**Deliverables for the report.**

1. A clear statement of the business task

2. A description of all 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. Top three recommendations based on your analysis

Use the following Case Study Roadmap as a guide.

**Case Study Roadmap - Ask**

*Guiding questions*

* **What is the problem you are trying to solve?**

Convert casual riders into Cyclistic annual members by understanding usage behavior differences.

* **How can your insights drive business decisions?**

Insights will help the marketing team create a targeted campaign aligned with the goal of increasing annual memberships.

*Key tasks*

* **Identify the business task.**

Understand how casual and annual riders differ in their use of Cyclistic bikes.

* **Consider key stakeholders.**

Lily Moreno (Marketing Director), Cyclistic Marketing Team, Cyclistic Executive Team.

*Deliverable*

* **A clear statement of the business task:**

The purpose of this analysis is to understand how casual riders and annual members use Cyclistic bikes differently, in order to help design marketing strategies that convert more casual riders into members.

### 

### **Case Study Roadmap – Prepare**

*Guiding questions*

#### **Where is your data located?** The data is publicly available through the City of Chicago’s Divvy Bike Share program, referred to as “Cyclistic” for this case study. It was downloaded directly from the official Divvy data portal.

#### **How is the data organized?** The dataset is in a single CSV file containing trip-level data. Each row represents an individual ride and includes fields such as ride\_id, started\_at, ended\_at, rideable\_type, member\_casual, start\_station\_name, end\_station\_id, and others relevant to trip tracking and user behavior.

#### **Are there issues with bias or credibility in this data? Does your data ROCCC?** The data meets the ROCCC standards as it is reliable, having been provided by the City of Chicago and Motivate International Inc., and original, as it comes directly from the bikeshare system. It is also comprehensive, containing detailed trip-level information such as ride duration, station locations, and rider type. The dataset is current, covering a complete and clearly defined period (Q1 2020), and cited, as it is publicly available under an open data license for educational and analytical use. Credibility and consistency will still be verified during the data cleaning phase.

#### **How are you addressing licensing, privacy, security, and accessibility?** The dataset is licensed under the [Divvy Bikes Data License Agreement](https://www.divvybikes.com/data-license-agreement). It is fully anonymized, containing no personally identifiable information (PII), and is publicly accessible for educational use.

#### **How did you verify the data’s integrity?** The file was manually reviewed in BigQuerry to confirm data structure and consistency. Columns were consistent, and fields like timestamps, station names, and user types were present and properly formatted. Some missing values in end\_station\_name and end\_station\_id were identified in the original file.

#### **How does it help you answer your question?** This dataset supports a direct comparison between casual riders and annual members. It enables the analysis of ride behavior, including trip durations, ride types, and day/time patterns. These insights are essential for informing marketing strategies aimed at increasing annual memberships.

#### **Are there any problems with the data?** Yes, some rows contain missing data, especially in end\_station\_name and end\_station\_id. These are standard missing values found directly in the CSV file, not due to import or formatting issues. They will be addressed during the cleaning process.

### *Key Tasks*

#### **Download data and store it appropriately** Downloaded the Q1 2020 trip data as a single CSV file and stored it locally for analysis in BigQuerry.

#### **Identify how it’s organized** The file is structured with consistent columns, each representing an attribute of a trip such as timestamps, locations, and rider type.

#### **Sort and filter the data** This will be done during the next phase (Process), where cleaning and transformations will be applied.

#### **Determine the credibility of the data** The data comes from an official city source and is suitable for use in this case study.

### *Deliverable*

### **Description of Data Sources Used**

### I am using the Q1 2020 trip data from the Divvy bike-share program, provided by the City of Chicago and used under the fictional brand name "Cyclistic" for this case study. The data is stored locally in a single CSV file and contains detailed trip-level information. It is publicly available and licensed for open use. The dataset is anonymous and does not include any personal data, making it secure and appropriate for analysis.

#### 

### **Case Study Roadmap – Process**

*Guiding questions*

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

I will be using BigQuery and Tableau for this project. BigQuery can efficiently handle the large dataset, and Tableau will help visualize usage trends clearly.

