Report

# Flight Prices Analysis
Data Management

Project in Data Management with a specialisation
in Data collection, Data cleaning, Data visualisation
and Modelling.

Student 1: Adelia Rahmawati Safitri

Student 2: Céline Lam

Student code 1: s175873

Student code 2: s191578

Course: ECON2306-1

Lecturer: Malka Guillot

June 2023

# Abstract

The goal of this project is to develop a model that can accurately predict the prices of round-trip airline tickets from Lufthansa and Swiss airlines to São Paulo and New York City. By determining the optimal time to purchase tickets for each destination, it is hoped that significant savings can be achieved, allowing for the maximisation of the travel budget. To accomplish this, historical data on ticket prices and other relevant factors such as holidays will be analysed in order to identify patterns and trends that can be used to forecast future ticket prices. This project is expected to provide valuable experience in working with time series data and developing predictive models, while also uncovering practical ways to save money on air travel during peak travel times.

**Keywords:** Destinations, Historical data, Ticket prices, Patterns, Trends, Forecasting, Time series data, Predictive models

# 1. Introduction

As students, we are constantly seeking opportunities to maximise our limited budgets. One area where we identified potential for cost savings is air travel, particularly during peak travel times such as holidays. During these periods, ticket prices can increase significantly, making it challenging for us to afford travel for visiting family or taking vacations.

With this in mind, we decided to undertake a project to develop a model capable of predicting the prices of round-trip tickets from Lufthansa and Swiss airlines to São Paulo and New York City. By analysing historical data on ticket prices and other relevant factors such as holidays, we aim to identify patterns and trends that can assist us in forecasting future ticket prices. Utilising this information, our goal is to determine the optimal time to purchase a ticket for each destination, allowing us to save money and make the most of our travel budget.

Through this project, we hope to gain valuable experience in working with time series data and developing predictive models while also discovering practical ways to save money on air travel during peak travel times. By focusing on two specific airlines and destinations, we aim to develop a model that is both accurate and applicable to our travel needs.

## 1.1 Research questions/hypotheses

The objective of this study is to create a model that can predict the prices of round-trip tickets for Lufthansa and Swiss flights to Sao Paulo and NYC. The first hypothesis posits that historical ticket prices for these flights exhibit identifiable patterns and trends. By examining ticket price data over time, it is anticipated that these patterns can be discerned and used to forecast future prices. To evaluate this hypothesis, historical ticket price data and other relevant factors will be analyzed to identify patterns and trends. This analysis will employ statistical methods to reveal relationships between ticket prices and other variables such as the time at which flight ticket data is collected.

The second hypothesis proposes that the time at which flight ticket data is collected significantly impacts ticket prices. It is anticipated that collecting data at certain times during the week will result in lower ticket prices.

The third hypothesis asserts that a model trained on historical data can accurately predict future ticket prices. To test these hypotheses, a predictive model will be created using machine learning techniques to forecast future ticket prices based on historical data. This model will consider a variety of factors known to affect ticket prices, including the time at which flight ticket data is collected and airline-specific factors.

The research questions are: What are the primary factors that influence ticket prices for Lufthansa and Swiss flights to Sao Paulo and NYC? How does the time at which flight ticket data is collected impact ticket prices? When is the optimal time to collect flight ticket data to minimize cost? And how accurately can future ticket

prices be predicted using a model trained on historical data? In order to answer these questions, the study will involve collecting and analyzing large amounts of data on ticket prices and other relevant factors. This data will be used to train and evaluate the predictive model, allowing us to determine its accuracy and effectiveness in forecasting future ticket prices. By identifying the key factors that influence ticket prices and determining the optimal time to collect flight ticket data, we hope to provide valuable insights for travelers looking to save money on air travel.

# 2. Data

In this project, data was collected from the online travel booking platform booking.com. This platform was chosen due to its extensive coverage of flights from Lufthansa and Swiss airlines to Sao Paulo and NYC. Data collection was conducted on a daily basis, with observations recorded every 2 hours between 8 am and 10 pm. This schedule was chosen to capture variations in ticket prices throughout the day. In total, 11619 data points were collected over the course of the project.

The data collected includes a range of variables relevant to the research questions. These include the price for round-trip tickets for the specified destinations using Lufthansa and Swiss airlines, as well as the departure and arrival times/dates for both the outbound and inbound flights. Additionally, data on the duration of the outbound and inbound flights and the number of stopovers for both flights were recorded. Since the weight allowance for luggage is consistent across both airlines, this variable was not included in the data collection.

Furthermore, the time at which each observation was recorded was also noted in order to facilitate analysis of trends in ticket prices relative to the day/time of the week. This information is expected to provide valuable insights into how ticket prices vary over time.

By analyzing this comprehensive dataset, it is anticipated that patterns and trends can be identified that will aid in predicting future ticket prices for these specific airlines and destinations. The data collected will be used to train a predictive model that can forecast future ticket prices with a high degree of accuracy. This model will be a valuable tool for travelers seeking to minimize costs and maximize their travel budgets.

## 2.1 Data collection

sdfghjklmqsdfghjklmsdfghjklmdfgvbhnj

## 2.2 Data preparation

Describe what you did in order to use the data for analysis.

**2.3**

# 3. Approach

Describe the techniques that you will use for managing and analyzing your data and the tool(s) that you will use for data management/data analysis and evaluation. Motivate why you have chosen a certain method over others and describe how you plan to evaluate your solution. This section should give a detailed account of what you did in order to answer your questions.

# 4. Results

Describe your final data analysis results and everything that you discovered on the way the put you forward and you think is worth describing here. Elaborate on the significance of your findings/results and convey your thoughts.

# 5. Discussion

What do the results mean (their significance and to whom) and how do they answer your research questions/hypotheses? From the evaluation and validation of the tool, what can you conclude?

## 5.1 Limitations and Challenges

What could have been investigated if given more time? What have been difficult when solving the problem and getting answers for your research questions/hypotheses?

# 6. Conclusion

Write the conclusion. What did you gain from the project assignment? Briefly explain your questions/hypotheses, findings, and meaningful discussion points in relation to the data collection, data management, data analysis, visualization and interaction concepts, and evaluation of your tool. What additional investigations need to be performed (or what is the limitation) in order to say that your solution is a good one for the problem?

# 7. Reflections on own work

- Describe how you decided to scope (and/or re-scope) your problem formulation during the work, and given the data you had access to.

- Describe how you searched for knowledge on how to scope the DS question you wanted to answer

- understand how to implement, test, and validate your results.

- Which sources helped you to get progress and how?

- What would you have done differently if you were to start over again, for understanding the problem faster and for understanding what could be done with the data you had access to?

- What else is there that you would have changed about this assignment?