CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter 3 will elaborate in detail the research method, design and approach used to attain and analyse all the data to accomplish the research objectives and answer the research questions. This research will follow the data science project lifecycle from the beginning until the end. The aspects that will be discussed and included in this chapter are research framework, problem formulation, data collection, data pre-processing and data analysis.

3.2 Research Framework

Research is an activity that involves observation, collecting more data and information while research methodology can show or describe how data is collected, analysed and data presented. As being mentioned, this research will follow the data science project life cycle phases which includes business problem understanding, data collection, data preparation, exploratory data analysis (EDA), data modelling, model evaluation and model deployment. All those steps must be followed and be done steps by steps to achieve meticulous results. Figure 3.1 below shows the typical data science project lifecycle process steps by steps.

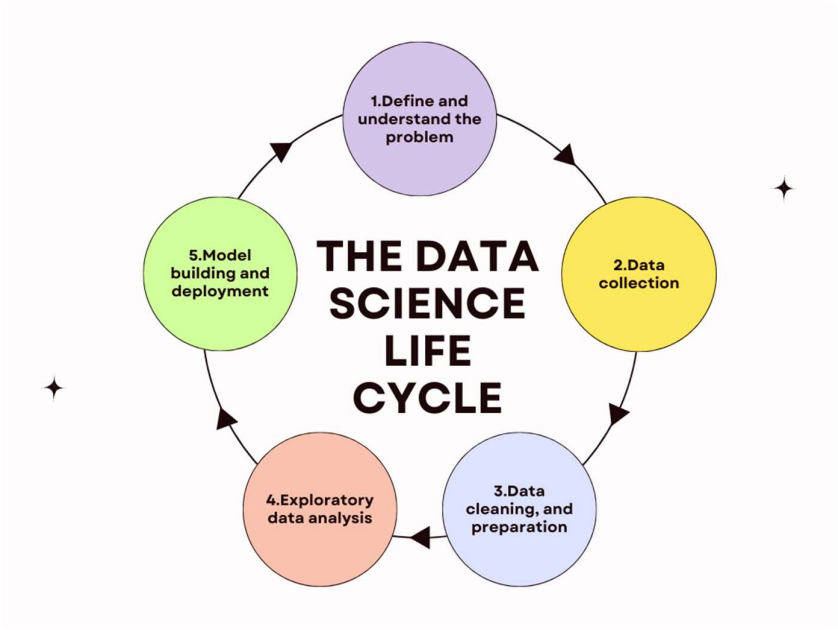

Figure 3.1 Data Science Project Life Cycles

A research framework is designed following the steps of typical data science life cycle starts from problem formulation, data collection method which involves on how the researcher find the sources of datasets that suitable for this research, data preparation, perform exploratory data analysis (EDA) acts as a preliminary analysis helps the researcher to get a comprehensive understanding of the dataset. This EDA process will achieve the first objective of this research. After exploratory data analysis (EDA) is performed, the suitable data modelling and deployment is identified to achieve the second and third of this research objectives which is to predict the future trends of fertility rates in Malaysia. The following figure 3.2 illustrates the process involved in the research frameworks.

