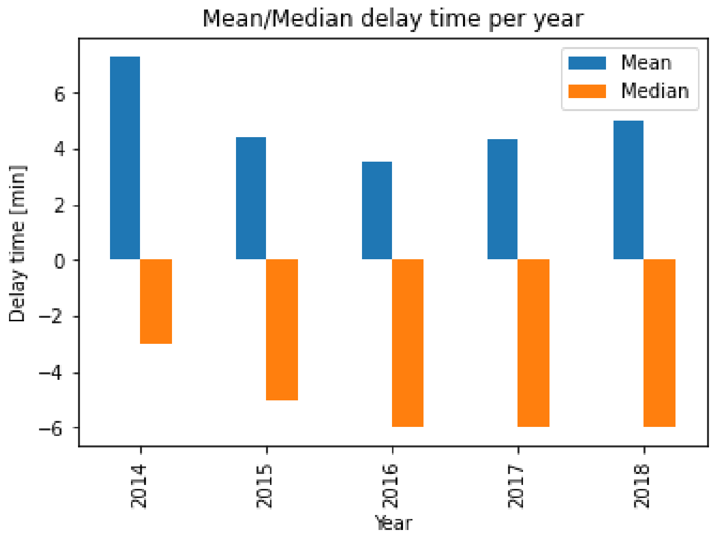
Case study

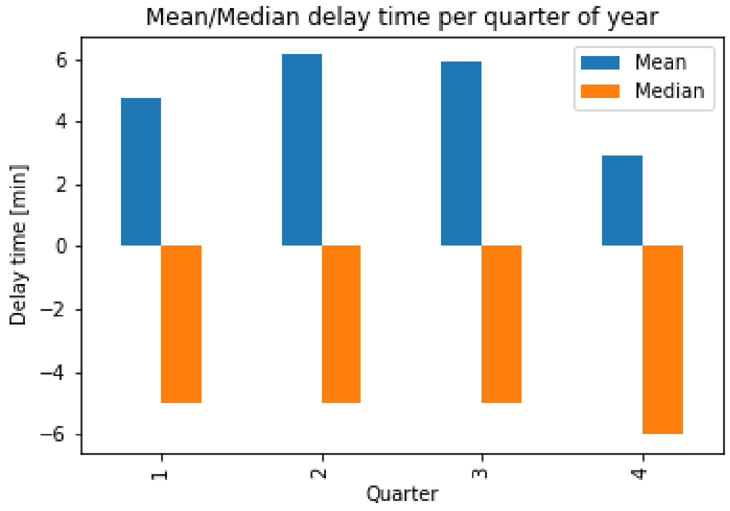
Vojislav Staletovic

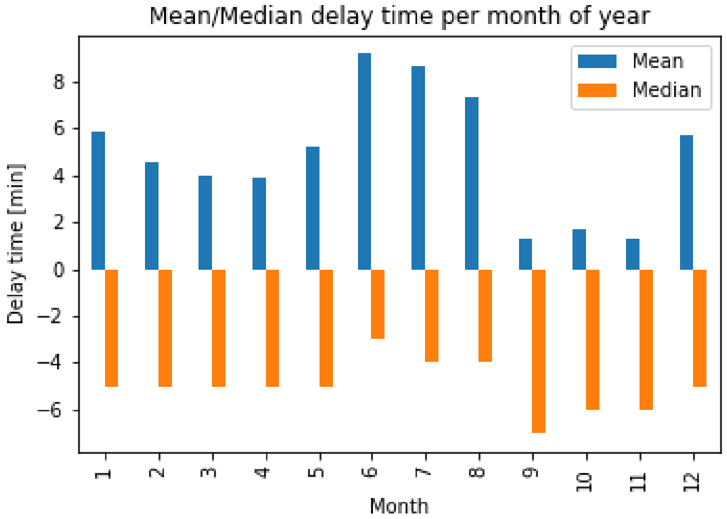
Mean/Median/Skewness/Kurtosis of delay time

