
Title: "DSC520 final"

author: "xin tang"

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

output: html_document

output:

  pdf_document: default

bibliography: flight_citations.bib

editor_options:

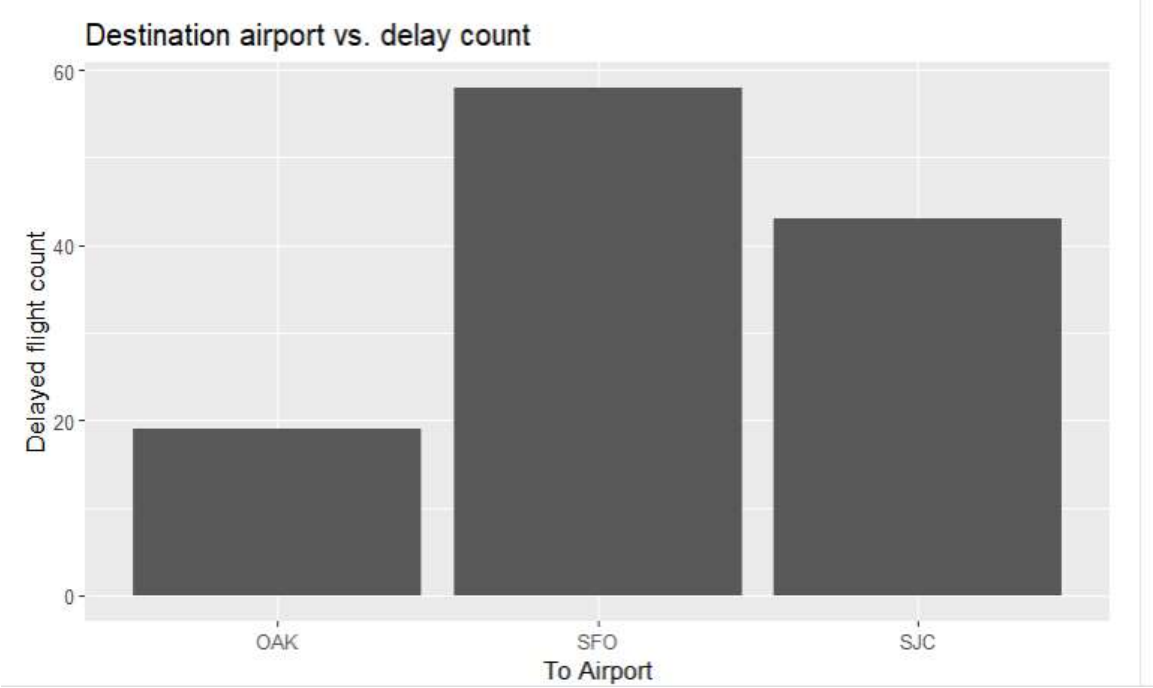
  markdown:

    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

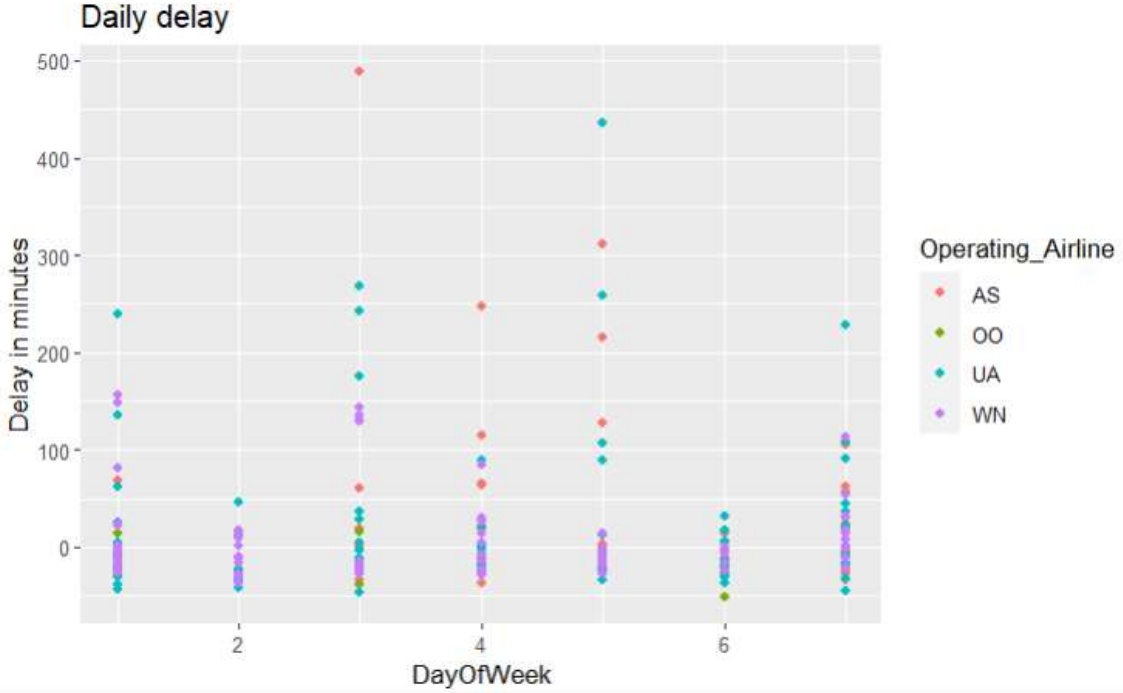
Origin <chr>	delay_count <int>	flight_count <int>	delay_percent <chr>
OAK	21	31	68%
SFO	58	154	38%
SJC	30	112	27%

