Cyclistic Case Study

Richard Laffar

2024-09-26

Introduction In this case study I worked for a fictional company called Cyclistic, a bike-share company, based in Chicago. I followed the process from the Google Data Analytics Course; Ask, Prepare, Process, Analyse, Share and Act.

Statement of Business task (Ask) To identify how annual members and casual riders use Cyclistic bikes differently enabling the business to design marketing strategies.

My analysis will consist of;

- analysis of temporal daily, weekly and monthly journey patterns relating to casual users, with comparison to subscribers when appropriate
- analysis of journey durations in conjunction with the temporal patterns
- insight into any geographical patterns

Source Data (Prepare) The historial trip data relates to Cyclistic Bike-Share, a fictional company.

The original data is publicly available and downloaded from Divvy, and made available by Motivate International Inc. under licence here.

As a newcomer to data analytics I decided to attempt the whole exercise in R to enhance my skills. To avoid errors due to available memory in RStudio I I followed the course guidelines to use the quarterly data, instead of more recent monthly data. This consisted of 4 files and, up to 15 columns of data, only 8 of which existed in all 4 files. The data is comprehensive and contained 3.76 million rows of data after cleaning each row relating to a single trip.

I checked the reliability of the data for missing values and anomalies. The steps I took to correct these are outlined below.

Organising and Cleaning the Data (Process) To ensure the data is reliable I carried out a number of steps to re-organise, standardise and clean the data. The steps are outlined below:

- some columns were renamed, to ensure consistency, prior to joining datasets
- some columns were re-ordered so the datasets could be joined
- columns not present in all 4 datasets were deleted prior to joining the datasets
- missing values, and related data, were deleted
- selected rows were deleted to improve data integrity (i.e. those rows with negative journey times)
- dates were reformatted, from characters to ddtm, data to enable data analysis
- new columns were created for trip duration, day, month and season
- data was joined into a single data set

Part One: Importing Packages and Data {# Check and install packages only if needed}
packages <- c("dplyr", "skimr", "janitor", "here", "lubridate", "ggplot2", "knitr")
new_packages <- packages[!(packages %in% installed.packages()[,"Package"])] if(length(new_packages))
install.packages(new_packages)

```
library(dplyr)
library(skimr)
library(janitor)
library(here)
library(lubridate)
library(ggplot2)
library(knitr)
```

```
Q1 <- read.csv("~/Desktop/Cyclistic Files/Divvy_Trips_2020_Q1.csv")
Q2 <- read.csv("~/Desktop/Cyclistic Files/Divvy_Trips_2019_Q2.csv")
Q3 <- read.csv("~/Desktop/Cyclistic Files/Divvy_Trips_2019_Q3.csv")
Q4 <- read.csv("~/Desktop/Cyclistic Files/Divvy_Trips_2019_Q4.csv")
```

Part Two: Cleaning, Organising and Standardising the Data R code to rename and standardise column names

```
Q1 <- rename(Q1,
        trip_id=ride_id,
        start_time_2 = started_at,
        end_time_2 = ended_at,
        usertype = member_casual
        )
Q4 <- rename(Q4,
        start_station_id = from_station_id,
        start_station_name = from_station_name,
        end_station_id  = to_station_id,
        end_station_name = to_station_name,
        end_time_2 = end_time,
        start_time_2 = start_time
        )
Q3 <- rename(Q3,
        start_station_id = from_station_id,
        start_station_name = from_station_name,
        end_station_id  = to_station_id,
        end_station_name = to_station_name,
        end_time_2 = end_time,
        start_time_2 = start_time
        )
Q2 <- rename(Q2,
        trip_id = X01...Rental.Details.Rental.ID,
        start_time_2 = X01...Rental.Details.Local.Start.Time,
        end_time_2 = X01...Rental.Details.Local.End.Time,
        bike_id = X01...Rental.Details.Bike.ID,
        trip_duration = X01...Rental.Details.Duration.In.Seconds.Uncapped,
        start_station_id = X03...Rental.Start.Station.ID,
        start_station_name = X03...Rental.Start.Station.Name,
        end_station_name = X02...Rental.End.Station.ID,
        end_station_id = X02...Rental.End.Station.Name,
```

```
        Member.Birth_Year = X05...Member.Details.Member.Birthday.Year,
        usertype = User.Type
)
```

R code to convert date from chr to ddtm

```
Q1$start_time <- as_datetime(Q1$start_time_2)
Q1$end_time <- as_datetime(Q1$end_time_2)
Q2$start_time <- as_datetime(Q2$start_time_2)
Q2$end_time <- as_datetime(Q2$end_time_2)
Q3$start_time <- as_datetime(Q3$start_time_2)
Q3$end_time <- as_datetime(Q3$end_time_2)
Q4$start_time <- as_datetime(Q4$start_time_2)
Q4$end_time <- as_datetime(Q4$end_time_2)
```

The R code below calculates trip duration. It added new columns for duration, as well as, day, month and season of travel to enable the creation of the charts below. It also renames usertypes to ensure consistency before joining the data sets.

```
Q1v2 <- Q1%>%
  mutate(
    trip_duration = end_time - start_time,
    hours = hour(start_time),
    weekday = weekdays(start_time),
    month = month(start_time),
    quarter = quarters(start_time),
    usertype = ifelse(as.character(usertype) == "member", "Subscriber", as.character(usertype)),
    usertype = ifelse(as.character(usertype) == "casual", "Casual", as.character(usertype))
  )
Q2v2 <- Q2%>%
  mutate(
    trip_duration = end_time - start_time,
    hours = hour(start_time),
    weekday = weekdays(start_time),
    month = month(start_time),
    quarter = quarters(start_time),
    usertype = ifelse(as.character(usertype) == "Customer", "Casual", as.character(usertype))
  )
Q3v2 <- Q3%>%
  mutate(
    trip_duration = end_time - start_time,
    hours = hour(start_time),
    weekday = weekdays(start_time),
    month = month(start_time),
    quarter = quarters(start_time),
    usertype = ifelse(as.character(usertype) == "Customer", "Casual", as.character(usertype))
  )
Q4v2 <- Q4%>%
  mutate(
    trip_duration = end_time - start_time,
    hours = hour(start_time),
    weekday = weekdays(start_time),
    month = month(start_time),
```

```
    quarter = quarters(start_time),
    usertype = ifelse(as.character(usertype) == "Customer", "Casual", as.character(usertype))
  )
```

The following code deletes those columns not present in all 4 files / datasets prior to joining the data sets

R code to reorder columns prior to joining the data sets

```
Q1v2 <- Q1v2 %>% relocate(start_station_id, .before=start_station_name)
Q1v2 <- Q1v2 %>% relocate(end_station_id, .before=end_station_name)
Q2v2 <- Q2v2 %>% relocate(trip_duration, .after=end_time)
```

R code to bind/join all 4 datasets together

```
all_qtrs <- rbind(Q1v2,Q2v2,Q3v2,Q4v2)
```

R code to delete rows with n/a (row 414427)

```
# Clean n/a
all_qtrs <- all_qtrs[complete.cases(all_qtrs), ]
```

R code to select rows with N/A ("1" searches by row, "2" by column

```
rows_with_na <- all_qtrs[apply(
  all_qtrs,
  1,
  function(x) any(is.na(x))
), ]
```

R code to remove rows with trip durations which are;

- equal to zero,
- negative,
- over 24 hours and
- journeys which begin and end at the same station

```
all_qtrs <- all_qtrs[!(all_qtrs$trip_duration == 0),]
all_qtrs <- all_qtrs[!(all_qtrs$trip_duration < 0),]
all_qtrs <- all_qtrs[!(all_qtrs$trip_duration > 86400),]
all_qtrs <- all_qtrs[!(all_qtrs$start_station_id == all_qtrs$end_station_id),]
```

