

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

Bhaskar Sutar

2022-10-26

Introduction

This exploratory analysis case study is towards Capstone 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 (January 2021 - December 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).

A brief explanation of APPAA:

Ask

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

Prepare

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

Process

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

Analyze

- Identify patterns
- Draw conclusions
- Make predictions

Share

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

Act

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

1. Ask

Scenario

Marketing team needs to design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ

Stakeholders:

- Director of marketing
- Cyclistic executive team

Objective

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

Deliverables:

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

2. Prepare

Data Sources

A total of 12 datasets have been made available for each month starting from January 2021 to December 2021. Each dataset captures the details of every ride logged by the customers of Cyclistic. This data that has been made publicly available has been scrubbed to omit rider's personal information.

Documentation, Cleaning and Preparation of data for analysis

The combined size of all the 12 datasets is close to 1 GB. Data cleaning in spreadsheets will be time-consuming and slow compared to SQL or R. I am choosing R simply because I could do both data wrangling and analysis/ visualizations in the same platform. It is also an opportunity for me to learn R better.

```
library(tidyverse)
```

```
## — Attaching packages — tidyverse 1.3.2 —
## ✓ ggplot2 3.3.6      ✓ purrr  0.3.5
## ✓ tibble  3.1.8      ✓ dplyr  1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## — Conflicts — tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

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

```
library(dplyr)
library(readr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##   chisq.test, fisher.test
```

```
library(data.table)
```

```
## data.table 1.14.2 using 2 threads (see ?getDTthreads). Latest news: r-datatable.com
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter,
##     second, wday, week, yday, year
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(tidyr)
```

Load datasets

```
trip21_Jan <- read_csv("D:\\files\\202101-divvy-tripdata.csv")
```

```
## Rows: 96834 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm   (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Feb <- read_csv("D:\\files\\202102-divvy-tripdata.csv")
```

```
## Rows: 49622 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm   (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Mar <- read_csv("D:\\files\\202103-divvy-tripdata.csv")
```

```
## Rows: 228496 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr (7): ride_id, rideable_type, start_station_name, star...
## dbl (4): start_lat, start_lng, end_lat, end_lng
## dtm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Apr <- read_csv("D:\\files\\202104-divvy-tripdata.csv")
```

```
## Rows: 337230 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr (7): ride_id, rideable_type, start_station_name, star...
## dbl (4): start_lat, start_lng, end_lat, end_lng
## dtm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_May <- read_csv("D:\\files\\202105-divvy-tripdata.csv")
```

```
## Rows: 531633 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr (7): ride_id, rideable_type, start_station_name, star...
## dbl (4): start_lat, start_lng, end_lat, end_lng
## dtm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Jun <- read_csv("D:\\files\\202106-divvy-tripdata.csv")
```

```
## Rows: 729595 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr (7): ride_id, rideable_type, start_station_name, star...
## dbl (4): start_lat, start_lng, end_lat, end_lng
## dtm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Jul <- read_csv("D:\\files\\202107-divvy-tripdata.csv")
```

```
## Rows: 822410 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm   (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Aug <- read_csv("D:\\files\\202108-divvy-tripdata.csv")
```

```
## Rows: 804352 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm   (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Sep <- read_csv("D:\\files\\202109-divvy-tripdata.csv")
```

```
## Rows: 756147 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm   (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Oct <- read_csv("D:\\files\\202110-divvy-tripdata.csv")
```

```
## Rows: 631226 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm  (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Nov <- read_csv("D:\\files\\202111-divvy-tripdata.csv")
```

```
## Rows: 359978 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm  (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
trip21_Dec <- read_csv("D:\\files\\202112-divvy-tripdata.csv")
```

```
## Rows: 247540 Columns: 13
## — Column specification —————
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, star...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dtm  (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Check column names of each dataset for consistency

```
colnames(trip21_Jan)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name" "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Feb)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Mar)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Apr)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_May)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Jun)
```



```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Jul)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Aug)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Sep)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Oct)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Nov)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

```
colnames(trip21_Dec)
```

```
## [1] "ride_id"          "rideable_type"
## [3] "started_at"       "ended_at"
## [5] "start_station_name" "start_station_id"
## [7] "end_station_name"  "end_station_id"
## [9] "start_lat"        "start_lng"
## [11] "end_lat"          "end_lng"
## [13] "member_casual"
```

Check data structures and data types for all data frames

```
str(trip21_Jan)
```

```
## spec_tbl_df [96,834 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27"
## $ rideable_type    : chr [1:96834] "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
## $ started_at       : POSIXct[1:96834], format: "2021-01-23 16:14:19" ...
## $ ended_at         : POSIXct[1:96834], format: "2021-01-23 16:24:44" ...
## $ start_station_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" ...
## $ start_station_id  : chr [1:96834] "17660" "17660" "17660" "17660" ...
## $ end_station_name  : chr [1:96834] NA NA NA NA ...
## $ end_station_id    : chr [1:96834] NA NA NA NA ...
## $ start_lat         : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng         : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
## $ end_lat           : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng           : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
## $ member_casual     : chr [1:96834] "member" "member" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Feb)
```

```
## spec_tbl_df [49,622 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91"
## $ rideable_type    : chr [1:49622] "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
## $ started_at       : POSIXct[1:49622], format: "2021-02-12 16:14:56" ...
## $ ended_at         : POSIXct[1:49622], format: "2021-02-12 16:21:43" ...
## $ start_station_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St" "Wood St & Chicago Ave" ...
## $ start_station_id  : chr [1:49622] "525" "525" "KA1503000012" "637" ...
## $ end_station_name  : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Randolph St" "Honore St & Division St" ...
## $ end_station_id    : chr [1:49622] "660" "16806" "TA1305000029" "TA1305000034" ...
## $ start_lat         : num [1:49622] 42 42 41.9 41.9 41.8 ...
## $ start_lng         : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
## $ end_lat          : num [1:49622] 42 42 41.9 41.9 41.8 ...
## $ end_lng          : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
## $ member_casual     : chr [1:49622] "member" "casual" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Mar)
```

