# Predicting Flight Prices Using Supervised Learning: A Linear Regression Approach

## Graduation Thesis: The Interim Report

**Chen Lv**
Student ID: 50079791
Program: CS

a35lc21@abdn.ac.uk

*Aberdeen-SCNU Joint Institute of Data Science and AI,*
*University of Aberdeen, Aberdeen AB24 3UE, UK*

# Introduction

Many factors, including travel dates, locations, carriers, and demand patterns, affect flight costs. Predicting flight prices is a difficult but essential task for passengers, airlines, and travel companies due to the constantly changing makeup of these components. Better customer planning and airline pricing strategies can be made possible by accurate flight price predictions.The goal of this research is to forecast flight prices by creating a predictive model through the use of supervised learning techniques. In particular, the model will assess and forecast flight prices using historical data via a linear regression approach[1]. The dataset, which was obtained via Kaggle, includes a number of features that are used as model inputs, including travel dates, destinations, and airlines.Because of its ease of use and interpretability, linear regression—a basic machine learning algorithm—will be used to better understand the relationship between the independent features and the dependent variable (flight price). This project's goal is to assess how well linear regression predicts flight costs and investigate the possibility of additional improvement through the use of more sophisticated methods in subsequent research.

# Goals

The main objective of this research is to use linear regression to create an accurate and effective predictive model for flight price forecasting[2]. The model seeks to pinpoint the main variables affecting changes in flight prices by utilizing historical data on travel dates, destinations, airlines, and demand patterns. The particular goals are:

- To preprocess and clean the dataset to ensure its suitability for modeling.
- To implement a linear regression model to predict flight prices based on the identified features.

- To evaluate the performance of the linear regression model using appropriate metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
- To identify the most significant features affecting flight price predictions, providing valuable insights for both consumers and airlines.
- To explore opportunities for improving the model's accuracy and efficiency, potentially through feature engineering or by applying more advanced machine learning techniques in future studies.

# Methodology

In order to develop an effective flight price prediction model using supervised learning techniques, a systematic approach will be followed. The methodology is outlined in the steps below:

- **Defining the Problem and Setting Objectives**
  The first step is to clearly define the problem of flight price prediction and set specific objectives for the project, such as understanding key factors influencing flight prices and evaluating the performance of the chosen model (linear regression[1]).
- **Data Collection and Preprocessing**
  The dataset from Kaggle will be used as the primary data source. The data will be collected and cleaned to ensure consistency and remove any noise. Preprocessing steps will include handling missing values, encoding categorical variables, and normalizing numerical features to prepare the dataset for modeling.
- **Literature Review and Model Selection**
  A literature review will be conducted to review existing works related to flight price prediction[3] and supervised learning techniques, with a focus on linear regression models. This will help identify the best practices and methodologies used in similar projects and refine the approach for this study.
- **Feature Selection and Engineering**
  Key features such as travel dates, destinations, airlines, and demand trends will be selected. Additional feature engineering techniques may be applied to improve the model, such as creating new features or transforming existing ones to better represent the underlying patterns in the data.
- **Model Development and Training**
  A linear regression model[1] will be developed and trained on the prepared dataset. The training process will involve splitting the data into training and testing sets to ensure that the model generalizes well to unseen data.
- **Model Evaluation and Tuning**
  The model's performance will be evaluated using standard regression metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). If the

model's performance is not satisfactory, hyperparameter tuning will be performed to improve its accuracy.

- **Interpretation of Results**
  The results from the linear regression model will be analyzed to understand the relationship between the predicted flight prices and the input features. The most significant features influencing the price predictions will be identified, providing valuable insights for further analysis and refinement.

# Resources Required

The resources essential to the successful delivery of the project include:

- **Hardware:**
  Laptop: A computer with sufficient processing power to run data analysis tasks, train machine learning models, and perform computational tasks efficiently.

- **Software:**
  Programming Environment: A suitable integrated development environment (IDE) such as Python with libraries like Pandas, NumPy, Scikit-learn, and Matplotlib for data analysis and machine learning.
  Data Processing and Machine Learning Tools: Python-based libraries for data manipulation (e.g., Pandas, NumPy), machine learning model implementation (e.g., Scikit-learn for linear regression), and visualization (e.g., Matplotlib, Seaborn).
  Documentation Tools: Microsoft Word or LaTeX for writing reports, and potentially Jupyter Notebook for documenting code and providing inline explanations during the analysis and model development stages.

- **Other Tools:**
  Data Visualization Software: For presenting the results and insights from the analysis (e.g., Tableau, Power BI, or Python's Matplotlib and Seaborn).

# Risk Assessment

Several potential risks have been identified in this project, along with strategies to mitigate them:

**1.Data Quality and Availability**

Risk: Missing values or inconsistencies in the dataset.

Mitigation: Apply data cleaning and imputation techniques. Seek additional data if necessary.

**2.Model Performance**

Risk: Linear regression may not provide accurate predictions.

