How Does a Bike-Share Navigate Speedy Success?

Oluwanifemi Aweda

2022-07-05

Introduction.

The analysis in this notebook was done to fulfill the requirements for getting the google data analytics certification hosted on coursera. The case study involves a bikeshare company's data of its customer's trip details over a 12 month period (June 2021 - May 2022). The data (https://divvy-tripdata.s3.amazonaws.com/index.html) has been made available by Motivate International Inc. under this license (https://ride.divvybikes.com/data-license-agreement).

The analysis follows the 6 phases of the Data Analysis process as a guideline which includes the; Ask, Prepare, Process, Analyze, and Act phases.

The Ask Phase

I was tasked with the director of marketing to find the differentiating factor between casual riders and annual members with the aim of converting casual riders into annual members. **Stakeholders** * The marketing director. * The executive team. * Fellow data analysts. * Casual riders. **Deliverable** * Identify the differentiating factor between casual riders and annual members. * Provide effective visuals and relevant data to support insights gotten.

The Prepare Phase

The data used in the analysis is a secondary data obtained from here (https://divvy-tripdata.s3.amazonaws.com/index.html). The dataset is a publicly available one that holds the records of different users of Cyclistic. The data has been made available by Motivate International Inc. under this license (https://ride.divvybikes.com/data-license-agreement). The data gotten includes 12 months of data (202106-divvy-tripdata to 202205-divvy-tripdata). The data has 13 features (columns) with multiple entries (rows) for each ride taken. **The following were done during the prepare stage** * Downloading of the data. * Extraction of data. * Combining data into the same folder structure. * The consistency of the data was checked across all files (through the column names). * The data type consistency across all files was also checked (data structure).

Import the data set

```
tripdata_202106 <- read.csv("202106-divvy-tripdata.csv")
tripdata_202107 <- read.csv("202107-divvy-tripdata.csv")
tripdata_202108 <- read.csv("202108-divvy-tripdata.csv")
tripdata_202109 <- read.csv("202109-divvy-tripdata.csv")
tripdata_202110 <- read.csv("202110-divvy-tripdata.csv")
tripdata_202111 <- read.csv("202111-divvy-tripdata.csv")
tripdata_202112 <- read.csv("202112-divvy-tripdata.csv")
tripdata_202201 <- read.csv("202201-divvy-tripdata.csv")
tripdata_202202 <- read.csv("202202-divvy-tripdata.csv")
tripdata_202203 <- read.csv("202203-divvy-tripdata.csv")
tripdata_202204 <- read.csv("202204-divvy-tripdata.csv")
tripdata_202205 <- read.csv("202205-divvy-tripdata.csv")
```

Check the column names to know if the follow same pattern

```
colnames(tripdata_202106)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202107)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202108)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202109)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202110)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202111)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202112)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202201)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202202)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202203)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202204)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(tripdata_202205)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

Check the structure of the data in each column to make sure they align

```
str(tripdata_202106)
```

```
## 'data.frame':    729595 obs. of  13 variables:
##  $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C
0FE48C412214" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-06-13 14:31:28" "2021-06-04 11:18:02" "2021-06-04 09:49:
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 09:55:
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" ...
```

```
str(tripdata_202107)
```

```
## 'data.frame':    822410 obs. of  13 variables:
##  $ ride_id           : chr  "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B
58EAB20E8AA5" ...
##  $ rideable_type     : chr  "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-07-02 14:44:36" "2021-07-07 16:57:42" "2021-07-25 11:30:
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 11:48:
45" "2021-07-08 22:23:32" ...
##  $ start_station_name: chr  "Michigan Ave & Washington St" "California Ave & Cortez St" "W
abash 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" "Rush 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" ...
```

```
str(tripdata_202108)
```

```
## 'data.frame':    804352 obs. of  13 variables:
##  $ ride_id           : chr  "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834
D3208BFAF1DA" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-08-10 17:15:49" "2021-08-10 17:23:14" "2021-08-21 02:34:
23" "2021-08-21 06:52:55" ...
##  $ ended_at          : chr  "2021-08-10 17:22:44" "2021-08-10 17:39:24" "2021-08-21 02:50:
36" "2021-08-21 07:08:13" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(tripdata_202109)
```

```
## 'data.frame':    756147 obs. of  13 variables:
##  $ ride_id           : chr  "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1
DE133B3DBF55" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-09-28 16:07:10" "2021-09-28 14:24:51" "2021-09-28 00:20:
16" "2021-09-28 14:51:17" ...
##  $ ended_at          : chr  "2021-09-28 16:09:54" "2021-09-28 14:40:05" "2021-09-28 00:23:
57" "2021-09-28 15:00:06" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(tripdata_202110)
```

```
## 'data.frame':    631226 obs. of  13 variables:
##  $ ride_id           : chr  "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "3629
47F0437E1514" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-10-22 12:46:42" "2021-10-21 09:12:37" "2021-10-16 16:28:
39" "2021-10-16 16:17:48" ...
##  $ ended_at          : chr  "2021-10-22 12:49:50" "2021-10-21 09:14:14" "2021-10-16 16:36:
26" "2021-10-16 16:19:03" ...
##  $ start_station_name: chr  "Kingsbury St & Kinzie St" "" "" "" ...
##  $ start_station_id  : chr  "KA1503000043" "" "" "" ...
##  $ 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 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(tripdata_202111)
```

```
## 'data.frame':    359978 obs. of  13 variables:
##  $ ride_id           : chr  "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061C" "2F3B
E33085BCFF02" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-11-27 13:27:38" "2021-11-27 13:38:25" "2021-11-26 22:03:
34" "2021-11-27 09:56:49" ...
##  $ ended_at          : chr  "2021-11-27 13:46:38" "2021-11-27 13:56:10" "2021-11-26 22:05:
56" "2021-11-27 10:01:50" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 42 42 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ end_lat           : num  42 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.8 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(tripdata_202112)
```

