Cyclistics

Last Modified by: Jayanth T

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

May 18, 2023

Overview

As a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, me and my team want to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, we will design a new marketing strategy to convert casual riders into annual members.

Job role: Junior Data Analyst

I need to perform real world tasks to find the story about the data to answer the questions of stakeholders. This is my process of analyzing the dataset given.

Insights 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.
- Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The bikes can be unlocked from one station and returned to any other station in the system anytime.
- Pricing plans of Cyclists: 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.
- The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Goals

1. Identifying patterns in the company's last year data related to casual and member raids, we need to discover trends and useful info from data enhance the company's sales by delivering insights about the data to make business decisions.

Milestones and Deliverables in each phase

- ☐ **Ask phase**: A clear statement of the business task
- ☐ **Prepare phase:** A description of all data sources used
- ☐ **Process phase**: Documentation of any cleaning or manipulation of data
- ☐ **Analyze phase**: summary of your analysis
- ☐ **Share phase**:Supporting visualizations and key findings
- ☐ **Act phase**: Your top three recommendations based on your analysis

Ask phase:

We need to 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.
We are focusing on the question

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

This question answers the difference between the annual and casual users and how they are using the bikes differently so we can find the trends between them and collect information which helps the stakeholders to better understand the situation and help them to make business decisions.
Primary Stakeholders : Lily Moreno - director, Cyclistic marketing analytics team.
Secondary Stakeholders: Cyclistic executive team.

Prepare Phase:

We need to get the data set. The data set is provided by the company and it contains 12 tables. Each table in the dataset is the data of one month about the raids. Each table contains 13 columns including start and end time, ride_id, membership and ride type.
Data set is downloaded and stored clearly and the tables are verified for consistency and clarity. All the tables are consistent with the same column names. Some contain null values and small errors which can be fixed in the processing phase.

Data Credibility:

- Reliability: Yes
- Consistency: Yes
- Original: Yes
- Comprehensive: Yes
- Cited: Yes

As the data is credible it's time for the next phase.For the next phases I am using R - Studio for further processing the data, analyzing and making data visualizations.

Process phase:

As the dataset is large, using spreadsheets/ Excel is not that helpful so I thought to use mySQL in BigQuery. But as it is a cloud platform and cannot process large amount of data I am using R-Studio. Using R-Studio for Processing:

Installing and Importing packages:

Importing packages tidyverse, lubridate, ggplot2, geosphere

```
# install packages:

install.packages("tidyverse")
install.packages("lubridate")
install.packages("geosphere")
install.packages("ggplot2")

# import libraries:

library(tidyverse)
library(lubridate)
library(geosphere)
library(ggplot2)
```

Loading the dataset and creating year_data:

```
# loading the dataset:

jan <- read.csv("dataset/202201-divvy-tripdata.csv")
feb <- read.csv("dataset/202202-divvy-tripdata.csv")
mar <- read.csv("dataset/202203-divvy-tripdata.csv")
apr <- read.csv("dataset/202204-divvy-tripdata.csv")
may <- read.csv("dataset/202205-divvy-tripdata.csv")
jun <- read.csv("dataset/202206-divvy-tripdata.csv")
jul <- read.csv("dataset/202207-divvy-tripdata.csv")
aug <- read.csv("dataset/202208-divvy-tripdata.csv")
sep <- read.csv("dataset/202209-divvy-tripdata.csv")
oct <- read.csv("dataset/202210-divvy-tripdata.csv")
nov <- read.csv("dataset/202211-divvy-tripdata.csv")
dec <- read.csv("dataset/202212-divvy-tripdata.csv")

# merging all the data to create year data:

year_data <- bind_rows(jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec)
```

Creating new var/ columns day and month in year_data:

```
# creating new columns day and month:

year_data$started_at <- as.POSIXct(year_data$started_at, format = "%Y-%m-%d %H:%M:%S")
year_data$ended_at <- strptime(year_data$ended_at, format = "%Y-%m-%d %H:%M:%S")

year_data <- year_data %>%
  mutate(ride_weekday = wday(started_at, label = TRUE, abbr = FALSE),
         ride_month = month(started_at, label = TRUE, abbr = FALSE))
```

