Mental health analysis

Prepared by:

yassin tamer	**23-101161**
youssef khaled	**23-101058**
aleyeldeen manzour	**23-101129**

About the data set.

The purpose of this data set is to capture and analyze survey responses related to mental health in the tech workplace.It includes demographic information, employment details, and a range of questions about mental health experiences, support systems, and workplace attitudes. The survey aims to explore how mental health is perceived, addressed, and accommodated in the technology industry, particularly in relation to employer benefits, openness in discussing mental health, and the impact of workplace environments on individuals seeking treatment. This data can be used to identify trends, challenges, and opportunities for improving mental health awareness and support within tech companies.

Research Question

"How do workplace factors influence the likelihood of employees in the tech industry seeking mental health treatment?"

Hypothesis Testing

1)Family History vs. Treatment
2)Gender vs. Treatment
3)Remote Work vs. Treatment

Population of interest

The population of interest in this study consists of people living in the USA who work in the technology industry, particularly those with experiences or perspectives related to mental health at work. This group includes employees in various tech roles, such as developers and engineers, as well as self-employed individuals and those working for companies of different sizes.

Sampling Method

This study employed a convenience sampling approach, likely recruiting participants through accessible online platforms such as social media, forums, and technology-focused communities. Participation was voluntary, with individuals opting in to complete the survey without being selected through a random or stratified process. As such, this method aligns with characteritics of convenience sampling.

Bias Identification

The dataset on mental health in the tech industry provides valuable insights but is subject to biases that may affect its validity and generalizability:

1. Self-Selection Bias: Voluntary survey participation likely overrepresents individuals with a strong interest in or experience with mental health, potentially misrepresenting the broader tech workforce.
2. Sampling Bias: Convenience-based sampling via tech forums, social media, or professional networks may overrepresent certain demographics (e.g., North American or younger individuals) while underrepresenting others, limiting generalizability.
3. Demographic Misreporting and Data Quality Issues: Inconsistent demographic data, such as implausible ages or varied gender responses, can introduce errors, necessitating thorough data cleaning for accurate analysis.
4. Social Desirability Bias: Respondents may provide socially acceptable answers due to the sensitive nature of mental health, potentially underreporting stigma and overreporting supportive environments.
5. Non-Response Bias: Missing responses in key fields may correlate with mental health status or workplace experiences, skewing results toward those with more significant issues.

Data collection

The dataset was collected through an anonymous online survey focused on mental health in the workplace. It includes 1,259 self-reported responses covering demographics, employment details, and mental health support. The presence of timestamps suggests the survey was conducted over time, with voluntary participation providing insights into workplace mental health experiences.

Data cleaning

The dataset underwent several data cleaning steps to ensure accuracy and consistency. First, the comments column was removed due to over 90% missing values. Next, the Age column was cleaned by filtering out unrealistic values, keeping only ages between 16 and 100. The Gender column was standardized by mapping various textual entries to three categories: "Male", "Female", and "Other", using pattern matching to handle inconsistencies in how respondents described their gender.

