correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

Project Checkpoint

Due: March 15th 11:59 pm PT on the Participant Portal (https://cor1.co/login/).
**Each member of the team needs to upload the project to the portal*

Problem Statement: *Mental Health in Tech Companies*

How is the productivity affected by the work environment in Tech companies for people with mental health problems?

Please provide an updated description of any datasets that you are using. Briefly mention any challenges you have faced in cleaning and transforming the data.

We are using the OSMI Mental Health in Tech Survey from 2014-2019, which consists of a set of survey questions asked to tech workers annually about mental health issues in the workplace; this dataset includes information on attitudes, employer support, and personal factors. One challenge we found was that data across years is not always consistent so matching variables can be challenging. There was also a lot of comments/discussions-based data which could be valuable, but would be difficult to quantify. The dataset also had a lot of missing values, so we have to decide how to interpret the significance of these missing values.

Briefly describe any findings and insights derived from exploratory data analysis (EDA). Has this sparked new questions that you want to explore or changed your goals entirely?

We found a clear difference between those who received treatment and those who didn't in terms of work productivity. We also identified several factors that might affect if someone seeks treatment, such as company support, work environment, insurance coverage, and medical leave. We plan to use these features together with some personal information to predict how likely someone is to seek treatment for their mental disorder, and relate this to their work performance.
While exploring the data, an interesting question sparked: how likely is for a person to develop a mental disease while working in a specific work environment ? A first approach on answering this question showed that we might be facing a lack of statistics (after a proper selection of our data).

Please provide an updated description of the methodology that you will use to answer your question (indicating any statistical techniques or software that you plan to use)

We plan to use Pandas for data manipulation. We will define a dataframe with specific features which we consider important for our prediction analysis. We also plan to apply a specific pre-selection criteria to our data, in order to restrict the study to a specific category (e.g. employees from US only).
We will split the entire dataset into two subsets: training and test datasets.
We will use Python as our primary software, and Scikit-learn for training the Machine Learning algorithm.