bike project

Askhat Bissembay

2022-10-05

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

Introduction

This exploratory analysis case study is towards Capstome project requirement for Google Data Analytics Professional Certificate. The case study involves a bikeshare company's data of its customer's trip details over a 12 month period (April 2020 - March 2021). The data has been made available by Motivate International Inc. under this license.

The analysis will follow the 6 phases of the Data Analysis process: Ask, Prepare, Process, Analyze, and Act (APPAA):

ASK

- Ask effective and narrative questions.
- Define the area of the analysis.
- Define what is our goal.

PREPARE

- Verify data's integrity.
- Check data reliability.
- Check data type.
- Merge datasets.

PROCESS

- Clean and Transform data.
- Remove unnecessary data.

ANALYZE

- Identify patterns.
- Draw coclusions.
- Make predictions.

SHARE

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

ACT

- Solve problems.
- Give recommendations based on insights.

1.ASK

Scenario

The marketing team needs to develop marketing strategies to turn casual riders into annual riders. However, to do this, the marketing analytics team needs to better understand the difference between casual riders and annual riders.

Consider key stakeholders:

- Team of marketing.
- Cyclistic team.

Deliverables

- Insights on how annual and casual riders differ.
- Provide visuals and relevant data to support insights.
- Use insights to give recommendations to convert casual riders to member riders.

2.PREPARE

Data sources

A total of **12 datasets** for each month starting from **April 2020 to February 2021**.This data that has been made publicly available has been scrubbed to omit rider's personal information.

General information.

The combined size of all the 12 datasets is 595 MB. Data cleaning in spreadsheets will be time-consuming and slow compared to R, cause totally datasets have >3M rows. That is why I choose R because I could do both data wrangling and visualization in R. It also give me opportunity for me to learn R better.

```
library(tidyverse)
```

Load libraries

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
```

```
##
##     chisq.test, fisher.test
```

```
library(dplyr)
library(skimr)
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```

str(bike04)
```

Load datasets

```
## 'data.frame':    84776 obs. of  13 variables:
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA7
##  $ 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 17:54:13" "2020-0
##  $ ended_at          : chr  "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 18:08:36" "2020-0
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "Califo
##  $ 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
##  $ 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" ...
```

```
str(bike05)
```

```
## 'data.frame':    200274 obs. of  13 variables:
##  $ ride_id           : chr  "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA
##  $ 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 14:11:03" "2020-0
##  $ ended_at          : chr  "2020-05-27 10:16:49" "2020-05-25 11:05:40" "2020-05-02 15:48:21" "2020-0
##  $ start_station_name: chr  "Franklin St & Jackson Blvd" "Clark St & Wrightwood Ave" "Kedzie Ave & Mi
##  $ 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
##  $ 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" ...
```

```
str(bike06)
```

```
## 'data.frame':    343005 obs. of  13 variables:
##  $ ride_id           : chr  "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85F
##  $ 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 17:12:41" "2020-0
##  $ ended_at          : chr  "2020-06-13 23:36:55" "2020-06-26 07:31:58" "2020-06-23 17:21:14" "2020-0
##  $ start_station_name: chr  "Wilton Ave & Belmont Ave" "Federal St & Polk St" "Daley Center Plaza" "
##  $ 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"
##  $ 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" ...
```

```
str(bike07)
```

```
## 'data.frame':    551480 obs. of  13 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B358
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-07-09 15:22:02" "2020-07-24 23:56:30" "2020-07-08 19:49:07" "2020-0
##  $ ended_at          : chr  "2020-07-09 15:25:52" "2020-07-25 00:20:17" "2020-07-08 19:56:22" "2020-0
##  $ start_station_name: chr  "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "Lake Shore Dr & Divers
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : chr  "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "Clark St & Wellington
##  $ 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     : chr  "member" "member" "casual" "casual" ...
```

```
str(bike08)
```

```
## 'data.frame':    622361 obs. of  13 variables:
##  $ ride_id           : chr  "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79FBBD412E578
##  $ rideable_type     : chr  "docked_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-08-20 18:08:14" "2020-08-27 18:46:04" "2020-08-26 19:44:14" "2020-0
##  $ ended_at          : chr  "2020-08-20 18:17:51" "2020-08-27 19:54:51" "2020-08-26 21:53:07" "2020-0
##  $ start_station_name: chr  "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Columbus Dr & 
##  $ 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 S
##  $ 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" ...
```

