DATA SQUAD 32

Exploratory Data Analysis

FRANCLIM CARDOZO

IRONHACK

2019

1. Main Goal

The dataset to be analysed is dated from 2014 and was based on a survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. The dataset contains many useful information that may lead to the answer of some relevant questions. The questions proposed in this report are the following:

- Considering Male and Female genders from the dataset, which gender has more mental health problems?
- If a person has a mental illness, it is better to discuss it with co-workers or supervisors?
- How many people are receiving treatment and are mentally ill?
- How many people have a family history of mental illness and are ill?
- How many people feel that a mental health condition might interfere with work?
- What is the country with more people with mental health conditions?
- Is there any correlation on working remotely and mental illness?

2. Description of dataset

This dataset has 27 columns and 1259 rows. Each column accounts for a different feature analysed after the survey was taken. According to the number of rows, 1259 people took the survey in 2014. There are many features that can be analysed in this dataset, including, gender, age, country, family history, treatment, interference of work, mental health consequence, co-workers, supervisors and a few others. These are the features that will be more relevant to answer to the proposed questions above.

3. Description of what you did

To get started, the file containing the dataset was imported and read thoroughly in order to have an idea of how many columns and rows are in the dataset and to understand the type of data to be analysed and the importance of it. Once the questions were proposed the dataset started to be cleaned according to each question. However, a few general steps were taken beforehand as follows:

1. Size of the dataset;
2. Getting the shape of dataset (number of rows and columns);
3. Understanding the descriptive statistics of the dataset;
4. Getting the column names for further reference;
5. Filtering the dataset according to the relevant question;
6. Looking at unique columns to see what type of answers were answered (some of the answers in Gender column were not written in a consistent and meaningful way, so some word cleaning was done);
7. It was decided not to consider any answer that is not Female or Male, because it is assumed people were not serious when taking this survey. To do this, data was grouped by mental health consequence and gender columns, and these out of the context answers were not considered;
8. Converting strings Yes and No to numbers such 1 and 0 in a few columns;
9. Creating a drop lists to remove rows that were no longer necessary for the analysis (e.g. in mental health consequence column the answer Maybe was reduced because it was considered as a vague answer);
10. Plotting mentally ill people by Gender;
11. Plotting mentally ill people by Country;
12. Getting different tables containing information that will help answering some questions;
13. Making a correlation between two features of the dataset (Remote Work and Mental health consequence).

More questions and analysis could have been done, but to meet the deadline, I decided to stick to the initial questions and try to answer all of them based on the data cleaning and manipulation done.

4. Description of results

After cleaning and manipulating the dataset, the first analysis done was the relationship of mental illness with people's gender. Fig. 1 clearly indicates that that people that are mentally ill tend to be more from the male gender than the feminine.

