Bikeshare

Pyimoe Than

6/17/2022

##Business problem

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

## Importing library

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(writexl)  
library(tidyr)  
library(ggplot2)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ tibble 3.1.7 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ✔ purrr 0.3.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(data.table)

##   
## Attaching package: 'data.table'

## The following object is masked from 'package:purrr':  
##   
## transpose

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(lubridate)#for datetime object

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':  
##   
## hour, isoweek, mday, minute, month, quarter, second, wday, week,  
## yday, year

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

## Load the data

Bike\_Share\_202004=read.csv("202004-divvy-tripdata.csv",header=T)  
Bike\_Share\_202005=read.csv("202005-divvy-tripdata.csv",header=T)  
Bike\_Share\_202006=read.csv("202006-divvy-tripdata.csv",header=T)  
Bike\_Share\_202007=read.csv("202007-divvy-tripdata.csv",header=T)  
Bike\_Share\_202008=read.csv("202008-divvy-tripdata.csv",header=T)  
Bike\_Share\_202009=read.csv("202009-divvy-tripdata.csv",header=T)  
Bike\_Share\_202010=read.csv("202010-divvy-tripdata.csv",header=T)  
Bike\_Share\_202011=read.csv("202011-divvy-tripdata.csv",header=T)  
Bike\_Share\_202012=read.csv("202012-divvy-tripdata.csv",header=T)

## Total Number of Columns and Width

dim(Bike\_Share\_202004)

## [1] 84776 13

dim(Bike\_Share\_202005)

## [1] 200274 13

dim(Bike\_Share\_202006)

## [1] 343005 13

dim(Bike\_Share\_202007)

## [1] 551480 13

dim(Bike\_Share\_202008)

## [1] 622361 13

dim(Bike\_Share\_202009)

## [1] 532958 13

dim(Bike\_Share\_202010)

## [1] 388653 13

dim(Bike\_Share\_202011)

## [1] 259716 13

dim(Bike\_Share\_202012)

## [1] 131573 13

## Check Column name of each dataset for consistency

colnames(Bike\_Share\_202004)

## [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(Bike\_Share\_202005)

## [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(Bike\_Share\_202006)

## [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(Bike\_Share\_202007)

## [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(Bike\_Share\_202008)

## [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(Bike\_Share\_202009)

## [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(Bike\_Share\_202010)

## [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(Bike\_Share\_202011)

## [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(Bike\_Share\_202012)

## [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"

## What kinds of Data we have in the dataset

str(Bike\_Share\_202004)

## 'data.frame': 84776 obs. of 13 variables:  
## $ ride\_id : chr "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...  
## $ rideable\_type : chr "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : chr "2020-04-26 17:45:14" "2020-04-17 17:08:54" "2020-04-01 17:54:13" "2020-04-07 12:50:19" ...  
## $ ended\_at : chr "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 18:08:36" "2020-04-07 13:02:31" ...  
## $ start\_station\_name: chr "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...  
## $ start\_station\_id : int 86 503 142 216 125 173 35 434 627 377 ...  
## $ end\_station\_name : chr "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...  
## $ end\_station\_id : int 152 499 255 657 323 35 635 382 359 508 ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.7 -87.7 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 42 ...  
## $ end\_lng : num -87.7 -87.7 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr "member" "member" "member" "member" ...

str(Bike\_Share\_202005)

