Capstone Project

Kiran Kour

MIS- 64099-030

Professor Dr. Rouzbeh Razavi

Kent State University

Introduction:

In the dynamic landscape of urban transportation, bike-sharing systems have emerged as a popular and sustainable mode of transit. This capstone project, conducted as part of the MSBA summer course, delves into a comprehensive analysis of a bike-sharing service dataset of a company named Cyclist. Leveraging the versatile capabilities of the R programming language and the intuitive RStudio IDE, this project aims to unravel key insights, patterns, and trends within the realm of bike ridership.

Bike-sharing services have become a flexible and environmentally friendly option for urban commuters and leisure riders. This analysis focuses on a dataset encompassing various ride-related attributes, membership details, and temporal variations. By employing R's robust statistical tools and visualization functionalities, this project seeks to extract actionable information to inform strategic decisions for optimizing the bike-sharing experience.

The primary objective is to discern distinct behaviors and preferences exhibited by two core user segments: **members and casual riders**. By unraveling their distinct characteristics and usage patterns, we aim to highlight potential opportunities for service enhancement, customer engagement strategies, and operational adjustments. This investigation contributes to a deeper understanding of the dynamics driving bike-sharing systems. It provides valuable insights that can be harnessed by the service provider to further refine their offerings.

The subsequent sections of this report will delve into a detailed analysis of the dataset, encompassing key observations, findings, and recommendations. By exploring user demographics, temporal trends, bike preferences, and geographical distribution, this project endeavors to uncover actionable insights to drive the bike-sharing service toward increased efficiency, user satisfaction, and sustainable growth.

For this project following data analysis steps will be followed:

- Ask
- Prepare

- Process
- Analyze
- Act

Scenario

As a junior data analyst within Cyclistic's marketing team, my role shapes the company's future success. Cyclists, a leading bike-share company in Chicago, aims to maximize annual memberships, and your team is tasked with understanding how casual riders and annual members use the service differently. By delving into data insights and crafting professional visualizations using R programming language and RStudio IDE, you will provide compelling evidence to support recommendations for a new marketing strategy. The goal is to convert casual riders into loyal annual members, ultimately driving Cyclistic's growth and market impact.

Ask:

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

- What are the distinct patterns of bike usage between annual members and casual riders?

- How do ride durations, frequency, and timing vary between these two user segments?

- Are there specific geographic areas where annual members and casual riders tend to ride more frequently?

2. Why would casual riders buy Cyclistic annual memberships?

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

The marketing director and your manager, Lily Moreno, have assigned me the first question: ***How do annual members and casual riders use Cyclistic bikes differently?***

My analysis of the first question will set the foundation for developing strategies to address the subsequent questions and guide Cyclistic's marketing initiatives.

Key tasks

- *Identify the Business task*
    - The main objective is to build the best marketing strategies to turn casual bike riders into annual members by analyzing how the 'Casual' and 'Annual' customers use Cyclistic bike share differently.
- *Consider key stakeholders*
    - Cyclistic executive team, Director of Marketing (Lily Moreno), Marketing Analytics team

Deliverable

- A clear statement of the business task
    - Find the differences between casual and member riders.

Prepare

I will use Cyclistic's historical trip data to analyze and identify trends. The data has been made available and downloaded from Kaggle.

Key tasks

❖ Download data and store it appropriately.

- Data has been downloaded from here, and copies have been stored securely on my computer and Kaggle.
- Identify how it's organized.
- All trip data is in comma-delimited (.CSV) format. 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" (Total 13 column)

- ❖ Sort and filter the data.
  - For this analysis, I'm going to use the last 12 months data of the year 2021.

- ❖ Determine the credibility of the data.
  - For the purposes of this case study, the datasets are appropriate and it will enable me to answer the business questions. But due to data privacy, I cannot use the rider's personally identification information, and this will prevent me from determining if a single user/rider taken several rides. All ride ids are unique in this data-set.

Deliverable

- ❖ A description of all data sources used
  - Main source of data provided by the Cylistic company.

```
#Installation of Packages

#install.packages("tidyverse")
#install.packages("janitor")
#install.packages("ggmap")
#install.packages("geosphere")
#install.packages("lubridate")

library(tidyverse)

library(janitor)

library(ggmap)

library(geosphere)

library(lubridate)

#Importing data

jan21<- read_csv("202101-divvy-tripdata.csv")

feb21<- read_csv("202102-divvy-tripdata.csv")
```

```r
mar21<- read_csv("202103-divvy-tripdata.csv")

apr21<- read_csv("202104-divvy-tripdata.csv")

may21<- read_csv("202105-divvy-tripdata.csv")

jun21<- read_csv("202106-divvy-tripdata.csv")

july21<- read_csv("202107-divvy-tripdata.csv")

sep21<- read_csv("202109-divvy-tripdata.csv")

oct21<- read_csv("202110-divvy-tripdata.csv")

nov21<- read_csv("202111-divvy-tripdata.csv")

dec21<- read_csv("202112-divvy-tripdata.csv")



#Checking data sets for consistency
colnames(jan21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```r
colnames(feb21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```r
colnames(mar21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```r
colnames(apr21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```r
colnames(may21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(jun21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(july21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(aug21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(sep21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(oct21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(nov21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(dec21)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
#Merging individual monthly data frames into one large data frame

tripdata<- bind_rows(jan21, feb21, mar21, apr21, may21, jun21, july21, aug21,
sep21, oct21, nov21, dec21)
```

Process

Cleaning and Preparation of data for analysis

Key tasks

- Check the data for errors.
- Choose your tools.
- Transform the data so you can work with it effectively.
- Document the cleaning process.

