Divvy Trip Data Analysis

Setup and Data Loading

Display the first 6 rows of the data frame (default behavior of head())

```
##    trip_id          start_time            end_time bikeid tripduration
## 1 21742443 2019-01-01 00:04:37 2019-01-01 00:11:07   2167        390.0
## 2 21742444 2019-01-01 00:08:13 2019-01-01 00:15:34   4386        441.0
## 3 21742445 2019-01-01 00:13:23 2019-01-01 00:27:12   1524        829.0
## 4 21742446 2019-01-01 00:13:45 2019-01-01 00:43:28    252      1,783.0
## 5 21742447 2019-01-01 00:14:52 2019-01-01 00:20:56   1170        364.0
## 6 21742448 2019-01-01 00:15:33 2019-01-01 00:19:09   2437        216.0
##   from_station_id                  from_station_name to_station_id
## 1             199              Wabash Ave & Grand Ave            84
## 2              44              State St & Randolph St           624
## 3              15                Racine Ave & 18th St           644
## 4             123      California Ave & Milwaukee Ave           176
## 5             173 Mies van der Rohe Way & Chicago Ave            35
## 6              98          LaSalle St & Washington St            49
##                   to_station_name   usertype gender birthyear
## 1      Milwaukee Ave & Grand Ave Subscriber   Male      1989
## 2 Dearborn St & Van Buren St (*) Subscriber Female      1990
## 3  Western Ave & Fillmore St (*) Subscriber Female      1994
## 4              Clark St & Elm St Subscriber   Male      1993
## 5        Streeter Dr & Grand Ave Subscriber   Male      1994
## 6        Dearborn St & Monroe St Subscriber Female      1983
```

You can also specify the number of rows you want to see, e.g., the first 10 rows:

print(head(divvy_data, n = 10))

*** Now I want to print the columns names of the data frame. ***

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

*** Now I want to print the structure of the data frame. ***

```
## 'data.frame':    365069 obs. of  12 variables:
##  $ trip_id          : int  21742443 21742444 21742445 21742446 21742447 21742448 21742449 21742450 2
##  $ start_time       : chr  "2019-01-01 00:04:37" "2019-01-01 00:08:13" "2019-01-01 00:13:23" "2019-0
##  $ end_time         : chr  "2019-01-01 00:11:07" "2019-01-01 00:15:34" "2019-01-01 00:27:12" "2019-0
##  $ bikeid           : int  2167 4386 1524 252 1170 2437 2708 2796 6205 3939 ...
```

```
##  $ tripduration     : chr  "390.0" "441.0" "829.0" "1,783.0" ...
##  $ from_station_id  : int  199 44 15 123 173 98 98 211 150 268 ...
##  $ from_station_name: chr  "Wabash Ave & Grand Ave" "State St & Randolph St" "Racine Ave & 18th St" "
##  $ to_station_id    : int  84 624 644 176 35 49 49 142 148 141 ...
##  $ to_station_name  : chr  "Milwaukee Ave & Grand Ave" "Dearborn St & Van Buren St (*)" "Western Ave
##  $ usertype         : chr  "Subscriber" "Subscriber" "Subscriber" "Subscriber" ...
##  $ gender           : chr  "Male" "Female" "Female" "Male" ...
##  $ birthyear        : int  1989 1990 1994 1993 1994 1983 1984 1990 1995 1996 ...
```

*** Now I want to print the summary of the data frame. ***

```
##     trip_id          start_time          end_time             bikeid
##  Min.   :21742443   Length:365069      Length:365069      Min.   :   1
##  1st Qu.:21848765   Class :character   Class :character   1st Qu.:1777
##  Median :21961829   Mode  :character   Mode  :character   Median :3489
##  Mean   :21960872                                         Mean   :3429
##  3rd Qu.:22071823                                         3rd Qu.:5157
##  Max.   :22178528                                         Max.   :6471
##
##  tripduration       from_station_id from_station_name  to_station_id
##  Length:365069      Min.   :  2.0   Length:365069      Min.   :  2.0
##  Class :character   1st Qu.: 76.0   Class :character   1st Qu.: 76.0
##  Mode  :character   Median :170.0   Mode  :character   Median :168.0
##                     Mean   :198.1                      Mean   :198.6
##                     3rd Qu.:287.0                      3rd Qu.:287.0
##                     Max.   :665.0                      Max.   :665.0
##
##  to_station_name      usertype            gender            birthyear
##  Length:365069      Length:365069      Length:365069      Min.   :1900
##  Class :character   Class :character   Class :character   1st Qu.:1975
##  Mode  :character   Mode  :character   Mode  :character   Median :1985
##                                                           Mean   :1982
##                                                           3rd Qu.:1990
##                                                           Max.   :2003
##                                                           NA's   :18023
```