Simple Calculations (Analyse) Below are some basic calculations for subscribers and casual users combined. It's apparent that casual users make longer journeys. This will be explored in greater depth later. On the basis of findings I removed outliers (see earlier step above). For example, those trips with a duration over 24 hours, were removed. This is because a day pass runs out after 24 hours (or 86,400 seconds). Also, those trips which began and ended at same station were removed (all under 32 seconds). It's also evident that more trips are made on weekdays. We'll also review this in more depth later.

```
kable(all_qtrs %>%
  summarise(
    Mean = mean(all_qtrs$trip_duration, na.rm = TRUE),
    Maximum = max(all_qtrs$trip_duration, na.rm = TRUE),
    Minimum = min(all_qtrs$trip_duration, na.rm = TRUE),
    Standard_Deviation = sd(all_qtrs$trip_duration,na.rm = TRUE)
  ),"pipe",col.names = c("Mean","Maximum","Minimum","Standard Deviation"))
```

Mean	Maximum	Minimum	Standard Deviation
1099.371 secs	86385 secs	32 secs	2014.439

```
kable(all_qtrs %>%
  group_by(usertype) %>%
  summarise(Mean = mean(trip_duration),
            Maximum = max(trip_duration),
            Minimum = min(trip_duration),
            Standard_Deviation = sd(trip_duration)
  ),"pipe",col.names = c("Usertype", "Mean","Maximum","Minimum","Standard Deviation"))
```

Mean, min and max trip duration by both subscriber and casual member

Usertype	Mean	Maximum	Minimum	Standard Deviation
Casual	2267.8891 secs	86349 secs	45 secs	3506.833
Subscriber	770.2211 secs	86385 secs	32 secs	1115.339

```
kable(all_qtrs %>%
  group_by(weekday, usertype) %>%
  summarise(
    Average_Trip = mean(trip_duration),
    Shortest_trip = min(trip_duration),
    Longest_trip = max(trip_duration),
    Standard_Deviation=sd(trip_duration)
),"pipe",col.names = c("Weekday","Usertype","Average Trip","Shortest Trip","Longest Trip","Standard Devi
```

Average trip duration by weekday and usertype (similar pattern across all days)

```
## `summarise()` has grouped output by 'weekday'. You can override using the
## `.groups` argument.
```

```
kable(all_qtrs %>%
  count(weekday),"pipe",col.names = c("Weekday","Number of Trips"))
```

Trips by week (mode)

Weekday	Number of Trips
Friday	560425
Monday	560892
Saturday	475681
Sunday	426671
Thursday	572814
Tuesday	585874
Wednesday	579337

Data Analysis (Analyse, Share and Act) I've shared the code for my first chart below which I've included as .jpg files

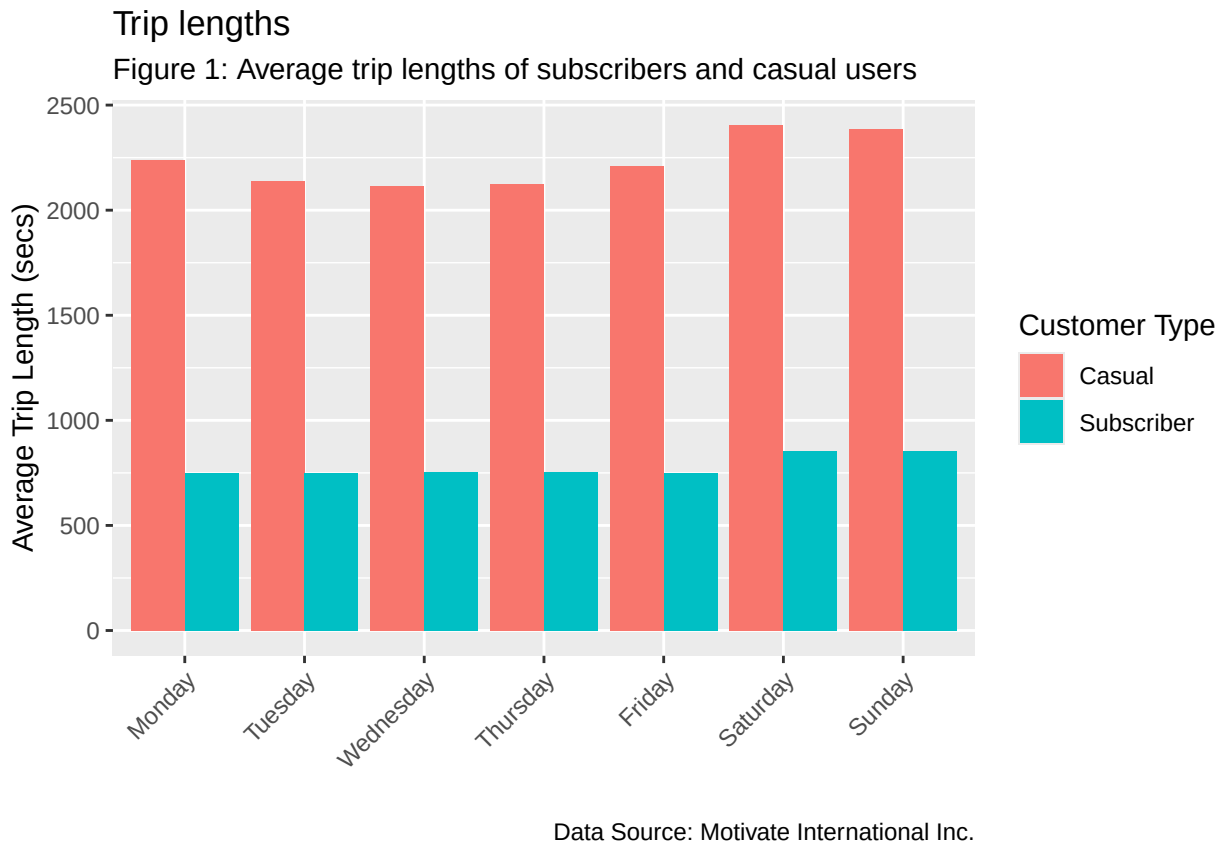
```
trip_times <- all_qtrs %>%
  select(usertype, weekday, trip_duration) %>%
  group_by(usertype, weekday) %>%
  summarise(average_trip_length = mean(trip_duration),
            total_trips = n())
```

```
## `summarise()` has grouped output by 'usertype'. You can override using the
## `.groups` argument.
```

```
  trip_times$weekday <- factor(trip_times$weekday, levels = c("Monday","Tuesday","Wednesday","Thursday"

ggplot(
  data = trip_times,
  aes(
    x = weekday,
    y= average_trip_length,
    fill = usertype,
  )) +
  labs(
    title = "Trip lengths",
    subtitle = "Figure 1: Average trip lengths of subscribers and casual users",
    caption = "Data Source: Motivate International Inc."
  )+
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle=45, hjust=1))+
  ylab("Average Trip Length (secs)")+
  xlab("")+
  labs(fill="Customer Type")
```

```
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```




```
ggsave("figure_1.jpg")
```

```
## Saving 6.5 x 4.5 in image
## Don't know how to automatically pick scale for object of type <difftime>.
## Defaulting to continuous.
```