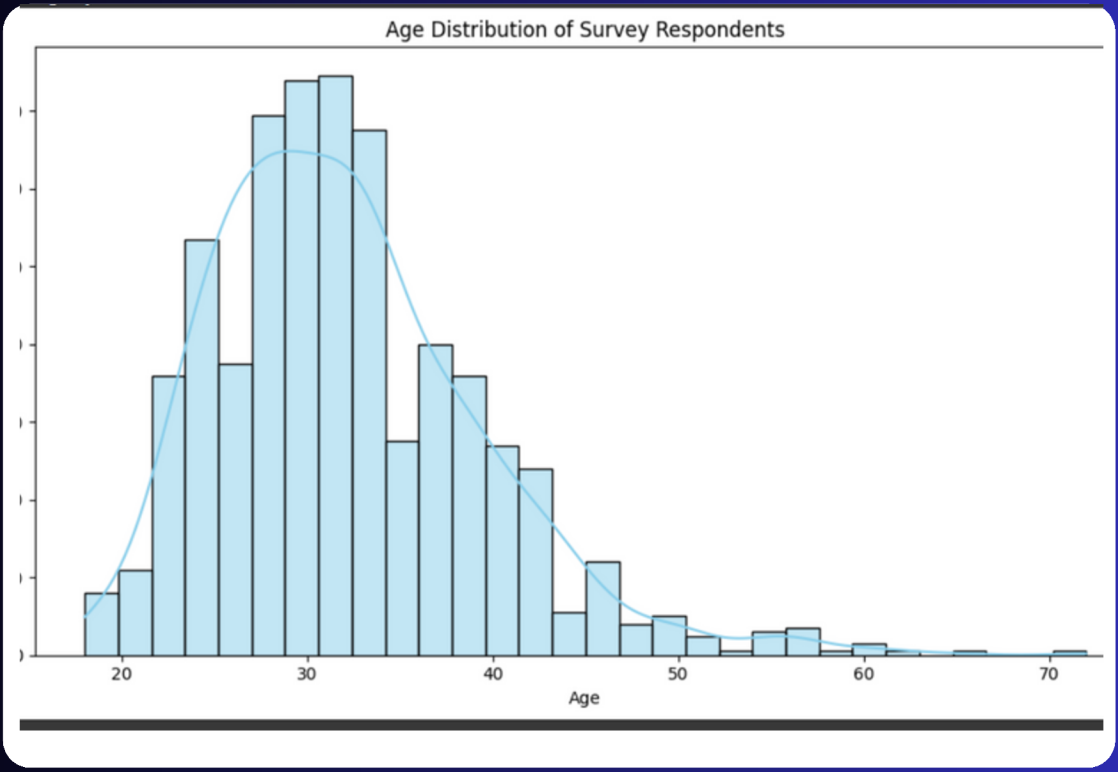
Corelation Analysis

A correlation analysis was conducted to explore the relationships between various features in the mental health survey dataset. Since the majority of the data consists of categorical variables, one-hot encoding was applied to convert these into a numerical format suitable for correlation analysis. Non-informative columns such as timestamps, free-text comments, and U.S. state identifiers were excluded to maintain analytical relevance. The Pearson correlation coefficient was calculated to quantify the linear relationship between all pairs of encoded features. To visualize the results, a heatmap was generated using the Seaborn library, employing a “coolwarm” color palette to distinguish positive and negative correlations. This visual representation provides a clear and concise overview of how different factors may be associated, supporting the identification of patterns that could inform further statistical or predictive modeling.

Data set content

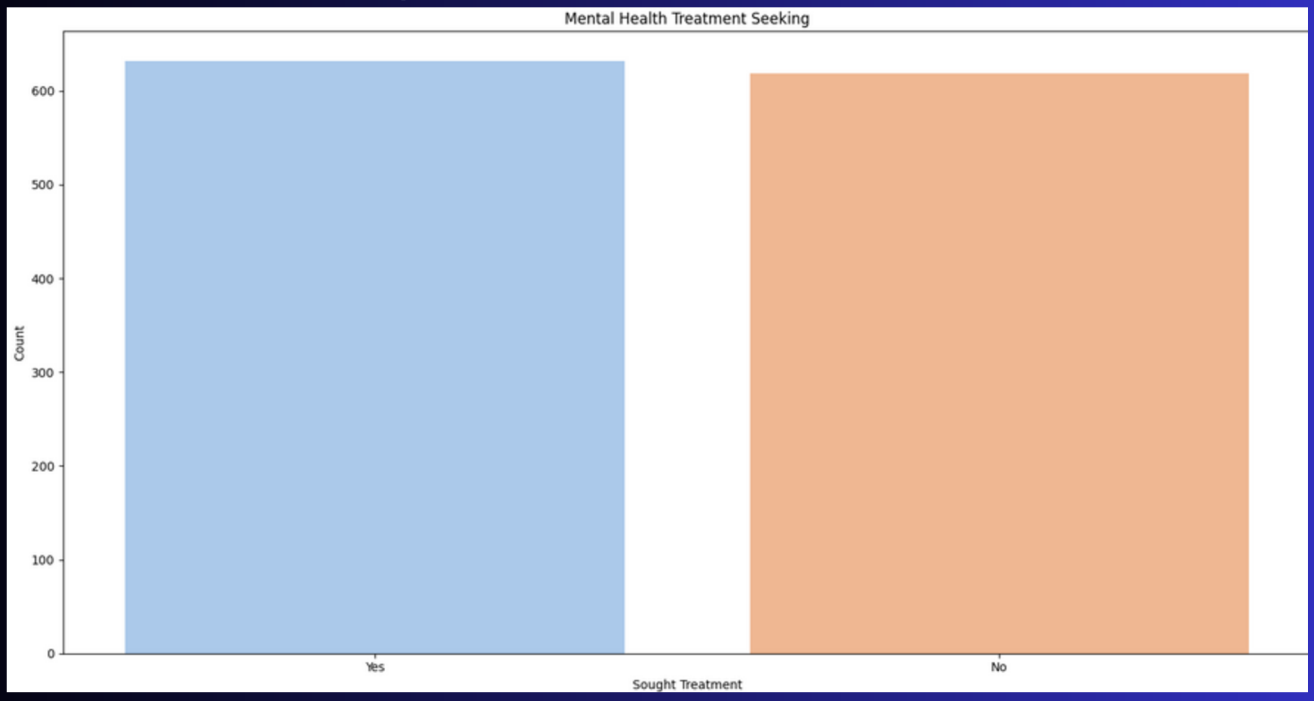
- Title: Mental Health in Tech Survey Dataset
- source:https://www.kaggle.com/datasets/osmi/mental health-in-tech-survey/data
- Rows: 1,259
- Columns: 27
- Key Columns:
  - Demographics: Age, Gender, Country, state
  - Employment: self_employed, no_employees, remote_work, tech_company
  - Mental Health: family_history, treatment, work_interfere, 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
  - Others: Timestamp, comments

