Cyclistic Case Study

Carolyn Chell

06/10/2021

## Case Study

This case study has been completed as a capstone project as part of the Google Data Analytics Certificate. It is based on a fictional company and used public data made available by Motivate International Inc.

## Introduction

Cyclistic is a bike sharing company that has 5,800 bicycles which are geo-tracked and locked into a network of 692 stations across Chicago. The company want to analyze the usage of their bikes in order to gain insights into how casual riders and members use bikes differently. Then, using this information the CEO would like high level recommendations for designing marketing strategies to convert casual riders into annual members to increase sales.

## Key Objectives:

1. Identify the business task

Cyclistic’s finance analysts have concluded that annual members are much more profitable than casual riders. With this in mind, the business objective is to increase sales by converting casual riders to members. My task as a junior data analyst is:

***Identify how annual members and casual riders use Cyclistic bikes differently.***

These insights will be used by the marketing team to develop a campaign aimed at converting casual riders to members by purchasing an annual membership.

The business task is to produce a report which outlines the following:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of the analysis
5. Supporting visualizations of key findings
6. Top 3 recommendations based on the analysis
   1. Consider the key stakeholders

Cyclistic - A bike share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day

Lily Moreno - Director of Marketing and my manager

Cyclistic Marketing Analytics Team - A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy

Cyclistic Executive Team – they will decide whether to approve the recommended marketing program

### *Hypothesis:*

* 1. Mean trip duration will be longer for casual riders than for members. Members are likely to use the bike for commuting to work, whereas casual riders are more likely to cycle for pleasure at a leisurely pace.
  2. Usage will spike for casual riders over the weekend.
  3. Usage will increase during the warmer months
  4. Start stations in the tourist areas will be popular for casual riders as tourists want to see the city by bike.

I would also expect members to cycle more frequently than casual riders however this can not be analyzed due to data-privacy issues prohibiting the collection of riders’ personally identifiable information.

## 

## Preparing the data

The data set is sourced from public data compiled by Motivate International Inc. 12 months of data has been used which is approximately 4 million records. Given the quantity of data, I have decided to use R as the main analytics tool.

Data has needed to be cleaned due to a number of duplicate records, missing data and invalid data.

A link to the data can be found here:

[http://divvy-tripdata.s3.amazonaws.com/index.html](about:blank)

### Packages to be installed

install.packages("tidyverse")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.1'  
## (as 'lib' is unspecified)

install.packages("janitor")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.1'  
## (as 'lib' is unspecified)

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(knitr)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(janitor)

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(ggplot2)

## Processing the data

### Upload 12 months of data from Motivate International Inc.

setwd('/cloud/project/data/')  
df\_2021\_08 <- read.csv("202108-divvy-tripdata.csv")  
df\_2021\_07 <- read.csv("202107-divvy-tripdata.csv")  
df\_2021\_06 <- read.csv("202106-divvy-tripdata.csv")  
df\_2021\_05 <- read.csv("202105-divvy-tripdata.csv")  
df\_2021\_04 <- read.csv("202104-divvy-tripdata.csv")  
df\_2021\_03 <- read.csv("202103-divvy-tripdata.csv")  
df\_2021\_02 <- read.csv("202102-divvy-tripdata.csv")  
df\_2021\_01 <- read.csv("202101-divvy-tripdata.csv")  
df\_2020\_12 <- read.csv("202012-divvy-tripdata.csv")  
df\_2020\_11 <- read.csv("202011-divvy-tripdata.csv")  
df\_2020\_10 <- read.csv("202010-divvy-tripdata.csv")  
df\_2020\_09 <- read.csv("202009-divvy-tripdata.csv")

### Combine the 12 data frames into 1 dataframe

bike\_rides <- rbind(df\_2021\_08,df\_2021\_07,df\_2021\_06, df\_2021\_05, df\_2021\_04, df\_2021\_03, df\_2021\_02, df\_2021\_01, df\_2020\_12, df\_2020\_11, df\_2020\_10, df\_2020\_09)

### Add ride length column for calculations

bike\_rides$ride\_length <- difftime(bike\_rides$ended\_at, bike\_rides$started\_at, units = c("mins"))

### Add columns that list the date, month, day and year for each ride

bike\_rides\_v2 <- bike\_rides  
bike\_rides\_v2$date <- as.Date(bike\_rides\_v2$started\_at)  
bike\_rides\_v2$month <- format(as.Date(bike\_rides\_v2$date), "%m")  
bike\_rides\_v2$day <- format(as.Date(bike\_rides\_v2$started\_at), "%d")  
bike\_rides\_v2$year <- format(as.Date(bike\_rides\_v2$started\_at), "%y")  
bike\_rides\_v2$day\_of\_week <- format(as.Date(bike\_rides\_v2$started\_at), "%A")

