Case study: Cyclistic report

Ben Tunnicliff

11/08/2021

Background

- The company is a fictional bike-share company based in Chicago, named Cyclistic. The company maintains a fleet of more than 5800 bicycles and 600 docking stations. They set themselves apart by offering different types of bikes such as reclining bikes, hand tricycles, and cargo bikes in order to make bike sharing more inclusive to people with diabilities and other riders who are unable to use a stadard bicycle.
- Cyclistic offers a few options for riders in terms of pricing plans. They offer single-ride passes, full-day passes, and annual memberships. those who purchase annual memberships are referred to as members, the others are referred to as casual riders.
- Cyclistic finance analysts have concluded that annual members are much more profitable than casual riders, and it is believed that maximising the number of annual members will be key to the company's future growth.
- The current marketing strategy revolves around building general awareness and appealing to broad consumer segments.

Goals/Business tasks

- The goal of the marketing team is to design marketing strategies aimed at converting casual riders into annual members.
- My task is to answer the question: **How do annual members and casual riders use Cyclistic bikes differently?**

Preparing data

Data source previous 12 months (07/2020 to 06/2021) of Cyclistic trip data downloaded from here (https://divvy-tripdata.s3.amazonaws.com/index.html)

- 12 .csv files.
- 13 variables including start/end station names and ID numbers, rideable types, start and end times etc.

Notes

- The datasets have a different name because Cyclistic is a fictional company.
- The data has been made available by Motivate International Inc. under this license. (https://www.divvybikes.com/data-license-agreement)
- This is public data.
- The data is unbiased, original and is not missing any important information.

Processing and preparing data

For this analysis I will be using R, making use of the tidyverse and lubridate packages

```
#loading necessary packages.
library(tidyverse)
library(lubridate)
```

Collecting the data

```
#Importing all datasets
divvy202007 <- read.csv("202007-divvy-tripdata.csv")
divvy202008 <- read.csv("202008-divvy-tripdata.csv")
divvy202009 <- read.csv("202009-divvy-tripdata.csv")
divvy202010 <- read.csv("202010-divvy-tripdata.csv")
divvy202011 <- read.csv("202011-divvy-tripdata.csv")
divvy202012 <- read.csv("202012-divvy-tripdata.csv")
divvy202101 <- read.csv("202101-divvy-tripdata.csv")
divvy202102 <- read.csv("202102-divvy-tripdata.csv")
divvy202103 <- read.csv("202103-divvy-tripdata.csv")
divvy202104 <- read.csv("202104-divvy-tripdata.csv")
divvy202105 <- read.csv("202105-divvy-tripdata.csv")
divvy202106 <- read.csv("202106-divvy-tripdata.csv")
```

