Cyclistic Analytics Report

Ruby Smith

2024-10-24

Purpose: Compare & Contrast Annual Members vs. Casual Riders

The Cyclistic Team is looking to increase annual memberships. Marketing team wants me to compare annual memberships to casual riders over the last year and come up with ideas to convert their casual rides to annual memberships. My goal is to analyze the data and compare casual rides and annual memberships over the last year to help with marketing strategies.

Key Tasks For Case Study:

- Locate and Analyze data of casual riders and annual memberships to understand trends.
- Create visualizations and show insights into analyzed data about riders and communicate it to all stakeholders (Cyclistic, executive team, Marketing manager, and marketing team).
- Help come up with ideas to maximize the number of annual memberships.

Data:
Data collected by Motivate International Inc.: Index of bucket "divvy-tripdata"
https://divvy-tripdata.s3.amazonaws.com/index.html

Data Useable by License: Data License Agreement | Divvys Bikes
https://divvybikes.com/data-license-agreement
The data itself cannot be shared stand-alone, but the data can be shared with my analysis and report.

Data Organized:
12 .csv files named by Year, Month, and file name (ex. 202403-divvy-tripdata). The .csv files contain data on each bike trip taken during the specified month in the name of the file. Each file contains 13 columns of data in wide format with several hundred thousands of rows. The Primary Key for each dataset is the 'ride_id' column.

Column Names:
ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng, and member_casual.

Bias:
I feel that there is little to no bias in the data. The datasets contain all rides, it is reliable, current, and from the company itself. The only issue is some locations for the starting_station name and id and also the ending_station name and id, there is nothing and is marked as 'null.'

Integrity:
It looks like all data imported correctly. All .csv files imported contain the same number of columns and are all named the same and have matching data types. Datasets are kept private and only I can have access to them.

Prepare Phase Conclusion and Insights:

The datasets contain information that will help me find trends between casual riders and memberships. I have data that shows types of bikes used, how long the rides took, and starting and ending locations.
There are some areas I have noticed that do not have names or id's for starting locations or ending locations and some latitude and longitude is not entered in as well. I will explore the datasets more thoroughly to understand. Possible hypothesis: error reading some locations and possibly not turning in bike at end of route.
I want to combine all data into one place so it can be easily accessed and updated when necessary. I want to get trip time and compare memberships and filter the dataset with other categories. I want to separate the date and time and look for trends in the days of the week. I want to compare latitude and longitude on Tableau for any geographical trends. I also plan to separate categories by bike type as well for any trends.

Tools used for Analysis:

- Google Cloud - imported .csv to cloud because of file size
- BigQuery - imported .csv from cloud to BigQuery to analyze data with SQL
- Tableau - to create visualizations
- Posit - for R programming creating graphs and report
- GitHub - to store queries used in SQL
- Google Sheets - to store tables created from SQL queries to use for visualizations
- Google Search - to help with query organization and R language when stuck
- Notes - from Coursera data analytics course

Changes Made to Original Data:

- Combine September 2023 - August 2024 into one dataset
  rs-project-01-429415.cyclistic_bike_data.annual_202309_202408_tripdata
- Remove latitude and longitude NULLs
- Check for duplicates with the ride_id column, found 171. Remove duplicates and create an updated dataset.
  rs-project-01-429415.cyclistic_bike_data.annual_202309_202408_tripdata2
- Created new column 'trip_time_seconds'
- Separated date from started_time and ended_time
- Created 2 new columns 'start_date' and 'end_date'.
- Create column 'day_of_ride' for weekday names.

Updated Column Names:
ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng, member_casual, trip_time_seconds, start_date, end_date, and day_of_ride.

https://console.cloud.google.com/bigquery?ws=!1m5!1m4!4m3!1srs-project-01-429415!2scyclistic_bike_data!3sannual_202309_202408_tripdata2

Process Phase Conclusion and Insights:

I decided to remove the rows that do not have latitude and longitude values. I want to compare locations geographically and without that information, I would be unable to do so. There were 7526 entries out of 5.6 million (<0.2%). I felt that with what I would like to analyze, taking out less than 0.2% of incomplete data would be ok and it would all still be in original data stored. I have removed 171 duplicated as well and restored integrity to the data. I have completed the tasks I wanted to do from the Prepare Phase and in doing so created 4 new columns.

Analysis:

GitHub Link For All SQL queries: Case_Study_Cyclistic
https://github.com/RubyRene90/Case_Study_Cyclistic/tree/main)

- Part 1: 1_Understanding_DataSet
- Part 2: 2_annual_tripdata_cleaning
- Part 3: 3_annual_tripdata_analyzing
- Part 4: 4_annual_tripdata_analyzing2
- Part 5: 5_annual_tripdata_analyzing3

Google Sheets Document: '202309_202408_CyclisticAnalyzedData'
https://docs.google.com/spreadsheets/d/1DlsFo9a9v79fZ12HbLlfPqG-iBwvY2XONY1JVraHPBs/edit?usp=sharing

- Sheet 1: Annual
- Sheet 2: Monthly
- Sheet 3: Weekday
- Sheet 4: Quarterly
- Sheet 5: Time_Intervals
- Sheet 6: Popular_Locations
- Sheet 7: Q2_Popular_Locations

Beginning analysis phase by running annual aggregate functions. When running MIN on the 'trip_time_seconds', the answer that came back was negative. Also, when running MAX, the answer was over 24 hours. I decided since both results were small when researched, I pulled all data for any time that was negative (344 results) and over 24 hours (389 rows) and removed them from the cleaned dataset.

I created a Table on Sheet 1: Annually with 3 rows and 8 columns.
Column Names: rider_status, rider_count, percent_of_riders, annual_avg_minutes, annual_min, annual_max_minutes, classic_count, electric_count

Continued with running monthly aggregate functions. I created a Table for casuals on Sheet 2: Monthly with 24 rows and 9 columns.
Column Names: month, rider_status, rider_count, percent_of_rider_type, monthly_avg_minutes, min, max_minutes, classic_count, electric_count

Continued with running weekly aggregate functions. I created a Table on Sheet 3: Weekday with 14 rows and 10 columns.
Column Names: weekday, total_weekday_count, rider_status, rider_count, percent_of_total, weekday_avg_minutes, min, max_minutes, classic_count, electric_count

Continued with running Count and Average for quarterly aggregate functions. I created a Table on Sheet 4: Quarterly with 56 rows and 5 columns.
Column Names: rider_status, weekday, rider_count, quarterly_avg_minutes, quarter

I checked different trip time intervals to see any differences between memberships and casuals. I created a Table on Sheet 4: Time_Intervals with 8 rows and 4 columns.
Column Names: time_interval, membership, casual, all_riders

I queried for the top 20 starting and ending stations. I was able to get the starting and ending latitude and longitudes for each of the top results. I created a Table on Sheet 5: Popular Locations with 80 rows and 6 columns.
Column Names: rider_status, location, station_name, rider_count, latitude, longitude

I also queried the top 20 starting and ending stations for the 2nd quarter (December - February). I created a Table on Sheet 6: Q2_Popular_Locations with 80 rows and 6 columns.
Column Names: rider_status, location, station_name, rider_count, latitude, longitude

Analysis Phase Conclusion and Insights:

After running queries in SQL and creating more manageable tables in google sheets, I can see trends in the data between September 2023 - August 2024. Membership rides currently count for 65% of all rides annually. Casual riders (non-membership riders) have a higher annual average in ride time by a 9 minute difference compared to membership riders. Also, there is no significant difference between all_riders, membership, or casual riders when it comes to using classic or electric bikes.
When separating the data on a monthly basis, I noticed both membership and casual ride count drop in the winter months between December 2023 and February 2024. The casual ride count drops more significantly than membership ride count. Also, the casual average ride time also drops to more closely resemble membership ride time average.
After separating and analyzing my data on a weekly basis, I noticed membership rider count was higher on weekdays than weekends while casual rider count was the complete opposite, higher on the weekends and lower on the weekdays. Both membership riders and casual riders have an increased trip time average on the weekends.
I also separated my data into quarterly aggregates to see if there was any difference between seasons (quarters and seasons matched for the most part). I noticed that rider_count was higher for both membership riders and casual riders in June - August. Both counts dropped in December - February but casual rider count dropped significantly everyday of the week. The average trip time stays about the same throughout the year with the exception of December - February, there is a small decrease in average trip time between both rider types.
I decided to separate the ride count into different time intervals. The majority of membership ride time is between 1-5 min and 5-10 min whereas the majority of casual ride time is between 5-10 min and 15-30 min. There is also a larger number of casual riders who have taken bike rides over 1 hr than member riders.
I analyzed the top 20 station locations for starting and ending locations. I noticed that a lot of membership trips and casual trips did not share the same station locations. I was able to pull the latitude and longitude for each station location to hopefully use to visualize data geographically.
After noticing a complete drop in casual rider count for 2nd quarter (December-February), I decided to also analyze the top 20 station locations for 2nd quarter to see if there were any differences. The results were more similar to overall membership station locations than overall casual station locations.

