NYC Yellow Cab Taxi Data Analysis Project Proposal

Team Members: Jacob Kuriakose (ASU ID: 1233481797), Jui-Chi Lui (ASU ID: 1238358162), Subash Lakshminarayanan (ASU ID: 1221461384), Thanyathorn Limsuvattanaphong (ASU ID: 1235549421)

Dataset Context and Overview

The New York City Yellow Cab taxi dataset for January 2024 contains 2,964,624 individual taxi trips, representing one of the largest transportation datasets available for urban mobility analysis. This dataset captures NYC's transportation ecosystem during January 2024, providing insights into travel patterns, fare structures, and operational dynamics.

The dataset includes 19 variables covering temporal, geographic, financial, and operational dimensions. Key variables include pickup/dropoff timestamps, passenger counts, trip distances, fare amounts, tip amounts, payment methods, vendor information, and location identifiers (PULocationID, DOLocationID).

Initial Data Investigation

Scale and Coverage: Nearly 3 million trips with 84,704 trips per day average, peaking at 110,515 trips on January 27, 2024. Busiest hour is 6:00 PM (212,788 trips), quietest is 4:00 AM (16,742 trips).

Financial Structure: Fare amounts range from -$899 to $5,000 (mean: $18.18, median: $12.80). 37,448 trips have negative fares (data errors), 893 have zero fares. Tips provided in 76% of trips (mean: $3.34, median: $2.70). Total amount averages $26.80.

Operational Patterns: Three vendors operate the fleet - Vendor 2 (75.4%), Vendor 1 (24.6%). Payment methods: Type 1 (78.2%, likely credit card), Type 2 (14.8%, likely cash), Type 0 (4.7%, no charge). Rate Code 1 (standard fare) represents 89.8% of trips.

Geographic Distribution: Uses location IDs (1-265) with 260 unique pickup and 261 unique dropoff locations covering NYC metropolitan area.

Data Quality Issues: 4.73% missing values in passenger_count, RatecodeID, store_and_fwd_flag, congestion_surcharge, and Airport_fee. Data anomalies include negative trip durations and extreme fare amounts requiring preprocessing.

Research Objectives and Hypotheses

Primary Research Objective

To analyze the relationships between temporal patterns, trip characteristics, and fare determinants in NYC Yellow Cab operations during January 2024, focusing on demand patterns, pricing dynamics, and operational efficiency factors.

Secondary Objectives

1. Identify peak demand periods and their relationship to fare amounts and tip behavior
2. Analyzing trip distance/duration impact on fare structure and customer satisfaction (measured through tips)
3. Examine vendor performance and differences in payment methods and customer preferences
4. Investigate geographic patterns in trip demand and fare variations across NYC locations

Research Hypotheses

Hypothesis 1: Temporal Demand and Pricing

H_1: Peak hours (6–8 PM) will have a significantly higher average total fare compared to off-peak hours, controlling for trip distance and duration.

Hypothesis 2: Trip Distance and Fare Elasticity

H_2: There is a positive correlation between trip distance and fare amount, which may be non-linear due to base fare structures and congestion pricing.

Hypothesis 3: Customer Satisfaction and Tip Behavior

H_3: Longer trip durations (controlling for trip distance) are associated with lower tip percentages, suggesting possible customer dissatisfaction with traffic delays.

Hypothesis 4: Vendor Performance and Customer Preferences

H_4: Different vendors show significant differences in average tip percentages, indicating varying service quality levels.

Hypothesis 5: Payment Method and Customer Preferences

H_5: Payment methods significantly influence customer preferences across different time periods compared to other payment methods, suggesting that upfront pricing is suitable for use at any time of day.

Hypothesis 6: Geographic Fare Variation

H_6: Certain pickup and drop-off location combinations show significant differences in fare structures, reflecting variations in market demand and operational complexity across NYC neighborhoods.

Research Utilization / Expected Application

For vendors: Identify low-demand periods suitable for targeted promotions or discounts to increase ridership and revenue, improve service efficiency, and attract more customers.

For customers: Determine the best times to travel in order to minimize fares and avoid extra charges, such as congestion fees.

Project Management Plan

Jacob Kuriakose: Data preprocessing and exploratory data analysis
Jui-Chi Lui: Vendor performance analysis
Subash Lakshminarayanan: Geographic analysis and fare structure modeling
Thanyathorn Limsuvattanaphong: Statistical modeling and hypothesis testing
Timeline: 4 weeks - Data cleaning (Week 1), Analysis (Weeks 2-3), Report (Week 4)

Methodology and Evaluation Metrics

The analysis will employ multiple statistical techniques: descriptive statistics and data visualization, correlation analysis, regression analysis, ANOVA/t-test comparisons, chi-square tests, and geographic analysis using location-based clustering techniques.

Data preprocessing will focus on cleaning anomalies, handling missing values, and creating derived variables such as trip duration, fare per mile, and tip percentage. The analysis will control confounding variables and use appropriate statistical tests to validate hypotheses.

Evaluation Metrics: Model performance will be evaluated using R^2 for regression models, accuracy and precision for classification tasks, and p-values for hypothesis testing. Statistical significance will be tested at $\alpha = 0.05$ level with appropriate effect size calculations (Cohen's d for t-tests, η^2 for ANOVA). Key performance indicators include fare prediction accuracy (target: $R^2 > 0.7$), tip percentage prediction (target: MAE < 2%), and vendor performance differentiation (target: significant differences at $p < 0.05$).

This comprehensive analysis will provide valuable insights into NYC's transportation ecosystem and contribute to understanding urban mobility patterns, customer behavior, and operational efficiency in the taxi industry.