# Introduction

Ride-hailing platforms like Uber have transformed urban transportation by providing convenient, real-time access to rides. At the core of this innovation is a dynamic pricing system that determines ride fares based on a multitude of factors. While distance and time are fundamental determinants of cost, numerous other variables influence pricing, including time of day, day of the week, and external conditions like weather. This project aims to build a predictive model for estimating Uber ride fares in New York City by analyzing such diverse factors and leveraging machine learning techniques to create an accurate and interpretable model.

Accurate fare prediction is not just a technical challenge but also a tool for increasing transparency and efficiency in the ride-hailing ecosystem. Riders benefit from knowing fare estimates in advance, while Uber and its drivers can optimize operations by anticipating demand and pricing patterns. This study dives into this predictive challenge, utilizing a comprehensive dataset that combines ride-specific details with external data sources like historical weather records to build a robust model. By integrating diverse data dimensions, this project sheds light on the complex interactions between geospatial, temporal, and environmental factors in determining ride fares.

The dataset for this study includes information on pickup and drop-off locations, timestamps, and fare amounts for rides in New York City. This is enriched with external weather data, such as temperature, precipitation, and snowfall, matched to the specific time and location of each ride. The dataset undergoes preprocessing to clean anomalies, handle missing values, and engineer features that enhance the model’s predictive power. Derived features include geospatial distances, cyclic encodings of time variables, and indicators of weather conditions.

To address the predictive task, multiple regression models are employed. Linear regression serves as a baseline, while more sophisticated approaches like Random Forest, Kernel Ridge Regression, and Lasso Regression capture complex interactions and account for overfitting. The models are evaluated on key metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics provide insights into the models’ accuracy, robustness, and generalizability.

The results reveal that distance remains the most influential factor in predicting fares, while temporal and weather features add incremental value. Among the models tested, Kernel Ridge Regression stands out with the best performance, highlighting the importance of non-linear relationships in the data. This suggests that while simpler models can explain the majority of fare variability, complex interactions between features require more nuanced approaches.

This project not only demonstrates the feasibility of accurate fare prediction but also underscores the potential for integrating additional external factors, such as real-time traffic or socioeconomic data, to enhance performance further. By bridging data-driven insights with actionable predictions, the study contributes to a deeper understanding of ride-hailing pricing dynamics and lays the groundwork for future innovations in urban transportation analytics.

In the sections that follow, we detail the data sources and cleaning processes, the feature engineering and modeling methods employed, the results and their interpretation, and potential avenues for further exploration. Through this analysis, we aim to illuminate the factors driving Uber ride fares and demonstrate the value of predictive modeling in addressing real-world challenges.

# Data Description

The success of a predictive model is heavily reliant on the quality, breadth, and relevance of the data it utilizes. For this project, three primary data sources were combined to create a comprehensive dataset, each contributing unique and critical information for the task of predicting Uber ride fares in New York City. Below is a detailed description of these data sources and their integration.

## Uber Pricing Data

The foundation of the dataset comes from a publicly available Kaggle dataset containing ride information for Uber trips in New York City. This dataset includes key features such as:

* **Pickup and Drop-off Coordinates**: Latitude and longitude values representing the exact geospatial points for ride start and end locations.
* **Timestamps**: Detailed information about the date and time of each ride, enabling the derivation of temporal features.
* **Passenger Count**: The number of passengers for each ride, which may affect the type of vehicle used and, consequently, the fare.
* **Fare Amount**: The target variable for this predictive task, representing the total cost of the ride.

### **Size and Scope:**

Initially, the dataset contained 200,000 rows of data. After preprocessing (which included cleaning invalid or incomplete entries, handling outliers, and ensuring logical consistency in the data), the final dataset was reduced to 13,953 rows. This curated dataset ensures high data quality, essential for robust model training and evaluation.

## Area Tags

To enhance the understanding of pickup and drop-off locations, the dataset was enriched with area tags derived from **Geopy**, a Python library for geocoding and geospatial data. Using the latitude and longitude values from the Uber dataset, Geopy was used to:

* **Identify Boroughs and Neighborhoods**: Assign descriptive area tags such as “Manhattan,” “Brooklyn,” or “Queens” to rides based on their pickup and drop-off coordinates.
* **Cluster Locations**: Enable the classification of rides into well-defined urban zones, which is valuable for understanding how fares vary across different parts of New York City.

The inclusion of geospatial area tags provides meaningful context to the dataset, helping to explain variations in ride fares due to factors like local traffic patterns, demand, or socioeconomic characteristics of the area.

