Analysis of Cyclistic Data

Michael-Philipp Stiebing

2023-03-02

Starting with processed dataframe all_trips_v5

- Conduct descriptive analysis as per script

```
# Descriptive analysis on ride_length (all figures in seconds)
mean(all_trips_v5$ride_length) #straight average (total ride length / rides)
```

```
## [1] 998.9599
```

```
median(all_trips_v5$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 634
```

```
max(all_trips_v5$ride_length) #longest ride
```

```
## [1] 86391
```

```
min(all_trips_v5$ride_length) #shortest ride
```

```
## [1] 61
```

```
# You can condense the four lines above to one line using summary() on the specific attribute
summary(all_trips_v5$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      61     366     634     999    1129   86391
```

```
# Compare members and casual users
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual, FUN = mean)
```

```
##   all_trips_v5$member_casual all_trips_v5$ride_length
## 1                     casual                1346.5537
## 2                     member                 758.8174
```

```
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual, FUN = median)
```

```
##   all_trips_v5$member_casual all_trips_v5$ride_length
## 1                     casual                     808
## 2                     member                     540
```

```
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual, FUN = max)
```

```
##   all_trips_v5$member_casual all_trips_v5$ride_length
## 1                     casual                    86391
## 2                     member                    86180
```

```
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual, FUN = min)
```

```
##   all_trips_v5$member_casual all_trips_v5$ride_length
## 1                     casual                       61
## 2                     member                       61
```

```
# See the average ride time by each day for members vs casual users
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual + all_trips_v5$day_of_week, FUN = mean)
```

```
##    all_trips_v5$member_casual all_trips_v5$day_of_week all_trips_v5$ride_length
## 1                      casual                  Friday                1261.1652
## 2                      member                  Friday                 745.0343
## 3                      casual                  Monday                1366.7836
## 4                      member                  Monday                 731.2731
## 5                      casual                Saturday                1503.8461
## 6                      member                Saturday                 848.3125
## 7                      casual                  Sunday                1545.3593
## 8                      member                  Sunday                 841.4815
## 9                      casual                Thursday                1185.8862
## 10                     member                Thursday                 728.3687
## 11                     casual                 Tuesday                1205.8529
## 12                     member                 Tuesday                 721.9373
## 13                     casual               Wednesday                1161.0311
## 14                     member               Wednesday                 721.9887
```

```
# Notice that the days of the week are out of order. Let's fix that.
all_trips_v5$day_of_week <- ordered(all_trips_v5$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Frid


# Now, let's run the average ride time by each day for members vs casual users
aggregate(all_trips_v5$ride_length ~ all_trips_v5$member_casual + all_trips_v5$day_of_week, FUN = mean)
```

```
##    all_trips_v5$member_casual all_trips_v5$day_of_week all_trips_v5$ride_length
## 1                      casual                  Sunday                1545.3593
## 2                      member                  Sunday                 841.4815
## 3                      casual                  Monday                1366.7836
## 4                      member                  Monday                 731.2731
## 5                      casual                 Tuesday                1205.8529
## 6                      member                 Tuesday                 721.9373
## 7                      casual               Wednesday                1161.0311
## 8                      member               Wednesday                 721.9887
## 9                      casual                Thursday                1185.8862
## 10                     member                Thursday                 728.3687
## 11                     casual                  Friday                1261.1652
## 12                     member                  Friday                 745.0343
## 13                     casual                Saturday                1503.8461
## 14                     member                Saturday                 848.3125
```

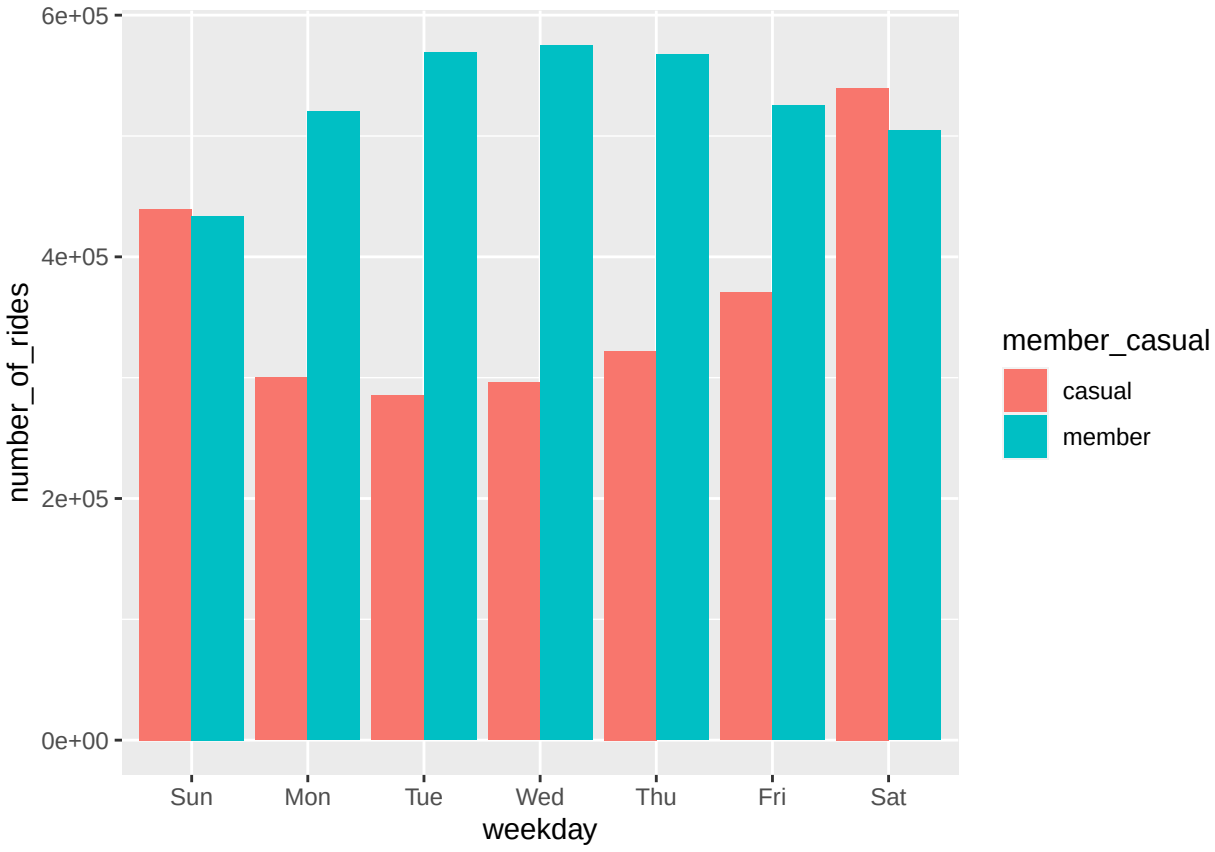
```
# analyze ridership data by type and weekday
all_trips_v5 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                                    #calculates the number of rides and average duration
  ,average_duration = mean(ride_length)) %>%              # calculates the average duration
  arrange(member_casual, weekday)                                                  # sorts
```

