**Mental Health Analysis in Technology Organizations**

**CIS 5810**

**Management Information Systems**

Project in IBM Watson – Part 2

Guided by Professor Shilpa Balan

**Shanmathi Arul Murugan**

CIN: 306594267

California State University, Los Angeles

**DATASET URL:**

<https://www.kaggle.com/osmi/mental-health-in-tech-survey/data>

**TWITTER HASHTAGS:**

#mentalhealth #mentalillness #mind #stress #anxiety #depression #mentalhealthmonday #sadness #psychology #selfhelp

**DATA DESCRIPTION:**

The dataset is a survey conducted from 2014 to 2016 among various people around the world about their mental health and how they deal with it. The dataset contains about 27 distinct columns and thousand plus responses from individuals majorly working in the technology organizations. The values in the survey aids in understanding the different attitudes towards mental health and how it is approached. The survey is conducted among individuals from various age groups and across 48 different countries ranging from Australia, China, United Kingdom, Canada, Brazil, France, India, New Zealand, etc. It first begins with questioning if the persons have a family history of mental health conditions, if so have they sought any treatment for it. The survey clarifies if the person is self-employed or working in a tech company. It also questions if their mental illness interferes with their work and in that case how hard or easy it is to work from home and to avail leave.

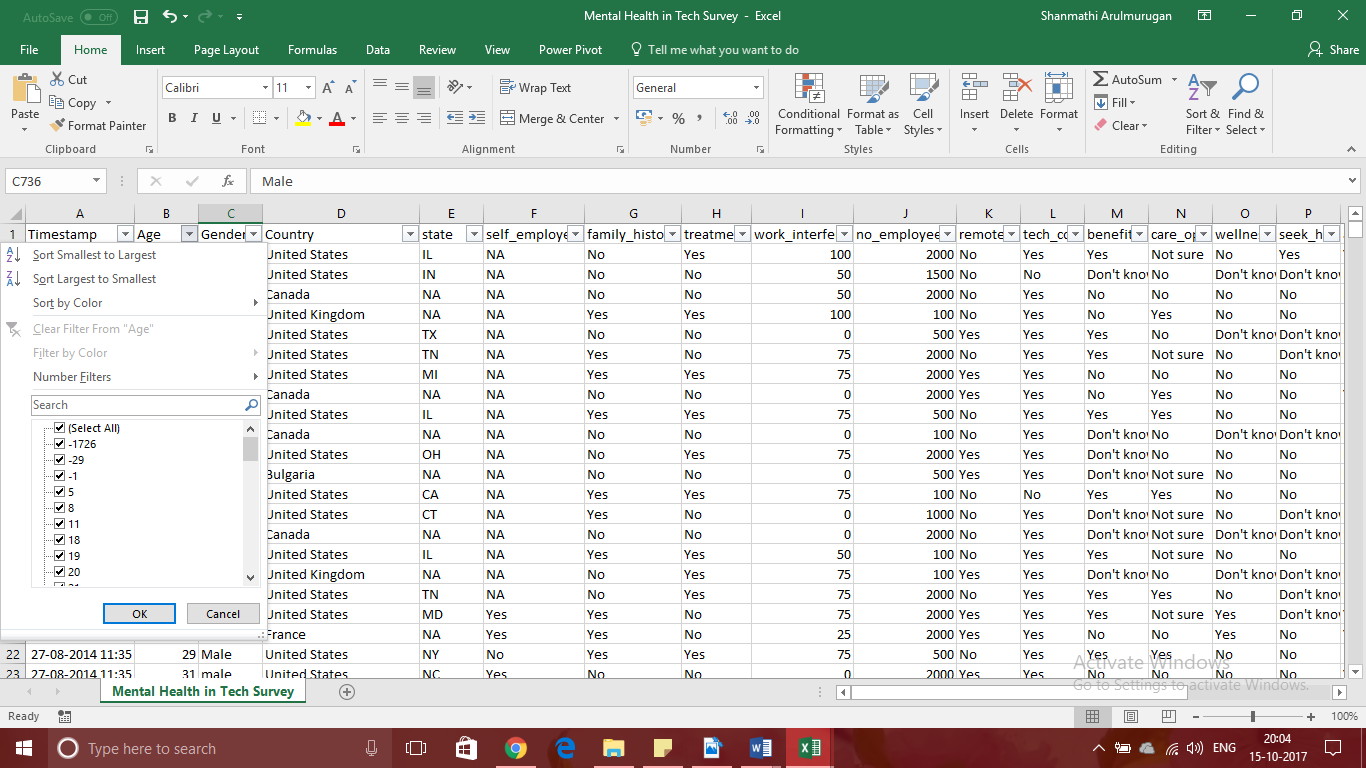
Since the main focus of the survey are the employees of technology organizations, the individuals suffering from mental illness are interviewed if their employers provide any mental health benefits, care options and wellness program. It anlyses if the employers provide benefits for the mental well being of their employees and if their anonymity is protected. It examines if the people having mental health condition reaches out to their co-workers and supervisors for help or assistance. The survey records if the people have faced or observed any pessimistic behaviour among their co-workers towards their mental health condition. With the help of IBM Watson Analytics this dataset can be analysed to understand how the individuals suffering from mental illness can be approached and treated.

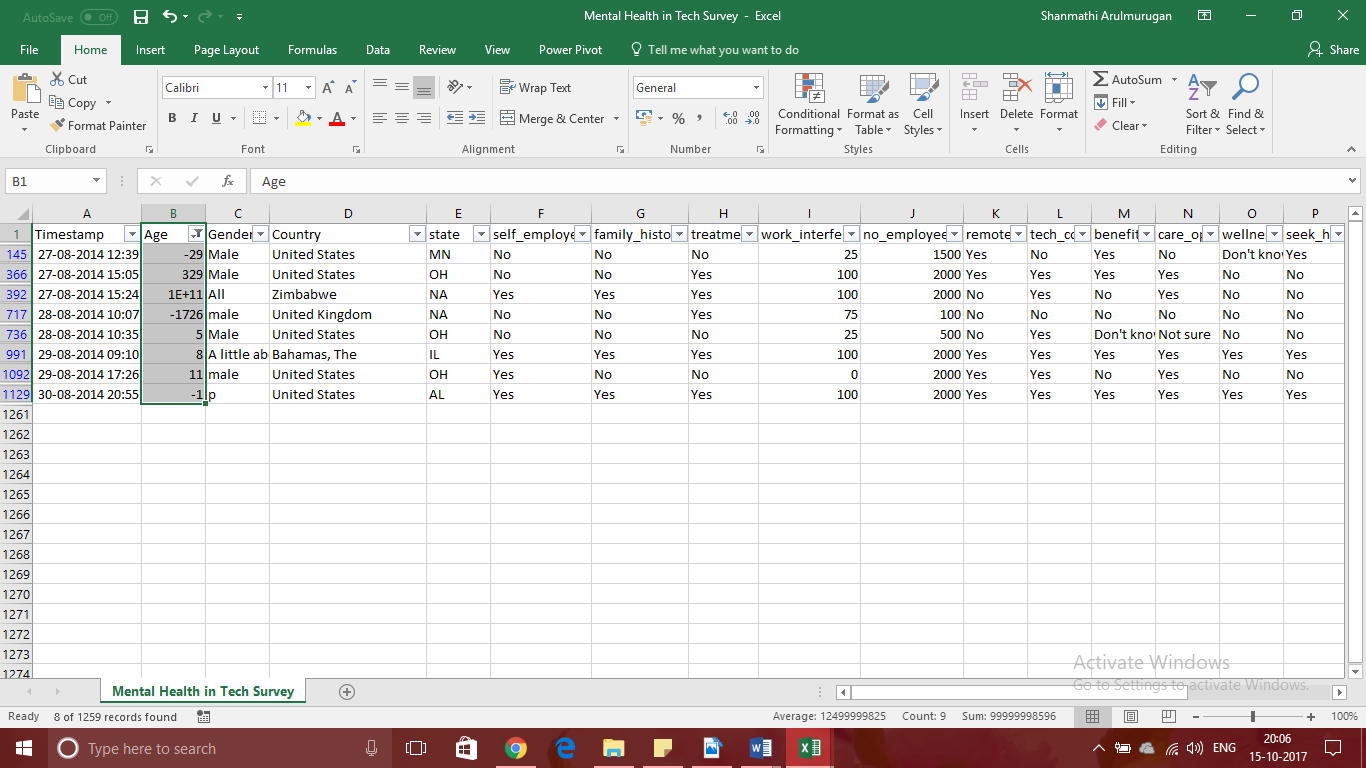
Along with this dataset, with the help of a set of ten hashtags a twitter dataset has been acquired for a period of one month in December 2016 containing forty two thousand plus values in it. This dataset contains columns such as Timestamp, Author country, Author state, hashtags used, sentiment, gender, retweet count, etc. The sentiment value will help in understanding if the tweet had a positive or negative response. Analyzing this dataset along with the previous one aid in discerning the mental health sickness better.

**DATA REFINEMENT:**

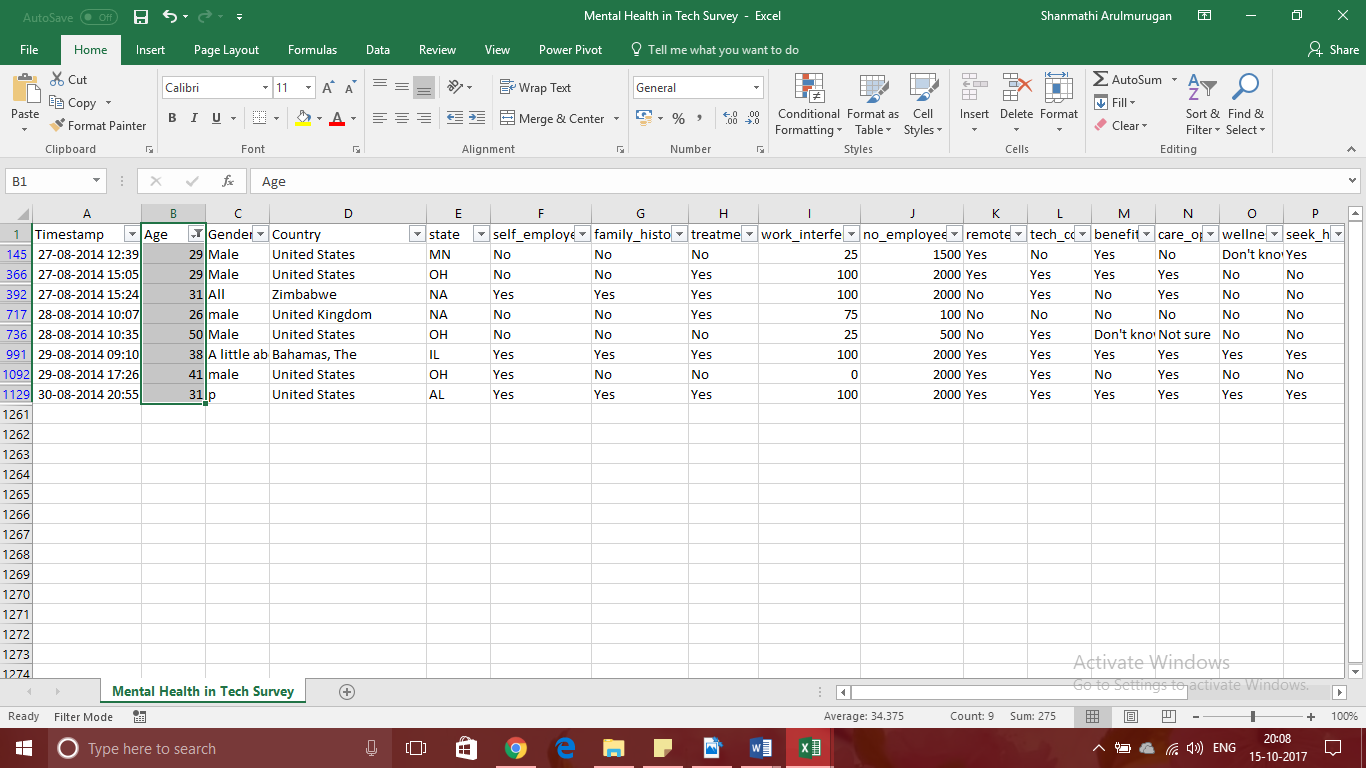
**Type 1:** Cleaning the values of Age column in MS Excel

