Cyclistic

Michael Porco

2022-10-29

Cyclystic Analysis

Stakeholder Lily Moreno, director of marketing, and my manager, has tasked me with finding the difference in usage of Cyclistic bikes between annual members, and casual riders.

Ask

How do the annual members and casual riders use bikes differently?

Prepare

I have downloaded the most recent 12 months of Cyclistic trip data to my computer. This data includes the start and end times of the ride, the day, month, and the year. This data is made available by Motivate International Inc. https://ride.divvybikes.com/data-license-agreement

Process

Within Excel, I have created a copy of the original data in csv format. I saved the original data and kept it in a separate folder. Next, I created two new columns of data in the excel spreadsheet. One for ride length, by comparing start and end times. And one for the day of the week.

I brought in the data to RStudio for cleaning and manipulation.

Importing Data

I installed and loaded the necessary packages. For this analysis I used tidyverse, janitor, lubridate, and skimr. Then, I load each months data into a data frame.

```
library(tidyverse)
library(janitor)
library(lubridate)
```

```
sep <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202109_tripdata.csv")
oct <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202110_tripdata.csv")
nov <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202111_tripdata.csv")
dec <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202112_tripdata.csv")
jan <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202201_tripdata.csv")
feb <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202202_tripdata.csv")
mar <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202203_tripdata.csv")
apr <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202204_tripdata.csv")
may <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202205_tripdata.csv")
jun <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202206_tripdata.csv")
jul <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202207_tripdata.csv")
aug <- read.csv("C:\\Users\\mikep\\Documents\\Cyclistic\\Cyclistic_csv\\202208_tripdata.csv")
```

Cleaning

Next, I make sure that all of the data frames are compatible, with the same number and type of columns. I used glimpse on each data frame to do this.

```
glimpse(sep)
```

```
## Rows: 756,147
## Columns: 15
## $ ride_id            <chr> "9DC7B962304CBFD8", "F930E2C6872D6B32", "6EF7213790~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "9/28/2021 16:07", "9/28/2021 14:24", "9/28/2021 0:~
## $ ended_at           <chr> "9/28/2021 16:09", "9/28/2021 14:40", "9/28/2021 0:~
## $ start_station_name <chr> "", "", "", "", "", "", "", "", "", "", "Clark St &~
## $ start_station_id   <chr> "", "", "", "", "", "", "", "", "", "", "TA13070001~
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",~
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",~
## $ start_lat          <dbl> 41.89000, 41.94000, 41.81000, 41.80000, 41.88000, 4~
## $ start_lng          <dbl> -87.68000, -87.64000, -87.72000, -87.72000, -87.740~
## $ end_lat            <dbl> 41.89, 41.98, 41.80, 41.81, 41.88, 41.88, 41.74, 41~
## $ end_lng            <dbl> -87.67, -87.67, -87.72, -87.72, -87.71, -87.74, -87~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
## $ ride_length        <chr> "0:02:00", "0:16:00", "0:03:00", "0:09:00", "0:10:0~
## $ day_of_week        <chr> "Tuesday", "Tuesday", "Tuesday", "Tuesday", "Tuesda~
```

After confirming the data frames compatibility, I combined them all into a single data frame.

```
df <- rbind(sep, oct, nov, dec, jan, feb, mar, apr, may, jun, jul, aug)
```

Now I make sure there are is no NA pieces of the data or anything that has characters which won't be accepted. I also filter out any ride where the ride length was zero. This could have been the company checking the systems.

```
df <- clean_names(drop_na(filter(df, ride_length != "0:00:00")))
```

Analyze

Add Data

I will be using lubridate functions to convert the start, end, and ride length columns to datetime figures. Some of the ride length times were coming up with NAs, but only a very small percentage that will have no effect on the analysis, so I will drop NAs again.

```
df <- mutate(df, started_at = mdy_hm(started_at))
df <- mutate(df, ended_at = mdy_hm(ended_at))
df <- mutate(df, ride_length = period_to_seconds(hms(ride_length)))
df <- drop_na(df)
```

Analysis

I will be using Tableau to analyze the data and to create supporting visualizations. So here, I will create a new csv file of the cleaned and analyzed data to be used in Tableau.

```
write.csv(df, "ride_data.csv")
```

Share

Here is a link to the analysis and visualization on tableau public.

https://public.tableau.com/app/profile/michael.porco/viz/CyclisticAnalysis_16673576046850/Story1

https://public.tableau.com/app/profile/michael.porco/viz/CyclisticAnalysismonthly/Story2

https://public.tableau.com/app/profile/michael.porco/viz/CyclisticAnalysishourly/Story3

https://public.tableau.com/app/profile/michael.porco/viz/CyclisticAnalysisridelength/Story4

https://public.tableau.com/app/profile/michael.porco/viz/CyclisticAnalysisbiketypes/Story5

Act

Based upon my analysis, I would like to give my top three recommendations.

1. Let users know when they have used the service enough that the membership would be less expensive.

2. Give a discount to the membership in spring so that people who would be riding more in the summer can feel like it is worth it.

3. Have different prices for casual riders based on the time of day, just like trains do with peak hours.