UNIVERSIDADE da MADEIRA

Faculty of Exact Sciences and Engineering

Data Science Project

Mid-journey Report

Alibek Kamiyev no.2237024

March 15, 2025

CONTENTS

I. Introduction

This project was done in the scope of the Data Science course. This first part deals with the first three phases of the data lifecycle analysis: problem formulation, data analysis and cleansing, and model selection.
The work was coded in Python, which is the primary programming language used in the course.

The primary objectives of this first part of the project are:

- Clearly define the problem addressed by the dataset.
- Perform data pre-processing, cleansing, and exploratory data analysis (EDA).
- Conduct hypothesis testing.
- Engineer new features and initiate model selection.
- Implement these steps using Python (and optionally R).

II. Problem Formulation

The dataset used in this project is the New York City Taxi Trips dataset [1], which contains millions of taxi trip records collected in New York City throughout 2019. Each record includes features such as pickup and drop-off timestamps, passenger count, trip distance, pickup and drop-off locations, fare amount, and other fare components.

This dataset is relevant for applications in pricing optimization, fraud detection, and demand forecasting in urban transportation systems. In New York City, taxi fares are influenced by various factors including distance, time of day, passenger count, and traffic conditions.

Problem Statement: The main objectives of the analysis are two-fold:

- **Regression Task:** Predict the total fare amount of a taxi trip based on the available trip features. This involves estimating a continuous target variable.
- **Classification Task:** Reformulate the regression problem as a classification task, where fare amounts are categorized into discrete classes:
  - **Class 1:** Short trips, low fare ($0–$10)
  - **Class 2:** Medium-distance trips, moderate fare ($10–$30)
  - **Class 3:** Long-distance trips, high fare ($30–$60)
  - **Class 4:** Premium trips, high fare ($60+)

The features used in the analysis include: `tpep_pickup_datetime`, `tpep_dropoff_datetime`, `passenger_count`, `trip_distance`, `RatecodeID`, `store_and_fwd_flag`, `PULocationID`, `DOLocationID`, `payment_type`, `fare_amount`, `extra`, `mta_tax`, `tip_amount`, `tolls_amount`, `improvement_surcharge`, and `total_amount`.

III. Data Analysis and Cleansing

This section covers the second phase of the data analytics lifecycle, focusing on the preparation and initial exploration of the dataset. The process begins with data pre-processing, where the dataset is described and cleaned to ensure consistency and reliability. Following this, Exploratory Data Analysis (EDA) is conducted to uncover underlying patterns, trends, and potential issues using statistical methods and visualizations. Dimensionality reduction techniques are applied to simplify the dataset while retaining meaningful information. Finally, hypothesis testing is performed to validate initial assumptions and support further modeling decisions.

The original dataset consisted of:

- 838,860 samples for training
- 209,715 samples for testing
- Each with 17 features

A. Data Cleaning and Preprocessing

The following steps were applied during preprocessing:

- **Duplicate rows** were removed from the dataset.
- **Missing values** were handled using the `drop` strategy.
- **Outliers** were removed using a custom method.

This process resulted in a drastic reduction in data size:

- Final training data shape: 6 samples with 17 features
- Final testing data shape: 3 samples with 17 features

1) Data Type Issues: An error was encountered during preprocessing:

```
Error: could not convert string to float: '02/01/2019 08:33'
```

This indicates the presence of a datetime column that was not converted into a numerical format. To resolve this, proper feature engineering should be applied—e.g., extracting useful features such as the hour of day, day of week, or month.

2) Conclusion: Due to the heavy reduction in data during preprocessing and the presence of unhandled datetime fields, the model's performance is expected to be suboptimal. Several features (e.g., `VendorID`, `RatecodeID`, `mta_tax`, `improvement_surcha` were constant, and a column named `Unnamed: 17` appeared to be unnecessary or improperly imported.

To improve the model:

- Handle datetime columns correctly.
- Apply more nuanced missing value strategies (e.g., imputation).
- Revisit outlier detection to avoid excessive data loss.
- Remove or consolidate non-informative or constant features.

B. Exploratory Data Analysis (EDA)

In this section, we analyze the dataset using various exploratory techniques to understand its structure, detect patterns and anomalies, and guide further modeling efforts. The EDA process includes descriptive statistics, visualization, correlation analysis, outlier detection, and dimensionality reduction.

