Flight delay Analysis using Data Science

Xin Tang

Bellevue University

Requirement:

1. Exploratory data analysis, transformations, and summary statistics on the data via R
2. A recommendation is required for a model or method you would implement to solve the problem.
3. Final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps.

INTRODUCTION

I am a frequent business travler. Flying is an unavoidable part of my travel. Flight delays and cancellations are annoying and a big waste of time. In a recent business trip, due to airline factor, my flight was delayed 1.5 hours from planned departure time; Furthermore, due to that delay, we run into severe weather condition at destination airport disrupted the normal airport operation, as the result, I had another 2 hours wait in taxiway after landing to get a gate assigned.

By using data science, I would like to conduct research on some available airport and airline flight performance data, I hope I could draw some conclusions using historic data, or even create a model, which could pick/recommend flights, airports, or airlines to reduce the risk of flight delay or cancellation.

1. Get dataset 1 for flight info:
    a. Find all flights from and arrived at AUS, my home airport.
    b. Filter the dataset to include only north CA as destinations, which I used most often.
    c. Include delay time, on time flag, cancellation, reason code of cancellation. Month and day, airline code, also include mileages.
    d. Several subset dataset will be created to assist the analysis.
2. Get dataset 2 for airfare price!
3. Get dataset for airline cancellation for additional analysis.

RESEARCH QUESTIONS

Based on my travel pattern, I can choose departure region and destination area. I would like to answer the following questions, the question list may change along my research process :

1. Which airport has more flights choice to destinations (area) I fly frequently?

2. Which airport has best record in term of delay?
3. Which airport has less cancellation record?
4. Which month is the best month with the least delays?
5. Which airline has the least chance of delays.
6. Rason of delay (weather, airline, air control etc.)
7. Which flight has the best on time record?
8. Which flight has the worst record of flight cancellation?
9. On average, which airport offers more affordable price?
10. Which flights can give me the most mileage points, based on mileage.
11. Which flights can give me the most mileage points, based on price.
12. Based on the above data, Could I find a best route?

APPROACH

With an accessible dataset, I will use the R knowledge gained from the DSC520 class and do some analysis. The analysis includes charts, statistical analysis and, if possible, a prediction using proper models.

There are major steps will be taken:

1. Find proper datasets and import data into R.
2. Tidy the data, examine the data structure on each dataset.
3. Transform the data, pick up useful data, create new datasets. Perform the analysis
4. Visualize the data.
5. If possible, create model or create a wish list of future research.
6. Draw conclusions.

How your approach addresses (fully or partially) the problem

Based on the data available, the chart will give a visual indication, the statistical analysis will give a mathematical analysis of the findings.

DATA:

Based on limited flight dataset for free public access, I will use the following datasets I could find:

1. Flight dataset on January 2023 for all Major US airports and airlines. - Bureau of Transportation Statistics

   https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGK
2. Quarterly flight airfare dataset in year 2022 - Bureau of Transportation Statistics, https://www.transtats.bts.gov/AverageFare/
3. Airline on time data for last 5 years. Data is from Bureau of Transportation Statistics

https://www.transtats.bts.gov/Tables.asp?QO_VQ=EFD&QO_anzr=Nv4yv0r%FDb0-gvzr%FDcr4s14zn0pr%FDQn6n&QO_fu146_anzr=b0-gvzr

Required Packages (preliminary)

- ggplot2
- dplyr
- tidyr
- writexl

Plots and Tables (preliminary)

- histogram chart
- scatter chart
- boxplot

Limitation / Questions for future steps (to be defined)

- How does this study could expand to other airports/destinations?
- How to add connection flight factors.