Case Study 1

How Does a Bike-Share Navigate Speedy Success?

Project Title: Cyclistic Bike Share Analysis

- This project was an interactive capstone on the Google Analytics course curriculum hosted by Coursera. This is my first project in which to showcase my initial workflow through a fictional data analytics business project.

Introduction

- The initial scenario is of a junior data analyst working for a marketing division is asked to provide information using the company's existing database to answer the following business questions:
    - The company's future success depends on increasing the number of riders 2 become annual members, by joining their membership program.
    - The team wants to understand the different and similar behaviors of the casual riders and their annual member riders.
    - The business goal is to develop clarity within the data for the business marketing side of the company to focus their efforts in converting casual riders to utilizing the membership program.

Data Acquisition and Preparation

- Data sources for this project were provided as first party data collected over 2014 to 2022. They were housed in a zip file consisting of CSV files.
- Upon further examination these files represented a single month's worth of data and ranged from 50 to 100 megabytes, which in programs like Google Sheets and Microsoft Excel represented 500,000 to 800,000 records.
- To develop analysis which could provide trends throughout a single year these monthly files were uploaded to Google BigQuery/SSMS and unions into a single year for 2021.

Data Cleaning

- Primary key right ID checked for duplicates.
- Vital fields related to ride start times, end times, and casual or member status checked for blank or null field.

  Analysis identified 1440 no start times which was deemed insignificant to affect overall analysis of over 5,500,000 records or 2.6 X $10^{-4}$and filtered out.

Exploratory Data Analysis (EDA)

- Create a new column titled ride time, derived from utilizing time and date functions in SQL increments in minutes.
- Provide additional columns for the month of the year, day of the week, and year extract for further analysis.
- Main data table divided into casual and member data tables for comparison analysis.
- Both members and casual tables further filtered why month, day of the week, bike type, and start station used
- Data filtered relative to ride stations, casual, members, and aggregated to count ride_id
- Negative ride_time records due to end time greater than start time, assuming an incorrect records.

    - 87 Records were identified out of 5,595,063 accounting for 1.6 X10-5 of all records in 2021, and considered insignificant to effect analysis but filter out for descriptive statistics.

Research Questions or Hypotheses Stop

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

Data Analysis

Descriptive Stats

- Calculated Avg, Min, and Max to ride_time for all data.
    - MIN = 0 minutes (excluding neg rides)
    - MAX = 55,944 (39 days possible outlier)
    - AVG = 21 minutes

- **Ride-Times = 0 min duration (51,137 records)**
    - Casual = 20,219 Records
    - Member = 30,918 Records
- **Ride-Times = 1 -5 min duration (833,193 records)**
    - Casual = 216,288 Records
    - Member = 616,905 Records
- **Ride-Times = 5 – 15 min duration (2,433,756 records)**
    - Casual = 969,384 Records
    - Member = 1,464,372 Records
- **Ride-Times = 15 – 30 min duration (1,333,782 records)**
    - Casual = 702,828 Records
    - Member = 630,954 Records
- **Ride-Times = 30 – 60 min duration (572,135 records)**
    - Casual = 375,747 Records
    - Member = 196,388 Records
- **Ride-Times = 60 – Day min duration (240,862 records)**
    - Casual = 217,379 Records
    - Member = 23,483 Records
- **Ride-Times = Greater than a Day min duration (70,198 records)**
    - Casual = 27,160 Records
    - Member = 43,038 Records

Conclusion

Thank you for the opportunity to present the findings from our recent data analysis of Cyclistic's rider behavior for the year 2021. This analysis provides valuable insights into the composition and trends within our rider community, helping us better understand the dynamics of our bike share program.

Casual Riders Comprise 45% of Annual Ride Volume

In 2021, casual riders accounted for a substantial portion of Cyclistic's ride volume, making up 45% of the total rides for the year. This indicates a significant presence and reliance on our bike-sharing services by non-member riders.

Seasonal Patterns in Rider Behavior

Our analysis revealed distinct seasonal patterns in rider behavior. Specifically, we observed a recurring pattern of reduced ride volume for both casual and member riders during the late fall and winter months, with peak activity occurring during the summer months. Notably, during the summer, casual rides exceeded member rides. However, for the remainder of the year, including the fall months leading into winter, member ride volume consistently exceeded that of casual riders.

