Cyclistic

```
installed.packages("tidyverse")
install.packages("tidyverse")
library("tidyverse")
install.packages("lubridate")
bike_share_202107 <- read.csv('202107-divvy-tripdata.csv')
library("lubridate")
install.packages("ggplot2")
library("ggplot2")
install.packages("dyplr")
library("dyplyr")
```

Cleaned and renamed .CSV files locally. trip_data_jan, trip_data_feb, trip_data_mar, , trip_data_april, trip_data_may, trip_data_jun, trip_data_july, trip_data_aug, trip_data_sept, trip_data_oct, trip_data_nov, trip_data_dec.

Merged into one data set named: trip_data:

```
trip_data <- bind_rows(trip_data_jan, trip_data_feb, trip_data_mar, trip_data_april,
trip_data_may, trip_data_jun, trip_data_july, trip_data_aug, trip_data_sept, trip_data_oct,
trip_data_nov, trip_data_dec)
```

Removed Duplicates:

```
> trip_data_no_dups <- trip_data[!duplicated(trip_data$ride_id), ]
> print(paste("Removed", nrow(trip_data) - nrow(trip_data_no_dups), "duplicated rows"))
[1] "Removed 209 duplicated rows"
```

Parsing Datetime Columns:

```
> trip_data_no_dups$started_at <- as.POSIXct(trip_data_no_dups$started_at, "%Y-%m-%d %H:%M:%S")
> trip_data_no_dups$ended_at <- as.POSIXct(trip_data_no_dups$ended_at, "%Y-%m-%d %H:%M:%S")
```

Create new column for ride length in minutes:

```
> trip_data_no_dups <- trip_data_no_dups %>%
+     mutate(ride_time_m = as.numeric(trip_data_no_dups$ended_at -
trip_data_no_dups$started_at) / 60)
> summary(trip_data_no_dups$ride_time_m)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
-29049.97      7.32     13.17     23.46     24.08  55944.15 
*notice outliers in min and max
```

Combine year and month into one column:

```
> trip_data_no_dups <- trip_data_no_dups %>%
+     mutate(year_month = paste(strftime(trip_data_no_dups$started_at, "%Y"),
+                              "-",
+                              strftime(trip_data_no_dups$started_at, "%m"),
+                              paste("(",strftime(trip_data_no_dups$started_at, "%b"), ")", sep="")))
> unique(trip_data_no_dups$year_month)
 [1] "2020 - 08 (Aug)" "2020 - 07 (Jul)" "2020 - 09 (Sep)" "2020 - 10 (Oct)" "2020 - 11 (Nov)"
"2020 - 12 (Dec)" "2021 - 01 (Jan)"
 [8] "2021 - 02 (Feb)" "2021 - 03 (Mar)" "2021 - 04 (Apr)" "2021 - 05 (May)" "2021 - 06 (Jun)"
"2021 - 07 (Jul)"
```

Create weekday column:

```
> trip_data_no_dups <- trip_data_no_dups %>%
+     mutate(weekday = paste(strftime(trip_data_no_dups$ended_at, "%u"), "-",
strftime(trip_data_no_dups$ended_at, "%a")))
> unique(trip_data_no_dups$weekday
+ )
[1] "4 - Thu" "3 - Wed" "2 - Tue" "1 - Mon" "5 - Fri" "7 - Sun" "6 - Sat"
```

Create start hour column:

```
> trip_data_no_dups <- trip_data_no_dups %>%
+     mutate(start_hour = strftime(trip_data_no_dups$ended_at, "%H"))
> unique(trip_data_no_dups$start_hour)
 [1] "14" "15" "17" "08" "12" "16" "18" "11" "04" "05" "19" "13" "09" "20" "21" "01" "10" "07" "06"
"22" "03" "00" "02" "23"
```

Create new cleaned .csv file:

```
> trip_data_no_dups %>%
+ write.csv("trip_data_clean.csv")
```

Analyze:

```
> trip_data <- trip_data_no_dups
> head(trip_data)

    ride_id rideable_type        started_at          ended_at          start_station_name
start_station_id
1 322BD23D287743ED   docked_bike 2020-08-20 18:08:14 2020-08-20 18:17:51 Lake Shore
Dr & Diversey Pkwy             329
2 2A3AEF1AB9054D8B electric_bike 2020-08-27 18:46:04 2020-08-27 19:54:51       Michigan
Ave & 14th St             168
3 67DC1D133E8B5816 electric_bike 2020-08-26 19:44:14 2020-08-26 21:53:07     Columbus
Dr & Randolph St             195
4 C79FBBD412E578A7 electric_bike 2020-08-27 12:05:41 2020-08-27 12:53:45          Daley
Center Plaza              81
5 13814D3D661ECADB electric_bike 2020-08-27 16:49:02 2020-08-27 16:59:49      Leavitt St &
Division St             658
6 56349A5A42F0AE51 electric_bike 2020-08-27 17:26:23 2020-08-27 18:07:50      Leavitt St &
Division St             658
        end_station_name end_station_id start_lat start_lng  end_lat   end_lng member_casual
date month day year day_of_week
1   Clark St & Lincoln Ave           141  41.93259 -87.63643 41.91569 -87.63460        member
2020-08-20    08  20 2020    Thursday
2   Michigan Ave & 14th St           168  41.86438 -87.62368 41.86422 -87.62344        casual
2020-08-27    08  27 2020    Thursday
3   State St & Randolph St            44  41.88464 -87.61955 41.88497 -87.62757        casual
2020-08-26    08  26 2020   Wednesday
4     State St & Kinzie St            47  41.88409 -87.62964 41.88958 -87.62754        casual
2020-08-27    08  27 2020    Thursday
5 Leavitt St & Division St           658  41.90299 -87.68377 41.90300 -87.68384        casual
2020-08-27    08  27 2020    Thursday
6 Leavitt St & Division St           658  41.90302 -87.68373 41.90309 -87.68363        casual
2020-08-27    08  27 2020    Thursday
  ride_length ride_time_m      year_month weekday start_hour
1    577 secs    9.616667 2020 - 08 (Aug) 4 - Thu         14
2   4127 secs   68.783333 2020 - 08 (Aug) 4 - Thu         15
3   7733 secs  128.883333 2020 - 08 (Aug) 3 - Wed         17
4   2884 secs   48.066667 2020 - 08 (Aug) 4 - Thu         08
```

```
5    647 secs   10.783333 2020 - 08 (Aug) 4 - Thu         12
6   2487 secs   41.450000 2020 - 08 (Aug) 4 - Thu         14
```

