Cyclistic Case Study

Ali Gruett

2023-06-02

Analysis of how annual members and casual riders use Cyclistic bikes differently.

Load relevant packages.

Load all packages necessary to complete the analysis. Either set the working directory or make sure the current working directory is correct.

```
library(tidyverse)  #data manipulation, exploration, and visualization
library(lubridate)  #works with dates and times
library(ggplot2)  #visualize data
library(dplyr) #good for data manipulation
library(ggmap) #visualizing data on a map
library(tidyr) #data cleaning
library(sf) #a standardized way to encode spatial vector data
library(mapview) #provides functions to create interactive visualizations of spatial data
library(reshape2)
getwd() #displays your working directory}
```

Import data into R

Read data into R from the working directory. Data is in the form of 12 csv files. Name all the data frames created. Each data frame represents all trips taken during a month. The time frame is May 2022 - April 2023.

```
May_2022 <- read_csv("202205-divvy-tripdata.csv")
Jun_2022 <- read_csv("202206-divvy-tripdata.csv")
Jul_2022 <- read_csv("202207-divvy-tripdata.csv")
Aug_2022 <- read_csv("202208-divvy-tripdata.csv")
Sep_2022 <- read_csv("202209-divvy-tripdata.csv")
Oct_2022 <- read_csv("202210-divvy-tripdata.csv")
Nov_2022 <- read_csv("202211-divvy-tripdata.csv")
Dec_2022 <- read_csv("202212-divvy-tripdata.csv")
Jan_2023 <- read_csv("202301-divvy-tripdata.csv")
Feb_2023 <- read_csv("202302-divvy-tripdata.csv")
Mar_2023 <- read_csv("202303-divvy-tripdata.csv")
Apr_2023 <- read_csv("202304-divvy-tripdata.csv")
```

Inspect dataframes to ensure data types and columns are correct and identical between all the data frames.This allows the data to append correctly.

```
str(May_2022)
```

```
## spc_tbl_ [634,858 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:634858] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415C
F" "6FF59852924528F8" ...
##  $ rideable_type     : chr [1:634858] "classic_bike" "classic_bike" "classic_bike" "classic_b
ike" ...
##  $ started_at        : POSIXct[1:634858], format: "2022-05-23 23:06:58" "2022-05-11 08:53:28"
...
##  $ ended_at          : POSIXct[1:634858], format: "2022-05-23 23:40:19" "2022-05-11 09:31:22"
...
##  $ start_station_name: chr [1:634858] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monro
e St" "Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr [1:634858] "TA1307000117" "13300" "TA1305000032" "TA1305000032"
...
##  $ end_station_name  : chr [1:634858] "Halsted St & Roscoe St" "Field Blvd & South Water St"
"Wood St & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr [1:634858] "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:634858] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:634858] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:634858] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:634858] "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>
```

Complete this step for all months.

Append into one data frame.

```
all_trips <- bind_rows(May_2022,Jun_2022,Jul_2022,Aug_2022,Sep_2022,Oct_2022,Nov_2022,Dec_2022,J
an_2023,Feb_2023,Mar_2023,Apr_2023)
```

Inspect the new data frame.

Observe the number of columns and rows, the structure of the data frame, get a preview of the data, names of the columns, and data type.

```
str(all_trips)
```

```
## spc_tbl_ [5,859,061 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5859061] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415C
F" "6FF59852924528F8" ...
##  $ rideable_type     : chr [1:5859061] "classic_bike" "classic_bike" "classic_bike" "classic_
bike" ...
##  $ started_at        : POSIXct[1:5859061], format: "2022-05-23 23:06:58" "2022-05-11 08:53:2
8" ...
##  $ ended_at          : POSIXct[1:5859061], format: "2022-05-23 23:40:19" "2022-05-11 09:31:2
2" ...
##  $ start_station_name: chr [1:5859061] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monr
oe St" "Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr [1:5859061] "TA1307000117" "13300" "TA1305000032" "TA1305000032"
...
##  $ end_station_name  : chr [1:5859061] "Halsted St & Roscoe St" "Field Blvd & South Water St"
"Wood St & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr [1:5859061] "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5859061] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5859061] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:5859061] "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>
```

For columns that are character data types, view the length. For datetime and double, view the five number summary.

```
summary(all_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5859061     Length:5859061     Min.   :2022-05-01 00:00:06.00
##  Class :character   Class :character   1st Qu.:2022-07-03 11:12:30.00
##  Mode  :character   Mode  :character   Median :2022-08-28 12:44:57.00
##                                        Mean   :2022-09-19 13:39:54.23
##                                        3rd Qu.:2022-11-08 06:30:21.00
##                                        Max.   :2023-04-30 23:59:05.00
##
##      ended_at                      start_station_name start_station_id
##  Min.   :2022-05-01 00:05:17.00   Length:5859061     Length:5859061
##  1st Qu.:2022-07-03 11:38:52.00   Class :character   Class :character
##  Median :2022-08-28 13:07:09.00   Mode  :character   Mode  :character
##  Mean   :2022-09-19 13:58:50.35
##  3rd Qu.:2022-11-08 06:43:39.00
##  Max.   :2023-05-03 10:37:12.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5859061     Length:5859061     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5859061
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.37   Max.   :  0.00
##  NA's   :5973    NA's   :5973
```

View the first several rows of data.

```
head(all_trips)
```

```
## # A tibble: 6 × 13
##   ride_id          rideable_type started_at          ended_at
##   <chr>            <chr>         <dttm>              <dttm>
## 1 EC2DE40644C6B0F4 classic_bike  2022-05-23 23:06:58 2022-05-23 23:40:19
## 2 1C31AD03897EE385 classic_bike  2022-05-11 08:53:28 2022-05-11 09:31:22
## 3 1542FBEC830415CF classic_bike  2022-05-26 18:36:28 2022-05-26 18:58:18
## 4 6FF59852924528F8 classic_bike  2022-05-10 07:30:07 2022-05-10 07:38:49
## 5 483C52CAAE12E3AC classic_bike  2022-05-10 17:31:56 2022-05-10 17:36:57
## 6 C0A3AA5A614DCE01 classic_bike  2022-05-04 14:48:55 2022-05-04 14:56:04
## # ℹ 9 more variables: start_station_name <chr>, start_station_id <chr>,
## #   end_station_name <chr>, end_station_id <chr>, start_lat <dbl>,
## #   start_lng <dbl>, end_lat <dbl>, end_lng <dbl>, member_casual <chr>
```

View the types of riders and bikes. Make sure there are no errors with data entry for these columns (ex: for members there should only be two types).

```
unique(all_trips$member_casual)
```

```
## [1] "member" "casual"
```

```
unique(all_trips$rideable_type)
```

```
## [1] "classic_bike"  "docked_bike"   "electric_bike"
```

Create a frequency table for the member_casual and rideable_type columns.

```
table(all_trips$member_casual)
```

```
##
##  casual  member
## 2358307 3500754
```

```
table(all_trips$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##       2642585        170518       3045958
```

Members have taken more rides than casual riders during the past year (May 2022-April 2023). The most common type of bike ridden is the electric bike.

Preparing for analysis

Create columns that list the date, month, day, year, day of week, and hour of each ride. Create a new column that combines the start and end station names.

```
all_trips$date <- as_datetime(all_trips$started_at)
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$hour <- format(as_datetime(all_trips$date), "%H")
all_trips$start_end <- paste(all_trips$start_station_name, ",", all_trips$end_station_name)
```

Add a calculated column for duration of the ride in minutes.

```
all_trips$ride_length <- difftime(all_trips$ended_at,all_trips$started_at, units=c("mins"))
```

Inspect the data frame to make sure the columns were added and appear correct.

```
str(all_trips)
```

```
## spc_tbl_ [5,859,061 × 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5859061] "EC2DE40644C6B0F4" "1C31AD03897EE385" "1542FBEC830415C
F" "6FF59852924528F8" ...
##  $ rideable_type     : chr [1:5859061] "classic_bike" "classic_bike" "classic_bike" "classic_
bike" ...
##  $ started_at        : POSIXct[1:5859061], format: "2022-05-23 23:06:58" "2022-05-11 08:53:2
8" ...
##  $ ended_at          : POSIXct[1:5859061], format: "2022-05-23 23:40:19" "2022-05-11 09:31:2
2" ...
##  $ start_station_name: chr [1:5859061] "Wabash Ave & Grand Ave" "DuSable Lake Shore Dr & Monr
oe St" "Clinton St & Madison St" "Clinton St & Madison St" ...
##  $ start_station_id  : chr [1:5859061] "TA1307000117" "13300" "TA1305000032" "TA1305000032"
...
##  $ end_station_name  : chr [1:5859061] "Halsted St & Roscoe St" "Field Blvd & South Water St"
"Wood St & Milwaukee Ave" "Clark St & Randolph St" ...
##  $ end_station_id    : chr [1:5859061] "TA1309000025" "15534" "13221" "TA1305000030" ...
##  $ start_lat         : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5859061] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:5859061] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5859061] -87.6 -87.6 -87.7 -87.6 -87.7 ...
##  $ member_casual     : chr [1:5859061] "member" "member" "member" "member" ...
##  $ date              : POSIXct[1:5859061], format: "2022-05-23 23:06:58" "2022-05-11 08:53:2
8" ...
##  $ month             : chr [1:5859061] "05" "05" "05" "05" ...
##  $ day               : chr [1:5859061] "23" "11" "26" "10" ...
##  $ year              : chr [1:5859061] "2022" "2022" "2022" "2022" ...
##  $ day_of_week       : chr [1:5859061] "Monday" "Wednesday" "Thursday" "Tuesday" ...
##  $ hour              : chr [1:5859061] "23" "08" "18" "07" ...
##  $ start_end         : chr [1:5859061] "Wabash Ave & Grand Ave , Halsted St & Roscoe St" "DuS
able Lake Shore Dr & Monroe St , Field Blvd & South Water St" "Clinton St & Madison St , Wood St
& Milwaukee Ave" "Clinton St & Madison St , Clark St & Randolph St" ...
##  $ ride_length       : 'difftime' num [1:5859061] 33.35 37.9 21.8333333333333 8.7 ...
##   ..- attr(*, "units")= chr "mins"
##  - 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>
```

Ride length and hour are currently a character data type. Convert them to numeric.

```
all_trips$ride_length <- as.numeric(as.character(all_trips$ride_length))
is.numeric(all_trips$ride_length)
```

```
## [1] TRUE
```

```
all_trips$hour <- as.numeric(as.character(all_trips$hour))
is.numeric(all_trips$hour)
```

```
## [1] TRUE
```

```
summary(all_trips$ride_length)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -10353.35      5.65      9.98     18.94     17.92  41387.25
```

The summary shows the minimum for ride_length is a negative value. Remove any negative values and create a new data set (version 2).

```
all_trips_v2 <- all_trips[!(all_trips$ride_length<0),]
```

```
summary(all_trips_v2)
```

```
##    ride_id          rideable_type        started_at
##  Length:5858958     Length:5858958     Min.   :2022-05-01 00:00:06.00
##  Class :character   Class :character   1st Qu.:2022-07-03 11:11:58.25
##  Mode  :character   Mode  :character   Median :2022-08-28 12:44:51.50
##                                        Mean   :2022-09-19 13:39:50.03
##                                        3rd Qu.:2022-11-08 06:32:21.75
##                                        Max.   :2023-04-30 23:59:05.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-05-01 00:05:17.00   Length:5858958     Length:5858958
##  1st Qu.:2022-07-03 11:38:33.75   Class :character   Class :character
##  Median :2022-08-28 13:07:00.50   Mode  :character   Mode  :character
##  Mean   :2022-09-19 13:58:46.32
##  3rd Qu.:2022-11-08 06:46:17.00
##  Max.   :2023-05-03 10:37:12.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5858958     Length:5858958     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5858958
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.37   Max.   :  0.00
##  NA's   :5973    NA's   :5973
##       date                          month              day
##  Min.   :2022-05-01 00:00:06.00   Length:5858958     Length:5858958
##  1st Qu.:2022-07-03 11:11:58.25   Class :character   Class :character
##  Median :2022-08-28 12:44:51.50   Mode  :character   Mode  :character
##  Mean   :2022-09-19 13:39:50.03
##  3rd Qu.:2022-11-08 06:32:21.75
##  Max.   :2023-04-30 23:59:05.00
##
##      year           day_of_week             hour        start_end
##  Length:5858958     Length:5858958     Min.   : 0.00   Length:5858958
##  Class :character   Class :character   1st Qu.:11.00   Class :character
##  Mode  :character   Mode  :character   Median :15.00   Mode  :character
##                                        Mean   :14.21
##                                        3rd Qu.:18.00
##                                        Max.   :23.00
##
##   ride_length
##  Min.   :    0.00
##  1st Qu.:    5.65
##  Median :    9.98
```

```
##  Mean   :   18.94
##  3rd Qu.:   17.92
##  Max.   :41387.25
##
```

Analysis

1. Comparison of ride length and total number of rides between members and casual riders.

Ride length analysis.

```
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                 28.49466
## 2                     member                 12.50031
```

```
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                41387.250
## 2                     member                 1559.667
```

```
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                        0
## 2                     member                        0
```

```
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                12.516667
## 2                     member                 8.666667
```

Count the total number of rides taken by members and casual riders.

```
nrow(subset(all_trips_v2,member_casual=="member"))
```

```
## [1] 3500705
```

```
nrow(subset(all_trips_v2,member_casual=="casual"))
```

```
## [1] 2358253
```

Member's trip lengths are shorter by an average of 15.99 minutes and members are using the bikes more than casual riders.

2. Comparison of rides taken each day of the week between members and casual riders.

Order the days of the week Sun-Sat.

