Case Study - Cyclistic Bike share (2017-2018)

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 2017-2018

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

```
q2_2017 <- read.csv("Divvy_Trips_2017_Q2.csv")
q3_2017 <- read.csv("Divvy_Trips_2017_Q3.csv")
q4_2017 <- read.csv("Divvy_Trips_2017_Q4.csv")
q1_2018 <- read.csv("Divvy_Trips_2018_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"           "start_time"        "end_time"
##  [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"
```

```
##  [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"
```

```
##  [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"
```

There are total 12 columns in every data set and the name of columns are same for three quarter but not consistent for q1_2018 on wards so we will change the column name to 2020-Q1 column names

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

```
q4_2017 <- rename(q4_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)
q3_2017 <- rename(q3_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)

q2_2017 <- rename(q2_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)

q1_2018 <- rename(q1_2018
                  ,ride_id = "X01...Rental.Details.Rental.ID"
                  ,rideable_type = "X01...Rental.Details.Bike.ID"
                  ,started_at = "X01...Rental.Details.Local.Start.Time"
                  ,ended_at = "X01...Rental.Details.Local.End.Time"
                  ,start_station_name = "X03...Rental.Start.Station.Name"
                  ,start_station_id = "X03...Rental.Start.Station.ID"
                  ,end_station_name = "X02...Rental.End.Station.Name"
                  ,end_station_id = "X02...Rental.End.Station.ID"
                  ,member_casual = "User.Type")
```

Checking data types of columns

```
## Rows: 1,119,814
## Columns: 12
## $ ride_id            <int> 14853213, 14853212, 14853210, 14853209, 14853208, 1~
## $ started_at         <chr> "6/30/2017 23:59:51", "6/30/2017 23:59:28", "6/30/2~
## $ ended_at           <chr> "7/1/2017 00:13:57", "7/1/2017 00:07:10", "7/1/2017~
## $ rideable_type      <int> 893, 1909, 2071, 101, 47, 973, 181, 829, 3376, 2712~
## $ tripduration       <int> 846, 462, 340, 427, 580, 363, 322, 1569, 199, 4830,~
## $ start_station_id   <int> 107, 165, 327, 192, 130, 313, 313, 307, 261, 6, 6, ~
## $ start_station_name <chr> "Desplaines St & Jackson Blvd", "Clark St & Grace S~
## $ end_station_id     <int> 56, 234, 327, 40, 331, 157, 127, 506, 320, 35, 35, ~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "Clark St & Montrose A~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ gender             <chr> "Male", "Female", "Female", "Male", "Male", "Male",~
## $ birthyear          <int> 1975, 1968, 1996, 1980, 1990, 1990, 1992, 1989, 198~
```

```
## Rows: 1,608,270
## Columns: 12
## $ ride_id            <int> 16734065, 16734064, 16734063, 16734062, 16734061, 1~
## $ started_at         <chr> "9/30/2017 23:59:58", "9/30/2017 23:59:53", "9/30/2~
## $ ended_at           <chr> "10/1/2017 00:05:47", "10/1/2017 00:05:47", "10/1/2~
## $ rideable_type      <int> 1411, 3048, 2590, 551, 1287, 6132, 5235, 54, 3823, ~
## $ tripduration       <int> 349, 354, 226, 521, 530, 1072, 497, 214, 1398, 1072~
## $ start_station_id   <int> 216, 216, 141, 96, 96, 478, 114, 87, 90, 296, 296, ~
## $ start_station_name <chr> "California Ave & Division St", "California Ave & D~
## $ end_station_id     <int> 259, 259, 144, 217, 217, 117, 296, 127, 86, 268, 26~
## $ end_station_name   <chr> "California Ave & Francis Pl", "California Ave & Fr~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Customer~
## $ gender             <chr> "Male", "Male", "Male", "", "Female", "Male", "Male~
## $ birthyear          <int> 1985, 1979, 1993, NA, 1994, 1980, 1988, 1977, NA, 1~
```

```
## Rows: 669,239
## Columns: 12
## $ ride_id            <int> 17536701, 17536700, 17536699, 17536698, 17536697, 1~
## $ started_at         <chr> "12/31/2017 23:58", "12/31/2017 23:54", "12/31/2017~
## $ ended_at           <chr> "1/1/2018 0:03", "1/1/2018 0:18", "1/1/2018 0:18", ~
## $ rideable_type      <int> 3304, 5975, 4906, 5667, 5353, 5840, 6351, 2562, 247~
## $ tripduration       <int> 284, 1402, 1441, 315, 272, 589, 301, 141, 615, 743,~
## $ start_station_id   <int> 159, 145, 145, 340, 240, 93, 337, 226, 49, 196, 59,~
## $ start_station_name <chr> "Claremont Ave & Hirsch St", "Mies van der Rohe Way~
## $ end_station_id     <int> 69, 145, 145, 143, 245, 343, 182, 117, 26, 255, 72,~
## $ end_station_name   <chr> "Damen Ave & Pierce Ave", "Mies van der Rohe Way & ~
## $ member_casual      <chr> "Subscriber", "Customer", "Customer", "Subscriber",~
## $ gender             <chr> "Male", "", "", "Male", "Male", "Male", "Male", "Ma~
## $ birthyear          <int> 1988, NA, NA, 1963, 1977, 1988, 1990, 1987, 1981, 1~
```

```
## Rows: 387,145
## Columns: 12
## $ ride_id                                           <int> 17536702, 17536703, ~
## $ started_at                                        <chr> "2018-01-01 00:12:00~
## $ ended_at                                          <chr> "2018-01-01 00:17:23~
## $ rideable_type                                     <int> 3304, 5367, 4599, 23~
## $ X01...Rental.Details.Duration.In.Seconds.Uncapped <chr> "323.0", "377.0", "2~
## $ start_station_id                                  <int> 69, 253, 98, 125, 12~
## $ start_station_name                                <chr> "Damen Ave & Pierce ~
## $ end_station_id                                    <int> 159, 325, 509, 364, ~
## $ end_station_name                                  <chr> "Claremont Ave & Hir~
## $ member_casual                                     <chr> "Subscriber", "Subsc~
## $ Member.Gender                                     <chr> "Male", "Male", "Mal~
## $ X05...Member.Details.Member.Birthday.Year         <int> 1988, 1984, 1989, 19~
```

