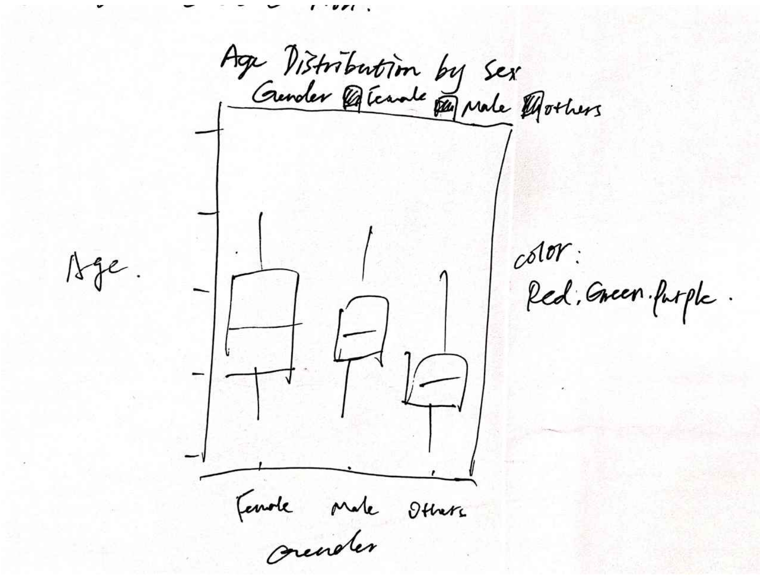
Data Visualization Process Book

Prevalence of Mental Health

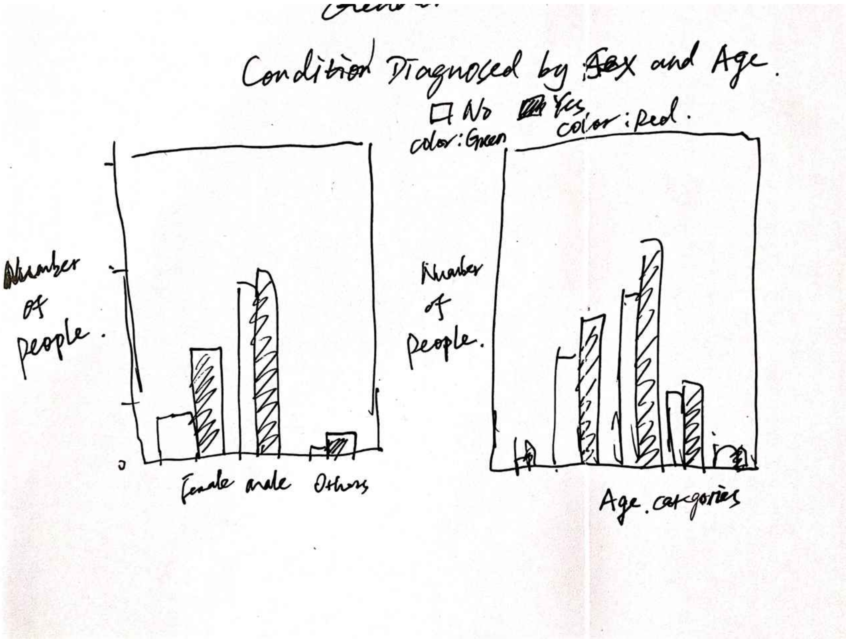
"Age Distribution by Sex"

First, we processed some data about gender information. Since our dataset is a survey, answers of the question "What is your gender?" are inconsistent. We classified those answers like Female, female, Woman ,etc. into "Female" category, and those answers like Male, and Man(regardless of upper and lower case) into "Male" category, and other types of answers such as Agender, cis gender in to "Others" category. And then with the information provided by the column "What is your age?", we made a box plot with "Age" on y-axis and "Gender" on x-axis for age distribution by sex to get a basic sense of age range of the industry.