Deliverable

- Documentation of any cleaning or manipulation of data

The following code chunks will be used for this 'Process' phase.

```
#Checking merged data frame

colnames(tripdata)  #List of column names
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
head(tripdata)  #See the first 6 rows of data frame.
```

```
## # A tibble: 6 × 13
##   ride_id        ridea…¹ started_at          ended_at            start…² start…³
##   <chr>          <chr>   <dttm>              <dttm>              <chr>   <chr>
## 1 E19E6F1B8D4C4… electr… 2021-01-23 16:14:19 2021-01-23 16:24:44 Califo… 17660
## 2 DC88F20C2C55F… electr… 2021-01-27 18:43:08 2021-01-27 18:47:12 Califo… 17660
## 3 EC45C94683FE3… electr… 2021-01-21 22:35:54 2021-01-21 22:37:14 Califo… 17660
## 4 4FA453A75AE37… electr… 2021-01-07 13:31:13 2021-01-07 13:42:55 Califo… 17660
## 5 BE5E8EB4E7263… electr… 2021-01-23 02:24:02 2021-01-23 02:24:45 Califo… 17660
## 6 5D8969F88C773… electr… 2021-01-09 14:24:07 2021-01-09 15:17:54 Califo… 17660
## # … with 7 more variables: end_station_name <chr>, end_station_id <chr>,
## #   start_lat <dbl>, start_lng <dbl>, end_lat <dbl>, end_lng <dbl>,
## #   member_casual <chr>, and abbreviated variable names ¹rideable_type,
## #   ²start_station_name, ³start_station_id
```

```
str(tripdata)  #See list of columns and data types (numeric, character, etc)
```

```
## spc_tbl_ [5,595,063 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5595063] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453
A75AE377DB" ...
##  $ rideable_type     : chr [1:5595063] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
...
##  $ started_at        : POSIXct[1:5595063], format: "2021-01-23 16:14:19" "2021-01-27 18:43:08" ...
##  $ ended_at          : POSIXct[1:5595063], format: "2021-01-23 16:24:44" "2021-01-27 18:47:12" ...
##  $ start_station_name: chr [1:5595063] "California Ave & Cortez St" "California Ave & Cortez St" "Calif
ornia Ave & Cortez St" "California Ave & Cortez St" ...
##  $ start_station_id  : chr [1:5595063] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5595063] NA NA NA NA ...
##  $ end_station_id    : chr [1:5595063] NA NA NA NA ...
##  $ start_lat         : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5595063] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5595063] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5595063] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(tripdata)   #Statistical summary of data. Mainly for numeric.
```

```
##    ride_id          rideable_type        started_at
##  Length:5595063     Length:5595063     Min.   :2021-01-01 00:02:05.00
##  Class :character   Class :character   1st Qu.:2021-06-06 23:52:40.00
##  Mode  :character   Mode  :character   Median :2021-08-01 01:52:11.00
##                                        Mean   :2021-07-29 07:41:02.63
##                                        3rd Qu.:2021-09-24 16:36:16.00
##                                        Max.   :2021-12-31 23:59:48.00
##
##    ended_at                      start_station_name start_station_id
##  Min.   :2021-01-01 00:08:39.00   Length:5595063     Length:5595063
##  1st Qu.:2021-06-07 00:44:21.00   Class :character   Class :character
##  Median :2021-08-01 02:21:55.00   Mode  :character   Mode  :character
##  Mean   :2021-07-29 08:02:58.75
##  3rd Qu.:2021-09-24 16:54:05.50
##  Max.   :2022-01-03 17:32:18.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5595063     Length:5595063     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##    end_lat         end_lng       member_casual
##  Min.   :41.39   Min.   :-88.97   Length:5595063
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
```

```
##  Max.   :42.17   Max.   :-87.49
##  NA's   :4771    NA's   :4771
```

```
#Adding date, month, year, day of week columns

tripdata <- tripdata %>%
  mutate(year = format(as.Date(started_at), "%Y")) %>% # extract year
  mutate(month = format(as.Date(started_at), "%B")) %>% #extract month
  mutate(date = format(as.Date(started_at), "%d")) %>% # extract date
  mutate(day_of_week = format(as.Date(started_at), "%A")) %>% # extract day o
f week
  mutate(ride_length = difftime(ended_at, started_at)) %>%
  mutate(start_time = strftime(started_at, "%H"))

#Converting 'ride_length' to numeric for calculation on data

tripdata <- tripdata %>%
  mutate(ride_length = as.numeric(ride_length))
is.numeric(tripdata$ride_length) # to check it is right format
```

```
## [1] TRUE
```

```
#Adding ride distance in km

tripdata$ride_distance <- distGeo(matrix(c(tripdata$start_lng, tripdata$start
_lat), ncol = 2), matrix(c(tripdata$end_lng, tripdata$end_lat), ncol = 2))

tripdata$ride_distance <- tripdata$ride_distance/1000 #Distance in km

# Remove "bad" data
# The dataframe includes a few hundred entries when bikes were taken out of d
ocks
# and checked for quality by Divvy where ride_length was negative or 'zero'

tripdata_clean <- tripdata[!(tripdata$ride_length <= 0),]
```

Analyze

Now all the required information are in one place and ready for exploration.

Key tasks

- Aggregate your data so it's useful and accessible.
- Organize and format your data.
- Perform calculations.

- Identify trends and relationships.

