Contents

[Introduction 1](#_Toc158384946)

[Data Preparation 1](#_Toc158384947)

Dataset Characterization [1](#_Toc158384948)

Data Cleaning/Preprocessing [3](#_Toc158384948)

Exploratory Data Analysis (EDA) [4](#_Toc158384948)

Encoding, Scaling, Engineering [12](#_Toc158384948)

Dimensional Reduction [12](#_Toc158384948)

[Statistical Techniques 14](#_Toc158384946)

Descriptive Statistics [14](#_Toc158384948)

Confidence Intervals [14](#_Toc158384948)

Hypothesis Testing [14](#_Toc158384948)

Correlation/Regression [15](#_Toc158384948)

[Machine Learning 18](#_Toc158384946)

ML Approaches Overview [18](#_Toc158384948)

Feature Selection/Hyperparameters [19](#_Toc158384948)

Model Training/Validation [20](#_Toc158384948)

Comparison/Discussion [21](#_Toc158384948)

[Conclusion 2](#_Toc158384950)3

[References 2](#_Toc158384950)4

# Introduction

The ability to leverage data analytics has become critical for organizations seeking to enhance decision-making and operational efficiency. This report focuses on analysing a dataset from a global bike rental company operating in major cities, including New York, Berlin, and London. The dataset captures variables such as rider demographics, ride characteristics, environmental conditions, and customer satisfaction, offering a rich source for uncovering patterns and insights.

The primary objective is to prepare and analyse the dataset to achieve key business goals: optimizing pricing strategies, enhancing customer satisfaction, and improving fleet management. The report is structured into three main areas: data preparation, statistical analysis, and machine learning. Through a structured approach, this analysis demonstrates how data-driven insights can address complex business challenges and inform strategic decisions for the bike rental company.

# Data Preparation

## Dataset Characterization

The dataset contains 900 records from a global bike rental company, operating in cities such as Berlin, London, New York, Chicago, San Francisco, and Dublin. This dataset provides a mix of numerical variables (e.g., ride duration, distance covered) and categorical variables (e.g., city, weather condition, bike model). The diversity of these variables makes the dataset a valuable resource for uncovering meaningful patterns and insights.

* **Dataset Overview:**
  + **Rows:** 900
  + **Columns:** 8
  + **Variable Types:**
    - **Numerical (4):** Rider Age, Ride Duration (minutes), Distance Covered (km), Bike Rental Cost ($)
    - **Categorical (4):** City, Weather Condition, Bike Model, Rider Satisfaction
* **Missing Values:**
  + Missing values were identified across all variables, as follows:
    - City: 10 missing values
    - Weather Condition: 11 missing values
    - Rider Age: 15 missing values
    - Bike Model: 10 missing values
    - Ride Duration: 13 missing values
    - Distance Covered: 13 missing values
    - Rider Satisfaction: 11 missing values
    - Bike Rental Cost: 9 missing values
* **Descriptive Statistics:**
  + **Rider Age:** Mean = 39.26 years; Range = 18–60 years
  + **Ride Duration (minutes):** Mean = 62.23 minutes; Range = 5–120 minutes
  + **Distance Covered (km):** Mean = 20.13 km; Range = 0.51–39.87 km
  + **Bike Rental Cost ($):** Mean = $26.58; Range = $7.33–$48.38

**Initial Insights from Categorical Variables**

* **City Distribution:** Most rentals occurred in London and Berlin, while Dublin had significantly fewer entries.
* **Weather Conditions:** Rentals were evenly distributed across sunny, rainy, snowy, and cloudy days, with slightly more rides recorded on rainy days.
* **Bike Models:** Rentals were distributed among Mountain, Standard, and Electric bikes, but the Manual bike had only one entry, likely a data entry error.
* **Rider Satisfaction:** Most customers reported medium or high satisfaction, with fewer entries for low satisfaction.

## Data Cleaning/Preprocessing

**Handling Missing Values**

The dataset exhibited missing values across both numerical and categorical columns. Specific strategies were employed to address these gaps:

* **Numerical Columns:** Mean imputation was applied to fill missing values in columns such as rider age, ride duration, distance covered, and bike rental cost. This method retained the dataset's overall statistical integrity while ensuring no loss of numerical information.
* **Categorical Columns:** Mode imputation was used for columns like city, weather condition, bike model, and rider satisfaction, as the mode reflects the most common and representative category.

**Outlier Detection and Removal**

Outliers in numerical variables were identified and removed using the Interquartile Range (IQR) method:

* **IQR Calculation:**
  + Formula: IQR = Q3 - Q1 (where Q1 is the 25th percentile and Q3 is the 75th percentile).
  + Threshold: Observations outside [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] were considered outliers.
* **Impact on Dataset:**
  + After outlier removal, the dataset was reduced from 900 rows to 763 rows. This process enhanced data quality by eliminating extreme values that could skew analysis.

**Column Renaming**

Column names were standardized to improve clarity and ease of use in analysis:

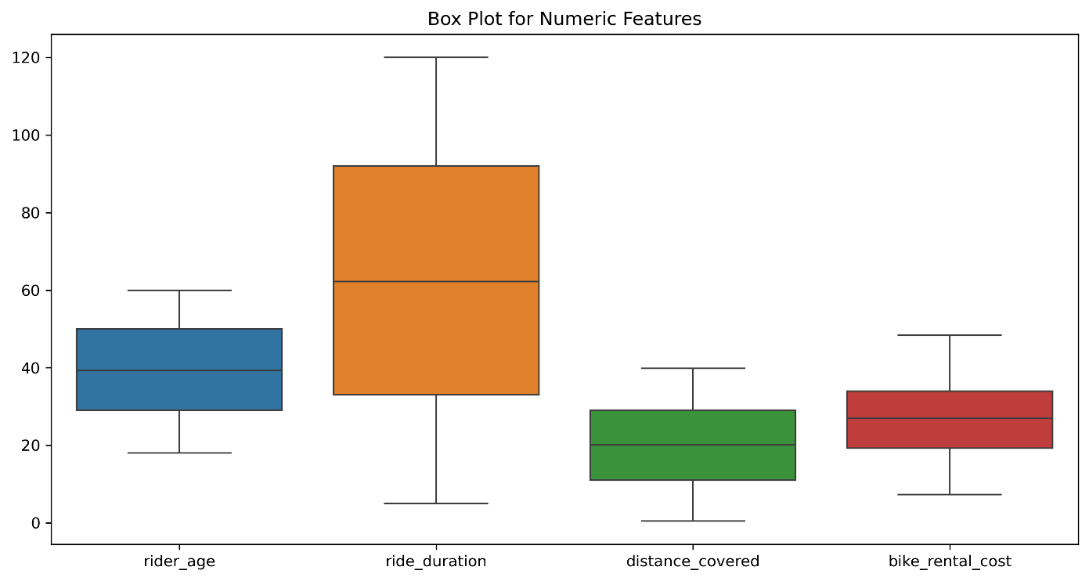
* **Old Format:** Names like "Ride Duration" and "Bike Rental Cost" had inconsistent capitalization and spacing.
* **New Format:** Columns were renamed to lowercase with underscores, e.g., 'ride\_duration' and 'bike\_rental\_cost.'

This formatting ensures compatibility with programming tools and enhances readability during data manipulation.

## Exploratory Data Analysis (EDA)

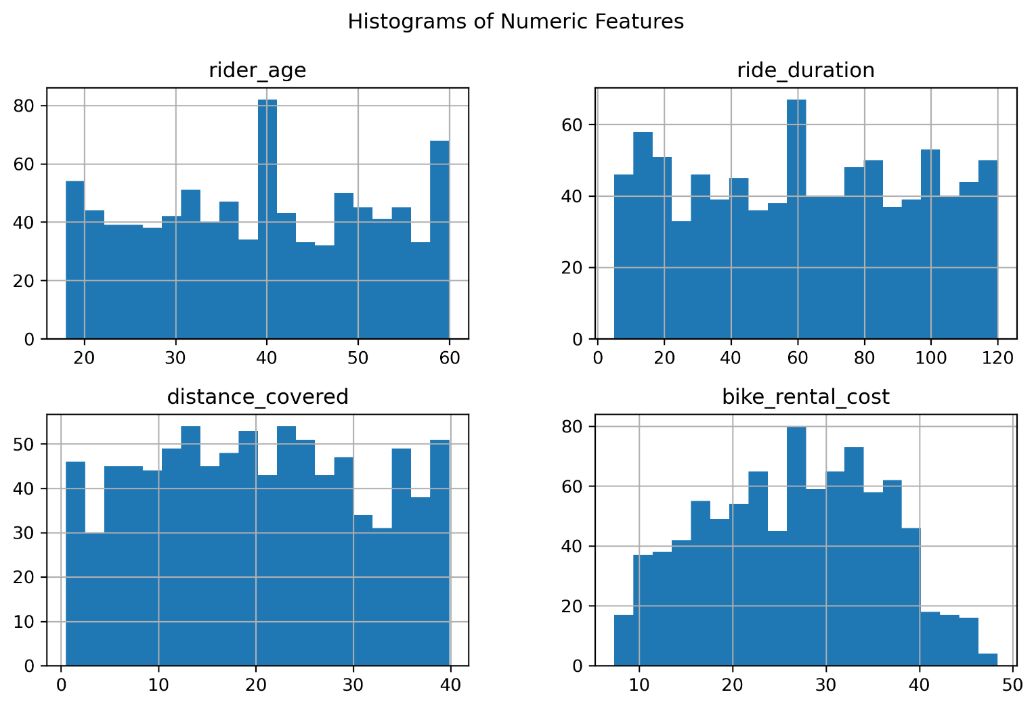
EDA provided critical insights into the dataset’s structure and relationships between variables:

