Data Visualization: Mental Health in Technical Workplace

Mohajit Banik

Data Analysis with Python

National Institute of Industrial Training

Ishapore, West Bengal

mohajit25@gmail.comArpita Goswami

Data Analysis with Python

National Institute of Industrial Training

Barrackpore, West Bengal

goswamiarpita872@gmail.com

*Abstract*—This study investigates the mental health trends among tech professionals. The primary objective is to identify significant patterns and correlations that can inform mental health support strategies within the tech industry. The dataset includes demographic information, job roles, work environments, mental health conditions, and access to mental health resources.

Advanced data processing techniques, including data cleaning, exploratory data analysis (EDA), and machine learning algorithms, are employed to ensure the accuracy and depth of the findings. Preliminary results indicate that factors such as job role, work environment, and access to mental health resources significantly influence mental health outcomes. For instance, individuals in high-pressure roles or those lacking access to mental health resources exhibit higher levels of anxiety and depression.

The study underscores the importance of targeted mental health interventions and policies tailored to the unique challenges of the tech industry. By providing a detailed analysis of the most recent data, this research offers valuable insights into the current state of mental health among tech professionals. The findings highlight the need for comprehensive mental health support systems and proactive measures to address the mental health challenges faced by tech employees.

This research contributes to the broader understanding of mental health in the workplace and supports the development of effective strategies to enhance the well-being of tech professionals. The insights gained from this study will be instrumental for employers, policymakers, and mental health professionals in creating a healthier and more supportive work environment in the tech industry.

Introduction

The rapid evolution of the tech industry has brought about significant advancements and opportunities, but it has also introduced unique challenges that impact the mental health of its workforce. This dataset provides a comprehensive overview of mental health trends among tech professionals. This period is particularly noteworthy due to the post-pandemic adjustments and the increasing demands for remote and hybrid work environments.

The dataset encompasses a wide range of variables, including demographic information, job roles, work environments, mental health conditions, and access to mental health resources. By analyzing these variables, we aim to uncover patterns and correlations that can inform better mental health support strategies within the tech industry. The analysis leverages advanced data processing techniques, including data cleaning, exploratory data analysis (EDA), and machine learning algorithms, to ensure robust and insightful findings.

The tech industry is known for its high-pressure environment, long working hours, and constant need for innovation, all of which can contribute to mental health issues such as anxiety, depression, and burnout. Understanding the prevalence and factors associated with these conditions is crucial for developing effective interventions and fostering a healthier work culture. This study not only highlights the current state of mental health in the tech industry but also provides actionable insights for employers, policymakers, and mental health professionals.

By focusing on the most recent data, this analysis captures the contemporary challenges faced by tech professionals and offers a timely perspective on mental health trends. The findings from this study will contribute to the ongoing discourse on mental health in the workplace and support the development of targeted initiatives to improve the well-being of tech employees.

**Abbreviations & Modules:**

**Abbreviations:**

* **q=**  A variable that is used to denote the CSV Dataset for mental health in the Technical workspace.
* **rgen=** Replace Gender, a function to replace misspelled and unwanted strings to correct genders in the *Gender* column.
* **ager=** Replace Age, a function to negative and over 100 integer values to “0” in the *Age* column.

## **Modules:**

* **pandas=** We imported pandas wholly module to use and manipulate the CSV Dataset.
* **seaborn**= We imported the seaborn wholly module to plot various types of graphs.
* **matplotlib=**We imported matplotlib for designing the graphs.

## **Data Cleaning**

Data cleaning is a pivotal step in data preprocessing, essential for transforming raw data into a polished, analysis-ready dataset. In this project, we meticulously re-entered integer values with their correct legends, replacing irrelevant entries with precise, meaningful data. This diligent process ensured that each element was accurately represented, enhancing the relevance and accuracy of our dataset. By aligning data with its correct legends, we’ve not only refined the dataset but also optimized it for insightful analysis and impactful visualization. This attention to detail lays a strong foundation for robust, data-driven decision-making and ensures that our results are both reliable and actionable.