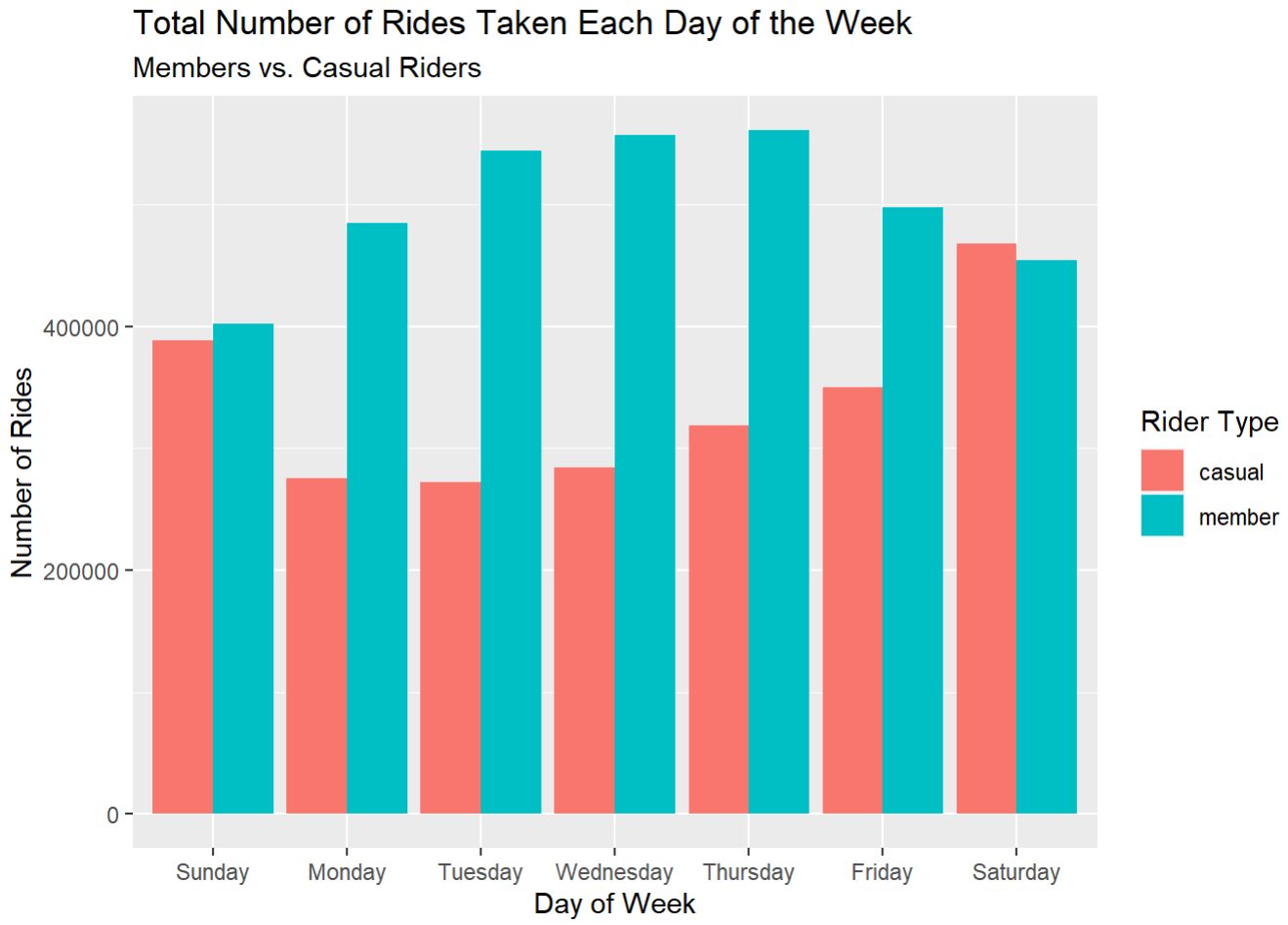
```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tues
day", "Wednesday", "Thursday", "Friday", "Saturday"))

all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)
```

```
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual day_of_week number_of_rides average_duration
##    <chr>         <ord>                 <int>            <dbl>
##  1 casual        Sunday               388811             33.4
##  2 casual        Monday               275748             28.4
##  3 casual        Tuesday              272648             25.3
##  4 casual        Wednesday            284575             24.2
##  5 casual        Thursday             318467             24.7
##  6 casual        Friday               350081             27.5
##  7 casual        Saturday             467923             32.2
##  8 member        Sunday               402066             13.8
##  9 member        Monday               484560             12.0
## 10 member        Tuesday              544393             12.0
## 11 member        Wednesday            556913             11.9
## 12 member        Thursday             560877             12.1
## 13 member        Friday               497473             12.4
## 14 member        Saturday             454423             13.9
```

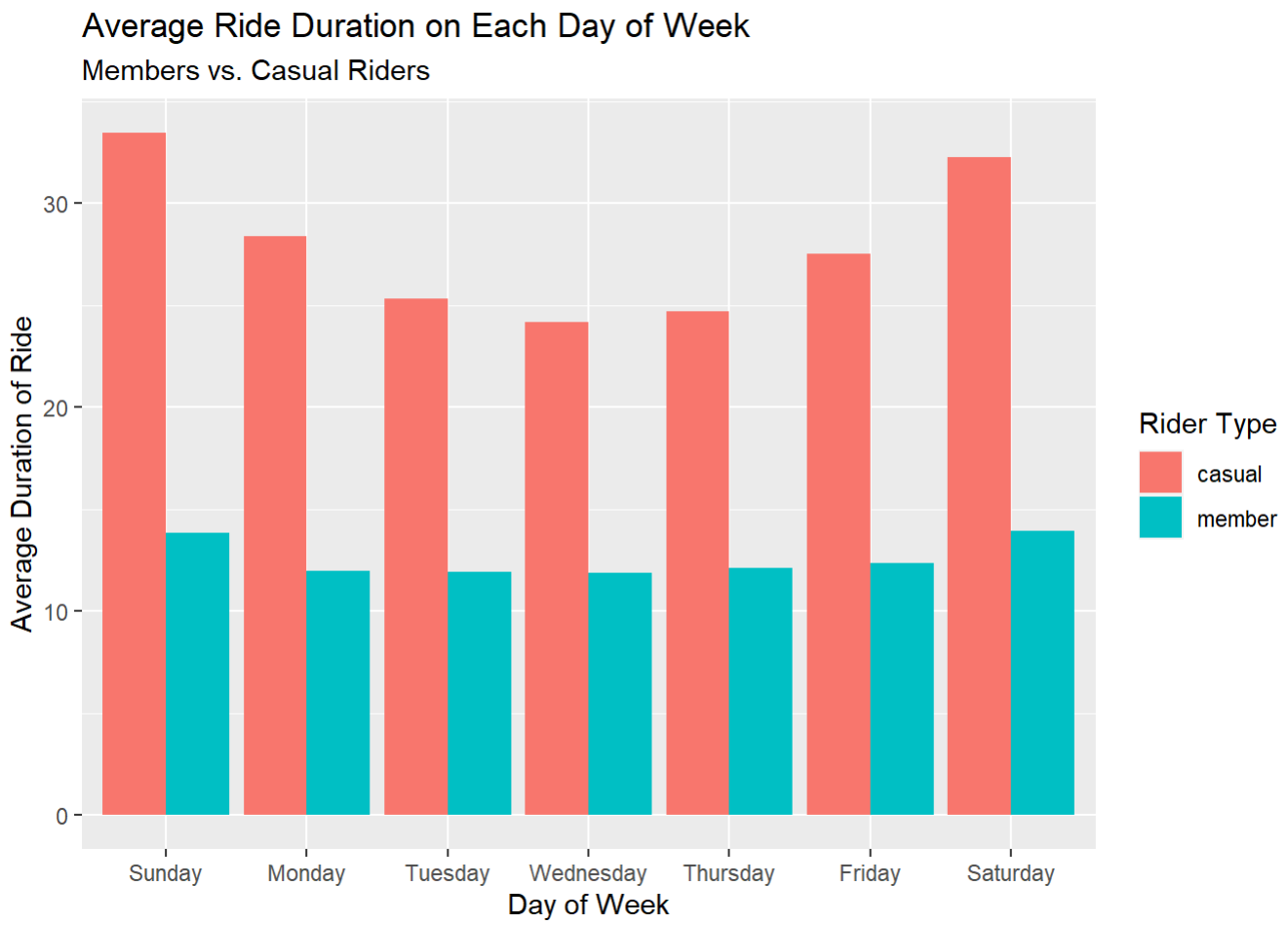
The n() function counts the number of observations by the groups identified in the summarize function

Graph as a column chart.

```
options(scipen=999)
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x="Day of Week", y="Number of Rides",
       title="Total Number of Rides Taken Each Day of the Week",
       subtitle="Members vs. Casual Riders") +
  guides(fill=guide_legend(title="Rider Type"))
```




```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x="Day of Week", y="Average Duration of Ride",
       title="Average Ride Duration on Each Day of Week",
       subtitle="Members vs. Casual Riders")+
  guides(fill=guide_legend(title="Rider Type"))
```



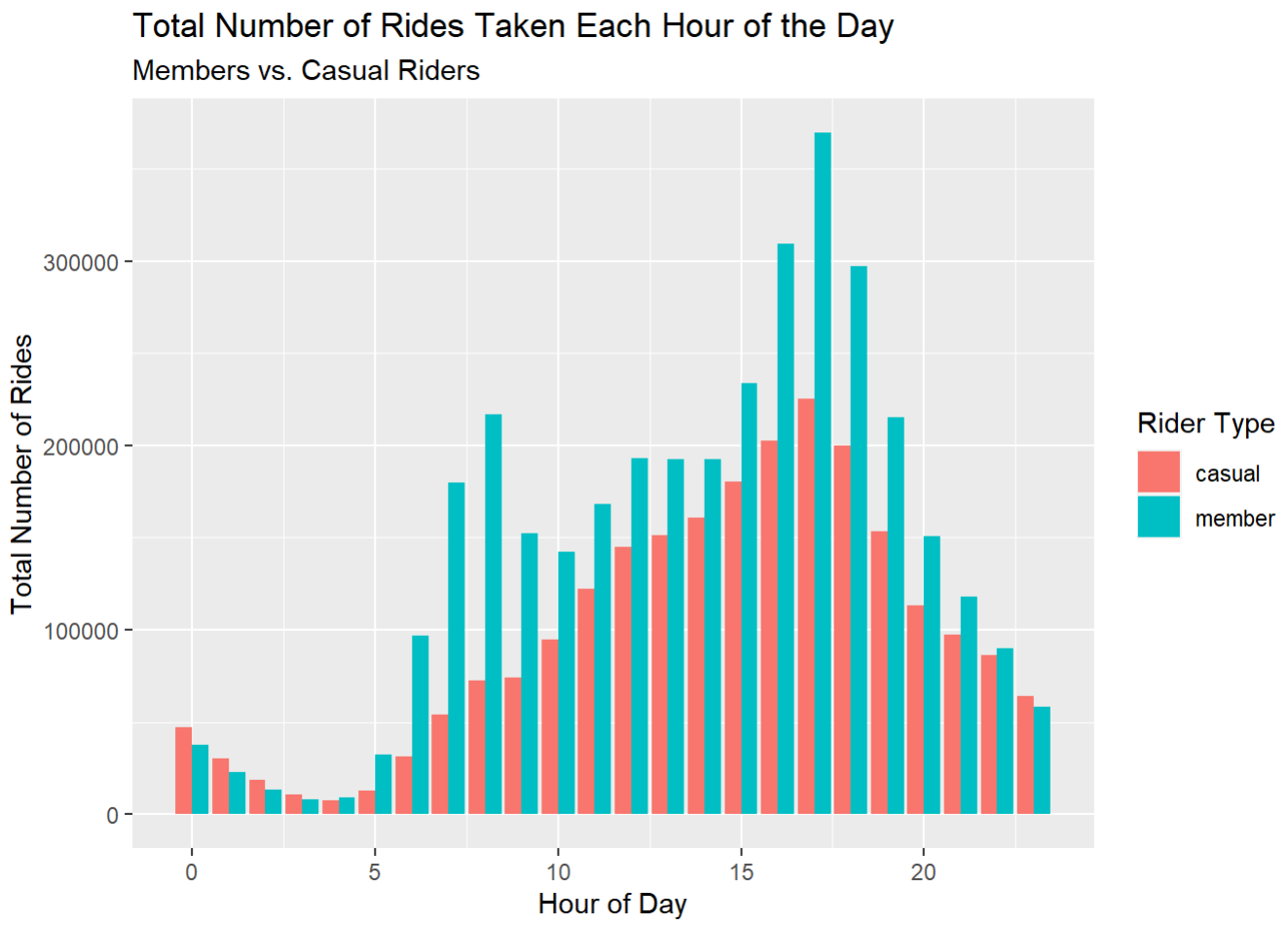
Casual riders use Cyclistic bikes more often on the weekends while members take more rides on weekdays. The duration of member's rides stays fairly consistent throughout the week with a slight increase in duration on the weekends. Casual riders take significantly longer rides on the weekends with their shortest rides being taken on Wednesdays.

