casestudy1v1

Kiet

2025-03-17

## Introduction

Welcome to the Cyclistic bike-share analysis case study! In this case study, I work for a fictional company, Cyclistic, along with some key team members. In order to answer the business questions, follow the steps of the data analysis process: Ask, Prepare, Process, Analyze, Share, and Act. Along the way, the Case Study Roadmap tables — including guiding questions and key tasks — will help me stay on the right path.

## Scenario

I’m a junior data analyst working on the marketing analyst team at Cyclistic, a bike-share company in Chicago.

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships.

Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic’s finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a solid opportunity to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs

## 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 the bikes to commute to work each day.

● Lily Moreno: The director of marketing and my 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 analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. I joined this team six months ago and have been busy learning about Cyclistic’s mission and business goals—as well as how I, 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 Phase

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics.We are interested in analyzing the Cyclistic historical bike trip data to identify trends. She believes the company’s future success depends on maximizing the number of annual memberships. Therefore, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my recommendations, so they must be backed up with compelling data insights and professional data visualizations.

## Data Preperation

Cyclistic’s historical data to analyse and identify trends is downloaded and stored appropriately in a folder.The data source here being the Motivate International Inc. The recommended cyclistic trip data to be downloaded is that of the previous 12 months which spans from January 2024 to December 2024. The data for each month stored in the folder are then named according to the necessary naming conventions with the usage of underscores, capital letters and numbers. An example being the data for the month of April 2021 stored as “202402-divvy-tripdata”. There are no issues with the data’s credibility and the data is good data and that is because it is reliable: comes from a good data source and under a valid license, the data is original: validated with the company’s original source, the data is comprehensive: contains all critical information we need, the data is current: downloaded the previous 12 months and so it is very useful at this particular time, the data is cited: from a source and a reliable one at that.

## Data Processing

Clean data is the launchpad for any strong analysis. The integrity of data is very important and so this is always in a direct relationship with clean data, one of the famous rules of a data’s integrity is it’s ability to not be compromised. Thus getting rid of duplicates, blanks, irrelevant data that do not align with objectives, inconsistent data, outdated data, inaccurate/incorrect data helps us ensure data integrity- assurance of, data accuracy and consistency over its entire life-cycle. This data must be unique , values must match prescribed patterns, the data must be consistent, the data must be relevant to determine our course of action to achieve our intended objective. This phase ensures that data meets all these requirements.

The preferred tool for data processing for this analysis is R and RStudio. This is often considered the best place for this task and that is because it handles large complex data which excel and SQL would not be able to handle, it’s convenient library of packages which offer a helpful combination of code and reusable R function among others. The processing of data using RStudio in R begins by installing the necessary packages. These packages include tidyverse-a collection of packages in R with a common design philosophy for data manipulation and exploration and visaulizations, lubridate-an R package that makes it easier to work with dates and times, dplyr- an R package that contains a set of functions (or “verbs”) that perform common data manipulation operations such as filtering for rows, selecting specific columns etc,ggplot2- an R package used for visualizations and readr- an R package for data importing. It’s important to know that all these pacakges are core packages under the tidyverse

install.packages("tidyverse")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

install.packages("lubridate")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'lubridate' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

install.packages("dplyr")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'dplyr' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

install.packages("ggplot2")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

install.packages("readr")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'readr' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

install.packages("hydroTSM")

## Installing package into 'C:/Users/Admin/AppData/Local/R/win-library/4.4'  
## (as 'lib' is unspecified)

## package 'hydroTSM' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Admin\AppData\Local\Temp\RtmpWUkwfj\downloaded\_packages

The next step of the data processing phase is to load the various packages using the library function

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(lubridate)  
library(dplyr)  
library(ggplot2)  
library(readr)  
library(hydroTSM)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric  
##   
##   
## Attaching package: 'hydroTSM'  
##   
## The following object is masked from 'package:tidyr':  
##   
## extract

I then imported the data into RStduio using the read\_csv function contained in the readr package. This function would be used for each month’s trip csv data

data\_24\_01 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202401-divvy-tripdata/202401-divvy-tripdata.csv")  
data\_24\_02 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202402-divvy-tripdata/202402-divvy-tripdata.csv")  
data\_24\_03 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202403-divvy-tripdata/202403-divvy-tripdata.csv")  
data\_24\_04 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202404-divvy-tripdata/202404-divvy-tripdata.csv")  
data\_24\_05 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202405-divvy-tripdata/202405-divvy-tripdata.csv")  
data\_24\_06 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202406-divvy-tripdata/202406-divvy-tripdata.csv")  
data\_24\_07 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202407-divvy-tripdata/202407-divvy-tripdata.csv")  
data\_24\_08 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202408-divvy-tripdata/202408-divvy-tripdata.csv")  
data\_24\_09 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202409-divvy-tripdata/202409-divvy-tripdata.csv")  
data\_24\_10 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202410-divvy-tripdata/202410-divvy-tripdata.csv")  
data\_24\_11 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202411-divvy-tripdata/202411-divvy-tripdata.csv")  
data\_24\_12 <- read.csv("D:/Hồ sơ/GG Data Analyst/Data/202412-divvy-tripdata/202412-divvy-tripdata.csv")

