# Case Study 1 - How Does a Bike-share navigate Speedy Success?

Serge Geukjian

9/24/2021

### Importing the data and renaming all columns

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(skimr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
q1_2018 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2018 (Full)/Divvy_Trips_2018_Q1.csv")
```

```
## Rows: 387145 Columns: 12
```

```
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): 03 - Rental Start Station Name, 02 - Rental End Station Name,
User...
## dbl  (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID, 03 -
R...
```

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

##
## 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.
```

```
q2_2018 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2018 (Full)/Divvy_Trips_2018_Q2.csv")
```

```
## Rows: 1059681 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
## 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.
```

```
q3_2018 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2018 (Full)/Divvy_Trips_2018_Q3.csv")
```

```
## Rows: 1513570 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
## 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_2018 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2018 (Full)/Divvy_Trips_2018_Q4.csv")
```

```
## Rows: 642686 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
## 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_2019 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2019 (Full)/Divvy_Trips_2019_Q1.csv")
```

```
## Rows: 365069 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
## 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.
```

```
q2_2019 <- read_csv("/Users/SG13/Desktop/Online Courses/Google Data Analytics
Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone
-  Complete a Case Study (COURSERA)/Case Study 1/Tables from the
Database/2019 (Full)/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...
## dbl  (5): 01 - Rental Details Rental ID, 01 - Rental Details Bike ID, 03 -
R...
## dttm (2): 01 - Rental Details Local Start Time, 01 - Rental Details Local
En...
##
## 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("/Users/SG13/Desktop/Online Courses/Google Data Analytics Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone -  Complete a Case Study (COURSERA)/Case Study 1/Tables from the Database/2019 (Full)/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
## 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("/Users/SG13/Desktop/Online Courses/Google Data Analytics Professional Certificate (COURSERA)/Course 8 - Google Data Analytics Capstone -  Complete a Case Study (COURSERA)/Case Study 1/Tables from the Database/2019 (Full)/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
## 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_2018 <- rename(q1_2018,
                  "BIKE_RIDE_ID" = "01 - Rental Details Rental ID",
                  "START_DATE" = "01 - Rental Details Local Start Time",
                  "END_DATE" = "01 - Rental Details Local End Time",
                  "BIKE_TYPE_ID" = "01 - Rental Details Bike ID",
                  "DURATION_sec" = "01 - Rental Details Duration In Seconds Uncapped",
                  "START_STATION_ID" = "03 - Rental Start Station ID",
                  "START_STATION_NAME" = "03 - Rental Start Station Name",
                  "END_STATION_ID" = "02 - Rental End Station ID",
                  "END_STATION_NAME" = "02 - Rental End Station Name",
                  "USER_TYPE" = "User Type",
                  "GENDER" = "Member Gender",
                  "MEMBER_BIRTHDAY" = "05 - Member Details Member Birthday
```

```
Year"
                  ))
```

```
## # A tibble: 387,145 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     17536702 2018-01-01 00:12:00 2018-01-01 00:17:23         3304
##  2     17536703 2018-01-01 00:41:35 2018-01-01 00:47:52         5367
##  3     17536704 2018-01-01 00:44:46 2018-01-01 01:33:10         4599
##  4     17536705 2018-01-01 00:53:10 2018-01-01 01:05:37         2302
##  5     17536706 2018-01-01 00:53:37 2018-01-01 00:56:40         3696
##  6     17536707 2018-01-01 00:56:15 2018-01-01 01:00:41         6298
##  7     17536708 2018-01-01 00:57:26 2018-01-01 01:02:40         1169
##  8     17536709 2018-01-01 01:00:29 2018-01-01 01:13:43         6351
##  9     17536710 2018-01-01 01:07:12 2018-01-01 01:31:53         1920
## 10     17536711 2018-01-01 01:07:54 2018-01-06 10:04:02         4783
## # … with 387,135 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q2_2018 <- rename(q2_2018,
                  "BIKE_RIDE_ID" = "trip_id",
