cyclist bike company

ali abdelfattah

2023-11-29

Introduction

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, your 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 our 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.

Key stakeholders

1- Cyclistic executive team (primary stakeholder)

2- Lily Moreno (director of marketing) & Marketing Analytics team (colleagues)

(seconday stakeholder)

Deliverables

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings

6.top three recommendations based on my analysis

business task

1-analyzing the different between the casual bike riders and annual members in using Cyclistic

2-identifing how to increase the number of annual members, by designing a new marketing strategy

3-produce a visualization to capture the analysis process and solution

statment

our study will analyze the difference between annual members and casual users in different ways to investigate the cause of casual users not subscribing to annual membership then we will recommend how to convince casual users to subscribe and how to promote our new marketing campaign

data sources used

Credible data = ROCCC: Reliable, Original, Comprehensive, Current, Cited

1. reliable:the data is not biased but the September month is missing (working with 11 months of data)

2.Original:the original data scores is known and can be located (created by the company)

3.Comprehensive:the data is missing information so it is not comprehensiv

4.current:the data is not updated last update

5.cited:the data is cited information scores is known

Cyclistic bike-share analysis

importing important libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
##
```

```
library(ggplot2)
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(forcats)
```

data exploration

In this part we import the monthly data and combining it into one data frame then we remove empty columns and rows,convert wrong data types and create new columns that will guide our analysis

Import 12 most recent monthly datasets (August 2022 to July 2023) into RStudio using read.csv function

```
df1 <- read.csv("data-cyclist/202004-divvy-tripdata.csv")
df2 <- read.csv("data-cyclist/202005-divvy-tripdata.csv")
df3 <- read.csv("data-cyclist/202006-divvy-tripdata.csv")
df4 <- read.csv("data-cyclist/202007-divvy-tripdata.csv")
df5 <- read.csv("data-cyclist/202008-divvy-tripdata.csv" )
df6 <- read.csv( "data-cyclist/202009-divvy-tripdata.csv")
df7 <- read.csv("data-cyclist/202010-divvy-tripdata.csv")
df8 <- read.csv("data-cyclist/202011-divvy-tripdata.csv")
df9 <- read.csv("data-cyclist/202012-divvy-tripdata.csv")
df10 <- read.csv( "data-cyclist/202101-divvy-tripdata.csv")
df11 <- read.csv("data-cyclist/202102-divvy-tripdata.csv")
df12 <- read.csv("data-cyclist/202103-divvy-tripdata.csv")
```

Combinig the data

You can also embed plots, for example:

```
bike_rides_for_year<-rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
dim(bike_rides_for_year)
```

```
## [1] 3489748      13
```

removing empty cols and rows

```
bike_rides_for_year <- janitor::remove_empty(bike_rides_for_year,which = c("cols"))
bike_rides_for_year <- janitor::remove_empty(bike_rides_for_year,which =c("rows"))

bike_rides_for_year <- bike_rides_for_year %>%
  filter(start_station_name!="") %>%
  filter(member_casual!="") %>%
  filter(rideable_type!="") %>%
  filter(start_station_id !="") %>%
  filter(end_station_name!="") %>%
  filter(end_station_id!="")


