Problem Framing and Dataset Analysis

Urban Mobility Data Explorer

1. Dataset Description and Context

This project uses three official NYC Taxi & Limousine Commission (TLC) datasets that form a relational data model (fact + dimension tables) for analyzing urban taxi mobility patterns in New York City.

1.1 yellow_tripdata Fact Table

The table containing 7.67 million raw trip-level records for January 2019. Each row represents a single yellow taxi trip with 18 fields:
This table acts as the fact table in a star schema—each trip references dimension tables via PULocationID, DOLocationID, Vendor_Id, Ratecode_Id, and payment_type

Field Category	Key Fields
Temporal	tpep_pickup_datetime, tpep_dropoff_datetime
Spatial	PULocationID (PickUp Zone), DOLocationID (Dropoff zone)
Trip Details	trip_distance, passenger_count, RatecodeID, store_and_fwd_flag
Fare Breakdown	fare_amount,extramta_tax, tip_amount, tolls_amount, improvement_surcharge, congestion_surcharges, total_amount
Metadata	Vendor_ID, Payment_type

This table acts as the fact table in a star schema—each trip references dimension tables via PULocationID, DOLocationID, Vendor_Id, Ratecode_Id, and payment_type

1.2 taxi_zone_lookup Dimension Table

Field	Description	Examples
locationID	Unique zone ID (1–265)	132
Borough	NYC borough name	Manhattan
Zone	Neighborhood/area name	JFK Airport
Service_zone	Taxi service classification	Yellow Zone, Boro Zone, EWR

Notable entries include LocationID 264 ("Unknown" / "N/A") for unresolved zones and LocationID 265 ("Outside of NYC") for trips beyond city boundaries — both critical for data quality assessment

1.3. taxi_zones Spatial Metadata

An ESRI Shapefile containing 265 polygon geometries defining the geographic boundaries of each taxi zone. Uses the NAD83 / New York Long Island (EPSG:2263 coordinate system . Attributes include : OBJECTID, shape_leng zone , locationID, and horough. These polygons enable spatial visualization area calcultaions and geographic analysis of trip patterns

1.3.1 Relational Model

The three datasets integrate as follows:

yellow_tripdata.PULocationID ──→ taxi_zone_lookup.LocationID ──→ taxi_zones.LocationID yellow_tripdata.DOLocationID ──→ taxi_zone_lookup.LocationID ──→ taxi_zones.LocationID

This star schema design separates high-volume transactional data (7.67M trips) from low-volume descriptive data (265 zones), enabling efficient joins, geographic enrichment, and spatial visualization of trip patterns.

2. Data Challenges

2.1Missing and Null Fields (yellow_tripdata)

The trip data contains several fields with missing or zero values that compromise analysis:

- Passenger_count: Approximately 2-3% of trips report 0 passengers — likely meter errors or data entry issues, not empty vehicles.
- Congestion_subcharge: Many records show blank/null values, as this surcharge was only introduced on January 1, 2019, and phased in during the month.
- Trip_distance: Thousands of trips report zero distance but non-zero fares, suggesting cancelled trips, meter resets, or short movements within a single zone
- Store and fwd flag: Some null entries indicate connectivity gaps where the vehicle's data transmission status was not recorded

2.2 Anomalies (yellow_tripdata)

The January 2019 file contains trips with pickup dates from November and December 2018 (eg: 2018-12-21, 2018-11-28) These are "store-and-forward" records trips recorded by the meter offline and transmitted late. This creates a data leakage issue where the file does not strictly represent January 2019 activity.

2.3 Outliers in Fare and Distance

Extreme values exist across multiple fields:

- Trip_distance: Ranges from 0 to hundreds of miles — some physically impossible within NYC
- Total_amount : Negative values (refunds/adjustments) and extreme positives (>$500) for short trips A small
- Fare_amount: number of negative fares indicate reversed or disputed charges
- Passenger_count: Values of 0 and 7+ are questionable (standard taxi seats 4, max legal is 5-6)

2.4 Orphan Location IDs

Some trips reference PULocationID = 264 (unknown) or refere PULocationID = 265 (Outside of NYC). Meaning the GPS-to Zone mapping failed or the triporiginated/ended beyond city limits these records cannot be geographically visualized and must be handled separately.

1.1 Spatial Data: Zone Size Dis parity

The taxi zones shapefile reveals a 30,000× size variation between the smallest zone (Little Italy, Manhattan: ~0.0015 sq km) and the largest (Newark Airport: ~45 sq km). This extreme range distorts choropleth maps and makes raw trip counts misleading without area normalization.

1.2 Temporal Limitation of Spatial Data

The shapefile was published in September 2015, while the trip data is from January 2019. Although TLC maintains stable zone definitions for data continuity, post- 2015 developments (e.g., Hudson Yards) may not be fully reflected in zone names or boundaries.

3.Data Cleaning Assumptions

3.1Temporal Filtering

Assumption: Only trips with pickup dates in January 2019 belong in this analysis.

