Google Study: How Does a Bike-Share Navigate Speedy Success?

Zach Mitchell

01-03-2023

**Introduction**

This analysis is for the capstone project that follows with the completion of the [Google Data Analytics professional Certificate](https://www.coursera.org/professional-certificates/google-data-analytics). This case study uses data provided by a bikeshare company that collects detailed reports about its customers trips over a twelve month period (January 2022-December 2022). The data has been made available by Motivate International Inc. under this [license](https://ride.divvybikes.com/data-license-agreement).

The analysis will adhere to the six main steps of the Data Analysis process: Ask, Prepare, Process, Analyze, and Act (APPAA).

**Summary of the Company**

Cyclistic is a bike-share company that started in 2016 in the Chicago area. The program includes 692 bike stations with a collection of 5,824 bicycles. It presents its customers with a variety of bikes to choose from that includes standard two-wheel bikes, reclining bikes, hand tricycles, and cargo bikes. The program includes multiple ride passes where riders can choose between single-ride passes, full-day passes, and annual memberships.

**Description of the Data**

Listed below is the column titles followed by a description of the data:

* **ride\_id**: Unique identifiable number assigned to the bike ride trip.
* **rideable\_type**: Type of bike used during the trip; standard two-wheel bike, reclining bike, hand tricycle, and cargo bike.
* **started\_at**: Start date and time for the trip.
* **ended\_at**: End date and time for the trip.
* **start\_station\_name**: Name of the station at the start of the trip.
* **start\_station\_id**: Unique identifiable code that corresponds with the start station.
* **end\_station\_name**: Name of the station at the end of the trip.
* **end\_station\_id**: Unique identifiable code that corresponds with the end station.
* **start\_lat**: Latitude coordinates at the start of the trip.
* **start\_lng**: Longitude coordinates at the start of the trip.
* **end\_lat**: Latitude coordinates at the end of the trip.
* **end\_lng**: Longitude coordinates at the end of the trip.
* **member\_casual**: Type of customer during the trip; (“member” = annual member, “casual” = Casual rider).

**1. Ask**

How do annual members and casual riders use Cyclistic bikes differently?

**Business Task**

The director of marketing at Cyclistic wants the marketing analytics team to create a strategy for converting casual riders into annual members. The finance analysts have concluded that by converting casual riders into annual members it will be much more profitable for the company. The goal for this business task is to find out how annual members and casual riders use the Cyclistic bikes differently. This will give the director of marking and the executive team a better understanding on how to convert casual riders into annual members to help the future success of Cyclistic.

**Stakeholders**

* Director of marketing
* Cyclistic executive team

**2. Prepare**

The [data](https://divvy-tripdata.s3.amazonaws.com/index.html) for the analysis is provided by Motivate International, inc. (A link to the license is included in the introduction labeled as “license”). There are twelve datasets used in this analysis each covering the time period of one month to equal a full year. The time period is January 2022 to December 2022. Each of the datasets includes thirteen columns of data all listed in the “Description of the Data” section. The data does not include personalized information about the customers to help safeguard the privacy of the riders. This means that credit card information is not included in the data, so the data will not be able to determine where casual riders reside and if they have purchased multiple passes.

Since the datasets are so large R was used in place of excel/google sheets to speed up cleaning and preparation of the data. R was also used for analysis and visualizations to keep all the data on the same platform.

**R Libraries**

── **Attaching packages** ─────────────────── tidyverse 1.3.2 ──

✔ ggplot2 3.4.0 ✔ purrr 1.0.0

✔ tibble 3.1.8 ✔ dplyr 1.0.10

✔ tidyr 1.2.1 ✔ stringr 1.5.0

✔ readr 2.1.3 ✔ forcats 0.5.2

── **Conflicts** ────────────────────── tidyverse\_conflicts() ──

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

Attaching package: ‘lubridate’

The following objects are masked from ‘package:base’:

date, intersect, setdiff, union

Attaching package: ‘janitor’

The following objects are masked from ‘package:stats’:

chisq.test, fisher.test

Attaching package: ‘data.table’

The following objects are masked from ‘package:lubridate’:

hour, isoweek, mday, minute, month, quarter, second, wday, week,

yday, year

**Loading Datasets**

tripdata\_202112 <- read\_csv("Desktop/bikeshare\_data/original\_data/202112-divvy-tripdata.csv")

tripdata\_202201 <- read\_csv("Desktop/bikeshare\_data/original\_data/202201-divvy-tripdata.csv")

tripdata\_202202 <- read\_csv("Desktop/bikeshare\_data/original\_data/202202-divvy-tripdata.csv")

tripdata\_202203 <- read\_csv("Desktop/bikeshare\_data/original\_data/202203-divvy-tripdata.csv")

tripdata\_202204 <- read\_csv("Desktop/bikeshare\_data/original\_data/202204-divvy-tripdata.csv")

tripdata\_202205 <- read\_csv("Desktop/bikeshare\_data/original\_data/202205-divvy-tripdata.csv")

tripdata\_202206 <- read\_csv("Desktop/bikeshare\_data/original\_data/202206-divvy-tripdata.csv")

tripdata\_202207 <- read\_csv("Desktop/bikeshare\_data/original\_data/202207-divvy-tripdata.csv")

tripdata\_202208 <- read\_csv("Desktop/bikeshare\_data/original\_data/202208-divvy-tripdata.csv")

tripdata\_202209 <- read\_csv("Desktop/bikeshare\_data/original\_data/202209-divvy-tripdata.csv")

tripdata\_202210 <- read\_csv("Desktop/bikeshare\_data/original\_data/202210-divvy-tripdata.csv")

tripdata\_202211 <- read\_csv("Desktop/bikeshare\_data/original\_data/202211-divvy-tripdata.csv")

**Column Name Check**

colnames(tripdata\_202112)

colnames(tripdata\_202201)

colnames(tripdata\_202202)

colnames(tripdata\_202203)

colnames(tripdata\_202204)

colnames(tripdata\_202205)

colnames(tripdata\_202206)

colnames(tripdata\_202207)

colnames(tripdata\_202208)

colnames(tripdata\_202209)

colnames(tripdata\_202210)

colnames(tripdata\_202211)

> colnames(tripdata\_202112)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202201)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202202)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202203)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202204)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202205)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202206)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202207)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202208)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202209)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202210)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

> colnames(tripdata\_202211)

[1] "ride\_id" "rideable\_type" "started\_at"

[4] "ended\_at" "start\_station\_name" "start\_station\_id"

[7] "end\_station\_name" "end\_station\_id" "start\_lat"

[10] "start\_lng" "end\_lat" "end\_lng"

[13] "member\_casual"

**Data Structure Check**

str(tripdata\_202112)

str(tripdata\_202201)

str(tripdata\_202202)

str(tripdata\_202203)

str(tripdata\_202204)

str(tripdata\_202205)

str(tripdata\_202206)

str(tripdata\_202207)

str(tripdata\_202208)

str(tripdata\_202209)

str(tripdata\_202210)

str(tripdata\_202211)

> str(tripdata\_202112)

spc\_tbl\_ [247,540 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:247540] "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278BA87BF698339" ...