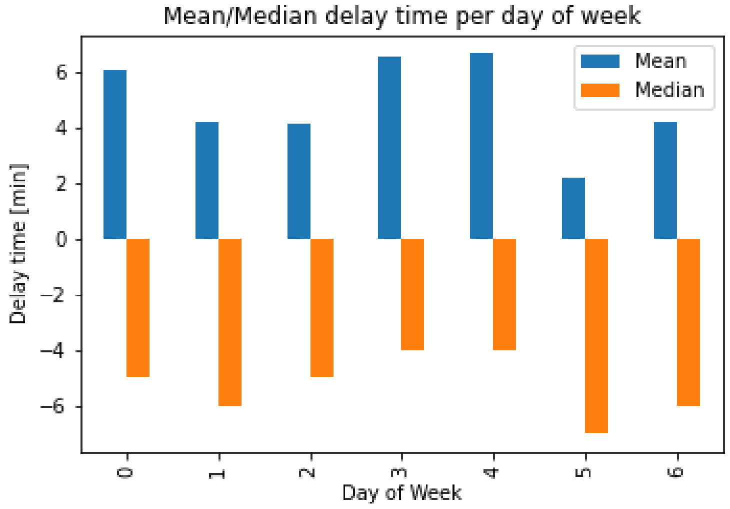
- Mean = 4.94 min
- Median = -5 min
- Skewness = 7.8
- Kurtosis = 132.56
- Comparing mean and median, mean is more sensitive to outliers, so that's the reason why mean has higher value.
- Skewness is a measure of the asymmetry of a distribution, and if it is zero, there is no asymmetry. Skewness is 0 for normal distribution. We have positive skew, and because of it, median has lower value than mean.
- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. Kurtosis is 0 for normal distribution when it is used Fisher's definition. Extreme positive kurtosis (our case) indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean. That means there are more chances of outliers. When kurtosis is positive, it has a leptokurtic distribution.
- Theoretically, I would fit beta distribution for delay time, because it can fit my skewness and kurtosis. It is right skewed.

Mean/Median

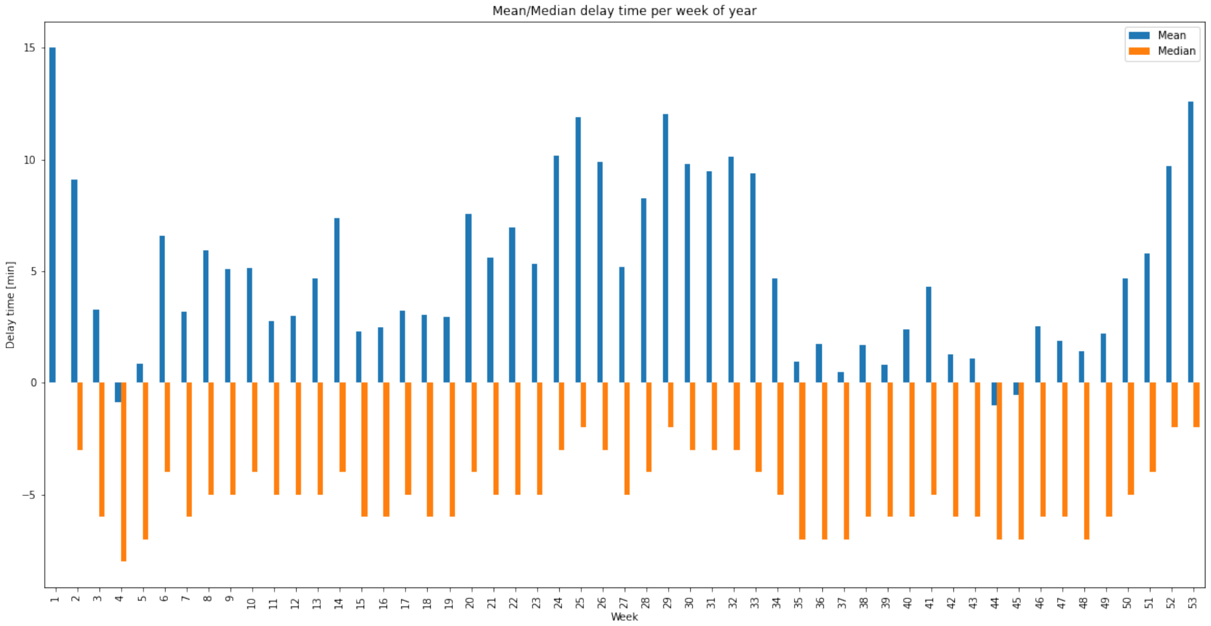

Mean/Median delay time per year
Mean
Median
Delay time [min]
6
4
2
0
−2
−4
−6
2014
2015
2016
2017
2018
Year

