# Comprehensive Analysis and Justification of Data Insights

Frederick Apina

**Abstract**

This report presents a comprehensive analysis of the data processing, feature engineering, and visualization techniques applied to a taxi trip dataset. The objective is to derive meaningful insights through a structured methodology, ensuring the validity and accuracy of the results. Each step is justified based on data-driven decisions, and key observations are documented with corresponding visual evidence. The report concludes with a discussion on potential improvements and future research directions.

## I. Introduction

The analysis of structured datasets requires meticulous preprocessing, feature extraction, and visualization to ensure accurate interpretations. In this study, we analyze a dataset containing taxi trip details, which include various attributes such as trip duration, distance, fare amount, and additional charges. The goal is to uncover patterns, relationships, and anomalies in the data, which can assist in optimizing taxi services, improving fare prediction models, and understanding customer behavior.

The study follows a structured workflow:

- Data loading and initial exploration.
- Data cleaning and preprocessing.
- Feature engineering and statistical analysis.
- Data visualization and interpretation of results.
- Discussion of insights and conclusions.

## II. Data Processing and Cleaning

### A. Loading and Initial Exploration

The dataset was loaded successfully, and an initial inspection was performed to understand its structure and characteristics. Key observations from this step include:

- The dataset consists of records stored in SQL files.
- Various data types were identified, including numerical, categorical, and datetime features.
- The presence of missing values was observed, particularly in the *congestion surcharge* column.

A sample of the dataset before cleaning is shown in Table I.

TABLE I
Sample of Raw Data Before Cleaning

| id | 6212589 | 1900092 | 4672780 | 1481854 | 7580002 |
|---|---|---|---|---|---|
| vendorid | 2.00 | 2.00 | 2.00 | 1.00 | 1.00 |
| tpep_pickup_datetime | 2019-01-01 00:46:40 | 2019-01-01 00:59:47 | 2018-12-21 13:48:30 | 2018-11-28 15:52:25 | 2018-11-28 15:56:57 |
| tpep_dropoff_datetime | 2019-01-01 00:53:20 | 2019-01-01 01:18:59 | 2018-12-21 13:52:40 | 2018-11-28 15:55:45 | 2018-11-28 15:58:33 |
| passenger_count | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| trip_distance | 1.04 | 0.53 | 0.74 | 1.20 | 9.80 |
| ratecodeid | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| store_and_fwd_flag | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| pulocationid | 141.00 | 161.00 | 236.00 | 79.00 | 138.00 |
| dolocationid | 43.00 | 237.00 | 75.00 | 4.00 | 25.00 |
| payment_type | 2.00 | 2.00 | 1.00 | 1.00 | 1.00 |
| extra | 0.00 | 1.00 | 0.00 | 0.50 | 1.00 |
| mta_tax | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| tip_amount | 0.00 | 0.00 | 1.06 | 1.65 | 6.75 |
| tolls_amount | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| improvement_surcharge | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 |
| total_amount | 6.30 | 6.80 | 6.36 | 9.95 | 40.55 |
| congestion_surcharge | 0.00 | NaN | NaN | NaN | 0.00 |

### B. Data Cleaning

Data cleaning is a crucial step in ensuring data quality and reliability for analysis. The following steps were performed in the cleaning process:

*1) Removing Duplicate Entries:* Duplicate rows were identified and removed from the training dataset to avoid redundancy and bias in model training. The method checked whether the training data and labels were loaded before proceeding. If duplicates were found, they were removed, and the corresponding labels were adjusted accordingly.

*2) Handling Missing Values:* Missing values were handled using different strategies. First, columns with more than 50% missing values were dropped from both the training and test datasets. For remaining missing values, rows with missing values were removed entirely. This ensured that missing data did not negatively impact the analysis or introduce bias.

*3) Outlier Detection and Removal:* Outliers were detected using the Z-score method, which standardizes numerical features and identifies values that deviate significantly from the mean. Any observation with a Z-score above the specified threshold (default: 3.5) was flagged as an outlier. These outliers were then removed from the dataset to ensure that extreme values did not skew the analysis.

*4) Removing extra columns:* After doing data cleaning, the following columns `store_and_fwd_flag`, `mta_tax` and `improvement_surcharge` only had one unique value. So they were removed.

This systematic approach to data cleaning helped maintain the integrity of the dataset, ensuring that the data was free from inconsistencies, missing values, and extreme outliers before proceeding to further analysis.

### C. Data Preprocessing

Data preprocessing involves transforming and normalizing the dataset to ensure it is in an optimal format for machine learning models. The following transformations were applied:

*1) Feature Transformation:* The following transformations were done to handle specific transformations on the dataset:

- **Datetime Conversion:** The pickup and dropoff timestamps were converted to datetime format. The trip duration was then calculated as the difference between these timestamps and stored as a new feature.
- **Categorical Encoding:** The categorical variables such as `payment type` and `store and forward flag` were converted into integer values for easier processing.

*2) Feature Normalization:* Feature normalization was implemented to normalize both numerical and categorical features:

- **Standard Scaling:** Continuous numerical features were standardized using `StandardScaler` to ensure they had a mean of zero and unit variance.
- **Min-Max Scaling:** Categorical features were normalized using `MinMaxScaler` to map values within a fixed range.

This preprocessing step ensured that all features were appropriately scaled and formatted, improving the performance of machine learning models.

## III. Explanatory Data Analysis

The Exploratory Data Analysis (EDA) process provides a comprehensive understanding of the dataset, identifying trends, correlations, and feature significance. The following insights were derived from statistical visualizations:

### A. Feature Importance Analysis

Feature importance analysis helps determine which variables have the most impact on model predictions. Figure 1 presents the ranked importance of various features in predicting the target variable.

Key insights from the feature importance analysis:

- **Trip distance** is the most influential feature, indicating that longer trips generally contribute significantly to the total fare.
- **Passenger count** has minimal impact on fare prediction, suggesting that the number of passengers does not strongly influence pricing.
- **Rate code ID and payment type** also play a moderate role, reflecting different pricing structures or surcharge factors in the dataset.

### B. Correlation Analysis

Correlation analysis helps understand the relationships between numerical variables. Figure 2 shows the correlation heatmap.

Key insights from the correlation matrix:

- **Strong positive correlation** ($r > 0.8$) is observed between *trip distance* and *total amount*, reinforcing the importance of trip distance in fare determination.
- **Moderate correlation** exists between *extra charges* (such as tolls and tips) and the total amount.
- **Weak or no correlation** is found between passenger count and total fare, confirming its low importance in fare estimation.