```
knitr::include_graphics("~/Desktop/Cyclistic Files/figure_1.jpg")
```

Trip lengths are significantly longer for casual users compared to those with a subscription. This *might* be because bikes are not always returned after each individual trip if 24-hour passes have been purchased. More information would be required to confirm this theory. Also, average journey times on saturdays and sundays are longer This reflects the existing knowledge that 30% of casual riders use the bikes for commuting.

The chart clearly shows that the number of trips by casual users increases at weekends. When considered in tandem with Figure 1 we can conclude that casual users make both more trips and longer trips at weekends. It also indicates that subscribers travel less, but make longer trips (figure 1) on saturdays and sundays.

Recommendation One: Consider offering a weekend only subscription for casual users.

The chart shows a similar pattern for subscribers and casual users alike. All users make more trips during the warmer months, peaking in August. The pattern is more pronounced for casual users with very little use between November and March.

The chart shows a breakdown of (figure 2) across each month (Figure 3) of the year.

Recommendations Two & Three:

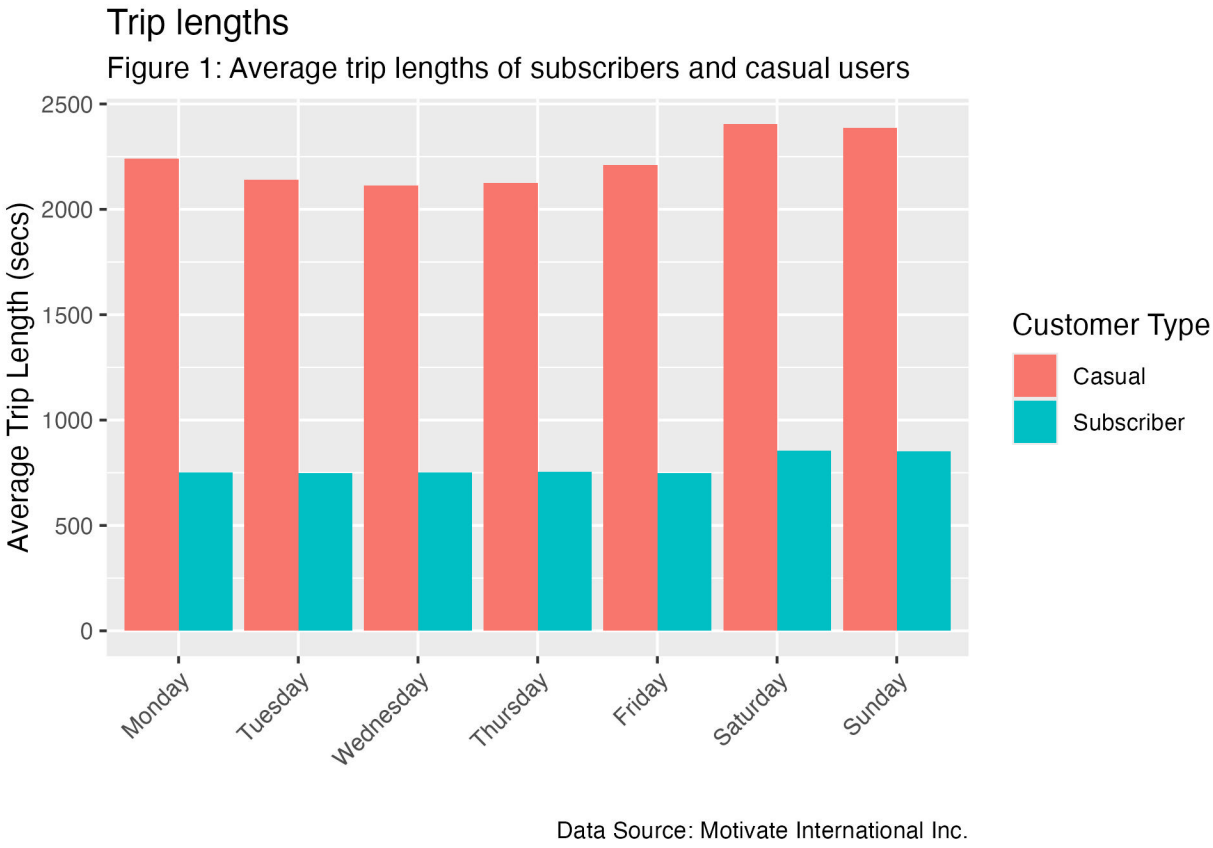


Figure 1: Average trip lengths of subscribers and casual users.

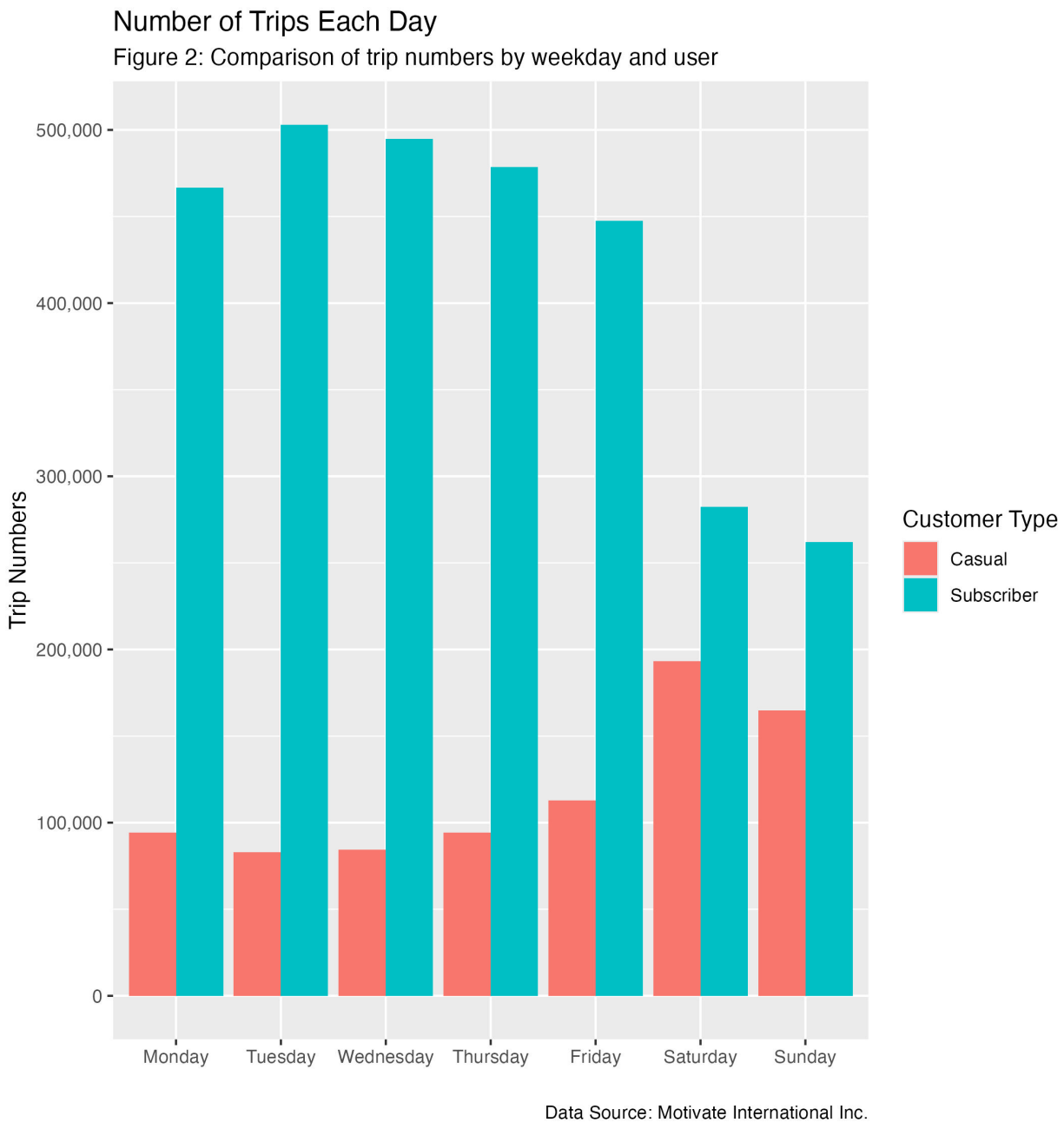


Figure 2: Comparison of trip numbers by weekday and user

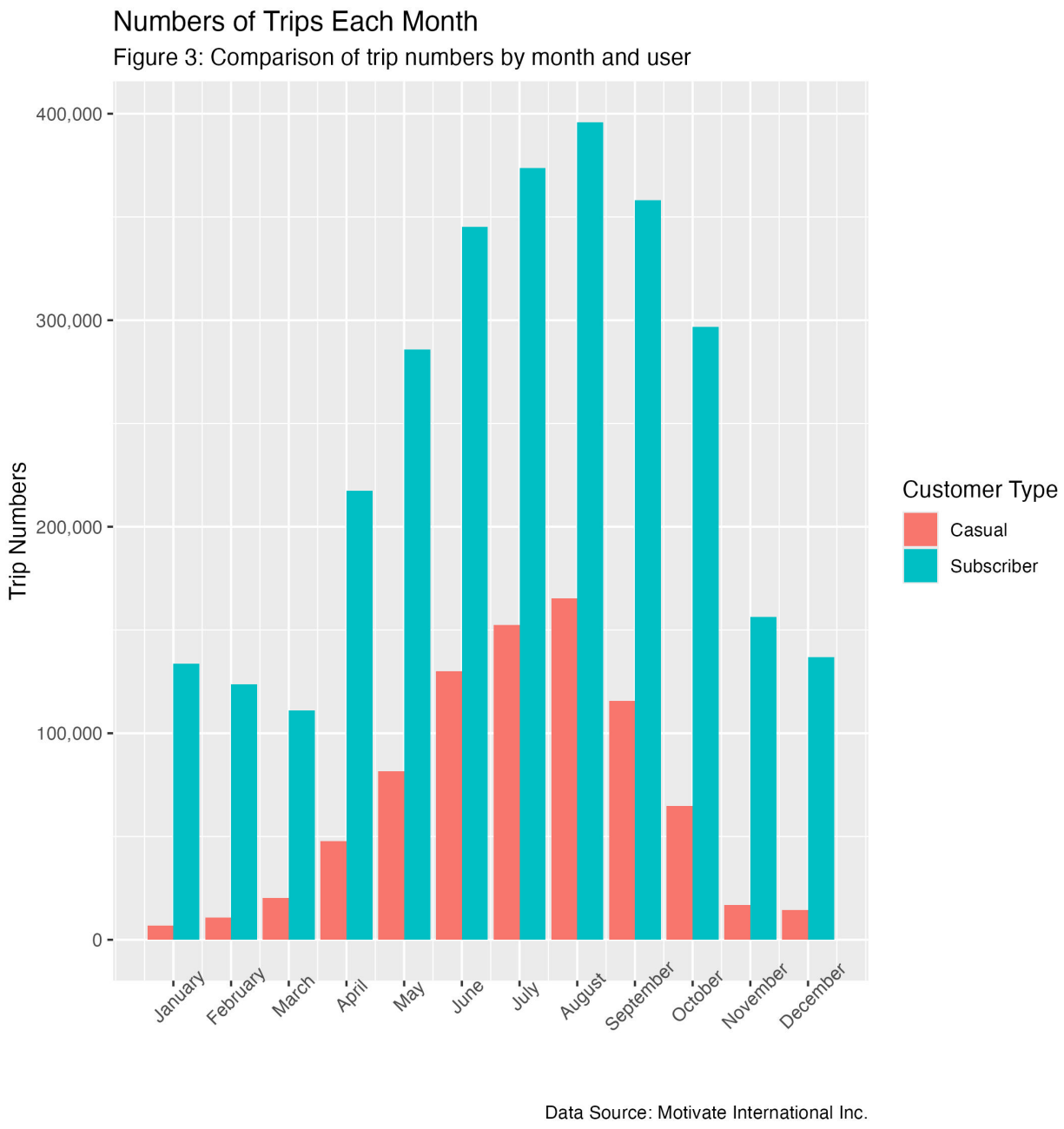


Figure 3: Comparison of trip numbers by month and user

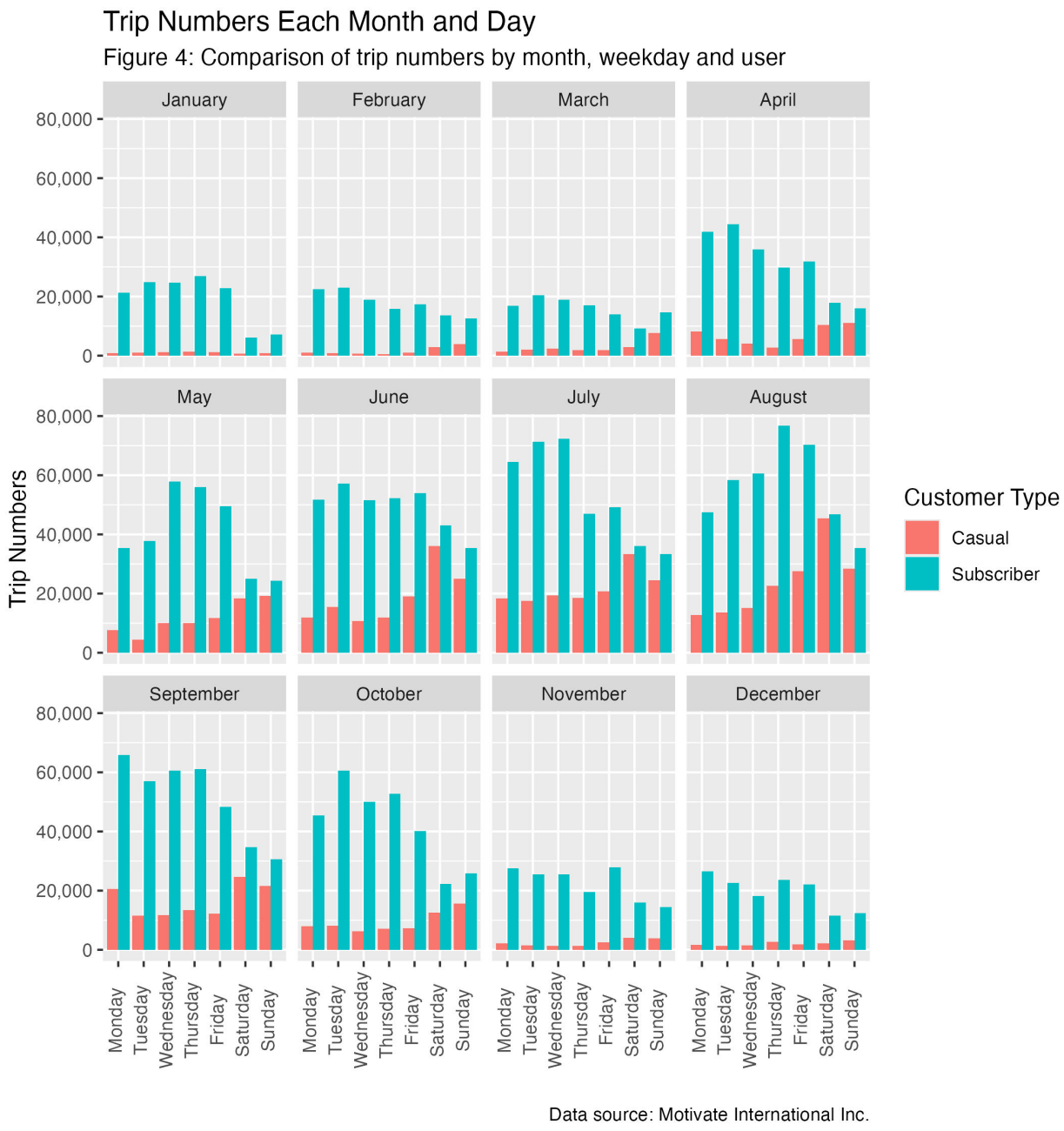