Data Analysis



The graph presents a histogram illustrating the age distribution of survey respondents, with a smooth line overlay that highlights the overall trend. The distribution is right-skewed, indicating that the majority of participants are concentrated in the 25 to 35 age range, with the frequency peaking around age 30. The number of respondents gradually decreases with age, highlighting a lower representation of older individuals.

Data Analysis

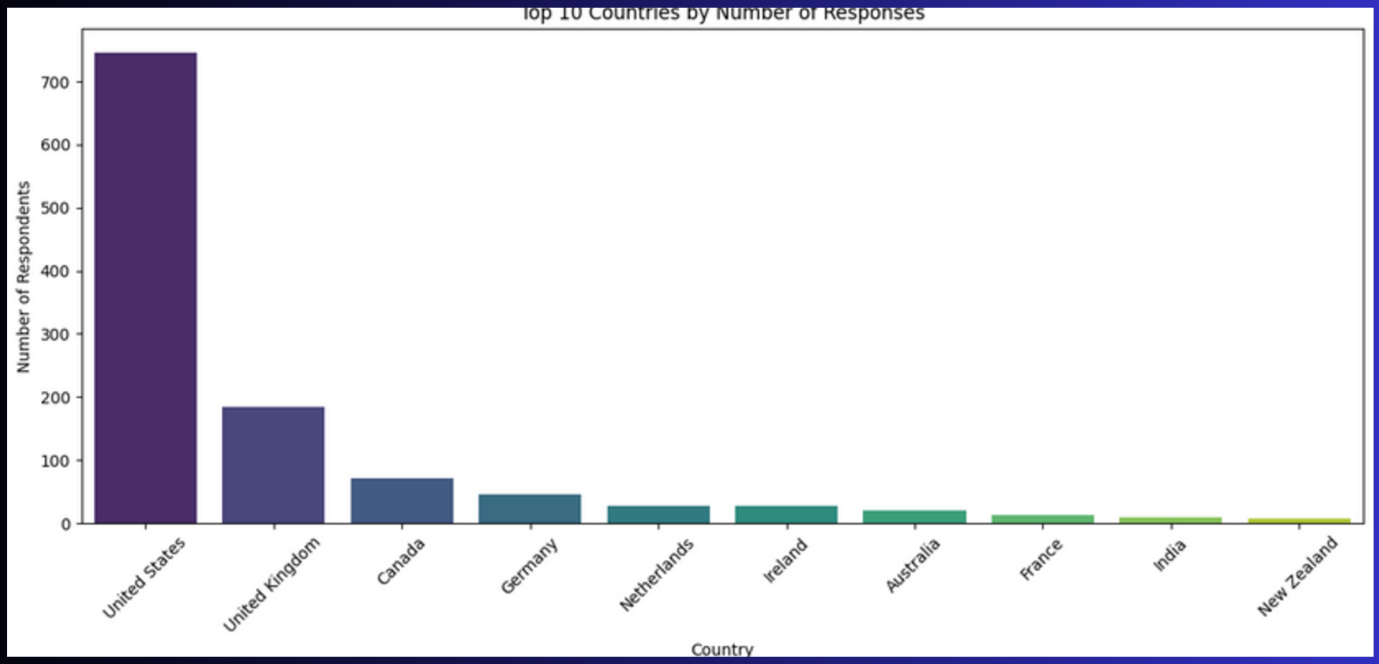


The chart titled "Mental Health Treatment Seeking" displays a histogram illustrating the count of individuals who have sought mental health treatment, categorized into "Yes" and "No." The x-axis lists the treatment-seeking categories, while the y-axis represents the count, ranging from 0 to 700.

