```
---
title: "Case Study Project 1"
output: html_document
date: "2023-02-01"
---

```{r setup, include = FALSE}
knitr::opts_chunk$set(message = FALSE, warning = FALSE, echo = FALSE)
```


```{r load packages, include = FALSE}
library(readr)
library(tidyverse)
library(skimr)
library(janitor)
library(dplyr)
library(lubridate)
library(scales)
library(hms)
library(ggplot2)
library(ggforce)
library(ggrepel)
library("googledrive")
library(huxtable)
```


Background information
Since 2016, Cyclistic has launched a successful bike-share programme. Since then, the program has grown to a fleet of 5824 bicycles into a network of 692 stations across Chicago. Until now, Cyclistic's approach has been relied on the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. The management of Cyclistic would like to improve overall market share of the bike-sharing scene. The finance analysts have concluded that annual members are much more profitable than casual riders. Rather than creating a marketing campaign that targets all-new customers, there is a good chance to convert casual riders into members.
The marketing analyst team now needs to better understand how annual members and casual riders differ, in order to create effective marketing strategies.

This report is based on analysis of Cyclistic rider data for the past year (August 2021 - July 2022).

The Objective
The objective of this report is to look at past year Cyclistic rider data to understand usage differences between member riders and casual riders and thereafter develop effective marketing strategy.

The Analysis


```{r data setup, include = FALSE}
drive_auth()
folder_url <- "https://drive.google.com/drive/folders/14Y8nI5uF_RV0iBPBmaRVgPYKr6g6DFwr"
drive_share_anyone(folder_url)
folder <- drive_get(as_id(folder_url))
csv_files <- drive_ls(folder, type = "csv")
map(csv_files$id, ~purrr::safely(drive_download)(as_id(.x), overwrite = FALSE))
TD_202108 <- read_csv("202108-divvy-tripdata.csv")
TD_202109 <- read_csv("202109-divvy-tripdata.csv")
TD_202110 <- read_csv("202110-divvy-tripdata.csv")
TD_202111 <- read_csv("202111-divvy-tripdata.csv")
TD_202112 <- read_csv("202112-divvy-tripdata.csv")
TD_202201 <- read_csv("202201-divvy-tripdata.csv")
TD_202202 <- read_csv("202202-divvy-tripdata.csv")
TD_202203 <- read_csv("202203-divvy-tripdata.csv")
TD_202204 <- read_csv("202204-divvy-tripdata.csv")
```
```

```
TD_202205 <- read_csv("202205-divvy-tripdata.csv")
TD_202206 <- read_csv("202206-divvy-tripdata.csv")
TD_202207 <- read_csv("202207-divvy-tripdata.csv")

tripdata <- bind_rows(TD_202108, TD_202109, TD_202110, TD_202111, TD_202112, TD_202201,
TD_202202, TD_202203, TD_202204, TD_202205, TD_202206, TD_202207)

```



```{r data cleaning, include=FALSE}
### check for missing data
skim_without_charts(tripdata)

### check for duplicates
tripdata %>%
get_dupes(ride_id)

```

The table below lists the variables available for analysis.

```{r variable list table}

Variable <- c("Variable name", "ride_id", "rideable_type", "started_at", "ended_at",
"start_station_name", "start_station_id", "end_station_name", "end_station_id",
"start_lat", "start_lng", "end_lat", "end_lng", "member_casual")
Description <- c("Description", "random ride ID number, not tied to customer details",
"bike type: Classic/Docked/Electric", "start date time", "end date time", "start station
name", "start station ID", "end station name", "end station ID", "start station
latitude", "start station longitude", "end station latitude", "end station longtidue",
"member type: Member/Casual")
Data <- c("Data available", "check data duplication", "differentiate bike type usage",
"calculate ride duration, ridership by time/day of the week/month", " ", "identify
stations with high usage", " ", " ", " ", " ", " ", " ", " ", "differentiate member
type")

variable_list_table <- data.frame(Variable, Description, Data)

### draw table
h_table1 <- as_hux(variable_list_table, add_colnames = FALSE) %>%
  merge_down(., 4:5, 3) %>%
  merge_down(., 6:13, 3) %>%
  set_all_borders(.,everywhere, everywhere, 1) %>%
  set_bold(1, 1:3, TRUE) %>%
  set_background_color(1, 1:3, "#EDEDED")

set_markdown(h_table1)
```

