---
title: "cyclistic_bike_data"
author: "Reynardt Cloete"
date: "11/09/2021"
output: html_document

About the company

Cyclistic is a bike-share program that features more than 5800 bicycles and 600 docking stations. Cyclistic users are more likely to ride for leisure, while about 30% use them to travel to work each day. Single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance office concluded that annual members are more profitable than casual riders. Moreno (Director of Marketing) believes maximizing the number of annual riders will be key to future growth. Moreno's goal is to convert casual riders to annual members.

Stakeholders and team

Key stakeholders include: Cyclistic executive team, Director of Marketing (Lily Moreno), Marketing Analytics team.

Ask

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual membership?
3. How can Cyclistic use digital media to influence casual riders to become members?

Moreno has assigned the data analyst to answer the first question, namely: How do annual members and casual riders use Cyclistic bikes differently?

Setting up my environment

Notes: Setting up my R environment by loading the 'tidyverse', 'lubridate', and 'janitor' packages.

```{r loading packages}
library(tidyverse)
library(lubridate)
library(janitor)
```

Importing the data

Notes: Cyclistic user data for the past 12 months has been downloaded. Local copies have been stored securely on my personal computer. All data is comma-delimited (.CSV) format. We are going to assume that the data is credible due to the fact that it is public data.
Here are the last 12 months of bike data.

```{r import the data}
df1 <- read.csv("202008-divvy-tripdata.csv")
df2 <- read.csv("202009-divvy-tripdata.csv")
df3 <- read.csv("202010-divvy-tripdata.csv")
df4 <- read.csv("202011-divvy-tripdata.csv")
df5 <- read.csv("202012-divvy-tripdata.csv")
df6 <- read.csv("202101-divvy-tripdata.csv")
df7 <- read.csv("202102-divvy-tripdata.csv")
df8 <- read.csv("202103-divvy-tripdata.csv")
df9 <- read.csv("202104-divvy-tripdata.csv")
df10 <- read.csv("202105-divvy-tripdata.csv")
df11 <- read.csv("202106-divvy-tripdata.csv")
df12 <- read.csv("202107-divvy-tripdata.csv")
```

Combine with rbind

Notes: We need to combine all 12 files into 1 data set in order to work with the data. Here we will have 4731081 rows.

```{r combine data}
yearly_bike_data <- rbind(df1, df2, df3, df4, df5, df6, df7, df8, df9, df10, df11, df12)
```

Data cleaning

Notes: Let us clean some data by checking for and removing empty columns and rows, as well as rows containing N/A values. This leaves us with 4595623 rows.

```{r clean data}
yearly_bike_data <- janitor::remove_empty(yearly_bike_data, which = c("cols"))
yearly_bike_data <- janitor::remove_empty(yearly_bike_data, which = c("rows"))
yearly_bike_data <- yearly_bike_data[complete.cases(yearly_bike_data), ]
```

Change date format

Notes: Convert started_at and ended_at dates to date/time stamps (ymd,hms).

```{r date format 1}
yearly_bike_data$started_at <- lubridate::ymd_hms(yearly_bike_data$started_at)
yearly_bike_data$ended_at <- lubridate::ymd_hms(yearly_bike_data$ended_at)
```

Calculate bike ride duration in seconds

Notes: Determine how long each bike ride was in seconds by subtracting the 'started_at' form 'ended_at'. We see here that some of them have negative values.

```{r duration in seconds}
yearly_bike_data <- yearly_bike_data %>%
  mutate(ride_length_in_seconds = yearly_bike_data$ended_at - yearly_bike_data$started_at)
```

Remove negative values

Notes: Now we can remove those rows with negative values by selecting all those rows that are greater than 0 seconds. This leaves us with 4587101 rows.

```{r negative values}
yearly_bike_data <- yearly_bike_data[yearly_bike_data$ride_length_in_seconds > 0, ]
```

Calculate bike ride duration in minutes and hours

Notes: Since the rides don't last longer than a day, we need to see how long they are in terms of minutes and hours in order to compare casual and member riders.

```{r duration in min and hr}
yearly_bike_data$hours <- difftime(yearly_bike_data$ended_at, yearly_bike_data$started_at, units = ("hours"))
yearly_bike_data$mins <- difftime(yearly_bike_data$ended_at, yearly_bike_data$started_at, units = ("mins"))
```

