Case Study

**How Does a Bike-Share Navigate Speedy Success?**

**Scenario:**

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company’s future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Lily Moreno, the Director of the Marketing team, has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

**Deliverables:**

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. Your top three recommendations based on your analysis

**Steps in the Analysis:**

***1. Ask***

**Guiding questions**

* What is the problem you are trying to solve?

We are trying to solve the problem statement: How do annual members and casual riders use Cyclistic bikes differently?

* How can your insights drive business decisions?

**Key tasks**

* Identify the business task
* Consider key stakeholders

**Deliverable -** A clear statement of the business task

***2. Prepare***

**Guiding questions**

* Where is your data located?

The data is located [here](https://divvy-tripdata.s3.amazonaws.com/index.html), where I have taken the data of 9 months i.e. files from Sept 2020 (202009) to May 2021 (202105).

* How is the data organized?

The data is organised into monthly bike trips in a csv format.

* Are there issues with bias or credibility in this data? Does your data ROCCC?

The data source is original and also updated till the date of starting this project. It includes bike trips of every single day of last 12 months so its 100% credible and reliable. Also the data doesn’t reveals any personal information of the riders.

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

The data is in public domain and has been officially made available by Motivate International Inc. under this [license](https://www.divvybikes.com/data-license-agreement).

* How did you verify the data’s integrity?

After performing the integrity checks on the cleaned data, I found that that there ae no duplicates and null values in the data which I am going to analyse.

**Key tasks**

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

**Deliverable -** A description of all data sources used

***3. Process***

**Guiding questions**

* What tools are you choosing and why?

I have used SQL standard dialect on BigQuery to clean and analyse the data, and Tableau to create visualizations from the analysed data.

* Have you ensured your data’s integrity?

The data’s integrity has been ensured by performing data integrity checks and cleaning the data.

* What steps have you taken to ensure that your data is clean? Have you documented your cleaning process so you can review and share those results?

Following is the documentation of my cleaning process:

Since the data is broken into multiple files by months, we first integrated all these files into one single table using the UNION ALL command to clean and analyze it efficiently. Also we added two columns: ride\_length and day\_of\_week, as we will require them both for the analyze process.

CREATE TABLE

Cyclistic\_bike\_share.tripdata\_union AS

SELECT DATETIME\_DIFF(ended\_at, started\_at, SECOND) AS `ride\_length`,

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

       \*

FROM (

SELECT 'sept\_2020' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.sept\_2020` UNION ALL

SELECT 'oct\_2020' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.oct\_2020` UNION ALL

SELECT 'nov\_2020' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.nov\_2020` UNION ALL

SELECT 'dec\_2020' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.dec\_2020` UNION ALL

SELECT 'jan\_2021' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.jan\_2021` UNION ALL

SELECT 'feb\_2021' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.feb\_2021` UNION ALL

SELECT 'mar\_2021' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.mar\_2021` UNION ALL

SELECT 'apr\_2021' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.apr\_2021` UNION ALL

SELECT 'may\_2021' AS filename, \* FROM `sage-artifact-322407.Cyclistic\_bike\_share.may\_2021` )

On successful merging the data, we obtained some 25,56,715 rows of raw data which needs to be processed and analyzed in order to create visualizations to draw compelling stories.

We obtained the following schema:

|  |  |
| --- | --- |
| Field name | Type |
| ride\_id | STRING |
| round\_trip | INTEGER |
| rideable\_type | STRING |
| start\_station\_id | STRING |
| start\_station\_name | STRING |
| end\_station\_id | STRING |
| end\_station\_name | STRING |
| member\_casual | STRING |
| ride\_length | INTEGER |
| day\_of\_week | INTEGER |
| filename | STRING |
| started\_at | TIMESTAMP |
| ended\_at | TIMESTAMP |
| start\_lat | FLOAT |
| start\_lng | FLOAT |
| end\_lat | FLOAT |
| end\_lng | FLOAT |

