Bike-share Case Study

Timur Rakhimyanov

2022-12-28

Dataset

Data, provided for this analysis, contains 12 months of Cyclistic's users bike-sharing history.
Data has been made available under this license.

Questions

Questions, that were asked for this analysis are:
1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Preparation and procession of data

Loading libraries and reading data

To begin with, I'll load the libraries I'm going to use.

```
library(tidyverse)
library(janitor)
library(lubridate)
library(pivottabler)
```

Then let's read datasets.

```
bike_trips_datasets = list.files(path = "datasets",
                                 recursive = TRUE,
                                 pattern = ".csv",
                                 full.names = TRUE)
bike_trips_dataframe <- readr::read_csv(bike_trips_datasets, id = "file_name")
```

Cleaning data

Now let's do standard cleaning and peak in the data.

```
clean_names(bike_trips_dataframe)
```

```
head(bike_trips_dataframe)
```

```
## # A tibble: 6 x 14
##   file_name      ride_id ridea~1 started_at          ended_at            start~2
##   <chr>          <chr>   <chr>   <dttm>              <dttm>              <chr>
## 1 datasets/2021~ 7C00A9~ electr~ 2021-11-27 13:27:38 2021-11-27 13:46:38 <NA>
## 2 datasets/2021~ 908548~ electr~ 2021-11-27 13:38:25 2021-11-27 13:56:10 <NA>
## 3 datasets/2021~ 0A7D10~ electr~ 2021-11-26 22:03:34 2021-11-26 22:05:56 <NA>
## 4 datasets/2021~ 2F3BE3~ electr~ 2021-11-27 09:56:49 2021-11-27 10:01:50 <NA>
## 5 datasets/2021~ D67B47~ electr~ 2021-11-26 19:09:28 2021-11-26 19:30:41 <NA>
## 6 datasets/2021~ 02F85C~ electr~ 2021-11-26 18:34:07 2021-11-26 18:52:49 Michig~
## # ... with 8 more variables: 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>, and abbreviated variable names
## #   1: rideable_type, 2: start_station_name
```

We can see, that there are values missing (shown as "NA"), let's check how much do we miss.

```
sapply(bike_trips_dataframe, function(x) sum(is.na(x)))
```

```
##          file_name            ride_id      rideable_type         started_at
##                  0                  0                  0                  0
##           ended_at start_station_name   start_station_id   end_station_name
##                  0             878177             878177             940010
##     end_station_id          start_lat          start_lng            end_lat
##             940010                  0                  0               5835
##            end_lng      member_casual
##               5835                  0
```

Missing data is start and end stations, so our choices are:
1. Populate the missing values with averages.
2. Delete rows with data missing.
3. Create sub-dataframes with no data missing for particular parts of analysis.
Third options is preferable, because resulting analysis would be more comprehensive.

So the plan is: we will look for insights in full dataset based on fields with no data missing, then we will create separate dataframes for stations insights.

Adding additional attributes

For now let's add durations of the trips.

```
bike_trips_dataframe <- transform(bike_trips_dataframe, duration = difftime(ended_at, started_at, units
```

