**Research Questions**

Although the dataset has responses to a wide range of questions been asked through this survey for people coming from different races age and gender, it’s very difficult to answer all the possible questions. Hence this research paper only looks to answer three questions but in detail with methodology so as to be as specific as possible in the assumptions and conclusions for the data.

1. Do people coming from a certain country feel that they have been supported and taken care of when they suffered with mental health issues?
2. Do People coming from a certain background, age group or sex tend to be open to their supervisor regarding their mental health?
3. What’s the general opinion difference between people coming from tech and non tech regarding mental health or in other words is Tech industry across gender, age and background creating an environment not suitable for better mental health?

**Data Preprocessing**

**Procedure:**

Mental health dataset here was taken from the website <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey/data> **, which is an extract from the source** <https://osmihelp.org/research.html>

Dataset here could be downloaded in various formats like xlss, csv and Json format but for the research paper csv format is chosen for the further data analysis process.

**Data Cleaning and Validation:**

For data cleaning, in this research paper R language is being used. To begin with the cleaning. First the data from the csv file was read into a data frame called “data” and since there were blank cells and R does not allow for blank cells to be read, these had to be converted into NA.

Now since there were null values in each column from Figure:1

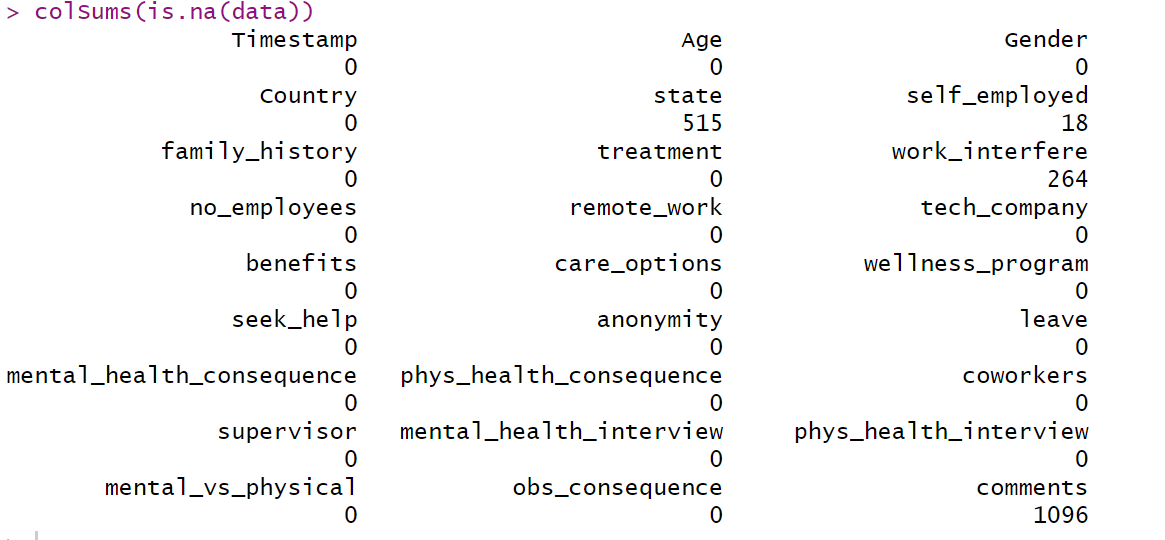


Fig:1

But specifically comments column having around 1096 NA values out of 1259 rows removing “comments” column from the cleaned dataset was to be done. Gender column had some misspelt values (Figure:2) which actually represented the same values and hence changes had to be done.

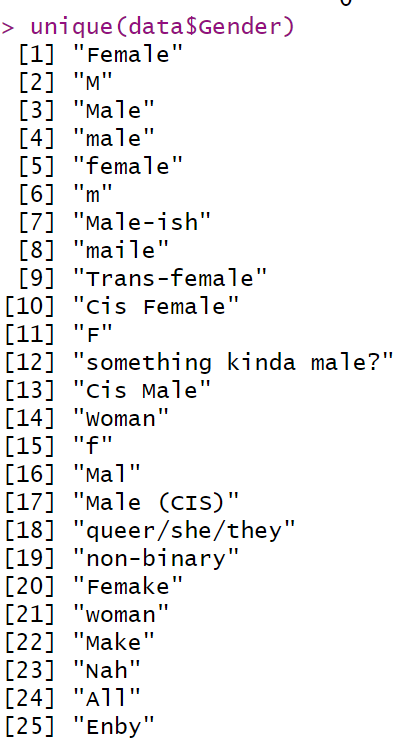


Fig :2

After standardizing values for male in Gender even standardizing values for females was also necessary as there were mistakes in that as well. So, the code for all the operations done until now is given in the figure 3.

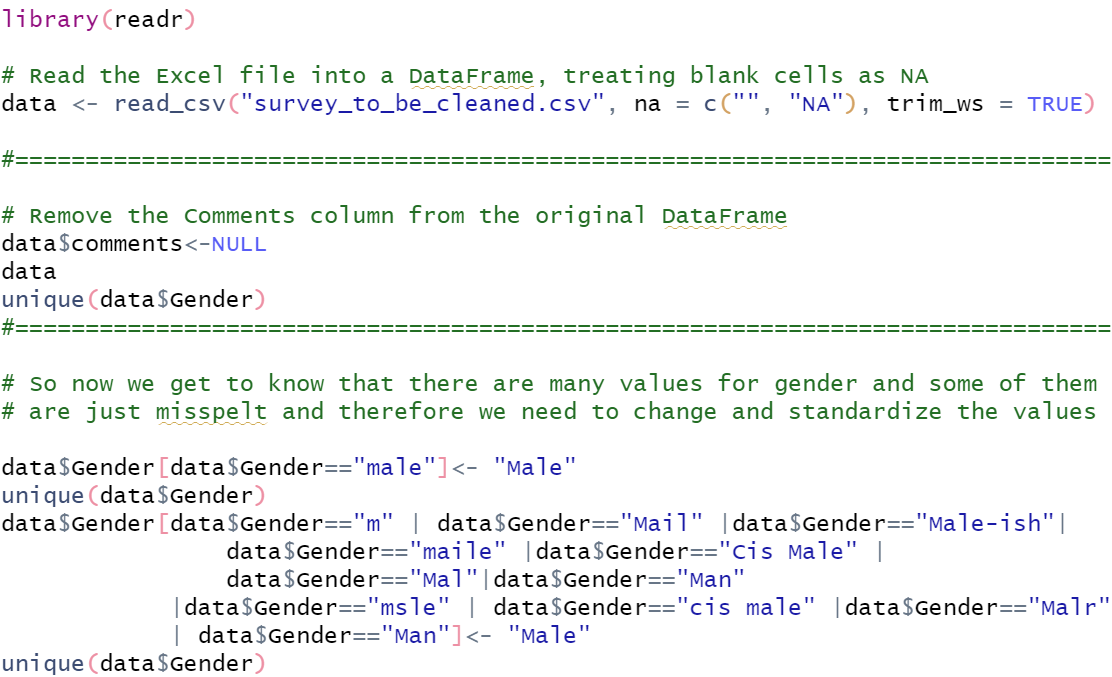


Figure:3

After that as we can see there are many absurd values for gender as “non-binary”,” Cis-Male” and” Trans-Female”. In this research we have decided not to categorize them into different and rather have placed them into a separate category called “**others**” as shown in Figure 4.

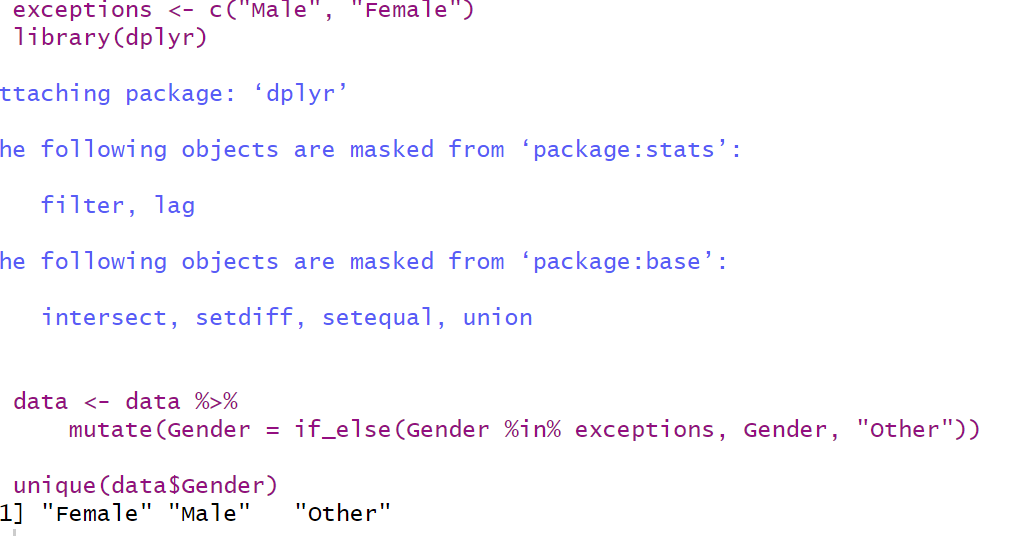
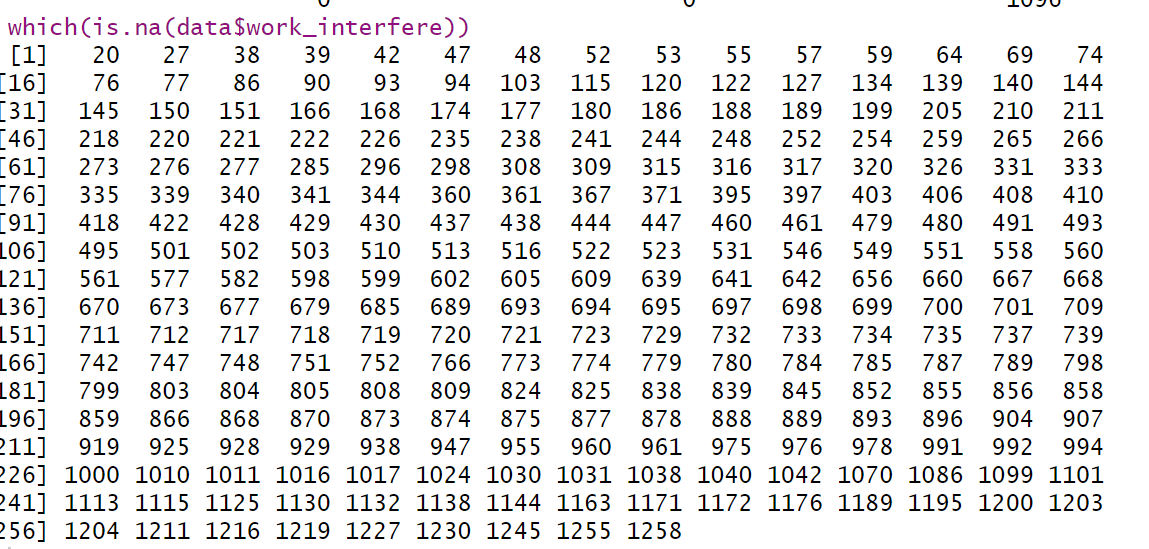


Figure:4

Going back to figure 1, what we observe that not only comments column had null values but also variables like state, work\_interference and self\_employed had null values. Moving ahead, as much of our analysis and research questions would not be involving state, it’s important to deal with the other variables so that data can be clean and be ready for analysis.

Although there are 264 values null in the dataset, which is 16% of the data, for “work\_interference” it’s important to understand how these values can be imputed and whether imputation of such values could cause any misrepresentation of data as it could lead to incorrect conclusions as well.

So, the first step in understanding whether there is a pattern in the missing data is to look for the rows with null values.  Fig:5

As we can see that the rows with null values seem to be random and don’t follow a pattern it gives an indication that the null values are just random and not due to reason but still to be sure it’s important to have a few more checks. For instance, since our first research question involves the discussion of mental health and its issues country wise, here we can make sure to see whether null values aren’t present for a particular country because if they were to be, then analysis won’t be able to stand on concrete conclusions.



Figure:6

Now through Figure:6 we get to know that countries that were categorized into “other” weren’t present. So, what this tells us is that: null values have been all random and moreover countries with less than 10 records are also not null values. Hence, it’s evident that these values can be imputed and can be used for further analysis.