```
## 'data.frame':    247540 obs. of  13 variables:
##  $ ride_id           : chr  "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278
BA87BF698339" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "classic_bike"
...
##  $ started_at        : chr  "2021-12-07 15:06:07" "2021-12-11 03:43:29" "2021-12-15 23:10:
28" "2021-12-26 16:16:10" ...
##  $ ended_at          : chr  "2021-12-07 15:13:42" "2021-12-11 04:10:23" "2021-12-15 23:23:
14" "2021-12-26 16:30:53" ...
##  $ start_station_name: chr  "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted St
& North Branch St" "Halsted St & North Branch St" ...
##  $ start_station_id  : chr  "13307" "KP1705001026" "KA1504000117" "KA1504000117" ...
##  $ end_station_name  : chr  "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Broadway &
Barry Ave" "LaSalle Dr & Huron St" ...
##  $ end_station_id    : chr  "TA1307000130" "TA1307000119" "13137" "KP1705001026" ...
##  $ 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 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```
str(tripdata_202201)
```

```
## 'data.frame':    103770 obs. of  13 variables:
##  $ ride_id           : chr  "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB8
0ED419105406" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "classic_bike" "classic_bike"
...
##  $ started_at        : chr  "2022-01-13 11:59:47" "2022-01-10 08:41:56" "2022-01-25 04:53:
40" "2022-01-04 00:18:04" ...
##  $ ended_at          : chr  "2022-01-13 12:02:44" "2022-01-10 08:46:17" "2022-01-25 04:58:
01" "2022-01-04 00:33:00" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffie
ld Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...
##  $ start_station_id  : chr  "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr  "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave &
Fullerton Ave" "Paulina St & Montrose Ave" ...
##  $ end_station_id    : chr  "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num  42 42 41.9 42 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 42 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "member" "casual" ...
```

```
str(tripdata_202202)
```

```
## 'data.frame':    115609 obs. of  13 variables:
##  $ ride_id           : chr  "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A178
9BDF844412BE" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike"
...
##  $ started_at        : chr  "2022-02-19 18:08:41" "2022-02-20 17:41:30" "2022-02-25 18:55:
56" "2022-02-14 11:57:03" ...
##  $ ended_at          : chr  "2022-02-19 18:23:56" "2022-02-20 17:45:56" "2022-02-25 19:09:
34" "2022-02-14 12:04:00" ...
##  $ start_station_name: chr  "State St & Randolph St" "Halsted St & Wrightwood Ave" "State
St & Randolph St" "Southport Ave & Waveland Ave" ...
##  $ start_station_id  : chr  "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...
##  $ end_station_name  : chr  "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Can
al St & Adams St" "Broadway & Sheridan Rd" ...
##  $ end_station_id    : chr  "13179" "TA1307000113" "13011" "13323" ...
##  $ 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 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(tripdata_202203)
```

```
## 'data.frame':    284042 obs. of  13 variables:
##  $ ride_id           : chr  "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F44
6FD9DEE3F389" ...
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "classic_bike" "classic_bike"
...
##  $ started_at        : chr  "2022-03-21 13:45:01" "2022-03-16 09:37:16" "2022-03-23 19:52:
02" "2022-03-01 19:12:26" ...
##  $ ended_at          : chr  "2022-03-21 13:51:18" "2022-03-16 09:43:34" "2022-03-23 19:54:
48" "2022-03-01 19:22:14" ...
##  $ start_station_name: chr  "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & B
erwyn Ave" "Wabash Ave & Wacker Pl" ...
##  $ start_station_id  : chr  "TA1307000131" "13042" "13109" "TA1307000131" ...
##  $ end_station_name  : chr  "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apt
s)" "Broadway & Ridge Ave" "Franklin St & Jackson Blvd" ...
##  $ end_station_id    : chr  "KA1503000043" "620" "15578" "TA1305000025" ...
##  $ start_lat         : num  41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(tripdata_202204)
```

```
## 'data.frame':    371249 obs. of  13 variables:
##  $ ride_id           : chr  "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F07FF" "84D4
751AEB31888D" ...
##  $ rideable_type     : chr  "electric_bike" "classic_bike" "classic_bike" "classic_bike"
...
##  $ started_at        : chr  "2022-04-06 17:42:48" "2022-04-24 19:23:07" "2022-04-20 19:29:
08" "2022-04-22 21:14:06" ...
##  $ ended_at          : chr  "2022-04-06 17:54:36" "2022-04-24 19:43:17" "2022-04-20 19:35:
16" "2022-04-22 21:23:29" ...
##  $ start_station_name: chr  "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted
St & Polk St" "Wentworth Ave & Cermak Rd" ...
##  $ start_station_id  : chr  "515" "13075" "TA1307000121" "13075" ...
##  $ end_station_name  : chr  "University Library (NU)" "Green St & Madison St" "Green St &
Madison St" "Delano Ct & Roosevelt Rd" ...
##  $ end_station_id    : chr  "605" "TA1307000120" "TA1307000120" "KA1706005007" ...
##  $ start_lat         : num  42 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  42.1 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "casual" ...
```

```
str(tripdata_202205)
```

```
## 'data.frame':    634858 obs. of  13 variables:
##  $ ride_id           : chr  "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF5
9852924528F8" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike"
...
##  $ started_at        : chr  "2022-05-23 23:06:58" "2022-05-11 08:53:28" "2022-05-26 18:36:
28" "2022-05-10 07:30:07" ...
##  $ ended_at          : chr  "2022-05-23 23:40:19" "2022-05-11 09:31:22" "2022-05-26 18:58:
18" "2022-05-10 07:38:49" ...
##  $ start_station_name: chr  "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St"
"Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr  "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...
##  $ end_station_name  : chr  "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood S
t & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr  "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

We can see that the data follows the same structure across the 12 months both in the naming of the columns and the data types that each of the column holds.