$ rideable\_type : chr [1:247540] "electric\_bike" "electric\_bike" "electric\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:247540], format: "2021-12-07 15:06:07" "2021-12-11 03:43:29" ...

$ ended\_at : POSIXct[1:247540], format: "2021-12-07 15:13:42" "2021-12-11 04:10:23" ...

$ start\_station\_name: chr [1:247540] "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted 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" "Broadway & 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>

> str(tripdata\_202201)

spc\_tbl\_ [103,770 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:103770] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB80ED419105406" ...

$ rideable\_type : chr [1:103770] "electric\_bike" "electric\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:103770], format: "2022-01-13 11:59:47" "2022-01-10 08:41:56" ...

$ ended\_at : POSIXct[1:103770], format: "2022-01-13 12:02:44" "2022-01-10 08:46:17" ...

$ start\_station\_name: chr [1:103770] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffield Ave & Fullerton Ave" "Clark St & Bryn Mawr Ave" ...

$ start\_station\_id : chr [1:103770] "525" "525" "TA1306000016" "KA1504000151" ...

$ end\_station\_name : chr [1:103770] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave & Fullerton Ave" "Paulina St & Montrose Ave" ...

$ end\_station\_id : chr [1:103770] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...

$ start\_lat : num [1:103770] 42 42 41.9 42 41.9 ...

$ start\_lng : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...

$ end\_lat : num [1:103770] 42 42 41.9 42 41.9 ...

$ end\_lng : num [1:103770] -87.7 -87.7 -87.7 -87.7 -87.6 ...

$ member\_casual : chr [1:103770] "casual" "casual" "member" "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(tripdata\_202202)

spc\_tbl\_ [115,609 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:115609] "E1E065E7ED285C02" "1602DCDC5B30FFE3" "BE7DD2AF4B55C4AF" "A1789BDF844412BE" ...

$ rideable\_type : chr [1:115609] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:115609], format: "2022-02-19 18:08:41" "2022-02-20 17:41:30" ...

$ ended\_at : POSIXct[1:115609], format: "2022-02-19 18:23:56" "2022-02-20 17:45:56" ...

$ start\_station\_name: chr [1:115609] "State St & Randolph St" "Halsted St & Wrightwood Ave" "State St & Randolph St" "Southport Ave & Waveland Ave" ...

$ start\_station\_id : chr [1:115609] "TA1305000029" "TA1309000061" "TA1305000029" "13235" ...

$ end\_station\_name : chr [1:115609] "Clark St & Lincoln Ave" "Southport Ave & Wrightwood Ave" "Canal St & Adams St" "Broadway & Sheridan Rd" ...

$ end\_station\_id : chr [1:115609] "13179" "TA1307000113" "13011" "13323" ...

$ start\_lat : num [1:115609] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:115609] -87.6 -87.6 -87.6 -87.7 -87.6 ...

$ end\_lat : num [1:115609] 41.9 41.9 41.9 42 41.9 ...

$ end\_lng : num [1:115609] -87.6 -87.7 -87.6 -87.6 -87.6 ...

$ member\_casual : chr [1:115609] "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(tripdata\_202203)

spc\_tbl\_ [284,042 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:284042] "47EC0A7F82E65D52" "8494861979B0F477" "EFE527AF80B66109" "9F446FD9DEE3F389" ...

$ rideable\_type : chr [1:284042] "classic\_bike" "electric\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:284042], format: "2022-03-21 13:45:01" "2022-03-16 09:37:16" ...

$ ended\_at : POSIXct[1:284042], format: "2022-03-21 13:51:18" "2022-03-16 09:43:34" ...

$ start\_station\_name: chr [1:284042] "Wabash Ave & Wacker Pl" "Michigan Ave & Oak St" "Broadway & Berwyn Ave" "Wabash Ave & Wacker Pl" ...

$ start\_station\_id : chr [1:284042] "TA1307000131" "13042" "13109" "TA1307000131" ...

$ end\_station\_name : chr [1:284042] "Kingsbury St & Kinzie St" "Orleans St & Chestnut St (NEXT Apts)" "Broadway & Ridge Ave" "Franklin St & Jackson Blvd" ...

$ end\_station\_id : chr [1:284042] "KA1503000043" "620" "15578" "TA1305000025" ...

$ start\_lat : num [1:284042] 41.9 41.9 42 41.9 41.9 ...

$ start\_lng : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.6 ...

$ end\_lat : num [1:284042] 41.9 41.9 42 41.9 41.9 ...

$ end\_lng : num [1:284042] -87.6 -87.6 -87.7 -87.6 -87.7 ...

$ member\_casual : chr [1:284042] "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(tripdata\_202204)

spc\_tbl\_ [371,249 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:371249] "3564070EEFD12711" "0B820C7FCF22F489" "89EEEE32293F07FF" "84D4751AEB31888D" ...

$ rideable\_type : chr [1:371249] "electric\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:371249], format: "2022-04-06 17:42:48" "2022-04-24 19:23:07" ...

$ ended\_at : POSIXct[1:371249], format: "2022-04-06 17:54:36" "2022-04-24 19:43:17" ...

$ start\_station\_name: chr [1:371249] "Paulina St & Howard St" "Wentworth Ave & Cermak Rd" "Halsted St & Polk St" "Wentworth Ave & Cermak Rd" ...

$ start\_station\_id : chr [1:371249] "515" "13075" "TA1307000121" "13075" ...

$ end\_station\_name : chr [1:371249] "University Library (NU)" "Green St & Madison St" "Green St & Madison St" "Delano Ct & Roosevelt Rd" ...

