Predicting trip duration from ride-hailing data in Chicago, US

Team 4

Louis, Malak, Lena and Eero

Business Understanding

Background

Prediction Task

Relevance

Business Understanding

Background

- City of Chicago
- Ride-hailing industry

Prediction Task

- Trip duration estimation

Relevancy

- Enhancing real time predictions
- Stakeholder Analysis

Stakeholder Analysis

Primary Stakeholders

- Operations Team
- Customer Service
- Product Management
- Data Science Team

Secondary Stakeholders

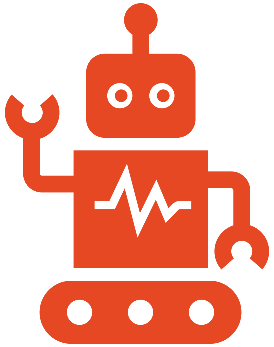
- Drivers
- Passengers
- Business Leadership

External Stakeholders

- Ride-Hailing Providers
- Reulators and City Planners

Data Understanding

Data Source Overview

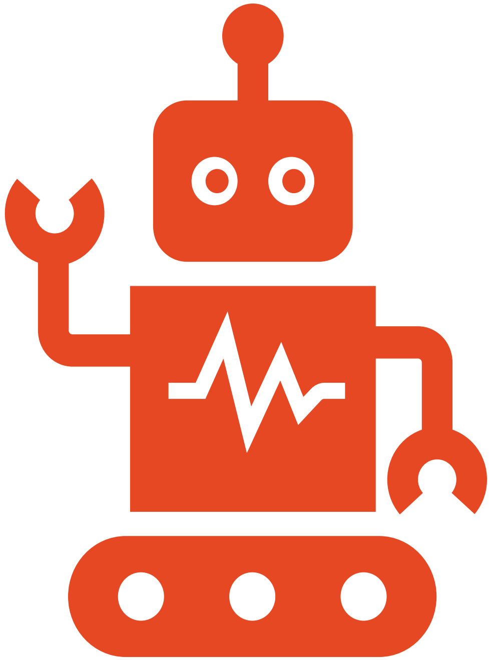


Data Acquisition & Loading

Data Source Overview

- City of Chicago Open Data Portal
- 2023-2024
- 174M rows of data
- 24 individual features
- Data collected via routine reporting by ride-hailing companies

Data Acquistion & Loading



- Filter application
- Reducing data to every 100th row
- 327,526 individual trips for further processing
- ~55 trips per unique pickup/dropoff area code combination

