HERALD COLLEGE KATHMANDU

UNIVERSITY PARTNER
UNIVERSITY OF WOLVERHAMPTON

Academic Year	Module	Assessment Number	Assessment Type
A22	6CS030 - Big Data	A1	Report

Big Data Analytics on the state of Mental Health Support in the Technology Industry

University Id: 2050062

Student Name: Sahaja Phuyal

Award/Course: Bsc. (hons) Computer Science

Cohort: 5

Module Leader: Jnaneshwar Bohara

Tutor: Yamu Poudel

Acknowledgement

I would like to express my deepest gratitude to those who have assisted me in successfully completing the project. Firstly, I would like to extend my gratitude to the University of Wolverhampton and Herald College Kathmandu for providing me with the opportunity to work on this project.

I am very grateful to my Tutor Mr. Yamu Poudel for his constant guidance, encouragement and support throughout this semester. It has certainly made me stay focused and motivated and helped me to successfully complete this project. I would also like to acknowledge my lecturer Mr. Jnaneshwar Bohara who has provided me with valuable insights and helped me understand the complex concepts.

Abstract

Generic Information

This project examines the state of mental health support for employees within the tech industry and how it has evolved over the years 2014 to 2019 using the survey datasets from OSMI, an organization that conducts annual surveys to generate a better understanding of how mental health is addressed within tech companies.

Problem Statement

Mental health disorders affect 990 million individuals worldwide. The tech industry is a very competitive sector and tech workers are more likely to be suffering from mental health issues compared to the wider population. But talking about mental health issues in the workplace is highly stigmatized in our society or even in the global context.

Aim

The main goal of this study is to know about the status of mental health support for tech personnel and how it has changed in recent years.

Contributions of the work

After the analysis, this project provides us with the number of people who are and aren't aware of their mental health benefits and helps to find effective ways to raise awareness among the ones who aren't much aware regarding their mental health benefits coverage

Social Impact of the work

This project can be very impactful to the number of tech companies and tech workers as it raises some level of awareness regarding mental health and help companies and employees to keep track of their mental health.

Contents

Table of Figures

1. Background of the Study

1.1 General Information

The term "big data" has been gaining quite a bit of momentum throughout the world in recent years. It has its application in so many areas including manufacturing, marketing, finance, education, etc. Especially, in the field of healthcare, Big Data has made tremendous advancements by improving public health strategies, conducting various medical research, finding better ways to conduct treatment, etc. (Ting, 2020) However, its application in Mental Health remains relatively low in comparison with other healthcare applications. (Davis, 2018) Nonetheless, research and big data analytics on mental health issues have begun to reach a turning point.

This paper focuses on performing big data analysis on the state of mental health support in the technology industry. For this, datasets are taken from the survey done by Open Source Mental Illness (OSMI) from the year 2014 to 2019. OSMI is an organization that conducts annual surveys to generate a better understanding of how mental health is addressed within tech companies. Different big data technologies like Machine Learning, Hadoop, Spark, ELK, etc. have been used to carry out effective data analysis and some valuable insights have been retrieved which will be discussed throughout this report.

1.2 Problem Statement

Over the past few years, due to the massive increase in mental health disorders among people and their major global effects, it has become necessary to better comprehend the disappointing consequences of mental health problems. (Q. Liu and X. Liao, 2021)

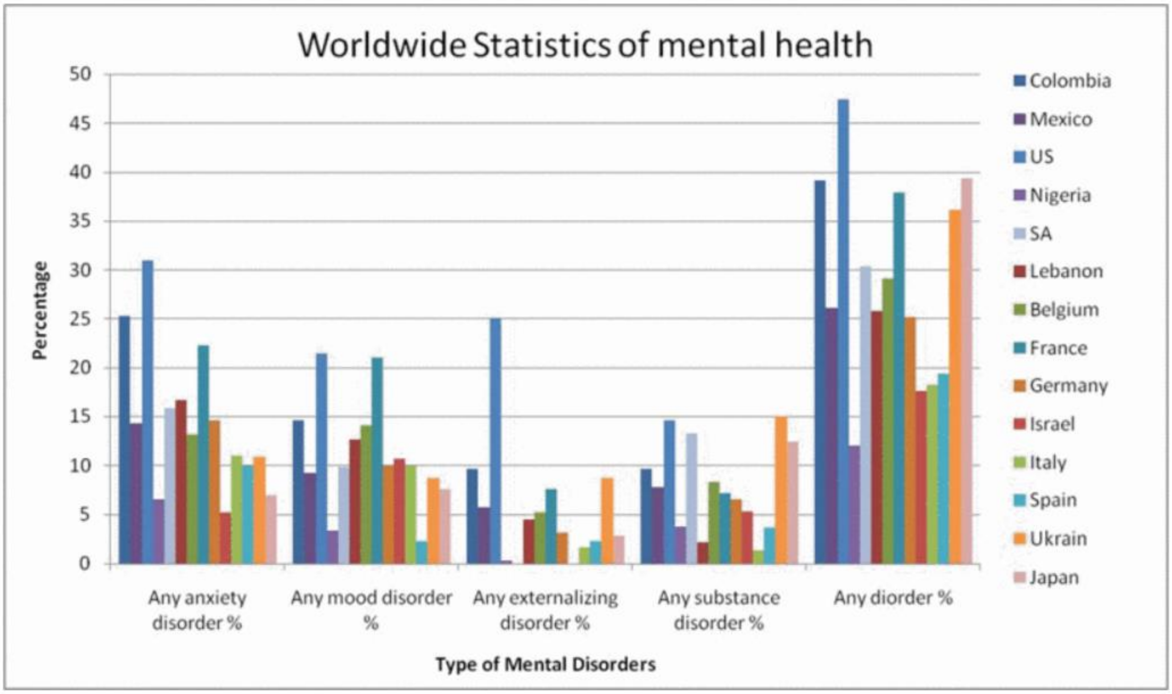


Figure 1: Worldwide Statistics of Mental Health (Ayesha Kamran Ul haq, 2020)

As shown in the figure above, people have been affected all around the world by mental health issues. Mental health disorders affect 990 million individuals worldwide. Especially, in the technology industry, people are often seemed to be dealing with mental health issues. According to recent statistics by Global Health Data exchange, mental health problem has been detected in 51% of tech employees and has impacted their performance by 71%. (Global Health Data Exchange, 2022) As technology is the most competitive industry, tech workers are often seen having irregular work hours, pressure to meet deadlines, and overestimated goals all of which can contribute to a high degree of stress and mental illness. But employees usually opt to keep their mental problems to themselves rather than seeking help as mental health disorders are stigmatized in our society, especially in the workplace. Therefore, this is a major problem that needs some thorough data analysis in order to produce useful outcomes and solutions.

1.2 Aims and Objectives

1.2.1 Aims

The major aim of this project is to examine the state of mental health support for employees within the tech industry and how it has evolved over recent years.

As mentioned in the problem statement above, the tech industry is a very competitive sector and tech workers are more likely to be suffering from mental health issues compared to the wider population. But talking about mental health issues in the workplace is highly stigmatized in our society or even in the global context. Thus, this project aims to somewhat remove the stigma regarding mental health issues and also raise awareness among tech workplaces through effective data analysis.

1.2.2 Objectives

The major objectives of this project are:

- To analyze the outlook and attitude of employees regarding mental health issues in the tech workspaces.
- To find out the number of employees who are and aren't aware of their mental health benefits in their workplace.
- To determine how comfortable employees feel to discuss mental health issues with their coworkers.
- To find out the number of employers who provide mental health benefits as part of healthcare coverage.
- To come up with some insightful findings and ways to make employees aware of mental health issues.

1.3 Contributions of the work connected with Methodology

This project has the potential to make a significant contribution to the field of mental health. They are as follows:

- This project helps in finding out effective ways to enhance mental health conditions in tech workplaces.
- After the analysis, this project provides us with the number of people who are and aren't aware of their mental health benefits and helps to find effective ways to raise awareness among the ones who aren't much aware regarding their mental health benefits coverage.
- It helps in finding ways to make employees feel comfortable with each other to talk about mental health problems.
- It helps in eradicating the stigma that prevails in society regarding mental health issues among people.

1.4 Organization of Report

Section 1 i.e background of the study has been completed. In the upcoming section of the report i.e Section 2, some related works will be presented and evaluated. Moreover, a comparison of existing works with my work will also be shown. Then in Section 3, the data analysis methodology will be discussed explaining each phase and step of the data analysis. Then the findings and the result will be presented in section 4 with the screenshots of code. In section 5, the conclusion of the report will be discussed.

2. Related work

Various related works have been found in the IEEE Xplore digital library. A few of them are presented below.