```
## spec_tbl_df [228,496 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A28
4" "994D05AA75A168F2" ...
## $ rideable_type    : chr [1:228496] "classic_bike" "classic_bike" "classic_bike" "classic_b
ike" ...
## $ started_at       : POSIXct[1:228496], format: "2021-03-16 08:32:30" ...
## $ ended_at         : POSIXct[1:228496], format: "2021-03-16 08:36:34" ...
## $ start_station_name: chr [1:228496] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitag
e Ave" "Shields Ave & 28th Pl" "Winthrop Ave & Lawrence Ave" ...
## $ start_station_id  : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...
## $ end_station_name  : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomingd
ale Ave" "Halsted St & 35th St" "Broadway & Sheridan Rd" ...
## $ end_station_id    : chr [1:228496] "13266" "18017" "TA1308000043" "13323" ...
## $ start_lat         : num [1:228496] 41.9 41.9 41.8 42 42 ...
## $ start_lng         : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...
## $ end_lat           : num [1:228496] 41.9 41.9 41.8 42 42.1 ...
## $ end_lng           : num [1:228496] -87.7 -87.7 -87.6 -87.6 -87.7 ...
## $ member_casual     : chr [1:228496] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Apr)
```

```
## spec_tbl_df [337,230 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:337230] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BA
D" "1887262AD101C604" ...
## $ rideable_type    : chr [1:337230] "classic_bike" "docked_bike" "docked_bike" "classic_bik
e" ...
## $ started_at       : POSIXct[1:337230], format: "2021-04-12 18:25:36" ...
## $ ended_at         : POSIXct[1:337230], format: "2021-04-12 18:56:55" ...
## $ start_station_name: chr [1:337230] "State St & Pearson St" "Dorchester Ave & 49th St" "Loo
mis Blvd & 84th St" "Honore St & Division St" ...
## $ start_station_id  : chr [1:337230] "TA1307000061" "KA1503000069" "20121" "TA1305000034"
...
## $ end_station_name  : chr [1:337230] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th S
t" "Loomis Blvd & 84th St" "Southport Ave & Waveland Ave" ...
## $ end_station_id    : chr [1:337230] "13235" "KA1503000069" "20121" "13235" ...
## $ start_lat         : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
## $ start_lng         : num [1:337230] -87.6 -87.6 -87.7 -87.7 -87.7 ...
## $ end_lat           : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
## $ end_lng           : num [1:337230] -87.7 -87.6 -87.7 -87.7 -87.7 ...
## $ member_casual     : chr [1:337230] "member" "casual" "casual" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_May)
```

```
## spec_tbl_df [531,633 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC
2" "7881AC6D39110C60" ...
## $ rideable_type    : chr [1:531633] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:531633], format: "2021-05-30 11:58:15" ...
## $ ended_at         : POSIXct[1:531633], format: "2021-05-30 12:10:39" ...
## $ start_station_name: chr [1:531633] NA NA NA NA ...
## $ start_station_id  : chr [1:531633] NA NA NA NA ...
## $ end_station_name  : chr [1:531633] NA NA NA NA ...
## $ end_station_id    : chr [1:531633] NA NA NA NA ...
## $ start_lat         : num [1:531633] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng         : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
## $ end_lat           : num [1:531633] 41.9 41.8 41.9 41.9 41.9 ...
## $ end_lng           : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
## $ member_casual     : chr [1:531633] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Jun)
```

```
## spec_tbl_df [729,595 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:729595] "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF
2" "B03C0FE48C412214" ...
## $ rideable_type    : chr [1:729595] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:729595], format: "2021-06-13 14:31:28" ...
## $ ended_at         : POSIXct[1:729595], format: "2021-06-13 14:34:11" ...
## $ start_station_name: chr [1:729595] NA NA NA NA ...
## $ start_station_id  : chr [1:729595] NA NA NA NA ...
## $ end_station_name  : chr [1:729595] NA NA NA NA ...
## $ end_station_id    : chr [1:729595] NA NA NA NA ...
## $ start_lat         : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
## $ start_lng         : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
## $ end_lat           : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
## $ end_lng           : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
## $ member_casual     : chr [1:729595] "member" "member" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Jul)
```



```
## spec_tbl_df [822,410 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427
A" "379B58EAB20E8AA5" ...
## $ rideable_type    : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bi
ke" ...
## $ started_at       : POSIXct[1:822410], format: "2021-07-02 14:44:36" ...
## $ ended_at         : POSIXct[1:822410], format: "2021-07-02 15:19:58" ...
## $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez
St" "Wabash Ave & 16th St" "California Ave & Cortez St" ...
## $ start_station_id  : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
## $ end_station_name  : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St"
"Rush St & Hubbard St" "Carpenter St & Huron St" ...
## $ end_station_id    : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
## $ start_lat         : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng         : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ end_lat           : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng           : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
## $ member_casual     : chr [1:822410] "casual" "casual" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Aug)
```

```
## spec_tbl_df [804,352 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234
D" "5834D3208BFAF1DA" ...
## $ rideable_type    : chr [1:804352] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:804352], format: "2021-08-10 17:15:49" ...
## $ ended_at         : POSIXct[1:804352], format: "2021-08-10 17:22:44" ...
## $ start_station_name: chr [1:804352] NA NA NA NA ...
## $ start_station_id  : chr [1:804352] NA NA NA NA ...
## $ end_station_name  : chr [1:804352] NA NA NA NA ...
## $ end_station_id    : chr [1:804352] NA NA NA NA ...
## $ start_lat         : num [1:804352] 41.8 41.8 42 42 41.8 ...
## $ start_lng         : num [1:804352] -87.7 -87.7 -87.7 -87.7 -87.6 ...
## $ end_lat           : num [1:804352] 41.8 41.8 42 42 41.8 ...
## $ end_lng           : num [1:804352] -87.7 -87.6 -87.7 -87.7 -87.6 ...
## $ member_casual     : chr [1:804352] "member" "member" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Sep)
```

```
## spec_tbl_df [756,147 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB91
0" "78D1DE133B3DBF55" ...
## $ rideable_type    : chr [1:756147] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:756147], format: "2021-09-28 16:07:10" ...
## $ ended_at         : POSIXct[1:756147], format: "2021-09-28 16:09:54" ...
## $ start_station_name: chr [1:756147] NA NA NA NA ...
## $ start_station_id  : chr [1:756147] NA NA NA NA ...
## $ end_station_name  : chr [1:756147] NA NA NA NA ...
## $ end_station_id    : chr [1:756147] NA NA NA NA ...
## $ start_lat         : num [1:756147] 41.9 41.9 41.8 41.8 41.9 ...
## $ start_lng         : num [1:756147] -87.7 -87.6 -87.7 -87.7 -87.7 ...
## $ end_lat           : num [1:756147] 41.9 42 41.8 41.8 41.9 ...
## $ end_lng           : num [1:756147] -87.7 -87.7 -87.7 -87.7 -87.7 ...
## $ member_casual     : chr [1:756147] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Oct)
```

