Cyclistic Bike-Share Analysis: How to Navigate Speedy Success (Capstone Case Study)

Tyrone Hubbard

2025-03-24

Introduction

Welcome to my Google Data Analytics Capstone Cyclistic bike-share analysis case study! In this analysis, I assume the role of a junior data analyst working for Cyclistic, a fictional bike-share company in Chicago. The primary goal is to understand how annual members and casual riders use Cyclistic bikes differently. The insights gained will be used to develop a new marketing strategy aimed at converting casual riders into annual members, a key objective for the company's future success.

Cyclistic launched its successful bike-share program in 2016 (Cyclistic Case Study). The program has since grown to include a fleet of 5,824 geotracked bicycles across 692 stations in Chicago (Cyclistic Case Study). Cyclistic distinguishes itself by offering a variety of bike options, including reclining bikes, hand tricycles, and cargo bikes, catering to a diverse range of riders (Cyclistic Case Study). While Cyclistic users are more likely to ride for leisure, about 30% use the bikes to commute to work each day (Cyclistic Case Study).

Cyclistic's marketing strategy has historically focused on building general awareness and appealing to broad consumer segments, enabled by flexible pricing plans such as single-ride passes, full-day passes, and annual memberships (Cyclistic Case Study). Casual riders are those who purchase single-ride or full-day passes, while annual members are those who purchase annual memberships (Cyclistic Case Study).

Business Task

The main business task is to analyze Cyclistic's historical bike trip data to identify trends and differences in how annual members and casual riders use Cyclistic bikes (Cyclistic Case Study). This analysis will inform the development of targeted marketing strategies to convert casual riders into annual members.

Data Sources and Preparation

The analysis uses Cyclistic's historical trip data. Specifically, the Divvy 2019 Q1 and Divvy 2020 Q1 datasets were used (Cyclistic Case Study). It is important to note that Cyclistic is a fictional company and the datasets are appropriate for this case study. The data was provided by Motivate International Inc. (Cyclistic Case Study). Data privacy considerations prevent the use of riders' personally identifiable information.

The datasets were prepared for analysis using R. This involved:

- Standardizing column names.
- Combining the datasets into a single dataframe.
- Converting date and time columns to appropriate formats.
- Calculating ride length.
- Creating columns for day of the week, month, and year.
- Removing or filtering out irrelevant or erroneous data (e.g., negative ride lengths, extremely long ride lengths).

```r
divvy_2019 <- read_csv("data/Divvy_2019_Q1.csv")
```

```
## Rows: 365069 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): start_time, end_time, start_station_name, end_station_name, user_ty...
## dbl (4): trip_id, bikeid, start_station_id, end_station_id
## num (1): tripduration
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
divvy_2020 <- read_csv("data/Divvy_2020_Q1.csv")
```

```
## Rows: 426886 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): trip_id, rideable_type, start_time, end_time, start_station_name, e...
## dbl (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, en...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
names(divvy_2019)
```

```
##  [1] "trip_id"            "start_time"         "end_time"
##  [4] "bikeid"             "tripduration"       "start_station_id"
##  [7] "start_station_name" "end_station_id"     "end_station_name"
## [10] "user_type"          "gender"             "birthyear"
```

```r
unique(divvy_2019$user_type)
```

```
## [1] "Member" "Casual"
```

```r
table(divvy_2019$gender)
```

```
##
##  Female    Male Unknown
##   66918  278440   19711
```

```r
table(divvy_2019$birthyear)
```

```
##
##    1900    1901    1918    1921    1931    1933    1934    1938    1939    1940
##      69       2      40      60       2       7       5       1      13      29
##    1941    1942    1943    1944    1945    1946    1947    1948    1949    1950
##       1      51      27      23     134     307      93     332     232     464
##    1951    1952    1953    1954    1955    1956    1957    1958    1959    1960
##    1132    1096     930    1384    1904    2052    2136    2376    2377    3336
##    1961    1962    1963    1964    1965    1966    1967    1968    1969    1970
##    3534    4508    3449    4569    3778    4663    3499    4408    5923    5789
##    1971    1972    1973    1974    1975    1976    1977    1978    1979    1980
##    4023    5031    3973    5842    4591    5897    8116    6523    7432    9247
##    1981    1982    1983    1984    1985    1986    1987    1988    1989    1990
##    9733   10641   10559   12835   13411   13210   17155   14588   21014   17410
##    1991    1992    1993    1994    1995    1996    1997    1998    1999    2000
##   16146   18924   14197   11450    8559    5335    1874    1672    1894     826
```

```
##    2001    2002    2003 Unknown
##     118      83       2   18023
```

```r
summary(divvy_2019)
```

```
##     trip_id           start_time          end_time             bikeid
##  Min.   :21742443   Length:365069      Length:365069      Min.   :   1
##  1st Qu.:21848765   Class :character   Class :character   1st Qu.:1777
##  Median :21961829   Mode  :character   Mode  :character   Median :3489
##  Mean   :21960872                                         Mean   :3429
##  3rd Qu.:22071823                                         3rd Qu.:5157
##  Max.   :22178528                                         Max.   :6471
##   tripduration      start_station_id start_station_name end_station_id
##  Min.   :      61   Min.   :  2.0    Length:365069      Min.   :  2.0
##  1st Qu.:     326   1st Qu.: 76.0    Class :character   1st Qu.: 76.0
##  Median :     524   Median :170.0    Mode  :character   Median :168.0
##  Mean   :    1016   Mean   :198.1                       Mean   :198.6
##  3rd Qu.:     866   3rd Qu.:287.0                       3rd Qu.:287.0
##  Max.   :10628400   Max.   :665.0                       Max.   :665.0
##  end_station_name    user_type            gender           birthyear
##  Length:365069      Length:365069      Length:365069      Length:365069
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
names(divvy_2020)
```

```
##  [1] "trip_id"            "rideable_type"      "start_time"
##  [4] "end_time"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "user_type"
```

```r
unique(divvy_2020$user_type)
```

```
## [1] "Casual" "Member"
```

```r
summary(divvy_2020)
```

```
##    trip_id          rideable_type       start_time          end_time
##  Length:426886      Length:426886      Length:426886      Length:426886
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  start_station_name start_station_id end_station_name   end_station_id
##  Length:426886      Min.   :  2.0    Length:426886      Min.   :  2.0
##  Class :character   1st Qu.: 77.0    Class :character   1st Qu.: 77.0
##  Mode  :character   Median :176.0    Mode  :character   Median :175.0
##                     Mean   :209.8                       Mean   :209.3
##                     3rd Qu.:298.0                       3rd Qu.:297.0
##                     Max.   :675.0                       Max.   :675.0
##    start_lat       start_lng         end_lat         end_lng
##  Min.   :41.74   Min.   :-87.77   Min.   :41.74   Min.   :-87.77
```

```
##  1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66
##  Median :41.89   Median :-87.64   Median :41.89   Median :-87.64
##  Mean   :41.90   Mean   :-87.64   Mean   :41.90   Mean   :-87.64
##  3rd Qu.:41.92   3rd Qu.:-87.63   3rd Qu.:41.92   3rd Qu.:-87.63
##  Max.   :42.06   Max.   :-87.55   Max.   :42.06   Max.   :-87.55
##   user_type
##  Length:426886
##  Class :character
##  Mode  :character
##
##
##
```

```r
divvy_2019 <- divvy_2019 %>%
  select(trip_id, start_time, end_time, start_station_name, start_station_id, end_station_name, end_sta

divvy_2019$rideable_type <- NA
divvy_2019$start_lat <- NA
divvy_2019$start_lng <- NA
divvy_2019$end_lat <- NA
divvy_2019$end_lng <- NA

divvy_2019 <- divvy_2019 %>%
  select(trip_id, rideable_type, start_time, end_time, start_station_name, start_station_id, end_statio

divvy_2019$trip_id <- as.character(divvy_2019$trip_id)
divvy_2020$trip_id <- as.character(divvy_2020$trip_id)
all_trips <- bind_rows(divvy_2019, divvy_2020)

str(all_trips$trip_id)
```

```
##  chr [1:791955] "21742443" "21742444" "21742445" "21742446" "21742447" ...
```

