Data-Driven Insights for HR: Understanding Mental Health Challenges in Technology Workplaces

Machine Learning - Unsupervised Learning and Feature Engineering DLBDSMLUSL01

Philipp Mc Guire

Abstract

This case study explores mental health challenges in technology-related jobs using survey data from employees. The analysis employs clustering techniques to categorize participants based on their responses, identifying 15 distinct groups with shared characteristics. Key findings highlight the prevalence of stress, social anxiety, and workplace discrimination fears, with specific job roles (e.g., backend developers, sales professionals) exhibiting unique mental health concerns. The study also reveals regional and age-related trends, emphasizing the need for targeted HR interventions. Correlation analysis further uncovers significant relationships between mental health diagnoses, workplace factors, and job roles. These results provide an initial foundation for identifying key areas for action, taking a closer look where necessary, and developing the first structured approaches to improving workplace mental health.

Table of Contents

List of Figures

1. Introduction

Mental health issues in the workplace are an increasing concern, especially in technology-related jobs where employees often experience high stress, long working hours, and burnout. The Human Resources (HR) department at our company has initiated a preemptive mental health program to mitigate these issues and enhance employee well-being. As part of this initiative, the HR department has tasked the data science team with conducting a quantitative analysis of a representative survey to identify key trends, categorize employees based on their responses, and provide interpretable visualizations to inform policy decisions.

The primary challenge in working with the provided dataset lies in its high dimensionality, complexity, and presence of missing or non-standardized textual inputs. Our objective is to develop a systematic approach to processing this dataset, reducing its complexity while preserving its core characteristics, and identifying distinct clusters of employees based on their responses.

2. Data Exploration and Challenges

The dataset analyzed in this study originates from a large-scale survey conducted mostly among technology professionals. It includes a diverse set of features covering demographics (age, gender, region), employment attributes (job role, employer type), and mental health-related responses (self-reported diagnoses, workplace support, personal experiences). Given the high dimensionality of the dataset and the variety of response types, an extensive data preprocessing and exploratory analysis was necessary to ensure meaningful insights.

2.1 Key Challenges

One of the primary challenges was the high dimensionality of the dataset, as it contained a mix of categorical and numerical variables. Many responses were in binary format, such as yes/no answers, while others were ordinal, requiring transformation for effective interpretation. Certain variables related to mental health diagnoses and workplace conditions were interdependent, making it necessary to reduce dimensionality while preserving key information. To address this, feature selection techniques and clustering were applied to retain only the most relevant features.

Another major issue was missing values, as many participants had skipped specific questions, leading to gaps in the dataset. Missing categorical responses were replaced with "unknown" to maintain consistency, while missing age values were imputed using the median after plausibility checks.

As seen in figure 1, missing responses frequently appeared in the following questions:

- Q19: *If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?*
- Q23: *If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?*
- Q18: *If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?*
- Q21: *If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?*
- Q17: *Do you know local or online resources to seek help for a mental health disorder?*
- Q20: *If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?*
- Q22: *Do you believe your productivity is ever affected by a mental health issue?*

These questions exhibited a high proportion of NaNs, which were subsequently replaced with *"unknown"* for analysis purposes to ensure consistency and maintain interpretability.

