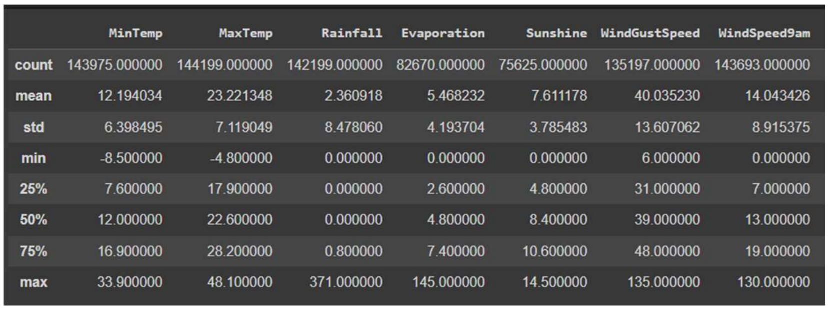
Data Collection and Preprocessing Phase

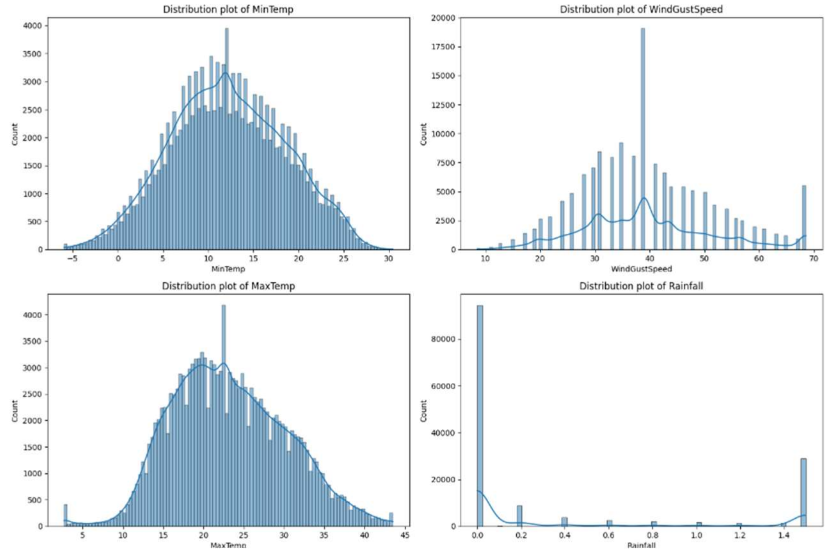
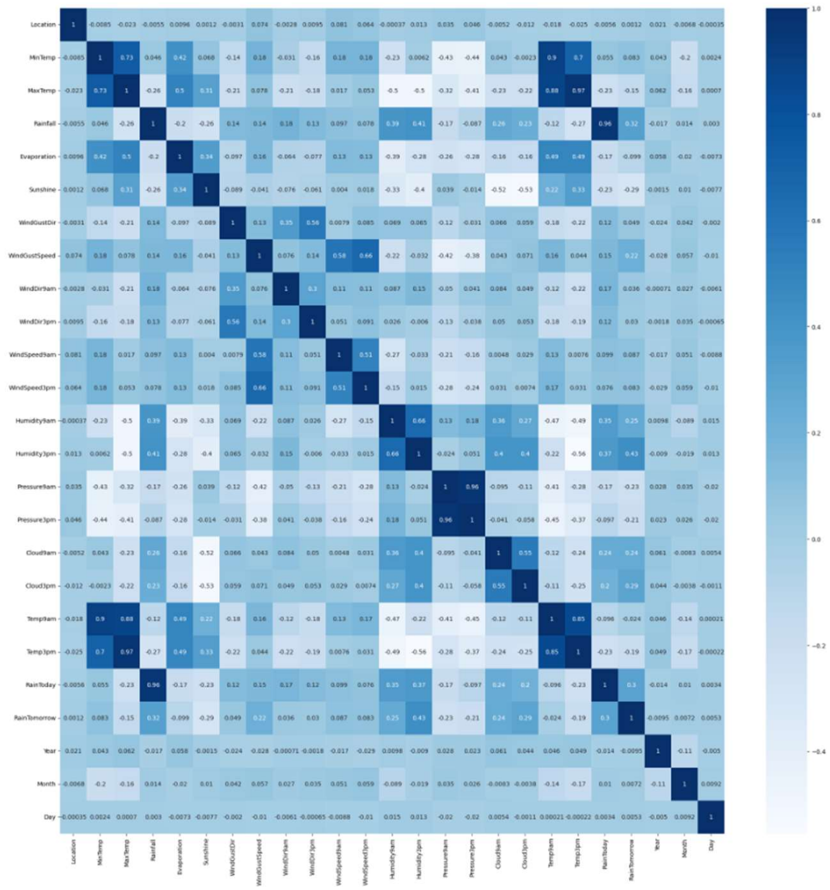
Date	23 September 2024
Team ID	LTVIP2024TMID24992
Project Title	Rainfall Prediction Using Machine Learning
Maximum Marks	6 Marks

