Case Study: How Does a Bike-Share Navigate Speedy Success?

Bhavesh Upadhyay

2022-03-15

Bike Share Report

Welcome to the Cyclistic bike-share analysis case study! In this case study, I will perform many real-world tasks of a junior data analyst. You will work for a fictional company, Cyclistic, and meet different characters and team members. In order to answer the key business questions.

The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, my team wants to **Understand how casual riders and annual members use Cyclistic bikes differently.** From these insights, my team will **design a new marketing strategy to convert casual riders into annual members.** But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

About the company

- In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.
- Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments.One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes,and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.
- Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.
- Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

- For the ask step, first let's get some context from the cyclistic document:

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Guiding questions

- **What is the problem you are trying to solve?**

Ans: Design a new marketing strategy to convert casual riders into annual members, because annual members are much more profitable than casual riders. Maximizing the number of annual members will be key to future growth.

- **How can your insights drive business decisions?**

Ans: Design marketing strategies aimed at converting casual riders into annual members.Maximizing the number of annual members will be key to future growth.

Prepare

- Data set provided from google.
- Download the previous 12 months of Cyclistic trip data https://divvy-tripdata.s3.amazonaws.com/index.html and store to local storage.
- Storing previous 12 months data in 12 variable with list format. After storing, We use glimpse function to view column and datatype format. I have found that column in each variable are same. So bind data in one variable. We perform analysis on that.
- That binded data is stored in RData format.
- Because RData is a format designed for use with R, a system for statistical computation and related graphics, for storing a complete R workspace or selected "objects" from a workspace in a form that can be loaded back by R.

Guiding questions

- **Where is your data located?**

Ans: Download the previous 12 months of Cyclistic trip data https://divvy-tripdata.s3.amazonaws.com/index.html and store to local storage.

- **How is the data organized?**

Ans: Data is organized in different .csv file and uploaded and ordered by year.

- **Are there issues with bias or credibility in this data? Does your data ROCCC?**

Ans: Data is not bias and collected using bikers consent. finally, it's ROCCC because it's reliable, original, comprehensive, current and cited.

- **How are you addressing licensing, privacy, security, and accessibility?**

Ans: The company has their own licence over the dataset. Besides that, the dataset doesn't have any personal information about the riders.

- **How did you verify the data's integrity?** Ans: All the files have consistent columns and each column has the correct type of data.
- **How does it help you answer your question?**

Ans: On that data we have make desicion to get business growth.

- **Are there any problems with the data?**

Ans: All csv files contains same colums. Data contains some missing cells, We have remove to it.

Process

This step will prepare the data for analysis. All the csv files will be merged into one file to improve workflow

Guiding questions

- **What tools are you choosing and why?**

Ans: I using R language and Tableau to make dashboad and easy to take insights. Because of the large dataset and to gather experience with the language.

- **Have you ensured your data's integrity?**

Ans: Yes, the data is consistent throughout the columns.

- **What steps have you taken to ensure that your data is clean?**

Ans: Removing nulls and assign correct datatype to each column.

- **Have you documented your cleaning process so you can review and share those results?**

Ans: Yes, it's all documented in this R notebook.

Setting up environment

Setting up environment by loading 'tidyverse','ggplot', 'lubridate','janitor','dplyr','plyr','rmarkdown','scales'

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
library(dplyr)
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.1.3
```

Extract data from database

Download the previous 12 months of Cyclistic trip data https://divvy-tripdata.s3.amazonaws.com/index.html and store to local storage.

Storing previous 12 months data in 12 variable with list format. After storing, We use glimpse function to view column and datatype format. I have found that column in each variable are same. So bind data in one variable. We perform analysis on that.

That binded data is stored in RData format.

Because RData is a format designed for use with R, a system for statistical computation and related graphics, for storing a complete R workspace or selected "objects" from a workspace in a form that can be loaded back by R.

```
df1 <- read.csv("H:\\analytics\\google\\final csv in r\\202011-divvy-tripdata.csv")
glimpse(df1)
```

```
## Rows: 222,789
## Columns: 19
## $ ride_id            <chr> "BD0A6FF6FFF9B921", "96A7A7A4BDE4F82D", "C61526D065~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <chr> "2020-11-01 13:36:00", "2020-11-01 10:03:26", "2020~
## $ started_at_date    <chr> "2020-11-01", "2020-11-01", "2020-11-01", "2020-11-~
## $ started_at_time    <chr> "13:36:00", "10:03:26", "00:34:05", "00:45:16", "15~
## $ ended_at           <chr> "2020-11-01 13:45:40", "2020-11-01 10:14:45", "2020~
## $ ended_at_date      <chr> "2020-11-01", "2020-11-01", "2020-11-01", "2020-11-~
## $ ended_at_time      <chr> "13:45:40", "10:14:45", "01:03:06", "00:54:31", "16~
## $ ride_length        <chr> "00:09:40", "00:11:19", "00:29:01", "00:09:15", "00~
## $ day_of_week        <int> 1, 1, 1, 1, 1, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ~
## $ start_station_name <chr> "Dearborn St & Erie St", "Franklin St & Illinois St~
## $ start_station_id   <int> 110, 672, 76, 659, 2, 72, 76, 58, 394, 623, 313, 17~
## $ end_station_name   <chr> "St. Clair St & Erie St", "Noble St & Milwaukee Ave~
## $ end_station_id     <int> 211, 29, 41, 185, 2, 76, 72, 288, 273, 2, 313, 301,~
## $ start_lat          <dbl> 41.89418, 41.89096, 41.88098, 41.89550, 41.87650, 4~
```

