R version 4.3.0 (2023-04-21 ucrt) -- "Already Tomorrow"

Copyright (C) 2023 The R Foundation for Statistical Computing

Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

> library(readr)

> Divvy\_Trips\_2019\_Q2 <- read\_csv("C:/Users/nicky/Downloads/google data analytics/Capstone/project/2019/Divvy\_Trips\_2019\_Q2/Divvy\_Trips\_2019\_Q2.csv")

**Rows:** 1108163 **Columns:** 12

── **Column specification** ────────────────────────────────────────────────

**Delimiter:** ","

chr (4): 03 - Rental Start Station Name, 02 - Rental End Station Na...

dbl (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID...

num (1): 01 - Rental Details Duration In Seconds Uncapped

dttm (2): 01 - Rental Details Local Start Time, 01 - Rental Details ...

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.

> View(Divvy\_Trips\_2019\_Q2)

> library(readr)

> Divvy\_Trips\_2019\_Q3 <- read\_csv("C:/Users/nicky/Downloads/google data analytics/Capstone/project/2019/Divvy\_Trips\_2019\_Q3/Divvy\_Trips\_2019\_Q3.csv")

**Rows:** 1640718 **Columns:** 12

── **Column specification** ────────────────────────────────────────────────

**Delimiter:** ","

chr (4): from\_station\_name, to\_station\_name, usertype, gender

dbl (5): trip\_id, bikeid, from\_station\_id, to\_station\_id, birthyear

num (1): tripduration

dttm (2): start\_time, end\_time

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.

> View(Divvy\_Trips\_2019\_Q3)

> library(readr)

> Divvy\_Trips\_2019\_Q4 <- read\_csv("C:/Users/nicky/Downloads/google data analytics/Capstone/project/2019/Divvy\_Trips\_2019\_Q4/Divvy\_Trips\_2019\_Q4.csv")

**Rows:** 704054 **Columns:** 12

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (4): from\_station\_name, to\_station\_name, usertype, gender

dbl (5): trip\_id, bikeid, from\_station\_id, to\_station\_id, birthyear

num (1): tripduration

dttm (2): start\_time, end\_time

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.

> View(Divvy\_Trips\_2019\_Q4)

> library(readr)

> Divvy\_Trips\_2020\_Q1 <- read\_csv("C:/Users/nicky/Downloads/google data analytics/Capstone/project/2020/Divvy\_Trips\_2020\_Q1/Divvy\_Trips\_2020\_Q1.csv")

**Rows:** 426887 **Columns:** 13

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, member\_casual

dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, end\_lng

dttm (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.

> View(Divvy\_Trips\_2020\_Q1)

> install.packages("tidyverse")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/nicky/AppData/Local/R/win-library/4.3’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/tidyverse\_2.0.0.zip'

Content type 'application/zip' length 430853 bytes (420 KB)

downloaded 420 KB

package ‘tidyverse’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\nicky\AppData\Local\Temp\RtmpqceYko\downloaded\_packages

> library(tidyverse)

── **Attaching core tidyverse packages** ───────────────────────────────────────────────────────────────── tidyverse 2.0.0 ──

✔ dplyr 1.1.3 ✔ purrr 1.0.2

✔ forcats 1.0.0 ✔ stringr 1.5.0

✔ ggplot2 3.4.4 ✔ tibble 3.2.1

✔ lubridate 1.9.3 ✔ tidyr 1.3.0

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

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

✖ dplyr::lag() masks stats::lag()

ℹ Use the conflicted package to force all conflicts to become errors

Warning messages:

1: package ‘tidyverse’ was built under R version 4.3.2

2: package ‘ggplot2’ was built under R version 4.3.2

3: package ‘purrr’ was built under R version 4.3.2

4: package ‘dplyr’ was built under R version 4.3.2

5: package ‘lubridate’ was built under R version 4.3.2

> library(lubridate)

> library(ggplot2)

> getwd()

[1] "C:/Users/nicky/OneDrive/Documents"

> q2\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q2.csv")

Error: 'Divvy\_Trips\_2019\_Q2.csv' does not exist in current working directory ('C:/Users/nicky/OneDrive/Documents').