### Delete bad data where ride length is < 1 minute

bike\_rides\_v2 <- bike\_rides\_v2[!(bike\_rides\_v2$ride\_length) <=1,]

### Remove empty rows and columns

bike\_rides\_v2 <- bike\_rides\_v2[!(bike\_rides\_v2$start\_station\_name == ""), ]  
  
bike\_rides\_v2 <- bike\_rides\_v2[!(bike\_rides\_v2$end\_station\_name == ""), ]

### Remove the variables not needed for analysis: start station id, end station id, start\_lat, start\_lng, end\_lat, end\_lng

bike\_rides\_v2 <- bike\_rides\_v2 %>%   
 select(-c(start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, end\_lng))

## 

## Analyzing and visualizing the data

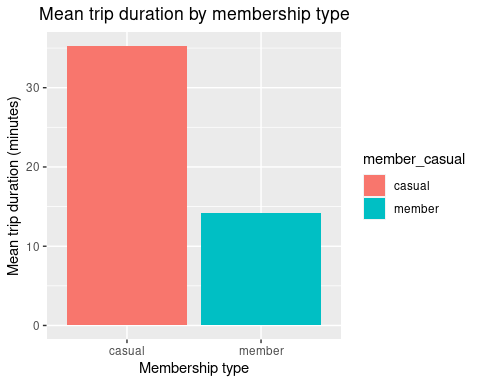
### Compare mean trip duration for casual riders and members

bike\_rides\_v2 %>%  
group\_by(member\_casual) %>%  
summarize(ride\_length\_min = mean(ride\_length))

## # A tibble: 2 × 2  
## member\_casual ride\_length\_min  
## <chr> <drtn>   
## 1 casual 35.30994 mins   
## 2 member 14.17794 mins

bike\_rides\_v2 %>%  
 na.omit() %>%  
 group\_by(member\_casual) %>%  
 summarize(mean\_trip\_duration = mean(ride\_length)) %>%  
ggplot(aes(x = member\_casual, y = mean\_trip\_duration, fill = member\_casual)) +  
 geom\_col() +  
 labs(title = "Mean trip duration by membership type", x = "Membership type",   
 y = "Mean trip duration (minutes)") +  
 theme(plot.title = element\_text(hjust = 0.5))

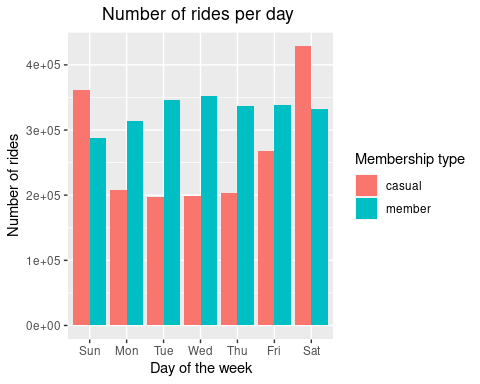
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.



### Plot the number of rides per day to identify which days are most popular

bike\_rides\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 summarise(number\_of\_rides = n()) %>%   
 arrange(member\_casual, weekday) %>%   
 ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +  
 geom\_col(position = "dodge")+  
 labs(title = "Number of rides per day ",x="Day of the week",y="Number of rides", fill="Membership type") +  
 theme(legend.position="right") +  
 theme(plot.title = element\_text(hjust = 0.5))

## `summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.



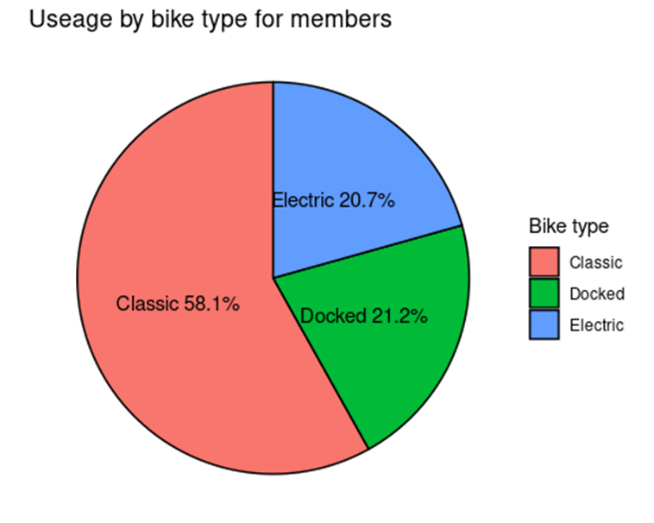
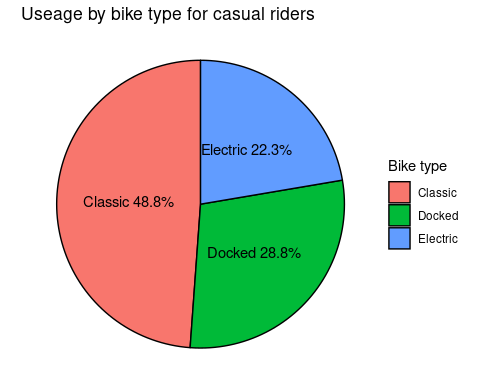
**Bike type usage for casual riders and members**