                  "START_DATE" = "start_time",
                  "END_DATE" = "end_time",
                  "BIKE_TYPE_ID" = "bikeid",
                  "DURATION_sec" = "tripduration" ,
                  "START_STATION_ID" = "from_station_id",
                  "START_STATION_NAME" = "from_station_name",
                  "END_STATION_ID" = "to_station_id",
                  "END_STATION_NAME" = "to_station_name",
                  "USER_TYPE" = "usertype",
                  "GENDER" = "gender",
                  "MEMBER_BIRTHDAY" = "birthyear"
                  ))
```

```
## # A tibble: 1,059,681 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     18000527 2018-04-01 00:04:44 2018-04-01 00:13:03         3819
##  2     18000528 2018-04-01 00:06:42 2018-04-01 00:27:07         5000
##  3     18000529 2018-04-01 00:07:19 2018-04-01 00:23:19         5165
##  4     18000530 2018-04-01 00:07:33 2018-04-01 00:14:47         3851
##  5     18000531 2018-04-01 00:10:23 2018-04-01 00:22:12         5065
##  6     18000532 2018-04-01 00:11:29 2018-04-01 00:22:28         5962
##  7     18000533 2018-04-01 00:15:49 2018-04-01 00:19:47         4570
##  8     18000534 2018-04-01 00:17:00 2018-04-01 00:22:53         1323
##  9     18000535 2018-04-01 00:18:24 2018-04-01 00:23:06         1977
## 10     18000536 2018-04-01 00:20:00 2018-04-01 00:26:22         2602
```

```
## # … with 1,059,671 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q3_2018 <- rename(q3_2018,
                  "BIKE_RIDE_ID" = "trip_id",
                  "START_DATE" = "start_time",
                  "END_DATE" = "end_time",
                  "BIKE_TYPE_ID" = "bikeid",
                  "DURATION_sec" = "tripduration" ,
                  "START_STATION_ID" = "from_station_id",
                  "START_STATION_NAME" = "from_station_name",
                  "END_STATION_ID" = "to_station_id",
                  "END_STATION_NAME" = "to_station_name",
                  "USER_TYPE" = "usertype",
                  "GENDER" = "gender",
                  "MEMBER_BIRTHDAY" = "birthyear"
                  ))
```

```
## # A tibble: 1,513,570 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     19244622 2018-07-01 00:00:03 2018-07-01 23:56:11         5429
##  2     19244623 2018-07-01 00:00:13 2018-07-01 00:06:39           93
##  3     19244624 2018-07-01 00:00:15 2018-07-01 00:23:26         2461
##  4     19244625 2018-07-01 00:00:25 2018-07-01 00:23:31         2991
##  5     19244626 2018-07-01 00:00:27 2018-07-01 00:11:23         2851
##  6     19244627 2018-07-01 00:00:35 2018-07-01 00:16:09         5980
##  7     19244628 2018-07-01 00:00:37 2018-07-01 00:10:14         3132
##  8     19244629 2018-07-01 00:00:55 2018-07-01 00:09:20         2281
##  9     19244630 2018-07-01 00:01:38 2018-07-01 00:25:25         3465
## 10     19244631 2018-07-01 00:01:44 2018-07-01 00:25:25         3873
## # … with 1,513,560 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q4_2018 <- rename(q4_2018,
                  "BIKE_RIDE_ID" = "trip_id",
                  "START_DATE" = "start_time",
                  "END_DATE" = "end_time",
                  "BIKE_TYPE_ID" = "bikeid",
                  "DURATION_sec" = "tripduration" ,
                  "START_STATION_ID" = "from_station_id",
                  "START_STATION_NAME" = "from_station_name",
                  "END_STATION_ID" = "to_station_id",
                  "END_STATION_NAME" = "to_station_name",
```

```
                "USER_TYPE" = "usertype",
                "GENDER" = "gender",
                "MEMBER_BIRTHDAY" = "birthyear"
                ))
```

```
## # A tibble: 642,686 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     20983530 2018-10-01 00:01:17 2018-10-01 00:29:35         4551
##  2     20983531 2018-10-01 00:03:59 2018-10-01 00:10:55          847
##  3     20983532 2018-10-01 00:05:14 2018-10-01 00:14:08         6188
##  4     20983533 2018-10-01 00:05:48 2018-10-01 00:18:46         6372
##  5     20983534 2018-10-01 00:07:29 2018-10-01 00:25:51         1927
##  6     20983535 2018-10-01 00:07:36 2018-10-01 00:11:25         2392
##  7     20983536 2018-10-01 00:08:09 2018-10-01 00:58:48          308
##  8     20983537 2018-10-01 00:09:29 2018-10-01 00:15:23         1187
##  9     20983538 2018-10-01 00:09:33 2018-10-01 00:12:27         6247
## 10     20983539 2018-10-01 00:09:44 2018-10-01 00:21:06         3083
## # … with 642,676 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q1_2019 <- rename(q1_2019,
                "BIKE_RIDE_ID" = "trip_id",
                "START_DATE" = "start_time",
