Case Study: How Does a Bike-Share Navigate Speedy Success?

Luke Tozier

2022-11-07

Business Objective

Cyclistic is a *(fictional)* bike-share company based in Chicago. The marketing team believes that their future success depends on maximizing the number of annual memberships. Pursuant to this goal they would like the following question answered

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

To answer this, we're going to look at the last 12 months of trip data - 12 csv files split per month.

Installing and Loading Libraries

tidyverse/readr/dplyr/janitor for data import and manipulation, lubridate for date functions, and ggplot2 for visualization

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(readr)
library(lubridate)
library(janitor)
```

Collect, wrangle, and combine the data

Compare column names for each of the files to make sure they match before merging

```
# Make a variable list of all csv dataframes
csv.list <- list.files(path='C:/Users/Luke/Desktop/Cycle/data') %>%
  lapply(read_csv)
```

```
# Check if any columns don't match using this janitor function
compare_df_cols(csv.list, return = c("mismatch"))
```

```
##  [1] column_name csv.list_1  csv.list_2  csv.list_3  csv.list_4  csv.list_5
##  [7] csv.list_6  csv.list_7  csv.list_8  csv.list_9  csv.list_10 csv.list_11
## [13] csv.list_12
## <0 rows> (or 0-length row.names)
```

Combine into a single dataframe

```
all_trips <- list.files(path='C:/Users/Luke/Desktop/Cycle/data', full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
```

Clean, Organize and Prep Data

Inspect the new table that has been created

```
# Structure of the dataframe
str(all_trips)
```

```
## spc_tbl_ [5,828,235 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5828235] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "3629
##  $ rideable_type     : chr [1:5828235] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:5828235], format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct[1:5828235], format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr [1:5828235] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:5828235] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:5828235] NA NA NA NA ...
##  $ end_station_id    : chr [1:5828235] NA NA NA NA ...
##  $ start_lat         : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5828235] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Compare member riders vs casual
table(all_trips$member_casual)
```

```
##
```

```
##  casual  member
## 2401286 3426949
```

```
# Compare rideable options
table(all_trips$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##       2740516        192475       2895244
```

```
# Check trips that ended before starting
sum(all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at) < 0 )
```

```
## [1] 108
```

```
# Check for duplicate Ride IDs
sum(duplicated(all_trips$ride_id) > 0 )
```

```
## [1] 0
```

Problems to address:

(1) The data can only be aggregated at the ride-level. It needs some additional columns of data such as day, month, and year that provide additional opportunities for aggregation
(2) There is no trip duration column.
(3) There are 108 rides where the ended_at time is before the started_at time resulting in a negative trip duration.

```
# Add columns that list the date, month, day, and year of each ride. Also add a "ride_length" calculati
all_trips$date <- as.Date(all_trips$started_at) #The default format is yyyy-mm-dd
all_trips$month <- format(as.Date(all_trips$date), "%m")
all_trips$day <- format(as.Date(all_trips$date), "%d")
all_trips$year <- format(as.Date(all_trips$date), "%Y")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at)
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
all_trips$starttime <- format(as.POSIXct(all_trips$started_at), format = "%H:%M")
```

Inspect the structure of the newly added columns