Let's check if we have any errors - negative durations.

```
head(filter(bike_trips_dataframe, duration < 0))
```

```
##                               file_name          ride_id rideable_type
## 1 datasets/202111-divvy-tripdata.csv B029250A1EFF2975   docked_bike
## 2 datasets/202111-divvy-tripdata.csv D631251FA9C7FC03  classic_bike
```

```
## 3 datasets/202111-divvy-tripdata.csv 021DC77C70A3E367  classic_bike
## 4 datasets/202111-divvy-tripdata.csv 235ACD294AFB837F electric_bike
## 5 datasets/202111-divvy-tripdata.csv 6A2DCA5CB1596CA6  classic_bike
## 6 datasets/202111-divvy-tripdata.csv E89DD4EBFBD231E3  classic_bike
##            started_at            ended_at         start_station_name
## 1 2021-11-07 01:40:02 2021-11-07 01:05:46   Halsted St & Dickens Ave
## 2 2021-11-07 01:52:53 2021-11-07 01:05:22      Clark St & Newport St
## 3 2021-11-07 01:40:13 2021-11-07 01:00:29       New St & Illinois St
## 4 2021-11-07 01:34:03 2021-11-07 01:17:13 Sheridan Rd & Lawrence Ave
## 5 2021-11-07 01:54:25 2021-11-07 01:03:44  Franklin St & Illinois St
## 6 2021-11-07 01:54:04 2021-11-07 01:25:57    Orleans St & Hubbard St
##   start_station_id                    end_station_name end_station_id
## 1            13192             Leavitt St & Division St            658
## 2              632           Racine Ave & Fullerton Ave   TA1306000026
## 3     TA1306000013                Michigan Ave & 8th St            623
## 4     TA1309000041 Damen Ave & Thomas St (Augusta Blvd)   TA1307000070
## 5              RN-  Mies van der Rohe Way & Chicago Ave          13338
## 6              636                 Clark St & Drummond Pl   TA1307000142
##   start_lat start_lng  end_lat   end_lng member_casual       duration
## 1  41.91994 -87.64883 41.90300 -87.68382        casual -34.26667 mins
## 2  41.94454 -87.65468 41.92556 -87.65840        member -47.51667 mins
## 3  41.89085 -87.61862 41.87277 -87.62398        casual -39.73333 mins
## 4  41.96948 -87.65473 41.90143 -87.67743        member -16.83333 mins
## 5  41.89102 -87.63548 41.89691 -87.62174        casual -50.68333 mins
## 6  41.89003 -87.63662 41.93125 -87.64434        casual -28.11667 mins
```

Now let's clear the errors.

```
bike_trips_dataframe <- filter(bike_trips_dataframe, duration >= 0)
```

Let's add days of week too, Monday would be 1, Sunday would be 7 and so on.

```
bike_trips_dataframe <- transform(bike_trips_dataframe, day_of_week = wday(started_at, week_start = 1))
```

Analysis

The goal of analysis is to find behavior patters of casual members.
I'd like to find some correlations between **membership type** and following attributes: **bike types**, **trip durations** and **days of week**.

Bike types correlations

Let's make a pivot table on membership type vs. bike type.