**The cleaning: -**

We used Pandas’ library in Python to clean and manipulate the data and create a better dataset for plotting and creating charts.

def ager(a):

    if a<0 or a>100:

        return 0

    elif isnull(a):

        return 0

    else:

        return a

q=q[q["Age"] != 0]

The **ager** function is created to process values from the 'Age' column by ensuring that all entries fall within the valid range of 0 to 100. Values outside this range are replaced with “0.” After applying this function, only rows where the 'Age' column does not contain “0” are retained. This approach effectively cleans the dataset by removing irrelevant and out-of-range entries, resulting in a more accurate and reliable dataset for analysis.

def rgen(a):

    if a in ["male", "Male", "something kinda male?", "M", "m", "Male ", "Man", "Male-ish", "Make", "Mail", "Mal", "maile", "male leaning androgynous" "Malr", "msle"]:

        return "male"

    elif a in ["Female", "female", "F", "f", "Woman", "femail", "Femake", "Female ", "Malr", "woman"]:

        return "female"

    elif a in ["Trans woman", "Trans-female", "Female (trans)"]:

        return "transgender"

    elif a in ["Cis Man", "Cis Male", "cis male", "Male (CIS)", "Cis Female", "cis-female/femme", "Female (cis)"]:

        return "cisgender"

    else:

        return "other"

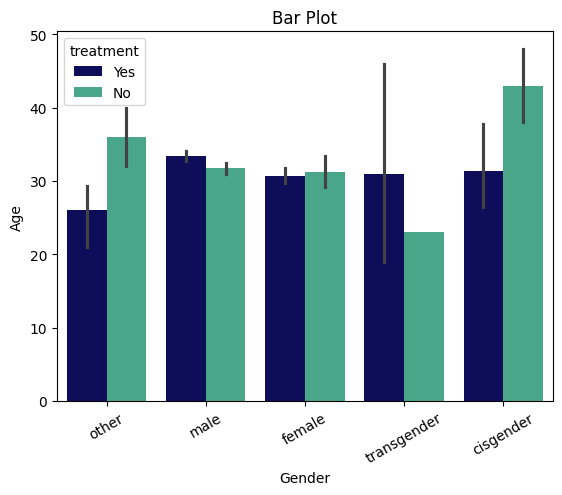
The **rgen** function is tailored to standardize values in the ‘Gender’ column for enhanced clarity and consistency. It replaces any male-related terms with “Male” and female-related terms with “Female.” Additionally, it recognizes and categorizes entries as “Cisgender” or “Transgender” as applicable. All other non-standard gender identities are labeled as “Other.” This transformation ensures a more insightful and streamlined dataset by consolidating gender information into clear, uniform categories, facilitating more accurate analysis and interpretation.

q=q.dropna()

The robust **dropna()** function is employed to eliminate residual NULL values across all columns, particularly in “self\_employed” and “work\_interfere,” where such values were prominent. This critical step ensures that the dataset is both meaningful and clean, paving the way for more accurate analysis and effective visualization. By removing these NULL entries, we achieve a dataset that is ready for insightful exploration and robust data-driven decisions.

## **Data Visualizations**

Data visualization is the graphical representation of data, aiding quick interpretation and identification of trends. It encompasses various techniques like bar charts, histograms, and scatter plots to enhance communication and facilitate informed decision-making. Ultimately, it makes complex data accessible, understandable, and actionable, benefiting both technical and non-technical users.

* **Age and Gender Distribution for Treatment History**

This bar plot visualizes the age distribution across various gender identities for individuals who have either received (Yes) or not received (No) a specific treatment. The x-axis categorizes gender into groups such as ‘other’, ‘male’, ‘female’, ‘transgender’, and ‘cisgender’, while the y-axis represents age, ranging from 0 to 50. The bars are color-coded to distinguish between those who have received treatment (dark blue) and those who have not (teal).

