LA METRO BIKE SHARE

HOW DATA ENHANCES STATION ANALYTICS

by Oliver Bohler

Capstone Project 2

Springboard DataScience

WORKFLOW

The data for this project was collected from public available websites and databases. the main objectives are:

01 Merging Data from various tables

02 Cleaning the Data

03 Exploratory Data Analysis

04 Splitting the Datasets

05 Feature Engineering

06 Machine Learning and Modeling

THE DATA

Total Trips	Passes Sold
2,617,542	364,617

Calories Burned	Miles Traveled	Emissions Reduced Pounds of CO_2
365,230,608	9,130,765	8,674,227

For this project, I web scraped API data to gather real-time dock availability for Metro Bike Share stations. Additionally, I extracted Los Angeles bus and train station locations from publicly available databases to analyze transit connectivity. Using the Metro LA Bike Share trip and station tables, I integrated trip volume and station metadata to enhance the analysis

DATA CLEANING

Data Format

Each .csv file contains data for one quarter of the year. Each file contains the following data points:

- **trip_id:** Locally unique integer that identifies the trip
- **duration:** Length of trip in minutes
- **start_time:** The date/time when the trip began, presented in ISO 8601 format in local time
- **end_time:** The date/time when the trip ended, presented in ISO 8601 format in local time
- **start_station:** The station ID where the trip originated (for station name and more information on each station see th *Station Table*)
- **start_lat:** The latitude of the station where the trip originated
- **start_lon:** The longitude of the station where the trip originated
- **end_station:** The station ID where the trip terminated (for station name and more information on each station see th *Station Table*)
- **end_lat:** The latitude of the station where the trip terminated
- **end_lon:** The longitude of the station where the trip terminated
- **bike_id:** Locally unique integer that identifies the bike
- **plan_duration:** The number of days that the plan the passholder is using entitles them to ride; 0 is used for a single ride plan (Walk-up)
- **trip_route_category:** "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips
- **passholder_type:** The name of the passholder's plan
- **bike_type:** The kind of bike used on the trip, including standard pedal-powered bikes, electric assist bikes, or smart bikes.

01 Data Integration and Merging:

The trip data was merged with the station dataset using a left join on Station_ID to ensure all trip records were preserved while incorporating relevant station details. Additionally, API data on bike dock availability was integrated to provide real-time insights into station usage.

02 Data Cleaning and Standardization

Missing values were handled by imputing frequent categories, and station names were standardized for consistency. Anomalous records, such as virtual station trips and extremely short or long durations, were filtered out to ensure data quality.

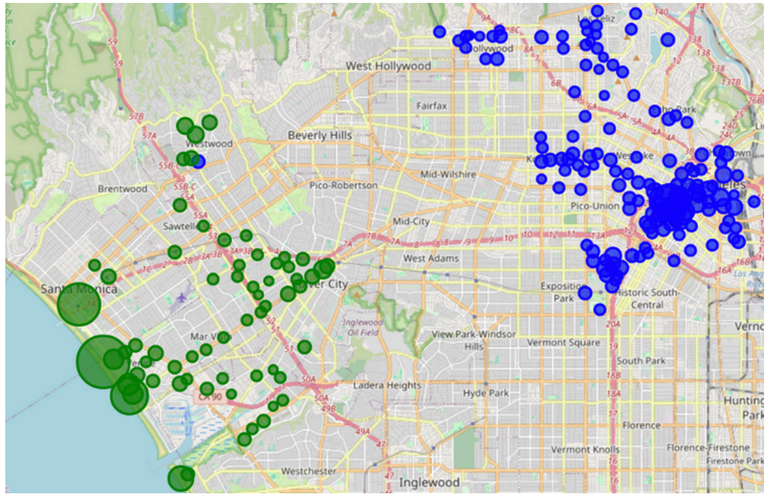
03 Feature Engineering and Preprocessing:

Date components like year, month, and day name were extracted and formatted for time-based analysis. Additional steps included deduplication, correcting inconsistencies, and computing missing pass holder types using mode imputation.

