Code

Case study: How Does a Bike-Share Navigate Speedy Success?

Akash Panwar

26-01-2022

Introduction

Capstone Project for the Google Data Analytics Professional Certificate

The analysis is been followed in the 6 phases of the Data Analysis process: Ask, Prepare, Process, Analyze, Share, and Act.

A brief explanation of above phases:

Ask

- Ask effective questions
- Define the scope of the analysis
- Define what success looks like

Prepare

- Verify data's integrity
- Check data credibility and reliability
- Check data types
- Merge datasets

Process

- Clean, Remove and Transform data
- Document cleaning processes and results

Analyze

- Identify patterns
- Draw conclusions
- Make predictions

Share

- Create effective visuals
- Create a story for data
- Share insights to stakeholders

Act

- Give recommendations based on insights
- Solve problems
- Create something new

1. Ask

Scenario

In an effort to increase revenue, the marketing department, would like to find a way to maximise the number of annual memberships to our service. In order to better understand the market, I have been asked to analyse the ways in which annual members and casual riders use Cyclistic bikes differently. Data-driven insight into these trends should help the marketing team determine what might make casual riders more likely to buy an annual membership, and this will ultimately shape their digital media strategy. Given that the executive team must approve the marketing strategy, I will include some recommendations that align with the goal of increasing annual memberships.

Stakeholders:

- Director of marketing
- Cyclistic executive team

Objective

Hence, the objective for this analysis is to throw some light on how the two types of customers: annual members and casual riders, use Cyclistic bikeshare differently, based on few parameters that can be calculated/ obtained from existing data.

Deliverables:

- Insights on how annual members and casual riders use Cyclistic bikes differently
- Provide effective visuals and relevant data to support insights
- Use insights to give three recommendations to convert casual riders to member riders

2. Prepare

Data Sources

I downloaded the data from the divvy trip data and stored in desktop.

- I renamed the folder to make it simple.
- I renamed all the files as per the standard naming conventions.
- I took 12 months of data and merged into the single folder.

Documentation, Cleaning and Preparation of data for analysis

The combined size of all the 16 datasets is close to 1 GB. Data cleaning in spreadsheets will be time-consuming and slow compared to SQL or R. I am choosing R simply because I could do both data wrangling and analysis/ visualizations in the same platform. It is also an opportunity for me to learn R better.

Load libraries

Hide

```
library(tidyverse)
library(ggplot2)
library(lubridate)
library(dplyr)
library(readr)
library(janitor)
library(data.table)
library(tidyr)
```

Load datasets

Hide

```
tripdata_202004 <- read.csv("Cyclistic Data Source/202004-divvy-tripdata.csv")
tripdata_202005 <- read.csv("Cyclistic Data Source/202005-divvy-tripdata.csv")
tripdata_202006 <- read.csv("Cyclistic Data Source/202006-divvy-tripdata.csv")
tripdata_202007 <- read.csv("Cyclistic Data Source/202007-divvy-tripdata.csv")
tripdata_202008 <- read.csv("Cyclistic Data Source/202008-divvy-tripdata.csv")
tripdata_202009 <- read.csv("Cyclistic Data Source/202009-divvy-tripdata.csv")
tripdata_202010 <- read.csv("Cyclistic Data Source/202010-divvy-tripdata.csv")
tripdata_202011 <- read.csv("Cyclistic Data Source/202011-divvy-tripdata.csv")
tripdata_202012 <- read.csv("Cyclistic Data Source/202012-divvy-tripdata.csv")
tripdata_202101 <- read.csv("Cyclistic Data Source/202101-divvy-tripdata.csv")
tripdata_202102 <- read.csv("Cyclistic Data Source/202102-divvy-tripdata.csv")
tripdata_202103 <- read.csv("Cyclistic Data Source/202103-divvy-tripdata.csv")
tripdata_202104 <- read.csv("Cyclistic Data Source/202104-divvy-tripdata.csv")
tripdata_202105 <- read.csv("Cyclistic Data Source/202105-divvy-tripdata.csv")
tripdata_202106 <- read.csv("Cyclistic Data Source/202106-divvy-tripdata.csv")
tripdata_202107 <- read.csv("Cyclistic Data Source/202107-divvy-tripdata.csv")
```

Check column names of each dataset for consistency

Hide

```
colnames(tripdata_202004)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202005)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202006)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202107)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202008)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202009)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202010)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202011)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202012)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202101)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202102)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202103)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202104)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202105)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202106)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Hide

```
colnames(tripdata_202107)
```

```
 [1] "ride_id"            "rideable_type"      "started_at"         "ended_at"
"start_station_name"
 [6] "start_station_id"   "end_station_name"   "end_station_id"     "start_lat"
"start_lng"
[11] "end_lat"            "end_lng"            "member_casual"
```

Check data structures and data types for all data frames

Hide

```
str(tripdata_202004)
```

```
'data.frame':   84776 obs. of  13 variables:
 $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4"
"2A59BBDF5CDBA725" ...
 $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike"
...
 $ started_at        : chr  "2020-04-26 17:45:14" "2020-04-17 17:08:54" "2020-04-01 1
7:54:13" "2020-04-07 12:50:19" ...
 $ ended_at          : chr  "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 1
8:08:36" "2020-04-07 13:02:31" ...
 $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct &
Erie St" "California Ave & Division St" ...
 $ start_station_id  : int  86 503 142 216 125 173 35 434 627 377 ...
 $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana
Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
 $ end_station_id    : int  152 499 255 657 323 35 635 382 359 508 ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

```
str(tripdata_202005)
```

```
'data.frame':   200274 obs. of  13 variables:
 $ ride_id           : chr  "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52"
"58991CF1DB75BA84" ...
 $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike"
...
 $ started_at        : chr  "2020-05-27 10:03:52" "2020-05-25 10:47:11" "2020-05-02 1
4:11:03" "2020-05-02 16:25:36" ...
 $ ended_at          : chr  "2020-05-27 10:16:49" "2020-05-25 11:05:40" "2020-05-02 1
5:48:21" "2020-05-02 16:39:28" ...
 $ start_station_name: chr  "Franklin St & Jackson Blvd" "Clark St & Wrightwood Ave"
"Kedzie Ave & Milwaukee Ave" "Clarendon Ave & Leland Ave" ...
 $ start_station_id  : int  36 340 260 251 261 206 261 180 331 219 ...
 $ end_station_name  : chr  "Wabash Ave & Grand Ave" "Clark St & Leland Ave" "Kedzie
Ave & Milwaukee Ave" "Lake Shore Dr & Wellington Ave" ...
 $ end_station_id    : int  199 326 260 157 206 22 261 180 300 305 ...
 $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 42 41.9 41.9 41.8 ...
 $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202006)
```

```
'data.frame':   343005 obs. of  13 variables:
 $ ride_id           : chr  "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D14B0B5659BCF"
"C41237B506E85FA1" ...
 $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike"
...
 $ started_at        : chr  "2020-06-13 23:24:48" "2020-06-26 07:26:10" "2020-06-23 1
7:12:41" "2020-06-20 01:09:35" ...
 $ ended_at          : chr  "2020-06-13 23:36:55" "2020-06-26 07:31:58" "2020-06-23 1
7:21:14" "2020-06-20 01:28:24" ...
 $ start_station_name: chr  "Wilton Ave & Belmont Ave" "Federal St & Polk St" "Daley
Center Plaza" "Broadway & Cornelia Ave" ...
 $ start_station_id  : int  117 41 81 303 327 327 41 115 338 84 ...
 $ end_station_name  : chr  "Damen Ave & Clybourn Ave" "Daley Center Plaza" "State St
& Harrison St" "Broadway & Berwyn Ave" ...
 $ end_station_id    : int  163 81 5 294 117 117 81 303 164 53 ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
 $ end_lng           : num  -87.7 -87.6 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "member" "member" "casual" ...
```

