LA Metro Bikes
Station Optimization

How data enables new insights into trip and station behaviour

by Oliver Bohler

The Data Sources

Metro LA:	*Official website with .csv quarterly trip data*
Metro Bike API	*Live Data of stations active and docks per station*
Additional Data Sets	**Stations** *A dataset containing station coordinates and names* **Metro Train:** *Bus and Train routes and stations with coordinates*

Introduction

The Los Angeles Metro Bike Stations have experienced rapid growth, particularly with the introduction of new electronic and smart bike stations in Downtown Los Angeles and the Westside, which have seen a steady increase in demand. As part of LA's broader initiative to promote sustainable public transportation, efforts are being made to encourage residents to choose bikes over cars, with frequent promotional campaigns designed to attract both new and returning riders.

Leveraging data analytics and machine learning can enhance these initiatives by providing deeper insights into trip patterns and station usage. To maximize new sign-ups and boost monthly pass sales, this segment of the project focuses on identifying the most common passholder types at each station and uncovering patterns that can help predict pass sales at specific locations. These insights can be used to optimize promotional strategies, making campaigns more targeted and ultimately increasing revenue.

Additionally, the dataset is structured to support further machine learning applications and feature engineering opportunities, including:

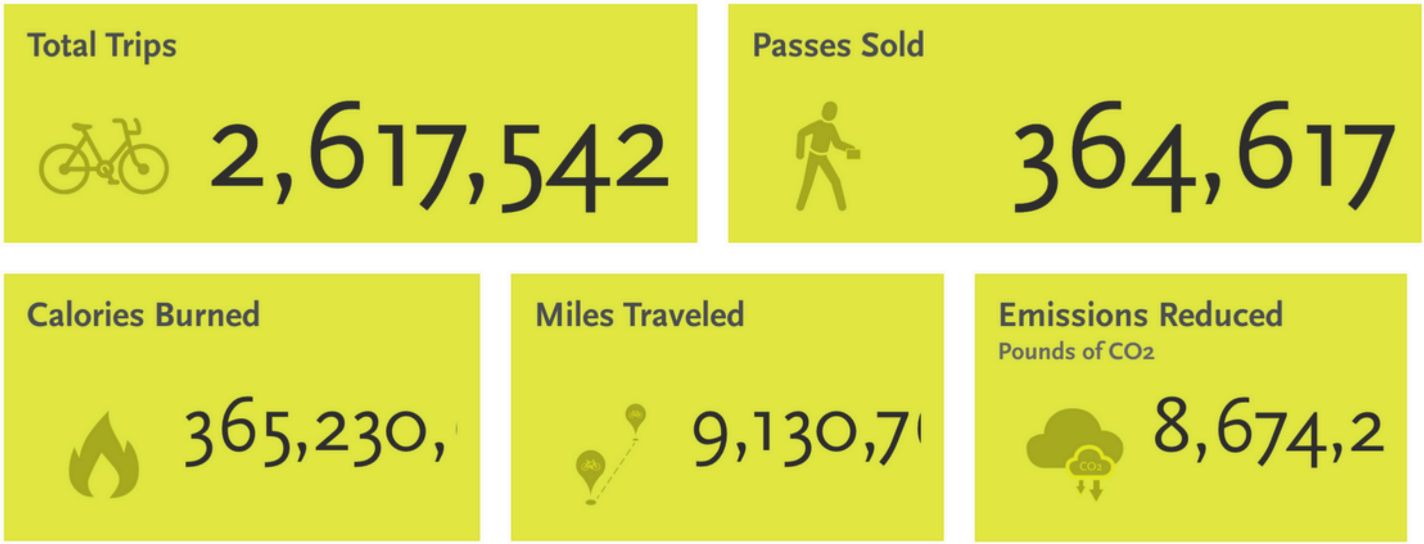
- **Station demand prediction** for efficient bike availability

- **Bike dock optimization** to reduce shortages and surpluses
- **Smart and E-bike distribution** strategies for better resource allocation
- **Region-specific modeling** to tailor services based on local commuting behaviors

By applying these advanced analytics, Metro Bike Share can continue improving its operational efficiency, user experience, and sustainability impact across Los Angeles.

Goals

Data



* Data is based on estimation using trip duration and origin/destination location.

By leveraging GeoPandas, transit station locations (train and bus), and daily usage patterns, the following objectives are established:

- **Understand trip patterns across different regions** to identify variations in rider behavior.
- **Analyze station-specific user and usage trends** to determine factors influencing demand.
- **Develop a predictive model for pass type classification** based on key insights.

A critical aspect of this analysis is identifying what makes a station busy and assessing rider habits. Are users frequent commuters (monthly/annual pass holders), or do they primarily rely on walk-up and one-day passes for convenience? Through in-depth data analysis, the goal is to uncover station-specific fluctuations, user attributes, and behavioral indicators that influence bike-sharing patterns. These insights will help optimize station planning, marketing strategies, and resource allocation, ultimately improving the efficiency and accessibility of Metro Bike Share services.

Data Cleaning

Overview of Cleaning Process

The initial objective was to merge the trip data with the station data to establish a unified dataset for analysis. To ensure accurate mapping of station names and corresponding IDs, a left join on 'Station_ID' was performed. This approach preserved all trip records while integrating relevant station information.

Following the merge, the resulting DataFrame was cleaned and analyzed through a structured data preparation process, which included:

- **Handling missing values** to maintain data integrity.
- **Standardizing station names** to ensure consistency.
- **Filtering and correcting inconsistencies** in trip records.
- **Converting date-time formats** to facilitate temporal analysis.

These preprocessing steps set the foundation for deeper insights into trip patterns, station demand, and passholder behavior, ultimately enhancing the predictive modeling and decision-making process for Metro Bike Share.

Further cleaning steps included:

Data Cleaning	Approach
Filtering Anomalous Data	Imputed missing pass holder values with most frequent category to preserve data
	Dropping the virtual station (3000) trips as they are inconclusive
Data Quality	Dropping duplicated entries and inactive stations
	Computed, analyzed missing values to ensure data quality and dropped missing values
Mode Computation	For missing pass holder types the mode was calculated and added
Date Components	Additional columns such as year, month, name of day were extracted and converted to date-time format
Filtering Trip Times	Trips with 1 min or 1440 (24h) duration were removed
API Integration	Bike dock distribution data was integrated from api

By implementing these data preparation steps, the dataset is now clean, well-structured, and optimized for further exploration and analysis. This ensures that the subsequent exploratory data analysis (EDA) will be based on accurate, high-quality data that reliably captures the true patterns and behaviors within the Metro Bike Share system. A well-prepared dataset not only enhances visualization and trend identification but also improves the effectiveness of predictive modeling, enabling deeper insights into station demand, rider preferences, and usage trends.

Exploratory Data Analysis

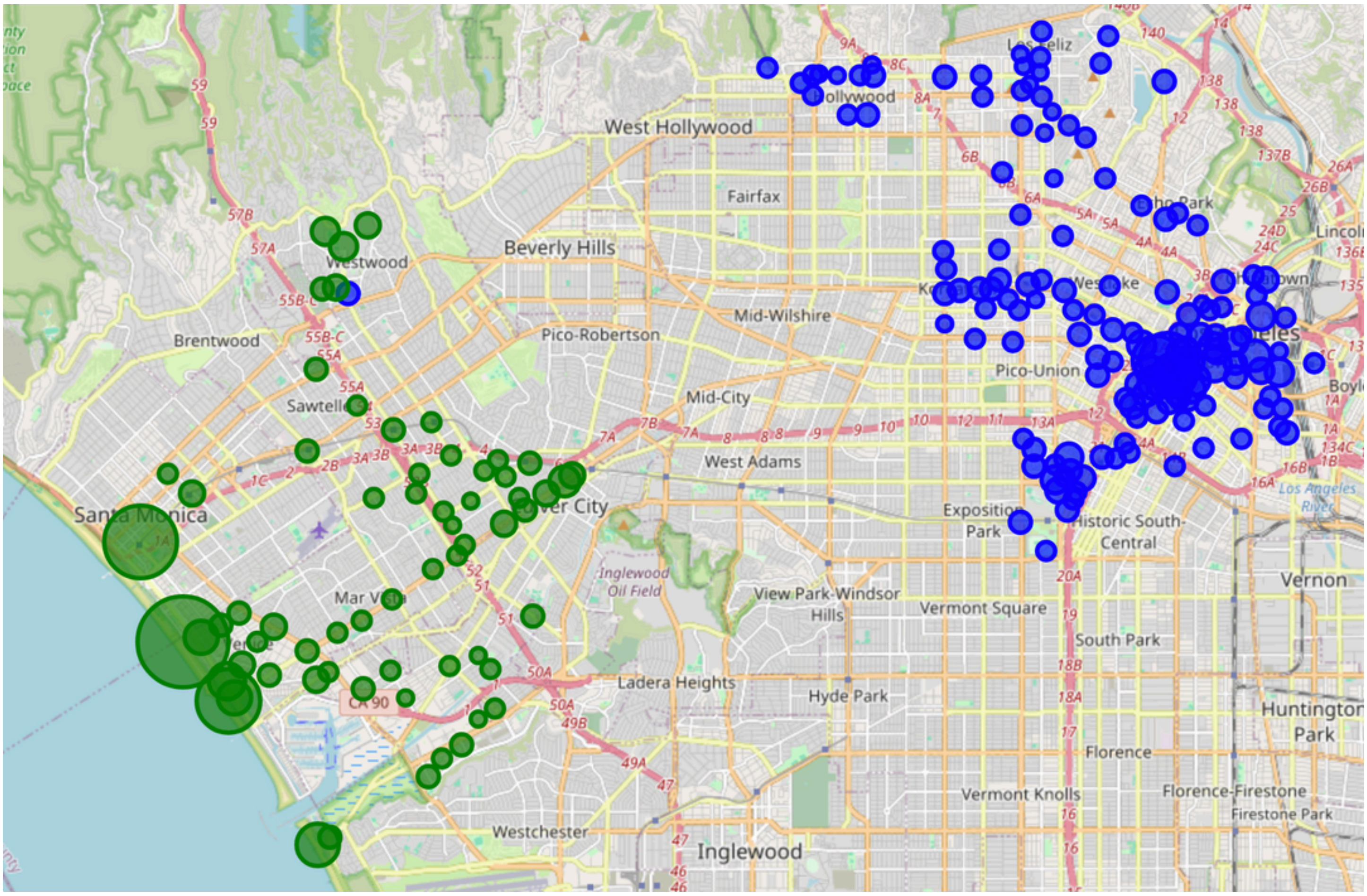
The Exploratory Data Analysis (EDA) was conducted to extract valuable insights that can be tailored to specific modeling tasks. For this phase of the project, the primary focus is on station activity and rider behavior, aiming to provide both a broad and detailed overview of the dataset while uncovering station-specific and seasonal usage patterns.