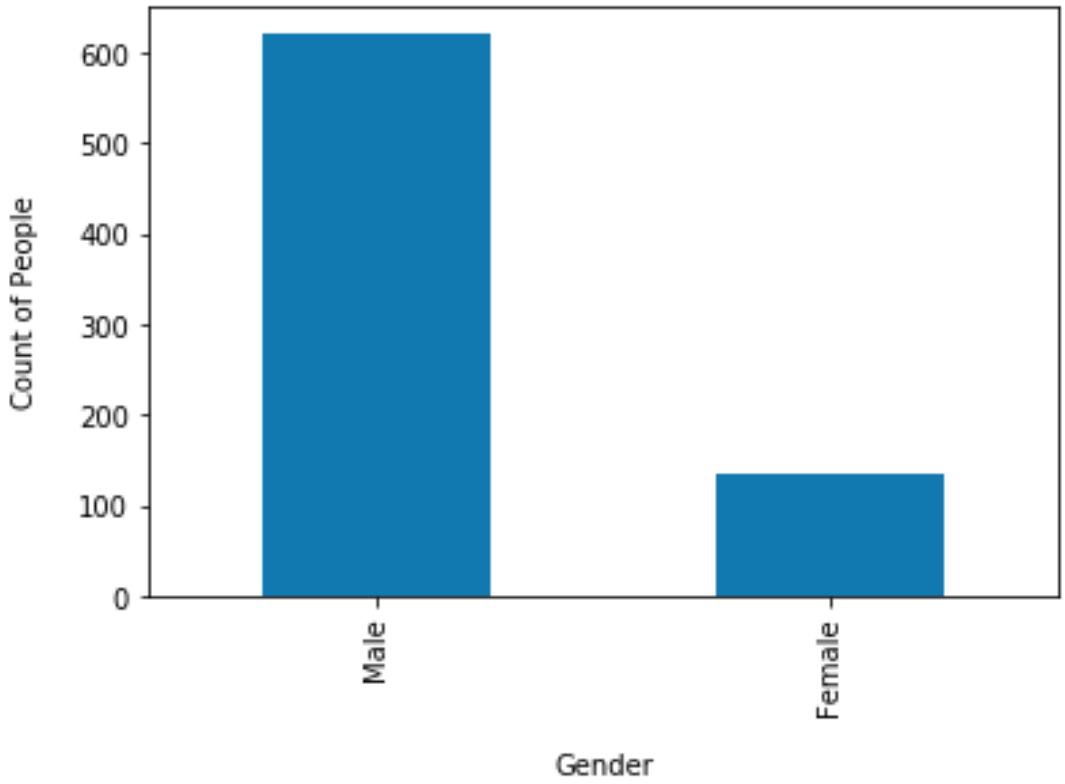

Fig. 1 – Number of mental ill people by Gender

From the total of people analysed and removing answers that were not relevant, approximately 620 ill people are male and 150 are female.

The next analysis was done on whether people tend to talk to their co-workers or supervisors if they have a mental condition (assuming that they were all diagnosed positive for this issue).

Table 1 – Number of people that prefer to rather discuss their mental condition with co-workers and supervisors

	mental_health_consequence	coworkers	supervisor
729	1	Yes	No
962	1	Yes	No
127	1	No	Yes
1118	1	No	Yes

As seen from Table 1 above, only two people prefer to talk to their co-workers and only two to their supervisor. Furthermore, there seems to be many people who have not answered yes or no to this question as this are the only yes and no responses for this case.

The next analysis consisted on answering the question related to how many people have a mental condition if they had a family background. To analyse this matter, the data was filtered for family history and mental health consequence and a total count of people who replied yes was done. In the Table 2 we can see that from the total people who have answered this question, 134 have confirmed that have a family background of mental disorders. It was assumed these are also the mental ill people.

Table 2 – Number of people that have a mental condition but had a family history of the same problem

	family_history	mental_health_consequence
3	Yes	1
12	Yes	1
25	Yes	1
31	Yes	1
59	Yes	1
...	...	...
1228	Yes	1
1238	Yes	1
1248	Yes	1
1252	Yes	1
1256	Yes	1

134 rows × 2 columns

The following analysis approach was related to the number of people who have a mental condition and have received or are receiving treatment.

Table 3 – Number of people that have received treatment and have a mental condition

	mental_health_consequence	treatment
3	1	Yes
12	1	Yes
25	1	Yes
59	1	Yes
60	1	Yes
...	...	...
1238	1	Yes
1247	1	Yes
1248	1	Yes
1252	1	Yes
1256	1	Yes

174 rows × 2 columns

From the results obtained of the Table 3 above, the total rows were counted, and it was concluded that 174 people that have a mental problem are receiving treatment.

In this survey, people were also asked about if they feel that work interfere if they have a mental condition. So, to know if work affects people with mental condition, the dataset was group by number of people and they different answers which were never, often, rarely and sometimes.

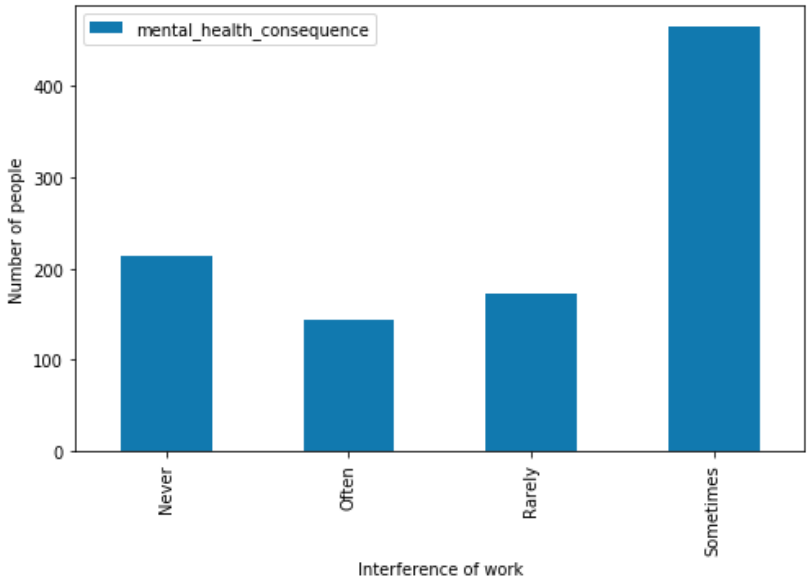

Fig. 4 – Number of people that believe work might interfere if they have a mental condition.