```
str(bike09)
```

```
## 'data.frame':    532958 obs. of  13 variables:
##  $ ride_id           : chr  "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB9
##  $ 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 15:09:04" "2020-0
##  $ ended_at          : chr  "2020-09-17 14:44:24" "2020-09-17 15:07:45" "2020-09-17 15:09:35" "2020-0
##  $ start_station_name: chr  "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N
##  $ 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
##  $ 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" ...
```

```
str(bike10)
```

```
## 'data.frame':    388653 obs. of  13 variables:
##  $ ride_id           : chr  "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE261B9E8
##  $ 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 23:00:01" "2020-
##  $ ended_at          : chr  "2020-10-31 19:57:12" "2020-11-01 00:04:16" "2020-10-31 23:08:22" "2020-
##  $ start_station_name: chr  "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave" "Stony Is
##  $ 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" "University Ave & 57
##  $ 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" ...
```

```
str(bike11)
```

```
## 'data.frame':    259716 obs. of  13 variables:
##  $ ride_id           : chr  "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533E89C32080B
##  $ 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 00:34:05" "2020-
##  $ ended_at          : chr  "2020-11-01 13:45:40" "2020-11-01 10:14:45" "2020-11-01 01:03:06" "2020-
##  $ start_station_name: chr  "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shore Dr & Mon
##  $ 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" "Federal St & Polk S
##  $ 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" ...
```

```
str(bike12)
```

```
## 'data.frame':    131573 obs. of  13 variables:
##  $ ride_id           : chr  "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE119628E44F87
##  $ 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 15:04:33" "2020-
##  $ ended_at          : chr  "2020-12-27 12:55:06" "2020-12-18 17:44:19" "2020-12-15 15:11:28" "2020-
##  $ 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" ...
```

```
str(bike13)
```

```
## 'data.frame':    96834 obs. of  13 variables:
##  $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377
##  $ 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 22:35:54" "2021-
##  $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-
##  $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St" "California Av
##  $ 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" ...
```

```
str(bike14)
```

```
## 'data.frame':    49622 obs. of  13 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E7
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 19:10:18" "2021-
##  $ ended_at          : chr  "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 19:19:10" "2021-
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake S
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Rand
##  $ 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" ...
```

```
str(bike15)
```

```
## 'data.frame':    228496 obs. of  13 variables:
##  $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-03-16 08:32:30" "2021-03-28 01:26:28" "2021-03-11 21:17:29" "2021-
##  $ ended_at          : chr  "2021-03-16 08:36:34" "2021-03-28 01:36:55" "2021-03-11 21:33:53" "2021-
##  $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields A
##  $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted
##  $ 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" ...
```

Data tramsformation and Cleaning We can see that datatypes of *start_station_id & end_start_station_id* in all datasets are not same. In some it *int* type and in some *char* type. Lets convert them from *int* to *char* datatype.

```
bike04 <- bike04 %>% mutate(start_station_id=as.character(start_station_id),
                         end_station_id=as.character(end_station_id))
bike05 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike06 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike07 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike08 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike09 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike10 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
bike11 <- bike05 %>% mutate(start_station_id=as.character(start_station_id),
                            end_station_id=as.character(end_station_id))
```

3.PROCESS

```

bike <- bind_rows(bike04, bike05, bike06, bike07, bike08,
                  bike09, bike10, bike11, bike12, bike13,
                  bike14, bike15)

str(bike)
```

Combine all the datasets into one dataframe.

```
## 'data.frame':    1993219 obs. of  13 variables:
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA7
##  $ 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 17:54:13" "2020-
##  $ ended_at          : chr  "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 18:08:36" "2020-
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "Calif
##  $ start_station_id  : chr  "86" "503" "142" "216" ...
##  $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt
##  $ 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" ...
```

Transform and Cleaning *started_at* & *ended_at* should be datetime datatype isntead of char. Lets convert it.

```
bike$started_at <- ymd_hms(bike$started_at)
bike$ended_at <- ymd_hms(bike$ended_at)
str(bike)
```

```
## 'data.frame':    1993219 obs. of  13 variables:
##  $ ride_id           : chr  "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA7
##  $ 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" ...
##  $ ended_at          : POSIXct, format: "2020-04-26 18:12:03" "2020-04-17 17:17:03" ...
##  $ start_station_name: chr  "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "Calif
##  $ start_station_id  : chr  "86" "503" "142" "216" ...
##  $ end_station_name  : chr  "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt
##  $ 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" ...
```

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

Remove columns which not required or beyond of the project

