Data Collection and Preprocessing Phase

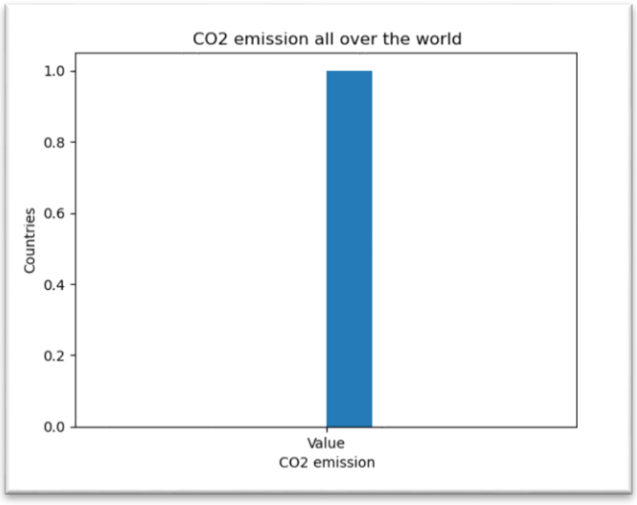
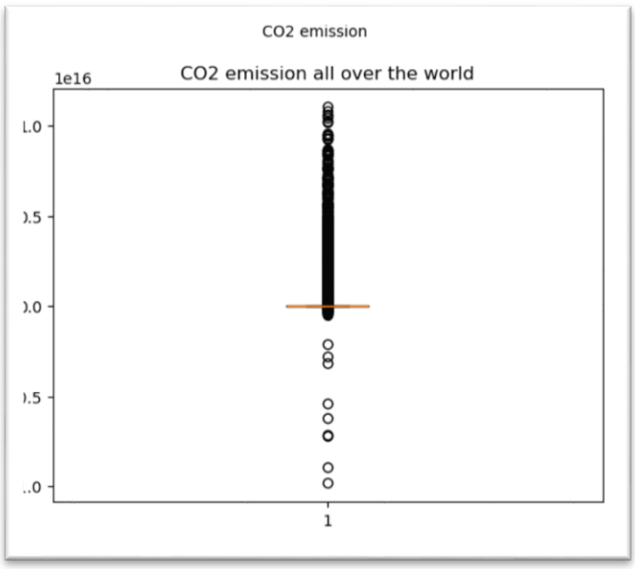
Date	6th July 2024
Team ID	SWTID1720000556
Project Title	Predicting Co2 Emission by Countries Using Machine Learning
Maximum Marks	6 Marks

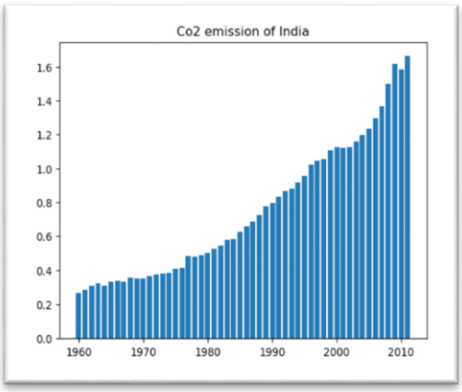
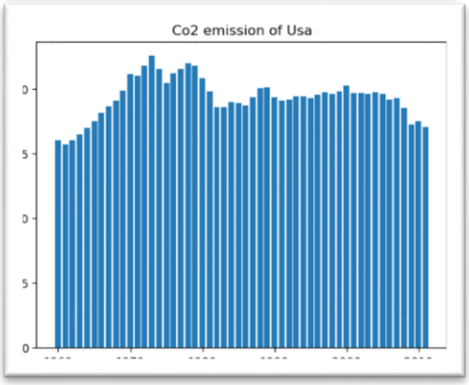
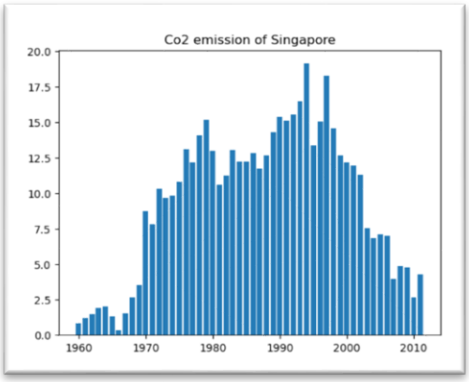
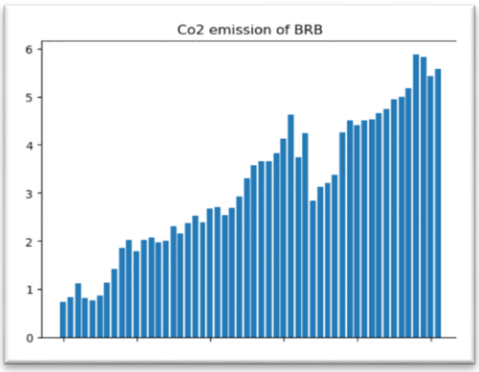
Data Exploration and Preprocessing Template