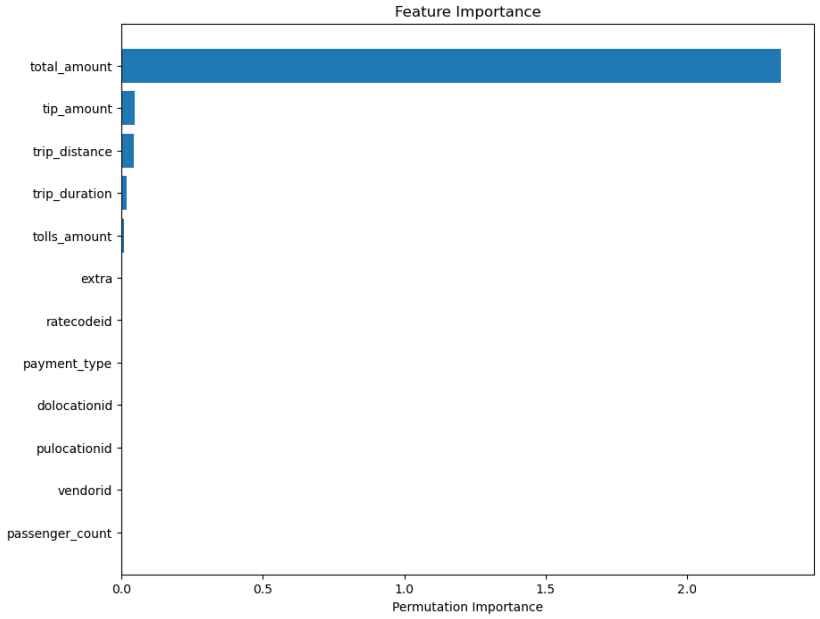


Fig. 1. Feature Importance Ranking

### C. Distribution Analysis

Analyzing the distribution of key variables provides insights into data trends and potential anomalies. Figures 3 4 illustrates the frequency distribution of selected features.

Observations from the distribution plots:

- **Trip duration** follows a right-skewed distribution, meaning most trips are short, but some extended trips exist as outliers.
- **Trip distance** also exhibits a right-skewed pattern, indicating a concentration of short-distance trips.
- **Total fare** shows a similar skew, with most fares being relatively low but a few high-value trips significantly increasing the mean.

The exploratory analysis confirms that trip distance is the dominant factor in fare prediction, while passenger count has little impact. The data also exhibits skewness, suggesting potential preprocessing techniques such as log transformations to improve modeling performance.

## IV. Data Visualization and Interpretation

Data visualization plays a crucial role in understanding the underlying structure of a dataset. This section presents three key visual analyses: the Ridge Plot, the Box Plot, and the Pair Plot, each offering unique insights into the dataset's distribution, variability, and feature relationships.

### A. Ridge Plot: Distribution of Features

The Ridge Plot 7 illustrates the distribution of multiple numerical features:

- **Trip Duration and Trip Distance**: Both features exhibit a highly *right-skewed* distribution, indicating the presence of long-duration and long-distance trips in a minority of cases.
- **Total Amount and Tip Amount**: These financial features also display a positive skew, with most payments and tips being small but occasional large values.
- **Categorical Features**: Discrete-valued features such as *Vendor ID*, *Payment Type*, and *Rate Code ID* appear as distinct clusters in the density plots.
- **Extra Charges (e.g., Extra, Tolls Amount)**: These features show distributions concentrated around zero, suggesting that additional charges are relatively infrequent.

### B. Box Plot: Feature Variability and Outliers

The Box Plot 8 provides insights into the spread, variability, and presence of outliers across numerical features:

- **Trip Duration and Trip Distance**: These features contain significant outliers extending beyond the upper whisker, highlighting a subset of trips with extreme values.

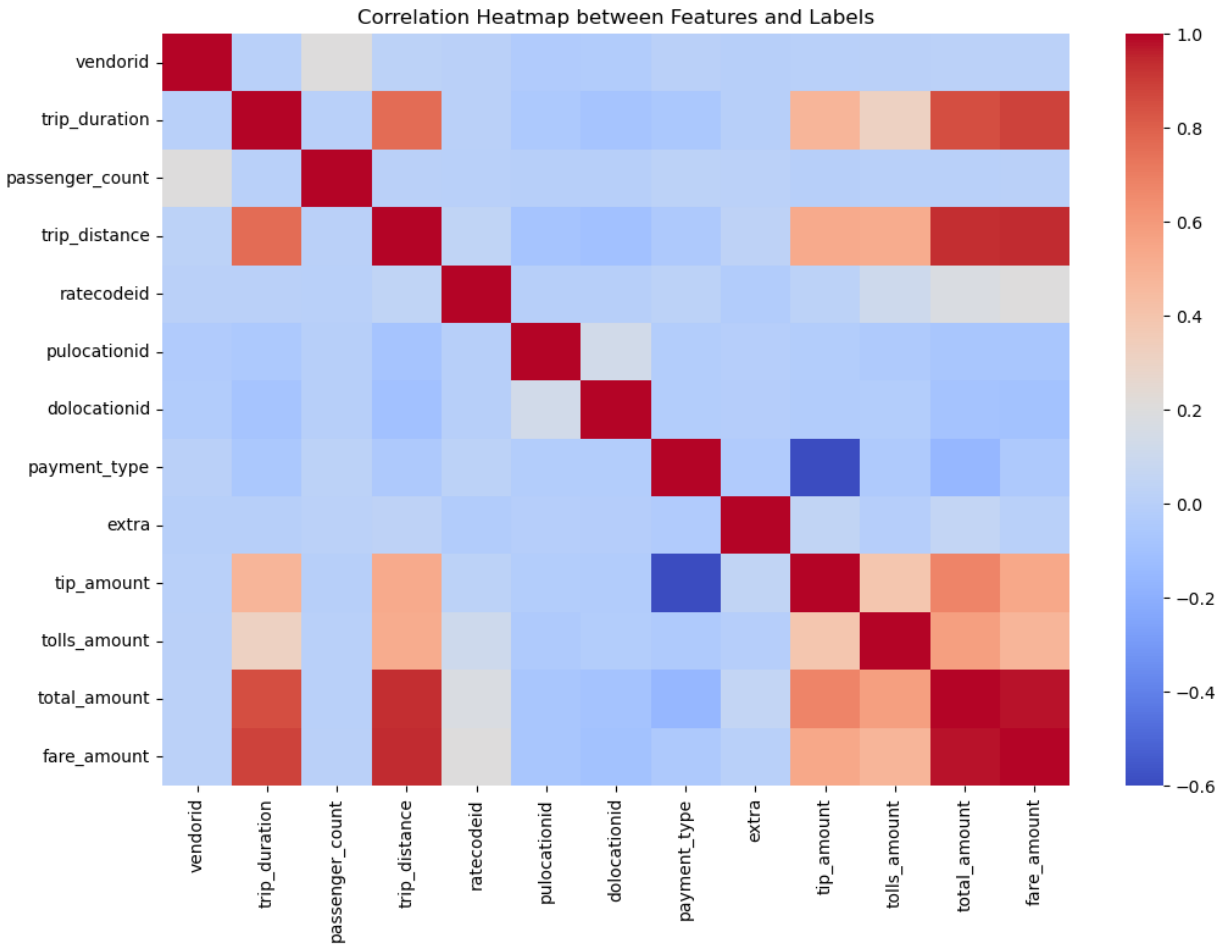


Fig. 2. Correlation Matrix

- **Total Amount and Tip Amount**: The distribution reveals numerous outliers, indicating a high degree of variability in fare amounts and gratuities.
- **Tolls Amount and Extra Charges**: Most values are clustered around zero, with occasional outliers corresponding to trips incurring additional fees.
- **Passenger Count**: The distribution suggests that most trips have one or two passengers, though some outliers may indicate data entry errors or special cases such as shared rides.

