# R Markdown – Data Analysis of Cyclistic

# LOAD PACKAGES

### TIDYVERSE

Load the tidyverse package as it includes the packages dplyr and readr which contains the function read_csv()

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------ tidyverse 1.3.1 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

### LUBRIDATE

Load the lubridate package to work with dates.

```
library(lubridate)
```

### SCALES

Load the scales package to work with plots.

```
library(scales)
```

### KNITR

Load the knitr package to work with R Markdown documents and to use the function kable()

```
library(knitr)
```

# IMPORT DATASET

The SQL Server table named “full_year_2021_V3” contained the Cyclistic dataset after data wrangling & data cleaning. This dataset was then exported as a CSV file to the desktop computer. This file is now being imported into R as a data frame titled “cyclistic_data”.

```r
cyclistic_data <- read_csv("sqlexport.csv")
```

```
## Rows: 5543989 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_..
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_datetime, ended_datetime
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types=FALSE` to quiet this message
```

# PREVIEW DATA FRAME

Preview the imported Cyclistic dataset such as the column names and data types by using the function glimpse().

```r
glimpse(cyclistic_data)
```

```
## Rows: 5,543,989
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_datetime   <dttm> 2021-01-23 16:14:00, 2021-01-27 18:43:00, 2021-01-~
## $ ended_datetime     <dttm> 2021-01-23 16:24:00, 2021-01-27 18:47:00, 2021-01-~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, "Wood St & Augusta ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, "657", "13258", "65~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90041, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.94000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.710~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

Function str() displays the structure of the cyclistic dataset including the class and data type of each column.

