IS 525

DATA WAREHOUSING AND BUSINESS INTELLIGENCE

FINAL PROJECT REPORT

New York City Taxi Trip Analysis Dashboard

Submitted By:
Aditi Patil (aditiap2)
Sahil Shimpi (sshimpi2)
Viral Sheth (vmsheth2)

Guided By:
Prof. Michael Wonderlich

School of Information Sciences
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Abstract

Our project investigates taxi transportation dynamics in New York City through the development of an interactive, data-driven dashboard. The analysis is based on publicly available trip records from the New York City Taxi and Limousine Commission (TLC) for February 2025, comprising over 21 million entries across Yellow Taxis, Green Taxis, For-Hire Vehicles (FHV), and High Volume FHVs.

To manage the scale and complexity of the dataset, Microsoft Fabric was employed for data preprocessing and transformation, followed by Python for schema validation, and Tableau Prep for final ETL processing. Tableau was used to design and implement the visual analytics environment. The dashboards provide comprehensive insights into trip efficiency, fare structures, tipping behavior, payment method patterns, and geospatial pickup/drop-off trends. These dashboards enable operational visibility and decision-making for transportation planning and urban analytics.

Introduction

New York City's taxi system plays a critical role in the city's public transportation infrastructure, serving millions of passengers each month. With services spanning Yellow Taxis, Green Taxis, For-Hire Vehicles (FHV), and High Volume FHVs, the NYC taxi network provides a unique opportunity to study large-scale urban mobility, payment behaviors, and geospatial ride distribution.

To support this analysis, we selected publicly available data from the **Taxi and Limousine Commission (TLC)**, a reliable, government-maintained source that offers detailed, trip-level records. The TLC dataset includes timestamps, trip distances, fare breakdowns, pickup and drop-off zones, payment types, and ride-sharing indicators, making it well-suited for building robust business intelligence dashboards.

We focused specifically on **February 2025** for three reasons:

1. The dataset was among the most recent TLC releases available in Parquet format.

2. It offered a **manageable time window** (one month) while still containing **over 21 million records**, ensuring scale for deep analytical insights.

3. The month was free from extreme seasonal or holiday disruptions (unlike December or July), providing a more **neutral baseline for taxi behavior**.

This project leverages February 2025 trip data to design dashboards that reveal patterns in ride activity, fare behavior, and spatial trends, enabling data-driven decision-making in urban transportation.

Data Source

This project uses publicly available trip data from the New York City Taxi & Limousine Commission (TLC), which publishes detailed records for all licensed taxi and for-hire vehicle services in NYC. It consists of:

- **Yellow Taxis**
- **Green Taxis**
- **For-Hire Vehicles (FHV)**
- **High Volume For-Hire Vehicles (HVFHV)**

Dataset Details:

- Format: Parquet
- Size: ~21.3 million trip records for Feb 2025
- Fields: ~28 attributes including pickup/dropoff timestamps, zones, distance, fare, tips, tolls, and taxi type

All data was ingested into Microsoft Fabric, then transformed using Tableau Prep into a single flat CSV table.

Data Dictionary References:

We used the official NYC TLC data dictionaries to map fields across:

- Yellow Taxi Trip Records
- Green Taxi Trip Records
- For-Hire Vehicle Trip Records
- High Volume FHV Trip Records
- Taxi Zone Lookup Table

These documents guided our field selection and were essential for consistent interpretation across data sources.

Tools Used – Platforms, Software, and Their Roles

This project involved handling high-volume trip data in Parquet format, requiring a multi-stage toolchain that could adapt to changing technical constraints. The tools we used were not only selected for their individual capabilities, but also for how they complemented each other in an evolving ETL and dashboarding pipeline.

Tool	Purpose
Microsoft Fabric	Used initially for Parquet ingestion, schema inspection, and early transformations. Dataflow Gen2 was employed to build ingestion pipelines for the 2024 dataset.
Power BI (Semantic Models)	Explored for connecting Fabric data to dashboards. Used to create a dashboard by connecting Fabric data via semantic models.
Python	Used to interact with Parquet files directly. Python scripts helped convert Parquet to CSV, perform schema validation, and inspect file-level metadata.
DASK (via Google Colab)	Enabled parallelized conversion of large Parquet datasets into multiple CSV chunks. DASK provided a scalable and efficient alternative to single-threaded pandas workflows, helping streamline the transformation of February 2025 data.
Tableau Prep Builder	Replaced Fabric in the final ETL stage. Tableau Prep was used to clean fields, normalize schema inconsistencies across taxi types, add calculated fields (e.g., Tip%, Distance Bins), and join lookup tables like Taxi Zones. Final unified CSVs were exported for visualization.

