Mental Health in Tech Industry

Tomeka N. Pena

Western Governors University

Table of Contents

[A. Proposal Overview 3](#_Toc2137083979)

[A.1 Research Question or Organizational Need 3](#_Toc1815560889)

[A.2 Context and Background 3](#_Toc1384268585)

A.3 and A3A Summary of Published Works and Their Relation to the Project 5

[Review of Work 1 4](#_Toc694593955)

[Review of Work 2 4](#_Toc475195288)

Review of Work 3 5

A.4 Summary of Data Analytics Solution 6

[A.5 Benefits and Support of Decision-Making Process 7](#_Toc1116633891)

[B. Data Analytics Project Plan 8](#_Toc1561262715)

[B.1 Goals, Objectives, and Deliverables 8](#_Toc1217623361)

[B.2 Scope of Project 9](#_Toc427690838)

[B.2.A Included in Project Scope 9](#_Toc1243851001)

[B.2.B Not included in Project Scope 9](#_Toc2036717847)

[B.3 Standard Methodology 10](#_Toc204762054)

[B.4 Timeline and Milestones 13](#_Toc1049840195)

[B.5 Resources and Costs 14](#_Toc313275103)

[B.6 Criteria for Success 14](#_Toc1337049307)

[C. Design of Data Analytics Solution 15](#_Toc764513894)

[C.1 Hypothesis 15](#_Toc1169566000)

[C.2 and C.2.A Analytical Method 15](#_Toc1622461191)

[C.3 Tools and Environments 16](#_Toc53928856)

[C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance 17](#_Toc1386867513)

[C.5 Practical Significance 18](#_Toc157862960)

[C.6 Visual Communication 20](#_Toc1263386613)

[D. Description of Dataset 21](#_Toc34323025)

[D.1 Source of Data 21](#_Toc1867631323)

[D.2 Appropriateness of Dataset 21](#_Toc418252330)

[D.3 Data Collection Methods 22](#_Toc1665164903)

[D.4 Observations on Quality and Completeness of Data 22](#_Toc57902398)

[D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances 22](#_Toc1649406101)

[References 25](#_Toc1322159537)

# A. Proposal Overview

## A.1 Research Question or Organizational Need

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 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.

## A.2 Context and Background

Mental health in the workplace has become a critical issue, with companies more gradually recognizing the importance of mental health support as employees increasingly are seeking support for mental health challenges. Many organizations now provide resources and policies, but questions remain about how awareness of these resources impacts employees' willingness to discuss their mental health. This research is rooted in the context of promoting well-being in a corporate environment and aims to explore how access to mental health information affects communication and comfort levels within a company. It is crucial to understand this dynamic to better tailor mental health programs that effectively serve employees and enhance comfort and inclusivity surrounding the topic of mental health.

## A.3 and A3A Summary of Published Works and Their Relation to the Project

### Review of Work 1

### Mind Share Partners' 2021 *Mental Health at Work Report* provides a deep analysis of mental health trends within various workplaces across different industries. The report states, “76% of employees experienced at least one mental health symptom in 2020,” emphasizing the urgency for businesses to prioritize mental health initiatives (Mind Share Partners, 2021). It highlights how the pandemic has exacerbated mental health concerns, with younger employees and employees of color disproportionately affected. The report outlines various solutions such as increasing access to mental health resources, fostering open dialogues about mental health, and promoting flexible work environments. The findings serve as a call to action for companies to adopt more inclusive mental health policies that cater to the growing needs of their workforce.

### The research question in this project focuses on how awareness of mental health resources impacts employee comfort in discussing mental health issues with supervisors and coworkers. The Mind Share Partners report directly relates to this by showing how companies that actively promote mental health awareness have better outcomes in employee engagement and overall well-being. The report also suggests that employees who know about available mental health resources are more likely to utilize them, which aligns with the project’s goal of understanding the relationship between awareness and open communication. This work provides a foundational perspective on the positive outcomes of increasing mental health awareness and resource accessibility in the workplace.