## 'data.frame': 200274 obs. of 13 variables:  
## $ ride\_id : chr "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA84" ...  
## $ rideable\_type : chr "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : chr "2020-05-27 10:03:52" "2020-05-25 10:47:11" "2020-05-02 14:11:03" "2020-05-02 16:25:36" ...  
## $ ended\_at : chr "2020-05-27 10:16:49" "2020-05-25 11:05:40" "2020-05-02 15:48:21" "2020-05-02 16:39:28" ...  
## $ start\_station\_name: chr "Franklin St & Jackson Blvd" "Clark St & Wrightwood Ave" "Kedzie Ave & Milwaukee Ave" "Clarendon Ave & Leland Ave" ...  
## $ start\_station\_id : int 36 340 260 251 261 206 261 180 331 219 ...  
## $ end\_station\_name : chr "Wabash Ave & Grand Ave" "Clark St & Leland Ave" "Kedzie Ave & Milwaukee Ave" "Lake Shore Dr & Wellington Ave" ...  
## $ end\_station\_id : int 199 326 260 157 206 22 261 180 300 305 ...  
## $ start\_lat : num 41.9 41.9 41.9 42 41.9 ...  
## $ start\_lng : num -87.6 -87.6 -87.7 -87.7 -87.7 ...  
## $ end\_lat : num 41.9 42 41.9 41.9 41.8 ...  
## $ end\_lng : num -87.6 -87.7 -87.7 -87.6 -87.6 ...  
## $ member\_casual : chr "member" "casual" "casual" "casual" ...

str(Bike\_Share\_202006)

## 'data.frame': 343005 obs. of 13 variables:  
## $ ride\_id : chr "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85FA1" ...  
## $ rideable\_type : chr "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : chr "2020-06-13 23:24:48" "2020-06-26 07:26:10" "2020-06-23 17:12:41" "2020-06-20 01:09:35" ...  
## $ ended\_at : chr "2020-06-13 23:36:55" "2020-06-26 07:31:58" "2020-06-23 17:21:14" "2020-06-20 01:28:24" ...  
## $ start\_station\_name: chr "Wilton Ave & Belmont Ave" "Federal St & Polk St" "Daley Center Plaza" "Broadway & Cornelia Ave" ...  
## $ start\_station\_id : int 117 41 81 303 327 327 41 115 338 84 ...  
## $ end\_station\_name : chr "Damen Ave & Clybourn Ave" "Daley Center Plaza" "State St & Harrison St" "Broadway & Berwyn Ave" ...  
## $ end\_station\_id : int 163 81 5 294 117 117 81 303 164 53 ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.7 -87.6 -87.6 -87.6 -87.7 ...  
## $ end\_lat : num 41.9 41.9 41.9 42 41.9 ...  
## $ end\_lng : num -87.7 -87.6 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr "casual" "member" "member" "casual" ...

str(Bike\_Share\_202007)

## 'data.frame': 551480 obs. of 13 variables:  
## $ ride\_id : chr "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...  
## $ rideable\_type : chr "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : chr "2020-07-09 15:22:02" "2020-07-24 23:56:30" "2020-07-08 19:49:07" "2020-07-17 19:06:42" ...  
## $ ended\_at : chr "2020-07-09 15:25:52" "2020-07-25 00:20:17" "2020-07-08 19:56:22" "2020-07-17 19:27:38" ...  
## $ start\_station\_name: chr "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "Lake Shore Dr & Diversey Pkwy" "LaSalle St & Illinois St" ...  
## $ start\_station\_id : int 180 299 329 181 268 635 113 211 176 31 ...  
## $ end\_station\_name : chr "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "Clark St & Wellington Ave" "Clark St & Armitage Ave" ...  
## $ end\_station\_id : int 291 461 156 94 301 289 140 31 191 142 ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ end\_lat : num 41.9 42 41.9 41.9 41.9 ...  
## $ end\_lng : num -87.6 -87.7 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr "member" "member" "casual" "casual" ...

str(Bike\_Share\_202008)

