# Case Study - Cyclistic Bike share (2020-2021)

Mohd Ali Ansari

15/06/2021

### 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 Dataset is from 2020-2021

### 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

```
df_202004<-read.csv("202004-divvy-tripdata.csv")
df_202005<-read.csv("202005-divvy-tripdata.csv")
df_202006<-read.csv("202006-divvy-tripdata.csv")
df_202007<-read.csv("202007-divvy-tripdata.csv")
df_202008<-read.csv("202008-divvy-tripdata.csv")
df_202009<-read.csv("202009-divvy-tripdata.csv")
df_202010<-read.csv("202010-divvy-tripdata.csv")
df_202011<-read.csv("202011-divvy-tripdata.csv")
df_202012<-read.csv("202012-divvy-tripdata.csv")
```

```
df_202101<-read.csv("202101-divvy-tripdata.csv")
df_202102<-read.csv("202102-divvy-tripdata.csv")
df_202103<-read.csv("202103-divvy-tripdata.csv")
df_202104<-read.csv("202104-divvy-tripdata.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] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

*There are total 13 columns in every dataset and also the name are same so we are good to move on to the next step*

**Checking data types of columns**

```
## Rows: 84,776
## Columns: 13
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-04-26 17:45:14", "2020-04-17 17:08:54", "2020~
## $ ended_at           <chr> "2020-04-26 18:12:03", "2020-04-17 17:17:03", "2020~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <int> 86, 503, 142, 216, 125, 173, 35, 434, 627, 377, 508~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <int> 152, 499, 255, 657, 323, 35, 635, 382, 359, 508, 37~
## $ start_lat          <dbl> 41.8964, 41.9244, 41.8945, 41.9030, 41.8902, 41.896~
## $ start_lng          <dbl> -87.6610, -87.7154, -87.6179, -87.6975, -87.6262, -~
## $ end_lat            <dbl> 41.9322, 41.9306, 41.8679, 41.8992, 41.9695, 41.892~
## $ end_lng            <dbl> -87.6586, -87.7238, -87.6230, -87.6722, -87.6547, -~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

```
## Rows: 200,274
## Columns: 13
## $ ride_id            <chr> "02668AD35674B983", "7A50CCAF1EDDB28F", "2FFCDFDB91~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-05-27 10:03:52", "2020-05-25 10:47:11", "2020~
## $ ended_at           <chr> "2020-05-27 10:16:49", "2020-05-25 11:05:40", "2020~
## $ start_station_name <chr> "Franklin St & Jackson Blvd", "Clark St & Wrightwoo~
## $ start_station_id   <int> 36, 340, 260, 251, 261, 206, 261, 180, 331, 219, 24~
## $ end_station_name   <chr> "Wabash Ave & Grand Ave", "Clark St & Leland Ave", ~
## $ end_station_id     <int> 199, 326, 260, 157, 206, 22, 261, 180, 300, 305, 14~
## $ start_lat          <dbl> 41.8777, 41.9295, 41.9296, 41.9680, 41.8715, 41.847~
## $ start_lng          <dbl> -87.6353, -87.6431, -87.7079, -87.6500, -87.6699, -~
## $ end_lat            <dbl> 41.8915, 41.9671, 41.9296, 41.9367, 41.8472, 41.869~
## $ end_lng            <dbl> -87.6268, -87.6674, -87.7079, -87.6368, -87.6468, -~
## $ member_casual      <chr> "member", "casual", "casual", "casual", "member", "~

## Rows: 343,005
## Columns: 13
## $ ride_id            <chr> "8CD5DE2C2B6C4CFC", "9A191EB2C751D85D", "F37D14B0B5~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-06-13 23:24:48", "2020-06-26 07:26:10", "2020~
## $ ended_at           <chr> "2020-06-13 23:36:55", "2020-06-26 07:31:58", "2020~
## $ start_station_name <chr> "Wilton Ave & Belmont Ave", "Federal St & Polk St",~
## $ start_station_id   <int> 117, 41, 81, 303, 327, 327, 41, 115, 338, 84, 317, ~
## $ end_station_name   <chr> "Damen Ave & Clybourn