

👉 **Portfolio Project** | Bay Wheels User Analysis with SQL

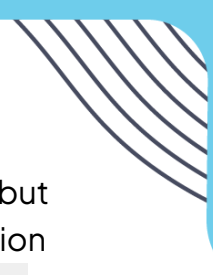
INTRODUCTION: To help the marketing team use data-driven approaches in their new marketing efforts. Investigate the differences between Lyft users and Ford users. Lyft wants to increase memberships in its rideshare program and needs to determine how their users, both past and present, use their product.

– Data Set **Description**

- 3 datasets: `lyft.baywheels`, `ford.gobike`, and the `sf.weather` dataset.

The `lyft.baywheels` dataset reports information about rentals made on the Bay Wheels bike share system. Each row represents a single rental; we will be making use of the following fields in this project:

- **started_date** - Date for start of rental
- **started_at** - Timestamp for start of rental
- **ended_at** - Timestamp for end of rental
- **start_station_name** - For rentals that started from a bike dock, the name of the dock.
- **end_station_name** - For rentals that ended at a bike dock, the name of the dock.
- **start_lat**, **start_lng** - Latitude and longitude, respectively, of the start of the rental.
- **end_lat**, **end_lng** - Latitude and longitude, respectively, of the end of the rental.
- **member_casual** - String indicating whether the rental was made by a system “member”, who has a monthly subscription with the bikeshare system, or by a “casual” user, who is making a one-time rental.



The `ford.gobike` dataset has information very similar to the `lyft.baywheels` table, but reports rides prior to Lyft's takeover of the bikeshare system. One major distinction between the two tables is different field names. The field names in the `ford.gobike` dataset will be explained through the course of the project tasks.

The `sf.weather` dataset contains daily weather statistics recorded at SF International Airport through 2020. We will be concerned with the following three features in this project:

- **date** - Date of weather recordings
 - **temperature_avg** - Average temperature in Fahrenheit
 - **precipitation** - Recorded precipitation in inches
-

– Task 1: Top User Engagement

First need to combine the data needed from Ford and Lyft. Below is a table of equivalent columns between the two datasets, detailing which columns in the `lyft.baywheels` data set matches which columns in the `ford.gobike` data table.

| Lyft Bay Wheels | Ford GoBike |
|--------------------|-------------------------|
| started_date | start_date |
| started_at | start_time |
| ended_at | end_time |
| start_station_name | start_station_name |
| end_station_name | end_station_name |
| start_lat | start_station_latitude |
| start_lng | start_station_longitude |
| end_lat | end_station_latitude |
| end_lng | end_station_longitude |
| member_casual | user_type |

A. Filter the `ford.gobike` data to only include data from the year 2020.

```
SELECT *  
FROM ford.gobike  
WHERE date_part('year', your_date_column) = 2020
```

- B. Write a query that unions the `ford.gobike` dataset and the `lyft.baywheels` still at the year 2020.

```
SELECT started_date, started_at, ended_at, start_station_name,
end_station_name, start_lat, start_lng, end_lat, end_lng,
member_casual
FROM lyft.baywheels
UNION
SELECT start_date, start_time, end_time, start_station_name,
end_station_name, start_station_latitude,
start_station_longitude, end_station_latitude,
end_station_longitude, user_type
FROM ford.gobike
WHERE date_part('year', start_date) = 2020
```

Create a new column called `data_source` that has the value 'Lyft' if the data came from the Lyft dataset and the value 'Ford' if it came from the Ford dataset.

```
SELECT 'Lyft' AS data_source, started_date, started_at,
ended_at, start_station_name, end_station_name, start_lat,
start_lng, end_lat, end_lng, member_casual
FROM lyft.baywheels
UNION
SELECT 'Ford' AS data_source, start_date, start_time,
end_time, start_station_name, end_station_name,
start_station_latitude, start_station_longitude,
end_station_latitude, end_station_longitude, user_type
FROM ford.gobike
WHERE date_part('year', start_date) = 2020
```

Store it specially in your schema. **For the remainder of this project, you'll query** `project.ford_lyft_analysis`.

– Task 2: Preparing the Data and Creating New Features

Create additional variables!

