Google\_Data\_Analytics\_Capstone\_Project\_on\_Cyclistic\_Bike\_Share

Arijit Roy

24/11/2021

# Cyclistic Bike Share: Google Data Analytics Course Capstone Project

## Business Task

By analyzing one year of data of the Cyclistic, a bike sharing company, to find out a way that how to convert the casual riders into members.

* Note:This analysis will be based on the “divvy-tripdata”, from the source <https://divvy-tripdata.s3.amazonaws.com/index.html>. From November of 2020 to October of 2021, 12 months of data are taken for this analysis.

## Ask

The question to be answered by the process:

* 1. How do annual members and casual riders use Cyclistic bikes differently?
  2. Why would casual riders buy Cyclistic annual memberships?
  3. How can Cyclistic use digital media to influence casual riders to become members?

## Prepare

In this phase the 12 different .csv type file was downloaded from the given data source. The name of the files were changed and were arranged in a folder so that the file could be recognized easily.

## Process

To conduct all the data clearing, manipulating and visualization process, the “RStudio” app has been used and for some additional data visualization, the “tableau-public” has been used.

### Setting up my environment:

Notes: setting up my R environment by loading the tidyverse,lubridate,janitor and ggplot2 packages:

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(janitor)

##   
## Attaching package: 'janitor'

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

library(ggplot2)

### \* STEP 1: COLLECT DATA:

r2011 <- read\_csv("bikeride\_20\_11.csv")

## Rows: 259716 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (5): ride\_id, rideable\_type, start\_station\_name, end\_station\_name, memb...  
## dbl (6): start\_station\_id, end\_station\_id, start\_lat, start\_lng, end\_lat, e...  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2012 <- read\_csv("bikeride\_20\_12.csv")

## Rows: 131573 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2101 <- read\_csv("bikeride\_21\_01.csv")

## Rows: 96834 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2102 <- read\_csv("bikeride\_21\_02.csv")

## Rows: 49622 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2103 <- read\_csv("bikeride\_21\_03.csv")

## Rows: 228496 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2104 <- read\_csv("bikeride\_21\_04.csv")

## Rows: 337230 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2105 <- read\_csv("bikeride\_21\_05.csv")

## Rows: 531633 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2106 <- read\_csv("bikeride\_21\_06.csv")

## Rows: 729595 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2107 <- read\_csv("bikeride\_21\_07.csv")

## Rows: 822410 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2108 <- read\_csv("bikeride\_21\_08.csv")

## Rows: 804352 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2109 <- read\_csv("bikeride\_21\_09.csv")

## Rows: 756147 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

r2110 <- read\_csv("bikeride\_21\_10.csv")

## Rows: 631226 Columns: 13

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (7): ride\_id, rideable\_type, start\_station\_name, start\_station\_id, end\_...  
## dbl (4): start\_lat, start\_lng, end\_lat, end\_lng  
## dttm (2): started\_at, ended\_at

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

### \* STEP 2:WRANGLE DATA AND COMBINE INTO A SINGLE FILE:

Compare column names each of the files:

colnames(r2011)

## [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"

colnames(r2012)

## [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"

colnames(r2101)

## [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"

colnames(r2102)

## [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"

colnames(r2103)

## [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"

colnames(r2104)

## [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"

colnames(r2105)

## [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"

colnames(r2106)

## [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"

colnames(r2107)

## [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"

colnames(r2108)

## [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"

colnames(r2109)

## [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"

colnames(r2110)

## [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 the data format of each column:

str(r2011)

## spec\_tbl\_df [259,716 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:259716] "D0C1A65AFC859916" "985AD418D8ED316C" "AB459A840CF1BBF8" "5239202FF2A38305" ...  
## $ rideable\_type : chr [1:259716] "docked\_bike" "docked\_bike" "docked\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:259716], format: "2020-11-01 00:00:00" "2020-11-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:259716], format: "2020-11-01 00:35:00" "2020-11-01 00:02:00" ...  
## $ start\_station\_name: chr [1:259716] "Broadway & Barry Ave" "Desplaines St & Kinzie St" "Broadway & Cornelia Ave" "Wells St & Evergreen Ave" ...  
## $ start\_station\_id : num [1:259716] 300 56 303 291 291 110 153 NA NA 344 ...  
## $ end\_station\_name : chr [1:259716] "Broadway & Barry Ave" "Desplaines St & Randolph St" "Sheridan Rd & Irving Park Rd" "Franklin St & Chicago Ave" ...  
## $ end\_station\_id : num [1:259716] 300 96 240 31 31 291 289 NA NA 226 ...  
## $ start\_lat : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:259716] -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:259716] 41.9 41.9 42 41.9 41.9 ...  
## $ end\_lng : num [1:259716] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:259716] "casual" "member" "member" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_double(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_double(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2012)

## spec\_tbl\_df [131,573 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:131573] "1C46BF5EB60CC524" "1405BFC02FDB5190" "892ECFAB44203EAD" "7F0119D219127E62" ...  
## $ rideable\_type : chr [1:131573] "electric\_bike" "electric\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : POSIXct[1:131573], format: "2020-12-01 00:01:00" "2020-12-01 00:01:00" ...  
## $ ended\_at : POSIXct[1:131573], format: "2020-12-01 00:06:00" "2020-12-01 00:06:00" ...  
## $ start\_station\_name: chr [1:131573] NA NA "Larrabee St & Armitage Ave" "Wabash Ave & Adams St" ...  
## $ start\_station\_id : chr [1:131573] NA NA "TA1309000006" "KA1503000015" ...  
## $ end\_station\_name : chr [1:131573] NA "Wentworth Ave & 63rd St" "Sedgwick St & Webster Ave" "Peoria St & Jackson Blvd" ...  
## $ end\_station\_id : chr [1:131573] NA "KA1503000025" "13191" "13158" ...  
## $ start\_lat : num [1:131573] 41.8 41.8 41.9 41.9 42 ...  
## $ start\_lng : num [1:131573] -87.6 -87.6 -87.6 -87.6 -87.7 ...  
## $ end\_lat : num [1:131573] 41.8 41.8 41.9 41.9 42 ...  
## $ end\_lng : num [1:131573] -87.6 -87.6 -87.6 -87.6 -87.7 ...  
## $ member\_casual : chr [1:131573] "member" "casual" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2101)

