CHAPTER 3

RESEARCH METHODOLOGY

3.0 Introduction

This chapter describes the research method that predicts the profitability of AirAsia using fuel price trends via ARIMA and Random Forest models. It includes a historical financial dataset, fuel price information, and passenger demand metrics to evaluate how volatility in fuel prices affects financial performance. The methodology comprises problem definition, data collection and preprocessing, feature engineering, model building, and assessment. ARIMA captures trend and seasonality effects for fuel prices and profitability while the Random Forest manages multivariate and non-linear associations by using engineered features such as lagged values, moving averages, and seasonality indicators. In turn, both models balance each other by pairing conventional statistical forecasting with machine learning strength to further make a more robust and precise forecast regarding AirAsia's profitability when there are high fluctuations in the prices of fuels.

3.1 Research Framework

This research framework includes the following steps:

1. **Problem Formulation:** State the research aims and find the effect of fuel price changes on AirAsia's profits.
2. **Data Collection:** Historical financial data of AirAsia, fuel price data from the Ministry of Energy and Natural Resources (KPKT), and passenger demand data from Malaysia Airports Authority (MAA) are compiled.

3. **Data Pre-processing and Exploratory Data Analysis (EDA):** Handle missing values and outliers, perform initial visualizations and summary statistics to understand the structure and anomalies in the dataset.
4. **Feature Engineering and Time Series Decomposition:** engineer features like lagged values, rolling statistics, and date-based indicators to enrich the modelling dataset and decompose the time series to extract trend and seasonality.
5. **Model Development:** Forecasting models under univariate time series analysis using ARIMA and multivariate modelling via Random Forest to capture non-linear relationships.
6. **Model Evaluation:** Compare model performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess their accuracy in forecasting AirAsia's profitability based on fuel price trends. To determine which model provides the most reliable and consistent predictions, and to potentially combine or select the best-performing model for final forecasting. This evaluation ensures that the chosen model can effectively capture both trend patterns which is ARIMA and non-linear relationships which is Random Forest leading to more accurate and robust forecasts.

The details of the research framework for this study are shown in Figure 3.1,

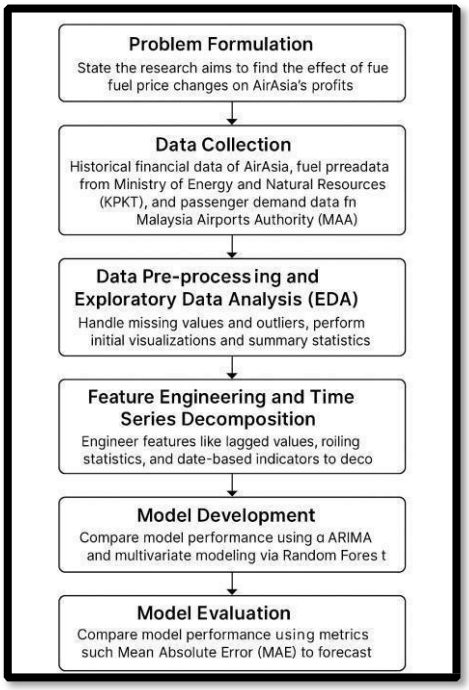

Figure 3.1 Research Framework for Sentiment Analysis

3.2 Problem Formulation

This study will predict the profitability of AirAsia by analysing the historical trends of fuel prices through a hybrid modelling approach that merges ARIMA with Random Forest techniques. This research, which will take advantage of both strengths ARIMA to capture time-based patterns and Random Forest to handle multivariate and non-linear relationships aims to deliver more accurate and robust profitability forecasts. To ensure reliable and meaningful results, several key challenges must be addressed:

a. To identify how Fuel Price Volatility affects the Financial Performance of AirAsia. That is, how changes in jet fuel prices affect indicators of profitability like net income, operating cost, and revenue.

b. To successfully combining Univariate Time Series Forecasting (ARIMA) with Multivariate Machine Learning Modelling (Random Forest) to improve forecasting accuracy and offer actionable insights for financial planning and strategic decision-making in the airline industry.

3.3 Data Collection

This project consists of data format to gather dataset that is accurate for this project. The Table 3.1 below, outlines the data format as the minimum requirements for a dataset, specifying that it should cover a time frame of at least 3 to 5 years, contain between 50,000 and 200,000 records, and include at least 5 to 7 key variables. These guidelines ensure that the dataset is sufficiently comprehensive to capture meaningful trends and patterns while remaining manageable in size and complexity.

