Cyclistic Bike-Share Q3-2023 Analysis Project

Hugo Villa

2023-11-8

#### A Google Data Analytics Professional Certificate Capstone Project.

This analysis is for case study 1 from the Google Data Analytics Certificate (Cyclistic). It’s originally based on the case study “‘Sophisticated, Clear, and Polished’: Divvy and Data Visualization” written by Kevin Hartman (found here:<https://artscience.blog/home/divvy-dataviz-case-study>). We will be using the Divvy dataset for the case study.

##### ***ASK***

The purpose of this script is to consolidate downloaded Q3-2023 Divvy data into a single dataframe and then conduct simple analysis to help answer the key question:

**“In what ways do member and casual riders use Divvy bikes differently?”**

**key Stakeholders**

* Cyclistic Executive Team
* Cyclistic Director of Marketing (Lily Moreno)
* Cyclistic Marketing Analytics Team

##### ***PREPARE***

**Setting up my environment**

Notes: Setting up my R environment by loading the following required packages:

library(tidyverse)  
library(janitor)  
library(tidyr)  
library(lubridate)   
library(ggplot2)

**Collect data**

Uploading 3 separate Cyclistic’s datasets (csv files) source: [Index of bucket "divvy-tripdata"](https://divvy-tripdata.s3.amazonaws.com/index.html)

ROCCC approach was implemented to determine the credibility of the data.

Reliable- the data has been verified and is accurate, represents all bike rides for the Q3 2023

Original – the data is made available by Motivate International Inc. under the following license: [Data License Agreement | Divvy Bikes](https://divvybikes.com/data-license-agreement)

Comprehensive- The datasets include all the bike ride details such as time, distance, station ID, membership type, etc.

Current- The Third Quarter of 2023 includes the latest data available at the time this report mas made.

Cited- The data is cited under the data license agreement.

jul23\_df <- read.csv("C:/Users/ronin/Desktop/Q3 2023 data/Divvy\_tripdata\_07\_2023.csv")  
aug23\_df <- read.csv("C:/Users/ronin/Desktop/Q3 2023 data/Divvy\_tripdata\_08\_2023.csv")  
sep23\_df <- read.csv("C:/Users/ronin/Desktop/Q3 2023 data/Divvy\_tripdata\_09\_2023.csv")

##### ***PROCESS***

**Wrangle and Cleaning the data frame**

As the data was being cleaned, it was noted that some of the data is missing specially in the “start and end station names and IDs” this could potentially hinder my ability to perform a thorough analysis; nonetheless, the data was clean and ready for its analysis.

Combining individual month datasets into a single data frame

##Combine 3 data frames into 1 data frame  
  
usage <- rbind(jul23\_df,aug23\_df, sep23\_df)

Remove empty spaces, nulls, duplicates

usage <- janitor::remove\_empty(usage,which = c("cols"))  
usage <- janitor::remove\_empty(usage,which = c("rows"))  
usage <- usage[!(usage$ride\_length <=0),]

Creating dataframe eliminating unnecessary columns

usage <- usage %>%   
 select(-c(start\_lat, end\_lat, start\_lng, end\_lng))

**Add data to prepare for analysis**

convert date/timestamp chr format to dttm format

usage$started\_at <- lubridate::mdy\_hm(usage$started\_at)   
usage$ended\_at <- lubridate::mdy\_hm(usage$ended\_at)

Create dataframe containing new columns (date, month, day, and year for each ride)

usage$date <- as.Date(usage$started\_at)  
usage$month <- format(as.Date(usage$date), "%m")  
usage$day <- format(as.Date(usage$date), "%d")  
usage$year <- format(as.Date(usage$date), "%Y")  
usage$day\_of\_week <- format(as.Date(usage$date), "%A")

Add a “ride\_length” calculation to usage (in seconds)

usage$ride\_length <- difftime(usage$ended\_at, usage$started\_at, unit = c("secs"))

Convert “ride\_length” from factor to numeric to run calculations on data

is.factor(usage$ride\_length)

## [1] FALSE

usage$ride\_length <- as.numeric(as.character(usage$ride\_length))  
is.numeric(usage$ride\_length)

## [1] TRUE

Remove “bad” data by creating a new version of the dataframe (v2)

usage\_v2 <- usage[!(usage$start\_station\_name == "HQ QR" | usage$ride\_length<0),]

##### ***ANALYZE***

**Descriptive analysis**

Descriptive analysis on ride\_length (all figures in seconds)

mean(usage\_v2$ride\_length) #straight average (total ride length / rides)

## [1] 1248.786

median(usage\_v2$ride\_length) #midpoint number in the ascending array of ride lengths

## [1] 660

max(usage\_v2$ride\_length) #longest ride

## [1] 5909340

min(usage\_v2$ride\_length) #shortest ride

## [1] 0

Compare member and casual riders

aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual, FUN = mean)

