Airline Delay Analysis

Big Data Spring 2022

Introduction

Information

- Analyze a series of data sets related to airlines and airports
- Create metrics to visualize what may cause delays, and their trend
- Provide a series of visualizations that show these trends
- This will be helpful as we all travel this season

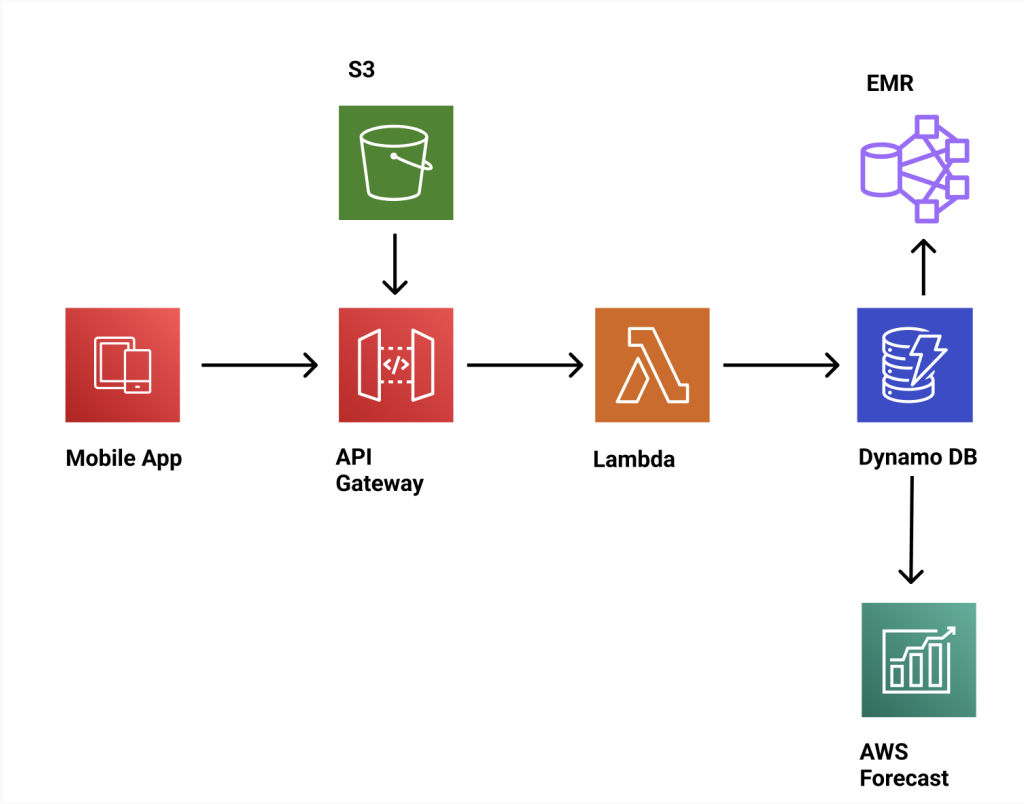
Architecture

- Used Kaggle, Amstat
- Spark Implementation, NumPy, Pandas
- Processing 10gb of data requires a distributed system
- Scalable programming that can be implemented on future data streams

Big Data

- Large Scale Data Set
- Attempted running our calculations locally
- Errors occurred and calculations were taking way to long
- Sparks distributed nature solved our problem
- Data cleansing/extraction was practically impossible locally

Mobile App/ Prediction



Preliminary Analysis

specifications and considerations

- representation of time
  - *points* vs *durations*
  - float data type
- standardization
  - centralized record-keeping
  - nomenclature consistency: IATA airport codes
  - objectivity of time data
  - time zone recording
- comprehensiveness of data fields
  - events/stages of flight process
  - *planned* and *actual* times

Preliminary Analysis (cont'd)

tradeoffs and implications

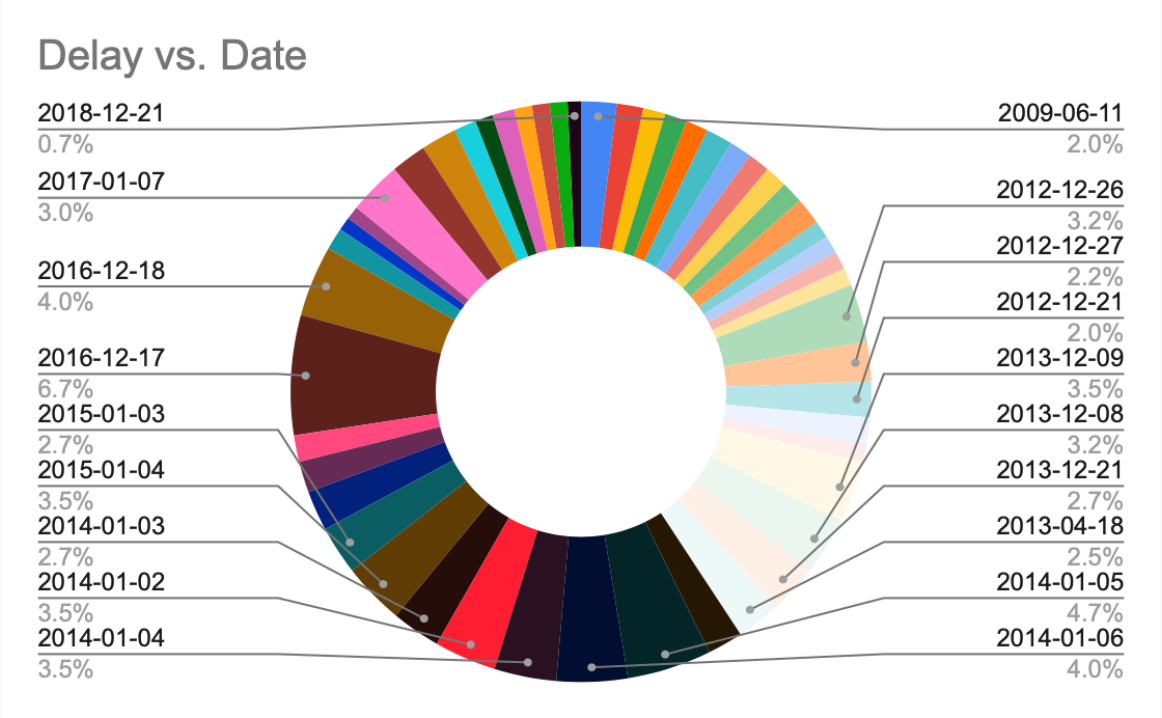
- choice of local time zones over single standard
  - difficult/unintuitive time comparisons
  - real-world practicality
- inclusion of seemingly-redundant fields
  - data volume
  - elimination of complicated calculations

Preliminary Analysis (cont'd)