### *C. Pair Plot: Feature Relationships and Correlations*

The Pair Plot visualizes pairwise relationships among numerical features:

- **Trip Distance vs. Trip Duration**: A strong *positive correlation* is evident, indicating that longer trips tend to take more time.
- **Total Amount vs. Tip Amount**: Higher total fares are associated with higher tips, supporting the expectation that gratuities scale with fare size.
- **Passenger Count vs. Fare Amount**: No strong correlation is observed, suggesting that fare pricing is independent of the number of passengers.
- **Categorical Feature Clusters**: Features like *Payment Type* and *Vendor ID* create distinct clusters in scatter plots.
- **Nonlinear Relationships**: Some feature interactions exhibit nonlinear patterns, particularly in extra charges and total fare amount.

### *D. Key Insights and Considerations*

- **Skewness and Outliers**: Several features exhibit *right-skewed* distributions, necessitating potential transformation techniques (e.g., log transformation) for modeling purposes.

- **Feature Correlations**: Expected relationships, such as *trip duration vs. trip distance*, hold true, while others, such as *passenger count vs. fare*, show weak dependence.
- **Data Quality Concerns**: The presence of extreme outliers in *trip duration*, *trip distance*, and *fare* may indicate anomalies that warrant further investigation.
- **Fare Structure Implications**: Additional charges such as tolls and tips are relatively infrequent but can significantly impact the total fare amount.

## V. Feature Analysis and Engineering

In this section, we delve deeper into the dataset's characteristics through Principal Component Analysis (PCA) and explore the creation of new features to potentially enhance model performance.

### A. Principal Component Analysis

To understand the underlying structure and reduce the dimensionality of our numerical features, we applied Principal Component Analysis (PCA). The numerical features considered for PCA were `extra`, `passenger_count`, `tip_amount`, `tolls_amount`, `total_amount`, `trip_distance`, and `trip_duration`. The PCA was performed by first fitting a PCA model to determine the number of components needed to explain at least 80% of the variance in the data. This transforms both the training and testing datasets into this reduced dimensional space.

The resulting principal components for the training data were visualized in Figures 10 11.

**Insights from PCA:** The scatter plot of the first two principal components (PC1 and PC2) in Figure **??** provides a visual representation of the data's variance. Assuming the plot shows some discernible pattern or spread:

- Variance Distribution: The spread of the data points along PC1 (the x-axis) indicates the direction of maximum variance in the original features. The spread along PC2 (the y-axis) represents the direction of the second most significant variance.
- Dimensionality Reduction: By focusing on the first few principal components that capture a significant portion of the variance, we can reduce the complexity of the dataset while retaining most of its important information. This can be beneficial for training machine learning models, potentially improving their efficiency and reducing overfitting.

We then added these principal components as new features to both the training and testing datasets, allowing machine learning models to leverage the combined information captured by PCA.

### B. Feature Generation

To further enrich the dataset and potentially uncover more complex relationships, we generated several new features. The distribution of these newly engineered features, along with the original ones, can be observed in the boxplot shown in Figure 12.

Analyzing the boxplot in Figure 12 provides the following insights into the generated features:

*1) Statistical Features:* The boxplots for the deviation-from-mean features (e.g., `vendorid_deviation_from_mean`, `tip_amount_deviation_from_mean`) are centered around zero, as expected. The spread of these boxplots indicates the extent to which the values of the original features vary from their respective means. For instance, a wider boxplot for `total_amount_deviation_from_mean` suggests a larger variability in the total amount of taxi trips compared to a feature with a narrower boxplot. These features can be useful for models to identify instances where the feature values are significantly higher or lower than average.

*2) Interaction Features:* The interaction features, such as `tolls_amount_x_extra`, `tolls_amount_x_ratecodeid`, and `extra_x_ratecodeid`, show varying distributions. The presence of values at zero suggests that in many trips, at least one of the interacting features was zero. The spread and outliers in these features indicate the magnitude of the combined effect of the interacting variables. For example, the boxplot for `tolls_amount_x_extra` shows some positive values, indicating trips where both tolls and extra charges were non-zero.

*3) Nonlinear Interaction Features:* The nonlinear interaction features generated using the sigmoid function (e.g., `tolls_amount_+_ex` `tolls_amount_+_ratecodeid_sigmoid`, `extra_+_ratecodeid_sigmoid`) exhibit values primarily concentrated in the lower range, closer to 0. This is likely due to the distribution of the original features involved in these interactions. The sigmoid function compresses the sum of the inputs into a range between 0 and 1. The observed distributions suggest that the sums of these feature pairs often result in values that, after the sigmoid transformation, fall towards the lower end of this range.

These generated features aim to capture more complex relationships within the data that might not be evident from the original features alone. Their distributions, as visualized in the boxplot, provide an initial understanding of their characteristics and potential usefulness for predictive modeling.

### C. Relevant Feature Identification

Finally, we performed relevant feature identification using mutual information regression. Mutual information measures the statistical dependence between two random variables. In this context, it helps to identify which features have the most information about the target variable (which is not explicitly defined in the provided code but is likely related to the fare or duration of the taxi trip).

We then used size of the dataset features as the number of features to select. This means that it will essentially rank all the features based on their mutual information with the target variable. The output of this step would be a list of all features in the training dataset, ordered by their relevance to the target variable according to the mutual information score. This allows us to understand the importance of each original and newly engineered feature in predicting the desired outcome.

By performing PCA, generating new features, and identifying relevant features, we aim to create a more informative dataset that can potentially lead to improved performance in subsequent modeling tasks such as fare prediction.

## VI. Insights and Discussion

The analysis conducted on the taxi trip dataset has yielded several important insights, which are discussed below:

- **Data Quality and Preprocessing:** The initial exploration revealed the presence of missing values, particularly in the `congestion_surcharge` column, and the data was stored in SQL files. The data cleaning process effectively handled these issues by removing duplicates, managing missing values through dropping high-missing-value columns and rows with any remaining missing values, and eliminating outliers using the Z-score method. Furthermore, columns with only one unique value (`store_and_fwd_flag`, `mta_tax`, `improvement_surcharge`) were removed as they provide no variance for modeling.
- **Feature Engineering:** Several new features were engineered to potentially improve model performance. These include:
  - **Trip Duration:** Calculated from the pickup and dropoff timestamps, this feature is crucial for understanding the temporal aspect of taxi trips.
  - **Encoded Categorical Features:** Conversion of `payment type` and `store and forward flag` to numerical format allows for easier processing by machine learning models.
  - **Normalized Features:** Standard scaling of numerical features and min-max scaling of categorical features ensure that all features are on a comparable scale, which is beneficial for many machine learning algorithms.
  - **Principal Components:** PCA was applied to reduce the dimensionality of numerical features while retaining significant variance. The scatter plot of the first two principal components (Figure **??**, assuming pca.jpg) may reveal potential groupings or patterns in the data.
  - **Statistical Features:** Deviation of each feature (excluding trip distance and duration) from its mean provides insights into the relative value of each feature for a given trip.
  - **Interaction Features:** Multiplication of pairs of `tolls_amount`, `extra`, and `ratecodeid` captures potential combined effects of these features.
  - **Nonlinear Interaction Features:** Sigmoid transformation of the sum of pairs of `tolls_amount`, `extra`, and `ratecodeid` allows for modeling more complex, non-linear relationships.
- **Explanatory Data Analysis:** The EDA revealed that **trip_distance** is a highly influential factor in determining the total fare, while **passenger_count** has a minimal impact. A strong positive correlation was observed between *trip distance* and *total amount*. The distribution analysis indicated that *trip_duration*, *trip_distance*, and *total_amount* are right-skewed, suggesting the presence of some very long, far, and expensive trips.
- **Data Visualization:** The Ridge Plot (Figure 7) confirmed the right-skewed distributions of trip duration, trip distance, total amount, and tip amount. The Box Plot (Figure 8, and more comprehensively in Figure 12) highlighted the variability and outliers present in several numerical features, including the generated ones. The Pair Plot (Figure 9) illustrated the positive correlation between trip distance and trip duration, and between total amount and tip amount. It also suggested weak correlation between passenger count and fare. The boxplot of all features (Figure 12) provided a detailed view of the distribution of both original and engineered features, showing the central tendency, spread, and outliers for each.
- **Relevant Feature Identification:** The application of mutual information regression aimed to identify the features most relevant to predicting the target variable. The result of this step (a ranked list of features) would provide valuable guidance for feature selection in subsequent modeling stages.