SHARE

After organizing and formatting data, I imported my google sheets document, '202309_202408_CyclisticAnalyzedData' https://docs.google.com/spreadsheets/d/1DlsFo9a9v79fZ12HbLlfPqG-iBwvY2XONY1JVraHPBs/edit?usp=sharing
to Tableau Public to create visuals for the analyzed data. https://public.tableau.com/app/profile/ruby.smith/vizzes

4 Vizzes available:

- Cyclistic-Location Popularity
- Cyclistic Annually, Monthly, and Weekday Dashboards -Tabs
- Cyclistic Time Intervals
- Cyclistic Quarterly - Tabs

I saved the vizzes as .jpeg files and used them to create a Case Study Slide Show.

Google Slides '2024_Cyclistic_Case_StudyRS'
https://docs.google.com/presentation/d/193xSKJ1-WgIpOcSgydBVMSnMISDIsP1sYla5yYSAQ9o/edit?usp=sharing

Create and Share Visuals through R

Install and load packages that I used for creating visuals through R

```
install.packages("tidyverse")
install.packages("formatR")
library(tidyverse)
library(ggplot2)
library(readxl)
library(formatR)
library(knitr)
```

Import analyzed data from spreadsheet and create data frames

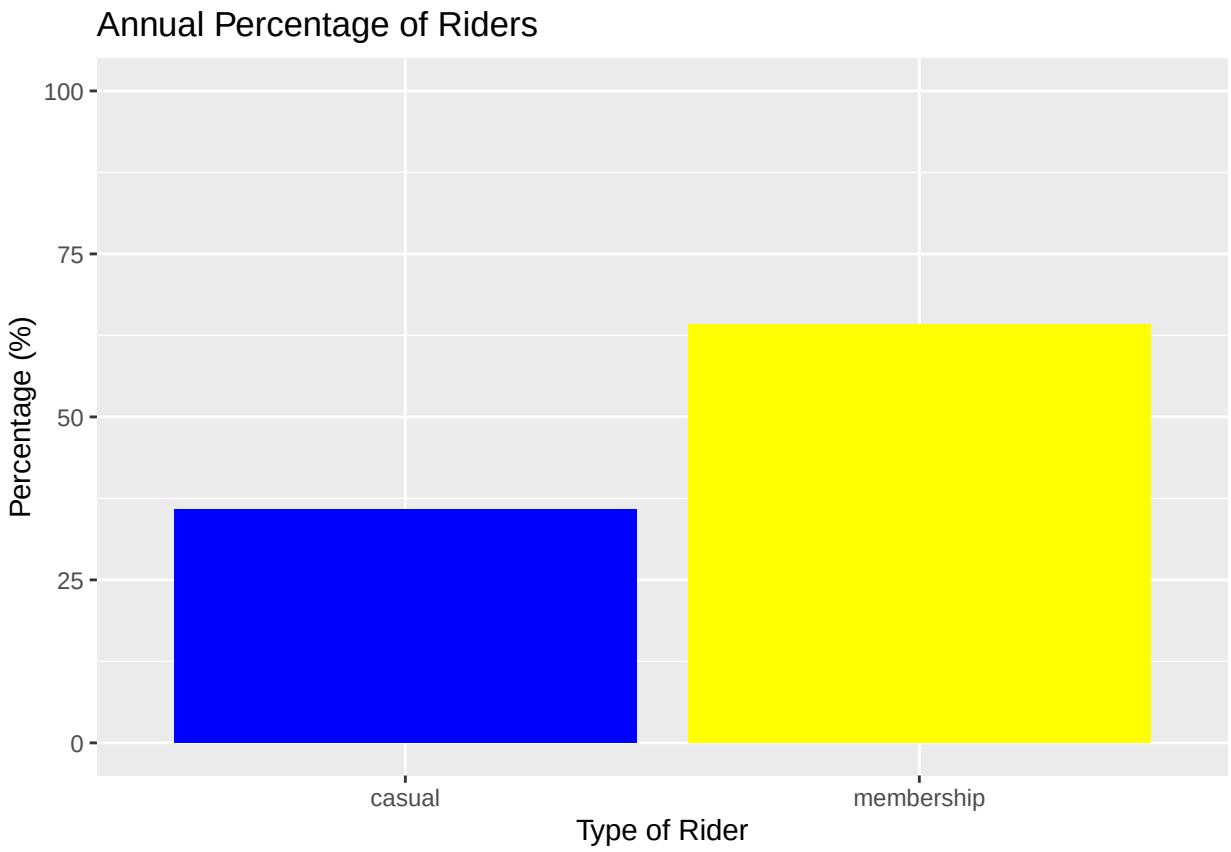
```
Annual <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Annual")
Monthly <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Monthly")
Weekday <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Weekday")
Quarterly <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Quarterly")
Pop_Locations <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Popular_Locations")
Q2Pop_locations <- read_excel("202309_202408_CyclisticAnalyzedData.xlsx",
    sheet = "Q2_Popular_Locations")
time <- read_excel("time.xlsx")
filtered_annual <- subset(Annual, percent_of_riders < 100)
filtered_quarterly <- subset(Quarterly, quarter == "December - February")
```

Factors needed to create order within columns that were utilized

```
Weekday$weekday <- factor(Weekday$weekday, levels = c("Sunday", "Monday", "Tuesday",
    "Wednesday", "Thursday", "Friday", "Saturday"))

Quarterly$quarter <- factor(Quarterly$quarter, levels = c("September - November",
    "December - February", "March - May", "June - August"))

Quarterly$weekday <- factor(Quarterly$weekday, levels = c("Sunday", "Monday", "Tuesday",
    "Wednesday", "Thursday", "Friday", "Saturday"))

time$time_interval <- factor(time$time_interval, levels = c("< 1 min", "1 < n > 5 min",
    "5 < n > 10 min", "10 < n > 15 min", "15 < n > 30 min", "30 min < n > 1 hr",
    "1 hr < n > 3 hr", "< 3 hr"))
```