dim(bike_rides_for_year)
```

```
## [1] 3294691      13
```

changing data types to date/time and creating new coloumns for houers and dates

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



bike_rides_for_year$start_hour <- lubridate::hour(bike_rides_for_year$started_at)
bike_rides_for_year$end_hour <- lubridate::hour(bike_rides_for_year$ended_at)

bike_rides_for_year$start_date <- as.Date(bike_rides_for_year$started_at)
bike_rides_for_year$end_date <- as.Date(bike_rides_for_year$ended_at)

bike_rides_for_year$week_days <- as.character(
lubridate::wday(bike_rides_for_year$start_date,label = T ,abbr = F))


bike_rides_for_year$trip_duration_second <-
  difftime(bike_rides_for_year$ended_at,bike_rides_for_year$started_at)

bike_rides_for_year$trip_duration_miniuts <-
  difftime(bike_rides_for_year$ended_at,bike_rides_for_year$started_at,units="mins")

bike_rides_for_year$trip_duration_hours <-
  difftime(bike_rides_for_year$ended_at,bike_rides_for_year$started_at,units="hours")


glimpse(bike_rides_for_year)
```

```
## Rows: 3,294,691
## Columns: 21
## $ ride_id              <chr> "A847FADBBC638E45", "5405B80E996FF60D", "5DD24A7…
## $ rideable type        <chr> "docked bike", "docked bike", "docked bike", "do…
```

```
## $ started_at            <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54, 2020-…
## $ ended_at              <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03, 2020-…
## $ start_station_name    <chr> "Eckhart Park", "Drake Ave & Fullerton Ave", "Mc…
## $ start_station_id      <chr> "86", "503", "142", "216", "125", "173", "35", "…
## $ end_station_name      <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko Park"…
## $ end_station_id        <chr> "152", "499", "255", "657", "323", "35", "635", …
## $ start_lat             <dbl> 41.8964, 41.9244, 41.8945, 41.9030, 41.8902, 41.…
## $ start_lng             <dbl> -87.6610, -87.7154, -87.6179, -87.6975, -87.6262…
## $ end_lat               <dbl> 41.9322, 41.9306, 41.8679, 41.8992, 41.9695, 41.…
## $ end_lng               <dbl> -87.6586, -87.7238, -87.6230, -87.6722, -87.6547…
## $ member_casual         <chr> "member", "member", "member", "member", "casual"…
## $ start_hour            <int> 17, 17, 17, 12, 10, 17, 14, 12, 10, 15, 15, 15, …
## $ end_hour              <int> 18, 17, 18, 13, 11, 18, 14, 13, 10, 15, 15, 15, …
## $ start_date            <date> 2020-04-26, 2020-04-17, 2020-04-01, 2020-04-07,…
## $ end_date              <date> 2020-04-26, 2020-04-17, 2020-04-01, 2020-04-07,…
## $ week_days             <chr> "Sunday", "Friday", "Wednesday", "Tuesday", "Sat…
## $ trip_duration_second  <drtn> 1609 secs, 489 secs, 863 secs, 732 secs, 3175 s…
## $ trip_duration_miniuts <drtn> 26.816667 mins, 8.150000 mins, 14.383333 mins, …
## $ trip_duration_hours   <drtn> 0.44694444 hours, 0.13583333 hours, 0.23972222 …
```

removing duplicates and wrong trip_duration and null values

```
df<-bike_rides_for_year %>% filter(trip_duration_second>0)%>% filter(trip_duration_miniuts<(720)
) %>% na.omit()
dim(df)
```

```
## [1] 3278631      21
```

```
df<- distinct(df,ride_id,.keep_all = TRUE)
dim(df)
```