**Imputation:**

For “**work\_interference**” and “**self\_employed**” being two categorical columns it was quite clear that the values with highest frequency had to be imputed for both the columns. But after getting to know that all the values within “**work\_interfence”** column are very close to each other in terms of count as shown in figure 7, it was important to fill the rest of null values randomly, as bias towards the highest frequency value had to be countered.

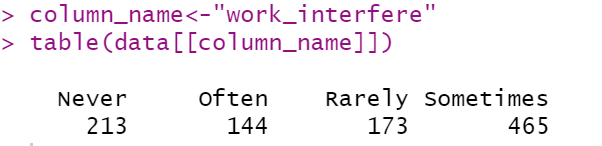


Figure:7

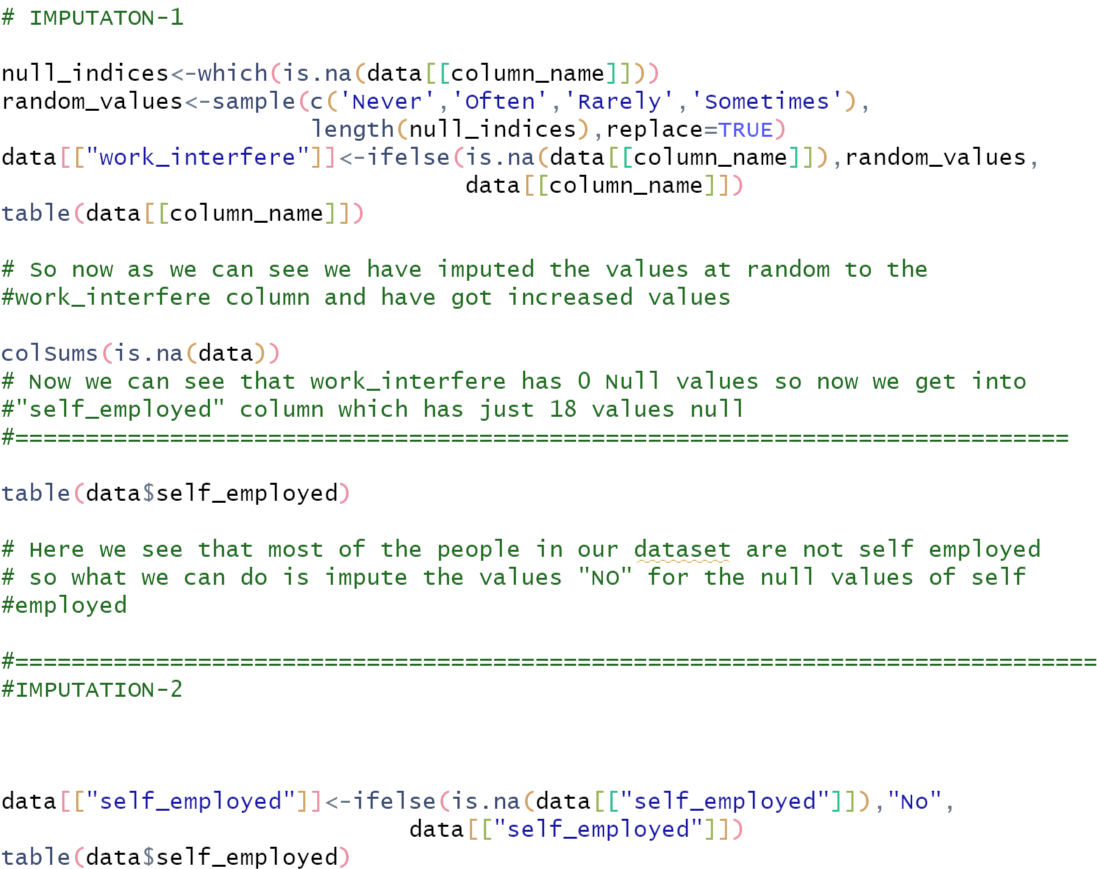


Figure:8

This figure depicts the coding for both the imputations and their explanations. On the other hand, for “self\_employed” column, which had only values either” Yes” or “**No**” and distribution among yes and no was not equally poised, therefore null values were imputed with “**No**”.

Moving towards age column, From the summary statistics in figure9, its apparent that age has some negative values as well as some really high values like 10^11.

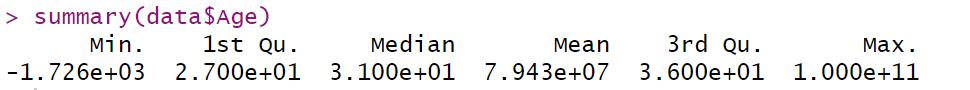


Figure:9

After treatment of the age column summary statistics looked like

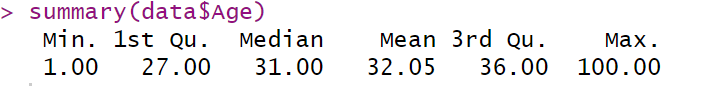


Figure:10

Since the mean value has changed, minimum value has been made to one (positive value) and median has remained the same closer to mean it appears to be a right decision of changing the maximum value to 100, which is still rare but doesn’t still look to harm the dataset as much.

**Analysis through Python**

After Data cleaning through R the file survey\_to\_ be\_cleaned is been read into a new csv file survey\_cleaned and is thus then used in Jupyter notebook for further analysis of the research questions. As we can see in Figure 18, File is been written to a csv and been provided a path so as to be retrieved by python.

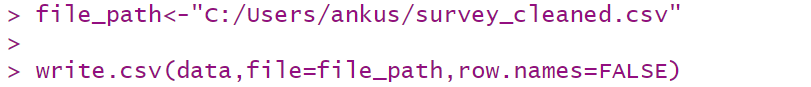


Figure:18

So, to begin with in Python, first the all the modules required for answering the research questions through analysis are been imported. Then the csv file survey\_cleaned is been read into a data frame “df” as we can see through snapshot of code in Figure 19.



Figure 19

Once the data is imported in a data frame, further wrangling operations and if any cleaning needed operations can be done. Although the data that we loaded into this data frame “df” is after cleaning data through R and shouldn’t be having any null values, we can still have a look at it (Figure 20).

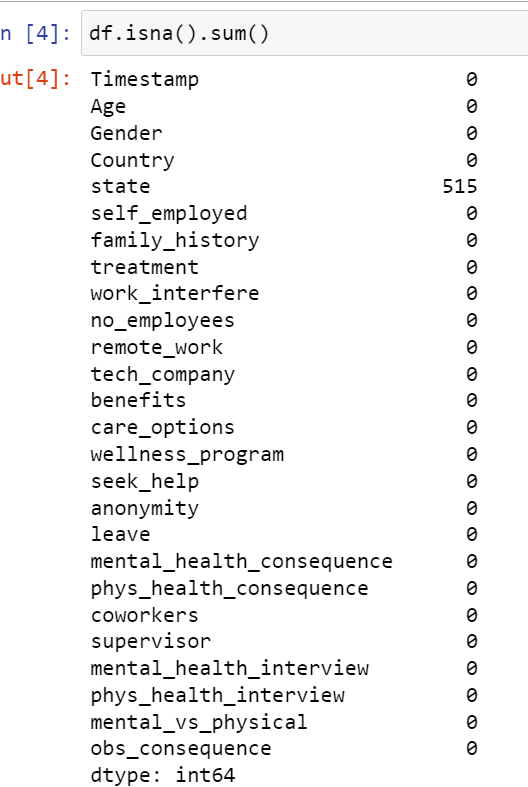
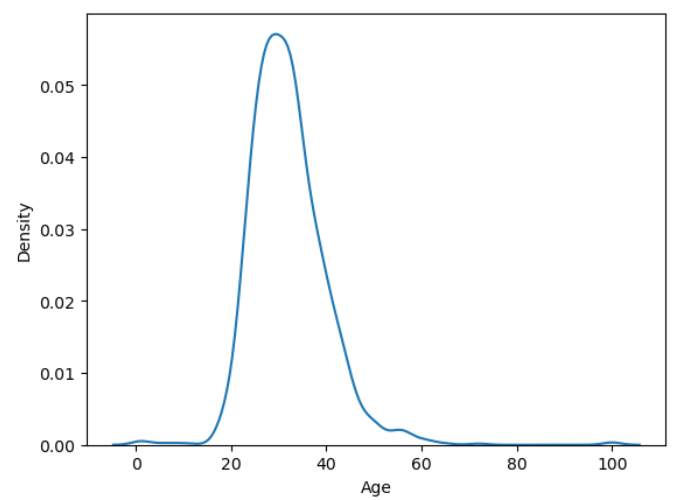


Figure: 20

It’s quite obvious that “**State**” is the only column which was supposed to be displaying null values as it was not previously dealt.

