



**5DATA006C**

**Data Visualization**

**PORTFOLIO**

**Module Leader:**

**Fouzul Hassan**

**Full Name:** Sithuli Nayanama Kothalwala

**Blackboard Name:** Sithuli Kothalawala

**IIT ID:** 20230336

**UOW ID:** 20521288

**UOW username :** w2052128

## Table of Contents

Research Question and Data Sourcing.....	3
Data Preparation.....	4
Tidiness Evaluation.....	4
Observations .....	4
Variables .....	4
Assessment of Tidiness.....	4
Tidy Dataset Transformation .....	5
Exploratory Data Analysis (EDA) .....	18
Univariate Analysis.....	18
Insights of Univariate Analysis.....	19
Bivariate Analysis.....	23
Multivariate Analysis.....	23
Insights of Bivariate and Multivariate Analysis.....	24
Data Storytelling .....	29
The Silent Struggle of Emma : A Story of Workplace Mental Health .....	29
Why this research was important? (Gap analysis ) .....	33
References.....	34

# Research Question and Data Sourcing

**Research Question :** “How does workplace culture influence employees seeking treatment for mental health issues?”

## **Relevance:**

Understanding the influence of workplace culture on mental health treatment decision is crucial in promoting supportive work environments. Mental health challenges are common among employees but often remain unaddressed due to workplace barriers, such as unsupportive co-workers, managers or inadequate leave policies. Addressing these factors can help organizations implement policies that encourage mental well-being, improve employee satisfaction, and enhance productivity.

## **Dataset used :**

The analysis utilizes the “*Mental Health in Tech Survey*” dataset provided by Open Sourcing Mental Illness on Kaggle. This dataset contains responses over 1200 individuals, primarily working in technological-relates fields. It includes information on demographics, workplace characteristics, and mental health-related experiences.

The dataset is ideal for exploring this research question as it captures variables related to workplace culture (e.g., support from co-workers and supervisors, leave policies, remote work policies ) and mental health treatment-seeking behaviour. For this study, key variables include whether employees sought treatment for mental health ( treatment ) and workplace-related factors (co-workers, supervisor, leave, benefits, remote work, work interference).

## **Reference:**

Open Sourcing Mental Illness, LTD (2014). *Mental Health in Tech Survey*. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey> [Accessed 30 Dec. 2024].

# Data Preparation

## Tidiness Evaluation

The dataset is particularly tidy, but improvements are needed.

### Observations

1. The dataset has 1260 entries and 27 columns
2. Columns include demographics details, treatment status, physical and mental health interview, and workplace related factors.
3. Some columns have irrelevant data, missing values, and format issues that need to be dealt with.

### Variables

Dependant variable - treatment

Independent -

- Demographic information : Age, Gender, Country
- Health Interview: mental\_health\_interview, seek\_help
- Workplace related factors : work\_interfere, remote\_work, coworkers, supervisor, leave, benefits )

### Assessment of Tidiness

Some columns, such as ‘comments’, have excessive missing values, making them less useful for analysis. Since this column is not directly relevant to the research question or the analysis it could be dropped.

Some columns have mixed data types (e.g, categorical values stored as strings )

Missing values in columns such as ‘self\_employed’ and ‘work\_interfere’ need handling.

The same word has been written in different ways which causes errors in analysis ( In Gender column Male and Female are written in many different ways and with typos.)

There is one duplicate row that need to be removed.

# Tidy Dataset Transformation

## 1. Removing irrelevant columns

This screenshot shows a Microsoft Excel spreadsheet titled "survey - Excel". The data consists of approximately 30 rows of responses. The first few rows are labeled with codes like "1 comments", "2 NA", etc. Rows 16 through 28 contain various statements such as "I'm not on my company's health insurance which could be part of the reason I answered Don't know to so many questions.", "I have chronic low-level neurological issues that have mental health side effects. One of my supervisors has also experienced similar neurological problems so I feel more comfortable being open about my issues than I would with someone without that experience.", and "Relatively new job. Ask again later". The "Ready" status bar at the bottom indicates there are 1260 cells in the sheet.

1	comments
2	NA
3	NA
4	NA
5	NA
6	NA
7	NA
8	NA
9	NA
10	NA
11	NA
12	NA
13	NA
14	NA
15	I'm not on my company's health insurance which could be part of the reason I answered Don't know to so many questions.
16	NA
17	I have chronic low-level neurological issues that have mental health side effects. One of my supervisors has also experienced similar neurological problems so I feel more comfortable being open about my issues than I would with someone without that experience.
18	My company does provide healthcare but not to me as I'm on a fixed-term contract. The mental healthcare I use is provided entirely outside of my work.
19	NA
20	NA
21	NA
22	NA
23	NA
24	NA
25	NA
26	Relatively new job. Ask again later
27	Sometimes I think about using drugs for my mental health issues. If I use drugs I feel better
28	NA
29	NA

## 2. Formatting no\_employees column

This formular is used to get the mid-point of ranges in the column. Phrases like 'More than 1000' are replaced with a numeric approximation like 1001. Values with neither range nor descriptor remain unchanged.

This screenshot shows the same Excel spreadsheet as above, but with a formula applied to the "no\_employees" column. The formula is =IF(ISNUMBER(FIND("-",J2)),(LEFT(J2,FIND("-",J2)-1)+MID(J2,FIND("-",J2)+1,LEN(J2)-FIND("-",J2))),SUBSTITUTE(J2,"More than","",1,J2))". This formula checks if the cell contains a dash, indicating a range. If it does, it extracts the first part of the range (before the dash) and adds the length of the dash plus one to it. If it doesn't, it replaces the phrase "More than" with an empty string and adds one to the result. The formula is applied to the entire column, changing values like "More than 1000" to "1001". The "Ready" status bar at the bottom indicates there are 1260 cells in the sheet.

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	Jun-25						
2	Female	United States	IL	NA	No	Yes	Often								
3	M	United States	IN	NA	No	No	Rarely	More than 1000							
4	Male	Canada	NA	NA	No	No	Rarely		Jun-25						
5	Male	United Kingdom	NA	NA	Yes	Yes	Often	26-100							
6	Male	United States	TX	NA	No	No	Never	100-500							
7	Male	United States	TN	NA	Yes	No	Sometimes		Jun-25						
8	Female	United States	MI	NA	Yes	Yes	Sometimes		01-May						
9	M	Canada	NA	NA	No	No	Never		01-May						
10	Female	United States	IL	NA	Yes	Yes	Sometimes	100-500							
11	Male	Canada	NA	NA	No	No	Never	26-100							
12	Male	United States	OH	NA	No	Yes	Sometimes		Jun-25						
13	male	Bulgaria	NA	NA	No	No	Never	100-500							
14	female	United States	CA	NA	Yes	Yes	Sometimes	26-100							
15	Male	United States	CT	NA	Yes	No	Never	500-1000							
16	Male	Canada	NA	NA	No	No	Never		Jun-25						
17	female	United States	IL	NA	Yes	Yes	Rarely	26-100							
18	Male	United Kingdom	NA	NA	No	Yes	Sometimes	26-100							
19	Male	United States	TN	NA	No	Yes	Sometimes		Jun-25						
20	male	United States	MD	Yes	Yes	No	Sometimes	01-May							
21	Male	France	NA	Yes	Yes	No	NA		Jun-25						
22	Male	United States	NY	No	Yes	Yes	Sometimes	100-500							
23	male	United States	NC	Yes	No	No	Never		01-May						
24	Male	United States	MA	No	No	Yes	Often	26-100							
25	Male	United States	IA	No	No	Yes	Never	More than 1000							
26	male	United States	CA	No	Yes	Yes	Rarely	26-100							
27	male	United States	TN	No	Yes	Yes	Sometimes	More than 1000							
28	male	United States	TN	No	No	No	NA	01-May							
29	Female	United States	CA	Min	Yes	Yes	Rarely		Jun-25						

survey - Excel

Average: 24428.59614 Count: 2519 Sum: 41797328

Then the values in the column are formatted into Numeric values as follows.

survey - Excel

Average: 16803.66322 Count: 1260 Sum: 21155812

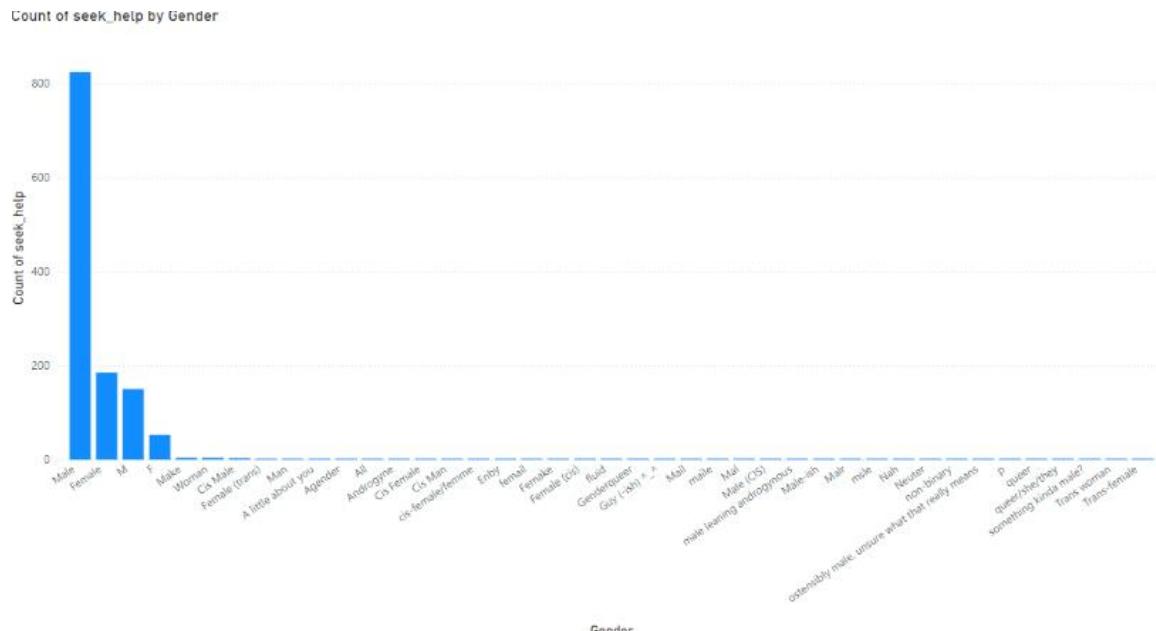
survey - Excel

Average: 16804 Count: 1260 Sum: 21155812

### 3. Cleaning the Gender column

In the Gender column Male was written in many different ways such as M, male and even with typos like make and Mal. Female was also written in many different ways such as female and femake. These were cleaned using Find & Replace dialogue box ( Ctrl + H ).

Other than that, there were invalid values which were manually replaced to closest value or assigned as ‘Other’.



survey - Excel

File Home Insert Page Layout Formulas Data Review View Help

Clipboard Cut Copy Format Painter

Font Calibri 11 A A

Wrap Text Alignment Conditional Formatting Table Styles Cell Insert Delete Format Cells Editing

General Number Styles Cells Editing

Find and Replace

All done. We made 150 replacements.

OK

Find what: M

Replace with: Male

Format Set

Replace All

Replace

Find All

Find Next

Close

Microsoft Excel

Count: 1260

Ready

survey - Excel

File Home Insert Page Layout Formulas Data Review View Help

Clipboard Cut Copy Format Painter

Font Calibri 11 A A

Wrap Text Alignment Conditional Formatting Table Styles Cell Insert Delete Format Cells Editing

General Number Styles Cells Editing

Find and Replace

All done. We made 971 replacements.

OK

Find what: male

Replace with: Male

Format Set

Replace All

Replace

Find All

Find Next

Close

Microsoft Excel

Count: 1260

Ready

survey - Excel

Find and Replace

All done. We made 53 replacements.

OK

Find what: F

Replace with: Male

Within: Sheet Match case

Search: By Rows Match entire cell contents

Look in: Formulas

Options <>

Replace All Replace Find All Find Next Close

Count: 1260

Ready

survey - Excel

Find and Replace

All done. We made 4 replacements.

OK

Find what: Woman

Replace with: Female

Within: Sheet Match case

Search: By Rows Match entire cell contents

Look in: Formulas

Options <>

Replace All Replace Find All Find Next Close

Count: 1260

Ready

survey - Excel

File Home Insert Page Layout Formulas Data Review View Help

Clipboard Cut Copy Format Painter

Font Calibri 11 A A Wrap Text General Conditional Formatting Table Cell Styles Insert Delete Format AutoSum Sort & Filter Select

C95 something kinda male?

