Mental Health in Tech - Data Analytics Report

Rezwanul Islam Khan

Domain Introduction

Mental health has become an increasingly important concern in the tech industry, where high performance expectations, long working hours, and a culture of constant innovation can contribute to psychological stress. Despite rising awareness, stigma around mental health remains a barrier to open discussion, especially in professional environments. The Open Sourcing Mental Illness (OSMI) initiative has conducted annual surveys from 2017 to 2023 to understand mental health experiences in the tech workforce. This project analyzes and models responses from those surveys to explore how personal demographics and workplace support systems influence individuals' willingness to discuss mental health concerns. The findings aim to inform better workplace practices and promote a more inclusive, mentally healthy tech culture.

Objectives

1. Analyze gender distribution in the tech workforce to understand representation and how it may influence reported mental health outcomes.
2. Examine the relationship between mental health benefits and openness to sharing mental health concerns at work.
3. Compare the prevalence of reported mental health disorders across gender groups.
4. Assess how workplace support influences employees' mental health experiences — including employer discussions, resources, and medical coverage.
5. Evaluate trends over time in attitudes toward mental health openness from 2017 to 2023.
6. Identify country-level differences, with a focus on the United Kingdom, in mental health transparency and support.
7. Assess the impact of the COVID-19 pandemic (2020–2021) on mental health in the tech sector by comparing changes in disorder prevalence and willingness to share.

Data Preparation

The dataset used in this project was sourced from the Open Sourcing Mental Illness (OSMI) initiative on Kaggle, which collects annual surveys on mental health in the tech industry. For this analysis, all the datasets were combined from seven survey years — 2017 through 2023 — to enable longitudinal analysis and enhance the model's generalizability.

Key Preparation Steps:

Loading and Cleaning:

Each yearly dataset was read using read_csv() and cleaned individually to standardize column names and formats.

Column Harmonization:

Because column naming conventions varied between years, key variables (e.g., age, gender, willing_to_share, tech_company) were renamed consistently across datasets.

Variable Overview

As part of the data harmonization process, key survey variables were identified and standardized across multiple years. The original datasets (2017–2023) used varying column names, which were cleaned and aligned to create a consistent structure. Below is an overview of the primary variables used in the analysis:

Variable Name Description

tech_company: Indicates whether the respondent's employer is primarily a technology company or organization. mental_health_benefits: Whether the employer provides mental health benefits as part of their healthcare coverage.

resources_for_mh: Whether the employer offers resources to help employees learn about mental health disorders and seek help.

mh_employer_discussion: Whether the respondent has ever discussed their mental health with their employer.

mh_coworker_discussion: Whether the respondent has ever discussed mental health with coworkers.

medical_coverage: Indicates whether the respondent has medical insurance (private or state-provided) that covers mental health treatment.

currently_has_mh_disorder: Whether the respondent currently identifies as having a mental health disorder.

willing_to_share: A self-reported score (0–10) indicating the respondent's willingness to share mental health issues with friends and family.

age: Age of the respondent (cleaned to include values between 18 and 75 only).

gender: Self-identified gender of the respondent, standardized into "Male", "Female", and "Other".

country: The country where the respondent resides. (Countries with very low representation were grouped as "Other"for modeling).

Merging:

Cleaned data from all years were merged using bind_rows() into a unified final_data dataframe.

Missing Data Handling:

Rows with critical missing values were dropped.

Since the objective of this analysis is to understand the impact of mental health issues specifically within tech companies, we excluded all records where the tech_company field was missing or indicated a non-tech employer (i.e., 0 or false). By retaining only responses from individuals employed in tech organizations (where tech_company equals 1 or true), we ensured that the analysis remained focused and relevant to the target population.

Age values outside 18–75 were replaced with the dataset mean.

Gender entries were normalized into Male, Female, or Other.

For U.S. respondents, a lot of medical_coverage values were missing.According to OECD, 90 out of 100 americans health insurance cover mental illness(OECD,2017). As it is a very important factor in the data analysis the missing values were imputed to reflect a realistic 90% coverage assumption.

Feature Creation:

A categorical variable willing_to_share_level was created by segmenting the 0–10 willingness score into Low, Medium, and High levels for classification and visualization purposes.

Loading necessary libraries

Loading data

```
df2017 <- read_csv("data/2017.csv")
```

```
## New names:
## Rows: 756 Columns: 123
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (81): #, How many employees does your company or organization have?, Do...
dbl
## (25): <strong>Are you self-employed?</strong>, Is your employer primari...
lgl
## (15): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood Di...
dttm
## (2): Start Date (UTC), Submit Date (UTC)
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
```

```
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`
```

```
df2018 <- read_csv("data/2018.csv")
```

```
## New names:
## Rows: 417 Columns: 123
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (81): #, How many employees does your company or organization have?, Do...
dbl
## (25): <strong>Are you self-employed?</strong>, Is your employer primari...
lgl
## (15): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood Di...
dttm
## (2): Start Date (UTC), Submit Date (UTC)
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
```

```
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
```

```
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`

df2019 <- read_csv("data/2019.csv")

## New names:
## Rows: 352 Columns: 82
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (55): How many employees does your company or organization have?, Does
y... dbl
## (9): Overall, how much importance does your employer place on physical ...
lgl
## (18): *Are you self-employed?*, Is your employer primarily a tech
compan...
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....19`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....45`
## • `Why or why not?` -> `Why or why not?...61`
## • `Why or why not?` -> `Why or why not?...63`

df2020 <- read_csv("data/2020.csv")

## New names:
## Rows: 180 Columns: 120
## ── Column specification
```

```
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (78): #, How many employees does your company or organization have?,
Doe... dbl
## (25): *Are you self-employed?*, Is your employer primarily a tech
compan... lgl
## (17): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood
Dis...
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
```

```
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`
```

```
df2021 <- read_csv("data/2021.csv")
```

```
## New names:
## Rows: 131 Columns: 124
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (76): #, How many employees does your company or organization have?, Do...
dbl
## (25): *Are you self-employed?*, Is your employer primarily a tech compa...
lgl
## (21): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood Di...
dttm
```

```
## (2): Start Date (UTC), Submit Date (UTC)
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
```

```
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`
```

```
df2022 <- read_csv("data/2022.csv")
```

```
## New names:
## Rows: 164 Columns: 126
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (81): #, How many employees does your company or organization have?, Do...
dbl
## (25): *Are you self-employed?*, Is your employer primarily a tech compa...
lgl
## (18): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood Di...
dttm
## (2): Start Date (UTC), Submit Date (UTC)
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
```

```
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
```

```
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`

df2023 <- read_csv("data/2023.csv")

## New names:
## Rows: 6 Columns: 126
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (64): #, How many employees does your company or organization have?, Do...
dbl
## (25): *Are you self-employed?*, Is your employer primarily a tech compa...
lgl
## (35): Anxiety Disorder (Generalized, Social, Phobia, etc)...51, Mood Di...
dttm
## (2): Start Date (UTC), Submit Date (UTC)
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....20`
## • `Describe the conversation your coworker had with you about their mental
##   health (please do not use names).` -> `Describe the conversation your
##   coworker had with you about their mental health (please do not use
##   names)....46`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
```

```
##   (Generalized, Social, Phobia, etc)...51`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...52`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...53`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...54`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...55`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...56`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...57`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...59`
## • `Dissociative Disorder` -> `Dissociative Disorder...60`
## • `Substance Use Disorder` -> `Substance Use Disorder...61`
## • `Addictive Disorder` -> `Addictive Disorder...62`
## • `Other` -> `Other...63`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...64`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...65`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...66`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...67`
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...68`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...69`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...70`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...71`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...72`
## • `Dissociative Disorder` -> `Dissociative Disorder...73`
## • `Substance Use Disorder` -> `Substance Use Disorder...74`
## • `Addictive Disorder` -> `Addictive Disorder...75`
## • `Other` -> `Other...76`
## • `Anxiety Disorder (Generalized, Social, Phobia, etc)` -> `Anxiety
Disorder
##   (Generalized, Social, Phobia, etc)...77`
## • `Mood Disorder (Depression, Bipolar Disorder, etc)` -> `Mood Disorder
##   (Depression, Bipolar Disorder, etc)...78`
## • `Psychotic Disorder (Schizophrenia, Schizoaffective, etc)` -> `Psychotic
##   Disorder (Schizophrenia, Schizoaffective, etc)...79`
## • `Eating Disorder (Anorexia, Bulimia, etc)` -> `Eating Disorder
(Anorexia,
##   Bulimia, etc)...80`
```

```
## • `Attention Deficit Hyperactivity Disorder` -> `Attention Deficit
##   Hyperactivity Disorder...81`
## • `Personality Disorder (Borderline, Antisocial, Paranoid, etc)` ->
##   `Personality Disorder (Borderline, Antisocial, Paranoid, etc)...82`
## • `Obsessive-Compulsive Disorder` -> `Obsessive-Compulsive Disorder...83`
## • `Post-traumatic Stress Disorder` -> `Post-traumatic Stress
Disorder...84`
## • `Stress Response Syndromes` -> `Stress Response Syndromes...85`
## • `Dissociative Disorder` -> `Dissociative Disorder...86`
## • `Substance Use Disorder` -> `Substance Use Disorder...87`
## • `Addictive Disorder` -> `Addictive Disorder...88`
## • `Other` -> `Other...89`
## • `Why or why not?` -> `Why or why not?...98`
## • `Why or why not?` -> `Why or why not?...100`
## • `Other` -> `Other...118`
```

