Project Roadmap: Employee Wellness Analysis

Here is a step-by-step plan for your project.

Step 1: Understand the Dataset (Data Loading and Initial Exploration)

First things first, we need to get a feel for the data. Once you share the CSV file, we will:

- **Load the Data**: We'll use the pandas library in Python to load your CSV file into a DataFrame.
- **First Look**: We'll use functions like `head()` to see the first few rows, `info()` to understand the data types and non-null values for each column, and `shape` to see the number of rows and columns.
- **Descriptive Statistics**: For numerical columns like 'Age', we'll use `describe()` to get a statistical summary (mean, median, standard deviation, etc.).

Step 2: Data Preprocessing (Data Cleaning)

Raw data is rarely perfect. So, our next step will be to clean it up to make it ready for analysis.

- **Handle Missing Values**: We'll check for any missing data in columns and decide on a strategy to handle them (e.g., filling them with a specific value or dropping them).
- **Check for Duplicates**: We'll look for and remove any duplicate entries to ensure data accuracy.
- **Correct Data Types**: We will make sure each column has the correct data type (e.g., 'Age' should be a number, 'Timestamp' should be a datetime object).
- **Address Inconsistencies**: We'll check for and correct any inconsistencies in categorical data (e.g., "Male", "male", "M" should all be standardized to "Male").

Step 3: Exploratory Data Analysis (Diving Deep with Visualizations)

This is the most exciting part! We'll start exploring the data to find patterns, relationships, and insights.

- **Univariate Analysis (Analyzing one variable at a time)**:
    - For **categorical variables** (like 'Gender', 'family_history', 'treatment'), we will use bar charts and pie charts to see the distribution of values.
    - For **numerical variables** (like 'Age'), we will use histograms and box plots to understand their distribution and identify any outliers.

- **Bivariate and Multivariate Analysis (Analyzing relationships between variables)**:
    - We will explore relationships between different variables. For example:
        - How does `treatment` vary with `family_history`?
        - Is there a connection between `work_interfere` and `mental_health_consequence`?
        - Does the `wellness_program` availability impact whether employees `seek_help`?
    - We'll use visualizations like stacked bar charts, grouped bar charts, scatter plots, and heatmaps to uncover these relationships.

Step 4: Insights and Interpretation (Connecting the Dots)

As we perform the analysis, we'll be noting down the key findings. The goal here is to answer the main project question: **"Who are the employees who are in need or may be in need of treatment?"**. We'll summarize our findings in a clear and concise way, for example:

- "Employees with a family history of mental illness are X% more likely to need treatment."
- "A majority of employees who feel their mental health interferes with their work are not seeking help due to concerns about anonymity."
- "In companies with a formal wellness program, employees are more open to discussing mental health."

Step 5: Presentation Report (Putting it all together)

Finally, we will compile all our work into a well-documented Jupyter Notebook. This will include:

- The Python code we used for analysis and visualization.
- The visualizations themselves (the plots and charts).
- Clear and concise text that explains the insights we've derived from each step of the analysis.

This will be your final submission for the internship project.