```
## $ start_lng          <dbl> -87.62913, -87.63534, -87.61675, -87.68201, -87.620~
## $ end_lat            <dbl> 41.89443, 41.90067, 41.87205, 41.91774, 41.87645, 4~
## $ end_lng            <dbl> -87.62338, -87.66248, -87.62955, -87.69139, -87.620~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
df2 <- read.csv("H:\\analytics\\google\\final csv in r\\202012-divvy-tripdata.csv")
df3 <- read.csv("H:\\analytics\\google\\final csv in r\\202101-divvy-tripdata.csv")
df4 <- read.csv("H:\\analytics\\google\\final csv in r\\202102-divvy-tripdata.csv")
df5 <- read.csv("H:\\analytics\\google\\final csv in r\\202103-divvy-tripdata.csv")
df6 <- read.csv("H:\\analytics\\google\\final csv in r\\202104-divvy-tripdata.csv")
df7 <- read.csv("H:\\analytics\\google\\final csv in r\\202105-divvy-tripdata.csv")
df8 <- read.csv("H:\\analytics\\google\\final csv in r\\202106-divvy-tripdata.csv")
df9 <- read.csv("H:\\analytics\\google\\final csv in r\\202107-divvy-tripdata.csv")
df10 <- read.csv("H:\\analytics\\google\\final csv in r\\202108-divvy-tripdata.csv")
df11 <- read.csv("H:\\analytics\\google\\final csv in r\\202109-divvy-tripdata.csv")
df12 <- read.csv("H:\\analytics\\google\\final csv in r\\202110-divvy-tripdata.csv")
bike_ride <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
```

That bike_ride binded data is stored in RData format.

Because RData is a format designed for use with R, a system for statistical computation and related graphics, for storing a complete R workspace or selected "objects" from a workspace in a form that can be loaded back by R.

```
save(bike_ride,file = "H:\\pravin\\bike_ride.RData")
```

Again we use glimpse function on bike_ride and I have found that we have to assign appropriate datatype for each column

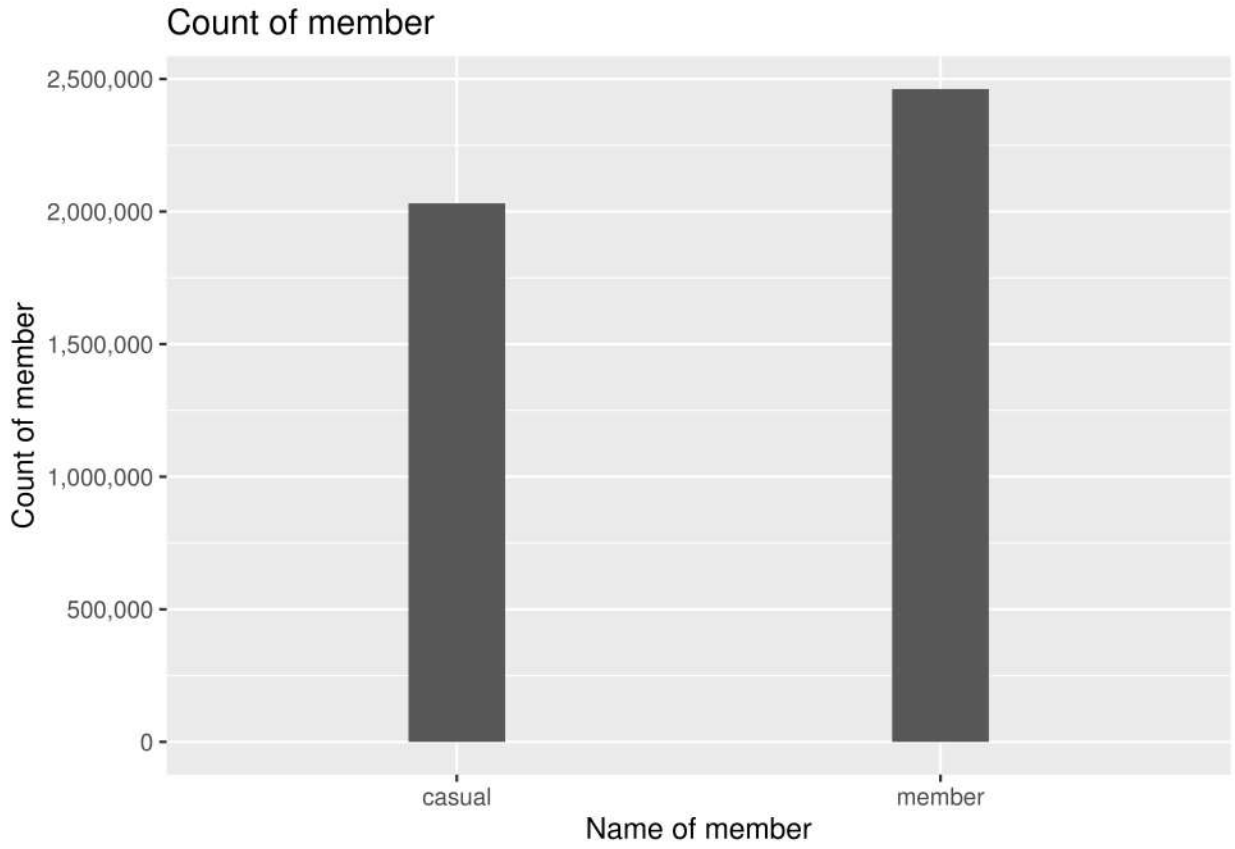
```
bike_ride$started_at <- lubridate::ymd_hms(bike_ride$started_at)
bike_ride$ended_at <- lubridate::ymd_hms(bike_ride$ended_at)

