

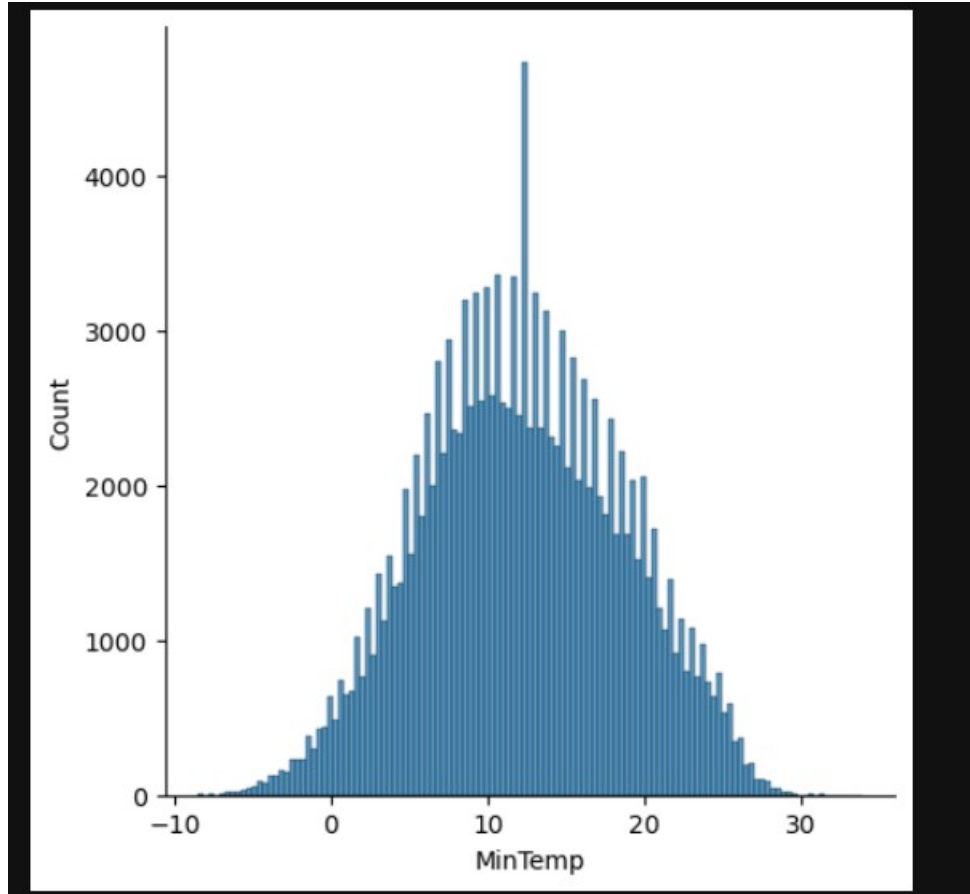
## Data Collection and Preprocessing Phase