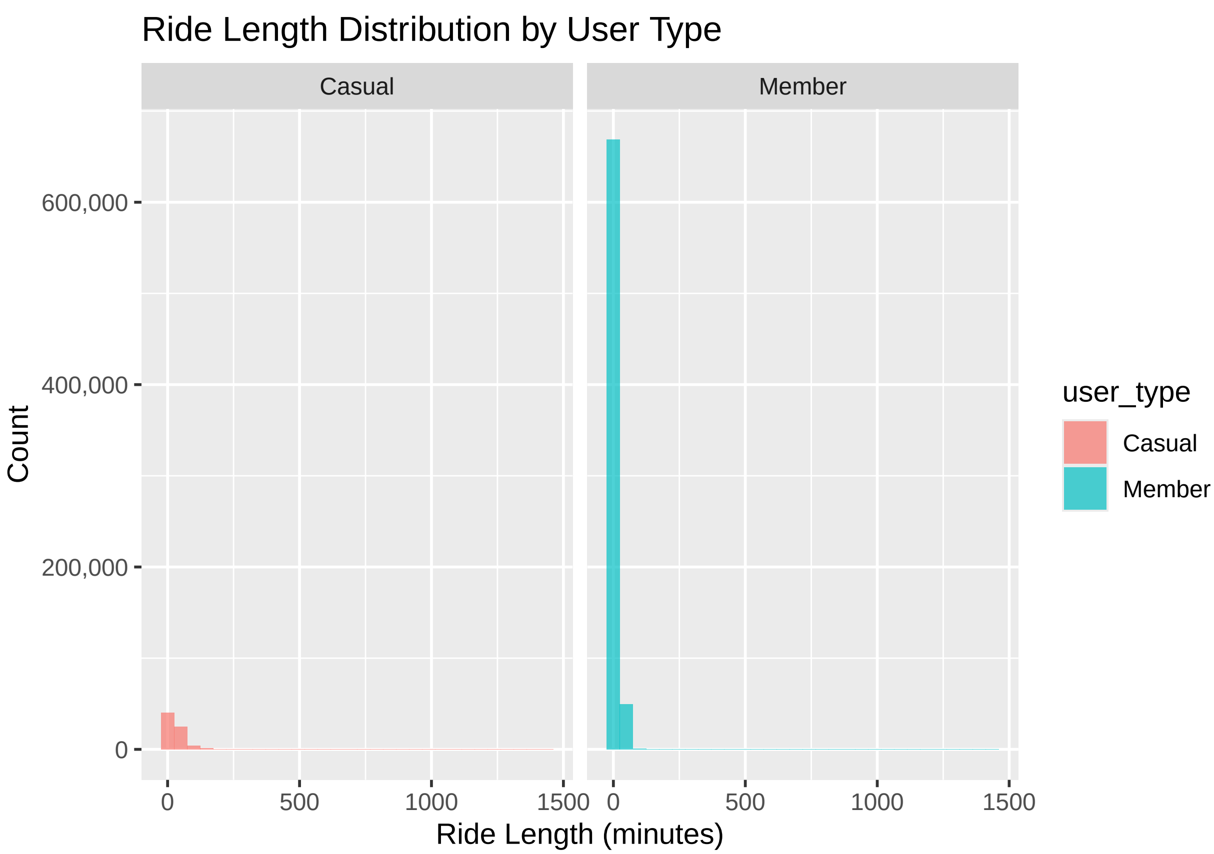
```r
#Convert start_time and end_time to datetime objects:
all_trips$start_time <- ymd_hms(all_trips$start_time)
all_trips$end_time <- ymd_hms(all_trips$end_time)

#Calculate ride_length in minutes:
all_trips$ride_length <- difftime(all_trips$end_time, all_trips$start_time, units = "mins")
all_trips$ride_length <- as.numeric(all_trips$ride_length)

#Create columns for day of week, month, and year:
all_trips$day_of_week <- wday(all_trips$start_time, label = TRUE)
all_trips$month <- month(all_trips$start_time, label = TRUE)
all_trips$year <- year(all_trips$start_time)

#Remove negative ride_length values (if any):
all_trips <- all_trips %>% filter(ride_length >= 0)

#Remove any rows where the ride_length is extremely long. These trips are likely data errors.
all_trips <- all_trips %>% filter(ride_length < 1440)
```

