



MEXICO CITY ECOBIKES

PROJECT 4
BOOTCAMP

BYRONT URCID
CARLOS SÁNCHEZ
LUIS GALÍNDEZ





ECOBICI: Mexico City's Bike-Sharing System

- **Launched in 2010** with 85 stations and 1,114 bicycles
- **Operated by** Mexico City's government through the Ministry of Mobility
- **Goal:** To integrate bicycles into the public transport system as a sustainable and efficient option



How ECOBICI Has Grown

2025 stats:

- 689 stations
- 9,300 bicycles
- 70,000 average weekday trips
- Over 165,000 active yearly memberships
- **Coverage:** Active in 4 boroughs: Cuauhtémoc, Miguel Hidalgo, Benito Juárez, and Coyoacán



Where Our Data Comes From

Source: [Datos Abiertos CDMX \(Mexico City Open Data Portal\)](#)

Dataset: ECOBICI trip data

Includes:

- Start & end time, duration
- Origin & destination stations
- User type (annual/pass, occasional)
- Gender and age (when available)

Time ranges used:

- 1 year of data for machine learning models
- 1 month of data for Tableau visualizations

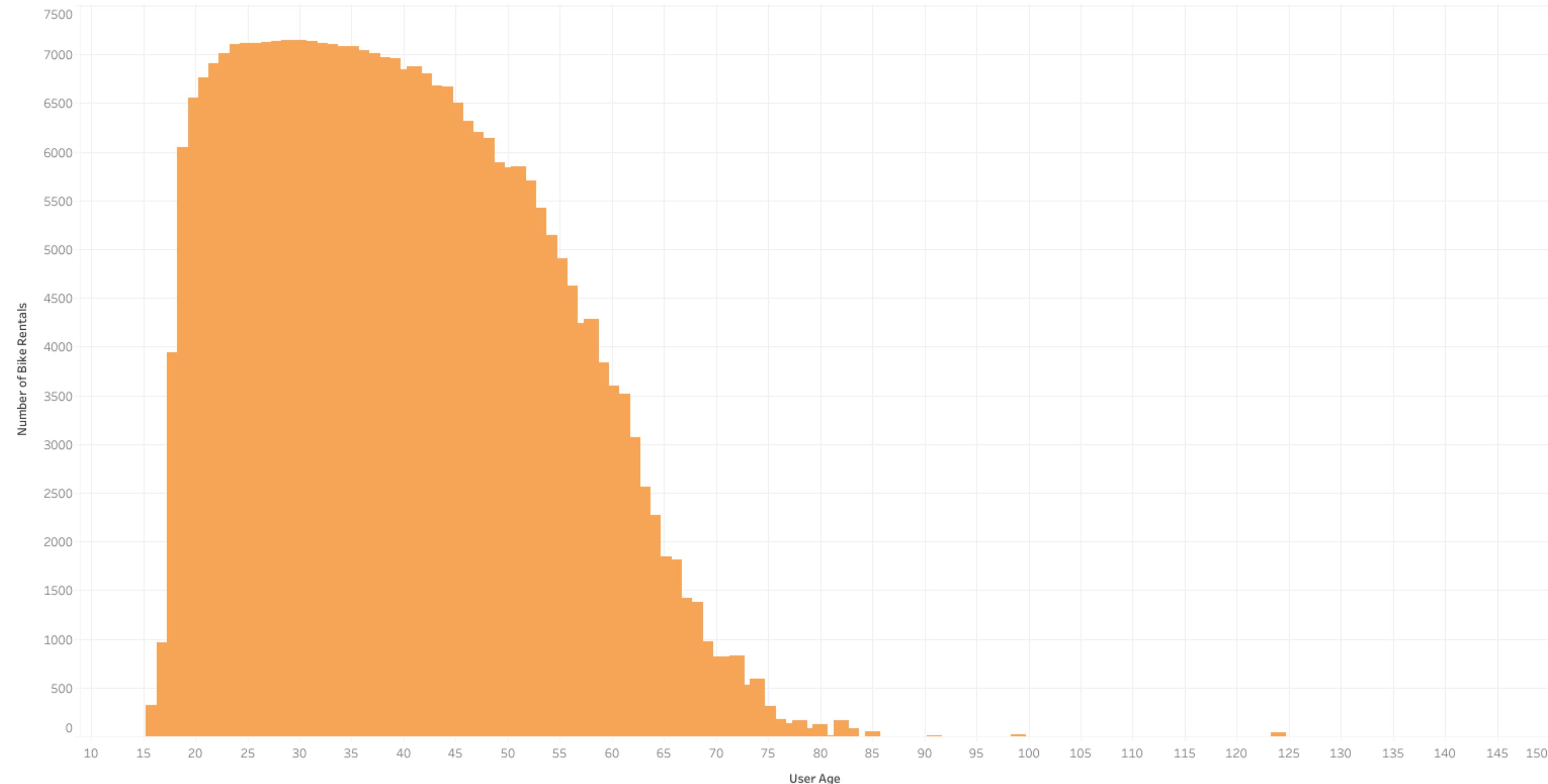




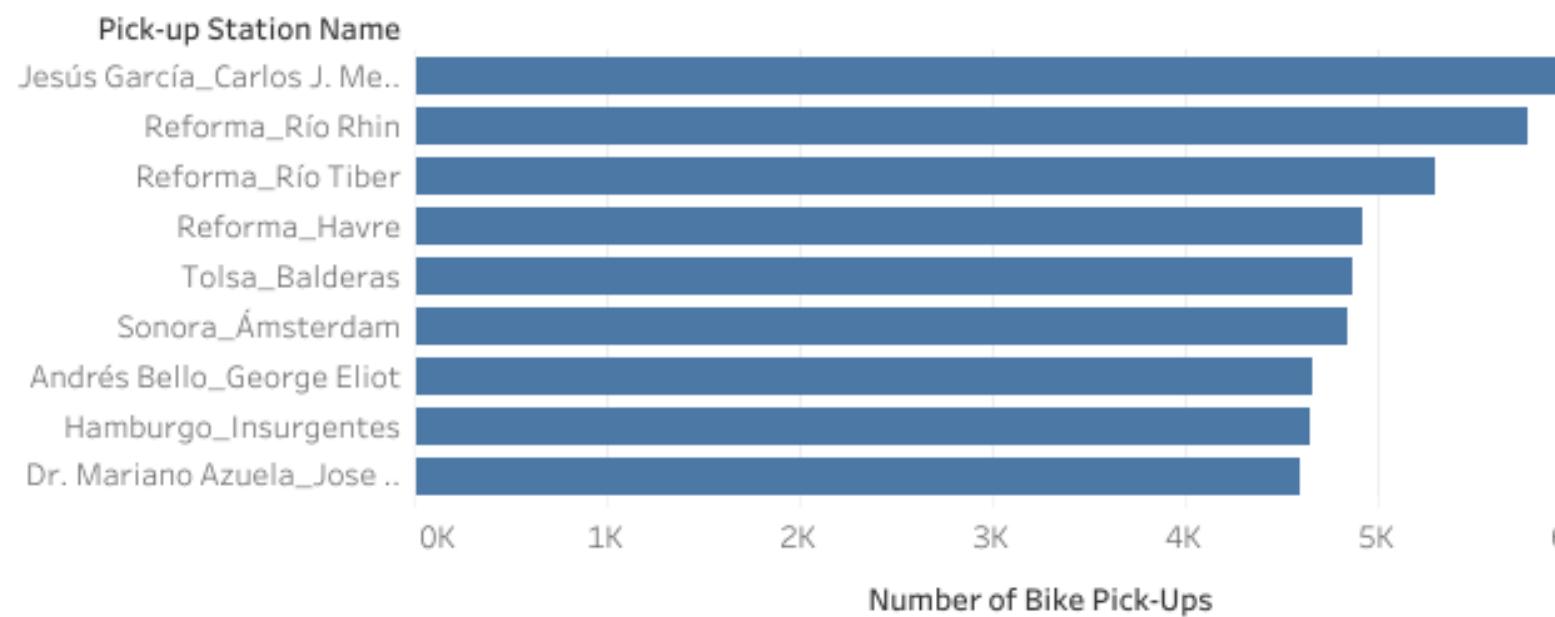
How is ECOBICI being used?

To better understand how the system is being used across the city, we analyzed recent usage data. We explored user behavior, peak hours, the most active stations, and user demographics.