Merging data

```
df2017_clean <- df2017 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you currently have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you <strong>live</strong> in?`
  ) %>%
  transmute(
    year = 2017,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
```

```
    gender,
    country
  )


df2018_clean <- df2018 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you currently have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you <strong>live</strong> in?`
  ) %>%
  transmute(
    year = 2018,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )

df2019_clean <- df2019 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you *currently* have a mental health
```

```
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you *live* in?`
  ) %>%
  transmute(
    year = 2019,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )


df2020_clean <- df2020 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you *currently* have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you *live* in?`
  ) %>%
```

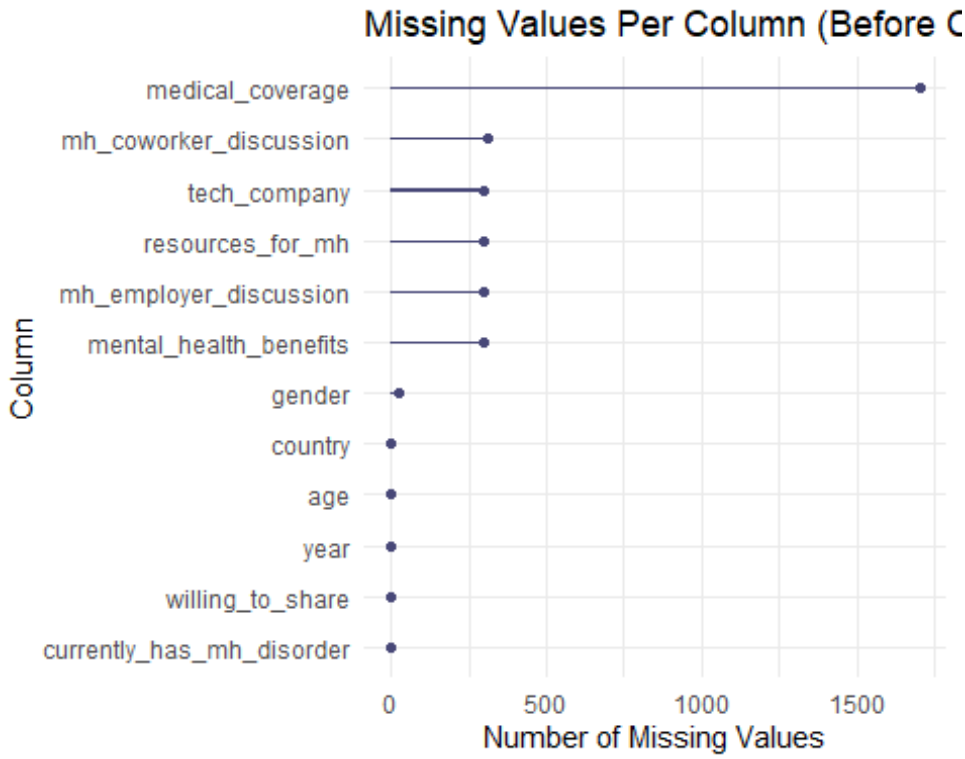
```r
  transmute(
    year = 2020,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )

df2021_clean <- df2021 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you *currently* have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you *live* in?`
  ) %>%
  transmute(
    year = 2021,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )
```

```r
df2022_clean <- df2022 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you *currently* have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you *live* in?`
  ) %>%
  transmute(
    year = 2022,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )

df2023_clean <- df2023 %>%
  rename(
    tech_company = `Is your employer primarily a tech company/organization?`,
    mental_health_benefits = `Does your employer provide mental health
benefits as part of healthcare coverage?`,
    resources_for_mh = `Does your employer offer resources to learn more
about mental health disorders and options for seeking help?`,
    currently_has_mh_disorder = `Do you *currently* have a mental health
disorder?`,
    mh_employer_discussion = `Have you ever discussed your mental health with
your employer?`,
    mh_coworker_discussion = `Have you ever discussed your mental health with
coworkers?`,
```

```
    medical_coverage = `Do you have medical coverage (private insurance or
state-provided) that includes treatment of mental health disorders?`,
    willing_to_share = `How willing would you be to share with friends and
family that you have a mental illness?`,
    age = `What is your age?`,
    gender = `What is your gender?`,
    country = `What country do you *live* in?`
  ) %>%
  transmute(
    year = 2023,
    tech_company,
    mental_health_benefits,
    resources_for_mh,
    currently_has_mh_disorder,
    mh_employer_discussion,
    mh_coworker_discussion,
    medical_coverage,
    willing_to_share,
    age,
    gender,
    country
  )


final_data <- bind_rows(
  df2017_clean,
  df2018_clean,
  df2019_clean,
  df2020_clean,
  df2021_clean,
  df2022_clean,
  df2023_clean
)
```

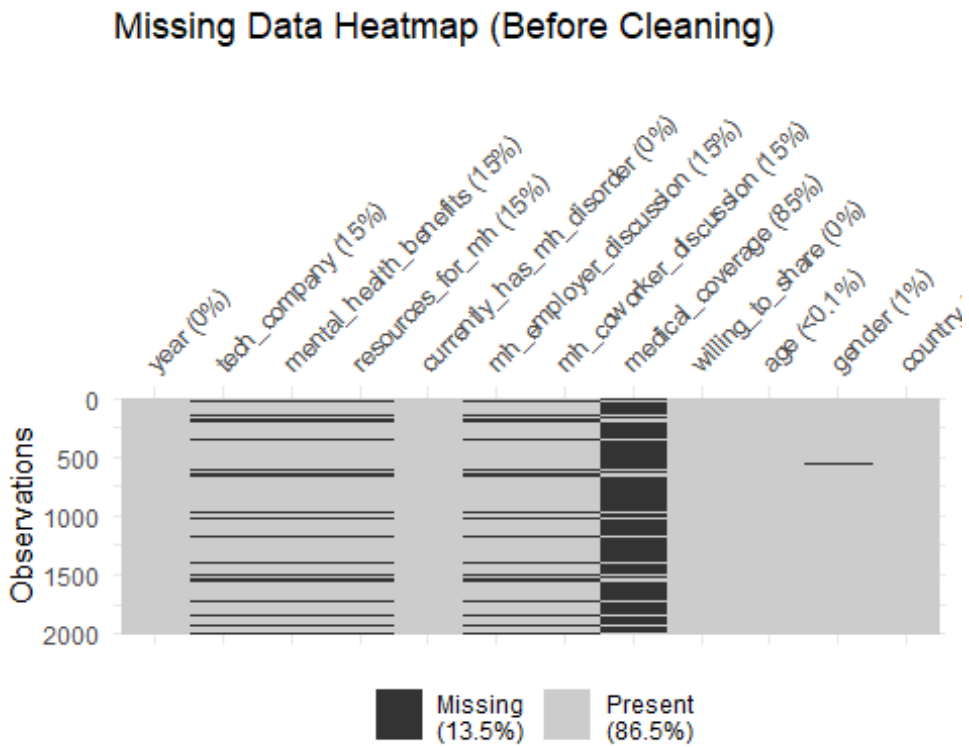
Missing value bar chart

```
# Missing value bar chart
gg_miss_var(final_data) +
  labs(
    title = "Missing Values Per Column (Before Cleaning)",
    x = "Column",
    y = "Number of Missing Values"
  )
```



Missing data heatmap

```
# Missing data heatmap
vis_miss(final_data, warn_large_data = FALSE) +
  labs(title = "Missing Data Heatmap (Before Cleaning)")
```



As we can see there is some data missing we will now start cleaning the data.

