Portfolio Project | Bay Wheels 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.

RESOURCES: If you need hints on the Milestone or are feeling stuck, there are multiple ways of getting help. Attend Drop-In Hours to work on these problems with your peers, or reach out to the HelpHub if you have questions. Good luck!

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 Bay Wheels 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 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

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 Bay Wheels	Ford GoBike
Started_date → date	Start_date → date
Started_at → time	Start_time → time
Ended_at → time	End_time → time
Start_station_name → text	Start_station_name → text
End_station_name → text	End_station_name → text
Start_lat → numeric	Start_station_latitude → numeric
Start_lng → numeric	Start_station_longitude → numeric
End_lat → numeric	End_station_latitude → numeric
End_lng → numeric	End_station_longitude → numeric
Member_casual → text	User_type → text

A. Write a query that filters the `ford.gobike` data to only include data from the year 2020. HINT: Use the `date_part` function in SQL!

```
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

--Only grab data from the year 2020.

WHERE
  date_part('year', start_date) = 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.

Note: You will want the Lyft data to be the first table in your query so that the column names from the Lyft dataset become the standard ones for the remainder of your analysis.

```
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
```

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. Here is an example of this for the Lyft table:

```
SELECT
  *,
  'Lyft' AS data_source
FROM lyft_baywheels
```

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

```
SELECT
  started_date,
  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_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,
```

```
  'Ford' AS data_source
FROM
  ford.gobike
WHERE
  date_part('year', start_date) = 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'**.

In other words, if member_casual is equal to 'Subscriber' your member_type field should be the string 'member' and if member_casual is equal to 'Customer', your member_type field should be the string 'casual'. Remember SQL is case sensitive!

```
SELECT
  DISTINCT member_casual,
  CASE
    WHEN member_casual = 'Subscriber' THEN 'member'
    WHEN member_casual = 'Customer' THEN 'casual'
    else member_casual
  END AS member_type
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_date` field, truncated to the day level. From the `sf_weather` table, return the average daily temperature, and the amount of precipitation.

```
SELECT
  b.temperature_avg,
  b.precipitation,
  CASE
    WHEN member_casual = 'Subscriber' THEN 'member'
    WHEN member_casual = 'Customer' THEN 'casual'
    else member_casual
  END AS member_type
FROM
  project.ford_lyft_analysis AS a
  INNER JOIN sf.weather AS b ON
date_trunc('day',a.started_date) = date_trunc('day',b.date)
```

```
--This query counts and double checks that it does return
--“almost 2 million records for the year 2020”.