- A.** The `member_casual` column is supposed to indicate whether the rental was made by a system “member”, who has a monthly subscription, or by a “casual” user, who is making a one-time rental. The `member_casual` column actually has *four* different values: ‘member’, ‘Subscriber’, ‘casual’, and ‘Customer’. This is because Ford referred to its members as ‘Subscribers’ and its casual users as ‘Customer’ in its data.

Return all the variables from `project.ford_lyft_analysis`, plus a new variable called “`member_type`”, that contains **only values that match the Lyft classifications: ‘member’ or ‘casual’**.

```
WITH project_ford_lyft_analysis AS (  
  SELECT 'Lyft' AS data_source, started_date, started_at,  
    ended_at, start_station_name, end_station_name, start_lat,  
    start_lng, end_lat, end_lng, member_casual  
  FROM lyft.baywheels  
  UNION  
  SELECT 'Ford' AS data_sourcetwo, start_date, start_time,  
    end_time, start_station_name, end_station_name,  
    start_station_latitude, start_station_longitude,  
    end_station_latitude, end_station_longitude, user_type  
  FROM ford.gobike  
  WHERE date_part('year', start_date) = 2020)  
  
SELECT *  
FROM project.ford_lyft_analysis  
WHERE member_casual = 'member'  
OR member_casual = 'Subscriber'
```

B. Incorporate weather data into the analysis.

San Francisco's average daily temperature and amount of precipitation are the best metrics to base the weather analysis on. These are located in the `temperature_avg` and `precipitation` columns, respectively, of the `sf.weather` table. Join the table with the `sf_weather` data on the `started_date` field. From the `sf_weather` table, return the average daily temperature, and the amount of precipitation.

```
WITH project_ford_lyft_analysis AS (  
  SELECT 'Lyft' AS data_source, started_date, started_at,  
    ended_at, start_station_name, end_station_name, start_lat,  
    start_lng, end_lat, end_lng, member_casual  
  FROM lyft.baywheels  
  UNION  
  SELECT 'Ford' AS data_sourcetwo, start_date, start_time,  
    end_time, start_station_name, end_station_name,  
    start_station_latitude, start_station_longitude,  
    end_station_latitude, end_station_longitude, user_type  
  FROM ford.gobike  
  WHERE date_part('year', start_date) = 2020)  
  
SELECT *  
FROM project.ford_lyft_analysis AS f  
INNER JOIN  
  sf.weather AS w  
ON f.started_date = w.date  
WHERE member_casual = 'member'  
OR member_casual = 'Subscriber'
```

This query will result in almost 2 million records for the year 2020! It was loaded into a Tableau Workbook, where the rest of the project will take place.

– Task 3: Visualizing and Analyzing Using Tableau

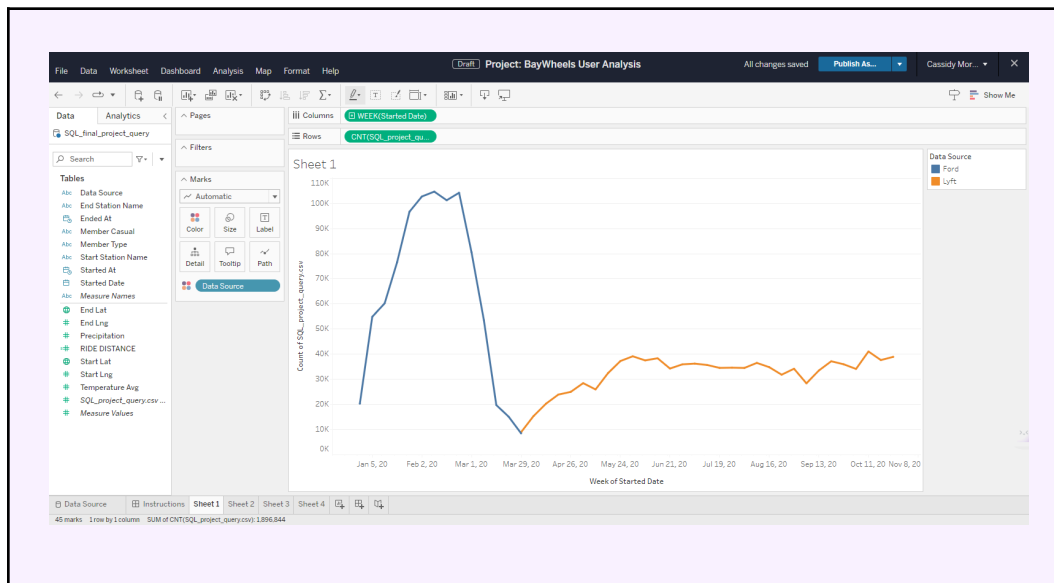
Investigating the differences between Lyft users and Ford users so that the marketing team at Lyft can make the best plan possible to help increase memberships in its rideshare program. The remaining Tasks will be completed in Tableau and will focus on visualizing and analyzing the results.