## Weather Data

Weather conditions play a significant role in ride-hailing dynamics, influencing both demand and driving conditions. To incorporate this critical factor, historical weather data was obtained from the **National Centers for Environmental Information (NCEI)**. This data was matched to the Uber ride data based on timestamps and geolocation to provide accurate weather details for each trip.

### **Weather Features**:

* **Temperature**: Maximum and minimum temperatures on the ride date, indicating seasonal and daily variations.
* **Precipitation**: Rainfall amount, which often correlates with increased demand and slower travel times.
* **Snowfall**: Indicators of challenging driving conditions and potential fare surcharges during severe weather.

The weather data was preprocessed to ensure alignment with the Uber dataset by aggregating it to an hourly resolution and assigning it to rides based on their nearest timestamp and location.

# Models and Methods

To enhance predictive performance, comprehensive feature engineering was performed to extract meaningful information from the dataset. The steps included:

## Feature Engineering

To enhance predictive performance, comprehensive feature engineering was performed to extract meaningful information from the dataset. The steps included:

**Geospatial Features**:

* **Haversine Distance**: Computed the great-circle distance between pickup and drop-off coordinates to represent the straight-line travel distance.
* **Coordinate Differences**: Derived features from the absolute differences in latitude and longitude, which provided additional spatial context.

**Temporal Features**:

* **Cyclic Encodings**: Represented periodic variables like month, hour, and weekday as sine and cosine transformations to account for their cyclical nature (e.g., January follows December).
* **Time of Day Categories**: Segmented hour into bins such as morning, afternoon, evening, and night to capture demand patterns.
* **Day of Week and Holidays**: Incorporated weekday and holiday indicators to identify peak periods.

**Weather Features**:

* Matched hourly weather data (temperature, precipitation, snow) to each ride using timestamps and geolocation.

**Scaling**:

* Standardized continuous features using StandardScaler to ensure numerical stability and compatibility across models.

## Model Selection

To determine the best predictive approach, a combination of linear and non-linear regression models was tested. Each model was chosen for its ability to handle specific data characteristics and relationships.

### Linear Regression

Linear regression was used as the baseline model, providing a straightforward interpretation of relationships between features and fare amounts.

**Advantages:**

* Simple and interpretable.
* Serves as a benchmark for evaluating more complex models.

**Limitations:**

* Assumes linear relationships, which may not fully capture the interactions between features.

### Lasso Regression

Lasso regression introduces L1 regularization to shrink irrelevant feature coefficients to zero, effectively performing feature selection.

**Advantages**:

* Reduces overfitting by selecting only the most relevant features.
* Enhances interpretability by simplifying the model.

**Limitations:**

* May underperform if important features are highly correlated.

### Random Forest Regressor

Random Forest is an ensemble learning method that builds multiple decision trees and averages their predictions, capturing complex non-linear relationships.

**Advantages**:

* Robust to overfitting and noise due to ensemble averaging.
* Captures feature interactions effectively.

### Kernel-Ridge Regression

Kernel Ridge Regression extends linear models by incorporating non-linear transformations through kernels like the Radial Basis Function (RBF).

**Advantages**:

* Captures nonlinear dependencies in data with the kernel trick
* Highly flexible, with the ability to model complex relationships.