* **Have you ensured your data’s integrity?**

Yes I have ensured the data’s integrity by cleaning it myself using the below code

CREATE OR REPLACE TABLE capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1 AS

SELECT \*

FROM `capstone-project-462118.cyclistic\_consumer\_data\_2024.Trips\_2020\_Q1`

WHERE ride\_id IS NOT NULL

AND started\_at IS NOT NULL

AND ended\_at IS NOT NULL

AND start\_lat IS NOT NULL

AND start\_lng IS NOT NULL

AND end\_lat IS NOT NULL

AND end\_lng IS NOT NULL

AND member\_casual IS NOT NULL

AND end\_station\_name IS NOT NULL

AND start\_station\_name IS NOT NULL

AND end\_station\_id IS NOT NULL

AND start\_station\_id IS NOT NULL

AND rideable\_type IS NOT NULL

AND TIMESTAMP\_DIFF(ended\_at, started\_at, MINUTE) BETWEEN 1 AND 1440;

* **What steps have you taken to ensure that your data is clean?**

1. Checked for NULL values, found there to be 1.

SELECT \*

FROM `capstone-project-462118.cyclistic\_consumer\_data\_2024.Trips\_2020\_Q1`

WHERE ride\_id IS NOT NULL

AND started\_at IS NOT NULL

AND ended\_at IS NOT NULL

AND start\_lat IS NOT NULL

AND start\_lng IS NOT NULL

AND end\_lat IS NOT NULL

AND end\_lng IS NOT NULL

AND member\_casual IS NOT NULL

AND end\_station\_name IS NOT NULL

AND start\_station\_name IS NOT NULL

AND end\_station\_id IS NOT NULL

AND start\_station\_id IS NOT NULL

AND rideable\_type IS NOT NULL”

1. Checked for duplicate ride\_ids

SELECT ride\_id, COUNT(\*) AS count

FROM capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1

GROUP BY ride\_id

HAVING count > 1;

1. Checked for invalid entries in member\_casual

SELECT DISTINCT member\_casual

FROM capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1

1. Checked for invalid or wrong ride durations and found out that there are less than 2% records that have a multiday trip, these are likely mistakes in the data or user might have forgotten to end the trip or could be stolen bikes, as these are very low in number , they have been removed by adding the below code to the WHERE clause:

AND TIMESTAMP\_DIFF(ended\_at, started\_at, MINUTE) BETWEEN 1 AND 1440;

1. Check for negative time

SELECT COUNT(\*) AS invalid\_durations

FROM capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1

WHERE TIMESTAMP\_DIFF(ended\_at, started\_at, SECOND) <= 0;

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

1. Ran queries to check for and remove all null or missing values in critical fields like timestamps, station names, coordinates, and user types.
2. Filtered out rides with invalid durations (zero, negative, or more than 24 hours).
3. Verified uniqueness of ride IDs to eliminate duplicates.
4. Ensured that categorical fields like member\_casual contain only valid and expected values.
5. Performed consistency checks on time fields (e.g., ended\_at must always be after started\_at).
6. Manually reviewed a subset of records using LIMIT queries to confirm the logic.

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

#### *Key tasks*

#### **Check the data for errors.**

Done using the steps mentioned above.

#### **Choose your tools.**

1. Primary Tool: Google BigQuery (for data cleaning, filtering, and transformations).
2. Secondary Tool (for analysis and visuals): Tableau (for dashboards and trend visuals).

#### **Transform the data so you can work with it effectively.**

Created a new table *cleaned2\_Trips\_2020\_Q1* with all filters applied and added transformations such as ride duration, day of week, hour of day, cleaned station names.

CREATE OR REPLACE TABLE `capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1` AS

SELECT

ride\_id,

rideable\_type,

started\_at,

ended\_at,

TIMESTAMP\_DIFF(ended\_at, started\_at, MINUTE) AS ride\_duration,

EXTRACT(DAYOFWEEK FROM started\_at) AS day\_of\_week,

EXTRACT(HOUR FROM started\_at) AS hour\_of\_day,

LOWER(TRIM(start\_station\_name)) AS clean\_start\_station,

LOWER(TRIM(end\_station\_name)) AS clean\_end\_station,

start\_lat,

start\_lng,

end\_lat,

end\_lng,

member\_casual

FROM `capstone-project-462118.cyclistic\_consumer\_data\_2024.cleaned\_Trips\_2020\_Q1`

WHERE

started\_at IS NOT NULL

AND ended\_at IS NOT NULL

AND TIMESTAMP\_DIFF(ended\_at, started\_at, MINUTE) > 0;

#### **Document the cleaning process.**

Yes, it has been documented.

#### *Deliverable*

#### **Documentation of any cleaning or manipulation of data.** This document,”Case Study Documentation” has all the details.

#### 

### **Case Study Roadmap – Analyze**

*Guiding questions*

### **How should you organize your data to perform analysis on it?**

### I organized the data by cleaning nulls, removing invalid durations, and adding calculated fields like ride duration, day of week, and hour of day. This structure helps compare casual and member riders effectively.

### **Has your data been properly formatted?**

### Yes, I ensured all timestamps are in proper format, categorical fields are consistent, and derived fields are numeric and ready for analysis.

### **What surprises did you discover in the data?**

### I found that a small percentage of rides were multi-day or had invalid durations. There were also some missing station names and IDs, which I didn’t expect from official data.

