Flight Analysis using Data Science

Xin Tang

Bellevue University

INTRODUCTION

I am a frequent business traveler. 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.

PROBLEM STATEMENT

By using data science, pick a flight route and time to travel between Austin, TX and Bay-area, California. Also analysis the airfare data to find the fare variation among months.

HOW TO ADDRESS THE PROBLEM

There are plenty of dataset available from Bureau of Transportation Statistics. I downloaded 3 dataset, flight info data, cancellation dataset and airfare dataset. From the datasets, I will use the R knowledge gained from the DSC520 class and do some analysis. The analysis includes charts, statistical analysis.

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. Draw conclusions.

Analysis

Below are the analysis conducted. Originally script is in R markdown. It can run in R, however, I run into a knit error as *yaml.load(..., eval.expr = TRUE)* : *Duplicate map key: 'output'*, which I do not know how to fix, so copied the output in here. Please refer to repo location below for original R file and rmd file:

https://github.com/tang-xin72/DSC520/tree/main/scratch/project

- get_hflight.R: the R code to analysis the flight info, (I am not using hflight dataset)
- flight -cancel.R: the R code to analysis the flight cancel data
- fare-check.R: the R code to analysis the flight airfare data
- final_xin_tang.rmd: R markdown file

see following page for R mark down output

Title: "DSC520 final"

author: "xin tang"

date: "`r Sys.Date()`"

output: html_document

output:

  pdf_document: default

bibliography: flight_citations.bib

editor_options:

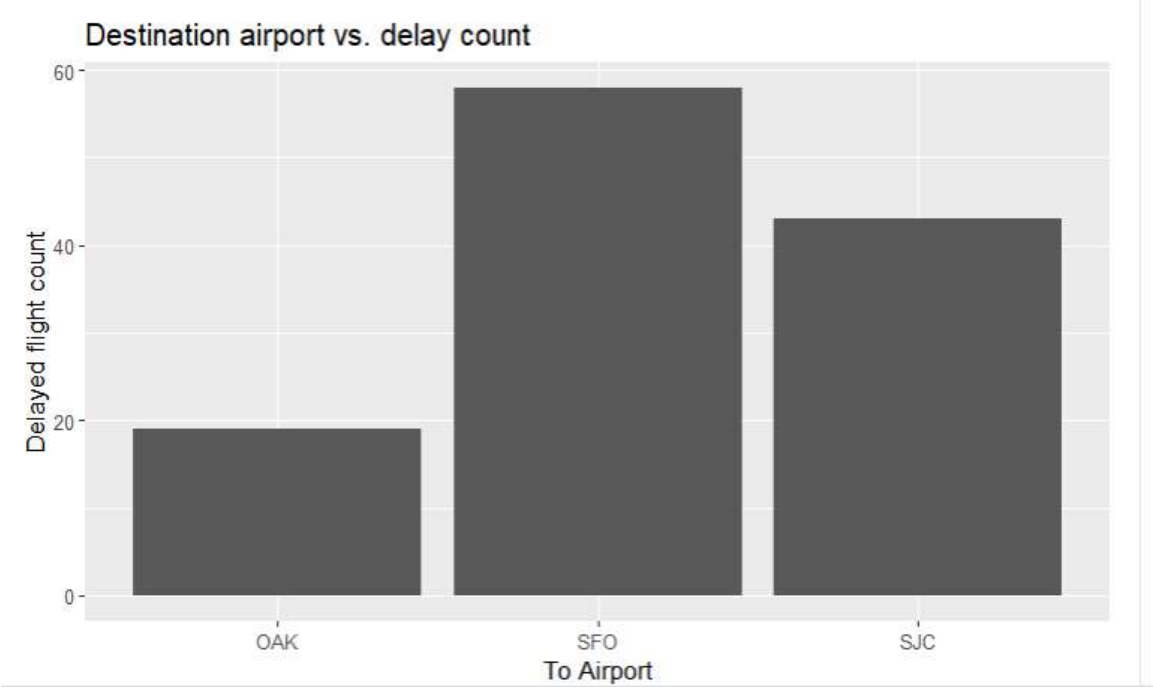
  markdown:

    wrap: 72

My home airport is Austin, Texas (AUS), due to work, I fly to bay area in California frequently. I can choose from 3 airprots to use: San Francisco airport (SFO), SanJose airport (SJC) and Oakland airport(OJC). I am interested to find how to pick the flight out and back flight to avoid delay, with least cancellation risk and possible best fare.I only studied direct flight since it will be most time saving options.

