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

Introduction

This is the final capstone project in my Google Data Analytic Professional Certificate course. In this case study, I will work for a fictional company, Cyclistic, as a junior data analyst. In order to answer the key business questions, I will follow the steps of the data analysis process: **ask**, **prepare**, **process**, **analyze**, **share**, and **act**. To ensure professionalism, I applied the following techniques to complete this project: **R** programming language (RStudio), **Microsoft office** (EXCEL), and **data visualization** tool (Tableau).

Scenario

As 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, our team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve team's recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Characters and teams

Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.
Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
Cyclistic marketing analytic team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.
Cyclistic executive team: The notoriously detail-oriented executive team will decide

whether to approve the recommended marketing program.

Ask (Deliverable: A clear statement of the business task)

Moreno (director of marketing and my manager) has assigned me the question: **How do annual members and casual riders use Cyclistic bikes differently?**

Key tasks:

1. **Identify the business task**
    - The goal is to develop a marketing strategy to be able to convert casual riders into annual members by analyzing the difference in product usage between casual riders and annual members.

2. **Consider key stakeholders**
    - Director of Marketing and my manager (Lily Moreno), Marketing Analytics team, and Cyclistic Executive team.

Prepare (Deliverable: A description of all data sources used)

Key tasks

1. **Determine the credibility of the data**
    - Data is from Cyclistic's own database and is updated on a monthly basis. Since it is first-hand data collected by Cyclistic itself, it can be considered credible.

2. **Download data and store it appropriately.**
    - I will use Cyclistic's historical trip data to analyze and identify trends. Downloaded the previous 12 months (2021/04 to 2022/03) of Cyclistic trip data here, and stored in local file. The data has been made available by Motivate International Inc. under this license.

3. **Identify how it's organized.**
    - Data names: YYYYMM-divvy-tripdata
    - Data Types: .CSV (UTF-8).
    - Contents: There are total 13 columns containing information related to ride_id, rideable_type, started_at, ended_at, etc.

4. **Sort and filter the data.**
    - Upload all datasets in R. After confirming that each column represents the same kind of information, combine all datasets into one data frame and export as .csv.

Process (Deliverable: Documentation of any cleaning or manipulation of data)

```
> summary(Alltrip_data)
   ride_id          rideable_type        started_at
 Length:5723532     Length:5723532     Min.   :2021-04-01 00:03:18
 Class :character   Class :character   1st Qu.:2021-06-22 15:20:26
 Mode  :character   Mode  :character   Median :2021-08-17 18:25:49
                                       Mean   :2021-08-26 22:25:18
                                       3rd Qu.:2021-10-14 19:48:10
                                       Max.   :2022-03-31 23:59:47

    ended_at                    start_station_name start_station_id
 Min.   :2021-04-01 00:14:29   Length:5723532     Length:5723532
 1st Qu.:2021-06-22 15:47:37   Class :character   Class :character
 Median :2021-08-17 18:44:32   Mode  :character   Mode  :character
 Mean   :2021-08-26 22:46:50
 3rd Qu.:2021-10-14 20:03:28
 Max.   :2022-04-01 22:10:12

 end_station_name   end_station_id       start_lat       start_lng
 Length:5723532     Length:5723532     Min.   :41.64   Min.   :-87.84
 Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
 Mode  :character   Mode  :character   Median :41.90   Median :-87.64
                                       Mean   :41.90   Mean   :-87.65
                                       3rd Qu.:41.93   3rd Qu.:-87.63
                                       Max.   :45.64   Max.   :-73.80

    end_lat          end_lng        member_casual
 Min.   :41.39   Min.   :-88.97   Length:5723532
 1st Qu.:41.88   1st Qu.:-87.66   Class :character
 Median :41.90   Median :-87.64   Mode  :character
 Mean   :41.90   Mean   :-87.65
 3rd Qu.:41.93   3rd Qu.:-87.63
 Max.   :42.17   Max.   :-87.49
 NA's   :4716    NA's   :4716
```

Key tasks

1. **Check the data for errors.**
   - Drop all “NA” columns
2. **Choose your tools.**
   - Due to the large number of datasets, I choose to work with **R** here.
3. **Document the cleaning process.**
   - Create a column called “ride_length.” Calculate the length of each ride by subtracting the column “started_at” from the column “ended_at”.
   - Separate the dates into month, day, year, and create a column called

“day_of_week”.

