Case study: How does a bike-share navigate speedy success?

Sutiasmoko

2024-01-22

Scenario

I am a junior data analyst working on the marketing analyst team at Cyclistic, a bike-share company in Chicago *(fictional company)*. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my recommendations, so they must be backed up with compelling data insights and professional data visualizations.

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the *flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships.* Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. *Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Manager believes that maximizing the number of annual members will be key to future growth.* Rather than creating a marketing campaign that targets all-new customers, Manager believes there is a solid opportunity to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Manager has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Manager and team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

Three questions will guide the future marketing program:

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?

Manager has assigned me the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

I will produce a report with the following deliverables:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of my analysis
5. Supporting visualizations and key findings
6. My top three recommendations based on my analysis

Produce a Report

1. A clear statement of the business task

- The primary goal is to understand and analyze the differences in the usage patterns of Cyclistic bikes between annual members and casual riders.
- Identify key insights that can help improve services, marketing, or user experience based on these differences.

2. A description of all data sources used

- Summary of data

```r
library(skimr)
skim_without_charts(appended_2023All)
```

Table 1: Data summary

Name	appended_2023All
Number of rows	5719877
Number of columns	13
Column type frequency:	
character	9
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	5719877	0
rideable_type	0	1	11	13	0	3	0
started_at	0	1	19	19	0	4823909	0
ended_at	0	1	19	19	0	4835702	0
start_station_name	0	1	0	64	875716	1593	0
start_station_id	0	1	0	35	875848	1517	0
end_station_name	0	1	0	64	929202	1598	0
end_station_id	0	1	0	36	929343	1521	0
member_casual	0	1	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.05	41.63	41.88	41.90	41.93	42.07
start_lng	0	1	-87.65	0.03	-87.94	-87.66	-87.64	-87.63	-87.46
end_lat	6990	1	41.90	0.05	0.00	41.88	41.90	41.93	42.18
end_lng	6990	1	-87.65	0.07	-88.16	-87.66	-87.64	-87.63	0.00

```
glimpse(appended_2023All)
```

```
## Rows: 5,719,877
## Columns: 13
## $ ride_id            <chr> "F96D5A74A3E41399", "13CB7EB698CEDB88", "BD88A2E670~
## $ rideable_type      <chr> "electric_bike", "classic_bike", "electric_bike", "~
## $ started_at         <chr> "2023-01-21 20:05:42", "2023-01-10 15:37:36", "2023~
## $ ended_at           <chr> "2023-01-21 20:16:33", "2023-01-10 15:46:05", "2023~
## $ start_station_name <chr> "Lincoln Ave & Fullerton Ave", "Kimbark Ave & 53rd ~
## $ start_station_id   <chr> "TA1309000058", "TA1309000037", "RP-005", "TA130900~
## $ end_station_name   <chr> "Hampden Ct & Diversey Ave", "Greenwood Ave & 47th ~
## $ end_station_id     <chr> "202480.0", "TA1308000002", "599", "TA1308000002", ~
## $ start_lat          <dbl> 41.92407, 41.79957, 42.00857, 41.79957, 41.79957, 4~
## $ start_lng          <dbl> -87.64628, -87.59475, -87.69048, -87.59475, -87.594~
## $ end_lat            <dbl> 41.93000, 41.80983, 42.03974, 41.80983, 41.80983, 4~
## $ end_lng            <dbl> -87.64000, -87.59938, -87.69941, -87.59938, -87.599~
## $ member_casual      <chr> "member", "member", "casual", "member", "member", "~
```

3. Documentation of any cleaning or manipulation of data

- Create and update columns part1

```
appended_2023All <- appended_2023All %>%
  mutate(
    started_at = ymd_hms(started_at),
    ended_at = ymd_hms(ended_at),
    dist_km = geosphere::distGeo(
      p1 = cbind(start_lng, start_lat),
      p2 = cbind(end_lng, end_lat)) / 1000,
    started_date = as.Date(started_at),
    ended_date = as.Date(ended_at),
    Started_time = format(started_at, format="%H:%M:%S"),
    Ended_time = format(ended_at, format="%H:%M:%S"),
    Duration = as.numeric(difftime(ended_at, started_at, units = "hours")),
    start_location = paste(start_lat, start_lng, sep = "_")
  )
```

- Create discrete bins

```
duration_bins <- c(0, 0.25, 0.5, Inf)
dist_km_bins <- c(0, 3, 5, Inf)
```

- Create and update columns part2

```
appended_2023All <- appended_2023All %>%
  mutate(
    duration_category = cut(Duration, breaks = duration_bins,
    labels = c("0-0.25 hour", "0.25-0.5 hour", "more than 0.5 hour")),
    dist_km_category = cut(dist_km, breaks = dist_km_bins,
    labels = c("0-3 km", "3-5 km", "more than 5 km")),
    started_year = year(started_date),
    started_month = month(started_date),
    started_year_month = format(started_date,"%Y-%m"),
    started_day = format(started_at,"%d"),
    started_hour = format(started_at,"%H"),
    Started_dayOfWeek = weekdays(started_at),
    started_hour = format(started_at,"%H")
  )
```