```
## Rows: 1,993,219
## Columns: 9
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04--~
## $ ended_at           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04--~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "38~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

```
bike <- bike %>% rename(ride_type=rideable_type,
                        start_time=started_at,
                        end_time=ended_at,
                        customer_type=member_casual)
bike$timedif.min <- round(difftime(bike$end_time,bike$start_time, units = "mins"), digit=2)
```

```
glimpse(bike)
```

Rename columns for better readability

```
## Rows: 1,993,219
## Columns: 10
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ ride_type          <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ start_time         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04--~
## $ end_time           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04--~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "38~
## $ customer_type      <chr> "member", "member", "member", "member", "casual", "~
## $ timedif.min        <drtn> 26.82 mins, 8.15 mins, 14.38 mins, 12.20 mins, 52.~
```

```
bike$day_of_the_week <- format(as.Date(bike$start_time),'%A')
```

```
bike$month <- format(as.Date(bike$start_time),'%B_%Y')
bike$time <- as.POSIXct(bike$start_time, format="%H:%M")
```

Lets change time datatype into more useful type *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), the date is of no interest to us,only the hours-minutes-seconds are*

```
glimpse(bike)
```

```
## Rows: 1,993,219
## Columns: 13
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ ride_type          <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ start_time         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-~
## $ end_time           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-04-~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635", "38~
## $ customer_type      <chr> "member", "member", "member", "member", "casual", "~
## $ timedif.min        <drtn> 26.82 mins, 8.15 mins, 14.38 mins, 12.20 mins, 52.~
## $ day_of_the_week    <chr> "Sunday", "Friday", "Wednesday", "Tuesday", "Saturd~
## $ month              <chr> "April_2020", "April_2020", "April_2020", "April_20~
## $ time               <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-04-~
```

```

nrow(subset(bike,timedif.min < 0))
```

How many trips lenghts less than 0

```
## [1] 1693
```

1693 rows with time trip is less than 0 and lets delete them from our dataframe as they contribute only less than 0.001% of the total rows

```
bike <- bike %>% filter(timedif.min > 0)
```

4&5 Analyze and Share the Data

The dataframe is now ready for descriptive analysis and we are ready to search insights, lets look casual riders and member use bicycle. Firstly, lets look commonly to statistical summary of trip duration and number of trips.

```
#Group data by customer type
table(bike$customer_type)
```

```
##
##  casual  member
##  773740 1217668
```

```
#Statistical summary
summary(as.integer(bike$timedif.min))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     8.00    16.00    29.57    29.00 58720.00
```

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

```
## # A tibble: 2 x 5
##   customer_type min_trip_duration max_trip_duration median_trip_duration mean_~1
##   <chr>         <drtn>            <drtn>            <drtn>               <drtn>
## 1 casual        0.02 mins         55683.88 mins     24.87 mins           48.904~
## 2 member        0.02 mins         58720.03 mins     13.08 mins           18.084~
## # ... with abbreviated variable name 1: mean_trip_duration
```

The mean trip duration of member riders(18.08) is lower than the mean trip duration of all trips(29.57 min), while casual riders(48.90 min) is exactly opposite.This tells us that casual riders usually take the bikes out for a longer duration compared to members.

Average duration and number of trips by day of the week Firstly, lets fix the order of the day of the week

```
## fix for the day_of_the_week variable so that they show up
bike$day_of_the_week <- factor(bike$day_of_the_week, levels= c("Sunday", "Monday", "Tuesday",
                                                                "Wednesday", "Thursday", "Friday",
```

```
#statistical summary of average trip duration by day of the week and customer type
bike %>% group_by(customer_type,day_of_the_week) %>%
  summarise(number_of_rides = n(),average_duration_mins=mean(timedif.min)) %>%
  arrange(customer_type,desc(number_of_rides))
```