```
## spec_tbl_df [631,226 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id           : chr [1:631226] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE1
4" "362947F0437E1514" ...
## $ rideable_type     : chr [1:631226] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:631226], format: "2021-10-22 12:46:42" ...
## $ ended_at         : POSIXct[1:631226], format: "2021-10-22 12:49:50" ...
## $ start_station_name: chr [1:631226] "Kingsbury St & Kinzie St" NA NA NA ...
## $ start_station_id  : chr [1:631226] "KA1503000043" NA NA NA ...
## $ end_station_name  : chr [1:631226] NA NA NA NA ...
## $ end_station_id    : chr [1:631226] NA NA NA NA ...
## $ start_lat         : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng         : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
## $ end_lat           : num [1:631226] 41.9 41.9 41.9 41.9 41.9 ...
## $ end_lng           : num [1:631226] -87.6 -87.7 -87.7 -87.7 -87.7 ...
## $ member_casual     : chr [1:631226] "member" "member" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Nov)
```

```
## spec_tbl_df [359,978 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id          : chr [1:359978] "7C00A93E10556E47" "90854840DFD508BA" "0A7D10CDD144061
C" "2F3BE33085BCFF02" ...
## $ rideable_type    : chr [1:359978] "electric_bike" "electric_bike" "electric_bike" "electr
ic_bike" ...
## $ started_at       : POSIXct[1:359978], format: "2021-11-27 13:27:38" ...
## $ ended_at         : POSIXct[1:359978], format: "2021-11-27 13:46:38" ...
## $ start_station_name: chr [1:359978] NA NA NA NA ...
## $ start_station_id  : chr [1:359978] NA NA NA NA ...
## $ end_station_name  : chr [1:359978] NA NA NA NA ...
## $ end_station_id    : chr [1:359978] NA NA NA NA ...
## $ start_lat         : num [1:359978] 41.9 42 42 41.9 41.9 ...
## $ start_lng         : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ end_lat           : num [1:359978] 42 41.9 42 41.9 41.9 ...
## $ end_lng           : num [1:359978] -87.7 -87.7 -87.7 -87.8 -87.6 ...
## $ member_casual     : chr [1:359978] "casual" "casual" "casual" "casual" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
str(trip21_Dec)
```

```
## spec_tbl_df [247,540 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ ride_id           : chr [1:247540] "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C
5" "3278BA87BF698339" ...
## $ rideable_type     : chr [1:247540] "electric_bike" "electric_bike" "electric_bike" "classi
c_bike" ...
## $ started_at       : POSIXct[1:247540], format: "2021-12-07 15:06:07" ...
## $ ended_at         : POSIXct[1:247540], format: "2021-12-07 15:13:42" ...
## $ start_station_name: chr [1:247540] "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Hal
sted St & North Branch St" "Halsted St & North Branch St" ...
## $ start_station_id  : chr [1:247540] "13307" "KP1705001026" "KA1504000117" "KA1504000117"
...
## $ end_station_name  : chr [1:247540] "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Bro
adway & Barry Ave" "LaSalle Dr & Huron St" ...
## $ end_station_id    : chr [1:247540] "TA1307000130" "TA1307000119" "13137" "KP1705001026"
...
## $ start_lat         : num [1:247540] 41.9 41.9 41.9 41.9 41.9 ...
## $ start_lng         : num [1:247540] -87.7 -87.6 -87.6 -87.6 -87.7 ...
## $ end_lat           : num [1:247540] 41.9 42 41.9 41.9 41.9 ...
## $ end_lng           : num [1:247540] -87.7 -87.7 -87.6 -87.6 -87.6 ...
## $ member_casual     : chr [1:247540] "member" "casual" "member" "member" ...
## - attr(*, "spec")=
## .. cols(
## ..   ride_id = col_character(),
## ..   rideable_type = col_character(),
## ..   started_at = col_datetime(format = ""),
## ..   ended_at = col_datetime(format = ""),
## ..   start_station_name = col_character(),
## ..   start_station_id = col_character(),
## ..   end_station_name = col_character(),
## ..   end_station_id = col_character(),
## ..   start_lat = col_double(),
## ..   start_lng = col_double(),
## ..   end_lat = col_double(),
## ..   end_lng = col_double(),
## ..   member_casual = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

Combine all the datasets into one single dataframe to consolidate analysis

```
trips21fill<- rbind(trip21_Jan, trip21_Feb, trip21_Mar, trip21_Apr, trip21_May, trip21_Jun, trip
21_Jul, trip21_Aug, trip21_Sep, trip21_Oct, trip21_Nov, trip21_Dec)
```

View newly created dataset

```
View(trips21fill)
```

All looks good!

Remove columns not required or beyond the scope of project

```
trips21fill <- trips21fill %>%
  select(-c(start_lat:end_lng))
glimpse(trips21fill)
```

```
## Rows: 5,595,063
## Columns: 9
## $ ride_id          <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55...
## $ rideable_type    <chr> "electric_bike", "electric_bike",...
## $ started_at       <dtm> 2021-01-23 16:14:19, 2021-01-27 ...
## $ ended_at         <dtm> 2021-01-23 16:24:44, 2021-01-27 ...
## $ start_station_name <chr> "California Ave & Cortez St", "Ca...
## $ start_station_id  <chr> "17660", "17660", "17660", "17660...
## $ end_station_name  <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ end_station_id    <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ member_casual     <chr> "member", "member", "member", "me...
```

Rename columns for better readability

```
trips21fill <- trips21fill %>%
  rename(ride_type = rideable_type,
         start_time = started_at,
         end_time = ended_at,
         customer_type = member_casual)
glimpse(trips21fill)
```

```
## Rows: 5,595,063
## Columns: 9
## $ ride_id          <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55...
## $ ride_type        <chr> "electric_bike", "electric_bike",...
## $ start_time       <dtm> 2021-01-23 16:14:19, 2021-01-27 ...
## $ end_time         <dtm> 2021-01-23 16:24:44, 2021-01-27 ...
## $ start_station_name <chr> "California Ave & Cortez St", "Ca...
## $ start_station_id  <chr> "17660", "17660", "17660", "17660...
## $ end_station_name  <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ end_station_id    <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ customer_type     <chr> "member", "member", "member", "me...
```

