CHAPTER 4

EXPLORATORY DATA ANALYSIS (EDA) / INITIAL FINDING

4.0 Introduction

The Exploratory Data Analysis (EDA) phase is vital for understanding the dataset, identifying patterns, and gaining insight into the subsequent modelling phase. This chapter will outline the expected outcomes from the EDA process, including visualization, descriptive statistics, initial findings, and future engineering.

4.1 Visualizations

Visualization is one of the most potent tools for uncovering patterns and anomalies within the data. The following explanation of the visualization is expected to be produced during the EDA process:

1) **Time Series Plots**
   - **Purpose:** Time series plots are fundamental to understanding the temporal dynamics of energy consumption. They help visualize how energy consumption has changed, identifying trends and seasonality.
   - **Implementation**: The historical energy consumption data such as the amount of electricity consumed in Malaysia, total electricity consumption, energy demand by sector in Malaysia, household electricity per capita in Malaysia, and urbanization in Malaysia based on year and primary energy supply in Malaysia. Separate plots for different sectors will be created to identify the consumption patterns. For instance, a time series for energy demands by sector in Malaysia may reveal increases in usage during a specific time. Similarly, the household per capita plots may show the correlations between energy used within the year.

2) **Correlation Heatmaps:**

- **Purpose:** Correlation heatmaps visualize the strength and direction of the relationships between multiple variables. They are instrumental in identifying which variables are vital and correlated with energy consumption.
- **Implementation**: A heatmap will display the correlation between energy consumption in Malaysia and various economic, demographic, and climatic variables. For example, a strong positive correlation between GDP and energy consumption would indicate that economic growth drives higher energy usage, and population growth leads to higher energy consumption. The negative correlation with temperature might suggest that energy consumption will decrease during milder weather conditions**.** This insight will help to select relevant features for the regression model that will be implemented later.

3) **Scatter Plots**:
   - **Purpose**: Scatter plots effectively explore the relationship between two variables and identify potential linear and non-linear relationships.
   - **Implementation**: This will examine the relationship between energy consumption and key predictors such as GDP, population growth, temperature, and humidity. For example, plotting energy consumption in Malaysia against GDP can reveal whether there is a linear relationship (indicating GDP increases so with the energy consumption) or to see if there are non-linear patterns. This help to understand the relationship and determine whether polynomial or interaction terms are needed in the regression models.

4) **Box Plots**:
   - **Purpose**: The box plot provides a visual summary of the dataset's distribution, which helps identify outliers and compare distributions across different groups.
   - **Implementation**: The box plot will visualize the distribution of variables such as energy consumption, GDP, economic factors, and climate variables. It will help to identify the outliers and understand the spread and central tendency of the data. For example, the box plot energy consumption for various factors such as residential, industrial, and commercial can highlight variations and outliers within each sector to focus on specific analysis and feature engineering.

5) **Histograms**:

- **Purpose**: Histograms visualize the distribution of a single variable, showing the frequency of different values and helping to identify patterns such as skewness or kurtosis.
- **Implementation**: Histograms will assess the distribution of individual variables, such as energy consumption, GDP, and temperature. For instance, a histogram of energy consumption might reveal a right-skewed distribution, indicating that a few periods have exceptionally high consumption. This insight may prompt transformations to normalize the data. Similarly, GDP, temperature and humidity histograms will help understand their distributions and any potential need for scaling or transformation.

4.2 Descriptive Statistics

Descriptive analysis will summarize the dataset distribution's central tendency, dispersion, and shape, which are crucial to understanding the data and informing further analysis. The following statistics are expected to be calculated in Table 4.1 below:

Table 4.1 Measures of Descriptive Analysis

Measures	**Description**
Central Tendency	1) **Mean:** The average energy consumption, GDP, economic factors, and climate variables provide a central point around which data points are distributed. 2) **Median**: The middle value of the average energy consumption, GDP, economic factors, and climate variables dataset, offering a robust measure of central tendency that is less affected by outliers. 3) **Mode:** The mode is the most frequently occurring value, which helps identify common patterns or behavior in the data.

Dispersion	1) Range, interquartile range (IQR), variance, and standard deviation of energy consumption, GDP, economic factors, and climate variables. 2) Assess variability within energy consumption data and other predictors.
Statistics	1) Minimum, maximum, and quartile values for energy consumption and predictors. 2) Identify any significant deviations or anomalies in the data.

4.3 Initial Insights

Initial insights from EDA are expected to provide a deeper understanding of the data and guide the feature engineering and modelling processes. Key insights may include:

1. **Trends and Seasonality**:
    - Identification of long-term trends in various energy consumption in Malaysia, such as the pattern of increases or decreases.
2. **Relationships Between Variables**:
    - Strong correlations between energy consumption in Malaysia and economic factors such as GDP, industrial and residential.
    - The impact of climatic factors such as temperature and humidity on energy usage in Malaysia.
3. **Sector-Specific Insights**:
    - Differences in energy consumption patterns across residential, industrial, and commercial sectors.
    - Identification of high-consumption periods and potential causes.

4.4 Feature Engineering

Feature engineering involves creating new variables that capture additional information or enhance the predictive power of the models. The Expected feature engineering steps include:

1. **Lagged Variables**:
    - Create lagged versions of energy consumption and critical predictors to capture temporal dependencies.
    - Test different lag periods to find the most predictive lags.
2. **Interaction Terms**:
    - Generate interaction terms between variables to capture combined effects such as GDP growth and energy consumption.
3. **Polynomial Features**:
    - Create polynomial features for variables with non-linear relationships to energy consumption.
4. **Rolling Statistics**:
    - Calculate rolling means, variances, and other statistics to smooth out short-term fluctuations and highlight longer-term trends.