Mental Health in Tech Industry

Tomeka N. Pena

Western Governors University

Table of Contents

[A. Project Highlights 3](#_Toc177923213)

[B. Project Execution 4](#_Toc177923214)

[C. Data Collection Process 6](#_Toc177923215)

[C.1 Advantages and Limitations of Data Set 6](#_Toc177923216)

[D. Data Extraction and Preparation 7](#_Toc177923217)

[E. Data Analysis Process 9](#_Toc177923218)

[E.1 Data Analysis Methods 9](#_Toc177923219)

[E.2 Advantages and Limitations of Tools and Techniques 10](#_Toc177923220)

[E.3 Application of Analytical Methods 11](#_Toc177923221)

[F. Data Analysis Results 14](#_Toc177923222)

[F.1 Statistical Significance 14](#_Toc177923223)

[F.2 Practical Significance 16](#_Toc177923224)

[F.3 Overall Success 17](#_Toc177923225)

[G. Conclusion 18](#_Toc177923226)

[G.1 Summary of Conclusions 18](#_Toc177923227)

[G.2 Effective Storytelling 20](#_Toc177923228)

[G.3 Recommended Courses of Action 21](#_Toc177923229)

[H. Panopto Presentation 22](#_Toc177923230)

[Appendix A 24](#_Toc177923231)

[Supporting Files 24](#_Toc177923232)

# A. Project Highlights

The focus of this project seeks to gain understanding of the relationship between access to mental health resources and employees’ comfort in discussing mental health issues at work by answering the following research question: Is there a correlation between the availability of mental health resources and employees’ willingness to discuss mental health issues at work, and is there a difference in employee comfort levels in discussing mental health issues between those who know their mental health care options and those who do not? Specifically, the research aims to determine if knowing mental health care options influences employees' comfort in discussing these issues with coworkers and supervisors. Addressing this question is vital for creating a supportive workplace, as open communication about mental health can lead to better mental health outcomes and overall productivity

The scope of this project will encompass the collection and analysis of employee data related to mental health resource awareness and communication comfort in the workplace. It will involve cleaning and preparing the dataset for analysis, followed by exploratory data analysis to identify trends and patterns. Statistical modeling will be employed to assess the correlation between the awareness of mental health resources and comfort in discussing mental health with supervisors and coworkers. Additionally, the project will include visualizations of key findings and actionable recommendations for improving organizational mental health policies based on the results of the analysis. These elements are essential to addressing the research question effectively

The methodology used when creating this project was the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as it is a well-established, iterative methodology that provides a structured approach for tackling data-driven projects. It is widely used because of its flexibility and comprehensiveness, making it applicable across various industries. CRISP-DM comprises of six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase focused on different aspects of the data analytics process, ensuring that the project remains goal-oriented, the data is properly handled, and the final analysis produces actionable insights. By following CRISP-DM, we achieved clarity in the objectives of the data analysis and ensured that each step was aligned with solving the research question.

The tools used to complete this project are Jupyter Notebook and Python. Jupyter Notebook was used to perform exploratory data analysis (EDA), data cleaning and transformation, data analysis, and data visualization. Python is the programming language that was used to clean, structure, and preprocess the dataset. It was well-suited for this analysis due to its wide range of statistical and visualization libraries that are efficient for processing large datasets. The libraries Pandas and NumPy in Python helped in handling the dataset by filtering relevant columns, dealing with missing values, and performing group-by operations. Visualization libraries such as Matplotlib or Seaborn was used to graphically represent the data and identify trends.

# B. Project Execution

The primary goal of this project is to assess the impact of mental health resource awareness on employees' comfort in discussing mental health issues in the workplace. The key objective is to determine if employees who are more informed about mental health resources feel more comfortable engaging in mental health-related conversations with their peers and supervisors. Another objective is to identify specific factors, such as company size or industry, that may influence this relationship. The major deliverables include a cleaned dataset, comprehensive statistical analysis, detailed visualizations, and a final report summarizing findings with actionable recommendations for organizations. These deliverables will support the primary goal by providing data-driven insights that organizations can use to enhance communication about mental health resources and foster a supportive work environment. Each deliverable will be aligned with the overall objective of helping organizations make evidence-based decisions to improve workplace mental health awareness.