Tableau Desktop	Served as the core dashboarding platform. Used to build all KPI tiles, reference line charts, dot plots, treemaps, heatmaps, bubble plots, and efficiency index visuals. Tableau's slicers and interactivity features helped explore taxi data across multiple dimensions.
ClickUp	Managed task delegation and planning. We used ClickUp to define dashboard ownership, track deliverables, prioritize enhancements, and align on shared deadlines during the final sprint.
Google Drive, Docs & Sheets	Facilitated collaborative planning and data validation. Used for versioning data files, documenting transformation steps, tracking schema changes, and ensuring seamless handoffs between ETL and dashboarding stages.

Each tool played a critical role in adapting to evolving requirements across ingestion, transformation, analysis, and visualization, culminating in dashboards that meet both technical and storytelling goals.

ETL and Data Modeling Strategy – Adapting to Technical Constraints

---

Initial Approach: Fabric and Semantic Models

We began by using Microsoft Fabric's Dataflow Gen2 to load and transform Parquet data for all taxi types across the year 2024. Our plan was to create a semantic model and connect it to Power BI dashboards. We successfully built a Power BI dashboard using this approach, showcasing fare breakdowns and borough-level revenue insights.

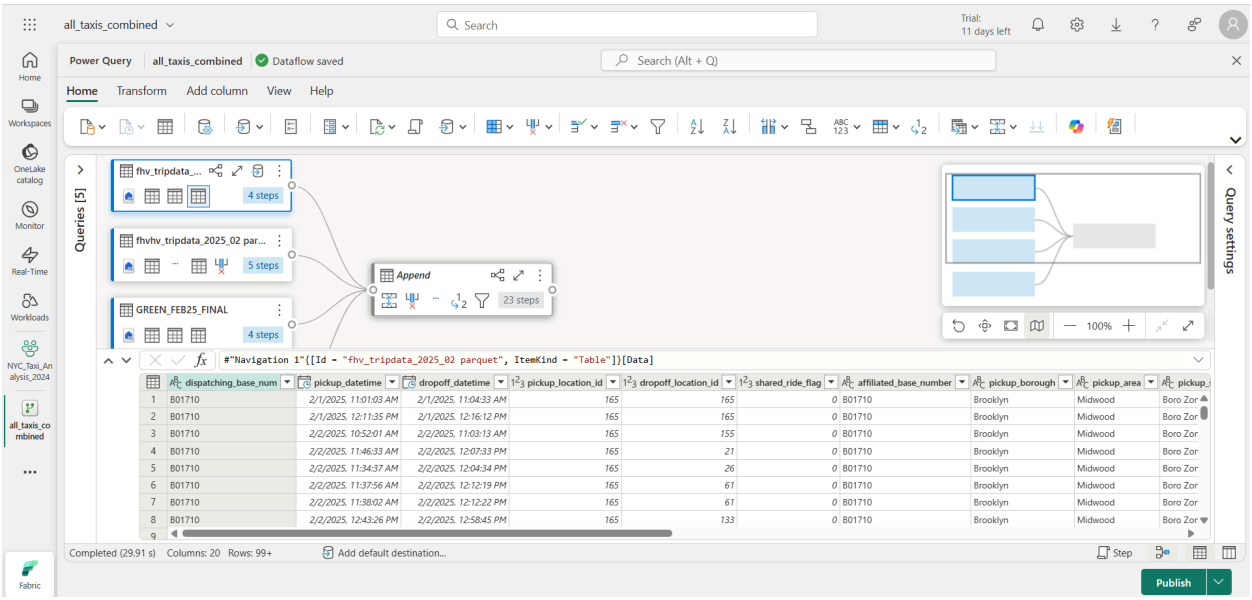


Fig. Dataflow Gen2 setup showing the schema and staging configuration

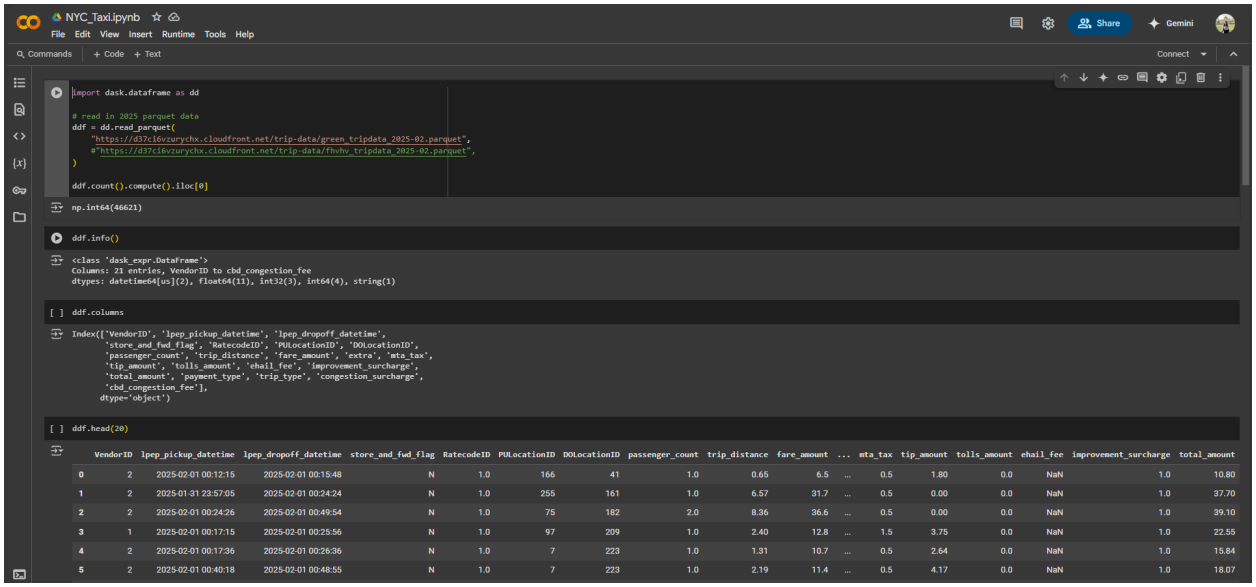
However, we quickly encountered the following issues:

- **Performance Bottlenecks:** Join operations failed due to large row volumes.
- **Staging Issues:** Only the schema was published, not the data rows.
- **Power BI Limitations:** Some semantic models only loaded the top N rows, which led to inaccurate dashboards.

Despite reducing the dataset to a single month (February 2025) and splitting the data by taxi type, these problems persisted.

Revised Approach: DASK and Tableau Prep

We pivoted to using DASK in Google Colab to convert the Parquet files into chunked CSVs. This approach allowed for parallel processing and manageable file sizes.