Los Angeles is an extensive city, not only in terms of population but also geographic scale. Following the data cleaning process—where inactive stations were removed—it became evident that the most effective approach would be to analyze Downtown and Westside stations separately. This decision allows for more accurate regional modeling, as each area exhibits unique demand and usage characteristics.

To visualize the spatial distribution of stations, **GeoPandas** was used to create an **interactive map** displaying key insights such as:

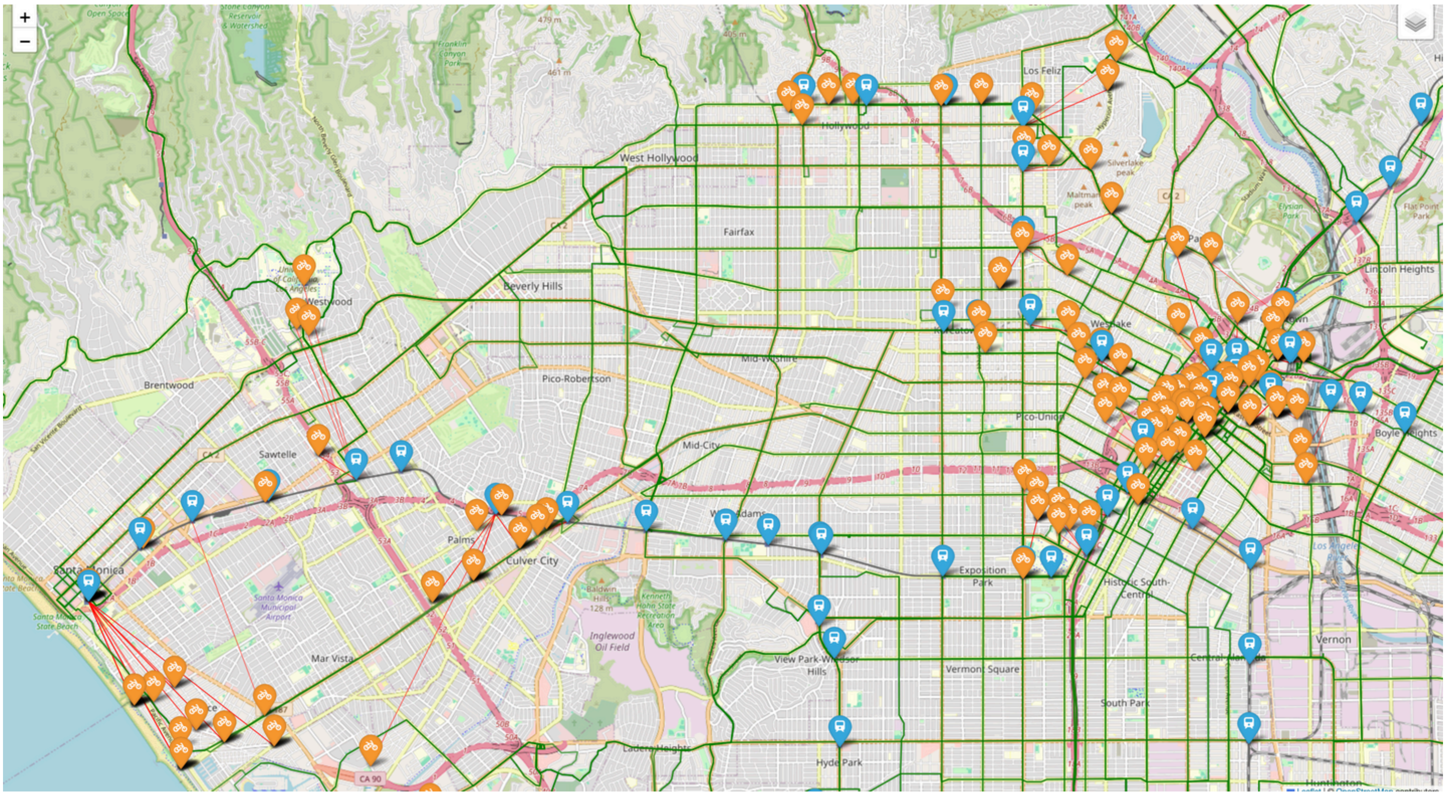
- **Total trip count per station**
- **Most traveled routes**
- **Overall demand distribution**

On the map, the size of each station's circle represents total trip volume—the larger the circle, the higher the trip count. This visualization provides an intuitive way to identify high-demand stations, frequently used routes, and potential areas for bike redistribution or station expansion.

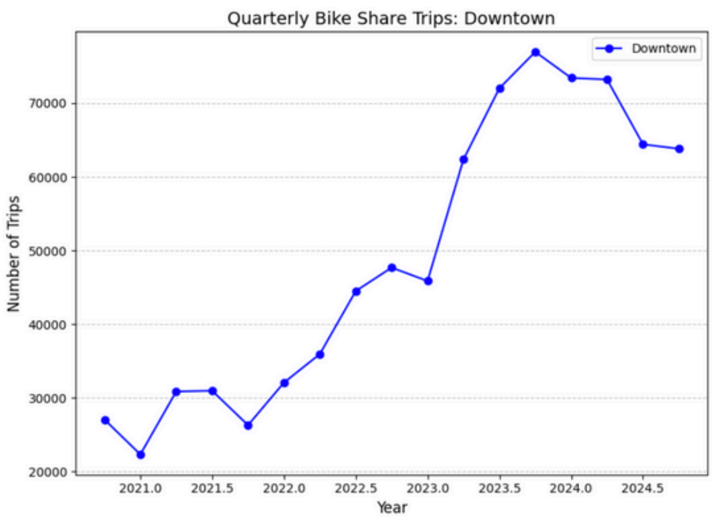


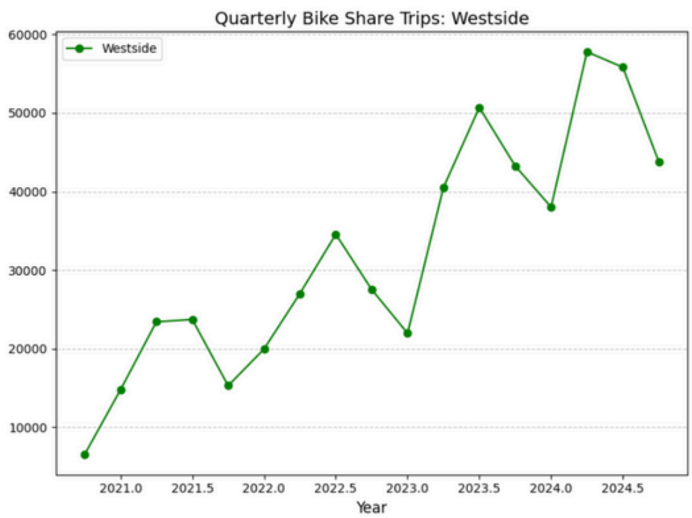
The westside is clearly dominated by the main bike stations on the ocean boardwalk. It is easy to conclude that there will be a significant seasonal influence on the westside, as this is a very attractive travel and vacation destination for people around the world.

A second map was created, incorporating Metro train stations and bus routes to better understand the distribution of bike stations. The results indicate that Metro follows a clear strategy when selecting locations for new bike stations. Additionally, the distance between each bike station and the nearest Metro train station was calculated and visualized using red lines.

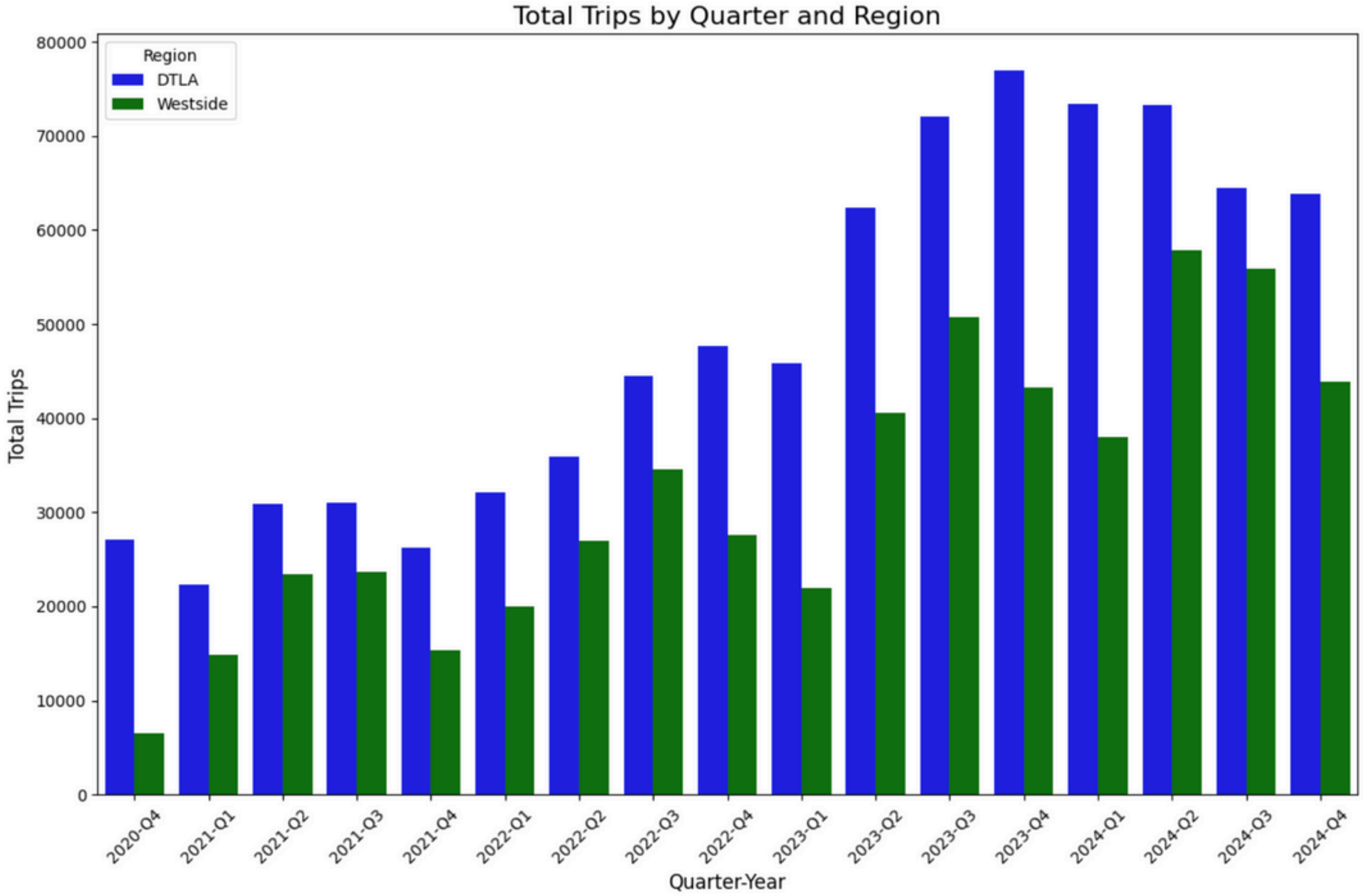


Downtown, on the other hand, exhibits a different usage pattern despite having several key stations. While the stations are clustered, they serve a higher proportion of commuters and local users compared to the Westside. Seasonality may play a less significant role in Downtown than in the Westside. The following charts provide a quarterly breakdown of trips for each region, supporting the hypothesis that seasonal trends and user demographics vary by location.