1.Cleaning the gender Column

```
#  Cleaning the `gender` column
# We found a variety of gender entries (e.g., "male", "MALE", "I identify as man", etc.)
# To standardize them, we grouped all variations into three categories: "Male", "Female", "Other"
# Missing values are assigned to "Other" to preserve the record without biasing distribution.

final_data <- final_data %>%
  mutate(
    gender = ifelse(is.na(gender), "Other", tolower(trimws(gender))),
    gender = case_when(
      str_detect(gender, "female") ~ "Female",
      str_detect(gender, "male") ~ "Male",
      TRUE ~ "Other"
    )
  )
```

2. Cleaning the age Column

```
#  Cleaning the `age` column
# Age values below 18 or above 75 are considered invalid.
# We replace these with the mean age (calculated from valid values only).
```

```r
final_data <- final_data %>%
  mutate(age = as.numeric(age)) %>%
  mutate(age = ifelse(age < 18 | age > 75, NA, age))

# Replace NA ages with mean age of valid range
mean_age <- mean(final_data$age, na.rm = TRUE)
final_data <- final_data %>%
  mutate(age = ifelse(is.na(age), round(mean_age), age))
```

3. Cleaning the mental_health_benefits Column

```r
# Standardizing `mental_health_benefits`
# "No" and "Not eligible" both mean the respondent does not have coverage.
# We group all such responses under "No", and others under "Yes".

final_data <- final_data %>%
  mutate(mental_health_benefits = tolower(trimws(mental_health_benefits)))
%>%
  mutate(mental_health_benefits = case_when(
    mental_health_benefits %in% c("no", "not eligible") ~ "No",
    TRUE ~ "Yes"
  ))
```

4. Imputing medical_coverage Based on Country

```r
# Step 1: Subset ALL USA respondents
usa_all <- which(final_data$country == "United States of America")

# Step 2: Count how many should be "Yes" (90%)
total_usa <- length(usa_all)
target_yes_count <- round(total_usa * 0.9)

# Step 3: Find who already said "Yes"
already_yes <- usa_all[final_data$medical_coverage[usa_all] == "Yes"]

# Step 4: Determine how many more we need to flip to "Yes"
remaining_needed <- max(0, target_yes_count - length(already_yes))

# Step 5: Candidates to flip to "Yes" (those with "No" or NA)
flip_candidates <- usa_all[
  is.na(final_data$medical_coverage[usa_all]) |
  final_data$medical_coverage[usa_all] == "No"
]

# Step 6: Randomly sample from flip_candidates
set.seed(123)
flip_to_yes <- sample(flip_candidates, size = remaining_needed)

# Step 7: Assign values
final_data$medical_coverage[flip_to_yes] <- "Yes"
```

```r
# All remaining USA records (not "Yes") will be "No"
usa_still_na_or_no <- setdiff(usa_all, c(already_yes, flip_to_yes))
final_data$medical_coverage[usa_still_na_or_no] <- "No"

final_data <- final_data %>%
  mutate(medical_coverage = ifelse(is.na(medical_coverage), "Yes",
medical_coverage))
```

5. Dropping Records with Critical Missing Values and non_techcompanies

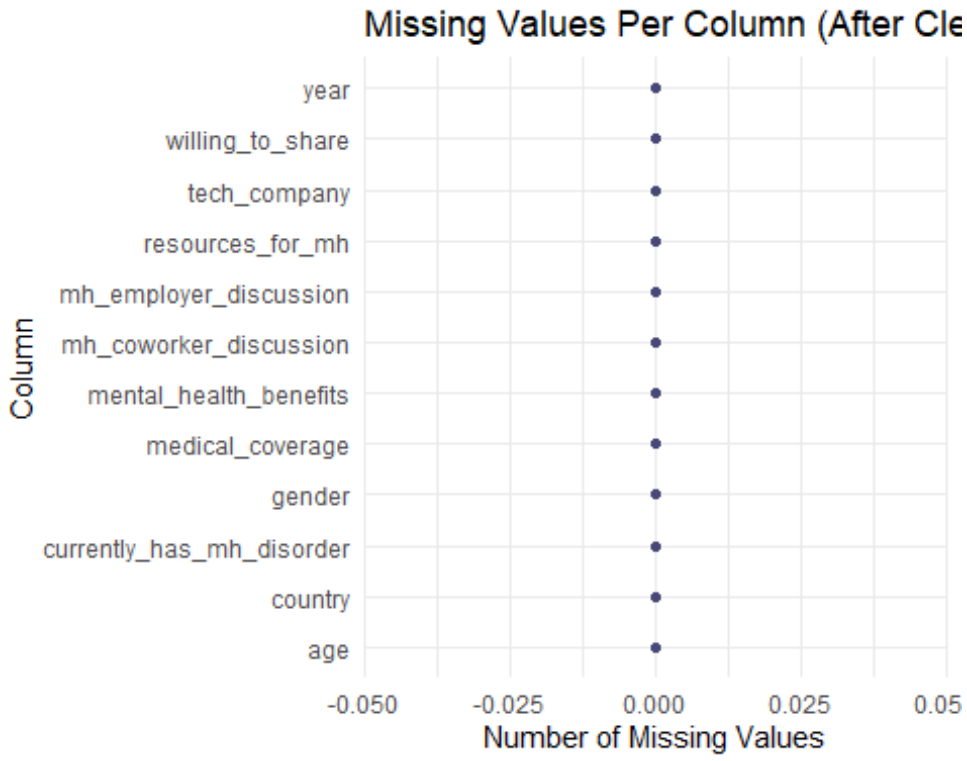
```r
# Dropping records with critical missing values and non-tech companies
final_data <- final_data %>%
  drop_na() %>%
  filter(tech_company == 1)
```

Saving the final clensed data

```r
write.csv(final_data, "C:/Users/rez12/Downloads/Data
Analytics/final_cleaned_data.csv", row.names = FALSE)
```

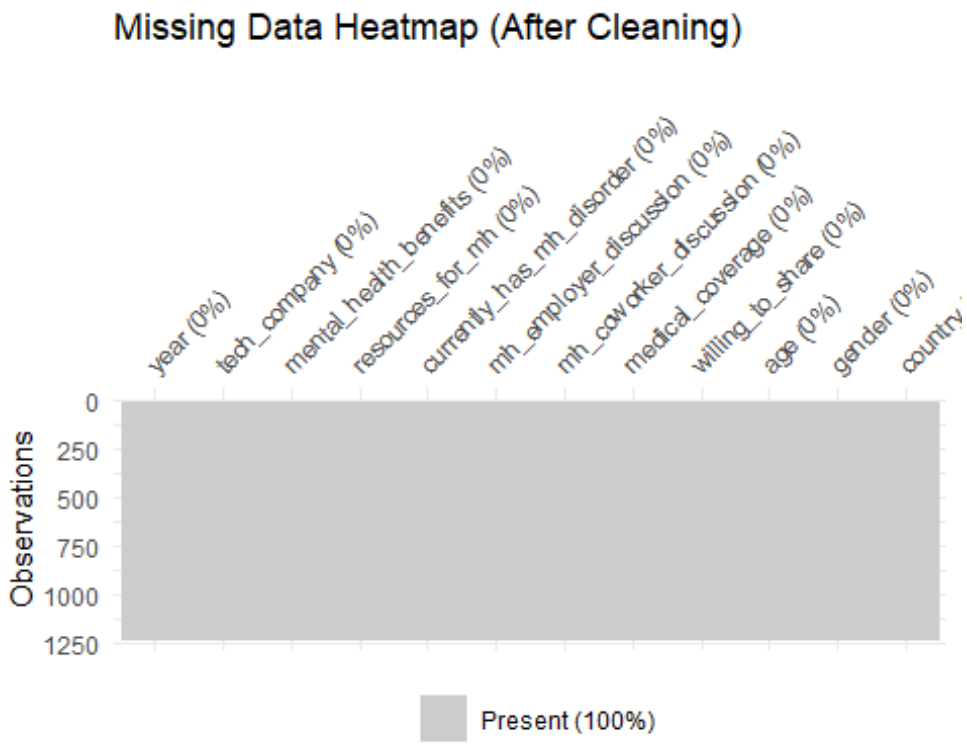
Bar Plot: Missing Values per Column (After Cleaning)

```r
library(naniar)

gg_miss_var(final_data) +
  labs(
    title = "Missing Values Per Column (After Cleaning)",
    x = "Column",
    y = "Number of Missing Values"
  )
```



Heatmap View of Missing Data (After Cleaning)

```
vis_miss(final_data, warn_large_data = FALSE) +
  labs(title = "Missing Data Heatmap (After Cleaning)")
```



Creating a New Leveled Column

```
library(dplyr)

final_data <- final_data %>%
  mutate(willing_to_share_level = case_when(
    willing_to_share <= 3 ~ "Low",
    willing_to_share <= 6 ~ "Medium",
    willing_to_share <= 10 ~ "High",
    TRUE ~ NA_character_
  )) %>%
  mutate(willing_to_share_level = factor(willing_to_share_level,
                                         levels = c("Low", "Medium",
"High")))
```