### Review of Work 2

### The article *“Mental Health in Tech: Analysis of Workplace Risk Factors and Impact of*

*COVID-19"* presents a comprehensive analysis of the factors influencing mental health in the tech industry, with an emphasis on workplace conditions and the exacerbating impact of COVID-19. The authors highlight the unique stressors faced by tech workers, such as extended working hours, unrealistic deadlines, and a competitive culture, which often lead to burnout and mental health deterioration. The article further explores how the pandemic intensified these challenges, as tech workers faced isolation, increased workloads, and blurred boundaries between work and personal life. According to the authors, "COVID-19 has disrupted workplace dynamics and has increased stress levels, leading to higher rates of anxiety and depression among tech professionals" (Mitravinda et al., 2023). The study also provides recommendations for employers, suggesting the implementation of mental health resources and flexible working conditions to mitigate these challenges.

### This article closely relates to the project's research question as it emphasizes the importance of awareness in reducing stigma and promoting open communication about mental health in the workplace. The tech industry, known for high stress, is a prime example of how knowledge of mental health resources can directly influence employee comfort in discussing these issues with supervisors and peers. The article reinforces the need for companies to actively promote mental health awareness, echoing the project's objective of investigating the impact of resource knowledge on workplace discussions about mental health.

### Review of Work 3

Deloitte’s *Mental Health and Employers: The Case for Investment* focuses on the economic benefits of mental health support in the workplace, demonstrating that companies investing in employee mental health see significant returns. The report shows that “for every £1 spent on mental health support, businesses see a return of £5,” highlighting the financial incentives for prioritizing mental health (Abraham et al., 2022). The report also discusses the effectiveness of mental health programs in improving productivity, reducing absenteeism, and increasing employee engagement. By making mental health a core part of their business strategy, organizations can create more supportive and productive work environments.

This report is relevant to the project, because it provides a financial perspective on the research question, which focuses on awareness of mental health resources and its impact on communication in the workplace in the United Kingdom. In addition, the United Kingdom had the second largest participants in the study and including research from that geographical location is pertinent. Deloitte’s findings underline the importance of promoting mental health programs, not only for employee well-being but also for company productivity and ROI. By linking mental health support to business outcomes, the report strengthens the case for increasing awareness and communication about mental health resources, aligning directly with the project's goal of assessing how knowledge of these resources affects workplace comfort in discussing mental health issues.

## A.4 Summary of Data Analytics Solution

I will first begin with data wrangling and will download the dataset from Kaggle.com as a csv file in a zip file. Afterwards, I will upload the dataset into a dataframe in Jupyter Notebook to perform exploratory data analysis and data cleaning to prepare the data for analysis.

The data analytic solution for this project focuses on conducting correlation analysis and t-tests to address two key questions: the relationship between the availability of mental health resources and willingness to discuss mental health issues, and the difference in comfort levels in discussing mental health based on knowledge of care options. The first analysis will use Pearson Correlation to assess the relationship between the availability of mental health resources and employees' willingness to discuss mental health issues. This test will provide a correlation coefficient that indicates the strength and direction of the linear relationship between these two variables. The second analysis will utilize a t-test to determine if there is a statistically significant difference in comfort levels between employees who are aware of their mental health care options and those who are not. Both methods will be implemented using Python and appropriate packages like pandas, scipy, and seaborn for data manipulation, statistical tests, and visualization.

For visualization, correlation matrices and box plots will be used to provide a clear understanding of the relationships between the variables. Tools like Matplotlib and Seaborn will help in plotting these relationships, ensuring that findings are clearly communicated. This approach ensures a robust examination of the data, allowing for both statistical significance and practical insights into how mental health resources and knowledge affect workplace discussions. By integrating these methods, the solution will directly address the research questions and help organizations make data-driven decisions about mental health support initiatives.