**Key observations-**

1. **Overview:**
   * The bar plot compares the average age of individuals across different gender categories, segmented by treatment status (Yes or No).
2. **Axes:**

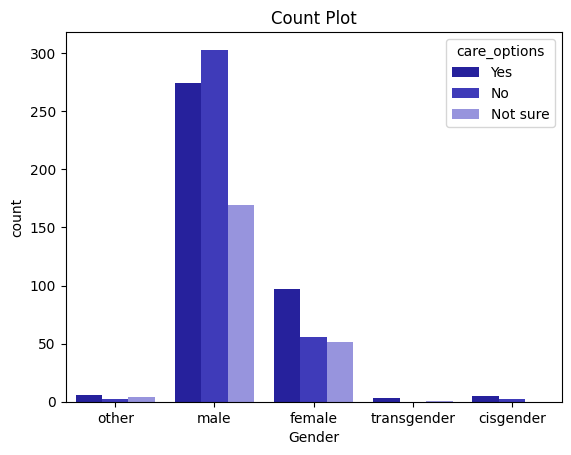


* + X-axis: Represents different gender categories, including "other," "male," "female," "transgender," and "cisgender."
  + Y-axis: Represents the average age, likely ranging from a minimum of 0 to a maximum of around 50.

1. **Gender Categories:**
   * The plot includes five gender categories: other, male, female, transgender, and cisgender.
2. **Treatment Status:**
   * Yes (Treatment): Represented by the darker bars.
   * No (No Treatment): Represented by the lighter bars.
3. **Age Distribution:**
   * Other: This group has the highest average age, indicating potential unique treatment needs.
   * Male and Female: Both categories show lower average ages compared to "other," with treatment individuals slightly older.
   * Transgender: This group has a moderate average age, with treated individuals being older.
   * Cisgender: Similar to females, with treated individuals showing a higher average age.
4. **Variability:**
   * Error bars indicate variability in age within each category. Larger error bars in some categories suggest a wider age range among individuals.
5. **Data Interpretation:**
   * The plot shows that individuals who received treatment generally have a higher average age compared to those who did not, across most gender categories.
   * The "cisgender" category shows the highest average age for individuals who received treatment, while the "transgender" category has a notable difference between treatment statuses.
6. **Trends:**
   * The "other" gender category has the highest average age among those who did not receive treatment, suggesting that this group may have unique characteristics or circumstances.
   * The error bars indicate variability in the data, showing that there is some fluctuation in average ages within each category.
7. **Clinical Relevance:**
   * Understanding the relationship between gender, treatment status, and age can provide insights into the demographics of individuals seeking treatment.
   * This information could be valuable for healthcare providers in tailoring interventions and support systems for different gender groups.
8. **Implications:**
   * The data may indicate differing treatment needs across gender categories, particularly for the "other" group.
   * Understanding age distribution can help tailor healthcare services and interventions.

**Conclusion:**

The graph indicates that there is a variation in age among those who have received treatment across different gender identities. Some gender groups show a higher average age for those receiving treatment compared to those who do not. This suggests that age and gender may influence the likelihood of receiving treatment, highlighting the need for targeted mental health interventions that consider both age and gender factors.

* 

Bottom of Form

* **Gender-Based Preferences for Care Options**

This count plot visualizes the distribution of preferences for care options across different gender identities. The x-axis categorizes gender into groups such as ‘other’, ‘male’, ‘female’, ‘transgender’, and ‘cisgender’. The y-axis represents the count of individuals in each category. The bars are color-coded to indicate responses: ‘Yes’ (dark blue), ‘No’ (blue), and ‘Not sure’ (light blue).

**Key observations:**

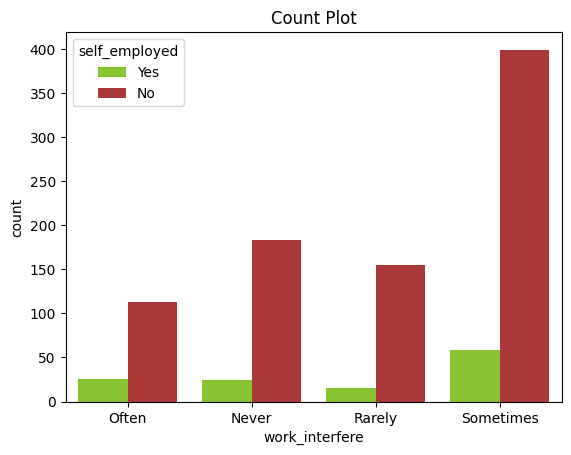
1. **Axes:**
   * X-axis: Represents different gender categories, including "other," "male," "female," "transgender," and "cisgender."
   * Y-axis: Represents the count of individuals, indicating how many responses fall into each category.
2. **Bars:**
   * The bars are color-coded to represent responses to care options:

* Dark Blue Bars: Indicate individuals who answered "Yes" to care options.
* Medium Blue Bars: Indicate individuals who answered "No."
* Light Blue Bars: Indicate individuals who answered "Not sure."
  + Each gender category has stacked bars representing the different responses.