1. Ridership Breakdown

1.1 Ridership breakdown by rider type
``` {r rider breakdown by rider type table code, include = FALSE}
### Order Ridertype
tripdata$member_casual <- ordered(tripdata$member_casual, levels = c("member",
"casual"))

### Table of Ridership breakdown by rider type
ridertype <- tripdata %>%
group_by(member_casual) %>%
summarise(percent = 100 * n() / nrow(tripdata))
```

```{r rider breakdown by rider type pie chart code, include = FALSE}
### draw pie chart
p1.1_ridership_ridertype <- ggplot(data=ridertype, aes(x= "", y = percent, fill =
```

```
member_casual)) +
geom_col(color = "white") +
geom_label_repel(aes(label = paste(round(percent, digits=2), "%", sep=""), group =
factor(member_casual), fill = factor(member_casual)), color = "black", fill = "white",
position = position_stack(vjust = 0.5)) +
coord_polar(theta = "y") +
scale_fill_manual(values = c("#FFCC66", "#99CCFF")) +
ggtitle("Ridership breakdown by Rider Type") +
theme_void() +
  theme(panel.background = element_rect(colour = "black"))+
          theme(plot.title = element_text(size = 12)) +
                  labs(fill = "Rider type")
```


```{r pie chart of ridership rider type breakdown}
p1.1_ridership_ridertype
```

Based on the past year data, currently the Members make up 57% of the riders, while Casual riders constitutes 43%.

\ \n

1.2 Bike type breakdown

```{r member biketype table code, include = FALSE}
### member rider biketype table
member_biketype <- tripdata %>%
filter(member_casual == "member") %>%
group_by(rideable_type, member_casual) %>%
summarise(percent = 100 * n() / nrow(.))
```

```{r member biketype chart code, include = FALSE}
### draw pie chart
p1.2_member_biketype <- ggplot(data=member_biketype, aes(x= "", y = percent, fill =
rideable_type)) +
geom_col(color = "white") +
geom_label_repel(size = 5, aes(label = paste(round(percent, digits=2), "%", sep="")),
group = factor(rideable_type), fill = factor(rideable_type)), color = "black", fill =
"white", position = position_stack(vjust = 0.5)) +
coord_polar(theta = "y") +
ggtitle("Member Riders Bike Type Breakdown") +
scale_fill_brewer(palette="Blues") +
theme_void() +
theme(panel.background = element_rect(colour = "black")) +
theme(legend.text = element_text(size = 14), legend.title = element_text(size = 16),
plot.title = element_text(size = 17)) +
        labs(fill = "Bike type")
```

```{r casual rider biketype table code, include = FALSE}
### causal rider biketype table

tripdata$rideable_type <- str_replace(tripdata$rideable_type, "docked_bike",
"classic_bike")
casual_biketype <- tripdata %>%
filter(member_casual == "casual") %>%
group_by(rideable_type, member_casual) %>%
summarise(percent = 100 * n() / nrow(.))
```

```{r casual rider biketype chart code, include = FALSE}
### draw pie chart
p1.3_casual_biketype <- ggplot(data=casual_biketype, aes(x= "", y = percent, fill =
rideable_type)) +
geom_col(color = "white") +
geom_label_repel(size = 5, aes(label = paste(round(percent, digits=2), "%", sep="")),
```

```
group = factor(rideable_type), fill = factor(rideable_type)), color = "black", fill =
"white", position = position_stack(vjust = 0.5)) +
coord_polar(theta = "y") +
ggtitle("Casual Riders Bike Type Breakdown") +
scale_fill_brewer(palette="Oranges") +
theme_void() +
  theme(panel.background = element_rect(colour = "black")) +
theme(legend.text = element_text(size = 14), legend.title = element_text(size = 16),
plot.title = element_text(size = 17)) +
        labs(fill = "Bike type")
```

```{r biketype chart, fig.show = "hold", out.width = "50%"}
p1.2_member_biketype
p1.3_casual_biketype
```

The choice of bikes between the 2 rider groups are similar. There were slightly more take up for classic bikes (~50% range) compared to electric bikes (~40% range) for both groups.
\ \n

 ^Casual riders have 3 categories of bike: Docked bike, Classic bike and Electric bike. Docked bike and classic bike have been grouped together, to simplify comparison across Member and Causal riders.