**Pre-Refinement:**





**Post-Refinement:**

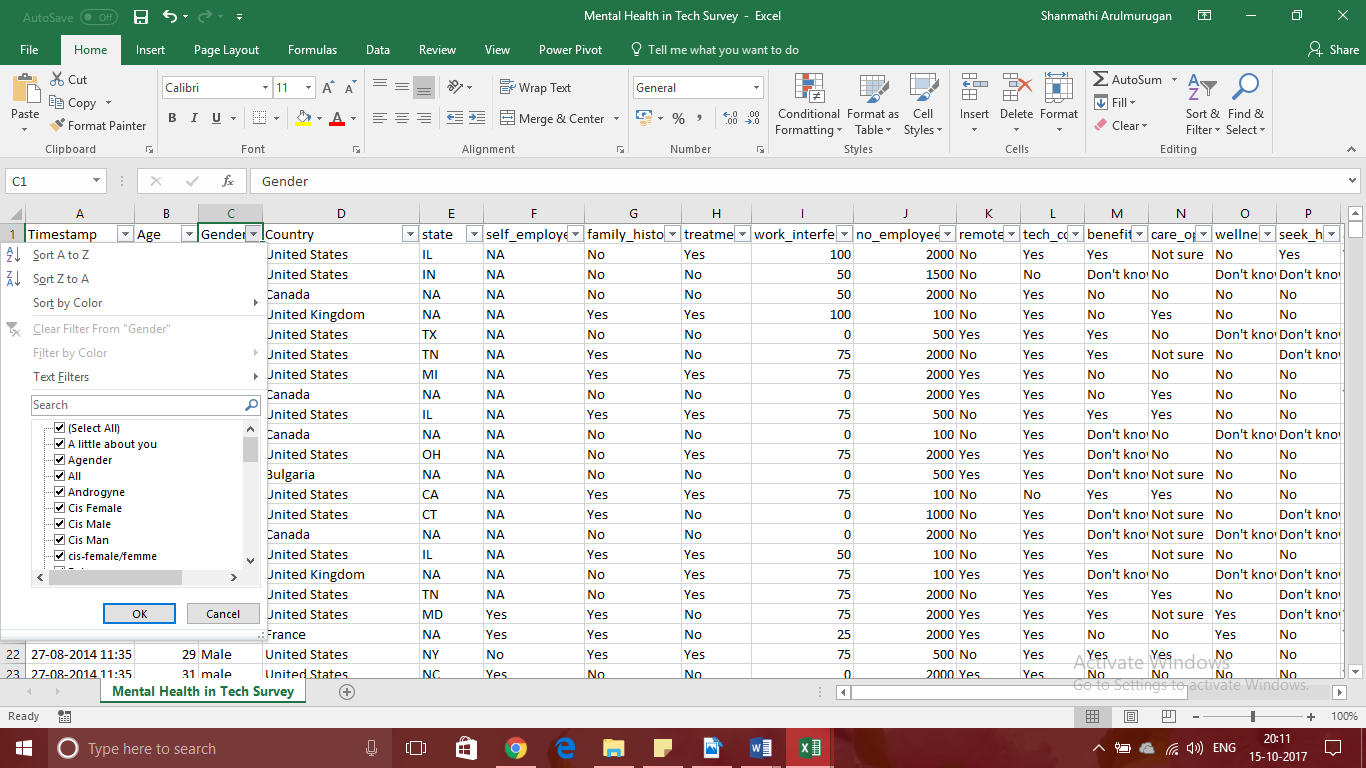


**Description:**

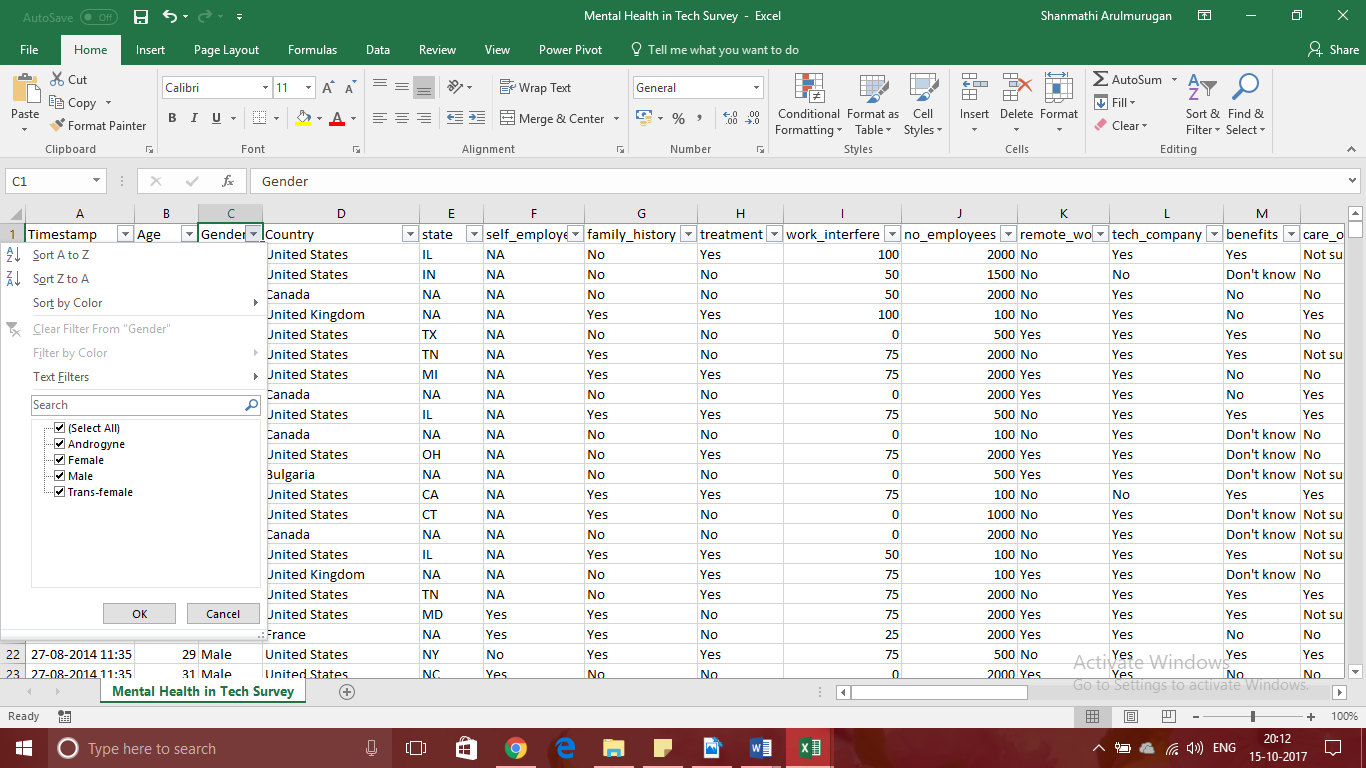
From the pre-refinement screenshot it can be identified that the dataset contains negative and obscure values for the Age column. For example, 329, 5 and -1726 cannot be the values of the age of a person employed in a company. Hence, such values are selected and corrected accordingly in the dataset.