The Process Phase

The data processing stage helps to rectify identified issues with the data to be used and also interact to get a feel of the data. I will be using R for the analysis. The reason why I choose R for the analysis is because we have a lot of data and combining the data together is faster and simpler using a programming language. To ensure that the data is clean, I got myself familiar with the data and performed some operations as documented in the Deliverable section. After processing the data, I check over again to make sure the data is clean enough, making sure the data is correct, reliable and relevant to the business problem.

For the processing phase, each of the following activities were carried out. * Combine data sets to a single dataframe. * Rename columns for better readability. * Delete unnecessary columns. * Generate aggregate columns that can help bring insight to the analysis. * Perform some basic statistics to get familiar with the data.

Combine the 12 data sets into one

```
combined_data <- bind_rows(tripdata_202106, tripdata_202107, tripdata_202108, tripdata_20210
9,
                           tripdata_202110, tripdata_202111, tripdata_202112, tripdata_20220
1,
                           tripdata_202202, tripdata_202203, tripdata_202204, tripdata_20220
5)
colnames(combined_data)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
str(combined_data)
```

```
## 'data.frame':    5860776 obs. of  13 variables:
##  $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C
0FE48C412214" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ started_at        : chr  "2021-06-13 14:31:28" "2021-06-04 11:18:02" "2021-06-04 09:49:
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 09:55:
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" ...
```

Rename the columns for readability

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

```
## Rows: 5,860,776
## Columns: 13
## $ ride_id            <chr> "99FEC93BA843FB20", "06048DCFC8520CAF", "9598066F68…
## $ ride_type          <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ start_time         <chr> "2021-06-13 14:31:28", "2021-06-04 11:18:02", "2021…
## $ end_time           <chr> "2021-06-13 14:34:11", "2021-06-04 11:24:19", "2021…
## $ start_station_name <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",…
## $ start_station_id   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",…
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Michigan Ave &…
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "13042", "", ""…
## $ start_lat          <dbl> 41.80, 41.79, 41.80, 41.78, 41.80, 41.78, 41.79, 41…
## $ start_lng          <dbl> -87.59, -87.59, -87.60, -87.58, -87.59, -87.58, -87…
## $ end_lat            <dbl> 41.80000, 41.80000, 41.79000, 41.80000, 41.79000, 4…
## $ end_lng            <dbl> -87.6000, -87.6000, -87.5900, -87.6000, -87.5900, -…
## $ customer_type      <chr> "member", "member", "member", "member", "member", "…
```

Remove unnecessary columns

```
combined_data <- combined_data %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
glimpse(combined_data)
```

```
## Rows: 5,860,776
## Columns: 9
## $ ride_id            <chr> "99FEC93BA843FB20", "06048DCFC8520CAF", "9598066F68…
## $ ride_type          <chr> "electric_bike", "electric_bike", "electric_bike", …
## $ start_time         <chr> "2021-06-13 14:31:28", "2021-06-04 11:18:02", "2021…
## $ end_time           <chr> "2021-06-13 14:34:11", "2021-06-04 11:24:19", "2021…
## $ start_station_name <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",…
## $ start_station_id   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "",…
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Michigan Ave &…
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "13042", "", ""…
## $ customer_type      <chr> "member", "member", "member", "member", "member", "…
```

Generate new columns for better analysis and aggregation

```
combined_data[['start_time']] <- ymd_hms(combined_data[['start_time']])
combined_data[['end_time']] <- ymd_hms(combined_data[['end_time']])
str(combined_data)
```

```
## 'data.frame':    5860776 obs. of  9 variables:
##  $ ride_id           : chr  "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C
0FE48C412214" ...
##  $ ride_type         : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bik
e" ...
##  $ start_time        : POSIXct, format: "2021-06-13 14:31:28" "2021-06-04 11:18:02" ...
##  $ end_time          : POSIXct, format: "2021-06-13 14:34:11" "2021-06-04 11:24:19" ...
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ customer_type     : chr  "member" "member" "member" "member" ...
```

```
combined_data$date <- as.Date(combined_data$start_time)
combined_data$month <- format(as.Date(combined_data$date), "%m")
combined_data$day <- format(as.Date(combined_data$date), "%d")
combined_data$year <- format(as.Date(combined_data$date), "%Y")
combined_data$day_of_week <- format(as.Date(combined_data$date), "%A")

combined_data$ride_length <- as.numeric(difftime(combined_data$end_time,combined_data$start_t
ime))
```

See an overview of the data

```
dim(combined_data)
```

```
## [1] 5860776      15
```

```
summary(combined_data)
```

```
##    ride_id           ride_type           start_time
##  Length:5860776     Length:5860776     Min.   :2021-06-01 00:00:38.00
##  Class :character   Class :character   1st Qu.:2021-07-29 10:43:32.00
##  Mode  :character   Mode  :character   Median :2021-09-23 17:33:23.00
##                                        Mean   :2021-10-26 03:44:54.77
##                                        3rd Qu.:2022-01-14 10:59:12.75
##                                        Max.   :2022-05-31 23:59:56.00
##     end_time                      start_station_name start_station_id
##  Min.   :2021-06-01 00:06:22.00   Length:5860776     Length:5860776
##  1st Qu.:2021-07-29 11:02:56.50   Class :character   Class :character
##  Median :2021-09-23 17:49:29.50   Mode  :character   Mode  :character
##  Mean   :2021-10-26 04:05:36.14
##  3rd Qu.:2022-01-14 11:14:48.50
##  Max.   :2022-06-02 11:35:01.00
##  end_station_name   end_station_id     customer_type           date
##  Length:5860776     Length:5860776     Length:5860776     Min.   :2021-06-01
##  Class :character   Class :character   Class :character   1st Qu.:2021-07-29
##  Mode  :character   Mode  :character   Mode  :character   Median :2021-09-23
##                                                           Mean   :2021-10-25
##                                                           3rd Qu.:2022-01-14
##                                                           Max.   :2022-05-31
##     month              day                year           day_of_week
##  Length:5860776     Length:5860776     Length:5860776     Length:5860776
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   ride_length
##  Min.   :  -3482
##  1st Qu.:    382
##  Median :    680
##  Mean   :   1241
##  3rd Qu.:   1236
##  Max.   :3356649
```

```
length(unique(combined_data$customer_type))
```

```
## [1] 2
```