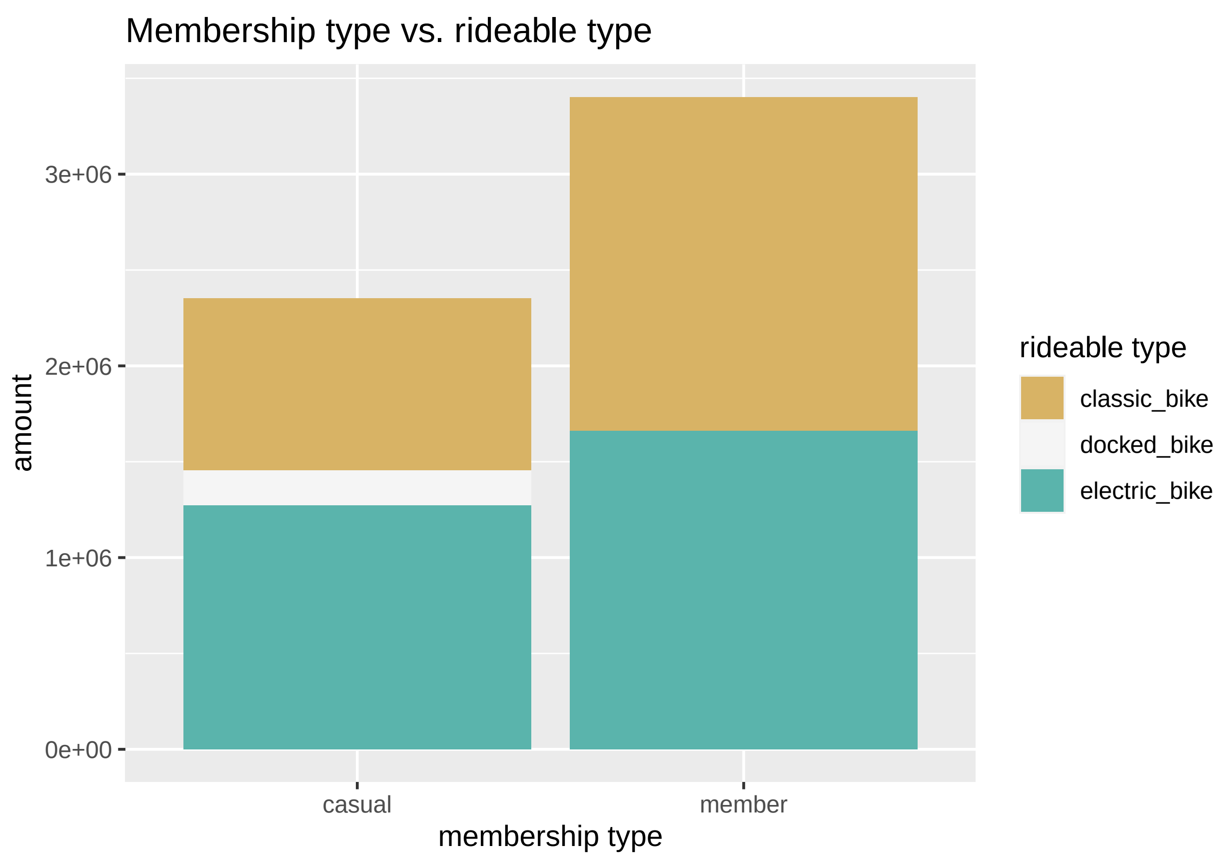
```
qpvt(bike_trips_dataframe, "rideable_type", "member_casual","n()")
```

```
##                 casual   member   Total
## classic_bike     897411  1740493  2637904
## docked_bike      182200            182200
## electric_bike   1273361  1662117  2935478
## Total           2352972  3402610  5755582
```

Two insights I could see are:
1. Casual users tend to use electric bikes much more than classic ones, where members prefer the other option.
2. There are no data of member's usage of docked bikes which can be explained in a way that there are no such thing as docks for subscribers.

Let's visualize it.

```
ggplot(data=bike_trips_dataframe)+
  geom_bar(mapping=aes(x=member_casual, fill=rideable_type))+
  scale_fill_brewer(type = "div")+
  labs(x="membership type", y="amount", fill="rideable type", title="Membership type vs. rideable type")
```



Days of week correlations

Let's make a pivot table on membership type vs. day of week.

```
qpvt(bike_trips_dataframe, "day_of_week", "member_casual","n()")
```