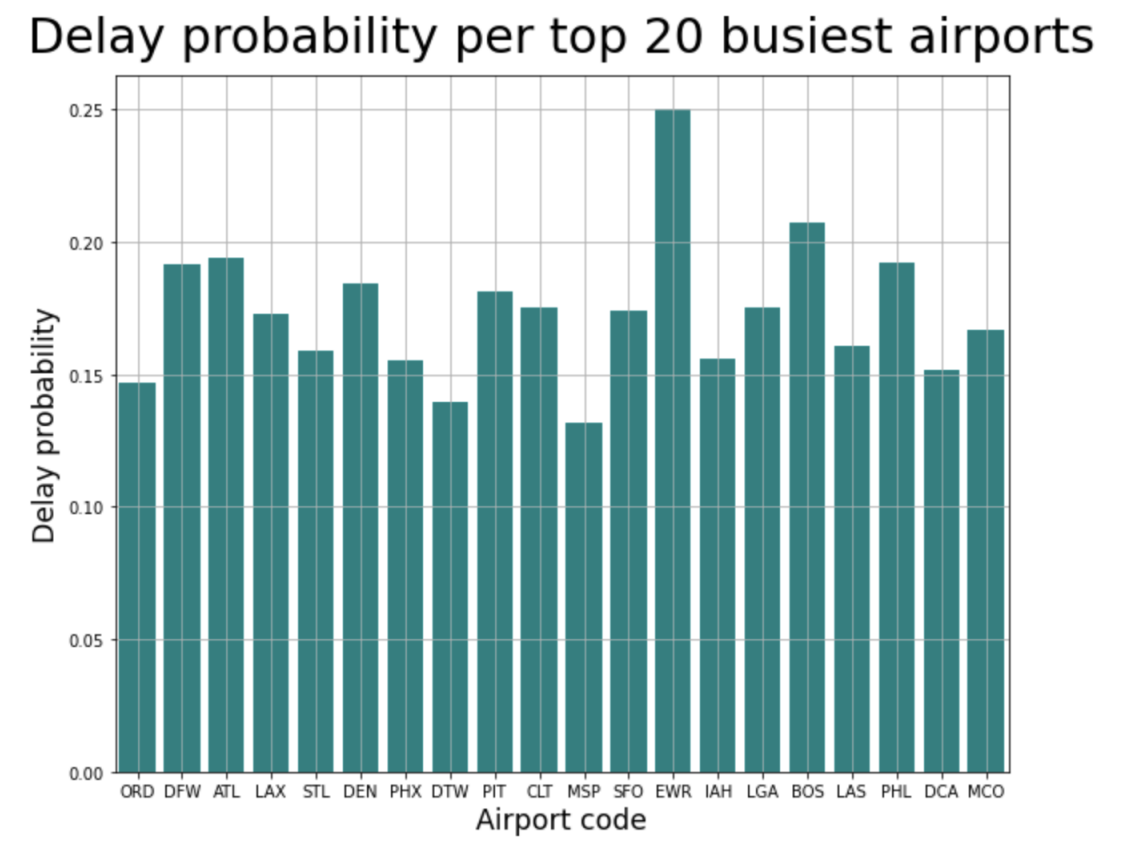
recognizing missing/ incomplete data

- "expected"/necessary fields
- cancelled flights
- exceptions

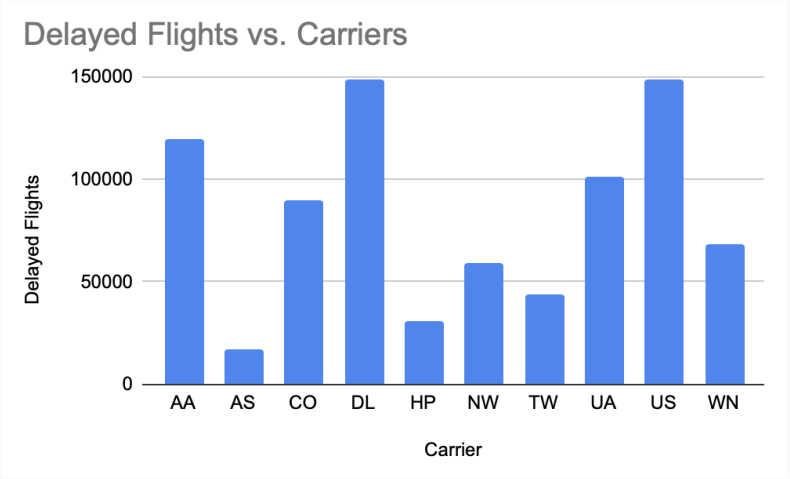
Least Reliable Day to Fly

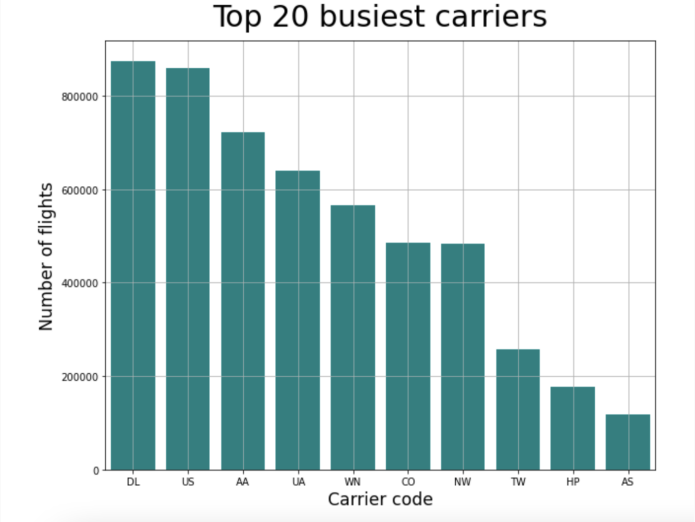


1994

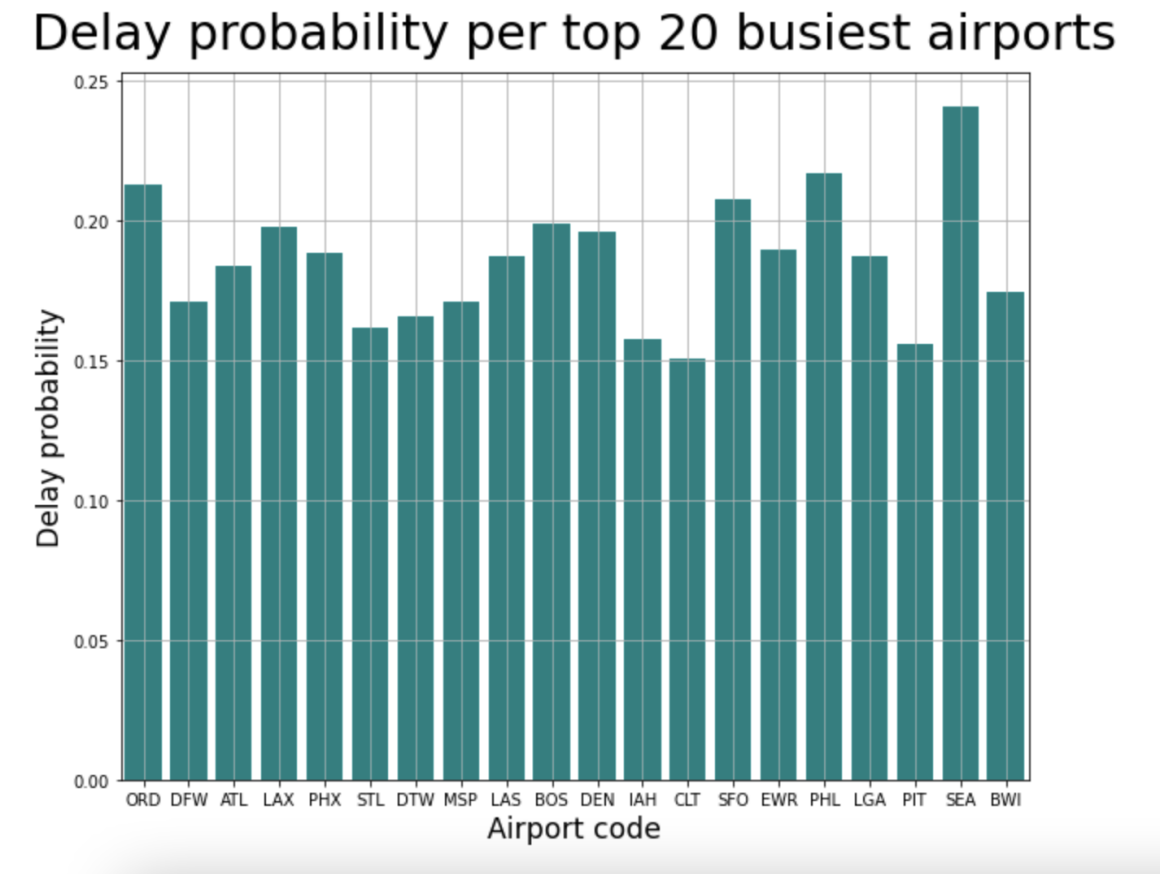

Delay probability per top 20 busiest airports
Delay probability
0.25
0.20
0.15
0.10
0.05
0.00
ORD DFW ATL LAX STL DEN PHX DTW PIT CLT MSP SFO EWR IAH LGA BOS LAS PHL DCA MCO
Airport code

