Case study 1: How Does a Bike-Share Navigate Speedy Success

Soa-Yu Chan

2021/6/19

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

Contents of Document

1. Data Source

This is a public data came from **Google Data Analytics** course in coursera. Cyclistic is a fictional bike-share company located in Chicago with more than 5,800 bicycles and 600 docking stations. The data has been made available by Motivate International Inc. under this **license**. This case study hope to find the difference between annual members and casual riders.

2. Data Structure and Manipulate

- `X.U.FEFF.`: Index
- `ride_id`: ID attached to each trip taken
- `rideable_type`: rideable type
- `start_at`: day and time trip started, in CST
- `ended_at`: day and time trip ended, in CST
- `start_station_name`: name of station where trip originated
- `start_station_id`: ID of station where trip originated
- `end_station_name`: name of station where trip terminated
- `end_station_id`: ID of station where trip terminated
- `start_lat`: station latitude where trip originated
- `start_lng`: station longitude where trip originated
- `end_lat`: station latitude where trip terminated
- `end_lng`: station longitude where trip terminated

- `member_casual`: “casual” is a rider who purchased single-ride passes or full-day passes; “member” is a rider who purchased an Annual Membership

- `ride_length_second`: each ride time from trip originated to trip terminated, in second

- `ride_length`: each ride time from start to end, in hh:mm:ss
- `day_of_week`: the day of the week that each ride started

From **2020-04** to **2021-05**.

```
library(table1)
library(tidyverse)
library(ggplot2)
library(lubridate)
library(scales)

bike_share <- read.csv(file = "D:/case_study_2021_06_13/202004_202105_divvy_tripdata.csv",
                header = T, na.strings = c("", "NA"), encoding = "UTF-8", sep = ",")
glimpse(bike_share)
```

```
## Rows: 4,348,052
## Columns: 17
## $ X.U.FEFF.          <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A79A4~
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike", "docke~
## $ started_at         <chr> "2020-04-26 17:45:14", "2020-04-17 17:08:54", "2020~
## $ ended_at           <chr> "2020-04-26 18:12:03", "2020-04-17 17:17:03", "2020~
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "McClu~
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35", "434~
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park", "~
## $ end_station_id     <chr> "152.0", "499.0", "255.0", "657.0", "323.0", "35.0"~
## $ start_lat          <dbl> 41.8964, 41.9244, 41.8945, 41.9030, 41.8902, 41.896~
## $ start_lng          <dbl> -87.6610, -87.7154, -87.6179, -87.6975, -87.6262, -~
## $ end_lat            <dbl> 41.9322, 41.9306, 41.8679, 41.8992, 41.9695, 41.892~
## $ end_lng            <dbl> -87.6586, -87.7238, -87.6230, -87.6722, -87.6547, -~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
## $ ride_length_second <dbl> 1609, 489, 863, 732, 3175, 324, 313, 4549, 344, 103~
## $ ride_length        <chr> "0:26:49", "0:08:09", "0:14:23", "0:12:12", "0:52:5~
## $ day_of_week        <int> 1, 6, 4, 3, 7, 5, 5, 3, 4, 7, 7, 7, 6, 7, 2, 7, 1, ~
```

- Add `ride_length_hour`: ride length, in hour
- Add `started_date`: started date, in yyyy-mm-dd

- Add `start_mm_yyyy`: started year-month, in yyyy-mm

```
bike_share$ride_length_hour <- ((bike_share$ride_length_second + 0.001)/3600) - (0.001 / 3600)
bike_share$started_date <- as_date(ymd_hms(bike_share$started_at))
bike_share$start_mm_yyyy <- format_ISO8601(bike_share$started_date, precision = "ym")
```

- Some trip characteristic between casual riders and annual members

```
table1(~ factor(day_of_week) + ride_length_hour + factor(rideable_type) + start_mm_yyyy | factor(member
```

