Group A8 Report

Tech-Health

Analysing the factors affecting mental health in the tech industry

Magnar Salei
Hendrik Ojamaa
https://github.com/EnriqueOjamaa/mental-health.git

1. Business understanding

1.1 Identifying our business goals

- Background

  We believe that mental health in the tech-industry is a relevant topic for many people attending this course and who are active in the sector in general. Additionally, people working in IT have the highest percentage of remote workers[1], which might differentiate the nature of mental health issues for people involved with it.

  The topic has not yet been thoroughly studied, which might make it a bit difficult for us to analyse, since there's not much material to rely on. On the other hand, that increases the value of the research, and hopefully, our findings are able to provide valuable insight for people within the industry.

- Business goals

  As we briefly mentioned beforehand, our goal conclusively is to detect places of concern regarding the hazards of working in the tech industry. More specifically, our target is to clearly and concisely visualise these findings and come up with practical suggestions on how to prevent such issues from arising.

[1] https://workplaceinsight.net/top-global-industries-leading-the-way-in-remote-work/

- Business success criteria

  Since the study deals with a not easily and quantitatively measurable topic, we find it not purposeful to outline specific KPI's to follow. However, we would hope to measure the success of our work based on the feedback we receive - since the project is directly targeted to our fellow students and lectors, then we would measure the outcome by how useful others find it to be.

1.2 Assessing our situation

- Inventory of resources

  The project is carried out by two students of the course. One of us, Magnar, has extensive experience in programming, whereas Hendrik's professional background is closely involved with IT project management.

  We are using a Kaggle dataset. It includes an SQLite file with 3 tables - *survey*, *question*, and *answer*. The raw data was originally processed using Python, SQL and Excel and contains all information needed for data analysis.

  The course materials are our main source of information for visualising the results and creating different models. We also intend to seek further insight online, if needed.

- Requirements, assumptions, and constraints

  The main point of concern regarding the project seems to be differentiating mental health issues regarding remote work from the tech-industry itself, as the tech-industry has the highest percentage of remote workers. Meaning that the problems might be derived simply from the isolation and not be related to IT.

- Risks and contingencies

  We find the greatest risk to be being unable to follow the deadlines and stay on track with the project, since we have to share our attention with different projects, both academically and professionally. However, we are countering it with a concise timeplan for executing the project.

- Terminology and costs and benefits

  We find it not necessary to list out these points. Firstly, coming up with terminology for the project doesn't seem to simplify our project, as there are just two of us and communication is effortless.

  Secondly, there are no financially measurable costs nor benefits, and qualitatively these would be difficult to define.

1.3 Defining our data-mining goals

- Data-mining goals

  Our goal is to find correlations between different parameters of IT-industry specifics and mental health. Based on these findings, we will create models, compare the results and eventually, visualise the outcome. We will then conclude the information in a presentation and create a report.

- Data-mining success criteria

  Since we do not set specific KPI's, we will try to find as many significant correlations as possible from our dataset. There are three correlation coefficient values that we are going to consider as our threshold values: < 0.4 is weak/insignificant, 0.4-0.7 is moderate and > 0.7 is strong/significant.

2. Data understanding

2.1 Gathering data - data requirements, availability and selection criteria

We are using an SQLite database available on Kaggle. The original, raw data is from the Open Source Mental Illness (OSMI) surveys from five different year groups, which has been processed using Python, SQL and Excel for cleaning and manipulation. Therefore there is no need to perform further data mining with the dataset. Additionally, it has the necessary information to carry out data analysis on the subject, meaning that we don't intend to use other datasets for the project.

2.2 Describing data

As mentioned earlier, the data originates from the Open Source Mental Illness (OSMI) surveys from years 2014, 2016, 2017, 2018 and 2019. The dataset and the survey includes 105 questions to firstly, identify and group the respondents; and secondly, to get insight about the state of their mental health and professional career.

The data is layed out followingly:

Table	Total Rows	Total Columns
Answer	236898	4
Question	105	2
Survey	5	2

2.3 Exploring data

Since the data is already processed and usable for data analysis, there is no need to prepare the data ourselves. When looking through the data, we were convinced that it is indeed properly cleaned and processed.

After further analysis of the data and mostly the nature of the questions, the dataset seems to be very well associated with the topic and goals of our study. That is why we don't find it necessary to involve additional datasets.

2.4 Verifying data quality

The dataset is readily available for data analysis. The author of the dataset carried out the following steps to clean and prepare the data:

- Similar questions were grouped together
- Values for answers were made consistent (for instance 1 == 1.0)
- Spelling errors were fixed

3. Planning of the project

3.1 Further studying the nature of mental health issues and the peculiarities of the tech industry

In order to understand the detected correlations between parameters after the data analysis, it is necessary to understand the nature of the industry that the respondents of the survey are engaged in. We are basically analysing their evaluations about their work life and mental state, which would be more meaningful if we comprehend the context of it.

Time cost for Magnar: 5 hours
Time cost for Hendrik: 5 hours

3.2 Finding correlations between parameters, creating different models, visualising the findings

This part covers the technical execution of the project.

Time cost for Magnar: 15 hours
Time cost for Hendrik: 7 hours

3.3 Analysing the outcomes for conclusions and insightful graphs

We will interpret our findings of the data analysis, visualise them and come up with conclusions. We find the first phase of the project, where we're getting acquainted with the specifics of the topic, being very important for interpreting the results. We also intend to include previous studies and statistics for explaining the wider background.

Time cost for Magnar: 5 hours
Time cost for Hendrik: 8 hours

3.4 Creating a video and a poster about the project (14th of December)

Eventually, we will be looking to forward our results to others, which needs to be done in a format of a 3-minute video and a “poster”, which will basically be our main slide.

Time cost for Magnar: 5 hours
Time cost for Hendrik: 10 hours

3.5 Preparation for the presentation (17th of December)

In addition to the forementioned poster, we will also create a full slideshow describing our project and prepare for the final presentation. We are also looking forward to the chance to see and discuss each others' work :)

Time cost for Magnar: 3 hours
Time cost for Hendrik: 3 hours

Total time cost for Magnar: 33 hours
Total time cost for Hendrik: 33 hours