1994

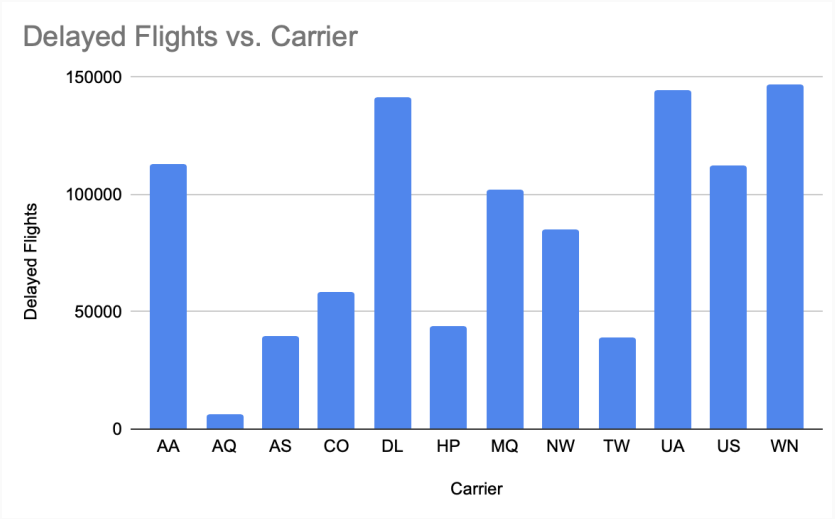

Delayed Flights vs. Carriers
Delayed Flights
150000
100000
50000
0
AA
AS
CO
DL
HP
NW
TW
UA
US
WN
Carrier

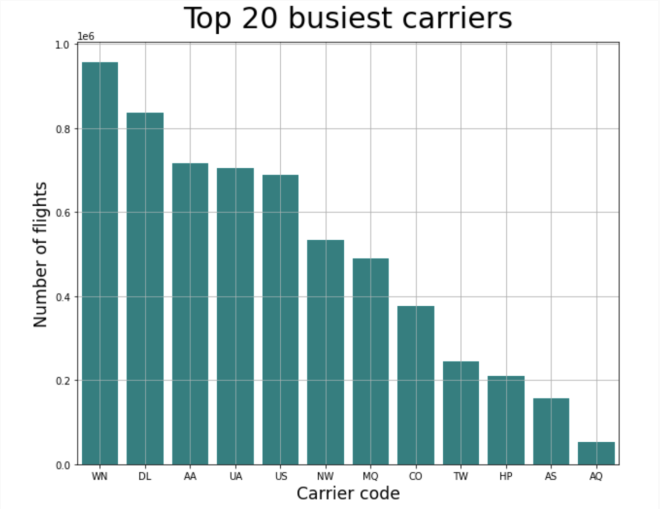

Top 20 busiest carriers
Number of flights
800000
600000
400000
200000
0
DL
US
AA
UA
WN
CO
NW
TW
HP
AS
Carrier code

2001

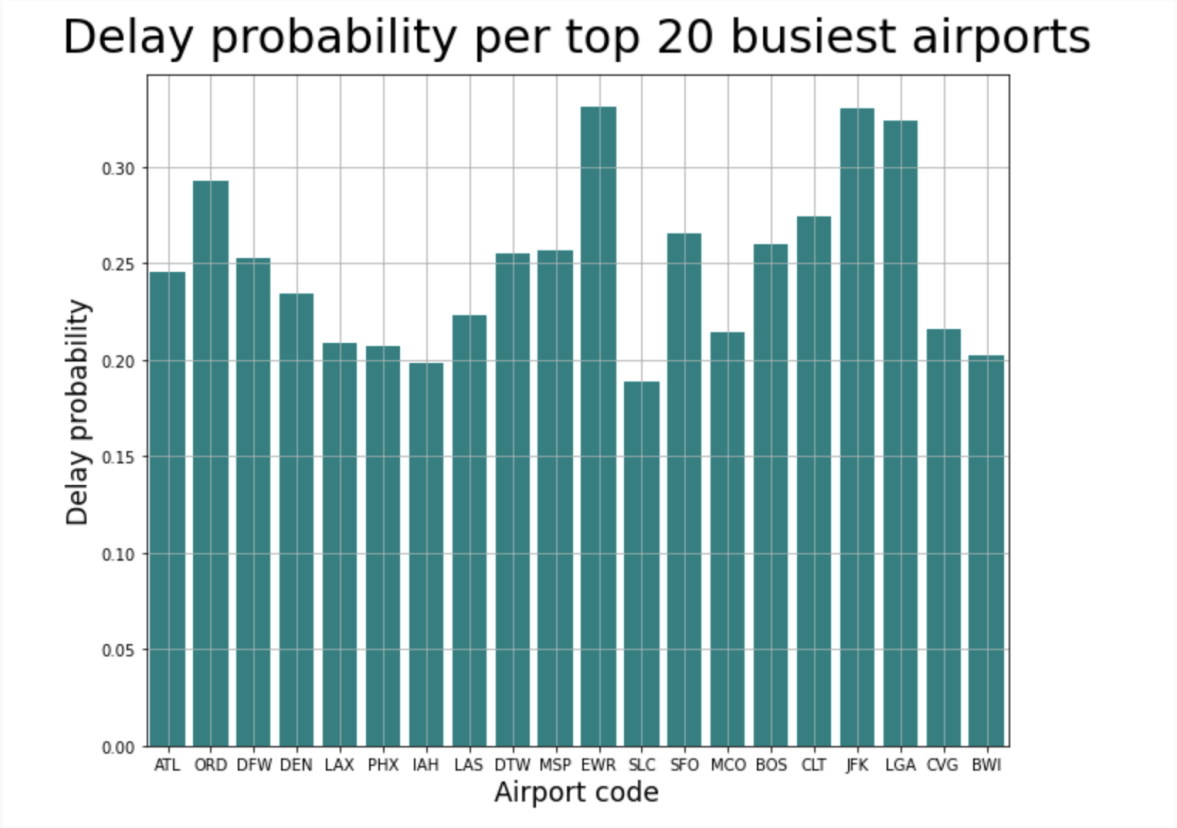

Delay probability per top 20 busiest airports
Delay probability
0.25
0.20
0.15
0.10
0.05
0.00
ORD DFW ATL LAX PHX STL DTW MSP LAS BOS DEN IAH CLT SFO EWR PHL LGA PIT SEA BWI
Airport code