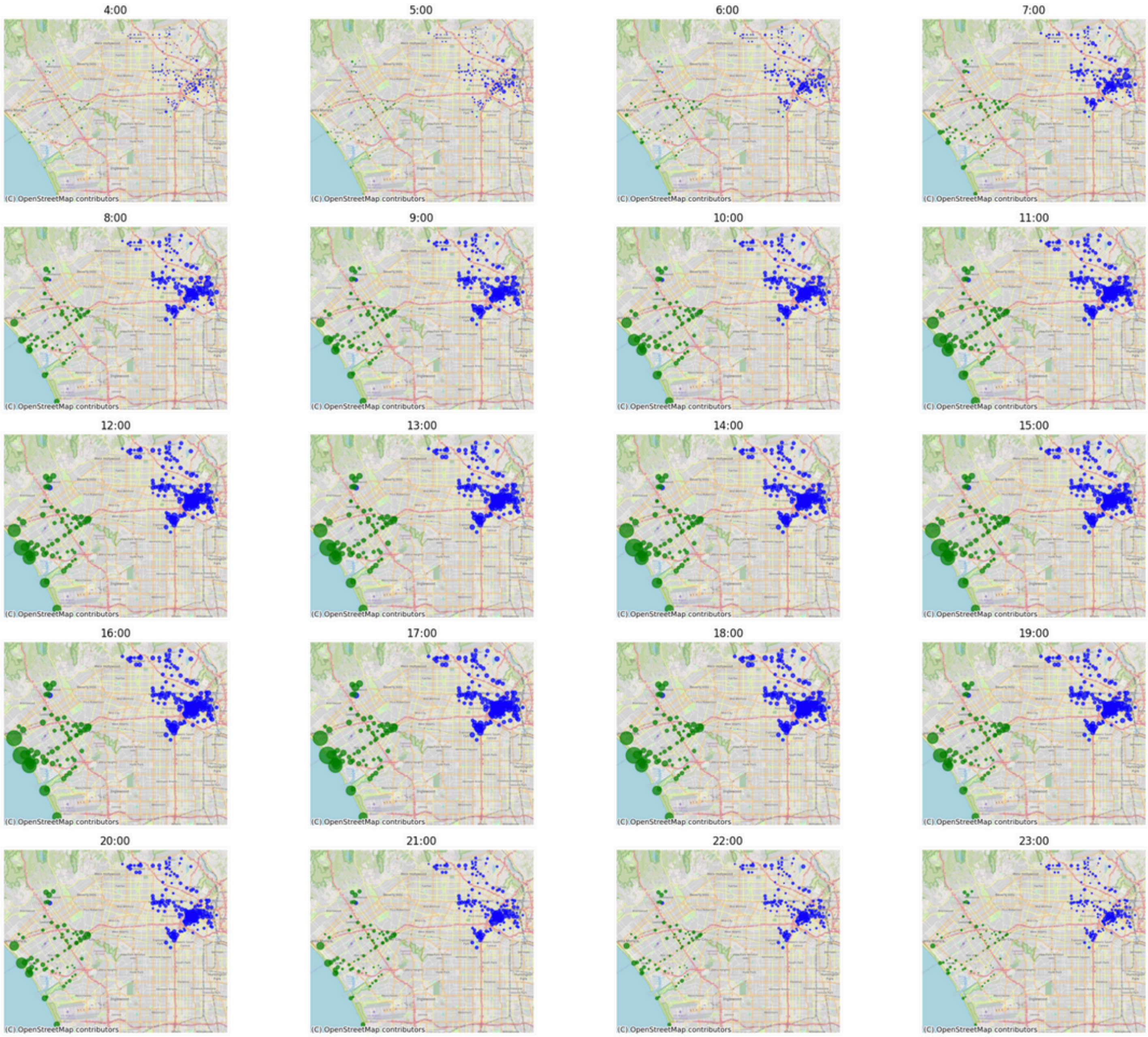




Plotting demand per region on a bar chart reveals a steady growth trend since 2021. While Downtown leads in total trip volume, it's important to consider that the number of stations is significantly higher in this region—146 stations in Downtown compared to just 62 in the Westside. This disparity in station availability plays a key role in the difference in trip counts between the two areas.

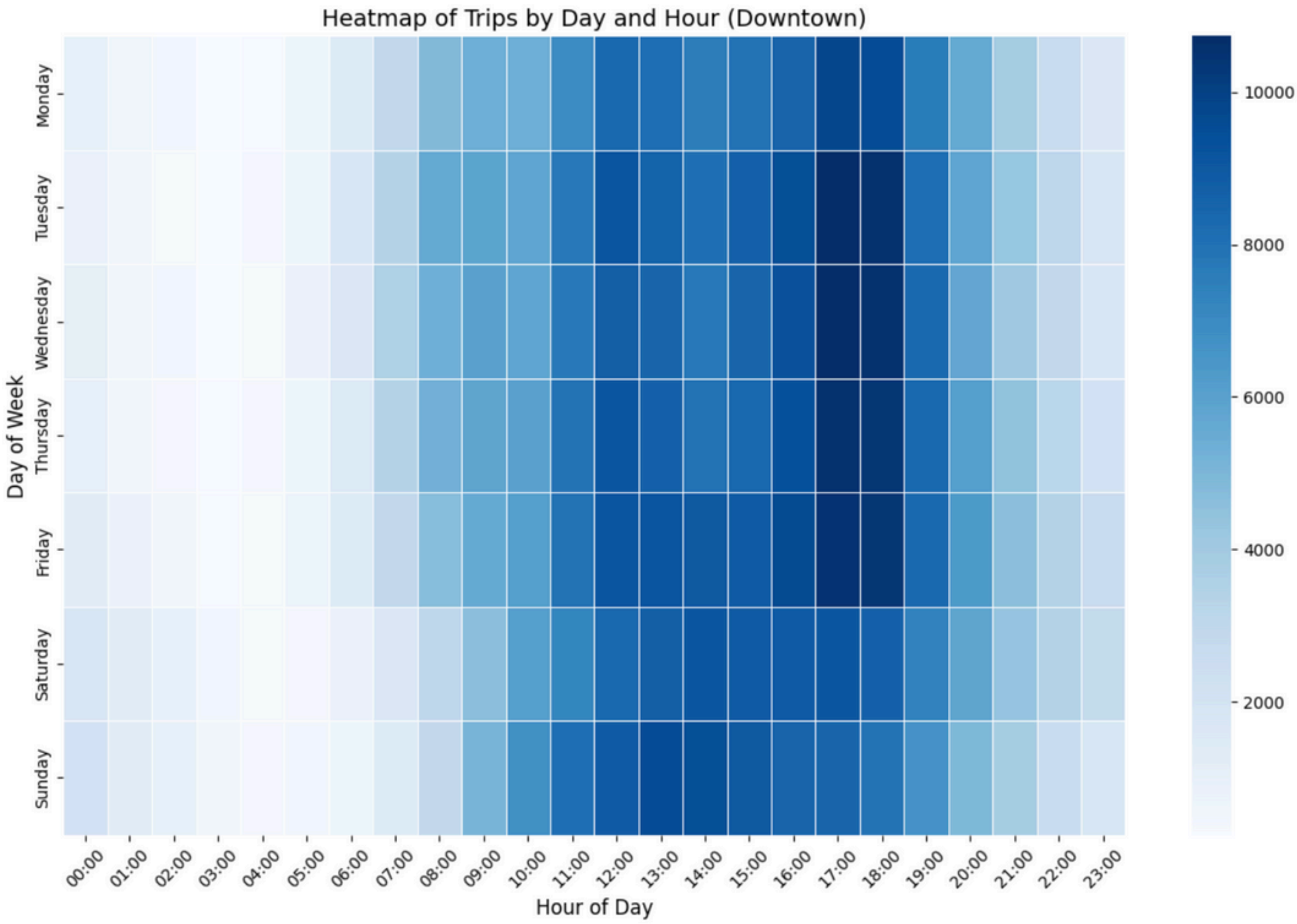


Visualizing trip volumes by the hour provided deeper insights into how each region operates over a 24-hour period. The data clearly shows that most demand occurs between 7 AM and 7 PM, with Downtown experiencing higher activity in the morning, likely due to commuter traffic. The next step was to analyze whether usage patterns differ between weekdays and weekends.