```r
str(cyclistic_data)
```

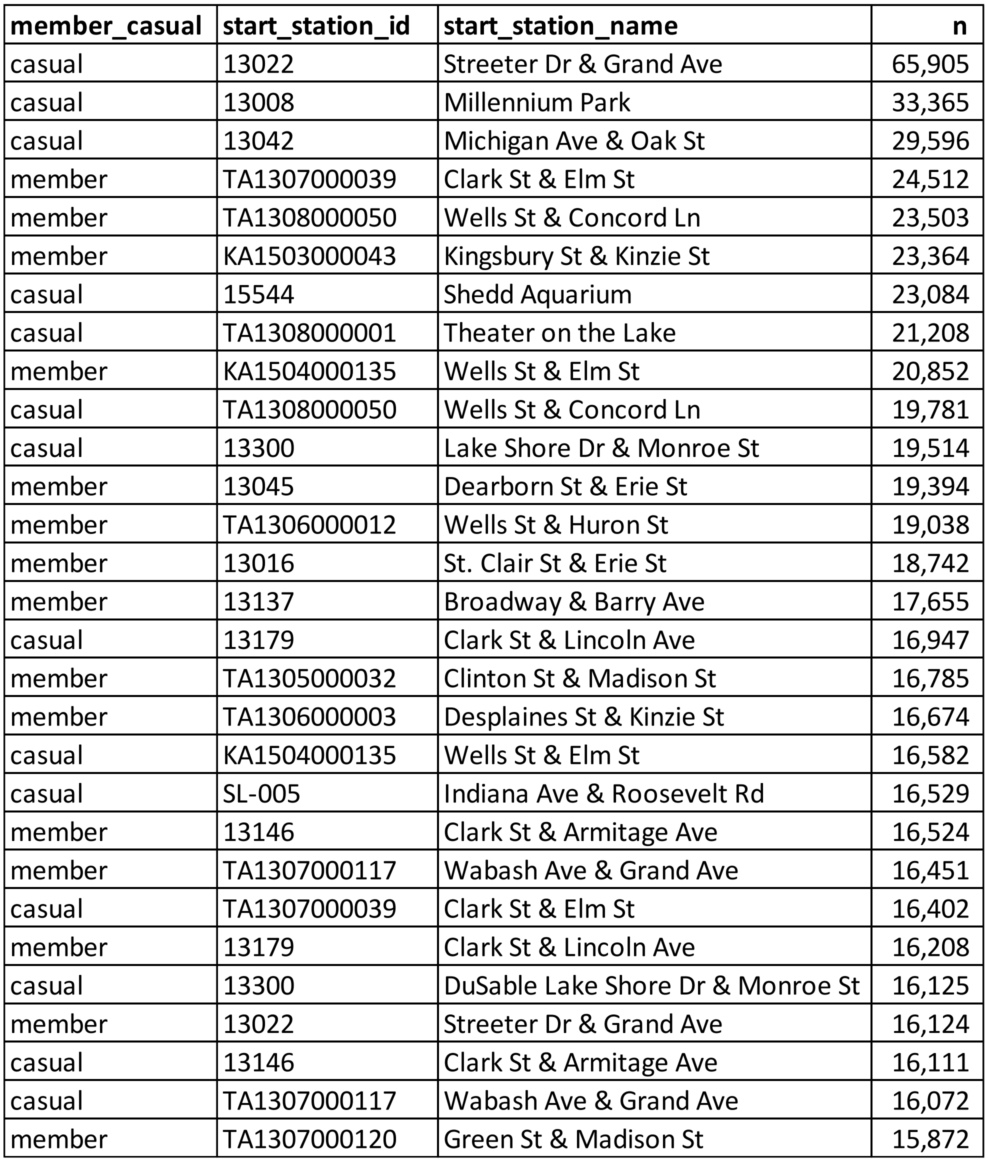
```
## spec_tbl_df [5,543,989 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5543989] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C946
83FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr [1:5543989] "electric_bike" "electric_bike" "electric_bike"
"electric_bike" ...
##  $ started_datetime  : POSIXct[1:5543989], format: "2021-01-23 16:14:00" "2021-01-27 1
8:43:00" ...
##  $ ended_datetime    : POSIXct[1:5543989], format: "2021-01-23 16:24:00" "2021-01-27 1
8:47:00" ...
##  $ start_station_name: chr [1:5543989] "California Ave & Cortez St" "California Ave &
Cortez St" "California Ave & Cortez St" "California Ave & Cortez St" ...
##  $ start_station_id  : chr [1:5543989] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5543989] NA NA NA NA ...
##  $ end_station_id    : chr [1:5543989] NA NA NA NA ...
##  $ start_lat         : num [1:5543989] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5543989] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5543989] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5543989] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5543989] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_datetime = col_datetime(format = ""),
##   ..   ended_datetime = 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>
```

# START STATIONS FREQUENTLY USED

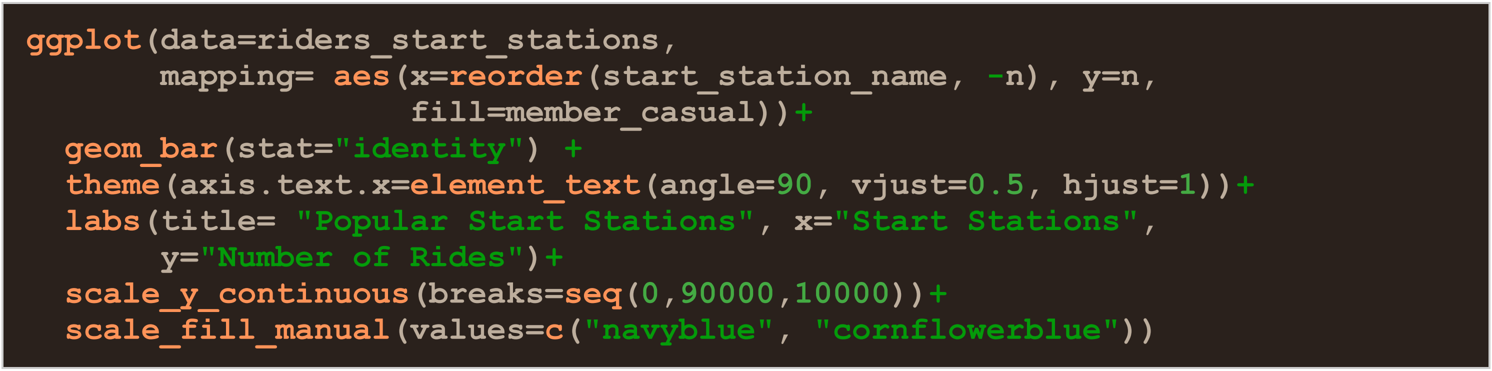
This Stacked Bar Chart displays the Top 10 start stations most frequently used by both casual & member riders. Some start stations are used solely by member riders or solely by casual riders. The X axis of this chart shows the start staton names while the Y axis displays the number of rides for each start station.