First I will do an analysis on delay using flight data.

Dest <chr>	delay_count <int>	flight_count <int>	delay_percent <chr>
OAK	19	31	61.29%
SFO	58	154	37.66%
SJC	43	112	38.39%



Now analysis delay condition from those airports

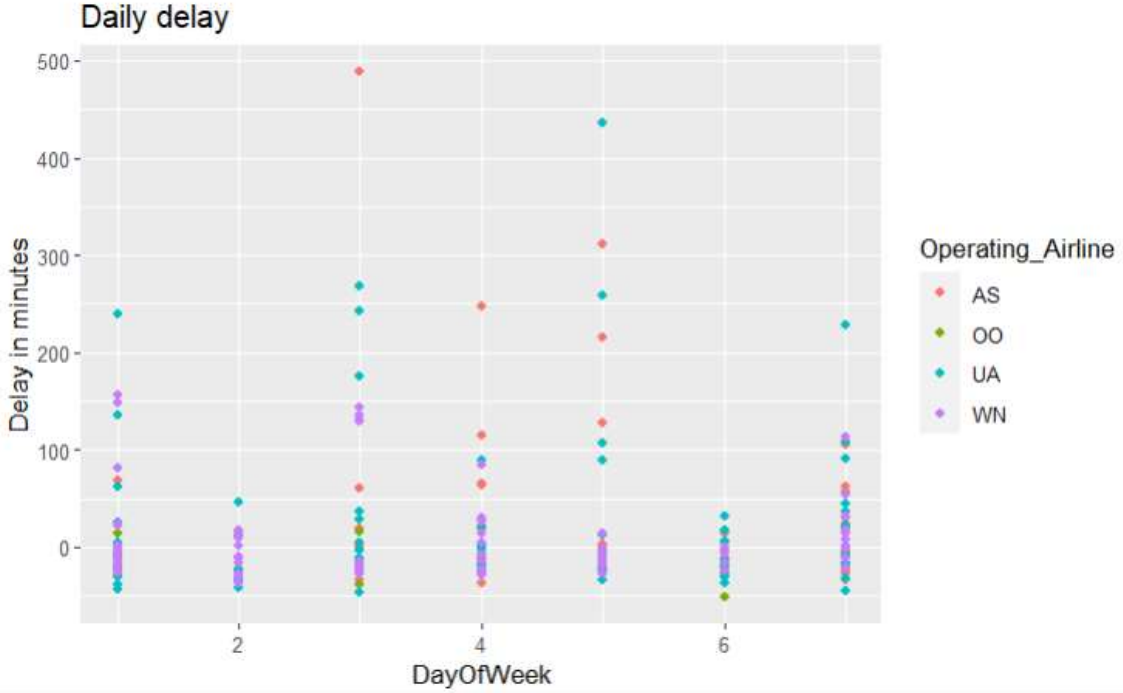
Origin <chr>	delay_count <int>	flight_count <int>	delay_percent <chr>
OAK	21	31	68%
SFO	58	154	38%
SJC	30	112	27%