**Type 2:** Changing the spelling mistakes and warped values in the Gender column in MS Excel

**Pre-Refinement:**



**Post-Refinement:**

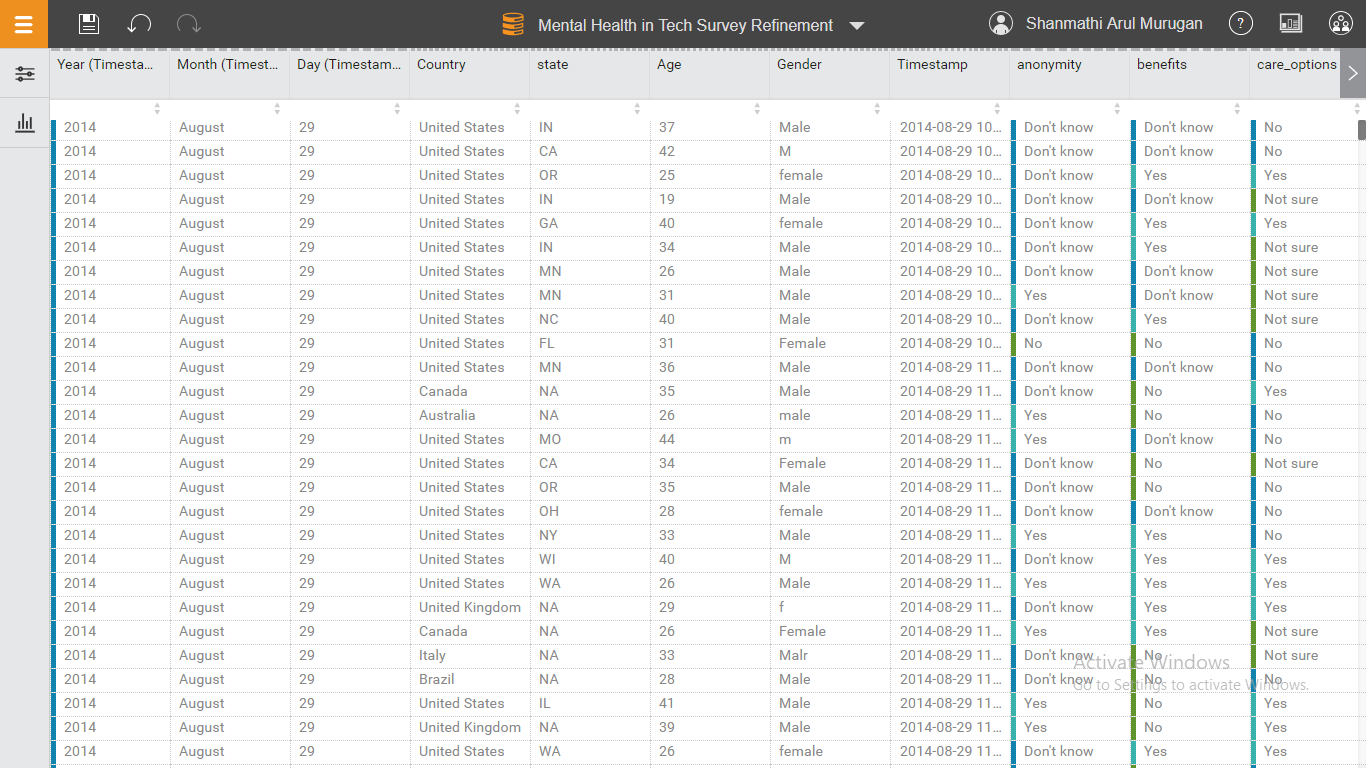


**Description:**

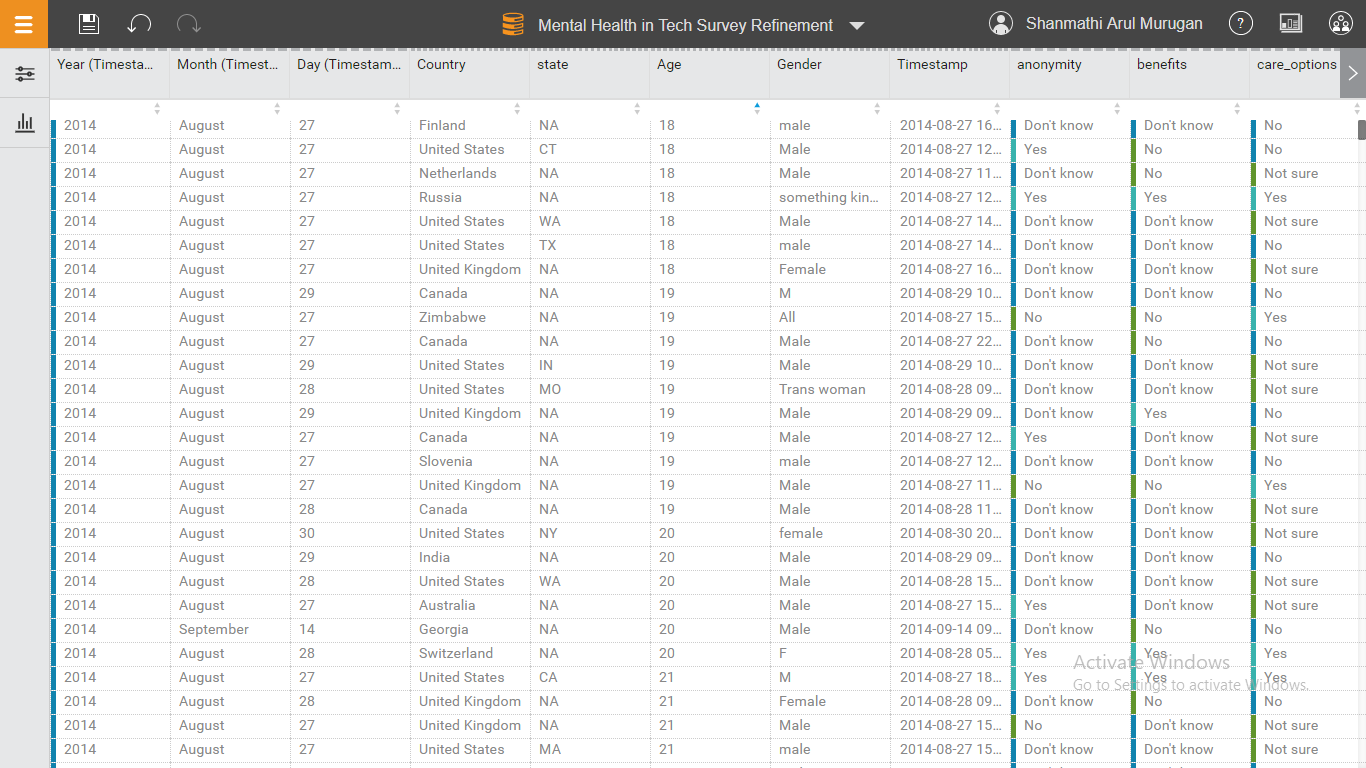
In the Gender column of the dataset there are misspelled and warped values such as ‘Cis Male’ which also refers to ‘Male’ in general. Hence, such errors are corrected as part of this refinement to improve the data quality.

**Type 3:** Sorting the Age column in ascending order in IBM Watson Analytics

**Pre-Refinement:**



**Post-Refinement:**

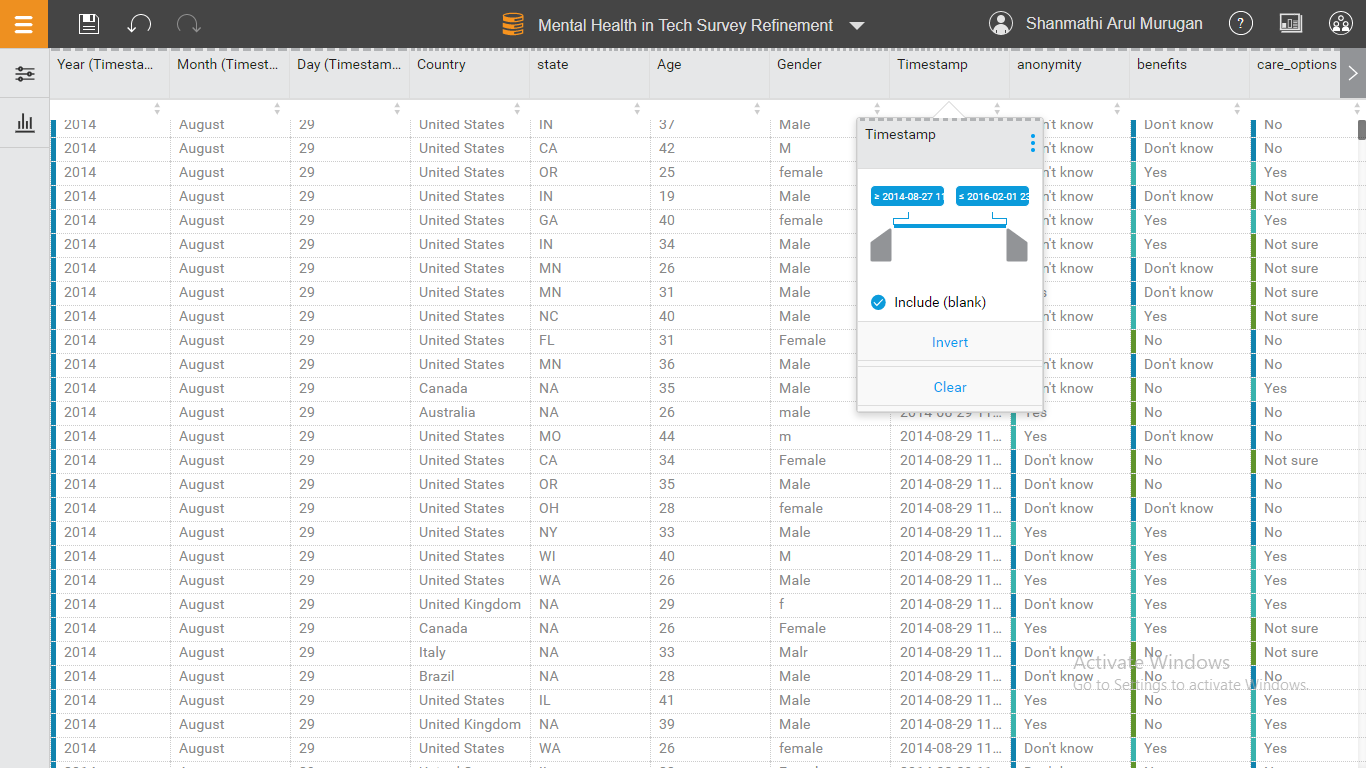


**Description:**

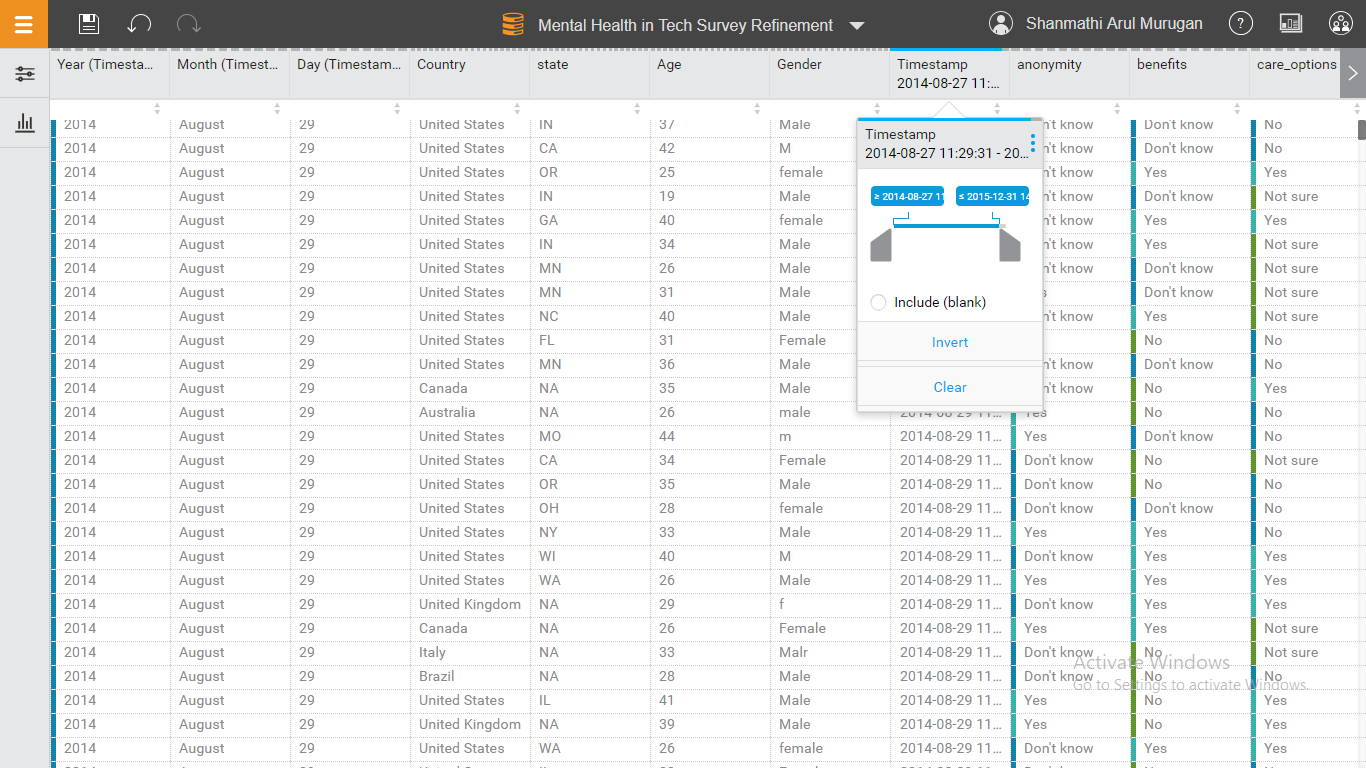
In the dataset, the Age column has values ranging from 18 to 65 and it is presented assorted. To acquire better idea on the range of values, the column is sorted in Ascending order.

