Cyclistic: Google Data Analytics Professional Certificate Capstone

Marcelo Henrique Guidini Angeli

2025-07-31

Analysis Context

The Company

Founded in 2016, Cyclistic is a company that is specialized in sharing bicycles, tricycles, scooters and other similar means of transportation. Today the business fleet has approximately 5.800 vehicles and over 690 stations concentrated mainly in the Chicago region.

Revenue Model

Cyclistic has three rental options: single ride pass, daily pass and annual pass (Cyclistic member).

The Goal

The company already has a solid client base, but most of the clients use either the daily or single ride pass both of which are significantly less profitable to Cyclistic. In order to solve this the marketing manager requested an analysis to understand the difference between members and casual users. These findings will be used to guide the next ad campaign that aims specifically at converting the non members. The following report must answer the question: *1. What are the main differences in vehicle usage between members and casual riders?*

Analysis of Cyclistic's different users profile

Introduction

This report aims to document the analysis process between the main differences in casual (single ride/daily pass) and members(annual pass). Furthermore, it intends to give *insights* to the next company's marketing campaign based on such findings.

The Analysis

Data Preparation

The data used is pubic and was made available by a real company called Divvy. You can get the data here.Since the study is hypothetical the data will be assumed to be from Cyclistics.

The data used for the analysis was gathered from July 2024 to June 2025. The subsequent code imports the data and creates a single data frame containing the year around information. Two extras columns named "ride_year" and "ride_month" were added to help the analysis process.

```
return_csv <- function(file_name){
  return(read.csv(paste(getwd(),file_name, sep="/")));
}

return_col_names <- function(file_name){
  if(!length(file_name)){
    stop("No .csv file found in the current directory. Verify the current directory.");
  }
  return(colnames(return_csv(file_name)))
}

file_names <- sort(list.files(pattern = "\\.csv$"))

# Verifying whether the files can be merged

col_names <- return_col_names(file_names[1])
for (file_name in file_names) {
  if(!identical(col_names, return_col_names(file_name))){
    stop("Impossible to merge. The files have different columns.")
  }
}

# Creating a column two columns to store the trip month and the trip year, respectively.
# Assuming the files are names following the pattern: YYYYMM-divvy-tripdata.csv
dataframe_list <- list()
for(file_name in file_names){
  month <- as.numeric(substr(file_name, start=5, stop=6))
  year <- as.numeric(substr(file_name, start=1, stop=4))
  new_dataframe <- cbind(data.frame(return_csv(file_name),
                      travel_month=month, travel_year=year))
  dataframe_list[[length(dataframe_list) + 1]] <- new_dataframe
}
dataframe_merged <- do.call(rbind, dataframe_list)
```

After this another column with the trip elapsed time was created. See code below:

```
# Creating the column travel_time (seconds)

dataframe_merged <- dataframe_merged %>%
  mutate(travel_time = abs(
    as.numeric(difftime(
    as.POSIXct(ended_at), as.POSIXct(started_at), units = "secs"))))
```

Data Cleaning

Now the data will be cleaned. There are 43 trips with negative time - where the beginning time is later than the end time. These occurrences will be removed in order to maintain data consistency. This represents less then 0.000008% of the total trips.

```
removed_trips <- dataframe_merged %>%
  filter(started_at > ended_at)

dataframe_merged <-  dataframe_merged %>%
  filter(started_at < ended_at)
```

After removing such rows all possible duplicates were also removed.

```
# Removing duplicate rows
dataframe_merged <- dataframe_merged %>%
  distinct(ride_id, .keep_all = TRUE)
```

Other than that, the column "rideable_type" was verified since it could only contain the values: "electric_bike", "classic_bike" and "electric_scooter". The same process was applied to the "member_casual" column but with the words: "member" and "casual". No such error was found.

```
# Validating the columns "rideable_type" and "member_casual"
vehicle_types <- unique(dataframe_merged$rideable_type)
print(vehicle_types)
```

```
## [1] "electric_bike"    "classic_bike"     "electric_scooter"
```

