Case Study 1

Tina

2021/5/24

Install Packages

```
install.packages("tidyverse")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library("tidyverse")
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
install.packages("lubridate")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(dplyr)
```

Read Data

```
bike_share_202104 <- read.csv('202104-divvy-tripdata.csv')
```

Glance of the Data

```
head(bike_share_202104)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 6C992BD37A98A63F  classic_bike 2021-04-12 18:25:36 2021-04-12 18:56:55
## 2 1E0145613A209000   docked_bike 2021-04-27 17:27:11 2021-04-27 18:31:29
## 3 E498E15508A80BAD   docked_bike 2021-04-03 12:42:45 2021-04-07 11:40:24
## 4 1887262AD101C604  classic_bike 2021-04-17 09:17:42 2021-04-17 09:42:48
## 5 C123548CAB2A32A5   docked_bike 2021-04-03 12:42:25 2021-04-03 14:13:42
## 6 097E76F3651B1AC1  classic_bike 2021-04-25 18:43:18 2021-04-25 18:43:59
##         start_station_name start_station_id             end_station_name
## 1    State St & Pearson St     TA1307000061 Southport Ave & Waveland Ave
## 2 Dorchester Ave & 49th St     KA1503000069     Dorchester Ave & 49th St
## 3    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 4  Honore St & Division St     TA1305000034 Southport Ave & Waveland Ave
## 5    Loomis Blvd & 84th St            20121        Loomis Blvd & 84th St
## 6     Clinton St & Polk St            15542         Clinton St & Polk St
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1          13235  41.89745 -87.62872 41.94815 -87.66394        member
## 2   KA1503000069  41.80577 -87.59246 41.80577 -87.59246        casual
## 3          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 4          13235  41.90312 -87.67394 41.94815 -87.66394        member
## 5          20121  41.74149 -87.65841 41.74149 -87.65841        casual
## 6          15542  41.87147 -87.64095 41.87147 -87.64095        casual
```

```
glimpse(bike_share_202104)
```

```
## Rows: 337,230
## Columns: 13
## $ ride_id            <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508~
## $ rideable_type      <chr> "classic_bike", "docked_bike", "docked_bike", "clas~
## $ started_at         <chr> "2021-04-12 18:25:36", "2021-04-27 17:27:11", "2021~
## $ ended_at           <chr> "2021-04-12 18:56:55", "2021-04-27 18:31:29", "2021~
## $ start_station_name <chr> "State St & Pearson St", "Dorchester Ave & 49th St"~
## $ start_station_id   <chr> "TA1307000061", "KA1503000069", "20121", "TA1305000~
## $ end_station_name   <chr> "Southport Ave & Waveland Ave", "Dorchester Ave & 4~
## $ end_station_id     <chr> "13235", "KA1503000069", "20121", "13235", "20121",~
## $ start_lat          <dbl> 41.89745, 41.80577, 41.74149, 41.90312, 41.74149, 4~
## $ start_lng          <dbl> -87.62872, -87.59246, -87.65841, -87.67394, -87.658~
## $ end_lat            <dbl> 41.94815, 41.80577, 41.74149, 41.94815, 41.74149, 4~
## $ end_lng            <dbl> -87.66394, -87.59246, -87.65841, -87.66394, -87.658~
## $ member_casual      <chr> "member", "casual", "casual", "member", "casual", "~
```

```
colnames(bike_share_202104)
```

```
##  [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"
```

Data frame after clean

```
bike_share_202104_clean <- read.csv("bike_share - 202104-divvy-tripdata(filtered).csv")
```

Numbers of trips for members and casual riders

```
ggplot(data = bike_share_202104_clean) +
  geom_bar(mapping = aes(x = member_casual, fill =      member_casual)) +
  labs(title="Numbers of trips for members and casual riders",
       x="Type of Riders",
       y="Number of Trips")
```

Calculate ride length

```
bike_share_202104_clean$ride_length <- difftime(bike_share_202104_clean$ended_at,bike_share_202104_clean
```

Transfer data type

```
bike_share_202104_clean$ride_length <-
as.numeric(as.character(bike_share_202104_clean$ride_length))
glimpse(bike_share_202104_clean)
```

```
## Rows: 296,057
## Columns: 8
## $ ride_id       <chr> "19EC33CD4797D240", "8491214C7F46B7A9", "7EC7EE9B698661D~
## $ rideable_type <chr> "classic_bike", "classic_bike", "classic_bike", "classic~
## $ started_at    <chr> "2021/4/1 8:42", "2021/4/1 9:30", "2021/4/1 10:18", "202~
## $ ended_at      <chr> "2021/4/1 8:43", "2021/4/1 9:31", "2021/4/1 10:19", "202~
## $ member_casual <chr> "member", "member", "member", "casual", "member", "membe~
## $ ride_length   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
```