	casual	member	Overall
	(N=1820638)	(N=2527414)	(N=4348052)
factor(day_of_week)			
<U+00A0><U+00A0>1	349624 (19.2%)	332561 (13.2%)	682185 (15.7%)
<U+00A0><U+00A0>2	201620 (11.1%)	336647 (13.3%)	538267 (12.4%)
<U+00A0><U+00A0>3	184237 (10.1%)	349638 (13.8%)	533875 (12.3%)
<U+00A0><U+00A0>4	193556 (10.6%)	369882 (14.6%)	563438 (13.0%)
<U+00A0><U+00A0>5	201598 (11.1%)	364234 (14.4%)	565832 (13.0%)
<U+00A0><U+00A0>6	261909 (14.4%)	377981 (15.0%)	639890 (14.7%)
<U+00A0><U+00A0>7	428094 (23.5%)	396471 (15.7%)	824565 (19.0%)
ride_length_hour			
<U+00A0><U+00A0>Mean (SD)	0.725 (6.07)	0.264 (1.40)	0.457 (4.08)
<U+00A0><U+00A0>Median [Min, Max]	0.344 [0, 928]	0.188 [0, 979]	0.238 [0, 979]
factor(rideable_type)			
<U+00A0><U+00A0>classic_bike	265526 (14.6%)	578053 (22.9%)	843579 (19.4%)
<U+00A0><U+00A0>docked_bike	1181475 (64.9%)	1434720 (56.8%)	2616195 (60.2%)
<U+00A0><U+00A0>electric_bike	373637 (20.5%)	514641 (20.4%)	888278 (20.4%)
start_mm_yyyy			
<U+00A0><U+00A0>2020-04	23610 (1.3%)	61115 (2.4%)	84725 (1.9%)
<U+00A0><U+00A0>2020-05	86844 (4.8%)	113258 (4.5%)	200102 (4.6%)
<U+00A0><U+00A0>2020-06	154551 (8.5%)	187985 (7.4%)	342536 (7.9%)
<U+00A0><U+00A0>2020-07	268688 (14.8%)	281047 (11.1%)	549735 (12.6%)
<U+00A0><U+00A0>2020-08	288639 (15.9%)	330953 (13.1%)	619592 (14.2%)
<U+00A0><U+00A0>2020-09	230072 (12.6%)	300754 (11.9%)	530826 (12.2%)
<U+00A0><U+00A0>2020-10	144529 (7.9%)	242213 (9.6%)	386742 (8.9%)
<U+00A0><U+00A0>2020-11	87911 (4.8%)	170940 (6.8%)	258851 (6.0%)
<U+00A0><U+00A0>2020-12	29997 (1.6%)	101142 (4.0%)	131139 (3.0%)
<U+00A0><U+00A0>2021-01	18117 (1.0%)	78715 (3.1%)	96832 (2.2%)
<U+00A0><U+00A0>2021-02	10131 (0.6%)	39491 (1.6%)	49622 (1.1%)
<U+00A0><U+00A0>2021-03	84032 (4.6%)	144462 (5.7%)	228494 (5.3%)
<U+00A0><U+00A0>2021-04	136601 (7.5%)	200624 (7.9%)	337225 (7.8%)
<U+00A0><U+00A0>2021-05	256916 (14.1%)	274715 (10.9%)	531631 (12.2%)

- Casual Riders Top 10 Trip

```
bike_share %>%
  group_by(start_station_name, end_station_name) %>%
  filter(member_casual == 'casual') %>%
  drop_na() %>%
  summarize(count_start_end = n(), average_rider_length = mean(ride_length_hour) * 60) %>%
  arrange(desc(count_start_end)) %>%
  `colnames<-`(c("Start station name", "End station name", "Count", "Average minutes per ride")) %>%
  head(n=10)
```

```
## # A tibble: 10 x 4
## # Groups:   Start station name [10]
##    ‘Start station name‘    ‘End station name‘      Count ‘Average minutes per r~
##    <chr>                   <chr>                   <int>                   <dbl>
##  1 Streeter Dr & Grand Ave Streeter Dr & Grand Ave  8341                    56.9
##  2 Lake Shore Dr & Monroe~ Lake Shore Dr & Monroe~  7937                    51.4
##  3 Millennium Park         Millennium Park          6528                    57.4
##  4 Buckingham Fountain     Buckingham Fountain      5999                    75.0
##  5 Michigan Ave & Oak St   Michigan Ave & Oak St    4842                    56.3
##  6 Indiana Ave & Roosevel~ Indiana Ave & Roosevel~  4584                    62.6
##  7 Fort Dearborn Dr & 31s~ Fort Dearborn Dr & 31s~  3917                    69.8
##  8 Michigan Ave & 8th St   Michigan Ave & 8th St    3795                    62.5
##  9 Theater on the Lake     Theater on the Lake      3634                    54.8
## 10 Shore Dr & 55th St      Shore Dr & 55th St       3610                    68.7
```