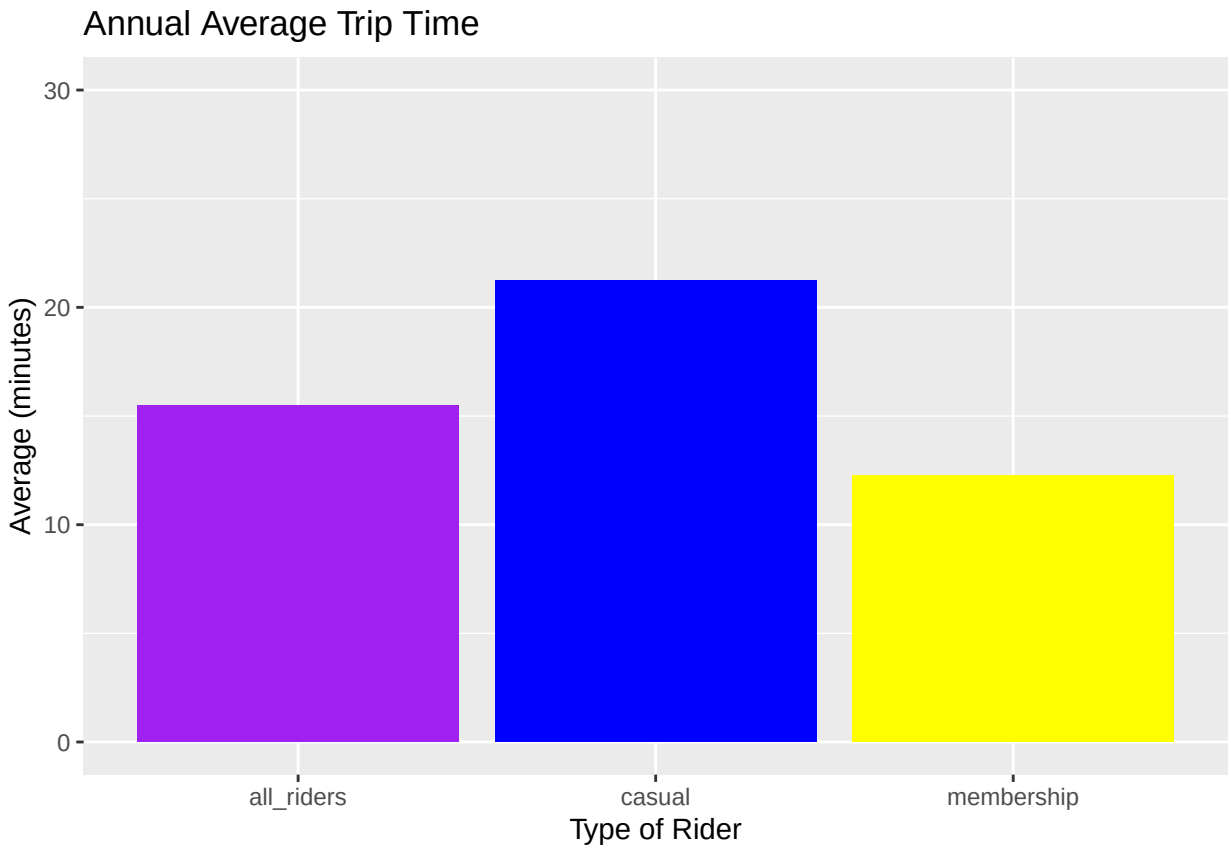
Analyzing Cyclistic Data Annually

```
ggplot(data = filtered_annual) + geom_col(mapping = aes(x = rider_status,
    y = percent_of_riders, fill = rider_status)) + scale_fill_manual(values = c("blue",
    "yellow")) + labs(title = "Annual Percentage of Riders",
    x = "Type of Rider", y = "Percentage (%)") + ylim(0, 100) +
    theme(legend.position = "none")
```



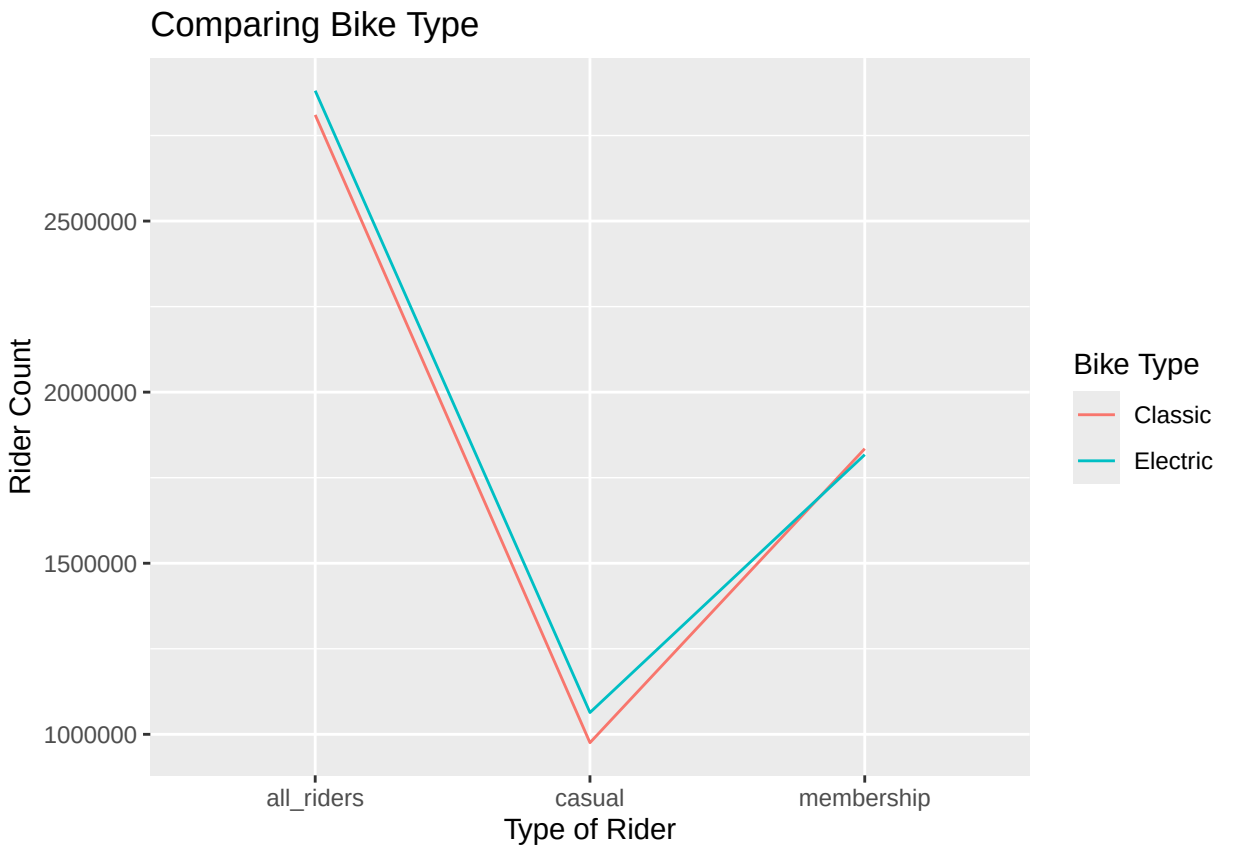
From September to August, 65% of Cyclistic riders held an annual membership, whereas approximately 35% were casual users who rented a bike for each use.

```
ggplot(data = Annual) + geom_col(mapping = aes(x = rider_status,
    y = annual_avg_minutes, fill = rider_status)) + scale_fill_manual(values = c("purple",
    "blue", "yellow")) + labs(title = "Annual Average Trip Time",
    x = "Type of Rider", y = "Average (minutes)") + ylim(0, 30) +
    theme(legend.position = "none")
```



When examining trip duration, I found that casual riders averaged 9 minutes longer per trip compared to annual members. Annual members typically rode for just over 10 minutes per trip, suggesting short-distance travel likely for commuting, errands, or visits. In contrast, casual riders averaged around 20 minutes per trip, indicating longer rides potentially for leisure or exploration.