Hide

```
str(tripdata_202107)
```

```
'data.frame':   822410 obs. of  13 variables:
 $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A"
"379B58EAB20E8AA5" ...
 $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bik
e" ...
 $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 1
1:30:55" "2021-07-08 22:08:30" ...
 $ ended_at          : chr  "2021-07-02 15:19:58" "2021-07-07 17:16:09" "2021-07-25 1
1:48:45" "2021-07-08 22:23:32" ...
 $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez S
t" "Wabash Ave & 16th St" "California Ave & Cortez St" ...
 $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
 $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Ru
sh St & Hubbard St" "Carpenter St & Huron St" ...
 $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

Hide

```
str(tripdata_202008)
```

```
'data.frame':   622361 obs. of  13 variables:
 $ ride_id           : chr  "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816"
"C79FBBD412E578A7" ...
 $ rideable_type     : chr  "docked_bike" "electric_bike" "electric_bike" "electric_b
ike" ...
 $ started_at        : chr  "2020-08-20 18:08:14" "2020-08-27 18:46:04" "2020-08-26 1
9:44:14" "2020-08-27 12:05:41" ...
 $ ended_at          : chr  "2020-08-20 18:17:51" "2020-08-27 19:54:51" "2020-08-26 2
1:53:07" "2020-08-27 12:53:45" ...
 $ start_station_name: chr  "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St"
"Columbus Dr & Randolph St" "Daley Center Plaza" ...
 $ start_station_id  : int  329 168 195 81 658 658 196 67 153 177 ...
 $ end_station_name  : chr  "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State
St & Randolph St" "State St & Kinzie St" ...
 $ end_station_id    : int  141 168 44 47 658 658 49 229 225 305 ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
 $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202009)
```

```
'data.frame':   532958 obs. of  13 variables:
 $ ride_id           : chr  "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E"
"57F6DC9A153DB98C" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2020-09-17 14:27:11" "2020-09-17 15:07:31" "2020-09-17 1
5:09:04" "2020-09-17 18:10:46" ...
 $ ended_at          : chr  "2020-09-17 14:44:24" "2020-09-17 15:07:45" "2020-09-17 1
5:09:35" "2020-09-17 18:35:49" ...
 $ start_station_name: chr  "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W
Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...
 $ start_station_id  : int  52 NA NA 246 24 94 291 NA NA NA ...
 $ end_station_name  : chr  "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W
Oakdale Ave & N Broadway" "Montrose Harbor" ...
 $ end_station_id    : int  112 NA NA 249 24 NA 256 NA NA NA ...
 $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
 $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202010)
```

```
'data.frame':   388653 obs. of  13 variables:
 $ ride_id           : chr  "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF"
"44A4AEE261B9E854" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2020-10-31 19:39:43" "2020-10-31 23:50:08" "2020-10-31 2
3:00:01" "2020-10-31 22:16:43" ...
 $ ended_at          : chr  "2020-10-31 19:57:12" "2020-11-01 00:04:16" "2020-10-31 2
3:08:22" "2020-10-31 22:19:35" ...
 $ start_station_name: chr  "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland
Ave" "Stony Island Ave & 67th St" "Clark St & Grace St" ...
 $ start_station_id  : int  313 227 102 165 190 359 313 125 NA 174 ...
 $ end_station_name  : chr  "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Univ
ersity Ave & 57th St" "Broadway & Sheridan Rd" ...
 $ end_station_id    : int  125 260 423 256 185 53 125 313 199 635 ...
 $ start_lat         : num  41.9 41.9 41.8 42 41.9 ...
 $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.8 42 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202011)
```

```
'data.frame':   259716 obs. of  13 variables:
 $ ride_id           : chr  "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5"
"E533E89C32080B9E" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2020-11-01 13:36:00" "2020-11-01 10:03:26" "2020-11-01 0
0:34:05" "2020-11-01 00:45:16" ...
 $ ended_at          : chr  "2020-11-01 13:45:40" "2020-11-01 10:14:45" "2020-11-01 0
1:03:06" "2020-11-01 00:54:31" ...
 $ start_station_name: chr  "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake
Shore Dr & Monroe St" "Leavitt St & Chicago Ave" ...
 $ start_station_id  : int  110 672 76 659 2 72 76 NA 58 394 ...
 $ end_station_name  : chr  "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Fede
ral St & Polk St" "Stave St & Armitage Ave" ...
 $ end_station_id    : int  211 29 41 185 2 76 72 NA 288 273 ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.6 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202012)
```

```
'data.frame':   131573 obs. of  13 variables:
 $ ride_id           : chr  "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A"
"BE119628E44F871E" ...
 $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_
bike" ...
 $ started_at        : chr  "2020-12-27 12:44:29" "2020-12-18 17:37:15" "2020-12-15 1
5:04:33" "2020-12-15 15:54:18" ...
 $ ended_at          : chr  "2020-12-27 12:55:06" "2020-12-18 17:44:19" "2020-12-15 1
5:11:28" "2020-12-15 16:00:11" ...
 $ start_station_name: chr  "Aberdeen St & Jackson Blvd" "" "" "" ...
 $ start_station_id  : chr  "13157" "" "" "" ...
 $ end_station_name  : chr  "Desplaines St & Kinzie St" "" "" "" ...
 $ end_station_id    : chr  "TA1306000003" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.8 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.8 ...
 $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

```
str(tripdata_202101)
```

```
'data.frame':   96834 obs. of  13 variables:
 $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27"
"4FA453A75AE377DB" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 2
2:35:54" "2021-01-07 13:31:13" ...
 $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 2
2:37:14" "2021-01-07 13:42:55" ...
 $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St"
"California Ave & Cortez St" "California Ave & Cortez St" ...
 $ start_station_id  : chr  "17660" "17660" "17660" "17660" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

```
str(tripdata_202102)
```

```
'data.frame':   49622 obs. of  13 variables:
 $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91"
"B32D3199F1C2E75B" ...
 $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bi
ke" ...
 $ started_at        : chr  "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 1
9:10:18" "2021-02-02 17:49:41" ...
 $ ended_at          : chr  "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 1
9:19:10" "2021-02-02 17:54:06" ...
 $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Cl
ark St & Lake St" "Wood St & Chicago Ave" ...
 $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
 $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St"
"State St & Randolph St" "Honore St & Division St" ...
 $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
 $ start_lat         : num  42 42 41.9 41.9 41.8 ...
 $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  42 42 41.9 41.9 41.8 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ member_casual     : chr  "member" "casual" "member" "member" ...
```

Hide

```
str(tripdata_202103)
```

```
'data.frame':   228496 obs. of  13 variables:
 $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284"
"994D05AA75A168F2" ...
 $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bik
e" ...
 $ started_at        : chr  "2021-03-16 08:32:30" "2021-03-28 01:26:28" "2021-03-11 2
1:17:29" "2021-03-11 13:26:42" ...
 $ ended_at          : chr  "2021-03-16 08:36:34" "2021-03-28 01:36:55" "2021-03-11 2
1:33:53" "2021-03-11 13:55:41" ...
 $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage
Ave" "Shields Ave & 28th Pl" "Winthrop Ave & Lawrence Ave" ...
 $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
 $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdal
e Ave" "Halsted St & 35th St" "Broadway & Sheridan Rd" ...
 $ end_station_id    : chr  "13266" "18017" "TA1308000043" "13323" ...
 $ start_lat         : num  41.9 41.9 41.8 42 42 ...
 $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.8 42 42.1 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202104)
```

