CHAPTER 3

METHODOLOGY

This chapter outlines the methodologies that will be employed in the analysis of GDP and unemployment trends in Malaysia. The chapter is structured to detail the Data Science Project Life Cycle, data sources and collection methods, data pre-processing steps, and the analytical techniques to be used for this study.

3.1 Data Science Project Life Cycle

The Data Science Project Life Cycle encompasses the following stages: Data Collection, Data Pre-processing, Exploratory Data Analysis (EDA), Model Building, Evaluation, and Deployment. Each stage is important in guaranteeing that the data is accurately analysed and that the results are reliable and actionable.

1. **Data Collection**: The initial stage involves gathering all relevant data required for analysis. This includes sourcing data from website data.gov.my.
2. **Data Pre-processing**: This stage involves cleaning and transforming the raw data into a usable format. It includes handling missing values, normalising data, and creating new features.
3. **Exploratory Data Analysis (EDA)**: EDA is the process of describing the data in a simplified manner using graphical techniques. It assists in comprehending the patterns, identifying outliers, and developing hypotheses.
4. **Model Building**: In this stage, statistical models and machine learning algorithms will be developed to analyse the data and generate predictions.
5. **Evaluation**: The models will be assessed using different performance measures to confirm that the results are accurate and relevant.

6. **Deployment**: The last stage involves deploying and leveraging the models in a live setting environment where they can be resourceful for actionable insights.

3.2 Data Sources and Collection Methods

This study will utilise a comprehensive set of datasets from Data.gov.my and other relevant sources. The primary datasets include:

1. **Monthly Unemployment Data** (lfs_month.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides monthly data on overall unemployment rates in Malaysia.
    - **Columns:**
        i. date: The date of the record.
        ii. lf: The number (in thousands) of employed and unemployed individuals. This figure also represents the number of people participating in the labour force.
        iii. lf_employed: The number (in thousands) of people who worked at least one hour for pay, profit or family gain, in thousands of people
        iv. lf_unemployed: The number (in thousands) of people who did not work but were looking for work or available to work
        v. lf_outside: The number (in thousands) of people not classified as employed or unemployed, including housewives, students, early retired, disabled persons and those not interested in looking for a job.
        vi. p_rate: Ratio of the labour force size to the working-age (15-64) population
        vii. ep_ratio: Ratio of the number of employed people to the working-age (15-64) population
        viii. u_rate: Ratio of unemployed to labour force size
2. **Seasonally Adjusted Monthly Unemployment Data** (lfs_month_sa.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides seasonally adjusted unemployment data to remove seasonal effects.

- **Columns:**
    i. date: The date of the record.
    ii. lf: Seasonally adjusted number (in thousands) of employed and unemployed individuals. This figure also represents the number of people participating in the labour force.
    iii. lf_employed: Seasonally adjusted number (in thousands) of people who worked at least one hour for pay, profit or family gain, in thousands of people
    iv. lf_unemployed: Seasonally adjusted number (in thousands) of people who did not work but were looking for work or available to work
    v. p_rate: Seasonally adjusted ratio of the labour force size to the working-age (15-64) population
    vi. u_rate: Seasonally adjusted ratio of unemployed to labour force size

3. **Monthly Youth Unemployment Data** (lfs_month_youth.csv)
    - **Source**: Data.gov.my
    - **Description**: Focuses on unemployment rates among the youth.
    - **Columns:**
        i. date: The date of the record.
        ii. unemployed_15_24: Number (in thousands) of people aged 15-24 who did not work but were looking for work or available to work
        iii. u_rate_15_24: Ratio of unemployed aged 15-24 to labour force participants aged 15-24
        iv. unemployed_15_30: Number (in thousands) of people aged 15-30 who did not work but were looking for work or available to work
        v. u_rate_15_30: Ratio of unemployed aged 15-30 to labour force participants aged 15-30
4. **Monthly Unemployment Duration Data** (lfs_month_duration.csv)
    - **Source**: Data.gov.my
    - **Description**: Contains data on the duration of unemployment periods.

- **Columns:**
    i. date: The date of the record.
    ii. unemployed: The number (in thousands) of people who did not work but were looking for work or available to work
    iii. unemployed_active: The number (in thousands) of people who were unemployed, but were actively looking for employment
    iv. unemployed_inactive: The number (in thousands) of people who were unemployed, but were not actively looking for employment due to waiting for ongoing job application results, being ill or believing that there was no work they could have successfully applied for
    v. unemployed_active_3mo: The number (in thousands) of people who have been actively unemployed for less than 3 months
    vi. unemployed_active_6mo: The number (in thousands) of people who have been actively unemployed for between 3 to 6 months
    vii. unemployed_active_12mo: The number (in thousands) of people who have been actively unemployed for between 6 to 12 months
    viii. unemployed_active_long: The number (in thousands) of people who have been actively unemployed for more than 12 months