4. A summary of my analysis

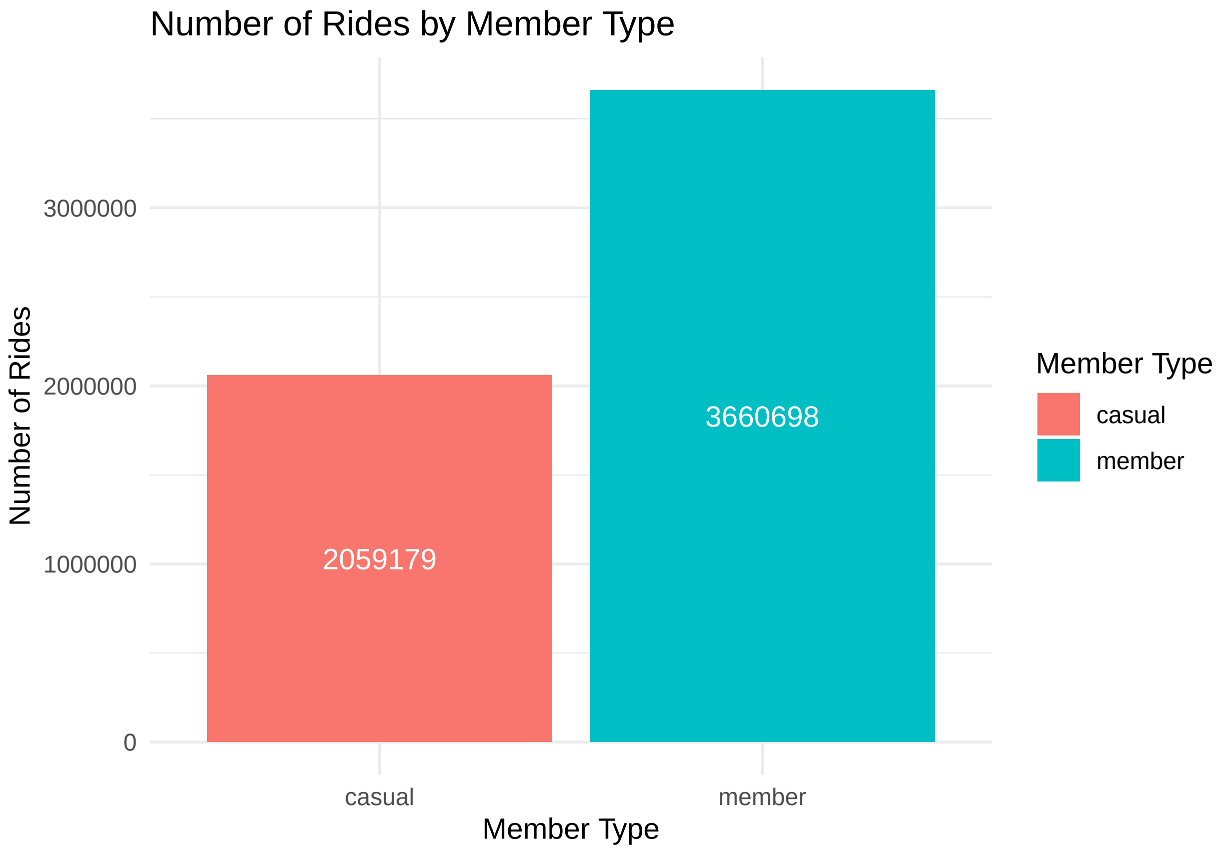
- Based on the member type, **member riders outnumber casual riders.**
- Considering the rideable type, **the utilization is dominated by classic and electric bikes**, with almost equal usage. However, docked bikes have the least utilization and are primarily used by casual riders.
- Analyzing the duration category, the majority of bike usage falls within the 0 to 0.25-hour range for both member and casual riders. Interestingly, **the distribution of the more than 0.5-hour category is higher among casual riders. Further investigation reveals that this usage pattern is prominent on Saturdays and Sundays, indicating recreational use.**
- Examining the distance category, **most riders cover distances of 0-3 km**, and the distribution is similar across member types.
- Daily riding trends reveal significant differences between casual and member riders. **Member riders predominantly use bikes on weekdays, suggesting commuting to work, while casual riders exhibit the opposite trend.** This indicates that a majority of member riders rent bikes for commuting to their workplaces.
- Hourly riding trends strengthen the observation that member riders predominantly use bikes for commuting. **Peak usage occurs at 8 AM and 5 PM, corresponding to typical work commute hours.**
- Monthly riding trends show that **casual riders peak in July**, while member riders peak in August.
- Among the top 3 start stations for casual riders, **Streeter Dr & Grand Ave has the highest ride count, followed by DuSable Lake Shore Dr & Monroe St. Michigan Ave & Oak St takes the third spot.**

5. Supporting visualizations and key findings

- **Number of Rides Distribution**

R Script: Number of Rides by Member Type

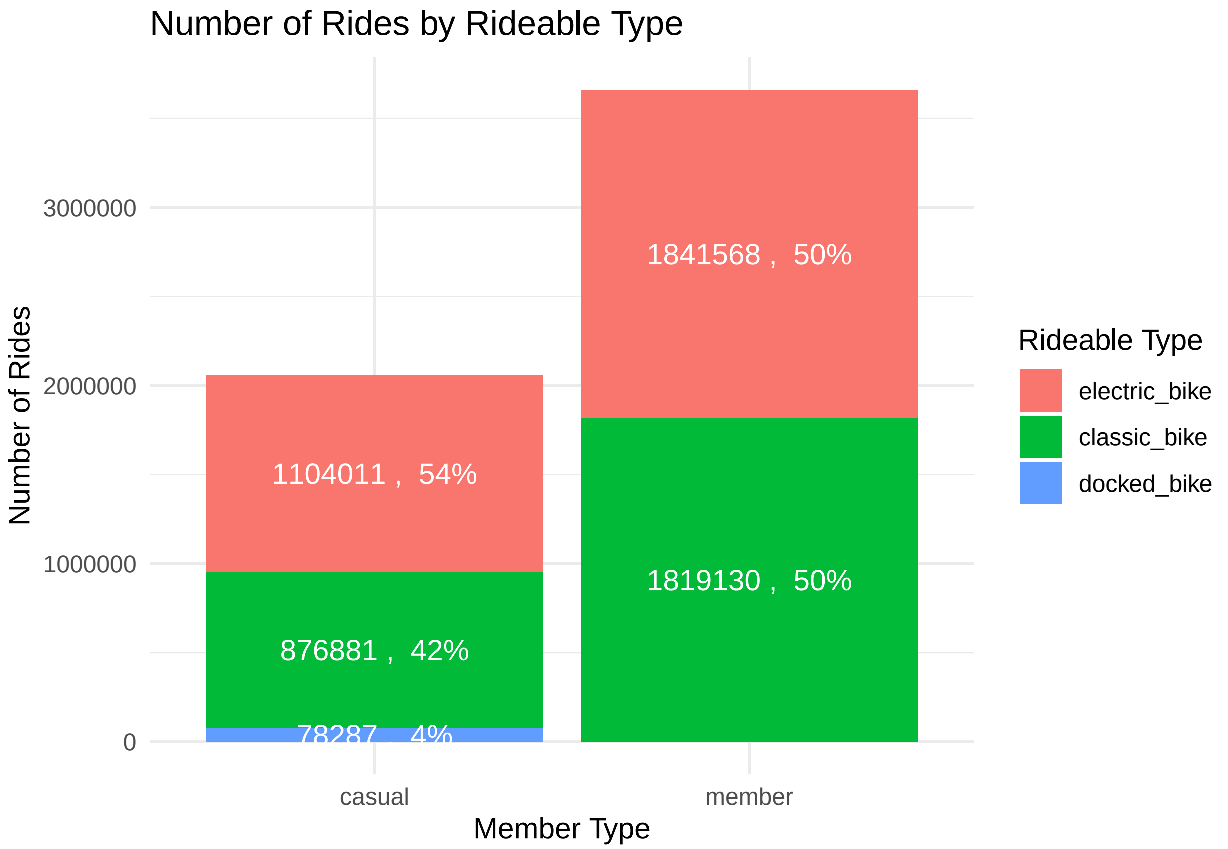
```
ride_frequency <- appended_2023All %>%
  group_by(member_casual) %>%
  summarise(count = n())

ggplot(ride_frequency, aes(x = member_casual, y = count, fill = member_casual)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = count), position=position_stack(vjust = 0.5), colour = "white") +
  labs(title = "Number of Rides by Member Type",
       x = "Member Type",
       y = "Number of Rides",
       fill = "Member Type") +
  theme_minimal()
```



