Cyclist Project

Setup

```
##installing packages
install.packages("tidyverse")
install.packages("dplyr")
install.packages("janitor")
install.packages("tidyr")
install.packages("lubridate")
##loading packages
library(tidyverse)
library(dplyr)
library(janitor)
library(tidyr)
library(lubridate)
```

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

About the company

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 very good chance 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.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst 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 thei

---

Ask

My project will answer these three overarching questions:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?
4. What additional data is needed for further research to expand on findings?

To answer these questions, I will create a series of charts depicting the differences in bike type, trip frequency, trip duration, trip start time, and trip starting location between members and casual users.

Prepare

The data:

- To prepare the data, I first downloaded it from the source website and then examined them. This allowed me to get a better understanding of the data and how best to perform analysis on it.
- To narrow the scope of the project, I only used 10-month's worth of data from January 2021 to July 2021. This was `202101-divvy-tripdata.zip` through to `202110-divvy-tripdata.zip` from the website.
- I imported the data into R studio to begin the data cleaning process:

```
##import the data
Data1 <- read_csv("202101-divvy-tripdata.csv")
Data2 <- read_csv("202102-divvy-tripdata.csv")
data3 <- read_csv("202103-divvy-tripdata.csv")
data4 <- read_csv("202104-divvy-tripdata.csv")
data5 <- read_csv("202105-divvy-tripdata.csv")
data6 <- read_csv("202106-divvy-tripdata.csv")
data7 <- read_csv("202107-divvy-tripdata.csv")
data8 <- read_csv("202108-divvy-tripdata.csv")
data9 <- read_csv("202109-divvy-tripdata.csv")
```

```
data10 <- read_csv("202110-divvy-tripdata.csv")
```

- I joined the data together using `full_join()`:

```
##join dataframes together
data_2021 <- data_202101 %>%
  full_join(., data_202102) %>%
  full_join(., data_202103) %>%
  full_join(., data_202104) %>%
  full_join(., data_202105) %>%
  full_join(., data_202106) %>%
  full_join(., data_202107) %>%
  full_join(., data_202108) %>%
  full_join(., data_202109) %>%
  full_join(., data_202110)
```

- This was possible as all 10 datasets had the same columns.
- However, the dataset was too large to present on R markdown, so I performed the steps using a small segment of the data to show the process.
- I used `glimpse()` to examine the data:

```
##examine data
glimpse(data_2021)
```

```
## Rows: 96,834
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45
C94683…
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_b
ike", …
## $ started_at         <dttm> 2021-01-23 16:14:19, 2021-01-27 18:43:08, 20
21-01-…
## $ ended_at           <dttm> 2021-01-23 16:24:44, 2021-01-27 18:47:12, 20
21-01-…
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave
& Cor…
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660",
"17660…
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Wood St
& Augu…
```

```
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "657", "13258",…
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4…
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696…
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4…
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700…
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "…
```

Process

- I started parsing the data by choosing only the `ride_id`, `rideable_type`, `started_at`, `ended_at`, `start_station_name`, `end_station_name`, and `member_casual` columns with `select()`.
- I then created a `trip_duration_sec`, `trip_duration_min` and `start_hour` columns using `mutate()`. These are columns depicting the duration of each bike trip in seconds, minutes, and the hour that each trip was started at.
- Finally, I filtered the data to only consist of trips which were between 1 minute and 60 minutes long, and also had data for the start and end station with `filter()`.
- A large proportion of trips were recorded as being under 1 minute long. I did not feel they counted as a full bike trip, so I removed them.
- I also removed trips above 60 minutes in duration as they were few, and there were some trips which were extremely long in duration. This also allowed me to limit the scope of the project.
- Filtering out data which had no start-station or end-station recorded also allowed me to remove data which was incomplete.

```
## parsing data
data_2021_new <- data_2021 %>%
  select(ride_id, rideable_type, started_at, ended_at, start_station_name, end_station_name, member_casual) %>% #select columns
  mutate(trip_duration_sec=(ended_at-started_at), #add trip_duration_sec
         trip_duration_min=(trip_duration_sec/60), #add trip_duration_min
         start_hour=hour(started_at)) %>% #add start_hour
  filter(trip_duration_min>1 & #filter to find trips longer than 1 min
           trip_duration_min<60 & #filter to find trips shorter than 60 mins
           start_station_name!="NA" & #filter out blank station names
           end_station_name!="NA")
```

- I then separated the new dataset into two separate datasets depending on whether trips were started by `member` or `casual` users.

```
##separate members and casual users
data_2021_member <- filter(data_2021_new, member_casual == "member")
data_2021_casual <- filter(data_2021_new, member_casual == "casual")
```

- After these changes, the new dataset was reduced to 10 columns (from 13) and 81,501 rows (from 96,834).

```
glimpse(data_2021_new)
```

```
## Rows: 81,501
## Columns: 10
## $ ride_id            <chr> "B9F73448DFBE0D45", "457C7F4B5D3DA135", "57C7
50326F…
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bi
ke", "…
## $ started_at         <dttm> 2021-01-24 19:15:38, 2021-01-23 12:57:38, 20
21-01-…
## $ ended_at           <dttm> 2021-01-24 19:22:51, 2021-01-23 13:02:10, 20
21-01-…
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave
& Cor…
## $ end_station_name   <chr> "Wood St & Augusta Blvd", "California Ave & N
orth A…
## $ member_casual      <chr> "member", "member", "casual", "casual", "casu
al", "…
## $ trip_duration_sec  <drtn> 433 secs, 272 secs, 587 secs, 537 secs, 609
secs, …
## $ trip_duration_min  <drtn> 7.216667 secs, 4.533333 secs, 9.783333 secs,
8.950…
## $ start_hour         <int> 19, 12, 15, 15, 15, 15, 10, 11, 7, 8, 8, 13,
9, 11,…
```

#Writing the cleaned data into a csv file so I can do the visualization

write.csv(data_2021_new,file = "CyclistCleanedDataNumber2.csv",row.names = FALSE)

1. **How do annual members and casual riders use Cyclistic bikes differently?**

- Highest number of trips for members and causal users peaks at 5pm.

- Members choose classic bikes more frequently than casual users.
- Stations with the highest proportions are **Lake Shore Dr & Monroe St** and **Streeter Dr & Grand Ave**
- Members start trips between 5am-7am more often than casual users. The difference is about 6 times more on the members side.

2. **Why would casual riders buy Cyclistic annual memberships?**

- I believe that casual users would be most likely to buy Cyclistic annual memberships if they decide to start using the bikes more often, such as for daily routines like commuting to and from work.

3. **How can Cyclistic use digital media to influence casual riders to become members?**

    - Marketing the healthy lifestyle and how using bicycle for transportation within the city could help to improve their health and save our planet.
    - Plus, focusing on the stations which has a great amount of proportion of casual users