MTA Daily Ridership Data Cleaning & Preprocessing

Preparing data for analysis and forecasting

O

- **OBJECTIVE:**

– Analyze and forecast daily ridership data from the Metropolitan Transportation Authority (MTA).
– Build a Tableau dashboard and comprehensive report to highlight trends and insights.

D

- **DATASET CONTEXT:**

– Covers multiple MTA agencies including NYCT, MTABC, LIRR, Metro-North, Access-A-Ride, and Bridges & Tunnels.
– “The Metropolitan Transportation Authority (MTA) is a public-benefit corporation responsible for public transportation in New York…”

Data Challenges & Initial Issues

- **Common Data Problems Identified:**

– Missing values (e.g., ride_date_str, station_id, ridership counts)
– Negative or out-of-range numeric values
– Extra spaces in text fields
– Duplicate records

- **Impact:**

– Inconsistent, unreliable raw CSV input requires extensive cleaning before analysis.

Staging & Simulating Raw Data

- **Dropping Existing Tables:**

– Ensured no residual data by dropping staging and cleaned tables.

- **Inserting Sample Data:**

– Simulated raw CSV import with fields like ride_date_str, station_id, daily_ridership, etc.

- **Creating a Staging Table:**

– Sample rows purposely included missing data, negative values, extra spaces, and duplicates

– This mimics real-world issues encountered during data collection.

Data Cleaning & Transformation Process

01

• CONVERSION & FILTERING:

– Converted ride_date_str to DATE (excluding rows with missing or empty dates).
– Filtered out rows missing critical identifiers such as station_id.

02

• NUMERIC & TEXT CLEANING:

– Used CASE statements to set NULL/negative ridership values to 0 and cap extreme values (e.g., daily_ridership > 100000 capped at 100000).
– Employed TRIM() to remove extra spaces from text fields.

03

• DUPLICATE REMOVAL:

– Applied a window function (ROW_NUMBER()) to remove duplicate entries.
– Final cleaned data inserted into a table with a composite primary key (ride_date, station_id).
– “SQL allows users to extract specific information from large datasets efficiently by using commands

Performance Optimization & Quality Checks

- **Indexing:**

– Created indexes on key columns (station_id and ride_date) to enhance query performance.

- **Data Quality Checks:**

– Verified no default dates (e.g., '1900-01-01') remain.
– Ensured text fields are properly trimmed (no leading/trailing spaces).
– Identified any anomalous ridership values (0 or capped at 100000) for further review.

Data Analysis & Forecasting Roadmap

- **CLEANED DATASET:**

– Ready for in-depth analysis using SQL and Python (pandas, Matplotlib).

- **ANALYSIS GOALS:**

– Uncover trends in ridership across different MTA agencies.

– Build a forecasting model to predict future ridership patterns.

- **FINAL DELIVERABLES:**

– A Tableau dashboard showcasing key insights.

– A comprehensive report summarizing the methodology, findings, and forecasting outcomes.

Summary & Next Steps

RECAP:

- Successfully tackled major data challenges with a robust SQL cleaning process, addressing issues like missing values, negative/out-of-range numbers, extra spaces, and duplicates.
- Utilized powerful SQL tools (SELECT, WHERE, ORDER BY, subqueries) to ensure data integrity and prepare a reliable dataset.

CONCLUSION:

- Clean, reliable data is the foundation for extracting meaningful insights.
- With a clear set of analysis questions and initial visualizations, we’re poised to uncover trends that will drive informed decisions and successful forecasting.