```
riders_start_stations <-
  cyclistic_data[! is.na(cyclistic_data$start_station_id),] %>%
  count(member_casual,start_station_id, start_station_name) %>%
  arrange(-n) %>%
  filter(n >= 15725)
```

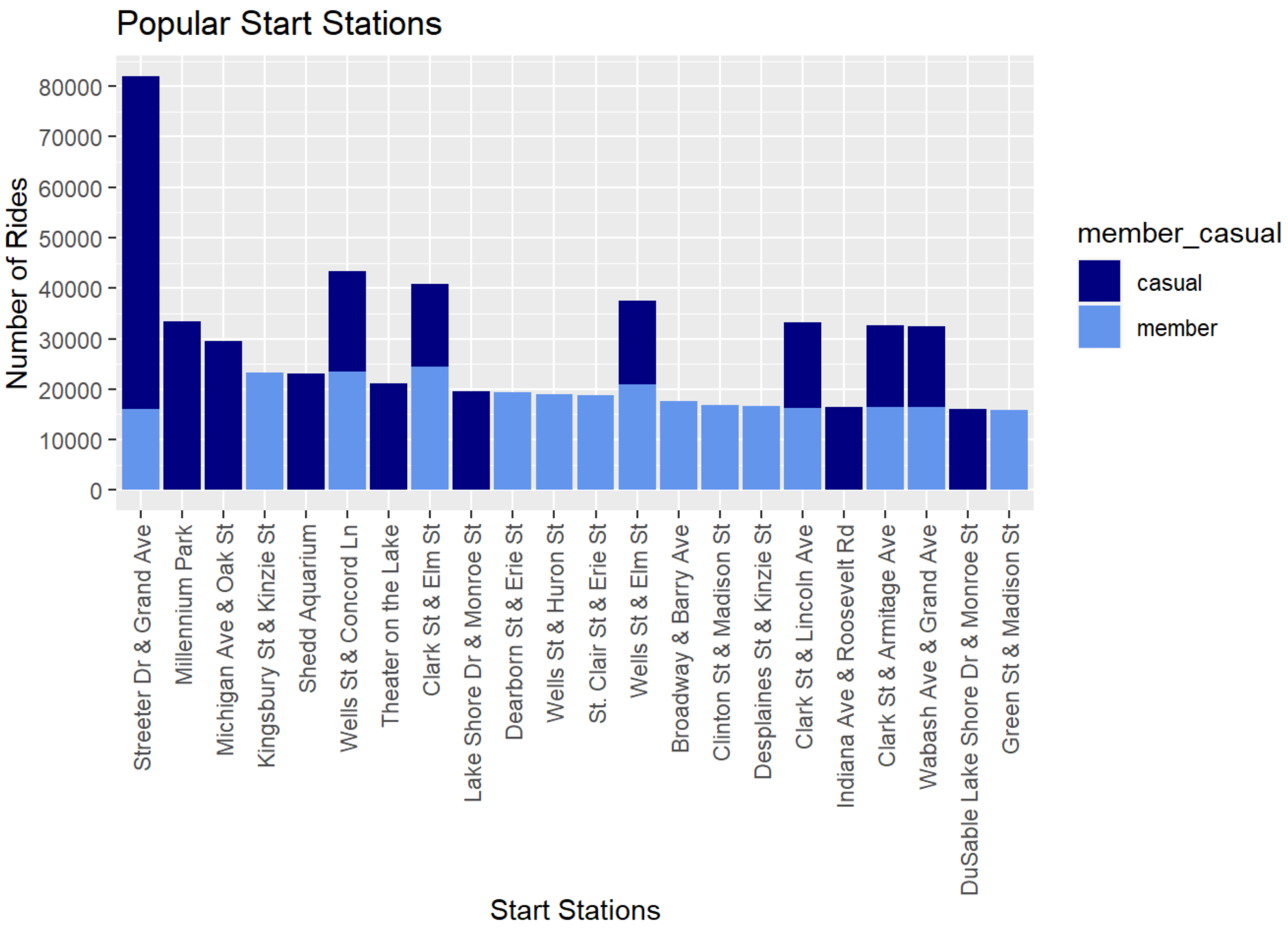
```
knitr::kable(riders_start_stations, "pipe")
```



| member_casual | start_station_id | start_station_name | n |
|---|---|---|---:|
| casual | 13022 | Streeter Dr & Grand Ave | 65,905 |
| casual | 13008 | Millennium Park | 33,365 |
| casual | 13042 | Michigan Ave & Oak St | 29,596 |
| member | TA1307000039 | Clark St & Elm St | 24,512 |
| member | TA1308000050 | Wells St & Concord Ln | 23,503 |
| member | KA1503000043 | Kingsbury St & Kinzie St | 23,364 |
| casual | 15544 | Shedd Aquarium | 23,084 |
| casual | TA1308000001 | Theater on the Lake | 21,208 |
| member | KA1504000135 | Wells St & Elm St | 20,852 |
| casual | TA1308000050 | Wells St & Concord Ln | 19,781 |
| casual | 13300 | Lake Shore Dr & Monroe St | 19,514 |
| member | 13045 | Dearborn St & Erie St | 19,394 |
| member | TA1306000012 | Wells St & Huron St | 19,038 |
| member | 13016 | St. Clair St & Erie St | 18,742 |
| member | 13137 | Broadway & Barry Ave | 17,655 |
| casual | 13179 | Clark St & Lincoln Ave | 16,947 |
| member | TA1305000032 | Clinton St & Madison St | 16,785 |
| member | TA1306000003 | Desplaines St & Kinzie St | 16,674 |
| casual | KA1504000135 | Wells St & Elm St | 16,582 |
| casual | SL-005 | Indiana Ave & Roosevelt Rd | 16,529 |
| member | 13146 | Clark St & Armitage Ave | 16,524 |
| member | TA1307000117 | Wabash Ave & Grand Ave | 16,451 |
| casual | TA1307000039 | Clark St & Elm St | 16,402 |
| member | 13179 | Clark St & Lincoln Ave | 16,208 |
| casual | 13300 | DuSable Lake Shore Dr & Monroe St | 16,125 |
| member | 13022 | Streeter Dr & Grand Ave | 16,124 |
| casual | 13146 | Clark St & Armitage Ave | 16,111 |
| casual | TA1307000117 | Wabash Ave & Grand Ave | 16,072 |
| member | TA1307000120 | Green St & Madison St | 15,872 |