Analysis

The analysis focused on comparing ride patterns between annual members and casual riders. Key metrics and visualizations include:

- **Ride Length Distribution:** Histograms were used to visualize the distribution of ride lengths for each user type.

```
#   Ride Length Distribution by User Type
ggplot(all_trips, aes(x = ride_length, fill = user_type)) +
  geom_histogram(bins = 30, alpha = 0.7) +
  facet_wrap(~user_type) +
  labs(title = "Ride Length Distribution by User Type", x = "Ride Length (minutes)", y = "Count") +
  scale_y_continuous(labels = scales::comma)
```

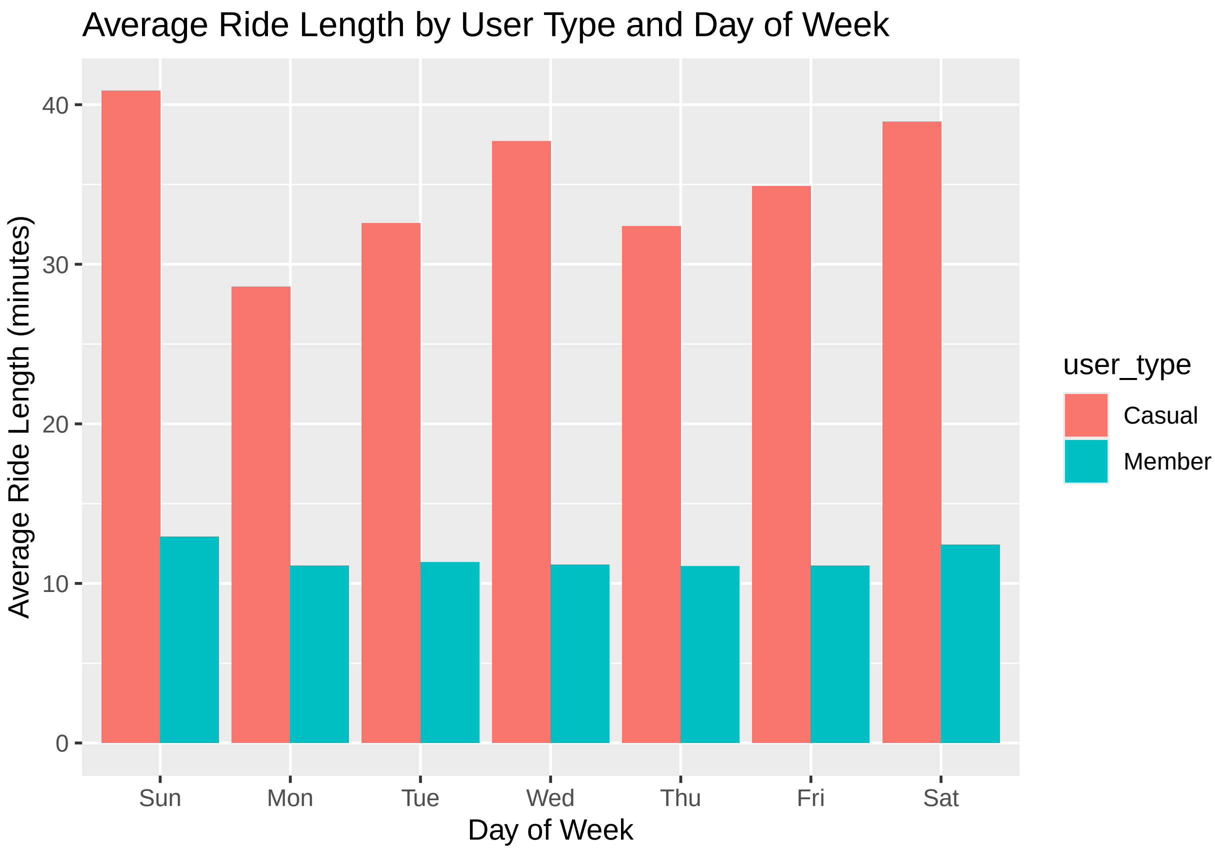


- **Average Ride Length by Day of Week:** Bar plots were used to compare average ride lengths by user type across different days of the week.

```
#   Average Ride Length by User Type and Day of Week
average_ride_length_by_day <- all_trips %>%
  group_by(user_type, day_of_week) %>%
  summarize(avg_ride_length = mean(ride_length))
```