2001

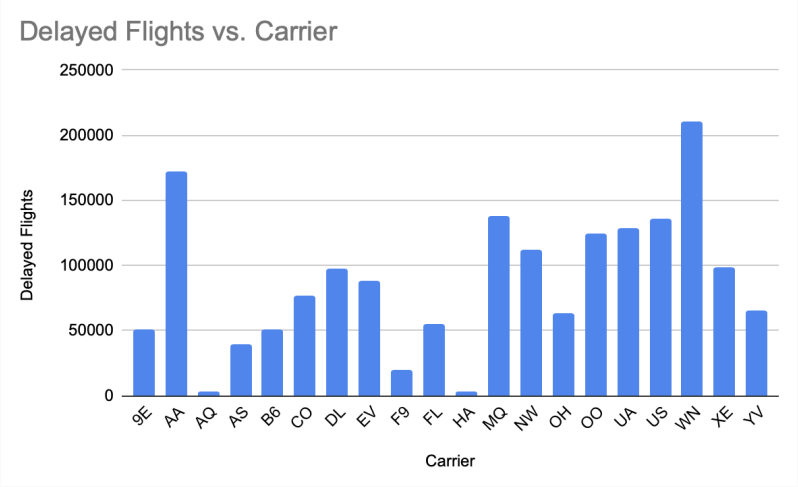

Delayed Flights vs. Carrier
150000
100000
50000
0
Delayed Flights
AA
AQ
AS
CO
DL
HP
MQ
NW
TW
UA
US
WN
Carrier

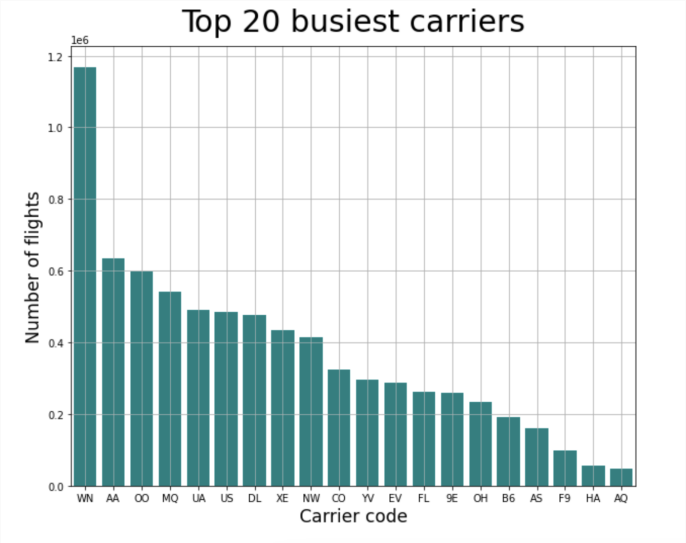

Top 20 busiest carriers
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Number of flights
WN
DL
AA
UA
US
NW
MQ
CO
TW
HP
AS
AQ
Carrier code

2007

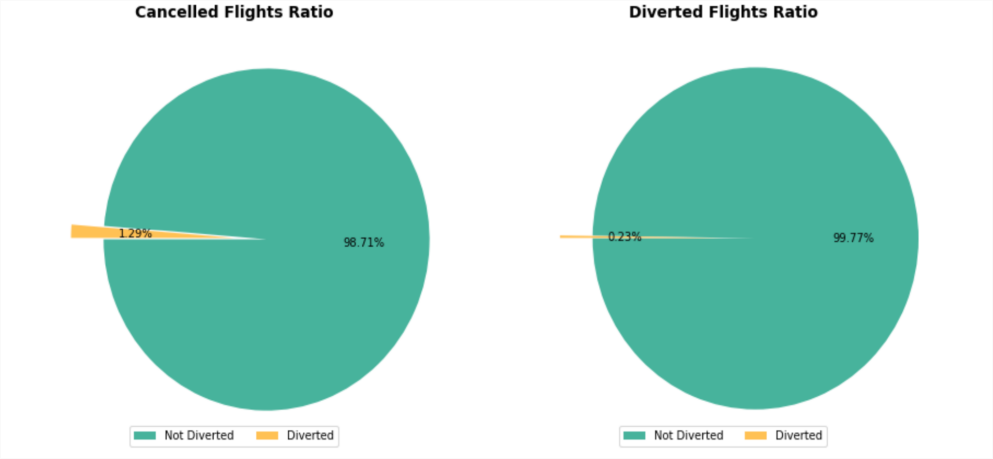

Delay probability per top 20 busiest airports
Delay probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
ATL ORD DFW DEN LAX PHX IAH LAS DTW MSP EWR SLC SFO MCO BOS CLT JFK LGA CVG BWI
Airport code

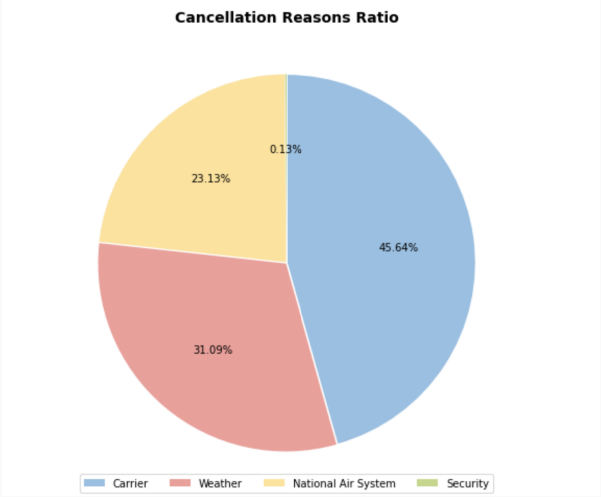
2007


Delayed Flights vs. Carrier
Delayed Flights
250000
200000
150000
100000
50000
0
9E AA AQ AS B6 CO DL EV F9 FL HA MQ NW OH OO UA US WN XE YV
Carrier

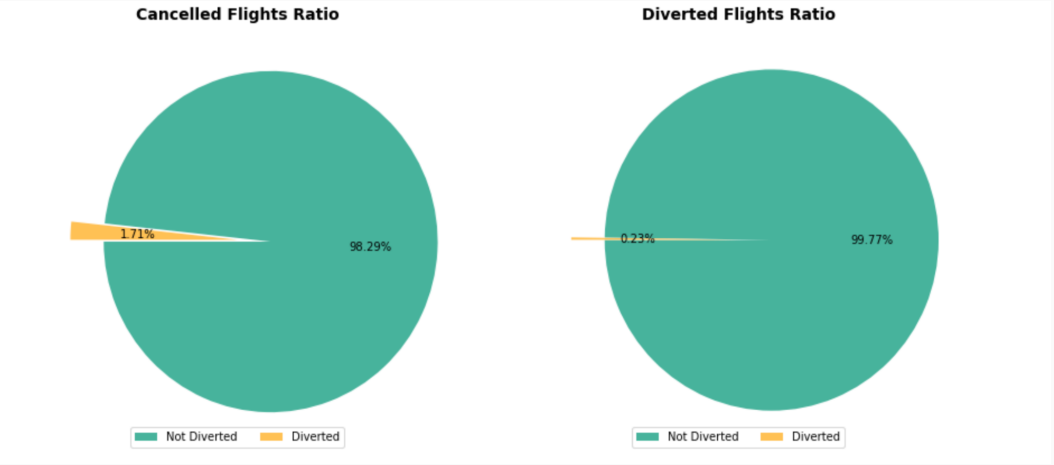

Top 20 busiest carriers
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Number of flights
WN AA OO MQ UA US DL XE NW CO YV EV FL 9E OH B6 AS F9 HA AQ
Carrier code