## A.5 Benefits and Support of Decision-Making Process

The data analytics solution proposed will offer significant benefits by providing actionable insights into how mental health awareness programs influence workplace culture. The findings will empower companies to make informed decisions regarding the implementation and communication of mental health resources. Specifically, this analysis can reveal whether improving awareness of mental health resources directly leads to higher levels of comfort in employees discussing mental health issues. This understanding can influence decisions regarding investment in mental health programs, communications strategy, and the types of resources that are most effective. In addition, by identifying gaps in resource awareness, the solution can guide companies on how to better promote their offerings, potentially leading to a healthier and more transparent work environment. The analysis also serves as a preventive measure by showing organizations areas where employees might not feel supported, enabling them to address concerns before they affect productivity or employee retention. Moreover, these data-driven insights offer a concrete justification for adjusting policies or allocating resources toward improving mental health support. Overall, this research will aid in strategic decision-making that prioritizes employee well-being, leading to a more engaged and resilient workforce.

# B. Data Analytics Project Plan

## B.1 Goals, Objectives, and Deliverables

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.

* Goal 1: To analyze the impact of mental health resource awareness on employees' comfort in discussing mental health issues at work.
  + Objective 1.1: Investigate the correlation between mental health resource availability and employee comfort with coworkers and supervisors
    - Deliverable 1.1.1: A detailed report on the correlation analysis.
  + Objective 1.2: Compare comfort levels between employees who know their mental health options and those who do not.
    - Deliverable 1.2.1: T-test results comparing the two employee groups.

## B.2 Scope of Project

### B.2.A Included in Project Scope

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.

### B.2.B Not included in Project Scope

The project will not include the development of new mental health resources or direct implementation of recommended changes in organizations. While the analysis will provide suggestions for improving communication strategies, it will not involve conducting interviews with employees or testing the implementation of these strategies in a real-world setting. The project also does not account for longitudinal studies to assess how changes in awareness or communication impact mental health outcomes over time. Furthermore, it will not delve into the specific mental health diagnoses of employees, focusing instead on broader trends in resource awareness and communication comfort. These elements are outside the scope of this analysis but may be relevant for future studies.

## B.3 Standard Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework 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 six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase focuses 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 can achieve clarity in the objectives of the data analysis and ensure that each step is aligned with solving the research question.

Below will list how these six phases will be implemented to complete this project using the CRISP-DM framework.

**Business Understanding**  
 In this phase, we will begin by clearly defining the objectives of the research and understanding how mental health resource awareness affects workplace communication. The goal is to understand whether employees who are more knowledgeable about their company’s mental health resources feel more comfortable discussing mental health issues with their coworkers and supervisors. During this phase, we will work closely with stakeholders to refine the research question and establish the specific hypotheses to be tested. We will also outline the key performance indicators (KPIs) that will measure the project’s success. This step ensures alignment between the project’s goals and the needs of stakeholders, guiding the subsequent phases toward answering the business or research questions.

**Data Understanding**  
 In this phase, the focus will be on collecting and exploring the data to develop an initial understanding of its structure, quality, and relevance to the research question. The dataset for this project includes employee survey responses regarding mental health resource awareness, communication comfort, and other contextual variables such as company size and industry. The data will be assessed for completeness, missing values, and outliers. Initial exploratory data analysis (EDA) will be conducted using Python libraries such as Pandas and Seaborn to identify trends and relationships between the variables. Additionally, descriptive statistics and visualizations will be used to provide a high-level overview of the dataset and any immediate insights into the relationships that could be explored further during modeling. The insights gained here will inform how to proceed with data cleaning and preparation.

**Data Preparation**  
 Data preparation is one of the most critical steps, where the data will be cleaned and transformed to ensure it is ready for analysis. During this phase, we will handle missing values by either imputing data or excluding incomplete rows, depending on the distribution of missing information. Categorical variables, such as company size and industry, will be encoded into numerical format using techniques like one-hot encoding to make them suitable for machine learning models. We will also perform feature scaling to normalize variables, ensuring they are comparable across different ranges. Additionally, this phase will involve splitting the data into training and testing sets, which will be essential during the modeling phase. The end result will be a clean, structured dataset ready for analysis.