- Annual Members Top 10 Trip

```
bike_share %>%
  group_by(start_station_name, end_station_name) %>%
  filter(member_casual == 'member') %>%
  drop_na() %>%
  summarize(count_start_end = n(), average_rider_length = mean(ride_length_hour)*60) %>%
  arrange(desc(count_start_end)) %>%
  `colnames<-`(c("Start station name", "End station name", "Count", "Average minutes per ride")) %>%
  head(n=10)
```

```
## # A tibble: 10 x 4
## # Groups:   Start station name [10]
##    ‘Start station name‘     ‘End station name‘       Count ‘Average minutes per~
##    <chr>                    <chr>                    <int>                 <dbl>
##  1 MLK Jr Dr & 29th St      State St & 33rd St        1519                  7.55
##  2 Ellis Ave & 60th St      Ellis Ave & 55th St       1416                  5.21
##  3 State St & 33rd St       MLK Jr Dr & 29th St       1350                  9.52
##  4 Ellis Ave & 55th St      Ellis Ave & 60th St       1328                  6.00
##  5 Clark St & Elm St        Clark St & Elm St         1253                 17.0
##  6 Lake Shore Dr & Welling~ Lake Shore Dr & Welling~  1231                 23.4
##  7 Lakefront Trail & Bryn ~ Lakefront Trail & Bryn ~  1220                 27.3
##  8 Lake Shore Dr & Belmont~ Lake Shore Dr & Belmont~  1200                 27.7
##  9 Burnham Harbor           Burnham Harbor            1167                 25.7
## 10 Streeter Dr & Grand Ave  Streeter Dr & Grand Ave   1146                 23.1
```

3. Data Visualizations (Bar Chart - Annual Member vs. Casual rider)

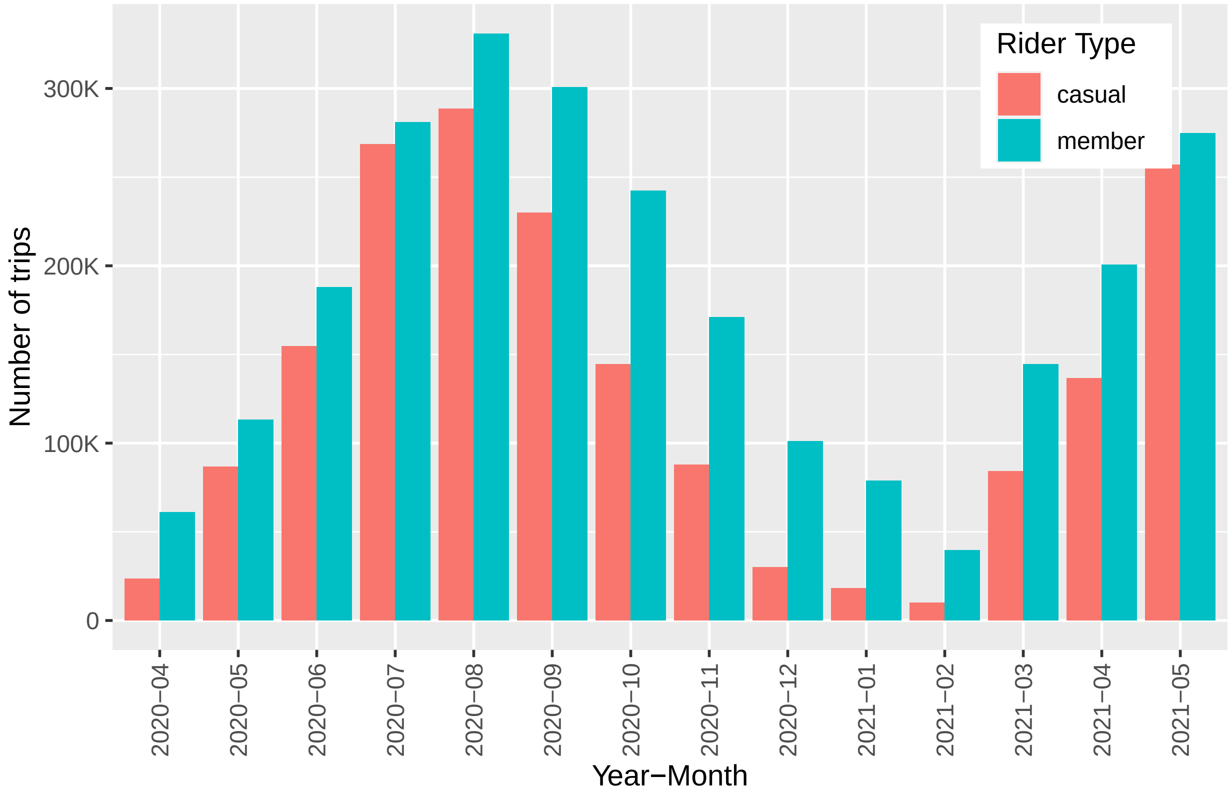
Visualization on Month Year

- Casual rider has less trip during winter season compare to annual member(Fig.1).

