Case Study - Cyclistics Bike share (2016-17)

Mohd Ali Ansari

13/06/2021

Problem Statement

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

Objective

This documents contains all the cleaning steps taken in order to clean and transform the datasets and preparing it for next step i.e. analysis Cleaning dataset is from the year 2016-2017

About dataset

Cyclistic's historical trip data to analyze and identify trends. Download data from here. (Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable to answer the business questions. The data has been made available by Motivate International Inc. under this license.)

This is public data that can use to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit from using riders' personally identifiable information. This means that one wouldn't connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes

```
library(tidyverse)
library(tidyr)
library(dplyr)
library(geosphere)
library(lubridate)
```

Importing the libraries

Loading the dataset The dataset is available in csv format after downloading so we will clean it simultaneously for merging them into one fiscal year Since the companies year starts from April month we will load all the data sets and then after checking for consistency we will merge them to make a complete one year tripdata

```
april_2016 <- read.csv("Divvy_Trips_2016_04.csv")
may_2016 <- read.csv("Divvy_Trips_2016_05.csv")
june_2016 <- read.csv("Divvy_Trips_2016_06.csv")
q3_2016 <- read.csv("Divvy_Trips_2016_Q3.csv")
```

```r
q4_2016 <- read.csv("Divvy_Trips_2016_Q4.csv")
q1_2017 <- read.csv("Divvy_Trips_2017_Q1.csv")
```

Checking for consistency We have to check for consistency as we have to merge all the datasets into one dataset. So the column names and columns data type should be same for all the datasets

Checking for Column name

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

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

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

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

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

```
##  [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"
```

There are total 12 columns in every data set and the name of columns are same for three quarter but not consistent so we will change the column name to 2020 column names

Renaming columns We will rename the columns into same format as of 2020 data because it is the latest format

```r
q4_2016 <- rename(q4_2016
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = starttime
                  ,ended_at = stoptime
                  ,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)
q3_2016 <- rename(q3_2016
                  ,ride_id = trip_id
```

```
                  ,rideable_type = bikeid
                  ,started_at = starttime
                  ,ended_at = stoptime
                  ,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)
june_2016 <- rename(june_2016
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = starttime
                  ,ended_at = stoptime
                  ,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)
may_2016 <- rename(may_2016
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = starttime
                  ,ended_at = stoptime
                  ,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)
april_2016 <- rename(april_2016
                  ,ride_id = trip_id
                  ,rideable_type = bikeid
                  ,started_at = starttime
                  ,ended_at = stoptime
                  ,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)
q1_2017 <- rename(q1_2017
                  ,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)
```

