Analysis of Cyclistic User Data

Michael-Philipp Stiebing

2023-03-21

Starting with processed dataframe all_trips_v5

- Conduct descriptive analysis as per script

```
# Set up two color palette that are compatible with all kinds of color vision,
# from http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/

cbPaletteMin <- c("#E69F00", "#0072B2")

# 2700s = 45 min 10800s = 3 hours

table(all_trips_v5$member_casual)
```

```
##
##  casual  member
## 2553025 3695371
```

```
summary(all_trips_v5$ride_length)
```

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

```
summary(all_trips_v5$geodist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   919.4  1608.6  2173.3  2820.9 42319.5
```

```
# 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
```

```r
# casual users seem to take longer rides


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",])
```

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

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

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

```r
# 10 percent of casual users take rides that are longer than 45 minutes, whereas only 1.7 percent of members do

nrow(filter(all_trips_v5,geodist <= 10)) / nrow(all_trips_v5)
```

```
## [1] 0.05118626
```

```r
nrow(all_trips_v5[all_trips_v5$member_casual == "casual" & all_trips_v5$geodist <= 10,]) /
  nrow(all_trips_v5[all_trips_v5$member_casual == "casual",])
```

```
## [1] 0.07263854
```

```r
nrow(all_trips_v5[all_trips_v5$member_casual == "member" & all_trips_v5$geodist <= 10,]) /
  nrow(all_trips_v5[all_trips_v5$member_casual == "member",])
```

```
## [1] 0.0363655
```

```r
# 7.3 percent of casual users 'take round trips', whereas only 3.6 percent of members do

# 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", "Friday", "Saturday"))

# 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
```

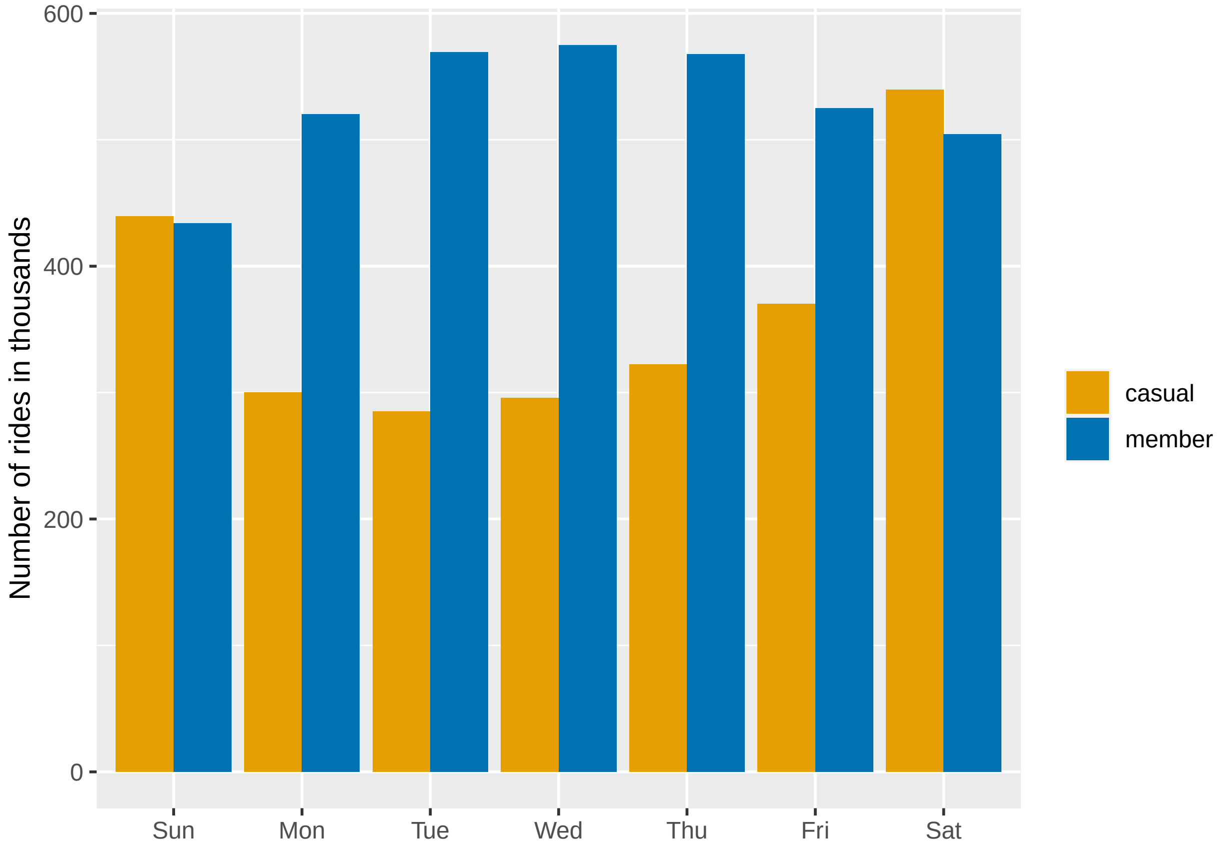
```
# analyze ridership data by type and weekday
all_trips_v5 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%    # creates weekday field using wday()
  group_by(member_casual, weekday) %>%                    # groups by usertype and weekday
  summarise(number_of_rides = n()                         # 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.
```

