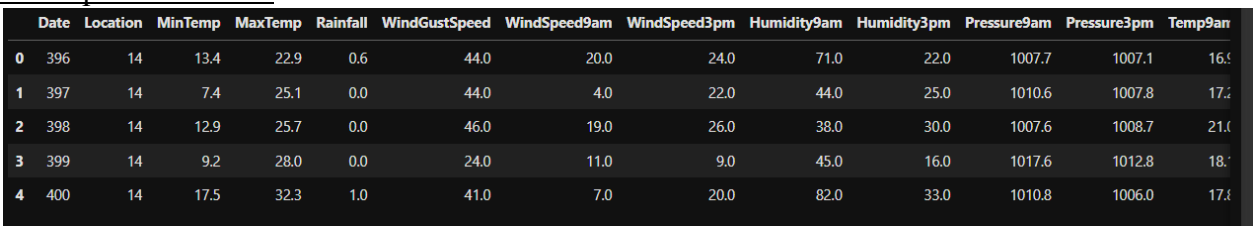
Data Collection and Preprocessing Phase

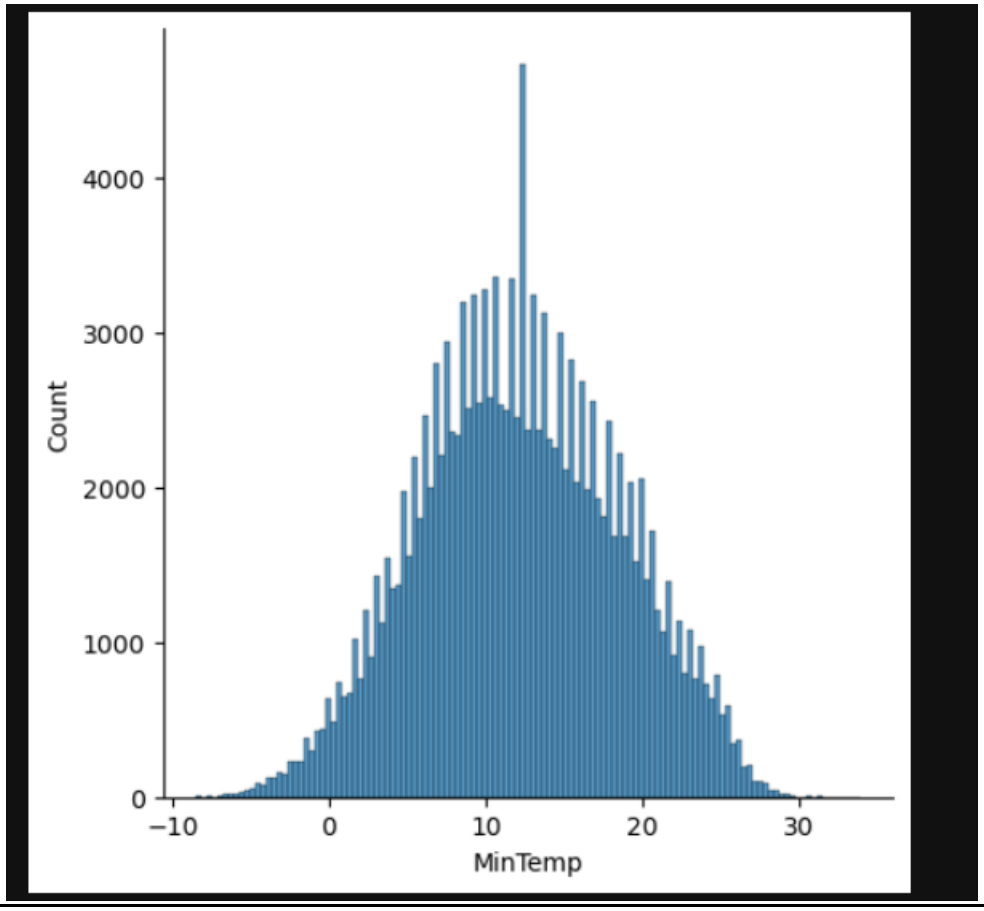
Date	13 July 2024
Team ID	740045
Project Title	Exploratory Analysis of Rain Fall Data in India for Agriculture
Maximum Marks	6 Marks

