Mental Health Clustering based on Tech Jobs

Case Study for Course DLBDSMLUSL01

Machine Learning – Unsupervised Learning and Feature Engineering

20.12.2024

Mohsen Nasiri

Matriculation Number: 32206338

Tutor's Name: Christian Müller-Kett

Table of Contents

1. Introduction 1

1.1. Context 1

1.2. Objectives 1

2. Methodology 1

2.1. Tools and Libraries 2

2.2. Simplifying 2

2.3. Removing Outliers 3

2.4. Remapping 3

2.6. Handling Missing Values 5

2.7. Dropping Columns 6

2.8. Encoding the Dataset 6

2.10. Data Scaling 7

2.11. Dimensionality Reduction 7

2.12. Clustering 7

3. Results and Conclusion 8

3.1. Clusters 8

3.2. Conclusion 10

References 11

Appendices 12

1. Introduction

This study analyzes survey data from employees in technology roles to identify clusters based on mental health responses. The goal is to provide actionable insights for Human Resources (HR) to design targeted mental health initiatives, fostering a healthier workplace environment.

1.1. Context

The HR department of a tech company aims to address mental health challenges among employees through focused interventions. To support this effort, survey data capturing various mental health factors, workplace dynamics, and individual experiences were analyzed. The dataset, however, presents challenges such as:

- High dimensionality
- Complex and inconsistent categorical responses
- Presence of missing values and unstandardized inputs

These issues make direct interpretation difficult, necessitating preprocessing and clustering techniques to uncover meaningful patterns.

1.2. Objectives

The primary objectives of this study are:

1. Simplify the dataset while retaining key information to enable clear analysis.
2. Cluster survey participants based on their mental health-related responses and workplace factors.
3. Extract cluster-specific insights to guide HR in developing targeted interventions.
4. Visualize the results to highlight distinct groups and actionable patterns.

By achieving these objectives, the study aims to provide HR with concrete leverage points for designing effective mental health programs tailored to employees' needs.

2. Methodology

The methodology outlines the steps taken to preprocess the dataset, reduce dimensionality, and perform clustering.

2.1. Tools and Libraries

The Pandas library was utilized for reading, cleaning, and manipulating the dataset and loading the CSV file into a data frame. For further handling of arrays and other operations related to the data frame, NumPy was used.

For visualizing and plotting data and trends, Matplotlib was implemented alongside Seaborn, which is built on Matplotlib to enhance visualizations.

During the data preprocessing stage, several tools from Scikit-learn (Sklearn) were employed. Sklearn's Impute module managed missing data by filling in missing values using strategies like the mean, median, or most frequent value. For encoding and scaling data, the Sklearn Preprocessing module provided:

- Multi Label Binarizer: Encodes categorical variables with multiple labels into binary indicators.
- Label Encoder: Encodes categorical labels into numeric values.
- Standard Scaler: Standardizes features by removing the mean and scaling them to unit variance.

For dimensionality reduction, UMAP was used. However, in earlier versions of the code, PCA from Sklearn's Decomposition module was implemented, which was not optimal.

To determine the optimal number of clusters, the KelbowVisualizer from Yellowbrick was utilized, to use the elbow method.

For clustering, the KMeans algorithm was applied. To evaluate the clustering performance, metrics such as the Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score were used. These metrics were implemented using Scikit-learn (Sklearn).

2.2. Simplifying

The original column names were complex and unclear. Each column was renamed to a more concise and descriptive format that aligns with the data it represents. In addition, the text data was converted to lowercase to ensure consistency, and any extra spaces at the ends of strings were removed to avoid mismatches.

- **Name Normalization:** "Are you self-employed?" → “is_self_employed”
- **Conversion to lowercase:** "United Kingdom" → "united kingdom"
- **Whitespace:** "yes " → "yes"

2.3. Removing Outliers

The dataset included outliers specifically the "age" column, which contained values outside the realistic working age range. To address this, the entries below 18 and above 80 were removed.

2.4. Remapping

Remapping was essential to create consistent patterns across similar responses and standardize the dataset particularly where multiple terms expressed the same idea or where values needed to be transformed. The following actions were taken in this phase:

2.4.1. Standardizing Ambiguous and Inapplicable Responses:

All responses such as "maybe," "I don't know," and "unsure" were replaced with "uncertain". Additionally, entries like "not applicable to me" were replaced with "inapplicable", to reduce redundancy.

2.4.2. Simplifying Responses for Binary Consistency:

Responses containing variations of "yes" (e.g., "yes, definitely") were simplified to "yes" to maintain binary consistency.

