

KGISL INSTITUTE OF TECHNOLOGY

**NAAN MUDHALVAN**

**DIVISION:** APPLIED DATA SCIENCE

**PROJECT TITLE: ELECTRICITY PRICE PREDICTION**

**PHASE 3**

**PROBLEM STATEMENT:**

The problem is to develop a predictive model that uses historical electricity prices and relevant factors to forecast future electricity prices. The objective is to create a tool that assists both energy providers and consumers in making informed decisions regarding consumption and investment by predicting future electricity prices. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

**TEAM MEMBERS:**

Shreevarsha G

Shereena A

Nishanthi Akshayaa S

Guhapriyan K

**TEAM MENTOR:**

Indu Poornima R

**DEVELOPMENT PART – 1**

Importing Required Libraries

The code begins by importing several essential libraries:

PANDAS (as pd):

This library is used for data manipulation and analysis. It provides data structures and functions for working with structured data.

NUMPY (as np):

NumPy is used for numerical operations and arrays. It is often used for mathematical and numerical computations.

MATPLOTLIB.PYPLOT (as plt):

Matplotlib is a data visualization library, and pyplot is a sub-library that provides a convenient interface for creating various types of plots and charts.

SEABORN (as sns):

Seaborn is another data visualization library that enhances the aesthetics and visual appeal of data visualizations.

LOADING THE DATASETS:

The code loads a dataset from a CSV file named "MSFT.csv" into a Pandas DataFrame, which is essentially a structured table of data. This DataFrame is named df. The dataset likely contains historical Microsoft stock data, and it's important for the subsequent data analysis.

**UNDERSTANDING THE DATASETS:**

To better understand the data, the code performs the following operations:

**df.describe() (Electricity) :**

This function provides summary statistics for numerical columns in the DataFrame. It gives information such as the mean, standard deviation, minimum, maximum, and quartiles for each numeric attribute.

**df.info() (Electricity):**

This function provides information about the DataFrame, including the data types of each column (e.g., integer, float, string).

**df.isnull().sum() (Electricity):**

This code counts the number of missing values (NaN) in each column of the DataFrame. Identifying missing data is crucial for data cleaning and imputation.

**EXCESS:**

* Electricity.shape()
* Electricity.mean()
* Electricity.median()
* Electricity.mode()
* Electricity.std()
* Electricity.var()
* Electricity.skew()
* Electricity.describe(include=’all’)
* Electricity.kurt()
* Sum(Electricity.duplicated())

**VISUALIZING THE DATASET:**

The code proceeds to visualize the data to gain insights into its distribution and relationships between variables.

**HISTOGRAMS FOR NUMERICAL COLUMNS:**

The code creates histograms for a set of specified numerical columns. A histogram is a graphical representation of the distribution of data. It helps visualize how values are spread across the range of each attribute. Each histogram is displayed with 20 bins, and it uses a blue color with black edges for aesthetics. Titles, x-axis labels, and y-axis labels are set to provide context for each histogram.