```
str(all_trips)
```

```
## spc_tbl_ [5,828,235 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5828235] "620BC6107255BF4C" "4471C70731AB2E45" "26CA69D43D15EE14" "3629
##  $ rideable_type     : chr [1:5828235] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:5828235], format: "2021-10-22 12:46:42" "2021-10-21 09:12:37" ...
##  $ ended_at          : POSIXct[1:5828235], format: "2021-10-22 12:49:50" "2021-10-21 09:14:14" ...
##  $ start_station_name: chr [1:5828235] "Kingsbury St & Kinzie St" NA NA NA ...
##  $ start_station_id  : chr [1:5828235] "KA1503000043" NA NA NA ...
##  $ end_station_name  : chr [1:5828235] NA NA NA NA ...
##  $ end_station_id    : chr [1:5828235] NA NA NA NA ...
##  $ start_lat         : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5828235] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5828235] -87.6 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5828235] "member" "member" "member" "member" ...
##  $ ride_length       : num [1:5828235] 188 97 467 75 496 861 161 501 448 509 ...
##  $ date              : Date[1:5828235], format: "2021-10-22" "2021-10-21" ...
##  $ month             : chr [1:5828235] "10" "10" "10" "10" ...
##  $ day               : chr [1:5828235] "22" "21" "16" "16" ...
##  $ year              : chr [1:5828235] "2021" "2021" "2021" "2021" ...
##  $ day_of_week       : chr [1:5828235] "Friday" "Thursday" "Saturday" "Saturday" ...
##  $ starttime         : chr [1:5828235] "12:46" "09:12" "16:28" "16:17" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Remove bad data

The dataframe includes a few hundred entries where ride_length was a negative or 0 second duration

```r
# Builds a new dataframe without the invalid rows
all_trips_v2 <- all_trips[!(all_trips$ride_length<1),]
print(paste("Removed", nrow(all_trips) - nrow(all_trips_v2), "invalid trips"))
```

```
## [1] "Removed 571 invalid trips"
```

Descriptive Analysis

Summarize ride length

```r
summary(all_trips_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       1     356     629    1176    1131 2442301 
```