```
import dask.dataframe as dd

# read in 2025 parquet data
ddf = dd.read_parquet(
    "https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2025-02.parquet",
    #"https://d37ci6vzurychx.cloudfront.net/trip-data/fhvhv_tripdata_2025-02.parquet",
)

ddf.count().compute().iloc[0]
```

```
np.int64(46621)
```

```
ddf.info()
```

```
<class 'dask_expr.DataFrame'>
Columns: 21 entries, VendorID to cbd_congestion_fee
dtypes: datetime64[us](2), float64(11), int32(3), int64(4), string(1)
```

```
ddf.columns
```

```
Index(['VendorID', 'lpep_pickup_datetime', 'lpep_dropoff_datetime',
       'store_and_fwd_flag', 'RatecodeID', 'PULocationID', 'DOLocationID',
       'passenger_count', 'trip_distance', 'fare_amount', 'extra', 'mta_tax',
       'tip_amount', 'tolls_amount', 'ehail_fee', 'improvement_surcharge',
       'total_amount', 'payment_type', 'trip_type', 'congestion_surcharge',
       'cbd_congestion_fee'],
      dtype='object')
```

```
ddf.head(20)
```

	VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag	RatecodeID	PULocationID	DOLocationID	passenger_count	trip_distance	fare_amount	...	mta_tax	tip_amount	tolls_amount	ehail_fee	improvement_surcharge	total_amount
0	2	2025-02-01 00:12:15	2025-02-01 00:15:48	N	1.0	166	41	1.0	0.65	6.5	...	0.5	1.80	0.0	NaN	1.0	10.80
1	2	2025-01-31 23:57:05	2025-02-01 00:24:24	N	1.0	255	161	1.0	6.57	31.7	...	0.5	0.00	0.0	NaN	1.0	37.70
2	2	2025-02-01 00:24:26	2025-02-01 00:49:54	N	1.0	75	182	2.0	8.36	36.6	...	0.5	0.00	0.0	NaN	1.0	39.10
3	1	2025-02-01 00:17:15	2025-02-01 00:25:56	N	1.0	97	209	1.0	2.40	12.8	...	1.5	3.75	0.0	NaN	1.0	22.55
4	2	2025-02-01 00:17:36	2025-02-01 00:26:36	N	1.0	7	223	1.0	1.31	10.7	...	0.5	2.64	0.0	NaN	1.0	15.84
5	2	2025-02-01 00:40:18	2025-02-01 00:48:55	N	1.0	7	223	1.0	2.19	11.4	...	0.5	4.17	0.0	NaN	1.0	18.07

Fig. Python-based Parquet to CSV chunking workflow

We then transitioned to Tableau Prep Builder for final data cleaning and preparation:

- Unified field formatting across datasets
- Calculated derived fields such as Tip Percentage and Trip Efficiency
- Joined taxi zone lookup tables
- Removed corrupt or incomplete records