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

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>           <int>            <dbl>
##  1 casual        Sun            439601            1545.
##  2 casual        Mon            300340            1367.
##  3 casual        Tue            285059            1206.
##  4 casual        Wed            295860            1161.
##  5 casual        Thu            322175            1186.
##  6 casual        Fri            370221            1261.
##  7 casual        Sat            539769            1504.
##  8 member        Sun            433825             841.
##  9 member        Mon            520207             731.
## 10 member        Tue            569216             722.
## 11 member        Wed            575028             722.
## 12 member        Thu            567557             728.
## 13 member        Fri            524988             745.
## 14 member        Sat            504550             848.
```

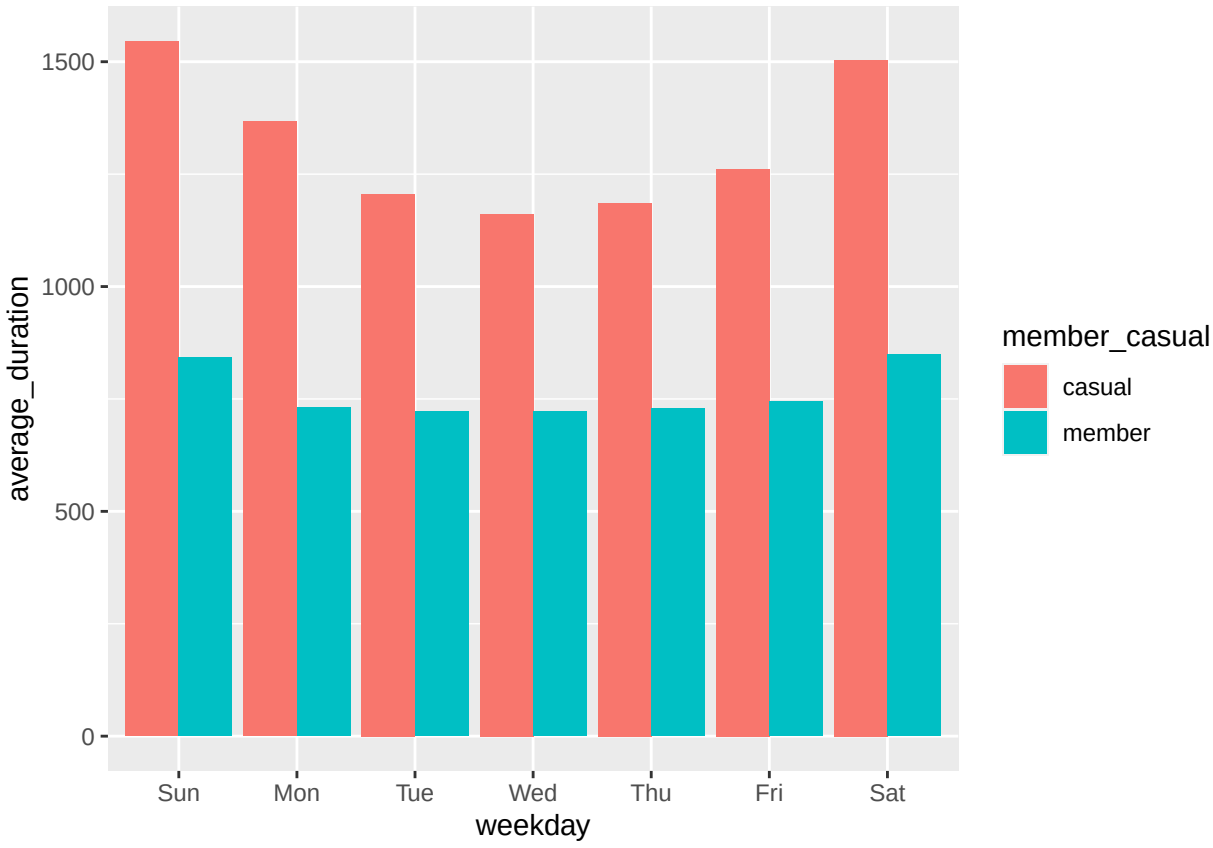
```
# Let's visualize the number of rides by rider type
all_trips_v5 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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




```
# Let's create a visualization for average duration
all_trips_v5 %>%
  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 = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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




```
#mean(round_trips$ride_length)

#mean(all_trips_v5[all_trips_v5$geodist<=0, 'ride_length'])

#mean(filter(all_trips_v5,geodist == 0)$ride_length)
```

```
# Descriptive analysis on ride_length (all figures in seconds)
mean(filter(all_trips_v5,geodist == 0)$ride_length) #straight average (total ride length / rides)
```

```
## [1] 1725.811
```

```
median(filter(all_trips_v5,geodist == 0)$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 818
```

```
max(filter(all_trips_v5,geodist == 0)$ride_length) #longest ride
```

```
## [1] 86329
```

```
min(filter(all_trips_v5,geodist == 0)$ride_length) #shortest ride
```

```
## [1] 61
```

```
# You can condense the four lines above to one line using summary() on the specific attribute
summary(filter(all_trips_v5,geodist == 0)$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    61.0   214.2   818.0  1725.8  2191.0 86329.0
```

```
# Compare members and casual users
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual, FUN = mean)
```

```
##   filter(all_trips_v5, geodist == 0)$member_casual
## 1                                           casual
## 2                                           member
##   filter(all_trips_v5, geodist == 0)$ride_length
## 1                                      2322.7035
## 2                                       936.9457
```

```
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual, FUN = median)
```

```
##   filter(all_trips_v5, geodist == 0)$member_casual
## 1                                           casual
## 2                                           member
##   filter(all_trips_v5, geodist == 0)$ride_length
## 1                                             1349
## 2                                              401
```

```
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual, FUN = max)
```

```
##   filter(all_trips_v5, geodist == 0)$member_casual
## 1                                           casual
## 2                                           member
##   filter(all_trips_v5, geodist == 0)$ride_length
## 1                                            86329
## 2                                            85136
```

```
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual, FUN = min)
```

```
##   filter(all_trips_v5, geodist == 0)$member_casual
## 1                                           casual
## 2                                           member
##   filter(all_trips_v5, geodist == 0)$ride_length
## 1                                               61
## 2                                               61
```

```
# See the average ride time by each day for members vs casual users
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual + filter(all_trips_v5,geodist == 0)
```

```
##    filter(all_trips_v5, geodist == 0)$member_casual
## 1                                            casual
## 2                                            member
## 3                                            casual
## 4                                            member
## 5                                            casual
## 6                                            member
```

```
## 7                                    casual
## 8                                    member
## 9                                    casual
## 10                                   member
## 11                                   casual
## 12                                   member
## 13                                   casual
## 14                                   member
##    filter(all_trips_v5, geodist == 0)$day_of_week
## 1                                   Sunday
## 2                                   Sunday
## 3                                   Monday
## 4                                   Monday
## 5                                   Tuesday
## 6                                   Tuesday
## 7                                 Wednesday
## 8                                 Wednesday
## 9                                  Thursday
## 10                                 Thursday
## 11                                   Friday
## 12                                   Friday
## 13                                 Saturday
## 14                                 Saturday
##    filter(all_trips_v5, geodist == 0)$ride_length
## 1                                2559.3728
## 2                                1059.8977
## 3                                2421.0331
## 4                                 940.2018
## 5                                2189.9315
## 6                                 899.6108
## 7                                2144.9871
## 8                                 887.2209
## 9                                2141.5593
## 10                                882.4784
## 11                               2107.7701
## 12                                890.4508
## 13                               2426.2851
## 14                               1007.8698
```

```
# Notice that the days of the week are out of order. Let's fix that.
#filter(all_trips_v5,geodist == 0)$day_of_week <- ordered(filter(all_trips_v5,geodist == 0)$day_of_week, levels=c("Sunday", "Monday", "Tu



# Now, let's run the average ride time by each day for members vs casual users
aggregate(filter(all_trips_v5,geodist == 0)$ride_length ~ filter(all_trips_v5,geodist == 0)$member_casual + filter(all_trips_v5,geodist == 0)
```

```
##    filter(all_trips_v5, geodist == 0)$member_casual
## 1                                  casual
## 2                                  member
## 3                                  casual
## 4                                  member
## 5                                  casual
## 6                                  member
## 7                                  casual
## 8                                  member
## 9                                  casual
## 10                                 member
## 11                                 casual
```

```
## 12                                member
## 13                                casual
## 14                                member
##    filter(all_trips_v5, geodist == 0)$day_of_week
## 1                               Sunday
## 2                               Sunday
## 3                               Monday
## 4                               Monday
## 5                              Tuesday
## 6                              Tuesday
## 7                            Wednesday
## 8                            Wednesday
## 9                             Thursday
## 10                            Thursday
## 11                              Friday
## 12                              Friday
## 13                            Saturday
## 14                            Saturday
##    filter(all_trips_v5, geodist == 0)$ride_length
## 1                           2559.3728
## 2                           1059.8977
## 3                           2421.0331
## 4                            940.2018
## 5                           2189.9315
## 6                            899.6108
## 7                           2144.9871
## 8                            887.2209
## 9                           2141.5593
## 10                           882.4784
## 11                          2107.7701
## 12                           890.4508
## 13                          2426.2851
## 14                          1007.8698
```