3. Comparison of start times between members and casual riders

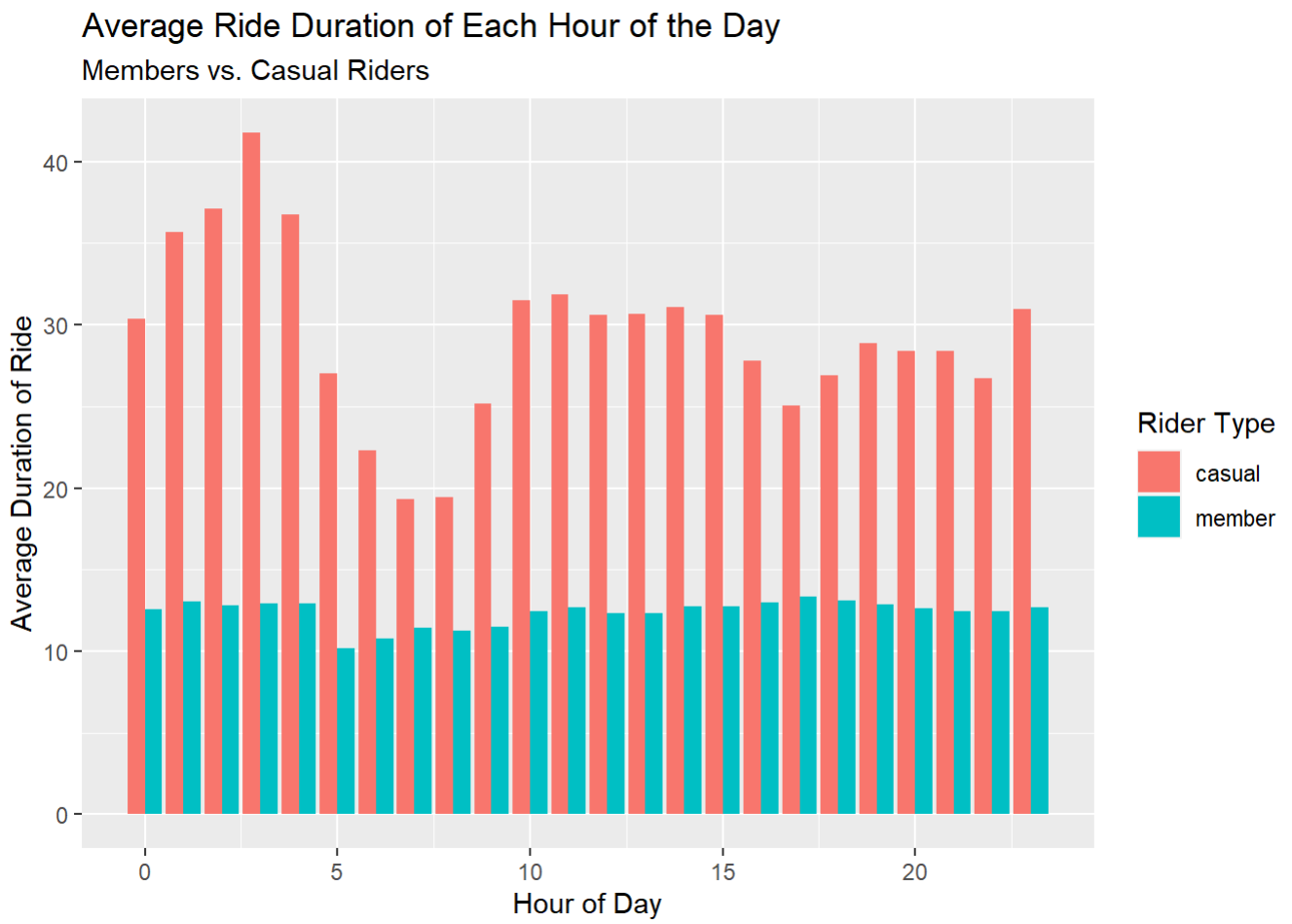
```
all_trips_v2 %>%
  group_by(member_casual,hour) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, hour)
```

```
## # A tibble: 48 × 4
## # Groups:   member_casual [2]
##    member_casual  hour number_of_rides average_duration
##    <chr>         <dbl>           <int>            <dbl>
##  1 casual            0           47112             30.4
##  2 casual            1           30385             35.7
##  3 casual            2           18802             37.1
##  4 casual            3           11007             41.8
##  5 casual            4            7680             36.8
##  6 casual            5           12848             27.0
##  7 casual            6           31240             22.3
##  8 casual            7           54058             19.3
##  9 casual            8           72823             19.4
## 10 casual            9           74458             25.2
## # ℹ 38 more rows
```

```
options(scipen=999)
all_trips_v2 %>%
  group_by(member_casual,hour) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, hour) %>%
  ggplot(aes(x=hour, y=number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")+
  labs(x="Hour of Day", y="Total Number of Rides",
       title="Total Number of Rides Taken Each Hour of the Day",
       subtitle="Members vs. Casual Riders")+
  guides(fill=guide_legend(title="Rider Type"))
```




```
all_trips_v2 %>%
  group_by(member_casual,hour) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, hour) %>%
  ggplot(aes(x=hour, y=average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x="Hour of Day", y="Average Duration of Ride",
       title="Average Ride Duration of Each Hour of the Day",
       subtitle="Members vs. Casual Riders")+
  guides(fill=guide_legend(title="Rider Type"))
```



Casual riders take their longest rides early in the morning (3am), though this is also the time when the least number of rides are taken. Their shortest trips start at 7 or 8am. The duration is fairly consistent from 10am-midnight, with a slight dip from 4-6 pm. The number of trips taken by casual riders is lowest at 4am, then gradually increases until it peaks at 5pm, followed by a gradual decrease.

For members, the duration of trips stays fairly consistent across different times of the day, with a slight decrease in the morning hours (5-9am). The number of trips taken for members peaks in the morning (8am) and early in the evening (5pm).

This suggests many members are using the bikes to commute while casual riders may take the bikes for leisure rides mostly in the evening.

4. Comparison of start station and end station usage between members and casual riders

Remove all trips that don't have either a start or end station name.

```
all_trips_v3 <- all_trips_v2[!(is.na(all_trips_v2$start_station_name) | is.na(all_trips_v2$end_station_name)), ]
```

```
nrow(subset(all_trips_v3,member_casual=="member"))
```

```
## [1] 2743031
```

```
nrow(subset(all_trips_v3,member_casual=="casual"))
```

```
## [1] 1791144
```

Determine the most frequently used start and end stations grouped by rider type:

```
all_trips_v3 %>%
  group_by(member_casual, start_station_name) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_trips=n()) %>%
  arrange(desc(number_of_trips))
```

```
## # A tibble: 1,431 × 3
## # Groups:   member_casual [1]
##    member_casual start_station_name           number_of_trips
##    <chr>         <chr>                                  <int>
##  1 member        Kingsbury St & Kinzie St               23814
##  2 member        Clark St & Elm St                      21656
##  3 member        Clinton St & Washington Blvd           20489
##  4 member        Wells St & Concord Ln                  20226
##  5 member        Loomis St & Lexington St               19832
##  6 member        University Ave & 57th St               19663
##  7 member        Ellis Ave & 60th St                    19517
##  8 member        Clinton St & Madison St                18338
##  9 member        Wells St & Elm St                      18221
## 10 member        Broadway & Barry Ave                   16817
## # ℹ 1,421 more rows
```

```
all_trips_v3 %>%
  group_by(member_casual, start_station_name) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_trips=n()) %>%
  arrange(desc(number_of_trips))
```

```
## # A tibble: 1,511 × 3
## # Groups:   member_casual [1]
##    member_casual start_station_name                number_of_trips
##    <chr>         <chr>                                       <int>
##  1 casual        Streeter Dr & Grand Ave                     54340
##  2 casual        DuSable Lake Shore Dr & Monroe St           30409
##  3 casual        Michigan Ave & Oak St                       23851
##  4 casual        Millennium Park                             23723
##  5 casual        DuSable Lake Shore Dr & North Blvd          22156
##  6 casual        Shedd Aquarium                              19567
##  7 casual        Theater on the Lake                         17321
##  8 casual        Wells St & Concord Ln                       15077
##  9 casual        Dusable Harbor                              13385
## 10 casual        Indiana Ave & Roosevelt Rd                  12913
## # ℹ 1,501 more rows
```

Casual riders use less stations than members. The most popular station makes up 3.03% (54,340/1,791,144) of the total number of trips (using v3 data) of casual riders. The top 10 most popular stations make up 12.99% (232,742/1,791,144) of the total number of trips.

In comparison, the most popular station for members makes up 0.87% (23,814/2,743,031) of the total number of trips taken by members. The top 10 most popular stations make up 7.24% (23,814/2,743,031) of the total number of trips.

Please note these numbers exclude trips where the start station or end station name was missing.

Determine the most common trips members and casual riders are taking:

```
all_trips_v3 %>%
  group_by(start_end) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
## # A tibble: 144,280 × 3
##    start_end                                    number_of_rides average_duration
##    <chr>                                                  <int>            <dbl>
##  1 Ellis Ave & 60th St , University Ave & 57th…            6548             4.41
##  2 University Ave & 57th St , Ellis Ave & 60th…            6179             4.58
##  3 Ellis Ave & 60th St , Ellis Ave & 55th St                5953             4.88
##  4 Ellis Ave & 55th St , Ellis Ave & 60th St                5531             5.14
##  5 State St & 33rd St , Calumet Ave & 33rd St               3980             4.32
##  6 Calumet Ave & 33rd St , State St & 33rd St               3838             4.03
##  7 Loomis St & Lexington St , Morgan St & Polk…            3515             4.92
##  8 Morgan St & Polk St , Loomis St & Lexington…            3342             5.20
##  9 University Ave & 57th St , Kimbark Ave & 53…            2696             7.08
## 10 Kimbark Ave & 53rd St , University Ave & 57…            2409             6.78
## # ℹ 144,270 more rows
```

```
all_trips_v3 %>%
  group_by(start_end) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
## # A tibble: 133,576 × 3
##    start_end                                      number_of_rides average_duration
##    <chr>                                                    <int>            <dbl>
##  1 Streeter Dr & Grand Ave , Streeter Dr & Gra…           10478             40.0
##  2 DuSable Lake Shore Dr & Monroe St , DuSable…            6618             33.5
##  3 DuSable Lake Shore Dr & Monroe St , Streete…            5105             27.3
##  4 Michigan Ave & Oak St , Michigan Ave & Oak …            4643             44.9
##  5 Millennium Park , Millennium Park                       4031             37.7
##  6 Montrose Harbor , Montrose Harbor                       2982             50.1
##  7 Streeter Dr & Grand Ave , DuSable Lake Shor…            2772             27.7
##  8 Streeter Dr & Grand Ave , Millennium Park               2684             33.6
##  9 Shedd Aquarium , Shedd Aquarium                         2570             22.4
## 10 DuSable Lake Shore Dr & North Blvd , DuSabl…            2451             36.6
## # ℹ 133,566 more rows
```

The top 10 trip routes make up 1.60% (43,991/2,743,031) of total member rides and 2.48% (44,334/1,791,144) of total casual rides. The top route for casual riders makes up .58% of total casual rides. The top route for members is .24% (6,548/2,743,031). Casual riders are taking the same routes more often than members.

5. Comparison of the number of rides and ride duration by day of the week for each rider type.

A. CASUAL RIDERS

Create new data frame with the total number of trips broken out by day of the week. It will contain casual rider trips only.

```
casual_start_end <- all_trips_v3 %>%
  group_by(start_end) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length),
            sat=length(which(day_of_week=="Saturday")),
            mon=length(which(day_of_week=="Monday")),
            tue=length(which(day_of_week=="Tuesday")),
            wed=length(which(day_of_week=="Wednesday")),
            thu=length(which(day_of_week=="Thursday")),
            fri=length(which(day_of_week=="Friday")),
            sun=length(which(day_of_week=="Sunday"))) %>%
  arrange(desc(number_of_rides))
```

Create data frame for the top 10 rows of the casual_start_end data frame.

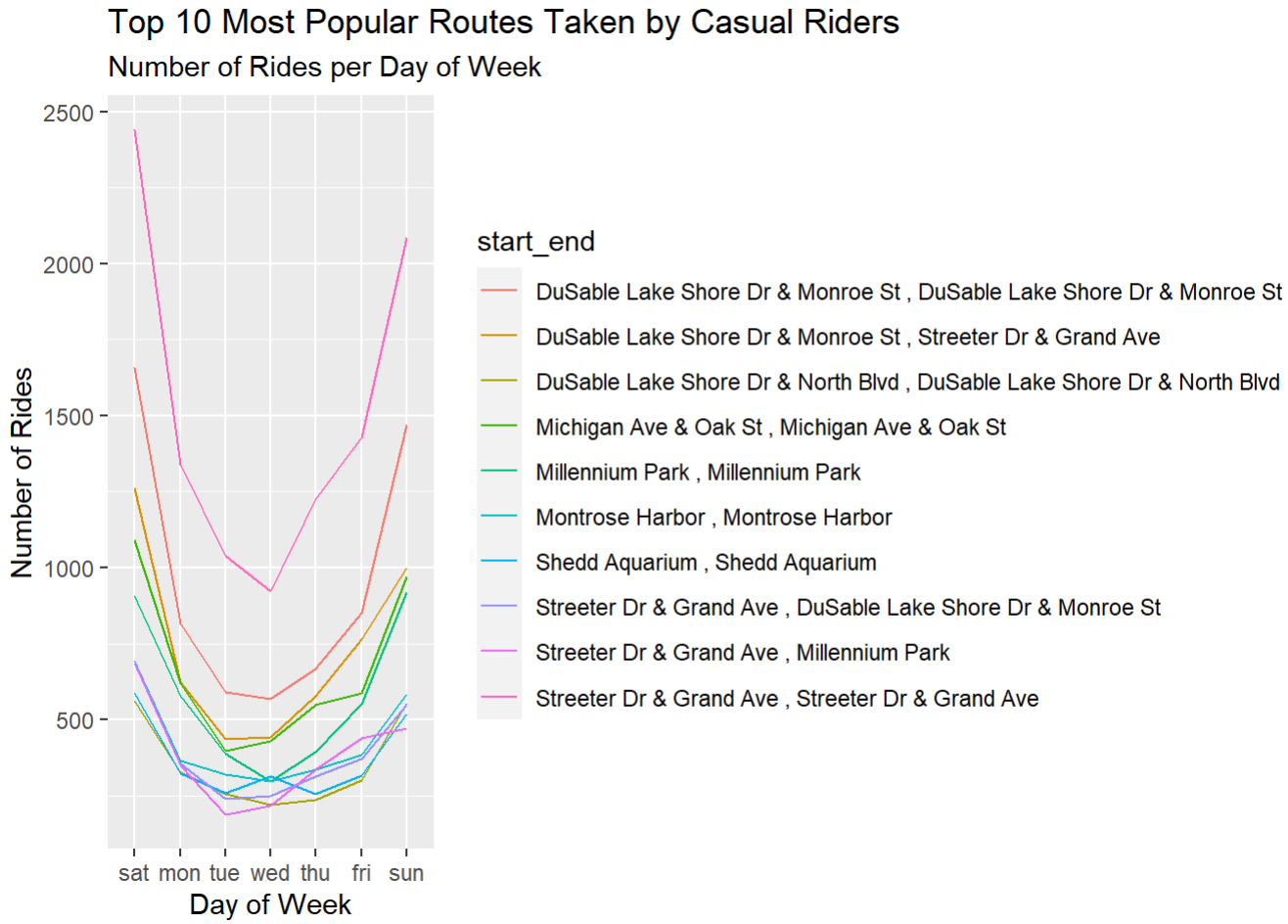
```
top10_casual_start_end <- casual_start_end %>%
  slice(1:10)
```