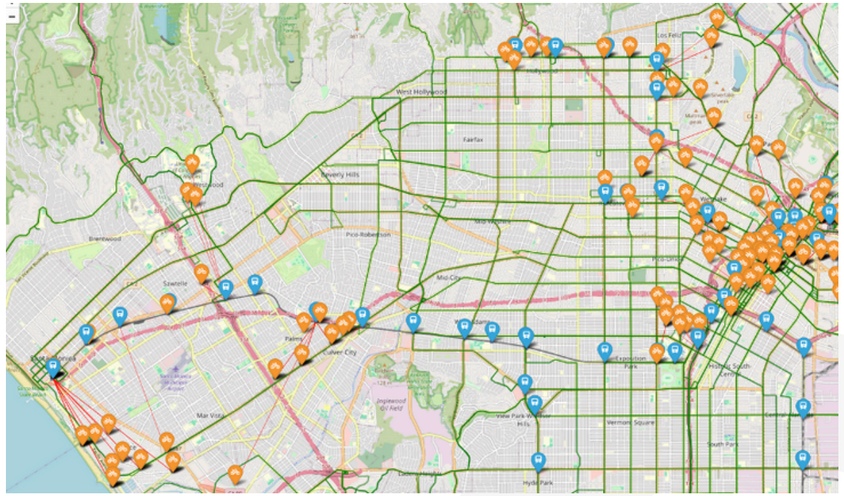
EXPLORATORY DATA ANALYSIS

Map Visualization of the Data

The map shows every active station in the region of Westside (Green) and Downtown LA (Blue) and it's trip volume. This offers a first look into distinct station behavior.



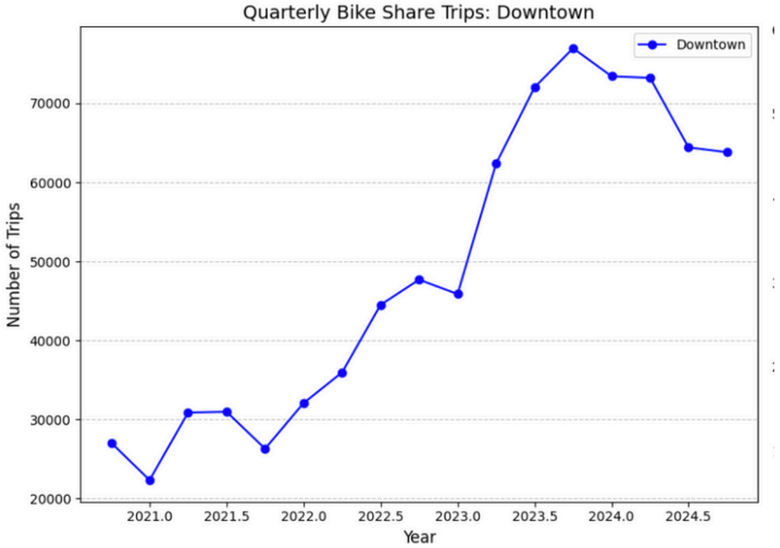
The map shows every active station, including Metro Train Stations (Blue) and Bus Lines(Green). Here the distance of each station to it's nearest Metro station was visualized.

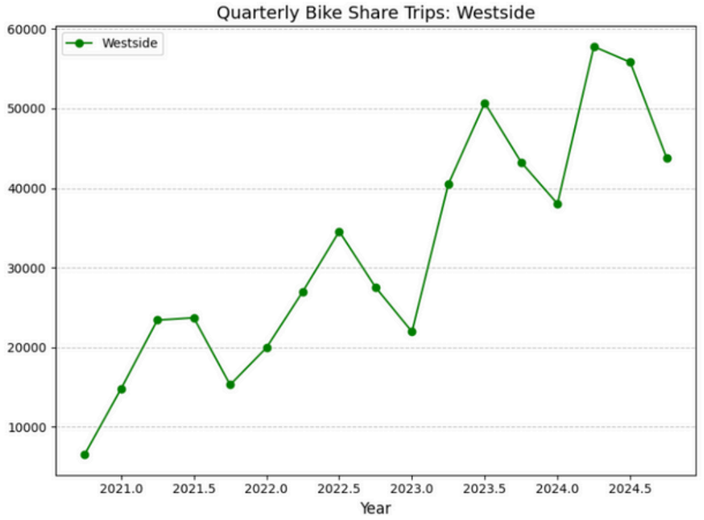


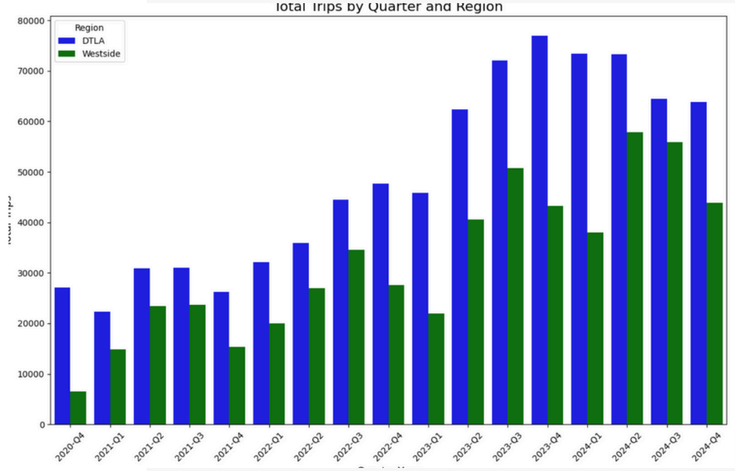
"Public transportation is not a business designed for profit. It exists—especially bike-sharing programs—to encourage citizens to adopt alternative transportation methods and, in some cases, to provide individuals with their only means of mobility."