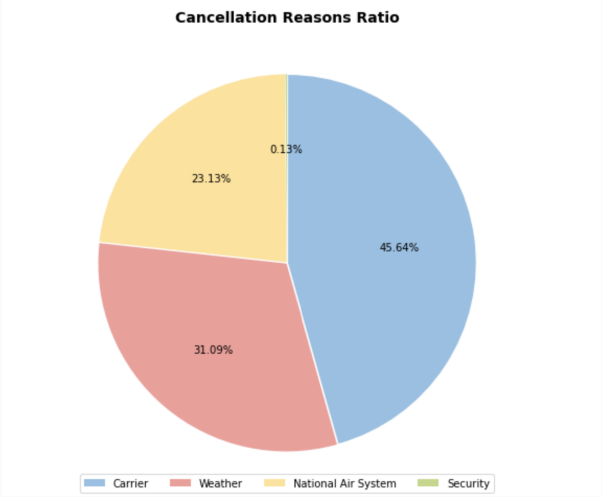
1994


Cancelled Flights Ratio
1.29%
98.71%
Not Diverted
Diverted
Diverted Flights Ratio
0.23%
99.77%
Not Diverted
Diverted

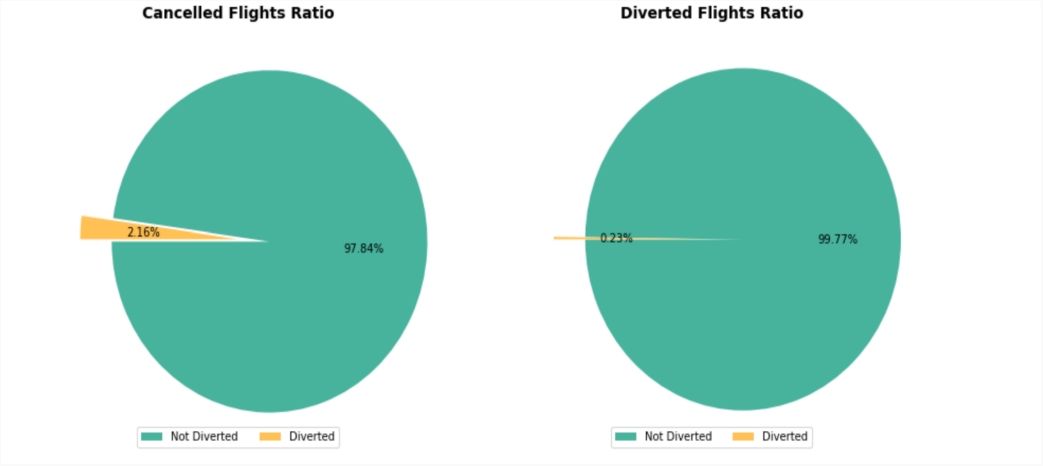

Cancellation Reasons Ratio
0.13%
23.13%
45.64%
31.09%
Carrier
Weather
National Air System
Security

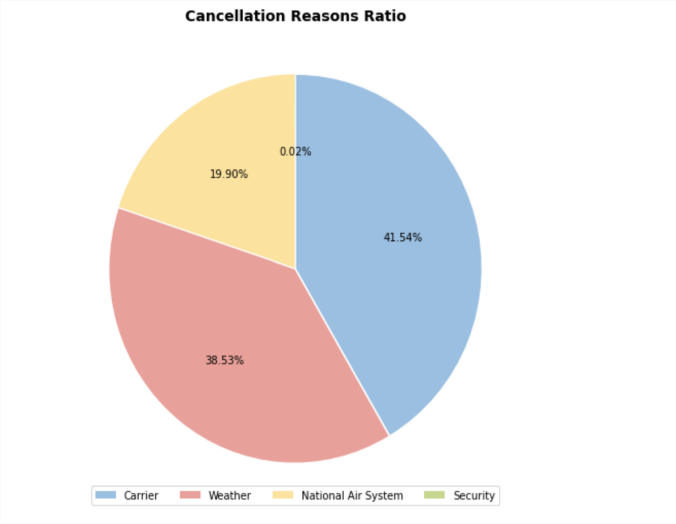
2006


Cancelled Flights Ratio
1.71%
98.29%
Not Diverted
Diverted
Diverted Flights Ratio
0.23%
99.77%
Not Diverted
Diverted

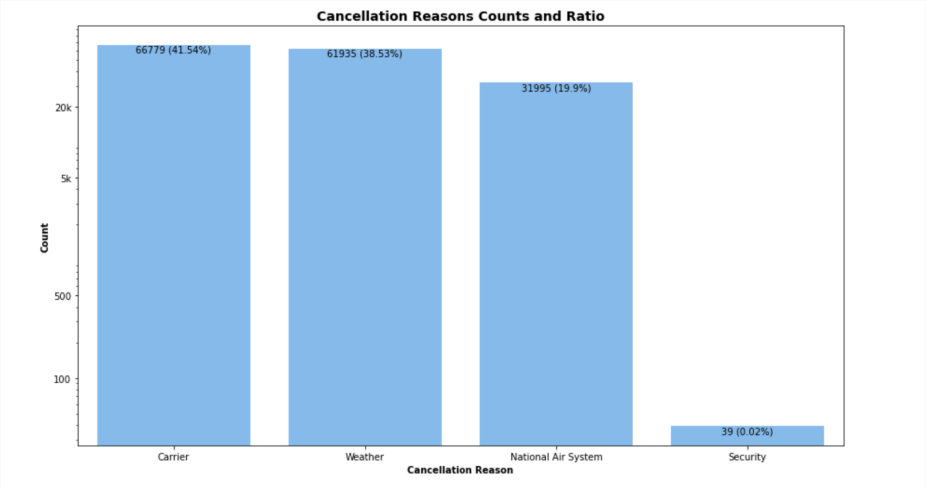

Cancellation Reasons Ratio
0.13%
23.13%
45.64%
31.09%
Carrier
Weather
National Air System
Security

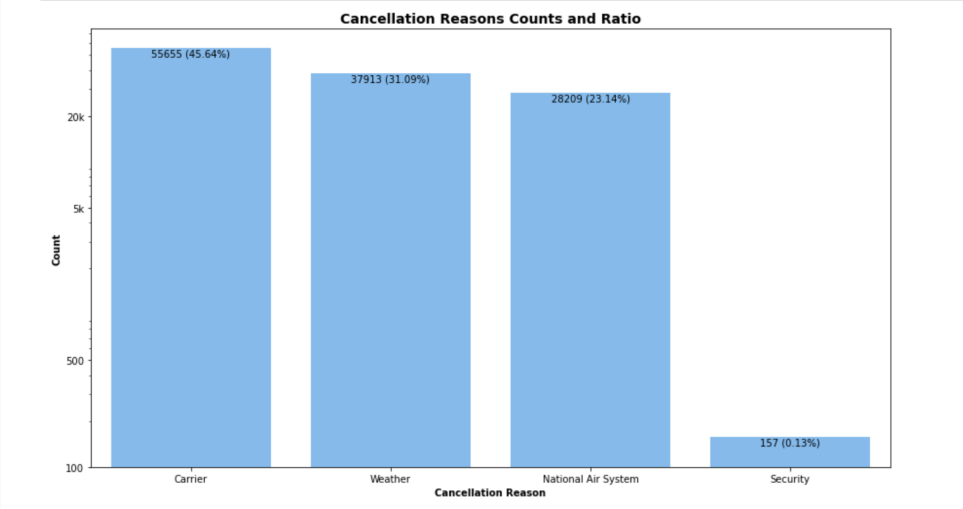
2007


Cancelled Flights Ratio
2.16%
97.84%
Not Diverted
Diverted
Diverted Flights Ratio
0.23%
99.77%
Not Diverted
Diverted