\ \n

\ \n

2. Ridership breakdown by duration

```{r duration calculation code, include = FALSE}
### calculate ride duration
duration_min = difftime(as.POSIXct(tripdata$ended_at), as.POSIXct(tripdata$started_at),
unit = "mins")
df <- cbind(tripdata, duration_min)

### group duration by category
df2 <- df %>%
mutate(
category = case_when(
duration_min <= 1 ~"less than 1 min",
duration_min > 1 & duration_min <= 10 ~"more than 1 min up to 10 min",
duration_min >10 & duration_min <= 20 ~"more than 10 min up to 20 min",
duration_min >20 & duration_min <= 30 ~"more than 20 min up to 30 min",
duration_min >30 & duration_min <= 60 ~"more than 30 min up to 1 hour",
duration_min >60 & duration_min <= 180 ~"more than 1 hour up to 3 hours",
duration_min >180 ~"more than 3 hours"))

### arrange time category in ascending order
df2$category <- ordered(df2$category, levels = c("less than 1 min", "more than 1 min up
to 10 min", "more than 10 min up to 20 min", "more than 20 min up to 30 min", "more than
30 min up to 1 hour", "more than 1 hour up to 3 hours", "more than 3 hours"))


```



```{r member rider duration table code, include = FALSE}
### member rider duration table
member_duration <- df2 %>%
filter(member_casual == "member", category != "less than 1 min") %>%
group_by(category) %>%
summarise(percent = 100 * n() / nrow(.))
```

```{r member rider pie chart code, include = FALSE}
```

```
### calculate the start and end angles for each pie
dat_pies_m <- member_duration %>%
mutate(end_angle = 2*pi*cumsum(percent)/sum(percent),
start_angle = lag(end_angle, default = 0),
mid_angle = 0.5*(start_angle + end_angle))


rpie = 1                          ### pie radius
rlabel = 1 * rpie                 ### radius of the labels; a number slightly larger than
0.5 seems to work better, but 0.5 would place it exactly in the middle as the question
asks for. 0.5 for center of the pie, 1 for edge of the pie

### draw the pie
p2_member_duration <- ggplot(dat_pies_m) +
  geom_arc_bar(aes(x0 = 0, y0 = 0, r0 = 0, r = rpie,
                   start = start_angle, end = end_angle, fill = category)) +
  geom_label_repel(size = 5, aes(x = rlabel*sin(mid_angle), y = rlabel*cos(mid_angle),
label = paste(round(percent, digits=2), "%", sep="")),
             hjust = 0.9, vjust = 0.5) +
  coord_fixed() +
  ggtitle("Member Riders Duration Breakdown") +
  scale_fill_brewer(palette="Oranges") +
  scale_x_continuous(limits = c(-1, 1), name = "", breaks = NULL, labels = NULL) +
  scale_y_continuous(limits = c(-1, 1), name = "", breaks = NULL, labels = NULL) +
  theme_void() +
  theme(panel.background = element_rect(colour = "black")) +
theme(legend.text = element_text(size = 12), legend.title = element_text(size = 14),
plot.title = element_text(size = 17)) +
  labs(fill = "Duration category")
```

```{r casual rider duration table code, include = FALSE}
### casual rider duration table
casual_duration <- df2 %>%
filter(member_casual == "casual", category != "less than 1 min") %>%
group_by(category) %>%
summarise(percent = 100 * n() / nrow(.))
```

```{r casual rider duration chart code, include = FALSE}
### calculate the start and end angles for each pie
dat_pies_c <- casual_duration %>%
mutate(end_angle = 2*pi*cumsum(percent)/sum(percent),
start_angle = lag(end_angle, default = 0),
mid_angle = 0.5*(start_angle + end_angle))

rpie = 1                          ### pie radius
rlabel = 1 * rpie                 ### radius of the labels; a number slightly larger than
0.5 seems to work better, but 0.5 would place it exactly in the middle as the question
asks for. 0.5 for center of the pie, 1 for edge of the pie

## draw the pie
p2_casual_duration <- ggplot(dat_pies_c) +
  geom_arc_bar(aes(x0 = 0, y0 = 0, r0 = 0, r = rpie,
                   start = start_angle, end = end_angle, fill = category)) +
  geom_label_repel(size = 5, aes(x = rlabel*sin(mid_angle), y = rlabel*cos(mid_angle),
label = paste(round(percent, digits=2), "%", sep="")),
            hjust = 0.5, vjust = 0.5) +
  coord_fixed() +
  ggtitle("Casual Riders Duration Breakdown") +
  scale_fill_brewer(palette="Blues") +
  scale_x_continuous(limits = c(-1, 1), name = "", breaks = NULL, labels = NULL) +
  scale_y_continuous(limits = c(-1, 1), name = "", breaks = NULL, labels = NULL)  +
  theme_void() +
  theme(panel.background = element_rect(colour = "black")) +
theme(legend.text = element_text(size = 12), legend.title = element_text(size = 14),
plot.title = element_text(size =17)) +
  labs(fill = "Duration category")
```

```{r rider duration pie chart, fig.show = "hold", out.width = "50%"}
p2_member_duration
p2_casual_duration
```