$ end\_station\_id : chr [1:371249] "605" "TA1307000120" "TA1307000120" "KA1706005007" ...

$ start\_lat : num [1:371249] 42 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...

$ end\_lat : num [1:371249] 42.1 41.9 41.9 41.9 41.9 ...

$ end\_lng : num [1:371249] -87.7 -87.6 -87.6 -87.6 -87.6 ...

$ member\_casual : chr [1:371249] "member" "member" "member" "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(),

.. end5\_lat = col\_double(),

.. end\_lng = col\_double(),

.. member\_casual = col\_character()

.. )

- attr(\*, "problems")=<externalptr>

> str(tripdata\_202205)

spc\_tbl\_ [634,858 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415CF" "6FF59852924528F8" ...

$ rideable\_type : chr [1:634858] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28" ...

$ ended\_at : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22" ...

$ start\_station\_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monroe St" "Clinton St & Madison St" "Clinton St & Madison St" ...

$ start\_station\_id : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032" ...

$ end\_station\_name : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St" "Wood St & Milwaukee Ave" "Clark St & Randolph St" ...

$ end\_station\_id : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...

$ start\_lat : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...

$ end\_lat : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...

$ end\_lng : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...

$ member\_casual : chr [1:634858] "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(tripdata\_202206)

spc\_tbl\_ [769,204 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:769204] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD771D2" "C9C320375DE1D5C6" ...

$ rideable\_type : chr [1:769204] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...

$ started\_at : POSIXct[1:769204], format: "2022-06-30 17:27:53" "2022-06-30 18:39:52" ...

$ ended\_at : POSIXct[1:769204], format: "2022-06-30 17:35:15" "2022-06-30 18:47:28" ...

$ start\_station\_name: chr [1:769204] NA NA NA NA ...

$ start\_station\_id : chr [1:769204] NA NA NA NA ...

$ end\_station\_name : chr [1:769204] NA NA NA NA ...

$ end\_station\_id : chr [1:769204] NA NA NA NA ...

$ start\_lat : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...

$ start\_lng : num [1:769204] -87.6 -87.6 -87.7 -87.7 -87.6 ...

$ end\_lat : num [1:769204] 41.9 41.9 41.9 41.8 41.9 ...

$ end\_lng : num [1:769204] -87.6 -87.6 -87.6 -87.7 -87.6 ...

$ member\_casual : chr [1:769204] "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(tripdata\_202207)

spc\_tbl\_ [823,488 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:823488] "954144C2F67B1932" "292E027607D218B6" "57765852588AD6E0" "B5B6BE44314590E6" ...

$ rideable\_type : chr [1:823488] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:823488], format: "2022-07-05 08:12:47" "2022-07-26 12:53:38" ...

$ ended\_at : POSIXct[1:823488], format: "2022-07-05 08:24:32" "2022-07-26 12:55:31" ...

$ start\_station\_name: chr [1:823488] "Ashland Ave & Blackhawk St" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)" "Buckingham Fountain (Temp)" ...

$ start\_station\_id : chr [1:823488] "13224" "15541" "15541" "15541" ...

$ end\_station\_name : chr [1:823488] "Kingsbury St & Kinzie St" "Michigan Ave & 8th St" "Michigan Ave & 8th St" "Woodlawn Ave & 55th St" ...

$ end\_station\_id : chr [1:823488] "KA1503000043" "623" "623" "TA1307000164" ...

$ start\_lat : num [1:823488] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:823488] -87.7 -87.6 -87.6 -87.6 -87.6 ...

$ end\_lat : num [1:823488] 41.9 41.9 41.9 41.8 41.9 ...

$ end\_lng : num [1:823488] -87.6 -87.6 -87.6 -87.6 -87.7 ...

$ member\_casual : chr [1:823488] "member" "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(tripdata\_202208)

spc\_tbl\_ [785,932 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:785932] "550CF7EFEAE0C618" "DAD198F405F9C5F5" "E6F2BC47B65CB7FD" "F597830181C2E13C" ...

$ rideable\_type : chr [1:785932] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...

$ started\_at : POSIXct[1:785932], format: "2022-08-07 21:34:15" "2022-08-08 14:39:21" ...

$ ended\_at : POSIXct[1:785932], format: "2022-08-07 21:41:46" "2022-08-08 14:53:23" ...

$ start\_station\_name: chr [1:785932] NA NA NA NA ...

$ start\_station\_id : chr [1:785932] NA NA NA NA ...

$ end\_station\_name : chr [1:785932] NA NA NA NA ...

$ end\_station\_id : chr [1:785932] NA NA NA NA ...

$ start\_lat : num [1:785932] 41.9 41.9 42 41.9 41.9 ...

$ start\_lng : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...

$ end\_lat : num [1:785932] 41.9 41.9 42 42 41.8 ...

$ end\_lng : num [1:785932] -87.7 -87.6 -87.7 -87.7 -87.7 ...

$ member\_casual : chr [1:785932] "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(tripdata\_202209)

spc\_tbl\_ [701,339 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:701339] "5156990AC19CA285" "E12D4A16BF51C274" "A02B53CD7DB72DD7" "C82E05FEE872DF11" ...

$ rideable\_type : chr [1:701339] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...

$ started\_at : POSIXct[1:701339], format: "2022-09-01 08:36:22" "2022-09-01 17:11:29" ...

$ ended\_at : POSIXct[1:701339], format: "2022-09-01 08:39:05" "2022-09-01 17:14:45" ...

$ start\_station\_name: chr [1:701339] NA NA NA NA ...

$ start\_station\_id : chr [1:701339] NA NA NA NA ...

$ end\_station\_name : chr [1:701339] "California Ave & Milwaukee Ave" NA NA NA ...

$ end\_station\_id : chr [1:701339] "13084" NA NA NA ...

$ start\_lat : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...

$ end\_lat : num [1:701339] 41.9 41.9 41.9 41.9 41.9 ...

$ end\_lng : num [1:701339] -87.7 -87.6 -87.6 -87.7 -87.7 ...

$ member\_casual : chr [1:701339] "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(tripdata\_202210)

spc\_tbl\_ [558,685 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:558685] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E653FDC2D99236" ...

$ rideable\_type : chr [1:558685] "classic\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...

$ started\_at : POSIXct[1:558685], format: "2022-10-14 17:13:30" "2022-10-01 16:29:26" ...