Related work 1: "Big data application: Study and archival of mental health data, using MongoDB"

The authors of this paper have mostly talked about the application of Big Data in Mental Health. This paper presents plans and suggestions to use data mining methods such as genetic algorithms on data obtained from global mental health statistical data, and deploy this data to a big data platform like MongoDB to retrieve organized data sets that might assist in understanding which treatments work for which patients. They have built it on the requirement for large-scale data storage and information retrieval, and it combines the MongoDB system with the tool's storing capabilities. When a big amount of consumer data is available, a genetic algorithm is a viable method for refining the data using a fitness function. Achievement, storage systems, and processors are the three categories of information saved in an extraction system. This technique suggests that mental disorders can be treated more effectively in very little time and at a lower cost using this approach. This will aid doctors in providing therapy with more precise data, as well as those who are last in receiving proper care due to the greater expense.

(Dhaka & Johari, 2018)

Related work 2: "A survey on big data-driven digital phenotyping of mental health"

This paper offers a general summary of digital phenotyping of mental health (DPMH) from embedded sensor viewpoints in this review article that outlines the goal of DPMH by combining richer data from prominent and popular sensors, social networks, and medical systems. The paper begins by doing a thorough study and proposing a study paradigm that emphasizes the important features of mental health, as well as discussing the issues that richer information for digital phenotyping has prompted.

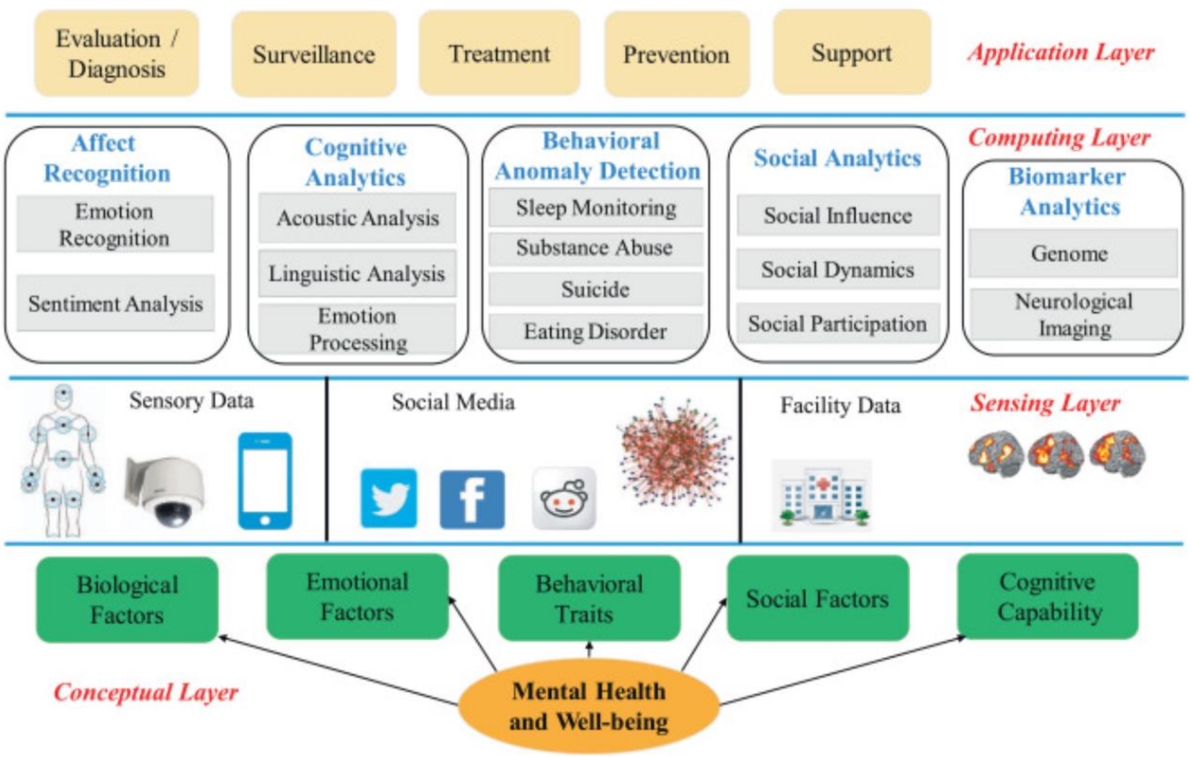


Figure 2: Digital phenotyping framework for Mental Health (D.Zeng, 2019)

As shown in the figure above, social, biological, emotional, behavioral, and cognitive are some of the aspects of mental situations. With transparent sensors and evidentiary mining through large data, digital phenotyping will change the research of mental health and therefore will play a critical role in the therapeutic approach for brain conditions.

This paper expands the scope of digital phenotyping in this study, offers a general model, and highlights essential sensory and computational approaches in the five dimensions of emotion identification as shown in the figure above. With a shortage of basic scientific concepts and restricted research activities, we discovered that digital phenotyping for mental wellbeing is still in its early stages. Despite the fact that several elements connected to mental health have been discovered, neither of them have been used in real diagnosis or therapy. Big data, on the other side, present significant obstacles to digital phenotyping because of its variability, large volume, noise and vibration, and sparseness. (D.Zeng, 2019)

2.1 Difference between related existing works and my work

The related works discussed above have presented excellent ideas and methodologies to improve the sector of Mental Health. However, the content seems a bit vague as it hasn't particularly focused on one field. My work focuses particularly on finding out the state of mental health support within the technology industry and trying to find solutions to enhance the condition of mental health in the workspaces.

3. Methodology

The data analysis is done by following the essential data science steps. They are briefly discussed below with the help of a block diagram:

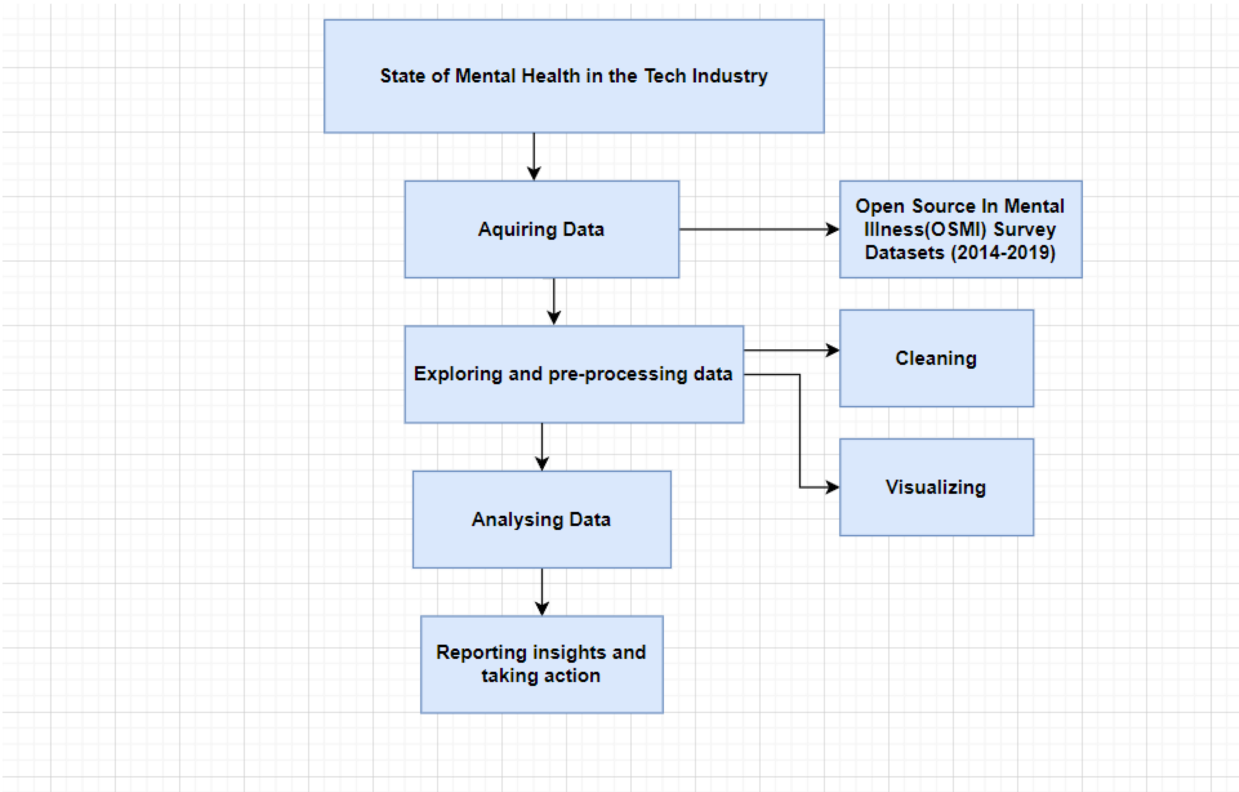


Figure 3: Block Diagram of the Methodology

1. Acquiring the Data

The datasets are taken from the survey done by the organization Open Source Mental Illness (OSMI). OSMI conducts a survey on Mental Health in the tech industry every year to generate a better understanding of how mental health is addressed within tech companies. Data from the years 2014, 2016, 2017, 2018, and 2019 were compiled from individual reports sourced from OSMI. The datasets from 2020 onwards have been omitted due to a low number of respondents. The datasets were acquired from the website of OSMI.