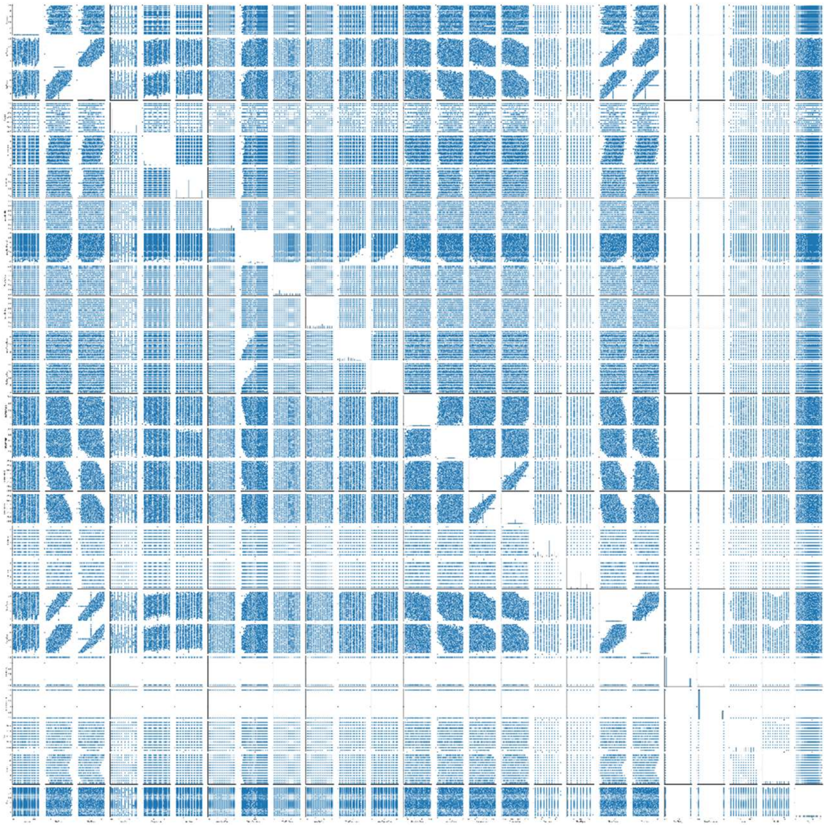
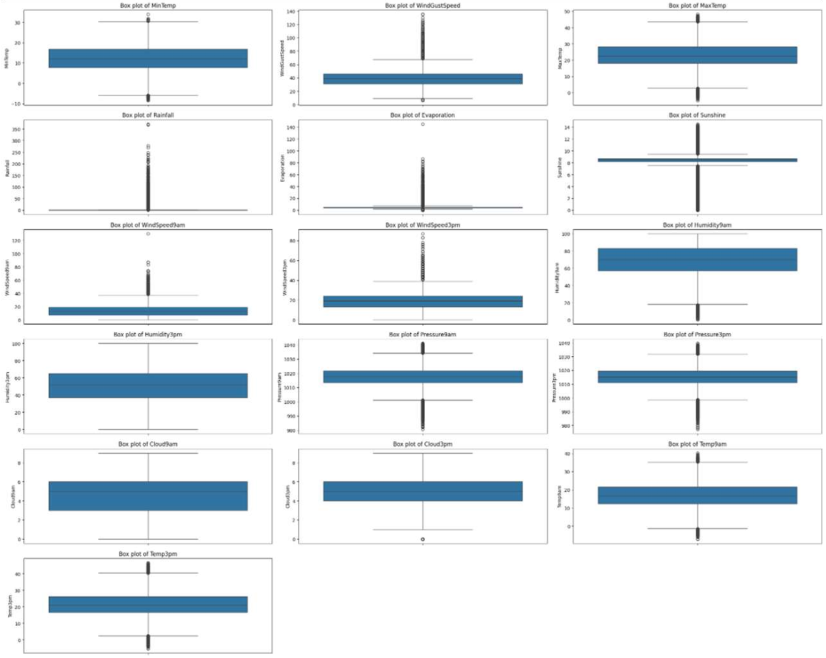
Data Exploration and Preprocessing Template