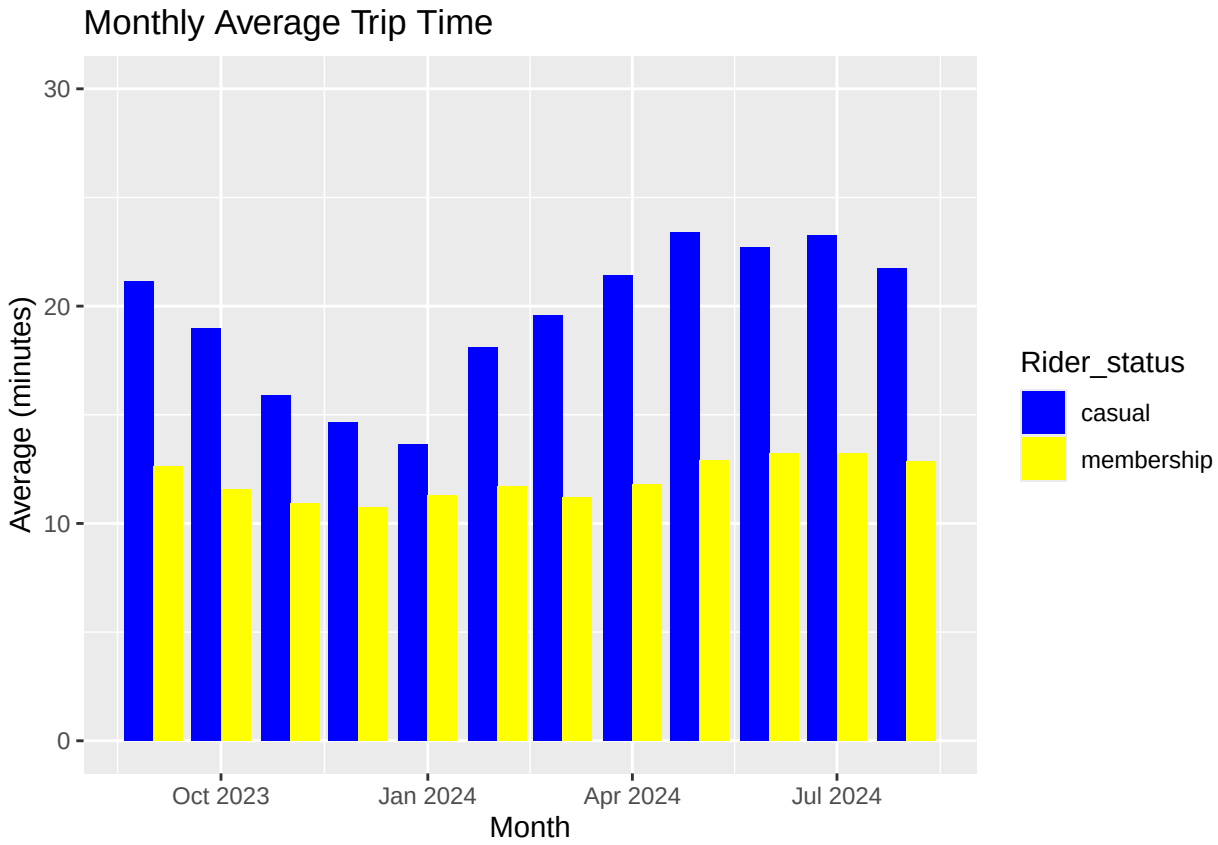
```r
ggplot(data = Annual) + geom_line(mapping = aes(x = rider_status,
    y = classic_count, group = 3, color = "green")) +
    geom_line(mapping = aes(x = rider_status, y = electric_count,
        group = 3, color = "purple")) + labs(title = "Comparing Bike Type",
    x = "Type of Rider", y = "Rider Count") + scale_color_discrete(name = "Bike Type",
    labels = c("Classic", "Electric"))
```



The choice between classic manual bikes and electric bikes showed no significant difference, though casual riders exhibited a slight preference for electric bikes. A periodic breakdown of bike type choices reinforced these findings.

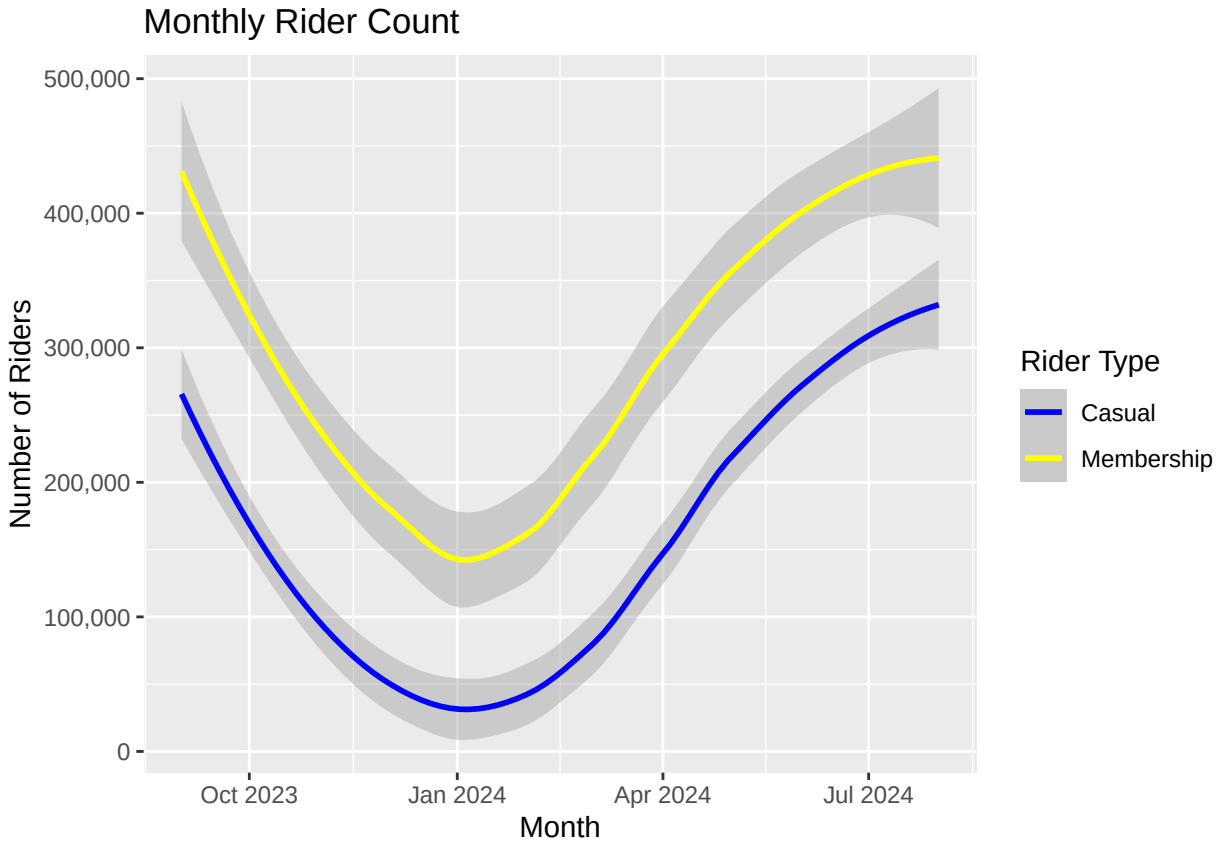
Analyzing Cyclistic data Monthly

```
ggplot(data = Monthly, aes(x = month, y = monthly_avg_minutes,
    fill = Rider_status)) + geom_col(position = "dodge") +
    scale_fill_manual(values = c("blue", "yellow")) +
    ylim(0, 30) + labs(title = "Monthly Average Trip Time",
    x = "Month", y = "Average (minutes)") + scale_color_discrete(name = "Rider Type",
    labels = c("Casual", "Membership"))
```



Membership riders typically maintain an average trip duration of around 12 minutes throughout the year, showing stability with minimal fluctuation. Casual riders consistently take longer trips, averaging about 20 minutes or more, except during the winter months (November to February), when a decline is observed.

```r
ggplot(data = Monthly) + geom_smooth(mapping = aes(x = month,
    y = rider_count, color = Rider_status)) + scale_y_continuous(label = scales::comma) +
    labs(title = "Monthly Rider Count", x = "Month", y = "Number of Riders") +
    scale_color_manual(values = c("blue", "yellow"), name = "Rider Type",
        labels = c("Casual", "Membership"))
```