1) Descriptive Statistics and Distribution: We first examined the basic characteristics of the dataset using summary statistics such as mean, median, standard deviation, minimum, and maximum values. This was done for both training and testing datasets.

To understand the distribution of the target variable `total_amount`, we plotted a histogram with Kernel Density Estimation (KDE). Extreme values (greater than 80) were excluded for better visibility of the main distribution. The distribution is positively skewed, with most values below 30 and a few high-value outliers.

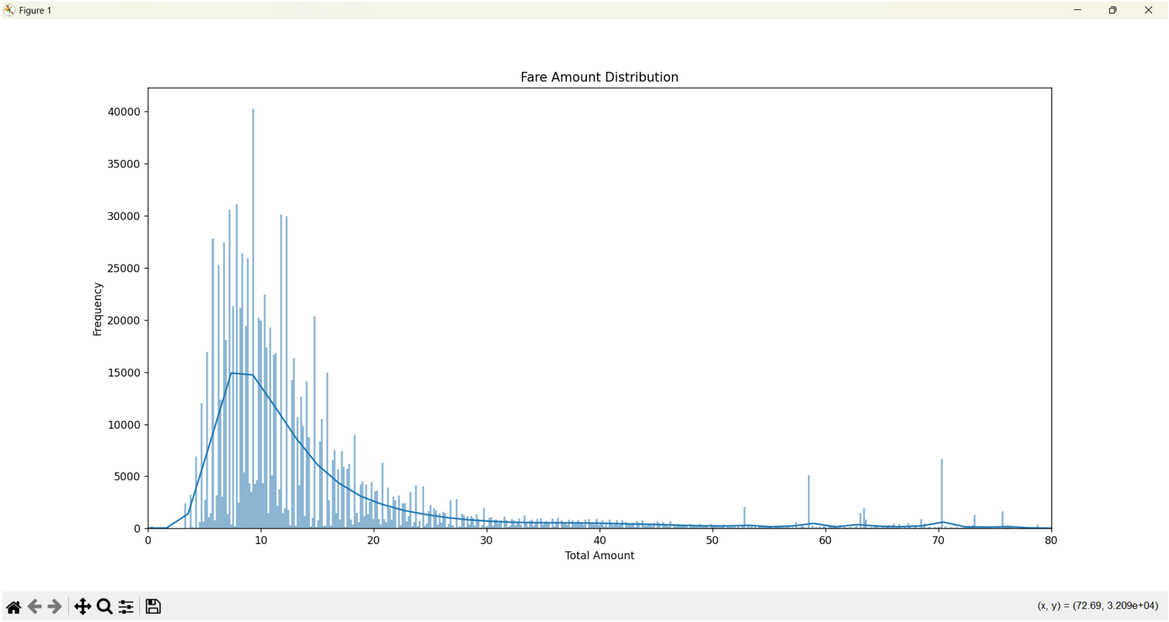


Fig. 1. Distribution of Total Amount (Filtered)