2.4.3. Remapping Positive and Negative Responses:

The following changes were made for categorical columns, ensuring consistent phrasing while maintaining the meaning.

- "yes, I think they would" → "yes, I think so"
- "yes, they do" → "yes, I experienced it"
- "no, they do not" → "no, I didn't experience it"
- "no, I don’t think they would" → "no, I don’t think so"

2.4.4. Unifying Multilevel Categorical Responses:

For categorical columns with multilevel responses, the following was applied to enable ordinal encoding while preserving the hierarchical structure:

- "all", "yes", and "always" → "all"
- "some" and "sometimes" → "some"
- "none", "no", and "never" → "none"

2.4.5. Gender Data Standardization:

Variations of "male" (e.g., "m", "mail") were replaced with "man". Similarly, variations of "female" (e.g., "f", "woman") were replaced with "woman". Entries that did not fit into these categories were labeled as "diverse".

2.4.6. Handling Missing Values of the States Columns:

Missing or null values in the United States columns were replaced with "not specified" to account for individuals outside the United States.

2.4.7. Grouping Age Data:

The "age" column was grouped into ranges (e.g., 18–30, 30–40) for better interpretability and analysis.

2.4.8. Grouping Reason Responses:

Responses about willingness to discuss mental or physical health were grouped into broader categories to reduce variability and create consistency. The groups were created by responses containing the following terms:

- Terms related to stigma or acceptance were grouped as "stigma"
- Privacy-related terms were grouped as "wont_discuss"
- Open and transparent discussions were grouped as "honesty"
- Responses that didn’t contain any of the above terms were grouped as "other"

2.5. Creating New Columns

For a clearer and more consistent analysis, it was necessary to group and categories responses and create new columns for them, since each person could relate to multiple groups and these groups need to be treated separately for analysis. This step also reduced variables and enhanced the results. However, at this point the responses of each column were separated and transformed into a list. The new columns will be created during encoding.

Conditions

Columns containing varied textual descriptions of mental health conditions were grouped into general categories:

Categories Defined:

- **Anxiety and Mood Disorders:** Includes generalized anxiety disorder, depression, and seasonal affective disorder.
- **Neurodevelopmental and Behavioral Disorders:** Includes ADHD, Autism spectrum disorders, and related impairments.

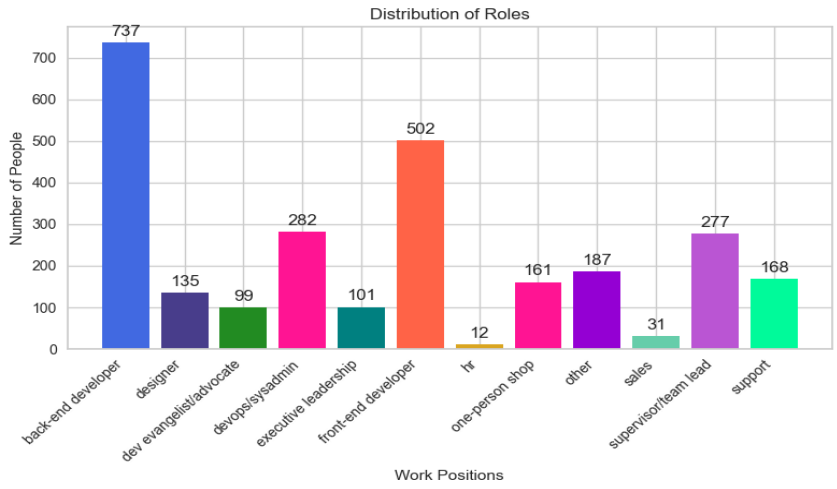
- **Trauma and Stress-Related Disorders:** Includes PTSD, burnout, and dissociative disorders.
- **Substance and Addiction Disorders:** Includes substance use and sexual addiction disorders.
- **Personality and Psychotic Disorders:** Includes borderline personality disorder and schizophrenia.
- **Sexual and Gender Identity Disorders:** Includes gender dysphoria and related conditions.
- **Neurological Disorders:** Includes traumatic brain injuries and tinnitus.
- **Crisis and Mental Health Conditions:** Includes suicidal ideation.
- **Other:** Includes sleep and eating disorders.

Example Transformation:

- Response: "tinnitus | anxiety disorder (generalized, social, phobia, etc.)"
- Transformed: "neurological_disorders" and "anxiety_and_mood_disorders"

Work Positions

The “work_position” column was separated into multiple individual roles. This transformation ensured each role was treated separately for analysis. The separation resulted in the following roles forming:



2.6. Handling Missing Values

Missing values significantly affect the performance of the algorithm. The following strategies were employed:

- **Numerical Columns:** Missing values were replaced with the median of their respective column. This approach was chosen to minimize the influence of outliers.