## spec\_tbl\_df [96,834 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:96834] "A3F8D895163BBB49" "0D139A3203274B87" "C7AE8E9CDB197A8E" "3097EF26414C7016" ...  
## $ rideable\_type : chr [1:96834] "electric\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:96834], format: "2021-01-01 00:02:00" "2021-01-01 00:02:00" ...  
## $ ended\_at : POSIXct[1:96834], format: "2021-01-01 00:12:00" "2021-01-01 00:08:00" ...  
## $ start\_station\_name: chr [1:96834] NA "State St & 33rd St" "Lakeview Ave & Fullerton Pkwy" "Montrose Harbor" ...  
## $ start\_station\_id : chr [1:96834] NA "13216" "TA1309000019" "TA1308000012" ...  
## $ end\_station\_name : chr [1:96834] NA "MLK Jr Dr & 29th St" "Ritchie Ct & Banks St" "Montrose Harbor" ...  
## $ end\_station\_id : chr [1:96834] NA "TA1307000139" "KA1504000134" "TA1308000012" ...  
## $ start\_lat : num [1:96834] 42 41.8 41.9 42 42 ...  
## $ start\_lng : num [1:96834] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:96834] 42 41.8 41.9 42 42 ...  
## $ end\_lng : num [1:96834] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:96834] "member" "member" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2102)

## spec\_tbl\_df [49,622 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:49622] "19A112436EC92A7E" "05A7BDE8A0D909FC" "CAC54BCCD36F6B15" "B12AD1843BE99A2C" ...  
## $ rideable\_type : chr [1:49622] "classic\_bike" "electric\_bike" "electric\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:49622], format: "2021-02-01 00:55:00" "2021-02-01 01:07:00" ...  
## $ ended\_at : POSIXct[1:49622], format: "2021-02-01 10:44:00" "2021-02-01 01:47:00" ...  
## $ start\_station\_name: chr [1:49622] "Kimbark Ave & 53rd St" "California Ave & Fletcher St" "Canal St & Adams St" "Canal St & Adams St" ...  
## $ start\_station\_id : chr [1:49622] "TA1309000037" "15642" "13011" "13011" ...  
## $ end\_station\_name : chr [1:49622] NA "Maplewood Ave & Peterson Ave" NA "Desplaines St & Jackson Blvd" ...  
## $ end\_station\_id : chr [1:49622] NA "KA1504000130" NA "15539" ...  
## $ start\_lat : num [1:49622] 41.8 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:49622] -87.6 -87.7 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:49622] NA 42 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:49622] NA -87.7 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:49622] "member" "casual" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2103)

## spec\_tbl\_df [228,496 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:228496] "A4F3120AD663324C" "30B35AAB2C588648" "CE1DDE1F3CAD840D" "C2310B7A78A87BD2" ...  
## $ rideable\_type : chr [1:228496] "classic\_bike" "electric\_bike" "classic\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:228496], format: "2021-03-01 00:01:00" "2021-03-01 00:05:00" ...  
## $ ended\_at : POSIXct[1:228496], format: "2021-03-01 02:27:00" "2021-03-01 00:06:00" ...  
## $ start\_station\_name: chr [1:228496] "Peoria St & Jackson Blvd" "Western Ave & Howard St" "LaSalle St & Illinois St" "Western Ave & Howard St" ...  
## $ start\_station\_id : chr [1:228496] "13158" "527" "13430" "527" ...  
## $ end\_station\_name : chr [1:228496] "Wood St & Taylor St (Temp)" "Western Ave & Howard St" "Wentworth Ave & Cermak Rd" "Campbell Ave & Fullerton Ave" ...  
## $ end\_station\_id : chr [1:228496] "13285" "527" "13075" "15648" ...  
## $ start\_lat : num [1:228496] 41.9 42 41.9 42 41.9 ...  
## $ start\_lng : num [1:228496] -87.6 -87.7 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num [1:228496] 41.9 42 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:228496] "casual" "member" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2104)

## spec\_tbl\_df [337,230 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:337230] "0DA46DD62F111D09" "FF009F812ADAFB3E" "8089677A9443153D" "623C22F8DEC21AE3" ...  
## $ rideable\_type : chr [1:337230] "electric\_bike" "classic\_bike" "classic\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:337230], format: "2021-04-01 00:03:00" "2021-04-01 00:03:00" ...  
## $ ended\_at : POSIXct[1:337230], format: "2021-04-01 00:30:00" "2021-04-01 00:23:00" ...  
## $ start\_station\_name: chr [1:337230] "Western Ave & Howard St" "LaSalle St & Illinois St" "Wabash Ave & Wacker Pl" NA ...  
## $ start\_station\_id : chr [1:337230] "527" "13430" "TA1307000131" NA ...  
## $ end\_station\_name : chr [1:337230] "Campbell Ave & Fullerton Ave" "Wentworth Ave & 24th St (Temp)" "Dearborn St & Erie St" NA ...  
## $ end\_station\_id : chr [1:337230] "15648" "TA1308000026" "13045" NA ...  
## $ start\_lat : num [1:337230] 42 41.9 41.9 41.7 42 ...  
## $ start\_lng : num [1:337230] -87.7 -87.6 -87.6 -87.6 -87.7 ...  
## $ end\_lat : num [1:337230] 41.9 41.9 41.9 41.7 42 ...  
## $ end\_lng : num [1:337230] -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:337230] "member" "member" "member" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2105)

## spec\_tbl\_df [531,633 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:531633] "015AD83E58838178" "850563C43E02CBAC" "20252EF674FE87AA" "70F741E61F1342B5" ...  
## $ rideable\_type : chr [1:531633] "classic\_bike" "classic\_bike" "docked\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:531633], format: "2021-05-01 00:00:00" "2021-05-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:531633], format: "2021-05-01 00:12:00" "2021-05-01 00:04:00" ...  
## $ start\_station\_name: chr [1:531633] "LaSalle St & Illinois St" "Halsted St & Clybourn Ave" "Racine Ave & Fullerton Ave" "Larrabee St & Menomonee St" ...  
## $ start\_station\_id : chr [1:531633] "13430" "331" "TA1306000026" "TA1306000007" ...  
## $ end\_station\_name : chr [1:531633] "Racine Ave & Randolph St" "Larrabee St & Division St" "Lake Shore Dr & North Blvd" "Halsted St & Wrightwood Ave" ...  
## $ end\_station\_id : chr [1:531633] "13155" "KA1504000079" "LF-005" "TA1309000061" ...  
## $ start\_lat : num [1:531633] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:531633] -87.6 -87.6 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:531633] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:531633] -87.7 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr [1:531633] "member" "member" "casual" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2106)