Wrangling data and combining into a single file

Examining the structure of each dataset to look for inconsistencies.

```
str(divvy202007)
```

```
## 'data.frame':    551480 obs. of  13 variables:
##  $ ride_id           : Factor w/ 551480 levels "000001004784CD35",..: 254094 410914 454389 181590 180120 21785
9 74704 434522 542626 297608 ...
##  $ rideable_type     : Factor w/ 2 levels "docked_bike",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ started_at        : Factor w/ 468441 levels "2020-07-01 00:00:14",..: 123022 355167 114644 246083 48091 413
950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 467938 levels "2020-07-01 00:03:01",..: 122578 354581 114058 245800 47706 413
288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 620 levels "","2112 W Peterson Ave",..: 468 271 322 342 324 229 44 518 126 23
8 ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : Factor w/ 623 levels "","2112 W Peterson Ave",..: 581 54 142 120 139 579 192 240 89 382
...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ 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 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2 2 2 2 ...
```

```
str(divvy202008)
```

```
## 'data.frame':    622361 obs. of  13 variables:
##  $ ride_id           : Factor w/ 622361 levels "0000376F8A298CB2",..: 121677 102355 251960 485097 47413 209041
572253 439485 27195 74727 ...
##  $ rideable_type     : Factor w/ 2 levels "docked_bike",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ started_at        : Factor w/ 514094 levels "2020-08-01 00:00:01",..: 317981 436824 423318 430476 434240 43
4950 423882 425170 422733 432804 ...
##  $ ended_at          : Factor w/ 512691 levels "2020-08-01 00:04:41",..: 316439 436534 423696 429594 432800 43
4140 422666 423915 421382 431654 ...
##  $ start_station_name: Factor w/ 637 levels "","2112 W Peterson Ave",..: 331 393 157 172 360 360 116 496 527 5
61 ...
##  $ start_station_id  : int  329 168 195 81 658 658 196 67 153 177 ...
##  $ end_station_name  : Factor w/ 637 levels "","2112 W Peterson Ave",..: 135 393 545 542 360 360 198 525 274 6
09 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 1 1 1 1 1 1 1 1 1 ...
```

```
str(divvy202009)
```

```
## 'data.frame':    532958 obs. of  13 variables:
##  $ ride_id           : Factor w/ 532958 levels "00002279D7D315A5",..: 89779 349629 278773 182749 386687 115699
26535 312995 175935 303278 ...
##  $ rideable_type     : Factor w/ 2 levels "docked_bike",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ started_at        : Factor w/ 446551 levels "2020-09-01 00:00:07",..: 235638 236129 236145 240029 236240 24
0731 228515 233987 238344 234011 ...
##  $ ended_at          : Factor w/ 443261 levels "2020-09-01 00:04:43",..: 234031 234291 234313 238558 234975 23
9709 226851 232211 236941 232322 ...
##  $ start_station_name: Factor w/ 665 levels "","2112 W Peterson Ave",..: 417 599 599 24 245 126 621 1 1 1 ...
##  $ start_station_id  : int  52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : Factor w/ 664 levels "","2112 W Peterson Ave",..: 263 598 598 437 245 1 59 1 1 1 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(divvy202010)
```

```
## 'data.frame':    388653 obs. of  13 variables:
##  $ ride_id           : Factor w/ 388653 levels "00004BC36B8BA585",..: 262383 339326 276737 104040 25355 332030
296134 32903 61070 343472 ...
##  $ rideable_type     : Factor w/ 2 levels "docked_bike",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ started_at        : Factor w/ 338923 levels "2020-10-01 00:00:06",..: 337794 338892 338737 338589 337784 31
6661 312746 315903 311227 315918 ...
##  $ ended_at          : Factor w/ 335877 levels "2020-10-01 00:05:09",..: 334752 335814 335651 335484 334730 31
3668 309860 313012 308330 312991 ...
##  $ start_station_name: Factor w/ 670 levels "","2112 W Peterson Ave",..: 356 549 578 135 551 361 356 507 1 95
...
##  $ start_station_id  : int  313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : Factor w/ 670 levels "","2112 W Peterson Ave",..: 506 326 594 61 571 627 506 356 610 24
8 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(divvy202011)
```

```
## 'data.frame':    259716 obs. of  13 variables:
##  $ ride_id           : Factor w/ 259716 levels "000058F6DBC33A30",..: 192009 153049 201131 232599 29373 116211
148308 160945 168141 132269 ...
##  $ rideable_type     : Factor w/ 2 levels "docked_bike",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ started_at        : Factor w/ 233064 levels "2020-11-01 00:00:08",..: 2723 1201 102 135 3822 142727 143204
142809 142987 137913 ...
##  $ ended_at          : Factor w/ 231222 levels "2020-11-01 00:02:20",..: 2653 1202 156 119 3979 141945 142131
141729 142012 136908 ...
##  $ start_station_name: Factor w/ 666 levels "","2112 W Peterson Ave",..: 204 255 348 376 64 601 348 1 405 125
...
##  $ start_station_id  : int  110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : Factor w/ 658 levels "","2112 W Peterson Ave",..: 546 444 246 564 64 346 594 1 355 410
...
##  $ 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     : Factor w/ 2 levels "casual","member": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(divvy202012)
```

```
## 'data.frame':    131573 obs. of  13 variables:
##  $ ride_id           : Factor w/ 131573 levels "000048F4A10F09FA",..: 57568 11004 42073 97616 53928 99637 9031
5 15315 3091 96244 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 1 3 3 3 3 3 3 3 3 3 ...
##  $ started_at        : Factor w/ 125672 levels "2020-12-01 00:01:15",..: 112797 83380 70156 70391 99448 99894
12737 12159 56580 81420 ...
##  $ ended_at          : Factor w/ 125413 levels "2020-11-25 07:40:56",..: 112530 83129 69917 70154 99144 99618
12903 12348 56687 81134 ...
##  $ start_station_name: Factor w/ 645 levels "","2112 W Peterson Ave",..: 5 1 1 1 1 1 1 1 1 1 ...
##  $ start_station_id  : Factor w/ 643 levels "","13001","13006",..: 59 1 1 1 1 1 1 1 1 1 ...
##  $ end_station_name  : Factor w/ 644 levels "","2112 W Peterson Ave",..: 205 1 1 1 1 1 1 1 1 1 ...
##  $ end_station_id    : Factor w/ 642 levels "","13001","13006",..: 507 1 1 1 1 1 1 1 1 1 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 2 2 2 2 2 2 2 2 2 ...
```

```
str(divvy202101)
```

```
## 'data.frame':    96834 obs. of  13 variables:
##  $ ride_id           : Factor w/ 96834 levels "0000CB9B87AECF12",..: 85247 83334 89312 30159 71828 35268 3440
84454 73654 70200 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ started_at        : Factor w/ 93736 levels "2021-01-01 00:02:05",..: 74985 85029 69525 17741 72665 26048 56
53 45379 24251 78630 ...
##  $ ended_at          : Factor w/ 93582 levels "2021-01-01 00:08:39",..: 74802 84840 69352 17713 72501 26319 56
29 45291 24187 78449 ...
##  $ start_station_name: Factor w/ 641 levels "","2112 W Peterson Ave",..: 72 72 72 72 72 72 72 72 72 72 ...
##  $ start_station_id  : Factor w/ 639 levels "","13001","13006",..: 212 212 212 212 212 212 212 212 212 212 ...
##  $ end_station_name  : Factor w/ 633 levels "","2112 W Peterson Ave",..: 1 1 1 1 1 1 1 1 1 625 ...
##  $ end_station_id    : Factor w/ 630 levels "","13001","13006",..: 1 1 1 1 1 1 1 1 1 327 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 2 2 2 1 1 2 2 2 2 ...
```

```
str(divvy202102)
```

```
## 'data.frame':    49622 obs. of  13 variables:
##  $ ride_id           : Factor w/ 49622 levels "000113AD1F802A92",..: 26634 3153 44708 34798 25468 36799 32495
8073 10501 6558 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 1 1 3 1 3 3 1 1 1 1 ...
##  $ started_at        : Factor w/ 48139 levels "2021-02-01 00:55:44",..: 16968 18803 13136 3408 26702 29923 128
3 15569 41907 21999 ...
##  $ ended_at          : Factor w/ 48035 levels "2021-02-01 01:22:48",..: 16930 18774 13123 3368 26657 29841 122
9 15530 41749 21974 ...
##  $ start_station_name: Factor w/ 583 levels "","2112 W Peterson Ave",..: 233 233 121 576 487 218 316 218 316 3
16 ...
##  $ start_station_id  : Factor w/ 583 levels "","13001","13006",..: 242 242 305 275 74 202 404 202 404 404 ...
##  $ end_station_name  : Factor w/ 585 levels "","2112 W Peterson Ave",..: 462 43 500 257 211 315 315 215 315 31
5 ...
##  $ end_station_id    : Factor w/ 585 levels "","13001","13006",..: 291 190 441 444 575 406 406 205 406 406 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 1 2 2 2 1 1 2 2 2 ...
```

```
str(divvy202103)
```

```
## 'data.frame':    228496 obs. of  13 variables:
##  $ ride_id           : Factor w/ 228496 levels "0000465CF790D824",..: 185433 43666 118021 136656 199406 184597
37162 217393 187975 48848 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 1 1 1 1 1 1 1 3 3 1 ...
##  $ started_at        : Factor w/ 209025 levels "2021-03-01 00:01:09",..: 91217 179105 65206 61056 120307 10990
1 112767 143727 206151 63471 ...
##  $ ended_at          : Factor w/ 208629 levels "2021-03-01 00:06:28",..: 91076 178832 65051 61023 120104 10981
0 112393 143522 205693 63322 ...
##  $ start_station_name: Factor w/ 674 levels "","2112 W Peterson Ave",..: 307 307 541 661 264 264 571 543 154 4
21 ...
##  $ start_station_id  : Factor w/ 674 levels "","13001","13006",..: 176 176 136 614 301 301 18 609 8 524 ...
##  $ end_station_name  : Factor w/ 674 levels "","2112 W Peterson Ave",..: 576 105 279 61 113 113 352 237 259 49
8 ...
##  $ end_station_id    : Factor w/ 674 levels "","13001","13006",..: 96 218 623 115 379 379 496 393 529 369 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 1 1 1 1 1 1 2 2 2 2 ...
```

```
str(divvy202104)
```

```
## 'data.frame':    337230 obs. of  13 variables:
##  $ ride_id           : Factor w/ 337230 levels "00001A81D056B01B",..: 143082 39569 300958 32373 254792 12428 1
10411 280871 145203 5400 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 1 2 2 1 2 1 1 3 1 1 ...
##  $ started_at        : Factor w/ 298722 levels "2021-04-01 00:03:18",..: 121625 266503 17983 160271 17977 2437
56 23625 61937 118383 229855 ...
##  $ ended_at          : Factor w/ 298625 levels "2021-04-01 00:14:29",..: 121877 267738 72363 160453 19437 2436
43 22860 61951 118872 229458 ...
##  $ start_station_name: Factor w/ 682 levels "","2112 W Peterson Ave",..: 579 221 402 309 402 160 19 221 19 221
...
##  $ start_station_id  : Factor w/ 682 levels "","13001","13006",..: 572 438 242 541 242 148 210 438 210 438 ...
##  $ end_station_name  : Factor w/ 682 levels "","2112 W Peterson Ave",..: 559 219 400 559 400 158 18 219 18 219
...
##  $ end_station_id    : Factor w/ 682 levels "","13001","13006",..: 81 438 242 81 242 148 210 438 210 438 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 1 1 2 1 1 1 1 1 1 ...
```

```
str(divvy202105)
```

```
## 'data.frame':    531633 obs. of  13 variables:
##  $ ride_id           : Factor w/ 531633 levels "00007D79F5A24D07",..: 415202 459780 22324 249843 276260 510537
416197 150842 473165 140090 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ started_at        : Factor w/ 447224 levels "2021-05-01 00:00:11",..: 413510 412810 417621 417678 424275 41
2914 412027 61000 59542 55426 ...
##  $ ended_at          : Factor w/ 447217 levels "2021-05-01 00:03:26",..: 413102 413177 416700 417169 423690 41
2805 411781 61095 59534 55633 ...
##  $ start_station_name: Factor w/ 688 levels "","2112 W Peterson Ave",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ start_station_id  : Factor w/ 687 levels "","13001","13006",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ end_station_name  : Factor w/ 684 levels "","2112 W Peterson Ave",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ end_station_id    : Factor w/ 683 levels "","13001","13006",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(divvy202106)
```

```
## 'data.frame':    729595 obs. of  13 variables:
##  $ ride_id           : Factor w/ 729595 levels "000002EBE159AE82",..: 439058 17159 426509 502460 530070 281532
359740 176414 108703 499913 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ started_at        : Factor w/ 589805 levels "2021-06-01 00:00:38",..: 261995 56935 55912 48849 59887 48225
188094 188917 184050 473563 ...
##  $ ended_at          : Factor w/ 589069 levels "2021-06-01 00:06:22",..: 260960 56690 55680 48814 59482 47748
187603 188374 183662 472558 ...
##  $ start_station_name: Factor w/ 690 levels "","2112 W Peterson Ave",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ start_station_id  : Factor w/ 690 levels "","13001","13006",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ end_station_name  : Factor w/ 691 levels "","2112 W Peterson Ave",..: 1 1 1 1 1 1 1 1 1 1 432 ...
##  $ end_station_id    : Factor w/ 691 levels "","13001","13006",..: 1 1 1 1 1 1 1 1 1 16 ...
##  $ 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     : Factor w/ 2 levels "casual","member": 2 2 2 2 2 2 2 2 2 2 ...
```