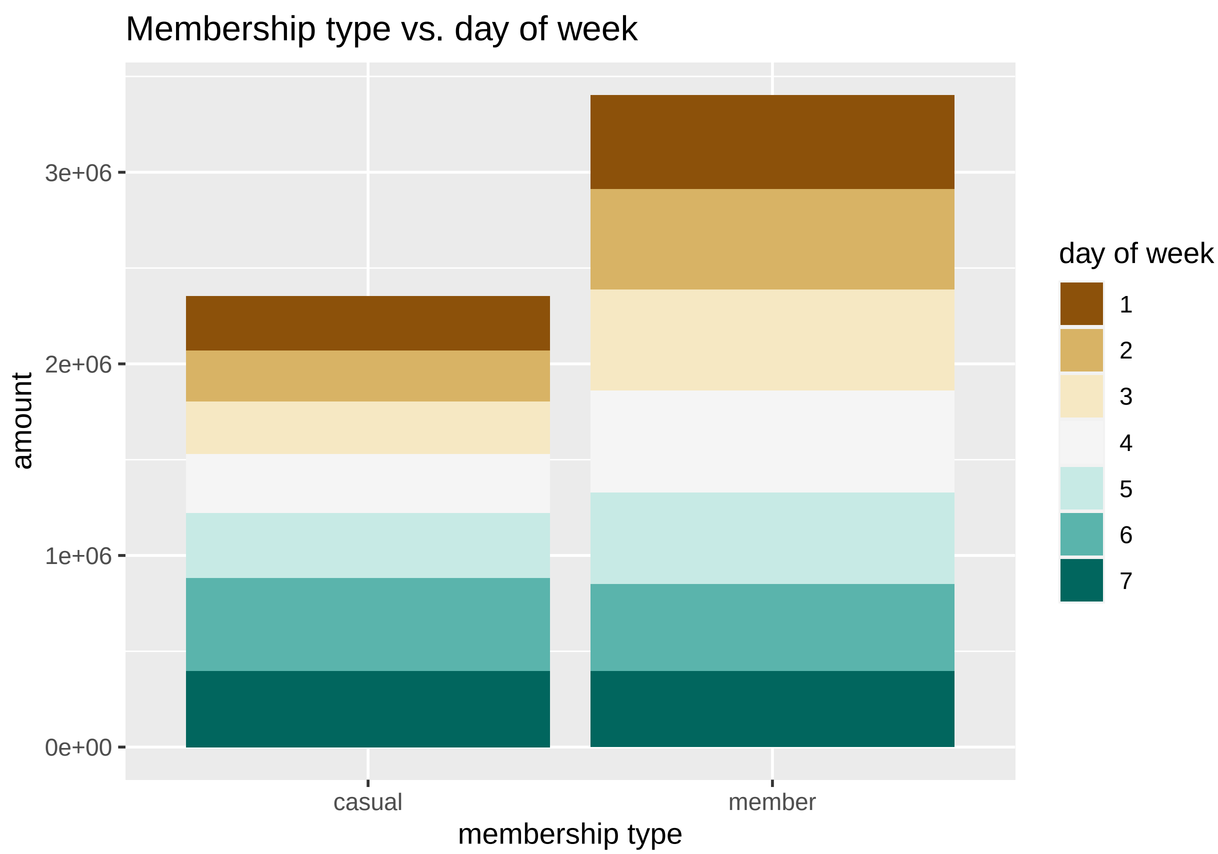
```
##        casual   member   Total
## 1       284967   489683   774650
## 2       264384   523703   788087
## 3       275394   529655   805049
## 4       306947   532664   839611
## 5       338956   476693   815649
```

```
## 6        485276   454383   939659
## 7        397048   395829   792877
## Total  2352972  3402610  5755582
```

Key insight I could see is:
Casual users tend to use bikes much more on weekends as opposed to members that are using bikes on weekdays more frequently.

Let's visualize it.

```
ggplot(data=bike_trips_dataframe)+
  geom_bar(mapping=aes(x=member_casual, fill=factor(day_of_week)))+
  scale_fill_brewer(type = "div")+
  labs(x="membership type", y="amount", fill="day of week", title="Membership type vs. day of week")
```



Days of week and mean duration correlations

Let's make a pivot table on membership type vs. day of week but this time include mean duration.

```
qpvt(bike_trips_dataframe, "day_of_week", "member_casual", "mean(as.numeric(duration))", format="%.1f")
```

