Data Collection and Preprocessing Phase

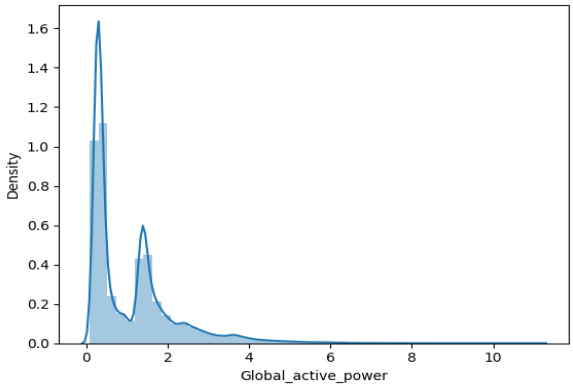
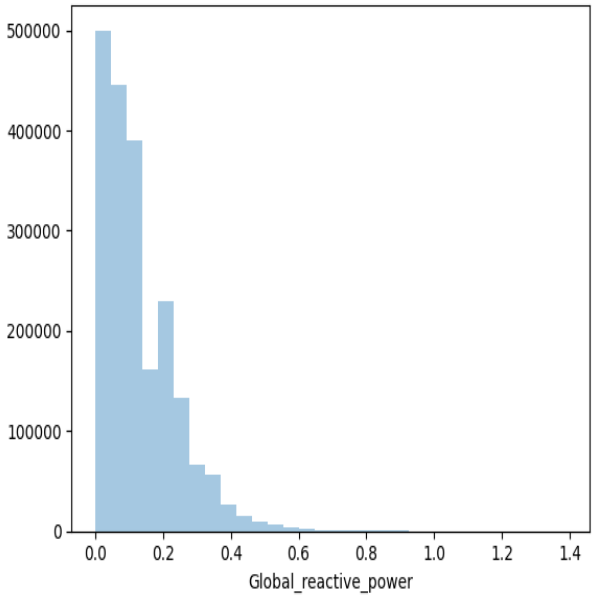
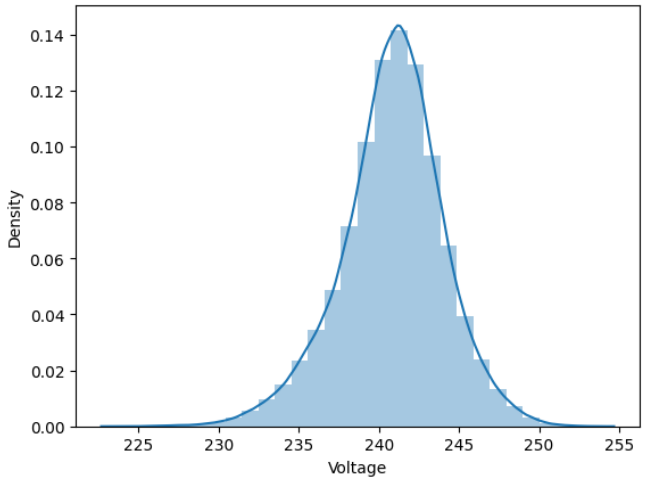
Date	31 June 2024
Team ID	740147
Project Title	Power Consumption Analysis for Households
Maximum Marks	6 Marks