```
# analyze ridership data by type and weekday
filter(all_trips_v5,geodist == 0) %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                                    #calculates the number of rides and average duration
  ,average_duration = mean(ride_length)) %>%              # calculates the average duration
  arrange(member_casual, weekday)                                                  # sorts
```

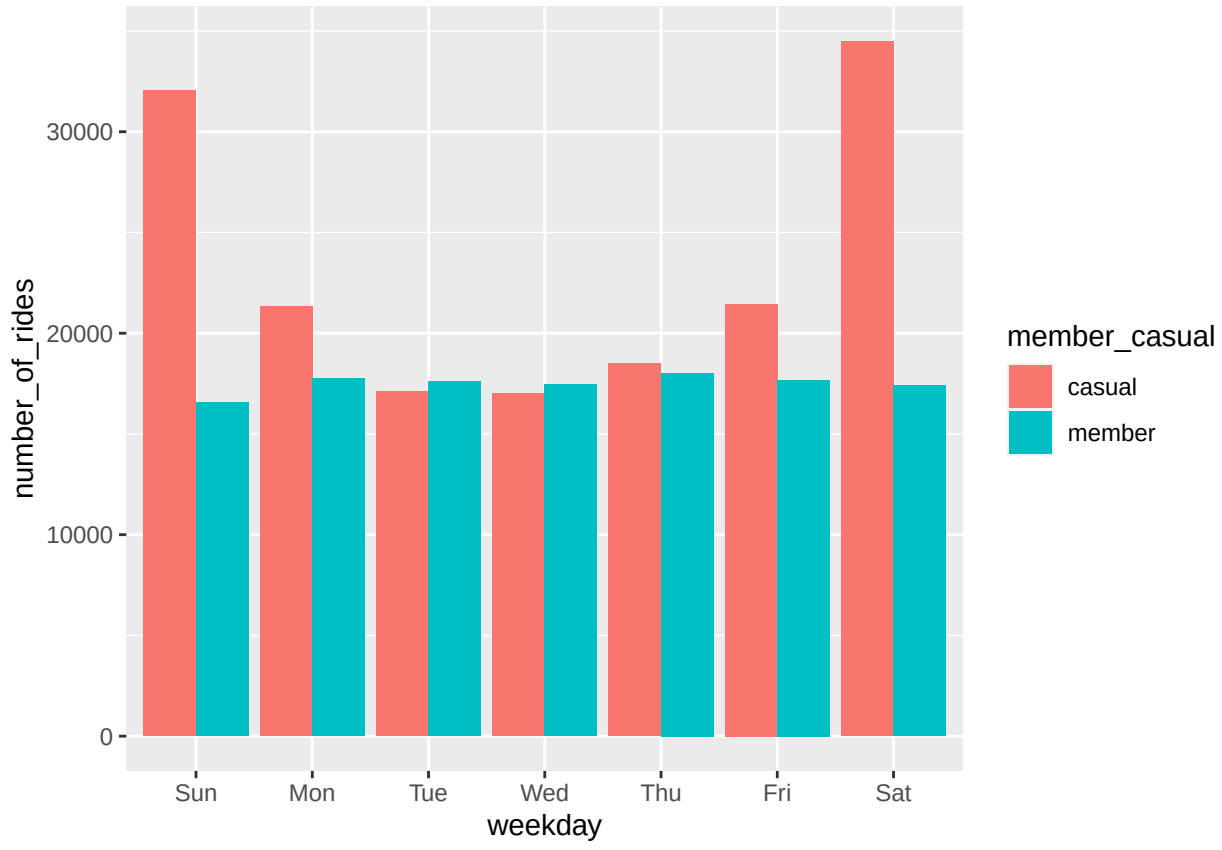
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>           <int>            <dbl>
##  1 casual        Sun             32049            2559.
##  2 casual        Mon             21330            2421.
##  3 casual        Tue             17116            2190.
##  4 casual        Wed             17018            2145.
##  5 casual        Thu             18485            2142.
##  6 casual        Fri             21445            2108.
##  7 casual        Sat             34487            2426.
##  8 member        Sun             16564            1060.
##  9 member        Mon             17766             940.
```

```
## 10 member      Tue          17598         900.
## 11 member      Wed          17467         887.
## 12 member      Thu          18027         882.
## 13 member      Fri          17685         890.
## 14 member      Sat          17417        1008.
```

```
# Let's visualize the number of rides by rider type
filter(all_trips_v5,geodist == 0) %>%
  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")
```

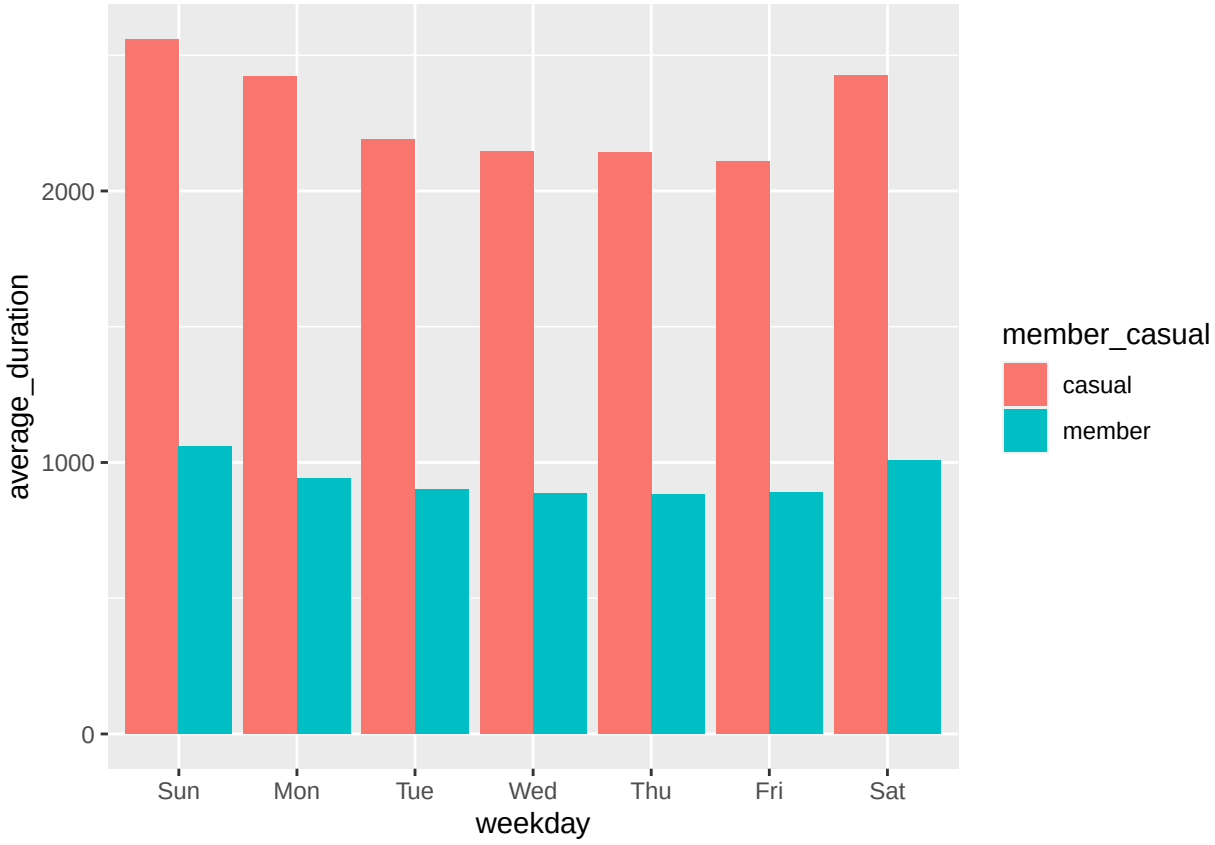
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```




```
# Let's create a visualization for average duration
filter(all_trips_v5,geodist == 0) %>%
  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 = average_duration, fill = member_casual)) +
geom_col(position = "dodge")
```

