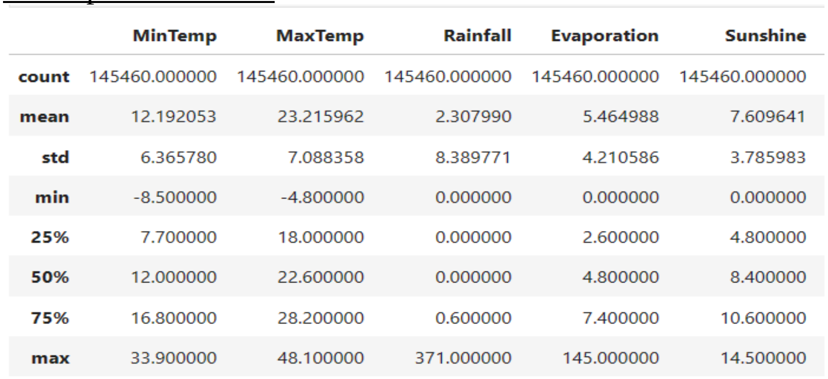
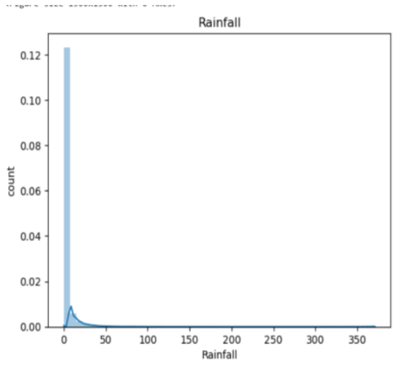
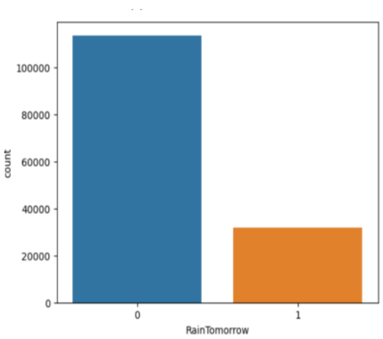
Data Collection and Preprocessing Phase

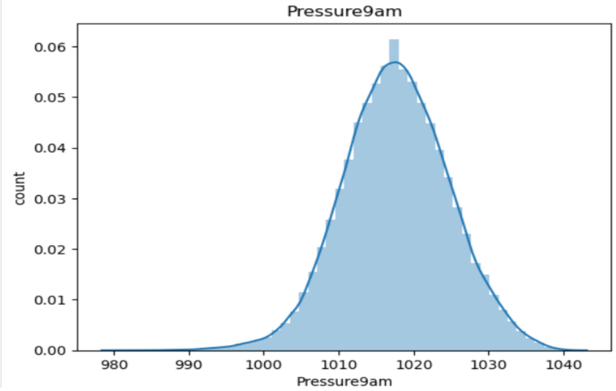
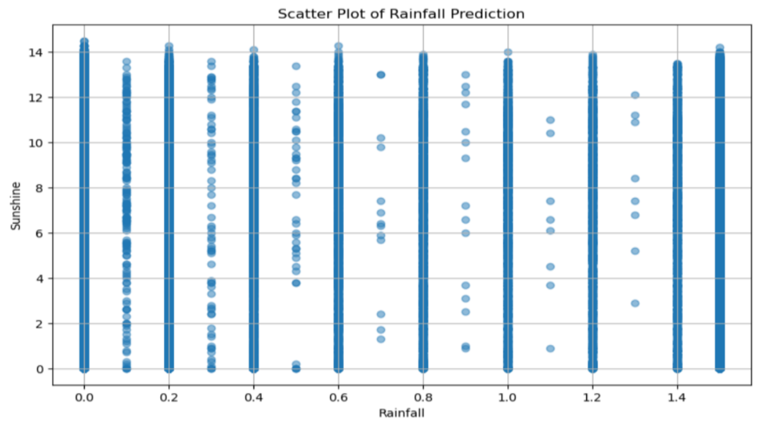
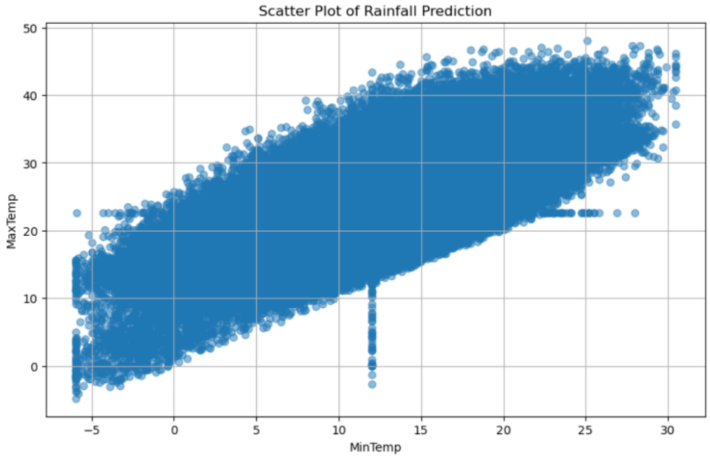
Date	24 April 2024
Team ID	Team-738169
Project Title	Rainfall Prediction Using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

