Portfolio Project | BayWheels User Analysis

INTRODUCTION: Here's what you need to know: Lyft purchased its bike share program from Ford (who owned GoBike) and needs a data analyst – that's you! – to help the marketing team use data-driven approaches in their new marketing efforts. You've been tasked by your manager to 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.

HOW IT WORKS: Follow the prompts in the questions below to investigate your data. Post your answers in the provided boxes: the **yellow boxes** for the queries you write, **purple boxes** for visualizations and **blue boxes** for text-based answers. When you're done, export your document as a pdf file and submit it on the Milestone page – see instructions for creating a PDF at the end of the Milestone.

PROMPT: Congratulations are in order! You've been hired as an intern by Lyft, one of the largest ride-sharing transportation providers in the country. In your new role, you'll be working on the Lyft BayWheels product: their latest initiative that provides rental bikes all across San Francisco through the Lyft app.

SQL App: Here's that link to our specialized SQL app, where you'll write your SQL queries and interact with the data.

— Data Set **Description**

To begin, you'll query a total of 3 datasets. You'll start with the `lyft.baywheels` and `ford.gobike` datasets available in your schema. Later, you will join the sf.weather dataset.

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

- **`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

These datasets are currently captured in your SQL database in separate tables, but your manager has told you that they are indeed the same data, just with different names.

Before you can start analyzing customer activity, you first need to combine the data needed from Ford and Lyft. While the datasets are currently captured in your SQL database in separate data tables, your manager has assured you that they are the same data, though with different variable names. Below is a table of equivalent columns between the two datasets, detailing which columns in the `lyft.baywheels` data set match which columns in the `ford.gobike` data table.

Lyft BayWheels	Ford GoBike
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. Write a query that filters the `ford.gobike` data to only include data from the year 2020.

```
SELECT
  *
FROM
  ford.gobike
WHERE
  DATE_PART('YEAR', start_time) = 2020
```

B. Write a query that unions the `ford.gobike` dataset and the `lyft.baywheels` dataset using the corresponding columns above. Make sure that you are still filtering to the year 2020 on the Ford data.

```
SELECT
  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_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_time) = 2020
```

After showing the result of the query to your manager, she tells you that she wants to know which data source is attributed to each row. She asks you to 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.

A colleague teaches you a simple method to do this. When writing your query, add an additional column after your select statement.

C. Modify your query from part B to include the `data_source` column.

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

Great! Since you and other members on your team will be referencing the output of your query for deeper analysis, your manager asked the Engineering team to 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

Now that we have combined and joined our three data tables together, you'll need to create additional variables so that you can perform the analysis your manager is asking from you.

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. You notice that 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.

Write a query that returns 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'**.

```
SELECT
  started_at,
  ended_at,
  start_station_name,
  end_station_name,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  CASE
  WHEN member_casual= 'Subscriber'
  THEN 'member'
  WHEN member_casual= 'Customer'
  THEN 'casual'
  ELSE member_casual
  END AS member_type,
  data_source
```

```
FROM
  project.ford_lyft_analysis
```

B. Almost there! After going over the table with your manager, she hypothesized that patterns are driven by changes in weather and wants you to incorporate weather data into your analysis.

You both decide San Francisco's average daily temperature and amount of precipitation are the best metrics to base your weather analysis on. These are located in the `temperature_avg` and `precipitation` columns, respectively, of the `sf.weather` table.

Modify your query from part B once more to join the table with the `sf_weather` data on the started_at field, truncated to the day level. From the sf_weather table, return the weather date, the average daily temperature, and the amount of precipitation.

```
SELECT
  started_at,
  ended_at,
  start_station_name,
  end_station_name,
  start_lat,
  start_lng,
  end_lat,
  end_lng,
  data_source,
  CASE
    WHEN member_casual = 'Subscriber' THEN 'member'
    WHEN member_casual = 'Customer' THEN 'casual'
    ELSE member_casual
  END AS member_type,
   precipitation,
```

```
  temperature_avg,
  date
FROM
  project.ford_lyft_analysis
INNER JOIN sf.weather
        ON sf.weather.date = date_trunc('day', started_at)
```

That's it! Now this query will result in almost 2 million records for the year 2020! Since SQLPad will only let you download 150,000 records in a .csv, the engineering team used some extra tools they have to download the result of your query. It's loaded for you in a Tableau Workbook, where you'll complete the rest of your project.