Deliverable

- A summary of the analysis

The following code chunks will be used for this 'Analyze' phase

Comparing members and casual users :

```
#First lets check the cleaned data frame

str(tripdata_clean)
```

```
## tibble [5,594,410 × 20] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5594410] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453
A75AE377DB" ...
##  $ rideable_type     : chr [1:5594410] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
...
##  $ started_at        : POSIXct[1:5594410], format: "2021-01-23 16:14:19" "2021-01-27 18:43:08" ...
##  $ ended_at          : POSIXct[1:5594410], format: "2021-01-23 16:24:44" "2021-01-27 18:47:12" ...
##  $ start_station_name: chr [1:5594410] "California Ave & Cortez St" "California Ave & Cortez St" "Calif
ornia Ave & Cortez St" "California Ave & Cortez St" ...
##  $ start_station_id  : chr [1:5594410] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:5594410] NA NA NA NA ...
##  $ end_station_id    : chr [1:5594410] NA NA NA NA ...
##  $ start_lat         : num [1:5594410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:5594410] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:5594410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:5594410] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:5594410] "member" "member" "member" "member" ...
##  $ year              : chr [1:5594410] "2021" "2021" "2021" "2021" ...
##  $ month             : chr [1:5594410] "January" "January" "January" "January" ...
##  $ date              : chr [1:5594410] "23" "27" "21" "07" ...
##  $ day_of_week       : chr [1:5594410] "Saturday" "Wednesday" "Thursday" "Thursday" ...
##  $ ride_length       : num [1:5594410] 625 244 80 702 43 ...
##  $ start_time        : chr [1:5594410] "11" "13" "17" "08" ...
##  $ ride_distance     : num [1:5594410] 2.246 0.558 0.281 2.246 0.276 ...
```

```
#Lets check summarised details about the cleaned dataset

summary(tripdata_clean)
```

```
##    ride_id          rideable_type        started_at
##  Length:5594410     Length:5594410     Min.   :2021-01-01 00:02:05.00
##  Class :character   Class :character   1st Qu.:2021-06-06 23:48:04.75
##  Mode  :character   Mode  :character   Median :2021-08-01 01:47:45.00
##                                        Mean   :2021-07-29 07:38:35.31
##                                        3rd Qu.:2021-09-24 16:34:56.75
##                                        Max.   :2021-12-31 23:59:48.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2021-01-01 00:08:39.00   Length:5594410     Length:5594410
```

```
##  1st Qu.:2021-06-07 00:39:47.25   Class :character   Class :character
##  Median :2021-08-01 02:16:53.00   Mode  :character   Mode  :character
##  Mean   :2021-07-29 08:00:31.61
##  3rd Qu.:2021-09-24 16:52:38.00
##  Max.   :2022-01-03 17:32:18.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5594410     Length:5594410     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual          year
##  Min.   :41.39   Min.   :-88.97   Length:5594410     Length:5594410
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character   Class :character
##  Median :41.90   Median :-87.64   Mode  :character   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.49
##  NA's   :4770    NA's   :4770
##     month               date           day_of_week         ride_length
##  Length:5594410     Length:5594410     Length:5594410     Min.   :      1
##  Class :character   Class :character   Class :character   1st Qu.:    405
##  Mode  :character   Mode  :character   Mode  :character   Median :    720
##                                                           Mean   :   1316
##                                                           3rd Qu.:   1307
##                                                           Max.   :3356649
##
##   start_time        ride_distance
##  Length:5594410     Min.   :  0.000
##  Class :character   1st Qu.:  0.901
##  Mode  :character   Median :  1.639
##                     Mean   :  2.187
##                     3rd Qu.:  2.883
##                     Max.   :114.567
##                     NA's   :4770
```

Descriptive Analysis :

```
## Conduct descriptive analysis
# descriptive analysis on 'ride_length'
# mean = straight average (total ride length / total rides)
# median = midpoint number of ride length array
# max = longest ride
# min = shortest ride

tripdata_clean %>%
  summarise(average_ride_length = mean(ride_length), median_length = median(r
ide_length),
            max_ride_length = max(ride_length), min_ride_length = min(ride_le
ngth))
```

```
## # A tibble: 1 × 4
##   average_ride_length median_length max_ride_length min_ride_length
```

```
##                 <dbl>         <dbl>           <dbl>           <dbl>
## 1               1316.           720         3356649               1
```

- The provided dataset pertains to ride lengths over the course of the entire year 2021. Notably, the minimum ride length (min_ride_length) and maximum ride length (max_ride_length) exhibit values that could be described as unusual or anomalous. Due to the limitations of the available information, pinpointing the precise cause behind these values is a challenge. However, a thorough analysis is warranted to better understand this situation.

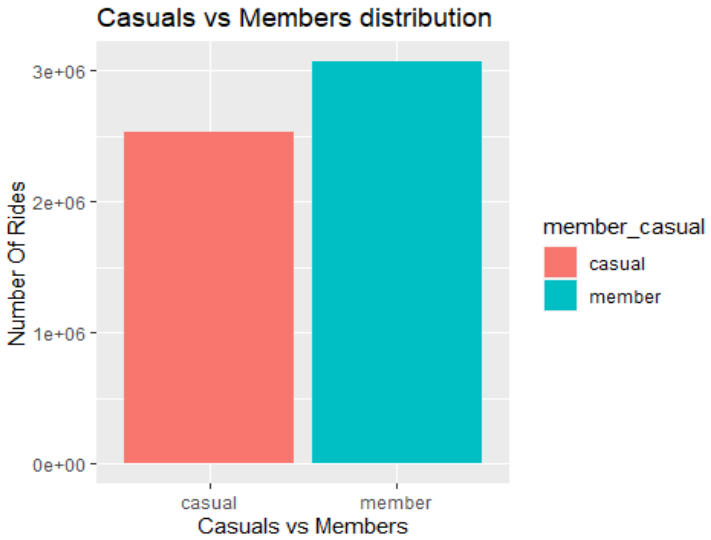
Comparing Casual riders and the Members

1. Members vs casual riders difference depending on total rides taken.

```
# members vs casual riders difference depending on total rides taken
tripdata_clean %>%
    group_by(member_casual) %>%
    summarise(ride_count = length(ride_id), ride_percentage = (length(ride_id
) / nrow(tripdata_clean)) * 100)
```

```
## # A tibble: 2 × 3
##   member_casual ride_count ride_percentage
##   <chr>              <int>           <dbl>
## 1 casual           2528664            45.2
## 2 member           3065746            54.8
```

```
ggplot(tripdata_clean, aes(x = member_casual, fill=member_casual)) +
    geom_bar() +
    labs(x="Casuals vs Members", y="Number Of Rides", title= "Casuals vs Memb
ers distribution")
```



- In the Casuals vs Members distribution chart, it is evident that members account for approximately 55% of the dataset, while casual riders represent around 45%. This disparity in distribution indicates a noticeable trend: throughout the entire year of 2021, members of the ride-sharing service utilized the platform approximately 10% more frequently than casual riders.