**Type 4:** Selecting a range of values in the Timestamp column and uncheck Include(blank) field

**Pre-Refinement:**



**Post-Refinement:**

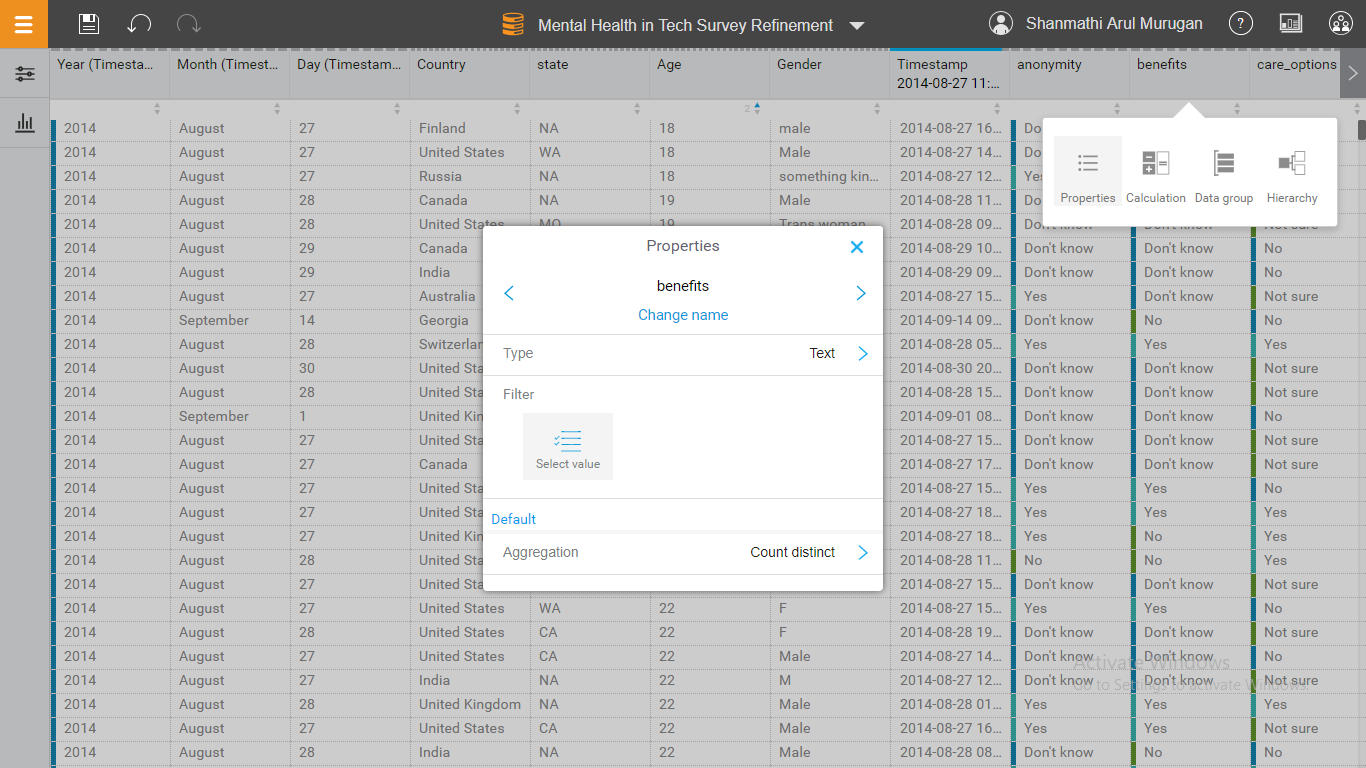


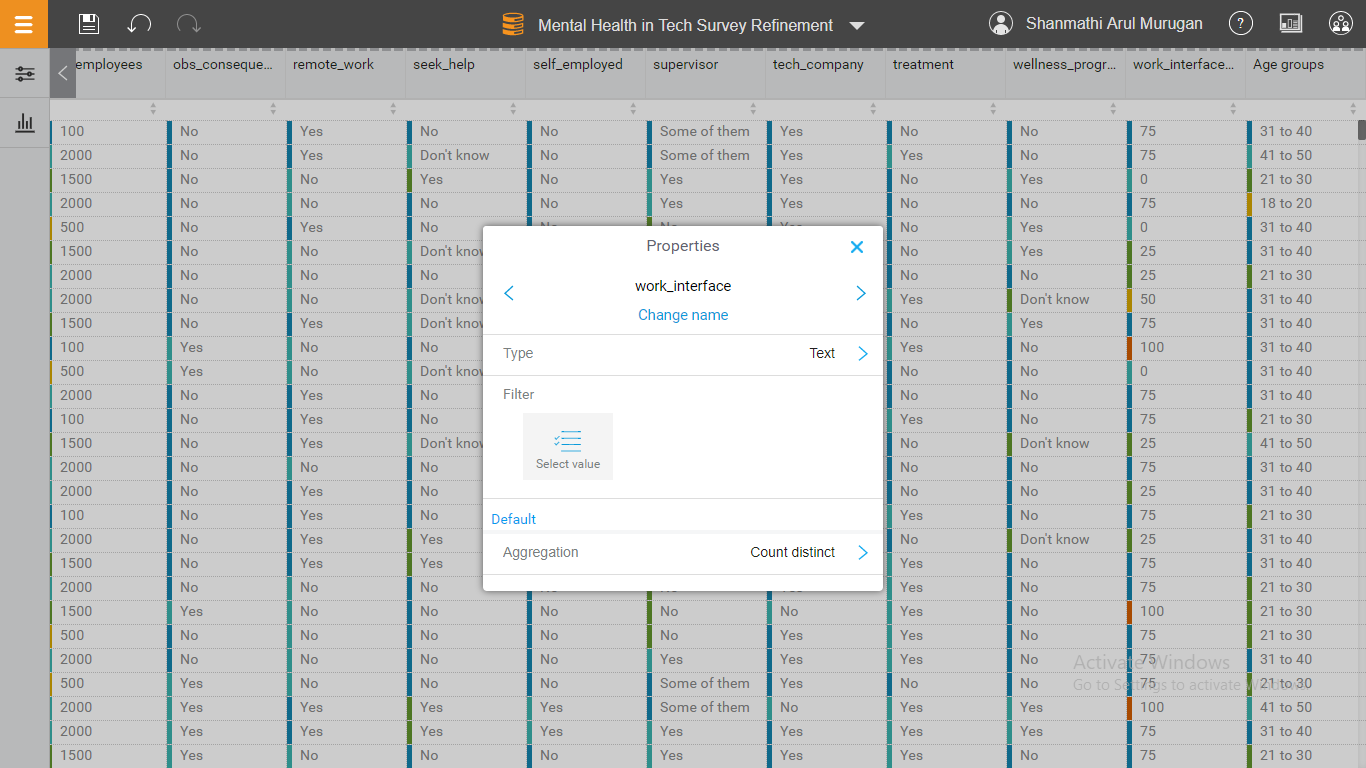
**Description:**

In the Timestamp column of the dataset a range of values have been selected, starting from the year 2014 to 2015, eliminating 2016 since it does not contribute much value to the dataset. Also, to refine the data better, the blank or empty values in the columns have been removed.

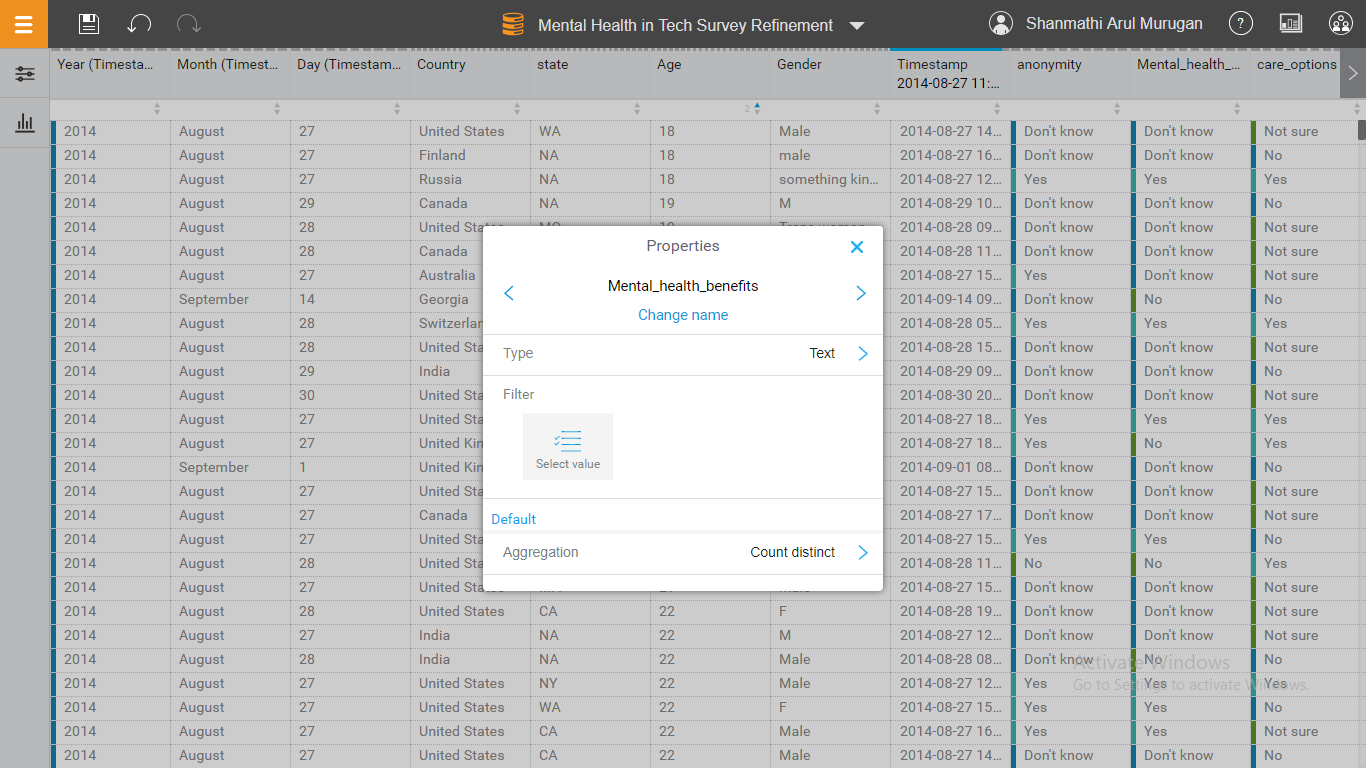
**Type 5:** Renaming column names from ‘benefits’ to ‘Mental\_health\_benefits’ and ‘work\_interface’ to ‘work\_interface\_percent’