- **Categorical Columns:** Missing values in categorical features were imputed with the most frequent category. This ensured that the data distribution remained consistent.

2.7. Dropping Columns

To address issues related to high cardinality, two columns were removed due to high cardinality which complicated the analysis affected the clusters negatively. The columns **“country_of_residence”** and **“country_of_work”** were more than 95 percent similar. Similarly, **“us_state_or_territory_of_residence”** and **“us_state_or_territory_of_work”** were also more than 95 percent similar.

2.8. Encoding the Dataset

Encoding was necessary to transform raw categorical data into a structured numerical format suitable for the algorithms. Different strategies were applied based on the type of data and its complexity. Below is the explanation:

2.8.1. Text Range Conversion

The column that represented percentages was mapped to its numerical equivalents, for example "1-25%" was encoded as 0.25.

2.8.2. Binary Encoding

Columns with distinct categories like "yes", "no", "maybe", and "uncertain" binary encoding were applied. Responses were mapped to numerical values. However, the encoding process was tailored for each column based on its set of responses. For example, columns with 'yes' and 'no' were encoded separately from columns with 'yes', 'no', and 'maybe', or 'yes', 'no', 'maybe', and 'uncertain'. This method maintains data integrity, avoids misinterpretation, and enhances the model by preserving the distinctiveness of the responses.

2.8.3. Ordinal Encoding

For columns with ordered categories (e.g., "none", "some", "all"), ordinal encoding was applied to assign numerical values based on their order while differentiating special cases like 'inapplicable' and 'uncertain' to maintain order and integrity. For example, values were assigned as "none" = 0, "some" = 1, "all" = 2, "inapplicable" = -2, "uncertain" = -1.

2.8.4. Multi-label Binarization

The columns with multiple categories in a single response, such as job roles or the conditions columns, unique categories were transformed into a binary column, where 1 indicated the presence of the category, and 0 for absence. This method increased the columns of the dataset; however, it was necessary to distinguish the importance of each element in the responses.

2.8.5. Frequency Encoding

When dealing with columns that had many unique values and no clear ordering, like the reasons or countries columns, frequency encoding was used which replaces each category in a column with its frequency count within the dataset, preserving the significance of the categories without expanding dimensionality.

2.10. Data Scaling

After encoding every column, the encoded dataset was scaled using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1. This equalized their influence, preventing larger values from dominating the analysis and making the data compatible with the KMeans algorithm.

2.11. Dimensionality Reduction

UMAP (Uniform Manifold Approximation and Projection) was used for dimensionality reduction, transforming the data into two principal components which provided slightly better metric scores compared to three components.

UMAP was selected over PCA (Principal Component Analysis) because PCA emphasizes maximizing variance, while UMAP retains the overall shape of the data and separates points more effectively because most columns were categorical. Therefore, UMAP offered superior metric scores, and better separation of data points compared to PCA.

2.12. Clustering

After reducing the data, KMeans clustering was used to group participants based on the data. This final part consisted of the following steps:

2.12.1. Determining Optimal Clusters

The "elbow method" was used with a visualizer to identify the ideal number of clusters, also known as ‘k’.

2.12.2. Applying KMeans

KMeans partitioned the dataset into k clusters by minimizing the variance within each group.

2.12.3. Evaluating Clusters

The quality of the clusters was assessed using several metrics:

- **Silhouette Score (0.48):** Measures how well the clusters are separated.
- **Calinski Harabasz (1934.77):** Evaluates the compactness and separation of clusters.

- **Davies Bouldin (0.75):** Measures similarities within a group and differences between groups.