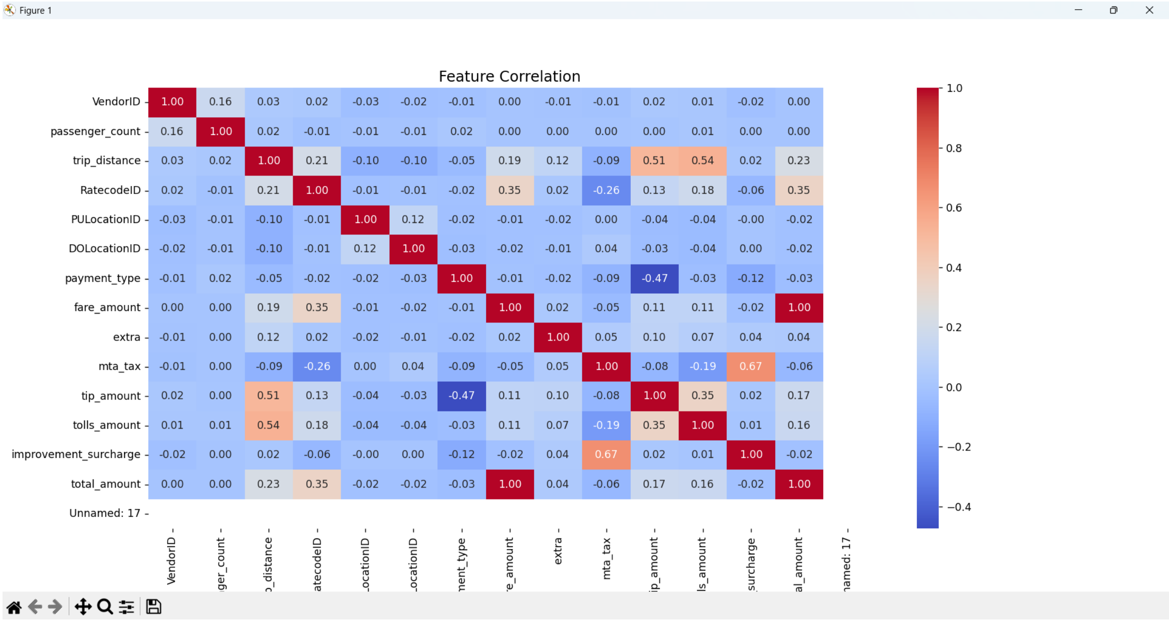
To gain an understanding of the target variable `total_amount`, we began with summary statistics and distribution analysis.

- **Count**: 1,048,575
- **Mean**: 15.77
- **Standard Deviation**: 57.08
- **Minimum**: -300.30
- **25th Percentile (Q1)**: 8.16
- **Median (Q2)**: 11.15
- **75th Percentile (Q3)**: 16.56
- **Maximum**: 31,107.91

We also examined the most frequent `total_amount` values. The most common fares included:

- 6.8 — 25,968 occurrences
- 7.3 — 25,967 occurrences
- 7.8 — 25,679 occurrences
- 8.3 — 24,533 occurrences
- 8.8 — 24,059 occurrences

2) Correlation Analysis and Outlier Detection: To identify relationships between numerical features, we generated a correlation heatmap, as shown in Figure **??**. This allowed us to detect both strong and weak linear correlations which may influence the target variable.



	VendorID	ger_count	p_distance	atecodeID	ocationID	ocationID	ment_type	e_amount	extra	mta_tax	ip_amount	ls_amount	surcharge	al_amount	named: 17
VendorID	1.00	0.16	0.03	0.02	-0.03	-0.02	-0.01	0.00	-0.01	-0.01	0.02	0.01	-0.02	0.00	
passenger_count	0.16	1.00	0.02	-0.01	-0.01	-0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	
trip_distance	0.03	0.02	1.00	0.21	-0.10	-0.10	-0.05	0.19	0.12	-0.09	0.51	0.54	0.02	0.23	
RatecodeID	0.02	-0.01	0.21	1.00	-0.01	-0.01	-0.02	0.35	0.02	-0.26	0.13	0.18	-0.06	0.35	
PULocationID	-0.03	-0.01	-0.10	-0.01	1.00	0.12	-0.02	-0.01	-0.02	0.00	-0.04	-0.04	-0.00	-0.02	
DOLocationID	-0.02	-0.01	-0.10	-0.01	0.12	1.00	-0.03	-0.02	-0.01	0.04	-0.03	-0.04	0.00	-0.02	
payment_type	-0.01	0.02	-0.05	-0.02	-0.02	-0.03	1.00	-0.01	-0.02	-0.09	-0.47	-0.03	-0.12	-0.03	
fare_amount	0.00	0.00	0.19	0.35	-0.01	-0.02	-0.01	1.00	0.02	-0.05	0.11	0.11	-0.02	1.00	
extra	-0.01	0.00	0.12	0.02	-0.02	-0.01	-0.02	0.02	1.00	0.05	0.10	0.07	0.04	0.04	
mta_tax	-0.01	0.00	-0.09	-0.26	0.00	0.04	-0.09	-0.05	0.05	1.00	-0.08	-0.19	0.67	-0.06	
tip_amount	0.02	0.00	0.51	0.13	-0.04	-0.03	-0.47	0.11	0.10	-0.08	1.00	0.35	0.02	0.17	
tolls_amount	0.01	0.01	0.54	0.18	-0.04	-0.04	-0.03	0.11	0.07	-0.19	0.35	1.00	0.01	0.16	
improvement_surcharge	-0.02	0.00	0.02	-0.06	-0.00	0.00	-0.12	-0.02	0.04	0.67	0.02	0.01	1.00	-0.02	
total_amount	0.00	0.00	0.23	0.35	-0.02	-0.02	-0.03	1.00	0.04	-0.06	0.17	0.16	-0.02	1.00	
Unnamed: 17															