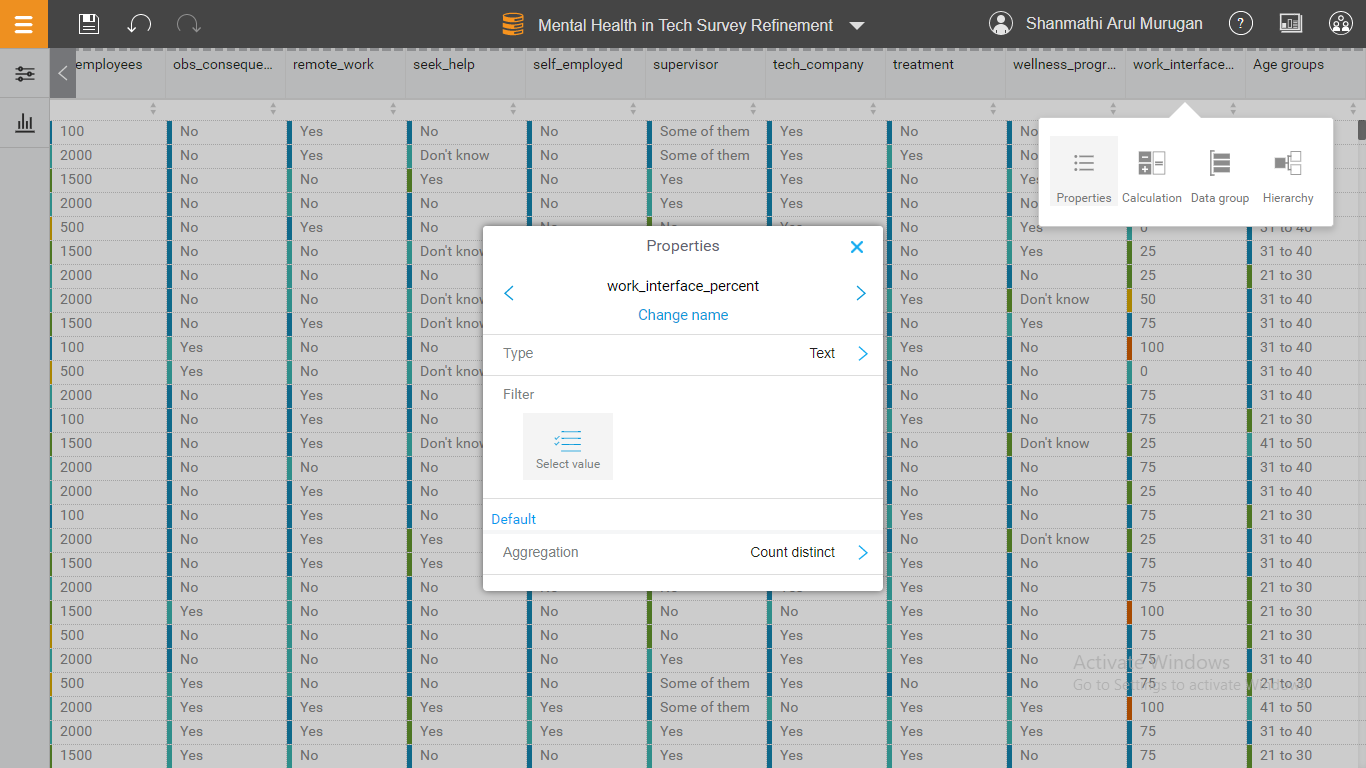
**Pre-Refinement:**





**Post-Refinement:**





**Description:**

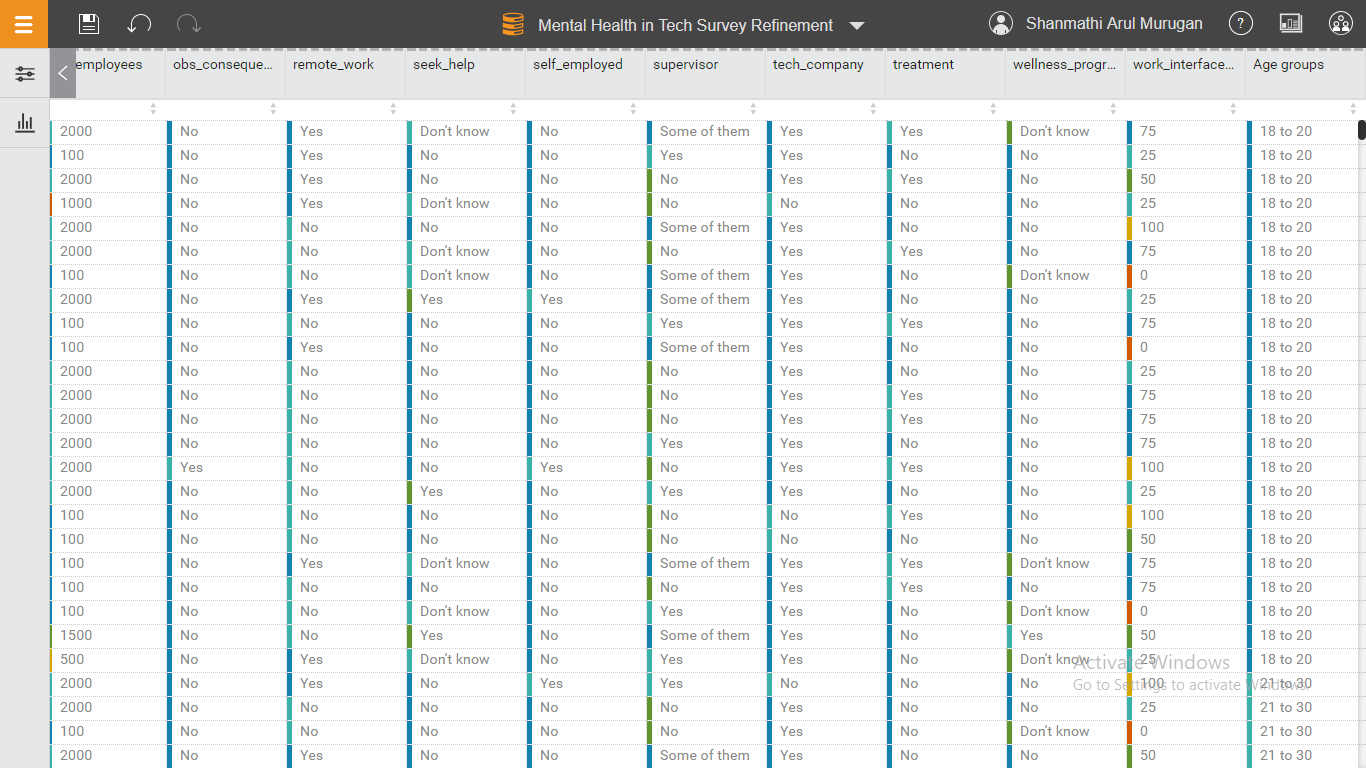
Renaming the column names helps in better understanding of the dataset. It also helps in performing any calculations or analysis based on them easier.

**DATA GROUPING:**

**Description:**

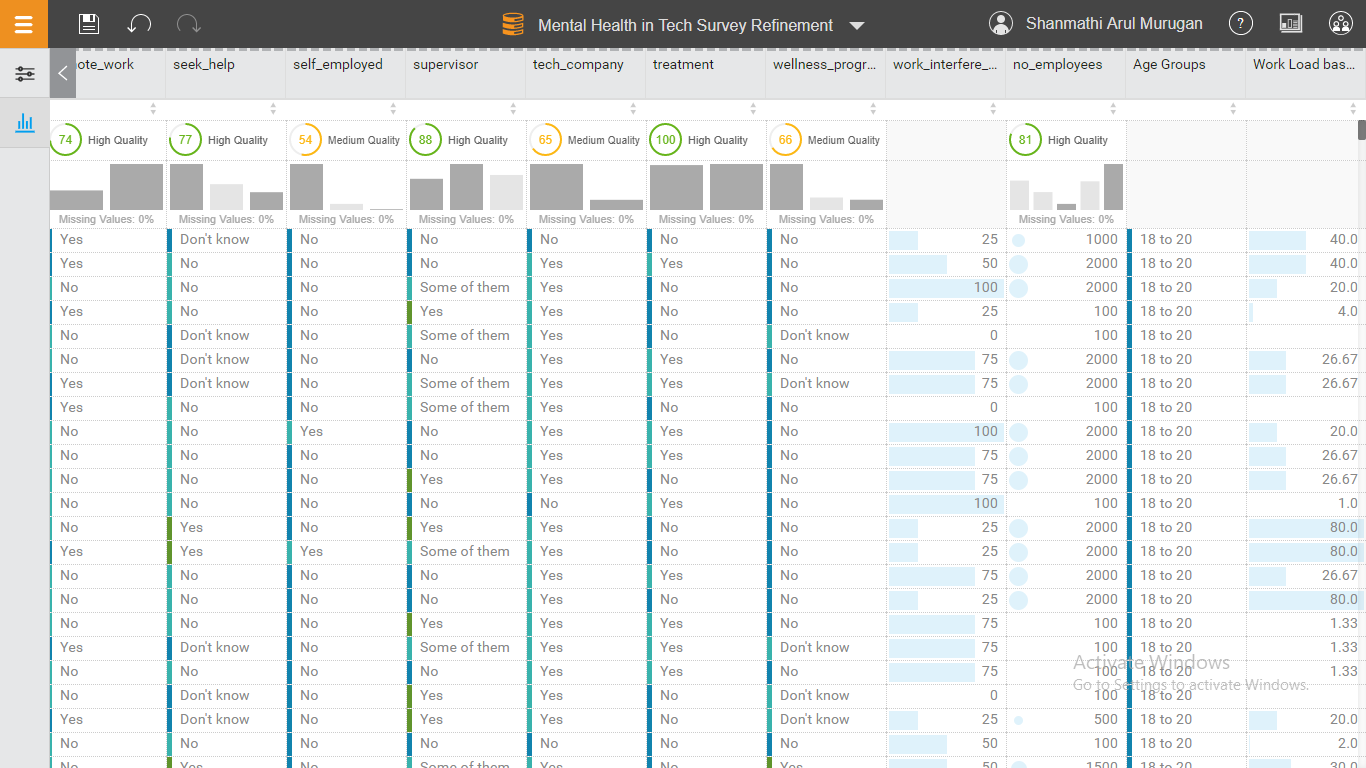
As explained in the refinement section, the Age column contains values varying from 18 to 65 and it would be easy to analyse or perform any calculations if the values are grouped together. Hence the Age column is grouped in range of ten such as 21 to 30, 31 to 40, etc.