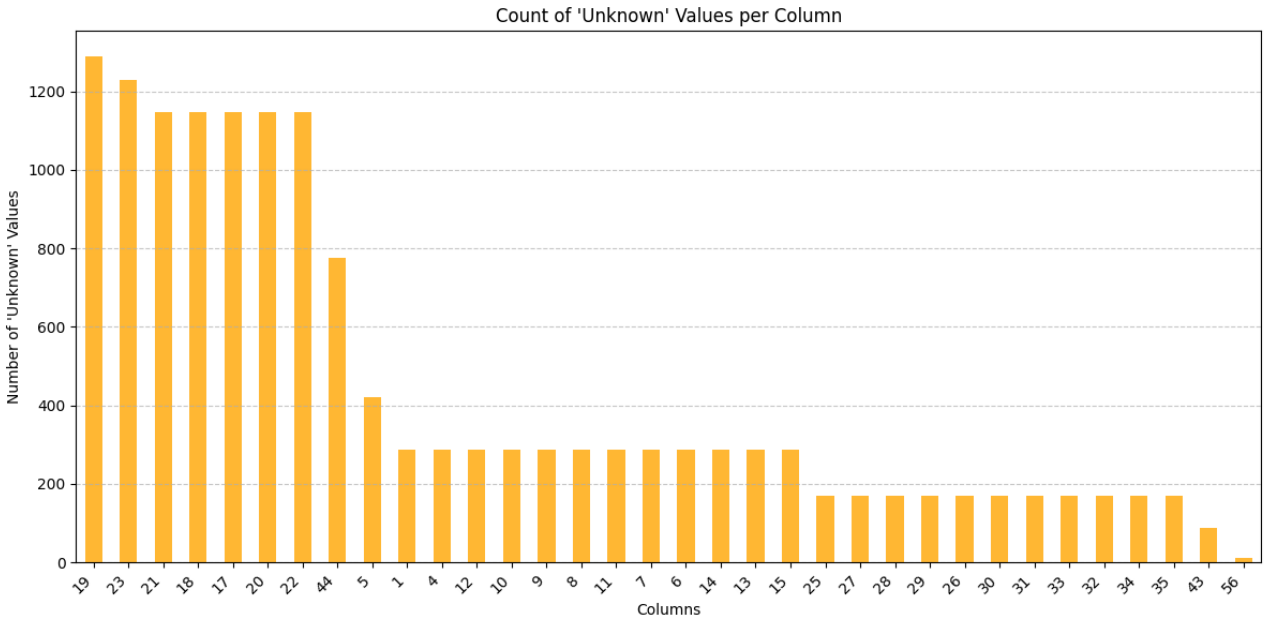

Figure 1: A bar chart was generated to visualize the distribution of missing data across features.

A further challenge was the unstructured textual responses within open-ended survey questions. Many participants provided free-text answers to questions such as *"Would you be willing to bring up a physical health issue with a potential employer in an interview? Why or why not?"*. These responses were highly variable, making direct analysis difficult. Instead of manually standardizing the text, Latent Dirichlet Allocation (LDA) was applied to extract underlying topics from these responses.

TF-IDF vectorization and stopword removal were applied before running Latent Dirichlet Allocation (LDA) to extract meaningful topics. The top 10 most significant words per topic were used for

manual labeling, ensuring interpretability. This data-driven approach, combined with stopword removal, ensured that themes emerged organically from the survey responses rather than being predefined, allowing for a more accurate differentiation of topics.

Another key consideration was the complex correlation structure within the dataset. Many features were highly interrelated, making it challenging to isolate independent effects. The correlation analysis focused on identifying relationships between clusters and the top 10 correlated features within the dataset.

While correlation analysis provides valuable insights into linear relationships between features, it does not inherently reveal complex, nonlinear patterns or subgroups within the dataset. This is where clustering analysis, explained later, adds further depth. By grouping individuals based on similar response patterns, clustering enables the identification of distinct employee profiles, allowing for a more nuanced understanding of how different workplace factors interact with mental health conditions. For instance, clustering can reveal hidden groupings where individuals with similar workplace experiences and stressors tend to share specific mental health challenges, even if these relationships are not apparent through simple correlation analysis.

Following the preprocessing steps, exploratory data analysis (EDA) was conducted to uncover distribution patterns before applying clustering techniques. The age distribution analysis revealed that most participants fell within the 26-35 and 36-45 age groups, with very few respondents over 55. This suggested that mental health concerns might be more prevalent among younger professionals. A histogram was generated to visualize this trend.