**Univariate Analysis**

Distribution plot for age

 Figure:21

**Value counts for all the categorical columns**

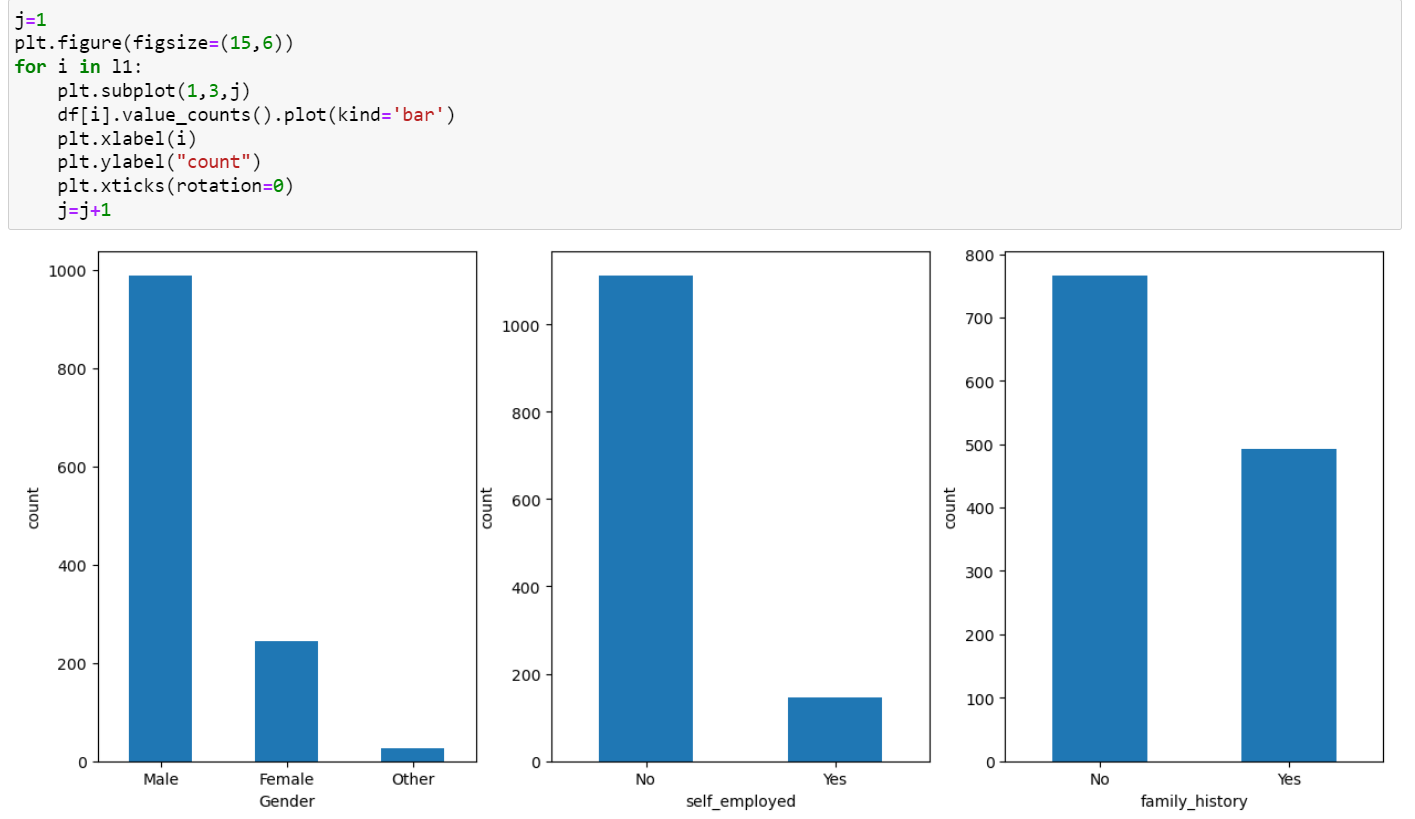
****

Figure:22 Count plot for Gender, Self\_employed and family history

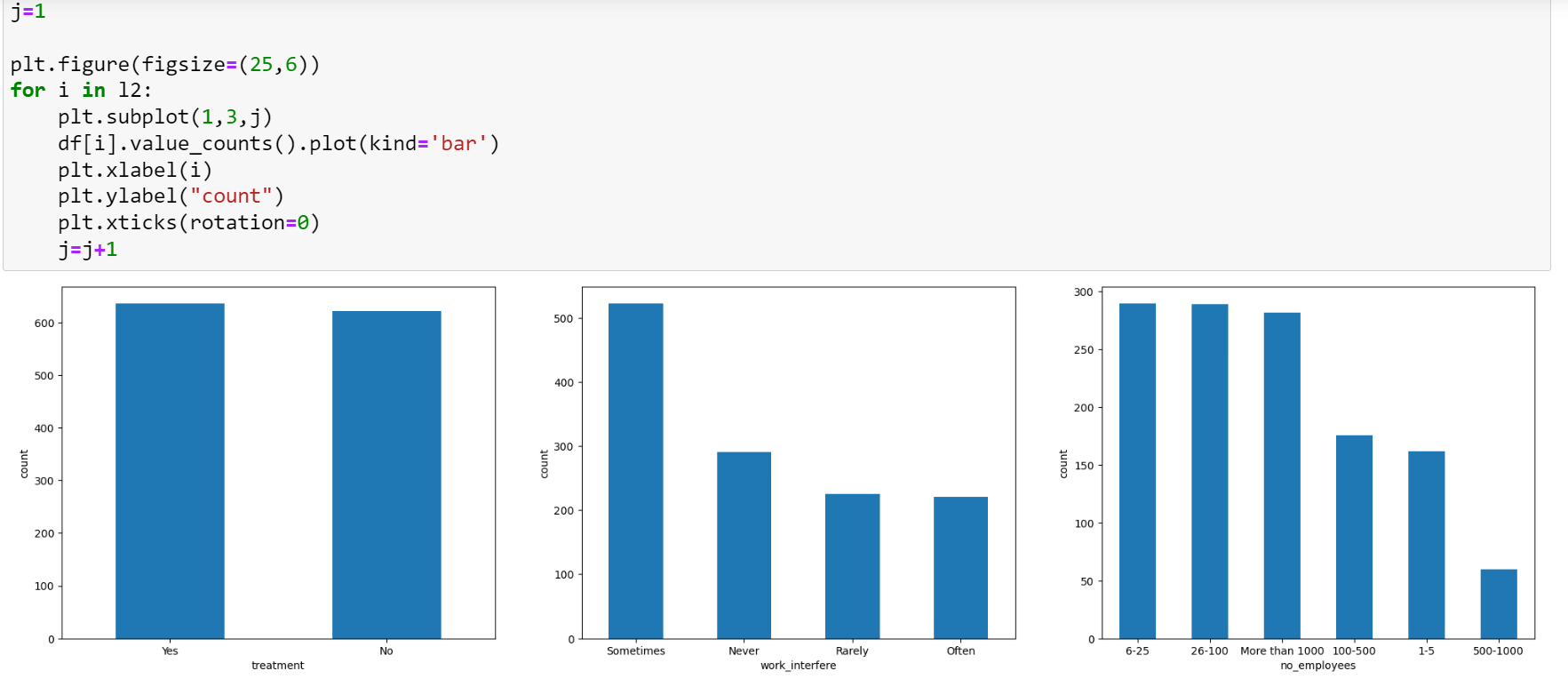


Fig:23

From Figure 23 we can see that Treatment, work\_interfere and number of employees count plots, Treatment column here signifies the responses of people as “Yes” and “No” for whether a person has taken treatment or not for mental health. Whereas opinion on work\_interfere talks about the opinion of people on whether work could interfere and cause mental stress to a person. And through the count plot its evident most of the people believe that mental health has much more to do in life other than work.

Whereas the count plot for number of employees is a subjective question, which talks about the range of employees present at their work place. Now it appears that most of the people surveyed either belong to very small-scale companies or belong to big companies and hence the data looks nicely distributed as most often it’s the variation in the extremities of an attribute, which provides some interesting results.

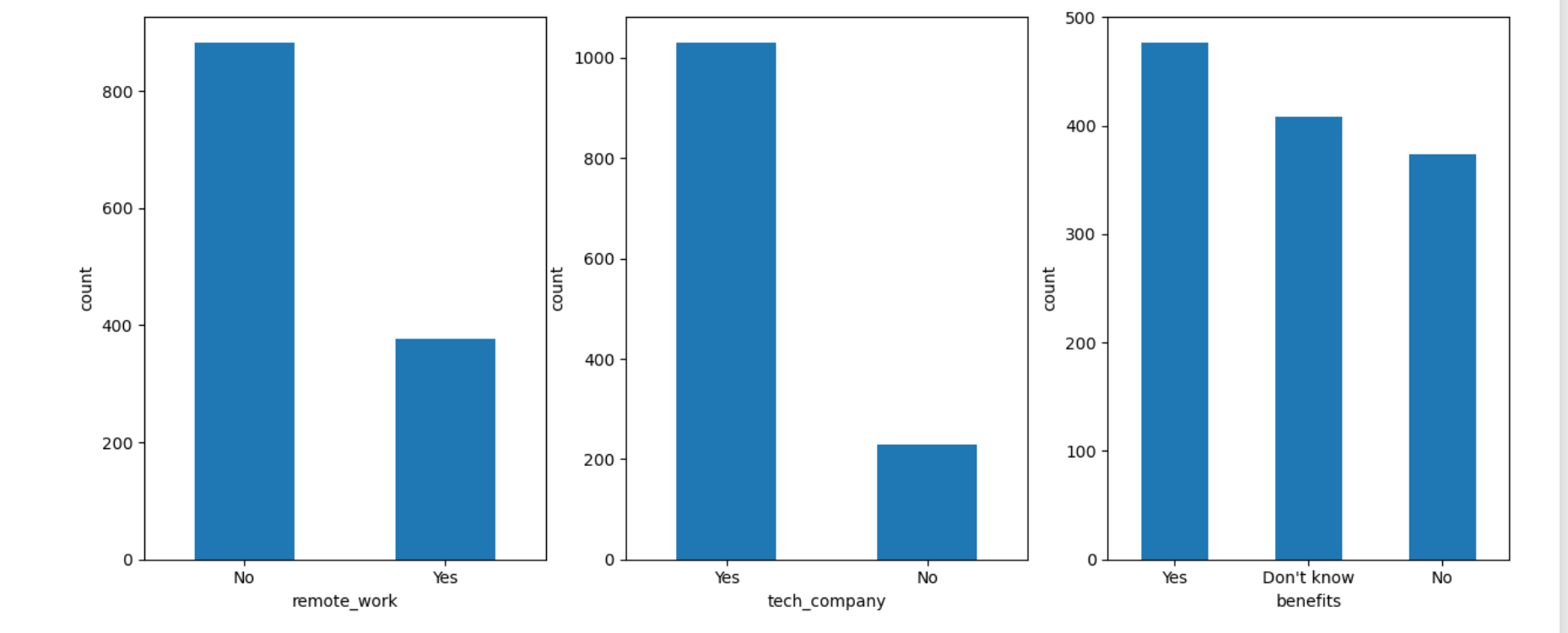


Figure:24

Here in Fig:24, Out of these three columns, it is the “benefits” column which asks an opinion question on whether benefits would be given by a company if the person suffered with mental health. To surprise, we can understand that most of the people think they will be benefitted but at the same time around 1/3rd of the surveyed people is clueless about the benefits and since they are not aware about it, it becomes even more important to understand “benefits” attribute from a mental health concern point of view. Further analysis on this attribute will be done to answer one of the research questions chosen for this dataset.

Similarly Care options, wellness programs and seek help are also some of the columns which describe the feeling of people regarding mental health through their work place.

Some other responses to the questionnaire of survey are about whether the person feels free to speak about mental health issue with coworkers and supervisors, Will a person bring up mental health issue or a physical health issue to a potential employer in an interview (Mental\_health\_interview, Physical\_health\_interview). Moreover, the survey also digs in to understand whether leave for mental health condition is granted easily and are there any observed negative consequences by a person for another coworker for mental health.