Summary Tables

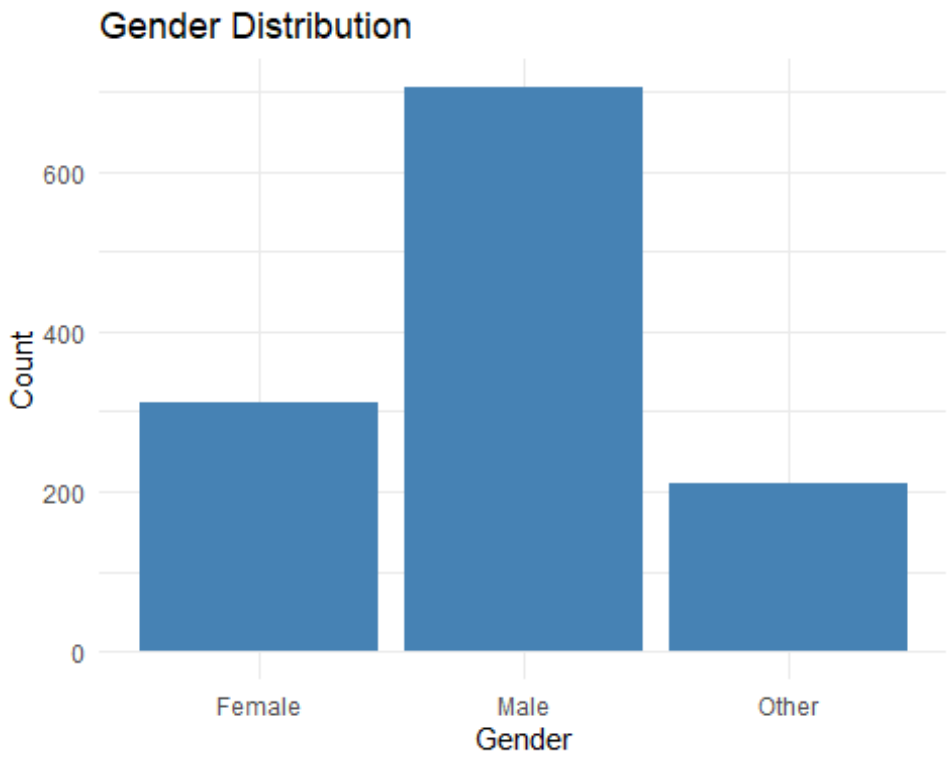
Gender Distribution

```
table(final_data$gender)

##
## Female   Male  Other
##    312    706    210

prop.table(table(final_data$gender)) * 100

##
##   Female     Male    Other
## 25.40717 57.49186 17.10098
```

Visualizations

Gender Distribution Bar Plot

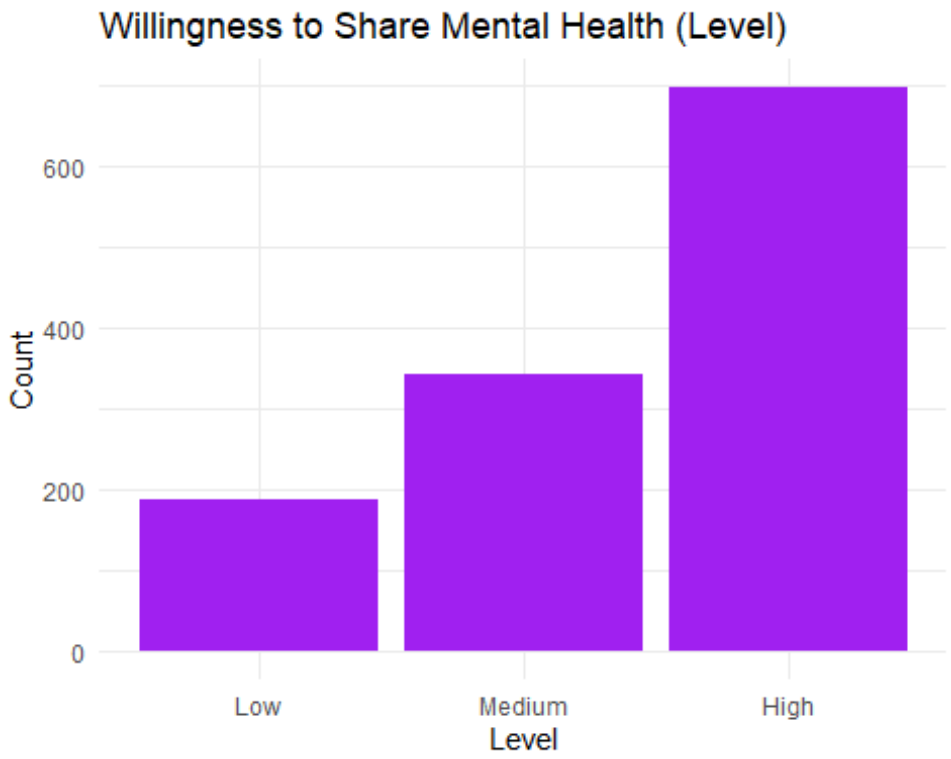
```
ggplot(final_data, aes(x = gender)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Gender Distribution", x = "Gender", y = "Count") +
  theme_minimal()
```



Insight: The dataset shows that male respondents significantly outnumber female and other gender identities. This suggests a gender imbalance among tech professionals, which may affect how mental health experiences are reported and perceived.

Willingness to Share (Level)

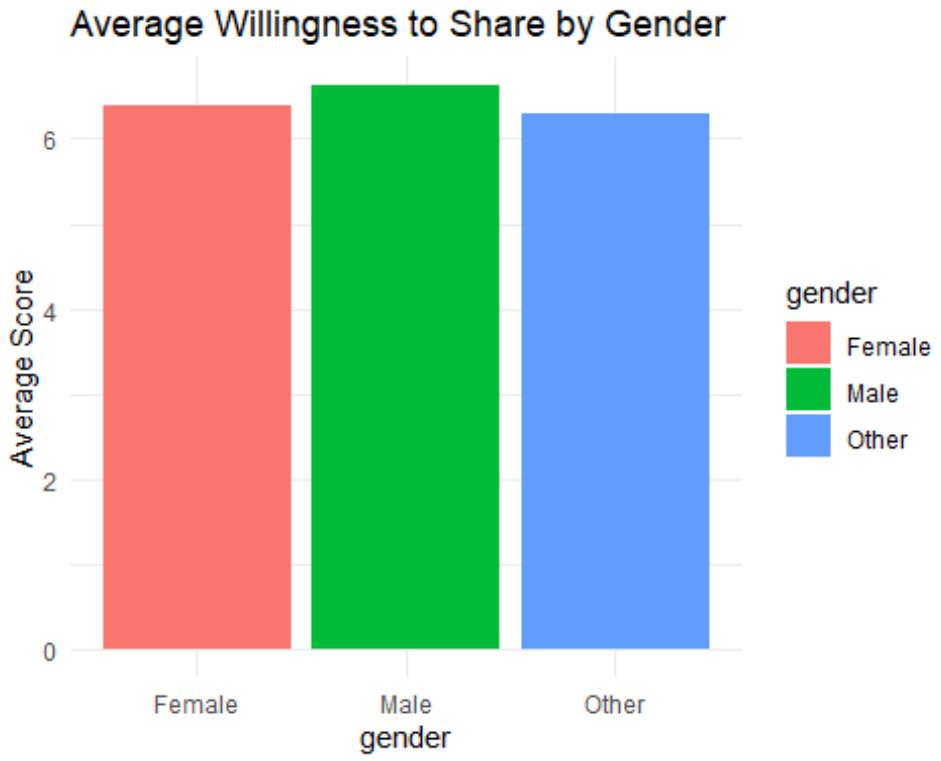
```
ggplot(final_data, aes(x = willing_to_share_level)) +
  geom_bar(fill = "purple") +
  labs(title = "Willingness to Share Mental Health (Level)", x = "Level", y =
"Count") +
  theme_minimal()
```



Insight: A large portion of respondents fall into the “High” and “Medium” willingness levels, indicating a generally positive attitude toward discussing mental health. However, a notable share still reports “Low” willingness, suggesting stigma or fear may persist in some environments.

Average Willingness to Share by Gender

```
final_data %>%
  group_by(gender) %>%
  summarise(avg_willingness = mean(willing_to_share)) %>%
  ggplot(aes(x = gender, y = avg_willingness, fill = gender)) +
  geom_col() +
  labs(title = "Average Willingness to Share by Gender", y = "Average Score")
+
  theme_minimal()
```



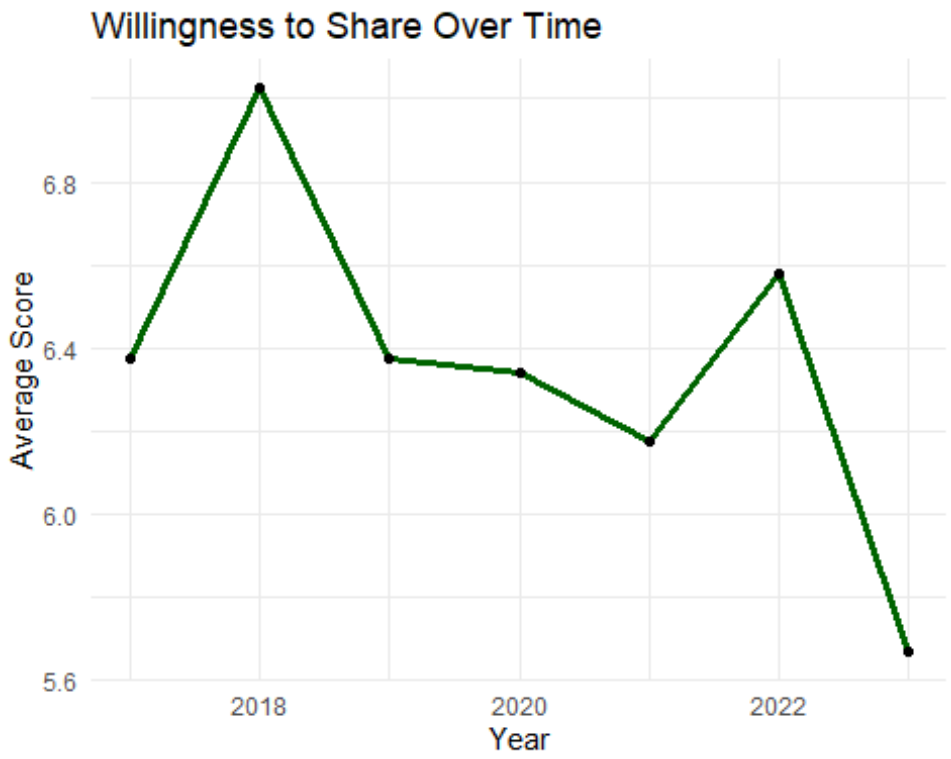
Insight: Male respondents reported a slightly higher average willingness to share mental health concerns compared to females and others. This could reflect differences in openness, social expectations, or workplace experience among gender groups.