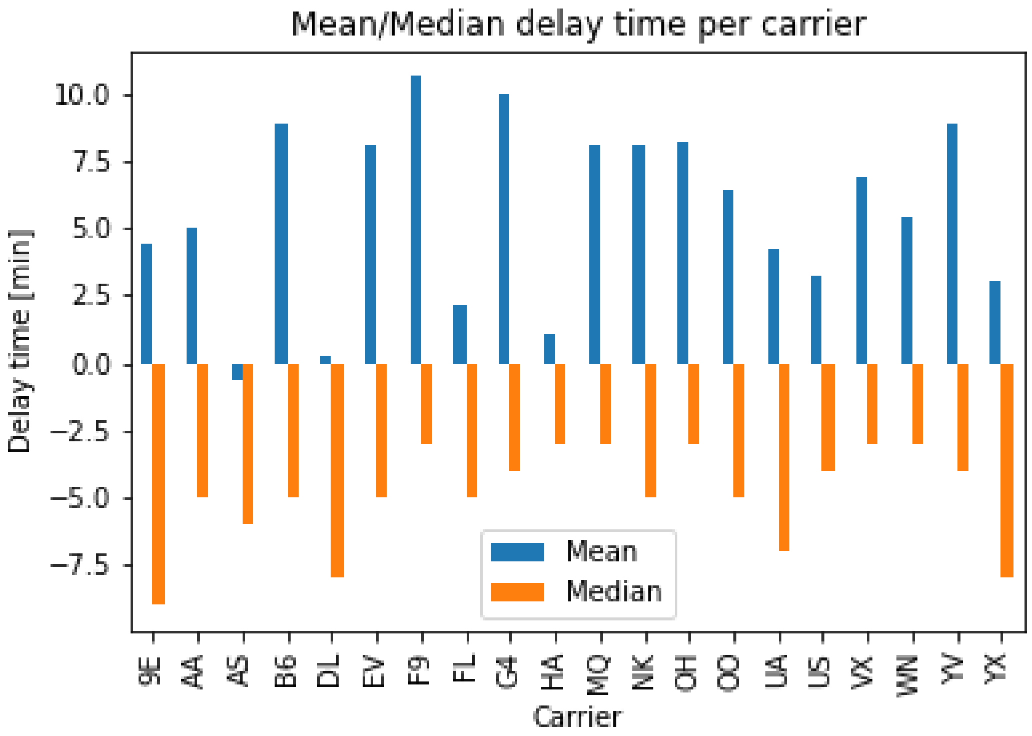

Mean/Median delay time per quarter of year
Mean
Median
Delay time [min]
6
4
2
0
−2
−4
−6
1
2
3
4
Quarter


Mean/Median delay time per month of year
Mean
Median
Delay time [min]
8
6
4
2
0
−2
−4
−6
1
2
3
4
5
6
7
8
9
10
11
12
Month