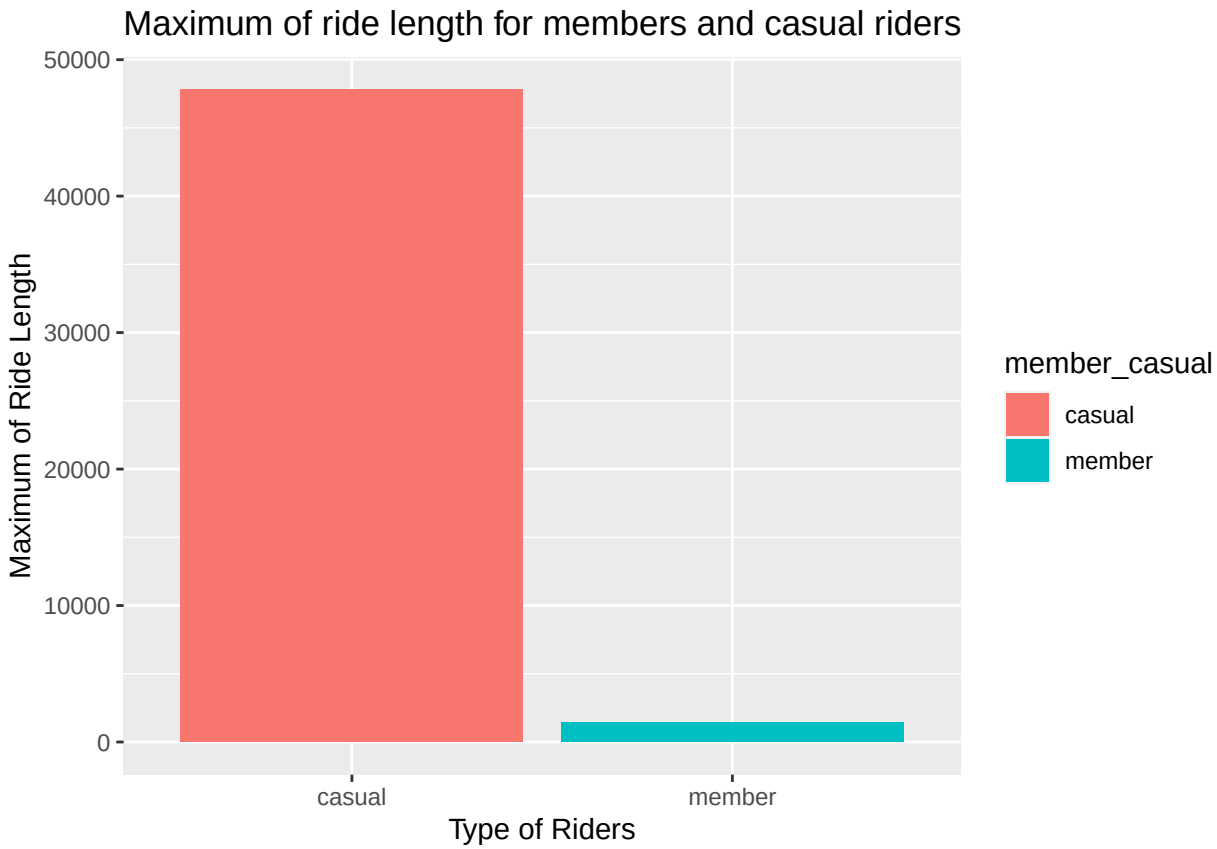
```
## $ day_of_week   <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5,~
## $ start_time    <chr> "8:42", "9:30", "10:18", "12:11", "13:12", "15:04", "16:~
```

Average of ride length for members and casual riders

```
bike_share_202104_clean %>%
  group_by(member_casual) %>%
  summarise(average_duration = mean(ride_length)) %>%
  ggplot(aes(x = member_casual, y = average_duration, fill =      member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Average of ride length for members and casual riders",
       x="Type of Riders",
       y="Average of Ride Length")
```

Maximum of ride length for members and casual riders

```
bike_share_202104_clean %>%
  group_by(member_casual) %>%
  summarise(max_duration = max(ride_length)) %>%
  ggplot(aes(x = member_casual, y = max_duration, fill =      member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Maximum of ride length for members and casual riders",
       x="Type of Riders",
       y="Maximum of Ride Length")
```



