# Big Data - Practice 1

Oleksandr Severhin

October 2025

## Contents

# 1 Introduction

For the Big Data course, I have chosen a NYC Taxi dataset. The dataset was downloaded from here: *dataset*. It provides NYC taxi rides data that is a perfect fit for our task of working with the Big Data dataset, EDA, and creating various models for the data analysis.

# 2 Setup

In order to work with R and the chosen dataset, we can use RStudio or VSCode (with R extension). Firstly, we have to download R on the machine, then we can choose the program we want to code into and set everything up.

I have chosen VSCode with R Extensions. Also, I have chosen *pacman* for package manager.

```
# Installing pacman and all other necessary packages:
if (!require("pacman")) install.packages("pacman")
pacman::p_load(arrow, dplyr, lubridate, ggplot2, ggcorrplot, broom, MASS, patchwork)

# Creating folders data and plots folders:
if (!dir.exists("data")) dir.create("data")
if (!dir.exists("plots")) dir.create("plots")

# Reading the data:
tryCatch({
  taxi_data_raw <- read_parquet("data/yellow_tripdata_2025-01.parquet")
  cat("Dataset loaded successfully.\n")
}, error = function(e) {
  stop("Failed to load the dataset. Make sure 'yellow_tripdata_2025-01.parquet'
  is in the 'data' directory.")
})
```

# 3 An Exploratory Data Analysis

Before building any models, we conducted an exploratory data analysis (EDA) to understand the data's fundamental characteristics and quality. The purpose of this phase was to uncover irregularities, spot key trends, and develop an intuition for how the variables interact.

Columns overview:

```
# Filtering out outliers and illogical values,
# engineer new features from existing columns,
# filter based on the newly created features to remove more outliers,
# select the final set of columns for the analysis:

taxi_data_clean <- taxi_data_raw %>%
```

```
  filter(
    total_amount > 2.5 & total_amount < 250,
    fare_amount > 0 & fare_amount < 200,
    trip_distance > 0.1 & trip_distance < 50,
    passenger_count > 0 & passenger_count < 7,
    RatecodeID == 1,
    payment_type %in% c(1, 2)
  ) %>%
  mutate(
    payment_method = factor(payment_type, levels = c(1, 2), labels = c("Card", "Cash")),
    pickup_hour = hour(tpep_pickup_datetime),
    day_of_week = wday(tpep_pickup_datetime, label = TRUE, week_start = 1),
    trip_duration_mins = as.numeric(difftime(tpep_dropoff_datetime,
    tpep_pickup_datetime, units = "mins")),
    average_speed_mph = trip_distance / (trip_duration_mins / 60)
  ) %>%
  filter(
    trip_duration_mins > 1 & trip_duration_mins < 120,
    !is.na(average_speed_mph) & average_speed_mph > 1 & average_speed_mph < 70
  ) %>%
  dplyr::select(
    total_amount,
    trip_distance,
    passenger_count,
    payment_method,
    pickup_hour,
    day_of_week,
    trip_duration_mins,
    fare_amount
  )

# Saving the cleaned dataset:
write_parquet(taxi_data_clean, "data/cleaned_yellow_tripdata_2025-01.parquet")
```

### 3.1 Distribution of Key Variables

We first examined the distributions of our target variable, `total_amount`, and a primary predictor, `trip_distance`. Both variables showed a significant **rightward skew**, a common feature in datasets related to financial transactions and travel. This indicates that most taxi rides are relatively short and inexpensive, with a long tail of less frequent, expensive long-distance trips. This skewness suggested that a logarithmic transformation would be a valuable step for improving model performance and satisfying the assumption of normally distributed residuals.

```
dist_total_amount <- ggplot(taxi_data_clean, aes(x = total_amount)) +
```

```
  geom_histogram(aes(y = ..density..), bins = 50, fill = "steelblue",
  color = "white", alpha = 0.8) +
  geom_density(color = "darkred", size = 1) +
  labs(title = "Distribution of Total Fare Amount",
  x = "Total Amount ($)", y = "Density")

dist_trip_distance <- ggplot(taxi_data_clean, aes(x = trip_distance)) +
  geom_histogram(aes(y = ..density..),
  bins = 50, fill = "skyblue", color = "white", alpha = 0.8) +
  geom_density(color = "darkblue", size = 1) +
  labs(title = "Distribution of Trip Distance",
  x = "Trip Distance (miles)", y = "Density")
```