## 'data.frame': 622361 obs. of 13 variables:  
## $ ride\_id : chr "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79FBBD412E578A7" ...  
## $ rideable\_type : chr "docked\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2020-08-20 18:08:14" "2020-08-27 18:46:04" "2020-08-26 19:44:14" "2020-08-27 12:05:41" ...  
## $ ended\_at : chr "2020-08-20 18:17:51" "2020-08-27 19:54:51" "2020-08-26 21:53:07" "2020-08-27 12:53:45" ...  
## $ start\_station\_name: chr "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Columbus Dr & Randolph St" "Daley Center Plaza" ...  
## $ start\_station\_id : int 329 168 195 81 658 658 196 67 153 177 ...  
## $ end\_station\_name : chr "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St & Randolph St" "State St & Kinzie St" ...  
## $ end\_station\_id : int 141 168 44 47 658 658 49 229 225 305 ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.6 -87.6 -87.6 -87.6 -87.7 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num -87.6 -87.6 -87.6 -87.6 -87.7 ...  
## $ member\_casual : chr "member" "casual" "casual" "casual" ...

str(Bike\_Share\_202009)

## 'data.frame': 532958 obs. of 13 variables:  
## $ ride\_id : chr "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...  
## $ rideable\_type : chr "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2020-09-17 14:27:11" "2020-09-17 15:07:31" "2020-09-17 15:09:04" "2020-09-17 18:10:46" ...  
## $ ended\_at : chr "2020-09-17 14:44:24" "2020-09-17 15:07:45" "2020-09-17 15:09:35" "2020-09-17 18:35:49" ...  
## $ start\_station\_name: chr "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Ashland Ave & Belle Plaine Ave" ...  
## $ start\_station\_id : int 52 NA NA 246 24 94 291 NA NA NA ...  
## $ end\_station\_name : chr "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway" "Montrose Harbor" ...  
## $ end\_station\_id : int 112 NA NA 249 24 NA 256 NA NA NA ...  
## $ start\_lat : num 41.9 41.9 41.9 42 41.9 ...  
## $ start\_lng : num -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num 41.9 41.9 41.9 42 41.9 ...  
## $ end\_lng : num -87.6 -87.6 -87.6 -87.6 -87.6 ...  
## $ member\_casual : chr "casual" "casual" "casual" "casual" ...

str(Bike\_Share\_202010)

## 'data.frame': 388653 obs. of 13 variables:  
## $ ride\_id : chr "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE261B9E854" ...  
## $ rideable\_type : chr "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2020-10-31 19:39:43" "2020-10-31 23:50:08" "2020-10-31 23:00:01" "2020-10-31 22:16:43" ...  
## $ ended\_at : chr "2020-10-31 19:57:12" "2020-11-01 00:04:16" "2020-10-31 23:08:22" "2020-10-31 22:19:35" ...  
## $ start\_station\_name: chr "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave" "Stony Island Ave & 67th St" "Clark St & Grace St" ...  
## $ start\_station\_id : int 313 227 102 165 190 359 313 125 NA 174 ...  
## $ end\_station\_name : chr "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "University Ave & 57th St" "Broadway & Sheridan Rd" ...  
## $ end\_station\_id : int 125 260 423 256 185 53 125 313 199 635 ...  
## $ start\_lat : num 41.9 41.9 41.8 42 41.9 ...  
## $ start\_lng : num -87.6 -87.7 -87.6 -87.7 -87.7 ...  
## $ end\_lat : num 41.9 41.9 41.8 42 41.9 ...  
## $ end\_lng : num -87.6 -87.7 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr "casual" "casual" "casual" "casual" ...

str(Bike\_Share\_202011)

## 'data.frame': 259716 obs. of 13 variables:  
## $ ride\_id : chr "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533E89C32080B9E" ...  
## $ rideable\_type : chr "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2020-11-01 13:36:00" "2020-11-01 10:03:26" "2020-11-01 00:34:05" "2020-11-01 00:45:16" ...  
## $ ended\_at : chr "2020-11-01 13:45:40" "2020-11-01 10:14:45" "2020-11-01 01:03:06" "2020-11-01 00:54:31" ...  
## $ start\_station\_name: chr "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shore Dr & Monroe St" "Leavitt St & Chicago Ave" ...  
## $ start\_station\_id : int 110 672 76 659 2 72 76 NA 58 394 ...  
## $ end\_station\_name : chr "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal St & Polk St" "Stave St & Armitage Ave" ...  
## $ end\_station\_id : int 211 29 41 185 2 76 72 NA 288 273 ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.6 -87.6 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ end\_lng : num -87.6 -87.7 -87.6 -87.7 -87.6 ...  
## $ member\_casual : chr "casual" "casual" "casual" "casual" ...