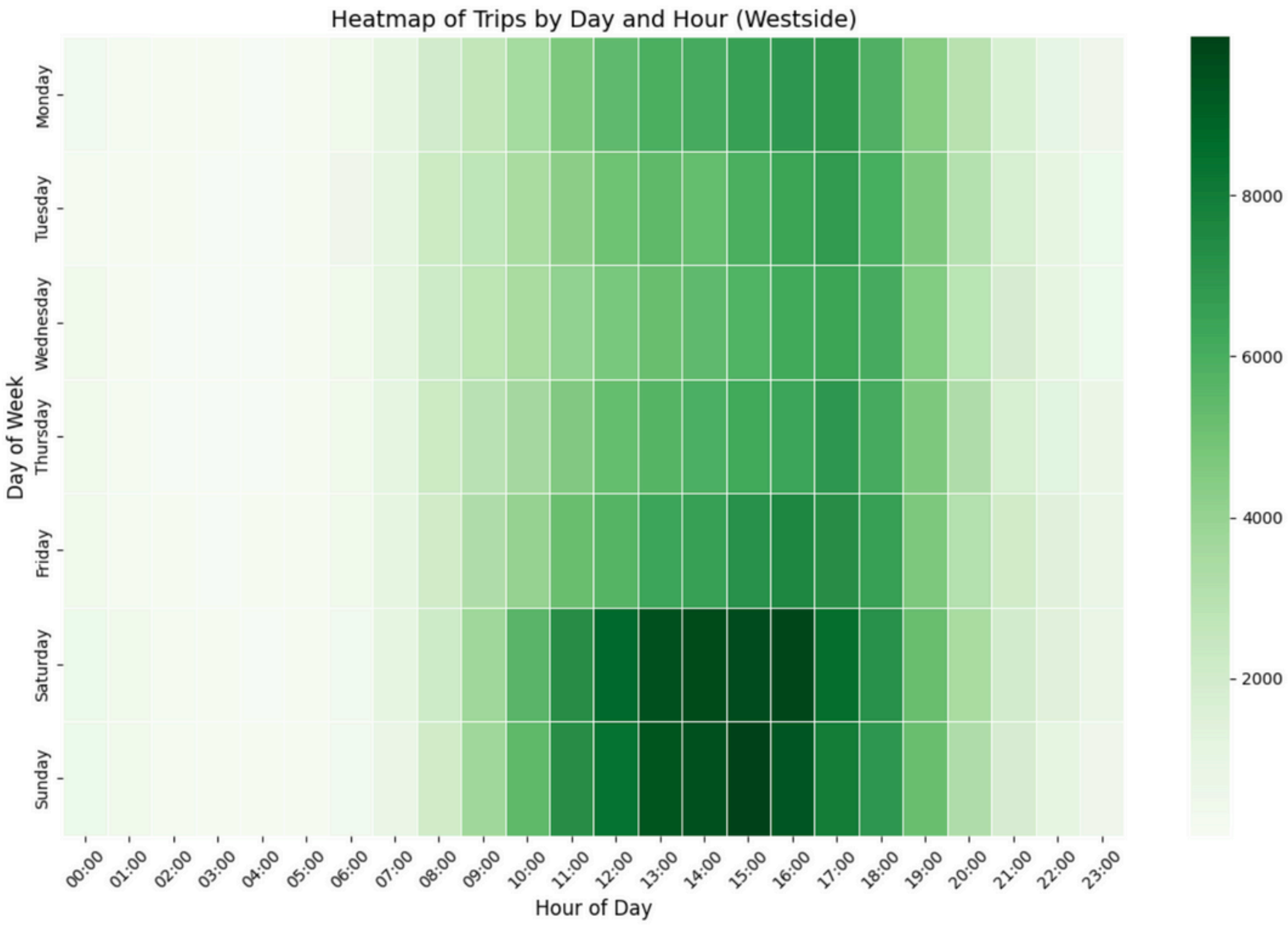


A heat-map is an effective way to visualize daily variations in trip volumes, with darker squares indicating higher trip activity.

For Downtown, the commuting patterns are clearly visible. The most active time slot is between 12 PM and 7 PM, with strong indications that bikes are also used for morning commutes. Weekdays dominate overall usage, while weekend activity is comparatively lower.



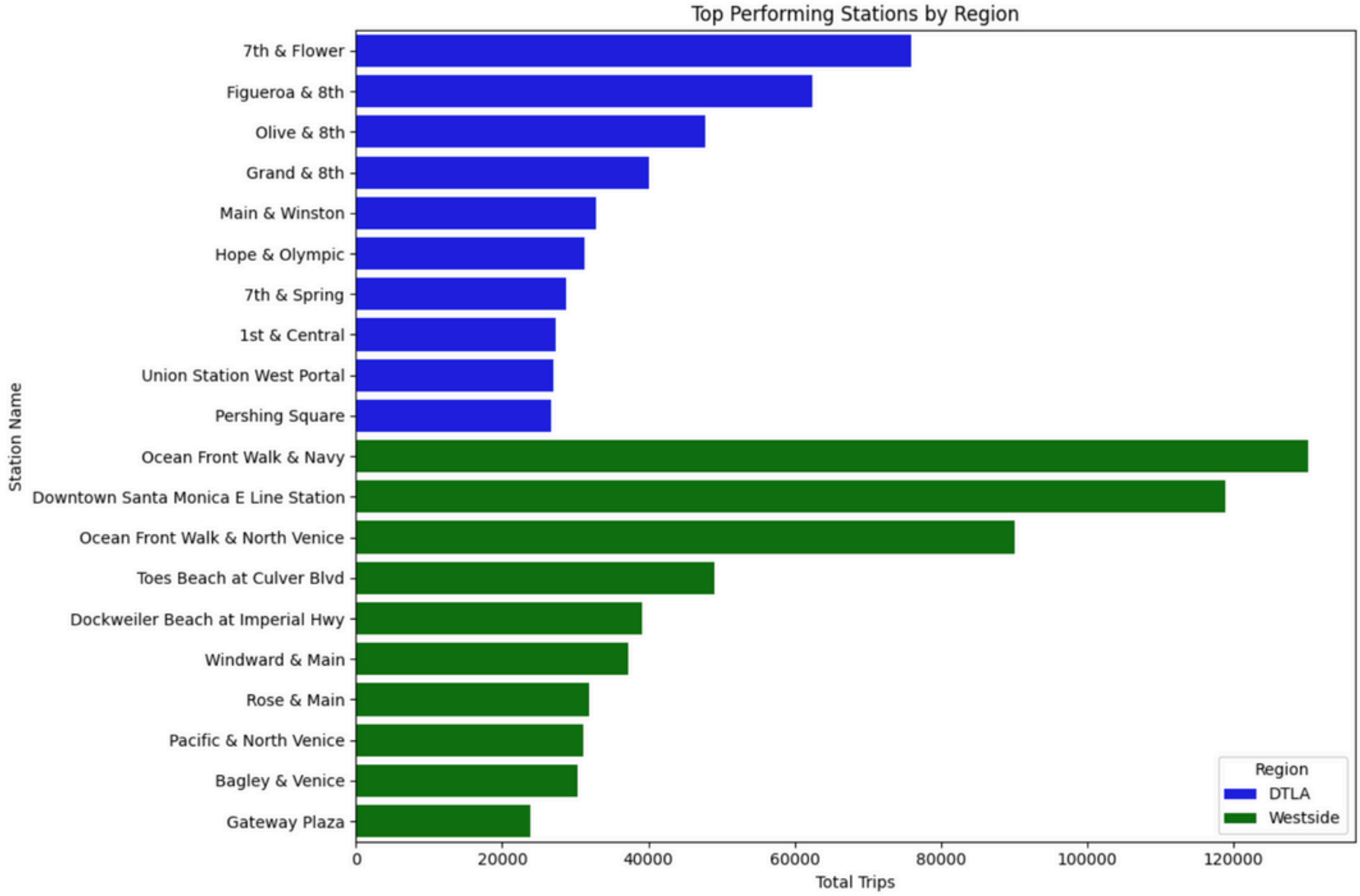
For the Westside: Here we can see a clear trend of recreational and leisure use in this region. While we have some higher demand during the week it is imminent that the main usage here is focused on the weekend.



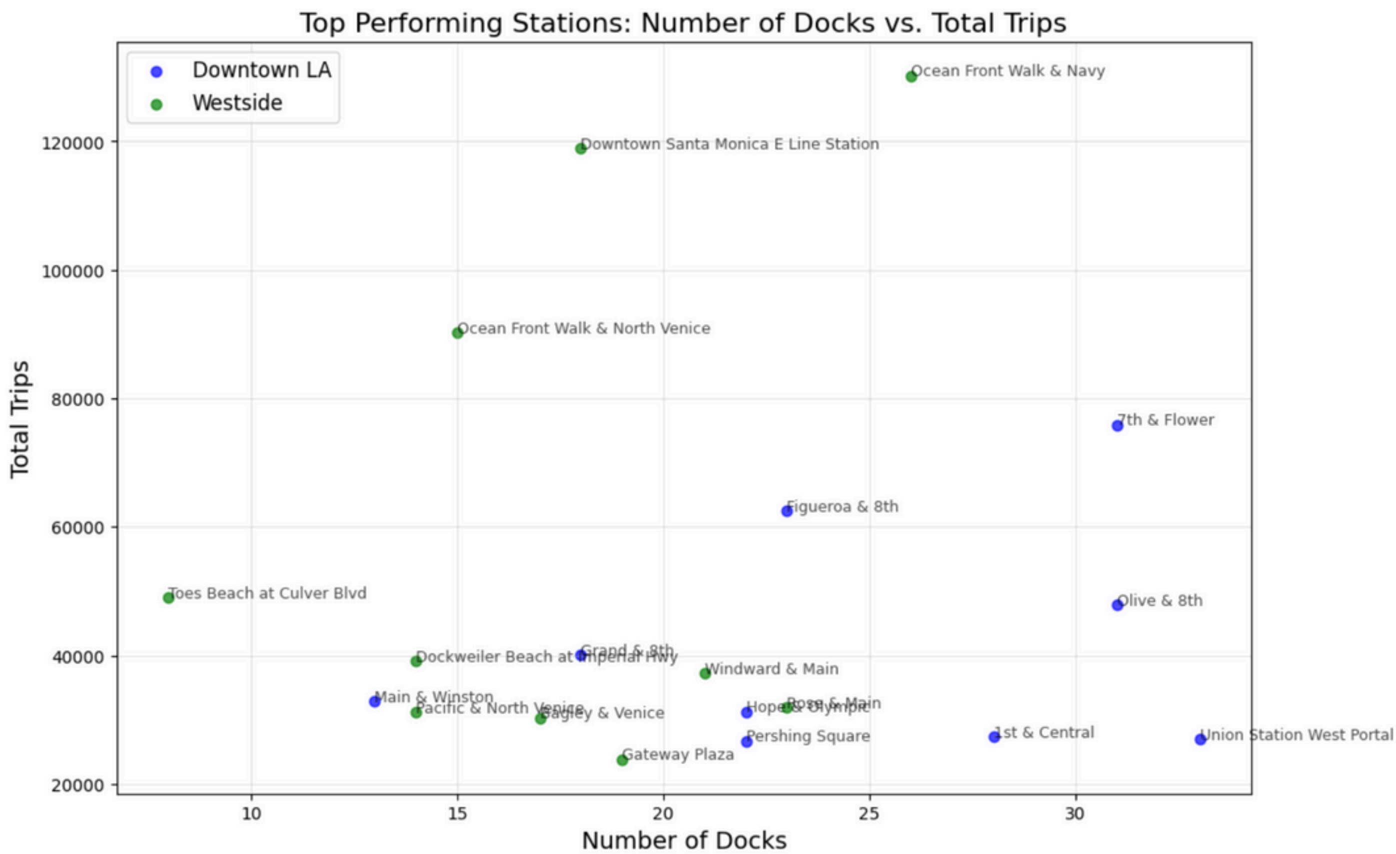
As previously mentioned, station distribution varies significantly across regions, with some areas having a higher concentration of stations than others. To gain an initial understanding of regional station behavior, stations were categorized into top and underperformers, specifically identifying the 10 highest and 10 lowest-performing stations based on total trips.

The chart below highlights the top-performing stations, revealing key insights into demand concentration and regional dependencies. Notably, the Westside relies heavily on its primary boardwalk stations, where trip volumes are significantly higher compared to surrounding stations. In contrast, Downtown stations exhibit a more balanced distribution of trip volume, with less variation between high and low-performing locations.

These findings suggest that station placement, local infrastructure, and accessibility play a crucial role in shaping bike-sharing demand patterns, particularly in areas with concentrated tourism and recreational usage, such as the Westside.

