Flight cancellations and delays cost many billion $ for US economy. Estimating the rate of flight delays/cancellations is important to assess system performance, to reduce the unwanted cost to passengers, airports and airlines, and to guide investment decisions and policy development for the US National Air Transportation System as well as the US Bureau of Transportation.

The dataset is from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics for all domestic flights operated by large air carriers in 2015:

<https://www.kaggle.com/usdot/flight-delays>

It includes three datasets and the main one is flight.csv. Each entry of the **flights.csv** file corresponds to a delayed or cancelled flight. In total, the dataset comprises **5,819,079 flights through 31 variables.**

The dataset is quite big and contains many “NAs” as well as some missing values that requires cleaning. Dates are coded according to 4 key variables: YEAR, MONTH, DAY, and DAY\_OF\_WEEK, which should be properly formatted; moreover; the hour of the take-off is coded as a float which is not convenient to use. The data need to be merged with other potentially useful datasets that has exhaustive description of the airports. The third dataset is very small and is just contains the full name of airlines along with their abbreviated versions.

> summary (df)

YEAR MONTH DAY DAY\_OF\_WEEK AIRLINE FLIGHT\_NUMBER

Min. :2015 Min. : 1.000 Min. : 1.0 Min. :1.000 WN :1261855 Min. : 1

1st Qu.:2015 1st Qu.: 4.000 1st Qu.: 8.0 1st Qu.:2.000 DL : 875881 1st Qu.: 730

Median :2015 Median : 7.000 Median :16.0 Median :4.000 AA : 725984 Median :1690

Mean :2015 Mean : 6.524 Mean :15.7 Mean :3.927 OO : 588353 Mean :2173

3rd Qu.:2015 3rd Qu.: 9.000 3rd Qu.:23.0 3rd Qu.:6.000 EV : 571977 3rd Qu.:3230

Max. :2015 Max. :12.000 Max. :31.0 Max. :7.000 UA : 515723 Max. :9855

(Other):1279306

TAIL\_NUMBER ORIGIN\_AIRPORT DESTINATION\_AIRPORT SCHEDULED\_DEPARTURE DEPARTURE\_TIME

: 14721 ATL : 346836 ATL : 346904 Min. : 1 Min. : 1

N480HA : 3768 ORD : 285884 ORD : 285906 1st Qu.: 917 1st Qu.: 921

N484HA : 3723 DFW : 239551 DFW : 239582 Median :1325 Median :1330

N488HA : 3723 DEN : 196055 DEN : 196010 Mean :1330 Mean :1335

N493HA : 3585 LAX : 194673 LAX : 194696 3rd Qu.:1730 3rd Qu.:1740

N478HA : 3577 SFO : 148008 SFO : 147966 Max. :2359 Max. :2400

(Other):5785982 (Other):4408072 (Other):4408015 NA's :86153

DEPARTURE\_DELAY TAXI\_OUT WHEELS\_OFF SCHEDULED\_TIME ELAPSED\_TIME

Min. : -82.00 Min. : 1.00 Min. : 1 Min. : 18.0 Min. : 14

1st Qu.: -5.00 1st Qu.: 11.00 1st Qu.: 935 1st Qu.: 85.0 1st Qu.: 82

Median : -2.00 Median : 14.00 Median :1343 Median :123.0 Median :118

Mean : 9.37 Mean : 16.07 Mean :1357 Mean :141.7 Mean :137

3rd Qu.: 7.00 3rd Qu.: 19.00 3rd Qu.:1754 3rd Qu.:173.0 3rd Qu.:168

Max. :1988.00 Max. :225.00 Max. :2400 Max. :718.0 Max. :766

NA's :86153 NA's :89047 NA's :89047 NA's :6 NA's :105071

AIR\_TIME DISTANCE WHEELS\_ON TAXI\_IN SCHEDULED\_ARRIVAL

Min. : 7.0 Min. : 21.0 Min. : 1 Min. : 1.00 Min. : 1

1st Qu.: 60.0 1st Qu.: 373.0 1st Qu.:1054 1st Qu.: 4.00 1st Qu.:1110

Median : 94.0 Median : 647.0 Median :1509 Median : 6.00 Median :1520