Add new columns that can be used for aggregate functions

```

#column for day of the week the trip started
trips21fill$day_of_the_week <- format(as.Date(trips21fill$start_time), '%a')

#column for month when the trip started
trips21fill$month <- format(as.Date(trips21fill$start_time), '%b_%y')

#column for time of the day when the trip started
#Time element needs to be extracted from start_time. However, as the times must be in POSIXct
#(only times of class POSIXct are supported in ggplot2), a two-step conversion is needed.
#First the time is converted to a character vector, effectively stripping all the date information.
#The time is then converted back to POSIXct with today's date - the date is of no interest to us,
#only the hours-minutes-seconds are.
trips21fill$time <- format(trips21fill$start_time, format = "%H:%M")
trips21fill$time <- as.POSIXct(trips21fill$time, format = "%H:%M")

#column for trip duration in min
trips21fill$trip_duration <- (as.double(difftime(trips21fill$end_time, trips21fill$start_time)))/60

# check the dataframe
glimpse(trips21fill)

```

```

## Rows: 5,595,063
## Columns: 13
## $ ride_id          <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55...
## $ ride_type        <chr> "electric_bike", "electric_bike",...
## $ start_time       <dtm> 2021-01-23 16:14:19, 2021-01-27 ...
## $ end_time         <dtm> 2021-01-23 16:24:44, 2021-01-27 ...
## $ start_station_name <chr> "California Ave & Cortez St", "Ca...
## $ start_station_id  <chr> "17660", "17660", "17660", "17660...
## $ end_station_name  <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ end_station_id    <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ customer_type     <chr> "member", "member", "member", "me...
## $ day_of_the_week   <chr> "Sat", "Wed", "Thu", "Thu", "Sat"...
## $ month             <chr> "Jan_21", "Jan_21", "Jan_21", "Ja...
## $ time              <dtm> 2022-10-27 16:14:00, 2022-10-27 ...
## $ trip_duration     <dbl> 10.4166667, 4.0666667, 1.3333333,...

```

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

```

# checking for trip lengths less than 0
nrow(subset(trips21fill, trip_duration < 0))

```

```
## [1] 147
```



```
#checking for testrides that were made by company for quality checks
nrow(subset(trips21fill, start_station_name %like% "TEST"))
```

```
## [1] 0
```

```
nrow(subset(trips21fill, start_station_name %like% "test"))
```

```
## [1] 0
```

```
nrow(subset(trips21fill, start_station_name %like% "Test"))
```

```
## [1] 0
```

As there are 147 rows with trip_dration less than 0 mins. we will remove these observations from our dataframe. We will create a new dataframe deviod of these obseravtions without making any changes to the existing dataframe.

```
# remove negative trip durations
trips21fill_v2 <- trips21fill[!(trips21fill$trip_duration < 0),]

#check dataframe
glimpse(trips21fill_v2)
```

```
## Rows: 5,594,916
## Columns: 13
## $ ride_id          <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55...
## $ ride_type        <chr> "electric_bike", "electric_bike",...
## $ start_time        <dtm> 2021-01-23 16:14:19, 2021-01-27 ...
## $ end_time          <dtm> 2021-01-23 16:24:44, 2021-01-27 ...
## $ start_station_name <chr> "California Ave & Cortez St", "Ca...
## $ start_station_id  <chr> "17660", "17660", "17660", "17660...
## $ end_station_name  <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ end_station_id    <chr> NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ customer_type     <chr> "member", "member", "member", "me...
## $ day_of_the_week    <chr> "Sat", "Wed", "Thu", "Thu", "Sat"...
## $ month              <chr> "Jan_21", "Jan_21", "Jan_21", "Ja...
## $ time              <dtm> 2022-10-27 16:14:00, 2022-10-27 ...
## $ trip_duration     <dbl> 10.4166667, 4.0666667, 1.3333333,...
```

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

```
# checking count of distinct values
table(trips21fill_v2$customer_type)
```

```
##
## casual member
## 2528946 3065970
```

```
#aggregating total trip duration by customer type
setNames(aggregate(trip_duration ~ customer_type, trips21fill_v2, sum), c("customer_type", "total_trip_duration(mins)"))
```

```
## customer_type total_trip_duration(mins)
## 1 casual 80931864
## 2 member 41800052
```

4&5. Analyze and Share the Data

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

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

```
# statistical summary of trip_duration for all trips
summary(trips21fill_v2$trip_duration)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
##      0.00     6.75    12.00    21.94    21.78 55944.15
```

```
#statistical summary of trip_duration by customer_type
trips21fill_v2 %>%
  group_by(customer_type) %>%
  summarise(min_trip_duration = min(trip_duration), max_trip_duration = max(trip_duration),
            median_trip_duration = median(trip_duration), mean_trip_duration = mean(trip_duration))
```

```
## # A tibble: 2 × 5
##   customer_type min_trip_duration max_trip_d...1 media...2 mean...3
##   <chr>          <dbl>          <dbl>    <dbl>    <dbl>
## 1 casual          0          55944.    16.0    32.0
## 2 member          0          1560.     9.6    13.6
## # ... with abbreviated variable names 1max_trip_duration,
## # 2median_trip_duration, 3mean_trip_duration
```

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

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

```
# fix the order for the day_of_the_week and month variable so that they show up
# in the same sequence in output tables and visualizations
trips21fill_v2$day_of_the_week <- ordered(trips21fill_v2$day_of_the_week, levels=c("Mon", "Tue",
"Wed", "Thu", "Fri", "Sat", "Sun"))
trips21fill_v2$month <- ordered(trips21fill_v2$month, levels=c("Jan_21", "Feb_21", "Mar_21", "Apr_21",
"May_21", "Jun_21", "Jul_21", "Aug_21", "Sep_21", "Oct_21", "Nov_21", "Dec_21" ))
trips21fill_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(number_of_rides = n(), average_duration_mins = mean(trip_duration)) %>%
  arrange(customer_type, desc(number_of_rides))
```