All of the column names match, however there are some variables that are under different classes which will need to be corrected.

```
#converting ride_id and rideable_type to character so that they stack correctly. also converting start and end st
ation ids to int for 202012 onward.
divvy202007 <- mutate(divvy202007, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
divvy202008 <- mutate(divvy202008, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
divvy202009 <- mutate(divvy202009, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
divvy202010 <- mutate(divvy202010, ride_id = as.character(ride_id), rideable_type = as.character(rideable_type))
divvy202011 <- divvy202011 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
))
divvy202012 <- divvy202012 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202101 <- divvy202101 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202102 <- divvy202102 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202103 <- divvy202103 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202104 <- divvy202104 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202105 <- divvy202105 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
divvy202106 <- divvy202106 %>% mutate(ride_id = as.character(ride_id), rideable_type = as.character(rideable_type
), start_station_id = as.integer(start_station_id), end_station_id = as.integer(end_station_id))
```

Now that the formatting is consistent the data can be merged into one table, and unnecessary data can be removed.

```
#Stacking individual month's data frames into one large data frame, and removing longitude and latitude columns.
all_trips <- bind_rows(divvy202007, divvy202008, divvy202009, divvy202010, divvy202011, divvy202012, divvy202101,
                       divvy202102, divvy202103, divvy202104, divvy202105, divvy202106)


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

Cleaning and adding new data

Starting by looking at the dimensions and structure of the new dataset.

```
str(all_trips)
```

```
## 'data.frame':    4460151 obs. of  9 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : Factor w/ 3813396 levels "2020-07-01 00:00:14",..: 123022 355167 114644 246083 48091 41
3950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 3801221 levels "2020-07-01 00:03:01",..: 122578 354581 114058 245800 47706 41
3288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 713 levels "","2112 W Peterson Ave",..: 468 271 322 342 324 229 44 518 126 23
8 ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : Factor w/ 714 levels "","2112 W Peterson Ave",..: 581 54 142 120 139 579 192 240 89 382
...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2 2 2 2 ...
```

- The all_trips dataset has 4460151 rows and 9 columns.
- A few more columns will be added to make the analysis more in depth.

Columns will be added for the date, day, month, year, day of the week, and length of each ride.