After having a look at the data there were empty cells especially in the start_station column. There were also values with ride length less than 0. **Drop rows with empty value and ride_length less than 0**

```
nrow(combined_data[combined_data$start_station_name == "", ])
```

```
## [1] 823167
```

```
combined_data_clean <- combined_data[!(combined_data$start_station_name == "" | combined_data
$ride_length<0),]
```

See an overview of the data

```
dim(combined_data_clean)
```

```
## [1] 5037488      15
```

```
glimpse(combined_data_clean)
```

```
## Rows: 5,037,488
## Columns: 15
## $ ride_id            <chr> "0D904FEC5F84A538", "C4185F300D6B552B", "60F97090AC…
## $ ride_type          <chr> "classic_bike", "classic_bike", "classic_bike", "cl…
## $ start_time         <dttm> 2021-06-04 07:29:18, 2021-06-23 08:39:36, 2021-06-…
## $ end_time           <dttm> 2021-06-04 07:45:34, 2021-06-23 08:41:37, 2021-06-…
## $ start_station_name <chr> "Orleans St & Elm St", "Desplaines St & Kinzie St",…
## $ start_station_id   <chr> "TA1306000006", "TA1306000003", "TA1307000127", "KA…
## $ end_station_name   <chr> "Orleans St & Elm St", "Kingsbury St & Kinzie St", …
## $ end_station_id     <chr> "TA1306000006", "KA1503000043", "TA1309000014", "TA…
## $ customer_type      <chr> "member", "member", "member", "member", "member", "…
## $ date               <date> 2021-06-04, 2021-06-23, 2021-06-27, 2021-06-01, 20…
## $ month              <chr> "06", "06", "06", "06", "06", "06", "06", "06", "06…
## $ day                <chr> "04", "23", "27", "01", "01", "17", "14", "17", "14…
## $ year               <chr> "2021", "2021", "2021", "2021", "2021", "2021", "20…
## $ day_of_week        <chr> "Friday", "Wednesday", "Sunday", "Tuesday", "Tuesda…
## $ ride_length        <dbl> 976, 121, 467, 158, 146, 218, 1954, 1112, 166, 1244…
```

```
head(combined_data_clean)
```

```
##             ride_id    ride_type          start_time            end_time
## 51 0D904FEC5F84A538 classic_bike 2021-06-04 07:29:18 2021-06-04 07:45:34
## 52 C4185F300D6B552B classic_bike 2021-06-23 08:39:36 2021-06-23 08:41:37
## 54 60F97090AC85F55E classic_bike 2021-06-27 12:26:58 2021-06-27 12:34:45
## 57 FBC7B1F0160AA304 classic_bike 2021-06-01 12:30:24 2021-06-01 12:33:02
## 58 37A52001AEEFA4E5 classic_bike 2021-06-01 11:32:17 2021-06-01 11:34:43
## 67 E49E5426F0B74023 classic_bike 2021-06-17 17:55:12 2021-06-17 17:58:50
##           start_station_name start_station_id          end_station_name
## 51       Orleans St & Elm St     TA1306000006       Orleans St & Elm St
## 52 Desplaines St & Kinzie St     TA1306000003  Kingsbury St & Kinzie St
## 54       Clark St & Grace St     TA1307000127     Clark St & Leland Ave
## 57  Kingsbury St & Kinzie St     KA1503000043 Desplaines St & Kinzie St
## 58 Desplaines St & Kinzie St     TA1306000003  Kingsbury St & Kinzie St
## 67  Kingsbury St & Kinzie St     KA1503000043 Desplaines St & Kinzie St
##    end_station_id customer_type       date month day year day_of_week
## 51   TA1306000006        member 2021-06-04    06  04 2021      Friday
## 52   KA1503000043        member 2021-06-23    06  23 2021   Wednesday
## 54   TA1309000014        member 2021-06-27    06  27 2021      Sunday
## 57   TA1306000003        member 2021-06-01    06  01 2021     Tuesday
## 58   KA1503000043        member 2021-06-01    06  01 2021     Tuesday
## 67   TA1306000003        member 2021-06-17    06  17 2021    Thursday
##    ride_length
## 51         976
## 52         121
## 54         467
## 57         158
## 58         146
## 67         218
```

The Analyze Phase

For the analyze phase we try to derive as much insights as we can to drive decision (mostly using descriptive analysis). We will be sorting, filtering, groups, and aggregating the date. Exploratory Data Analysis (EDA).

