Project Description:

The project that I studied is mental health in the tech area. The main focus point is the importance's level of mental health in the employers' perspective. Whether they include mental health benefits as part of healthcare coverage? Do they provide any resources about mental health? Or Do they willing to talk about employees' mental health conditions? Since mental health is a vital part of our daily lives and work, it can impact our thoughts, behaviours, and emotions. It is essential and necessary to keep mental health and to promote effectiveness. Especially in the tech area, people are highly stressed, and mental health becomes even more critical. In this project, data scientists are the people who are telling a story to audiences. Data scientists need to clean the data first, remove N/A values, find and process outliers, and transfer the data into useable type based on the data collected from the past five years of questionnaires. Then, they need to use visualizing tools to visualize the data and tell audiences what problems happened. Finally, they may make predictions by building prediction models and tell audiences what may happen next.

Data:

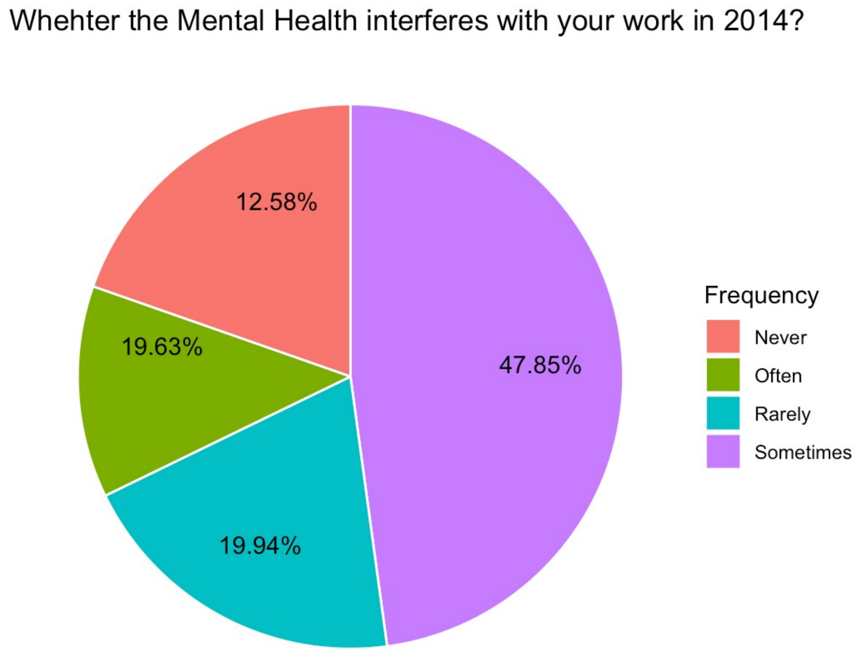
In this project, data comes from the Open Source Mental Illness (OSMI). They did surveys from 2014 to 2019, and each survey towards the mental health problem in the tech workplace.

The data's total size is 4218 objects and approximately 80 variables, which means the data volume is enormous. Hence, when I used the R Programming read_csv function to read the data, some data will be lost because of duplicates or other reasons. Also, the data's velocity is uncertain since the questionnaires changed every year, and participants are different, which means the data change varies for each year. And the data itself has problems as well. Because the data comes from questionnaires, which means they are object answers and only represent their personal view. And all the data are characters, so they cannot be quantified. Therefore, when I tried to build a prediction model to process data, I could use the models very limited. Finally, there are different forms of data, which are called data variety. In each questionnaire, there are multiple-choice questions, open questions, rating questions, etc. Hence, when I processed the data, I need to remove useless answers first by using R. Then, I need to make sure answers are consistent; for example, "I don't know
"equals "don't know."

The analysis tool I used in this project is R studio. Because R studio can clean, manipulate, and visualise data very quickly. It also performs perfectly for statistical computing. However, there are limited tools built with or connect to R when comparing to Python.

Business Analysis:

This project sits in the tech area, which is now the rapidly developed area. People have to keep productive and creative in the tech area because they are making deals with technologies. However, high demands and strict deadlines increase employees' stress levels, which causes the mental health problem to become more and more critical. According to (team, 2019), 39% of people said there is no mental health support in their workplace. We can also see the pie chart about whether the Mental Health interferes with your work based on the 2014 questionnaire. 19.63% of people said often, and 47.85% of people said sometimes, which means mental health exactly affects people's work in real life. I used a pie chart to show the result because it can easily show each part's percentage.



