Data Collection and Pre-processing Phase

Date	15 March 2024
Team ID	SWTID1749622322
Project Title	Mental Health Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

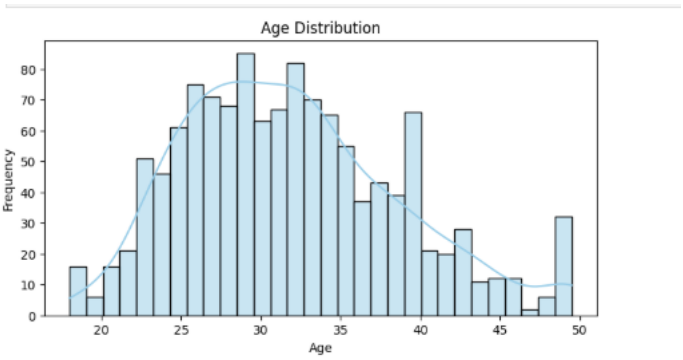
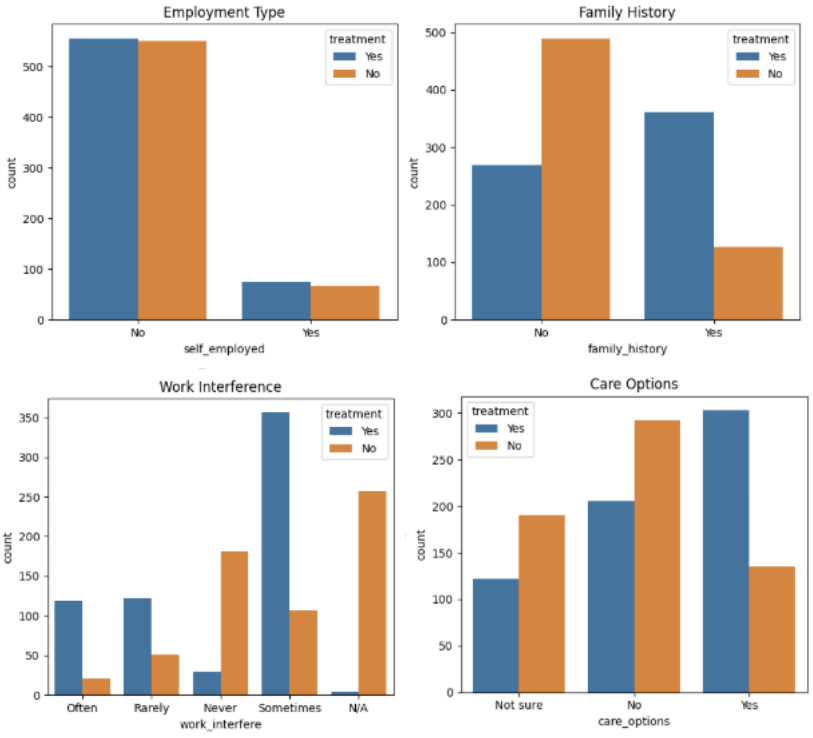
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