*** Now I need to check the number of usertypes in the data frame. ***

```
## [1] "Subscriber" "Customer"
```

*** I need to see the number of unique Subscribers vs Customers in the data frame. ***

```
##
##   Customer Subscriber
##      23163     341906
```

I need to load the csv file named "Divvy_Trips_2019_Q2.csv" and 2019_Q3 2019_Q4 into R and print the first 6 rows of the data frame.

```
##   X01...Rental.Details.Rental.ID X01...Rental.Details.Local.Start.Time
## 1                       22178529                    2019-04-01 00:02:22
## 2                       22178530                    2019-04-01 00:03:02
## 3                       22178531                    2019-04-01 00:11:07
## 4                       22178532                    2019-04-01 00:13:01
## 5                       22178533                    2019-04-01 00:19:26
## 6                       22178534                    2019-04-01 00:19:39
##   X01...Rental.Details.Local.End.Time X01...Rental.Details.Bike.ID
## 1                 2019-04-01 00:09:48                         6251
```

```
## 2                 2019-04-01 00:20:30                          6226
## 3                 2019-04-01 00:15:19                          5649
## 4                 2019-04-01 00:18:58                          4151
## 5                 2019-04-01 00:36:13                          3270
## 6                 2019-04-01 00:23:56                          3123
##   X01...Rental.Details.Duration.In.Seconds.Uncapped
## 1                                             446.0
## 2                                           1,048.0
## 3                                             252.0
## 4                                             357.0
## 5                                           1,007.0
## 6                                             257.0
##   X03...Rental.Start.Station.ID X03...Rental.Start.Station.Name
## 1                            81              Daley Center Plaza
## 2                           317             Wood St & Taylor St
## 3                           283       LaSalle St & Jackson Blvd
## 4                            26        McClurg Ct & Illinois St
## 5                           202            Halsted St & 18th St
## 6                           420             Ellis Ave & 55th St
##   X02...Rental.End.Station.ID X02...Rental.End.Station.Name  User.Type
## 1                          56     Desplaines St & Kinzie St Subscriber
## 2                          59     Wabash Ave & Roosevelt Rd Subscriber
## 3                         174         Canal St & Madison St Subscriber
## 4                         133      Kingsbury St & Kinzie St Subscriber
## 5                         129     Blue Island Ave & 18th St Subscriber
## 6                         426           Ellis Ave & 60th St Subscriber
##   Member.Gender X05...Member.Details.Member.Birthday.Year
## 1          Male                                      1975
## 2        Female                                      1984
## 3          Male                                      1990
## 4          Male                                      1993
## 5          Male                                      1992
## 6          Male                                      1999
```

```
##    trip_id          start_time            end_time bikeid tripduration
## 1 23479388 2019-07-01 00:00:27 2019-07-01 00:20:41   3591      1,214.0
## 2 23479389 2019-07-01 00:01:16 2019-07-01 00:18:44   5353      1,048.0
## 3 23479390 2019-07-01 00:01:48 2019-07-01 00:27:42   6180      1,554.0
## 4 23479391 2019-07-01 00:02:07 2019-07-01 00:27:10   5540      1,503.0
## 5 23479392 2019-07-01 00:02:13 2019-07-01 00:22:26   6014      1,213.0
## 6 23479393 2019-07-01 00:02:21 2019-07-01 00:07:31   4941        310.0
##   from_station_id             from_station_name to_station_id
## 1             117      Wilton Ave & Belmont Ave           497
## 2             381       Western Ave & Monroe St           203
## 3             313 Lakeview Ave & Fullerton Pkwy           144
## 4             313 Lakeview Ave & Fullerton Pkwy           144
## 5             168        Michigan Ave & 14th St            62
## 6             300          Broadway & Barry Ave           232
##                 to_station_name   usertype gender birthyear
## 1    Kimball Ave & Belmont Ave Subscriber   Male      1992
## 2        Western Ave & 21st St   Customer               NA
## 3    Larrabee St & Webster Ave   Customer               NA
## 4    Larrabee St & Webster Ave   Customer               NA
## 5               McCormick Place   Customer               NA
```

```
## 6 Pine Grove Ave & Waveland Ave Subscriber   Male      1990
```

```
##    trip_id          start_time            end_time bikeid tripduration
## 1 25223640 2019-10-01 00:01:39 2019-10-01 00:17:20   2215        940.0
## 2 25223641 2019-10-01 00:02:16 2019-10-01 00:06:34   6328        258.0
## 3 25223642 2019-10-01 00:04:32 2019-10-01 00:18:43   3003        850.0
## 4 25223643 2019-10-01 00:04:32 2019-10-01 00:43:43   3275      2,350.0
## 5 25223644 2019-10-01 00:04:34 2019-10-01 00:35:42   5294      1,867.0
## 6 25223645 2019-10-01 00:04:38 2019-10-01 00:10:51   1891        373.0
##   from_station_id              from_station_name to_station_id
## 1              20   Sheffield Ave & Kingsbury St           309
## 2              19 Throop (Loomis) St & Taylor St           241
## 3              84      Milwaukee Ave & Grand Ave           199
## 4             313  Lakeview Ave & Fullerton Pkwy           290
## 5             210      Ashland Ave & Division St           382
## 6             156      Clark St & Wellington Ave           226
##               to_station_name   usertype gender birthyear
## 1   Leavitt St & Armitage Ave Subscriber   Male      1987
## 2         Morgan St & Polk St Subscriber   Male      1998
## 3      Wabash Ave & Grand Ave Subscriber Female      1991
## 4      Kedzie Ave & Palmer Ct Subscriber   Male      1990
## 5 Western Ave & Congress Pkwy Subscriber   Male      1987
## 6    Racine Ave & Belmont Ave Subscriber Female      1994
```