Remove columns

Notes: Let's remove the columns we won't be working with by selecting them and adding negative values before them.

```{r unnecessary columns}
yearly_bike_data <- yearly_bike_data %>%
  select(-ride_id, -end_station_name, -end_station_id, -start_lat, -end_lat, -start_lng, -end_lng)
```

Seperate date and time

Notes: It will probably simplify the analysis process if we separate date from time. Now we have 4 columns.

```{r seperate date and time}
yearly_bike_data <- separate(yearly_bike_data,"started_at",into=c('start_date','start_time'), sep=' ')
yearly_bike_data <- separate(yearly_bike_data,"ended_at",into=c('end_date','end_time'), sep=' ')
```

Change date format

Notes: Change date format for new separated date/time columns.

```{r date format 2}
yearly_bike_data$start_date <- lubridate::ymd(yearly_bike_data$start_date)
yearly_bike_data$end_date <- lubridate::ymd(yearly_bike_data$end_date)
```

Create weekdays

Notes: Create a column to display weekdays in order to compare weekday bike use between groups.

```{r create weekdays}
yearly_bike_data <- yearly_bike_data %>%
  mutate(day_of_week = weekdays(yearly_bike_data$start_date))
```

Remove minutes with values < 1

Notes: We're not going to be working with rides less than one minute. So let's select only those that are > 1. We will now have 4517784 rows of data.

```{r remove values < 1}
yearly_bike_data <- yearly_bike_data[yearly_bike_data$mins > 1, ]
```

Determine how many are casual riders and how many are members

Notes: Let us count the number of users from the casual group and the member group. As we will see, there are more members than casual users.

```{r count casual_member}
table(yearly_bike_data['member_casual'])
```

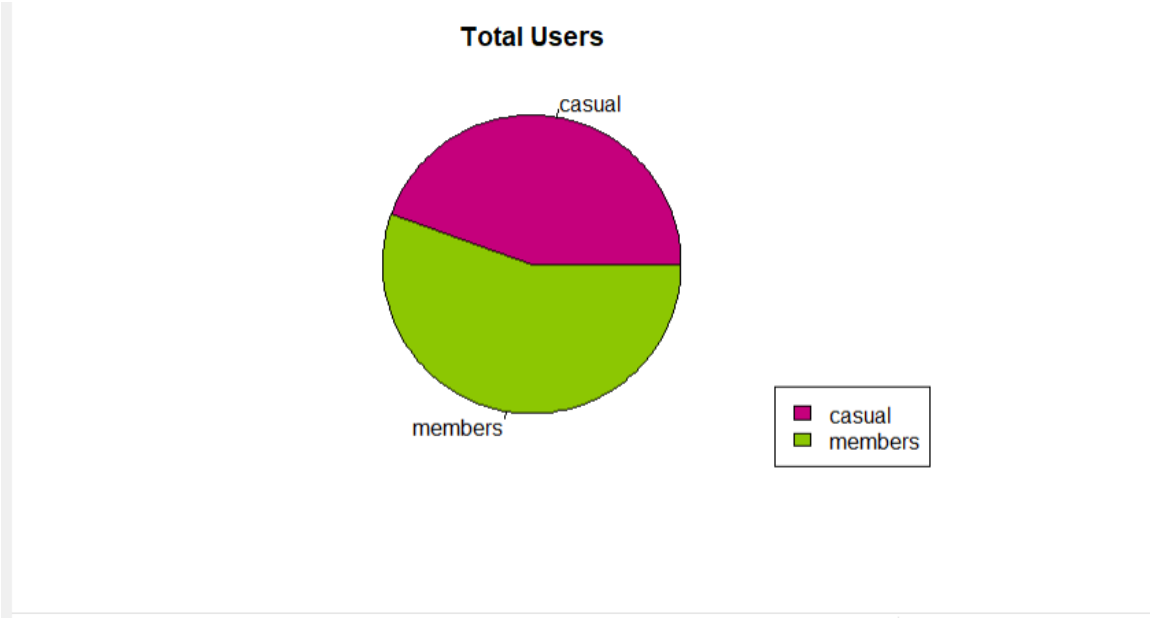
casual member
2010677 2507107

Pie chart

Notes: A pie chart will be a good visual display of a comparison between the 2 groups. First we create the values.

```{r pie chart}
total_number_of_users <- c(2010677, 2507107)
pie_labels <- c("casual", "members")
colors <- c("#C5007C", "#8CC702")

pie(total_number_of_users, label=pie_labels, main = "Total Users", col=colors)
legend("bottomright", pie_labels, fill = colors)
```



