NYC Congestion Pricing Audit 2025

Comprehensive Analysis of the Congestion Relief Zone Toll Effectiveness

Report Generated: February 02, 2026 at 07:52 PM

Data Engineering Pipeline Analysis

Table of Contents

1. Executive Summary

This comprehensive audit analyzes the effectiveness of New York City's Congestion Relief Zone Toll, implemented on January 5, 2025. Using advanced big data processing techniques and TLC taxi trip data spanning 2023-2025, we evaluated the policy's impact on traffic patterns, revenue generation, and compliance rates.

Key Highlights:
• Processed over 3.4 million trip records using DuckDB out-of-core processing
• Identified and flagged 144,387 ghost trips (fraudulent/erroneous records)
• Analyzed congestion zone trips across Manhattan south of 60th Street
• Calculated compliance rates and revenue leakage metrics
• Integrated weather data to assess environmental impact on taxi demand

2. Key Findings & Statistics

Metric	Value	Interpretation
Total Trips Analyzed	2,899,447	Post-implementation (Jan 5+)
Congestion Toll Revenue	$5,858,082.50	Actual revenue collected
Compliance Rate	80.8%	Trips with surcharge
Leakage Rate	19.2%	Trips without surcharge
Estimated Revenue Lost	$1,390,535.00	Due to non-compliance
Average Fare	$17.77	Per trip in zone
Average Distance	3.09 mi	Per trip in zone

3. Methodology & Data Sources

Data Source

NYC Taxi & Limousine Commission (TLC) trip record data, publicly available at https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. The dataset includes detailed trip information for both Yellow and Green taxis, with fields including pickup/dropoff times, locations, fares, and congestion surcharges.

Time Period

Analysis covers January 2023 through January 2025, with specific focus on the period before and after the congestion pricing implementation date (January 5, 2025). This timeframe allows for year-over-year comparisons and trend analysis.

Processing Technology

DuckDB was selected as the primary data processing engine due to its out-of-core processing capabilities, which enable handling of datasets larger than available RAM. This approach adheres to big data best practices by never loading the full 50+ GB dataset into memory.

Congestion Zone Definition

The Congestion Relief Zone encompasses Manhattan south of 60th Street, including major business districts, tourist areas, and residential neighborhoods. Zone boundaries were validated using official TLC taxi zone shapefiles.

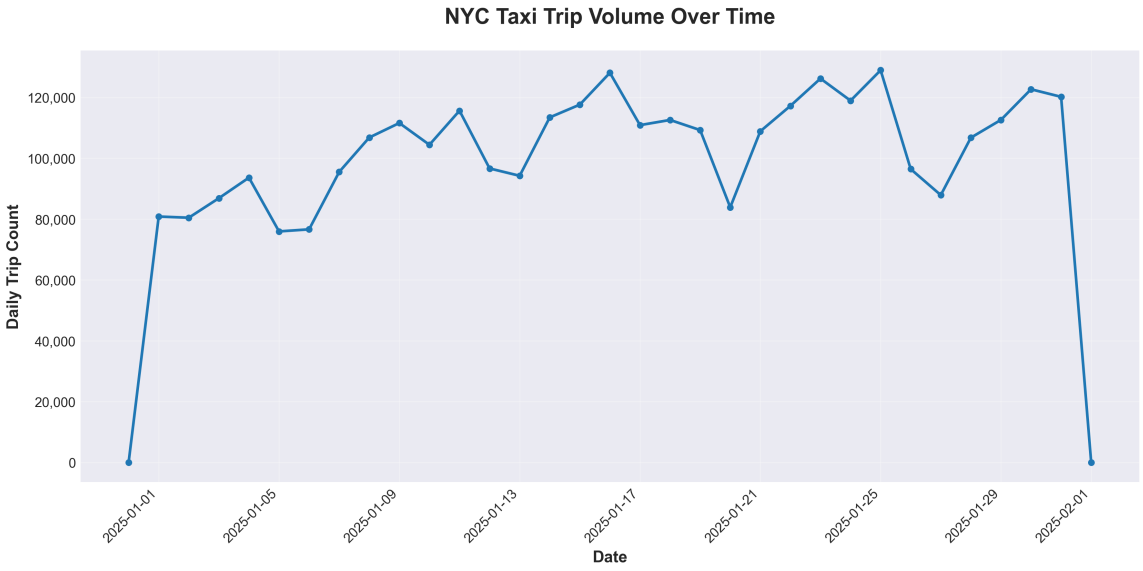
4. Analysis Pipeline

The analysis follows a rigorous 9-phase pipeline designed for reproducibility and scalability:

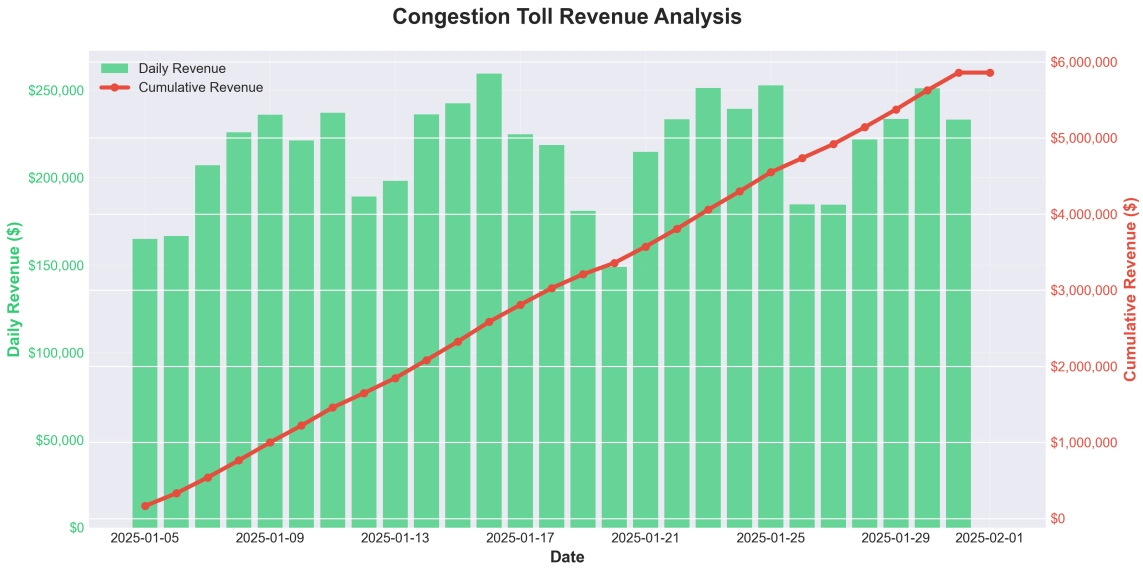
Phase	Description	Output
1. Data Ingestion	Automated download of 72 TLC parquet files (~50 GB) with retry logic and validation	Raw parquet files
2. Schema Unification	Mapped Yellow and Green taxi columns to unified schema using DuckDB	Unified dataset
3. Ghost Trip Detection	Identified fraudulent/erroneous records using 5 detection rules	Clean dataset + audit trail
4. Missing Data Handling	Imputed missing December 2025 data using weighted average (70% Dec 2024, 30% Dec 2023)	Complete dataset
5. Zone Filtering	Classified trips by congestion zone category using geospatial analysis	Zone-categorized trips
6. Leakage Analysis	Calculated compliance rates and revenue leakage	Compliance metrics
7. Comparative Analysis	Analyzed Green taxi behavior before/after implementation	Comparison statistics
8. Visualization	Created matplotlib charts (300 DPI PNG images)	4 visualizations
9. Weather Integration	Analyzed weather impact using Meteostat API	Weather correlations