Link to the datasets: https://osmhhelp.org/research

2. Exploring and Pre-processing Data

Each dataset was downloaded and compiled into a unified spreadsheet, which is where they were concatenated and a few of the data cleaning process was also done there. Data visualization has been done to better understand the project by following various data visualization techniques which will be shown in detail in the upcoming sections.

Dataset Observation Definitions:

The headers to the columns in the original dataset featured full questions. Abbreviated versions were required to complete an efficient analysis. The following shows the abbreviated version on the left and the original question on the right.

- Mental health benefits: Does your employer provide mental health benefits as part of healthcare coverage?
- Mental health benefits awareness: Do you know the options for mental health care available under your employer-provided coverage?
- Employer mental health discussion: Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?
- Employer mental health learning resources: Does your employer offer resources to learn more about mental health concerns and options for seeking help?
- Mental health treatment anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?
- Mental health leave accessibility: If a mental health issue prompted you to request a medical leave from work, how easy or difficult would it be to ask for that leave?
- Mental health discussion comfort with coworkers: Would you feel comfortable discussing a mental health issue with your coworkers?

- Mental health discussion comfort with supervisors: Would you feel comfortable discussing a mental health issue with your direct supervisors?

3. **Analyzing Data**

   After acquiring, exploring, and preprocessing the data, the in-depth analysis is done in all of the above-mentioned data science tools and technologies like Machine Learning, MongoDB, Hadoop, Spark, and Kibana. The analysis process will be discussed in detail in the results and finding section with the evidence of the code too.

4. **Reporting Insights and Taking Actions**

   After the in-depth analysis of the data, various insights were retrieved.
   It appears that there is a considerable segment of employees that are unaware if they have access to mental health benefits. If they do have the coverage they are unsure of the programs that are available. Additionally, employees are not confident that their anonymity will be protected if they decide to use a benefit or program to address their mental health. These concerns and lack of awareness could be addressed by conducting an assessment of the current onboarding process, the methods used to communicate, and the frequency of communication. A survey could also be conducted to determine any support, training, or education needed to improve communication between the employer and employee regarding mental health.

5. Results and Discussions

Data analysis is done using various tools and technologies. Each one of them will be discussed in detail with the evidence of the code.

5.1 Using various python libraries

Various python libraries like NumPy, pandas, and matplotlib have been used to perform various data analytics. The code is done in the tool called “Google colab”. Analysis and findings of most of the survey questions (which is the column of our dataset) have been done using this technology.

[2]
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

file = pd.read_csv('Mental Health.csv')
file.head()
```

	year	self_employed	number_of_employees	tech_company	mental_health_benefits	mental_health_benefits_awareness
0	2014	No	100-500	Yes	Yes	Yes
1	2014	No	26-100	Yes	Yes	Yes
2	2014	No	More than 1000	No	Don't know	No
3	2014	No	26-100	Yes	Yes	Not sure
4	2014	No	More than 1000	No	Yes	Yes

Figure 4: Importing necessary libraries

Firstly, necessary libraries have been imported and dataset visualization is done.

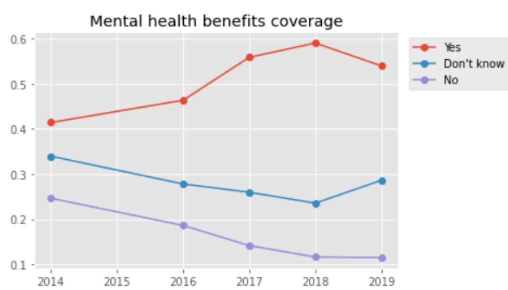
Column 4: Mental health benefits coverage

Survey question of this column: Does your employer provide mental health benefits as part of healthcare coverage?

```
data_mental_health_benefits = file.groupby('year')['mental_health_benefits'].value_counts(normalize=True).unstack()[["Yes", "Don't know",
calc_mean_mental_health_benefits_dont_know = round( data_mental_health_benefits["Don't know"].mean() * 100)
calc_difference_mental_health_benefit_coverage_2019_2014 = round((data_mental_health_benefits.iloc[-1]['Yes'] - data_mental_health_benefits.iloc
calc_average_annual_change = calc_difference_mental_health_benefit_coverage_2019_2014 / 5

plt.style.use('ggplot')
data_mental_health_benefits.plot(marker='o')
plt.title('Mental health benefits coverage')
plt.ylabel("")
plt.xlabel("")
plt.legend(loc='upper right',bbox_to_anchor = (1.32, 1.0))
plt.show()

print("2014 vs 2019 benefit access: {}% increase ".format(calc_difference_mental_health_benefit_coverage_2019_2014))
print("Annual change to employees that have coverage: {}% average".format(calc_average_annual_change))
print("Employees that don't know their coverage: {}% average".format(calc_mean_mental_health_benefits_dont_know))
```




```
2014 vs 2019 benefit access: 13% increase
Annual change to employees that have coverage: 2.6% average
Employees that don't know their coverage: 28% average
```

Figure 5: Analysing Mental health benefits coverage

Over 5 years there was a 13% increase in access to mental health benefit coverage (up from 41% in 2014). However, an average of 28% of tech employees don't know if they have any mental health benefits coverage.

If a large portion of employees remains unaware of their coverage, any progress that is made in employers providing additional support will not be fully utilized.

Column 5. Mental health benefits awareness

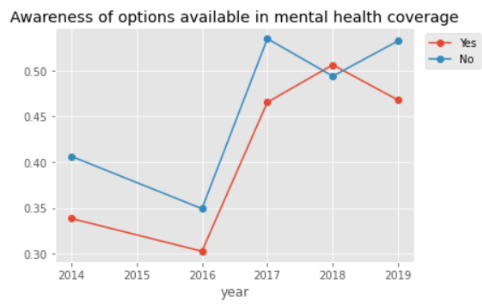
Survey question of this column: Do you know the options for mental health care available under your employer-provided coverage?

```
data_mental_health_benefits_awareness = file.groupby('year')['mental_health_benefits_awareness'].value_counts(normalize=True).unstack()[['Yes','
calc_difference_mental_health_benefits_2019_2014 = round((data_mental_health_benefits_awareness.iloc[-1]['No'] - data_mental_health_benefits_awa
calc_mental_health_benefits_awareness_2019_no = round(data_mental_health_benefits_awareness.iloc[-1]['No'] * 100)

data_mental_health_benefits_awareness.plot(marker='o')
plt.style.use('ggplot')
plt.legend(loc='upper right',bbox_to_anchor = (1.2, 1.0))
plt.title("Awareness of options available in mental health coverage")
plt.show()

print("Employees unaware of benefits or options in their coverage: {}%".format(calc_mental_health_benefits_awareness_2019_no))
print("2014 vs 2019 employees not aware: {}% increase".format(calc_difference_mental_health_benefits_2019_2014))

#data chart
#print(data_mental_health_benefits_awareness)
```




```
Employees unaware of benefits or options in their coverage: 53%
2014 vs 2019 employees not aware: 13% increase
```

Figure 6: Analysing Mental health benefits awareness

In 2019, 53% of employees were unaware of the details or options of their mental health benefits coverage, which is a 13% increase from 2014.

Even if an employee has coverage, they are unlikely to get the support they need if they don't know what options or programs are available.

Column 6: Employer mental health discussion

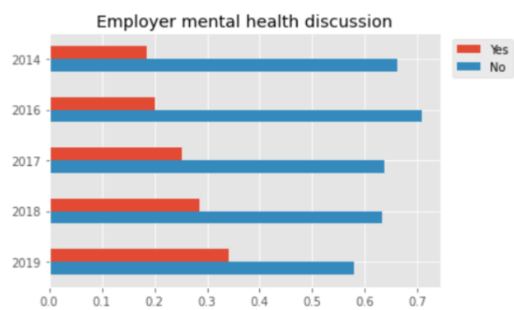
Survey question of this column: Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?

```
[5] data_mental_health_discussion = file.groupby('year')['employer_mental_health_discussion'].value_counts(normalize=True).unstack()[['Yes', 'No']]
    calc_average_annual_improvement_employer_mental_health_discussion = round(((data_mental_health_discussion.iloc[-1]['Yes'] - data_mental_health_c
    calc_employer_mental_health_discussion_2019_no = round(data_mental_health_discussion.iloc[-1]['No'] * 100)

    data_mental_health_discussion.plot(kind='barh').invert_yaxis()
    plt.legend(loc='upper right',bbox_to_anchor = (1.2, 1.0))
    plt.title("Employer mental health discussion")
    plt.ylabel("")
    plt.style.use('ggplot')
    plt.show()

    print("Employers that do not formally discuss mental health in 2019: {}%".format(calc_employer_mental_health_discussion_2019_no))
    print("Average change in employer driven discussion: {}% increase each year.".format(calc_average_annual_improvement_employer_mental_health_disc

    # data chart
    # print(data_mental_health_discussion)
```




```
Employers that do not formally discuss mental health in 2019: 58%
Average change in employer driven discussion: 3% increase each year.
```

Figure 7: Analysing Employer mental health discussion

As of 2019, 58% of employers do not have any formal discussions regarding mental health benefits or support. Over the course of 5 years, employers discussing mental health wellness or support has increased by an average of 3% each year.

If the employers are willing to take additional initiative in initiating and following up on these conversations, awareness of available benefits and support among employees can improve. This is an essential and crucial step as mental health can negatively affect employees in unexpected ways. Constant support, and mitigation can come in the form of providing avenues for conversation, and learning resources.

Column 7: Employer mental health learning resources

Survey question of this column: Does your employer offer resources to learn more about mental health concerns and options for seeking help?