Ave", "Daley Center Plaza", "~
## $ end_station_id     <int> 163, 81, 5, 294, 117, 117, 81, 303, 164, 53, 168, 1~
## $ start_lat          <dbl> 41.94018, 41.87208, 41.88424, 41.94553, 41.92154, 4~
## $ start_lng          <dbl> -87.65304, -87.62954, -87.62963, -87.64644, -87.653~
## $ end_lat            <dbl> 41.93193, 41.88424, 41.87405, 41.97835, 41.94018, 4~
## $ end_lng            <dbl> -87.67786, -87.62963, -87.62772, -87.65975, -87.653~
## $ member_casual      <chr> "casual", "member", "member", "casual", "casual", "~

## Rows: 551,480
## Columns: 13
## $ ride_id            <chr> "762198876D69004D", "BEC9C9FBA0D4CF1B", "D2FD8EA432~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-07-09 15:22:02", "2020-07-24 23:56:30", "2020~
## $ ended_at           <chr> "2020-07-09 15:25:52", "2020-07-25 00:20:17", "2020~
## $ start_station_name <chr> "Ritchie Ct & Banks St", "Halsted St & Roscoe St", ~
## $ start_station_id   <int> 180, 299, 329, 181, 268, 635, 113, 211, 176, 31, 14~
## $ end_station_name   <chr> "Wells St & Evergreen Ave", "Broadway & Ridge Ave",~
## $ end_station_id     <int> 291, 461, 156, 94, 301, 289, 140, 31, 191, 142, 31,~
## $ start_lat          <dbl> 41.90687, 41.94367, 41.93259, 41.89076, 41.91172, 4~
## $ start_lng          <dbl> -87.62622, -87.64895, -87.63643, -87.63170, -87.626~
## $ end_lat            <dbl> 41.90672, 41.98404, 41.93650, 41.91831, 41.90799, 4~
## $ end_lng            <dbl> -87.63483, -87.66027, -87.64754, -87.63628, -87.631~
## $ member_casual      <chr> "member", "member", "casual", "casual", "member", "~

## Rows: 622,361
## Columns: 13
## $ ride_id            <chr> "322BD23D287743ED", "2A3AEF1AB9054D8B", "67DC1D133E~
## $ rideable_type      <chr> "docked_bike", "electric_bike", "electric_bike", "e~
## $ started_at         <chr> "2020-08-20 18:08:14", "2020-08-27 18:46:04", "2020~
## $ ended_at           <chr> "2020-08-20 18:17:51", "2020-08-27 19:54:51", "2020~
## $ start_station_name <chr> "Lake Shore Dr & Diversey Pkwy", "Michigan Ave & 14~
## $ start_station_id   <int> 329, 168, 195, 81, 658, 658, 196, 67, 153, 177, 313~
## $ end_station_name   <chr> "Clark St & Lincoln Ave", "Michigan Ave & 14th St",~
## $ end_station_id     <int> 141, 168, 44, 47, 658, 658, 49, 229, 225, 305, 296,~
## $ start_lat          <dbl> 41.93259, 41.86438, 41.88464, 41.88409, 41.90299, 4~
## $ start_lng          <dbl> -87.63643, -87.62368, -87.61955, -87.62964, -87.683~
## $ end_lat            <dbl> 41.91569, 41.86422, 41.88497, 41.88958, 41.90300, 4~
## $ end_lng            <dbl> -87.63460, -87.62344, -87.62757, -87.62754, -87.683~
## $ member_casual      <chr> "member", "casual", "casual", "casual", "casual", "~

## Rows: 532,958
## Columns: 13
## $ ride_id            <chr> "2B22BD5F95FB2629", "A7FB70B4AFC6CAF2", "86057FA01B~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2020-09-17 14:27:11", "2020-09-17 15:07:31", "2020~
## $ ended_at           <chr> "2020-09-17 14:44:24", "2020-09-17 15:07:45", "2020~
## $ start_station_name <chr> "Michigan Ave & Lake St", "W Oakdale Ave & N Broadw~
## $ start_station_id   <int> 52, NA, NA, 246, 24, 94, 291, NA, NA, NA, 273, 145,~
## $ end_station_name   <chr> "Green St & Randolph St", "W Oakdale Ave & N Broadw~
## $ end_station_id     <int> 112, NA, NA, 249, 24, NA, 256, NA, NA, NA, 273, NA,~
## $ start_lat          <dbl> 41.88669, 41.94000, 41.94000, 41.95606, 41.89186, 4~
## $ start_lng          <dbl> -87.62356, -87.64000, -87.64000, -87.66892, -87.621~
## $ end_lat            <dbl> 