Based on the member type, **member riders outnumber casual riders.**

R Script: Number of Rides by Rideable Type

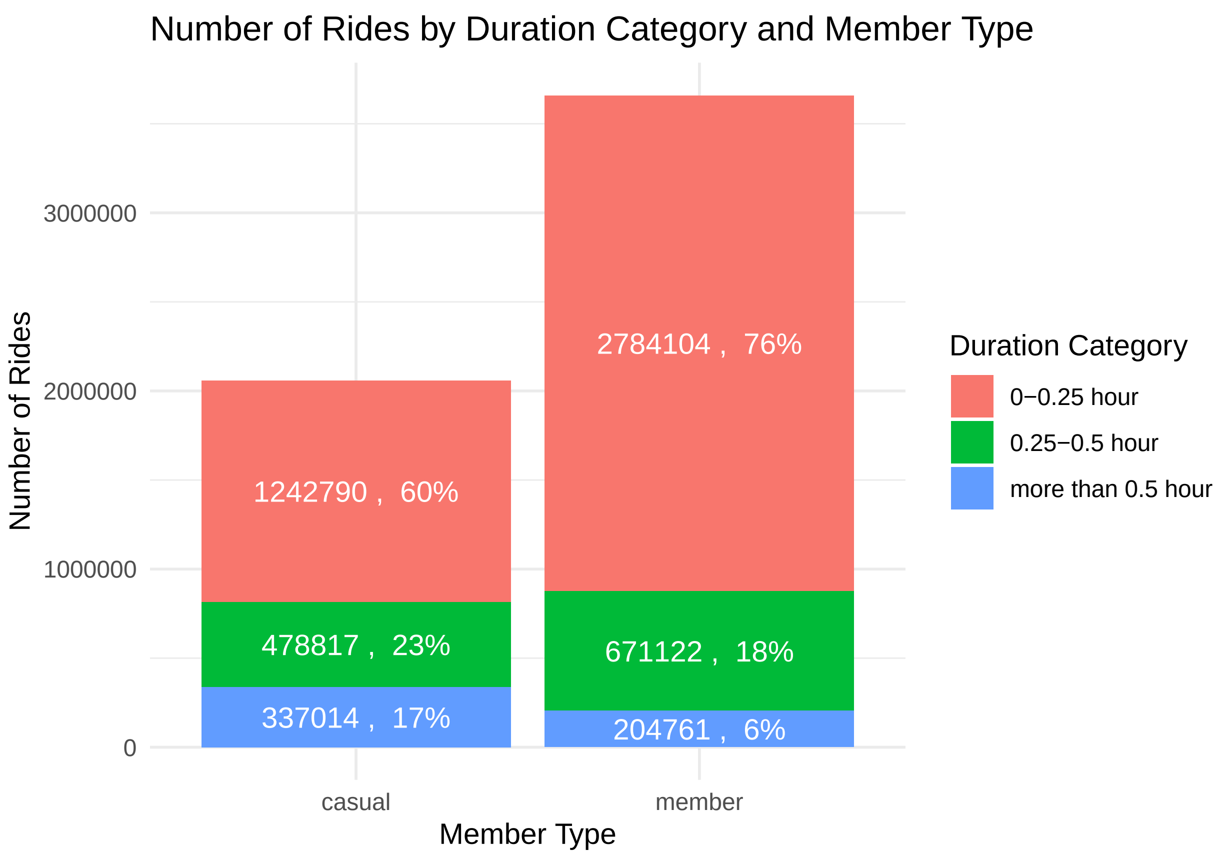
```
rides_by_rideableType <- appended_2023All %>%
  group_by(member_casual, rideable_type) %>%
  summarise(count = n()) %>%
  mutate(percentage = paste0(round_percent(count / sum(count) * 100), "%")) %>%
  mutate(label_text = paste(count, ", ", percentage))

ggplot(rides_by_rideableType,
       aes(x = member_casual, y = count, fill = reorder(rideable_type, -count))) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = label_text), position=position_stack (vjust = 0.5),
            colour = "white") +
  labs(title = "Number of Rides by Rideable Type",
       x = "Member Type",
       y = "Number of Rides",
       fill = "Rideable Type") +
  theme_minimal()
```



Considering the rideable type, **the utilization is dominated by classic and electric bikes**, with almost equal usage. However, docked bikes have the least utilization and are primarily used by casual riders.