bike\_rides\_v2 %>%  
 group\_by(rideable\_type, member\_casual) %>%  
 count(rideable\_type)

## # A tibble: 6 × 3  
## # Groups: rideable\_type, member\_casual [6]  
## rideable\_type member\_casual n  
## <chr> <chr> <int>  
## 1 classic\_bike casual 911402  
## 2 classic\_bike member 1340584  
## 3 docked\_bike casual 537848  
## 4 docked\_bike member 490027  
## 5 electric\_bike casual 416634  
## 6 electric\_bike member 476535

bikes\_pct <- data.frame(value = c(911402, 537848, 416634),  
 Type = c("Classic", "Docked", "Electric")) %>%  
 # factor levels need to be the opposite order of the cumulative sum of the values  
 mutate(Type = factor(Type, levels = c("Classic", "Docked", "Electric")),  
 cumulative = cumsum(value),  
 midpoint = cumulative - value / 2,  
 label = paste0(Type, " ", round(value / sum(value) \* 100, 1), "%"))  
  
bikes\_pct %>%   
 ggplot(aes(x=1, y=value, fill= Type)) +  
 geom\_col(colour = "black") +  
 geom\_text(aes(label = label), position = position\_stack(vjust = 0.50))+  
 coord\_polar(theta = "y") +   
 ggtitle("Useage by bike type for casual riders") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 guides(fill = guide\_legend(title = "Bike type")) +  
 theme\_void()

bikes\_pct\_members <- data.frame(value = c(1340584, 490027, 476535),  
 Type = c("Classic", "Docked", "Electric")) %>%  
 # factor levels need to be the opposite order of the cumulative sum of the values  
 mutate(Type = factor(Type, levels = c("Classic", "Docked", "Electric")),  
 cumulative = cumsum(value),  
 midpoint = cumulative - value / 2,  
 label = paste0(Type, " ", round(value / sum(value) \* 100, 1), "%"))  
  
bikes\_pct\_members %>%   
 ggplot(aes(x=1, y=value, fill= Type)) +  
 geom\_col(colour = "black") +  
 geom\_text(aes(label = label), position = position\_stack(vjust = 0.50))+  
 coord\_polar(theta = "y") +   
 ggtitle("Useage by bike type for members") +  
 theme(plot.title = element\_text(hjust = .75)) +  
 guides(fill = guide\_legend(title = "Bike type")) +  
 theme\_void()



### Create column for ride length (minutes) and change data type from difftime to numeric to enable graphs

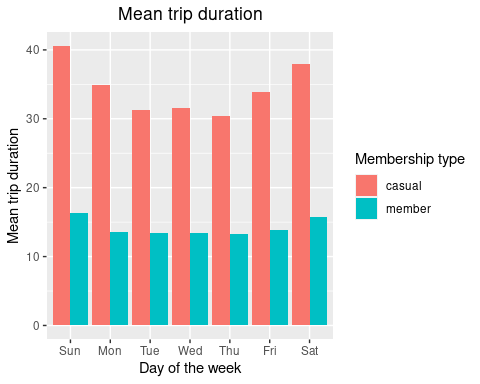
mean\_time <- bike\_rides\_v2 %>% summarize(mean\_time = mean(ride\_length))  
  
bike\_rides\_v2$ride\_length <- as.numeric(as.character(bike\_rides\_v2$ride\_length))  
is.numeric(bike\_rides\_v2$ride\_length)

## [1] TRUE

### Trip duration (minutes) by day of the week by membership

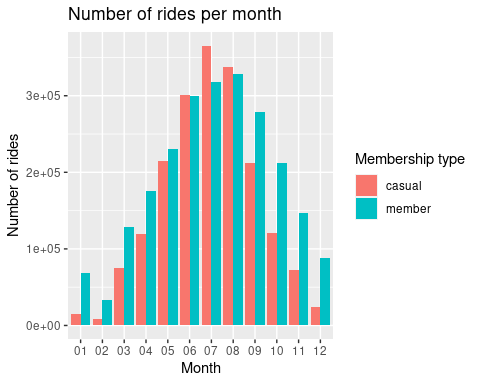
bike\_rides\_v2 %>%   
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 summarise(mean\_ride\_duration = mean(ride\_length)) %>%   
 arrange(member\_casual, weekday) %>%   
 ggplot(aes(x = weekday, y = mean\_ride\_duration, fill = member\_casual)) +  
 geom\_col(position = "dodge") +  
 labs(title = "Mean trip duration", x="Day of the week",y= "Mean trip duration", fill="Membership type") +  
 theme(legend.position="right") +  
 theme(plot.title = element\_text(hjust = 0.5))