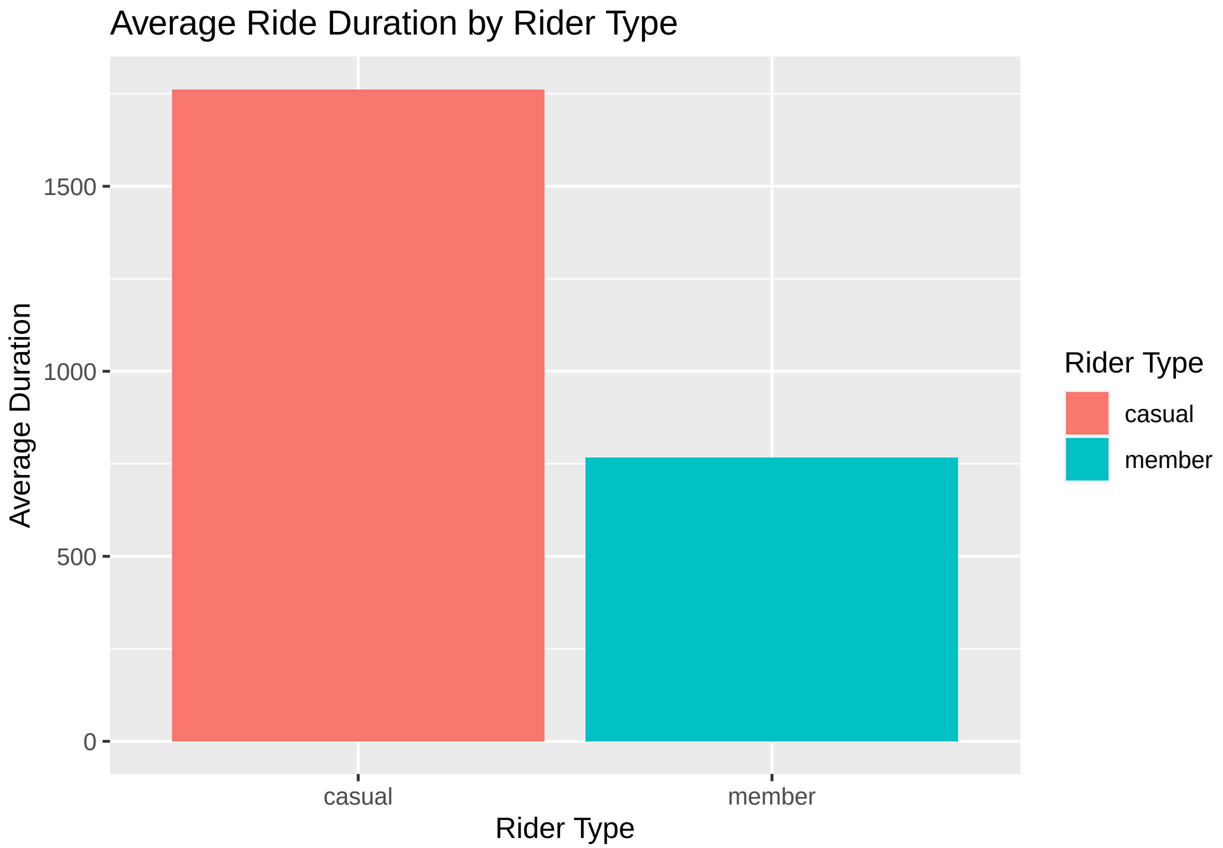
Compare member and casual users

Mean, median, min, and max ride length per rider type

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                1761.8174
## 2                     member                 766.1685
```

```
#Bar plot to visualize average ride duration
all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(number_of_rides = n(),average_duration = mean(ride_length)) %>%
  ggplot(aes(x = member_casual, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Average Ride Duration by Rider Type", x="Rider Type", y="Average Duration")+
  scale_fill_discrete(name = "Rider Type")
```



```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                      807
