Project 2 Proposal: Domestic Flight Patterns & Delay Trends in the US Airline Industry

W200 Section 06 Benoit
Team 4: Aashray Puri, Isaac Madera, Raymond Hung

Team Members

- Aashray Puri, Isaac Madera, Raymond Hung

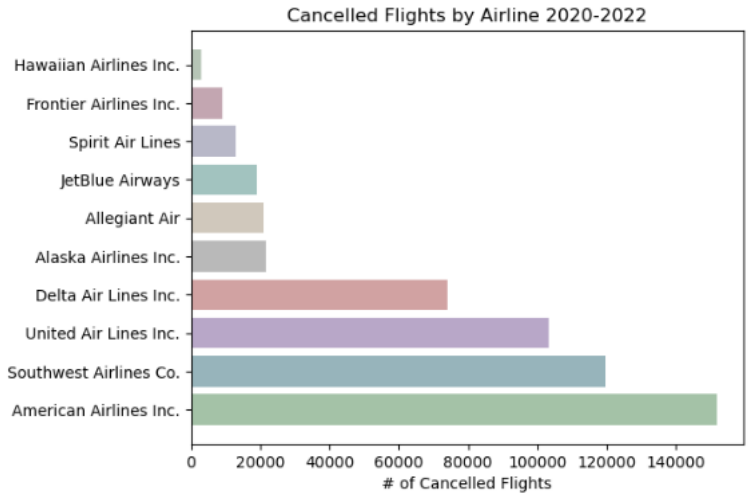
GitHub Repository

- Location: https://github.com/UC-Berkeley-I-School/Project2_Puri_Madera_Hung

Primary Dataset

- Original source appears to be the Bureau of Transportation Statistics
- 5 files for each of the last 5 years (2018, 2019, 2020, 2021, Jan-Sep 2022)
- Approximately 29 million rows (5-9 million per year)
- Columns include airline information, geographic location and delay statistics
- Data source 1 Link [Kaggle]
    - Dataset only has partial data for 2022 (up to Sept 2022), and is missing 2018 data for some airlines. We will account for this in our final report.
- Data source 2 Link [Kaggle]
    - This dataset contains a 2-letter code that maps to a major airline. This table will be useful for mapping the 'Marketing_Airline_Network' field in the Combined_Flights datasets.

Plots, Figures, Tables



Marketing_Airline_Network	delay_percent	total_flights
AA	0.201	1043530
AS	0.180	223287
B6	0.319	156793
DL	0.167	849158
F9	0.312	86557
G4	0.343	73504
HA	0.172	42212
NK	0.258	129208
UA	0.192	742144
WN	0.242	731925

[Left graph] From 2020-2022, American Airlines, Southwest, and United had the most flight cancellations. [Right table] In 2022, G4 (Allegiant Airlines) had the most arrival delays (**as a percentage**) of all domestic airlines. HA (Hawaiian Airlines) has the lowest arrival delay percentage.

Variables & Exploration

We see three distinct opportunities for analysis from this dataset:

1. Airline route patterns
    a. **Relevant columns: Airline, Marketing_Airline_Network, Operating_Airline, FlightDate**
    b. This dataset is a library of all domestic US routes flown by domestic airlines. We believe there are compelling insights on flying patterns and the competitive landscape. For example, which airline flies the most, which airports each airline flies to, monthly seasonality, etc…
2. Delays and cancellations
    a. **Relevant columns: DepDelay, ArrDelay, DepTime, ArrTime, Distance, Canceled, Diverted**
    b. This dataset has detailed records on departure and arrival delays. We plan to evaluate among other things which airlines are responsible for the most delays and cancellations? Any time of day or time of year correlation we can identify?
3. Geographic / weather
    a. **Relevant columns: FlightDate, Origin, Destination, OriginStateName, DestinationStateName**
    b. We believe we can connect our dataset to historical weather data for high-volume airports, to understand the impact weather has on flights. For example, do summer storms or winter snow have a heavier impact on operations?

Supplemental Datasets

- Weather Data by Airport:
    - We plan on evaluating the impact of weather on delays and cancellations. It has information on snow, rain, thunderstorms by airport, by date. We would join this to our flight dataset **by airport code and date**.

Final Report Plan

We plan to cover the areas specified in the Variables section and go deep on the insights. We will organize the work as follows:

1. Identify flying patterns:
    a. [Airline] Which airlines have the largest flight networks?
    b. [Airline] Airline airport hubs & destinations
    c. [Seasonality] Flight patterns by monthly, day of week, etc.
    d. [Trends] Did flight patterns change post-COVID?
2. Evaluate delay & cancellation data:
    a. [Airline] Best and worst performing on-time airlines
    b. [Airline] Difference between legacy (American, Delta, United) and low-cost?
    c. [Geography] Are certain airports more susceptible to delays?
    d. [Seasonality] Delay distribution and seasonality (hourly, day of week, monthly)
    e. [Trends] How have delay patterns shifted from pre vs post COVID?
3. Compare to weather data
    a. Impact of weather on delays or cancellations to major airports