Transform the top 10 rows from wide to long format.

```
casual_start_end_long <- melt(top10_casual_start_end,
                              id.vars = c("start_end", "number_of_rides","average_duration"))
casual_start_end_long$average_duration=NULL
casual_start_end_long$number_of_rides=NULL
```

Graph

```
casual_start_end_long %>%
  group_by(start_end) %>%
  ggplot(aes(x=variable,y=value,group=start_end, color=start_end))+
  geom_line() +
  labs(x="Day of Week", y="Number of Rides",
       title="Top 10 Most Popular Routes Taken by Casual Riders",
       subtitle="Number of Rides per Day of Week") +
  guides(fill=guide_legend(title="Start, End Station Name"))
```



B. MEMBERS

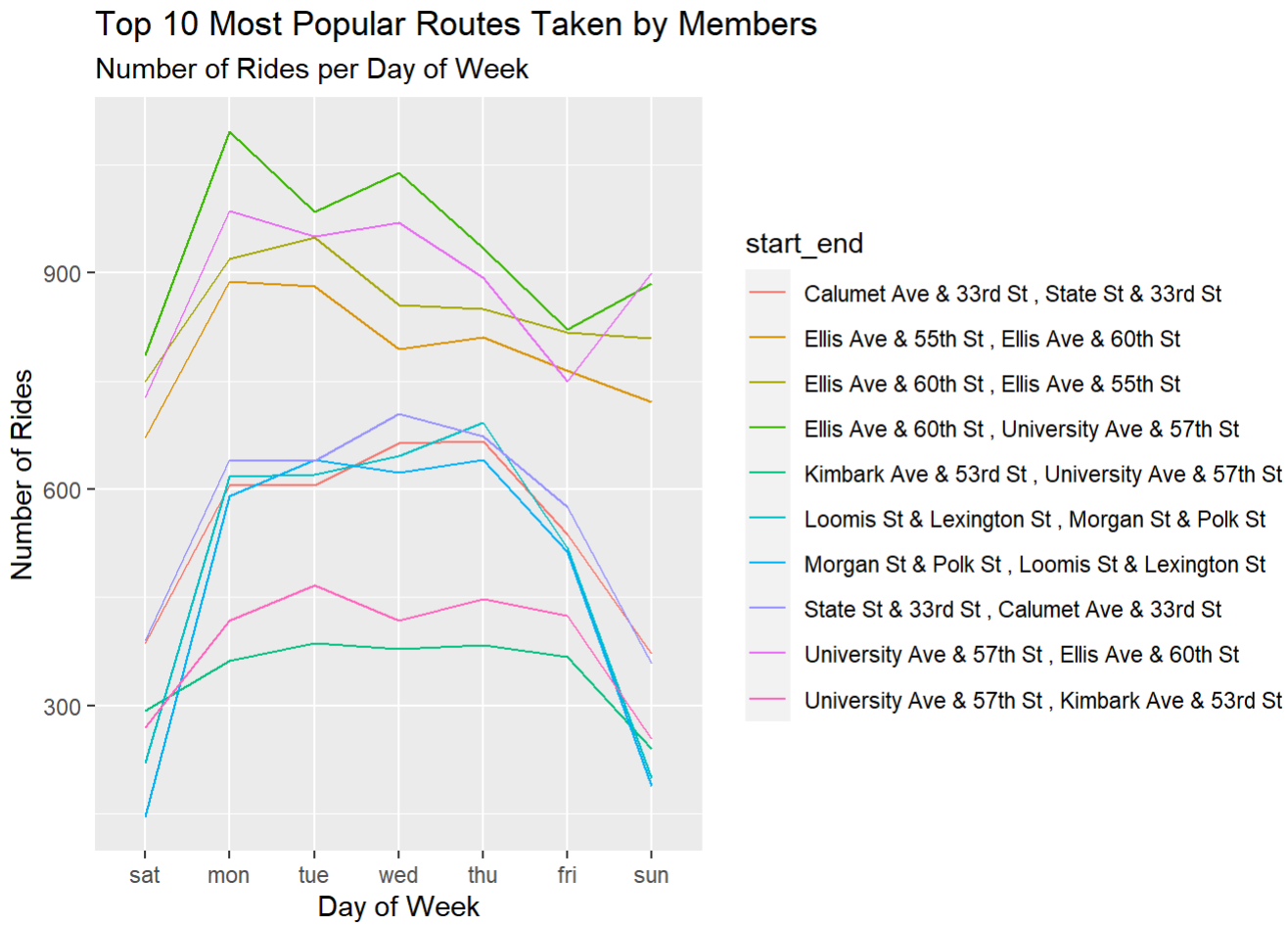
Repeat the same steps for members.

```
member_start_end<- all_trips_v3 %>%
  group_by(start_end) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length),
            sat=length(which(day_of_week=="Saturday")),
            mon=length(which(day_of_week=="Monday")),
            tue=length(which(day_of_week=="Tuesday")),
            wed=length(which(day_of_week=="Wednesday")),
            thu=length(which(day_of_week=="Thursday")),
            fri=length(which(day_of_week=="Friday")),
            sun=length(which(day_of_week=="Sunday"))) %>%
  arrange(desc(number_of_rides))
```

```
top10_member_start_end <- member_start_end %>%
  slice(1:10)
```

```
member_start_end_long <- melt(top10_member_start_end,
                              id.vars = c("start_end", "number_of_rides","average_duration"))
member_start_end_long$average_duration=NULL
member_start_end_long$number_of_rides=NULL
```

```
member_start_end_long %>%
  group_by(start_end) %>%
  ggplot(aes(x=variable,y=value,group=start_end, color=start_end))+
  geom_line()+
  labs(x="Day of Week", y="Number of Rides",
       title="Top 10 Most Popular Routes Taken by Members",
       subtitle="Number of Rides per Day of Week") +
  guides(fill=guide_legend(title="Start, End Station Name"))
```



At the top 10 most popular stations, casual riders mostly ride on the weekends while members mostly ride on weekdays.

```
summary(member_start_end$average_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.64   19.12   21.95   27.64 1455.88
```

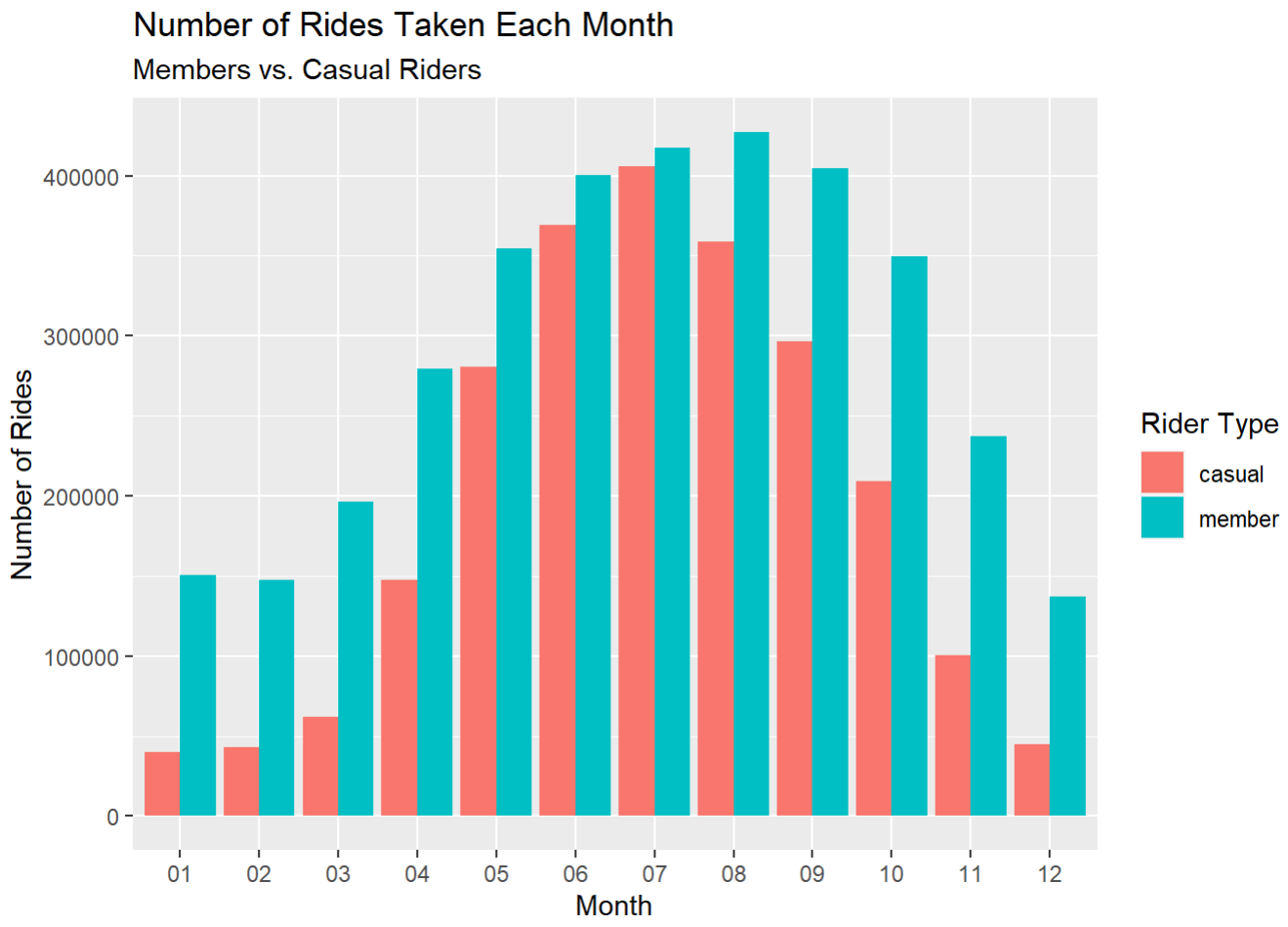
```
summary(casual_start_end$average_duration)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     0.017    14.533    21.942    30.171    33.726 10722.967
```

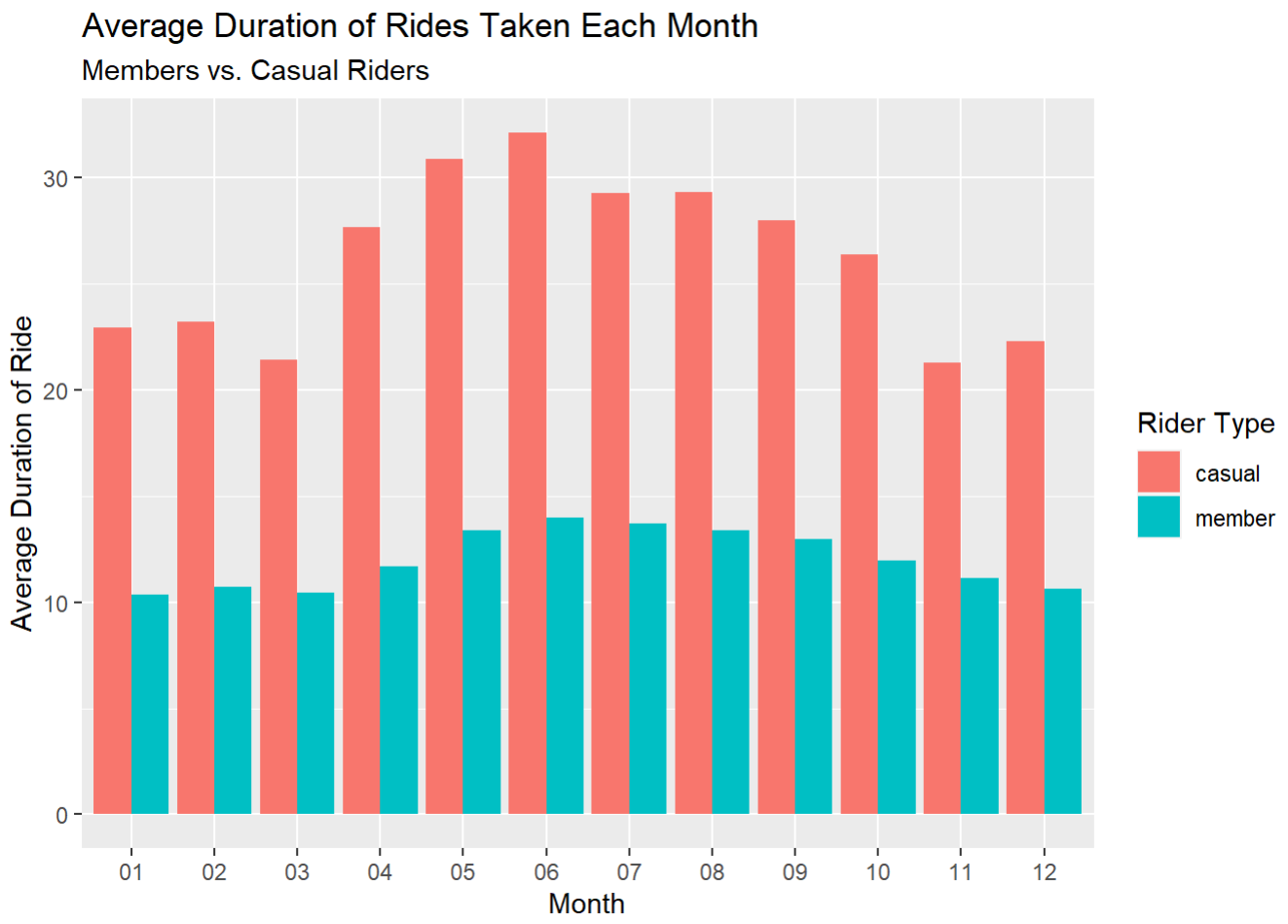
As seen previously, member's rides are, on average, shorter in duration than casual riders.