Weekly Patterns: Casual vs. Member Riders

On a weekly basis, we noticed interesting disparities between casual and member riders. Casual rider volume was notably higher than member rider volume during the weekends. In contrast, member rider volume remained consistently higher and relatively stable throughout the workweek. This indicates that while casual riders are more active on weekends, our membership base contributes to consistent weekday ridership.

Most Popular Ride Duration

When analyzing ride duration, we found that both casual and member riders predominantly opted for rides lasting between 5 to 15 minutes. These rides accounted for a substantial 43.5% of all annual ride volume, highlighting a preference for short-duration trips among our riders.

Stations with Casual Rider Dominance

A station-level analysis identified top locations where casual rider volume significantly exceeded that of member riders by approximately 5,000 rides annually. In total, these stations contributed to a variance of 179,000 rides. This data suggests that certain stations are particularly attractive to casual riders, possibly due to their relevance for tourism or recreational purposes.

Implications and Further Considerations

The similarities in ride duration across both rider groups suggest that the system may serve as a commuting tool for short-distance travel. However, without access to financial data, we cannot determine how many daily commutes a single user is making.

In conclusion, our data analysis reveals that Cyclistic's casual riders exhibit behaviors that align with tourism and recreational use of our bike share program. Additionally, our membership base contributes to consistent ridership throughout the workweek. Understanding these trends and patterns allows us to tailor our marketing efforts and services more effectively to cater to the needs of both casual and member riders.

As we move forward, further exploration into rider preferences, additional data sources, and financial information could provide us with deeper insights to refine our strategies and improve the overall experience for all Cyclistic riders.

Future Directions

- If an invoice or customer id become part of the record then further analysis could be on the behavior of Casual users riding behaviors allowing for a more targeted business plan