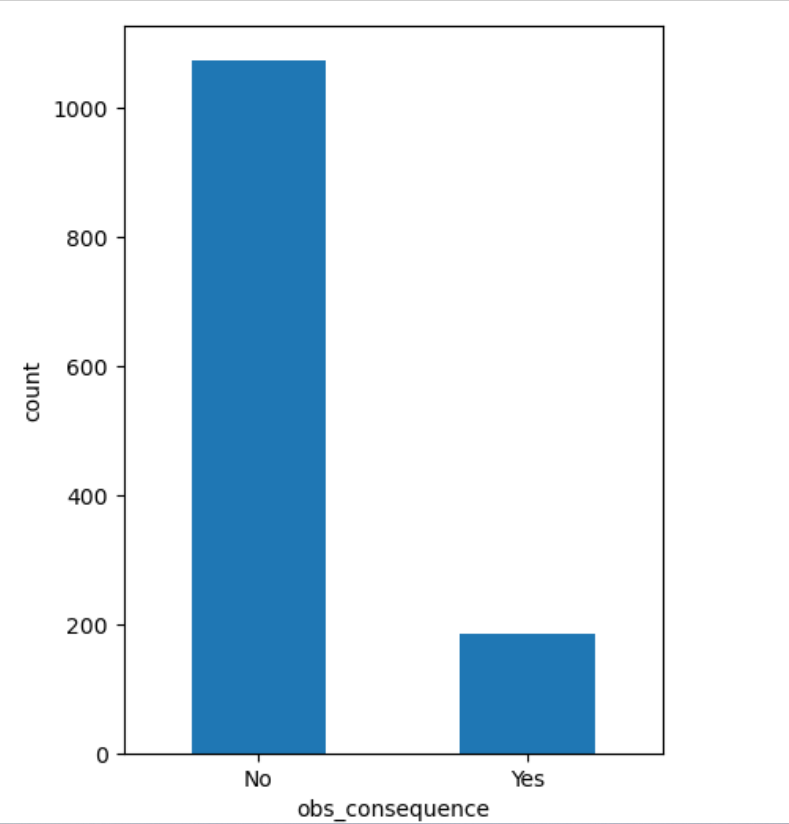


Figure 25

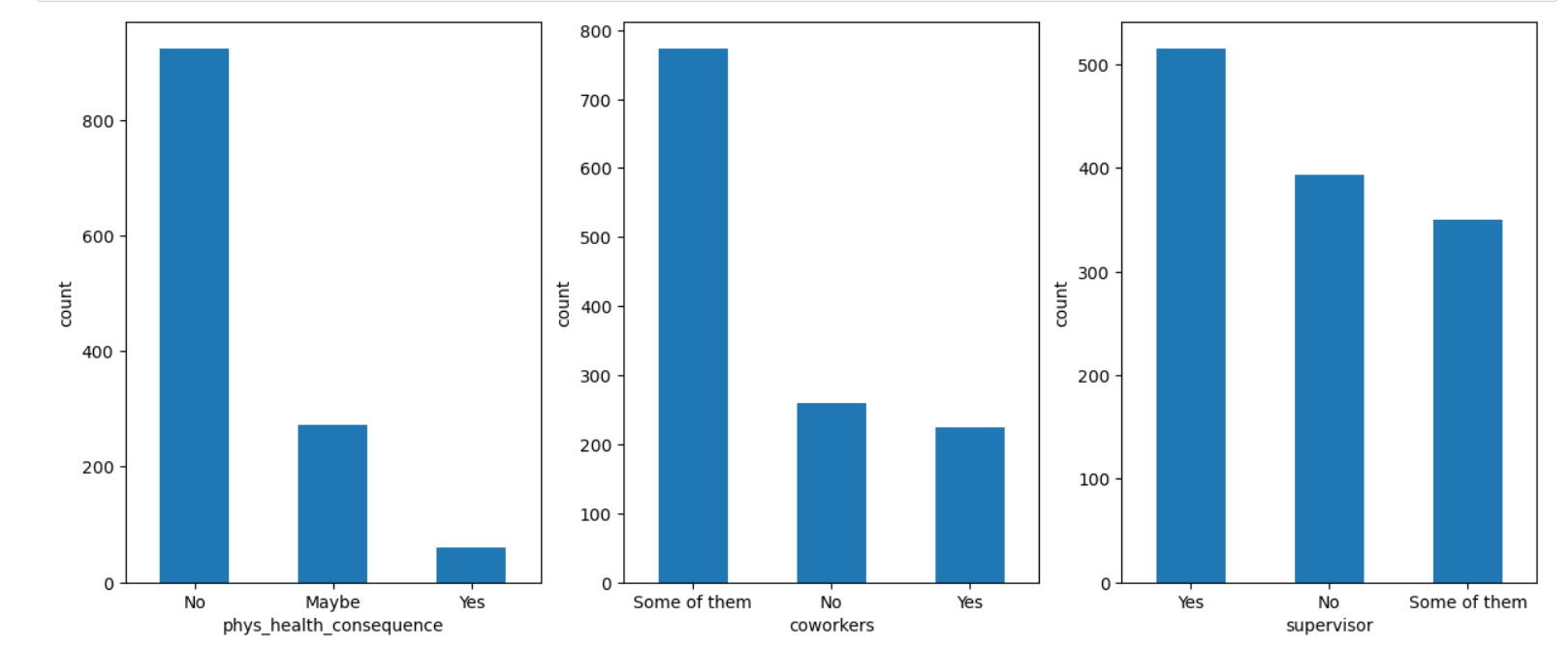


Figure: 26

**Multivariate analysis**

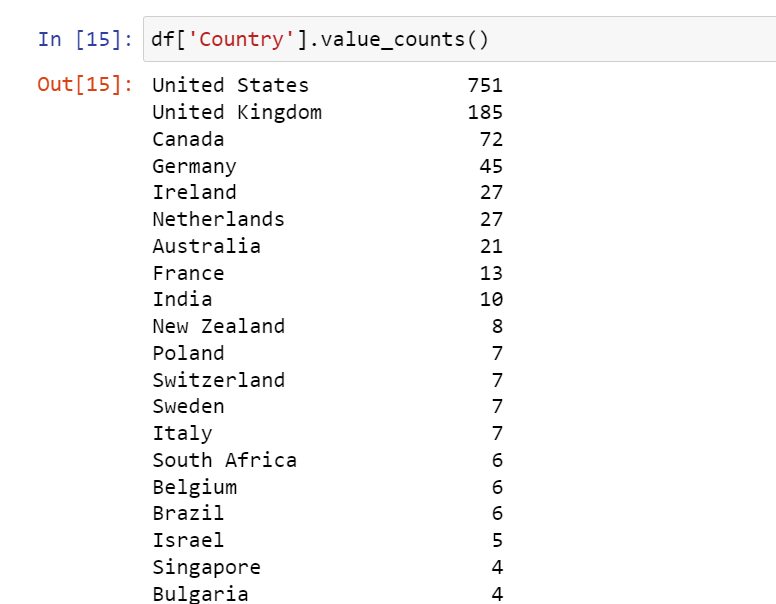
**First Research question:**

1. Do people coming from a certain country feel that they have been supported and taken care of when they suffered with mental health issues?

Procedure:

Moving ahead with this research question, we intend to find on whether each country has its own way of handling mental health issue. Since we have two to three questionaries corresponding to the given research questions. Analysis for each of these would be done considering the grouping variable to be country in the all the cases.

Now if we remember discussing about country column in the previous sections of the research paper, we could recall that “Country”, as per the survey has about 48 different countries present in the dataset with around 35-40 of countries having only single digit number of responses, which would have to be acting as a representative for a country in the analysis. Therefore, it’s important to understand this issue and resolve it before proceeding to the next steps.

Figure :27 & 28

**Strategy:**

To work out the country column, it is better to group them into three different categories; One of them being “United states” as it has the highest number of responses, the next category being Tier 2 where countries like U.K Canada, Germany, Ireland, Netherlands, Australia France and India could be clubbed up and then around 35-40 countries with less than 10 responses for each country could be grouped into a single category called tier3 countries.

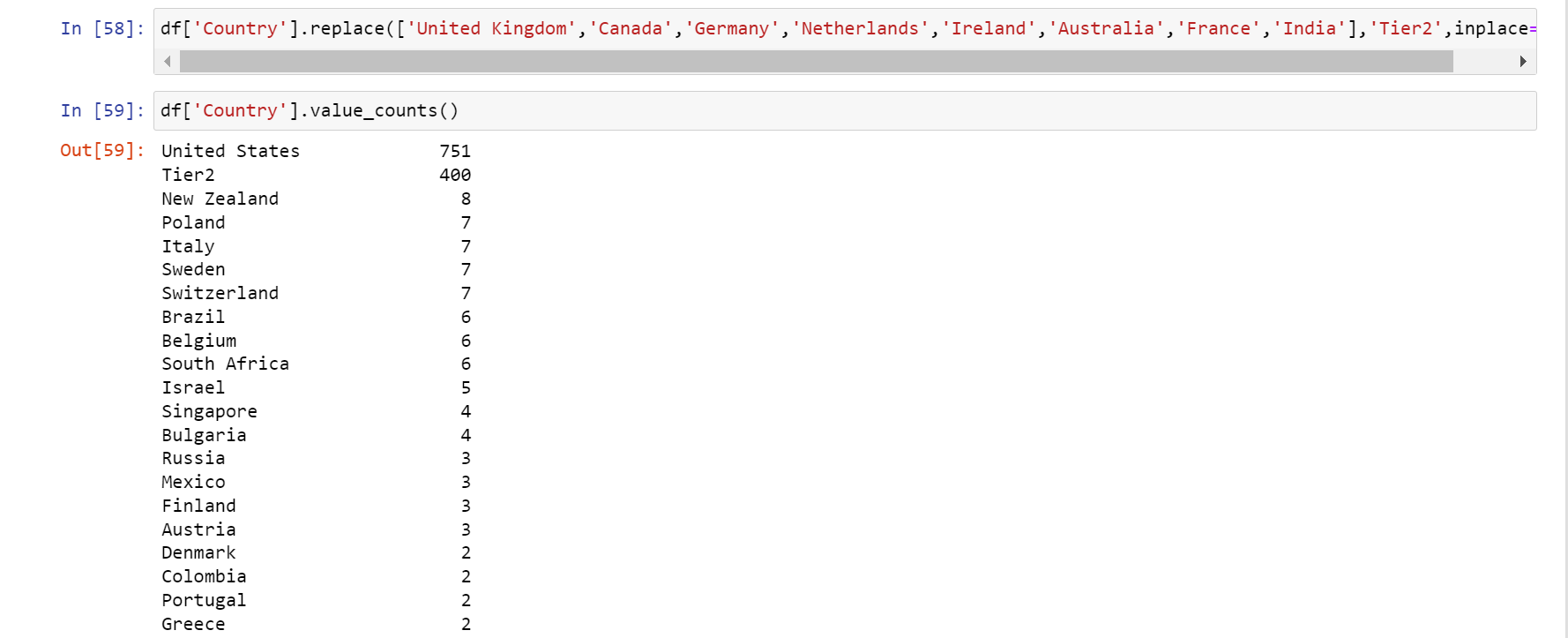
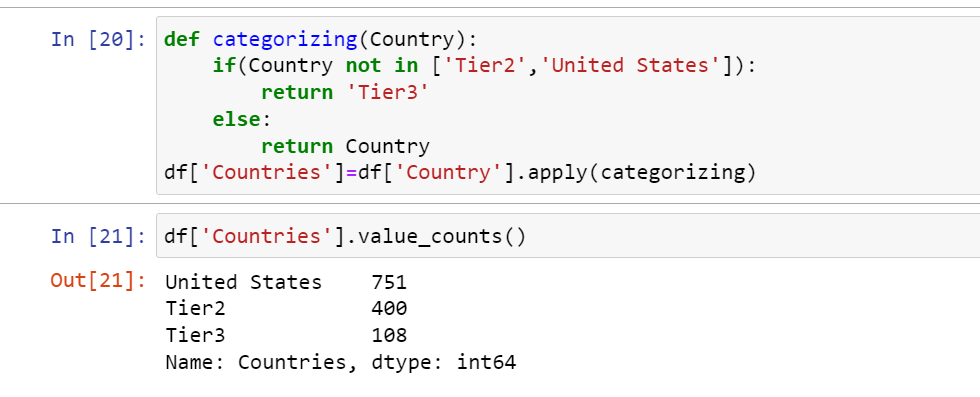
 Fig:29

Fig:29 & 30 describes the code related to changing the categories in the country column as per mentioned strategy

 Fig:30

**Insights on Benefits**

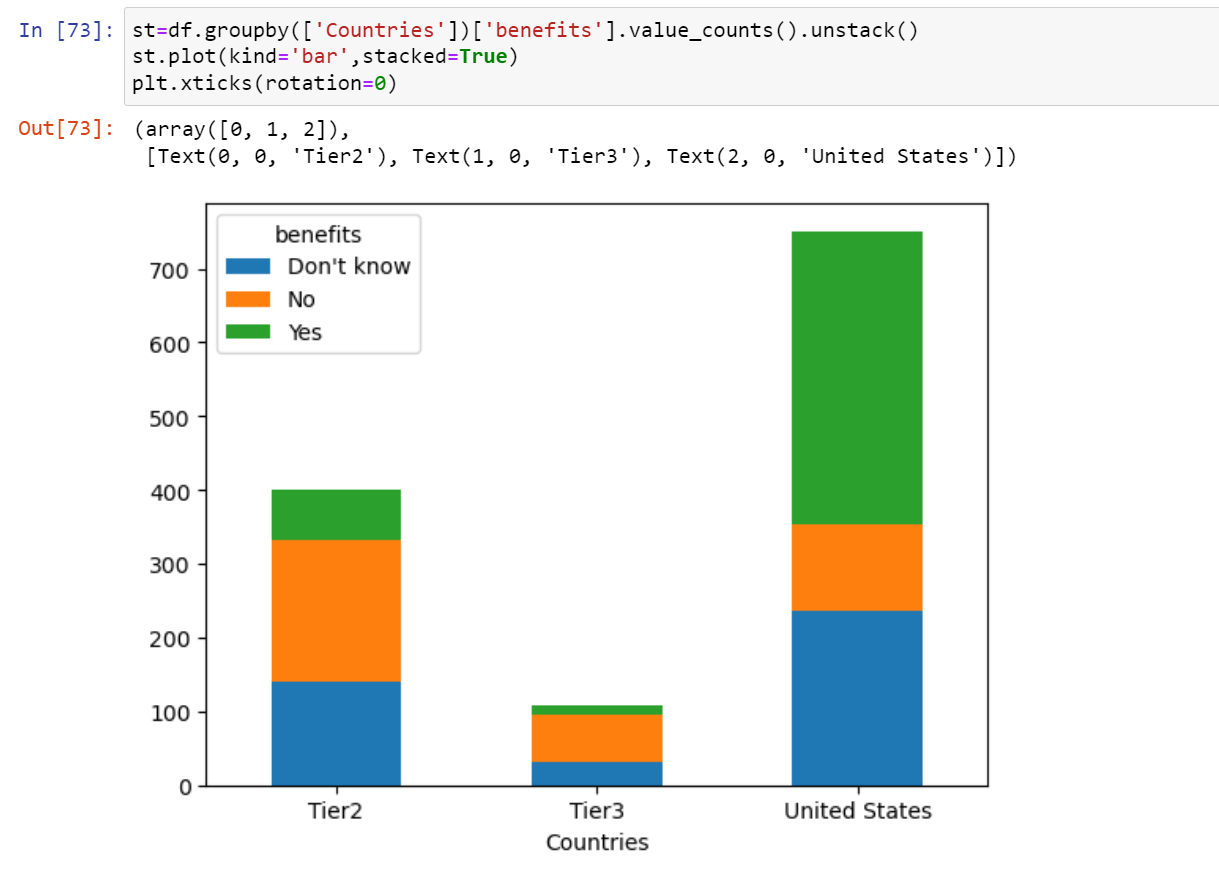
****

Fig:31

In the above visualization, we can straight away draw some sharp insights; Firstly, Compared to Tier 2 and Tier 3 countries, where it appears that only around 10-20% working population believe they would be getting benefits, In United states its more than 50% of the respondents. Another fascinating aspect of the visualization is that respondents of Tier 2 and Tier 3 would probably not be getting benefits as most of them are clear that there are no benefits. Whereas on the other hand in United states the proportion of people wouldn’t be getting benefits is lesser compared to the people unaware about which is also an advantageous point as part of the research.