*** adding a collumn for quarter to all four datasets indicating the quarter ***

*** Check the structure difference between the four dataframes ***

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
## [13] "quarter"
```

```
##  [1] "X01...Rental.Details.Rental.ID"
##  [2] "X01...Rental.Details.Local.Start.Time"
##  [3] "X01...Rental.Details.Local.End.Time"
##  [4] "X01...Rental.Details.Bike.ID"
##  [5] "X01...Rental.Details.Duration.In.Seconds.Uncapped"
##  [6] "X03...Rental.Start.Station.ID"
##  [7] "X03...Rental.Start.Station.Name"
##  [8] "X02...Rental.End.Station.ID"
##  [9] "X02...Rental.End.Station.Name"
## [10] "User.Type"
## [11] "Member.Gender"
## [12] "X05...Member.Details.Member.Birthday.Year"
## [13] "quarter"
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
## [13] "quarter"
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
```

```
## [10] "usertype"          "gender"            "birthyear"
## [13] "quarter"
```

*** Renaming the columns of Q2 data frame to match the other dataframes ***

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
## [13] "quarter"
```

*** Now combine all four quarters of Divvy trip data into a single data frame again. ***

```
##    trip_id          start_time            end_time bikeid tripduration
## 1 21742443 2019-01-01 00:04:37 2019-01-01 00:11:07   2167        390.0
## 2 21742444 2019-01-01 00:08:13 2019-01-01 00:15:34   4386        441.0
## 3 21742445 2019-01-01 00:13:23 2019-01-01 00:27:12   1524        829.0
## 4 21742446 2019-01-01 00:13:45 2019-01-01 00:43:28    252      1,783.0
## 5 21742447 2019-01-01 00:14:52 2019-01-01 00:20:56   1170        364.0
## 6 21742448 2019-01-01 00:15:33 2019-01-01 00:19:09   2437        216.0
##   from_station_id                  from_station_name to_station_id
## 1             199              Wabash Ave & Grand Ave            84
## 2              44              State St & Randolph St           624
## 3              15                Racine Ave & 18th St           644
## 4             123      California Ave & Milwaukee Ave           176
## 5             173 Mies van der Rohe Way & Chicago Ave            35
## 6              98          LaSalle St & Washington St            49
##                   to_station_name   usertype gender birthyear quarter
## 1      Milwaukee Ave & Grand Ave Subscriber   Male      1989      Q1
## 2 Dearborn St & Van Buren St (*) Subscriber Female      1990      Q1
## 3  Western Ave & Fillmore St (*) Subscriber Female      1994      Q1
## 4              Clark St & Elm St Subscriber   Male      1993      Q1
## 5        Streeter Dr & Grand Ave Subscriber   Male      1994      Q1
## 6        Dearborn St & Monroe St Subscriber Female      1983      Q1
```