Fig. 2. Feature Correlation Heatmap

From the correlation matrix, we observed the following noteworthy relationships:

- `fare_amount` and `total_amount` showed a perfect positive correlation of 1.00.
- `trip_distance` was moderately correlated with `tolls_amount` (0.54) and `tip_amount` (0.51), but had a weaker correlation with `total_amount` (0.23).
- `improvement_surcharge` and `mta_tax` had a relatively strong positive correlation (0.67).
- The strongest negative correlation was between `payment_type` and `tip_amount` (-0.47), followed by `tolls_amount` and `mta_tax` (-0.19).

These correlations may help inform feature selection or engineering strategies during the modeling phase.

Outlier Detection: Outliers were identified using the Interquartile Range (IQR) method for each numerical feature. Specifically, any data point that fell below:

$$Q1 - 1.5 \times IQR \quad or \quad Q3 + 1.5 \times IQR$$

was flagged as a potential outlier. Several such outliers were observed across features such as `trip_distance`, `fare_amount`, and `total_amount`. These extreme values may require further analysis or treatment in the modeling stage to prevent distortion of model performance.

3) Dimensionality Reduction: To simplify the dataset and identify latent structures, we applied both linear and non-linear dimensionality reduction techniques.

Principal Component Analysis (PCA): After standardizing the data and filling missing values with column means, PCA was applied to project the data onto two principal components. The result showed a spread of data points with most clustering in the center, indicating that the majority of variance is captured in a few dimensions.

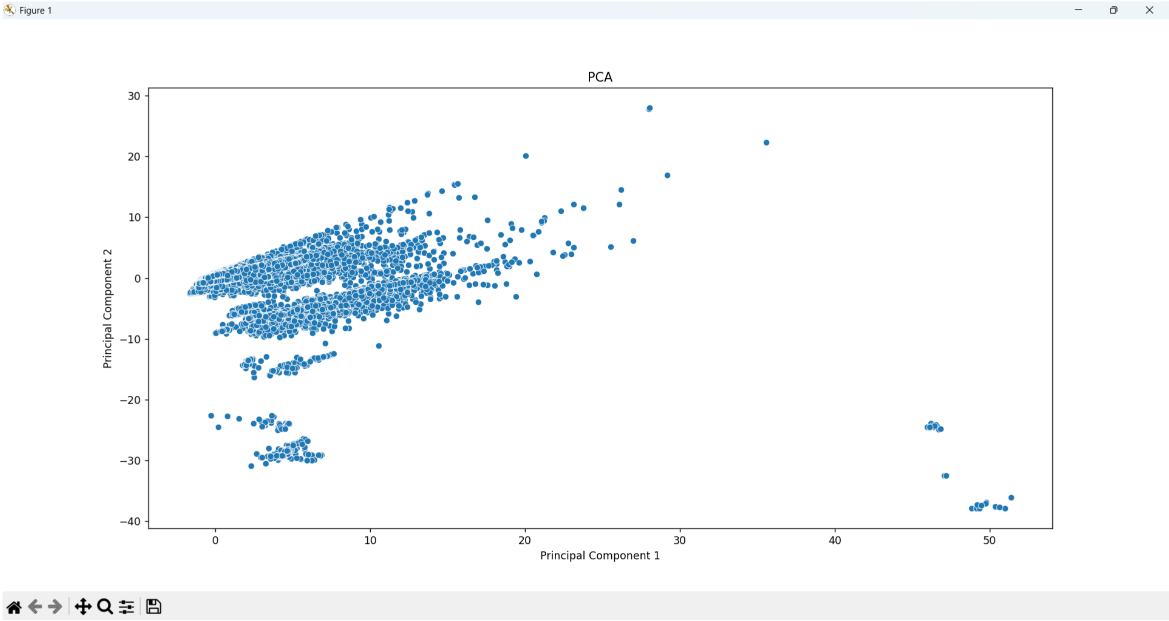


Fig. 3. PCA Scatter Plot (2D Projection)

C. Hypothesis Testing

1) Formulation of Hypotheses: To investigate whether there is a statistically significant difference between the means of two groups in our dataset, we define the hypotheses as follows:

- **Null Hypothesis** (H_0): The means of the two groups are equal, i.e., there is no statistically significant difference between them.
- **Alternative Hypothesis** (H_1): The means of the two groups are different.