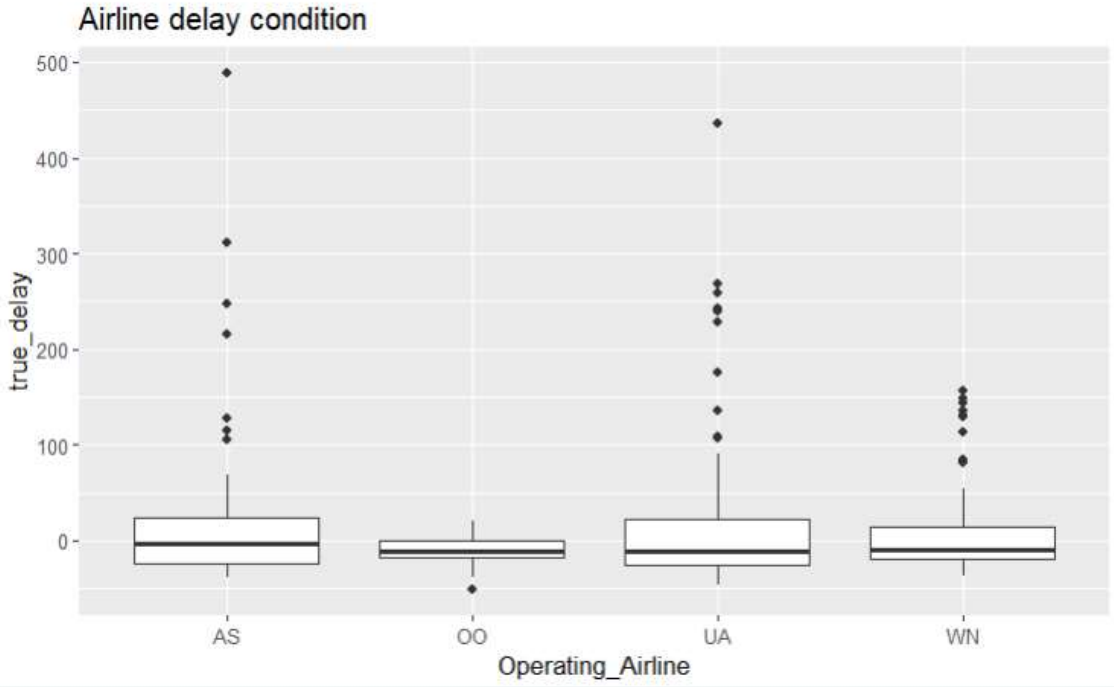
From this analysis, no matter as Origin or Destination airport, OAK airport have twice as much delay as other 2 airports. so it is not a good choice to fly into

The rest analysis will not include OAK, since it is not a preferred place to fly to/from.

Now look at the delay from other 2 aspects(by daily and by airline).

looks like Friday may have more chances to delay but every Airline have similar chance of delay


Daily delay
Delay in minutes
500
400
300
200
100
0
2
4
6
DayOfWeek
Operating_Airline
AS
OO
UA
WN


Airline delay condition
true_delay
500
400
300
200
100
0
AS
OO
UA
WN
Operating_Airline

Excluding early arrival, now answer the question: who may have least delay.

Airline OO (skywest) has least delay and rest are about same when fly into Austin.
Airline UA (United) has worst delay when fly into bay area.
Fly into and from SJC is always a good choice.

Operating_Airl... <chr>	Dest <chr>	avg_delay <dbl>
AS	SFO	78.50000
OO	SJC	75.45455
UA	SFO	132.68966
WN	SJC	59.93548

Operating_Airl... <chr>	Origin <chr>	avg_delay <dbl>
AS	SFO	80.07407
OO	SJC	12.00000
UA	SFO	83.85714
WN	SJC	52.07143