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

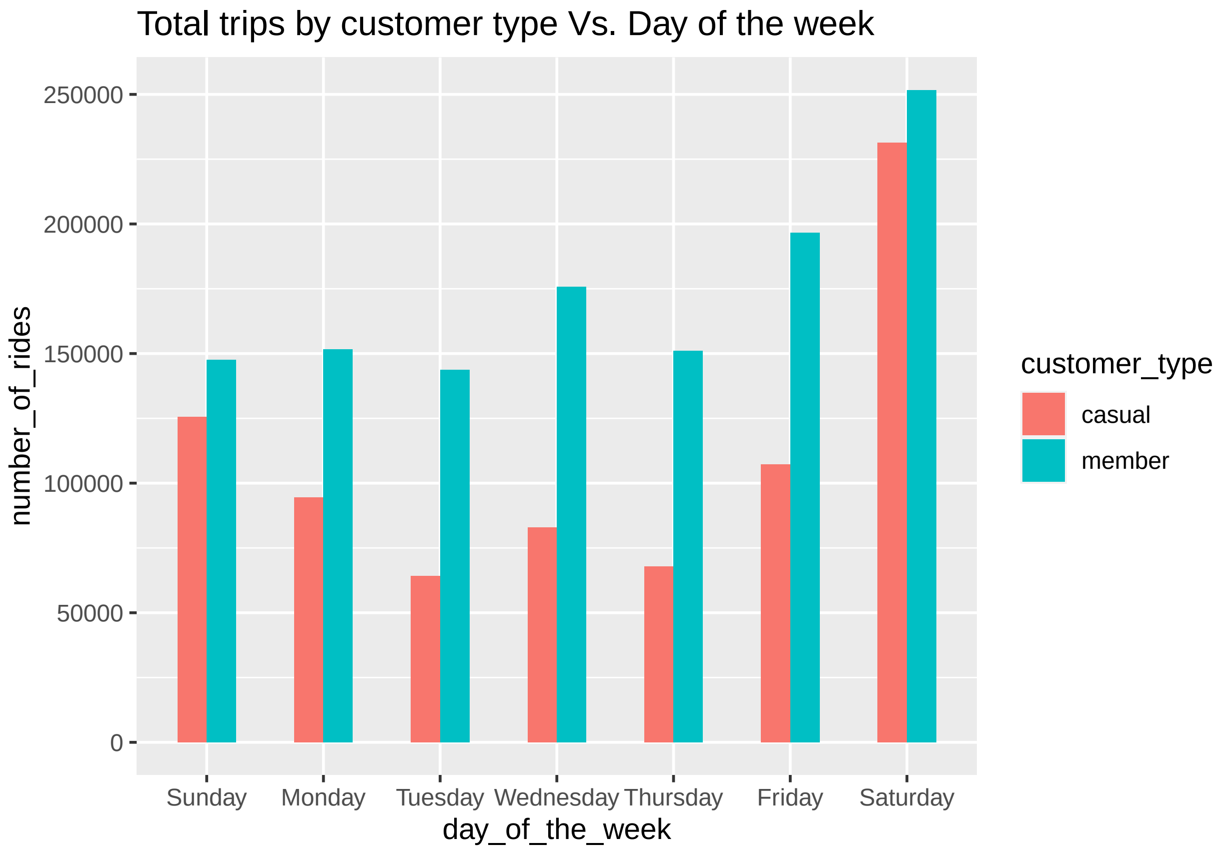
```
## # A tibble: 14 x 4
## # Groups:   customer_type [2]
##    customer_type day_of_the_week number_of_rides average_duration_mins
##    <chr>         <fct>                     <int> <drtn>
##  1 casual        Saturday                 231414 48.27960 mins
##  2 casual        Sunday                   125592 52.96178 mins
##  3 casual        Friday                   107306 49.14532 mins
##  4 casual        Monday                    94566 51.21581 mins
##  5 casual        Wednesday                 82879 45.33693 mins
##  6 casual        Thursday                  67805 45.25676 mins
##  7 casual        Tuesday                   64178 47.87430 mins
##  8 member        Saturday                 251697 20.53166 mins
##  9 member        Friday                   196518 17.69571 mins
## 10 member        Wednesday                175659 16.94463 mins
## 11 member        Monday                   151639 17.32118 mins
## 12 member        Thursday                 150971 16.56836 mins
## 13 member        Sunday                   147537 20.03839 mins
## 14 member        Tuesday                  143647 16.11489 mins
```

```

bike %>%
  group_by(customer_type,day_of_the_week) %>%
  summarise(number_of_rides = n()) %>%
  arrange(customer_type,desc(number_of_rides)) %>%
  ggplot(aes(x = day_of_the_week,y = number_of_rides,fill=customer_type))+geom_bar(stat='identity') %>%
  labs(title ="Total trips by customer type Vs. Day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5))
```

Visualization

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



From the bar graph above, member customers are most busy on Saturday, while casual members use mostly on Saturday and Sunday. Interesting pattern to note though is the member users during week use bikeshare services approximately evenly except on Saturday.

```
bike$month <- as.factor(bike$month)
#Average number of trips by customer type and month
unique(bike$month)
```