**Histogram of the 'Close' Column (Target):**

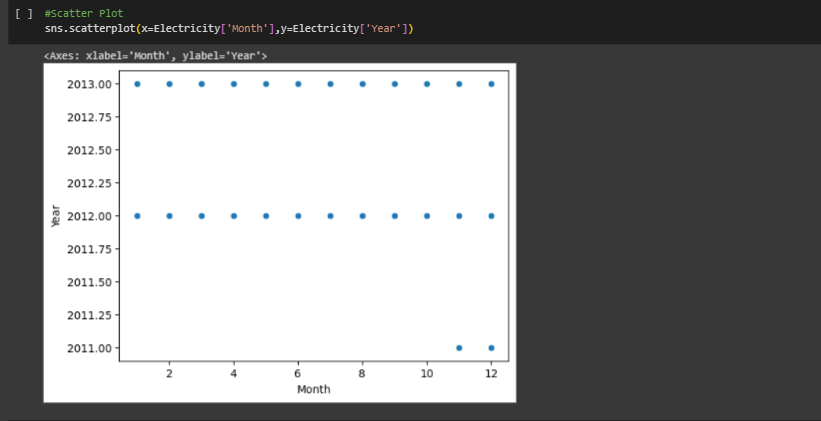
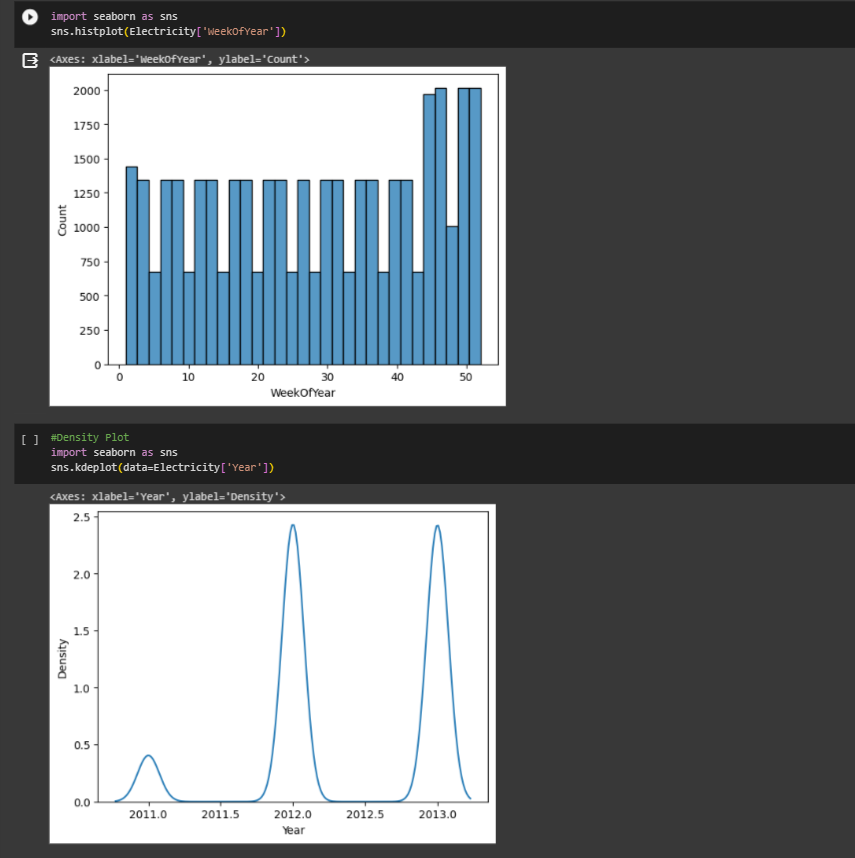
