**DATA CLEANING AND JOINING**

**About Dataset**

Context

**TECH MENTAL DATA ANALYSIS**

**ABOUT DATASET**

This dataset contains a survey about how employees see, understand and handle mental health.

In the eyes of the society, there is a conception that tech industry is a very demanding industry and there by affects mental health. Interestingly, the importance of data analytics is to be able to support or dispute some of these assumptions with data.

According to Aditi Muyle, Kaggle, Mental health affects your emotional, psychological and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make choices. In the workplace, communication and inclusion are keys skills for successful high performing teams or employees. The impact of mental health to an organization can mean an increase of absent days from work and a decrease in productivity and engagement. In the United States, approximately 70% of adults with depression are in the workforce. Employees with depression will miss an estimated 35 million workdays a year due mental illness. Those workers experiencing unresolved depression are estimated to encounter a 35% drop in their productivity, costing employers $105 billion dollars each year.

Ultimately, we are looking at using this data to measure the most prominent factor that affects mental health.

Secondly, to determine what ways employer can help and support the cause of mental health.

Created a new table that filtered out only thos who are employed

**Open Sourcing Mental Illness** is a non-profit, corporation dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech and open source communities. OSMI began in 2013, with Ed Finkler speaking at tech conferences about his personal experiences as a web developer and open source advocate with a mental health disorder. The response was overwhelming, and thus OSMI was born.

* Every year, **OSMI** came out with a new survey to see how employees want to get mental health treatment in tech companies around the world and I pick the survey from 2014.
* This survey is filled by respondents who suffer from mental health disorders (diagnose or un-diagnosed by medical, even it's just a feeling) in tech companies and see if any factors can affect the employee to get treatment or not.
* From this research, this m

Survey data (2014, 2016, 2017 and 2018)

The aim of this dataset is to provide access to the raw survey data from the 2016, 2017 and 2018 OSMI mental health in technology surveys used to facilitate analysis e.g [my kernel fusing the OSMI surveys across time periods](https://www.kaggle.com/ekwiecinska96/dataset-creation-fusing-surveys-from-2014-2018).

This is due to the fact that the popular 2014 dataset uploaded onto Kaggle has already been pre-processed and cleaned (and the only other 2016 upload does not play nice with kernels). Whilst this is useful, many columns were renamed into simple attributes e.g 'Are you self-employed?' is standardised to 'self\_employed'. As none of the  
surveys from the following years have had this treatment, it was difficult to reverse-engineer the processing steps to make the attributes match. Also, it's great to have all the data in one place.

Similarity matrix

The associated similarity matrix, stored as a numpy-readable file (.npy) is a supplementary file for the previously mentioned kernel. This was uploaded due to the unfortunate fact that any [SpaCy models](https://spacy.io/usage/models)that are contain word vectors (aka any model other than *sm*) are not supported by Kaggle on the date of writing (Jun 2019). Please see the associated kernel for more information on how this matrix was created.

Acknowledgements

The original data collection and hosting has all been provided by [Open-Sourcing Mental Illness (OSMI).](https://osmihelp.org/) you can find all of the datasets (including 2016, 2017 and 2018) [here](https://osmihelp.org/research).

Inspiration

The inspiration for uploading these datasets was to allow Kaggle users such as myself to have greater control over the pre-processing and standardisation of attributes.

Data Preparation

Dataset Original has 63 columns and 1434 rows

Drops

Some columns that are not relevant to this particular analysis have been reduced, resulting into 30 Columns and 1433 rows

Column names

The column titles are in question format and lengthy as well. So, renaming the columns will be best advised so as to allow easy manipulations and less clumsy exercises.

Age

We will be removing outliers and these are ages below 15 and above 74

Gender inconsistencies

There are two observations and cleaning we have to do here. First will be the typos and secondly, reducing gender types to the three major categories.

Missing values/ Nulls

Although some of the columns were blank. These shall be treated as missing and considered during analysis.

State column

The missing values in the columns are as a result of the other countries present

It looks more like this survey was based in the USA (as it has 70% of all the entries) but with other nationals in the industry

Therefore, it will not be fair to make a substantial demographic conclusion as some countries have just single or very insignificant number of entries.

Analysis and Inferences

Gender- There are more males than any other genders in this data-

Therefore, we must be careful not to infer that the male gender is more susceptible to mental health issues

However, it is safe to conclude that male workers are more than the female or non-binary in the industry.

The Question

To help employers find the most prominent areas or ways they can help employee maintain or manage mental health which revolves around mental, psychological and social wellbeing.

EDA

58.52 % of the entire population has sought treatment for mental health

Out of these 58.52% people who sought treatment, 36% of them has a family history of mental health issue. It is acceptable to say the family history factor influenced their willingness to sought treatment as there is a decrease in treatment among the percentage of people who do not have family history.