bike_ride$started_at_date <- lubridate::ymd(bike_ride$started_at_date)
bike_ride$ended_at_date <- lubridate::ymd(bike_ride$ended_at_date)

bike_ride$started_at_time <- lubridate::hms(bike_ride$started_at_time)
bike_ride$ended_at_time <- lubridate::hms(bike_ride$ended_at_time)

bike_ride$ride_length <- lubridate::hms(bike_ride$ride_length)

bike_ride$start_hour <-  lubridate::hour(bike_ride$started_at_time)
bike_ride$end_hour <- lubridate::hour(bike_ride$ended_at_time)

bike_ride$start_hour <- as.integer(bike_ride$start_hour)
bike_ride$end_hour <- as.integer(bike_ride$end_hour)
```

The bike_ride RData is cleaned and ready to perform analysis.

Plot of Count of casual and annual member in previous 12 months.

```
ggplot(data = bike_ride)+
  geom_bar(mapping=aes(x=bike_ride$member_casual),width = 0.2)+
  scale_y_continuous(labels = comma)+
  labs(title ="Count of member",x="Name of member",y="Count of member")
```

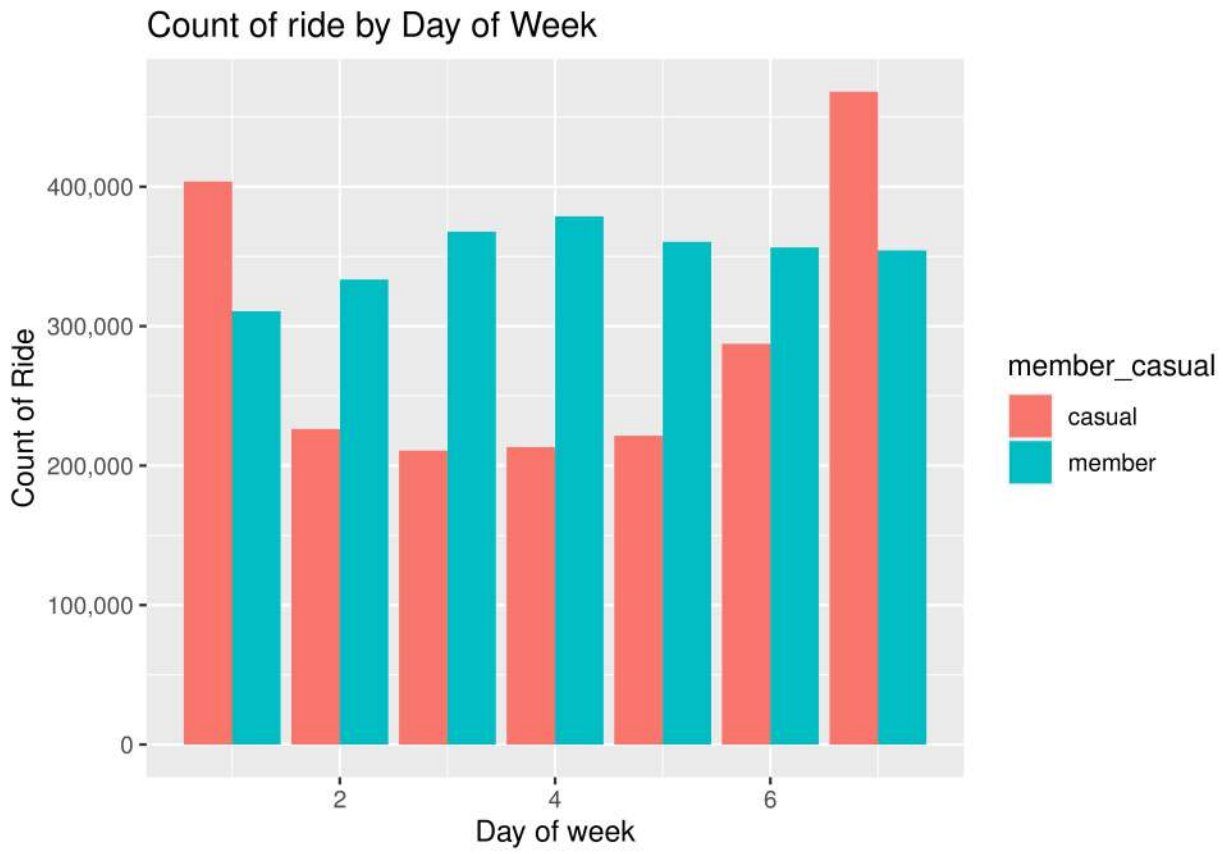
```
## Warning: Use of `bike_ride$member_casual` is discouraged. Use `member_casual`
## instead.
```



Plot of count of ride in Day of week in previous 12 months.

