会员和临时使用者在哪些方面以不同的方式使用 Divvy 自行车

qinggeli

2022-11-10

在这个案例分析中，我们根据共享自行车公司 Cyclistic 提供的数据，对会员和非会员的共享自行车使用进行分析

首先加载必要的 function

```
library(ggplot2)
library(tibble)
library(tidyr)
library(readr)
library(purrr)
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

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

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

```
library(stringr)
library(forcats)
library(lubridate)
```

```
##
## 载入程辑包：'lubridate'
```

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

再加载原始数据

```
inf_2021_1 <- read_csv("202101-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_2 <- read_csv("202102-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_3 <- read_csv("202103-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_4 <- read_csv("202104-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.
```

```
inf_2021_5 <- read_csv("202105-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.
```

```
inf_2021_6 <- read_csv("202106-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.
```

```
inf_2021_7 <- read_csv("202107-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_8 <- read_csv("202108-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_9 <- read_csv("202109-divvy-tripdata.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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_10 <- read_csv("202110-divvy-tripdata.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 d
ata.
```

```
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_11 <- read_csv("202111-divvy-tripdata.csv")

## Rows: 359978 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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.

inf_2021_12 <- read_csv("202112-divvy-tripdata.csv")

## Rows: 247540 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 d
ata.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet
 this message.
```

将原来 12 格表格融合成为 1 个 full_year 表格

```
full_year <- bind_rows(
                      inf_2021_1,
                      inf_2021_2,
                      inf_2021_3,
                      inf_2021_4,
                      inf_2021_5,
                      inf_2021_6,
                      inf_2021_7,
                      inf_2021_8,
                      inf_2021_9,
                      inf_2021_10,
                      inf_2021_11,
                      inf_2021_12,
                      )
```

将 ride_id 和 rideable_type 转换为字符，以便它们可以正确堆叠

```
full_year <-  mutate(full_year, ride_id = as.character(ride_id)
                   ,rideable_type = as.character(rideable_type))

full_year <- mutate(full_year, start_station_id = as.numeric(start_stat
ion_id),
                   end_station_id = as.numeric(end_station_id))

## Warning in mask$eval_all_mutate(quo): 强制改变过程中产生了 NA

## Warning in mask$eval_all_mutate(quo): 强制改变过程中产生了 NA
```

增加"ride_length" 列

```
full_year$ride_length <- difftime(full_year$ended_at,full_year$started_
at)
```

转换 "ride_length" 从因子转换成数字，这样我们就可以计算

```
is.factor(full_year$ride_length)#判断是否是因子

## [1] FALSE

full_year$ride_length <- as.numeric(as.character(full_year$ride_length))
#转换
is.numeric(full_year$ride_length)#判断是否是数字

## [1] TRUE
```

对 member_casual 列的原有数据进行编辑进行编辑

```
full_year <-full_year%>%mutate(member_casual = recode(member_casual
                                  ,"Subscriber" = "member"
                                  ,"Customer" = "casual"))
table(full_year$member_casual)

##
##  casual  member
## 2529005 3066058
```

查看临时和会员的平均骑乘距离

```
filter_casual<-filter(full_year,member_casual=="casual")
mean_casual_ride_length<-mean(filter_casual$ride_length)
filter_member<-filter(full_year,member_casual=="member")
mean_member_ride_length<-mean(filter_member$ride_length)
rm(filter_member,filter_casual)
view(mean_member_ride_length)
view(mean_casual_ride_length)
```

去除坏数据

```
full_year_v2 <- full_year[!(full_year$start_station_name == "HQ QR" | f
ull_year$ride_length<0),]
```

增加 weekday 因子，并计算平均使用者类型的骑乘数和平均时间

```
full_year_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #使用 wday()创建
weekday 列
  group_by(member_casual, weekday) %>%  #groups by usertype & weekday
  summarise(number_of_rides = n()                              #计算骑乘
  ,average_duration = mean(ride_length)) %>%        # 计算平均时间
  arrange(member_casual, weekday)