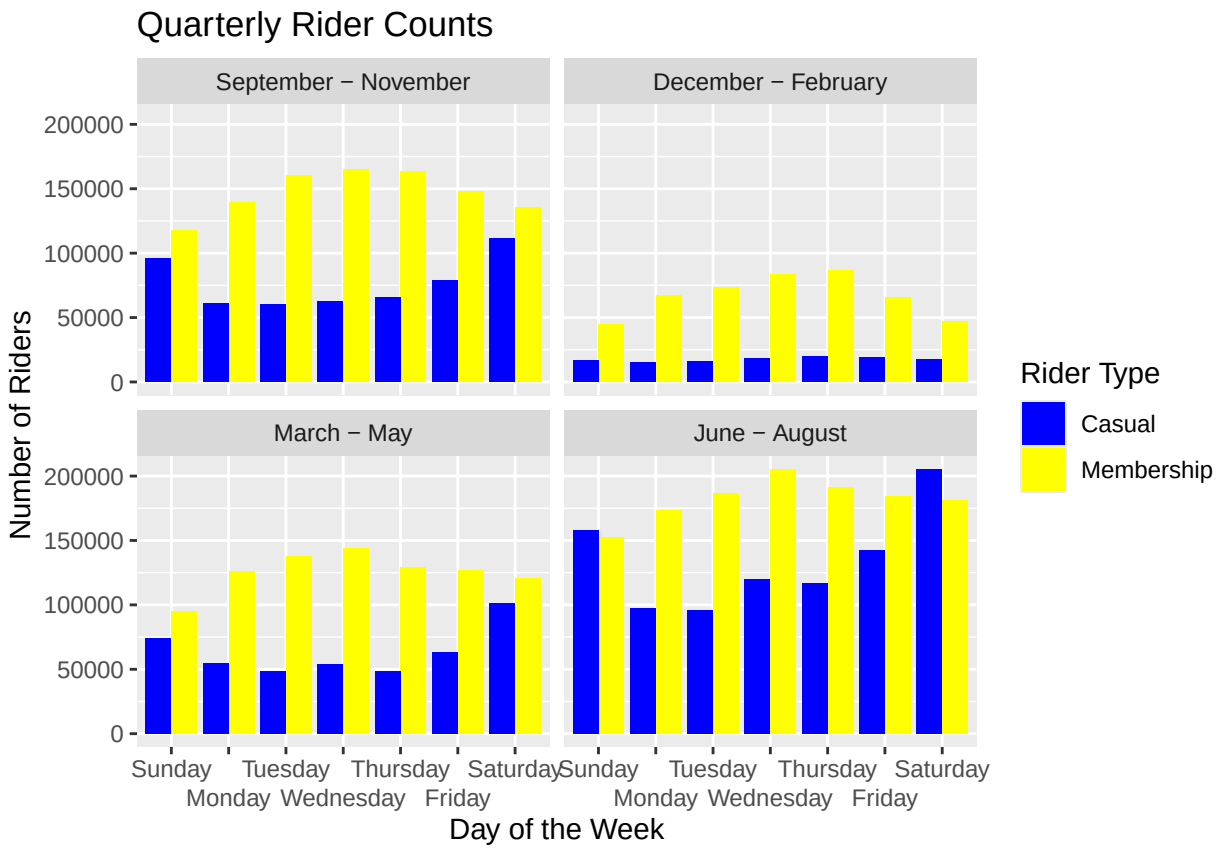
When examining monthly data, it is evident that riders with annual memberships typically take trips averaging around 12 minutes throughout the year, with minimal fluctuation. Casual riders, on the other hand, consistently average longer trips, beginning at approximately 20 minutes or more for most of the year. A notable decrease in trip duration is observed from November to February, a trend that holds true when analyzed quarterly.

Additionally, seasonal trends indicate a higher number of trips during the summer months and fewer during the winter. This pattern remains consistent when broken down on a monthly basis.

The data suggests that bike rentals are influenced by seasonal changes, regardless of membership status. Annual members tend to use bike rentals more routinely for shorter distances, likely for commuting, errands, or social visits. Conversely, casual riders appear to use bikes less frequently but for longer durations, possibly for exercise, sightseeing, or extended outings.

Analyzing Cyclistic data by Quarter

```
ggplot(data = Quarterly, aes(x = weekday, y = rider_count,
    fill = rider_status)) + geom_col(position = "dodge") +
    facet_wrap(~quarter) + scale_fill_manual(values = c("blue",
    "yellow"), name = "Rider Type", labels = c("Casual",
    "Membership")) + labs(title = "Quarterly Rider Counts",
    x = "Day of the Week", y = "Number of Riders") +
    scale_x_discrete(guide = guide_axis(n.dodge = 2))
```

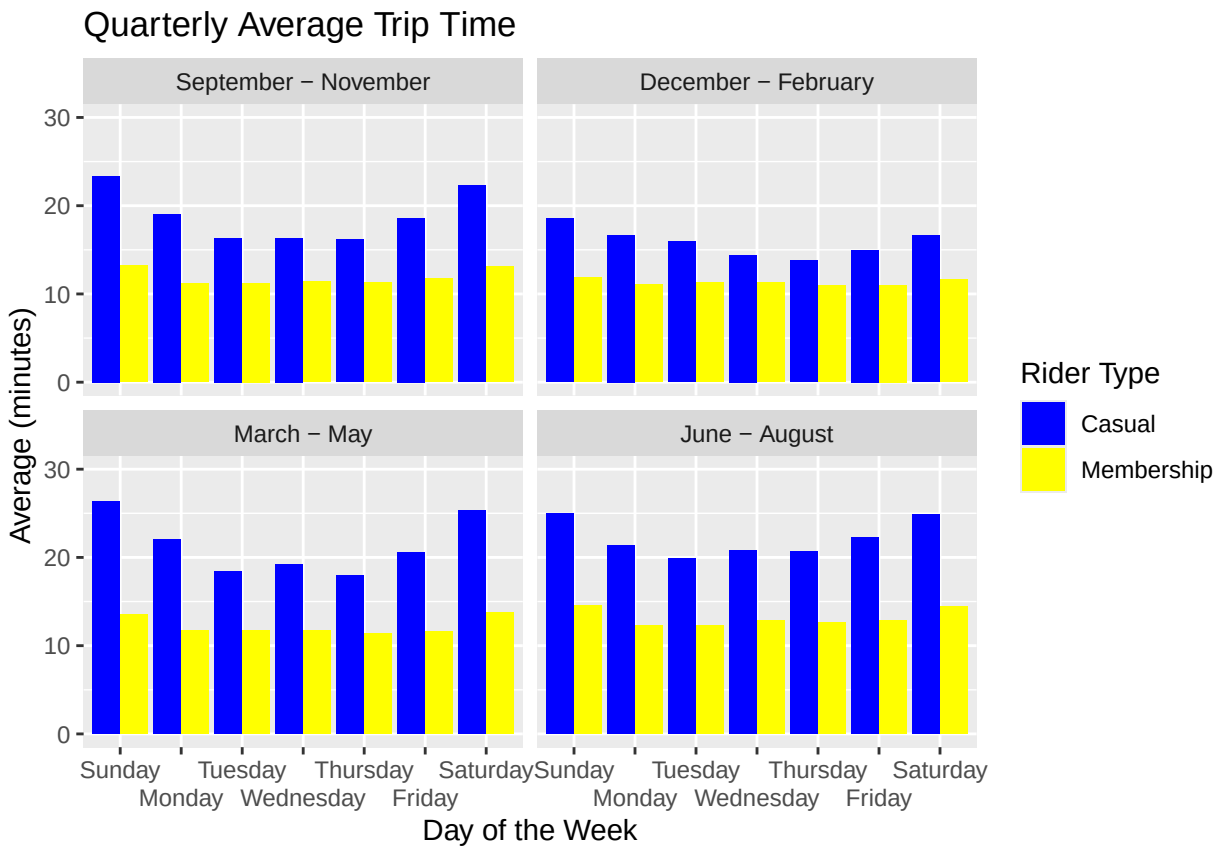


When dividing the year into quarters, it became clear that each quarter aligns well with the seasons—Fall, Winter, Spring, and Summer. I further organized the data by days of the week to identify any differences. Throughout the year, annual membership riders consistently use the Cyclistic Bike-Share service during the week, whereas casual riders predominantly use it on weekends.

From December to February, the coldest months, we observed the slowest quarter with the fewest trips taken. Casual trips were notably low, and there was a decrease in trips by annual members as well. Conversely, June to August, the warmest months, saw the highest number of trips. Casual trips peaked on weekends and were also higher during weekdays. Membership trips reached their highest numbers in the summer quarter.