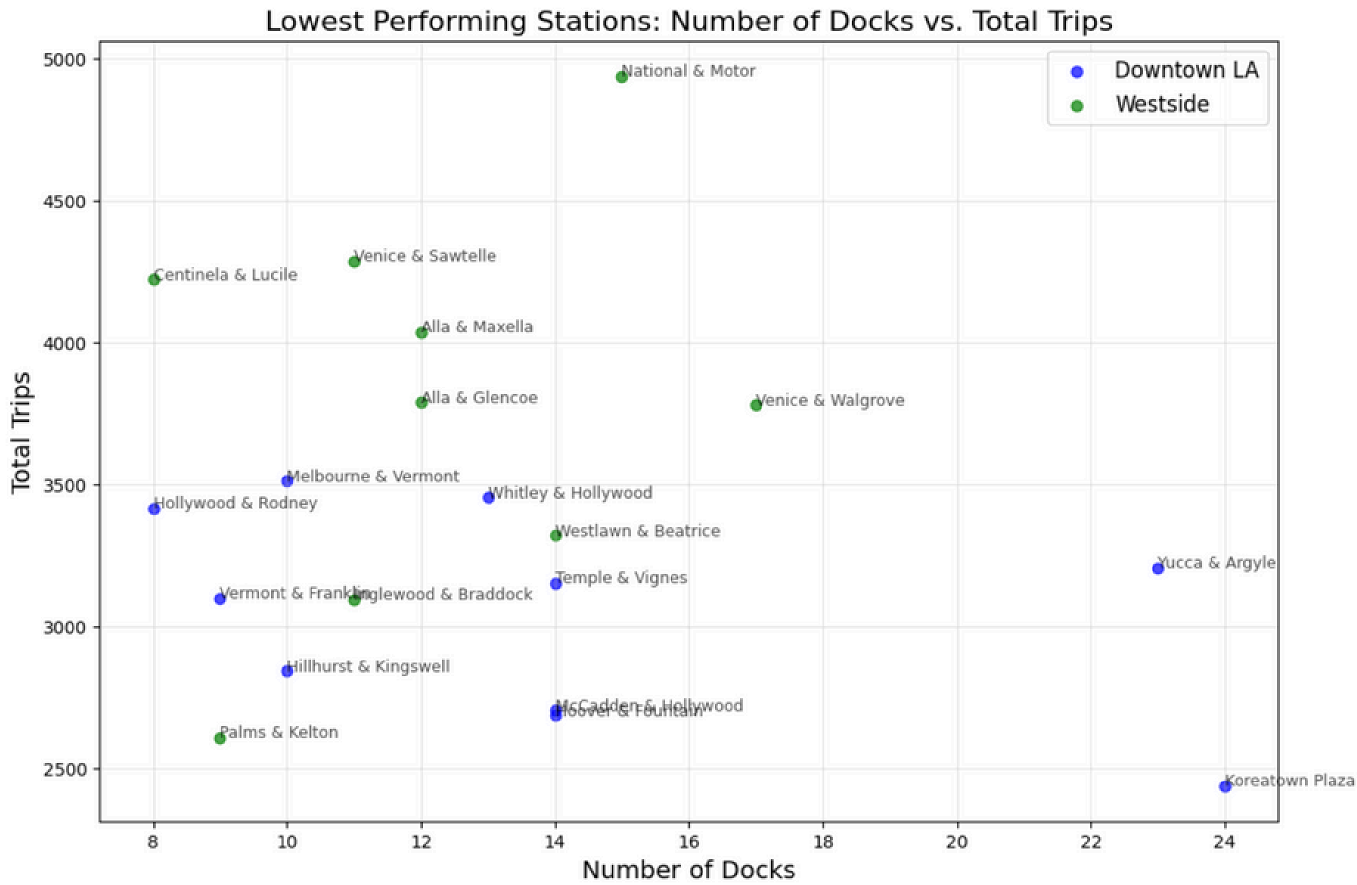


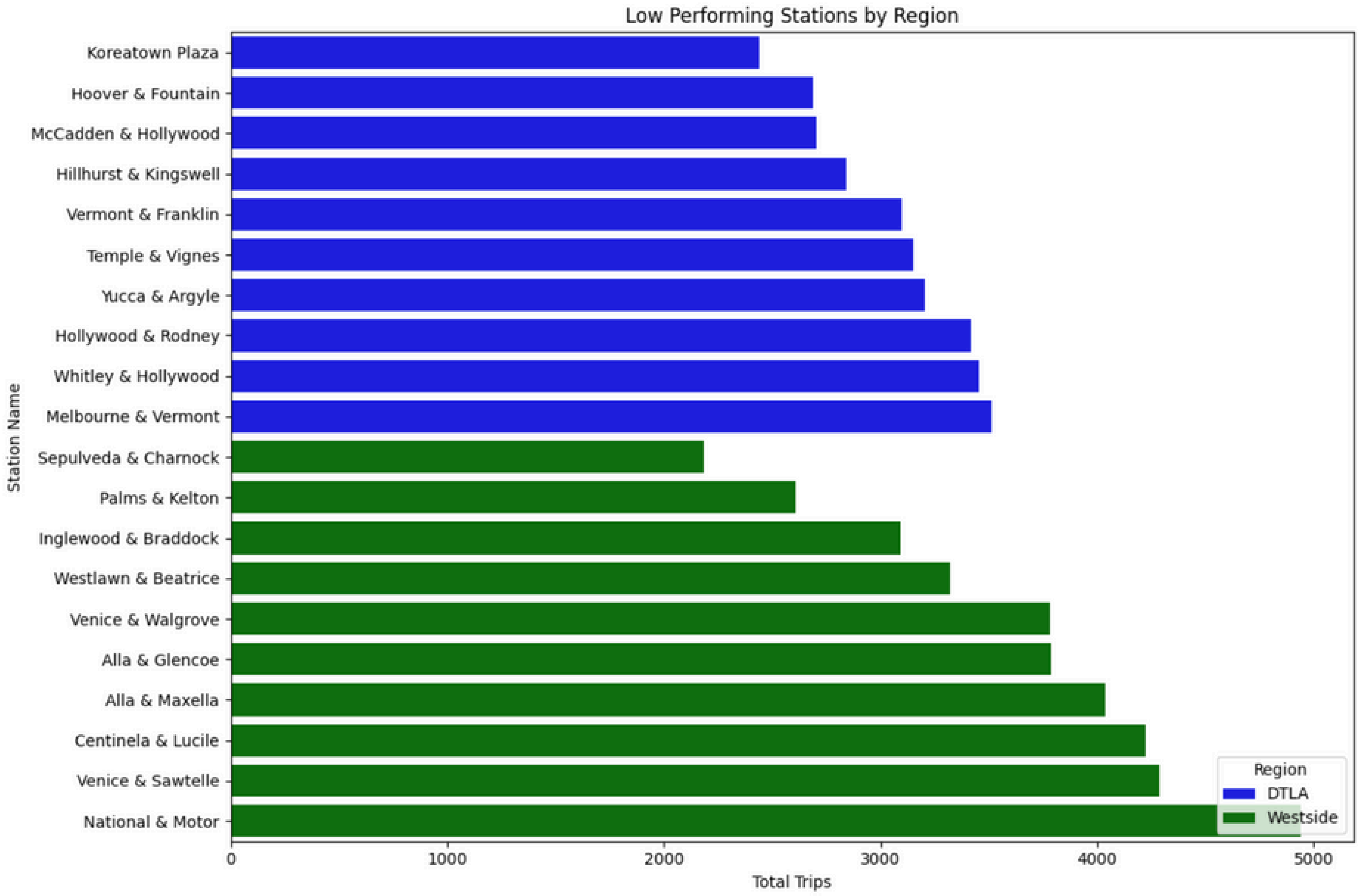
The distribution of bike docks can help to see how well stations and dock stations match the demand. It seems like the average number of docks is between 15-20 docks per station.



Interestingly, the placement of low-performing stations appears to have been strategically planned. The data suggests that these stations were expected to have lower traffic, as they are equipped with fewer bikes, typically ranging from 8 to 14. This indicates a deliberate effort to allocate resources efficiently based on anticipated demand.

The bar chart below further supports this observation by showing that most underperforming stations exhibit consistently low trip volumes, with little variation among them. This pattern suggests that while these stations may not see high daily usage, they still serve a purpose—likely providing last-mile connectivity, filling gaps in coverage, or supporting localized demand in residential or less tourist-heavy areas.





While working on this project, I had the opportunity to speak with an urban planner for the City of Los Angeles. When I asked why stations with low trip volumes weren't simply closed down, his response provided valuable insight into the broader purpose of public transportation. He explained,

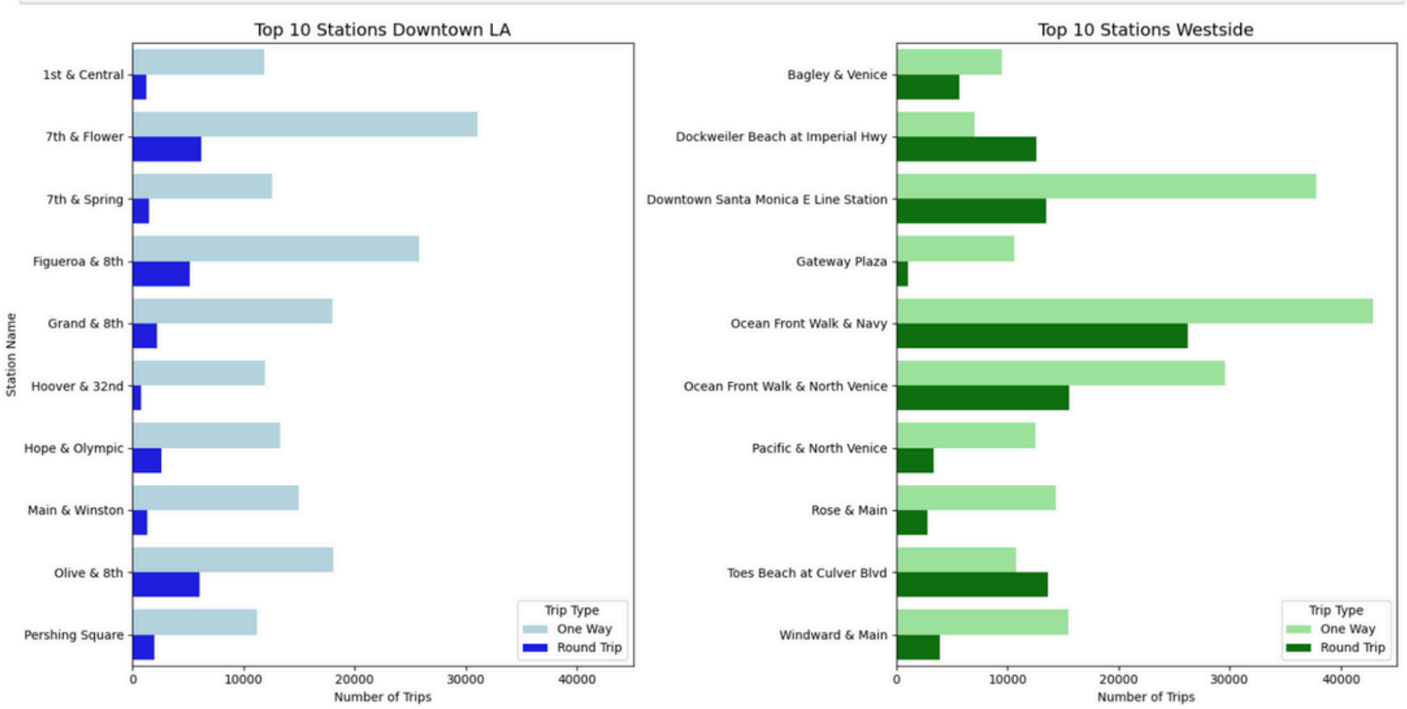
"Public transportation is not a business designed for profit. It exists—especially bike-sharing programs—to encourage citizens to adopt alternative transportation methods and, in some cases, to provide individuals with their only means of mobility."