```
bike_ride %>%
  group_by(member_casual,day_of_week)%>%
  dplyr::summarise(no_of_rides = n())%>%
  arrange(member_casual,day_of_week)%>%
  ggplot(aes(x=day_of_week,y=no_of_rides,fill=member_casual))+ geom_col(position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title ="Count of ride by Day of Week",x="Day of week",y="Count of Ride")
```

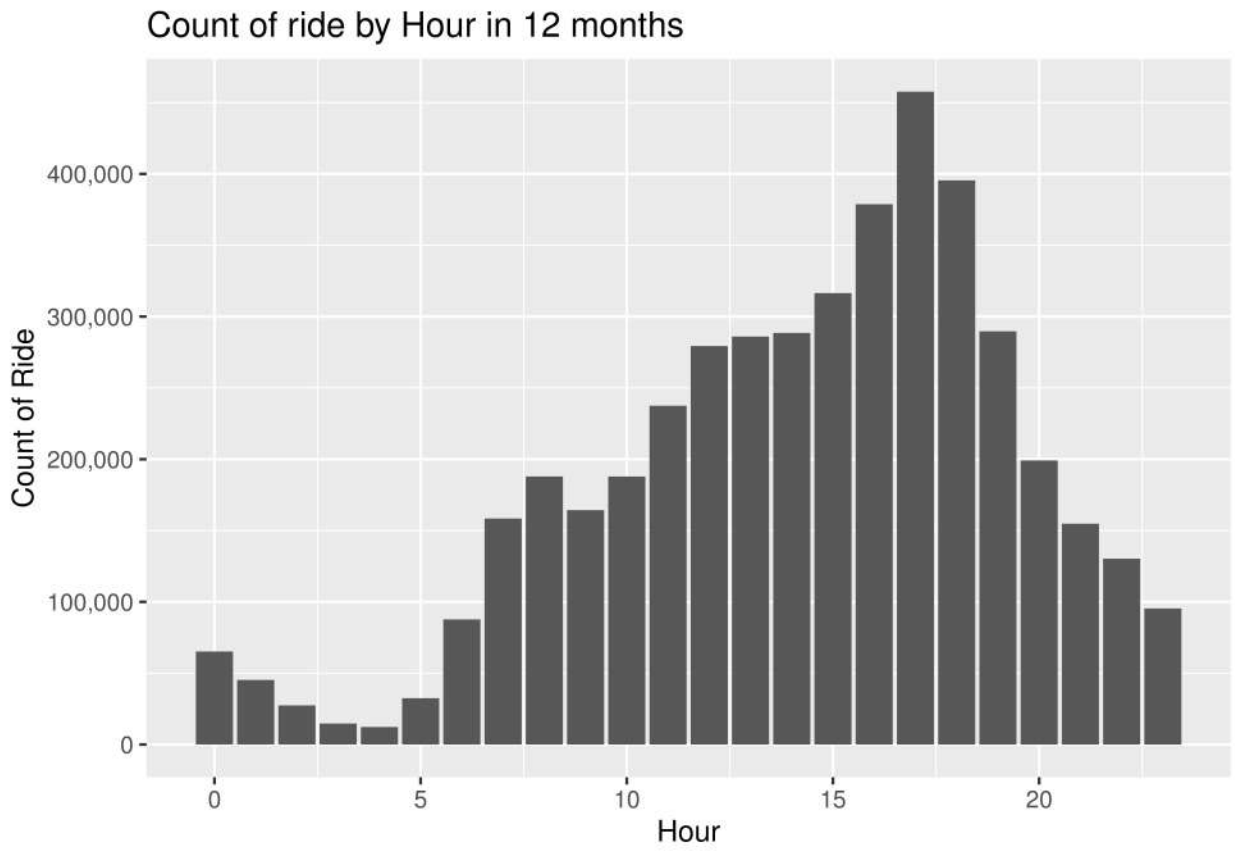
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Plot of count of ride in Hour in previous 12 months.