Duration breakdown for Member riders show 54% of Member riders cycle for less than 10 minutes, 28% cycle between 10 to 20 minutes. Short commute of up to 20 minutes total to 82% of Member ridership. This suggests that Member riders use the bike mainly for short commute.

There is an equal mix of Casual riders riding for less than 10 minutes (32%) and between 10 to 20 minutes (31%). The next 2 duration categories "between 20 to 30 minutes" and "30 minutes up to an hour" makes up 15% and 14% of Casual rider ridership. More Casual riders are riding for a longer duration compared to Member riders, indicating that there are more Causal riders cycle for leisure.

\ \n

3. Ridership grouped by Day of the Week

```{r ridership grouped by day of the week table code, include = FALSE}
### extract day of the week
day = strftime(tripdata$started_at, '%A')
df3 <- cbind(tripdata, day)
### arrange day
df3$member_casual <- ordered(df3$member_casual, levels = c("member", "casual"))
df3$day <- ordered(df3$day, levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))

### ridership breakdown by day
rider_day <- df3 %>%
group_by(member_casual, day) %>%
summarise(num_of_rides = n())
```

```{r ridership breakdown by day chart code, include = FALSE}
p3_ridership_day <- ggplot(data=rider_day, aes(x = day, y = num_of_rides, fill = member_casual)) +
geom_bar(color = "white", stat = "identity", position = position_dodge(), size = .3) +
ggtitle("Ridership grouped by Day of the Week") +
scale_fill_manual(values = c("#FFCC66", "#99CCFF")) +
xlab("Day of the Week") +
ylab("Number of rides") +
theme_bw() +
theme(axis.text.x = element_text(angle = 45), axis.text = element_text(size = 10)) +
        theme(plot.title = element_text(size = 12)) +
  labs(fill = "Rider type")
```

```{r ridership breakdown by day chart}
p3_ridership_day
```

The ridership of Member riders are generally more consistent throughout the week compared to Casual riders, with more riders using the bike during weekdays. Casual riders' bike usage is significantly higher during weekends.

\ \n

4. Ridership breakdown by Month

```{r ridership breakdown by month table code, include = FALSE}
### Extract month from date ##
df4 <- df %>%
mutate(month = format(df$started_at, "%B"))
### arrange month
df4$member_casual <- ordered(df4$member_casual, levels = c("member", "casual"))
```

```
df4$month <- ordered(df4$month, levels = c("August", "September", "October", "November",
"December", "January", "February", "March", "April", "May", "June", "July"))

### Ridership breakdown by month
rider_month <- df4 %>%
group_by(month, member_casual) %>%
summarise(num_of_rides = n())
```

```{r ridership breakdown by month chart code, include = FALSE}
p4_ridership_month <- ggplot(data=rider_month, aes(x = month, y = num_of_rides, fill =
member_casual)) +
geom_bar(color = "white", stat = "identity", position = position_dodge(), size = .3) +
ggtitle("Ridership breakdown by Month") +
scale_fill_manual(values = c("#FFCC66", "#99CCFF")) +
xlab("Month") +
ylab("Number of Rides") +
theme_bw() +
theme(axis.text.x = element_text(angle = 45), axis.text = element_text(size = 10)) +
  theme(plot.title = element_text(size = 12)) +
  labs(fill = "Rider type")
```

```{r ridership breakdown by month chart}
p4_ridership_month
```