                "END_DATE" = "end_time",
                "BIKE_TYPE_ID" = "bikeid",
                "DURATION_sec" = "tripduration" ,
                "START_STATION_ID" = "from_station_id",
                "START_STATION_NAME" = "from_station_name",
                "END_STATION_ID" = "to_station_id",
                "END_STATION_NAME" = "to_station_name",
                "USER_TYPE" = "usertype",
                "GENDER" = "gender",
                "MEMBER_BIRTHDAY" = "birthyear"
                ))
```

```
## # A tibble: 365,069 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     21742443 2019-01-01 00:04:37 2019-01-01 00:11:07         2167
##  2     21742444 2019-01-01 00:08:13 2019-01-01 00:15:34         4386
##  3     21742445 2019-01-01 00:13:23 2019-01-01 00:27:12         1524
##  4     21742446 2019-01-01 00:13:45 2019-01-01 00:43:28          252
##  5     21742447 2019-01-01 00:14:52 2019-01-01 00:20:56         1170
##  6     21742448 2019-01-01 00:15:33 2019-01-01 00:19:09         2437
##  7     21742449 2019-01-01 00:16:06 2019-01-01 00:19:03         2708
##  8     21742450 2019-01-01 00:18:41 2019-01-01 00:20:21         2796
```

```
##  9     21742451 2019-01-01 00:18:43 2019-01-01 00:47:30         6205
## 10     21742452 2019-01-01 00:19:18 2019-01-01 00:24:54         3939
## # … with 365,059 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q2_2019 <- rename(q2_2019,
                  "BIKE_RIDE_ID" = "01 - Rental Details Rental ID",
                  "START_DATE" = "01 - Rental Details Local Start Time",
                  "END_DATE" = "01 - Rental Details Local End Time",
                  "BIKE_TYPE_ID" = "01 - Rental Details Bike ID",
                  "DURATION_sec" = "01 - Rental Details Duration In Seconds
Uncapped",
                  "START_STATION_ID" = "03 - Rental Start Station ID",
                  "START_STATION_NAME" = "03 - Rental Start Station Name",
                  "END_STATION_ID" = "02 - Rental End Station ID",
                  "END_STATION_NAME" = "02 - Rental End Station Name",
                  "USER_TYPE" = "User Type",
                  "GENDER" = "Member Gender",
                  "MEMBER_BIRTHDAY" = "05 - Member Details Member Birthday
Year"
                  ))
```

```
## # A tibble: 1,108,163 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     22178529 2019-04-01 00:02:22 2019-04-01 00:09:48         6251
##  2     22178530 2019-04-01 00:03:02 2019-04-01 00:20:30         6226
##  3     22178531 2019-04-01 00:11:07 2019-04-01 00:15:19         5649
##  4     22178532 2019-04-01 00:13:01 2019-04-01 00:18:58         4151
##  5     22178533 2019-04-01 00:19:26 2019-04-01 00:36:13         3270
##  6     22178534 2019-04-01 00:19:39 2019-04-01 00:23:56         3123
##  7     22178535 2019-04-01 00:26:33 2019-04-01 00:35:41         6418
##  8     22178536 2019-04-01 00:29:48 2019-04-01 00:36:11         4513
##  9     22178537 2019-04-01 00:32:07 2019-04-01 01:07:44         3280
## 10     22178538 2019-04-01 00:32:19 2019-04-01 01:07:39         5534
## # … with 1,108,153 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q3_2019 <- rename(q3_2019,
                  "BIKE_RIDE_ID" = "trip_id",
                  "START_DATE" = "start_time",
                  "END_DATE" = "end_time",
                  "BIKE_TYPE_ID" = "bikeid",
                  "DURATION_sec" = "tripduration" ,
```

```
                  "START_STATION_ID" = "from_station_id",
                  "START_STATION_NAME" = "from_station_name",
                  "END_STATION_ID" = "to_station_id",
                  "END_STATION_NAME" = "to_station_name",
                  "USER_TYPE" = "usertype",
                  "GENDER" = "gender",
                  "MEMBER_BIRTHDAY" = "birthyear"
                  ))
```

```
## # A tibble: 1,640,718 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     23479388 2019-07-01 00:00:27 2019-07-01 00:20:41         3591
##  2     23479389 2019-07-01 00:01:16 2019-07-01 00:18:44         5353
##  3     23479390 2019-07-01 00:01:48 2019-07-01 00:27:42         6180
##  4     23479391 2019-07-01 00:02:07 2019-07-01 00:27:10         5540
##  5     23479392 2019-07-01 00:02:13 2019-07-01 00:22:26         6014
##  6     23479393 2019-07-01 00:02:21 2019-07-01 00:07:31         4941
##  7     23479394 2019-07-01 00:02:24 2019-07-01 00:23:12         3770
##  8     23479395 2019-07-01 00:02:26 2019-07-01 00:28:16         5442
##  9     23479396 2019-07-01 00:02:34 2019-07-01 00:28:57         2957
## 10     23479397 2019-07-01 00:02:45 2019-07-01 00:29:14         6091
## # … with 1,640,708 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