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




```
short_ride <- (filter(all_trips_v5,ride_length < 86400 & !is.na(geodist)))
long_ride <- (filter(all_trips_v5,ride_length > 86400 & !is.na(geodist)))
nadist_ride <- (filter(all_trips_v5,is.na(geodist)))

summary(filter(short_ride)$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      61     366     634     999    1129   86391
```

```
summary(filter(long_ride)$ride_length)
```

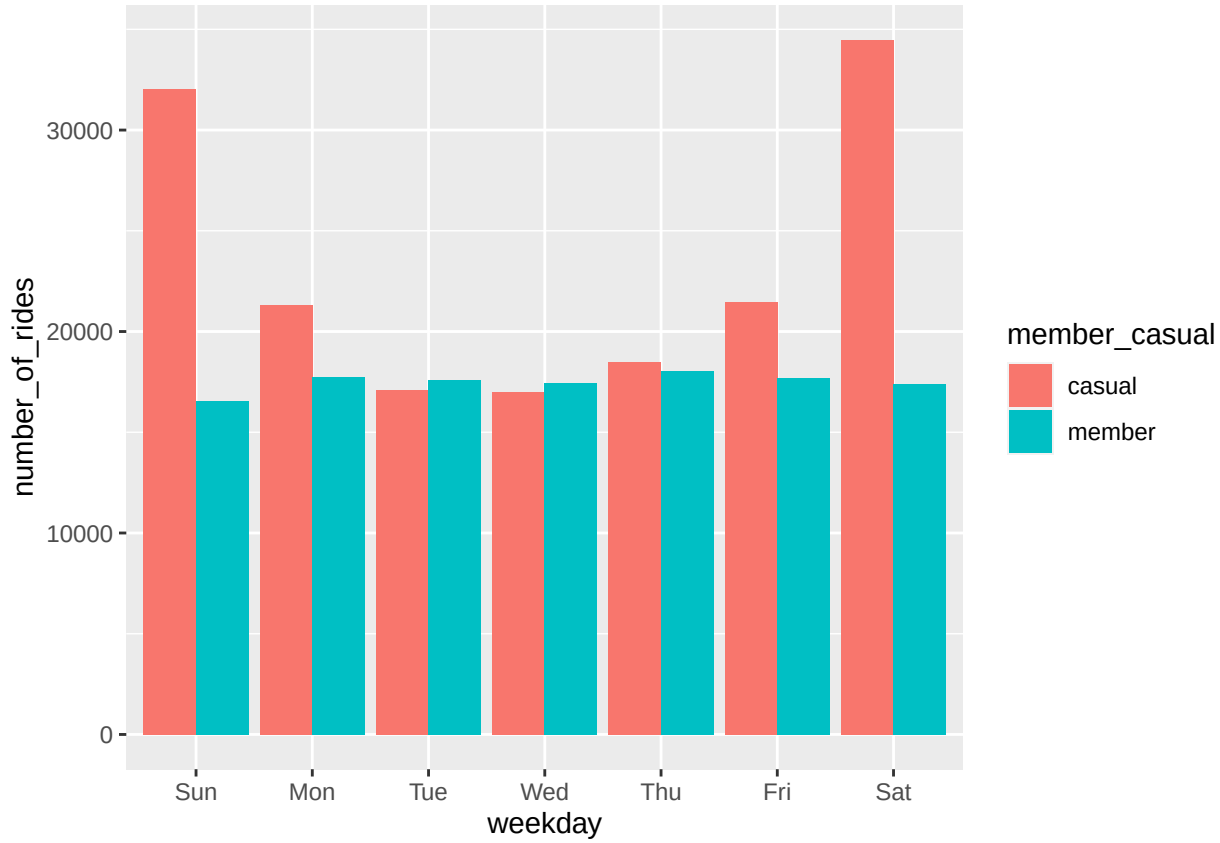
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```
summary(filter(nadist_ride)$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

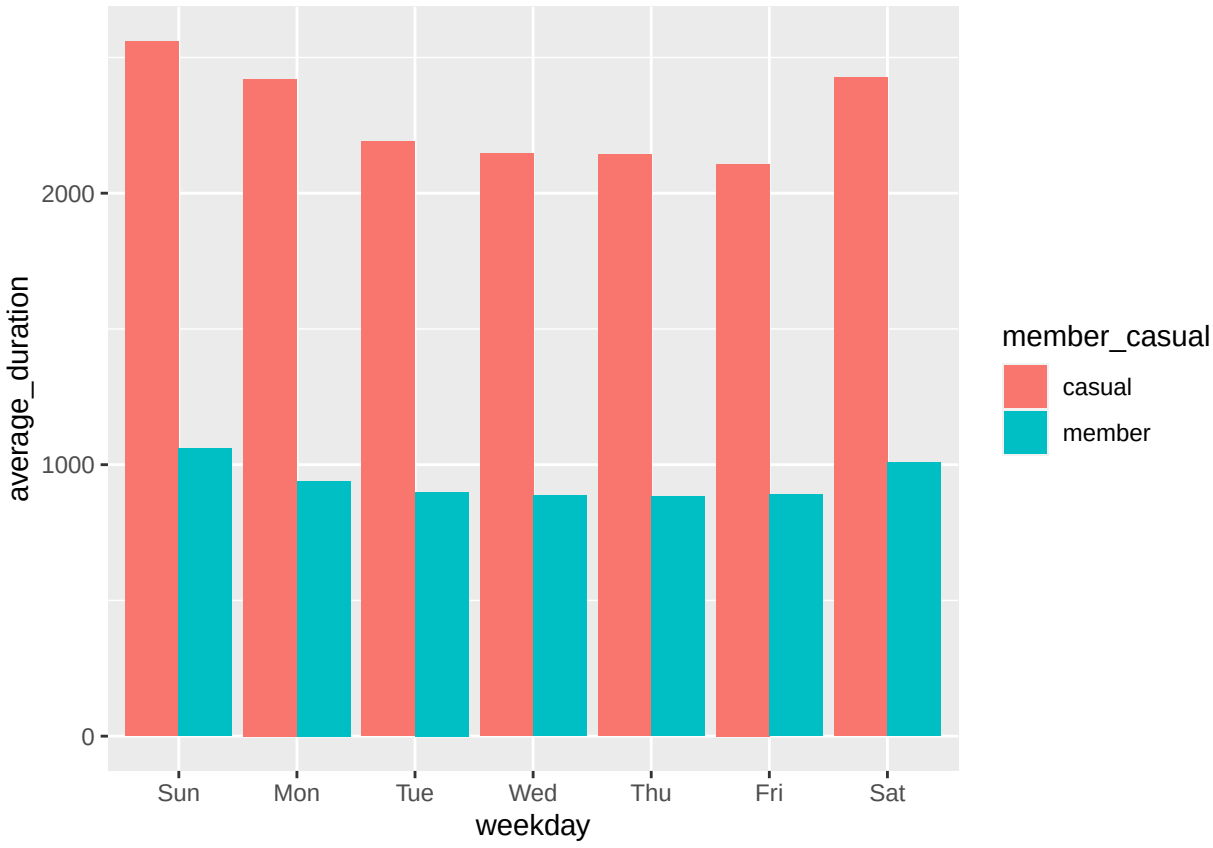
```
# Let's visualize the number of rides by rider type
filter(short_ride,geodist == 0) %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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




```
# Let's create a visualization for average duration
filter(short_ride,geodist == 0) %>%
  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 = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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




```
nrow(all_trips_v5[all_trips_v5$member_casual == "casual",])
```

```
## [1] 2553025
```

```
nrow(all_trips_v5[all_trips_v5$member_casual == "member",])
```

```
## [1] 3695371
```

```
#nrow(all_trips_v5[all_trips_v5$member_casual == "casual" & all_trips_v5$ride_length > 2700,])
nrow(all_trips_v5[all_trips_v5$member_casual == "casual" & all_trips_v5$ride_length > 2700,]) / nrow(all_trips_v5[all_trips_v5$member_
```

```
## [1] 0.1008094
```

```
nrow(all_trips_v5[all_trips_v5$member_casual == "member" & all_trips_v5$ride_length > 2700,]) / nrow(all_trips_v5[all_trips_v5$membe
```

```
## [1] 0.01653609
```

```
nrow(all_trips_v5[all_trips_v5$member_casual == "casual" & all_trips_v5$ride_length > 10800,]) / nrow(all_trips_v5[all_trips_v5$member
```

```
## [1] 0.004855025
```