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

```
ggplot(average_ride_length_by_day, aes(x = day_of_week, y = avg_ride_length, fill = user_type)) +
  geom_col(position = "dodge") +
  labs(title = "Average Ride Length by User Type and Day of Week", x = "Day of Week", y = "Average Ride
```

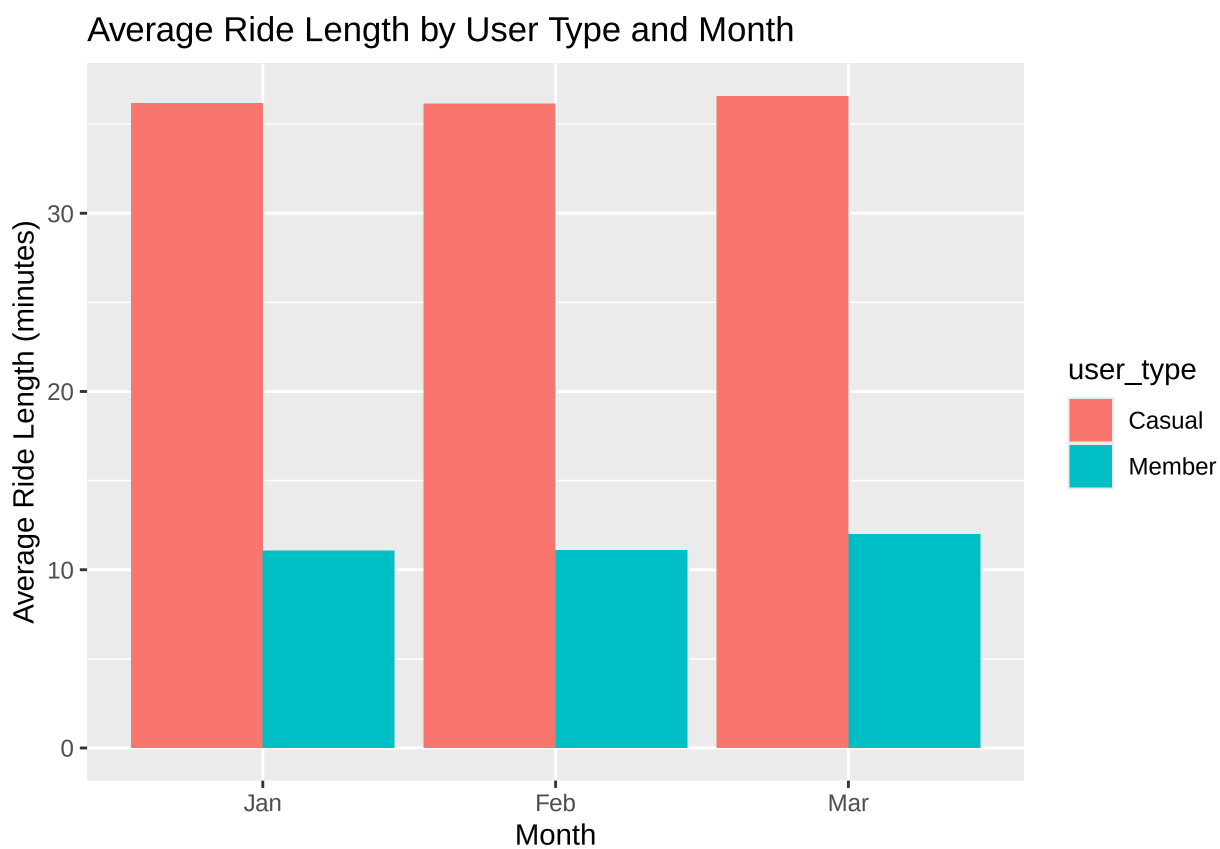


- **Average Ride Length by Month:** Bar plots were used to compare average ride lengths by user type across different months.

```
#   Average Ride Length by User Type and Month
average_ride_length_by_month <- all_trips %>%
  group_by(user_type, month) %>%
  summarize(avg_ride_length = mean(ride_length))
```

