Ask

A clear statement of the business task.

**There are three questions to guide the future marketing program, of which the first I have to answer:**

1. **How do annual members and casual riders use Cyclistic bikes differently?**
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

**A description of all data sources used.**

The data is sources from [Index of bucket "divvy-tripdata"](https://divvy-tripdata.s3.amazonaws.com/index.html)

This data has been shared by by Lyft Bikes and Scooters, LLC (“Bikeshare”). They operate the City of Chicago’s (“City”) Divvy bicycle sharing service.

And that data has been downloaded and locally stored at C:\Users\Frank\Documents\Google Course Data\Capstone Project Cyclistic.

The data has been made available by Motivate International Inc.

**Documentation of any cleaning or manipulation of data.**

**Data exploration**

Checked for distinct member\_casual values:

* Verified that the column contains only the expected values (member and casual).

Checked for null values in started\_at column:

* Ensured that there are no missing values in the started\_at column.

Checked for null values in ended\_at column:

* Confirmed that there are no missing values in the ended\_at column.

Checked for missing values in start\_station\_name:

* Identified rows where the start\_station\_name is either null or empty.
* Found that 1,080,470 rows (18.3% of total) have missing values.

Checked for missing values in end\_station\_name:

* Identified rows where the end\_station\_name is either null or empty.
* Found that 1,108,680 rows (18.8% of total) have missing values.

Checked for missing values in member\_casual:

* Confirmed that there are no missing or empty values in the member\_casual column.

Checked for missing values in start\_lat, start\_lng, end\_lat, and end\_lng:

* Verified that there are no null values in the latitude and longitude columns for both start and end points.

Checked for duplicate values in ride\_id:

* Found 5,902,977 distinct ride\_id values compared to a total of 5,903,148 rows, indicating a difference of 171 duplicate entries.

**Data Cleaning**

Created a table for duplicate records:

* Identified and grouped ride\_id values with a count greater than 1, indicating duplicate records. The duplicates were saved into a new table Cyclistic.Duplicates.

Checked details of duplicate records:

* Performed a join to examine details of duplicate records by comparing the Duplicates table with the original Merge table.
* Found that the duplicates arise due to different time notations, e.g., 2024-05-31 23:56:47 UTC versus 2024-05-31 23:56:47.556000 UTC.

Created a deduplicated table:

* Used a ROW\_NUMBER() function to assign a unique row number to each ride\_id within duplicates.
* Retained only the first occurrence of each ride\_id, effectively removing duplicates.

Created a table without missing station names:

* Filtered rows to exclude records with null or empty values in the start\_station\_name and end\_station\_name columns.
* Saved the cleaned dataset into a new table Cyclistic.NoMissingStations.

Added a Travel\_time column and created a new table:

* Calculated travel time in minutes by subtracting started\_at from ended\_at and converting the result into minutes (rounded to two decimal places).
* Cast latitude and longitude columns as FLOAT64 to ensure consistent data types and stored this information in a new table Cyclistic.Cleaned\_Travel\_Time.

Checked Travel\_time values:

* Analyzed calculated travel times and identified 44 records with negative values (indicating ended\_at was earlier than started\_at).
* Decided to exclude these records to ensure data accuracy.

Created a table without negative travel times:

* Removed records with negative Travel\_time\_in\_Min values from the dataset.
* Saved the filtered dataset into a new table Cyclistic.Cleaned\_No\_Wrong\_Rents\_No\_Minus.

Removed wrongly rented bikes:

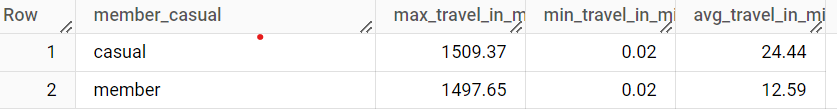
* Identified trips with travel times below 2 minutes where the start\_station\_name and end\_station\_name were the same.
* Saved these "wrong rents" into a new table Cyclistic.Cleaned\_Wrong\_Rents to address potential errors.

**Summary Of Analysis**

Calculated maximum and minimum travel time per user type:

Created a new table to calculate the maximum and minimum travel times (Travel\_time\_in\_Min) for each member\_casual type (member vs. casual).

Used the MAX() and MIN() aggregate functions to get the maximum and minimum travel times, grouped by user type.



Created a table displaying usage per member type per month:

* Created a new table showing the count of rides (count\_of\_rides) for each member\_casual (member or casual) by year and month.
* Extracted the year and month from the started\_at timestamp.
* Grouped the data by year, month, and member\_casual to track the number of rides for each member type over time.

Created a table displaying the most used beginning stations per membership type:

* Created a table to compare the most used starting stations for both member and casual riders.
* Used two subqueries (member\_usage and casual\_usage) to calculate the count of rides (rides\_per\_station) for both user types per starting station.
* Joined these two subqueries on start\_station\_name to get the number of rides for both member and casual users per station.
* Ordered the results by the number of rides for both member and casual users in descending order.