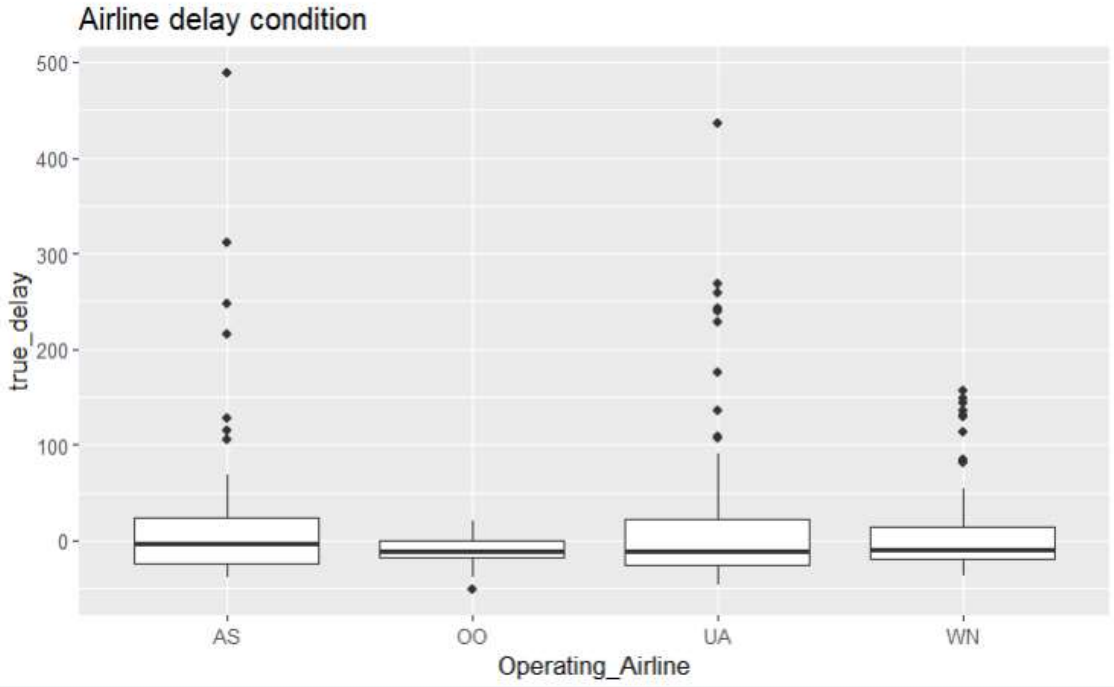
From this analysis, no matter as Origin or Destination airport, OAK airport have twice as much delay as other 2 airports. so it is not a good choice to fly into