*** saving the new dataframe in a csv file ***

*** Using janitor package to clean the column names of the combined data frame ***

```
## Installing package into '/home/hamza/R/library'
## (as 'lib' is unspecified)
```

```
##
## Attaching package: 'janitor'
```

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

*** It is time for the data cleaning phase now. First, I will check for missing values in the combined data frame. ***

```
##           trip_id        start_time          end_time            bikeid
##                 0                 0                 0                 0
##      tripduration   from_station_id from_station_name     to_station_id
##                 0                 0                 0                 0
##   to_station_name          usertype            gender         birthyear
##                 0                 0                 0            538751
##           quarter
```

```
##                 0
```

*** Some entries are note empty in the sense of "na" but they are just blank. I will check for those as well. ***

```
## [1] 559206
```

```
## Installing package into '/home/hamza/R/library'
## (as 'lib' is unspecified)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

*** Correcting the data type of the "tripduration" column from character to numeric ***

```
## Warning: NAs introduced by coercion
```

```
##  num [1:3818004] 390 441 829 NA 364 216 177 100 NA 336 ...
```

*** checking the number of customers vs subscribers in the combined data frame ***

```
##
##   Customer Subscriber
##     880637    2937367
```

*** converting start_time and end_time columns to date-time format ***

```
##  POSIXct[1:3818004], format: "2019-01-01 00:04:37" "2019-01-01 00:08:13" "2019-01-01 00:13:23" ...
```

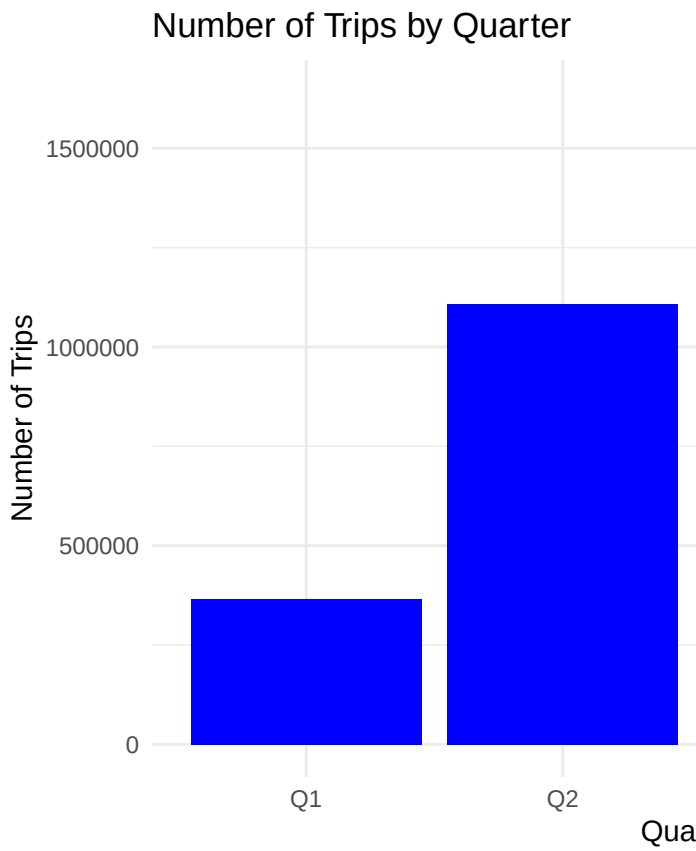
```
##  POSIXct[1:3818004], format: "2019-01-01 00:11:07" "2019-01-01 00:15:34" "2019-01-01 00:27:12" ...
```

*** Average duration of trips by usertype ***

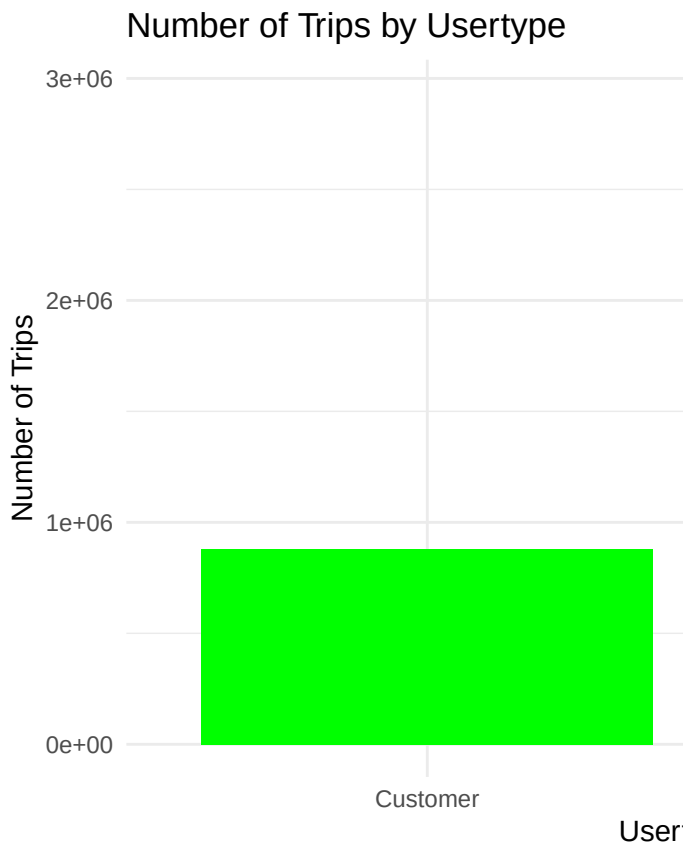
```
## # A tibble: 2 x 2
##   usertype   average_duration
##   <chr>                 <dbl>
## 1 Customer               644.
## 2 Subscriber             501.
```