Trends Over Time

Average Willingness to Share Per Year

```
final_data %>%
  group_by(year) %>%
  summarise(avg = mean(willing_to_share)) %>%
  ggplot(aes(x = as.numeric(year), y = avg)) +
  geom_line(color = "darkgreen", size = 1.2) +
  geom_point() +
  labs(title = "Willingness to Share Over Time", x = "Year", y = "Average
Score") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

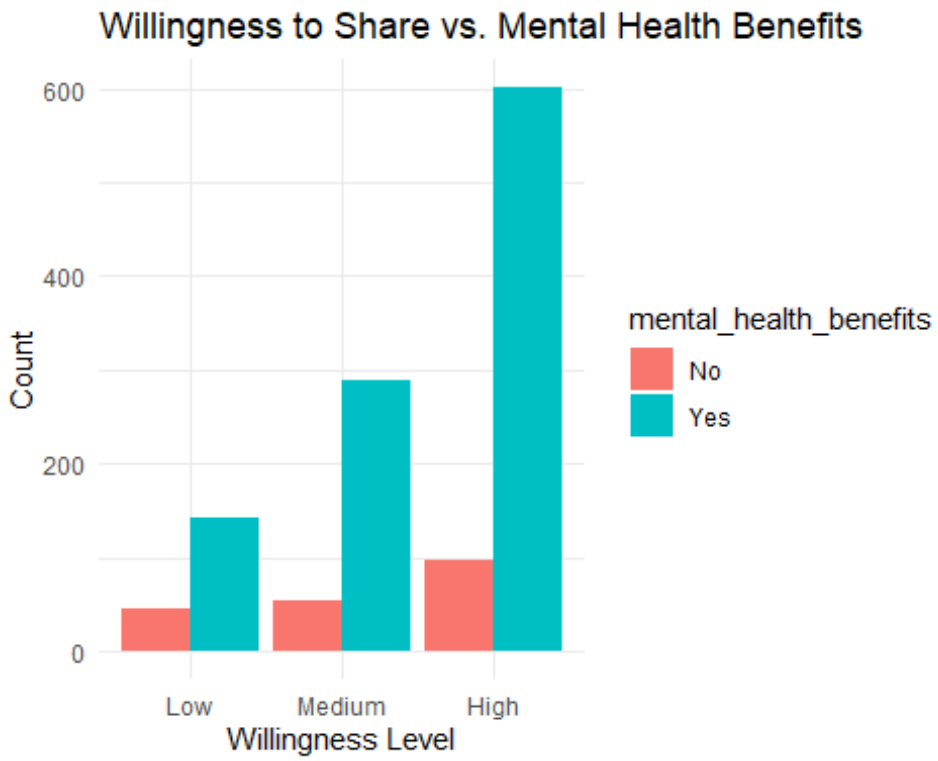


Insight: The willingness to share mental health concerns peaked in 2018, showing early positive momentum. A noticeable dip occurred during and after the COVID-19 period, with partial recovery in 2022. However, 2023 saw a significant drop, possibly reflecting emerging stigma or reduced organizational focus

Additional Insights

Willingness to Share vs. Mental Health Benefits

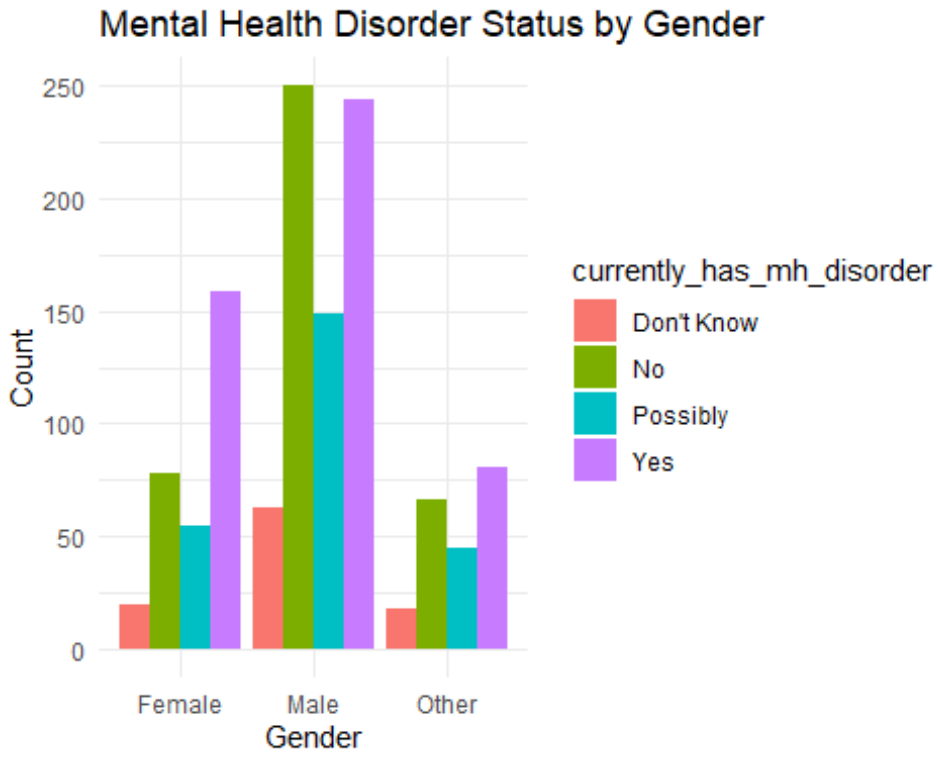
```
final_data %>%
  group_by(willing_to_share_level, mental_health_benefits) %>%
  summarise(count = n(), .groups = "drop") %>%
  ggplot(aes(x = willing_to_share_level, y = count, fill =
mental_health_benefits)) +
  geom_col(position = "dodge") +
  labs(title = "Willingness to Share vs. Mental Health Benefits",
       x = "Willingness Level", y = "Count") +
  theme_minimal()
```



Insight: Respondents who report having mental health benefits are more likely to fall into higher willingness levels. This highlights the impact of employer support structures in promoting openness around mental health.

Mental Health Disorder Status by Gender

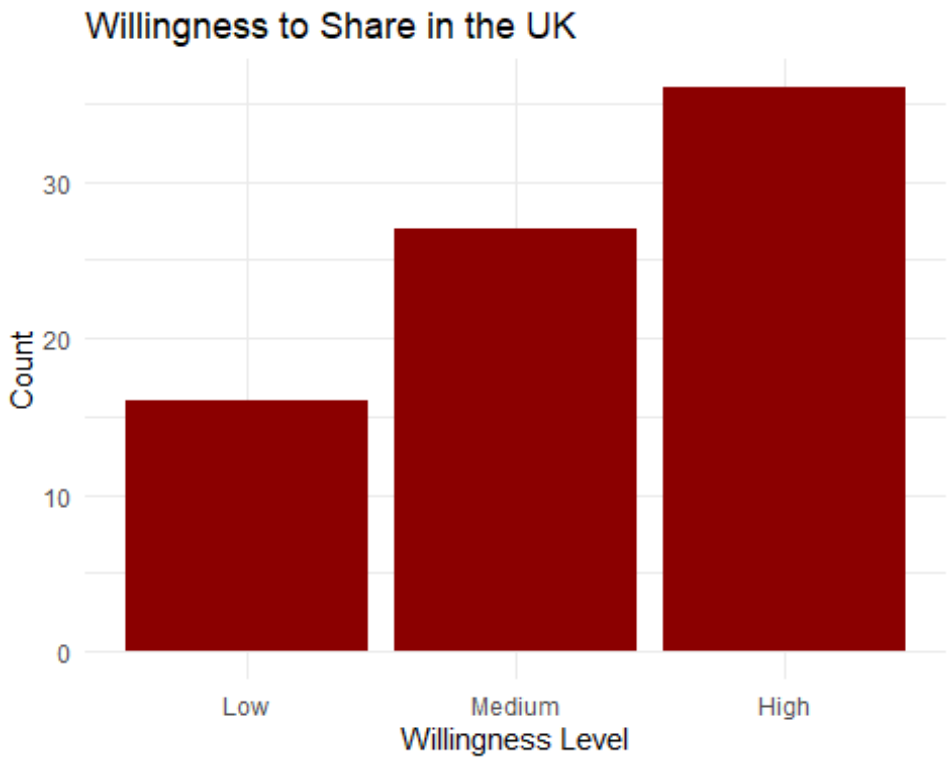
```
final_data %>%
  group_by(gender, currently_has_mh_disorder) %>%
  summarise(count = n(), .groups = "drop") %>%
  ggplot(aes(x = gender, y = count, fill = currently_has_mh_disorder)) +
  geom_col(position = "dodge") +
  labs(title = "Mental Health Disorder Status by Gender",
       x = "Gender", y = "Count") +
  theme_minimal()
```



Insight: Female and other-gender respondents report current mental health disorders more frequently than males. This could indicate actual prevalence differences or greater willingness among these groups to acknowledge mental health conditions.