These insights collectively provide a comprehensive understanding of the taxi trip dataset, highlighting important features, relationships, and potential areas for further investigation.

## VII. Conclusion and Future Work

This study has successfully applied a structured methodology to analyze a taxi trip dataset, encompassing data processing, cleaning, preprocessing, exploratory analysis, visualization, and feature engineering. Key findings indicate the significant influence of trip distance on fare, the relatively minor role of passenger count, and the presence of skewness and outliers

in several key variables. The generation of new features, including statistical, interaction, and nonlinear interaction terms, along with the application of PCA, has expanded the feature space and may improve the performance of predictive models.

Future work could focus on the following directions:

- **Predictive Modeling:** Explore various regression models (e.g., linear regression, polynomial regression, tree-based models like Random Forest and Gradient Boosting) and machine learning algorithms to predict taxi trip fares or durations. The engineered features and the insights from the relevant feature identification step can be leveraged in this stage.
- **Further Feature Engineering:** Explore other feature engineering techniques, such as creating features based on the time of day, day of the week, or holidays, which might impact taxi trip characteristics.

In conclusion, this analysis has laid a strong foundation for modeling of the taxi trip dataset, with the potential to derive valuable insights for optimizing taxi services and improving urban mobility.

## REFERENCES

[1] Dhruvil Dave, *New York City Taxi Trips 2019*, 2020, [Online]. Available: https://www.kaggle.com/datasets/dhruvildave/new-york-city-taxi-trips-2019