A B C D E F G H I J K L M N O P Q R S

70 27-08-2014 11:52 31 Female United Sta NM No No No NA 126 Yes No Don't know, No Don't know, Don't know, Don't know, Maybe  
71 27-08-2014 11:52 34 Male United Sta NY Yes No No Rarely 45413 Yes Yes No No No No Don't know, Don't know, Maybe  
72 27-08-2014 11:53 28 Male France NA No No No Never 126 No Yes No No No No Don't know, Don't know, Maybe  
73 27-08-2014 11:53 34 Male Canada NA No No No Never 45809 No Yes Don't know, Not sure Don't know, Don't know, Don't know, Somewhat No  
74 27-08-2014 11:54 23 Trans-female United Sta MA No No No Rarely 1001 No Yes Yes No No Yes Somewhat Maybe  
75 27-08-2014 11:56 38 Male United Kin NA No No No NA 126 No Yes No No No No Don't know, Yes Somewhat No  
76 27-08-2014 11:56 33 Male United Sta CA No No No Never 1001 No Yes Don't know, Not sure Yes Don't know, Don't know, Don't know, Maybe  
77 27-08-2014 11:57 19 Male United Kin NA No No No NA 45413 No Yes No Yes No No No Don't know, Yes Somewhat No  
78 27-08-2014 11:57 25 Male United Sta WA No No No NA 1001 Yes Yes Yes Yes Yes Yes Don't know, Don't know, No  
79 27-08-2014 11:57 31 Male United Sta WA Yes Yes No Sometimes 45413 Yes Yes No No No No Yes Somewhat No  
80 27-08-2014 11:59 32 Male United Sta UT No Yes Yes Sometimes 126 Yes No No No No No Don't know, Somewhat Yes  
81 27-08-2014 12:00 28 Male Germany NA No No No Never 45809 Yes Yes Don't know, No Don't know, Don't know, Very easy, No  
82 27-08-2014 12:01 38 Male United Sta NY No Yes No Sometimes 600 Yes Yes Yes Yes Yes Yes Don't know, Don't know, No  
83 27-08-2014 12:02 23 Male United Kin NA No No No Never 126 Yes Yes Not sure Yes Yes Don't know, Very easy, No  
84 27-08-2014 12:02 30 Male Canada NA No No No Never 126 No No No No No No Don't know, Don't know, No  
85 27-08-2014 12:03 27 Cis Female United Sta NY Yes No Yes Offer 45413 Yes Yes Yes Yes Yes Yes Very easy, Maybe  
86 27-08-2014 12:03 33 Male United Sta CA No Yes No Never 1001 No Yes No No No No Don't know, Don't know, No  
87 27-08-2014 12:05 31 Male United Sta TX No No No NA 126 No Yes Don't know, Not sure No No Don't know, Don't know, No  
88 27-08-2014 12:07 39 Male United Kin NA Yes No Yes Often 45809 No Yes No No No No Don't know, Very diffi, Maybe  
89 27-08-2014 12:10 34 Female United Sta OR No Yes Yes Rarely 1500 Yes Yes Not sure No Don't know, Don't know, No  
90 27-08-2014 12:11 29 Female United Sta FL No No Yes Sometimes 126 No Yes Yes Yes No No Don't know, Don't know, Maybe  
91 27-08-2014 12:11 32 Male United Sta IL No No No NA 1500 No Yes Don't know, No Don't know, Don't know, No  
92 27-08-2014 12:12 31 Male United Sta NY No No No Never 1500 No Yes Yes Yes Yes Somewhat No  
93 27-08-2014 12:13 40 Male United Sta TX No No Yes Sometimes 126 Yes No Yes Yes Very diffi, No  
94 27-08-2014 12:14 34 Male United Sta OH No No No NA 126 No Yes Don't know, No No No Don't know, Somewhat Maybe  
95 27-08-2014 12:15 18 something kinda male? Russia NA No No No NA 126 Yes Yes Yes No Yes Somewhat No  
96 27-08-2014 12:15 25 Female Canada NA No No Yes Sometimes 126 No Yes Yes Yes No Don't know, Don't know, No  
97 27-08-2014 12:15 29 Male United Sta MN No No No Never 126 No Yes Don't know, No No Don't know, Don't know, No  
98 27-08-2014 12:15 24 Male United Sta MO No Yes No Rarely 126 No Yes Don't know, Not sure No Don't know, Somewhat Maybe

Ready survey

survey - Excel

File Home Insert Page Layout Formulas Data Review View Help

Clipboard Cut Copy Format Painter

Font Calibri 11 A A Wrap Text General Conditional Formatting Table Cell Styles Insert Delete Format AutoSum Sort & Filter Select

C388 Male

A B C D E F G H I J K L M N O P Q R S

388 27-08-2014 15:23 42 Male New Zeala NA No No No Never 1500 No Yes Don't know, Not sure No No No No Don't know, Don't know, No  
389 27-08-2014 15:24 29 Nah United Sta CA Yes Yes Yes Sometimes 45413 Yes Yes Yes Yes No No Don't know, Very diffi, Yes  
390 27-08-2014 15:24 25 Female United Sta CA No No Yes Sometimes 45809 No Yes Don't know, Not sure No No No No Don't know, Somewhat Yes  
391 27-08-2014 15:24 33 Female Sweden NA No Yes Yes Rarely 1001 No Yes Yes Yes Yes Yes Don't know, No  
392 27-08-2014 15:24 1E+11 All Zimbabwe NA Yes Yes Yes Often 45413 No Yes No Yes No No No Don't know, Maybe  
393 27-08-2014 15:24 40 Female United Sta PA No Yes Yes Rarely 1001 No No Yes No No No Don't know, Don't know, Somewhat Maybe  
394 27-08-2014 15:24 31 Male United Sta SC No No No Never 1001 No Yes Don't know, No No Don't know, Don't know, Somewhat Yes  
395 27-08-2014 15:24 26 Male Canada NA No Yes Yes Often 126 Yes Yes Don't know, No No Don't know, Don't know, No  
396 27-08-2014 15:24 24 Female United Sta TX No Yes No NA 1500 No No No No No No Very diffi, No  
397 27-08-2014 15:25 29 Male United Sta TX No No No Never 1001 No No Yes Not sure No Don't know, Don't know, No  
398 27-08-2014 15:25 48 Male United Kin NA No No No NA 126 No Yes Don't know, No No Don't know, Don't know, No  
399 27-08-2014 15:25 35 Male United Kin NA No No No Sometimes 45809 No Yes No No No Don't know, Very diffi, Maybe  
400 27-08-2014 15:25 32 Female United Sta AL No No No Never 600 No Yes Don't know, Not sure No Don't know, Don't know, No  
401 27-08-2014 15:25 29 Male United Sta TX No Yes No NA 1500 No Yes Don't know, Not sure Don't know, Don't know, Somewhat Maybe  
402 27-08-2014 15:26 26 Male United Sta TX No No No Never 1001 No No Yes Yes No No Don't know, Somewhat Maybe  
403 27-08-2014 15:26 28 Male United Sta TX No No No Never 126 No Yes Don't know, No No Don't know, Don't know, No  
404 27-08-2014 15:26 23 Male United Sta TX No Yes No NA 1500 No Yes Don't know, No No Don't know, Don't know, No  
405 27-08-2014 15:26 35 Male United Sta TX No Yes No NA 126 No Yes Don't know, No No Don't know, Very diffi, Maybe  
406 27-08-2014 15:27 29 Male United Sta TX No Yes No NA 1001 No Yes Don't know, No No Don't know, Very easy, Maybe  
407 27-08-2014 15:27 26 Male United Sta TX No Yes No NA 126 Yes Yes Yes Yes Yes Yes Don't know, No  
408 27-08-2014 15:27 33 Male United Sta TX No Yes No NA 1500 Yes Yes Yes Yes Yes Yes Don't know, Don't know, Yes  
409 27-08-2014 15:27 33 Male United Sta TX No Yes No NA 126 Yes Yes Yes Yes Yes Yes Don't know, No  
410 27-08-2014 15:28 22 Male United Sta TX No Yes No NA 1001 Yes Yes Don't know, No No No Don't know, Somewhat No  
411 27-08-2014 15:28 30 Female United Sta TX No Yes No NA 126 Yes Yes Don't know, No No No Don't know, Don't know, No  
412 27-08-2014 15:29 33 Male United Sta TX No Yes No NA 1500 Yes Yes Don't know, No No No Don't know, Somewhat Yes  
413 27-08-2014 15:29 31 Female United Sta TX No Yes No NA 126 Yes Yes Don't know, Not sure No Don't know, Very diffi, Yes  
414 27-08-2014 15:29 21 Male United Kin NA No Yes Yes Sometimes 1001 No Yes Don't know, Not sure Don't know, Don't know, No  
415 27-08-2014 15:30 31 Enby United Sta MA No Yes No Never 1001 No Yes Don't know, No No No Don't know, Don't know, No  
416 27-08-2014 15:30 26 Female United Sta TX No Yes No NA 600 Yes Yes Don't know, Not sure Don't know, Don't know, Somewhat No

Find and Replace

Find what: Nah

Replace with: Other

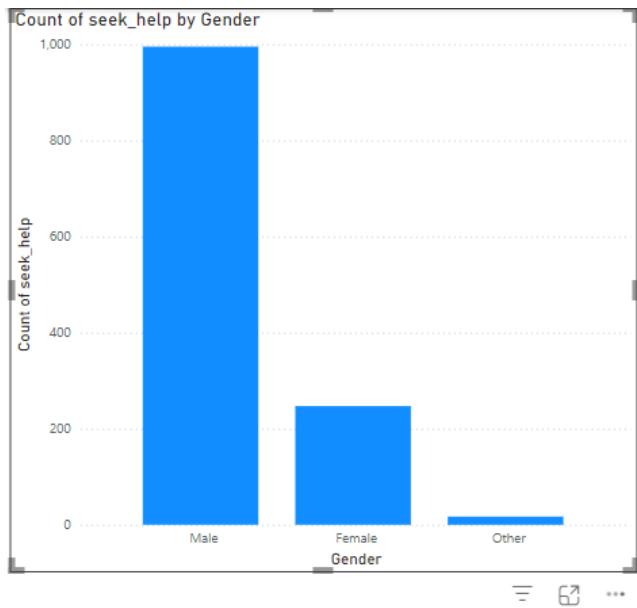
Within: Sheet Match entire cell contents