```
(q4_2019 <- rename(q4_2019,
                  "BIKE_RIDE_ID" = "trip_id",
                  "START_DATE" = "start_time",
                  "END_DATE" = "end_time",
                  "BIKE_TYPE_ID" = "bikeid",
                  "DURATION_sec" = "tripduration" ,
                  "START_STATION_ID" = "from_station_id",
                  "START_STATION_NAME" = "from_station_name",
                  "END_STATION_ID" = "to_station_id",
                  "END_STATION_NAME" = "to_station_name",
                  "USER_TYPE" = "usertype",
                  "GENDER" = "gender",
                  "MEMBER_BIRTHDAY" = "birthyear"
                  ))
```

```
## # A tibble: 704,054 × 12
##    BIKE_RIDE_ID START_DATE          END_DATE            BIKE_TYPE_ID
##           <dbl> <dttm>              <dttm>                     <dbl>
##  1     25223640 2019-10-01 00:01:39 2019-10-01 00:17:20         2215
##  2     25223641 2019-10-01 00:02:16 2019-10-01 00:06:34         6328
##  3     25223642 2019-10-01 00:04:32 2019-10-01 00:18:43         3003
##  4     25223643 2019-10-01 00:04:32 2019-10-01 00:43:43         3275
```

```
##  5     25223644 2019-10-01 00:04:34 2019-10-01 00:35:42         5294
##  6     25223645 2019-10-01 00:04:38 2019-10-01 00:10:51         1891
##  7     25223646 2019-10-01 00:04:52 2019-10-01 00:22:45         1061
##  8     25223647 2019-10-01 00:04:57 2019-10-01 00:29:16         1274
##  9     25223648 2019-10-01 00:05:20 2019-10-01 00:29:18         6011
## 10     25223649 2019-10-01 00:05:20 2019-10-01 02:23:46         2957
## # … with 704,044 more rows, and 8 more variables: DURATION_sec <dbl>,
## #   START_STATION_ID <dbl>, START_STATION_NAME <chr>, END_STATION_ID
<dbl>,
## #   END_STATION_NAME <chr>, USER_TYPE <chr>, GENDER <chr>,
## #   MEMBER_BIRTHDAY <dbl>
```

### Combining all tables into one master table

```
MASTER_DATA <-
bind_rows(q1_2018,q2_2018,q3_2018,q4_2018,q1_2019,q2_2019,q3_2019,q4_2019)
glimpse(MASTER_DATA)
```

```
## Rows: 7,421,086
## Columns: 12
## $ BIKE_RIDE_ID       <dbl> 17536702, 17536703, 17536704, 17536705,
17536706, 1…
## $ START_DATE         <dttm> 2018-01-01 00:12:00, 2018-01-01 00:41:35,
2018-01-…
## $ END_DATE           <dttm> 2018-01-01 00:17:23, 2018-01-01 00:47:52,
2018-01-…
## $ BIKE_TYPE_ID       <dbl> 3304, 5367, 4599, 2302, 3696, 6298, 1169, 6351,
192…
## $ DURATION_sec       <dbl> 323, 377, 2904, 747, 183, 266, 314, 794, 1481,
4641…
## $ START_STATION_ID   <dbl> 69, 253, 98, 125, 129, 304, 164, 182, 99, 99,
99, 1…
## $ START_STATION_NAME <chr> "Damen Ave & Pierce Ave", "Winthrop Ave &
Lawrence …
## $ END_STATION_ID     <dbl> 159, 325, 509, 364, 205, 299, 174, 142, 99, 99,
99,…
## $ END_STATION_NAME   <chr> "Claremont Ave & Hirsch St", "Clark St &
Winnemac A…
## $ USER_TYPE          <chr> "Subscriber", "Subscriber", "Subscriber",
"Subscrib…
## $ GENDER             <chr> "Male", "Male", "Male", "Male", "Male",
"Female", "…
## $ MEMBER_BIRTHDAY    <dbl> 1988, 1984, 1989, 1983, 1989, 1994, 1998, 1990,
NA,…
```

```
skim_without_charts(MASTER_DATA)
```

*Data summary*

| | |
|---|---|
| Name | MASTER_DATA |
| Number of rows | 7421086 |

| | |
|---|---|
| Number of columns | 12 |

---

Column type frequency:

| | |
|---|---|
| character | 4 |
| numeric | 6 |
| POSIXct | 2 |

---

| | |
|---|---|
| Group variables | None |

**Variable type: character**

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| START_STATION_NAME | 0 | 1.00 | 10 | 43 | 0 | 664 | 0 |
| END_STATION_NAME | 0 | 1.00 | 10 | 43 | 0 | 664 | 0 |
| USER_TYPE | 0 | 1.00 | 8 | 10 | 0 | 2 | 0 |
| GENDER | 1121711 | 0.85 | 4 | 6 | 0 | 2 | 0 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|
| BIKE_RIDE_ID | 0 | 1.00 | 21843132.28 | 2452330.54 | 1753 6702 | 1971 5491 | 2186 9996 | 2401 9004 | 2596 2904 |
| BIKE_TYPE_ID | 0 | 1.00 | 3430.34 | 1918.58 | 1 | 1753 | 3523 | 5124 | 6946 |
| DURATION_sec | 0 | 1.00 | 1432.22 | 32953.47 | 61 | 403 | 691 | 1247 | 1433 6400 |
| START_STATION_ID | 0 | 1.00 | 195.73 | 148.86 | 1 | 77 | 172 | 287 | 673 |
| END_STATION_ID | 0 | 1.00 | 196.55 | 148.93 | 1 | 77 | 172 | 287 | 673 |
| MEMBER_BIRTHDAY | 1093 960 | 0.85 | 1983.41 | 10.92 | 1759 | 1978 | 1987 | 1991 | 2014 |

**Variable type: POSIXct**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| START_DATE | 0 | 1 | 2018-01-01 | 2019-12-31 | 2019-02-03 | 6454879 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | 00:12:00 | 23:57:17 | 10:51:44 | |
| END_DATE | 0 | 1 | 2018-01-01 00:17:23 | 2020-01-21 13:54:35 | 2019-02-03 11:08:58 | 6287326 |

### Data manipulation and additional calculations