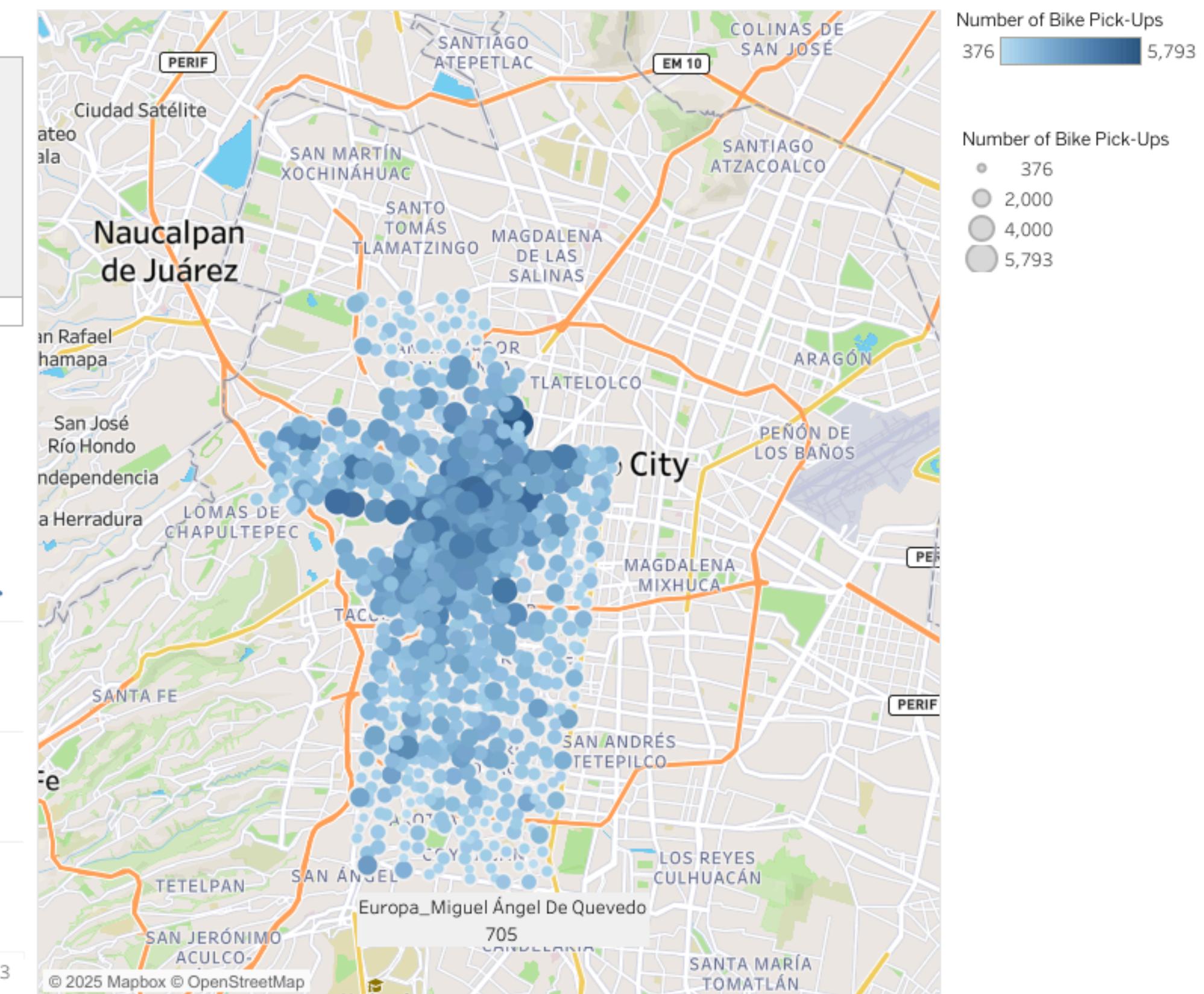
Age Distribution of Bike Share Users



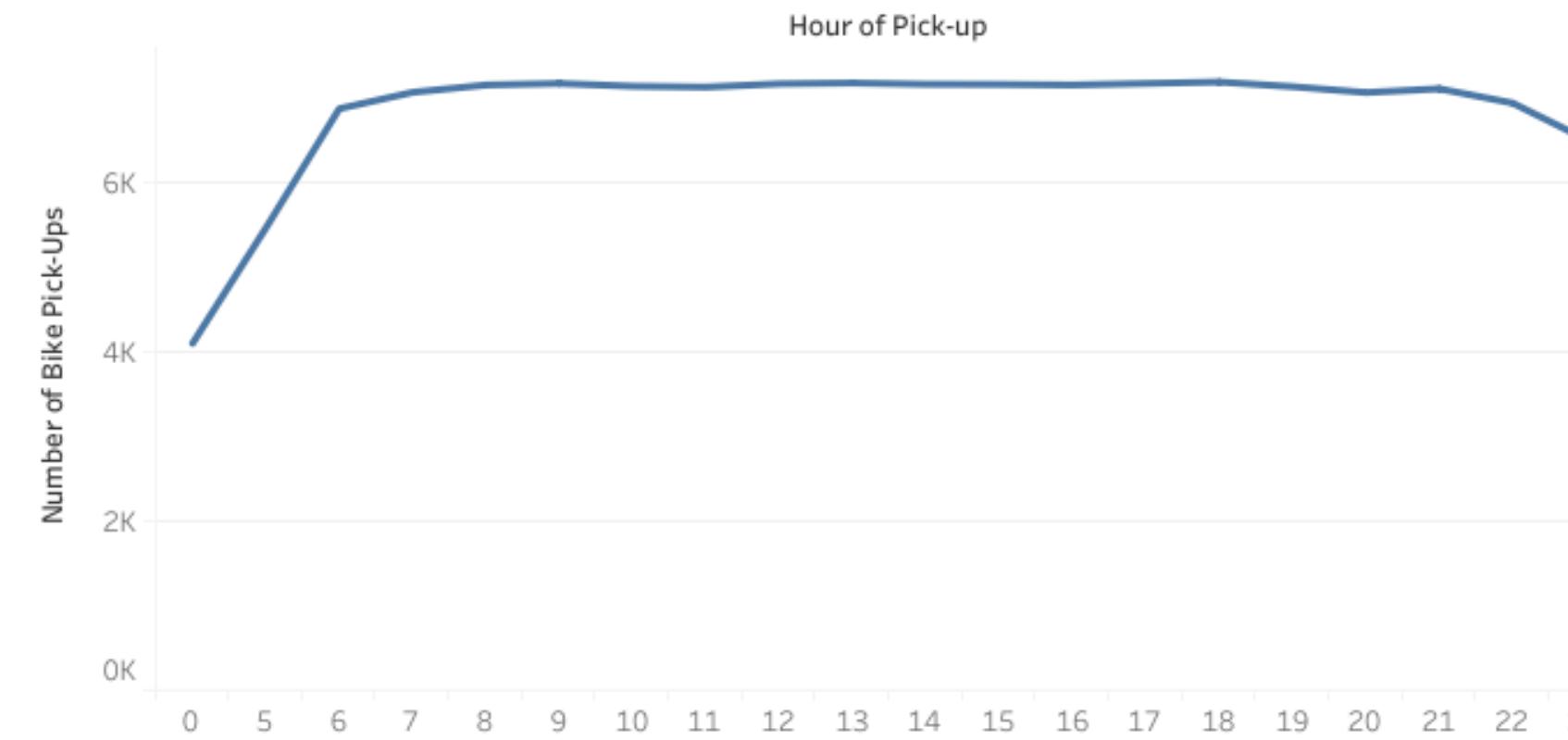
Top 10 Most Frequent Bike Pick-Up Stations



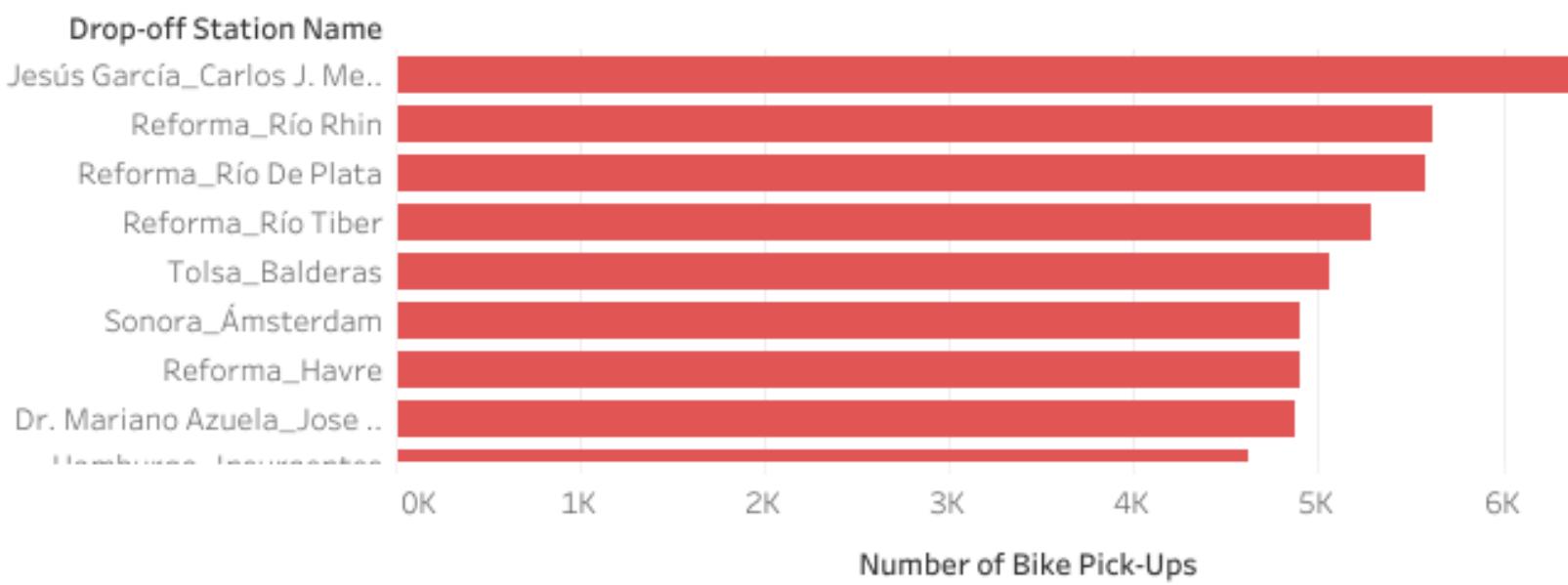
Spatial Distribution of Bike Pick-Ups in Mexico City