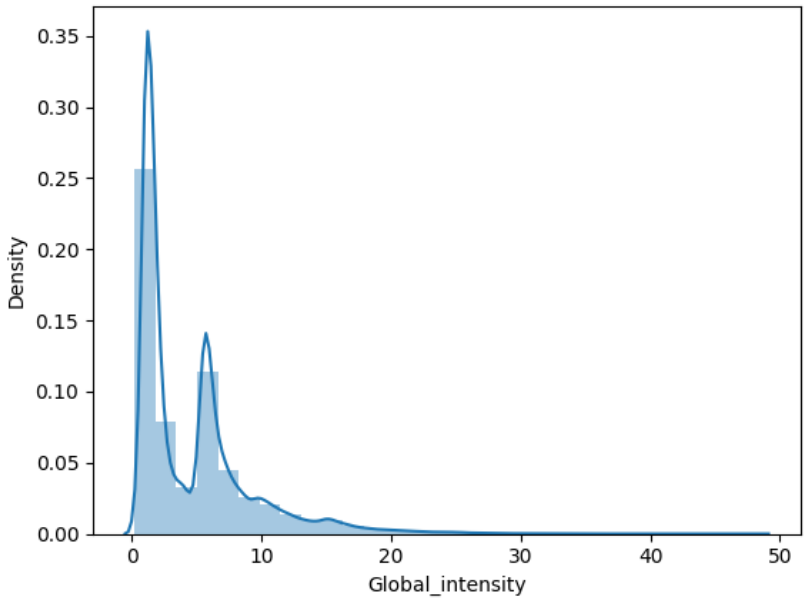
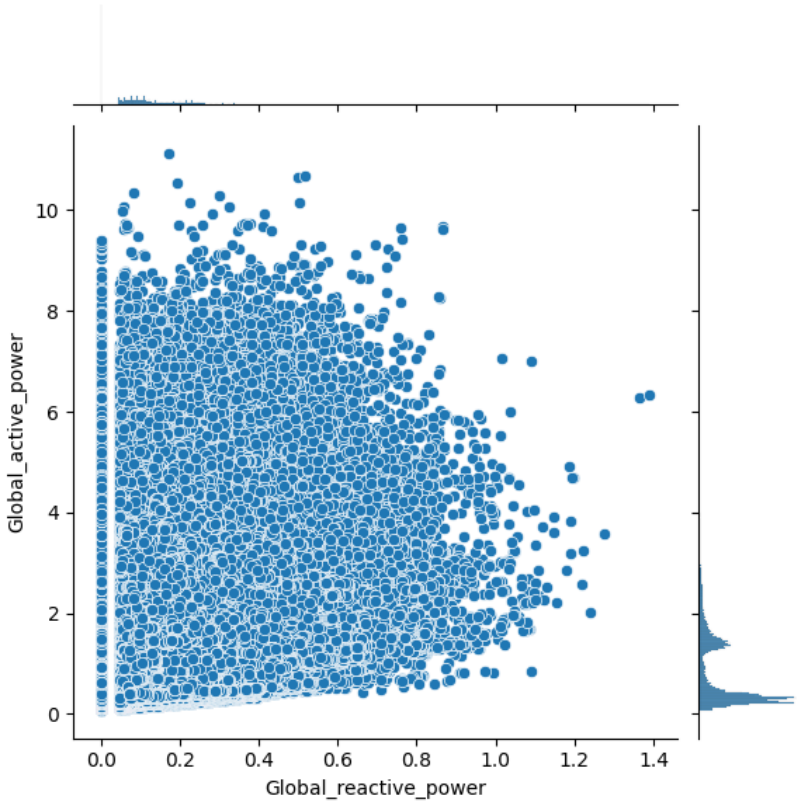
Data Exploration and Preprocessing