# Results and Interpretation

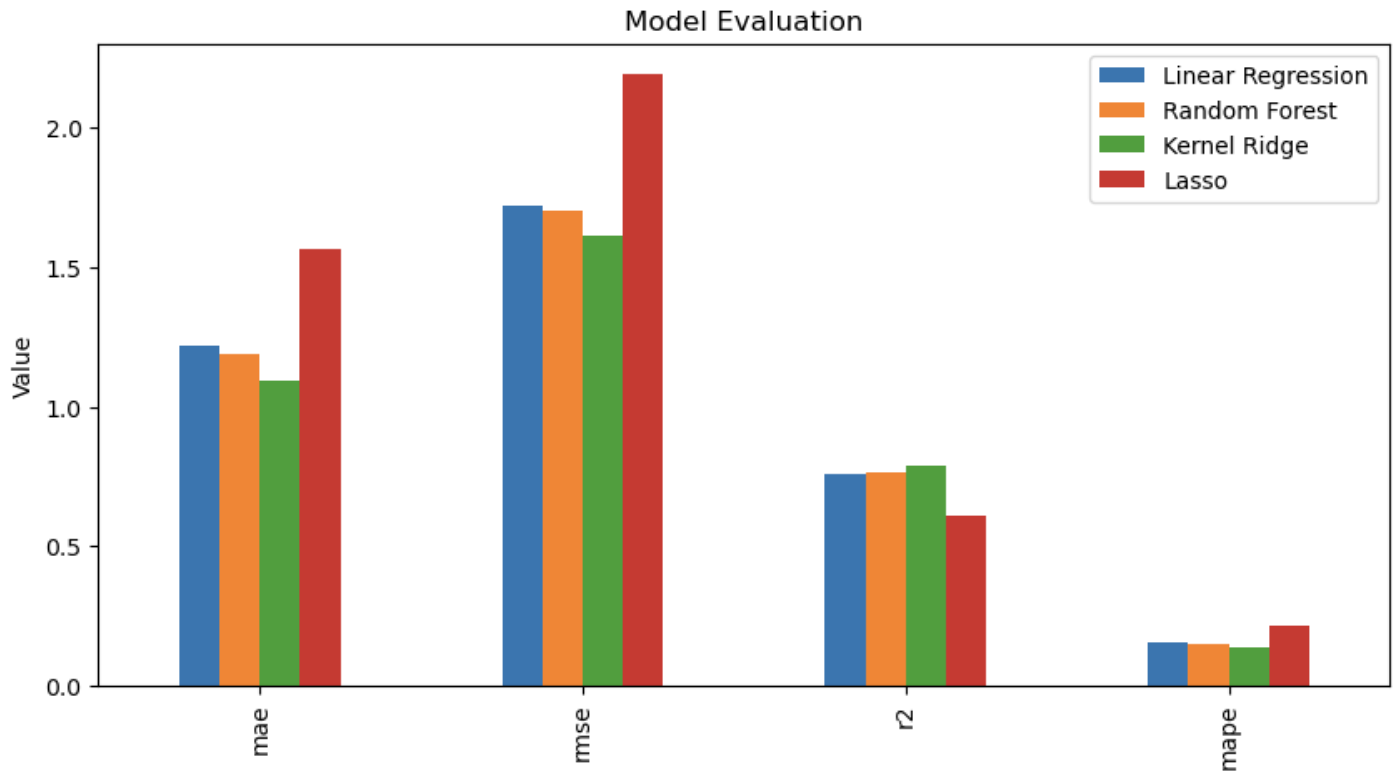
## Performance Metrics

The models were evaluated on the test set using the following metrics:

* **Mean Absolute Error (MAE) -** MAE measures the average absolute difference between the predicted and actual values. It provides a straightforward, easy-to-interpret measure of error.
* **Root Mean Squared Error (RMSE) -** RMSE calculates the square root of the average squared differences between predicted and actual values. It places greater emphasis on larger errors due to squaring.
* **(Coefficient of Determination) -**  measures the proportion of variance in the target variable y that is explained by the model. The closer we get to 1 the better the model is.
* **Mean Absolute Percentage Error (MAPE) -** MAPE expresses the prediction error as a percentage of the actual value. It measures the average percentage difference between predictions and actual values.

## Model Performance

The 4 models performed as follows:



And numerically:



### Kernel Ridge Regression

* Achieved the best performance across all metrics, with the lowest MAE (1.09), RMSE (1.61), MAPE (13.59%) and the highest (0.79).
* This indicates that Kernel Ridge effectively captures the non-linear relationships between features and fare amounts, making it the most suitable model for this dataset.

### Random Forest Regressor:

* Performed slightly worse than Kernel Ridge but still achieved strong results, with an RMSE of 1.70 and an score of 0.76.
* The model is robust to overfitting, thanks to ensemble averaging, and excels in capturing complex feature interactions. However, its slightly higher RMSE suggests it may not fully model non-linear relationships as effectively as Kernel Ridge.

### 

### Linear Regression

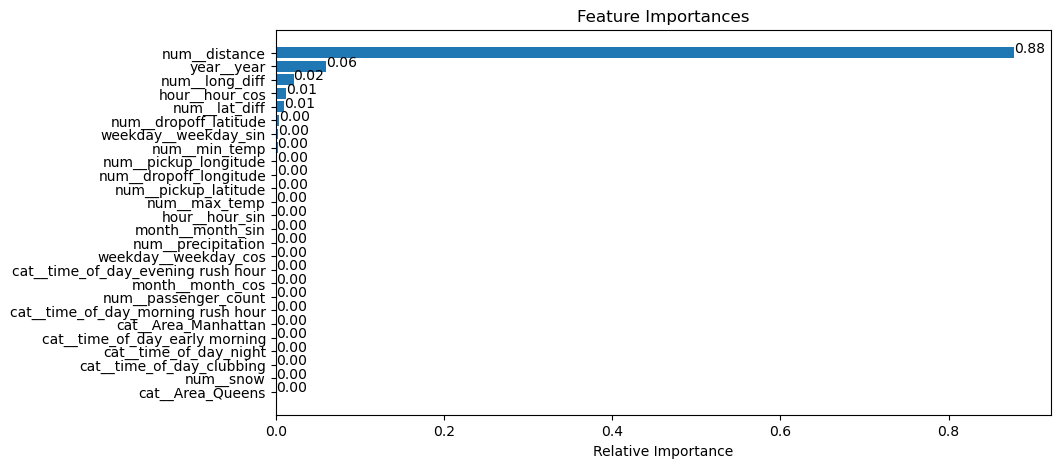
* Provided a reasonable baseline with an of 0.76, indicating that a significant portion of the fare variability can be explained by linear relationships.
* However, the model’s higher MAE and RMSE (1.22 and 1.72, respectively) highlight its limitations in capturing non-linear dependencies and interactions between features.