41.88357, 41.94000, 41.94000, 41.96398, 41.89135, 4~
## $ end_lng            <dbl> -87.64873, -87.64000, -87.64000, -87.63822, -87.620~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~

## Rows: 388,653
## Columns: 13
## $ ride_id            <chr> "ACB6B40CF5B9044C", "DF450C72FD109C01", "B6396B54A1~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2020-10-31 19:39:43", "2020-10-31 23:50:08", "2020~
## $ ended_at           <chr> "2020-10-31 19:57:12", "2020-11-01 00:04:16", "2020~
## $ start_station_name <chr> "Lakeview Ave & Fullerton Pkwy", "Southport Ave & W~
## $ start_station_id   <int> 313, 227, 102, 165, 190, 359, 313, 125, NA, 174, 11~
## $ end_station_name   <chr> "Rush St & Hubbard St", "Kedzie Ave & Milwaukee Ave~
## $ end_station_id     <int> 125, 260, 423, 256, 185, 53, 125, 313, 199, 635, 30~
## $ start_lat          <dbl> 41.92610, 41.94817, 41.77346, 41.95085, 41.92886, 4~
## $ start_lng          <dbl> -87.63898, -87.66391, -87.58537, -87.65924, -87.663~
## $ end_lat            <dbl> 41.89035, 41.92953, 41.79145, 41.95281, 41.91778, 4~
## $ end_lng            <dbl> -87.62607, -87.70782, -87.60005, -87.65010, -87.691~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~

## Rows: 259,716
## Columns: 13
## $ ride_id            <chr> "BD0A6FF6FFF9B921", "96A7A7A4BDE4F82D", "C61526D065~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2020-11-01 13:36:00", "2020-11-01 10:03:26", "2020~
## $ ended_at           <chr> "2020-11-01 13:45:40", "2020-11-01 10:14:45", "2020~
## $ start_station_name <chr> "Dearborn St & Erie St", "Franklin St & Illinois St~
## $ start_station_id   <int> 110, 672, 76, 659, 2, 72, 76, NA, 58, 394, 623, NA,~
## $ end_station_name   <chr> "St. Clair St & Erie St", "Noble St & Milwaukee Ave~
## $ end_station_id     <int> 211, 29, 41, 185, 2, 76, 72, NA, 288, 273, 2, 506, ~
## $ start_lat          <dbl> 41.89418, 41.89096, 41.88098, 41.89550, 41.87650, 4~
## $ start_lng          <dbl> -87.62913, -87.63534, -87.61675, -87.68201, -87.620~
## $ end_lat            <dbl> 41.89443, 41.90067, 41.87205, 41.91774, 41.87645, 4~
## $ end_lng            <dbl> -87.62338, -87.66248, -87.62955, -87.69139, -87.620~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~

## Rows: 131,573
## Columns: 13
## $ ride_id            <chr> "70B6A9A437D4C30D", "158A465D4E74C54A", "5262016E0F~
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "~
## $ started_at         <chr> "2020-12-27 12:44:29", "2020-12-18 17:37:15", "2020~
## $ ended_at           <chr> "2020-12-27 12:55:06", "2020-12-18 17:44:19", "2020~
## $ start_station_name <chr> "Aberdeen St & Jackson Blvd", "", "", "", "", "", "~
## $ start_station_id   <chr> "13157", "", "", "", "", "", "", "", "", "", "", ""~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", "", "", "", "", "", ""~
## $ end_station_id     <chr> "TA1306000003", "", "", "", "", "", "", "", "", "",~
## $ start_lat          <dbl> 41.87773, 41.93000, 41.91000, 41.92000, 41.80000, 4~
## $ start_lng          <dbl> -87.65479, -87.70000, -87.69000, -87.70000, -87.590~
## $ end_lat            <dbl> 41.88872, 41.91000, 41.93000, 41.91000, 41.80000, 4~
## $ end_lng            <dbl> -87.64445, -87.70000, -87.70000, -87.70000, -87.590~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~

## Rows: 96,834
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2021-01-23 16:14:19", "2021-01-27 18:43:08", "2021~