```
#adding columns to list the day, month, and year of each ride.
all_trips$date <- as.Date(all_trips$started_at)
all_trips$month <- format(as.Date(all_trips$date), "%B")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year <- format(as.Date(all_trips$date), "%y")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%a")

#adding a ride length column (in seconds)
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)

#checking the structure of the added columns.
str(all_trips)
```

```
## 'data.frame':    4460151 obs. of  15 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : Factor w/ 3813396 levels "2020-07-01 00:00:14",..: 123022 355167 114644 246083 48091 41
3950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 3801221 levels "2020-07-01 00:03:01",..: 122578 354581 114058 245800 47706 41
3288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 713 levels "","2112 W Peterson Ave",..: 468 271 322 342 324 229 44 518 126 23
8 ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : Factor w/ 714 levels "","2112 W Peterson Ave",..: 581 54 142 120 139 579 192 240 89 382
...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2 2 2 2 ...
##  $ date              : Date, format: "2020-07-09" "2020-07-24" ...
##  $ month             : chr  "July" "July" "July" "July" ...
##  $ day               : chr  "09" "24" "08" "17" ...
##  $ year              : chr  "20" "20" "20" "20" ...
##  $ day_of_week       : chr  "Thu" "Fri" "Wed" "Fri" ...
##  $ ride_length       : 'difftime' num  0 86400 0 0 ...
##   ..- attr(*, "units")= chr "secs"
```

```
#converting ride length to a numeric value for calculations.
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

```
## [1] TRUE
```

```
str(all_trips)
```

```
## 'data.frame':    4460151 obs. of  15 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : Factor w/ 3813396 levels "2020-07-01 00:00:14",..: 123022 355167 114644 246083 48091 41
3950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 3801221 levels "2020-07-01 00:03:01",..: 122578 354581 114058 245800 47706 41
3288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 713 levels "","2112 W Peterson Ave",..: 468 271 322 342 324 229 44 518 126 23
8 ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : Factor w/ 714 levels "","2112 W Peterson Ave",..: 581 54 142 120 139 579 192 240 89 382
...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2 2 2 2 ...
##  $ date              : Date, format: "2020-07-09" "2020-07-24" ...
##  $ month             : chr  "July" "July" "July" "July" ...
##  $ day               : chr  "09" "24" "08" "17" ...
##  $ year              : chr  "20" "20" "20" "20" ...
##  $ day_of_week       : chr  "Thu" "Fri" "Wed" "Fri" ...
##  $ ride_length       : num  0 86400 0 0 0 0 0 0 0 0 ...
```

In the dataset there are instances of “bad data” such as trips with negative ride length, as well as bikes being taken for quality tests which need to be removed. A new data frame will be created for this.

```
#removing "bad" data, where bikes were taken out of docks for quality checks, or ride_length was negative.
all_trips_v2 <- all_trips[!(all_trips$start_station_name == "HQ QR" | all_trips$ride_length < 0),]
```

Another column will be added to include the start time of each ride for further analysis.

```
all_trips_v2$start_time <- as.POSIXct(all_trips_v2$started_at, "%Y-%m-%d %H:%M:%S")
str(all_trips_v2)
```

```
## 'data.frame':    4459772 obs. of  16 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : Factor w/ 3813396 levels "2020-07-01 00:00:14",..: 123022 355167 114644 246083 48091 41
3950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 3801221 levels "2020-07-01 00:03:01",..: 122578 354581 114058 245800 47706 41
3288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 713 levels "","2112 W Peterson Ave",..: 468 271 322 342 324 229 44 518 126 23
8 ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : Factor w/ 714 levels "","2112 W Peterson Ave",..: 581 54 142 120 139 579 192 240 89 382
...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2 2 2 2 ...
##  $ date              : Date, format: "2020-07-09" "2020-07-24" ...
##  $ month             : chr  "July" "July" "July" "July" ...
##  $ day               : chr  "09" "24" "08" "17" ...
##  $ year              : chr  "20" "20" "20" "20" ...
##  $ day_of_week       : chr  "Thu" "Fri" "Wed" "Fri" ...
##  $ ride_length       : num  0 86400 0 0 0 0 0 0 0 0 ...
##  $ start_time        : POSIXct, format: "2020-07-09 15:22:02" "2020-07-24 23:56:30" ...
```

The all_trips_v2 dataset has 4459772 rows with 16 columns, and it is now cleaned and ready for analysis.

Analysis of the data

Starting at a broad level by taking the mean of ride length, as well as the number of rides for members and casual riders.

```
#Getting descriptive statistics.
all_trips_v2 %>% summarise(mean_ride_length = mean(ride_length))
```

```
##   mean_ride_length
## 1         928.9072
```

```
table(all_trips_v2$member_casual)
```

```
##
##  casual  member
## 1929240 2530532
```

From the table we can see that members make up around 57% of total rides. We can then look at the mean ride length for both members and casual riders.

```
#comparing members and casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                1808.8953
## 2                     member                 258.0174
```

We can also look at average ride length for members and casual riders for each day of the week.

```
#change the days of the week to be in order.
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri",
"Sat"))

#average ride time by day for members vs casual users
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                      Sun                1491.5833
## 2                      member                      Sun                 181.6585
## 3                      casual                      Mon                1578.2741
## 4                      member                      Mon                 183.5504
## 5                      casual                      Tue                1428.2457
## 6                      member                      Tue                 170.4273
## 7                      casual                      Wed                1543.1744
## 8                      member                      Wed                 194.6631
## 9                      casual                      Thu                1729.7799
## 10                     member                      Thu                 237.6705
## 11                     casual                      Fri                2564.7638
## 12                     member                      Fri                 410.5104
## 13                     casual                      Sat                2045.2234
## 14                     member                      Sat                 408.2599
```

We can then create a table showing average ride length and number of rides for each day of the week, for members and casual riders.

```
# analyze ridership data by type and weekday
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <fct>         <ord>             <int>            <dbl>
##  1 casual        Sun              366593            1492.
##  2 casual        Mon              209391            1578.
##  3 casual        Tue              204222            1428.
##  4 casual        Wed              215607            1543.
##  5 casual        Thu              212529            1730.
##  6 casual        Fri              281188            2565.
##  7 casual        Sat              439710            2045.
##  8 member        Sun              323420             182.
##  9 member        Mon              337503             184.
## 10 member        Tue              364505             170.
## 11 member        Wed              389251             195.
## 12 member        Thu              363892             238.
## 13 member        Fri              375057             411.
## 14 member        Sat              376904             408.
```

