# Final Project Report

Subject: **OPS-640-01: Data Analysis**

# Project: Flight Price Prediction

## Introduction:

The flight prediction dataset is a collection of information encompassing various aspects of air travel, including airline details, journey dates, source and destination cities, specific routes, departure and arrival times, total stops, and ticket prices. The dataset captures the complexity and diversity of flight itineraries, providing a rich landscape for analysis. Particularly intriguing are the patterns that may emerge from examining the relationships between key variables, such as impact of flight duration and total number of stops on ticket prices.

As a recommendation for the business, a thorough analysis of the dataset can inform airlines and travel agencies about the factors influencing the pricing and customer decision-making. This understanding can guide targeted marketing efforts, optimize route planning, and contribute to the development of competitive pricing strategies, ultimately improving overall business performance and customer satisfaction

## Data Overview:

The flight prediction dataset exhibits a structured format with 10,683 records, each capturing a unique flight itinerary.

Key columns include:

Airline, Date of journey, Source, Destination, Route, Duration, Total stops, Price.

Furthermore, the dataset has been strategically divided into separate datasets, each focusing on flights operated by the airline. This way of splitting the data helps us look at the details and patterns specific to airline.

To split the dataset into different datasets we used Python programming, The code can be viewed in https://colab.research.google.com/drive/1MCy8EBTXuQOGB2UtOpdCHhoUjs9qnYS_?usp=sharing

## Descriptive Statistics:

In examining flight prediction dataset, we dive into descriptive statistics to gain comprehensive understanding of key numerical variables. The mean price provides insight into average cost, while median provides robust measure unaffected by extreme values. Additionally standard deviation illuminates the variability in prices. Visual representations including histograms and boxplots gives distribution of flight prices.

The results are shown in the table below.

| Airline | Mean | Median | Standard Deviation | Covarience | Corelation |
|---|---|---|---|---|---|
| AirAsia | 5590.26 | 5162 | 2027.36 | | |
| Air India | 9611.21 | 9443 | 3900.95 | | |
| Go air | 5861.05 | 5135 | 2703.58 | | |
| Indigo | 5673.68 | 5000 | 2264.14 | | |
| Jet Airways | 11643.92 | 11467 | 4258.94 | | |
| Jet Airways Business | 58358.66 | 54747 | 11667.59 | | |
| Multiple Carriers | 10902.67 | 10197 | 3721.23 | | |
| Multiple Carriers Premium | 11418.84 | 11269 | 1717.15 | | |
| Multiple Carriers Economy | 4338.28 | 3873 | 1849.92 | | |
| Vistara | 7796.34 | 7980 | 2914.3 | | |
| Total | | | | 1189048 | 0.5069 |

## Hypothesis Testing:

In our exploration of the flight prediction dataset, we set out to examine whether there is a significant difference in the mean ticket prices between short-duration and long duration flights. **The null hypothesis (H0) posited that the mean prices for short-duration flights are equal to the mean prices for long-duration flights. Conversely, the alternative hypothesis (H1) suggested that the mean prices for short-duration flights are different from the mean prices for long-duration flights.** Through statistical analysis, utilizing an appropriate test at a significance level of 0.05, we evaluated the evidence to either accept or reject the null hypothesis, shedding light on the potential impact of flight duration on ticket prices.

we explored the mean flight price, obtaining a sample mean of **$9086.443** with a standard deviation of $44.619. The test, comparing against a null hypothesis

of $9000, yielded a p-value of 0.0527, suggesting insufficient evidence to reject the hypothesis at a 0.05 significance level. The 95% confidence interval for the true mean price is **[$8998.991, $9173.895]."**

## Linear Regression:

We conducted linear regression to understand the relationship between flight prices and the predictor variables—Duration and Total Stops for 10 different airline datasets. The results are as follows:

| Airline | Intercept | Duration | Total Stops |
|---|---|---|---|
| AirAsia | 2692.76 | +1.022 per minute | +1739.50 per stop |
| Air India | 2852.48 | +0.2502 per minute | +2914.06 per stop |
| GoAir | 1945.37 | -8.308 per minute | +4198.64 per stop |
| IndiGo | 2905.37 | +2.302 per minute | +1502.55 per stop |
| Jet Airways | 4574.00 | +0.09361 per minute | +3464.00 per stop |
| Jet Airways Business | 90284.78 | +84.33 per minute | -27686.60 per stop |
| Multiple Carriers | 4369.72 | -1.845 per minute | +3753.26 per stop |
| Premium Carriers | 12392.51 | -1.601 per minute | NA (Not Available) |
| SpiceJet | 2148.58 | +19.22 per minute | -1051.74 per stop |
| Vistara | 2550.49 | -0.4917 per minute | +3824.03 per stop |

| **Airline** | **Intercept** | **Duration** | **Total Stops** |
|---|---|---|---|
| **AirAsia** | **Base price** | **Increase in flight duration adds to ticket price** | **More stops->High price** |
| **Air India** | **Base price** | **Slight increase in price with long duration** | **More stops->High price** |
| **GoAir** | **Base price** | **Longer durations result lower price** | **More stops->High price** |
| **Indigo** | **Base price** | **Longer durations add to ticket price** | **More stops->High price** |
| **Jet Airways** | **Base price** | **Slight increase in price with long duration** | **More stops->High price** |
| **Jet Airways Business** | **Base price** | **Significant increase in price with long duration** | **More stops->High price** |
| **Multiple Carriers** | **Base price** | **slight decrease in price with long duration** | **More stops->High price** |
| **Premium Carriers** | **Base price** | **slight decrease in price with long duration** | **Impact on stops not available** |
| **Spice Jet** | **Base price** | **Longer durations add toi ticket price** | **More stops decrease price** |

| | | | |
|---|---|---|---|
| **Vistara** | **Base price** | **slight decrease in price with long duration** | **More stops increase ticket prce** |