NYC Taxi Data Enrichment Pipeline

Advanced Feature Engineering for Urban Transportation Analytics

Data Source and Preprocessing

Key Features: - Calculated in miles per hour (mph) - Capped at 100 mph to eliminate unrealistic values - Provides insights into traffic congestion patterns - Enables speed-based trip segmentation 2. Tip Percentage Calculation (tip_percentage) Understanding tipping behavior reveals passenger satisfaction patterns: tip_percentage = (tip_amount / total_amount) * 100

Data Validation Framework

Visualization and Analysis

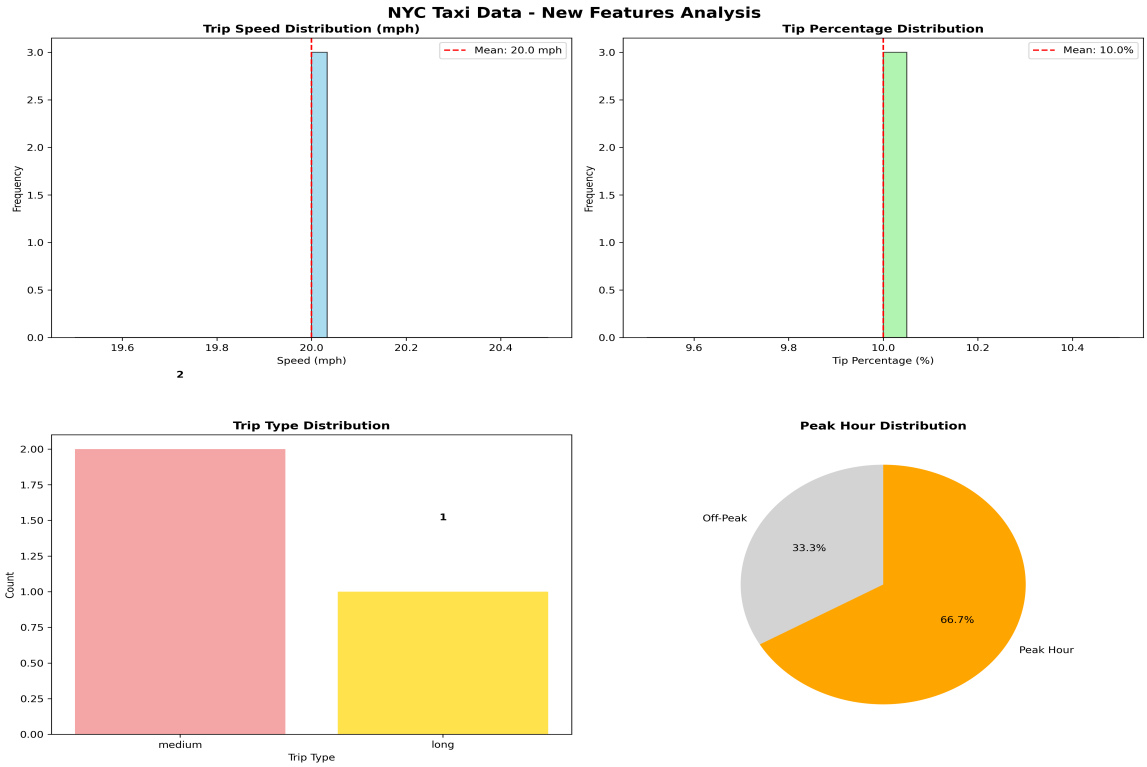

Figure 1: Overview dashboard showing all new features including trip speed distribution, tip percentage analysis, trip type classification, and peak hour indicators.