```
ggplot(data=riders_start_stations,
       mapping= aes(x=reorder(start_station_name, -n), y=n,
                    fill=member_casual))+
  geom_bar(stat="identity") +
  theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))+
  labs(title= "Popular Start Stations", x="Start Stations",
       y="Number of Rides")+
  scale_y_continuous(breaks=seq(0,90000,10000))+
  scale_fill_manual(values=c("navyblue", "cornflowerblue"))
```



# BIKE ROUTES FREQUENTLY USED

There are 246,265 bike routes. A route is defined as the pairing of a start station and end station for each bike ride. This bar chart shows the most popular bike routes with the highest number of rides.

```r
cyclistic_data2 <-
  cyclistic_data %>%
  drop_na() %>%
  select(start_station_id, end_station_id, start_station_name,
         end_station_name, member_casual) %>%
  filter(start_station_id != end_station_id) %>%
  unite ("id_start_end",
         start_station_id, end_station_id, sep = "-") %>%
  unite ("name_start_end",
         start_station_name, end_station_name, sep = "-") %>%
  count(id_start_end, name_start_end, member_casual) %>%
  arrange(-n)
```

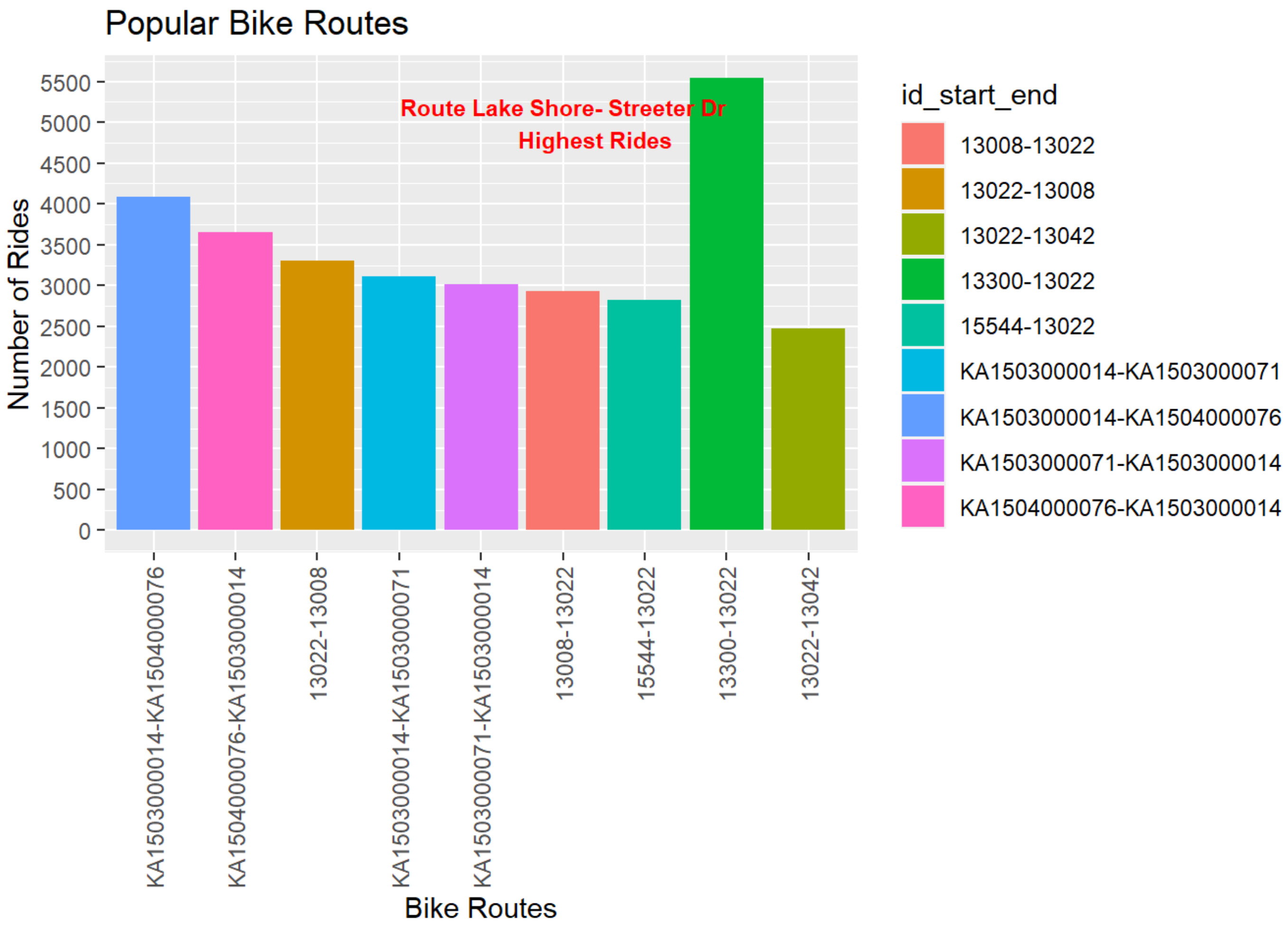
```r
cyclistic_data3 <- cyclistic_data2 %>%
                   top_n(10)
```

```
## Selecting by n
```

```r
knitr::kable(cyclistic_data3, "pipe")
```

| id_start_end | name_start_end | member_casual | n |
|---|---|---|---:|
| KA1503000014-KA1504000076 | Ellis Ave & 60th St-Ellis Ave & 55th St | member | 4,082 |
| KA1504000076-KA1503000014 | Ellis Ave & 55th St-Ellis Ave & 60th St | member | 3,652 |
| 13022-13008 | Streeter Dr & Grand Ave-Millennium Park | casual | 3,309 |
| KA1503000014-KA1503000071 | Ellis Ave & 60th St-University Ave & 57th St | member | 3,109 |
| KA1503000071-KA1503000014 | University Ave & 57th St-Ellis Ave & 60th St | member | 3,010 |
| 13008-13022 | Millennium Park-Streeter Dr & Grand Ave | casual | 2,927 |
| 15544-13022 | Shedd Aquarium-Streeter Dr & Grand Ave | casual | 2,822 |
| 13300-13022 | Lake Shore Dr & Monroe St-Streeter Dr & Grand Ave | casual | 2,811 |
| 13300-13022 | DuSable Lake Shore Dr & Monroe St-Streeter Dr & Grand Ave | casual | 2,736 |
| 13022-13042 | Streeter Dr & Grand Ave-Michigan Ave & Oak St | casual | 2,478 |

```
ggplot(data=cyclistic_data3,
       mapping= aes(x=reorder(id_start_end, -n), y=n,
                    fill=id_start_end))+
  geom_bar(stat="identity") +
  theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))+
  labs(title= "Popular Bike Routes", x="Bike Routes",
       y="Number of Rides")+
  scale_y_continuous(breaks=seq(0,6000,500))+
  annotate("text", x="13008-13022", y=5000, label="Route Lake Shore-Stree
           ter Dr Highest Rides", color="red", fontface="bold", size=3)
```



# SEASON WITH HIGHEST NUMBER OF RIDES

This Scatter Plot displays number of rides by month for both casual and member riders. The summer season in Chicago, June to September had the highest number of rides. This is highlighted in the chart by creating an annotation of a yellow rectangle.

The data type of column started_datetime is POSIXct and Double which is a DateTime