**Numerical Variables**



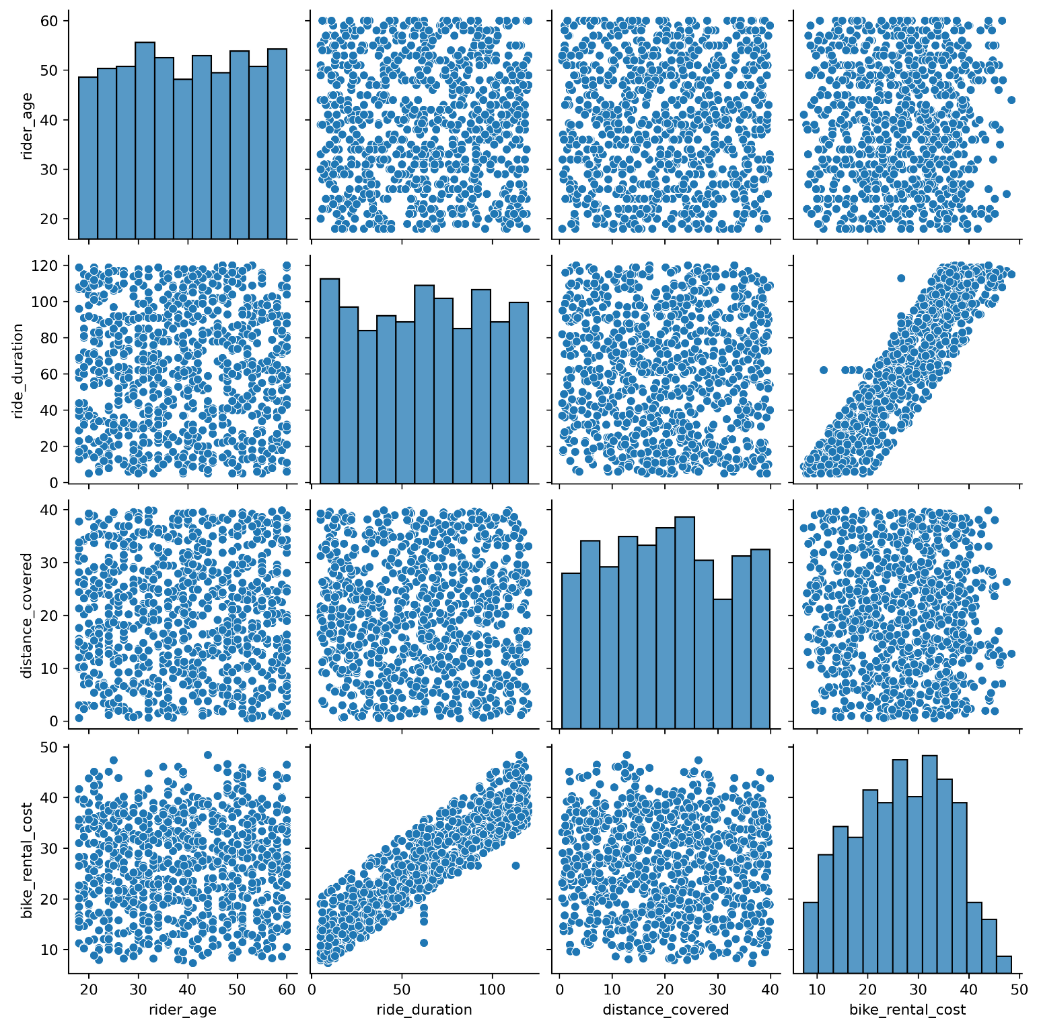
These box plots reveal the distribution and range of the numerical variables (rider\_age, ride\_duration, distance\_covered, and bike\_rental\_cost):

* ride\_duration exhibits a wider range and greater variability compared to the other features.
* No significant outliers are visible after applying the IQR-based outlier removal.



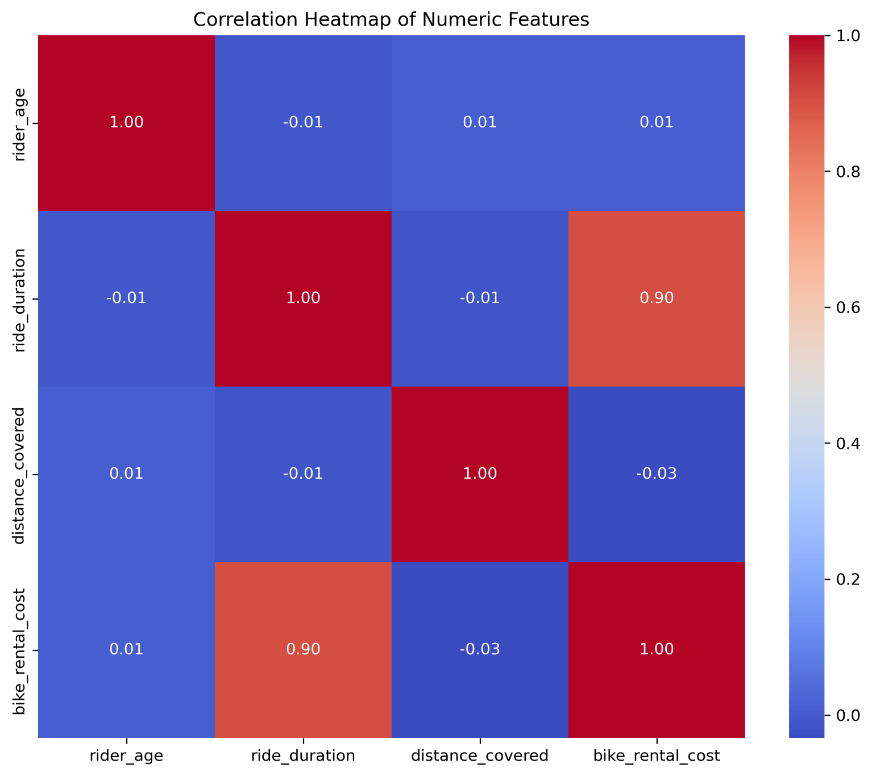
Histograms provide insight into the frequency distribution of the numerical features:

* Ride duration showed a peak at 60 minutes, reflecting standard rental periods.
* Distance covered and bike rental cost exhibited right-skewed distributions, with most values concentrated around typical averages.



The pair plots highlight the relationships between numerical features:

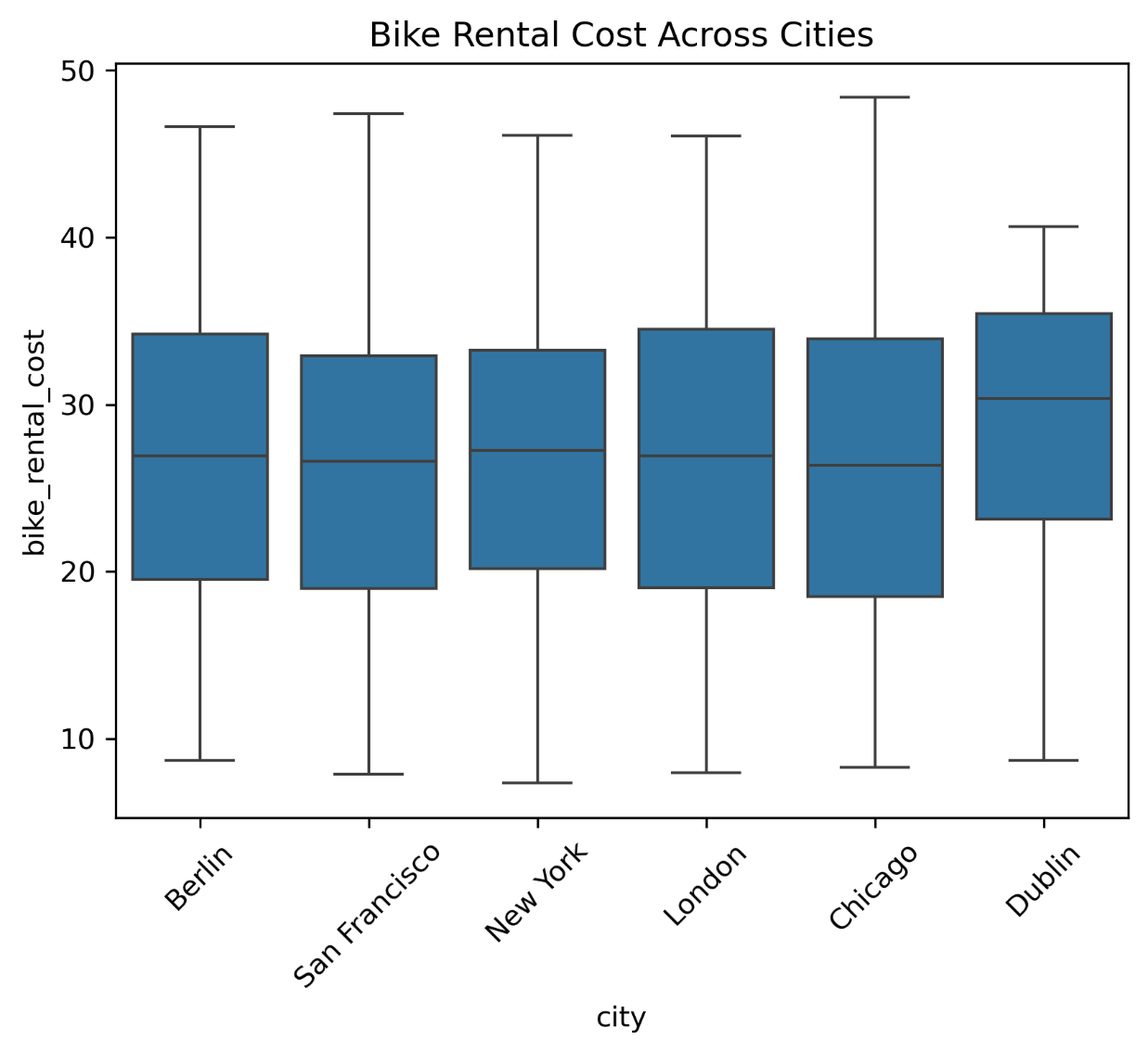
* **ride\_duration and bike\_rental\_cost**: A strong positive correlation is visible, indicating that longer rides generally cost more.
* Other numerical features (e.g., rider\_age and distance\_covered) do not exhibit strong pairwise relationships.



The heatmap quantitatively represents correlations:

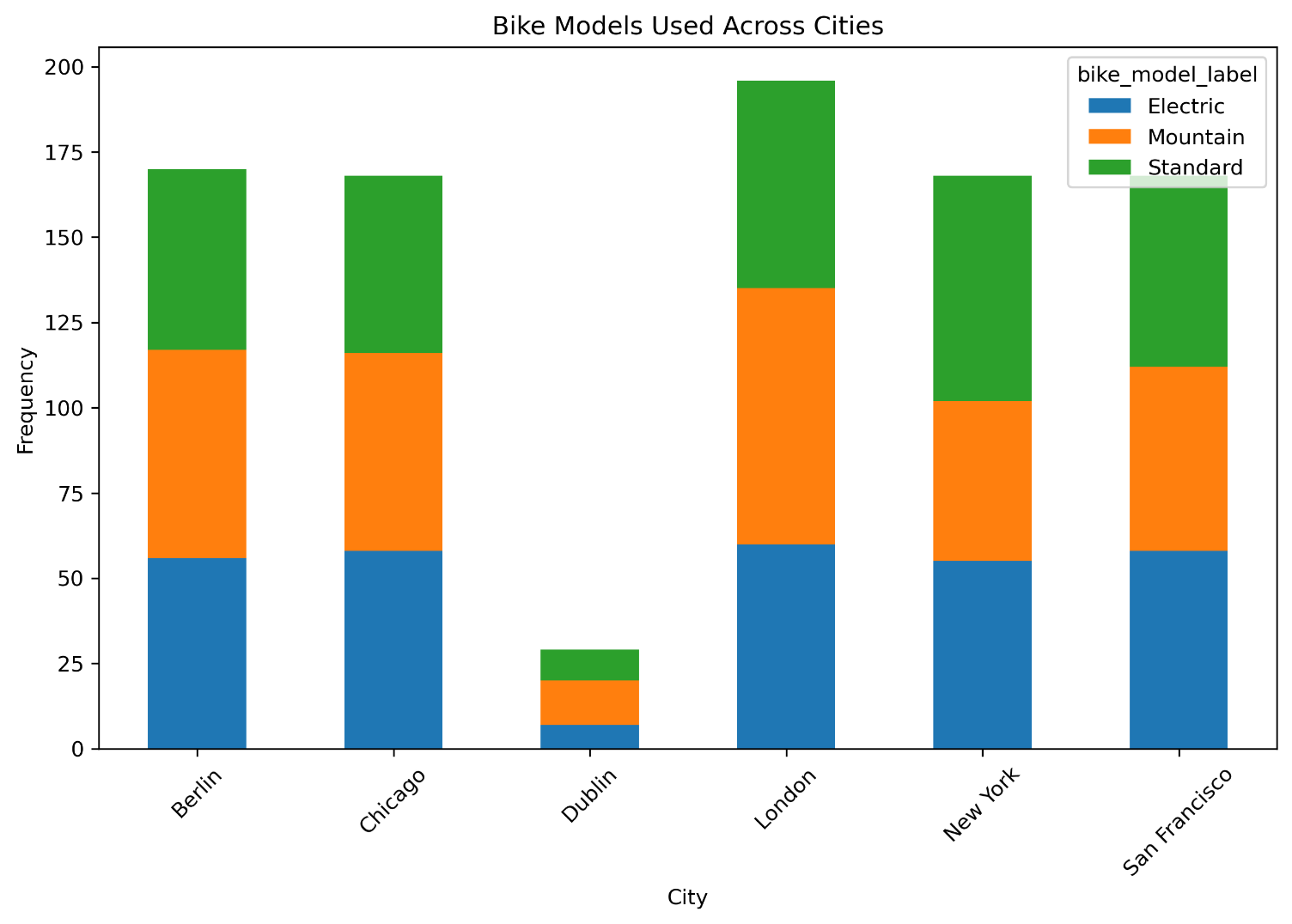
* **ride\_duration and bike\_rental\_cost** have a high correlation (0.90), confirming the trend observed in the pair plot.

**Categorical Variables**



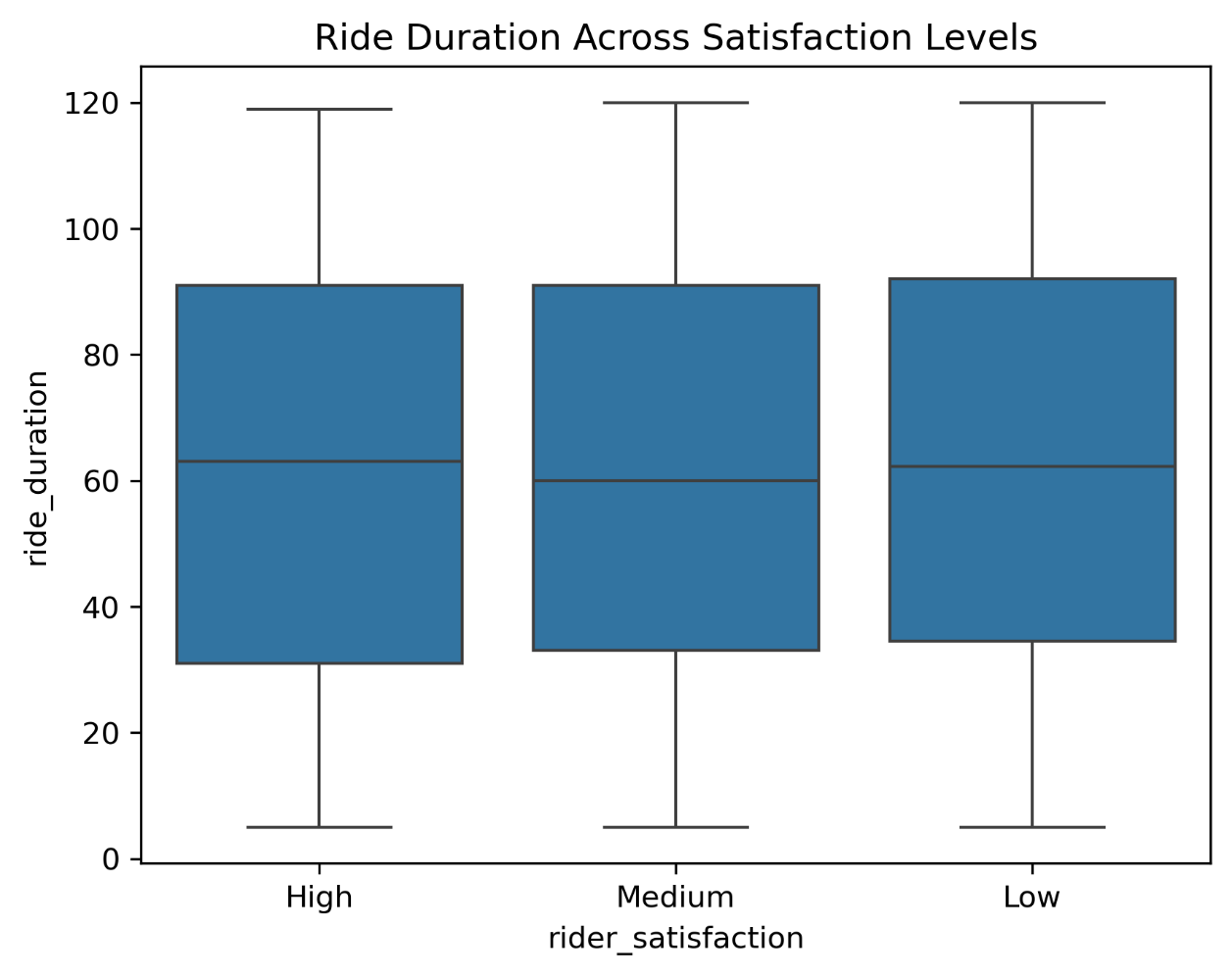
These box plots illustrate the variation in bike rental costs across cities:

* The median rental cost is similar across most cities, except for Dublin, which has slightly higher costs. This could be partly due to the smaller sample size.
* Variability in costs is consistent, with no city showing extreme outliers.