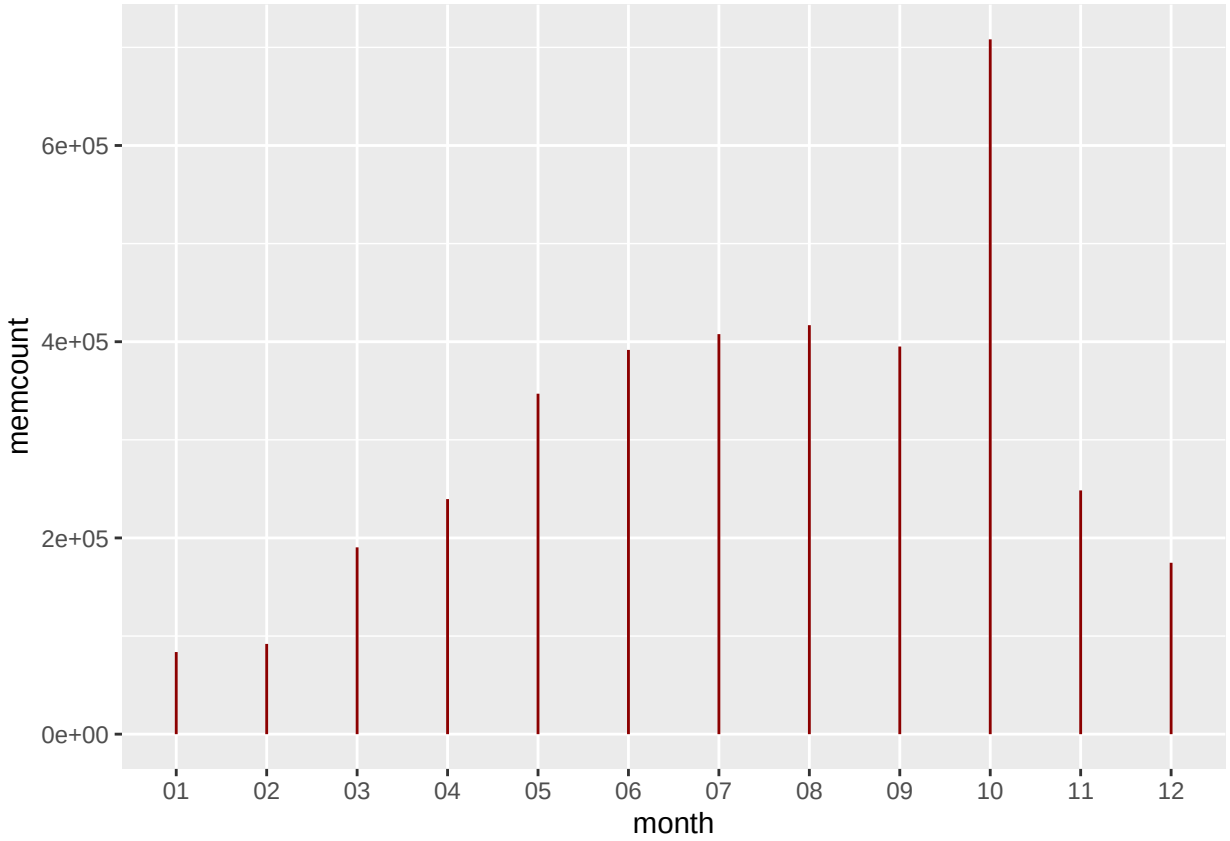
```
all_trips_v5 %>%
  group_by(month, member_casual) %>%
  summarize(cascount = sum(member_casual == "casual"), memcount = sum(member_casual == "member"))
```

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

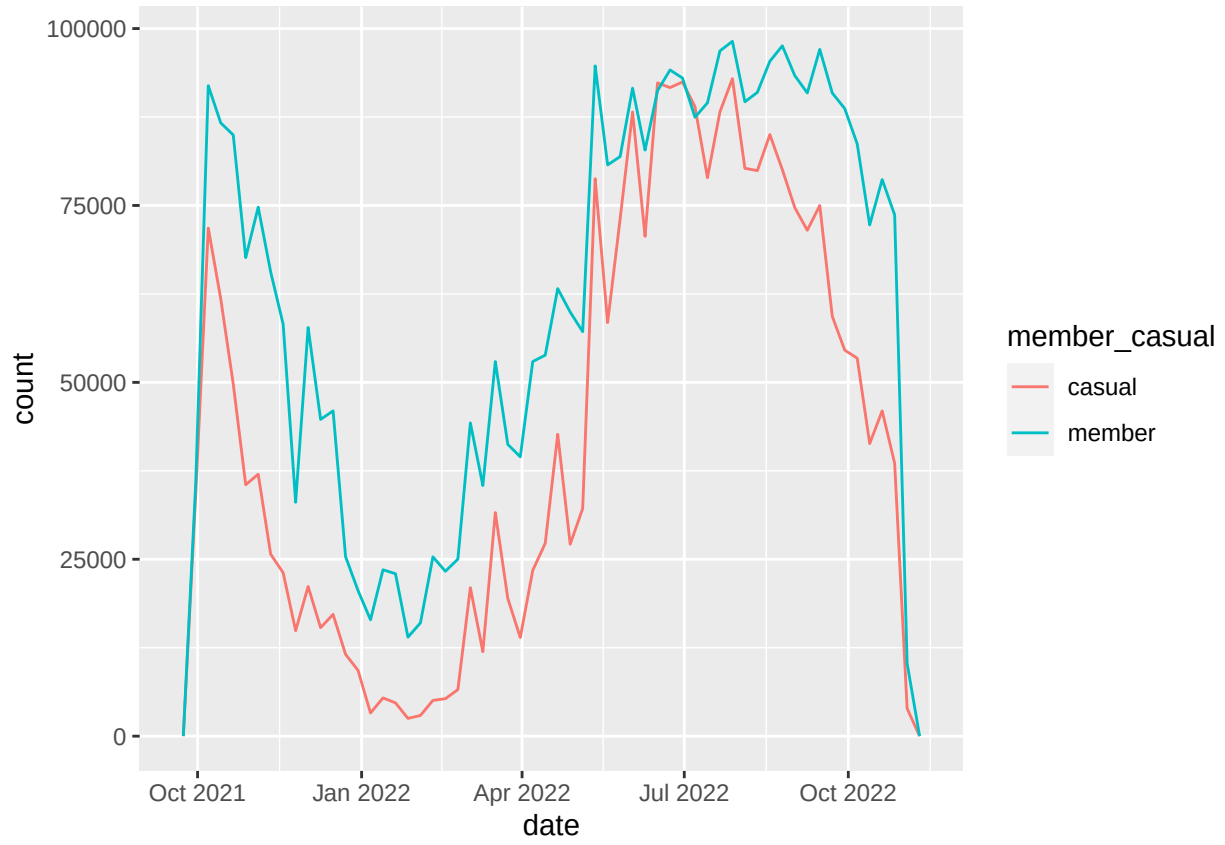
```
## # A tibble: 24 x 4
## # Groups:   month [12]
##    month member_casual cascount memcount
##    <chr> <chr>           <int>    <int>
##  1 01    casual          18069        0
##  2 01    member              0    83700
##  3 02    casual          20900        0
##  4 02    member              0    91984
##  5 03    casual          88033        0
##  6 03    member              0   190401
##  7 04    casual         123721        0
##  8 04    member              0   239611
##  9 05    casual         274367        0
## 10 05    member              0   347015
## # ... with 14 more rows
```

```
all_trips_v5 %>%
  group_by(month, member_casual) %>%
  summarize(cascount = sum(member_casual == "casual"), memcount = sum(member_casual == "member")) %>%
  arrange(month,cascount,memcount)  %>%
  ggplot(aes(x=month)) + geom_line(aes(y=memcount), color = "darkred")
```