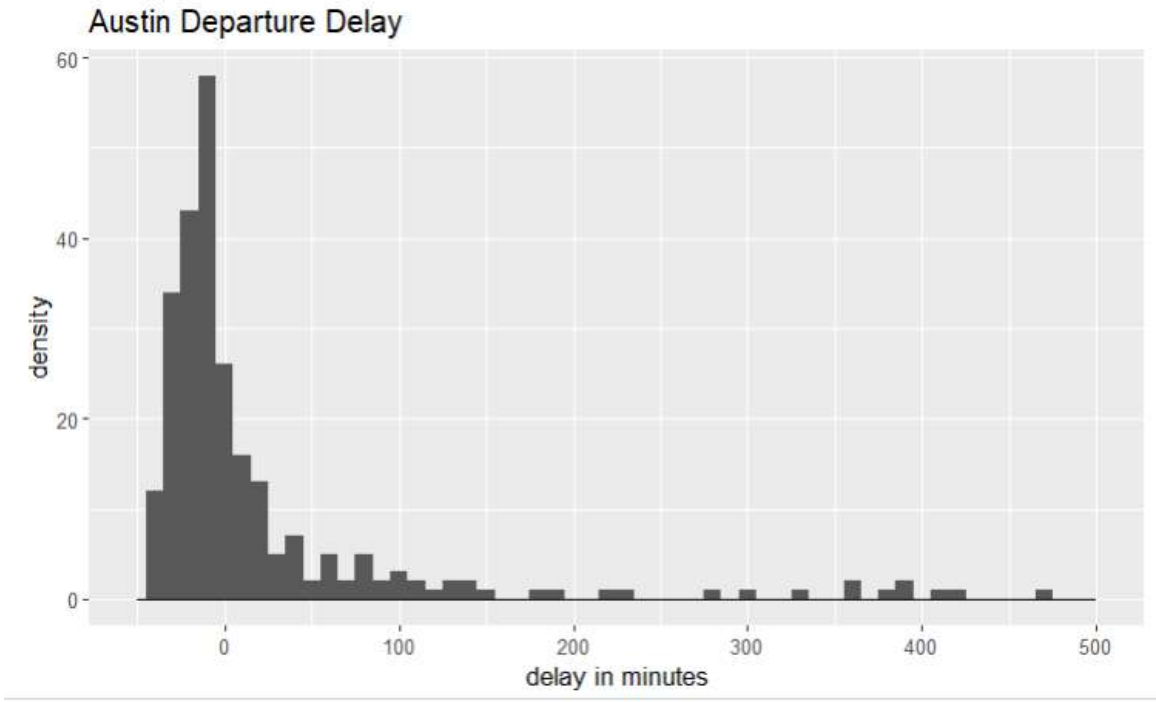
Check which delay factor is the most significant when fly to bay area

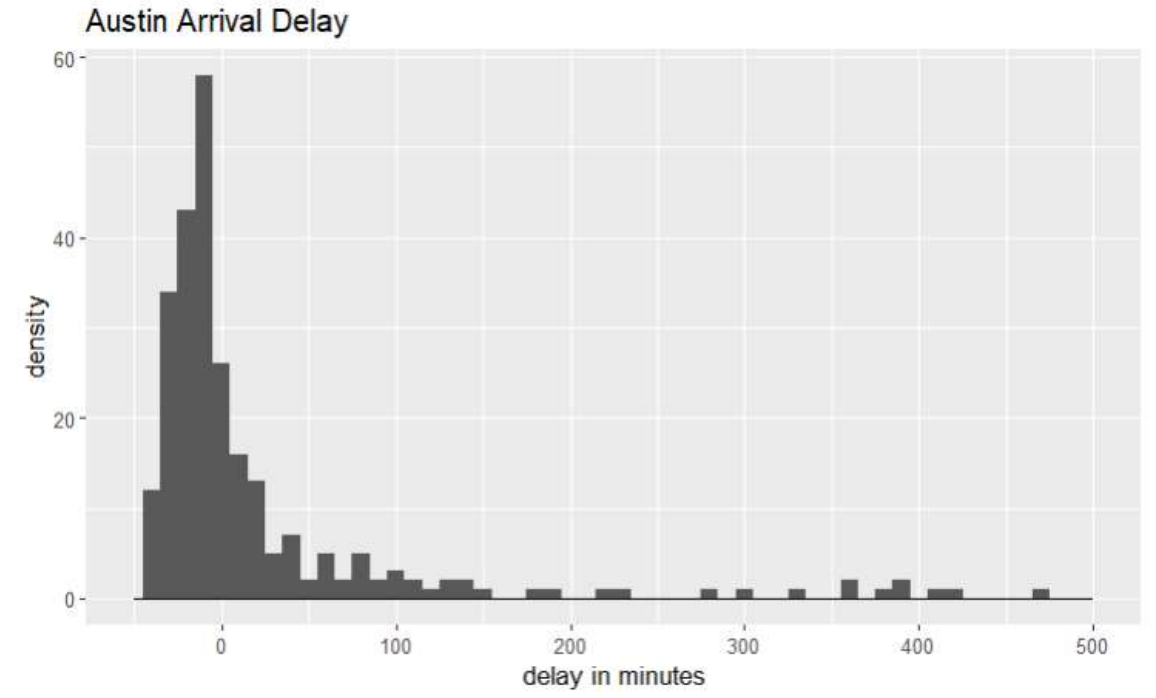
it shows the prior flight delay and control tower caused most delay, which I can not control