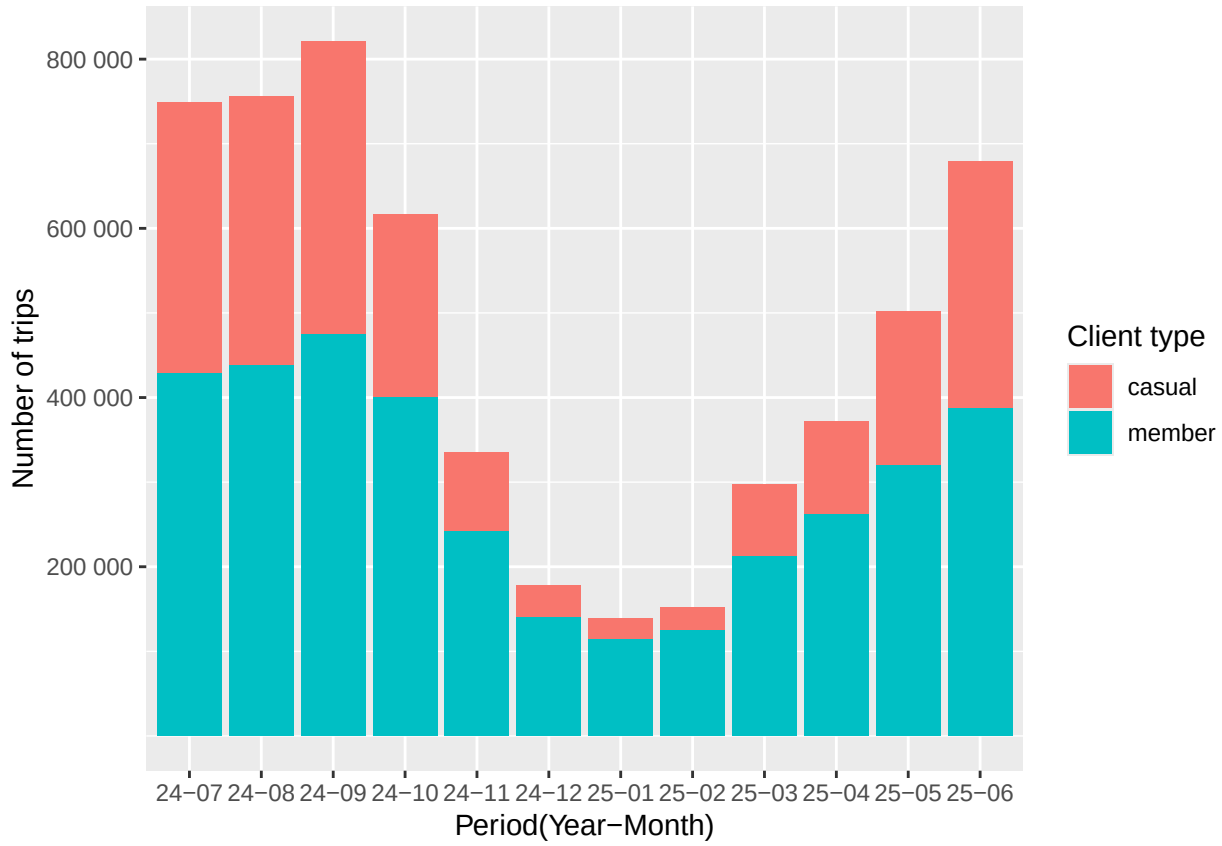
```
client_type <- unique(dataframe_merged$member_casual)
print(client_type)
```

```
## [1] "casual" "member"
```

Analysis

Initially the team verified the trips distribution along the year categorized by members and non members. In the chart it is possible to notice that the members are far more consistent over the year compared to the casuals. Both groups have a huge increase in activities during the months of June to October. The main difference lies in the months from December to February in which the members are responsible for almost 90% of the rides compared to the approximately 60% in the other months. This suggests that this scenario is correlated to the north-american seasons. This is based in the fact that both the representativeness and the number of rides of the non members increases dramatically during the summer period compared to what would be the winter.

```r
#Finding the number of trips by month
ggplot(dataframe_merged) +
  geom_bar(mapping = aes( x = ifelse(travel_month < 10,
        paste(substr(as.character(travel_year), 3, 4),
        travel_month, sep = "-0"),
        paste(substr(as.character(travel_year), 3, 4),
        travel_month, sep = "-")),
  fill = member_casual)) +
  scale_y_continuous(breaks = c(200000, 400000, 600000, 800000),
  labels = label_number()) +
  xlab("Period(Year-Month)") + ylab("Number of trips") +  labs(fill="Client type")
```