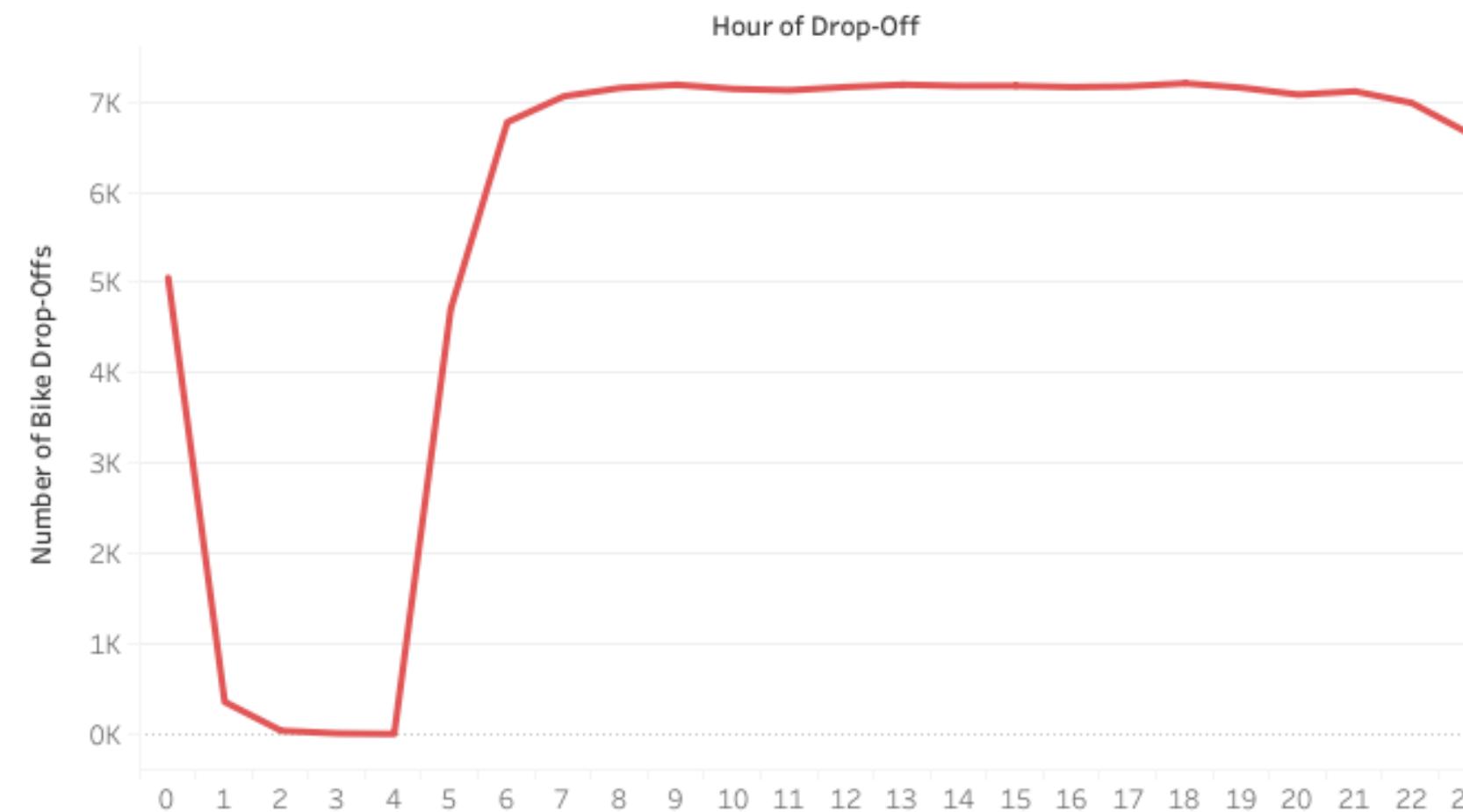
Hourly Distribution of Bike Pick-Ups



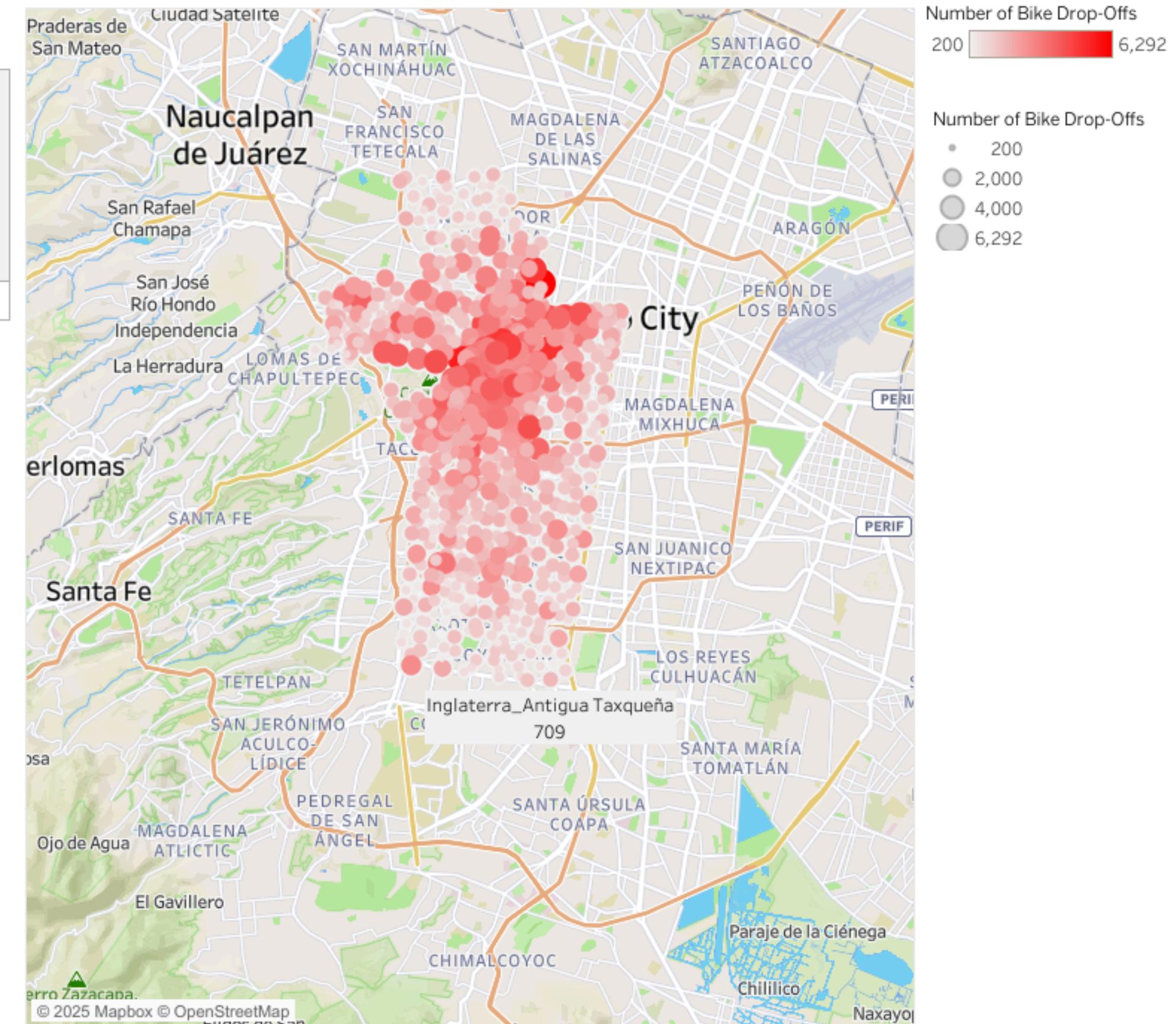
Top 10 Most Frequent Bike Drop-Off Stations



Hourly Distribution of Bike Drop-Offs



Spatial Distribution of Bike Drop-Offs in Mexico City



Machine Learning Applications

1. Bicycle Demand Prediction: Analyzes pickup and drop-off patterns to predict which stations will require the most bikes at certain times of the day.
2. Route Analysis: We will attempt to create a profile of the most used routes and how they vary by time of day, which could help improve the station system.
3. User Classification: Uses demographic data (gender and age) to predict user categories and usage patterns (for example, is a younger user more likely to use a bike at night?).
4. User Segmentation: Groups users into different segments based on their usage behavior, which can help with marketing campaigns or service improvements.
5. Trip duration



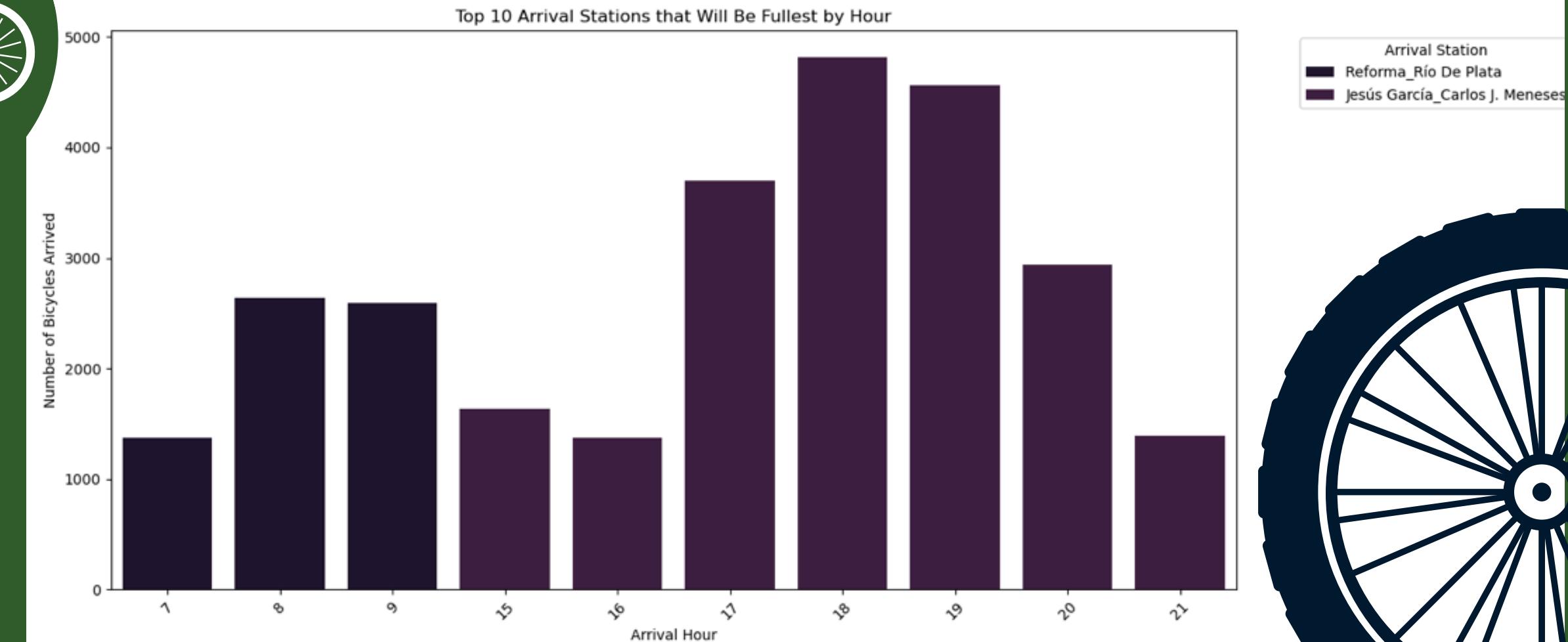
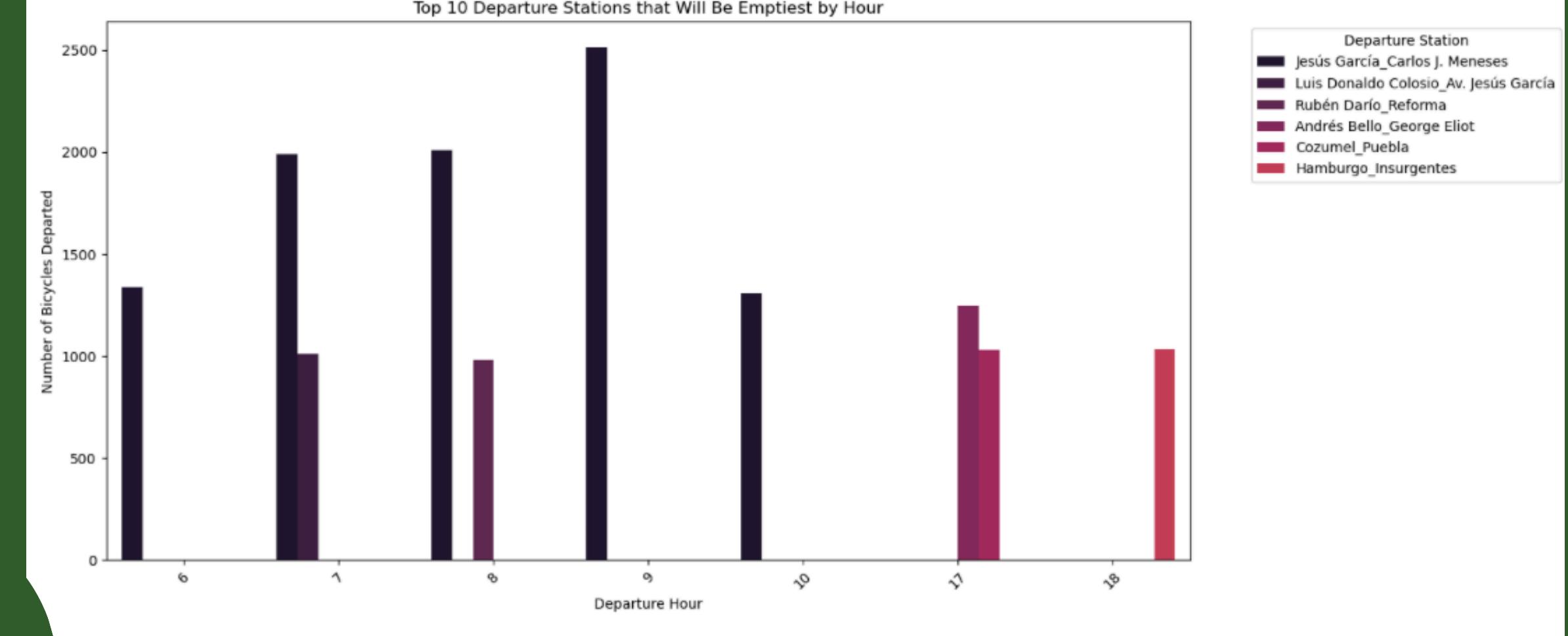
1. Bicycle Demand Prediction

For this model we used the Random Forest Regressor Method



The model was trained with information of 1,745,620 millions of trips

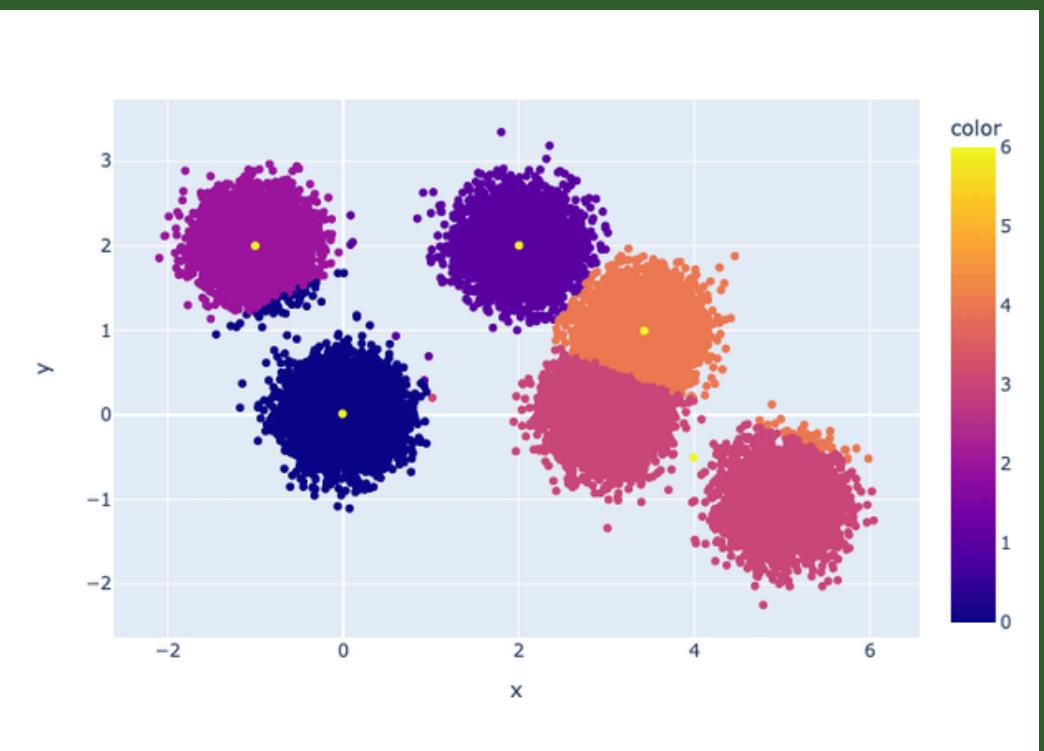
Departure Stations Model Accuracy: **77%**
Arrival Stations Model Accuracy: **83%**