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

```
ggplot(average_ride_length_by_month, aes(x = month, y = avg_ride_length, fill = user_type)) +
  geom_col(position = "dodge") +
  labs(title = "Average Ride Length by User Type and Month", x = "Month", y = "Average Ride Length (min
```

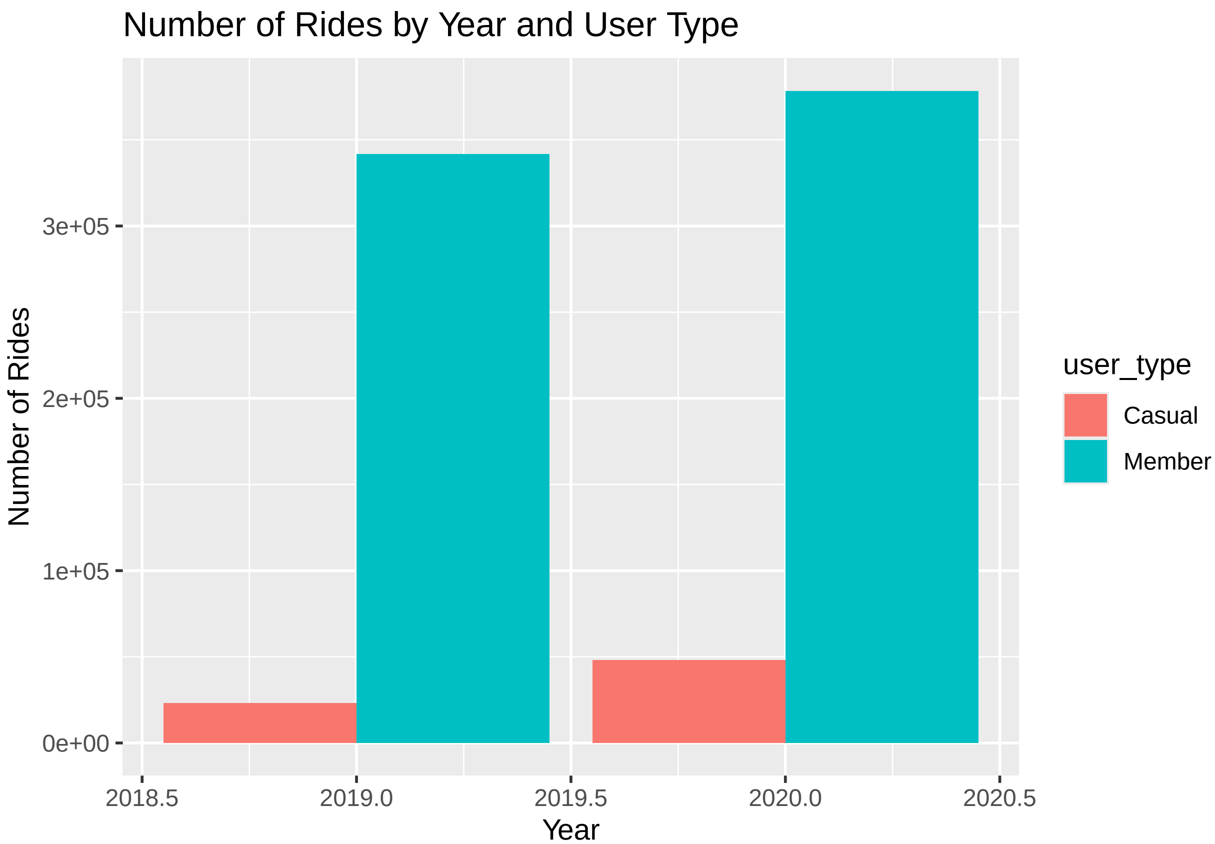


- **Number of Rides by Year:** Bar plots were used to visualize the number of rides by user type for 2019 and 2020.

```r
#   Number of Rides by Year and User Type
rides_by_year <- all_trips %>%
  group_by(user_type, year) %>%
  summarize(number_of_rides = n())
```