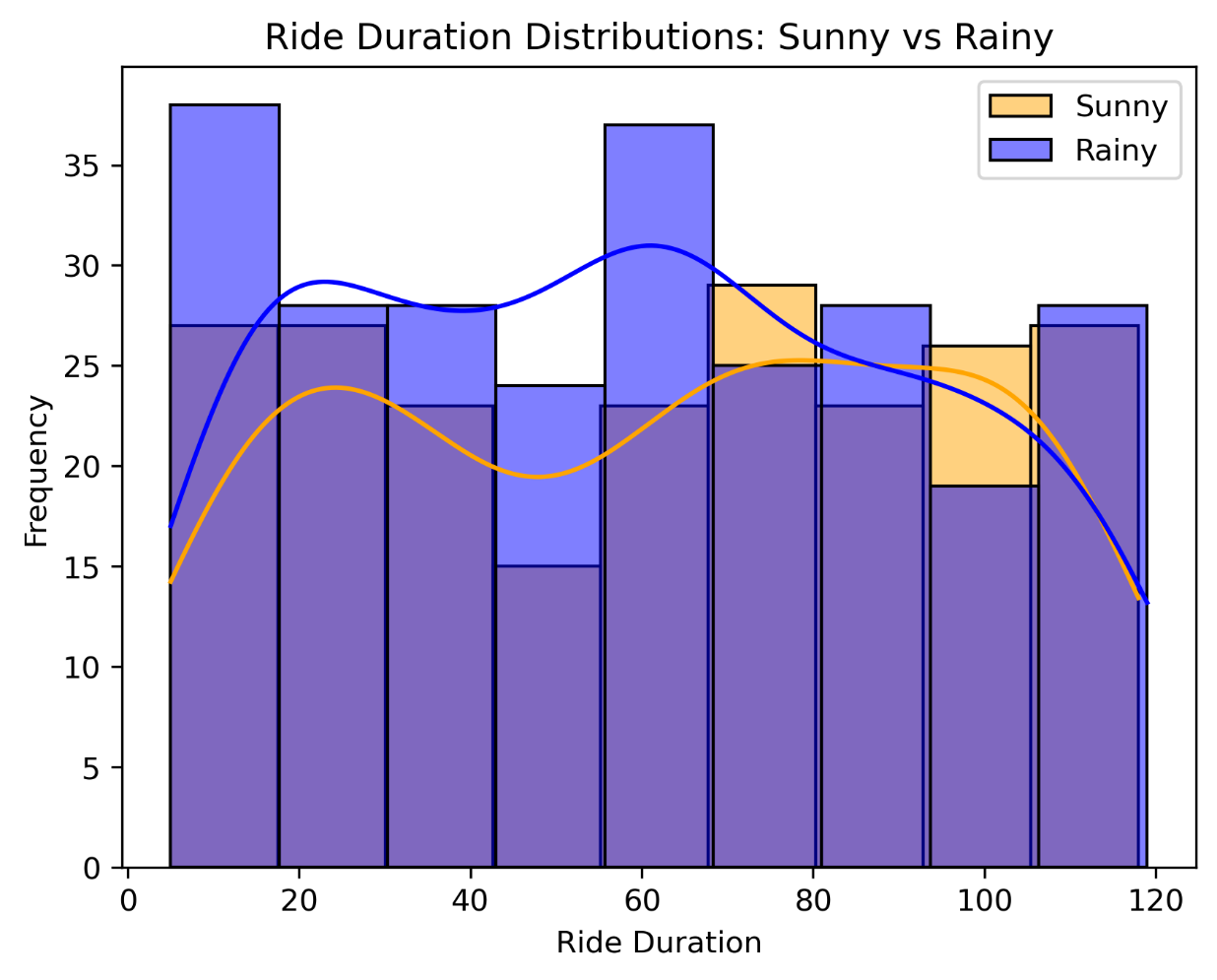
The stacked bar chart highlights the distribution of bike models in each city:

* Standard, Mountain, and Electric bikes are evenly distributed across Berlin, Chicago, London, New York, and San Francisco.
* Dublin stands out due to its significantly lower rental counts, which again can be explained by its relatively small sample size.



These plots compare ride durations across satisfaction levels:

* There is no significant difference in median ride duration for High, Medium, and Low satisfaction levels.
* Variability in ride durations is consistent, suggesting that factors beyond duration influence rider satisfaction.



This overlaid histogram compares ride duration distributions under sunny and rainy conditions:

* **Sunny days:** Ride durations show a peak around 20 minutes, with a gradual decline for longer rides.
* **Rainy days:** The distribution is more uniform, with a slight increase in shorter durations.
* This suggests that weather conditions may slightly influence ride duration patterns, with shorter rides being more common in the rain.

## Encoding, Scaling, Engineering

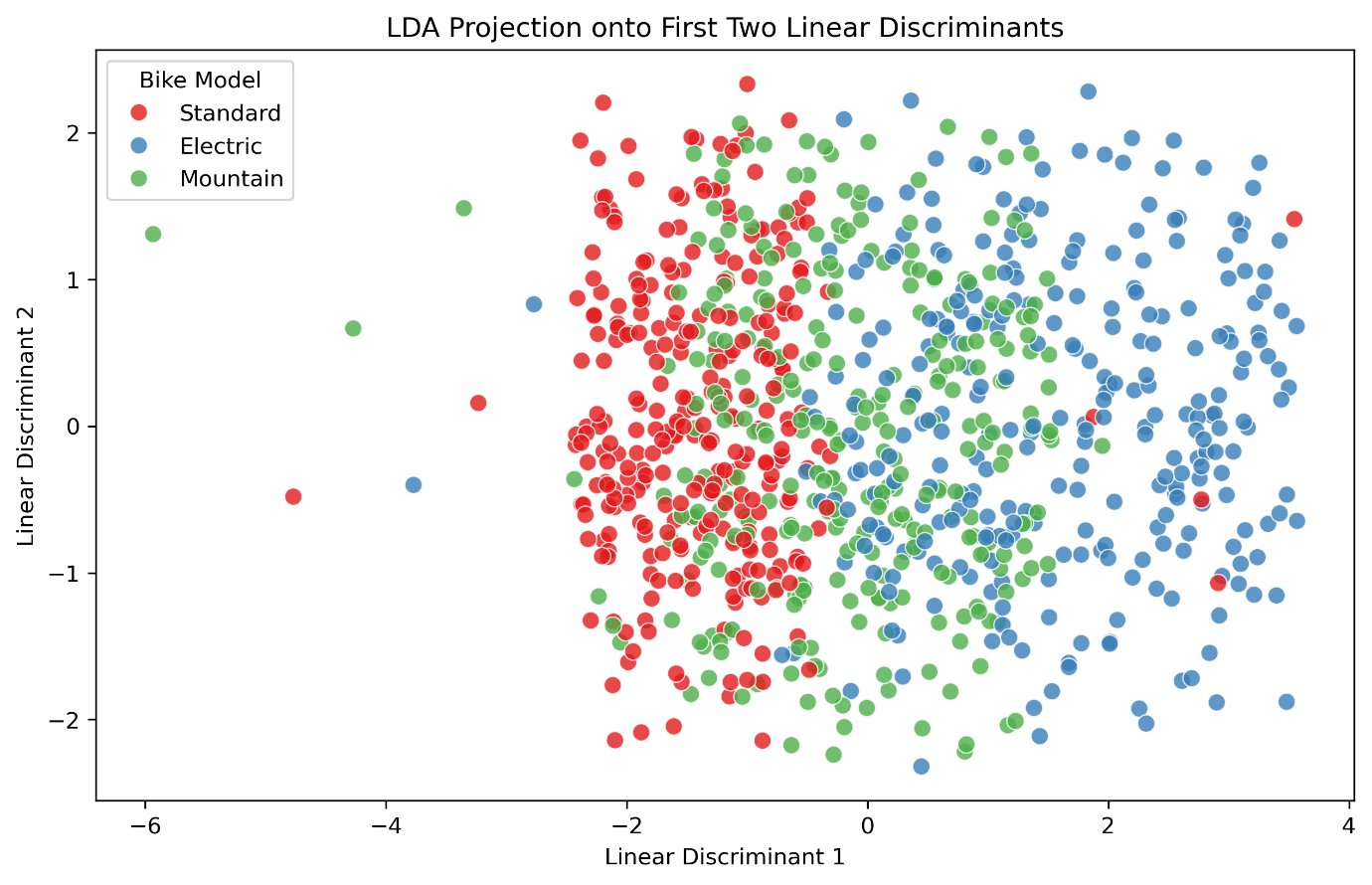
To prepare the data for machine learning models, additional processing steps were undertaken:

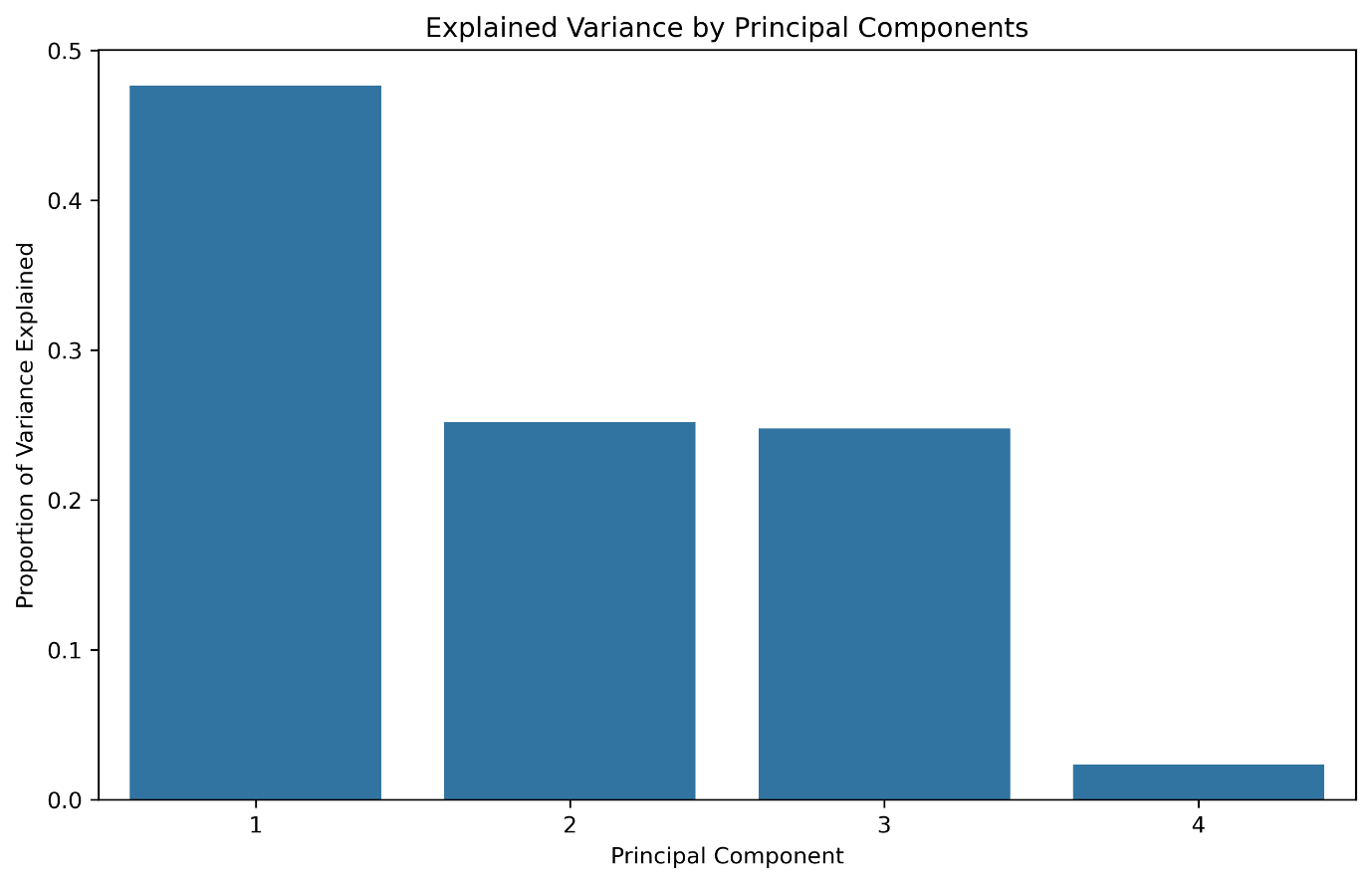
* **Encoding Categorical Variables:**
  + One-hot encoding transformed categorical variables, such as city and weather condition, into binary columns.
  + For example, a ‘city’ variable with values (e.g., London, Berlin) became separate columns, each representing a binary indicator.
* **Scaling Numerical Variables:**
  + Z-score normalization was applied to numerical features like ride duration and rental cost, ensuring consistent scales and improving model performance.
* **Feature Engineering:**
  + Interaction terms were created, such as the interaction between weather and satisfaction levels, to capture non-linear relationships.

## Dimensionality Reduction

Dimensionality reduction techniques were applied to simplify the dataset and improve interpretability:

* **Principal Component Analysis (PCA):**
  + PCA transformed the numerical data into three principal components, explaining 98% of the variance.
  + The first component emphasized rider age and distance covered, while the second and third highlighted ride duration and rental cost.
  + Visualizations revealed clustering by rider satisfaction categories.
* **Linear Discriminant Analysis (LDA):**
  + LDA optimized class separation for rider satisfaction levels.
  + Projections showed clear separations between high, medium, and low satisfaction groups, aiding classification tasks.





# Statistical Techniques

## Descriptive Statistics

Descriptive statistics summarize the dataset's main characteristics, including central tendency, variability, and probabilities:

* **Central Tendency and Variability:**
  + Ride Duration: Mean = 62.23 min, Std = 33.71 min, Range = [5, 120].
  + Distance Covered: Mean = 20.13 km, Std = 11.10 km, Range = [0.51, 39.87].
  + Bike Rental Cost: Mean = $26.58, Std = $9.39, Range = [$7.33, $48.38].
* **Normality Test:**
  + A Shapiro-Wilk test revealed that ride\_duration and bike\_rental\_cost do not follow a normal distribution (p < 0.05).

The graphs provided in the earlier EDA section provide visualizations for these statistics.

## Confidence Intervals

Confidence intervals were calculated for two comparisons:

1. **Sunny vs. Rainy Ride Duration:**
   * Sunny: Mean = 61.92 min, 95% CI = [57.40, 66.44].
   * Rainy: Mean = 59.36 min, 95% CI = [55.28, 63.44].
   * Interpretation: Overlapping intervals suggest no significant difference in ride duration between sunny and rainy conditions.
2. **Electric vs. Standard Bike Rental Cost:**
   * Electric: Mean = $31.00, 95% CI = [$30.03, $31.97].
   * Standard: Mean = $22.17, 95% CI = [$21.15, $23.19].
   * Interpretation: Non-overlapping intervals confirm a significant difference in rental costs.

## Hypothesis Testing

Several hypothesis tests were conducted to explore relationships within the data. The following are three examples of results from these tests:

1. **T-Test (Sunny vs. Rainy Ride Duration):**
   * t-statistic = 0.82, p = 0.41.
   * Result: Fail to reject the null hypothesis; no significant difference in ride duration.
2. **ANOVA (Bike Rental Cost Across Satisfaction Levels):**
   * F-statistic = 0.04, p = 0.96.
   * Result: Fail to reject the null hypothesis; no significant difference in rental cost.
3. **Chi-Square Test (Bike Model vs. Satisfaction Levels):**
   * χ² = 3.57, p = 0.47.
   * Result: Fail to reject the null hypothesis; no significant association between bike model and satisfaction.

## Correlation/Regression

**Correlation Analysis**

Correlation analysis was performed to investigate relationships between numerical features, including ride\_duration, distance\_covered, and bike\_rental\_cost. Key findings include:

* **Strong Positive Correlation:**
  + Ride Duration and Bike Rental Cost: r = 0.90, indicating a strong linear relationship where longer rides generally result in higher rental costs.
* **Weak Correlations:**
  + Other numerical features, such as rider\_age, showed minimal correlations with bike\_rental\_cost, suggesting limited predictive power.

The correlation heatmap above confirmed these relationships, with ride\_duration being the dominant numerical predictor of bike\_rental\_cost.