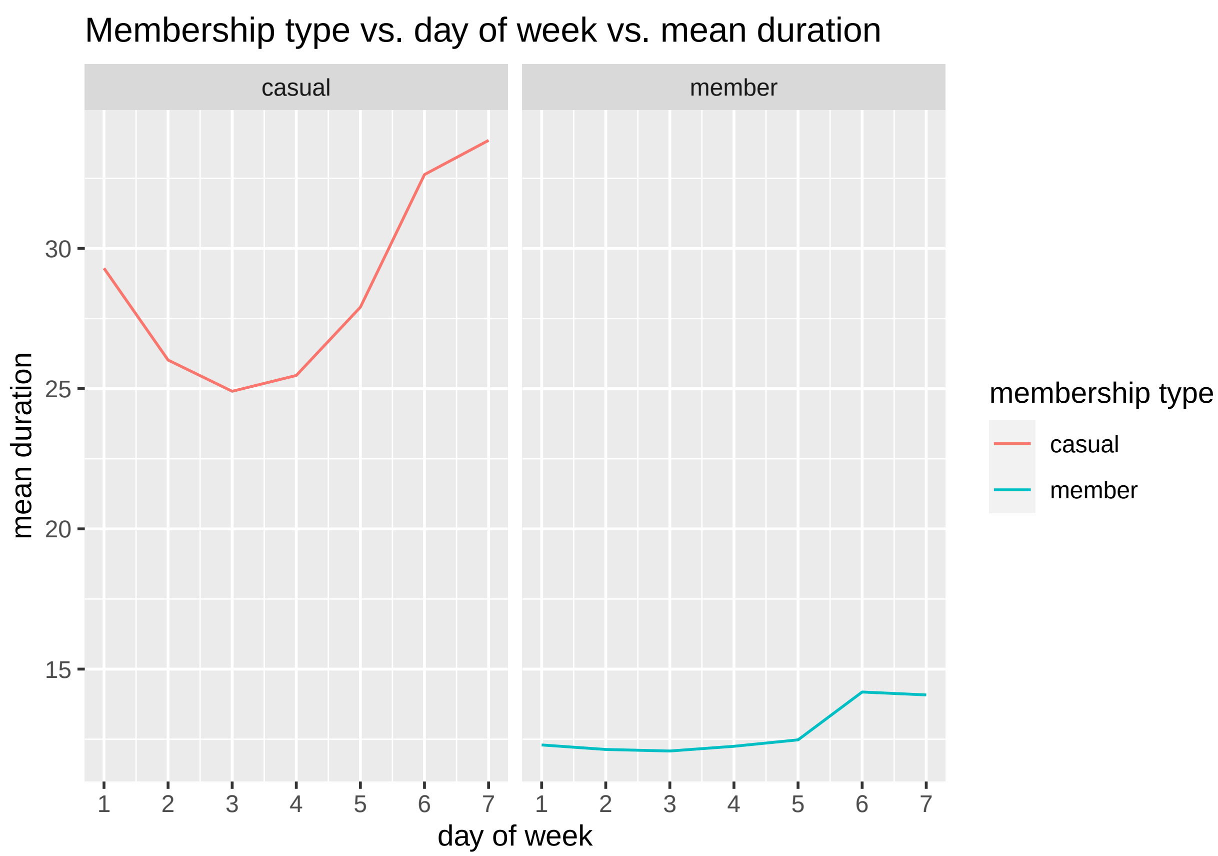
```
##        casual  member  Total
## 1        29.3    12.3   18.5
## 2        26.0    12.1   16.8
```

```
## 3         24.9    12.1   16.5
## 4         25.5    12.2   17.1
## 5         27.9    12.5   18.9
## 6         32.6    14.2   23.7
## 7         33.9    14.1   24.0
## Total     29.2    12.7   19.4
```

Two insight I could see are:
1. Average trip duration is much higher for casual members.
2. Distribution of duration on days of week is similar with subscribed members with pretty high increase on Mondays and Sundays.

