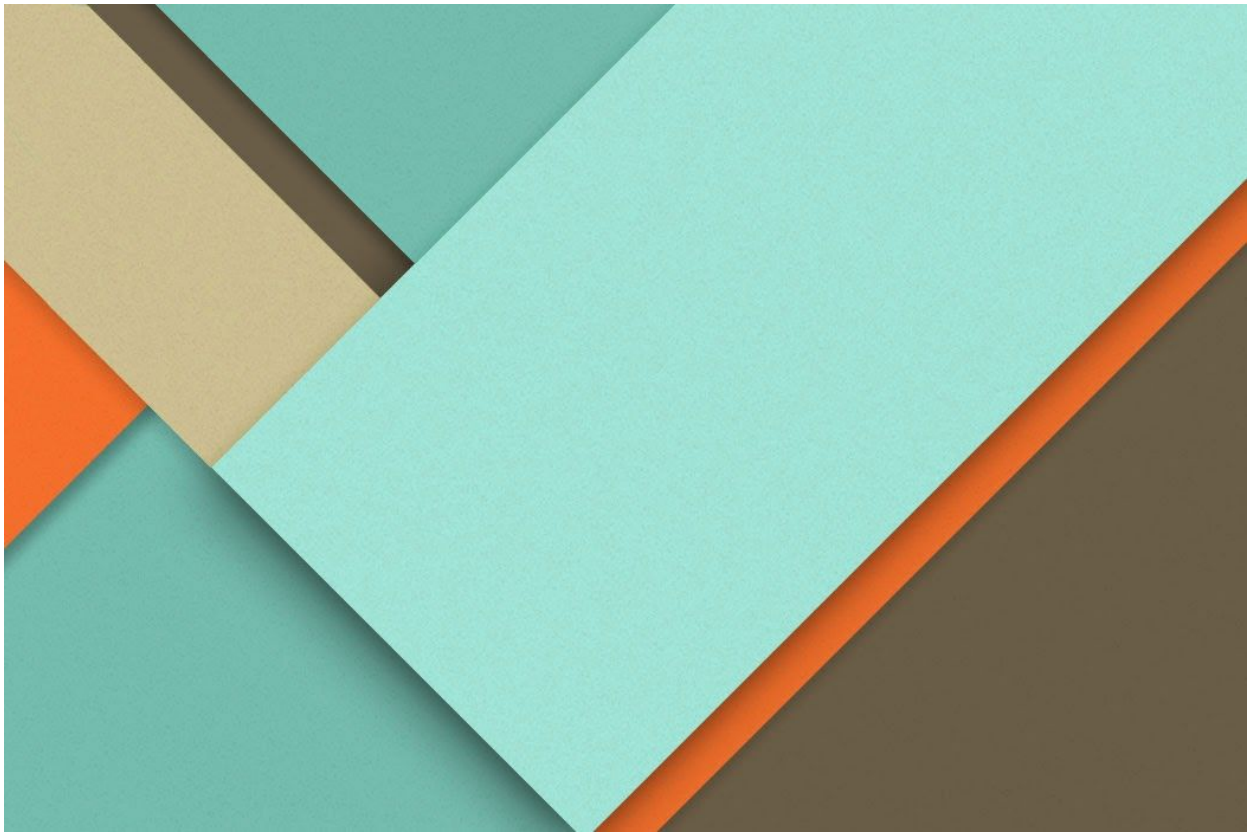

Project 1: Verina, Kenny, Shivani and Eline

June 2020

Data Engineering Bootcamp

Columbia University

“Mental Health in Tech”

Overview

In this first project, we want to find an answer to the questions stated in the Goals section,, applied to the tech sector. For this we will use a dataset retrieved via kaggle.com.

Url: https://www.kaggle.com/osmi/mental-health-in-tech-survey

Authors: Verina Hanien, Shivani Chopra, Kenny Wu, Eline Van Eldere

Goals

In order to get an understanding of the mental health status quo in the tech sector, which was specified by the survey cohort in a dropdown menu where they indicated whether or not they were working in a tech company.

In order to formulate hypotheses we want to test, we center our analysis around the following three questions:

1. In terms of seeking help or reaching out, are women or people who identify as female more vocal than men or people who identify as male?
2. Is there a significant difference between age groups when it comes to using or reaching out for mental help services in the workplace?
3. When reaching out for mental health in the workplace, is guaranteed anonymity a deciding factor?

We work around three central hypotheses:

1. When seeking mental health in the workplace, women or people who identify as female are more likely to reach out for help
2. Age is a decisive factor when it comes to seeking mental health
3. Guaranteed anonymity is a decisive factor for people to seek mental health care in the workplace

Specifications

Justification of dataset used

Dataset: OSMI Mental Health in Tech Survey from Kaggle public data website. This dataset consists of 1260 rows and 27 columns. The dataset is comprised of responses from individuals in the tech industry about mental health.

Milestones

I. Data cleanup

- We decide to analyse the 2014 part of the database as it was the most suited for the purpose
- We use both Excel (and Python) to clean data

We determine both the gender and date columns needed some cleaning up, we decided to do this by performing an Excel

II. Data aggregation

In order to adapt the dataset to our needs, we choose to aggregate the data to be able to manipulate and navigate them in Python. This is a fairly straightforward task since we are dealing with one single dataset.

III. Data analysis

Since we are aware of the mostly categorical nature of our dataset, we keep this in mind when heading into the data analysis. We will adapt our visualisations and analysis tools to suit the data type we are working with i.e. when it comes to company size, we see a number of categories which will be visualized in histograms rather than scatter plots.

IV. Final Conclusions

Our visualisation and analysis is tested with our original hypotheses, both of them are to be presented using Powerpoint, paving the way for further discussion and research suggestions.