```
'data.frame':   337230 obs. of  13 variables:
 $ ride_id           : chr  "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD"
"1887262AD101C604" ...
 $ rideable_type     : chr  "classic_bike" "docked_bike" "docked_bike" "classic_bike"
...
 $ started_at        : chr  "2021-04-12 18:25:36" "2021-04-27 17:27:11" "2021-04-03 1
2:42:45" "2021-04-17 09:17:42" ...
 $ ended_at          : chr  "2021-04-12 18:56:55" "2021-04-27 18:31:29" "2021-04-07 1
1:40:24" "2021-04-17 09:42:48" ...
 $ start_station_name: chr  "State St & Pearson St" "Dorchester Ave & 49th St" "Loomi
s Blvd & 84th St" "Honore St & Division St" ...
 $ start_station_id  : chr  "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
 $ end_station_name  : chr  "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St"
"Loomis Blvd & 84th St" "Southport Ave & Waveland Ave" ...
 $ end_station_id    : chr  "13235" "KA1503000069" "20121" "13235" ...
 $ start_lat         : num  41.9 41.8 41.7 41.9 41.7 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.8 41.7 41.9 41.7 ...
 $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "casual" "casual" "member" ...
```

Hide

```
str(tripdata_202105)
```

```
'data.frame':   531633 obs. of  13 variables:
 $ ride_id           : chr  "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2"
"7881AC6D39110C60" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2021-05-30 11:58:15" "2021-05-30 11:29:14" "2021-05-30 1
4:24:01" "2021-05-30 14:25:51" ...
 $ ended_at          : chr  "2021-05-30 12:10:39" "2021-05-30 12:14:09" "2021-05-30 1
4:25:13" "2021-05-30 14:41:04" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.8 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

Hide

```
str(tripdata_202106)
```

```
'data.frame':   729595 obs. of  13 variables:
 $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2"
"B03C0FE48C412214" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric
_bike" ...
 $ started_at        : chr  "2021-06-13 14:31:28" "2021-06-04 11:18:02" "2021-06-04 0
9:49:35" "2021-06-03 19:56:05" ...
 $ ended_at          : chr  "2021-06-13 14:34:11" "2021-06-04 11:24:19" "2021-06-04 0
9:55:34" "2021-06-03 20:21:55" ...
 $ start_station_name: chr  "" "" "" "" ...
 $ start_station_id  : chr  "" "" "" "" ...
 $ end_station_name  : chr  "" "" "" "" ...
 $ end_station_id    : chr  "" "" "" "" ...
 $ start_lat         : num  41.8 41.8 41.8 41.8 41.8 ...
 $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
 $ end_lat           : num  41.8 41.8 41.8 41.8 41.8 ...
 $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Hide

```
str(tripdata_202107)
```

```
'data.frame':   822410 obs. of  13 variables:
 $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A"
"379B58EAB20E8AA5" ...
 $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bik
e" ...
 $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 1
1:30:55" "2021-07-08 22:08:30" ...
 $ ended_at          : chr  "2021-07-02 15:19:58" "2021-07-07 17:16:09" "2021-07-25 1
1:48:45" "2021-07-08 22:23:32" ...
 $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez S
t" "Wabash Ave & 16th St" "California Ave & Cortez St" ...
 $ start_station_id  : chr  "13001" "17660" "SL-012" "17660" ...
 $ end_station_name  : chr  "Halsted St & North Branch St" "Wood St & Hubbard St" "Ru
sh St & Hubbard St" "Carpenter St & Huron St" ...
 $ end_station_id    : chr  "KA1504000117" "13432" "KA1503000044" "13196" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
 $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "casual" "casual" "member" "member" ...
```

Data transformation and cleaning

start_station_id & end_station_id are not consistent in all datasets. The ones in tripdata_202004, tripdata_202005, tripdata_202006, tripdata_202008, tripdata_202009, tripdata_202010, tripdata_202011 are int vs. the others are char. Convert the inconsistent ones from int to char datatype.

Hide

```
tripdata_202004 <- tripdata_202004 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202005 <- tripdata_202005 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202006 <- tripdata_202006 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202008 <- tripdata_202008 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202009 <- tripdata_202009 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202010 <- tripdata_202010 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
tripdata_202011 <- tripdata_202011 %>% mutate(start_station_id = as.character(start_s
tation_id), end_station_id = as.character(end_station_id))
```

3. Process

Combine all the datasets into one single dataframe

Hide

```
all_trips <- bind_rows(tripdata_202004,tripdata_202005,tripdata_202006,tripdata_20210
7,tripdata_202008,tripdata_202009,tripdata_202010,tripdata_202011,tripdata_202012,
                tripdata_202101,tripdata_202102,tripdata_202103,tripdata_202104,tripd
ata_202105,tripdata_202106,tripdata_202107)
str(all_trips)
```

```
'data.frame':   6181546 obs. of  13 variables:
 $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4"
"2A59BBDF5CDBA725" ...
 $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike"
...
 $ started_at        : chr  "2020-04-26 17:45:14" "2020-04-17 17:08:54" "2020-04-01 1
7:54:13" "2020-04-07 12:50:19" ...
 $ ended_at          : chr  "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 1
8:08:36" "2020-04-07 13:02:31" ...
 $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct &
Erie St" "California Ave & Division St" ...
 $ start_station_id  : chr  "86" "503" "142" "216" ...
 $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana
Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
 $ end_station_id    : chr  "152" "499" "255" "657" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

Clean-up further!

Hold on! started_at & ended_at should be in datetime datatype instead of char. Convert all from char to datetime.

Hide

```
all_trips[['started_at']] <- ymd_hms(all_trips[['started_at']])
all_trips[['ended_at']] <- ymd_hms(all_trips[['ended_at']])

str(all_trips)
```

```
'data.frame':   6181546 obs. of  13 variables:
 $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4"
"2A59BBDF5CDBA725" ...
 $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike"
...
 $ started_at        : POSIXct, format: "2020-04-26 17:45:14" "2020-04-17 17:08:54"
"2020-04-01 17:54:13" ...
 $ ended_at          : POSIXct, format: "2020-04-26 18:12:03" "2020-04-17 17:17:03"
"2020-04-01 18:08:36" ...
 $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct &
Erie St" "California Ave & Division St" ...
 $ start_station_id  : chr  "86" "503" "142" "216" ...
 $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana
Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
 $ end_station_id    : chr  "152" "499" "255" "657" ...
 $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
 $ end_lat           : num  41.9 41.9 41.9 41.9 42 ...
 $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
 $ member_casual     : chr  "member" "member" "member" "member" ...
```

All looks good!