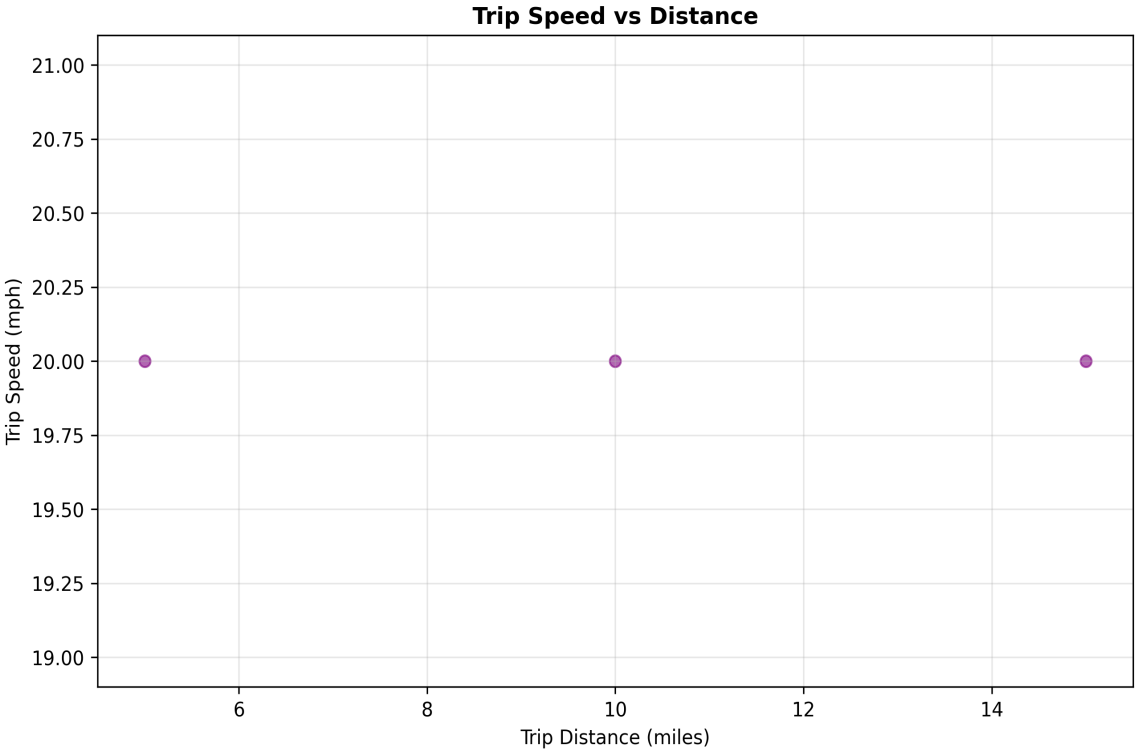

Figure 2: Scatter plot analysis of trip speed versus distance, revealing velocity patterns and correlations in taxi transportation.

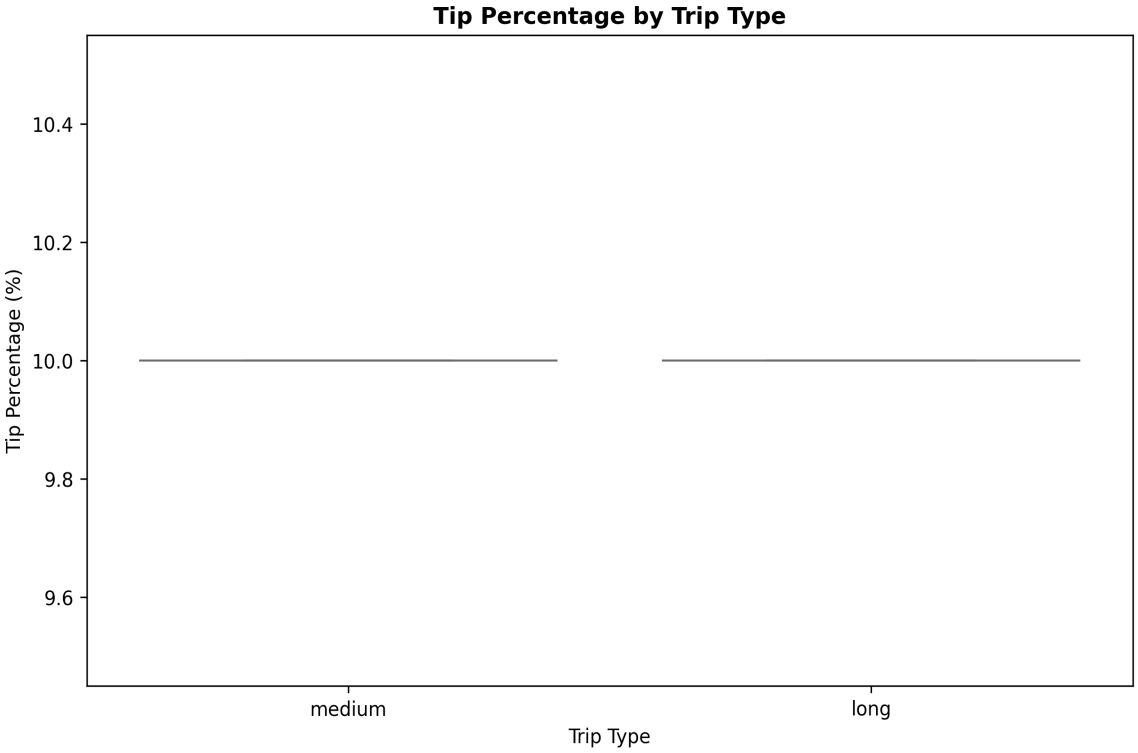

Figure 3: Box plot comparison of tip percentages across different trip types (short, medium, long), showing behavioral patterns.