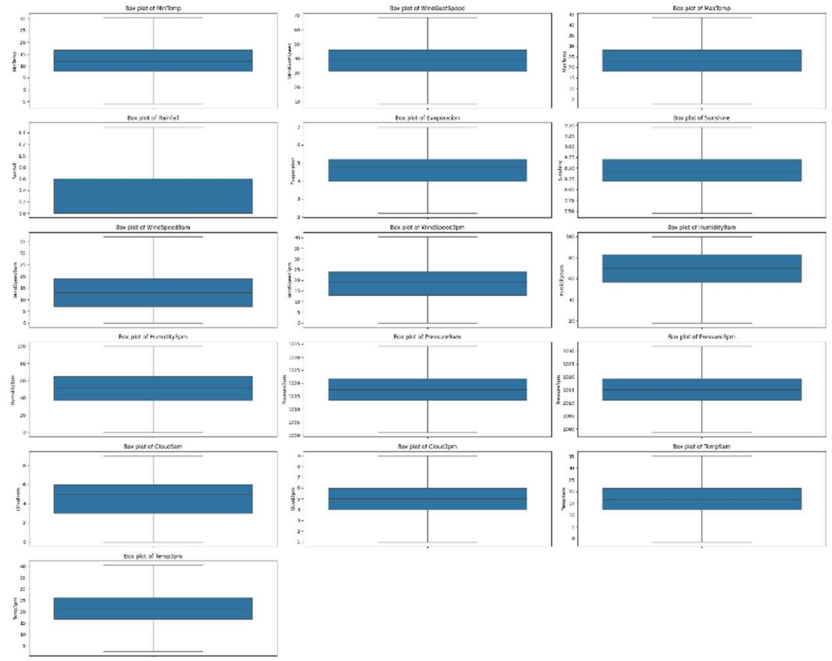
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Dimension: 145460 rows × 23 columns 

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am
count	143975.000000	144199.000000	142199.000000	82670.000000	75625.000000	135197.000000	143693.000000
mean	12.194034	23.221348	2.360918	5.468232	7.611178	40.035230	14.043426
std	6.398495	7.119049	8.478060	4.193704	3.785483	13.607062	8.915375
min	-8.500000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000
25%	7.600000	17.900000	0.000000	2.600000	4.800000	31.000000	7.000000
50%	12.000000	22.600000	0.000000	4.800000	8.400000	39.000000	13.000000
75%	16.900000	28.200000	0.800000	7.400000	10.600000	48.000000	19.000000
max	33.900000	48.100000	371.000000	145.000000	14.500000	135.000000	130.000000

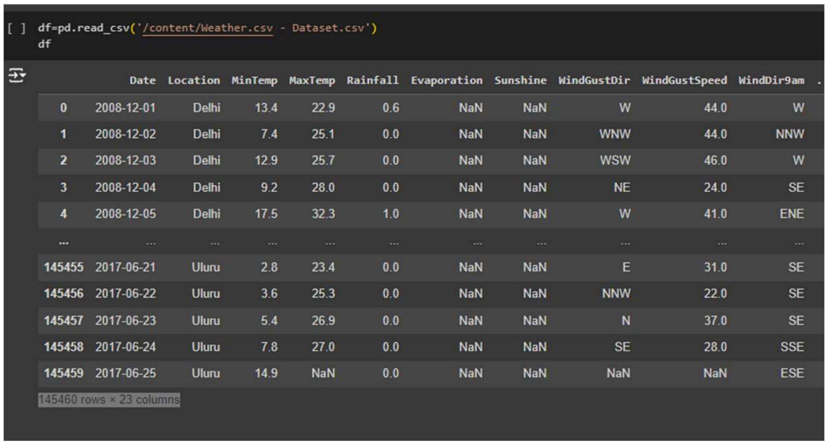
Univariate Analysis	
Bivariate Analysis	

Multivariate Analysis	
Outliers and Anomalies	Identification of outliers  Treatment of outliers.