```
data_employer_mental_health_learning_resources = file.groupby('year')['employer_mental_health_learning_resources'].value_counts(normalize=True)
data_employer_mental_health_learning_resources_filtered_yes = data_employer_mental_health_learning_resources[['Yes']]
calc_difference_employer_mental_health_learning_resources_2019_2014 = int(round(((data_employer_mental_health_learning_resources_filtered_yes.i
calc_employer_mental_health_learning_resources_2019_yes = int(round(data_employer_mental_health_learning_resources_filtered_yes.iloc[-1]*100))

data_employer_mental_health_learning_resources_filtered_yes.plot(marker='o', legend= None)
plt.style.use('ggplot')
plt.ylim(0.2,1.0)
plt.xlabel("")
plt.title('Employers that provide learning resources to help address mental health')
plt.show()

print("Employers that provide educational resources in 2019: {}%".format(calc_employer_mental_health_learning_resources_2019_yes))
print("Average change in employers providing resources: {}% average increase per year".format(calc_difference_employer_mental_health_learning_r

#data chart
#print(data_employer_mental_health_learning_resources)
```

Employers that provide learning resources to help address mental health


```
Employers that provide educational resources in 2019: 34%
Average change in employers providing resources: 3% average increase per year
```

Figure 8: Analysing Employer mental health learning resources

In 2019, 34% of employers provided educational resources to help employees gain a better understanding of the available opportunities to address or support their mental health through company benefits. Employers that provide educational resources have increased by an average of 3% each year.

Column 8: Mental health treatment anonymity

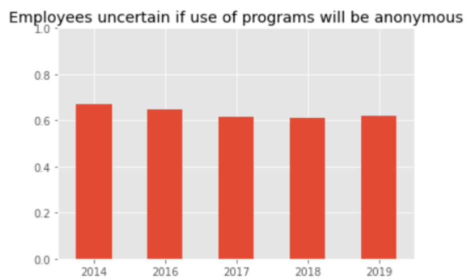
Survey question of this column: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?

```
data_treatment_anonymity = file.groupby('year')['mental_health_treatment_anonymity'].value_counts(normalize=True).ur
data_treatment_anonymity_filtered_dont_know = data_treatment_anonymity["Don't know"]
calc_mean_data_treatment_anonymity_filtered_dont_know = round(data_treatment_anonymity_filtered_dont_know.mean() * 100)

data_treatment_anonymity_filtered_dont_know.plot(kind='bar',legend = None)
plt.style.use('ggplot')
plt.title("Employees uncertain if use of programs will be anonymous")
plt.ylabel("")
plt.xlabel("")
plt.xticks(rotation = 0)
plt.ylim(0,1)
plt.show()

print("Employees that don't know if treatment will remain anonymous: {}% average".format(calc_mean_data_treatment_anonymity_filtered_dont_know))

#data chart
#print(data_treatment_anonymity)
```




```
Employees that don't know if treatment will remain anonymous: 63% average
```

Figure 9: Analysing mental health treatment anonymity

From 2014 to 2019, an average of 63% of employees don't know if their anonymity will be protected if they were to use mental health or substance abuse treatment programs provided by their employer.

It appears that the majority of employees either don't trust or are unaware of the process involved with using benefits or programs for treating mental health or addiction. This could potentially be addressed by developing trust with employees right from the beginning during the onboarding process, and through developing consistent awareness of resources and opportunities to maintain or address mental health. It will be important to clearly walk employees through each step of the process so they can have a clear understanding of what to expect.

Column 9: Mental health discussion comfort with direct supervisor

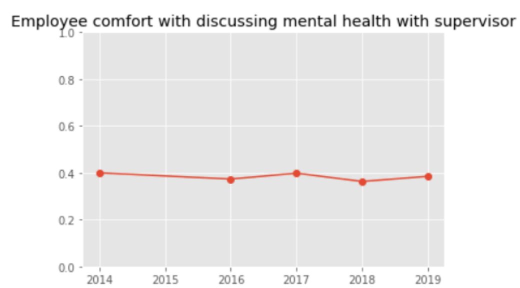
Survey question of this column: Would you feel comfortable discussing a mental health issue with your direct supervisor?

```
data_discussion_comfort_supervisor = file.groupby('year')['mental_health_discussion_comfort_supervisor'].value_counts(normalize=True).unstack()
data_discussion_comfort_supervisor_filtered = data_discussion_comfort_supervisor[['Yes']]
calc_mean_discussion_comfort_supervisor_yes = int(round(data_discussion_comfort_supervisor_filtered.mean() * 100))
calc_difference_discussion_comfort_supervisor = int(round((data_discussion_comfort_supervisor_filtered.iloc[-1] - data_discussion_comfort_superv

data_discussion_comfort_supervisor_filtered.plot(legend=None, marker='o')
plt.style.use('ggplot')
plt.title("Employee comfort with discussing mental health with supervisor")
plt.xlabel("")
plt.ylim(0,1)
plt.show()

print("Employees comfortable discussing mental health with their supervisor: {}% average".format(calc_mean_discussion_comfort_supervisor_yes))
print("2014 vs 2019 comfort level: {}%".format(calc_difference_discussion_comfort_supervisor))

#data chart
#print(data_discussion_comfort_supervisor)
```




```
Employees comfortable discussing mental health with their supervisor: 38% average
2014 vs 2019 comfort level: -1%
```

Figure 10: Analysing Mental health discussion comfort with direct supervisor

An average of 38% of employees are comfortable discussing mental health with their direct supervisor. Employee comfort in 2019 has decreased by 1% when compared to levels in 2014.

5.2 Using MongoDB

```
C:\Users\User>mongoimport --db Mental_health --collection healthData_csv --file "Downloads/Mental_health.csv" --type csv --headerline
2022-05-04T12:57:21.644+0545    connected to: mongodb://localhost/
2022-05-04T12:57:21.729+0545    3483 document(s) imported successfully. 0 document(s) failed to import.

C:\Users\User>mongo
```

Figure 11: Creating Database and collection

Creating a database named “Mental_health” and a collection named “healthData_csv” to import the dataset.

```
> show collections
healthData_csv
> db.healthData_csv.count()
3483
>
```

Figure 12: Counting the total data

Checking the collection which was just created and counting the total number of data in the dataset. Thus, the total number of respondents in the survey is 3483.

```
> db.healthData_csv.findOne()
{
        "_id" : ObjectId("62722755445b40b5f1e55777"),
        "year" : 2014,
        "self_employed" : "No",
        "number_of_employees" : "More than 1000",
        "tech_company" : "No",
        "mental_health_benefits" : "Don't know",
        "mental_health_benefits_awareness" : "No",
        "employer_mental_health_discussion" : "No",
        "employer_mental_health_learning_resources" : "Don't know",
        "mental_health_treatment_anonymity" : "Don't know",
        "mental_health_leave_accessibility" : "Don't know",
        "mental_health_discussion_comfort_coworkers" : "No",
        "mental_health_discussion_comfort_supervisor" : "No"
}
> db.healthData_csv.distinct("tech_company")
[ "No", "Yes" ]
>
```

Figure 13: Visualization of columns

Visualizing all the columns and each of their data and checking the distinct values of the column “tech_company”.