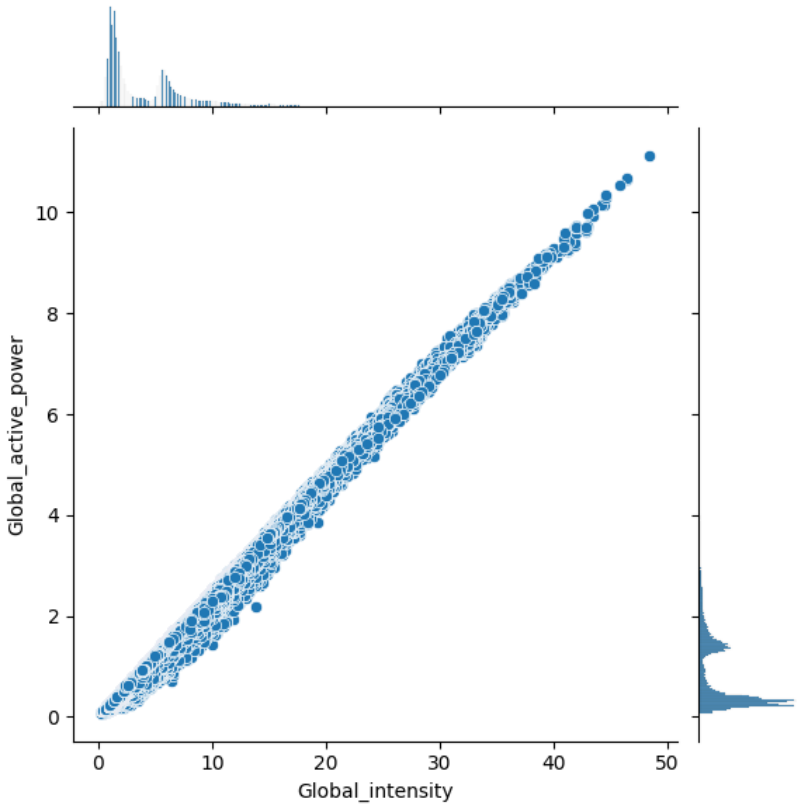
Dataset gets statistically analyzed to identify patterns and outliers. Data preprocessing addresses missing values, improving data quality for further analysis and modelling, and forming a strong foundation for insights and predictions.

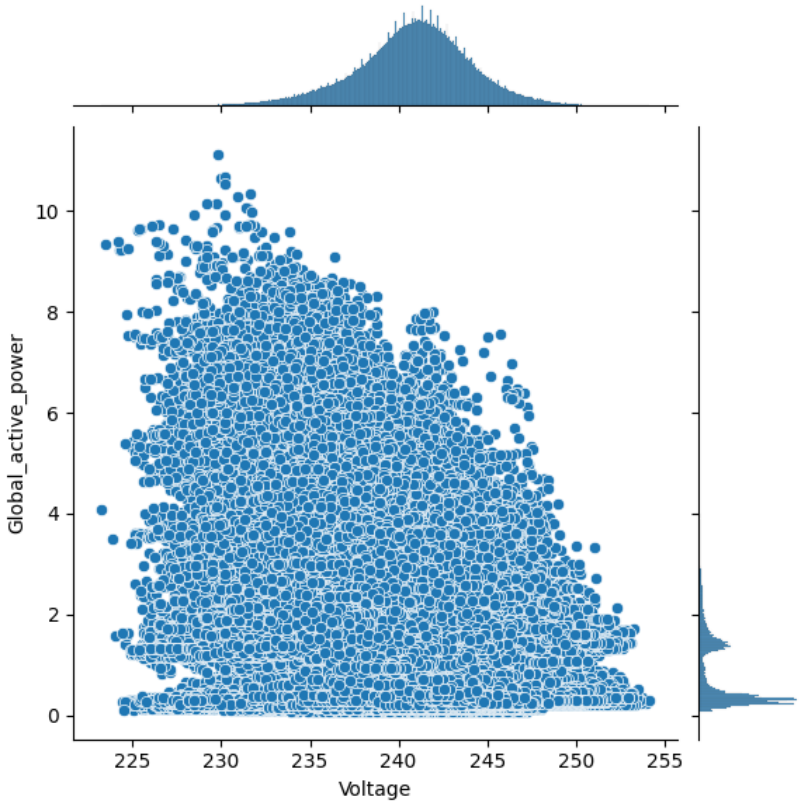
Section	Description
Data Overview	Dimensions: 2049280 x 7 Descriptive Statistics:

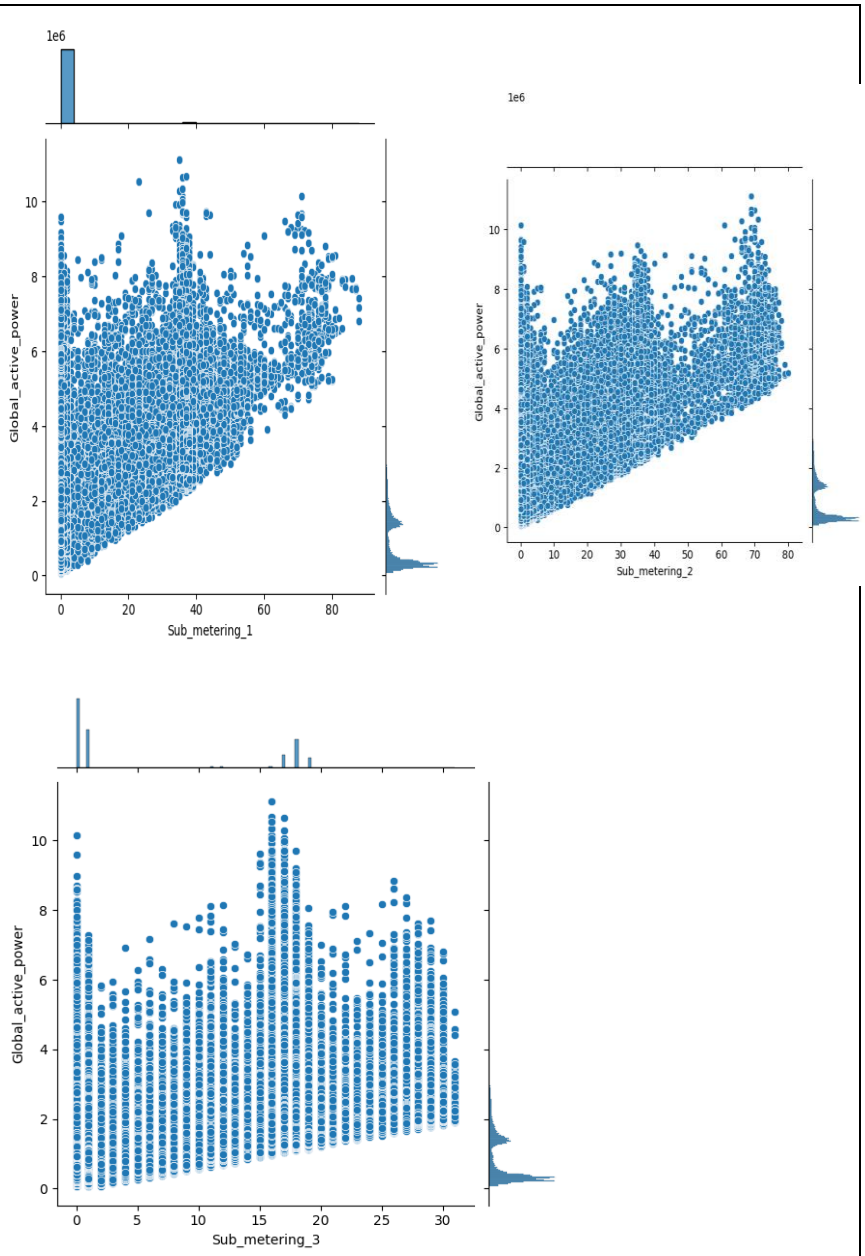
	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
count	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06
mean	1.091615e+00	1.237145e-01	2.408399e+02	4.627759e+00	1.121923e+00	1.298520e+00	6.458447e+00
std	1.057294e+00	1.127220e-01	3.239987e+00	4.444396e+00	6.153031e+00	5.822026e+00	8.437154e+00
min	7.600000e-02	0.000000e+00	2.232000e+02	2.000000e-01	0.000000e+00	0.000000e+00	0.000000e+00
25%	3.080000e-01	4.800000e-02	2.389900e+02	1.400000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	6.020000e-01	1.000000e-01	2.410100e+02	2.600000e+00	0.000000e+00	0.000000e+00	1.000000e+00
75%	1.528000e+00	1.940000e-01	2.428900e+02	6.400000e+00	0.000000e+00	1.000000e+00	1.700000e+01
max	1.112200e+01	1.390000e+00	2.541500e+02	4.840000e+01	8.800000e+01	8.000000e+01	3.100000e+01

