GDAC Case Study 1

Terry Li

2024-04-16

```
library(tidyr)
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(lubridate)
```

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

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

```
library(readr)
library(ggplot2)
```

1. Processing Stage

1.1 Import Data

```
a <- read.csv("202304.csv")
b <- read.csv("202305.csv")
c <- read.csv("202306.csv")
d <- read.csv("202307.csv")
e <- read.csv("202308.csv")
f <- read.csv("202309.csv")
```

```r
g <- read.csv("202310.csv")
h <- read.csv("202311.csv")
i <- read.csv("202312.csv")
j <- read.csv("202401.csv")
k <- read.csv("202402.csv")
l <- read.csv("202403.csv")
```

```r
data <- bind_rows(a, b, c, d, e, f, g, h, i, j, k, l)
```

```r
str(data)
```

```
## 'data.frame':    5750177 obs. of  13 variables:
##  $ ride_id           : chr  "8FE8F7D9C10E88C7" "34E4ED3ADF1D821B" "5296BF07A2F77CB5" "40759916B76D5D5
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2023-04-02 08:37:28" "2023-04-19 11:29:02" "2023-04-19 08:41:22" "2023-0
##  $ ended_at          : chr  "2023-04-02 08:41:37" "2023-04-19 11:52:12" "2023-04-19 08:43:22" "2023-0
##  $ start_station_name: chr  "" "" "" "" ...
##  $ start_station_id  : chr  "" "" "" "" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.8 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.8 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

1.2 Arrange and Mutate Data

```r
ordered_data <- data %>%
  mutate(started_at = as_datetime(started_at), ended_at = as_datetime(ended_at)) %>%
  arrange(started_at)
```

```r
mutated_data <- ordered_data %>%
  mutate(ride_length = as.numeric(ended_at - started_at), day_of_week = wday(started_at, label = TRUE),
```

1.3 Verify and Cleaning Data

```r
mutated_data %>%
  filter(is.na(ride_id) | is.na(started_at) | is.na(ended_at) | is.na(start_lat) | is.na(start_lng) | is
  count()
```

```
##      n
## 1 7566
```

```r
mutated_data %>%
  filter(is.na(ride_id)) %>%
  count()
```

```
##   n
## 1 0
```

```
mutated_data %>%
  filter(is.na(started_at) | is.na(ended_at)) %>%
  count()
```

```
##   n
## 1 0
```

```
mutated_data %>%
  filter(is.na(start_lat) | is.na(start_lng)) %>%
  count()
```

```
##   n
## 1 0
```

```
mutated_data %>%
  filter(is.na(end_lat) | is.na(end_lng)) %>%
  count()
```

```
##      n
## 1 7566
```

```
mutated_data %>%
  filter(is.na(member_casual)) %>%
  count()
```

```
##   n
## 1 0
```

```
# We can see that under end_lat and end_lng there has 7566 observations contain null values.
```

```
mutated_data %>%
  reframe(range(ride_length))
```

```
##   range(ride_length)
## 1            -999391
## 2            5909344
```

```
mutated_data %>%
  filter(ride_length <= 0) %>%
  count()
```

```
##      n
## 1 1464
```

```
# Ride_length of a set of observations are negative or zero which require removal.
```

```
processed_data <- mutated_data %>%
  filter(!is.na(end_lat), !is.na(end_lng), ride_length > 0)
```

```
processed_data %>%
  distinct(ride_id) %>%
  count()
```

```
##         n
## 1 5741147
```

```
str(processed_data)
```

```
## 'data.frame':    5741147 obs. of  16 variables:
##  $ ride_id           : chr  "563BB19A89F51F15" "AD304476EF192169" "F4490F618609D351" "08848F48F7ACF6
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : POSIXct, format: "2023-04-01 00:00:02" "2023-04-01 00:00:07" ...
##  $ ended_at          : POSIXct, format: "2023-04-01 00:07:04" "2023-04-01 00:03:10" ...
##  $ start_station_name: chr  "Wentworth Ave & 35th St" "Sheffield Ave & Wrightwood Ave" "Stave St & A
##  $ start_station_id  : chr  "KA1503000005" "TA1309000023" "13266" "" ...
##  $ end_station_name  : chr  "Halsted St & 35th St" "Sheffield Ave & Webster Ave" "" "" ...
##  $ end_station_id    : chr  "TA1308000043" "TA1309000033" "" "" ...
##  $ start_lat         : num  41.8 41.9 41.9 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.8 41.9 41.9 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "member" "casual" "casual" ...
##  $ ride_length       : num  422 183 233 327 931 263 948 244 893 36 ...
##  $ day_of_week       : Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ month             : Ord.factor w/ 12 levels "Jan"<"Feb"<"Mar"<..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
write.csv(processed_data, "Processed data.csv")
```

2. Analysis Stage

2.1 Summarise and Aggregate Data

```
summarised_data <- processed_data %>%
  summarise(avg_ride = mean(ride_length), med_ride = median(ride_length), max_ride = max(ride_length), m
summarised_data
```

```
##   avg_ride med_ride max_ride min_ride
## 1 921.5734      577   728178        1
```

```
group_compare <- processed_data %>%
  group_by(member_casual) %>%
  summarise(ride_count = n(), avg_ride = mean(ride_length), med_ride = median(ride_length), max_ride = m
group_compare
```

```
## # A tibble: 2 x 6
##   member_casual ride_count avg_ride med_ride max_ride min_ride
##   <chr>              <int>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 casual           2061399    1255.      717   728178        1
## 2 member           3679748     735.      517    89996        1
```