R Script: Number of Rides by Duration Category and Member Type

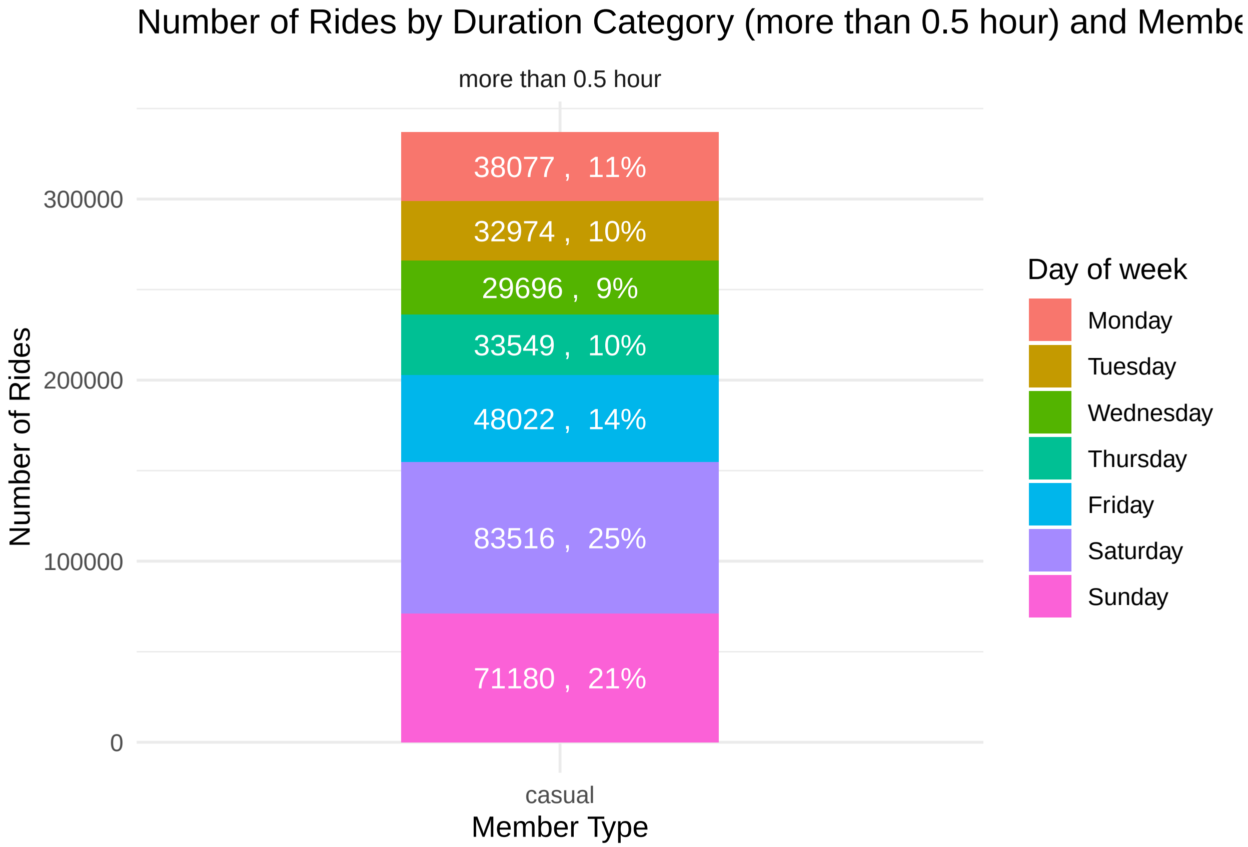
```r
rides_by_duration <- appended_2023All %>%
  filter(!is.na(duration_category)) %>%
  group_by(member_casual, duration_category) %>%
  summarise(count = n()) %>%
  mutate(percentage = paste0(round_percent(count / sum(count) * 100), "%")) %>%
  mutate(label_text = paste(count, ", ", percentage))


ggplot(rides_by_duration,
       aes(x = member_casual, y = count, fill = reorder(duration_category,-count))) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = label_text), position=position_stack(vjust = 0.5),
            colour = "white") +
  labs(title = "Number of Rides by Duration Category and Member Type",
       x = "Member Type",
       y = "Number of Rides",
       fill = "Duration Category") +
  theme_minimal()
```



Analyzing the duration category, the majority of bike usage falls within the 0 to 0.25-hour range for both member and casual riders. Interestingly, **the distribution of the more than 0.5-hour category is higher among casual riders.**

R Script: Number of Rides by Duration Category (more than 0.5 hour) and Member Type