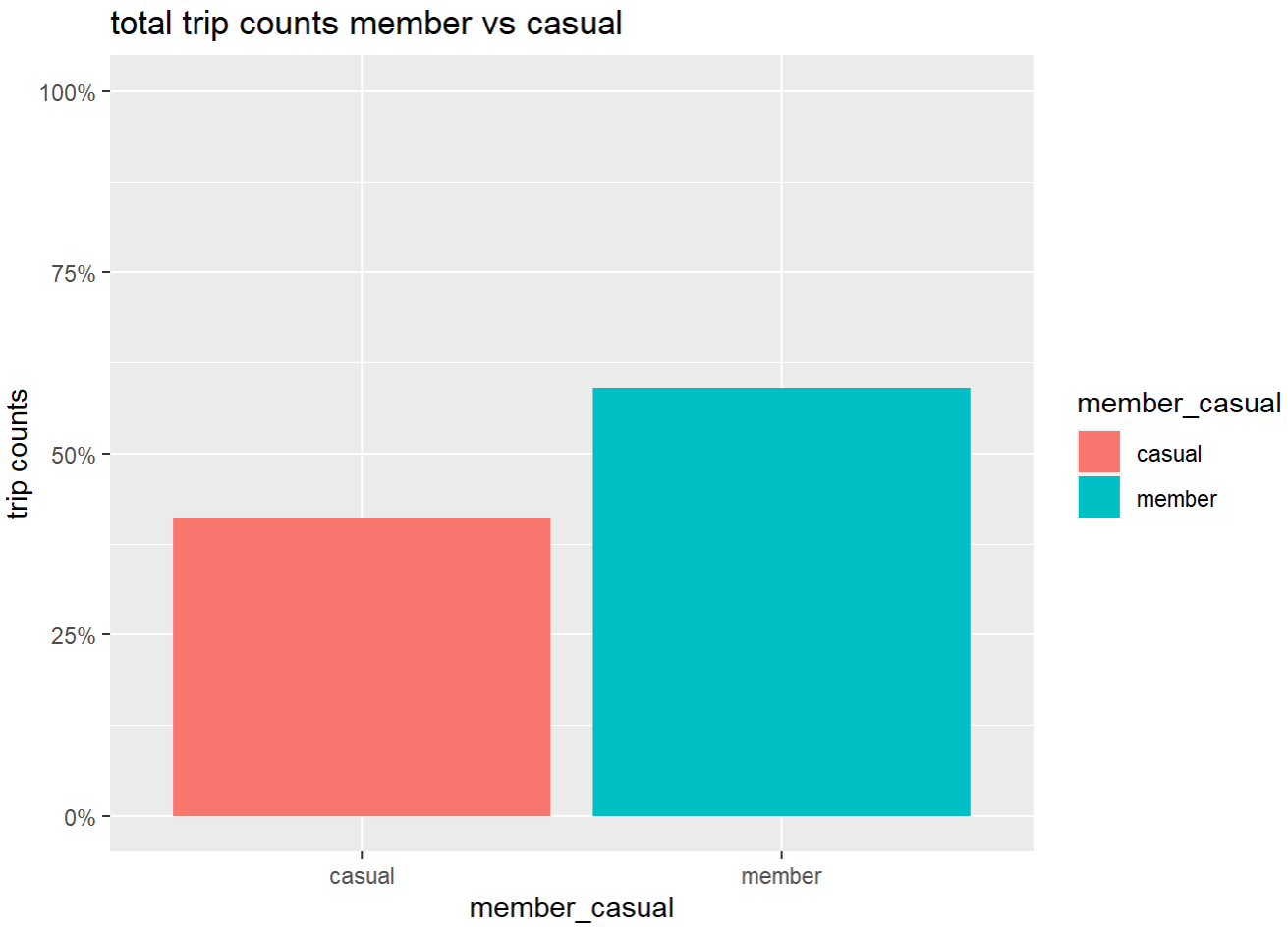
```
## [1] 3278631      21
```

exploring the relation betwwen casual and members for total number of trips

```
 df %>% count(member_casual) %>% mutate(percent=n/sum(n)*100) %>% arrange(desc(n))
```

```
##   member_casual       n  percent
## 1        member 1935509 59.03406
## 2        casual 1343122 40.96594
```

```
df %>% count(member_casual) %>% mutate(percent=n/sum(n)) %>% arrange(desc(n)) %>%  ggplot()+ge
om_col(aes(member_casual,percent,fill=member_casual))+
  labs(title="total trip counts member vs casual",y="trip counts")+
   scale_y_continuous(labels = percent,limits = c(0,1))
```