The methodology used when creating this project was the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as it is a well-established, iterative methodology that provides a structured approach for tackling data-driven projects. It is widely used because of its flexibility and comprehensiveness, making it applicable across various industries. CRISP-DM comprises of six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase focused on different aspects of the data analytics process, ensuring that the project remains goal-oriented, the data is properly handled, and the final analysis produces actionable insights. By following CRISP-DM, we achieved clarity in the objectives of the data analysis and ensured that each step was aligned with solving the research question.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Projected Time** | **Start Date** | **End Date** |
| Retrieve and Clean Dataset | 8 Hrs. | 09/16/2024 | 09/16/2024 |
| Evaluate and Analyze Dataset | 8 Hrs. | 09/17/2024 | 09/17/2024 |
| Create Visualizations | 4 Hrs. | 09/19/2024 | 09/19/2024 |
| Create Slideshow of Analysis Results | 4 Hrs. | 09/19/2024 | 09/19/2024 |
| Present Slideshow of Analysis Results | 2 Hrs. | 09/20/2024 | 09/20/2024 |

The plan and methodology did not vary from Task 2. However, there were some minute changes in the timeline than what was projected in Task 2. Some of the work took less time than projected to complete, but the work was performed on the projected start date.

# C. Data Collection Process

The data collection process did not vary from Task 2 nor did I encounter any unplanned issues. I was able to collect the data by downloading the survey.csv file from Kaggle.com. The dataset was downloaded as a zip file and was uploaded into Jupyter Notebook using Panda read\_csv function as outline and planned in Task 2.

Fortunately, I did not run into any obstacles during the data collection process. The dataset was made publicly available to me as a Kaggle member. I was able to download the dataset for free from Kaggle.com and was able to save the it as a CSV file to prepare it for my data analysis.

In addition, I did not run into any unplanned data governance issues during the data collection process. The dataset is unrestricted and free to both use and share without licensing. Furthermore, all personal information that could reveal participants identities were removed before downloading the dataset, thus keeping with governance regulations such as GDPR and HIPPA.

## C.1 Advantages and Limitations of Data Set

**Advantages of the Dataset**

One key advantage of the dataset is the breadth and relevance of variables related to mental health in the workplace. It contains a variety of columns capturing different aspects of employees' attitudes and experiences with mental health, such as access to resources, willingness to discuss mental health, and specific demographic details.

For instance, the inclusion of the care\_options column provides critical information on whether employees know about available mental health resources. This allows us to explore relationships between the availability of resources and the willingness to discuss mental health, answering research questions about the role of workplace support systems in mental health conversations.

**Disadvantage of the Dataset**

A disadvantage of the dataset is the presence of subjective and categorical data, which can introduce bias or ambiguity in interpretation. Many responses use non-standard categories such as "maybe" or "don't know", which may complicate quantitative analysis.

For example, the seek\_help column, with values such as "yes," "no," and "maybe", presents a challenge for analysis because the interpretation of "maybe" is inherently subjective and can differ between respondents. This can introduce noise in the data when performing correlation analysis, as it's unclear how to categorize or treat such intermediate responses quantitatively.

# D. Data Extraction and Preparation

The dataset was made publicly available at Kaggle.com to it members. First, I created an account with Kaggle.com to be able to download the dataset onto my computer. The data was then extracted as a CSV file within a zip file and uploaded into Jupyter Notebook as a dataframe using Panda’s read\_csv function. I then performed exploratory data analysis (EDA) to investigate what necessities were needed to be executed in the data preparation step. I began by using the head() function to view the first 5 rows of the dataset as well as the columns and format of the values in each column. Next, info() function was used to obtain more information such as total number of rows and the data type of the columns, followed by the isnull().sum() functions to gather the sum of null values in each column.