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

```r
ggplot(rides_by_year, aes(x = year, y = number_of_rides, fill = user_type)) +
  geom_col(position = "dodge") +
  labs(title = "Number of Rides by Year and User Type", x = "Year", y = "Number of Rides")
```



- **Top 10 Start Stations:** Bar plots were used to visualize the top 10 start stations for each user type.

```
#   Top 10 Start Stations by User Type
top_start_stations <- all_trips %>%
  group_by(user_type, start_station_name) %>%
  summarize(number_of_rides = n()) %>%
  arrange(desc(number_of_rides)) %>%
  filter(start_station_name != "NA") %>%
  top_n(10, number_of_rides)
```

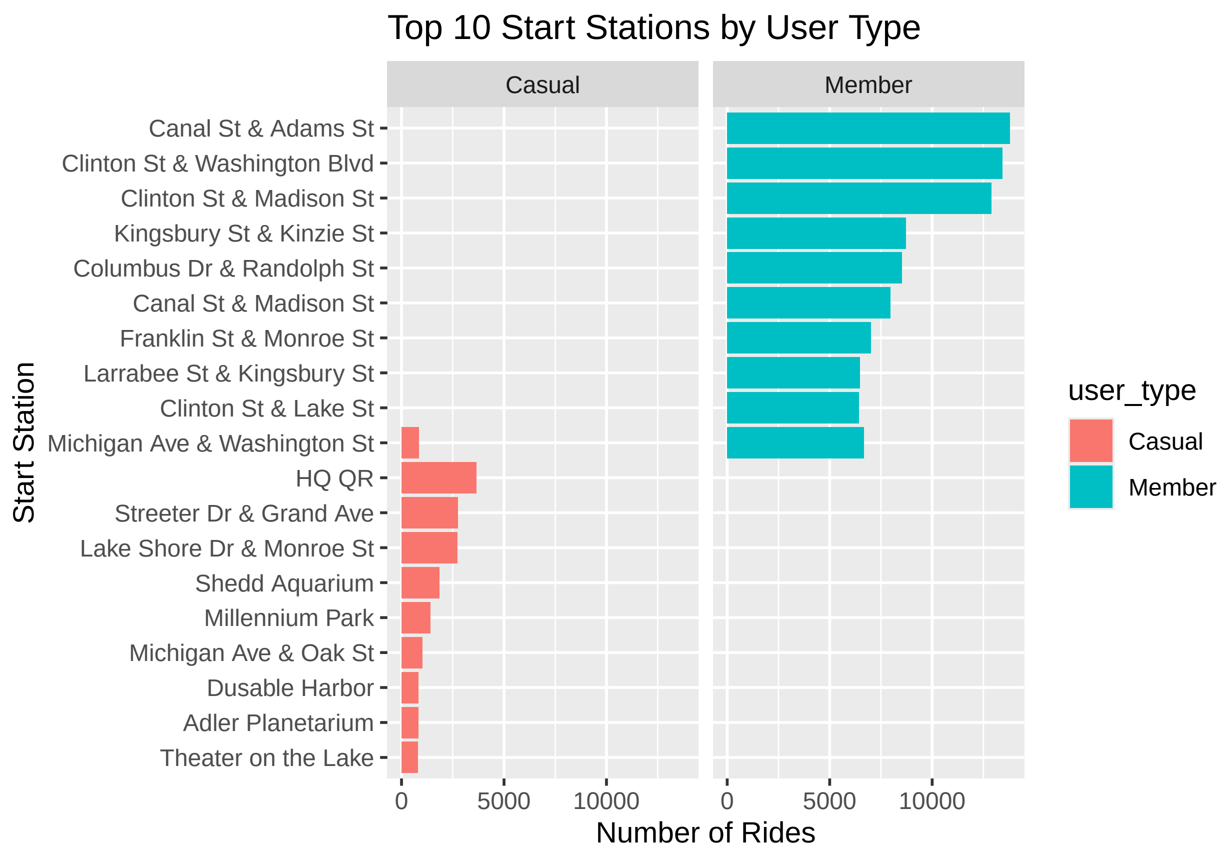
```
## `summarise()` has grouped output by 'user_type'. You can override using the
## `.groups` argument.
```

```
print(top_start_stations)
```

```
## # A tibble: 20 x 3
## # Groups:   user_type [2]
##    user_type start_station_name           number_of_rides
##    <chr>     <chr>                                  <int>
##  1 Member    Canal St & Adams St                    13797
##  2 Member    Clinton St & Washington Blvd           13434
##  3 Member    Clinton St & Madison St                12889
##  4 Member    Kingsbury St & Kinzie St                8718
##  5 Member    Columbus Dr & Randolph St               8515
##  6 Member    Canal St & Madison St                   7956
##  7 Member    Franklin St & Monroe St                 7008
##  8 Member    Michigan Ave & Washington St            6684
##  9 Member    Larrabee St & Kingsbury St              6466
## 10 Member    Clinton St & Lake St                    6437
```

```
## 11 Casual    HQ QR                                      3649
## 12 Casual    Streeter Dr & Grand Ave                    2741
## 13 Casual    Lake Shore Dr & Monroe St                  2727
## 14 Casual    Shedd Aquarium                             1831
## 15 Casual    Millennium Park                            1403
## 16 Casual    Michigan Ave & Oak St                      1017
## 17 Casual    Michigan Ave & Washington St                835
## 18 Casual    Dusable Harbor                              829
## 19 Casual    Adler Planetarium                           826
## 20 Casual    Theater on the Lake                         794
```

```r
ggplot(top_start_stations, aes(x = reorder(start_station_name, number_of_rides), y = number_of_rides, fi
  geom_col(position = "dodge") +
  coord_flip() +
  facet_wrap(~user_type) +
  labs(title = "Top 10 Start Stations by User Type", x = "Start Station", y = "Number of Rides")
```