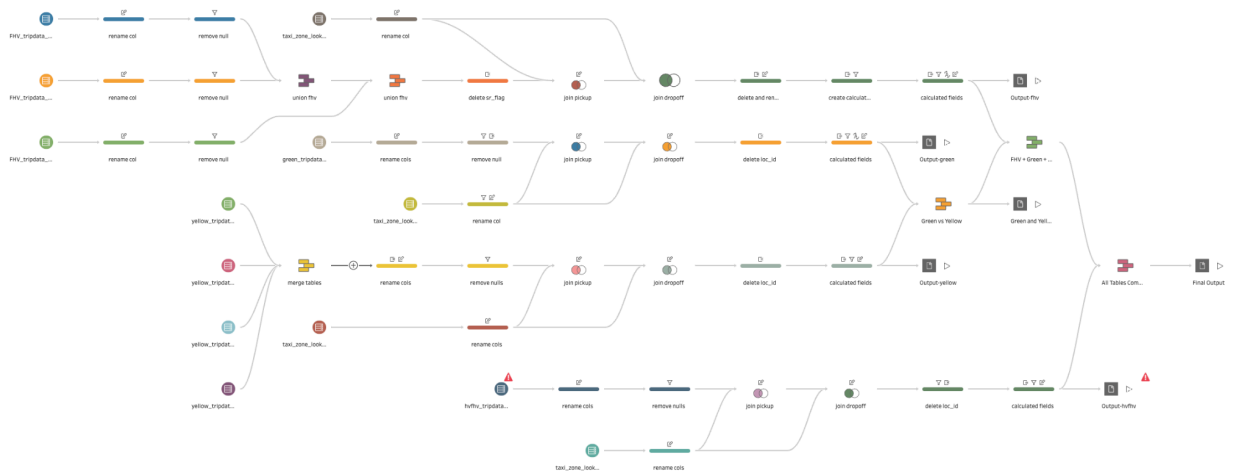


Fig. Final ETL flow illustrating joins, filters, and calculated fields

The output was a clean, flattened CSV that could be directly consumed in Tableau Desktop.

Data Modeling Decision

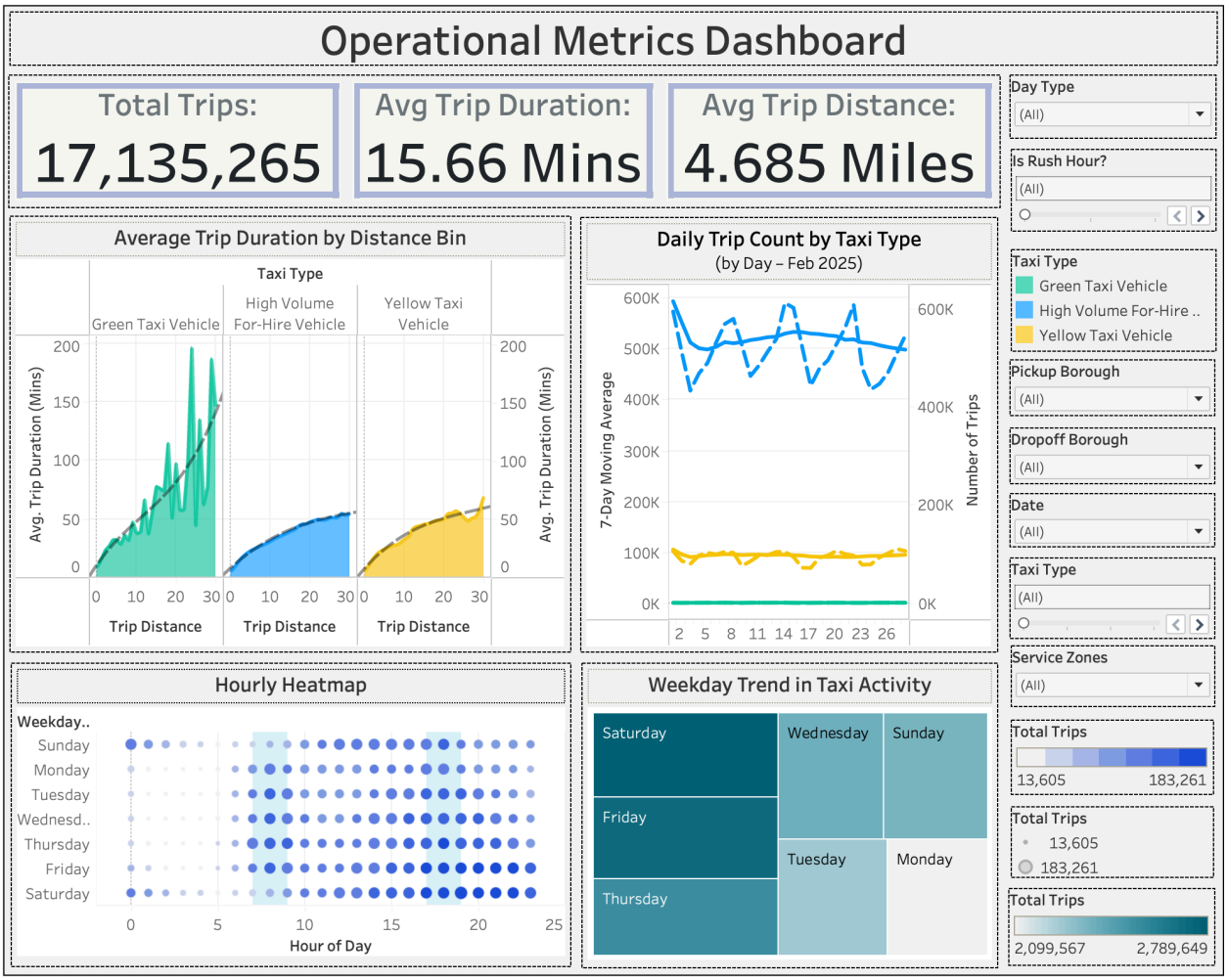
Due to the semantic model's limitations, we adopted a flat-table model for all visualizations. This decision was driven by:

- Better dashboard performance and responsiveness
- Full control over included fields and transformations
- Compatibility with Tableau's data handling capabilities

This approach ensured that we had a scalable and reliable pipeline for visualization, even if it meant stepping away from formal star schema practices.

Dashboard Overview

Operational Metrics Dashboard (Tableau):



Tools:

Tableau Desktop with full CSV data exported from *Tableau Prep Builder*

Overview:

The **Operational Metrics Dashboard** offers a focused, high-level summary of **trip volume**, **temporal demand**, and **efficiency trends** for New York City taxi services in **February 2025**. Built in Tableau using a fully preprocessed and schema-normalized flat CSV, the dashboard presents **eight coordinated visualizations** that emphasize operational patterns and peak activity

windows. It integrates **global filters** - including taxi type, rush hour flag, day type, pickup/dropoff boroughs, and service zones - allowing users to perform **multi-dimensional analysis** with ease.

Key Insights:

- **Over 17 million total trips** were recorded across all taxi types, with an **average duration of 15.66 minutes** and **average distance of 4.685 miles**, indicating the high operational scale of NYC's taxi ecosystem.

- The **Average Trip Duration by Distance Bin** visual clearly shows that **trip time grows with distance**, although **variability differs by taxi type**, with Green Taxis displaying greater inconsistency at longer ranges.

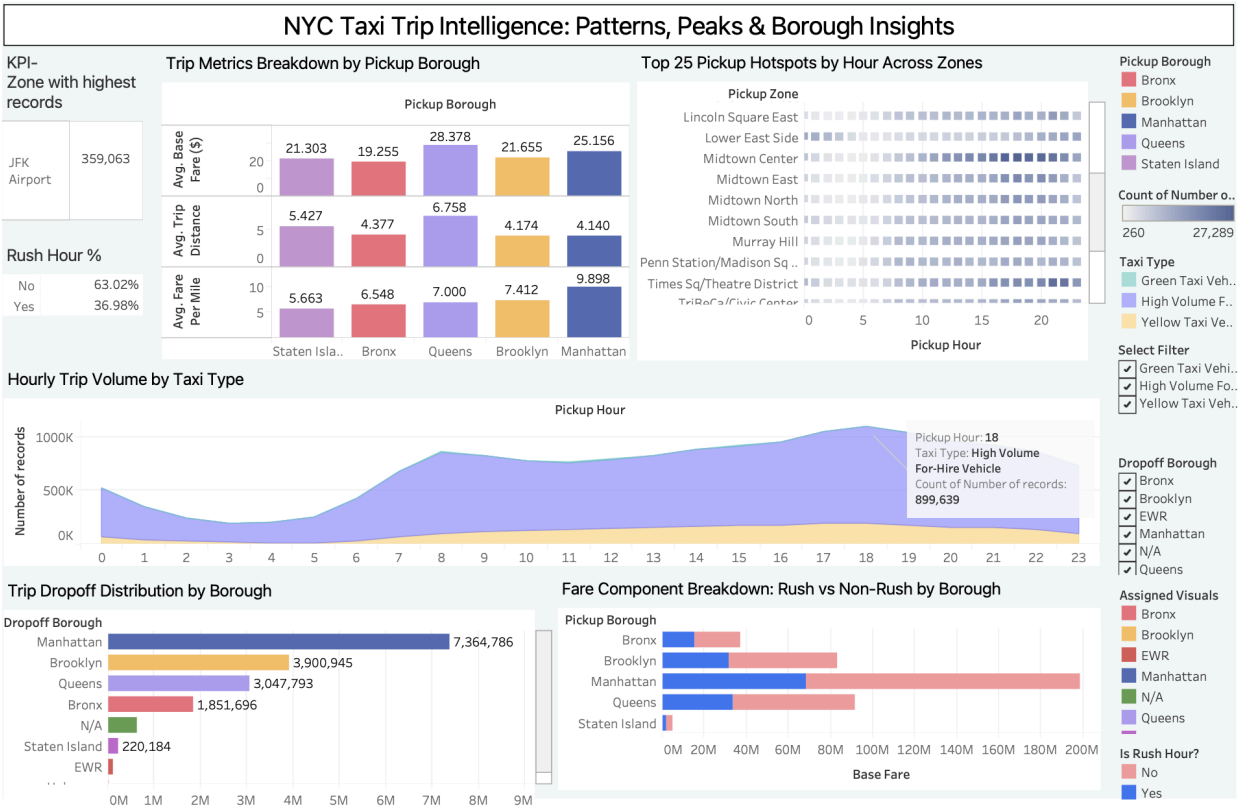
- The **Daily Trip Count by Taxi Type** line chart reveals **predictable weekday cycles**, with **high-volume FHVs dominating daily traffic** and slight dips during weekends.