From Fig 4, we can conclude that only 144 people consider that work might interfere with their condition whereas 465 people answered sometimes. This indicates that most of the people strongly disagree that work affects their mental condition.

The last analysis was to see if there was any trend on mental illness and countries around the world. After filtering the data, the data was plotted using a bar graph and the results were presented in Fig 2.

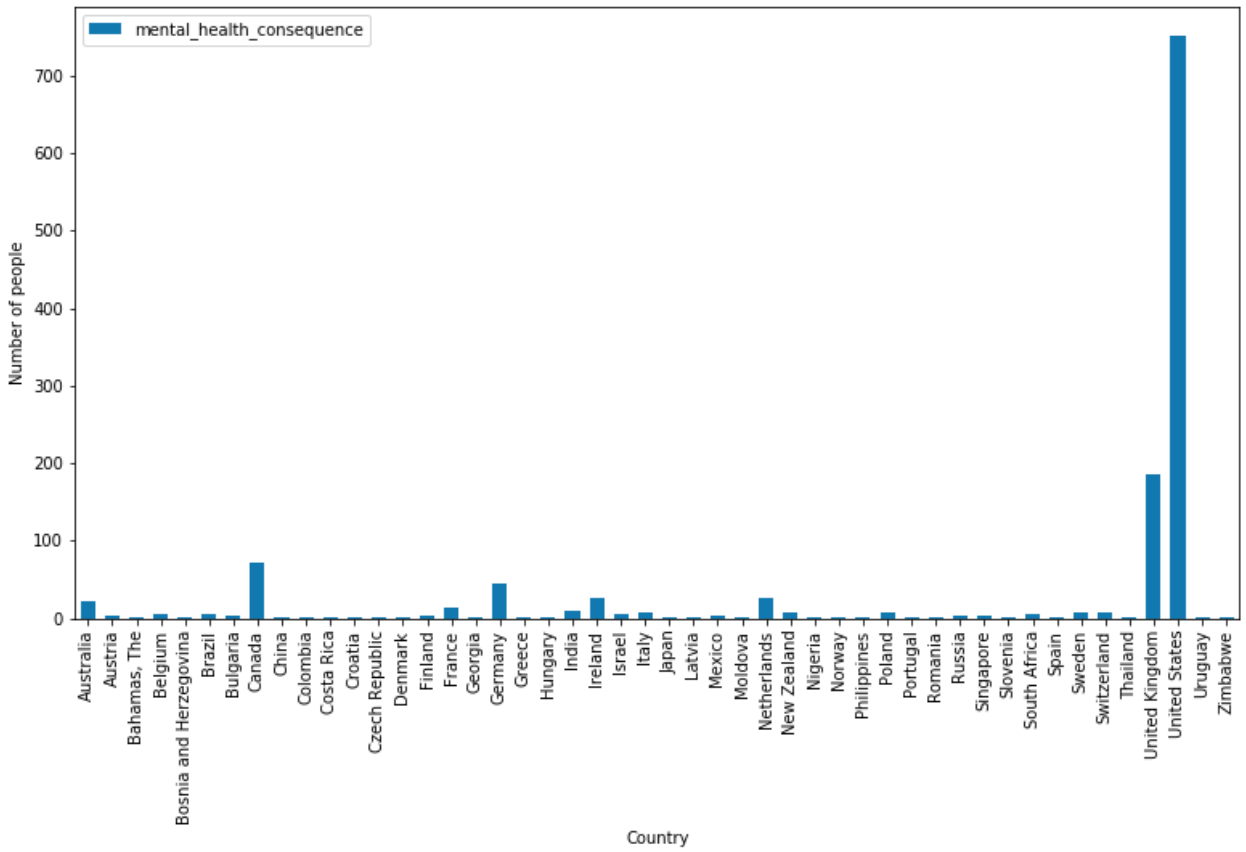

Fig. 2 – Number of people mentally ill sorted by country

As observed in Fig. 2, United States is the country which have more reports on people being mentally ill with a total of 751 people. After the US, United Kingdom takes the second place with a large difference on totals. There are 185 mentally ill in the UK.

Lastly, a correlation between two features was done. The features to test a correlation were remote work and mental ill problems. The correlation obtained was 0.00156 (0.156 %) which indicates a weak positive correlation between these two variables. According to the value obtained, there is no correlation between working remotely and having a mental problem, as this value is very close to zero.

This dataset had man other interesting features to be explore and compared and many other further analysis could have been done.