## spec\_tbl\_df [729,595 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:729595] "91DD8E98C53BA95B" "AA39E61C85E5B4E9" "07DDAE10F8C4A9B7" "0E9702B5AFFC73D0" ...  
## $ rideable\_type : chr [1:729595] "docked\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:729595], format: "2021-06-01 00:00:00" "2021-06-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:729595], format: "2021-06-01 00:41:00" "2021-06-01 00:20:00" ...  
## $ start\_station\_name: chr [1:729595] "Lake Shore Dr & Monroe St" NA "Wabash Ave & 9th St" "Perry Ave & 69th St" ...  
## $ start\_station\_id : chr [1:729595] "13300" NA "TA1309000010" "KA1503000047" ...  
## $ end\_station\_name : chr [1:729595] "Streeter Dr & Grand Ave" NA "Wabash Ave & 9th St" "May St & 69th St" ...  
## $ end\_station\_id : chr [1:729595] "13022" NA "TA1309000010" "567" ...  
## $ start\_lat : num [1:729595] 41.9 42 41.9 41.8 41.9 ...  
## $ start\_lng : num [1:729595] -87.6 -87.7 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:729595] 41.9 41.9 41.9 41.8 41.9 ...  
## $ end\_lng : num [1:729595] -87.6 -87.7 -87.6 -87.7 -87.6 ...  
## $ member\_casual : chr [1:729595] "casual" "member" "casual" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2107)

## spec\_tbl\_df [822,410 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:822410] "B8BA05857944F445" "B68C8C61487B5C15" "3FBF4B489DE3F24E" "8A1D30385234151D" ...  
## $ rideable\_type : chr [1:822410] "electric\_bike" "classic\_bike" "docked\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:822410], format: "2021-07-01 00:00:00" "2021-07-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:822410], format: "2021-07-01 00:04:00" "2021-07-01 00:46:00" ...  
## $ start\_station\_name: chr [1:822410] NA "Bissell St & Armitage Ave" "Clark St & Winnemac Ave" "Michigan Ave & Lake St" ...  
## $ start\_station\_id : chr [1:822410] NA "13059" "TA1309000035" "TA1305000011" ...  
## $ end\_station\_name : chr [1:822410] "State St & 33rd St" "Lincoln Ave & Diversey Pkwy" "Broadway & Thorndale Ave" "Wabash Ave & Wacker Pl" ...  
## $ end\_station\_id : chr [1:822410] "13216" "TA1307000064" "15575" "TA1307000131" ...  
## $ start\_lat : num [1:822410] 41.8 41.9 42 41.9 41.9 ...  
## $ start\_lng : num [1:822410] -87.6 -87.7 -87.7 -87.6 -87.6 ...  
## $ end\_lat : num [1:822410] 41.8 41.9 42 41.9 41.9 ...  
## $ end\_lng : num [1:822410] -87.6 -87.7 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:822410] "member" "casual" "casual" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2108)

## spec\_tbl\_df [804,352 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:804352] "D6AC43863387B96A" "00015E7CD37F64C4" "3B9554C394136D13" "3F35D7C35A1DB902" ...  
## $ rideable\_type : chr [1:804352] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:804352], format: "2021-08-01 00:00:00" "2021-08-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:804352], format: "2021-08-01 00:13:00" "2021-08-01 00:05:00" ...  
## $ start\_station\_name: chr [1:804352] "Clark St & Wrightwood Ave" "Racine Ave & Fullerton Ave" "Wells St & Huron St" "Halsted St & Dickens Ave" ...  
## $ start\_station\_id : chr [1:804352] "TA1305000014" "TA1306000026" "TA1306000012" "13192" ...  
## $ end\_station\_name : chr [1:804352] "Ashland Ave & Wrightwood Ave" "Halsted St & Wrightwood Ave" "Wells St & Evergreen Ave" "Wilton Ave & Belmont Ave" ...  
## $ end\_station\_id : chr [1:804352] "13296" "TA1309000061" "TA1308000049" "TA1307000134" ...  
## $ start\_lat : num [1:804352] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:804352] -87.6 -87.7 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:804352] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:804352] -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:804352] "member" "member" "casual" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2109)

## spec\_tbl\_df [756,147 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:756147] "D3B0587EBDF843D5" "15D5830C421DD198" "997B20717E3C8E38" "880070A73B4EB011" ...  
## $ rideable\_type : chr [1:756147] "classic\_bike" "electric\_bike" "electric\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:756147], format: "2021-09-01 00:00:00" "2021-09-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:756147], format: "2021-09-01 00:03:00" "2021-09-01 00:00:00" ...  
## $ start\_station\_name: chr [1:756147] "Clifton Ave & Armitage Ave" NA "Desplaines St & Jackson Blvd" "Racine Ave & Randolph St" ...  
## $ start\_station\_id : chr [1:756147] "TA1307000163" NA "15539" "13155" ...  
## $ end\_station\_name : chr [1:756147] "Southport Ave & Clybourn Ave" NA "Clark St & North Ave" "Ashland Ave & Lake St" ...  
## $ end\_station\_id : chr [1:756147] "TA1309000030" NA "13128" "13073" ...  
## $ start\_lat : num [1:756147] 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num [1:756147] -87.7 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num [1:756147] 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num [1:756147] -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr [1:756147] "member" "casual" "casual" "casual" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(r2110)

## spec\_tbl\_df [631,226 x 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:631226] "ABA2BDC3595FC3E9" "0BE9C131A5705D92" "74483AC18C8C6B90" "F8DBF095F01E1B68" ...  
## $ rideable\_type : chr [1:631226] "classic\_bike" "classic\_bike" "classic\_bike" "classic\_bike" ...  
## $ started\_at : POSIXct[1:631226], format: "2021-10-01 00:00:00" "2021-10-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:631226], format: "2021-10-01 00:10:00" "2021-10-01 00:05:00" ...  
## $ start\_station\_name: chr [1:631226] "Morgan St & Lake St" "Damen Ave & Cortland St" "Halsted St & Roscoe St" "Pine Grove Ave & Irving Park Rd" ...  
## $ start\_station\_id : chr [1:631226] "TA1306000015" "13133" "TA1309000025" "TA1308000022" ...  
## $ end\_station\_name : chr [1:631226] "Noble St & Milwaukee Ave" "Winchester Ave & Elston Ave" "Greenview Ave & Diversey Pkwy" "Clarendon Ave & Junior Ter" ...  
## $ end\_station\_id : chr [1:631226] "13290" "KA1504000140" "13294" "13389" ...  
## $ start\_lat : num [1:631226] 41.9 41.9 41.9 42 41.8 ...  
## $ start\_lng : num [1:631226] -87.7 -87.7 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num [1:631226] 41.9 41.9 41.9 42 41.9 ...  
## $ end\_lng : num [1:631226] -87.7 -87.7 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr [1:631226] "casual" "casual" "casual" "member" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. ride\_id = col\_character(),  
## .. rideable\_type = col\_character(),  
## .. started\_at = col\_datetime(format = ""),  
## .. ended\_at = col\_datetime(format = ""),  
## .. start\_station\_name = col\_character(),  
## .. start\_station\_id = col\_character(),  
## .. end\_station\_name = col\_character(),  
## .. end\_station\_id = col\_character(),  
## .. start\_lat = col\_double(),  
## .. start\_lng = col\_double(),  
## .. end\_lat = col\_double(),  
## .. end\_lng = col\_double(),  
## .. member\_casual = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