**Linear Regression**

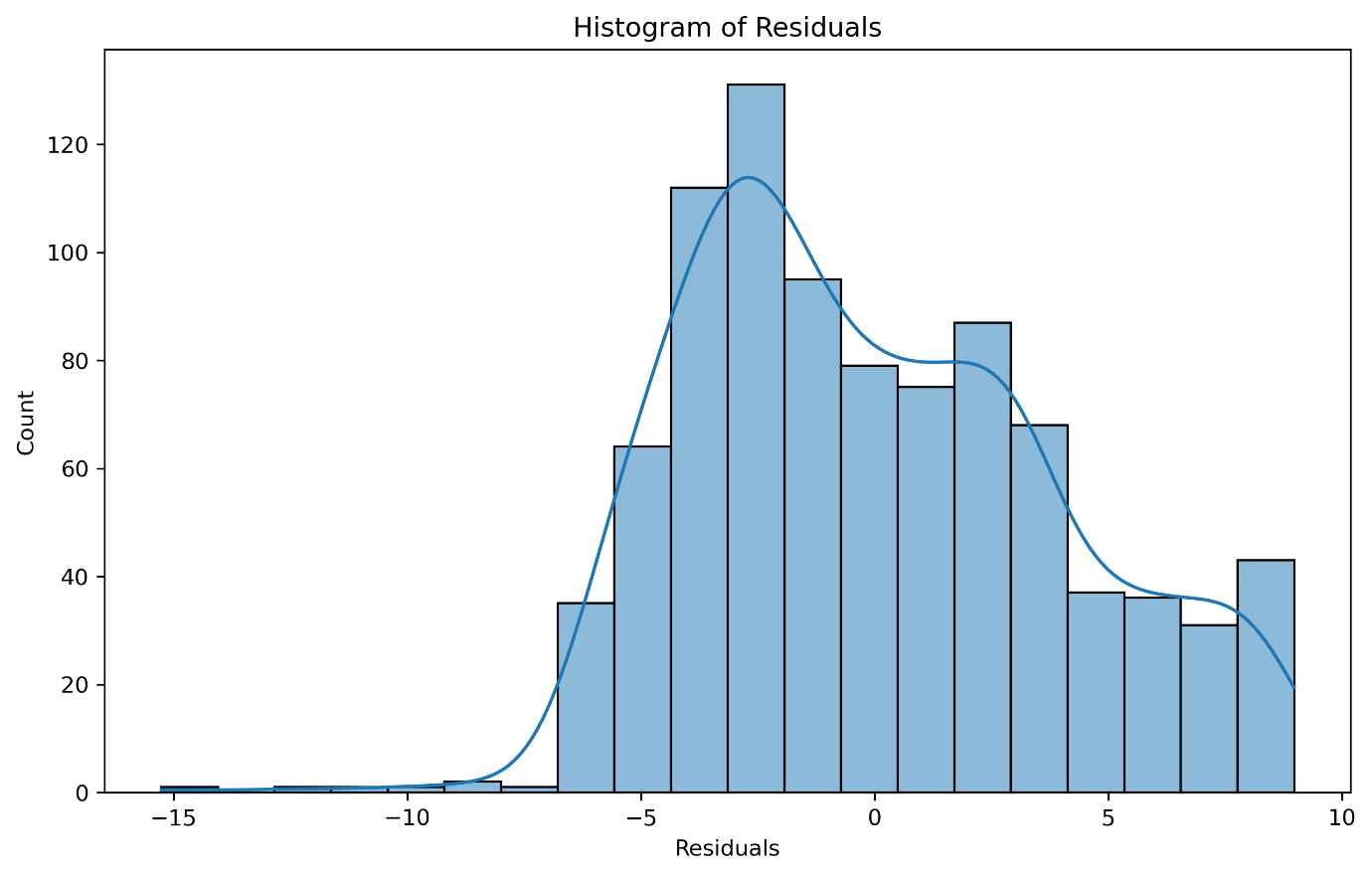
A simple linear regression model was constructed to quantify the relationship between ride\_duration (independent variable) and bike\_rental\_cost (dependent variable).

* **Model Summary:**
  + **R-squared:** 0.819, indicating that 81.9% of the variance in bike\_rental\_cost is explained by ride\_duration.
  + **Adjusted R-squared:** 0.819, confirming the model's robustness with a single predictor.
  + **F-statistic:** 4058, p < 0.001, highlighting the model's statistical significance.
* **Coefficients:**
  + **Intercept (const):** 10.90 (p < 0.001). This represents the baseline rental cost for a ride of zero duration.
  + **Ride Duration:** 0.2521 (p < 0.001). For every additional minute of ride duration, the rental cost increases by $0.252.
* **Statistical Tests:**
  + **Durbin-Watson:** 2.021, indicating no significant autocorrelation in residuals.
  + **Omnibus Test (p = 0.000):** Residuals deviate slightly from normality, which may affect the model's predictive accuracy for extreme values.

**Residual Analysis**

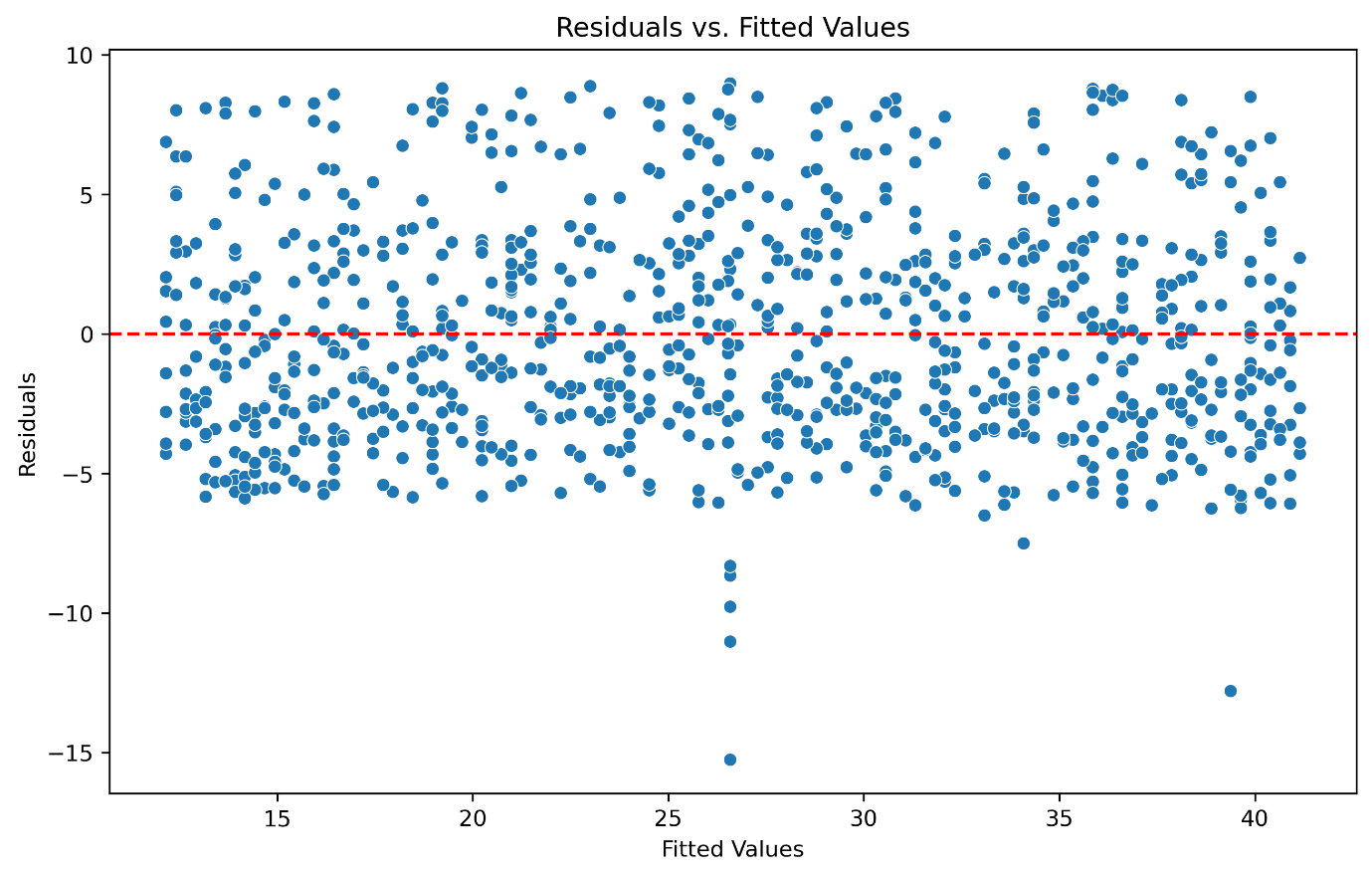
Residual analysis was conducted to evaluate the performance and validity of the regression model predicting bike\_rental\_cost using ride\_duration.