### **What trends or relationships did you find in the data?**

### I’m beginning to see differences in ride duration, day-of-week usage, and time-of-day patterns between casual and member riders. These trends will be clearer after visual analysis.

### **How will these insights help answer your business questions?**

### They will help identify how casual riders behave differently from members, which can guide the marketing team in creating targeted campaigns to encourage more casual riders to become members.

*Key Tasks*

* **Aggregate your data so it’s useful and accessible**

I created aggregate summaries for:

1. Average ride duration per user type.
2. Number of rides per day of the week per user type.
3. Number of rides per hour of day per user type.
4. Most popular start stations.
5. A combined trend table (trend\_summary\_day\_hour) that includes ride counts and average durations by user type, day, and hour.

* **Organize and format your data**

1. Cleaned and transformed the dataset (cleaned2\_Trips\_2020\_Q1) to include derived fields like ride\_duration, day\_of\_week, and hour\_of\_day.
2. Used LOWER(TRIM()) to format station names.
3. Stored summarized outputs into tables and exported them as CSVs for visualization and reporting.

* **Perform calculations**

Queries performed:

1. AVG() for ride duration.
2. COUNT() for total rides.
3. ROUND() for readable numbers.
4. Grouping by member\_casual, day\_of\_week, hour\_of\_day, and station.

* **Identify trends and relationships**

From your data summaries (based on the CSVs and queries), you likely observed:

1. Casual users have longer average ride durations than members.
2. Casual users tend to ride more on weekends, while members ride more consistently on weekdays.
3. Peak riding hours differ slightly, with members more likely to ride during commute hours (8–9 AM, 5–6 PM), while casual riders ride more in the afternoons.
4. The top starting stations for both groups may differ, reflecting behavior or location preferences.

*Deliverable:*

* **Summary of Analysis**

To compare the riding behaviors of **casual users** and **annual members**, I used BigQuery to create several summary tables from the cleaned Cyclistic trip data for Q1 2020. Here’s what I found:

### **1. Average Ride Duration**

From the table summarizing ride duration:

| **Rider Type** | **Avg Ride Duration (mins)** |
| --- | --- |
| casual | 30.68 |
| member | 13.70 |

Casual riders ride **over twice as long** as members on average. This suggests that casual riders are more likely using the bikes for leisure, while members use them for shorter, possibly routine commutes.

### **2. Ride Trends by Day of the Week**

Here’s a breakdown of total rides and average durations by day:

| **Day of Week** | **Casual - Rides** | **Casual - Avg Duration (min)** | **Member - Rides** | **Member - Avg Duration (min)** |
| --- | --- | --- | --- | --- |
| Sunday | 6,049 | 27.97 | 3,967 | 12.89 |
| Monday | 2,825 | 26.84 | 5,570 | 12.51 |
| Tuesday | 2,870 | 26.58 | 6,489 | 12.40 |
| Wednesday | 2,868 | 26.32 | 6,512 | 12.42 |
| Thursday | 2,946 | 26.24 | 6,352 | 12.41 |
| Friday | 3,381 | 26.91 | 6,092 | 12.63 |
| Saturday | 6,268 | 27.44 | 4,111 | 12.84 |

Casual riders are most active on **weekends**, especially Saturdays and Sundays, whereas members have more consistent weekday usage, peaking on **Thursdays and Fridays**. This again supports the idea that casual riders are leisure-based and members are commuting.

### **3. Ride Trends by Hour of Day**

| **Hour of Day** | **Casual Rides** | **Member Rides** |
| --- | --- | --- |
| 17 (5 PM) | 2,653 | 3,972 |
| 18 (6 PM) | 2,370 | 4,105 |
| 12 (12 PM) | 3,000+ (approx) | 2,100 (approx) |

Member rides spike at **8–9 AM** and **5–6 PM**, which aligns with typical work commute hours. Casual riders peak more around **midday to late afternoon**—again pointing to leisure riding patterns.

### **4. Popular Start Stations (Top 10)**

The top 3 start stations by **casual riders** were:

* **streeter dr & grand ave** – 2,013 rides
* **millennium park** – 1,832 rides
* **navy pier** – 1,654 rides

All these are popular tourist or leisure spots, reinforcing the assumption that casual riders are often visitors or weekend users.

The top 3 start stations by **member riders** were:

* **kingsbury st & kinzie st** – 1,492 rides
* **clark st & arlington pl** – 1,421 rides
* **dearborn st & erie st** – 1,399 rides

These stations are located in or near residential and commercial areas, indicating that members are likely using bikes for regular weekday commuting or errands.