1. **Data Interpretation:**
   * The "male" category has the highest count of responses, particularly for the "Yes" option, suggesting that a significant number of males are seeking or have access to care options.
   * The "female" category shows a lower count compared to males, with fewer individuals indicating "Yes" to care options.
   * The "other" category has very few responses, indicating that this group may be underrepresented or less likely to respond.
2. **Trends:**
   * The "transgender" and "cisgender" categories have notably lower counts across all responses, suggesting that these groups may have different experiences or access to care options.
   * The "Not sure" responses are minimal across all categories, indicating a general clarity among respondents regarding their care options.
3. **Clinical Relevance:**
   * Understanding the distribution of care options by gender can provide insights into the accessibility and utilization of healthcare services among different demographic groups.
   * This information could help healthcare providers identify areas for improvement in service delivery and outreach.

**Conclusion:**

The graph reveals significant variations in care option preferences across different gender identities. The ‘male’ category shows the highest count for ‘Yes’, indicating a strong preference for care options. Other gender categories exhibit diverse responses, with notable counts for ‘No’ and ‘Not sure’. This suggests that gender may influence preferences for care options, highlighting the need for tailored approaches to address the unique needs and preferences of different gender groups.

* **Impact of Work Interference on Self-Employed vs. Non-Self-Employed Individuals**

This count plot visualizes the frequency of work interference among self-employed and non-self-employed individuals. The x-axis categorizes the frequency of work interference into four groups: Often, Never, Rarely, and Sometimes. The y-axis represents the count of individuals in each category. The bars are color-coded to indicate self-employment status: green for ‘Yes’ (self-employed) and red for ‘No’ (non-self-employed).

**Key observations:**

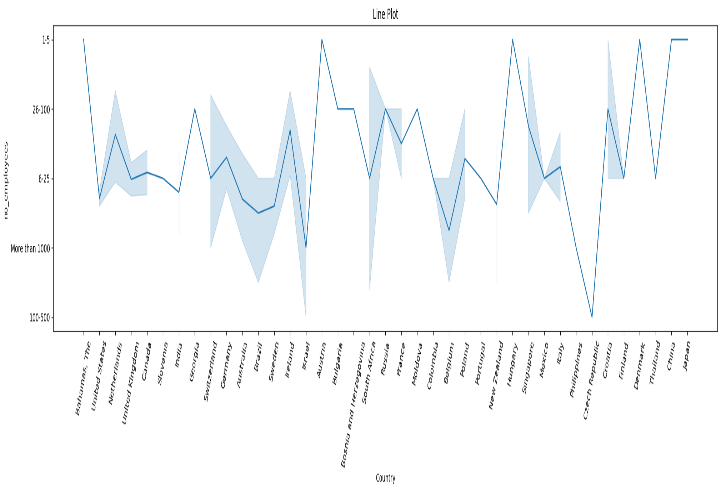
1. **Axes:**
   * X-axis: Represents the frequency of work interference, categorized into "Often," "Never," "Rarely," and "Sometimes."
   * Y-axis: Represents the count of individuals, indicating how many responses fall into each category of work interference.
2. **Bars:**
   * The bars are color-coded to represent self-employment status:

* Green Bars: Indicate individuals who are self-employed (Yes).
* Red Bars: Indicate individuals who are not self-employed (No).
  + Each work interference category has two bars: one for each self-employment status.

