D.Chakravardhan Reddy

IMT2014015

Data Understanding:

Collecting Initial Data:

The dataset has been collected from a survey done by Open Source Mental illness corporation. This is an existing data provided by the people in the form of the survey. We collected two data sets one is the survey done in 2014 and other is in 2016. Both of the data sets have almost similar attributes. We, are planning to integrate the two data sets so it could be helpful for the further analytics. Some of the promising attributes from the database are age, gender, country, no_of_employees, family_history, work_interfere, care_options treatment, mental_illness. Attributes like year, comments, self-employed seems irrelevant and can be excluded for the analytics. This existing data is enough for our analysis and there is no need of additional data.

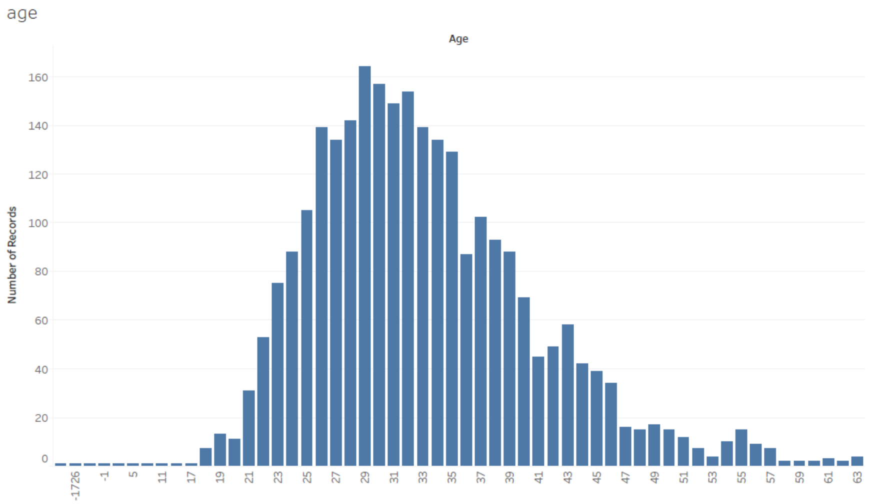
Data Description:

The data set is in the CSV format which can be used readily. The data set consists of 2692 rows and 28 attributes.

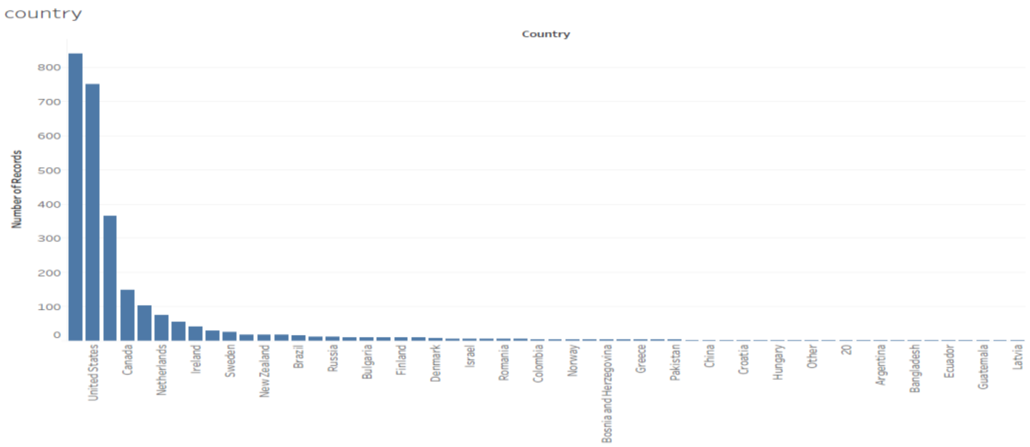
Most of the data types in the data set are symbolic, whether they are dates and times, or answers to multiple-choice questions from the Survey questionnaire. Some of these variables can be used to create new variables such as year (after merging), Age_group into categorical. Most the attributes are Boolean which takes values either yes/no or 1/0.

We were able to prioritize the attributes by determining the covariances between the attribute and the target attribute. The attributes with high variability are selected for the further analysis.