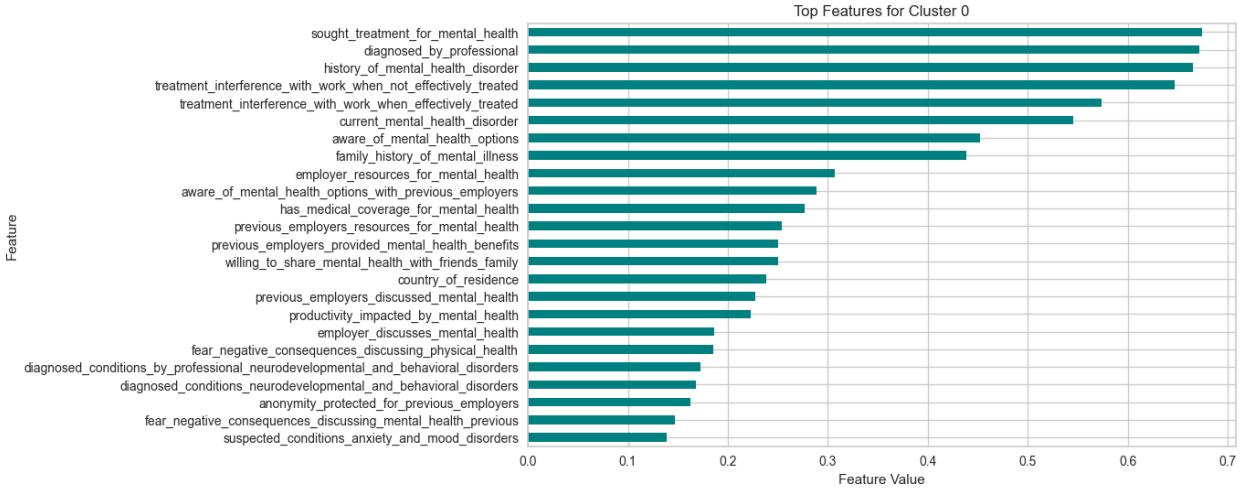
3. Results and Conclusion

The KMeans clustering algorithm identified four distinct clusters. The clusters and keys that define them are explained below:

3.1. Clusters

Cluster 0: Employees in this cluster focus a lot on getting mental health treatment and diagnoses.

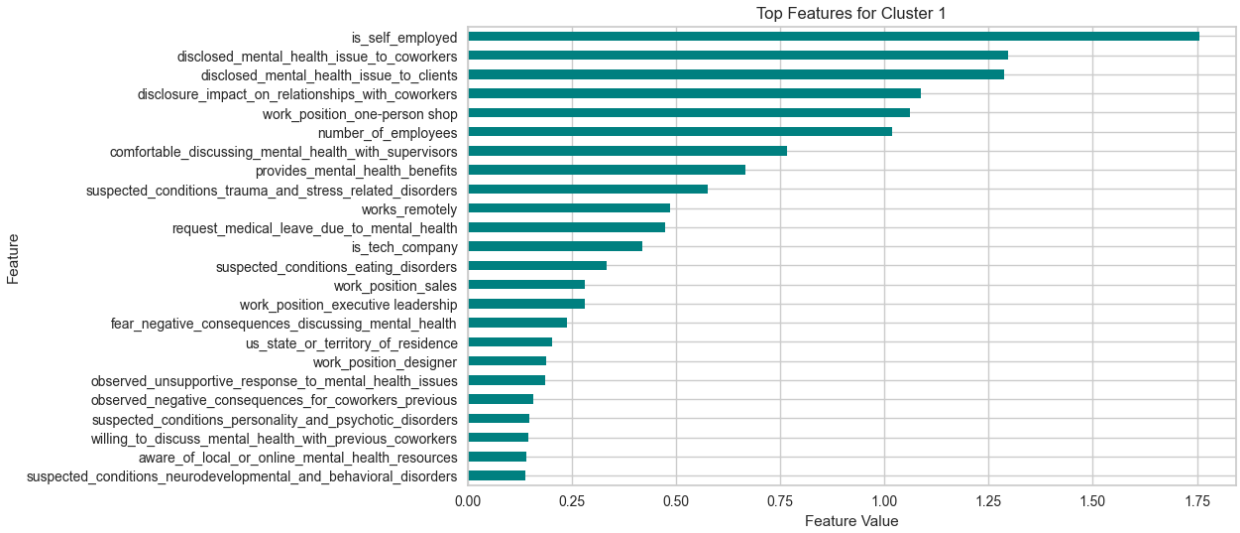
1. **Seeking Treatment:** Sought treatments for mental health issues.
2. **Diagnosis History:** Had a history of mental health disorders and professional diagnoses.
3. **Work Interference:** Treatments significantly affected their work productivity, especially when untreated.
4. **Aware of Resources:** Aware of mental health resources provided by employer.
5. **Family History:** Their family has a history of mental health issues.



Cluster 1: Employees in this cluster are self-employed, work remotely or are part of smaller teams and they are open to talking about their conditions.