Checking data types of columns

```
## Rows: 231,635
## Columns: 12
## $ ride_id            <int> 9379901, 9379900, 9379899, 9379898, 9379897, 937989~
```

```
## $ started_at         <chr> "4/30/2016 23:59", "4/30/2016 23:58", "4/30/2016 23~
## $ ended_at           <chr> "5/1/2016 0:11", "5/1/2016 0:07", "5/1/2016 0:02", ~
## $ rideable_type      <int> 21, 3583, 4557, 2443, 50, 1676, 3619, 2200, 693, 16~
## $ tripduration       <int> 733, 556, 253, 802, 1146, 1291, 451, 1954, 226, 954~
## $ start_station_id   <int> 123, 349, 59, 289, 239, 239, 56, 129, 300, 287, 131~
## $ start_station_name <chr> "California Ave & Milwaukee Ave", "Halsted St & Wri~
## $ end_station_id     <int> 374, 165, 273, 199, 227, 227, 186, 36, 299, 56, 308~
## $ end_station_name   <chr> "Western Ave & Walton St", "Clark St & Grace St", "~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Male", "Male", "Male", "", "", "Male", "Ma~
## $ birthyear          <int> 1982, 1991, 1984, 1978, NA, NA, 1988, 1992, 1984, 1~
```

```
## Rows: 363,319
## Columns: 12
## $ ride_id            <int> 9835709, 9835708, 9835707, 9835706, 9835705, 983570~
## $ started_at         <chr> "5/31/2016 23:57", "5/31/2016 23:57", "5/31/2016 23~
## $ ended_at           <chr> "6/1/2016 0:14", "6/1/2016 0:14", "6/1/2016 0:15", ~
## $ rideable_type      <int> 609, 1207, 4369, 2703, 3828, 4377, 3115, 2489, 4711~
## $ tripduration       <int> 1045, 1035, 1166, 1348, 1101, 199, 439, 239, 271, 5~
## $ start_station_id   <int> 22, 22, 90, 174, 90, 283, 164, 349, 33, 26, 26, 130~
## $ start_station_name <chr> "May St & Taylor St", "May St & Taylor St", "Millen~
## $ end_station_id     <int> 282, 282, 320, 273, 22, 81, 110, 343, 321, 90, 90, ~
## $ end_station_name   <chr> "Halsted St & Maxwell St", "Halsted St & Maxwell St~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Male", "Male", "Female", "Male", "Male", "~
## $ birthyear          <int> 1993, 1993, 1988, 1992, 1987, 1984, 1984, 1990, 199~
```

```
## Rows: 477,873
## Columns: 12
## $ ride_id            <int> 10426657, 10426656, 10426655, 10426654, 10426653, 1~
## $ started_at         <chr> "6/30/2016 23:59", "6/30/2016 23:58", "6/30/2016 23~
## $ ended_at           <chr> "7/1/2016 0:02", "7/1/2016 0:14", "7/1/2016 0:16", ~
## $ rideable_type      <int> 1508, 1858, 3076, 4031, 4199, 858, 1083, 68, 5346, ~
## $ tripduration       <int> 190, 967, 1082, 1445, 974, 1081, 446, 2265, 1004, 1~
## $ start_station_id   <int> 93, 90, 36, 258, 90, 57, 20, 177, 26, 259, 177, 133~
## $ start_station_name <chr> "Sheffield Ave & Willow St", "Millennium Park", "Fr~
## $ end_station_id     <int> 113, 90, 69, 251, 90, 215, 289, 340, 36, 123, 340, ~
## $ end_station_name   <chr> "Bissell St & Armitage Ave", "Millennium Park", "Da~
## $ member_casual      <chr> "Subscriber", "Customer", "Subscriber", "Customer",~
## $ gender             <chr> "Male", "", "Male", "", "", "", "Female", "", "", "~
## $ birthyear          <int> 1993, NA, 1992, NA, NA, NA, 1992, NA, NA, 1986, NA,~
```

```
## Rows: 1,441,811
## Columns: 12
## $ ride_id            <int> 12150160, 12150159, 12150158, 12150157, 12150156, 1~
## $ started_at         <chr> "9/30/2016 23:59:58", "9/30/2016 23:59:58", "9/30/2~
## $ ended_at           <chr> "10/1/2016 00:04:03", "10/1/2016 00:04:09", "10/1/2~
## $ rideable_type      <int> 4959, 2589, 3656, 3570, 3158, 1026, 4895, 5851, 279~
## $ tripduration       <int> 245, 251, 1500, 245, 1638, 980, 205, 1331, 283, 170~
## $ start_station_id   <int> 69, 383, 302, 475, 302, 302, 117, 77, 222, 302, 183~
## $ start_station_name <chr> "Damen Ave & Pierce Ave", "Ashland Ave & Harrison S~
## $ end_station_id     <int> 17, 320, 334, 471, 492, 180, 300, 26, 309, 492, 350~
## $ end_station_name   <chr> "Wood St & Division St", "Loomis St & Lexington St"~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Customer", "Subscriber~
## $ gender             <chr> "Male", "Female", "", "Female", "", "Male", "Male",~
```

```
## $ birthyear          <int> 1988, 1990, NA, 1988, NA, 1991, 1993, 1987, 1988, N~
```

```
## Rows: 683,832
## Columns: 12
## $ ride_id            <int> 12979228, 12979227, 12979226, 12979225, 12979224, 1~
## $ started_at         <chr> "12/31/2016 23:57:52", "12/31/2016 23:53:18", "12/3~
## $ ended_at           <chr> "1/1/2017 00:06:44", "1/1/2017 00:08:13", "1/1/2017~
## $ rideable_type      <int> 5076, 5114, 1026, 504, 4451, 5643, 48, 2865, 1779, ~
## $ tripduration       <int> 532, 895, 931, 970, 980, 179, 1863, 1867, 1656, 108~
## $ start_station_id   <int> 502, 195, 195, 199, 199, 47, 177, 177, 195, 264, 15~
## $ start_station_name <chr> "California Ave & Altgeld St", "Columbus Dr & Rando~
## $ end_station_id     <int> 258, 25, 25, 35, 35, 125, 140, 140, 195, 52, 42, 77~
## $ end_station_name   <chr> "Logan Blvd & Elston Ave", "Michigan Ave & Pearson ~
## $ member_casual      <chr> "Customer", "Customer", "Customer", "Subscriber", "~
## $ gender             <chr> "", "", "", "Male", "Female", "Male", "", "", "", "~
## $ birthyear          <int> NA, NA, NA, 1985, 1985, 1970, NA, NA, NA, 1986, 199~
```

```
## Rows: 431,691
## Columns: 12
## $ ride_id            <int> 13518905, 13518904, 13518903, 13518902, 13518901, 1~
## $ started_at         <chr> "3/31/2017 23:59:07", "3/31/2017 23:56:25", "3/31/2~
## $ ended_at           <chr> "4/1/2017 00:13:24", "4/1/2017 00:00:21", "4/1/2017~
## $ rideable_type      <int> 5292, 4408, 696, 4915, 4247, 3536, 5111, 1579, 3914~
## $ tripduration       <int> 857, 236, 348, 288, 415, 242, 361, 354, 299, 1173, ~
## $ start_station_id   <int> 66, 199, 520, 110, 327, 143, 81, 56, 210, 117, 117,~
## $ start_station_name <chr> "Clinton St & Lake St", "Wabash Ave & Grand Ave", "~
## $ end_station_id     <int> 171, 26, 432, 142, 331, 289, 41, 77, 69, 29, 29, 30~
## $ end_station_name   <chr> "May St & Cullerton St", "McClurg Ct & Illinois St"~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Male", "Female", "Male", "Female", "Male",~
## $ birthyear          <int> 1989, 1990, 1979, 1985, 1989, 1988, 1987, 1981, 199~
```