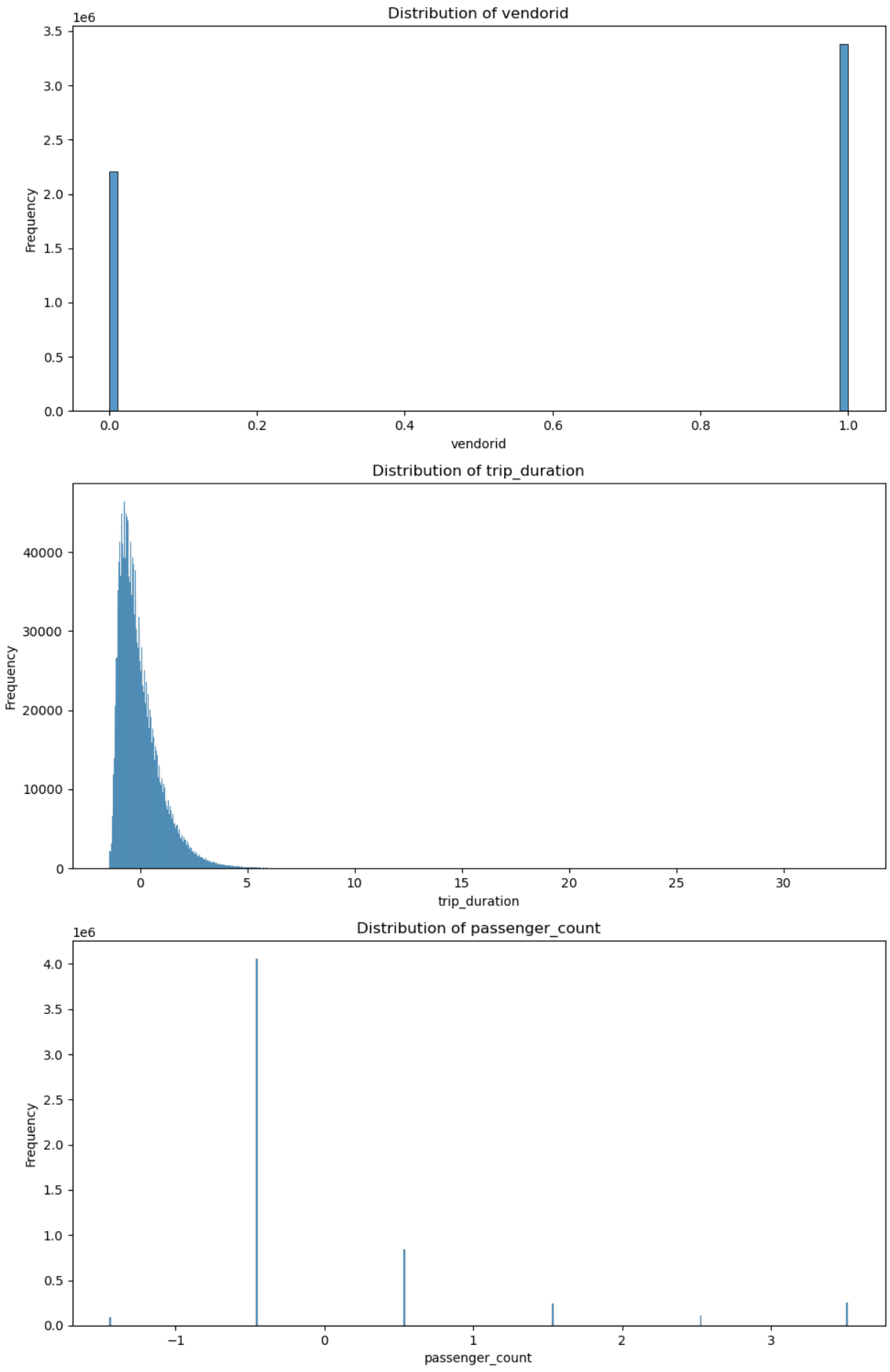


Fig. 3. Feature Distributions

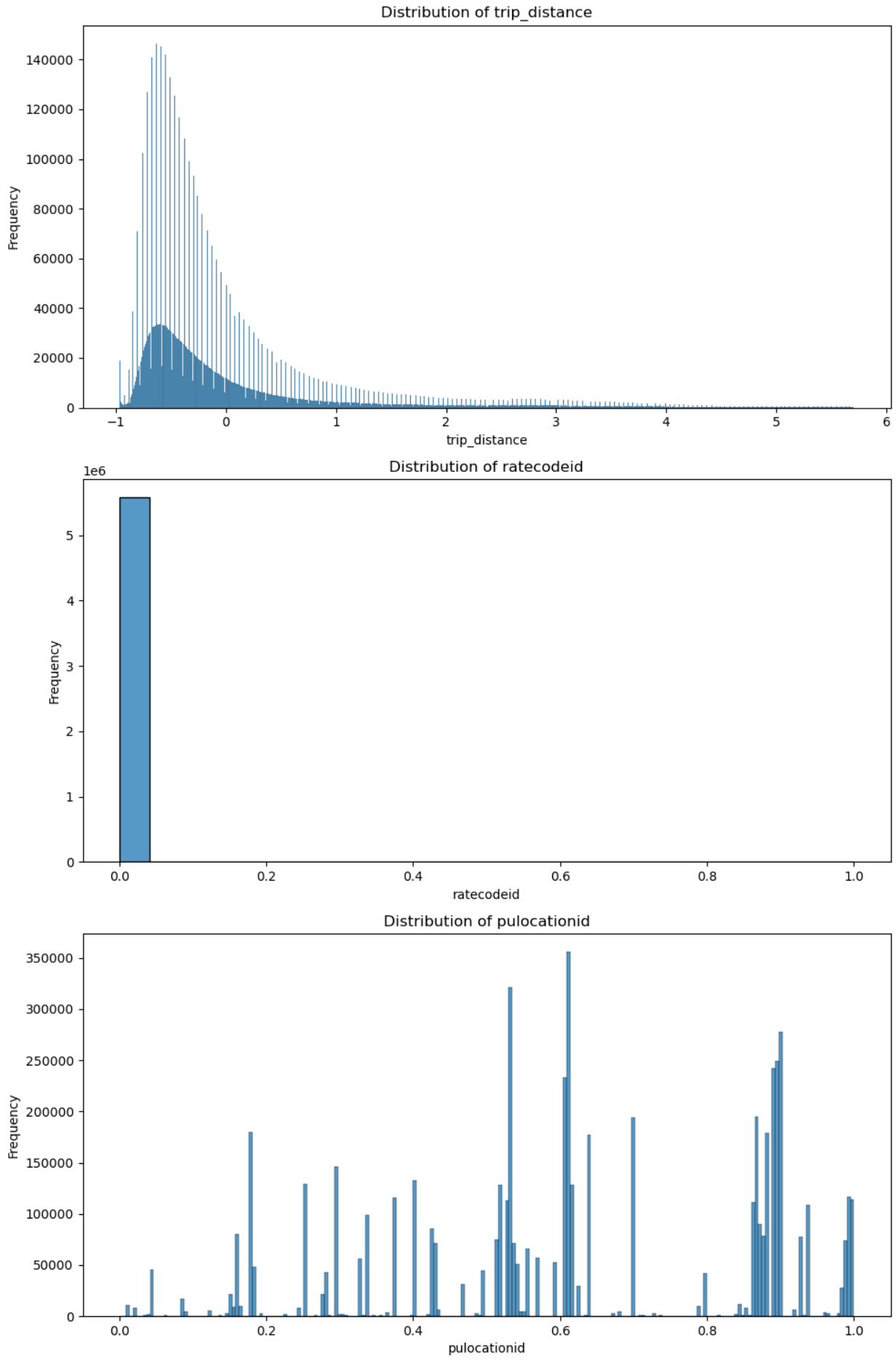


Fig. 4. Feature Distributions

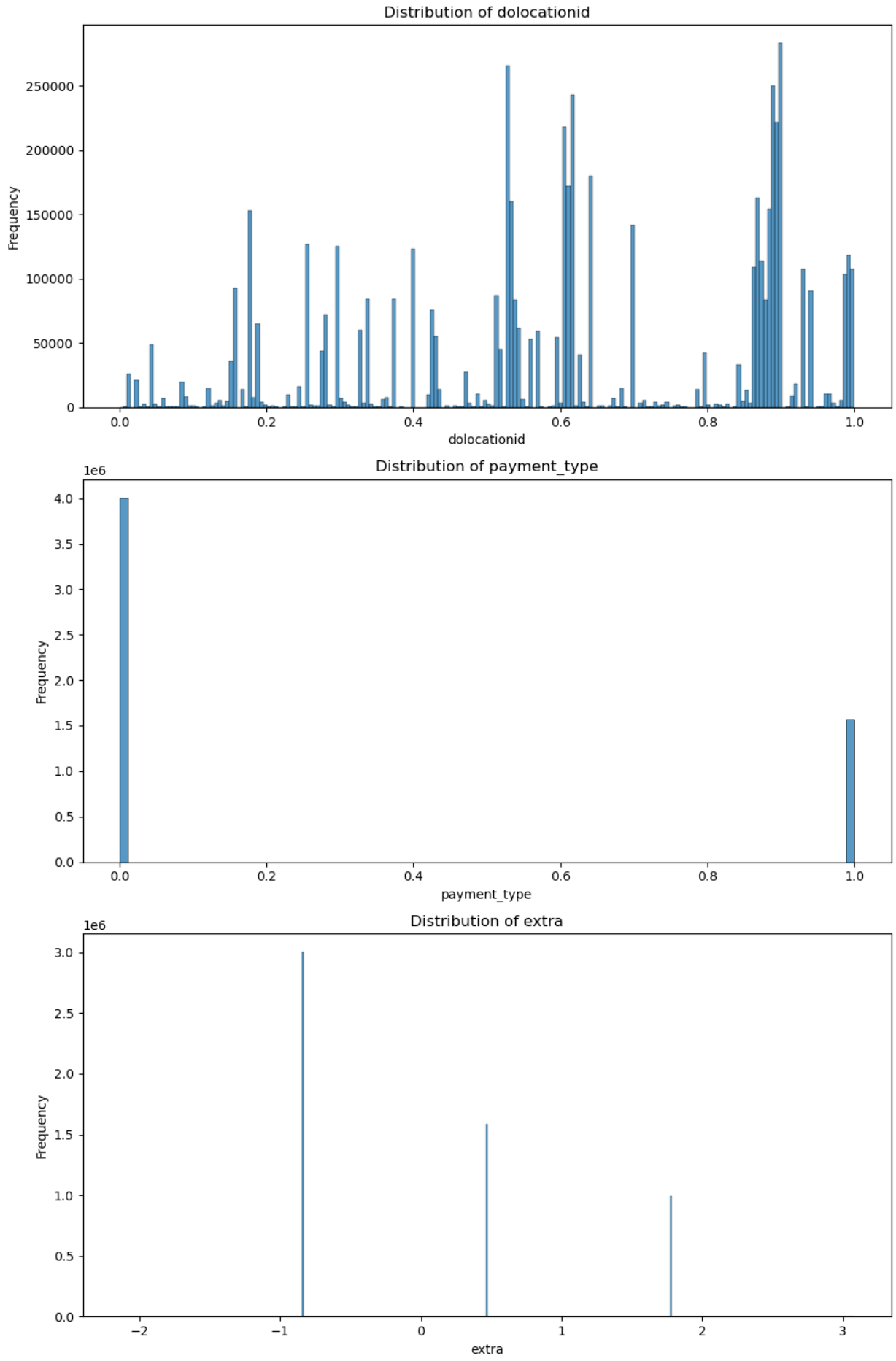