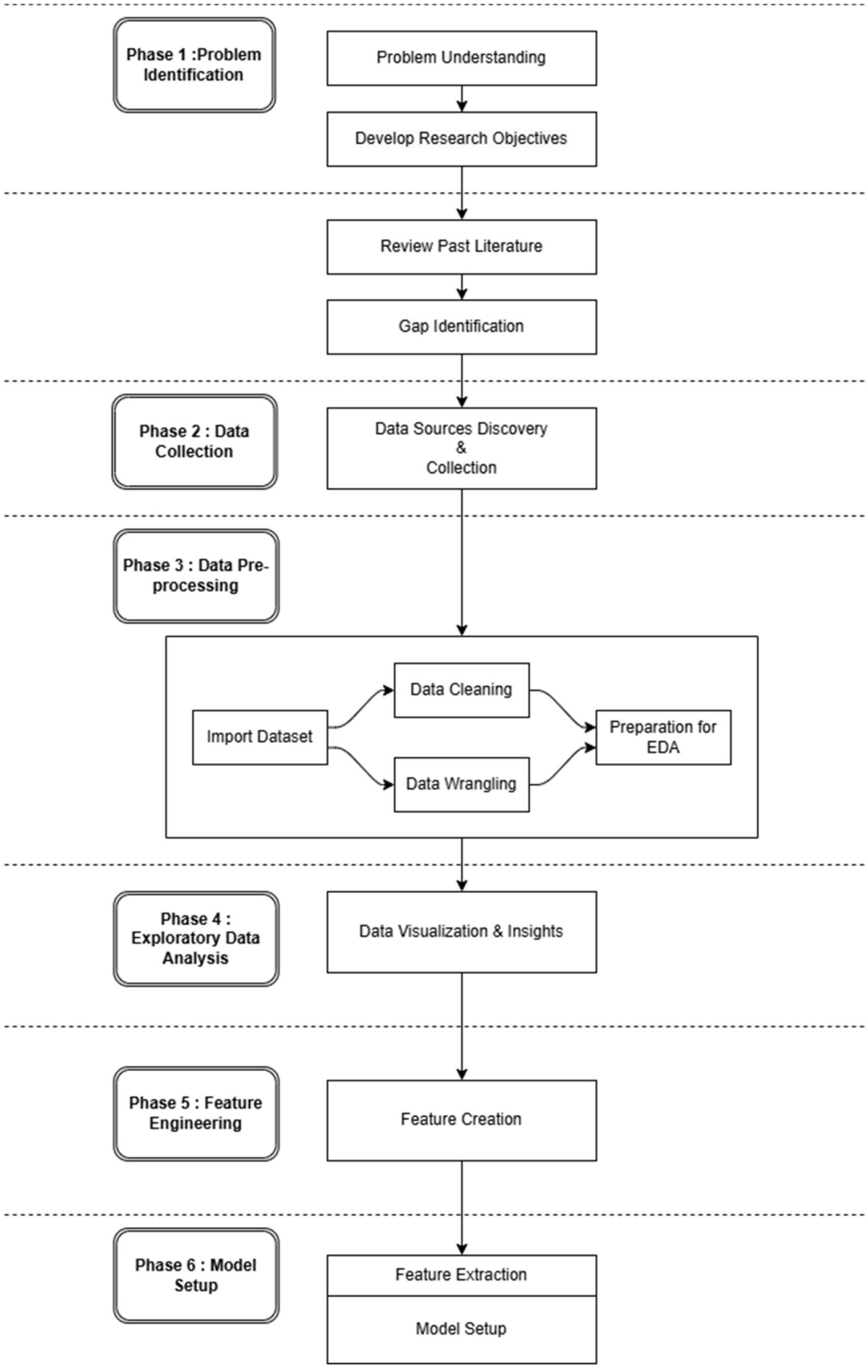

Figure 3.2 Research Framework

3.2.1 Problem Understanding

The initial of problem understanding and identification has been identified during literature review phase where the global trends of fertility has been declined, and Malaysian population has been labelled as ageing nation in 2022. So, the researchers are intended to find the root causes of declination trends and intended to predict the future trends in Malaysian populations to mitigate the prolonged effects cause by drop of fertility rates as well as to help the policy makers to refrain this problem from continues to happen.