## $ ended_at           <chr> "2021-01-23 16:24:44", "2021-01-27 18:47:12", "2021~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> "", "", "", "", "", "", "", "", "", "Wood St & Augu~
## $ end_station_id     <chr> "", "", "", "", "", "", "", "", "", "657", "13258",~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~

## Rows: 49,622
## Columns: 13
## $ ride_id            <chr> "89E7AA6C29227EFF", "0FEFDE2603568365", "E6159D746B~
## $ rideable_type      <chr> "classic_bike", "classic_bike", "electric_bike", "c~
## $ started_at         <chr> "2021-02-12 16:14:56", "2021-02-14 17:52:38", "2021~
## $ ended_at           <chr> "2021-02-12 16:21:43", "2021-02-14 18:12:09", "2021~
## $ start_station_name <chr> "Glenwood Ave & Touhy Ave", "Glenwood Ave & Touhy A~
## $ start_station_id   <chr> "525", "525", "KA1503000012", "637", "13216", "1800~
## $ end_station_name   <chr> "Sheridan Rd & Columbia Ave", "Bosworth Ave & Howar~
## $ end_station_id     <chr> "660", "16806", "TA1305000029", "TA1305000034", "TA~
## $ start_lat          <dbl> 42.01270, 42.01270, 41.88579, 41.89563, 41.83473, 4~
## $ start_lng          <dbl> -87.66606, -87.66606, -87.63110, -87.67207, -87.625~
## $ end_lat            <dbl> 42.00458, 42.01954, 41.88487, 41.90312, 41.83816, 4~
## $ end_lng            <dbl> -87.66141, -87.66956, -87.62750, -87.67394, -87.645~
## $ member_casual      <chr> "member", "casual", "member", "member", "member", "~

## Rows: 228,496
## Columns: 13
## $ ride_id            <chr> "CFA86D4455AA1030", "30D9DC61227D1AF3", "846D87A156~
## $ rideable_type      <chr> "classic_bike", "classic_bike", "classic_bike", "cl~
## $ started_at         <chr> "2021-03-16 08:32:30", "2021-03-28 01:26:28", "2021~
## $ ended_at           <chr> "2021-03-16 08:36:34", "2021-03-28 01:36:55", "2021~
## $ start_station_name <chr> "Humboldt Blvd & Armitage Ave", "Humboldt Blvd & Ar~
## $ start_station_id   <chr> "15651", "15651", "15443", "TA1308000021", "525", "~
## $ end_station_name   <chr> "Stave St & Armitage Ave", "Central Park Ave & Bloo~
## $ end_station_id     <chr> "13266", "18017", "TA1308000043", "13323", "E008", ~
## $ start_lat          <dbl> 41.91751, 41.91751, 41.84273, 41.96881, 42.01270, 4~
## $ start_lng          <dbl> -87.70181, -87.70181, -87.63549, -87.65766, -87.666~
## $ end_lat            <dbl> 41.91774, 41.91417, 41.83066, 41.95283, 42.05049, 4~
## $ end_lng            <dbl> -87.69139, -87.71676, -87.64717, -87.64999, -87.677~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
## Rows: 337,230
## Columns: 13
## $ ride_id            <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508~
## $ rideable_type      <chr> "classic_bike", "docked_bike", "docked_bike", "clas~
## $ started_at         <chr> "2021-04-12 18:25:36", "2021-04-27 17:27:11", "2021~
## $ ended_at           <chr> "2021-04-12 18:56:55", "2021-04-27 18:31:29", "2021~
## $ start_station_name <chr> "State St & Pearson St", "Dorchester Ave & 49th St"~
## $ start_station_id   <chr> "TA1307000061", "KA1503000069", "20121", "TA1305000~
## $ end_station_name   <chr> "Southport Ave & Waveland Ave", "Dorchester Ave & 4~
## $ end_station_id     <chr> "13235", "KA1503000069", "20121", "13235", "20121",~
## $ start_lat          <dbl> 41.89745, 41.80577, 41.74149, 41.90312, 41.74149, 4~
## $ start_lng          <dbl> -87.62872, -87.59246, -87.65841, -87.67394, -87.658~
## $ end_lat            <dbl> 41.94815, 41.80577, 41.74149, 41.94815, 41.74149, 4~
## $ end_lng            <dbl> -87.66394, -87.59246, -87.65841, -87.66394, -87.658~
## $ member_casual      <chr> "member", "casual", "casual", "member", "casual", "~
```