Creating new column time in ride time hours in year_data:

```
# add time in hours form start to end

year_data <- year_data %>%
  mutate(time_hours = as.numeric(difftime(ended_at, started_at, units="hours")))
```

```
$ ride_weekday      <ord> Thursday, Monday, Tuesday, Tuesday, Thursday, Tuesday, Sunday, Saturday, Mon…
$ ride_month        <ord> January, January, January, January, January, January, January, January, Janu…
$ time_hours        <dbl> 0.049166667, 0.072500000, 0.072500000, 0.248888889, 0.100555556, 0.056111111…
```

Creating new column ride_distance

```
## ride_distance:

year_data$ride_distance <- distHaversine(matrix(c(year_data$start_lng, year_data$start_lat)
                                                , ncol = 2),
                                         matrix(c(year_data$end_lng, year_data$end_lat), ncol = 2))

year_data$ride_distance <- year_data$ride_distance/ 1000
```

Basic cleaning:

```
# basic cleaing:
# 1. removing nulls

year_data <- drop_na(year_data)

# 2. removing negative distances

year_data <- year_data %>%
  filter(ride_distance>0)
```

After this process we get a table containing 17 columns

```
> str(year_data)
'data.frame':   5349132 obs. of  17 variables:
 $ ride_id           : chr  "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB80ED419105406" ...
 $ rideable_type     : chr  "electric_bike" "electric_bike" "classic_bike" "classic_bike" ...
 $ started_at        : POSIXct, format: "2022-01-13 11:59:47" "2022-01-10 08:41:56" "2022-01-25 04:53:40"
...
 $ ended_at          : POSIXlt, format: "2022-01-13 12:02:44" "2022-01-10 08:46:17" "2022-01-25 04:58:01"
...
 $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffield Ave & Fullerton
Ave" "Clark St & Bryn Mawr Ave" ...
 $ start_station_id  : chr  "525" "525" "TA1306000016" "KA1504000151" ...
 $ end_station_name  : chr  "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave & Fullerton Ave" "P
aulina St & Montrose Ave" ...
 $ end_station_id    : chr  "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
 $ start_lat         : num  42 42 41.9 42 41.9 ...
 $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ end_lat           : num  42 42 41.9 42 41.9 ...
 $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
 $ member_casual     : chr  "casual" "casual" "member" "casual" ...
 $ ride_weekday      : Ord.factor w/ 7 levels "Sunday"<"Monday"<..: 5 2 3 3 5 3 1 7 2 6 ...
 $ ride_month        : Ord.factor w/ 12 levels "January"<"February"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ time_hours        : num  0.0492 0.0725 0.0725 0.2489 0.1006 ...
 $ ride_distance     : num  0.7 0.695 1.002 2.466 0.815 ...
>
```

As the data is ready and clean now it's time to analyze phase.

Analyze phase:

Data is ready and it is clean. For the analysis phase I am using R-studio.

For analyze phase,

Let us first see how many casual and member rides took place in the last year.

```
> year_data %>% summarize(
+   casual_percentage = sum(member_casual=="casual")*100/n(),
+   member_percentage = sum(member_casual=="member")*100/n())
  casual_percentage member_percentage
1          40.02238          59.97762
>
```

This tells us there are many membership rides than casual rides. It varied with a margin of approximately 20%..

Now let's find the different rideable types used in an year

```
> rideable_types <- unique(year_data$rideable_type)
> print(rideable_types)
[1] "electric_bike" "classic_bike"  "docked_bike"
> year_data %>% summarize(
+   electric_bike_percentage = sum(rideable_type=="electric_bike")*100/n(),
+   classic_bike_percentage = sum(rideable_type=="classic_bike")*100/n(),
+   docked_bike_percentage = sum(rideable_type=="docked_bike")*100/n()
+ )
  electric_bike_percentage classic_bike_percentage docked_bike_percentage
1                  51.7019                45.61628                2.681818
>
```

This tells us there are mainly 3 bike rides and electric bikes and classic bikes have more usage than docked bikes. Classic bikes are used 45% and electric bikes are used 5 more percent by the users.

