Predicting Uber Fare Prices

A Data Science Project In Python

-Apala Mishra

Introduction:

This project's goal is to create a model that can precisely forecast the cost of Uber and Lyft rides depending on a number of factors, such as the weather, distance, cab type, surge multiplier, destination, and time stamp. The model can assist ride-sharing companies in estimating the cost of a ride, enhancing their pricing strategy, and increasing customer transparency.

Data Source

Source:
https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices

Project Outline

1. Data exploration
2. Checking for null values
3. Data cleaning
4. Important visualizations
5. Regression Model Selection and Training
6. Model Evaluation

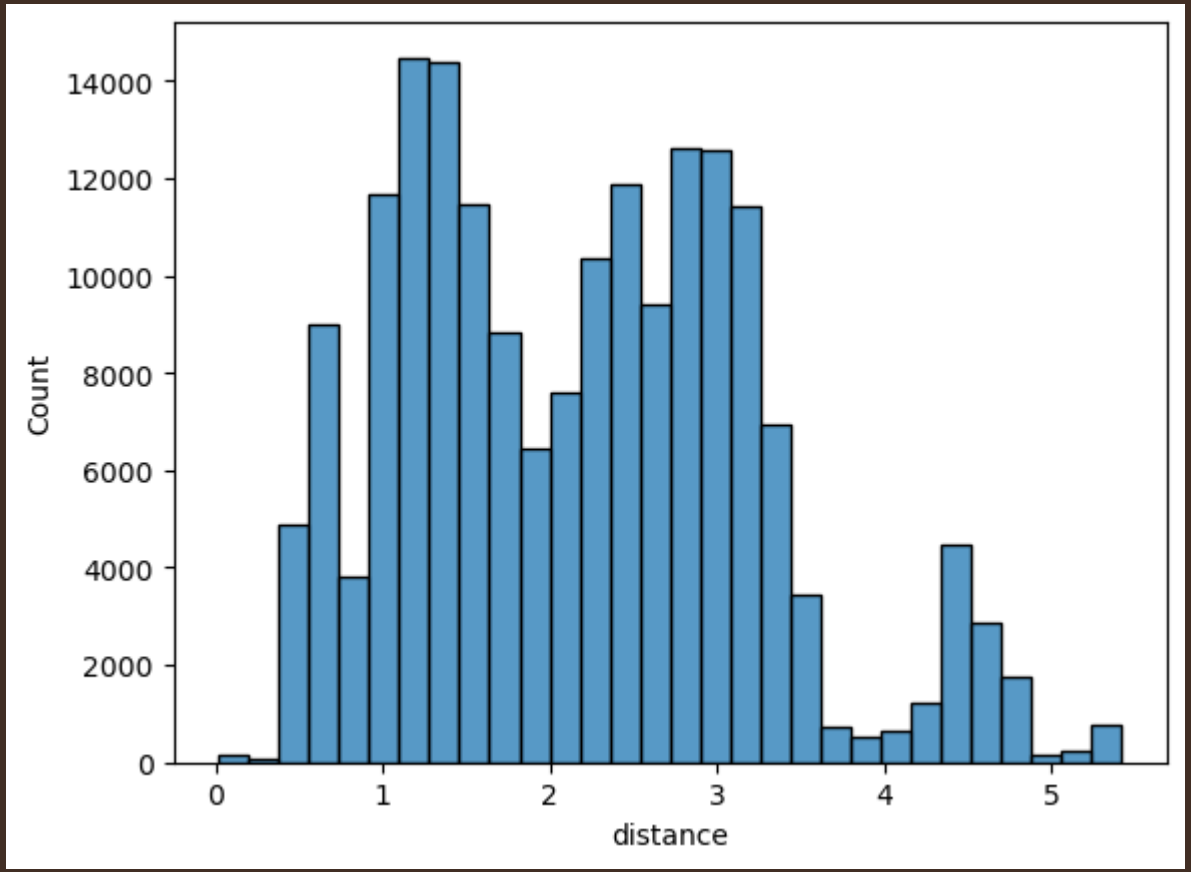
1. Data Exploration

This stage provides a good starting point for exploring the dataset and getting an initial understanding of its structure and content. It helps to identify the number of variables, their data types, and if there are any missing values. These initial explorations can help guide further analysis and modeling decisions.