Create Summary of Data:

```
> summary(trip_data)
   ride_id             rideable_type       started_at                  ended_at                  
start_station_name
 Length:4730872     classic_bike :1785514   Min.   :2020-08-01 00:00:01   Min.   :2020-08-01 
00:04:41   Length:4730872   
 Class :character   docked_bike  :1558141   1st Qu.:2020-10-03 08:45:45   1st Qu.:2020-10-03 
09:08:11   Class :character  
 Mode  :character   electric_bike:1387217   Median :2021-04-05 13:52:15   Median :2021-04-05 
14:16:31   Mode  :character  
                                            Mean   :2021-02-17 10:26:13   Mean   :2021-02-17 10:49:41                     
                                            3rd Qu.:2021-06-15 05:56:21   3rd Qu.:2021-06-15 06:21:59                     
                                            Max.   :2021-07-31 23:59:58   Max.   :2021-08-12 17:45:41                     

 start_station_id   end_station_name   end_station_id      start_lat       start_lng         end_lat     
end_lng      
 Length:4730872     Length:4730872     Length:4730872     Min.   :41.64   Min.   :-87.87   Min.   
:41.51   Min.   :-88.07  
 Class :character   Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66   1st 
Qu.:41.88   1st Qu.:-87.66  
 Mode  :character   Mode  :character   Mode  :character   Median :41.90   Median :-87.64   
Median :41.90   Median :-87.64  
                                                          Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   
:-87.64  
                                                          3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93   3rd 
Qu.:-87.63  
                                                          Max.   :42.08   Max.   :-87.52   Max.   :42.16   Max.   :-87.44  
                                                                                           NA's   :5246    NA's   :5246    
 member_casual        date               month               day                year             day_of_week    
ride_length      
 casual:2102054   Min.   :2020-08-01   Length:4730872     Length:4730872     Length:4730872     
Sunday   :723892   Length:4730872   
 member:2628818   1st Qu.:2020-10-03   Class :character   Class :character   Class :character   
Monday   :579158   Class :difftime  
                  Median :2021-04-05   Mode  :character   Mode  :character   Mode  :character   
Tuesday  :604193   Mode  :numeric   
                  Mean   :2021-02-16                                                            Wednesday:631066                    
                  3rd Qu.:2021-06-15                                                            Thursday :612640                    
                  Max.   :2021-07-31                                                            Friday   :692953                    
```

```
                                                          Saturday :886970
  ride_time_m         year_month          weekday           start_hour
 Min.   :-29049.97   Length:4730872     Length:4730872     Length:4730872
 1st Qu.:     7.32   Class :character   Class :character   Class :character
 Median :    13.17   Mode  :character   Mode  :character   Mode  :character
 Mean   :    23.46
 3rd Qu.:    24.08
 Max.   : 55944.15
```

