SMOOTHFLY

MATTHEW BLAKE, SOHAM BUTALA, AADITYA PARTHASARATHY, FRANCES SCOTT-WEIS

PROJECT DESIGN

Background

- Covid-19 led to an increase in delays, cancellations, and other disruptions to air travel
- Experiencing such disruptions adds stress and uncertainty to traveling which is already heightened due to the pandemic
- The Bureau of Transportation Statistics (BTS) holds a wealth of information on flight history in the US, but the data is largely inaccessible and lacks interpretability

Project Goals

- Increase the accessibility of flight data from the BTS to allow users to gain insights into future travel plans with their own analysis
- Provide interpretable analysis of key factors that a user may consider when booking air travel:
  - Fares
  - Denied boarding
  - Delays

DATA USED

- Airline On-time Statistics
  - This data source was used for the Data Pipeline and provides an interface for selecting a variety of fields and obtaining data on relevant flight routes
  - Limitation: Data must be manually queried by selecting airline, origin/destination city, and time frame, thus there is no way to obtain sets of data for more than one carrier or route in one query
- Airline On-Time Statistics and Delay Causes
  - This data was used for the Delays visualizations and contains information about the flight routes and carriers which were delayed for more than 30 minutes aggregated by month.
  - Limitations: Delays dataset did not contain information on flights by quarters as the consumer fares dataset did, thus preventing us from accounting for seasonality. Delays also only contained information on arrival delays, not departure delays.
- Domestic Airline Consumer Airfare Report
  - This data was used for the Fares visualizations and contains information on fare prices of short and long-haul flights in the US aggregated by quarter
  - Limitation: Carrier information is not available
- Passengers Denied Confirmed Space Report
  - This data was used for the Denied Boarding visualizations and contains information on involuntary and voluntary denied boarding by carrier that is aggregated on a quarterly level
  - Limitations: Formatting for years prior to 2018 is inconsistent, column definitions and airline naming conventions are inconsistent, and data on flight route is not available

All the data for this project was obtained from the Bureau of Transportation Statistics.

USE CASES

Data Pipeline Use Case

- User Objective: A user needs domestic airlines data to work on any analysis related to flights arrival, departure, and delay
- Expected Interaction:
  - The user will select any or all the parameters like source airport, destination airport, and airline carrier.
  - The tool will then fetch the data from the Bureau of Transportation Statistics and provide the data in multiple formats based on the user's chosen criteria.

Visualization Use Case

- User Objective: A user needs to travel from LA to NYC to participate in an important conference. Due to work conflicts, they will need to fly out on the same day as they are presenting and thus, they want to take all possible steps to avoid disruptions to their intended travel plans. They are specifically interested in which airline will provide them the most security when booking.
- Expected Interaction:
  - The user will select denied boarding as their factor of interest.
  - The tool will output visualizations with insights into denied boardings over time and by carrier.
  - The user will use the interactive tools to help choose which carrier they should fly with.

DESIGN

Data Pipeline

Purpose: Scrapes data from BTS to create full dataset of flight routes

Input: BTS departures website

Output: Dataset of all flight route information available through BTS

Interactions: Dataset is queried by the web app to produce datasets for users that match their specifications

Data Manager

Purpose: Holds cleaned datasets containing information on denied boarding, delays, and fares

Input: Raw datasets from BTS

Output: Cleaned datasets for denied boarding, fares, and delays

Interactions: Output is used by the Visualization Manager to create figures

Visualization Manager

Purpose: Produces visualizations to be output by the web application

Input: Datasets on delays, fares, and denied boarding from data manager

Output: Collection of functions to produce visualizations

Interactions: Datasets produced by the Data Manager as used as input and the function outputs are used by the Website to produce figures

Website

Purpose Provides accessible and interpretable flight data to users

Input: Dataset created by the data pipeline, user input on specifications for their dataset, and visualization functions

Output: Datasets filtered to users' specifications and dashboard of visualizations for each variable being considered, delays, fares, and denied boarding

Interactions: Functions from the Visualization Manager are used to output dashboards and output from the Data Pipeline is queried to obtain filtered datasets

DEMO

LESSONS LEARNED AND FUTURE WORK

- Lessons Learned
  - Using logging and error handling techniques
  - Handling edge cases when scraping data from a website
  - Working with a variety of datasets, even from the same source, necessitates a substantial amount of data cleaning and processing
    - Datasets were often incomplete i.e. they contained discontinuous data collected at various timepoints
    - Needed to reformat/transform variables in the datasets in order to create our visualizations

- Future Work
  - Expand dashboard to query data from the pipeline to allow for further analysis and specifications from a user
  - Automate the process for our dashboard to query the pipeline to ensure that our visualizations are up-to-date.
  - Incorporate forecasting modeling into our dashboards i.e. predict the likelihood of a user being denied boarding or their flight being delayed
  - Our datasets only contained data on flights within the United States. We could expand our dashboard to include flights outside of the states.