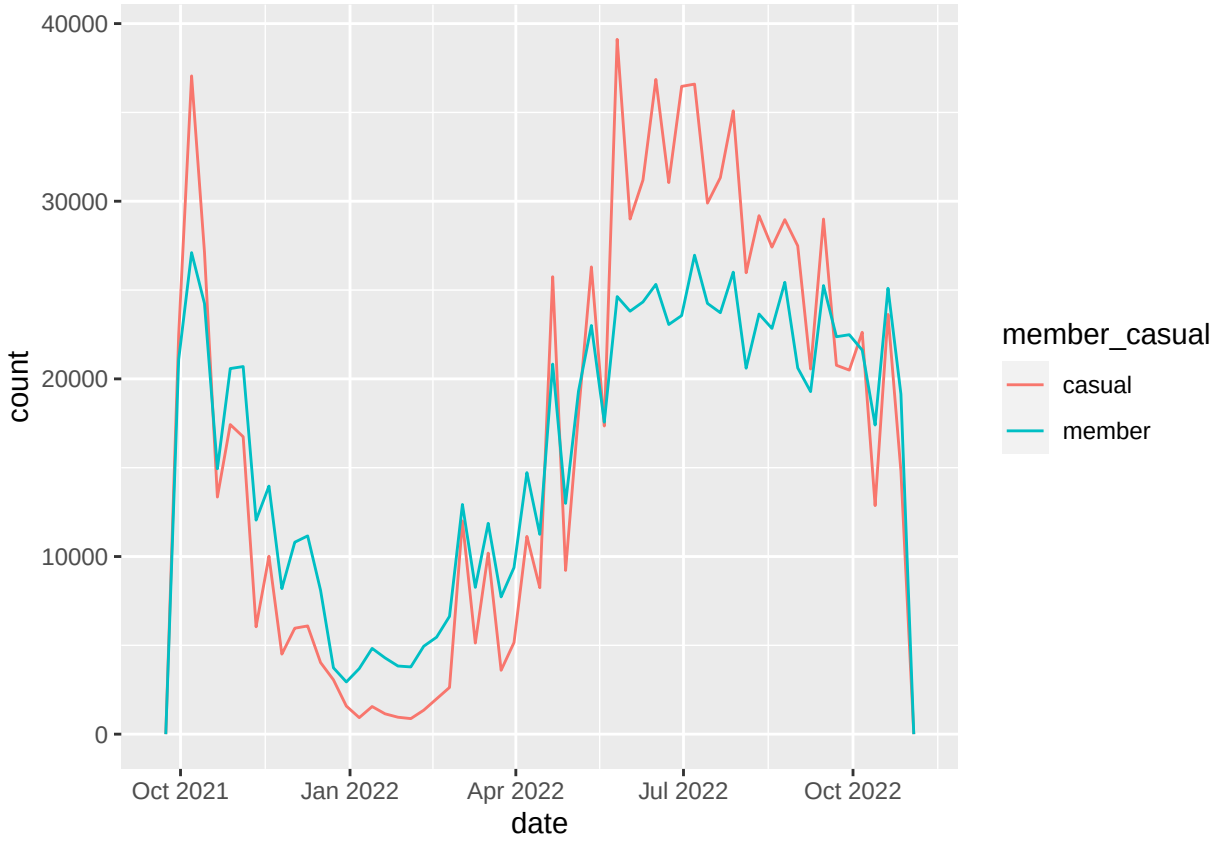
```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```



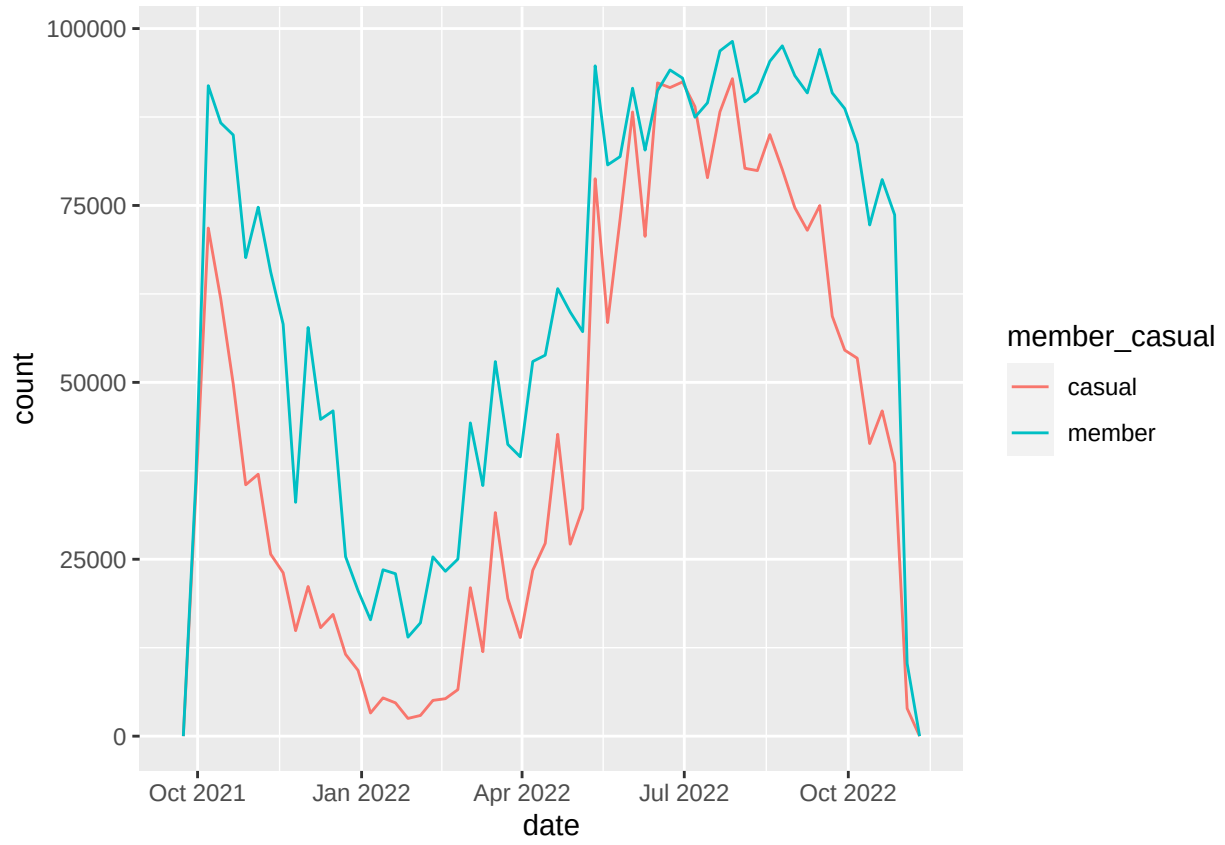

```
ggplot(all_trips_v5, aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```