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

```
q4_2017 <-  mutate(q4_2017, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q3_2017 <-  mutate(q3_2017, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q2_2017 <-  mutate(q2_2017, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))
q1_2018 <-  mutate(q1_2018, 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_2017)
```

```
## Rows: 1,608,270
## Columns: 12
## $ ride_id            <chr> "16734065", "16734064", "16734063", "16734062", "16~
## $ started_at         <chr> "9/30/2017 23:59:58", "9/30/2017 23:59:53", "9/30/2~
## $ ended_at           <chr> "10/1/2017 00:05:47", "10/1/2017 00:05:47", "10/1/2~
## $ rideable_type      <chr> "1411", "3048", "2590", "551", "1287", "6132", "523~
## $ tripduration       <int> 349, 354, 226, 521, 530, 1072, 497, 214, 1398, 1072~
## $ start_station_id   <int> 216, 216, 141, 96, 96, 478, 114, 87, 90, 296, 296, ~
## $ start_station_name <chr> "California Ave & Division St", "California Ave & D~
## $ end_station_id     <int> 259, 259, 144, 217, 217, 117, 296, 127, 86, 268, 26~
## $ end_station_name   <chr> "California Ave & Francis Pl", "California Ave & Fr~
## $ member_casual      <chr> "Subscriber", "Subscriber", "Subscriber", "Customer~
## $ gender             <chr> "Male", "Male", "Male", "", "Female", "Male", "Male~
## $ birthyear          <int> 1985, 1979, 1993, NA, 1994, 1980, 1988, 1977, NA, 1~
```