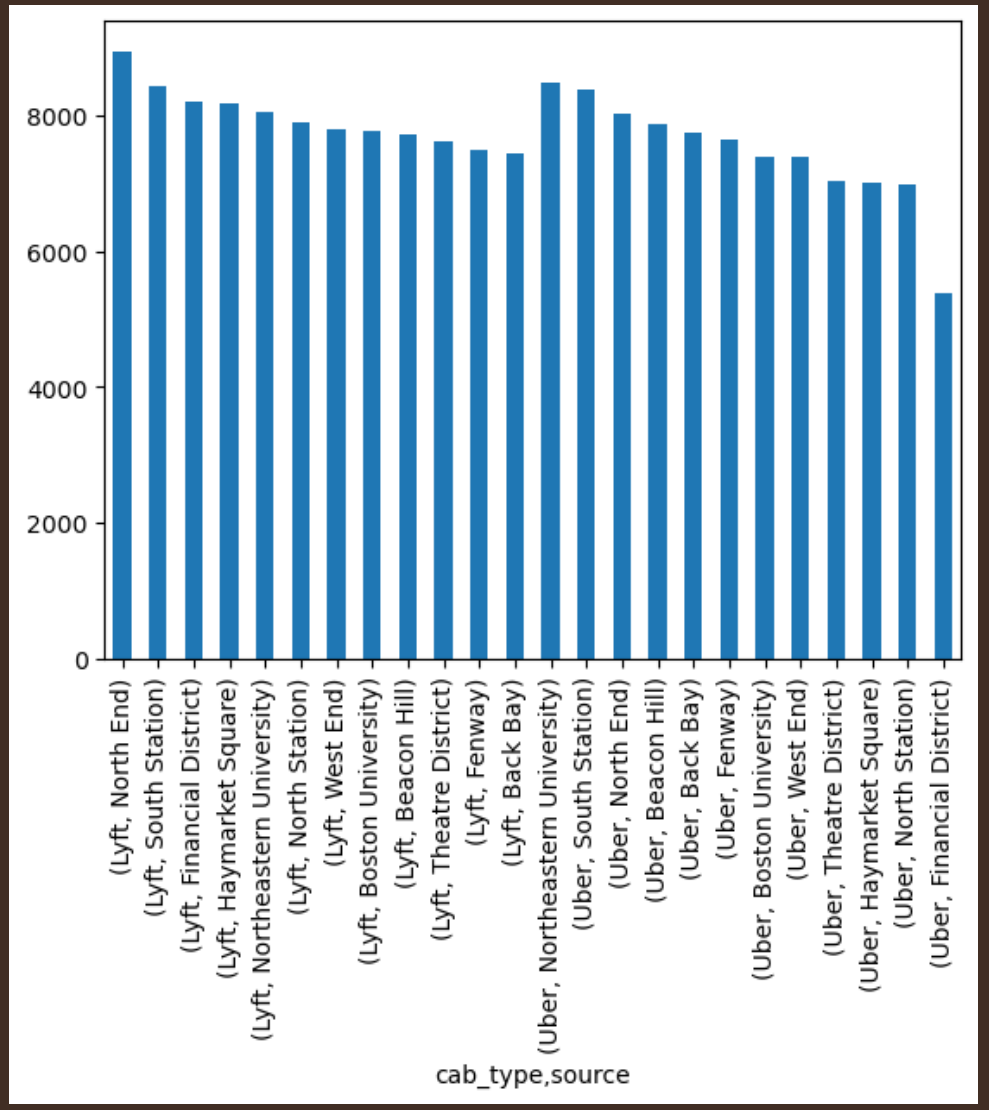
2, 3. Checking for Null Values & Data Cleaning