- The "Yes" category, depicted in light blue, shows a count of approximately 600 individuals, indicating a substantial portion of the population has sought treatment.
- The "No" category, shown in a peach shade, has a count of around 550 individuals, slightly lower than the "Yes" group but still significant.

The chart highlights a nearly balanced distribution, with a slight edge toward those who have sought mental health treatment, suggesting a relatively high engagement with mental health services within the sampled population.

Data Analysis

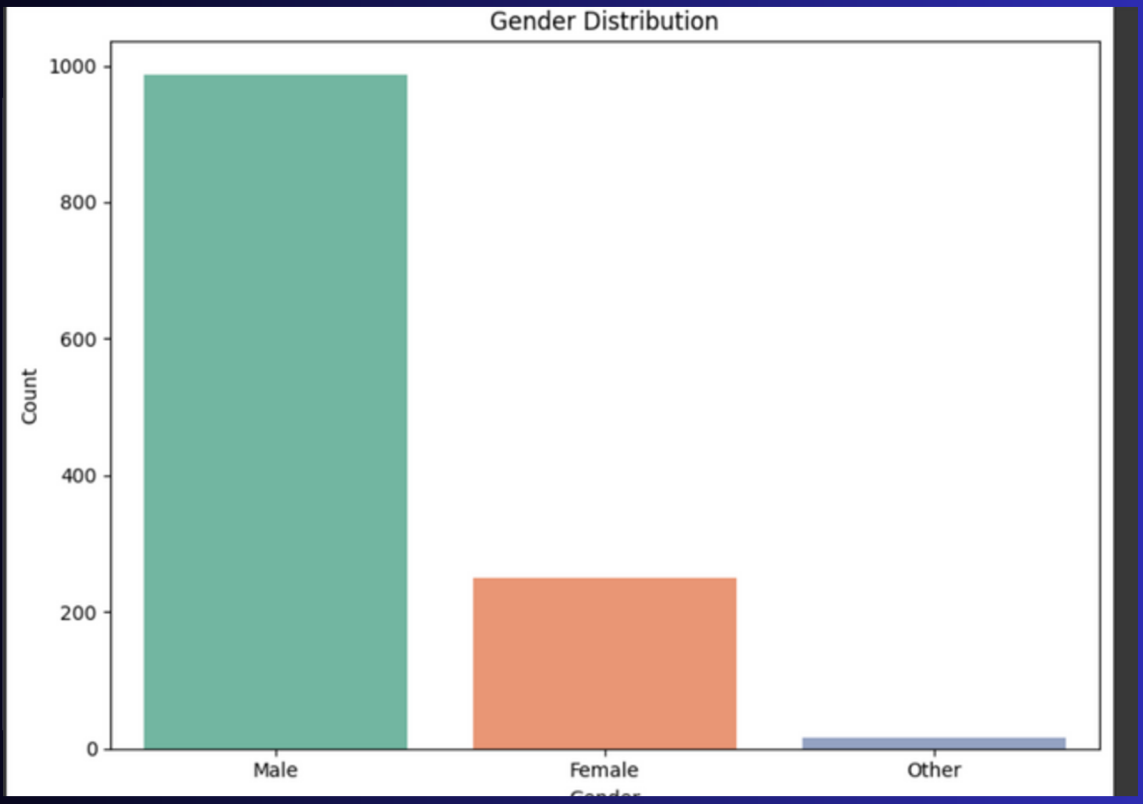


The chart titled "Top 10 Countries by Number of Responses" displays a histogram illustrating the number of respondents across ten countries. The x-axis lists the countries, while the y-axis represents the number of respondents, ranging from 0 to 700.

- The "United States," in purple, has the highest number of respondents, approximately 700, indicating it leads significantly.
- The "United Kingdom," in dark blue, follows with around 200 respondents, showing a notable but much smaller contribution.
- "Canada," in medium blue, has about 100 respondents, while "Germany," in teal, is close behind with a similar count.
- "Netherlands," in green, "Ireland," in light green, "Australia," in olive, "France," in lime, "India," in yellow-green, and "New Zealand," in yellow, each have progressively fewer respondents, with counts ranging from approximately 50 down to near 10.