Average duration of trips aby month

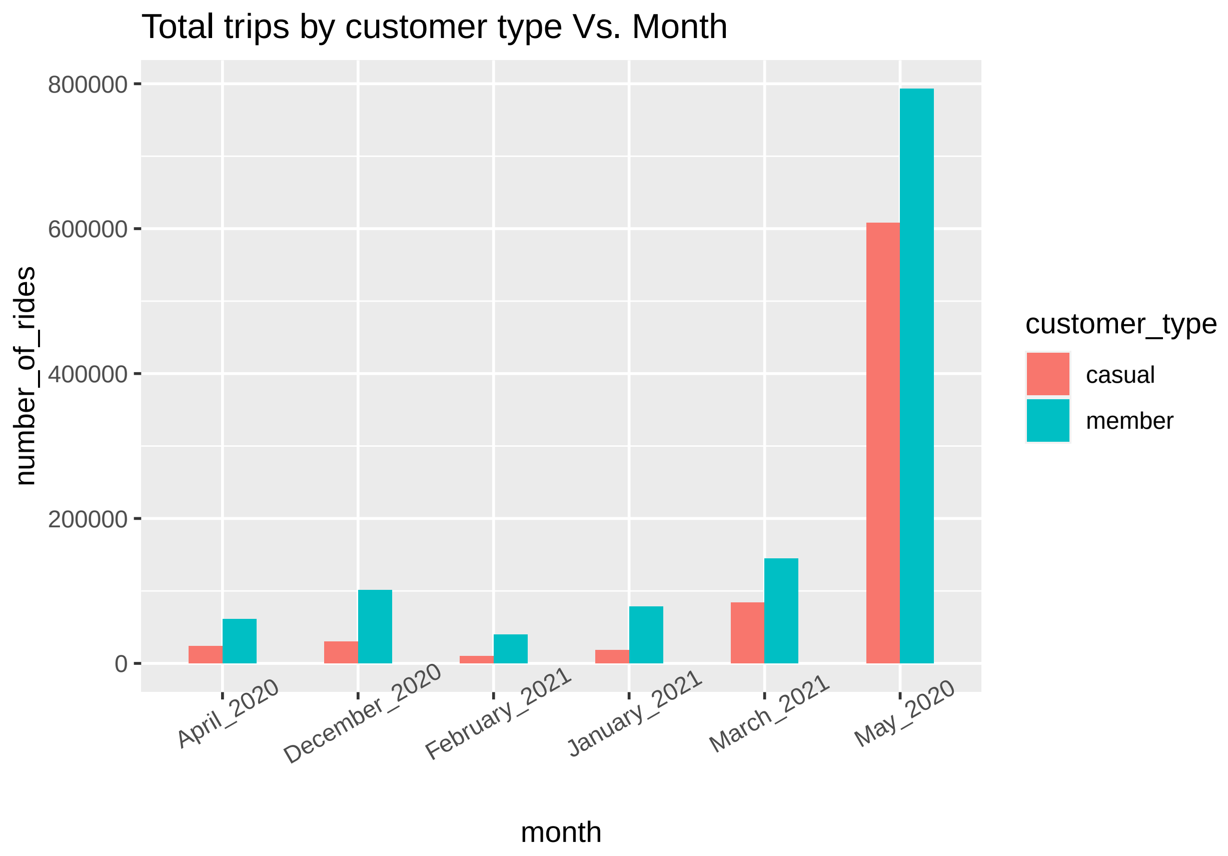
```
## [1] April_2020    May_2020      December_2020 January_2021  February_2021
## [6] March_2021
## 6 Levels: April_2020 December_2020 February_2021 January_2021 ... May_2020
```

```
bike %>% group_by(customer_type,month) %>%
  summarise(number_of_riders=n(),average_duration_mins=mean(timedif.min)) %>%
  arrange(customer_type,desc(number_of_riders))
```

```
## `summarise()` has grouped output by 'customer_type'. You can override using the
## `.groups` argument.
## # A tibble: 12 x 4
## # Groups:   customer_type [2]
##    customer_type month         number_of_riders average_duration_mins
##    <chr>         <fct>                    <int> <drtn>
##  1 casual        May_2020                607866 51.22107 mins
##  2 casual        March_2021               84028 38.16100 mins
##  3 casual        December_2020            29994 26.85230 mins
##  4 casual        April_2020               23605 73.14255 mins
##  5 casual        January_2021             18117 25.68458 mins
##  6 casual        February_2021            10130 49.37811 mins
##  7 member        May_2020                792764 19.77339 mins
##  8 member        March_2021              144456 13.97064 mins
##  9 member        December_2020           101137 12.74996 mins
## 10 member        January_2021             78711 12.87297 mins
## 11 member        April_2020               61112 21.48035 mins
## 12 member        February_2021            39488 18.02346 mins
```

```
# Visualization by month
bike %>% 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))+geom_bar(stat = 'identity') %>%
  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
## `.groups` argument.
```