## usage\_v2$member\_casual usage\_v2$ride\_length  
## 1 casual 1876.1507  
## 2 member 813.0188

aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual, FUN = median)

## usage\_v2$member\_casual usage\_v2$ride\_length  
## 1 casual 780  
## 2 member 540

aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual, FUN = max)

## usage\_v2$member\_casual usage\_v2$ride\_length  
## 1 casual 5909340  
## 2 member 90000

aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual, FUN = min)

## usage\_v2$member\_casual usage\_v2$ride\_length  
## 1 casual 0  
## 2 member 0

Set days in order

usage\_v2$day\_of\_week <- ordered(usage\_v2$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

Average ride time by each day for member vs casual users

aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual + usage\_v2$day\_of\_week, FUN = mean)

## usage\_v2$member\_casual usage\_v2$day\_of\_week usage\_v2$ride\_length  
## 1 casual Sunday 2154.7946  
## 2 member Sunday 909.3473  
## 3 casual Monday 1872.8536  
## 4 member Monday 776.1469  
## 5 casual Tuesday 1714.9172  
## 6 member Tuesday 780.4218  
## 7 casual Wednesday 1638.3320  
## 8 member Wednesday 768.1304  
## 9 casual Thursday 1686.1597  
## 10 member Thursday 772.0713  
## 11 casual Friday 1803.0793  
## 12 member Friday 818.9641  
## 13 casual Saturday 2034.1997  
## 14 member Saturday 887.6893

Analyze ridership data by type and weekday

usage\_v2 %>%   
 mutate(day\_of\_week = wday(started\_at)) %>%   
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(number\_of\_rides = n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, day\_of\_week)

## # A tibble: 14 × 4  
## # Groups: member\_casual [2]  
## member\_casual day\_of\_week number\_of\_rides average\_duration  
## <chr> <dbl> <int> <dbl>  
## 1 casual 1 150203 2155.  
## 2 casual 2 101500 1873.  
## 3 casual 3 106220 1715.  
## 4 casual 4 99979 1638.  
## 5 casual 5 111784 1686.  
## 6 casual 6 139417 1803.  
## 7 casual 7 194939 2034.  
## 8 member 1 151270 909.  
## 9 member 2 168533 776.  
## 10 member 3 196309 780.  
## 11 member 4 197334 768.  
## 12 member 5 204805 772.  
## 13 member 6 194017 819.  
## 14 member 7 189264 888.

in order to visualize the busiest hour by rider type, an “hour” column has to be added

usage\_v2$hour <- hour(usage\_v2$started\_at)

**Analysis summary**

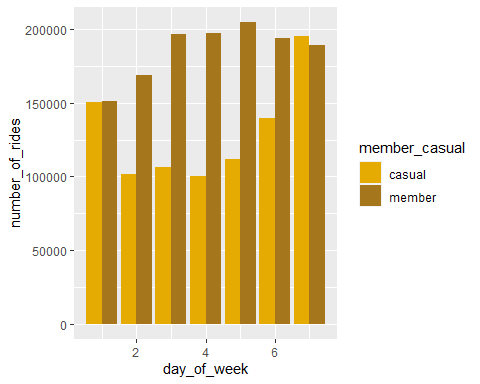
After analyzing the differences between annual members and casual riders, I have found that annual members have a longer ride duration and cover more distance (on average) especially during weekdays compared to casual riders. Casual riders’ average ride length is more than twice of that of members throughout the week, especially on Saturday and Sundays, this leads me to believe that casual riders use bike share more for leisure than for communing. It has also come to my attention that the most popular bike for both types of riders is the classic, rather than electric; at the same time, the busiest time for both riders is at 5 pm.

##### ***SHARE***

**Visualizing**

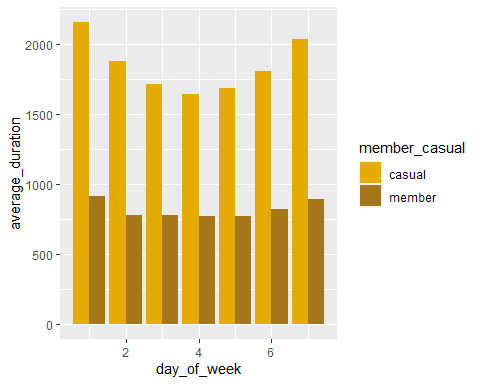
Number of rides by rider type and busiest weekday

my\_colors <- RColorBrewer::brewer.pal(8, "Dark2")[6:7]  
usage\_v2 %>%   
 mutate(day\_of\_week = wday(started\_at)) %>%   
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(number\_of\_rides = n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, day\_of\_week) %>%   
 ggplot(aes(x = day\_of\_week, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 scale\_fill\_manual(values = my\_colors)