2. Comparison between Members Causal riders depending on ride length (mean, median, minimum, maximum)

```
tripdata_clean %>%
  group_by(member_casual) %>%
  summarise(average_ride_length = mean(ride_length), median_length = median(ride_leng
th),
            max_ride_length = max(ride_length), min_ride_length = min(ride_length))
```

```
## # A tibble: 2 × 5
##   member_casual average_ride_length median_length max_ride_length min_ride_len…¹
##   <chr>                       <dbl>         <dbl>           <dbl>          <dbl>
## 1 casual                      1920.           959         3356649              1
## 2 member                       818.           576           93596              1
## # … with abbreviated variable name ¹min_ride_length
```

- Based on the data presented in the table above, we can conclude that casual riders take longer bike rides than members. This inference is supported by the fact that casual riders' average trip duration or average ride length is higher than member riders. In other words, casual riders, on average, spend more time on each bike trip than members.

3. See total rides and average ride time by each day for members vs. casual riders.

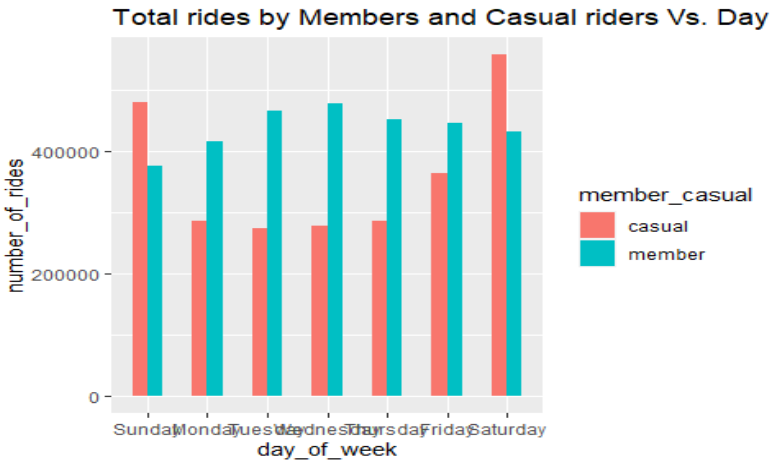
```
#Lets fix the days of the week order.
tripdata_clean$day_of_week <- ordered(tripdata_clean$day_of_week,
                                      levels=c("Sunday", "Monday", "Tuesday", "Wednesda
y", "Thursday", "Friday", "Saturday"))

tripdata_clean %>%
  group_by(member_casual, day_of_week) %>%  #groups by member_casual
```

```
  summarise(number_of_rides = n() #calculates the number of rides and average duratio
n
  ,average_ride_length = mean(ride_length),.groups="drop") %>% # calculates the avera
ge duration
  arrange(member_casual, day_of_week) #sort
```

```
## # A tibble: 14 × 4
##    member_casual day_of_week number_of_rides average_ride_length
##    <chr>         <ord>                <int>               <dbl>
##  1 casual        Sunday              481048               2254.
##  2 casual        Monday              286340               1913.
##  3 casual        Tuesday             274357               1679.
##  4 casual        Wednesday           278910               1660.
##  5 casual        Thursday            286038               1662.
##  6 casual        Friday              364037               1821.
##  7 casual        Saturday            557934               2083.
##  8 member        Sunday              376086                940.
##  9 member        Monday              416181                795.
## 10 member        Tuesday             465474                767.
## 11 member        Wednesday           477117                769.
## 12 member        Thursday            451490                767.
## 13 member        Friday              446384                800.
## 14 member        Saturday            433014                916.
```

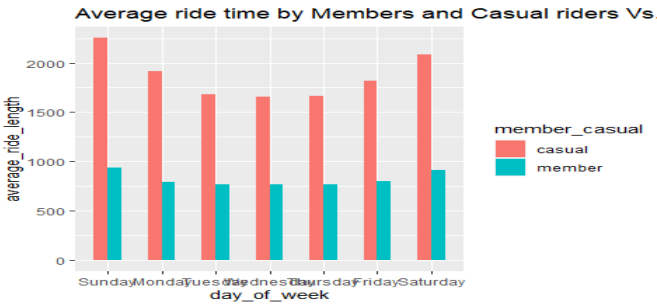
4. Visualizing total rides data by type and day of week

```
tripdata_clean %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), .groups="drop") %>%
  arrange(member_casual, day_of_week)  %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  labs(title ="Total rides by Members and Casual riders Vs. Day of the week") +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



5. Visualizing average ride time data by type and day of week.

```
tripdata_clean %>%
  group_by(member_casual, day_of_week) %>%
  summarise(average_ride_length = mean(ride_length), .groups="drop") %>%
  ggplot(aes(x = day_of_week, y = average_ride_length, fill = member_casual)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title ="Average ride time by Members and Casual riders Vs. Day of the week")
```



Analyzing the first chart provided, it becomes evident that members maintain a consistent frequency of bike trips across the entire week, with a slight decline in rides observed on Sundays. In contrast, casual riders exhibit a distinct pattern: the highest number of rides is recorded over the weekends, with a noticeable surge starting on Fridays and continuing through Saturdays and Sundays.

Furthermore, a substantial discrepancy regarding the average ride length between members and casual riders is noticeable. Members' average ride length is significantly shorter than that of casual riders. This distinction is further highlighted by the fact that the average ride length for casual riders experiences a substantial increase over the weekends, coinciding with the surge in total rides during that period.

This pattern **divergence for casual riders** can be attributed to two interrelated factors: the comparatively **longer average ride lengths on weekends** and the **increased total number of rides** during this period. **For members**, the **average ride length remains relatively constant** throughout the week, with values consistently below 1000 seconds.

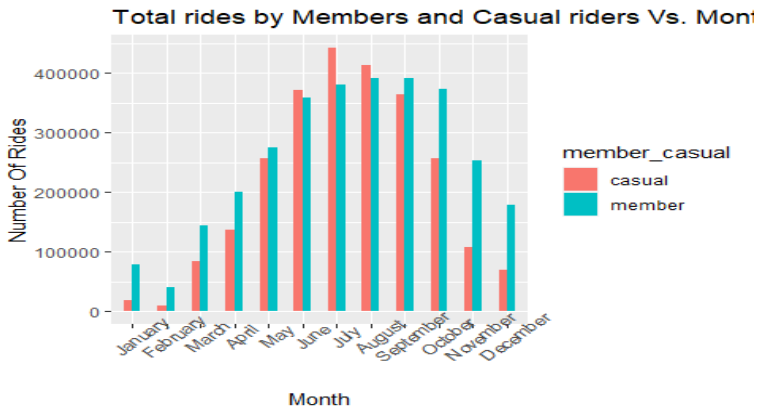
6. See each month's total rides and average ride time for members vs. casual riders.