KEY INSIGHTS

REGIONAL DIFFERENCES IN DATA







Seasonality Downtown LA

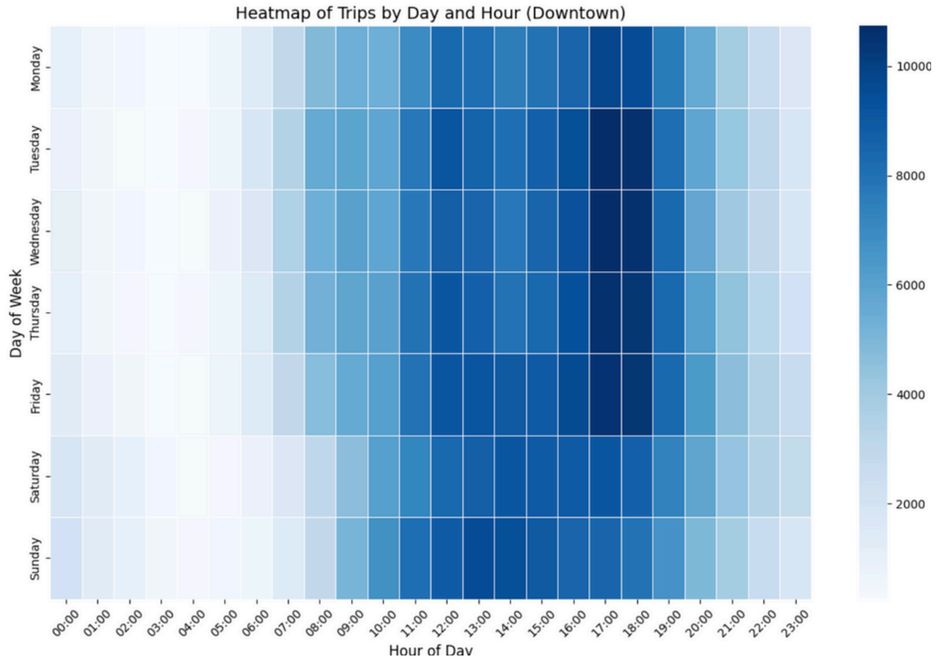
Bike share usage in Downtown LA has shown a steady upward trend over time, with notable peaks in ridership after 2023. This suggests a growing reliance on bike share services in this region, likely influenced by urban mobility initiatives and increased commuter adoption.

Seasonality Westside

The Westside bike share system also demonstrates an increasing trend in usage, but with more fluctuations compared to Downtown LA. These variations may be influenced by tourism patterns, weather conditions, or local events.

Trip Demand per Region

A comparison of quarterly trip volumes highlights that Downtown LA consistently outperforms the Westside in terms of bike share trips. However, the gap has narrowed over time, indicating a rising demand in the Westside.