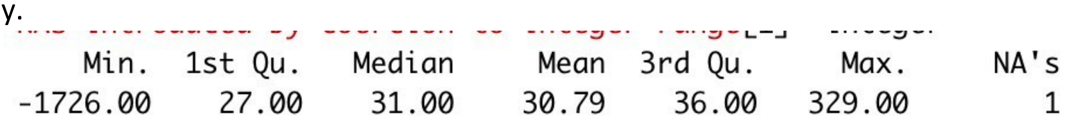
And the challenge of this project is we can only get object answers from participants, which cannot represent the whole area. Some participants also may not take the survey seriously, so we cannot get precisely accurate data. Moreover, there is no data in 2015, which means the data is not complete.

This project's primary value is to let audiences know the current condition of the tech area's mental health problem. And how employers treat mental health problems on employers. Because many tech companies only focus on employees' work and physical health but ignore their mental health.

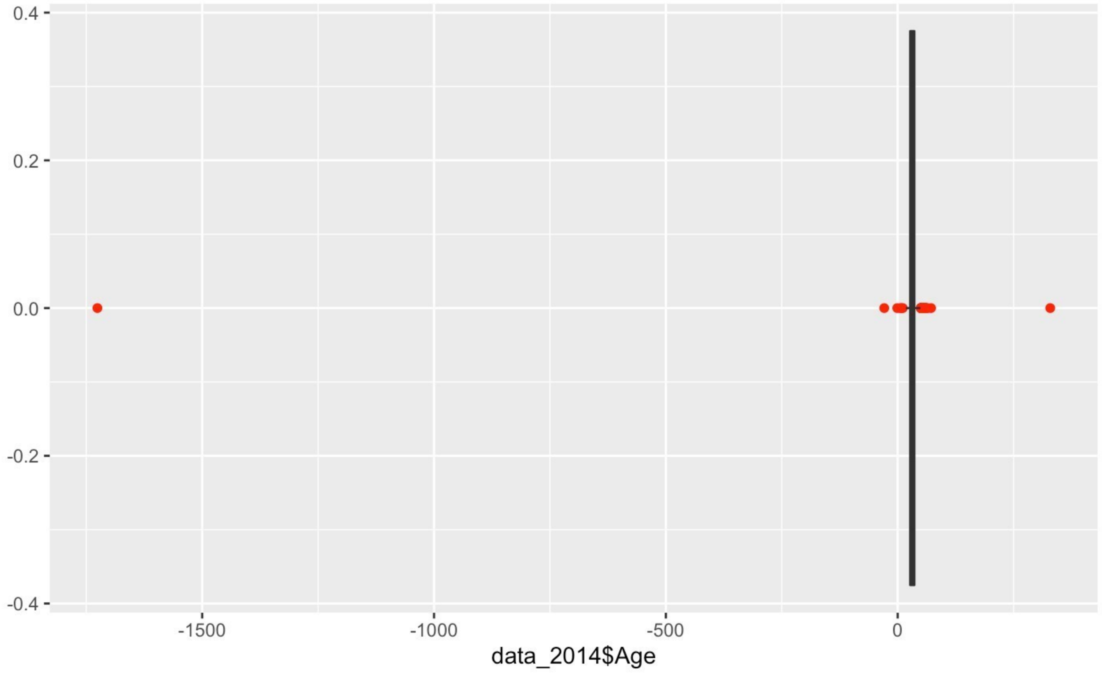
After discussing the importance of mental health in this project, they can realize and create a better work environment, and more mental health problems could be treated as well. Furthermore, in this project, there are two data curation issues. First, data quality cannot maintain in each year. For example, in 2014, there are 1260 objects, and data quality was high. But in 2019, there are only 352 objects, and data quality was low. Also, there are too many useless values and N/A values in the data, which means the data management team may not realize its problems. So, we need to clean and add more valuable value to perform better analysis.