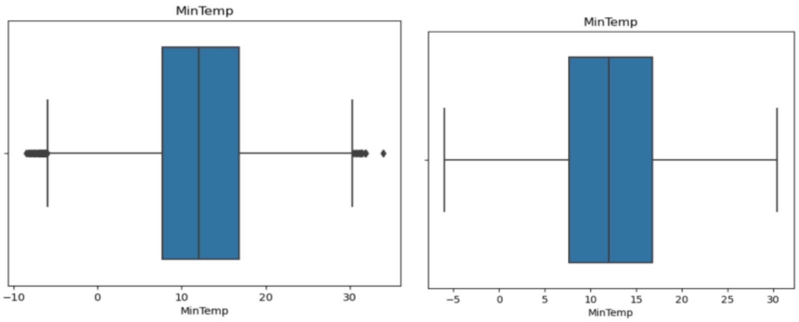
This report presents the findings and steps undertaken during the exploration and preprocessing of the rainfall dataset. The primary objectives were to gain insights into the data distribution, identify potential issues, and prepare the data for further analysis and modeling.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:
(145460, 23)
Descriptive Statistics

<table>
<tr><th></th><th>MinTemp</th><th>MaxTemp</th><th>Rainfall</th><th>Evaporation</th><th>Sunshine</th></tr>
<tr><td>count</td><td>145460.000000</td><td>145460.000000</td><td>145460.000000</td><td>145460.000000</td><td>145460.000000</td></tr>
<tr><td>mean</td><td>12.192053</td><td>23.215962</td><td>2.307990</td><td>5.464988</td><td>7.609641</td></tr>
<tr><td>std</td><td>6.365780</td><td>7.088358</td><td>8.389771</td><td>4.210586</td><td>3.785983</td></tr>
<tr><td>min</td><td>-8.500000</td><td>-4.800000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>25%</td><td>7.700000</td><td>18.000000</td><td>0.000000</td><td>2.600000</td><td>4.800000</td></tr>
<tr><td>50%</td><td>12.000000</td><td>22.600000</td><td>0.000000</td><td>4.800000</td><td>8.400000</td></tr>
<tr><td>75%</td><td>16.800000</td><td>28.200000</td><td>0.600000</td><td>7.400000</td><td>10.600000</td></tr>
<tr><td>max</td><td>33.900000</td><td>48.100000</td><td>371.000000</td><td>145.000000</td><td>14.500000</td></tr>
</table></td></tr>
<tr><td>Univariate Analysis</td><td>

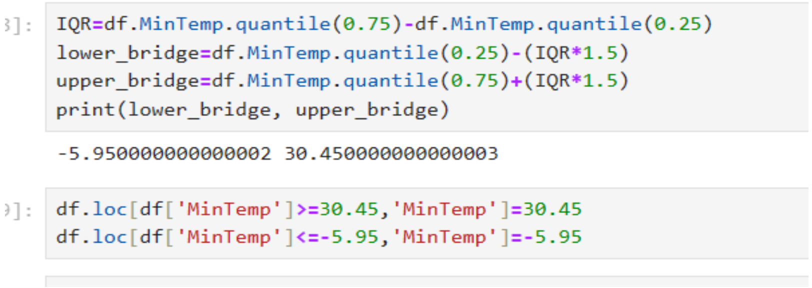

</td></tr>
</table>

	
Bivariate Analysis	 
Multivariate Analysis	-

Outliers and Anomalies



Handeling Outliers by IQR(Inter Quartile Range).