The Share Link is in the box below.

<https://prod-useast-b.online.tableau.com/t/globaltech/views/PortfolioProject1/Sheet3>

A. Plot the number of rentals made each week.

Using the visualization, when did operations transfer over from Ford to Lyft?
Are there any major differences in the volume of rentals before and after the transfer?

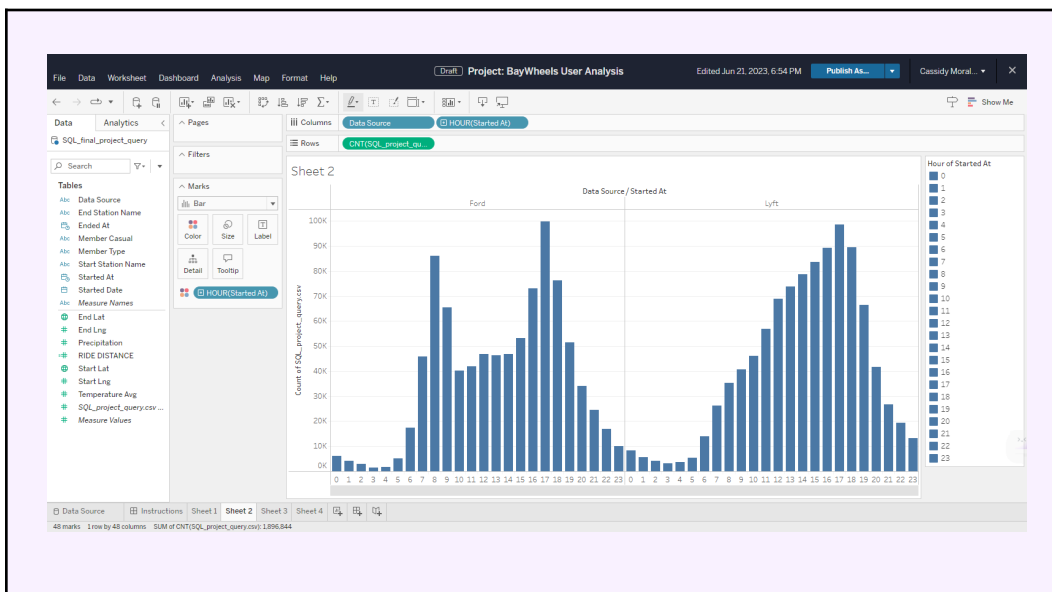


The switch was made from Ford to Lyft on week 14. There has been a steady positive increase since then but not as many as Ford had before the switch.

- B.** Create a bar chart to depict the total number of rides during each hour of the day. During which hours of the day are customers most likely to rent a bike?

They are most likely to rent a bike from 14:00–17:00 with 17:00 being the latest. That is 2pm–5pm.