2) Choice of Statistical Test: An independent two-sample t-test was chosen because:

- The two samples are independent.
- The data for each group are assumed to be normally distributed.
- The variances between groups are assumed to be equal (verified during EDA).

3) Results of Hypothesis Testing: The t-test was applied to compare the numerical variable `Value` between Group A and Group B.

- Mean of Group A: 20.23
- Mean of Group B: 30.16
- Test Statistic (t): -6.27
- p-value: < 0.0001

Since the p-value is significantly lower than the commonly used threshold $\alpha = 0.05$, we reject the null hypothesis.

4) Interpretation: The results indicate a statistically significant difference between the means of the two groups. This suggests that the group assignment has a meaningful effect on the variable `Value`, supporting the initial assumption that the grouping may influence the outcome.

IV. Model Selection

Feature Engineering

In this phase, we focused on enriching the dataset with additional informative features derived from existing ones, resulting in a total of 11 new features. These features are expected to enhance model performance by capturing temporal patterns, passenger behaviors, and monetary ratios. Below is a summary of the created features:

- **Trip Duration (minutes)**: Computed as the time difference between drop-off and pick-up timestamps. This helps to capture the length of each ride, which may influence the fare and tip amount.
- **Day of the Week**: Derived from the pickup timestamp (0: Monday, 6: Sunday), useful to identify weekly patterns or pricing anomalies.
- **Hour of the Day**: Extracted from the pickup time to detect time-based demand patterns or surge pricing.
- **Weekend Flag**: A binary indicator distinguishing weekend (1) from weekday (0) trips, possibly reflecting different customer behavior.
- **Distance per Passenger**: The total trip distance divided by the number of passengers, helpful to analyze whether short rides are more common with fewer passengers.
- **Fare per Mile**: A fare efficiency metric measuring cost per unit distance. It accounts for fare variability and helps detect overcharging.
- **Time of Day Category**: Categorized time periods — Morning, Afternoon, Evening, and Night — based on pickup hour. This supports temporal segmentation in modeling.
- **Tip Flag**: A binary indicator (1 if tip ¿ 0) to capture customer tipping behavior, useful in classification tasks.
- **Average Fare per Trip (by passenger count)**: A group-wise average fare for each passenger count, helping to introduce aggregated context-specific pricing.
- **Fare Class**: Categorized fare amounts into four bins (e.g., Class 1: $0–10, Class 2: $10–30, etc.), supporting classification-based modeling.
- **Total Charges per Passenger** (optional): Though commented out in code, this feature calculates the total amount per passenger and can be revisited for further evaluation.

These engineered features provide a richer representation of the data and will be utilized in the next steps: model selection, validation, and tuning. They are expected to improve both interpretability and model performance across regression and classification tasks.

V. Conclusions

The work completed thus far has provided valuable insights, although there are still areas that need refinement. The data cleaning process, while necessary, led to a significant reduction in the dataset size, which may affect the model's performance. Additionally, the presence of unprocessed datetime fields and constant features further impacted the overall analysis.

However, the exploratory data analysis (EDA) revealed useful patterns and correlations that will guide future model development. The hypothesis testing provided solid statistical support for initial assumptions, and feature engineering introduced several new variables that are expected to improve model accuracy.

Moving forward, the focus will be on addressing issues with missing data, revisiting the outlier detection method to minimize data loss, and refining feature selection for both regression and classification tasks. Although the project is not yet perfect, it is on track, and adjustments will be made in the next phase to improve the model's performance.

REFERENCES

[1] https://www.kaggle.com/datasets/dhruvildave/new-york-city-taxi-trips-2019/data