**Insights on Care Options**

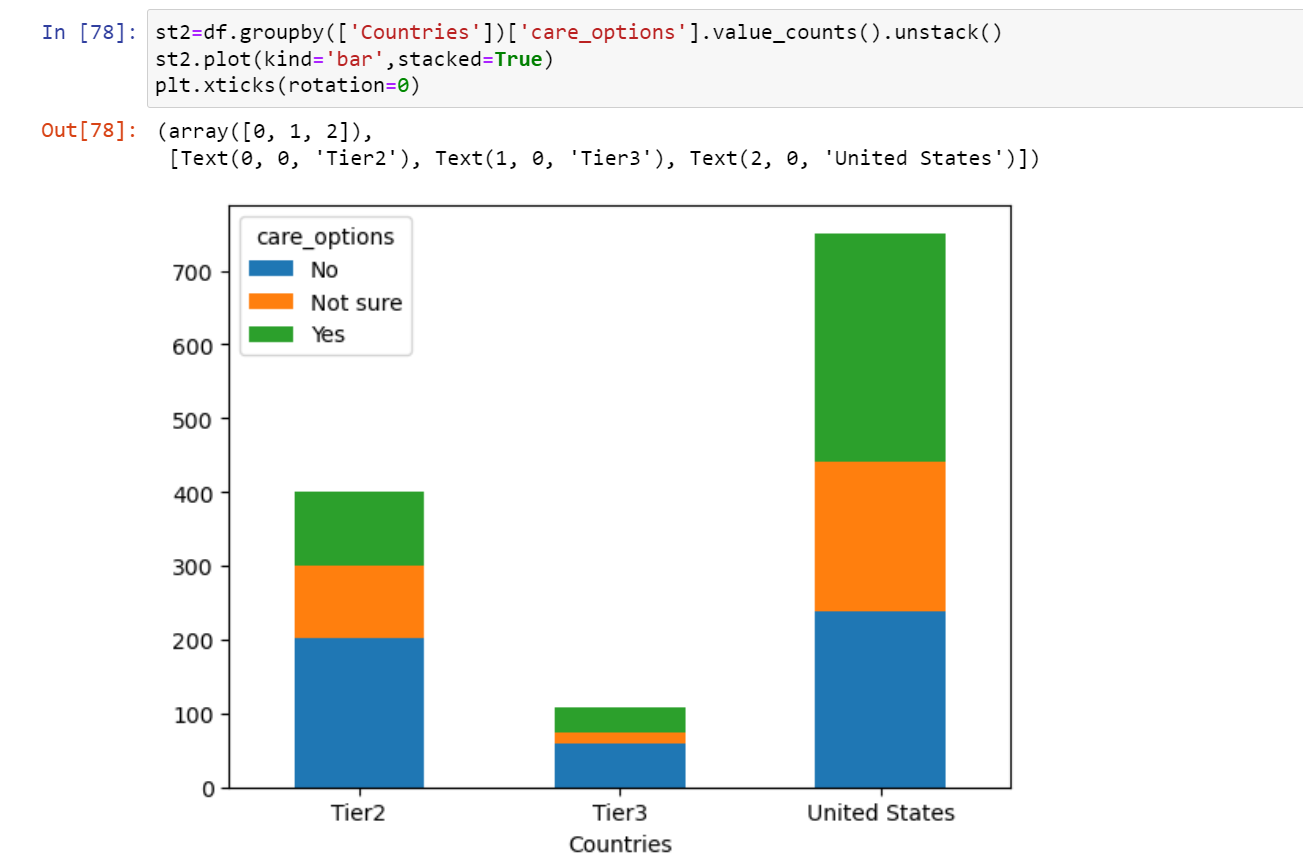
****

Fig:32

Second Visualization, Care options, it talks about whether an employer of a particular country would be provided with support in the form of care if the person suffered with mental health concern.

Stacked bar chart in Fig:32 hints us once again at similar conclusion as that of in benefits but in a different way. Although the Tier2 countries and United States look to have the same proportion of respondents for “Not sure” as an opinion, it’s surprisingly the Tier3 countries where in people are much sure that no care options would be given to them in case of any mental health concern.

**Insights on Wellness Program**

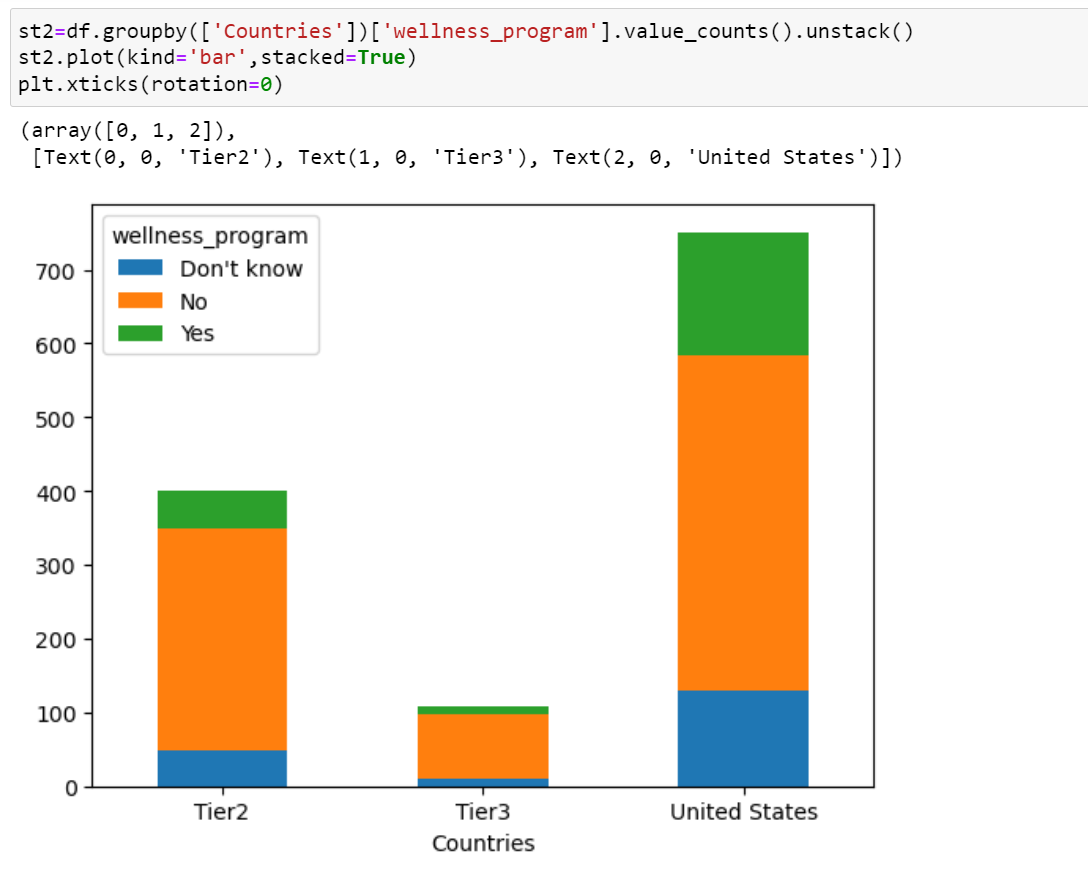
****

Figure:33

When it comes to Wellness program, we see a totally different picture be it Tier2, Tier3 or United states. It’s clear that people across the world have a common issue or at least have a perception that employers wouldn’t be discussing mental health as part of wellness programs

**Combining effects of benefits and wellness program to understand mental health support across different countries.**

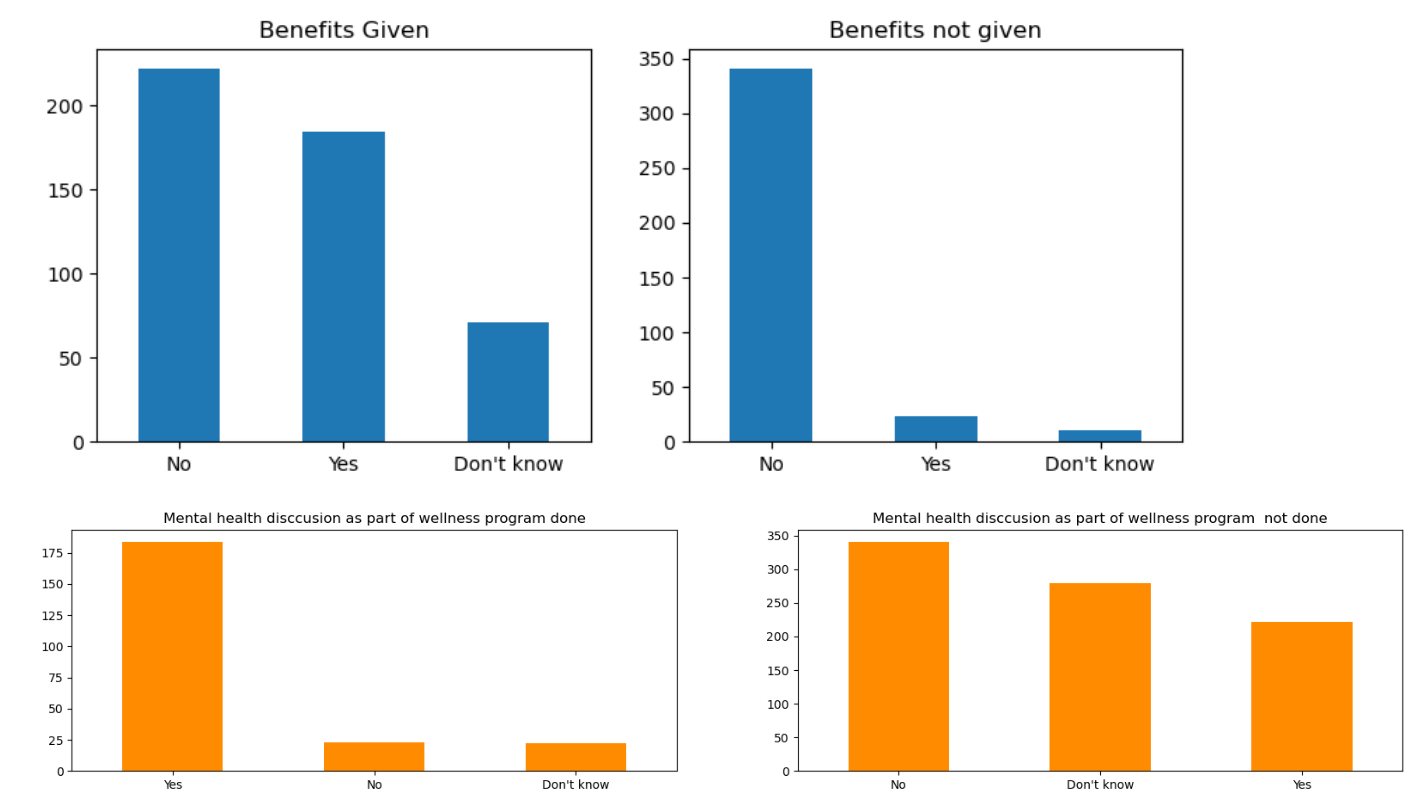


Figure: 34

Here, Fig 34 indicates four different bar plots showcasing an angle to the mental health issue; First two graphs talk about how if Benefits given or not given, changes the dynamics of wellness program options provided by the employer.

One of the most interesting aspects of this visualization is, it talks about how mental health is perceived from both the aspects and whether these two attributes have a combined effect on deciding whether a person will suffer mental health concern.

And from the above two graphs what we understand is that either benefits are been given or are not given by an employer in any of the country It has no effect on an opinion given by the respondents on the wellness programs. In other words, irrespective of benefits given or not mental health getting included in wellness programs remains to be point of concern for the people across countries.

On the other hand, if we look at the below two graphs in Fig 34, Even though mental health discussion becomes part of wellness program the benefits remain at a particular proportion but it’s the mental health discussion which when not included by the employer in the wellness program also has an impact on the benefits given to a person in a company.