**Screenshot:**



**CALCULATION:**

**Screenshot:**



**Description:**

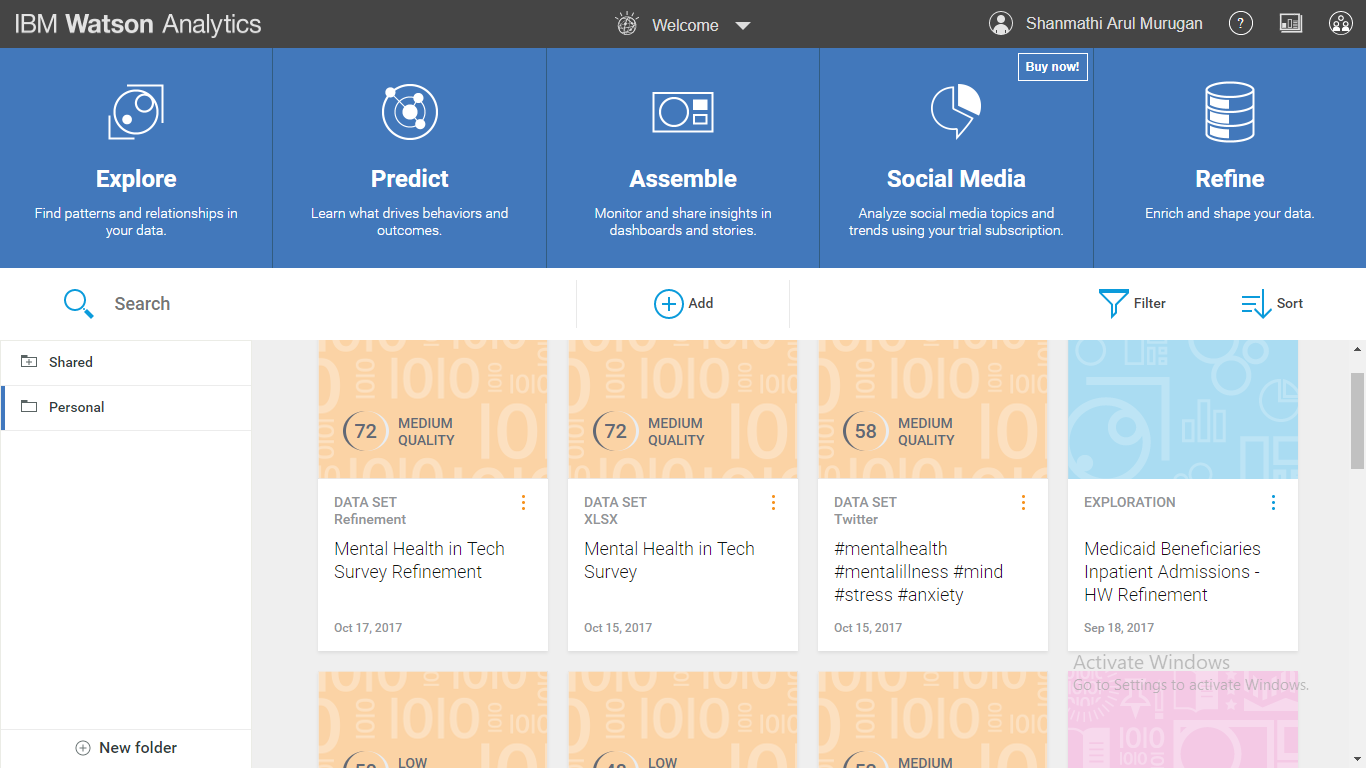
The ‘Work Load based on employees’ is calculated by dividing the number of employees working in an organization by the work interface. This gives an average value of how the individuals are affected mentally based on the work load they receive.

**DATA QUALITY:**

**Pre-Refinement:** The quality of the dataset is 71%



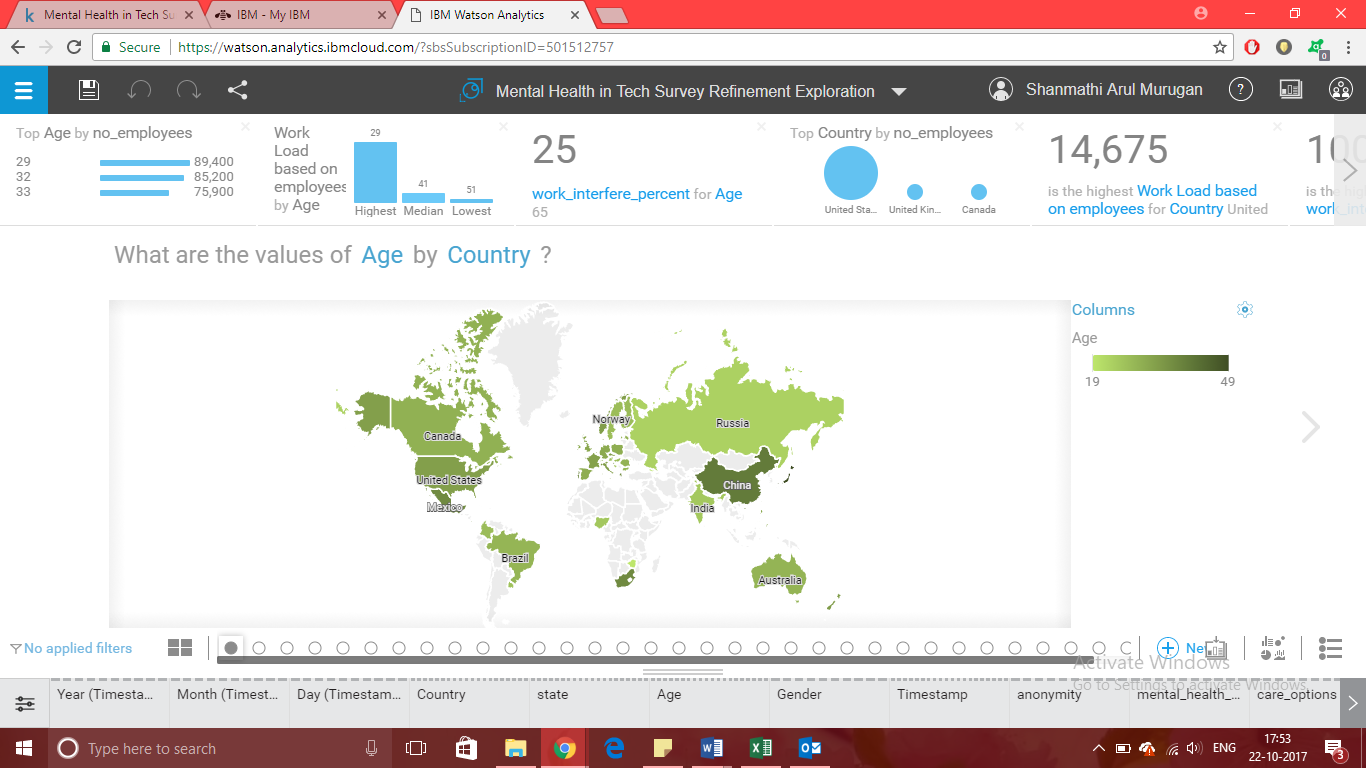
**Post-Refinement:** The quality of the dataset is 72%



**DATA EXPLORATION:**

**Question 1:** What are the values of Age by Country?

**Screenshot:**

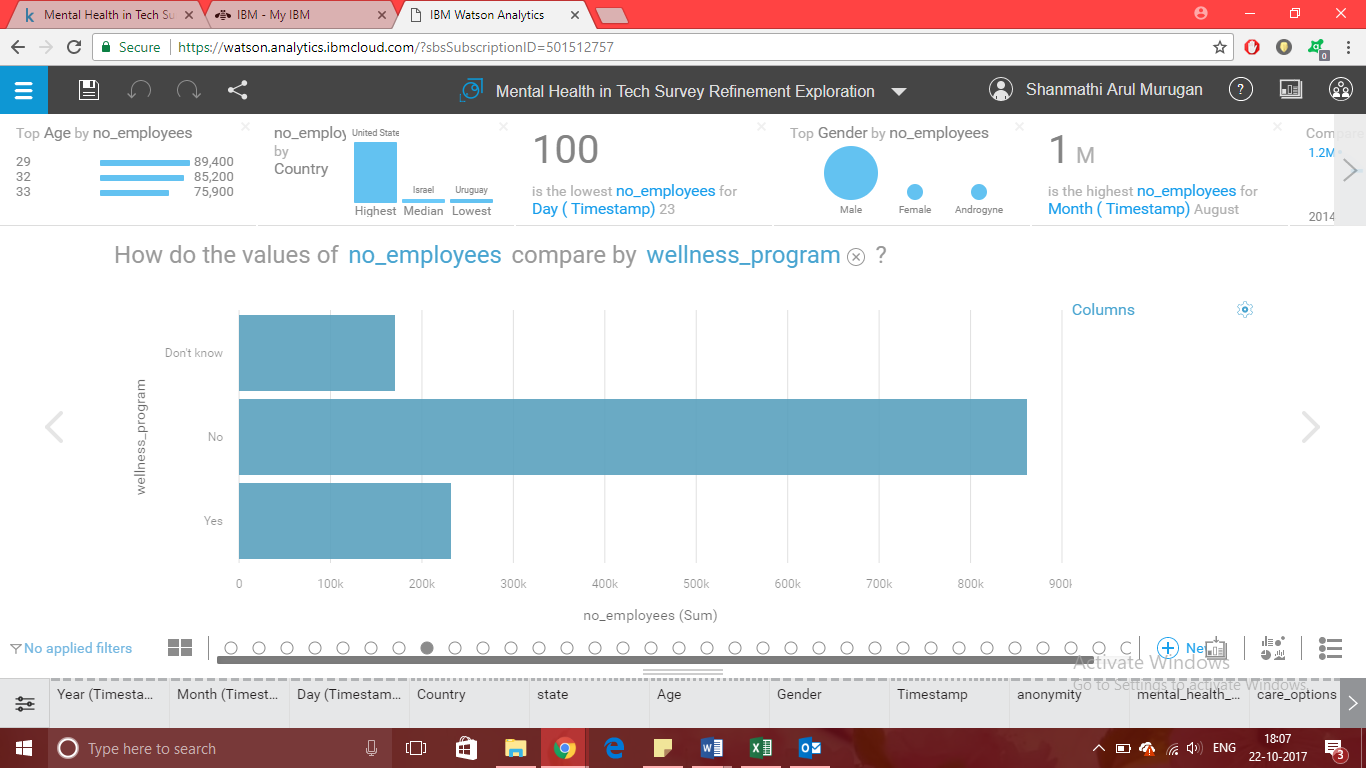


**Description:**

From the visualization it can be observed that people suffering from mental illness are spread among almost all around the world, for about 47 countries to be specific. Based on the colour coding it can be examined that most of the population is in China and United States of America. It is evident that these are the countries that has numerous number of tech industries affecting the mental health of the employees due to the stress level in the work environment.