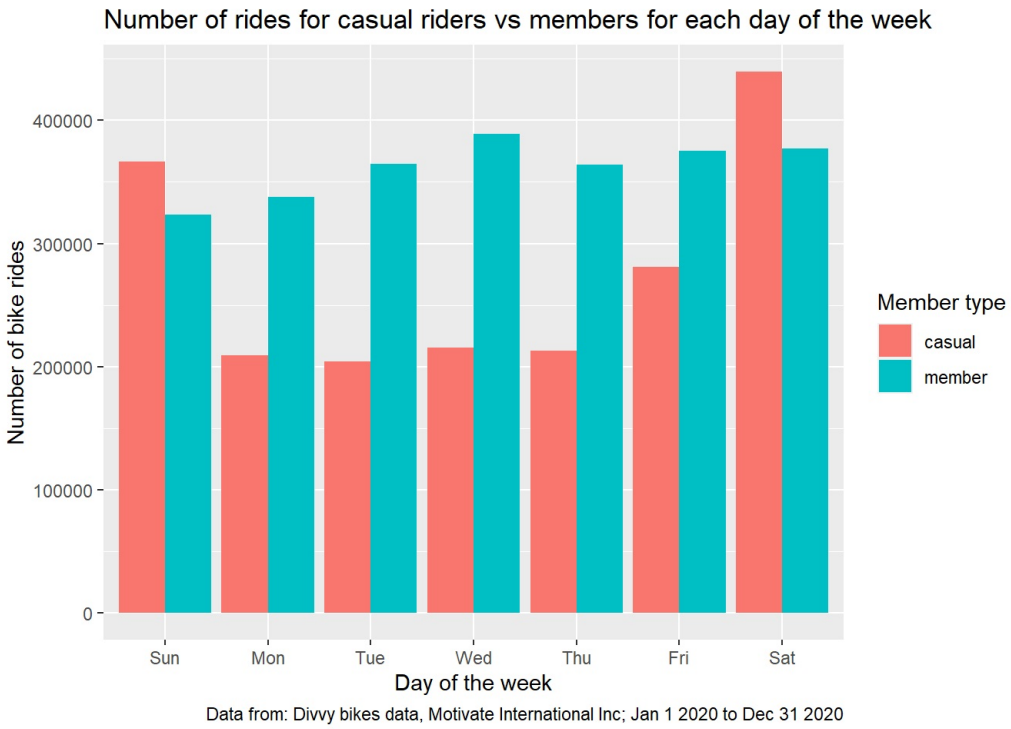
From this table it can be seen that:

- The number of rides through the week for members is relatively consistent, while for casual riders there are more rides on weekend days.
- The average ride length for members is much lower than the average ride length for casual riders.
- The average ride length for casual riders is higher on weekends.

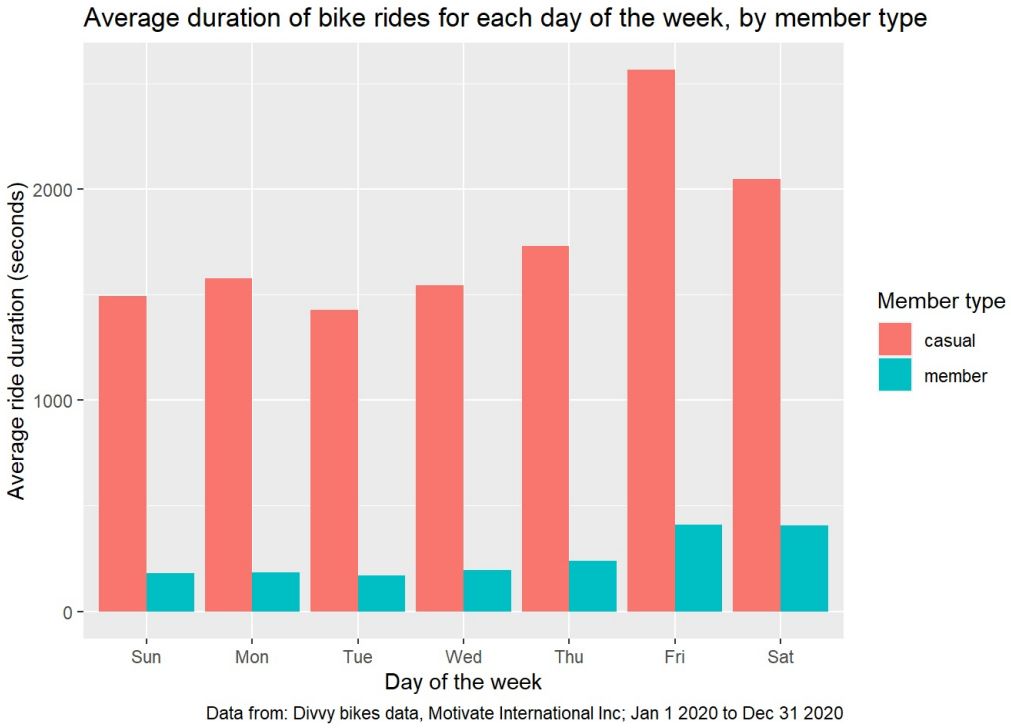
Visualising the data

Number of rides and Average ride length for each day of the week.

```
#visualise number of rides by rider type
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  labs(title = "Number of rides for casual riders vs members for each day of the week",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Day of the week",
       y = "Number of bike rides",
       fill = "Member type")
```




```
#Create a visualisation for average ride length.
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Average duration of bike rides for each day of the week, by member type",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Day of the week",
       y = "Average ride duration (seconds)",
       fill = "Member type")
```