str(Bike\_Share\_202012)

## 'data.frame': 131573 obs. of 13 variables:  
## $ ride\_id : chr "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE119628E44F871E" ...  
## $ rideable\_type : chr "classic\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...  
## $ started\_at : chr "2020-12-27 12:44:29" "2020-12-18 17:37:15" "2020-12-15 15:04:33" "2020-12-15 15:54:18" ...  
## $ ended\_at : chr "2020-12-27 12:55:06" "2020-12-18 17:44:19" "2020-12-15 15:11:28" "2020-12-15 16:00:11" ...  
## $ 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 : num 41.9 41.9 41.9 41.9 41.8 ...  
## $ start\_lng : num -87.7 -87.7 -87.7 -87.7 -87.6 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 41.8 ...  
## $ end\_lng : num -87.6 -87.7 -87.7 -87.7 -87.6 ...  
## $ member\_casual : chr "member" "member" "member" "member" ...

## Convert numeric ID into Categorical ID

Bike\_Share\_202004$start\_station\_id=as.character(Bike\_Share\_202004$start\_station\_id)  
Bike\_Share\_202004$end\_station\_id=as.character(Bike\_Share\_202004$end\_station\_id)  
  
Bike\_Share\_202005$start\_station\_id=as.character(Bike\_Share\_202005$start\_station\_id)  
Bike\_Share\_202005$end\_station\_id=as.character(Bike\_Share\_202005$end\_station\_id)  
  
Bike\_Share\_202006$start\_station\_id=as.character(Bike\_Share\_202006$start\_station\_id)  
Bike\_Share\_202006$end\_station\_id=as.character(Bike\_Share\_202006$end\_station\_id)  
  
  
Bike\_Share\_202007$start\_station\_id=as.character(Bike\_Share\_202007$start\_station\_id)  
Bike\_Share\_202007$end\_station\_id=as.character(Bike\_Share\_202007$end\_station\_id)  
  
Bike\_Share\_202008$start\_station\_id=as.character(Bike\_Share\_202008$start\_station\_id)  
Bike\_Share\_202008$end\_station\_id=as.character(Bike\_Share\_202008$end\_station\_id)  
  
Bike\_Share\_202009$start\_station\_id=as.character(Bike\_Share\_202009$start\_station\_id)  
Bike\_Share\_202009$end\_station\_id=as.character(Bike\_Share\_202009$end\_station\_id)  
  
Bike\_Share\_202010$start\_station\_id=as.character(Bike\_Share\_202010$start\_station\_id)  
Bike\_Share\_202010$end\_station\_id=as.character(Bike\_Share\_202010$end\_station\_id)  
  
Bike\_Share\_202011$start\_station\_id=as.character(Bike\_Share\_202011$start\_station\_id)  
Bike\_Share\_202011$end\_station\_id=as.character(Bike\_Share\_202011$end\_station\_id)  
  
Bike\_Share\_202012$start\_station\_id=as.character(Bike\_Share\_202012$start\_station\_id)  
Bike\_Share\_202012$end\_station\_id=as.character(Bike\_Share\_202012$end\_station\_id)

## Combine all data and make it all trips

Trips=rbind(Bike\_Share\_202004,Bike\_Share\_202005,Bike\_Share\_202006,Bike\_Share\_202007,Bike\_Share\_202008,Bike\_Share\_202009,Bike\_Share\_202010,Bike\_Share\_202011,Bike\_Share\_202012)  
str(Trips)