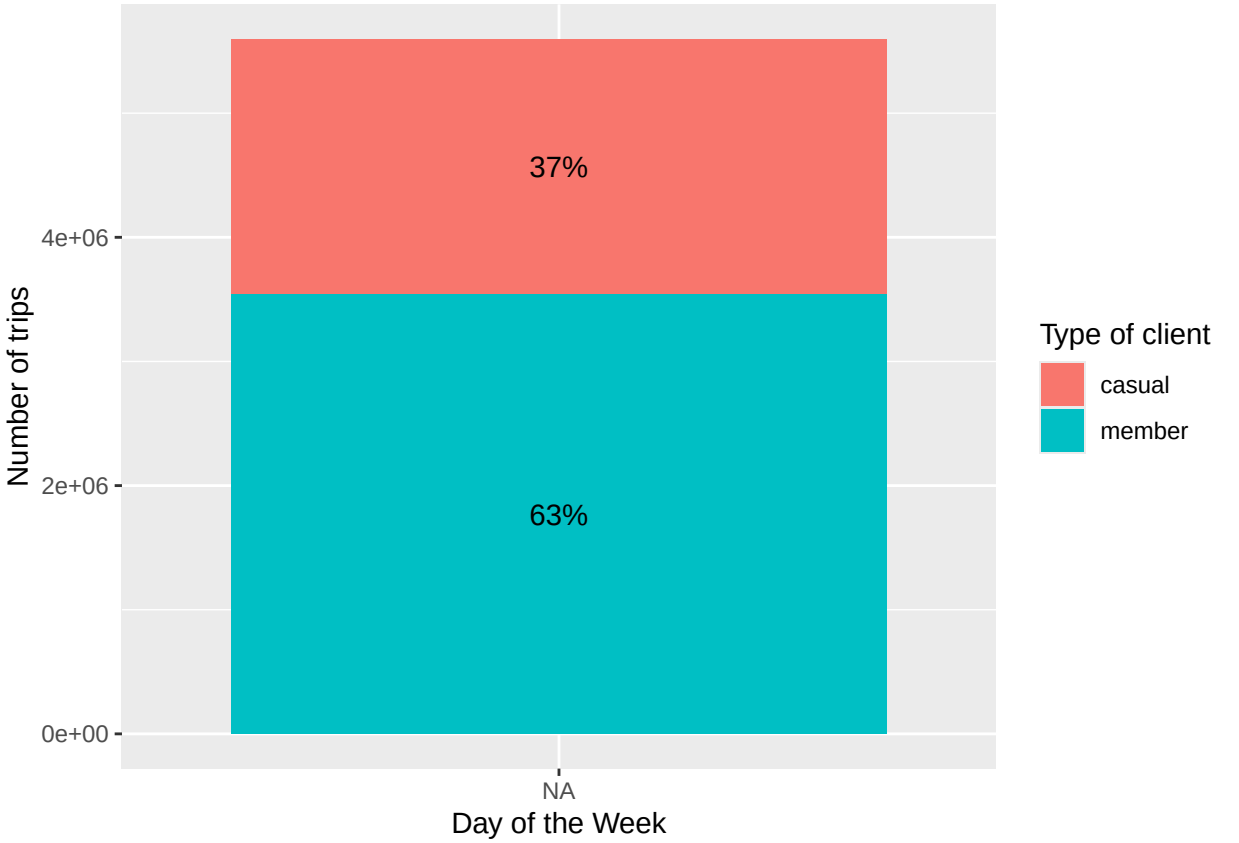


Later, it was decided to test which days of the week would have the biggest number of trips per client type. The graph show that the day with least amount of trips is the Sunday. The number of trips grows as the week progresses and reach a peak in Saturday. Again, the difference lies in the representativeness of the casual and members. The members are responsible for approximately 70% of the trips during the weekdays value that falls to about 50% during the weekends. This shows that the casuals tends to prefer Saturdays and Sundays, in comparison with the other week days. This situation is probably related to the way each group uses the vehicles. The members likely use them for their commute or day to day tasks. On the other hand, the casuals see the vehicles as a recreational activity.

```r
#Finding the number of trip categorized by day of the week and user type
trips_per_category_per_weekday <- dataframe_merged %>%
    group_by(member_casual = dataframe_merged$member_casual,
    weekday = factor(weekdays(as.Date(started_at)),
    levels = c("Sunday", "Monday", "Tuesday",