$ ended\_at : POSIXct[1:558685], format: "2022-10-14 17:19:39" "2022-10-01 16:49:06" ...

$ start\_station\_name: chr [1:558685] "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne Ave & Balmoral Ave" "Rush St & Cedar St" ...

$ start\_station\_id : chr [1:558685] "13290" "13288" "655" "KA1504000133" ...

$ end\_station\_name : chr [1:558685] "Larrabee St & Division St" "Damen Ave & Cullerton St" "Western Ave & Leland Ave" "Orleans St & Chestnut St (NEXT Apts)" ...

$ end\_station\_id : chr [1:558685] "KA1504000079" "13089" "TA1307000140" "620" ...

$ start\_lat : num [1:558685] 41.9 41.9 42 41.9 41.9 ...

$ start\_lng : num [1:558685] -87.7 -87.7 -87.7 -87.6 -87.6 ...

$ end\_lat : num [1:558685] 41.9 41.9 42 41.9 41.9 ...

$ end\_lng : num [1:558685] -87.6 -87.7 -87.7 -87.6 -87.6 ...

$ member\_casual : chr [1:558685] "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(tripdata\_202211)

spc\_tbl\_ [337,735 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:337735] "BCC66FC6FAB27CC7" "772AB67E902C180F" "585EAD07FDEC0152" "91C4E7ED3C262FF9" ...

$ rideable\_type : chr [1:337735] "electric\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:337735], format: "2022-11-10 06:21:55" "2022-11-04 07:31:55" ...

$ ended\_at : POSIXct[1:337735], format: "2022-11-10 06:31:27" "2022-11-04 07:46:25" ...

$ start\_station\_name: chr [1:337735] "Canal St & Adams St" "Canal St & Adams St" "Indiana Ave & Roosevelt Rd" "Indiana Ave & Roosevelt Rd" ...

$ start\_station\_id : chr [1:337735] "13011" "13011" "SL-005" "SL-005" ...

$ end\_station\_name : chr [1:337735] "St. Clair St & Erie St" "St. Clair St & Erie St" "St. Clair St & Erie St" "St. Clair St & Erie St" ...

$ end\_station\_id : chr [1:337735] "13016" "13016" "13016" "13016" ...

$ start\_lat : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...

$ end\_lat : num [1:337735] 41.9 41.9 41.9 41.9 41.9 ...

$ end\_lng : num [1:337735] -87.6 -87.6 -87.6 -87.6 -87.6 ...

$ member\_casual : chr [1:337735] "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>

**Combining all Datasets**

> all\_trips <- bind\_rows(tripdata\_202112, tripdata\_202201, tripdata\_202202, tripdata\_202203, tripdata\_202204, tripdata\_202205, tripdata\_202206, tripdata\_202207, tripdata\_202208, tripdata\_202209, tripdata\_202210, tripdata\_202211)

> str(all\_trips)

spc\_tbl\_ [5,733,451 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:5733451] "46F8167220E4431F" "73A77762838B32FD" "4CF42452054F59C5" "3278BA87BF698339" ...

$ rideable\_type : chr [1:5733451] "electric\_bike" "electric\_bike" "electric\_bike" "classic\_bike" ...

$ started\_at : POSIXct[1:5733451], format: "2021-12-07 15:06:07" "2021-12-11 03:43:29" ...

$ ended\_at : POSIXct[1:5733451], format: "2021-12-07 15:13:42" "2021-12-11 04:10:23" ...

$ start\_station\_name: chr [1:5733451] "Laflin St & Cullerton St" "LaSalle Dr & Huron St" "Halsted St & North Branch St" "Halsted St & North Branch St" ...

$ start\_station\_id : chr [1:5733451] "13307" "KP1705001026" "KA1504000117" "KA1504000117" ...

$ end\_station\_name : chr [1:5733451] "Morgan St & Polk St" "Clarendon Ave & Leland Ave" "Broadway & Barry Ave" "LaSalle Dr & Huron St" ...

$ end\_station\_id : chr [1:5733451] "TA1307000130" "TA1307000119" "13137" "KP1705001026" ...

$ start\_lat : num [1:5733451] 41.9 41.9 41.9 41.9 41.9 ...

$ start\_lng : num [1:5733451] -87.7 -87.6 -87.6 -87.6 -87.7 ...

$ end\_lat : num [1:5733451] 41.9 42 41.9 41.9 41.9 ...

$ end\_lng : num [1:5733451] -87.7 -87.7 -87.6 -87.6 -87.6 ...

$ member\_casual : chr [1:5733451] "member" "casual" "member" "member"

**3. Process**

**Cleaning and Formatting the Data**

* Remove columns that are not needed for the analysis. (start\_lat, start\_lng, end\_lat, and end\_lng).

> all\_trips <- all\_trips %>%

+ select(-c(start\_lat:end\_lng))

> glimpse(all\_trips)

Rows: 5,733,451

Columns: 9

$ ride\_id *<chr>* "46F8167220E4431F", "73A77762838B32FD", "4CF42452054F…

$ rideable\_type *<chr>* "electric\_bike", "electric\_bike", "electric\_bike", "c…

$ started\_at *<dttm>* 2021-12-07 15:06:07, 2021-12-11 03:43:29, 2021-12-15…

$ ended\_at *<dttm>* 2021-12-07 15:13:42, 2021-12-11 04:10:23, 2021-12-15…

$ start\_station\_name *<chr>* "Laflin St & Cullerton St", "LaSalle Dr & Huron St", …

$ start\_station\_id *<chr>* "13307", "KP1705001026", "KA1504000117", "KA150400011…

$ end\_station\_name *<chr>* "Morgan St & Polk St", "Clarendon Ave & Leland Ave", …

$ end\_station\_id *<chr>* "TA1307000130", "TA1307000119", "13137", "KP170500102…

$ member\_casual *<chr>* "member", "casual", "member", "member", "member", "me…

**Adding New Columns**

* New columns to be added (days\_of\_the\_week, month, time, ride\_length)

all\_trips$day\_of\_the\_week <- format(as.Date(all\_trips$started\_at), "%a")

all\_trips$month <- format(as.Date(all\_trips$started\_at), "%b\_%y")

all\_trips$time <- format(all\_trips$started\_at, format = "%H:%M")

all\_trips$time <- as.POSIXct(all\_trips$time, format = "%H:%M")

* Time needs to be changed from a character to POSIXct. The time will show today's date, but we are only interested in using the hours, minutes, and seconds.

all\_trips$ride\_length <- (as.double(difftime(all\_trips$ended\_at, all\_trips$started\_at)))/60

* The time will be in minutes.