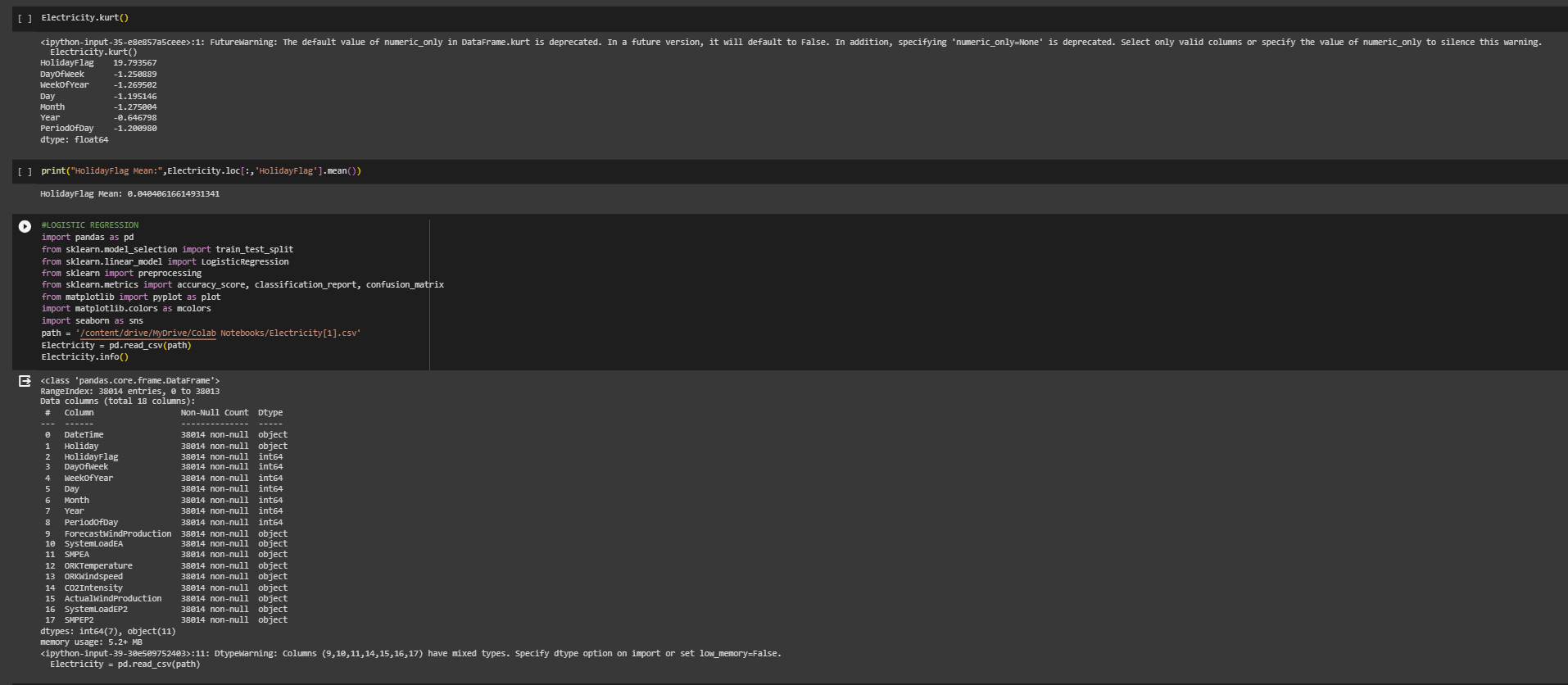
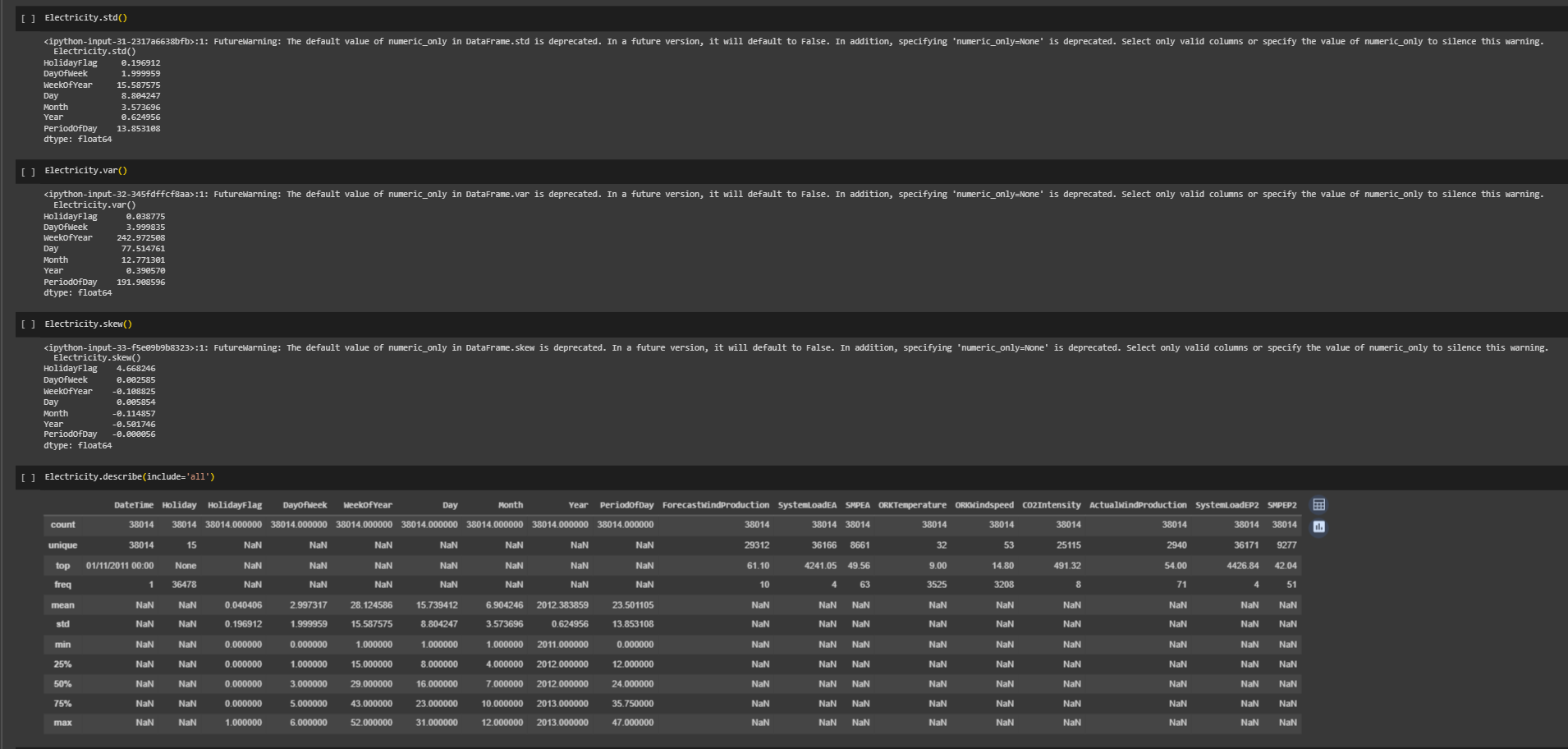
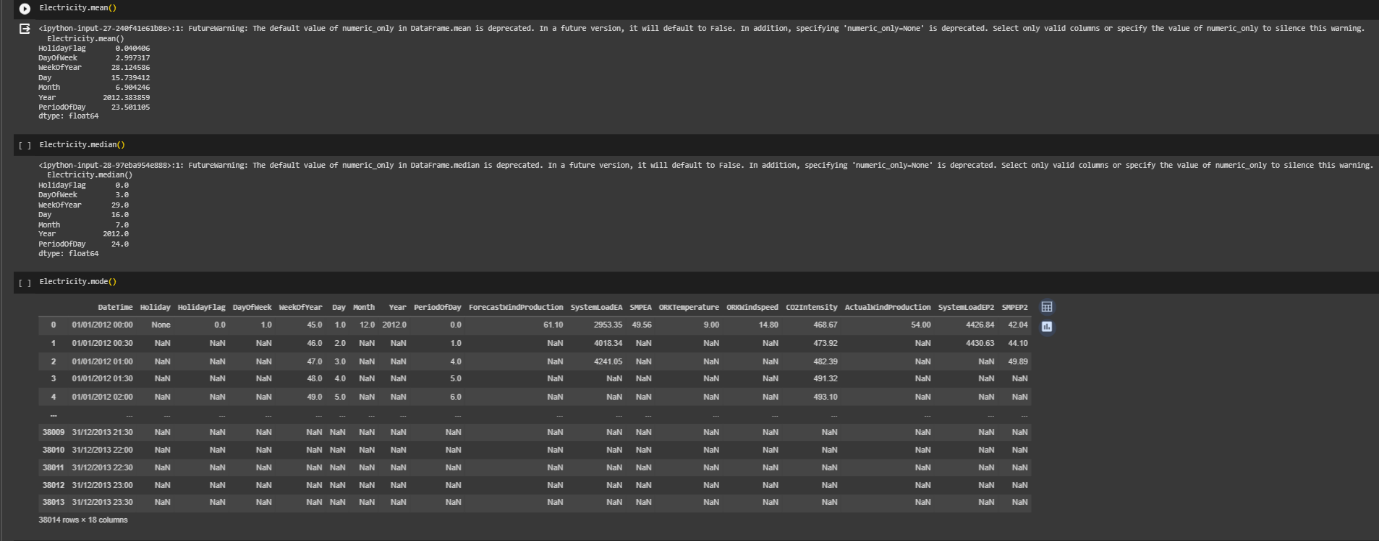
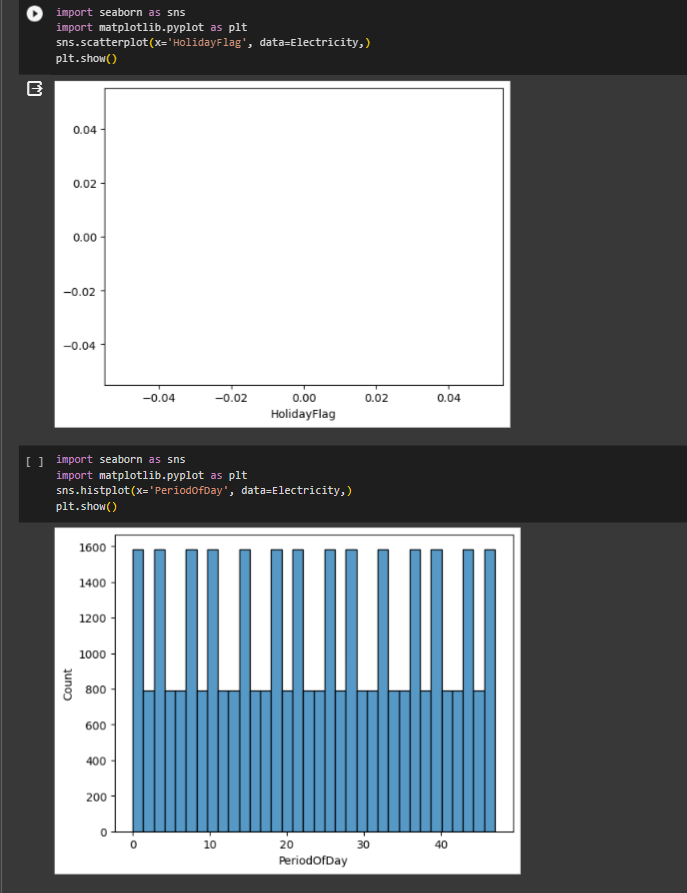