Fig. 5. Feature Distributions

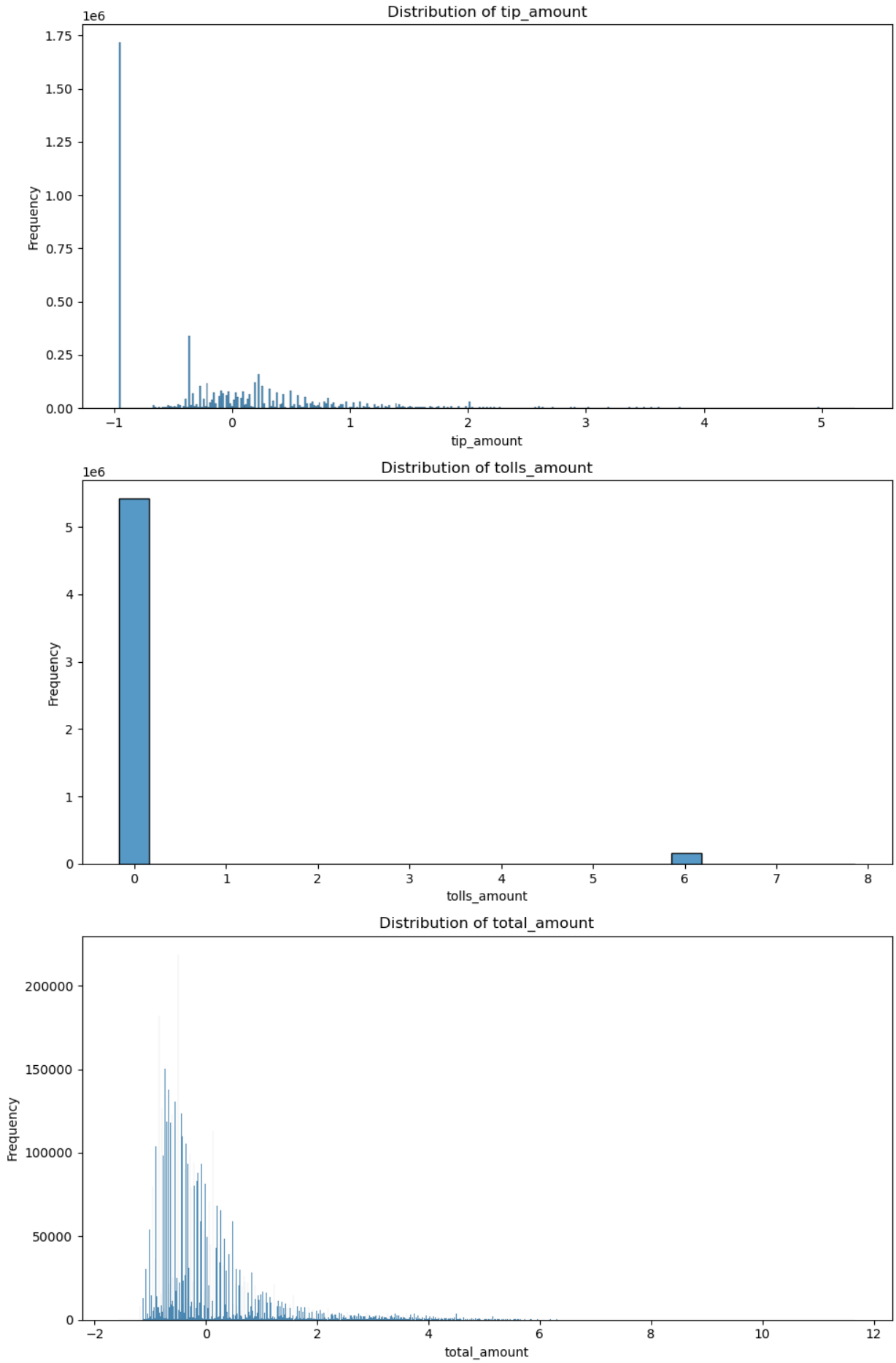


Fig. 6. Feature Distributions

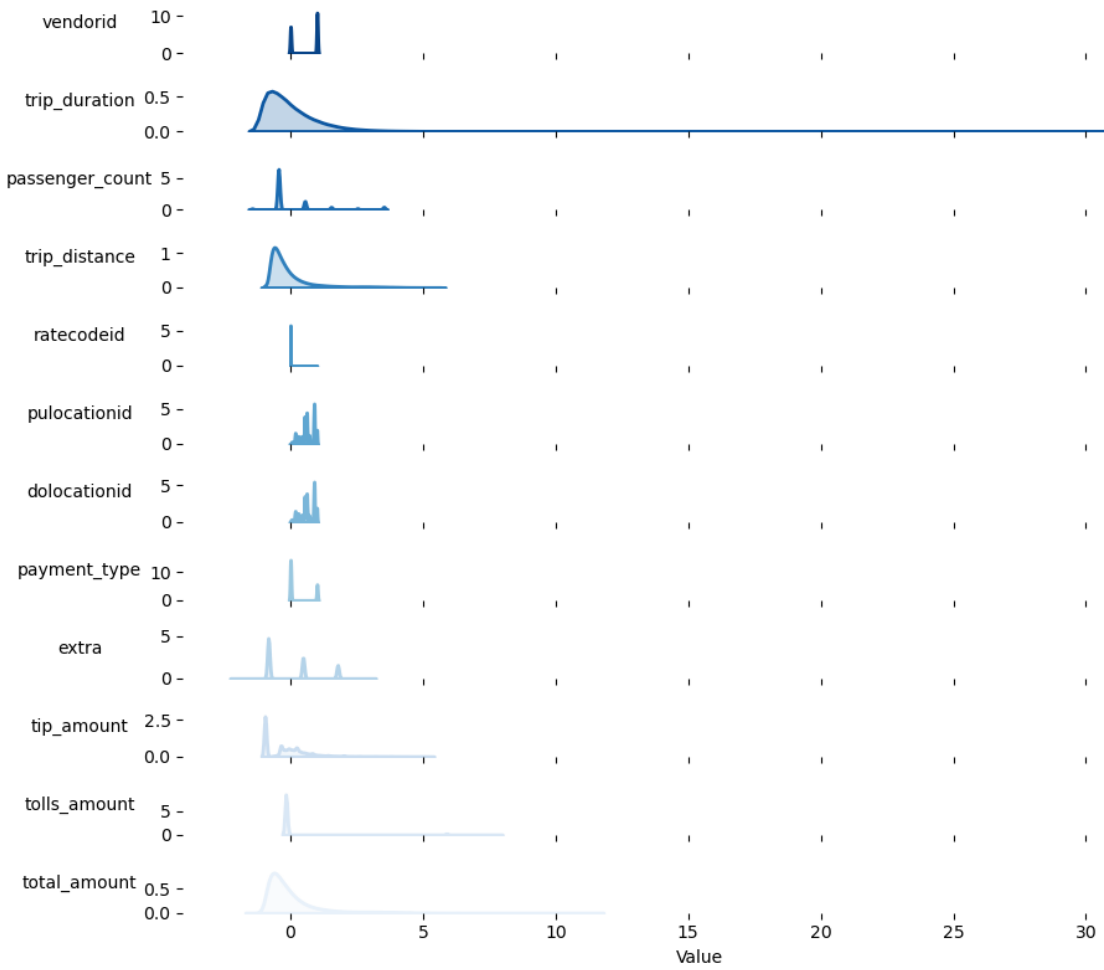


Fig. 7. Ridge Plot