5. **Monthly Unemployment Status Data** (lfs_month_status.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides detailed status information of the unemployed.
    - **Columns:**
        i. date: The date of the record.
        ii. variable: Value type; thousands of people (persons) or share of employed persons (share)
        iii. employed: The number of people who worked at least one hour for pay, profit or family gain, in thousands of people
        iv. employed_employer: The number of employed persons people who operate a business, a plantation or other trade and employ one or more workers to help them

v. employed_employee: The number of employed persons who work for a public or private employer and receive regular remuneration in wages, salary, commission, tips or payment in kind
vi. employed_own_account: The number of employed persons who operate their own farm, business or trade without employing any paid workers in their operations
vii. employed_unpaid_family: The number of employed persons who work without pay or wages on a farm, business or trade operated by another member of their family

6. **Annual Real GDP Data** (gdp_gni_annual_real.csv)
    - **Source**: Data.gov.my and World Bank
    - **Description**: Contains annual real GDP data.
    - **Columns:**
        i. date: The date of the record.
        ii. series: Series type, either absolute values ('abs') or annual growth ('growth_yoy')
        iii. gdp: The total value of goods and services produced within that year, after deducting the cost of goods and services used in production, but before deducting the consumption of fixed capital. The values are in constant 2015 prices, i.e. with base 2015 = 100.
        iv. gni: The total value of the production of the nationals of the country, whether they are in residence in Malaysia or not. Mechnically, it is defined as GDP plus net factor incomes from abroad. The values are in constant 2015 prices, i.e. with base 2015 = 100.
        v. gdp_capita: The ratio of GDP to the total population of Malaysia in that year
        vi. gni_capita: The ratio of GNI to the total population of Malaysia in that year
7. **Annual Nominal GDP Data** (gdp_gni_annual_nominal.csv)
    - **Source**: Data.gov.my and World Bank
    - **Description**: Contains annual nominal GDP data.

- **Columns:**
    i. date: The date of the record.
    ii. series: Series type, either absolute values ('abs') or annual growth ('growth_yoy')
    iii. gdp: The total value of goods and services produced within that year, after deducting the cost of goods and services used in production, but before deducting the consumption of fixed capital. The values are at current prices.
    iv. gni: The total value of the production of the nationals of the country, whether they are in residence in Malaysia or not. Mechnically, it is defined as GDP plus net factor incomes from abroad. The values are at current prices for that year.
    v. gdp_capita: The ratio of GDP to the total population of Malaysia in that year
    vi. gni_capita: The ratio of GNI to the total population of Malaysia in that year

8. **GDP Lookup Data** (gdp_lookup.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides additional context or conversion between nominal and real GDP.
    - **Columns:**
        i. method: Either the production approach ('production'), expenditure approach ('expenditure'), or income approach ('income')
        ii. code: Reference for the variable as contained within the dataset
        iii. variable_en: Definition of the variable in English
        iv. variable_bm: Definition of the variable in Malay
9. **GDP Nominal Supply Data** (gdp_annual_nominal_supply.csv)
    - **Source**: Data.gov.my
    - **Description**: Details GDP from the supply side.
    - **Columns:**

i. series: Series type, either absolute values ('abs') or annual growth ('growth_yoy')
ii. sector: Code for the economic sector, to be mapped using the Lookup Table. The lookup table will give you the English and Malay definitions.
iii. value: Either the value of GDP in RM millions (for rows with series type 'abs') or GDP growth as a percentage (for rows with series type 'growth_yoy')

10. **Malaysia Economic Indicator Data** (malaysia_economic_indicator.csv)
    - **Source**: Data.gov.my
    - **Description**: Includes various economic indicators for Malaysia.
    - Columns:
        i. leading: The Leading Index measures anticipations of the overall economic activity in the months ahead. The index tells us where the economy is going.
        ii. coincident: The Coincident Index is a comprehensive measure of the overall current economic performance.
        iii. lagging: The Lagging Index is to validate the signal of the Leading and Coincident Indexes
        iv. leading_diffusion: The Diffusion Index is a complement to the Composite Index. It is used to assist in making a decision especially in determining the turning point of an economic cycle. The value of 100 for Diffusion Index implies that all components are increasing and the value of zero shows that all components are decreasing.
        v. coincident_diffusion: The Diffusion Index is a complement to the Composite Index. It is used to assist in making a decision especially in determining the turning point of an economic cycle. The value of 100 for Diffusion Index implies that all components are increasing and the value of zero shows that all components are decreasing

3.3 Data Pre-processing

Data pre-processing is an important phase of the data science project life cycle which involves cleaning of the collected raw data into a convenient form for analysis. The pre-processing steps undertaken in this study will include:

1. **Data Cleaning**
    - **Handling Missing Values**: Datasets with missing values will be identified and handled by either using suitable methods of removing the records if they are deemed unnecessary for the analysis.
    - **Outlier Detection and Treatment**: Outliers will be identified and resolved to prevent them from skewing the analysis results.
    - **Duplicate Removal**: Duplicate records will be identified and removed to ensure each data entry is unique.