1. **Data Interpretation:**
   * The "Sometimes" category shows the highest count of individuals, particularly among those who are not self-employed, indicating that work interference is a significant issue for this group.
   * The "Never" category has the lowest counts for both self-employed and non-self-employed individuals, suggesting that very few people report no work interference.
2. **Trends:**
   * There is a noticeable trend where non-self-employed individuals report higher counts of work interference, especially in the "Sometimes" and "Never" categories.
   * Self-employed individuals show a more balanced distribution across the categories, but still have fewer counts in the "Often" category compared to non-self-employed individuals.
3. **Clinical Relevance:**
   * Understanding the relationship between work interference and self-employment can provide insights into how work conditions affect mental health and productivity.
   * This information could be valuable for employers and policymakers in creating supportive work environments.

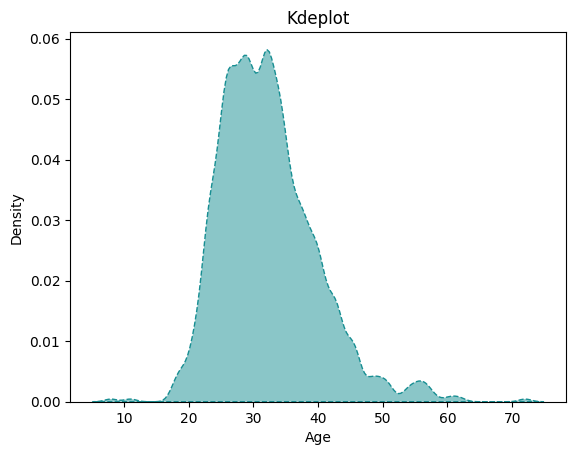
**Conclusion:**

The graph reveals that non-self-employed individuals report more frequent work interference compared to their self-employed counterparts. The most significant difference is observed in the ‘Sometimes’ category, where non-self-employed individuals show a higher count. This suggests that self-employment may offer more flexibility, potentially reducing the frequency of work interference. However, further analysis is needed to understand the underlying factors contributing to these differences.

* **Interest Rate Trends Across Different Countries**

This line plot illustrates the fluctuations in interest rates across various countries. The x-axis represents different countries, while the y-axis shows the interest rates in percentage, ranging from approximately 0.25% to 1.5%. The line plot, along with the shaded area, indicates the variability and trends in interest rates over a specified period.

**Key observations**

1. **Axes:**
   * X-axis: Represents different countries, including Bahamas, United States, Netherlands, United Kingdom, Canada, India, and others.
   * Y-axis: Represents the number of employees, with categories such as "More than 1000," "100-500," and "6-25."
2. **Data Representation:**
   * The blue line represents the average number of employees for each country.
   * The shaded area around the line indicates variability or uncertainty in the data, showing the range of employee counts for each country.
3. **Trends:**
   * The line plot shows fluctuations in the number of employees across different countries, with some countries exhibiting higher employee counts than others.
   * Notably, certain countries like the United States and the Netherlands appear to have higher employee counts, while others like India show more variability.
4. **Countries of Interest:**
   * Certain countries, such as the United States and the United Kingdom, may show higher employee counts, indicating larger companies or a greater number of businesses.
   * Countries like the Bahamas and smaller European nations may show lower employee counts, reflecting smaller business sizes or fewer companies.
5. **Variability:**
   * The shaded areas indicate that there is considerable variability in the number of employees in certain countries, particularly in the middle range of employee counts.
   * This variability may suggest differences in labor markets, economic conditions, or industry presence.
6. **Clinical Relevance:**
   * Understanding the distribution of employees across countries can provide insights into workforce demographics and economic health.
   * This information could be valuable for businesses and policymakers in workforce planning and resource allocation.
7. **Implications:**
   * Understanding the distribution of employee counts across countries can inform business strategies, workforce planning, and economic analysis.
   * Countries with higher employee counts may present more opportunities for employment and economic growth, while those with lower counts may indicate smaller markets or industries.
8. **Clinical Relevance:**
   * Understanding the distribution of employees across countries can provide insights into workforce demographics and economic health.
   * This information could be valuable for businesses and policymakers in workforce planning and resource allocation.

**Conclusion:**

The graph reveals significant variations in interest rates among different countries, reflecting diverse economic policies and conditions. Peaks and troughs in the plot suggest that while some countries experience higher interest rates, others maintain lower rates. This variability highlights the dynamic nature of global financial markets and the impact of national economic strategies on interest rates.