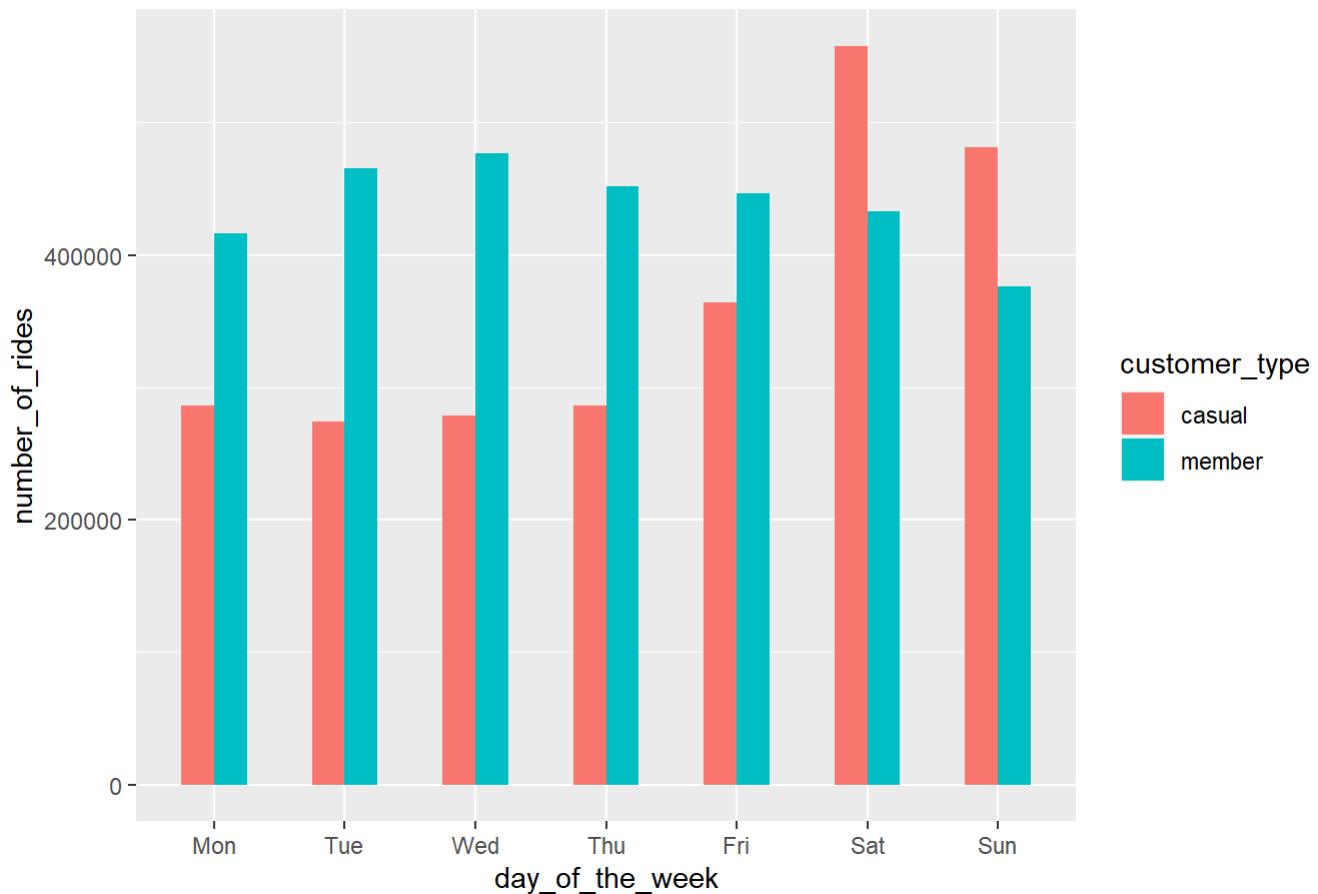
```
## `summarise()` has grouped output by 'customer_type'. You can
## override using the `.groups` argument.
```

```
## # A tibble: 14 × 4
## # Groups:   customer_type [2]
##   customer_type day_of_the_week number_of_rides average_dur...1
##   <chr>         <ord>                <int>         <dbl>
## 1 casual       Sat                557994         34.7
## 2 casual       Sun                481104         37.6
## 3 casual       Fri                364075         30.3
## 4 casual       Mon                286373         31.9
## 5 casual       Thu                286064         27.7
## 6 casual       Wed                278948         27.7
## 7 casual       Tue                274388         28.0
## 8 member       Wed                477156         12.8
## 9 member       Tue                465509         12.8
## 10 member      Thu                451520         12.8
## 11 member      Fri                446423         13.3
## 12 member      Sat                433041         15.3
## 13 member      Mon                416204         13.2
## 14 member      Sun                376117         15.7
## # ... with abbreviated variable name 1average_duration_mins
```

```
trips21fill_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(number_of_rides = n()) %>%
  arrange(customer_type, day_of_the_week) %>%
  ggplot(aes(x = day_of_the_week, y = number_of_rides, fill = customer_type)) +
  labs(title = "Total trips by customer type Vs. Day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```

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

Total trips by customer type Vs. Day of the week



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

Average number of trips by customer type and month

```
unique(trips21fill$month)
```

```
## [1] "Jan_21" "Feb_21" "Mar_21" "Apr_21" "May_21" "Jun_21"
## [7] "Jul_21" "Aug_21" "Sep_21" "Oct_21" "Nov_21" "Dec_21"
```

