STAT 628 Airline Performance

2025-03-31

Project Introduction

The period between Memorial Day (the last Monday of May) and Labor Day (the first Monday of September) is the peak summer travel season in the US, making it one of the busiest times for the airline industry. Recognizing important patterns in flight delays and cancellations can help passengers mitigate potential disruptions to their travel plans. In this project, you will use data from the U.S. Department of Transportation to **discover meaningful patterns about delays and cancellations during the summer travel season.** You will leverage these data to build a Shiny application that allows users to input relevant flight details (e.g., origin & destination, airline, scheduled departure & arrival times, etc.) and then receive predictions (and uncertainties) about:

- Whether their flight will be cancelled
- Whether their flight will be depart on-time and, if not, the length of the departure delay
- Whether their flight will be arrive on-time and, if not, the length of the arrival delay The app must also include some visualization of these predictions and the associated uncertainties.

Data Resources

Every month, the U.S. Department of Transportation compiles data from U.S. commercial, non-chartered passenger flights and computers key statistics about the on-time performance of each flight. These monthly data can be downloaded from this API. Because downloading these data can take a long time (~5 minutes for a single month's worth of data), we will share compressed files containing data from the last several summers on Box. The data is available here: https://uwmadison.box.com/s/u5gmh68dz8rar24q1zckzy2p9r0iuz27.

Note, you will need to sign in with your UW credentials to access the files. After downloading each compressed archive, you will need to un-zip the archive and manually rename each un-zipped CSV.

Descriptions of all the fields in the dataset are available here.

It may also be useful to utilize the list of all U.S. airports (including their location), which can be downloaded here.

We will additional use information about airport elevantion and time zone. CSV files containing this information will be posted on Canvas.

Project Timeline

Scoping

A key first step in your analysis will be to define its scope, in terms of space, time, and airline. You could limit your analysis to flights between a small subset of cities or other geographic regions; flights operated by a select subset of airlines (e.g., American Airlines or Southwest Airlines); or flights operated within a certain window (e.g., all flights from summer 2024).

At minimum, you need to work with at (i) least 3 months of data; (ii) at least 2 airlines; and (iii) flights between at least 10 different cities. Of course, you are welcome to use more data.

Data Preprocessing

The Department of Transportation provides flight data broken down by month. Among many other things, the data records the schedule departure, actual departure, schedule arrival, and actual arrival times of every flight. Unfortunately, these times are recorded in the local timezone of the origin (for departure) and destination (for arrivals) airports.

Red-eye flights introduce additional complications: they may depart from their origin late at night and arrive at their destination early in the morning. During lecture, we will discuss how to convert all times to a single timezone (UTC) and other important pre-processing steps. You must then repeat these pre-processing steps on each data file you include in your analysis.

Modeling

Once you have pre-processed the data, you may want to build different predictive models. There are several possibilities. You could, for instance, build a single model to predict the actual gate arrival time based on the scheduled itinerary. You could alternatively build separate models to predict cancellation, delays (both departure and arrival), and the actual amount of delay (both departure and arrival).

At a minimum, your models should take as input the origin and destination airports, the airline, and the scheduled arrival and departure times. You can leverage additional inputs from the Department of Transportation data. You are welcome to incorporate additional, outside data but these data must be well-documented in your technical report and must be publicly available.

Building an application

In many companies, data scientists are expected to make "actionable'' prototypes/products based on their data analyses. To mimic this practice from industry, you will create a web-based application that will demonstrate your analysis in real-time.

Shiny is an easy-to-use, R-based platform to turn your R code into a web application. While you do not have to use Shiny (if you have app development experience, feel free to use other languages/platforms!), all applications must run on the latest Chrome browser and be accessible to the teaching staff for grading. For more information about Shiny, visit: https://shiny.posit.co/.

We'll leave the user-interface and other graphical specifications of the Shiny app up to you. But, the Shiny app must contain the following elements, at a minimum:

- An interface for users to enter the origin and destination airports, the airline, and the scheduled departure times for a particular flight.
- Predictions for cancellation and any dealys (both departure and arrival) along with uncertainties
- Useful, interactive visualization about your predictions and the associated uncertainties

Project Deliverables

You will upload two documents to Canvas by 11:59pm Central Time on Friday April 25: a one-page executive summary and a technical report that includes all relevant code. In addition to these items, you will give an in-class presentation in Week 13.

Executive Summary

The executive summary should present your conclusions and be free of technical jargon, figures, graphs, and R code. In other words, you should describe and interpret your results and should not describe the process by which you obtained your results. The executive summary should convey the main conclusions of your modeling efforts in language that is accessible to someone who may not have taken STAT 628 before. Your executive summary should

- Clearly define the scope of your analysis

- Briefly describe the statistical model you built
- Describe the main drivers of flight cancellation and on-time (or early) arrivals
- Provide between 1 and 3 easy-to-understand recommendations for passengers about how best to avoid cancellations and delayed arrivals.
- Briefly discuss the limitations of your analysis.

Your executive summary should not be longer than one page. Do not adjust the page margins or use an extremely small font size to achieve this. It should be free of technical jargon and make minimal use of equations. Your executive summary will be graded on a 0–5 point scale based on the extent to which it delivers the above requested items in a concise and clear fashion.

Technical Report

The technical report is meant to convey to your peers that your analysis was sound. It is not, however, meant to be a step-by-step chronology of what you did to arrive at your final model. Instead, you should summarize the most important steps of your modelling process.

You are strongly encouraged to prepare your technical report using RMarkdown. While this allows you to include all relevant R code, you should take care not to include excessive output in your report. You should also ensure that code and output do not extend beyond the page margin and that your figures are appropriately sized. For each plot that you include, you must explain its relevance in the exposition. See Section 5.3 of the RMarkdown Cookbook for information about controlling page margins and Section 5.4 of the same book for information about sizing figures. **The instructor and TA will run the code that you submit and will deduct marks if they are unable to reproduce the results of your analysis.** You should divided your Technical Summary into several sections, one for each phase of the project.

Your technical summary will be graded on a 0–5 point scale based on its clarity and style and the degree, the technical soundness of your analysis, and the extent to which it includes the requested items.

The instructor and TA will also run the code provided in the technical summary. If they able to run the code successfully and reproduce all reported numerical summaries and visualizations, you will receive 5 points. If they are able to reproduce most but not all of those results, you will receive between 3 and 4 points and if they are unable to less than half of the results, you will receive between 1 and 2 points for the code.

Presentation

You will give a short presentation demonstrating your application to the whole class. Further details about the presentation will be shared later in April.