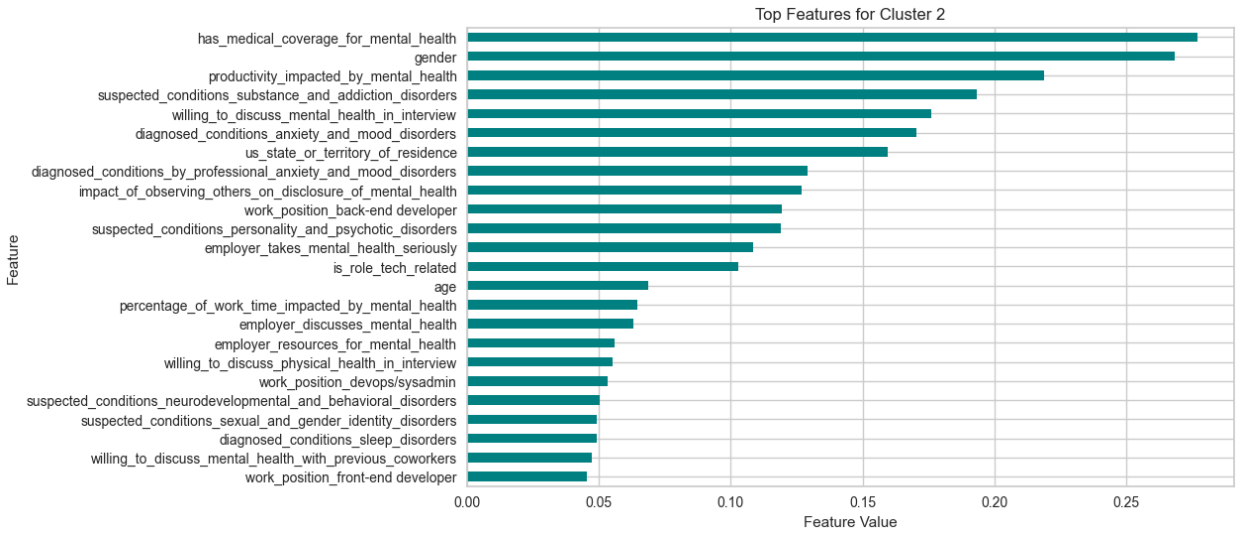
1. **Self-Employed:** is self-employed.
2. **Openness to Discussion:** Were comfortable discussing mental health with coworkers, supervisors, and clients.
3. **Impact on Relationships:** Sharing mental health information affected their work relationships.

4. **Non-IT Roles:** Working as executive leadership, designer, one person shop or sales.
5. **Remote Work:** Works remotely.



Cluster 2: Employees in this cluster work in IT roles and rely on employer mental health resources.

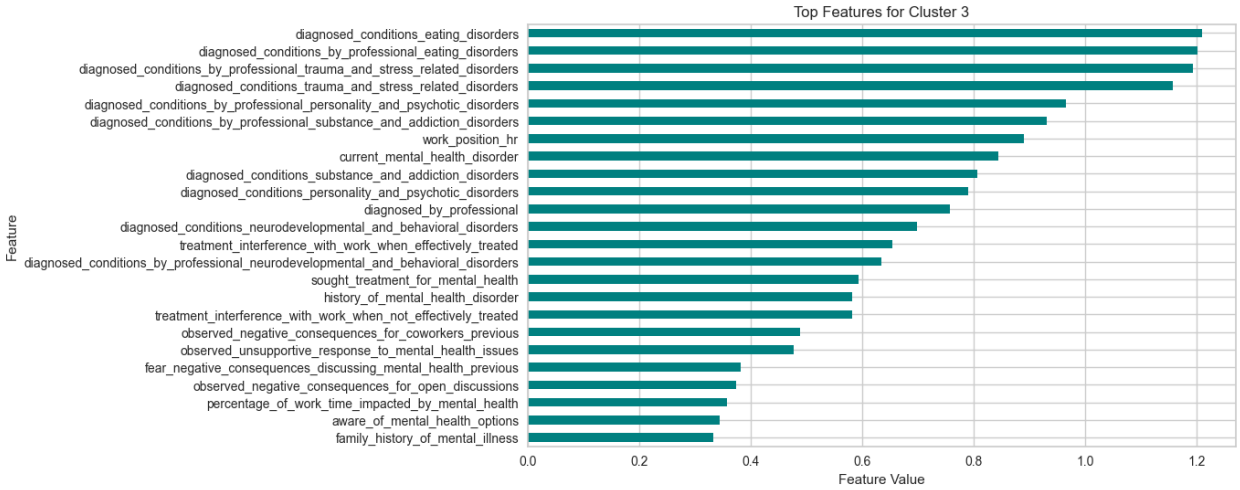
1. **Medical Coverage:** Access to mental health care was important.
2. **Employer Involvement:** Very involved in mental health discussions and resources.
3. **Workplace Culture:** Their workplace took mental health seriously.
4. **IT Roles:** Working in back-end, front-end, or DevOps.
5. **Productivity Impact:** Mental health had a lower impact on their productivity.



Cluster 3: Employees in this cluster have many serious and complex mental health issues.

1. **Serious Diagnoses:** eating disorders, trauma, and stress-related and other disorders.