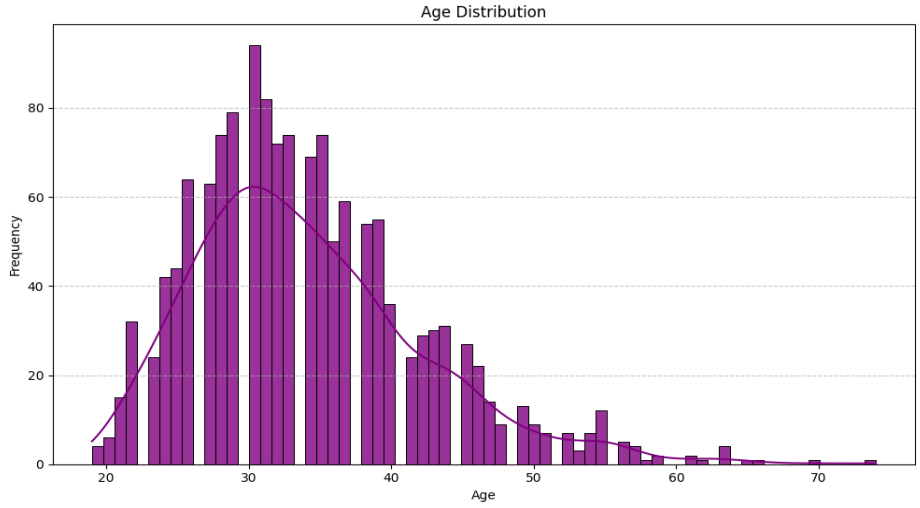

Figure 2: Age Distribution of Survey Participants

Additionally, regional distribution analysis showed that the majority of respondents were from North America and Europe, with smaller representation from other continents. This was an important factor in understanding workplace mental health trends across different regions, which was visualized through bar charts and pie charts.

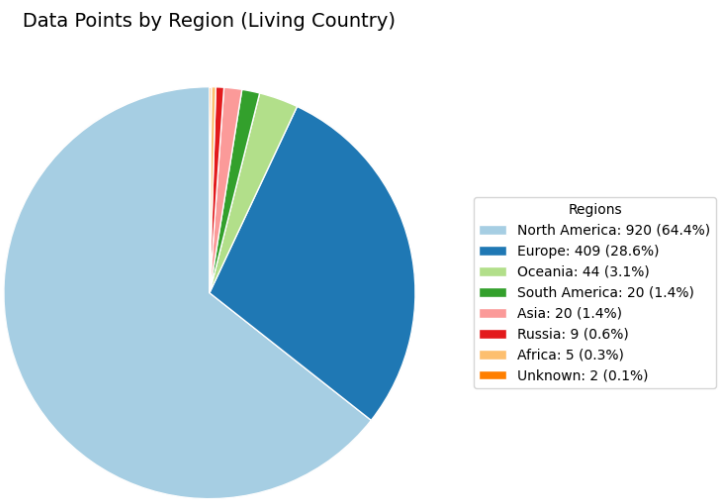

Figure 3: Regional Distribution of Survey Participants

An analysis of self-reported mental health diagnoses revealed that mood disorders, anxiety disorders, and ADHD were the most commonly reported conditions among respondents, reinforcing the prevalence of stress-related conditions in technology professions. These insights were summarized using a bar chart to depict the proportion of each diagnosis in the dataset.

