CASE STUDY 1

**Ask**:

Business Task:

The stakeholders' main goal is to maximize the number of annual members + design appropriate marketing strategies aimed at making and converting casual riders into annual members.

Our business task is to identify the difference between casual and annual member riders, like how they use the company’s bikes differently, etc. To do that we use the database to accomplish analysis to give some insights to our stakeholders.

**Prepare**:

Data is retrieved from Divvy Bakes datasets, which is public data licensed by Lift Inc. Data is located at <https://divvy-tripdata.s3.amazonaws.com/index.html>. Data is sorted by the date and it’s monthly based. I retrieved datasets starting from November 2023 and ending with April 2024. They all are in Excel format and by checking data it meets our business task and answers the questions asked by our stakeholders. It can help to find the differences between riders. The data also meets the basic standard, and the data’s integrity can be verified.

NOTE: I chose only 6-month data because I don’t have the opportunity to use full-version of my data analytical tools, like Bigquery (SQL), Tableau, and R. So for a meaningful and complete analysis, I choose data that can be edited and imported in my available tools.

**Process:**

I choose SQL because it’s the most convenient and quickest tool for a process and analyzing huge amounts of data. I use some SQL functions to clean and prepare the data for more usable and easily accessible analysis. But from the start, I unite and combine all data. I use UNION ALL and CREATE functions to combine all data and create a new table.

CREATE OR REPLACE TABLE `eminent-card-429217-h8.trip\_data.all\_trip` AS

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_nov\_2023`

UNION ALL

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_dec\_2023`

UNION ALL

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_jan\_2024`

UNION ALL

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_feb\_2024`

UNION ALL

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_mar\_2024`

UNION ALL

SELECT \*

FROM `eminent-card-429217-h8.trip\_data.trip\_data\_apr\_2024`

Then I use functions like DISTINCT and WHERE clauses to correct and check data’s integrity and make sure the data is clean and doesn’t contain any errors and missing information.

WHERE

length(trip\_durations.ride\_id) = 16

AND trip\_durations.ride\_length\_sec > -1

AND (trip\_durations.member\_casual = 'casual' OR trip\_durations.member\_casual = 'member')

The next steps: I decided to prepare and ease the further analysis of my project, so I created a column ride\_length and day\_of\_week (1 = Sunday, 7 = Saturday).

WITH `trip\_durations` AS

(

SELECT \*,

DATE\_DIFF(ended\_at,started\_at,second) AS ride\_length\_sec

FROM `eminent-card-429217-h8.trip\_data.all\_trip`

)

SELECT

DISTINCT (trip\_durations.ride\_id),

trip\_durations.rideable\_type,

trip\_durations.member\_casual,

trip\_durations.started\_at,

`trip\_durations`.ended\_at,

EXTRACT(DAYOFWEEK from started\_at) as day\_of\_week,

ride\_length\_sec,

FORMAT(

'%d Days, %t',

DIV(ride\_length\_sec, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(ride\_length\_sec, 86400)))

) AS ride\_length

FROM `trip\_durations`

**Analyze:**

NOTE: I analyze and take into account trip lengths that have 0 or similar durations because I don’t know if it’s a simple error in the system database or if it’s just riders who end their trip immediately. I don’t have the opportunities and resources to find out the real reason behind it.

First of all, I do simple calculations, like mean, max, and min to find a general knowledge about the data.

SELECT

calcul\_1.mean\_ride\_length,

calcul\_1.max\_ride\_length,

FORMAT(

'%d Days, %t',

DIV(calcul\_1.mean\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(calcul\_1.mean\_ride\_length, 86400)))

) AS mean\_ride\_length\_t,

FORMAT(

'%d Days, %t',

DIV(calcul\_1.max\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(calcul\_1.max\_ride\_length, 86400)))

) AS max\_ride\_length\_t,

FORMAT(

'%d Days, %t',

DIV(calcul\_1.min\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(calcul\_1.min\_ride\_length, 86400)))

) AS min\_ride\_length\_t,

FROM (

SELECT

CAST(ROUND(AVG(ride\_length\_sec)) AS INT64) as mean\_ride\_length,

MAX(ride\_length\_sec) as max\_ride\_length,

MIN(ride\_length\_sec) as min\_ride\_length,

FROM `eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

) as calcul\_1

GROUP BY

calcul\_1.max\_ride\_length,

calcul\_1.mean\_ride\_length,

calcul\_1.min\_ride\_length

Using some formatting, I also made a more complex and well-viewed time that represents ride length not only by seconds, but by an hour, minutes, and seconds (HH: MM: SS). In some cases, even days.