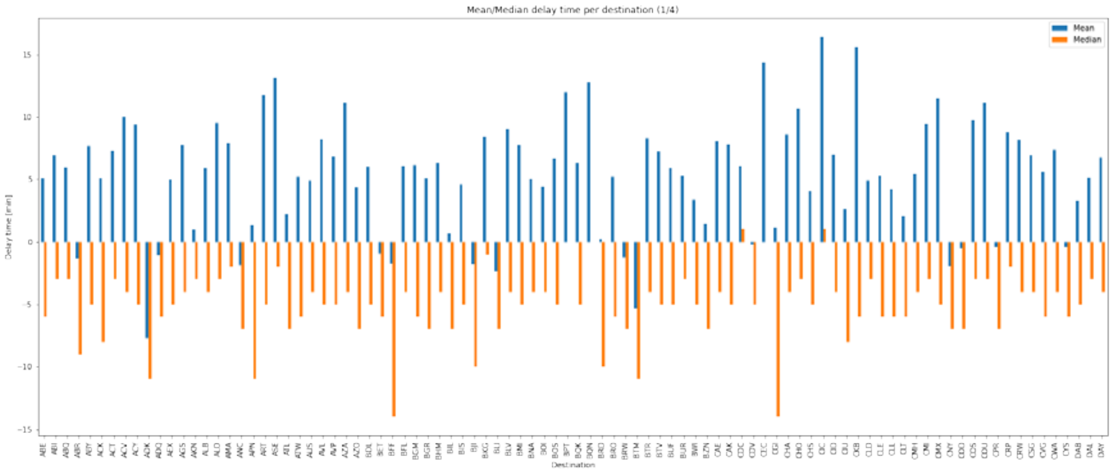

Mean/Median delay time per day of week
Mean
Median
Delay time [min]
6
4
2
0
−2
−4
−6
0
1
2
3
4
5
6
Day of Week