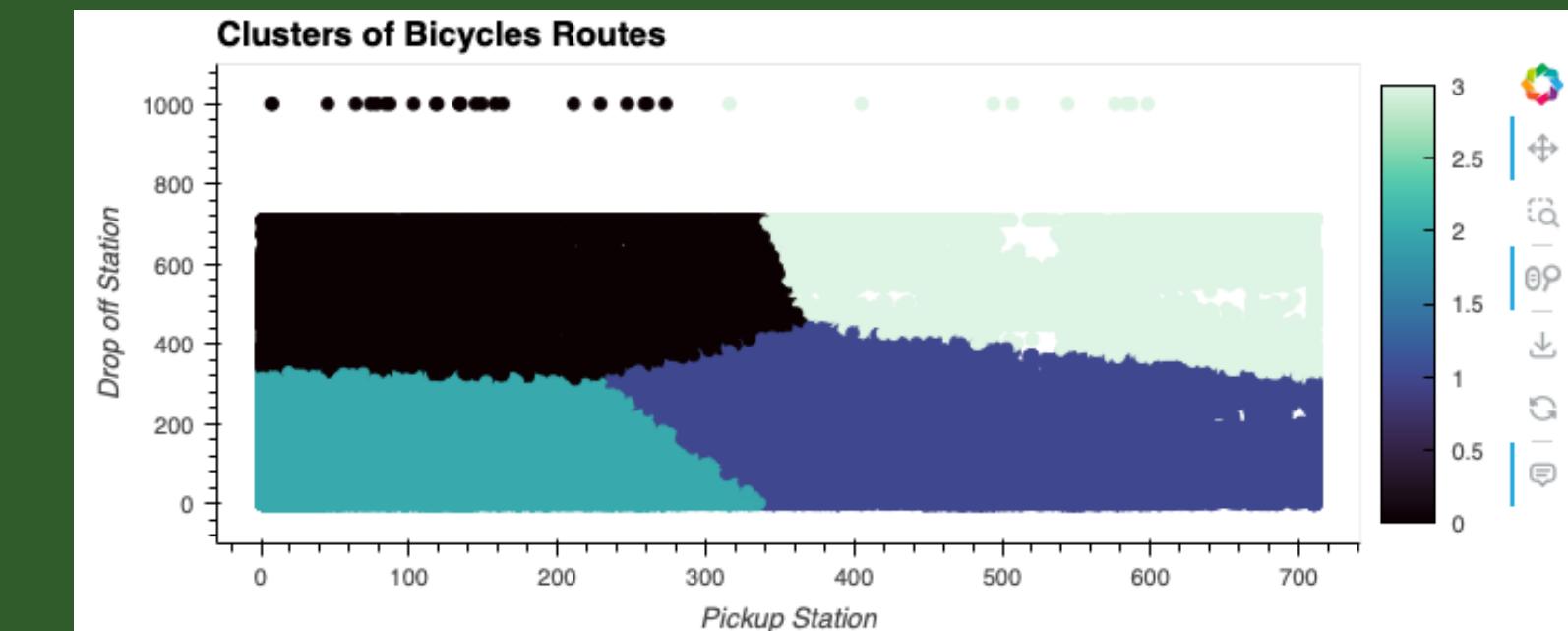
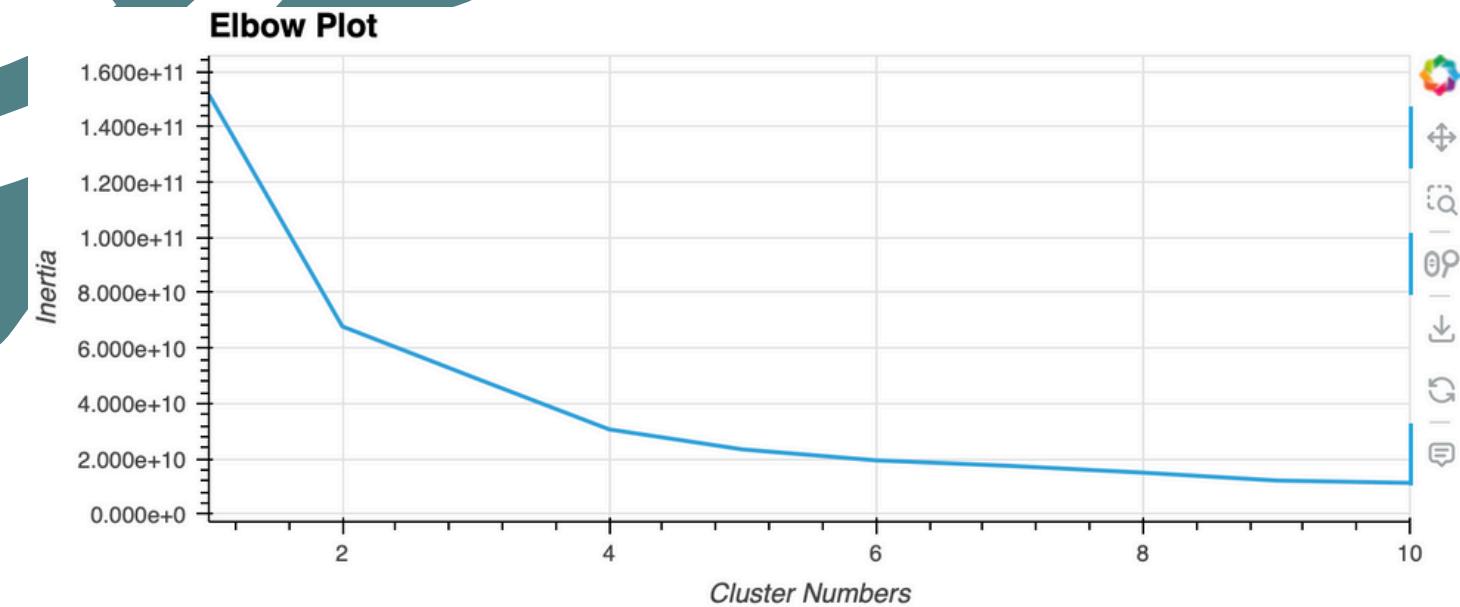
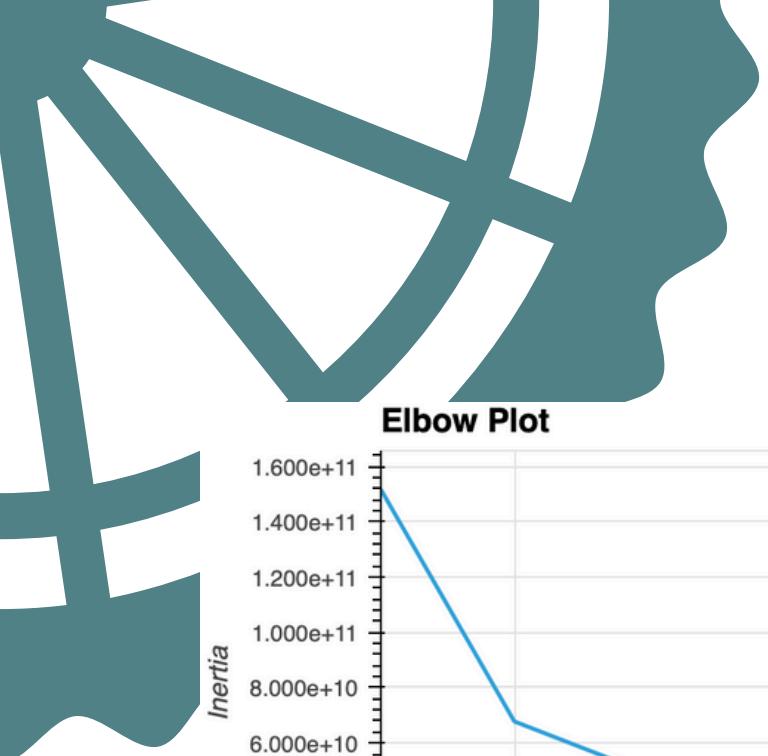
2. Route Analysis



For this model we used the K Means clustering model.



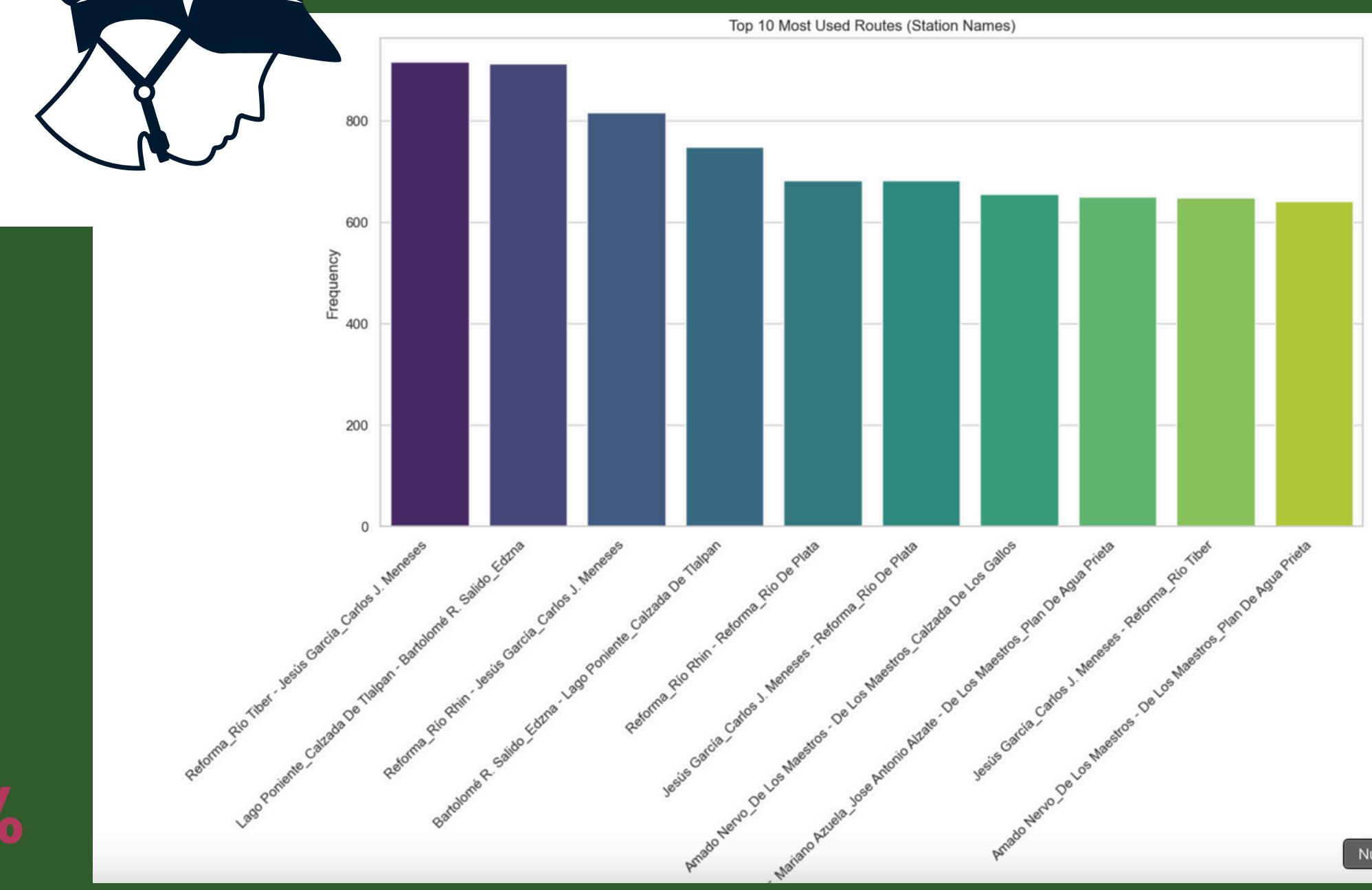
Most used routes prediction



Top 10 Most used predicted routes by the model

TOP 10 ROUTES		Frequency
Route_Names		Frequency
Reforma_Río Tiber - Jesús García_Carlos J. Meneses	917	
Lago Poniente_Calzada De Tlalpan - Bartolomé R. Salido_Edzna	913	
Reforma_Río Rhin - Jesús García_Carlos J. Meneses	817	
Bartolomé R. Salido_Edzna - Lago Poniente_Calzada De Tlalpan	748	
Reforma_Río Rhin - Reforma_Río De Plata	683	
Jesús García_Carlos J. Meneses - Reforma_Río De Plata	682	
Amado Nervo_De Los Maestros - De Los Maestros_Calzada De Los Gallos	656	
Dr. Mariano Azuela_Jose Antonio Alzate - De Los Maestros_Plan De Agua Prieta	651	
Jesús García_Carlos J. Meneses - Reforma_Río Tiber	648	
Amado Nervo_De Los Maestros - De Los Maestros_Plan De Agua Prieta	642	