#### \* Correcting the format of the odd columns:

r2011 <- mutate(r2011, end\_station\_id= as.character(end\_station\_id),  
 start\_station\_id= as.character(start\_station\_id))

#### \* Rechecking the format of the columns:

compare\_df\_cols(r2011, r2012, r2101, r2102, r2103, r2104, r2105, r2106, r2107,  
 r2108, r2109, r2110, return = "mismatch")

## [1] column\_name r2011 r2012 r2101 r2102 r2103   
## [7] r2104 r2105 r2106 r2107 r2108 r2109   
## [13] r2110   
## <0 rows> (or 0-length row.names)

### \* Converting the 12 dataftame into one by binding the rows:

year\_bikeride <- bind\_rows(r2011, r2012, r2101, r2102, r2103, r2104, r2105,  
 r2106, r2107, r2108, r2109, r2110)

## \* STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS:

\*Remove the unnecessary columns:

year\_bikeride<- year\_bikeride %>%   
 select(-c(start\_lat, start\_lng, end\_lat, end\_lng))

#### \* Check the summary of new data frame:

summary(year\_bikeride)

## ride\_id rideable\_type started\_at   
## Length:5378834 Length:5378834 Min. :2020-11-01 00:00:00   
## Class :character Class :character 1st Qu.:2021-05-17 12:45:00   
## Mode :character Mode :character Median :2021-07-13 22:33:00   
## Mean :2021-06-27 18:37:11   
## 3rd Qu.:2021-09-02 18:18:00   
## Max. :2021-10-31 23:59:00   
## ended\_at start\_station\_name start\_station\_id   
## Min. :2020-11-01 00:02:00 Length:5378834 Length:5378834   
## 1st Qu.:2021-05-17 13:07:00 Class :character Class :character   
## Median :2021-07-13 22:57:00 Mode :character Mode :character   
## Mean :2021-06-27 18:57:41   
## 3rd Qu.:2021-09-02 18:35:00   
## Max. :2021-11-03 21:45:00   
## end\_station\_name end\_station\_id member\_casual   
## Length:5378834 Length:5378834 Length:5378834   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

#### \* Dividing the date, month, year and day columns:

year\_bikeride$date <- as.Date(year\_bikeride$started\_at)  
year\_bikeride$month <- format(as.Date(year\_bikeride$started\_at), "%m")  
year\_bikeride$day <- format(as.Date(year\_bikeride$started\_at), "%d")  
year\_bikeride$year <- format(as.Date(year\_bikeride$started\_at), "%Y")  
year\_bikeride$day\_of\_week <- format(as.Date(year\_bikeride$started\_at), "%A")

#### \* Changing the month from number to abbreviated name:

year\_bikeride<- year\_bikeride %>%   
 mutate(month = month.abb[as.numeric(month)])

#### \* Finding out the journey time by differentiating the starting time and ending time of the ride:

year\_bikeride$ride\_length <- difftime(year\_bikeride$ended\_at,year\_bikeride$started\_at)

#### \* Checking the new data frame with additional columns:

str(year\_bikeride)

## tibble [5,378,834 x 15] (S3: tbl\_df/tbl/data.frame)  
## $ ride\_id : chr [1:5378834] "D0C1A65AFC859916" "985AD418D8ED316C" "AB459A840CF1BBF8" "5239202FF2A38305" ...  
## $ rideable\_type : chr [1:5378834] "docked\_bike" "docked\_bike" "docked\_bike" "electric\_bike" ...  
## $ started\_at : POSIXct[1:5378834], format: "2020-11-01 00:00:00" "2020-11-01 00:00:00" ...  
## $ ended\_at : POSIXct[1:5378834], format: "2020-11-01 00:35:00" "2020-11-01 00:02:00" ...  
## $ start\_station\_name: chr [1:5378834] "Broadway & Barry Ave" "Desplaines St & Kinzie St" "Broadway & Cornelia Ave" "Wells St & Evergreen Ave" ...  
## $ start\_station\_id : chr [1:5378834] "300" "56" "303" "291" ...  
## $ end\_station\_name : chr [1:5378834] "Broadway & Barry Ave" "Desplaines St & Randolph St" "Sheridan Rd & Irving Park Rd" "Franklin St & Chicago Ave" ...  
## $ end\_station\_id : chr [1:5378834] "300" "96" "240" "31" ...  
## $ member\_casual : chr [1:5378834] "casual" "member" "member" "casual" ...  
## $ date : Date[1:5378834], format: "2020-11-01" "2020-11-01" ...  
## $ month : chr [1:5378834] "Nov" "Nov" "Nov" "Nov" ...  
## $ day : chr [1:5378834] "01" "01" "01" "01" ...  
## $ year : chr [1:5378834] "2020" "2020" "2020" "2020" ...  
## $ day\_of\_week : chr [1:5378834] "Sunday" "Sunday" "Sunday" "Sunday" ...  
## $ ride\_length : 'difftime' num [1:5378834] 2100 120 360 240 ...  
## ..- attr(\*, "units")= chr "secs"

head(year\_bikeride)

## # A tibble: 6 x 15  
## ride\_id rideable\_type started\_at ended\_at start\_station\_n~  
## <chr> <chr> <dttm> <dttm> <chr>   
## 1 D0C1A6~ docked\_bike 2020-11-01 00:00:00 2020-11-01 00:35:00 Broadway & Barr~  
## 2 985AD4~ docked\_bike 2020-11-01 00:00:00 2020-11-01 00:02:00 Desplaines St &~  
## 3 AB459A~ docked\_bike 2020-11-01 00:00:00 2020-11-01 00:06:00 Broadway & Corn~  
## 4 523920~ electric\_bike 2020-11-01 00:00:00 2020-11-01 00:04:00 Wells St & Ever~  
## 5 587849~ electric\_bike 2020-11-01 00:00:00 2020-11-01 00:04:00 Wells St & Ever~  
## 6 33A07F~ docked\_bike 2020-11-01 00:01:00 2020-11-01 00:09:00 Dearborn St & E~  
## # ... with 10 more variables: start\_station\_id <chr>, end\_station\_name <chr>,  
## # end\_station\_id <chr>, member\_casual <chr>, date <date>, month <chr>,  
## # day <chr>, year <chr>, day\_of\_week <chr>, ride\_length <drtn>