Data Preprocessing Code Screenshots

Loading Data



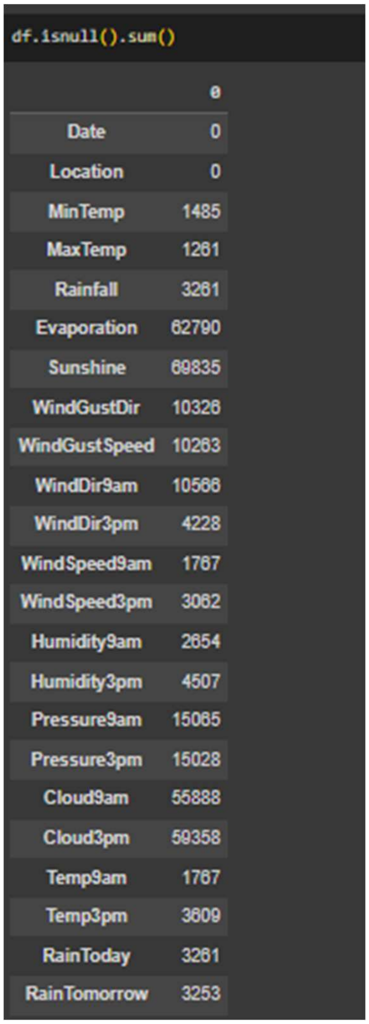
```
[ ] df=pd.read_csv('/content/Weather.csv - Dataset.csv')
    df
```

```
              Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed WindDir9am .
     0  2008-12-01    Delhi    13.4    22.9      0.6         NaN      NaN           W          44.0          W
     1  2008-12-02    Delhi     7.4    25.1      0.0         NaN      NaN         WNW          44.0        NNW
     2  2008-12-03    Delhi    12.9    25.7      0.0         NaN      NaN         WSW          46.0          W
     3  2008-12-04    Delhi     9.2    28.0      0.0         NaN      NaN          NE          24.0         SE
     4  2008-12-05    Delhi    17.5    32.3      1.0         NaN      NaN           W          41.0        ENE
   ...         ...      ...     ...     ...      ...         ...      ...         ...           ...        ...
145455  2017-06-21    Uluru     2.8    23.4      0.0         NaN      NaN           E          31.0         SE
145456  2017-06-22    Uluru     3.6    25.3      0.0         NaN      NaN         NNW          22.0         SE
145457  2017-06-23    Uluru     5.4    26.9      0.0         NaN      NaN           N          37.0         SE
145458  2017-06-24    Uluru     7.8    27.0      0.0         NaN      NaN          SE          28.0        SSE
145459  2017-06-25    Uluru    14.9     NaN      0.0         NaN      NaN         NaN           NaN        ESE

145460 rows × 23 columns
```