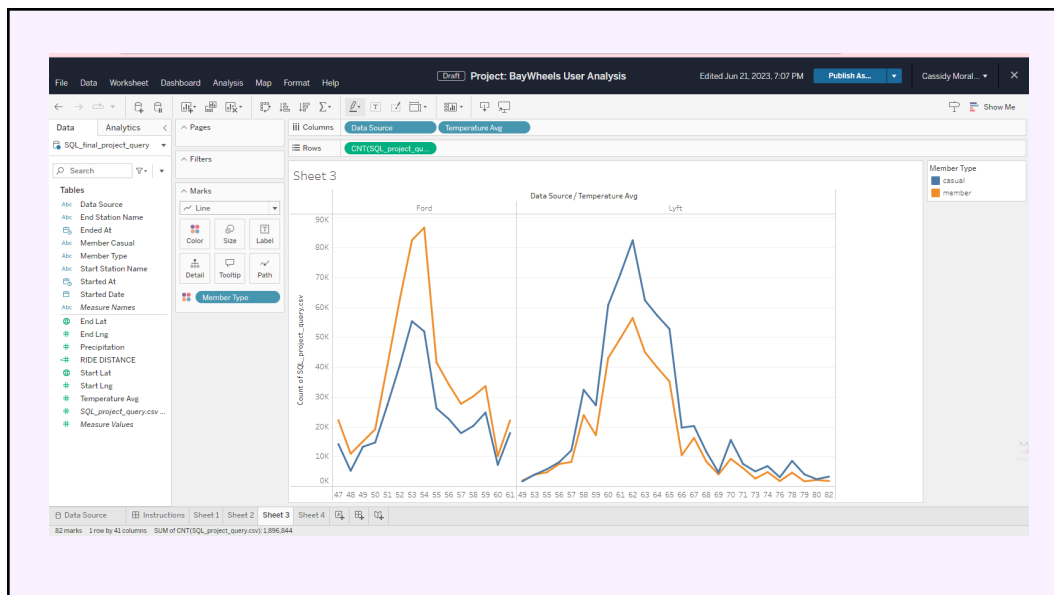
- C.** Create two side-by-side bar charts: one to illustrate the total number of rides during each hour of the data for Ford GoBike data, and the other for Lyft Baywheels. Regarding popular hours of the day, what are the differences between Lyft users and Ford users?



The data from Lyft has a fairly normal distribution with the peak hour being 17:00 (5pm). The riders are steadily using more and more rides per hour. However, for Ford it seems the rides peak at 8am and 5pm. Leading me to believe that working class use Ford more than Lyft.

- D. Create a line plot of the average temperature on the horizontal-axis and the number of rides taken on the vertical-axis. Plot one line for each Member Type. Finally, add **Data Source** to the column in order to compare Ford ridership with Lyft ridership.

How does the temperature affect ridership? Which riders are more willing to use a bike on cold days, and which riders are more likely to ride on warmer days?



Between Lyft and Ford. Ford users are more willing to ride on colder days. The Ford members are willing to ride much more than non members. Lyft user ride on warmer days with casual members riding more.


– Task 4: Communicating Results

Show the visualizations with a short paragraph explaining what insights can be drawn from it and any data-based marketing strategies to recommend to increase ridership at Lyft Baywheels.

- A. In a single paragraph, summarize what can be gleaned from your visualizations. In particular, are there differences between the datasets representing Ford and Lyft riders? How might Lyft market to customers in order to build upon the success of the Ford's GoBike program?

Based on the visualizations, it can be observed that Lyft and Ford riders exhibit distinct patterns and preferences. Lyft riders show a fairly normal distribution of rides throughout the day, with the peak hour being 5pm). The number of rides per hour steadily increases, suggesting a growing usage trend. In contrast, Ford riders have peak hours at 8am and 5pm, showing a preference among working class individuals who may rely on Ford for their commute daily. Additionally, Ford riders seem to be more inclined to ride on colder days compared to Lyft riders. Ford members exhibit a higher willingness to ride, always surpassing non-members in ride frequency. In contrast, Lyft users tend to ride more on warmer days, with casual members contributing to the influx in rides.

Considering these differences, Lyft could market to customers by leveraging the success of Ford's program. One approach could be to target the working-class by highlighting the convenience and reliability of Lyft rides during peak commuting hours. They could run a discount during these hours to help drive more traffic. Additionally, Lyft can promote the advantages of its membership program, encouraging users to become members and enjoy exclusive benefits. To capitalize on Lyft's appeal on



warmer days, the company could also consider running weather-specific promotions or incentives to encourage more rides during such conditions. By effectively targeting different customer segments and capitalizing on Lyft's membership program, the company can build upon the success of Ford's program and further expand its market reach.

That's it!