Search: By Rows Look in: Formulas

Options <<

Replace All Replace Find All Find Next Close Count: 1260

Ready survey



#### 4. Handling Missing Values and outliers

##### Handling missing values

The missing values in the columns, self\_employed and work\_interfere are replaced with mode which is 'No' in this case and the cleaned dataset is saved into a new .csv file using an R code.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Timestamp	Age	Gender	Country	state	self_employed	family_his_treatment	work_interfere	no_employees	remote_wtech	combenefits	care_optic	wellness	seek_help	leave	mental_health	phys_health	coworkers	supervisor	ment		
2	27-08-2014 11:29	37	Female	United Sta IL	NA	No	Yes	Often	45809	No	Yes	Yes	Yes	Yes	Yes	Yes	Some of t	Yes	No			
3	27-08-2014 11:29	44	Male	United Sta IN	NA	No	No	Rarely	1001	No	No	Don't know	No	Don't know	Don't know	Don't know	No	No	No	No	No	
4	27-08-2014 11:29	32	Male	Canada	NA	NA	No	No	Rarely	45809	No	Yes	No	No	No	No	Don't know	Somewhat No	No	Yes	Yes	
5	27-08-2014 11:29	31	Male	United Kin NA	NA	Yes	Yes	Often	126	No	Yes	No	Yes	No	No	No	Somewhat Yes	Yes	Some of t	No	Mayb	
6	27-08-2014 11:30	31	Male	United Sta TX	NA	No	No	Never	600	Yes	Yes	No	Don't know	Don't know	Don't know	Don't know	No	No	Some of t	Yes	Yes	
7	27-08-2014 11:31	33	Male	United Sta TN	NA	Yes	No	Sometime	45809	No	Yes	Yes	Not sure	No	No	Don't know	Don't know	No	No	Yes	Yes	No
8	27-08-2014 11:31	35	Female	United Sta MI	NA	Yes	Yes	Sometime	45413	Yes	Yes	No	No	No	No	No	Somewhat Maybe	Maybe	Some of t	No	No	
9	27-08-2014 11:32	39	Male	Canada	NA	No	No	Never	45413	Yes	Yes	No	Yes	No	No	Yes	Don't know	No	No	No	No	
10	27-08-2014 11:32	42	Female	United Sta IL	NA	Yes	Yes	Sometime	600	Yes	Yes	Yes	No	No	No	No	Very diffic	Maybe	No	Yes	Yes	
11	27-08-2014 11:32	23	Male	Canada	NA	NA	No	No	Never	126	No	Don't know	No	Don't know	Don't know	Don't know	No	No	Yes	Yes	Mayb	
12	27-08-2014 11:32	31	Male	United Sta OH	NA	No	Yes	Sometime	45809	Yes	Yes	Don't know	No	No	No	Don't know	Don't know	No	Some of t	Yes	No	
13	27-08-2014 11:32	29	Male	Bulgaria	NA	NA	No	No	Never	600	Yes	Don't know	Not sure	No	No	Don't know	Don't know	No	Yes	Yes	Yes	
14	27-08-2014 11:33	42	Female	United Sta CA	NA	Yes	Yes	Sometime	126	No	Yes	Yes	No	No	Don't know	Somewhat Yes	Yes	Yes	Yes	Mayb		
15	27-08-2014 11:33	36	Male	United Sta CT	NA	Yes	No	Never	1500	No	Yes	Don't know	Not sure	No	No	Don't know	Don't know	No	Yes	Yes	No	
16	27-08-2014 11:33	27	Male	Canada	NA	No	No	Never	45809	No	Yes	Don't know	Not sure	No	No	Don't know	Don't know	No	Some of t	Some of t	Mayb	
17	27-08-2014 11:34	29	Female	United Sta IL	NA	Yes	Yes	Rarely	126	No	Yes	Yes	Not sure	No	No	Don't know	Somewhat No	No	Yes	Some of t	Mayb	
18	27-08-2014 11:34	23	Male	United Kin NA	NA	No	Yes	Sometime	126	Yes	Yes	Don't know	No	Don't know	Don't know	Don't know	Very easy	Maybe	No	Some of t	No	Mayb
19	27-08-2014 11:34	32	Male	United Sta TN	NA	No	Yes	Sometime	45809	No	Yes	Yes	No	No	Don't know	Don't know	Don't know	Maybe	No	Some of t	Yes	No
20	27-08-2014 11:34	46	Male	United Sta MD	Yes	Yes	No	Sometime	45413	Yes	Yes	Not sure	Yes	Don't know	Yes	Very easy	No	No	Yes	Yes	No	
21	27-08-2014 11:35	36	Male	France	NA	Yes	Yes	No	NA	45809	Yes	Yes	No	Yes	No	Yes	Somewhat No	No	Some of t	Some of t	Mayb	
22	27-08-2014 11:35	29	Male	United Sta NY	No	Yes	Yes	Sometime	600	No	Yes	Yes	Yes	No	No	No	Somewhat Maybe	No	Some of t	Some of t	No	
23	27-08-2014 11:35	31	Male	United Sta NC	Yes	No	No	Never	45413	Yes	Yes	No	No	No	No	Yes	Somewhat No	No	Some of t	Some of t	No	
24	27-08-2014 11:35	46	Male	United Sta MA	No	No	Yes	Often	126	Yes	Yes	Yes	No	No	No	Don't know	Don't know	Maybe	No	Some of t	Yes	No
25	27-08-2014 11:36	41	Male	United Sta IA	No	No	Yes	Never	1001	No	Don't know	No	No	Don't know	Don't know	Don't know	Maybe	No	No	No	No	
26	27-08-2014 11:36	33	Male	United Sta CA	No	Yes	Yes	Rarely	126	No	Yes	Yes	Not sure	Don't know	Yes	Don't know	No	No	Yes	Yes	No	
27	27-08-2014 11:37	35	Male	United Sta TN	No	Yes	Yes	Sometime	1001	No	Yes	Yes	No	Don't know	No	Very easy	Yes	No	Some of t	Yes	No	
28	27-08-2014 11:37	33	Male	United Sta TN	No	No	No	NA	45413	No	Yes	Don't know	Not sure	No	No	Don't know	Don't know	Don't know	Maybe	Some of t	No	
29	27-08-2014 11:37	25	Female	United Sta CA	No	Yes	Yes	Yes	Response	45809	Yes	Yes	Yes	Yes	Don't know	Don't know	Don't know	No	No	Yes	Yes	

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source on Save Go to file/function Addins
Multiple linear regression.R Untitled1* Untitled2* Untitled3* Untitled4* Untitled5* Untitled6* Run Source
1 # Install and load required packages
2 install.packages("tidyverse")
3
4 library(dplyr) # For data manipulation
5 library(tidyr) # For handling missing values
6 library(ggplot2) # For visualization
7
8 # Load the dataset
9 data <- read.csv("D:\\Data Visualization\\final portfolio\\survey.csv", stringsAsFactors = FALSE)
10
11 # Check for missing values in the dataset
12 summary(data) # Quick summary to identify NA values
13 colSums(is.na(data)) # Count missing values per column
14
15 # Handling Missing Values
16 # 1. Remove rows with missing values (if data loss is acceptable)
17 data_cleaned <- na.omit(data)
18
19 # 2. Replace missing values in numerical columns with the column mean
20 data <- data %>%
21   mutate(across(where(is.numeric), ~ ifelse(is.na(), mean(., na.rm = TRUE), .)))
22
23 (Top Level) R Script

```

Console Terminal Background Jobs

```

R 4.4.1 - /-/ family_history treatment work_interfere
no_employees remote_work tech_company
benefits care_options wellness_program
seek_help anonymity leave
mental_health_consequence phys_health_consequence coworkers
supervisor mental_health_interview phys_health_interview
mental_vs_physical obs_consequence
> | 

```

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source on Save Go to file/function Addins
Multiple linear regression.R Untitled1* Untitled2* Untitled3* Untitled4* Untitled5* Untitled6* Run Source
15 # Handling Missing Values
16 # 1. Remove rows with missing values (if data loss is acceptable)
17 data_cleaned <- na.omit(data)
18
19 # 2. Replace missing values in numerical columns with the column mean
20 data <- data %>%
21   mutate(across(where(is.numeric), ~ ifelse(is.na(), mean(., na.rm = TRUE), .)))
22
23 # 3. Replace missing values in categorical columns with the mode
24 replace_mode <- function(x) {
25   if (is.character(x) || is.factor(x)) {
26     mode_value <- names(which.max(table(x, useNA = "no")))
27     x[is.na(x)] <- mode_value
28   }
29   return(x)
30 }
31 data <- data %>%
32   mutate(across(where(is.character), replace_mode))
33
34 # Verify that missing values are handled
35 rm(data)
36
325 (Top Level) R Script

```

Console Terminal Background Jobs

```

R 4.4.1 - /-/ # Remove rows with missing values (if data loss is acceptable)
# data_cleaned <- na.omit(data)
# 2. Replace missing values in numerical columns with the column mean
# data <- data %>%
#   mutate(across(where(is.numeric), ~ ifelse(is.na(), mean(., na.rm = TRUE), .)))
# 3. Replace missing values in categorical columns with the mode
replace_mode <- function(x) {
  if (is.character(x) || is.factor(x)) {
    mode_value <- names(which.max(table(x, useNA = "no")))
    x[is.na(x)] <- mode_value
  }
  return(x)
}
data <- data %>%
  mutate(across(where(is.character), replace_mode))
> | 

```

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. Below the menu is a toolbar with various icons. The main area contains an R script titled "Multiple linear regression.R". The script includes several code snippets for handling missing values, such as removing rows with missing values, replacing numerical columns with their mean, and replacing categorical columns with their mode. A specific section is highlighted in blue:

```

16 # 1. Remove rows with missing values (if data loss is acceptable)
17 data_cleaned <- na.omit(data)
18
19 # 2. Replace missing values in numerical columns with the column mean
20 data <- data %>%
21   mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))
22
23 # 3. Replace missing values in categorical columns with the mode
24 replace_mode <- function(x) {
25   if (is.character(x) || is.factor(x)) {
26     mode_value <- names(which.max(table(x, useNA = "no")))
27     x[is.na(x)] <- mode_value
28   }
29   return(x)
30 }
31 data <- data %>%
32   mutate(across(where(is.character), replace_mode))
33
34 # Verify that missing values are handled
35 colSums(is.na(data))
36
37 <--
```

The status bar at the bottom indicates "35.21 (Top Level)". To the right of the script is a "Data" pane titled "msleep\_clean" showing a table with 56 observations and 2 variables. The table includes columns like "carsByEngine", "colours", "invalid\_rows", "invalid\_timestamps", "labels", "platform", and "platforms". Below the Data pane are tabs for Files, Plots, Packages, Help, Viewer, and Presentation.

The screenshot shows a Microsoft Excel spreadsheet titled "cleaned\_survey". The table has approximately 30 columns and 200 rows of data. The columns include timestamp, age, gender, country, state, and various survey responses. The data spans from row 1 to row 200, with columns labeled A through V. The Excel ribbon at the top includes File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Help. The "Home" tab is selected, showing standard tools for cutting, pasting, and formatting. The "Data" tab is also visible in the ribbon.

## Handling outliers

In the Age column of the dataset there are a couple of outliers as follows. There are extremely unrealistic values and even negative values. Other than that, since this is about a work environment ages like 5 are also invalid. Therefore, ages above 65 and below 18 are considered outliers and are removed.