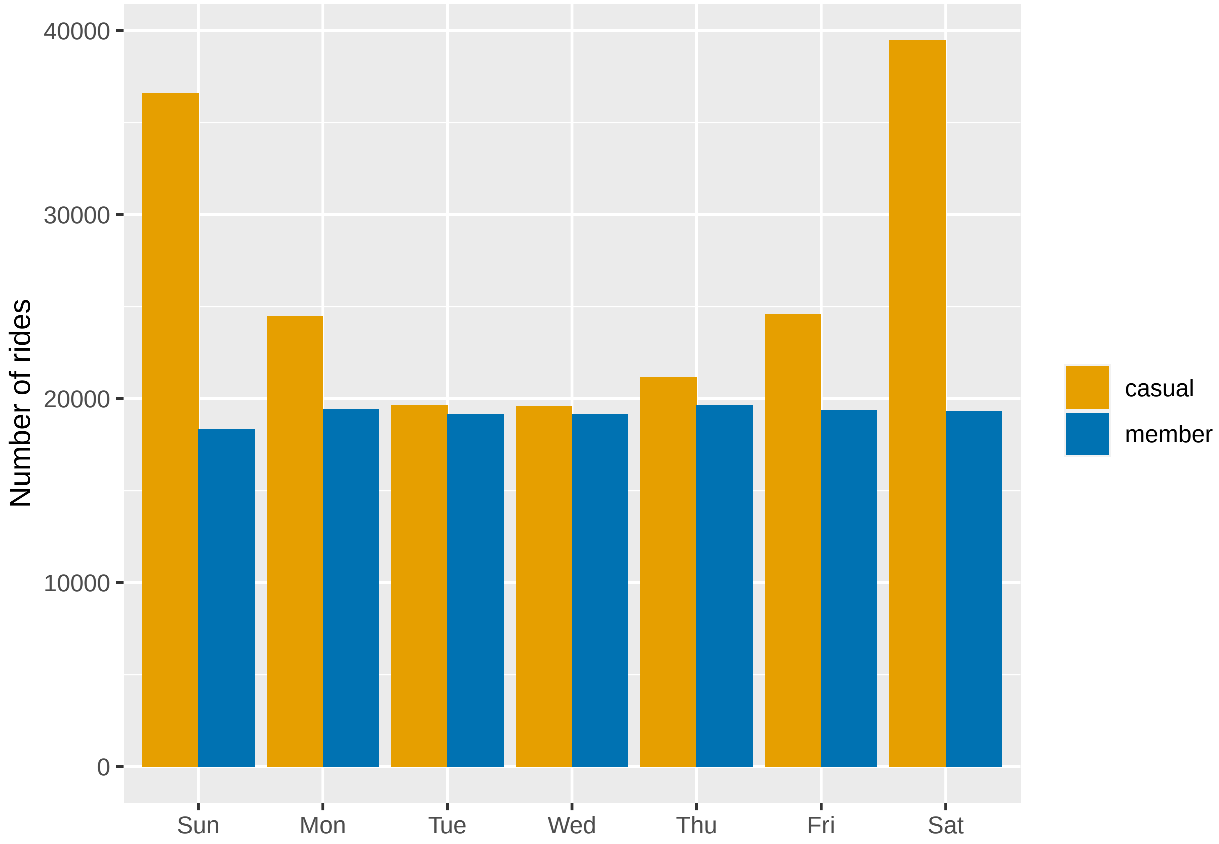
```
# Let's visualize the number of rides by rider type
all_trips_v5 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()/1000
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "Number of rides in thousands", x = NULL, fill = NULL)
```