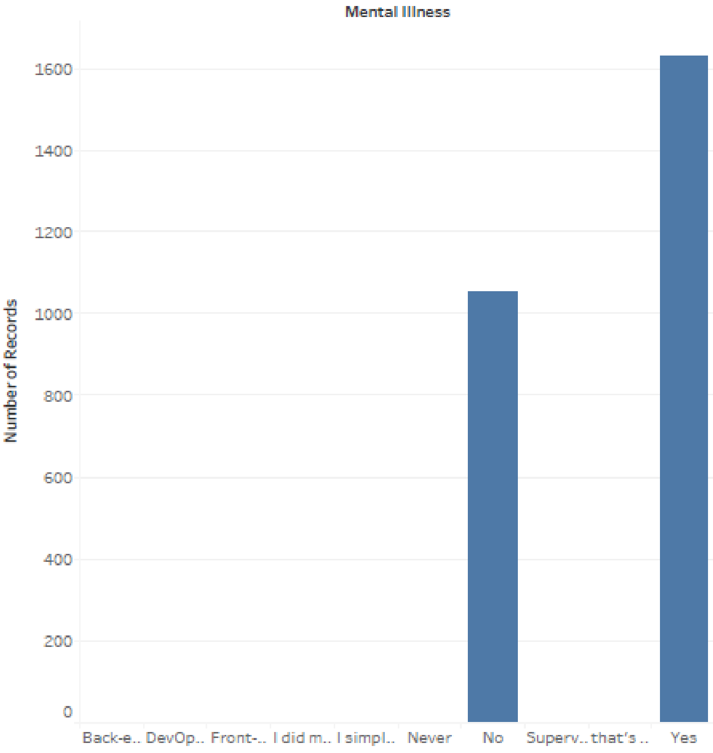
Data exploration

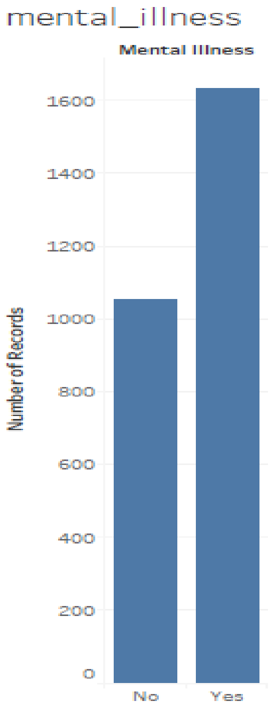


From the above graph, we can see that most of the people's age is between 20 and 50. So, categorizing age variable would be helpful in performing age wise analysis. For example, in particular age group how many people have mental illness. This type of analysis will become helpful to the people in that respective age group. And also, we see some outliers here. We can filter the age between 18 and 65. So, our analysis will be primarily helpful to the people between 18 and 65.

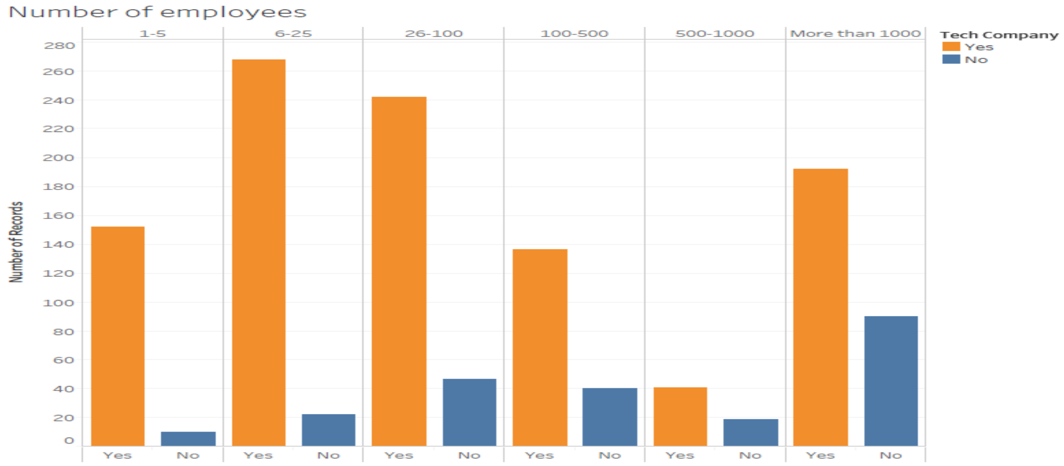


From the above graph, we can see that most of the people are from united states, United Kingdom and Canada. So, our analysis will be meaningful to the people living in these countries. Since most of the records are from united states we can say that our predictions will be accurate to the people living in united states.