```
group_compare_wk <- processed_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(ride_count = n(), avg_ride = mean(ride_length), med_ride = median(ride_length), max_ride = m
```

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

```
group_compare_wk
```

```
## # A tibble: 14 x 7
## # Groups:   member_casual [2]
##    member_casual day_of_week ride_count avg_ride med_ride max_ride min_ride
##    <chr>         <ord>            <int>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 casual        Sun             334990    1459.      841   229104        1
##  2 casual        Mon             235568    1242.      687    89997        1
##  3 casual        Tue             243441    1135.      647   728178        1
##  4 casual        Wed             246653    1073.      623   322740        1
##  5 casual        Thu             272137    1090.      635   413473        1
##  6 casual        Fri             311607    1216.      701   198050        1
##  7 casual        Sat             417003    1416.      838   669136        1
##  8 member        Sun             407224     814.      554    89994        1
##  9 member        Mon             500232     702.      493    89996        1
## 10 member        Tue             571182     710.      507    89993        1
## 11 member        Wed             589657     705.      505    89996        1
## 12 member        Thu             601523     704.      507    89995        1
## 13 member        Fri             529866     728.      509    89995        1
## 14 member        Sat             480064     812.      564    89995        1
```

```
group_compare_mth <- processed_data %>%
  group_by(member_casual, month) %>%
  summarise(ride_count = n(), avg_ride = mean(ride_length), med_ride = median(ride_length), max_ride = m
```

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

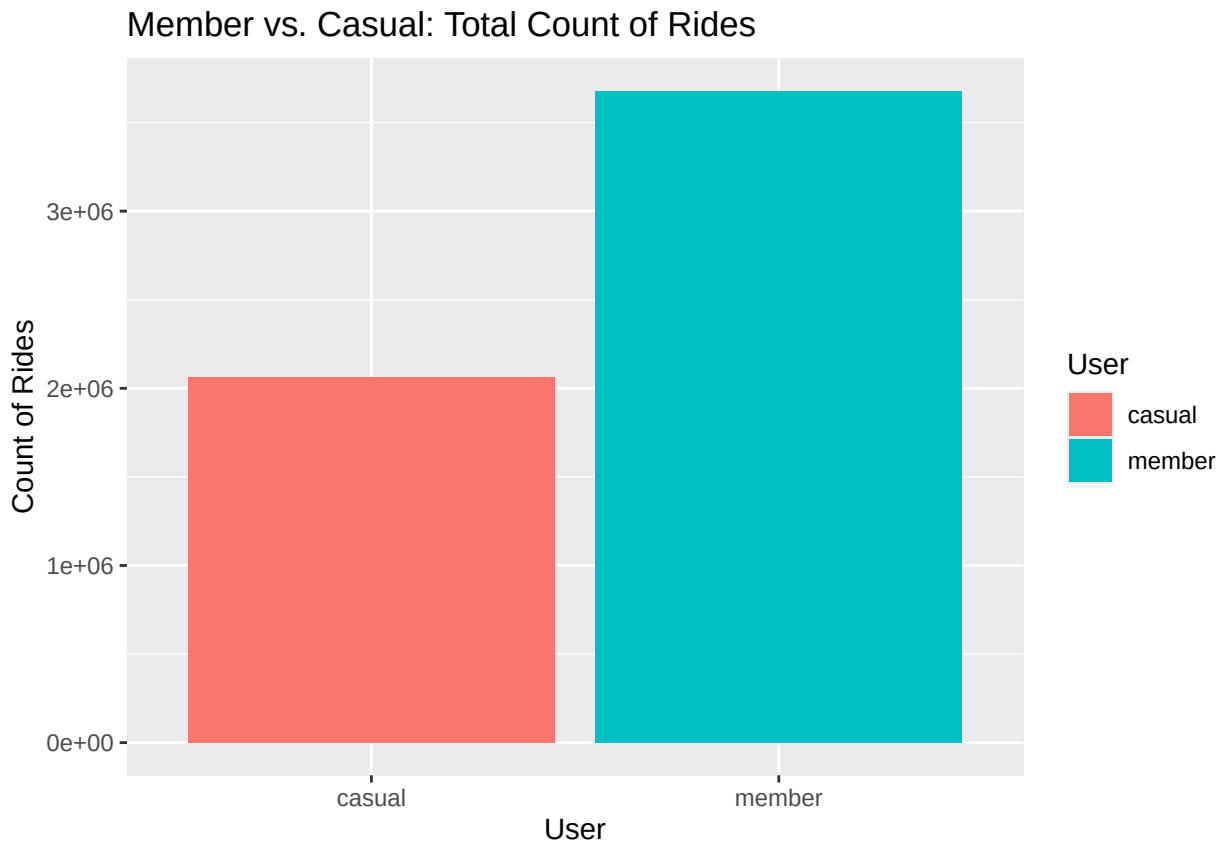
```
group_compare_mth
```