**Question 2:** How do the values of no\_employees compare by wellness\_program?

**Screenshot:**

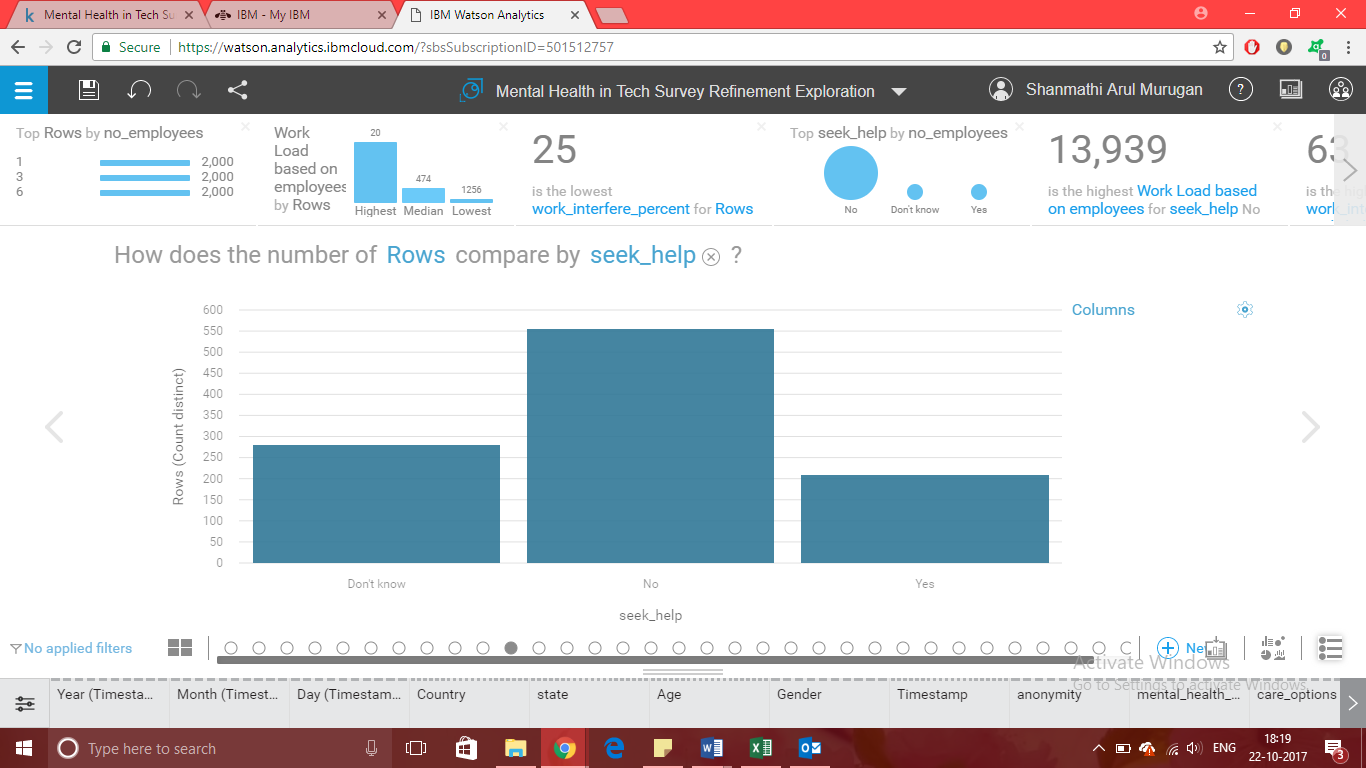


**Description:**

The above visualization is the comparison between the number of employees working in a tech company and the wellness program provided by them for their employees having mental sickness. In the survey, the majority of the crowd has stated that their employers do not provide a wellness program or guidance to the mentally affected workers. It can also be observed that larger the number of workers lesser the guidance provided. As a result, lot of people remain untreated of their mental health issues.

**Question 3:** How does the number of Rows compare by seek\_help?

**Screenshot:**

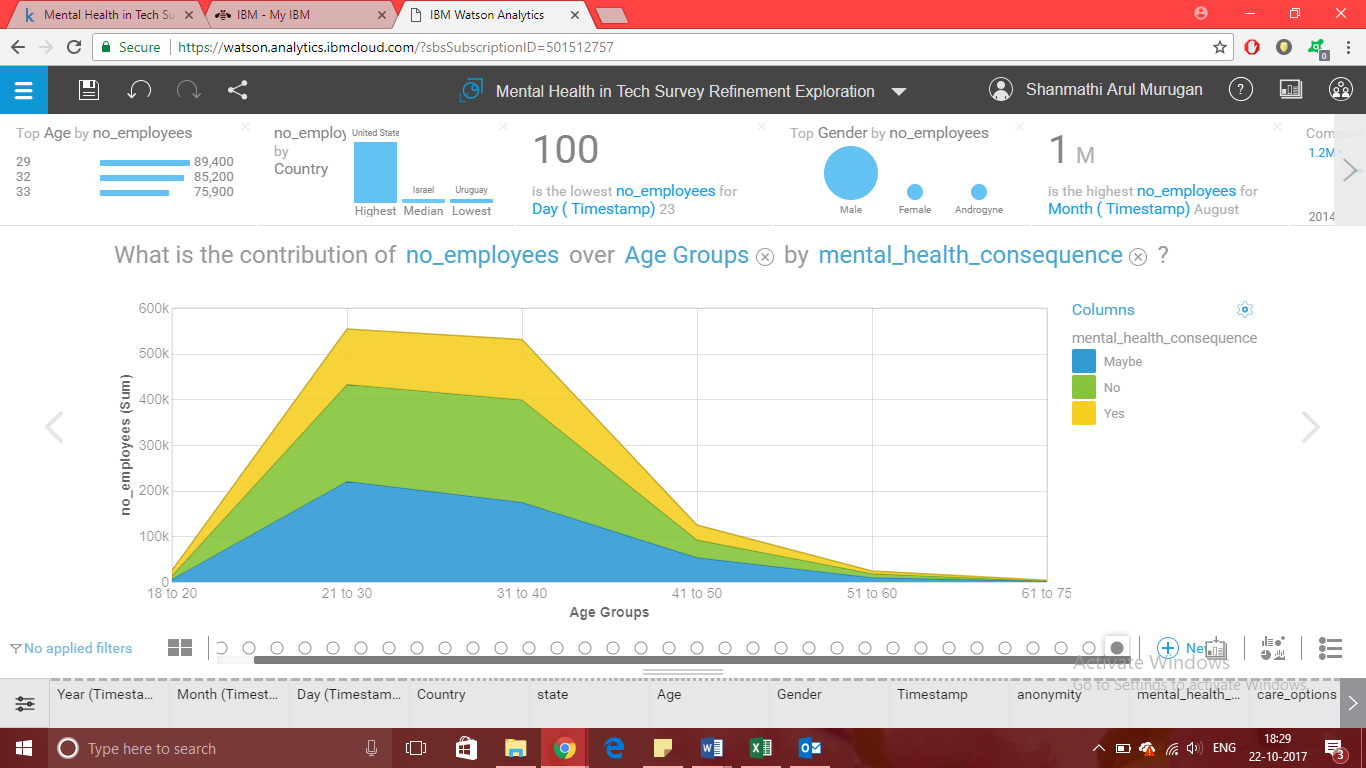


**Description:**

From the above visualization it can be determined that maximum number of patients are hesitant to reach out to their co-workers and supervisors for help or assistance. This is because of the common misconception that prevails in the society, that is people with mental illness are violent and the tendency to stay away from them. This indulges a fear of rejection among the patients and they are reluctant to seek professional help for fear of stigmatization. That attitude has to be changed in order to create an accessible environment for those who are affected mentally.

**Question 4:** What is the contribution of no\_employees over Age Groups by mental\_health\_consequence?

**Screenshot:**

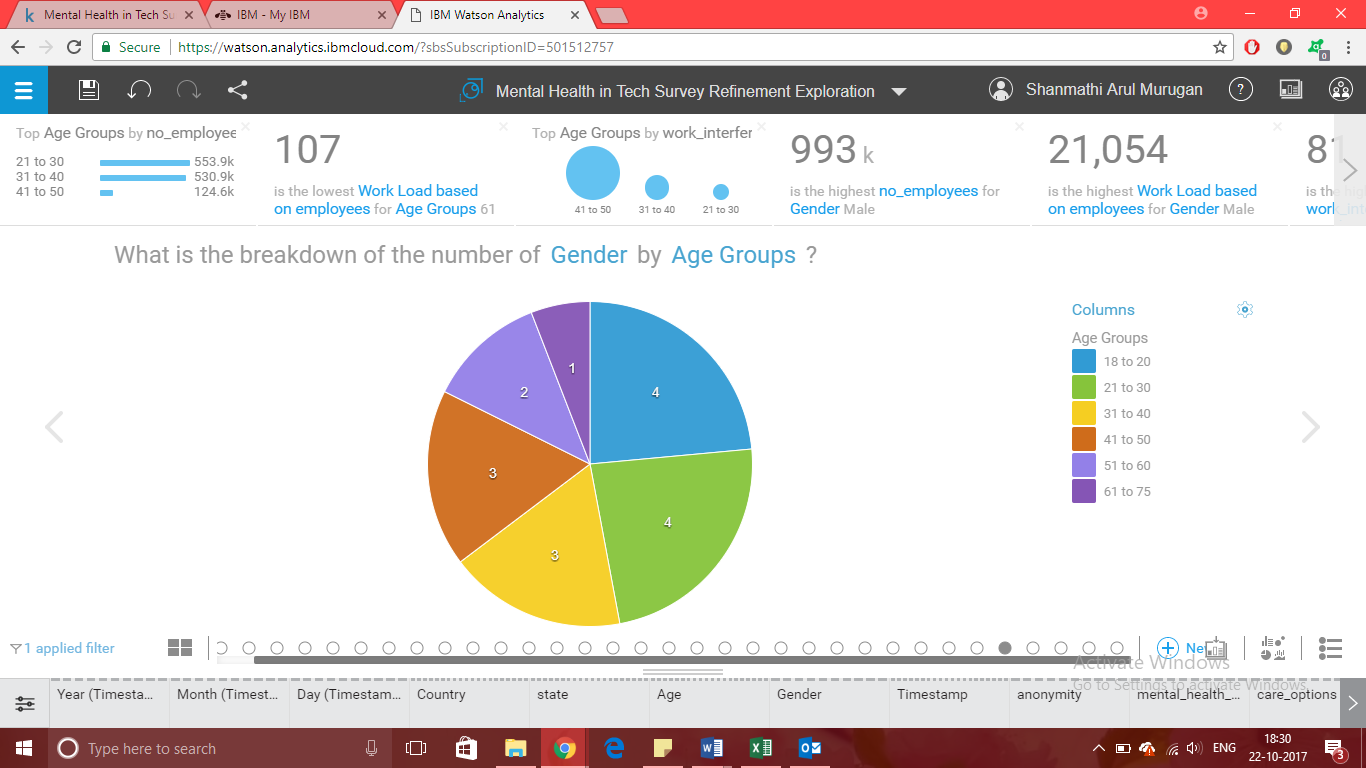


**Description:**

The above visualization specifies how a number of people spread across various age groups think about communicating their mental health issues with their employers or co-workers. It can be analysed that majority of the employees have said that speaking about their condition have had only negative impacts on them. As a result of it, they are not outspoken about their sickness and tend to keep it within themselves and this will only worsen their health condition.

**Question 5:** What is the breakdown of the number of Gender by Age Groups?

**Screenshot:**



**Description:**

The Pie chart about is just to get a clear idea on which age group is majorly affected. It is obvious that individuals aging from 18 to 30 suffer from mental illness. It is because in technology industries young adults are given more work compared to the people aging 30 plus, since most of them will be in managerial positions and lesser work to do. This analysis states that young adults should be given more attention concerning their mental health status.