Proceeded to merge every month’s data into a larger data set using the rbind function. merging data makes analysis easier

data\_12\_months <- rbind(data\_24\_01,data\_24\_02,data\_24\_03,data\_24\_04,data\_24\_05,data\_24\_06,data\_24\_07,data\_24\_08,data\_24\_09,data\_24\_10,data\_24\_11,data\_24\_12)

The ride length between the ended\_time and started\_time is calculated. Getting the ride length is important because it gives s fresh insights for our analysis

data\_12\_months$ride\_length <- difftime(data\_12\_months$ended\_at, data\_12\_months$started\_at, units = "mins")

Create a new version after update

data\_12\_months\_v2 <- data\_12\_months

I proceeded by extracting the day, month, year and day\_of\_week from the started\_at column which is in the POSIXct format. This helps in deriving more insights from this data set

data\_12\_months\_v2$date <- as.Date(data\_12\_months\_v2$started\_at)   
data\_12\_months\_v2$month <- format(as.Date(data\_12\_months\_v2$date), "%m")  
data\_12\_months\_v2$day <- format(as.Date(data\_12\_months\_v2$date), "%d")  
data\_12\_months\_v2$year <- format(as.Date(data\_12\_months\_v2$date), "%Y")  
data\_12\_months\_v2$day\_of\_week <- format(as.Date(data\_12\_months\_v2$date), "%A")

I now have a large data set with the necessary columns for analysis but first the data has to be cleaned; get rid of null values, get rid of duplicates, get rid of unnecessary information/columns.

data\_12\_months\_v2 <- distinct(data\_12\_months\_v2) # get ride of duplicates  
data\_12\_months\_v2 <- data\_12\_months\_v2 %>%   
 drop\_na() # get rid of null values

Remove ride\_lengths of values less than a minute and more than an 1440 minutes/ an hour

data\_12\_months\_v2 <- data\_12\_months\_v2 %>% filter(ride\_length > 1) %>% filter(ride\_length < 1440) # get rid of ride\_lengths of more than a day (1440 minutes)

Create a new version dataset which select columns that would be useful for this analysis

data\_12\_months\_v3 <- data\_12\_months\_v2%>%  
 select(ride\_id, rideable\_type, started\_at, ended\_at, start\_station\_name,end\_station\_name, member\_casual, month, year, day\_of\_week, ride\_length)

unique(data\_12\_months\_v3$rideable\_type) # Check rideable types

## [1] "electric\_bike" "classic\_bike" "electric\_scooter"

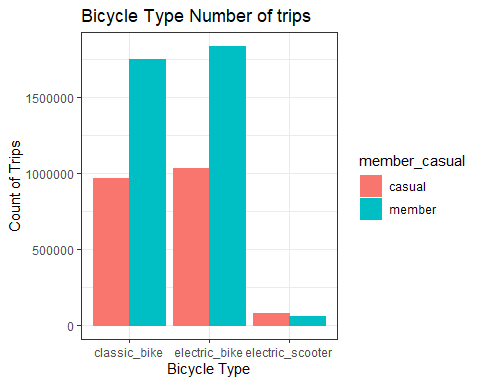
unique(data\_12\_months\_v3$member\_casual) # check member\_casual

## [1] "member" "casual"

## Analyse Phase

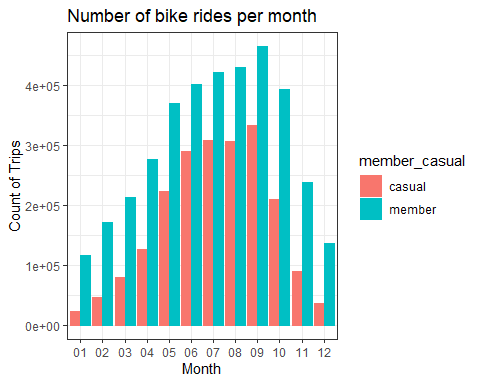
The next phase after the process phase(cleaning phase) is the analysis phase. Here calculations would take place which we would then use for data insights.