UK: Willingness to Share Mental Health (Level)

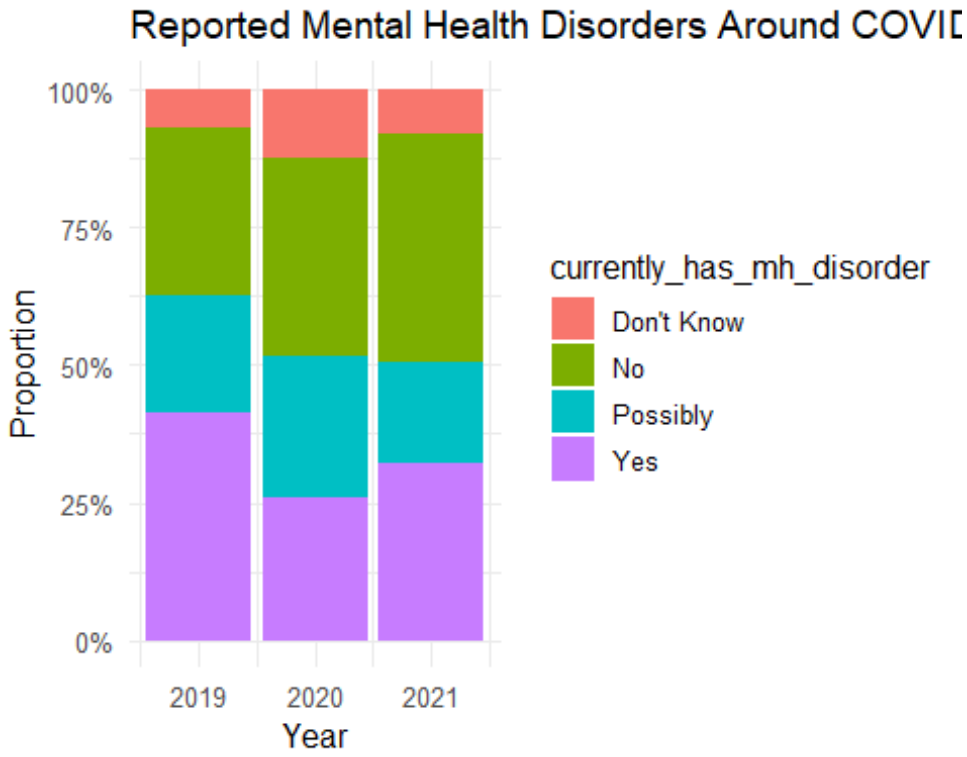
```
final_data %>%
  filter(country == "United Kingdom") %>%
  ggplot(aes(x = willing_to_share_level)) +
  geom_bar(fill = "darkred") +
  labs(title = "Willingness to Share in the UK", x = "Willingness Level", y =
"Count") +
  theme_minimal()
```



Insight: Respondents from the UK show a strong concentration in the “High” willingness level, suggesting a relatively open culture toward mental health in the tech sector. This aligns with the UK’s universal healthcare and strong mental health advocacy efforts.

COVID-19 Impact: Mental Health Disorder Prevalence

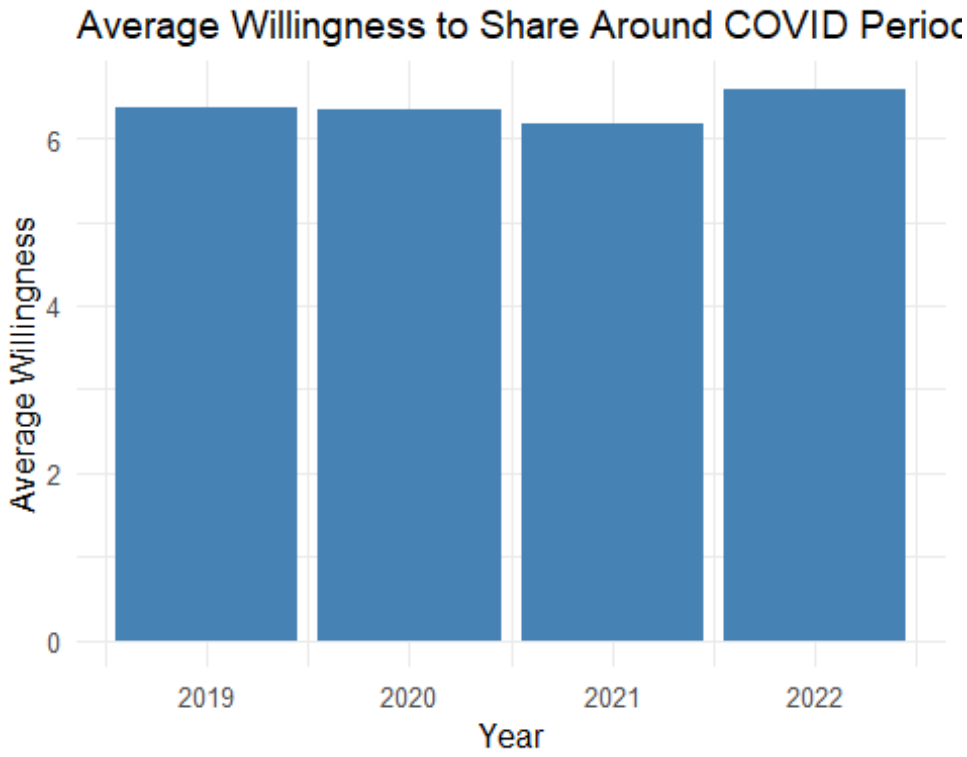
```
final_data %>%
  filter(year %in% c("2019", "2020", "2021")) %>%
  group_by(year, currently_has_mh_disorder) %>%
  summarise(count = n(), .groups = "drop") %>%
  ggplot(aes(x = year, y = count, fill = currently_has_mh_disorder)) +
  geom_col(position = "fill") +
  labs(title = "Reported Mental Health Disorders Around COVID Years",
       y = "Proportion", x = "Year") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme_minimal(base_size = 12)
```



Insight: The proportion of respondents reporting mental health disorders rose during the 2020–2021 period, suggesting that the COVID-19 pandemic may have intensified psychological challenges for tech employees.

COVID-19 Impact: Willingness to Share

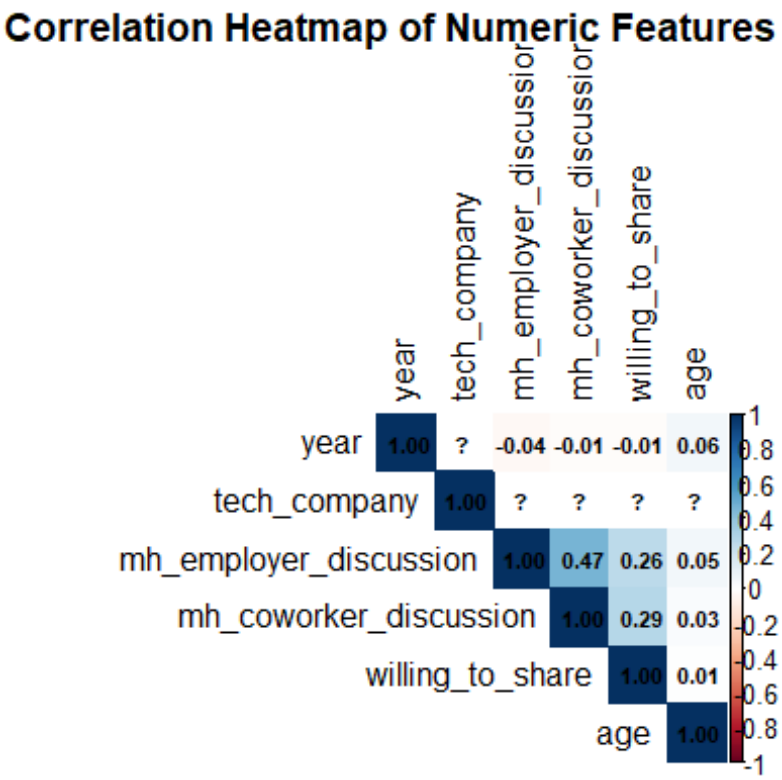
```
final_data %>%
  filter(year %in% c("2019", "2020", "2021", "2022")) %>%
  group_by(year) %>%
  summarise(avg = mean(willing_to_share)) %>%
  ggplot(aes(x = year, y = avg)) +
  geom_col(fill = "steelblue") +
  labs(title = "Average Willingness to Share Around COVID Period",
       y = "Average Willingness", x = "Year") +
  theme_minimal(base_size = 12)
```



Insight: There is a noticeable dip in willingness to share during 2020, with a recovery by 2022. This pattern suggests that the onset of the pandemic may have introduced uncertainty or reduced openness about mental health issues, which later improved as support systems adapted.

Correlation Heatmap of Numeric Features