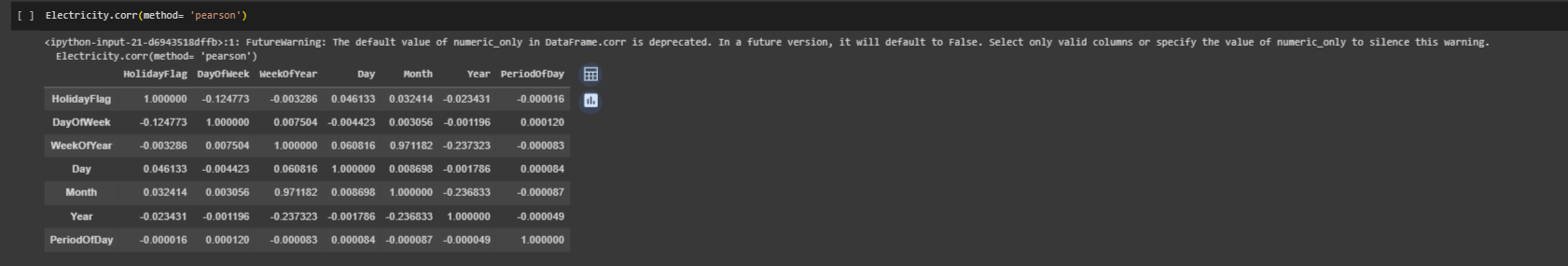
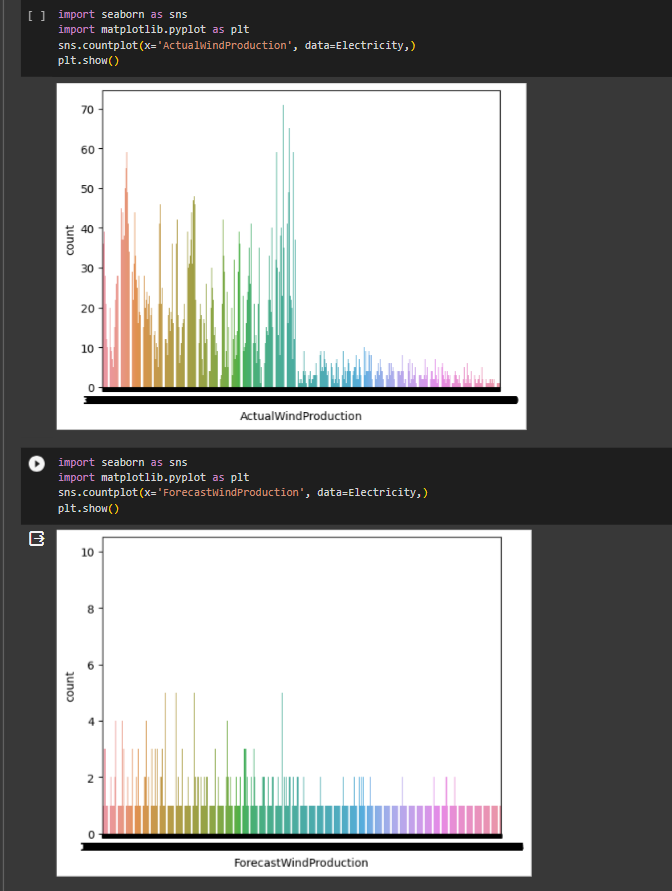
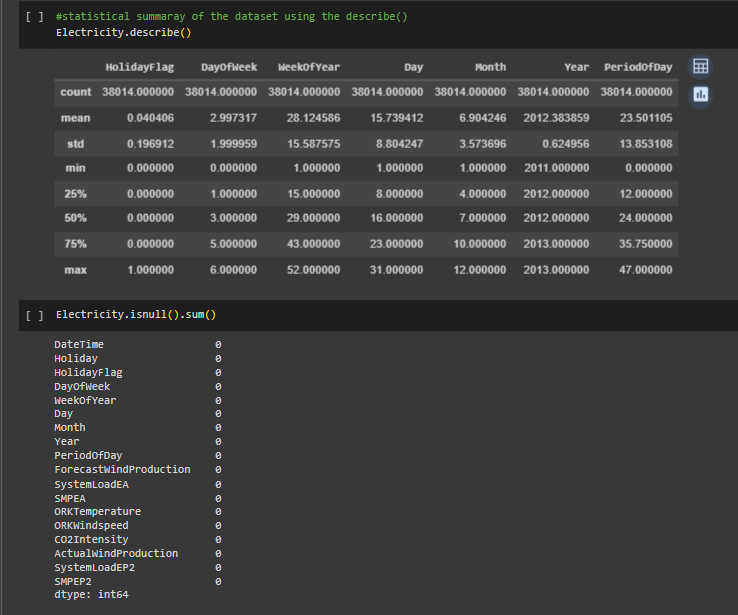