> glimpse(all\_trips)

Rows: 5,733,451

Columns: 13

$ ride\_id *<chr>* "46F8167220E4431F", "73A77762838B32FD"…

$ rideable\_type *<chr>* "electric\_bike", "electric\_bike", "ele…

$ started\_at *<dttm>* 2021-12-07 15:06:07, 2021-12-11 03:43…

$ ended\_at *<dttm>* 2021-12-07 15:13:42, 2021-12-11 04:10…

$ start\_station\_name *<chr>* "Laflin St & Cullerton St", "LaSalle D…

$ start\_station\_id *<chr>* "13307", "KP1705001026", "KA1504000117…

$ end\_station\_name *<chr>* "Morgan St & Polk St", "Clarendon Ave …

$ end\_station\_id *<chr>* "TA1307000130", "TA1307000119", "13137…

$ member\_casual *<chr>* "member", "casual", "member", "member"…

$ day\_of\_the\_week *<chr>* "Tue", "Sat", "Wed", "Sun", "Thu", "We…

$ month *<chr>* "Dec\_21", "Dec\_21", "Dec\_21", "Dec\_21"…

$ time *<dttm>* 2023-01-07 15:06:00, 2023-01-07 03:43…

$ ride\_length *<dbl>* 7.583333, 26.900000, 12.766667, 14.716…

**Removing Bad Data**

* Check for ride length entries that are < 0. The company also did quality tests on the bikes, these data entries are labeled as a string “test” in the start\_station\_name. We will check for “Test”, “TEST”, and “test” to make sure the label was not typed in differently.

nrow(subset(all\_trips,ride\_length < 0))

* 100 entries were < 0

nrow(subset(all\_trips, start\_station\_name %like% "TEST"))

* 0 “TEST” entries

nrow(subset(all\_trips, start\_station\_name %like% "test"))

* 0 “test” entries

nrow(subset(all\_trips, start\_station\_name %like% "Test"))

* 1 “Test” entry

> all\_trips\_v2 <- all\_trips[!(all\_trips$ride\_length < 0),]

* Removes all 100 entries that were < 0, and renames “all\_trips” to “all\_trips\_v2”.

> all\_trips\_v2 <- all\_trips\_v2[!(all\_trips\_v2$start\_station\_name %like% "Test"),]

* Removes 1 entry labeled “Test”.

> glimpse(all\_trips\_v2)

Rows: 5,733,350

Columns: 13

$ ride\_id *<chr>* "46F8167220E4431F", "73A77762838B32FD"…

$ rideable\_type *<chr>* "electric\_bike", "electric\_bike", "ele…

$ started\_at *<dttm>* 2021-12-07 15:06:07, 2021-12-11 03:43…

$ ended\_at *<dttm>* 2021-12-07 15:13:42, 2021-12-11 04:10…

$ start\_station\_name *<chr>* "Laflin St & Cullerton St", "LaSalle D…

$ start\_station\_id *<chr>* "13307", "KP1705001026", "KA1504000117…

$ end\_station\_name *<chr>* "Morgan St & Polk St", "Clarendon Ave …

$ end\_station\_id *<chr>* "TA1307000130", "TA1307000119", "13137…

$ member\_casual *<chr>* "member", "casual", "member", "member"…

$ day\_of\_the\_week *<chr>* "Tue", "Sat", "Wed", "Sun", "Thu", "We…

$ month *<chr>* "Dec\_21", "Dec\_21", "Dec\_21", "Dec\_21"…

$ time *<dttm>* 2023-01-07 15:06:00, 2023-01-07 03:43…

$ ride\_length *<dbl>* 7.583333, 26.900000, 12.766667, 14.716…

> any(all\_trips\_v2$ride\_length < 0)

[1] FALSE

> any(all\_trips\_v2$start\_station\_name %like% "Test")

[1] FALSE

* Double checking to make sure all < 0 and “Test” have been removed.

**Checking member\_casual Column**

> table(all\_trips\_v2$member\_casual)

casual member

2346820 3386530

> setNames(aggregate(ride\_length ~ member\_casual, all\_trips\_v2, sum), c("member\_casual", "total\_ride\_length(mins)"))

member\_casual total\_ride\_length(mins)

1 casual 68313470

2 member 43039268

* With these two tests we can ensure that the member\_casual column only has two distinct values being “casual” and “member”. This also gives us a quick comparison between the total number of ‘members vs casuals’ and the amount of time each spent riding a bike.

**4. Analyze/Share**

**Summary of ride length**

> summary(all\_trips\_v2$ride\_length)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 5.83 10.30 19.42 18.50 41387.25

**Comparing Casual riders and Member riders**

* Mean, Median, Max, Min

> aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = mean)

member\_casual ride\_length

1 casual 29.10895

2 member 12.70896

> aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = median)

member\_casual ride\_length

1 casual 13.050000

2 member 8.833333

> aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = max)

member\_casual ride\_length

1 casual 41387.25

2 member 1559.90

> aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = min)

member\_casual ride\_length

1 casual 0

2 member 0

**Observation**

* In the “mean” and “median” tables, Casual riders have a longer ride length than member riders. Casual riders also have a higher mean ride length compared to the all\_trips\_v2 mean, while member riders have a lower mean. This shows that casual riders ride the bikes longer than member riders.

**Comparing total number of trips by customer type on each day of the week**

>all\_trips\_v2$day\_of\_the\_week <- ordered(all\_trips\_v2$day\_of\_the\_week, levels=c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))

> all\_trips\_v2$month <- ordered(all\_trips\_v2$month, levels=c("Dec\_21", "Jan\_22", "Feb\_22", "Mar\_22", "Apr\_22", "May\_22", "Jun\_22","Jul\_22","Aug\_22", "Sep\_22", "Oct\_22", "Nov\_22"))

* Add days of the week and the months together in the same sequence so they show up together in the tables.