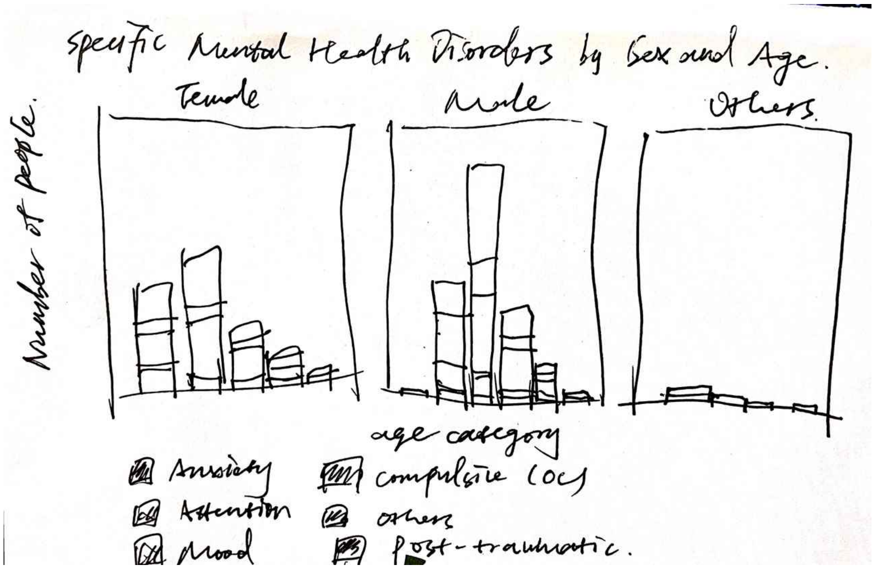
"Condition Diagnosed by sex and age"

We made two bar charts with "Number of People" on y-axis and "Gender"/"age-category" on x-axis to visualize the answer of the column "Have you been diagnosed with a mental health condition by a mental health professional?" We divided gender into female, male and others, and age into 6 sections to show different performances.



"Specific Mental Health Disorders by Sex and Age"

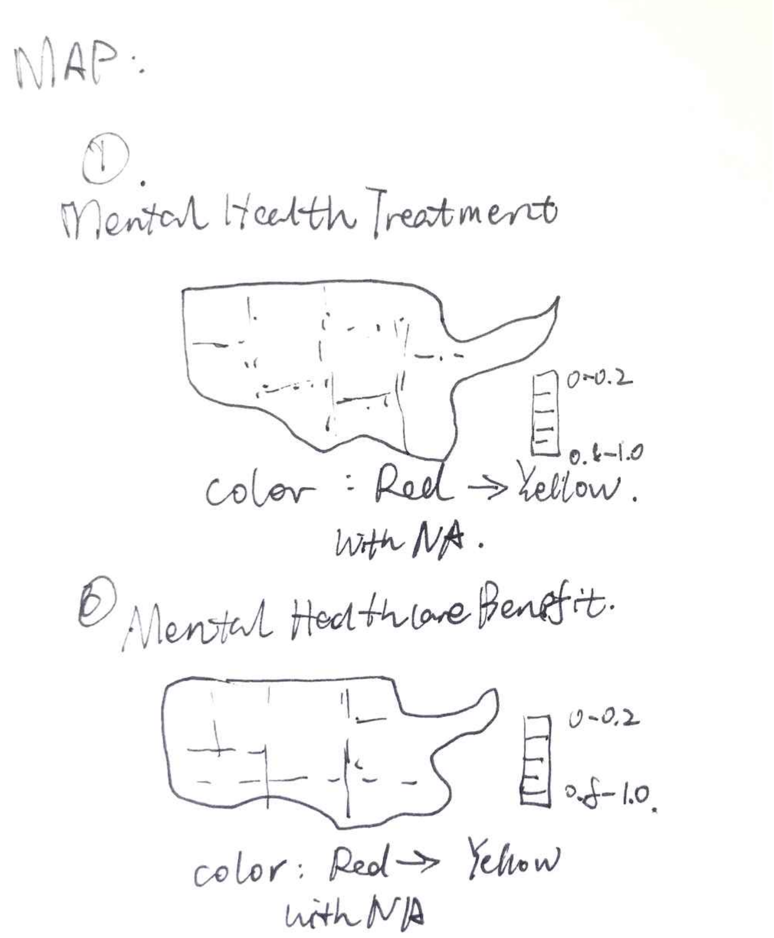
At the beginning, we visualized the answers for the question "If yes, what conditions have you been diagnosed with?" and we got 25 types of mental health disorders responded. We made them into a stacked bar chart but we discovered that some disorders had very few counts and some of them were basically similar conditions, which we thought the information provided by them were limited. In order to fix that, we identified 5 major categories of mental health disorders(counts>20) which are Anxiety Disorder(Generalized, Social, Phobia, etc.), Attention Deficit Hyperactivity Disorder, Mood Disorder(Depression, Bipolar Disorder, etc), Obsessive-Compulsive Disorder, and Post-traumatic Stress Disorder. And we put all other answers into "others". We had 3 charts separately for each gender category to show specific mental health disorders diagnosed, with number of people on y-axis and age sections on x-axis.