```
> db.healthData_csv.find({"mental_health_benefits":"Don't know"})
{ "_id" : ObjectId("62722755445b40b5f1e55777"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "More than 1000", "tech_company" : "No", "mental_health_benefits" : "Don't know", "mental_health_ben
efits_awareness" : "No", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "Don't know", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessibi
lity" : "Don't know", "mental_health_discussion_comfort_coworkers" : "No", "mental_health_discussion_comfort_supervisor" : "No" }
{ "_id" : ObjectId("62722755445b40b5f1e55779"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "1-5", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_awar
eness" : "Not sure", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "Don't know", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessibility
" : "Don't know", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "No" }
{ "_id" : ObjectId("62722755445b40b5f1e5577c"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "26-100", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_a
wareness" : "Not sure", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "Don't know", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessibil
ity" : "Somewhat difficult", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "No" }
{ "_id" : ObjectId("62722755445b40b5f1e5577f"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "500-1000", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits
_awareness" : "No", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "No", "mental_health_treatment_anonymity" : "Yes", "mental_health_leave_accessibility" : "Somewhat ea
sy", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "Yes" }
{ "_id" : ObjectId("62722755445b40b5f1e5578c"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "26-100", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_a
wareness" : "No", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "No", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessibility" : "Don't
know", "mental_health_discussion_comfort_coworkers" : "No", "mental_health_discussion_comfort_supervisor" : "No" }
{ "_id" : ObjectId("62722755445b40b5f1e5578e"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "26-100", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_a
wareness" : "Not sure", "employer_mental_health_discussion" : "No", "employer_mental_health_learning_resources" : "No", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessibility" : "
Don't know", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "Yes" }
{ "_id" : ObjectId("62722755445b40b5f1e55792"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "100-500", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_
awareness" : "Not sure", "employer_mental_health_discussion" : "Don't know", "employer_mental_health_learning_resources" : "Don't know", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_a
ccessibility" : "Don't know", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "Yes" }
{ "_id" : ObjectId("62722755445b40b5f1e55794"), "year" : 2014, "self_employed" : "No", "number_of_employees" : "100-500", "tech_company" : "Yes", "mental_health_benefits" : "Don't know", "mental_health_benefits_
awareness" : "No", "employer_mental_health_discussion" : "Don't know", "employer_mental_health_learning_resources" : "Don't know", "mental_health_treatment_anonymity" : "Don't know", "mental_health_leave_accessi
bility" : "Don't know", "mental_health_discussion_comfort_coworkers" : "Some of them", "mental_health_discussion_comfort_supervisor" : "Some of them" }
```

Figure 14: Using the find query

Finding the data of employees who answered “don’t know” when they were asked if they were aware of mental health benefits.

5.3 Using Hadoop

```
hdoop@sahaja-VirtualBox:~/hadoop-3.2.2$ cd sbin
hdoop@sahaja-VirtualBox:~/hadoop-3.2.2/sbin$ ls
distribute-exclude.sh  hadoop-daemons.sh  mr-jobhistory-daemon.sh  start-all.sh       start-dfs.sh         start-yarn.sh  stop-balancer.sh  stop-secure-dns.sh  workers.sh
FederationStateStore   httpfs.sh          refresh-namenodes.sh     start-balancer.sh  start-secure-dns.sh  stop-all.cmd   stop-dfs.cmd      stop-yarn.cmd       yarn-daemon.sh
hadoop-daemon.sh       kms.sh             start-all.cmd            start-dfs.cmd      start-yarn.cmd       stop-all.sh    stop-dfs.sh       stop-yarn.sh        yarn-daemons.sh
hdoop@sahaja-VirtualBox:~/hadoop-3.2.2/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [sahaja-VirtualBox]
2022-05-04 17:15:22,955 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers
hdoop@sahaja-VirtualBox:~/hadoop-3.2.2/sbin$
```

Figure 15: Starting the hadoop

First, I have started the hadoop server.

```
hdoop@sahaja-VirtualBox:~$ mkdir mental_health
hdoop@sahaja-VirtualBox:~$ cd mental_health
hdoop@sahaja-VirtualBox:~/mental_health$ cp /home/sahaja/Downloads/Mental_Health.csv
cp: missing destination file operand after '/home/sahaja/Downloads/Mental_Health.csv'
Try 'cp --help' for more information.
hdoop@sahaja-VirtualBox:~/mental_health$ cp /home/sahaja/Downloads/Mental_Health.csv /home/hdoop/mental_health
hdoop@sahaja-VirtualBox:~/mental_health$ ls
Mental_Health.csv
hdoop@sahaja-VirtualBox:~/mental_health$ cp /home/sahaja/Desktop/mentalHealth.java /home/hdoop/mental_health
hdoop@sahaja-VirtualBox:~/mental_health$ ls
Mental_Health.csv  mentalHealth.java
hdoop@sahaja-VirtualBox:~/mental_health$
```

Figure 16: Creating a new directory

Then, as shown in the screenshot above, I have made a new directory named "mental_health" and added the java file and csv file to this directory in the hdoop user from the main user.

```
hdoop@sahaja-VirtualBox:~/mental_health$ javac -classpath $(hadoop classpath) mentalHealth.java
hdoop@sahaja-VirtualBox:~/mental_health$ ls
'mentalHealth$csvReducer.class'  'mentalHealth$PopMapper.class'   mentalHealth.class   Mental_Health.csv   mentalHealth.java
hdoop@sahaja-VirtualBox:~/mental_health$
```

Figure 17: Compiling the java file

In the screenshot above, I have compiled the java file.

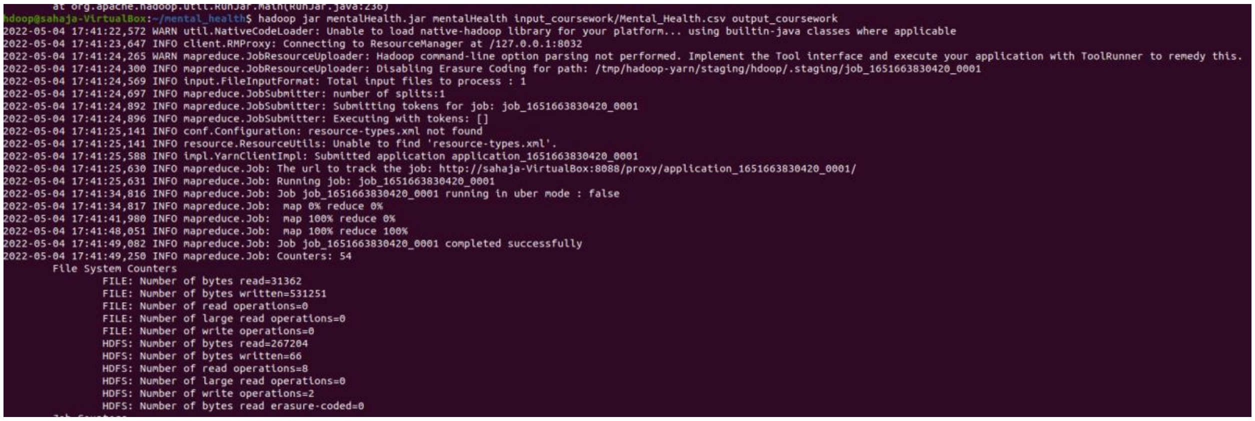


Figure 18: Creating and compiling the jar file

Then, in the screenshot above the jar file was created and compiled.

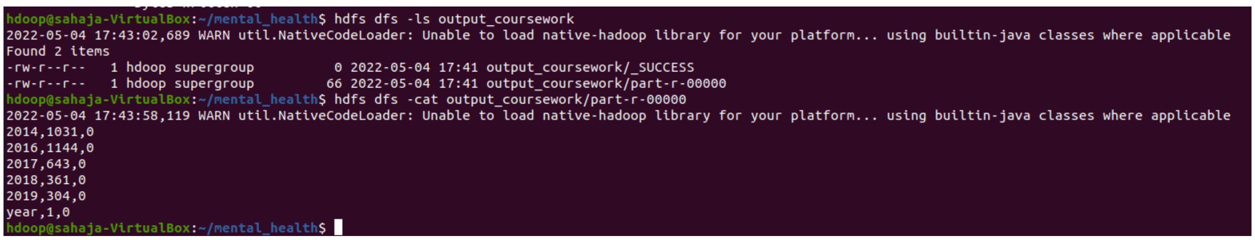


Figure 19: Generating the output

As shown in the figure above, the output has been generated and it can be seen that there are 1031 respondents in the year 2014, 1122 respondents in the year 2016, 643 respondents in the year 2017, 361 respondents in the year 2018, and 304 respondents in the year 2019.

5.4 Using Spark

```
hdoop@sahaja-VirtualBox:/home/sahaja$ spark-shell
22/05/03 23:06:10 WARN Utils: Your hostname, sahaja-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
22/05/03 23:06:10 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
22/05/03 23:06:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/03 23:06:21 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Spark context Web UI available at http://10.0.2.15:4041
Spark context available as 'sc' (master = local[*], app id = local-1651598481448).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.3
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_312)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val dataframe = spark.read.csv("Downloads/Mental_Health.csv")
dataframe: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 10 more fields]

scala>
```

Figure 20: Loading the dataset in the spark-shell

In the screenshot above, I have started the spark-shell and loaded the dataset as a data frame using spark.