Now let us group the rideable types by their membership

```
> year_data %>% group_by(member_casual) %>% summarize(
+   electric_bike_percentage = sum(rideable_type=="electric_bike")*100/n(),
+   classic_bike_percentage = sum(rideable_type=="classic_bike")*100/n(),
+   docked_bike_percentage = sum(rideable_type=="docked_bike")*100/n()
+ )
# A tibble: 2 × 4
  member_casual electric_bike_percentage classic_bike_percentage docked_bike_percentage
  <chr>                            <dbl>                   <dbl>                  <dbl>
1 casual                            55.7                    37.6                   6.70
2 member                            49.0                    51.0                   0
>
```

This tells us that casual members used electric bikes more often around 56% and then electric bikes around 38% and they also used docked bikes around 7%.

But members haven't used docked bikes at all, they only used classic and electric bikes but the difference is very less.

Now let us see which monthly rides and rides based on the weekday

For monthly:

```
> monthly_summarize_data <- year_data %>% summarise(
+   Jan = sum(ride_month=="January")*100/n(),
+   Feb = sum(ride_month=="February")*100/n(),
+   Mar = sum(ride_month=="March")*100/n(),
+   Apr = sum(ride_month=="April")*100/n(),
+   May = sum(ride_month=="May")*100/n(),
+   Jun = sum(ride_month=="June")*100/n(),
+   Jul = sum(ride_month=="July")*100/n(),
+   Aug = sum(ride_month=="August")*100/n(),
+   Sep = sum(ride_month=="September")*100/n(),
+   Oct = sum(ride_month=="October")*100/n(),
+   Nov = sum(ride_month=="November")*100/n(),
+   Dec = sum(ride_month=="December")*100/n(),
+ )
>
> print(monthly_summarize_data)
       Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
1 1.838579 2.031862 4.967572 6.453197 11.08053 13.53395 14.50834 13.90212 12.45529 9.950568 6.02881 3.249181
```

Group by membership:

```
> monthly_summarize_groupby_member_data <- year_data %>% group_by(member_casual) %>% summarise(
+   Jan = sum(ride_month=="January")*100/n(),
+   Feb = sum(ride_month=="February")*100/n(),
+   Mar = sum(ride_month=="March")*100/n(),
+   Apr = sum(ride_month=="April")*100/n(),
+   May = sum(ride_month=="May")*100/n(),
+   Jun = sum(ride_month=="June")*100/n(),
+   Jul = sum(ride_month=="July")*100/n(),
+   Aug = sum(ride_month=="August")*100/n(),
+   Sep = sum(ride_month=="September")*100/n(),
+   Oct = sum(ride_month=="October")*100/n(),
+   Nov = sum(ride_month=="November")*100/n(),
+   Dec = sum(ride_month=="December")*100/n(),
+ )
>
> monthly_summarize_groupby_member_data
# A tibble: 2 × 13
  member_casual   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
  <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 casual        0.795 0.915  3.80  5.37  11.9  15.8  17.5  15.5  12.9  9.11  4.38  1.96
2 member        2.53  2.78   5.74  7.17  10.5  12.0  12.5  12.8  12.2 10.5   7.13  4.11
```

By this we can conclude that the number of rides in January is less and then it increases gradually and peaks in summer(May, June and July) and then decreases by December. Same trend is followed by both casual and member raids.