this shows that aunal members has more trip counts than casual ones

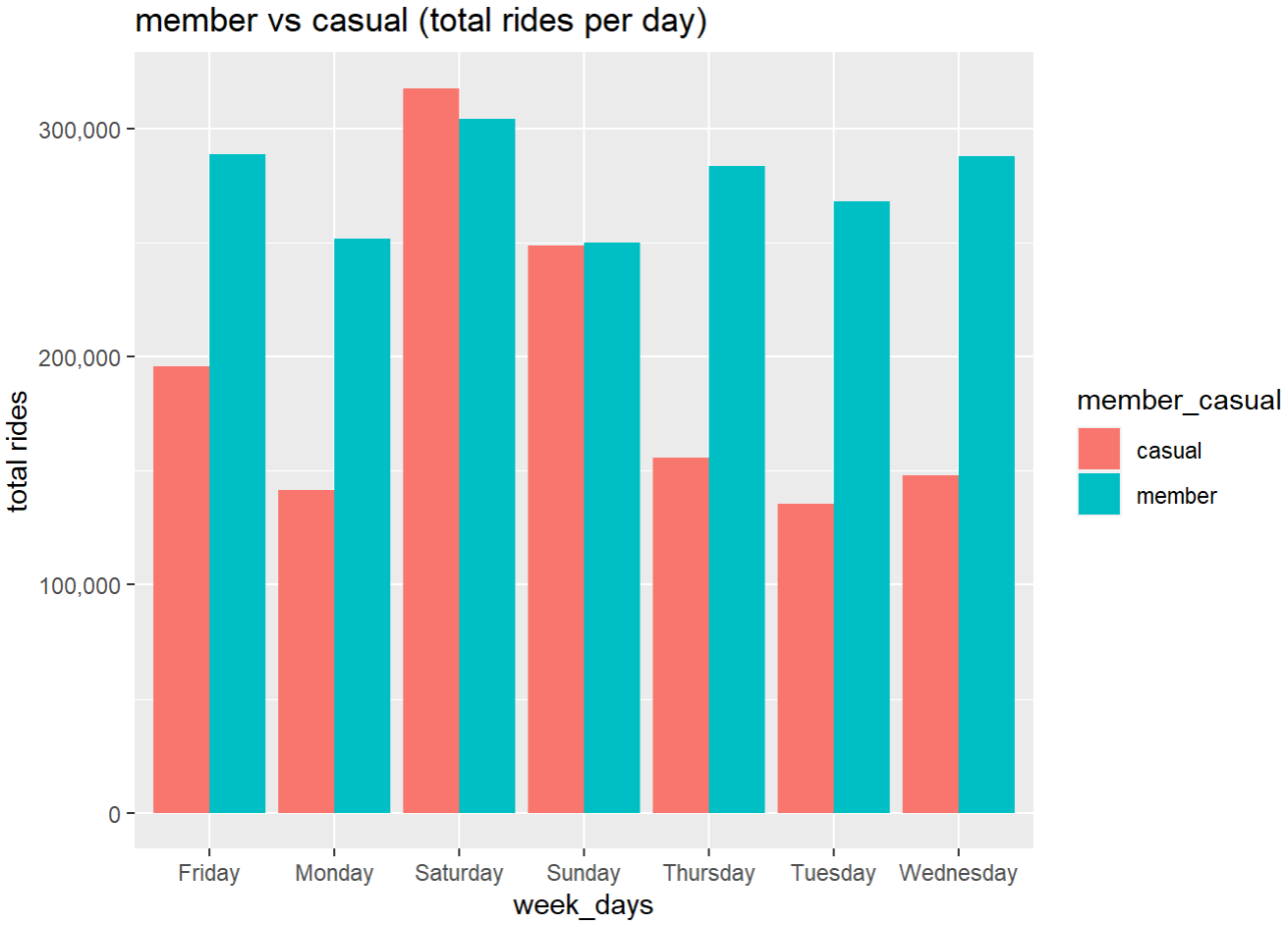
this graph shows the diffrent betwwen casual users and anual members for each day

```
df %>%group_by(member_casual,week_days) %>%  count(member_casual) %>% arrange(desc(n))
```

```
## # A tibble: 14 × 3
## # Groups:   member_casual, week_days [14]
##    member_casual week_days      n
##    <chr>         <chr>      <int>
##  1 casual        Saturday  317981
##  2 member        Saturday  304490
##  3 member        Friday    288755
##  4 member        Wednesday 288296
##  5 member        Thursday  283637
##  6 member        Tuesday   268123
##  7 member        Monday    251830
##  8 member        Sunday    250378
##  9 casual        Sunday    248850
## 10 casual        Friday    195683
## 11 casual        Thursday  155581
## 12 casual        Wednesday 147834
## 13 casual        Monday    141463
```

```
## 14 casual        Tuesday   135730
```