6. Comparison of rides taken each month between members and casual riders

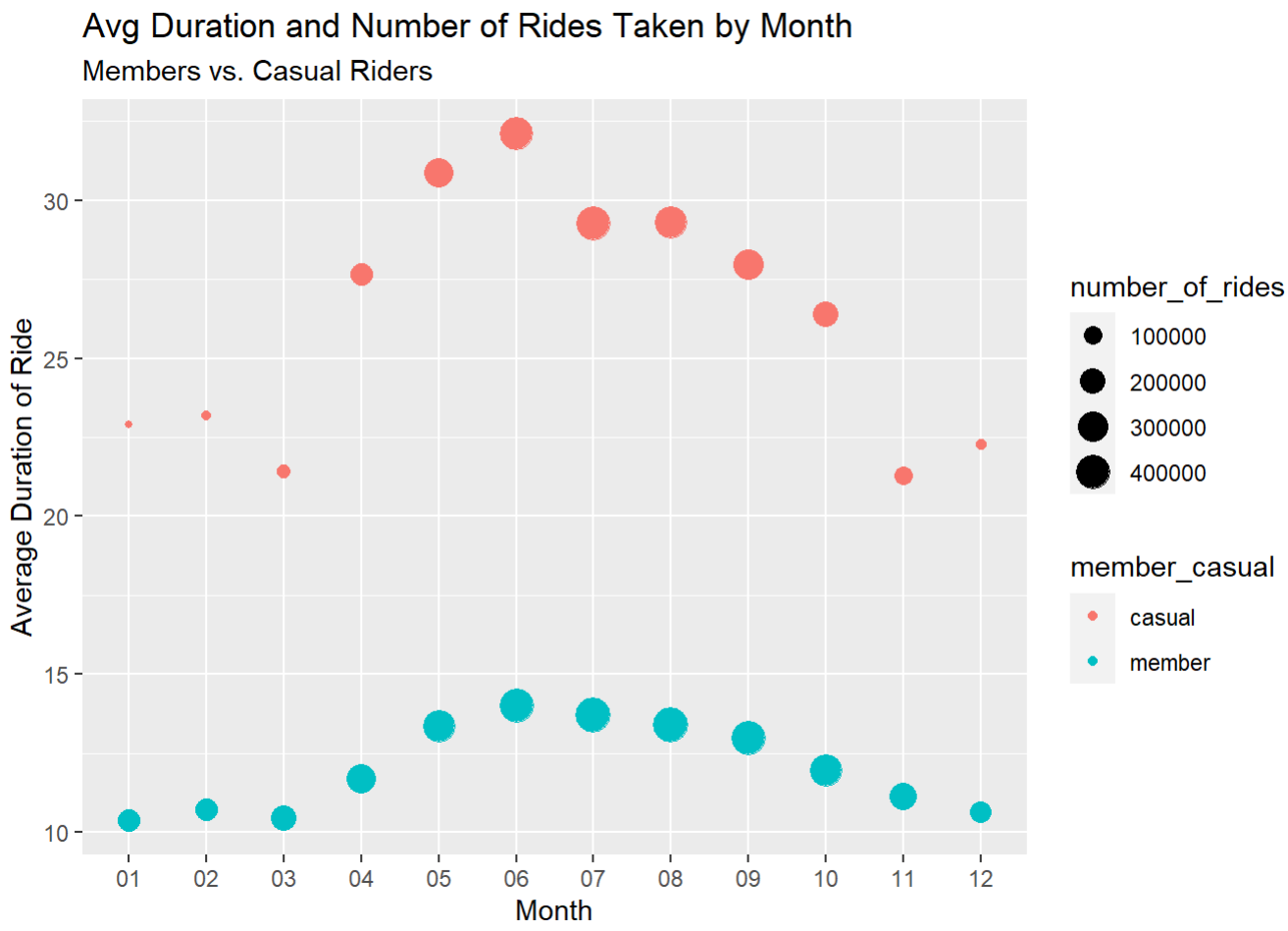
```
all_trips_v2 %>%
  group_by(member_casual,month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x="Month", y="Number of Rides",
       title="Number of Rides Taken Each Month",
       subtitle="Members vs. Casual Riders") +
  guides(fill=guide_legend(title="Rider Type"))
```



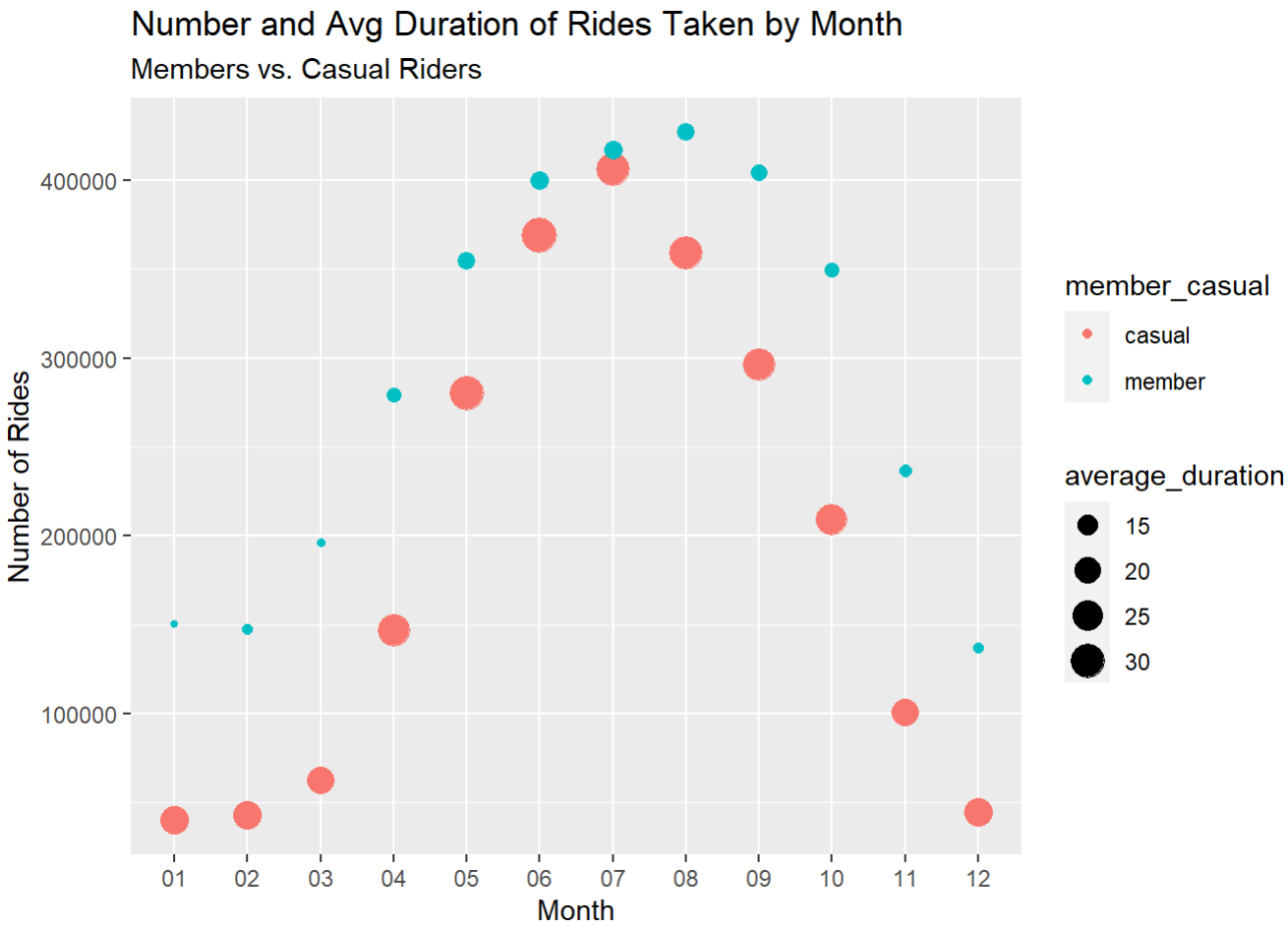

```
all_trips_v2 %>%
  group_by(member_casual,month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x="Month", y="Average Duration of Ride",
       title="Average Duration of Rides Taken Each Month",
       subtitle="Members vs. Casual Riders") +
  guides(fill=guide_legend(title="Rider Type"))
```




```
all_trips_v2 %>%
  group_by(member_casual,month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=average_duration, color = member_casual, size=number_of_rides)) +
  geom_point()+
  labs(x="Month", y="Average Duration of Ride",
       title="Avg Duration and Number of Rides Taken by Month",
       subtitle="Members vs. Casual Riders")
```




```
all_trips_v2 %>%
  group_by(member_casual,month) %>%
  summarise(number_of_rides = n()
            ,average_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x=month, y=number_of_rides, color = member_casual, size=average_duration)) +
  geom_point()+
  labs(x="Month", y="Number of Rides",
       title="Number and Avg Duration of Rides Taken by Month",
       subtitle="Members vs. Casual Riders")
```



The duration of rides each month for casual riders is significantly greater than members. For casual riders and members, the duration of rides is higher in the summer months, though this is more significant for casual riders.

For casual riders and members, there are a greater number of rides taken during the summer months than in the winter months, though the difference is more pronounced for casual riders. The number of rides taken by members and casual riders is closest during the month of July. Members are taking more rides than casual riders during the winter months.

This suggests that members are using the bikes year-round as a means of transportation while casual riders may use the bikes more for leisure during the summer months.

7. Average Duration vs. Number of Rides for Members and Casual Riders.

Used data within two standard deviations of the mean to be able to see trends more clearly.

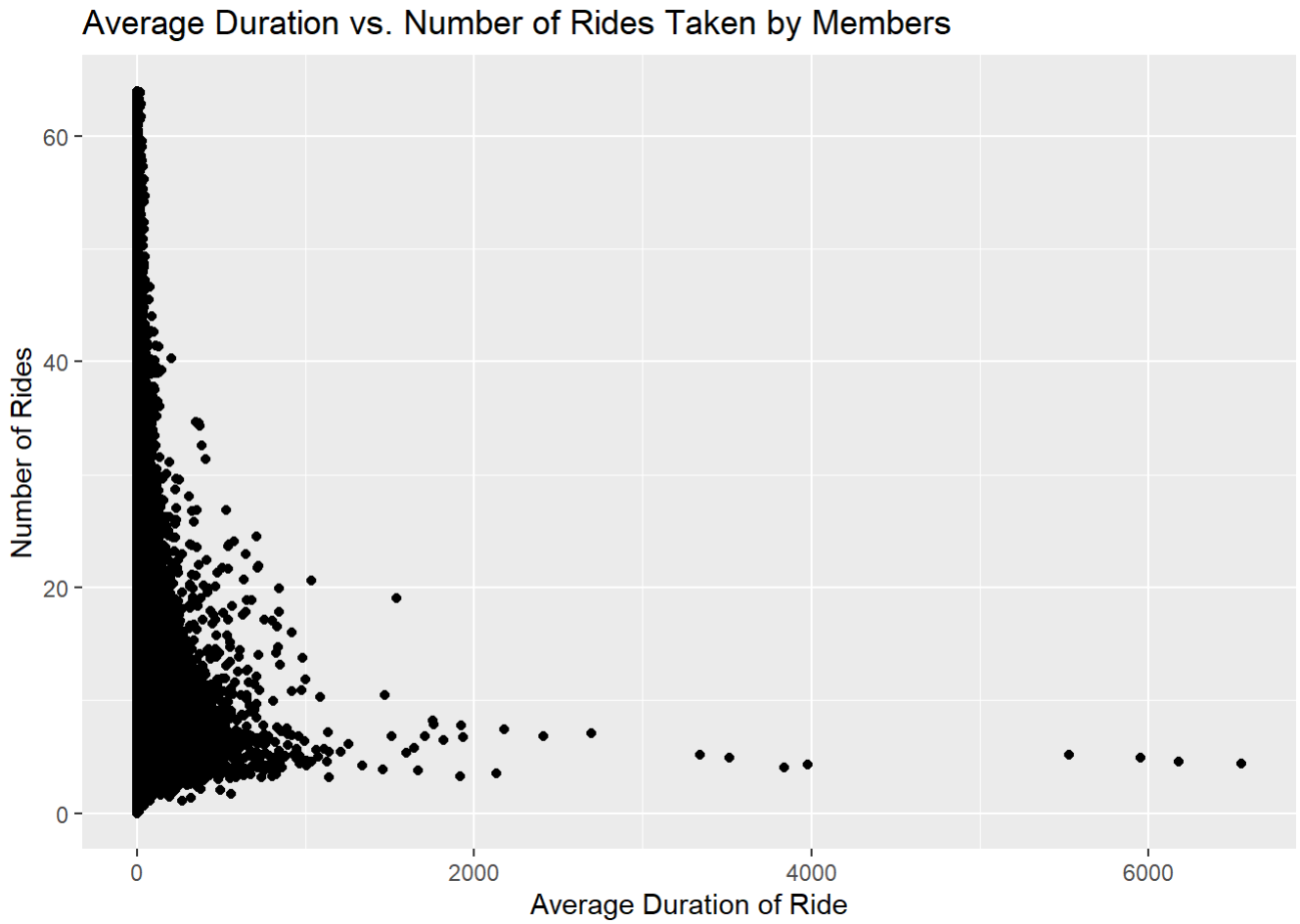
```
mean(member_start_end$average_duration)
```

```
## [1] 21.94725
```

```
sd(member_start_end$average_duration)
```

```
## [1] 20.32174
```

```
member_start_end %>%
  filter(average_duration<63.99886) %>%
  ggplot(aes(x=number_of_rides, y=average_duration))+
          geom_point() +
  labs(x="Average Duration of Ride", y="Number of Rides",
       title="Average Duration vs. Number of Rides Taken by Members")
```