```
ggplot(data = bike_ride)+
  geom_bar(mapping=aes(x=bike_ride$start_hour))+
  scale_y_continuous(labels = comma)+
  labs(title ="Count of ride by Hour in 12 months",x="Hour",y="Count of Ride")
```

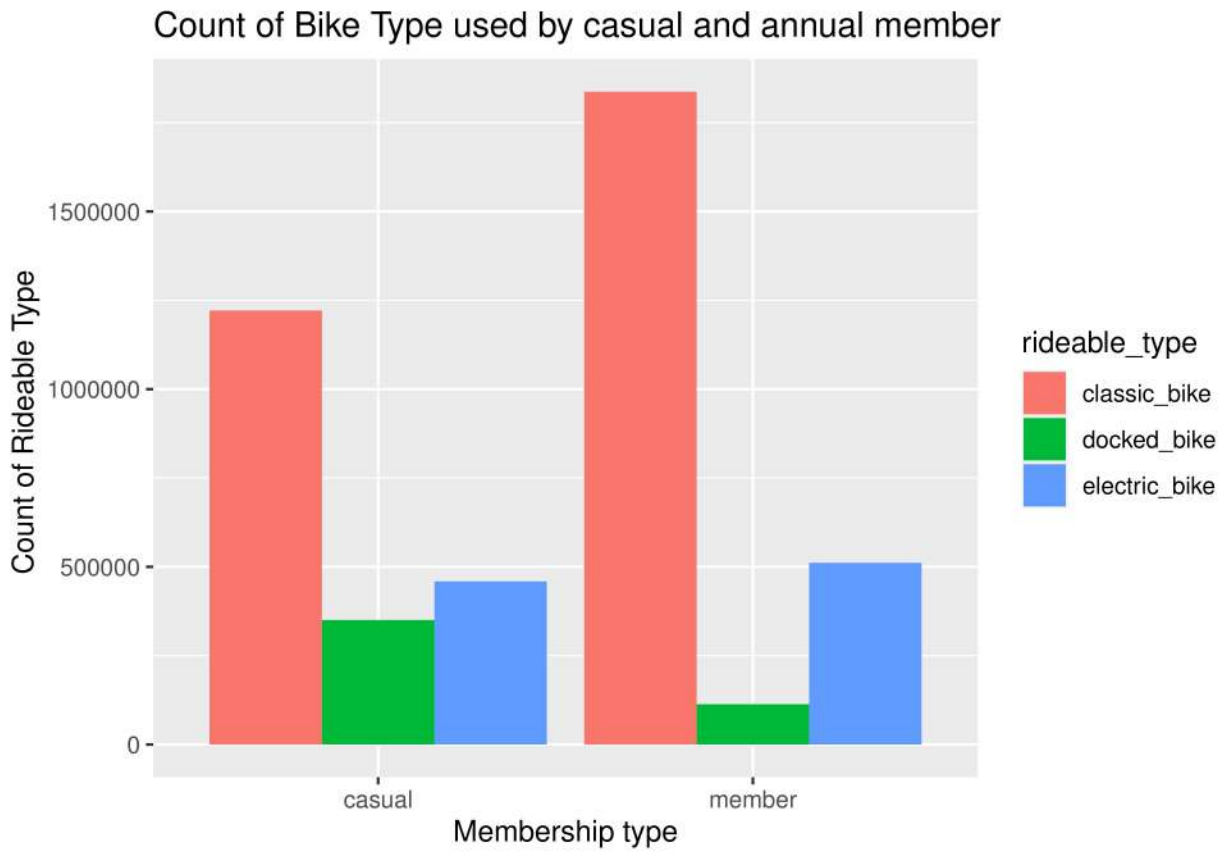
```
## Warning: Use of ‘bike_ride$start_hour‘ is discouraged. Use ‘start_hour‘ instead.
```



Count of Bike Type used by casual and annual member

```
bike_ride %>%
  group_by(member_casual,rideable_type)%>%
  dplyr::summarise(no_of_rides = n())%>%
  arrange(member_casual,rideable_type)%>%
  ggplot(aes(x=member_casual,y=no_of_rides,fill=rideable_type))+
  geom_col(position = "dodge")+
  labs(title ="Count of Bike Type used by casual and annual member",x="Membership type",y="Count of Rid
```

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



Descriptive Analysis / Analyze.

The data exploration will consist of building a profile for annual members and how they differ from casual riders.

```
summary(bike_ride)
```

```
##    ride_id          rideable_type        started_at
##  Length:4492760     Length:4492760     Min.   :2001-09-20 21:00:00
##  Class :character   Class :character   1st Qu.:2021-04-03 13:15:27
##  Mode  :character   Mode  :character   Median :2021-06-18 21:02:57
##                                        Mean   :2020-08-27 12:05:19
##                                        3rd Qu.:2021-08-10 19:36:05
##                                        Max.   :2030-09-20 21:23:59
##  started_at_date      started_at_time
##  Min.   :2020-11-01   Min.   :0S
##  1st Qu.:2021-05-14   1st Qu.:11H 37M 37S
##  Median :2021-07-10   Median :15H 33M 15S
##  Mean   :2021-06-24   Mean   :14H 45M 43.2782721534168S
##  3rd Qu.:2021-08-30   3rd Qu.:18H 19M 49S
##  Max.   :2021-10-31   Max.   :23H 59M 59S
##     ended_at                    ended_at_date
##  Min.   :2001-09-20 21:00:03   Min.   :2020-11-01
##  1st Qu.:2021-04-03 13:46:46   1st Qu.:2021-05-14
```