Figure 4: Comparison of trip numbers by month day and usertype

- Offer casual riders a seasonal pass perhaps spreading the 6 months from mid-April to mid-October as casual users tend to cycle during the warmest months of the year.
- Consider a hybrid of the two previous recommendations i.e. offer saturday and sunday membership for casual users for only 6 months of the year.

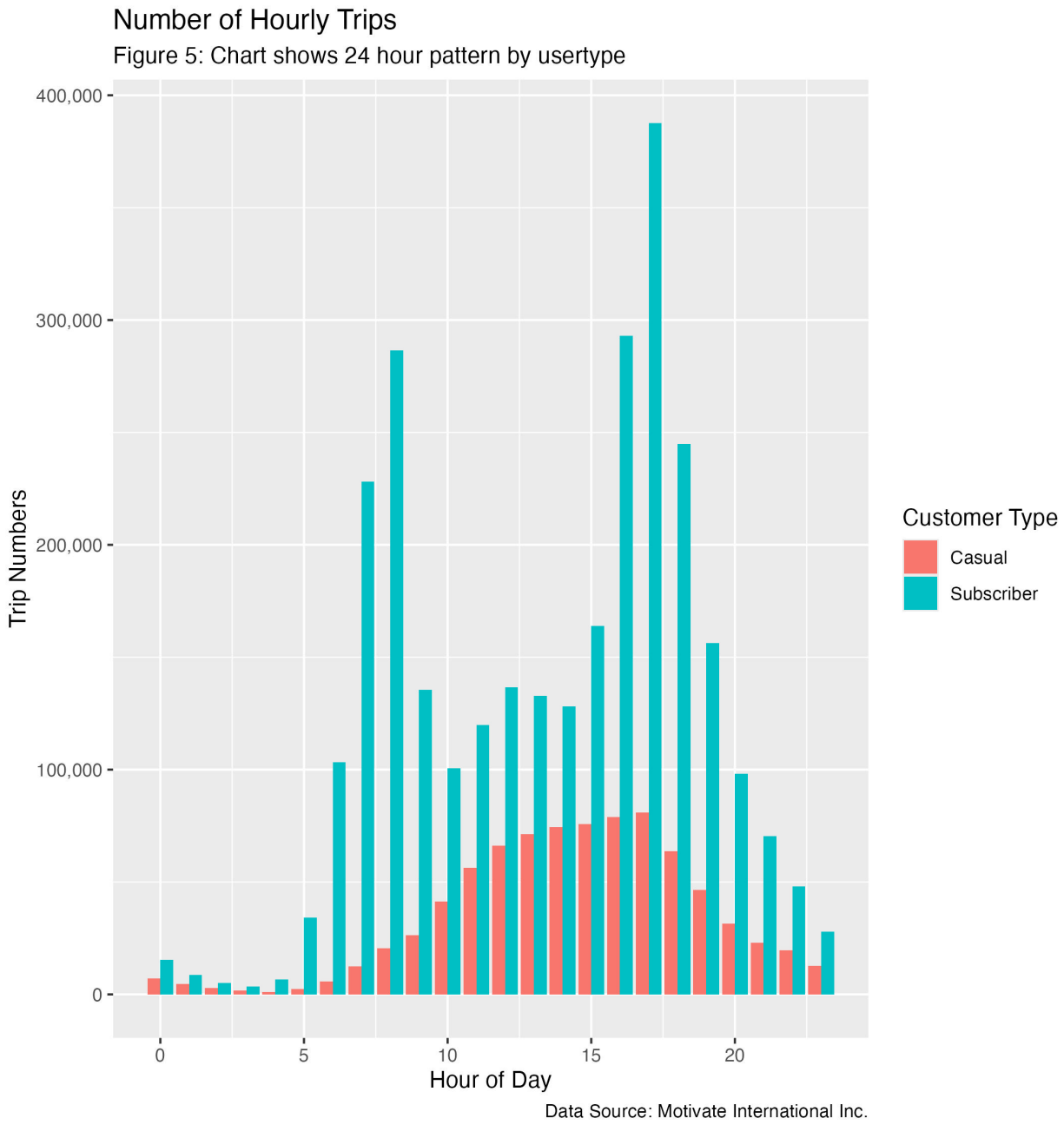


Figure 5: Chart shows number of trips during a 24 hour period

The chart shows two distinct patterns. First, a double peak during commuter hours, for subscribers. Secondly we see a hump (single peak) for casual users peaking in the middle of the day.

This chart is a faceted version of Figure 5. It's clear that user pattern is almost identical on saturdays and

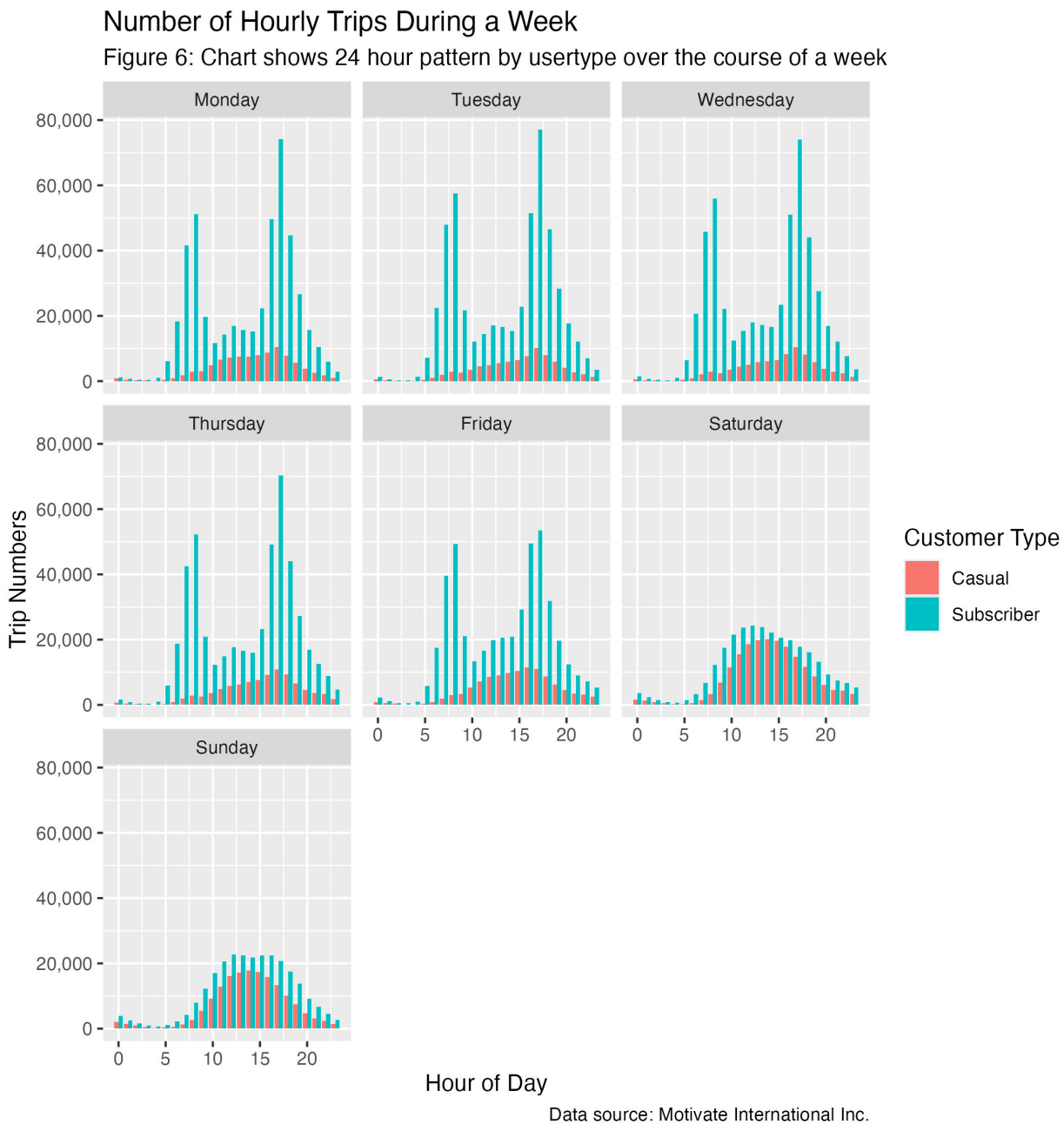