```
only showing top 3 rows

scala> dataframe.show(15)
```

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11
year	self_employed	number_of_employees	tech_company	mental_health_ben...	mental_health_ben...	employer_mental_h...	employer_mental_h...	mental_health_tre...	mental_health_lea...	mental_health_dis...	mental_health_dis...
2014	No	100-500	Yes	Yes	Yes	No	No	No	Somewhat difficult	Some of them	Some of them
2014	No	26-100	Yes	Yes	Yes	No	No	Don't know	Don't know	Some of them	Yes
2014	No	More than 1000	No	Don't know	No	No	Don't know	Don't know	Don't know	No	No
2014	No	26-100	Yes	Yes	Not sure	Don't know	Yes	Yes	Don't know	Yes	Yes
2014	No	More than 1000	No	Yes	Yes	No	Don't know	No	Very easy	Some of them	Yes
2014	No	1-5	Yes	Don't know	Not sure	No	Don't know	Don't know	Don't know	Some of them	No
2014	No	6-25	Yes	Yes	Yes	Don't know	Don't know	Don't know	Don't know	Yes	Yes
2014	No	26-100	Yes	Don't know	Not sure	No	Don't know	Don't know	Somewhat difficult	Some of them	No
2014	No	6-25	Yes	No	No	No	No	Don't know	Very difficult	Some of them	No
2014	No	6-25	Yes	No	No	No	No	Don't know	Don't know	Some of them	Some of them
2014	No	500-1000	Yes	Don't know	No	No	No	Yes	Somewhat easy	Some of them	Yes
2014	No	26-100	Yes	Yes	Yes	Yes	Yes	Yes	Very easy	Some of them	Some of them
2014	No	1-5	Yes	Yes	Yes	No	No	Don't know	Don't know	Some of them	No
2014	No	6-25	Yes	No	Yes	No	No	Don't know	Very easy	Yes	Yes

```
only showing top 15 rows
```

Figure 21: Visualizing the dataset

Then, as shown in the screenshot above, I have visualized the first 15 rows of the dataset.

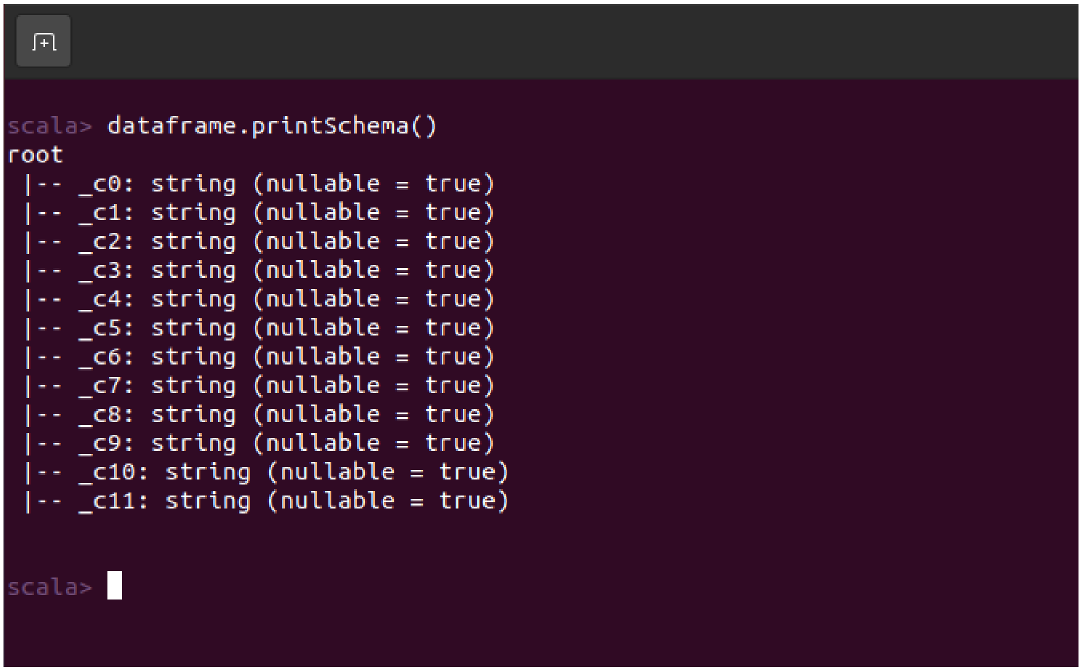


Figure 22: Visualizing the schema

The screenshot above shows the data type of each column.

```
scala> dataframe.select("_c0", "_c3","_c4").show(20)
+----+------------+--------------------+
| _c0|         _c3|                 _c4|
+----+------------+--------------------+
|year|tech_company|mental_health_ben...|
|2014|         Yes|                 Yes|
|2014|         Yes|                 Yes|
|2014|          No|          Don't know|
|2014|         Yes|                 Yes|
|2014|          No|                 Yes|
|2014|         Yes|          Don't know|
|2014|         Yes|                 Yes|
|2014|         Yes|          Don't know|
|2014|         Yes|                  No|
|2014|         Yes|                  No|
|2014|         Yes|          Don't know|
|2014|         Yes|                 Yes|
|2014|         Yes|                 Yes|
|2014|         Yes|                  No|
|2014|         Yes|                  No|
|2014|         Yes|                  No|
|2014|         Yes|                 Yes|
|2014|         Yes|                 Yes|
|2014|         Yes|                  No|
+----+------------+--------------------+
only showing top 20 rows
```

Figure 23: Visualizing the 3 specific columns

The screenshot above visualizes only the year, tech_company and mental_health_benefits column.

```
scala> val dataframe = spark.read.csv("Downloads/Mental_Health.csv")
dataframe: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 10 more fields]

scala> dataframe.filter($"_c0">2016).show()
+----+---+--------------+---+--------------------+---+----------+----------+----------+--------------------+-----+-----+
| _c0|_c1|           _c2|_c3|                 _c4|_c5|       _c6|       _c7|       _c8|                 _c9| _c10| _c11|
+----+---+--------------+---+--------------------+---+----------+----------+----------+--------------------+-----+-----+
|2017| No|       100-500|Yes|                  No|Yes|        No|Don't know|Don't know|          Don't know|  Yes|  Yes|
|2017| No|       100-500|Yes|                 Yes|Yes|        No|        No|Don't know|          Don't know|  Yes|Maybe|
|2017| No|          6-25|Yes|          Don't know| No|Don't know|        No|       Yes|           Difficult|Maybe|  Yes|
|2017| No|More than 1000|Yes|                 Yes|Yes|Don't know|Don't know|       Yes|           Difficult|  Yes|  Yes|
|2017| No|       100-500|Yes|                 Yes| No|        No|Don't know|       Yes|       Somewhat easy|Maybe|Maybe|
|2017| No|          6-25|Yes|                 Yes|Yes|        No|        No|       Yes|           Very easy|   No|  Yes|
|2017| No|        26-100|Yes|                 Yes| No|        No|        No|Don't know|       Somewhat easy|Maybe|  Yes|
|2017| No|       100-500| No|          Don't know| No|        No|        No|       Yes|           Very easy|Maybe|Maybe|
|2017| No|       100-500| No|                 Yes|Yes|        No|        No|Don't know|          Don't know|   No|   No|
|2017| No|       100-500|Yes|          Don't know| No|        No|Don't know|Don't know|           Difficult|   No|Maybe|
|2017| No|More than 1000|Yes|                  No| No|        No|        No|Don't know|Neither easy nor ...|  Yes|Maybe|
|2017| No|More than 1000|Yes|          Don't know| No|        No|Don't know|Don't know|Neither easy nor ...|Maybe|Maybe|
|2017| No|        26-100|Yes|          Don't know|N/A|Don't know|Don't know|       Yes|           Very easy|  Yes|Maybe|
|2017| No|       100-500| No|                 Yes| No|        No|Don't know|Don't know|       Somewhat easy|Maybe|  Yes|
|2017| No|        26-100| No|                  No| No|        No|        No|       Yes|Neither easy nor ...|  Yes|  Yes|
|2017| No|          6-25|Yes|          Don't know| No|        No|        No|Don't know|       Somewhat easy|Maybe|Maybe|
|2017| No|More than 1000| No|          Don't know| No|        No|Don't know|Don't know|Neither easy nor ...|Maybe|Maybe|
|2017| No|          6-25|Yes|                  No| No|        No|        No|Don't know|Neither easy nor ...|   No|   No|
|2017| No|       100-500| No|Not eligible for ...|N/A|       Yes|       Yes|       Yes|  Somewhat difficult|  Yes|  Yes|
|2017| No|        26-100| No|          Don't know| No|        No|        No|Don't know|           Difficult|  Yes|  Yes|
+----+---+--------------+---+--------------------+---+----------+----------+----------+--------------------+-----+-----+
only showing top 20 rows
```

Figure 24: Using SQL query in Spark

In the screenshot above, we have manipulated the CSV file using SQL query to show the values of the year greater than 2016.