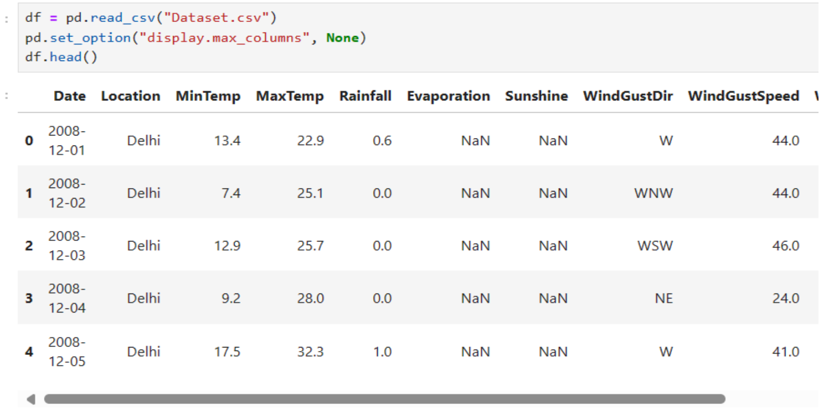
```
IQR=df.MinTemp.quantile(0.75)-df.MinTemp.quantile(0.25)
lower_bridge=df.MinTemp.quantile(0.25)-(IQR*1.5)
upper_bridge=df.MinTemp.quantile(0.75)+(IQR*1.5)
print(lower_bridge, upper_bridge)
```

```
-5.950000000000002 30.450000000000003
```

```
df.loc[df['MinTemp']>=30.45,'MinTemp']=30.45
df.loc[df['MinTemp']<=-5.95,'MinTemp']=-5.95
```

Data Preprocessing Code Screenshots

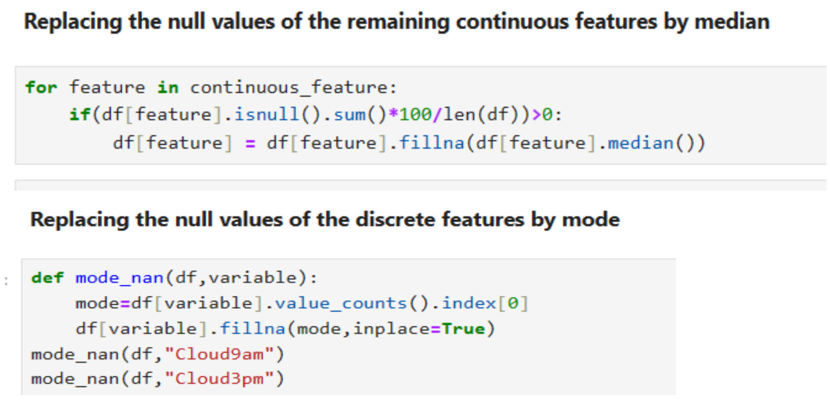
Loading Data



```
df = pd.read_csv("Dataset.csv")
pd.set_option("display.max_columns", None)
df.head()
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed
0	2008-12-01	Delhi	13.4	22.9	0.6	NaN	NaN	W	44.0
1	2008-12-02	Delhi	7.4	25.1	0.0	NaN	NaN	WNW	44.0
2	2008-12-03	Delhi	12.9	25.7	0.0	NaN	NaN	WSW	46.0
3	2008-12-04	Delhi	9.2	28.0	0.0	NaN	NaN	NE	24.0
4	2008-12-05	Delhi	17.5	32.3	1.0	NaN	NaN	W	41.0

Handling Missing Data

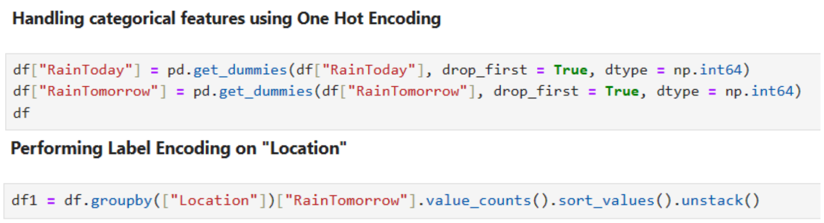


Replacing the null values of the remaining continuous features by median

```
for feature in continuous_feature:
    if(df[feature].isnull().sum()*100/len(df))>0:
        df[feature] = df[feature].fillna(df[feature].median())
```

Replacing the null values of the discrete features by mode

```
def mode_nan(df,variable):
    mode=df[variable].value_counts().index[0]
    df[variable].fillna(mode,inplace=True)
mode_nan(df,"Cloud9am")
mode_nan(df,"Cloud3pm")
```

Data Transformation	
	Handling categorical features using One Hot Encoding
	`df["RainToday"] = pd.get_dummies(df["RainToday"], drop_first = True, dtype = np.int64)`
	`df["RainTomorrow"] = pd.get_dummies(df["RainTomorrow"], drop_first = True, dtype = np.int64)`
	`df`
	Performing Label Encoding on "Location"
	`df1 = df.groupby(["Location"])["RainTomorrow"].value_counts().sort_values().unstack()`

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