For weekly:

```
> weekly_summarize_data <- year_data %>% summarise(
+   Mon = sum(ride_weekday=="Monday")*100/n(),
+   Tue = sum(ride_weekday=="Tuesday")*100/n(),
+   Wed = sum(ride_weekday=="Wednesday")*100/n(),
+   Thu = sum(ride_weekday=="Thursday")*100/n(),
+   Fri = sum(ride_weekday=="Friday")*100/n(),
+   Sat = sum(ride_weekday=="Saturday")*100/n(),
+   Sun = sum(ride_weekday=="Sunday")*100/n(),
+ )
> weekly_summarize_data
       Mon      Tue      Wed      Thu      Fri      Sat      Sun
1 13.22716 13.89498 14.19133 14.93743 14.17828 16.05743 13.51339
```

Group by membership:

```
> weekly_summarize_groupby_member_data <- year_data %>% group_by(member_casual) %>% summarise(
+   Mon = sum(ride_weekday=="Monday")*100/n(),
+   Tue = sum(ride_weekday=="Tuesday")*100/n(),
+   Wed = sum(ride_weekday=="Wednesday")*100/n(),
+   Thu = sum(ride_weekday=="Thursday")*100/n(),
+   Fri = sum(ride_weekday=="Friday")*100/n(),
+   Sat = sum(ride_weekday=="Saturday")*100/n(),
+   Sun = sum(ride_weekday=="Sunday")*100/n(),
+ )
>
> weekly_summarize_groupby_member_data
# A tibble: 2 × 8
  member_casual   Mon   Tue   Wed   Thu   Fri   Sat   Sun
  <chr>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 casual         11.9  11.4  11.9  13.4  14.5  20.3  16.6
2 member         14.1  15.6  15.7  16.0  14.0  13.2  11.5
```

This tells us that every day the rides are almost the same. It Should be noted that Saturday has slightly more rides than other weekdays.

Average ride time by all rides:

```
> avg_ridetime <- year_data %>% summarise(
+   avg_ride_time_in_mins = mean(ride_time_mins)
+ )
>
> avg_ridetime
  avg_ride_time_in_mins
1              15.93793
>
```

Average ride time group by membership and ride types:

```
> avg_ridetime_group <- year_data %>% group_by(member_casual, rideable_type) %>%
+   summarise(avg_ride_time_in_mins = mean(ride_time_mins)
+ )
`summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
>
> avg_ridetime_group
# A tibble: 5 x 3
# Groups:   member_casual [2]
  member_casual rideable_type avg_ride_time_in_mins
  <chr>         <chr>                         <dbl>
1 casual        classic_bike                   23.4
2 casual        docked_bike                    47.7
3 casual        electric_bike                  16.5
4 member        classic_bike                   13.2
5 member        electric_bike                  11.6
>
```

Now the analyze phase is completed,

These are some key points:

- Electric bikes and casual bikes are more used by the riders.
- Docked bikes are used by only casual members and these bikes are used in less percentage of rides.
- The total rides in a year start very low and increase monthly and have the highest number of rides in months of summer and then they gradually decrease.
- Casual riders use the bikes more than members in time.

Now as the analyze phase is completed it is time for Share phase,

Share phase:

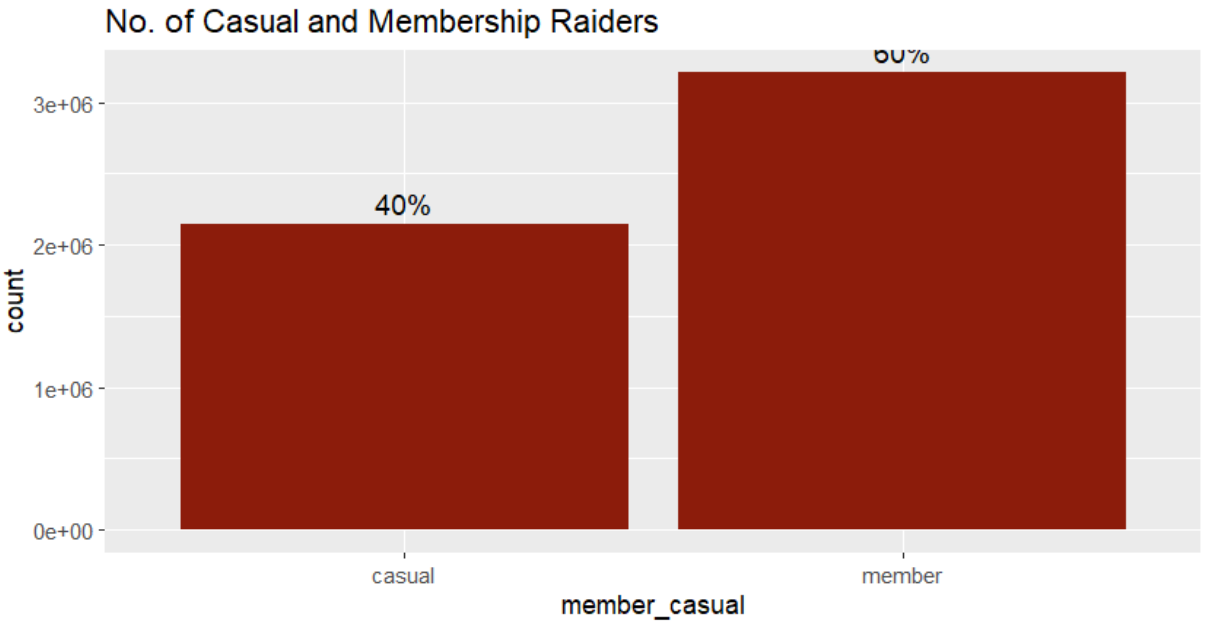
Now it's time for the share phase,

Tools used: R-Studio

Data is analyzed and we found some key insights and some major points from the data. It's time to make some visualizations to share these with the stakeholders. I am using R-studio to build these visualizations and after making them I will make a presentation to share it with stakeholders.