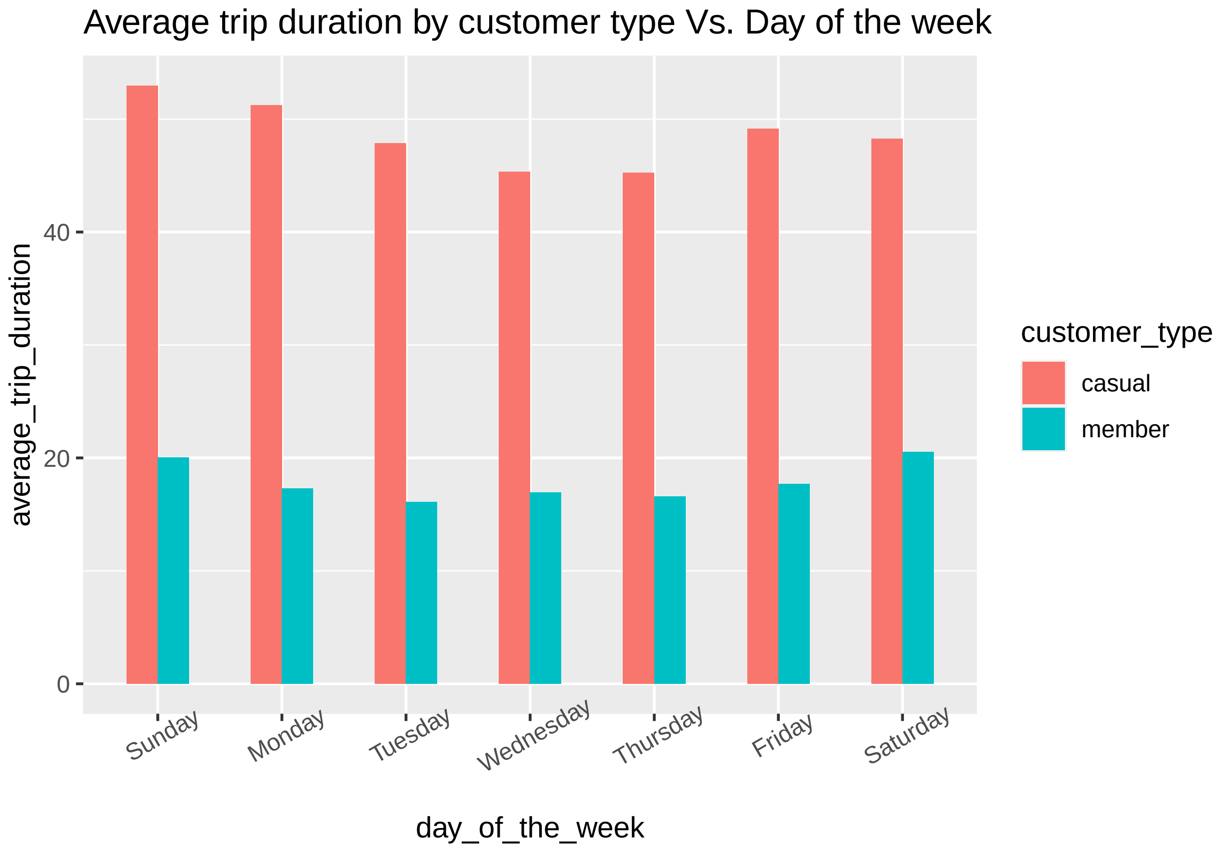


The data shows that the months of May 2020 are the most busy time of the year among both members and casual riders. This could be attributed to an external factor (eg. cold weather, major quality issue) that might have hindered with customer needs. 2021 is a tough year when Covid comes. People care more 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.

```
bike %>% group_by(customer_type,day_of_the_week) %>%
  summarise(average_trip_duration=mean(timedif.min)) %>%
  ggplot(aes(x=day_of_the_week,y=average_trip_duration,fill=customer_type))+geom_bar(stat='identity') %
  labs(title ="Average trip duration by customer type Vs. Day of the week") +
  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))
```

Visualization of average trip duration by day of the week

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



The average trip duration of a casual driver is more than twice that of a member. Note that this does not necessarily mean that casual riders travel a greater distance.It is also interesting that during all week average trip duration are approximately same for all types of riders about 50 min for casual and about 19 min for member