**Modeling**  
 This phase involves selecting and applying appropriate statistical models to analyze the relationship between mental health resource awareness and employees’ comfort in discussing mental health issues. We will use supervised machine learning models, such as logistic regression or decision trees, to predict communication comfort based on awareness levels and other contextual factors. The scikit-learn library in Python will be used for model development and evaluation. Additionally, statistical tests, such as correlation and regression analysis, will be employed to examine whether significant relationships exist between the variables. Model parameters will be tuned through techniques like cross-validation to ensure the models provide accurate and reliable predictions. At the end of this phase, we will evaluate the performance of these models using metrics like accuracy, precision, recall, or R-squared values.

**Evaluation**  
 Once the models are developed, they need to be evaluated to ensure they meet the project objectives. In this phase, the performance of each model will be assessed using the test data set to see how well it generalizes to new, unseen data. Evaluation metrics such as confusion matrices, F1-scores, and ROC-AUC curves will be calculated to determine the effectiveness of the models. Additionally, statistical significance tests will be applied to verify if the relationships identified in the modeling phase are not due to random chance. The evaluation process will also involve checking if the model aligns with the original business objectives and research questions, ensuring that the insights derived from the analysis are relevant and actionable for stakeholders. Any necessary adjustments or refinements to the models will be made based on the evaluation outcomes.

**Deployment**  
 In the deployment phase, the insights and models developed throughout the project will be shared with stakeholders in a clear, accessible format. This may include interactive dashboards created using Tableau, detailed reports summarizing the analysis, and presentations highlighting key findings and recommendations. The goal of this phase is to ensure that the results are actionable and can inform future decisions within the organization regarding mental health resource awareness and communication strategies. Additionally, any code or models developed during the project will be shared with the organization for future use or replication in similar analyses. This phase ensures the project’s outcomes are implemented in practice and integrated into the organization’s decision-making processes, providing lasting value.

Each of these phases is critical to ensuring a successful data analytics project that answers the research question and delivers actionable insights to stakeholders. The CRISP-DM framework provides a structured path that ensures the project remains focused and methodologically sound, leading to meaningful and reliable results.

## B.4 Timeline and Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Projected Time** | **Projected start date** | **Anticipated end date** |
| Retrieve and Clean Dataset | 8 Hrs. | 09/16/2024 | 09/16/2024 |
| Evaluate and Analyze Dataset | 12 Hrs. | 09/17/2024 | 09/18/2024 |
| Create Visualizations | 6 Hrs. | 09/19/2024 | 09/19/2024 |
| Create Slideshow of Analysis Results | 5 Hrs. | 09/19/2024 | 09/19/2024 |
| Present Slideshow of Analysis Results | 1 Hrs. | 09/20/2024 | 09/20/2024 |

## B.5 Resources and Costs

1. Hardware: $0
2. Python programming language: $0
3. Jupyter Notebook: $0
4. Tableau: $0
5. PowerPoint: $0
6. Labor hours: $0

There are no projected costs for this project. Hardware items that will be used are already owned by me. Python programming language, Jupyter Notebook and Tableau are available to for public use for free. PowerPoint is already installed on my computer and does not require additional cost for downloading. Lastly, I will be conducting all labor for this project, which will not incur additional cost as well.

## B.6 Criteria for Success

Success for this project will be evaluated based on specific, measurable criteria that ensure the analysis is both rigorous and actionable. One primary criterion is the completion of the data analysis process according to the CRISP-DM methodology, ensuring that each phase is executed thoroughly and correctly. Another key metric for success will be the statistical significance of the relationship between mental health resource awareness and employee comfort in discussing mental health issues. Success will also be determined by the clarity and usefulness of the visualizations and insights generated from the analysis. For instance, a model will be considered successful if it achieves an R-squared value of 0.6 or higher, indicating a strong correlation between the variables. Additionally, the project will be deemed successful if the final report includes actionable recommendations that align with the research question, offering tangible solutions for organizations. The ability to influence decision-making processes and improve workplace mental health policies based on these insights will be the ultimate measure of success.