Yet again let's visualize it.

```
ggplot(bike_trips_dataframe,aes(color=member_casual))+
  stat_summary(fun="mean", geom="line", mapping=aes(x=day_of_week, y=as.numeric(duration)))+
  scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7))+
  facet_wrap(~member_casual)+
  labs(x="day of week", y="mean duration", color="membership type", title="Membership type vs. day of we
```



Additional analysis on stations

Now I'll get back to stations data and create sub-dataframes without any missing values.

```
no_start_missing <- filter(bike_trips_dataframe, !is.na(start_station_id))
no_end_missing <- filter(bike_trips_dataframe, !is.na(end_station_id))
```

Let's limit dataframes to include only 10 most popular stations in both cases.

```
most_popular_starts <- c(head(no_start_missing %>% count (start_station_name, sort=TRUE), 10)['start_sta
popular_starts_df <- filter(no_start_missing, start_station_name %in% most_popular_starts$start_station_
most_popular_ends <- c(head(no_end_missing %>% count (end_station_name, sort=TRUE), 10)['end_station_nam
popular_ends_df <- filter(no_end_missing, end_station_name %in% most_popular_ends$end_station_name)
```

Start stations correlations

Let's make a pivot table on membership type vs. start stations.

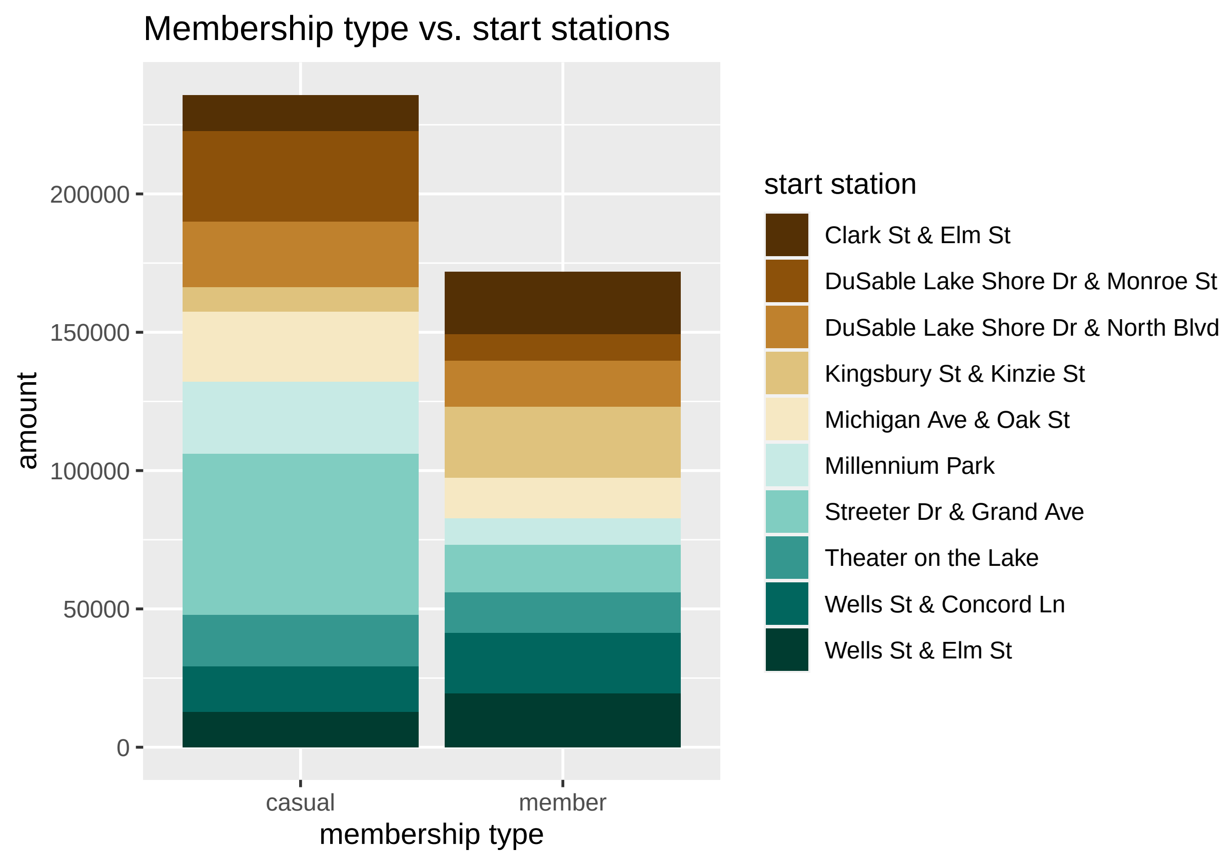
```
qpvt(popular_starts_df, "start_station_name", "member_casual", "n()")
```

```
##                                     casual  member  Total
## Clark St & Elm St                    13198   22665   35863
## DuSable Lake Shore Dr & Monroe St    32598    9595   42193
## DuSable Lake Shore Dr & North Blvd   23792   16554   40346
## Kingsbury St & Kinzie St              8776   25800   34576
## Michigan Ave & Oak St                25327   14515   39842
## Millennium Park                      25989    9559   35548
## Streeter Dr & Grand Ave              58382   17248   75630
## Theater on the Lake                  18574   14635   33209
## Wells St & Concord Ln                16391   21896   38287
## Wells St & Elm St                    12768   19426   32194
## Total                               235795  171893  407688
```