```
MASTER_DATA <- MASTER_DATA %>%
  mutate("DURATION_min" = DURATION_sec / 60)%>%
  mutate("QUARTER" = quarter(START_DATE))%>%
  mutate("YEAR" = year(START_DATE))%>%
  mutate("MONTH" = month(START_DATE))

MASTER_DATA$GENDER[is.na(MASTER_DATA$GENDER)] = "undefined"
MASTER_DATA$MEMBER_BIRTHDAY[is.na(MASTER_DATA$MEMBER_BIRTHDAY)] = 0

MASTER_DATA <- MASTER_DATA %>%
  filter(MASTER_DATA$MEMBER_BIRTHDAY > 1919 & MASTER_DATA$MEMBER_BIRTHDAY !=
0)

MASTER_DATA$WEEKDAY <- format(as.Date(MASTER_DATA$START_DATE),"%A")

aggregate(MASTER_DATA$DURATION_min ~ MASTER_DATA$USER_TYPE,FUN = mean)

##   MASTER_DATA$USER_TYPE MASTER_DATA$DURATION_min
## 1              Customer                 48.64189
## 2            Subscriber                 14.41251

aggregate(MASTER_DATA$DURATION_min ~ MASTER_DATA$USER_TYPE,FUN = median)

##   MASTER_DATA$USER_TYPE MASTER_DATA$DURATION_min
## 1              Customer                23.616667
## 2            Subscriber                 9.666667

aggregate(MASTER_DATA$DURATION_min ~ MASTER_DATA$USER_TYPE,FUN = max)

##   MASTER_DATA$USER_TYPE MASTER_DATA$DURATION_min
## 1              Customer                 132324.1
## 2            Subscriber                 225960.0

aggregate(MASTER_DATA$DURATION_min ~ MASTER_DATA$USER_TYPE,FUN = min)

##   MASTER_DATA$USER_TYPE MASTER_DATA$DURATION_min
## 1              Customer                 1.016667
## 2            Subscriber                 1.016667

MASTER_DATA$WEEKDAY <- ordered(MASTER_DATA$WEEKDAY,
levels=c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunda
y"))
```

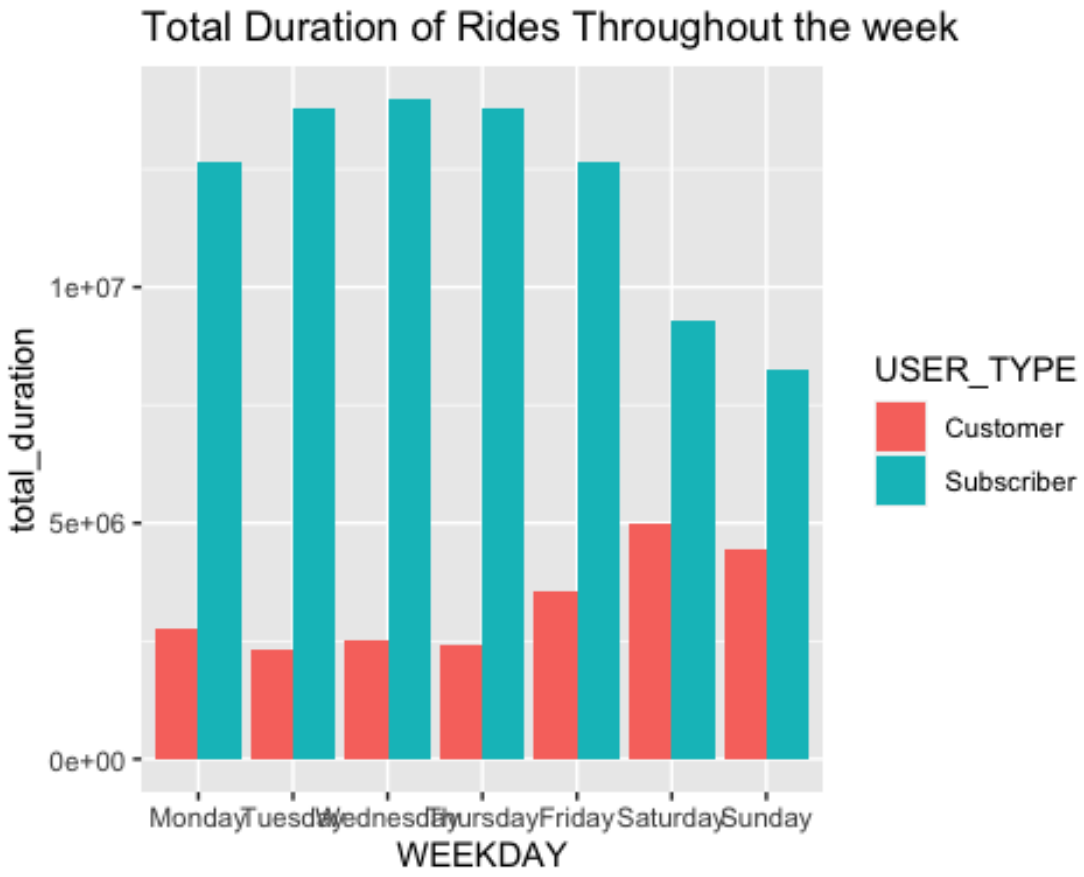
```
aggregate(MASTER_DATA$DURATION_min ~ MASTER_DATA$USER_TYPE +
MASTER_DATA$WEEKDAY, FUN = mean)
```