## `summarise()` has grouped output by 'member_casual'. You can overrid
e using the
## `.groups` argument.

## # A tibble: 15 × 4
## # Groups:   member_casual [3]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        周日             430241            2372.
##  2 casual        周一             248229            2030.
##  3 casual        周二             234951            1780.
##  4 casual        周三             238801            1762.
##  5 casual        周四             245095            1763.
##  6 casual        周五             314861            1937.
##  7 casual        周六             499089            2187.
##  8 member        周日             329940             948.
##  9 member        周一             366329             800.
## 10 member        周二             410609             771.
## 11 member        周三             420864             772.
## 12 member        周四             396134             770.
## 13 member        周五             389483             802.
## 14 member        周六             379501             924.
## 15 <NA>          <NA>             690789              NA
```

去掉 NA rows

```
full_year_v2<- drop_na(full_year_v2)
```

查看 full_year_v2 表格的数据结构

```
colnames(full_year_v2)

##  [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"      "ride_length"

nrow(full_year_v2)

## [1] 1040807
```

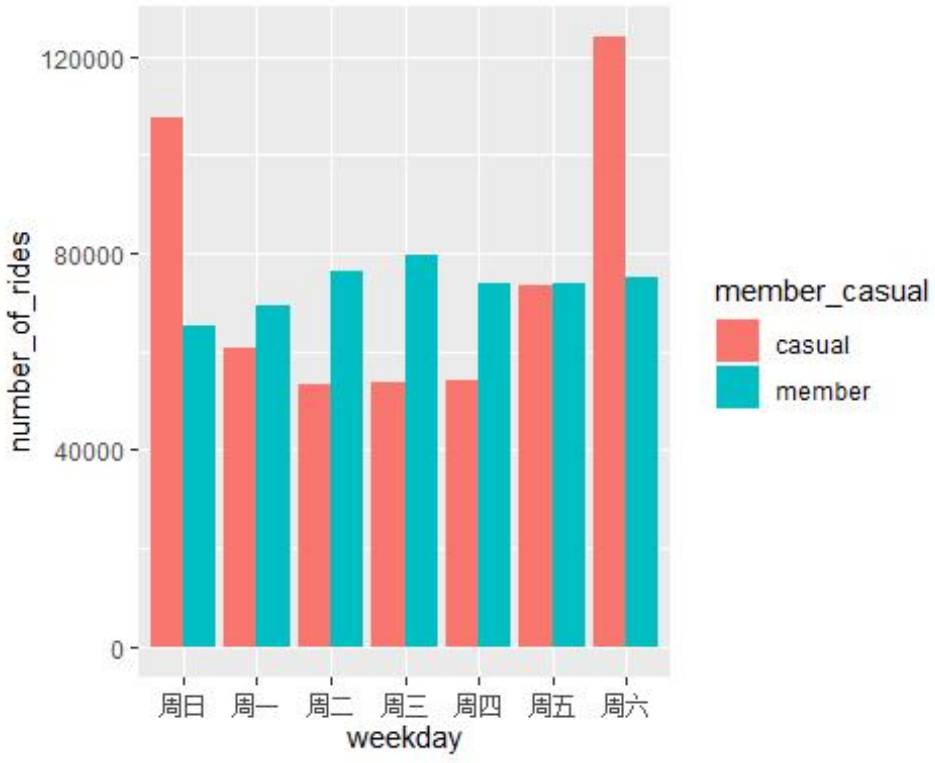
```
str(full_year_v2)
```

```
## tibble [1,040,807 × 14] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:1040807] "B9F73448DFBE0D45" "457C7F4B5
D3DA135" "57C750326F9FDABE" "4D518C65E338D070" ...