## 2                     member                      533
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                  2442301
## 2                     member                    93594
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                        1
## 2                     member                        1
```

Key Takeaway: *Casual riders have a much higher average trip time then members, especially on weekends*

Analyze the number of rides by rider type and day of the week

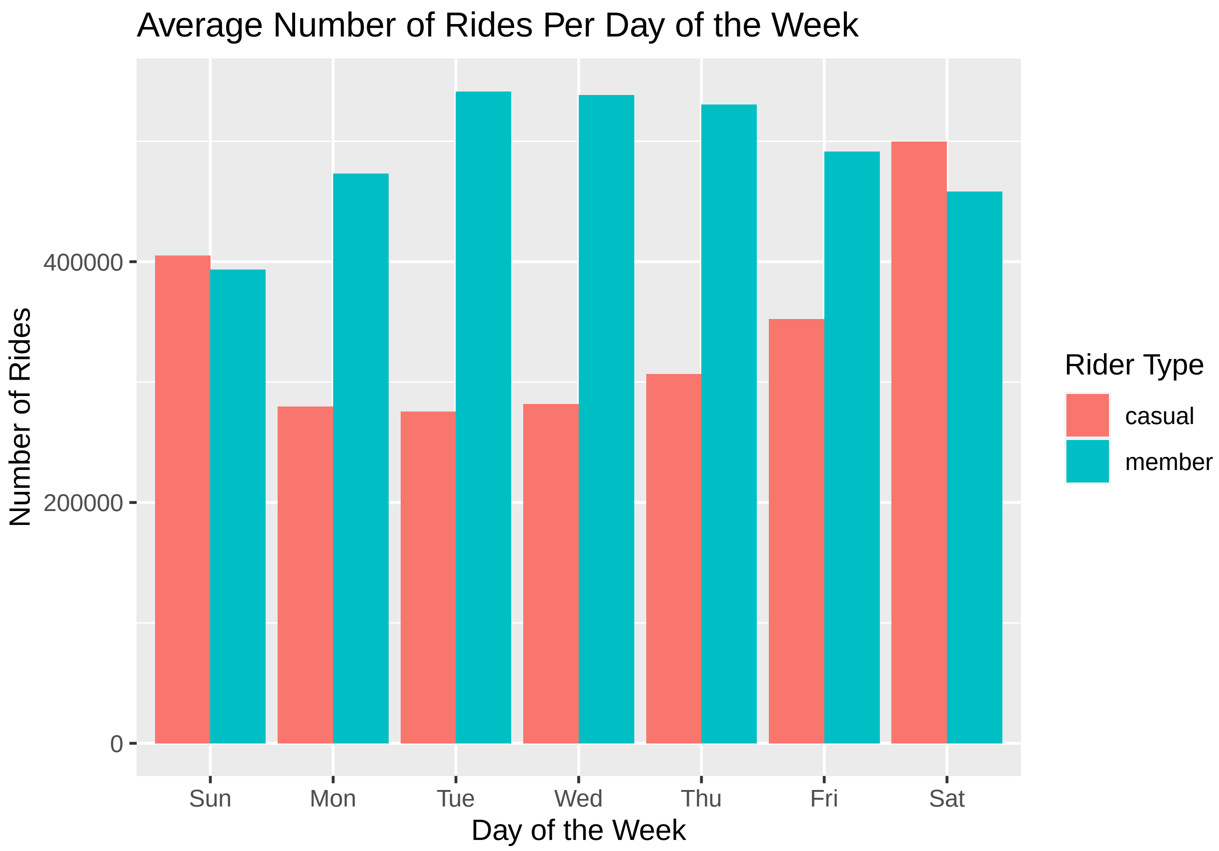
```
#Total rides by rider type
all_trips_v2 %>%
  group_by(member_casual) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual)
```