Dest <chr>	carrier_delay <dbl>	weather_delay <dbl>	NAS_delay <dbl>	Security_delay <dbl>
SFO	7.947368	0.1842105	41.473684	0
SJC	14.315789	1.0000000	5.315789	0

Origin <chr>	carrier_delay <dbl>	weather_delay <dbl>	NAS_delay <dbl>	Security_delay <dbl>
SFO	26.76316	0	8.868421	0
SJC	5.80000	0	1.266667	0

Get histogram of delay, inbound and outbound each, check if there are normal distribution.





Now check correlation between flight time and different delay factors. using flight data 'Austin to SFO'

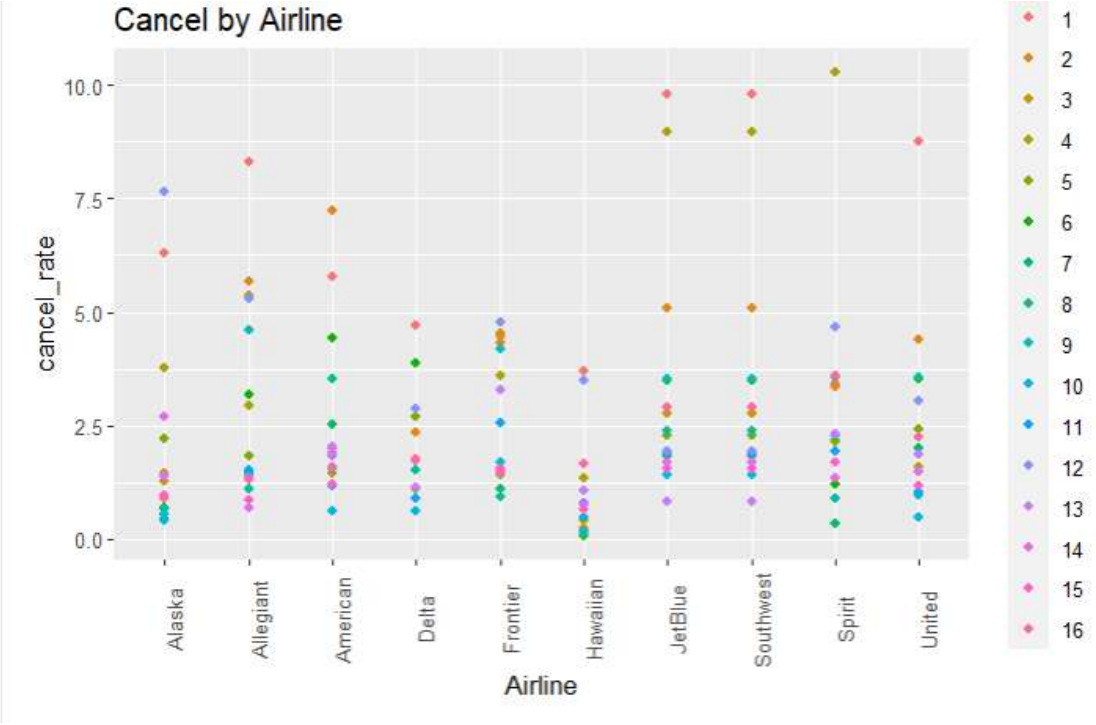
Again the air control caused delay is more important factor

```
                   ActualElapsedTime CarrierDelay   NASDelay LateAircraftDelay
ActualElapsedTime          1.0000000   -0.2300359  0.8807402        -0.3607977
CarrierDelay              -0.2300359    1.0000000 -0.2124367        -0.1899233
NASDelay                   0.8807402   -0.2124367  1.0000000        -0.3655875
LateAircraftDelay         -0.3607977   -0.1899233 -0.3655875         1.0000000
[1] 0.8807402
[1] -0.2300359
[1] -0.3607977
```