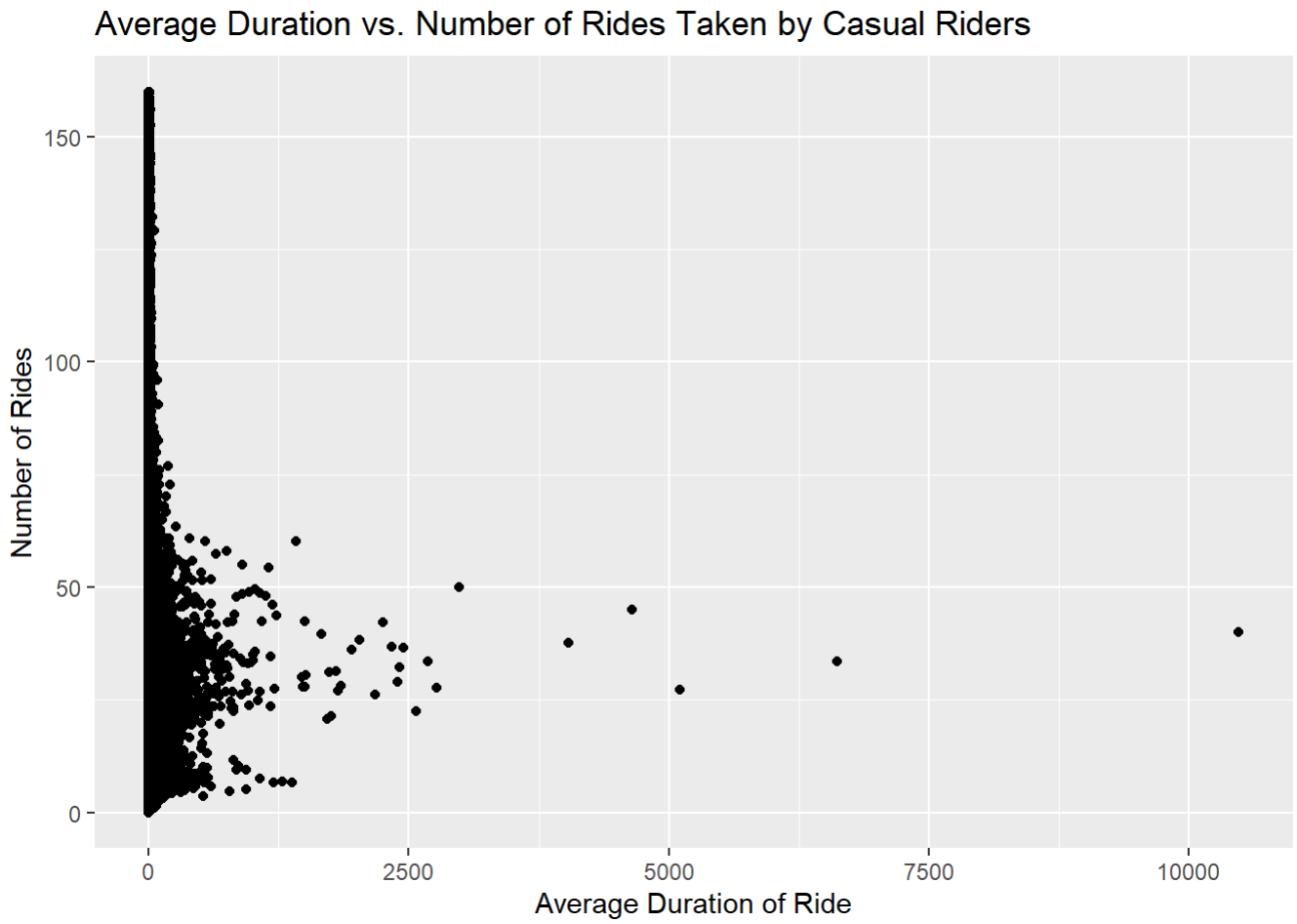

```
mean(casual_start_end$average_duration)
```

```
## [1] 30.17134
```

```
sd(casual_start_end$average_duration)
```

```
## [1] 54.47108
```

```
casual_start_end %>%
  filter(average_duration<160.05513) %>%
  ggplot(aes(x=number_of_rides, y=average_duration))+
  geom_point() +
  labs(x="Average Duration of Ride", y="Number of Rides",
       title="Average Duration vs. Number of Rides Taken by Casual Riders")
```



Each point represents a route (start station/end station combination). This is another way to graphically show insights seen previously. Members' rides are shorter on average. There are fewer popular routes for casual riders and they are more frequently used compared to the most popular routes taken by members. One route in particular was taken over 10,000 times by casual riders versus the most popular member route being taken only approximately 6,500 times.

Next, to visualize where the most popular start and end stations are located, the latitude and longitudes provided for the stations will be used to map out where they are in Chicago. The map will be compared to Google Maps to determine whether there are any landmarks which may provide greater context for why these locations are popular for members and casual riders.

8. Visualizing the Data on a Map of Chicago

A. MEMBERS

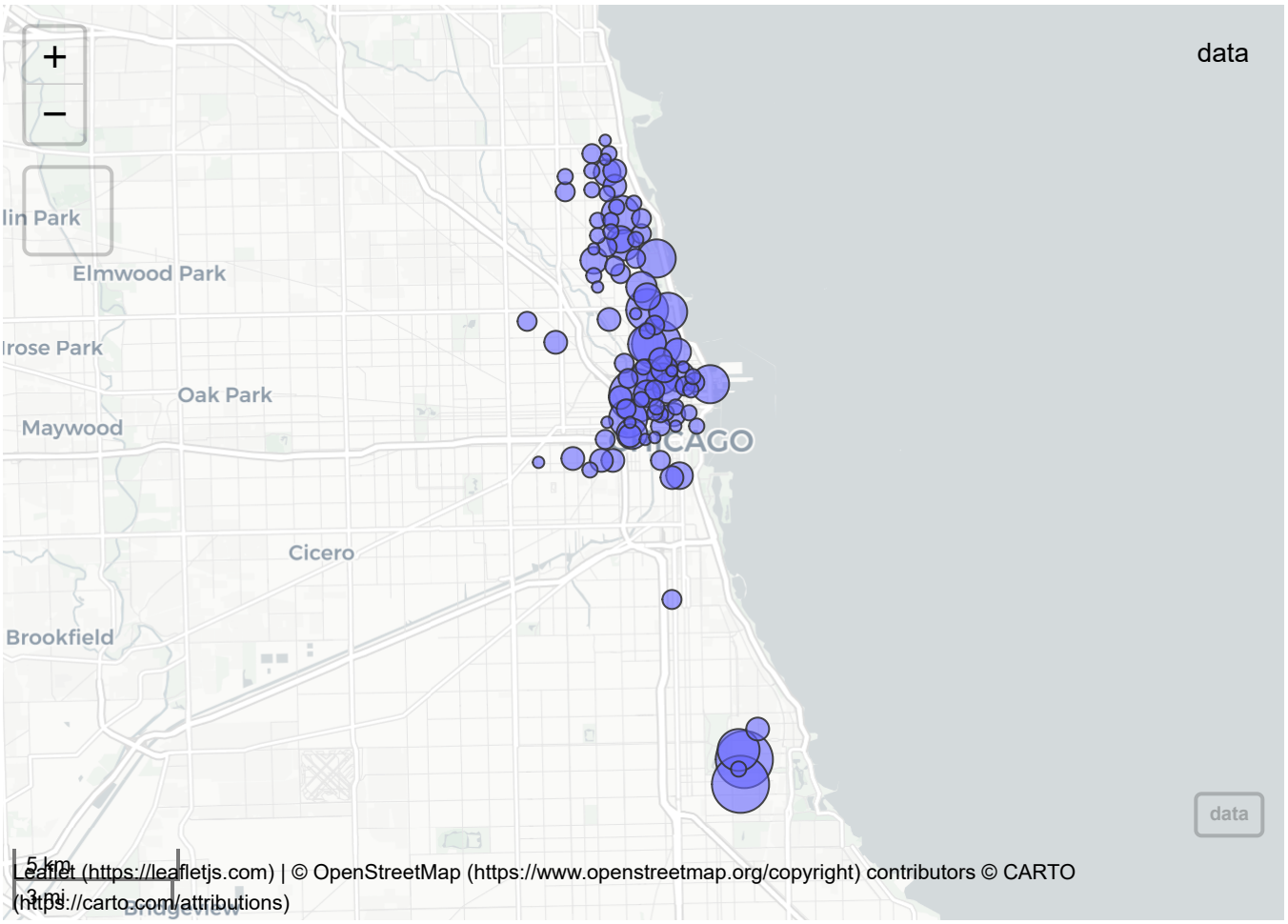
Most popular start locations.

```
map_member_start<-all_trips_v3 %>%
  group_by(start_station_name, start_lat,start_lng) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
FALSE `summarise()` has grouped output by 'start_station_name', 'start_lat'. You can
FALSE override using the `.groups` argument.
```

```
top100_map_member_start <- map_member_start[1:100, ]
```

```
mapview(top100_map_member_start,xcol="start_lng",ycol="start_lat", crs=4296, grid=FALSE, cex="nu
mber_of_rides")
```



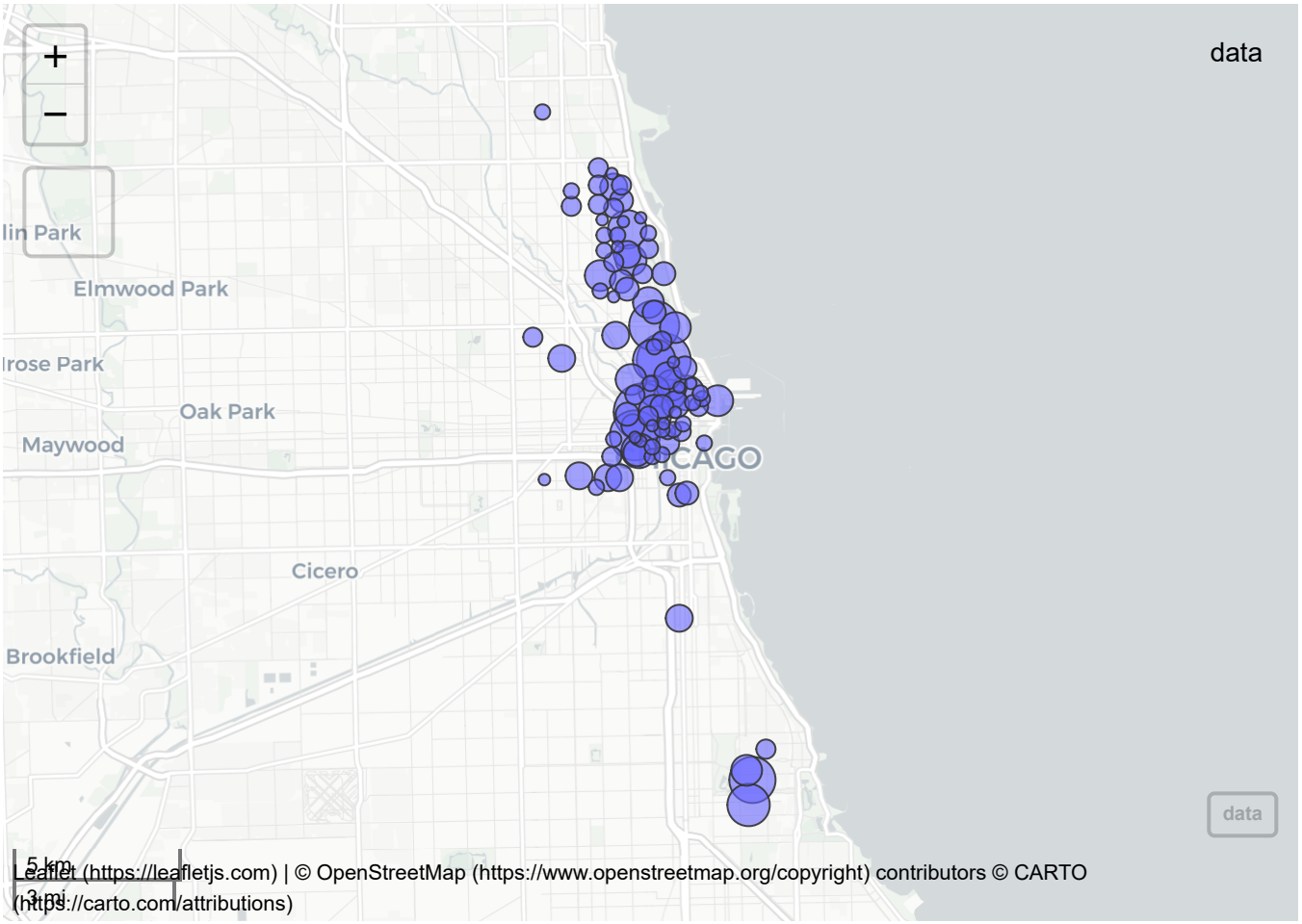
Most popular end locations.

```
map_member_end <-all_trips_v3 %>%
  group_by(end_station_name,end_lat,end_lng) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
FALSE `summarise()` has grouped output by 'end_station_name', 'end_lat'. You can
FALSE override using the `.groups` argument.
```

```
top100_map_member_end <- map_member_end[1:100, ]
```

```
mapview(top100_map_member_end,xcol="end_lng",ycol="end_lat",crs=4296, grid=FALSE,cex="number_of_
rides")
```