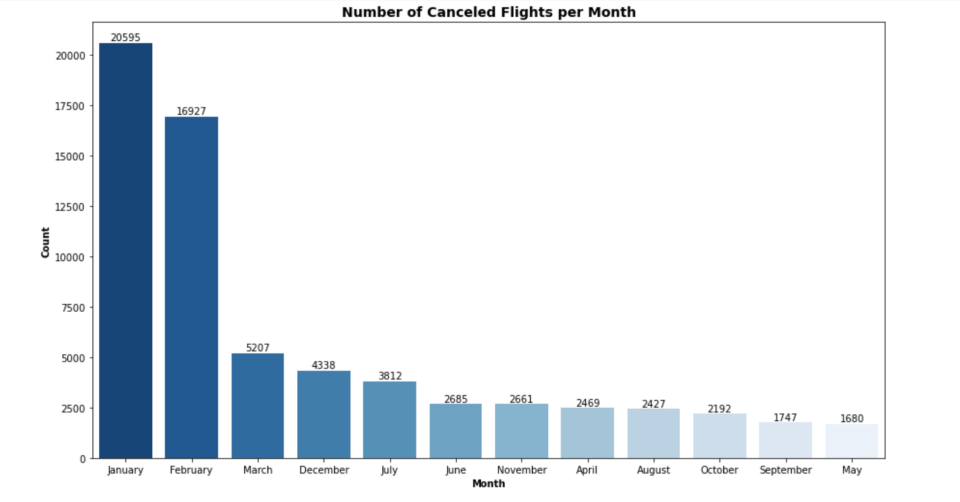

Cancellation Reasons Ratio
0.02%
19.90%
41.54%
38.53%
Carrier
Weather
National Air System
Security

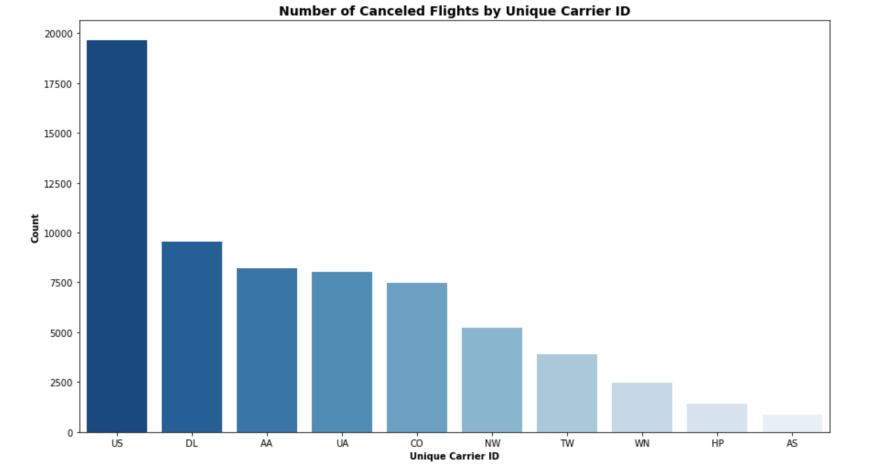
2006 v 2007


Cancellation Reasons Counts and Ratio
66779 (41.54%)
61935 (38.53%)
31995 (19.9%)
39 (0.02%)
20k
5k
500
100
Count
Carrier
Weather
National Air System
Security
Cancellation Reason

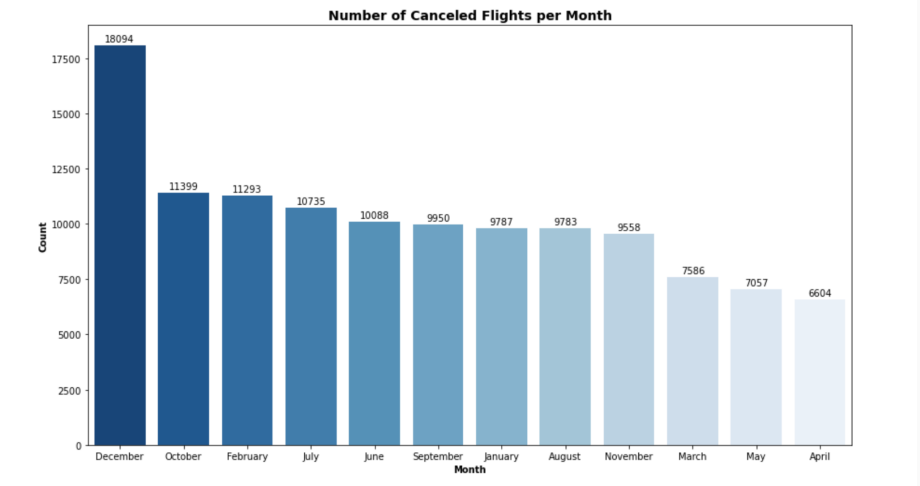

Cancellation Reasons Counts and Ratio
55655 (45.64%)
37913 (31.09%)
28209 (23.14%)
157 (0.13%)
20k
5k
500
100
Count
Carrier
Weather
National Air System
Security
Cancellation Reason

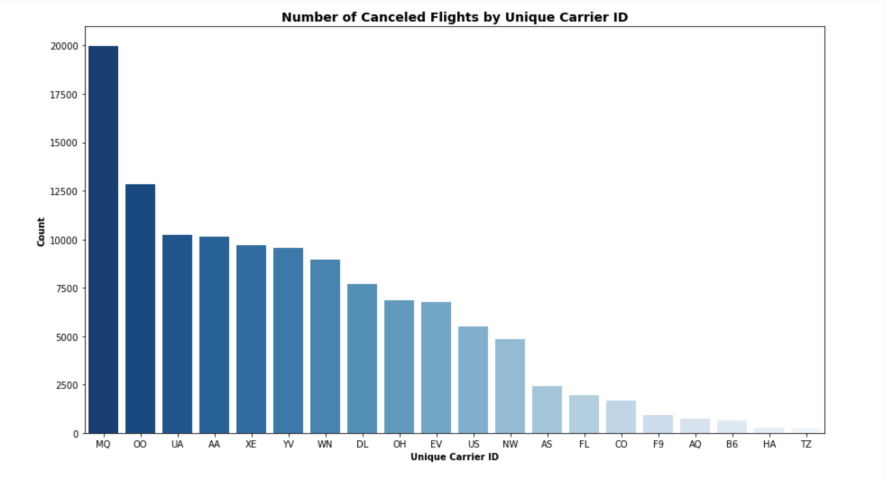
1994


Number of Canceled Flights per Month
Count
Month
20595
16927
5207
4338
3812
2685
2661
2469
2427
2192
1747
1680
January
February
March
December
July
June
November
April
August
October
September
May