```
# First lets fix the days of the week order.
tripdata_clean$month <- ordered(tripdata_clean$month,
                                levels=c("January", "February", "March", "April", "May",
"June", "July", "August", "September", "October", "November", "December"))

tripdata_clean %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(), average_ride_length = mean(ride_length), .groups="
drop") %>%
  arrange(member_casual, month)
```

```
## # A tibble: 24 × 4
##    member_casual month     number_of_rides average_ride_length
##    <chr>         <ord>               <int>               <dbl>
##  1 casual        January             18117               1541.
##  2 casual        February            10130               2963.
##  3 casual        March               84028               2290.
##  4 casual        April              136590               2282.
##  5 casual        May                256888               2294.
##  6 casual        June               370636               2228.
##  7 casual        July               442011               1968.
##  8 casual        August             412608               1727.
##  9 casual        September          363840               1669.
## 10 casual        October            257203               1721.
## # … with 14 more rows
```

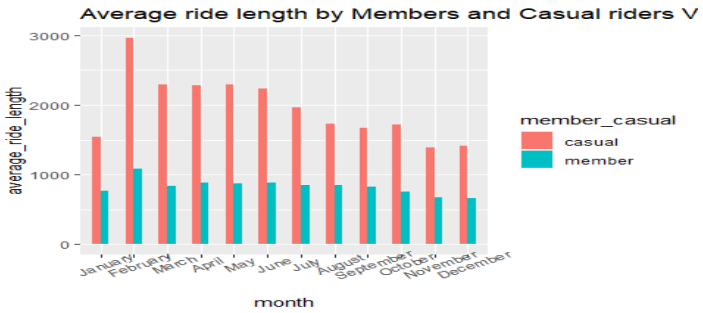
7. Visualizing total rides data by type and month

```
tripdata_clean %>%
  group_by(member_casual, month) %>%
  summarise(number_of_rides = n(),.groups="drop") %>%
  arrange(member_casual, month)  %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  labs(title ="Total rides by Members and Casual riders Vs. Month", x = "Month", y= "
Number Of Rides") +
  theme(axis.text.x = element_text(angle = 45)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
```



8. Visualizing average ride time data by type and month

```
tripdata_clean %>%
  group_by(member_casual, month) %>%
  summarise(average_ride_length = mean(ride_length),.groups="drop") %>%
  ggplot(aes(x = month, y = average_ride_length, fill = member_casual)) +
  geom_col(width=0.5, position = position_dodge(width=0.5)) +
  labs(title ="Average ride length by Members and Casual riders Vs. Month") +
  theme(axis.text.x = element_text(angle = 30))
```



The months of June, July, August, and September emerge as the busiest period of the year for both members and casual riders. This surge in activity is likely attributed to the warmer months, facilitating more favorable riding conditions. Conversely, the winter months of November, December, January, and February exhibit a substantial decline in total rides for both customer segments. It is plausible that the colder weather during this time discourages bike usage.

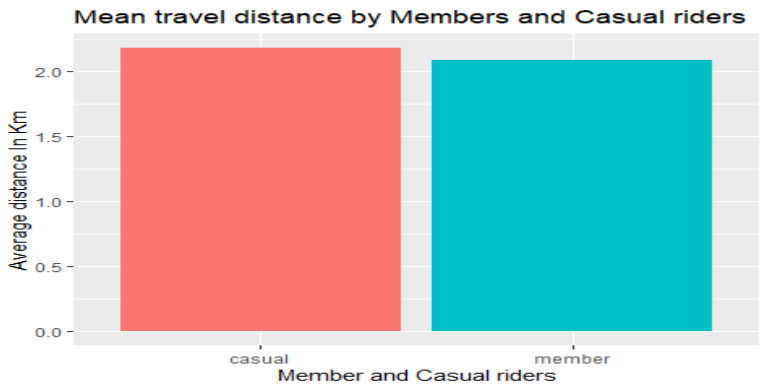
Members consistently maintain higher total ride numbers than casual riders throughout the year, except for June, July, and August. During these summer months, casual riders catch up, potentially owing to vacationers and seasonal riders.

The average ride length for members remains relatively consistent, remaining below 1000 seconds for the entirety of the year. Conversely, casual riders exhibit a broader range in their average ride lengths, typically falling between 1000 to 2000 seconds. An exception is observed in the month of February, where casual riders experience a notable increase in average ride length despite having the lowest total ride count compared to other months.

These observations provide valuable insights into the riding patterns of both member and casual riders throughout the year. The seasonality effect is apparent, with warmer months driving increased ridership, particularly among casual riders. Additionally, the divergence in average ride lengths between members and casual riders and the nuanced fluctuations in different months underscores these distinct customer groups' varying behaviors and preferences.

9. Comparison between Members and Casual riders depending on ride distance