# C. Design of Data Analytics Solution

## C.1 Hypothesis

Null hypothesis: There is no correlation between the availability of mental health resources and employees’ willingness to discuss mental health issues at work nor is there a significant difference in employee comfort levels in discussing mental health issues between those who know their mental health care options and those who do not.

Alternative hypothesis: There is a significant correlation between the availability of mental health resources and employees’ willingness to discuss mental health issues at work and employees who know their mental health care options have significantly different comfort levels in discussing mental health issues compared to those who do not.

## C.2 and C.2.A Analytical Method

The correlation analysis is an appropriate method for this project because it allows 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 provides a robust framework for understanding the relationship between them. This method is particularly effective for providing an overview of the data and identifying trends that can be further investigated. Additionally, the Pearson correlation coefficient offers an intuitive metric that stakeholders can easily interpret, making it a valuable tool for decision-making regarding mental health initiatives.

The t-test is justified as it directly compares the means of two independent groups, making 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 can determine if the mean comfort levels of employees who are aware of mental health resources differ significantly from those who are not. This method is particularly relevant because it not only tests for statistical significance but also provides practical insights into how interventions (such as increasing resource awareness) could improve employee well-being. The results from the t-test will offer concrete evidence for whether expanding mental health resources in the workplace can effectively enhance communication and openness around mental health issues.

## C.3 Tools and Environments

The tools that I will be using to complete this project are Jupyter Notebook and Python. Jupyter Notebook will be used to perform exploratory data analysis (EDA), data cleaning and transformation, data analysis, and data visualization. Python is the programming language that will be used to clean, structure, and preprocess the dataset. It is 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 will help in handling the dataset by filtering relevant columns, dealing with missing values, and performing group-by operations. Visualization libraries such as Matplotlib or Seaborn will be used to graphically represent the data and identify trends.

## C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

To evaluate the statistical significance of the findings, two methods will be employed: Pearson Correlation and an independent t-test. For the first research question, the null hypothesis will be that there is no correlation between the availability of mental health resources and employees' willingness to discuss mental health issues. The planned statistical test is Pearson Correlation, which generates a correlation coefficient (R-value) that ranges from -1 to 1, where values closer to 1 or -1 indicate a strong positive or negative correlation, respectively. The significance of the correlation will be assessed using a p-value, and an alpha value (α) of 0.05 will be used as the threshold for statistical significance. If the p-value is less than 0.05, the null hypothesis will be rejected, indicating a statistically significant relationship between the two variables.

For the second research question, an independent t-test will be conducted to compare the comfort levels of employees in discussing mental health issues based on whether they know their mental health care options or not. The null hypothesis for this test is that there is no significant difference between the two groups in terms of comfort levels. The test will generate a t-statistic and a p-value to determine whether the difference in means between the two groups is statistically significant. As with the correlation test, an alpha value of 0.05 will be used to evaluate the significance of the result. If the p-value is below this threshold, the null hypothesis will be rejected, indicating a significant difference in comfort levels based on knowledge of mental health care options. The metrics for these tests will be used to evaluate whether the findings are statistically significant and provide insight into the relationships explored in the project.

**C.4.A Justification for Methods and Metrics**

The Pearson Correlation test is an appropriate choice for the first research question because it specifically evaluates the linear relationship between two continuous or ordinal variables. In this case, the availability of mental health resources and willingness to discuss mental health issues are measured on scales that allow for linear relationship analysis. Since the goal is to assess whether there is a correlation between these variables, Pearson Correlation is the most suitable method. The correlation coefficient will indicate the strength of the relationship, while the p-value will determine whether the correlation is statistically significant.