Most popular routes.

```
map_member_start_end<- all_trips_v3 %>%
  group_by(start_end,start_lat,start_lng,end_lat, end_lng) %>%
  filter(member_casual=="member") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

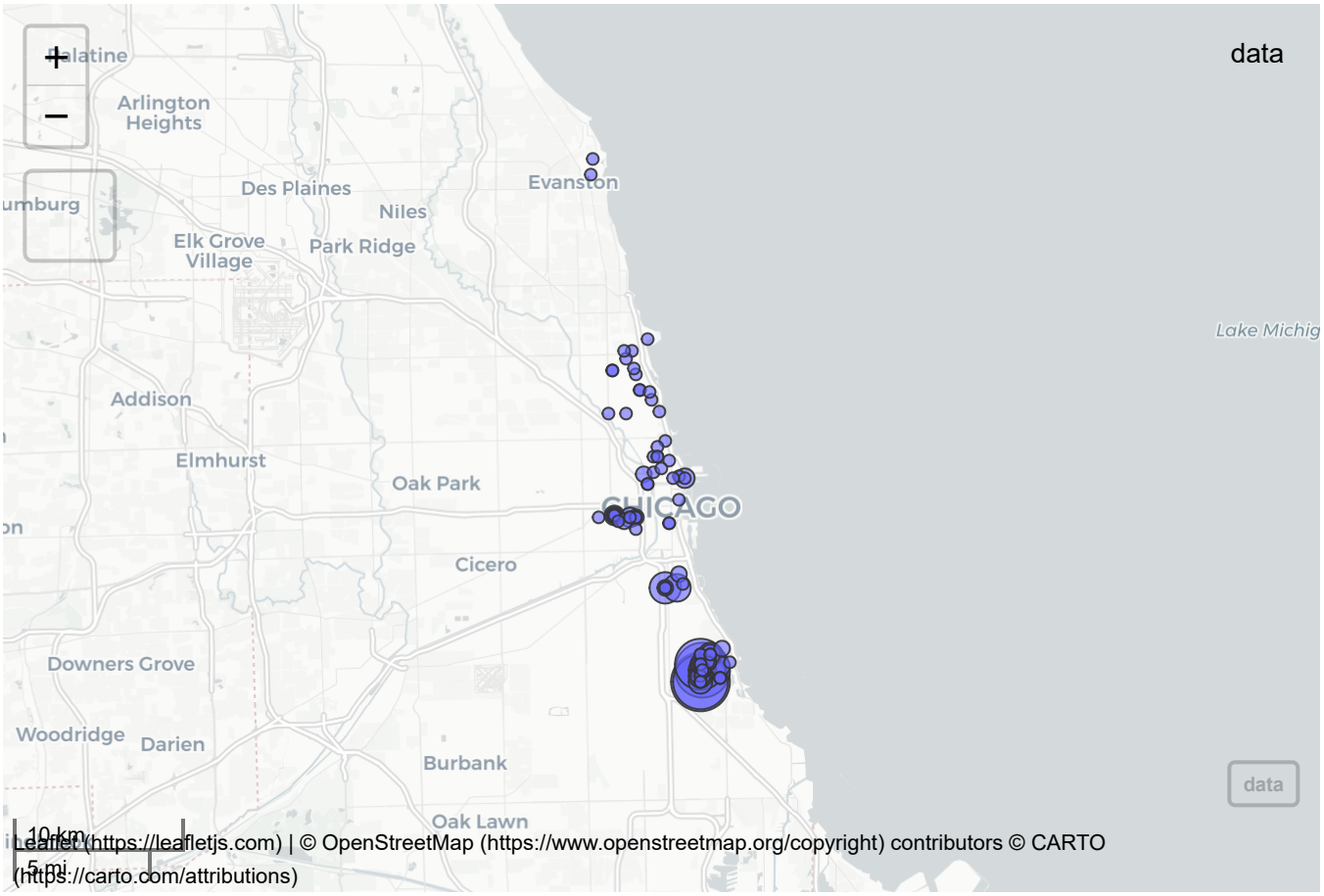
```
FALSE `summarise()` has grouped output by 'start_end', 'start_lat', 'start_lng',
FALSE 'end_lat'. You can override using the `.groups` argument.
```

```
top100_map_member_start_end <- map_member_start_end[1:100, ]
```

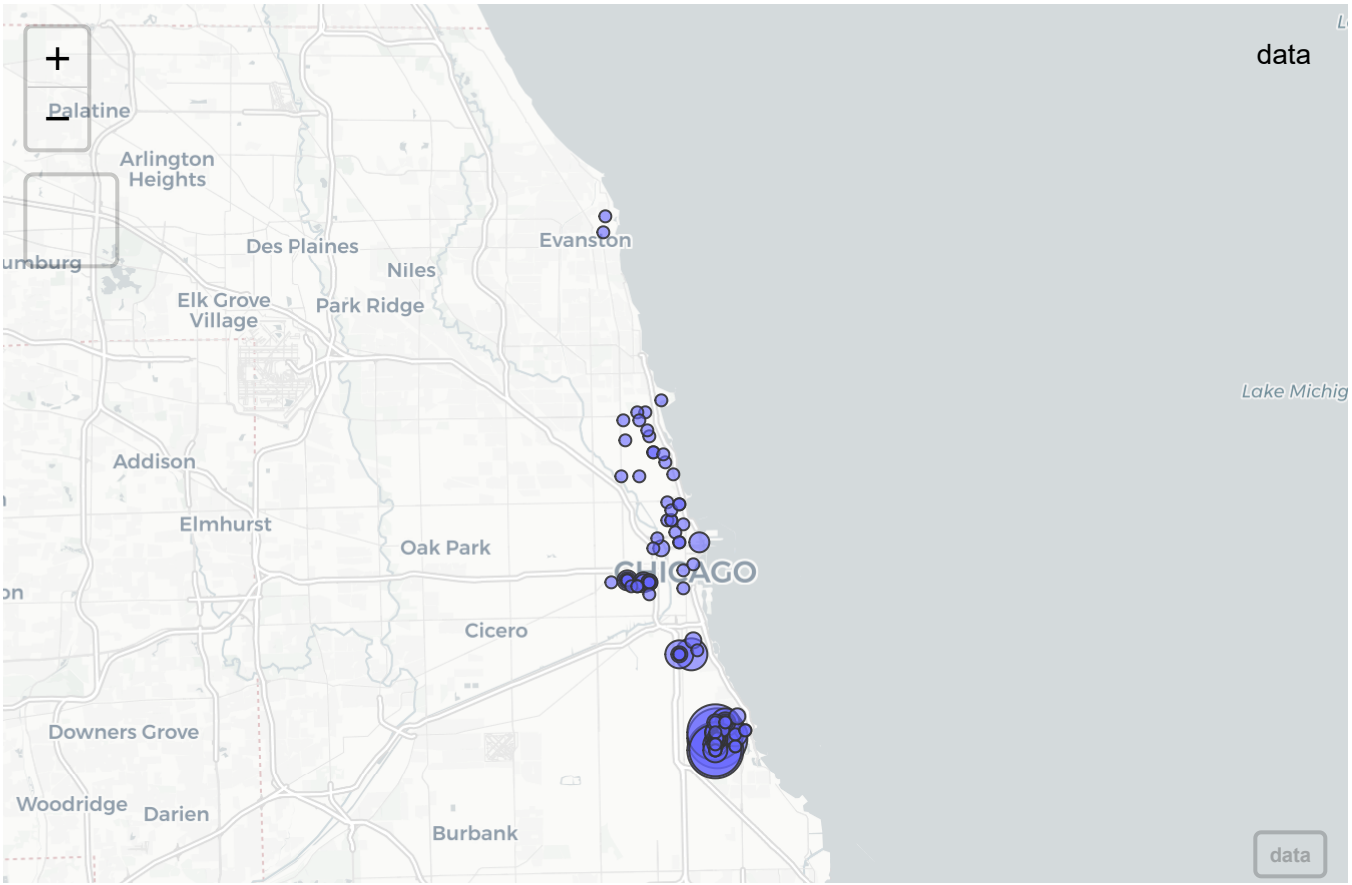
```
sum(top100_map_member_start_end$number_of_rides)/sum(map_member_start_end$number_of_rides)
```

```
## [1] 0.03585741
```

```
mapview(top100_map_member_start_end,xcol="start_lng", ycol="start_lat", crs=4296, grid=FALSE, ce
x="number_of_rides")
```




```
mapview(top100_map_member_start_end,xcol="end_lng", ycol="end_lat", crs=4296, grid=FALSE, cex="number_of_rides")
```



B. CASUAL RIDERS

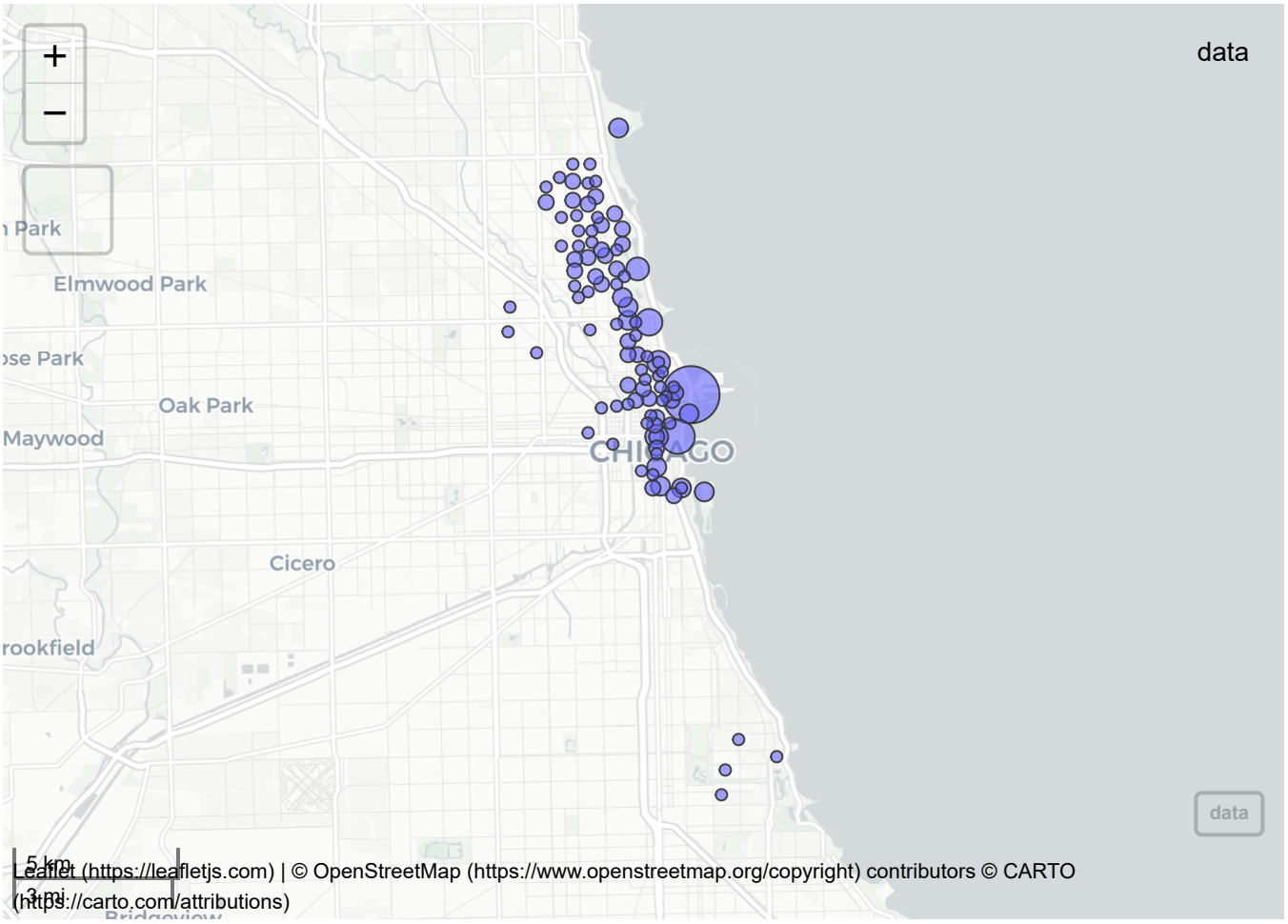
Most popular start locations.

