Flight delay Analysis using Data Science

Xin Tang

Bellevue University

(red text is update for project step2)

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, Get data ready and then analysis. the chart will give a visual indication, the statistical analysis will give a mathematical analysis of the findings.

- Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.
    1. Download the data. my original downloaded flight data is close to 260mb, too big to be handled by github or R. Besides, it contains airport info which I am intended to use,

       For my second set of data, the flight cancellation data. It had detail data and summary in the same tab, also the column is not a variable but combination of multiple variables. So the cleaning steps involved are:
        - Check any NA (empty cell) observation, remove if necessary.
        - Remove the column with duplicate information.
        - Select only interested origin and destination. Remove all info from airport I am not planning to analysis.
        - For cancel data, separate the detail and summary data into different tab.
        - Transform the data so each column only have one variable and each row is one observation.
    2. Show how the final data set looks like (2 different data sets shown)

Dataset 1: flight data

	Year	Quarter	Month	DayofMonth	DayOfWeek	FlightDate	Operating_Airline	Tail_Number	Flight_Number_Operating_Airline	Origin	OriginCi
1	2023	1	1	31	2	2023-01-31	AS	N459AS	534	SFO	San Franc
2	2023	1	1	31	2	2023-01-31	AS	N302AS	512	SFO	San Franc
3	2023	1	1	30	1	2023-01-30	AS	N491AS	534	SFO	San Franc
4	2023	1	1	30	1	2023-01-30	AS	N471AS	512	SFO	San Franc
5	2023	1	1	29	7	2023-01-29	AS	N464AS	534	SFO	San Franc
6	2023	1	1	29	7	2023-01-29	AS	N298AK	512	SFO	San Franc
7	2023	1	1	28	6	2023-01-28	AS	N553AS	534	SFO	San Franc
8	2023	1	1	27	5	2023-01-27	AS	N535AS	534	SFO	San Franc
9	2023	1	1	27	5	2023-01-27	AS	N492AS	512	SFO	San Franc
10	2023	1	1	26	4	2023-01-26	AS	N403AS	534	SFO	San Franc
11	2023	1	1	26	4	2023-01-26	AS	N518AS	512	SFO	San Franc
12	2023	1	1	25	3	2023-01-25	AS	N448AS	534	SFO	San Franc
13	2023	1	1	25	3	2023-01-25	AS	N318AS	512	SFO	San Franc
14	2023	1	1	24	2	2023-01-24	AS	N319AS	534	SFO	San Franc
15	2023	1	1	24	2	2023-01-24	AS	N283AK	512	SFO	San Franc

Dataset2: cancel rate

	Year	Month	Month_factor	Airline	cancel_rate
1	2022	1	1	Alaska	0.0628896903
2	2022	2	2	Alaska	0.0145088050
3	2022	3	3	Alaska	0.0128074423
4	2022	4	4	Alaska	0.0377584934
5	2022	5	5	Alaska	0.0221364985
6	2022	6	6	Alaska	0.0067613653
7	2022	7	7	Alaska	0.0043720191
8	2022	8	8	Alaska	0.0053229545
9	2022	9	9	Alaska	0.0065423618
10	2022	10	10	Alaska	0.0041548999
11	2022	11	11	Alaska	0.0136707487
12	2022	12	12	Alaska	0.0763461728
13	2023	1	13	Alaska	0.0137275607
14	2023	2	14	Alaska	0.0269904090
15	2023	3	15	Alaska	0.0097891325
16	2023	4	16	Alaska	0.0090912022
17	2022	1	1	Allegiant	0.0830846913
18	2022	2	2	Allegiant	0.0568496310

Dataset 3 flight price data per quarter

	Quarter	Airport_Code	Airport	City	State	Average_Fare	Adjusted_Average	Passe
1	Q1	AUS	Austin - Bergstrom International	Austin	TX	317.4099	332.3620	
2	Q1	SFO	San Francisco International	San Francisco	CA	364.0607	381.2102	
3	Q1	SJC	Norman Y. Mineta San Jose International	San Jose	CA	286.2944	299.7807	
4	Q1	OAK	Metro Oakland International	Oakland	CA	261.9732	274.3138	
5	Q2	AUS	Austin - Bergstrom International	Austin	TX	376.7318	383.0860	
6	Q2	SFO	San Francisco International	San Francisco	CA	498.5494	506.9582	
7	Q2	SJC	Norman Y. Mineta San Jose International	San Jose	CA	349.9405	355.8427	
8	Q2	OAK	Metro Oakland International	Oakland	CA	303.1360	308.2489	
9	Q3	AUS	Austin - Bergstrom International	Austin	TX	364.7552	366.0945	
10	Q3	SFO	San Francisco International	San Francisco	CA	456.8363	458.5136	
11	Q3	SJC	Norman Y. Mineta San Jose International	San Jose	CA	317.8025	318.9693	
12	Q3	OAK	Metro Oakland International	Oakland	CA	284.1627	285.2060	
13	Q4	AUS	Austin - Bergstrom International	Austin	TX	378.9604	378.9604	
14	Q4	SFO	San Francisco International	San Francisco	CA	476.8593	476.8593	
15	Q4	SJC	Norman Y. Mineta San Jose International	San Jose	CA	335.1386	335.1386	
16	Q4	OAK	Metro Oakland International	Oakland	CA	295.3193	295.3193	

3. What do you not know how to do right now that you need to learn to import and cleanup your dataset?
    - how to import large data file but avoid crash the system.
    - How to turn dataset into standard dataset(row=observation, column=variable)
4. Discuss how you plan to uncover new information in the data that is not self-evident.
    - Mainly use dplyr package to divide and form new dataset, each serve one purpose.
    - Using plot to help visualize the data so to decide next step and do further data transformation (like remove outliner etc.)
5. Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
    - See answer related to question 4. I will slice data till it serves only one purpose. Like calculate which airline had most flight, most delay etc.
    - Also I will join different columns to form a more focused dataset, to serve the purpose I want.

6. How could you summarize your data to answer key questions?
    - Summarize() will be used frequently to get average data. Like airline average cancellations rate per month.
7. What type of plot to help you to illustrate the findings?
    - Scatter plot with color as second filter
    - Histogram chart
    - Box chart
8. What do you not know right now may help you to answer the question? Or what info is not self-evident?
    - For now I am not sure if I should build a logic model (like if the flight will be delayed) or a liner model (to predict how much delay). And if I have enough or proper data to do so.
    - Second all average, max. min info are not self-evident, so I need to use data manipulate (group by, filter, mutate etc) to get to a result.
9. Do you plan to use any machine learning techniques to answer your question?
    - For now, at week9, I am not planning to. First I do not know any machine learning skills. Second, I only focus on a small portion of dataset. If I would like to expand my study to a national area, or incorporate any transfer flight, I may need to.
10. Question for future steps.
    - Do I have enough data (data type) to build a regression model
    - Does my analysis could apply to a larger scope (like national flight level)

- How much data can R handle from a personal computer (since my 200Mb data frozen my program for at least 1 hour)

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.