Data Limitations & Modeling Considerations

Deficient Data Quality

Data Representation Challenges

Modeling Considerations

Data Preparation

Data Exploration*

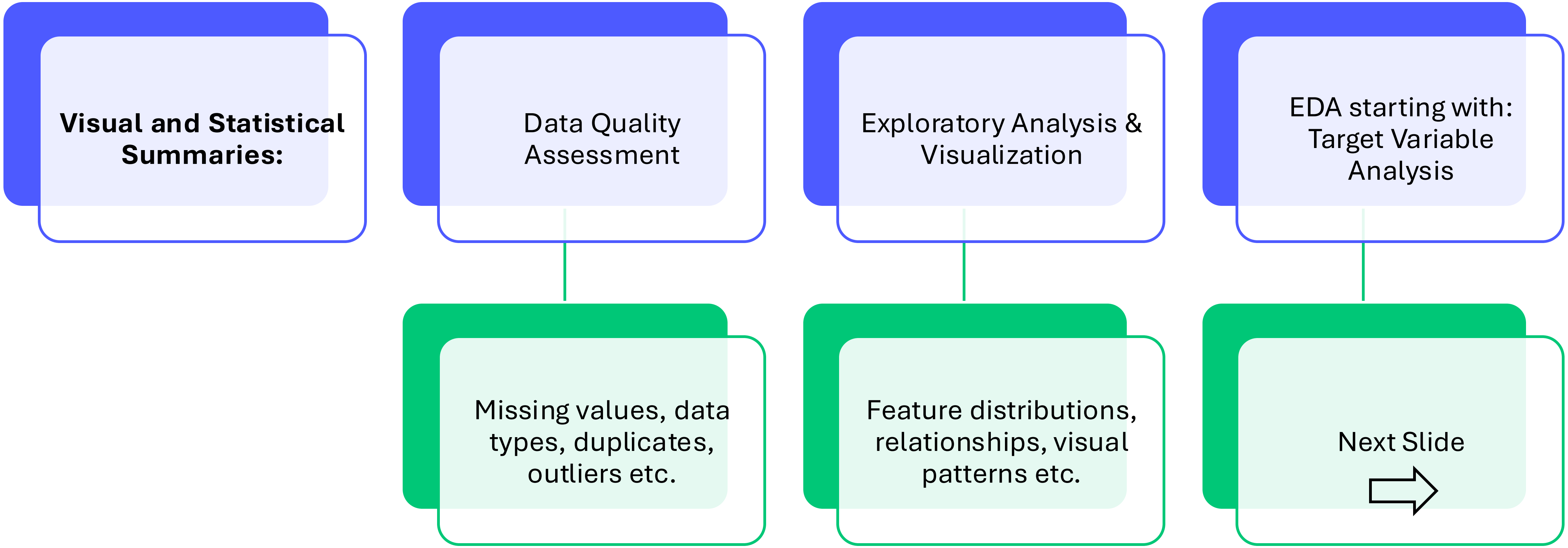
Data Cleaning

Feature Engineering

Feature Selection

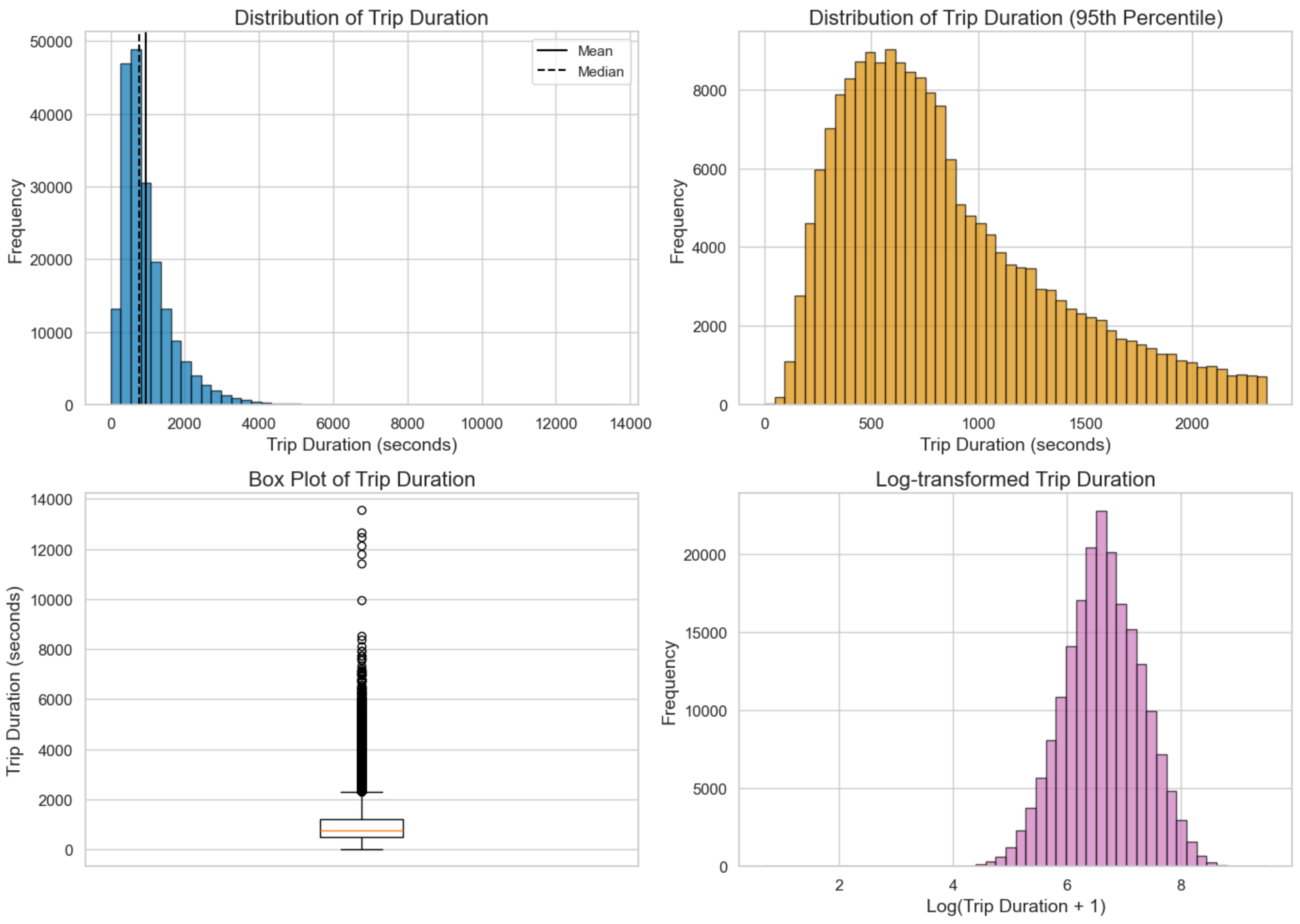
Post-Prep EDA & Visualizations

Data Exploration



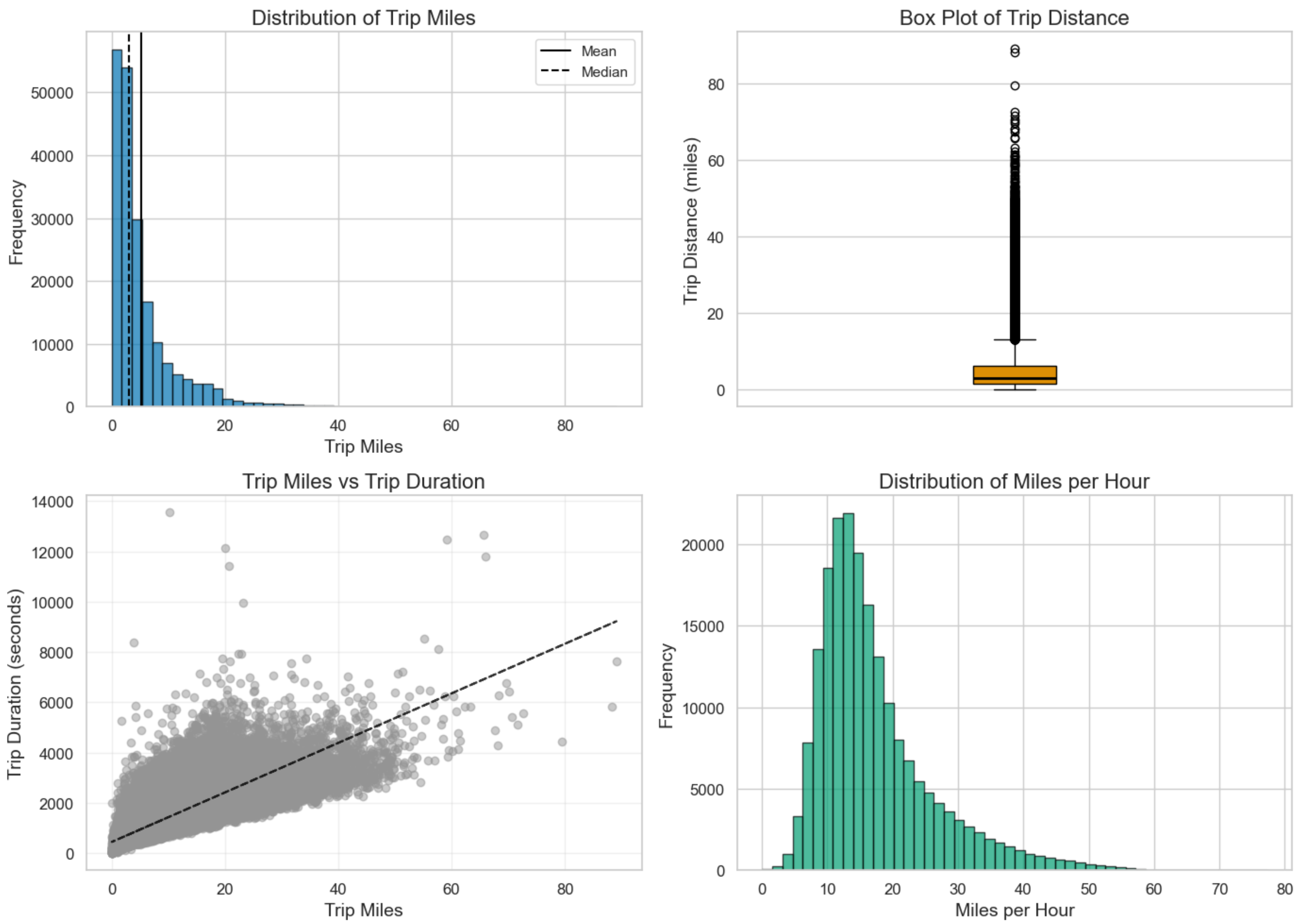
Various graphs regarding trip duration

- Mean: 16 minutes
- Median: 12.6 minutes
- Distribution heavily right-skewed
- Most trips between 4-16 minutes



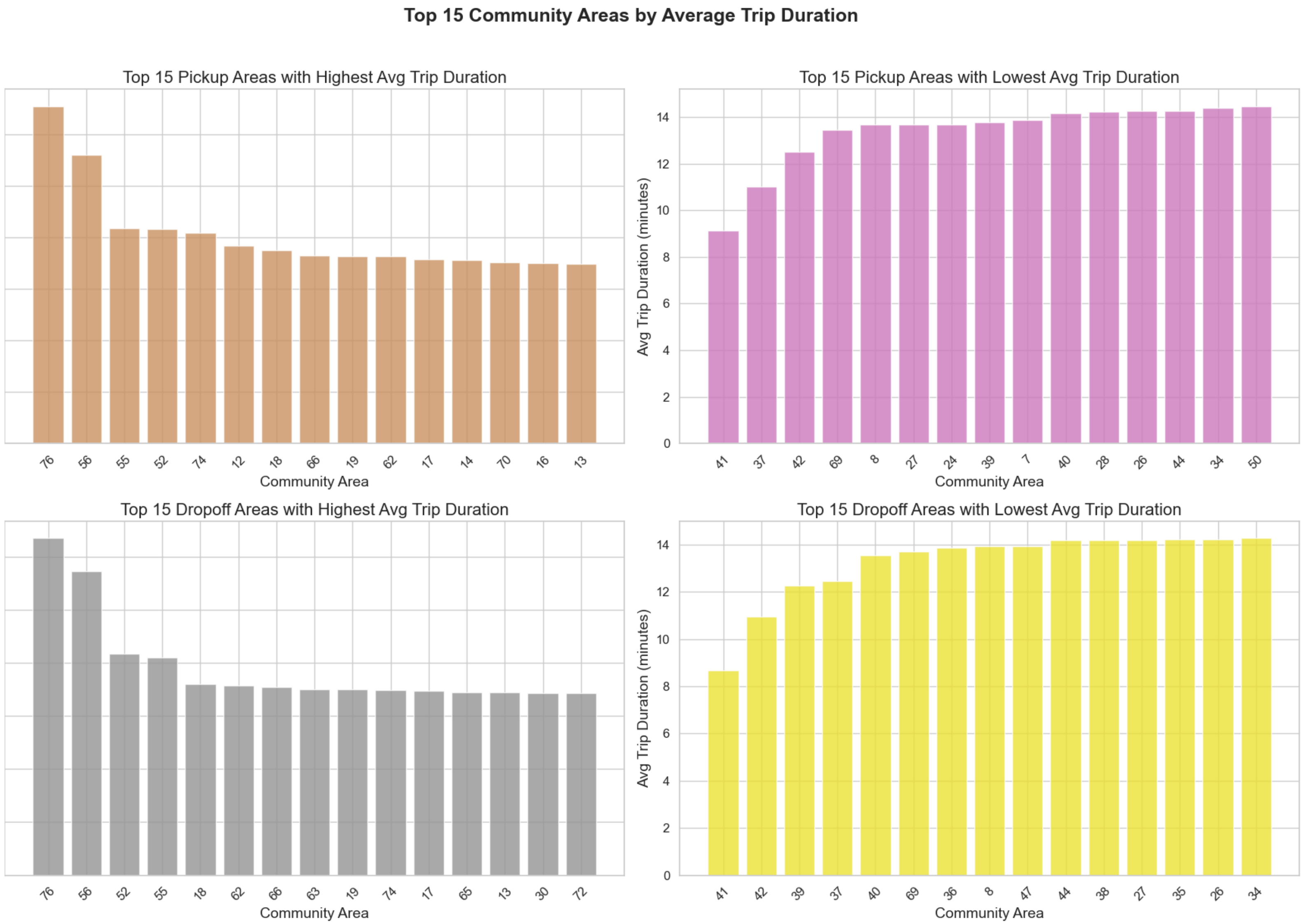
Trip distance vs. Trip duration analysis