2. **Professional Treatment:** sought help from mental health professionals.
3. **Workplace Interference:** Mental health issues disrupted their work.
4. **Negative Consequences:** observed negative workplace responses to their mental health issues.
5. **Currently Undiagnosed:** Has a history of illnesses and currently has untreated conditions.



3.2. Conclusion

With the help of unsupervised learning, we identified four key clusters each representing unique mental health that can help create a healthier workplace environment. We identified the following groups:

1. Employees that actively care and seek treatment.
2. Self-employed, remote, and non-IT workers who are open about their conditions.
3. IT workers that rely on work-provided mental health resources.
4. Employees with severe mental health conditions requiring specialized care.

By understanding the clusters, the HR department can implement strategies to improve the health and productivity of their employees. All the possible graphs of the clusters in relation to all the columns are included in the full code found on the GitHub link below:

- https://github.com/NasiriMohsen/Unsupervised-Learning-Mental-Health-Dataset

References

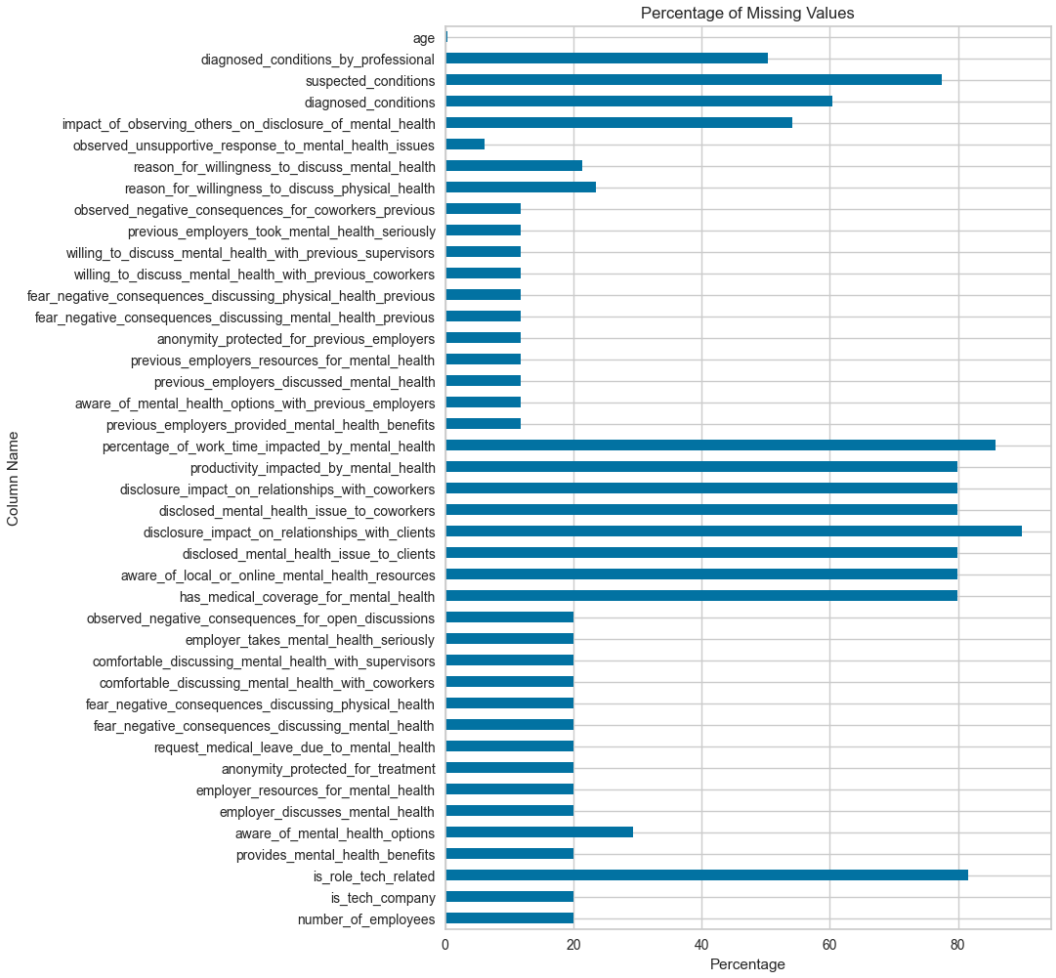
Venkatesh, D. (2016). *Mental health in tech survey 2016* [Data set]. Open Sourcing Mental Illness. https://www.kaggle.com/osmi/mental-health-in-tech-2016