```
class(cyclistic_data$started_datetime)
```

```
## [1] "POSIXct" "POSIXt"
```

```
typeof(cyclistic_data$started_datetime)
```

```
## [1] "double"
```

Data frame copy is created called cyclistic_data5

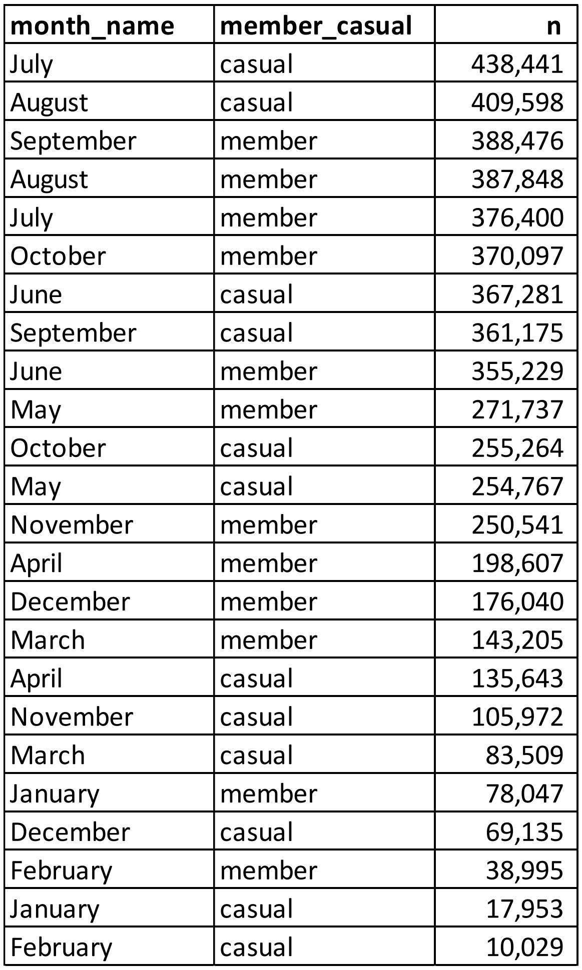
```
cyclistic_data5 <- cyclistic_data
```

Month name for column started_datetime is included in the data frame.

```
cyclistic_data5$month_name <-
  month(ymd_hms(cyclistic_data5$started_datetime),
        label=TRUE, abbr=FALSE)
```

```
season_rides <- cyclistic_data5 %>%
                count(month_name, member_casual) %>%
                arrange(-n)
```

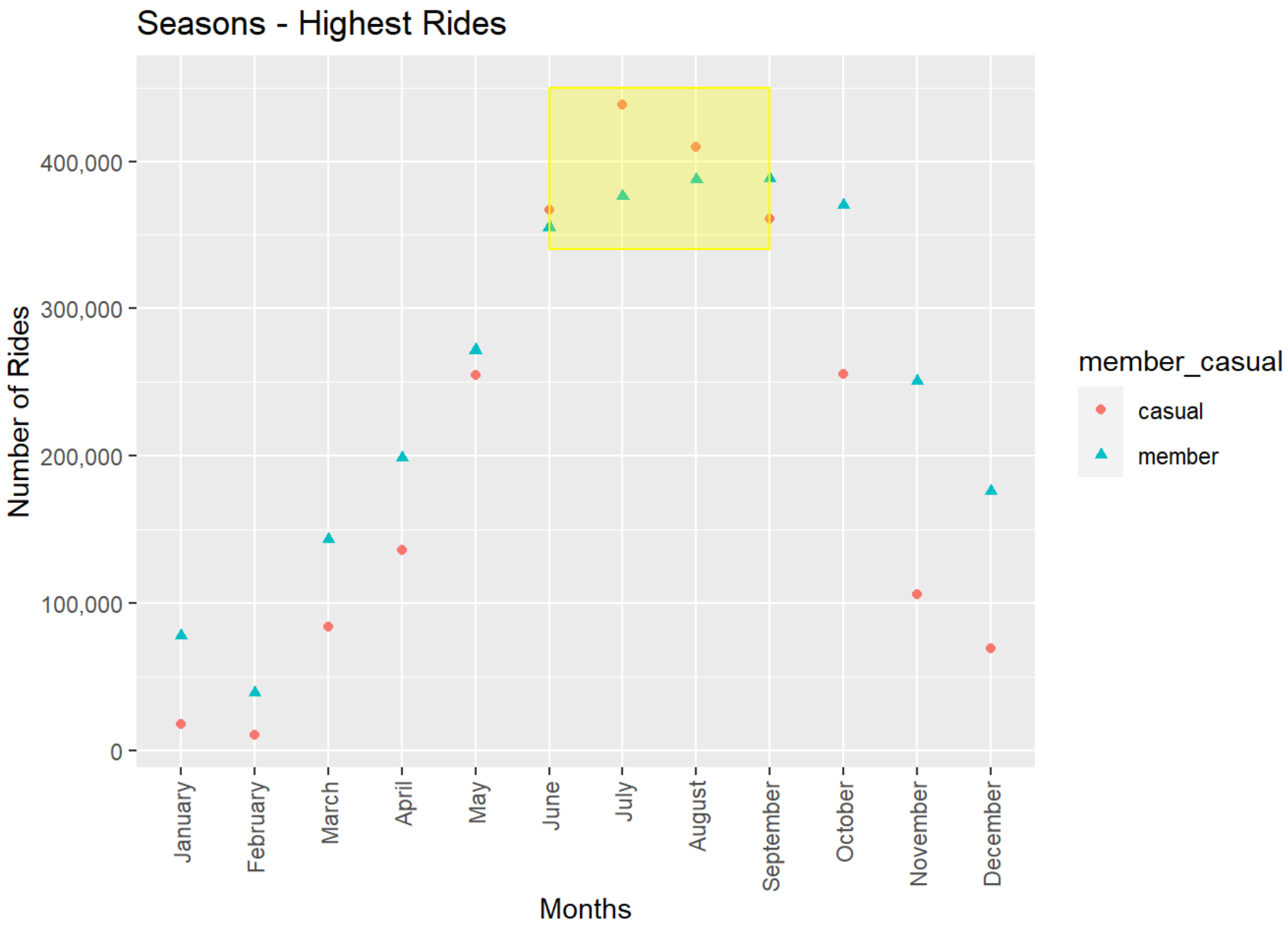
```
knitr::kable(season_rides, "pipe")
```



| month_name | member_casual | n |
|---|---|---:|
| July | casual | 438,441 |
| August | casual | 409,598 |
| September | member | 388,476 |
| August | member | 387,848 |
| July | member | 376,400 |
| October | member | 370,097 |
| June | casual | 367,281 |
| September | casual | 361,175 |
| June | member | 355,229 |
| May | member | 271,737 |
| October | casual | 255,264 |
| May | casual | 254,767 |
| November | member | 250,541 |
| April | member | 198,607 |
| December | member | 176,040 |
| March | member | 143,205 |
| April | casual | 135,643 |
| November | casual | 105,972 |
| March | casual | 83,509 |
| January | member | 78,047 |
| December | casual | 69,135 |
| February | member | 38,995 |
| January | casual | 17,953 |
| February | casual | 10,029 |

