Bikeshare

Pyimoe Than

6/17/2022

Business problem

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

Importing library

```
library(dplyr)
library(writexl)
library(tidyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.1
```

```
library(tidyverse)
library(data.table)
library(lubridate)#for datetime object
```

Load the data

```
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_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_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_202010)
```

```
## 'data.frame':    388653 obs. of  13 variables:
##  $ ride_id           : chr  "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE
261B9E854" ...
##  $ 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" "S
tony 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 A
ve & 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" "E533E89
C32080B9E" ...
##  $ 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 D
r & 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" "BE11962
8E44F871E" ...
##  $ 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_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_202010,Bike_Share_202011,Bike_Share_202012)
str(Trips)
```

Check Missing Values

```
##            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              55839                  0              62613 
##          start_lat          start_lng            end_lat            end_lng 
##                  0                  0                869                869 
##      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
<int>
779942

1 row

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 <lgl>	n <int>
FALSE	776732
TRUE	3210

2 rows

```
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())
```

member_type <chr>	Num_Ride <int>
casual	262407
member	514249

2 rows

Average length of ride

```
ave=Trips%>%
  group_by(member_type)%>%
  summarize(Average_Ride_length=mean(ride_length))
ave
```

member_type <chr>	Average_Ride_length <drtn>
casual	30.40575 mins
member	13.64512 mins

2 rows

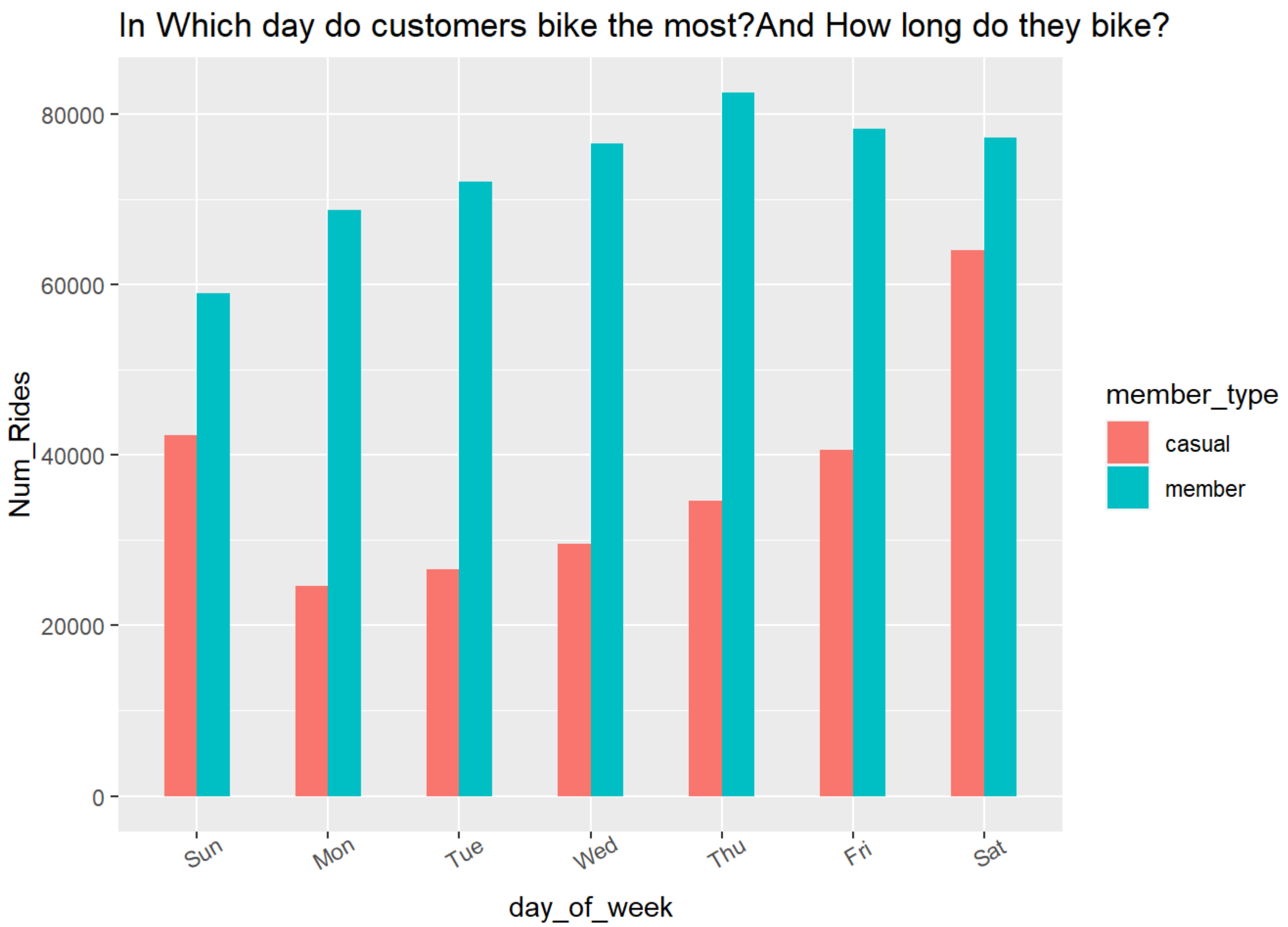
On average, paid member rides the bike less minutes than casual member. This mean that casual members ride the bike longer duration than paid members.

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

member_type <chr>	day_of_week <ord>	Num_Rides <int>	average_ride <drtn>
casual	Sun	42334	36.45422 mins
casual	Mon	24635	25.78331 mins
casual	Tue	26546	27.40715 mins
casual	Wed	29599	25.45775 mins

member_type <chr>	day_of_week <ord>	Num_Rides <int>	average_ride <drtn>
casual	Thu	34682	25.41700 mins
casual	Fri	40541	30.34712 mins
casual	Sat	64070	34.45245 mins
member	Sun	59007	14.65787 mins
member	Mon	68677	12.42571 mins
member	Tue	72018	12.98060 mins

1-10 of 14 rows Previous **1** 2 Next


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

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.

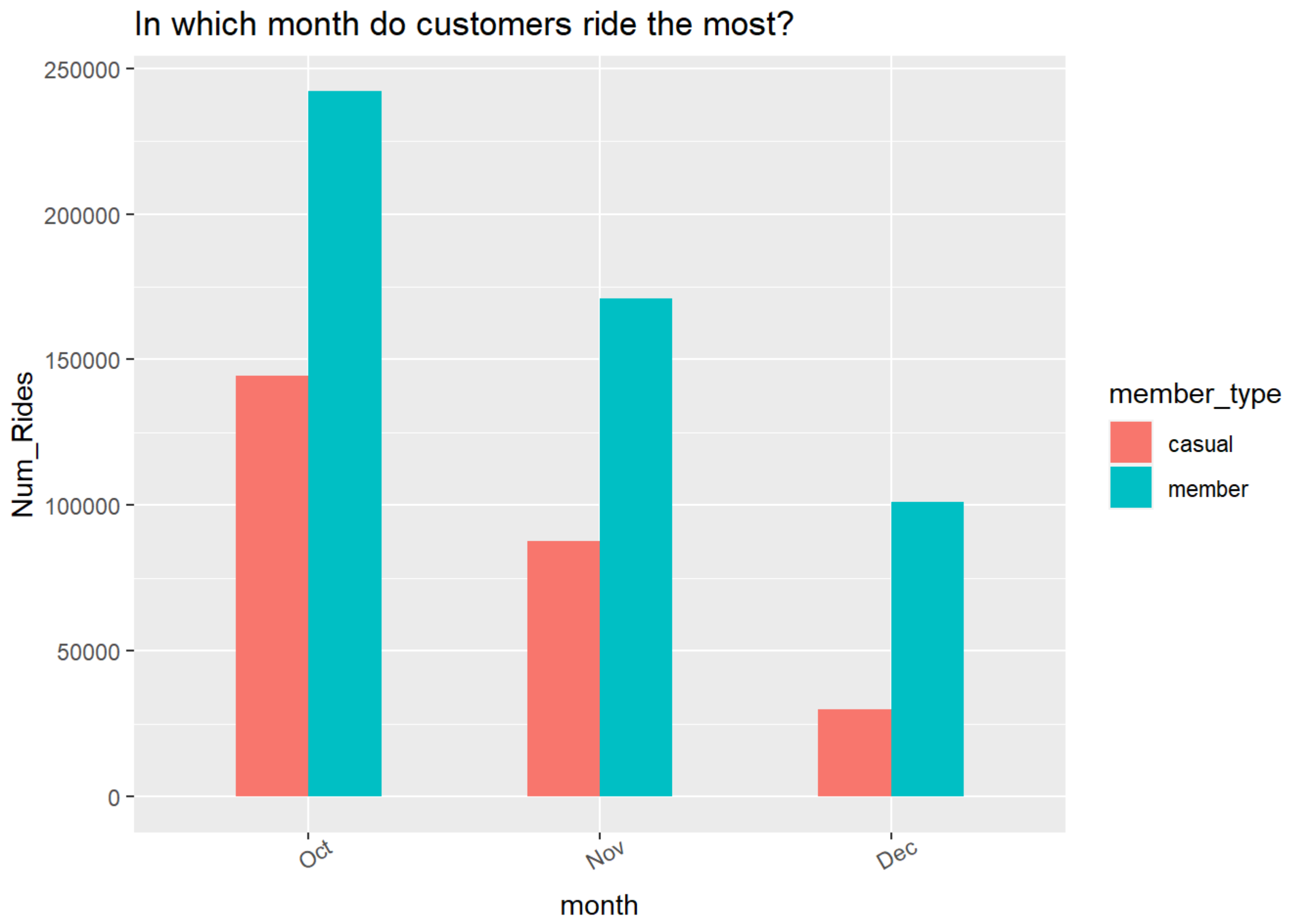
In which month do customers ride the most?

member_type	month	Num_Rides	average_ride
<chr>	<ord>	<int>	<drtn>
casual	Oct	144511	30.26893 mins
casual	Nov	87902	31.84320 mins
casual	Dec	29994	26.85229 mins
member	Oct	242191	14.05165 mins
member	Nov	170921	13.59875 mins
member	Dec	101137	12.74999 mins

6 rows

