Assessment: Data Analysis

Airline On-Time Performance

A dataset containing on-time performance of domestic flights which operated in United States during August 2018 is available at the following link: https://data.world/dot/airline-on-time-performance-statistics

(Note: you may need to log in with a Google account to view / download the data)

The “August 2018 Nationwide” file contains the delay information, while “Air Carriers” file contains the mapping of Airline ID to Airline Name. “Airports” file can be ignored.

Data description (you can ignore the other columns):

- `fl_date`: Flight date
- `op_carrier_airline_id`: Airline ID (use the Air Carriers file to map to airline name)
- `tail_num`: Aircraft registration number, also referred to as tail number
- `op_carrier_flight_num`: Flight number
- `origin`: Origin airport IATA code
- `dest`: Destination airport IATA code
- `crs_dep_time`: Scheduled local departure time (CRS refers to carrier reservation system)
- `dep_time`: Actual local departure time
- `dep_delay`: Departure time delay in minutes, which is the difference between scheduled and actual departure time
- `dep_delay_new`: Adjusted departure time delay in minutes, where the negative `dep_delay` value (flights which departed early) is set to 0.
- `crs_arr_time`: Scheduled local arrival time (CRS refers to carrier reservation system)
- `arr_time`: Actual local arrival time
- `arr_delay`: Arrival time delay in minutes, which is the difference between scheduled and actual arrival time
- `arr_delay_new`: Adjusted arrival time delay in minutes, where the negative `arr_delay` value (flights which arrived early) is set to 0.
- `cancelled`: A binary value which indicates if the flight was cancelled
- `crs_elapsed_time`: Scheduled flight duration in minutes
- `actual_elapsed_time`: Actual flight duration in minutes
- `carrier_delay`: Delay in minutes caused by factors within the airline’s control, such as maintenance, equipment change, crew, boarding, etc.
- `weather_delay`: Delay in minutes caused by bad weather
- `nas_delay`: Delay in minutes caused by airspace or ground air traffic congestion (NAS – National Airspace System of the United States)
- `security_delay`: Delay in minutes caused due to security check or processing of passengers
- `late_aircraft_delay`: Delay in minutes caused due to late arrival of aircraft from previous flight

Download the data and carry out the following tasks to the best of your ability, in any tool of your choice (Excel, Tableau, PowerBI, Python, etc.). Feel free to play around with the data, slice and dice it as much as you like – this is an open-ended exercise.

Summarize your results / answers in a Word document, PowerPoint presentation, Jupyter notebook, or any other medium of your choice.

1. Generate descriptive statistics report (table) for average, minimum, and maximum departure delay, arrival delay, and cancellation rate aggregated by origin airport, destination airport, and airline.
2. Which airline has the best on-time performance (OTP) measured by median departure delay and median arrival delay in minutes?

3. Which airline has the best OTP measured by proportion of departures within 15 minutes of scheduled departure time?
4. Rank the 5 delay types by overall number of occurrences, and by total amount of delay caused.
5. Create a scatterplot to examine the relationship between scheduled and actual flight duration. Do you notice anything interesting?
6. Consider the flights departing from Atlanta (ATL). During which hours of the day are the flights most delayed? During which hours of the day are you most likely to leave on time?
7. Consider the flights departing from 3 New York area airports: John F. Kennedy (JFK), Newark Liberty (EWR), and LaGuardia (LGA). If you took a flight departing from one of these airports, in which airport would you have the highest likelihood of delay if your flight departed between 7 am – 8 am on a Monday?
8. Plot the distribution of departure delay and arrival delay in minutes for all United Airlines (UA) flights. Now, plot it separately for flights which departed from Houston (IAH) and Chicago O'Hare (ORD) for each of the 5 delay types. Are there any differences in distribution of delay duration by delay type between these two airports?
9. Carrier delay represents the delay caused due to factors in control of the operating airline. Which airline has the highest rate of carrier-induced delays? What proportion of total delay duration suffered by this airline is due to factors within its control?
10. Late Aircraft delay is caused by a previous flight operated by the same aircraft (tail number) arriving late, thereby causing a delay on the subsequent flight. Based on this, I might conclude that the frequency of this type of delay would increase during the course of the day – so that early morning departures are less likely to be delayed because of late aircraft compared to flights scheduled later in the day. Can you prove or disprove this hypothesis using the data provided?