```r
ggplot(data=season_rides)+
 geom_point(mapping =aes(x=month_name, y=n, color=member_casual,
                          shape=member_casual))+
 geom_smooth(mapping =aes(x=month_name, y=n))+
 theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))+
 labs(title= "Seasons - Highest Rides", x="Months",
      y="Number of Rides")+
 scale_y_continuous(labels=comma) +
 annotate("rect", xmin="June", xmax="September", ymin=340000, ymax=450000,
          fill="yellow", color="yellow", alpha=0.3)
```

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



# DAYS WITH HIGHEST NUMBER OF RIDES

This Scatter Plot displays number of rides taken for each day of the week during the year 2021 for both casual and member riders.

The name of the day for each date in the column started_datetime has been included in the data frame.

The size and shape of the data points for member riders are different from the casual riders.

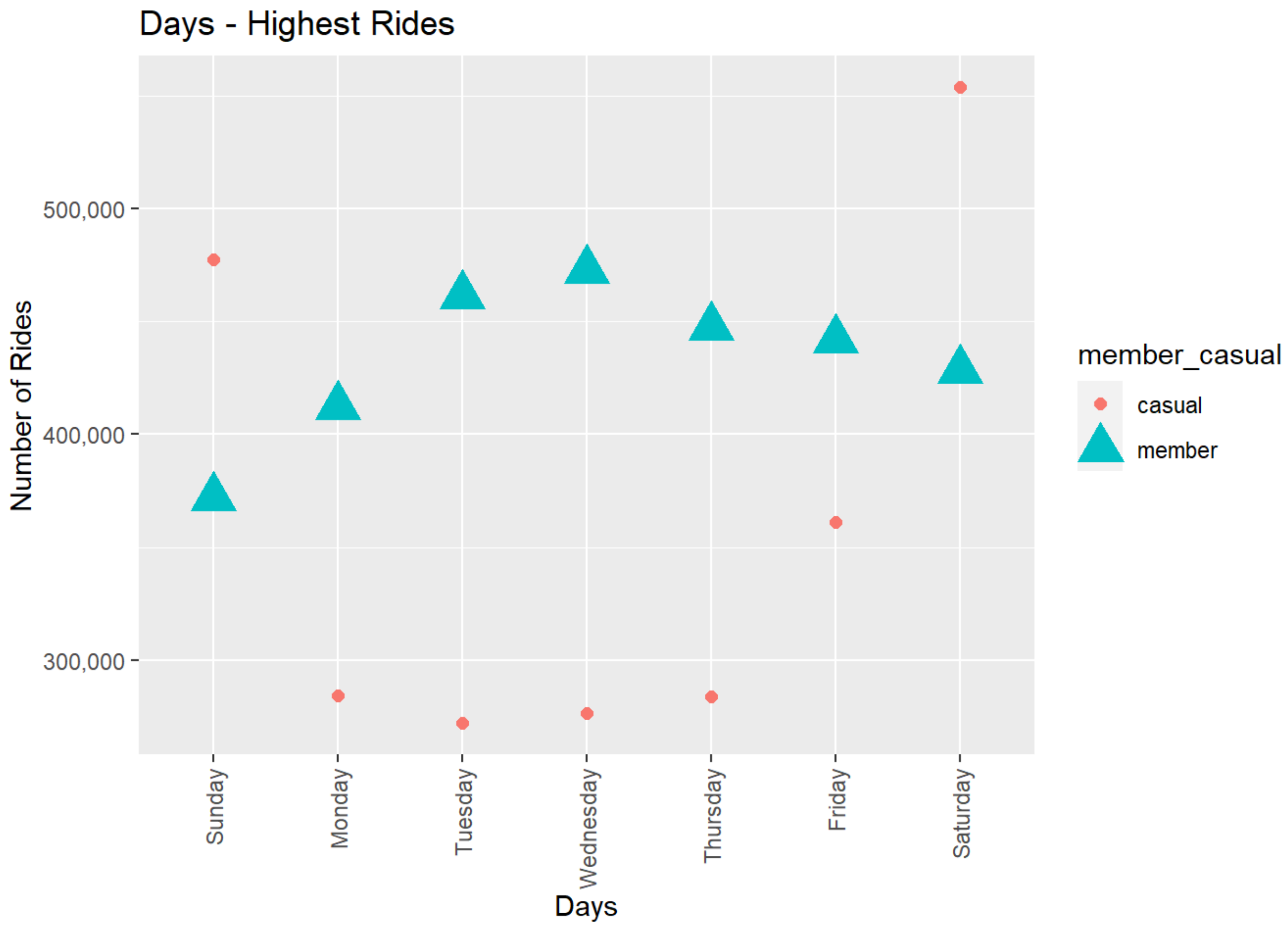
```
cyclistic_data5$day_name <-
  wday(ymd_hms(cyclistic_data5$started_datetime),
        label=TRUE, abbr=FALSE)
```