Key insights would be stations that tend to be a trip start more frequently (relatively) for casual member; these would be following stations:
1. Streeter Dr & Grand Ave with great increase.
2. DuSable Lake Shore Dr & Monroe St with great increase.
3. Millennium Park with high increase.
4. Michigan Ave & Oak St with slight increase.

Let's visualize it.

```
ggplot(data=popular_starts_df)+
  geom_bar(mapping=aes(x=member_casual, fill=start_station_name))+
  scale_fill_brewer(type = "div")+
  labs(x="membership type", y="amount", fill="start station", title="Membership type vs. start stations
```



End stations correlations

Let's make a pivot table on membership type vs. end stations.

```
qpvt(popular_ends_df, "end_station_name", "member_casual", "n()")
```

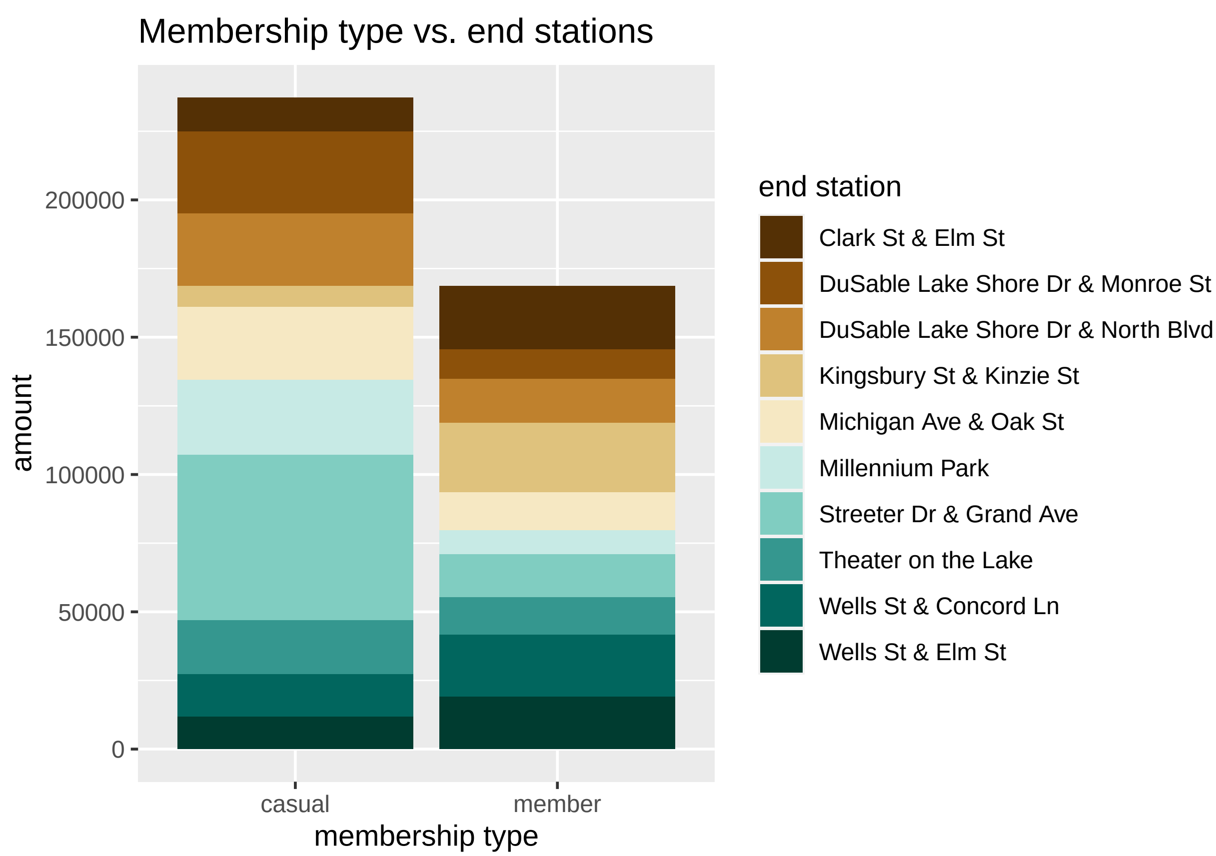
```
##                                      casual  member  Total
## Clark St & Elm St                     12317   23045  35362
## DuSable Lake Shore Dr & Monroe St     30012   10719  40731
## DuSable Lake Shore Dr & North Blvd    26252   16066  42318
## Kingsbury St & Kinzie St               7714   25277  32991
## Michigan Ave & Oak St                 26599   13799  40398
## Millennium Park                       27184    8714  35898
## Streeter Dr & Grand Ave               60254   15746  76000
## Theater on the Lake                   19563   13667  33230
## Wells St & Concord Ln                 15646   22490  38136
## Wells St & Elm St                     11745   19080  30825
## Total                                237286  168603 405889
```