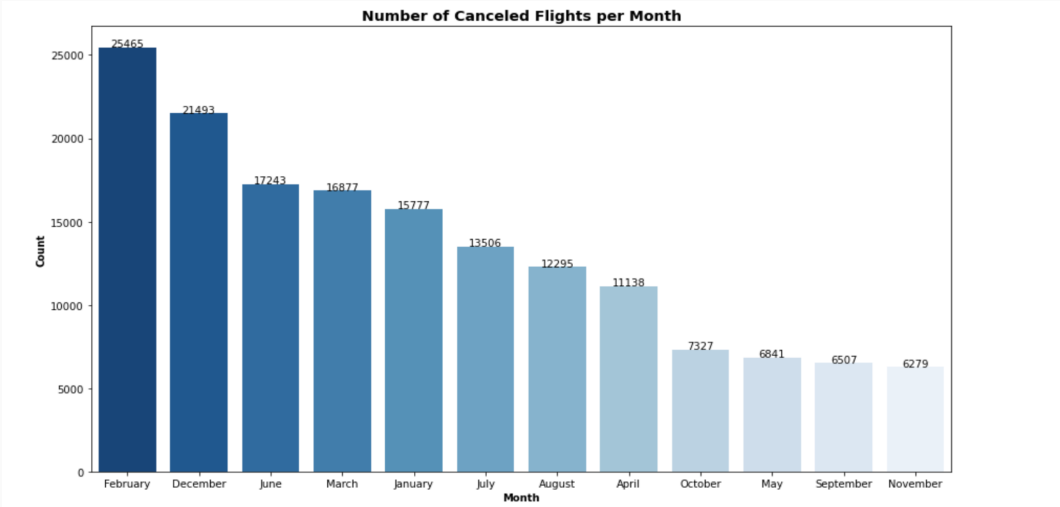

Number of Canceled Flights by Unique Carrier ID
Count
Unique Carrier ID
US
DL
AA
UA
CO
NW
TW
WN
HP
AS

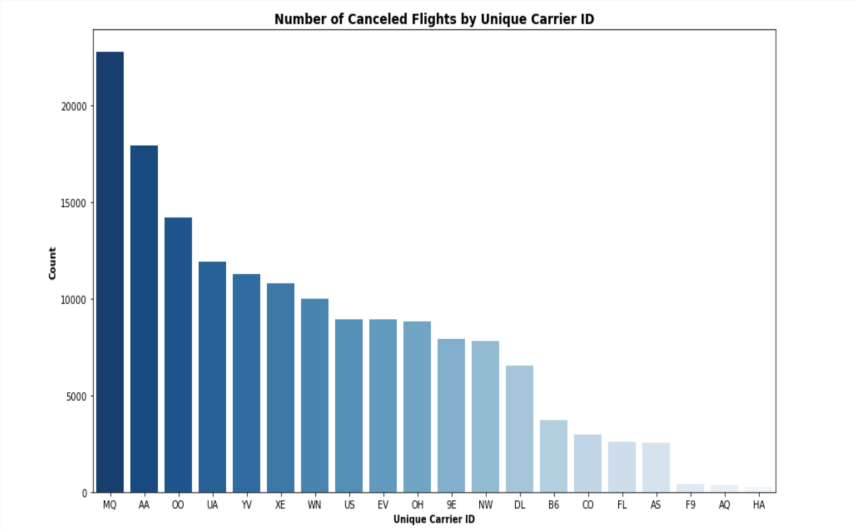
2006


Number of Canceled Flights per Month
Count
Month
18094
11399
11293
10735
10088
9950
9787
9783
9558
7586
7057
6604
December
October
February
July
June
September
January
August
November
March
May
April

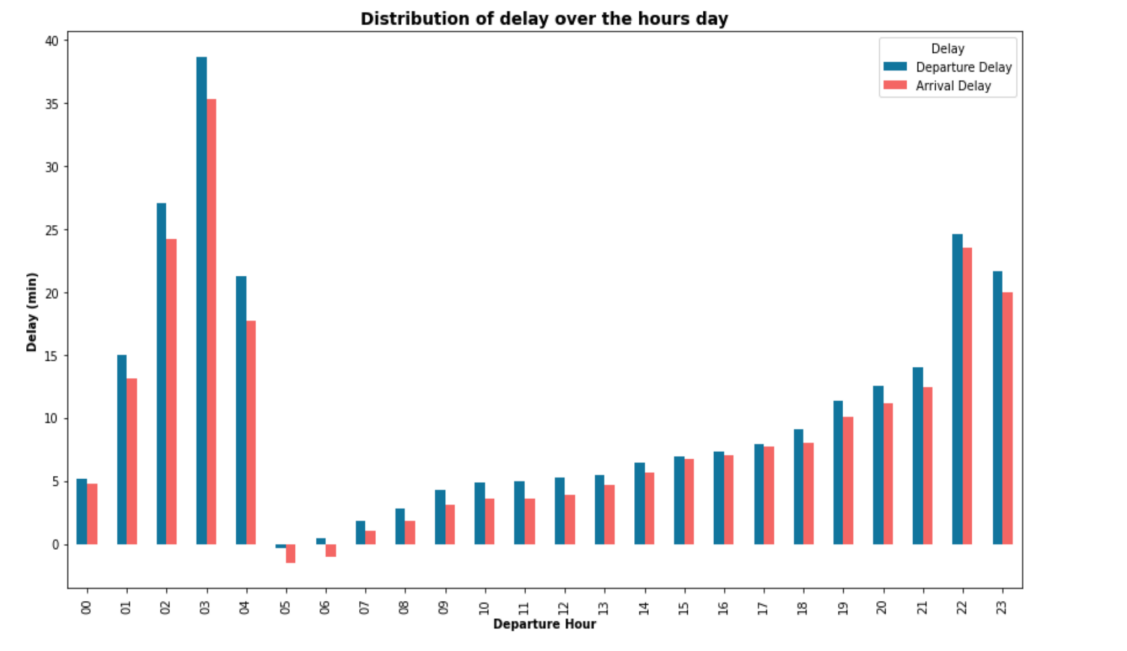

Number of Canceled Flights by Unique Carrier ID
Count
Unique Carrier ID
MQ
OO
UA
AA
XE
YV
WN
DL
OH
EV
US
NW
AS
FL
CO
F9
AQ
B6
HA
TZ

2007

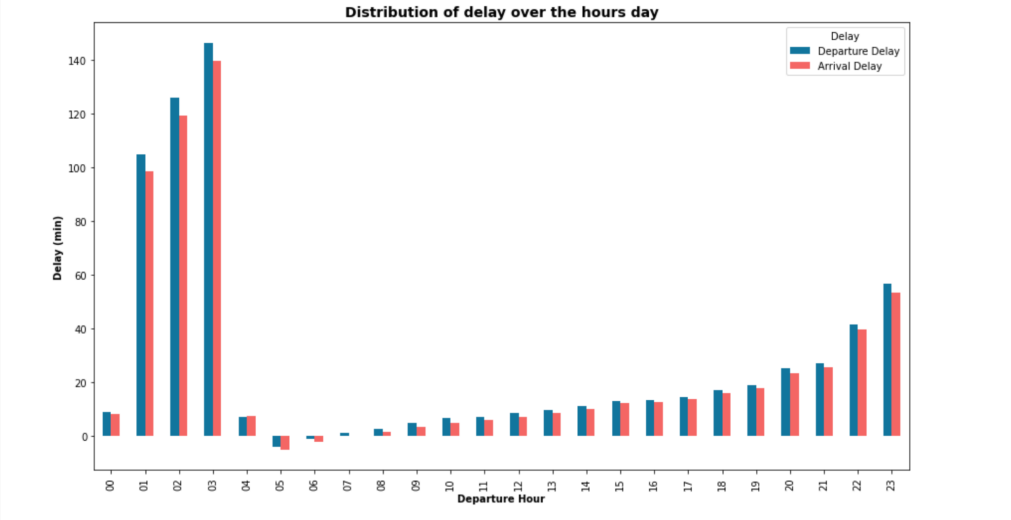

Number of Canceled Flights per Month
Count
Month
25465
21493
17243
16877
15777
13506
12295
11138
7327
6841
6507
6279
February
December
June
March
January
July
August
April
October
May
September
November