Now, we will clean the data using some SQL functions such as DISTINCT(), LENGTH(), TRIM(), CAST(), etc. and keywords such as NULL, LIKE, etc. in order to filter the invalid data. It is also mentioned here that the staff also uses the bikes for a time duration of less than a minute for testing purposes. These tests start at or end at “Base - 2132 W Hubbard Warehouse”. Similarly, if a user won’t return the bike within a day (24 hrs), then the company deems this case as stolen or lost bikes. So we will filter the following data:

* Unique ride\_id whose character length is not equal to 16
* start\_station\_name, start\_station\_id, end\_station\_name , end\_station\_id, start\_lat , end\_lat , start\_lng , end\_lng whose values are NULL.
* Any column data with a blank space present
* ride\_length less than 60 secs and greater then (60\*60\*24) secs
* start\_station\_name and end\_station\_name with ‘TEST’ and ‘Base’ as a value.(since I assume from my data scanning that ‘Base’ and ‘TEST’ won’t be the name of any station in the dataset)

We will add a column of round\_trip to know how many round trips have been completed by the users and we will also extract the DATE from the DATETIME of started\_at column.

CREATE TABLE

Cyclistic\_bike\_share.tripdata\_cleaned AS

SELECT

    DISTINCT TRIM(ride\_id) AS ride\_id,

    CASE WHEN UPPER(start\_station\_name) = UPPER(end\_station\_name) THEN 1 ELSE 0 END AS round\_trip,

    TRIM(rideable\_type) AS rideable\_type,

    TRIM(start\_station\_name) AS start\_station\_name,

    TRIM(end\_station\_name) AS end\_station\_name,

    TRIM(member\_casual) AS member\_casual,

    CAST(started\_at AS DATE) AS started\_date,

    \* EXCEPT (ride\_id, rideable\_type, start\_station\_name, end\_station\_name, member\_casual)

FROM `sage-artifact-322407.Cyclistic\_bike\_share.tripdata\_union`

WHERE

    LENGTH(ride\_id) = 16

    AND ride\_length BETWEEN 60 AND (60\*60\*24)

    AND started\_at IS NOT NULL

    AND ended\_at IS NOT NULL

    AND start\_station\_name IS NOT NULL

    AND end\_station\_name IS NOT NULL

AND start\_station\_id IS NOT NULL

    AND end\_station\_id IS NOT NULL

    AND start\_lat IS NOT NULL

    AND end\_lat IS NOT NULL

    AND start\_lat IS NOT NULL

    AND end\_lng IS NOT NULL

    AND UPPER(start\_station\_name) NOT LIKE '%BASE%'

    AND UPPER(end\_station\_name) NOT LIKE '%BASE%'

    AND UPPER(start\_station\_name) NOT LIKE '%TEST%'

    AND UPPER(end\_station\_name) NOT LIKE '%TEST%'

On successfully executing this, we got the following results:

|  |  |  |
| --- | --- | --- |
| Raw records | Cleaned records | Invalid/duplicate records |
| 25,56,715 | 22,22,490 | 3,34,225 |