```
##    MASTER_DATA$USER_TYPE MASTER_DATA$WEEKDAY MASTER_DATA$DURATION_min
## 1               Customer              Monday                 49.05163
## 2             Subscriber              Monday                 13.95018
## 3               Customer             Tuesday                 47.15669
## 4             Subscriber             Tuesday                 13.92186
## 5               Customer           Wednesday                 49.19079
## 6             Subscriber           Wednesday                 14.02002
## 7               Customer            Thursday                 44.22185
## 8             Subscriber            Thursday                 14.11735
## 9               Customer              Friday                 56.28054
## 10            Subscriber              Friday                 14.05666
## 11              Customer            Saturday                 45.95604
## 12            Subscriber            Saturday                 16.34197
## 13              Customer              Sunday                 49.48059
## 14            Subscriber              Sunday                 15.97212
```

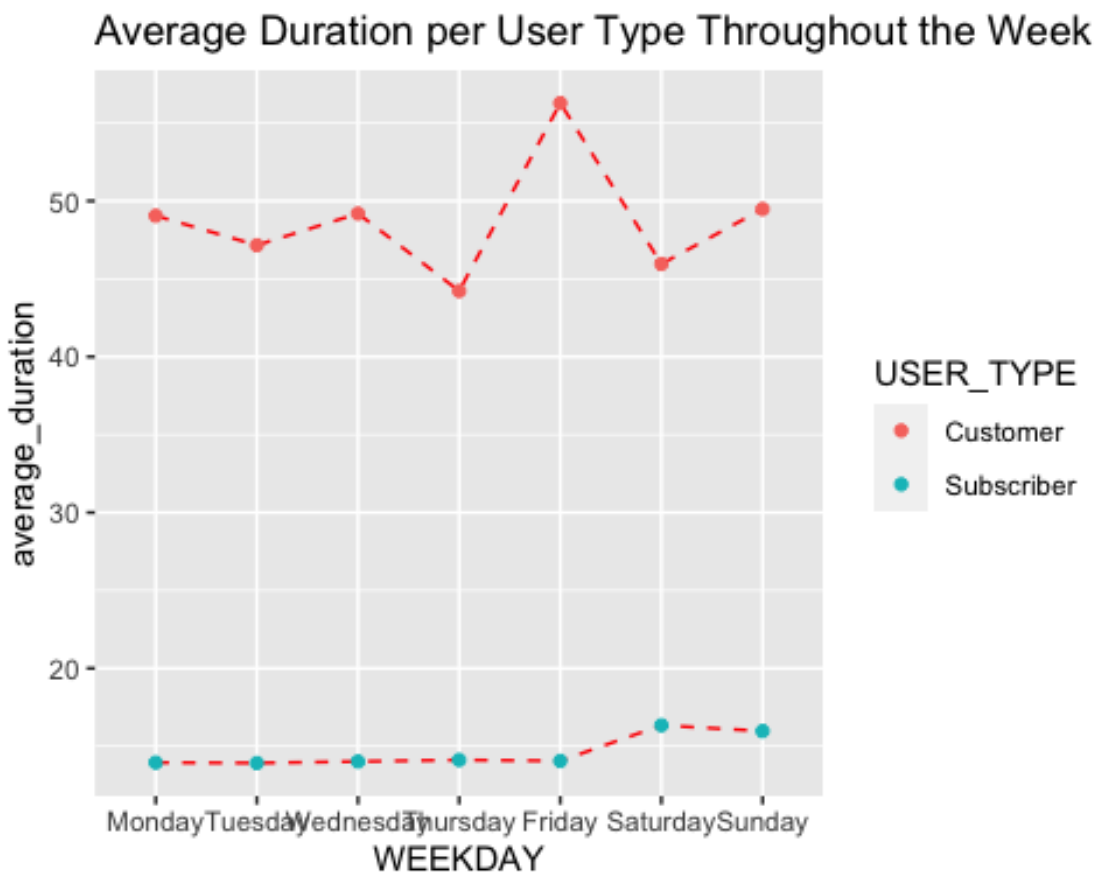
### Analysis Visualizations