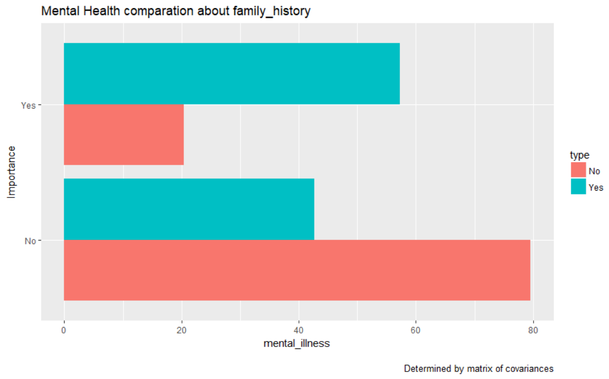


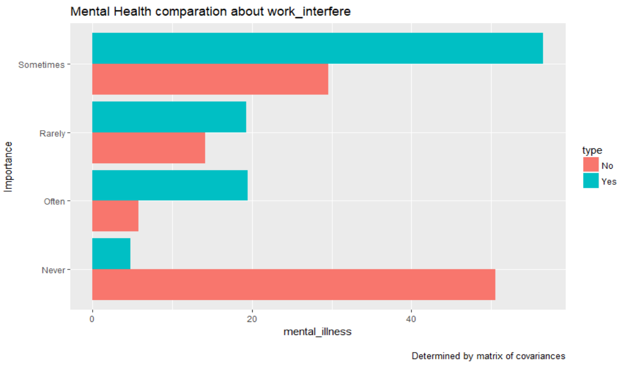


Looks like there is no class imbalance in the data. The number of people with mental illness (positive) and without mental illness(negative) are in good number to build the model and predict the mental illness.



The above graph shows number of employees who work in tech/non-tech companies. In our data set, we can see that the number of employees' range is more in 6-25,26-100 and more than 1000. And always the tech employees are more than that of non-tech. So, our analysis will be primarily useful to small or medium scale tech companies.





The above graphs indicate the covariances between the attribute mental_illness and work_interfere which helps in deciding whether the attribute is important for analysis or not. For example, we can see that people who said work_intefere never have very less mental_illness and very high mental stability, which is indicating the importance of the attribute.

From the above graph, we can see that people with family history are more likely to have mental illness compared to those who don't have family history. So, family history seems like a promising attribute in our analysis.

Data Quality

Since this is a survey, there will be unanswered questions by many people. Hence missing fields are very common in the data set. The number of NA's in the following attributes are: state – 515, comments – 1095, work_interfere – 264, self-employed – 18. We can see that the attributes state, work_interfere, comments, self-employed have null values. Since the percentage of null values for comments are more we can simply removes the attributes as it is not so relevant for our analysis. For the analysis which includes the state and work_interfere, self_employed attributes we just remove the rows where values are null. Most of the attributes have Coding inconsistencies such as for examples Gender has Male, M etc. And some of the fields consists of Bad data i.e mismatch between meaning of the field and the value. For example, a Boolean attribute has non-numeric value. Also, we can see measurement errors in the no_employees and mental_illness fields.

Data Preparation

Integrate Data

We merged the two datasets by the column names. i.e if the attribute name is same for both values, append those values to form a new data set.

Selecting Data

Important attributes which will play key role in the analysis are identified in the data understanding phase by determining the matrix of covariances. Some of the important attributes are: Gender, family_history, work_interfere, benefits, care_options etc.

Cleaning Data

We prepared the data required for initial analysis by cleaning the gender and categorizing the age attribute. For gender attribute, we found how many unique values are there and then we changed accordingly. For age, we categorized into three groups which helps us to perform some age wise analysis. For attributes like tech_company, mental_vs_physical, care_options we just changed “1” to “yes” and “0” to “No”.

Format Data

We are planning to use models like random forest, KNN to predict the mental_illness. For those models, the current format of the data is sufficient and the data need not be reformatted.