1. **Histogram of Residuals**

****

* The histogram shows the distribution of residuals, which represent the difference between observed and predicted values.
* **Findings:**
  + The residuals are roughly centred around 0, with a slight skew toward negative values.
  + A minor deviation from normality is observed, as indicated by the Omnibus test in the regression summary (p < 0.001).

1. **Residuals vs. Fitted Values**

****

The residuals vs. fitted values plot evaluates whether the residuals are randomly distributed:

**Findings:**

* The residuals scatter randomly around 0, indicating no major issues with heteroscedasticity (variance inconsistency).
* A few extreme outliers are present, suggesting potential leverage points or influential observations.

**3. Extreme Residuals**

Observations with the largest residuals were examined:

**Top 5 rows with extreme residuals:**

ride\_duration bike\_rental\_cost residuals

* 159 62.23 11.32 -15.26
* 630 113.00 26.58 -12.80
* 664 62.23 15.55 -11.03
* 607 62.23 16.81 -9.77
* 295 62.23 35.56 8.98

These rows indicate mismatches where the observed bike\_rental\_cost deviates significantly from predictions. This could suggest data entry issues or unmodeled variables affecting the rental cost.

**Key Insights**

1. Strong Predictor: Ride\_duration is the strongest predictor of bike\_rental\_cost, contributing significantly to the model's high R² value.
2. Cost Impact: The regression coefficient indicates a linear cost increase, providing a clear pricing structure for longer rides.
3. Model Strength: While the model explains a significant portion of the variance, deviations in residual normality should be addressed for more complex models.

The regression analysis highlights the critical role of ride duration in determining bike rental costs, aligning with business expectations.

# Machine Learning

## ML Approaches Overview

The dataset provided includes a mix of numerical and categorical features, with a defined target variable for prediction tasks. Thus, a supervised learning approach was the most logical choice for the primary analysis. Supervised learning is particularly well-suited to this dataset, as the labelled target variables (e.g., bike rental cost, electric bike usage) allow for effective regression and classification tasks. Additionally, unsupervised learning techniques, such as dimensionality reduction using Principal Component Analysis (PCA), were employed to enhance exploratory analysis and uncover hidden patterns.

**Supervised Learning**

Supervised learning was applied in two main areas:

1. **Regression Tasks:**
   * Models predicted continuous variables such as bike rental cost using independent features.
   * Metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) evaluated model performance.
2. **Classification Tasks:**
   * Models predicted whether an electric bike was used during a ride (binary classification).
   * Metrics such as accuracy, ROC-AUC, and F1-score assessed classification models.

**Unsupervised Learning**

Dimensionality reduction techniques, such as PCA and Linear Discriminant Analysis (LDA), were utilized to reduce data complexity, identify relationships between features, and enhance visualization. These methods proved instrumental in efficient data preprocessing and model training.

**Pros and Cons of ML Approaches**

* **Supervised Learning:**
  + **Pros:** Clear performance metrics, straightforward interpretation, and direct applicability to predictive tasks.
  + **Cons:** Requires labelled data, demanding preprocessing, and feature engineering.
* **Unsupervised Learning:**
  + **Pros:** Excellent for exploratory data analysis, dimensionality reduction, and visualization.
  + **Cons:** Limited interpretability and lack of direct optimization for prediction tasks.

**Independent and Dependent Variables**

* **Independent Variables:** Ride duration, distance covered, rider age, weather conditions (e.g., sunny, rainy, snowy), bike model types (e.g., electric, mountain, standard), and rental cost.
* **Dependent Variables:**
  + Regression: Bike rental cost.
  + Classification: Electric bike usage (binary variable).

## Feature Selection/Hyperparameters

**Feature Selection**

Feature selection was conducted using statistical analysis and model-based approaches to identify the most relevant predictors for regression and classification tasks. Significant predictors included:

* **Regression Models:** Ride duration, distance covered, rider age, and bike model types (e.g., electric, mountain, standard).
* **Classification Models:** Ride duration, distance covered, rider age, weather conditions, and bike rental cost.

Feature importance was assessed through:

* **Linear Regression Coefficients:** Identified the direct impact of independent variables on rental cost.
* **Random Forest Feature Importance:** Highlighted ride duration as the most influential feature for both regression and classification.
* **XGBoost Feature Importance:** Revealed bike rental cost and ride duration as dominant predictors in classifying electric bike usage.

**Hyperparameter Tuning**

To optimize model performance, hyperparameter tuning was conducted using GridSearchCV and cross-validation:

* **Random Forest Regression:**
  + Tuned parameters: Number of estimators, maximum depth, and minimum samples split.
  + Best parameters: n\_estimators=200, max\_depth=20, min\_samples\_split=10.
  + Result: Improved MAE (2.36) and RMSE (2.96).
* **XGBoost Classification:**
  + Tuned parameters: Number of estimators, learning rate, maximum depth, subsample, and column sampling by tree.
  + Best parameters: n\_estimators=200, learning\_rate=0.1, max\_depth=7, subsample=0.8, colsample\_bytree=1.0.
  + Result: High accuracy (0.79) and ROC-AUC score (0.86).

**Methodology**

* **GridSearchCV:** Exhaustively searched parameter ranges using 5-fold cross-validation to ensure robust evaluation.

## Model Training/Validation

**Training and Testing**

The dataset was split into training (80%) and testing (20%) subsets. Key steps included:

1. **Standardization:** All numerical features were scaled using StandardScaler for uniformity.
2. **Resampling:** Synthetic Minority Oversampling Technique (SMOTE) balanced classes in classification tasks.
3. **Model Training:** Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, and Gradient Boosting were trained on the processed data.
4. **Model Testing:** Predictions were generated for the test set, and metrics such as MAE, RMSE, accuracy, and ROC-AUC were computed.

**Cross-Validation**

5-fold cross-validation was used to validate models and prevent overfitting:

* **Linear Regression:** Average cross-validation score of 0.92.
* **Random Forest:** Consistent scores of 0.90 for regression and classification.
* **XGBoost:** Cross-validation ROC-AUC score of 0.86.
* **Stacking Classifier:** Achieved the highest ROC-AUC score of 0.89.

**Performance Metrics**

* **Regression Metrics:**
  + Mean Absolute Error (MAE): Average magnitude of errors.
  + Root Mean Squared Error (RMSE): Interpretable error measure in the same units as the target variable.
* **Classification Metrics:**
  + Accuracy: Percentage of correctly predicted instances.
  + ROC-AUC Score: Model’s ability to distinguish between classes.
  + F1-Score: Balances precision and recall for imbalanced datasets.

**Results**

* **Regression Tasks:**
  + Linear Regression performed best with MAE = 2.02 and RMSE = 2.53.
  + Optimized Random Forest Regression achieved MAE = 2.36 and RMSE = 2.96.
* **Classification Tasks:**
  + XGBoost achieved accuracy = 0.79 and ROC-AUC = 0.86.
  + Stacking Classifier outperformed with accuracy = 0.81 and ROC-AUC = 0.89.

## Comparison/Discussion

**Regression Models**

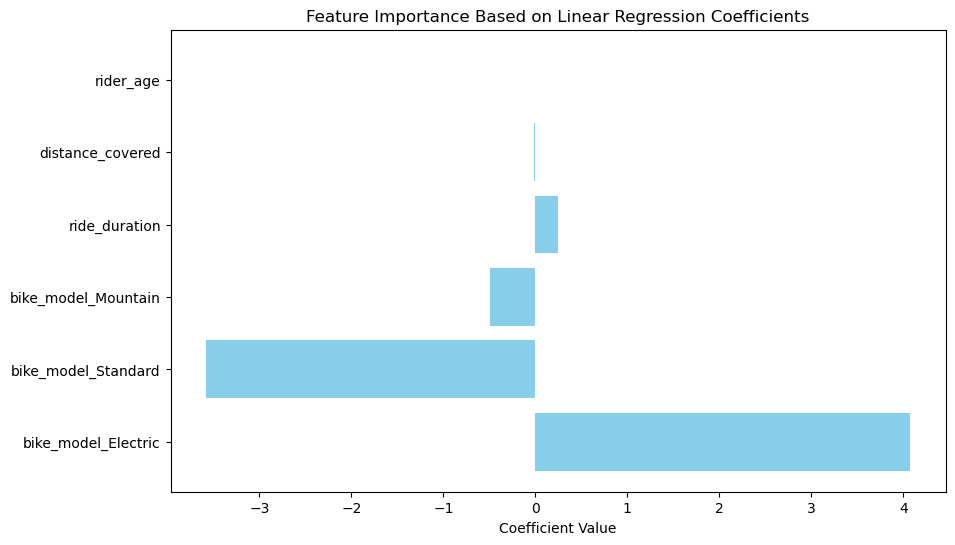
1. **Linear Regression:** Best performance with high accuracy and minimal residual variance.
2. **Ridge and Lasso Regression:** Comparable to Linear Regression, reduced the impact of less significant features.
3. **Random Forest:** Performed well but did not surpass Linear Regression for linear tasks.
4. **Optimized Random Forest:** Showed improved metrics after hyperparameter tuning.