Figure 6: Number of Hourly Trips During a Week

sundays for casual users and subscribers.

Recommendation Four: Consider offering a discounted subscription aimed at casual users which would allow full weekend use and weekday use between peak hours. This will ensure cycles are always available for existing subscribers (commuters) whilst providing a second membership option for casual users.

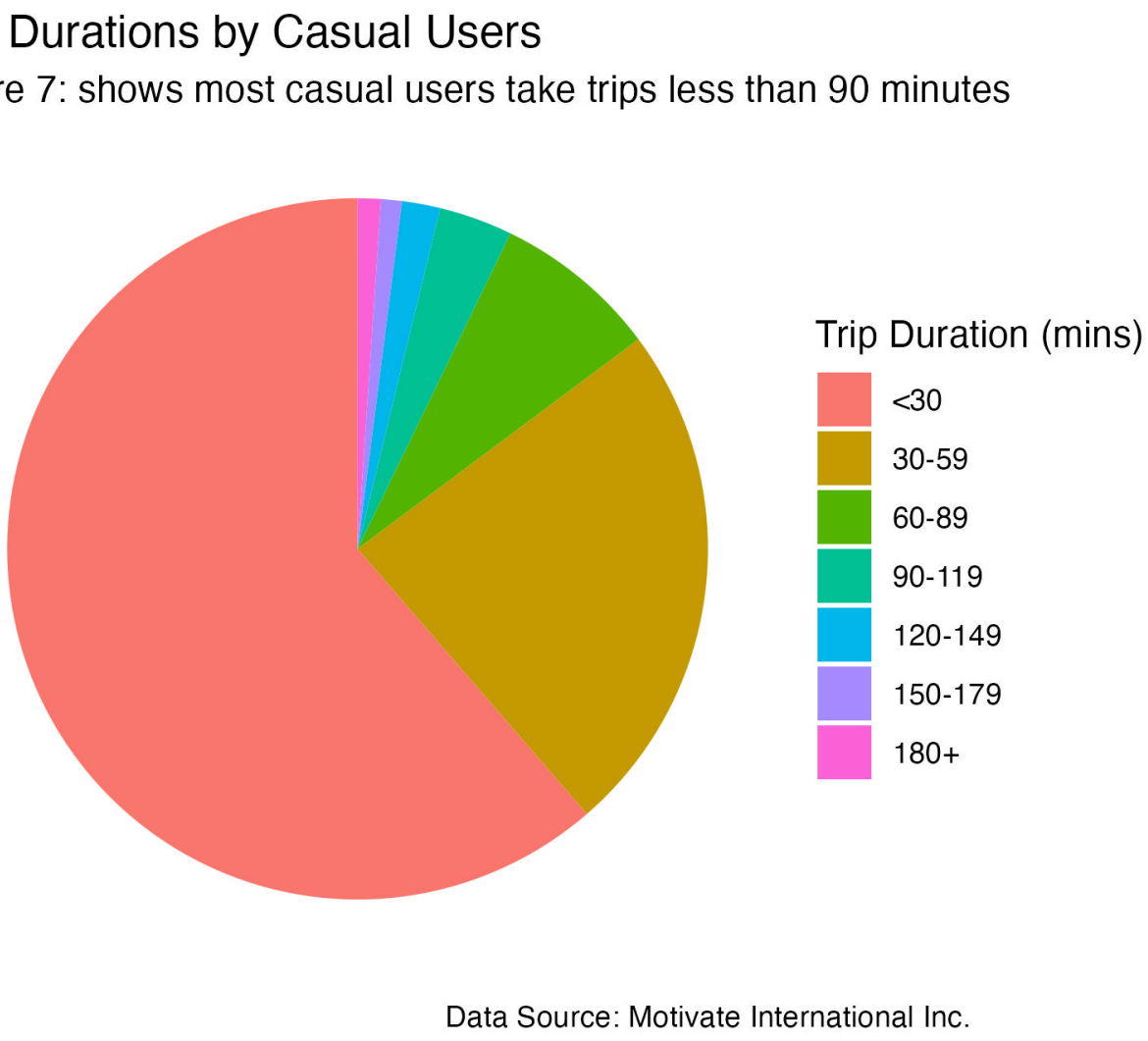


Figure 7: Trip Durations of Casual Users

The chart shows the vast majority of trips by casual users are less than an hour long, with very few journeys over 1.5 hours. This may indicate that most casual users do not buy the 24 hour pass but more data would be needed to confirm this observation.

The chart shows that most journeys originate from only a few stations for casual users, one in particular. The charts below identify these stations.