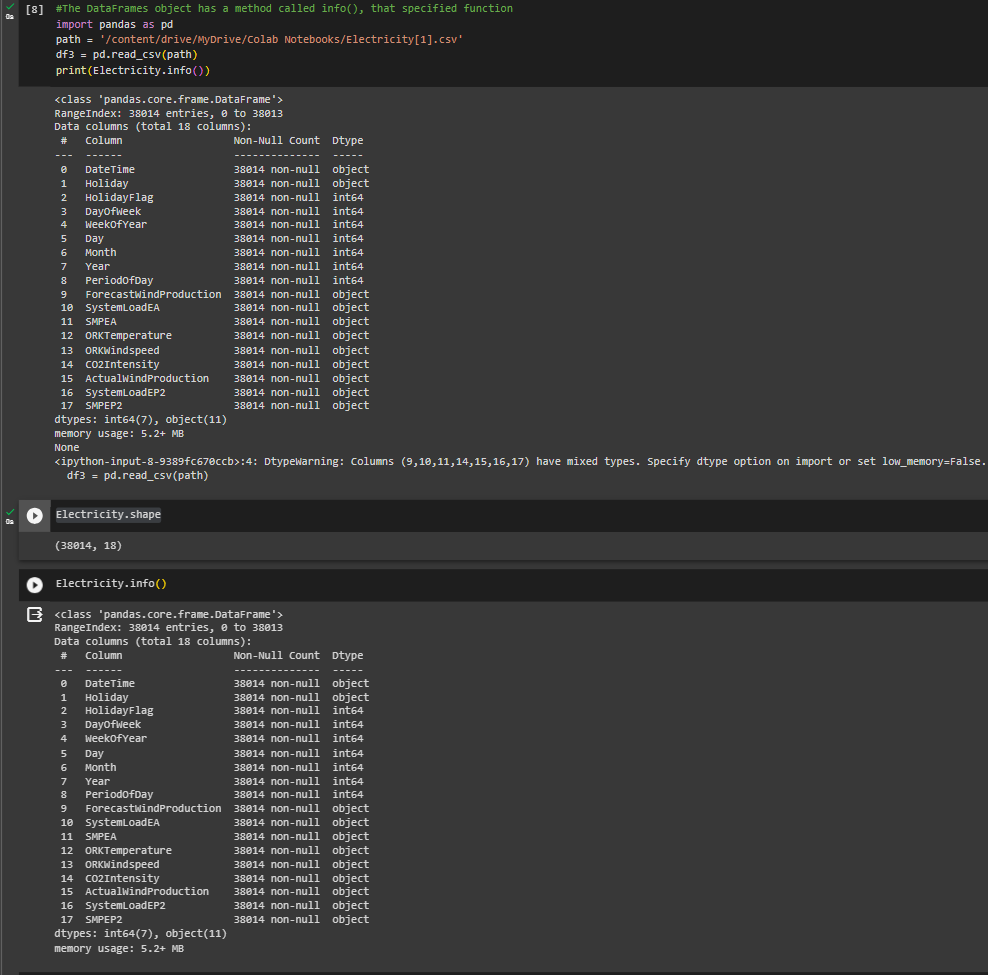
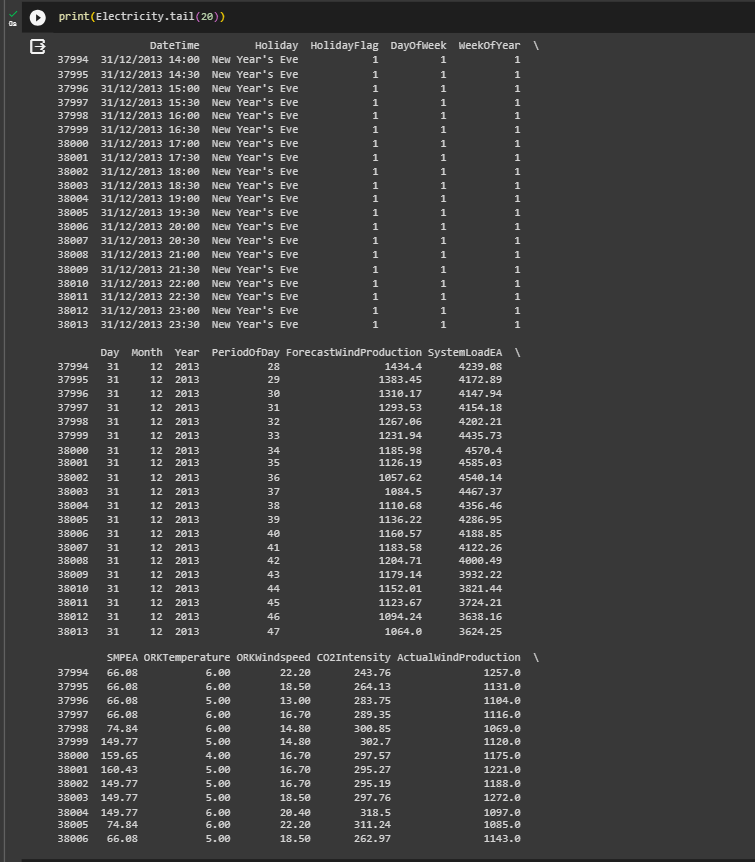
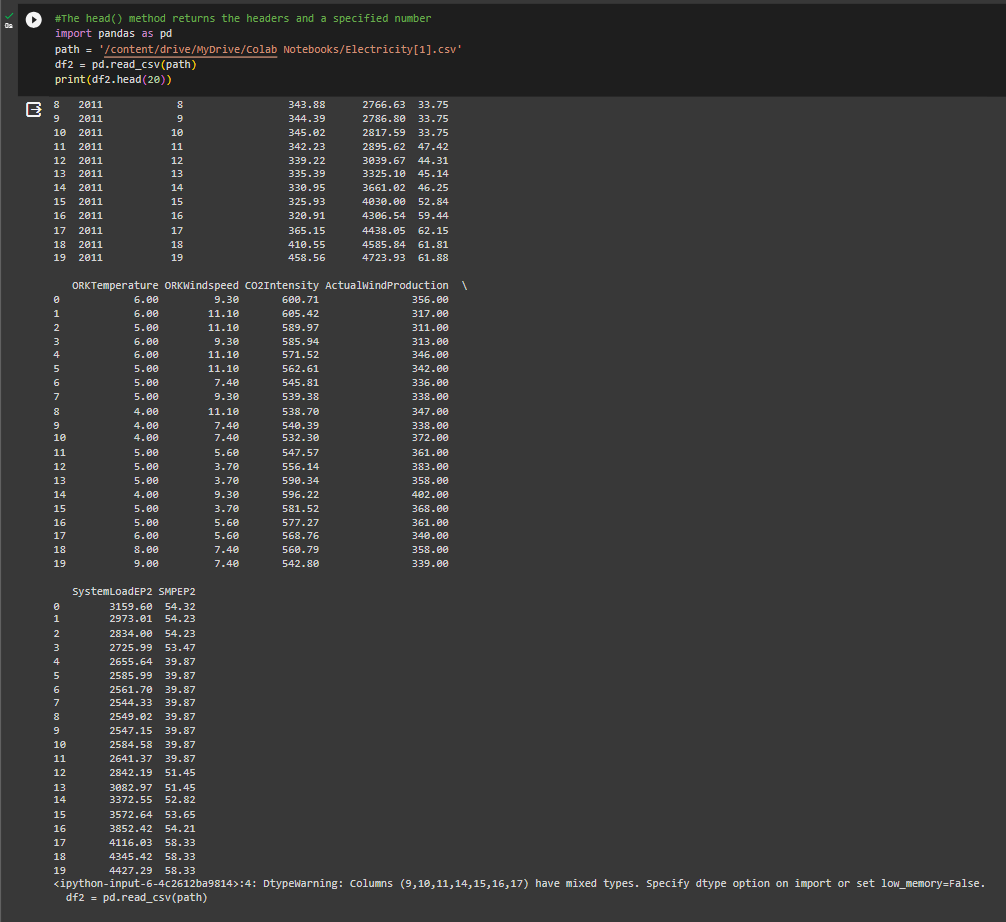
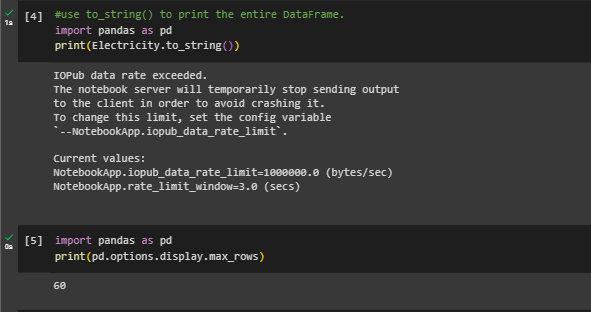