Remove columns not required or beyond the scope of project

Hide

```
all_trips <- all_trips %>%
    select(-c(start_lat:end_lng))
glimpse(all_trips)
```

```
Rows: 6,181,546
Columns: 9
$ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4E006F
4", "2A59BBDF5CDBA725", "…
$ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docked_bik
e", "docked_bike", "docke…
$ started_at         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-01 17:5
4:13, 2020-04-07 12:50:…
$ ended_at           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04-01 18:0
8:36, 2020-04-07 13:02:…
$ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClurg Ct &
Erie St", "California …
$ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434", "62
7", "377", "508", "374",…
$ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "Indiana
Ave & Roosevelt Rd", "…
$ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "382", "35
9", "508", "374", "128"…
$ member_casual      <chr> "member", "member", "member", "member", "casual", "membe
r", "member", "casual", "…
```

Rename columns for better readability

Hide

```
all_trips <- all_trips %>%
    rename(ride_type = rideable_type,
           start_time = started_at,
           end_time = ended_at,
           customer_type = member_casual)
glimpse(all_trips)
```

```
Rows: 6,181,546
Columns: 9
$ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4E006F
4", "2A59BBDF5CDBA725", "…
$ ride_type          <chr> "docked_bike", "docked_bike", "docked_bike", "docked_bik
e", "docked_bike", "docke…
$ start_time         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-01 17:5
4:13, 2020-04-07 12:50:…
$ end_time           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04-01 18:0
8:36, 2020-04-07 13:02:…
$ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClurg Ct &
Erie St", "California …
$ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434", "62
7", "377", "508", "374",…
$ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "Indiana
Ave & Roosevelt Rd", "…
$ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "382", "35
9", "508", "374", "128"…
$ customer_type      <chr> "member", "member", "member", "member", "casual", "membe
r", "member", "casual", "…
```

Add new columns that can be used for aggregate functions

Hide

```
#column for day of the week the trip started
all_trips$day_of_the_week <- format(as.Date(all_trips$start_time),'%a')

#column for month when the trip started
all_trips$month <- format(as.Date(all_trips$start_time),'%b_%y')

#column for time of the day when the trip started
#Time element needs to be extracted from start_time. However, as the times must be in POSIXct
#(only times of class POSIXct are supported in ggplot2), a two-step conversion is needed.
#First the time is converted to a character vector, effectively stripping all the date information.
#The time is then converted back to POSIXct with today's date – the date is of no interest to us,
#only the hours-minutes-seconds are.
all_trips$time <- format(all_trips$start_time, format = "%H:%M")
all_trips$time <- as.POSIXct(all_trips$time, format = "%H:%M")

#column for trip duration in min
all_trips$trip_duration <- (as.double(difftime(all_trips$end_time, all_trips$start_time)))/60

# check the dataframe
glimpse(all_trips)
```

```
Rows: 6,181,546
Columns: 13
$ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4E006F4", "2A59BBDF5CDBA725", "…
$ ride_type          <chr> "docked_bike", "docked_bike", "docked_bike", "docked_bike", "docked_bike", "docke…
$ start_time         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-01 17:54:13, 2020-04-07 12:50:…
$ end_time           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04-01 18:08:36, 2020-04-07 13:02:…
$ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClurg Ct & Erie St", "California …
$ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434", "627", "377", "508", "374",…
$ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "Indiana Ave & Roosevelt Rd", "…
$ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "382", "359", "508", "374", "128"…
$ customer_type      <chr> "member", "member", "member", "member", "casual", "member", "member", "casual", "…
$ day_of_the_week    <chr> "Sun", "Fri", "Wed", "Tue", "Sat", "Thu", "Thu", "Tue", "Wed", "Sat", "Sat", "Sat…
$ month              <chr> "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_20", "…
$ time               <dttm> 2022-01-26 17:45:00, 2022-01-26 17:08:00, 2022-01-26 17:54:00, 2022-01-26 12:50:…
$ trip_duration      <dbl> 26.816667, 8.150000, 14.383333, 12.200000, 52.916667, 5.400000, 5.216667, 75.8166…
```

Let's check to see if the trip_duration column has any negative values, as this may cause problem while creating visualizations. Also, we do not want to include the trips that were part of quality tests by the company. These trips are usually identified by string 'test' in the start_station_name column.

Hide

```
# checking for trip lengths less than 0
nrow(subset(all_trips,trip_duration < 0))
```

```
[1] 8845
```

Hide

```
#checking for testrides that were made by company for quality checks
nrow(subset(all_trips, start_station_name %like% "TEST"))
```

```
[1] 3220
```

Hide

```
nrow(subset(all_trips, start_station_name %like% "test"))
```

```
[1] 0
```

Hide

```
nrow(subset(all_trips, start_station_name %like% "Test"))
```

```
[1] 0
```

As there are 8845 rows with trip_dration less than 0 mins and 3220 trips that were test rides, we will remove these observations from our dataframe as they contribute to only about 0.3% of the total rows. We will create a new dataframe deviod of these obseravtions without making any changes to the existing dataframe.

Hide

```
# remove negative trip durations
all_trips_v2 <- all_trips[!(all_trips$trip_duration < 0),]

#remove test rides
all_trips_v2<- all_trips_v2[!((all_trips_v2$start_station_name %like% "TEST" | all_tr
ips_v2$start_station_name %like% "test")),]


#check dataframe
glimpse(all_trips_v2)
```

```
Rows: 6,169,494
Columns: 13
$ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4E006F
4", "2A59BBDF5CDBA725", "…
$ ride_type          <chr> "docked_bike", "docked_bike", "docked_bike", "docked_bik
e", "docked_bike", "docke…
$ start_time         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-01 17:5
4:13, 2020-04-07 12:50:…
$ end_time           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04-01 18:0
8:36, 2020-04-07 13:02:…
$ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClurg Ct &
Erie St", "California …
$ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434", "62
7", "377", "508", "374",…
$ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "Indiana
Ave & Roosevelt Rd", "…
$ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "382", "35
9", "508", "374", "128"…
$ customer_type      <chr> "member", "member", "member", "member", "casual", "membe
r", "member", "casual", "…
$ day_of_the_week    <chr> "Sun", "Fri", "Wed", "Tue", "Sat", "Thu", "Thu", "Tue", "W
ed", "Sat", "Sat", "Sat…
$ month              <chr> "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_20", "Apr_2
0", "Apr_20", "Apr_20", "…
$ time               <dttm> 2022-01-26 17:45:00, 2022-01-26 17:08:00, 2022-01-26 17:5
4:00, 2022-01-26 12:50:…
$ trip_duration      <dbl> 26.816667, 8.150000, 14.383333, 12.200000, 52.916667, 5.40
0000, 5.216667, 75.8166…
```

It is important to make sure that customer_type column has only two distinct values. Let's confirm the same.

Hide

```
# checking count of distinct values
table(all_trips_v2$customer_type)
```

```
 casual  member
2803618 3365876
```

Hide

```
#aggregating total trip duration by customer type
setNames(aggregate(trip_duration ~ customer_type, all_trips_v2, sum), c("customer_typ
e", "total_trip_duration(mins)"))
```

customer_type	total_trip_duration(mins)
<chr>	<dbl>
casual	105835525
member	51144366

2 rows

4&5. Analyze and Share the Data

The dataframe is now ready for descriptive analysis that will help us uncover some insights on how the casual riders and members use Cyclistic rideshare differently.

First, let's try to get some simple statistics on trip_duration for all customers, and do the same by customer_type.

Hide

```
# statictical summary of trip_duration for all trips
summary(all_trips_v2$trip_duration)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.00     7.58    13.67    25.44    24.88 58720.03
```

Hide

```
#statistical summary of trip_duration by customer_type
all_trips_v2 %>%
    group_by(customer_type) %>%
    summarise(min_trip_duration = min(trip_duration),max_trip_duration = max(trip_dur
ation),
              median_trip_duration = median(trip_duration), mean_trip_duration = mean
(trip_duration))
```

customer_type <chr>	min_trip_duration <dbl>	max_trip_duration <dbl>	median_trip_duration <dbl>	mean_t
casual	0	55944.15	18.43333	
member	0	58720.03	10.83333	

2 rows

The mean trip duration of member riders is lower than the mean trip duration of all trips, while it is exactly the opposite for casual riders, whose mean trip duration is higher than the the mean trip duration of all trips. This tells us that casual riders usually take the bikes out for a longer duration compared to members.