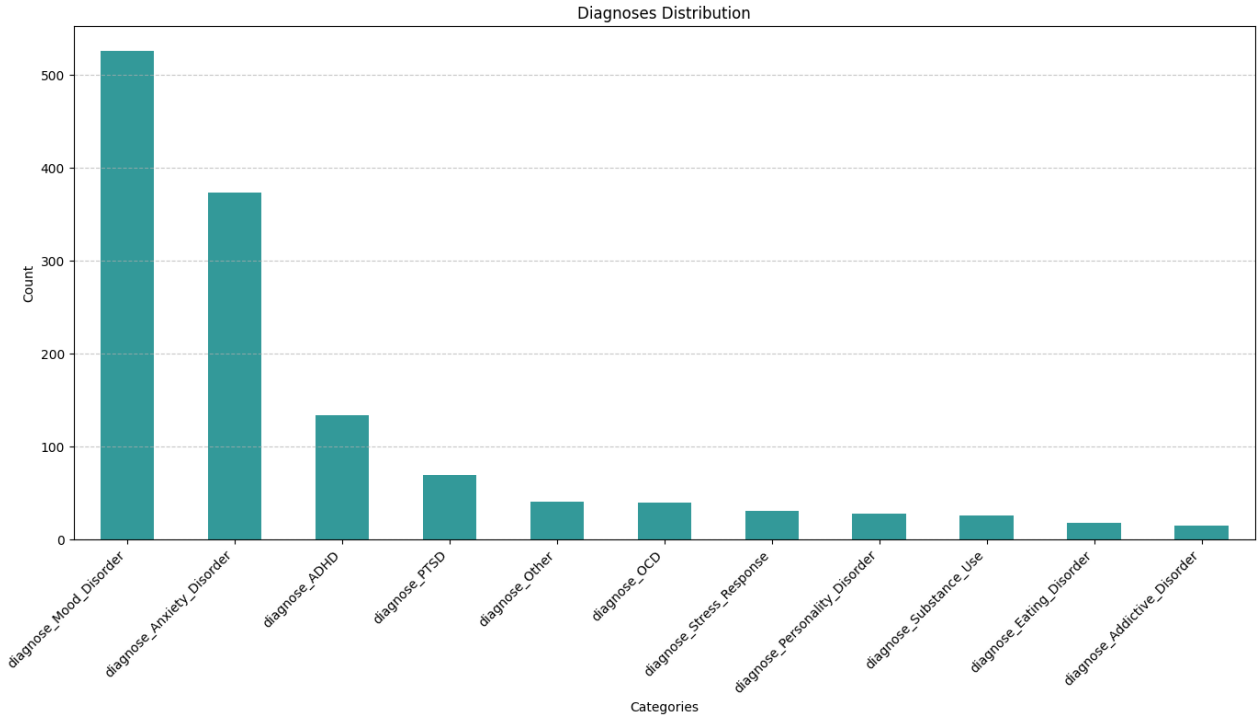

Figure 4: Diagnoses Distribution of Survey Participants

In summary, the challenges posed by high dimensionality, missing values, unstructured text inputs, and correlation complexity were systematically addressed through advanced preprocessing techniques and exploratory analysis. These efforts ensured that the dataset was clean, structured, and suitable for clustering, ultimately leading to meaningful insights that could support HR strategies for improving workplace mental health policies.

3. Clustering Approach and Methodology

To identify distinct groups within the dataset, an unsupervised machine learning approach was applied, utilizing clustering techniques to segment participants based on their responses. Given the complexity of the dataset, a structured methodology was followed to ensure meaningful clustering results. This involved several key steps, including data preprocessing, model selection, cluster evaluation, and visualization to enhance interpretability.

3.1 Data Preprocessing

Before applying clustering algorithms, extensive preprocessing was required to standardize the dataset and optimize performance. Normalization was performed on continuous variables, such as age, to ensure that numerical features were comparable and did not disproportionately influence the clustering process. This was particularly important for avoiding bias in distance-based algorithms like K-Means is.

Since the dataset contained numerous categorical variables, categorical encoding was necessary. Text-based responses related to job roles, living/working country, and mental health topics were transformed using one-hot encoding, allowing categorical data to be represented numerically. This enabled the clustering algorithm to recognize patterns in non-numerical features without introducing artificial ordering. Additionally, Highly correlated or redundant features were removed to enhance model performance.

3.1.1 Why Excluding Column 24: "Do you have previous employers?"

Column 24 ("Do you have previous employers?") was excluded from the dataset to improve the quality of the clustering analysis. Since this variable was present across all clusters, it did not contribute to distinguishing meaningful groupings within the data. Instead, its inclusion could have added unnecessary noise, potentially influencing the clustering algorithm without providing useful differentiation. By removing such broadly present variables, the clustering process focused more effectively on features that revealed distinct mental health and workplace patterns, leading to better-defined and more interpretable clusters.

3.2 Clustering Algorithm

To determine the optimal number of clusters, Silhouette Analysis was employed to assess cluster cohesion and separation, ensuring that each cluster was distinct while maintaining internal consistency. Additionally, Principal Component Analysis (PCA) was applied for dimensionality reduction, enhancing interpretability by projecting high-dimensional data onto a lower-dimensional

space, which further supported the clustering process and visualization. These combined techniques provided a robust foundation for meaningful segmentation.

Silhouette Analysis evaluates how similar an object is to its own cluster compared to other clusters, with a higher silhouette score indicating better-defined clusters. This method is commonly used to determine the optimal number of clusters in K-Means clustering (scikit-learn developers).

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional form while preserving as much variance as possible. Combining PCA with K-Means clustering can improve segmentation results by reducing noise and improving performance (Kaloyanova, 2024).