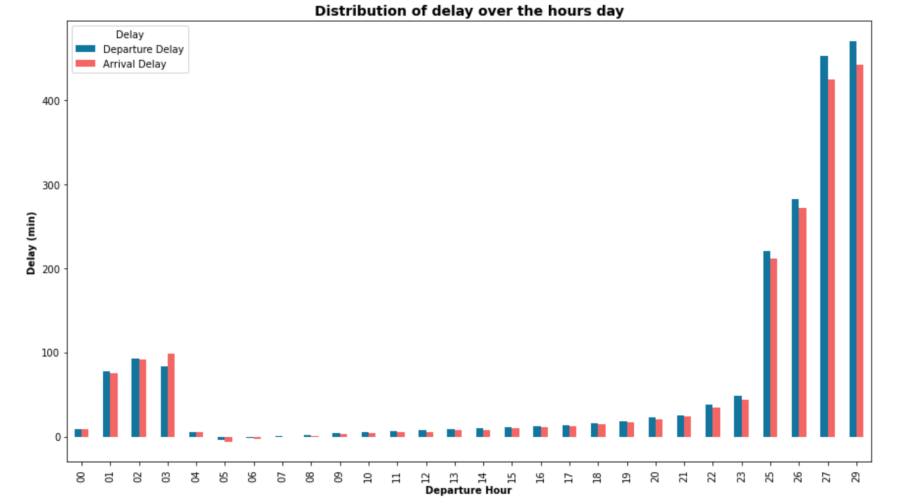

Number of Canceled Flights by Unique Carrier ID
Count
Unique Carrier ID
MQ
AA
OO
UA
YV
XE
WN
US
EV
OH
9E
NW
DL
B6
CO
FL
AS
F9
AQ
HA

1994

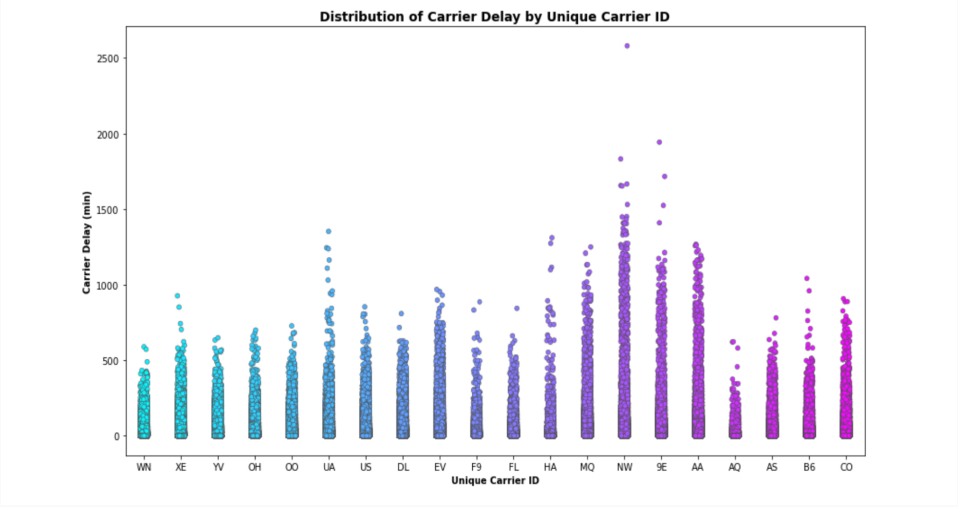

Distribution of delay over the hours day
Delay
Departure Delay
Arrival Delay
Delay (min)
40
35
30
25
20
15
10
5
0
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Departure Hour

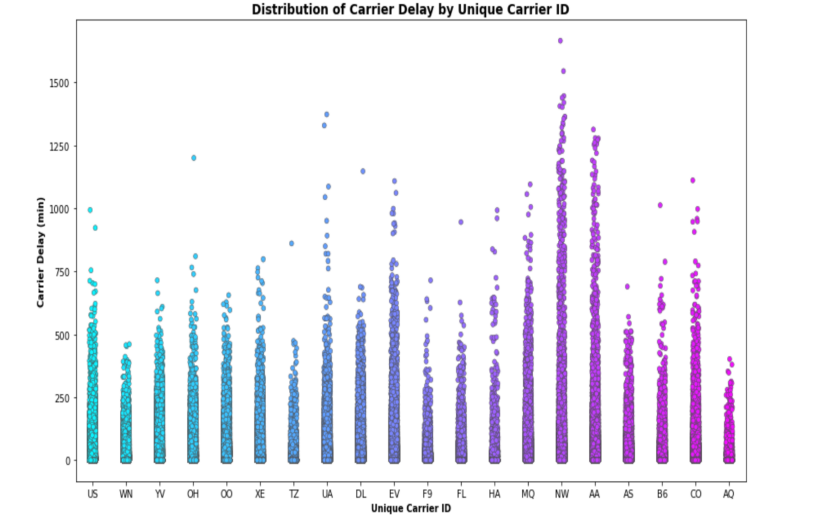
2006 v 2007


Distribution of delay over the hours day
Delay
Departure Delay
Arrival Delay
Delay (min)
Departure Hour


Distribution of delay over the hours day
Delay
Departure Delay
Arrival Delay
Delay (min)
Departure Hour

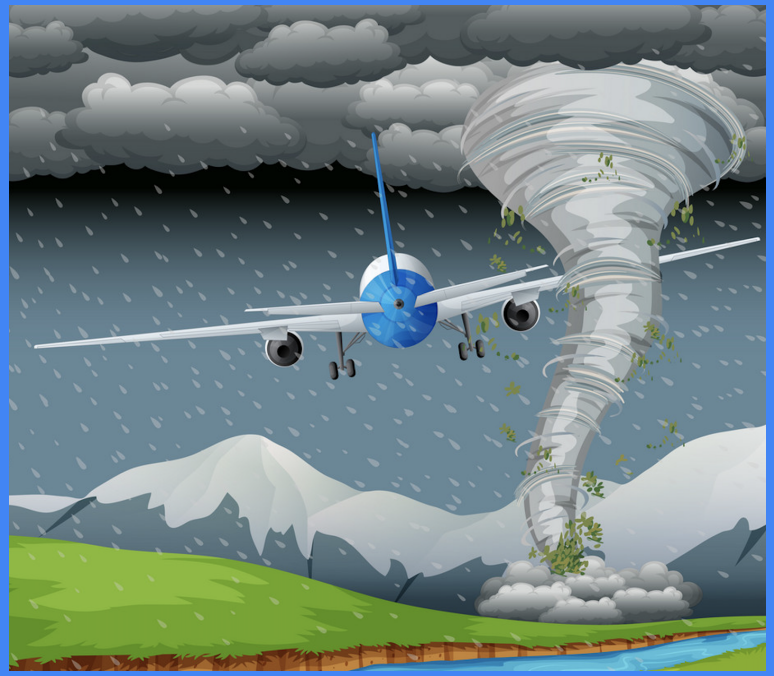
2006 v 2007


Distribution of Carrier Delay by Unique Carrier ID
Carrier Delay (min)
2500
2000
1500
1000
500
0
WN XE YV OH OO UA US DL EV F9 FL HA MQ NW 9E AA AQ AS B6 CO
Unique Carrier ID


Distribution of Carrier Delay by Unique Carrier ID
Carrier Delay (min)
1500
1250
1000
750
500
250
0
US WN YV OH OO XE TZ UA DL EV F9 FL HA MQ NW AA AS B6 CO AQ
Unique Carrier ID

Conclusion

- Climate And Carrier



Thanks!