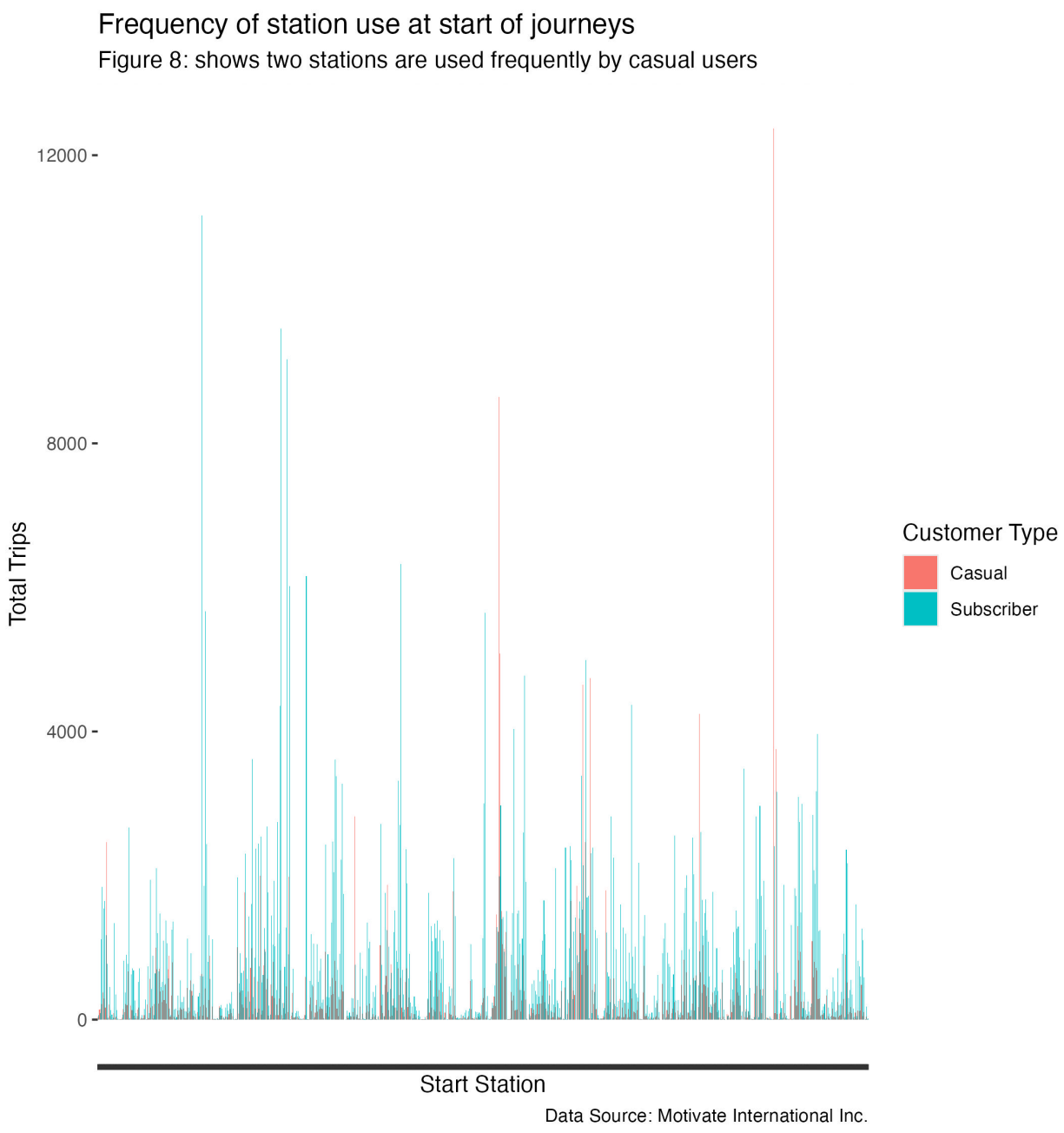


Figure 8: Most used stations at start of journeys (left)

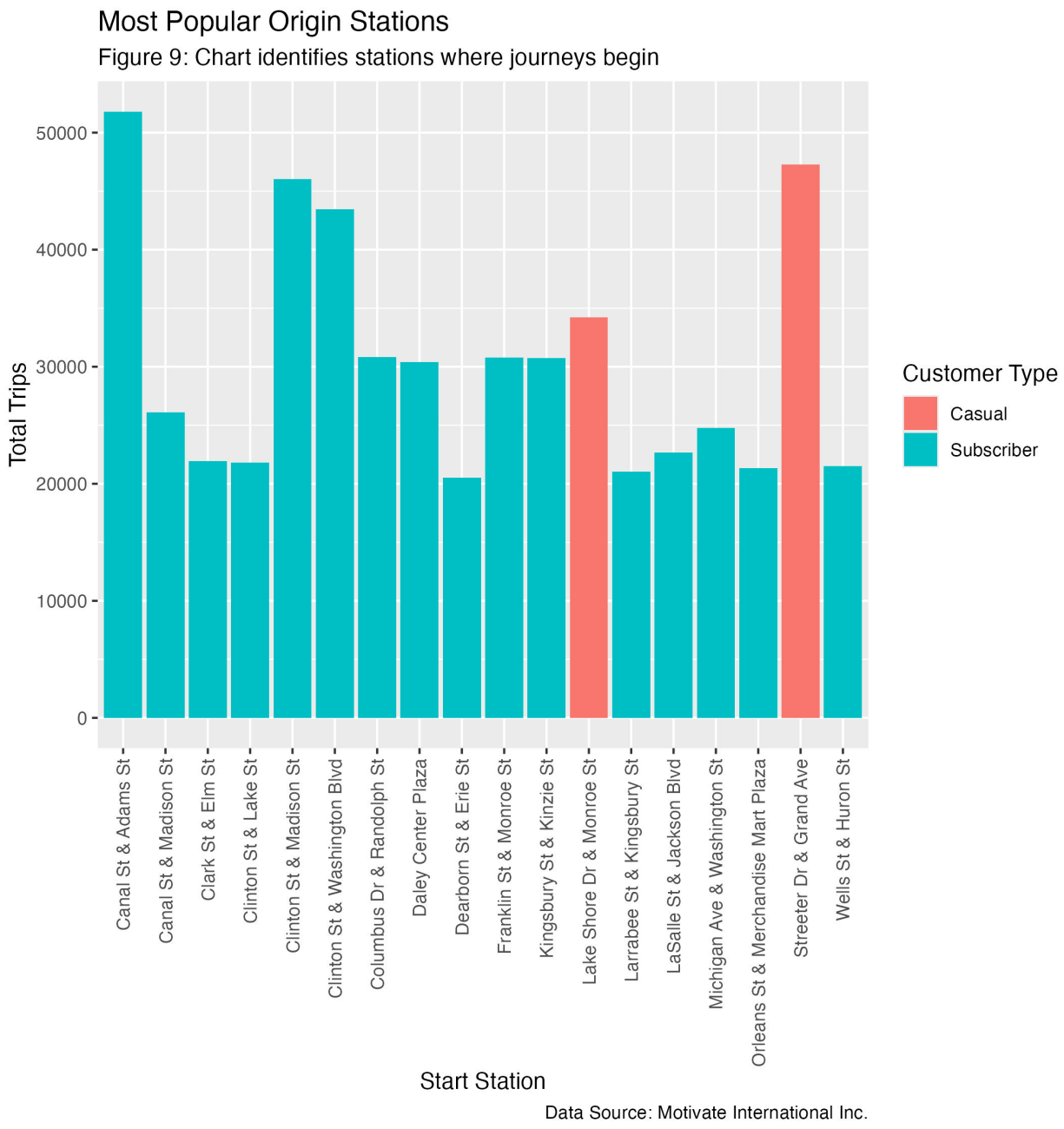


Figure 9: Most popular stations of origin

This chart identifies the most most popular stations of origin observed in Figure 8, as Streeter Drive & Grand Avenue and Lake Shore & Monroe Street.

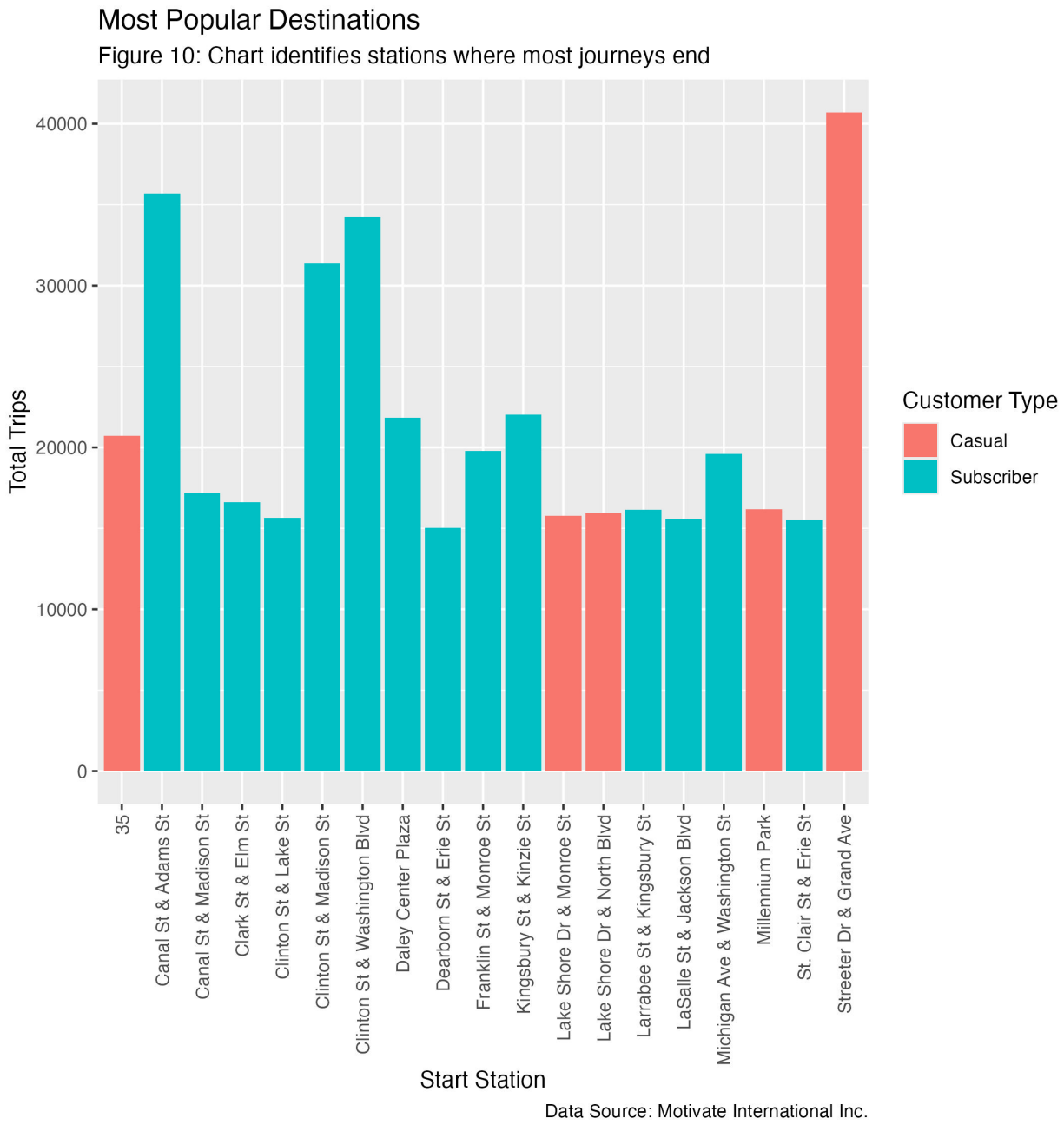


Figure 10: Most popular destination stations

The chart roughly mirrors Figure 9 as the most popular origin is also the most popular destination.

Chart shows a weekly analysis of Figure 9.

Recommendation Five: Consider focusing face-to-face customer engagement at the most popular times, and geographical locations, identified in Figures 8 - 11. For example, at Lake Shore Dr & Monroe St on Fridays, Saturdays and Sundays only.