Create column for months

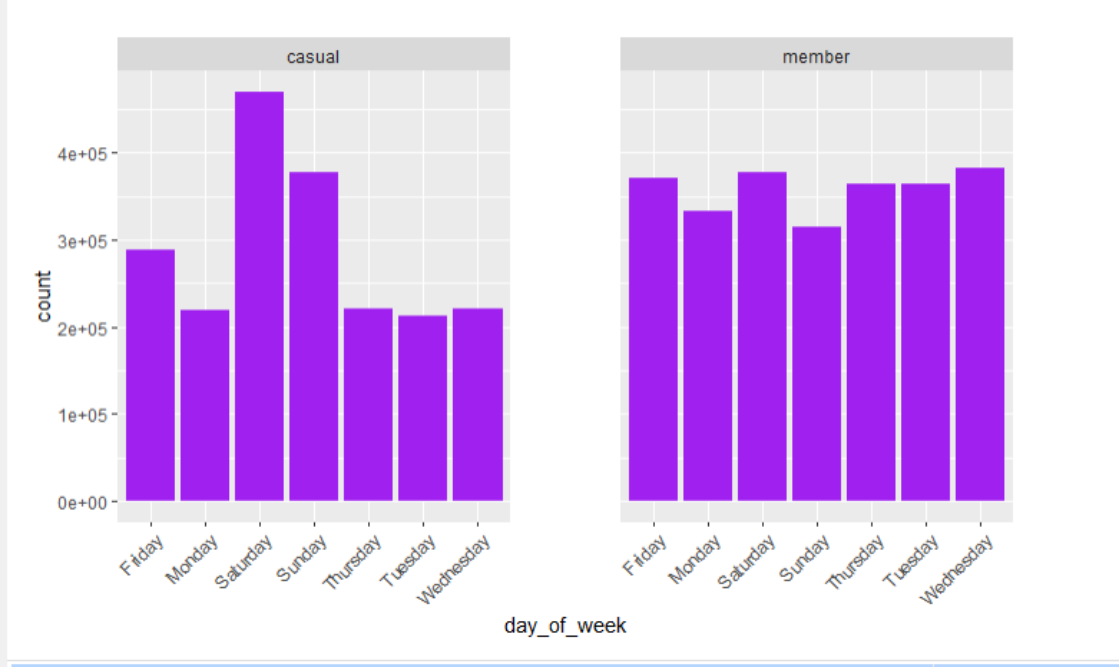
Notes: We want to see how these 2 groups make use of the bikes throughout the year.

```{r create months}
yearly_bike_data <- yearly_bike_data %>%
  mutate(months = case_when(
    month(yearly_bike_data$start_date)==1 ~ "Jan",
    month(yearly_bike_data$start_date)==2 ~ "Feb",
    month(yearly_bike_data$start_date)==3 ~ "Mar",
    month(yearly_bike_data$start_date)==4 ~ "Apr",
    month(yearly_bike_data$start_date)==5 ~ "May",
    month(yearly_bike_data$start_date)==6 ~ "Jun",
    month(yearly_bike_data$start_date)==7 ~ "Jul",
    month(yearly_bike_data$start_date)==8 ~ "Aug",
    month(yearly_bike_data$start_date)==9 ~ "Sep",
    month(yearly_bike_data$start_date)==10 ~ "Oct",
    month(yearly_bike_data$start_date)==11 ~ "Nov",
    month(yearly_bike_data$start_date)==12 ~ "Dec",
  ))
```

Weekday bike usage

Notes: First let's look at how these groups differ in terms of weekday usage by means of a bar chart.

```{r bar chart 1}
ggplot(data=yearly_bike_data) +
  geom_bar(mapping=aes(x=day_of_week), fill="purple") +
  facet_wrap(~member_casual) +
  theme(panel.spacing = unit(4, "lines"), axis.text.x=element_text(angle=45, hjust=1))
```



Create seasons using months table

Notes: We can also create a column that display which seasons correspond to which month.