```
map_casual_start<-all_trips_v3 %>%
  group_by(start_station_name, start_lat,start_lng) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
FALSE `summarise()` has grouped output by 'start_station_name', 'start_lat'. You can
FALSE override using the `.groups` argument.
```

```
top100_map_casual_start <- map_casual_start[1:100, ]
```

```
mapview(top100_map_casual_start,xcol="start_lng",ycol="start_lat", crs=4296, grid=FALSE, cex="nu
mber_of_rides")
```



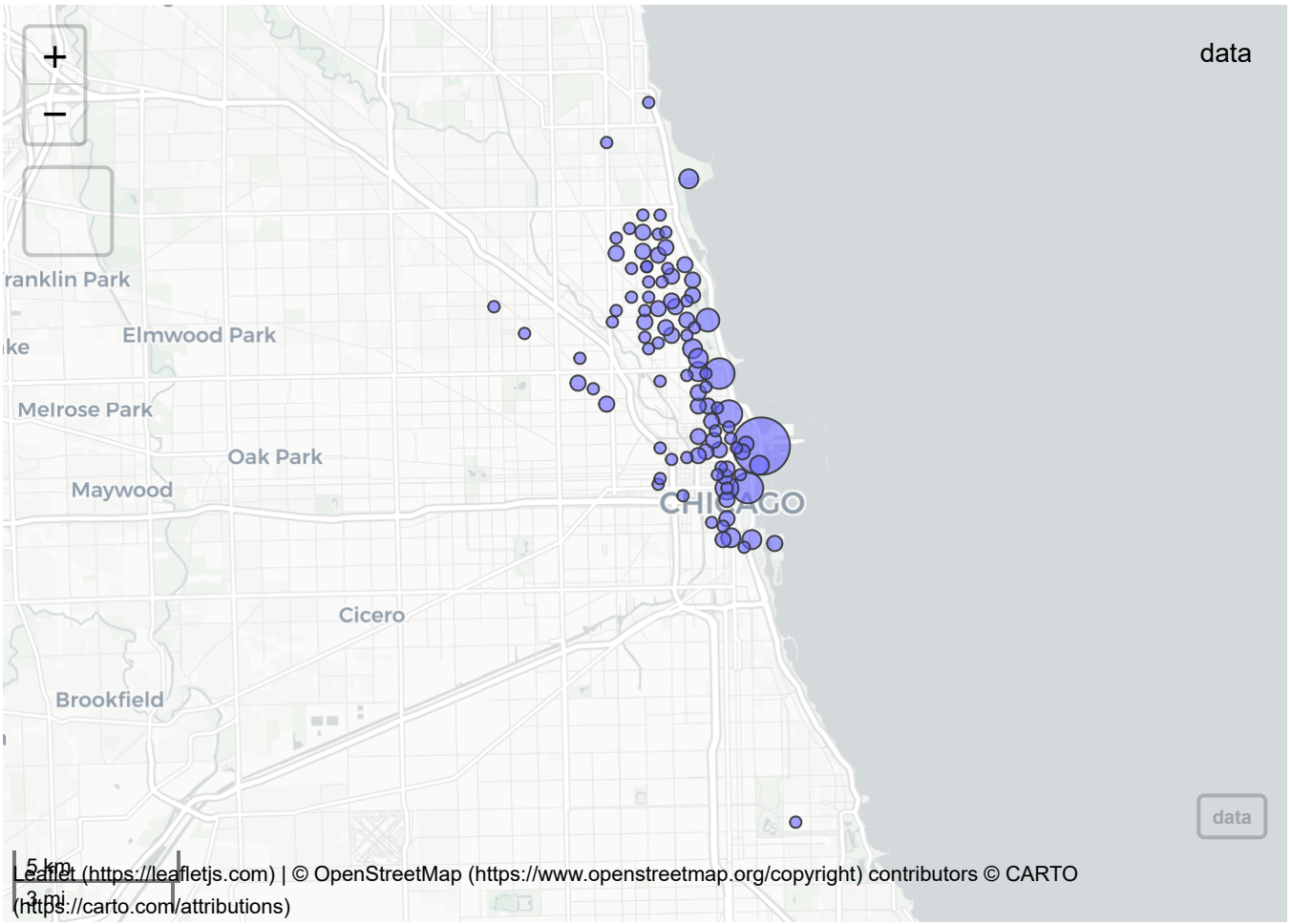
Most popular end locations.

```
map_casual_end <-all_trips_v3 %>%
  group_by(end_station_name,end_lat,end_lng) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

```
FALSE `summarise()` has grouped output by 'end_station_name', 'end_lat'. You can
FALSE override using the `.groups` argument.
```

```
top100_map_casual_end <- map_casual_end[1:100, ]
```

```
mapview(top100_map_casual_end,xcol="end_lng",ycol="end_lat",crs=4296, grid=FALSE,cex="number_of_
rides")
```



Most popular routes.

```
map_casual_start_end<- all_trips_v3 %>%
  group_by(start_end,start_lat,start_lng,end_lat, end_lng) %>%
  filter(member_casual=="casual") %>%
  summarize(number_of_rides=n(),
            average_duration=mean(ride_length)) %>%
  arrange(desc(number_of_rides))
```

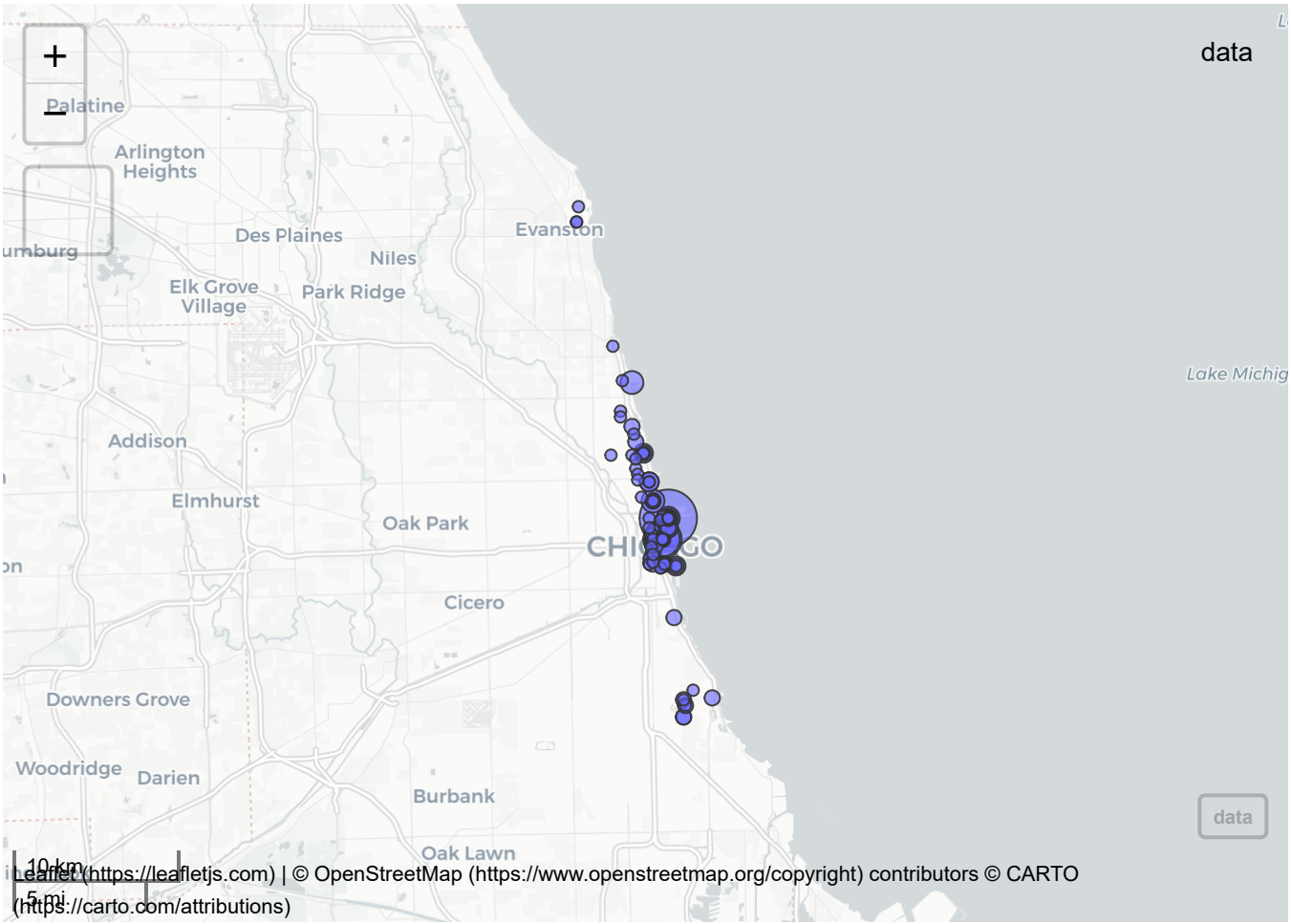
```
FALSE `summarise()` has grouped output by 'start_end', 'start_lat', 'start_lng',
FALSE 'end_lat'. You can override using the `.groups` argument.
```

```
top100_map_casual_start_end <- map_casual_start_end[1:100, ]
```

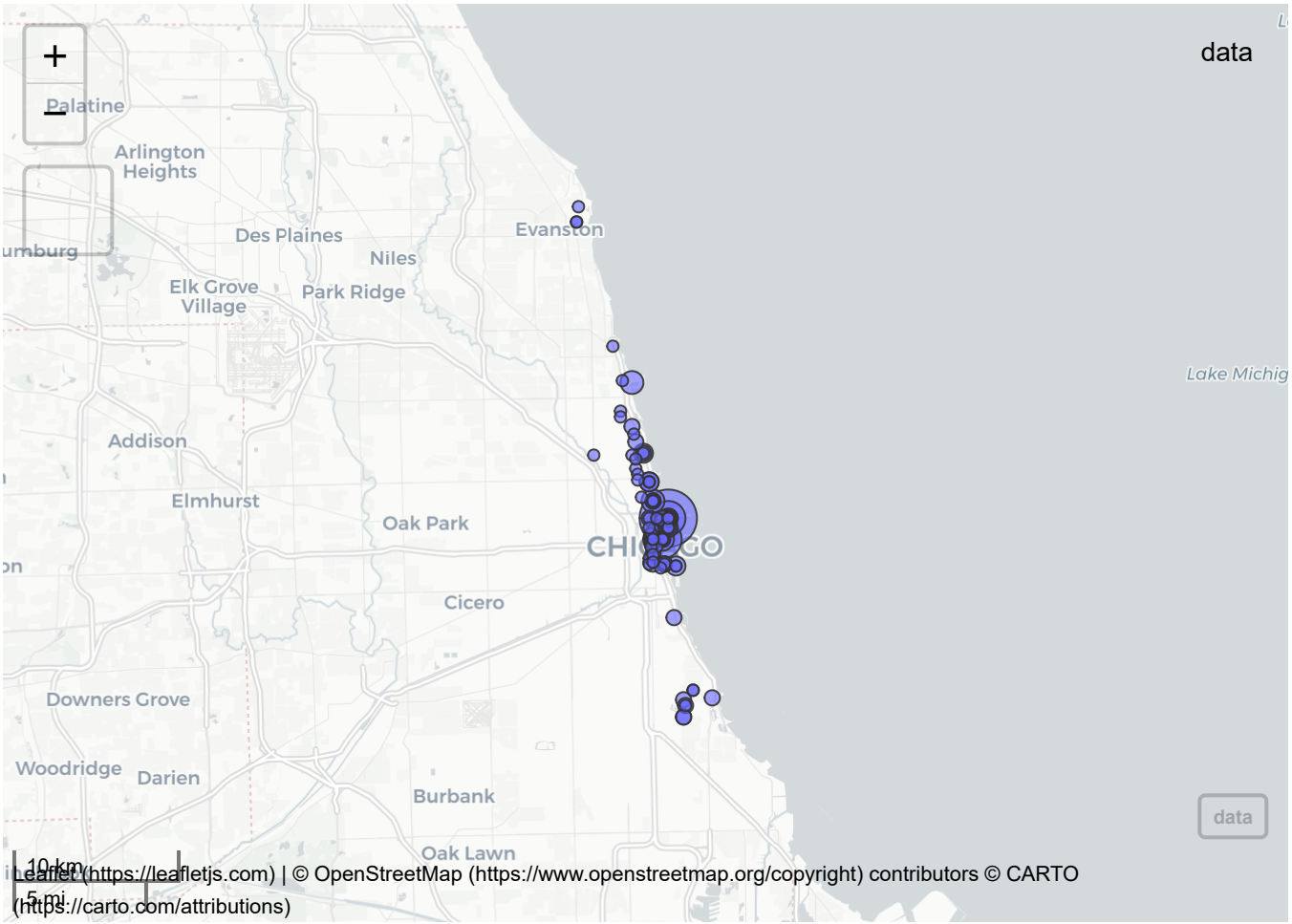
```
sum(top100_map_casual_start_end$number_of_rides)/sum(map_casual_start_end$number_of_rides)
```

```
## [1] 0.05339325
```

```
mapview(top100_map_casual_start_end,xcol="start_lng", ycol="start_lat", crs=4296, grid=FALSE, ce
x="number_of_rides")
```



```
mapview(top100_map_casual_start_end,xcol="end_lng", ycol="end_lat", crs=4296, grid=FALSE, cex="number_of_rides")
```



When compared to Google Maps landmarks, member trips are concentrated around colleges while casual rider trips are concentrated around tourist landmarks.

This supports the hypothesis that members are primarily using the bikes as a means of reliable transportation while casual users are using them to sight-see and visit popular locations. Since many members start and end locations are near colleges, the demographic of members may be college students.