```
scala> dataframe.groupBy("_c0", "_c4").count().show()
+----+-------------------+-----+
| _c0|                _c4|count|
+----+-------------------+-----+
|year|mental_health_ben...|    1|
|2019|                Yes|  164|
|2016|         Don't know|  318|
|2019|                 No|   35|
|2018|         Don't know|   85|
|2017|                Yes|  359|
|2018|Not eligible for ...|   21|
|2018|                 No|   42|
|2019|         Don't know|   87|
|2016|Not eligible for ...|   83|
|2018|                Yes|  213|
|2019|Not eligible for ...|   18|
|2017|         Don't know|  167|
|2014|                 No|  254|
|2014|                Yes|  427|
|2016|                 No|  213|
|2017|                 No|   91|
|2014|         Don't know|  350|
|2017|Not eligible for ...|   26|
|2016|                Yes|  530|
+----+-------------------+-----+
```

Figure 25: Using the groupBy query

In the above screenshot, we have found the count of people who answered “yes”, “No”, “Don’t know”, and “not eligible” in each year when asked if they are aware of mental health benefits in their workplace.

In 2014, 427 people answered “Yes”, 254 people answered “No” and 350 people answered, “Don’t know”.

In 2016, 530 people answered “Yes”, 213 people answered “No”, 318 people answered “Don’t know” and 83 people answered “Not eligible”.

In 2017, 359 people answered “Yes”, 91 people answered “No”, 167 people answered “Don’t know” and 26 people answered “Not eligible”.
In 2018, 213 people answered “Yes”, 42 people answered “No”, 85 people answered “Don’t know” and 21 people answered “Not eligible”.

In 2019, 164 people answered “Yes”, 35 people answered “No”, 87 people answered “Don’t know” and 18 people answered “Not eligible”.

5.5 Using Kibana

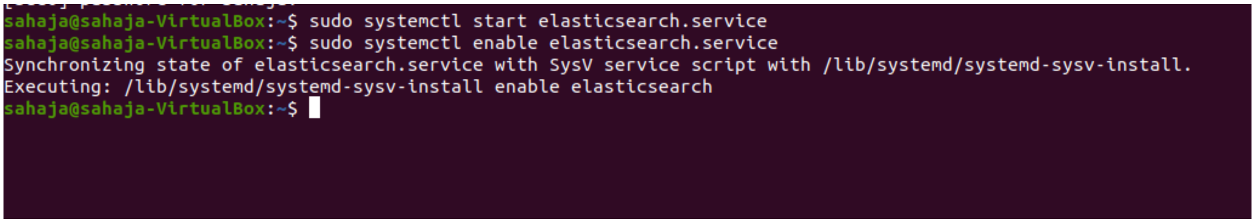


Figure 26: Starting elastic search

In the screenshot above, I have started the Elastic Search.

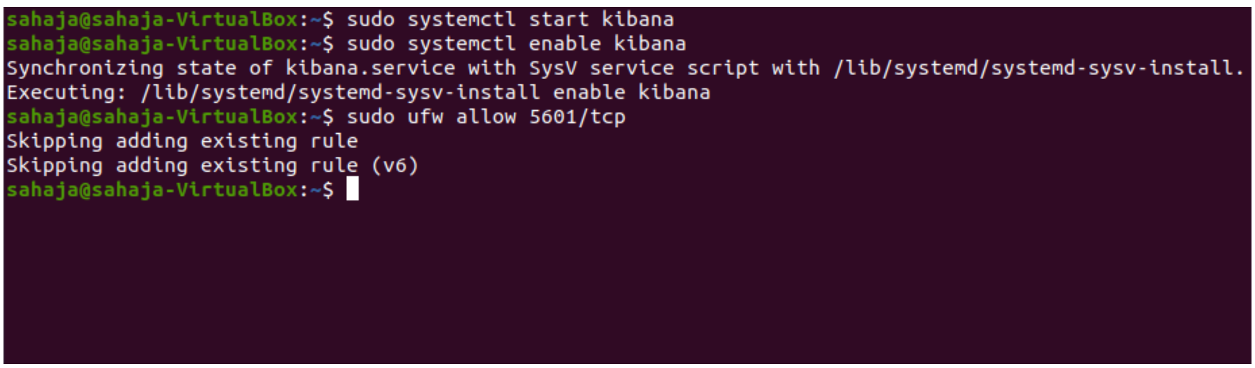


Figure 27: Starting Kibana

Now, I have started the Kibana at port 5601 as shown in the screenshot above.

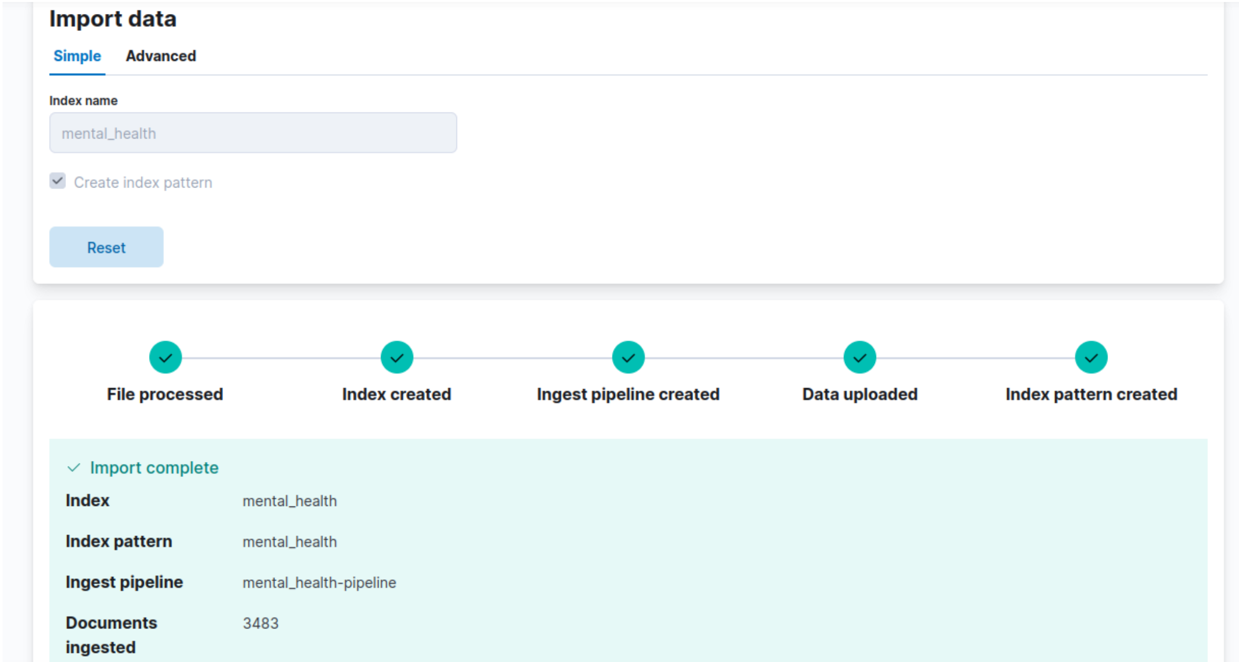


Figure 28: Importing dataset to Kibana

Here, I have imported the dataset to Kibana.

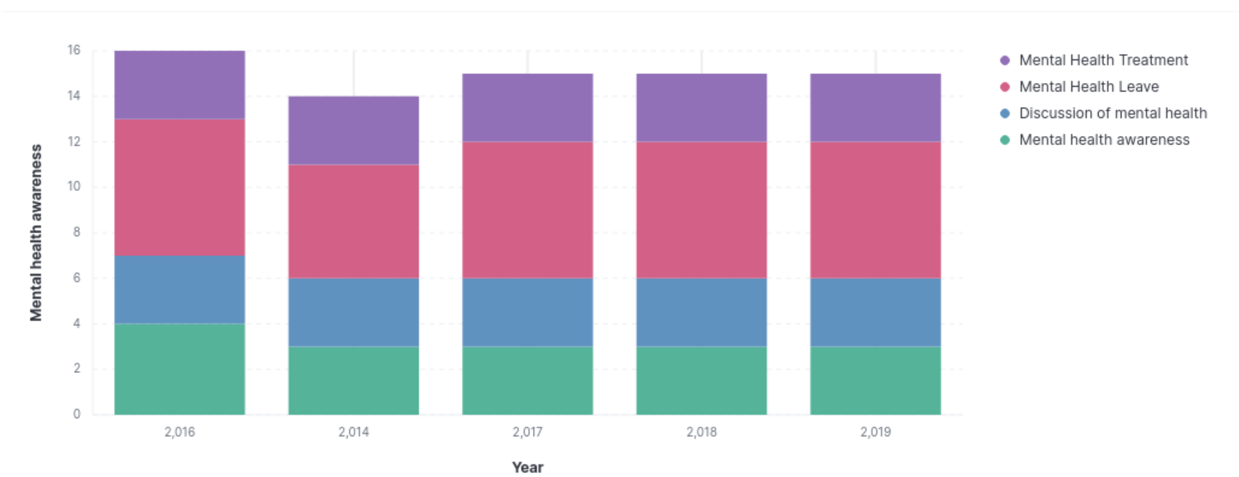


Figure 29: Vertically stacked bar chart

The figure above shows the vertically stacked bar chart which shows the number of employees who are aware of mental health benefits, who have access to mental health treatment, who are comfortable discussing mental health with their coworkers, and who have access to mental health leave in their workplace in each year.

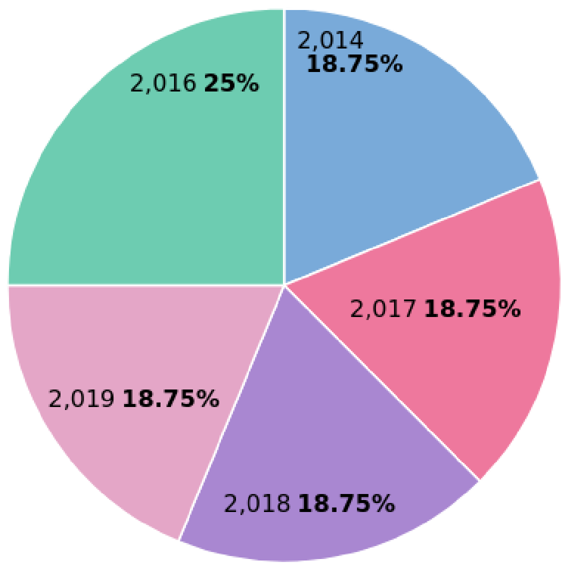


Figure 30: Pie Chart

The figure above is the pie chart that shows the percentage of employees each year who are aware of mental health benefits. 2016 has the highest percentage i.e 25%, so we can say that the awareness regarding mental health was at its peak in 2016 compared to other years.

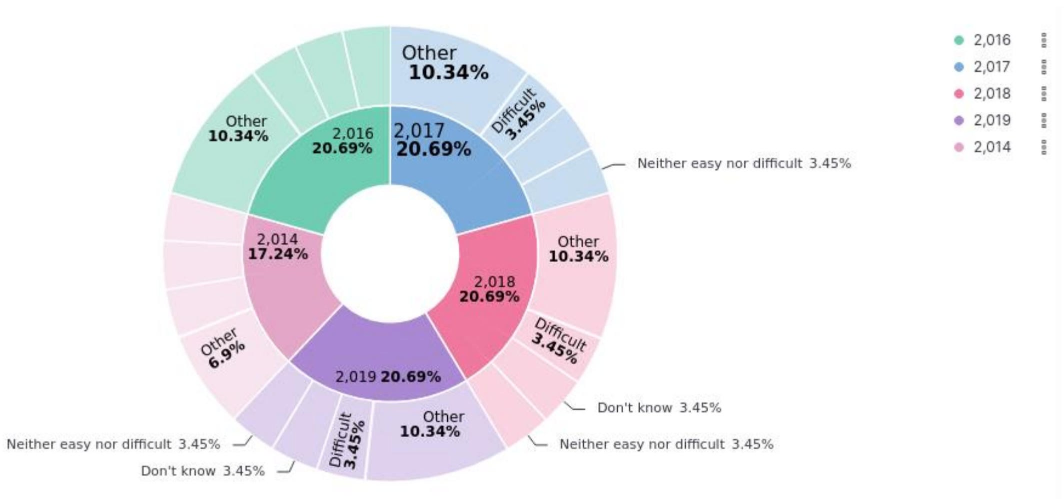


Figure 31: Donut Chart

The figure above is the donut chart which shows the number of employees have access to mental health leave. It seems that most of the employees find it difficult to get a mental health leave from their work.

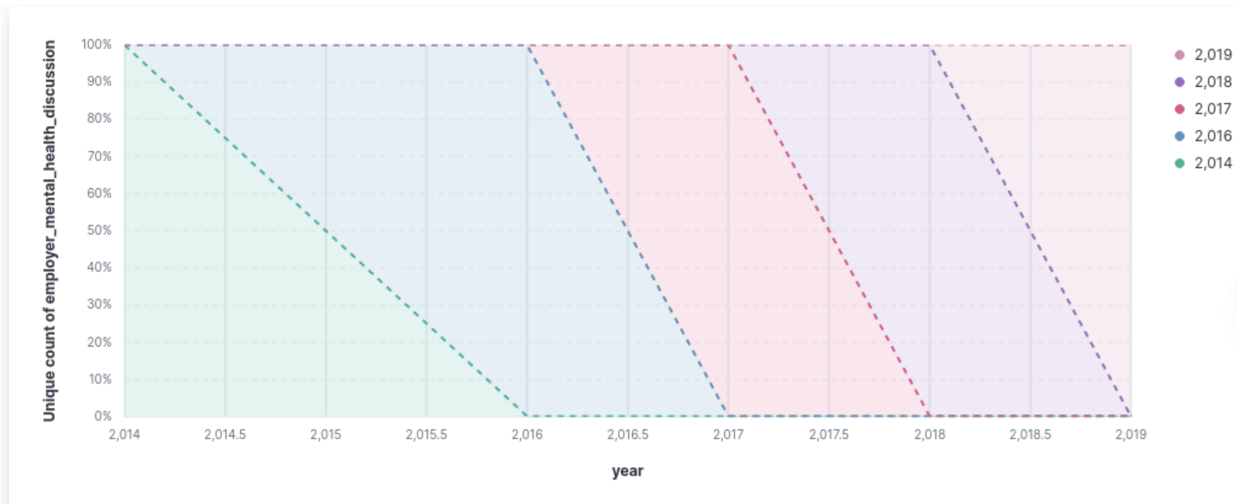


Figure 32: Line Graph

The figure above is the line graph of area percentage which shows the percentage of employees who can discuss mental health issues with their coworkers each year. It seems that the year 2016 has the highest percentage of all the other years.

5.6 Result and Discussion Summary