Maps

"Frequency of Employee Who Sought Mental Health Treatment" & "Frequency of Employer with Mental Health Benefit"

At first, we actually wanted to visualize the count of employees for this plot. However, we then noticed that the distribution of survey participants for each state is not quite even. Therefore we transformed it to the percentage/frequency analysis. To make the plot more readable, all decimals are rounded. The shape file of the States is downloaded from www.census.gov. To make the plots interactive, the leaflet package is used. We made 5 bins, broken at 0, 0.2, 0.4, 0.6, 0.8, 1, to better show the data distribution. The colors for bins are also arranged, we managed to show the color from lighter yellow to darker red indicating the frequency increase accordingly. Legend title is added to show the purpose clearly. To make the analysis look more consistent and comparable, we used the same template for the 2 map plots.



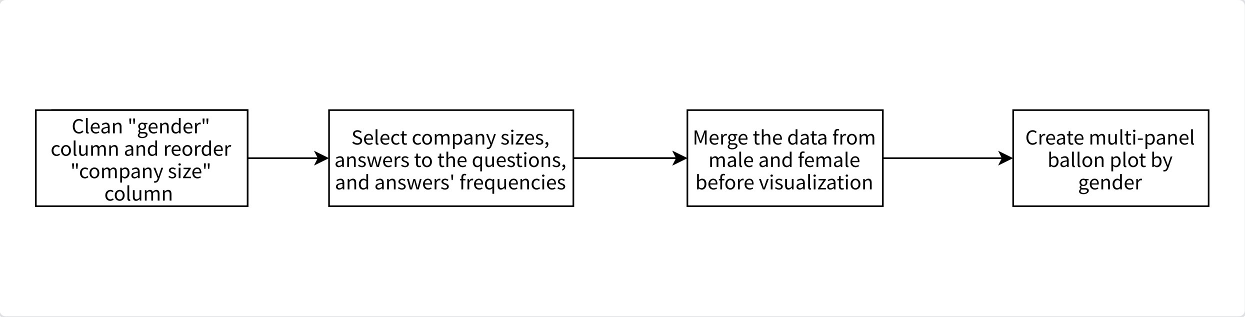
Mental Health in TECH/IT Workplace

Following our analysis of employees who sought mental treatment and had mental healthcare benefit, we would like to continue with relationship between mental issues and reactions in the workplace. In the first 3 plots, we would put emphasis on three dimensions to evaluate the mental health issues, which are influences on interview results, career outcomes and coworker relationships. We used gg

Multi-panel Ballon Plot by Gender

Balloon plot is an alternative to bar plot for visualizing a large categorical data. We' ll use the function `ggballoonplot()` [in ggpubr], which draws a graphical matrix of a contingency table, where each cell contains a dot whose size reflects the relative magnitude of the corresponding component.

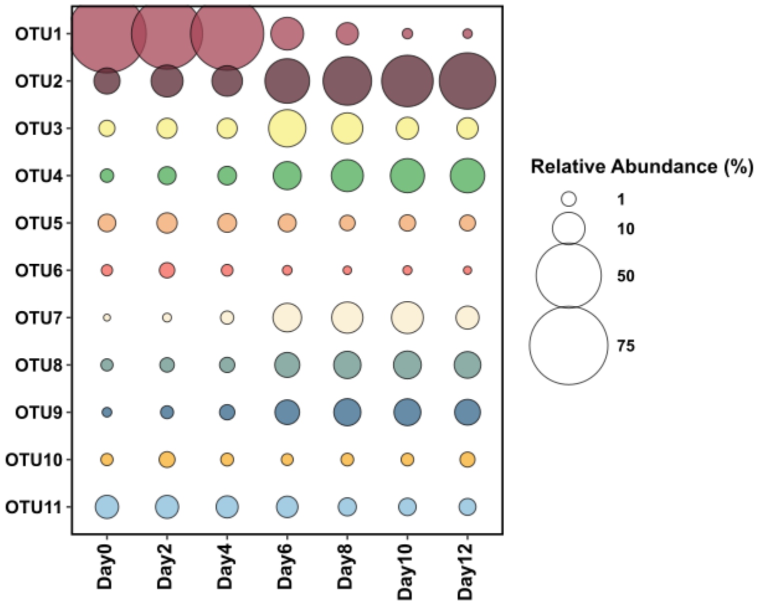
Before processing the balloon plot, we need to first clean the "gender" and "company_size" columns by filtering meaningless words and repetitive contents. After that, we include company sizes, answers to the questions, and answers' frequencies into the panel split by gender to visualize a grouped frequency table.