After examining some geographic data, I organized the data by season and found the results insightful. Fall appeared busy, Winter showed significantly reduced activity, Spring indicated a resurgence, and Summer was exceedingly busy. The drop in casual ridership during winter may be attributed to less favorable weather conditions, while the increased activity on weekends suggests that casual riders use the service primarily for leisure or exercise. The steady weekday usage by members reinforces the idea that they rely on the service for routine activities, such as commuting to work or school.

```
ggplot(data = Quarterly, aes(x = weekday, y = quarterly_avg_minutes,
    fill = rider_status)) + geom_col(position = "dodge") + facet_wrap(~quarter) +
    ylim(0, 30) + scale_fill_manual(values = c("blue", "yellow"),
    name = "Rider Type", labels = c("Casual", "Membership")) +
    labs(title = "Quarterly Average Trip Time", x = "Day of the Week",
        y = "Average (minutes)") + scale_x_discrete(guide = guide_axis(n.dodge = 2))
```



When examining the average length of each trip on a quarterly basis, we observed that the period from December to February was the least busy, while June to August was the busiest. This led us to analyze potential differences in trip durations across these periods.

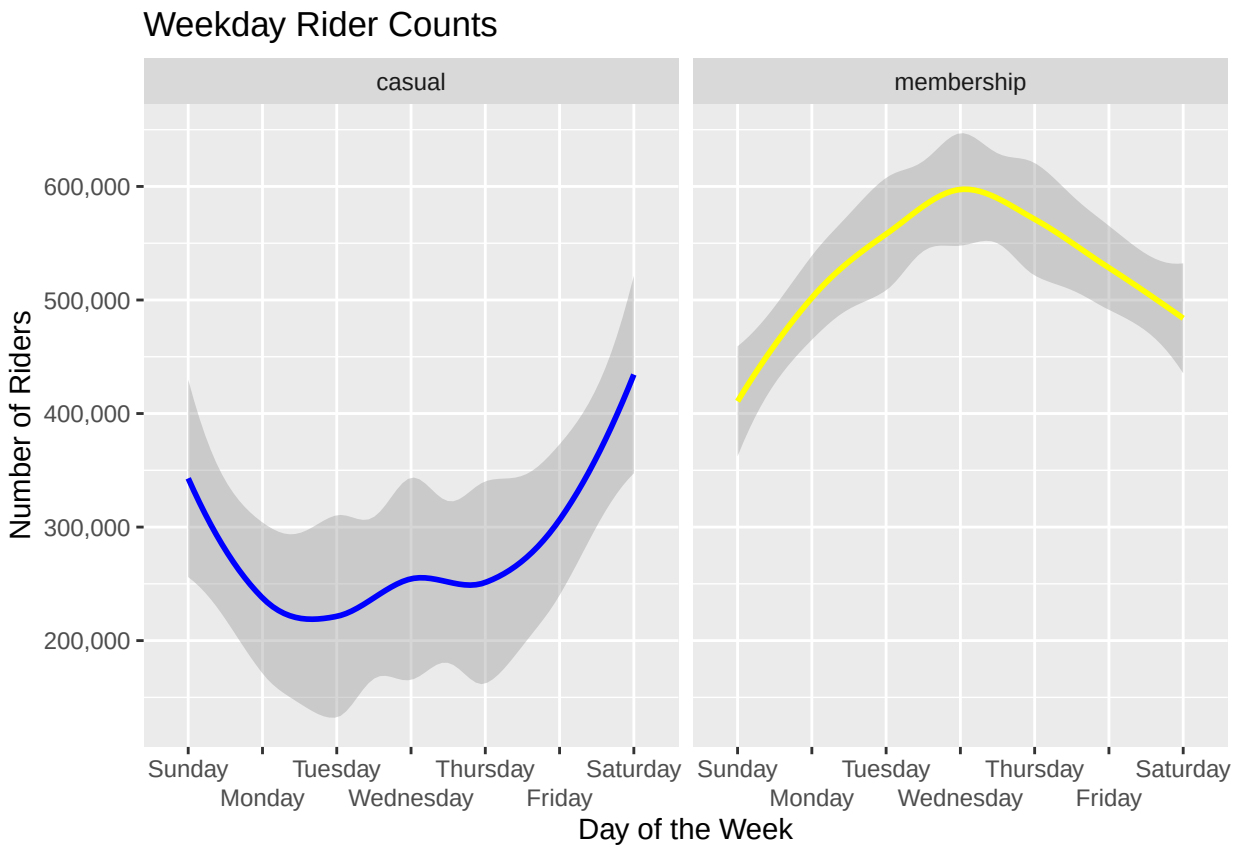
The analysis revealed that annual membership riders maintained a consistent average trip duration throughout the year. In contrast, casual riders experienced a decrease in trip time during the winter months and an increase in the summer months.

This quarterly breakdown reinforces the notion that members likely use the bikes for routine purposes, as their average trip duration remains stable year-round. Casual riders, however, show a slight decrease in average trip duration during winter, indicating less frequent but longer rides for purposes such as exercise or leisure.

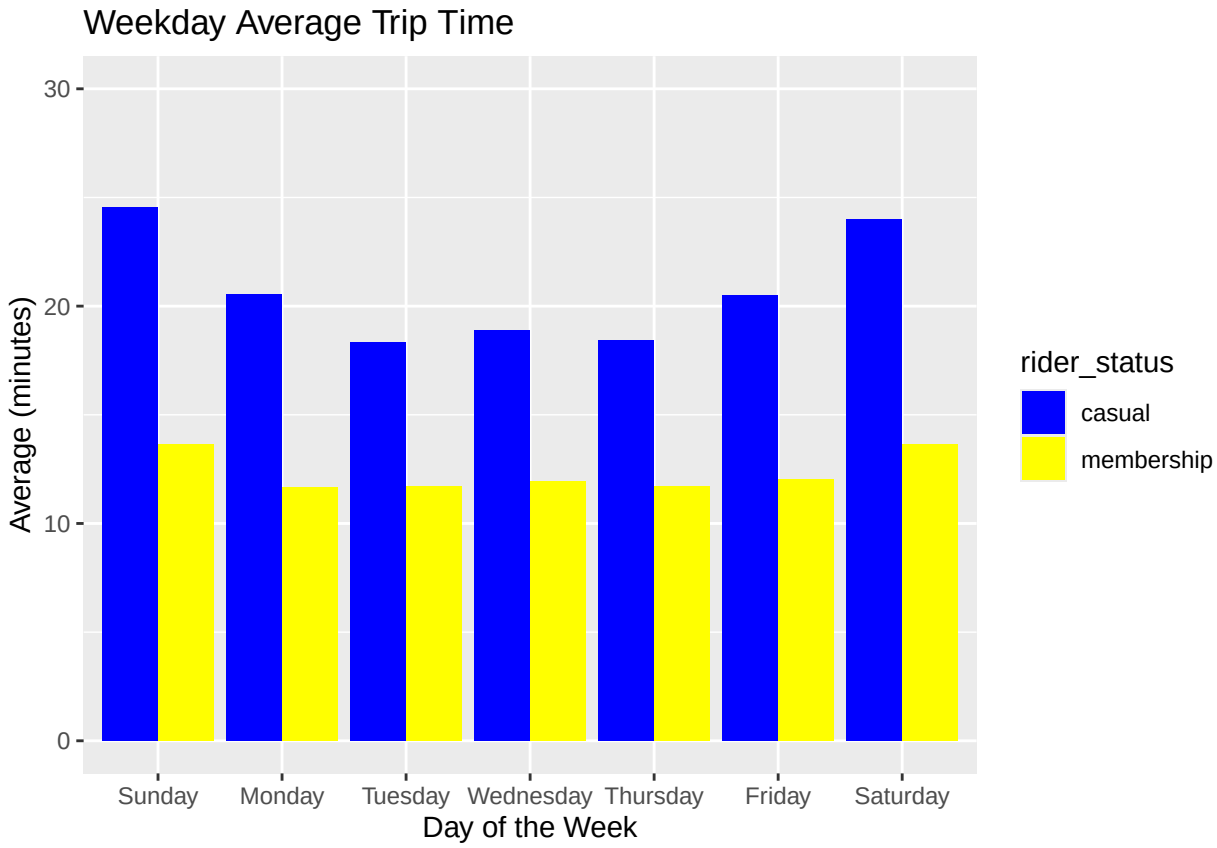
This data suggests that bike rentals are influenced by seasonal changes, regardless of membership status. Annual members tend to use bike rentals more routinely for shorter distances, likely for commuting, errands, or social visits. Conversely, casual riders appear to use bikes less frequently but for longer durations, possibly for exercise, sightseeing, or extended outings.