Total number of trips by customer type and day of the week

Hide

```
# fix the order for the day_of_the_week and month variable so that they show up
# in the same sequence in output tables and visualizations
all_trips_v2$day_of_the_week <- ordered(all_trips_v2$day_of_the_week, levels=c("Mon",
"Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))
all_trips_v2$month <- ordered(all_trips_v2$month, levels=c("Apr_20", "May_20", "Jun_2
0", "Jul_20", "Aug_20", "Sep_20", "Oct_20",
                                                           "Nov_20", "Dec_20", "Jan_2
1", "Feb_21", "Mar_21",
                                                            "Apr_21", "May_21", "Jun_2
1", "Jul_21"))
all_trips_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(number_of_rides = n(),average_duration_mins = mean(trip_duration)) %>%
  arrange(customer_type, desc(number_of_rides))
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```

customer_type <chr>	day_of_the_week <ord>	number_of_rides <int>	average_duration_mins <dbl>
casual	Sat	656413	40.04901
casual	Sun	514661	43.17128
casual	Fri	408144	35.77946
casual	Thu	318956	34.39917
casual	Mon	304957	37.94956
casual	Wed	304118	33.35507
casual	Tue	296369	33.86455
member	Sat	523719	16.87145
member	Fri	504817	14.80364
member	Wed	498728	14.36992

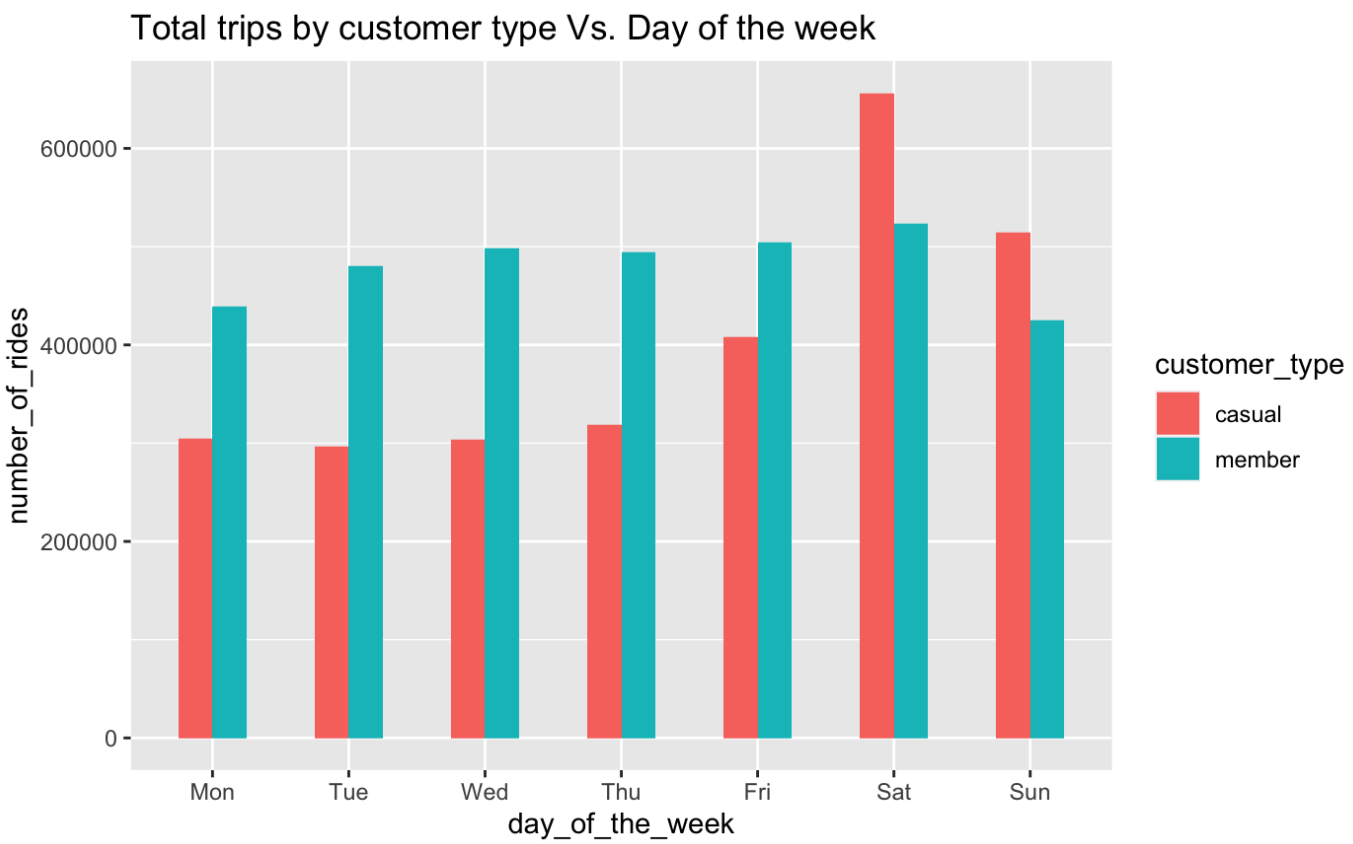
1-10 of 14 rows Previous **1** 2 Next

Visualization:

Hide

```
all_trips_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(number_of_rides = n()) %>%
  arrange(customer_type, day_of_the_week)  %>%
  ggplot(aes(x = day_of_the_week, y = number_of_rides, fill = customer_type)) +
  labs(title ="Total trips by customer type Vs. Day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```



From the table and graph above, casual customers are most busy on Sundays followed by Saturdays, while members are most busy on later half of the week extending into the weekend. Interesting pattern to note though is the consistent trip numbers among members with less spread over entire week as compared to casual riders who don't seem to use the bikeshare services much during weekdays.

Average number of trips by customer type and month

Hide

```
unique(all_trips$month)
```

```
 [1] "Apr_20" "May_20" "Jun_20" "Jul_21" "Aug_20" "Sep_20" "Oct_20" "Nov_20" "Dec_20"
"Jan_21" "Feb_21"
[12] "Mar_21" "Apr_21" "May_21" "Jun_21"
```