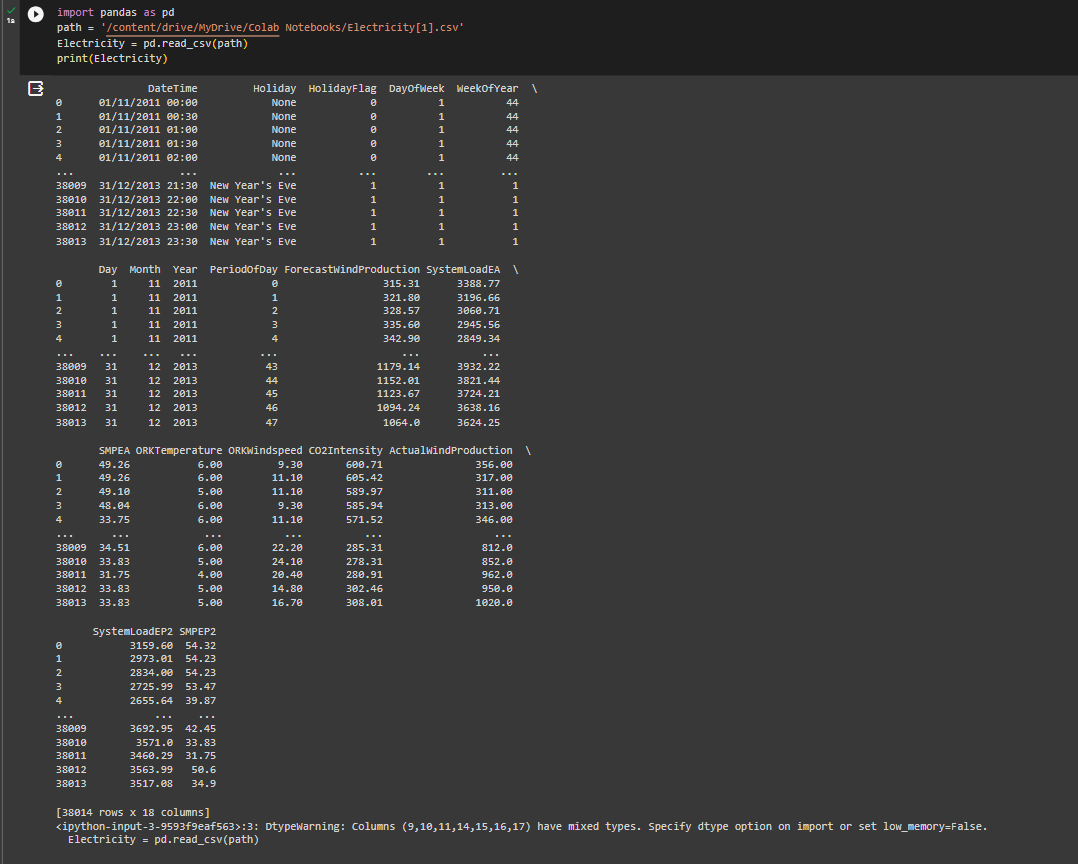
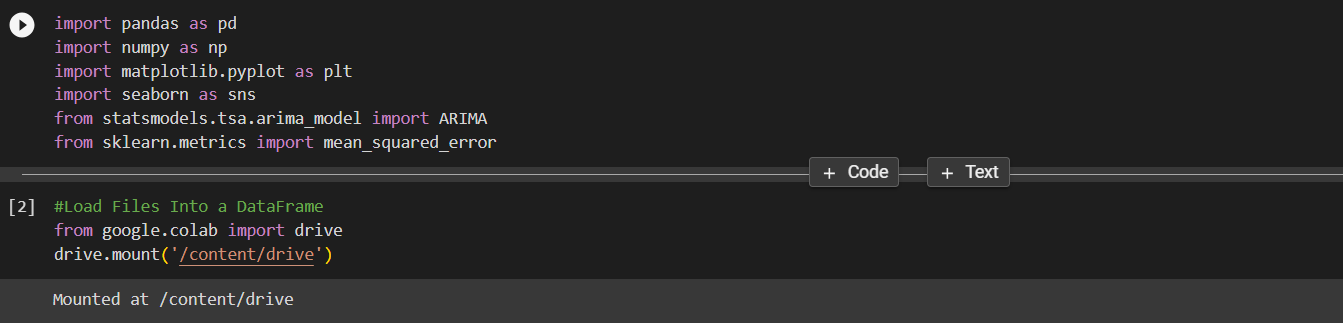
A separate histogram is created specifically for the 'Close' column, which is likely the target variable of interest. This histogram visualizes how the closing prices are distributed.