Key insights would be stations that tend to be a trip end more frequently (relatively) for casual member; these would be following stations:
1. Streeter Dr & Grand Ave with great increase.
2. Millennium Park with high increase.
3. DuSable Lake Shore Dr & Monroe St with high increase.

4. DuSable Lake Shore Dr & North Blvd with slight increase.
5. Michigan Ave & Oak St with slight increase.

Let's visualize it.

```
ggplot(data=popular_ends_df)+
  geom_bar(mapping=aes(x=member_casual, fill=end_station_name))+
  scale_fill_brewer(type = "div")+
  labs(x="membership type", y="amount", fill="end station", title="Membership type vs. end stations")
```



Answering questions

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

Key findings are:
1. Casual users tend to use electric bikes much more than classic ones, where members prefer the other option.
2. Casual users tend to use bikes much more on weekends as opposed to members that are using bikes on weekdays more frequently.
3. Average trip duration is much higher for casual members.
4. Distribution of duration on days of week is similar with subscribed members with pretty high increase on Mondays and Sundays.
*5. **Streeter Dr & Grand Ave**, **DuSable Lake Shore Dr & Monroe St**, **Millennium Park** and **Michigan Ave & Oak St** are much more popular stations for casual members.*

Why would casual riders buy Cyclistic annual memberships?

Based on findings, my suggestions would be:

1. *Some benefits on electric bikes usage for subscribed members.*
2. *Better plans on weekend trips for subscribed members.*
3. *Profitable offers on longer trips (let's say, more than 15 or 20 minutes) for subscribed members.*

How can Cyclistic use digital media to influence casual riders to become members?

If promoted via app itself, advertisements will be more successful on weekends and Mondays than on weekdays.

If promoted via street screens or banners, advertisements will be more useful in following areas: **Streeter Dr & Grand Ave**, **DuSable Lake Shore Dr & Monroe St**, **Millennium Park** and **Michigan Ave & Oak St**.