> setwd("C:\Users\nicky\Downloads\google data analytics\Capstone\project\cyclistic\R")

Error: '\U' used without hex digits in character string (<input>:1:11)

> setwd("C:/Users/nicky/Downloads/google data analytics/Capstone/project/cyclistic/R")

> getwd()

[1] "C:/Users/nicky/Downloads/google data analytics/Capstone/project/cyclistic/R"

STEP 1 : DATA COLLECTION

> q2\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q2.csv")

**Rows:** 1108163 **Columns:** 12

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (4): 03 - Rental Start Station Name, 02 - Rental End Station Name, User Type, Member Gender

dbl (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID, 03 - Rental Start Station ID, 02 - Rental End S...

num (1): 01 - Rental Details Duration In Seconds Uncapped

dttm (2): 01 - Rental Details Local Start Time, 01 - Rental Details Local End Time

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.

> q3\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q3.csv")

**Rows:** 1640718 **Columns:** 12

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (4): from\_station\_name, to\_station\_name, usertype, gender

dbl (5): trip\_id, bikeid, from\_station\_id, to\_station\_id, birthyear

num (1): tripduration

dttm (2): start\_time, end\_time

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.

> q4\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q4.csv")

**Rows:** 704054 **Columns:** 12

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (4): from\_station\_name, to\_station\_name, usertype, gender

dbl (5): trip\_id, bikeid, from\_station\_id, to\_station\_id, birthyear

num (1): tripduration

dttm (2): start\_time, end\_time

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.

> q1\_2020 <- read\_csv("Divvy\_Trips\_2020\_Q1.csv")

**Rows:** 426887 **Columns:** 13

── **Column specification** ─────────────────────────────────────────────────────────────────────────────────────────────────

**Delimiter:** ","

chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, member\_casual

dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, end\_lng

dttm (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.

STEP 2 : WRANGLE DATA AND COMBINE INTO A SINGLE FILE

> colnames(q3\_2019)

[1] "trip\_id" "start\_time" "end\_time" "bikeid" "tripduration"

[6] "from\_station\_id" "from\_station\_name" "to\_station\_id" "to\_station\_name" "usertype"

[11] "gender" "birthyear"

> colnames(q4\_2019)

[1] "trip\_id" "start\_time" "end\_time" "bikeid" "tripduration"

[6] "from\_station\_id" "from\_station\_name" "to\_station\_id" "to\_station\_name" "usertype"

[11] "gender" "birthyear"

> colnames(q2\_2019)

[1] "01 - Rental Details Rental ID" "01 - Rental Details Local Start Time"

[3] "01 - Rental Details Local End Time" "01 - Rental Details Bike ID"

[5] "01 - Rental Details Duration In Seconds Uncapped" "03 - Rental Start Station ID"

[7] "03 - Rental Start Station Name" "02 - Rental End Station ID"

[9] "02 - Rental End Station Name" "User Type"

[11] "Member Gender" "05 - Member Details Member Birthday Year"

> colnames(q1\_2020)

[1] "ride\_id" "rideable\_type" "started\_at" "ended\_at" "start\_station\_name"

[6] "start\_station\_id" "end\_station\_name" "end\_station\_id" "start\_lat" "start\_lng"

[11] "end\_lat" "end\_lng" "member\_casual"

> (q4\_2019 <- rename(q4\_2019

+ ,ride\_id = trip\_id

+ ,rideable\_type = bikeid

+ ,started\_at = start\_time

+ ,ended\_at = end\_time

+ ,start\_station\_name = from\_station\_name

+ ,start\_station\_id = from\_station\_id

+ ,end\_station\_name = to\_station\_name

+ ,end\_station\_id = to\_station\_id

+ ,member\_casual = usertype))

# A tibble: 704,054 × 12

ride\_id started\_at ended\_at rideable\_type tripduration start\_station\_id start\_station\_name

*<dbl>* *<dttm>* *<dttm>* *<dbl>* *<dbl>* *<dbl>* *<chr>*

1 25223640 2019-10-01 00:01:39 2019-10-01 00:17:20 2215 940 20 Sheffield Ave & Kingsbur…

2 25223641 2019-10-01 00:02:16 2019-10-01 00:06:34 6328 258 19 Throop (Loomis) St & Tay…