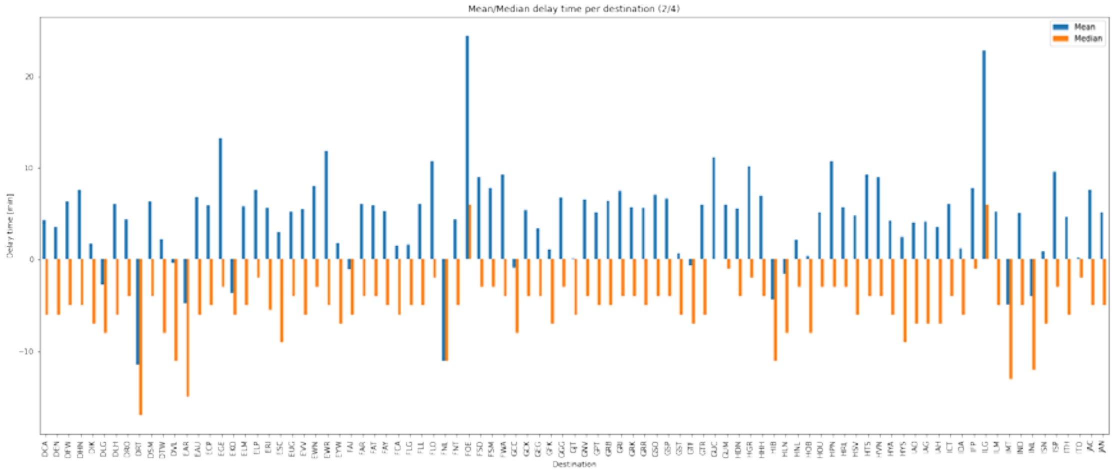
Mean/Median

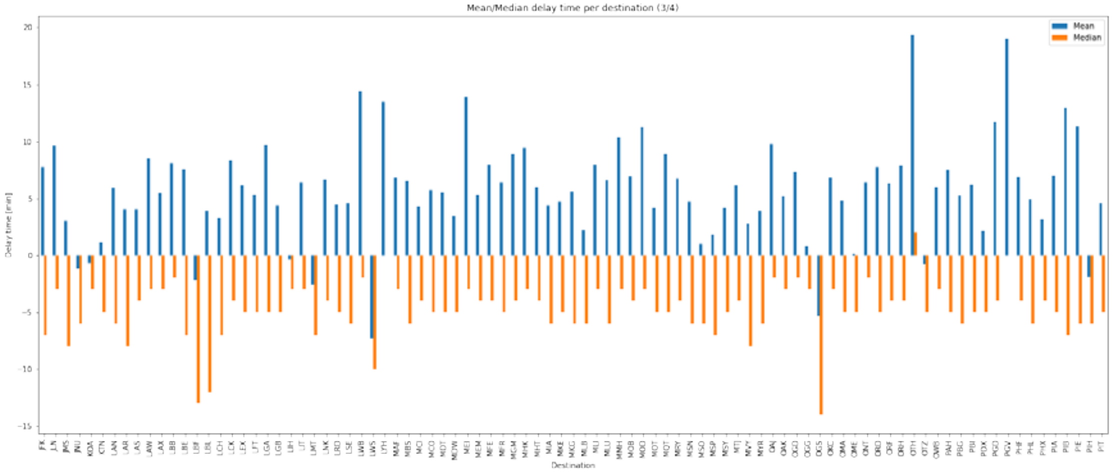

Mean/Median delay time per week of year
Mean
Median
Delay time [min]
15
10
5
0
−5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
Week

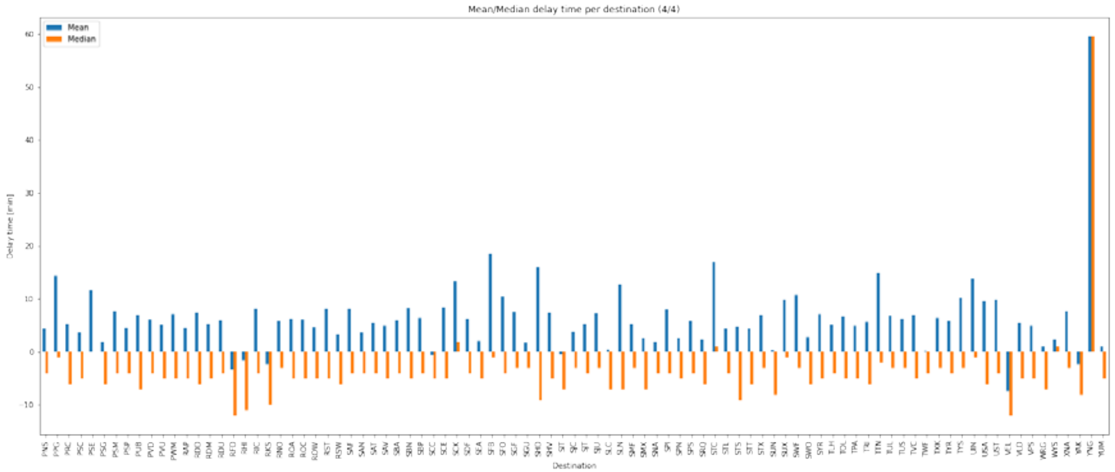

Mean/Median delay time per carrier
Delay time [min]
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
Mean
Median
9E
AA
AS
B6
DL
EV
F9
FL
G4
HA
MQ
NK
OH
OO
UA
US
VX
WN
YV
YX
Carrier

Mean/Median

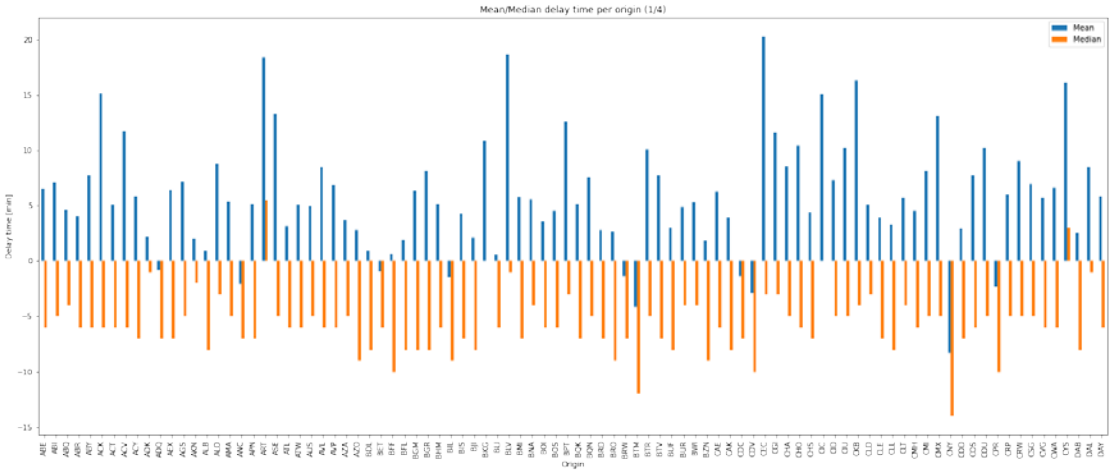

Mean/Median delay time per destination (1/4)

