UMBC

Urban Mobility Trends Unveiled: A Comprehensive Analysis of Lyft Bikes in San Francisco

Meta Team Members:

Divya Dodla (UA03330)

Naren Kandregula (PW71937)

Sachin Dudam (GG42052)

Project Description

GOAL : To analyze lyft Bay Wheels System Data of San Francisco Bay Area from the years 2021-2023 by performing Big Data operations, feature engineering, generating visualizations (EDA), and building prediction model based on the principles of Big Data to predict the potential bike availability at specific stations.

DATA :

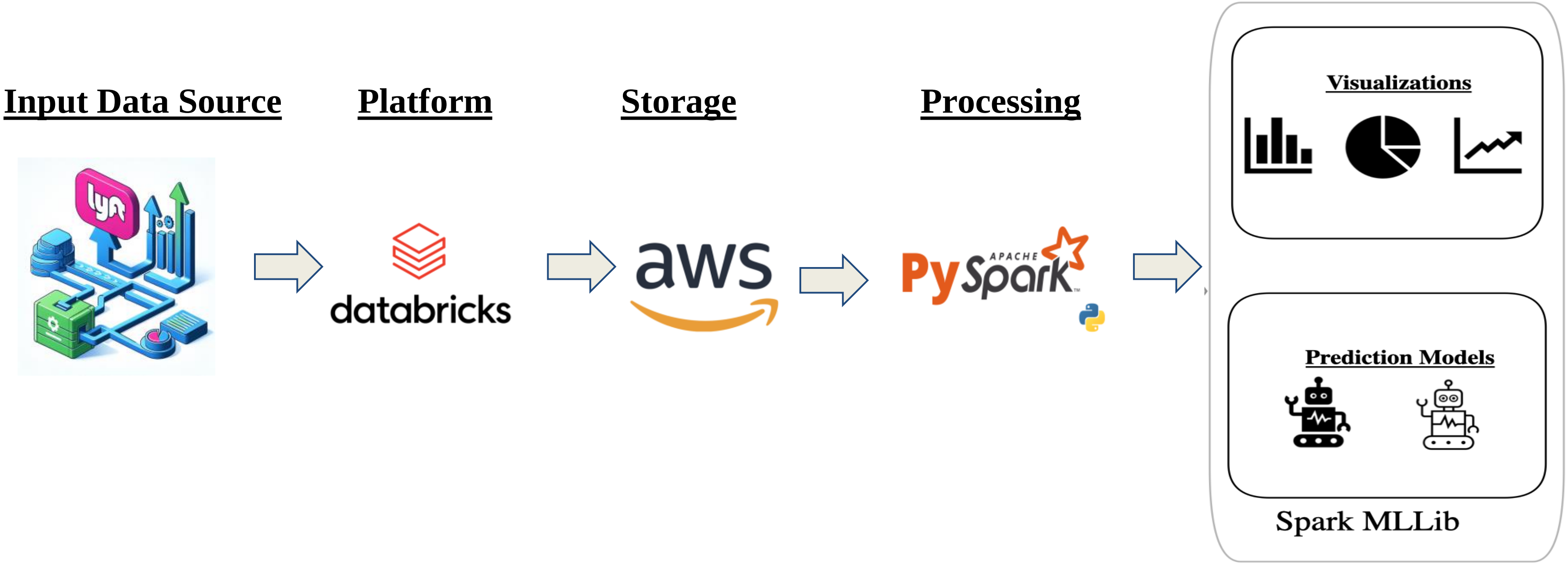
- Size of dataset: Approximately 1.24 GB.
- Number of Records: 6,625,912 individual records.

The large dataset size necessitates leveraging distributed computing capabilities, and the complexity of the analysis demands scalable solutions to extract actionable insights from the wealth of information within the Lyft Bay Wheels System Data

Motivation

- The increasing popularity of bike-sharing systems contributes significantly to sustainable urban transportation.
- Analyzing the Bay Wheels System Data allows us to gain insights into user behaviors, station utilization, and broader trends, aiding urban planners and policymakers in making informed decisions to enhance the efficiency and accessibility of bike-sharing services.

Implementation of Architecture



Input Data Source

Platform

Storage

Processing

Visualizations

Prediction Models

Spark MLLib

Feature Engineering & Exploratory Data Analysis (EDA)

Missing values in categorical columns (Station Names and IDs) are filled with "Unknown" and rows with missing values in geographic coordinates (Latitude and Longitude) are removed.

