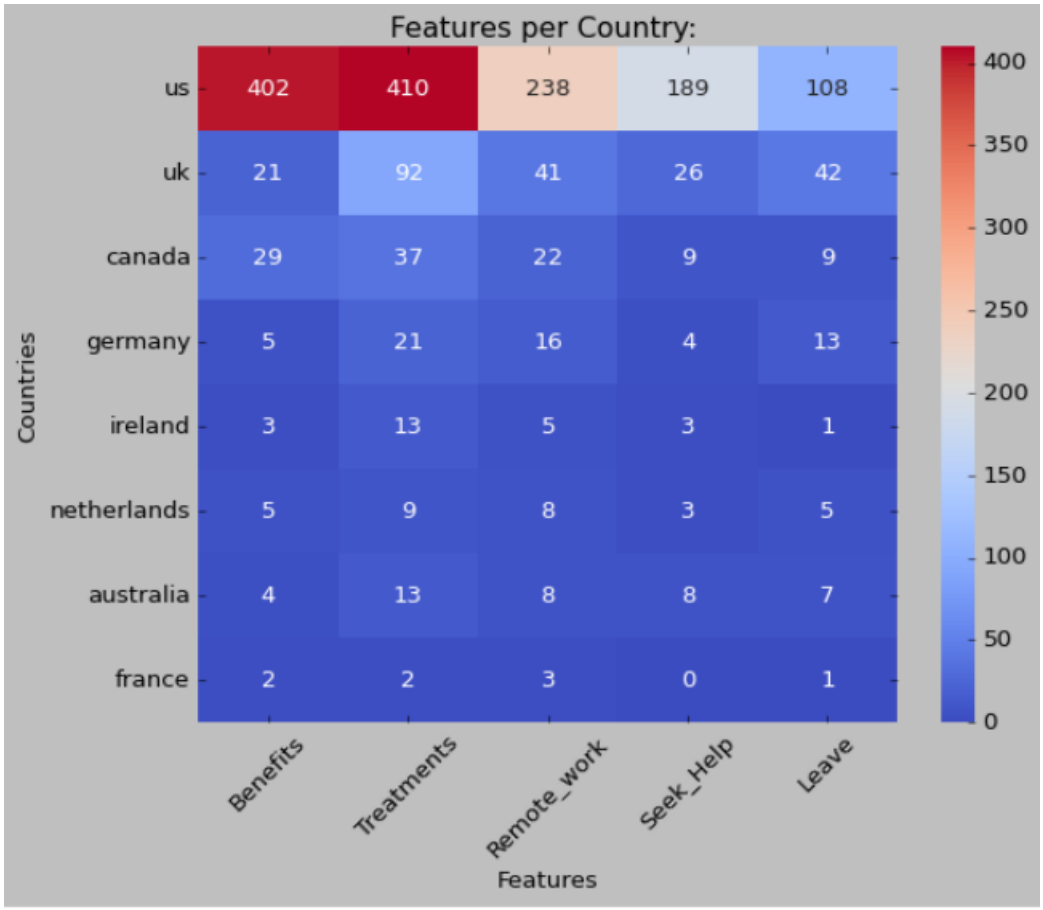
Also, people would be more willing to share they have a physical illness more than a mental illness, again this just shows the stigma attached to mental illness.

We can conclude that based of people's experiences it seems that companies do discriminate based on mental/physical illness.



Data Observations:

From the data above we can see that tech companies do offer slightly more benefits and care options, but the ratio between the company types is very similar. Therefore, we cannot say that tech companies necessarily are more opened minded or accommodating towards mental illness with their employees.



Data Observations:

From the data above we can see the differences of features compared to one another based per Country. The **US** an **UK** offer more benefits and the **leave** ration is also high. The benefits might therefore be part of extra leave days. **France** has the least amount of leave days and also people are not confident to ask for leave. Those who sought for treatments were higher in comparison to the benefits and seek help offers from the companies. This means more people **were seeking outside help due to less benefits or help offered by companies**.

Data Observations:

From the visualisation above we can see what word stand out the most, and what words were generally used to get an overall sense of what people's emotions are towards mental health in the workspace.
The words expressed are generally **negative.**

Words that stand out:

- Mental Health
- Employer
- Health issue
- Problem
- People
- Issue

Conclusion:

We can see that with the dataset that there needs to be more a tolerance and awareness of mental illness in the workplace. If companies embraced it more openly and got rid of the stigma's there would be more acceptance. By doing this is would cause less issues and interferences in the workplace. Companies should also allow more benefits, leave and care options for people suffering with mental illness. This dataset was conducted in 2014 and it would be interesting to see if any recent surveys has shown an improvement.

THIS REPORT WAS WRITTEN BY : Warrick Sabatta