```
## # A tibble: 24 x 7
## # Groups:   member_casual [2]
##    member_casual month ride_count avg_ride med_ride max_ride min_ride
##    <chr>         <ord>      <int>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 casual        Jan        24339     889.      451    89997        1
##  2 casual        Feb        46957    1135.      577    89996        1
##  3 casual        Mar        82218    1194.      650    90562        1
##  4 casual        Apr       146878    1224.      675   140961        1
```

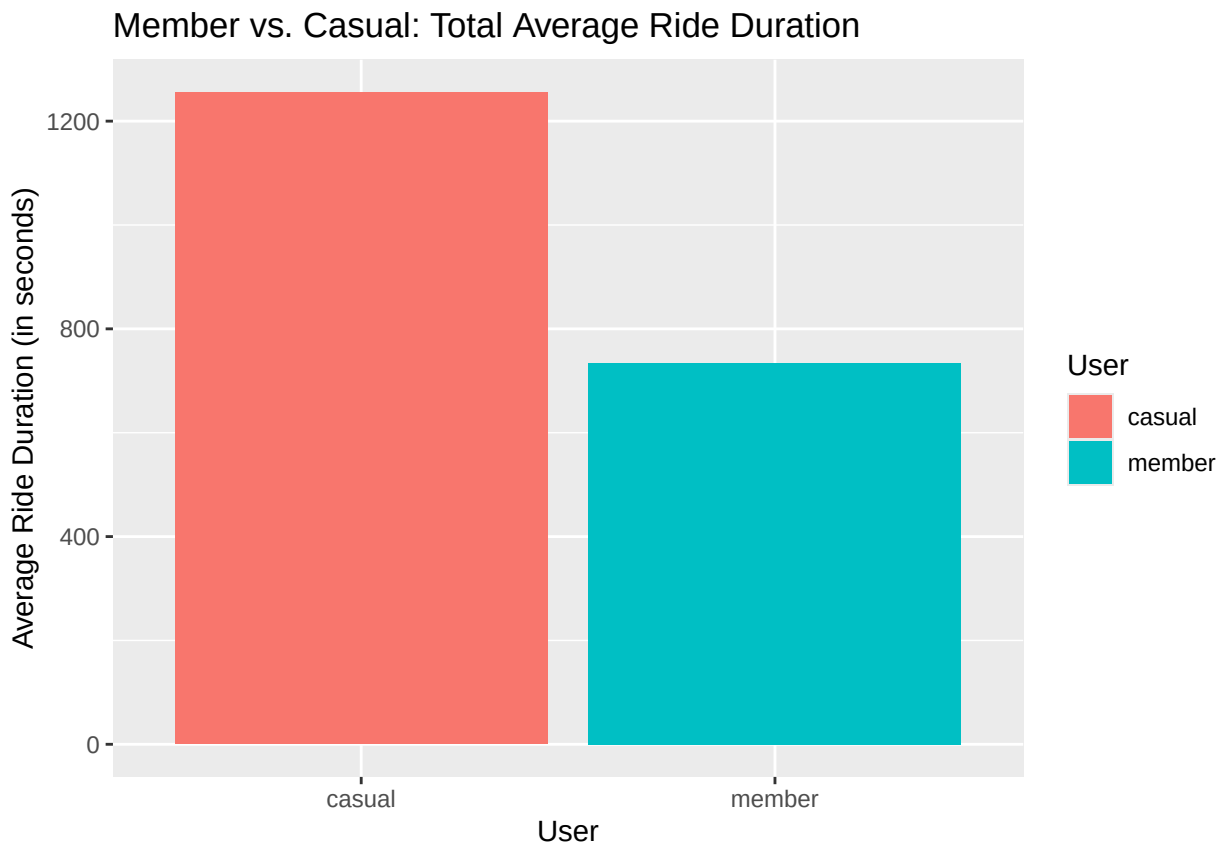
```
##  5 casual        May       233563    1321.      766   728178        1
##  6 casual        Jun       300408    1303.      773   669136        1
##  7 casual        Jul       330142    1363.      802   147471        1
##  8 casual        Aug       309931    1318.      771   413473        1
##  9 casual        Sep       260836    1275.      738    89994        1
## 10 casual        Oct       176553    1145.      634    89996        1
## # i 14 more rows
```

2.2 Visualisations

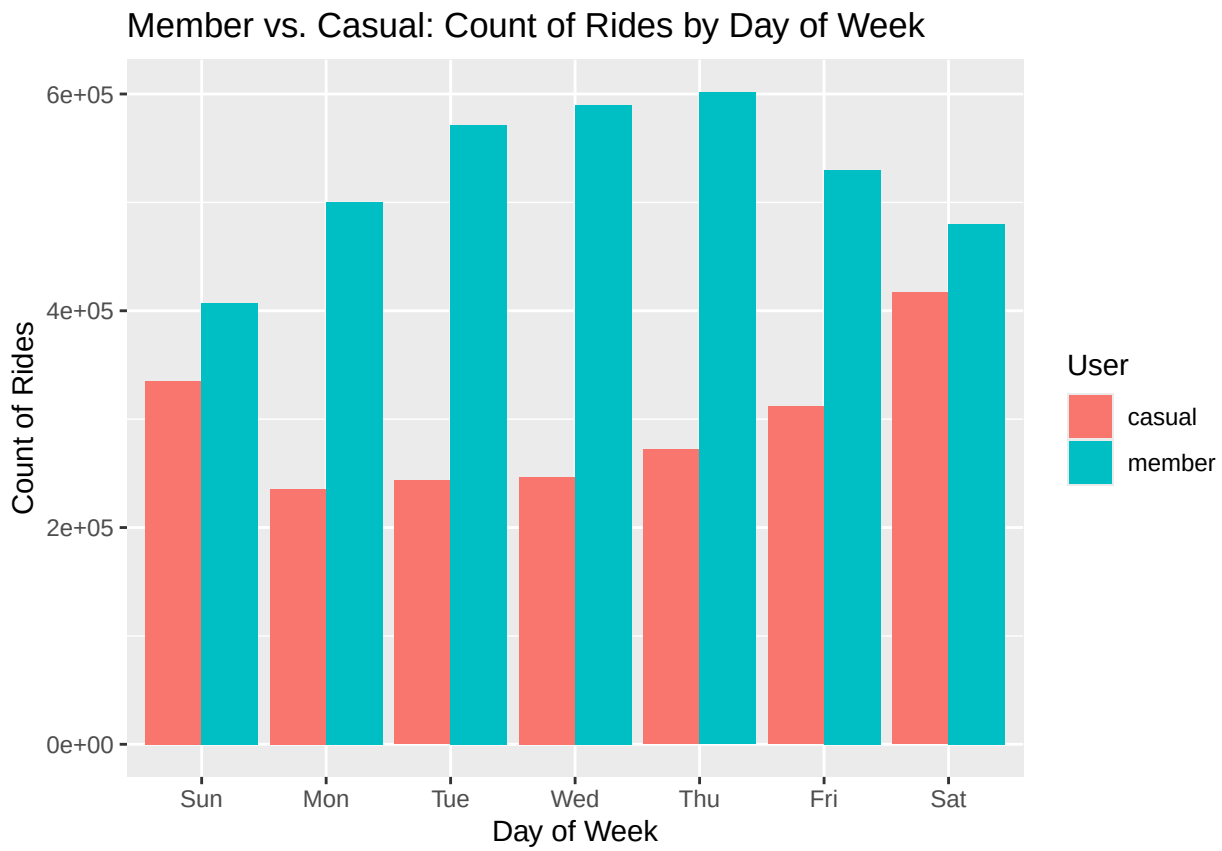
```
ggplot(group_compare, aes(member_casual, ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Total Count of Rides", x = "User", y = "Count of Rides", fill = "Use
```



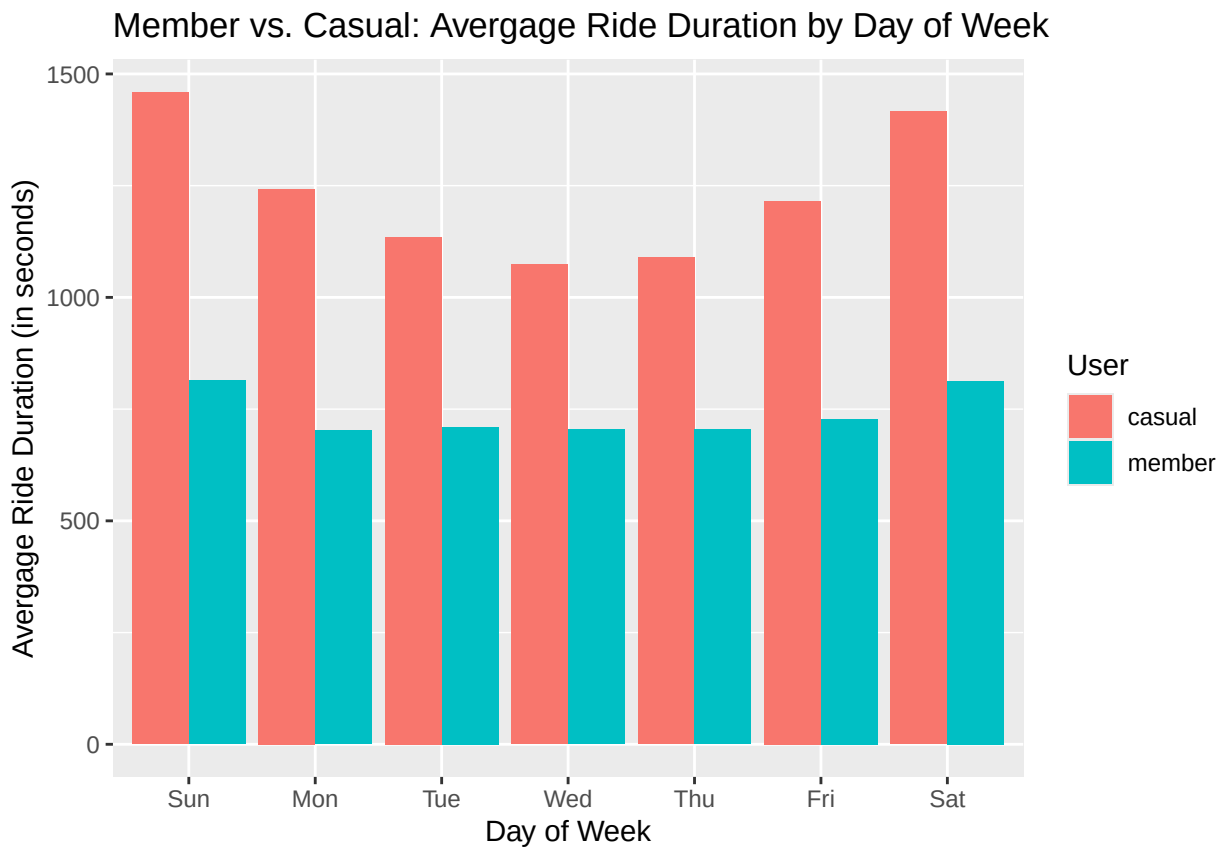