Calculated casual rides per station:

* Created a table to track the number of casual rides per starting station.
* Used a subquery to count the number of rides (rides\_per\_station) for casual riders (member\_casual = 'casual') per start\_station\_name.
* Ordered the results by the number of casual rides in descending order.

Filtered casual rides at the top 10 most popular stations for annual users:

* Queried the Calculations\_casual\_rides\_per\_station table to filter out casual rides at specific stations that are also popular for annual users (e.g., Kingsbury St & Kinzie St, Clinton St & Washington Blvd, etc.).
* Retrieved data for the casual rides at these stations based on the start\_station\_name condition.

**Supporting visualizations and key findings.**

**Your top three recommendations based on your analysis.**

Prepare

### **Guiding Questions**

* **Where is your**
* **data located?**
* **How is the data organized?**

Per month and also collected per quarter. Since the last 12 months h,ave been asked and the last quarterly report is from 2020, while the monthly reports are available till 2024-11 I will use the monthly reports.

* **Are there issues with bias or credibility in this data? Does your data ROCCC (Reliable, Original, Comprehensive, Current, Cited)?**

It is nowhere mentioned how the data has been gathered. However due to the nature of the data it seems likely that this data has been automatically generated by a system that tracks the incoming and outgoing bikes, time and date and the type of user etc..

The data is unchanged and from the source, therefore reliable and original. It seems like it tracked all the in and out going bikes. We can not know this for 100% sure, because there might have been a temporary or long term broken tracking system at one or more locations. However, this has not been reported. The data is current, since the most recent data is from the previous months from now, which is November 2024.

Made available by Motivate International Inc.

Sources:

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202411-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202410-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202409-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202408-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202407-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202406-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202405-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202404-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202403-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202402-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2024). [202401-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

Lyft Bikes and Scooters, LLC (“Bikeshare”). (2023). [202312-divvy-tripdata.zip. Divvy-tripdata]. Index of bucket "divvy-tripdata".

* **How are you addressing licensing, privacy, security, and accessibility?**

**License.** Bikeshare hereby grants to you a non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyze, copy, modify, distribute in your product or service and use the Data for any lawful purpose (“License”).

Prohibited Conduct. The License does not authorize you to do, and you will not do or assist others in doing, any of the following

* Use the Data in any unlawful manner or for any unlawful purpose;
* Host, stream, publish, distribute, sublicense, or sell the Data as a stand-alone dataset; provided, however, you may include the Data as source material, as applicable, in analyses, reports, or studies published or distributed for non-commercial purposes;
* Access the Data by means other than the interface Bikeshare provides or authorizes for that purpose;
* Circumvent any access restrictions relating to the Data;
* Use data mining or other extraction methods in connection with Bikeshare's website or the Data;
* Attempt to correlate the Data with names, addresses, or other information of customers or Members of Bikeshare; and
* State or imply that you are affiliated, approved, endorsed, or sponsored by Bikeshare.
* Use or authorize others to use, without the written permission of the applicable owners, the trademarks or trade names of Lyft Bikes and Scooters, LLC, the City of Chicago or any sponsor of the Divvy service. These marks include, but are not limited to DIVVY, and the DIVVY logo, which are owned by the City of Chicago.
* **How did you verify the data’s integrity?**

Data integrity refers to the accuracy, consistency, and reliability of data throughout its lifecycle, ensuring that it remains unaltered and trustworthy. All the files are similarly structured, so therefore consistent.

* **How does it help you answer your question?**

By finding out the different characteristics of the different user types, one can search for patterns.

Things I can check out per type of user

Rideable\_type

Started at/ended at. Date, popularity per day of the week, duration per ride, location by coordinates of popular stations, popular starting/ending stations.

* **Are there any problems with the data?**

Yes, some of the station\_ID’s are missing.

The bike\_ID format is also not consistent. Some of them are just numbers, others have letters and numbers. Doesn’t have to be a problem, but it is at least less structured than just having one type of format.

The databases also contain a lot of rows, so Google Sheets might not be sufficient, and I will have to use SQL. In fact I had to split some of them in two using R

**202011-divvy-tripdata.csv is too large to import into sheets.**

### **Key Tasks**

* **Download data and store it appropriately.**
* **Identify how it’s organized.**
* **Sort and filter the data.**
* **Determine the credibility of the data.**

### **Deliverable**

* **A description of all data sources used.**

Process

### **Guiding Questions**

* **What tools are you choosing and why?**

I will use SQL, because the files are too large for Google Sheets and Excel

The files are also too large to upload to SQL [Divvy\_Trips\_2019\_Q1 - Google Spreadsheets](https://docs.google.com/spreadsheets/d/1uCTsHlZLm4L7-ueaSLwDg0ut3BP_V4mKDo2IMpaXrk4/template/preview?pli=1&resourcekey=0-dQAUjAu2UUCsLEQQt20PDA#gid=1797029090). Therefore I split some files into two, so I could upload them.