Examining Data by Weekday

```
ggplot(data = Weekday) + geom_smooth(mapping = aes(x = weekday,
    y = rider_count, group = 1, color = rider_status)) +
    facet_wrap(~rider_status) + scale_y_continuous(label = scales::comma) +
    scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
    scale_color_manual(values = c("blue", "yellow")) +
    labs(title = "Weekday Rider Counts", x = "Day of the Week",
        y = "Number of Riders") + theme(legend.position = "none")
```



```r
ggplot(data = Weekday, aes(x = weekday, y = weekday_avg_minutes,
    fill = rider_status)) + geom_col(position = "dodge") +
    scale_fill_manual(values = c("blue", "yellow")) +
    ylim(0, 30) + labs(title = "Weekday Average Trip Time",
    x = "Day of the Week", y = "Average (minutes)") +
    scale_color_discrete(name = "Rider Type", labels = c("Casual",
        "Membership"))
```

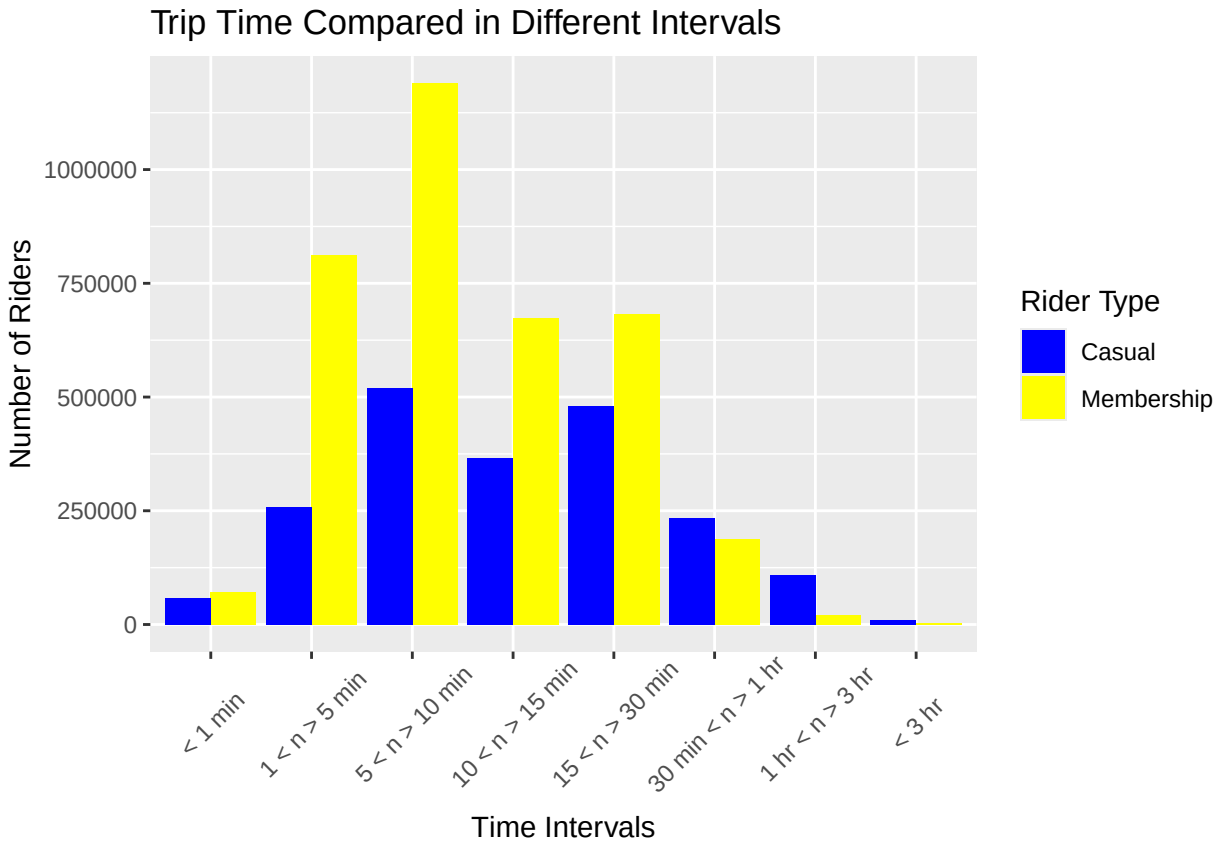


Upon analyzing the data on a weekly basis, it becomes evident that annual membership riders maintain a consistent average trip duration of 10 to 15 minutes, regardless of the day. However, their numbers are significantly higher during weekdays compared to weekends. In contrast, casual riders exhibit longer average trip durations, particularly on weekends, and maintain a higher average trip duration throughout the week compared to members. Casual riders are more active on weekends and less so on weekdays.

This suggests that membership riders utilize the service more consistently throughout the week, likely for commuting to work or school, as well as for weekend activities such as errands or exercise. Conversely, casual riders appear to use the service primarily for leisure activities or longer-distance trips, especially on weekends.

Breaking data down into time intervals

```
colnames(time) <- c("time_interval", "rider_status", "rider_count")

ggplot(data = time, aes(x = time_interval, y = rider_count, fill = rider_status)) +
    geom_col(position = "dodge") + theme(axis.text.x = element_text(angle = 45,
    vjust = 0.6)) + scale_fill_manual(values = c("blue", "yellow"),
    name = "Rider Type") + labs(title = "Trip Time Compared in Different Intervals",
    x = "Time Intervals", y = "Number of Riders")
```



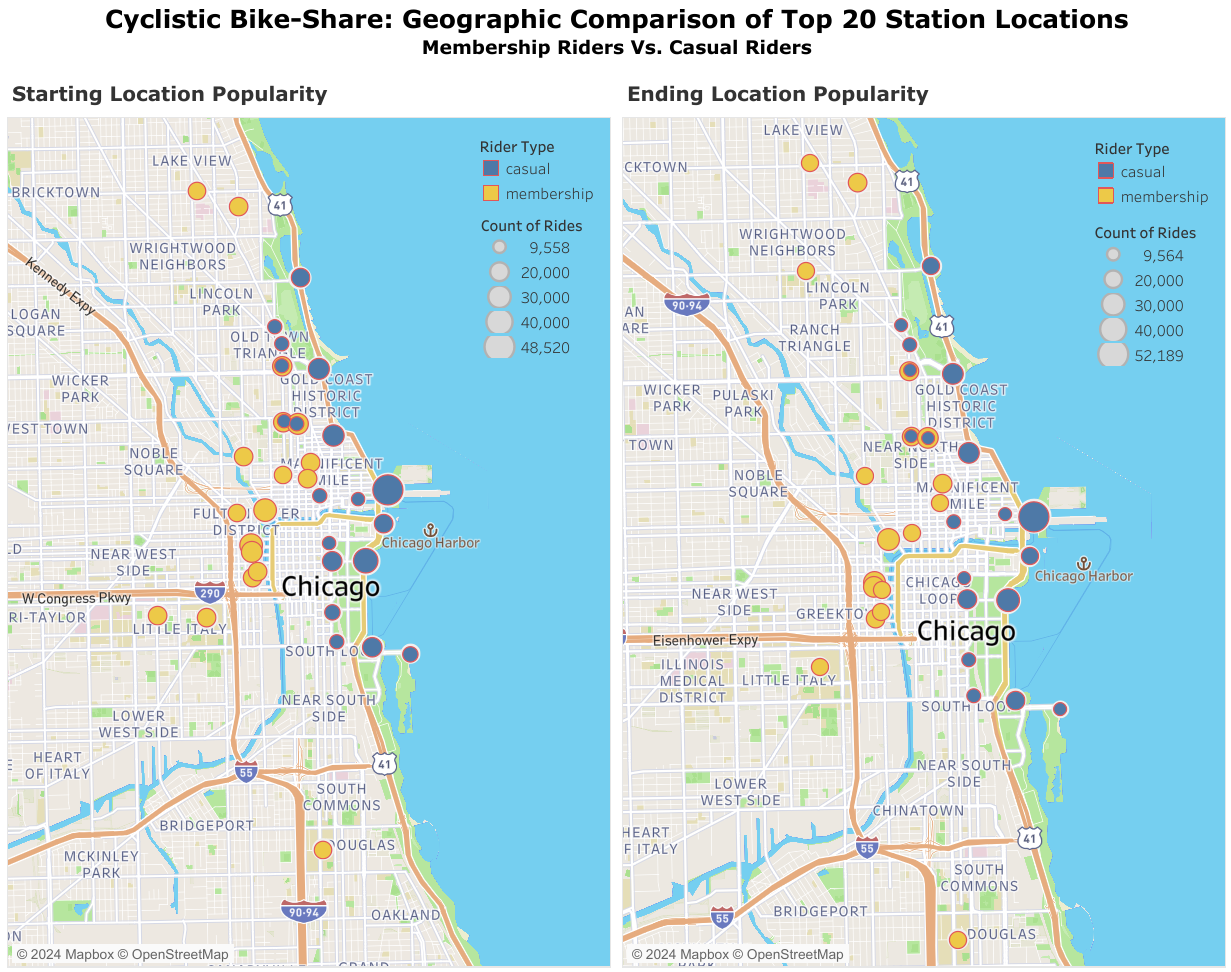
The data indicates that the most popular trip durations fall within the 5 to 10-minute range, with the highest number of riders at approximately 1200K. Following closely, the 10 to 15-minute interval also sees significant usage, with both yellow and blue bars reaching around 800K riders. Conversely, the least popular durations are trips less than 1 minute and those between 1 to 3 hours, reflected by much shorter bars. Additionally, a higher percentage of casual riders take trips exceeding 30 minutes compared to membership riders.