### Lasso Regression

* Performed the worst among all models, with an of 0.61 and a significantly higher RMSE of 2.19.
* The L1 regularization used in Lasso effectively reduced feature complexity, but in this case, it may have over-penalized relevant features, leading to suboptimal performance.

## Feature Importances

We derived feature importance scores based on the ranking by the random forest regressor. These are shown as follows:



### Distance as the Primary Predictor:

* The Haversine distance between pickup and drop-off locations emerged as the most significant predictor of fare amounts. This aligns with intuitive expectations, as ride distance is directly correlated with cost.
* Furthermore the absolute value of the difference between pickup and dropoff longitude and latitude emerged as weak predictors (denoted num\_long\_diff, num\_lat\_diff)

### Impact of Temporal Features:

* The year of the ride also emerged as a predictor. Intuitively we know that Ubers have gotten more expensive over time so this makes sense.
* The hour of day is also shown as a weak predictor of prices. Specifically, the cos component appears to be a predictor showing that our encoding strategy was fruitful. It would make sense for fares to be higher during peak hours.

### Weather Influence:

* Surprisingly, the weather data we combined with our original dataset made virtually no difference according to the Random Forest Regressor.

## Limitations

### Sparse Weather Data

* The weather data was matched to rides based on geospatial proximity and timestamps, but variations in weather conditions across different areas of New York City may not have been fully captured.

### Unobserved features

* Factors such as real-time traffic conditions, rider-specific preferences, and dynamic surge pricing were not included, which could further improve prediction accuracy.

### Data Quality

* During our cleaning process, we dropped approximately 80% of the rides due to either missing values, or unrealistic values.
* It is possible that useful information was removed due to the cleaning process.

# Conclusion and Next Steps

## Conclusion

This study explored the development of a predictive model for Uber ride fares in New York City, leveraging a dataset enriched with geospatial, temporal, and weather-related features. The analysis aimed to identify the factors influencing fare amounts and to build models capable of making accurate predictions.

The results demonstrate that fare amounts are predominantly driven by ride distance, while temporal and weather conditions provide significant additional predictive power. Features such as pickup and drop-off locations, time of day, day of the week, and weather events (e.g., precipitation and snow) collectively contribute to variations in fare pricing. Among the models tested, Kernel Ridge Regression emerged as the most effective, achieving the lowest error rates and the highest R^2, showcasing its ability to capture complex, non-linear relationships in the data. Random Forest Regressor also performed well, highlighting the value of ensemble methods for robust predictions.

However, the study also revealed certain limitations. The weather data, although valuable, was matched based on location and timestamp, which may not fully capture localized variations in conditions. Additionally, factors such as real-time traffic conditions, dynamic surge pricing, and rider-specific preferences were not included in the dataset, leaving room for further improvement in prediction accuracy.

The project highlights the importance of rigorous feature engineering, thoughtful model selection, and robust evaluation in predictive modeling tasks. By integrating multiple data sources and testing diverse approaches, the study offers a comprehensive understanding of the factors influencing Uber fare dynamics.

### Areas for Improvement

* Incorporating Traffic variables may help predict surges
* Neural networks may be able to capture the nonlinear aspects of Uber Pricing and could allow for improved accuracy
* While we did incorporate Geo-spacial variables, it may be useful to have even more granular data. For example, instead of Manhattan, we have East Village, Soho, etc.

# References

[Sklearn](https://scikit-learn.org/stable/)

[Geopy](https://geopy.readthedocs.io/en/stable/)

[Weather](https://www.ncei.noaa.gov/cdo-web/orders?id=3865537&email=hatimqr02@gmail.com)

[Kaggle](https://www.ncei.noaa.gov/cdo-web/orders?id=3865537&email=hatimqr02@gmail.com)