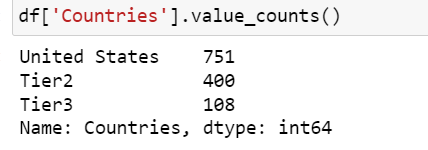
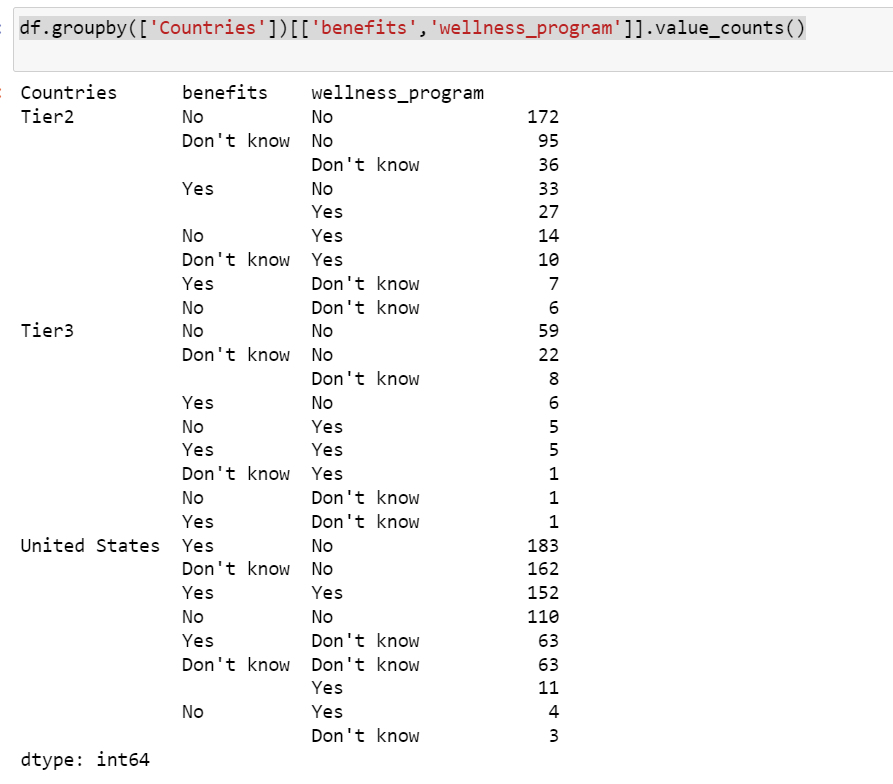


Fig:35

To talk about the countries when we compare United states to tier two and tier 3 countries, what we can deduce is that; Although United States has a greater proportion of people believing that support relating to mental health will be given be it wellness programs or benefits, it’s the tier two and tier three countries where people are more clearer regarding both the aspects being an issue for mental health. Whereas for U.S respondents, it looks like the lack of clarity through the employers to employees regarding mental health support is an issue, which still is a small issue compared to the one for tier2 and tier3 countries.

**Modelling as per the research question:**

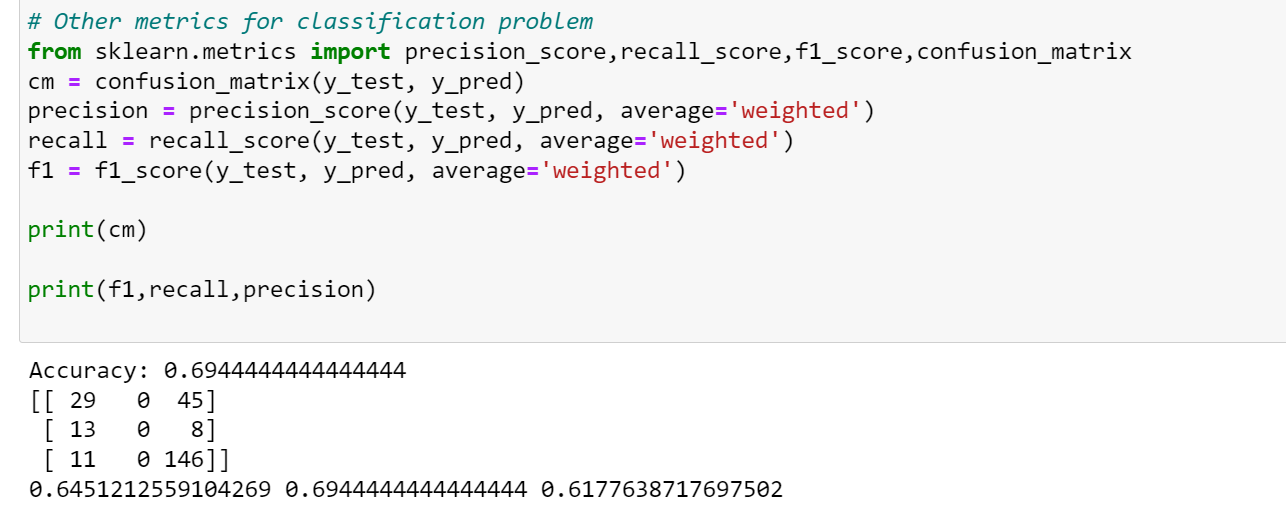
Now for this, I have selected three input variables benefits, care\_options and wellness\_programs and an output variable countries. Using decision tree algorithm, the following results were obtained ****

Fig 35 (a)

Later cross validation was performed to make sure the results obtained had a sense to it and was not just because random selection in the figure Fig: 35(b)

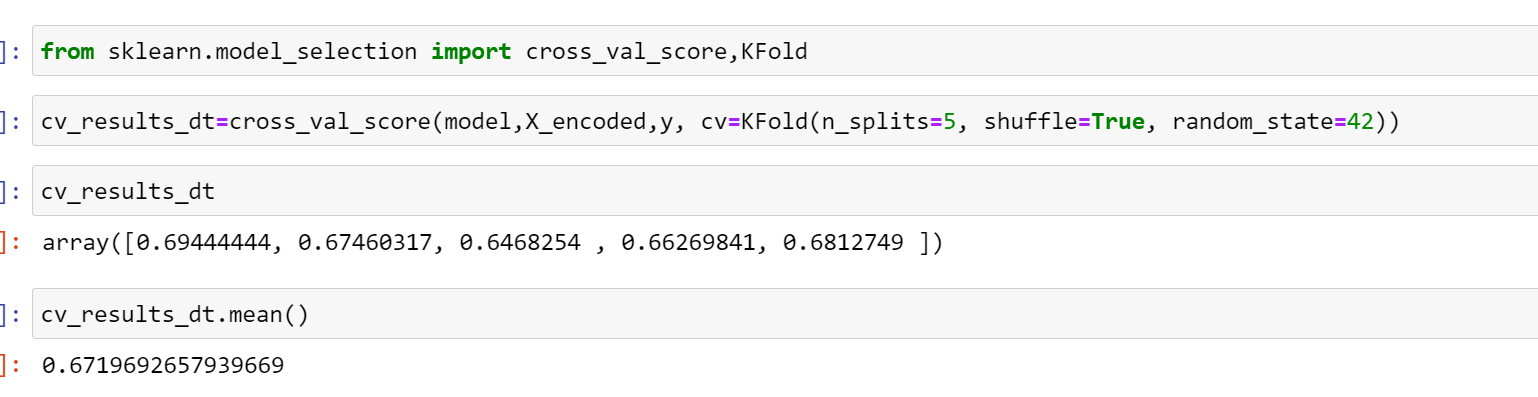


Fig: 35(b)

**Second Research question**

2) Do People coming from a certain background, age group or sex tend to be open to their supervisor regarding their mental health?

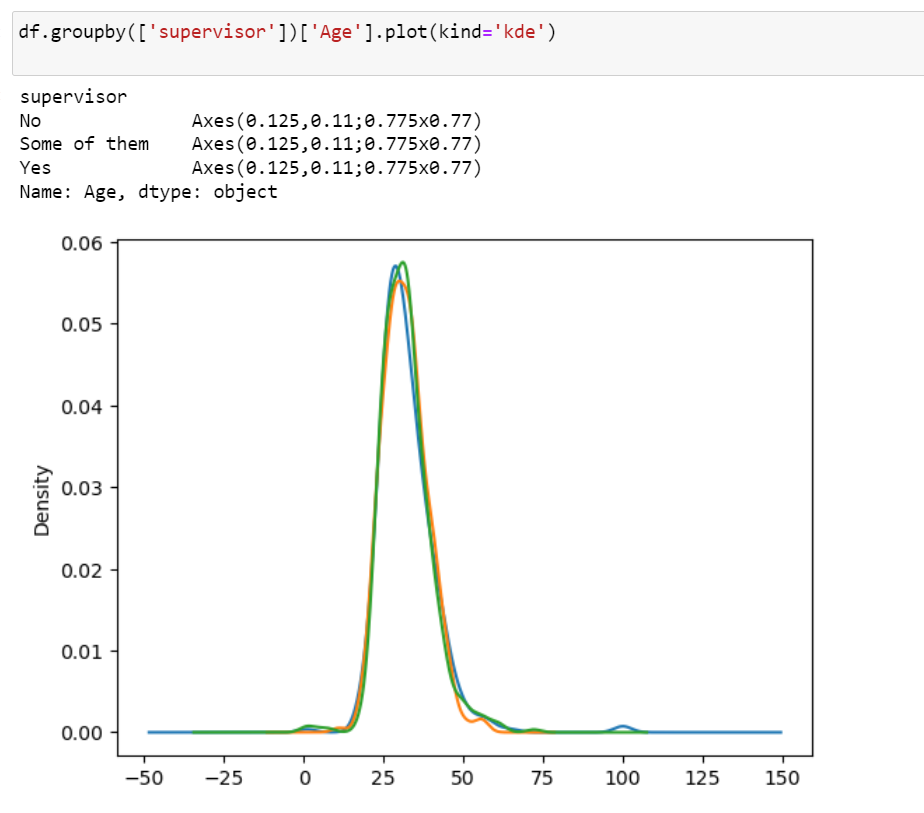


Fig: 36

So, to understand how age varies for responses regarding supervisor, a distribution plot is done in figure 36.

And it’s very interesting to see that the distributions for all the three responses of “Yes”,” No” and “Some of them” are on overlapping on each other stating a fact that age has never been a differentiator between whether a person discusses about mental health concern with his supervisor or not.

**Insights through Countries**

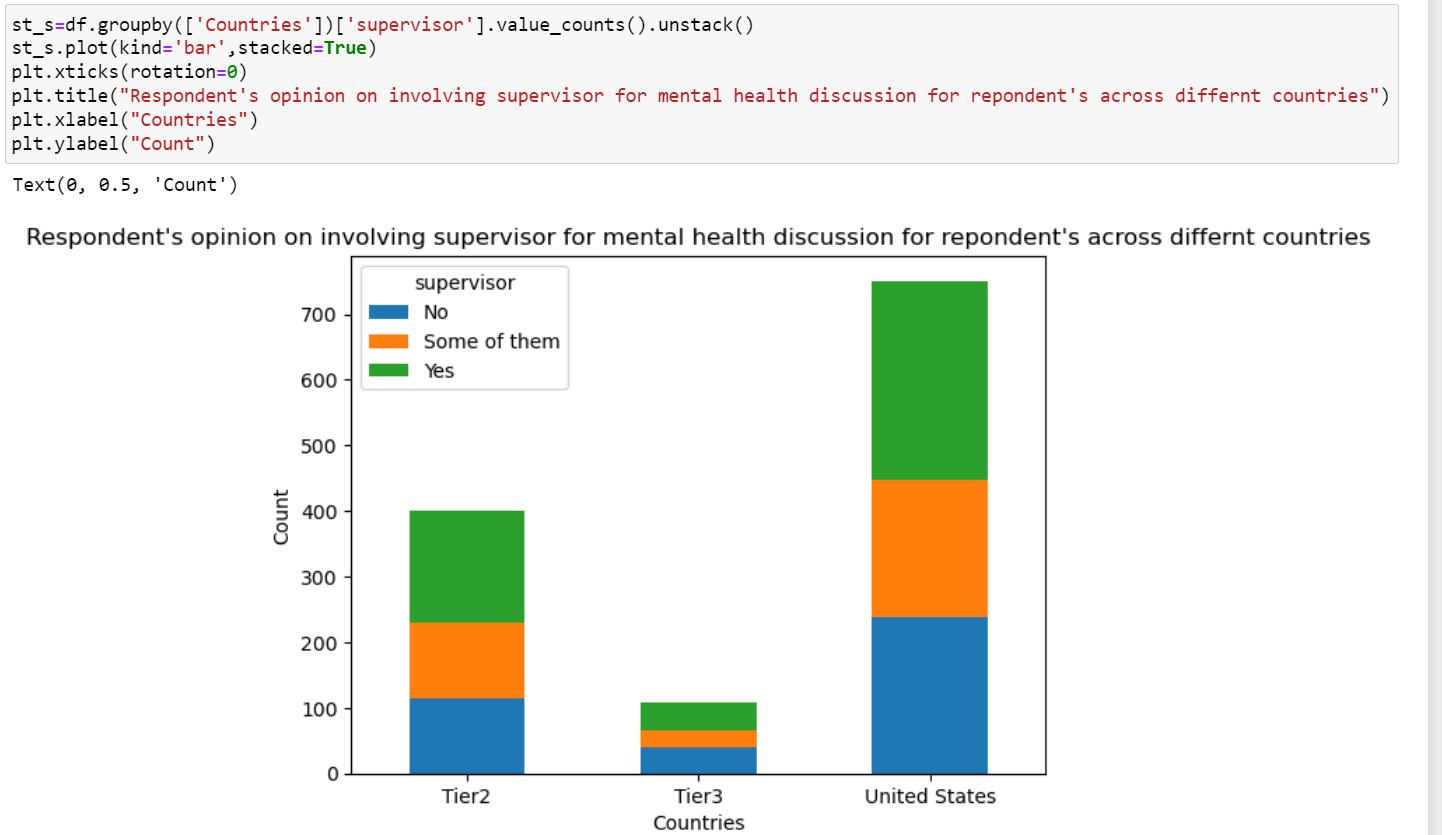
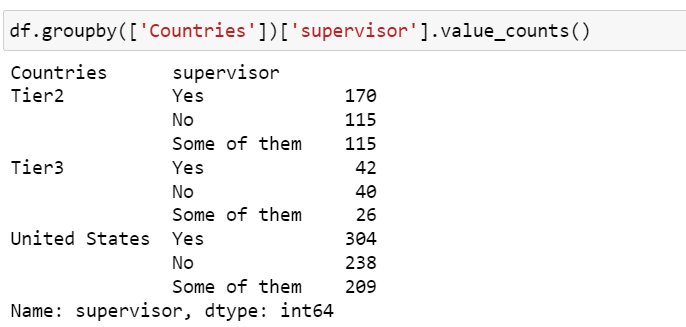


Fig:37

The above visualization, Figure 37 describes about how a response towards supervisor of a company changes from country to country regarding mental health and we can clearly see that the role of a tier2, tier3 countries or even U.S is not much when it comes to openness of a respondent to the supervisor.

