Notebook 06: CSV Generation (Dashboard Data Builder)

Notebook 06 is exclusively dedicated to creating tidy, final-state CSV files for the Streamlit application. It takes the comprehensive analytical frameworks built earlier and refines them into easy-to-read, pre-aggregated tables, ensuring the dashboard loads quickly and starts instantly without needing to recompute millions of records.

1. Goal and Data Loading

The objective of this notebook is described as the **CUNY export builder**. It focuses on loading bus stops and routes data (GTFS), building proximity buffers around campuses, and mapping stops to routes. It also integrates external data by **fetching ridership via SoQL** (Socrata Query Language) for the selected CUNY routes and creates a clean fact table for the period of **January to August 2025**.

The notebook relies on the established data pipeline and analytical outputs, including the comprehensive violation records and CUNY campus references.

2. Core Methodological Steps (CUNY Focus)

This notebook executes several data transformation steps to create the final CUNY insights used in the `dashboard/insights/CUNY_Insights/` folder.

A. GTFS Data Integration and Mapping

The process starts by loading bus infrastructure data:

• It loads **GTFS stops and routes**. This data is essential for accurately mapping where bus stops are located and which routes they serve.

• It then **builds buffers around campuses with dynamic sizes**. This leverages the spatial analysis logic (like the 500-meter buffer zone established in Notebook 01) to link violations directly to the CUNY environment.

• The final step here is mapping the identified CUNY-proximate stops to the specific bus routes that serve them.

B. External Ridership Data Fetching (SoQL Integration)

A key feature of Notebook 06 is the integration of *external* data to add real-world context and impact metrics, particularly addressing Datathon Question 1 (CUNY utilization).

• The notebook uses the `fetch_ridership_for_routes` function which is designed to connect to the ridership endpoint (e.g., `https://data.ny.gov/resource/gxb3-akrn.csv`).

• It fetches ridership via **SoQL** for the identified CUNY routes. The notebook mentions using the **Socrata Query Language (SoQL)** for filtering and aggregating large datasets, including bus ridership data, using Python to submit queries.

• The query is executed in chunks of **300 routes** at a time (`chunk_size = 300`) to manage the volume of data requested via the endpoint. The query uses parameters such as `$select`, `$where` (filtering by `bus_route` and `transit_timestamp`), `$group`, and `$limit` to efficiently gather `total_ridership` data.

• The gathered ridership data is then aggregated by campus name, calculating the total ridership and merging it with violation data for a comprehensive `campus_summary`.

C. Final Fact Table Generation

This notebook creates a clean **tidy fact table** covering the period **January to August 2025**. This is the master dataset used to derive all the specific CUNY insight exports.

3. Dashboard Output Files Generated

Notebook 06 exports several dedicated CSV files into the `dashboard/insights/CUNY_Insights/` subdirectory, specifically designed to populate CUNY-focused pages in the Streamlit application:

CSV File Output Path	Content & Purpose	Derived Insights Supported

`campus_summary.csv`	Summary statistics for each CUNY campus, including **violation counts** and **total ridership**.	Provides immediate campus-level performance metrics, directly supporting the quantification of CUNY student impact.
`monthly_campus_trend.csv`	Monthly trend data for CUNY campuses.	Visualizes how violations change over time specifically within CUNY corridors, identifying temporal predictability patterns (e.g., semester peaks).
`violations_by_type_campus.csv`	Violation counts broken down **by type** for each CUNY campus.	Necessary for policy analysis, showing the distribution of exempt vs. non-exempt violations at critical locations (Datathon Question 2).
`routes_per_campus.csv`	Bus routes that serve each CUNY campus, integrated with **ridership data**.	Directly answers Datathon Question 1 ("Which MTA bus routes are highly utilized by CUNY students?").
`campus_route_ridership.csv`	Detailed ridership data for routes serving CUNY campuses.	Provides the granular data underpinning the summary files, allowing for deep dives in the dashboard.
`violations_monthly_trend.csv` (General)	Monthly trend of total violations aggregated by `year_month`, sorted for time series visualization.	Supports a general temporal trend dashboard component, providing context beyond just the CUNY subset.

4. Integration with the Streamlit Dashboard

The final export action of Notebook 06 ensures the Streamlit dashboard is immediately runnable and reflective of the latest analysis:

• **Instant Loading:** By exporting pre-aggregated CSVs, the dashboard can start instantly without expensive recomputation, which is critical for good dashboard user experience.

• **Narrative Support:** The generated CSVs feed the dashboard sections, notably the page called **"The Rolling Study Hall"** which introduces the student-centric frame, and the page called **"The ClearLane Solution"** which displays the finalized, actionable target list table (`clear_lane_target_list.csv` from Notebook 05).

• **Data Structure:** The dashboard relies on these CSVs for drawing visualizations, including weekday bars, hourly lines, and heatmaps, ensuring the visualization reflects the specific CUNY-centric temporal and spatial patterns discovered in the analysis.

In essence, Notebook 06 is the **production factory** that packages the complex analytical outputs into the accessible, high-impact data format required by the final presentation layer, making the data transparent and immediately available for visualization and executive review.