```
##  Median :2021-06-18 21:28:51   Median :2021-07-10
##  Mean   :2020-08-27 16:24:42   Mean   :2021-06-24
##  3rd Qu.:2021-08-10 19:51:02   3rd Qu.:2021-08-30
##  Max.   :2030-09-20 21:23:59   Max.   :2021-11-03
##  ended_at_time                        ride_length
##  Min.   :0S                          Min.   :0S
##  1st Qu.:11H 50M 55S                 1st Qu.:7M 7S
##  Median :15H 49M 43S                 Median :12M 31S
##  Mean   :14H 56M 4.91495984650101S   Mean   :20M 33.7039149654111S
##  3rd Qu.:18H 34M 24S                 3rd Qu.:22M 40S
##  Max.   :23H 59M 59S                 Max.   :23H 59M 57S
##   day_of_week   start_station_name start_station_id   end_station_name
##  Min.   :1.00   Length:4492760     Length:4492760     Length:4492760
##  1st Qu.:2.00   Class :character   Class :character   Class :character
##  Median :4.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4.11
##  3rd Qu.:6.00
##  Max.   :7.00
##  end_station_id       start_lat       start_lng         end_lat
##  Length:4492760     Min.   :41.65   Min.   :-87.83   Min.   :41.65
##  Class :character   1st Qu.:41.88   1st Qu.:-87.66   1st Qu.:41.88
##  Mode  :character   Median :41.90   Median :-87.64   Median :41.90
##                     Mean   :41.90   Mean   :-87.64   Mean   :41.90
##                     3rd Qu.:41.93   3rd Qu.:-87.63   3rd Qu.:41.93
##                     Max.   :42.06   Max.   :-87.53   Max.   :42.17
##     end_lng       member_casual        start_hour       end_hour
##  Min.   :-87.83   Length:4492760     Min.   : 0.00   Min.   : 0.00
##  1st Qu.:-87.66   Class :character   1st Qu.:11.00   1st Qu.:11.00
##  Median :-87.64   Mode  :character   Median :15.00   Median :15.00
##  Mean   :-87.64                      Mean   :14.26   Mean   :14.43
##  3rd Qu.:-87.63                      3rd Qu.:18.00   3rd Qu.:18.00
##  Max.   :-87.52                      Max.   :23.00   Max.   :23.00
```

Mean and Max of Ride Lenght.

```
summary(bike_ride$ride_length)
```

```
##                   Min.                 1st Qu.                  Median
##                   "0S"                 "7M 7S"               "12M 31S"
##                   Mean                 3rd Qu.                    Max.
## "20M 33.7039149654111S"               "22M 40S"           "23H 59M 57S"
```

```
aggregate(bike_ride$ride_length ~ bike_ride$member_casual+bike_ride$day_of_week,FUN = mean )
```

```
##    bike_ride$member_casual bike_ride$day_of_week bike_ride$ride_length
## 1                   casual                     1              29.40965
## 2                   member                     1              29.38682
## 3                   casual                     2              29.39388
## 4                   member                     2              29.39341
## 5                   casual                     3              29.42607
## 6                   member                     3              29.43181
```

```
## 7                  casual                     4              29.44348
## 8                  member                     4              29.30196
## 9                  casual                     5              29.46907
## 10                 member                     5              29.34649
## 11                 casual                     6              29.43078
## 12                 member                     6              29.38046
## 13                 casual                     7              29.38325
## 14                 member                     7              29.38239
```

Top Ten Start Stations for Casual Riders.