```
## # A tibble: 2 x 2
##   member_casual number_of_rides
##   <chr>                   <int>
## 1 casual                2400991
## 2 member                3426673
```

```
#Rides and avg duration per rider type and day of the week
all_trips_v2 %>%
  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
  average_duration = mean(ride_length)) %>%         # calculates the average duration
  arrange(member_casual, weekday)                                # sorts
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              404977            2062.
##  2 casual        Mon              279762            1784.
##  3 casual        Tue              275745            1549.
##  4 casual        Wed              281640            1502.
##  5 casual        Thu              306662            1541.
##  6 casual        Fri              352466            1681.
##  7 casual        Sat              499739            1963.
##  8 member        Sun              393568             853.
##  9 member        Mon              473027             740.
## 10 member         Tue                 541484              730.
## 11 member         Wed                 538459              727.
## 12 member         Thu                 530510              738.
## 13 member         Fri                 491436              752.
## 14 member         Sat                 458189              856.
```

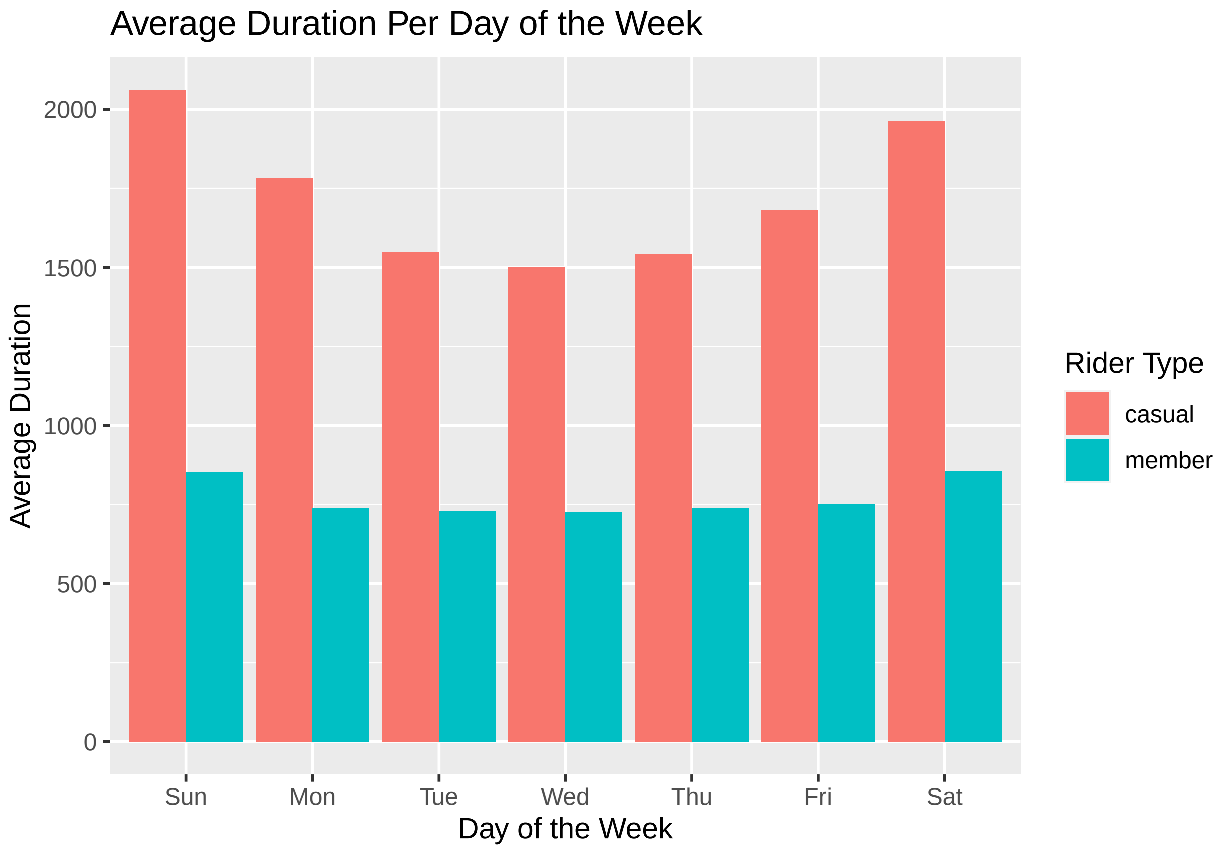
```
#Plot the number of rides by rider type and day of the week
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(title="Average Number of Rides Per Day of the Week",x="Day of the Week",y="Number of Rides")+
  scale_fill_discrete(name = "Rider Type")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