Check the summary of the ride_length column

```
summary(combined_data_clean$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     394     694    1299    1259 3356649
```

```
# summarise(combined_data_clean, mean_rd = mean(ride_length), min_re = min(ride_length),
          # median_rd = median(ride_length), max_rd = max(ride_length))
```

Compare members and casual riders

```
aggregate(combined_data_clean$ride_length ~ combined_data_clean$customer_type, FUN = mean)
```

```
##   combined_data_clean$customer_type combined_data_clean$ride_length
## 1                            casual                        1953.9121
## 2                            member                         789.8676
```

```
aggregate(combined_data_clean$ride_length ~ combined_data_clean$customer_type, FUN = median)
```

```
##   combined_data_clean$customer_type combined_data_clean$ride_length
## 1                            casual                             946
## 2                            member                             556
```

```
aggregate(combined_data_clean$ride_length ~ combined_data_clean$customer_type, FUN = max)
```

```
##   combined_data_clean$customer_type combined_data_clean$ride_length
## 1                            casual                          3356649
## 2                            member                            93594
```

```
aggregate(combined_data_clean$ride_length ~ combined_data_clean$customer_type, FUN = min)
```

```
##   combined_data_clean$customer_type combined_data_clean$ride_length
## 1                            casual                               0
## 2                            member                               0
```

See the average ride time by each day for members vs casual users

```
combined_data_clean$day_of_week <- ordered(combined_data_clean$day_of_week, levels=c("Sunday"
, "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
combined_data_clean$month <- ordered(combined_data_clean$month, levels=c("06", "07", "08", "0
9", "10", "11", "12", "01", "02", "03", "04", "05"))

aggregate(combined_data_clean$ride_length ~ combined_data_clean$customer_type + combined_data
_clean$day_of_week, FUN = mean)
```

```
##    combined_data_clean$customer_type combined_data_clean$day_of_week
## 1                             casual                           Sunday
## 2                             member                           Sunday
## 3                             casual                           Monday
## 4                             member                           Monday
## 5                             casual                          Tuesday
## 6                             member                          Tuesday
## 7                             casual                        Wednesday
## 8                             member                        Wednesday
## 9                             casual                         Thursday
## 10                            member                         Thursday
## 11                            casual                           Friday
## 12                            member                           Friday
## 13                            casual                         Saturday
## 14                            member                         Saturday
##    combined_data_clean$ride_length
## 1                        2244.7748
## 2                         900.3071
## 3                        1959.8245
## 4                         766.3789
## 5                        1680.2109
## 6                         742.1139
## 7                        1709.3799
## 8                         743.2990
## 9                        1778.4904
## 10                        753.0999
## 11                       1845.2443
## 12                        772.5013
## 13                       2124.5475
## 14                        888.4656
```

Analyze ridership data by type and weekday

```
combined_data_clean %>%
  mutate(weekday = wday(start_time, label = TRUE)) %>%
  group_by(customer_type, weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(customer_type, weekday)
```

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

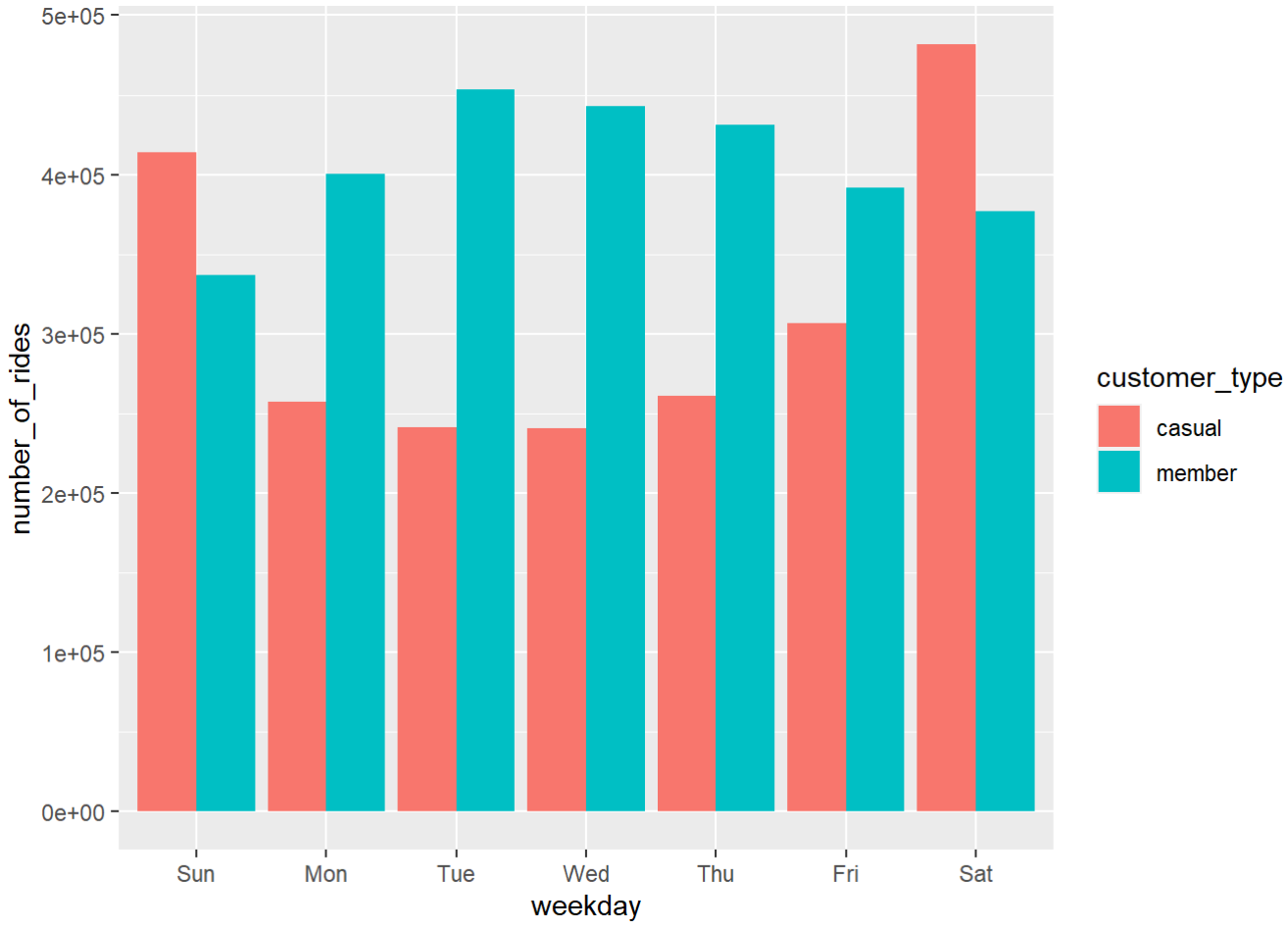
```
## # A tibble: 14 × 4
## # Groups:   customer_type [2]
##    customer_type weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              414136            2245.
##  2 casual        Mon              257509            1960.
##  3 casual        Tue              241481            1680.
##  4 casual        Wed              240842            1709.
##  5 casual        Thu              261274            1778.
##  6 casual        Fri              306918            1845.
##  7 casual        Sat              481689            2125.
##  8 member        Sun              336759             900.
##  9 member        Mon              400647             766.
## 10 member        Tue              453466             742.
## 11 member        Wed              443085             743.
## 12 member        Thu              431189             753.
## 13 member        Fri              391590             773.
## 14 member        Sat              376903             888.
```