The independent t-test is chosen for the second research question because it compares the means of two independent groups. In this case, the two groups are employees who know their mental health care options and those who do not, while the dependent variable is their comfort level in discussing mental health issues. The t-test will allow for a direct comparison of comfort levels across the two groups, determining whether the observed differences are statistically significant. The t-statistic will help evaluate the size of the difference, while the p-value will indicate whether this difference is statistically significant at the chosen alpha level of 0.05..

## C.5 Practical Significance

The practical significance of the data analytics solution will be assessed by determining how meaningful the findings are in a real-world context. In this project, the focus is on how mental health resource awareness influences employee comfort in discussing mental health issues with coworkers and supervisors. A finding may be statistically significant, but it is also crucial that the results translate into actionable changes for organizations. For example, if a significant relationship is found between resource awareness and communication comfort, 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.

To judge the practical significance, we will consider whether the analysis leads to actionable recommendations that can improve workplace culture around mental health. This could include 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 will be measured by how well these recommendations align with improving communication and comfort levels in real-world work settings. A large enough difference in communication comfort between employees who are aware of mental health resources versus those who are not will be considered practically significant, as it directly supports the goal of creating a supportive workplace environment.

Finally, practical significance will be evaluated by determining the extent to which these results support decision-making processes within organizations. If the analysis can inform better policies around mental health resource distribution and engagement, then the findings will have provided the expected benefits and 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 will be 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

## C.6 Visual Communication

To effectively communicate the findings of this project, two key visualizations will be created, each designed to summarize important aspects of the data analysis and modeling phases. The first visualization will be a bar chart that displays the distribution of employees’ awareness of mental health resources by company size. This visualization will clearly show how resource awareness varies across different organizations, helping to identify whether larger companies provide more visible resources compared to smaller ones. This bar chart will be generated using Python's Matplotlib library and will serve as an important tool for understanding the initial data distribution.

The second visualization will be a confusion matrix to depict the performance of the logistic regression model used to predict employees' comfort in discussing mental health based on resource awareness. The confusion matrix will visually communicate the accuracy of the model, showing the true positives, false positives, true negatives, and false negatives. This visualization will help stakeholders assess the model’s effectiveness in predicting communication comfort and provide insight into where the model might be underperforming. This chart will be produced using Seaborn in Python, a powerful tool for creating statistical plots.

Both visualizations will play crucial roles in the final project report and presentations. The bar chart will highlight key insights from the data, while the confusion matrix will allow for an easy understanding of the model’s performance and accuracy. Together, these tools will visually support the project’s narrative and findings, making it easier for stakeholders to grasp the key outcomes and implications of the data analysis solution.

# D. Description of Dataset

## D.1 Source of Data

I will be using a dataset called Mental Health in Tech Survey: Survey on Mental Health in the Tech Workplace in 2014. This dataset consists of 27 columns and 1259 rows that contains the necessary fields to answer my research question. The dataset is made publicly available and is provided by Kaggle. It is accessible as unrestricted data to Kaggle users and is downloadable as a zip file directly from the site page. The dataset can be found at: <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey>

## D.2 Appropriateness of Dataset

The dataset selected for this project is appropriate for the goals outlined in the research. It contains the necessary information regarding employees' mental health, awareness of mental health resources, and comfort levels discussing mental health in the workplace, which are central to the research question. The dataset includes a variety of variables, such as employee demographics, awareness of resources, and workplace factors, allowing for both exploratory and inferential analyses. The inclusion of continuous variables, such as comfort levels, supports correlation and regression analyses, while categorical variables, such as resource awareness, allow for hypothesis testing using t-tests. This wide range of data supports the project’s goal of understanding how resource awareness impacts mental health communication, making it a robust choice for answering the research questions. Moreover, the dataset is of sufficient size and quality to ensure statistical validity. The large sample size increases the power of the statistical tests, reducing the likelihood of Type I and Type II errors.

## D.3 Data Collection Methods

The data was collected by downloading the survey.csv file from Kaggle.com. The dataset is downloadable as a zip file and is public accessible to Kaggle members. After downloading, the dataset will be uploaded into Jupyter Notebook using panda read\_csv.