Handling Missing Data

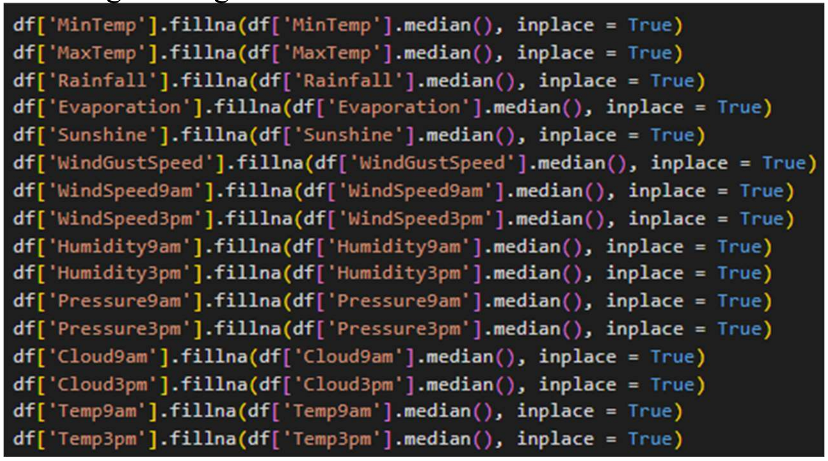
Identifying missing values



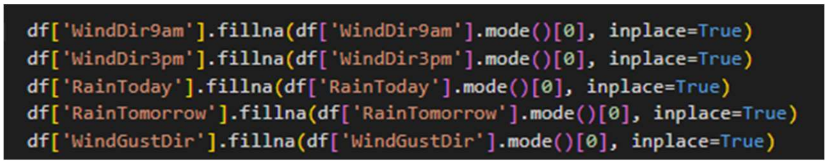
```
df.isnull().sum()
```

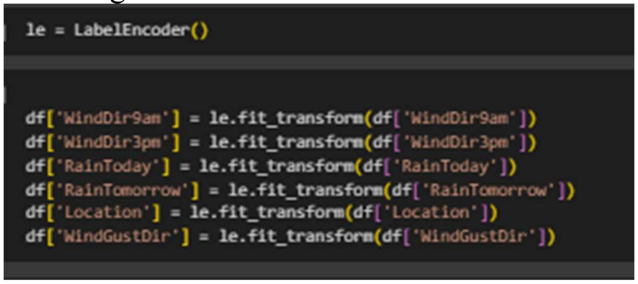
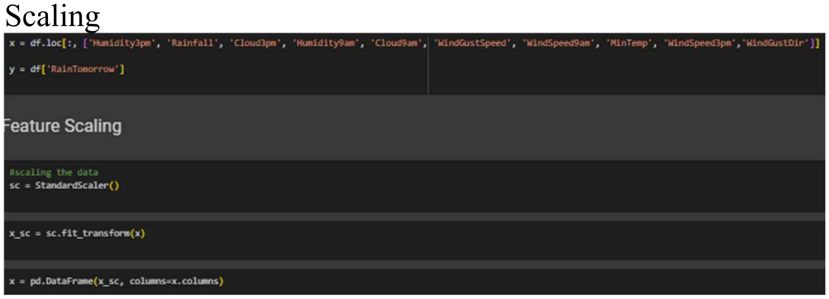
```
                   0
         Date      0
     Location      0
      MinTemp   1485
      MaxTemp   1261
     Rainfall   3261
  Evaporation  62790
     Sunshine  69835
  WindGustDir  10326
WindGustSpeed  10263
   WindDir9am  10566
   WindDir3pm   4228
 WindSpeed9am   1767
 WindSpeed3pm   3062
  Humidity9am   2654
  Humidity3pm   4507
  Pressure9am  15065
  Pressure3pm  15028
     Cloud9am  55888
     Cloud3pm  59358
      Temp9am   1767
      Temp3pm   3609
    RainToday   3261
 RainTomorrow   3253
```