#### \* Checking the ride length format and changing it into numeric:

is.factor(year\_bikeride$ride\_length)

## [1] FALSE

year\_bikeride$ride\_length <- as.numeric(as.character(year\_bikeride$ride\_length))  
  
is.numeric(year\_bikeride$ride\_length)

## [1] TRUE

### \* Checking and removing “bad” data:

table(year\_bikeride$ride\_length<0)

##   
## FALSE TRUE   
## 5378100 734

year\_bikeride\_v2 <- year\_bikeride[!(year\_bikeride$ride\_length<0),]

## Analysis

### CONDUCT DESCRIPTIVE ANALYSIS:

\*Analysis on ride\_length(in second):

mean(year\_bikeride\_v2$ride\_length)

## [1] 1350.136

median(year\_bikeride\_v2$ride\_length)

## [1] 720

max(year\_bikeride\_v2$ride\_length)

## [1] 3356640

min(year\_bikeride\_v2$ride\_length)

## [1] 0

#### \* Compare members and casual riders:

\*Comparison in members and casual risers on ride\_length(in second):

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual, FUN= mean)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$ride\_length  
## 1 casual 1953.3318  
## 2 member 837.6711

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual, FUN= median)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$ride\_length  
## 1 casual 960  
## 2 member 600

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual, FUN= max)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$ride\_length  
## 1 casual 3356640  
## 2 member 93600

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual, FUN= min)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$ride\_length  
## 1 casual 0  
## 2 member 0

#### \* Arranging the sequence of the days of the week:

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual+year\_bikeride\_v2$day\_of\_week, FUN= mean)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$day\_of\_week  
## 1 casual Friday  
## 2 member Friday  
## 3 casual Monday  
## 4 member Monday  
## 5 casual Saturday  
## 6 member Saturday  
## 7 casual Sunday  
## 8 member Sunday  
## 9 casual Thursday  
## 10 member Thursday  
## 11 casual Tuesday  
## 12 member Tuesday  
## 13 casual Wednesday  
## 14 member Wednesday  
## year\_bikeride\_v2$ride\_length  
## 1 1854.4354  
## 2 819.5990  
## 3 1938.9977  
## 4 811.0220  
## 5 2109.3536  
## 6 933.3074  
## 7 2286.2175  
## 8 955.0786  
## 9 1692.2934  
## 10 786.0227  
## 11 1717.1691  
## 12 786.2029  
## 13 1688.9647  
## 14 789.8731

year\_bikeride\_v2$day\_of\_week <- ordered(year\_bikeride\_v2$day\_of\_week, levels= c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday"))  
  
aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual+year\_bikeride\_v2$day\_of\_week, FUN= mean)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$day\_of\_week  
## 1 casual Sunday  
## 2 member Sunday  
## 3 casual Monday  
## 4 member Monday  
## 5 casual Tuesday  
## 6 member Tuesday  
## 7 casual Wednesday  
## 8 member Wednesday  
## 9 casual Thursday  
## 10 member Thursday  
## 11 casual Friday  
## 12 member Friday  
## 13 casual Saturday  
## 14 member Saturday  
## year\_bikeride\_v2$ride\_length  
## 1 2286.2175  
## 2 955.0786  
## 3 1938.9977  
## 4 811.0220  
## 5 1717.1691  
## 6 786.2029  
## 7 1688.9647  
## 8 789.8731  
## 9 1692.2934  
## 10 786.0227  
## 11 1854.4354  
## 12 819.5990  
## 13 2109.3536  
## 14 933.3074

#### \* Arranging the sequence of the month:

aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual+year\_bikeride\_v2$month, FUN= mean)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$month  
## 1 casual Apr  
## 2 member Apr  
## 3 casual Aug  
## 4 member Aug  
## 5 casual Dec  
## 6 member Dec  
## 7 casual Feb  
## 8 member Feb  
## 9 casual Jan  
## 10 member Jan  
## 11 casual Jul  
## 12 member Jul  
## 13 casual Jun  
## 14 member Jun  
## 15 casual Mar  
## 16 member Mar  
## 17 casual May  
## 18 member May  
## 19 casual Nov  
## 20 member Nov  
## 21 casual Oct  
## 22 member Oct  
## 23 casual Sep  
## 24 member Sep  
## year\_bikeride\_v2$ride\_length  
## 1 2281.3559  
## 2 881.3960  
## 3 1727.1880  
## 4 846.0600  
## 5 1610.8440  
## 6 764.6700  
## 7 2962.3275  
## 8 1081.2945  
## 9 1541.0830  
## 10 772.3189  
## 11 1967.4182  
## 12 854.3795  
## 13 2227.2776  
## 14 880.6820  
## 15 2289.6369  
## 16 838.1514  
## 17 2293.8561  
## 18 878.3820  
## 19 1906.5047  
## 20 813.2745  
## 21 1720.4843  
## 22 750.1765  
## 23 1669.0003  
## 24 824.1791

year\_bikeride\_v2$month <-  
 ordered(year\_bikeride\_v2$month, levels= c("Jan","Feb","Mar","Apr","May","Jun",  
 "Jul","Aug","Sep","Oct","Nov","Dec"))  
aggregate(year\_bikeride\_v2$ride\_length ~ year\_bikeride\_v2$member\_casual+year\_bikeride\_v2$month, FUN= mean)