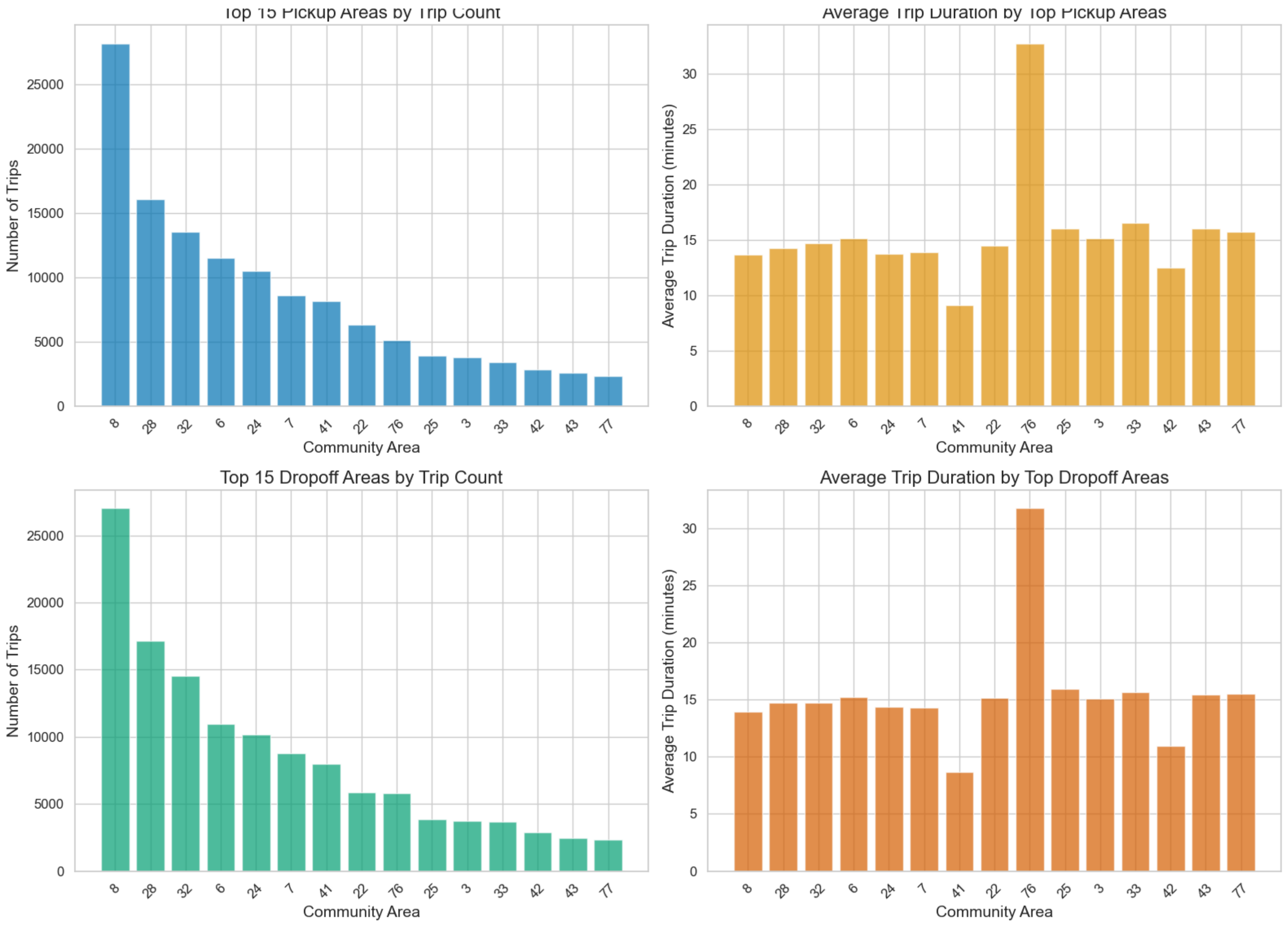
- Positive correlation with high variability
- Frequent delays relative to distance
- Speed distribution reflects urban conditions
- Distribution consistency confirms data quality



Strategic takeaways for stakeholders via spatial analysis

- Commuter Zones (12&18)
- Peripheral Areas (76&56)
- Short-Trip Neighborhoods (41,37,42)
- Service Planning





Insights from Top 15 community areas by trip count

- Busy areas (8,28,32)
- Range of trip duration
- Statistical anomalies (41&76)

Stakeholder Reflections

Are long trips from peripheral areas a result of poor infrastructure, or natural commuting patterns?

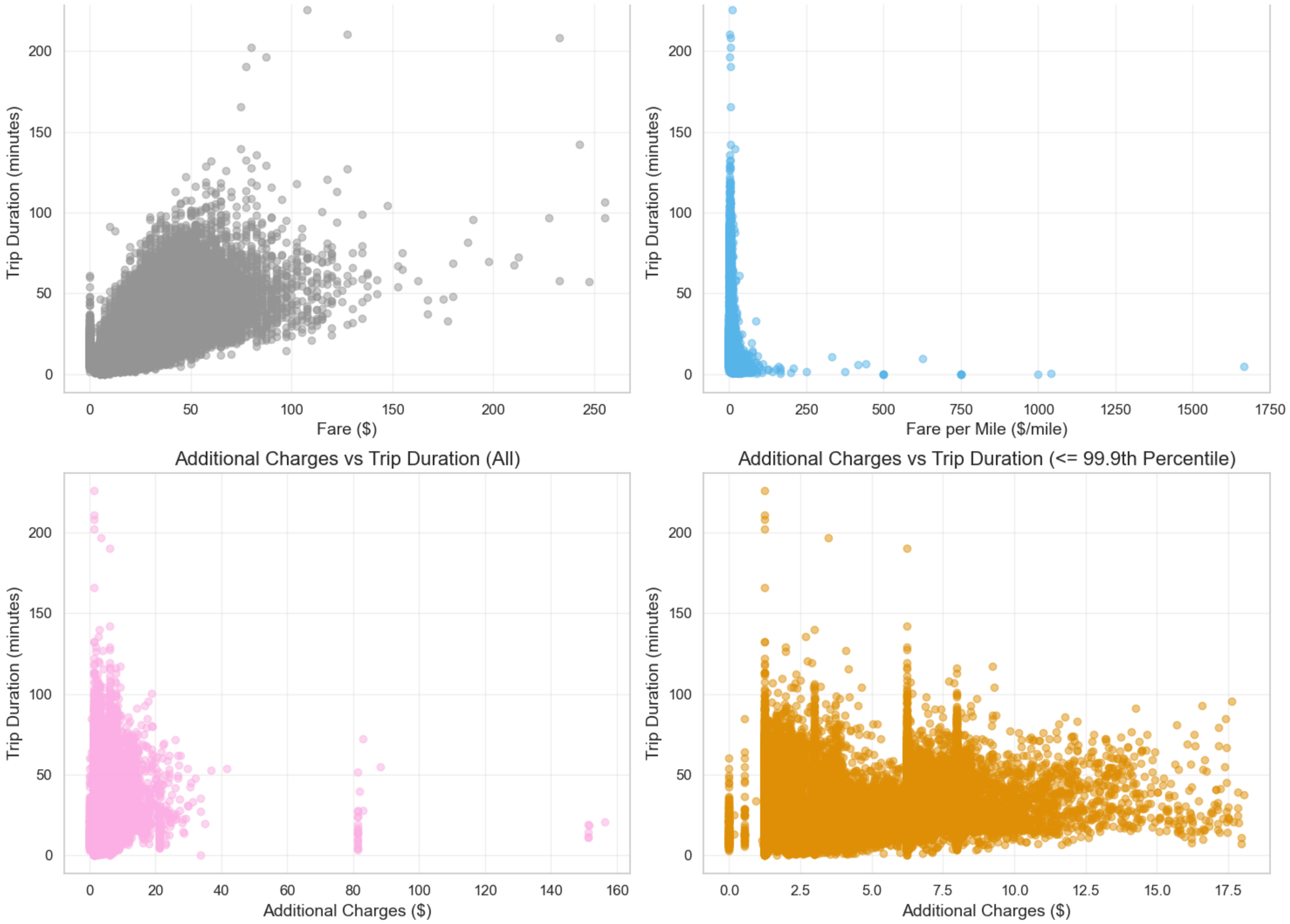
Should we target **service improvements** (e.g., transit or rideshare coverage) in areas with longer average trip durations?

Can we **optimize dispatch or resource allocation** based on highly local vs. outbound trip behaviors?

What **socioeconomic or geographical features explain why** some areas generate shorter vs. longer trips?

