**Project Proposal**

Topic: Customer Segmentation for Flight Bookings

Team Information: Yinan Wang

**Problem Statement**

**Motivation:**

Airlines need to continuously understand customer behavior to enhance and improve customer service and operational efficiency. In order to develop better marketing strategies, provide personalized service, or optimize flight scheduling, airlines need to analyze customer booking and travel patterns.

This project, which aims to group customers based on their booking and travel behavior, can help airlines:

1. Understand key customer segments.
2. Tailor marketing efforts to specific groups.
3. Improve user experience by offering personalized services such as extra baggage, preferred seating.

**Data Introduction:**

This set of data contains about 50,000 customer booking details. The data set includes the number of passengers, sales channels, type of trips, etc. The goal is to use this data to categorize and understand which types of services the customer base prefers.

**Project Scope:**

1. Analyze the dataset using PySpark on the Databricks platform.
2. Conduct exploratory data analysis (EDA) to understand customer behavior.
3. Use the K-Means clustering algorithm to segment customers based on their preferences such as extra baggage, flight meals, seat preference.

**Data scource:**

<https://www.kaggle.com/datasets/praveensaik/proactive-flight-booking-prediction-for-holidays>

**Data Dictionary**

num\_passengers: The number of passengers in the booking.

sales\_channel: The platform used for booking (Internet or other).

trip\_type: Indicates whether the trip is a round trip or one-way.

purchase\_lead: Number of days in advance the ticket was purchased before the flight date.

length\_of\_stay: Duration of the trip in days.

flight\_hour: Hour of the day the flight departs.

flight\_day: Day of the week the flight departs.

route: The origin-destination pair for the flight.

booking\_origin: Country or region from which the booking was made.

wants\_extra\_baggage: A binary variable indicating if the customer opted for extra baggage (1 = Yes, 0 = No).

wants\_preferred\_seat: A binary variable indicating if the customer opted for a preferred seat (1 = Yes, 0 = No).

wants\_in\_flight\_meals: A binary variable indicating if the customer opted for in-flight meals (1 = Yes, 0 = No).

flight\_duration: Duration of the flight in hours.

booking\_complete: A binary variable indicating whether the booking was completed successfully (1 = Yes, 0 = No).