Appendices

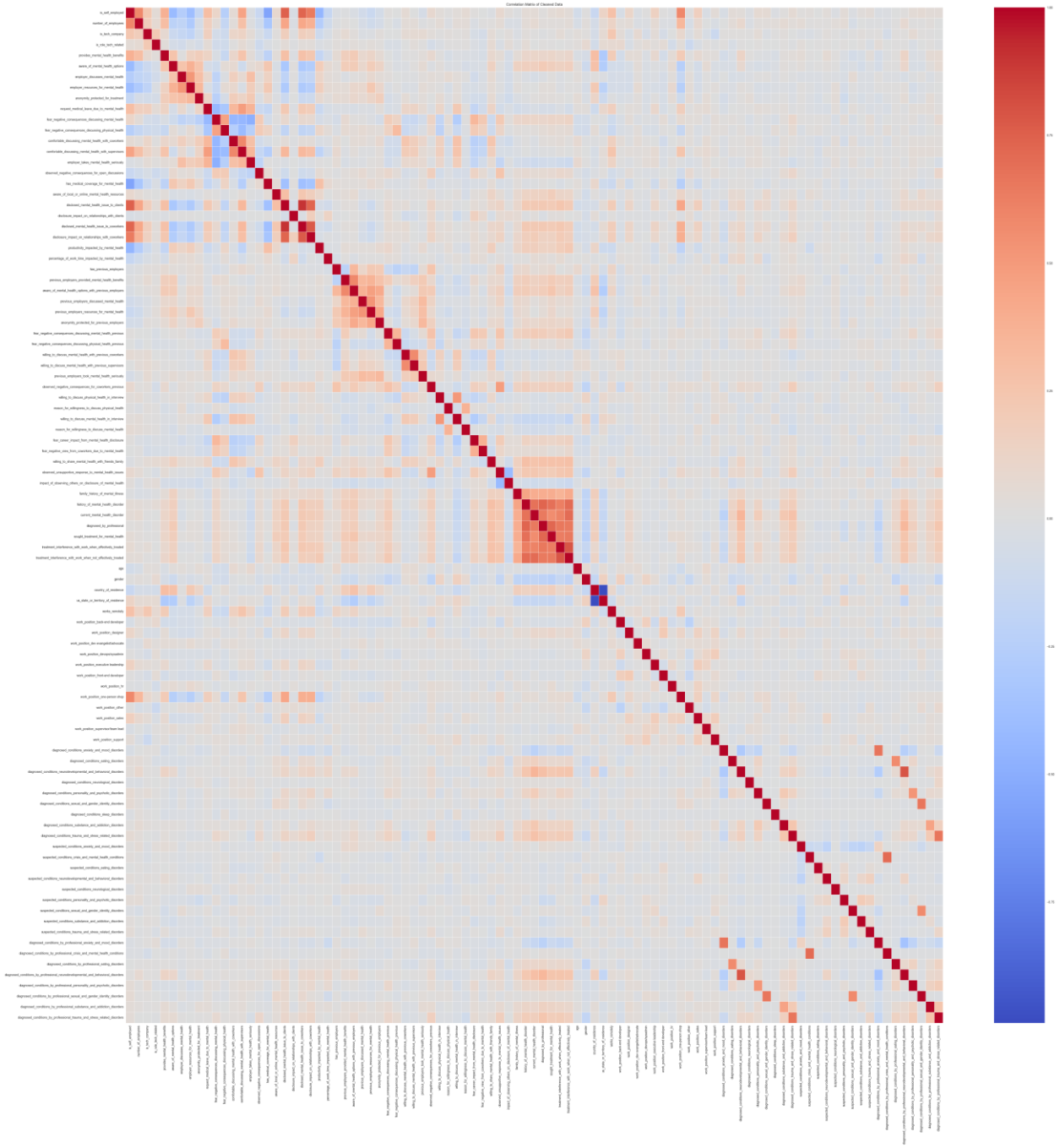
Tools and Libraries:

```
import pandas as pd
import numpy as np
from math import floor

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MultiLabelBinarizer
from sklearn.preprocessing import LabelEncoder, StandardScaler

from umap import UMAP
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, calinski_harabasz_score, davies_bouldin_score
```

