**Google Data Analytics Certificate**

**Capstone Project**

**Case Study: How Does a Bike-Share Navigate Speedy Success?**

As a junior data analyst, I was a part of a marketing analyst team at Cyclistic, a fictional bike-share company from Chicago. My task was to assist my team 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.

To complete my task successfully, I used six data analysis phases including: Ask, prepare, process, analyze, share, and act.

**Statement of the business task:**

Conduct a comprehensive analysis and comparative assessment of the usage patterns demonstrated by Cyclistic bike company’s annual members and casual customers, aiming to discover disparities in their behavioral characteristics and preferences.

**Key Stakeholders:**

1. Lily Moreno - Director of Marketing and the manager of the junior data analyst team. She is responsible for developing campaigns and initiatives to promote the bike-share program.

2. Cyclistic Marketing Analytics Team - The team of data analysts responsible for collecting, analyzing, and reporting data to guide Cyclistic marketing strategy.

3. Cyclistic Executive Team - The detail-oriented executive team who will ultimately decide whether to approve the recommended marketing program.

4. Cyclistic Users - The customers who use the bike-share program and whose behavior and preferences will be analyzed to design a new marketing strategy to convert casual riders into annual members.

**Dataset:** The data for this study was made available by Motivate International Inc. <https://divvy-tripdata.s3.amazonaws.com/index.html>

**Documentation of the cleaning process:**

**Excel:**

* Removed all the duplicates.

A screenshot of a computer

Description automatically generated

* Made sure that all the tables are consistent (column names etc.)
* Removed all the extra spaces using TRIM function. (Where applicable)
* A screenshot of a computer

  Description automatically generated
* Made sure that all dates are in the same format.
* Used FILTER function to be sure that there is not any error or unusual values.
* A screenshot of a computer

  Description automatically generatedA screenshot of a computer

  Description automatically generated
* I created new column called “ride length” and set up values in “time” HH: MM: SS 37:30:55
* I created new column called “day of week” Where Sunday is 1 and Saturday is 7
* I deleted columns which I will not use for the analysis, making table easy to read and understand.
* Using the function “find and replace”, I located all blanks in spreadsheet and deleted them.
* Calculated mean(average) and max values in column “ride\_length”.
* Calculated mode in column “day\_of\_week”, to discover the most frequently occurring value that appears in this column.
* Using Custom Filter function, I found outliers and deleted all the rows that are showing ride length <01.00 min and 24.00 h>
* Made pivot tables for each of twelve files to get initial insights of how these two types of riders use bikes differently.
* To be specific, for each month, I calculated average ride length, average ride length by day, and number of rides per day, and displayed results through charts to make it easier for stakeholders to understand.

Average ride length

A screenshot of a computer

Description automatically generated

Average ride length by day (Where #1 is Sunday, #7 is Saturday)

A screenshot of a computer

Description automatically generated

Number of rides per day (Where #1 is Sunday, #7 is Saturday)

A screenshot of a computer

Description automatically generated

**SQL**

For further analysis I imported 12 files into Microsoft Sequel Server and got 12 tables (Jan-Dec)

* Using Union All I made one table called “Full\_Year” combining 12 tables.
* Next, I calculated the total number of rides throughout the year. I was interested to know how many rides were done by casual, and how many members (In percentages)
* Calculated total number of rides for each month for both type of riders showed in percentages. I wanted to see how they use bikes differently each month and is there any significant difference.
* Calculated the average ride length.
* Calculated total number of rides per day for each group. To discover which days are the busiest.
* Calculated how different types of bikes are used among two groups, showed in percentages.

**Visualization is done in Power BI**

Insights form Power BI:

1. In consideration of the aggregate rides conducted throughout the year 2022, it is observed that 59.04% of the total rides were availed by registered members, whereas 40.96% were attributed to non-registered or casual riders.
2. Throughout the duration of the year, an analysis of the data reveals that casual riders enjoyed an averaged ride duration of approximately 22 minutes, which was nearly double the average duration of 12 minutes observed for member riders.
3. By carefully analyzing the data on the number of rides per day, an interesting pattern emerges. It is evident that most rides during weekdays (Monday to Friday) were taken by members, whereas on weekends, casual riders had a slight edge. This suggests that casual riders tend to use bikes more for leisurely purposes, while members are more likely to use them for commuting to work or running personal errands.
4. Upon analyzing the line chart depicting the total rides per month, it becomes evident that member riders consistently utilize bicycles for a longer duration throughout the year, with their usage remaining relatively steady from May to October. In contrast, the season for casual riders is comparatively shorter. Notably, there is a substantial surge in casual ridership observed in May, which reaches its peak in July before experiencing a rapid decline starting in September.
5. When analyzing the data on bike usage by different rider types, we can observe that member riders tend to lean slightly towards classic bikes, while casual riders show a stronger preference for electric bikes.

**Conclusion:**

In conclusion, member and casual riders differ in terms of the proportion of rides taken, ride duration, riding patterns based on weekdays vs weekends, seasonal variation in bike usage, and bike preferences. Member riders appear to be more consistent and focused on commuting or personal errands, while casual riders take longer rides, use bikes more for leisure, and exhibit preferences toward electric bikes.

**Recommendation:**

Following are the reasons why casual riders should buy annual membership, and steps that company could take to initiate that move.

