Data Collection and Preprocessing Phase

Date	15 july 2024
Team ID	740039
Project Title	Price prediction of natural gas using machine learning approach.
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	Basic statistics, dimensions, and structure of the data.
Univariate Analysis	Exploration of individual variables (missing values).
Bivariate Analysis	Relationships between two variables (boxplot, scatter plots).
Multivariate Analysis	Patterns and relationships involving multiple variables.

Outliers and Anomalies	Identification and treatment of outliers.

Data Preprocessing

Code Screenshots

Loading Data	loading the datset `[ ] data=pd.read_csv('/content/daily_csv.csv')`

Handling Missing Data	checking null values and filling missing values
	`[ ] data.isnull().any()`
	`Date False`
	`Price True`
	`dtype: bool`

Data Transformation	finding outliers
	IQR=q3-q1,upperbound=q3+1.5*IQR,lowerbound=q1-1.5*IQR
	`[ ] IQR=data['Price'].quantile(0.75)-data['Price'].quantile(0.25)`
	`upperbound=data['Price'].quantile(0.75)+1.5*IQR`
	`[ ] IQR`
	`2.58`
	`[ ] Lowerbound=data['Price'].quantile(0.25)-1.5*IQR`
	`Lowerbound`
	`-1.21`
	`[ ] upperbound=data['Price'].quantile(0.75)+1.5*IQR`
	`upperbound`
	`9.11`

Feature Engineering	Splitting the data
	`[ ] from sklearn.model_selection import train_test_split`
	`x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)`

Save Processed Data	training the model with decision tree
	`[ ] import numpy as np`
	`import pandas as pd # Use pandas as pd, not np`
	`import matplotlib.pyplot as plt`
	`import seaborn as sns`
	`from sklearn.model_selection import train_test_split`
	`from sklearn.tree import DecisionTreeRegressor`