Mean :113.5 Mean : 822.4 Mean :1471 Mean : 7.43 Mean :1494

3rd Qu.:144.0 3rd Qu.:1062.0 3rd Qu.:1911 3rd Qu.: 9.00 3rd Qu.:1918

Max. :690.0 Max. :4983.0 Max. :2400 Max. :248.00 Max. :2400

NA's :105071 NA's :92513 NA's :92513

ARRIVAL\_TIME ARRIVAL\_DELAY DIVERTED CANCELLED CANCELLATION\_REASON

Min. : 1 Min. : -87.00 Min. :0.00000 Min. :0.00000 :5729195

1st Qu.:1059 1st Qu.: -13.00 1st Qu.:0.00000 1st Qu.:0.00000 A: 25262

Median :1512 Median : -5.00 Median :0.00000 Median :0.00000 B: 48851

Mean :1476 Mean : 4.41 Mean :0.00261 Mean :0.01545 C: 15749

3rd Qu.:1917 3rd Qu.: 8.00 3rd Qu.:0.00000 3rd Qu.:0.00000 D: 22

Max. :2400 Max. :1971.00 Max. :1.00000 Max. :1.00000

NA's :92513 NA's :105071

AIR\_SYSTEM\_DELAY SECURITY\_DELAY AIRLINE\_DELAY LATE\_AIRCRAFT\_DELAY WEATHER\_DELAY

Min. : 0 Min. : 0 Min. : 0 Min. : 0 Min. : 0

1st Qu.: 0 1st Qu.: 0 1st Qu.: 0 1st Qu.: 0 1st Qu.: 0

Median : 2 Median : 0 Median : 2 Median : 3 Median : 0

Mean : 13 Mean : 0 Mean : 19 Mean : 23 Mean : 3

3rd Qu.: 18 3rd Qu.: 0 3rd Qu.: 19 3rd Qu.: 29 3rd Qu.: 0

Max. :1134 Max. :573 Max. :1971 Max. :1331 Max. :1211

NA's :4755640 NA's :4755640 NA's :4755640 NA's :4755640 NA's :4755640

Additional datasets that may be linked to the flight delays/cancellations dataset:

The literature describes three main reasons for delays/cancellations: severe weather, airline issues (e.g. inadequate staffing) and air traffic congestion. We have attempted to identify datasets that provide markers of these 3 reasons for delay/cancellation.

Severe weather: We have obtained a csv file with records of severe weather warnings from the National Oceanic and Atmospheric Administration’s Severe weather database (NCDC/NOAA <https://www.ncdc.noaa.gov/swdi/#Intro>). The dataset of interest includes

-the severe weather event date and time as MM/DD/YYYY 12 hour time AM/PM

-the state name and FIPS (i.e. state identification code)

-the event\_type e.g. hail, wind, ice, etc as text

Air traffic: Data on airport-specific air traffic is provided by the Bureau of Transportation Statistics at <https://www.transtats.bts.gov/Data_Elements.aspx?Data=2>. Unfortunately these data are provided overall and individually by airport and is only available at the monthly temporal resolution. We will continue to search for alternatives but this may be the best we can do. Below is the 2015 cut of the data for all airports combined. Individual airports will have to be downloaded one at a time if this resolution is deemed adequate for prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Flights (All Carriers - All Airports )** | | |  |  |
| **Year** | **Month** | **DOMESTIC** | **INTERNATIONAL** | **TOTAL** |
| 2015 | 1 | 634,868 | 120,629 | 755,497 |
| 2015 | 2 | 570,076 | 107,407 | 677,483 |
| 2015 | 3 | 684,009 | 124,658 | 808,667 |
| 2015 | 4 | 669,870 | 122,749 | 792,619 |
| 2015 | 5 | 690,230 | 125,341 | 815,571 |
| 2015 | 6 | 697,511 | 127,980 | 825,491 |
| 2015 | 7 | 732,894 | 137,374 | 870,268 |
| 2015 | 8 | 724,134 | 134,508 | 858,642 |
| 2015 | 9 | 658,506 | 115,656 | 774,162 |
| 2015 | 10 | 687,829 | 117,640 | 805,469 |
| 2015 | 11 | 651,293 | 112,641 | 763,934 |
| 2015 | 12 | 659,939 | 125,856 | 785,795 |