File		Home	Insert	Page Layout	Formulas	Data	Review	View	Help																		
Calibri												11	A	A'	Wrap Text	General	Conditional	Format as	Cell Styles	Insert	Delete	Format	AutoSum	Fill	Sort & Filter	Clear	Find & Select
Paste		Cut		Copy		Format Painter		B	I	U	V	W	X	Y	Z	Y	Find	Share									
B392	:	x	f	fx	9999999999																						
379	27-08-2014	15:22	25	Female	United Sta WA	No	Yes	Yes	Sometime:	126 Yes	Yes	Don't know	No	No	No	Don't know	Don't know	Maybe	No	Some of t	Yes	N					
380	27-08-2014	15:22	34	Female	United Sta NY	No	Yes	Yes	Rarely	1001 No	No	Yes	Yes	Yes	Yes	Yes	Very diffic	Maybe	Maybe	Some of t	Some of t	N					
381	27-08-2014	15:23	26	Female	United Sta CA	No	Yes	Yes	Never	1001 No	Yes	Yes	Yes	No	Yes	Yes	Don't know	Maybe	No	No	No	N					
382	27-08-2014	15:23	41	Male	Canada NA	No	No	Yes	Never	1500 No	Yes	Yes	Not sure	No	Don't know	Don't know	Somewhat No	No	Some of t	Some of t	N						
383	27-08-2014	15:23	27	Male	United Sta CA	No	No	Yes	Rarely	1001 No	Yes	Yes	No	Yes	Yes	Don't know	No	No	Some of t	Yes	N						
384	27-08-2014	15:23	31	Male	United Sta IN	No	Yes	Yes	Sometime:	126 Yes	Yes	Yes	No	No	No	Don't know	Don't know	No	Yes	Yes	N						
385	27-08-2014	15:23	25	Male	Germany NA	No	Yes	Yes	Often	126 Yes	Yes	No	No	No	Yes	Somewhat No	No	Yes	Yes	N							
386	27-08-2014	15:23	26	Male	United Sta NV	No	No	Yes	Sometime:	4580 No	Yes	Don't know	No	No	No	Don't know	Don't know	Yes	Some of t	No	N						
387	27-08-2014	15:23	27	Female	United Sta CO	No	Yes	Yes	Rarely	1001 Yes	Yes	Yes	Yes	No	Don't know	Don't know	Maybe	Yes	Some of t	No	N						
388	27-08-2014	15:23	42	Male	New Zeala NA	No	No	No	Never	1500 No	Yes	Don't know	Not sure	No	Don't know	Don't know	No	No	Some of t	Yes	N						
389	27-08-2014	15:24	29	Other	United Sta CA	Yes	Yes	Yes	Sometime:	45413 Yes	Yes	Yes	Yes	No	No	Don't know	Very diffic	No	Some of t	No	N						
390	27-08-2014	15:24	25	Female	United Sta CA	No	No	Yes	Sometime:	45809 No	Yes	Don't know	Not sure	No	Don't know	Somewhat Yes	No	Some of t	No	N							
391	27-08-2014	15:24	33	Female	Sweden NA	No	Yes	Yes	Rarely	1001 No	Yes	Yes	Yes	Yes	Yes	Don't know	Maybe	No	Some of t	No	N						
392	27-08-2014	15:24	9999999999	Other	Zimbabwe NA	Yes	Yes	Yes	Often	45413 No	Yes	No	No	No	No	Very diffic	Yes	No	No	Yes	N						
393	27-08-2014	15:24	40	Female	United Sta PA	No	Yes	Yes	Rarely	1001 No	No	Yes	No	Don't know	Don't know	Somewhat Maybe	Maybe	No	No	No	N						
394	27-08-2014	15:24	31	Male	United Sta SC	No	No	No	Never	1001 No	Yes	Don't know	No	No	Don't know	Don't know	Somewhat Yes	No	Some of t	No	N						
395	27-08-2014	15:24	26	Male	Canada NA	No	Yes	Yes	Often	126 Yes	Yes	Don't know	No	Don't know	Don't know	Don't know	Very easy	No	No	Yes	N						
396	27-08-2014	15:24	24	Female	United Sta TX	No	Yes	No	NA	1500 No	No	No	No	No	No	Very diffic	Yes	Maybe	No	No	N						
397	27-08-2014	15:25	29	Male	United Sta TX	No	No	No	Never	1001 No	Yes	Not sure	No	Don't know	Don't know	Don't know	Maybe	No	Some of t	Some of t	N						
398	27-08-2014	15:25	48	Male	United Kin NA	No	No	No	NA	126 No	Yes	Don't know	No	No	No	Don't know	Don't know	Maybe	No	No	N						
399	27-08-2014	15:25	35	Male	United Kin NA	No	No	No	Sometime:	45809 No	Yes	No	No	No	No	Don't know	Very diffic	Maybe	Some of t	Yes	N						
400	27-08-2014	15:25	32	Female	United Sta AL	No	No	No	Never	600 No	Yes	Don't know	Not sure	No	Don't know	Don't know	Maybe	No	No	Some of t	N						
401	27-08-2014	15:25	29	Male	Canada NA	No	No	Yes	Sometime:	600 No	Yes	Don't know	Not sure	No	Don't know	Don't know	Somewhat Maybe	No	Some of t	Some of t	N						
402	27-08-2014	15:26	26	Male	United Sta OR	No	Yes	Yes	Often	45809 No	Yes	Yes	Yes	No	No	Don't know	Somewhat Maybe	No	Some of t	Yes	N						
403	27-08-2014	15:26	28	Male	United Sta NJ	No	Yes	Yes	Sometime:	126 No	Yes	Yes	Yes	No	No	Don't know	Don't know	No	Some of t	No	N						
404	27-08-2014	15:26	23	Male	United Sta CA	No	No	No	NA	1001 No	Yes	Don't know	No	No	No	Don't know	Don't know	No	Some of t	Yes	N						
405	27-08-2014	15:26	35	Male	United Sta CA	No	Yes	Yes	Sometime:	600 No	Yes	Yes	Yes	No	No	No	Don't know	Very diffic	Maybe	Some of t	Some of t	N					
406	27-08-2014	15:27	29	Male	Germany NA	No	Yes	Yes	Sometime:	126 No	Yes	Don't know	No	No	No	Don't know	Very easy	Maybe	No	Some of t	Some of t	N					
407	27-08-2014	15:27	26	Male	United Sta CA	No	Yes	No	NA	45800 No	Yes	Yes	Yes	No	Yes	Very easy	No	No	Some of t	Yes	N						

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source on Save Go to file/function Addins
Multiple linear regression.R Untitled1* Untitled2* Untitled3* Untitled4* Untitled5* Untitled6* Source
35 colSums(is.na(data))
36
37 # Detect and Handle Outliers in the "Age" Column
38 # Visualize the outliers with a boxplot
39 ggplot(data, aes(x = "", y = Age)) +
40   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
41   labs(title = "Boxplot for Age with Outliers", x = "Age", y = "") +
42   theme_minimal()
43
44 # Handle Outliers: Remove or Cap Ages Outside 18-65
45 data <- data %>%
46   mutate(Age = ifelse(Age < 18, 18, ifelse(Age > 65, 65, Age)))
47
48 # Re-plot the boxplot after handling outliers
49 ggplot(data, aes(x = "", y = Age)) +
50   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
51   labs(title = "Boxplot for Age (Outliers Capped to 18-65)", x = "Age", y = "") +
52   theme_minimal()
53
54 # Save the cleaned dataset
55 write.csv(data, "D:\\Data Visualization\\final portfolio\\cleaned_survey.csv", row.names = FALSE)
56
42:18 (Top Level) R Script

```

Console Terminal Background Jobs

```

R - R 4.4.1 - ~/r
seek_help anonymity leave
0 0 0
mental_health_consequence phys_health_consequence coworkers
0 0 0
supervisor mental_health_interview phys_health_interview
0 0 0
mental_vs_physical obs_consequence
0 0
> # Detect and Handle Outliers in the "Age" Column
> # Visualize the outliers with a boxplot
> ggplot(data, aes(x = "", y = Age)) +
+   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
+   labs(title = "Boxplot for Age with Outliers", x = "Age", y = "") +
+   theme_minimal()
43
44 # Handle Outliers: Remove or Cap Ages Outside 18-65
45 data <- data %>%
46   mutate(Age = ifelse(Age < 18, 18, ifelse(Age > 65, 65, Age)))
47
48 # Re-plot the boxplot after handling outliers
49 ggplot(data, aes(x = "", y = Age)) +
50   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
51   labs(title = "Boxplot for Age (Outliers Capped to 18-65)", x = "Age", y = "") +
52   theme_minimal()
53
54 # Save the cleaned dataset
55 write.csv(data, "D:\\Data Visualization\\final portfolio\\cleaned_survey.csv", row.names = FALSE)
47:1 (Top Level) R Script

```

Environment History Connections Tutorial

Project: (None)

Global Environment

```

colours chr [1:2] "lightblue1" "lemonchiffon"
invalid_rows int [1:188] 244 245 246 247 248 249 250 251 252 253 ...
invalid_timestamps int [1:188] 244 245 246 247 248 249 250 251 252 253 ...
labels chr [1:2] "56%" "44%"
platform "flikTok"
Functions cap_outliers function (x)
detect_outliers function (x)
replace_mode function (x)

```

Files Plots Packages Help Viewer Presentation

Boxplot for Age with Outliers

Y-axis: 0.0e+00, 2.5e+10, 5.0e+10, 7.5e+10, 1.0e+11. X-axis: Age.

RStudio

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Source on Save Go to file/function Addins
Multiple linear regression.R Untitled1* Untitled2* Untitled3* Untitled4* Untitled5* Untitled6* Source
35 colSums(is.na(data))
36
37 # Detect and Handle Outliers in the "Age" Column
38 # Visualize the outliers with a boxplot
39 ggplot(data, aes(x = "", y = Age)) +
40   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
41   labs(title = "Boxplot for Age with Outliers", x = "Age", y = "") +
42   theme_minimal()
43
44 # Handle Outliers: Remove or Cap Ages Outside 18-65
45 data <- data %>%
46   mutate(Age = ifelse(Age < 18, 18, ifelse(Age > 65, 65, Age)))
47
48 # Re-plot the boxplot after handling outliers
49 ggplot(data, aes(x = "", y = Age)) +
50   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
51   labs(title = "Boxplot for Age (Outliers Capped to 18-65)", x = "Age", y = "") +
52   theme_minimal()
53
54 # Save the cleaned dataset
55 write.csv(data, "D:\\Data Visualization\\final portfolio\\cleaned_survey.csv", row.names = FALSE)
56
47:1 (Top Level) R Script

```

Console Terminal Background Jobs

```

R - R 4.4.1 - ~/r
mental_health_consequence phys_health_consequence coworkers
0 0 0
supervisor mental_health_interview phys_health_interview
0 0 0
mental_vs_physical obs_consequence
0 0
> # Detect and Handle Outliers in the "Age" Column
> # Visualize the outliers with a boxplot
> ggplot(data, aes(x = "", y = Age)) +
+   geom_boxplot(outlier.colour = "red", outlier.shape = 8) +
+   labs(title = "Boxplot for Age with Outliers", x = "Age", y = "") +
+   theme_minimal()
> # Handle Outliers: Remove or Cap Ages Outside 18-65
> data <- data %>%
+   mutate(Age = ifelse(Age < 18, 18, ifelse(Age > 65, 65, Age)))
47

```

Environment History Connections Tutorial

Project: (None)

Global Environment

```

data 1259 obs. of 26 variables
data_binary 481 obs. of 36 variables
data_cleaned 591 obs. of 26 variables
data_expanded 481 obs. of 29 variables
msleep 83 obs. of 13 variables
msleep_clean 56 obs. of 2 variables
outliers 0 obs. of 26 variables

```

Values

```

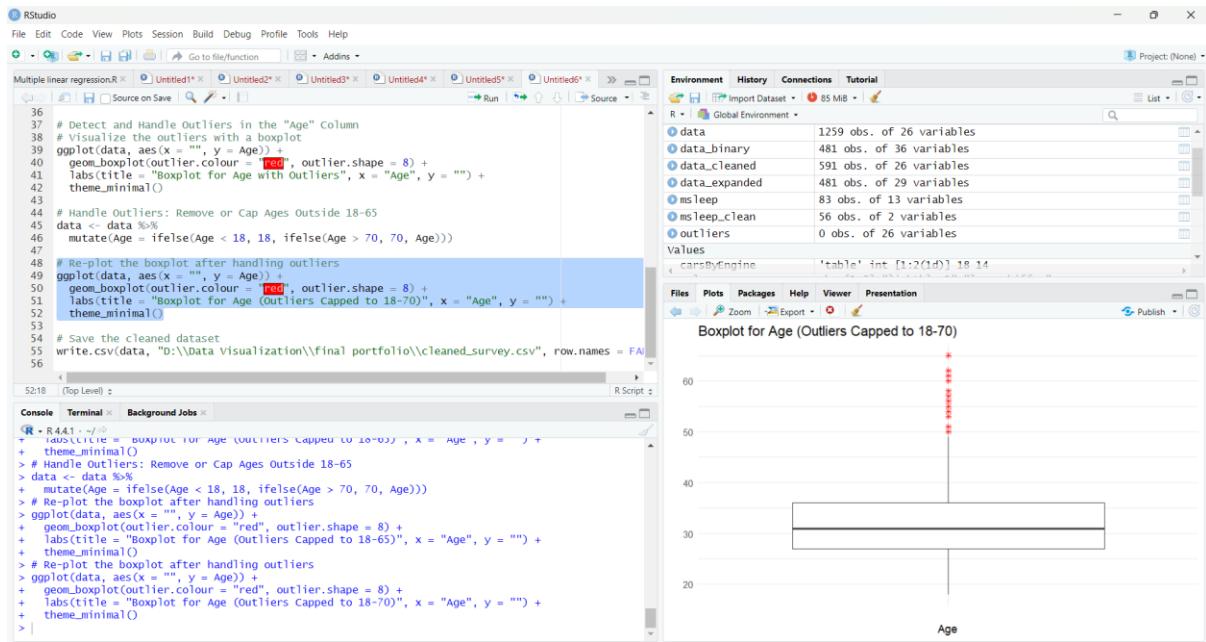
caranyaEngine 'table' int [1:2(3d)] 18 14