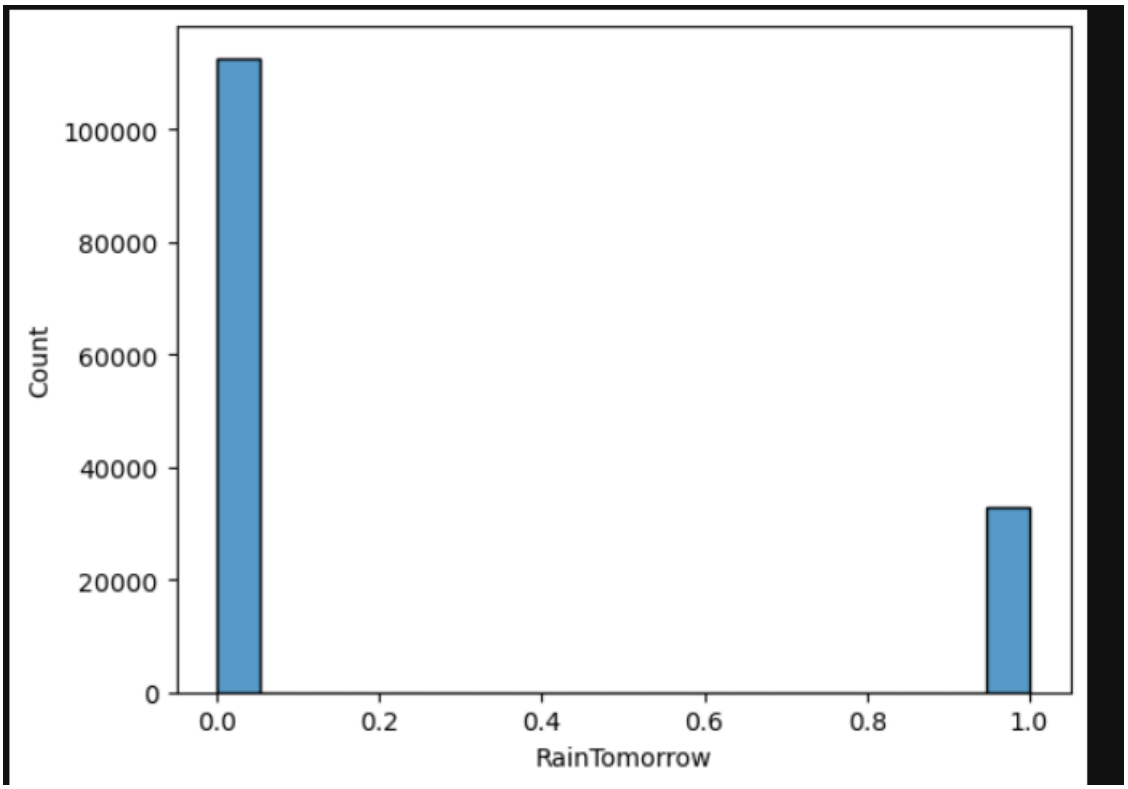
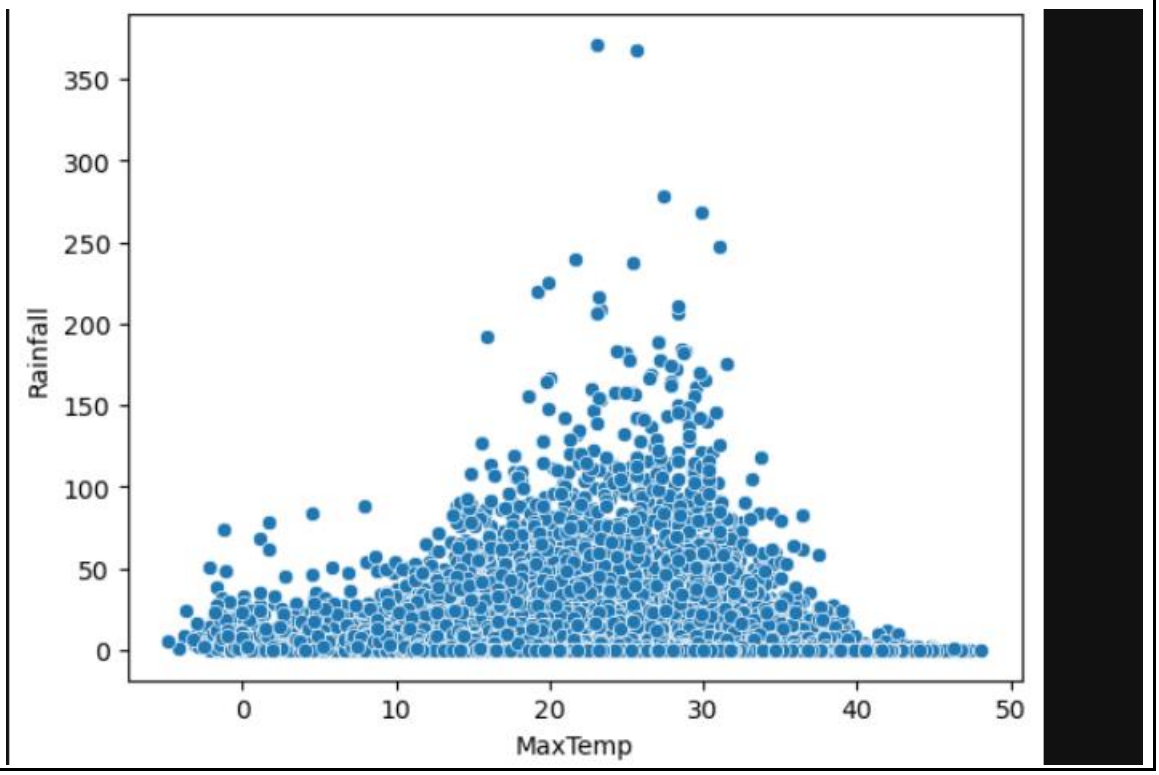
Data Exploration and Preprocessing Report