```
ggplot(group_compare, aes(member_casual, avg_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Total Average Ride Duration", x = "User", y = "Average Ride Duration
```



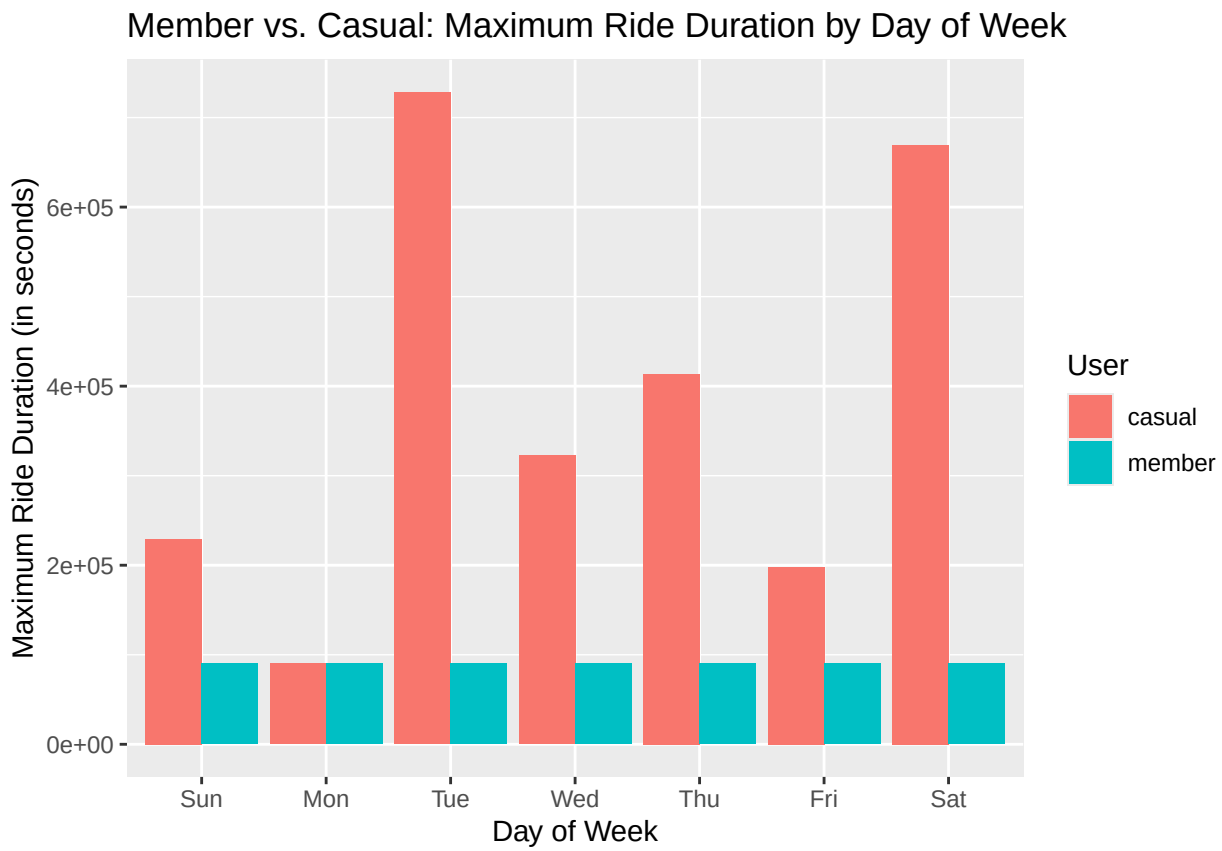