SELECT
  COUNT(*) AS result_count
FROM
  (
    SELECT
      b.temperature_avg,
      b.precipitation,
      CASE
        WHEN member_casual = 'Subscriber' THEN 'member'
        WHEN member_casual = 'Customer' THEN 'casual'
        ELSE member_casual
      END AS member_type
    FROM
      project.ford_lyft_analysis AS a
      INNER JOIN sf.weather AS b ON date_trunc('day',
a.started_date) = date_trunc('day', b.date)
  ) AS subquery;
```

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.**

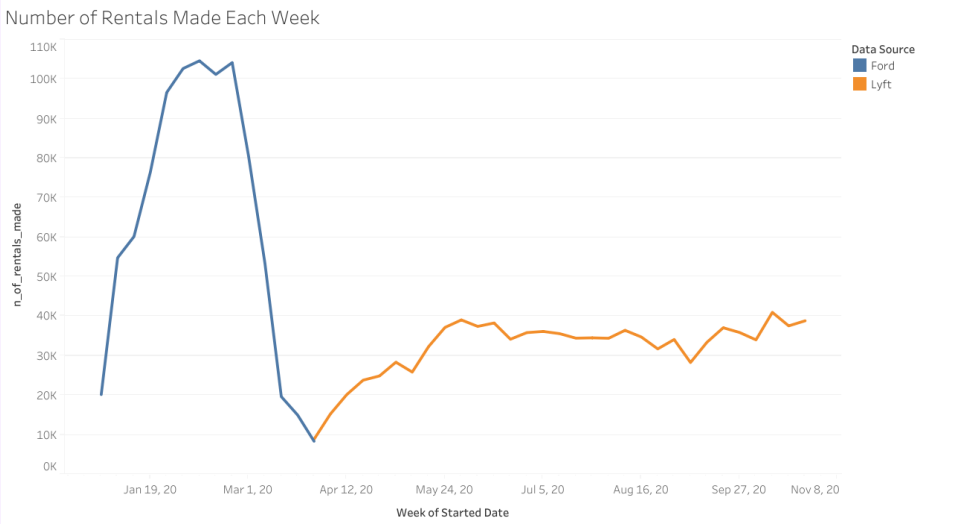
Once you've published your Tableau Workbook, paste the Share Link in the box below.

https://prod-useast-b.online.tableau.com/#/site/globaltech/workbooks/748254?: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. (Use the Started At field to determine each rental's 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?



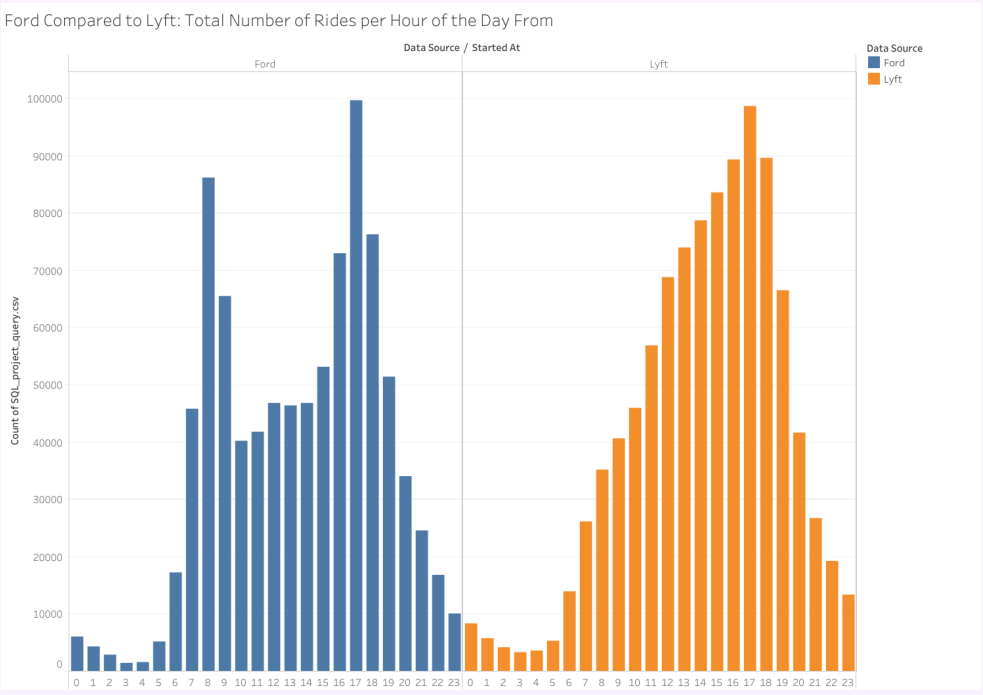
There is a major difference in the volume of rentals when the transfer occurred. This can be due to many different factors

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?

Ford data shows that 7:00AM, 8:00AM, and 9:00AM are the hours where customers are most likely to rent bikes in the morning hours, and 4:00PM, 5:00PM, and 6:00PM are the most likely evening hours for bike rentals.

Lyft data shows that 4:00PM, 5:00PM, and 6:00PM are the most popular and most likely hours for rental hours, although it's worth noting that there is a steady curve in the likelihood.

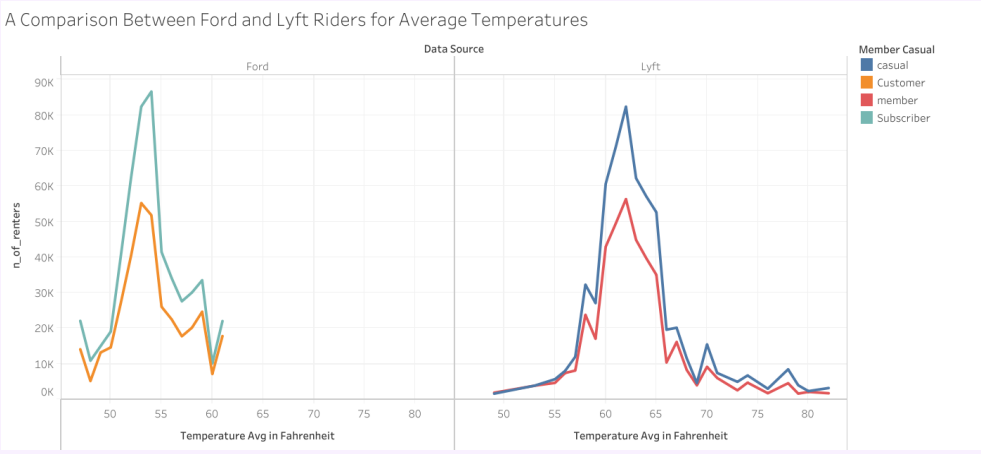
C. Let's break the hourly usage patterns down by data source. Using the **Data Source** field, modify your visualization from part B to create two side-by-side bar charts: one to illustrate the total number rides during each hour of the data for Ford GoBike data, and the other for Lyft Baywheels. Regarding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



Lyft Data shows that customers are renting bikes during the afternoon possibly for leisure time and exercise compared to Ford customers where there was a significant drop in customer use.

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. Note: you will have to convert the **Temperature Avg** feature into a Dimension first!

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?



Ford users are more likely to ride in the colder temperatures with Members being even more likely to ride as opposed to casual riders.

Lyft users are more likely to ride in the warmer temperatures with casual riders being even more likely to ride in warmer temperatures as opposed to 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. 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?

> The Lyft market can focus on promoting member deals for renters in colder climates. The Ford data shows that there is a stronger likelihood of this program succeeding. The main differences between datasets for Ford and Lyft riders show that Ford had a dramatic peak and yet no consistency, while Lyft had no dramatic peak and all of the consistency. It's worth noting that in April of 2020, was what I would determine as the official start of the pandemic and so this data is quite honestly obscured and an outlier for what most general years would look like. Thus, a limitation of data study is the data being limited to only the 2020 year.