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: 5 rows × 24 columns Descriptive statistics: 
Univariate Analysis	

	Date	Location	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Temp9am
0	396	14	13.4	22.9	0.6	44.0	20.0	24.0	71.0	22.0	1007.7	1007.1	16.9
1	397	14	7.4	25.1	0.0	44.0	4.0	22.0	44.0	25.0	1010.6	1007.8	17.2
2	398	14	12.9	25.7	0.0	46.0	19.0	26.0	38.0	30.0	1007.6	1008.7	21.0
3	399	14	9.2	28.0	0.0	24.0	11.0	9.0	45.0	16.0	1017.6	1012.8	18.1
4	400	14	17.5	32.3	1.0	41.0	7.0	20.0	82.0	33.0	1010.8	1006.0	17.8


4000
3000
2000
1000
0
Count
−10
0
10
20
30
MinTemp

Bivariate Analysis	
Multivariate Analysis	

Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data

```
[3]: data = pd.read_csv('weather.csv')

[4]: data.head()
```

[4]:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pressure9am	Pressure3p
0	2008-12-01	Delhi	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0	1007.7	1007
1	2008-12-02	Delhi	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0	1010.6	1007
2	2008-12-03	Delhi	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0	1007.6	1008
3	2008-12-04	Delhi	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0	1017.6	1012
4	2008-12-05	Delhi	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0	1010.8	1008

Handling Missing Data

```
# filling the missing data of numeric variables with mean
data['MinTemp'].fillna(data['MinTemp'].mean(),inplace=True)
data['MaxTemp'].fillna(data['MaxTemp'].mean(),inplace=True)
data['Rainfall'].fillna(data['Rainfall'].mean(),inplace=True)
data['WindGustSpeed'].fillna(data['WindGustSpeed'].mean(),inplace=True)
data['WindSpeed9am'].fillna(data['WindSpeed9am'].mean(),inplace=True)
data['WindSpeed3pm'].fillna(data['WindSpeed3pm'].mean(),inplace=True)
data['Humidity9am'].fillna(data['Humidity9am'].mean(),inplace=True)
data['Humidity3pm'].fillna(data['Humidity3pm'].mean(),inplace=True)
data['Pressure9am'].fillna(data['Pressure9am'].mean(),inplace=True)
data['Pressure3pm'].fillna(data['Pressure3pm'].mean(),inplace=True)
data['Temp9am'].fillna(data['Temp9am'].mean(),inplace=True)
data['Temp3pm'].fillna(data['Temp3pm'].mean(),inplace=True)
```

Data Transformation

```
Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'WindGustSpeed',
       'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
       'Pressure9am', 'Pressure3pm', 'Temp9am', 'Temp3pm', 'RainToday',
       'WindGustDir', 'WindDir9am', 'WindDir3pm'],
      dtype='object')
```

```
sc = StandardScaler()
x = sc.fit_transform(x)
```

Feature Engineering	Attached the codes in final submission.

Save Processed Data	-