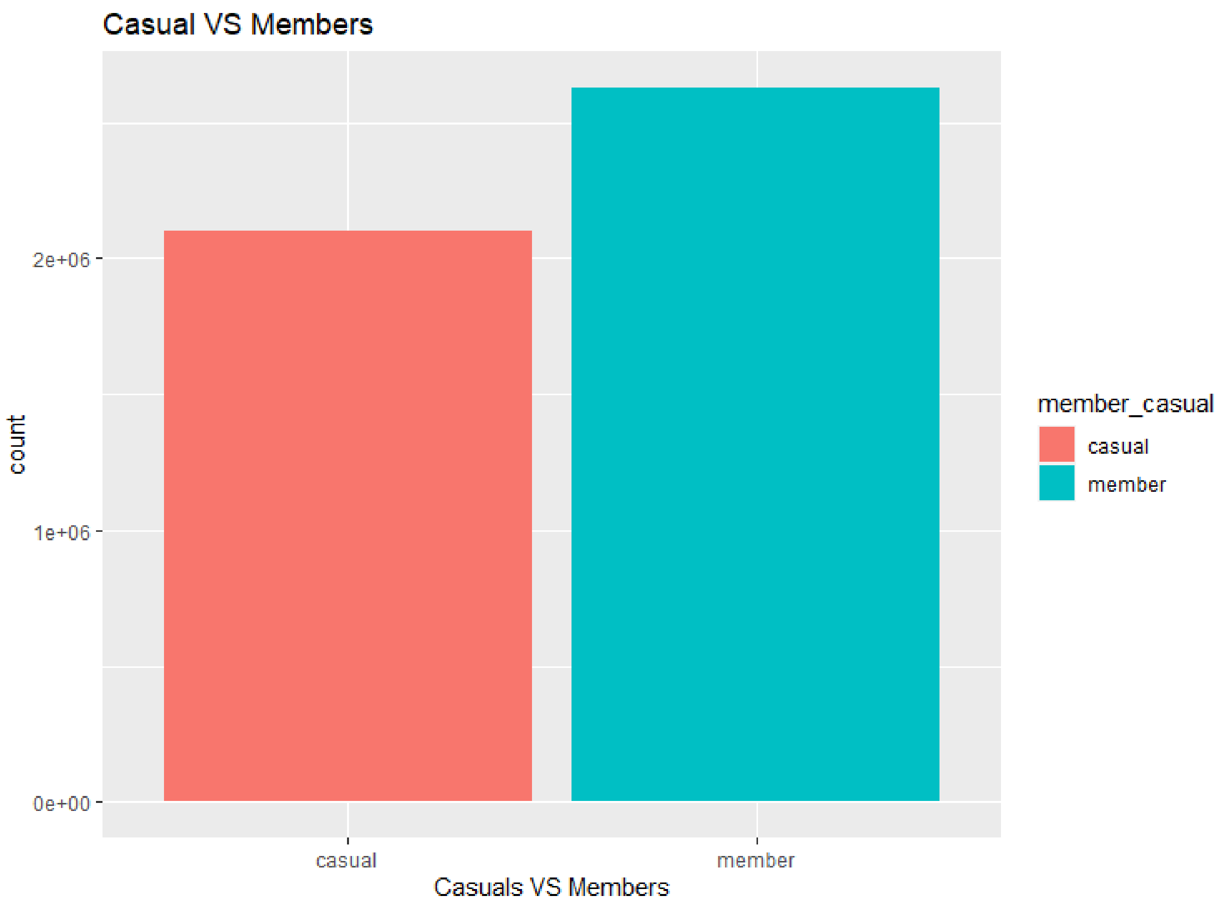
Visualize:

Casual vs Member Usage:

```
> trip_data %>%
+     group_by(member_casual) %>%
+     summarise(count = length(ride_id),
+               '%' = (length(ride_id) / nrow(trip_data)) * 100)
# A tibble: 2 x 3
  member_casual   count   `%`
  <fct>           <int> <dbl>
1 casual        2102054  44.4
2 member        2628818  55.6
```

Count of Rides Taken per Type of User:

```
> fig(16,8)
> ggplot(trip_data, aes(member_casual, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Casuals x Members", title="Chart 01 - Casuals x Members distribution")
> fig(16,8)
> ggplot(trip_data, aes(member_casual, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Casuals x Members", title="Casuals VS Members")
> fig(16,8)
> ggplot(trip_data, aes(member_casual, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Casuals VS Members", title="Casual VS Members")
> #Members account for 59% of rides
```