Hide

```
all_trips_v2 %>%
  group_by(customer_type, month) %>%
  summarise(number_of_rides = n(),`average_duration_(mins)` = mean(trip_duration)) %
>%
  arrange(customer_type,desc(number_of_rides))
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```

customer_type <chr>	month <ord>	number_of_rides <int>	average_duration_(mins) <dbl>
casual	Jul_21	884096	32.79077
casual	Jun_21	370678	37.12174
casual	Aug_20	287171	45.16081
casual	May_21	256916	38.23097
casual	Sep_20	229435	38.32516
casual	Jun_20	154401	51.71658
casual	Oct_20	143850	30.39722
casual	Apr_21	136601	38.02299
casual	Nov_20	87810	31.85596
casual	May_20	86786	51.25173

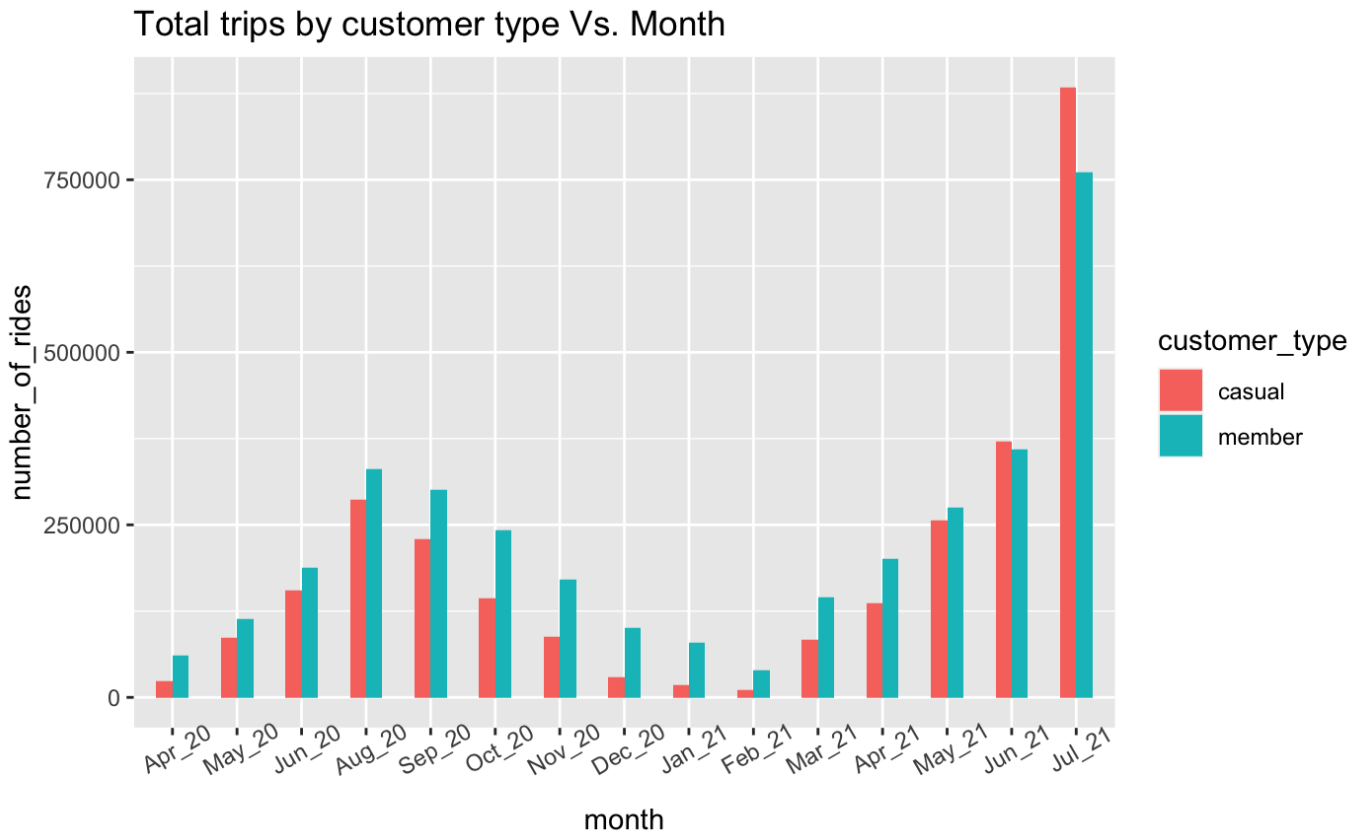
1-10 of 30 rows Previous **1** 2 3 Next

Visualization:

Hide

```
all_trips_v2 %>%
  group_by(customer_type, month) %>%
  summarise(number_of_rides = n()) %>%
  arrange(customer_type, month)  %>%
  ggplot(aes(x = month, y = number_of_rides, fill = customer_type)) +
  labs(title ="Total trips by customer type Vs. Month") +
  theme(axis.text.x = element_text(angle = 30)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```



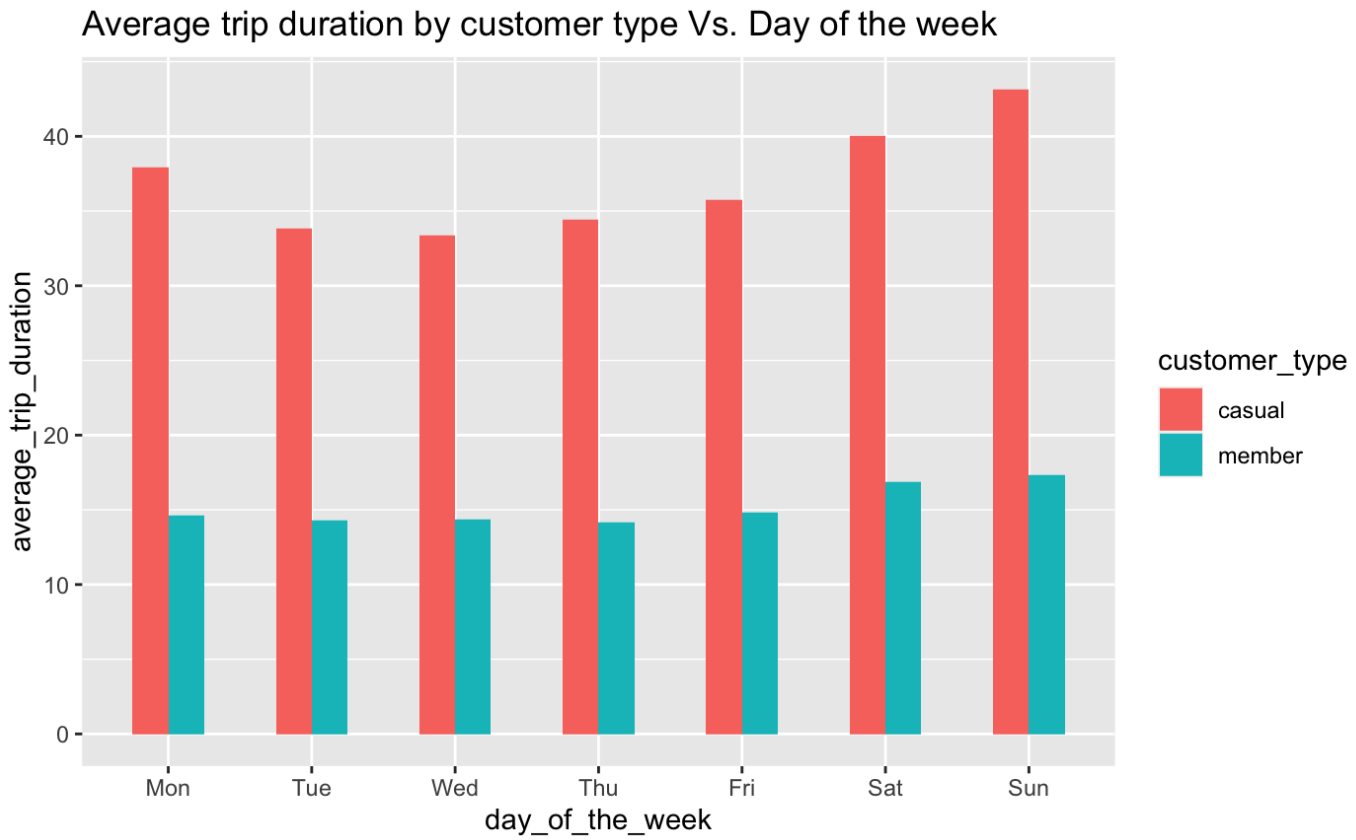
The data shows that the months of July, August and September are the most busy time of the year among both members and casual riders. This could be attributed to an external factor (eg. cold weather, major quality issue) that might have hindered with customer needs. 2021 is a tough year when Covid comes. People care more about their health. The charts shows that the no.of rides in 2021 is higher than 2020 in general. However, the number of trips made by members is always higher than the casual riders across all months of the year.