In essence, bike stations serve a purpose beyond ridership metrics. They contribute to reducing CO_2 emissions, promoting sustainable urban mobility, and increasing accessibility for those who may not have other transportation options. While some stations may have lower demand, their presence helps bridge gaps in transit coverage and aligns with the city's commitment to environmental sustainability and equitable transportation access.

To gain deeper insights into station usage, trip patterns were analyzed based on One-Way vs. Round-Trip rides across different regions. This analysis provided further evidence that demand patterns vary significantly between Downtown and the Westside.

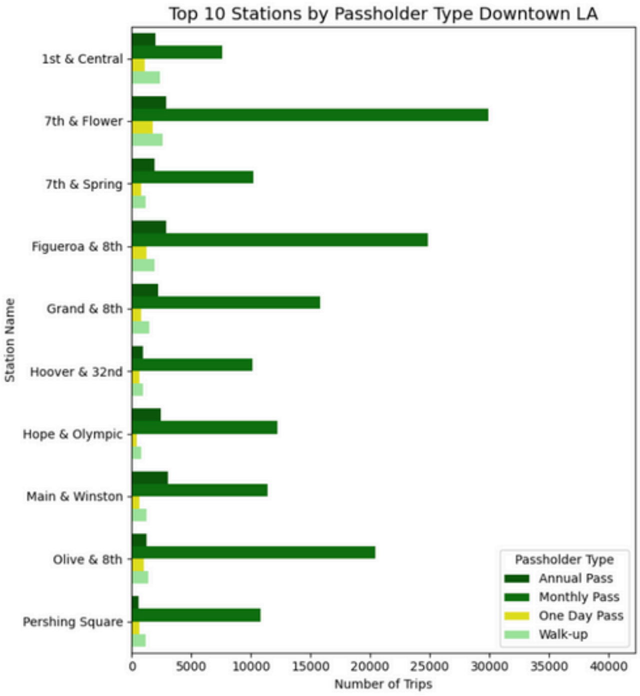
The findings confirm that Downtown serves primarily as a "Commuter Hub," with the majority of trips being One-Way, suggesting riders use the bikes for first-mile and last-mile connectivity to workplaces, transit stations, or other destinations. In contrast, the Westside exhibits a much higher proportion of Round-Trips, reinforcing the idea that bike usage in this region is more recreational—likely driven by tourism, leisure rides along the boardwalk, and casual users.

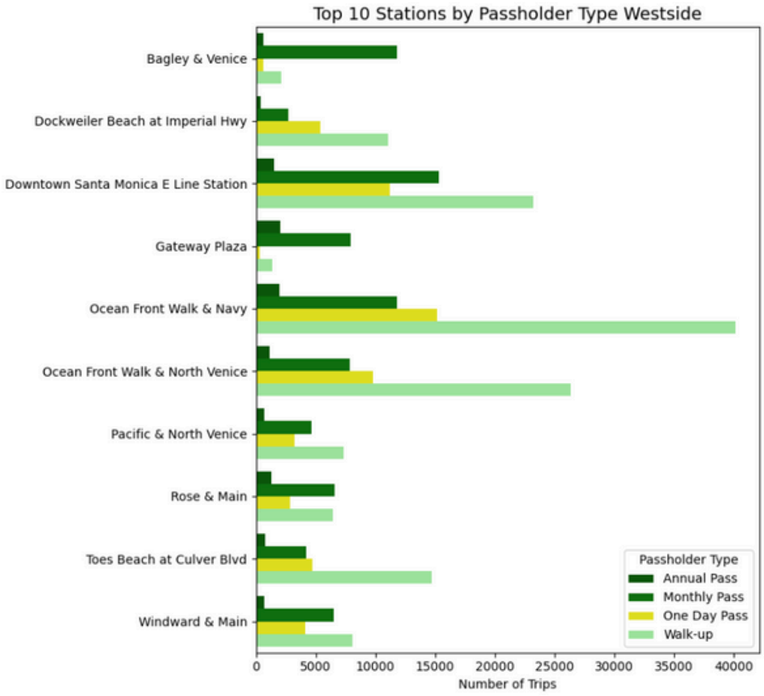
Interestingly, low-performing stations follow a similar trend but on a smaller scale. Even among stations with lower trip volumes, Downtown locations still favor One-Way trips, while Westside locations show a stronger Round-Trip presence.



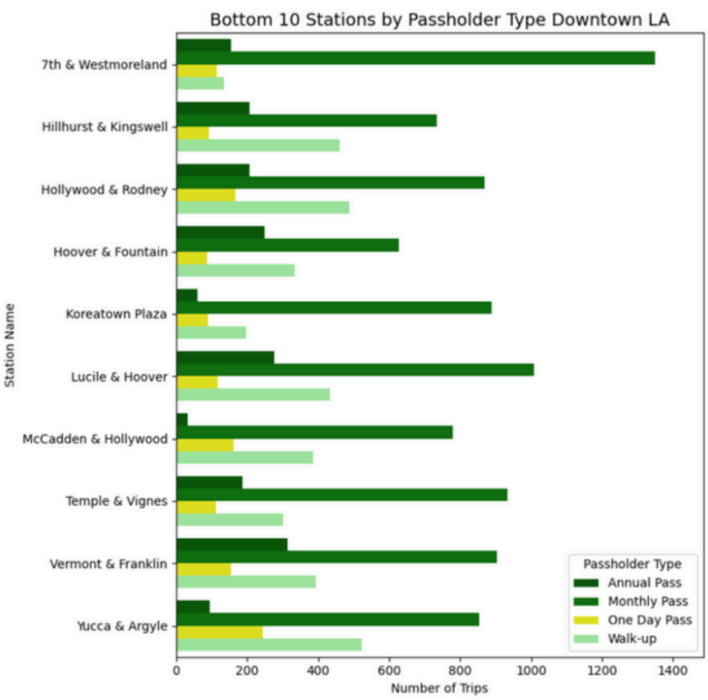
To gain deeper insights into user behavior, an analysis of pass types was conducted to understand the purchasing patterns of riders. The findings reveal a significant number of Walk-Up passes, which can likely be attributed to tourists visiting the area for a short period, often just for a day or two. This aligns with the high recreational usage observed on the Westside, where casual riders are more common, particularly near boardwalk and beachfront stations.

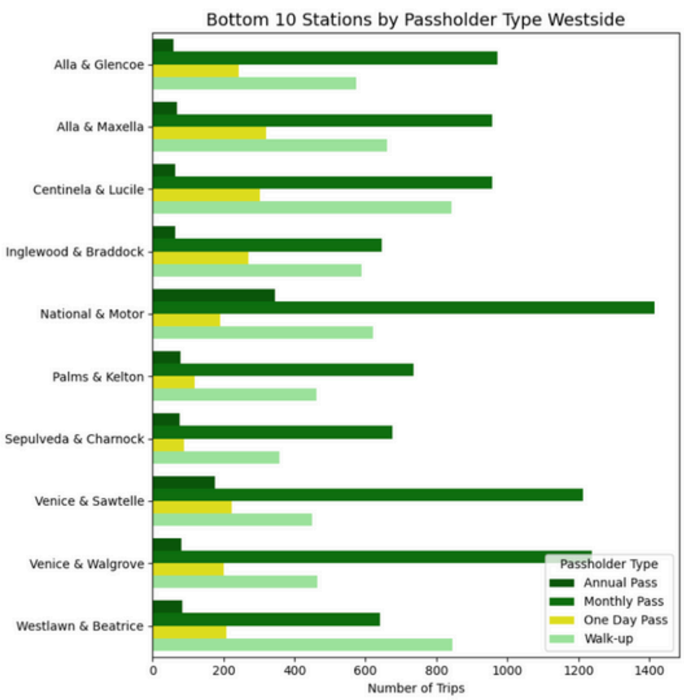
In contrast, Downtown's top stations exhibit a much higher demand for Monthly Pass users, reinforcing the hypothesis that Downtown serves as a commuter hub. This distinction between tourist-driven Walk-Up usage in the Westside and commuter-driven Monthly Pass usage in Downtown is visually confirmed through the data.





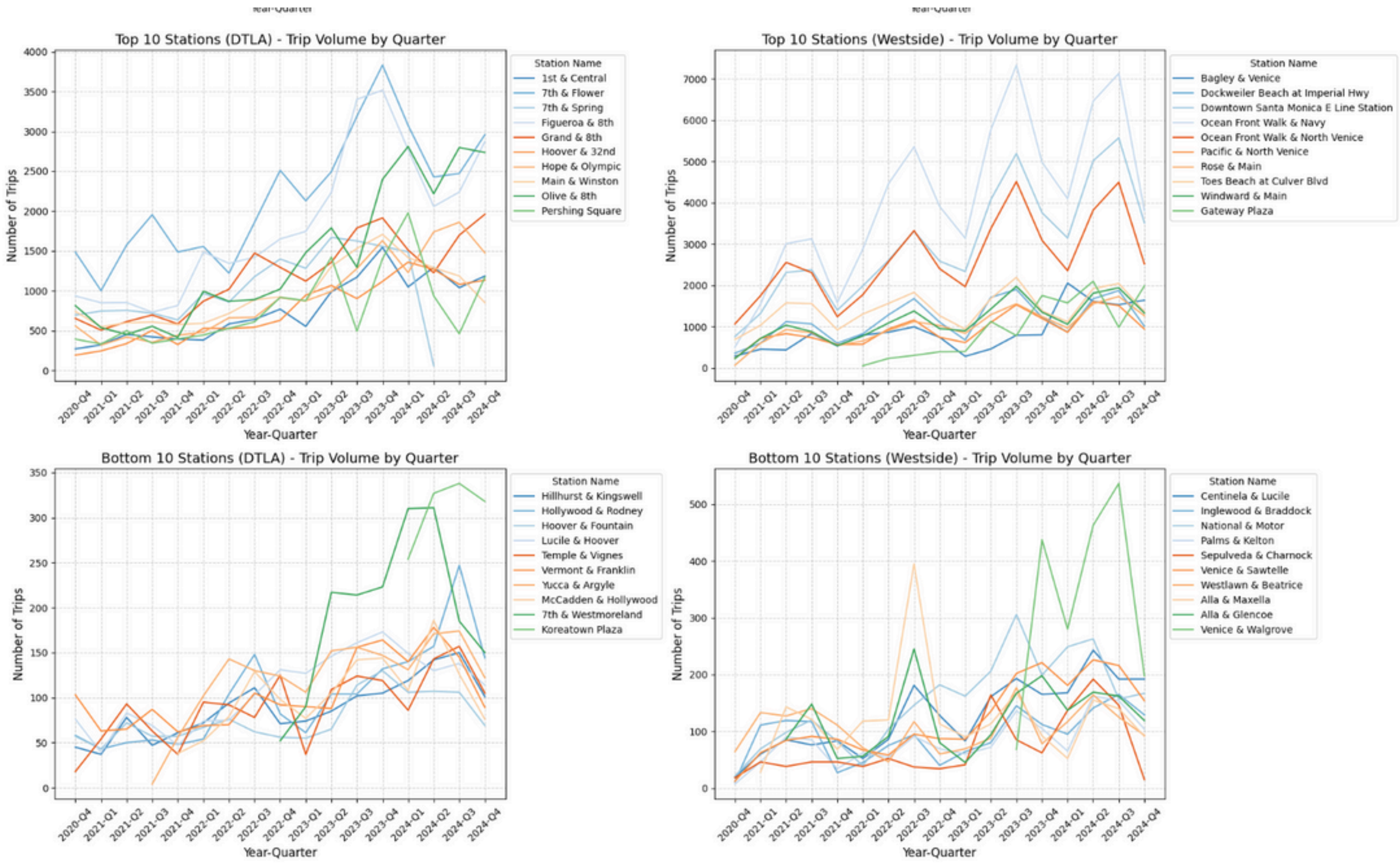
One of the more unexpected insights from the analysis is the distribution of passholder types at low-performing stations. Despite their lower trip volumes, these stations are primarily used by Monthly Pass holders in both regions, suggesting that they still serve a critical function for regular riders. This finding reinforces the idea that even low-demand stations contribute to the broader transportation network, providing access for commuters or residents who rely on bike-sharing as part of their routine travel.





However, while Monthly Pass usage dominates, Annual Pass adoption remains notably low across these stations. This suggests that long-term commitment to the bike-share program may not be as appealing to users—potentially due to pricing structures, limited perceived benefits over monthly plans, or the availability of alternative transportation options.

Finally, the seasonality of both top-performing and underperforming stations was analyzed by visualizing total trip counts across operational quarters. This approach highlights fluctuations in demand over time, providing insights into how ridership patterns shift seasonally.



For top-performing stations, we observe clear seasonal peaks, likely influenced by weather conditions, tourism trends, and commuter behavior. The Westside stations, in particular, show higher trip volumes during warmer months, reinforcing their recreational appeal. In contrast, Downtown stations maintain steadier demand year-round, aligning with their commuter-driven usage.

For underperforming stations, the seasonal trends are less pronounced, but there are still noticeable shifts in demand. These stations, while lower in total trip volume, may still experience modest increases during peak commuting or tourism seasons.

Conclusion:

The analysis clearly demonstrates that Downtown and the Westside exhibit distinct usage patterns, reinforcing the need to treat them as separate datasets for feature engineering and model selection. Downtown functions primarily as a commuter hub, where ridership is driven by monthly pass holders and weekday demand, while the Westside sees higher recreational usage, particularly from walk-up users and seasonal tourists. Given these differences, separating the data ensures that models can better capture region-specific trends and improve predictive accuracy.

For the purpose of this project, the focus will be on developing a feature engineering and model selection strategy aimed at understanding the key factors that drive monthly and annual pass sales. By identifying station-specific attributes, user behavior patterns, and seasonal influences, the goal is to optimize marketing strategies and operational decisions to increase long-term rider retention and improve Metro Bike Share's overall efficiency.

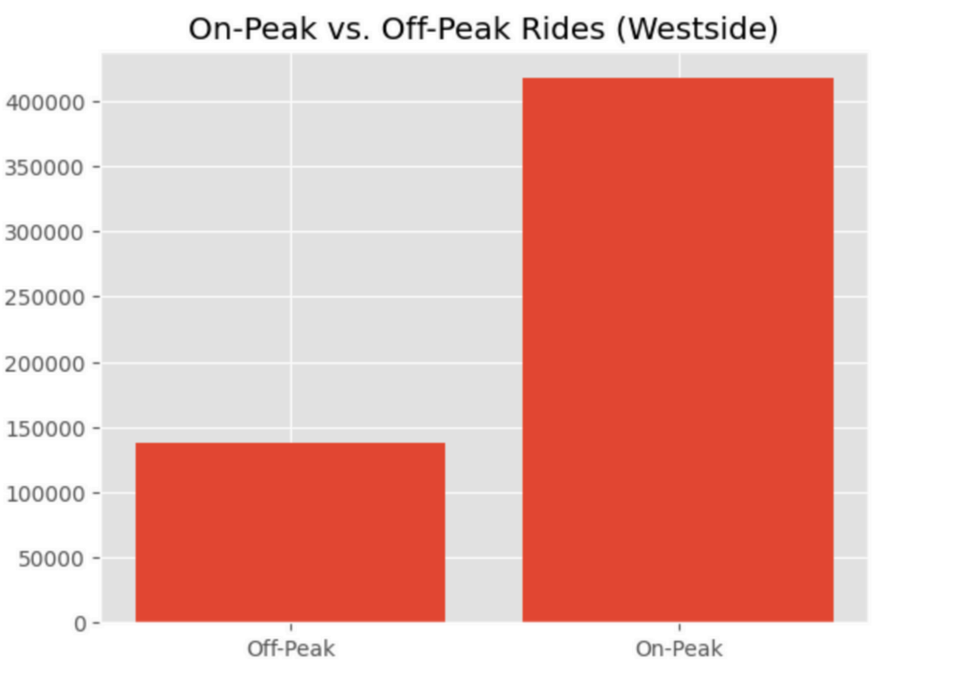
Feature Engineering

Goal

The objective is to predict which pass type a rider will purchase at any given station by engineering key features that capture trip behavior, station demand, and seasonal influences. Based on insights from the heatmap analysis, specific variables were created to simplify how trip timing affects model performance:

- **Peak Hours**: A binary column indicating whether a trip occurred during **high-demand or low-demand times**, derived from the heatmap findings.
- **Quarters**: A categorical variable representing the **season in which the trip occurred** (Winter, Spring, Summer, Fall), helping the model capture **seasonal variations in ridership**.
- **One-Hot Encoding**: All relevant **categorical variables** (such as station location, pass type, and trip category) were converted into a machine-readable format to ensure the model can effectively distinguish between different user behaviors.

These engineered features will enhance the predictive power of the model by incorporating temporal trends and behavioral patterns, allowing for a more precise understanding of pass holder preferences and targeted recommendations for Metro Bike Share operations.



Design

After analyzing the passholder types and their associated trip distributions, a clear pattern emerged: Annual and One-Day passes represent a significantly smaller portion of total trips, making them minority classes in the dataset. Including these underrepresented pass types in the model would likely lead to class imbalance issues, potentially skewing predictions and reducing overall model performance.

To ensure a more balanced and reliable classification, the decision was made to focus on the two dominant pass types:

- **Monthly Pass** – Representing **commuter and regular riders**, typically using the service for daily or frequent travel.
- **Walk-Up** – Representing **casual or tourist riders**, who generally purchase passes on an as-needed basis.

By filtering out Annual and One-Day passes, the model can focus on distinguishing between high-frequency and casual riders, leading to more meaningful predictions and ensuring that the insights gained remain actionable for Metro Bike Share operations.

	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

To enhance model performance and ensure **efficient feature selection**, the following steps were implemented:

Scaling & Normalization:

- Features were standardized using StandardScaler to ensure that all numerical variables have a mean of 0 and a standard deviation of 1.
- This step prevents features with larger magnitudes from dominating the model, ensuring a more balanced contribution of all input variables.

Principal Component Analysis (PCA) for Feature Selection & Optimization:

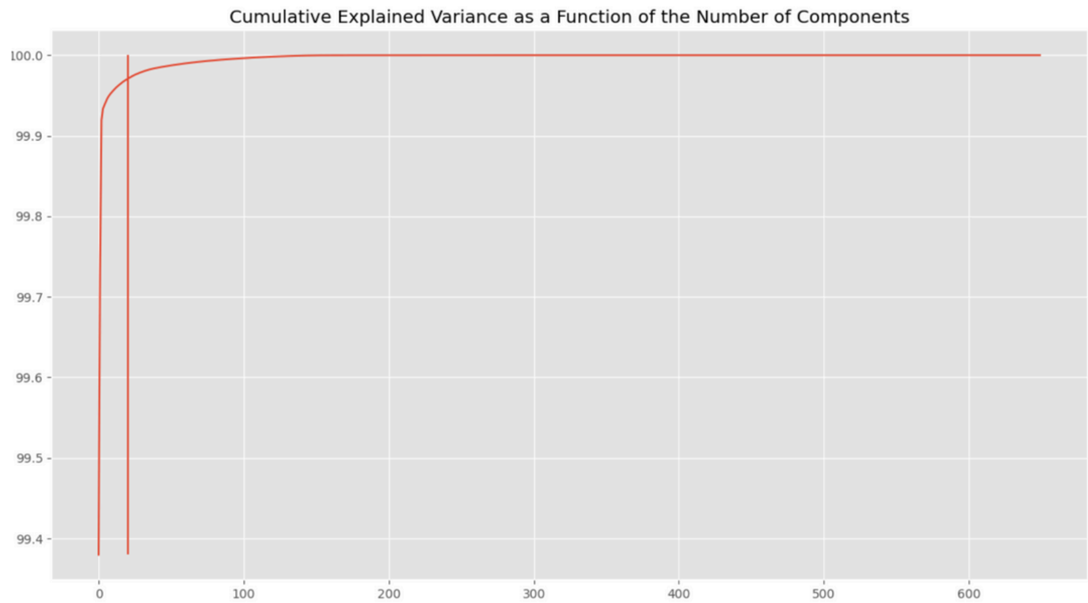
- PCA was applied to reduce dimensionality while retaining the most important variance-explaining features.

- This step optimizes model efficiency by removing redundant or less relevant features, improving generalization and reducing overfitting.
- The number of components was chosen based on the explained variance ratio, ensuring that most of the information is preserved while minimizing complexity.

Modeling Objective: Predicting Pass holder Type (Walk-Up vs. Monthly Pass):

  - The preprocessed dataset was used to build a classification model to predict whether a rider will choose a Walk-Up pass or a Monthly Pass at any given station.
  - The engineered features (Peak Hours, Seasonality, Categorical Encodings) combined with PCA-selected components provide a robust foundation for accurate predictions.
  - The resulting model aims to support Metro Bike Share in optimizing marketing strategies, station placement, and resource allocation based on regional user behavior.