```
tripdata_clean %>%
  group_by(member_casual) %>% drop_na() %>%
  summarise(average_ride_distance = mean(ride_distance)) %>%
  ggplot() +
  geom_col(mapping= aes(x= member_casual,y= average_ride_distance,fill=member_casual)
, show.legend = FALSE)+
  labs(title = "Mean travel distance by Members and Casual riders", x="Member and Cas
ual riders", y="Average distance In Km")
```

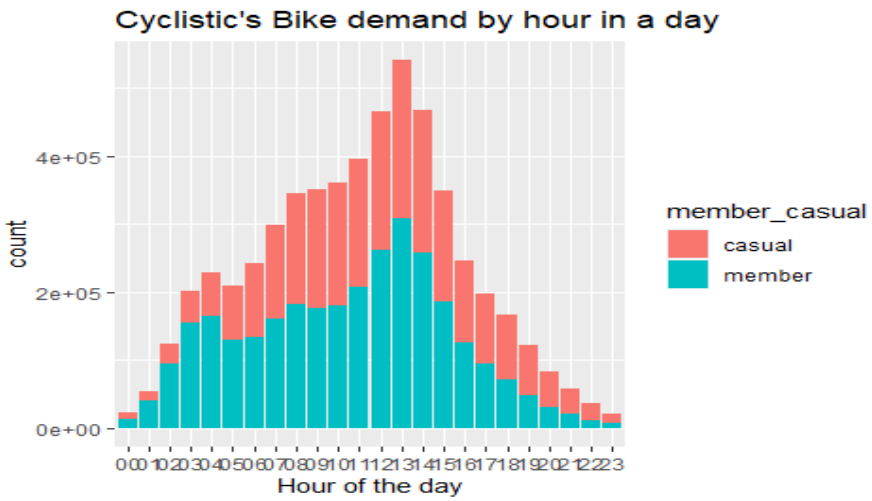


The chart above indicates a remarkable similarity in the average distance members and casual riders traveled. This resemblance in the average distance could be attributed to the following observation: while members engage in rides with consistent durations throughout the week, casual riders predominantly opt for rides during the weekends, typically characterized by longer ride times.

This pattern suggests that even though casual riders ride for longer durations during their weekend trips, the overall similarity in average distance traveled might result from members' consistent ride durations spread across the week. Combining these distinct riding behaviors contributes to the observed parity in average distances between the two rider categories.

10. Analysis and visualization of cyclist's bike demand by the hour in a day

```
tripdata_clean %>%
    ggplot(aes(start_time, fill= member_casual)) +
    labs(x="Hour of the day", title="Cyclistic's Bike demand by hour in a day") +
    geom_bar()
```



The chart above reveals notable differences in the riding patterns between members and casual riders based on the time of day. Specifically:

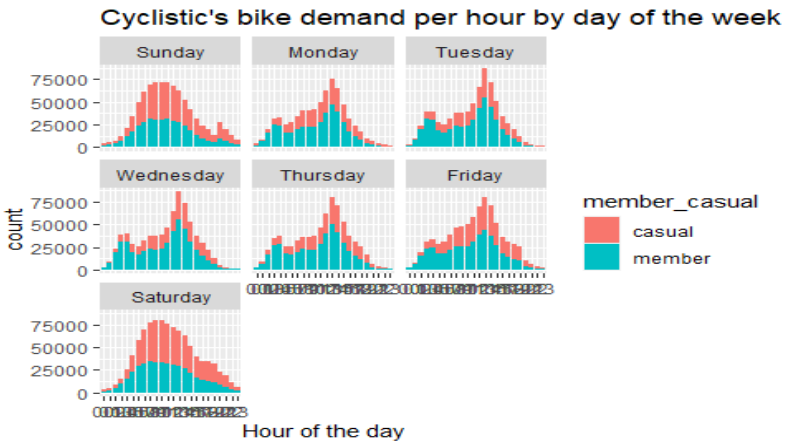
- The hours between 7 am and 11 am exhibit a higher concentration of member riders, indicating that members are more active in the morning hours.
- Conversely, casual riders are more prominent between 3 pm and 12 am, which signifies that they tend to favor the late afternoon and evening time slots for their rides.
- Interestingly, both member and casual riders experience a surge in ride volume during the afternoon hours, indicating a common trend of increased ride activity in the afternoon.

It is important to note that this analysis should be further examined daily to identify any potential variations or anomalies in the observed patterns. Investigating the day-to-day distribution within these time frames could provide deeper insights into the riding habits of both member and casual riders and help uncover any specific trends or fluctuations.

11. Analysis and visualization of cyclistic's bike demand per hour by day of the week

```
tripdata_clean %>%
    ggplot(aes(start_time, fill=member_casual)) +
    geom_bar() +
```

```
    labs(x="Hour of the day", title="Cyclistic's bike demand per hour by day of the w
eek") +
    facet_wrap(~ day_of_week)
```



Indeed, the observed disparities between weekdays and weekends are striking and provide insightful hypotheses about the distinct riding behaviors of members and casual riders:

1. Weekdays (Monday to Thursday)

- A substantial increase in ride volume is evident during the morning hours, specifically between 7 am and 10 am. This pattern suggests that members may use the bike-sharing service as part of their daily routine, potentially for commuting to work or other regular activities.

- Another surge in ride activity is observed between 5 pm and 7 pm, indicating a likely trend of members using the service for their commute back from work.

2. Weekends (Friday, Saturday, and Sunday)

- Notably, a significant spike in ride volume occurs during the weekends, particularly on Friday, Saturday, and Sunday.

- This weekend surge is more prominent among casual riders, which aligns with the hypothesis that casual riders are utilizing the bike-sharing service for leisure and recreational purposes during their weekends.

- The distinct weekend behavior further reinforces the notion that casual riders are more inclined to use the service for non-commuting activities, suggesting a leisure-oriented usage pattern.

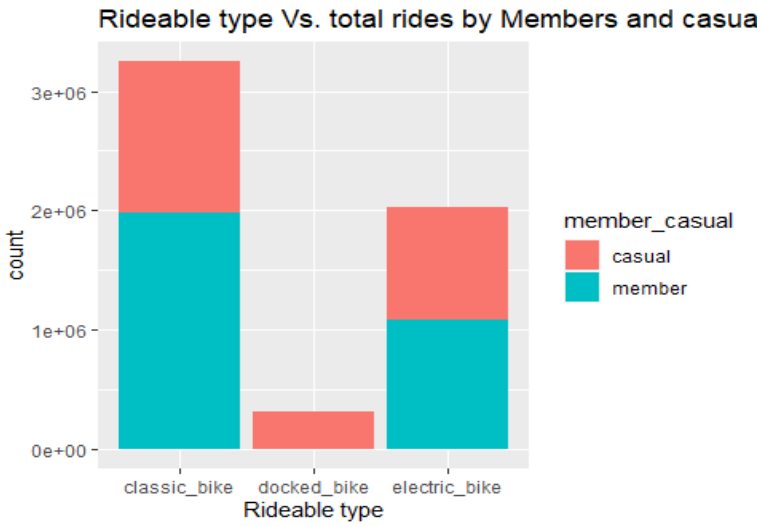
These observations provide a compelling framework for understanding how members and casual riders engage with the bike-sharing service during different days of the week. The clear patterns observed during weekdays and weekends offer valuable insights into these two rider groups' underlying motivations and preferences. Further analysis could validate these hypotheses and shed light on other potential factors influencing riders' behaviors.

12. Analysis and visualization of Rideable type Vs. Total rides by Members and casual riders