Number of trips after grouping

```
bike_share_202104_clean %>%
  group_by(member_casual, ride_length) %>%
  summarise(no_trip = n())
```

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

```
## # A tibble: 1,030 x 3
## # Groups:   member_casual [2]
##    member_casual ride_length no_trip
##    <chr>               <dbl>   <int>
##  1 casual                  1     884
##  2 casual                  2     784
##  3 casual                  3    1463
##  4 casual                  4    2457
##  5 casual                  5    3421
##  6 casual                  6    3981
##  7 casual                  7    4273
##  8 casual                  8    4555
##  9 casual                  9    4640
## 10 casual                 10    4443
## # ... with 1,020 more rows
```

Distribution of ride length for members and casual riders

```
bike_share_202104_clean %>%
  group_by(member_casual, ride_length) %>%
  summarise(no_trip = n()) %>%

  ggplot(aes(x = ride_length, y = no_trip, color = member_casual)) +
  facet_wrap(~member_casual) +
  theme(axis.text.x = element_text(angle = 45))+
  geom_point(position = "dodge")+
  labs(title="Distribution of ride length for members and casual riders",
       x="Ride Length",
       y="Number of Trips")
```

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

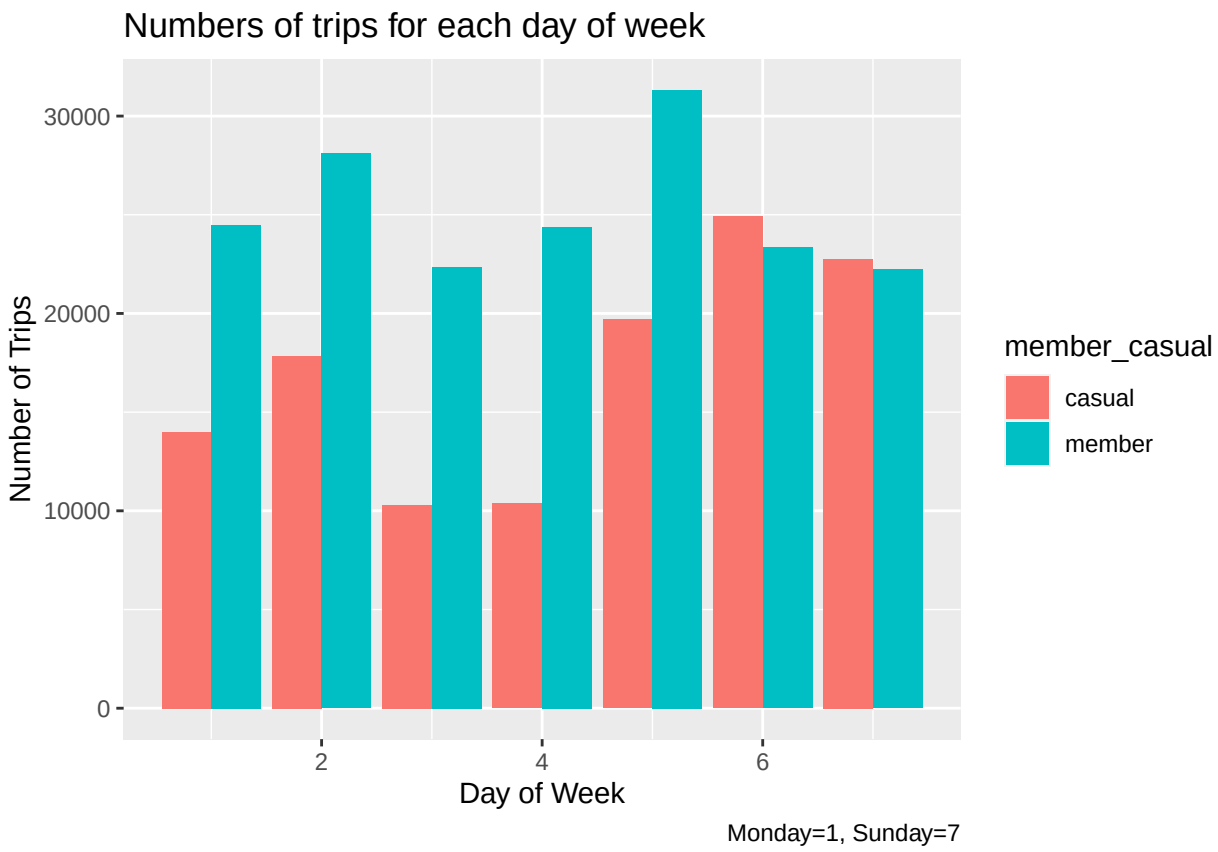
```
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```

Calculate the count of trips for each day of week (Monday=1, Sunday=7)