```

Files Plots Packages Help Viewer Presentation

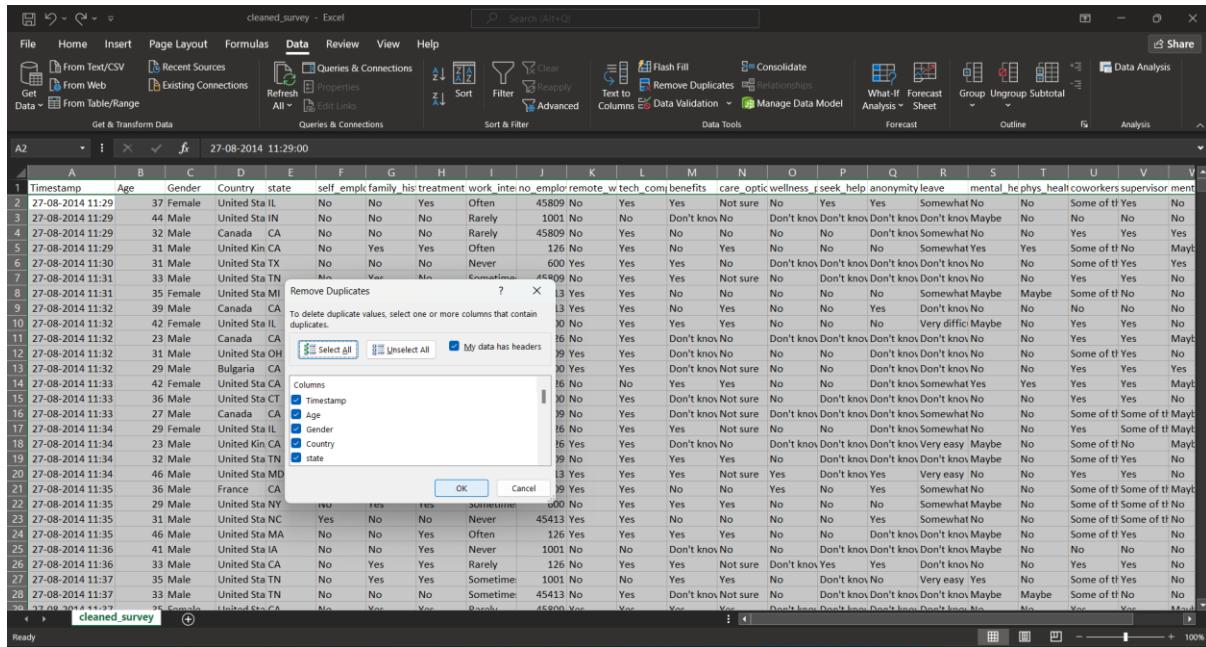
Boxplot for Age with Outliers

Y-axis: 0.0e+00, 2.5e+10, 5.0e+10, 7.5e+10, 1.0e+11. X-axis: Age.



## 5. Removing duplicates

Row number 861 and 862 are duplicated with exact same values. Therefore, one column is removed using ‘Remove duplicates’ option.



cleaned\_survey - Excel

The screenshot shows an Excel spreadsheet titled "cleaned\_survey". The Data tab is selected, displaying various data analysis tools like Sort, Filter, Flash Fill, and Consolidate. A dialog box is open in the center of the screen, showing a warning message: "1 duplicate values found and removed: 1258 unique values remain." There is an "OK" button at the bottom right of the dialog.

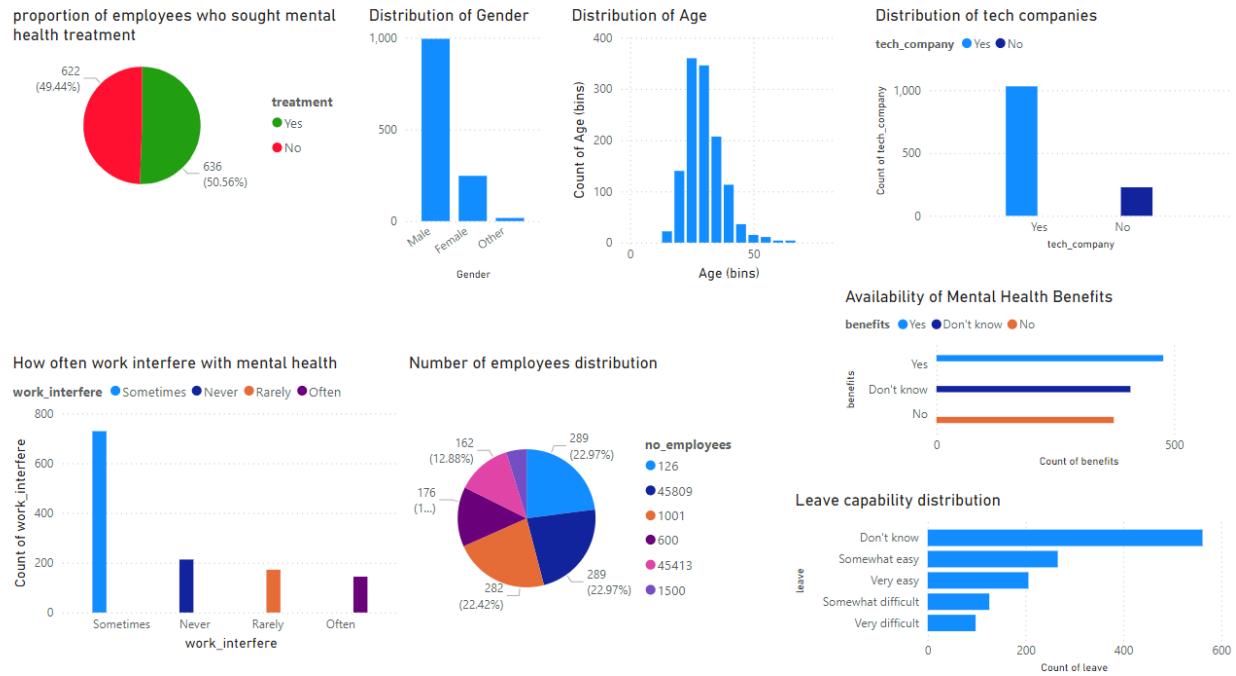
1	Timestamp	Age	Gender	Country	state	self_emplo family	his_treatment	work_inte_no	emplo_remote	w_tech	combenefits	care_optic	wellness	p_seek	help_anonymity	leave	mental	he_phys	health	coworkers	supervisor	ment								
2	27-08-2014 11:29	37	Female	United Sta IL	No	No	Yes	Often	45809	No	Yes	Not sure	No	Yes	Yes	Somewhat No	No	Some of t!	Yes	No	No	No								
3	27-08-2014 11:29	44	Male	United Sta IN	No	No	No	Rarely	1001	No	Don't know	No	Don't know	Don't know	Don't know	Don't know	Don't know	No	No	No	No	No								
4	27-08-2014 11:29	32	Male	Canada CA	No	No	No	Rarely	45809	No	Yes	No	No	No	No	Don't know	Somewhat No	No	Yes	Yes	Yes	Yes								
5	27-08-2014 11:29	31	Male	United Kin CA	No	Yes	Yes	Often	126	No	Yes	No	Yes	No	No	No	Somewhat Yes	Yes	Some of t!	No	Mayb	No								
6	27-08-2014 11:30	31	Male	United Sta TX	No	No	No	Never	600	Yes	Yes	No	Don't know	Don't know	Don't know	Don't know	Don't know	No	Some of t!	Yes	Yes	Yes								
7	27-08-2014 11:31	33	Male	United Sta TN	No	Yes	No	Sometime:	45809	No	Yes	Yes	Not sure	No	Don't know	Don't know	Don't know	No	Yes	Yes	No	No								
8	27-08-2014 11:31	35	Female	United Sta MI	No	Yes	Yes	Sometime:	45413	Yes	Yes	No	No	No	No	No	Somewhat Maybe	Maybe	Some of t!	No	No									
9	27-08-2014 11:32	39	Male	Canada CA	No	No	No	No	Microsoft Excel																					
10	27-08-2014 11:32	42	Female	United Sta IL	No	Yes	Yes	No	1 duplicate values found and removed: 1258 unique values remain.																					
11	27-08-2014 11:32	23	Male	Canada CA	No	No	No	No	OK																					
12	27-08-2014 11:32	31	Male	United Sta OH	No	No	Yes	No	Don't know																					
13	27-08-2014 11:32	29	Male	Bulgaria	CA	No	No	No	Don't know																					
14	27-08-2014 11:33	42	Female	United Sta CA	No	Yes	Yes	No	Very difficil																					
15	27-08-2014 11:33	36	Male	United Sta CT	No	Yes	No	Never	1500	No	Yes	Don't know	Not sure	No	No	Don't know	Don't know	No	Yes	Yes	No	No	No	No	No	No	No			
16	27-08-2014 11:33	27	Male	Canada CA	No	No	No	Never	45809	No	Yes	Don't know	Not sure	Don't know	Don't know	Don't know	Don't know	No	Yes	Yes	No	Some of t!	Some of t!	Mayb	No	No	No			
17	27-08-2014 11:34	29	Female	United Sta IL	No	Yes	Yes	Rarely	126	No	Yes	Not sure	No	No	No	Don't know	Somewhat No	No	Yes	Some of t!	Yes	No	No	No	No	No	No			
18	27-08-2014 11:34	23	Male	United Kin CA	No	No	Yes	Sometime:	126	Yes	Yes	Don't know	No	Don't know	Don't know	Don't know	Don't know	No	Some of t!	No	Mayb	No	No	No	No	No	No	No		
19	27-08-2014 11:34	32	Male	United Sta TN	No	No	Yes	Sometime:	45809	No	Yes	Yes	No	Don't know	Don't know	Don't know	Don't know	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No		
20	27-08-2014 11:34	46	Male	United Sta MD	Yes	Yes	No	Sometime:	45413	Yes	Yes	Not sure	Yes	Don't know	Yes	Very easy	No	No	Yes	Yes	No	Yes	Yes	No	No	No	No	No	No	
21	27-08-2014 11:35	36	Male	France CA	Yes	Yes	No	Sometime:	45809	Yes	Yes	No	No	Yes	No	Yes	Somewhat No	No	Some of t!	Some of t!	Mayb	No	No	No	No	No	No	No	No	No
22	27-08-2014 11:35	29	Male	United Sta NY	No	Yes	Yes	Sometime:	600	No	Yes	Yes	No	No	No	No	Somewhat Maybe	No	Some of t!	Some of t!	No	No	No	No	No	No	No	No	No	
23	27-08-2014 11:35	31	Male	United Sta NC	Yes	No	No	Never	45413	Yes	No	No	No	No	Yes	Somewhat No	No	Some of t!	Some of t!	No	No	No	No	No	No	No	No	No		
24	27-08-2014 11:35	46	Male	United Sta MA	No	No	Yes	Often	126	Yes	Yes	Yes	No	No	No	Don't know	Don't know	Maybe	Yes	Some of t!	Yes	No	No	No	No	No	No	No	No	
25	27-08-2014 11:36	41	Male	United Sta IA	No	No	Yes	Never	1001	No	Don't know	No	Don't know	Don't know	Don't know	Don't know	Don't know	No	No	No	No	No	No	No	No	No	No	No		
26	27-08-2014 11:36	33	Male	United Sta CA	No	Yes	Yes	Rarely	126	No	Yes	Not sure	Don't know	Yes	Yes	Don't know	No	No	Yes	Yes	No	Yes	Yes	No	No	No	No	No	No	
27	27-08-2014 11:37	35	Male	United Sta TN	No	Yes	Yes	Sometime:	1001	No	Yes	Yes	No	Don't know	Yes	Very easy	Yes	No	Some of t!	Yes	No	No	No	No	No	No	No	No	No	
28	27-08-2014 11:37	33	Male	United Sta TN	No	No	No	Sometime:	45413	No	Yes	Don't know	Not sure	No	Don't know	Don't know	Don't know	Maybe	Some of t!	No	No	No	No	No	No	No	No	No	No	
29	27-08-2014 11:37	25	Female	United Sta CA	No	Yes	Yes	Yes	45809	Yes	Yes	Yes	Yes	Yes	Yes	Don't know	Don't know	Don't know	Don't know	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

# Exploratory Data Analysis (EDA)

## Univariate Analysis

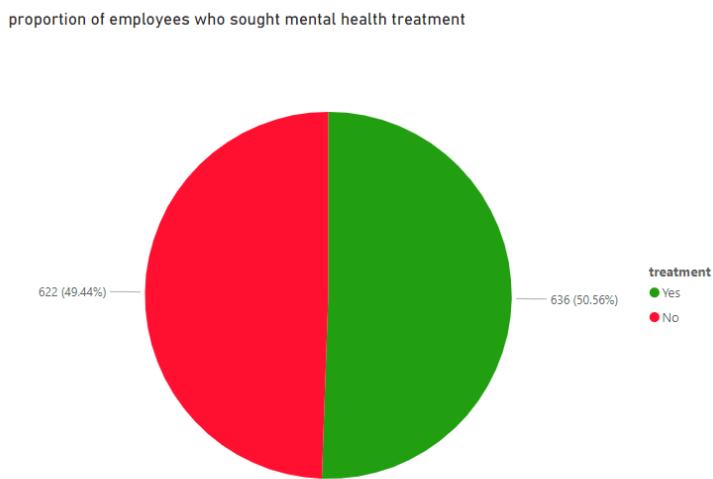
Univariate analysis examines the distribution of individual variables to understand their central tendency, dispersion and shape.