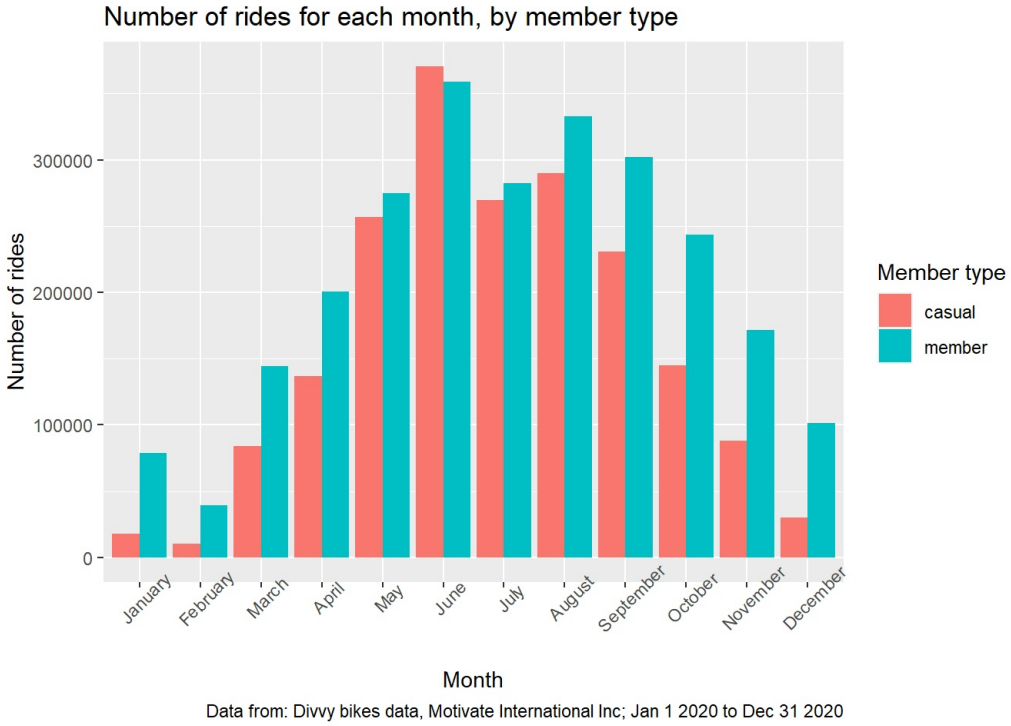
These two charts show the same thing as the previous table:

- Members tend to have a consistent number of rides for each day while casual riders use the service more on weekends.
- Average ride length for members is much lower than casual riders.
- For casual riders the average ride length peaks on weekends.
- Casual riders appear to use the service more than members on weekends.

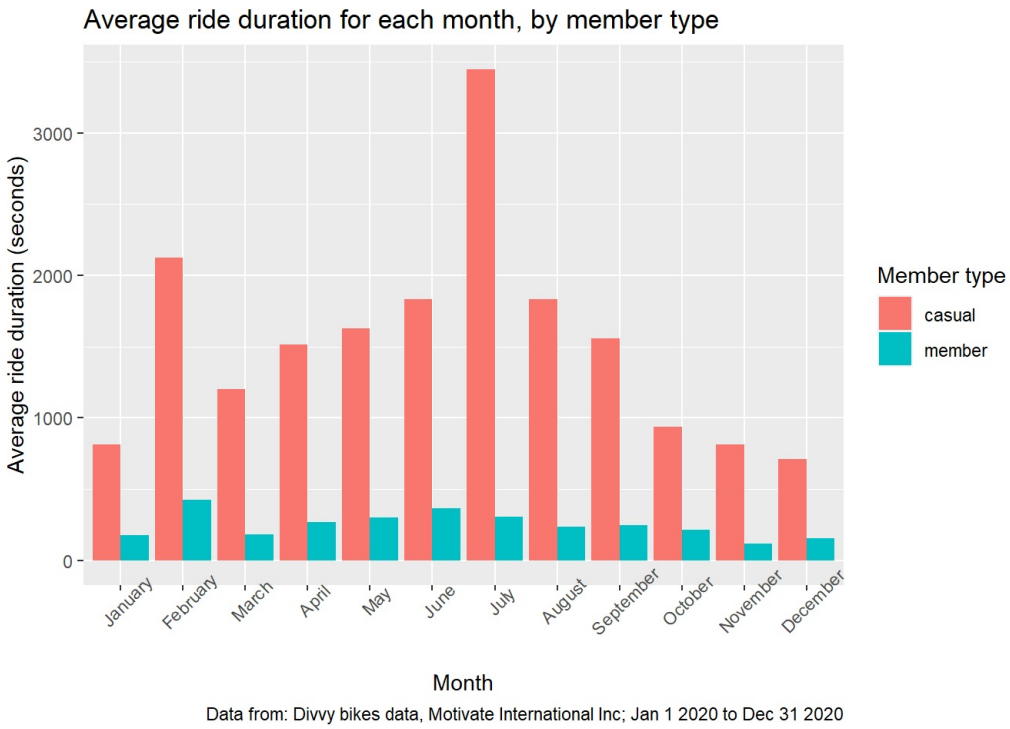
Number of rides and Average ride length for each month.

```
#Rider behaviour by month
all_trips_v2$month <- ordered(all_trips_v2$month, levels = c("January", "February", "March", "April",
                                                             "May", "June", "July", "August", "September", "Octob
er",
                                                             "November", "December"))


#Creating visualisations for number of rides and average ride length by month
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n()) %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Number of rides for each month, by member type",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Month",
       y = "Number of rides",
       fill = "Member type")
```




```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(),
            average_duration = mean(ride_length)) %>%
  ggplot(aes(x = month, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Average ride duration for each month, by member type",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Month",
       y = "Average ride duration (seconds)",
       fill = "Member type")
```