Clean "gender" column and reorder "company size" column
Select company sizes, answers to the questions, and answers' frequencies
Merge the data from male and female before visualization
Create multi-panel ballon plot by gender

Data Input

- Gender
- Company size
- Answers to the three questions
    - Would you bring up a mental health issue with a potential employer in an interview?
    - Do you feel that being identified as a person with a mental health issue would hurt your career?
    - Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?
- Answers' frequencies

Sample Plot

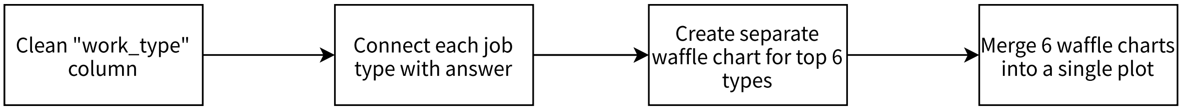


Waffle Chart by Job Types

Waffle Charts are a great way of visualizing data in relation to a whole, to highlight progress against a given threshold, or when dealing with populations too varied for pie charts. A lot of times, these are used as an alternative to the pie charts. It also has a niche for showing parts-to-whole contribution. It doesn’ t misrepresent or distort a data point (which a pie chart is sometimes guilty of doing). Waffle is a ggplot2 extension designed to create Waffle charts with a simple syntax. To install waffle package in R Studio use the following command:

```
install.packages("waffle")
```

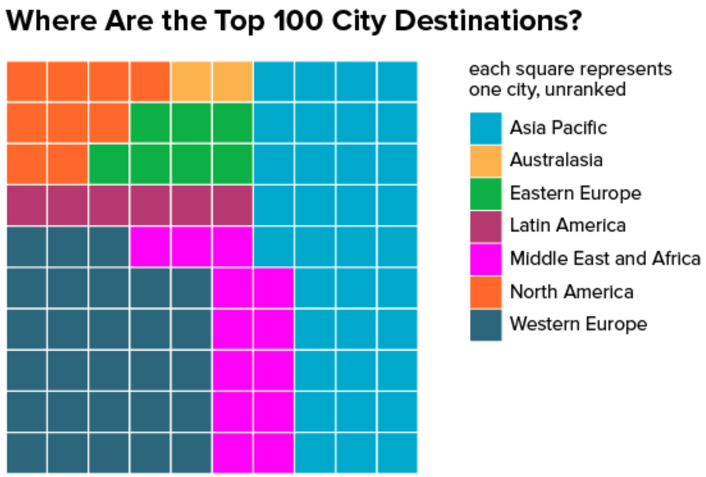
In our case, we evaluate "whether mental health issues would hurt your careers based on job types that are most common in the survey, including back-end engineers, front-end engineers, Team Lead, DevOps, Dev Evangelist and support. We first clean the "work_types"columns and extract each job type before the visualization. Then we separate them to see how respondents from each job type react to the statement that mental health issues would hurt their careers.


Clean "work_type" column
Connect each job type with answer
Create separate waffle chart for top 6 types
Merge 6 waffle charts into a single plot

Data Input

- Job types
  - Back-end engineers, front-end engineers, Team Lead, DevOps, Dev Evangelist and support
- Answers to the question
  - Do you feel that being identified as a person with a mental health issue would hurt your career?

