

Subject: Re: Examination of the Recent Airline Delays Data

Dear Lawrence,

Thank you for your recent communication outlining the investigative focus areas for the December 2019 and 2020 airline delay dataset.

I'm pleased to share that we have completed the analysis you requested. For detailed insights, please refer to the attached HTML file, which comprehensively documents our analytical process.

Additionally, I'd like to delve into our approach to data normalization in this email. Our main objective has been to minimize data duplication and uphold the integrity of the data. To achieve this, we have identified and separated key entities such as carriers, airports, and flights, with each category representing a distinct data set with shared characteristics. Below is an overview of the tables we created:

## Flights Table

- **Description:** Central to the database, it encompasses each flight's specifics, including year, month, carrier ID, and airport code.
- **Importance:** As a pivotal table, it logs critical flight details and connects to other tables like Carriers, Airports, and Delays through foreign keys, serving as the foundation for relational data analysis.

## Carriers Table

- **Description:** This table catalogs airline carriers, each marked by a unique carrier ID.
- **Importance:** Vital for sorting flight data by airline, it supports analyses such as delay frequency per carrier or their punctuality records.

## Airports Table

- **Description:** Details information about each airport, identified by a distinct airport code.
- **Importance:** Aids in geographic and specific airport-related analyses, like pinpointing airports with frequent delays or cancellations.

## Delays Table

- **Description:** Outlines various delay types for each flight, including categories like carrier, weather, NAS, security, and late aircraft, plus their durations.
- **Importance:** Complements the Flights table by offering detailed insights into delay causes and types.

Normalization offers several benefits, including:

1. **Scalability and Flexibility:** Our structure is designed for ease of updating and expanding, accommodating new data like additional carriers or airports or different delay types.

2. **Data Integrity and Validation:** By distributing data across multiple tables linked through keys, we ensure data accuracy and uniformity, reducing the likelihood of inconsistencies and errors.
3. **Query Efficiency:** This design enhances the efficiency of data queries, allowing for more effective extraction of insights like delay trends by carrier, airport performance, or monthly delay patterns.

Please feel free to reach out if you need further exploration of specific areas or modifications to the existing queries.

Kind regards,

Jun Hu