Table 3.1 The Data Format

Parameter	Value
Time Frame	Minimum of 3-5 years data
Data Size	Dataset with 50,000 until 200,000
Variables	Include at least 5 to 7 key variables

The information for this research was taken from an organized file from public dataset which has old notes about plane runs and passenger actions. This file has details like reservation kind, journey path, extra spending, and time-related factors that can be used to predict profit markers. It lays a base for knowing passenger want patterns, seasonality, and extra money factors, which are key parts in predicting airline profit. These ideas can be made better by adding outside data like fuel costs and money measures for fuller modelling using ARIMA and Random Forest methods.

Id	PAXCOUNT	SALESCHAN	TRIPTYPEDE	PURCHASEL	LENGTHOFS	flight_hour	flight_day	ROUTE	geoNetworl	BAGGAGE_(	SEAT_CATE	FNB_CATEG	INS_FLAG	flightDuration_hour
1	2	Internet	RoundTrip	262	19	7	Sat	AKLDEL	New Zealan	1	0	0	0	5.52
2	1	Internet	RoundTrip	112	20	3	Sat	AKLDEL	New Zealan	0	0	0	0	5.52
3	2	Internet	RoundTrip	243	22	17	Wed	AKLDEL	India	1	1	0	0	5.52
4	1	Internet	RoundTrip	96	31	4	Sat	AKLDEL	New Zealan	0	0	1	0	5.52
5	2	Internet	RoundTrip	68	22	15	Wed	AKLDEL	India	1	0	1	0	5.52
6	1	Internet	RoundTrip	3	48	20	Thu	AKLDEL	New Zealan	1	0	1	0	5.52
7	3	Internet	RoundTrip	201	33	6	Thu	AKLDEL	New Zealan	1	0	1	0	5.52
8	2	Internet	RoundTrip	238	19	14	Mon	AKLDEL	India	1	0	1	0	5.52
9	1	Internet	RoundTrip	80	22	4	Mon	AKLDEL	New Zealan	0	0	1	0	5.52
10	1	Mobile	RoundTrip	378	30	12	Sun	AKLDEL	India	0	0	0	0	5.52
11	2	Internet	RoundTrip	185	25	14	Tue	AKLDEL	United King	1	1	1	0	5.52
12	1	Internet	RoundTrip	8	43	2	Sat	AKLDEL	New Zealan	1	1	1	0	5.52
13	4	Internet	RoundTrip	265	24	19	Mon	AKLDEL	New Zealan	1	0	1	0	5.52
14	1	Internet	RoundTrip	185	17	14	Fri	AKLDEL	India	0	0	0	0	5.52
15	1	Internet	RoundTrip	245	34	4	Tue	AKLDEL	New Zealan	1	1	1	0	5.52
16	1	Internet	RoundTrip	197	18	14	Thu	AKLDEL	India	1	0	0	0	5.52

Figure 3.2 The dataset preview

As depicted in Figure 3.2, data for this study comprises historic records pertinent to airline operations and passenger behaviour. Features present in each record include booking type which is round-trip or one-way, travel route, ancillary spending, day of the week, country information, and numerical values that are most likely related to time or cost metrics. It is estimated that this data has 50,000 rows and 15 columns.

3.4 Data Pre-Processing

The initial analysis needs to be completed before moving on to further preprocessing. Data merging procedures are required to unify all the raw data into a single data frame once we have a good understanding of the features available in the dataset. Several data processing and transformation procedures will be used on the dataset to further unify the disorganized raw data. Table 3.2 lists every detail of the data pre-processing that was used, including handling missing values, outlier detection, normalization, and feature engineering, which are essential steps in preparing the data for modelling using both ARIMA and Random Forest techniques.

Table 3.2 Data Pre-Processing Methods

Data Pre-Processing	Purpose
Preliminary Analysis	To evaluate the provided dataset and to understand its structure and key variables like financial data, fuel prices, and passenger numbers.
Data Cleaning	Find and fix missing values in important data like revenue and fuel costs. Remove or fill in missing data as needed. Also, handle outliers to improve data quality.
Data Visualization	Charts like time series and pie charts to show trends and distributions of fuel prices, revenue, and passenger demand. This helps find patterns and issues that affect the models.

3.4.1 Preliminary Analysis

Preliminary analysis is an important step in any data analysis because it helps to become familiar with the dataset, understand its structure, format, and the types of variables it contains. It also helps identify issues that need to be addressed for reliable analysis, such as missing values, outliers, or inconsistencies.

In this project, the preliminary analysis includes two main stages:

a. Identify common patterns in the raw data, such as trends in fuel prices, revenue, and passenger demand.

b. Evaluate the data distribution over time and by key factors like fuel price changes and passenger load factors.

3.4.2 Data Cleaning

Data cleaning is an important process in time series analysis, especially to ensure that the data used is clean, relevant, and can be processed effectively by the model. Here are the data cleaning steps carried out on the dataset containing AirAsia's financial data, fuel prices, and passenger demand:

1. **Handle Missing Values:**

   Identify and fix missing values in key variables such as revenue, net income, fuel cost, and passenger load factor. Missing data will be removed or filled in using appropriate techniques like median imputation for skewed data.