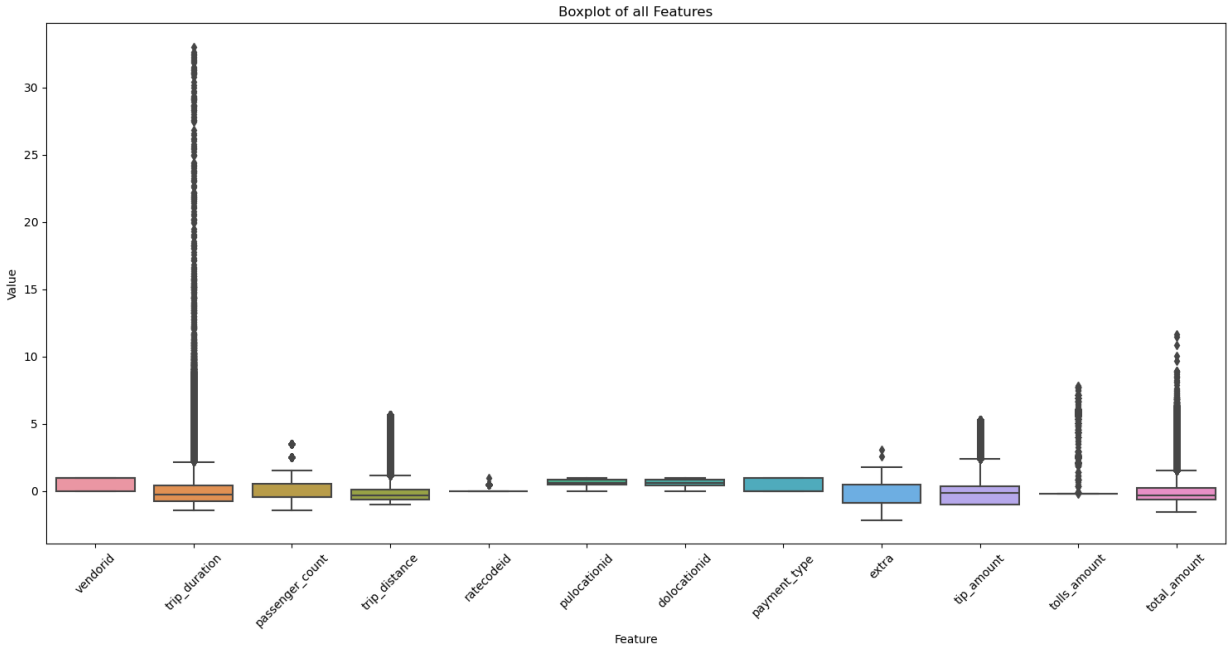


Fig. 8. Box Plot


Fig. 9. Pair Plot

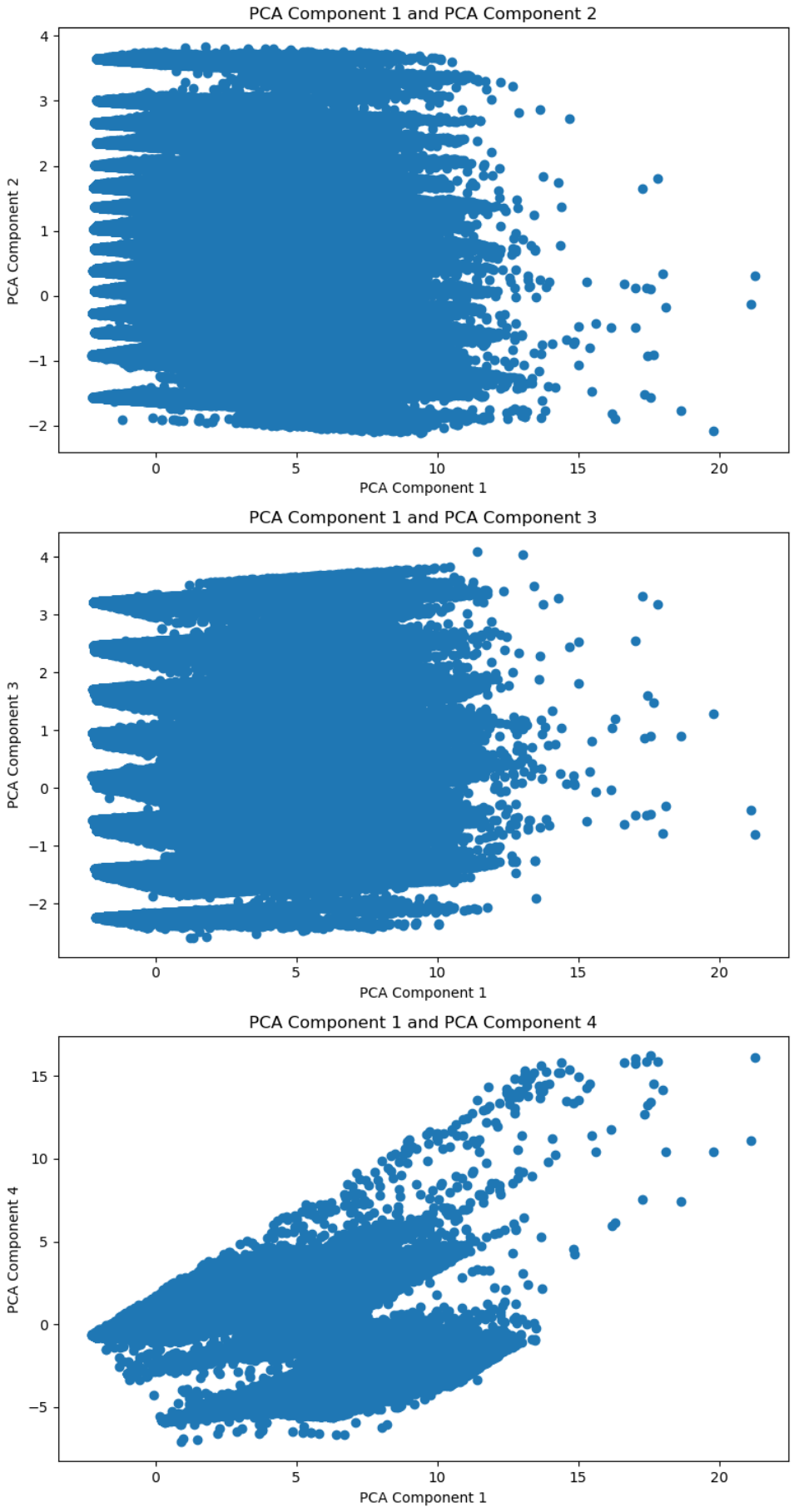


Fig. 10. Scatter Plot of the Principal Components