Financial Analysis

- Strong positive correlation between trip duration and fare
- Efficiency gains per-mile
- Charge Clustering
- Independent additional charges



Possible Business Implications for Stakeholders

Pricing Strategy

Distance-based efficiency

Short trip premium

Surge pricing impact

Operational Insights

Trip duration sweet spot

Pricing transparency

Market segmentation

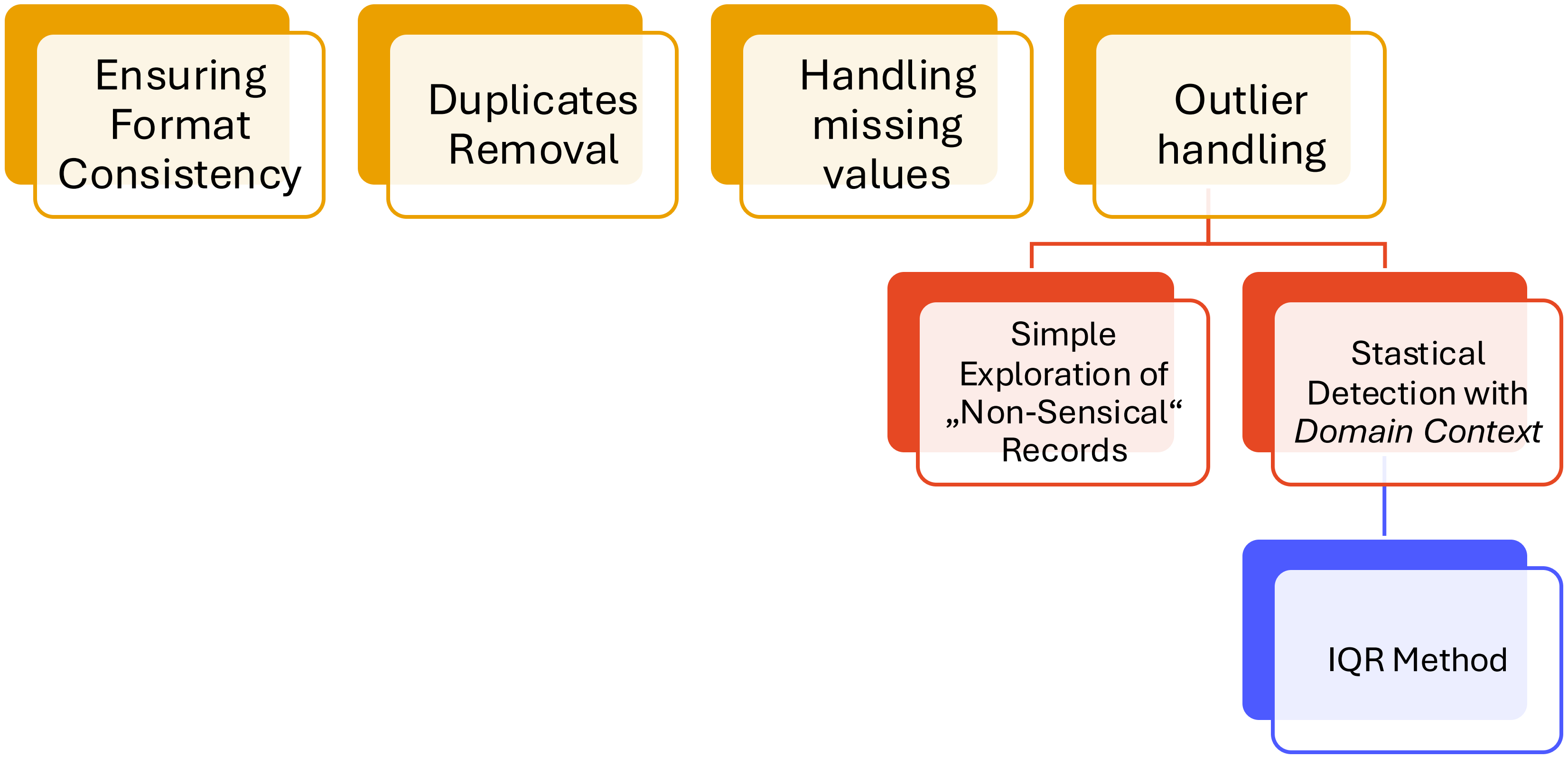
Data Quality Observations

Outlier Management

Missing values

Skewed distribution

Data Cleaning



Feature Engineering

Temporal Feature Extraction

- Extracting from *trip_start_timestamp* via Pandas

Merging Weather Data

- Weather conditions like *temperature, rainfall, wind etc.*
- Cleaning up Weather Data