Key Takeaway: *Members trend much higher on number of trips midweek, where casual riders take a small lead over members on weekends*

Visualize the average trip duration per day of week

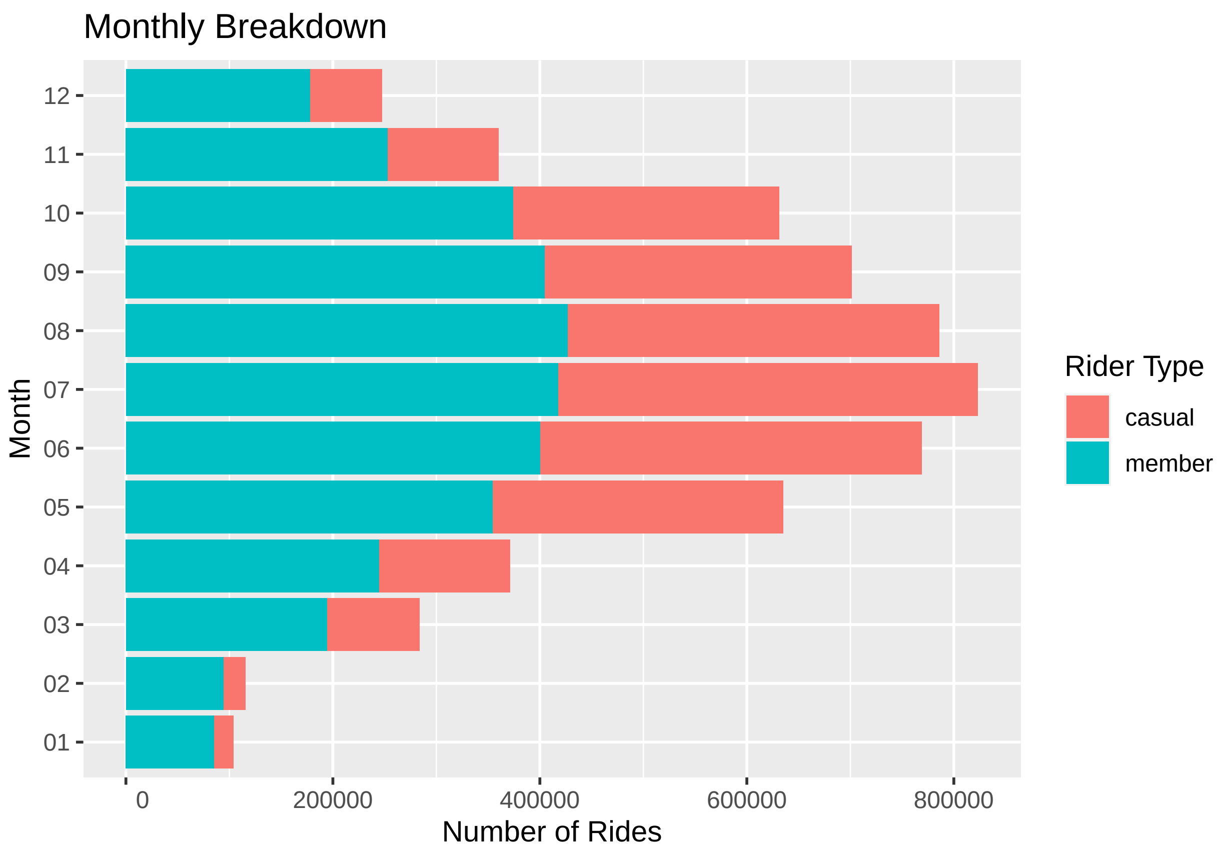
```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>% #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n(),      #calculates the number of rides and average duration
  average_duration = mean(ride_length)) %>%     # calculates the average duration
  arrange(member_casual, weekday)  %>%    # sorts
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual))+ #Plotting from here down
  geom_col(position = "dodge")+
  labs(title="Average Duration Per Day of the Week",x="Day of the Week",y="Average Duration")+
  scale_fill_discrete(name = "Rider Type")
```



Key Takeaway: *Casual Riders trend much higher on the average trip duration, especially on weekends*