Data Analysis:

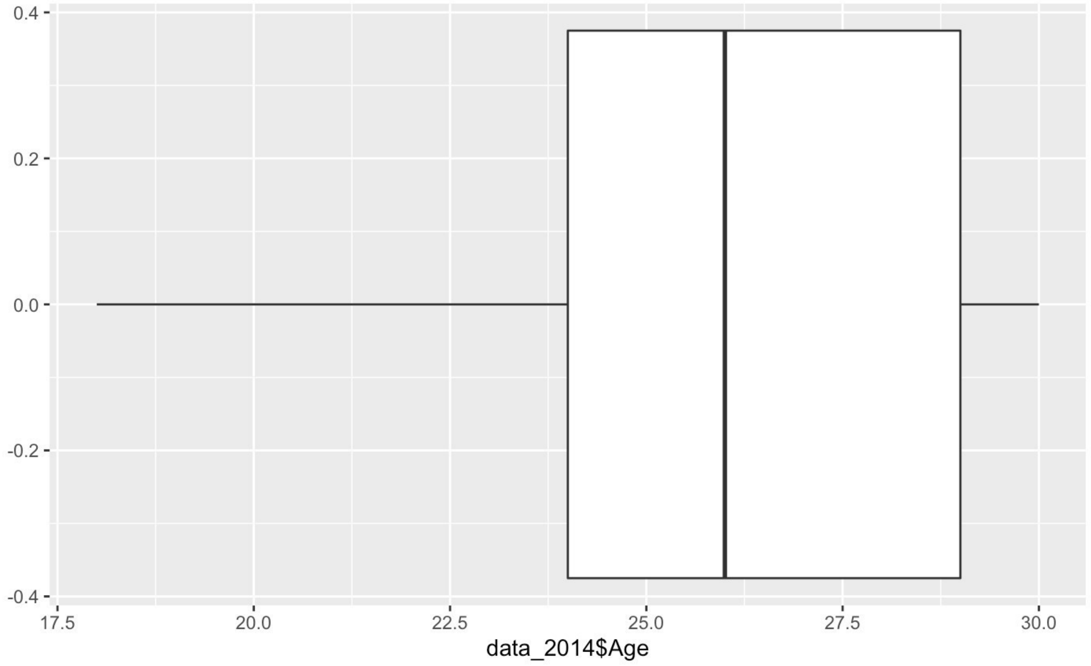
At first, I need to clean the data. In 2014, there are some outliers in the Age column. Here we can see. The minimum age is -1726, and the maximum value is 329, which is very unreasonable. Therefore, I used a boxplot to show the outliers, which is straightforward and easy.