## 'data.frame': 3114796 obs. of 13 variables:  
## $ ride\_id : chr "A847FADBBC638E45" "5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...  
## $ rideable\_type : chr "docked\_bike" "docked\_bike" "docked\_bike" "docked\_bike" ...  
## $ started\_at : chr "2020-04-26 17:45:14" "2020-04-17 17:08:54" "2020-04-01 17:54:13" "2020-04-07 12:50:19" ...  
## $ ended\_at : chr "2020-04-26 18:12:03" "2020-04-17 17:17:03" "2020-04-01 18:08:36" "2020-04-07 13:02:31" ...  
## $ start\_station\_name: chr "Eckhart Park" "Drake Ave & Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...  
## $ start\_station\_id : chr "86" "503" "142" "216" ...  
## $ end\_station\_name : chr "Lincoln Ave & Diversey Pkwy" "Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...  
## $ end\_station\_id : chr "152" "499" "255" "657" ...  
## $ start\_lat : num 41.9 41.9 41.9 41.9 41.9 ...  
## $ start\_lng : num -87.7 -87.7 -87.6 -87.7 -87.6 ...  
## $ end\_lat : num 41.9 41.9 41.9 41.9 42 ...  
## $ end\_lng : num -87.7 -87.7 -87.6 -87.7 -87.7 ...  
## $ member\_casual : chr "member" "member" "member" "member" ...

## Check Missing Values

#Margin=1 means check missing values in row  
#Margin=2 means check missing values in column  
apply(X=is.na(Trips),MARGIN=2,FUN=sum)

## ride\_id rideable\_type started\_at ended\_at   
## 0 0 0 0   
## start\_station\_name start\_station\_id end\_station\_name end\_station\_id   
## 0 83583 0 98104   
## start\_lat start\_lng end\_lat end\_lng   
## 0 0 4254 4254   
## member\_casual   
## 0

The data set has missing values in start\_station\_id column, end\_station\_id column, end\_lat column and end\_lng column.

## rename variable label

Trips=Trips%>%  
 rename(member\_type=member\_casual)

## Check duplicate data

count(distinct(Trips))

## n  
## 1 3114796

Based on the above results, the dataset has no duplicate rows.

## Right now started\_at column and ended\_at column are in character format. It should be in datetime format. Convert those columns into datetime format.

Trips$started\_at=strptime(Trips$started\_at,format="%Y-%m-%d %H: %M: %S")  
  
Trips$ended\_at=strptime(Trips$ended\_at,format="%Y-%m-%d %H: %M: %S")

## Remove start\_lat,start\_lng, end\_lat, and end\_lng columns since those columns are not useful in our analysis

Trips=Trips%>%  
 select(-c(start\_lat,start\_lng,end\_lat,end\_lng))

## Create the new column called ride\_length in minutes

Trips$ride\_length=difftime(Trips$ended\_at,Trips$started\_at,units="mins")

## calculate the day of the week that each ride started

Trips$day\_of\_week=wday(Trips$started\_at,label=TRUE)

## calculate the month that each ride started

Trips$month=month(Trips$started\_at,label=TRUE)

## Checking for negative values in ride\_length column. I Will remove those since it doesn’t make sense to have a negative length of the ride

Trips%>%count(ride\_length<0)

## ride\_length < 0 n  
## 1 FALSE 3104248  
## 2 TRUE 10548

Trips=Trips%>%filter(ride\_length>0)

Remove 10548 rows since they have a negative values in ride\_length column.

## Analyze Phase

## How many customers are casual and paid member

Trips%>%  
 group\_by(member\_type)%>%  
 summarize(Num\_Ride=n())

## # A tibble: 2 × 2  
## member\_type Num\_Ride  
## <chr> <int>  
## 1 casual 1314684  
## 2 member 1789196

## Average Ride\_length

ave=Trips%>%  
 group\_by(member\_type)%>%  
 summarize(Average\_Ride\_length=mean(ride\_length))  
ave

## # A tibble: 2 × 2  
## member\_type Average\_Ride\_length  
## <chr> <drtn>   
## 1 casual 45.64450 mins   
## 2 member 16.39175 mins

On average, paid member rides the bike less minutes than casual member. This mean that casual members ride the bike longer duration than paid members.