**Question 6 (Twitter Dataset):** How do the values of Author favourite count compare by Author gender and Sentiment?

**Screenshot:**



**Description:**

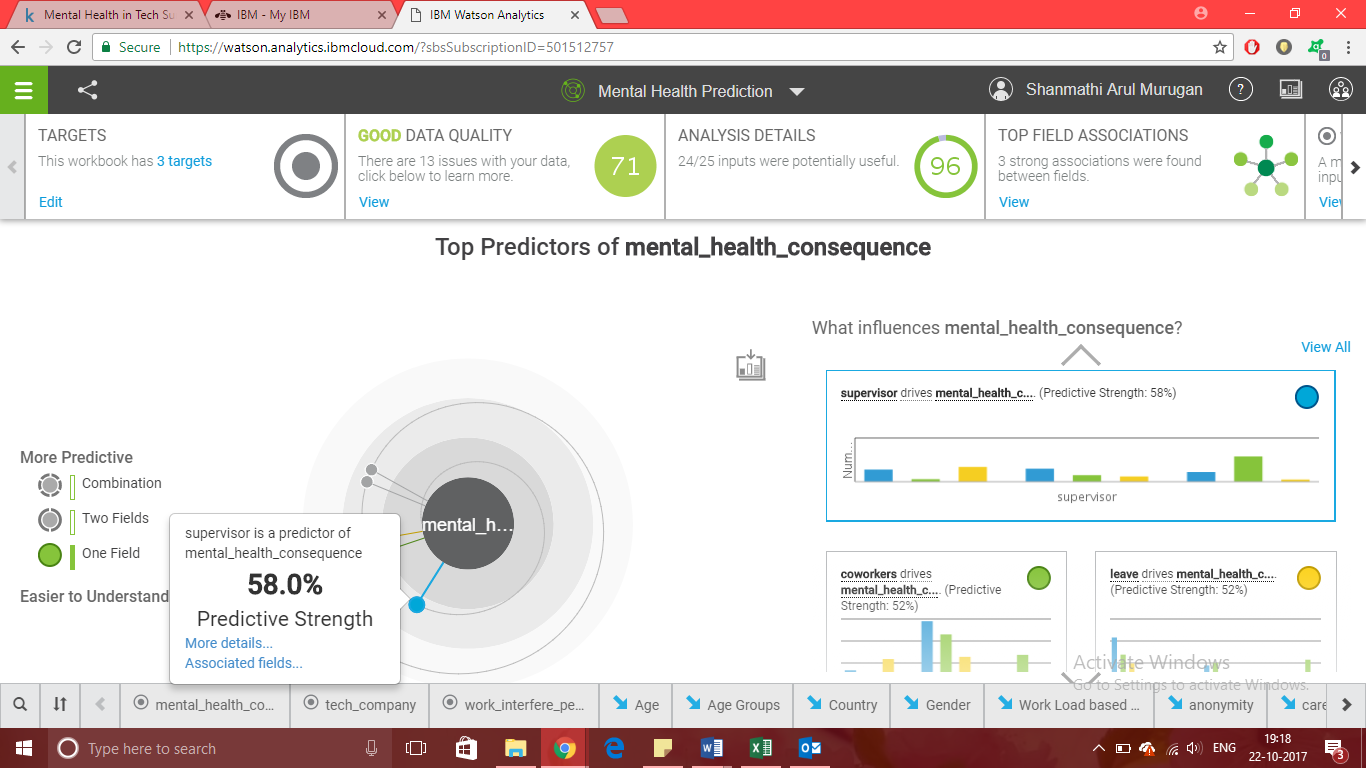
The above visualization is based on the twitter dataset acquired with the help of the hashtags such as #mentalhealth #mentalillness #mind #stress #anxiety #depression #mentalhealthmonday #sadness #psychology #selfhelp. The dataset gives an overall view on the approach towards mental illness among various number of people over a period of time in the year 2016. Based on the sentiment values it can be analysed that majorly there is only negative means of looking at the troubled people. This has to be reduced by spreading awareness on mental health conditions.

**PREDICTION:**

**Description:**

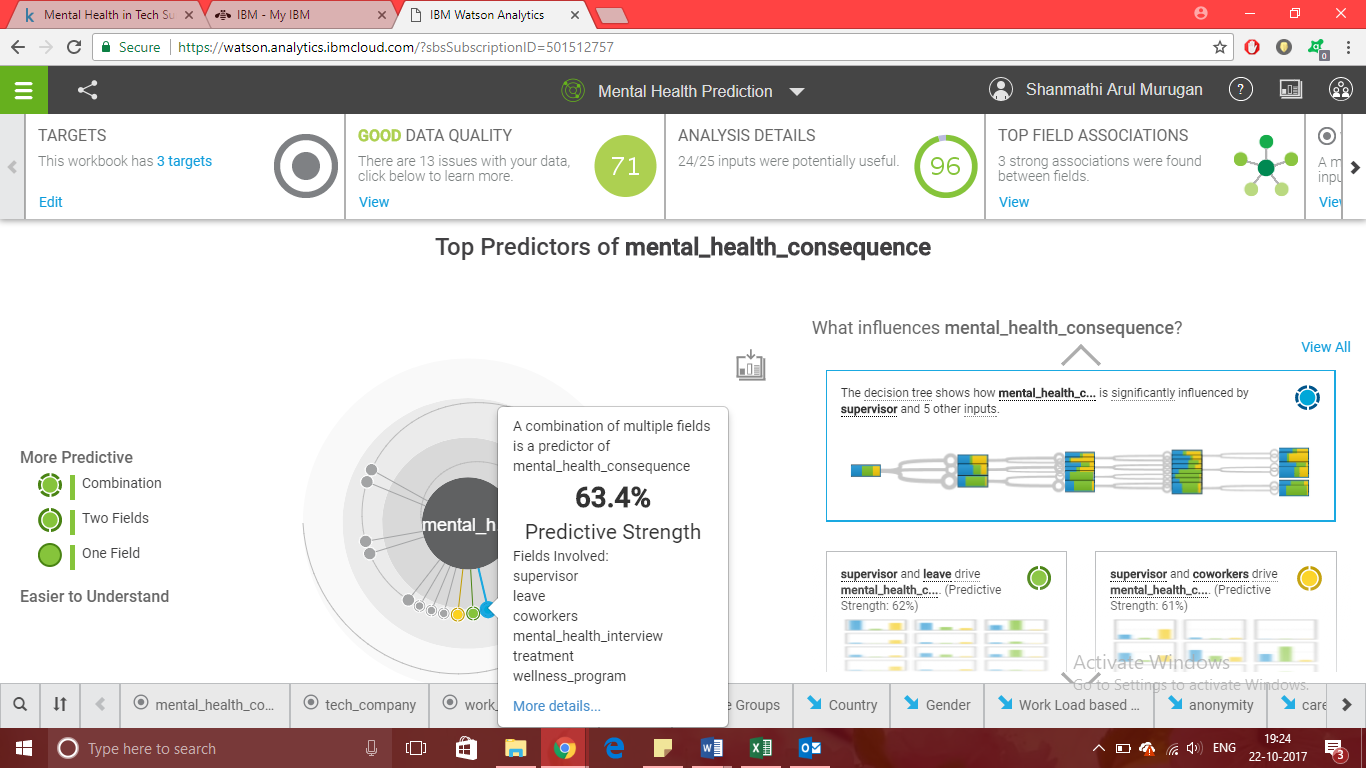
The prediction feature in Watson Analytics is based on 3 important targets of the dataset such as mental\_health\_consequence, tech\_company and work\_interface\_percent. These are the key factors that drives an individual on how they seek their mental health condition in their workplace.

**Bull’s Eye Screenshot:**



**Description:**

The above screenshot is the Bull’s eye of the prediction obtained with the help of just one field. It presents to 58.0% Predictive Strength. Whereas, when we include two or more fields we acquire a better Predictive strength of about 63.4% as shown in the below screenshot. It involves fields such as supervisor, leave, co-workers, mental\_health\_interview, treatment and wellness\_program.



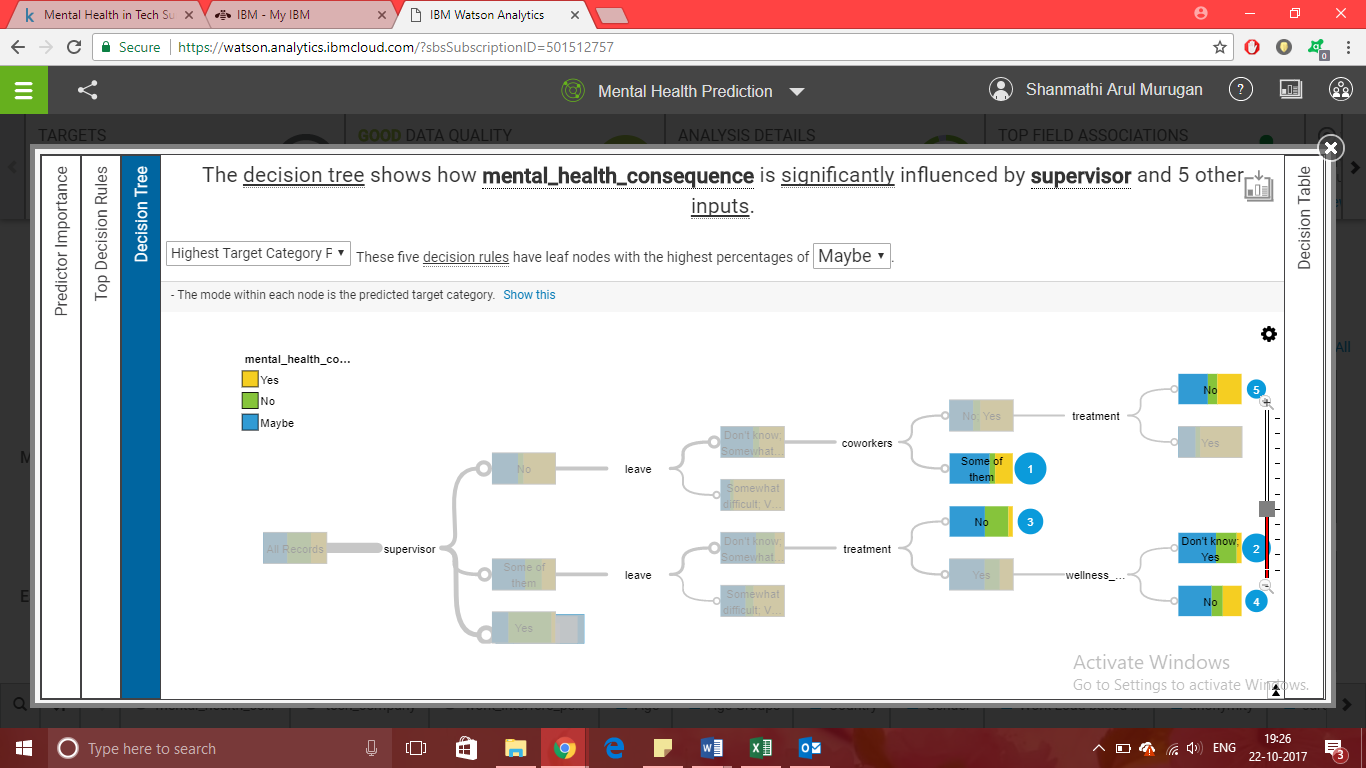
**Decision Table Screenshot:**



**Description:**

IBM Watson Analytics uses CHAID Classification tree to perform prediction with the given number targets and inputs. The result states that people suffering for mental sickness were not comfortable approaching their supervisors and co-workers for aid and in turn those two columns contribute higher values to the prediction. The decision tree shows the mental\_health\_consequence column that is if a person would be fine discussing their mental issue with their employers is influenced. It specifies that 6 inputs such supervisor, leave, coworkers, mental\_health\_interview, treatment and wellness\_program are accountable for it. It can be analysed that these are main factors that drives a person’s attitude towards their mental health. It also can be observed that there has been considerably only a negative effect on the affected individuals.

**Decision Tree Screenshot:**



**DASHBOARD:**

Below is the summary of Mental Health Analysis in Technology Organizations and Twitter Dataset on Mental Health. The assemble option in IBM Watson Analytics gives an overall picture of the analysis done with the help of the datasets. The view contains 3 visualizations from the original dataset and one from the twitter dataset which contributes primarily to the analysis performed.