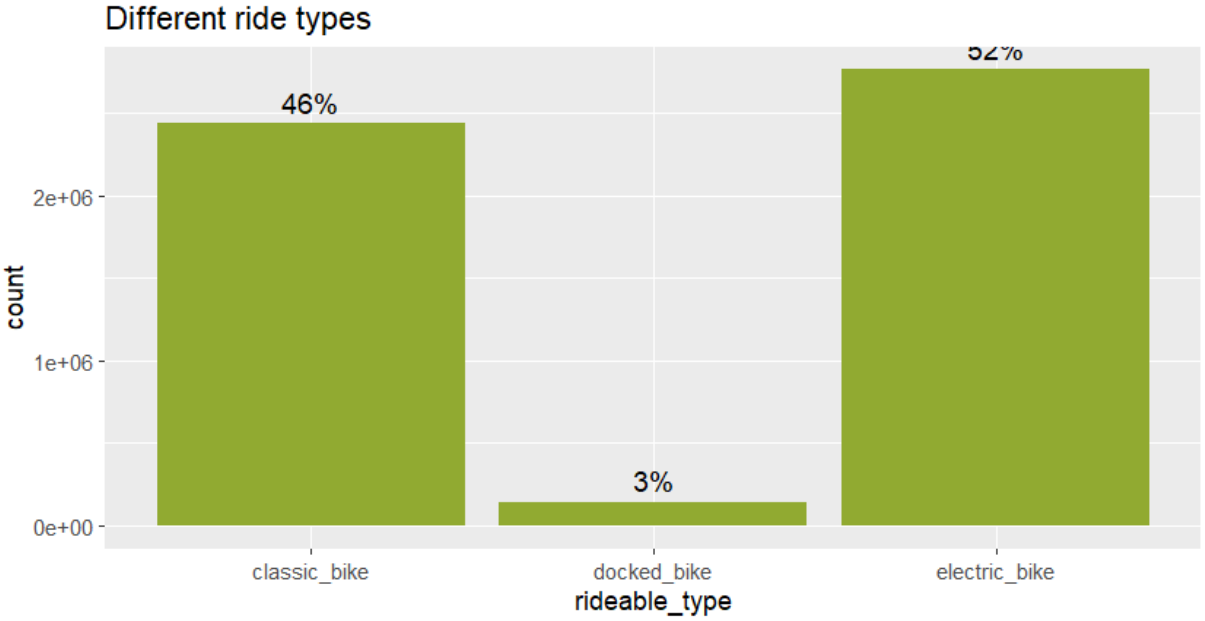
No.of casual and membership raiders

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x = member_casual, y = ..count..), stat = "count", fill="#8c1c0b") +
+   geom_text(mapping = aes(x = member_casual, y = ..count.., label = paste0(round((..count.. / sum
(..count..)) * 100), "%")),
+     stat = "count",vjust = -0.5,size = 4) +
+   labs(title = "No. of Casual and Membership Raiders")
>
```



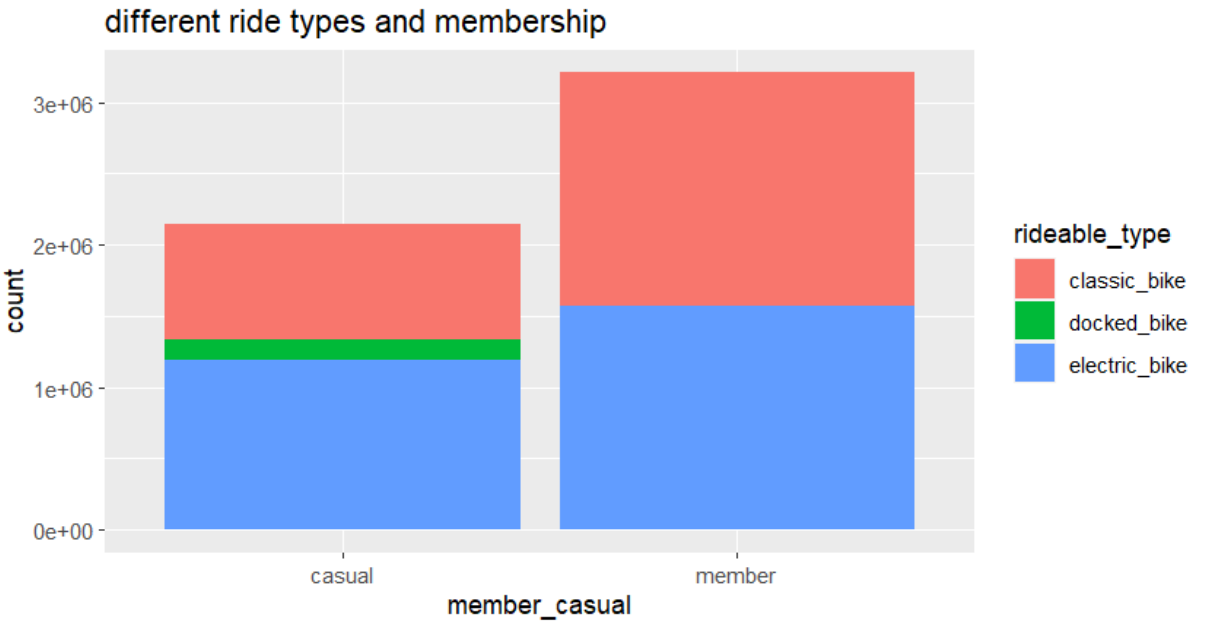
Different types of rides:

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x=rideable_type, y = ..count..), stat = "count" ,fill = "#91aa31") +
+   geom_text(mapping = aes(x=rideable_type, y = ..count.., label = paste0(round((..count.. / sum(..c
ount..)) * 100), "%")),
+             stat = "count",vjust = -0.5,size = 4)+
+   labs(title = "Different ride types")
>
```