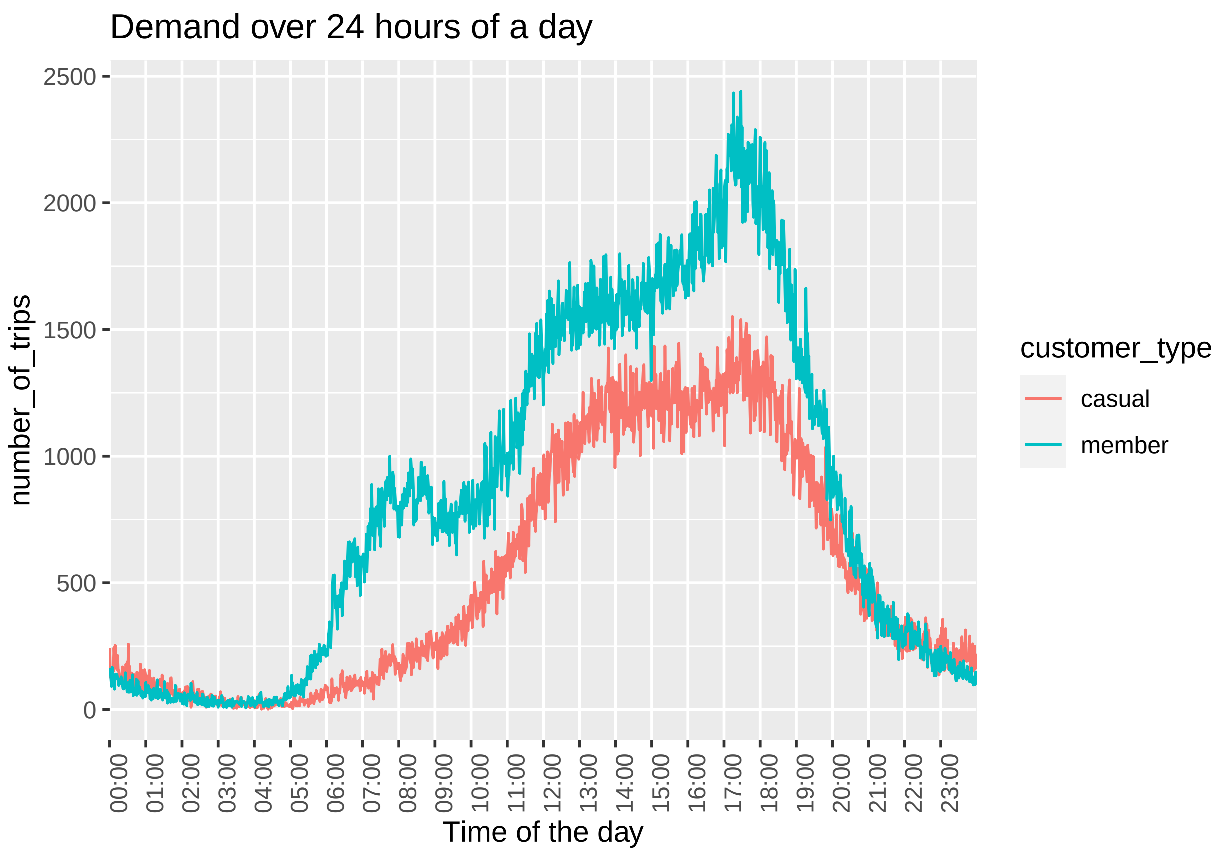
```
bike$time <- format(as.POSIXct(strptime(bike$time, "%Y-%m-%d  %H:%M:%S",tz="")) ,format = "%H:%M")
bike$time <- as.POSIXct(bike$time, format = "%H:%M")

bike %>%
  group_by(customer_type, time) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x = time, y = number_of_trips, color = customer_type, group = customer_type)) +
  geom_line(stat='identity') + scale_x_datetime(date_breaks = "1 hour", minor_breaks = NULL, date_label
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title ="Demand over 24 hours of a day", x = "Time of the day")
```