```
bike_ride %>%
  group_by(start_station_name)%>%
  filter(member_casual == "casual")%>%
  dplyr::summarise(n = n())%>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    start_station_name             n
##    <chr>                      <int>
##  1 Streeter Dr & Grand Ave    63193
##  2 Millennium Park            31976
##  3 Michigan Ave & Oak St      28904
##  4 Shedd Aquarium             21863
##  5 Theater on the Lake        21118
##  6 Lake Shore Dr & Monroe St  20808
##  7 Wells St & Concord Ln      18684
##  8 Clark St & Lincoln Ave     16259
##  9 Indiana Ave & Roosevelt Rd 16033
## 10 Wells St & Elm St          15657
```

Top Ten Start Stations for member Riders.

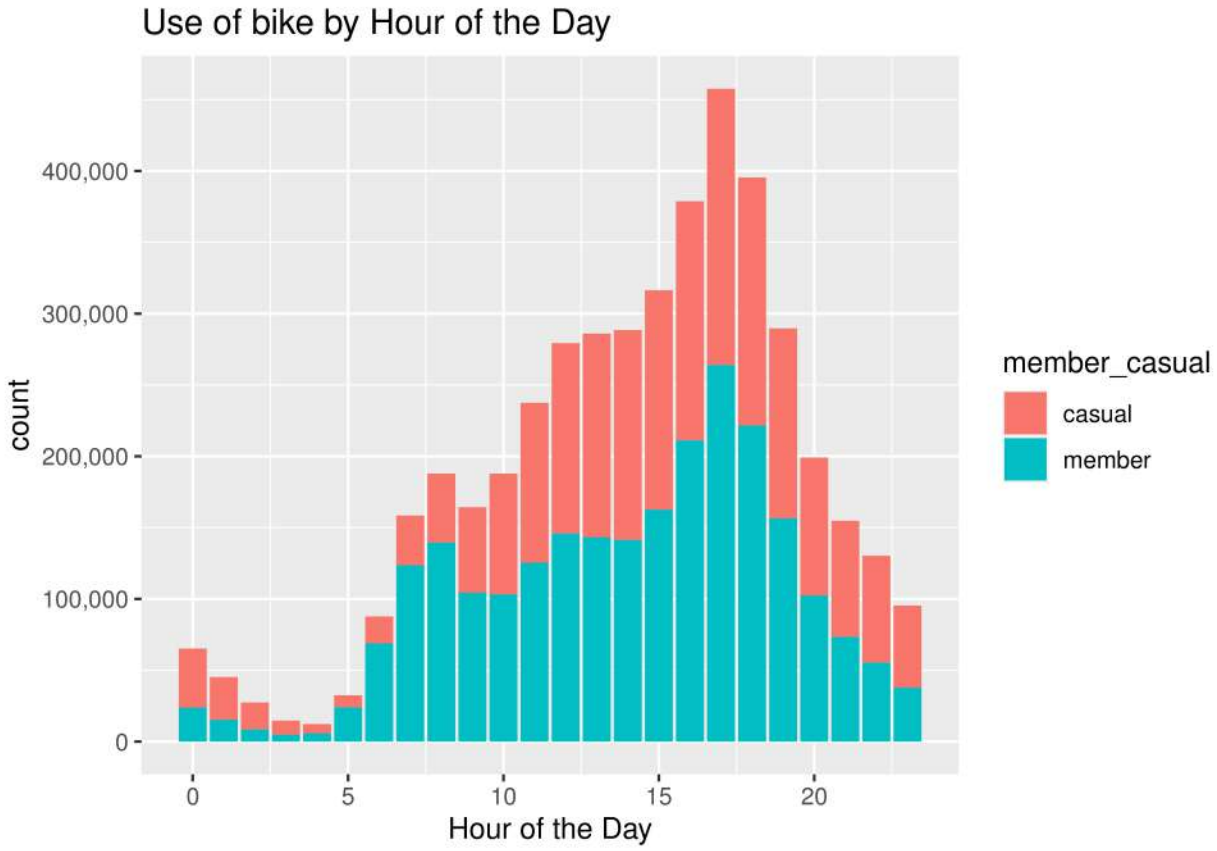
```
bike_ride %>%
  group_by(start_station_name)%>%
  filter(member_casual == "member")%>%
  dplyr::summarise(n = n())%>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    start_station_name           n
##    <chr>                    <int>
##  1 Clark St & Elm St        23520
##  2 Wells St & Concord Ln    22033
##  3 Kingsbury St & Kinzie St 21129
##  4 Wells St & Elm St        19524
##  5 Dearborn St & Erie St    18284
```

```
##  6 St. Clair St & Erie St   17816
##  7 Wells St & Huron St      17803
##  8 Broadway & Barry Ave     17092
##  9 Clark St & Armitage Ave  16102
## 10 Theater on the Lake      15857
```

Use of bike by Hour of the Day

```
bike_ride%>%
  ggplot(aes(start_hour,fill=member_casual))+
  scale_y_continuous(labels = comma)+
  labs(x="Hour of the Day", title="Use of bike by Hour of the Day")+
  geom_bar()
```

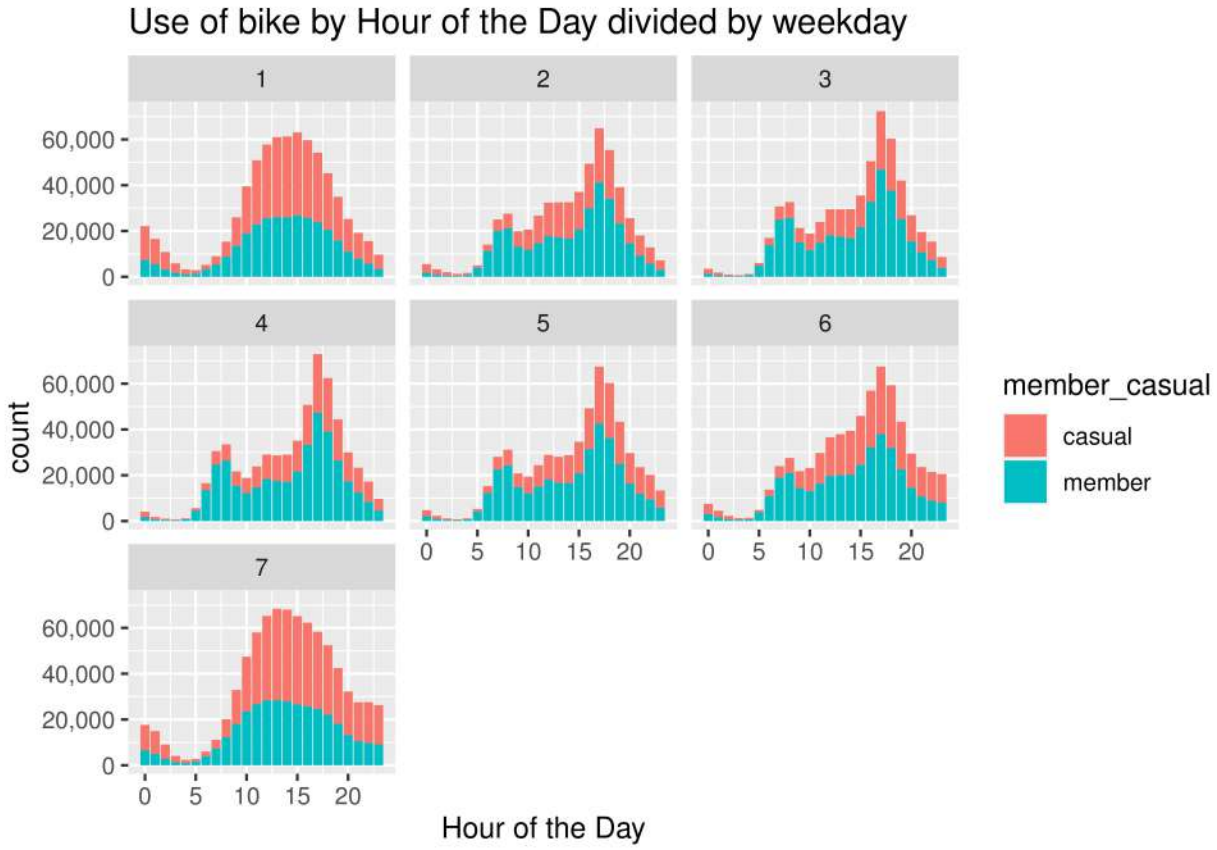


From this chart, we can see that There's a bigger volume of bikers in the 10 to 19 for both casual and member.

This chart can be expanded ween seen it divided by day of the week.

Use of bike by Hour of the Day divided by weekday