Even through the code in Fig:38 we can see and understand that the distribution across different countries has been uniform

Fig :38

**Role of Gender**

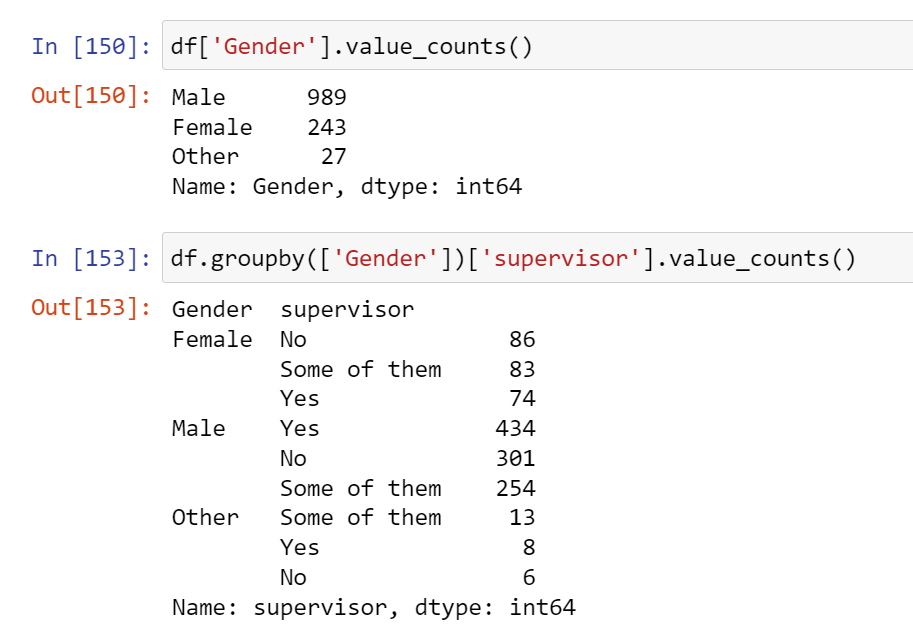
****

Fig:39

**Modelling using Random Forest for the 2nd research question:**

For this research question I chose random forest as the algorithm to predict whether a respondent of the survey would choose to talk to supervisor on mental health depending upon his age, sex or background.

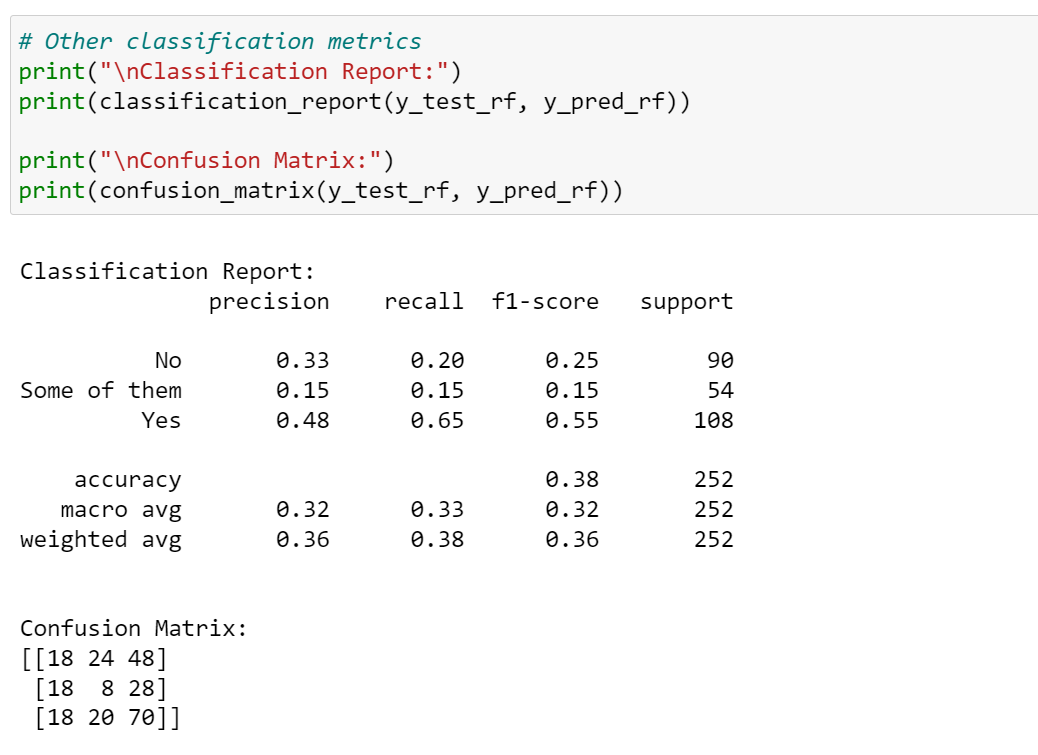
****

Fig: 35(c)

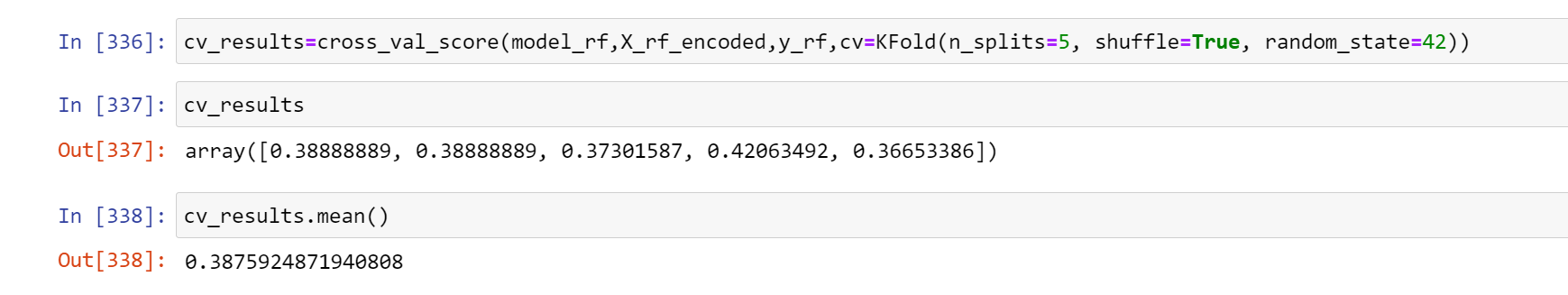


Fig:35(d)

It looks like only 38% of the times prediction of whether a person will be open to supervisor or not can be done using these three variables Although the visualization results suggest a strong relation between gender and supervisor attribute as we can see from Figure 39 that there are about 243 females out of 1259 respondents which makes it to about 19.3% of the respondents. Now as per the stats it’s clear that when it comes to Males, they tend to feel more open towards discussing mental health with any of their supervisor compared to women.

About 44% of males have a positive response towards supervisors, whereas only 31% of women feel they could talk to their supervisor’s. Moreover 35% of females feel they can’t talk about mental health to any of the supervisor, while only 30% of male respondents felt as such, which kind of reinstated the fact that across background and age females didn’t feel safe to talk about mental health to their supervisors

**Third Research Question**

3)What’s the general opinion difference between people coming from tech and non tech regarding mental health or in other words is Tech industry across gender, age and background creating an environment not suitable for better mental health?

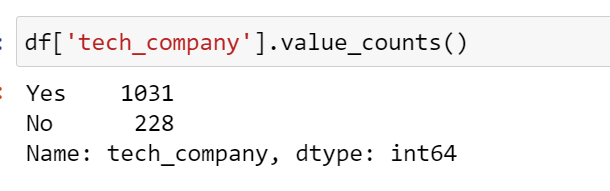


Fig:40

To begin with, this code tells us the number of tech and non -tech respondents in the survey

**Insights with Observed\_consequence**

This attribute talks about an observed consequence of mental health at the work place by a respondent, it’s a binary variable with either “Yes” or “No” response. So, to differentiate between tech and non tech groups here and to understand the reasons behind mental health issues for each of the group, we look to begin with this attribute.

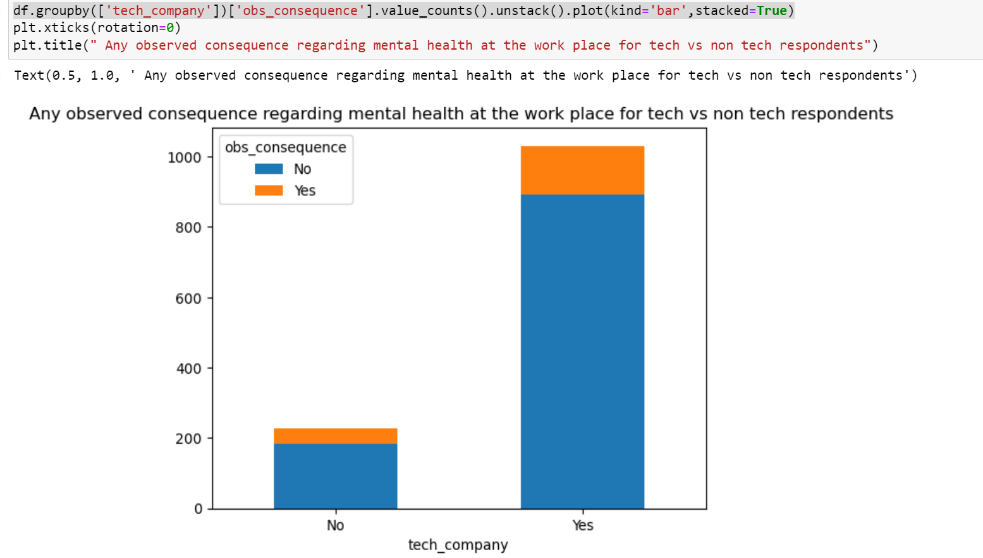


Fig:41

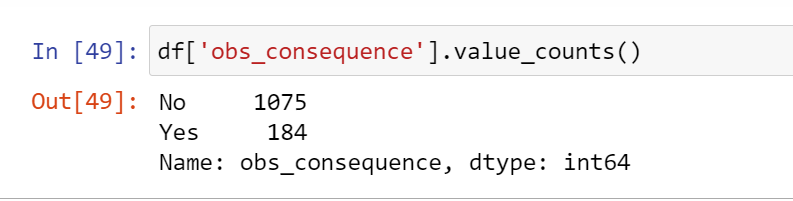


Fig:41 a)

So, the results that we get from the bar graph and Fig 41a) here is that there is proper distribution of “Yes” and “No” across both tech and non tech industry as per its value counts and hence “observed consequence” variable wouldn’t be involved in determining mental health concern in any capacity.

**Family History**

This attribute is also a binary variable either “Yes” or “No” to whether mental health issue is a concern because of family history. In other words, if a person responded “yes” it meant there was a mental health issue in his family and “No” if it was not.