```
ggplot(group_compare_wk, aes(day_of_week, ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Count of Rides by Day of Week", x = "Day of Week", y = "Count of Rid
```



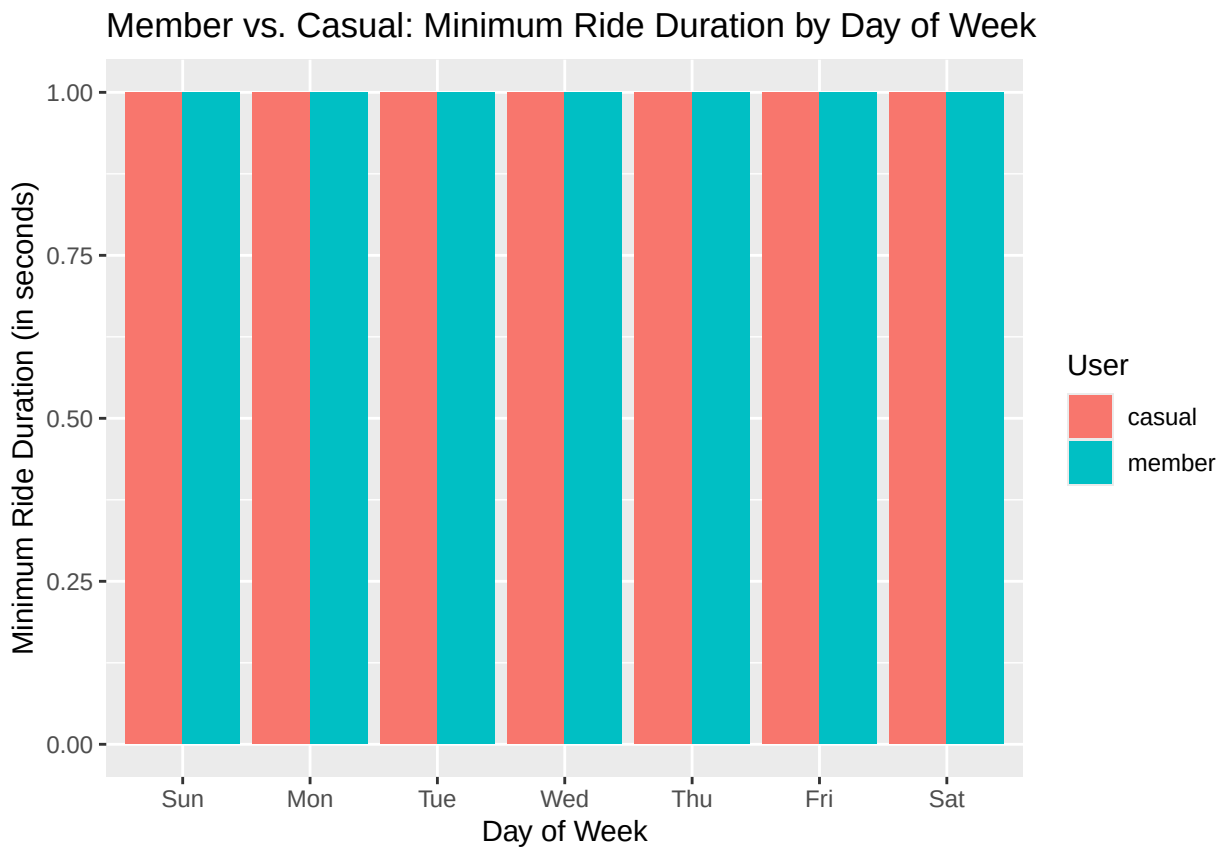