* **Age Distribution Analysis Using Kernel Density Estimation**

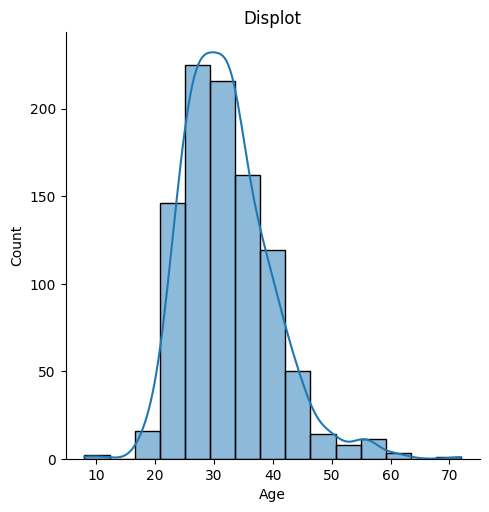
This graph represents a Kernel Density Estimate (KDE) plot, which visualizes the distribution of ages within a dataset. The x-axis represents age, ranging from 10 to 70 years, while the y-axis represents the density, indicating the probability of different age values. The KDE plot smooths out the data to provide an intuitive understanding of where age values are concentrated, highlighting the overall distribution pattern.

**Key observations:**

1. **Axes:**
   * X-axis: Represents age, ranging from approximately 10 to 75 years.
   * Y-axis: Represents the density of ages, indicating how frequently different age values occur in the dataset.
2. **Distribution Shape:**
   * The KDE plot shows a smooth curve representing the density of age values.
   * The peak of the distribution is around the age of 30, indicating that this is the most common age in the dataset.
3. **Density Interpretation:**
   * The highest density is observed around the ages of 25 to 35, suggesting that a significant portion of the population falls within this age range.
   * The density gradually decreases towards the lower (10-20 years) and higher (60-75 years) ends, indicating fewer occurrences of ages in these ranges.
4. **Lower Density Areas:**
   * There are very few individuals in the older age brackets (50-70), indicating that the dataset is skewed towards younger individuals.
   * The tail end of the distribution shows a very low density for ages above 40, suggesting a lack of representation in older demographics.
5. **Trends:**
   * The plot suggests a bimodal distribution, with a noticeable peak around the 30s and a smaller peak in the 20s.
   * This indicates that there may be two distinct groups within the population, possibly reflecting different life stages or demographic factors.
6. **Clinical Relevance:**
   * Understanding the age distribution can provide insights into the demographics of the population being studied, which can be important for healthcare planning and resource allocation.
   * This information may help identify age-related trends in health outcomes or service utilization.

**Conclusion:**

The KDE plot reveals that the age distribution within this dataset is unimodal, with a significant concentration of individuals around the age of 30. The density tapers off as age increases or decreases from this central peak, indicating fewer individuals at younger and older ages. This suggests that the dataset predominantly consists of young adults or middle-aged individuals, providing insights into the age demographics of the population sampled.

* **Age Distribution Analysis with Density Overlay**

The graph you've provided is a displot (distribution plot) that combines a histogram and a kernel density estimate (KDE) to visualize the distribution of ages in a dataset. The x-axis represents age, ranging from 0 to 70 years, while the y-axis shows the count of individuals in each age interval. The bars indicate the frequency of occurrences, and the smooth curve represents the probability density, providing a comprehensive view of the data’s distribution.

**Key observations:**

1. **Axes:**
   * X-axis: Represents age, ranging from approximately 10 to 70 years.
   * Y-axis: Represents the count of occurrences for each age group.
2. **Displot:**
   * The bars represent the frequency of individuals within specific age ranges. The histogram shows that the majority of individuals are concentrated in the 20-40 age range.
3. **Frequency Counts:**
   * The bars represent the number of individuals in each age group, with a noticeable decline in frequency as age increases beyond 40.
   * There are fewer individuals in the older age brackets (50-70), suggesting a younger demographic in the dataset.
4. **Kernel Density Estimate (KDE):**
   * The blue line represents the KDE, providing a smooth estimate of the distribution of ages.
   * The peak of the KDE is around the age of 30, indicating that this is the most common age in the dataset.
5. **Distribution Shape:**
   * The distribution appears to be roughly bell-shaped, suggesting a normal distribution centered around the 30s.
   * There is a gradual decline in counts as age increases beyond 40, indicating fewer occurrences in older age groups.
6. **Variability:**
   * The displot shows some variability in the younger age groups, with counts decreasing as age increases.
   * The KDE line indicates that while there are individuals in older age groups, they are less frequent compared to those in their 20s and 30s.
7. **Clinical Relevance:**
   * Understanding the age distribution can provide insights into the demographics of the population being studied, which can be important for healthcare planning and resource allocation.
   * This information may help identify age-related trends in health outcomes or service utilization