1. **Cost-effectiveness**: Highlight the cost benefits of becoming a member. Emphasize that regular riders can save money in the long run by opting for a membership instead of paying for the individual rides. Use digital media platforms to showcase membership plans, discounts, and promotions, making it clear that joining as a member offers financial advantages.
2. **Convenience and flexibility**: Position the membership as a convenient and flexible option for casual riders. Emphasize the ease of unlocking and using bikes, the freedom to pick up and drop off bikes at various locations, and the flexibility of riding anytime. Utilize digital media to demonstrate the user-friendly nature of the membership system, including more mobile apps, for easy booking and tracking.
3. **Community and social aspects**: Highlight the sense of community and belonging that comes with being a member. Showcase testimonials or stories from existing members, demonstrating the positive experiences and connections they have made. Utilize digital media platforms to foster engagement and interactions among members, such as creating online forums or social media groups for sharing experiences and organizing group rides.

**SQL Queries Explanation**

1. Total number of rides for whole year given in percentages for both rider types.

SELECT rider\_type,

COUNT(\*) AS Total\_rides,

CAST(ROUND(COUNT(\*) \* 100.0 / SUM(COUNT(\*)) OVER (), 2) AS DECIMAL(10, 2)) AS Percentage

FROM full\_year

GROUP BY rider\_type

**Explanation:** We could complete this query without using CAST function. However, the result was given with extra zeroes in percentage column.

To remove extra zeroes, we use CAST function to cast the result of the ROUND function to DECIMAL (10, 2). The Decimal (10, 2) means that there will be a maximum of 10 digits , with 2 of them representing decimal places.

1. Calculating average ride length for the whole year

SELECT

rider\_type AS RiderType,

AVG(DATEDIFF(minute, started\_at\_new, ended\_at\_new)) AS average\_ride\_length

FROM

Full\_Year

WHERE

rider\_type IN ('Casual', 'Member')

GROUP BY

rider\_type

In this query the “**DATEDIFF**” function is used to calculate the difference between started\_at\_new and ended\_at\_new timestamps in minutes. The **AVG** function is then applied to calculate the average duration for each distinct ‘rider\_type’. The result set will include two rows, one for each rider\_type., along with their corresponding average ride durations in minutes.

1. Calculating how often members and casual riders use different types of bikes.

SELECT

rider\_type AS RiderType,

CASE

WHEN rideable\_type = 'electric\_bike' THEN 'Electric'

WHEN rideable\_type = 'classic\_bike' THEN 'Classic'

WHEN rideable\_type = 'docked\_bike' THEN 'Docked'

ELSE rideable\_type

END AS BikeType,

COUNT(\*) AS Count,

CAST(COUNT(\*) \* 100.0 / SUM(COUNT(\*)) OVER (PARTITION BY rider\_type) AS DECIMAL(5, 2)) AS Percentage

FROM

Bike\_Share\_Project.dbo.full\_year

WHERE

rider\_type IN ('Casual', 'Member') AND

rideable\_type IN ('electric\_bike', 'classic\_bike', 'docked\_bike')

GROUP BY

rider\_type, rideable\_type

**Explanation**:

**CAST(COUNT(\*) \* 100.0 / SUM(COUNT(\*)) OVER (PARTITION BY rider\_type) AS DECIMAL(5, 2)) AS Percentage**:

This calculates the percentage of each group's count (defined by "rider\_type") relative to the total count of that rider type. The percentage is rounded to two decimal places.

1. --Calculating total number of rides by day of week for both types of riders

SELECT rider\_type,

CASE DATEPART(WEEKDAY, started\_at\_new)

WHEN 1 THEN 'Sunday'

WHEN 2 THEN 'Monday'

WHEN 3 THEN 'Tuesday'

WHEN 4 THEN 'Wednesday'

WHEN 5 THEN 'Thursday'

WHEN 6 THEN 'Friday'

WHEN 7 THEN 'Saturday'

END AS day\_of\_week,

COUNT(\*) AS total\_rides

FROM Bike\_Share\_Project.dbo.full\_year

GROUP BY DATEPART(WEEKDAY, started\_at\_new), rider\_type

ORDER BY DATEPART(WEEKDAY, started\_at\_new)

**Explanation:**

The **CASE** statement here is used to transform the numeric value of the day of the week (1 for Sunday, 2 for Monday, etc.) into a corresponding string representation (Sunday, Monday, etc.). The **DATEPART(WEEKDAY, started\_at\_new)** function returns the numeric representation of the day of the week for the **started\_at\_new** column, which likely contains the date and time when the bike ride started. The result of this **CASE** statement will be an additional column named **day\_of\_week.**

The **GROUP BY** clause groups the rows based on the **rider\_type** and the numeric representation of the day of the week (**DATEPART(WEEKDAY, started\_at\_new)**). This grouping allows the **COUNT(\*)** function to calculate the total number of rides for each combination of **rider\_type** and **day\_of\_week**

The **ORDER BY** clause orders the final result set based on the numeric representation of the day of the week. This ensures that the rows are presented in chronological order from Sunday to Saturday.

1. --Total number of rides each month for both types of riders

SELECT

DATENAME(MONTH, started\_at\_new) AS month,

rider\_type,

COUNT(\*) AS total\_rides

FROM Bike\_Share\_Project.dbo.full\_year

GROUP BY DATENAME(MONTH, started\_at\_new), rider\_type

ORDER BY MONTH

**Explanation:** In this query, we're using the **DATENAME** function with the **MONTH** parameter to get the month names from the **started\_at** column. This function will return the month name corresponding to the date in the **started\_at** column.