CHAPTER 4

EDA / INITIAL FINDINGS

In this chapter, we provide the first insights into the EDA of the datasets which is an essential first step for any data analysis. It is widely used in order to define the overall structure of the data, as well as possible and probable patterns and outliers, which in turn can be used to make hypotheses in order to guide further analysis. The EDA will entail computing the measures of central tendency and variability and making graphs of the data and any other preliminary findings that may be of help in the data analysis.

4.1 Descriptive Statistics

The first step of EDA will involve the calculation of summary statistics for the important variables in the datasets. This involves computation of measures of central tendency such as the mean and median as well as dispersion such as standard deviation variance, the range, minimum as well as the maximum values.

Expected Outcomes:

Summary Statistics: The descriptive statistics will therefore present basic information about the data by providing measures of central tendency, variability, and range in the context of the average unemployment rate and GDP.

4.2 Data Visualization

Descriptive visualisations will be employed to analyse the distribution of the data and the presence of patterns, as well as to detect outliers and anomalies. The following types of visualisations will be generated:

1. **Histograms**: Histograms are used to show the percent of GDP per capita and unemployment rate.

Expected Outcomes:

- **Unemployment Rate Histogram**: From the histogram, a pattern of the unemployment rate over the years will be displayed showing if it has a bell shaped curve or whether it is skewed.
- **GDP Histogram**: This will bring out the bar graph of the GDP values demonstrating when there was a rise or drop in the economy.

2. **Box Plots**: In order to perform robust regression, where the goal is to estimate the parameters of the linear model that are least sensitive to the outliers.

   Expected Outcomes:

   - **Unemployment Rate Box Plot**: This will be useful in tapping on the data in order to detect if there are any outliers or any very low or very high value of the unemployment rate.
   - **GDP Box Plot**: Similarly, it will display outliers in the raw GDP data and give an impression of the data distribution in the set.

3. **Time Series Plots**: To use trends, and patterns over time, to compare and contrast GDP and unemployment rates.

   Expected Outcomes:

   - **Unemployment Rate Time Series**: This type of graph shall help to identify its trend, possible cyclical behaviour, or other changes in the level of unemployment throughout the research period.
   - **GDP Time Series**: This will depict the growth of GDP over time; whether the economy is growing or shrinking.

4.3 Correlation Analysis

Inferential statistics shall be used, where correlation analysis shall be conducted in an effort to establish the correlation between the GDP and unemployment rates. In order to determine the nature and extent of the linear relationship between the variables, Pearson or Spearman correlation coefficient will be computed.

Expected Outcomes:

Correlation Matrix: The correlation analysis will give a quantitative estimate of the relationship between GDP and the variable of unemployment rates.

4.4 Initial Insights and Hypotheses

From the descriptive statistics, visualisations and correlation analysis, initial insights and hypotheses will be derived. These will be used in the subsequent steps of the analysis and model building process.

Expected Outcomes:

- **Trends and Patterns**: Estimation of important variables in the data together with periods of high and low unemployment and the corresponding GDP values.
- **Anomalies and Outliers**: Any anomalies or outliers that are seen during this process needs to be investigated further.
- **Hypotheses**: Creation of hypotheses in regards to the connection between GDP and rates of unemployment such as “higher GDP implies low unemployment rates.”

Conclusion

This chapter presents the data analysis results that help to identify the first conclusions and formulate hypotheses. This information will be then used in the analysis and model building phases of the project as the EDA provides a brief summary of the data and its patterns and relationships. The expected outcomes include descriptive analysis which helps to understand the distribution of the data, detect trends and outliers, and formulate preliminary hypotheses of the relationship between GDP and unemployment rates in Malaysia.