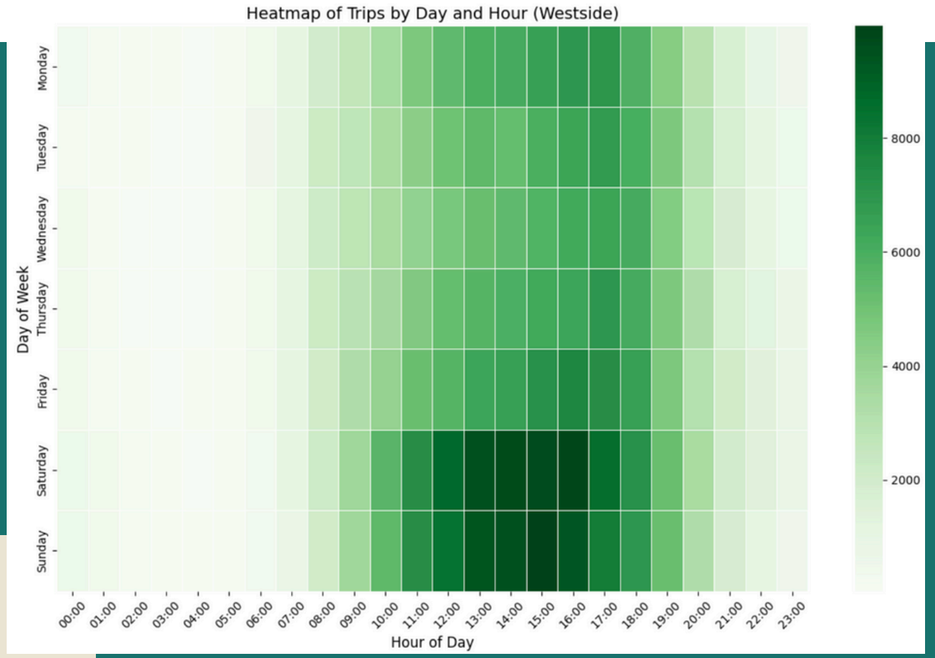


HEAT MAPS

VISUALIZATION OF TRIP DEMAND BY DAY AND HOUR

DOWNTOWN LA:

The heatmap indicates that peak hours for bike trips in Downtown LA are from 8:00 AM to 10:00 AM and 4:00 PM to 6:00 PM, aligning with typical commuter rush hours. The highest activity is observed on weekdays, especially Tuesday through Friday, with Wednesday and Thursday showing the most intense demand. Weekend usage is lower, with trips more evenly distributed throughout the day rather than concentrated in peak commuter hours.

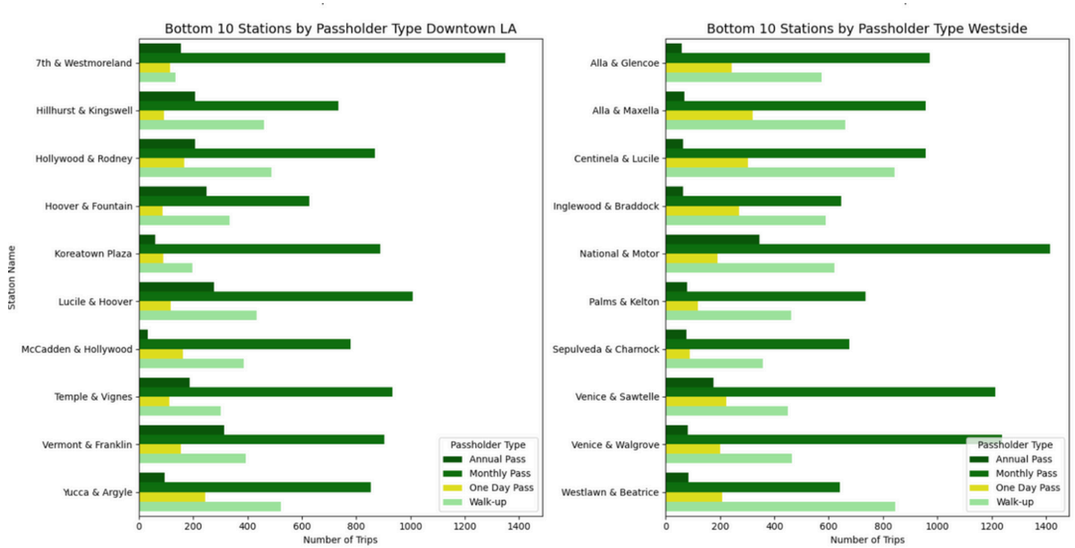


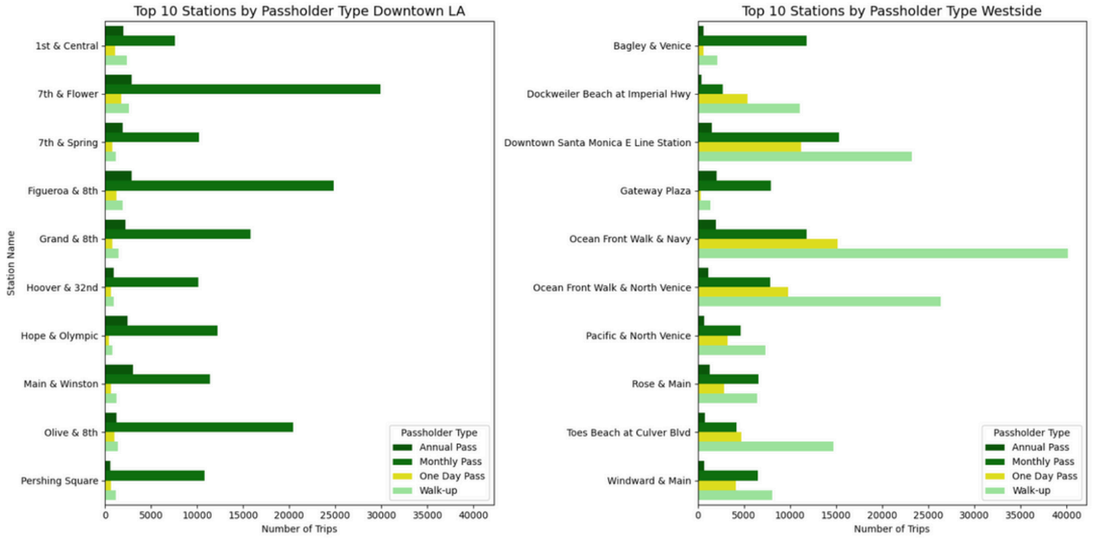
WESTSIDE:

The heatmap for the Westside indicates that bike trips peak in the afternoon, from 12:00 PM to 6:00 PM, especially on weekends, suggesting higher usage by tourists and leisure riders. Unlike Downtown, which follows a commuter pattern, the Westside sees steady activity throughout the day, with the highest demand on Saturday and Sunday, likely due to recreational trips near the beach and popular attractions.

RIDER DEMOGRAPHICS

PASS DISTRIBUTION





Low Traffic Stations

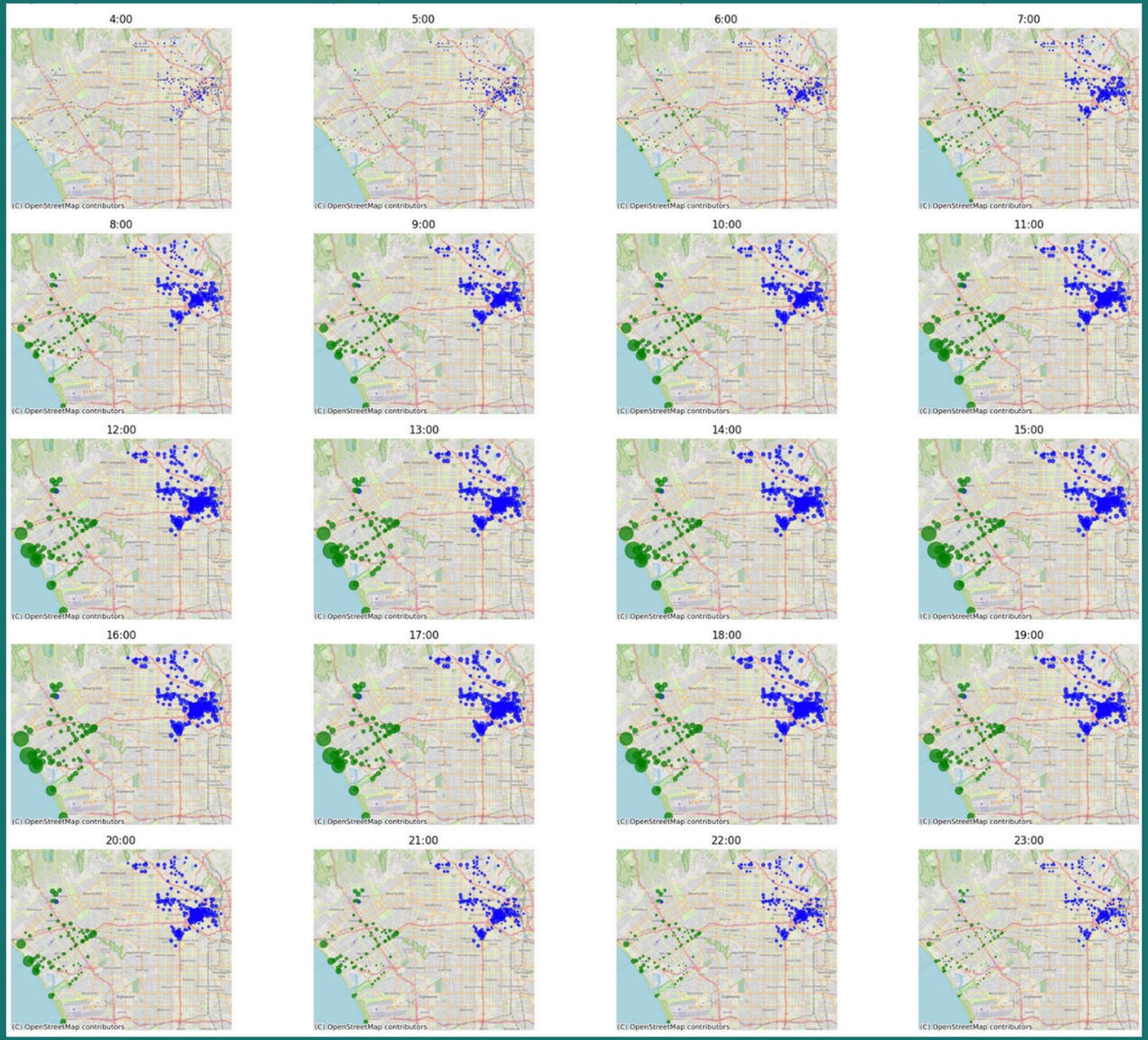
The low traffic stations pass distribution shows that it is designed for regular users. Monthly passes dominate while annual passes don’t seem to be that popular.