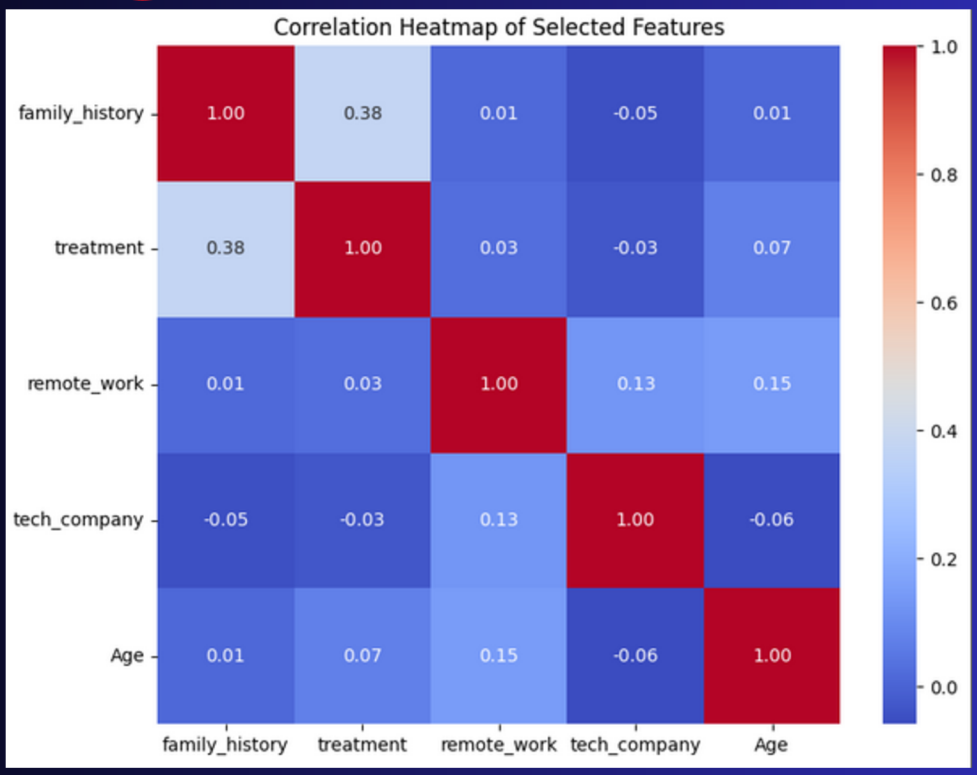
The chart highlights a steep decline in response numbers, with the United States dominating and the remaining countries showing a gradual decrease in participation.

Data Analysis



The chart titled "Gender Distribution" shows the count of individuals across three gender categories: Male, Female, and Other. Males, in teal, have the highest count at approximately 1000, followed by Females, in coral, at around 250, with Other, in light blue, at about 50. The data indicates a skewed distribution with Males being the dominant group.

Corelation Analysis



The heatmap titled "Correlation Heatmap of Selected Features" shows the correlation between family_history, treatment, remote_work, tech_company, and Age, with a color scale from -1.0 (blue, negative) to 1.0 (red, positive). Key correlations include:

- A moderate positive correlation (0.38) between family_history and treatment, indicating individuals with a family history are more likely to seek treatment.
- Weak positive correlations between remote_work and tech_company (0.13) and Age (0.15), suggesting a slight link to tech roles or older age.
- Weak negative correlations between tech_company and Age (-0.06) and family_history (-0.05), hinting at younger employees or less family history.
- Most other pairs, like remote_work with family_history (0.01) and treatment with tech_company (-0.03), show negligible correlation.

The strongest relationship is between family_history and treatment, while other variables exhibit weak or no significant correlations.

Hypothesis testing

Hypothesis Testing: Family History vs. Treatment

1. Our Hypotheses:
   - Null Hypothesis (H_0): Family history does NOT affect treatment seeking.
   - Alternative Hypothesis (H_1): Family history DOES affect treatment seeking.
2. Then created a Contingency Table:
   - Counted how many people fell into each combination of family history (Yes/No) and treatment seeking (Yes/No).
   - Then used the Chi-Squared Test for Independence, since we are testing the relationship between two categorical variables.
3. Ran the Test:
   - Calculated the Chi-squared statistic, degrees of freedom, and p-value using the contingency table.

   - Compared the p-value to 0.05 :
     - If $p < 0.05$, reject H_0 → There is a significant relationship.
     - If $p \geq 0.05$, fail to reject H_0 → No significant relationship.
   - result was $p < 0.05$ so indeed there is a significant relationship

4. Conclusion:
   - Based on results: Reject H_0 — Family history is significantly associated with treatment seeking.

Hypothesis testing

Hypothesis Testing: Gender vs. Treatment Seeking

1. Our Hypotheses:
   - H_0 (Null): Gender and treatment seeking are independent.
   - H_1 (Alternative): Gender and treatment seeking are associated.