`ride_id` is in integer form but it should be in character form for consistency

```
q1_2017 <-  mutate(q1_2017, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q3_2016 <-  mutate(q3_2016, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q4_2016 <-  mutate(q4_2016, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
april_2016 <-  mutate(april_2016, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
may_2016 <-  mutate(may_2016, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
june_2016 <-  mutate(june_2016, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
```

Changing data type *Checking one dataset if the conversion happens or not*

```
glimpse(q3_2016)
```

```
## Rows: 1,441,811
## Columns: 12
## $ ride_id            <chr> "12150160", "12150159", "12150158", "12150157", "12~
## $ started_at         <chr> "9/30/2016 23:59:58", "9/30/2016 23:59:58", "9/30/2~
```

```
## $ ended_at           <chr> "10/1/2016 00:04:03", "10/1/2016 00:04:09", "10/1/2~
## $ rideable_type      <chr> "4959", "2589", "3656", "3570", "3158", "1026", "48~
## $ tripduration       <int> 245, 251, 1500, 245, 1638, 980, 205, 1331, 283, 170~
## $ start_station_id   <int> 69, 383, 302, 475, 302, 302, 117, 77, 222, 302, 183~
## $ start_station_name <chr> "Damen Ave & Pierce Ave", "Ashland Ave & Harrison S~
## $ end_station_id     <int> 17, 320, 334, 471, 492, 180, 300, 26, 309, 492, 350~
## $ end_station_name   <chr> "Wood St & Division St", "Loomis St & Lexington St"~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Customer", "Subscriber~
## $ gender             <chr> "Male", "Female", "", "Female", "", "Male", "Male",~
## $ birthyear          <int> 1988, 1990, NA, 1988, NA, 1991, 1993, 1987, 1988, N~
```

Merging the datasets Now all the datasets is in the same order so we can now combine them into one single dataset to do the further cleaning and transform

```
all_trips <- bind_rows(april_2016,may_2016,june_2016, q3_2016, q4_2016, q1_2017)
```