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

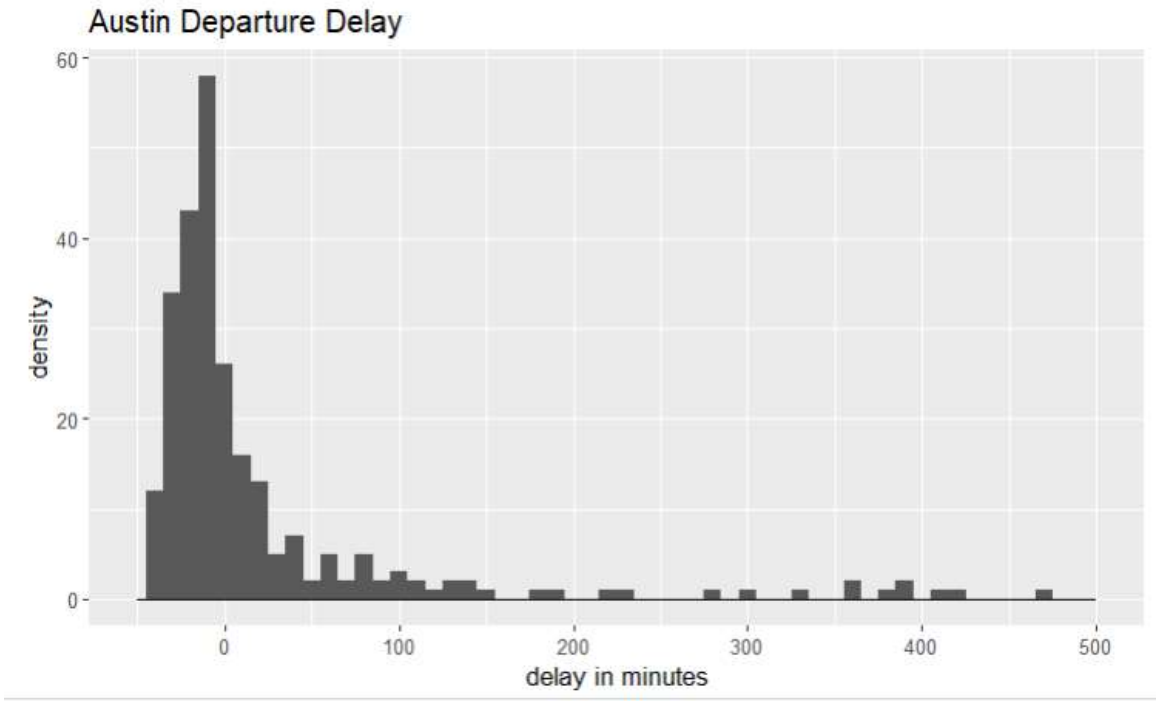
Check which delay factor is the most significant when fly to bay area

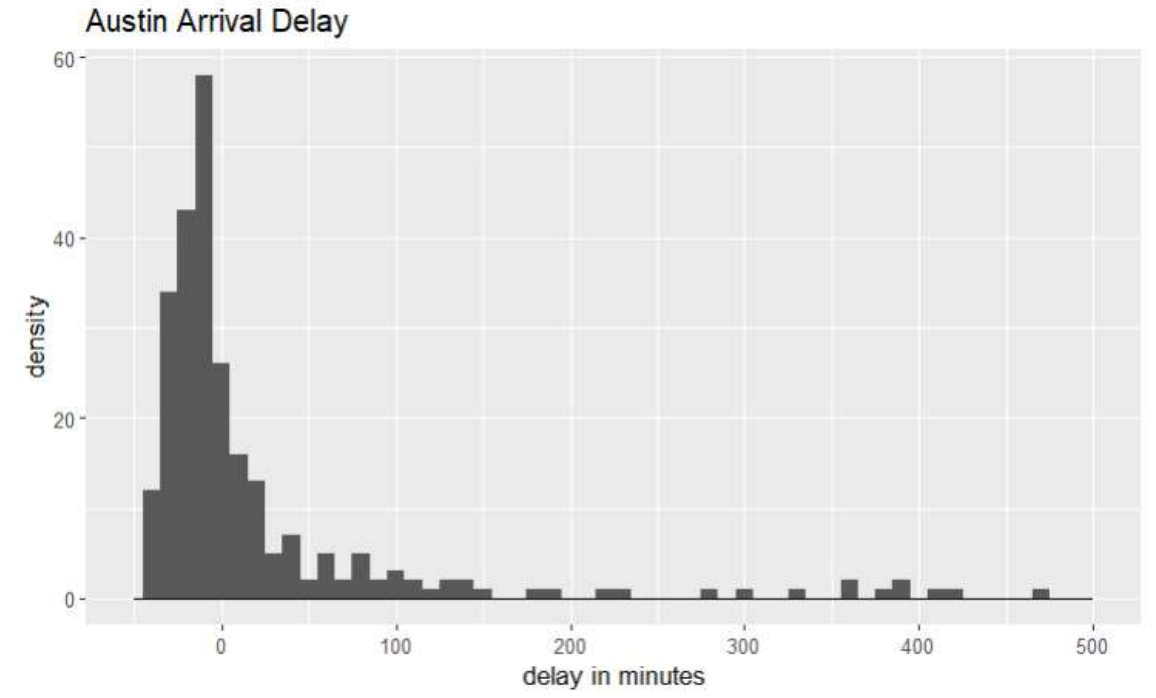
it shows the prior flight delay and control tower caused most delay, which I can not control

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'