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(q2_2017, q3_2017, q4_2017, q1_2018)
```

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"
##  [2] "started_at"
##  [3] "ended_at"
##  [4] "rideable_type"
##  [5] "tripduration"
##  [6] "start_station_id"
##  [7] "start_station_name"
##  [8] "end_station_id"
##  [9] "end_station_name"
## [10] "member_casual"
## [11] "gender"
## [12] "birthyear"
## [13] "X01...Rental.Details.Duration.In.Seconds.Uncapped"
## [14] "Member.Gender"
## [15] "X05...Member.Details.Member.Birthday.Year"
```

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

```
## Rows: 3,784,468
## Columns: 15
## $ ride_id                                           <chr> "14853213", "1485321~
## $ started_at                                        <chr> "6/30/2017 23:59:51"~
## $ ended_at                                          <chr> "7/1/2017 00:13:57",~
## $ rideable_type                                     <chr> "893", "1909", "2071~
## $ tripduration                                      <int> 846, 462, 340, 427, ~
## $ start_station_id                                  <int> 107, 165, 327, 192, ~
## $ start_station_name                                <chr> "Desplaines St & Jac~
## $ end_station_id                                    <int> 56, 234, 327, 40, 33~
## $ end_station_name                                  <chr> "Desplaines St & Kin~
## $ member_casual                                     <chr> "Subscriber", "Subsc~
## $ gender                                            <chr> "Male", "Female", "F~
## $ birthyear                                         <int> 1975, 1968, 1996, 19~
## $ X01...Rental.Details.Duration.In.Seconds.Uncapped <chr> NA, NA, NA, NA, NA, ~
## $ Member.Gender                                     <chr> NA, NA, NA, NA, NA, ~
## $ X05...Member.Details.Member.Birthday.Year         <int> NA, NA, NA, NA, NA, ~
```

Remove unnecessary columns Removing uncessary columns for consistency

```
all_trips <- all_trips %>%
  select(-c(birthyear, gender, "X01...Rental.Details.Duration.In.Seconds.Uncapped", "X05...Member.Detai
```

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:%S", "%y-%m-%d %H:%M:%S
all_trips$ended_at <- parse_date_time(all_trips$ended_at, c("%m/%d/%y %H:%M:%S", "%y-%m-%d %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] 14.100000  7.700000  5.666667  7.116667  9.666667  6.050000
```

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` 0

```
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"  "Yes" "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] "Friday" "Friday" "Friday" "Friday" "Friday" "Friday"
```

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] "June" "June" "June" "June" "June" "June"
```

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] 2017 2017 2017 2017 2017 2017
```

Deleting/Filtering bad data The `start_station_name = "TS ~ DIVVY PARTS TESTING"` is not relevant because it is the maintenance station for the bike so we have to remove it

`member_casual` has one extra value in three rows 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 | start_station_name=="TS ~ DIVVY PARTS TESTING" | member_casual=="De
```