Visualizaton of average trip duration by customer type on each day of the week

Hide

```
all_trips_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(average_trip_duration = mean(trip_duration)) %>%
  ggplot(aes(x = day_of_the_week, y = average_trip_duration, fill = customer_type)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title ="Average trip duration by customer type Vs. Day of the week")
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```



Hide

```
NA
```

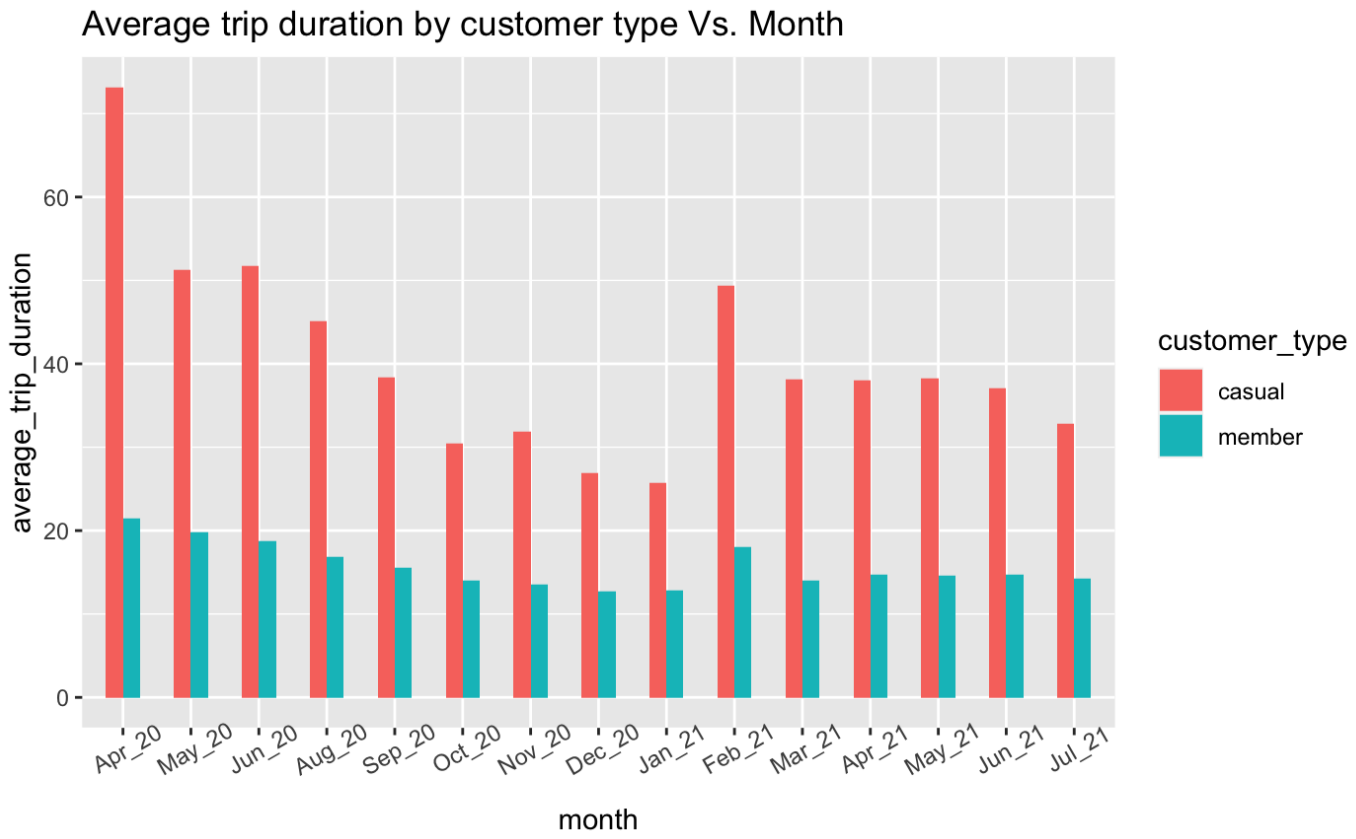
The average trip duration of a casual rider is more than twice that of a member. Note that this necessarily does not mean that casual riders travel farther distance. It is also interesting to note that weekends not only contribute to more number of trips but also longer trips on average when compared to weekdays.

Visualizaton of average trip duration by customer type Vs. month

Hide

```
all_trips_v2 %>%
  group_by(customer_type, month) %>%
  summarise(average_trip_duration = mean(trip_duration)) %>%
  ggplot(aes(x = month, y = average_trip_duration, fill = customer_type)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title ="Average trip duration by customer type Vs. Month") +
  theme(axis.text.x = element_text(angle = 30))
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```



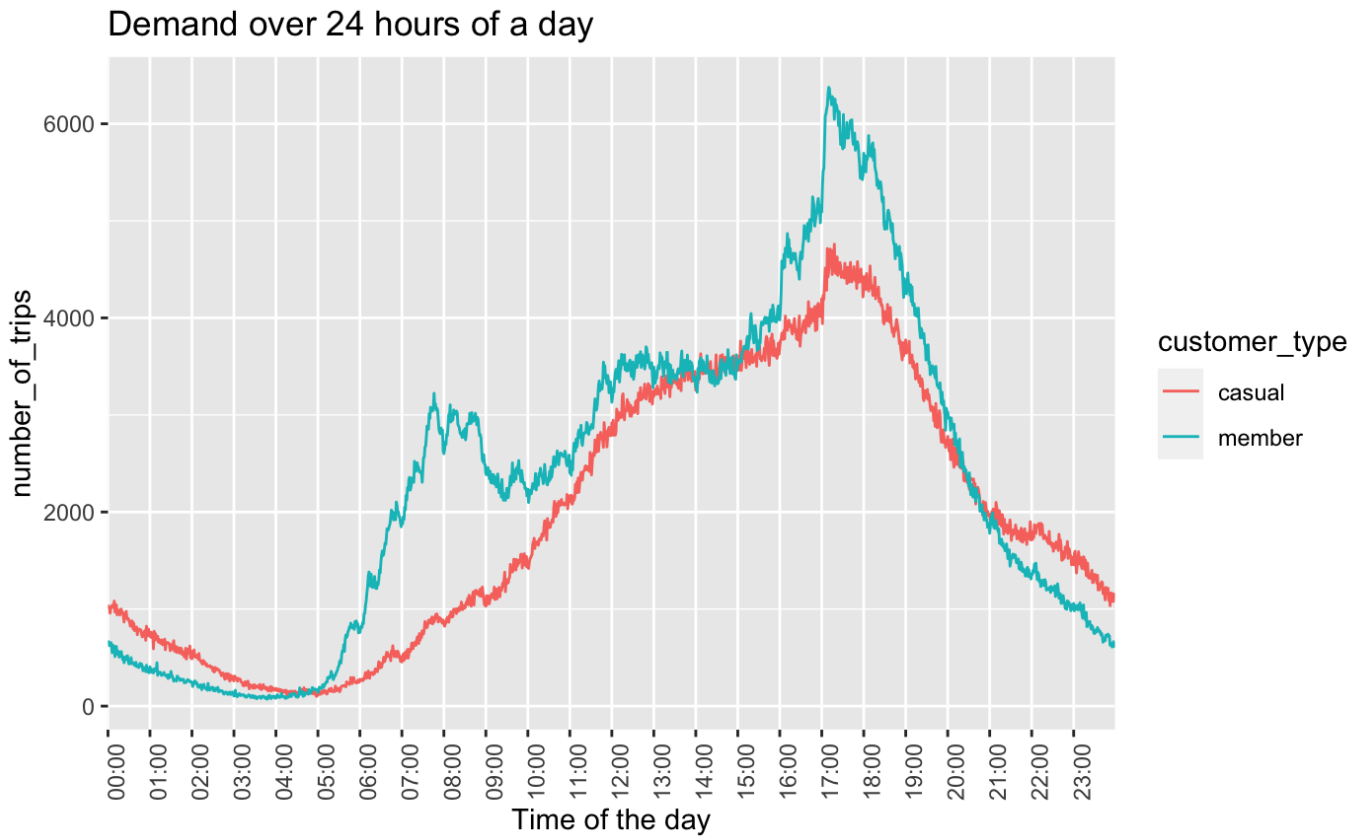
Average trip duration of member riders is anywhere between 10-20 minutes throughout the year, exception being April when it goes slightly over 20 minutes. However, there seems to be a distinct pattern when it comes to casual riders, whose average trip duration swings wildly from as low as ~25 minutes to more than an hour depending on time of the year. It is worth noting unusually long trip durations by casual riders in the month of April.