High Traffic Stations

The regional difference in usage patterns is very interesting. Most passholders in DTLA are monthly! While we can see a stark spike in walk up passes on the westside.

TRIP VOLUME

A MAP VISUALISATION BY TIME OF THE DAY



Peak Commute Patterns: Westside vs. Downtown

- Downtown (Blue) experiences the highest trip density during morning and evening rush hours, peaking around 8 AM and 5-7 PM.
- Westside (Green) sees more consistent, moderate trip volumes throughout the day, with noticeable spikes around midday and early evening.
- Evening hours (after 8 PM) show a decline in trips for both regions, but Downtown retains slightly more activity.

Trip Demand: Downtown Dominance vs. Westside Consistency

- Downtown (Blue): Clear demand spikes align with work commute hours, indicating a strong reliance on Metro Bike Share for work-related travel.
- Westside (Green): More evenly distributed trips throughout the day, suggesting diverse usage, including leisure and casual commuting.
- Late-Night Insights: Downtown maintains a slight edge in ridership past 10 PM, while Westside usage tapers off earlier.

Metro Bike Usage: Business vs. Leisure Trends

- Downtown (Blue): Heavy weekday traffic, particularly morning (6-9 AM) and evening (4-7 PM), driven by work-related commutes.
- Westside (Green): Higher midday and weekend usage, suggesting strong demand for leisure and recreational trips.
- Key Insight: Optimizing bike availability in Downtown during peak work hours and in the Westside during weekends could improve user satisfaction.

SPLITTING THE DATA

THE REASONING BEHIND REGIONAL DIFFERENTIATION

01 Distinct Seasonal Patterns

Downtown experiences high trip volumes on weekdays due to work commutes, while Westside trips remain steadier across the week, peaking on weekends. Treating them separately allows for better forecasting of usage trends.

02 Passholder vs. Walk-Up Behavior

Downtown stations are heavily used by monthly passholders, whereas Westside has a higher proportion of casual riders. This difference influences trip duration, station turnover, and bike availability strategies.

03 Imbalance in Station Density

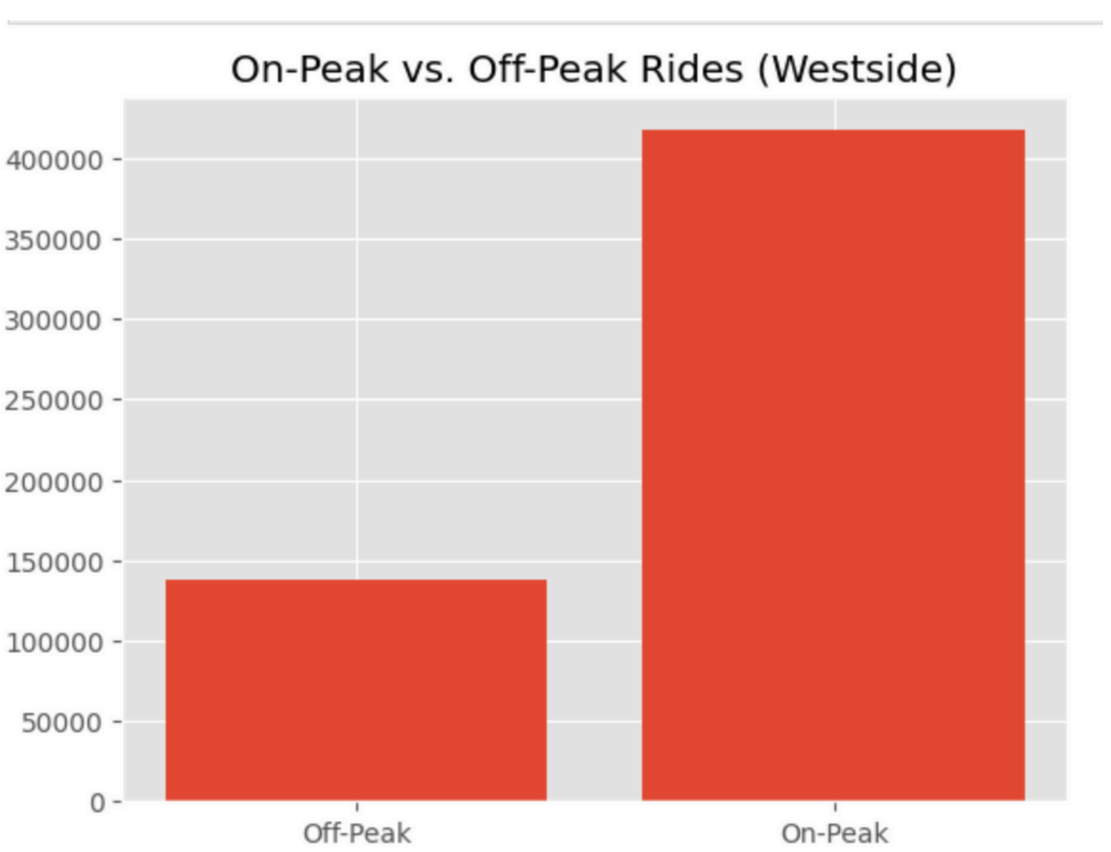
Downtown has significantly more stations than the Westside, leading to an uneven distribution of trip data. Analyzing both regions together skews insights, making it harder to identify underperforming locations accurately.

