Sophie Giacobbe

Dr. Asa Ben-Hur

DSCI 235

04 April 2023

Project Proposal

Since I stepped foot on my first airplane when I was just a few months old, my dad has kept track of every flight I've ever been on. 384 flights later, I've come to realize both my passion for travel and my passion for data. Though my personal flight data set isn't nearly large enough nor complex enough to build an entire project around, I knew immediately that I wanted to explore something similar.

Through Kaggle, I was able to find a dataset that fit my interests perfectly. The Department of Transportation keeps a record of all flight information by month. This information was compiled across the year 2015 into Flights.csv[1] and posted with the question: which airline should you fly on to avoid significant delays? Below is a list of included variables as well as their descriptions. Unless otherwise specified, time is measured in minutes and distance is measured in miles.

1. YEAR: year of flight
2. MONTH: month of flight
3. DAY: day of flight
4. DAY_OF_WEEK: weekday of flight
5. AIRLINE: IATA[1] airline code
6. FLIGHT_NUMBER: airline flight code
7. TAIL_NUMBER: aircraft identification number
8. ORIGIN_AIRPORT: IATA origin airport code
9. DESTINATION_AIRPORT: IATA destination airport code
10. SCHEDULED_DEPARTURE[3]: scheduled departure time
11. DEPARTURE_TIME[3]: actual departure time
12. DEPARTURE_DELAY: time between scheduled departure and actual departure
13. TAXI_OUT: time between gate departure and take off
14. WHEELS_OFF[3]: time of take off
15. SCHEDULED_TIME: scheduled duration
16. ELAPSED_TIME: time between gate departure and gate arrival
17. AIR_TIME: time between take off and landing
18. DISTANCE: distance between airports
19. WHEELS_ON[3]: time of landing
20. TAXI_IN: time between arrival and gate arrival
21. SCHEDULED_ARRIVAL[3]: scheduled arrival time
22. ARRIVAL_TIME[3]: actual arrival time
23. ARRIVAL_DELAY: time between scheduled arrival and actual arrival

[1] https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv

[2] International Air Transport Association

[3] Time in HHMM format, i.e. 1:07 pm is formatted 1307 & 12:05 am is formatted 0005 or 5

24. DIVERTED: equal to 1 if flight landed at a different airport than originally intended
25. CANCELLED: equal to 1 if flight was cancelled
26. CANCELLATION_REASON: equal to A if cancellation reason is due to the airline or carrier, B if due to weather, C if due to National Air System, and D if due to security
27. AIR_SYSTEM_DELAY: equal to 1 if delay occurs due to National Air System
28. SECURITY_DELAY: equal to 1 if delay occurs due to security
29. AIRLINE_DELAY: equal to 1 if delay occurs due to airline
30. LATE_AIRCRAFT_DELAY: equal to 1 if delay occurs due to late aircraft
31. WEATHER_DELAY: equal to 1 if delay occurs due to weather

I plan to categorize this project into two major sections: General and Timing. Within the Timing section, I will have three subsections: Delays, Taxiing, and Scheduled Time. Below is a brief outline of the questions I plan to answer, along with a tentative timeline. This will serve only as a starting point, leading to more questions and subsequently more opportunity for analysis.

General:

(complete by April 7)

- Which airport has the most departing flights? Arriving?
- What is the most popular cancellation reason?
- What is the most popular day to fly on?
- Which airline has the most flights in 2015?

Timing:

Delays (complete by April 14)

- Which airline has the most delays?
- What is the most popular reason for delay?
- Which destination airport has the most delays?

Taxiing (complete by April 21)

- Which airline taxis for the longest on average?
- Which airline taxis for the shortest on average?
- Which destination airport has the longest taxi time on average?
- Which destination airport has the shortest taxi time on average?

Scheduled Time (complete by April 28)

- Which airline has the longest scheduled flight between two airports?
- Which airline has the biggest gaps between scheduled arrival time and actual arrival time?
    - For both early arrival and late arrival
    - Is one airline more or less punctual than the rest?

[1] https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv
[2] International Air Transport Association
[3] Time in HHMM format, i.e. 1:07 pm is formatted 1307 & 12:05 am is formatted 0005 or 5