```
day_rides <- cyclistic_data5 %>%
             count(day_name, member_casual) %>%
             arrange(-n)
```

```
knitr::kable(day_rides, "pipe")
```

| day_name | member_casual | n |
|---|---|---:|
| Saturday | casual | 553,656 |
| Sunday | casual | 477,308 |
| Wednesday | member | 472,559 |
| Tuesday | member | 461,134 |
| Thursday | member | 447,235 |
| Friday | member | 441,939 |
| Saturday | member | 428,243 |
| Monday | member | 412,116 |
| Sunday | member | 371,996 |
| Friday | casual | 361,067 |
| Monday | casual | 284,134 |
| Thursday | casual | 283,737 |
| Wednesday | casual | 276,691 |
| Tuesday | casual | 272,174 |

```
ggplot(data=day_rides)+
  geom_point(mapping =aes(x=day_name, y=n, color=member_casual,
                          shape=member_casual, size=member_casual))+
  theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))+
  labs(title= "Days - Highest Rides", x="Days",
       y="Number of Rides")+
  scale_y_continuous(labels=comma)
```



# TIMES WITH HIGHEST NUMBER OF RIDES

This Scatter Plot displays number of rides by the time of day for both casual and member riders, during the year 2021.

Time by hour for column started_datetime is included in the data frame.

The geom_smooth function is used to show the regression line and patterns in the data.

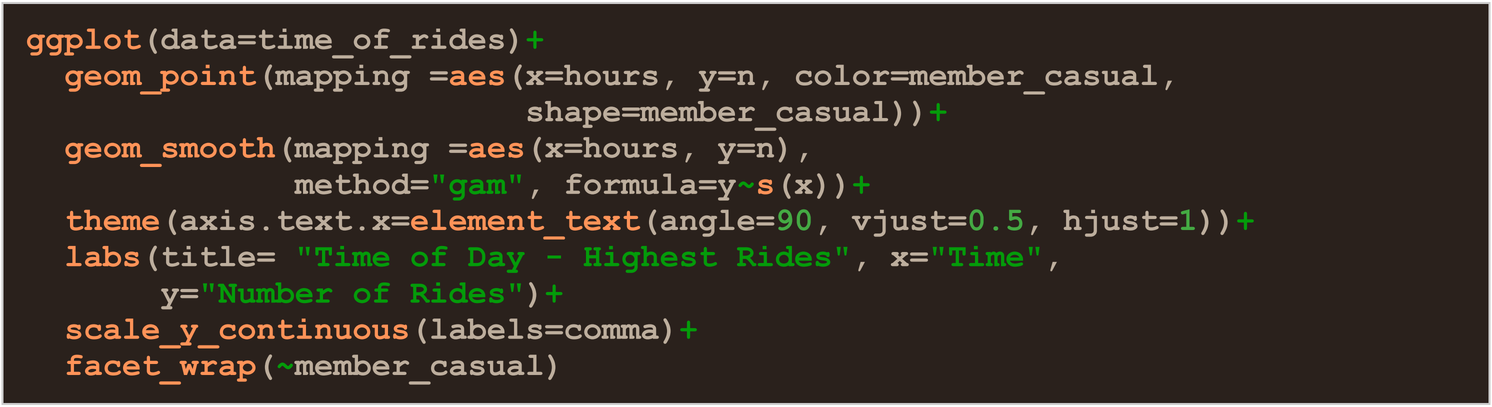
The facet_wrap function is used to display separate plots for member and casual riders which are subsets of the main dataset.