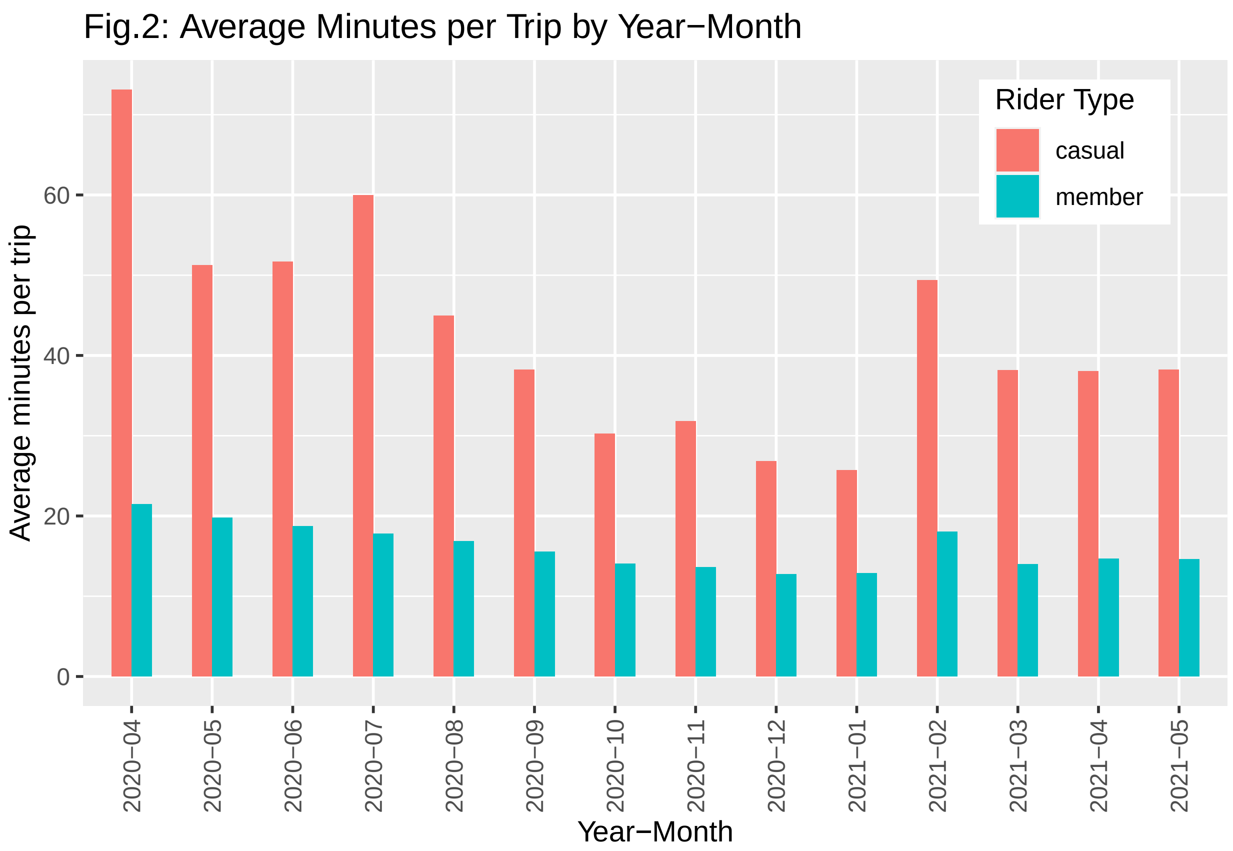
- Annual member spend less time than casual rider for each trip(Fig.2).

```
ggplot(data = bike_share)+
  geom_bar( position = 'dodge', mapping = aes(x = factor(start_mm_yyyy), fill = member_casual)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(title = 'Fig.1: Total number of trips according to Month',
       x = 'Year-Month', y = 'Number of trips', fill='Rider Type') +
  scale_y_continuous(breaks = c(0,100000,200000,300000,400000),
                     labels = c("0","100K","200K","300K","400K")) +
  theme(legend.position = c(.95, .97),
        legend.justification = c("right", "top"),
        legend.box.just = "right",
        legend.margin = margin(2, 10, 2, 6))
```