```
trips21fill_v2 %>%
  group_by(customer_type, month) %>%
  summarise(number_of_rides = n(), `average_duration_(mins)` = mean(trip_duration)) %>%
  arrange(customer_type, desc(number_of_rides))
```

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

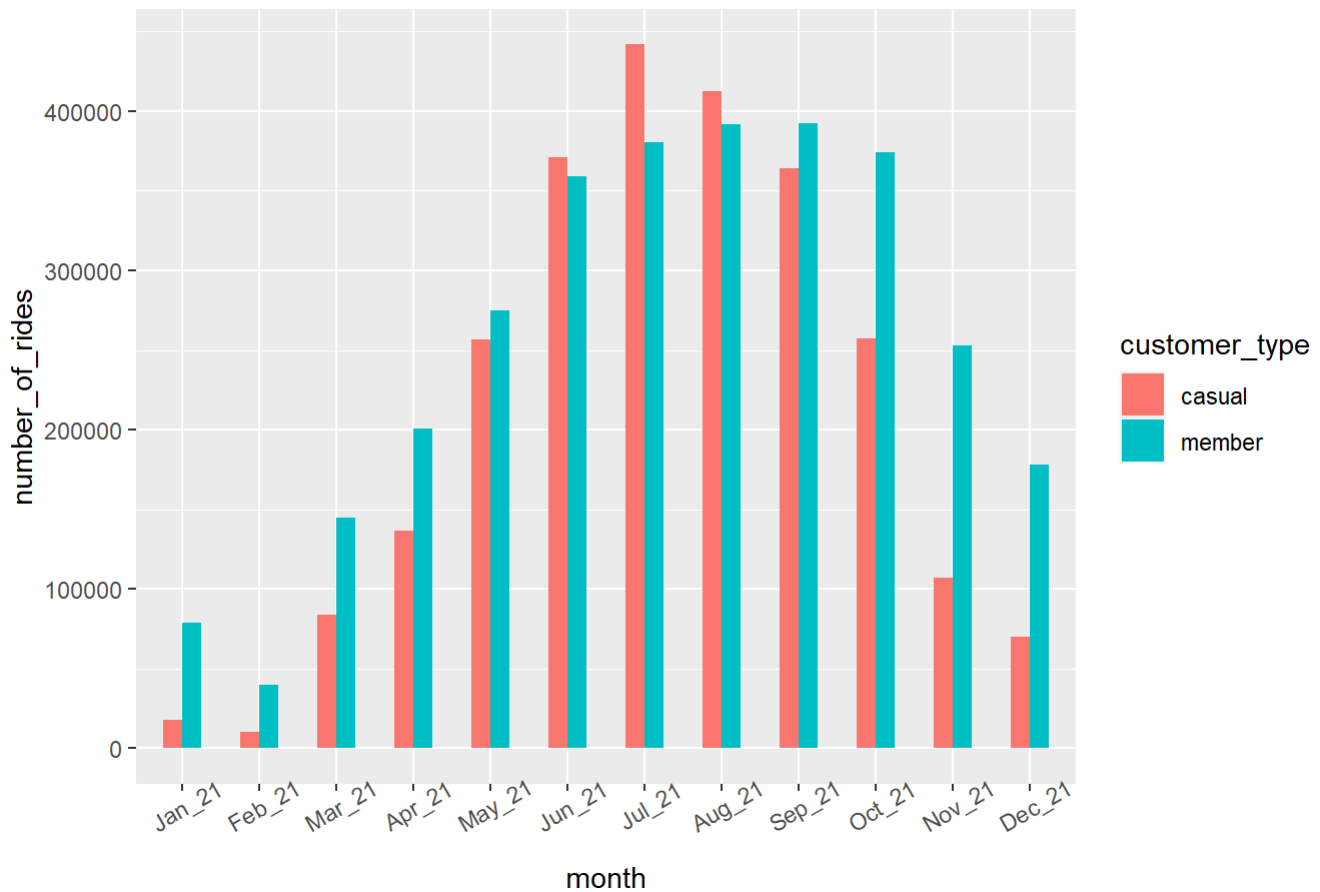
```
## # A tibble: 24 × 4
## # Groups:   customer_type [2]
##   customer_type month   number_of_rides average_duration_(mi...1
##   <chr>         <ord>         <int>             <dbl>
## 1 casual      Jul_21           442048             32.8
## 2 casual      Aug_21           412662             28.8
## 3 casual      Jun_21           370678             37.1
## 4 casual      Sep_21           363883             27.8
## 5 casual      Oct_21           257242             28.7
## 6 casual      May_21           256916             38.2
## 7 casual      Apr_21           136601             38.0
## 8 casual      Nov_21           106898             23.1
## 9 casual      Mar_21            84032             38.2
## 10 casual     Dec_21            69738             23.5
## # ... with 14 more rows, and abbreviated variable name
## #   1`average_duration_(mins)`
```

Visualization:

```
trips21fill_v2 %>%
  group_by(customer_type, month) %>%
  summarise(number_of_rides = n()) %>%
  arrange(customer_type, month) %>%
  ggplot(aes(x = month, y = number_of_rides, fill = customer_type)) +
  labs(title = "Total trips by customer type Vs. Month") +
  theme(axis.text.x = element_text(angle = 30)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```

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

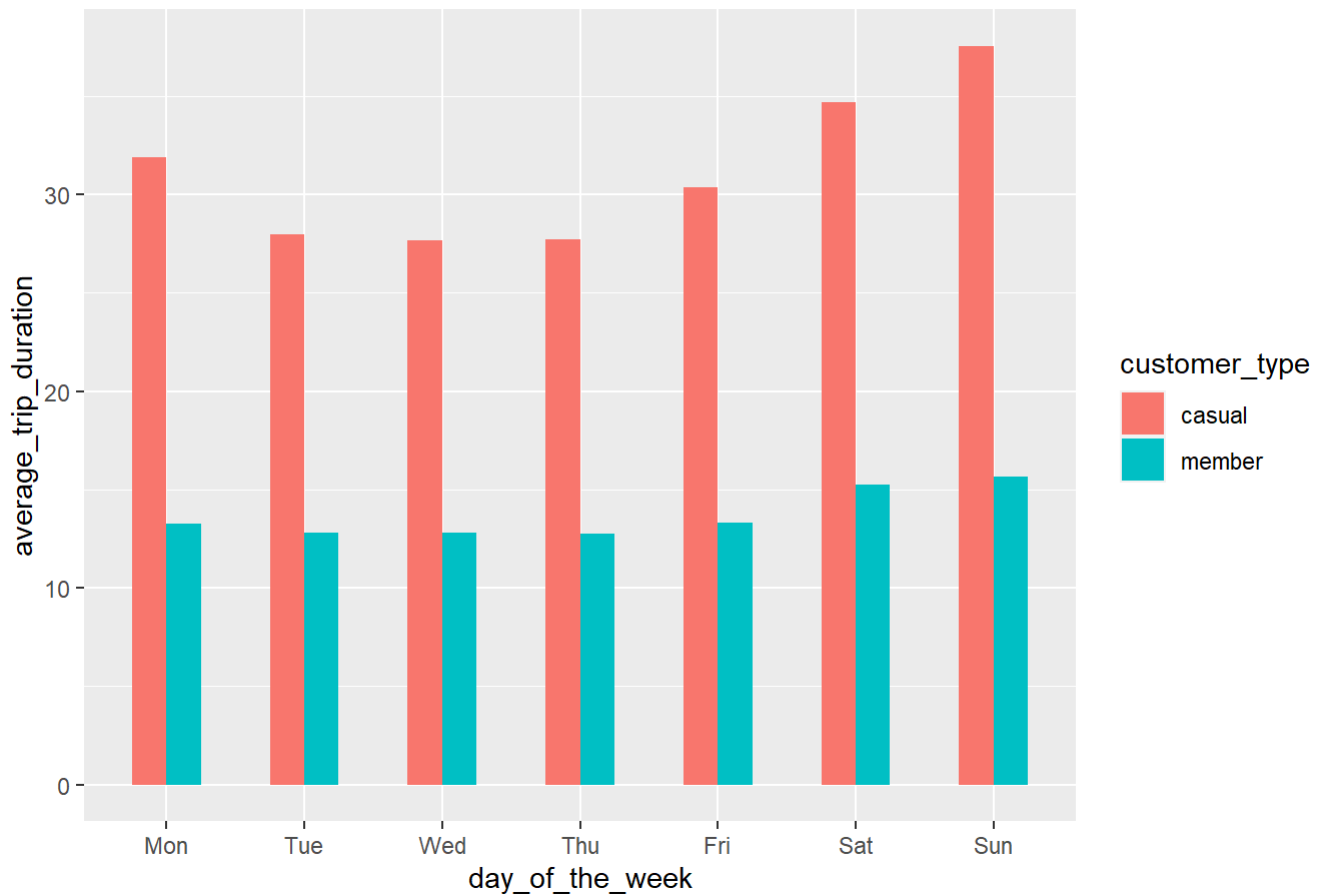
Total trips by customer type Vs. Month



```
trips21fill_v2 %>%
  group_by(customer_type, day_of_the_week) %>%
  summarise(average_trip_duration = mean(trip_duration)) %>%
  ggplot(aes(x = day_of_the_week, y = average_trip_duration, fill = customer_type)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title = "Average trip duration by customer type Vs. Day of the week")
```