After EDA was perform, I determined that I needed to drop unrelated columns, handle null values, and convert categorical data into numerical values in order to perform correlation analysis. Once unrelated columns were drop, the remaining columns left for data analysis were the following: Age, Gender, Country, state, care\_options, seek\_help, coworkers, supervisor, and mental\_health\_interview. Null values were found in the state column, which was expected because some participants did not reside in the United States and therefore would not be living in a U.S. state. For rows with country listed as United States and had a null value in the state column received a default value of ‘DC’ for the District of Columbia. I assumed that these participants resided in Washington, D.C. and did not list a state because D.C. is not considered a U.S. state. Next, I deleted rows that should have contained null values in the state column, but did not (non-U.S. countries with a U.S. state). I then used the unique() function to list all of the countries the participants were from. Since majority of the participants resided in the U.S. or U.K. and other countries had one participant, I changed the values of the country column to following three options: United States, United Kingdom, Other. The gender column allowed participants to enter their gender identity which totaled in numerous of entries, including misspelled entries and therefore required data normalization to be performed to obtain a consistent format of values. I used the unique() function to list all unique values then I fixed the column to have the following three values: Male, Female, Other. Lastly, I converted the categorical data need for the analysis into numerical values for correlation analysis. An example is the seek\_help, which had the values “Yes”, “No”, “Don’t Know”, that was changed to the following values respectively: 1, 2, 3. The categorical data had to be converted into numerical data to enable the calculation of correlation coefficients, as correlation analysis requires numerical inputs to quantify the strength and direction of relationships between variables.

# E. Data Analysis Process

## E.1 Data Analysis Methods

The correlation analysis was an appropriate method for this project because it allowed for the exploration of the linear relationship between two continuous variables: resource awareness and communication comfort. Since both of these are measurable on a continuous scale, correlation analysis provided a robust framework for understanding the relationship between them. This method was particularly effective for providing an overview of the data and identifying trends that would be further investigated. Additionally, the Pearson correlation coefficient offered an intuitive metric that stakeholders could easily interpret, making it a valuable tool for decision-making regarding mental health initiatives.

The t-test was justified as it directly compared the means of two independent groups, which made it well-suited for this project's goal of assessing differences in comfort levels based on awareness of mental health resources. By using a two-sample t-test, we determined if the mean comfort levels of employees who were aware of mental health resources differ significantly from those who were not. This method was particularly relevant because it not only tested for statistical significance but also provided practical insights into how interventions (such as increasing resource awareness) could improve employee well-being. The results from the t-test offered concrete evidence for whether expanding mental health resources in the workplace can effectively enhance communication and openness around mental health issues.

## E.2 Advantages and Limitations of Tools and Techniques

The tools that I used to complete this project were Jupyter Notebook and Python. Jupyter Notebook was used to perform exploratory data analysis (EDA), data cleaning and transformation, data analysis, and data visualization. Python is the programming language that was used to clean, structure, and preprocess the dataset. It was well-suited for this analysis due to its wide range of statistical and visualization libraries that are efficient for processing large datasets. The libraries Pandas and NumPy in Python helped in handling the dataset by filtering relevant columns, dealing with missing values, and performing group-by operations. Visualization libraries such as Matplotlib and Seaborn were used to graphically represent the data and identify trends.

**Advantages:**

* Jupyter Notebook provides an interactive environment that allows for step-by-step execution of code, making it highly suitable for exploratory data analysis (EDA) and iterative experimentation. This flexibility enhances the ability to visualize intermediate results and make adjustments in real time.
* Python is a versatile and powerful programming language with extensive libraries such as Pandas and NumPy for data manipulation, and Matplotlib and Seaborn for visualization, making it highly efficient for handling and processing large datasets in a wide range of analysis tasks.

**Limitations:**

* Jupyter Notebook can be resource-intensive and may become slow when handling very large datasets or performing complex operations, as it runs within a browser environment, which may not be optimized for high-performance computing.
* While Python offers excellent libraries for data analysis, it may not perform as well as lower-level languages like C or Java in terms of computational speed, especially when dealing with very large-scale data or highly complex algorithms, unless optimized libraries are used.

## E.3 Application of Analytical Methods

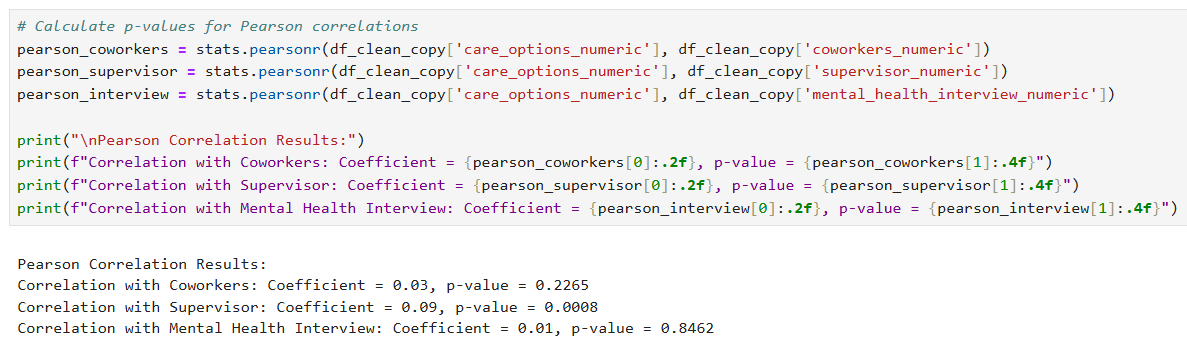
Below are the steps used in Jupyter Notebook to complete the data analysis in the research project:

1. Exploratory Data Analysis
   1. Imported necessary packages (Pandas, NumPy, scipy, matplotlib, seaborn)
   2. Imported dataset (survey.csv) using Pandas read\_csv.
   3. .head() function to view columns and first 5 rows of dataset
   4. .info() function to view number of total rows, column datatypes, and non-null value count
   5. .isnull().sum() functions to sum the null values in each column
   6. .value\_counts() to view each country participants were from and how many participants from each country
   7. Filtered the DataFrame for countries not equal to 'United States'
   8. Filtered the rows where 'Country' is 'United States' and 'state' is null
   9. Filtered for countries not equal to 'United States' and where 'state' is not null
2. Data Cleaning
   1. Dropped unrelated columns
      1. Created new dataframe with relevant columns
      2. Created a copy of new dataframe to perform data cleaning
      3. .head() function used to view first row to verify copy was created correctly
   2. Handled null values
      1. .isnull().sum() used to view sum of null values in each column
      2. Dropped rows of non-United States countries with values in 'state' column
      3. .info() function used to confirm rows were dropped
      4. Filtered for countries not equal to 'United States' and where 'state' is not null to confirm rows were dropped.
      5. Updated rows where Country is 'United States' and state is null with 'DC'
      6. .isnull().sum() function to check null values
      7. Filtered the DataFrame for countries not equal to 'United States'
   3. Converted categorical data into numerical data for correlation analysis
      1. Performed data normalization
         1. Gender column
            1. .unique() function to view unique values for 'Gender' column
            2. .replace() function to change values to Male, Female, or Other
            3. .unique() function to ensure gender values were changed
         2. Country column
            1. .apply(lambda x) function to change country to Other if not United States or United Kingdom
            2. .unique() function to check 'Country' column only contains values United States, United Kingdom, or Other
            3. .unique() to check 'state' column for data normalization
            4. .unique() to check 'care\_options' column for data normalization
            5. .unique() to check 'seek\_help' column for data normalization
            6. .unique() to check 'coworkers' column for data normalization
            7. .unique() to check 'supervisor' column for data normalization
            8. .unique() to check 'mental\_health\_interview' column for data normalization
      2. Converted Categorical Data
         1. .map() to convert columns to numerical values
         2. .head() to view first 5 rows to confirm numerical conversion
3. Correlation Analysis
   1. .corr(method=’pearson’) function to find correlation between availability of resources and willingness to discuss
   2. stats.pearsonr() used to calculate p-values for Pearson correlations
   3. if/else statements to compare p-values with alpha (0.05)
   4. Created matrix heatmap using matplotlib and seaborn
4. T-test
   1. Split data based on knowledge of care options (know and don’t know)
   2. .mean() to calculate mean comfort level for both groups (coworkers and supervisor)
   3. stats.ttest\_ind() to perform t-tests to compare comfort levels between the two groups
   4. if/else statements to compare the t-test p-values with alpha (0.05)
   5. Defined a function to create bar charts using def()
   6. Created bar charts using matplotlib
5. Saved cleaned dataframe as a CSV file
   1. .to\_csv() function

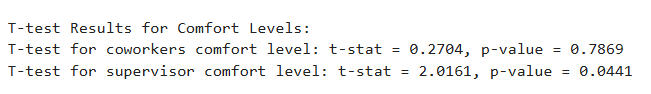
# F. Data Analysis Results

## F.1 Statistical Significance

* **Research Question Part A:** Correlation Between Mental Health Resources and Employees’ Willingness to Discuss Mental Health Issues
  + Null Hypothesis (H0): There is no correlation between the availability of mental health resources and employees' willingness to discuss mental health issues.
  + Statistical Test: Pearson Correlation
  + Metrics Generated:
    - Correlation Coefficient (R-value**)**: Measures the strength and direction of the linear relationship between the two variables (care options and willingness to discuss).
    - P-value: Assesses the statistical significance of the correlation.
  + Alpha Value (α): 0.05
  + Conclusion:
    - The following are the results of the Pearson Correlation test