3 25223642 2019-10-01 00:04:32 2019-10-01 00:18:43 3003 850 84 Milwaukee Ave & Grand Ave

4 25223643 2019-10-01 00:04:32 2019-10-01 00:43:43 3275 2350 313 Lakeview Ave & Fullerton…

5 25223644 2019-10-01 00:04:34 2019-10-01 00:35:42 5294 1867 210 Ashland Ave & Division St

6 25223645 2019-10-01 00:04:38 2019-10-01 00:10:51 1891 373 156 Clark St & Wellington Ave

7 25223646 2019-10-01 00:04:52 2019-10-01 00:22:45 1061 1072 84 Milwaukee Ave & Grand Ave

8 25223647 2019-10-01 00:04:57 2019-10-01 00:29:16 1274 1458 156 Clark St & Wellington Ave

9 25223648 2019-10-01 00:05:20 2019-10-01 00:29:18 6011 1437 156 Clark St & Wellington Ave

10 25223649 2019-10-01 00:05:20 2019-10-01 02:23:46 2957 8306 336 Cottage Grove Ave & 47th…

# ℹ 704,044 more rows

# ℹ 5 more variables: end\_station\_id <dbl>, end\_station\_name <chr>, member\_casual <chr>, gender <chr>, birthyear <dbl>

# ℹ Use `print(n = ...)` to see more rows

> (q3\_2019 <- rename(q3\_2019

+ ,ride\_id = trip\_id

+ ,rideable\_type = bikeid

+ ,started\_at = start\_time

+ ,ended\_at = end\_time

+ ,start\_station\_name = from\_station\_name

+ ,start\_station\_id = from\_station\_id

+ ,end\_station\_name = to\_station\_name

+ ,end\_station\_id = to\_station\_id

+ ,member\_casual = usertype))

# A tibble: 1,640,718 × 12

ride\_id started\_at ended\_at rideable\_type tripduration start\_station\_id start\_station\_name

*<dbl>* *<dttm>* *<dttm>* *<dbl>* *<dbl>* *<dbl>* *<chr>*

1 23479388 2019-07-01 00:00:27 2019-07-01 00:20:41 3591 1214 117 Wilton Ave & Belmont Ave

2 23479389 2019-07-01 00:01:16 2019-07-01 00:18:44 5353 1048 381 Western Ave & Monroe St

3 23479390 2019-07-01 00:01:48 2019-07-01 00:27:42 6180 1554 313 Lakeview Ave & Fullerton…

4 23479391 2019-07-01 00:02:07 2019-07-01 00:27:10 5540 1503 313 Lakeview Ave & Fullerton…

5 23479392 2019-07-01 00:02:13 2019-07-01 00:22:26 6014 1213 168 Michigan Ave & 14th St

6 23479393 2019-07-01 00:02:21 2019-07-01 00:07:31 4941 310 300 Broadway & Barry Ave

7 23479394 2019-07-01 00:02:24 2019-07-01 00:23:12 3770 1248 168 Michigan Ave & 14th St

8 23479395 2019-07-01 00:02:26 2019-07-01 00:28:16 5442 1550 313 Lakeview Ave & Fullerton…

9 23479396 2019-07-01 00:02:34 2019-07-01 00:28:57 2957 1583 43 Michigan Ave & Washingto…

10 23479397 2019-07-01 00:02:45 2019-07-01 00:29:14 6091 1589 43 Michigan Ave & Washingto…

# ℹ 1,640,708 more rows

# ℹ 5 more variables: end\_station\_id <dbl>, end\_station\_name <chr>, member\_casual <chr>, gender <chr>, birthyear <dbl>

# ℹ Use `print(n = ...)` to see more rows

> (q2\_2019 <- rename(q2\_2019

+ ,ride\_id = "01 - Rental Details Rental ID"

+ ,rideable\_type = "01 - Rental Details Bike ID"

+ ,started\_at = "01 - Rental Details Local Start Time"

+ ,ended\_at = "01 - Rental Details Local End Time"

+ ,start\_station\_name = "03 - Rental Start Station Name"

+ ,start\_station\_id = "03 - Rental Start Station ID"

+ ,end\_station\_name = "02 - Rental End Station Name"