Inspect the dataset Since now the dataset of all the month is transform into one dataset we will inspect it and then clean and process it to make ready for analysis

```
#checking column names
colnames(all_trips)
```

```
##  [1] "ride_id"            "started_at"         "ended_at"
##  [4] "rideable_type"      "tripduration"       "start_station_id"
##  [7] "start_station_name" "end_station_id"     "end_station_name"
## [10] "member_casual"      "gender"             "birthyear"
```

```
#checking data types
glimpse(all_trips)
```

```
## Rows: 3,630,161
## Columns: 12
## $ ride_id            <chr> "9379901", "9379900", "9379899", "9379898", "937989~
## $ started_at         <chr> "4/30/2016 23:59", "4/30/2016 23:58", "4/30/2016 23~
## $ ended_at           <chr> "5/1/2016 0:11", "5/1/2016 0:07", "5/1/2016 0:02", ~
## $ rideable_type      <chr> "21", "3583", "4557", "2443", "50", "1676", "3619",~
## $ tripduration       <int> 733, 556, 253, 802, 1146, 1291, 451, 1954, 226, 954~
## $ start_station_id   <int> 123, 349, 59, 289, 239, 239, 56, 129, 300, 287, 131~
## $ start_station_name <chr> "California Ave & Milwaukee Ave", "Halsted St & Wri~
## $ end_station_id     <int> 374, 165, 273, 199, 227, 227, 186, 36, 299, 56, 308~
## $ end_station_name   <chr> "Western Ave & Walton St", "Clark St & Grace St", "~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Male", "Male", "Male", "", "", "Male", "Ma~
## $ birthyear          <int> 1982, 1991, 1984, 1978, NA, NA, 1988, 1992, 1984, 1~
```

Remove unnecessary columns Removing uncessary columns for consistency

```
all_trips <- all_trips %>%
  select(-c(birthyear, gender,tripduration))
```

Converting data types Convert `started at` and `ended_at` to date and time

```
all_trips$started_at <- parse_date_time(all_trips$started_at, c("%m/%d/%y %H:%M","%m/%d/%y %H:%M:%S"))
all_trips$ended_at <- parse_date_time(all_trips$ended_at, c("%m/%d/%y %H:%M","%m/%d/%y %H:%M:%S"))
```

Removing inconsitency There are four unique values in member_casual `subscriber, member, customer, casual` but 2020 on wards these member has been changed into two unique values `member, casual`

```
all_trips <-  all_trips %>%
  mutate(member_casual = recode(member_casual
                                ,"Subscriber" = "member"
                                ,"Customer" = "casual"))
```

Ride length (new column) `ride_length` is the distance between started time and ended time

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at,units = "mins")
head(all_trips$ride_length)
```

```
## Time differences in mins
## [1] 12  9  4 14 19 22
```

Also we will convert the `ride_legnth` into numeric for further calculations

```
all_trips$ride_length <- round(as.numeric(as.character(all_trips$ride_length)),2)
```

Round trip (new column) We will produce a new column named `round_trip` = "Yes" where `start_station_name` is equal to `end_station_name`

```
all_trips <- all_trips %>%
  mutate(round_trip=case_when(
    start_station_name==end_station_name ~ "Yes",
    start_station_name!=end_station_name ~ "No"
  ))
head(all_trips$round_trip)
```

```
## [1] "No" "No" "No" "No" "No" "No"
```

Day (new column) calculating the `day` using the `started_date` column

```
all_trips$day <- day(all_trips$started_at)
head(all_trips$day)
```

```
## [1] 30 30 30 30 30 30
```

```
all_trips$day_of_week <- weekdays(all_trips$started_at)
head(all_trips$day_of_week)
```

Day of the week (new column)

```
## [1] "Saturday" "Saturday" "Saturday" "Saturday" "Saturday" "Saturday"
```

Month of the year (new column) calculating `month` using the `started_date` column

```
all_trips$month <- months.Date(all_trips$started_at)
head(all_trips$month)
```

```
## [1] "April" "April" "April" "April" "April" "April"
```

Year (new column) finally `year` column for summarizing the data by year

```
all_trips$year <- year(all_trips$started_at)
head(all_trips$year)
```

```
## [1] 2016 2016 2016 2016 2016 2016
```

Deleting/Filtering bad data `member_casual` has one extra value named `dependent` so we removed it

Also the negative `ride_length` is not good for analysation as the ended_time is less than the started time which is simply a bad data

```
all_trips<- all_trips %>%
  filter(!(all_trips$ride_length<0 | member_casual=="Dependent"))
```

Saving the transform data Finally saved the transform data for analysis

```
write.csv(all_trips,row.names=F,"Bike_sharing_clean/2016-17_tripdatas.csv")
```

Aggregating the file After cleaning, merging and saving all the file its time to aggregate them because the file size is too large to work with them so it is a must to agregate them into most suitable form

We will use `ride_length` for aggregating the data since it is a numerical column and it is most important for our analysis

Loading the cleaned data