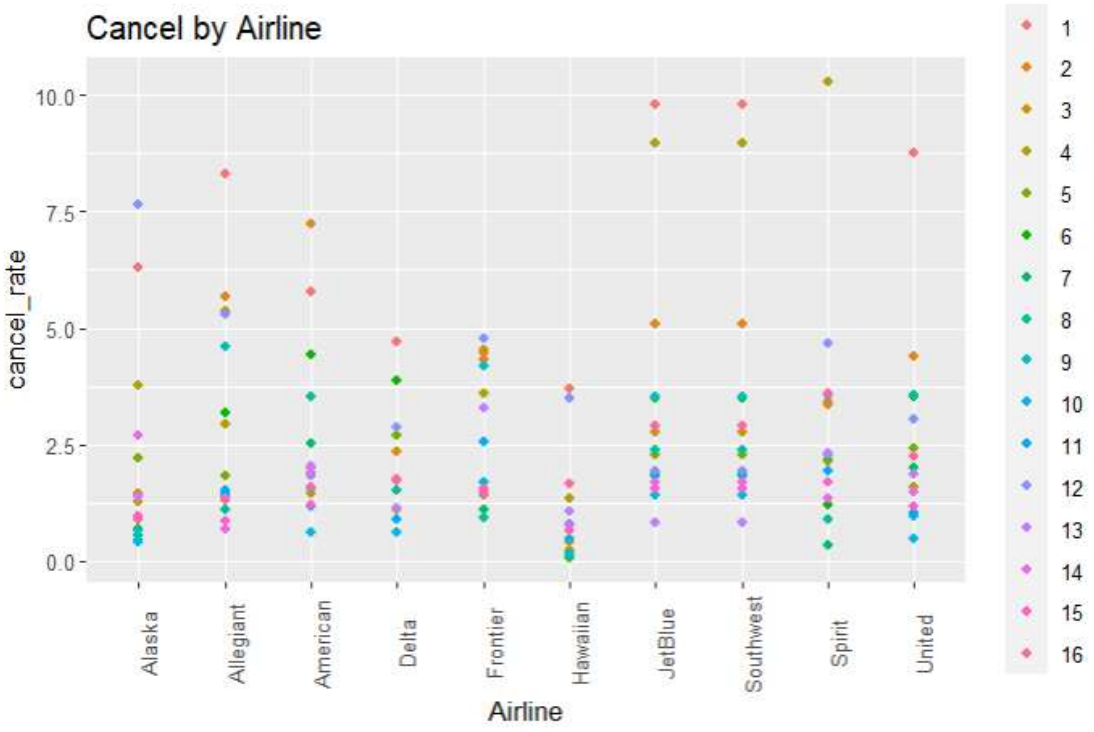
Now I want to look at cancellation data. Data is from Jan-2022 to April-2023

first check monthly cancel rate, then check airline performance.

looks like month Jan'2022, Feb'2022 and Dec'2022 are outliners (we know it from News)
Airline JetBlue, Southwest, Spirit and United have some bad records.