```
cyclistic_data5$hours <-
  hour(ymd_hms(cyclistic_data5$started_datetime))
```

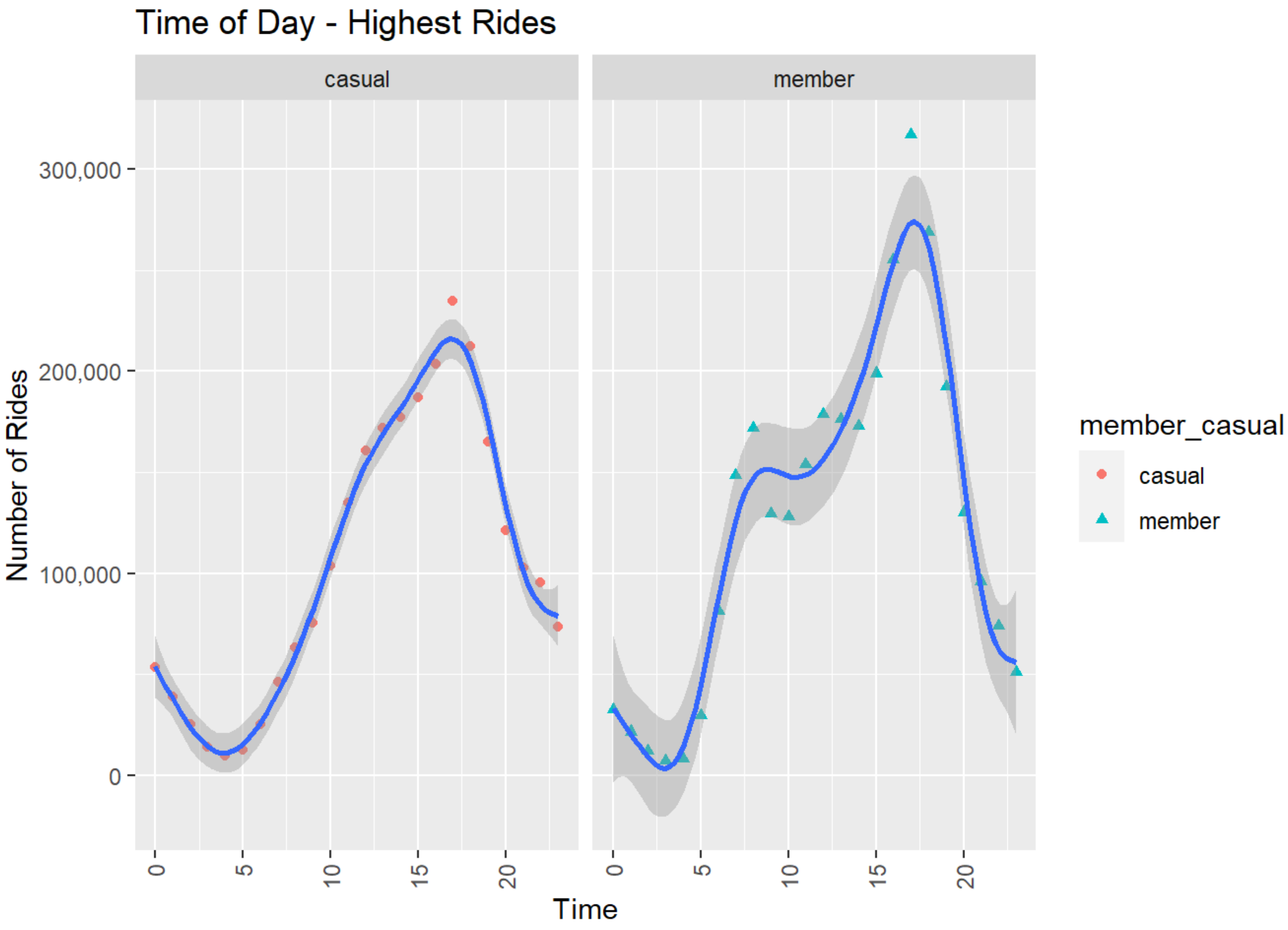
```
time_of_rides <- cyclistic_data5 %>%
                 count(hours, member_casual) %>%
                 arrange(-n)
```

```
knitr::kable(time_of_rides, "pipe")
```

| hours | member_casual | n |
|---:|:---:|---:|
| 17 | member | 317,113 |
| 18 | member | 268,810 |
| 16 | member | 254,992 |
| 17 | casual | 234,804 |
| 18 | casual | 212,186 |
| 16 | casual | 203,674 |
| 15 | member | 198,894 |
| 19 | member | 192,317 |
| 15 | casual | 187,124 |
| 12 | member | 178,795 |
| 14 | casual | 177,024 |
| 13 | member | 176,115 |
| 14 | member | 172,820 |
| 13 | casual | 172,049 |
| 8 | member | 171,715 |
| 19 | casual | 165,000 |
| 12 | casual | 160,721 |
| 11 | member | 153,783 |
| 7 | member | 148,403 |
| 11 | casual | 134,895 |
| 20 | member | 130,198 |
| 9 | member | 129,593 |
| 10 | member | 127,944 |
| 20 | casual | 121,097 |
| 10 | casual | 103,827 |
| 21 | casual | 102,903 |
| 21 | member | 96,132 |
| 22 | casual | 95,457 |
| 6 | member | 81,088 |
| 9 | casual | 75,615 |
| 22 | member | 73,774 |
| 23 | casual | 73,497 |
| 8 | casual | 63,048 |
| 0 | casual | 53,419 |
| 23 | member | 51,113 |
| 7 | casual | 46,360 |
| 1 | casual | 38,951 |
| 0 | member | 32,739 |
| 5 | member | 29,782 |
| 6 | casual | 25,513 |
| 2 | casual | 25,347 |
| 1 | member | 21,635 |
| 3 | casual | 13,961 |
| 5 | casual | 12,432 |
| 2 | member | 12,319 |
| 4 | casual | 9,863 |
| 4 | member | 8,058 |
| 3 | member | 7,090 |



```
ggplot(data=time_of_rides)+
  geom_point(mapping =aes(x=hours, y=n, color=member_casual,
                          shape=member_casual))+
  geom_smooth(mapping =aes(x=hours, y=n),
              method="gam", formula=y~s(x))+
  theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))+
  labs(title= "Time of Day - Highest Rides", x="Time",
       y="Number of Rides")+
  scale_y_continuous(labels=comma)+
  facet_wrap(~member_casual)
```


Time of Day - Highest Rides
casual
member
Number of Rides
Time
member_casual
casual
member