> all\_trips\_v2 %>%

group\_by(member\_casual, day\_of\_the\_week) %>%

summarise(number\_of\_rides = n(),average\_duration\_mins = mean(ride\_length)) %>%

arrange(member\_casual, desc(number\_of\_rides))

# A tibble: 14 × 4

# Groups: member\_casual [2]

member\_casual day\_of\_the\_week number\_of\_rides average\_duration\_mins

*<chr>* *<chr>* *<int>* *<dbl>*

1 casual Sat 476583 32.5

2 casual Sun 392105 34.1

3 casual Fri 340495 27.8

4 casual Thu 313736 25.6

5 casual Mon 280468 29.2

6 casual Wed 279380 24.7

7 casual Tue 264053 26.0

8 member Thu 540341 12.3

9 member Wed 537743 12.1

10 member Tue 518657 12.1

11 member Mon 476931 12.3

12 member Fri 476905 12.5

13 member Sat 445466 14.2

14 member Sun 390487 14.0

**Visualization for the table above**

all\_trips\_v2 %>%

group\_by(member\_casual, day\_of\_the\_week) %>%

summarise(number\_of\_rides = n()) %>%

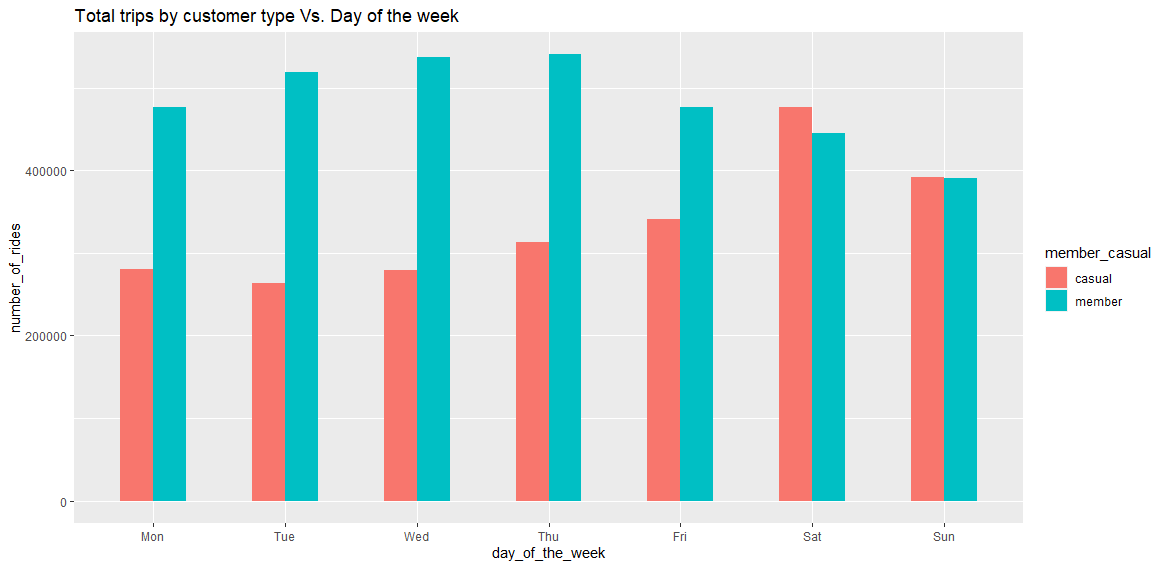
arrange(member\_casual, day\_of\_the\_week) %>%

ggplot(aes(x = day\_of\_the\_week, y = number\_of\_rides, fill = member\_casual)) +

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))

****

The graph shows that casual riders tend to ride more on the weekends than the weekdays. While member riders tend to ride more during the weekdays and slow down during the weekends.

**Comparing average number of trips by customer type and month**

all\_trips\_v2 %>%

group\_by(member\_casual, month) %>%

summarise(number\_of\_rides = n(),`average\_duration\_(mins)` = mean(ride\_length)) %>%

arrange(member\_casual,desc(number\_of\_rides))