### 3.2 Variable Relationships

Next, we visualized the relationships between variables. A clear and strong positive linear trend emerged between `trip_distance` and `total_amount`, confirming that as trip length increases, the fare predictably rises. We also observed a notable pattern with payment methods: credit card transactions consistently resulted in higher recorded fares compared to cash. This is likely because tips are officially logged on card payments but are often unrecorded when paid in cash.

```
scatter_dist_fare <- ggplot(sample_n(taxi_data_clean, 5000),
  aes(x = trip_distance, y = total_amount)) +
  geom_point(alpha = 0.4, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(
    title = "Trip Distance vs. Total Fare Amount",
    subtitle = "A strong positive linear relationship is evident",
    x = "Trip Distance (miles)",
    y = "Total Amount ($)"
  )

# Saving the plot:
ggsave("plots/distance_vs_fare.png", plot = scatter_dist_fare,
width = 8, height = 6, dpi = 300)
```

### 3.3 Correlation Insights

To quantify these relationships, we generated a correlation matrix for all numerical features. The matrix highlighted a very strong positive correlation between `trip_distance`, `fare_amount`, and `total_amount`, confirming they are all closely tied to the underlying fare structure of taxi rides. Conversely, the correlation with time-based variables like the pickup hour was weak, suggesting that while time of day has an effect, it is secondary to the trip's length.