Accuracy of the K Means model: **67%**

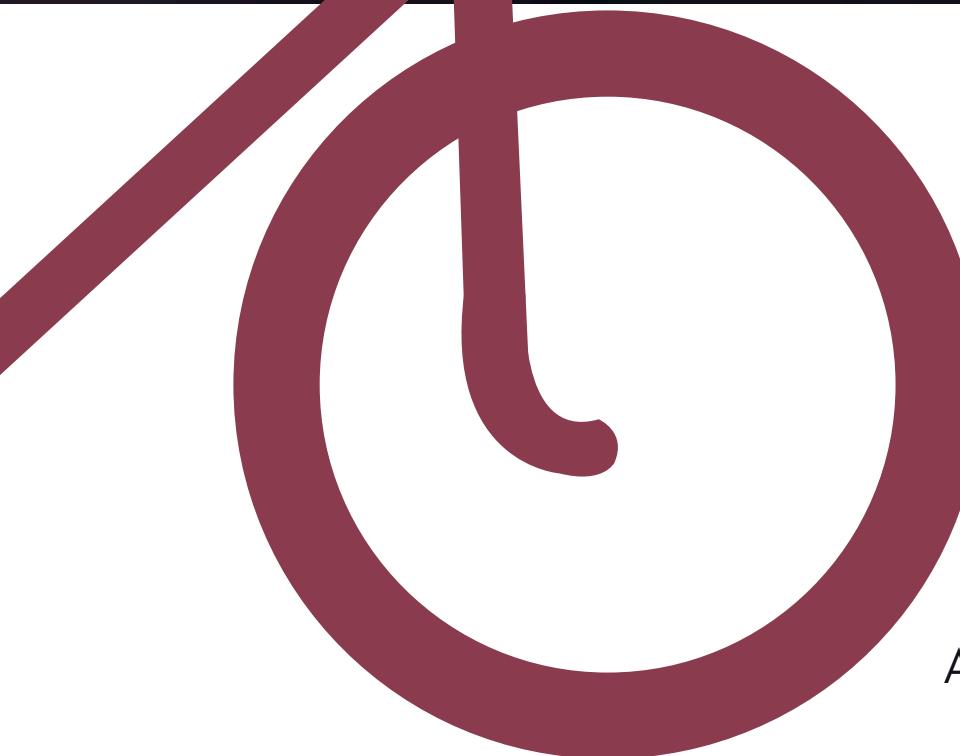


3. User Classification

For this model we used the Random Forest Regressor Method

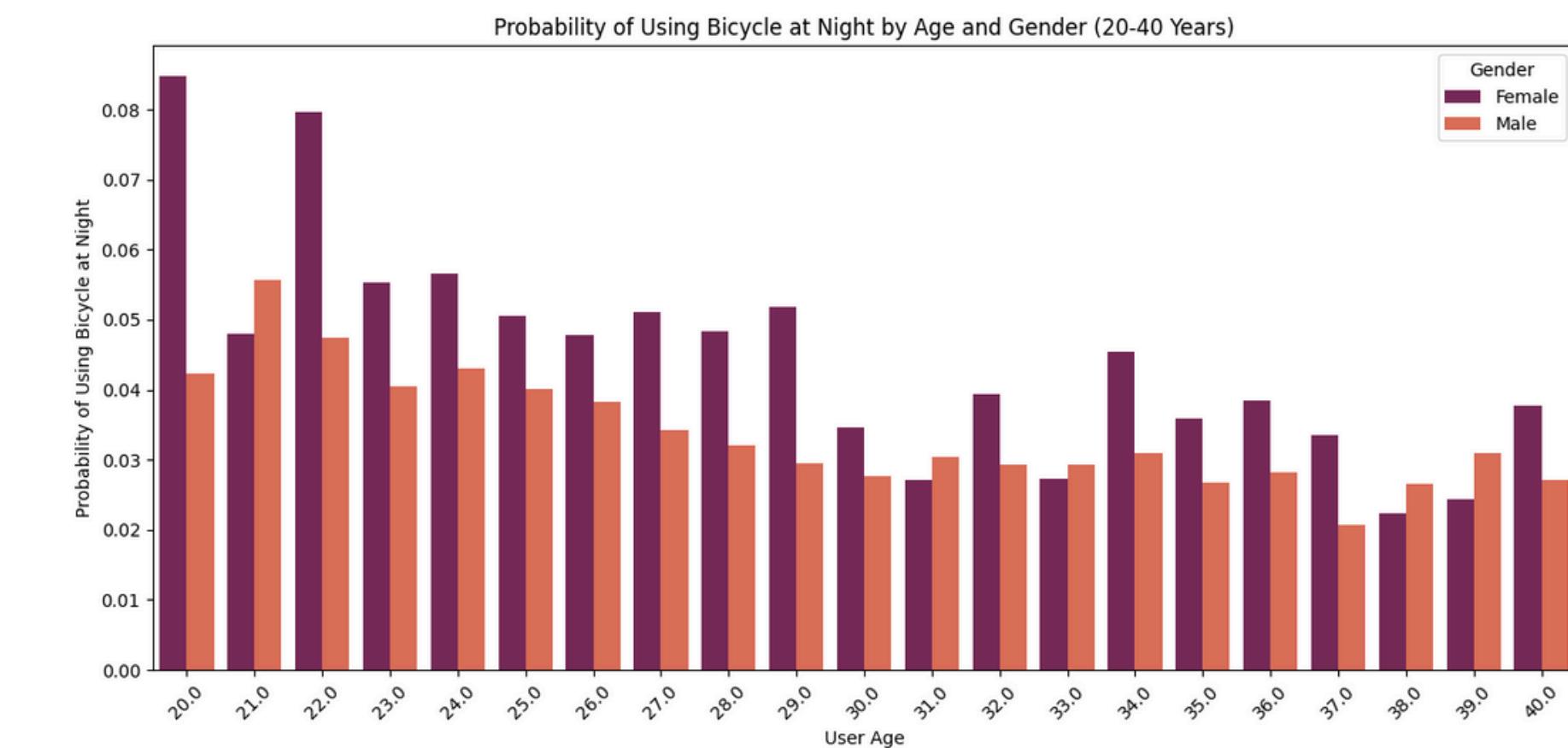


Usage Patterns predictions by demographics



Accuracy of the Random Forest
Regression Model:

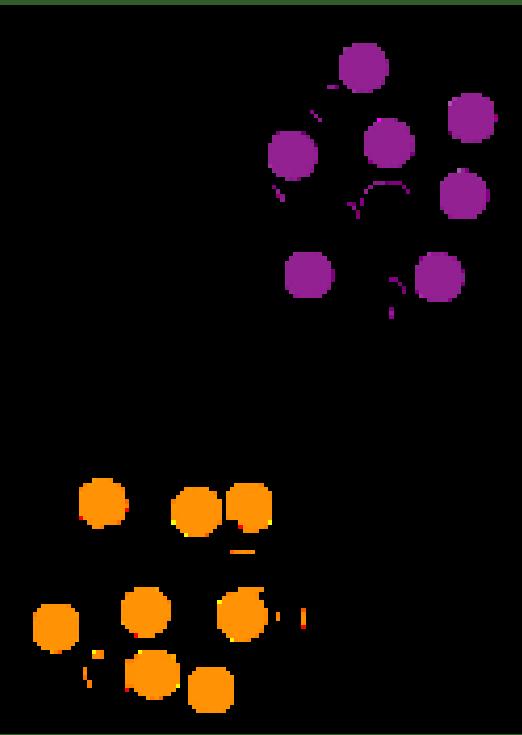
96%



To better understand how the system is being used across the city, we analyzed recent usage data. We explored user behavior, peak hours, the most active stations, and user demographics.

4. User Segmentation

For this model we used the K Means clustering model.



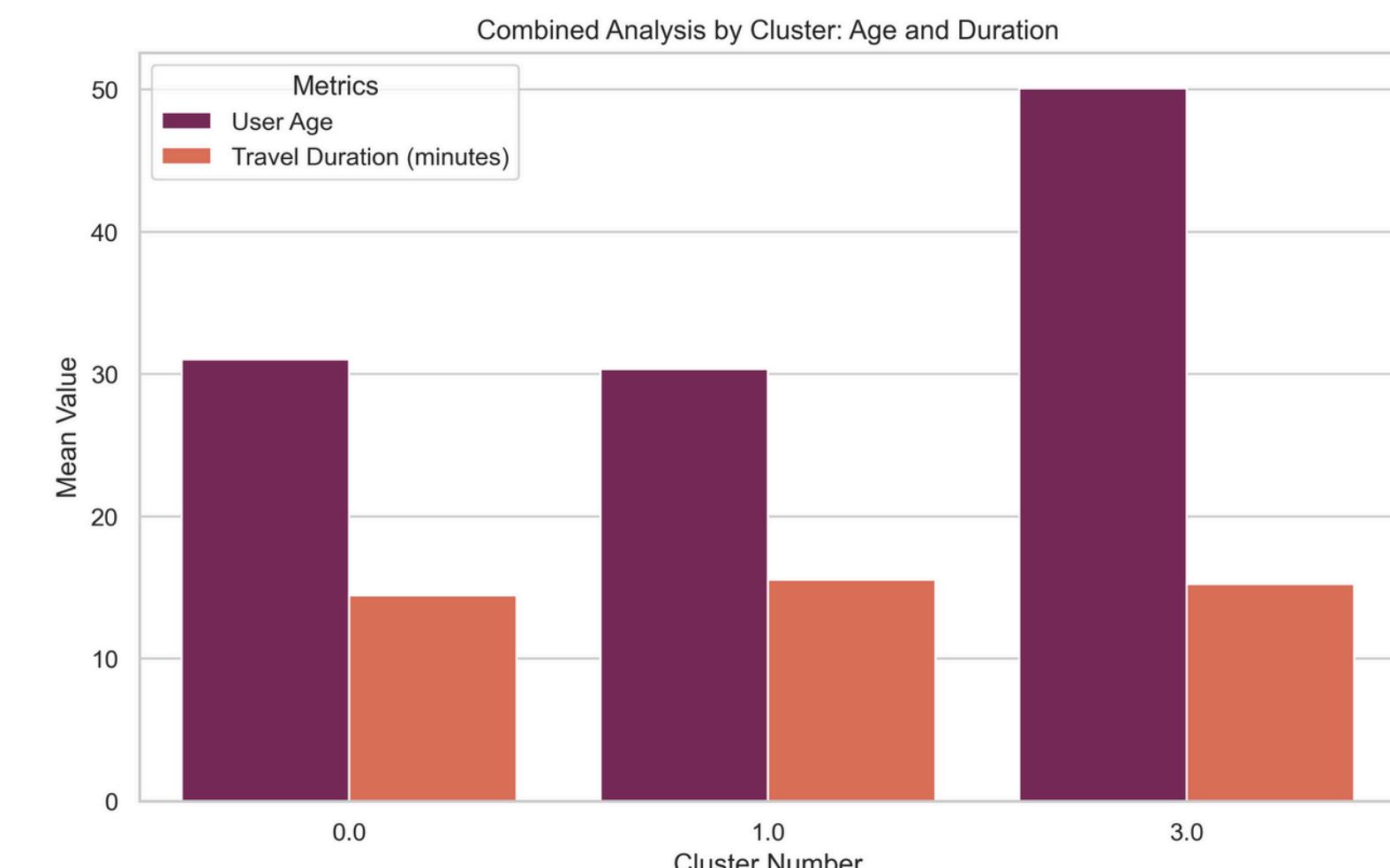
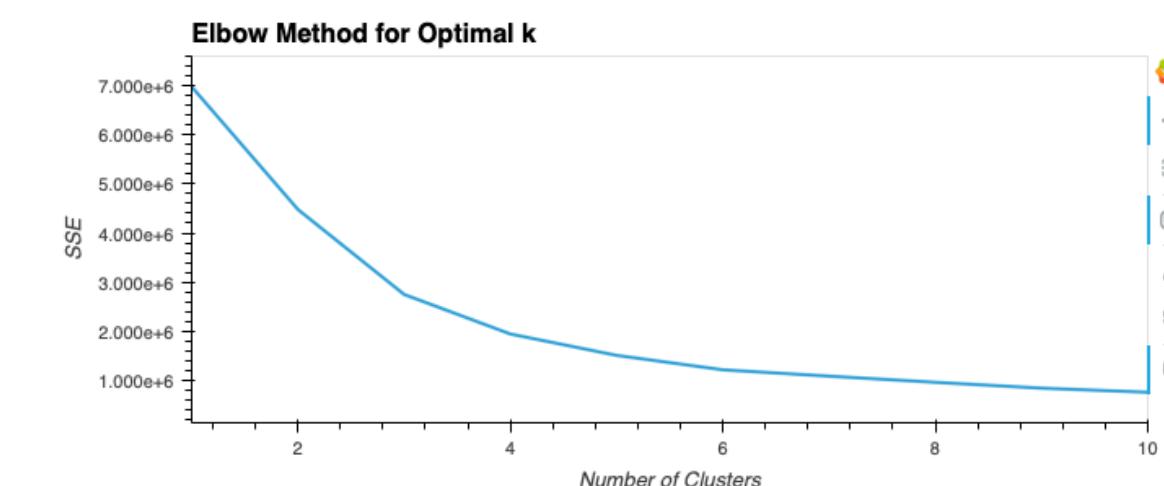
User Segmentation

Groups users into different segments based on their usage behavior, which can help with marketing campaigns or service improvements.