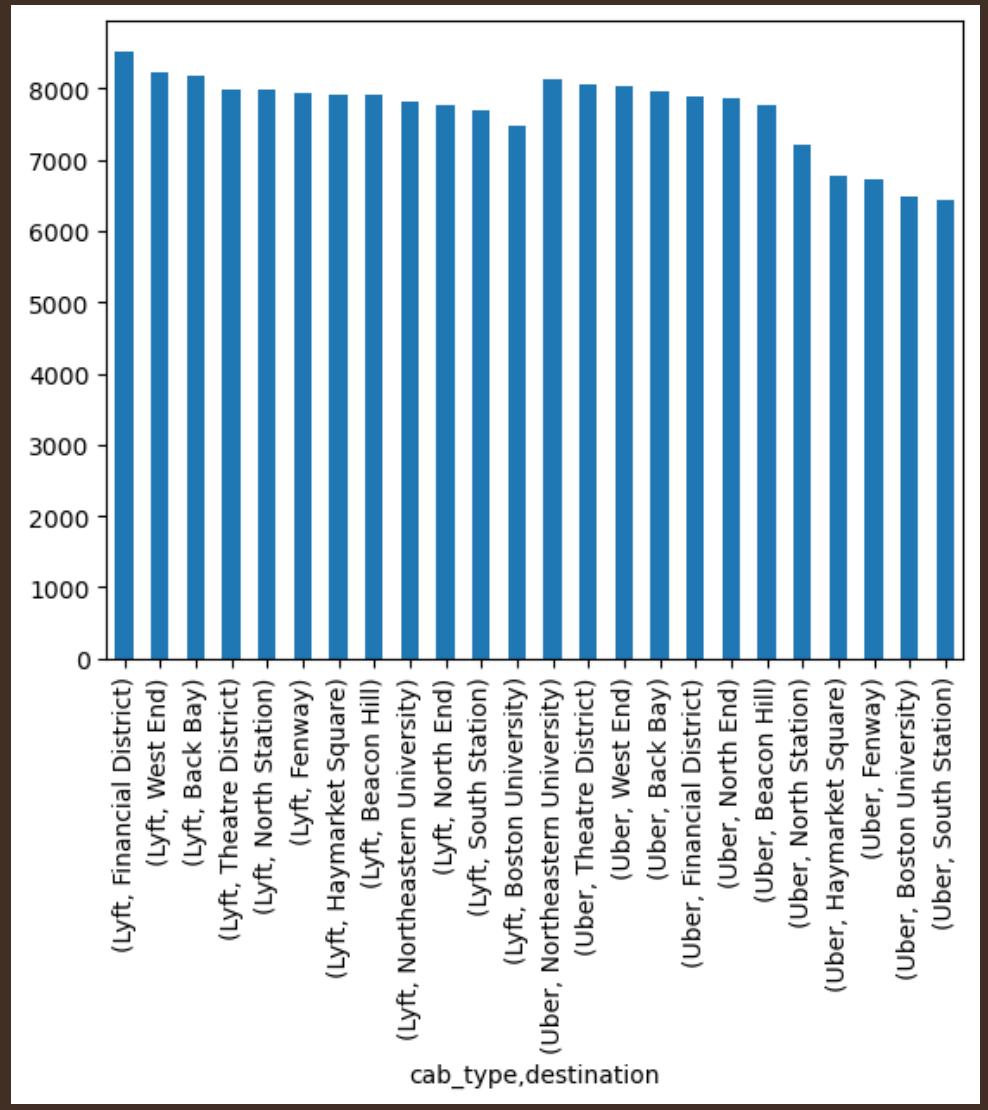
Cleaning the dataset is a crucial step in data analysis as it helps to ensure that the data is accurate, complete, and consistent. In this stage, we need to handle missing values, outliers, and other data quality issues that may affect the accuracy and reliability of our analysis.
By handling missing values, outliers, and other data quality issues, we can ensure that our dataset is accurate, complete, and consistent, which can improve the reliability and accuracy of our analysis.