+ ,end\_station\_id = "02 - Rental End Station ID"

+ ,member\_casual = "User Type"))

# A tibble: 1,108,163 × 12

ride\_id started\_at ended\_at rideable\_type 01 - Rental Details Duration In Seco…¹ start\_station\_id

*<dbl>* *<dttm>* *<dttm>* *<dbl>* *<dbl>* *<dbl>*

1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48 6251 446 81

2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30 6226 1048 317

3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19 5649 252 283

4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58 4151 357 26

5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13 3270 1007 202

6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56 3123 257 420

7 22178535 2019-04-01 00:26:33 2019-04-01 00:35:41 6418 548 503

8 22178536 2019-04-01 00:29:48 2019-04-01 00:36:11 4513 383 260

9 22178537 2019-04-01 00:32:07 2019-04-01 01:07:44 3280 2137 211

10 22178538 2019-04-01 00:32:19 2019-04-01 01:07:39 5534 2120 211

# ℹ 1,108,153 more rows

# ℹ abbreviated name: ¹​`01 - Rental Details Duration In Seconds Uncapped`

# ℹ 6 more variables: start\_station\_name <chr>, end\_station\_id <dbl>, end\_station\_name <chr>, member\_casual <chr>,

# `Member Gender` <chr>, `05 - Member Details Member Birthday Year` <dbl>

# ℹ Use `print(n = ...)` to see more rows

> str(q1\_2020)

spc\_tbl\_ [426,887 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:426887] "EACB19130B0CDA4A" "8FED874C809DC021" "789F3C21E472CA96" "C9A388DAC6ABF313" ...

$ rideable\_type : chr [1:426887] "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...

$ started\_at : POSIXct[1:426887], format: "2020-01-21 20:06:59" "2020-01-30 14:22:39" "2020-01-09 19:29:26" "2020-01-06 16:17:07" ...

$ ended\_at : POSIXct[1:426887], format: "2020-01-21 20:14:30" "2020-01-30 14:26:22" "2020-01-09 19:32:17" "2020-01-06 16:25:56" ...

$ start\_station\_name: chr [1:426887] "Western Ave & Leland Ave" "Clark St & Montrose Ave" "Broadway & Belmont Ave" "Clark St & Randolph St" ...

$ start\_station\_id : num [1:426887] 239 234 296 51 66 212 96 96 212 38 ...

$ end\_station\_name : chr [1:426887] "Clark St & Leland Ave" "Southport Ave & Irving Park Rd" "Wilton Ave & Belmont Ave" "Fairbanks Ct & Grand Ave" ...

$ end\_station\_id : num [1:426887] 326 318 117 24 212 96 212 212 96 100 ...

$ start\_lat : num [1:426887] 42 42 41.9 41.9 41.9 ...

$ start\_lng : num [1:426887] -87.7 -87.7 -87.6 -87.6 -87.6 ...

$ end\_lat : num [1:426887] 42 42 41.9 41.9 41.9 ...

$ end\_lng : num [1:426887] -87.7 -87.7 -87.7 -87.6 -87.6 ...

$ member\_casual : chr [1:426887] "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\_double(),

.. end\_station\_name = col\_character(),

.. end\_station\_id = col\_double(),

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

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

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

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

.. member\_casual = col\_character()

.. )

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

> str(q4\_2019)

spc\_tbl\_ [704,054 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : num [1:704054] 25223640 25223641 25223642 25223643 25223644 ...

$ started\_at : POSIXct[1:704054], format: "2019-10-01 00:01:39" "2019-10-01 00:02:16" "2019-10-01 00:04:32" "2019-10-01 00:04:32" ...

$ ended\_at : POSIXct[1:704054], format: "2019-10-01 00:17:20" "2019-10-01 00:06:34" "2019-10-01 00:18:43" "2019-10-01 00:43:43" ...

$ rideable\_type : num [1:704054] 2215 6328 3003 3275 5294 ...

$ tripduration : num [1:704054] 940 258 850 2350 1867 ...

$ start\_station\_id : num [1:704054] 20 19 84 313 210 156 84 156 156 336 ...