Action: Filtered out records with tpep pickup datetime outside the 2019-01-01 to 2029-01-31 range. Store and forward records from 2018 were excluded to maintain temporary consistency.

3.2. Zero-Distance and Zero-Fare Trips

Assumption: trips with trip_distance = 0 AND fare_amount = 0 are cancelled or erroneous
Action: Removed these records.However, trips with distance = 0 but a valid fare were retained, as they may represent short in-zone movement or flat rate trips

3.3. Passenger Count Imputation

Assumption: Trips with passenger_count = 0 are meter errors, not truly empty vehicles.
Action: Replaced 0 values with the mode (1 passenger), since single-rider trips dominate (~70% of all trips). Counts above 6 were capped at 6 (legal taxi maximum).

3.4 Outlier Boundaries for Fares and Distance

Assumption: Extreme fare/distance values beyond physically plausible ranges are data errors. Action: Applied domain-based thresholds:

Removed trips with: trip_distance > 100miles (NYC is 35miles long)
Removed trips with: total_amount < 0 tracked separately)
Removed trips with fare_amount < $2.50 (below based fare excluding zero-fare airport//negotiated trips)

3.5 Unknown and Out-of-NYC Zones

Assumption: LocationIDs 264 ("Unknown") and 265 ("Outside of NYC") represent unresolvable trips.
Action: Flagged and excluded from geographic analysis and choropleth maps. Retained in aggregate statistics (total trip counts, revenue summaries) since fares are still valid

3.6 Spatial Data Integrity

Assumption: All 265 zone geometries are valid and represent complete NYC coverage.
Verified: Zero missing values, zero invalid geometries, zero duplicates across all three datasets. The lookup table's 265 entries match the shapefile's 265 polygons exactly. No geometry repair (.buffer(0)) was needed.

3.7 Congestion Surcharge

Assumption: Null values in congestion_surcharge represent trips before the surcharge took effect (or exempt trips), not missing data.
Action: Filled nulls with $0.00 rather than dropping records, preserving the full trip dataset for non-surcharge analyses.

4. Unexpected Observation That Influenced Design

The Observation: Cross-Date Contamination in the Fact Table

The most unexpected finding was discovering that the yellow_tripdata_2029-01 file contains trips with pickup timestamps from November and December 2018 — not just January 2019. Specific examples from the raw data include pickups on legitimate 2018-12-21 13:48:30 and 2018-11-28 15:52:25 appaering alongside January 2019 records

Why This Was Unexpected:

A file named yellow_tripdata_2029-01 is reasonably assumed to contain only January 2019 data. Finding records from two months prior raised immediate concerns about data pipeline integrity, potential duplication with the November/December 2018 files, and whether the temporal scope of the analysis was compromised.

Root Cause Identified:
These are store-and-forward records — trips where the taximeter lost cellular connectivity and cached the data locally. When the vehicle later regained connectivity,
These state records were transmitted and ingested into the January file based on *upload date*, not *trip date*. The behavior.
Store_and_fwd_flag = ‘Y’ field confirms this

How This Influenced Design

This observation had four direct impacts on the system design:

This observation had four direct impacts on the system design:

1. Mandatory Temporal Validation Layer
Added an explicit date-range filter as the first step in the data pipeline before any joins, aggregations, or visualizations. Every query now enforces WHERE pickup_date BETWEEN ‘2019-01-01’ AND ‘2019-01-31’ to prevent temporal contamination. This was not originally planned, as the filename was assumed to be a sufficient scope guarantee.

2. Data Lineage Documentation
Implemented a data lineage tracker that records the number of records filtered at each cleaning stage. For transparency, the report now includes: "X records removed due to out-of-range pickup dates (store-and-forward lag)." This ensures reviewers understand why the final row count differs from the raw file.

3. Store-and-Forward Analysis Module
Created a dedicated analysis track examining store-and-forward trips separately. These trips reveal areas with poor cellular coverage (potential infrastructure gaps) and introduce systematic temporal bias — both valuable insights for urban mobility analysis that would have been missed if we simply discarded them.

4. Defensive Design Philosophy
This finding established a broader principle: never trust file-level metadata as a data quality guarantee . Every field is now validated independently against domain-specific rules (date ranges, fare bounds, geographic limits). This defensive approach caught additional anomalies (negative fares, impossible distances) that might otherwise have propagated silently through the analysis pipeline.

Summary

Dataset	Records	Key Challenge	Resolution
yellow_tripdata	7.67M trips	Cross-date contamination, outliers, nulls	Temporal filtering, domain-based thresholds, imputation
taxi_zone_lookup	265 zones	Special entries (Unknown, Outside NYC)	Flagged and excluded from spatial analysis
taxi_zones (shapefile)	265 polygons	30,000× zone size variation	Density normalization, stratified analysis

Data Quality: The dimension tables (lookup + shapefile) are production-quality with zero defects. The fact table requires significant cleaning but contains rich analytical value once properly validated. The unexpected store-and-forward contamination finding shaped the project's defensive, validation-first architecture