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



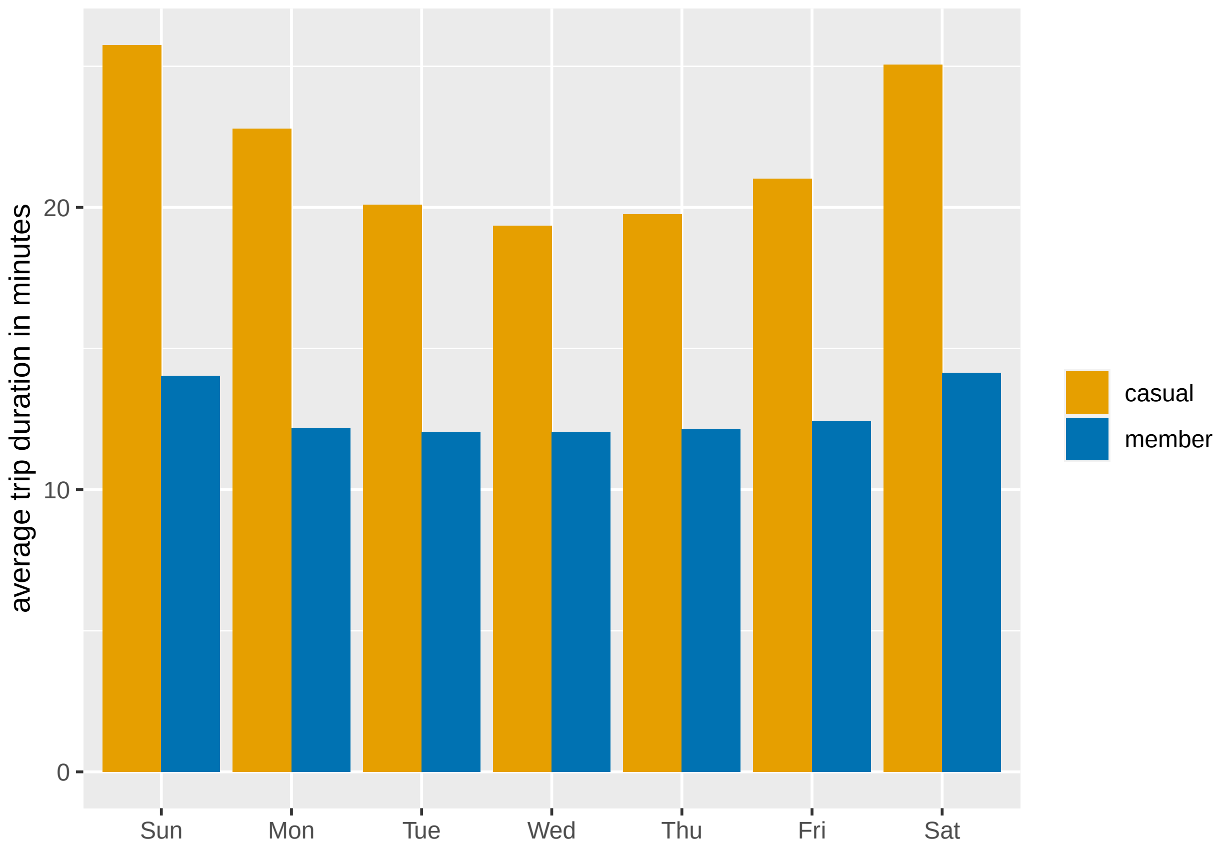
```
# ggsave("weeklyOverview.png",width=3840,height = 2160,unit="px")

# Let's visualize the number of rides by rider type for rides that begin and end within
# 10 meters of each other 'round trips'
filter(all_trips_v5,geodist <= 10) %>%
  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") +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "Number of rides", x = NULL, fill = NULL)
```