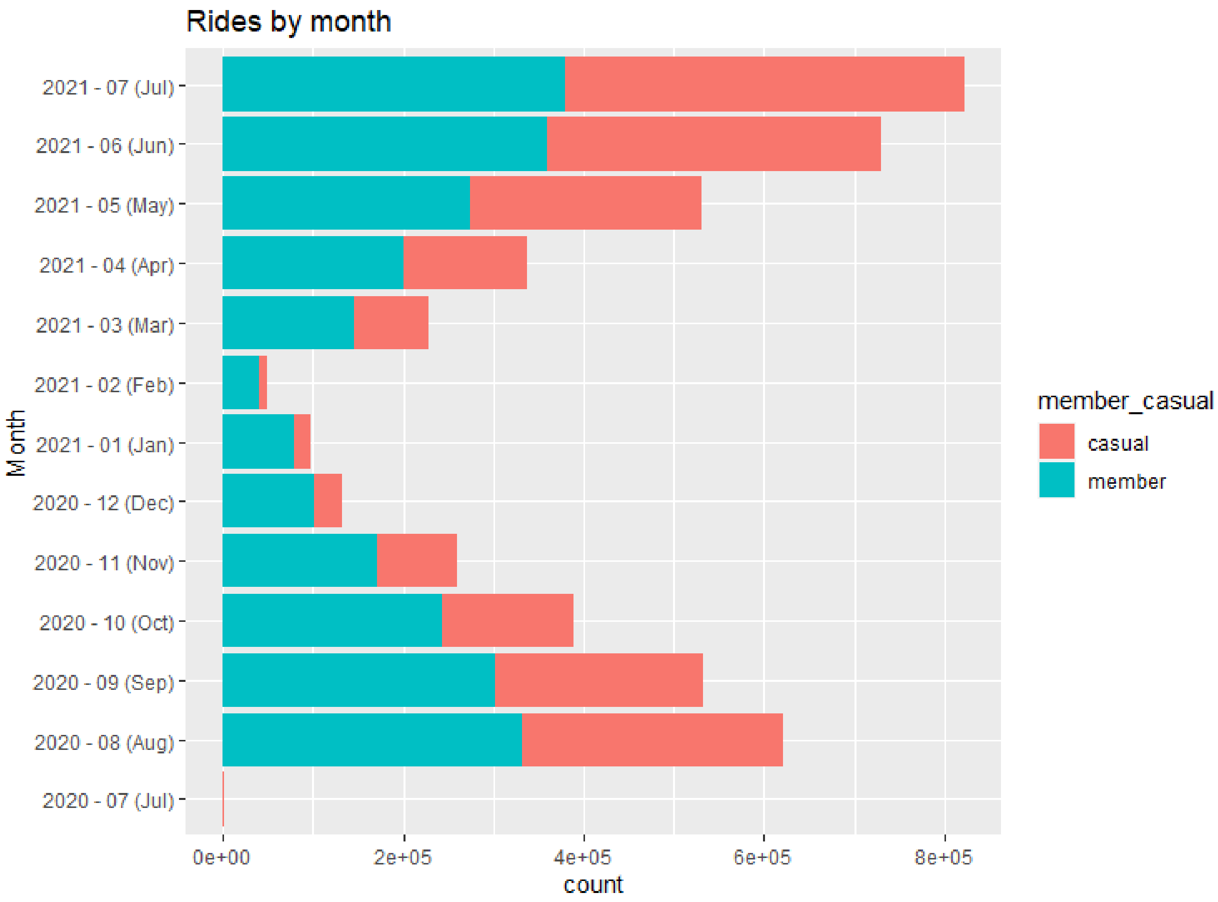


Usage by User per Month:

```
> trip_data %>%
+     group_by(year_month) %>%
+     summarise(count = length(ride_id),
+               '%' = (length(ride_id) / nrow(trip_data)) * 100,
+               'members_p' = (sum(member_casual == "member") / length(ride_id)) * 100,
+               'casual_p' = (sum(member_casual == "casual") / length(ride_id)) * 100,
+               'Member VS Casual Perc Difer' = members_p - casual_p)
```

```
# A tibble: 13 x 6
  year_month       count    `%` members_p casual_p `Member VS Casual Perc Difer`
  <chr>            <int>   <dbl>     <dbl>    <dbl>                         <dbl>
 1 2020 - 07 (Jul)   1383  0.0292      20.7     79.3                         -58.6
 2 2020 - 08 (Aug) 621136 13.1         53.5     46.5                           7.05
 3 2020 - 09 (Sep) 532946 11.3         56.7     43.3                          13.4
 4 2020 - 10 (Oct) 389046  8.22        62.7     37.3                          25.4
 5 2020 - 11 (Nov) 259230  5.48        66.1     33.9                          32.2
 6 2020 - 12 (Dec) 131482  2.78        77.1     22.9                          54.2
 7 2021 - 01 (Jan)  96673  2.04        81.3     18.7                         62.7
 8 2021 - 02 (Feb)  49648  1.05        79.6     20.4                         59.2
 9 2021 - 03 (Mar) 228526  4.83        63.2     36.8                         26.4
10 2021 - 04 (Apr) 337814  7.14        59.5     40.5                         18.9
11 2021 - 05 (May) 531266 11.2         51.7     48.3                          3.35
12 2021 - 06 (Jun) 729876 15.4         49.2     50.8                         -1.63
13 2021 - 07 (Jul) 821846 17.4         46.3     53.7                         -7.48
>

> trip_data %>%
+     ggplot(aes(year_month, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Month", title="Chart 02 - Distribution by month")
> trip_data %>%
+     ggplot(aes(year_month, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Month", title="Chart 02 - Distribution by month") +
+     coord_flip()
> trip_data %>%
+     ggplot(aes(year_month, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Month", title="Rides by month") +
+     coord_flip()
```