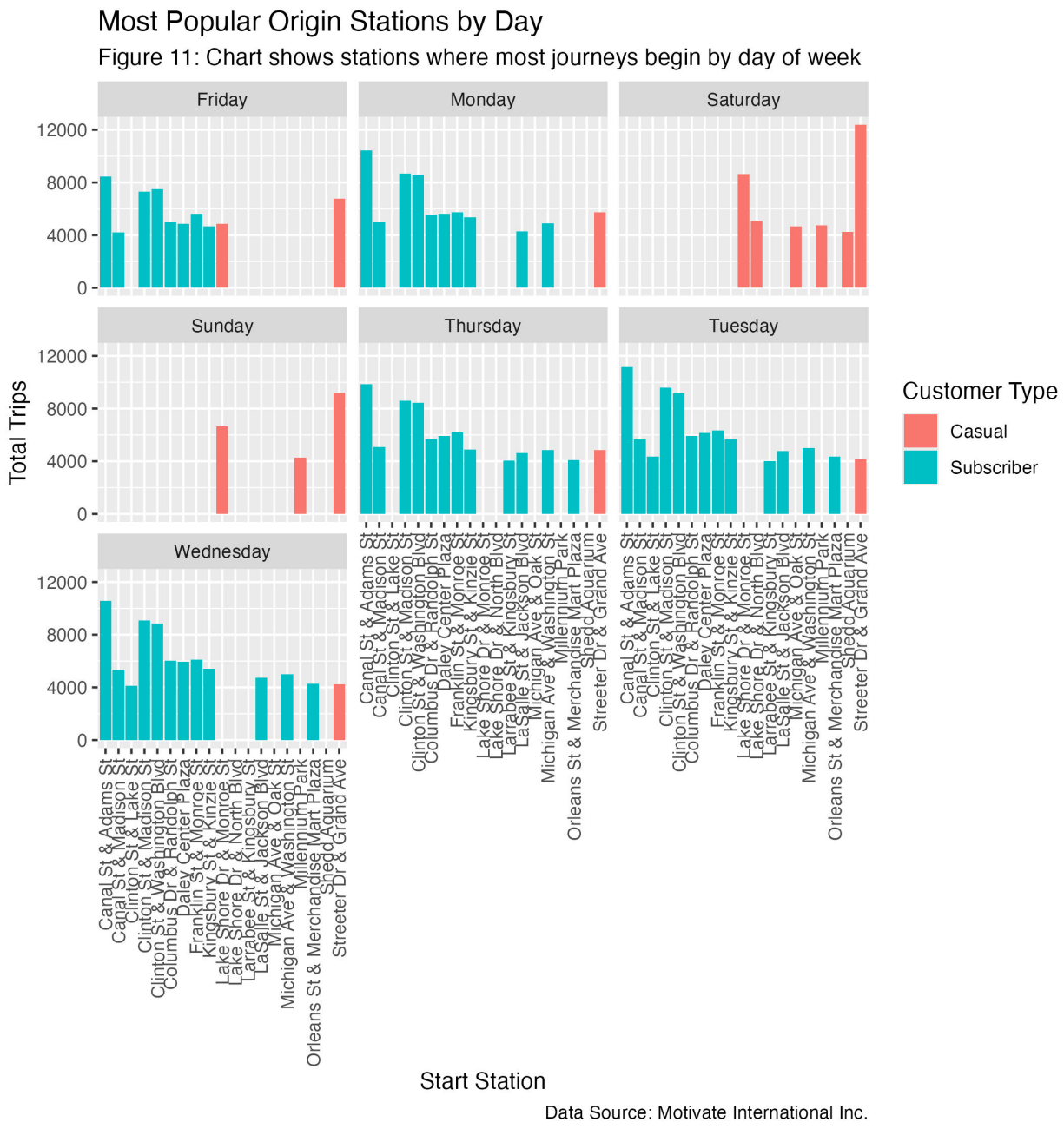


Figure 11: Identify most popular starting stations by day

Summary of Recommendations I made five recommendations in total but these tend to overlap so could be combined and condensed into fewer.

1. A weekend only annual subscription.

2. A 6 month, mid-April to mid-October, seasonal pass.

3. A hybrid of my first two recommendations. An annual membership for weekend use, during a 6 month period.

4. A discounted subscription, aimed at casual users, allowing full weekend use and weekday use outside peak (commuter) hours. As a bonus this should ensure enough cycles are always available to all users.

5. Target selected stations to engage with customers. Specifically, Streeter & Grand Avenue on any given day but focusing on Saturdays and Sundays.