The variables in this dataset include categorical (e.g., Gender, Country, treatment, work\_interfere) and numerical variables (eg., no\_employees)



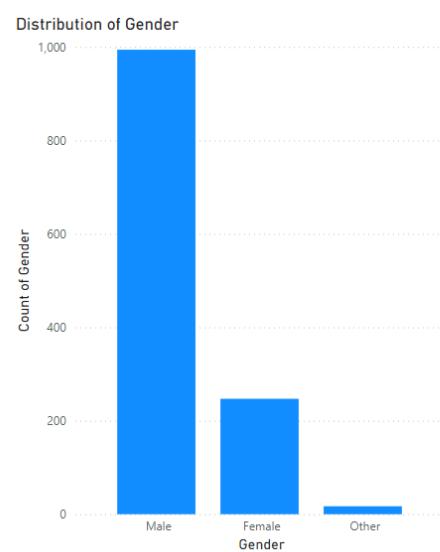
## Insights of Univariate Analysis

### I. Proportion of Employees who sought Mental Health Treatment (treatment)



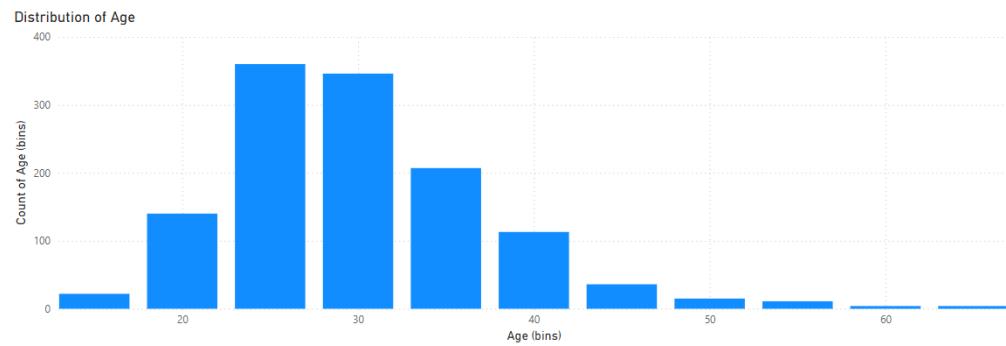
**Insights :** 49.44% of respondents sought mental health treatment while 50.56 did not. This indicates a fairly even split, highlighting that mental health issues are a significant concern among employees.

### II. Distribution of Gender



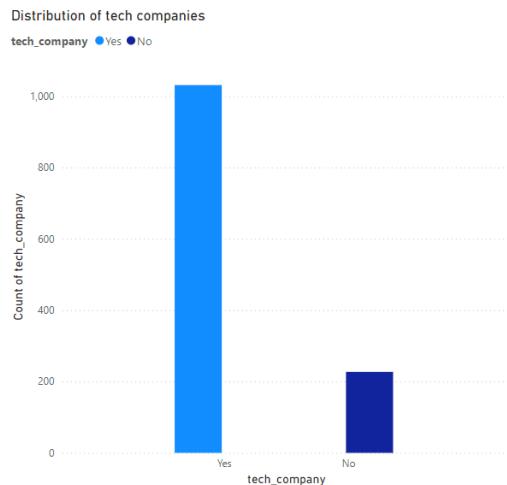
**Insights :** The majority of respondents identify as male. This distribution suggests a gender imbalance, which may influence how mental health issues are perceived in the workplace

### III. Distribution of Age



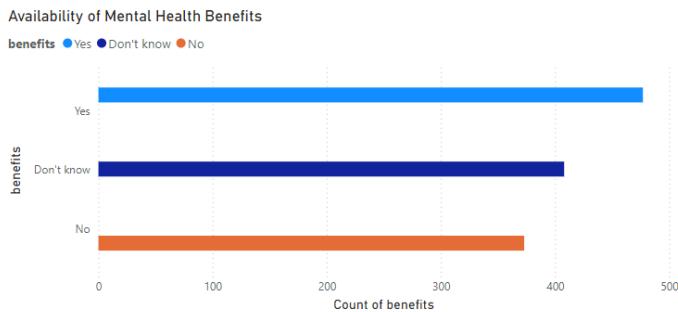
**Insights :** The average age of respondents is approximately 32 with most employees falling between the ages of 27 and 36 indicating that mental health issues are prevalent among younger professionals.

### IV. Distribution of Tech Companies (tech\_company)



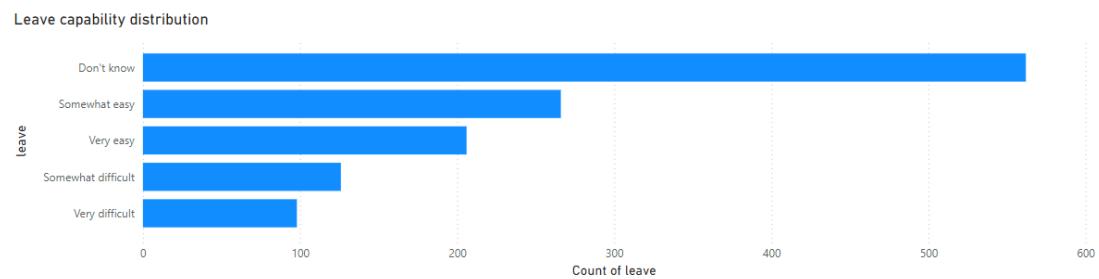
**Insights :** A significant portion of respondents work in tech companies aligning with the research focus.

## V. Availability of Mental Health Benefits ( benefits )



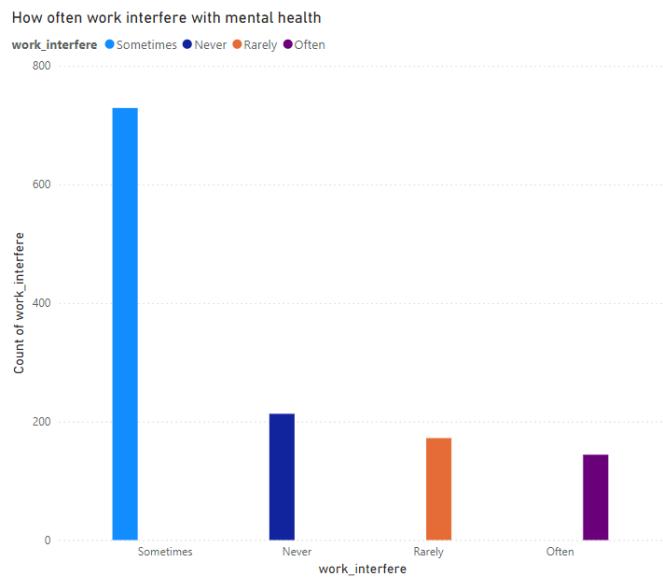
**Insights :** 47.7% respondents are provided with mental health benefits while a notable portion do not. This suggests that awareness and availability of mental health support could be improved.

## VI. Leave Capability Distribution



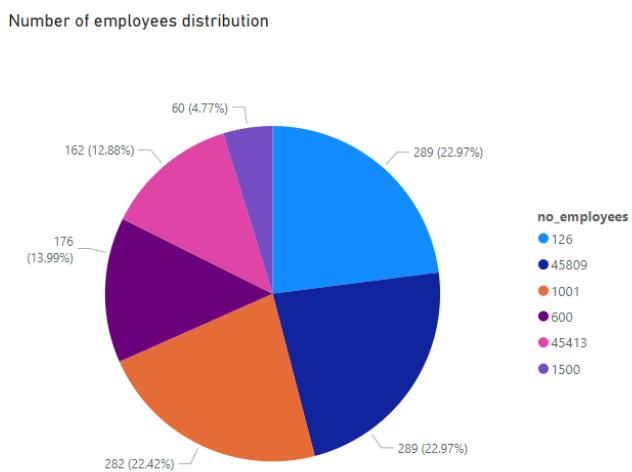
**Insights :** Many respondents indicated that taking leave for mental health reasons as “ Don’t know ” and “ Somewhat difficult”. This points to a potential stigma around taking a leave.

## VII. How often Work Interferes with Mental Health ( work\_interfere )



**Insights :** The majority reported that work “Sometimes” interfere with their mental health highlighting the need for workplace policies to be better accommodate mental well-being.

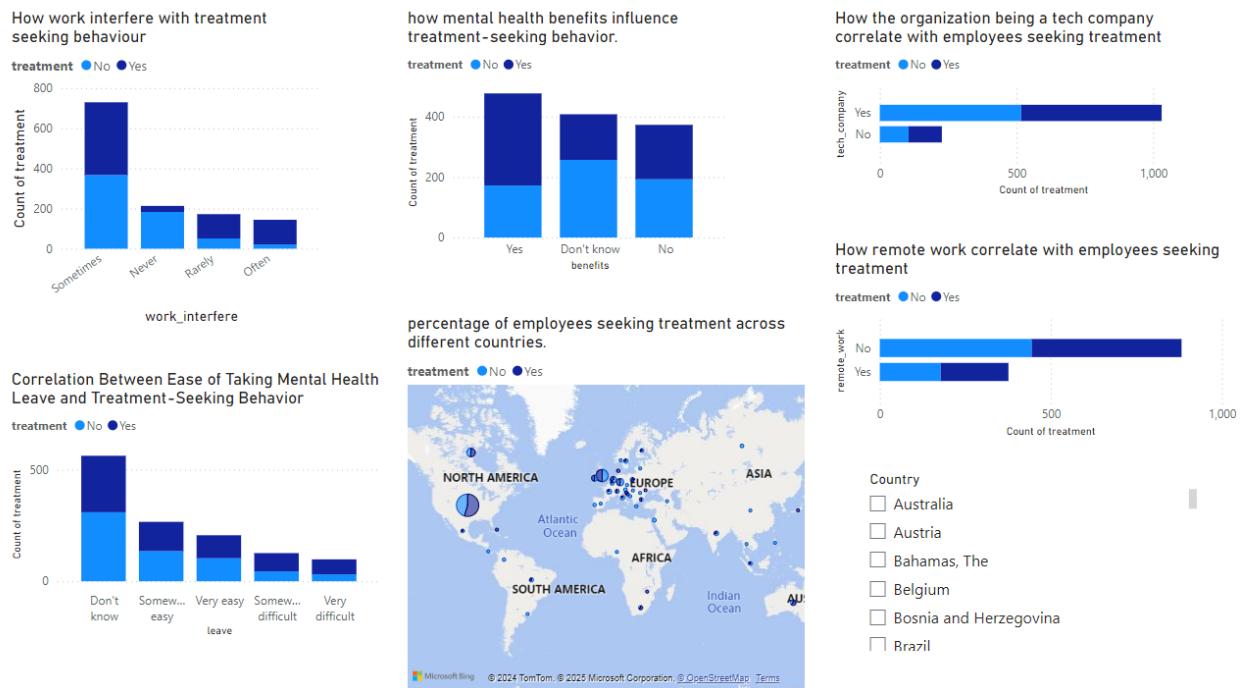
## VIII. Number of employees Distribution ( no\_employees )