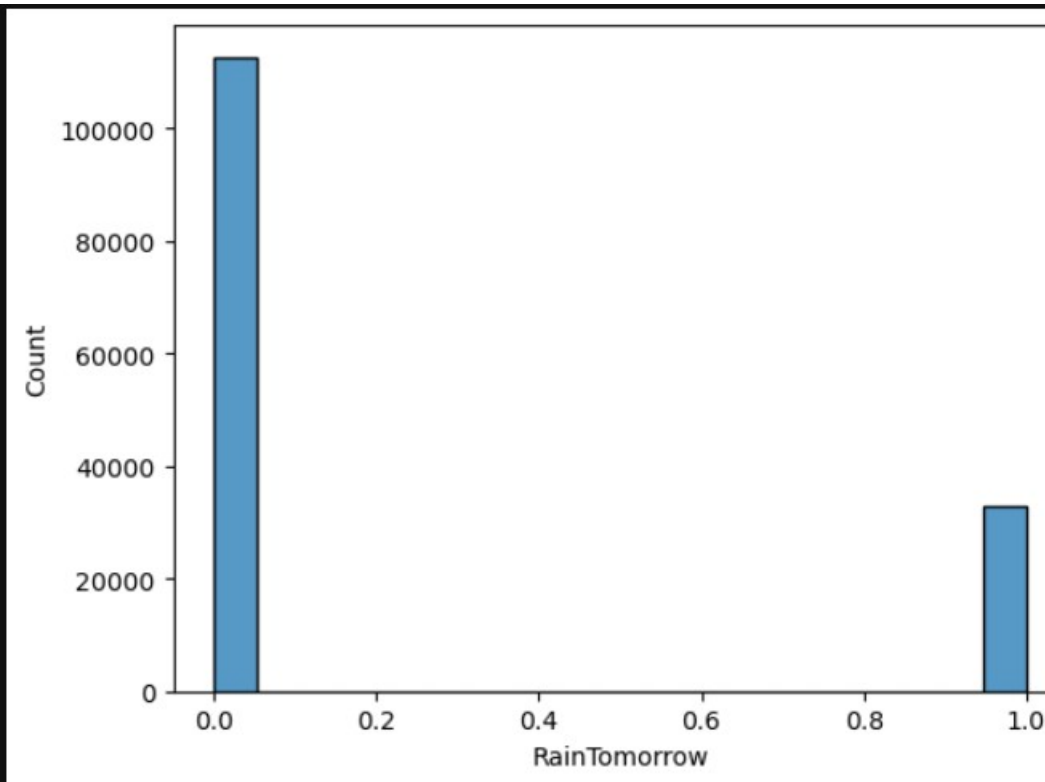
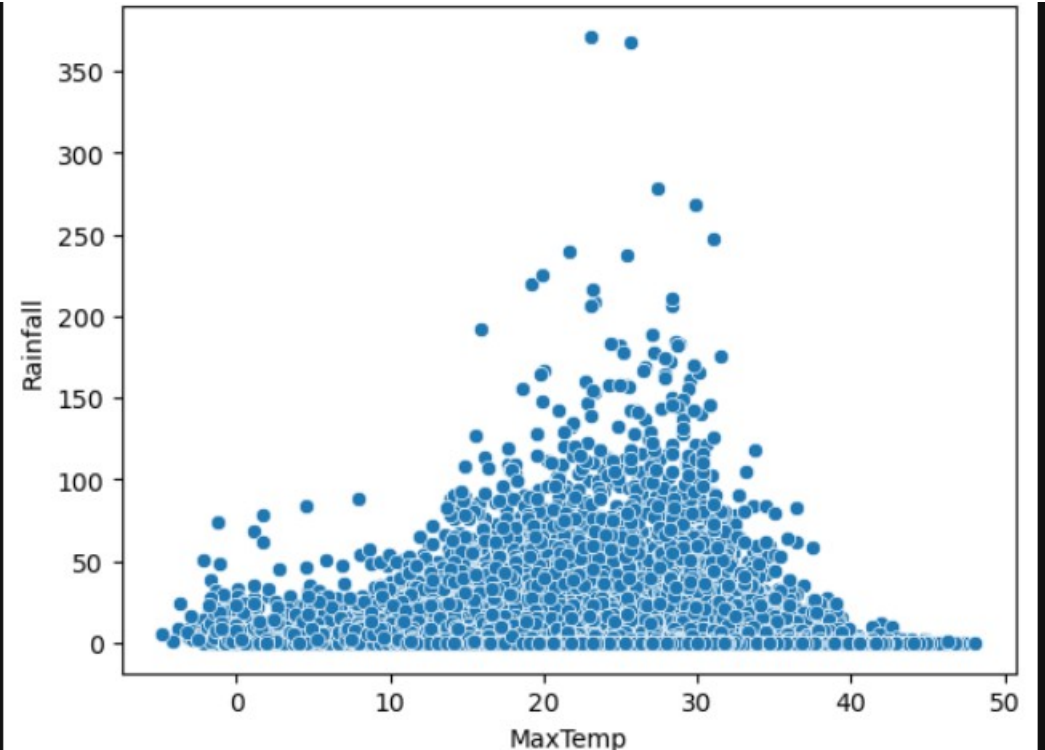
Date	13 July 2024
Team ID	739941
Project Title	Exploratory Analysis of Rain Fall Data in India for Agriculture
Maximum Marks	6 Marks

## 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	<u>Dimension:</u> 5 rows × 24 columns																																																																														
	<u>Descriptive statistics:</u>																																																																														
	<table><tr><th></th><th>Date</th><th>Location</th><th>MinTemp</th><th>MaxTemp</th><th>Rainfall</th><th>WindGustSpeed</th><th>WindSpeed9am</th><th>WindSpeed3pm</th><th>Humidity9am</th><th>Humidity3pm</th><th>Pressure9am</th><th>Pressure3pm</th></tr><tr><td>0</td><td>396</td><td>14</td><td>13.4</td><td>22.9</td><td>0.6</td><td>44.0</td><td>20.0</td><td>24.0</td><td>71.0</td><td>22.0</td><td>1007.7</td><td>1007.7</td></tr><tr><td>1</td><td>397</td><td>14</td><td>7.4</td><td>25.1</td><td>0.0</td><td>44.0</td><td>4.0</td><td>22.0</td><td>44.0</td><td>25.0</td><td>1010.6</td><td>1010.6</td></tr><tr><td>2</td><td>398</td><td>14</td><td>12.9</td><td>25.7</td><td>0.0</td><td>46.0</td><td>19.0</td><td>26.0</td><td>38.0</td><td>30.0</td><td>1007.6</td><td>1007.6</td></tr><tr><td>3</td><td>399</td><td>14</td><td>9.2</td><td>28.0</td><td>0.0</td><td>24.0</td><td>11.0</td><td>9.0</td><td>45.0</td><td>16.0</td><td>1017.6</td><td>1017.6</td></tr><tr><td>4</td><td>400</td><td>14</td><td>17.5</td><td>32.3</td><td>1.0</td><td>41.0</td><td>7.0</td><td>20.0</td><td>82.0</td><td>33.0</td><td>1010.8</td><td>1010.8</td></tr></table>		Date	Location	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	0	396	14	13.4	22.9	0.6	44.0	20.0	24.0	71.0	22.0	1007.7	1007.7	1	397	14	7.4	25.1	0.0	44.0	4.0	22.0	44.0	25.0	1010.6	1010.6	2	398	14	12.9	25.7	0.0	46.0	19.0	26.0	38.0	30.0	1007.6	1007.6	3	399	14	9.2	28.0	0.0	24.0	11.0	9.0	45.0	16.0	1017.6	1017.6	4	400	14	17.5	32.3	1.0	41.0	7.0	20.0	82.0	33.0	1010.8	1010.8
		Date	Location	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm																																																																		
	0	396	14	13.4	22.9	0.6	44.0	20.0	24.0	71.0	22.0	1007.7	1007.7																																																																		
	1	397	14	7.4	25.1	0.0	44.0	4.0	22.0	44.0	25.0	1010.6	1010.6																																																																		
	2	398	14	12.9	25.7	0.0	46.0	19.0	26.0	38.0	30.0	1007.6	1007.6																																																																		
3	399	14	9.2	28.0	0.0	24.0	11.0	9.0	45.0	16.0	1017.6	1017.6																																																																			
4	400	14	17.5	32.3	1.0	41.0	7.0	20.0	82.0	33.0	1010.8	1010.8																																																																			
Univariate Analysis																																																																															



