Phase 3: Development part 1

Title: Being the analysis by loading and preprocessing the Mental Health In Tech survey dataset

Introduction:

To begin building a project using the Mental Health in Tech survey dataset, you'll need to follow a series of steps for loading and preprocessing the data. Please note that I don't have access to specific datasets, so I'll provide a general outline of the process. You should replace "mental_health_in_tech_survey.csv" with the actual file path or URL of your dataset.

Dataset:

Age	Gender	Country	state	Self employed	Family_history	treatment	work_interfere	no_employees	Remote_work	Tech_company	benefits
37	Female	United States	IL	NA	No	Yes	Often	Jun-25	No	Yes	Yes
44	M	United States	IN	NA	No	No	Rarely	More than 1000	No	No	Don't know
32	Male	Canada	NA	NA	No	No	Rarely	Jun-25	No	Yes	No
31	Male	United Kingdom	NA	NA	Yes	Yes	Often	26-100	No	Yes	No
31	Male	United States	TX	NA	No	No	Never	100-500	Yes	Yes	Yes
33	Male	United States	TN	NA	Yes	No	Sometimes	Jun-25	No	Yes	Yes
35	Female	United States	MI	NA	Yes	Yes	Sometimes	01-May	Yes	Yes	No
39	M	Canada	NA	NA	No	No	Never	01-May	Yes	Yes	No
42	Female	United States	IL	NA	Yes	Yes	Sometimes	100-500	No	Yes	Yes
23	Male	Canada	NA	NA	No	No	Never	26-100	No	Yes	Don't know
31	Male	United States	OH	NA	No	Yes	Sometimes	Jun-25	Yes	Yes	Don't know
29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes	Yes	Don't know
42	female	United States	CA	NA	Yes	Yes	Sometimes	26-100	No	No	Yes
36	Male	United States	CT	NA	Yes	No	Never	500-1000	No	Yes	Don't know
27	Male	Canada	NA	NA	No	No	Never	Jun-25	No	Yes	Don't know
29	female	United States	IL	NA	Yes	Yes	Rarely	26-100	No	Yes	Yes
23	Male	United Kingdom	NA	NA	No	Yes	Sometimes	26-100	Yes	Yes	Don't know
32	Male	United States	TN	NA	No	Yes	Sometimes	Jun-25	No	Yes	Yes
46	male	United States	MD	Yes	Yes	No	Sometimes	01-May	Yes	Yes	Yes

Program:

Import Libraries:

First, you'll need to import the necessary Python libraries for data manipulation and analysis, such as Pandas, NumPy, and Matplotlib or Seaborn for visualization.

```
 import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

Load The Dataset:

Load the dataset into a Pandas DataFrame. Assuming you have the data in a CSV file, you can use the pd.read_csv() function.

```
df = pd.read_csv("mental_health_in_tech_survey.csv")
```

Exploratory Data Analysis (EDA):

It's crucial to perform EDA to get an understanding of the data. This includes looking at data summary statistics, data types, missing values, and visualizing the data.

```
print(df.head())

print(df.info())

print(df.describe())

print(df.isnull().sum())

sns.countplot(x="mental_health_condition", data=df)

plt.title("Distribution of Mental Health Conditions")