## In Which day do customers bike the most?And How long do they bike?

Day=Trips%>%  
 group\_by(member\_type,day\_of\_week)%>%  
 select(member\_type,day\_of\_week,ride\_length)%>%  
 summarize(Num\_Rides=n(),average\_ride=mean(ride\_length))%>%  
 arrange(member\_type,day\_of\_week)

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

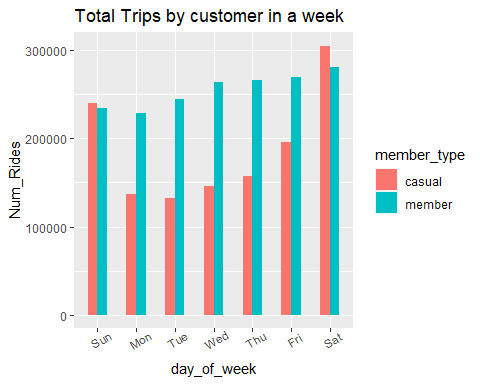
Day

## # A tibble: 14 × 4  
## # Groups: member\_type [2]  
## member\_type day\_of\_week Num\_Rides average\_ride   
## <chr> <ord> <int> <drtn>   
## 1 casual Sun 240645 51.72899 mins  
## 2 casual Mon 136489 45.46414 mins  
## 3 casual Tue 131901 41.03829 mins  
## 4 casual Wed 146284 41.28034 mins  
## 5 casual Thu 157497 43.87954 mins  
## 6 casual Fri 196411 43.44176 mins  
## 7 casual Sat 305457 47.33705 mins  
## 8 member Sun 234122 18.51921 mins  
## 9 member Mon 229339 15.48553 mins  
## 10 member Tue 244503 15.37219 mins  
## 11 member Wed 264465 15.54850 mins  
## 12 member Thu 266451 15.52433 mins  
## 13 member Fri 269247 16.13642 mins  
## 14 member Sat 281069 18.10636 mins

## Data visualizations

Trips%>%  
 group\_by(member\_type,day\_of\_week)%>%  
 summarize(Num\_Rides=n())%>%  
 ggplot(aes(x=day\_of\_week,y=Num\_Rides,fill=member\_type))+  
 theme(axis.text.x=element\_text(angle=30))+  
 labs(title="Total Trips by customer in a week")+  
 geom\_col(width=0.5, position=position\_dodge(width=0.5))+  
 scale\_y\_continuous(labels=function(x) format(x,scientific=FALSE))

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



On Weekend, Casual and paid member ride the bike the most. From Monday to Friday, casual member ride decrease. However, paid member ride is still close to weekend ride. On Sunday, casual member ride an average of 51.73 minute. Casual member rides more duration than paid member.

## In which month do customers ride the most?

month=Trips%>%  
 group\_by(member\_type,month)%>%  
 select(member\_type,month,ride\_length)%>%  
 summarize(Num\_Rides=n(),average\_ride=mean(ride\_length))%>%  
 arrange(member\_type,month)