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

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

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

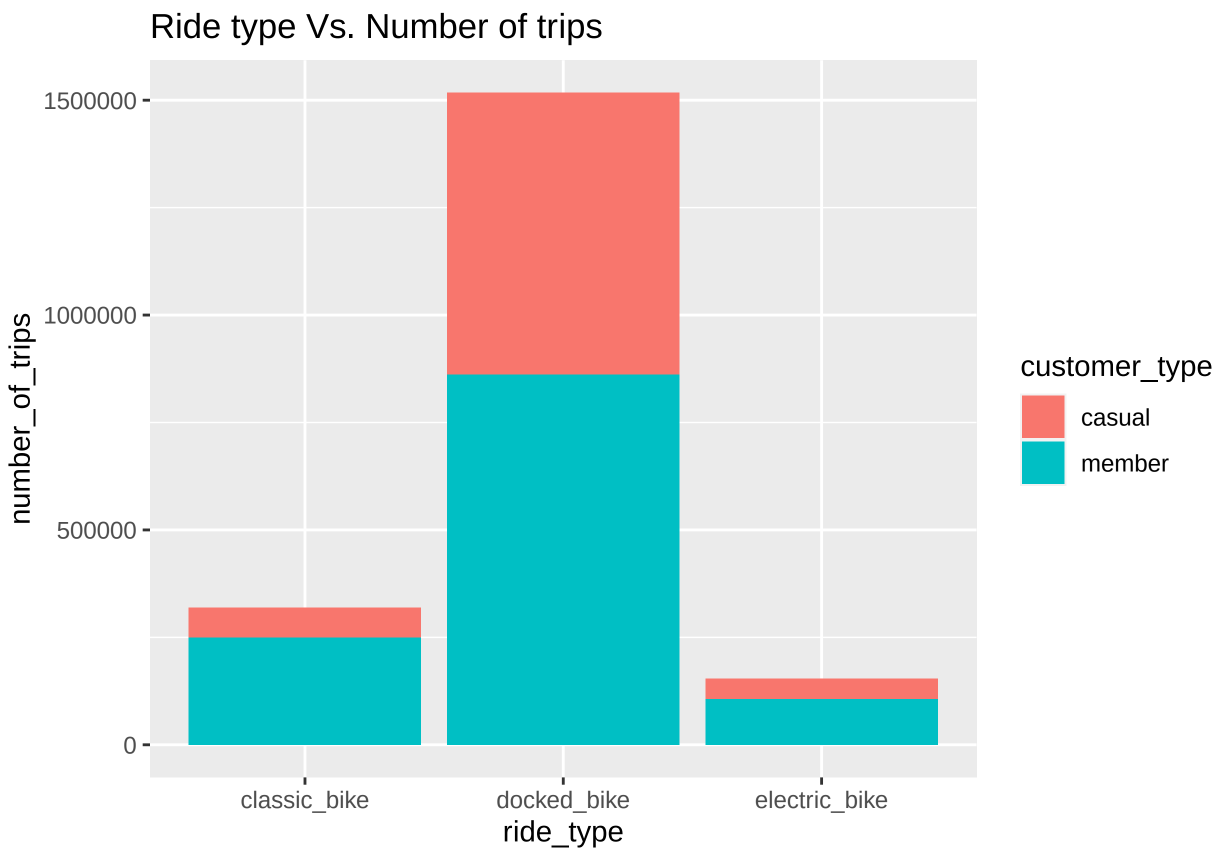


For the members, there seems to be distinct peak demand hour ***5-7 PM****, this also coincides with the peak demand hours of casual riders. It could probably be assumed that workers make up the majority of the participant profile due to the need for evening hours, but we need more data to substantiate this assumption.*

```
bike %>%
  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")
```

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

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



We see that Docked bike is much more used by customers, this confirms our suggestion that bicycles are mainly used by workers. And our task is to increase the number of member users of docked bike, cause we see casual users use more often docked bike

6.ACT

Key takeaways

- Casual customers use bikeshare services more during weekends, almost twice on Sunday and for 50% often than average weekday,while members use them consistently over the Sunday approximately for 30% more.
- Total trip by customer Vs. Month didn't give us important information, cause like we wrote before reason why May of 2020 has huge different may be (cold weather, major quality issue) that might have hindered with customer needs. 2021 is a tough year when Covid comes.
- Average trip duration of casual riders is more than twice that of member rider over any given day of the week.
- Both customers prefer ride a bike between 12.00 - 19.00.
- Both riders prefer docked bike, almost three times more.

Recommendations

- This data could be used to study the category of riders who can be targeted for attracting new members.
- Offer promotions to casual and member riders on weekdays to ensure that customers use bikeshare service evenly throught the week.
- Provide discount ride before 12.00 and after 19.00, so riders might choose to use bikes more often.

- Offer discounted membership fee for docked bike casual riders, it might nudge thisriders to buy membership with discount.

Additional data that could formore detailed analysis.

- Location of bike station - with this data we could eliminate the shortage or oversupply of bicycles.
- Pricing details - could build discount systems.
- Age and genders - help us for targeting for attracting new members.
- Weather data - how does the weather affect bike rental.

— END OF CASE STUDY—