**Link of the Dataset:**

<https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023?utm_source=chatgpt.com&select=flights_sample_3m.csv>

**About Dataset**

Airline Flight Delay and Cancellation Data, August 2019 - August 2023

US Department of Transportation, Bureau of Transportation Statistics [https://www.transtats.bts.gov](https://www.transtats.bts.gov/)

**Purpose**

The purpose of collecting this dataset was to facilitate programmatic retrieval via an API into an AWS EC2 instance for structured data analysis using Spark SQL. The dataset includes several versions, ranging from the complete set of 29 million rows to a reduced set of 3 million rows. This iterative refinement process was aimed at optimizing the use of AWS's distributed computing capabilities to achieve efficient querying and effective data manipulation. Best performance was achieved using the i3, m5, and c5 instances with upgraded storage.

**Context**

Source data related to flight delays and cancellations for January 2019 – August 2023 retireved from DOT *On-Time : Reporting Carrier On-Time Performance (1987-present)*

Variables include flight routes (origin, destination), time ranges for events (minutes, local time), delay and cancellation reasons/attributions (limited).

**Acknowledgements**

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report.

Retrieved in November 2023 using the application at [On-Time : Reporting Carrier On-Time Performance (1987-present)](https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr).

The source data was downloaded in subsets by month and joined by year. Most current available data for 2023 is from August. Data consolidation, transformation, wrangling, variable selection, and label updates were done in csvkit, Python and Excel.

**Attribution**

Original data sourced from the U.S. Department of Transportation, Bureau of Transportation Statistics [https://www.transtats.bts.gov](https://www.transtats.bts.gov/)

This transformed dataset adopts variable names from [Airline Delay and Cancellation Data, 2009 - 2018](https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data).