```
# Select only numeric columns
numeric_data <- final_data %>%
  select(where(is.numeric))

# Compute correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs")

## Warning in cor(numeric_data, use = "complete.obs"): the standard deviation
is
## zero

# Load package and plot heatmap

corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "black", tl.col = "black", number.cex = 0.7,
         title = "Correlation Heatmap of Numeric Features", mar = c(0,0,1,0))
```



Insight: The correlation heatmap reveals notable patterns among the numeric variables in the dataset. Most notably, there is a moderate positive correlation (r = 0.47) between respondents discussing mental health with employers and with coworkers, indicating that openness in one setting may reflect a generally supportive workplace culture. A weaker but still meaningful correlation (r = 0.26) exists between discussing mental health with employers and willingness to share such issues with friends or family, suggesting that workplace support may contribute to broader openness. Other variables such as age, year, and company type show minimal or unclear correlation with willingness to share, indicating their limited standalone predictive power. ## Predicting Mental Health Disclosure

Objective:

Build predictive models to identify factors influencing employees' willingness to discuss mental health.

Features Used:

- Age
- Gender
- Country
- mental_health_benefits
- medical_coverage
- mh_employer_discussion
- mh_coworker_discussion
- currently_has_mh_disorder

Target Variable:

- `willing_to_share` (numerical score from 0–10)

Data Preprocessing

```
# Standardize Yes/No strings for benefits and coverage
final_data <- final_data %>%
  mutate(
    mental_health_benefits = case_when(
      str_to_lower(mental_health_benefits) == "yes" ~ "Yes",
      TRUE ~ "No"
    ),
    medical_coverage = case_when(
      str_to_lower(medical_coverage) == "yes" ~ "Yes",
      TRUE ~ "No"
    ),
    mental_health_benefits = factor(mental_health_benefits),
    medical_coverage = factor(medical_coverage)
  )
# Merge rare countries into 'Other'
merge_rare_countries <- function(df, min_count = 5) {
  country_counts <- table(df$country)
  df$country <- ifelse(df$country %in% names(country_counts[country_counts >=
min_count]),
                       df$country, "Other")
  df$country <- factor(df$country)
  return(df)
}
model_data <- final_data %>%
  select(willing_to_share, age, gender, country,mh_employer_discussion,
mh_coworker_discussion, currently_has_mh_disorder, mental_health_benefits,
medical_coverage) %>%
  mutate(
    gender = factor(gender),
    country = factor(country),
    mental_health_benefits = factor(mental_health_benefits),
    medical_coverage = factor(medical_coverage)
```

```
  )

# Split data
set.seed(123)
train_index <- createDataPartition(model_data$willing_to_share, p = 0.8, list
= FALSE)
train_data <- model_data[train_index, ]
test_data <- model_data[-train_index, ]

# Apply rare-country merging to both sets
train_data <- merge_rare_countries(train_data)
test_data <- merge_rare_countries(test_data)
```

Linear Regression

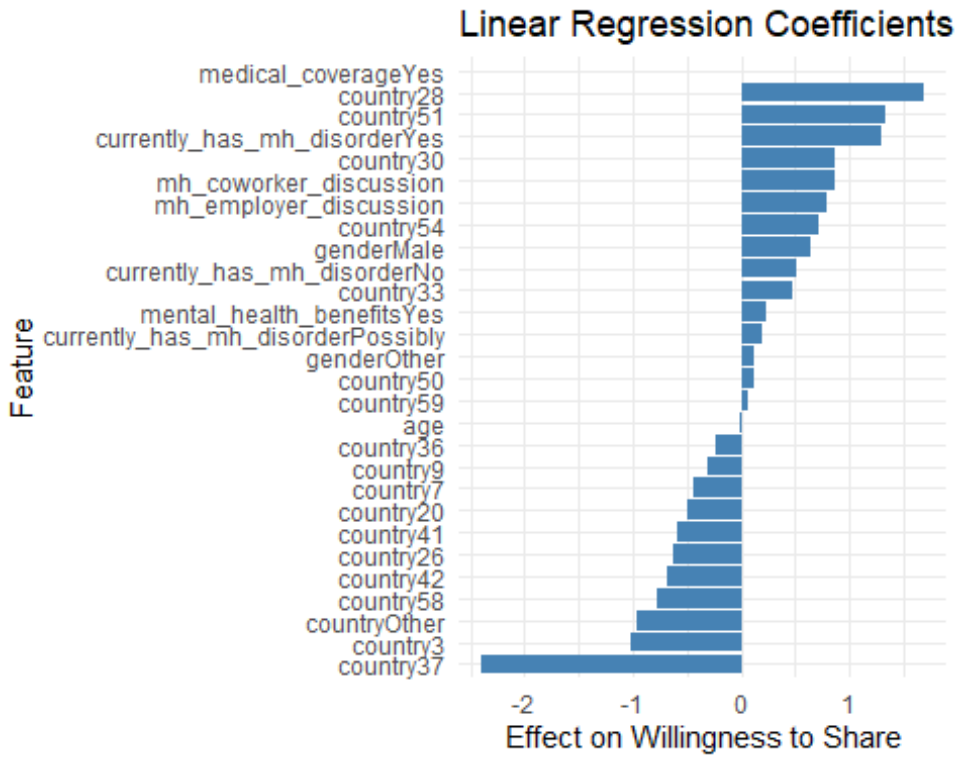
```
lr_model <- lm(willing_to_share ~ ., data = train_data)
lr_preds <- predict(lr_model, test_data)
postResample(lr_preds, test_data$willing_to_share)
```

```
##       RMSE   Rsquared        MAE
## 2.73192464 0.09963316 2.18190690
```

####Linear Regression Coefficient Plot

```
# Extract coefficients and prepare for plotting
coef_df <- data.frame(
  Feature = names(coef(lr_model))[-1],  # exclude intercept
  Coefficient = coef(lr_model)[-1]
)

# Load ggplot2 if needed
library(ggplot2)

# Plot coefficients
ggplot(coef_df, aes(x = reorder(Feature, Coefficient), y = Coefficient)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Linear Regression Coefficients",
       x = "Feature", y = "Effect on Willingness to Share") +
  theme_minimal()
```

```
## Warning: Removed 1 row containing missing values or values outside the
scale range
## (`geom_col()`).
```



Random Forest

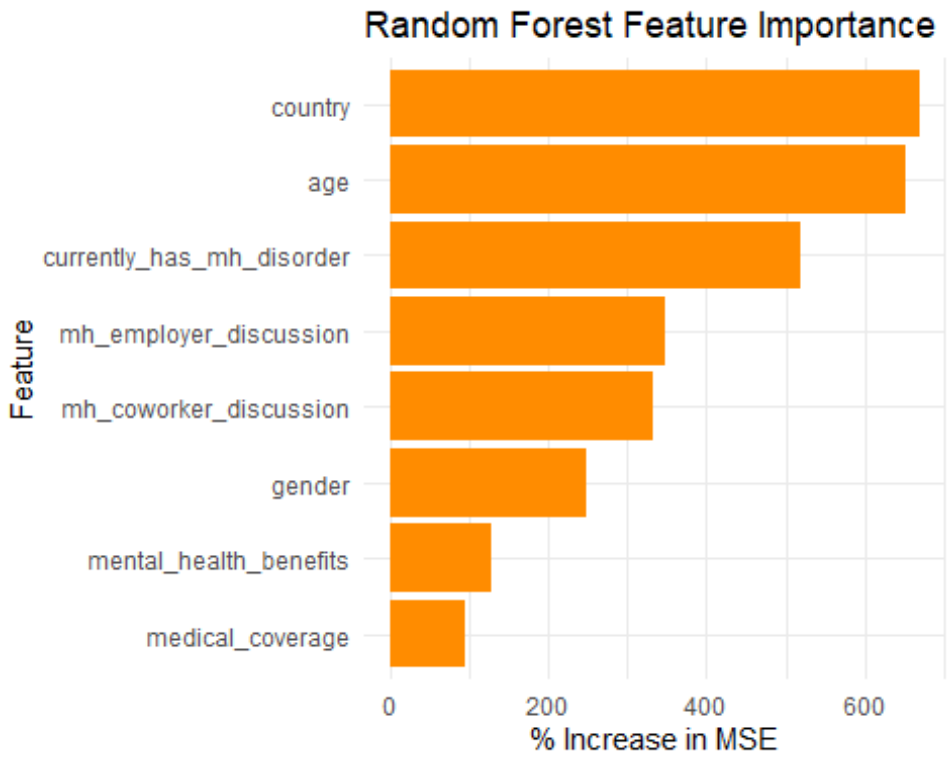
```
# Match factor levels in test set to training set
test_data$gender <- factor(test_data$gender, levels =
levels(train_data$gender))
test_data$country <- factor(test_data$country, levels =
levels(train_data$country))
test_data$mental_health_benefits <- factor(test_data$mental_health_benefits,
levels = levels(train_data$mental_health_benefits))
test_data$medical_coverage <- factor(test_data$medical_coverage, levels =
levels(train_data$medical_coverage))
rf_model <- randomForest(willing_to_share ~ ., data = train_data, ntree =
100)
rf_preds <- predict(rf_model, test_data)
postResample(rf_preds, test_data$willing_to_share)
```

