Introduction

In the last ten years, according to the Bureau of Transportation Statistics (BTS), only 79.63% [1] of all flights have performed on time. Only a few remaining percentage were cancelled or diverted, less than 2%; rest of them were delayed mainly due to late arriving aircraft followed by the cause of the national aviation system and air carrier. A flight is considered delayed when a flight arrives or departs 15 or more minutes than the scheduled time. Averagely speaking, 720 million people [2] were on board and 144 million of those were affected by flight delays caused by five main reasons. Those reasons are [3]:

1. **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
2. **Extreme Weather:** Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
3. **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
4. **Late-arriving aircraft:** A previous flight with the same aircraft arrived late, causing the present flight to depart late.
5. **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

These series of delays cause a serious financial burden on airlines. In 2010, the Federal Aviation Administration commission estimated that flight delays cost the airline companies $8 billion a year, most of which due to increased budget on crews, fuel and maintenance [4]. It is also worth noting that this cost adds up to $32 billion in a year by accounting for other lateral costs such as passenger cost and indirect effect on associated businesses.

By addressing delayed flights problems, it is possible to reduce its heavy financial load on airlines, passengers and GDP. That is,

(1) letting passengers know about the probability of a particular flight being delayed before booking can help them to plan on a better date and time. In addition, it will save passengers more than a million hours of unnecessary delays per year [5].

(2) booking websites, airline companies, travel agencies can provide better customer service. Providing likelihood of flight delays to clients can decrease the number of complains and increase customer satisfaction.

(3) airline companies can pinpoint important factors that cause flight delays and take precautions depending on this information. The principal benefit of this study to airlines is the reduction in additional operating cost caused by delays.

Overall, the findings of this study can provide a high-profile achievement by addressing aviation delay problems with a robust prediction model and helping people and businesses better on planning their flights

Data Acquiring & Wrangling

In this section, I will provide information about what data acquiring and wrangling processes were taken.

How data is gathered: For this study, I used four different datasets from various sources.

1) Airline on-time performance dataset is available on The Bureau of Transportation Statistics' website. In the process of data scraping, I used the selenium library in Python. This library works with a specific version of a Chrome driver and can be downloaded from this page. It downloads the monthly data to the ~/data/flight_data folder. From then on, a shell script concatenates these monthly flight data into a single csv file under the same folder.

2) Weather data was obtained from Iowa State University's Environmental Mesonet Platform. This platform works like an API service which needs modification of a requested URL for each inquiry. The python script for scraping data can be found on this GitHub repo.

3) Airport data was obtained from The Bureau of Transportation Statistics' website manually. This dataset has a substantial amount of information such as airport names, location, timezone etc.

4) ICAO data was acquired from OpenFlight dataset. This data is needed because The Environmental Mesonet Platform of Iowa State University requires ICAO codes as station IDs. This information does not exist in the airport dataset obtained from BTS.

How data is cleaned: Airline on-time performance dataset have redundant out-of-scope information, so I only kept 30 necessary features in the data gathering process. I created a dictionary called flag to identify null entries (condition 1), out of range entries (condition 3), and both null and out of range entries (condition 2). Certain implementations have been performed based on flag values. For instance, a week day column has to be in between 1 and 7. Any entry that has null in that column needs to be imputed using pandas dt accessor. Here is the list of ranges of certain columns.

Column	Minimum	Maximum	Treatment (or Remarks)
Date	Given	Given	Any date entries out of a given range is removed.
Week Day	1	7	Imputed if possible (using dt accessor of pandas)
IATA	N/A	N/A	Any null entries are removed.
TailNum	N/A	N/A	Any null entries are removed.
OrgAirID	10001	16878	Any out of range entries are removed.
DestAirID	10001	16878	Any out of range entries are removed.
OrgMarID	30001	36845	Imputed if possible. Otherwise removed.
DestMarID	30001	36845	Imputed if possible. Otherwise removed.
Div	0	1	Any wrong entries are imputed based on certain columns.
Cncl	0	1	Any wrong entries are imputed based on certain columns.
CnclCd	1	4	Any out of range entries are removed.
ScDepTime	0	2400	Any out of range entries are removed.
ScArrTime	0	2400	Any out of range entries are removed.
ScElaTime	N/A	N/A	Any null entries are removed.
DepTime	0	2400	Imputed if certain conditions met.
DepDelay	N/A	N/A	Imputed if certain conditions met.
TxO	N/A	N/A	Imputed if certain conditions met.
TxI	N/A	N/A	Imputed if certain conditions met.
WhOff	0	2400	Imputed if certain conditions met.
WhOn	0	2400	Imputed if certain conditions met.
ArrTime	0	2400	Imputed if certain conditions met.
ArrDelay	N/A	N/A	Imputed if certain conditions met.

In the final form of the data, null entries are only allowed when a flight is cancelled or diverted. The columns need to be null when a flight is cancelled are:

DepTime, DepDelay, TxO, WhOff, WhOn, TxI, ArrTime, ArrDelay, AcElaTime, AirTime

and when a flight is diverted are:

ArrTime, ArrDelay, AcElaTime, AirTime, TxI, WhOn

References

[1] https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1
[2] https://www.transtats.bts.gov/Data_Elements.aspx?Data=1
[3] https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
[4] https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf
[5] Sud, V. P., Tanino, M., Wetherly, J., Brennan, M., Lehky, M., Howard, K., & Oiesen, R. (2009). Reducing flight delays through better traffic management. *Interfaces*, *39*(1), 35-45.