- The **Hourly Heatmap** highlights **strong pickup concentrations between 8 AM and 6 PM**, with visible tapering after 9 PM, aligning with commuter patterns.

- The **Weekday Treemap** identifies **Friday and Saturday as peak days** for trip volume, while **Monday consistently underperforms**, reflecting reduced early-week demand.

- The interactive filters enable granular analysis by **rush hour**, **borough**, and **service zone**, giving users tools to isolate trends and anomalies in different urban subcontexts.

Spatial Insights Dashboard (Tableau):



Tools:

Tableau Desktop with full CSV data exported from *Tableau Prep Builder*

Overview:

The Tableau dashboard offers a comprehensive and interactive intelligence suite built on fully preprocessed CSV data. It contains eight coordinated visualizations that explore trip metrics, temporal trends, geographic pickup behavior, and rush hour effects. Designed for storytelling and dynamic exploration, it integrates filters for taxi type and borough, KPI tiles, and segmented visual panels.

Key Insights:

- **Manhattan leads in trip revenue per mile ($9.89)** while **Queens and Staten Island have longer average trip distances**, showing spatial travel variances across boroughs.

- **Heatmap analysis of pickups by hour** shows peak demand from 8 AM to 6 PM, with high-volume for-hire vehicles dominating afternoon and evening trips.

- **Top 25 pickup hotspots**, including JFK Airport and Midtown zones, show clustering of high-volume activity in specific city sectors.

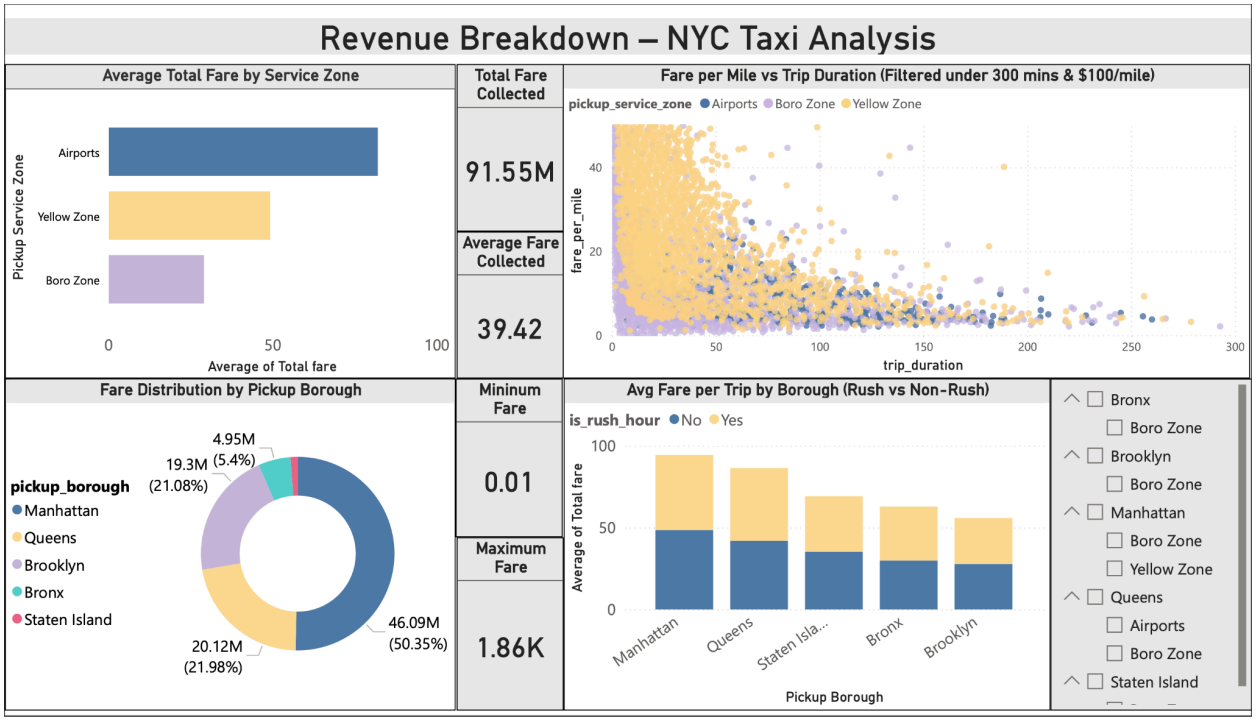
- **Fare components (tips, tolls, base fare)** increase significantly during **rush hours**, particularly in Manhattan and Queens, emphasizing time-based pricing effects.