Fig.1: Total number of trips according to Month



```
t2 <- bike_share %>%
  group_by(start_mm_yyyy, member_casual) %>%
  summarize(average_ride_length = mean(ride_length_hour*60), sum_ride_length = sum(ride_length_hour))

ggplot(t2, aes(x = start_mm_yyyy, y = average_ride_length,  fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge(), width =  0.5) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(title = 'Fig.2: Average Minutes per Trip by Year-Month',
       x = 'Year-Month', y = 'Average minutes per trip', fill = 'Rider Type') +
  theme(legend.position = c(.95, .97),
        legend.justification = c("right", "top"),
        legend.box.just = "right",
        legend.margin = margin(2, 10, 2, 6))
```



Visualization on Day of Week

- Much more trip during weekend for casual riders(Fig.3).
- No significant difference from sunday to saturday for annual members(Fig.3).
- Also, casual rider spend more time than annual member for each trip(Fig.4).

```
ggplot(data = bike_share) +
  geom_bar(mapping = aes(x = factor(day_of_week), fill = member_casual)) +
  facet_wrap(~member_casual) +
  labs(title = 'Fig.3: Total number of trips according to The Day of Week',
   x = 'The Day of Week', y = 'Number of trips', fill = 'Rider Type') +
  scale_x_discrete(breaks = 1:7,
                   labels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")) +
  scale_y_continuous(breaks = c(0, 100000, 200000, 300000, 400000),
                   labels = c("0", "100K", "200K", "300K", "400K")) +
  theme(legend.position = 'none')
```