Month_factor <fctr>	Avg_rate <dbl>
1	6.498719
2	3.917712
3	2.259518
4	4.414394
5	1.933109
6	2.510183
7	1.337138
8	1.700823
9	2.359296
10	1.086860



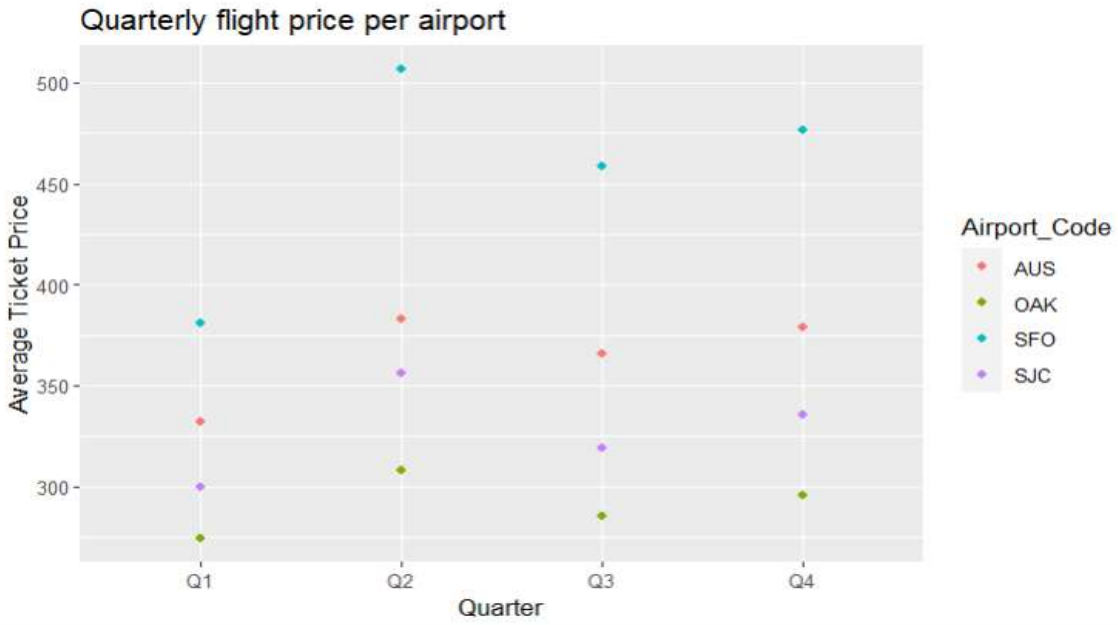
Next remove outliner month and try to rank the best and worst airline in term of cancellation.

Now looks like Hawaiian is the best, but it do not fly to Austin. the United airline looks better now, ranked #4.

Airline <chr>	avg <chr>
Alaska	133.2%
Allegiant	293.2%
American	252.6%
Delta	158.4%
Frontier	265.5%
Hawaiian	94.8%
JetBlue	273.3%
Southwest	327.2%
Spirit	282.5%
United	179.4%

Last, review the fare data. which contains fare info in whole year 2022

looks like in Q4 ticket is the most expensive, while Q1 is the cheapest



Conclusion

from Austin to Bay area and back. looks like taking skywest (now is called United Express) is likely had minimum delay and average risk of cancellation. the best airport to use is San Jose airport (SJC). Flying in Q1 is most likely to get cheap ticket.

References

@Flight Dataset @bureau_of_transportation_statistics_quarterly_nodate
@bureau_of_transportation_statistics_airline_nodate

Implications

Though my research is limited to my practical problem to select best flight route, it can be easily applied to any other airport and route. Furthermore, with no limitation of file size and process capability, the same principle could be expanded to a region or even a country.

Limitations:

Seems my computer and R studio will not be able to handle dataset too big, the code running performance will be expanded to a few hours. Also the free github account won't allow to upload file more than 200kb. So I need to select a dataset small enough.

Even had a few entry-level practices, I lack extensive understanding of the regression and machine learning algorithm. As well as time I wish to have to refine my research, I did not try to include too many variables and predict things I may not able to make judgement.

Concluding remarks:

I listed the research conclusion in page 12.

From this project, I see the power of applying R to analysis data. I also have good practice on most of the R topics covered in this class. I am interested in getting further study on some of the R topics, from next few classes and other resources.