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



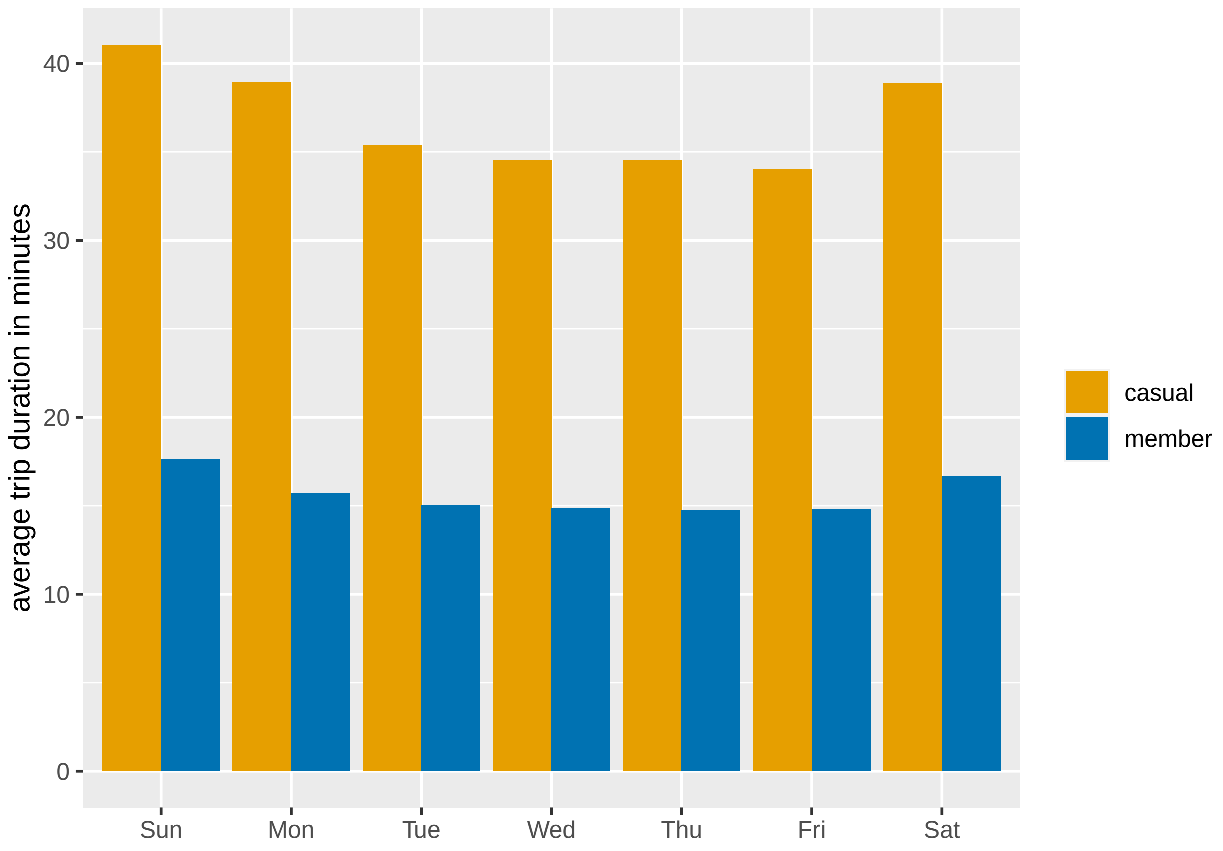
```
# ggsave("weeklyOverview-roundTrips.png",width=3840,height = 2160,unit="px")

# 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 / 60)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "average trip duration in minutes", x = NULL, fill = NULL)
```

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



```
# ggsave("weeklyOverview-duration.png",width=3840,height = 2160,unit="px")

# Let's create a visualization for average duration  for rides that begin and end within
# 10 meters of each other 'round trips'
filter(all_trips_v5,geodist <= 10) %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
          ,average_duration = mean(ride_length / 60)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "average trip duration in minutes", x = NULL, fill = NULL)
```