*So in from `df_202012` on wards the `start_station_id` and `end_station_id` is in character form but it should be in integer form*

**Converting data type** We will change the format of those columns from `character` to `integer`

```
df_202012 <- mutate(df_202012,start_station_id=as.integer(start_station_id), end_station_id=as.integer(
df_202101 <- mutate(df_202101,start_station_id=as.integer(start_station_id), end_station_id=as.integer(
df_202102 <- mutate(df_202102,start_station_id=as.integer(start_station_id), end_station_id=as.integer(
df_202103 <- mutate(df_202103,start_station_id=as.integer(start_station_id), end_station_id=as.integer(
df_202104 <- mutate(df_202104,start_station_id=as.integer(start_station_id), end_station_id=as.integer(
```

*Checking one dataset if the conversion happens or not*

```
glimpse(df_202104)
```

```
## Rows: 337,230
## Columns: 13
## $ ride_id            <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508~
## $ rideable_type      <chr> "classic_bike", "docked_bike", "docked_bike", "clas~
## $ started_at         <chr> "2021-04-12 18:25:36", "2021-04-27 17:27:11", "2021~
## $ ended_at           <chr> "2021-04-12 18:56:55", "2021-04-27 18:31:29", "2021~
## $ start_station_name <chr> "State St & Pearson St", "Dorchester Ave & 49th St"~
## $ start_station_id   <int> NA, NA, 20121, NA, 20121, 15542, 16948, NA, 16948, ~
## $ end_station_name   <chr> "Southport Ave & Waveland Ave", "Dorchester Ave & 4~
## $ end_station_id     <int> 13235, NA, 20121, 13235, 20121, 15542, 16948, NA, 1~
## $ start_lat          <dbl> 41.89745, 41.80577, 41.74149, 41.90312, 41.74149, 4~
## $ start_lng          <dbl> -87.62872, -87.59246, -87.65841, -87.67394, -87.658~
## $ end_lat            <dbl> 41.94815, 41.80577, 41.74149, 41.94815, 41.74149, 4~
## $ end_lng            <dbl> -87.66394, -87.59246, -87.65841, -87.66394, -87.658~
## $ member_casual      <chr> "member", "casual", "casual", "member", "casual", "~
```

**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_tripdata <- bind_rows(df_202004,df_202005,df_202006,df_202007,df_202008,df_202009,df_202010,df_2020
```

**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_tripdata)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

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

```
## Rows: 3,826,978
## Columns: 13
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-04-26 17:45:14", "2020-04-17 17:08:54", "2020~
## $ ended_at           <chr> "2020-04-26 18:12:03", "2020-04-17 17:17:03", "2020~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <int> 86, 503, 142, 216, 125, 173, 35, 434, 627, 377, 508~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <int> 152, 499, 255, 657, 323, 35, 635, 382, 359, 508, 37~
## $ start_lat          <dbl> 41.8964, 41.9244, 41.8945, 41.9030, 41.8902, 41.896~
## $ start_lng          <dbl> -87.6610, -87.7154, -87.6179, -87.6975, -87.6262, -~
## $ end_lat            <dbl> 41.9322, 41.9306, 41.8679, 41.8992, 41.9695, 41.892~
## $ end_lng            <dbl> -87.6586, -87.7238, -87.6230, -87.6722, -87.6547, -~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

**Remove unnecessary columns** Since in the old dataset i.e. before 2020 there is no record of latitude and longitude, so, we will remove them for consistency

```
all_tripdata <- all_tripdata %>%
  select(-c(start_lat,start_lng,end_lng,end_lat))
```

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

```
all_tripdata$started_at <- ymd_hms(all_tripdata$started_at)
all_tripdata$ended_at <- ymd_hms(all_tripdata$ended_at)
```

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

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

```
## Time differences in mins
## [1] 26.81667  8.15000 14.38333 12.20000 52.91667  5.40000
```

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

```
all_tripdata$ride_length <- round(as.numeric(as.character(all_tripdata$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`

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

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

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

```r
all_tripdata <- all_tripdata %>%
  mutate(day=day(started_at))
head(all_tripdata$day)
```