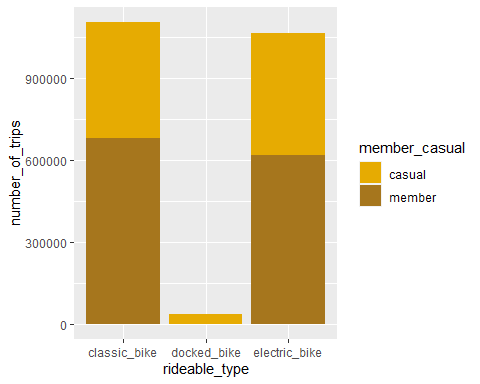
Average duration by rider type

my\_colors <- RColorBrewer::brewer.pal(8, "Dark2")[6:7]  
usage\_v2 %>%   
 mutate(day\_of\_week = wday(started\_at)) %>%   
 group\_by(member\_casual, day\_of\_week) %>%  
 summarise(number\_of\_rides = n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, day\_of\_week) %>%   
 ggplot(aes(x = day\_of\_week, y = average\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 scale\_fill\_manual(values = my\_colors)



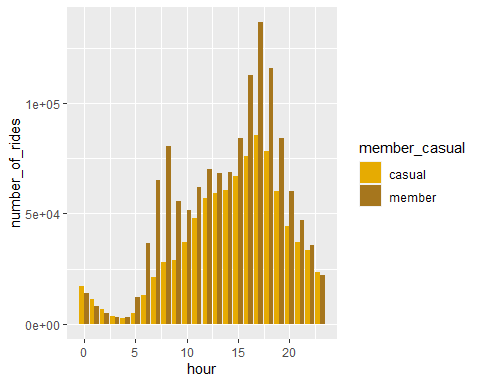
Total rides and preferred bicycle by rider type

my\_colors <- RColorBrewer::brewer.pal(8, "Dark2")[6:7]  
usage\_v2 %>%   
 group\_by(rideable\_type, member\_casual) %>%   
 summarize(number\_of\_trips = n())%>%  
 ggplot(aes(x = rideable\_type, y = number\_of\_trips, fill = member\_casual)) +  
 geom\_bar(stat = 'identity') +  
 scale\_y\_continuous(labels = function(x) format(x, scientific = FALSE)) +  
 scale\_fill\_manual(values = my\_colors)



busiest hour by rider type

my\_colors <- RColorBrewer::brewer.pal(8, "Dark2")[6:7]  
usage\_v2 %>%   
 mutate(day\_of\_week = wday(started\_at)) %>%   
 group\_by(member\_casual, hour) %>%  
 summarise(number\_of\_rides = n()  
 ,average\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, hour) %>%   
 ggplot(aes(x = hour, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 scale\_fill\_manual(values = my\_colors)



Export summary file for further analysis

counts <- aggregate(usage\_v2$ride\_length ~ usage\_v2$member\_casual + usage\_v2$day\_of\_week, FUN = mean)  
write.csv(counts, file = "C:\\Users\\ronin\\Desktop\\2023-Q3\\cyclistic\_Bike-Share\_ Q3-2023\_analysis.csv", row.names = FALSE)

##### ***ACT***

Conclusion

To answer Lily Moreno’s assigned question: In what ways do member and casual riders use Divvy bikes differently? I have come to the following conclusion.

This report confirms that in the third quarter of 2023 member riders have a longer ride duration as shown in bar charts, while casual riders use the bikes more for leisure than for commuting. Given its recent strategic initiatives, I expect Cyclistic Bike-Share to strengthen its market share, they offer excellent products and services at accessible rates. The review of the latest Q3 2023 report, as well as the analysis of the data reveals that Cyclistic Bike-Share has continued to prosper and expand immensely as a result of its successes. Furthermore, by looking at the “busiest hour by rider type” bar chart, we can see that member and casual riders use the bikes throughout the whole week regardless of the pick hours of usage.

Complete data under the “start station and end station, names and IDs” could enhance this report as this information is critical to evaluate the feasibility to establish new bike stations or remove those that are hardly used. Further analysis would have to be conducted. These findings can help Lily Moreno implement different marketing strategies and initiatives to promote and convert casual riders into annual riders.

1. Include relevant details

Give a clear picture of your thought process by adding details about the decisions you made and the steps you took to complete the project. It’s important to show your work!

1. 2

Lead with your key findings

When summarizing your project, focus on the results that show the most positive impact of your work. What benefits do your recommendations provide?

1. 3

Keep it succinct and professional

Keep your responses concise and to the point. Avoid slang or jargon that could make it difficult for the employer to understand what your solution has to offer.

R programming tool was utilized to analyze the raw data. The solution provided can positively impact the corporation's market share.

[Cyclistic Bike-Share Q3-2023 Project with R | Kaggle](https://www.kaggle.com/code/hugovillam/cyclistic-bike-share-q3-2023-project-with-r)

[Cyclistic Bike Share : Case study with R | Kaggle](https://www.kaggle.com/code/sayantanbagchi/cyclistic-bike-share-case-study-with-r)