Airline On-Time Streaming Data Analysis

Arwa EL-Hawwat
Rutgers University
Piscataway, NJ, USA
ame126@scarletmail.rutgers.edu

Jaini Patel
Rutgers University
Piscataway, NJ, USA
jp1891@scarletmail.rutgers.edu

Rahul Dev Ellezhuthil
Rutgers University
Piscataway, NJ, USA
re263@scarletmail.rutgers.edu

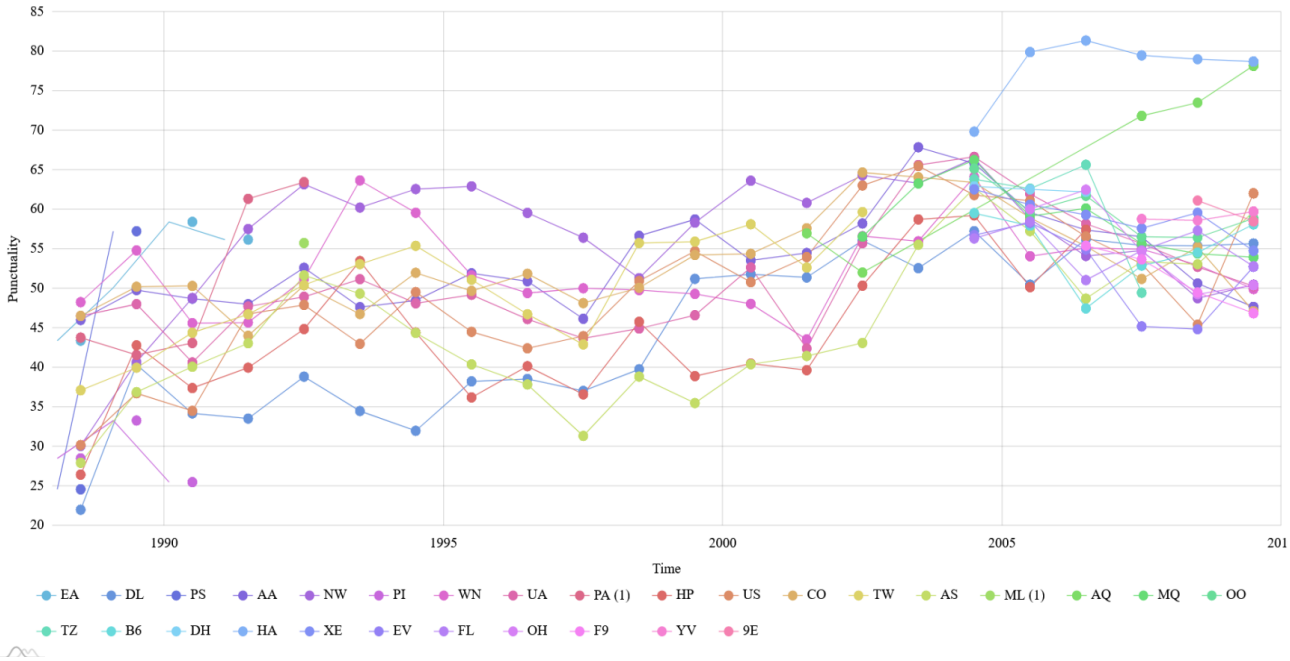
Supervised by Professor James Abello Mondero, Harshini Bonam, & Haoyang Zhang for CS543 Group 6, Fall 2021

ABSTRACT

Airline consumers today are overwhelmed by the sheer number of carriers and flight patterns offered to them when choosing flights. While they can easily compare costs and layovers at booking time, it can be much more difficult to predict delays on the day of departure. Utilizing past airline data, we can extract a model to formulate possible delays and when they are most likely to occur. In addition, we are able to determine the likelihood of certain carriers or trip patterns to experience difficulties. Our project will follow the format of database queries in a streaming fashion, and will therefore involve us maintaining a window of viewing the data while simultaneously communicating with our already existing analyses to draw conclusions based on real-world flight time and delay patterns in the United States.

GOALS

Our main objective for this project is to process the streaming flight data using Spark Structured Streaming and to be able to answer the following questions using database queries and visualizations on them: Which airline carrier is the most reliable in terms of punctuality? What were the worst months to fly historically? What are the busiest airports and paths in the United States? We aim to be able to organize and display our findings in a simple process and web application model.



Visualization of different US airline carriers and their calculated punctuality as a percentage from 1987 - 2008 using amChart

REFERENCES

[1] 2008, *Data Expo 2009 - Airline on-Time Performance*, American Statistical Association, https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009

[2] 2008, *Data Expo 2009: Airline on time data*, Harvard Dataverse, V1, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7