Accuracy of the K Means clustering model for user segmentation:

71%

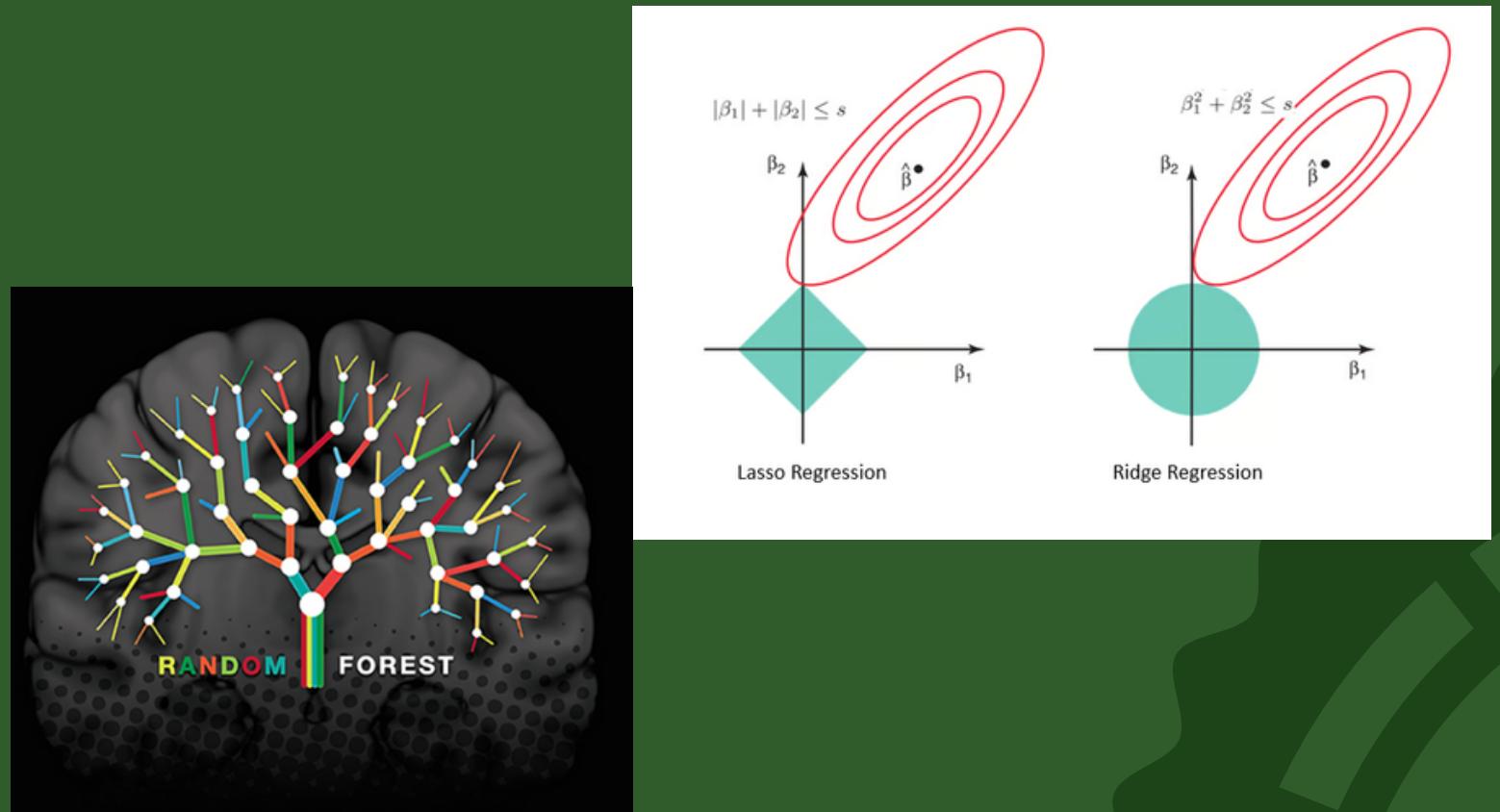


5. Trip duration



For this intend we use 5 different approaches to get a different perspective:

- a) Linear Regression
- b) Ridge
- c) Lasso Regression
- d) Decision Tree Regression
- e) Random Forest Regressor



SPLITTING THE DATASET

To compare how different machine learning models perform, we first split the dataset: one part for training and the other for testing.



pandas: For handling tabular data (reading CSVs, manipulating DataFrames, etc.).

LabelEncoder (from `sklearn.preprocessing`): Allows you to transform categorical variables into numeric values (e.g., convert 'M' / 'F' to 0 / 1).

`train_test_split` (from `sklearn.model_selection`): Facilitates randomly splitting a DataFrame (or arrays) into training and test sets.



After reading the dataset, information was taken from the trip start date and time columns, as well as the trip end date and time, to create a "trip duration" column.



Outlayers were removed from this dataset to improve travel time prediction results.

(Only trips longer than 10 minutes and shorter than 120 minutes)



Apply one-hot encoding to Hour_Day and Day_Week. This means that instead of having a column with numeric values (0–23 for the hour, 0–6 for the day), binary (dummy) columns are created that indicate "true/false" for each possible value.



The data was split into 80% training (X_{train} , y_{train}) and 20% test (X_{test} , y_{test}).

The data was then exported into pickle files.