Share

In the share phase of the data analysis process, Visualization is made to make the analysis easily digestible by the stakeholders. Plots make it easier to see trends, and relationships that exist in the data.

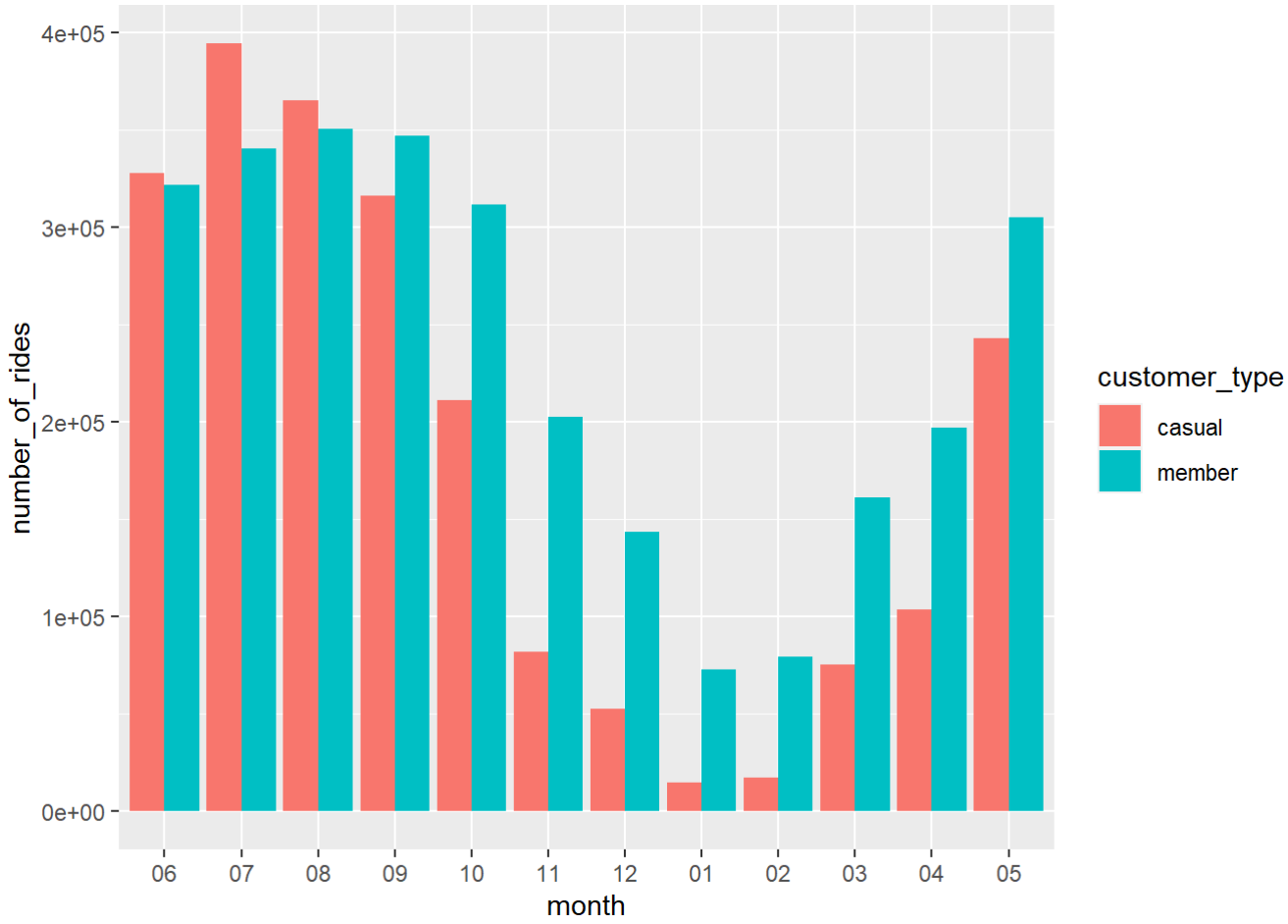
Visualize the number of rides by rider type by weekday