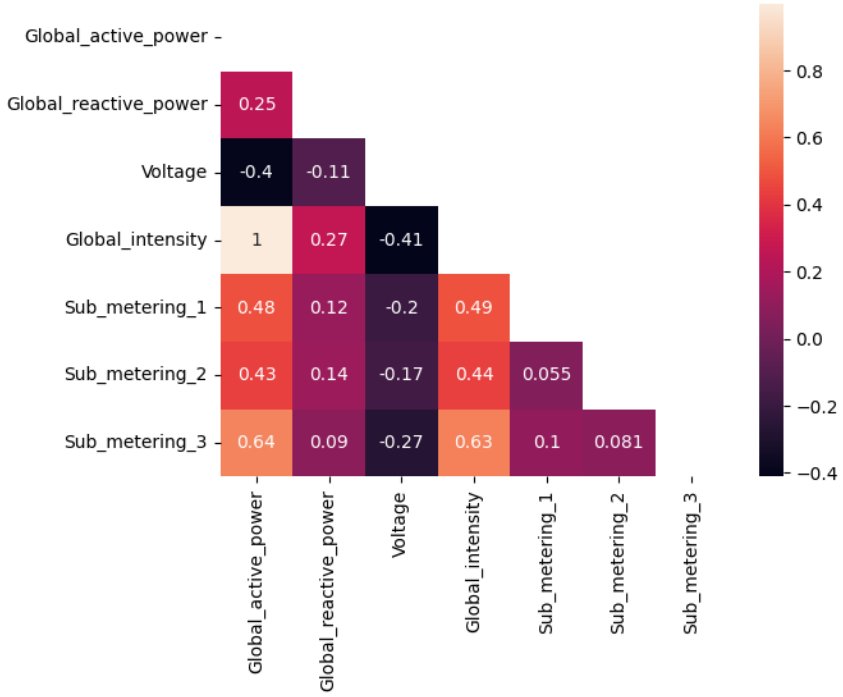
Univariate Analysis	  

	
Bivariate Analysis	


Global_active_power
Global_intensity
0
2
4
6
8
10
0
10
20
30
40
50


Global_active_power
Voltage
0
2
4
6
8
10
225
230
235
240
245
250
255


1e6
Global_active_power
Sub_metering_1
1e6
Global_active_power
Sub_metering_2
Global_active_power
Sub_metering_3

Multivariate Analysis	
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data	

```
#Loading the dataset
dataset = pd.read_csv("/content/drive/MyDrive/household_power_consumption.txt", sep=";", header=0, infer_datetime_format=Tru
dataset.head()
```

datetime	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Su
2006-12-16 17:24:00	4.216	0.418	234.840	18.400	0.000	1.000	
2006-12-16 17:25:00	5.360	0.436	233.630	23.000	0.000	1.000	
2006-12-16 17:26:00	5.374	0.498	233.290	23.000	0.000	2.000	
2006-12-16 17:27:00	5.388	0.502	233.740	23.000	0.000	1.000	
2006-12-16 17:28:00	3.666	0.528	235.680	15.800	0.000	1.000	

Handling Missing Data	

```
#checking for the null values
dataset.loc[dataset.Sub_metering_3.isnull()].head()
```

```
#replacing the null values
dataset.replace('?', np.nan, inplace=True)
```

```
#dropping the null values
dataset = dataset.dropna(how = 'all')
```

Data Transformation	`#changing the datatype of each column to float`
	`for i in dataset.columns:`
	`  dataset[i] = dataset[i].astype('float64')`

Feature Engineering	-
Save Processed Data	-