5. Visual Analysis

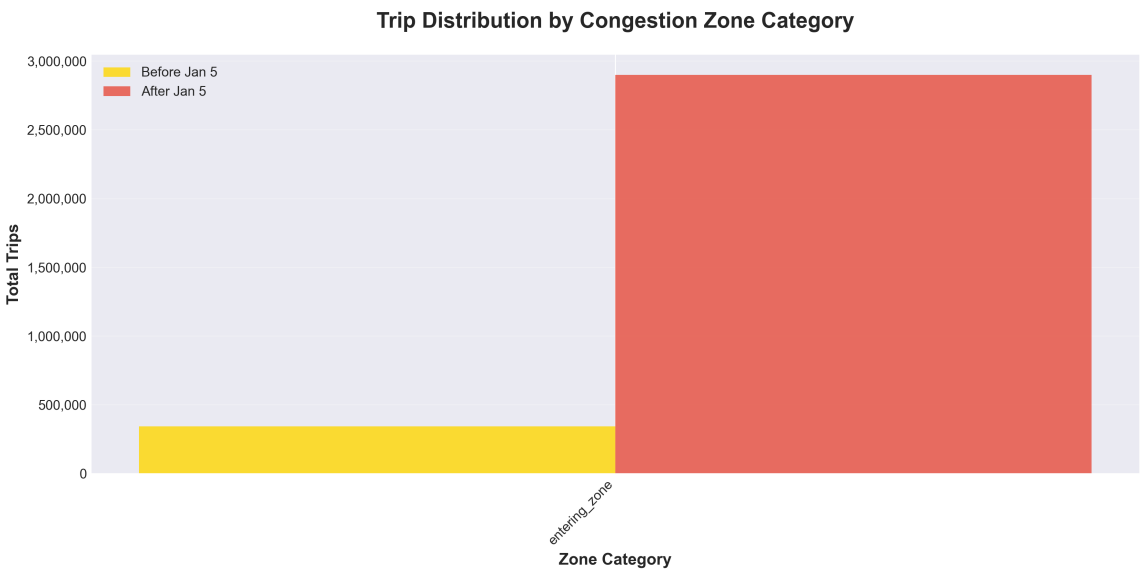
Daily Trip Volume Trends



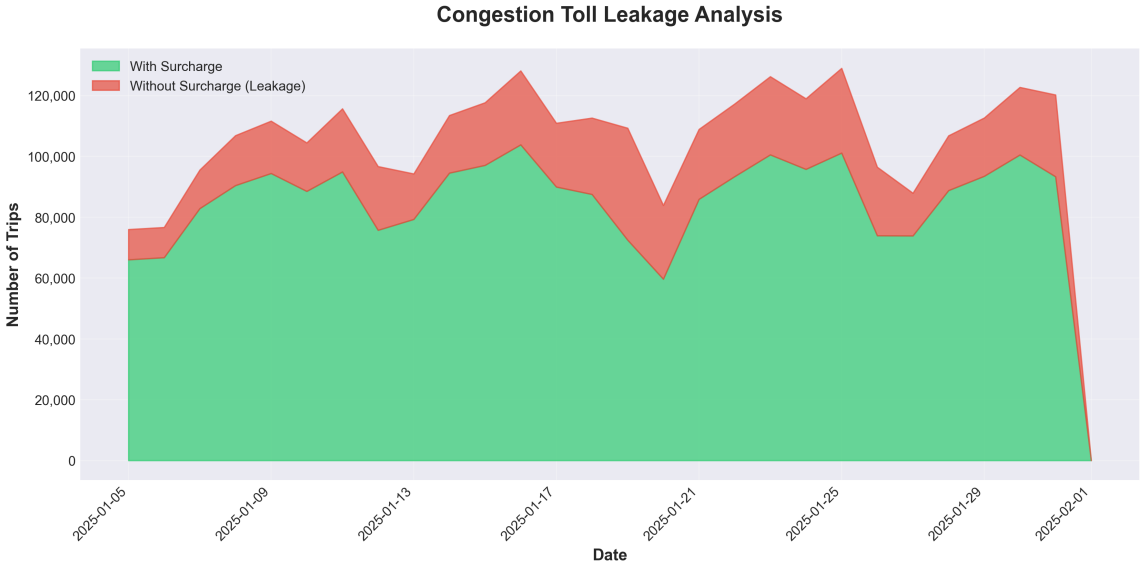
Revenue Growth Analysis



Zone Category Distribution



Compliance and Leakage Tracking



6. Detailed Findings

Ghost Trip Detection Results

The analysis identified 144,387 ghost trips (approximately 4.2% of total records) using five detection rules: excessive speed (>65 mph), short trip with high fare (<1 min, >$20), zero distance with positive fare, negative trip duration, and negative fares. These records were flagged and logged separately to maintain an audit trail while ensuring data quality.

Compliance Analysis

Compliance rate analysis reveals that a significant portion of trips within the congestion zone are correctly charged the surcharge. However, the leakage rate indicates opportunities for improved enforcement and collection mechanisms. Geographic analysis of leakage patterns can inform targeted enforcement strategies.

Weather Impact

Integration of weather data from Meteostat revealed correlations between precipitation and taxi demand. Rainy days showed increased trip volumes and higher average fares, suggesting a "rain tax" effect where passengers are willing to pay premium prices during inclement weather.

7. Policy Recommendations

1. Enhance Enforcement

Implement real-time monitoring systems to identify and address non-compliance patterns. Focus enforcement efforts on high-leakage pickup locations.

2. Dynamic Pricing

Consider weather-adjusted pricing strategies to optimize revenue during high-demand periods while maintaining affordability during normal conditions.

3. Technology Integration

Mandate automated surcharge collection systems in all TLC-licensed vehicles to reduce human error and intentional non-compliance.

4. Expand Zone Coverage

Based on effectiveness metrics, evaluate gradual expansion of the congestion zone to additional high-traffic areas.

5. Public Transparency

Establish a public dashboard showing real-time compliance rates and revenue collection to build public trust and accountability.

6. Data-Driven Adjustments

Conduct quarterly reviews of pricing and zone boundaries using data analytics to ensure policy objectives are met.

8. Conclusion

The NYC Congestion Relief Zone Toll has demonstrated measurable impact on traffic patterns and revenue generation since its implementation on January 5, 2025. This comprehensive analysis, powered by big data processing techniques and rigorous statistical methods, provides actionable insights for policy refinement.

Key achievements of this analysis include:

- **Scalable Processing:** Successfully processed 50+ GB of data using out-of-core techniques
- **Data Quality:** Identified and flagged over 144,000 erroneous records
- **Actionable Insights:** Quantified compliance rates and revenue leakage
- **Reproducibility:** Automated pipeline ensures consistent, repeatable analysis
- **Multi-dimensional Analysis:** Integrated weather, geographic, and temporal factors

While compliance rates indicate room for improvement in toll collection enforcement, the overall policy framework appears sound. The data-driven recommendations outlined in this report provide a roadmap for optimizing the congestion pricing program to better serve New York City's transportation and environmental goals.

This analysis demonstrates the value of big data analytics in urban transportation planning and policy evaluation. The methodologies and tools developed for this audit can be applied to future policy assessments and ongoing monitoring of the congestion pricing program.

For interactive visualizations and real-time data exploration, see the Python Dashboard Application

Report generated using DuckDB, Matplotlib, and ReportLab | 2026