# A tibble: 24 × 4

# Groups: member\_casual [2]

member\_casual month number\_of\_rides `average\_duration\_(mins)`

*<chr>* *<ord>* *<int>* *<dbl>*

1 casual Jul\_22 406046 29.3

2 casual Jun\_22 369044 32.1

3 casual Aug\_22 358917 29.3

4 casual Sep\_22 296694 28.0

5 casual May\_22 280414 30.9

6 casual Oct\_22 208988 26.4

7 casual Apr\_22 126417 29.5

8 casual Nov\_22 100747 21.3

9 casual Mar\_22 89880 32.6

10 casual Dec\_21 69738 23.5

11 casual Feb\_22 21416 26.7

12 casual Jan\_22 18519 30.4

13 member Aug\_22 427000 13.4

14 member Jul\_22 417426 13.7

15 member Sep\_22 404636 13.0

16 member Jun\_22 400148 14.0

17 member May\_22 354443 13.4

18 member Oct\_22 349693 12.0

19 member Apr\_22 244832 11.5

20 member Nov\_22 236947 11.1

21 member Mar\_22 194160 12.0

22 member Dec\_21 177802 11.0

23 member Feb\_22 94193 11.4

24 member Jan\_22 85250 12.0

**Visualization of the table above**

all\_trips\_v2 %>%

group\_by(member\_casual, month) %>%

summarise(number\_of\_rides = n()) %>%

arrange(member\_casual, month) %>%

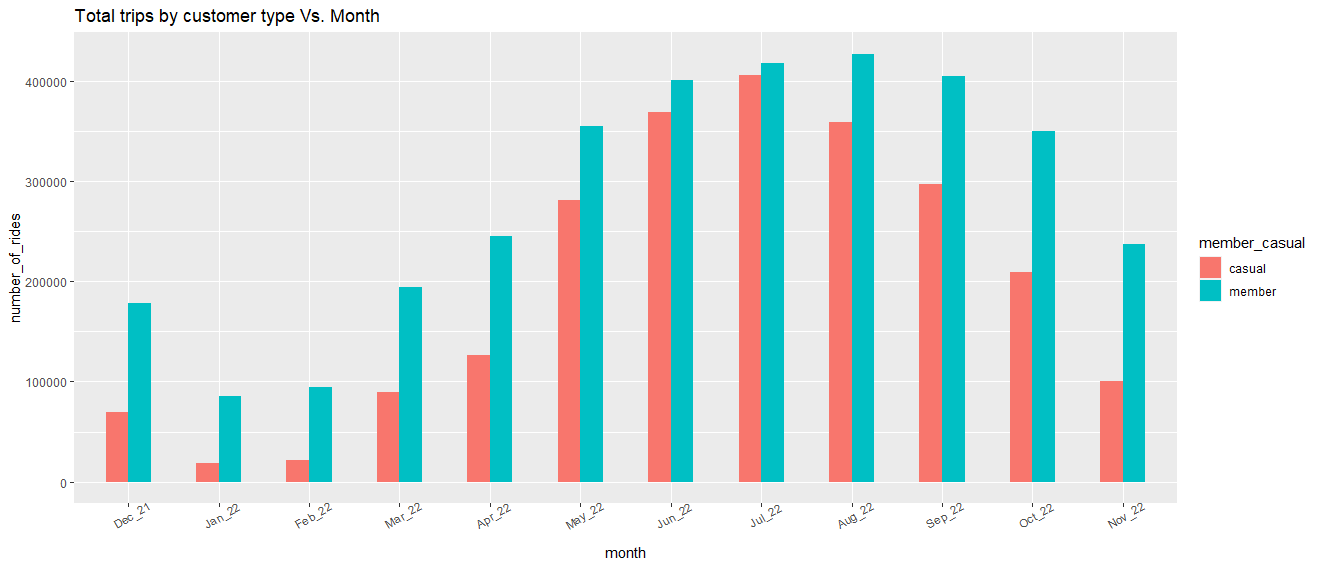
ggplot(aes(x = month, y = number\_of\_rides, fill = member\_casual)) +

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))



The graph shows that both member riders and casual riders ride more between May 2022 and October 2022, and less in December 2021 through April of 2022 and November 2022. This could be due to external factors like weather and temperature. During each of the months the member riders rode bikes more often than casual riders.

**Visualization for the average ride length for each customer type to each day of the week**

all\_trips\_v2 %>%

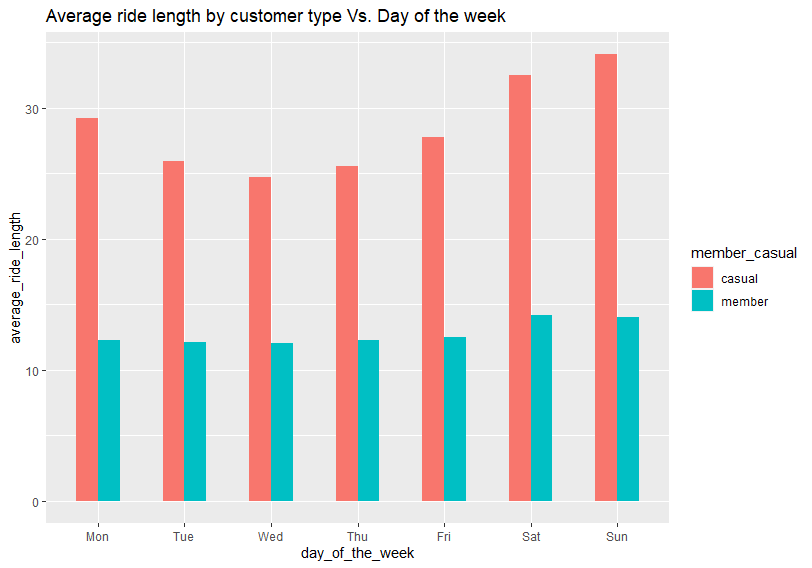
group\_by(member\_casual, day\_of\_the\_week) %>%

summarise(average\_ride\_length = mean(ride\_length)) %>%

ggplot(aes(x = day\_of\_the\_week, y = average\_ride\_length, fill = member\_casual)) +

geom\_col(width=0.5, position = position\_dodge(width=0.5)) +

labs(title ="Average ride length by customer type Vs. Day of the week")



The graph shows that on average casual riders' ride length (in minutes) is longer during everyday of the week compared to the member riders. Members' ride length continues to stay relatively the same throughout the week, while casual riders ride length becomes longer during saturday and sunday.

**Visualization for average ride length for each customer type to month**

all\_trips\_v2 %>%

group\_by(member\_casual, month) %>%

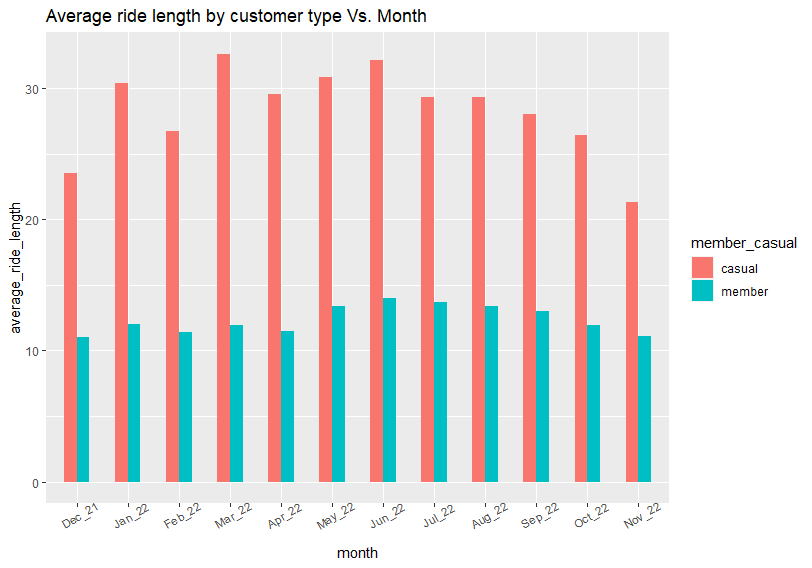
summarise(average\_ride\_length = mean(ride\_length)) %>%

ggplot(aes(x = month, y = average\_ride\_length, fill = member\_casual)) +

geom\_col(width=0.5, position = position\_dodge(width=0.5)) +

labs(title ="Average ride length by customer type Vs. Month") +

theme(axis.text.x = element\_text(angle = 30))



The graph shows that on average casual riders' ride length (in minutes) is longer during every month compared to the member riders. Casual riders ride length peaks out in the months January, March, and June. This could be due to the consistent weather patterns in those areas during those months. (Example: More sunny days, less rain, and temperature).