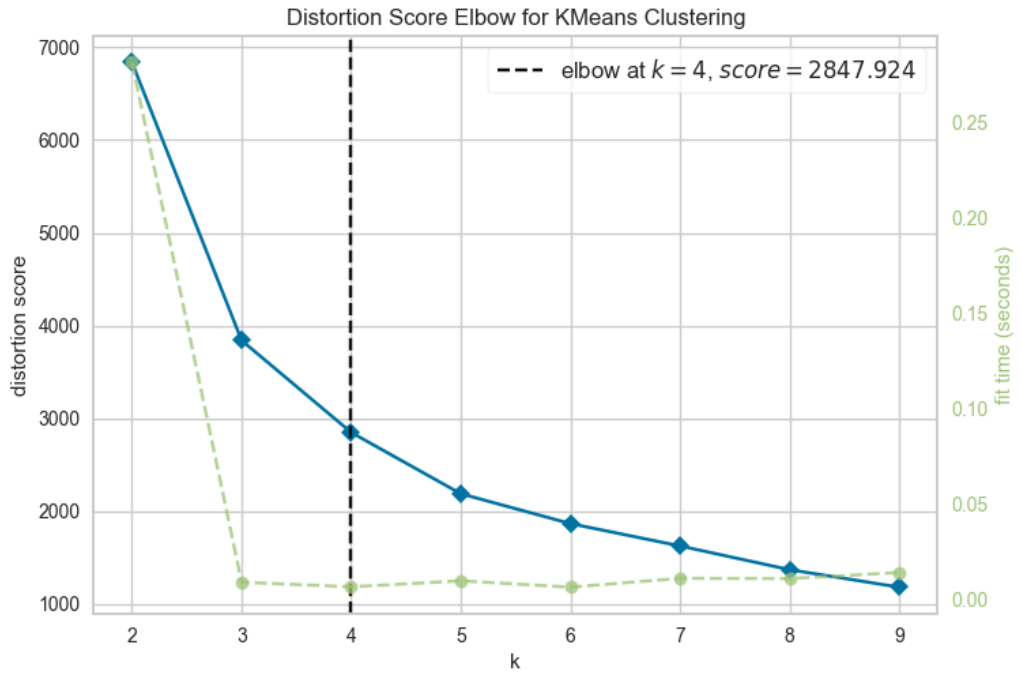
Columns that had Missing Values:



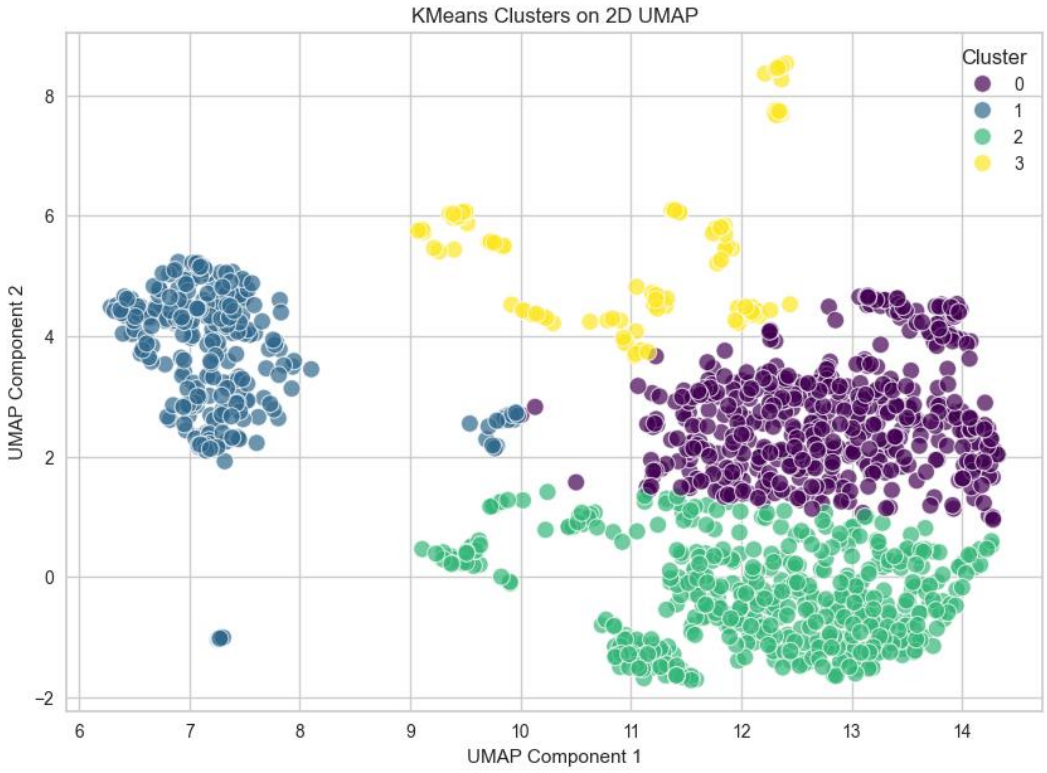
Coloration Heatmap:



Determining Optimal Clusters:



Evaluating Clusters:



Cluster Features Heatmap: