Capstone_Project_Google_Data_Analisys_Certificate

Karina Jiménez Piedra

2025-11-18

Context / Challenge

Cyclistic, a bike-rental company, serves two types of customers: casual riders, who rent bikes occasionally, and annual members, who pay a monthly fee and use the service regularly. The financial team has identified that members are more profitable. This project aims to describe the characteristics of each customer group and use those insights to support a marketing strategy that encourages casual riders to become members.

R work

Preparing the enviroment

Load libraries

```
# Load libraries ----

library(tidyverse)
library(here)
library(skimr) # Nicer summary
library(janitor)
library(dplyr)
```

Cleaning phase

Data integrity was ensured by following a guideline that checks the dataset across below key dimensions:

- Data Collection/Source
- Data Structure
- Data Completeness
- Data Consistency
- Data Accuracy/Validity
- Data Quality Assurance
- Security & Privacy

Several steps were carried out to meet these quality metrics. The most relevant ones were:

Several steps were performed to ensure the dataset was consistent and ready for analysis. The most relevant ones are described below.

1. Fixing date and time formats

To ensure consistency across both datasets, the time-related columns (`start_time, end_time, started_at, ended_at`) were converted to proper date-time objects using `ymd_hms()`. This step was necessary to calculate ride_length and perform time-based analyses accurately.

2. Handling missing values

The percentage of missing values in each dataset (data_2019 and data_2020) was evaluated, and rows containing NA were removed using `na.omit()`. Given the large size of the datasets, removing incomplete rows did not significantly affect the analysis and simplified the cleaning process.

3. Renaming columns

Column names in data_2019 were renamed using `rename()` to match the structure of data_2020. For example, `trip_id` was renamed to `ride_id`, `start_time` to `started_at`, and `usertype` to `member_casual`. This ensured the datasets could be combined without inconsistencies.

4. Stacking the datasets

After aligning column names and data types, the datasets were combined using `bind_rows()` to create a single unified data frame `all_trips`. This enabled consistent comparisons between member and casual riders.

5. Removing bad data

Rows with unrealistic values were removed, including rides with negative ride_length or test records from the `start_station_name` "HQ QR". This step ensured that the analysis was based on valid and meaningful data.

Descriptive phase

For descriptive analysis column `ride_length` was adding to data set and allowed identify ride times on each group of costumers.

Statistical metrics as **mean, median, maximum and minimum** were calculated:

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

Output ***ride_length****variable:*

Rider Type	Mean (secs)	Median (secs)	Max (secs)	Min (secs)
Casual	5761.0099	1366	9,387,024	2
Member	794.1507	507	6,096,428	1

Customer segments proportion:

```
all_trips_v2 %>%
  count(member_type = member_casual) %>%
  mutate(percent = n/sum(n))
```

Output:

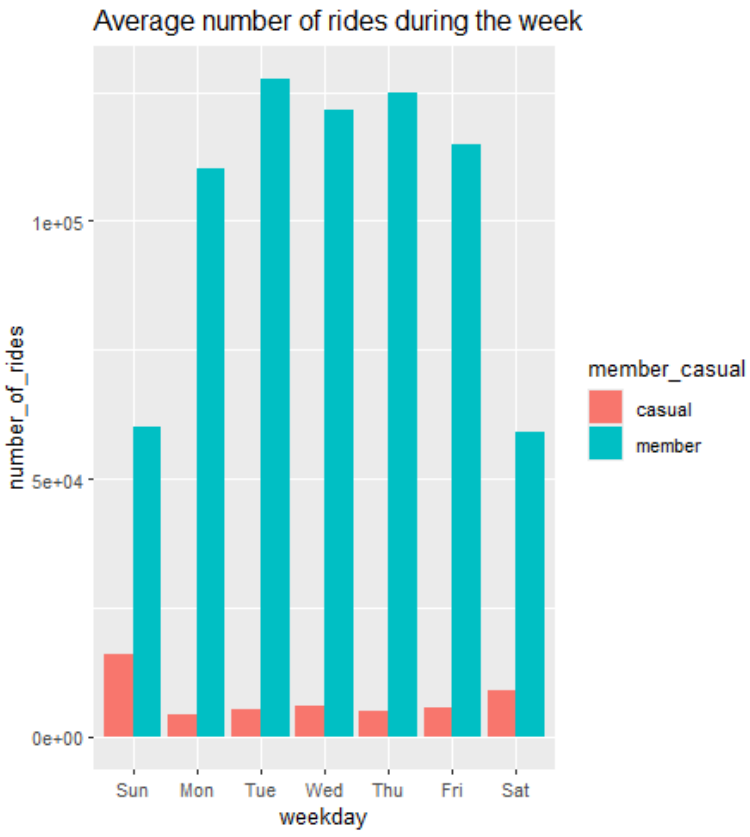
member_type	Quantity	Percent
casual	50,648	0.0659
member	717,829	0.9340