2. Created a Contingency Table:
   - Counted how many people in each gender group (Female, Male, Other) did or did not seek treatment.
   - then used the Chi-Square Test of Independence (because both variables are categorical).

3. Ran the tests:
   - Using scipy.stats.chi2_contingency() to get the Chi-square statistic, degrees of freedom, and p-value.
   - Common $\alpha = 0.05$
   - If p-value < 0.05, reject H_0 → There is a significant association.
   - If p-value ≥ 0.05, fail to reject H_0 → There is no significant association.
   - result was $p < 0.05$ so there is a signifiicant relation.

4. conclusion:

based on results: Reject H_0 — Gender is significantly associated with treatment seeking.

Hypothesis testing

Hypothesis Testing: Remote Work vs. Treatment Seeking

1. Our Hypotheses:
   - Null Hypothesis (H_0):Remote work does NOT affect treatment seeking.
   - Alternative Hypothesis (H_1): Remote work DOES affect treatment seeking.

2. Created a Contingency Table:
   - Counted how many people fall into each combination of remote work (Yes/No) and treatment seeking (Yes/No).
   - Used the Chi-Squared Test for Independence, since we're testing the relationship between two categorical variables.

3. Ran the Test:
   - Calculate the Chi-squared statistic, degrees of freedom, and p-value using the contingency table.
   - Compare the p-value to 0.05 (significance level):
   - If $p < 0.05$, reject H_0 → There is a significant relationship.
   - If $p \geq 0.05$, fail to reject H_0 → No significant relationship.
   - result was p>=0.05 so there's no significant relationship.

4. Conclusion:
   - Fail to reject H0. No significant association found between remote work and treatment seeking

Possible improvemenets

1. Targeted Awareness Campaigns
   - Focus on individuals with a family history of mental illness, as they are significantly more likely to seek treatment. Reinforce support systems and early intervention strategies for this group.

2. Inclusive and Universal Mental Health Initiatives
   - Since gender and remote work status are not significantly associated with treatment-seeking, avoid narrowly tailored programs. Instead, promote broadly accessible mental health resources that reach all employees or populations equally.

3. Normalize Mental Health Conversations
   - Although family history plays a role, stigma or lack of awareness may still deter others. Create environments (especially workplaces) that encourage open dialogue and reduce stigma around seeking help.

4. Strengthen Preventive Measures
   - Incorporate mental health screenings and educational tools into general health services, especially for those with known risk factors like family history.

5. Improve Data Collection
   - Encourage more complete and standardized responses in future surveys, particularly in areas like gender identity and workplace support systems, to enable deeper and more nuanced analysis.

Potential Issues

1.Mental Health Stigma

- Many people may still feel uncomfortable talking about or seeking help for mental health, especially those without a family history of mental illness.

2.Uneven Awareness and Support

- People without a family history may be less likely to seek help, possibly due to lower awareness or understanding of mental health problems.

3.Lack of Detailed Data

- Some survey answers, like gender identity and workplace policies, are not specific or complete enough, which makes it harder to get deep insights.

4.Limited Impact of Workplace Setup

- Working remotely doesn't seem to affect whether someone seeks mental health help, suggesting that workplace mental health support might not be reaching everyone effectively.

5.One-Size-Fits-All Programs

- Since gender and work type don't show strong effects, mental health programs may be too general and not tailored enough to individuals' needs.

Conclusion

This analysis investigated whether family history of mental illness, gender, and remote work status are associated with individuals seeking mental health treatment. Using chi-squared tests of independence on a survey dataset, the following conclusions were drawn:

- Family History: There is a statistically significant association between having a family history of mental illness and the likelihood of seeking mental health treatment ($\chi^2 = 175.96$, $p < 0.001$).

- Gender: Significant relationship was found between gender and treatment-seeking behavior since that we rejected the hypotheses ($\chi^2 = 50.9496$, $p < 0.001$).

- Remote Work: Working remotely did not show a statistically significant association with treatment-seeking behavior ($\chi^2 = 0.7667$, $p < 0.3812$).

In summary, the only factor among those analyzed that significantly influences mental health treatment seeking is a family history of mental illness. These findings support targeted mental health awareness and intervention efforts for individuals with such a background, while suggesting that gender and remote work status do not require differentiated approaches based on treatment trends.