## D.4 Observations on Quality and Completeness of Data

The quality of this dataset is suitable for the research project. It contains many usable columns need to conduct analysis such as: care\_options, wellness\_program, seek\_help, coworkers, and supervisors. The completeness of this dataset is relatively good overall, but does require some work to transform it into a suitable dataset for the analysis that will be performed. There are several columns that contains null values, the gender column contains multiple values (including mis-spelled values) and requires data normalization to be perform in order to create a consistent format. In addition, categorical data will need to be converted into numerical values to conduct correlation analysis.

## D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

**Data Governance**

Proper data governance ensures that the data is managed in a way that is compliant with legal, ethical, and security standards. For this project, the dataset will be stored and managed using a data governance framework that ensures accuracy, consistency, and accessibility for authorized individuals. Policies will be put in place to maintain the integrity of the data, with regular audits to ensure compliance with these governance principles. These practices guarantee that data is used in a way that aligns with organizational goals and ethical standards, preventing misuse or corruption of the dataset.

**Privacy**

Privacy is a significant consideration when working with sensitive employee data, especially when it involves mental health. To protect the privacy of individuals, all data will be anonymized before analysis, ensuring that personal identifiers, such as names and contact details, are removed or encrypted. Compliance with the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) will be ensured if the dataset involves sensitive health data. Anonymization will ensure that even if the data is exposed, individuals cannot be identified, reducing the risk of privacy violations.

**Security**

To protect the dataset from unauthorized access, robust security measures will be implemented. These include encryption of the data at rest and in transit, access control mechanisms, and regular security audits to identify potential vulnerabilities. Only authorized personnel will be able to access the data, and multi-factor authentication will be used to further safeguard access. Additionally, regular software updates will be applied to prevent security breaches, and an incident response plan will be developed in case of a data breach.

**Ethical, Legal, and Regulatory Compliance**

From an ethical standpoint, the research will ensure transparency and fairness in the use of data. Participants in the original dataset collection were likely informed about the purpose of the study, and their consent was obtained. The project will also adhere to any relevant legal frameworks, such as the GDPR for European data and HIPAA for health-related data in the U.S. The research will ensure that data is used ethically, respecting participants' rights and confidentiality. Regulatory compliance will be ensured by following guidelines such as the Institutional Review Board (IRB) protocols if applicable.

**D.5.A Precautions for Managing Risks**

Several precautions will be taken to manage risks related to data governance, privacy, security, and ethical considerations. For data governance, regular reviews will ensure data is consistently stored, updated, and maintained according to best practices. Clear policies will be established for accessing and modifying the data to prevent accidental alterations. In terms of privacy, anonymization and de-identification techniques will be employed, ensuring no sensitive information can be traced back to individuals. Consent forms will be reviewed to ensure participants were adequately informed during data collection. Security will be managed with encryption and strong access controls. All data will be encrypted both at rest and in transit, and access will be restricted to authorized personnel only. Additionally, routine security checks and audits will be conducted to ensure that vulnerabilities are detected early and addressed promptly. Lastly for ethical, legal, and regulatory compliance, regular audits will ensure that all relevant legal guidelines (such as GDPR or HIPAA) are being followed. Participants' data will only be used in ways that align with the consent they provided during data collection, and no data will be shared without proper anonymization or de-identification.

# References

Mind Share Partners. (2021). 2021 Mental Health at Work Report. Retrieved from <https://www.mindsharepartners.org/_files/ugd/94aa4e_936405820c69423992db0be60fdac232.pdf>

Mitravinda, K. M., Nair, D. S., & Srinivasacan, G. (2023). Mental health in tech: Analysis of workplace risk factors and impact of COVID-19. *Journal of Workplace Mental Health*, *13*(1), 24-37. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9907184/>

Abrahams, N., Assal, V., Hampson, E., Polner, A. (2022). Mental Health and Employers: The Case for Investment. Deloitte. Retrieved from <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-uk-mental-health-report-2022.pdf>