Removal of Irrelevant Features

- Based on previous descisions
- Prior final feature selection

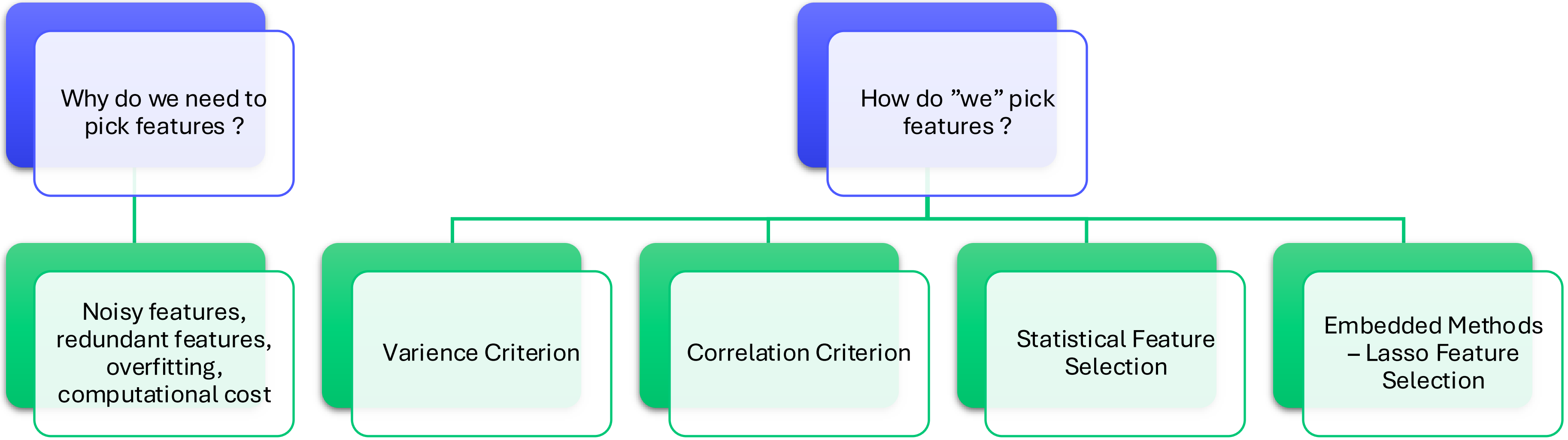
Feature Encoding

- For categorial (coordinate) variables

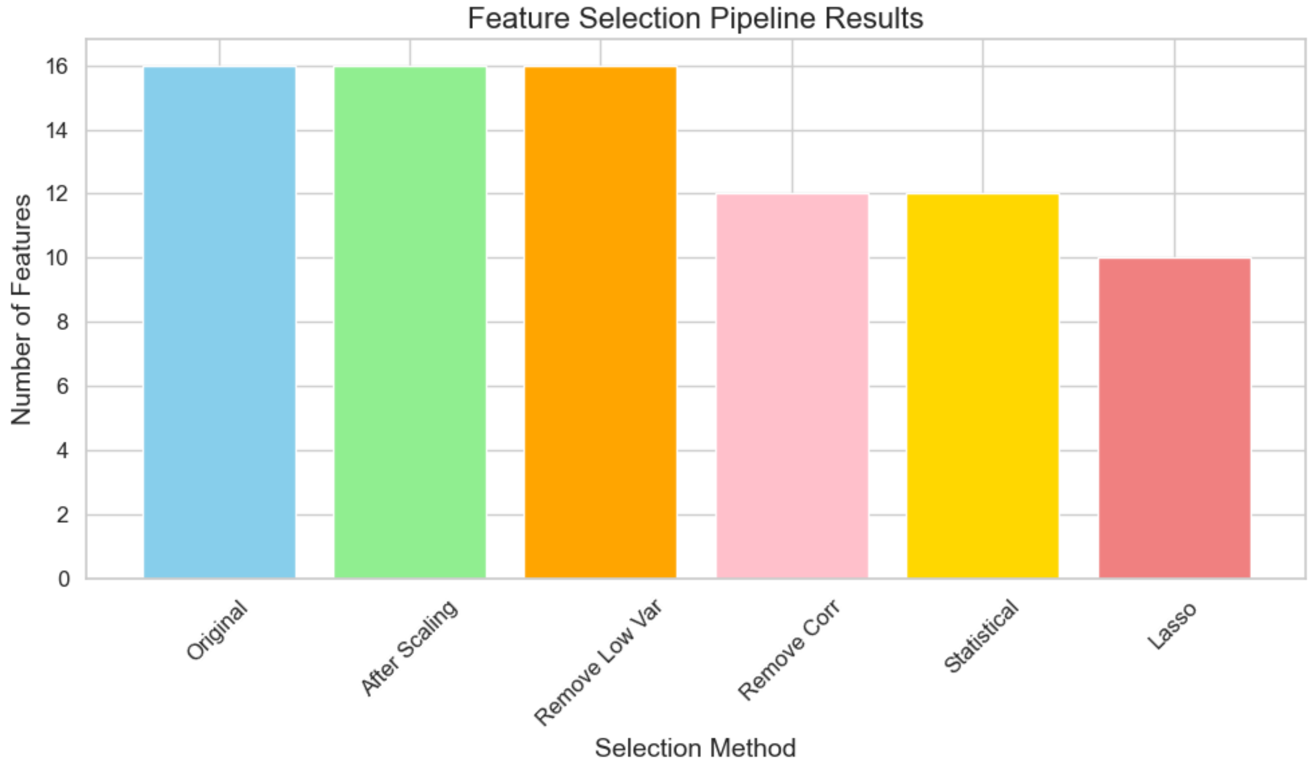
Feature Scaling

- Standardize numeric features using StandardScaler

Final Feature Selection



Final Feature Selection - Coding Results

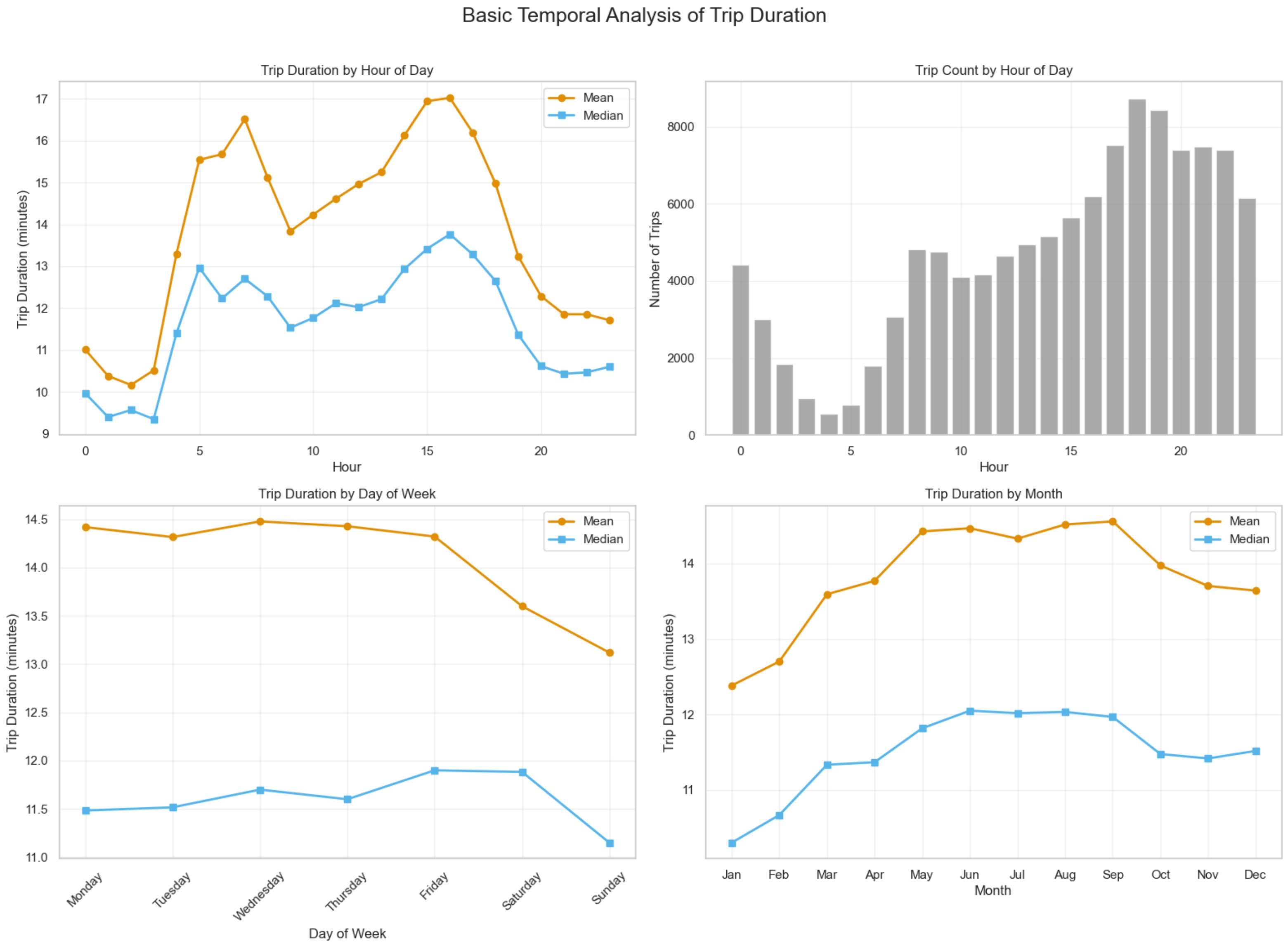


- Trip_miles
- Fare
- Additional_charges
- Hour
- Is_peak_hour
- Month
- Temp_pickup
- Prcp_pickup
- Wsdp_pickup
- Dropoff_community_area_target_encoded
- Trip_seconds

Post-Prep EDA & Visualizations

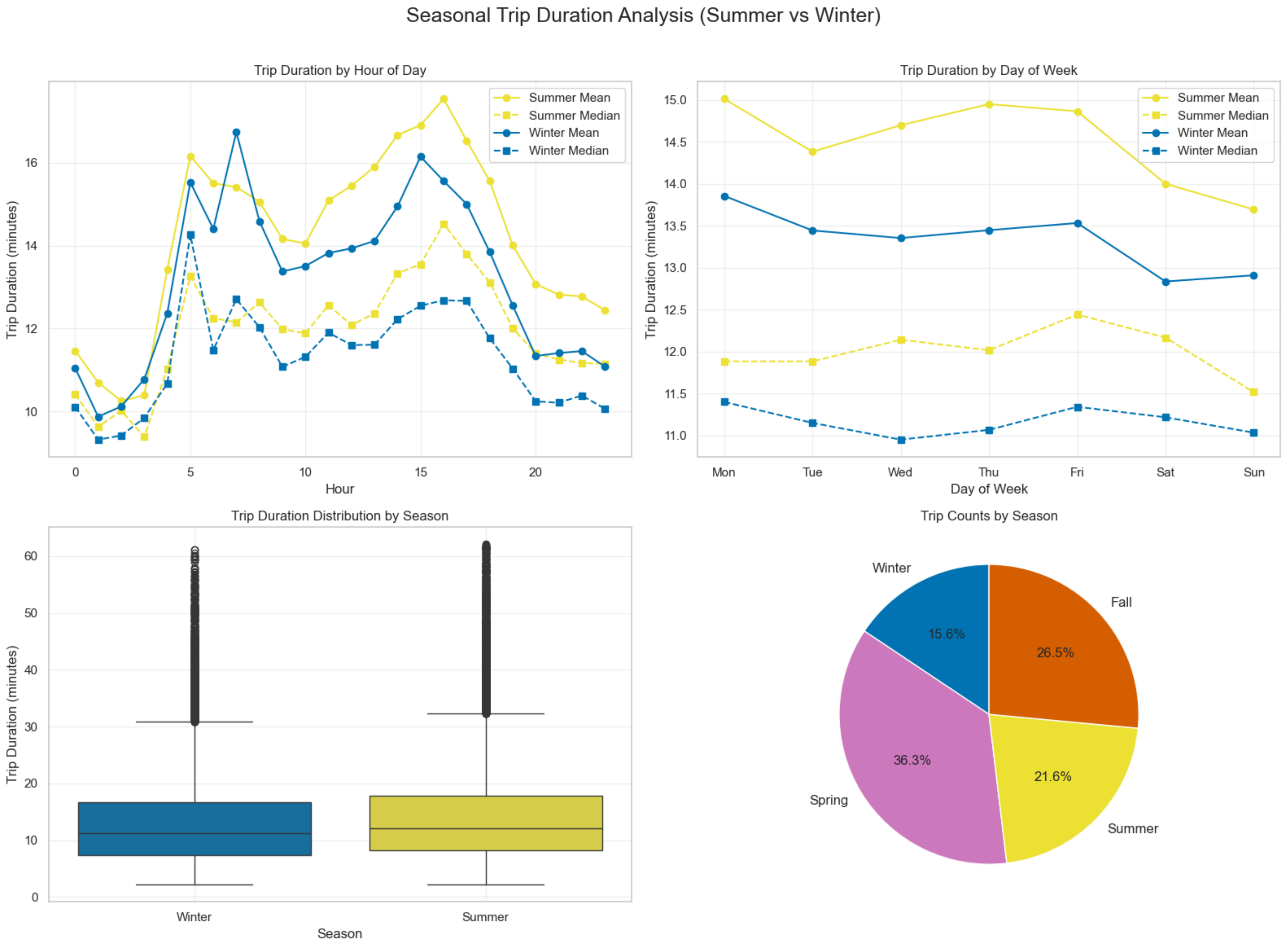
Weather data correlation

Temporal factors



Basic Temporal Analysis of Trip duration

- Peak hour congestion (4pm&7am)
- Increasing demand over the day
- Weekday vs. weekend patterns
- Seasonal Variation

