DSCI 510 Final Project Report: Flight & Airline Pricing Patterns

Team Member: Yajie (Elaine) Wan | USC Email: wanyajie@usc.edu | USC ID: 5126056352

In this project, U.S. airline ticket pricing patterns from 2023-2025 are analyzed by collecting real airfare data directly from the Bureau of Transportation Statistics (BTS) website. The goal is to identify long-term price trends and compare airfare levels across frequent routes and common carriers, as well as to visualize how the prices change over the quarters of each year. Through this report, my findings on which routes and airlines are more affordable or expensive and how airfares have evolved across recent years will be elaborated in detail.

My initial hypothesis is that the average airfare would gradually increase over the years, and there will be local maximum/peaks during the second and fourth quarters due to vacations and holidays, and the legacy carriers will consistently hold higher airfare than the other budget airlines.

The data is collected from the U.S. Department of Transportation Statistics (BTS) (https://transtats.bts.gov/DL_SelectFields.aspx), and the portion I retrieve consists of quarterly airfare data from 2023 Q1 through 2025 Q2. The method of collection is web-scraping. `requests` is used first to send a GET request to the data selection page, followed by using `BeautifulSoup` to extract the hidden ASP.NET form fields (__VIEWSTATE, __VIEWSTATEGENERATOR, __EVENTVALIDATION), which are required for each valid form submission. Then inside the nested loops over the years and quarters, I send POST request with the hidden fields and desired year, quarter, geography (California), as well as the checkbox fields for my selected columns. Each combination submission returns a zip file that's stored in data/raw, and the CSV inside is then extracted and stored in data/processed in the form of bts_YEAR_QUARTER.csv. There are total of 10 samples, from 2023 Q1 to 2025 Q2, in which 2025 Q2 is the latest updated on BTS' website.

There are some big changes from my original proposal, including the method of data collection, data source selection, as well as research questions. In the beginning, I was neglectful of how the instructions on the data collection should be completed and directly used a dataset found on Kaggle, which wasn't accepted, but I timely switched to web-scraping. Later, during data selection, I struggled to find an ample dataset with enough amount of data, but another dataset I browsed on Kaggle gave me inspiration to look up BTS' website, which was extremely

useful. As a result of changing to BTS' data which doesn't have data for each date of the year, I modified my research question to rather focus on the quarterly airfare data and average airfare data across frequent routes within the U.S. and common carriers: F9 (Frontier Airlines), NK (Spirit Airlines), WN (Southwest Airlines), HA (Hawaiian Airlines), AS (Alaska Airlines), B6 (JetBlue Airways), DL (Delta Air Lines), AA (American Airlines), and UA (United Airlines).

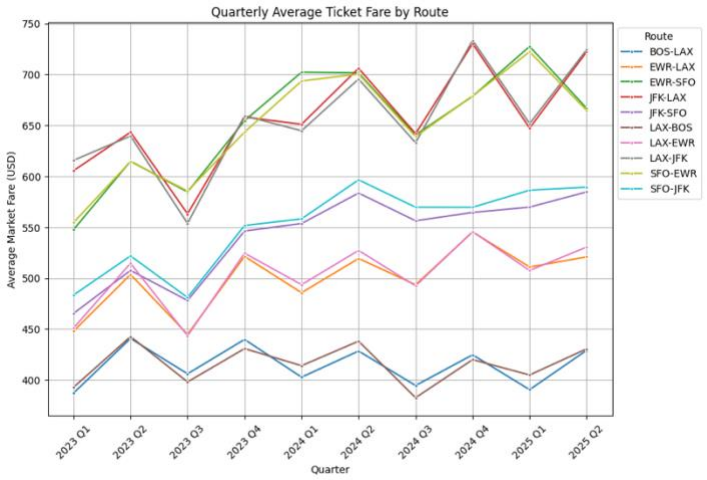
The analysis techniques I used include descriptive statistics (summary tables), group-by aggregations for average calculations, and linear regression for analyzing airfare patterns. The descriptive summaries yield with the help of group-by aggregations could be useful in line plotting for showing the trends of average airfare changes over the quarters of the years (table Quarterly Average Airfare by Route), giving detailed information of each route/carrier to be compared with each other — especially the column that shows the percentage differences against the route/carrier with the cheapest average airfare, and offering straightforward information of which carriers offer the cheapest airfare in general for each frequent route. Lastly, the linear regression model provides the quantitative information of the quarter indices on the long-term trend to indicate the extent of fluctuations in airfare.