Mitigation: Use multiple evaluation metrics (MAE, RMSE) and consider more complex models.

**3.Overfitting**

Risk: The model may overfit the training data, leading to poor generalization.
Mitigation: Use cross-validation and regularization techniques.

**4.Computational Limitations**

Risk: Insufficient resources to handle large datasets or complex models.
Mitigation: Optimize data processing and consider cloud-based resources if necessary.

**5.Time Constraints**

Risk: Delays in completing the project.
Mitigation: Set clear milestones and conduct regular progress reviews.

**6.Technical Issues**

Risk: Software bugs or system failures.
Mitigation: Use version control and regular backups. Test and debug at each stage.

**7.User Interpretation**

Risk: Users may find it difficult to interpret the model's results.
Mitigation: Provide clear documentation and visualizations of results.

**8.Ethical Considerations**

Risk: Privacy concerns related to travel data.
Mitigation: Ensure no personal data is used and follow ethical data guidelines.

# Timetable

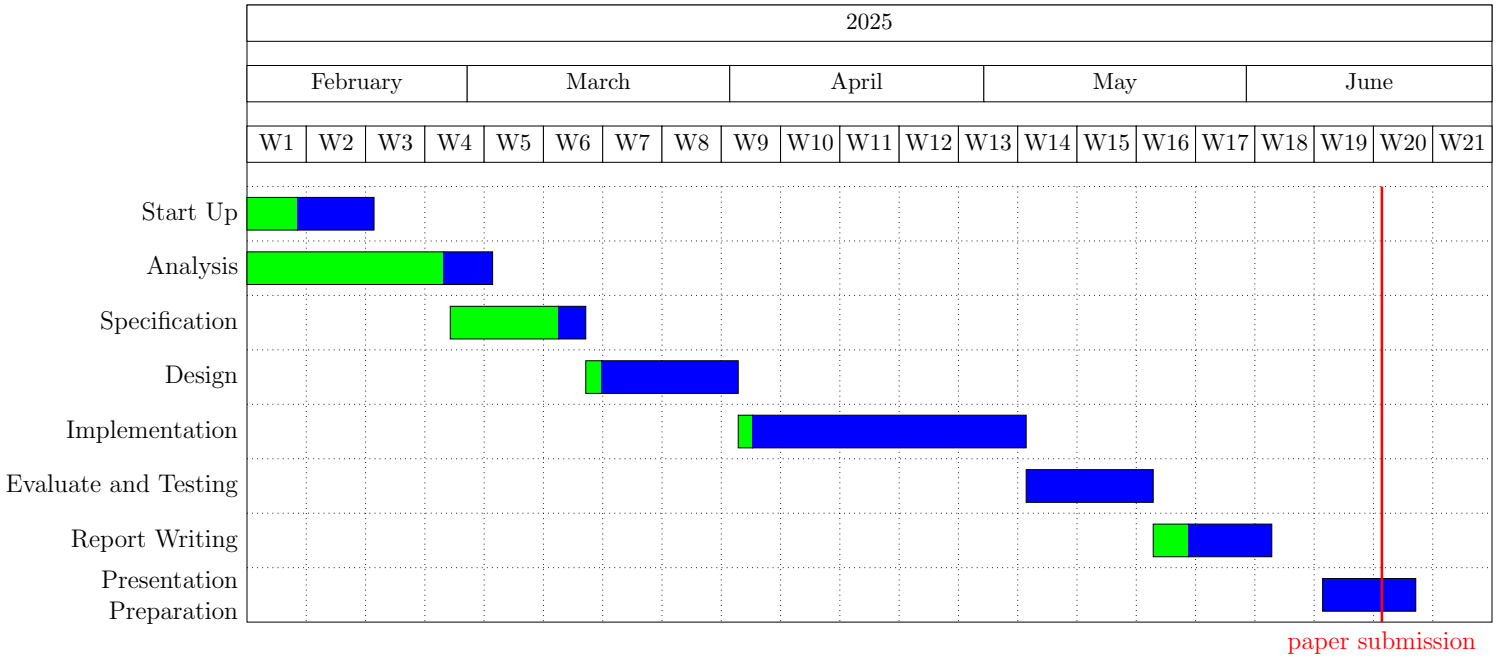


Figure 1: Main Project Activities.

# References

[1] Konstantinos Tziridis, Th Kalampokas, George A Papakostas, and Kostas I Diamantaras. Airfare prices prediction using machine learning techniques. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1036–1039. IEEE, 2017.

[2] Stacey Mumbower, Laurie A Garrow, and Matthew J Higgins. Estimating flight-level price elasticities using online airline data: A first step toward integrating pricing, demand, and revenue optimization. *Transportation Research Part A: Policy and Practice*, 66:196–212, 2014.

[3] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso, Steven Luis, and Shu-Ching Chen. A framework for airfare price prediction: a machine learning approach. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)*, pages 200–207. IEEE, 2019.