```

```
    "Wednesday", "Thursday", "Friday", "Saturday"))) %>%
  count() %>%
  group_by(weekday) %>%
  mutate(daily_percentage = n/sum(n))

ggplot(trips_per_category_per_weekday, aes(x = weekday, y = n, fill = member_casual)) +
  geom_col(position = "stack") +
  xlab("Day of the Week") + ylab("Number of trips") +  labs(fill="Type of client") +
  geom_text(aes(label = scales::percent(daily_percentage, accuracy = 1)),
            position = position_stack(vjust = 0.5))
```



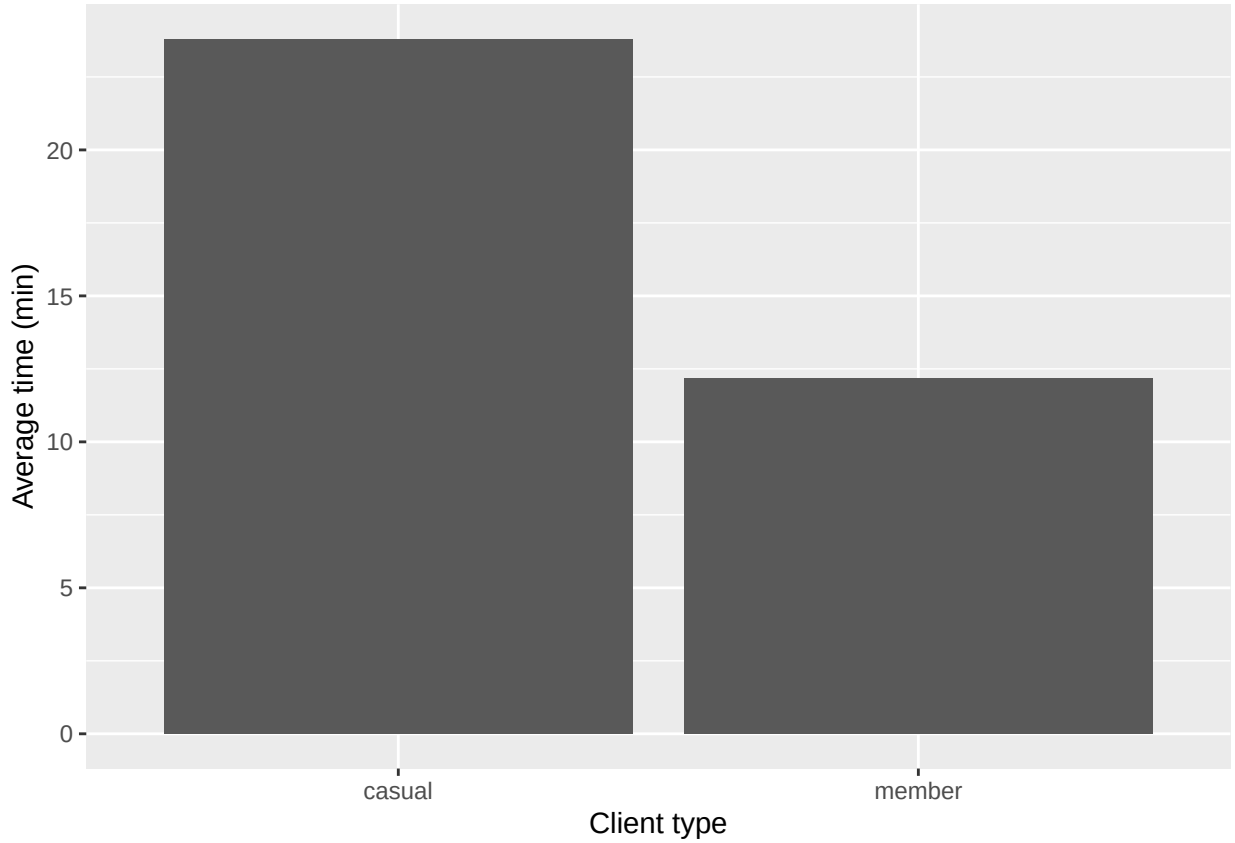
Regarding the average travel time the non members tend to have longer trips. Curiously, the average travel time for casual is about 24 minutes. This value is almost twice as big compared to the average 13 minutes of the members. See the chart bellow:

```
average_travel_time <- dataframe_merged %>%
  group_by(member_casual = dataframe_merged$member_casual) %>%
  summarise(average = mean(travel_time, na.rm=TRUE)/60)

print(average_travel_time)
```

```
## # A tibble: 2 x 2
##   member_casual average
##   <chr>           <dbl>
## 1 casual           23.8
## 2 member           12.2
```