TRAINING SOME MODELS

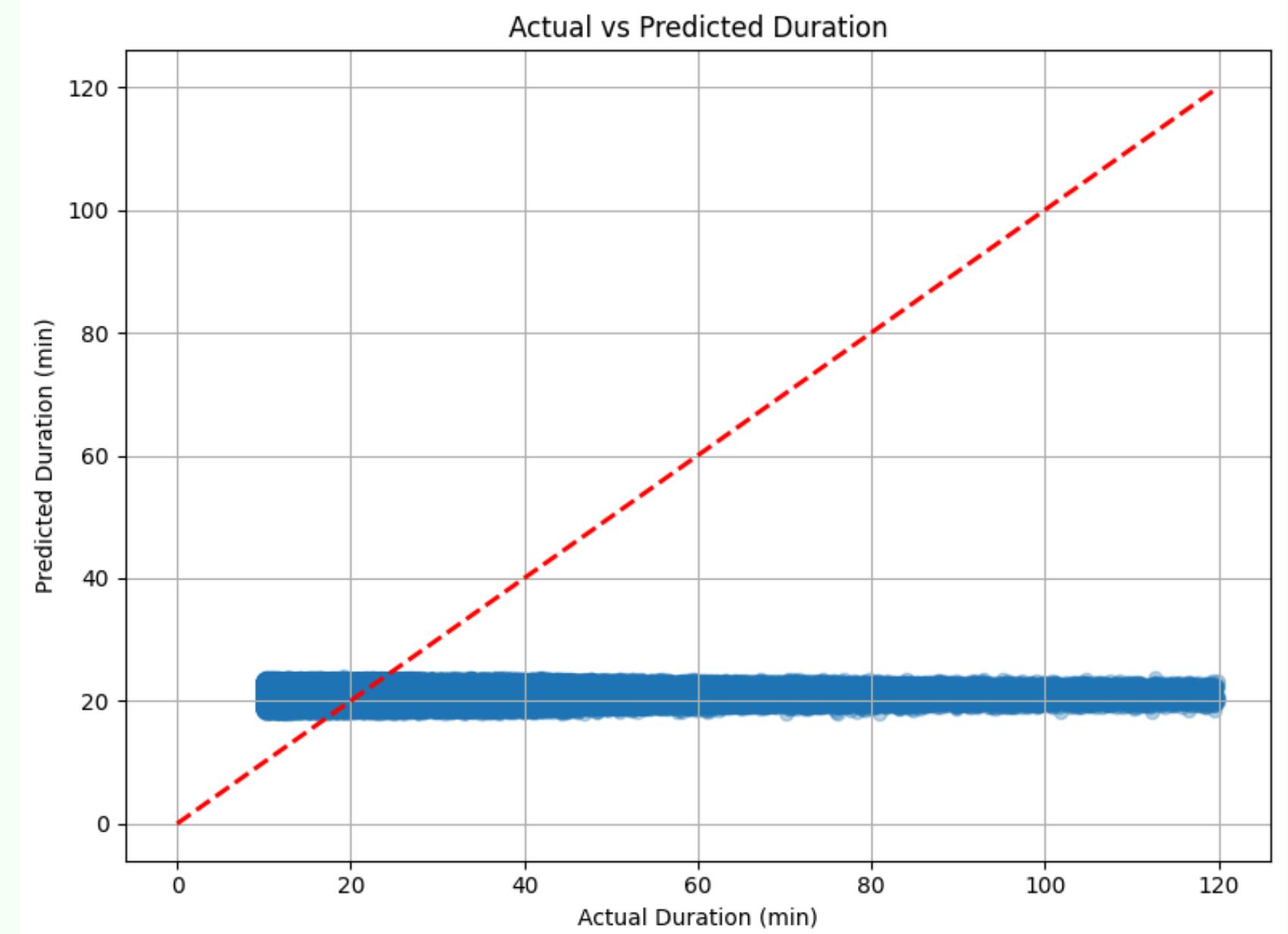
LINEAR REGRESSION

MAE: 7.68 minutes
RMSE: 102.57 minutes
R² Score: 0.0091

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

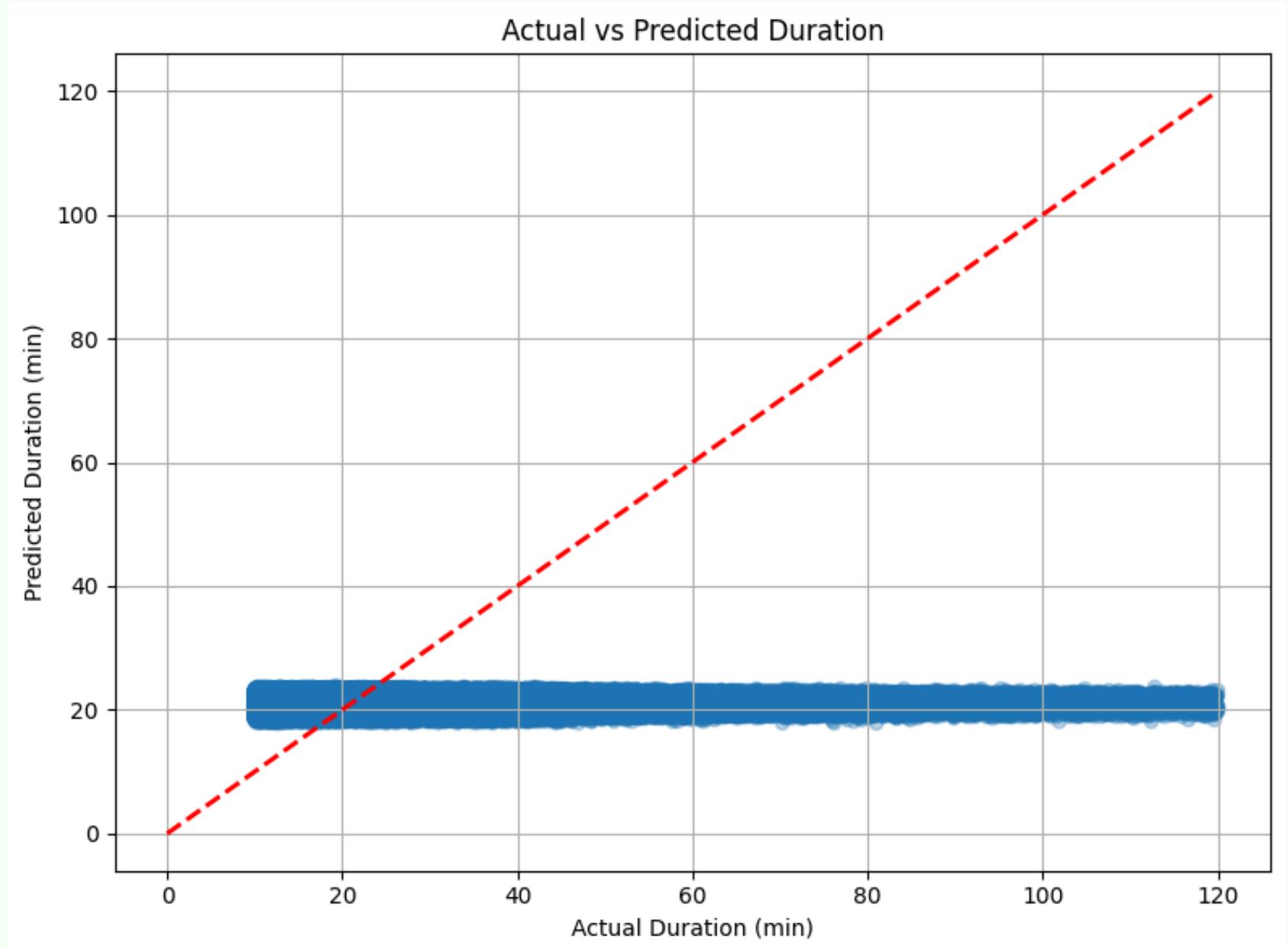
RIDGE

MAE: 7.68 minutes
RMSE: 102.57 minutes
R² Score: 0.0091

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

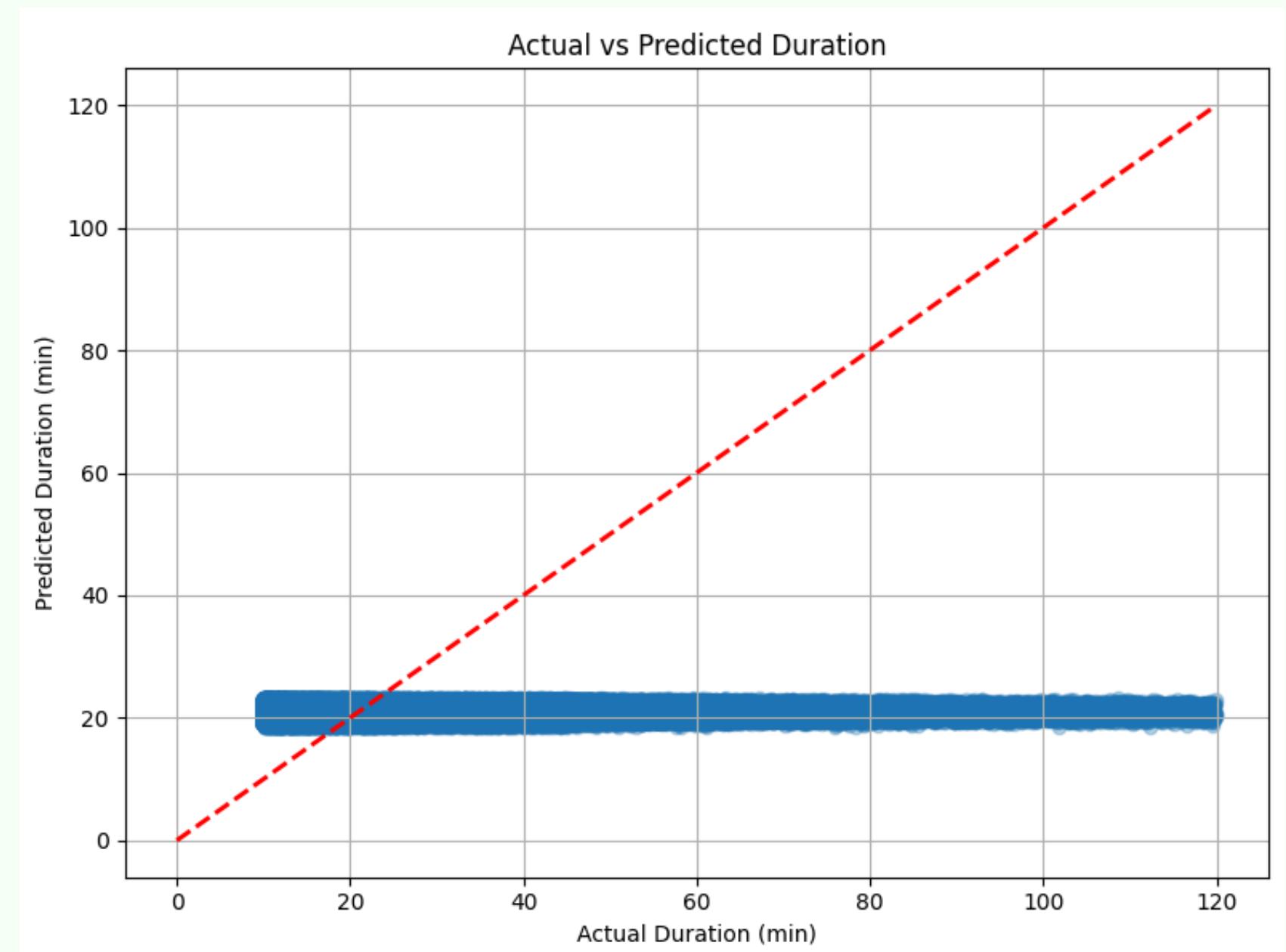
LASSO REGRESSION

MAE: 7.68 minutes
RMSE: 102.60 minutes
R² Score: 0.0088

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

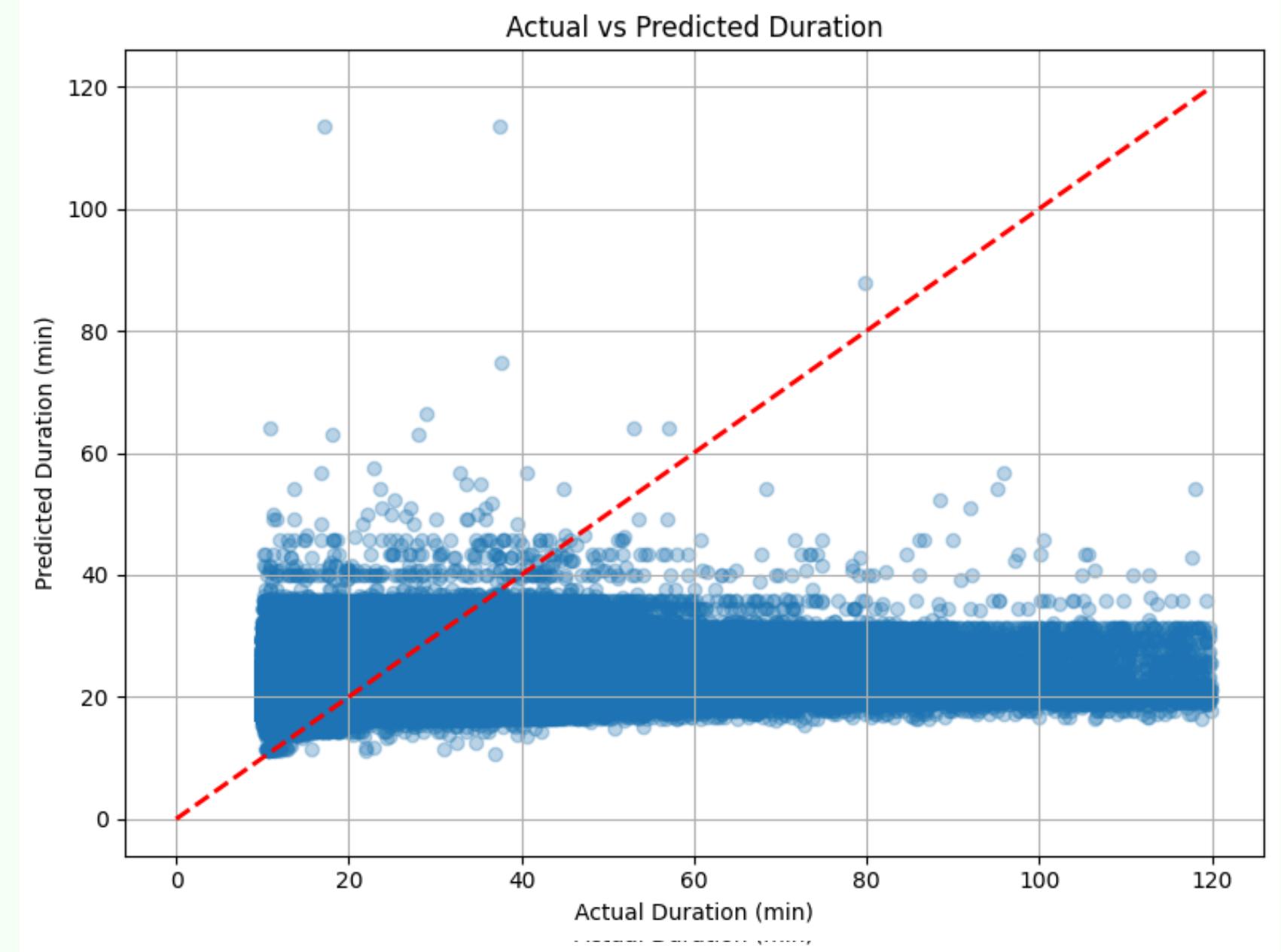
DECISION TREE REGRESSOR

MAE: 7.49 minutes
RMSE: 98.83 minutes
R² Score: 0.0453

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

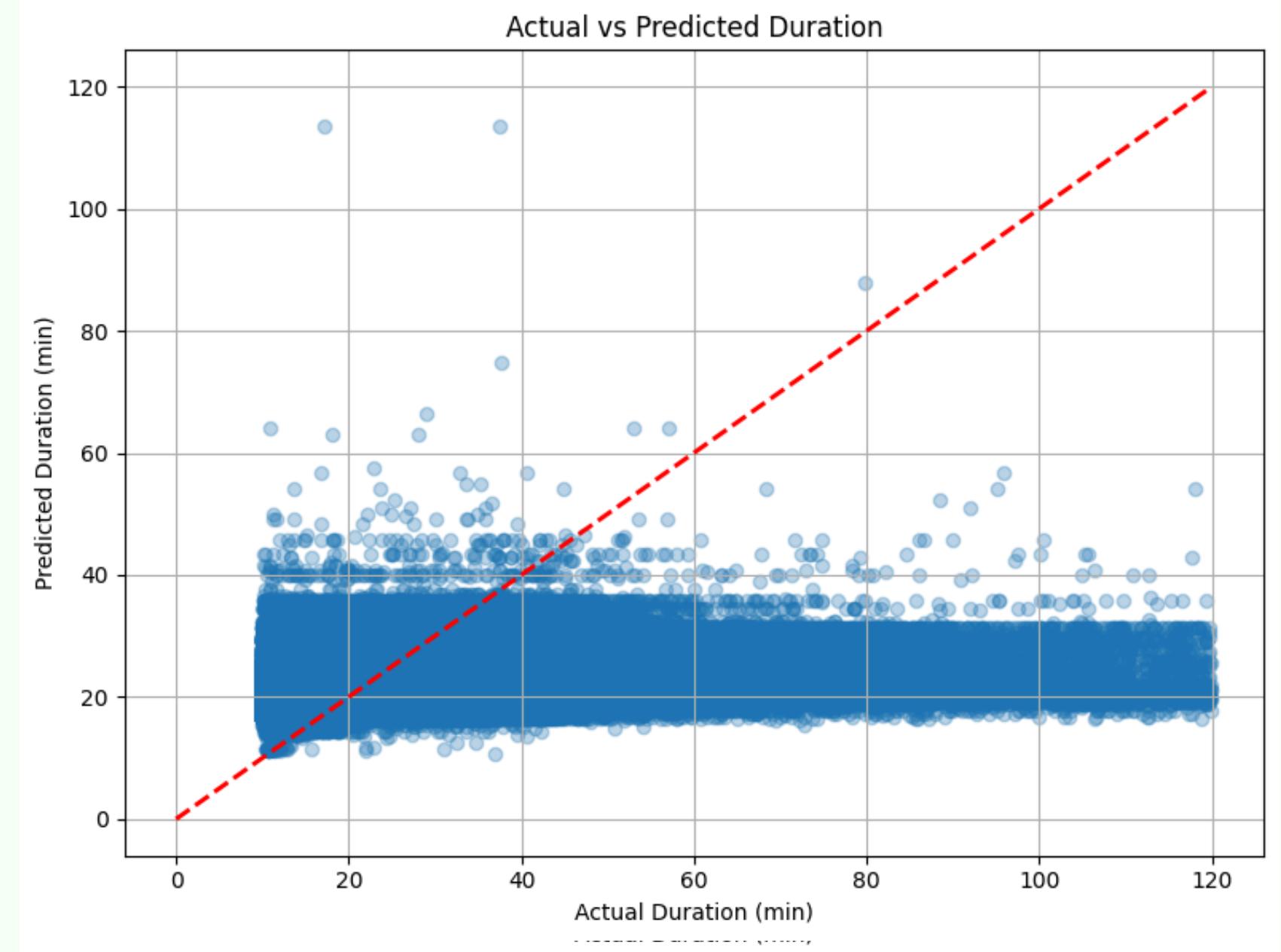
RANDOM FOREST REGRESSOR

MAE: 7.04 minutes
RMSE: 92.49 minutes
R² Score: 0.1065

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

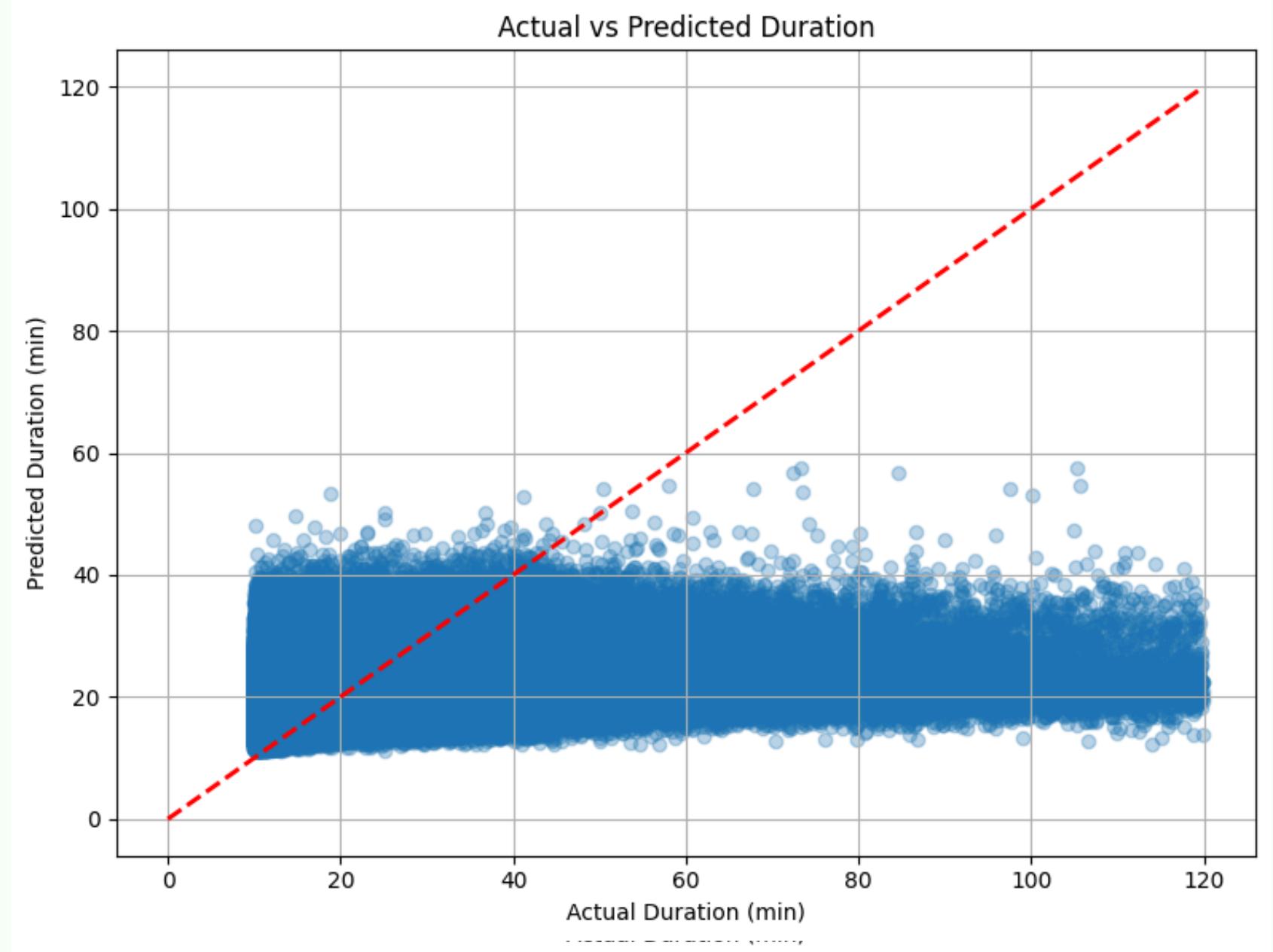
RANDOM FOREST REGRESSOR

MAE: 7.04 minutes
RMSE: 92.49 minutes
R² Score: 0.1065

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).



TRAINING SOME MODELS

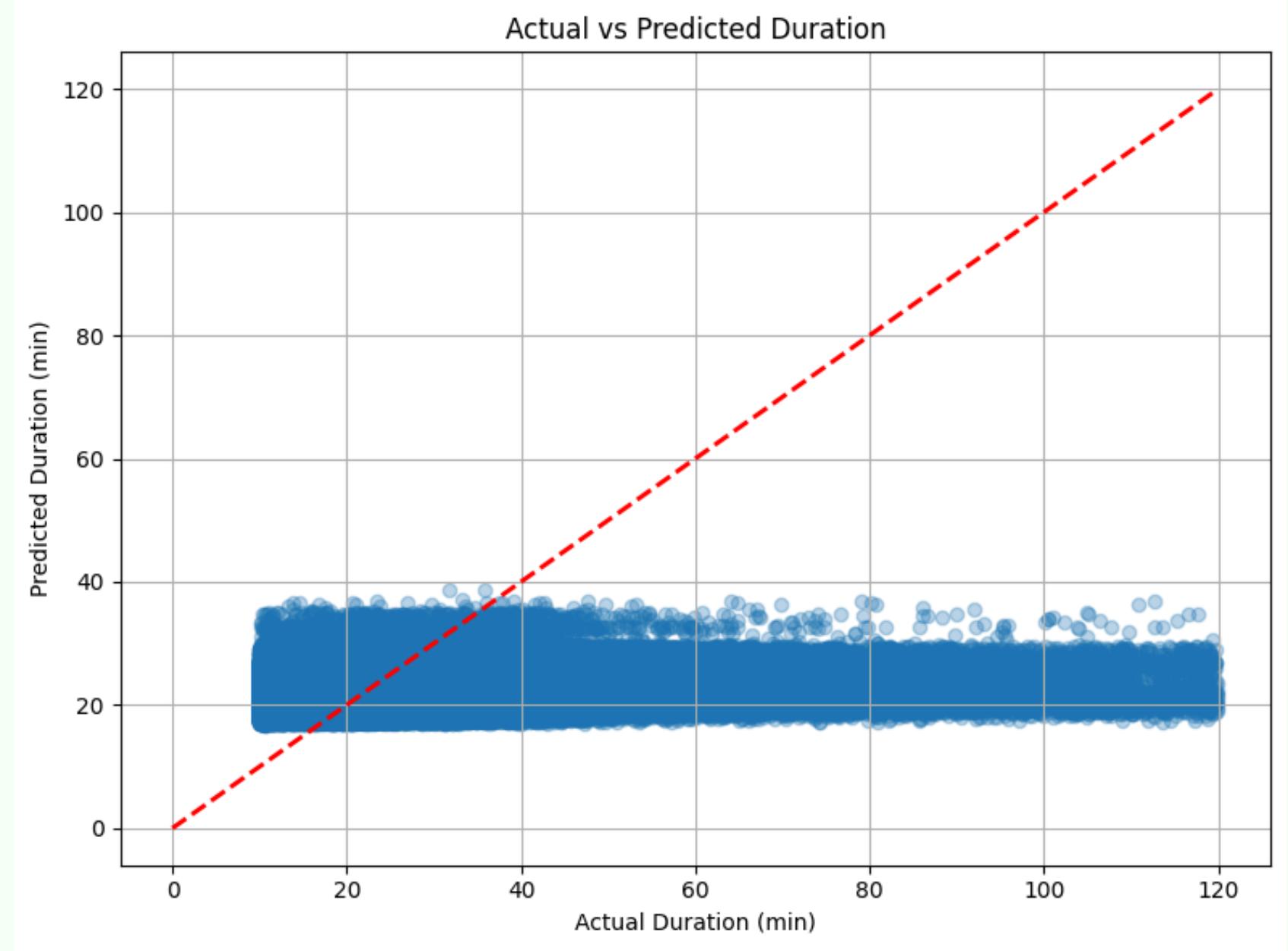
LGBM REGRESSOR

MAE: 7.46 minutes
RMSE: 98.25 minutes
R² Score: 0.0509

MAE (Mean Absolute Error) shows average error in the same units.

RMSE (Root Mean Squared Error) penalizes large errors more heavily.

R² Score (coefficient of determination) shows how much variance in duration is explained by the model (1.0 is perfect, 0.0 means "as good as mean", negative means worse).





Conclusions