**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

***4. Analyze***

**Guiding questions**

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

Organising the data included removing unwanted columns, correcting data types and sorting the rides according to the time. After that, this data is analysed and aggregated in the following way:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ridership | mean\_ride\_length\_mins | total\_trips | total\_round\_trips | total\_days |
| Casual Riders | 31.4 | 860738 | 118606 | 273 (9 months) |
| Member Riders | 14.1 | 1361752 | 49763 | 273 (9 months) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| months | ridership | total\_trips | total\_round\_trips | mean\_ride\_length\_mins | total\_days |
| Sep-2020 | casual | 212053 | 29974 | 32.6 | 30 |
| Sep-2020 | member | 279033 | 10988 | 15 | 30 |
| Oct-2020 | casual | 120625 | 14353 | 28.2 | 31 |
| Oct-2020 | member | 211758 | 7452 | 13.8 | 31 |
| Nov-2020 | casual | 72112 | 9876 | 29.4 | 30 |
| Nov-2020 | member | 147136 | 5642 | 13.5 | 30 |
| Dec-2020 | casual | 24282 | 2898 | 23.9 | 31 |
| Dec-2020 | member | 88022 | 3181 | 12.5 | 31 |
| Jan-2021 | casual | 14567 | 1600 | 22.4 | 31 |
| Jan-2021 | member | 68001 | 2193 | 12.1 | 31 |
| Feb-2021 | casual | 8484 | 1078 | 31.4 | 28 |
| Feb-2021 | member | 33779 | 1376 | 14.9 | 28 |
| Mar-2021 | casual | 74952 | 11559 | 32.4 | 31 |
| Mar-2021 | member | 128309 | 4663 | 13.8 | 31 |
| Apr-2021 | casual | 119253 | 18188 | 32.4 | 30 |
| Apr-2021 | member | 175238 | 6503 | 14.4 | 30 |
| May-2021 | casual | 214410 | 29080 | 33.4 | 31 |
| May-2021 | member | 230476 | 7765 | 14.5 | 31 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ridership | rideable\_type | mean\_ride\_length\_mins | total\_trips | total\_round\_trips | total\_days |
| casual | docked\_bike | 38.8 | 383460 | 67380 | 273 |
| casual | classic\_bike | 29.1 | 261699 | 31648 | 181 |
| casual | electric\_bike | 21.2 | 215579 | 19578 | 273 |
| member | docked\_bike | 14.7 | 489979 | 19526 | 96 |
| member | classic\_bike | 14.2 | 568551 | 20996 | 181 |
| member | electric\_bike | 12.9 | 303222 | 9241 | 273 |
|  |  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ridership | day\_of\_week | mean\_ride\_length\_mins | total\_trips | total\_round\_trips |
| casual | 1 | 36.2 | 172097 | 27999 |
| casual | 2 | 31.9 | 100512 | 15637 |
| casual | 3 | 28.3 | 85928 | 12133 |
| casual | 4 | 26.9 | 90291 | 11583 |
| casual | 5 | 25.5 | 86160 | 10433 |
| casual | 6 | 28.8 | 118041 | 13664 |
| casual | 7 | 34.5 | 207709 | 27157 |
| member | 1 | 16 | 170905 | 7592 |
| member | 2 | 13.6 | 187788 | 7025 |
| member | 3 | 13.3 | 196209 | 6847 |
| member | 4 | 13.3 | 208701 | 7243 |
| member | 5 | 13.1 | 194891 | 6590 |
| member | 6 | 13.8 | 201549 | 6931 |
| member | 7 | 15.7 | 201709 | 7535 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ridership | type\_of\_day | mean\_ride\_length\_mins | total\_trips | total\_round\_trips |
| casual | Weekday | 28.28 | 480932 | 63450 |
| casual | Weekend | 35.35 | 379806 | 55156 |
| member | Weekday | 13.42 | 989138 | 34636 |
| member | Weekend | 15.85 | 372614 | 15127 |

**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

***5. Share***

The data visualization and findings from it can be [seen here](https://public.tableau.com/views/GoogleDACapstone/Story1?:language=en-US&:display_count=n&:origin=viz_share_link).

**Key tasks**

* Determine the best way to share your findings.
* Create effective data visualizations.
* Present your findings.
* Ensure your work is accessible.

**Deliverable -** Supporting visualizations and key findings

***6. Act***

**Guiding questions**

Based on my findings, following are my top three recommendations:

1. Since most of the casuals love riding more on weekends, a weekend membership plan can be introduced in order to convert them to members. Also, those casuals who ride regularly throughout the week must be given trail membership in order to make them acquainted with annual membership plan.
2. Casuals make rides at their maximum in summers, so this might be the best time to offer them the above mentioned membership plans. This is perhaps the best time to pull larger number of causal users to member ones.
3. Identifying the areas where casuals work and increasing the number of stations in those areas may help turning them into members. This can be achieved by extracting their work info through feedback.

**Key tasks**

* Create your portfolio.
* Add your case study.
* Practice presenting your case study to a friend or family member.

**Deliverable -** Your top three recommendations based on your analysis