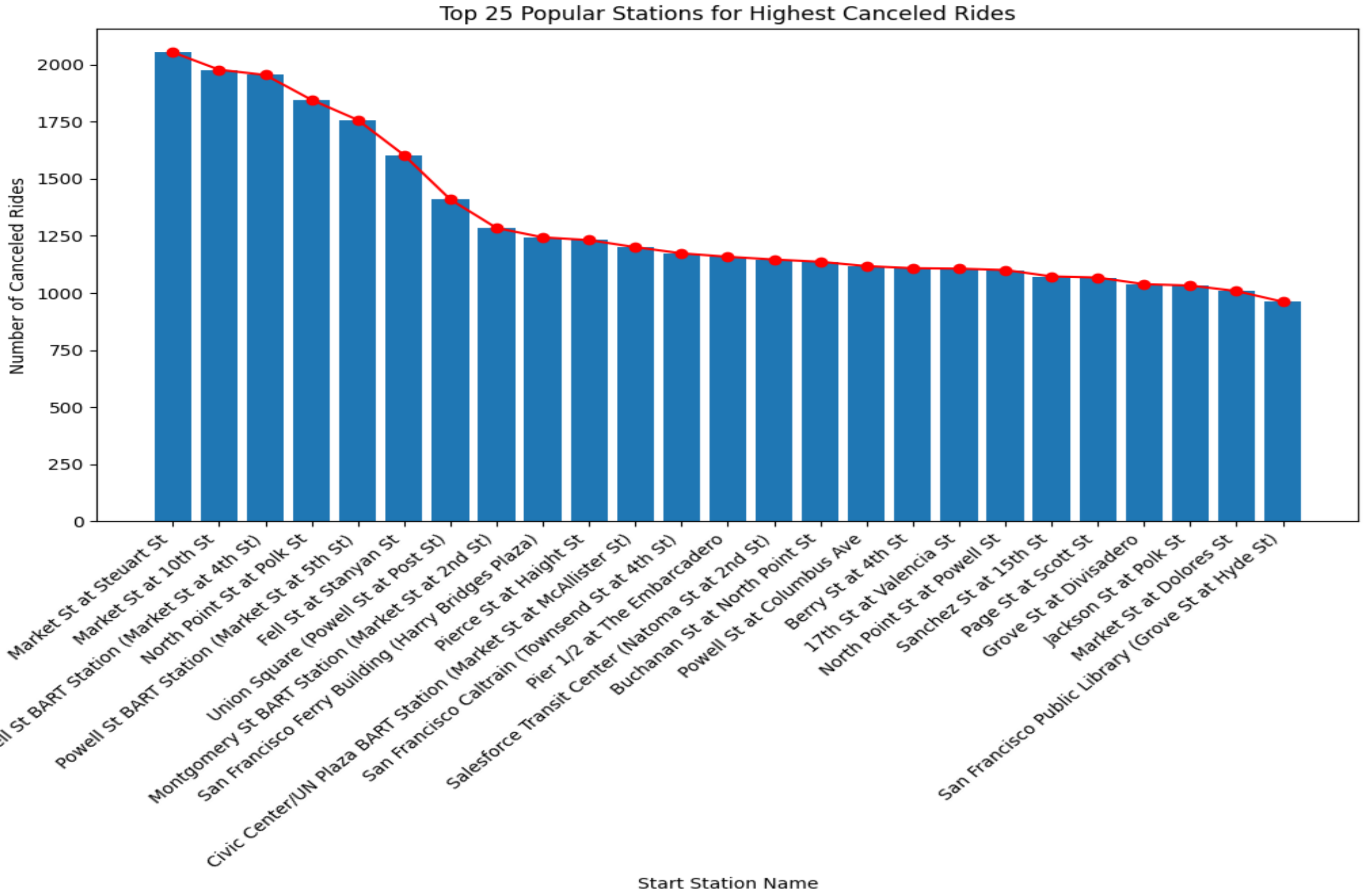
Potential Canceled Rides Analysis

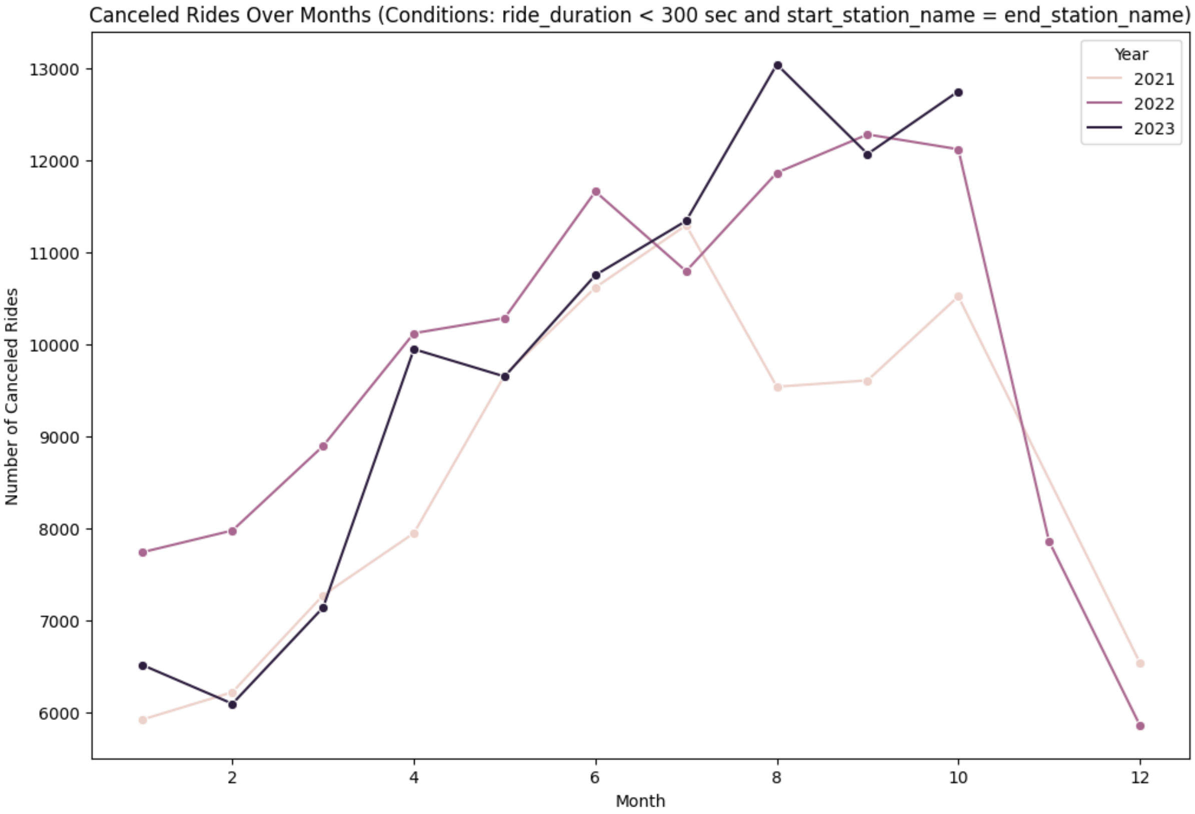
To filter the DataFrame to retain only valid rides by excluding two cases:

- Keeping rides with a duration of 5 minutes or more.
- Excluding rides with less than 5 minutes duration where the start and end stations are the same.

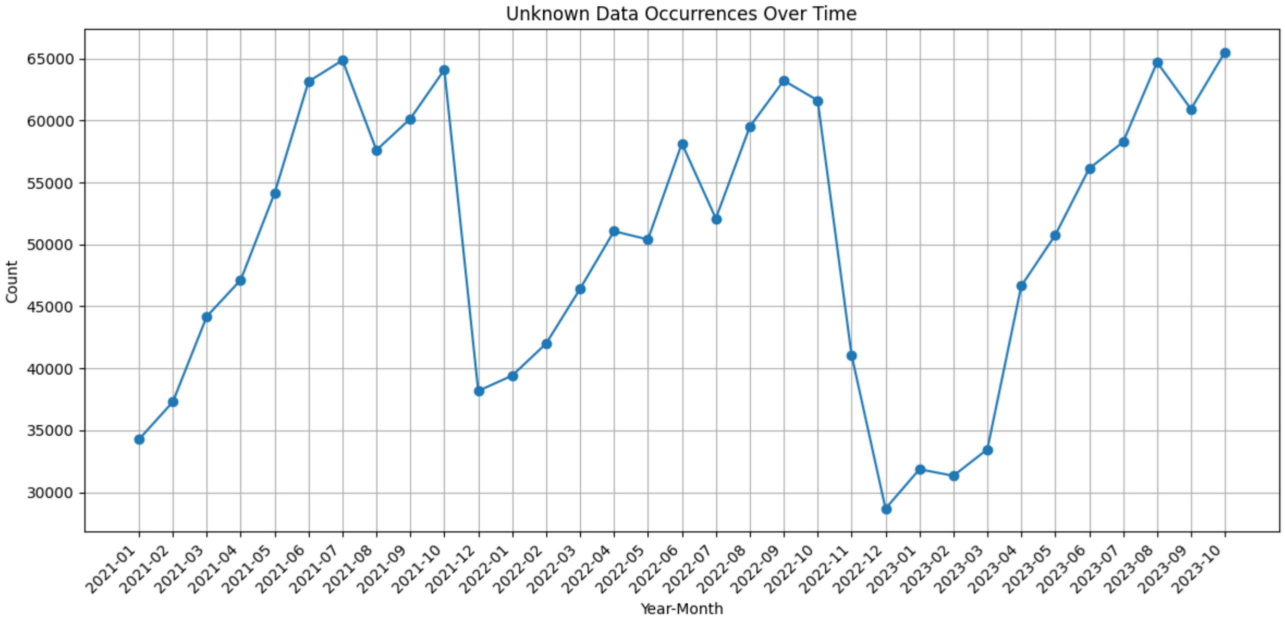
This ensures the Data Frame contains reliable ride data, filtering out potentially canceled or very short rides.

```
Number of potentially canceled rides: 311875
Total number of distinct stations: 603
```


Top 25 Popular Stations for Highest Canceled Rides
Number of Canceled Rides
0
250
500
750
1000
1250
1500
1750
2000
Market St at Steuart St
Market St at 10th St
Powell St BART Station (Market St at 4th St)
North Point St at Polk St
Powell St BART Station (Market St at 5th St)
Fell St at Stanyan St
Union Square (Powell St at Post St)
Montgomery St BART Station (Market St at 2nd St)
San Francisco Ferry Building (Harry Bridges Plaza)
Pierce St at Haight St
Civic Center/UN Plaza BART Station (Market St at McAllister St)
San Francisco Caltrain (Townsend St at 4th St)
Pier 1/2 at The Embarcadero
Salesforce Transit Center (Natoma St at 2nd St)
Buchanan St at North Point St
Powell St at Columbus Ave
Berry St at 4th St
17th St at Valencia St
North Point St at Powell St
Sanchez St at 15th St
Page St at Scott St
Grove St at Divisadero
Jackson St at Polk St
Market St at Dolores St
San Francisco Public Library (Grove St at Hyde St)
Start Station Name



The graph shows that there are more canceled rides in the summer months, and that the number of canceled rides has increased over time.



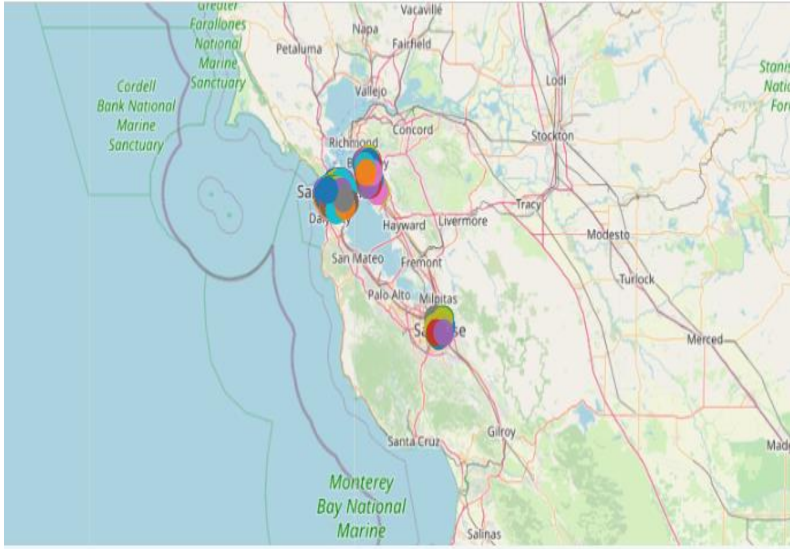
The unknown data is likely due to a number of factors, such as:

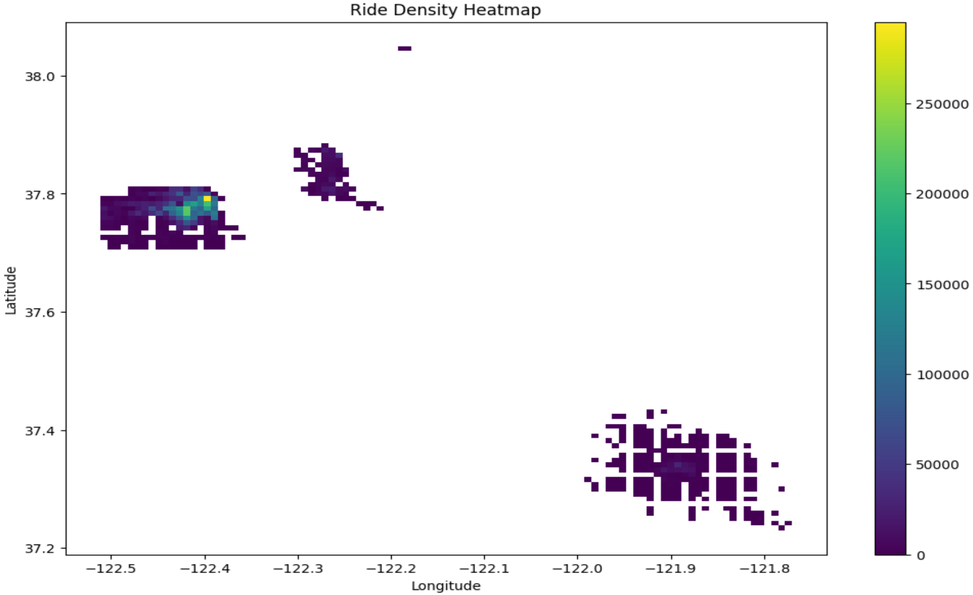
- GPS errors
- User input errors
- Stations that are no longer in service
- Stations that are not properly identified in the dataset

Unknown data occurrences in Lyft Bay wheels rides dataset decreasing over time, but still significant.

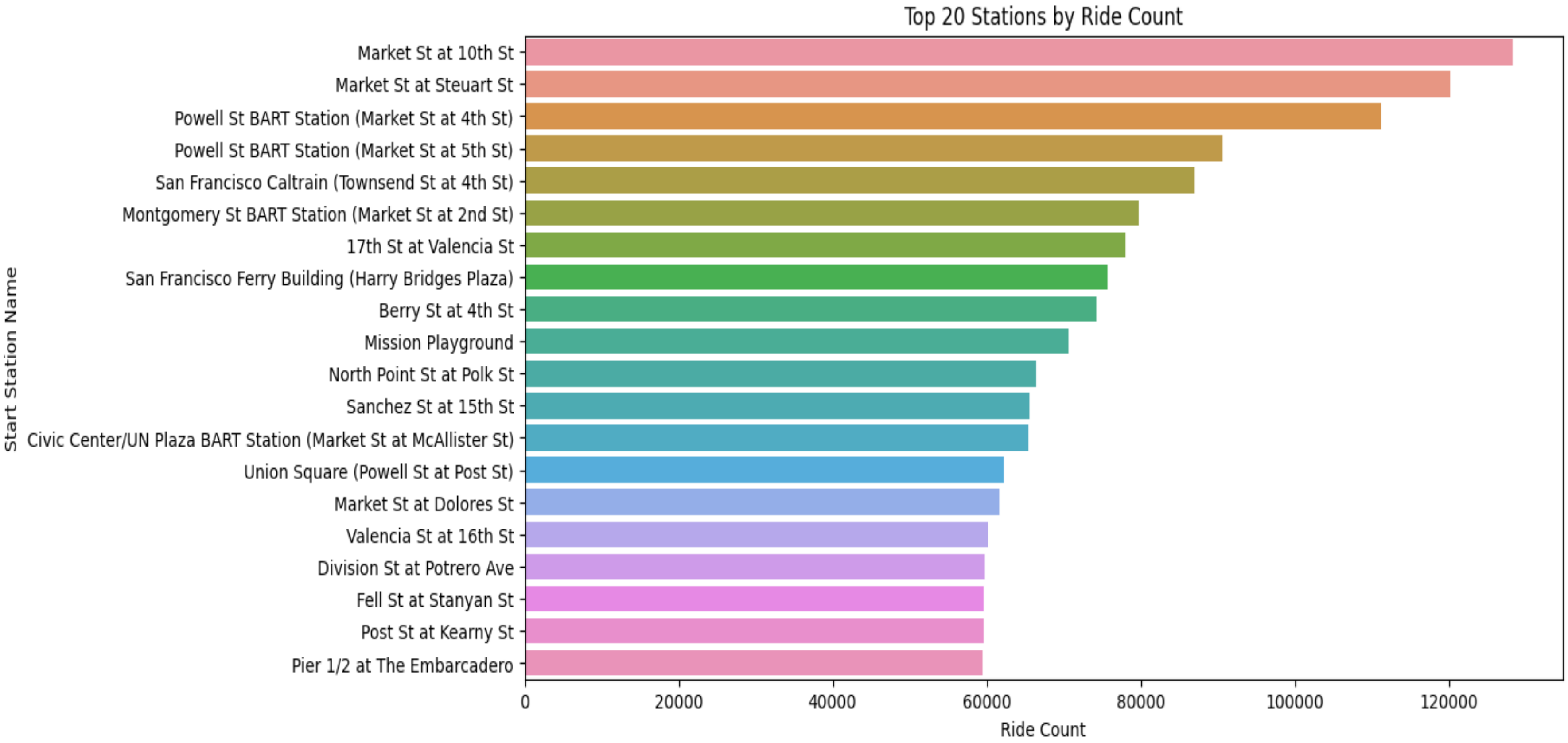
This insight suggests that Lyft is making progress in improving the accuracy and completeness of its dataset, but there is still room for improvement.

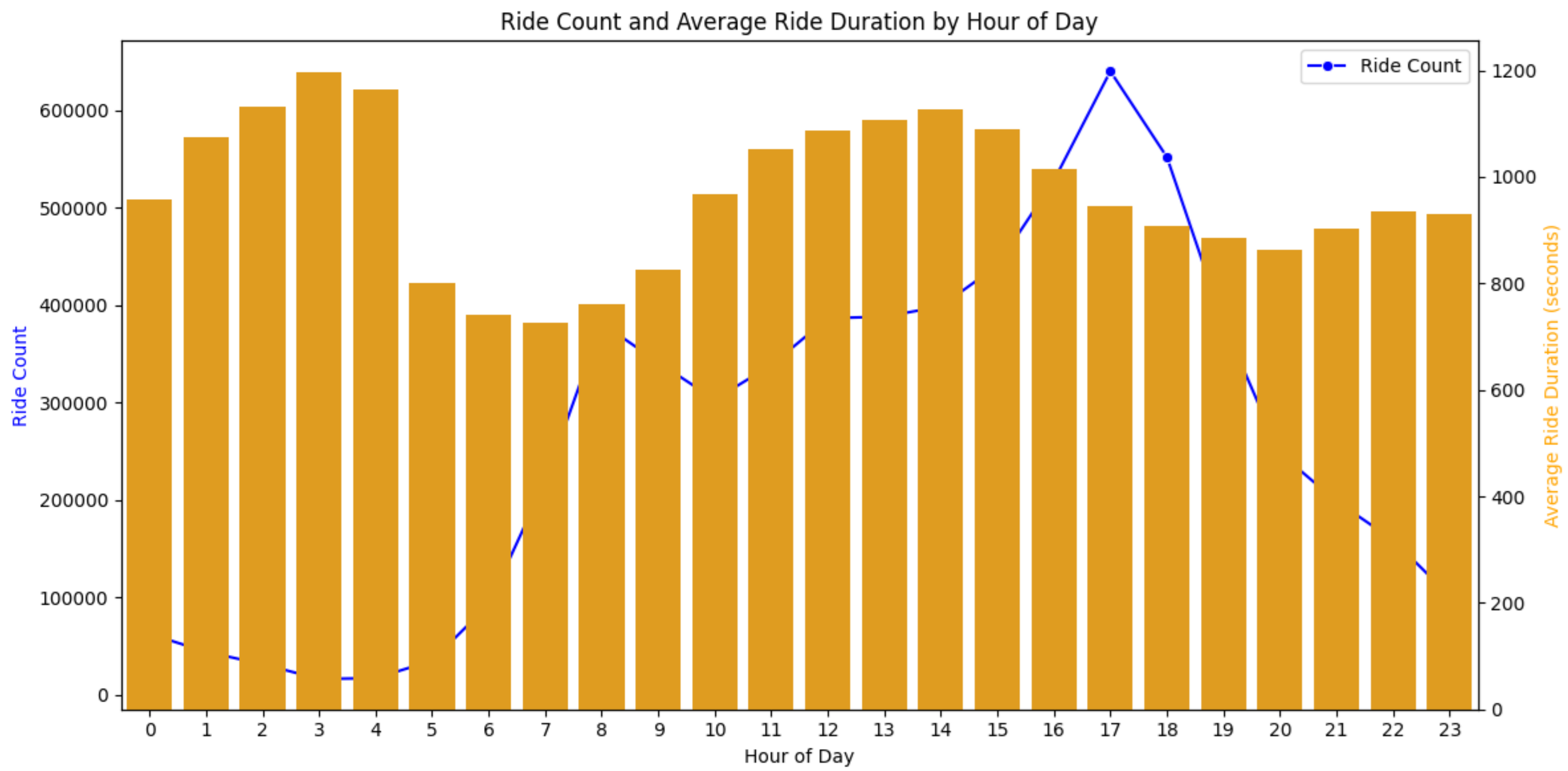
Stations



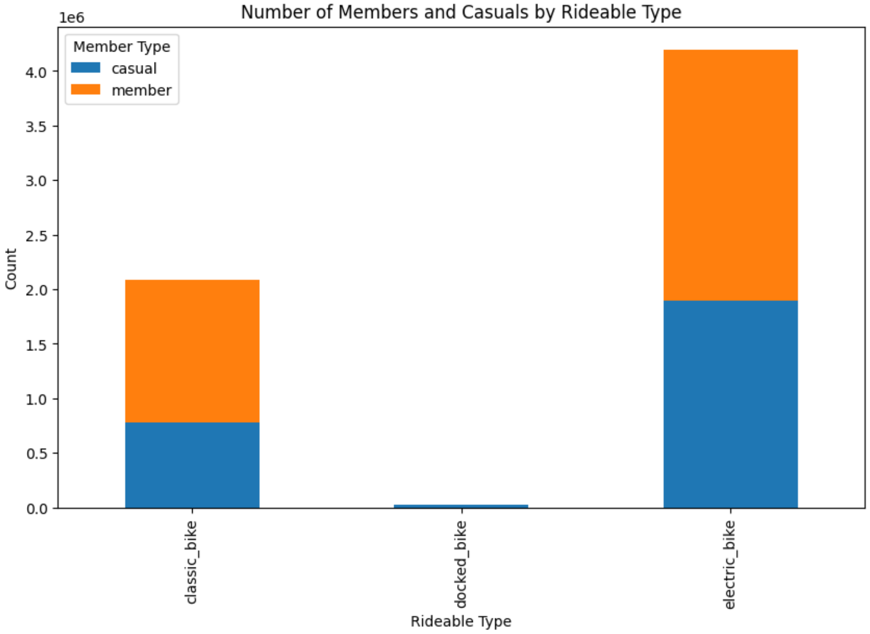


Lyft bike ride start stations are located throughout the San Francisco Bay Area, with a concentration in the downtown area.

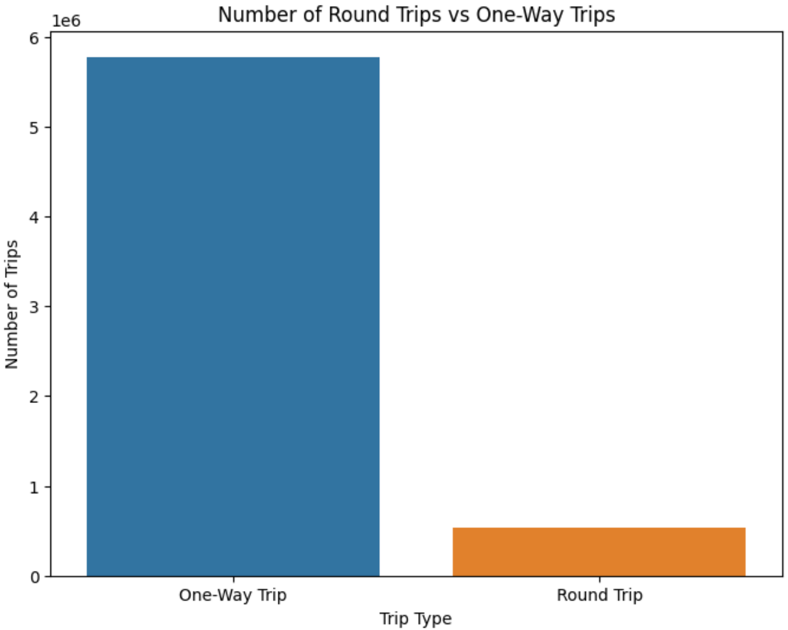

Top 20 Stations by Ride Count
Start Station Name
Market St at 10th St
Market St at Steuart St
Powell St BART Station (Market St at 4th St)
Powell St BART Station (Market St at 5th St)
San Francisco Caltrain (Townsend St at 4th St)
Montgomery St BART Station (Market St at 2nd St)
17th St at Valencia St
San Francisco Ferry Building (Harry Bridges Plaza)
Berry St at 4th St
Mission Playground
North Point St at Polk St
Sanchez St at 15th St
Civic Center/UN Plaza BART Station (Market St at McAllister St)
Union Square (Powell St at Post St)
Market St at Dolores St
Valencia St at 16th St
Division St at Potrero Ave
Fell St at Stanyan St
Post St at Kearny St
Pier 1/2 at The Embarcadero
0
20000
40000
60000
80000
100000
120000
Ride Count



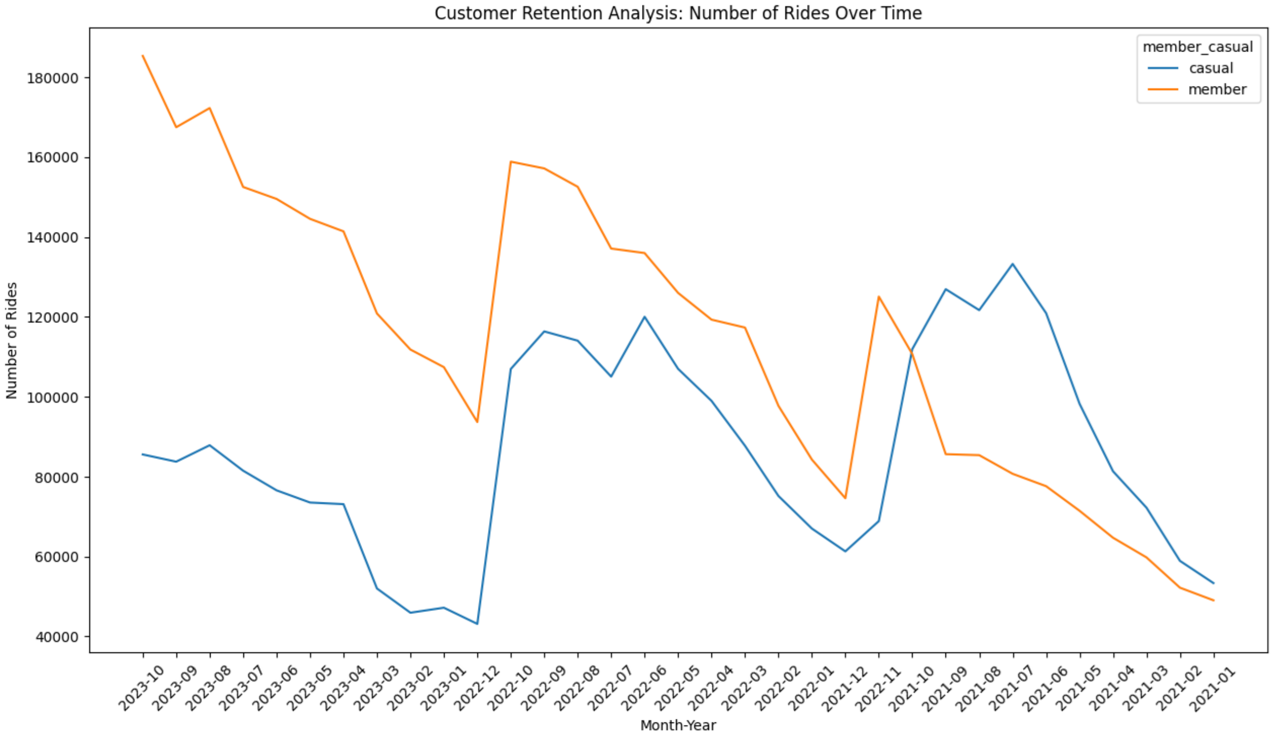
This insight suggests that people are more likely to take shorter rides during the day, and longer rides at night.



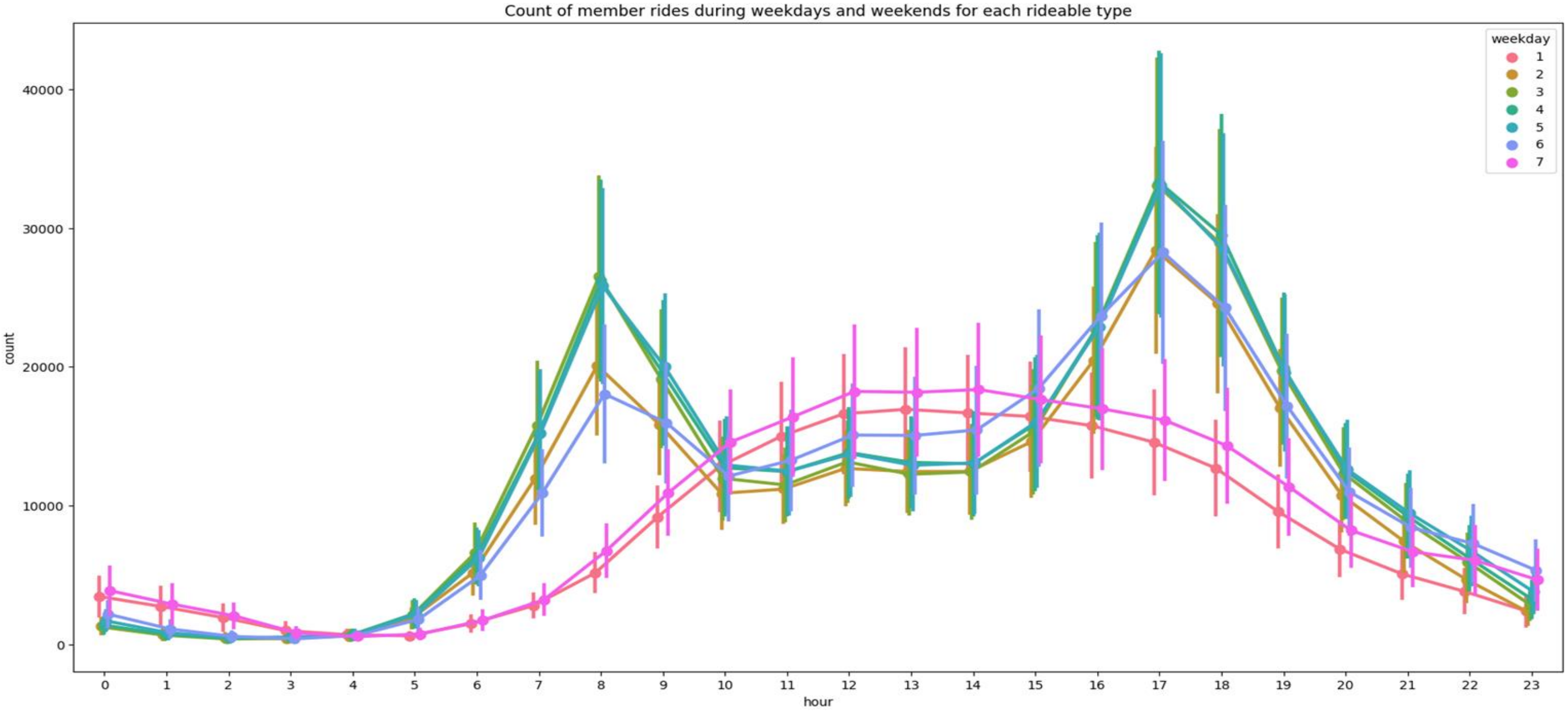
```
+-------------+-------+
|rideable_type|  count|
+-------------+-------+
|electric_bike|4197496|
| classic_bike|2085406|
|  docked_bike|  25252|
+-------------+-------+
```



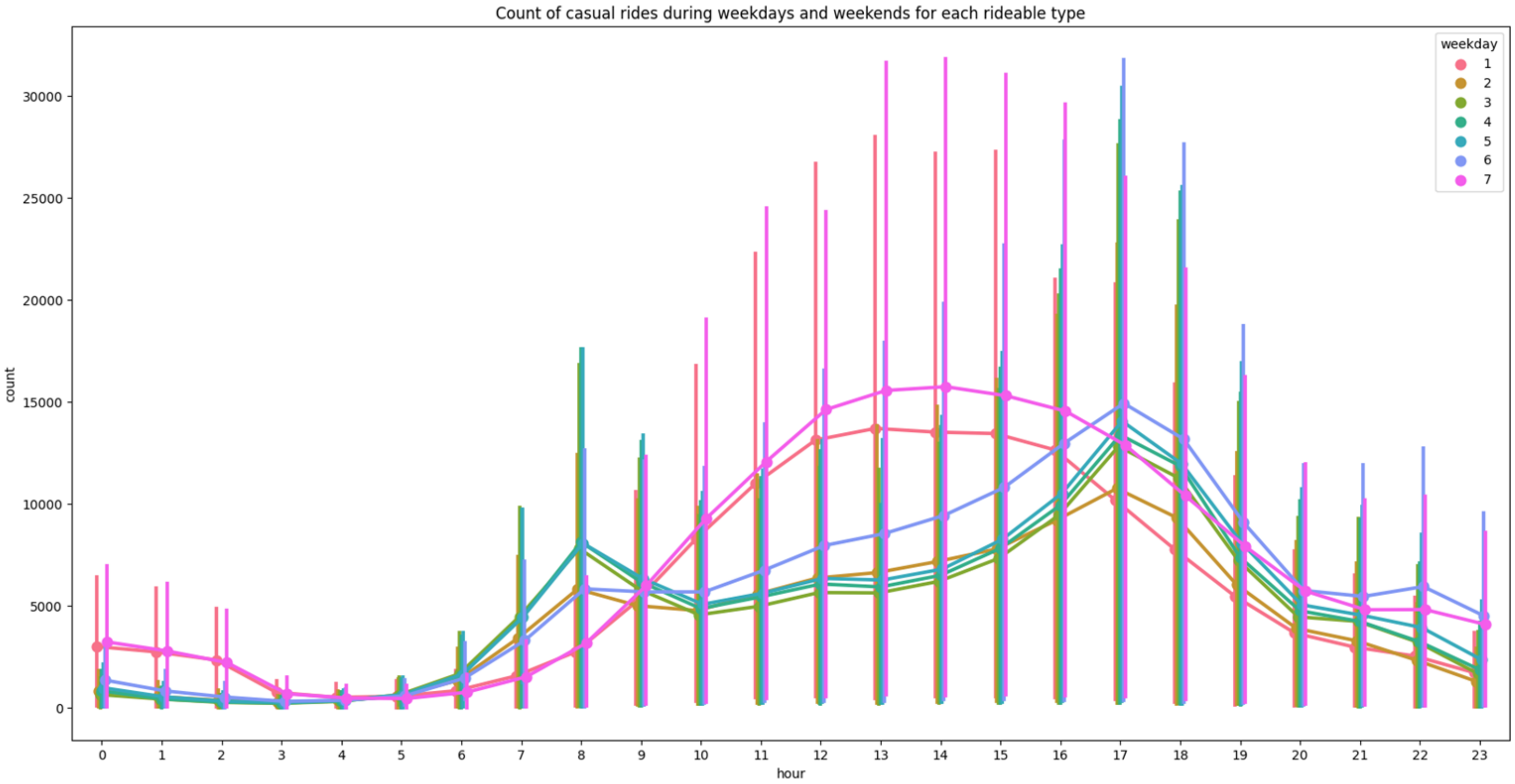
```
+-------------+-------+
|member_casual|  count|
+-------------+-------+
|       casual|2702435|
|       member|3605719|
+-------------+-------+
```



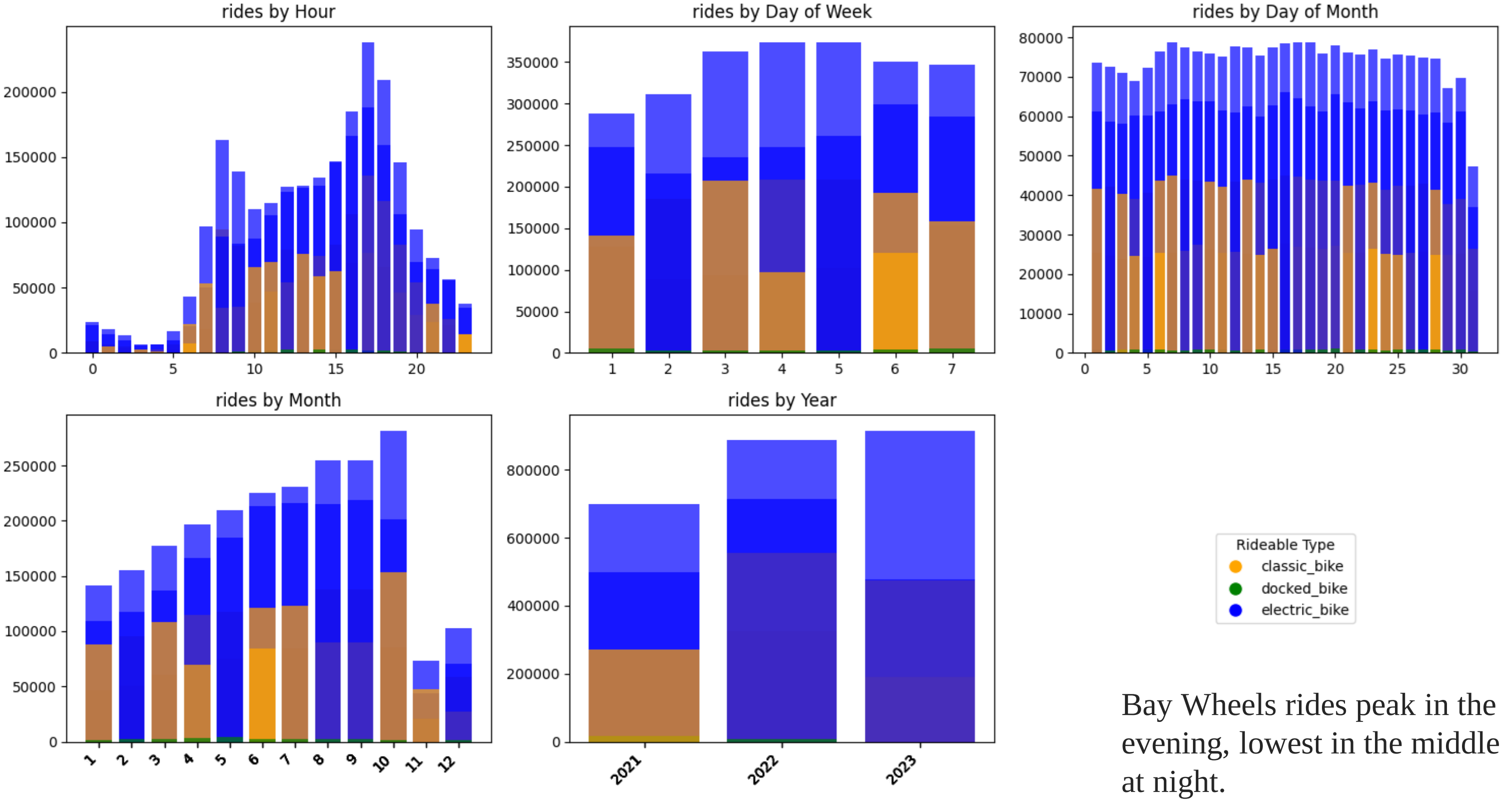
Lyft member rides increasing over time, while casual rider rides remain stable.
This insight suggests that Lyft is becoming more popular among regular riders, which could be due to a number of factors, such as the convenience and affordability of the service.



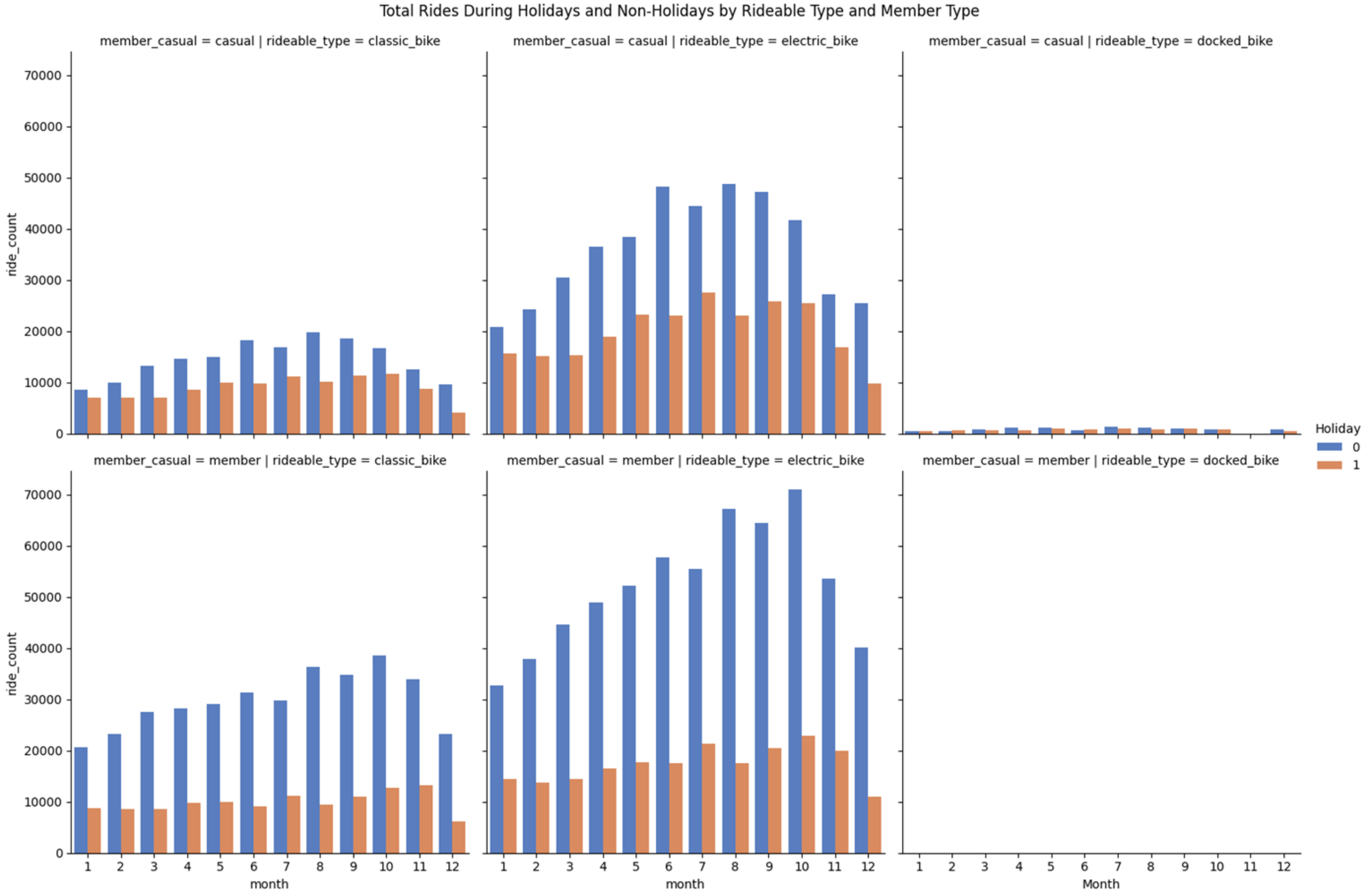
Average number of member rides higher on weekdays than on weekends.



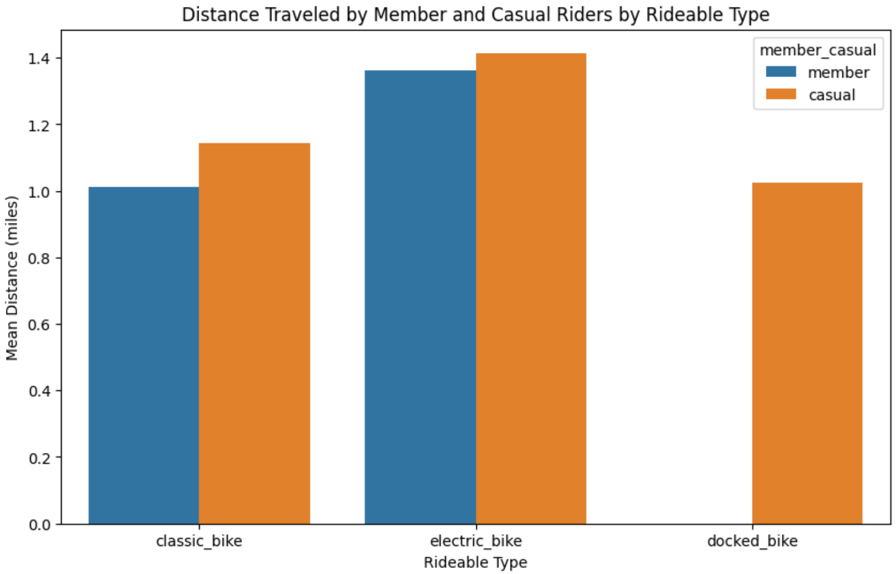
Average number of Casual rides higher on weekends than weekdays.

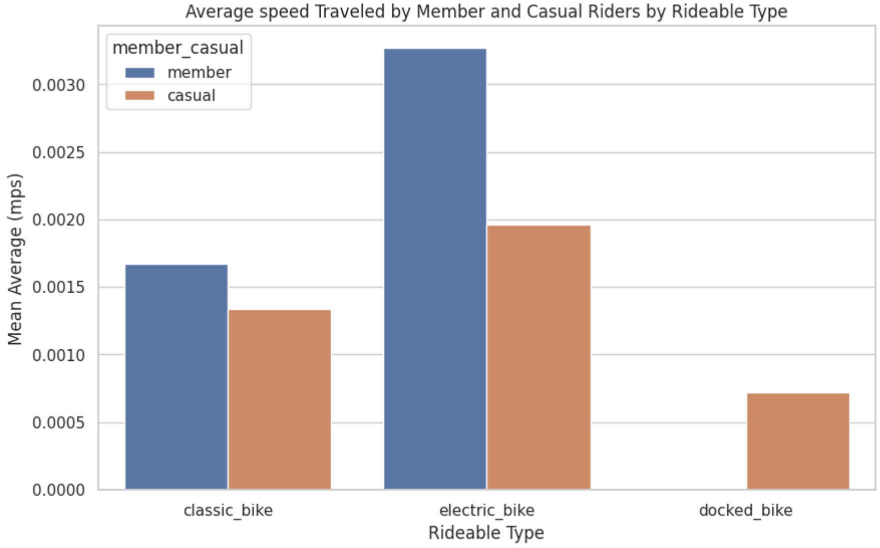


Bay Wheels rides peak in the morning and evening, lowest in the middle of the day and at night.



The holiday column includes weekends (Saturday and Sunday) and other specified holidays, adopted from [Office Holidays - California, USA](#).

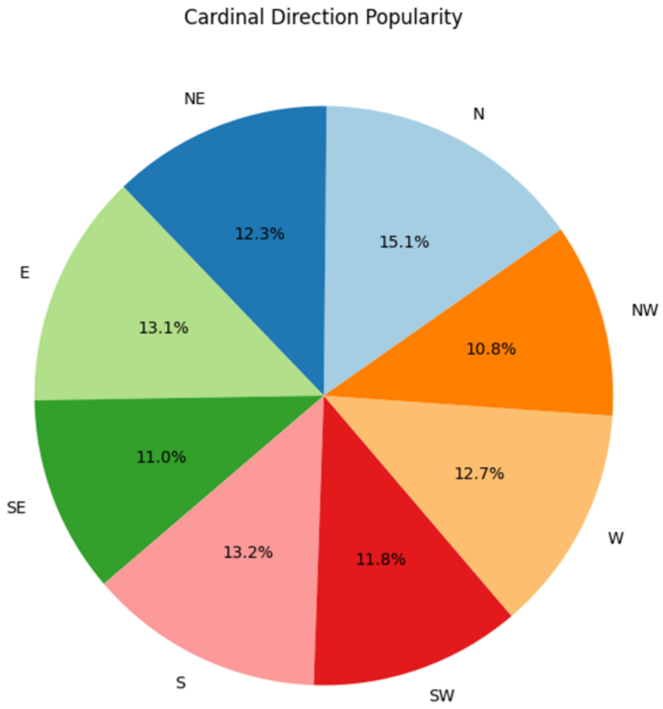




Casual riders travel further distances than member riders, regardless of rideable type.

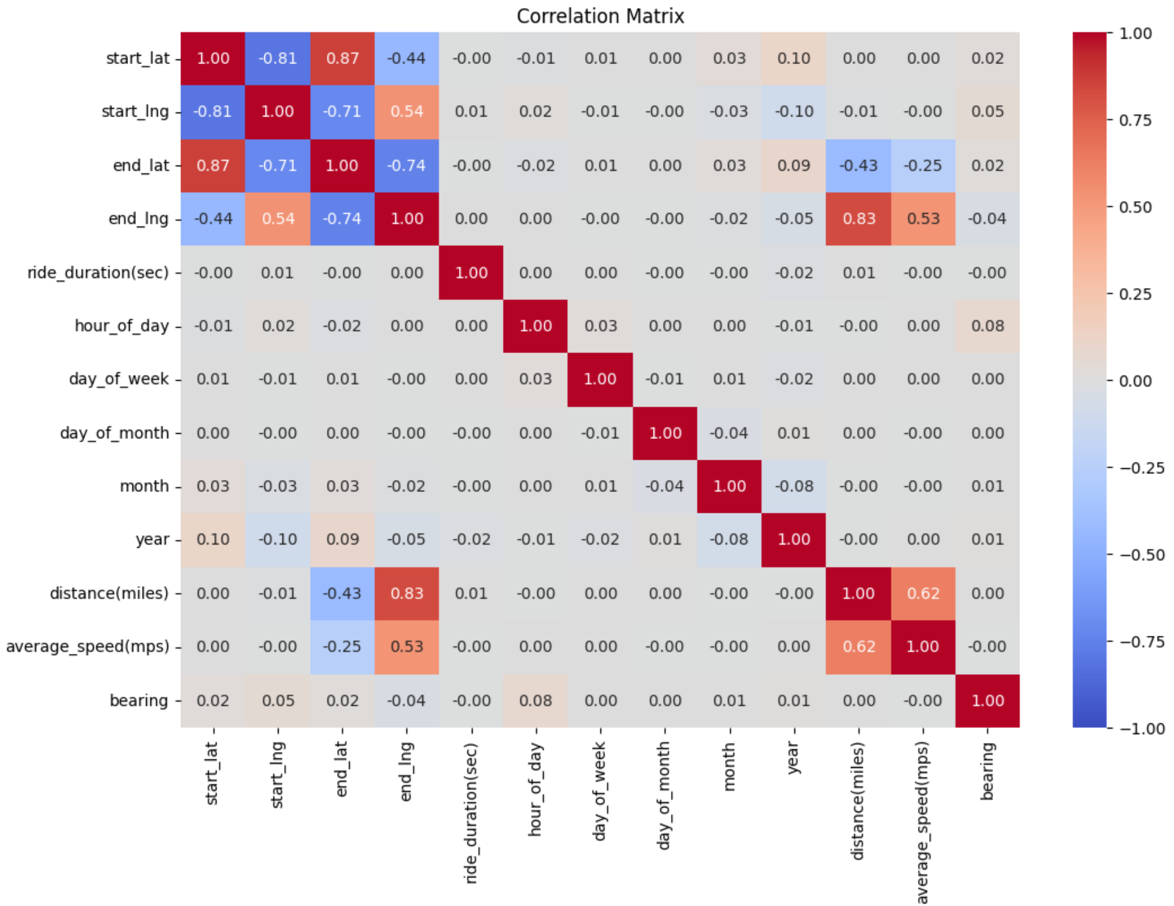
Member riders travel faster than casual riders.

The geolocation distance calculation using the Haversine formula, as implemented through a User-Defined Function (UDF) and inspired by the article 'Calculating Distance Between Two Geolocations in Python', adds a new 'distance(miles)' column to the DataFrame using the latitude and longitudinal information and duration of rides.



The largest segment, representing rides heading North, indicates a higher preference or need for travel in this direction. It might also reflect commuting patterns, which could be vital for urban planning and expansion of services. Understanding why certain directions are preferred could help Lyft optimize its network to better serve the community's transportation needs.

Inspired by the method detailed in the mapscaping reference, a User-Defined Function (UDF) is implemented to calculate the bearing between geographical coordinates, offering valuable insights into the direction of movement for analyzing ride patterns and trends in the dataset.



Correlation Matrix

	start_lat	start_lng	end_lat	end_lng	ride_duration(sec)	hour_of_day	day_of_week	day_of_month	month	year	distance(miles)	average_speed(mps)	bearing
start_lat	1.00	-0.81	0.87	-0.44	-0.00	-0.01	0.01	0.00	0.03	0.10	0.00	0.00	0.02
start_lng	-0.81	1.00	-0.71	0.54	0.01	0.02	-0.01	-0.00	-0.03	-0.10	-0.01	-0.00	0.05
end_lat	0.87	-0.71	1.00	-0.74	-0.00	-0.02	0.01	0.00	0.03	0.09	-0.43	-0.25	0.02
end_lng	-0.44	0.54	-0.74	1.00	0.00	0.00	-0.00	-0.00	-0.02	-0.05	0.83	0.53	-0.04
ride_duration(sec)	-0.00	0.01	-0.00	0.00	1.00	0.00	0.00	-0.00	-0.00	-0.02	0.01	-0.00	-0.00
hour_of_day	-0.01	0.02	-0.02	0.00	0.00	1.00	0.03	0.00	0.00	-0.01	-0.00	0.00	0.08
day_of_week	0.01	-0.01	0.01	-0.00	0.00	0.03	1.00	-0.01	0.01	-0.02	0.00	0.00	0.00
day_of_month	0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.01	1.00	-0.04	0.01	0.00	-0.00	0.00
month	0.03	-0.03	0.03	-0.02	-0.00	0.00	0.01	-0.04	1.00	-0.08	-0.00	-0.00	0.01
year	0.10	-0.10	0.09	-0.05	-0.02	-0.01	-0.02	0.01	-0.08	1.00	-0.00	0.00	0.01
distance(miles)	0.00	-0.01	-0.43	0.83	0.01	-0.00	0.00	0.00	-0.00	-0.00	1.00	0.62	0.00
average_speed(mps)	0.00	-0.00	-0.25	0.53	-0.00	0.00	0.00	-0.00	-0.00	0.00	0.62	1.00	-0.00
bearing	0.02	0.05	0.02	-0.04	-0.00	0.08	0.00	0.00	0.01	0.01	0.00	-0.00	1.00

The correlation matrix shows how different features of Lyft Bay wheels rides are related to each other.

Predicting Potential Bike Availability at Specific Stations

Bike Availability Prediction Workflow:

- Calculate hourly ride counts, create DataFrames, and join.
- Handle missing values, aggregate counts, and compute availability.
- Merge availability info with the original data.
- Select features, transform for ML, and train Machine Learning Algorithms.

Overall, the ML encompasses preprocessing, feature engineering, and machine learning to predict potential bike availability.

Availability column Assumptions:

- Positive differences imply increased availability, assuming immediate reflection.
- Negative differences set to 0 uniformly, potentially overlooking fluctuations.
- Binary classification (0 or 1) simplifies availability, missing nuanced variations.
- Null values filled with 0 assume no rides mean no availability, potentially inaccurately handling missing data.

Machine Learning Model Performance Evaluation

The ROC (Receiver Operating Characteristic) curve is a metric for evaluating the performance of classification models. Accuracy is the proportion of correctly classified instances out of the total instances. Here are the ROC curve scores and Accuracy scores for different machine learning models:

- Logistic Regression: 0.644 (ROC), 0.843(Accuracy)
- Random Forest: 0.688(ROC), 0.782(Accuracy)
- Multilayer Perceptron Classifier: 0.627(ROC), 0.843(Accuracy)

A higher ROC score and Higher Accuracy indicates better model performance in distinguishing between classes. The Logistic Regression Classifier demonstrates the highest ROC and Accuracy scores among the evaluated models, suggesting superior discriminatory power.

Insights & Urban Mobility Impact

- Lyft is making it easy for people to find a bike to ride, no matter where they are in the city. This can help to encourage people to choose biking as a transportation option, which can have a number of benefits, such as reducing traffic congestion and improving air quality.
- Electric bikes are popular among casual riders, covering longer distances at higher speeds compared to classic bikes.Peak ride hours and daily patterns reveal user behavior, aiding operational planning and resource allocation.
- Ride counts fluctuate during holidays; understanding these patterns helps optimize services and promotions.
- Cardinal direction analysis uncovers preferred routes, guiding station placements for enhanced accessibility.
- Multilayer Perceptron model outperforms, providing reliable predictions of ride availability.
- Data-driven insights reveal general user trends, guiding business strategies and user experience improvements.
- The Urban mobility trends from this analysis will contribute valuable insights to urban planners, policymakers, and transportation enthusiasts interested in understanding and optimizing the dynamics of bike-sharing systems.

Future Scope

- Apply the developed framework and insights to other cities with similar bike-sharing systems, adapting the model to local variations. This could contribute to a scalable solution applicable in diverse urban environments.
- Can develop a model to suggest optimal routes for rides based on historical traffic patterns, road closures, and other relevant factors.
- Can create a model to optimize the allocation of vehicles to different areas based on predicted demand and user behaviour.

By focusing on these areas, Lyft can potentially improve service reliability, customer satisfaction, and operational efficiency, ultimately enhancing the user experience and possibly reducing the rate of canceled rides.

References

System data: Bay Wheels. Lyft. (n.d.). https://www.lyft.com/bikes/bay-wheels/system-data.

Shaw, T. (2020, January 1). *Federal holidays in California in 2023*. Office Holidays. https://www.officeholidays.com/countries/usa/california.

Bhardwaj, A. (2020, June 14). *Calculating distance between two geolocations in python*. Medium. https://towardsdatascience.com/calculating-distance-between-two-geolocations-in-python-26ad3afe287b.

How to calculate bearing between two coordinates - December 4, 2023. mapscaping.com. (2023, November 7). https://mapscaping.com/how-to-calculate-bearing-between-two-coordinates/#:~:text=In%20Microsoft%20Excel%2C%20you%20can,x%2Daxis%20and%20the%20point.

Thank You…