That's it! Submit your final project for evaluation, and go celebrate your achievement! You just completed a rich, complex data analysis project representing real-world level work. You've gained some impressive skills! Well done, and never stop learning 😀

— LevelUp

The dataset in your Tableau workbook is rich – there's much more that can be done with the data! Below you'll find three additional LevelUp tasks. Have fun exploring them!

A. Your manager tells you that Lyft is interested in determining the distance riders travel between start and end points. Take a look in your Tableau notebook. You'll find a variable called RIDE DISTANCE that is the distance between the start and end points on a map.

   Note: this is not the same as the total distance traveled on the bike. For instance, if a ride began and ended at the same location, the distance would

show up as a zero in the data regardless of how long the bike was rented for. Instead, it lets Lyft know the typical distance riders travel when they start and end their rides at different points. The formula used is the Haversine distance. It calculates the distance between two GPS coordinates, taking Earth's curvature into consideration.

On Sheet 5, use this new calculated field to plot a histogram of the distance riders traveled. To make your visualization more useful, filter to values that are less than 7 miles and use a bin size of 0.1.

Analyze the histogram: how far do the majority of the rides typically go?

B. While you were assigned the analysis against temperature, one of your colleagues looked at the other weather feature you joined into the data: precipitation. She has interpreted the data to say that there's no major differences between Member Types in terms of ridership due to the weather.

   She's asked that you verify her work. Can you create a plot to illustrate how precipitation affects ridership? Compare between Ford and Lyft users and again between member and casual riders.

C. One of your colleagues has looked at the rentals by temperature plot you created and the rentals by precipitation plot your colleague created. With the approaching colder season in San Francisco, they're afraid of a dropoff in the amount of casual riders on the system and want to suggest additional marketing efforts to increase casual rider engagement over the next few months.

   How much do you agree with, or disagree with your colleague's assessment? Are there aspects of the data that they haven't considered in their analysis that can be addressed with other plots you created? Is there information outside of the available data that would be useful to make a better judgment of where to put the marketing focus for the next winter season?

— Submission

Great work completing your first Milestone! To submit your completed Milestone, you will need to download / export this document as a PDF and then upload it to the Milestone submission page. You can find the option to download as a PDF from the File menu in the upper-left corner of the Google Doc interface.