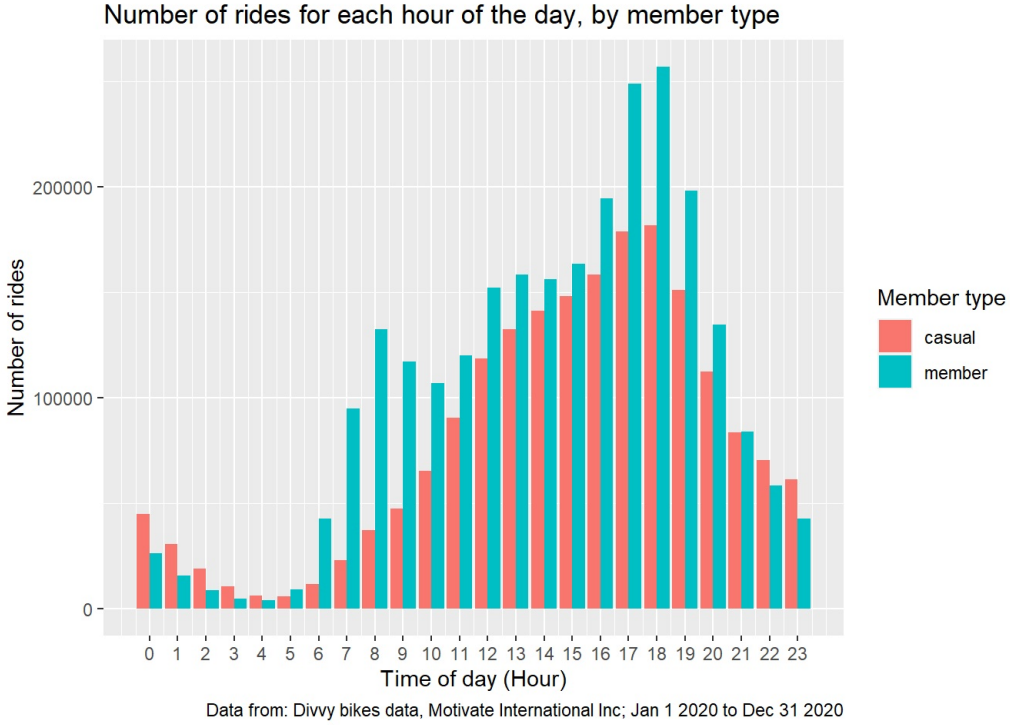


Several insights can be gained from these charts:

- Casual riders' usage of the service starts to increase in the Spring time, and peaks in the Summer months before decreasing when heading into Winter.
- Members tend to follow the same pattern, but with less change through the months when compared to casual riders.
- The average ride length for casual riders is much higher than average ride length for members.
- Casual riders' ride length increases in the Spring and Summer months and decreases towards Winter.

Number of rides beginning at each hour of the day.

```
all_trips_v2 %>%
  group_by(hour_of_day = hour(round_date(start_time, 'hour'))) %>%
  group_by(hour_of_day, member_casual) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  ggplot(aes(x = hour_of_day, y = number_of_rides, fill = member_casual)) +
  geom_col(position = 'dodge') +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  scale_x_continuous(breaks = 0:23) +
  labs(title = "Number of rides for each hour of the day, by member type",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Time of day (Hour)",
       y = "Number of rides",
       fill = "Member type")
```

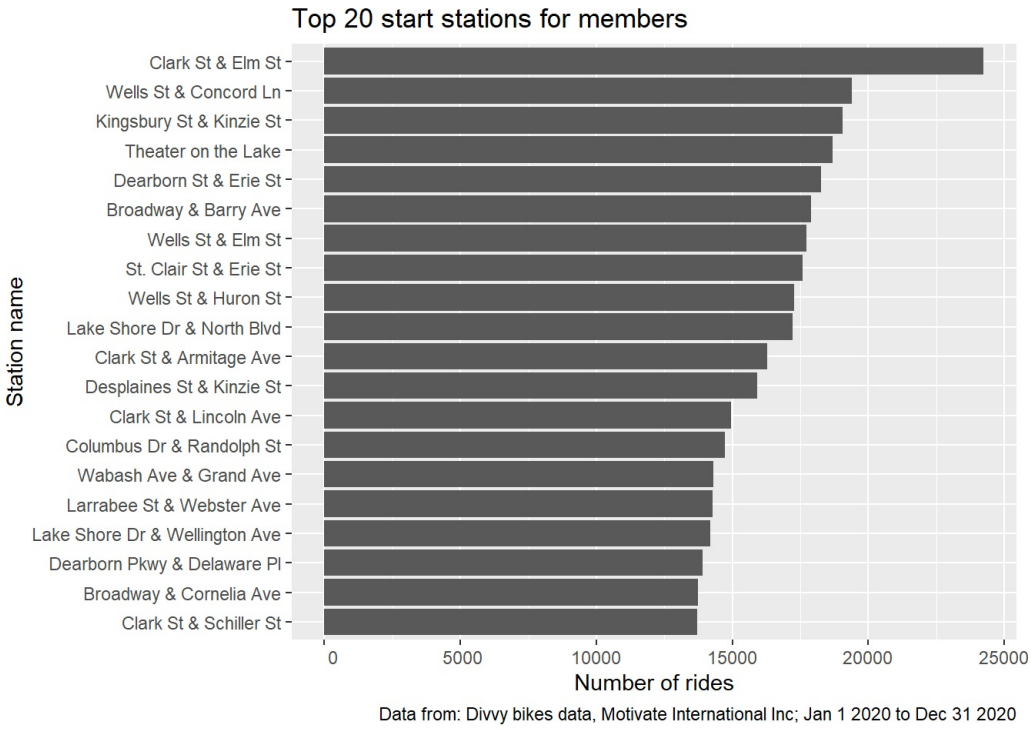


We can see from this chart that:

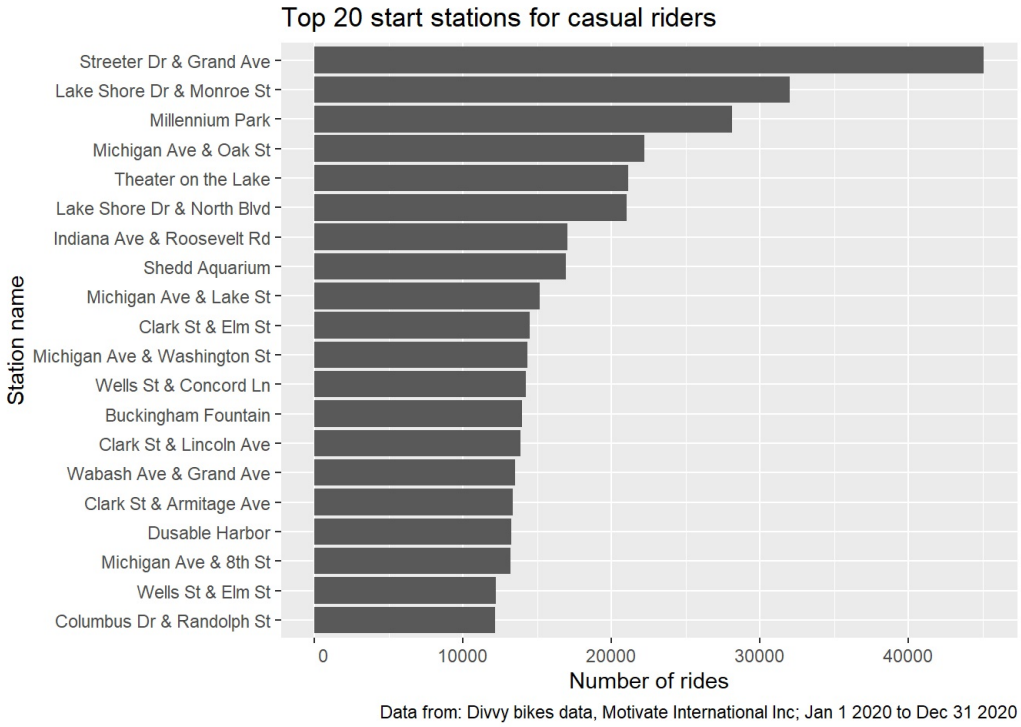
- The start times for members and casual riders begins to increase in the morning and peaks in the evening before decreasing overnight.
- Members ride start time increases heavily around 7-8am and 5-6pm which could mean that members are mostly using the service to commute to and from work.
- Casual riders seem to use the service more in the afternoon/evening.