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

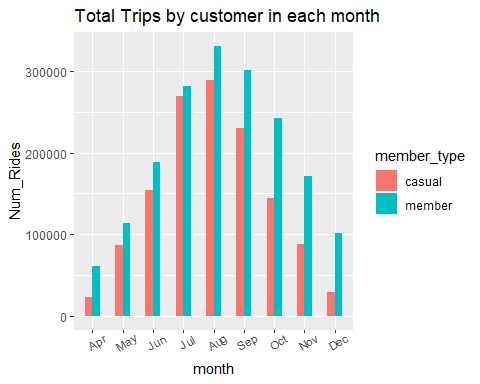
month

## # A tibble: 18 × 4  
## # Groups: member\_type [2]  
## member\_type month Num\_Rides average\_ride   
## <chr> <ord> <int> <drtn>   
## 1 casual Apr 23605 73.14255 mins  
## 2 casual May 86838 51.22108 mins  
## 3 casual Jun 154536 51.67146 mins  
## 4 casual Jul 268663 59.95475 mins  
## 5 casual Aug 288586 44.93976 mins  
## 6 casual Sep 230049 38.22331 mins  
## 7 casual Oct 144511 30.26893 mins  
## 8 casual Nov 87902 31.84320 mins  
## 9 casual Dec 29994 26.85229 mins  
## 10 member Apr 61112 21.48034 mins  
## 11 member May 113252 19.77340 mins  
## 12 member Jun 187968 18.73320 mins  
## 13 member Jul 281002 17.76842 mins  
## 14 member Aug 330895 16.83624 mins  
## 15 member Sep 300718 15.54198 mins  
## 16 member Oct 242191 14.05165 mins  
## 17 member Nov 170921 13.59875 mins  
## 18 member Dec 101137 12.74999 mins

## Data Visualizations of which month do customers ride the most.

Trips%>%  
 group\_by(member\_type,month)%>%  
 summarize(Num\_Rides=n())%>%  
 ggplot(aes(x=month,y=Num\_Rides,fill=member\_type))+  
 theme(axis.text.x=element\_text(angle=30))+  
 labs(title="Total Trips by customer in each month")+  
 geom\_col(width=0.5, position=position\_dodge(width=0.5))+  
 scale\_y\_continuous(labels=function(x) format(x,scientific=FALSE))

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

 During the peak of summer months, casual and paid member ride the most. After the summer season is over, number of rides for casual and paid member decrease significantly. And also, average number of rides also decrease after the summer is over.

## What kinds of ride do customer like the most?

Type=Trips%>%  
 group\_by(member\_type,rideable\_type)%>%  
 select(member\_type,rideable\_type,ride\_length)%>%  
 summarize(Num\_Rides=n(),average\_ride=mean(ride\_length))%>%  
 arrange(member\_type,-Num\_Rides)

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

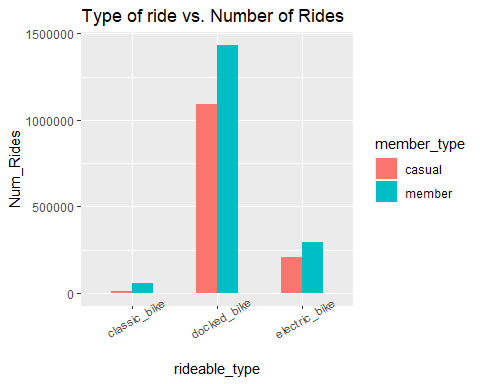
Type

## # A tibble: 6 × 4  
## # Groups: member\_type [2]  
## member\_type rideable\_type Num\_Rides average\_ride   
## <chr> <chr> <int> <drtn>   
## 1 casual docked\_bike 1094267 50.42243 mins  
## 2 casual electric\_bike 209098 21.68765 mins  
## 3 casual classic\_bike 11319 26.29597 mins  
## 4 member docked\_bike 1434532 17.13272 mins  
## 5 member electric\_bike 295369 13.42870 mins  
## 6 member classic\_bike 59295 13.22546 mins

## Data visualizations for what kinds of ride do customer like the most.

Trips%>%  
 group\_by(member\_type,rideable\_type)%>%  
 summarize(Num\_Rides=n())%>%  
 ggplot(aes(x=rideable\_type,y=Num\_Rides,fill=member\_type))+  
 labs(title="Type of ride vs. Number of Rides")+  
 theme(axis.text.x=element\_text(angle=30))+  
 geom\_col(width=0.5, position=position\_dodge(width=0.5))+  
 scale\_y\_continuous(labels=function(x) format(x,scientific=FALSE))

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

 Docked Bikes are the most popular type of bike for both casual and paid member. Both customers don’t usually use classic bike. #{r} #write.csv(Trips,'C:\\Users\\Aungkyaw\\Desktop\\data.csv') #