Findings

The analysis reveals several key differences in how annual members and casual riders use Cyclistic bikes:

- **Ride Length:** Casual riders take significantly longer rides than annual members. The ride length distribution for casual riders is right-skewed, indicating a mix of short and long rides, while annual members' rides are concentrated at shorter durations.
- **Day of Week Patterns:** Casual riders exhibit the longest rides on weekends, suggesting recreational use. Annual members maintain relatively consistent ride lengths throughout the week, indicating commuting to work patterns.
- **Monthly Patterns:** Ride length patterns remain consistent across months for both user types, with

casual riders consistently taking longer rides.
- **Ridership Growth:** Annual membership saw a substantial increase from 2019 to 2020, dwarfing the growth in casual ridership.
- **Start Station Locations:** Annual members primarily start their rides from stations in downtown Chicago, likely for commuting. Casual riders start from a wider range of stations, including those near recreational areas and tourist attractions.

Summary

This analysis of Cyclistic bike-share data highlights distinct usage patterns between annual members and casual riders. Annual members primarily use the service for short, consistent commutes, concentrated around downtown Chicago. Casual riders, on the other hand, use the service for longer, more varied rides, often for recreational purposes, particularly on weekends and near tourist locations. The significant growth in annual membership from 2019 to 2020 underscores the importance of this user group to Cyclistic's business.

Recommendations

Based on the analysis, the following recommendations are made to convert casual riders into annual members:

1. **Targeted Marketing Campaigns:** Develop marketing campaigns that appeal to the recreational and leisure-oriented usage patterns of casual riders. Highlight the value of annual memberships for weekend trips, exploring the city, and enjoying scenic rides.
2. **Incentivize Weekend/Leisure Use:** Offer membership perks or discounts specifically for weekend or off-peak usage, aligning with casual riders' longer, leisure-focused trips.
3. **Promote Convenience and Value:** Emphasize the convenience and value of annual memberships for regular use, even beyond commuting. This could include highlighting the ease of access, cost savings for frequent riders, and the availability of bikes for various purposes.

Conclusion

Understanding the distinct usage patterns of annual members and casual riders is crucial for Cyclistic's growth strategy. By tailoring marketing efforts to appeal to the specific needs and behaviors of casual riders, Cyclistic can effectively increase annual memberships and drive sustainable success.

References

Google Data Analytics Capstone: Cyclistic Case Study. (Accessed March 23, 2025). Case study: How does a bike-share navigate speedy success? Cyclistic.