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



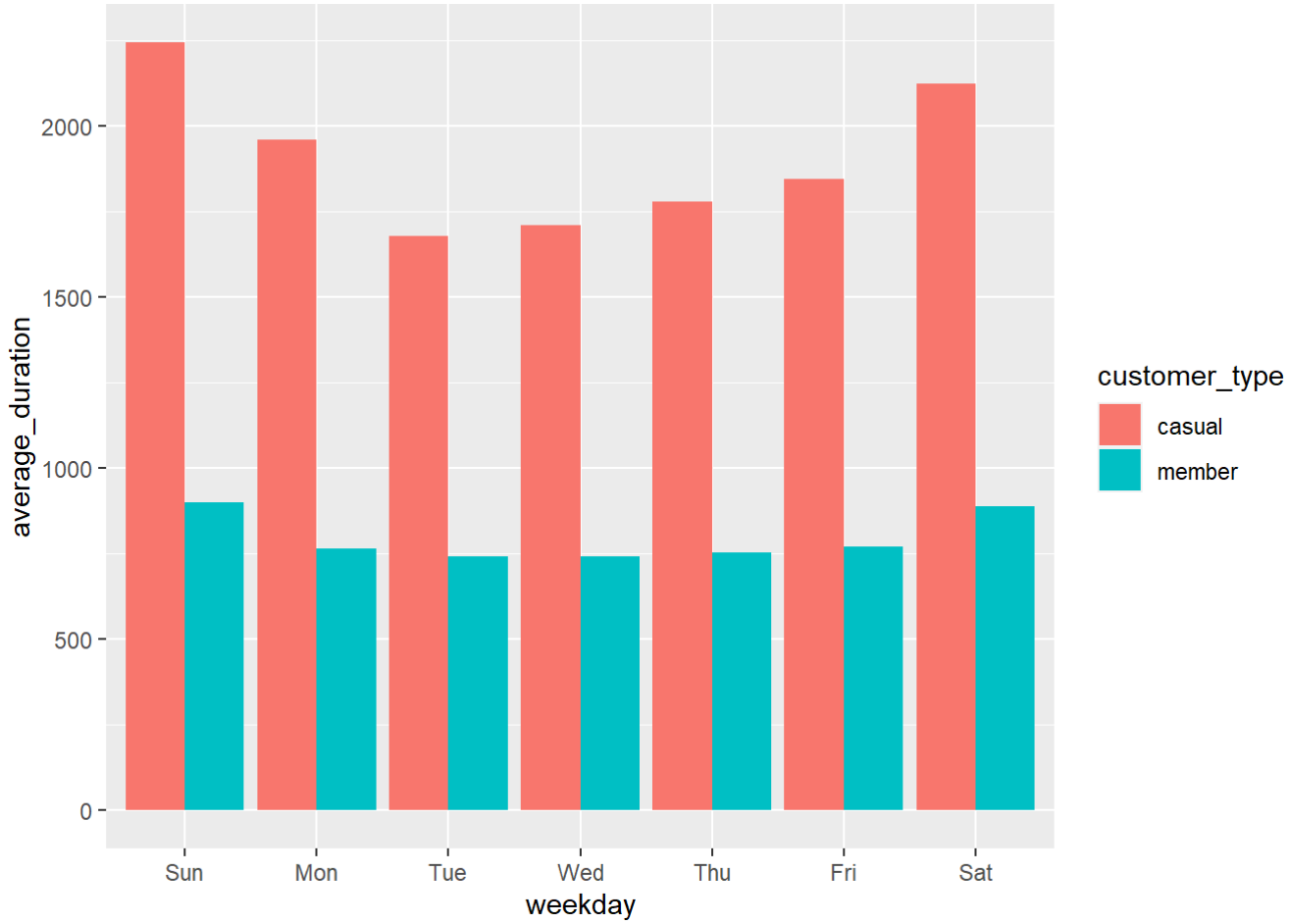
Visualize the number of rides by rider type by month

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



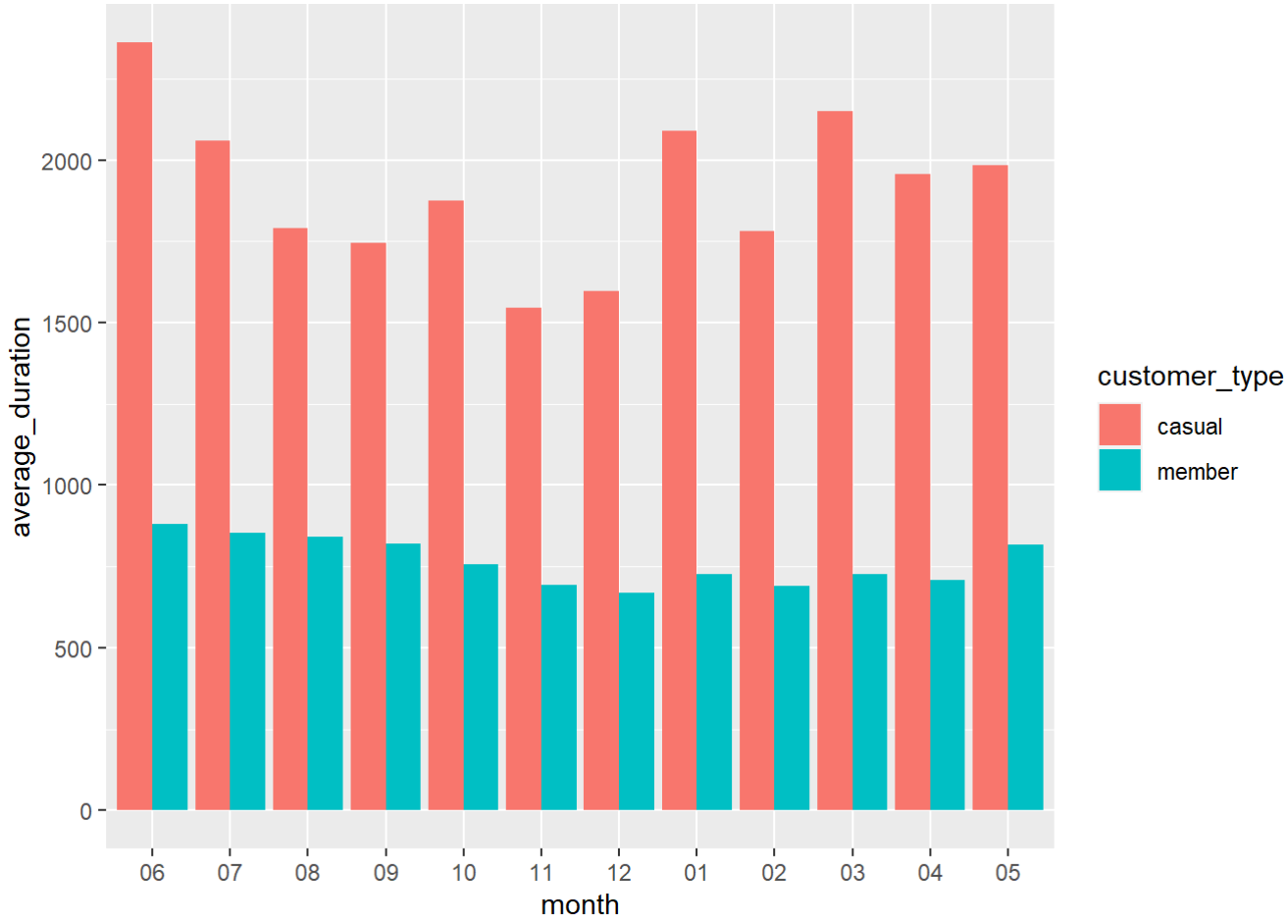
Visualization for average duration by weekday