##  $ rideable_type     : chr [1:1040807] "classic_bike" "electric_bike
" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct[1:1040807], format: "2021-01-24 19:15:
38" "2021-01-23 12:57:38" ...
##  $ ended_at          : POSIXct[1:1040807], format: "2021-01-24 19:22:
51" "2021-01-23 13:02:10" ...
##  $ start_station_name: chr [1:1040807] "California Ave & Cortez St"
"California Ave & Cortez St" "California Ave & Cortez St" "California A
ve & Cortez St" ...
##  $ start_station_id  : num [1:1040807] 17660 17660 17660 17660 17660
 ...
##  $ end_station_name  : chr [1:1040807] "Wood St & Augusta Blvd" "Cal
ifornia Ave & North Ave" "Wood St & Augusta Blvd" "Wood St & Augusta Bl
vd" ...
##  $ end_station_id    : num [1:1040807] 657 13258 657 657 657 ...
##  $ start_lat         : num [1:1040807] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:1040807] -87.7 -87.7 -87.7 -87.7 -87.7
 ...
##  $ end_lat           : num [1:1040807] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:1040807] -87.7 -87.7 -87.7 -87.7 -87.7
 ...
##  $ member_casual     : chr [1:1040807] "member" "member" "casual" "c
asual" ...
##  $ ride_length       : num [1:1040807] 433 272 587 537 609 ...
```

average duration 可视化

```
full_year_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")
```

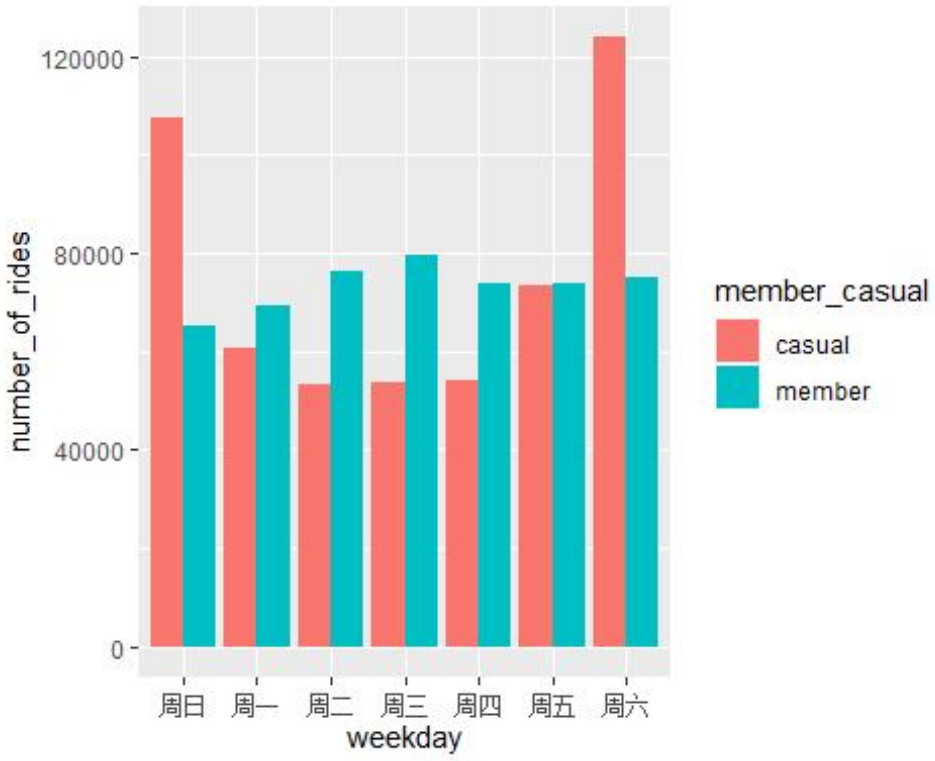
```
## `summarise()` has grouped output by 'member_casual'. You can overrid
e using the
## `.groups` argument.
```



##通过 rider type 可视化 number of rides

```
full_year_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 overrid
e using the
## `.groups` argument.
```


120000
80000
40000
0
number_of_rides
周日
周一
周二
周三
周四
周五
周六
weekday
member_casual
casual
member