From the summary table Route-Level Ticket Airfare, among the top 10 frequent routes (5 pairs/roundtrips), the round-trip route between BOS (Boston) and LAX (Los Angeles) is the cheapest, while the most expensive round-trip route is between JFK (New York) and LAX (Los Angeles), with around 58% higher prices in comparison. All five round-trip routes have their median airfares sit between $310 and $400, which are reasonable because these are all routes between the west and east coast of the US. The minimum airfare recorded could be because of the usage of mileage or discount coupons, while the maximum airfare recorded could be because of the premium-cabin tickets, last-minute bookings, flexible tickets, and multi-city itineraries, etc., within one trip. The number of observations stays fairly consistent for all routes, providing evidence of stable sampling.

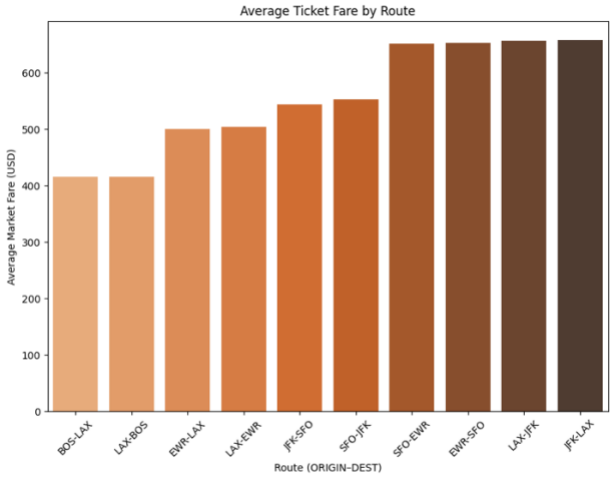
The table of route-carrier level gives us direct information that F9 (Frontier Airlines) is the carrier that offers the lowest price in general, and UA (United Airlines) is the one offering the most expensive airfare, all in general speaking.

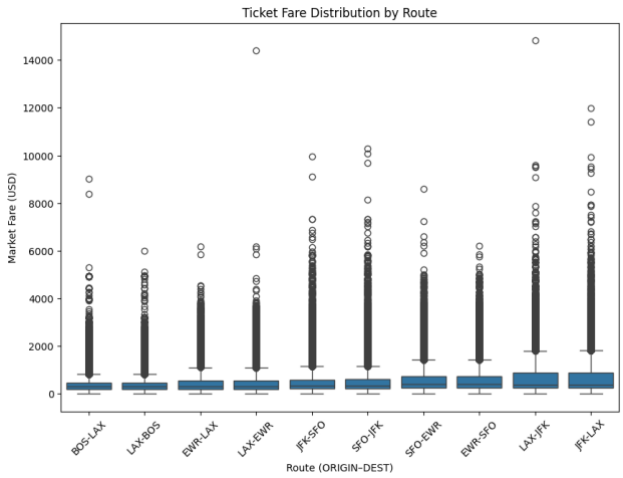
In the summary table of the fitted linear regression, the regression estimates the long-term trend in airfare using the quarter index as the predictor, with the coefficient being 9.5346, meaning that the ticket prices increase by about $9.53 per quarter, which is about $38.12 per

year. The intercept of around $515.19 represents the estimated base fare at the earliest quarter, which is 2023 Q1 in this setting. The p-value is extremely small ($p < 0.0001$), confirming the upward trend, although the increase per quarter is modest. The R-square is very low (0.002), indicating that the linear model only explains a really small portion of the overall variation, which means airfares could vary heavily due to other factors such as demand.

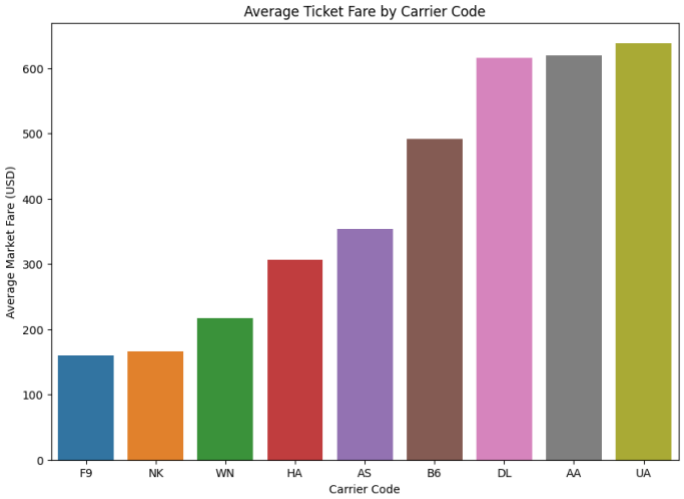


This figure on the left represents the quarterly average ticket fare via line plotting. The quarter names are displayed on the X-axis, while on the Y-axis, the quarterly average fare for each route per quarter is shown, and the data points are plotted as dots on the grid and connected by lines. Since there are five pairs of routes, each pair's line plots follow a similar pattern in terms of airfare changes. One interesting finding is that the second quarter (Q2) of each year is the period during which the average airfare for every route increase, which could be likely due to the time of summer vacations when people book flight tickets to travel around. The fourth quarter (Q4) is also another period that the data points displayed on the plot show as the peak for every year of 2023 and 2024, which is also reasonable because the fourth quarter contains the Thanksgiving break, as well as the Christmas and winter break, when many more people book tickets for family gatherings and travels.