– **Task 3:** Visualizing and Analyzing Using Tableau

Phew! Now that you've gotten the query out of the way, you're ready to dive into 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 your results. **Click this link to navigate to the workbook you'll use to complete the remainder of this Project.**

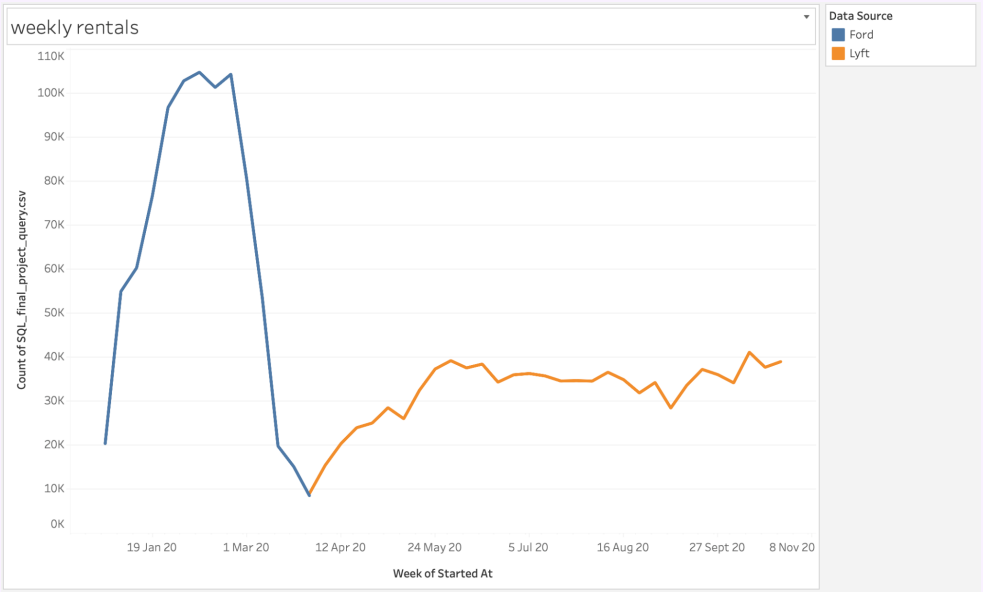
Once you've published your Tableau Workbook, paste the Share Link in the box below.

https://prod-useast-b.online.tableau.com/#/site/globaltech/collections/6a021436-5ac7-448d-a041-75792220ae29?:origin=card_share_link

Continue to post your answers in the provided boxes: purple boxes for your visualizations, and blue boxes for text-based answers.

A. On Sheet 1, start your exploration by plotting the number of rentals made each week. You should also add color to the chart so that you can clearly see when the Data Source changed over from Ford to Lyft.

Using your 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?

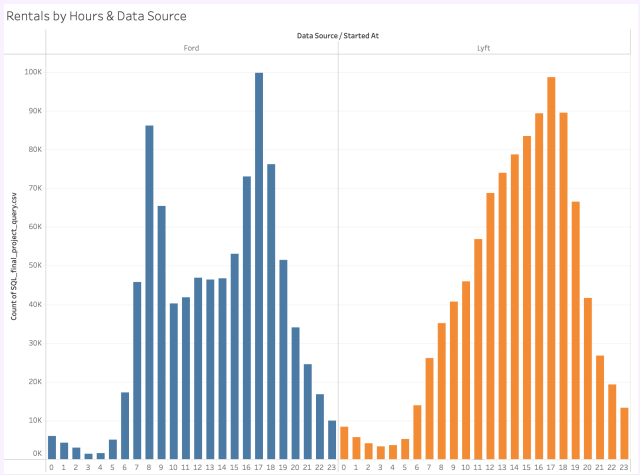


The transfer was made on March 29'th 2020. Ford had a lot more rentals in January and February (around 100k-110k weekly), but after that there was a fast decline from the end of February until the transfer was made. After the transfer there were less rentals but in a couple of months lyft stabilized and averaged 40k rentals.