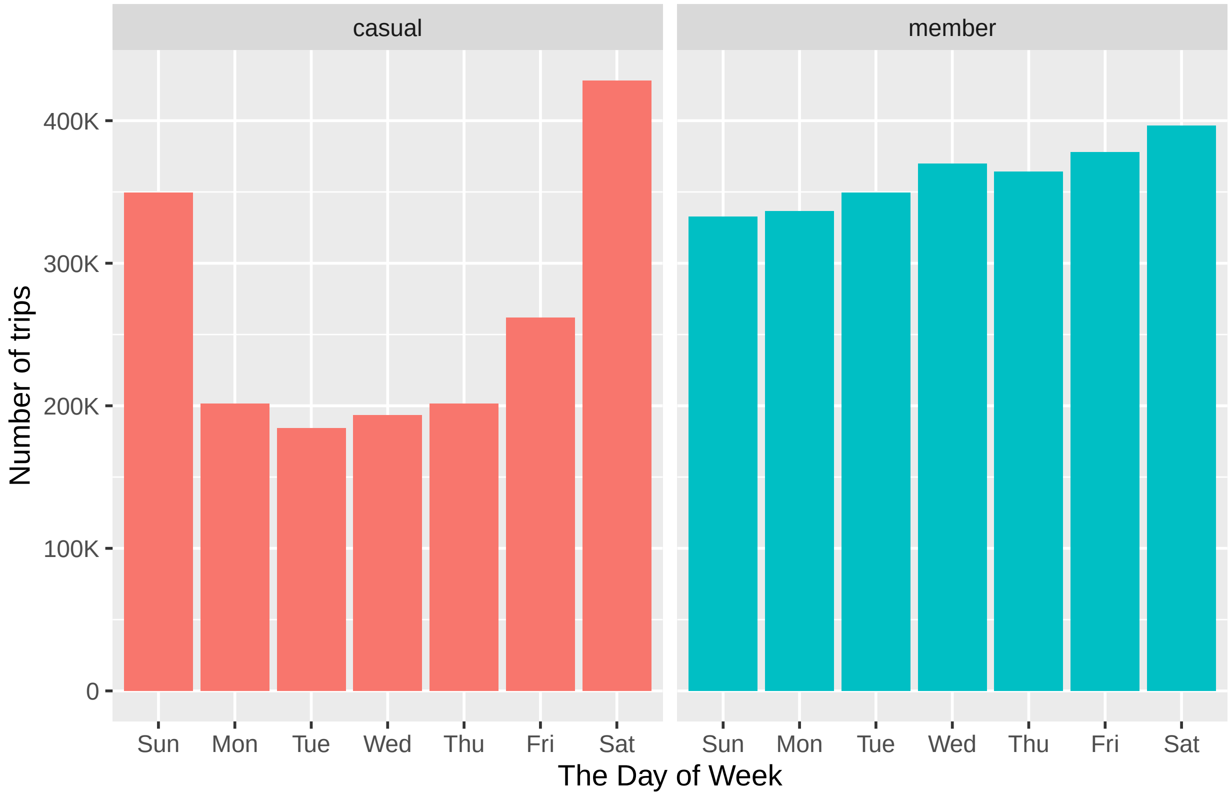
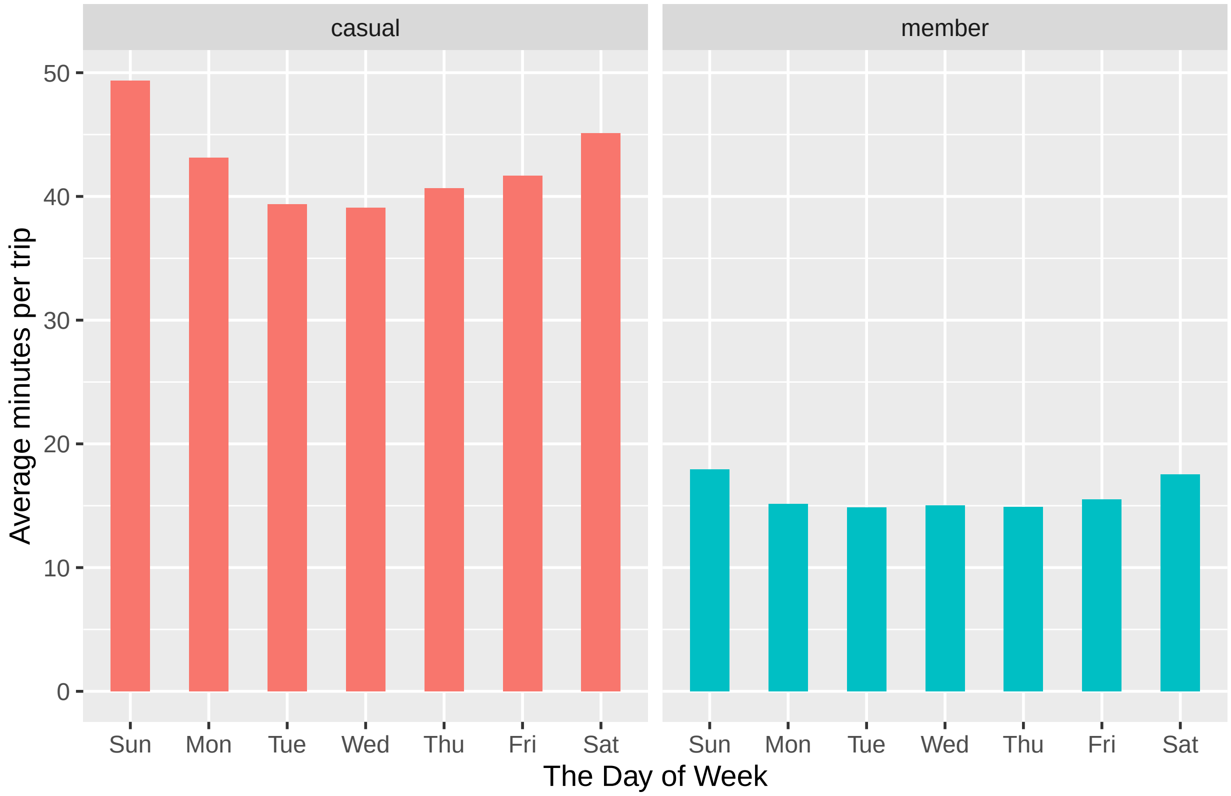


Fig.3: Total number of trips according to The Day of Week

```r
t3 <- bike_share %>%
  group_by(day_of_week, member_casual) %>%
  summarize(average_ride_length = mean(ride_length_hour*60), sum_ride_length = sum(ride_length_hour))

ggplot(t3, aes(x = factor(day_of_week), y = average_ride_length, fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge(), width = 0.5) +
  facet_wrap(~member_casual) +
  labs(title = 'Fig.4: Average Minutes per Trip by The Day of Week',
       x = 'The Day of Week', y = 'Average minutes per trip', fill = 'Rider Type') +
  scale_x_discrete(breaks = 1:7,
                   labels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")) +
  theme(legend.position = 'none')
```