```
bike_ride%>%
  ggplot(aes(start_hour,fill=member_casual))+
  scale_y_continuous(labels = comma)+
  labs(x="Hour of the Day", title="Use of bike by Hour of the Day divided by weekday")+
  geom_bar()+
  facet_wrap(~ day_of_week)
```



- It's important to note that:
- Classic bike is most used by casual and annual member.
- Members have a bigger preference for classic bikes.

Guiding questions

- **How should you organize your data to perform analysis on it?***

Ans: All dataset have identical colums, assign proper data-type to columns and concatenate into one dataset.

- **Has your data been properly formatted?**

Ans: Yes, all the columns have their correct data type.

- **What surprises did you discover in the data?**

Ans: One of the main surprises is for day of week 2 to 6 annual member ride bikes between 15 to 22 hour when analysed from weekdays, hours.

- **What trends or relationships did you find in the data?**

Ans:

1. Classic bike is most used by casual and annual member.
2. Casual member bike ride day and time is more on day of week 1 and 7 between 10 to 20 hour.
3. Members use bikes on schedules that differs from casual.
4. Streeter Dr & Grand Ave is most preffered start station for casual riders.
5. Clark St & Elm St is most preffered start station for annual members.

- **How will these insights help answer your business questions?*** Ans: Insights help to increase count of annual members.

Shre

The share phase is usually done by building a presentation/report.

Guiding questions

- **Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?**

Ans: Casual bikers use bike classic bike more. By advertising on digital platform, Posters and in news-paper to reach target audience. Also give discount on annual membership to increase count of annual members.

- **What story does your data tell?**

Ans: The main story the data tells is that members have set schedules, as seen on "Use of bike by Hour of the Day" and "Use of bike by Hour of the Day divided by weekday" charts.

- **How do your findings relate to your original question?**

Ans: By finding the how casual and annual member uses different types of bikes for different purposes.

- **Who is your audience? What is the best way to communicate with them?**

Ans: The main target audience is my Cyclistic Marketing Analytics Team and Lily Moreno. The best way to communicate is through a report presentation of the findings.

- **Can data visualization help you share your findings?**

Ans: Yes, the main core of the finds is through data visualization.

- **Is your presentation accessible to your audience?**

Ans: Yes, the plots were made using vibrant colors, and corresponding labels.

Act

The act phase would be done by the marketing team of the company. The main takeaway will be the top three recommendations for the marketing.

Guiding questions

- **What is your final conclusion based on your analysis?**

Ans: By finding the how casual and annual member uses different types of bikes for different purposes.

- **How could your team and business apply your insights?**

Ans: The insights could be implemented when preparing a marketing campaign for turning casual into members. The marketing can have a focus on workers as a green way to get to work.

- **What next steps would you or your stakeholders take based on your findings?**

Ans: Stakerholder consider insights which taken from visualization. Make strategy to increase annual member.

- **Is there additional data you could use to expand on your findings?** Ans: Climate data gives more help to find how casual and annual member uses bike as per climate change.

Deliverable

*Top three recommendations based on Analysis:

1. Build a marketing campaign focusing on show how bikes help people to get to work, while maintaining the planet green and avoid traffic. The ads could be show on professional social networks.
2. Increase benefits for riding during cold months. Coupons and discounts could be handed out.
3. Also provide advertising holders on public places or at starting and ending station, to reach target audience.

Conclusion

The Google Analytics Professional Certificate teached me a lot and the R language is really useful for analysing data. Rstudio is Great to handle large dataset easily.This took me more time than I expected, but it was extreamly fun.