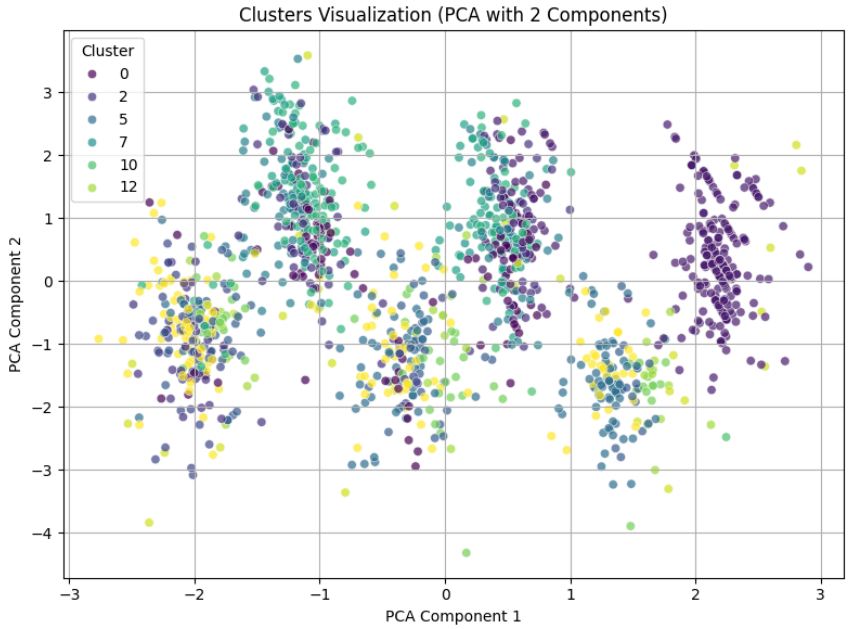

Figure 5: Cluster Distribution Visualization Using PCA

To determine the optimal number of principal components, we analyzed the cumulative explained variance ratio. The results showed that 17 principal components were required to retain 90% of the variance in the dataset, ensuring that the dimensionality reduction preserved the essential patterns. Figure 5 illustrates the distribution of clusters in a two-dimensional space using PCA. Each color represents a distinct cluster, highlighting how different groups of data points are distributed in the reduced space.

The distinct, well-separated groupings suggest that the clustering algorithm effectively identified patterns in the dataset. Some clusters (e.g., Cluster 2, shown in dark purple) appear to be more densely packed, indicating high intra-cluster similarity, while others, such as Cluster 12 (yellow), seem to be more dispersed, which might suggest a higher degree of variability among members of that cluster.

Additionally, overlapping regions between certain clusters suggest some shared characteristics among data points, which could indicate blended patterns in responses. The PCA transformation ensures that variance is maximized, but some information loss is expected since only two

components are displayed. Not all clusters may be distinctly visible due to dimensionality reduction, but further analysis in the full feature space can confirm their separability. Further analysis could explore higher-dimensional relationships to confirm the strength of the clusters.

3.3 Visualization Techniques

To improve the interpretability of clustering results, multiple visualization techniques were utilized. Principal Component Analysis (PCA) was applied for dimensionality reduction, projecting high-dimensional data onto a two-dimensional space to facilitate visualization of cluster separability. As previously mentioned, this approach was particularly useful for assessing whether the clusters formed distinct, non-overlapping groups.

Additionally, bar charts and pie charts were generated to illustrate the distribution of clusters across different attributes, including job roles, age groups, and mental health diagnoses. These visualizations helped to identify demographic trends and workplace patterns associated with each cluster. For instance, bar charts highlighted differences in the prevalence of certain mental health conditions among clusters. Meanwhile, a pie chart was specifically generated to represent the distribution of participants based on their living regions, providing insights into the geographical composition of the dataset. These plots, along with additional visualizations, can be found on GitHub. The repository link is provided in the reference section.

To further enhance the analysis, insights were primarily derived from the cluster analysis, where the two strongest defining parameters per cluster were extracted. Additionally, bar plots were generated to visualize the distribution of responses across all clusters for each individual survey question. This cross-checking approach helped to identify key associations between mental health diagnoses and workplace roles, ensuring that observed trends were well-supported by the underlying data.

These findings were crucial in pinpointing potential risk factors and workplace conditions that may contribute to mental health concerns, thereby guiding targeted HR interventions and mental health support programs within the organization.

Overall, the clustering methodology integrated robust preprocessing, optimal model selection, and comprehensive visualization techniques to uncover meaningful patterns within the dataset. This approach enabled a deeper understanding of how workplace factors, demographic characteristics, and mental health experiences interact, providing data-driven, actionable insights for HR initiatives aimed at improving employee well-being.

4. Results and Key Findings

The clustering analysis revealed 15 distinct clusters, each characterized by unique patterns related to mental health concerns, workplace conditions, and demographic attributes. The results provide a data-driven perspective on how different employee groups experience mental health challenges, offering valuable insights for targeted HR initiatives and workplace improvements.

4.1 Cluster Interpretation

A closer examination of the clusters uncovered several recurring themes. Fear of discrimination and trauma-related stress emerged as a key factor in multiple clusters, particularly among employees who hesitated to disclose mental health conditions due to concerns about negative consequences in the workplace. This was most prominent in Cluster 0, where fear of discrimination was the strongest parameter, and Cluster 1, which highlighted stress and uncertainty regarding disclosure. Cluster 3 and Cluster 6 also exhibited concerns about workplace bias, reinforcing the trend that employees with mental health struggles often worry about how they are perceived by colleagues and employers.