These preprocessing steps ensure that the dataset is well-structured, optimized for modeling, and capable of delivering actionable insights for improving bike-share accessibility and user engagement.




```
Training set: 260646 samples
Validation set: 86882 samples
Test set: 86882 samples
```

Model Selection and Training

Models used:

Model	Use	Accuracy
Random Forest	*Since pass holder type prediction involves categorical classification, Random Forest is valuable due to its ability to handle non-linearity and feature interactions effectively.*	*82%*
XGBoost	*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 based on features like station popularity, trip duration, and ride frequency.*	*85%*
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.*	*79%*

Model Evaluation:

Random Forest:

The Random Forest model achieves a validation accuracy of 81.87%, indicating strong performance in distinguishing between Monthly Pass and Walk-up users. The classification report highlights a balanced precision-recall tradeoff: Monthly Pass users are correctly identified with 85% recall, while Walk-up users have a recall of 79%, suggesting the model is slightly better at capturing Monthly Pass users. Precision is relatively balanced between the two classes, at 81% for Monthly Pass and 83% for Walk-up, leading to an overall F1-score of 0.82.

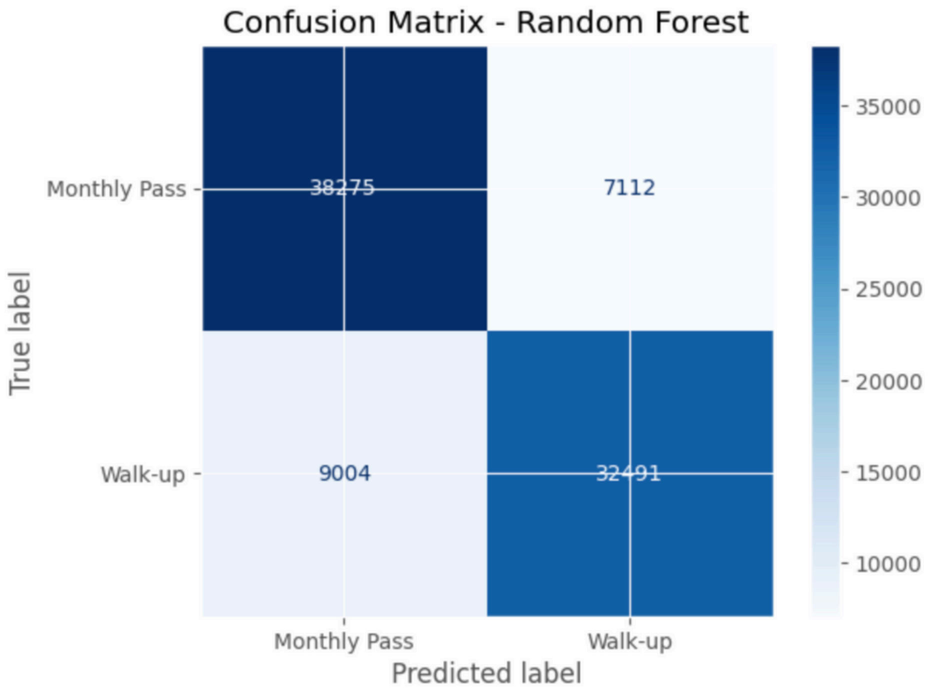
```
Validation Set Accuracy: 0.8187426624617297

Classification Report (Validation Set):
              precision    recall  f1-score   support

Monthly Pass       0.81      0.85      0.83     45444
     Walk-up       0.83      0.79      0.81     41438

    accuracy                           0.82     86882
   macro avg       0.82      0.82      0.82     86882
weighted avg       0.82      0.82      0.82     86882
```

The confusion matrix further confirms this trend. The model correctly classifies 38,175 Monthly Pass users but misclassifies 7,112 as Walk-up. Similarly, 32,091 Walk-up users are correctly identified, while 9,004 are misclassified as Monthly Pass users. The higher misclassification rate for Walk-up users suggests the model may slightly favor Monthly Pass predictions. This could be due to feature importance skewing toward characteristics more indicative of Monthly Pass users or an imbalance in feature representation.



XGBoost:

The XGBoost model demonstrates strong performance in classifying Monthly Pass and Walk-up users, with a validation accuracy of 79.89% and a test accuracy of 85.84%, indicating improved generalization compared to the Random Forest model. In the validation set, the model achieved 0.79 precision and 0.84 recall for Monthly Pass users, while Walk-up users had 0.81 precision and 0.76 recall, suggesting a slight bias toward correctly identifying Monthly Pass users.

```
XGBoost Validation Accuracy: 0.7988651274141939

XGBoost Classification Report:
              precision    recall  f1-score   support

Monthly Pass       0.79      0.84      0.81     45444
     Walk-up       0.81      0.76      0.78     41438

    accuracy                           0.80     86882
   macro avg       0.80      0.80      0.80     86882
weighted avg       0.80      0.80      0.80     86882
```

However, in the test set, performance significantly improved, with Monthly Pass achieving 0.84 precision and 0.89 recall, while Walk-up reached 0.88 precision and 0.82 recall, leading to an overall balanced F1-score of 0.86.

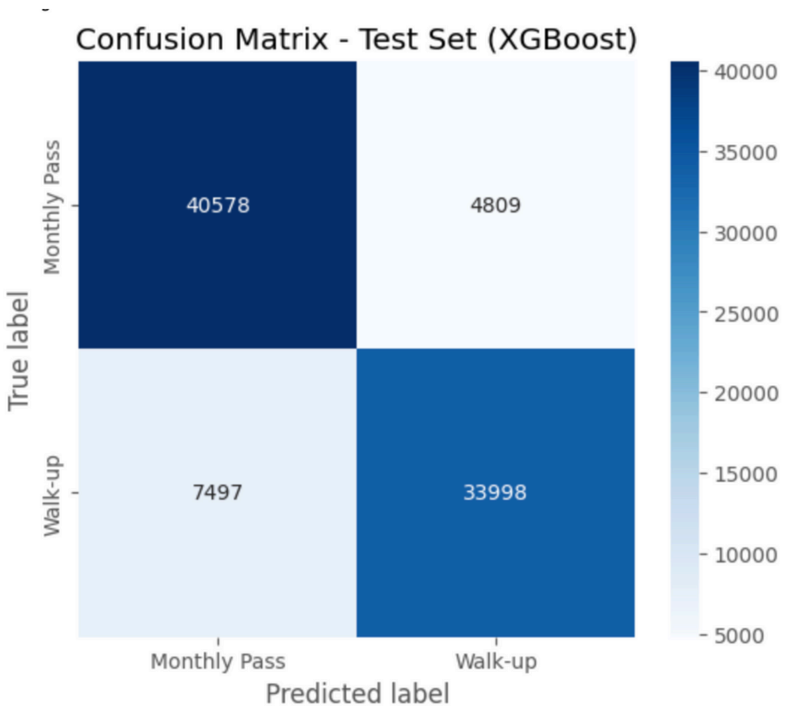
```
XGBoost Test Accuracy: 0.8583596141893602

XGBoost Classification Report:
              precision    recall  f1-score   support

Monthly Pass       0.84      0.89      0.87     45387
     Walk-up       0.88      0.82      0.85     41495

    accuracy                           0.86     86882
   macro avg       0.86      0.86      0.86     86882
weighted avg       0.86      0.86      0.86     86882
```

The confusion matrix further confirms that 40,578 Monthly Pass users and 33,998 Walk-up users were correctly classified, with misclassification rates lower than those observed in the Random Forest model. The improved performance on unseen data suggests that XGBoost effectively captures complex feature interactions and handles imbalanced distributions better.



FNN(Deep Learning):

The Feedforward Neural Network (FNN) model, consisting of four dense layers (128, 64, 32, and 1 neurons, respectively) with batch normalization and dropout layers, achieves a validation accuracy of 79.89%, which is on par with XGBoost but slightly lower than its test performance. The classification report indicates balanced precision and recall, with the Monthly Pass class (0) achieving 0.78 precision and 0.83 recall, while the Walk-up class (1) scores 0.80 precision and 0.74 recall, leading to an overall F1-score of 0.79.

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

Total params: 12,673 (49.50 KB)
Trainable params: 12,225 (47.75 KB)
Non-trainable params: 448 (1.75 KB)

This suggests that while the FNN captures meaningful patterns in the data, it slightly struggles with classifying Walk-up users, similar to previous models. The inclusion of batch normalization and dropout layers likely aids in generalization and stabilizing training, while the model's 12,225 trainable parameters keep it lightweight. However, compared to XGBoost, which achieved an F1-score of 0.86 on the test set, the FNN appears to underperform slightly, potentially due to hyperparameter choices or the need for deeper feature extraction. Further improvements could involve adjusting dropout rates, increasing model complexity, fine-tuning learning rates, or leveraging more advanced architectures such as CNNs or LSTMs if temporal dependencies exist in the data.

```
2716/2716 ━━━━━━━━━━━━━━━━━━━━ 1s 533us/step - accuracy: 0.7877 - loss: 0.4630
Validation Accuracy: 0.7865
2716/2716 ━━━━━━━━━━━━━━━━━━━━ 2s 552us/step
FNN Validation Accuracy: 0.7988651274141939

Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.83      0.80     45444
           1       0.80      0.74      0.77     41438

    accuracy                           0.79     86882
   macro avg       0.79      0.78      0.79     86882
weighted avg       0.79      0.79      0.79     86882
```

While the FNN demonstrates promising performance, XGBoost remains the stronger candidate for this classification task based on current results.

Model Performance Comparison Table

The table below summarizes the key performance metrics for each model:

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

Conclusion on Pass-holder Type Classification and Model Insights

The models developed for classifying Metro Bike Share users into Monthly Pass and Walk-up users have demonstrated strong performance, with XGBoost emerging as the best-performing model due to its high test accuracy (85.84%) and balanced precision-recall scores. The Random Forest models performed well, particularly in validation, but had slightly lower generalization capability. The Feedforward Neural Network (FNN), while achieving similar validation accuracy to XGBoost, struggled with distinguishing Walk-up users, suggesting that deep learning approaches might require additional feature engineering or architectural changes to be competitive.

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.

Further Model Improvements

While the current models perform well, there are several ways to improve prediction accuracy and usability:

1. Feature Engineering: 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.
2. 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.

3. Temporal and Sequential Modeling: 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.

Expanding the Use of This Dataset for Other Machine Learning Applications

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.
- How?
  - Implement time-series forecasting (ARIMA, LSTMs, XGBoost with lag features) to predict usage spikes.
  - Use clustering (e.g., K-Means, DBSCAN) to group stations based on similar demand patterns, helping with fleet redistribution.

2. Optimizing Bike Dock Distribution

- Goal: Determine the optimal number of bike docks per station based on actual usage trends.
- How?
  - Train a regression model (Random Forest Regressor, XGBoost, or Neural Networks) to estimate required docks at each station.
  - Use historical data to identify stations with frequent bike shortages or excess availability.
  - Example: If a model predicts that a station runs out of bikes frequently at 8 AM, Metro could increase dock capacity at that location.