Code

```
-----------------------------------------------------------------
/* Cyclistic data for 2021, imported CSV, primary source*/

-----------------------------------------------------------------
 -- Performed a union of 12 months of data into a single year
INSERT INTO [Cyclistic].[dbo].[2021TotalRaw]
```

```
SELECT *
FROM[Cyclistic].[dbo].[2021_01]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_02]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_03]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_04]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_05]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_06]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_07]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_08]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_09]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_10]
UNION
SELECT *
FROM[Cyclistic].[dbo].[2021_11]
```

```
UNION

SELECT *

FROM[Cyclistic].[dbo].[2021_12]

--------------------------------------------------------------------
-- Verification of complete data set

SELECT *
FROM [Cyclistic].[dbo].[2021TotalRaw]

SELECT COUNT(ride_id),
       member_casual
FROM [Cyclistic].[dbo].[2021TotalRaw]
GROUP BY member_casual

-- Result identify Casual riders at 2,529,005
-- Result identify Member riders at 3,066,058
-- Total records 5,595,063
----------------------------------------------------------------------
-- ID duplication (records will not be erased)

SELECT ride_id,
          started_at,
          COUNT(ride_id)

FROM [Cyclistic].[dbo].[2021TotalRaw]

GROUP BY ride_id, started_at

HAVING COUNT(ride_id) > 1

-----------------------------------------------------------------------
-- Adding filtering columnes
                     --ride_time (minutes), month, day of the week
                     -- using top 20 rows for quick iterations
SELECT TOP 20
       ride_id,
       rideable_type,
       started_at,
       ended_at,
       member_casual,
       start_station_name,
       end_station_name,
 DATEDIFF(minute,started_at,ended_at) AS ride_time,
 DATENAME(MONTH,DATEADD(month, 0, started_at)) AS month_Name,
 DATENAME(weekday, started_at) AS day_of_week_Name

FROM [Cyclistic].[dbo].[2021TotalRaw]

-----------------------------------------------------------------------
-- Inserting results into a new table

INSERT INTO [Cyclistic].[dbo].[2021TotalFiltered]

SELECT --TOP 20
       ride_id,
       rideable_type,
       started_at,
       ended_at,
       member_casual,
       start_station_name,
       end_station_name,
 DATEDIFF(minute,started_at,ended_at) AS ride_time,
 DATENAME(MONTH,DATEADD(month, 0, started_at)) AS month_Name,
 DATENAME(weekday, started_at) AS day_of_week_Name

FROM [Cyclistic].[dbo].[2021TotalRaw]


----------------------------------------------------------------------
-- Filter data by member and casual by month to see if there is
       -- any annual behaviors between the groups

SELECT
       month_Name,
       COUNT(ride_id) AS rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'member'

GROUP BY month_Name
----------------------
SELECT
       month_Name,
       COUNT(ride_id) AS rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'casual'

GROUP BY month_Name

----------------------------------------------------------------
-- Filter data by member and casual by weekdays to see if there is
       -- any weekly behaviors between the groups
SELECT
       day_of_the_week,
       COUNT(ride_id) AS rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'casual'

GROUP BY day_of_the_Week
----------------
SELECT
       day_of_the_week,
       COUNT(ride_id) AS rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'member'

GROUP BY day_of_the_Week

-------------------------------------------------------
--- Adhoc filtering to explore variance and similarity
-- between the casual and member riders
SELECT
       start_station_name,
       ride_Time,
       month_Name,
       day_of_the_Week,
       ride_id
```

```
FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'casual'
       AND start_station_name IS NOT NULL
              AND ride_Time >=0
                     AND month_Name IN ('January','February','March','April','May')
----------------------------------------
-- Count the numher of rides logged by the station
-- for the casual riders
SELECT
       start_station_name,
       COUNT(ride_id) AS num_Rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'Casual'


GROUP BY start_station_name
--------------------------------------------
-- Count the numher of rides logged by the station
-- for the casual riders

SELECT
       start_station_name,
       COUNT(ride_id) AS num_Rides

FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'member'


GROUP BY start_station_name

Order BY num_Rides DESC

----------------------------------------------------
--Investigating Ride time data, Found neg values


SELECT
       ride_id,
       AVG(ride_time) AS Average_Ride,
       MIN(ride_time) AS Min_Ride,
       MAX(ride_time) AS Max_Ride
FROM [Cyclistic].[dbo].[2021TotalFiltered]

------------------------------------------------------
--Time duration by groups per Member and Casual

SELECT
       ride_id,
       ride_Time,
       member_casual,
       CASE
              WHEN ride_Time =0 THEN '0'
              WHEN ride_Time >1 AND ride_Time <=5 THEN '1-5'
              WHEN ride_Time >5 AND ride_Time <=15 THEN '5-15'
              WHEN ride_Time >15 AND ride_Time <=30 THEN '15-30'
              WHEN ride_Time >30 AND ride_Time <=60 THEN '30-60'
              WHEN ride_Time >60 AND ride_Time <=1440 THEN '60-Day'
                     ELSE 'Greater than Day'
                            END AS 'ride_group'
```

```
FROM [Cyclistic].[dbo].[2021TotalFiltered]

WHERE member_casual = 'Casual'
```

Visualizations

ADD Link

Acknowledgments and Cited Materials linked

- https://d3c33hcgiwev3.cloudfront.net/ymogSWd_R2ujQawZle3_rQ_12891ea7af0a487bad109a95d513b2f1_DA-C8-Case-Study-1-PDF.pdf?Expires=1694390400&Signature=FEml-Jivto5S7LFy8gZD9c~gNm6G-bZJDlPvTKPJcJYP-fPKyXWiumDXa0xYZL8VM3YHHd~WW-u~6561xhb~kBUQcublet6Exv6I5EOQ4auucQCO-98HpM8~1LlxNZqeTZhDe1Y1S9ZihionogiVs8QUJO9KvKt-Kzp0meMVBzU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A

Data

- https://divvy-tripdata.s3.amazonaws.com/index.html

Tools

- ○ BigQuery
- ○ SSMS
- ○ Power BI

Contact Information

- Provide your contact information for potential collaboration or inquiries.

Remember to tailor your project to your specific interests and the dataset you are working with. Your portfolio should showcase your data analysis skills, from data acquisition and preparation to visualization and interpretation of results. Additionally, make sure to highlight your ability to communicate your findings effectively through clear explanations and visualizations.