## year\_bikeride\_v2$member\_casual year\_bikeride\_v2$month  
## 1 casual Jan  
## 2 member Jan  
## 3 casual Feb  
## 4 member Feb  
## 5 casual Mar  
## 6 member Mar  
## 7 casual Apr  
## 8 member Apr  
## 9 casual May  
## 10 member May  
## 11 casual Jun  
## 12 member Jun  
## 13 casual Jul  
## 14 member Jul  
## 15 casual Aug  
## 16 member Aug  
## 17 casual Sep  
## 18 member Sep  
## 19 casual Oct  
## 20 member Oct  
## 21 casual Nov  
## 22 member Nov  
## 23 casual Dec  
## 24 member Dec  
## year\_bikeride\_v2$ride\_length  
## 1 1541.0830  
## 2 772.3189  
## 3 2962.3275  
## 4 1081.2945  
## 5 2289.6369  
## 6 838.1514  
## 7 2281.3559  
## 8 881.3960  
## 9 2293.8561  
## 10 878.3820  
## 11 2227.2776  
## 12 880.6820  
## 13 1967.4182  
## 14 854.3795  
## 15 1727.1880  
## 16 846.0600  
## 17 1669.0003  
## 18 824.1791  
## 19 1720.4843  
## 20 750.1765  
## 21 1906.5047  
## 22 813.2745  
## 23 1610.8440  
## 24 764.6700

#### \* Finding the number of casual and member rides:

table(year\_bikeride$member\_casual)

##   
## casual member   
## 2470517 2908317

#### \* Finding the relation between the rider type, average ride duration, number of rides and the day of the week:

year\_bikeride\_v2 %>%   
 mutate(weekday= wday(started\_at, label= TRUE)) %>%   
 group\_by(member\_casual,weekday) %>%   
 summarise(number\_of\_rides= n(), average\_duration= mean(ride\_length)) %>%   
 arrange(member\_casual,weekday)

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

## # A tibble: 14 x 4  
## # Groups: member\_casual [2]  
## member\_casual weekday number\_of\_rides average\_duration  
## <chr> <ord> <int> <dbl>  
## 1 casual Sun 476167 2286.  
## 2 casual Mon 278266 1939.  
## 3 casual Tue 264266 1717.  
## 4 casual Wed 267481 1689.  
## 5 casual Thu 277330 1692.  
## 6 casual Fri 354988 1854.  
## 7 casual Sat 551867 2109.  
## 8 member Sun 368442 955.  
## 9 member Mon 391354 811.  
## 10 member Tue 431509 786.  
## 11 member Wed 444351 790.  
## 12 member Thu 425623 786.  
## 13 member Fri 425209 820.  
## 14 member Sat 421247 933.

#### \* Finding the relation between the rider type, average ride duration, number of rides and the months:

year\_bikeride\_v2 %>%   
 group\_by(member\_casual,month) %>%   
 summarise(number\_of\_rides= n(), average\_duration= mean(ride\_length)) %>%   
 arrange(member\_casual, month)

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

## # A tibble: 24 x 4  
## # Groups: member\_casual [2]  
## member\_casual month number\_of\_rides average\_duration  
## <chr> <ord> <int> <dbl>  
## 1 casual Jan 18117 1541.  
## 2 casual Feb 10131 2962.  
## 3 casual Mar 84033 2290.  
## 4 casual Apr 136601 2281.  
## 5 casual May 256916 2294.  
## 6 casual Jun 370680 2227.  
## 7 casual Jul 442056 1967.  
## 8 casual Aug 412666 1727.  
## 9 casual Sep 363888 1669.  
## 10 casual Oct 257242 1720.  
## # ... with 14 more rows

#### \* Creating a different data frame for the “casual rider”:

casual\_rider\_data<- filter(year\_bikeride\_v2, member\_casual == "casual")

#### \* Getting elaborate data about casual rider:

summary(casual\_rider\_data)

## ride\_id rideable\_type started\_at   
## Length:2470365 Length:2470365 Min. :2020-11-01 00:00:00   
## Class :character Class :character 1st Qu.:2021-05-31 14:52:00   
## Mode :character Mode :character Median :2021-07-18 12:42:00   
## Mean :2021-07-06 17:47:20   
## 3rd Qu.:2021-09-01 14:35:00   
## Max. :2021-10-31 23:57:00   
##   
## ended\_at start\_station\_name start\_station\_id   
## Min. :2020-11-01 00:04:00 Length:2470365 Length:2470365   
## 1st Qu.:2021-05-31 15:33:00 Class :character Class :character   
## Median :2021-07-18 13:16:00 Mode :character Mode :character   
## Mean :2021-07-06 18:19:53   
## 3rd Qu.:2021-09-01 15:01:00   
## Max. :2021-11-03 21:45:00   
##   
## end\_station\_name end\_station\_id member\_casual date   
## Length:2470365 Length:2470365 Length:2470365 Min. :2020-11-01   
## Class :character Class :character Class :character 1st Qu.:2021-05-31   
## Mode :character Mode :character Mode :character Median :2021-07-18   
## Mean :2021-07-06   
## 3rd Qu.:2021-09-01   
## Max. :2021-10-31   
##   
## month day year day\_of\_week   
## Jul :442056 Length:2470365 Length:2470365 Sunday :476167   
## Aug :412666 Class :character Class :character Monday :278266   
## Jun :370680 Mode :character Mode :character Tuesday :264266   
## Sep :363888 Wednesday:267481   
## Oct :257242 Thursday :277330   
## May :256916 Friday :354988   
## (Other):366917 Saturday :551867   
## ride\_length   
## Min. : 0   
## 1st Qu.: 540   
## Median : 960   
## Mean : 1953   
## 3rd Qu.: 1800   
## Max. :3356640   
##

#### \* Getting the specific data about the different types of bike preference by the “casual rider”:

table(casual\_rider\_data$rideable\_type)

##   
## classic\_bike docked\_bike electric\_bike   
## 1226300 350700 893365

#### \* 10 most popular stations from where the casual riders had started their journies:

drop\_na(head(count(casual\_rider\_data,start\_station\_name, sort = T), n=11),start\_station\_name)

## # A tibble: 10 x 2  
## start\_station\_name n  
## <chr> <int>  
## 1 Streeter Dr & Grand Ave 65019  
## 2 Millennium Park 33331  
## 3 Michigan Ave & Oak St 29988  
## 4 Shedd Aquarium 22521  
## 5 Theater on the Lake 21854  
## 6 Lake Shore Dr & Monroe St 21385  
## 7 Wells St & Concord Ln 19633  
## 8 Clark St & Lincoln Ave 16982  
## 9 Indiana Ave & Roosevelt Rd 16840  
## 10 Wells St & Elm St 16459

### **Core observation for “casual rider” from the analysis**

name\_of\_obsevation<- c("number\_of\_casual\_rides","average\_ride\_duration","max\_ride\_duration","most\_preferable\_ride\_type","more\_popular\_time\_of\_week","more\_popular\_time\_in\_year","most\_started\_journey\_station")  
result\_of\_observation<- c("2470517(45.93%)","(38.1-28.15)min","38.85 days","classic\_bike","Saturday,Saturday","summer months","Streeter Dr & Grand Ave")  
summary\_observation<- data.frame(name\_of\_obsevation,result\_of\_observation)  
head(summary\_observation,n=7)