*** installing ggplot2 package for visualization ***

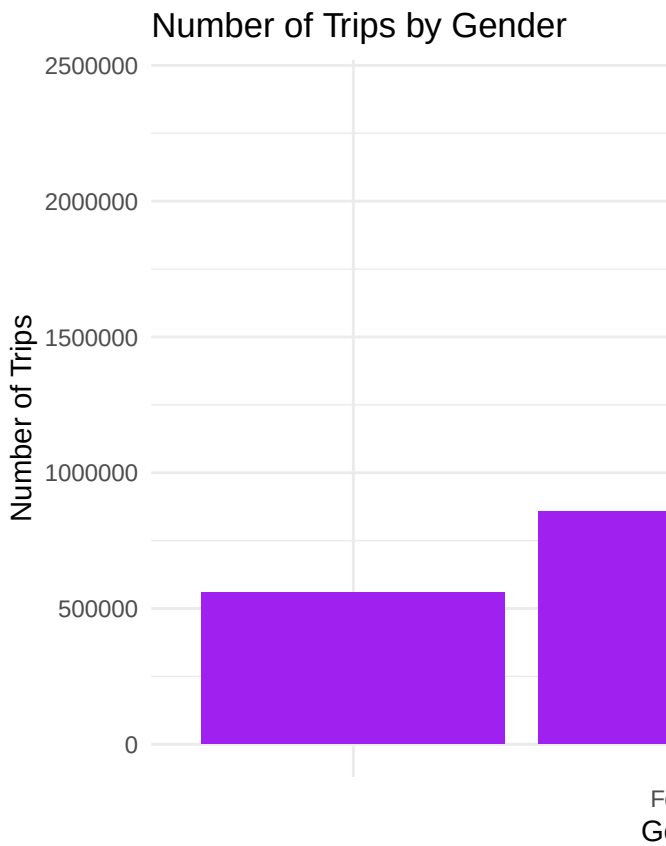
```
## Installing package into '/home/hamza/R/library'
## (as 'lib' is unspecified)
```



*** Number of trips by quarter shown in a bar chart using ggplot2
Now I will check the number of trips by usertype (Subscriber vs Customer) in ggplot2***



*** Number of trips by usertype shown in a bar chart using ggplot2
Now I will check the gender of subscribers vs customers in ggplot2***

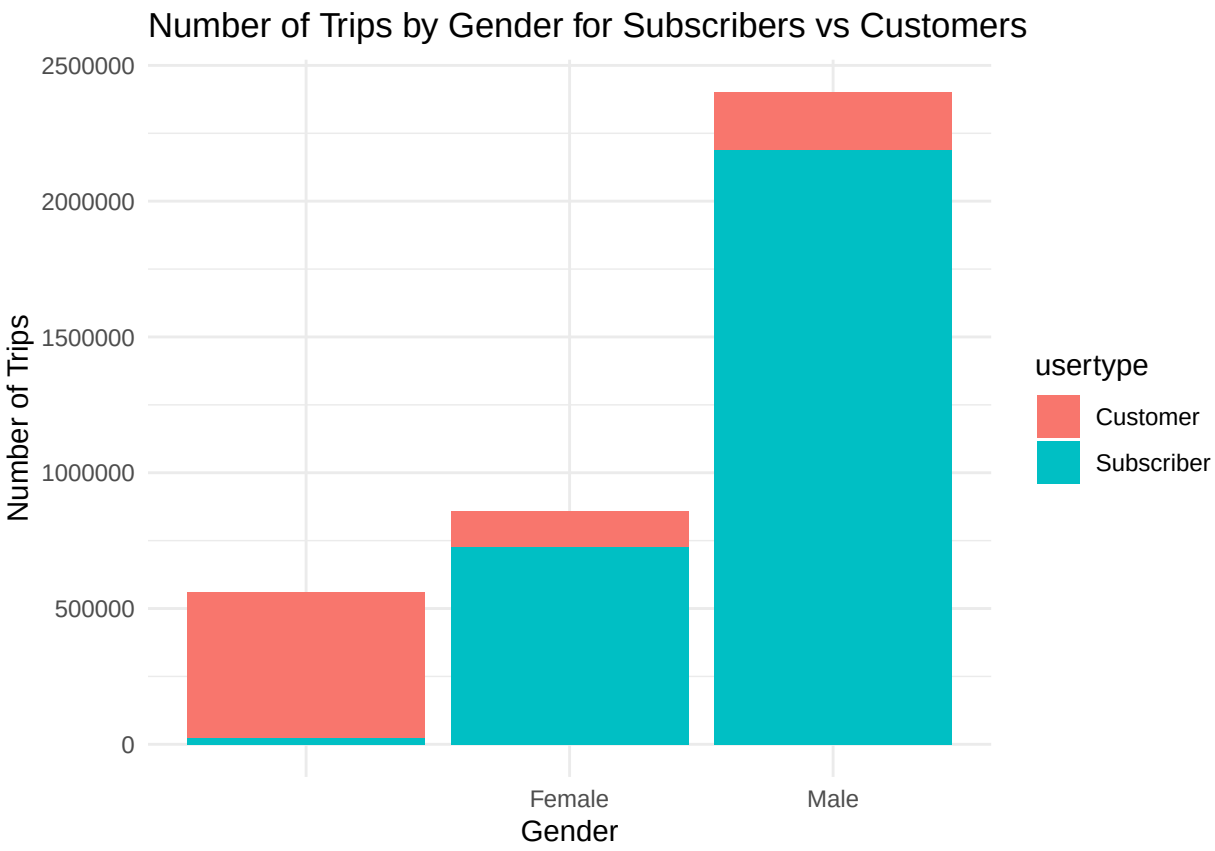


*** Number of trips by gender shown in a bar chart using ggplot2 ***

*** Now I will check the gender distribution for subscribers vs customers in ggplot2***

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

*** Number of trips by gender for subscribers vs customers shown in a bar chart using ggplot2 ***


Number of Trips by Gender for Subscribers vs Customers
2500000
2000000
1500000
1000000
500000
0
Number of Trips
Female
Male
Gender
usertype
Customer
Subscriber