2. **Detect and Handle Outliers:**

   Use methods like Z-score or IQR to detect outliers in features such as fuel prices and revenue. Outliers will be either removed or capped to improve data quality.

3. **Normalize Features:**

   Scale numerical features such as fuel price, and revenue to a standard range using Min-Max Scaling or Standardization to ensure all features contribute equally to the model.

4. **Encode Categorical Variables:**

   Convert categorical variables such as month and quarter into numerical formats using One-Hot Encoding or Label Encoding to prepare the data for modelling.

5. **Feature Engineering:**

   Create new features such as lagged values, moving averages, and seasonal indicators to enrich the dataset and improve model performance.

6. **Plot Time Series:**

   Visualize the time series of key variables such as fuel prices, revenue, and passenger demand to identify patterns, trends, and seasonality.

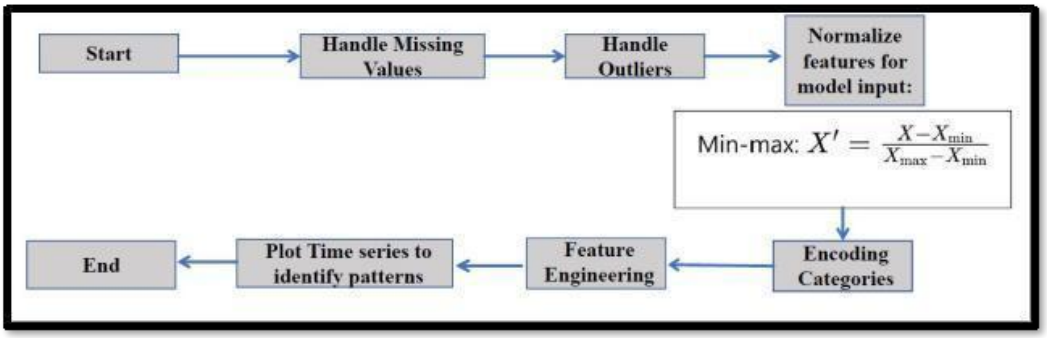

Figure 3.3 Flow Data Cleaning and Preparation

Figure 3.3 illustrates the data cleaning and preparation workflow for forecasting AirAsia's profitability based on fuel price trends using ARIMA and Random Forest models. The flowchart outlines a systematic sequence of steps to ensure that the dataset is clean, well-structured, and ready for modelling.

3.4.2.1 Handle Missing Values

This section explains how to find and fix missing data in the dataset to make sure the data is reliable for analysis. Missing data can make models less accurate and affect the results. These issues need to be fixed by using methods like median imputation for skewed data like fuel prices and mean imputation for normally distributed data. This step is important to prepare good-quality data for forecasting and modelling.

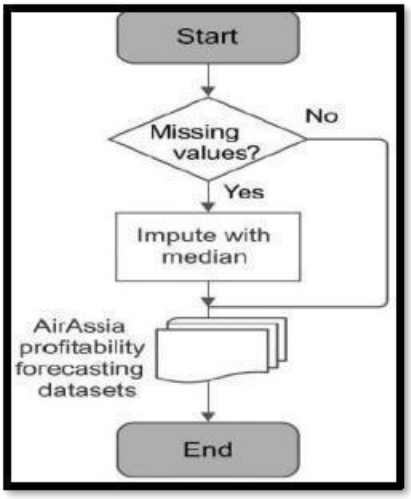

Figure 3.4 Flowchart for Handling Missing Value

Based on Figure 3.4, the flowchart illustrates the procedure for handling missing values in the AirAsia profitability forecasting dataset. The process begins with checking for the presence of any missing values. If such values are detected, they are imputed using the median of the respective feature, to maintain the distributional characteristics of the data. Once the imputation is completed, the cleaned dataset is prepared for subsequent analysis or model development.

3.4.2.2 Detect and Handle Outliers

This step is important because it helps improve the accuracy of forecasting models by identifying and handling extreme values which is the outliers in the data. In this project on forecasting AirAsia's profitability using fuel price trends, outliers in variables like fuel prices or revenue can negatively affect model performance specifically for ARIMA, which is sensitive to extreme values. By using methods like Z-score or IQR to detect and manage these outliers, which help in enhancing data quality and make both ARIMA and Random Forest models more reliable and accurate.