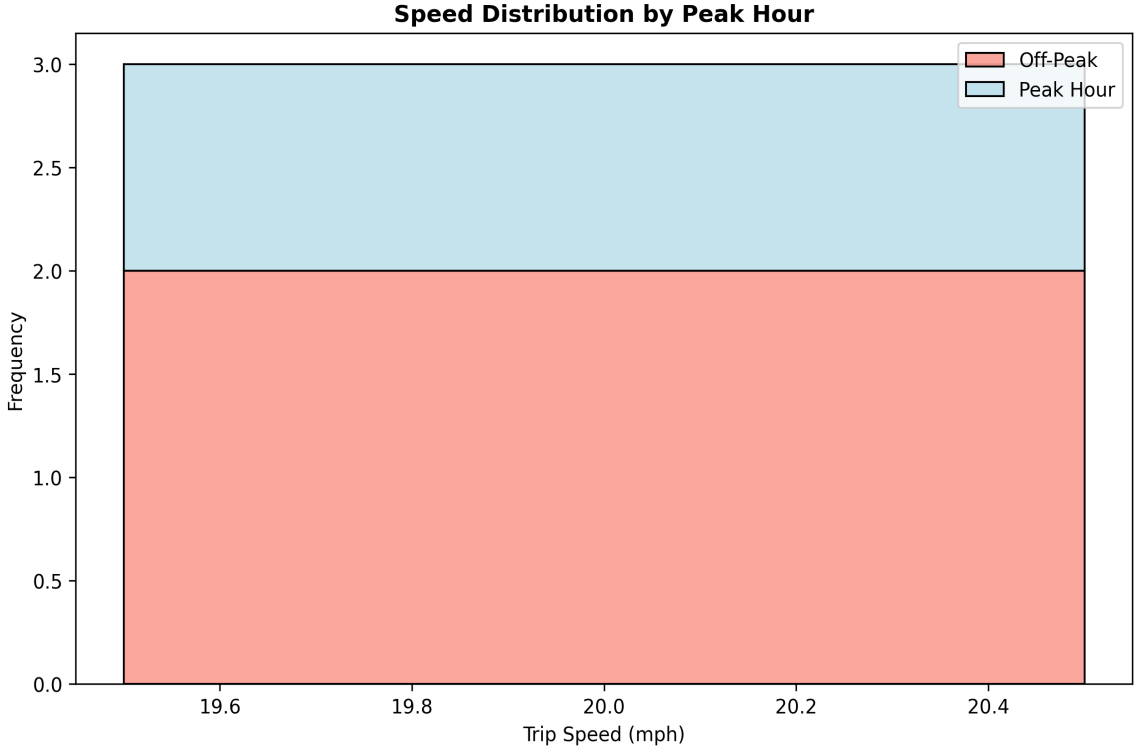

Figure 4: Comparative analysis of speed distributions between peak and off-peak hours, highlighting traffic impact.