```
ggplot(group_compare_wk, aes(day_of_week, avg_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Avergage Ride Duration by Day of Week", x = "Day of Week", y = "Averg
```



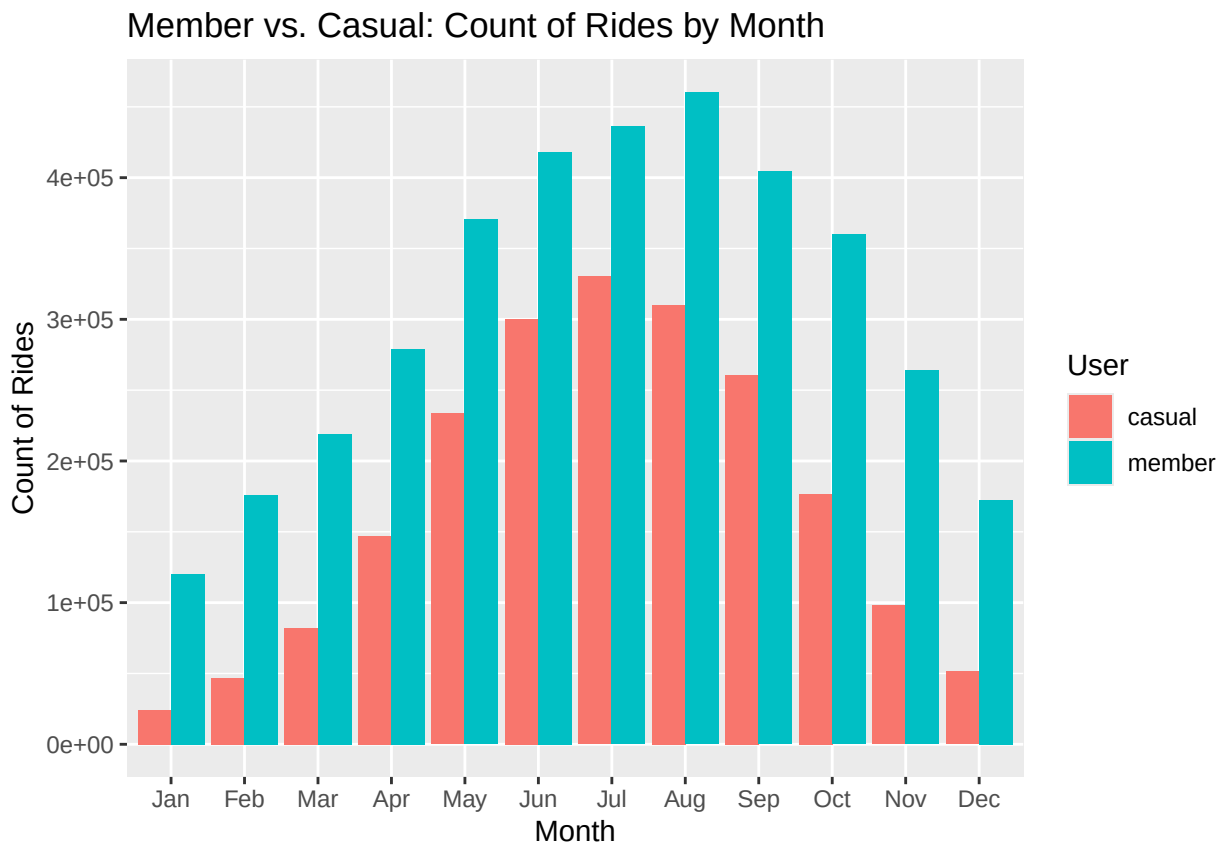