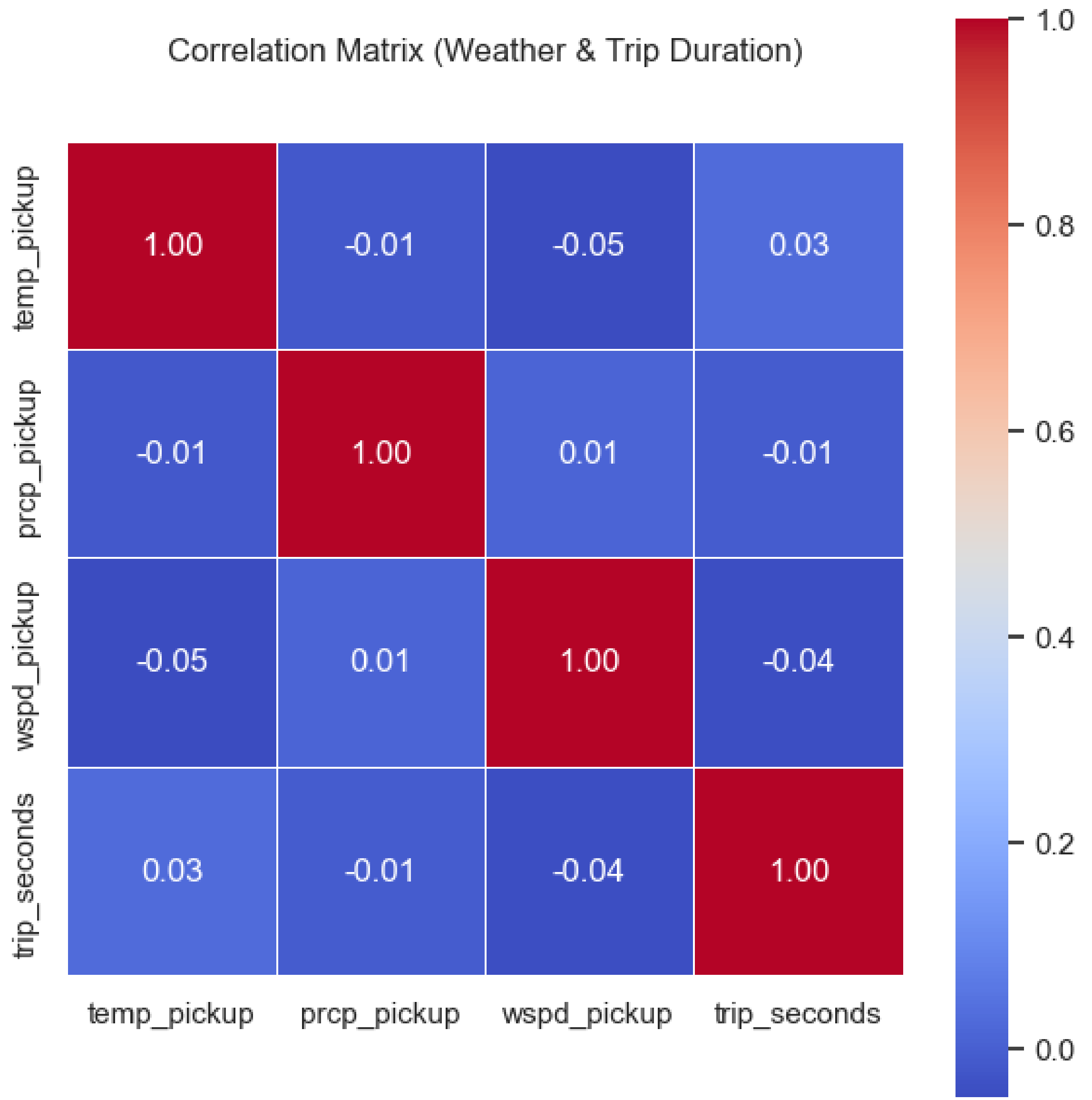


Seasonal Insights

- Higher demand and longer trips during summer
- Stronger peak hours during summer
- ➡ Seasonal marketing and off-season opportunities

Correlation Insights: Weather vs. Trip duration

Weather variables show very weak linear relationships with trip duration



Modeling

The Machine Learning Framework

- Hypothesis function
- Loss function
- Optimization method

Model Selection

- Linear Regression
- Random Forest Regression

Evaluation Strategy

Data Splitting

- Training Set (60%)
- Validation Set (20%)
- Test Set (20%)

Complementary regression metrics

- RMSE
- MAE
- R^2 Score

Multiple Polynomial Linear Regression

Baseline model

- Identifies non-linear Relationships between certain features and target while remaining linear in its parameters