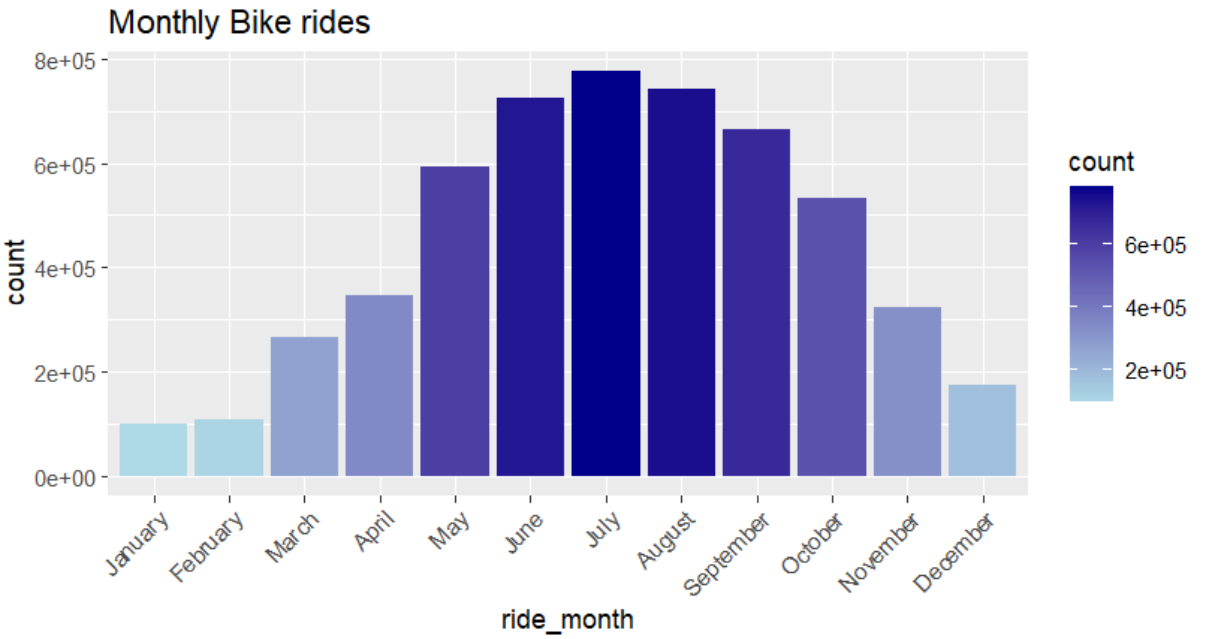
Different ride types that are used by members and casual riders:

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x = member_casual, fill = rideable_type))+
+   labs(title = "different ride types and membership")
> |
```



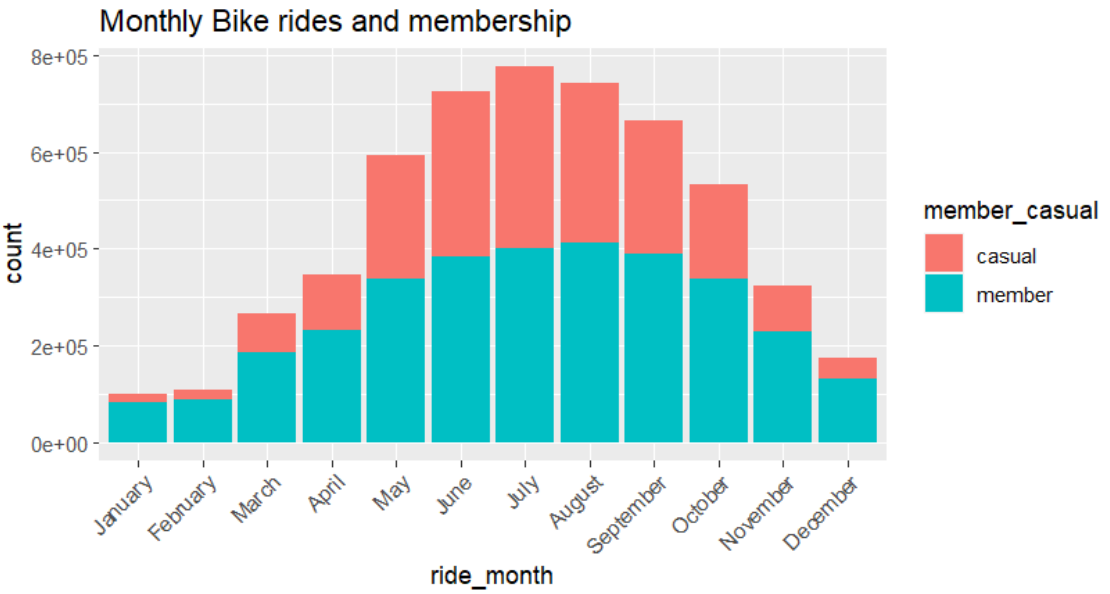
Monthly - bike rides:

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x=ride_month, y = ..count.., fill = after_stat(count))
+            , stat = "count") +
+   scale_fill_gradient(low = "lightblue", high = "darkblue") +
+   labs(title = "Monthly Bike rides")+
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
> |
```