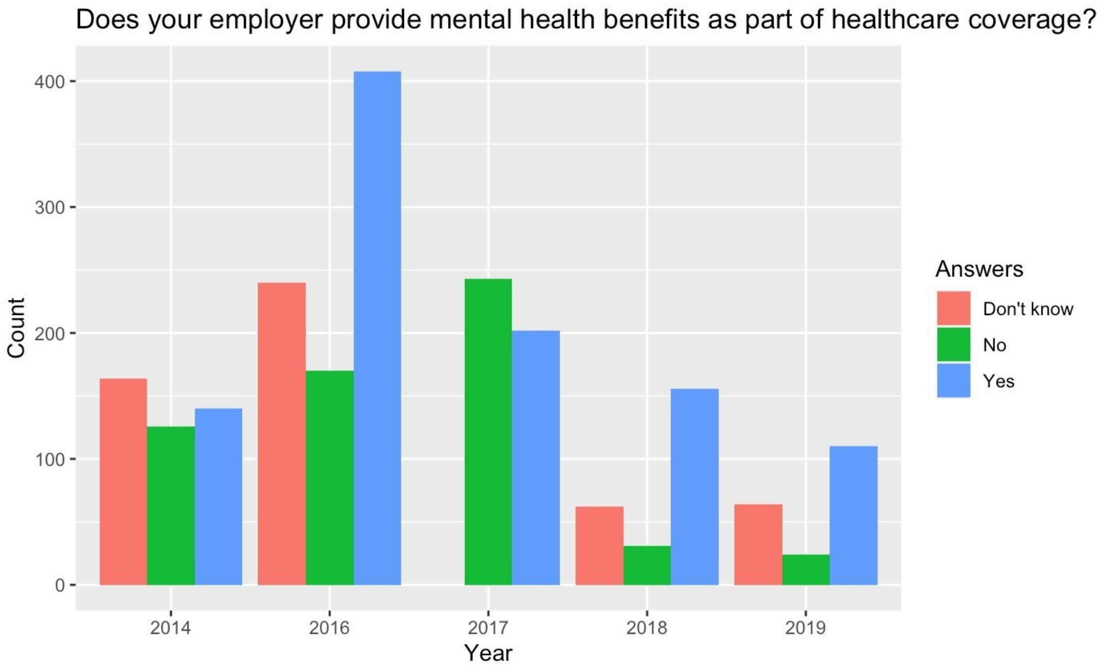
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
-1726.00    27.00    31.00    30.79    36.00   329.00        1 
```



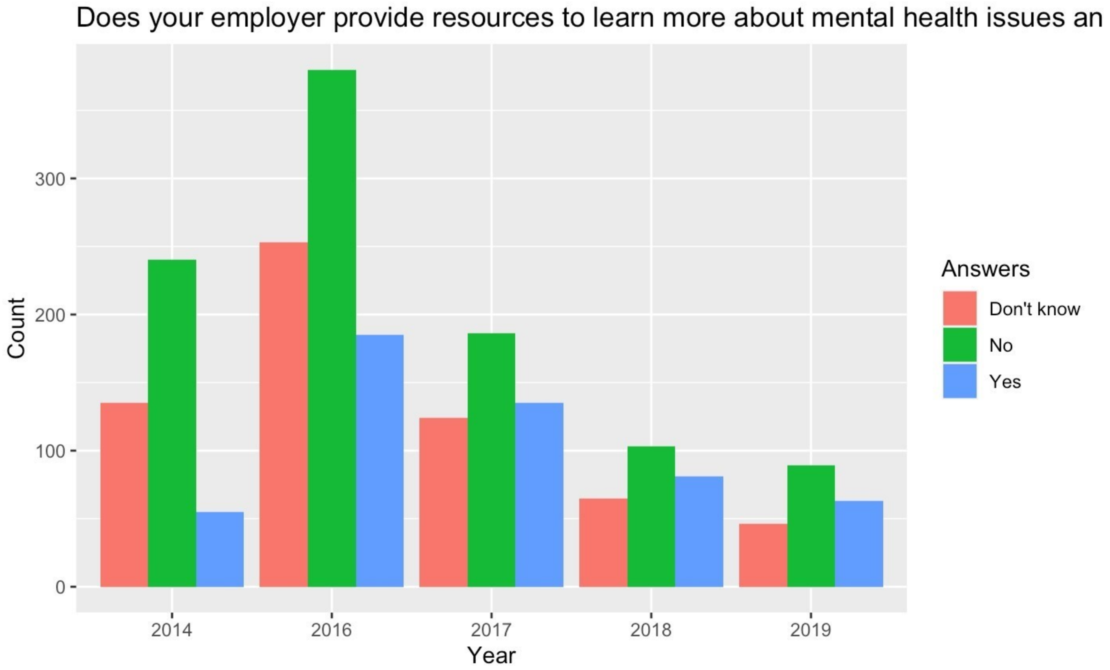
Now, we can see some outliers exist, and they remarkably affect mean value. Hence, I removed age, which is smaller than 15 and greater than 50. Because I only consider people in the age group between 15 and 50, they are closer to mental health in tech areas. Here is the plot after cleaning.



And now, we can see after 2016, and there is a sharp decrease in the answer "Yes." In 2016, the answer "Yes" is very high because it has a large sample size than other years. However, we still can see fewer and fewer employees responded "yes" from 2017 to 2019. I used a bar plot to show the tendency since, in real life, many companies only provide physical health benefits in healthcare coverage but ignore mental health. Physical health can directly affect people's work; mental health is usually a long-term problem that is hard to treat. Therefore, employers should provide health benefits as part of healthcare coverage.



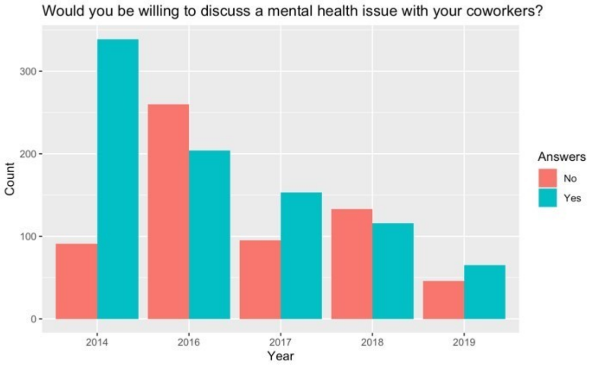
Next, we can see whether many employers would be willing to provide resources to learn more about mental health and how to ask for help.

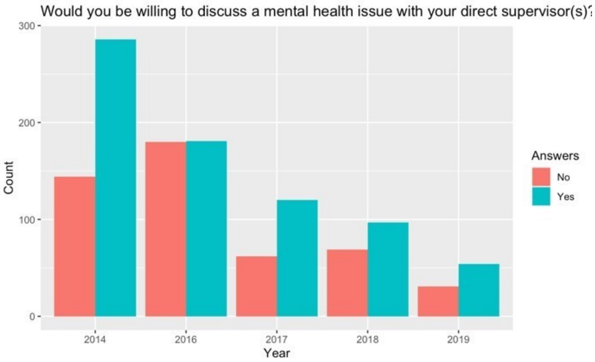