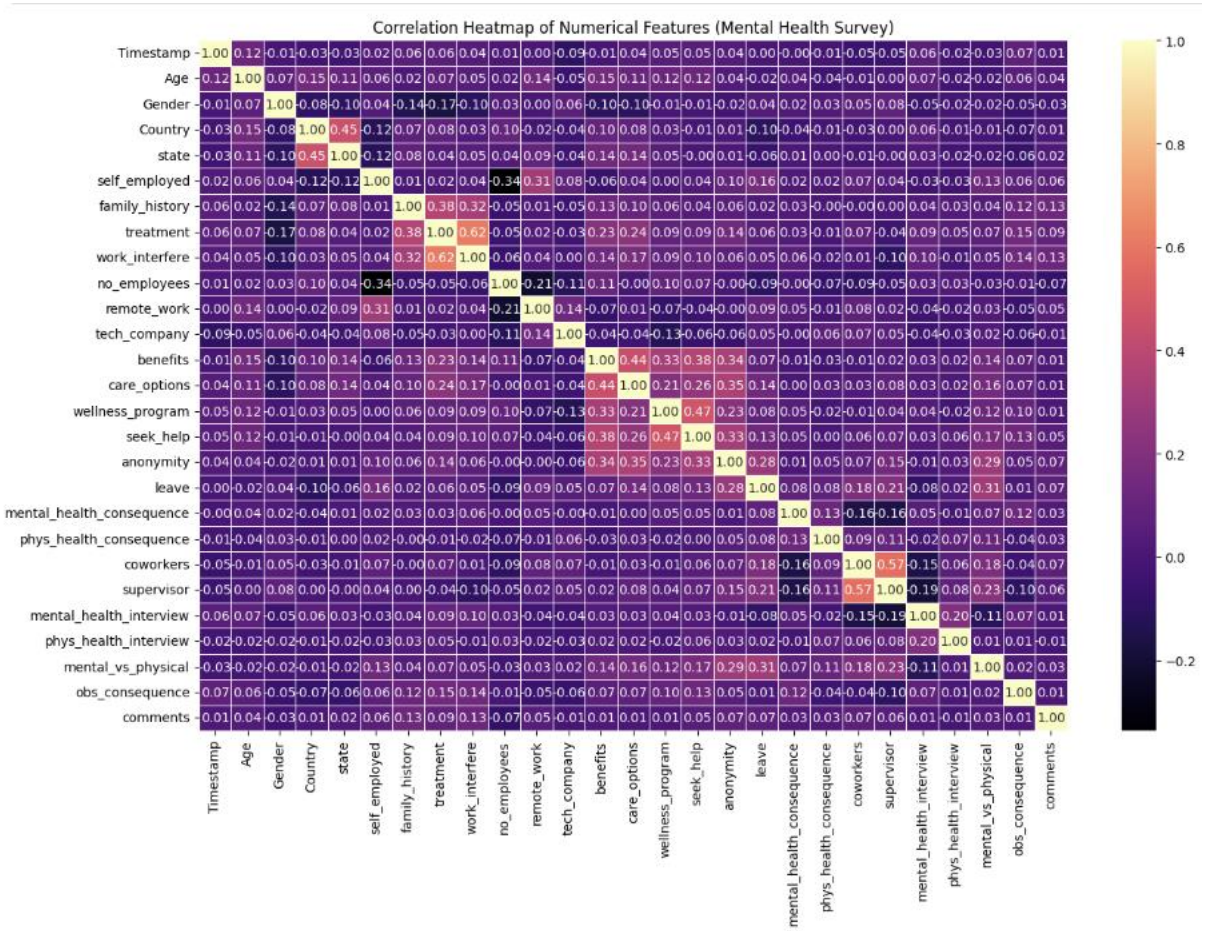
Section	Description
Data Overview	

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employe
count	1247	1247.000000	1247	1247	735	1247	1247	1247	1247	12
unique	1235	NaN	5	46	45	2	2	2	5	
top	2014-08-27 12:43:28	NaN	Male	United States	CA	No	No	Yes	Sometimes	6-2
freq	2	NaN	979	743	137	1107	759	630	463	28
mean	NaN	31.873296	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
std	NaN	6.756424	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
min	NaN	18.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
25%	NaN	27.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
50%	NaN	31.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
75%	NaN	36.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
max	NaN	49.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

11 rows × 27 columns

Univariate Analysis	

<table>
<tr>
<td></td>
<td> 
</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td> 
</td>
</tr>
</table>

Multivariate Analysis	
Outliers and Anomalies	-
Data Pre-processing Code Screenshots	

Loading Data

```
data = pd.read_csv(r'D:\Anshveer - Git - Mental Health\Mental_Health_FINAL\data\survey.csv')
```

```
data.head()
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	I
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Some
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	D k
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Some diff
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Some diff
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	D k

5 rows × 27 columns

Handling Missing Data

```
data['self_employed'].value_counts()
```

```
self_employed
No     1095
Yes     146
Name: count, dtype: int64
```

```
data['self_employed'].fillna('No',inplace=True)
```

```
data['work_interfere'].value_counts()
```

```
work_interfere
Sometimes    465
Never        213
Rarely       173
Often        144
Name: count, dtype: int64
```

```
data['work_interfere'].fillna('N/A',inplace=True)
```

```
data.drop(data[(data['Age']>60)| (data['Age']<18)].index,inplace=True)
```

Data Transformation

```
data['Gender'].replace(['Male ', 'male', 'M', 'm', 'Male', 'Cis Male',
                        'Man', 'cis male', 'Mail', 'Male-ish', 'Male (CIS)',
                        'Cis Man', 'msle', 'Malr', 'Mal', 'maile', 'Make '],
                       'Male', inplace=True)

data['Gender'].replace(['Female ', 'female', 'F', 'f', 'Woman', 'Female',
                        'femail', 'Cis Female', 'cis-female/femme', 'Femake', 'Female (cis)',
                        'woman'],
                       'Female', inplace=True)

data["Gender"].replace(['Female (trans)', 'queer/she/they', 'non-binary',
                        'fluid', 'queer', 'Androgyne', 'Trans-female', 'male leaning androgynous',
                        'Agender', 'A little about you', 'Nah', 'All',
                        'Ostensibly male, unsure what that really means',
                        'Genderqueer', 'Enby', 'p', 'Neuter', 'something kinda male?',
                        'Guy (-ish) ^_^', 'Trans woman',],
                       'Non-Binary', inplace=True)
```

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