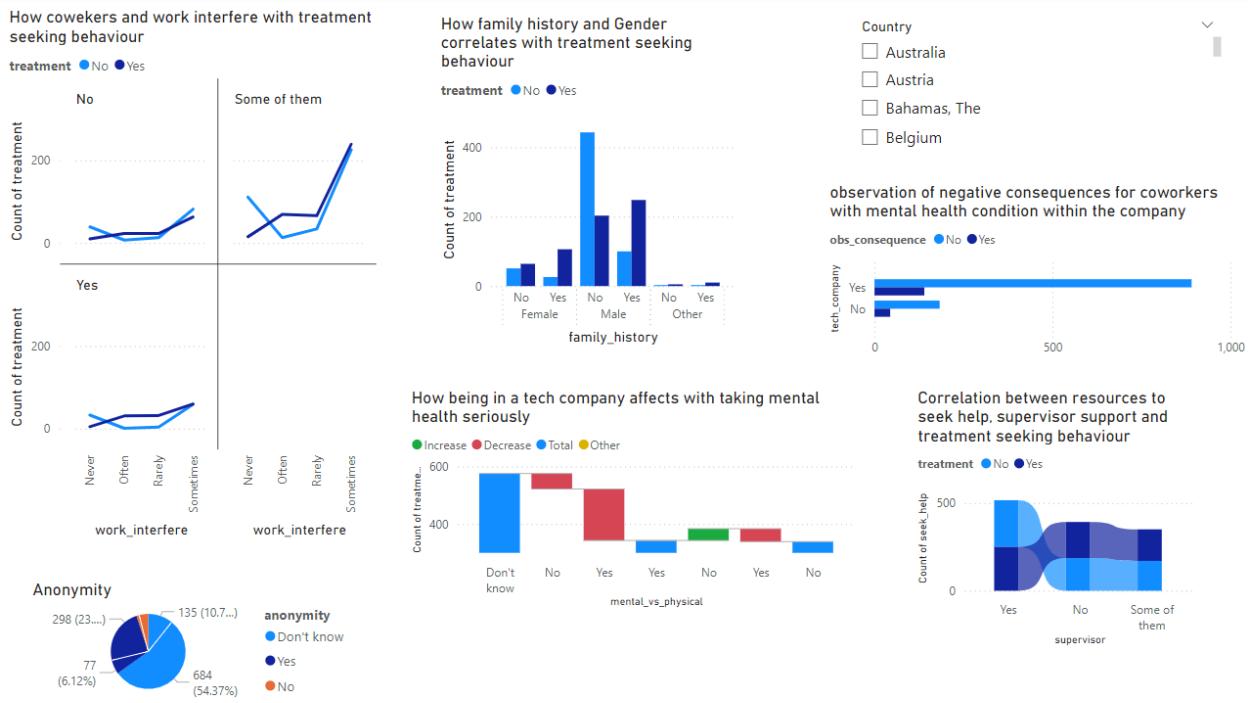
**Insights :** Respondents work across a range of company sizes with a significant proportion in smaller companies with around 120 to 1000 employees.

## Bivariate Analysis

The relationship between variables like work\_interfere, benefits, leave and remote\_work with the dependant variable treatment are analyzed.



## Multivariate Analysis



The screenshot shows the RStudio interface. In the top-left pane, there is an R script with the following code:

```

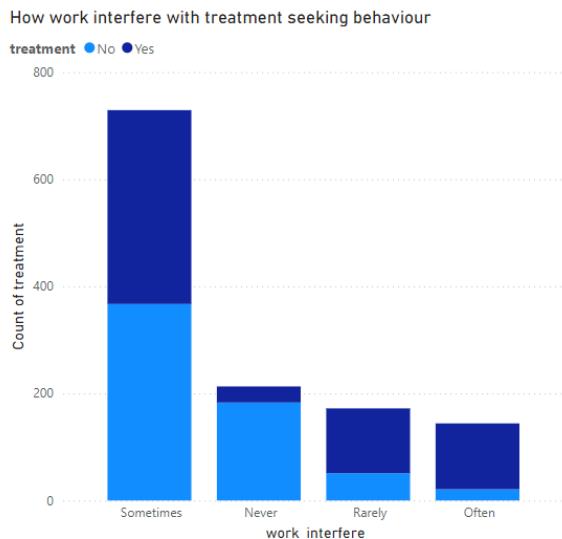
48 install.packages("corrplot")
49
50 # Load the package
51 library(corrplot)
52
53 # Correlation Between Variables
54 # Correlation matrix (numerical variables only)
55 numerical_vars <- data %>% select_if(is.numeric)
56 cor_matrix <- cor(numerical_vars, use = "complete.obs")
57
58 # Heatmap of correlations
59 corrplot(
60   cor_matrix,
61   method = "color",
62   type = "upper",
63   tl.col = "black",
64   tl.srt = 45,
65   title = "Correlation Between Variables"
66 )
67
68
69:1 (Top Level) <-- Background Jobs

```

In the top-right pane, the Global Environment tab is selected, showing objects like CO2, cor\_matrix, data, etc. In the bottom-right pane, a heatmap titled "Correlation Between Variables" is displayed, showing a correlation matrix between Age and no\_employees.

## Insights of Bivariate and Multivariate Analysis

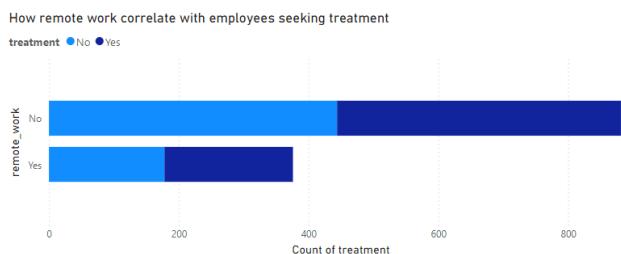
### 1. Work Interference vs Treatment-seeking behaviour



Employees experiencing higher level of mental health interference with work are most likely to seek treatment.

**Insight :** Work interference could be a significant stressor, pushing employees to seek help.

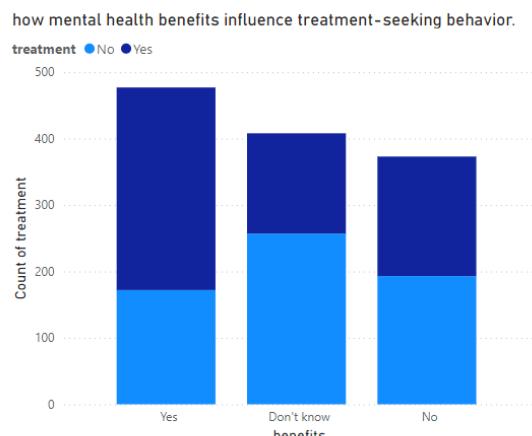
## 2. Remote work vs Treatment-seeking behaviour



Employees who do not work remotely tend to seek treatment more often.

**Insight :** Remote work may reduce workplace stressors potentially lowering treatment needs

## 3. Mental Health benefits vs Treatment-seeking behaviour



Employees who know their company provides mental health benefits are more likely to seek treatment than those who are unsure or not know.

**Insight :** Clear communication and provision of mental health benefits promote a culture of seeking help when needed.

#### **4. Country-wise Distribution**

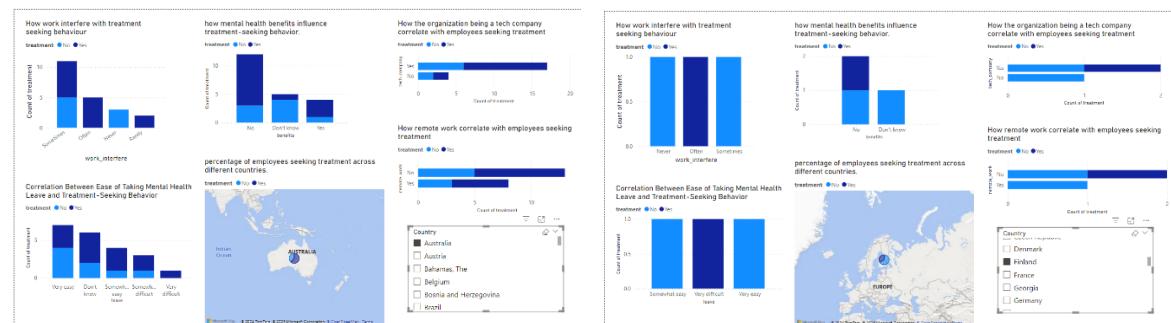
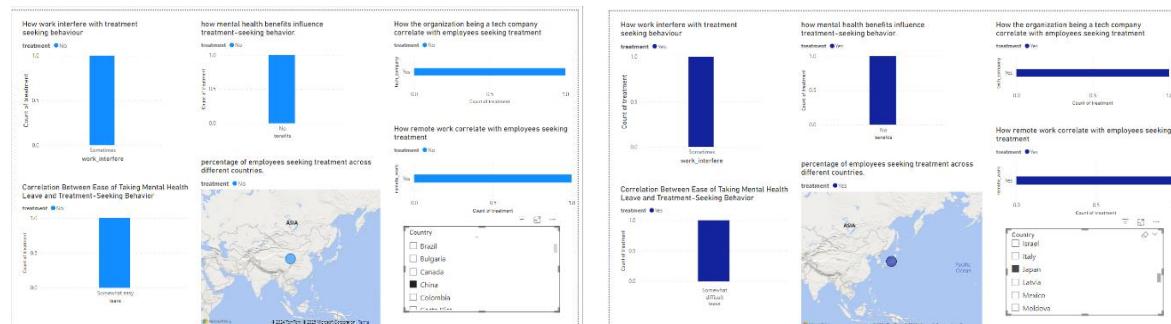
percentage of employees seeking treatment across different countries



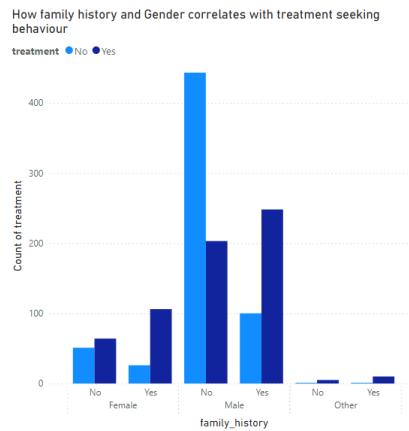
The map visualization highlights geographical variations in treatment-seeking behaviour. North America and Europe have higher concentration of employees seeking help.

**Insight :** Cultural and regional attitudes towards mental health likely influence these trends.

Below are examples of how this differs from Country to Country.



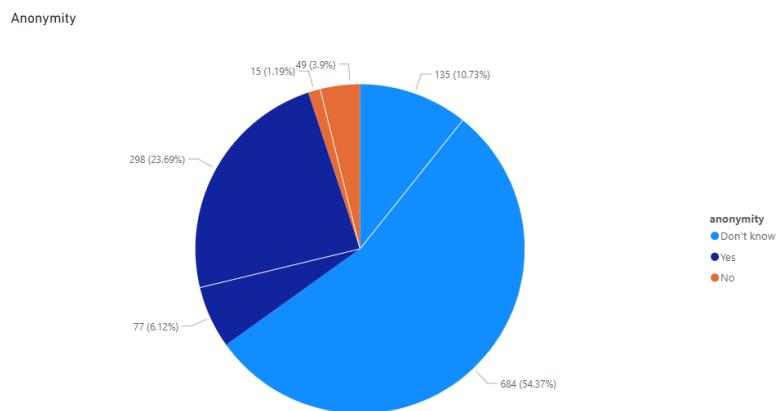
## 5. Family History vs Gender Correlation with treatment



Employees with a family history of mental health issues show a higher tendency to seek treatment

**Insight :** Genetic tendency and gender-specific openness to mental health issues effect treatment-seeking behaviour

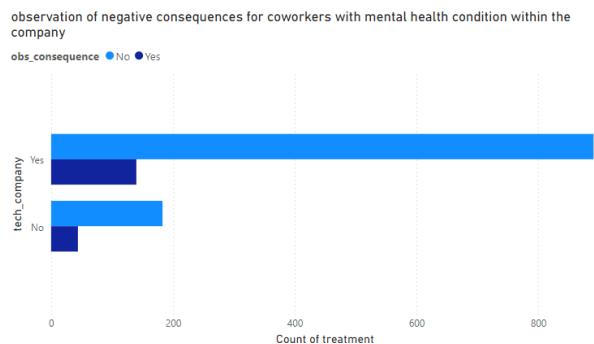
## 6. Anonymity Concerns



A significant proportion of employees prioritize anonymity when seeking treatment

**Insight :** Ensuring confidentiality could encourage more employees to seek mental health support.

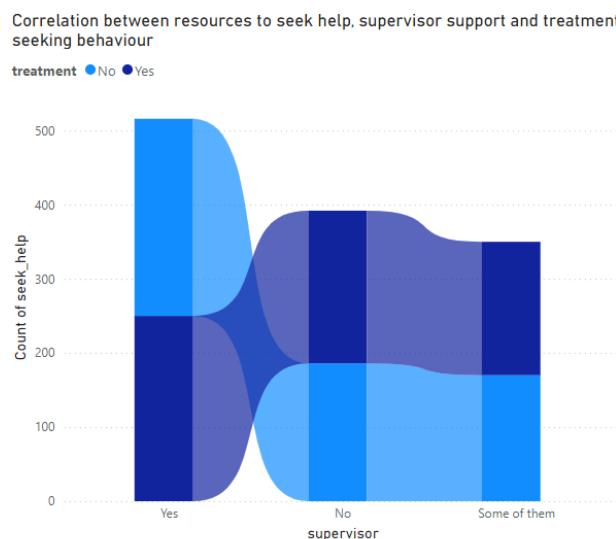
## 7. Organizational Culture and Consequences



Employees in tech companies and those observing negative consequences for co-workers with mental health issues are less likely to seek treatment.