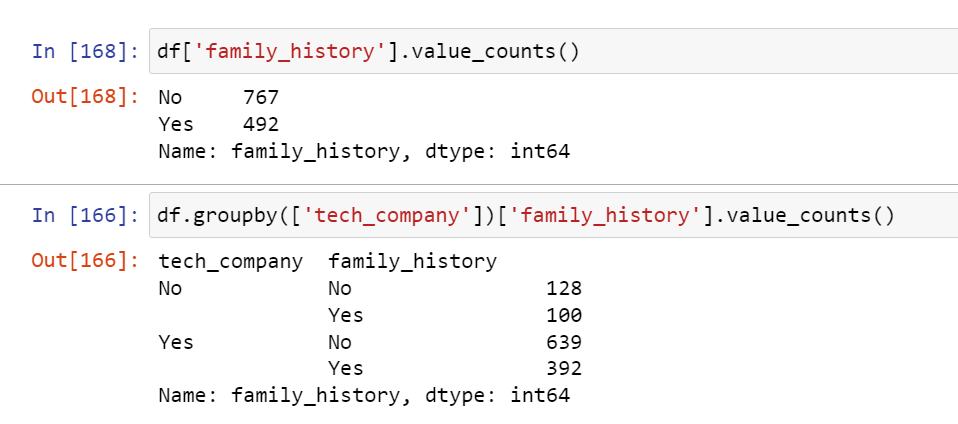


Fig: 42

Now the code in the figure gives the value counts for each of its value “Yes” and “No”. Then the next piece of code below it talks about the combination of family history value counts with tech company value counts. In other words, there are 128 respondents with no family history for mental health and also do not belong a tech company. Similarly, there are 392 respondents having a family history of mental health issue who work at a tech\_company.

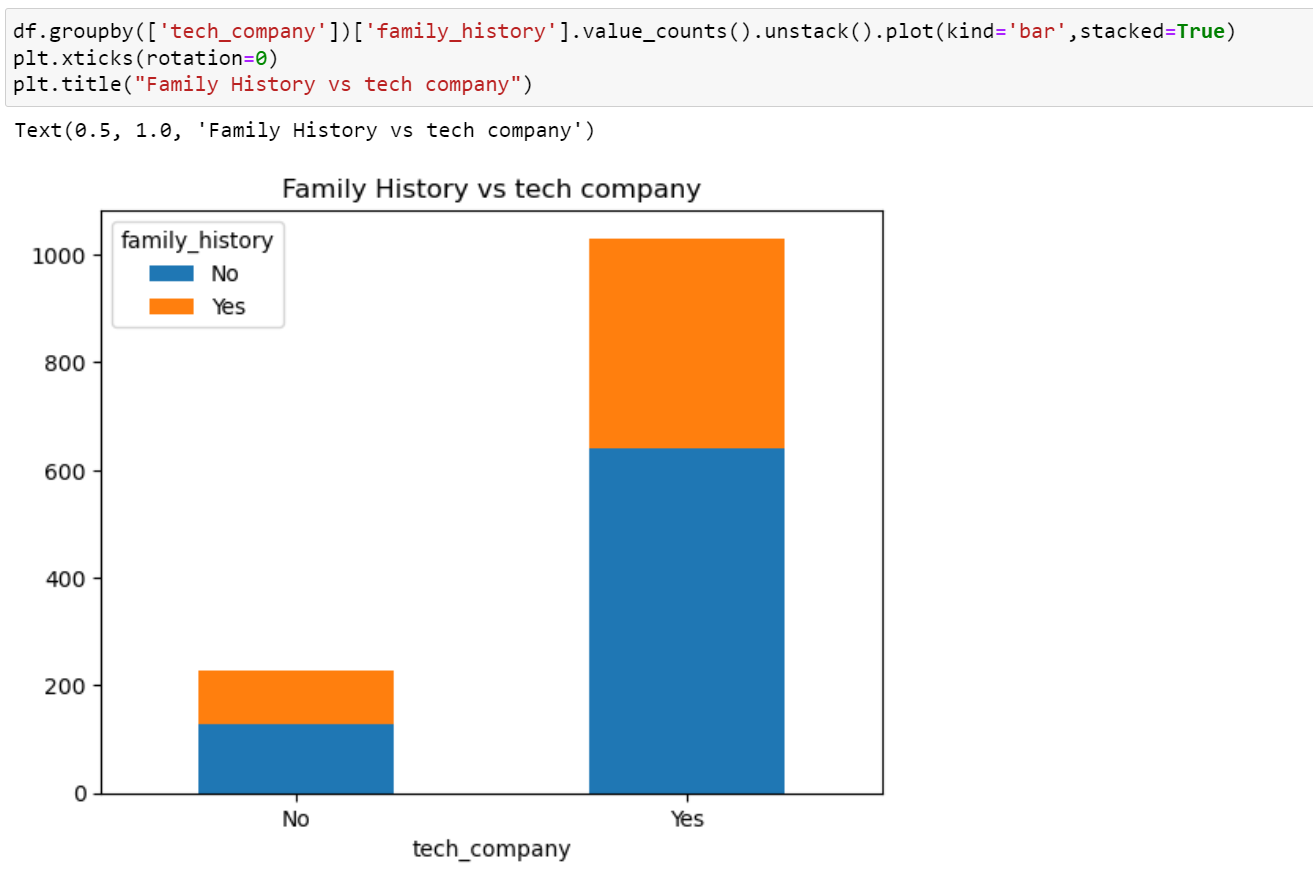
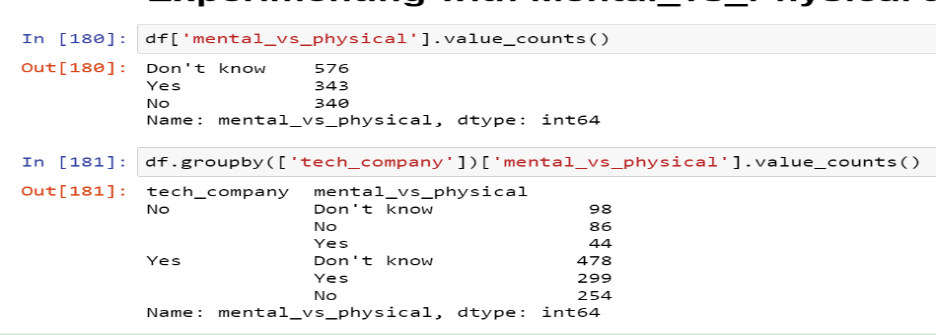


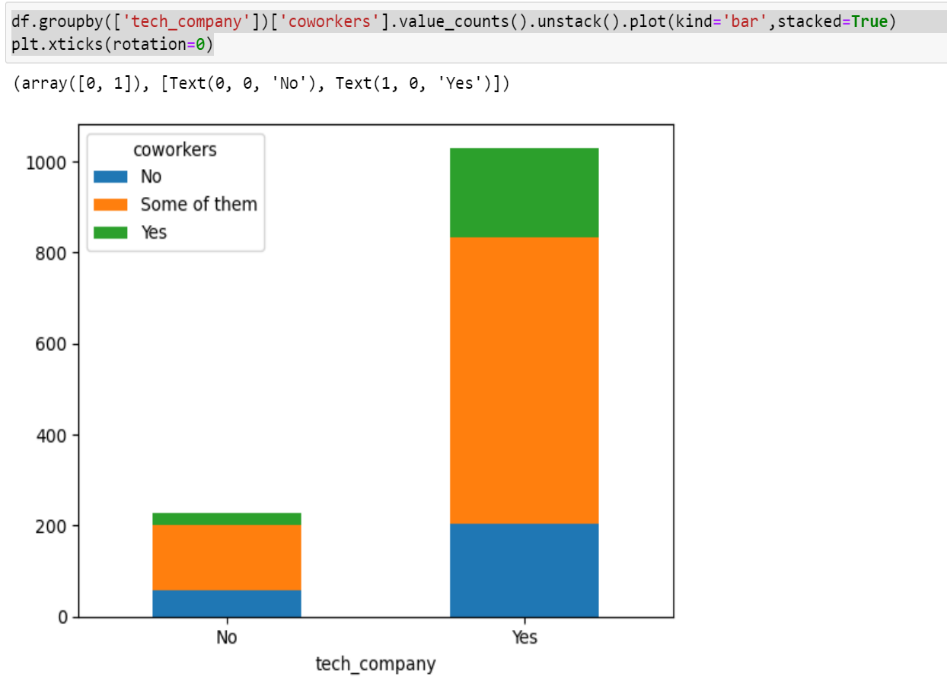
Fig :43

Now after understanding the data from figure 42 and looking at the bar chart in figure 43, we can deduce that the people not working at a tech company are more likely to have a family history of mental health issue, for non tech respondents it’s about 43% chance and for a tech respondent to be 38%.

**Mental\_vs\_physical**

**** Fig :44

Now from the stats for the mental\_vs\_physical attribute, which talks about whether an employer takes mental health as serious as physical health. It is clear that people working at non-tech have a different opinion when it comes to whether an employer will take mental health as seriously as physical health. As far as the figures are concerned, it is Fig:45 which gives us a better understanding.

 Fig:45

**Modelling a Decision tree to understand whether a person’s type of workplace could be predicted**

Results obtained using four variables as input; ‘obs\_consequence', 'mental\_vs\_physical', 'coworkers', 'family\_history' and the target variable being ‘Tech\_company’ whose response is either “yes” or “No” can be seen in figure 45(a)

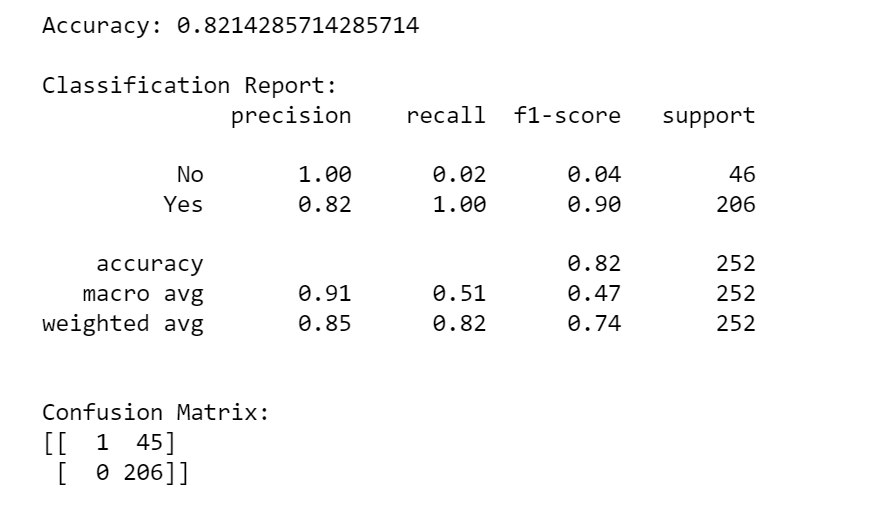


Fig:45(a)

**Insights through visualizations**

So, the insight that we get through this opinion is in comparison to tech companies non tech companies employers tend to give physical health more importance than mental health, which is the opposite for tech companies.

**Conclusions**

After analysing three variables (Benefits, Care options and wellness programmes) to understand, how does support regarding mental health vary for each country, we learn that almost half of the working population of U.S believe they would get benefits for their mental health concern. Whereas only 1/5th of the working population in other countries believe so. On similar lines, Care options also does follow a similar trend but the only difference is that the proportion of not sure people across the countries is high, which indicates unawareness among people on treatment methodologies for mental health. But when we discuss about mental health included in wellness programs and benefits opinions as a single entity and look to see their combined effect it’s quite evident that benefits look to be not part of wellness programs or in other words the former has no effect on the latter and Wellness programs remain to be point of concern for majority of the countries.

Moving on to other findings about the survey, sharing mental distress and concerns in regards to it with the supervisor in an organization appears to be irrespective of the respondent’s residing country or even age for that matter. Whereas, gender of a person looked to be an attribute of a respondent for it. 13% more males than females are open to their supervisor if they go through any mental health issues.

Finally, in determining whether a respondent belonged to tech or non tech company basing on his responses on few opinion questions like; observation. So, when an opinion was taken through the survey on whether an incident of mental health was observed at the work place or not it turned out that most of the people didn’t watch any of such incidents and the distribution of the people was also uniform across industries. Family history was another aspect through which different companies were to be distinguished in this paper and the results got from it were astonishing. Although there is popular belief that people from tech companies look to have high chance of mental health concerns, its observed that people working in non tech were more likely to have a family history than the people working in tech. Whereas when it comes to having a balance between physical health and mental health by the employer, we learn that non -tech companies are better at this than the tech companies