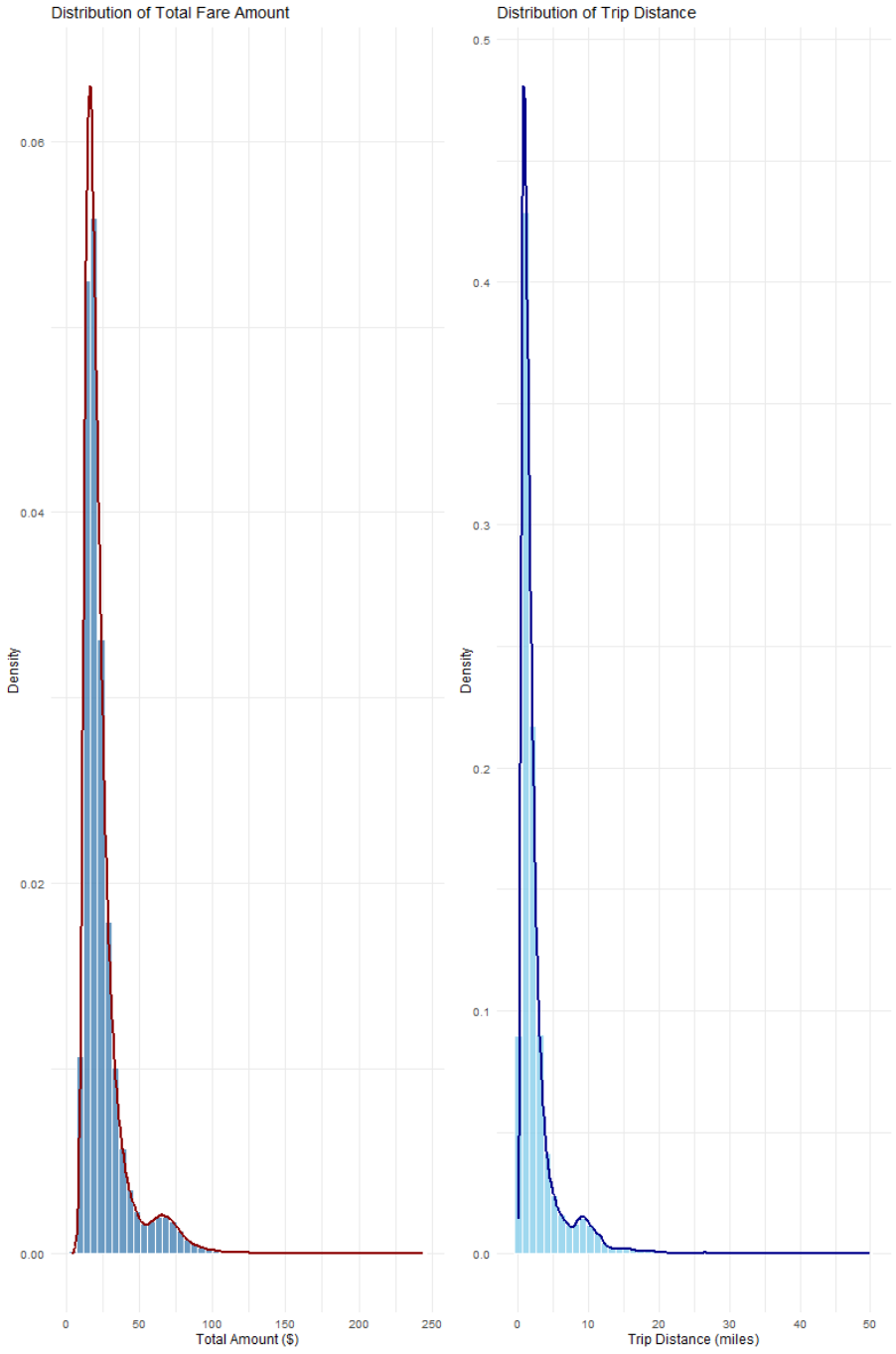


Figure 1: Enter Caption

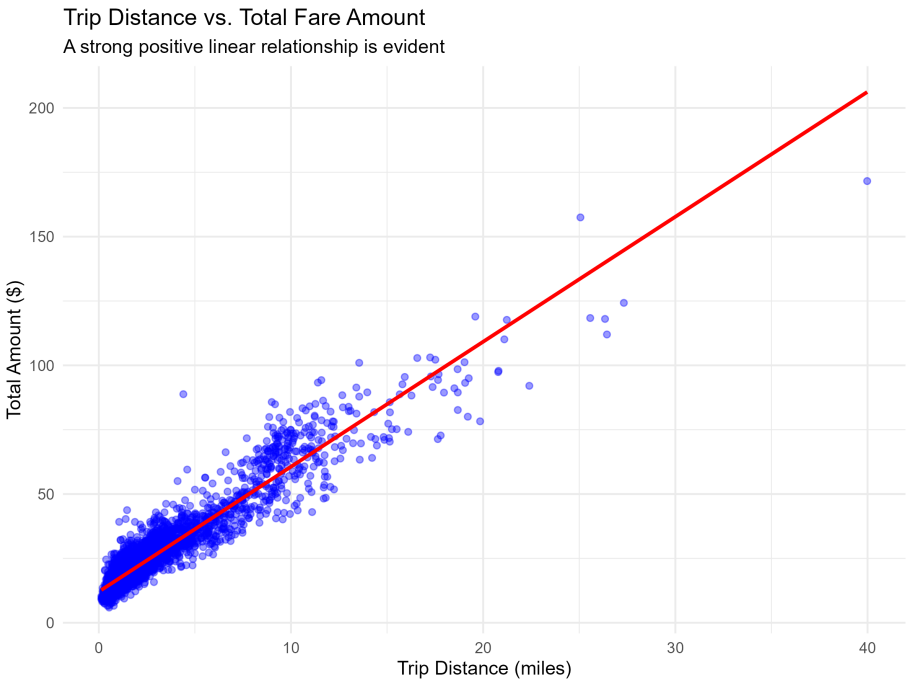


Figure 2: Distance vs Fare

```
numeric_vars <- taxi_data_clean %>% dplyr::select(where(is.numeric))
cor_matrix <- cor(numeric_vars)

corr_plot <- ggcorrplot(cor_matrix,
  method = "square",
  type = "lower",
  lab = TRUE,
  lab_size = 3,
  colors = c("#6D9EC1", "white", "#E46726")
) +
labs(title = "Correlation Matrix of Numeric Variables")
ggsave("plots/correlation_matrix.png", plot = corr_plot,
width = 8, height = 6, dpi = 300)
```

# 4 Developing a Predictive Model

Informed by our exploratory findings, we developed a series of linear regression models to predict the `total_amount`. We started with a simple model and progressively added complexity to improve its explanatory power.

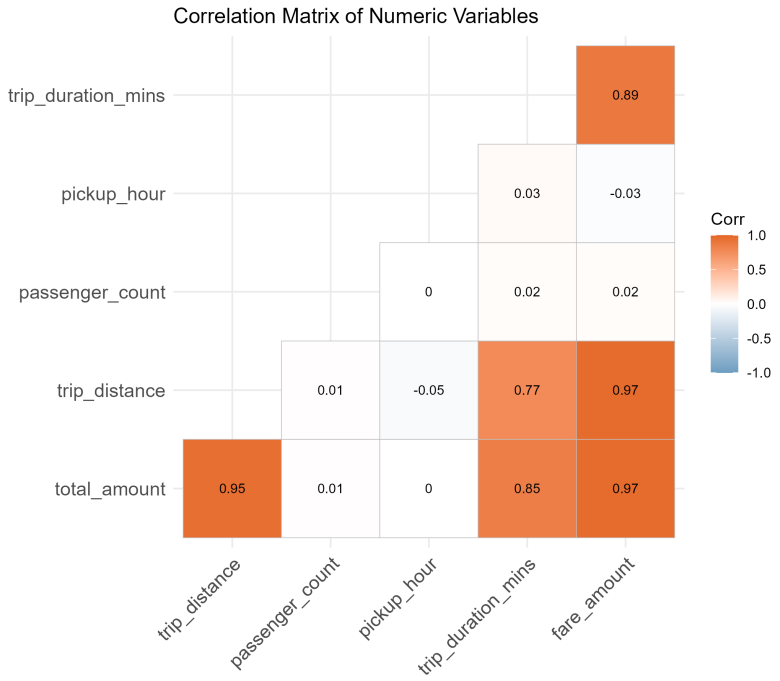


Figure 3: Correlation Matrix

## 4.1 Model 1: A Simple Distance-Based Model

Our initial model used only `trip_distance` to predict the fare. The model estimated a base fare of about $12.14 (the intercept) and an additional cost of approximately $4.84 for each mile traveled (the slope). Impressively, this simple model could account for 89.7% of the variance in taxi fares (Adjusted $R^2 = 0.897$), underscoring how central distance is to NYC taxi pricing.

```
model1 <- lm(total_amount ~ trip_distance, data = taxi_data_clean)
cat("\n--- Summary of Model 1: Simple Linear Regression ---\n")
print(summary(model1))
```

```
# Model 1 Output
> print(summary(model1))

Call:
lm(formula = total_amount ~ trip_distance, data = taxi_data_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-109.326   -2.509   -0.502    1.998  220.891
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.135668   0.003876    3131   <2e-16 ***
trip_distance  4.837876   0.001016    4760   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.722 on 2609482 degrees of freedom
Multiple R-squared:  0.8967,    Adjusted R-squared:  0.8967
F-statistic: 2.266e+07 on 1 and 2609482 DF,  p-value: < 2.2e-16
```

## 4.2 Model 2: A More Comprehensive Multiple Regression Model

To refine our prediction, we introduced three additional predictors: `passenger_count`, `payment_method`, and `pickup_hour`.