**Visualization of the demand for bikes during a 24 hour period**

all\_trips\_v2 %>%

group\_by(member\_casual, time) %>%

summarise(number\_of\_trips = n()) %>%

ggplot(aes(x = time, y = number\_of\_trips, color = member\_casual, group = member\_casual)) +

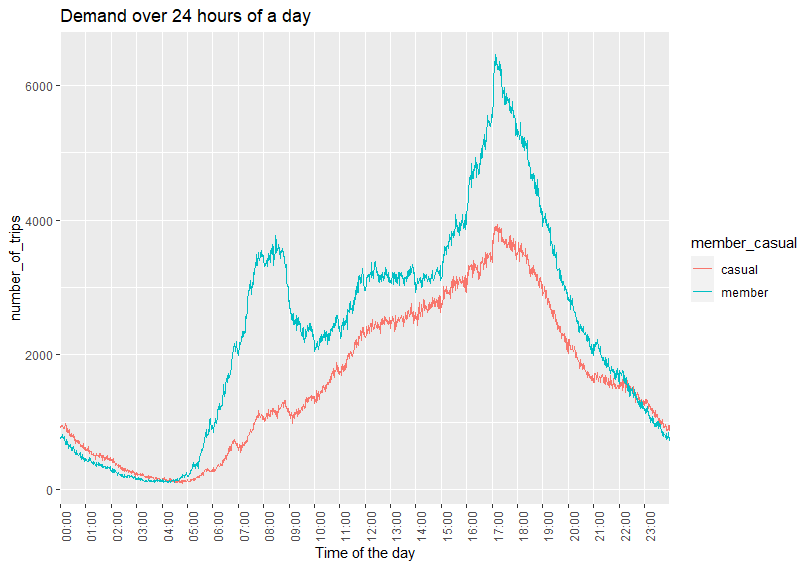
geom\_line() +

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")



The graph shows that during a normal 24 hour period of the day member riders' bike demand is higher than casual riders demand, except for between the hours of 10pm and 4am. During those house casual riders demand is higher. During the day member riders have two strong peaks of demand during 7am - 8am, and 5pm. These hours tend to follow times when people are traveling between work and home. The casual riders bike demand starts low in the morning around 5am, but continues to gradually increase up till 5pm. After 5pm both members and casual riders demand start to fall.

**Visualization comparing rideable type to the number of trips along with customer type**

all\_trips\_v2 %>%

group\_by(rideable\_type, member\_casual) %>%

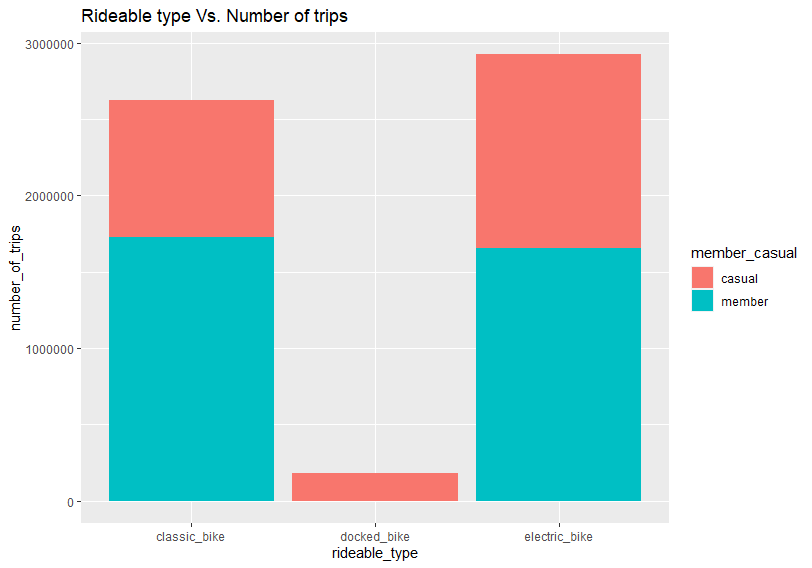
summarise(number\_of\_trips = n()) %>%

ggplot(aes(x= rideable\_type, y=number\_of\_trips, fill= member\_casual))+

geom\_bar(stat='identity') +

scale\_y\_continuous(labels = function(x) format(x, scientific = FALSE)) +

labs(title ="Rideable type Vs. Number of trips")



The graph shows that both classic bikes and electric bikes are the most frequently ridden types of bikes by both members and casual riders. Most of the classic bikes and electric bikes are ridden by member riders. The docked bikes were hardly ridden at all, and mostly ridden by only casual riders. (Although the number of each bike that is available to ride is not provided. So we do not know if a cause for less people riding docked bikes is because of a lack of docked bikes.)

**Export summary file for further analysis**

clean\_data <- aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$day\_of\_the\_week, FUN = mean)

write.csv(clean\_data, "Clean Data.csv", row.names = F)

* File is now labeled as clean\_data

**5. Act**

**Key findings**

* Casual riders have a longer mean ride length than member riders.
* Casual riders tend to ride more on the weekends than the weekdays. While member riders tend to ride more during the weekdays and slow down during the weekends.
* Both member riders and casual riders ride more between May 2022 and October 2022, and less in December 2021 through April of 2022 and November 2022.
* Casual riders' ride length (in minutes) is longer during everyday of the week compared to the member riders.
* Casual riders' ride length (in minutes) is longer during every month compared to the member riders.
* During a normal 24 hour period of the day member riders' bike demand is higher than casual riders demand, except for between the hours of 10pm and 4am.
* Both classic bikes and electric bikes are the most frequently ridden types of bikes by both members and casual riders.

**Recommendations**

* Offer monthly deals for new customers to push for new people to join as members. We see this a lot with workout gyms. Sometimes gyms will offer a join for $1 this month promotion to boost memberships.
* Offer discounts or promotions during the none busy months for casual riders to promote more people becoming members.
* Offer a free trial for a specific period of time.

**Reminder**

* The data we used does have some missing components that makes the analysis a little more difficult. We were not provided any personal data on the customers, so we do not know if the same person rode multiple times. We also do not know where these people reside. During the analysis, we also ran into an issue where we did not know the amount of each bike type that is available.

**References/Resources**

* [Google Data Analytics Professional Certificate](https://www.coursera.org/professional-certificates/google-data-analytics)
* [License](https://ride.divvybikes.com/data-license-agreement)
* [Data](https://divvy-tripdata.s3.amazonaws.com/index.html)
* [Rstudio](https://community.rstudio.com/)
* [Stack Overflow](https://stackoverflow.com/)
* [Kaggle](https://www.kaggle.com/)

**Contact**

* [LinkedIn](https://www.linkedin.com/in/zachary-mitchell-213608141/)