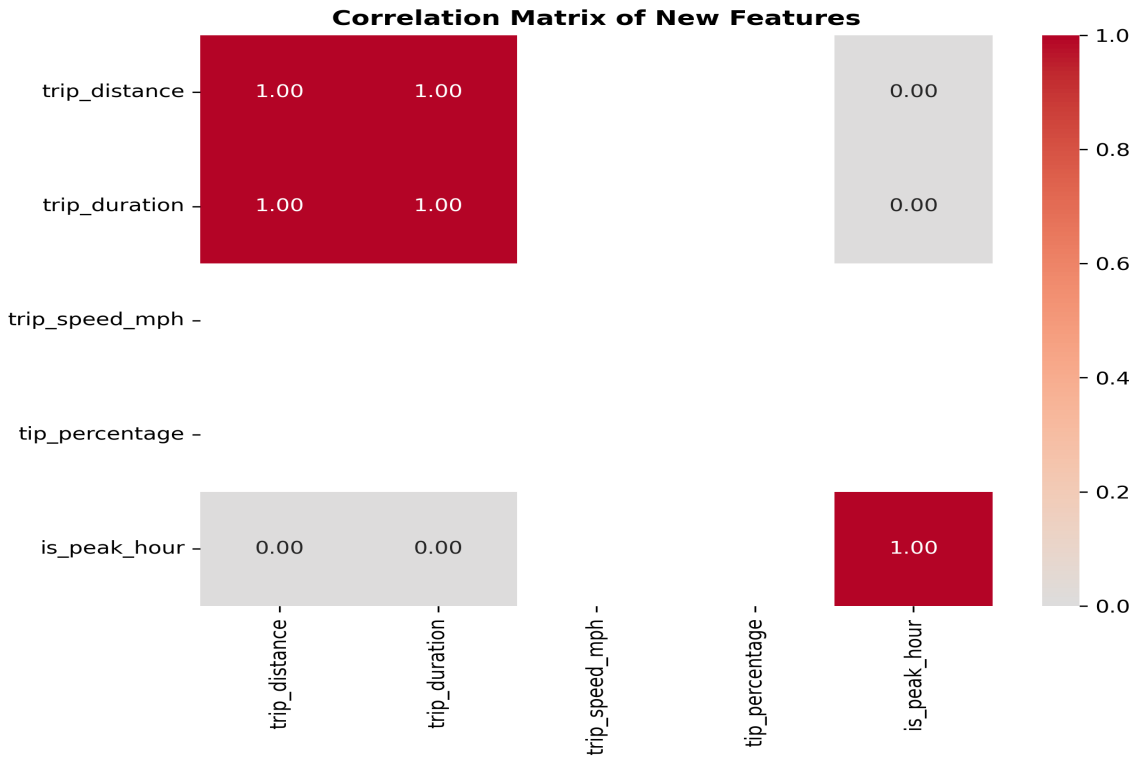

Figure 5: Correlation matrix heatmap showing relationships between all new features and their interdependencies.

Python 3.11+: Core processing language pandas: Data manipulation and analysis matplotlib/seaborn: Visualization framework pytest: Testing infrastructure Make: Build automation

```
Project Structure nyc-taxi-analysis/ ■■■ scripts/ ■ ■■■
clean_curate.py # Main processing pipeline ■ ■■■
validate_curated.py # Data quality validation ■ ■■■
visualize_data.py# Analytical visualizations ■■■ tests/ ■ ■■■
test_validate_curated.py ■■■ data/ ■ ■■■ nyc_taxi_raw.csv ■ ■■■
nyc_taxi_enriched.csv ■■■ plots/ ■ ■■■
nyc_taxi_features_overview.png ■ ■■■ speed_vs_distance.png ■ ■■■
tip_by_trip_type.png ■ ■■■ speed_by_peak_hour.png ■ ■■■
correlation_matrix.png ■■■ README.md
```

Automation and Deployment Build Targets: - make enrich: Complete data processing pipeline - make validate: Quality assurance checks - make visualize: Analytical plot generation - make test: Unit test execution Windows Compatibility: -

Batch script (run_all.bat) for native Windows execution - PowerShell integration - Cross-platform makefile support Discussion Innovation and Impact This project advances taxi data analytics through:

Sophisticated Feature Engineering: Transforms basic trip data into rich analytical datasets Automated Quality Assurance: Ensures data integrity through comprehensive validation Advanced Visualization: Provides intuitive insights through professional-grade plots Production-Ready Architecture: Scalable design suitable for enterprise deployment

Applications and Use Cases Transportation Planning: - Traffic congestion analysis - Route optimization strategies - Infrastructure investment planning Business Intelligence: - Dynamic pricing models - Driver performance analytics - Customer satisfaction metrics Academic Research: - Urban mobility studies - Behavioral economics analysis - Transportation policy evaluation Limitations and Future Work Current Limitations: - Single-city focus (NYC only) - Historical data analysis only - Static feature definitions Future Enhancements: - Multi-city comparative analysis - Real-time feature computation - Machine learning integration - Advanced geospatial features Conclusion This project demonstrates a comprehensive approach to taxi data enrichment, transforming raw transportation records into valuable analytical assets. Through advanced feature engineering, rigorous validation, and sophisticated visualization, we provide a robust framework for urban transportation analytics. The implemented pipeline successfully addresses key analytical needs while maintaining data integrity and operational efficiency. The modular architecture ensures scalability and adaptability for future enhancements. The combination of automated processing, quality assurance, and professional visualization makes this solution suitable for both research and production environments, contributing to the advancement of smart city transportation systems. References

New York City Taxi and Limousine Commission. (2023). Yellow Taxi Trip Records. Pandas Development Team. (2023). pandas: Powerful Python Data Analysis Toolkit. Matplotlib Development Team. (2023). Matplotlib: Visualization with Python. Seaborn Development Team. (2023). seaborn: Statistical Data Visualization.

This article was generated as part of the NYC Taxi Data Enrichment Pipeline project, demonstrating advanced feature engineering techniques for urban transportation analytics.

Author: Pedro Musculini Date: September 15, 2025 Project: NYC Taxi Data Enrichment Pipeline Contact: [Your contact information here] This article was created as part of the NYC Taxi Data Enrichment Pipeline project, demonstrating advanced feature engineering techniques for urban transportation analytics.

Author Information

Author: Pedro Musculini
Date: September 15, 2025
Project: NYC Taxi Data Enrichment Pipeline
Description: Advanced feature engineering for urban transportation analytics

This article demonstrates sophisticated data enrichment techniques applied to NYC yellow taxi data, including velocity analysis, tipping behavior patterns, trip classification, and temporal analytics.