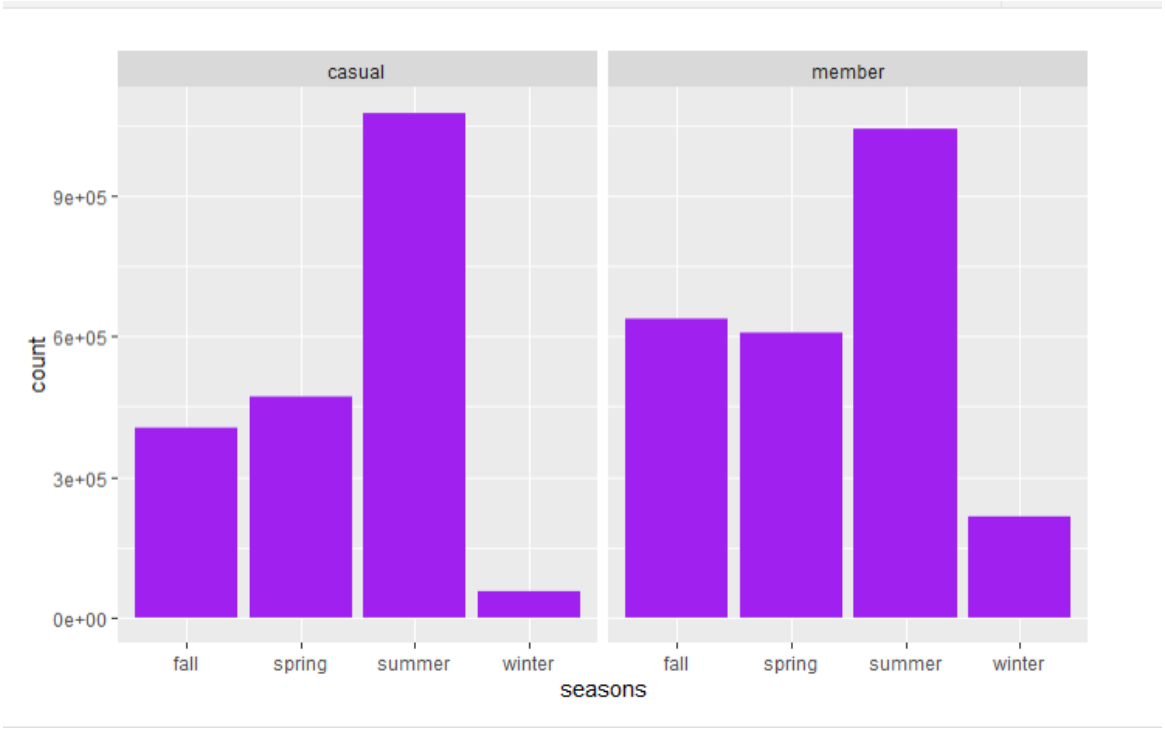
```{r create seasons}
yearly_bike_data <- yearly_bike_data %>%
  mutate(seasons = case_when(
    month(yearly_bike_data$start_date)==12|
      month(yearly_bike_data$start_date)==1|
      month(yearly_bike_data$start_date)==2 ~ "winter",
      month(yearly_bike_data$start_date)==3|
      month(yearly_bike_data$start_date)==4|
      month(yearly_bike_data$start_date)==5 ~ "spring",
      month(yearly_bike_data$start_date)==6|
```

```
    month(yearly_bike_data$start_date)==7|
    month(yearly_bike_data$start_date)==8 ~ "summer",
    month(yearly_bike_data$start_date)==9|
    month(yearly_bike_data$start_date)==10|
    month(yearly_bike_data$start_date)==11 ~ "fall"
  ))
```

Seasonal bike usage

Notes: Let's look at how these groups differ in terms of seasonal bike usage by means of a bar chart.

```{r bar chart 2}
ggplot(data=yearly_bike_data) +
  geom_bar(mapping=aes(x=seasons), fill="purple") +
  facet_wrap(~member_casual)
```