Objective

- Enhance capabilities through polynomial features and ridge regularization

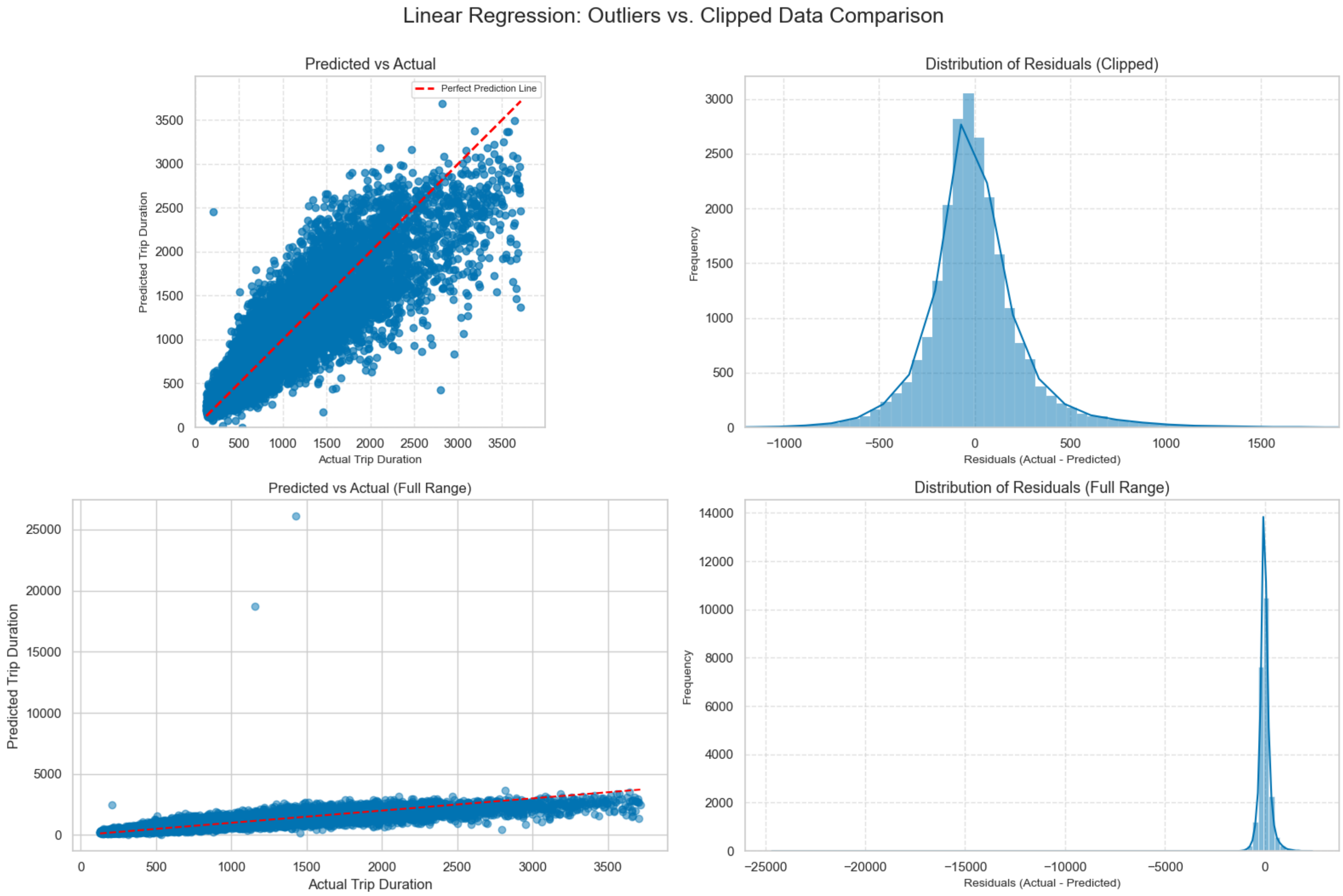
Implementation Considerations

- Streamlining with Pipeline
- Hyperparameter Tuning with GridSearchCV

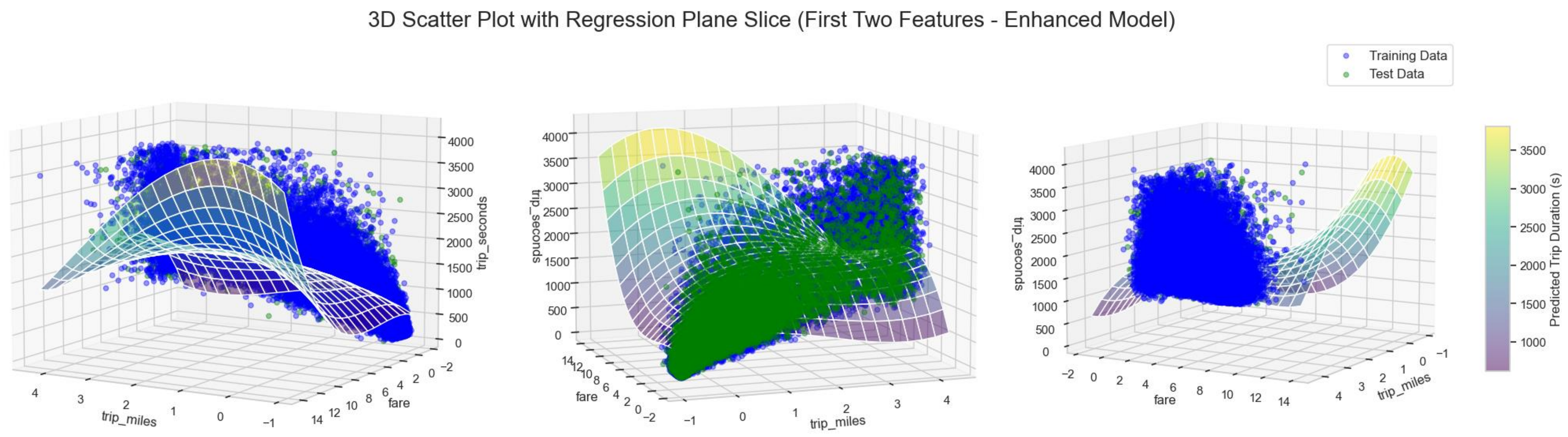
Results

- R^2 score of 0.6546
- RMSE of 328.301 seconds

Multiple Polynomial Linear Regression



Multiple Polynomial Linear Regression



Random Forest Regressor

Key Concepts

- Bagging (Bootstrap aggregation)
- Feature Randomness
- Prediction aggregation

Benefits to our Prediction Task

- Complex feature interactions in urban mobility
- Ability to handle mixed data
- Robustness to Outliers

Limitations

- Interpretability
- Performance on small datasets

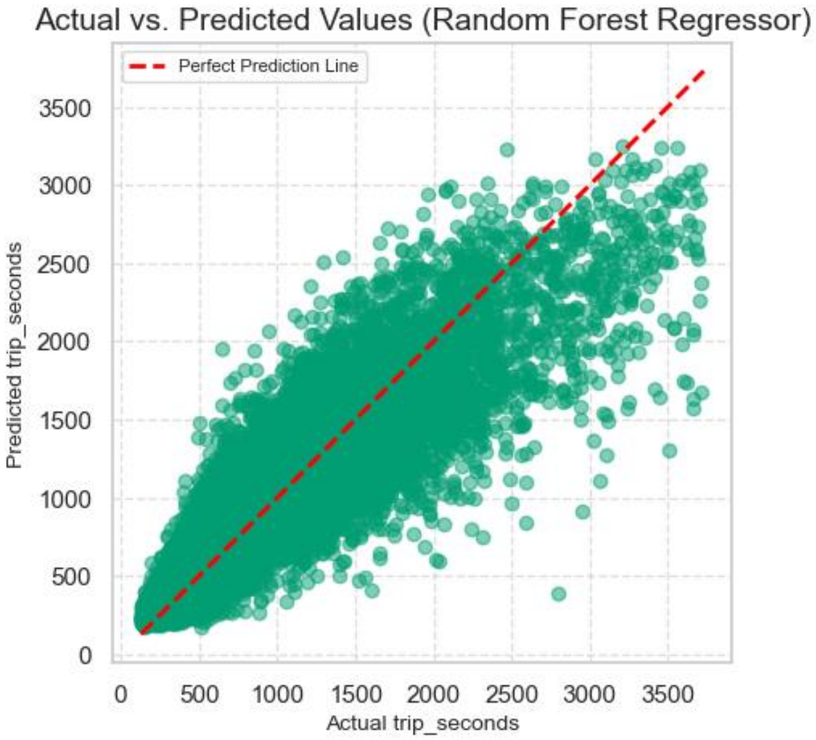
Results

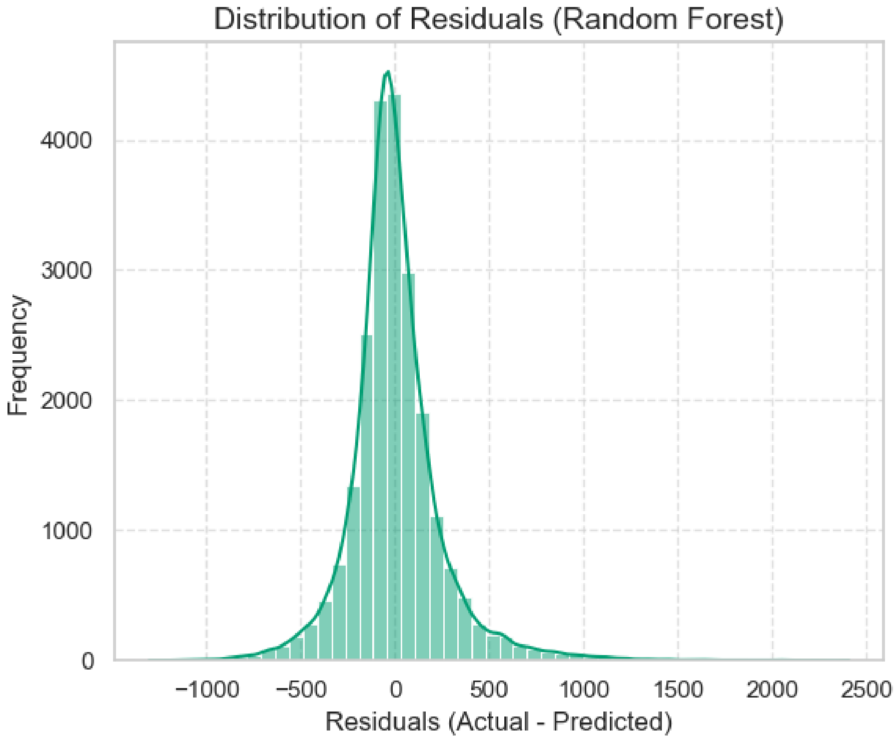
- RMSE of 246.283 seconds
- R^2 score of 0.8056

Hyperparameter Tuning with RandomizedSearchCV

- Wide range of hyperparameter values
- Faster than GridSearchCV
- Tests fixed number of random combinations
- Very good approximation of the best model
- Reduces training time