\ \n

Warmer months such as August to October 2021 and May to July 2022 have higher higher ridership compared to colder months like Noverber 2021 to April 2022, which is in line with the general trend worldwide^1^.
A steeper decrease is observed in ridership for Casual riders when the climate transits to the colder months.

\ \n

5. Ridership breakdown by Time

```{r ridership breakdown by time table code, include = FALSE}
### extract time
start_hour <- hour(as_hms(tripdata$started_at))
df5 <- cbind(tripdata, start_hour)  ## add column without mutate() ##

### Time category
df6 <- df5 %>%
mutate(
category_time = case_when(
start_hour == 5 ~"5am",
start_hour == 6 ~"6am",
start_hour == 7 ~"7am",
start_hour == 8 ~"8am",
start_hour == 9 ~"9am",
start_hour == 10 ~"10am",
start_hour == 11 ~"11am",
start_hour == 12 ~"12pm",
start_hour == 13 ~"1pm",
start_hour == 14 ~"2pm",
start_hour == 15 ~"3pm",
start_hour == 16 ~"4pm",
start_hour == 17 ~"5pm",
start_hour == 18 ~"6pm",
start_hour == 19 ~"7pm",
start_hour == 20 ~"8pm",
start_hour == 21 ~"9pm",
start_hour == 22 ~"10pm",
start_hour == 23 ~"11pm",
start_hour >= 0 & start_hour <= 5 ~"between 12am to 5am"))
```

```
### arrange time category
df6$member_casual <- ordered(df6$member_casual, levels = c("member", "casual"))
df6$category_time <- ordered(df6$category_time, levels = c("5am", "6am", "7am", "8am",
"9am", "10am", "11am", "12pm", "1pm", "2pm", "3pm", "4pm", "5pm", "6pm", "7pm", "8pm",
"9pm", "10pm", "11pm", "between 12am to 5am"))

## Ridership breakdown by Time
rider_time <- df6%>%
group_by(category_time, member_casual) %>%
summarise(num_of_rides = n())
```

```{r ridership breakdown by time chart code, include = FALSE}
p5_ridership_time <- ggplot(data = rider_time, aes(x = category_time, y = num_of_rides,
color = member_casual, group = member_casual))+
geom_line() +
ggtitle("Ridership breakdown by Time") +
  scale_color_manual(values = c("#FFCC66", "#99CCFF")) +
xlab("Time") +
ylab("Number of Rides") +
theme_bw() +
theme(axis.text.x = element_text(angle = 45), axis.text = element_text(size = 10)) +
  theme(plot.title = element_text(size = 12)) +
  labs(colour = "Rider type")
```

```{r ridership breakdown by time chart}
p5_ridership_time
```
\ \n

The Member riders dataset has 3 peak period that stands out:

* 7 - 8am, which coincides with the morning rush hour
* 12pm, the lunch hour
* 3-7pm , which overlaps with the evening rush hours

Member riders are using the bike for short commute to-and-fro work/school and during lunch hour.

The Casual riders dataset shows ridership increases through the day from 10 am on and peak at 5pm, suggesting that Casual riders are using the bike for leisure and/or running errands.