```
bike_share_202104_clean %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
```

```
  ggplot(aes(x = day_of_week, y = number_of_rides, fill =      member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Numbers of trips for each day of week",
       caption="Monday=1, Sunday=7",
       x="Day of Week",
       y="Number of Trips")
```

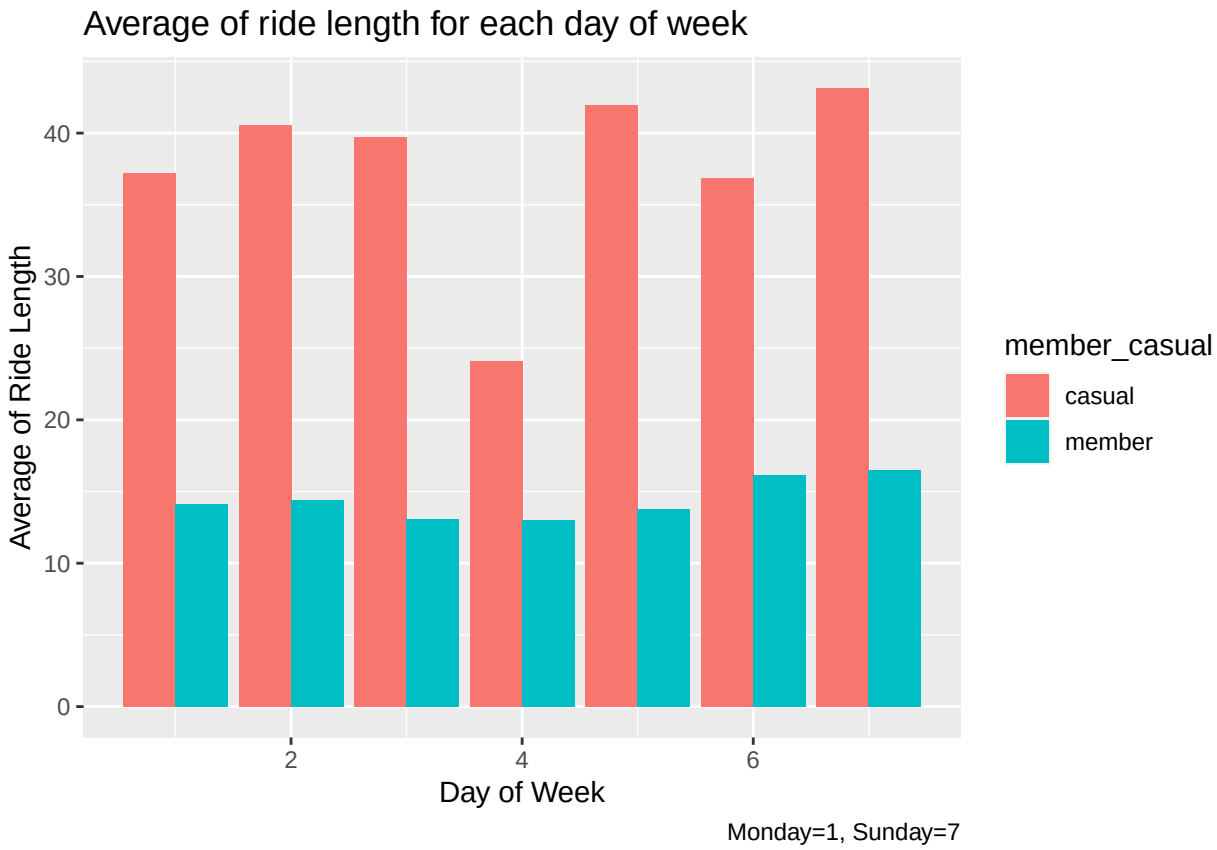
```
## `summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
```



Calculate the average of ride length for each day of week (Monday=1, Sunday=7)

```
bike_share_202104_clean %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill =      member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Average of ride length for each day of week",
       caption="Monday=1, Sunday=7",
       x="Day of Week",
       y="Average of Ride Length")
```

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


Average of ride length for each day of week
Average of Ride Length
40
30
20
10
0
2
4
6
Day of Week
Monday=1, Sunday=7
member_casual
casual
member