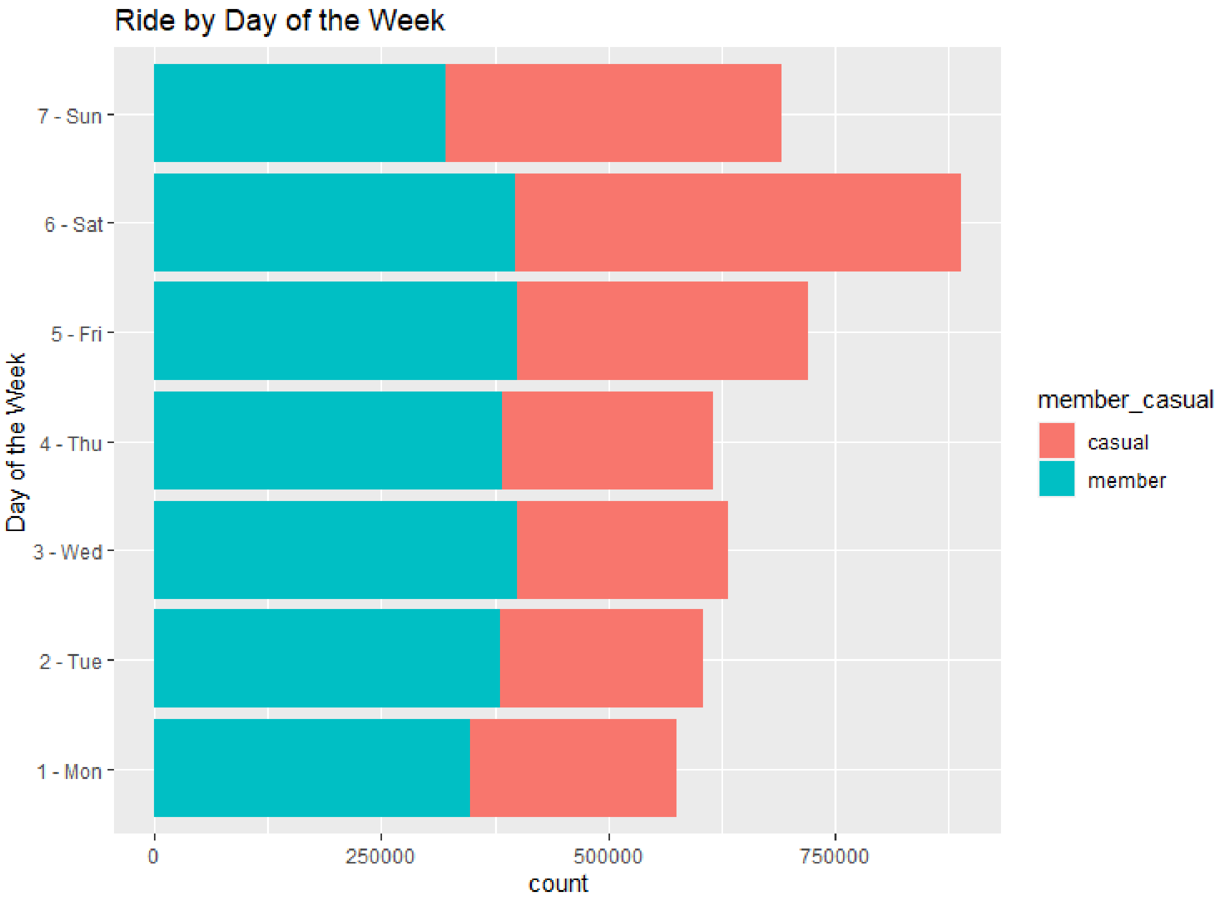


*Takeaway: Weather in Chicago impacts Usage

Rides by Day of Week:

```
> trip_data %>%
+     group_by(weekday) %>%
+     summarise(count = length(ride_id),
+               '%' = (length(ride_id) / nrow(trip_data)) * 100,
+               'members_p' = (sum(member_casual == "member") / length(ride_id)) * 100,
+               'casual_p' = (sum(member_casual == "casual") / length(ride_id)) * 100,
+               'Member x Casual Perc Difer' = members_p - casual_p)
# A tibble: 7 x 6
  weekday  count   `%` members_p casual_p `Member x Casual Perc Difer`
  <chr>    <int> <dbl>     <dbl>    <dbl>                        <dbl>
1 1 - Mon 575507  12.2      60.6     39.4                         21.1
```

```
2 2 - Tue 604943  12.8      63.0     37.0                       26.0
3 3 - Wed 632701  13.4      63.2     36.8                       26.5
4 4 - Thu 616709  13.0      62.0     38.0                       24.0
5 5 - Fri 721004  15.2      55.3     44.7                      10.6
6 6 - Sat 889065  18.8      44.7     55.3                      -10.5
7 7 - Sun 690943  14.6      46.4     53.6                       -7.30
>
> #use coor_flip again to view axis titles.
>
> ggplot(trip_data, aes(weekday, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Day of the Week", title="Ride by Day of the Week") +
+     coord_flip()
```