```
##       RMSE   Rsquared        MAE
## 2.78252764 0.07123889 2.23527374
```

####Feature Importance (Random Forest)

```
# Extract feature importance from trained Random Forest model
importance_scores <- importance(rf_model)

# Convert to a tidy data frame
importance_df <- data.frame(
  Feature = rownames(importance_scores),
```

```r
  Importance = importance_scores[, 1]  # %IncMSE
) %>%
  arrange(desc(Importance))

# Plot using ggplot2
library(ggplot2)

ggplot(importance_df, aes(x = reorder(Feature, Importance), y = Importance))
+
  geom_col(fill = "darkorange") +
  coord_flip() +
  labs(title = "Random Forest Feature Importance",
       x = "Feature", y = "% Increase in MSE") +
  theme_minimal()
```



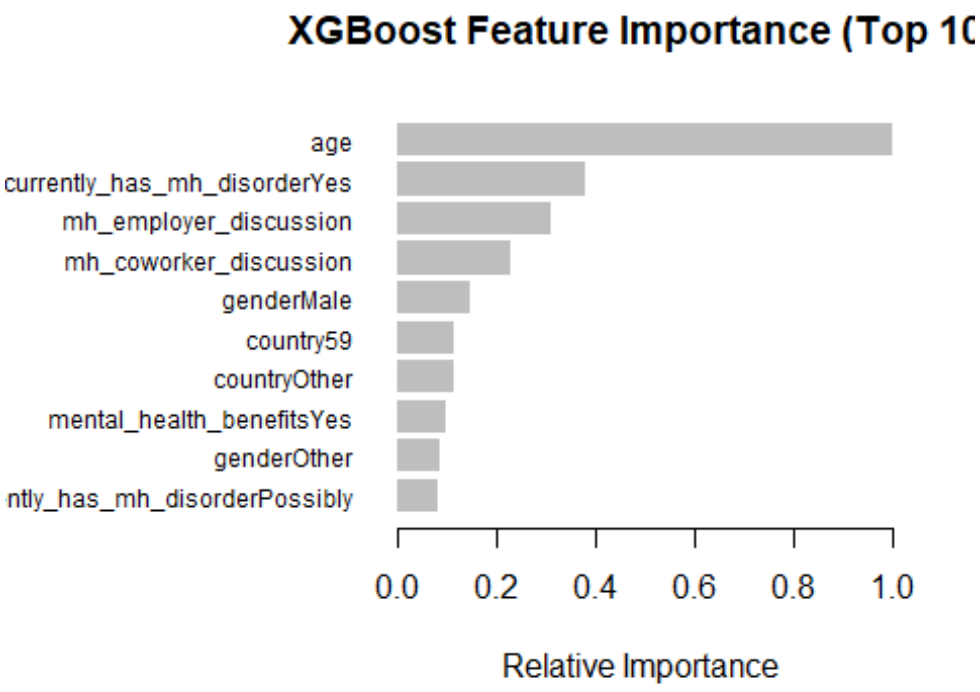
XGBoost

```r
# Convert factors to numeric for XGBoost
x_train <- model.matrix(willing_to_share ~ ., data = train_data)[, -1]
y_train <- train_data$willing_to_share
x_test <- model.matrix(willing_to_share ~ ., data = test_data)[, -1]
y_test <- test_data$willing_to_share

xgb_model <- xgboost(data = x_train, label = y_train, nrounds = 50, objective
= "reg:squarederror", verbose = 0)
xgb_preds <- predict(xgb_model, x_test)
postResample(xgb_preds, y_test)
```

```
##       RMSE   Rsquared        MAE
## 2.94704177 0.05021177 2.35780739
```

####XGBoost Feature Importance Plot

```
feature_names <- colnames(x_train)

# Compute and plot importance
importance_matrix <- xgb.importance(feature_names = feature_names, model =
xgb_model)

# Plot
xgb.plot.importance(importance_matrix,
                    top_n = 10,
                    measure = "Gain",
                    rel_to_first = TRUE,
                    xlab = "Relative Importance",
                    main = "XGBoost Feature Importance (Top 10)")
```



Model Interpretation Across all three models, several consistent patterns emerged. Linear Regression highlighted that having medical coverage and a current mental health disorder positively influenced respondents' willingness to share, while certain countries showed strong negative effects. Random Forest emphasized the importance of demographic variables such as country and age, along with key workplace interactions like employer and coworker discussions. XGBoost, meanwhile, placed the greatest weight on age and current mental health status, suggesting that individual traits may play a stronger

predictive role than workplace policies alone. Together, these models suggest that both personal context and organizational support significantly shape mental health openness in the tech industry.

Comparison of Models

```
results <- tibble(
  Model = c("Linear Regression", "Random Forest", "XGBoost"),
  RMSE = c(
    RMSE(lr_preds, test_data$willing_to_share),
    RMSE(rf_preds, test_data$willing_to_share),
    RMSE(xgb_preds, y_test)
  ),
  Rsquared = c(
    R2(lr_preds, test_data$willing_to_share),
    R2(rf_preds, test_data$willing_to_share),
    R2(xgb_preds, y_test)
  )
)

results
```

```
## # A tibble: 3 × 3
##   Model              RMSE Rsquared
##   <chr>             <dbl>    <dbl>
## 1 Linear Regression  2.73   0.0996
## 2 Random Forest      2.78   0.0712
## 3 XGBoost            2.95   0.0502
```

Model Performance – Key Insights Linear Regression performed best with the lowest RMSE (2.73) and highest R^2 (0.01), meaning it explained around 10% of the variance in willingness to share.

Random Forest followed closely (RMSE 2.78, R^2 0.07), offering better feature insights through its importance scores.

XGBoost had the weakest performance (RMSE 2.95, R^2 0.05), suggesting it may have overfit or lacked enough signal from the available features.

Overall, the low R^2 values across all models indicate that willingness to share is likely influenced by additional unmeasured psychological or environmental factors.

Skewness Check

```
skewness(train_data$willing_to_share)
```

```
## [1] -0.5920246
```

Insight The skewness value of -0.59 for the willing_to_share variable indicates a moderate left (negative) skew. This suggests that most respondents reported higher willingness to

share mental health issues, but a smaller group gave lower scores. While not severely skewed, the distribution is slightly unbalanced, which may impact linear model assumptions and suggests caution when interpreting regression-based results.

Methodology Overview

This project analyzed multi-year mental health survey data from the tech industry (2017–2023), sourced from the OSMI Kaggle repository. The goal was to investigate factors that influence an individual's willingness to discuss mental health challenges and to build predictive models around this behavior. After cleaning and standardizing the data—including harmonizing column names, filtering only tech company employees, and managing missing or inconsistent values—the dataset was enriched by creating a categorical version of the willingness score (willing_to_share_level) and imputing relevant fields like age and medical_coverage.

Three machine learning models were implemented to predict willing_to_share as a numeric outcome: Linear Regression, Random Forest, and XGBoost. Model performance was evaluated using RMSE and R^2 metrics. Linear Regression performed best with the lowest RMSE and highest R^2, suggesting that while the data can explain some variation in mental health openness, much of the behavior is influenced by unmeasured or external factors. Feature importance and coefficient plots revealed that personal mental health status, access to employer-provided resources, and country of residence were among the most influential predictors. Overall, the project achieved its objective of identifying meaningful patterns in workplace mental health disclosure and demonstrated how machine learning can assist in analyzing behavioral trends in organizational settings.

References

Osmihelp> (2024) Osmi: Datasets expert, Kaggle. Available at: https://www.kaggle.com/osmihelp/datasets (Accessed: 05 May 2025).

OECD (2017) Health at a glance 2017, Health at a Glance 2017. Available at: https://www.oecd.org/en/publications/health-at-a-glance-2017_health_glance-2017-en.html (Accessed: 05 May 2025).