$ start\_station\_name: chr [1:704054] "Sheffield Ave & Kingsbury St" "Throop (Loomis) St & Taylor St" "Milwaukee Ave & Grand Ave" "Lakeview Ave & Fullerton Pkwy" ...

$ end\_station\_id : num [1:704054] 309 241 199 290 382 226 142 463 463 336 ...

$ end\_station\_name : chr [1:704054] "Leavitt St & Armitage Ave" "Morgan St & Polk St" "Wabash Ave & Grand Ave" "Kedzie Ave & Palmer Ct" ...

$ member\_casual : chr [1:704054] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...

$ gender : chr [1:704054] "Male" "Male" "Female" "Male" ...

$ birthyear : num [1:704054] 1987 1998 1991 1990 1987 ...

- attr(\*, "spec")=

.. cols(

.. trip\_id = col\_double(),

.. start\_time = col\_datetime(format = ""),

.. end\_time = col\_datetime(format = ""),

.. bikeid = col\_double(),

.. tripduration = col\_number(),

.. from\_station\_id = col\_double(),

.. from\_station\_name = col\_character(),

.. to\_station\_id = col\_double(),

.. to\_station\_name = col\_character(),

.. usertype = col\_character(),

.. gender = col\_character(),

.. birthyear = col\_double()

.. )

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

> str(q3\_2019)

spc\_tbl\_ [1,640,718 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : num [1:1640718] 23479388 23479389 23479390 23479391 23479392 ...

$ started\_at : POSIXct[1:1640718], format: "2019-07-01 00:00:27" "2019-07-01 00:01:16" "2019-07-01 00:01:48" "2019-07-01 00:02:07" ...

$ ended\_at : POSIXct[1:1640718], format: "2019-07-01 00:20:41" "2019-07-01 00:18:44" "2019-07-01 00:27:42" "2019-07-01 00:27:10" ...

$ rideable\_type : num [1:1640718] 3591 5353 6180 5540 6014 ...

$ tripduration : num [1:1640718] 1214 1048 1554 1503 1213 ...

$ start\_station\_id : num [1:1640718] 117 381 313 313 168 300 168 313 43 43 ...

$ start\_station\_name: chr [1:1640718] "Wilton Ave & Belmont Ave" "Western Ave & Monroe St" "Lakeview Ave & Fullerton Pkwy" "Lakeview Ave & Fullerton Pkwy" ...

$ end\_station\_id : num [1:1640718] 497 203 144 144 62 232 62 144 195 195 ...

$ end\_station\_name : chr [1:1640718] "Kimball Ave & Belmont Ave" "Western Ave & 21st St" "Larrabee St & Webster Ave" "Larrabee St & Webster Ave" ...

$ member\_casual : chr [1:1640718] "Subscriber" "Customer" "Customer" "Customer" ...

$ gender : chr [1:1640718] "Male" NA NA NA ...

$ birthyear : num [1:1640718] 1992 NA NA NA NA ...

- attr(\*, "spec")=

.. cols(

.. trip\_id = col\_double(),

.. start\_time = col\_datetime(format = ""),

.. end\_time = col\_datetime(format = ""),

.. bikeid = col\_double(),

.. tripduration = col\_number(),

.. from\_station\_id = col\_double(),

.. from\_station\_name = col\_character(),

.. to\_station\_id = col\_double(),

.. to\_station\_name = col\_character(),

.. usertype = col\_character(),

.. gender = col\_character(),

.. birthyear = col\_double()

.. )

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

> str(q2\_2019)

spc\_tbl\_ [1,108,163 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : num [1:1108163] 22178529 22178530 22178531 22178532 22178533 ...

$ started\_at : POSIXct[1:1108163], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" "2019-04-01 00:11:07" "2019-04-01 00:13:01" ...

$ ended\_at : POSIXct[1:1108163], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" "2019-04-01 00:15:19" "2019-04-01 00:18:58" ...

$ rideable\_type : num [1:1108163] 6251 6226 5649 4151 3270 ...

$ 01 - Rental Details Duration In Seconds Uncapped: num [1:1108163] 446 1048 252 357 1007 ...

$ start\_station\_id : num [1:1108163] 81 317 283 26 202 420 503 260 211 211 ...

$ start\_station\_name : chr [1:1108163] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jackson Blvd" "McClurg Ct & Illinois St" ...