3.2.2 Data Collection Method

Data collection is an activity of gathering data and information from all compatible sources to find an answer for the issue of research, test hypotheses, and analyze the results. Collecting data is an essential element for a research study. This section will explain how the data will be evaluated in this study. Basically, there are two types of approaches when collecting the data which are primary data and secondary data collection. The researcher decided to implement the primary data collection in this research to fulfill the research objectives.

a) Sources of Data

Table 3.1 Sources of data according to

Research Objectives	Data Required	Sources of Data
i) To analyze the current population trends in Malaysia from year 1970 to 2023.	Historical data of Malaysian population from year 1970 to 2023.	• Department of Statistics Malaysia (DOSM) • Malaysia Official Open Data Portal (data.gov.my)
ii) To determine the direction of the causal relationship between declining fertility rate (TFR) to economic performance of a country.	• Gross Domestic Product (GDP) • Total Fertility Rate (TFR)	• World Bank Open Data • Department of Statistics Malaysia (DOSM)
iii) To forecast future fertility trends in Malaysia through regression model approach.	Historical data of Malaysian birth rates from year 1920 to 2023	• Malaysia Official Open Data Portal (data.gov.my)

The dataset obtained from the website of Malaysia's Open Data Portal (data.gov.my) with collaboration of Department of Statistics Malaysia is enough to represent the current population of Malaysia. The dataset that suitable for the analysis is births data from 1920 to 2023 and population by states from 1970 until 2023. These two datasets are considered suitable to forecast the future trends of birth rates in Malaysia.

3.2.3 Data Preparation

Data preparation phase considered as elementary phases where the dataset obtained need to go through the process of data cleaning, processing of raw data so that it is suitable for further processing and analysis. The researcher will import the dataset into Jupyter Notebook using Python to condition of the data and detect missing values. The researcher also intended to combine the two datasets to go through exploratory data analysis process (EDA) to easier the researcher to find the data information and conditions for further analysis. The researcher intends to do the following data pre-processing as below:

Table 3.1 Sources of data according to

Data Wrangling	Purposes
1) Convert "date" to a datetime format.	To ensure the "date" column is in the correct format for analysis.
2) Handle categorical variables	Convert categorical columns like "state", "age", "sex", and "ethnicity" into a format suitable for modelling.
3) Check for missing values	Handle missing values appropriately
4) Feature engineering	• Extract year or other useful features from the "date" column. • Group and aggregate data based on the analysis needed such as "total population per state"
5) Remove or transform outliers	Inspect the "population" column for outliers and handle them.
6) Normalize/scale numerical data	Scale the "population" values for use in machine learning models.

3.2.4 Preliminary Analysis

Preliminary analysis in data science lifecycles usually known as Exploratory Data Analysis (EDA). This phase needs to be performed to help the researcher to identify and understand the condition of the datasets and get the initial insights and findings through the datasets. The researcher intends to do the following preliminary analysis that is suitable.

Table 3.1 Sources of data according to

Analysis	Data Visualizations	Details
Population Distribution by Age Group	Bar Chart	Visualize the population across different age groups.
Gender-Based Population Analysis	Pie Chart	Display the proportion of each gender within a specific state or overall.
State-Wise Population Comparison	Bar Chart	Compare the total population across different states.
Time Series Analysis	Line Chart	Show how population changes over time for specific age groups, genders, or states.

3.2.5 Model Building & Deployment

To achieve the second and third objectives, two types of models have been identified by past literature review by (Mohd et al., 2021) which is using Autoregressive Distributed Lag (ARDL) and (Muadz Bin Zulqarnain & Md Yusuf, 2022) which is using Granger Causality test to identify the direction of causal relationship between socioeconomic factors and fertility rates.

Table 3.1 Sources of data according to

Research Objectives	Types of Analysis	Methodology/ Modelling
i) To analyze the current population trends in Malaysia from year 1970 to 2023.	Exploratory Data Analysis (EDA) • Descriptive Statistics & Analysis	
ii) To identify the direction of the causal relationship between declining fertility rate (TFR) to economic performance of a country.	• Diagnostic Analysis	Granger Causality Test Model
iii) To forecast future fertility trends in Malaysia through regression model approach.	• Predictive Analysis	Autoregressive Distributed Lag (ARDL) Model