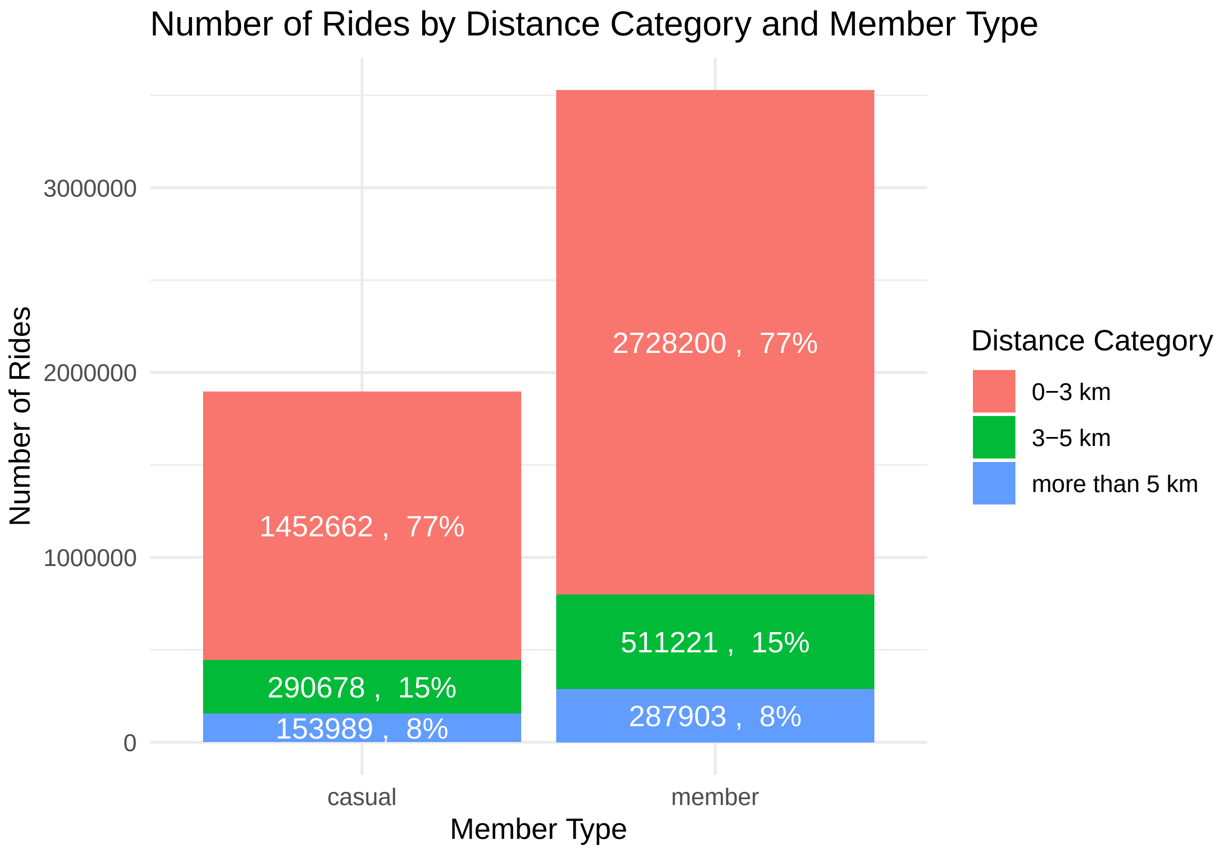
```
rides_by_duration_cek <- appended_2023All %>%
  filter(!is.na(duration_category),member_casual=="casual") %>%
  filter(duration_category=="more than 0.5 hour") %>%
  group_by(member_casual, duration_category,
           Started_dayOfWeek = factor(Started_dayOfWeek,
           levels = c("Monday", "Tuesday", "Wednesday", "Thursday",
                      "Friday", "Saturday", "Sunday"))) %>%
  summarise(count = n()) %>%
  mutate(percentage = paste0(round_percent(count / sum(count) * 100), "%")) %>%
  mutate(label_text = paste(count, ", ", percentage))

ggplot(rides_by_duration_cek, aes(x = member_casual, y = count, fill = Started_dayOfWeek)) +
  geom_bar(stat = "identity", width = 0.45) +
  geom_text(aes(label = label_text), position=position_stack(vjust = 0.5), colour = "white") +
  labs(title = "Number of Rides by Duration Category (more than 0.5 hour) and Member Type",
       x = "Member Type",
       y = "Number of Rides",
       fill = "Day of week") +
  theme_minimal()+
  facet_wrap(~duration_category)
```



Further investigation reveals that this usage pattern is prominent on Saturdays and Sundays, indicating recreational use.

R Script: Number of Rides by Distance Category and Member Type

```r
rides_by_distance <- appended_2023All %>%
  filter(!is.na(dist_km_category)) %>%
  group_by(member_casual, dist_km_category) %>%
  summarise(count = n()) %>%
  group_by(member_casual) %>%
  mutate (percentage = paste0(round_percent(count / sum(count) * 100), "%")) %>%
  mutate(label_text = paste(count, ", ", percentage))


ggplot(rides_by_distance, aes(x = member_casual, y = count, fill = dist_km_category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = label_text), position=position_stack(vjust = 0.5),
            colour = "white") +
  labs(title = "Number of Rides by Distance Category and Member Type",
       x = "Member Type",
       y = "Number of Rides",
       fill = "Distance Category") +
  theme_minimal()
```