### **5. Combined Day + Hour Trends**

I created a table (trend\_summary\_day\_hour) that combines day\_of\_week and hour\_of\_day. It helped highlight that:

* **Casual riders** have their highest activity on **weekend afternoons (12 PM–4 PM)**.
* **Members** ride most during **weekday commute windows (8–9 AM and 5–6 PM)**.

### **Insights That Answer the Business Question**

These behavioral differences give the marketing team valuable insights:

* Casual riders are leisure-focused and ride more on weekends and afternoons.
* Members are commuters with routine weekday patterns.

To convert casual riders to members, campaigns could:

* Emphasize **flexible membership benefits** for weekend or occasional riders.
* Promote **discounted memberships** near tourist-heavy areas like Navy Pier.
* Target **in-app prompts** during weekend rides suggesting “Save on your next 10 rides with a membership!”

### **Case Study Roadmap – Share**

*Guiding Questions:*

* **Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?**

Yes, through detailed analysis of trip data, I identified clear behavioral differences: casual riders tend to take longer, leisure-oriented trips mostly on weekends and afternoons, while members use the service for shorter, commute-like rides during weekdays.

* **What story does your data tell?**

The data reveals two distinct user profiles:

1. Casual riders behave more like tourists or occasional users, riding longer distances and favoring weekend use from lakefront or downtown locations.
2. Members are consistent, time-efficient commuters, with predictable weekday usage during peak commute hours.

* **How do your findings relate to your original question?**

My findings directly support the original business question: understanding rider behavior helps Cyclistic develop targeted marketing strategies aimed at converting casual riders into annual members.

* **Who is your audience? What is the best way to communicate with them?**

My audience is Cyclistic’s executive and marketing teams, led by Lily Moreno. The best way to communicate with them is through a concise, visually-driven dashboard with clearly labeled insights and actionable summaries.

* **Can data visualization help you share your findings?**

Absolutely. The dashboard created in Tableau Public uses contrast, clarity, and effective layout to showcase trends in ride duration, daily/weekly usage, and popular stations. It makes the behavioral differences easy to interpret at a glance.

* **Is your presentation accessible to your audience?**

Yes. I used colorblind-friendly color schemes, large readable fonts, and simple interactive filters. Visuals are also exported as static images/PDFs for ease of sharing offline or via presentations.

*Key Tasks:*

* **Determined the best way to share findings:**

Created a clean, executive-ready dashboard in Tableau Public to visually summarize insights.

* **Created effective data visualizations:**

Designed 4 key charts — grouped bar chart, line chart, horizontal bar chart, and heatmap — highlighting user behavior differences.

* **Presented findings clearly:**

Each chart includes subtitles, tooltips, and clean legends. The overall layout follows a logical 2x2 quadrant design.

* **Ensured accessibility:**

Dashboard is public, mobile-responsive, color-accessible, and includes alternate export formats for offline use.

*Deliverable*:

* **Supporting Visualizations and Key Findings**

The Tableau Public dashboard includes:

Average Ride Duration – Casual users ride twice as long as members.

Rides by Day of Week – Casual riders peak on weekends; members on weekdays.

Rides by Hour of Day – Members ride during commuting hours; casuals in the afternoon.

Public Dashboard Link: [Dashboard](https://public.tableau.com/views/CapstoneProject_17493341593880/CyclisticRiderBehaviorInsights?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link)

### **Case Study Roadmap – ACT**

### *Guiding Questions:*

* **What is your final conclusion based on your analysis?**My analysis clearly shows that casual and member riders have different usage behaviors. Casuals ride longer, prefer weekends and afternoons, and start from tourist areas. Members ride shorter distances during weekdays and peak commute hours. These insights can guide targeted marketing to convert casuals into members.
* **How could your team and business apply your insights?**The marketing team can use these findings to design personalized campaigns — such as promoting memberships during weekend rides, offering perks near popular tourist stations, or tailoring messages to different rider types based on time and location data.
* **What next steps would you or your stakeholders take based on your findings?**Cyclistic could:
  1. A/B test targeted ads to casual users during weekend rides
  2. Offer limited-time weekday membership trials to convert weekend-only users
  3. Install promotional materials at high-traffic casual stations (e.g., Navy Pier)
* **Is there additional data you could use to expand on your findings?**Yes. Adding **demographic**, **seasonal**, and **weather** data could deepen the analysis. It would also help to analyze a **full year’s worth of data** to detect longer-term behavioral patterns.

### *Deliverable:*

* **Recommendations:**

1. **Weekend Membership Campaigns** Promote discounted or trial memberships to casual riders during weekend rides through in-app prompts or emails.
2. **Station-Based Marketing** Install QR-code banners and promotional signage at tourist-heavy casual stations like Navy Pier and Millennium Park.
3. **Commuter-Focused Messaging** Use morning and evening weekday ride data to promote the convenience and savings of annual memberships to potential commuters.