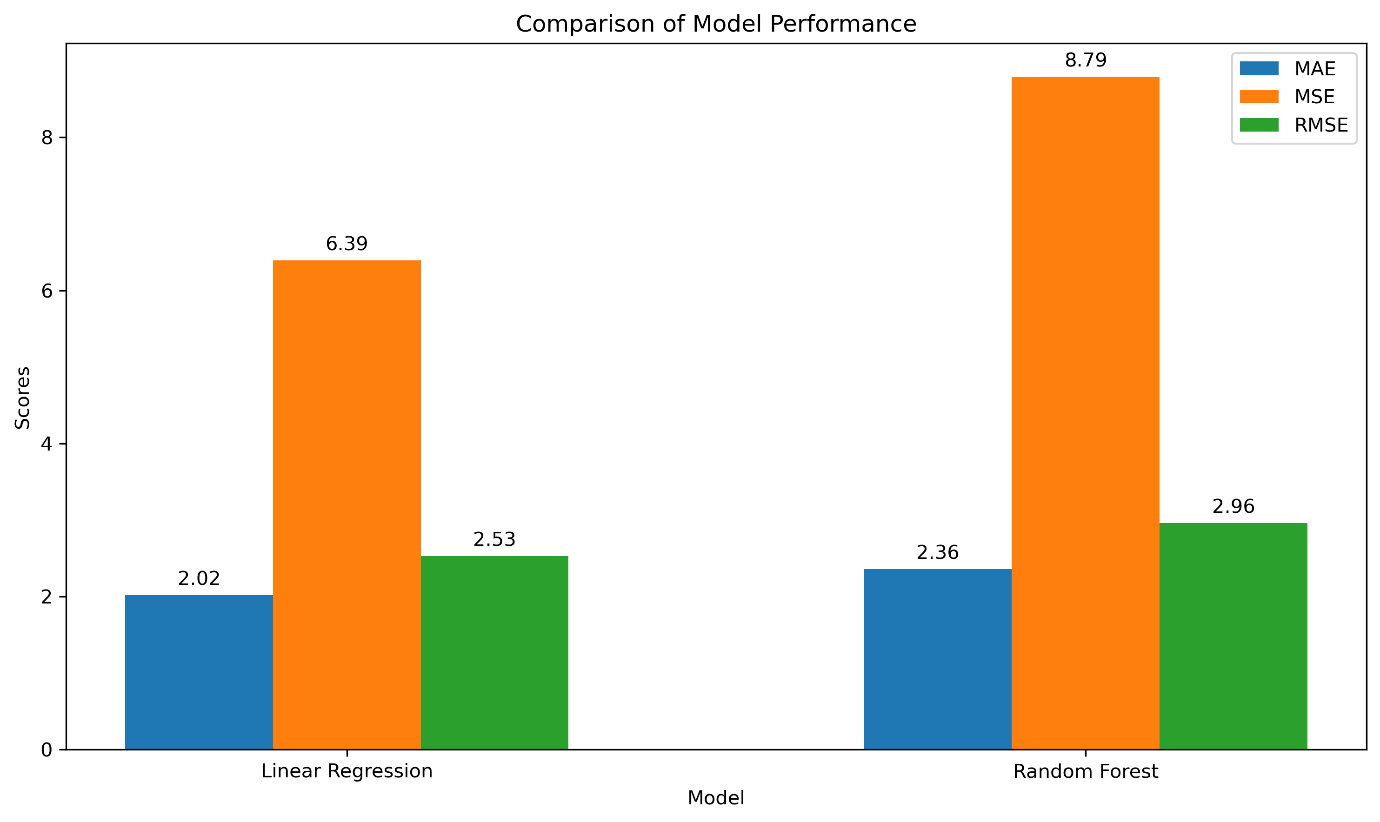
**Classification Models**

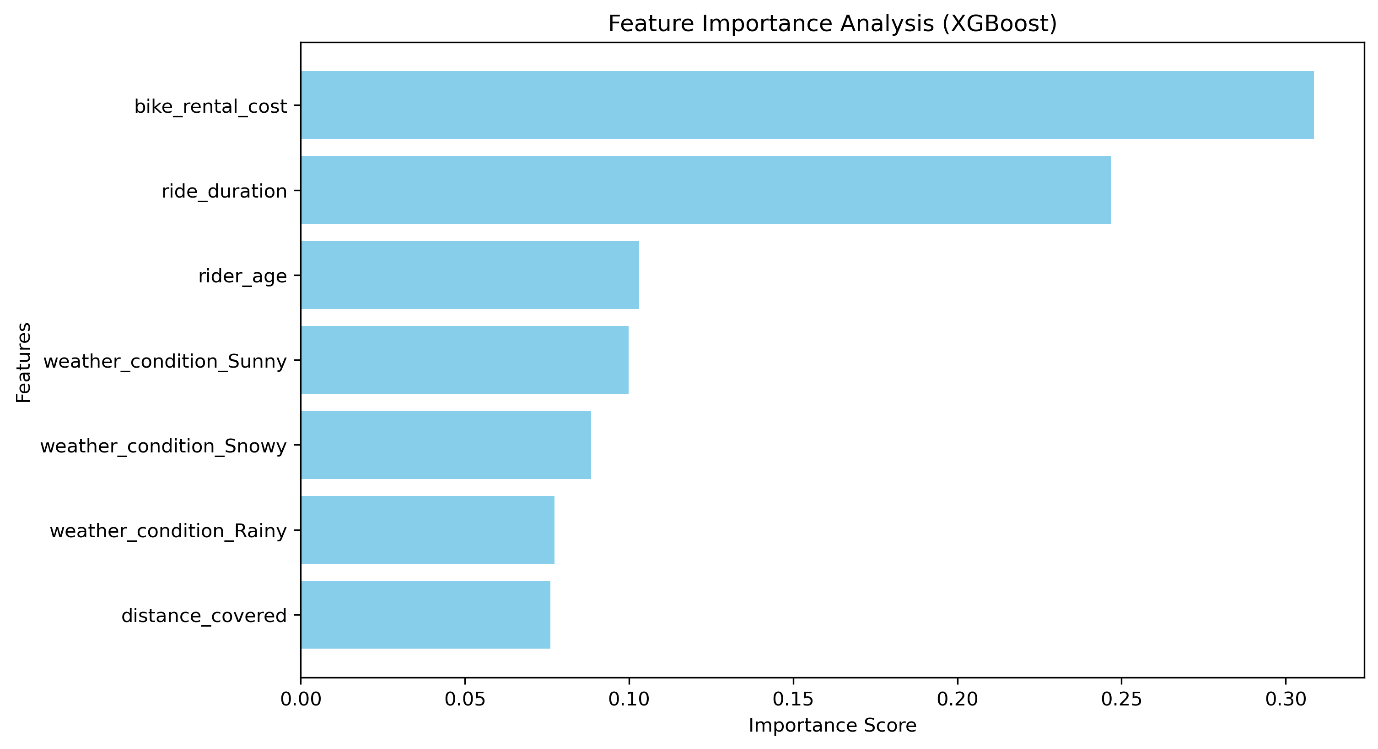
1. **Random Forest Classifier:** Struggled with imbalanced data.
2. **XGBoost:** Excelled in handling imbalances and capturing complex patterns.
3. **Gradient Boosting:** Underperformed due to limited parameter tuning.
4. **Stacking Classifier:** Combined predictions from multiple models, achieving the highest performance.

**Key Insights**

* Simpler models like Linear Regression excelled in regression tasks due to linear relationships.
* Advanced models like XGBoost and Stacking Classifier performed best for classification, particularly with imbalanced datasets.
* Hyperparameter tuning and feature selection were critical for maximizing model performance.





****

# Conclusion

This study demonstrated the value of machine learning in analysing bike rental data to improve operational efficiency and customer satisfaction. Key findings highlight the effectiveness of Linear Regression for predicting rental costs and the superiority of ensemble methods like the Stacking Classifier for classifying electric bike usage. Ride duration and rental cost consistently emerged as the most impactful features, while dimensionality reduction techniques like PCA and LDA enhanced preprocessing and visualization.

Hyperparameter tuning and dataset balancing with SMOTE significantly improved model performance, especially for imbalanced classification tasks. However, limitations such as dataset size and feature constraints suggest opportunities for further exploration. Future studies could incorporate larger datasets, additional features, and real-time predictive systems to enhance decision-making.

These findings underscore the transformative potential of machine learning in driving data-driven strategies for the bike rental industry, optimizing both customer experience and operational practices.

# References

* Bhattacharyya, S., Ohlhorst, F., & Xu, J. (2019). *Machine Learning for Big Data Analysis*. De Gruyter. Available from EBSCO eBook IT Core.
* Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools*. Manning Publications.
* Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
* Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson.
* Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers* (6th ed.). Wiley.
* Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage Publications.
* Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.
* Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.
* Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
* Thakur, A. (2020). *Approaching (Almost) Any Machine Learning Problem*. Abhishek Thakur.