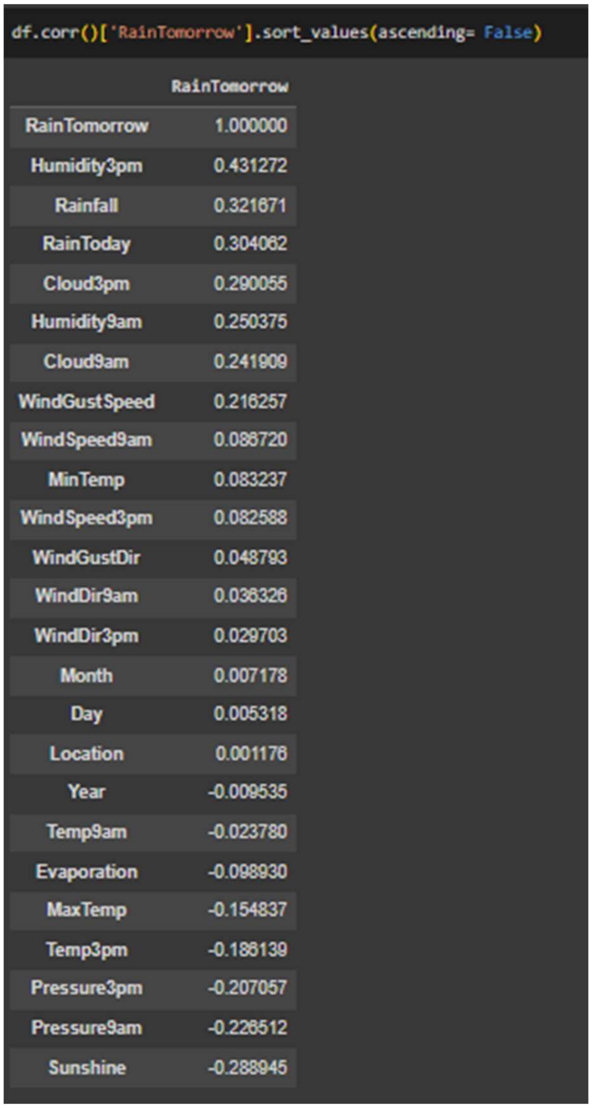
Handling missing values



```
df['MinTemp'].fillna(df['MinTemp'].median(), inplace = True)
df['MaxTemp'].fillna(df['MaxTemp'].median(), inplace = True)
df['Rainfall'].fillna(df['Rainfall'].median(), inplace = True)
df['Evaporation'].fillna(df['Evaporation'].median(), inplace = True)
df['Sunshine'].fillna(df['Sunshine'].median(), inplace = True)
df['WindGustSpeed'].fillna(df['WindGustSpeed'].median(), inplace = True)
df['WindSpeed9am'].fillna(df['WindSpeed9am'].median(), inplace = True)
df['WindSpeed3pm'].fillna(df['WindSpeed3pm'].median(), inplace = True)
df['Humidity9am'].fillna(df['Humidity9am'].median(), inplace = True)
df['Humidity3pm'].fillna(df['Humidity3pm'].median(), inplace = True)
df['Pressure9am'].fillna(df['Pressure9am'].median(), inplace = True)
df['Pressure3pm'].fillna(df['Pressure3pm'].median(), inplace = True)
df['Cloud9am'].fillna(df['Cloud9am'].median(), inplace = True)
df['Cloud3pm'].fillna(df['Cloud3pm'].median(), inplace = True)
df['Temp9am'].fillna(df['Temp9am'].median(), inplace = True)
df['Temp3pm'].fillna(df['Temp3pm'].median(), inplace = True)
```

	 `df['WindDir9am'].fillna(df['WindDir9am'].mode()[0], inplace=True)`
	`df['WindDir3pm'].fillna(df['WindDir3pm'].mode()[0], inplace=True)`
	`df['RainToday'].fillna(df['RainToday'].mode()[0], inplace=True)`
	`df['RainTomorrow'].fillna(df['RainTomorrow'].mode()[0], inplace=True)`
	`df['WindGustDir'].fillna(df['WindGustDir'].mode()[0], inplace=True)`

Data Transformation	Encoding
	 `le = LabelEncoder()`
	`df['WindDir9am'] = le.fit_transform(df['WindDir9am'])`
	`df['WindDir3pm'] = le.fit_transform(df['WindDir3pm'])`
	`df['RainToday'] = le.fit_transform(df['RainToday'])`
	`df['RainTomorrow'] = le.fit_transform(df['RainTomorrow'])`
	`df['Location'] = le.fit_transform(df['Location'])`
	`df['WindGustDir'] = le.fit_transform(df['WindGustDir'])`
	Scaling
	 `x = df.loc[:, ['Humidity3pm', 'Rainfall', 'Cloud3pm', 'Humidity9am', 'Cloud9am', 'WindGustSpeed', 'WindSpeed9am', 'MinTemp', 'WindSpeed3pm','WindGustDir']]`
	`y = df['RainTomorrow']`
	Feature Scaling
	`#scaling the data`
	`sc = StandardScaler()`
	`x_sc = sc.fit_transform(x)`
	`x = pd.DataFrame(x_sc, columns=x.columns)`

Feature Engineering	 df.corr()['RainTomorrow'].sort_values(ascending= False) RainTomorrow RainTomorrow 1.000000 Humidity3pm 0.431272 Rainfall 0.321671 RainToday 0.304062 Cloud3pm 0.290055 Humidity9am 0.250375 Cloud9am 0.241909 WindGustSpeed 0.216257 WindSpeed9am 0.086720 MinTemp 0.083237 WindSpeed3pm 0.082588 WindGustDir 0.048793 WindDir9am 0.036326 WindDir3pm 0.029703 Month 0.007178 Day 0.005318 Location 0.001176 Year -0.009535 Temp9am -0.023780 Evaporation -0.098930 MaxTemp -0.154837 Temp3pm -0.186139 Pressure3pm -0.207057 Pressure9am -0.226512 Sunshine -0.288945 Selecting which are highly corelated to the target column and relatable.
Save Processed Data	-