**Key Findings:**

- **Trip distance** remained the most powerful predictor, with its coefficient holding steady.
- **Payment method** proved significant. Payments made in cash were associated with a recorded total fare that was, on average, $4.37 lower than card payments, likely due to unrecorded cash tips.
- The number of **passengers** and the **pickup hour** had small but statistically significant effects, indicating that fares are slightly higher for trips with more people or those occurring later in the day.
- The model's overall fit improved, reaching an adjusted $R^2$ of 0.909.

```
model2 <- lm(total_amount ~ trip_distance + passenger_count
+ payment_method + pickup_hour, data = taxi_data_clean)
cat("\n--- Summary of Model 2: Multiple Linear Regression ---\n")
print(summary(model2))

#  Model 2 Output
> print(summary(model2))

Call:
lm(formula = total_amount ~ trip_distance + passenger_count +
    payment_method + pickup_hour, data = taxi_data_clean)

Residuals:
     Min      1Q  Median      3Q     Max
-106.448  -2.389  -0.462   1.812 220.207

Coefficients:
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.0291596  0.0087285 1263.58   <2e-16 ***
trip_distance       4.8568348  0.0009553 5084.14   <2e-16 ***
passenger_count     0.1020476  0.0037144   27.47   <2e-16 ***
payment_methodCash -4.3650052  0.0081049 -538.56   <2e-16 ***
pickup_hour         0.1038806  0.0004218  246.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.432 on 2609479 degrees of freedom
Multiple R-squared:  0.909,     Adjusted R-squared:  0.909
F-statistic: 6.519e+06 on 4 and 2609479 DF,  p-value: < 2.2e-16
```

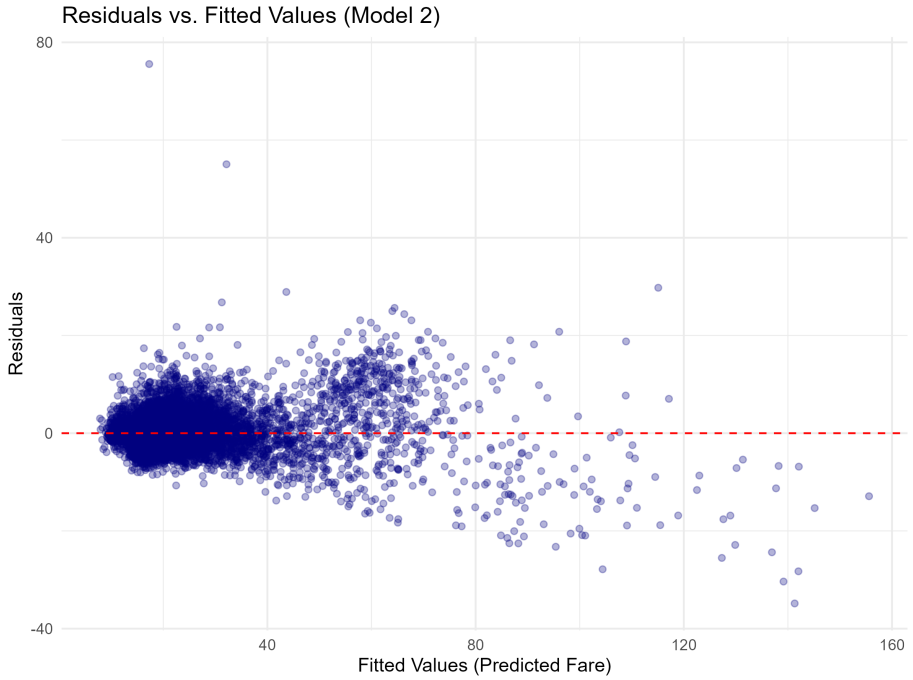


Figure 4: Residuals vs Fitted

## 4.3 Evaluating Model Assumptions

An analysis of the model's residuals and diagnostic plots pointed to areas for improvement. A "Residuals vs. Fitted" plot showed a subtle curve, suggesting some uncaptured non-linear relationships. Furthermore, the "Normal Q-Q" plot revealed that the residuals were not perfectly normally distributed, especially