This analysis suggests that the majority of riders prefer shorter trips, likely for quick commutes or errands. The high usage within the 5 to 10-minute range highlights the efficiency of the bike-share program for short-distance travel.

This breakdown reveals that most membership riders tend to use the service for shorter durations. While many casual riders share a similar average trip time of around 10 minutes, a significant proportion of them engage in longer rides. This raises the question of why casual riders, on average, undertake longer trips more frequently.

Analyzing Data Geographically

```
knitr::include_graphics("/cloud/project/Cyclistic_Geographic_Viz_RS.png")
```



Lastly, I examined popular station locations to see if any patterns emerged. I found that riders with annual memberships and casual riders typically do not share starting and ending locations. Most membership riders tend to stay inland, frequenting residential areas, businesses, train stations, hospitals, and colleges. In contrast, casual riders are more concentrated around Lake Michigan, choosing stations near attractions such as beaches, theaters, museums, Millennium Park, Navy Pier, and Soldier Field.

Geographically, this data suggests that casual riders are heading to recreational areas along the lake, likely for outings or special occasions. Given the less frequent usage compared to membership riders, it is reasonable to infer that casual riders are predominantly tourists or individuals engaging in leisure activities.

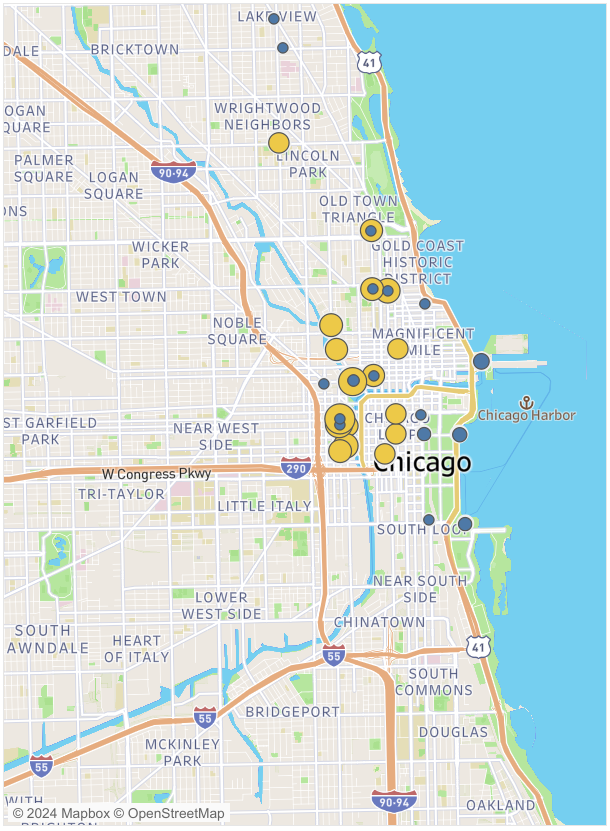
```
knitr::include_graphics("/cloud/project/geo_q2.png")
```

Cyclistic Bike-Share: Geographic Comparison of Top 20 Station Locations

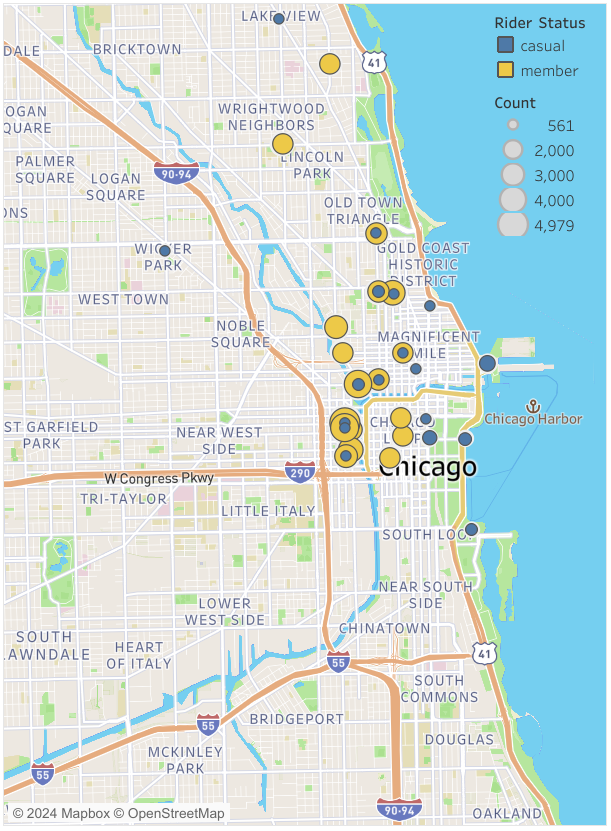
Quarter 2: December - February

Membership Riders Vs. Casual Riders

Starting Location Popularity



Ending Location Popularity



I analyzed the most popular station locations during the second quarter (December to February) and observed that casual riders frequently shared station locations with membership riders during this period. The average trip duration for casual riders was slightly higher but generally comparable to that of membership riders in Q2.

Upon further geographical analysis of Q2, it becomes apparent that casual riders are likely to be tourists or individuals partaking in special outings. During the winter months, the number of trips remains consistent daily, with a lower volume that closely matches the average trip time of members. Additionally, the geographical routes are more similar between casual and membership riders during this period compared to other months. This suggests that casual riders in winter are more likely to be local residents or individuals who use the Cyclistic Bike-Share service more regularly.

Act

After analyzing Cyclistic bike riders and identifying differences in how they utilize the Bike-Share service, I have several recommendations. Our data indicates that casual riders frequently visit leisure areas and are likely tourists or individuals on special outings. By segmenting these casual riders from those who exhibit behaviors similar to annual members, we can tailor our marketing strategies accordingly.
Identifying the most popular station locations allows us to target casual riders effectively. Advertising in these areas, especially near the Metra train station, which is also popular among casual riders, could attract more annual memberships and encourage casual riders to try Cyclistic Bike-Share.
Additionally, offering annual memberships as gifts during the slower winter months could boost sales. Parents could purchase these for their college-bound children, or spouses could gift them to support their partner's commute. This strategy could increase casual rider conversion to annual memberships.

Top 3 Recommendations:
1. Market to casual riders at stations not located in sightseeing or leisure areas.
2. Promote marketing around Metra train stations to attract new casual riders and potential annual memberships.
3. Offer annual memberships as a gift idea during the winter holiday months.