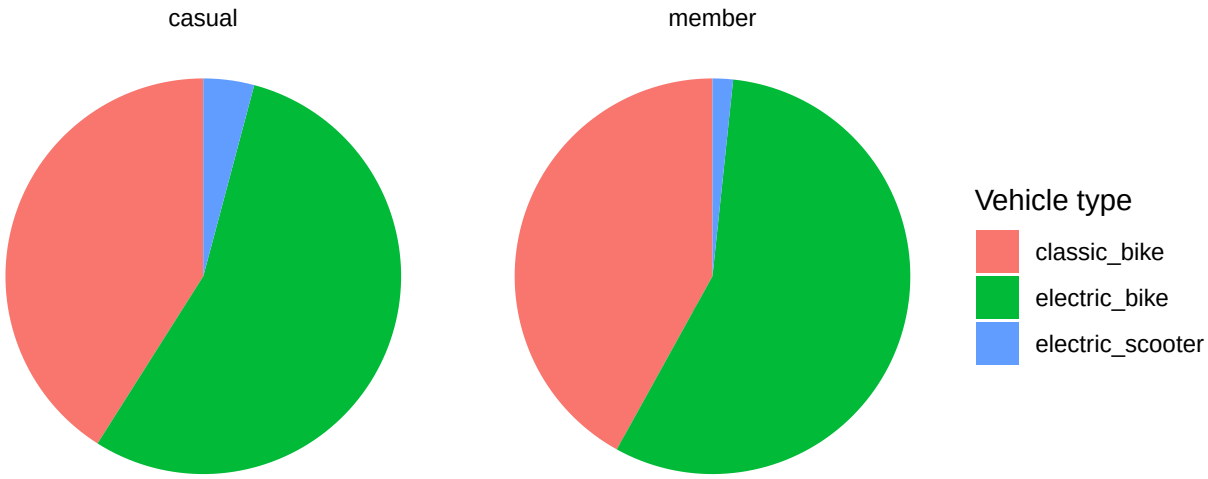
```r
ggplot(average_travel_time) + geom_col(mapping = aes(x = member_casual, y = average)) +
  xlab("Client type") + ylab("Average time (min)")
```



Finally, the most common vehicle was examined considering each client category. Both have shown a similar profile with great interest towards electric vehicles. This preference seems to be even stronger in the casuals.

```r
ride_counts_percent <- dataframe_merged %>%
  count(rideable_type, member_casual) %>%
  group_by(member_casual) %>%
  mutate(percent = n / sum(n))

ggplot(ride_counts_percent, aes(x = "", y = percent, fill = rideable_type)) +
  geom_col(width = 1) +
  coord_polar("y", start = 0) +
  facet_wrap(~member_casual) +
  xlab("Vehicle type") +
  ylab("Trip Proportion") +
  labs(fill = "Vehicle type") +
  theme_void()
```



Results

In general the clients prefer to rent their vehicles during the months of June to November.During these months the proportion of trips by members and non members is similar. However, during the off-season (December to March) the members become a huge majority. Based on this and aiming to convert the casuals into members, it is recommended that ads be displayed during the summer and spring.

Other than that the non members tend to use the bicycles during the weekends. In this context the ads should be exhibited during these days. Another idea would be to create promotions for clients that are riding in both days. Oddly enough the members have lower travel times on average.

Ultimately, both groups prefer the electric vehicles. Such trend is even stronger in the casual group. Therefore, it is recommend that the electric bicycle should be the most advertised vehicle in the campaign.