```
ggplot(group_compare_wk, aes(day_of_week, max_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Maximum Ride Duration by Day of Week", x = "Day of Week", y = "Maximu
```



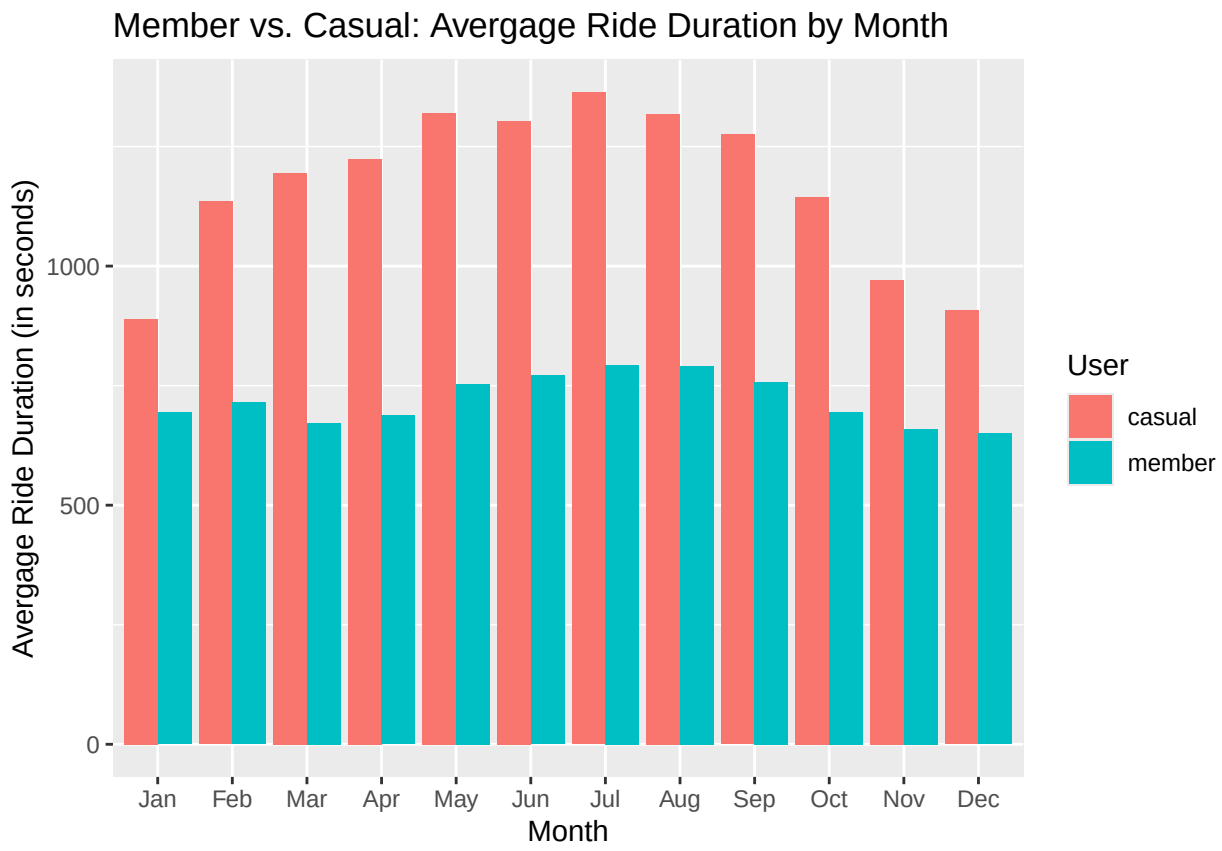

```
ggplot(group_compare_wk, aes(day_of_week, min_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Minimum Ride Duration by Day of Week", x = "Day of Week", y = "Minim
```



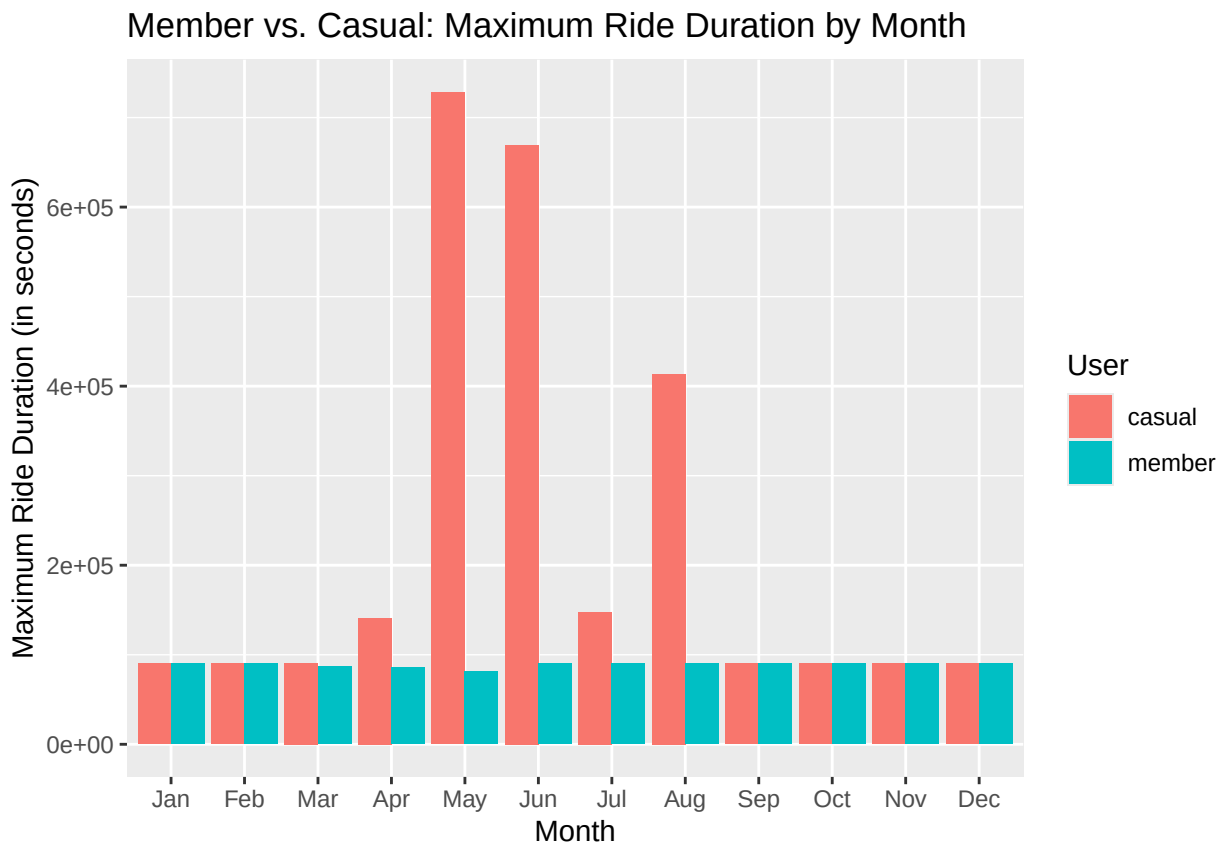

