Executive Summary: Data Preparation and Strategic Readiness

To: DeShawn Washington (Manager), Luana Rodriquez (Senior Analyst)

From: Htoo Pyae Shan, Data Analyst Intern

Date: October 30, 2025

Subject: Completion of Data Integrity and Feature Engineering Phase

1. Status and Key Accomplishments

The initial phase of the Predictive Regression Model project—Data Preparation and Integrity—is now complete. The **2017 Yellow Cab Trip Data** has been cleaned, validated, and enriched, ensuring the foundation for the upcoming Exploratory Data Analysis (EDA) and model training is robust.

Phase Deliverable	Result	Value to Project
Data Cleaning	Systematically filtered non-sensical data, including negative fares, zero distances, and trips with impossible average speeds (>60 MPH).	The data integrity check successfully removed **0.847%** of the original records, eliminating errors that would have biased the final model.
Feature Engineering	Created two critical predictors: trip_duration_minutes and avg_speed_mph.	Provided essential time and speed context, which is mandatory for a distance-based fare regression model.

2. Analytical Findings for Model Strategy

Exploratory Data Analysis (EDA) has yielded two critical findings that must inform the subsequent model building process:

- **Primary Predictors Confirmed:** Correlation analysis confirms that **trip_distance** (0.93 correlation) and **trip_duration_minutes** (0.17 correlation) are the strongest non-financial predictors of the target variable, total_amount. These features will be central to the regression model.
- **Transformation is Mandatory:** Visual analysis confirms that the target variable (total_amount) and all key predictors are **highly right-skewed**. This violates the assumption of normality required by linear regression, which would lead to an unreliable model.

3. Next Steps (Pre-Modeling)

Based on these findings, I recommend the following immediate action before moving into model fitting:

The next mandatory step is to apply a **logarithmic transformation** to the highly skewed features (specifically total_amount, trip_distance, and trip_duration_minutes). This will normalize their distributions, satisfying the fundamental assumptions of linear regression and maximizing the predictive performance of the final model.

The dataset is now technically prepared and ready for this final transformation and subsequent model construction.