Visual Creations

Let's visualize the number of rides by rider type:

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  ggtitle("Average number of rides during the week") +
  geom_col(position = "dodge")
```

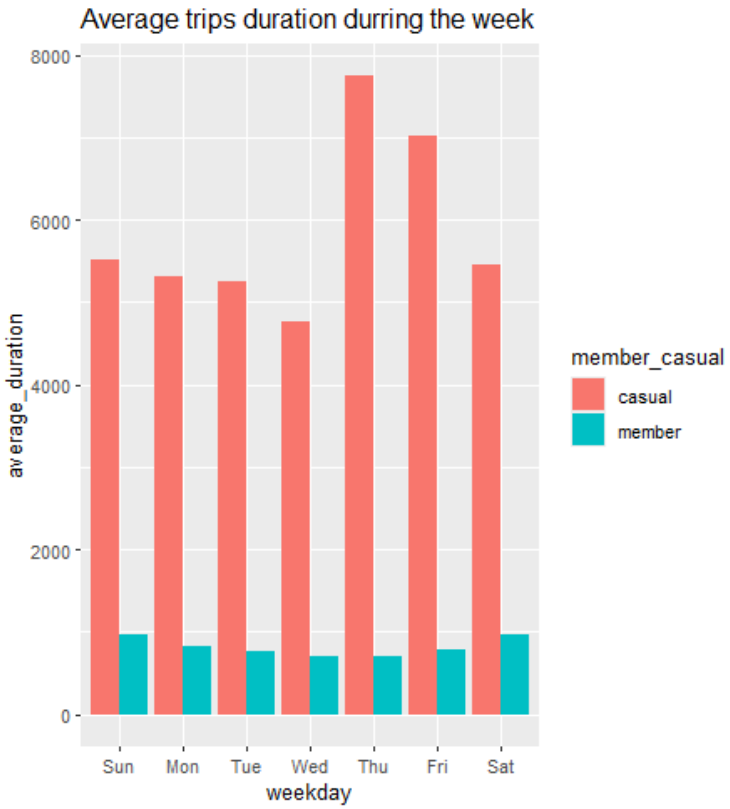
Output:



Let's create a visualization for average duration:

```r
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  ggtitle("Average trips duration durring the week") +
  geom_col(position = "dodge")
```

Output:



Insights generations

Member rides represent 93% of all trips, significantly higher than casual riders (7%). However, casual riders tend to take longer trips than members. More details below:

There are clear differences in ride frequency throughout the week between both groups. For members, the number of rides increases toward the middle of the week, reaching its highest point on Wednesday with 121,498 rides in average. Ride frequency then decreases during the weekend, reaching its lowest levels on Saturday with 59,166 rides more or less and Sunday with a mean of 59,952 rides.

In contrast, casual riders show the opposite pattern. Their highest ride activity occurs on Sundays (15,862 rides overall), followed by Saturdays (8,801 rides more or less). For the rest of the week, their ride counts generally range between 4,200 and 5,830.

Ride duration also varies between the two groups. Roughly, members ride for 794.15 seconds, while casual riders average 5,761 seconds. Members show a more stable pattern, with ride times consistently between

700 and 980 seconds. The longest overall duration for members occurs on Saturdays (973 seconds), while the shortest occurs on Thursdays (706 seconds).

Casual riders, on the other hand, register much longer ride duration. Their behavior can be split into two parts: 1. Sunday to Wednesday: average duration between 4,700 and 5,200 seconds 2. Thursday to Saturday: a noticeable increase, peaking on Thursday with 7,758 seconds, followed by a decrease toward Saturday (5,460 seconds)

As fun facts:

1. The maximum recorded ride time was 9,387,024 seconds, equivalent to 108.65 days.
2. It is noteworthy that for casual riders, the shortest average ride time occurs on Wednesday, immediately followed by the longest ride time on Thursday.
3. Overall, there is no clear relationship between the number of rides and the duration of rides. In other words, more rides do not necessarily mean longer distances.

Hypothesis: Based on the findings, members most likely use bikes for routine activities, as reflected in their shorter ride distances during the middle of the week. In contrast, many casual riders appear to use bikes for tourism or leisure, evidenced by their longer rides concentrated on weekends.

Next Steps

This analysis provides a general overview, but conducting deeper research is recommended to build a detailed customer profile. Future studies should explore the motivations behind each ride, age groups, price perceptions, and other behavioral insights that can support more precise marketing strategies.