Based on the analysis carried out with the different machine learning models, these are some of the areas of opportunity that we believe can be obtained by training and improving the algorithms presented.





Bicycle Demand Prediction



Bicycle redistribution. Understanding which stations tend to be used together or share usage patterns can help with transport logistics and bicycle redistribution.

Identification of high-demand areas

Areas with high concentrations of usage can be identified for the location of new stations or the adjustment of the size of current ones.

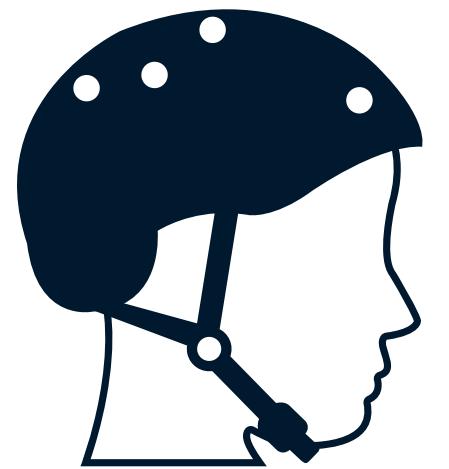




Targeted marketing

Understanding the demographics and patterns of users who use stations in a cluster can inform more effective marketing campaigns.





Ride Duration Insights

Predicting trip duration, you could explore factors influencing ride length, such as distance, time of day, or user experience, indicating potential areas for service improvements.



RESOURCES PAGE

1. Gobierno de la Ciudad de México.
(n.d.). Ecobici.
<https://ecobici.cdmx.gob.mx/>
2. Gobierno de la Ciudad de México.
(n.d.). Mapa de Ecobici.
<https://ecobici.cdmx.gob.mx/mapa/>
3. ECOBICI. (n.d.). Datos abiertos.
Gobierno de la Ciudad de México.
<https://ecobici.cdmx.gob.mx/datos-abiertos/>
4. Wikipedia contributors. (n.d.).
Ecobici (Ciudad de México).
Wikipedia.
[https://es.wikipedia.org/wiki/Ecobici_\(Ciudad_de_M%C3%A9xico\)](https://es.wikipedia.org/wiki/Ecobici_(Ciudad_de_M%C3%A9xico))





THANK YOU!

[HTTPS://GITHUB.COM/LORD-BYRONT/ML-MEXICO-CITY-BIKES/](https://github.com/lord-byront/ml-mexico-city-bikes/)