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

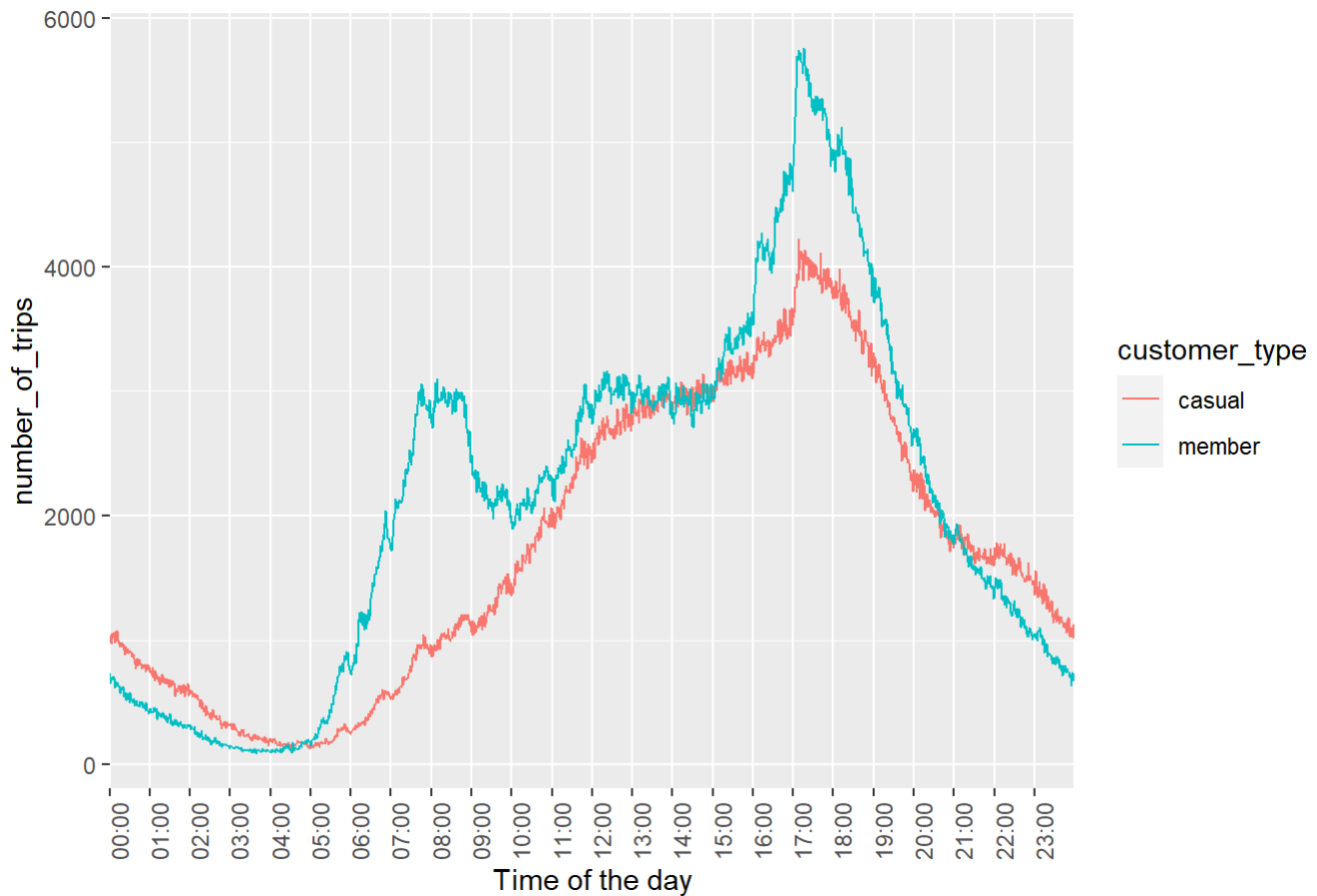
Average trip duration by customer type Vs. Day of the week



```
trips21fill_v2 %>%
  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() +
  scale_x_datetime(date_breaks = "1 hour", minor_breaks = NULL,
                   date_labels = "%H:%M", expand = c(0,0)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Demand over 24 hours of a day", x = "Time of the day")
```