FEATURE ENGINEERING

PASSHOLDER TYPE PREDICTION

At the core of our data strategy lies a refined approach to understanding Metro Bike usage. By engineering seasonal quarters, we identified shifts in demand across winter, spring, summer, and fall, allowing for precise trend analysis.We extracted peak-hour insights, revealing Downtown's dominance during morning and evening commutes, while Westside trips peaked midday and weekends.

To predict passholder types, we incorporated trip duration, time of day, and station usage, distinguishing monthly subscribers from casual riders.



- Seasonal quarter encoding helps capture time-based demand shifts, reducing noise and improving model generalization for forecasting.
- Extracting hourly trends enhances temporal segmentation, allowing for better classification of user behavior and improving model accuracy
- Engineering trip duration, start time, and station usage improves feature separability, increasing classification accuracy

MODEL SELECTION

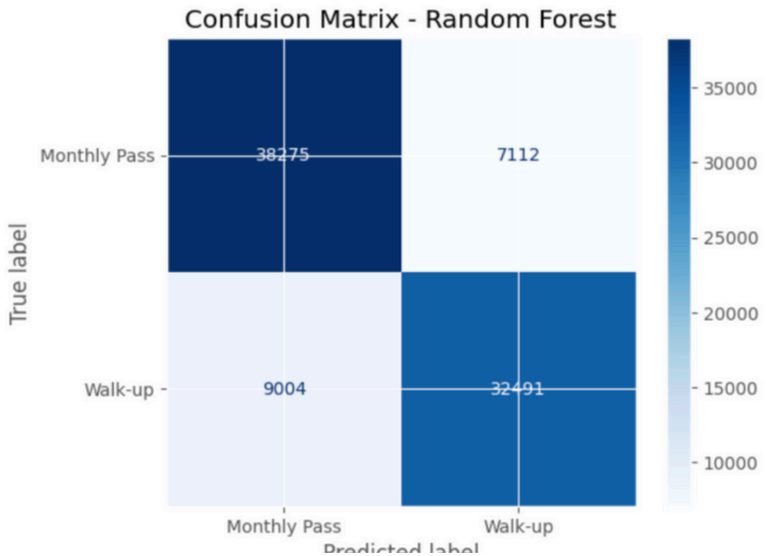
CLASS IMBALANCE

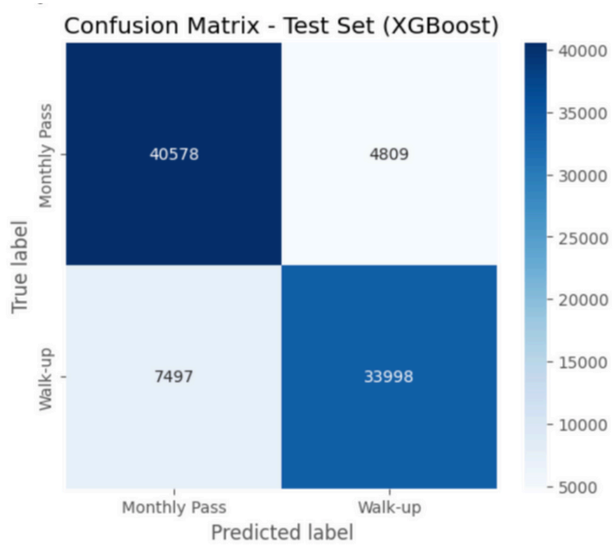
The model's accuracy may be impacted by class imbalance, as "Monthly Pass" and "Walk-up" trips significantly outnumber other categories, potentially leading to biased predictions. To improve model performance and maintain balance, we will focus on predicting "Walk-up" vs. "Monthly Pass" trips.

	passholder_type	trip_route_category	count
0	Annual Pass	One Way	31378
1	Annual Pass	Round Trip	3634
2	Monthly Pass	One Way	194810
3	Monthly Pass	Round Trip	31619
4	One Day Pass	One Way	60252
5	One Day Pass	Round Trip	25327
6	Walk-up	One Way	128880
7	Walk-up	Round Trip	79101

MODEL SELECTION

MACHINE LEARNING





Layer (type)	Output Shape	Param #
dense_8 (Dense)	(None, 128)	1,408
batch_normalization_3 (BatchNormalization)	(None, 128)	512
dropout_4 (Dropout)	(None, 128)	0
dense_9 (Dense)	(None, 64)	8,256
batch_normalization_4 (BatchNormalization)	(None, 64)	256
dropout_5 (Dropout)	(None, 64)	0
dense_10 (Dense)	(None, 32)	2,080
batch_normalization_5 (BatchNormalization)	(None, 32)	128
dense_11 (Dense)	(None, 1)	33

Random Forest Classifier

Since pass holder type prediction involves categorical classification, Random Forest is valuable due to its ability to handle non-linearity and feature interactions effectively.

XGBoost Classifier

XGBoost excels in handling imbalanced datasets and noisy data, making it ideal for predicting pass holder types. It can optimize decision boundaries, ensuring more accurate classifications between Monthly Pass and Walk-up users