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



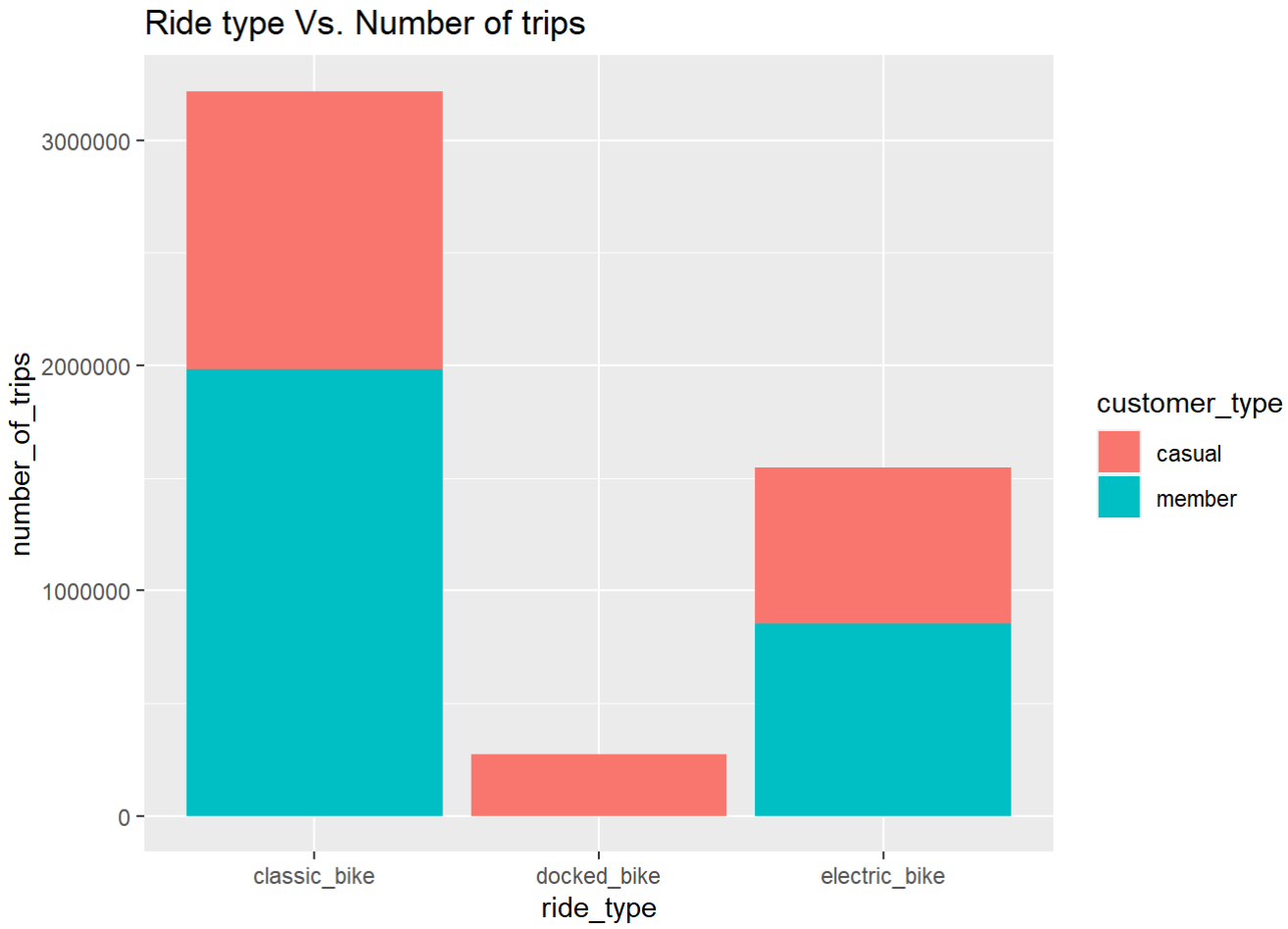
Visualization for average duration by month

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



Visualize the ride_type by number of trips

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



The Act Phase

From the analysis and visualization produces, we can have the following takeaways to answer the business question: **How do annual members and casual riders use Cyclistic bikes differently?**.

- Casual riders use the bikes for longer duration (per ride).
- Asides from weekends (Saturdays and Sundays), Members use more of the service during the weekdays.
- We also see that over the last course of year, most rides are by members and not casual riders.

Recommendations * Make a plan to let go of docked bike and focus more on the electric and classic ride types. * Reduce the number of available bikes for casual users during the weekends or increase the on the go price for bikes during the weekend. * Cap the duration a casual user can make use of bikes.