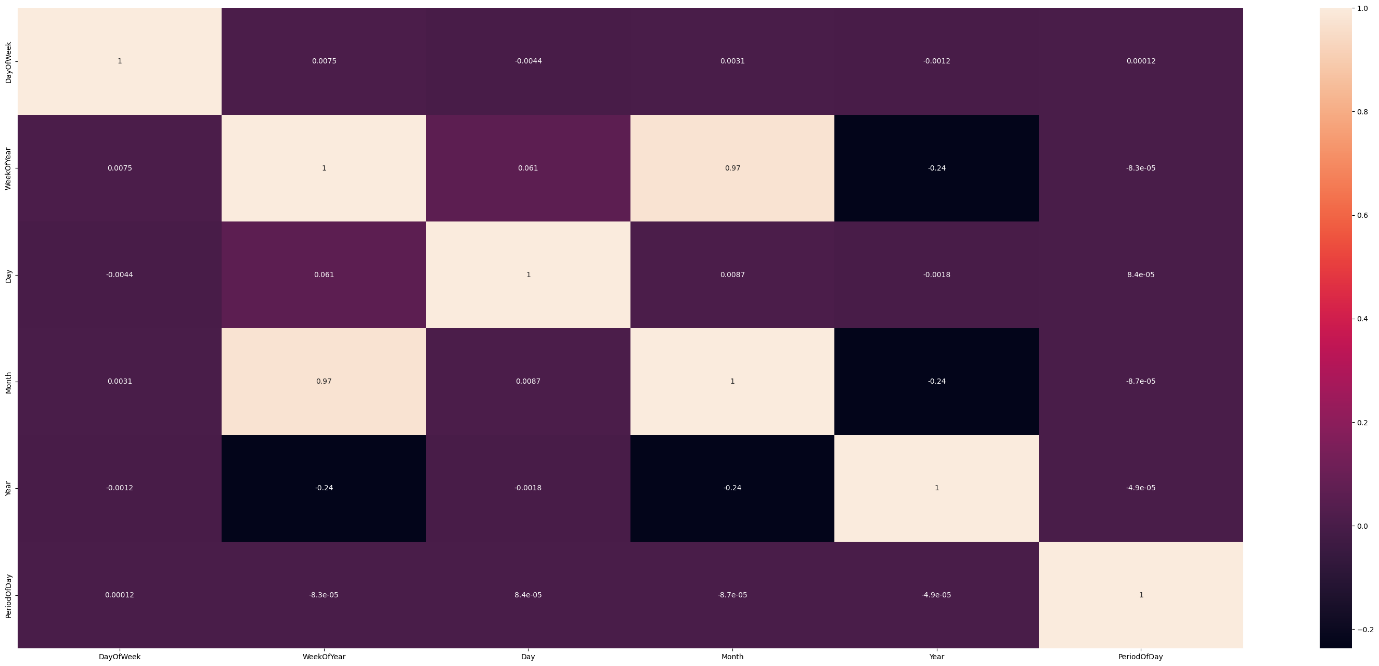
**Pairplot:**

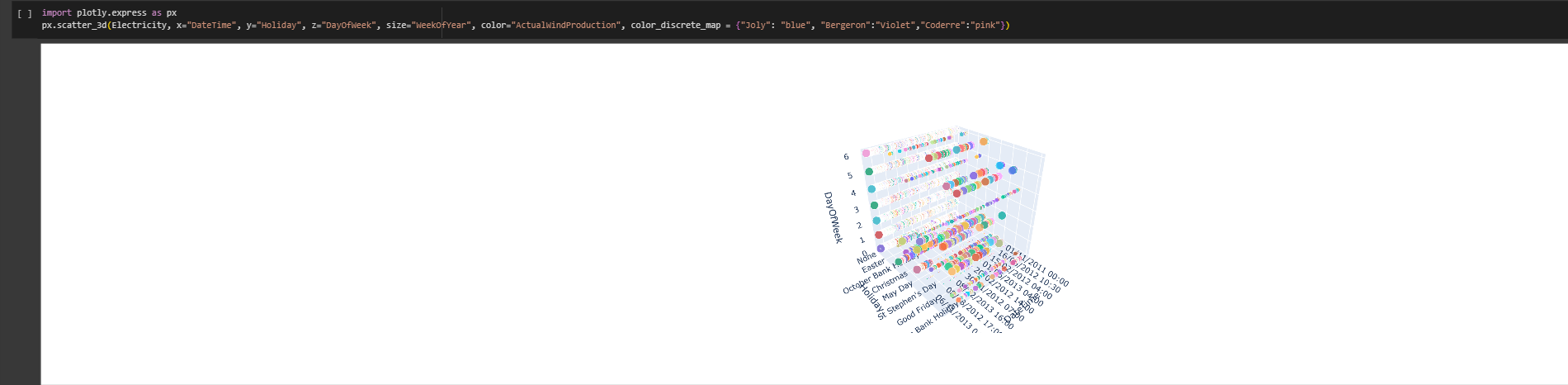
A pairplot is generated using Seaborn. This visualization creates scatterplots for combinations of numerical columns in the dataset. It helps identify potential relationships and correlations between different attributes. Pairplots are a useful tool for exploring multivariate data.

In conclusion, this code is a basic but essential data analysis pipeline for exploring a dataset containing Microsoft stock data. It includes data loading, summary statistics, and visualizations to gain initial insights and prepare the data for more in-depth analysis and modeling.

**MODEL:**

****

****

****

**COLAB LINK:**

**https://colab.research.google.com/drive/1931heJS5wtqcO8HQeeEdSRmoOFMss9Ze?usp=share\_link**