**Exploratory Data Analysis (EDA) and Pre-processing Report of Carbon Emission Prediction and Comparative Analysis.**

**Introduction**

This document presents an exploratory data analysis (EDA) and preprocessing overview for the carbon emissions dataset. The aim is to provide insights into the data, identify patterns, and prepare the data for further analysis or modeling.

**Dataset Description**

The dataset contains carbon emissions data categorized by state, sector, fuel type, and year. Key variables include:

* **year**: Year of emission observation (1970 – 2021)
* **state-name**: 53 states in America
* **sector-name**: Sector responsible for emissions (6 sectors)
* **fuel-name**: Type of fuel (4 types)
* **value**: Carbon emission value

**1. Data Loading and Overview**

The dataset initially consisted of [59901] entries and [5] columns. After preprocessing, it now contains [10367] entries and [5] columns.

**Summary:**

* **Total Records**: 10367
* **Data Columns**:
  + year: int64
  + state-name: object
  + sector-name: object
  + fuel-name: object
  + value: float64
* **Null Values**: No null values present.
* **Duplicates**: No duplicate records found.

**2. AutoEDA with Sweetviz**

Sweetviz library is used to generate automated exploratory data analysis reports, providing insights into the dataset's characteristics, distributions, and relationships.

**3. Descriptive Statistics**

Descriptive statistics are calculated for the emissions data, including mean, median, mode, standard deviation, variance, range, skewness, and kurtosis.

**4. Outlier Detection and Treatment**

Outliers are detected using boxplot visualization and treated using Winsorization method to ensure data integrity.

**5. Time Series Analysis**

**5.1 Shapiro-Wilk Test for Normality**

Shapiro-Wilk tests are conducted to assess the normality of emission data. Results indicate deviations from normality.

**5.2 Random Walk Analysis**

Augmented Dickey-Fuller (ADF) and KPSS tests are performed to assess stationarity and presence of random walk behaviour in time series data. Time series are categorized as either stationary or exhibiting random walk behaviour.

**6. Data Preprocessing**

**6.1 Filtering Non-Stationary Data**

Non-stationary time series are filtered out based on ADF and KPSS test results.

**6.2 Saving Preprocessed Data**

Preprocessed data, including stationary and non-stationary time series, are saved into separate CSV files for further analysis.

**Conclusion**

The EDA and preprocessing steps provide valuable insights into the carbon emissions dataset, facilitating subsequent analysis or modeling tasks.

This document summarizes the EDA and preprocessing steps performed on the carbon emissions dataset. It includes loading the data, generating automated exploratory analysis, calculating descriptive statistics, detecting and treating outliers, conducting time series analysis, and preprocessing the data for further analysis.