data\_12\_months\_v3 %>%  
 group\_by(rideable\_type, member\_casual) %>%  
 dplyr::summarize(count\_trips = n(), .groups = 'drop') %>% # Use .groups = 'drop' to ungroup the result  
 ggplot(aes(x = rideable\_type, y = count\_trips, fill = member\_casual, color = member\_casual)) +  
 geom\_bar(stat = 'identity', position = 'dodge') +  
 theme\_bw() +  
 labs(title = "Bicycle Type Number of trips", x = "Bicycle Type", y = "Count of Trips")



library("dplyr")  
data\_12\_months\_v3$month <- ordered(data\_12\_months\_v3$month, levels=c("01", "02", "03", "04", "05", "06", "07","08","09","10","11","12"))

data\_12\_months\_v3 %>%  
 group\_by(member\_casual,month) %>%  
 dplyr::summarize(count\_trips = n(), .groups = 'drop') %>%   
 ggplot(aes(x= month, y=count\_trips, fill=member\_casual, color=member\_casual)) +  
 geom\_bar(stat='identity', position = 'dodge') +  
 theme\_bw() +  
 labs(title ="Number of bike rides per month", x = "Month", y = "Count of Trips")



data\_12\_months\_v3 %>%  
 group\_by(member\_casual,start\_station\_name) %>%  
 dplyr::summarise(number\_of\_ride = n(), .groups = 'drop') %>%  
 filter(start\_station\_name != "", "casual"== member\_casual) %>%  
 arrange(-number\_of\_ride) %>%  
 head(n=30) %>%  
 select(-member\_casual)

## # A tibble: 30 × 2  
## start\_station\_name number\_of\_ride  
## <chr> <int>  
## 1 Streeter Dr & Grand Ave 50264  
## 2 DuSable Lake Shore Dr & Monroe St 33484  
## 3 Michigan Ave & Oak St 24738  
## 4 DuSable Lake Shore Dr & North Blvd 22667  
## 5 Millennium Park 22177  
## 6 Shedd Aquarium 20747  
## 7 Dusable Harbor 18173  
## 8 Theater on the Lake 16562  
## 9 Michigan Ave & 8th St 13322  
## 10 Adler Planetarium 12670  
## # ℹ 20 more rows

data\_12\_months\_v3 %>%  
 group\_by(member\_casual,start\_station\_name) %>%  
 dplyr::summarise(number\_of\_ride = n(), .groups = 'drop') %>%  
 filter(start\_station\_name != "", "member" == member\_casual) %>%  
 arrange(-number\_of\_ride) %>%  
 head(n=30) %>%  
 select(-member\_casual)

## # A tibble: 30 × 2  
## start\_station\_name number\_of\_ride  
## <chr> <int>  
## 1 Kingsbury St & Kinzie St 29212  
## 2 Clinton St & Washington Blvd 27434  
## 3 Clinton St & Madison St 24562  
## 4 Clark St & Elm St 24427  
## 5 Wells St & Concord Ln 20488  
## 6 Wells St & Elm St 20215  
## 7 Clinton St & Jackson Blvd 20046  
## 8 Dearborn St & Erie St 19196  
## 9 Canal St & Madison St 18901  
## 10 University Ave & 57th St 18808  
## # ℹ 20 more rows

## Data Visualization

Below is a dashboard that perfectly visualizes the data for which we can draw insights

## Data Insights

Below are the data insights retrieved from the visualization of the Cyclistic data

From the number of rides per season chart it is observed that members take more rides than casual over every season except during the summer period and that’s because for casual riders the appeal of the summer is forever enticing for a stroll on the bike as compared to every other season. The members however dominate other seasons because it’s very clear that the rides are their source of transportation to and from work. This is the same reason why in the number of rides per weekday line chart, we see the members dominate over every weekday and casuals over the weekend.

As earlier seen, although the casual dominates number of rides only over the weekends the last line chart shows a total domination of average ride lengths for casual over members across every weekday. This means the casual members have a greater ride length with a fewer rides than the members who have more rides with lesser ride length. This proves that for members, their rides are a routine one comprising of similar ride lengths but for casuals we have rides of different increased durations.

### Recommendations

A digital ad named “Passion Into Members” should be launched where we convince the casual riders to switch to members. How do we do this? For the members, its just a routine but for casaul riders it’s something they are passionate about and hence we put out an ad that speaks to their passion, something they can relate to.

We know of the decreasing number of rides during the winter, spring and autumn therefore we should provide for them a discount for their membership conversion, this shows how much we care and how much we know of the obsatcles they face in the form of weather.

Still on luring them, we can organise bicycle ride competitions for them every summer period. This lures other passionate bicylcle riders to enlist our services and thus be members.