plt.show()
```

Data Preprocessing:

Depending on the dataset's quality, you may need to perform data preprocessing tasks. This can include handling missing values, dealing with outliers, and encoding categorical variables. Here are some common preprocessing tasks:

Handle missing values (e.g., impute or remove rows/columns).

Encode categorical variables (e.g., using one-hot encoding or label encoding).

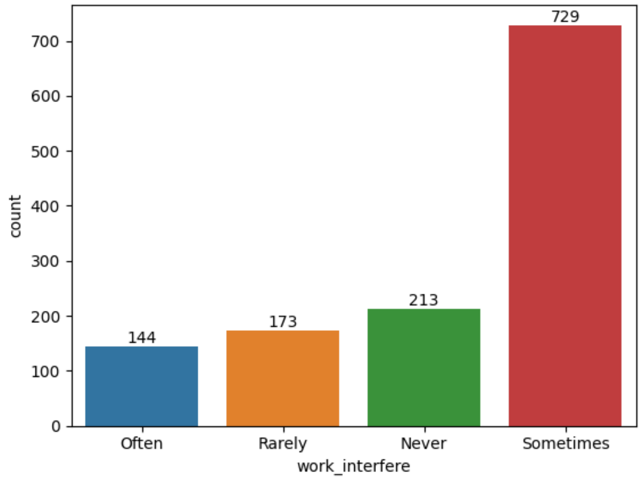
Standardize or normalize numerical features.

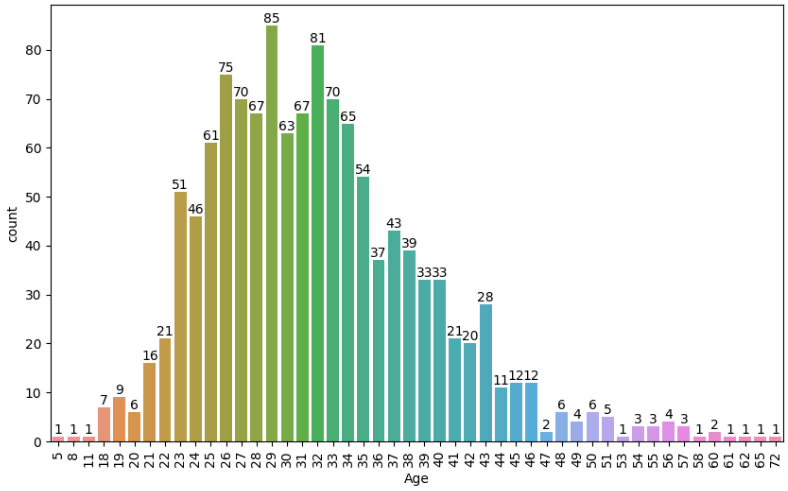
Remove outliers.

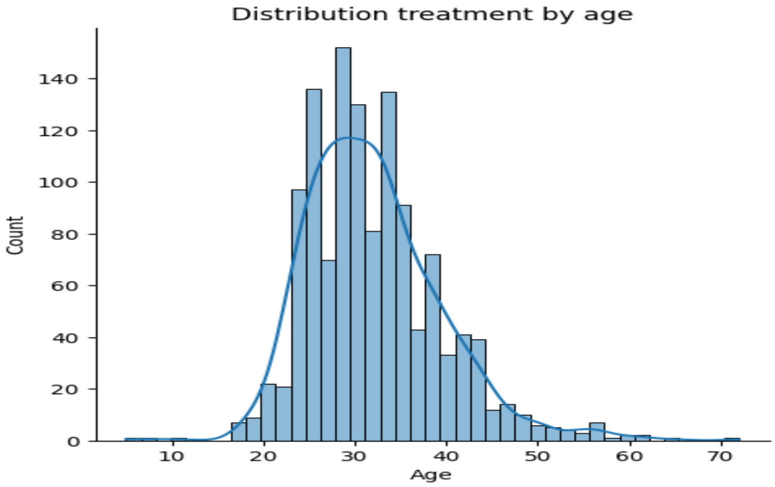
Program:

```
ax = sns.countplot(data = data , x = 'work_interfere');
ax.bar_label(ax.containers[0]);
ax = sns.countplot(data=data, x='work_interfere');
ax.bar_label(ax.containers[0]);

ax = sns.countplot(data=data, x='Gender');
ax.bar_label(ax.containers[0]);

plt.figure(figsize = (10,6))
age_range_plot = sns.countplot(data = data, x = 'Age');
age_range_plot.bar_label(age_range_plot.containers[0]);
plt.xticks(rotation=90);
plt.figure(figsize = (10,6))

age_range_plot = sns.countplot(data = data, x = 'Age');
age_range_plot.bar_label(age_range_plot.containers[0]);
plt.xticks(rotation=90);

plt.figure(figsize = (10,6));

treat = sns.countplot(data = data,  x = 'treatment');

treat.bar_label(treat.containers[0]);

plt.title('Total number of individuals who received treatment or not');
```

Output:







Conclusion:

1**. Loading the Dataset:** Use libraries like Pandas to load your air quality dataset

from a file (e.g., CSV) or another data source.

2. Exploratory Data Analysis (EDA): Conduct basic exploratory data analysis

to understand the structure and characteristics of your data, including checking the first few rows and obtaining summary statistics.

3. Handling Missing Values: Identify and handle missing values in the dataset.

You can choose to remove rows with missing values or impute missing values using appropriate strategies like mean, median, or custom methods.

4. Data Preprocessing: Depending on the nature of your data, perform preprocessing tasks such as encoding categorical variables, scaling numerical features, and creating new features. This step can be tailored to the specific requirements of your dataset and the machine learning model you intend to use.

5. Splitting Data: Split your data into features (X) and the target variable (y). This separation is essential for supervised machine learning tasks.

6. Train-Test Split: Further split your data into training and testing sets, allowing you to evaluate the performance of machine learning models accurately.

Once you've completed these steps, you'll be ready to proceed with the public health awareness campaign analysis using IBM Cognos for visualization. Define your analysis objectives for the campaign data and customize your data preprocessing steps as needed to achieve your specific goals.