* **Have you ensured your data’s integrity?**
* **What steps have you taken to ensure that your data is clean?**

**202409-divvy-tripdata-less-columns1**

All the ride\_id’s are unique, so there are no duplicates.

I checked for distinct values in Rideable\_typeand got the three expected results: electric\_bike, electric\_scooter and classic\_bike.

I checked whether any started\_at (=start date and time) had NULL or empty data, it didn’t.

I checked whether any ended\_at (=end date and time) had NULL or empty data, it didn’t.

I checked whether start\_station\_name had any empty data: 50767 records were empty of the total 410638 records, which is **12.36%.** There was no starting\_station added in this data. On an actual project in real life I would ask why this has happened.

Where the start\_station\_name is missing, also the start\_station\_id is missing.

39653/410638 = 9,65% end\_station\_names were empty.

However, whenever a start\_station\_name is empty I could use the start\_lat and start\_lng to identify the location and the team could later figure out which station it was.

225/410638 = 0.0547% of the end\_lat and end\_lng fields were ‘NA’, also the length of the coordinates differs, e.g. 42.07, 87.7 instead of the expected 6 digit 42.064854, -87.715297. Two digit coordinates still give about 1.1 km of accuracy, so this is probably enough to identify the station.

I checked for distinct member\_casual values, and received the expected ‘member’ and ‘casual’.

All the data types are correct, except the ‘end\_lat’ and ‘end\_lng’ column, which are STRINGs instead of FLOATs. This won’t have any effect on the analysis. But may affect the method to search for empty values.

The maximum travel times are all just below 1500 min. They all have in common that they have no end station. These are probably bikes which were picked up by the company because they weren’t returned by the driver. I will need to remove these rows, so keep the data accurate.

Of the rows which have no end\_station\_value, the highest 259 travel time all have a travel\_time\_in\_minutes in between 1499.62 and 1500. The starting and ending dates show no pattern. I probably have to remove them.

Also, I might have to remove all the columns without an end or starting station.

**255 rows have no end station AND now end coordinates, these records all have a travel time between 1499,62 and 1500.**

* **How can you verify that your data is clean and ready to analyze?**

**Check data types**

**Data range**

**Data constraints**

**Data consistency**

**Data structure**

**Code validation**

* **Have you documented your cleaning process so you can review and share those results?**

**Yes**

### 

### **Key Tasks**

* **Check the data for errors.**
* **Choose your tools.**
* **Transform the data so you can work with it effectively.**
* **Document the cleaning process.**

### **Deliverable**

* **Documentation of any cleaning or manipulation of data.**

Add column ride\_length

Average bike duration **this is still with the 1500 min. values!**

Casual: 21.76 min

Member: 12.04 min

**Add column day\_of\_week**

Analyze

### **Guiding Questions**

* **How should you organize your data to perform analysis on it?**
* **Has your data been properly formatted?**
* **What surprises did you discover in the data?**
* **What trends or relationships did you find in the data?**
* **How will these insights help answer your business questions?**

### **Key Tasks**

* **Aggregate your data so it’s useful and accessible.**
* **Organize and format your data.**
* **Perform calculations.**
* **Identify trends and relationships.**

### **Deliverable**

* **A summary of your analysis.**

**Type of bike per user!!**

**Popular stations per user type**

**Average travel length per user type**

**Possibly something with the coordinates of the stations.**

**Mayse use actual numbers instead of percentages so different databases can be aggregated.**

**Share**

**Act**

Variabelen waarop casual en annual kunnen verschillen:

Start/ end station

Tijd

Dag vd week

Rideable bike

Stakeholders: Marketing team

ROCC

Reliable

Original

Comprehensive

Current – van 2020-2021, dus wel een paar jaar oud

Cited

**Accuracy**: Ensuring that data correctly represents the real-world entity or event it is supposed to depict.

**Consistency**: Data must remain consistent across different databases, tables, or systems, even during updates or modifications.

**Reliability**: Data must remain intact and unchanged unless appropriately authorized, avoiding corruption or unintended alterations.

Legality

Lyft Bikes and Scooters, LLC (“Bikeshare”) operates the City of Chicago’s (“City”) Divvy bicycle sharing service. Bikeshare and the City are committed to supporting bicycling as an alternative transportation option. As part of that commitment, the City permits Bikeshare to make certain Divvy system data owned by the City (“Data”) available to the public, subject to the terms and conditions of this License Agreement (“Agreement”). By accessing or using any of the Data, you agree to all of the terms and conditions of this Agreement.

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?

2. Whywouldcasualriders buy Cyclistic annual memberships?

3. How can Cyclists use digital media to influence casual riders to become members?

You will produce a report with the following deliverables:

1. A clear statement of the business task

2. A description of al data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top three recommendations based on your analysis