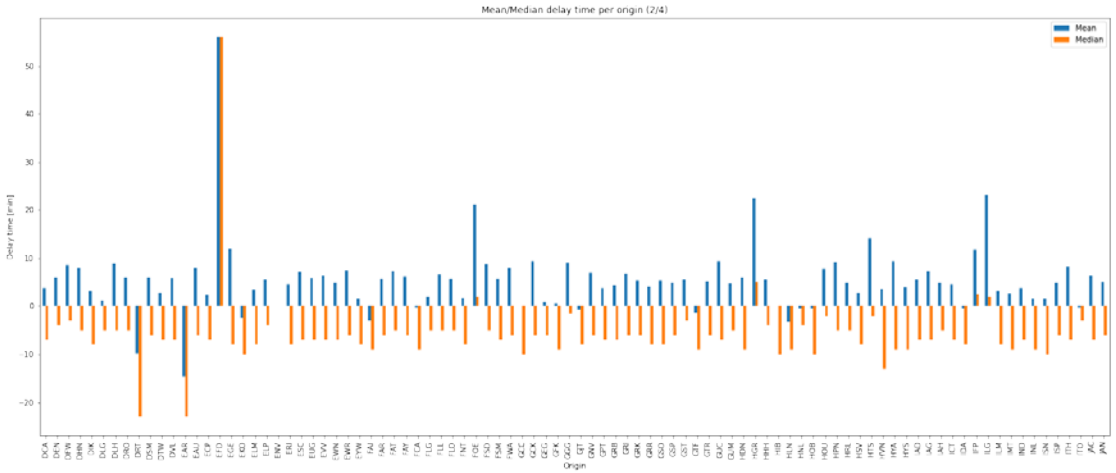

Mean/Median delay time per destination (2/4)

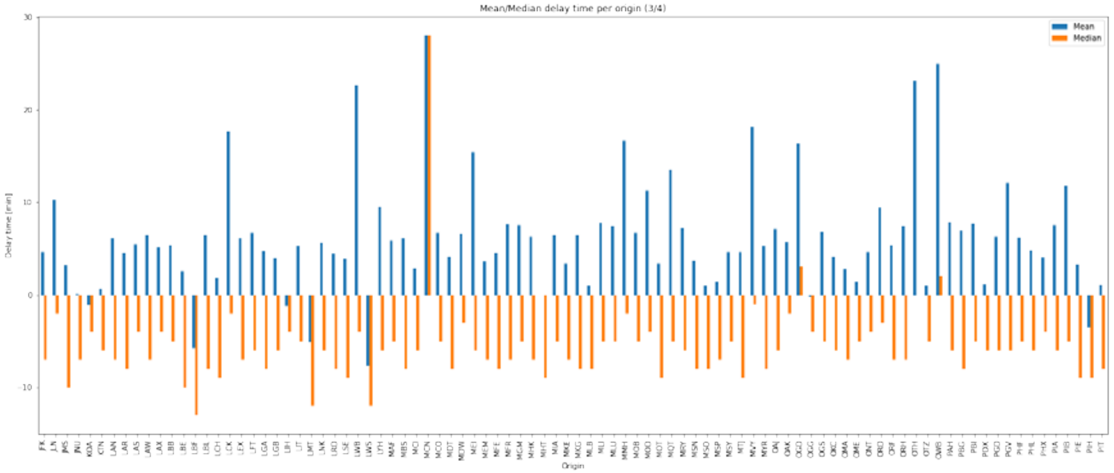

Mean/Median delay time per destination (3/4)