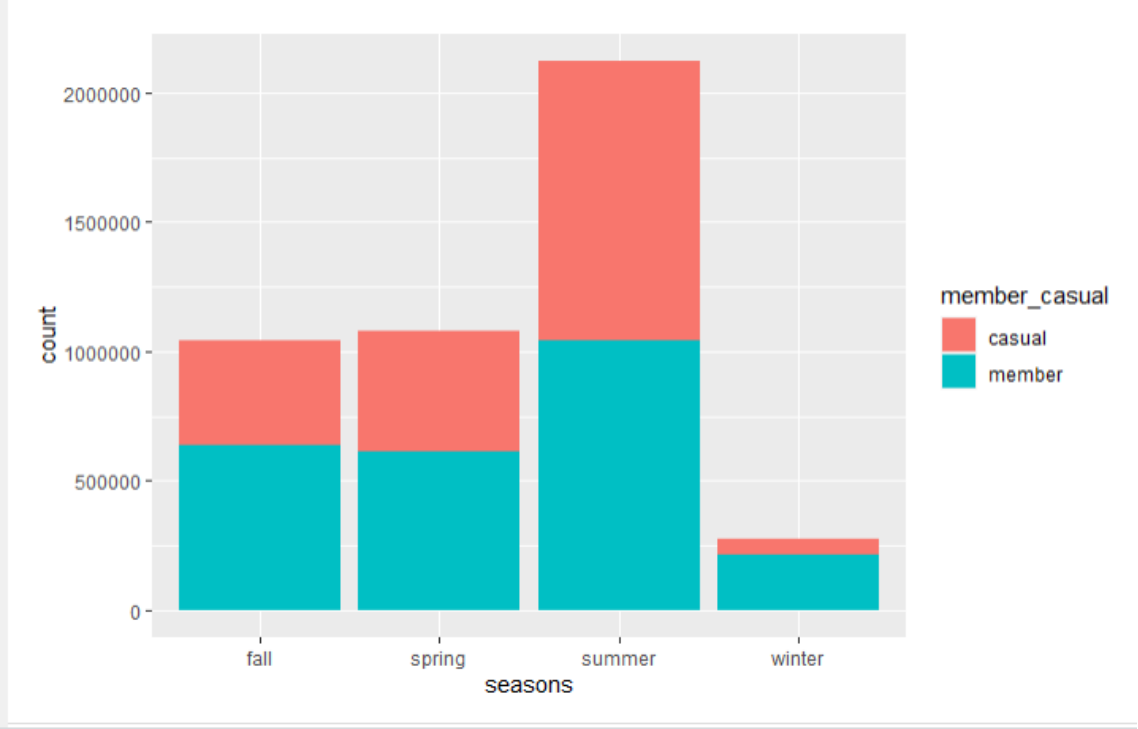
determine average, minimum and maximum ride_length of members and casual users

Notes: It appears here that casual riders in fact ride for longer periods of time compared to members. The average_ride_time of a casual user is 36.45 mins, and 14.70 mins for a member. The min_ride_length is 1 minute for both groups.

```{r min, max, avr}
ride_summary <- yearly_bike_data %>%
  group_by(member_casual) %>%
  summarise(average_ride_time = mean(mins),
            min_ride_length = min(mins),
            max_ride_length = max(mins))
```

lets display the member and casual riders within the seasons in a bar chart

```{r bar chart 3}
ggplot(data=yearly_bike_data) +
  geom_bar(mapping=aes(x=seasons,fill=member_casual))
```



Findings

Notes: Based on these findings, the following recommendations can be made.

- The average ride time of a casual user is 36.45 minutes.
- The average ride time of a member is 14.70 minutes.
- The minimum ride length for both groups is 1 minute.

- Casual riders ride more over the weekends compared to members.
- Members have a more consistent use of bikes throughout the week compared to casual users.
- Casual riders make use of the bikes for recreational purposes, as we can see by the usage spike on weekends.
- Members use the bikes as a means of transport for travelling to work and back.
- For both groups, winter is an unpopular season to use bikes, whereas summer is most popular.
- Members also seem to use the bikes more in fall and spring compared to casual users.

Casual users should be approached during the summer time, when bike use is at its peak and encouraged to become members as the data shows they use Cyclistic more than members on average, in order for them to save money.
Alternatively, weekend-only membership can be offered to casual users allowing them to use bikes only on weekends, whilst week-day use will include additional costs.

In order to increase usage for both groups in winter months, rewards can be offered for reaching certain goals in those months.