```
tripdata_clean %>%
    group_by(rideable_type) %>%
    summarise(count = length(ride_id))
```

```
## # A tibble: 3 × 2
##   rideable_type   count
##   <chr>           <int>
## 1 classic_bike  3250746
## 2 docked_bike    312335
## 3 electric_bike 2031329
```

```
ggplot(tripdata_clean, aes(x=rideable_type, fill=member_casual)) +
    labs(x="Rideable type", title="Rideable type Vs. total rides by Members and casua
l riders") +
    geom_bar()
```



The visualization provided above offers valuable insights into the bike preferences of members and casual riders:

1. Bike Preference for Members:

- Classic bikes are the most commonly used type among members, reflecting a clear preference for this traditional bike option

- Electric bikes are the second most popular choice for members, indicating that these riders also have an affinity for the convenience and advantages electric bikes offer.

2. Bike Preference for Casual Riders:

- Notably, casual riders predominantly opt for docked bikes, suggesting that this type of bike is more favored among this group.

- Electric bikes also see usage among casual riders, although to a lesser extent than members.

These insights provide a nuanced understanding of the bike choices made by members and casual riders. Members' higher utilization of classic bikes and the preference for docked bikes among casual riders may be influenced by factors such as riding preferences, convenience, familiarity, and the specific requirements of their trips. Further analysis and exploration could uncover the motivations and considerations behind these distinct bike preferences among the two rider categories.

13. Analyzing and visualizing the dataset on a coordinate basis

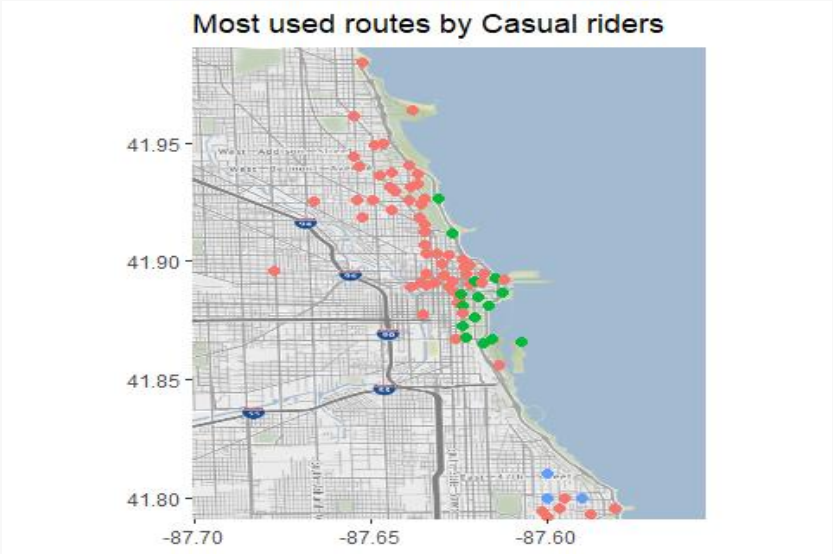
```
#Lets check the coordinates data of the rides.
#adding a new data frame only for the most popular routes >200 rides
coordinates_df <- tripdata_clean %>%
filter(start_lng != end_lng & start_lat != end_lat) %>%
group_by(start_lng, start_lat, end_lng, end_lat, member_casual, rideable_type) %>%
summarise(total_rides = n(),.groups="drop") %>%
filter(total_rides > 200)

# now lets create two different data frames depending on rider type (member_casual)

casual_riders <- coordinates_df %>% filter(member_casual == "casual")
member_riders <- coordinates_df %>% filter(member_casual == "member")
```

Let's setup ggmap and store map of Chicago (bbox, stamen map)

```
chicago <- c(left = -87.700424, bottom = 41.790769, right = -87.554855, top = 41.990119)

chicago_map <- get_stamenmap(bbox = chicago, zoom = 12, maptype = "terrain")

## ℹ Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.
```

Visualization on the Map

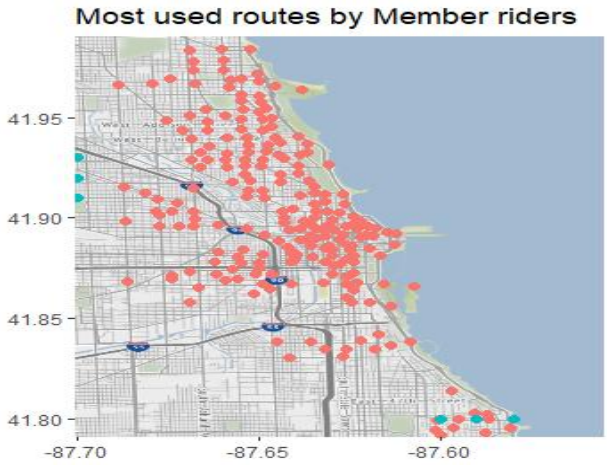
```
# maps on casual riders
ggmap(chicago_map,darken = c(0.1, "white")) +
   geom_point(casual_riders, mapping = aes(x = start_lng, y = start_lat, color=rideable_type), size = 2) +
   coord_fixed(0.8) +
   labs(title = "Most used routes by Casual riders",x=NULL,y=NULL) +
   theme(legend.position="none")

## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.