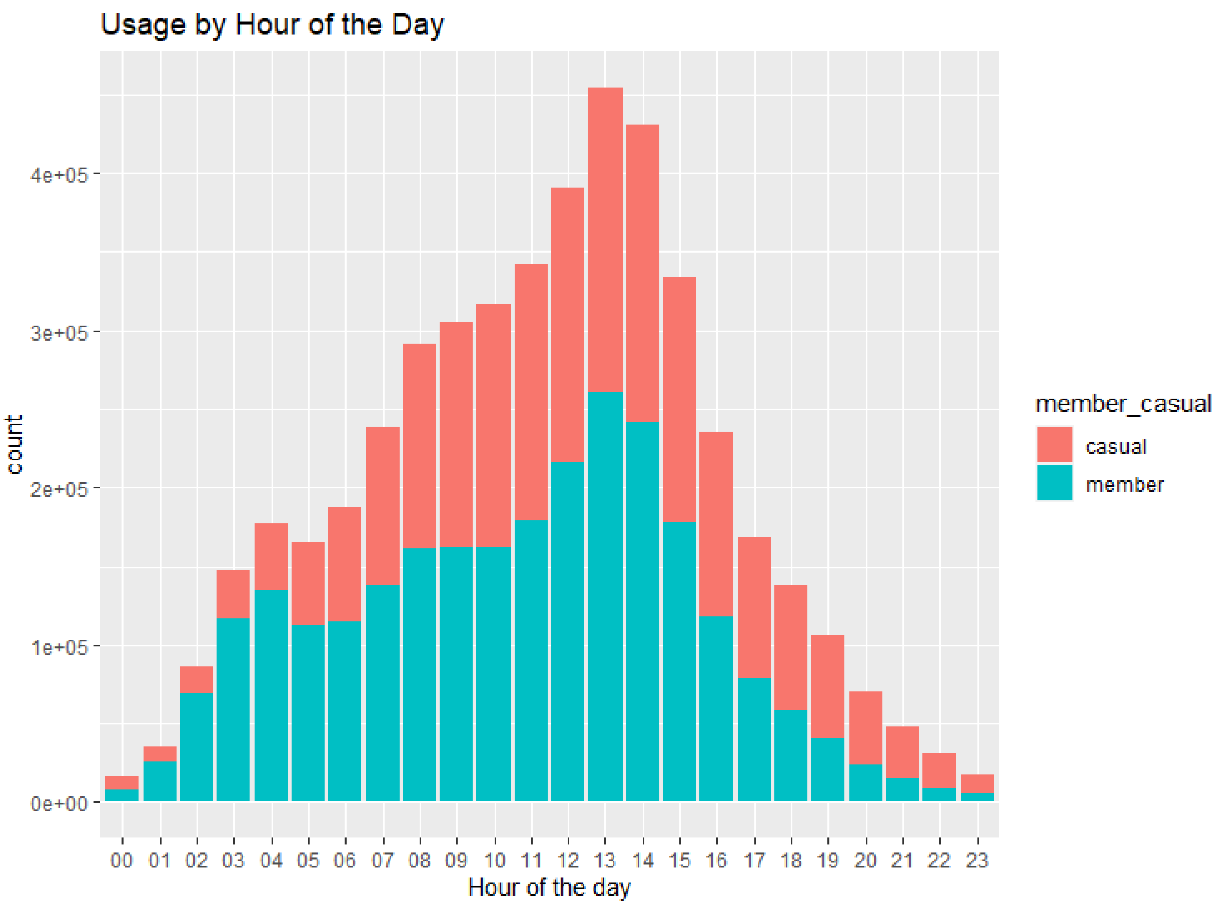


*Takeaway: Members are more consistent across the week, use bikes more on weekdays. Saturday is by far the most popular day.

Usage by Hour of the Day:

```
> trip_data %>%
+     group_by(start_hour) %>%
+     summarise(count = length(ride_id),
+               '%' = (length(ride_id) / nrow(trip_data)) * 100,
+               'members_p' = (sum(member_casual == "member") / length(ride_id)) * 100,
+               'casual_p' = (sum(member_casual == "casual") / length(ride_id)) * 100,
+               'member_casual_perc_difer' = members_p - casual_p)
# A tibble: 24 x 6
   start_hour  count   `%` members_p casual_p member_casual_perc_difer
   <chr>       <int> <dbl>     <dbl>    <dbl>                    <dbl>
 1 00          16087 0.340      49.5     50.5                    -1.06
 2 01          34851 0.737      75.1     24.9                    50.3
 3 02          86553 1.83       79.8     20.2                    59.7
 4 03         146921 3.11       79.6     20.4                    59.1
 5 04         177474 3.75       75.7     24.3                    51.3
 6 05         165275 3.49       68.3     31.7                    36.6
 7 06         187341 3.96       61.4     38.6                    22.7
 8 07         238752 5.05       57.8     42.2                    15.6
 9 08         290672 6.14       55.5     44.5                    10.9
10 09         305313 6.45       53.0     47.0                     5.95
# ... with 14 more rows
>
```

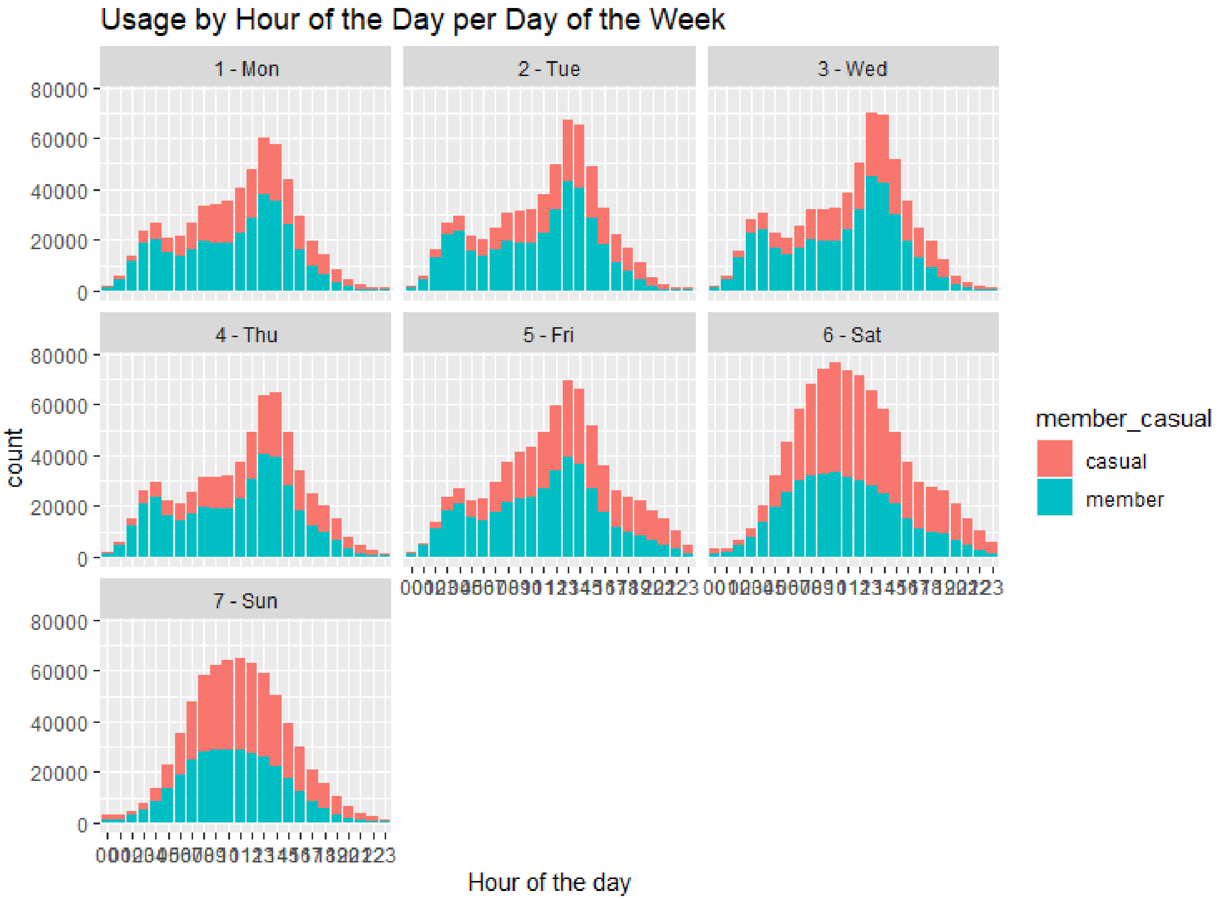
```
> trip_data %>%
+     ggplot(aes(start_hour, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Hour of the day", title="Usage by Hour of the Day per Day of the Week") +
+     facet_wrap(~ weekday)
```



*Takeaway: Members use bikes a bit at the start of workday, but definitely more at the end of the workday, assuming for commuting home or elsewhere, also perhaps for exercise.