```
trip <- read.csv("Bike_sharing_clean/2016-17_tripdatas.csv")
head(trip)
```

```
##   ride_id          started_at            ended_at rideable_type
## 1 9379901 2016-04-30 23:59:00 2016-05-01 00:11:00            21
## 2 9379900 2016-04-30 23:58:00 2016-05-01 00:07:00          3583
## 3 9379899 2016-04-30 23:58:00 2016-05-01 00:02:00          4557
## 4 9379898 2016-04-30 23:54:00 2016-05-01 00:08:00          2443
## 5 9379897 2016-04-30 23:52:00 2016-05-01 00:11:00            50
## 6 9379896 2016-04-30 23:49:00 2016-05-01 00:11:00          1676
##   start_station_id             start_station_name end_station_id
## 1              123 California Ave & Milwaukee Ave            374
## 2              349    Halsted St & Wrightwood Ave            165
## 3               59      Wabash Ave & Roosevelt Rd            273
## 4              289          Wells St & Concord Ln            199
## 5              239       Western Ave & Leland Ave            227
## 6              239       Western Ave & Leland Ave            227
##               end_station_name member_casual ride_length round_trip day
## 1      Western Ave & Walton St        member          12         No  30
## 2          Clark St & Grace St        member           9         No  30
## 3       Michigan Ave & 18th St        member           4         No  30
## 4       Wabash Ave & Grand Ave        member          14         No  30
## 5 Southport Ave & Waveland Ave        casual          19         No  30
## 6 Southport Ave & Waveland Ave        casual          22         No  30
##   day_of_week month year
## 1    Saturday April 2016
## 2    Saturday April 2016
## 3    Saturday April 2016
```

```
## 4    Saturday April 2016
## 5    Saturday April 2016
## 6    Saturday April 2016
```

Checking the data type of data

```
glimpse(trip)
```

```
## Rows: 3,630,117
## Columns: 15
## $ ride_id            <int> 9379901, 9379900, 9379899, 9379898, 9379897, 937989~
## $ started_at         <chr> "2016-04-30 23:59:00", "2016-04-30 23:58:00", "2016~
## $ ended_at           <chr> "2016-05-01 00:11:00", "2016-05-01 00:07:00", "2016~
## $ rideable_type      <int> 21, 3583, 4557, 2443, 50, 1676, 3619, 2200, 693, 16~
## $ start_station_id   <int> 123, 349, 59, 289, 239, 239, 56, 129, 300, 287, 131~
## $ start_station_name <chr> "California Ave & Milwaukee Ave", "Halsted St & Wri~
## $ end_station_id     <int> 374, 165, 273, 199, 227, 227, 186, 36, 299, 56, 308~
## $ end_station_name   <chr> "Western Ave & Walton St", "Clark St & Grace St", "~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
## $ ride_length        <dbl> 12, 9, 4, 14, 19, 22, 8, 33, 4, 16, 6, 3, 9, 11, 20~
## $ round_trip         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ day                <int> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,~
## $ day_of_week        <chr> "Saturday", "Saturday", "Saturday", "Saturday", "Sa~
## $ month              <chr> "April", "April", "April", "April", "April", "April~
## $ year               <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 201~
```

Changing the data_type We will convert `year` data type since we will consider it as the categorical data when we will merge all the dataset when aggregating

```
trip$year <- as.character(trip$year)
```

Checking the Statistics Since we will aggregate the data based on `ride_length`, its important to check its statistics to decide the aggregate parameter

```
summary(trip$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.98    6.87   11.72   16.50   19.52 1439.40
```

We can easily see that the ***ride_length*** *is right-skewed since there is a BIG difference between Q3 and Max value. So we will use median instead of mean for aggregation which is more ideal in this case of skewed column*

Aggregating the data

```
trip_data<-aggregate(trip$ride_length~trip$member_casual+trip$round_trip+trip$day_of_week+trip$month+tri
head(trip_data)
```

```
##   trip$member_casual trip$round_trip trip$day_of_week trip$month trip$year
## 1             casual              No           Friday      April      2016
## 2             member              No           Friday      April      2016
## 3             casual             Yes           Friday      April      2016
## 4             member             Yes           Friday      April      2016
## 5             casual              No           Monday      April      2016
## 6             member              No           Monday      April      2016
##   trip$ride_length
## 1               19
## 2                9
## 3               22
```

```
## 4                10
## 5                22
## 6                10
```

Saving the aggregate

Finally, last step is to save the data so we can use this data to merge all other aggregates data

```
write.csv(trip_data,row.names = F,"Bike_sharing_clean/tripdata_aggregate.csv")
```