**Insight :** Stigma and workplace culture in tech companies may discourage employees from accessing mental health support. Awareness may help reduce stigma.

## 8. Supervisor support and Resources



Employees who feel supported by supervisors or have access to resources are more likely to seek treatment.

**Insight :** Supervisor training and resource allocation play a vital role in addressing mental health needs.

# Data Storytelling

## The Silent Struggle of Emma : A Story of Workplace Mental Health

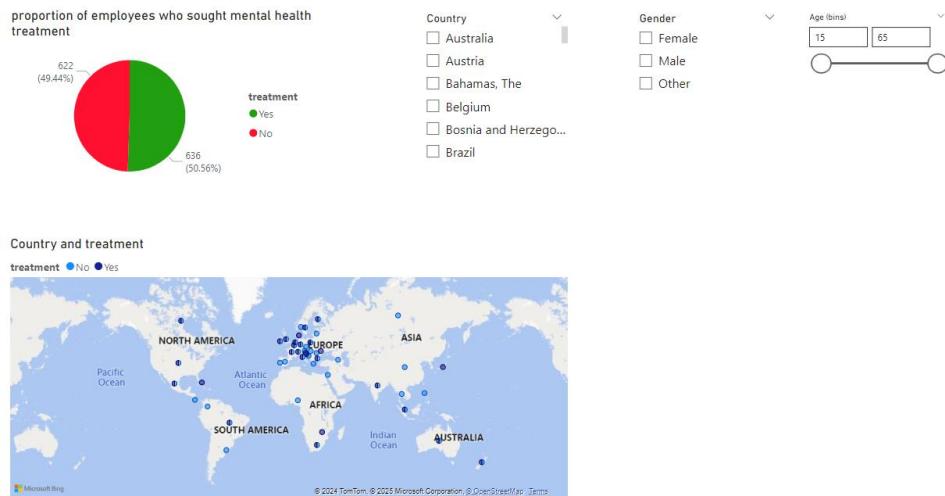
Emma, a 32 year-old marketing manager at a mid-sized tech company, was the kind of employee every organization hopes for. She was hardworking, innovative, and always the first to volunteer for new projects. On the surface, Emma seemed to have it all together – a stable career, a supportive family, and a bright future ahead. But beneath the surface, she was fighting an invisible battle that no one in her workplace knew about.

For months, Emma had been struggling with anxiety and overwhelming stress. Work had started to interfere with her mental health in ways she couldn't have imagined. She used to love the challenge of managing complex projects but now deadlines felt suffocating, meetings drained her energy, and even small tasks felt overwhelming. It didn't help that her company, though progressive in many ways, didn't openly discuss mental health. There was stigma around employees who admitted that they were struggling, and Emma feared being labelled as "weak" or "incapable".

Emma's struggle was a microcosm of a larger issue plaguing in workplaces worldwide. As businesses push for innovation and productivity, the invisible cost, employee mental health, often goes unnoticed. In Emma's case, her tech company's restless pace and lack of supportive policies left her feeling isolated and afraid to seek help. Emma's story is just one among many, but it sheds light on the challenges that employees face and the systemic change needed to create healthier workplaces

## The Landscape of Mental Health in Workplace

To understand Emma's struggle, we turn to data. A global survey revealed that 49.44% of employees sought mental health treatment while 50.56% did not.



This near-even split underscores how widespread mental health challenges are among working professionals.

Emma's tech industry background isn't surprising as 82% of respondents in the survey worked in tech companies.

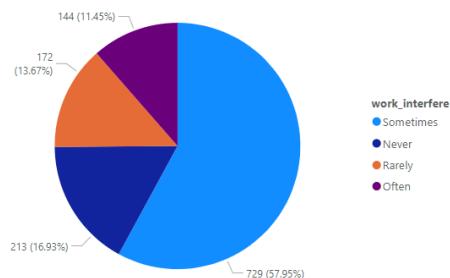


Tech's high-pressure culture often aggravates mental health issues, with employees reporting that work "sometimes" interfere with their well-being.

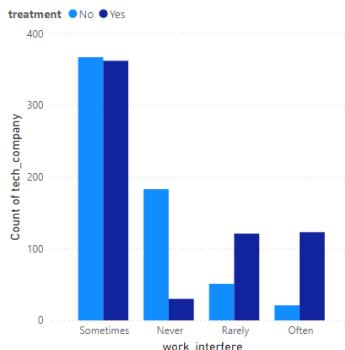
## Work Interference and its Impact

For Emma, work interference was a daily struggle. She wasn't alone, 58% of respondents reported that work sometimes interferes with their mental health.

How often work interfere with mental health



How often work interfere with mental health in tech company

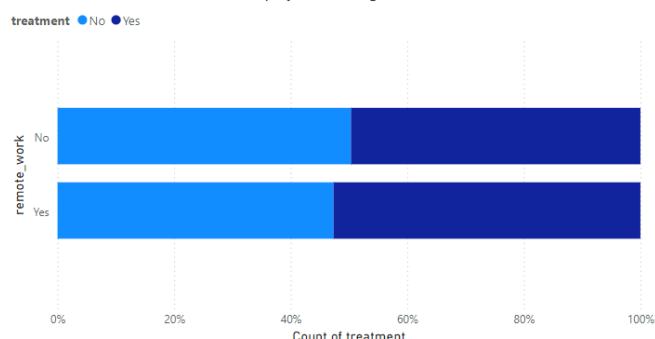


Data also revealed a significant insight : Employees who experience work interfering their mental health are most likely to seek treatment.

Companies must address the root causes of such interference, from unrealistic deadlines to insufficient support systems.

Emma often wondered if working remotely might have helped her manage stress. The data suggests that employees who do not work remotely are more likely to seek treatment.

How remote work correlate with employees seeking treatment



## Cultural and Regional Influence

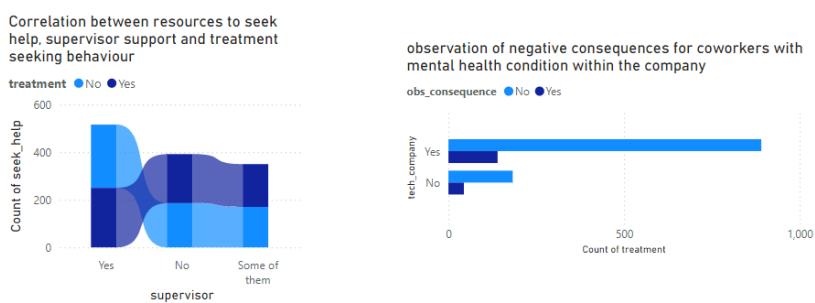
Emma often found solace in reading about global workplace trends. She was fascinated to learn that mental health attitudes vary widely by region. In countries like United States and parts of Europe, employees are more likely to seek help, reflecting cultural openness towards mental health.



## Breaking Stigma

One of Emma's biggest fears was the stigma associated with seeking help. The data revealed a troubling trend : employees in tech companies and those who observed negative consequences for seeking help.

These insights reinforced Emma's beliefs that the change has to start with leadership. Supervisors must create an environment where employees feel safe discussing mental health without fear of judgement.



## The Path Forward

Emma's Journey reached a turning point when she finally reached out to a trusted co-worker. To her surprise, they shared similar struggles and pointed her towards resources outside the company. While it wasn't the ideal solution, it was a step in the right direction.

For companies to truly support employees like Emma, the data offers a clear guidance :

- **Ensure anonymity** : Many employees prioritize confidentiality when seeking help.
- **Provide accessible resources** : Employees with supervisor support and mental health resources are more likely to seek treatment.
- **Train managers** : Supervisors play a significant role in fostering a supportive culture

Emma's story is a call to action. It is a reminder that behind every statistic, is a real person. By listening to the data and implementing changes, workplaces can become spaces where employees thrive – not just professionally, but personally.

## Why this research was important? (Gap analysis )

Despite significant strides in understanding workplace mental health, critical gaps remain. Existing research highlights the importance of leadership, workplace benefits and stigma in shaping employees' mental health outcomes. For instance, (*Smith et al., 2022*) found out that organizations offering mental health support 25% increase in treatment-seeking behaviour. However, factors like work interference, flexibility policies and remote work remain unexplored. Additionally, small-to-medium enterprises and non-Western workplaces are underrepresented, limiting the scope of current findings. Most studies also focus broadly on mental health resources without examining specific triggers for treatment-seeking behaviour.

This research addresses these gaps by leveraging a comprehensive global dataset, particularly from the tech industry, to analyse treatment-seeking behaviour. It examines underexplored factors like work interference, company size, remote work policies while considering diverse cultural contexts. By combining quantitative methods with actionable insights, this study contributes to a deeper understanding of how workplace culture influence mental health treatment, surfacing the way for inclusive, evidence-based interventions.

## References

- Open Sourcing Mental Illness, LTD (2014). *Mental Health in Tech Survey*. [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey> [Accessed 30 Dec. 2024].
- Harvey, P.D., Depp, C.A., Rizzo, A.A., Strauss, G.P., Spelber, D., Carpenter, L.L., Kalin, N.H., Krystal, J.H., McDonald, W.M., Nemerooff, C.B., Rodriguez, C.I., Widge, A.S. and Torous, J. (2022). Technology and Mental Health: State of the Art for Assessment and Treatment. *American Journal of Psychiatry*, [online] 179(12), pp.897–914. doi:<https://doi.org/10.1176/appi.ajp.21121254>.
- Nyblade, L., Stockton, M.A., Giger, K., Bond, V., Ekstrand, M.L., Lean, R.M., Ellen, Nelson, R.E., Sapag, J.C., Taweesap Siraprapasiri, Turan, J. and Wouters, E. (2019). Stigma in health facilities: why it matters and how we can change it. *BMC Medicine*, [online] 17(1). doi:<https://doi.org/10.1186/s12916-019-1256-2>.
- Ackerman, N. (2024). *How to support employee wellbeing at work*. [online] Thetimes.com. Available at: <https://www.thetimes.com/business-money/entrepreneurs/article/how-to-support-employee-wellbeing-at-work-enterprise-network-p2kf29566>? [Accessed 5 Jan. 2025].
- Smith, J.M., Smith, J., McLuckie, A., Szeto, A.C.H., Choate, P., Birks, L.K., Burns, V.F. and Bright, K.S. (2022). Exploring Mental Health and Well-Being Among University Faculty Members: A Qualitative Study. *Journal of Psychosocial Nursing and Mental Health Services*, 60(11), pp.17–25. doi:<https://doi.org/10.3928/02793695-20220523-01>.
- Vonderlin, R., Schmidt, B., Müller, G., Biermann, M., Kleindienst, N., Bohus, M. and Lyssenko, L. (2021). Health-Oriented Leadership and Mental Health From Supervisor and Employee Perspectives: A Multilevel and Multisource Approach. *Frontiers in Psychology*, [online] 11. doi:<https://doi.org/10.3389/fpsyg.2020.614803>.
- Elraz, H. (2018). Identity, mental health and work: How employees with mental health conditions recount stigma and the pejorative discourse of mental illness. *Human Relations*, [online] 71(5), pp.722–741. doi:<https://doi.org/10.1177/0018726717716752>.

Wu, A., Roemer, E.C., Kent, K.B., Ballard, D.W. and Goetzel, R.Z. (2021). Organizational Best Practices Supporting Mental Health in the Workplace. *Journal of Occupational and Environmental Medicine*, [online] 63(12), pp.e925–e931. doi:<https://doi.org/10.1097/jom.0000000000002407>.

Song, Z. and Baicker, K. (2019). Effect of a Workplace Wellness Program on Employee Health and Economic Outcomes. *JAMA*, [online] 321(15), pp.1491–1491. doi:<https://doi.org/10.1001/jama.2019.3307>.