$ end\_station\_id : num [1:1108163] 56 59 174 133 129 426 500 499 211 211 ...

$ end\_station\_name : chr [1:1108163] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...

$ member\_casual : chr [1:1108163] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...

$ Member Gender : chr [1:1108163] "Male" "Female" "Male" "Male" ...

$ 05 - Member Details Member Birthday Year : num [1:1108163] 1975 1984 1990 1993 1992 ...

- attr(\*, "spec")=

.. cols(

.. `01 - Rental Details Rental ID` = col\_double(),

.. `01 - Rental Details Local Start Time` = col\_datetime(format = ""),

.. `01 - Rental Details Local End Time` = col\_datetime(format = ""),

.. `01 - Rental Details Bike ID` = col\_double(),

.. `01 - Rental Details Duration In Seconds Uncapped` = col\_number(),

.. `03 - Rental Start Station ID` = col\_double(),

.. `03 - Rental Start Station Name` = col\_character(),

.. `02 - Rental End Station ID` = col\_double(),

.. `02 - Rental End Station Name` = col\_character(),

.. `User Type` = col\_character(),

.. `Member Gender` = col\_character(),

.. `05 - Member Details Member Birthday Year` = col\_double()

.. )

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

> q4\_2019 <- mutate(q4\_2019, ride\_id = as.character(ride\_id)

+ ,rideable\_type = as.character(rideable\_type))

> q3\_2019 <- mutate(q3\_2019, ride\_id = as.character(ride\_id)

+ ,rideable\_type = as.character(rideable\_type))

> q2\_2019 <- mutate(q2\_2019, ride\_id = as.character(ride\_id)

+ ,rideable\_type = as.character(rideable\_type))

> all\_trips <- bind\_rows(q2\_2019, q3\_2019, q4\_2019, q1\_2020)

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

+ select(-c(start\_lat, start\_lng, end\_lat, end\_lng, birthyear, gender, "01 - Rental Details Duration In Seconds Uncapped", "05 - Member Details Member Birthday Year", "Member Gender", "tripduration"))

STEP 3 : CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

> colnames(all\_trips)

[1] "ride\_id" "started\_at" "ended\_at" "rideable\_type" "start\_station\_id"

[6] "start\_station\_name" "end\_station\_id" "end\_station\_name" "member\_casual"

> nrow(all\_trips)

[1] 3879822

> dim(all\_trips)

[1] 3879822 9

> head(all\_trips)

# A tibble: 6 × 9

ride\_id started\_at ended\_at rideable\_type start\_station\_id start\_station\_name end\_station\_id

*<chr>* *<dttm>* *<dttm>* *<chr>* *<dbl>* *<chr>* *<dbl>*

1 22178529 2019-04-01 00:02:22 2019-04-01 00:09:48 6251 81 Daley Center Plaza 56

2 22178530 2019-04-01 00:03:02 2019-04-01 00:20:30 6226 317 Wood St & Taylor St 59

3 22178531 2019-04-01 00:11:07 2019-04-01 00:15:19 5649 283 LaSalle St & Jackson Bl… 174

4 22178532 2019-04-01 00:13:01 2019-04-01 00:18:58 4151 26 McClurg Ct & Illinois St 133

5 22178533 2019-04-01 00:19:26 2019-04-01 00:36:13 3270 202 Halsted St & 18th St 129

6 22178534 2019-04-01 00:19:39 2019-04-01 00:23:56 3123 420 Ellis Ave & 55th St 426

# ℹ 2 more variables: end\_station\_name <chr>, member\_casual <chr>

> str(all\_trips)

tibble [3,879,822 × 9] (S3: tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...

$ started\_at : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" "2019-04-01 00:11:07" "2019-04-01 00:13:01" ...

$ ended\_at : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" "2019-04-01 00:15:19" "2019-04-01 00:18:58" ...

$ rideable\_type : chr [1:3879822] "6251" "6226" "5649" "4151" ...

$ start\_station\_id : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...

$ start\_station\_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jackson Blvd" "McClurg Ct & Illinois St" ...

$ end\_station\_id : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...

$ end\_station\_name : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...

$ member\_casual : chr [1:3879822] "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...