Most frequently used start stations for members and casual riders.

```
# Creating a plot of popular start stations for members and casual riders.
all_trips_v2 %>%
  group_by(start_station_name, member_casual) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  filter(start_station_name != "", member_casual == "member") %>%
  arrange(-number_of_rides) %>%
  head(n = 20) %>%
  ggplot(aes(x = reorder(start_station_name, number_of_rides), y = number_of_rides)) +
  geom_col() +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  coord_flip() +
  labs(title = "Top 20 start stations for members",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Station name",
       y = "Number of rides")
```




```
all_trips_v2 %>%
  group_by(start_station_name, member_casual) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  filter(start_station_name != "", member_casual == "casual") %>%
  arrange(-number_of_rides) %>%
  head(n = 20) %>%
  ggplot(aes(x = reorder(start_station_name, number_of_rides), y = number_of_rides)) +
  geom_col() +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  coord_flip() +
  labs(title = "Top 20 start stations for casual riders",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "Station name",
       y = "Number of rides")
```



For members it should be noted that:

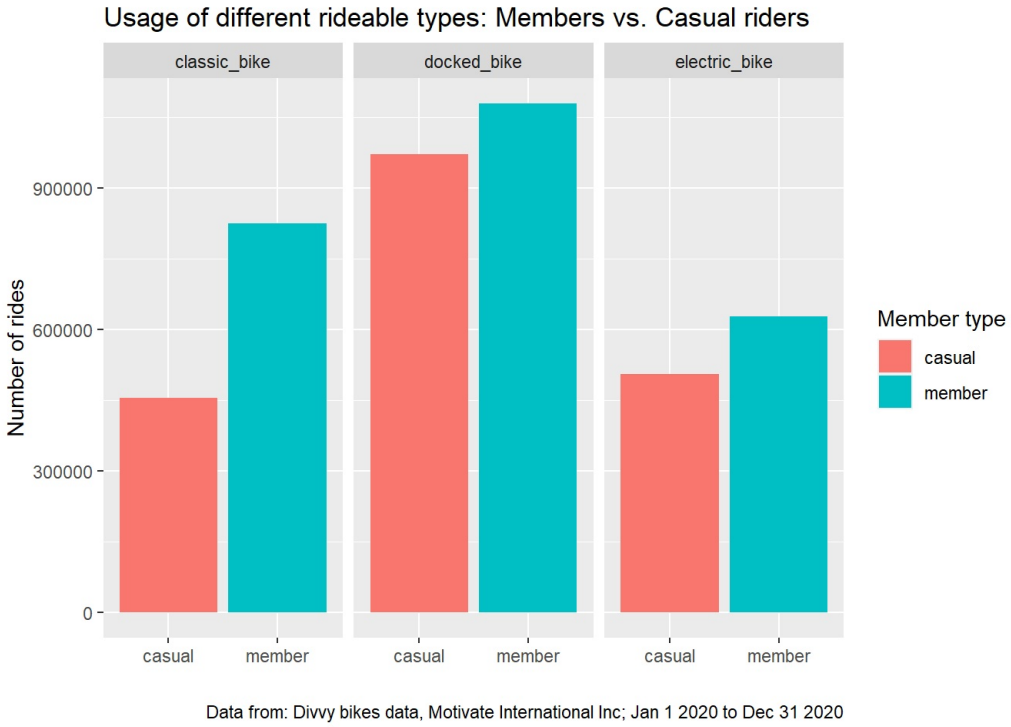
- A lot more members started trips at Clark St & Elm St than the next highest stations.
- The rest of the top 10 stations are relatively consistent.

For casual riders it should be noted that:

- The top 3 casual rider start stations have much more trips started than the others in the list.
- Theater on the lake is in the top 5 start stations for both members and casual riders.

Different rideable types for members and casual riders.

```
#Plotting usage of different rideable types for members and casual riders.
all_trips_v2 %>%
  group_by(rideable_type, member_casual) %>%
  summarise(number_of_rides = n(), .group = 'drop') %>%
  ggplot(aes(x = member_casual, y = number_of_rides, fill = member_casual)) +
  geom_col(position = 'dodge') +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
  facet_wrap(~rideable_type) +
  labs(title = "Usage of different rideable types: Members vs. Casual riders",
       caption = "Data from: Divvy bikes data, Motivate International Inc; Jan 1 2020 to Dec 31 2020",
       x = "",
       y = "Number of rides",
       fill = 'Member type')
```



This chart gives a few pieces of information:

- The docked bike is the most popular bike type among both members and casual riders.
- The classic bike is popular among members but is not very popular among casual riders.
- The electric bike is slightly more popular than the classic bike among casual riders, and is the least poular option for members.

Key findings

- The docked bike is very popular among casual riders.
- Casual riders usage of the service increase substantially in the Summer months, and on weekends in general.
- The afternoon and evening time periods are when the number of casual riders is highest.
- The top three start stations for casual riders are Skeeter Dr & Grand Ave, Lake Shore Dr & Monroe St, and Millenium park.
- It would appear that casual riders tend to use the service for leisure rides with some using it for commutes etc. while members seem to use the service mostly for work commutes with some leisure rides.

Reccomended actions

- Advertise membership to casual riders, mainly placing advertsements at heavy traffic starting stations (top 10 possibly).
- Offering a membership discount through Summer or on holiday weekends could be a good way to get more casual riders to sign up for memberships.
- Develop a marketing campaign focussing on the benefits of riding Cyclistic bikes as a way of getting to and from work or other repeat destinations as a way to get casual riders to use the service more, and possibly getting them to sign up for membership.