From the plot, we can see that answer "No" is the dominant role, which means employees are not focused on the mental health problem. However, the excellent point is that the difference between "Yes" and "No" decreases from 2014 to 2019, which shows employers start to take care of employees' mental health problems. Another good thing is that answer

"Don't Know" decreased after 2016. It means employees get to know their companies more than before. Moreover, the bar plot is still an excellent method to make a comparison between the five years.

The following two plots show whether people are willing to discuss a mental health issue with co-workers or direct supervisors. At first, we can see that answer "Yes" is the dominant role, which means people are not shy to discuss mental health with co-workers or supervisors. However, the answer "Yes" trend decreases, showing that the discussion's willingness level declined in the past five years. Overall, the count of discussing with coworkers is higher than discussing with supervisors. It represents that people feel more comfortable discussing with peers, but they may feel stressed when talking with supervisors.

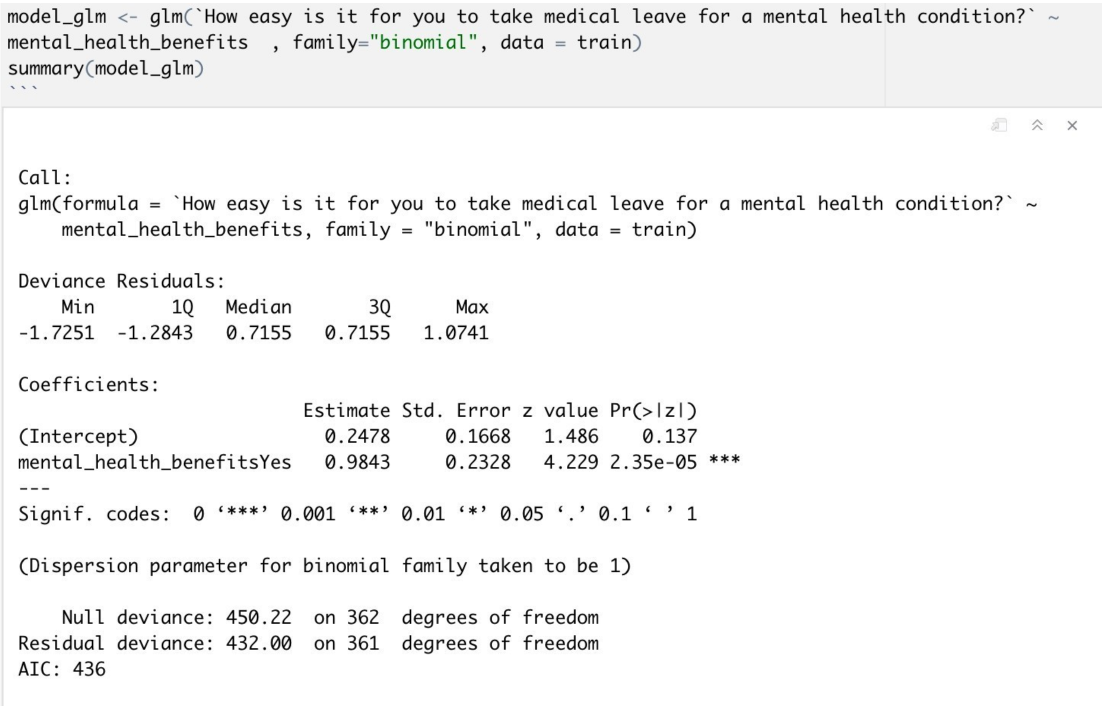




Finally, I built a prediction model to predict how easy to take medical leave for a mental health condition. The relationship factor that I chose is mental health benefits, representing the importance of employers treating mental health problems. And the model I chose is the character model since the data value I have are all characters, only Boolean values. At first, I split data into training and test, and the split ratio is 0.7.