```
MASTER_DATA %>%
  group_by(USER_TYPE,WEEKDAY)%>%
  summarise(number_of_rides = n(),average_duration =
mean(DURATION_min),max_duration = max(DURATION_min),total_duration =
sum(DURATION_min))%>%
  arrange(USER_TYPE,WEEKDAY) %>%
  ggplot(aes(x=WEEKDAY,y=total_duration,fill=USER_TYPE)) + geom_col(position
= "dodge") + labs(title = "Total Duration of Rides Throughout the week")
```

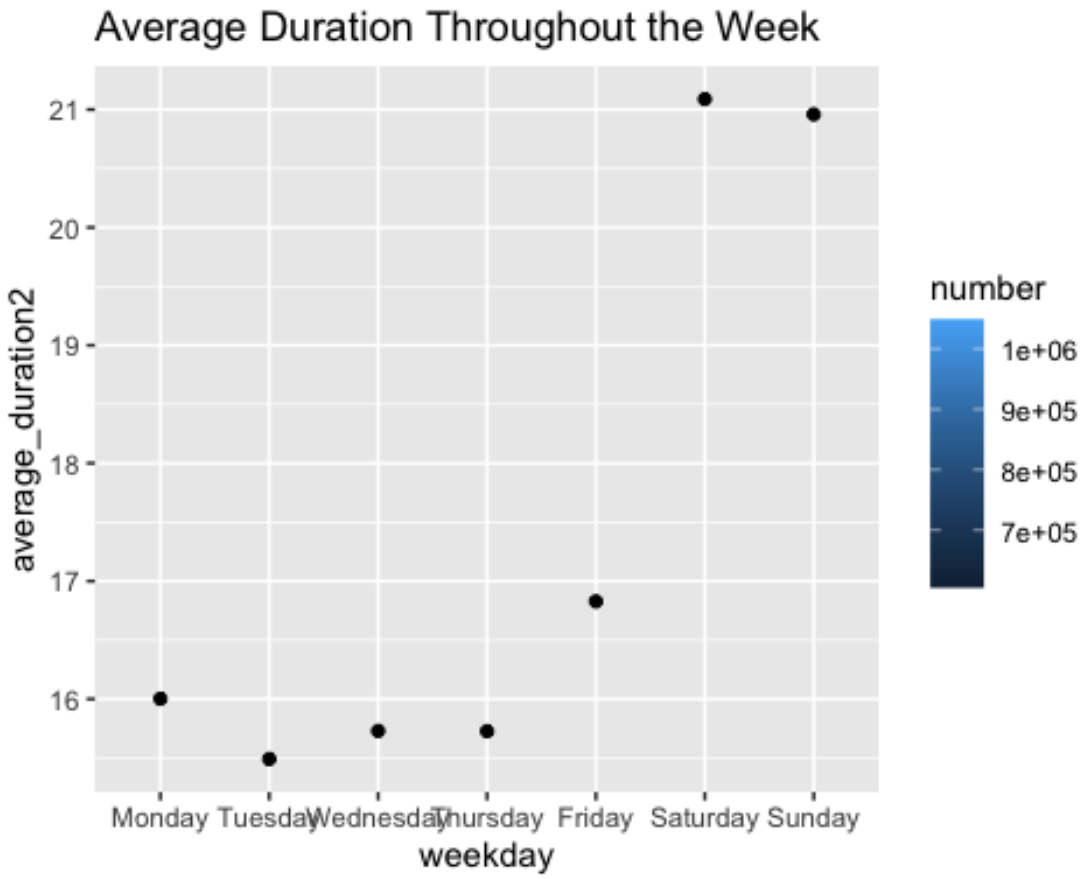
```
## `summarise()` has grouped output by 'USER_TYPE'. You can override using
the `.groups` argument.
```



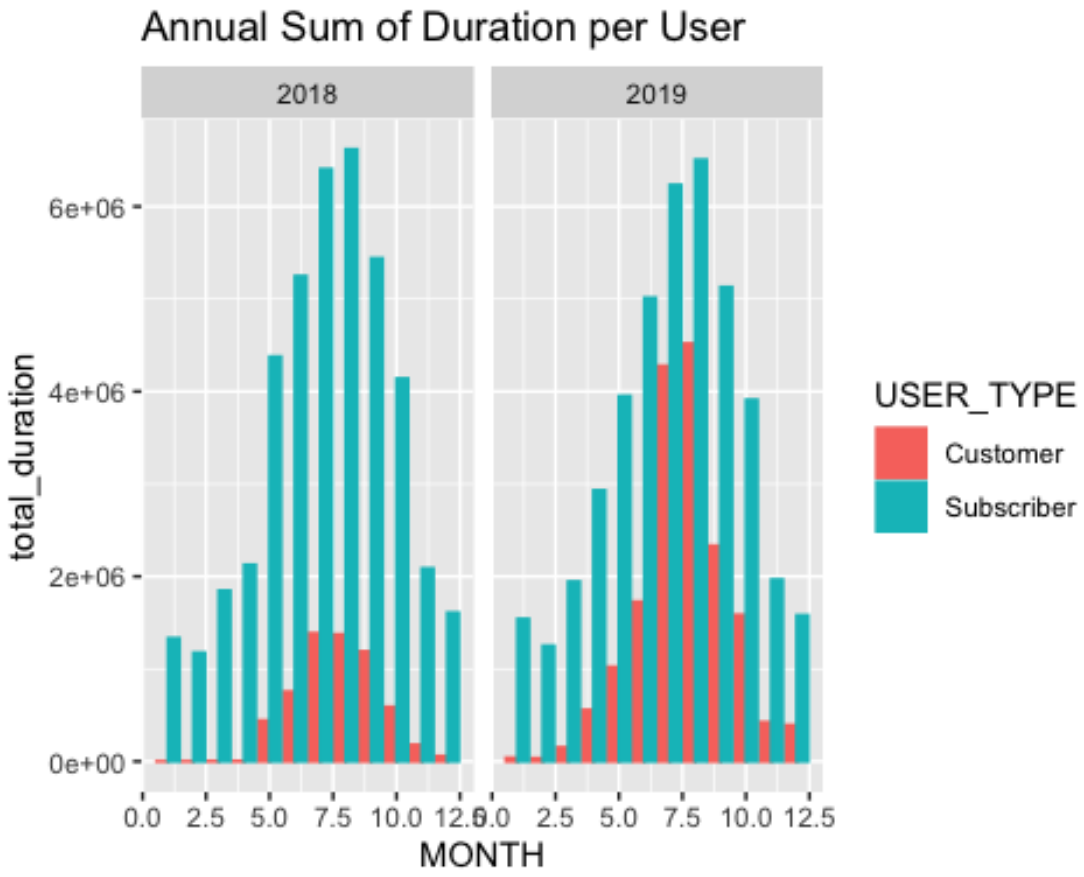