Analyze (Deliverable: A summary of your analysis)

```
> summary(Cleaned_Alltrip_data)
   ride_id          rideable_type        started_at                    ended_at
 Length:4641395     Length:4641395     Min.   :2021-04-01 00:03:18   Min.   :2021-04-01 00:14:29
 Class :character   Class :character   1st Qu.:2021-06-19 15:15:57   1st Qu.:2021-06-19 15:45:10
 Mode  :character   Mode  :character   Median :2021-08-13 18:15:54   Median :2021-08-13 18:35:09
                                       Mean   :2021-08-22 17:46:59   Mean   :2021-08-22 18:08:21
                                       3rd Qu.:2021-10-09 00:04:35   3rd Qu.:2021-10-09 00:23:48
                                       Max.   :2022-03-31 23:59:47   Max.   :2022-04-01 22:10:12
 start_station_name start_station_id   end_station_name   end_station_id       start_lat       start_lng
 Length:4641395     Length:4641395     Length:4641395     Length:4641395     Min.   :41.65   Min.   :-87.83
 Class :character   Class :character   Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :41.90   Median :-87.64
                                                                            Mean   :41.90   Mean   :-87.64
                                                                            3rd Qu.:41.93   3rd Qu.:-87.63
                                                                            Max.   :45.64   Max.   :-73.80
    end_lat         end_lng       member_casual       ride_length      ride_distance
 Min.   :41.65   Min.   :-87.83   Length:4641395     Min.   : -3354   Min.   :   0.0000
 1st Qu.:41.88   1st Qu.:-87.66   Class :character   1st Qu.:   407   1st Qu.:   0.8972
 Median :41.90   Median :-87.64   Mode  :character   Median :   717   Median :   1.6076
 Mean   :41.90   Mean   :-87.64                      Mean   :  1281   Mean   :   2.1164
 3rd Qu.:41.93   3rd Qu.:-87.63                      3rd Qu.:  1300   3rd Qu.:   2.7961
 Max.   :42.17   Max.   :-87.52                      Max.   :3356649   Max.   :1192.2456
```

Key tasks

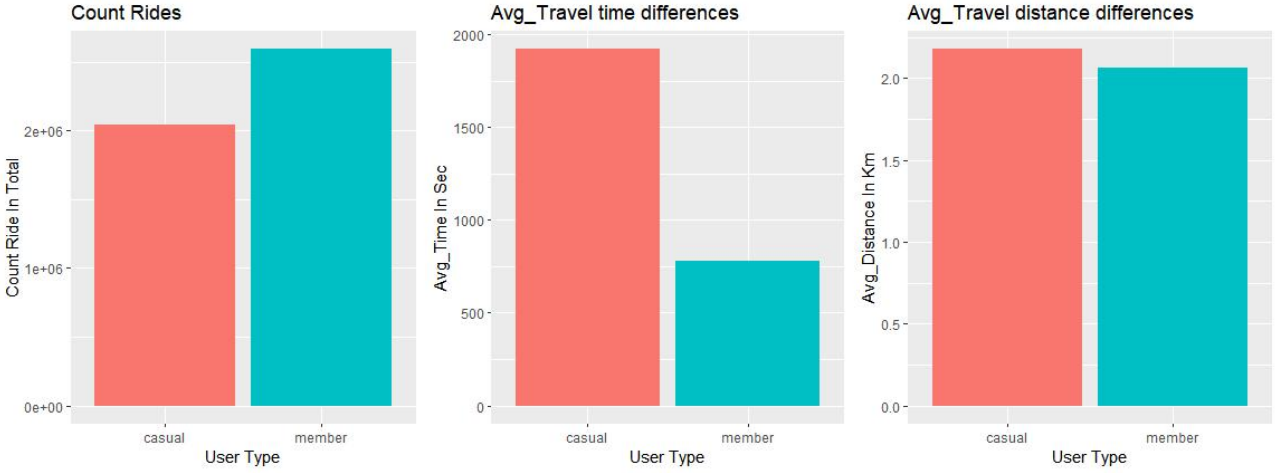
1. Aggregate data so it’s useful and accessible.

