**Case study: How does a bike-share navigate speedy success during the pandemic?**

**By Chinmay Dande**

**Phase One – Ask**

**Business Task**

The objective of this study is to identify how casual members and annual members use the Cyclistic App during the summer months of the pandemic.

**Key Stakeholders**

* Lily Moreno - Director of Marketing
* Cyclistic Marketing team
* Cyclistic Executive team

**Phase Two – Prepare**

**Summary of Preparation**

The data we will be referring to will be the during the Spring time of the pandemic (April to June 2020). The data is organized in a spreadsheet with various attributes and columns. It is unorganized and contains duplicates with mix of NA values. The data fits the ROCCC criteria and contains bias in terms of the areas being surveyed.

The data is open source and is available on the company website and is constantly updated. Once the data is organized properly with relevant summary statistics, we can then create visualizations for the data. First, we will filter and sort the data and then begin to process it.

**Phase Three – Process**

**Tools**

* SQL using BigQuery
* Excel for spreadsheet analysis
* R for summary stats and visualizations
* Tableau

**Cleaning Process**

* For the cleaning process to ensure data integrity, I am checking for duplicates, missing values, nulls and inconsistent formatting.
* Using Excel’s built in functions I was able to delete a bunch of duplicates and make the formatting more universal for each dataset.
* Next, I created a column for “rider\_length” and “day\_of\_week” to length of each bike ride and on what day of the week. This will help in understanding the difference in terms of frequency of use when comparing casual members and annual members.
* Additionally, I separated the characters of start and end times/dates to create a better picture of when riders used the bike and how long.

**Phase Four – Analyze**

**Trends and Relationships**

**Spreadsheet**

* Calculate; mean, max, min, mode, average and sum of rider length.

**April Summary Stats**

|  |  |
| --- | --- |
| **Attributes** | **Summary Stats** |
| Average | 0.02489584234 |
| Min | 0.002037037 |
| Max | 25.6380787 |
| Count for negative values | 51 |
| Mode | 0.004537037 |
| Median | 0.011296296 |
| Average (Wednesday) | 3.791556573 |

**May Summary Stats**

|  |  |
| --- | --- |
| Attributes | Summary Stats |
| Average | 0.01382577934 |
| Min | -0.9974189815 |
| Max | 0.4423148148 |
| Count for negative values | 1659 |
| Mode | 0.00003472222222 |
| Median | 0.01268518519 |
| Average (Thursday) | 4.527966686 |

**June Summary Stats**

|  |  |
| --- | --- |
| Attributes | Summary Stats |
| Average | 0.01156969796 |
| Min | -0.9984953704 |
| Max | 0.7821296296 |
| Count for negative values | 3533 |
| Mode | 0.005914351852 |
| Median | 0.01270833333 |
| Average (Thursday) | 4.061698809 |

**SQL**

In this phase, we are looking at various trends the data indicates:

* We start off by checking the aggregate functions in SQL to check; Distinct, mean, max, min and mode.
* The average rider length, average rider length during a specific day of the week and average customer type based on length and day of week.

**Queries used**

**#Average Rider Length**

SELECT

rider\_length,

(SELECT

AVG(rider\_length)

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`) AS avg\_rider\_length

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

**#Minium Rider Length**

SELECT

rider\_length,

(SELECT

MIN(rider\_length)

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`) AS min\_rider\_length

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

**#Max Rider Length**

SELECT

rider\_length,

(

SELECT

Max(rider\_length)

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`) AS max\_rider\_length

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

**# Number of Average Rides based on customer type**

SELECT

rider\_length,

COUNT(member\_casual) AS number\_of\_rides,

day\_of\_week,

(

SELECT

AVG(day\_of\_week)

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`) AS max\_rider\_length

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

GROUP BY

rider\_length,

day\_of\_week

**# Number of Average Rides based on Casual members**

SELECT

rider\_length,

day\_of\_week,

(

SELECT

COUNT(member\_casual) AS avg\_rides\_casual

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

WHERE

member\_casual = 'casual') AS casual\_member

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

GROUP BY

rider\_length,

day\_of\_week,

member\_casual

ORDER BY

member\_casual

**# Number of Average Rides based on annual members**

SELECT

rider\_length,

day\_of\_week,

(

SELECT

COUNT(member\_casual) AS avg\_rides\_casual

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

WHERE

member\_casual = 'member') AS annual\_member

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

GROUP BY

rider\_length,

day\_of\_week,

member\_casual

ORDER BY

member\_casual

#Average usage of annual members based of day of week

SELECT

AVG(rider\_length) AS average\_rider\_length,

AVG(day\_of\_week) AS average\_usage\_per\_day\_of\_week,

(

SELECT

COUNT(member\_casual) AS avg\_rides\_casual

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

WHERE

member\_casual = 'member') AS annual\_member

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

GROUP BY

rider\_length,

day\_of\_week,

member\_casual

ORDER BY

average\_usage\_per\_day\_of\_week DESC,

member\_casual

LIMIT 10

#Number of rides based on time of day for casual members

SELECT DISTINCT

start\_time,

member\_casual,

CASE

When start\_time >= "6:00" Then 'Morning'

When start\_time >= '12:00'Then 'Afternoon'

When start\_time >= '17:00' Then 'Evening'

Else 'Night'

End as time\_of\_day

FROM

`bamboo-weft-412703.Cyclist\_data.June 2020`

WHERE

member\_casual = "casual"

GROUP BY

start\_time,

member\_casual

ORDER BY

time\_of\_day

#Average day of week where bikes are used for casual and annual members

SELECT DISTINCT

AVG(day\_of\_week) as day\_casual,

member\_casual,

FROM

`bamboo-weft-412703.Cyclist\_data.April 2020`

GROUP BY

member\_casual

ORDER BY

member\_casual,

day\_casual

**Analysis**

Based on stats and queries, the average rental is on Wednesday and Thursday afternoons for both members during the springtime of the pandemic. The most popular time of rental was Sundays for both users. Next, we will display visually the findings and insights we found.

**Phase Five – Share**

**Visualization**

Using RStudio, I will prepare a data viz by installing the ggplot2 package. Next, using the bar graph feature I will outline which day has the highest usage. Lastly, the fill function to show which type of customer uses the service the most.

Both visuals below executed with the following code:

#ggplot 1

ggplot(data=cyclistic\_data) + geom\_bar(mapping = aes(x=day\_of\_week, fill=member\_casual)) + facet\_wrap(~member\_casual)

#ggplot2

ggplot(data=cyclistic\_data) + geom\_point(mapping = aes(x=member\_casual, y= rider\_length)) + facet\_wrap(~day\_of\_week)

**A graph of a group of people

Description automatically generated with medium confidenceA graph of a group of people

Description automatically generated**

**#ggplot 1 #ggplot2**

Both Plots showcase casual and annual members most popular day of the week being Sunday during the Spring time of the pandemic.