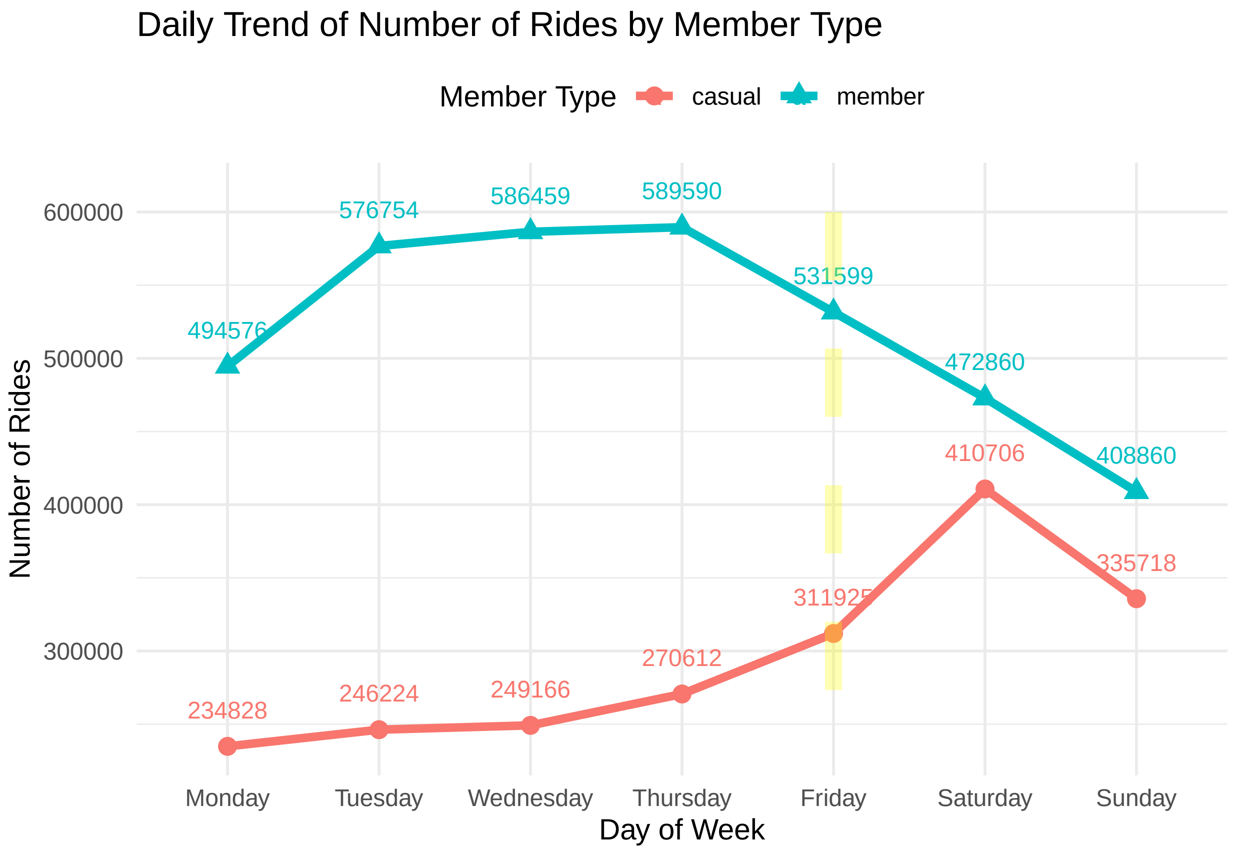


Examining the distance category, **most riders cover distances of 0-3 km**, and the distribution is similar across member types.

- **Number of Rides Trend**

R Script: Daily Trend of Number of Rides by Member Type

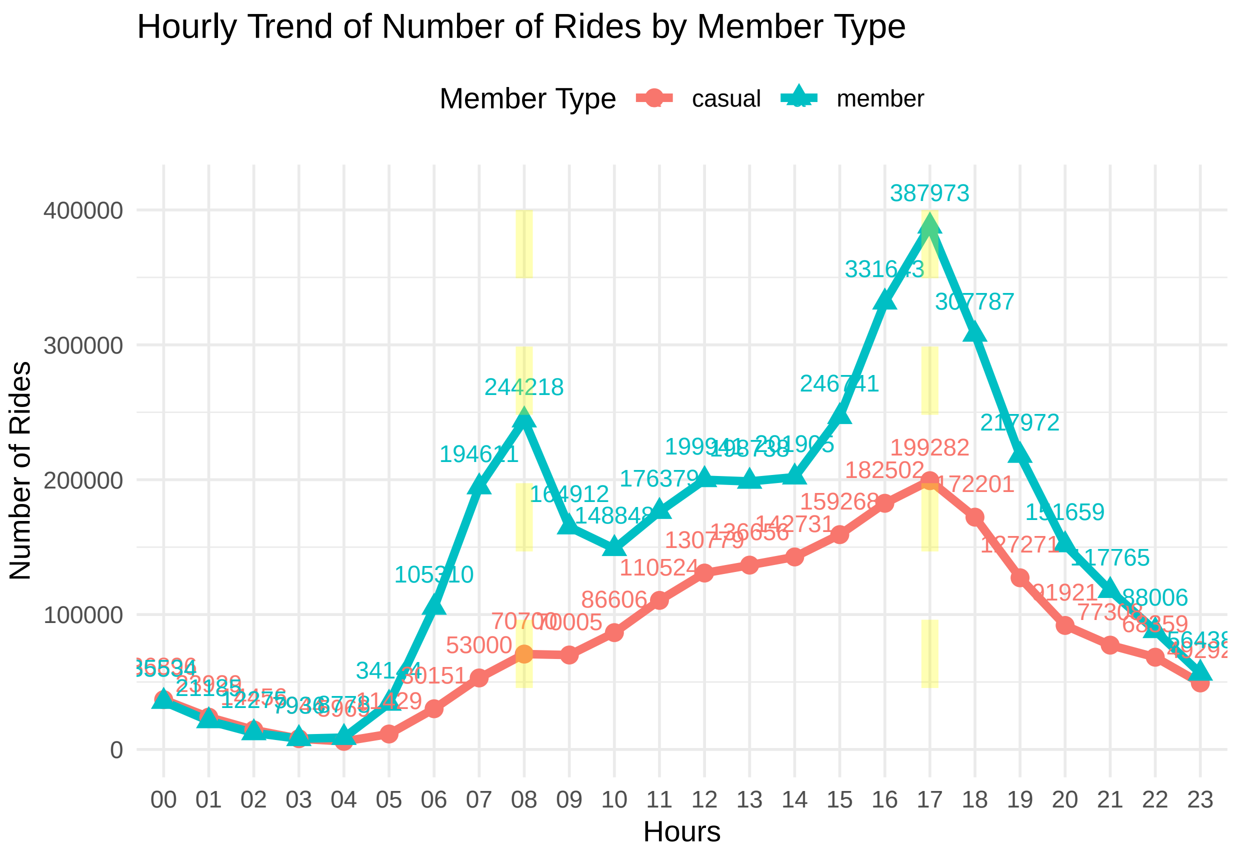
```r
rides_daily <- appended_2023All %>%
  mutate(Started_dayOfWeek = factor(Started_dayOfWeek,
  levels = c("Monday", "Tuesday", "Wednesday", "Thursday",
             "Friday", "Saturday", "Sunday"))) %>%
  group_by(member_casual, Started_dayOfWeek) %>%
  summarise(count = n()) %>% arrange(Started_dayOfWeek)

ggplot(rides_daily, aes(x = Started_dayOfWeek, y = count, group = member_casual,
                        shape = member_casual, color = member_casual)) +
  geom_line (stat = "identity", linewidth= 1.5) + geom_point(size= 3) +
  geom_text(aes(label = count), position = position_nudge(y=25000), size=3) +
  labs(title = "Daily Trend of Number of Rides by Member Type",
       x = "Day of Week",
       y = "Number of Rides", shape = "Member Type", color = "Member Type") +
  annotate("segment", x = "Friday", xend = "Friday", y = 600000, yend = 234000,
           colour = "yellow", linewidth=3, alpha=0.3, linetype = "dashed")+
  theme_minimal() + theme(legend.position = "top")
```



Daily riding trends reveal significant differences between casual and member riders. **Member riders predominantly use bikes on weekdays, suggesting commuting to work, while casual riders exhibit the opposite trend.** This indicates that a majority of member riders rent bikes for commuting to their workplaces.