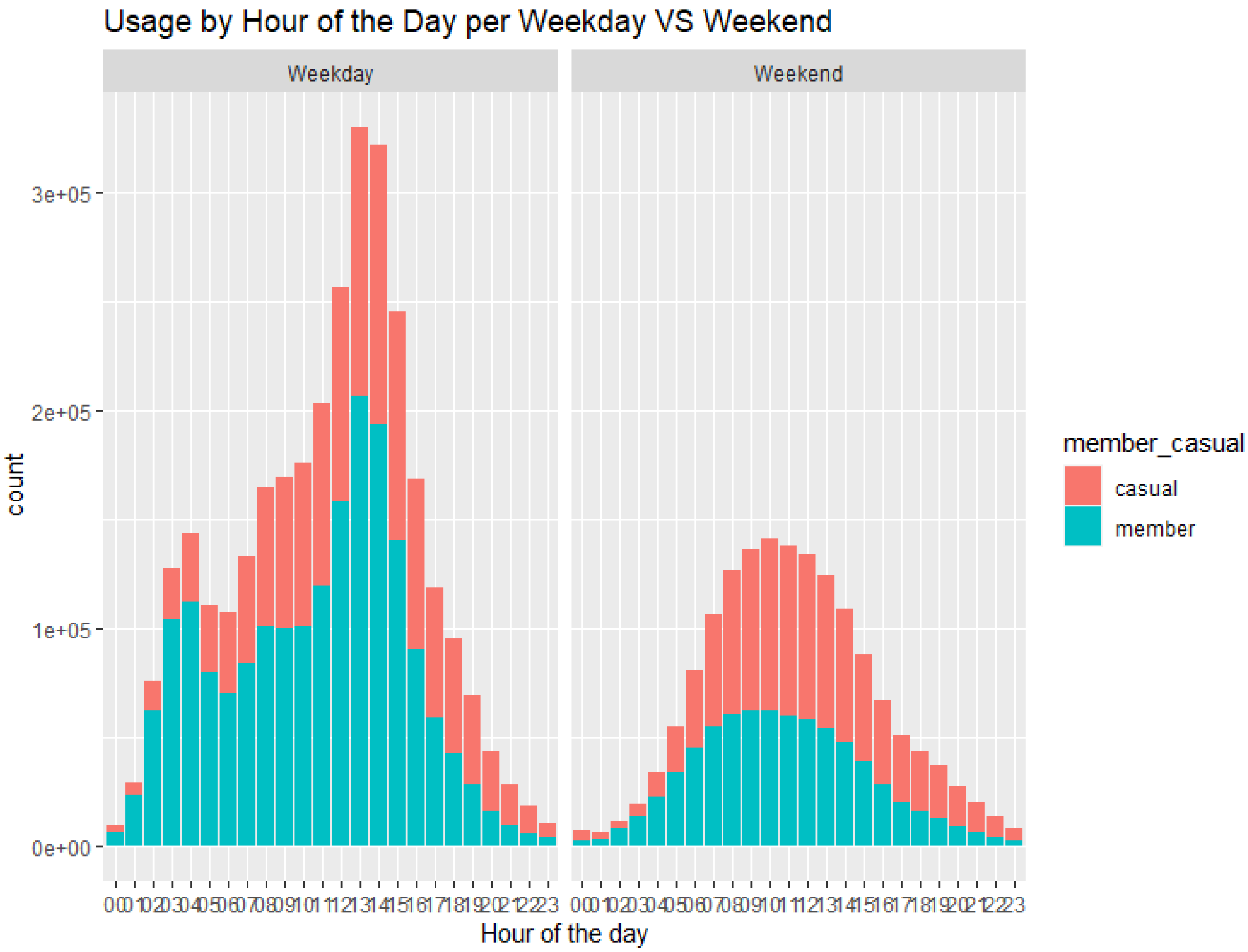
Usage per Hour of the Day per Day of the Week:

```
> trip_data %>%
+     ggplot(aes(start_hour, fill=member_casual)) +
+     geom_bar() +
+     labs(x="Hour of the day", title="Usage by Hour of the Day per Day of the Week") +
+     facet_wrap(~ weekday)
```



Comparing Usage Per Hour of the Day on Weekdays VS Weekends:

```
> trip_data %>%
+     mutate(type_of_weekday = ifelse(weekday == '6 - Sat' | weekday == '7 - Sun',
+                                     'Weekend',
+                                     'Weekday')) %>%
+     ggplot(aes(start_hour, fill=member_casual)) +
+     labs(x="Hour of the day", title="Usage by Hour of the Day per Weekday VS Weekend") +
+     geom_bar() +
+     facet_wrap(~ type_of_weekday)
>
```