> summary(all\_trips)

ride\_id started\_at ended\_at rideable\_type

Length:3879822 Min. :2019-04-01 00:02:22.00 Min. :2019-04-01 00:09:48.00 Length:3879822

Class :character 1st Qu.:2019-06-23 07:49:09.25 1st Qu.:2019-06-23 08:20:27.75 Class :character

Mode :character Median :2019-08-14 17:43:38.00 Median :2019-08-14 18:02:04.00 Mode :character

Mean :2019-08-26 00:49:59.38 Mean :2019-08-26 01:14:37.06

3rd Qu.:2019-10-12 12:10:21.00 3rd Qu.:2019-10-12 12:36:16.75

Max. :2020-03-31 23:51:34.00 Max. :2020-05-19 20:10:34.00

start\_station\_id start\_station\_name end\_station\_id end\_station\_name member\_casual

Min. : 1.0 Length:3879822 Min. : 1.0 Length:3879822 Length:3879822

1st Qu.: 77.0 Class :character 1st Qu.: 77.0 Class :character Class :character

Median :174.0 Mode :character Median :174.0 Mode :character Mode :character

Mean :202.9 Mean :203.8

3rd Qu.:291.0 3rd Qu.:291.0

Max. :675.0 Max. :675.0

NA's :1

> table(all\_trips$member\_casual)

casual Customer member Subscriber

48480 857474 378407 2595461

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

+ mutate(member\_casual = recode(member\_casual

+ ,"Subscriber" = "member"

+ ,"Customer" = "casual"))

> table(all\_trips$member\_casual)

casual member

905954 2973868

> all\_trips$date <- as.Date(all\_trips$started\_at)

> all\_trips$month <- format(as.Date(all\_trips$date), "%m")

> all\_trips$day <- format(as.Date(all\_trips$date), "%d")

> all\_trips$year <- format(as.Date(all\_trips$date), "%Y")

> all\_trips$day\_of\_week <- format(as.Date(all\_trips$date), "%A")

> all\_trips$ride\_length <- difftime(all\_trips$ended\_at,all\_trips$started\_at)

> str(all\_trips)

tibble [3,879,822 × 15] (S3: tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:3879822] "22178529" "22178530" "22178531" "22178532" ...

$ started\_at : POSIXct[1:3879822], format: "2019-04-01 00:02:22" "2019-04-01 00:03:02" "2019-04-01 00:11:07" "2019-04-01 00:13:01" ...

$ ended\_at : POSIXct[1:3879822], format: "2019-04-01 00:09:48" "2019-04-01 00:20:30" "2019-04-01 00:15:19" "2019-04-01 00:18:58" ...

$ rideable\_type : chr [1:3879822] "6251" "6226" "5649" "4151" ...

$ start\_station\_id : num [1:3879822] 81 317 283 26 202 420 503 260 211 211 ...

$ start\_station\_name: chr [1:3879822] "Daley Center Plaza" "Wood St & Taylor St" "LaSalle St & Jackson Blvd" "McClurg Ct & Illinois St" ...

$ end\_station\_id : num [1:3879822] 56 59 174 133 129 426 500 499 211 211 ...

$ end\_station\_name : chr [1:3879822] "Desplaines St & Kinzie St" "Wabash Ave & Roosevelt Rd" "Canal St & Madison St" "Kingsbury St & Kinzie St" ...

$ member\_casual : chr [1:3879822] "member" "member" "member" "member" ...

$ date : Date[1:3879822], format: "2019-04-01" "2019-04-01" "2019-04-01" "2019-04-01" ...

$ month : chr [1:3879822] "04" "04" "04" "04" ...

$ day : chr [1:3879822] "01" "01" "01" "01" ...

$ year : chr [1:3879822] "2019" "2019" "2019" "2019" ...

$ day\_of\_week : chr [1:3879822] "Monday" "Monday" "Monday" "Monday" ...

$ ride\_length : 'difftime' num [1:3879822] 446 1048 252 357 ...