Monthly breakdown of rides per user type

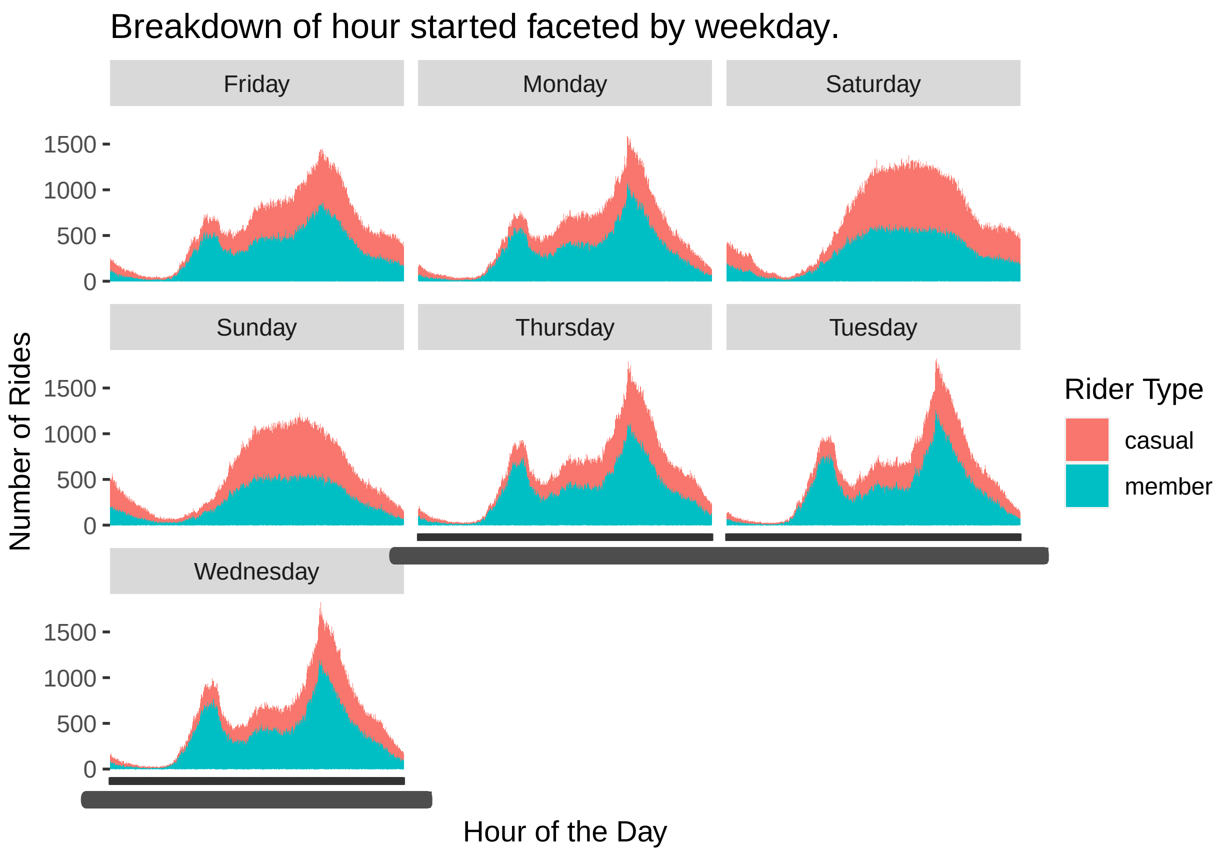
```
all_trips_v2 %>%
  ggplot(aes(x=month, fill=member_casual)) +
  geom_bar() +
  coord_flip()+
  labs(x="Month", y="Number of Rides",title="Monthly Breakdown")+
  scale_fill_discrete(name = "Rider Type")+
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



Key Takeaway: *The warmer months clearly have the highest amount of trips for obvious reasons. The colder months show a much higher ratio of member trips to casuals however, with that ratio evening out into the warmer months*

Distribution by hour of day facetted by weekday

```
all_trips_v2 %>%
  ggplot(aes(starttime, fill=member_casual)) +
  geom_bar() +
  labs(x="Hour of the Day",y="Number of Rides", title="Breakdown of hour started faceted by weekday.")
  facet_wrap(~ day_of_week)+
  scale_fill_discrete(name = "Rider Type")
```



Key Takeaway: *Traditional weekday commute times show a large amount of member trips, but also a hike in casual trips too. It looks like a decent amount of casual trips are to commute to and from work. Weekend afternoons show a big boost in casual trips, trumping member trips*

Overall Summary of analysis

- Casual riders take up about 70% of the total trip lengths on average.
- Member trips make up about 59% of the total trips.
- Members take the most trips on weekdays, Casuals on weekends.
- Most trips are within the warmer months, with mostly members taking trips during the winter.
- Both rider types trend high on weekday commute times.

My top three recommendations based on this analysis

- Incentivize casual riders to sign up for membership based on some type of ride length benefit. Perhaps a rewards program that keeps track of total trip time and offers some type of bonus based on time accrued.
- A marketing campaign to advertise to casual riders how much they could save on their daily work commute.
- Offer some type of discount to members that only take weekend trips to help convert casual riders that are only looking for weekend fun on their time off.

Exporting a summary file for further analysis and visualization

This summary file will be useful to create visualizations in any other software (Tableau, Excel, ect.) to be used in the final presentation.

```
counts <- all_trips_v2 %>%
  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
  ,average_duration = mean(ride_length)) %>%        # calculates the average duration
  arrange(member_casual, weekday)                               # sorts

write.csv(counts, file = "C:\\Users\\Luke\\Desktop\\Cycle Capstone\\summary.csv")
```