By analyzing, I found that the minimum ride length is 0 seconds, the average is 15:07, max is 1 day, 01:59:56. Take into consideration that it’s general results from a whole 6 months. To find more precise and by month, I sometimes use

WHERE

(EXTRACT(MONTH FROM started\_at) = 11 AND EXTRACT(YEAR FROM started\_at) = 2023)

OR

(EXTRACT(MONTH FROM ended\_at) = 11 AND EXTRACT(YEAR FROM ended\_at) = 2023)

Other things like the number of rides per day of the week, and the mode of this, were found also quite easy.

SELECT

day\_of\_week,

COUNT(day\_of\_week) AS number\_of\_rides

FROM

`eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

GROUP BY

day\_of\_week

ORDER BY

number\_of\_rides DESC

LIMIT 1

Thursday has the most number of rides.

Quite interesting trends and insights were found in the average ride length between member and casual riders. Casual riders are longer than members.

SELECT

FORMAT(

'%d Days, %t',

DIV(trip\_data.avg\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(trip\_data.avg\_ride\_length, 86400)))

) AS avg\_ride\_length\_t,

member\_casual,

trip\_data.avg\_ride\_length

FROM (

SELECT

member\_casual,

CAST(ROUND(AVG(ride\_length\_sec)) AS INT64) as avg\_ride\_length

FROM `eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

GROUP BY member\_casual

) as trip\_data

GROUP BY

trip\_data.member\_casual,

trip\_data.avg\_ride\_length

For casual riders, 23:22 is the average time. And for member riders - 12:13.

A good difference between them is also shown in the average time by the day of the week. Casual riders are using bikes more time on average on weekends. On the other hand, members’ ridel length is balanced at the same time every day of the week.

SELECT

trip\_data.member\_casual,

trip\_data.day\_of\_week,

avg\_ride\_length,

FORMAT(

'%d Days, %t',

DIV(trip\_data.avg\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(trip\_data.avg\_ride\_length, 86400)))

) AS avg\_ride\_length\_t,

FROM (

SELECT

member\_casual,

day\_of\_week,

CAST(ROUND(AVG(ride\_length\_sec)) AS INT64) as avg\_ride\_length

FROM `eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

GROUP BY

member\_casual,

day\_of\_week

) as trip\_data

GROUP BY

trip\_data.member\_casual,

trip\_data.day\_of\_week,

trip\_data.avg\_ride\_length

ORDER BY

trip\_data.member\_casual,

trip\_data.day\_of\_week;

Overall, if we don’t diversify between them, the average time will be longer on Saturday and Sunday.

SELECT

trip\_data.day\_of\_week,

avg\_ride\_length,

FORMAT(

'%d Days, %t',

DIV(trip\_data.avg\_ride\_length, 86400),

TIME(TIMESTAMP\_SECONDS(MOD(trip\_data.avg\_ride\_length, 86400)))

) AS avg\_ride\_length\_t,

FROM (

SELECT

day\_of\_week,

CAST(ROUND(AVG(ride\_length\_sec)) AS INT64) as avg\_ride\_length

FROM `eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

GROUP BY

day\_of\_week

) as trip\_data

GROUP BY

trip\_data.day\_of\_week,

trip\_data.avg\_ride\_length

ORDER BY

trip\_data.day\_of\_week;

The number of rides, between member and casual riders is also different.

SELECT

member\_casual,

day\_of\_week,

COUNT(ride\_id) AS ride\_count

FROM

`eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

WHERE

(EXTRACT(MONTH FROM started\_at) = 11 AND EXTRACT(YEAR FROM started\_at) = 2023)

OR

(EXTRACT(MONTH FROM ended\_at) = 11 AND EXTRACT(YEAR FROM ended\_at) = 2023)

GROUP BY

member\_casual,

day\_of\_week

ORDER BY

member\_casual,

day\_of\_week;

Casual riders have more rides on the weekends, and members surprisingly have it the least here among the other days of the week. The Middle of the week is the highest and busiest ride for member riders.

SELECT

member\_casual,

COUNT(ride\_id) AS ride\_count

FROM

`eminent-card-429217-h8.trip\_data.all\_trip\_clean\_v2`

GROUP BY

member\_casual

ORDER BY

member\_casual

Overall, we can see that in the data most of the riders are members, and only around 25% are casual riders.

I also use functions, like this

EXTRACT(YEAR FROM started\_at) AS year,

EXTRACT(MONTH FROM started\_at) AS month

And then group them, so I can easily find what amount of something happened each month.

**Share:**

I use Tableau because it’s a powerful BI tool, that can easily show my findings and visualize data coherently and interactively. Plus Tableau can be easily presentable and accessible to my audience. I also created a Dashboard that includes all my visualizations, I also added some in-Tableau calculations and formulas for more presentable visuals. <https://public.tableau.com/views/Book3_17245137324840/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>

**Share (cont.) and Act:**

Of course, there are some assumptions and limitations that analysis and data give. For instance, it’s only 6 month period, we also don’t include other factors, like competitors in the analysis. So presentation and recommendation are not the most professional or right-through decision. The presentation is presented and created in Google Slides/Google Sheets.