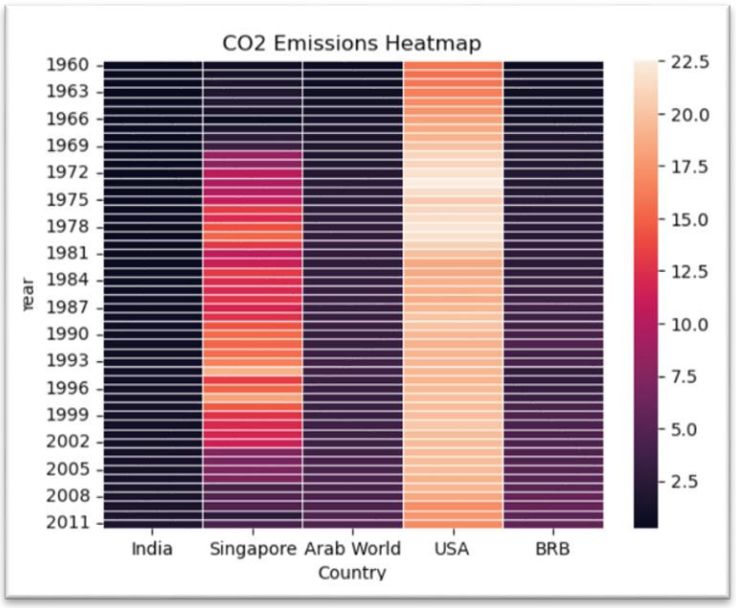
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	**Dimensions**: Rows - 5656458, columns - 6 **Basic structure of the data.**

	Year	Value
count	5.656458e+06	5.656458e+06
mean	1.994464e+03	1.070501e+12
std	1.387895e+01	4.842469e+13
min	1.960000e+03	-9.824821e+15
25%	1.984000e+03	5.566242e+00
50%	1.997000e+03	6.357450e+01
75%	2.006000e+03	1.346722e+07
max	2.015000e+03	1.103367e+16

Univariate Analysis	 

Bivariate Analysis	   

Multivariate Analysis	

Data Preprocessing Code Screenshots

Loading Data

```
data=pd.read_csv('indicators.csv')
```

```
data
```

	CountryName	CountryCode	IndicatorName	IndicatorCode	Year	Value
0	7	5	44	1195	1960	1.335609e+02
1	7	5	48	1218	1960	8.779760e+01
2	7	5	49	1219	1960	6.634579e+00
3	7	5	50	1220	1960	8.102333e+01
4	7	5	90	640	1960	3.000000e+06
...	...	...	...	...	...	...
5656453	246	246	1258	556	2015	3.600000e+01
5656454	246	246	1259	559	2015	9.000000e+01
5656455	246	246	1263	561	2015	2.420000e+02
5656456	246	246	1264	553	2015	3.300000e+00
5656457	246	246	1274	568	2015	3.280000e+01

5656458 rows × 6 columns

<table>
<tr>
<td>Handling Missing Data</td>
<td>

```
data.isnull().any()
```

```
CountryName      False
CountryCode      False
IndicatorName    False
IndicatorCode    False
Year             False
Value            False
dtype: bool
```

```
data['IndicatorName'].fillna(data['IndicatorName'].mode(),inplace=True)
data['CountryName'].fillna(data['CountryName'].mode(),inplace=True)
data['CountryCode'].fillna(data['CountryCode'].mode(),inplace=True)
```

```
if data['Value'].isnull().any().sum()!=0:
    data["Value"].fillna(data['Value'].mean(),inplace=True)
    print("Null values Removed")
else:
    print("No null Values")
```

```
No null Values
```

</td>
</tr>
<tr>
<td>Data Transformation</td>
<td>

```
from sklearn.preprocessing import LabelEncoder
data1 = data.copy()

for col in categorical:
    print("LABEL ENCODING OF:", col)
    LE = LabelEncoder()
    data[col] = LE.fit_transform(data[col])
    print(col,"is Encoded")
```

```
LABEL ENCODING OF: CountryName
CountryName is Encoded
LABEL ENCODING OF: CountryCode
CountryCode is Encoded
LABEL ENCODING OF: IndicatorName
IndicatorName is Encoded
LABEL ENCODING OF: IndicatorCode
IndicatorCode is Encoded
```

</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>Attached codes in the final folder</td>
</tr>
<tr>
<td>Save Processed Data</td>
<td>

```
le1=LabelEncoder()
le1.fit_transform(countriesNameColumn)
pickle.dump(le1,open("CountryName","wb"))
```

```
le2=LabelEncoder()
le2.fit_transform(countriesCodeColumn)
pickle.dump(le2,open("CountryCode","wb"))
```

```
le3=LabelEncoder()
le3.fit_transform(data1['IndicatorName'])
pickle.dump(le3,open("IndicatorName","wb"))
```

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