FNN (Deep Learning)

FNNs can model complex, non-linear relationships between input features and passholder types. Given enough data, they can capture subtle trends that may not be evident in traditional machine learning models.

MODEL COMPARISON

PASSHOLDER PREDICTION

These results indicate that the model is effective at predicting passholder type, meaning it can be used for operational decision-making by Metro Bike Share. The ability to accurately differentiate between commuters (Monthly Pass holders) and casual riders (Walk-up users) enables targeted improvements, such as adjusting pricing strategies, modifying marketing campaigns, and optimizing bike availability based on user demand trends.

Model	Validation Accuracy	Test Accuracy	Precision (0)	Recall (0)	F1-score (0)	Precision (1)	Recall (1)	F1-score (1)
Random Forest	**81.87%**	**0.81**	0.81	0.85	0.83	0.83	0.79	0.81
Optimized RF	**81.41%**	**81.45%**	0.81	0.84	0.83	0.82	0.78	0.80
XGBoost	**79.89%**	**85.84%**	0.84	0.89	0.87	0.88	0.82	0.85
Feedforward Neural Network (FNN)	**79.86%**	**N/A**	0.78	0.83	0.80	0.80	0.74	0.77

FUTURE IMPROVEMENTS & CONCLUSION

Beyond pass holder classification, this dataset can be leveraged for multiple machine learning applications that enhance Metro Bike Share operations and urban mobility planning. Some key areas include:

1. Predicting Station Demand & Redistribution Strategies

·Goal: Forecast the demand for bikes at each station throughout the day to ensure optimal bike availability and reduce station overcrowding.

2. Optimizing Bike Dock Distribution

·Goal: Determine the optimal number of bike docks per station based on actual usage trends.

01 Feature Engineering

1.Incorporating additional behavioral data, such as time-based trends (e.g., rush hour vs. off-peak usage) and ride duration patterns, could improve model accuracy.

02 Class Imbalance Handling

The Walk-up category has a slightly lower recall, meaning some casual users are misclassified as Monthly Pass users. Techniques like cost-sensitive learning, SMOTE (Synthetic Minority Over-sampling Technique), or ensemble balancing can help address this.

03 Temporal and Sequential Modeling

1.: Since passholder behavior is likely influenced by time-based trends, using models like LSTMs (Long Short-Term Memory networks) or time-series forecasting methods could yield further insights.

THANK YOU

FOR YOUR ATTENTION

Oliver Bohler **Capstone 2 Project** **Springboard Data Science**