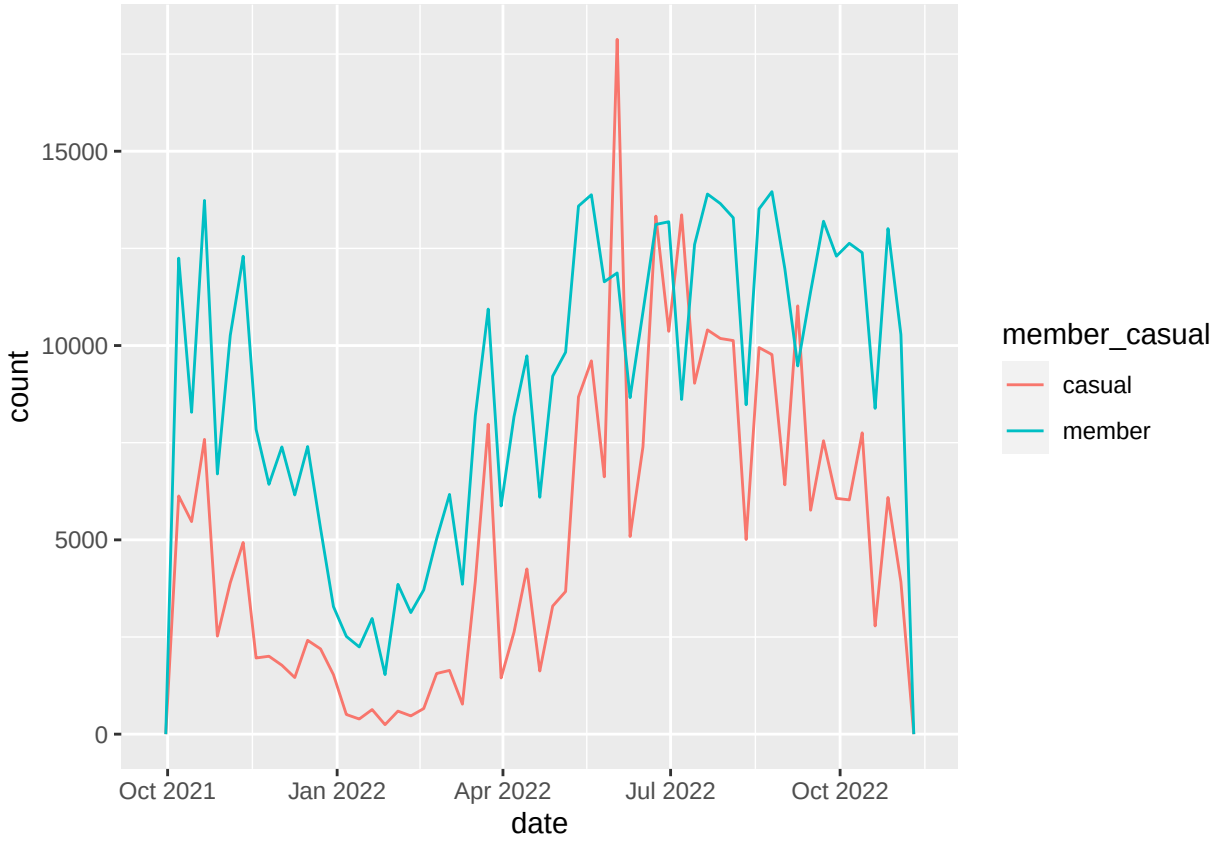

```
all_trips_v5 %>%
  filter(day_of_week == "Saturday" | day_of_week == "Sunday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```



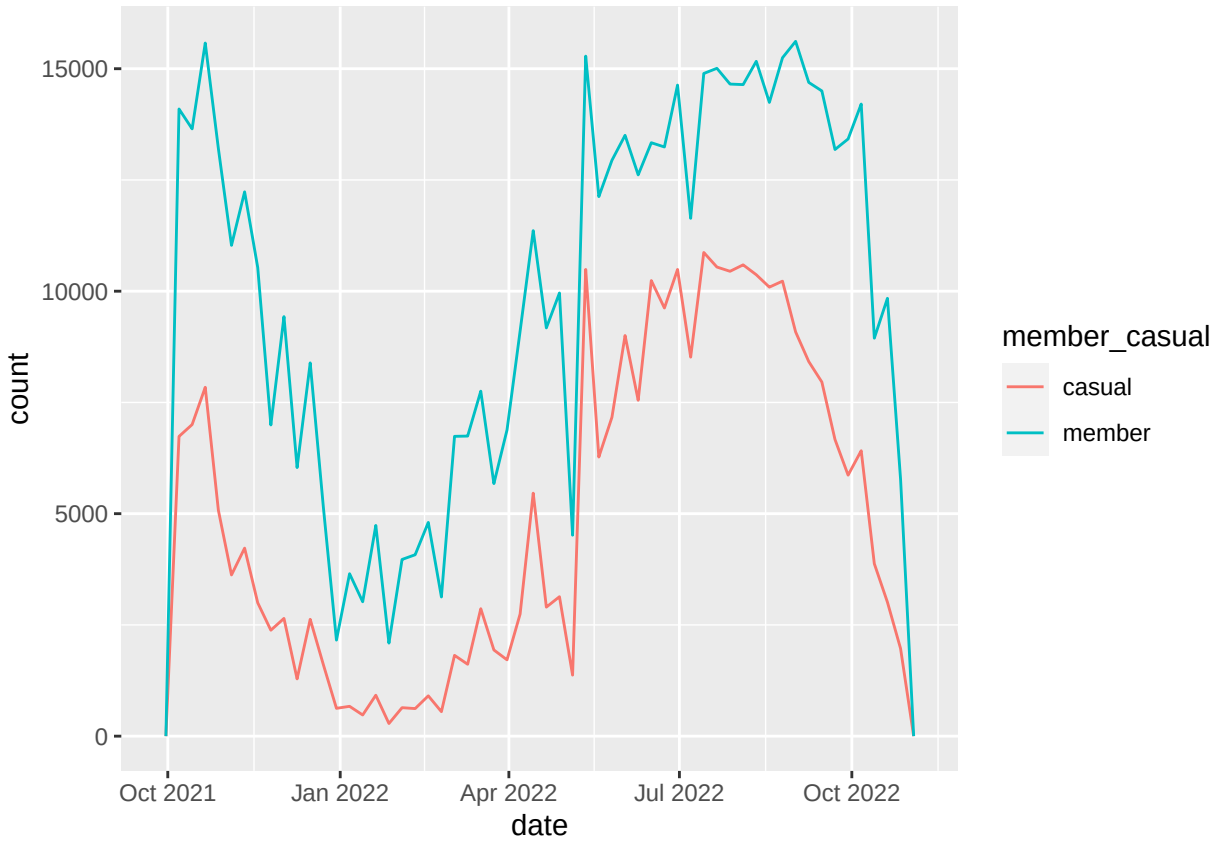

```
all_trips_v5 %>%
  filter(day_of_week != "Saturday" | day_of_week != "Sunday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```