B. Next, on Sheet 2, create a bar chart to depict the total number of rides during each hour of the day. No need to include this visualization in this report just yet! During which hours of the day are customers most likely to rent a bike?

> Between 7:00 and 21:00
> with best chances between 16:00 and 19:00

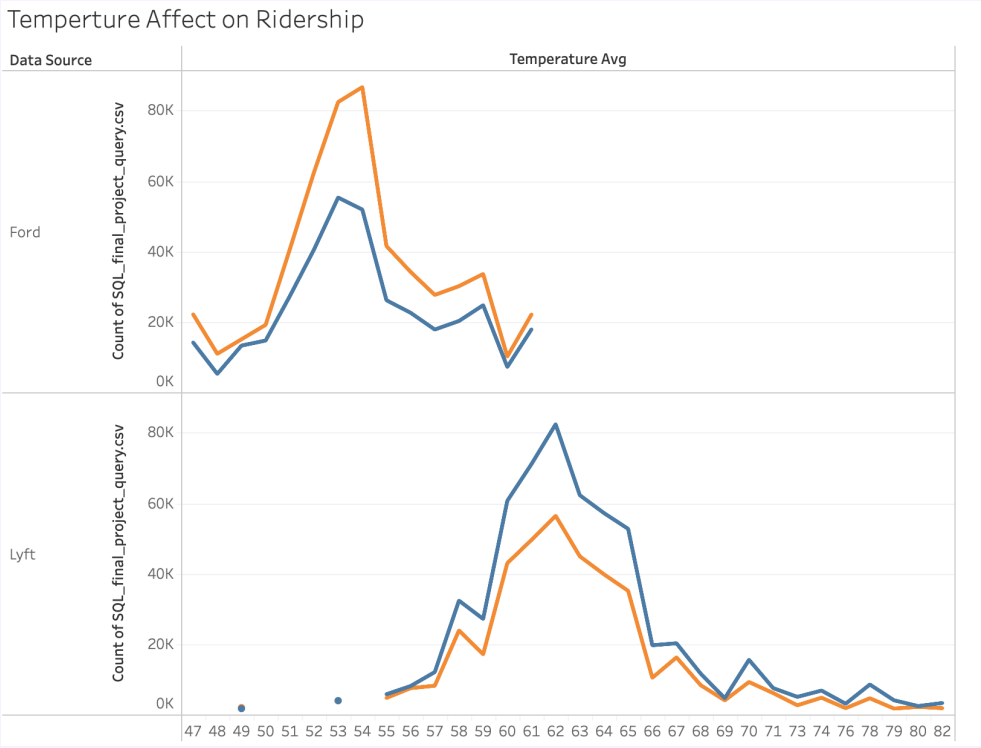
C. Let's break the hourly usage patterns down by data source. Modify your visualization from part B to create two side-by-side bar charts: one to illustrate the total number rides during each hour of the data for Ford GoBike data, and the other for Lyft Baywheels. Regarding popular hours of the day, what differences do you notice between Lyft users and Ford users?



> When Ford was leasing the bikes they had two peak hours: a morning rush at 8:00 and an evening rush at 17:00, while Lyft had escalating hours from morning until the rush hour at 17:00.

D. On Sheet 3, 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?



It seems Ford operated in 2020 mostly in the winter months, during their ownership in the very cold days members were more likely to rent bikes.

Lyft's period is the warmer season, but even on the colder days they shared with ford (55–61 degrees), and especially in the warmer days members were less likely to rent, and the top renters were casual and not members.

– **Task 4:** Communicating Results

Your manager wants you to share the visualizations you created in parts C and D of Task 4 with the marketing team for visibility. She asks you to email the visualizations to the team with a short paragraph explaining what insights can be drawn from it and any data-based marketing strategies you might recommend to increase ridership at Lyft Baywheels.

A. In a single paragraph, summarize what can be gleaned from your visualizations.

Looks like Ford had a lot more renters, so the potential for Lyft to grow is good. It seems like they have some work to do with the morning rush as it is not reaching its potential, they could create a campaign focusing on the benefits of renting the bikes to get to work and get ahead of the traffic. As members are the more loyal renters anc they seem to be disliking the warm season, i'd suggest some sort of perk for members in the warm season, maybe a sale on membership or discount for a member renting daily (rent all week get the weekend free or something of that sort).