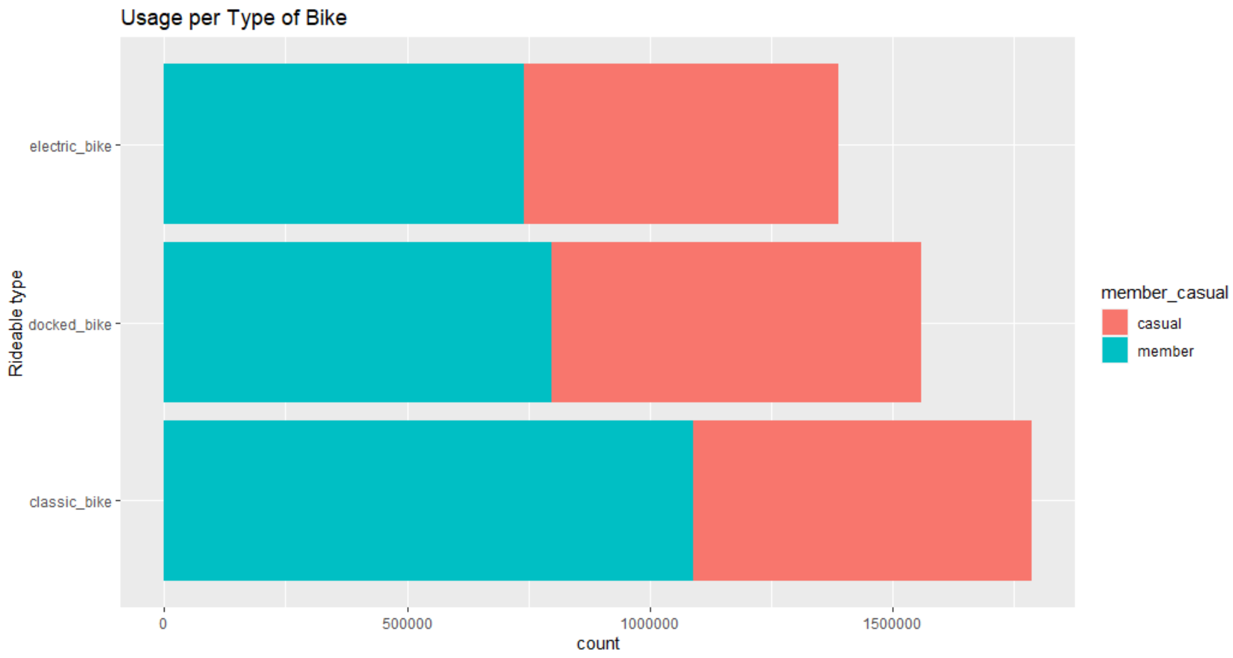


[Apologies for the clustered data points on the X axis. A better storyboard will be added at a later date.]

Types of Bikes Used:

```
> trip_data %>%
+     group_by(rideable_type) %>%
+     summarise(count = length(ride_id),
+               '%' = (length(ride_id) / nrow(trip_data)) * 100,
+               'members_p' = (sum(member_casual == "member") / length(ride_id)) * 100,
+               'casual_p' = (sum(member_casual == "casual") / length(ride_id)) * 100,
+               'member_casual_perc_difer' = members_p - casual_p)
# A tibble: 3 x 6
  rideable_type   count  `%` members_p casual_p member_casual_perc_difer
  <fct>           <int> <dbl>     <dbl>    <dbl>                    <dbl>
1 classic_bike  1785514  37.7      61.1     38.9                     22.1
```

```
2 docked_bike   1558141  32.9      51.2     48.8                   2.41
3 electric_bike 1387217  29.3      53.4     46.6                   6.80
>
>
> ggplot(trip_data, aes(rideable_type, fill=member_casual)) +
+     labs(x="Rideable type", title="Usage per Type of Bike") +
+     geom_bar() +
+     coord_flip()
```



Finally, I'll be honest here. The difficult part of taking online classes is not having anyone to ask for help. I felt I went above what was asked for a junior analyst role I stopped attempting to fix the outlier problem of the ride time. When I look at the summary it shows a negative number for the minimum (-29049.97 in minutes) and a huge number for the maximum (55944.15 in minutes). I spent a lot of time on StackOverflow attempting to fixed this and my best an final attempt still showed an error:

```
> #outliers
>
> summary(trip_data$ride_time_m)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-29049.97       7.32     13.17     23.46     24.08  55944.15
>
>
> #remove 5%
>
```

```
> trip_data_without_outliers <- trip_data %>%
+     filter(ride_time_m > as.numeric(ventiles['5%'])) %>%
+     filter(ride_time_m < as.numeric(ventiles['95%']))
Error: Problem with `filter()` input `..1`.
i Input `..1` is `ride_time_m > as.numeric(ventiles["5%"])`.
x object 'ventiles' not found
Run `rlang::last_error()` to see where the error occurred.
```

Main Takeaways:

- Members use the bikes more often, but for a lesser amount of time.This shows they are using it for the purpose of commuting or exercising during the week at commute time and after work time.
- Members like the classic bike more, perhaps showing the desire for exercise.
- Casual riders use it for fun on weekends and near tourist locations(I spend a lot of time in Chicago). Will include that data in a nicer presentation.
- Members make up approximately 59% of the usage.
- July ‘21 was the most popular month
- Weather greatly affects usage. Not much can be done about that.
- Saturday is the most popular day
- More usage overall on weekends
- More usage in the afternoon.

Act:
In order to switch casual user to members:

Show a cost savings for members.
Target advertising for “being green”.
Target advertising on the benefits of regular exercise.
Show hip, young, and healthy professionals using the bikes to avoid being stuck in traditional car traffic or waiting for the ‘L’ train.