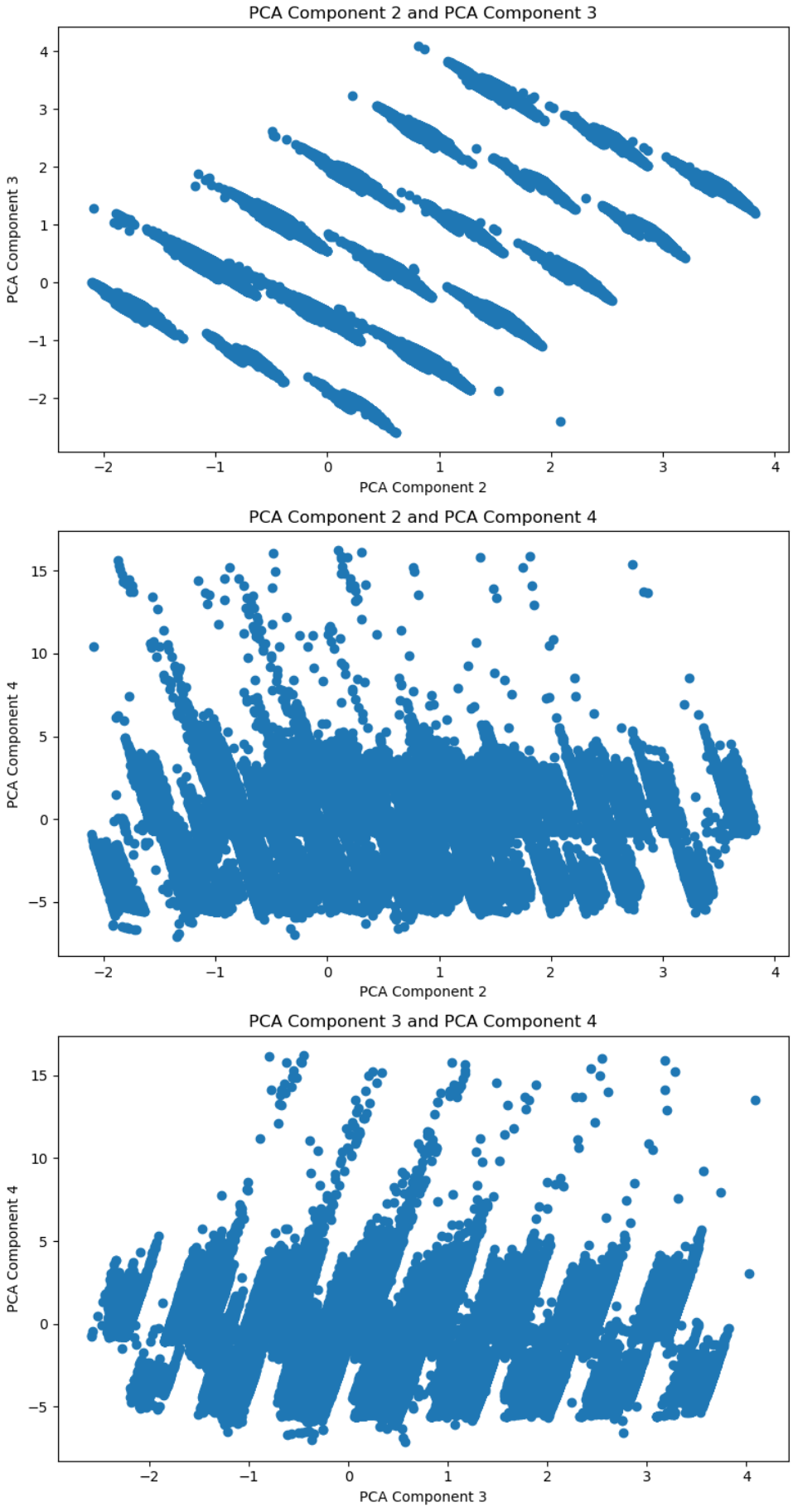


Fig. 11. Scatter Plot of the Principal Components

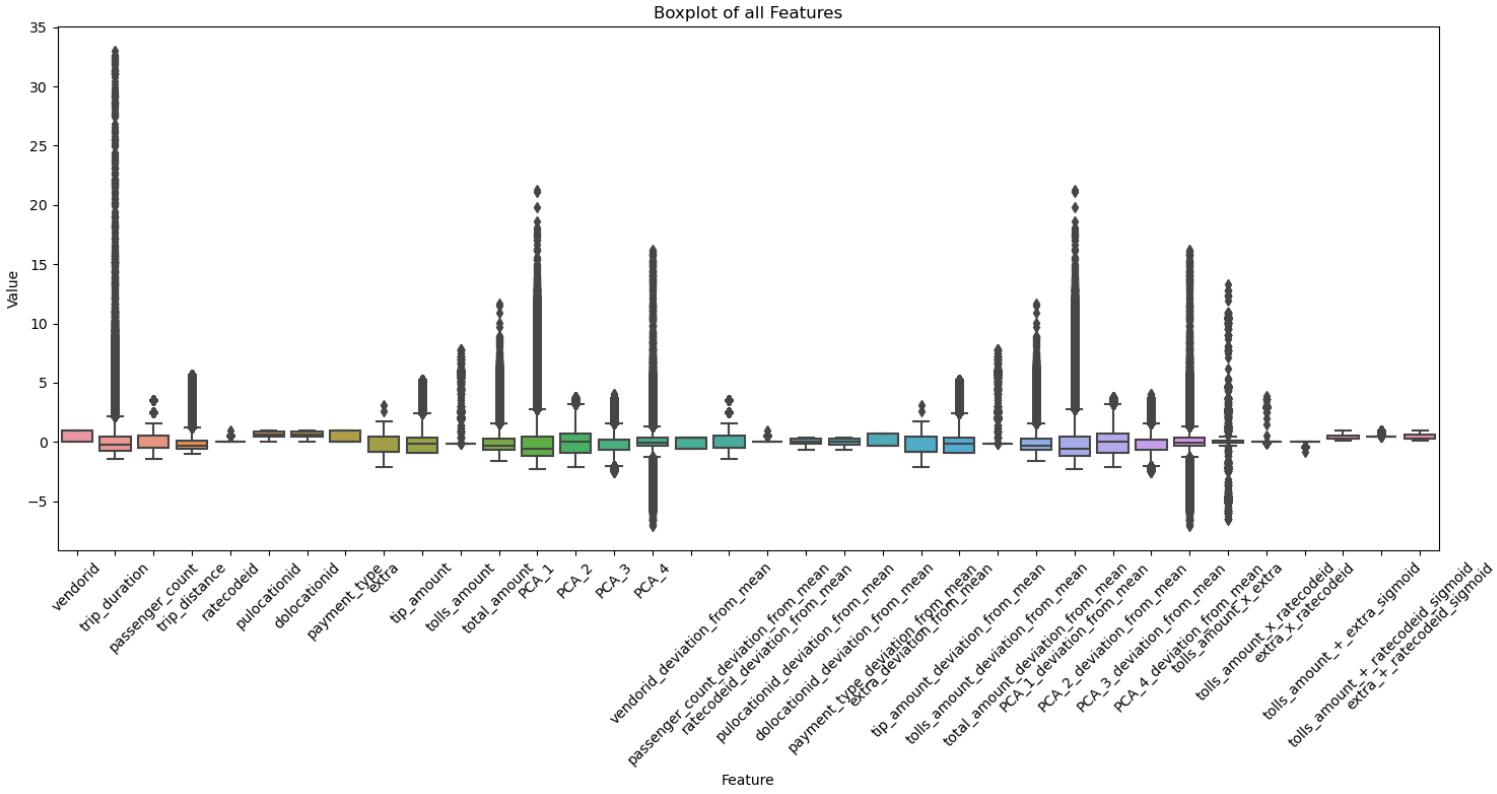


Fig. 12. Boxplot of All Features