R Script: Hourly Trend of Number of Rides by Member Type

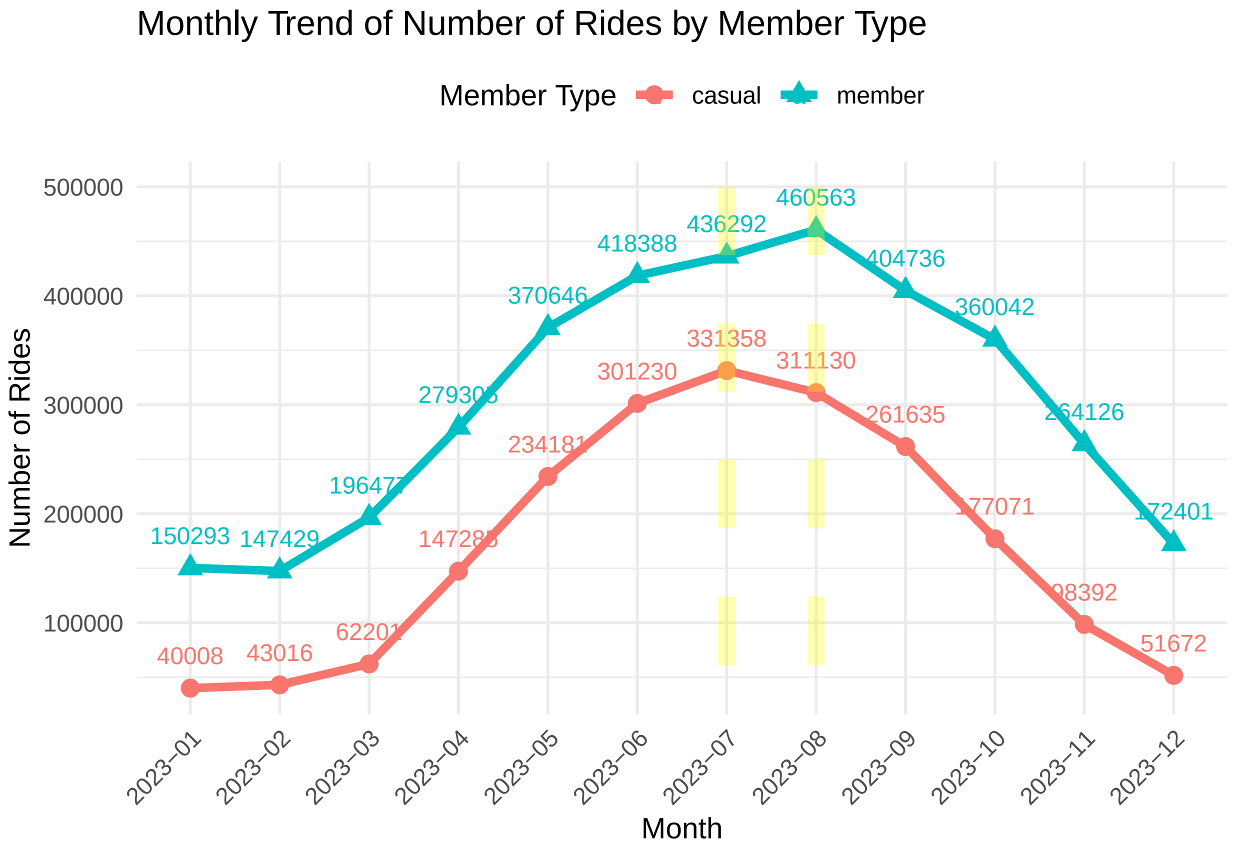
```
rides_hourly <- appended_2023All %>%
  mutate(member_casual, started_hour) %>%
  group_by(member_casual, started_hour) %>%
  summarise(count = n()) %>%
  arrange(started_hour)

ggplot(rides_hourly, aes(x = started_hour, y = count, group = member_casual,
                         shape = member_casual, color = member_casual)) +
  geom_line (stat = "identity", linewidth= 1.5) +
  geom_point(size= 3) +
  geom_text(aes(label = count), position = position_nudge(y=25000), size=3) +
  labs(title = "Hourly Trend of Number of Rides by Member Type",
       x = "Hours",
       y = "Number of Rides", shape = "Member Type", color = "Member Type") +
  annotate("segment", x = 9, xend = 9, y = 400000, yend = 0, colour = "yellow",
           linewidth=3, alpha=0.3, linetype = "dashed")+
  annotate("segment", x = 18, xend = 18, y = 400000, yend = 0, colour = "yellow",
           linewidth=3, alpha=0.3, linetype = "dashed")+
  theme_minimal() +
  theme(legend.position = "top")
```



Hourly riding trends strengthen the observation that member riders predominantly use bikes for commuting. **Peak usage occurs at 8 AM and 5 PM, corresponding to typical work commute hours.**