- **Rush hour trips account for 37%** of all rides, suggesting that more than one-third of taxi demand occurs during peak traffic conditions.

- **Drop-off frequencies mirror pickup trends**, with Manhattan also being the leading drop-off borough.

- **KPI cards reveal** that JFK Airport is the **zone with the highest pickups** across all taxi types.

Revenue Breakdown Dashboard (PowerBI):



Tools:

Power BI via Microsoft Fabric

Overview:

The **Revenue Breakdown Dashboard** offers a **concise, executive-oriented snapshot** of New York City taxi revenues, built using Power BI connected to data staged through Microsoft Fabric. The dashboard is designed for financial insights and aggregates key fare metrics across **pickup boroughs** and **service zones**. With clearly defined KPI tiles and five core visualizations - bar chart, donut chart, scatter plot, stacked bar chart, and numerical indicators - it prioritizes **clarity and comparability** in revenue trends.

Key Insights:

- **Airports lead** in **average total fare**, driven by longer trip distances and additional toll charges.

- **Fare per mile** shows **significant volatility** at lower trip durations, particularly in the **Boro Zone**, which may reflect inconsistent pricing for shorter, local rides.

- **Manhattan dominates fare contribution**, accounting for over **50% of the total revenue** - a reflection of its central economic role in the city.

- **Rush hour fares are consistently higher**, regardless of borough, suggesting congestion pricing or elevated demand influences.

- **Extreme fare values**, such as the **maximum fare of $1.86K**, underscore potential edge cases, such as airport transfers, traffic-heavy premium rides, or outliers.

Challenges & Solutions

Challenge	Solution
Parquet data is too large for local processing	Used Google Colab with DASK to convert Parquet to chunked CSVs
Microsoft Fabric failed on the full 2024 dataset	Scoped down to February 2025 with smaller dataflows
The dataflow schema was published, but the rows weren't	Enabled data staging explicitly within Fabric
Semantic models in Power BI displayed only sample data	Switched to using flattened CSVs from Tableau Prep
Tableau Desktop is unable to read raw Parquet	Moved transformation layer to CSV-compatible tools
Schema inconsistencies across taxi types	Normalized fields during ETL in Tableau Prep

This iterative approach to tooling and modeling not only helped us succeed technically - it also deepened our practical understanding of data warehouse architecture, ETL resilience, and the critical importance of flexibility in BI workflows.

Learnings & Reflections

This project served as a **practical learning experience** in managing large-scale data, choosing appropriate tools, and adapting to unexpected challenges in a business intelligence environment.

Technical Takeaways

- **Tool Selection Matters:** We learned that not all enterprise tools are suited for high-volume public datasets. While Microsoft Fabric and Power BI offered powerful features, they had clear limitations in handling row volumes and joining large datasets.

- **Data Format Handling:** Working with **Parquet files** introduced us to the challenges of large, columnar data and taught us the importance of understanding efficient data storage formats and schema consistency.

- **ETL Strategy:** Experimenting with **DASK for parallelized transformation** and **Tableau Prep for cleaning and joining** enabled us to build a streamlined and scalable ETL pipeline.

- **Flat vs. Semantic Models:** While **semantic models** are ideal in theory, they proved less reliable under scale and time constraints. A **flat-table structure** in Tableau delivered superior accuracy, speed, and control.

- **Dashboard Design & Logic:** We gained hands-on experience building **calculated fields**, **custom measures**, and **interactive dashboards** in both **Power BI** and **Tableau** - each with its own learning curve and value proposition.

New Technical Skills Gained

- **First-time exposure to .parquet files**, which helped us understand their advantages in **columnar storage and large-scale data retrieval**.

- **Introduced to Microsoft Fabric**, where we learned how to configure data lakes, manage schema staging, and build Dataflow Gen2 pipelines.

- **Compared BI platforms** - Power BI vs. Tableau - for **flexibility, usability, and analytical depth**.

- **Created responsive dashboards** that blended business insights with narrative clarity using filters, slicers, and visual patterns.

- The project culminated in a **portfolio-ready analytics solution** that bridged data engineering and visualization.