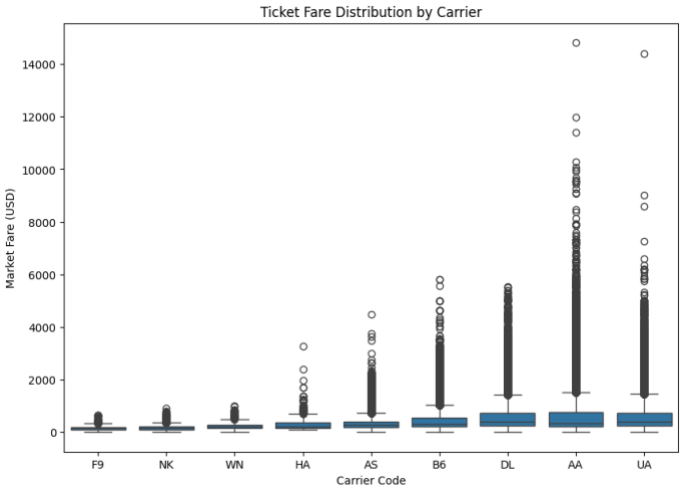




Both the above two figures have their X-axes filled with route names, with average market fare and market fare scatters respectively appearing on their Y-axes. With the visualization of the routes, it's obvious that there are only subtle differences between each pair of round-trip routes, indicating that the case where there is a great airfare difference only by switching the departure and arrival location is often rare and might only differ if the passenger chooses to choose another level of flight ticket, or chooses multi-city itineraries, etc.



This bar plot has the carrier codes displayed on the X-axis and their corresponding average market fares displayed as bars along the Y-axis. From the plot, F9 (Frontier Airlines) has the cheapest average market fare, which matches the previous observation that it offers the lowest price in general. It's also worth noticing that the carrier with the highest average market fare also matches the previous finding that UA (United Airlines) offers the most expensive price in general, and DL (Delta) and AA (American Airlines) have relatively close average market fares.



In this boxplot, same as the above one, the carrier codes fill the X-axis, but market fares are displayed as boxplots along the Y-axis with a lot of outliers. By looking at the IQR (interquartile range) boxes, although they are "suppressed" by the outliers, it's still interesting to see that the well-known budget airlines: F9 (Frontier Airlines), NK (Spirit Airlines), and WN (Southwest Airlines), has much less airfare variations compare to the common legacy carriers: DL (Delta Air Lines), AA (American Airlines), and UA (United Airlines).

The above analysis shows that airfare varies substantially across both routes and carriers. Among all the frequent routes, BOS-LAX (and LAX-BOX) is consistently the cheapest, while JFK-LAX (and LAX-JFK) is the most expensive, which also implies high demands on major business routes. The distinctions between airfare distribution of budget airlines and legacy carriers are revealed by the carrier-level results, in which budget airlines such as Frontier Airlines offer the lowest fares, and legacy carriers, especially United Airlines, show greater price variability — this confirms my initial hypothesis that the legacy carriers will consistently hold higher airfare. The overall linear regression fitted also aligns with my statement earlier that the average airfare would gradually increase over the years, and there will be price peaks during the second and fourth quarters, aligning with major travel periods. These findings help a lot in explaining why airfare varies so widely, offering travelers initial directions of when prices would typically rise, which routes and carriers tend to be cheaper, and giving them a general sense of how the prices vary throughout an ordinary year. These suggestions could assist in more informed travel planning and self-predictions of future airfares.

If given more time on the project, I would first expand it by increasing the dataset to include nationwide routes, rather than only those with either their departure or destination location in California, and by including more airline carriers. Additional improvements might include adding the column `NonStopMiles` to analyze the relationship between airfare and flight distance. It is also worth using more advanced modeling, such as route-specific regressions, to detect whether certain routes' pricing dominates the behavior of a carrier’s average pricing. Lastly, a user interface could be implemented for dynamic searches and comparisons for user-selected time ranges, locations, routes, and carriers.