at the tails. These observations reinforce the idea that transforming the target variable (e.g., using a logarithm) could create a more robust and valid model.

```
model2_augmented <- augment(model2)

residuals_plot <- ggplot(sample_n(model2_augmented, 10000),
  aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.3, color = "navy") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Residuals vs. Fitted Values (Model 2)",
    x = "Fitted Values (Predicted Fare)",
    y = "Residuals"
  )
ggsave("plots/residuals_vs_fitted.png", plot = residuals_plot,
width = 8, height = 6, dpi = 300)

qq_plot <- ggplot(sample_n(model2_augmented, 10000), aes(sample = .resid)) +
  stat_qq(alpha = 0.3) +
  stat_qq_line(color = "red", size = 1) +
  labs(
    title = "Normal Q-Q Plot of Residuals (Model 2)",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  )
ggsave("plots/qq_plot.png", plot = qq_plot, width = 8, height = 6, dpi = 300)
cat("Saved plot: plots/qq_plot.png\n")
```

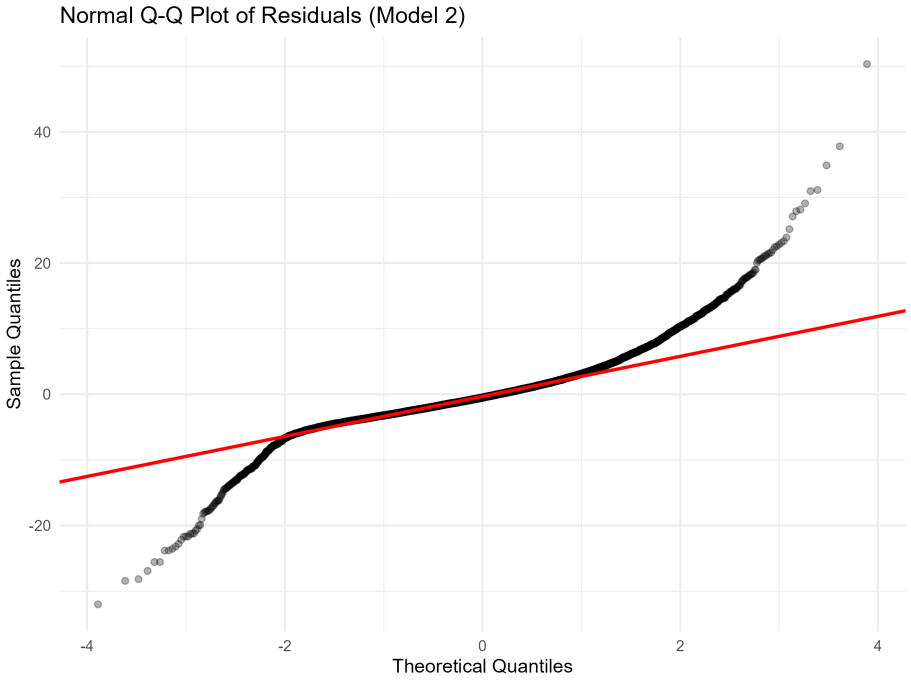


Figure 5: QQ Plot

# 5 Project Appendix: Dataset and Analysis Summary

## 5.1 Dataset Description

The dataset consists of trip records from NYC Yellow Taxis for January 2025. The goal of this analysis is to model the `total_amount` of a trip.

### Original Dataset Columns

- `VendorID`: An identifier for the taxi service provider.
- `tpep_pickup_datetime`: The date and time when the passenger was picked up.
- `tpep_dropoff_datetime`: The date and time when the passenger was dropped off.
- `passenger_count`: The number of passengers in the vehicle.
- `trip_distance`: The total distance of the trip, measured in miles.