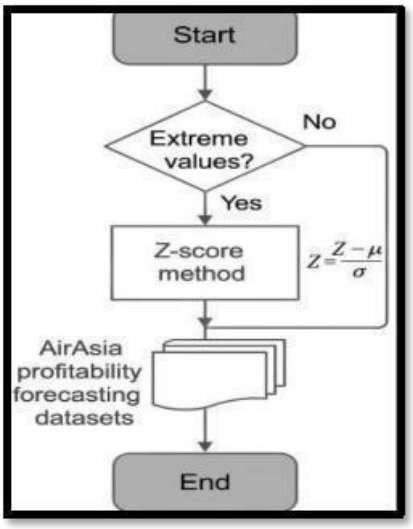

Figure 3.5 Flowchart for Detecting and Handle Outlier

Figure 3.5 presents a systematic flowchart for detecting and handling outliers in the dataset used to forecast AirAsia's profitability based on fuel price trends. It begins by checking for extreme values, then applies the Z-score method to identify outliers, and finally implements strategies such as removal or capping to improve data quality. This process is essential for ensuring accurate and reliable forecasting using ARIMA and Random Forest models by minimizing the impact of abnormal data points on model performance.

3.5 Time Series Decomposition and Feature Engineering

This step begins with feature engineering, where new variables are created to enrich the dataset and improve model performance. These features include lagged values, moving averages, seasonal indicators, and derived metrics. Each engineered feature is designed to capture temporal patterns and relationships between key variables such as fuel prices, revenue, and passenger demand.

1. **Lagged Features:**

$$\text{Fuel\_Price}_{t-1}, \quad \text{Fuel\_Cost}_{t-1}$$

Previous values of fuel price and cost. Formula:

2. **Moving Averages:**

$$\text{MA\_Fuel\_Price}_t = \frac{\text{Fuel\_Price}_t + \text{Fuel\_Price}_{t-1} + \text{Fuel\_Price}_{t-2}}{3}$$

Rolling averages of fuel price and cost. Formula:

3. **Seasonal Features:**

$$\text{Month}_t = \text{Month}(\text{Date}_t)$$

Month and quarter of the year. Formula:

4. Derived Metrics:

Fuel cost and revenue per passenger. Formula:

$$\text{Fuel\_Cost\_Per\_Passenger}_t = \frac{\text{Fuel\_Cost}_t}{\text{Passengers}_t}, \quad \text{Revenue\_Per\_Passenger}_t = \frac{\text{Revenue}_t}{\text{Passengers}_t}$$

Following feature engineering, time series decomposition is applied to break down key variables such as AirAsia's profitability or fuel prices into their core components which is trend long-term movement, seasonality repeating patterns, and random noise. This decomposition helps uncover hidden patterns in the data, such as seasonal fluctuations in fuel prices or long-term trends in profitability. The decomposition process can be represented as:

Feature engineering and time series decomposition enhance the predictive power of both ARIMA and Random Forest models by improving how temporal dependencies and variable relationships are represented in the modelling process.

3.6 Model Development and Evaluation

This In this phase, two forecasting models ARIMA AutoRegressive Integrated Moving Average (ARIMA) and Random Forest are developed and compared to evaluate their effectiveness in forecasting AirAsia's profitability based on fuel price trends. The ARIMA model is employed for univariate time series forecasting, particularly to capture patterns such as trends, seasonality, and autocorrelation in historical fuel price and profitability data. This model is suitable for datasets where future values depend linearly on past values and previous forecast errors (Yunos et al., 2024).

On the other hand, the Random Forest model is used for multivariate regression forecasting, allowing the inclusion of multiple input features such as lagged fuel prices, moving averages, seasonal indicators, and derived metrics like fuel cost per passenger. Random Forest leverages ensemble learning by combining predictions from multiple decision trees, which helps reduce variance and improve prediction accuracy (Yunos et al., 2024). Both models are

trained using historical data after feature engineering, normalization, and time series decomposition steps. The dataset is split into training and testing sets to validate model performance using evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). This comparative analysis aims to identify which model provides more accurate and reliable forecasts for AirAsia's profitability under fluctuating fuel price conditions. Below is the formula for Mean Absolute Error measurement:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

MAE measures the average magnitude of errors in a set of predictions, without considering their direction (positive or negative). It gives equal weight to all individual differences between actual values (*yi*) and predicted values (*y^i*). A lower MAE indicates better model performance, as it means the predictions are closer to the actual values. Other than that, below is the formula for RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE is the square root of the average of squared differences between predicted and actual values. Unlike MAE, RMSE penalizes larger errors more heavily due to the squaring of differences, making it more sensitive to outliers. This makes RMSE a good metric when large errors are particularly undesirable. Like MAE, a lower RMSE value indicates better predictive accuracy.

3.7 Summary

This chapter explains the research methodology in detail, from data collection to evaluation of the classification model. This process ensures that the forecasting of AirAsia's profitability based on fuel price trends is conducted systematically and data driven.