```
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="In which month do customers ride the most?")+
  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.
```

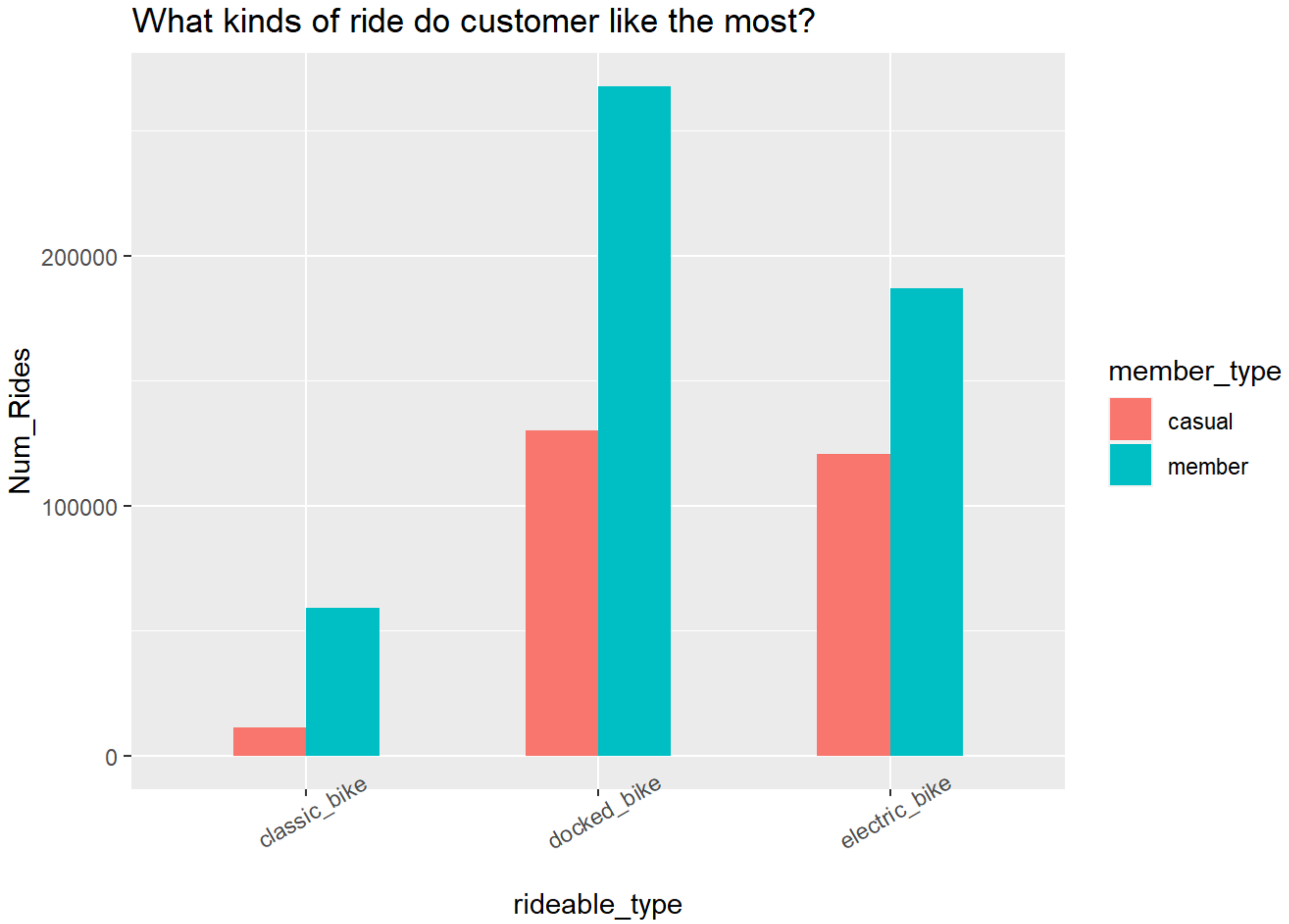


During the peak of summer months, casual and paid member ride the most. After the summer season is over, number of rides for casual and paid member decrease significantly. And also, average number of rides also decrease after the summer is over.

What kinds of ride do customer like the most?

member_type <chr>	rideable_type <chr>	Num_Rides <int>	average_ride <drtn>
casual	docked_bike	130164	40.77916 mins
casual	electric_bike	120924	19.62438 mins
casual	classic_bike	11319	26.29597 mins
member	docked_bike	267839	14.63814 mins
member	electric_bike	187115	12.35670 mins
member	classic_bike	59295	13.22546 mins

6 rows



What kinds of ride do customer like the most?

Docked Bikes are the most popular type of bike for both casual and paid member. Both customers don’t usually use classic bike.