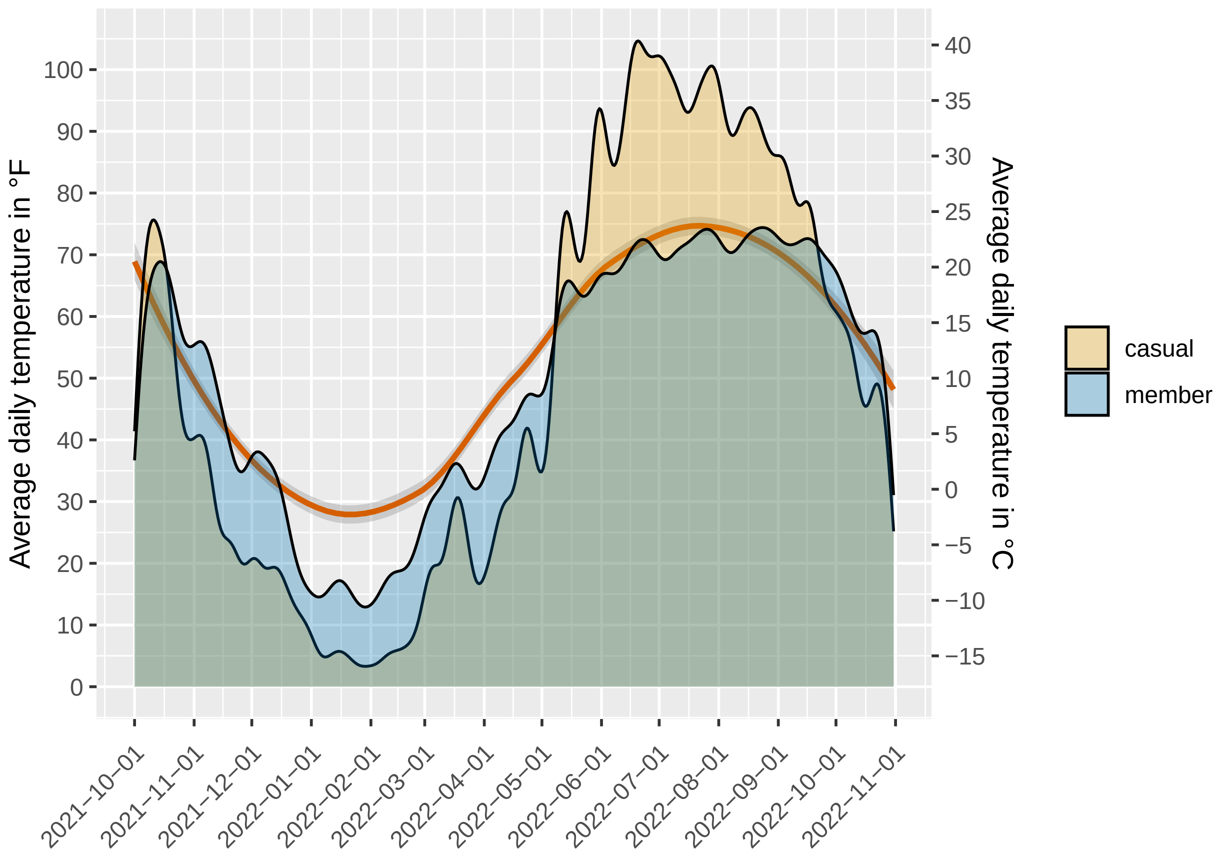
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