* + - With p-values of 0.2265 and 0.8462, there is insufficient evidence to reject the null hypothesis for the correlation between mental health resources and employees’ willingness to discuss mental health issues with coworkers and interview. However, with a p-value of 0.0008, the correlation of employees’ willingness to discuss mental health issues with supervisors is statistically significant, supporting the claim that the availability of mental health resources has a significant but weak relationship with the willingness to discuss mental health issues with supervisors.
* **Research Question Part B**: Comfort Level Differences Based on Knowledge of Mental Health Care Options
  + Null Hypothesis (H0): There is no significant difference in comfort levels between employees who know their mental health care options and those who do not.
  + Statistical Test: Independent t-test
  + Metrics Generated:
    - T-statistic: Measures the difference in means between the two groups (those who know care options vs. those who do not).
    - P-value: Assesses the statistical significance of the difference between the means.
  + Alpha Value (α): 0.05
  + Conclusion:
    - The following are the results of the independent t-test



* + - "There is no significant difference in comfort levels with coworkers based on knowledge of mental health care options, as the t-test yielded a p-value of 0.7869 which is above 0.05. However, with a p-value of 0.0441. there is sufficient evidence to reject the null hypothesis for the comfort level with supervisors, suggesting that employees who know their mental health care options feel significantly more comfortable discussing mental health issues with supervisors.

## F.2 Practical Significance

The practical significance of the data analytics solution is to be assessed by determining how meaningful the findings are in a real-world context. In this project, the focus was on how mental health resource awareness influences employee comfort in discussing mental health issues with coworkers and supervisors. There was a finding that was statistically significant, but it is also crucial that the results translate into actionable changes for organizations. For example, since there was a significant relationship is found between resource awareness and communication comfort between employees and their supervisor, this would suggest that improving the visibility and accessibility of mental health resources can have a direct impact on workplace communication and overall employee well-being. Such results would have clear practical implications for how organizations structure and promote their mental health programs as well as structure the roles supervisors play within the mental health programs.

To judge the practical significance, I considered whether the analysis led to actionable recommendations that can improve workplace culture around mental health. This includes revising the way organizations communicate mental health benefits, training programs for managers, or developing anonymous reporting tools for employees to express their concerns. Practical significance was measured by how well these recommendations align with improving communication and comfort levels in real-world work settings. The difference in communication comfort between employees who are aware of mental health resources versus those who are not, was considered practically significant, as it directly supports the goal of creating a supportive workplace environment.

Finally, practical significance was evaluated by determining the extent to which these results support decision-making processes within organizations. Since the analysis can inform better policies around mental health resource distribution and engagement, then the findings have provided the expected benefits and have met the objectives of the project. For instance, companies may adopt new wellness programs or communication strategies based on the findings, helping foster an open and inclusive work culture. The ultimate measure of practical significance was whether the project’s insights can be meaningfully applied in practice to benefit both the organization and its employees, particularly in the context of improving mental health awareness and communication.

## F.3 Overall Success

The Pearson Correlation analysis revealed a statistically significant, although weak, correlation between knowing mental health care options and comfort in discussing mental health issues with supervisors. This supports the hypothesis that access to mental health resources positively influences employees' comfort levels with supervisors, demonstrating the project’s success in identifying a meaningful relationship. While the correlation with coworkers and during interviews was not significant, this outcome still contributes to understanding how resource availability affects employee behavior differently depending on the context.

The independent t-test confirmed that employees who know their mental health care options feel significantly more comfortable discussing mental health issues with their supervisors. The t-test results for discussing mental health with coworkers showed no significant difference, which aligns with the data from the correlation analysis. The project successfully identified a key finding: knowing mental health care options primarily affects comfort with supervisors, validating the hypothesis for this part of the research.

Overall, the project was successful in addressing the research question and providing statistically supported insights into the relationship between the availability of mental health resources and employees' comfort in discussing mental health issues at work. The findings can inform future organizational policies and strategies aimed at improving workplace mental health discussions.