Another significant pattern involved hiring concerns and workplace fairness. Cluster 2 revealed that employees were particularly anxious about how their mental health history might affect job prospects. Similarly, Cluster 3 and Cluster 6 included themes related to bias and fairness, suggesting that employees may experience systemic challenges when seeking opportunities or career growth.

Role-specific mental health trends were also evident. Cluster 10 identified backend developers as a key group affected by developmental disorders, highlighting the need for workplace flexibility. Meanwhile, Cluster 13 was dominated by sales professionals, a role linked to high-stress levels, and Cluster 4 and Cluster 7 were characterized by individuals working in tech companies who experience personality disorders and social challenges. Clusters 7 and 10 stood out due to their restricted range of categorical responses, indicating that individuals in these groups exhibit highly specific workplace experiences or constraints. The consistency in responses for Cluster 7, which is associated with autism spectrum traits and social challenges, may indicate a uniform need for structured work environments or specialized support. Cluster 7 exhibited a striking pattern of uniform responses, for example in question 35, where all participants answered "none of them" (shown in figure 6) when asked if they had observed negative consequences for co-workers with mental health issues. Given that this cluster is associated with autism spectrum traits and social challenges, this consistency may be explained by the well-documented difficulty individuals with autism often have in perceiving and interpreting the emotional experiences of others (WHO, 2014). This aligns with broader findings in the dataset, where Cluster 7 repeatedly showed a limited range of categorical responses, suggesting a strong presence of distinct, highly structured cognitive

patterns (WHO, 2014). The uniformity across multiple questions further reinforces the validity of the clustering analysis, as it reflects the characteristic traits of individuals on the autism spectrum—such as consistent thinking patterns and a strong preference for structured environments.

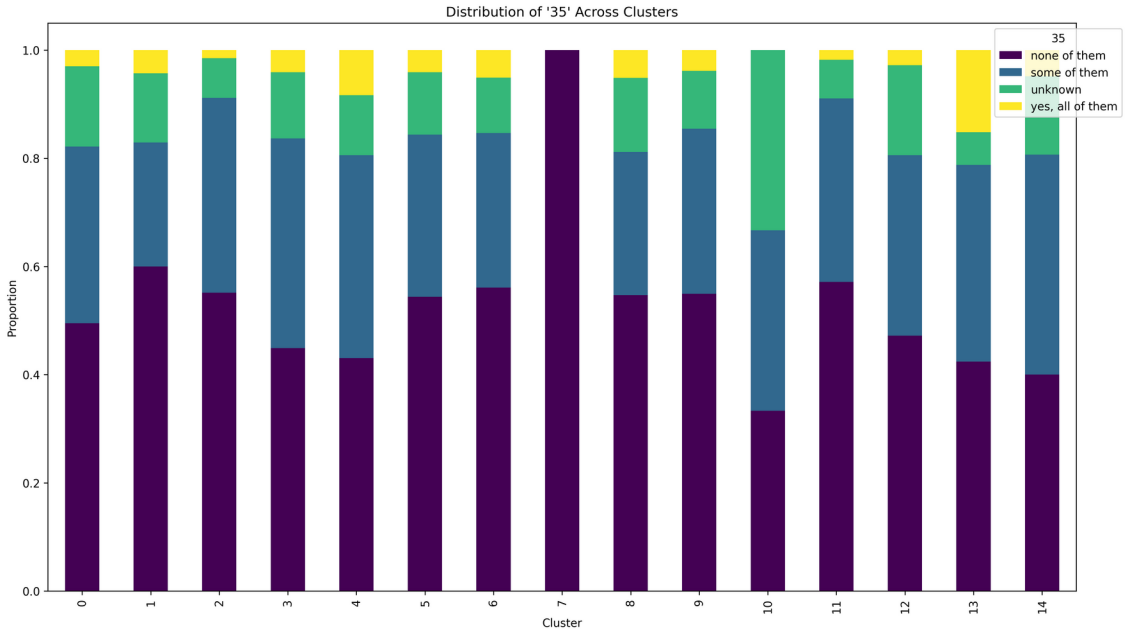

Figure 6: Answer Distribution across Clusters (Did you hear of or observe negative consequences for co-workers with mental health issues in your previous workplaces?)