2. Organize and format data.

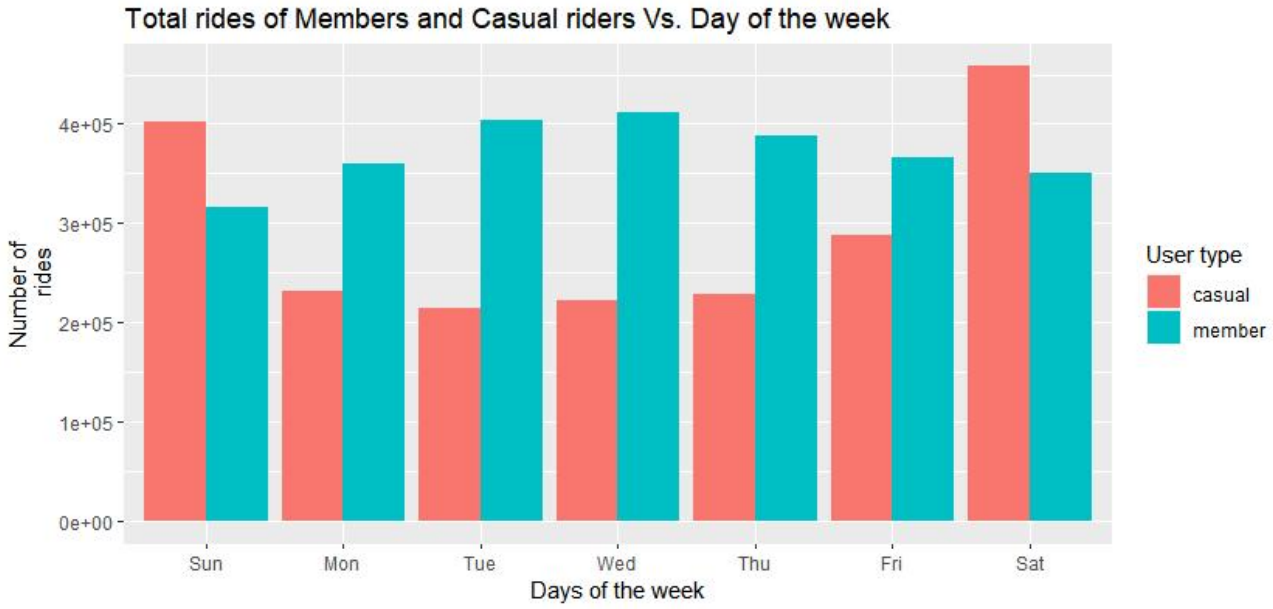
3. Perform calculations.

4. Identify trends and relationships.

- Count total users
- Plotting differences between usage of member vs. Casual



- Get differences in numbers between member and casual by days of the week



Summary:

From the bar chart we can see that:

- The total number of rides by members is higher than that of casual riders;
- The average ride time of casual riders is more than double that of members;
- The average ride distance of casual riders is slightly higher than that of members;
- Member riders ride more evenly throughout the week, while casual riders ride significantly more on weekends than weekdays.

Preliminary inference: Casual riders are more likely to use cycling as a form of exercise or sightseeing rather than a daily commute.

Share (Deliverable: Supporting visualizations and key findings)

Key tasks

1. **Create effective data visualizations.**
   - I'm going to use **Tableau** here to do visualizations to help with deeper analysis.
2. **Ensure your work is accessible**
   - I will upload all relevant data, analysis, r coding, sql, etc. to the Guihub repository and make it public.

Act **(Deliverable: Top three recommendations based on analysis)**

Key findings:

1. Random rider activity. Keywords: summer, weekend, afternoon

- From April to the peak in July and then gradually decline.
- The weekends of the week are most active.
- Be more active on weekends from 1pm to 3pm.
- On average, they are more active around 5:00 pm each day.

2. Casual riders average just over 30 minutes per ride.

Suggests:

1. Reduce the total marketing investment before 12:00 noon every day, because most of the riders are not active before the afternoon.
2. Significantly increase the marketing delivery from 1:00 pm to 3:00 pm on weekends and 5:00 pm on weekdays. The investment intensity can be gradually increased from April to July and then reduced to the average.
3. While increasing the price by a small amount according to the riding time, provide members with a discount for each ride of more than 30 minutes.