**Conclusion:**

The displot effectively visualizes age distribution, combining a histogram and KDE for a comprehensive view. The graph shows an approximately normal distribution, peaking around the late twenties to early thirties, indicating a dataset predominantly of young to middle-aged adults. The KDE overlay highlights the central tendency and spread of ages, offering valuable insights into the demographic composition of the population under study. This understanding is crucial for analysing health outcomes and other relevant factors.

* **Age Distribution Analysis with Density Overlay**

This scatter plot visualizes the relationship between the founder’s age and company size across various countries. The x-axis represents the age of the founders, ranging from 0 to over 20 years, while the y-axis lists different countries. The dots are color-coded to indicate different company sizes: 1-5 employees, 6-20 employees, 21-100 employees, 101-500 employees, and more than 1000 employees.

**Key observations-**

1. **Axes:**
   * X-axis: Represents age, ranging from approximately 10 to 70 years.
   * Y-axis: Represents different countries, listed vertically.
2. **Data Points:**
   * Each point on the scatter plot represents a specific country and is color-coded based on the number of employees:

* Red: 1-5 employees
* Pink: 6-25 employees
* Blue: 26-100 employees
* Cyan: 100-500 employees
* Purple: More than 1000 employees
* Yellow: 500-1000 employees

1. **Trends:**
   * The scatter plot shows a concentration of data points around the ages of 20 to 40, indicating that many employees fall within this age range.
   * There are fewer data points for older age groups, suggesting a potential trend where younger individuals are more prevalent in the workforce.
2. **Insights:**
   * Countries with 1-5 employees (pink) are prevalent across various ages, indicating small businesses or startups.
   * Countries with 1-5 employees (pink) are prevalent across various ages, indicating small businesses or startups.
   * The More than 1000 employees (light blue) category has fewer data points, indicating that large companies are less common across the represented age range.
3. **Implications:**
   * Understanding the age distribution of employees in relation to the number of employees can provide insights into workforce demographics, which can be valuable for businesses and policymakers.
   * This information could help in workforce planning, recruitment strategies, and addressing age-related workforce challenges.

**Conclusion:**

The scatter plot effectively visualizes the relationship between age and the number of employees across countries, revealing significant trends and concentrations. It suggests diverse company sizes founded by individuals of varying ages, indicating that entrepreneurship transcends age and geography. The widespread distribution underscores the global and inclusive nature of business creation, highlighting a universal entrepreneurial spirit. This insight is crucial for understanding the factors influencing these patterns and the diverse scales of businesses worldwide.

**OVERALL CONCLUSION**

The analysis of the mental health in tech dataset reveals critical mental health challenges faced by tech professionals. Key factors such as job role, work environment, and access to mental health resources significantly influence outcomes, highlighting the need for targeted interventions and support systems within the industry.

Individuals in high-pressure roles or lacking mental health resources are more prone to anxiety, depression, and burnout. Employers must prioritize mental health support, foster a supportive work environment, and promote work-life balance to improve the well-being of tech professionals.

Continuous monitoring and assessment of mental health trends in the tech industry are essential. Leveraging data processing techniques and machine learning algorithms provides a robust framework for understanding and addressing mental health challenges, benefiting employers, policymakers, and mental health professionals.

In conclusion, this study contributes valuable insights for improving tech professionals' mental health. Addressing the unique challenges of the tech industry can create a healthier work environment that fosters innovation and productivity. These findings will shape future mental health initiatives and policies within the tech sector.