R Script: Monthly Trend of Number of Rides by Member Type

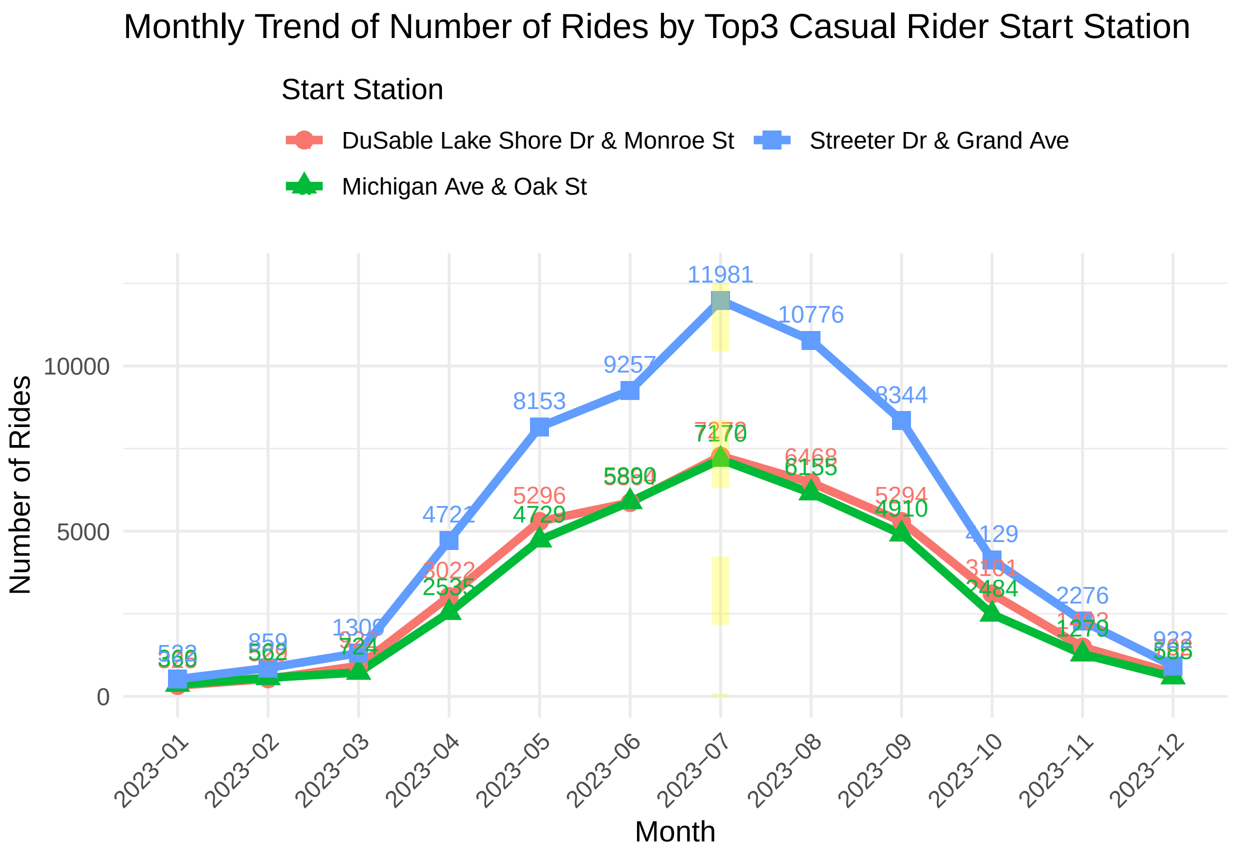
```r
rides_monthly <- appended_2023All %>%
  group_by(member_casual, started_year_month) %>%
  summarise(count = n())


ggplot(rides_monthly, aes(x = started_year_month, y = count, group = member_casual,
                          shape = member_casual, color = member_casual)) +
  geom_line (stat = "identity", linewidth= 1.5) +
  geom_point(size= 3) +
  geom_text(aes(label = count), position = position_nudge(y=30000), size=3) +
  labs(title = "Monthly Trend of Number of Rides by Member Type",
       x = "Month",
       y = "Number of Rides", shape = "Member Type", color = "Member Type") +
  annotate("segment", x = "2023-07", xend = "2023-07", y = 500000, yend = 39000,
           colour = "yellow", linewidth=3, alpha=0.3, linetype = "dashed")+
  annotate("segment", x = "2023-08", xend = "2023-08", y = 500000, yend = 39000,
           colour = "yellow", linewidth=3, alpha=0.3, linetype = "dashed")+
  theme_minimal() +
  theme(legend.position = "top") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Monthly riding trends show that **casual riders peak in July**, while member riders peak in August.

R Script: Monthly Trend of Number of Rides by Top3 Casual Rider Start Station

```r
# Start station name top 3 casual rider only

rides_by_startStation_top3_casual <- appended_2023All %>%
  filter(start_station_name != "", member_casual=="casual") %>%
  group_by(member_casual, start_station_name) %>%
  summarise(count = n()) %>%
  top_n(3, wt = count)

# monthly ride by station

rides_monthly_byStation <- appended_2023All %>%
  filter(!is.na(start_station_name) & start_station_name != "") %>%
  group_by(start_station_name, started_year_month) %>%
  summarise(count = n())


# Inner join rides_by_startStation_top3 casual & rides_monthly_byStation

rides_monthly_byStationTop3 <- inner_join(rides_by_startStation_top3_casual,
                               rides_monthly_byStation, by = "start_station_name")


ggplot(rides_monthly_byStationTop3, aes(x = started_year_month,
  y = count.y, group = start_station_name, shape = start_station_name,
  color = start_station_name)) +
  geom_line (stat = "identity", linewidth= 1.5) +
  geom_point(size= 3) +
  geom_text(aes(label = count.y), position = position_nudge(y=800), size=3) +
  labs(title = "Monthly Trend of Number of Rides by Top3 Casual Rider Start Station",
       x = "Month",
       y = "Number of Rides",
       shape = "Start Station",
       color = "Start Station") +
  annotate("segment", x = "2023-07", xend = "2023-07", y = 12500, yend = 0,
           colour = "yellow", linewidth=3, alpha=0.3, linetype = "dashed")+
  theme_minimal() +
    theme(legend.position = "top") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  guides(
    shape = guide_legend (title = "Start Station", direction = "vertical", ncol = 2),
    color = guide_legend (title = "Start Station", direction = "vertical", ncol = 2)
  )
```



Among the top 3 start stations for casual riders, **Streeter Dr & Grand Ave has the highest ride count, followed by DuSable Lake Shore Dr & Monroe St. Michigan Ave & Oak St takes the third spot.**

6. My top three recommendations based on my analysis

- **Based on the daily and hourly ride trends, where member riders predominantly bike on weekdays:** To encourage casual riders to convert to members, it is recommended to introduce weekday passes in the pricing plans. These passes should be priced more affordably compared to daily rentals for casual riders. The hope is that after trying this pricing plan, casual riders will find it more economical and eventually convert to becoming member riders.

- **Considering the daily trends, especially the higher and longer bike usage by casual riders on Saturdays and Sundays:** It is recommended to offer discounts specifically on these two days. This strategy aims to attract new casual customers or even individuals who may not have a strong inclination towards biking. Offering discounts on weekends can create a sense of community and encourage more people to take up biking, especially on Saturdays and Sundays.

- **Examining the monthly trends, with casual riders reaching their peak in July:** The recommendation is to implement a substantial discount during this month. The goal is to acquire a significant number of new casual customers. A proactive promotional campaign on social media, starting one month in advance (in June), can highlight the benefits of biking for cardiovascular health. If targeting all stations is not feasible, focus on the top 3 stations with the highest ride frequency.