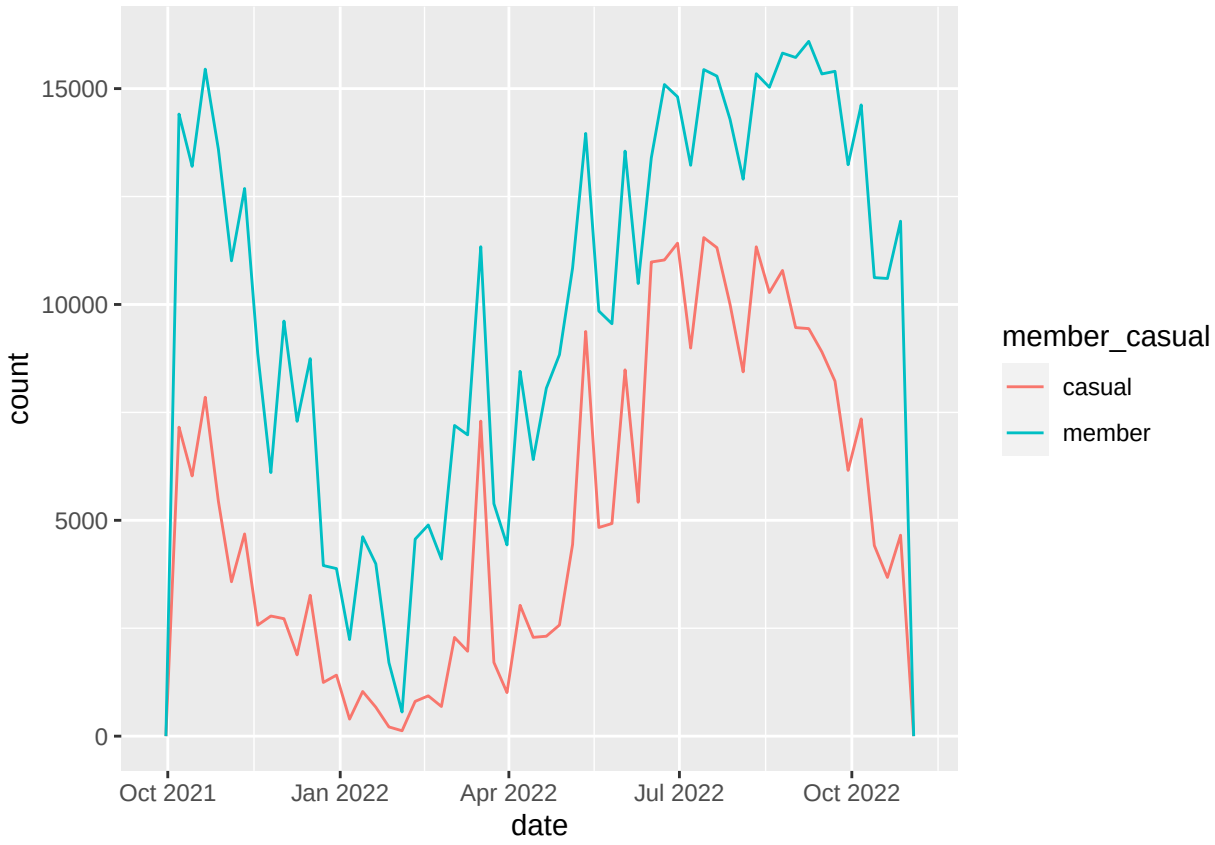

```
all_trips_v5 %>%
  filter(day_of_week == "Monday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```



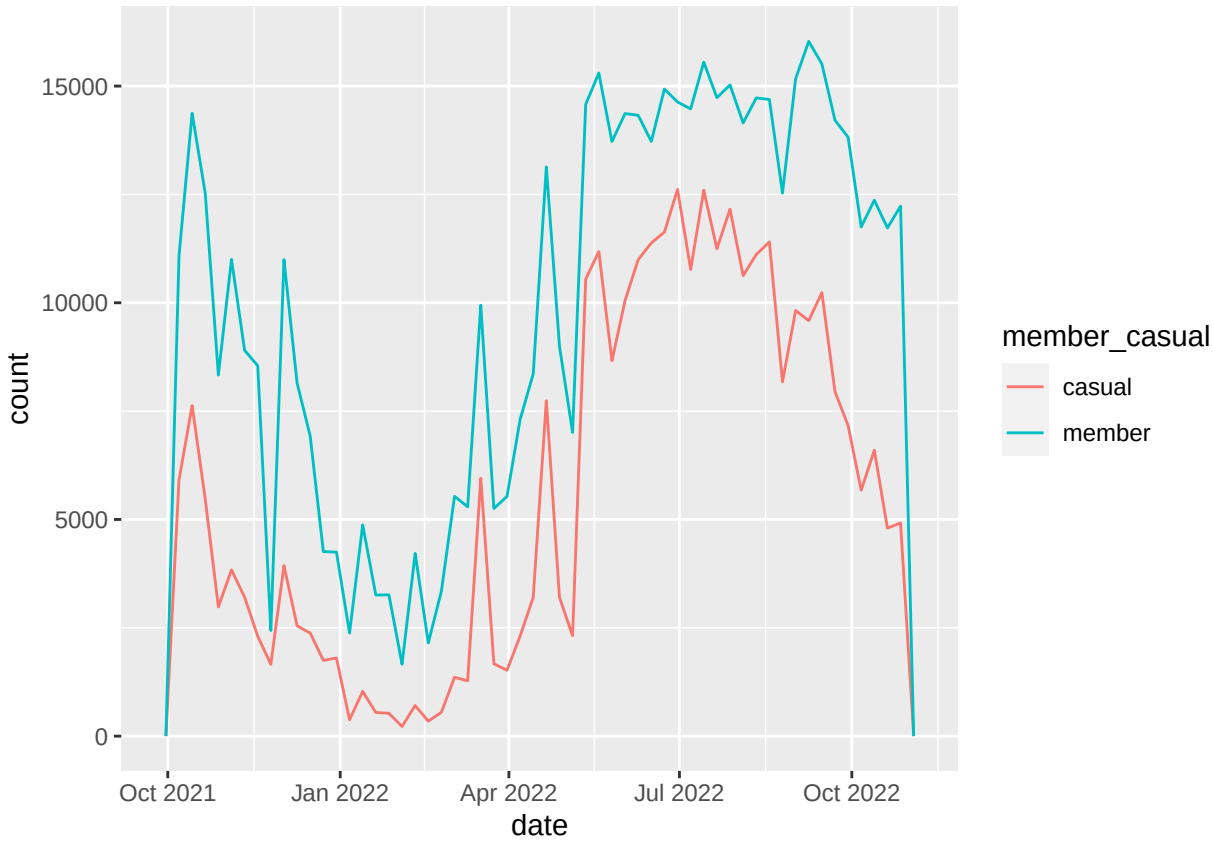

```
all_trips_v5 %>%
  filter(day_of_week == "Tuesday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```



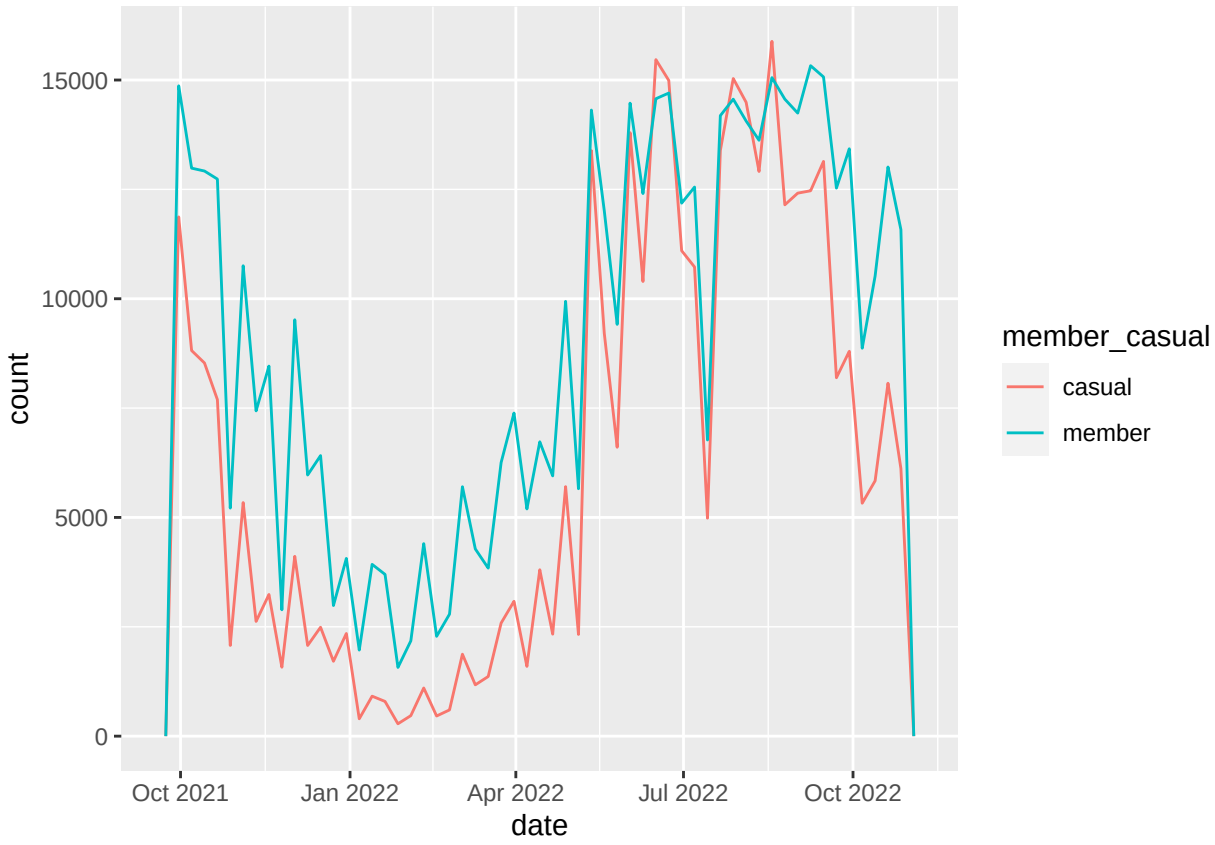

```
all_trips_v5 %>%
  filter(day_of_week == "Wednesday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```




```
all_trips_v5 %>%
  filter(day_of_week == "Thursday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```




```
all_trips_v5 %>%
  filter(day_of_week == "Friday") %>%
ggplot(aes(date, color = member_casual)) +
  geom_freqpoly(binwidth=7)
```




```
# 2700s > 45 min 10800s > 3 hours
```