```
[1] 363   6
[1] 154   6
```

Now, I get 363 training data and 154 test data. Then, I used the glm function to build a model that relates to mental health benefits. Here is the output that I got.



```
model_glm <- glm(`How easy is it for you to take medical leave for a mental health condition?` ~
mental_health_benefits  , family="binomial", data = train)
summary(model_glm)
```
```

```
Call:
glm(formula = `How easy is it for you to take medical leave for a mental health condition?` ~
    mental_health_benefits, family = "binomial", data = train)

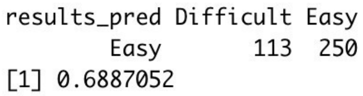
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7251  -1.2843   0.7155   0.7155   1.0741

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.2478     0.1668   1.486    0.137
mental_health_benefitsYes   0.9843     0.2328   4.229 2.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

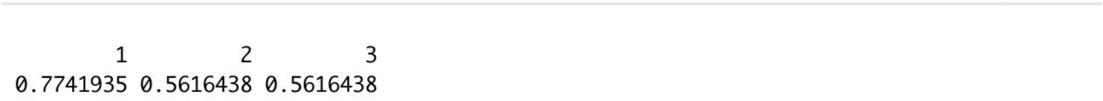
    Null deviance: 450.22  on 362  degrees of freedom
Residual deviance: 432.00  on 361  degrees of freedom
AIC: 436
```

Now I need to check this model's accuracy rate since the accuracy rate is low, which means the model is not accurate., which means the model is not accurate. The accuracy rate I got for this model is approximately 68%, which is acceptable in this project. Therefore, the model is suitable and correct.



```
results_pred Difficult Easy
        Easy       113  250
[1] 0.6887052
```

Next, I will make the prediction using this model. I assign new values to mental health, which are "Yes"," No" and "No". Then, I got the results below:



```
        1         2         3
0.7741935 0.5616438 0.5616438
```

If employers provide mental health benefits as part of healthcare coverage, there is a 77.42% chance to take medical leave for a mental health condition easily. If they don't provide mental health benefits, then only approximately 56% chance that they can take medical leave for a mental health condition.

Based on the data analysis I did before, we can suggest tech companies include mental health benefits in their cultures. The tech area is a high competition area; each tech company has to create its own culture to build a profitable and attractive reputation. Since more people worry about their mental health, a company that has a good reputation for mental health benefits may become more attractive.

Conclusion:
```

In this project, I mainly discussed the mental health problem in the tech workplace. I used data collected from Open Source Mental Illness (OSMI) organization and R Programming to manipulate and visualise the data. The problem we can see is that employers do not treat mental health as important as physical health. Most employers do not provide mental health as part of healthcare coverage, and they do not provide resources about mental health. Through this project, we could suggest to employers that mental health is significant for employees and that building a comfortable workplace is also essential for a good company.

Appendix

Bibliography

team, m. (2019, Oct). *Mental Health in the Tech Industry*. Retrieved from Modis: https://www.modis.com/en-au/news-and-insights/articles/mental-health-in-thetech-industry/

YouTube video link:

https://youtu.be/cQBRSMlyLvk