## Warning: Removed 12 rows containing missing values (`geom_point()`).
```




```
#map on member riders
ggmap(chicago_map,darken = c(0.1, "white")) +
    geom_point(member_riders, mapping = aes(x = start_lng, y = start_lat, color=rideable_type), size = 2) +
    coord_fixed(0.8) +
```

```
    labs(title = "Most used routes by Member riders",x=NULL,y=NULL) +
    theme(legend.position="none")
```

```
## Coordinate system already present. Adding new coordinate system, which will
## replace the existing one.

## Warning: Removed 53 rows containing missing values (`geom_point()`).
```



The geographical distribution depicted in the visualization underscores compelling disparities in the bike usage patterns of casual riders and members:

1. Casual Riders:

- Notably, casual riders predominantly concentrate their bike trips around the central area of the town or the bay area.

- This clustering suggests a distinct pattern of leisure-oriented bike usage, potentially encompassing activities like tourism, sightseeing, and recreational outings.

- The prominent presence of rides around the central area aligns with the hypothesis that casual riders are more likely to engage in bike-sharing for non-work-related activities.

2. Members:

- In contrast, members exhibit a broader spread of bike trips **encompassing** various regions, including the main city area and locations beyond the central zone.

- The wide-ranging distribution hints at a different usage pattern, potentially aligned with daily commuting or other work-related travel needs.

- The diversified geographic coverage further supports the hypothesis that members use bike-sharing for utilitarian purposes, which may involve work-related commutes or travel to different parts of the city.

These observations provide a comprehensive understanding of the spatial behavior of casual riders and members, shedding light on their motivations and intentions behind bike usage. The distinct clustering of casual rides in the central area versus the more dispersed usage by members in various city zones further reinforces the hypotheses of leisurely and utilitarian bike usage, respectively.

Main Insights and Conclusions:

1. Dominance of Membership: Members constitute the largest segment of total rides, surpassing casual riders by approximately 10%. This underscores the significance of the member base in the bike-sharing ecosystem.

2. Consistent Membership Superiority: Across all months, the prevalence of members utilizing the bike-sharing service remains consistently higher than that of casual riders. This showcases the service's enduring appeal to its member community.

3. Weekend Emphasis for Casual Riders: Data analysis highlights a notable concentration of casual rider activity during weekends. This suggests that weekends are a favored period for leisure or recreational bike rides, while members may utilize the service for distinct purposes throughout the weekdays.

4. Afternoon Usage Trends: The discernible spike in bike usage during the afternoon hours points towards a distinct usage pattern. This phenomenon indicates heightened activity during these hours, possibly linked to member commutes or daytime engagements.

5. Work-Related Bike Utilization: The consistent bike usage by members during colder months supports the hypothesis that members could be utilizing the bikes for work-related purposes. In contrast, the decline in casual ridership during colder periods implies reduced leisure usage, reinforcing the likelihood of members relying on the service for practical commuting needs.

Now for how members differ from casuals:

1. Usage Patterns on Weekends: Members generally contribute more data, except on Saturdays and Sundays, where casual riders lead in data points. Weekends are a prime time for casual riders, indicating leisurely or recreational bike usage.

2. Ride Length Disparity: Casual riders tend to have longer ride durations than members. Although the average ride times for members remain relatively consistent, there is a slight increase towards the end of the week.

3. Distinct Time-of-Day Preferences: The morning hours, particularly between 7 am and 10 am, see a higher presence of members. In contrast, casual riders are more active between 3 pm and 12 am, suggesting varying usage patterns throughout the day.

4. Bike Preference and Usage: Members exhibit a stronger inclination towards classic bikes, often followed by electric bikes. Their bike usage appears more routine-oriented. Conversely, casual riders demonstrate a different trend, with a concentration of activities primarily during the weekends, indicating a potential focus on recreational or spontaneous biking.

5. Geographical Distribution: Casual riders tend to spend time near the city center or the bay area, while members are dispersed across the city. This geographical contrast could reflect members' more diversified usage spread versus the casual riders' concentration in specific areas.

Act

Based on my analysis, Cyclistic's executive team, the Director of Marketing (Lily Moreno), and the Marketing Analytics team will do the act phase. (Data-driven decision-making)

The Top three recommendations based on my analysis:

1. Weekend-Exclusive Membership Option: Introduce a specialized membership tier designed exclusively for weekends, distinct from the full annual membership. This can cater to the preferences of casual riders who predominantly use the service on weekends. Pricing this membership differently could attract a new segment of users.

2. Promotions for Electric Bikes and Membership Bundles: Enhance the appeal of annual subscriptions and weekend-only memberships by bundling them with coupons or discounts for electric bike usage. Given the popularity of electric bikes among members and their potential for generating higher revenue, incentivizing casual riders to opt for electric bikes through discounts could drive engagement. Consider reallocating bike inventory by reducing classic bikes in favor of more electric bikes if electric bike usage comes at a premium cost.

3. Strategic Marketing Campaigns: Develop targeted marketing campaigns to emphasize the benefits of annual memberships. These campaigns could be disseminated through email or advertisements at docking stations, timed to coincide with peak months of the year. Highlight the cost savings, convenience, and exclusive perks that annual members enjoy, encouraging potential users to opt for a long-term commitment.

Note: All ride ids are unique, so we cannot conclude if the same rider has taken several rides—more rider data is needed for further analysis.

Additional data that could expand the scope of analysis:

1. Pricing Details and Cost Optimization: Gathering comprehensive pricing details for members and casual riders would allow for a more in-depth cost-benefit analysis. This data could help optimize the cost structure for casual riders by identifying potential areas for discounts or adjustments while ensuring profitability is maintained.

2. Geographical Analysis with Addresses/Neighborhoods: By incorporating address or neighborhood information of members, you could conduct a spatial analysis to identify any location-specific factors influencing membership adoption. This could reveal insights into the demographics, accessibility, or preferences of members based on their residential areas.

3. Recurring User Identification: Introducing a method to determine recurring bike users through payment information or personal identification would enable you to distinguish habitual users from occasional ones. This could provide insights into user loyalty, usage patterns, and preferences, aiding in targeted marketing and service enhancements.