- The pattern emerging seems to be the following:
  - Members take shorter rides
  - Casual users take longer rides, especially on weekends
  - Casual users take more round trips than members

```
# Let's create a visualization for number of rides by User type, over the whole timeframe.
# Overlay average daily temperature in Fahrenheit. ( Has to be scaled to match the geom_density plot)
all_trips_v5 %>%
  ggplot() +
  geom_smooth(data = climate_chicago_202110_202210, aes(x = date, y = as.double(avgtemp) / 20000),method=loess,color="#D55E00") +
  geom_density( aes(x = date,fill = member_casual),alpha = .3) +
  scale_y_continuous(
    breaks=c(seq(0.000,0.005,by=0.0005)),
    labels=c(seq(0,100,10)),
    sec.axis = sec_axis(~ (((. * 20000) - 32) * 5/9), name = "Average daily temperature in °C",breaks=c(seq(-15,40,5)))
    ) +
  scale_x_date(date_breaks = "1 month") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.9, hjust=1)) +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "Average daily temperature in °F", x = NULL, fill = NULL)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

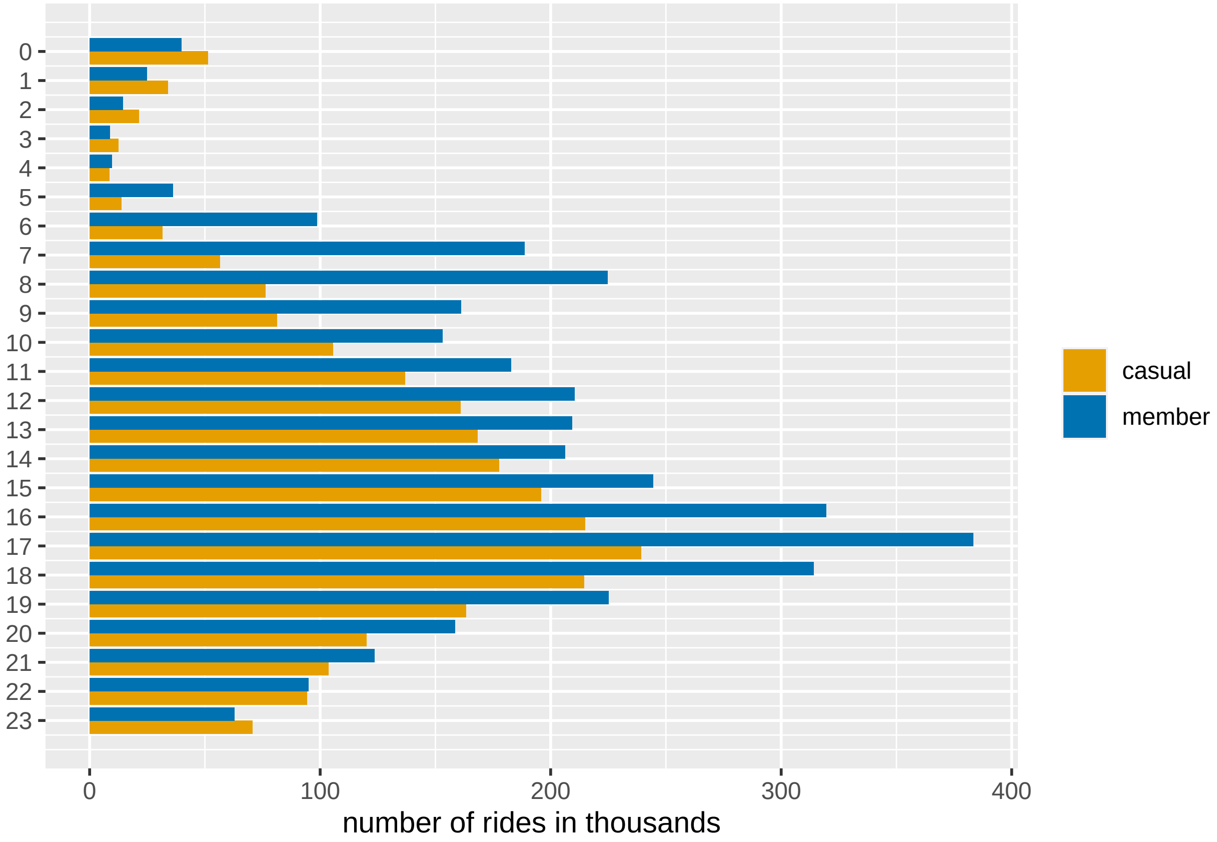