Visualizaton of bike demand over 24 hr period (a day)

Hide

```
all_trips_v2 %>%
  group_by(customer_type, time) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x = time, y = number_of_trips, color = customer_type, group = customer_t
ype)) +
  geom_line() +
  scale_x_datetime(date_breaks = "1 hour", minor_breaks = NULL,
                   date_labels = "%H:%M", expand = c(0,0)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title ="Demand over 24 hours of a day", x = "Time of the day")
```

```
`summarise()` has grouped output by 'customer_type'. You can override using the `.gro
ups` argument.
```



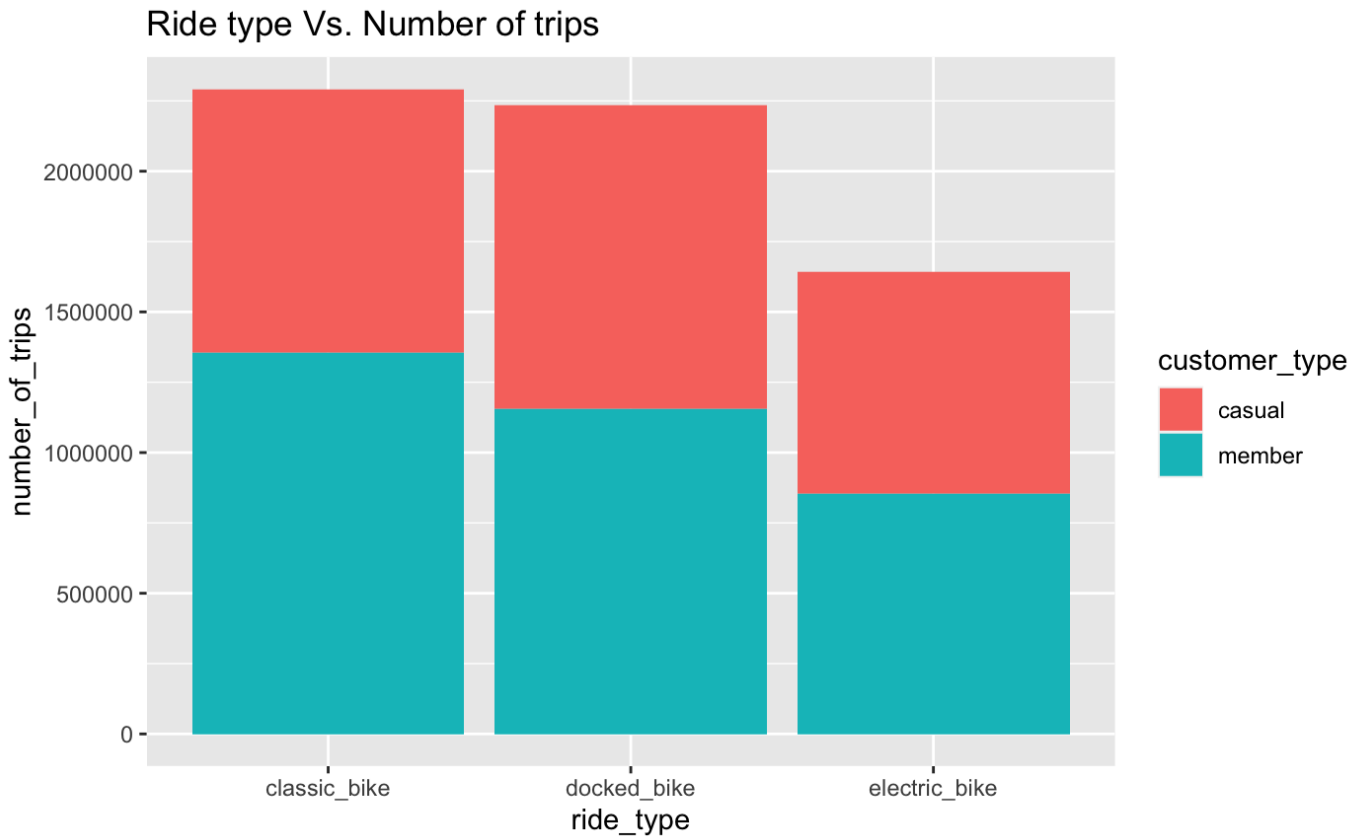
For the members, there seems to be two distict peak demand hours: 7-9 AM and 5-7 PM, the latter one coinciding with the peak demand hours of casual riders as well. One could probably hypothesize that office-goers make up majority of the members profile due to demand in both morning and evening hours, but we need more data to substabtiate this assumption.

Visualizaton of ride type Vs. number of trips by customer type

Hide

```
all_trips_v2 %>%
  group_by(ride_type, customer_type) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x= ride_type, y=number_of_trips, fill= customer_type))+
              geom_bar(stat='identity') +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  labs(title ="Ride type Vs. Number of trips")
```

```
`summarise()` has grouped output by 'ride_type'. You can override using the `.groups`
argument.
```



Classic bikes are predominantly used by members. Docked bikes are in most demand and equally used by both members as well as casual riders. Electric bikes are more favored by members. If electric bikes costs the highest among all 3 types, it would be a financially sound move to increase their fleet while reducing docked bikes, as they are already preferred by members who make up for the majority of the trips.

Note: Data is not available on the quantity of fleet across each type of bikes.

Creating a csv file of the clean data for futher analysis or visualizations in other tools like SQL, Tableau, Power BI, etc.

Hide

```
clean_data <- aggregate(all_trips_v2$trip_duration ~ all_trips_v2$customer_type + all
_trips_v2$day_of_the_week, FUN = mean)
write.csv(clean_data, "Clean Data.csv", row.names = F)
```

6. Act

Key Takeaways

- Casual riders made 41% of total trips contributing to 66% of total trip duration between Apr'20 - Mar'21. Member riders make up 59% of total trips contributing to 34% of total trip duration between Apr'20 - Mar'21

  Usage (based on trip duration) of bikes by casual riders is almost twice that of member riders.

- Casual customers use bikeshare services more during weekends, while members use them consistently over the entire week.

- Average trip duration of casual riders is more than twice that of member rider over any given day of the week cumulatively.

- Casual riders ride longer during first half of the year compared to the second half, while members clock relatively similar average trip duration month over month.
- Casual riders prefer docked bikes the most while classic bikes are popular among members.

Recommendations

- Provide attractive promotions for casual riders on weekdays so that casual members use the bikeshare services ore uniformly across the entire week.
- Offer discounted membership fee for renewals after the first year. It might nudge casual riders to take up membership.
- Offer discounted pricing during non-busy hours so that casual riders might choose to use bikes more often and level out demand over the day.

Additonal data that could expand scope of analysis

- Occupation of member riders - this data could be used to target non-members who come under similar occupation
- Age and gender profile - Again, this data could be used to study the category of riders who can be targeted for attracting new members.
- Pricing details for members and casual riders - Based on this data, we might be to optimize cost structure for casual riders or provide discounts without affecting the profit margin.
- Address/ neighborhood details of members to investigate if there are any location specific parameters that encourage membership.

——End of case study——