## name\_of\_obsevation result\_of\_observation  
## 1 number\_of\_casual\_rides 2470517(45.93%)  
## 2 average\_ride\_duration (38.1-28.15)min  
## 3 max\_ride\_duration 38.85 days  
## 4 most\_preferable\_ride\_type classic\_bike  
## 5 more\_popular\_time\_of\_week Saturday,Saturday  
## 6 more\_popular\_time\_in\_year summer months  
## 7 most\_started\_journey\_station Streeter Dr & Grand Ave

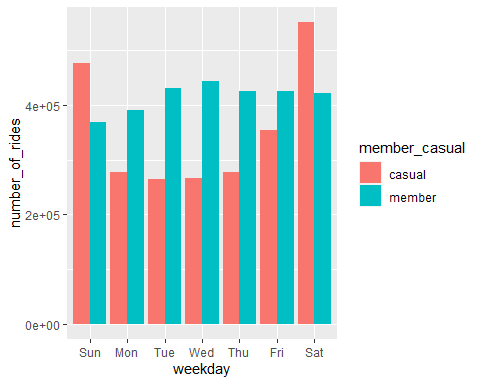
## Share

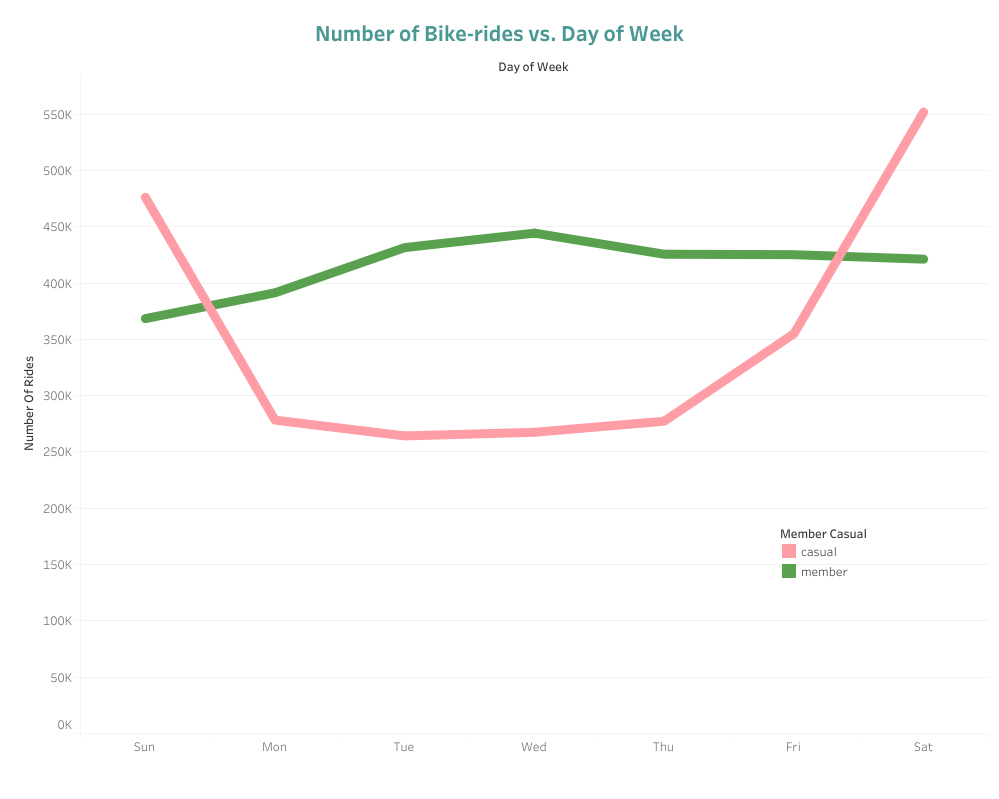
In this phase all the data that are derived through this analysis will be expressed through art of visualization.

### \* Day of the Week vs. Number of Rides Graph for Casual-riders and Member-riders

year\_bikeride\_v2 %>%   
 mutate(weekday= wday(started\_at, label= TRUE)) %>%   
 group\_by(member\_casual,weekday) %>%   
 summarise(number\_of\_rides= n(), average\_duration= mean(ride\_length)) %>%   
 arrange(member\_casual,weekday) %>%   
 ggplot(aes(x= weekday, y= number\_of\_rides, fill= member\_casual))+  
 geom\_col(position = "dodge")

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

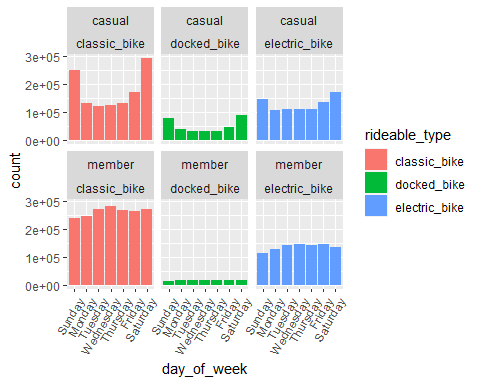




Line Graph : Day of the Week vs. Number of Rides Graph for Casual-riders and Member-riders

### \* Day of the Week vs. Number of Rides Graph for Casual-riders and Member-riders with different bike preferences

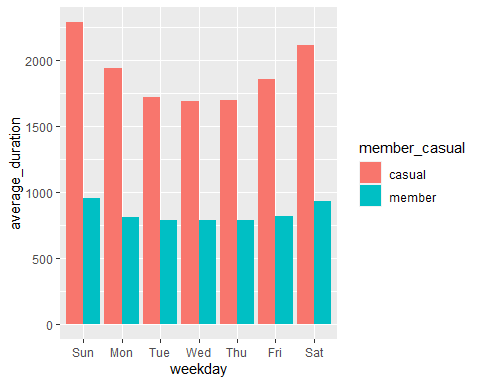
year\_bikeride\_v2 %>%   
 ggplot(aes(x= day\_of\_week, fill= rideable\_type))+  
 geom\_bar()+facet\_wrap(member\_casual~rideable\_type)+  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



### \* Day of the Week vs. Average duration of Ride Graph for Casual-riders and Member-riders

year\_bikeride\_v2 %>%   
 mutate(weekday= wday(started\_at, label= TRUE)) %>%   
 group\_by(member\_casual,weekday) %>%   
 summarise(number\_of\_rides= n(), average\_duration= mean(ride\_length)) %>%   
 arrange(member\_casual,weekday) %>%   
 ggplot(aes(x= weekday, y= average\_duration, fill= member\_casual))+  
 geom\_col(position = "dodge")

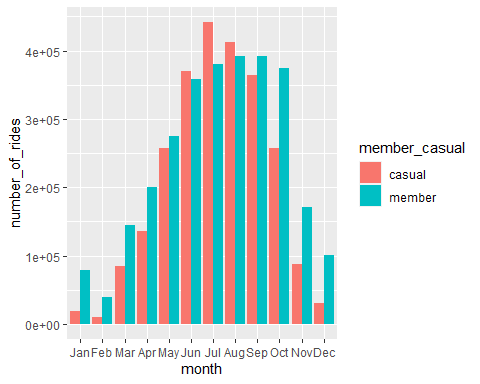
## `summarise()` has grouped output by 'member\_casual'. You can override using the `.groups` argument.