## `summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

 ### Number of rides per month for casual riders and members

bike\_rides\_v2 %>%  
group\_by(member\_casual, month) %>%  
summarize(num\_rides = n()) %>%  
arrange(member\_casual, month) %>%  
ggplot(aes(x = month, y = num\_rides, fill = member\_casual)) +  
geom\_col(position = "dodge") +  
labs(title = "Number of rides per month", x = "Month", y = "Number of rides", colour = "Membership Type") +  
guides(fill = guide\_legend(title = "Membership type"))

## `summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

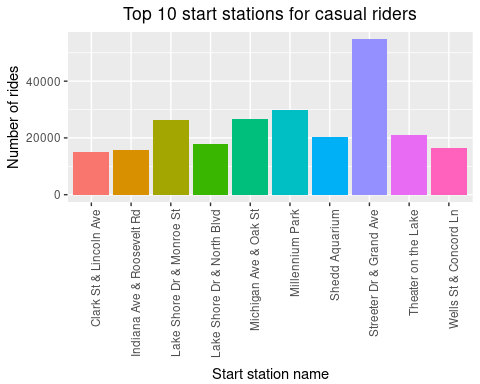


### Top 10 start stations for casual riders

bike\_rides\_v2 %>%  
 group\_by(start\_station\_name, member\_casual) %>%  
 filter(member\_casual == "casual") %>%  
 summarize(number\_of\_rides = n()) %>%  
 arrange(desc(number\_of\_rides)) %>%  
 head(10) %>%  
 ggplot(aes(x = start\_station\_name, y = number\_of\_rides, fill = start\_station\_name)) +  
 geom\_col() +  
 labs(title = "Top 10 start stations for casual riders", x = "Start station name", y = "Number of rides") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 guides(fill=FALSE) +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme(axis.title.x = element\_text(vjust=-0.5))

## `summarise()` has grouped output by 'start\_station\_name'. You can override using the `.groups` argument.

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =  
## "none")` instead.



## Key Findings

1. Mean trip length is significantly higher for casual riders (35 mins) than for members (14 mins). This is likely to be due to casual riders cycling at a leisurely pace taking breaks, versus members who may be commuting to work.
2. For casual riders the weekend is the most popular day for trips and the least popular days are Tuesday to Thursday. The variance between days of the week fluctuates significantly more for casual riders than for members. The trip count for members is relatively consistent, irrespective of the day of the week, with the slight exception for Sunday - traditionally a day of rest for week day workers.
3. Similarly, the mean trip length for casual riders increases for Saturday and Sundays.
4. The most popular bike type for casual riders and members is the classic, followed by docked then electric.
5. The most popular start stations for casual riders are Streeter Dr/ Grand Ave, Millennium Park and Michigan Ave/ Oak Street.
6. Demand for Cyclistic bikes is seasonal with demand increasing during Summer and decreasing during Winter. This is the case for both casual riders and members. We can say there is a correlation between usage and seasons for casual riders and members.

## Recommendations

Based on the above findings my top 3 recommendations are:

* + 1. Streeter Drive/ Grand Ave is by far the most popular start station for casual riders. Cyclistic could have an advertising presence in this area eg. digital bus advertising or billboards. Alternatively, special deal brochures promoting the benefits of annual membership could be handed out at popular stations during peak season.
    2. It would be good to know whether the increased demand during Summer is due to an increased number of visitors or more locals enjoying rides. If it is mainly due to visitors, it’s unlikely these riders will convert to an annual membership. A survey could be taken with riders near popular stations and analysis done to assess what proportion of casual riders are locals vs visitors. If the outcome of such analysis reveals increased usage is mainly due to locals, increased marketing spend could be employed during the summer months. Price reductions on memberships could be offered during Winter to increase usage and reduce seasonal fluctuations, thus increasing the probability of a user financially benefiting from an annual membership.
    3. The ride length for casual members is significantly lower for week days compared to weekends. This suggests that some riders are using the bikes to commute to work. Perhaps these riders are not aware there is an option to purchase an annual membership which would be more cost effective. A marketing campaign could be employed to raise awareness of this option and include the cost benefit of becoming a member.