Similarly, Cluster 10, dominated by backend developers, reflects a role that typically involves well-defined, standardized workflows, leading to more homogeneous workplace interactions. These findings underscore the importance of tailored interventions that address the specific needs of employees within these clusters.

An important demographic insight was the correlation between age and workplace experience. Younger employees, particularly those in early-career stages, expressed greater uncertainty about employer support, while older employees demonstrated a higher prevalence of burnout. Cluster 9, for example, was strongly characterized by personal feelings of uncertainty, while Cluster 8 revealed that many employees feel their mental health directly affects their job performance. These trends emphasize the importance of age-specific mental health policies that cater to employees at different career stages.

For a full breakdown of each cluster, including its defining parameters and associated insights, refer to the detailed cluster analysis available on GitHub. The repository link is provided in the reference section.

4.2 Notable Correlations

The analysis also revealed the strongest correlations between different variables, highlighting key patterns in the data. The most significant correlations included connections between mental health diagnoses, workplace conditions, and self-reported experiences. Notably, "topic_Attention_and_Hyperactivity_Disorders" - extracted from responses to questions about current mental health conditions (Q47), self-reported diagnoses (Q48), and suspected conditions (Q49) - showed a high correlation with the cluster variable (0.636). Additionally, it was strongly correlated with "diagnose_ADHD" (0.621), which was extracted from questions about formal medical diagnoses (Q50, Q51). This indicates a strong association between ADHD-related topics in self-reported responses and official medical diagnoses of ADHD.

Similarly, "topic_Trauma_and_Stress", which emerged from self-reported experiences, had a strong correlation with "topic_Social_Phobias_and_Trauma" (0.570), suggesting overlapping themes in responses related to trauma and social anxiety. This reinforces the idea that employees with trauma-related concerns frequently mention social phobias in their responses, indicating an intersection between these issues in workplace environments.

Additional insights emerged regarding workplace-related factors. For instance, the variable "Are you self-employed?" (Q0) showed a strong positive correlation with "Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?" (Q16) (0.769). This suggests that self-employed individuals in the dataset are more likely to have medical coverage, including mental health treatment, compared to those employed by companies. One possible explanation is that self-employed individuals, lacking employer-sponsored healthcare, may actively seek out private insurance options that include mental health benefits. Additionally, in some European countries where public healthcare systems cover mental health treatment, self-employed individuals may have access to similar or even better coverage than those relying on employer-provided plans. However, in Austria, for example, it can be difficult to access mental health treatment quickly through standard public healthcare, as there are not enough publicly funded mental health professionals available. In contrast, private mental health services are more readily accessible, providing faster treatment options for those who can afford them.

"dominant_topic_39_Relevance_to_Job_or_Business" correlated with "dominant_topic_39_Reasons_for_Disclosure_or_Non-Disclosure" (0.535), suggesting that concerns about mental health disclosure are strongly influenced by its perceived impact on professional reputation and career prospects.

Furthermore, the Python script "Correlation.py" can be extended to gather more detailed insights if needed, allowing for a deeper exploration of feature relationships and patterns within the dataset.

5. Discussion and Implications

The findings from this clustering analysis provide valuable insights for HR departments aiming to enhance workplace mental health policies. By identifying key concerns across different employee segments, companies can take targeted actions to create a more supportive and inclusive work environment.

5.1 Recommendations for HR Policy

To address the challenges identified in the analysis, several strategic interventions are recommended. Confidential mental health support services should be established to encourage employees to seek professional help without fear of stigma or workplace repercussions. Given the high prevalence of presenteeism among employees with mental health conditions - which often results in reduced productivity - offering anonymous counseling or access to external mental health professionals can foster a culture of openness and psychological safety (OECD, 2015).

Considering the role-specific stressors identified in various clusters, tailored mental health initiatives should be implemented for backend developers, sales teams, and high-pressure leadership positions. Research suggests that job-specific interventions, such as stress management training and flexible work arrangements, can effectively mitigate work-related psychological strain and improve employee well-being (OECD, 2015). Implementing specialized support groups and adjustments to work expectations can further contribute to better mental health outcomes, particularly for employees in highly demanding roles.

To combat workplace bias, particularly regarding mental health disclosure and hiring concerns, organizations should introduce mandatory bias-awareness training for hiring managers and team leads. Studies emphasize the need for balanced workplace interventions that not only reduce discrimination but also equip employers with strategies to support employees experiencing mental health challenges (OECD, 2015). Structured bias-awareness training can alleviate fears of career disadvantages due to mental health disclosure.