```
## [1] 26 17  1  7 18 30
```

```r

all_tripdata <- all_tripdata %>%
  mutate(day_of_week=weekdays(started_at))
head(all_tripdata$day_of_week)
```

**Day of the week (new column)**

```
## [1] "Sunday"    "Friday"    "Wednesday" "Tuesday"   "Saturday"  "Thursday"
```

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

```r
all_tripdata <- all_tripdata %>%
  mutate(month=months.Date(started_at))
head(all_tripdata$month)
```

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

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

```r
all_tripdata <- all_tripdata %>%
  mutate(year=year(started_at))
head(all_tripdata$year)
```

```
## [1] 2020 2020 2020 2020 2020 2020
```

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

```r
all_tripdata <- all_tripdata %>%
  filter(!(ride_length<0 | start_station_name =="WATSON TESTING - DIVVY"))
```

**Saving the transform data** Finally saved the transform data for analysis

```r
write.csv(all_tripdata,row.names=F,"Bike_sharing_clean/2020-21_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/2020-21_tripdatas.csv")
head(trip)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 A847FADBBC638E45   docked_bike 2020-04-26 17:45:14 2020-04-26 18:12:03
## 2 5405B80E996FF60D   docked_bike 2020-04-17 17:08:54 2020-04-17 17:17:03
## 3 5DD24A79A4E006F4   docked_bike 2020-04-01 17:54:13 2020-04-01 18:08:36
## 4 2A59BBDF5CDBA725   docked_bike 2020-04-07 12:50:19 2020-04-07 13:02:31
## 5 27AD306C119C6158   docked_bike 2020-04-18 10:22:59 2020-04-18 11:15:54
## 6 356216E875132F61   docked_bike 2020-04-30 17:55:47 2020-04-30 18:01:11
##                      start_station_name start_station_id
## 1                        Eckhart Park               86
## 2           Drake Ave & Fullerton Ave              503
## 3                McClurg Ct & Erie St              142
## 4        California Ave & Division St              216
## 5                Rush St & Hubbard St              125
## 6 Mies van der Rohe Way & Chicago Ave              173
##              end_station_name end_station_id member_casual ride_length
## 1 Lincoln Ave & Diversey Pkwy            152        member       26.82
## 2             Kosciuszko Park            499        member        8.15
## 3  Indiana Ave & Roosevelt Rd            255        member       14.38
## 4      Wood St & Augusta Blvd            657        member       12.20
## 5  Sheridan Rd & Lawrence Ave            323        casual       52.92
## 6     Streeter Dr & Grand Ave             35        member        5.40
##   round_trip day day_of_week month year
## 1         No  26      Sunday April 2020
## 2         No  17      Friday April 2020
## 3         No   1   Wednesday April 2020
## 4         No   7     Tuesday April 2020
## 5         No  18    Saturday April 2020
## 6         No  30    Thursday April 2020
```

### Checking the data type of data

```
glimpse(trip)
```

```
## Rows: 2,999,812
## Columns: 15
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-04-26 17:45:14", "2020-04-17 17:08:54", "2020~
## $ ended_at           <chr> "2020-04-26 18:12:03", "2020-04-17 17:17:03", "2020~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <int> 86, 503, 142, 216, 125, 173, 35, 434, 627, 377, 508~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <int> 152, 499, 255, 657, 323, 35, 635, 382, 359, 508, 37~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
## $ ride_length        <dbl> 26.82, 8.15, 14.38, 12.20, 52.92, 5.40, 5.22, 75.82~
## $ round_trip         <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No~
## $ day                 <int> 26, 17, 1, 7, 18, 30, 2, 7, 15, 4, 4, 25, 24, 11, 2~
## $ day_of_week         <chr> "Sunday", "Friday", "Wednesday", "Tuesday", "Saturd~
## $ month               <chr> "April", "April", "April", "April", "April", "April~
## $ year                <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 202~
```

**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.
##     0.00     8.43    15.40    29.24    27.85 58720.03
```

*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$year
## 1             casual              No           Friday      April      2020
## 2             member              No           Friday      April      2020
## 3             casual             Yes           Friday      April      2020
## 4             member             Yes           Friday      April      2020
## 5             casual              No           Monday      April      2020
## 6             member              No           Monday      April      2020
##   trip$ride_length
## 1            19.69
## 2            12.23
## 3            33.18
## 4            21.20
## 5            20.11
## 6            12.47
```

**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")
```