Fig.4: Average Minutes per Trip by The Day of Week

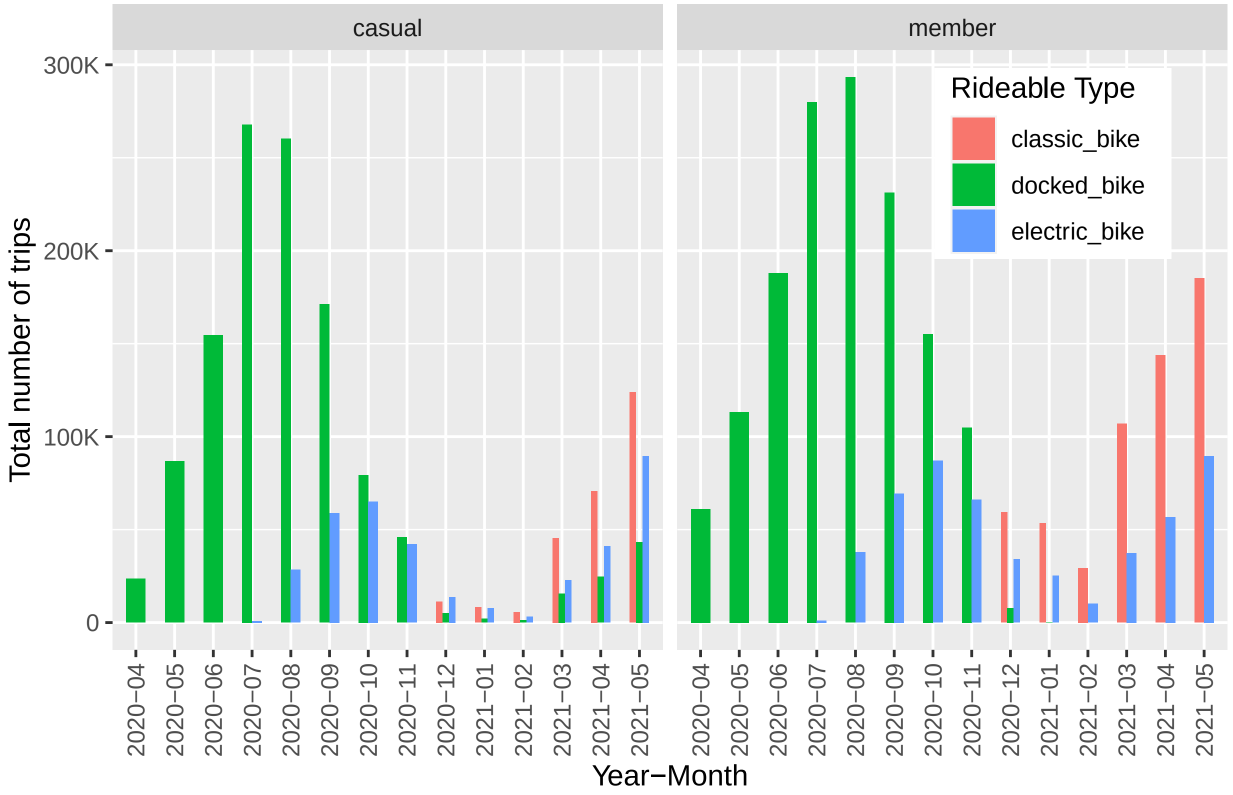


Visualization on Rideable Type

- For annual members: No docked_bike trip after 2021-01. Only classic_bike and electric_bike(Fig.5).
- For casual members: Total number of Classic_bike trip portion larger than docked_bike trip since 2020-12(Fig.5).
- docked_bike spend more average time than classic_bike per trip(Fig.6).