<p>Bivariate Analysis</p>	 <p>A histogram showing the distribution of 'RainTomorrow' values. The x-axis is labeled 'RainTomorrow' and ranges from 0.0 to 1.0 with major ticks every 0.2. The y-axis is labeled 'Count' and ranges from 0 to 100,000 with major ticks every 20,000. There are two bars: a very tall bar at 0.0 with a count of approximately 110,000, and a much shorter bar at 1.0 with a count of approximately 35,000.</p>
<p>Multivariate Analysis</p>	 <p>A scatter plot showing the relationship between 'MaxTemp' (x-axis) and 'Rainfall' (y-axis). The x-axis ranges from 0 to 50 with major ticks every 10. The y-axis ranges from 0 to 350 with major ticks every 50. The plot contains a large number of blue circular data points. Most points are clustered at low rainfall values (below 50) across the temperature range. There is a distinct upward trend in rainfall as temperature increases, particularly between 15 and 35 degrees, with several points reaching rainfall values between 200 and 350.</p>
<p>Outliers and</p>	<p>-</p>

Anomalies																																																																																					
Data Preprocessing Code Screenshots																																																																																					
Loading Data	<pre>[3]: data = pd.read_csv('weather.csv')  [4]: data.head()</pre> <table><thead><tr><th></th><th>Date</th><th>Location</th><th>MinTemp</th><th>MaxTemp</th><th>Rainfall</th><th>Evaporation</th><th>Sunshine</th><th>WindGustDir</th><th>WindGustSpeed</th><th>WindDir9am</th><th>...</th><th>Humidity3pm</th><th>Pr</th></tr></thead><tbody><tr><td>0</td><td>2008-12-01</td><td>Delhi</td><td>13.4</td><td>22.9</td><td>0.6</td><td>NaN</td><td>NaN</td><td>W</td><td>44.0</td><td>W</td><td>...</td><td>22.0</td><td></td></tr><tr><td>1</td><td>2008-12-02</td><td>Delhi</td><td>7.4</td><td>25.1</td><td>0.0</td><td>NaN</td><td>NaN</td><td>WNW</td><td>44.0</td><td>NNW</td><td>...</td><td>25.0</td><td></td></tr><tr><td>2</td><td>2008-12-03</td><td>Delhi</td><td>12.9</td><td>25.7</td><td>0.0</td><td>NaN</td><td>NaN</td><td>WSW</td><td>46.0</td><td>W</td><td>...</td><td>30.0</td><td></td></tr><tr><td>3</td><td>2008-12-04</td><td>Delhi</td><td>9.2</td><td>28.0</td><td>0.0</td><td>NaN</td><td>NaN</td><td>NE</td><td>24.0</td><td>SE</td><td>...</td><td>16.0</td><td></td></tr><tr><td>4</td><td>2008-12-05</td><td>Delhi</td><td>17.5</td><td>32.3</td><td>1.0</td><td>NaN</td><td>NaN</td><td>W</td><td>41.0</td><td>ENE</td><td>...</td><td>33.0</td><td></td></tr></tbody></table>		Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pr	0	2008-12-01	Delhi	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0		1	2008-12-02	Delhi	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0		2	2008-12-03	Delhi	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0		3	2008-12-04	Delhi	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0		4	2008-12-05	Delhi	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0	
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pr																																																																								
0	2008-12-01	Delhi	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0																																																																									
1	2008-12-02	Delhi	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0																																																																									
2	2008-12-03	Delhi	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0																																																																									
3	2008-12-04	Delhi	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0																																																																									
4	2008-12-05	Delhi	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0																																																																									
Handling Missing Data	<pre># 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)</pre>																																																																																				
Data Transformat ion	<pre>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)</pre>																																																																																				
Feature Engineering	Attached the codes in final submission.																																																																																				
Save Processed Data	-																																																																																				