2. **Data Transformation**
    - **Date Conversion:** Date columns will be converted to datetime format to facilitate time series analysis.
    - **Resampling and Interpolation:** Annual GDP data will be resampled to a monthly frequency to align with the monthly unemployment data. Missing values will be interpolated.

3. **Feature Engineering**
    - **New Features**: New features will be created to further enhance the analysis. For example, seasonally adjusted unemployment rates will be used to remove seasonal effects.
    - **Normalisation and Scaling**: The data will be normalised and scaled to ensure that all features contribute equally to the analysis.
4. **Data Merging**
    - **Combining Datasets**: The datasets will be merged based on the common date index to form a comprehensive dataset for analysis.

With these pre-processing steps, the data will be ready for the next stages of analysis, ensuring that it is clean, suitable, and consistent for the descriptive and time series analysis.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis or EDA is the process of describing the data, identifying the patterns, outliers, and correlations. EDA will involve the use of descriptive statistics and density plots, histograms, box plots, and time plots. This stage is useful for formulating hypotheses and determining important trends for additional analysis.

1. **Descriptive Statistics:** Summary statistics such as mean, median, standard deviation, and variance will be calculated for key variables.
2. **Data Visualization:** Histograms and box plots will be used to visualize the distribution of unemployment rates and GDP.
3. **Time Series Plots:** Time series plots will be created to visualize the trends in unemployment rates and GDP over time.

3.5 Time Series Analysis

Time series analysis involves the application of statistical techniques to analyse time-ordered data points. This study will employ time series decomposition, ARIMA, and Prophet models to analyse and forecast unemployment rates based on GDP and other economic indicators.

1. **Time Series Decomposition**

The unemployment rate time series will be decomposed into trend, seasonal, and residual components.

2. **ARIMA Model**
    - **Model Selection:** The future unemployment rates are going to be forecasted with the help of the ARIMA model which stands for AutoRegressive Integrated Moving Average. It is suitable for univariate time series data and flexible enough

to incorporate other components of the time series like the AR, the differencing and the MA, the ARIMA model is chosen.

- **Parameter Estimation:** The values of p, d and q in the ARIMA model will be selected using ACF/PACF diagrams or the Grid search or any other method such as using the pmdarima library.
- **Model Fitting:** Next, the analysis of the unemployment rate data will be carried out using the ARIMA model to identify the patterns of the data and forecast.
- **Model Evaluation:** The performance of the forecast obtained from the developed ARIMA model will be evaluated by Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). However, the diagnostic tests for the residuals will also be conducted to determine if there is White noise.
- **Forecasting:** The future unemployment rates will be forecasted through the application of the ARIMA model. When explaining the model to the audience, the forecast will be plotted along with records to show the output of the model.

3. **Prophet Model**
    - **Model Introduction:** Prophet is an open source forecasting tool created by Facebook for time series data with strong seasonal patterns and gaps. Prophet will be used as the second model to forecast the data in order to compare the results with the ARIMA model.
    - **Data Preparation:** The data will be transformed in a way that Prophet requires the time series to have columns named ‘ds’ (date) and ‘y’ (value).
    - **Model Fitting:** The Prophet model will be estimated on the unemployment rate data.
    - **Forecasting:** Forecasts will be made using the Prophet model. The next data frames will be generated for the forecast horizon and the predictions will be displayed.
    - **Model Evaluation:** The performance of the Prophet model will be assessed with the help of the same parameters as in the case of the ARIMA model (MAE, MSE, RMSE). Furthermore, the breakdown of the model into its components (trend,

seasonal, and holiday) will be conducted to determine the contribution of each component towards the forecast.

Therefore, using these time series analysis techniques, the study seeks to forecast future unemployment rates given past GDP and other economic variables. The findings from both ARIMA and Prophet models will be compared to identify the best method of forecasting in this case.

Conclusion

This chapter has described the approach that will be used for examining the trends of GDP and unemployment rate in Malaysia. The Data Science Project Life Cycle will be used as a guide for the research including data collection and preparation, EDA, modelling and evaluation, and deployment. To achieve the research objectives, this research will use multiple datasets and apply several analytical tools like time series decomposition, ARIMA models, and Prophet models to examine the GDP and unemployment interaction and present the findings that can be helpful for policymakers and economists.