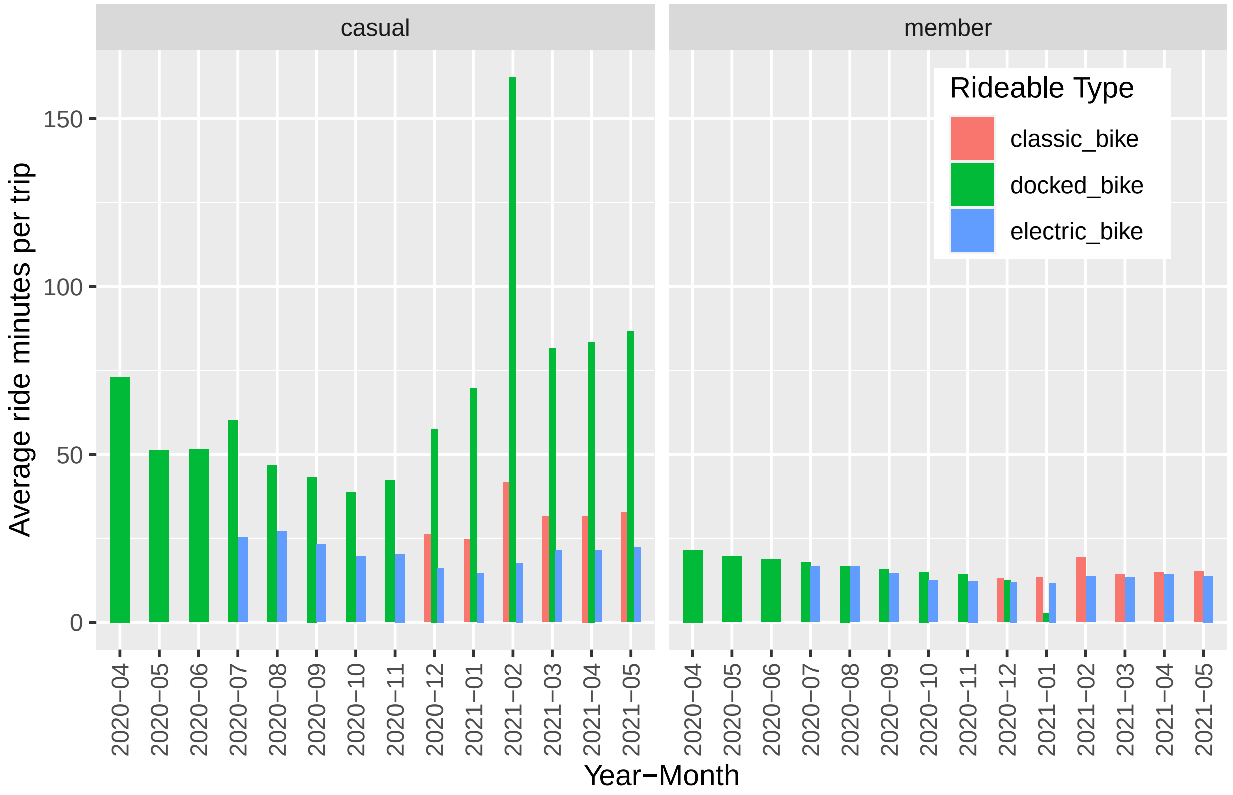
```
t5 <- bike_share %>%
  group_by(start_mm_yyyy, member_casual, rideable_type) %>%
  summarize(average_ride_length = mean(ride_length_hour * 60), sum_ride_length = sum(ride_length_hour),

ggplot(t5, aes(x = start_mm_yyyy, y = sum_trip,  fill = rideable_type)) +
  geom_bar(stat = "identity", position = position_dodge(), width =  0.5) +
  facet_wrap(~member_casual) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(title = 'Fig.5: Total Number of Trips by Year-Month for different Rideable Type',
       x = 'Year-Month', y = 'Total number of trips', fill = 'Rideable Type') +
  scale_y_continuous(breaks = c(0, 100000, 200000, 300000, 400000),
                     labels = c("0", "100K", "200K", "300K", "400K")) +
  theme(legend.position = c(.95, .97),
        legend.justification = c("right", "top"),
        legend.box.just = "right",
        legend.margin = margin(2, 10, 2, 6))
```

Fig.5: Total Number of Trips by Year−Month for different Rideable Type



```
ggplot(t5, aes(x = start_mm_yyyy, y = average_ride_length,  fill = rideable_type)) +
  geom_bar(stat = "identity", position = position_dodge(), width = 0.5) +
  facet_wrap(~member_casual) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(title = 'Fig.6: Average Ride Minutes per Trip by Year-Month for different Rideable Type',
       x = 'Year-Month', y = 'Average ride minutes per trip', fill = 'Rideable Type') +
  theme(legend.position = c(.95, .97),
        legend.justification = c("right", "top"),
        legend.box.just = "right",
        legend.margin = margin(2, 10, 2, 6))
```

Fig.6: Average Ride Minutes per Trip by Year−Month for different Rideable



---

4. Preliminary results

- Change all docked_bike to classic_bike. It can save time for riders.
- Add a half-year annual member to pricing plans. Because many casual riders not use in winter.
- Set more docking station near office zone. Annual member use Cyclistic to commute to work each day.