Project Management Insights

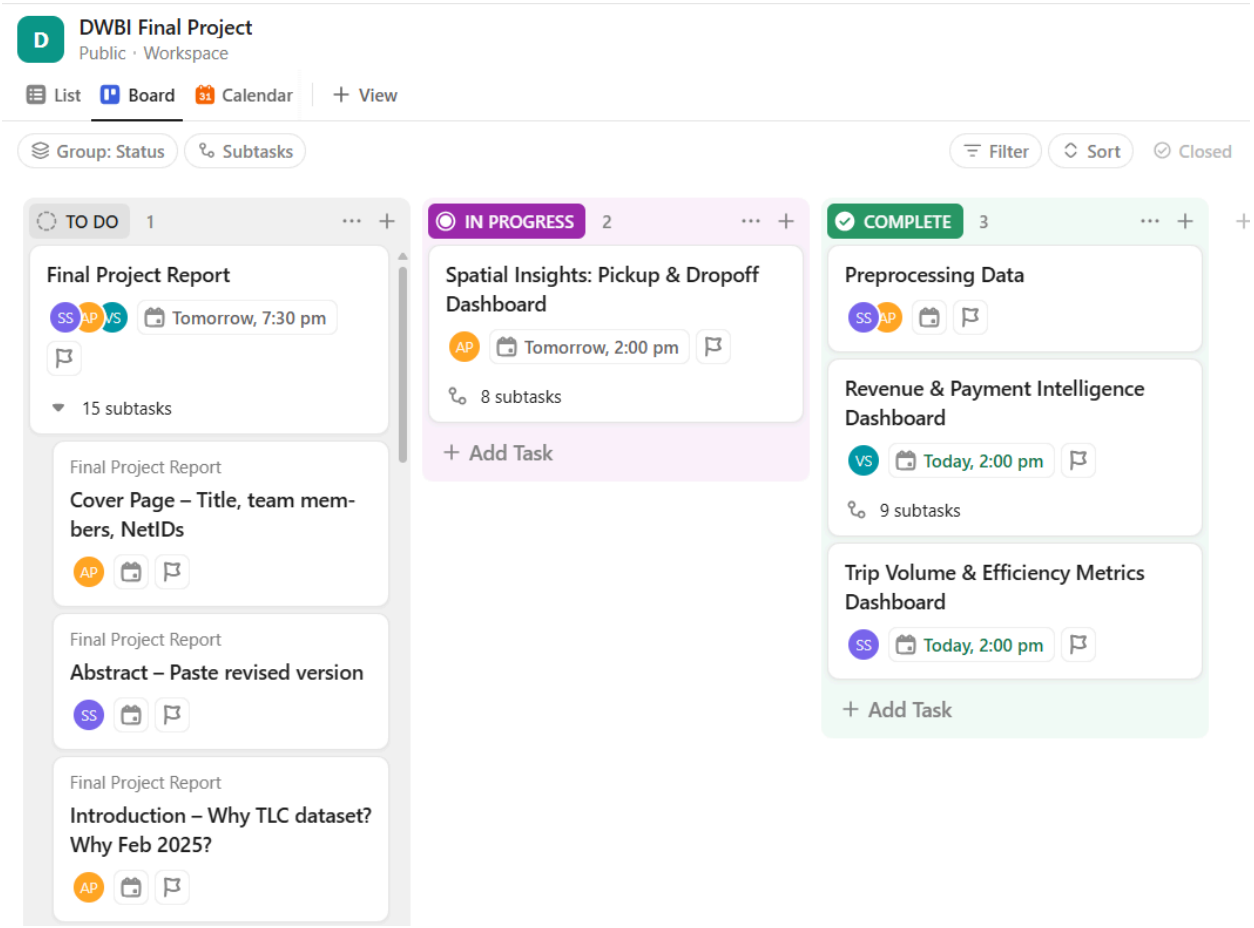


Fig. ClickUp's Task management board with assignments and deadlines

- **Iterative Development:** Our architecture evolved significantly across phases. Initial failures led us to modular experimentation and rapid pivots, allowing the team to

remain agile.

- **Collaboration Tools:** Tools like **ClickUp and Google Drive** allowed us to coordinate ETL handoffs, dashboard assignments, and documentation asynchronously.

Overall, the project taught us how to balance ambition with pragmatism - designing around limitations while still delivering a robust business intelligence solution under academic and time constraints.

Overview of Deliverables

Our final solution includes **one Power BI dashboard** built using Microsoft Fabric and semantic models, along with **two Tableau Desktop dashboards** developed using preprocessed CSV data output from Tableau Prep. Each dashboard was designed around a distinct analytical theme - revenue insights, operational efficiency, or spatial behavior and reflects deliberate choices in tool usage based on stakeholder needs and platform capabilities. This modular structure enabled tailored visual storytelling while showcasing diverse data visualization techniques.
