Data Collection and Preprocessing Phase

Date	19 May 2025
Team ID	SWTID1750233055
Project Title	Mental Health Prediction
Maximum Marks	6 Marks

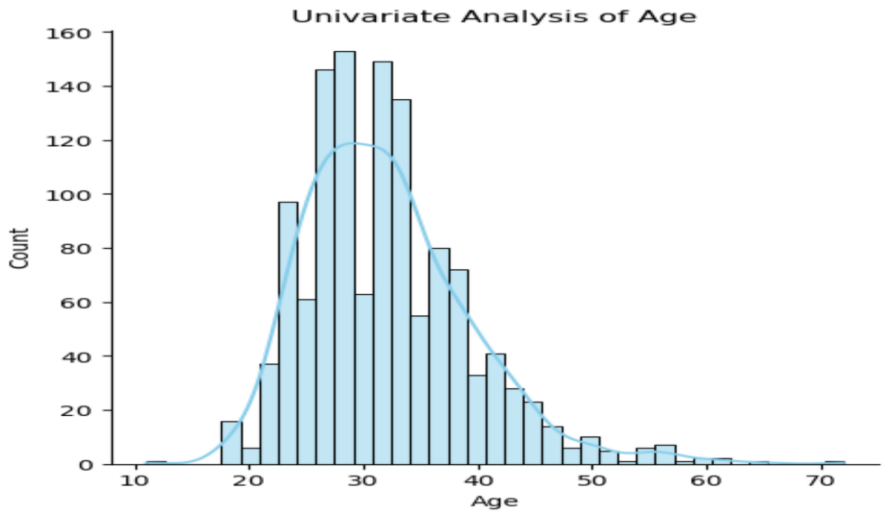
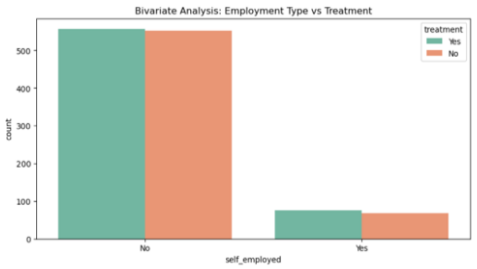
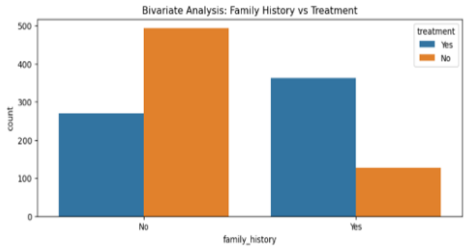
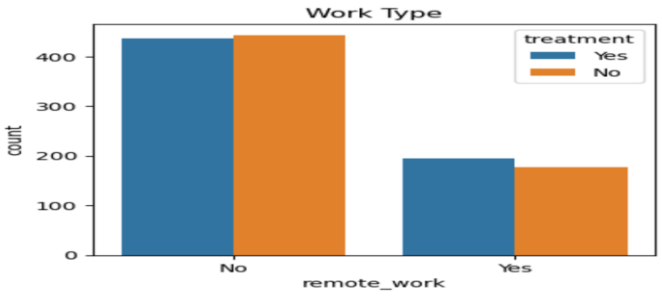
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 1259 rows × 27 columns Descriptive statistics:

	Age	Gender	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	benefits
count	1252.000000	1252	1252	1252	1252	1252	1252	1252	1252	1252
unique	NaN	38	2	2	2	4	6	2	2	3
top	NaN	Male	No	No	Yes	Sometimes	6-25	No	Yes	Yes
freq	NaN	820	1109	763	632	726	289	880	1026	473
mean	32.059904	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	7.309669	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	11.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	27.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	31.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	36.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	72.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Univariate Analysis	

	
Bivariate Analysis	  

Multivariate Analysis	-

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
Loading Data	

```
#1.Loading Data Set

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
data = pd.read_csv(r"C:\Users\jaind\OneDrive\Desktop\ML_Project\data\survey.csv")

data
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	leave	mental_health_consequence
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Somewhat easy	No
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	Don't know	Maybe
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Somewhat difficult	No
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Somewhat difficult	Yes
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	Don't know	No

Handling Missing Data	`data.isnull().sum()`
	
	`#we Have 4 Columns Which Have Null Values`
	`#Change The Value not available to anything Efficient To Reduce overfitting And Increase Accuracy`
	`data['self_employed'].value_counts()`
	
	`self_employed`
	`No 1095`
	`Yes 146`
	`Name: count, dtype: int64`
	
	`#As majority self employed Come with No Replace Null values To No`
	`data['self_employed'] = data['self_employed'].fillna('no')`

Data Transformation	`# Convert 'Age' to numeric and remove rows with invalid age`
	`data['Age'] = pd.to_numeric(data['Age'], errors='coerce')`
	`data.dropna(subset=['Age'], inplace=True)`
	
	`# Drop unnecessary columns (if they exist)`
	`data.drop(columns=['Country', 'Timestamp', 'state', 'comments', 'leave'], errors='ignore', inplace=True)`
	
	`# Split features and label`
	`x = data.drop('treatment', axis=1)`
	`y = data['treatment']`
	
	`# Encode label (target variable)`
	`le = LabelEncoder()`
	`y = le.fit_transform(y)`
	
	`# Identify numerical and categorical columns`
	`num_cols = ['Age']`
	`cat_cols = [col for col in x.columns if col not in num_cols]`
	
	`# Define column transformer`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-