```
ggplot(group_compare_mth, aes(month, ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Count of Rides by Month", x = "Month", y = "Count of Rides", fill =
```



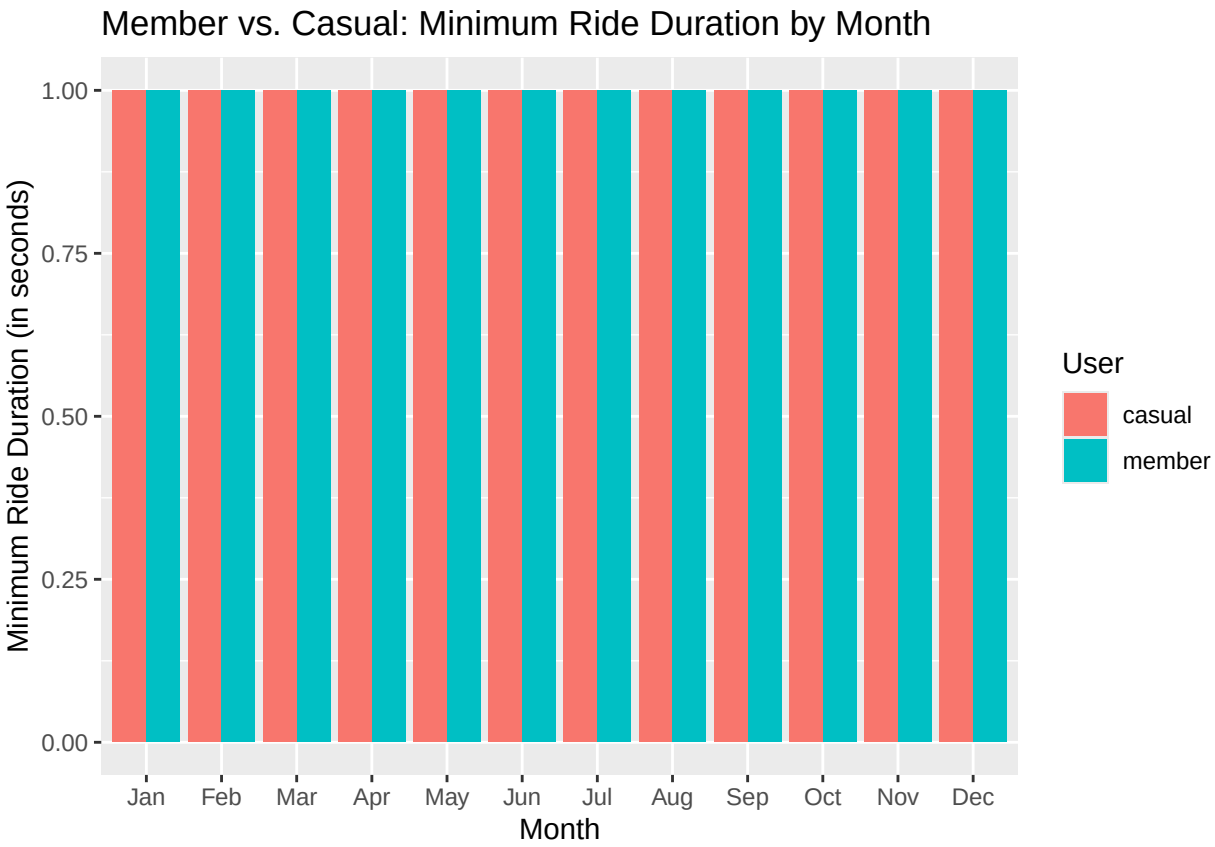

```
ggplot(group_compare_mth, aes(month, avg_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Avergage Ride Duration by Month", x = "Month", y = "Avergage Ride Du
```




```
ggplot(group_compare_mth, aes(month, max_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Maximum Ride Duration by Month", x = "Month", y = "Maximum Ride Dura
```




```
ggplot(group_compare_mth, aes(month, min_ride, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Member vs. Casual: Minimum Ride Duration by Month", x = "Month", y = "Minimum Ride Dura
```


Member vs. Casual: Minimum Ride Duration by Month
Minimum Ride Duration (in seconds)
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
User
casual
member