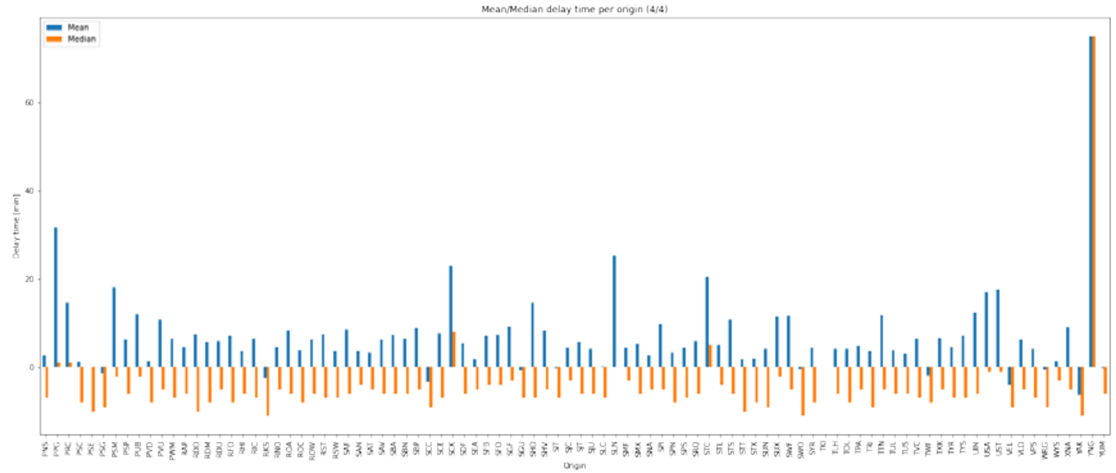

Mean/Median delay time per destination (4/4)

Mean/Median


Mean/Median delay time per origin (1/4)
Mean
Median
Delay time [min]
20
15
10
5
0
-5
-10
-15
Origin


Mean/Median delay time per origin (2/4)
Mean
Median
Delay time [min]
50
40
30
20
10
0
-10
-20
Origin


Mean/Median delay time per origin (3/4)
Mean
Median
Delay time [min]
30
20
10
0
-10
Origin


Mean/Median delay time per origin (4/4)
Mean
Median
Delay time [min]
60
40
20
0
Origin

- Year: minimum delay time was in 2016, and maximum delay time was in 2014
- Quarter of year: minimum delay time is in the fourth quarter of year, and maximum delay time is in the second quarter of year
- Month of year: minimum delay time is in September, and maximum delay time is in June
- Week of year: minimum delay time is in the 44. week, and maximum delay time is in the 1. week of year
- Day of week: minimum delay time is on Saturday, and maximum delay time is on Friday
- Carrier: minimum delay time is for carrier AS, and maximum delay time is for carrier F9
- Origin: minimum delay time is for origin EAR, and maximum delay time is for origin YNG
- Destination: minimum delay time is for destination DRT, and maximum delay time is for destination YNG

Gradient Boosting Regressor

- Gradient Boosting method is used for the prediction of delay time, and also for the importance of features
- All features are used except Cancellation code because most of the data are missing values
- Month and Day of Week are used as features instead of Date
- Metrics for model evaluation:
  - $R^2 = 0.993$
  - Mean Squared Error = 39.46
  - Mean Absolute Error = 4.19

Gradient Boosting Regressor

- Order of features, given by measuring the impact on delay time (Gini importance)
  1. DEP_DELAY - departure delay in minutes
  2. NAS_DELAY - delay due to National Aviation System in minutes
  3. TAXI_OUT - taxi out time in minutes
  4. LATE_AIRCRAFT_DELAY - previous flights caused delays (in minutes)
  5. CARRIER_DELAY - delay occurred due to carrier in minutes
  6. TAXI_IN - taxi in time in minutes
  7. WEATHER_DELAY - delay due to extreme weather conditions in minutes
  8. ACTUAL_ELAPSED_TIME - actual elapsed time in minutes
  9. CRS_ELAPSED_TIME - planned elapsed time in minutes
  10. OP_CARRIER - carrier
  11. CRS_DEP_TIME - planned departure time
- Other features don’t have any impact on delay time