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

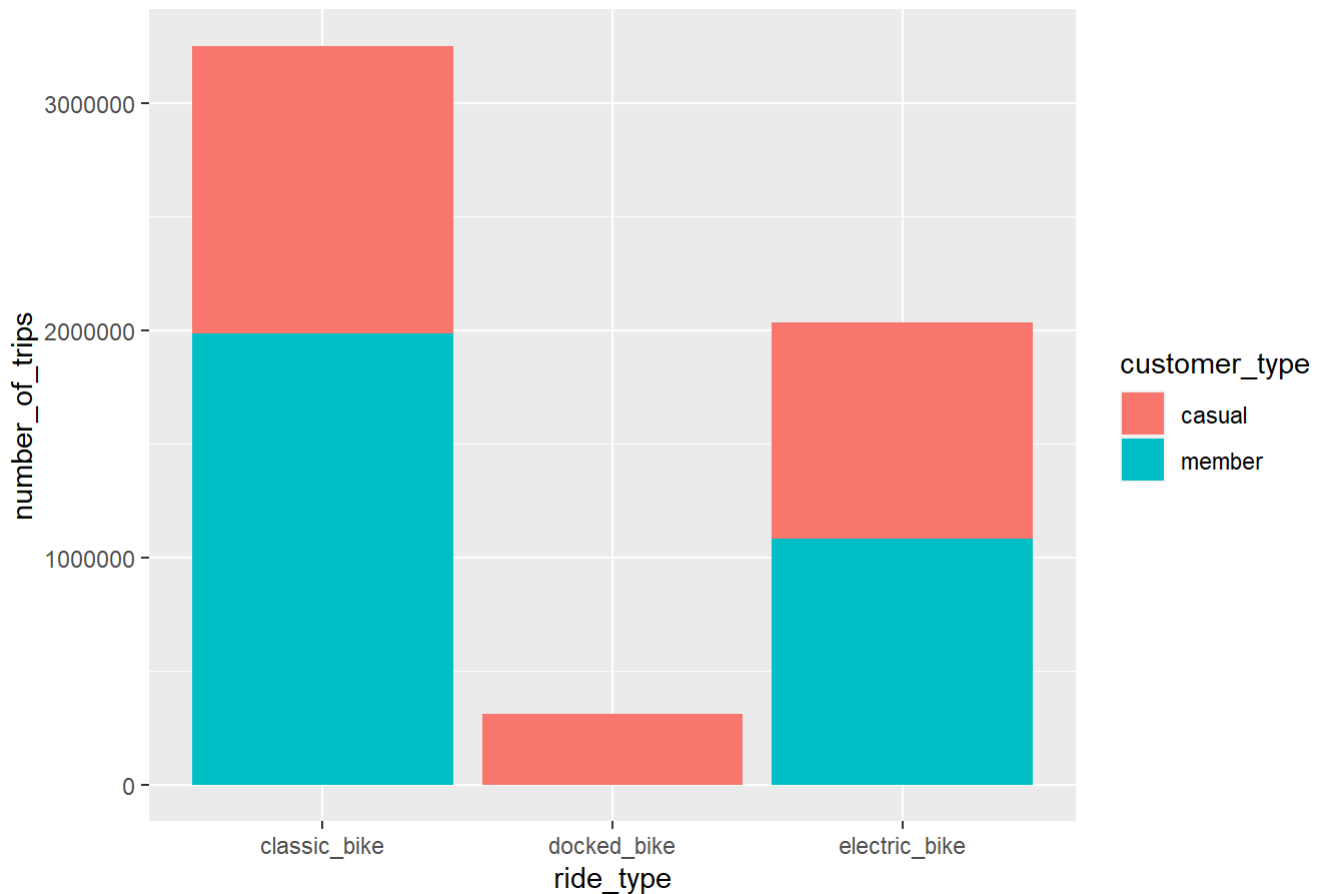
Demand over 24 hours of a day



```
trips21fill_v2 %>%
  group_by(ride_type, customer_type) %>%
  summarise(number_of_trips = n()) %>%
  ggplot(aes(x= ride_type, y=number_of_trips, fill= customer_type))+
    geom_bar(stat='identity') +
    scale_y_continuous(labels = function(x) format(x, scientific = FALSE)) +
    labs(title = "Ride type Vs. Number of trips")
```

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


Ride type Vs. Number of trips



Classic bikes are predominantly used by members. Classic bikes are in most demand and equally used by both members as well as casual riders. Electric bikes are more favored by members and casual, but Docked bikes are the less used biked from member, some Casual riders are using Docked bike.

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

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

```
clean_data <- aggregate(trips21fill_v2$trip_duration ~ trips21fill_v2$customer_type + trips21fill_v2$day_of_the_week, FUN = mean)
write.csv(clean_data, "Clean Data.csv", row.names = F)
```

6. Act

The average ride time shows a stark difference between the casuals and members. Casuals overall spend more time using the service than their full time member counterparts.

what does the data tell us?

key takeaways

- Casual users tended to ride more so in the warmer months of Chicago, namely June- August. Their participation exceeded that of the long term members.
- To further that the Casual demographic spent on average a lot longer time per ride than their long-term counter-parts.
- The days of the week also further shows that casual riders prefer to use the service during the weekends as their usage peaked then. The long term members conversly utilised the service more-so throughout the typical work week i.e (Monday- friday)
- Long term riders tended to stick more so to classic bikes as opposed to the docked or electric bikes.

Recommendations

- This report recommends the following: *

Introducing plans that may be more appealing to casuals for the summer months. This marketing should be done during the winter months in preperation. The casual users might be more interested in a membership option that allows for per-use balance card. Alternatively, the existing payment structure may be altered in order to make single-use more costly to the casual riders as well as lowering the long-term membership rate. Membership rates specifically for the warmer months as well as for those who only ride on the weekends would assist in targeting the casual riders more specifically

Things to Consider

Additional points that were not examined

The report understands the scope of this analysis is extremely limited and because of that fact, additional data, as well as data points may have been able to contribute to this report offering an even more granular analysis. The following are data points that could have enhanced the report:

- Age and gender: This would add a dynamic to whether or not customers are being targeted across demographic lines. Is the existing marketing effective? Is there potential for more inclusive targeting?
- Pricing structure: The actual pricing plans data was not provided and would give further insight to which plans are the most popular and by (how much) when comparing them. It would also be effective to understanding the spending behaviour of casual user.
- Household income data: Pinpointing the average income of the long-term members as compared to the casual counter-parts would allow for further analysis of what is the typical economic standing of each type of member, as well as providing the ability to analysis overall price sensitivity between the two different membership types.

Thank you for your time!