Random Forest Regressor





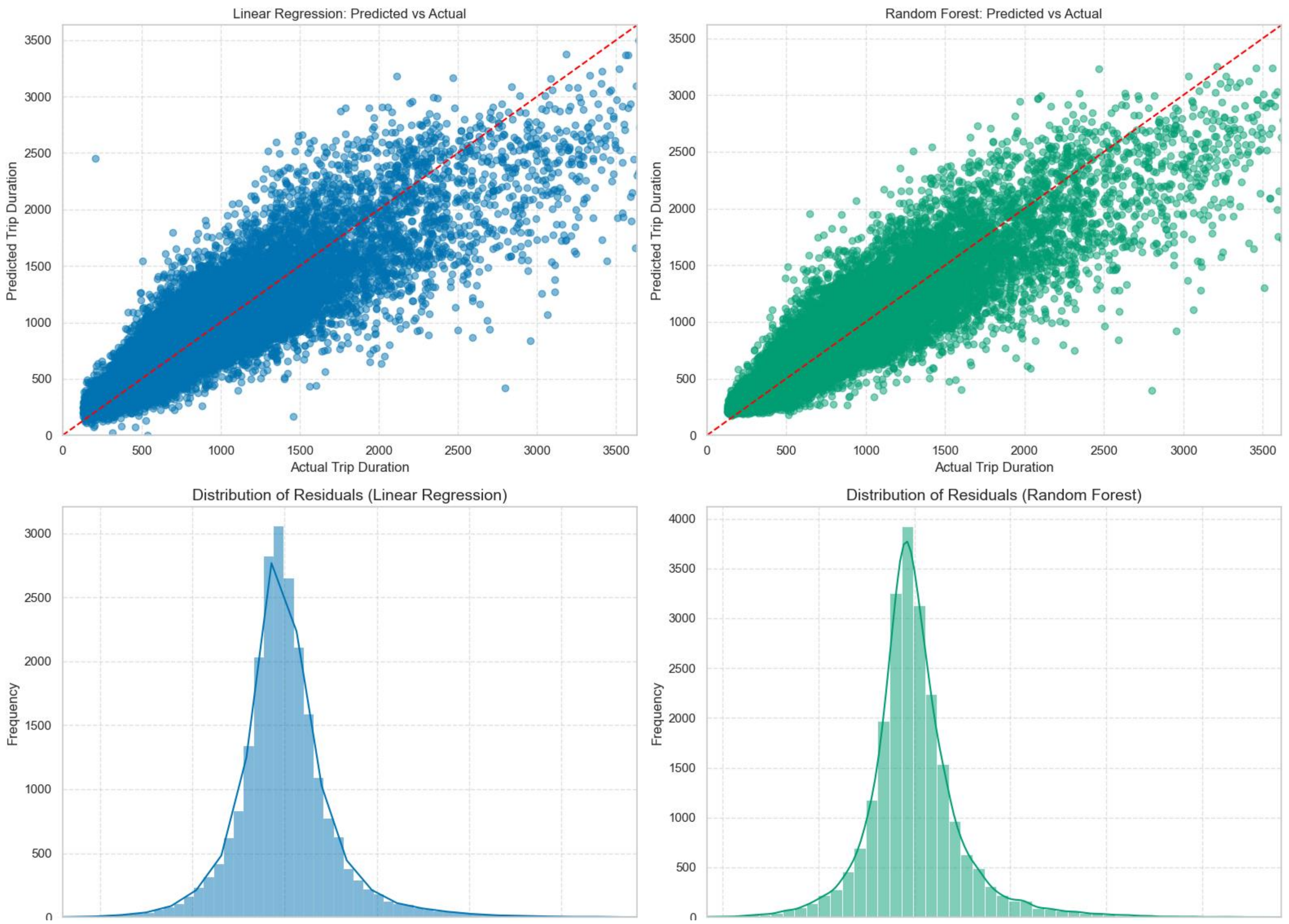
Evaluation of Trip Duration Prediction Models

Model Performance Comparison

- Random Forest Regressor Outperforms Linear Regression

Model	RMSE (min)	MAE (min)	R^2
Linear Regression	5.47	3.51	0.65
Random Forest	4.10	2.76	0.81

- **Random Forest:** Lower errors (RMSE, MAE) & better fit (Higher R^2)
- Visually, Random Forest shows tighter predictions & better residual distribution.



Visual Comparison of Predictions

- **Actual vs. Predicted Scatter Plots & Residuals**
- The performance is also visually evident

Feature Importance - Key Differences

Understanding Feature Influence

Linear Regression: Coefficients for *transformed* & *scaled* polynomial features.

Interpretation is complex; shows influence of combined/curved relationships.

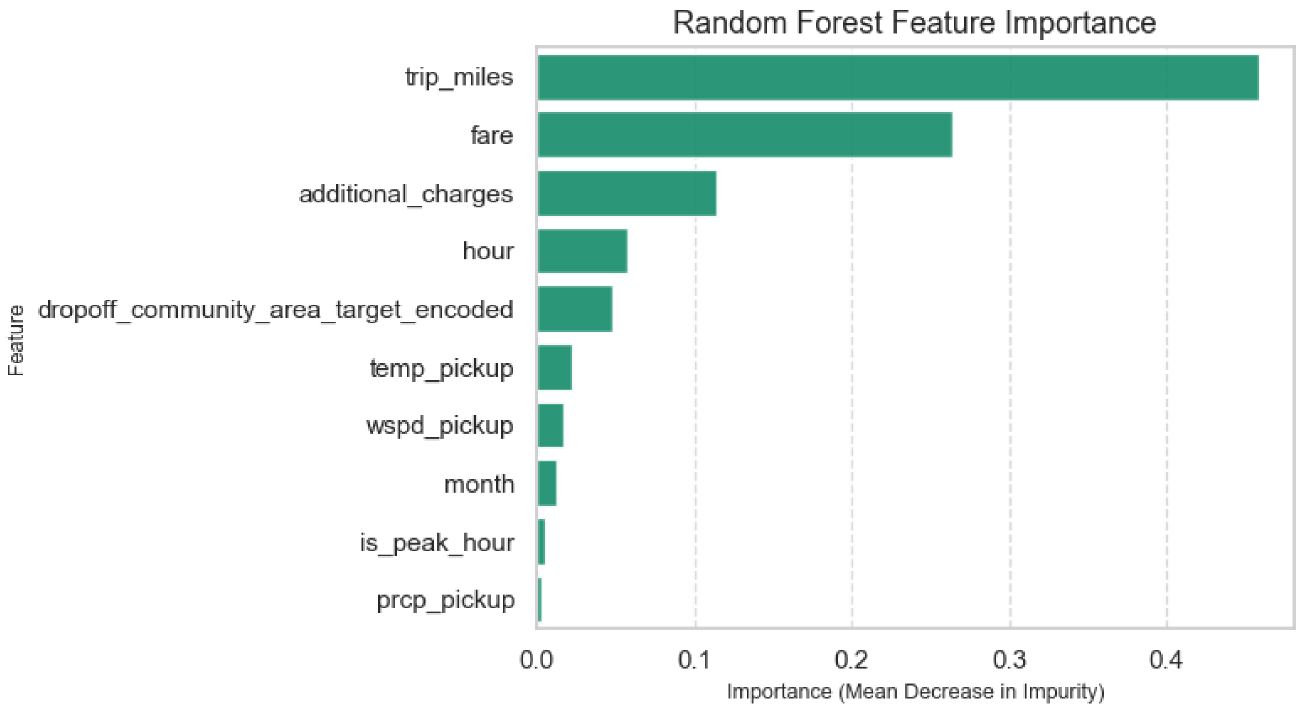
Random Forest: Importance for *original* input features (e.g., Gini importance).

Clear, direct, and intuitive measure of impact.

[redacted]

```
Feature Importance (Linear Regression) ["Top 15)"]:
                                             Feature  Coefficient
0                                         trip_miles   330.439342
1                            trip_miles is_peak_hour^2   217.994425
2                                       trip_miles^2  -185.832697
3                             trip_miles is_peak_hour  -153.049650
4                                             hour^2  -143.471992
5                                               fare   106.217965
6                                    trip_miles fare    94.832009
7                                  fare is_peak_hour^2    75.091321
8    trip_miles dropoff_community_area_target_encoded    53.166722
9                                             hour^3   -51.767111
10                                              hour    51.702226
11                                 hour is_peak_hour   -47.474875
12                                additional_charges    46.638869
13                                trip_miles hour^2   -45.665637
14                                 fare is_peak_hour   -42.366717
```

Feature Importance - Top Drivers



- **Both Models Highlight trip_miles and fare**
- **Random Forest:** trip_miles and fare are overwhelmingly most important (**72**% combined).
- additional_charges and hour also significant.

Model Selection 1

Multiple Polynomial Linear Regression

Pros:

- non-linear relationships through feature engineering.
- faster prediction
- informative coefficients

Cons

- Assumes a functional form for non-linearity
- Sensitive to outliers
- Interpretation of coefficients
- Less robust
- Lower performance than Random Forest

Model Selection 2

Random Forest Regressor

Pros:

- Captures non-linear and complex feature relationships.
- Handling of a mix of numerical and categorical features
- Robust to outliers and noisy data
- Interpretable feature importance scores for original features.
- Generalization performance
- Superior performance metrics

Cons:

- Black box model
- Training is computationally more intensive
- Hyperparameter tuning

Final Selection = Random Forest

- significantly better performance metrics
- better generalization to unseen data
- ability to handle non-linearity
- Robustness to outliers
- Black box
- No informative coefficients
- Slower prediction

Business Perspective

01

Value for Stakeholders

- Accurate trip_seconds prediction offers substantial value across the ride-hailing ecosystem.
- A comprehensive perspective, accounting for all stakeholders, is vital for business professionals.

02

Domain knowledge is indispensable for feature identification, result interpretation, and real-world implications.

Business Perspective

Company (Operations & Strategy):	Optimizes resource allocation, dynamic pricing, … , strategic growth.
Drivers:	Clearer expectations for time and income management. Potential increase in earning.
Passengers (Customers):	Delivers improved ETA accuracy and a better overall experience.
Investors/Shareholders:	Demonstrates a strong, data-driven business model and competitive advantage

Reflections & Potential Improvements

Reflections & Potential Improvements

- **Key Learnings:** CRISP-DM's value, outlier impact, crucial data prep, ensemble model power, and the performance vs. interpretability trade-off.
- **Improvements:**
    - Refine workflow planning.
    - More advanced feature engineering (temporal, geospatial, interactions).
    - Deeper dive into encoding and scaling strategies.
    - Advanced outlier handling.
    - Explore other (ensemble) models (e.g., Gradient Boosting).
    - Detailed error analysis for targeted enhancements.