\ \n

6. Summary of Rider Behaviour

Below is a table that summarises and compares rider behaviour:

```{r summary table}
##data frame
Member <- c("57% of riders are Members. 82% of members uses the bike for 20 minutes or
less
**→ short commute** ", "Highest ridership occur during weekdays
   **→ mainly for commute to work/school**", "3 peak period for rides: 7-8am, 12pm, 3-
7pm
   **→ commute to-and-fro workplace/school and lunch**")
Casual <- c("32% of riders cycle for <10 minutes, 31% ride for 10~20 minutes
   **→ a mix of short and medium duration commute**", "Highest ridership occur during
weekends
   **→ mainly for leisure**", "Ridership increases steadily from 10am and peak at 5pm
   **→ running errands/possibly tourist**")
summary_table <- data.frame(Member, Casual)

### draw table
h_table2 <- as_hux(summary_table, add_colnames = TRUE) %>%
 set_all_borders(.,everywhere, everywhere, 1) %>%
  set_bold(1, 1:2, TRUE) %>%
```

```
  set_background_color(1, 1:2, "#EDEDED")

set_markdown(h_table2)
```
\ \n

Limitations

Below is a table that summarises limitations of the dataset:

```{r limitations table}

Limitation <- c("Data-privacy prohibits the team from using riders' personally identifiable information", "2.2% of the data contained rides with less than 1 minute duration, which equates to a travel distance of ~0.3km -> possibly faulty bikes?")
Consequence <- c("-Unable to determine if users are residents or tourists
                 \ \n-Unable to identify multiple trips on the same day", "Unable to determine the cause of these rides, thus unable to rectify.")
Action <- c("Additional data required to enable the team to develop better targeted strategies", "Additional data required to ascertain faulty bikes. For this analysis, data with ride duration under 1 minute has been excluded")

limitation_table <- data.frame(Limitation, Consequence, Action)

### draw table
h_table3 <- as_hux(limitation_table, add_colnames = TRUE) %>%
 set_all_borders(.,everywhere, everywhere, 1) %>%
  set_bold(1, 1:3, TRUE) %>%
  set_background_color(1, 1:3, "#EDEDED")

set_markdown(h_table3)

```
\ \n

Recommendation

As a refresher, the finance team has concluded that annual members are much more profitable than casual riders. Thus the management are looking to maximise the number of annual memberships.

Before designing a new marketing strategy to convert casual riders to annual members, we need to understand the bike usage behaviour difference between the 2 groups.
Through the analysis, we have learnt that Member riders use the bike mainly for daily commute: weekday rides, duration within 10 minutes, peak period of morning rush hours, lunch hour and evening rush hours. Casual riders mainly use the bike for leisure purpose or running errands: weekend rides, duration of up to 20 minutes, riding between time period of 10am to evening. It has been determined that the bike usage pattern of Member riders and Casual riders are different. Therefore, it might take more than marketing strategies to entice Casual riders to commit to subscription.

Firstly, there is a need to take a closer look at Casual riders. There is a need to differentiate Casual riders who resides in Chicago from Tourists, as it is plausible for residents to commit to an annual subscription. As the frequency of Casual riders riding the bikes are lower than Member riders, the existing annual subscription package may be deemed as uneconomical for their usage.
A suggestion would be to devise a new shorter ride subscription pass for these Casual riders.
Another group of people who may use the Single Ride options are the Tourists. A subscription plan may not be enticing to Tourist, it might be worthwhile to explore multi-day short commute pass for Tourist.
An important next step would be to conduct a survey to hear directly from the users themselves: what would entice them to commit to subscription plan, what features do they want, how can we improve the user experience, in order to better understand their needs.

Summary

```{r report summary table}
```

```
Objective <- c("* Understand usage differences between Member riders and Casual riders \
\n* Develop effective marketing strategy")
Analysis <- c("* Member riders use the bike mainly for daily commute \ \n* Casual riders
use the bike mainly for leisure/running errands")
Recommendation <- c("* Deeper analysis on Casual riders required:
                     \ \n   + Resident Casual riders -> devise new shorter ride
subscription pass/weekend subscription pass
                     \ \n    + Tourist Casual riders -> multi-day pass
                    \ \n* Conduct survey:
                      \ \n    + What would entice Casual riders to commit to
subscription?")

report_summary_table <- data.frame(Objective, Analysis, Recommendation)

### draw table
h_table4 <- as_hux(report_summary_table, add_colnames = TRUE) %>%
  set_all_borders(.,everywhere, everywhere, 1) %>%
  set_bold(1, 1:3, TRUE) %>%
  set_background_color(1, 1:3, "#EDEDED")

set_markdown(h_table4)
```