- `RatecodeID`: A code indicating the fare type applied to the trip (e.g., 1 for standard rate).
- `store_and_fwd_flag`: Indicates if the trip data was stored locally before server upload.
- `PULocationID`: The identifier for the NYC Taxi Zone where the trip began.
- `DOLocationID`: The identifier for the NYC Taxi Zone where the trip ended.
- `payment_type`: A numeric code for the payment method used.
- `fare_amount`: The base cost of the trip, calculated by time and distance.
- `extra`: Additional charges for factors like rush hour or overnight trips.
- `mta_tax`: A mandatory tax of $0.50 from the Metropolitan Transportation Authority (MTA).
- `tip_amount`: The amount of tip provided, typically recorded only for credit card payments.
- `tolls_amount`: The total cost of all tolls paid during the trip.
- `improvement_surcharge`: A mandatory surcharge for infrastructure improvements.
- `total_amount`: The **target variable**; the total amount paid by the passenger.
- `congestion_surcharge`: An additional fee for trips in high-traffic zones.
- `Airport_fee`: A fee for pickups or drop-offs at airports.

#### 5.1.1 Engineered Features

- `payment_method`: The numeric `payment_type` was converted into a descriptive factor (Card, Cash) for clearer interpretation in the model.
- `pickup_hour`: A numeric feature (0-23) extracted from the pickup timestamp to analyze the effect of the time of day on fare prices.
- `day_of_week`: A factor (e.g., "Mon", "Tue") extracted from the pickup timestamp to capture any weekly travel patterns.
- `trip_duration_mins`: A numeric feature calculated from the pickup and dropoff times to measure the trip's duration, which was also used for cleaning.
- `average_speed_mph`: A numeric feature calculated from distance and duration, used to filter out unrealistic trips (e.g., walking-speed or impossibly fast journeys).

## 5.2 Summary of Modeling Steps

The analysis began by loading the raw Parquet file using the efficient `arrow` package. A multi-stage cleaning process was applied, filtering records based on logical bounds for fare, distance, and passenger count. New features were then engineered to capture temporal patterns and trip dynamics. A second filtering step used these new features, like trip duration and average speed, to further refine the dataset. Exploratory analysis visualized variable distributions and relationships, confirming a strong linear trend between distance and fare. Finally, two linear regression models were built and evaluated: a simple model using only `trip_distance` and a multiple regression model incorporating additional predictors.

## 5.3 Interpretation of Results

The final multiple regression model (Model 2) demonstrated a strong fit, explaining approximately 90.9% of the variance in the total trip amount (Adjusted $R^2 = 0.909$). `Trip_distance` was the most dominant predictor, with each additional mile increasing the expected fare by about $4.86. The model also quantified other effects: cash payments were associated with a recorded fare that was, on average, $4.37 lower than card payments, likely due to unrecorded tips. Fares also saw a minor increase of about $0.10 per additional passenger and $0.10 for each hour later in the day. Diagnostic plots revealed that while the model is highly predictive, the residuals exhibit non-normality and slight non-linearity, indicating opportunities for future model improvement.

## 5.4 Challenges Encountered

The primary challenge was the computational demand of processing a dataset with over three million initial records. This was managed by using the high-performance `arrow` package for data loading and the `dplyr` library for efficient data manipulation. For visualization and diagnostics, which can be slow with large datasets, a key strategy was to use random sampling (with `sample_n`) to generate plots from a smaller, representative subset of the data. This approach provided accurate insights into data patterns without causing performance issues.

## 5.5 Question for Feedback

The diagnostic plots for Model 2 show evidence of heteroskedasticity (a fanning pattern in the residuals) and non-normal residuals, which aligns with the right-skewed distribution of `total_amount` found during the EDA. Given these findings, would applying a logarithmic transformation to the `total_amount` variable be the most effective next step to improve the model's adherence to the core assumptions of linear regression?