4. Visualizations

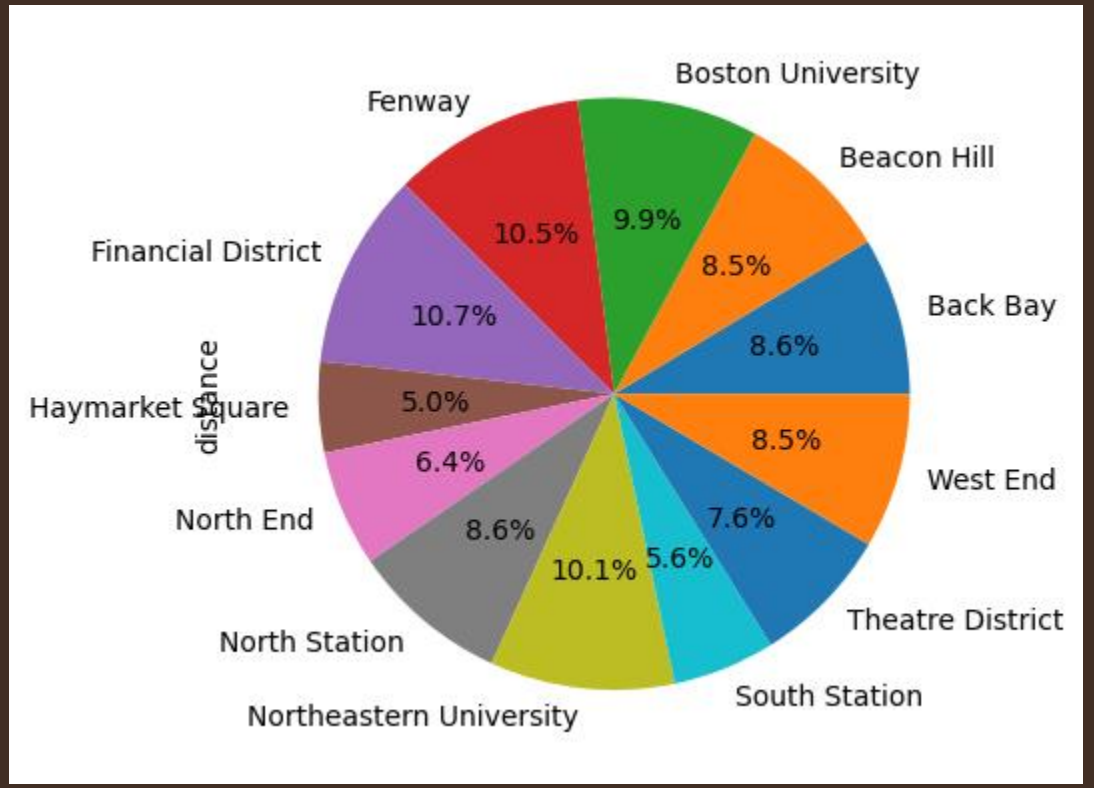


4. Visualizations

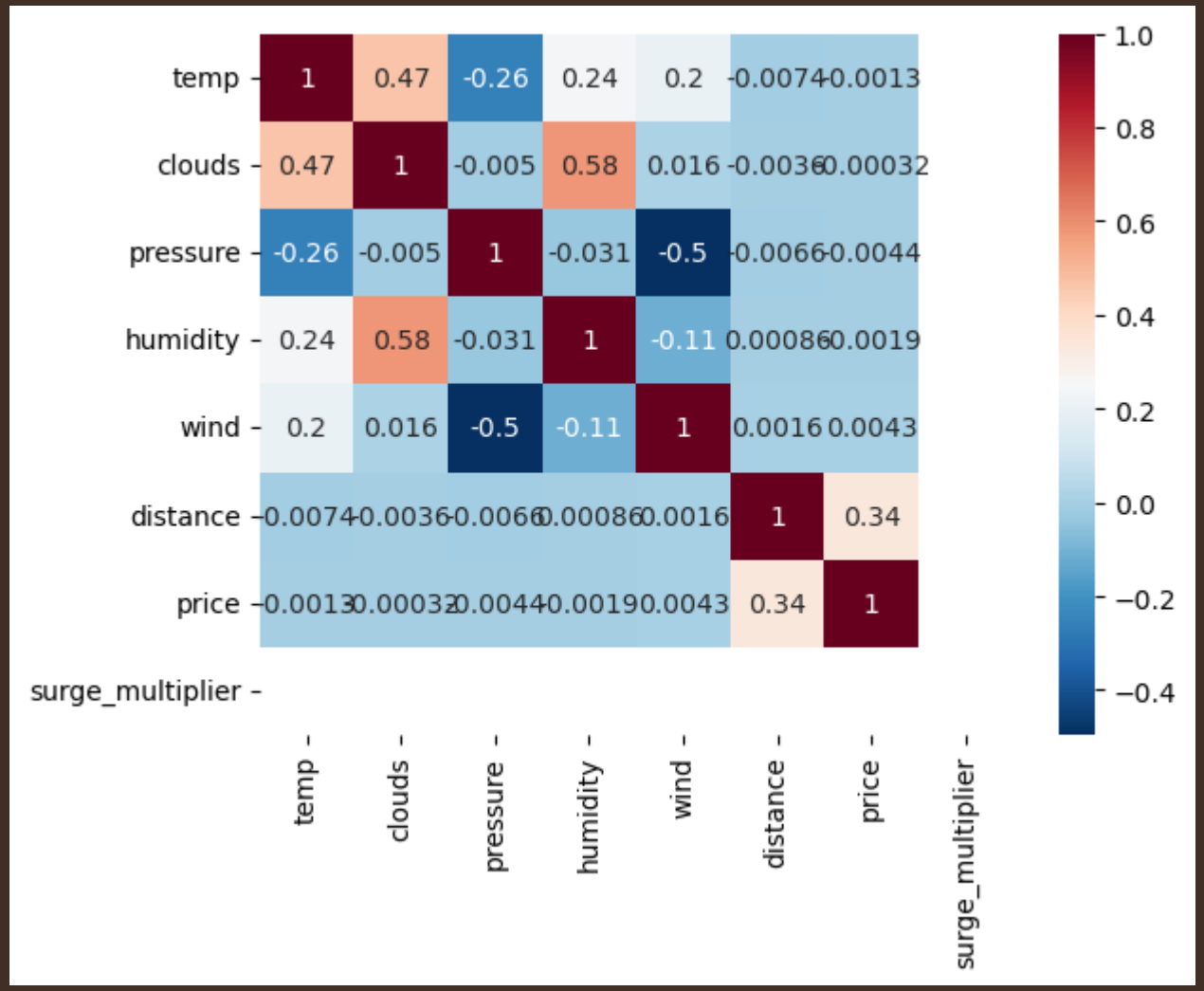




4. Visualizations



4. Visualizations



5. Model Selection:

Linear Regression:
Linear Regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input features and the output variable, and aims to find the best-fit line that minimizes the sum of squared errors between the predicted and actual values.

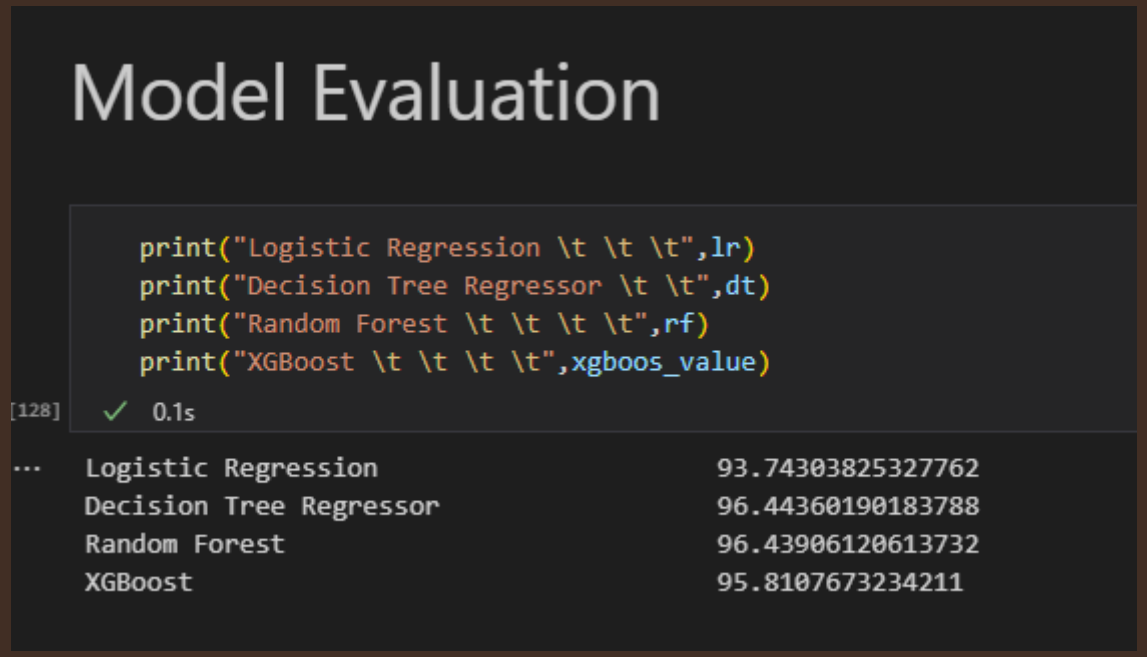
Decision Tree:
Decision Tree is a supervised machine learning algorithm that partitions the data into subsets based on the values of the input features. It creates a tree-like model of decisions and their possible consequences, and uses this model to predict the value of the target variable for new data.

5. Model Selection:

XGBoost:
XGBoost (Extreme Gradient Boosting) is a powerful ensemble machine learning algorithm that combines the predictions of multiple weak models to produce a strong model. It uses gradient boosting and a clever regularization technique to improve the accuracy and speed of the model.

Random Forest:
Random Forest is also an ensemble machine learning algorithm that combines the predictions of multiple decision trees to produce a strong model. It creates multiple decision trees on different subsets of the data and averages the results to reduce overfitting and improve accuracy.

6. Model Evaluation



Thank you