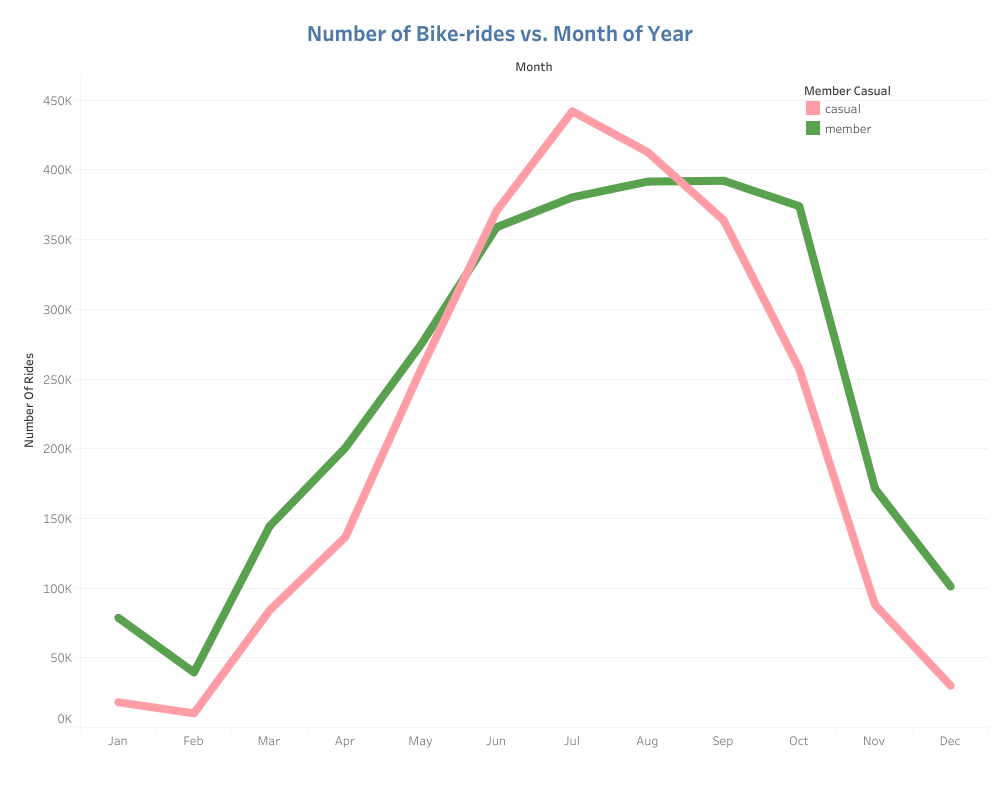


### \* Month vs. Number of Rides Graph for Casual-riders and Member-riders

year\_bikeride\_v2 %>%  
 group\_by(member\_casual,month) %>%   
 summarise(number\_of\_rides= n(), average\_duration= mean(ride\_length)) %>%   
 arrange(member\_casual,month) %>%   
 ggplot(aes(x= month, y= number\_of\_rides, fill= member\_casual))+  
 geom\_col(position = "dodge")

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

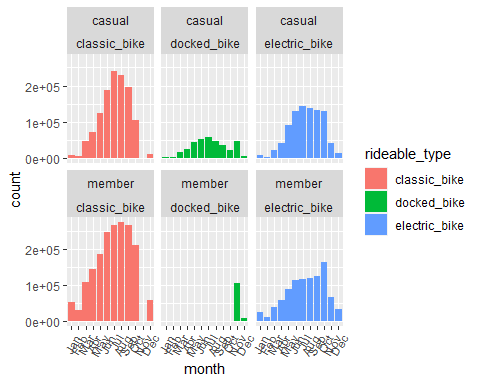




Line Graph : Month vs. Number of Rides Graph for Casual-riders and Member-riders

### \* Month vs. Number of Rides Graph for Casual-riders and Member-riders with different bike preferences

year\_bikeride\_v2 %>%   
 ggplot(aes(x= month, fill= rideable\_type))+  
 geom\_bar()+facet\_wrap(member\_casual~rideable\_type)+  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



## \* Conclusion:

#### *By looking at the business goal, this analysis has been conducted and here some key factors that is coming out of the process.*

#### \* 1.The casual users and the member users riding parttern is very different form one another. The casual users usually ride more in the weekends while the members almost ride equally throughout the week.

#### \* 2.As the months of winter approaches, both the casual and members bike ridings decreases drastically.

#### \* 3.Casual riders have a tendency of riding long hours than the members.

#### \* 4.Some stations are more popular in the casual riders.

#### \* 5.Casual riders mostly prefer to ride by “classic bike” and then “electric bike” and “docked bike” is least favourite among them.

#### ……………………………………………………………………………………………………………………..

#### *The above-mentioned factors are the key to meet the business goal. The casual riders riding patterns should be analyse in order to convert the casual riders into members.*

#### The number of bikes should be increase in weekends to meet the need of the casual riders as in the weekends the casual riding increases drastically.

#### As the electrical bike and the docked bike is more popular in the casual riders, therefore we can increase the price of usage of casual riding and can offer some membership benefits. So more casual members would be interested to become a member.

#### As the summer months are the popular time to the bike riders and moreover the casual riders usually like to take a long duration trip, more membership benefits can be offered for those months for long duration bike ride.

#### And lastly, we can use the digital media in the more popular stations for advertising the offers and benefits and the company name.

Note: *The data is data is limited in some way.Firstly, for some unknown reason the data for electric bike is not complete. Secondly, I don’t have the personal data about the rider so we cannot conclude whether the bike riding are enjoyed by unique riders or not. This unavailable data can affect the analysis.*