```
# ggsave("yearlyOverview.png",width=3840,height = 2160,unit="px")
```

- Looking at usage across the year, during the summer months, casual users overtake members
- Usage, especially by casual users, seems to correlate with average temperature

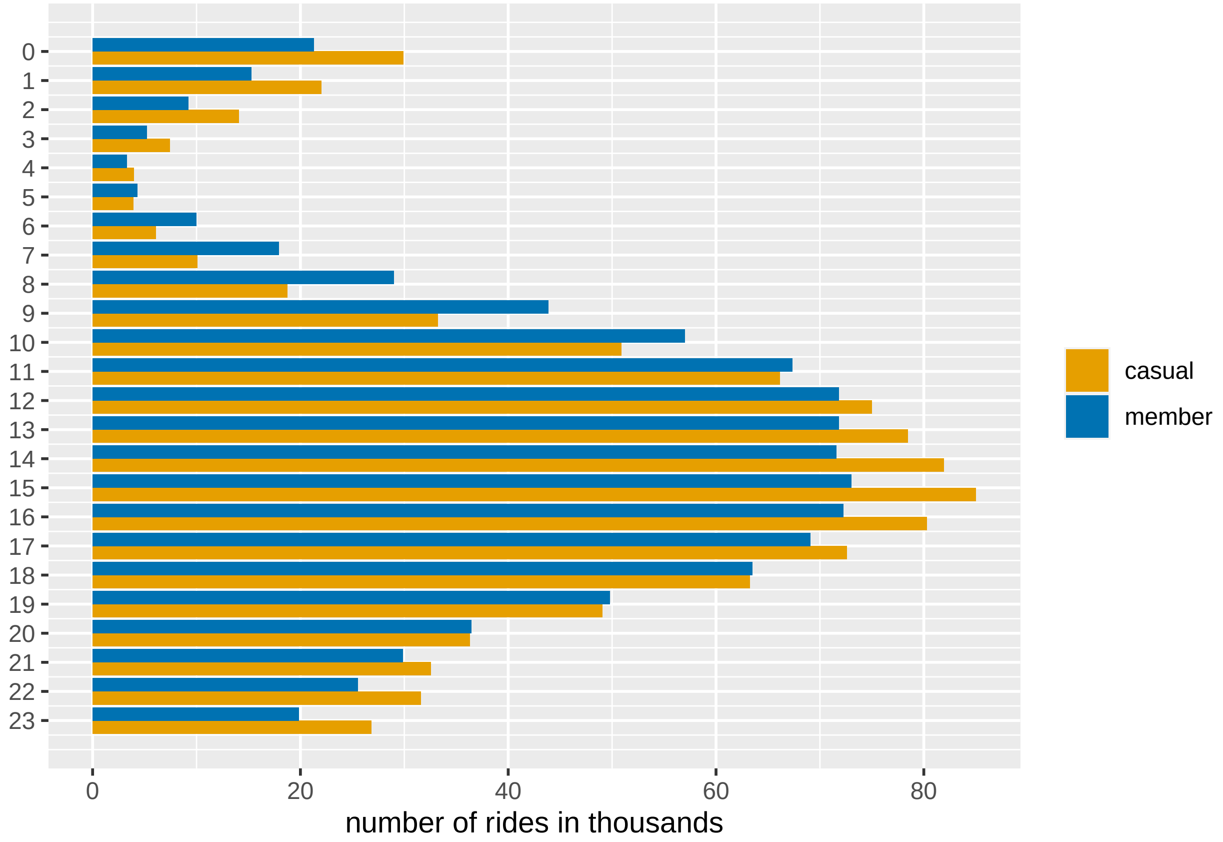
```
# Let's create a visualization for number of rides by User type, only looking at time of day
all_trips_v5 %>%
  group_by(member_casual, hours) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, as.numeric(hours))  %>%
  ggplot( aes(x = as.numeric(hours), y = number_of_rides / 1000, fill = member_casual)) +
  geom_col(position = "dodge") +
  coord_flip()  +
  scale_x_reverse(breaks = (0:23)) +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "number of rides in thousands", x = NULL, fill = NULL)
```

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



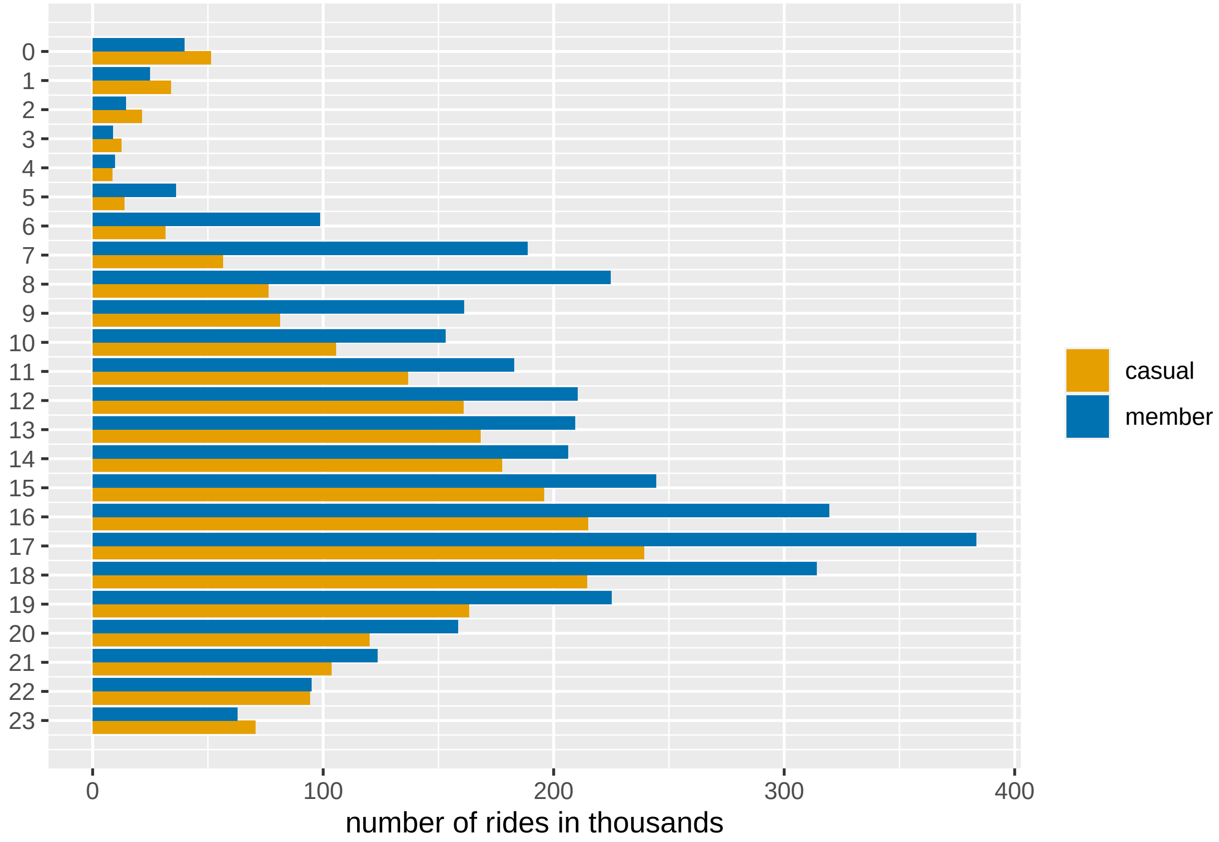
```
# ggsave("dailyOverview.png",width=3840,height = 2160,unit="px")

all_trips_v5 %>%
  filter(day_of_week == "Saturday" | day_of_week == "Sunday" ) %>%
  group_by(member_casual, hours) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, as.numeric(hours))  %>%
  ggplot( aes(x = as.numeric(hours), y = number_of_rides / 1000, fill = member_casual)) +
  geom_col(position = "dodge") +
  coord_flip()  +
  scale_x_reverse(breaks = (0:23)) +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "number of rides in thousands", x = NULL, fill = NULL)
```

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



```
all_trips_v5 %>%
  filter(day_of_week != "Saturday" | day_of_week != "Sunday" ) %>%
  group_by(member_casual, hours) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, as.numeric(hours))  %>%
  ggplot( aes(x = as.numeric(hours), y = number_of_rides / 1000, fill = member_casual)) +
  geom_col(position = "dodge") +
  coord_flip()  +
  scale_x_reverse(breaks = (0:23)) +
  scale_fill_manual(values=cbPaletteMin) +
  labs(y = "number of rides in thousands", x = NULL, fill = NULL)
```

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



- Looking at usage over hours of the day, members are clustered around morning and afternoon, preceding the start of a typical workday and following the end of it, which seems to support the theory that members mainly use the service to commute