NYC Yellow Cab Taxi Data Analysis Project Proposal

**Team Members:** Jacob Kuriakose (ASU ID: 1233481797), [Member 2 Name (ASU ID)], [Member 3 Name (ASU ID)], [Member 4 Name (ASU ID)]

# 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).

**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.

# Research Objectives and Hypotheses

## Primary Research Objective

Analyze 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. Analyze trip distance/duration impact on fare structure and customer satisfaction (measured through tips)
3. Examine vendor performance differences and customer payment preferences
4. Investigate geographic patterns in trip demand and fare variations across NYC locations

## Research Hypotheses

**Hypothesis 1: Temporal Demand and Pricing** H1: Peak demand hours (6-8 PM) will show significantly higher average fares compared to off-peak hours, controlling for distance and duration.

**Hypothesis 2: Trip Distance and Fare Elasticity** H2: Positive correlation between trip distance and fare amount, with non-linear relationship due to base fare structures and congestion pricing.

**Hypothesis 3: Customer Satisfaction and Tip Behavior** H3: Longer trip durations (controlling for distance) will be associated with lower tip percentages, suggesting customer dissatisfaction with traffic delays.

**Hypothesis 4: Vendor Performance and Customer Preferences** H4: Different vendors will show significant differences in average tip percentages and customer payment method preferences, indicating varying service quality levels.

**Hypothesis 5: Geographic Fare Variation** H5: Certain pickup/dropoff location combinations will show significantly different fare structures, reflecting market demand and operational complexity variations across NYC neighborhoods.

# Project Management Plan

Jacob Kuriakose: Data preprocessing and exploratory data analysis

Member 2: Vendor performance analysis

Member 3: Geographic analysis and fare structure modeling

Member 4: 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, 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 for confounding variables and use appropriate statistical tests to validate hypotheses.

**Evaluation Metrics**: Model performance will be evaluated using R² for regression models, accuracy and precision for classification tasks, and p-values for hypothesis testing. Statistical significance will be tested at α = 0.05 level with appropriate effect size calculations (Cohen's d for t-tests, η² for ANOVA). Key performance indicators include fare prediction accuracy (target: R² > 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.