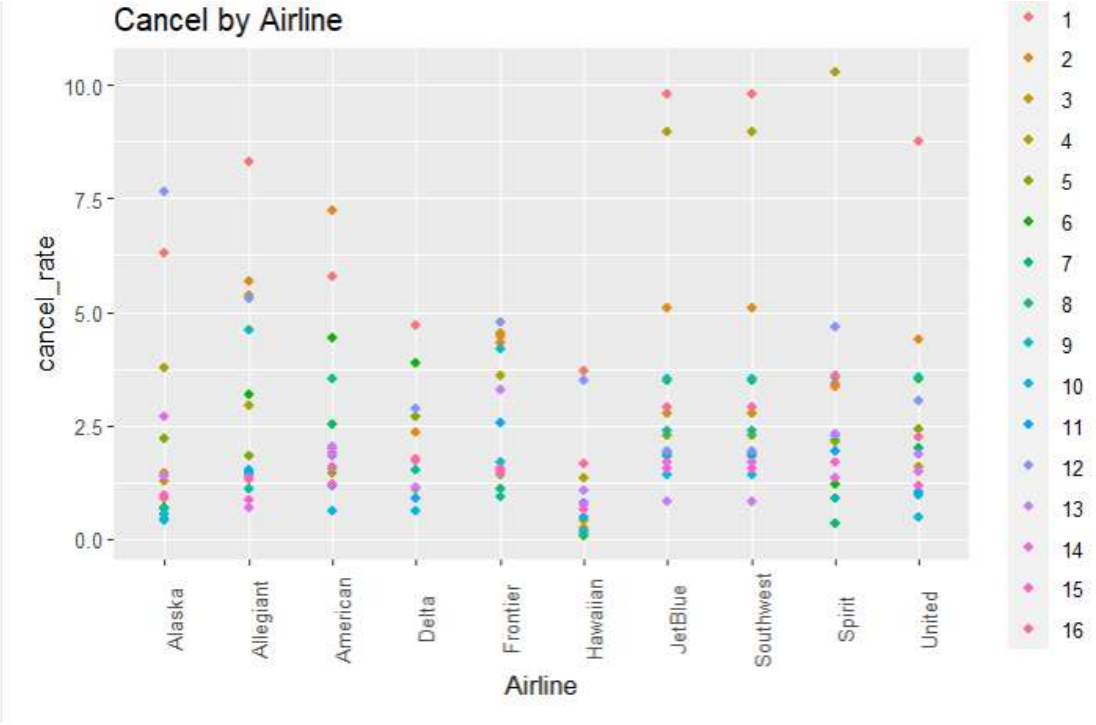
Again the air control caused delay is more important factor

```
                   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
```

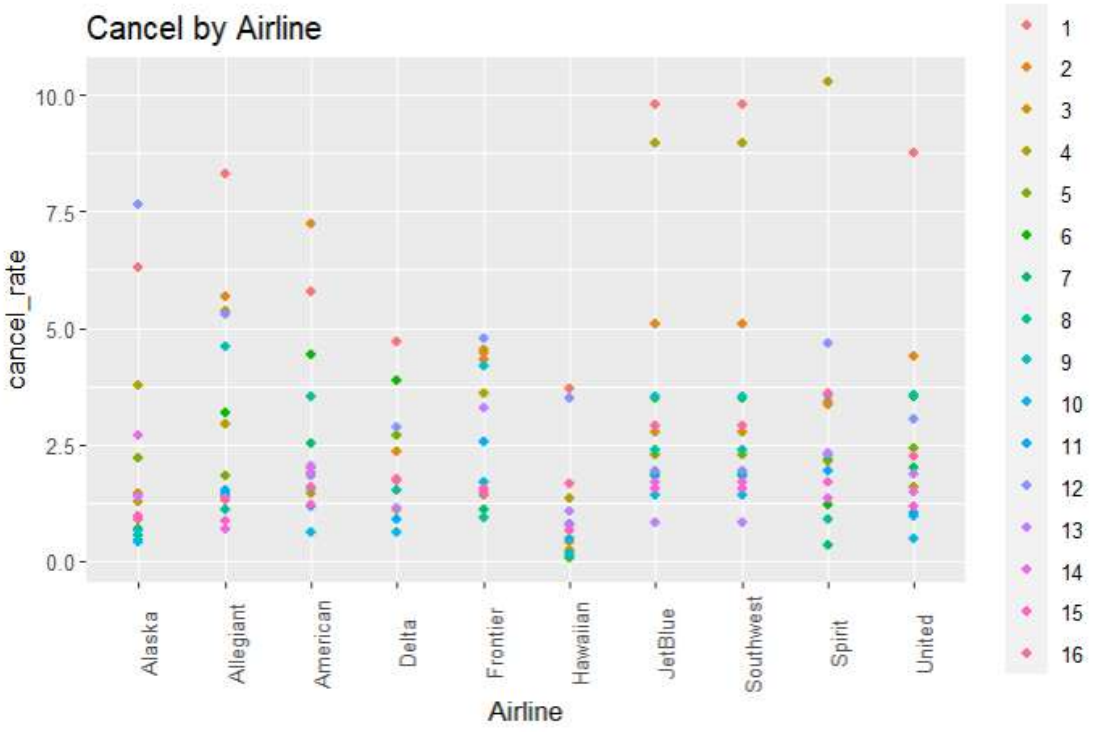
Now I want to look at cancellation data. Data is from Jan-2022 to April-2023

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