..- attr(\*, "units")= chr "secs"

> is.factor(all\_trips$ride\_length)

[1] FALSE

> all\_trips$ride\_length <- as.numeric(as.character(all\_trips$ride\_length))

> is.numeric(all\_trips$ride\_length)

[1] TRUE

> all\_trips\_v2 <- all\_trips[!(all\_trips$start\_station\_name == "HQ QR" | all\_trips$ride\_length<0),]

STEP 4 : CONDUCT DESCRIPTIVE ANALYSIS

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

[1] 1479.139

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

[1] 712

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

[1] 9387024

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

[1] 1

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

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

1 412 712 1479 1289 9387024

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

all\_trips\_v2$member\_casual all\_trips\_v2$ride\_length

1 casual 3552.7502

2 member 850.0662

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

all\_trips\_v2$member\_casual all\_trips\_v2$ride\_length

1 casual 1546

2 member 589

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

all\_trips\_v2$member\_casual all\_trips\_v2$ride\_length

1 casual 9387024

2 member 9056634

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

all\_trips\_v2$member\_casual all\_trips\_v2$ride\_length

1 casual 2

2 member 1

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

all\_trips\_v2$member\_casual all\_trips\_v2$day\_of\_week all\_trips\_v2$ride\_length

1 casual Friday 3773.8351

2 member Friday 824.5305

3 casual Monday 3372.2869

4 member Monday 842.5726

5 casual Saturday 3331.9138

6 member Saturday 968.9337

7 casual Sunday 3581.4054

8 member Sunday 919.9746

9 casual Thursday 3682.9847

10 member Thursday 823.9278

11 casual Tuesday 3596.3599

12 member Tuesday 826.1427

13 casual Wednesday 3718.6619

14 member Wednesday 823.9996

> all\_trips\_v2$day\_of\_week <- ordered(all\_trips\_v2$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

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

all\_trips\_v2$member\_casual all\_trips\_v2$day\_of\_week all\_trips\_v2$ride\_length

1 casual Sunday 3581.4054

2 member Sunday 919.9746

3 casual Monday 3372.2869

4 member Monday 842.5726

5 casual Tuesday 3596.3599

6 member Tuesday 826.1427

7 casual Wednesday 3718.6619

8 member Wednesday 823.9996

9 casual Thursday 3682.9847

10 member Thursday 823.9278

11 casual Friday 3773.8351

12 member Friday 824.5305

13 casual Saturday 3331.9138

14 member Saturday 968.9337

> all\_trips\_v2 %>%

+ mutate(weekday = wday(started\_at, label = TRUE)) %>%

+ group\_by(member\_casual, weekday) %>%

+ summarise(number\_of\_rides = n()

+ ,average\_duration = mean(ride\_length)) %>%

+ arrange(member\_casual, weekday)

`summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

# A tibble: 14 × 4

# Groups: member\_casual [2]

member\_casual weekday number\_of\_rides average\_duration

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

1 casual Sun 181293 3581.

2 casual Mon 103296 3372.

3 casual Tue 90510 3596.

4 casual Wed 92457 3719.

5 casual Thu 102679 3683.

6 casual Fri 122404 3774.

7 casual Sat 209543 3332.

8 member Sun 267965 920.

9 member Mon 472196 843.

10 member Tue 508445 826.

11 member Wed 500329 824.

12 member Thu 484177 824.

13 member Fri 452790 825.

14 member Sat 287958 969.

> all\_trips\_v2 %>%

+ mutate(weekday = wday(started\_at, label = TRUE)) %>%

+ group\_by(member\_casual, weekday) %>%

+ summarise(number\_of\_rides = n()

+ ,average\_duration = mean(ride\_length)) %>%

+ arrange(member\_casual, weekday) %>%

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

+ geom\_col(position = "dodge")

`summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

> all\_trips\_v2 %>%

+ mutate(weekday = wday(started\_at, label = TRUE)) %>%

+ group\_by(member\_casual, weekday) %>%

+ summarise(number\_of\_rides = n()

+ ,average\_duration = mean(ride\_length)) %>%

+ arrange(member\_casual, weekday) %>%

+ ggplot(aes(x = weekday, y = average\_duration, fill = member\_casual)) +

+ geom\_col(position = "dodge")

`summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

STEP 5 : EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

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

> write.csv(counts, file = 'C:/Users/nicky/Downloads/google data analytics/Capstone/project/cyclistic/R/exported data')