```
MASTER_DATA %>%
  group_by(USER_TYPE,WEEKDAY)%>%
  summarise(number_of_rides = n(),average_duration =
mean(DURATION_min),max_duration = max(DURATION_min),total_duration =
sum(DURATION_min))%>%
  arrange(USER_TYPE,WEEKDAY) %>%

ggplot(aes(x=WEEKDAY,y=average_duration,fill=USER_TYPE,group=USER_TYPE,color=
USER_TYPE)) + geom_line(linetype = "dashed", color = "red") + geom_point() +
labs(title = "Average Duration per User Type Throughout the Week")

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




```
df <- data.frame(weekday = MASTER_DATA$WEEKDAY,duration =
MASTER_DATA$DURATION_min)

df %>%
  group_by(weekday)%>%
  summarise(average_duration2 = mean(duration),number = n())%>%
  ggplot(aes(x=weekday,y=average_duration2, fill=number,group=number)) +
geom_point() + labs(title = "Average Duration Throughout the Week")
```




```
MASTER_DATA %>%
  group_by(USER_TYPE,QUARTER,YEAR,MONTH)%>%
  summarise(number_of_rides = n(),total_duration = sum(DURATION_min))%>%
  arrange(USER_TYPE,QUARTER,YEAR,MONTH) %>%

ggplot(aes(x=MONTH,y=total_duration,fill=USER_TYPE,group=USER_TYPE,color=USER
_TYPE)) + geom_col(position = "dodge")+ facet_wrap(~YEAR) + labs(title =
"Annual Sum of Duration per User")

## `summarise()` has grouped output by 'USER_TYPE', 'QUARTER', 'YEAR'. You
can override using the `.groups` argument.
```


Annual Sum of Duration per User
2018
2019
6e+06
4e+06
2e+06
0e+00
total_duration
0.0
2.5
5.0
7.5
10.0
12.5
MONTH
USER_TYPE
Customer
Subscriber