Additionally, flexible work arrangements should be promoted to reduce burnout and workplace stress, both of which were prevalent across multiple clusters. Expanding remote work options, introducing more flexible scheduling, and implementing mental health leave policies can provide employees with the necessary flexibility to manage their well-being while maintaining productivity. According to the OECD (2015), such policies are particularly beneficial for employees with mild to moderate mental health conditions, improving both job satisfaction and retention.

Beyond these direct interventions, a systematic evaluation of work-related stressors is crucial for long-term prevention of psychological strain in the workplace. Occupational psychology literature highlights that time pressure, lack of support from supervisors, and limited employee autonomy are

central risk factors contributing to stress and burnout. Addressing these structural issues requires improving leadership culture, increasing employee influence in decision-making, and fostering a supportive and inclusive work environment (Arbeiterkammer, 2018). When employees feel heard and empowered, workplace stressors are significantly reduced, contributing to overall mental well-being.

In summary, tackling workplace mental health challenges demands a multifaceted approach, combining enhanced mental health support services, role-specific interventions, bias-awareness training, and structural improvements to workplace policies. These measures, rooted in evidence-based strategies, can help organizations create a more inclusive, productive, and psychologically safe work environment. Further details, visualizations, and supporting data can be found in the referenced GitHub repository.

5.2 Limitations of the Study

While this study provides valuable insights, several limitations should be acknowledged. The dataset relies on self-reported survey responses, meaning that some mental health conditions may be underreported due to stigma or personal reluctance to disclose sensitive information. This introduces potential bias and may impact the accuracy of certain findings.

Additionally, the dataset primarily represents North American (+Japan) and European respondents, limiting its applicability to global workforce trends. While many insights may be generalizable, workplace mental health experiences can vary significantly based on cultural, economic, and policy differences in different regions. Future research should incorporate broader geographical representation to enhance the global relevance of these findings.

Despite these limitations, the study provides actionable insights that organizations can use to refine their mental health policies, workplace accommodations, and HR initiatives. By addressing the key concerns identified across different clusters, companies can foster a healthier, more supportive, and productive work environment for all employees.

6. Conclusion

This study presents a data-driven approach to analyzing mental health challenges in technology-focused workplaces. By applying clustering and correlation analysis, we successfully identified key mental health trends, workplace stressors, and demographic influences, enabling a more targeted approach to HR policy-making. The findings highlight the importance of addressing workplace bias, providing role-specific mental health support, and implementing flexible work policies to foster a more inclusive and supportive environment.

Moving forward, ongoing assessments using updated datasets will be crucial to monitor trends, evaluate the effectiveness of interventions, and adapt mental health strategies to evolving workplace needs. Continuous feedback enables organizations to enhance employee well-being, reduce stigma, and foster psychological safety.

7. References/Bibliography

Kaggle Dataset: *Mental Health in Tech 2016* (Dataset Source)

GitHub Repository: Unsupervised ML for Mental Health Analysis

Relevant Research Papers on Workplace Mental Health Strategies:

Dr. Glaser Jürgen, Mag. Molnar Martina (Arbeiterkammer Wien 2018). Psychische Belastung und Stress in der Arbeit: Ursache, Folgen, Lösungen [Mental strain and stress at work: causes, consequences, solutions] [pdf]. Retrieved from: https://wien.arbeiterkammer.at/service/broschueren/Arbeitnehmerschutz/broschueren/Psychische_Belastung_und_Stress_in_der_Arbeit_2018.pdf

Kaloyanova, E.(2024). How to Combine PCA and K-means Clustering in Python?. Retrieved from: https://365datascience.com/tutorials/python-tutorials/pca-k-means/

OECD (2015), Mental Health and Work: Austria, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264228047-en [pdf]. Retrieved from: https://www.oecd.org/en/publications/mental-health-and-work-austria_9789264228047-en

scikit-learn developers (BSD License). Selecting the number of clusters with silhouette analysis on KMeans clustering. Retrieved from: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

WHO, (2014). Comprehensive and coordinated efforts for the managment of autism spectrum disorders. Sixty-Seventh World Health Assembly, A67/17 [pdf]. Retrieved from: https://apps.who.int/gb/ebwha/pdf_files/WHA67/A67_17-en.pdf