Saving the transform data Finally saved the transform data for analysis

```
write.csv(all_trips,row.names=F,"Bike_sharing_clean/2017-18_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/2017-18_tripdatas.csv")
head(trip)
```

```
##    ride_id          started_at            ended_at rideable_type
## 1 14853213 2017-06-30 23:59:51 2017-07-01 00:13:57           893
## 2 14853212 2017-06-30 23:59:28 2017-07-01 00:07:10          1909
## 3 14853210 2017-06-30 23:59:18 2017-07-01 00:04:58          2071
## 4 14853209 2017-06-30 23:59:14 2017-07-01 00:06:21           101
## 5 14853208 2017-06-30 23:59:01 2017-07-01 00:08:41            47
## 6 14853206 2017-06-30 23:58:21 2017-07-01 00:04:24           973
##   start_station_id            start_station_name end_station_id
## 1              107  Desplaines St & Jackson Blvd             56
## 2              165           Clark St & Grace St            234
## 3              327   Sheffield Ave & Webster Ave            327
## 4              192           Canal St & Adams St             40
## 5              130       Damen Ave & Division St            331
## 6              313 Lakeview Ave & Fullerton Pkwy            157
##                 end_station_name member_casual ride_length round_trip day
## 1      Desplaines St & Kinzie St        member       14.10         No  30
## 2        Clark St & Montrose Ave        member        7.70         No  30
## 3    Sheffield Ave & Webster Ave        member        5.67        Yes  30
## 4          LaSalle St & Adams St        member        7.12         No  30
## 5  Halsted St & Blackhawk St (*)        member        9.67         No  30
## 6 Lake Shore Dr & Wellington Ave        member        6.05         No  30
##   day_of_week month year
## 1      Friday  June 2017
## 2      Friday  June 2017
## 3      Friday  June 2017
## 4      Friday  June 2017
## 5      Friday  June 2017
## 6      Friday  June 2017
```

Checking the data type of data

```
glimpse(trip)
```

```
## Rows: 3,784,426
## Columns: 15
## $ ride_id            <int> 14853213, 14853212, 14853210, 14853209, 14853208, 1~
## $ started_at         <chr> "2017-06-30 23:59:51", "2017-06-30 23:59:28", "2017~
## $ ended_at           <chr> "2017-07-01 00:13:57", "2017-07-01 00:07:10", "2017~
## $ rideable_type      <int> 893, 1909, 2071, 101, 47, 973, 181, 829, 3376, 2712~
## $ start_station_id   <int> 107, 165, 327, 192, 130, 313, 313, 307, 261, 6, 6, ~
## $ start_station_name <chr> "Desplaines St & Jackson Blvd", "Clark St & Grace S~
## $ end_station_id     <int> 56, 234, 327, 40, 331, 157, 127, 506, 320, 35, 35, ~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "Clark St & Montrose A~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
## $ ride_length        <dbl> 14.10, 7.70, 5.67, 7.12, 9.67, 6.05, 5.37, 26.15, 3~
## $ round_trip         <chr> "No", "No", "Yes", "No", "No", "No", "No", "No", "N~
## $ day                <int> 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30,~
## $ day_of_week        <chr> "Friday", "Friday", "Friday", "Friday", "Friday", "~
## $ month              <chr> "June", "June", "June", "June", "June", "June", "Ju~
## $ year               <int> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 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 but we have already aggregate the 2016-17 data based on median due to skewed column, we will consider this parameter for all the aggregation for consistency

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       4      10     109      18 5162393
```

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$year
## 1             casual              No          Tuesday      April      2010
## 2             member              No          Tuesday      April      2010
## 3             casual             Yes          Tuesday      April      2010
## 4             member             Yes          Tuesday      April      2010
## 5             casual              No           Friday     August      2010
## 6             member              No           Friday     August      2010
##   trip$ride_length
## 1            0.370
## 2            0.150
## 3            0.645
## 4            0.190
## 5            0.420
## 6            0.180
```

Saving the aggregate

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

We will merge the data with the old data we saved while aggregating

```
trip_old <- read.csv("Bike_sharing_clean/tripdata_aggregate.csv")
head(trip_old)
```

```
##   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
```

Tranforming the old data to merge perfectly We have to make column name and type consistent before merging

```
trip_old$trip.year <- as.character(trip_old$trip.year)
trip_old <- rename(trip_old,
                   "trip$member_casual"=trip.member_casual,
                   "trip$round_trip"=trip.round_trip,
                   "trip$day_of_week"=trip.day_of_week,
                   "trip$month"=trip.month,
                   "trip$year"=trip.year,
                   "trip$ride_length"=trip.ride_length
)
```

```
trip_merged <- bind_rows(trip_old, trip_data)
write.csv(trip_merged,row.names = F,"Bike_sharing_clean/tripdata_aggregate.csv")
```