DSCI 510 Final Project Proposal – Flight & Airline Pricing Patterns

1. **Name of the project and team members**
   Yajie(Elaine) Wan
2. **What problem are you trying to solve?**
   How have airline ticket fares changed over time across frequent U.S. routes from 2023 to 2025 (e.g., long-term pricing trends across quarters over the years)? Do some routes or airlines consistently show higher or lower average fares than others?
3. **How will you collect data and from where?**
   - The data will be collected from the U.S. Department of Transportation Statistics (BTS) by web-scraping quarterly fare data from 2023 Q1 through 2025 Q2, with the base URL being "https://transtats.bts.gov/DL_SelectFields.aspx"
   - Use `requests` and `BeautifulSoup` to simulate the BTS download form submission (with first and second-layer filters selected)
   - Loop from 2023 Q1 through 2025 Q2 (latest update) by:
     - Loading the web form through GET
     - Extracting ASP.NET hidden fields
     - Submitting the request via POST
     - Extracting CSV files from ZIP archives
4. **What analysis will you do and what visualizations will you create?**
   - Analysis – 1st question:
     - Create a time index column of quarter index for regression purposes
     - Compute quarterly average ticket fare for each selected route
     - Fit an overall linear regression model (e.g., ticket fare = $\beta 0 + \beta 1 *$ quarter_index $+ \varepsilon$) to display long-term fare trends across all selected carriers and routes
     - Filter to nine well-known carriers in the U.S.
   - Visualization – 1st question:
     - Line chart, with x-axis being the quarter & y-axis being the average fare of each selected route, to show how these routes' fares change over time
   - Analysis – 2nd question:
     - Route-level – for each selected route, compute:
       - Average, median, min, max, standard deviation, and number of observations of its fare
       - Percentage fare difference vs. cheapest-average-fare route
     - Carrier-level – for each selected carrier, compute:
       - Average, median, min, max, std, and count
       - Percentage fare difference vs. cheapest-average-fare carrier
     - Route-Carrier-level – for each selected route, compute:
       - The cheapest & most expensive carrier, as a concatenated data frame
   - Visualization – 2nd question:
     - Route-level:
       - Bar plot, with x-axis being the routes and y-axis being the average fare
       - Boxplot, to show variations and outliers for each route
     - Carrier-level:
       - Bar plot, with x-axis being the carriers and y-axis being the average fare
       - Boxplot, to show the extent of fare distribution for each carrier