From the years 2014 to 2019, employees that have access to mental health benefits have increased by an average of 2.6% per year (41% to 54%). However, an average of 28% of tech employees don't know if they have any mental health benefits coverage. In 2019, 53% of employees were unaware of the details or options of their mental health benefits coverage, which is a 13% increase from 2014. As of 2019, 58% of employers do not have any formal discussions regarding mental health benefits or support. Over the course of 5 years, employers discussing mental health wellness or support has increased by an average of only 3% each year. In 2019, 34% of employers provided educational resources to help employees gain a better understanding of the available opportunities to address or support their mental health through company benefits. Employers that provide educational resources have increased by an average of 3% each year. From 2014 to 2019, an average of 63% of employees don't know if their anonymity will be protected if they were to use mental health or substance abuse treatment programs provided by their employer. An average of 38% of employees are comfortable discussing mental health with their direct supervisor. Compared to comfort levels in 2014, results from 2019 reported a decrease of 1%.

6. Conclusion

To sum up, successful data analysis has been done to examine the state of mental health support for employees within the tech industry using the survey datasets from the year 2014 to 2019. The datasets are taken from the survey conducted by Open Source Mental Illness (OSMI), an organization that conducts annual surveys to generate a better understanding of how mental health is addressed within tech companies. The analysis has been done by following the essential data science steps i.e acquiring data, exploring and preprocessing data, analyzing data, and finally reporting insights and taking action. These steps have been completed using the tools and technologies like python libraries, MongoDB, Hadoop, Spark, and Kibana. The annual percentage of participants that responded with 'Yes', 'Don't know", and 'No' to mental health-related questions have been examined. The increase or decrease in the percentage of perceived employers regarding mental health support from year to year and the number of employees that can confidently respond 'Yes' when asked regarding mental health support have been measured.

It appears that there is a considerable segment of employees that are unaware if they have access to mental health benefits. If they do have the coverage, they are unsure of the programs that are available. Additionally, employees are not confident that their anonymity will be protected if they decide to use a benefit or program to address their mental health. These concerns and lack of awareness could be addressed by conducting an assessment of the current onboarding process, the methods used to communicate, and the frequency of communication. A survey could also be conducted to determine any support, training, or education needed to improve communication between the employer and employee regarding mental health. In order to dramatically increase the number of employers that provide mental health benefits, providing data and evidence that supports potential incentives for both the employer and employee, could drive adoption. This could be accomplished by designing and implementing new research projects with the goal of determining if mental health benefits have a direct impact on employee retention, creativity, or productivity.

References

Ayesha Kamran Ul haq, A. K. N. J. M. A. N. F. M., 2020. Data Analytics in Mental Healthcare. p. 9.

D.Zeng, Y. L. X. Z. D., 2019. A survey on big data-driven digital phenotyping of mental health. *Information Fusion,* pp. 290-307.

Davis, R. S. &. K., 2018. 'Big data' in mental health research: current status and emerging possibilities. *SpringerLink.*

Dhaka, P. & Johari, R., 2018. Big data application: Study and archival of mental health data, using MongoDB.

Global Health Data Exchange, 2022. [Online]
Available at: https://ghdx.healthdata.org/gbd-results-tool

Q. Liu and X. Liao, 2021. *Research on University Mental Health Education Based on Computer Big Data Statistical Analysis.* s.l., 2nd International Conference on Big Data and Informatization Education (ICBDIE).

Ting, G. H. a. W., 2020. *Analysis of Community Mental Health Services in the Context of Big Data.* s.l., International Conference on Big Data and Social Sciences.