Sample Plot



Text Analysis

Wordcloud

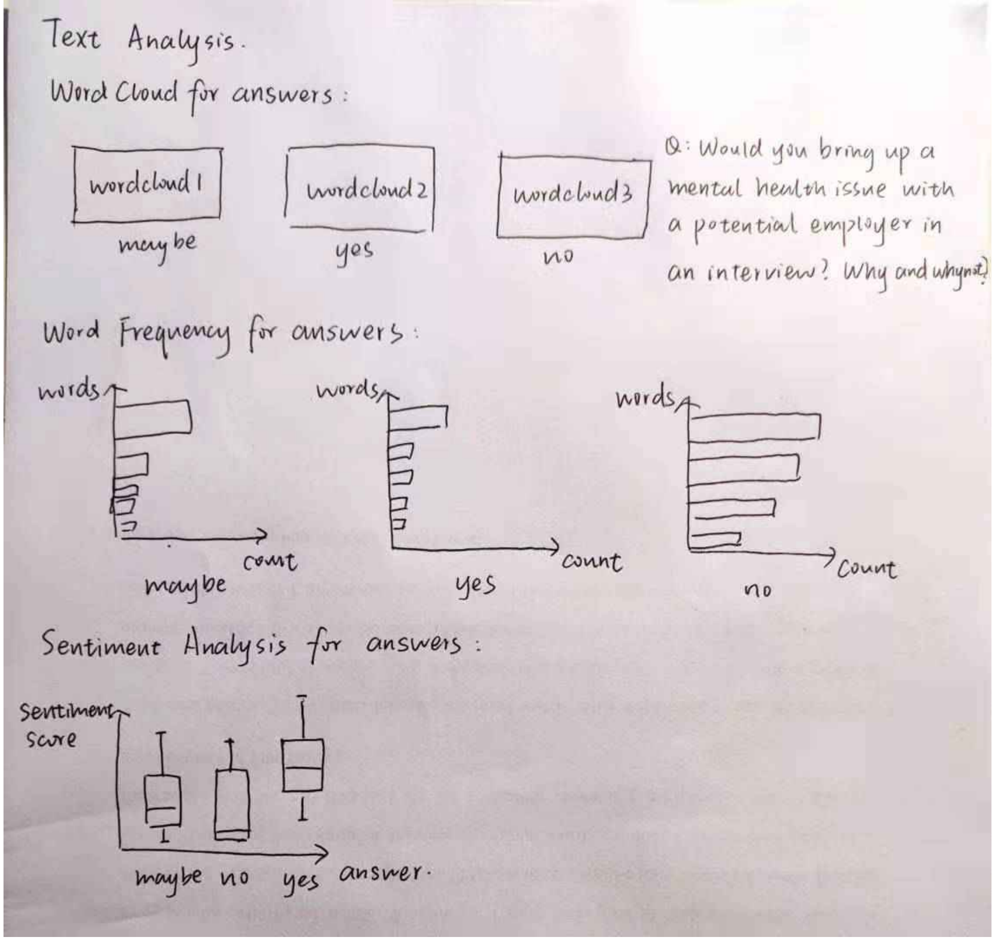
We did text analysis using participants’ answer for *"Would you bring up a mental health issue with a potential employer in an interview", "Why and why not"*. We drew word clouds to show frequent words used by participants for different answers. Firstly, we cleaned the answer by removing punctuations, stopwords and other meaningless words, numbers, white spaces and turning all words into lower cases. Then we stemmed the corpus and completed the stemmings. Finally, we built word clouds for each answer with tf scores.

Top 10 Most Frequent Words for Each Answer

We also showed the top 10 most frequent words for each answer by calculating each word frequency in all answers. The results showed that the most frequent words used in answers were verbs like feel and know and nouns like issue and job. As a result, we removed meaningless words that don't influence sentiment scoring and focused on words that conveyed some emotions.

Sentiment Analysis

We calculated sentiment scores relying on the Hu & Liu dictionary to analyze if there're sentiment differences among answers and used interactive boxplot to show it. We also added mean scores for each answer in the plot.


Text Analysis.
Word Cloud for answers:
wordcloud1
maybe
wordcloud2
yes
wordcloud3
no
Q: Would you bring up a mental health issue with a potential employer in an interview? Why and why not?
Word Frequency for answers:
words
count
maybe
words
count
yes
words
Count
no
Sentiment Analysis for answers:
Sentiment score
maybe
no
yes
answer.