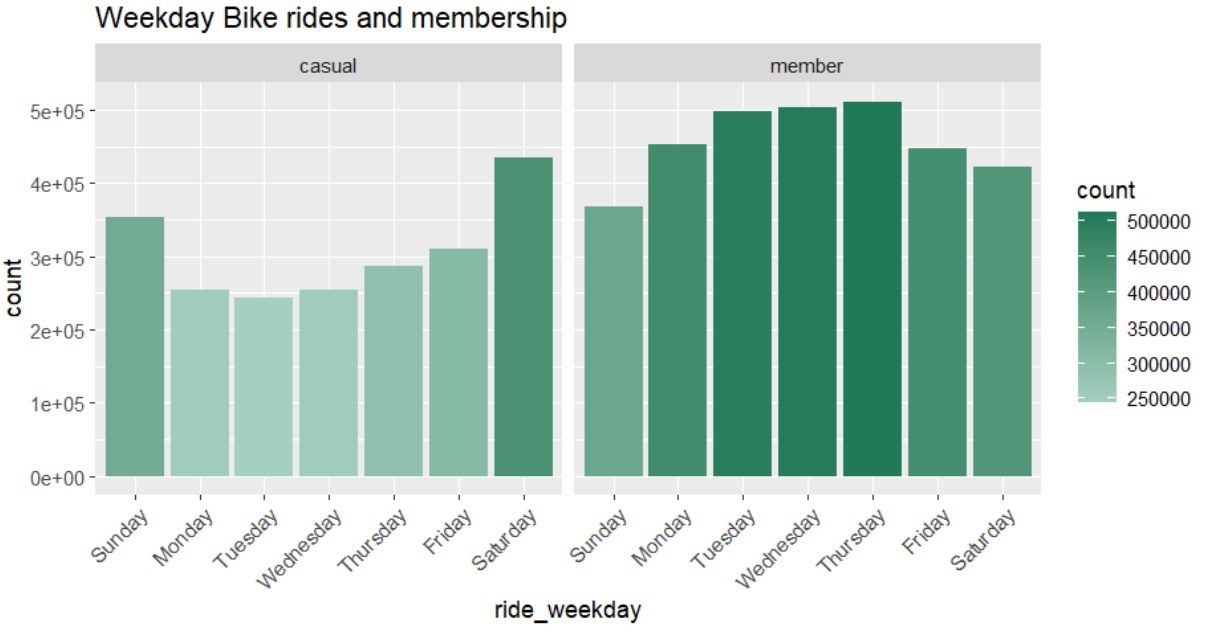
Monthly - bike rides and Membership:

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x=ride_month, fill=member_casual),
+            stat = "count") +
+   labs(title = "Monthly Bike rides and membership")+
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Weekday - no.of raids and membership:

```
> ggplot(data = year_data) +
+   geom_bar(mapping = aes(x=ride_weekday, y = ..count.., fill = after_stat(count)),
+            stat = "count") +
+   labs(title = "Weekday Bike rides and membership")+
+   scale_fill_gradient(low = "#a5cfbf", high = "#207a58") +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))+
+   facet_wrap(~member_casual)
>
```



With this we wrap up the Share phase.
Now it's time for the act phase.

Act phase:

In the act phase the stakeholders take action based on the data and recommendations we provided.
Recommendations:

- Annual members will be more likely to use bikes every day.
- Adding docked bikes to annual membership may increase people to buy the annual membership. As its stake is very low, the company can make decisions
- Providing more bikes in the summer season will be helpful to the company.
- Providing free trails and some discounts in the summer season.

This might help the company to increase annual memberships.

Thank you.