# G. Conclusion

## G.1 Summary of Conclusions

The project produced several key conclusions regarding the relationship between mental health resources and employees' comfort in discussing mental health issues at work. First, there is a weak but statistically significant correlation between awareness of mental health care options and comfort levels when discussing mental health issues with supervisors. Employees who are aware of these resources feel more comfortable having these conversations with their supervisors, indicating the importance of resource availability in fostering open communication with management.

Second, no significant correlation was found between resource awareness and comfort levels when discussing mental health issues with coworkers or during interviews. This suggests that while employees may value mental health resources, they may not see them as directly influencing their comfort levels in peer or interview settings.

Finally, the project revealed a significant difference in comfort levels with supervisors between employees who knew their mental health care options and those who did not. Employees who were informed about available resources felt notably more comfortable discussing mental health with their supervisors. However, this difference was not observed in comfort levels with coworkers, where knowledge of resources did not significantly impact discussions.

Overall, the conclusions highlight the critical role that mental health resource awareness plays in improving communication with supervisors and suggest that more can be done to extend this comfort to other workplace relationships. Employers should consider prioritizing the promotion and visibility of mental health resources to ensure employees are well-informed, as this can significantly enhance trust and openness, especially with management. Moreover, while the impact on peer-level discussions and interviews was not as pronounced, these areas may benefit from additional interventions, such as mental health training, team-building exercises, or fostering a more inclusive culture where mental health is openly discussed and supported across all levels of the organization. This could ultimately contribute to a healthier, more supportive work environment, where employees feel comfortable discussing mental health regardless of who they are interacting with.

## G.2 Effective Storytelling

In this research project, two key visualizations were used to effectively convey insights: a bar chart and a heat map. Both visualizations were developed using Python's Matplotlib and Seaborn libraries, which are effective tools for creating clean, customizable, and insightful graphs. These visualizations were essential in supporting the storytelling process by clearly representing the relationships and patterns in the data.

The bar chart was used to compare the mean comfort levels of employees when discussing mental health issues with coworkers and supervisors. This chart visually highlighted the differences in comfort between employees who were aware of their mental health care options and those who were not. The distinction between the two groups was clearly illustrated through side-by-side bars, making it easy to see the significant difference in comfort levels with supervisors, while the negligible difference in comfort with coworkers was also apparent. This visualization supported the conclusion that awareness of mental health resources significantly impacts comfort levels with supervisors but not with coworkers.

The heat map, on the other hand, was utilized to display the correlations between variables such as mental health resource availability and willingness to discuss mental health issues at work. By using color gradients to represent the strength of correlations, the heat map provided an intuitive way to grasp the relationships between variables at a glance. The heat map clearly demonstrated the weak, yet statistically significant, correlation between mental health resource awareness and comfort levels with supervisors, while showing the lack of significant correlation with coworkers or during interviews.

## G.3 Recommended Courses of Action

Based on the results and conclusions of the data analysis, two key recommendations can be made to improve workplace discussions around mental health, expand mental health resource awareness programs and enhance supervisor training on mental health conversations.

The data showed a significant relationship between employees' knowledge of mental health resources and their comfort in discussing mental health issues with supervisors. To build on this finding, organizations should focus on expanding awareness programs that educate employees about available mental health care options. This could involve regular communication through internal newsletters, workshops, and employee handbooks that clearly outline these resources. By increasing awareness, employees are more likely to feel comfortable addressing mental health concerns with their supervisors, fostering a more open and supportive workplace culture. This recommendation aligns with the research finding that greater awareness of mental health resources significantly increases employees' comfort in discussing mental health issues with their supervisors.

The analysis also revealed that employees are more comfortable discussing mental health issues with supervisors when they are aware of mental health care options. Organizations should provide targeted training for supervisors to improve their ability to handle mental health conversations empathetically and effectively. This training should include how to approach such discussions, offer appropriate support, and guide employees to available resources. By equipping supervisors with the necessary skills, organizations can strengthen the trust between employees and management, encouraging more open dialogue around mental health. This recommendation is rooted in the need identified by the research to improve comfort levels in supervisor-employee interactions regarding mental health.

# H. Panopto Presentation

**Presentation Video Link**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7e22c89f-f778-4a70-a321-b1f301570857#>   
 

**References**

*Mental Health in Tech Survey*. (n.d.). Kaggle. https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

# Appendix A

# Supporting Files

Jupyter Notebook with data analysis code - [View Capstone Notebook](Capstone%20Notebook.html)