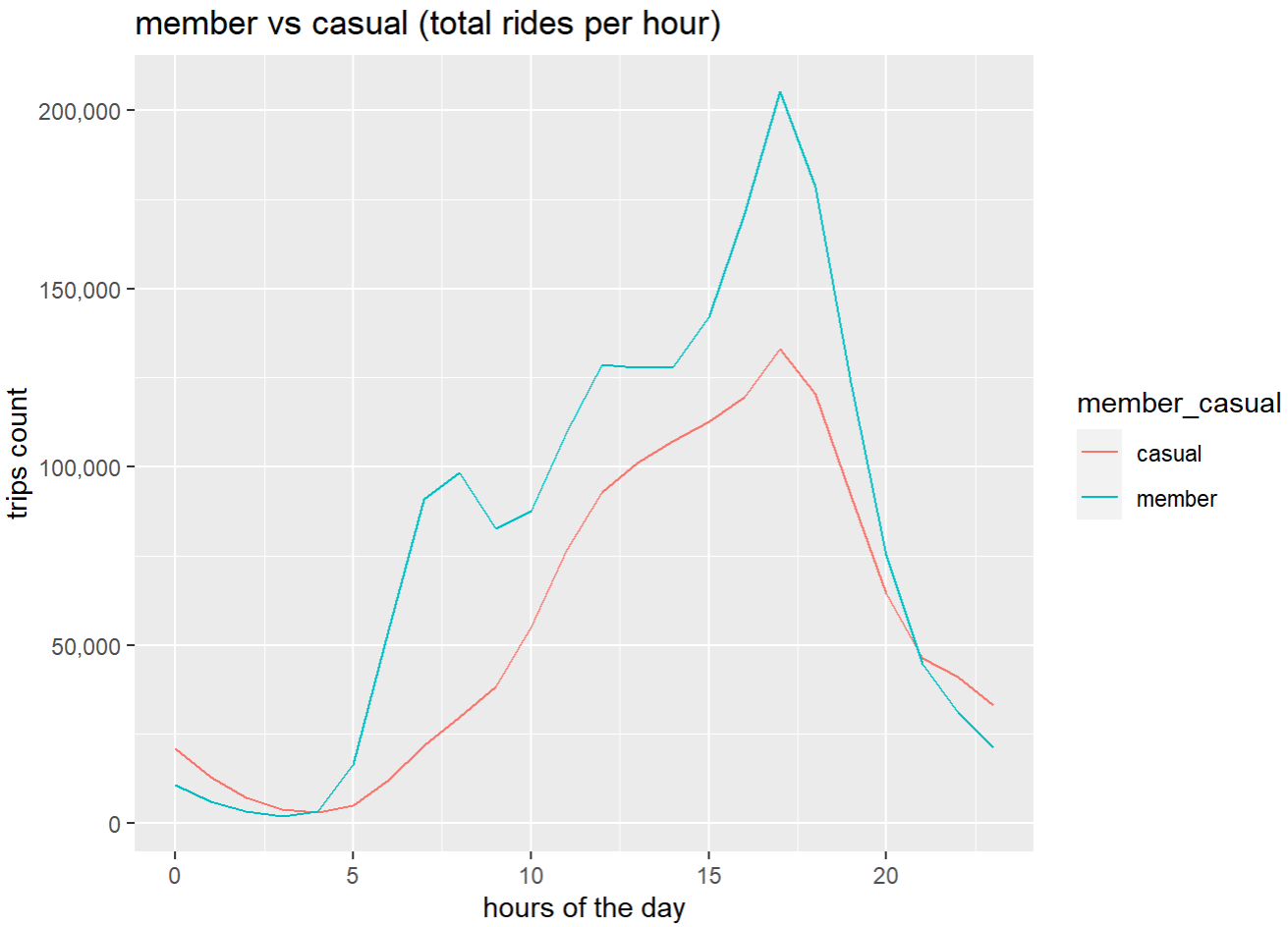
```
df %>%group_by(week_days) %>%  count(member_casual) %>% arrange(desc(n)) %>%
  ggplot()+geom_bar(aes(week_days,n,fill=member_casual),stat="identity", position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "member vs casual (total rides per day)",y="total rides")
```



this graph shows that (casual users) use cyclysits more on the weekend while anuual member use it more in the weekdays

it also shows that anual members has a close amout of trips each day while casual users has a big diffrent betwwen usage in weekend than weekday

```
bike_rides_for_year %>%group_by(member_casual) %>%  count(start_hour) %>%
  ggplot()+geom_line(aes(x=start_hour,y=n,colour=member_casual))+scale_y_continuous(labels = co
mma)+
  labs(title ="member vs casual (total rides per hour)" ,y="trips count",x = "hours of the day
" )
```



this graph shows that their is no diffrent in the hours which both members and casual use cyclistics they both start a 5 AM and pick at mid day till 8 pm

CREATING SUMMARY STATISCTICS DATA

```
bike_rides_stats <- df %>%
  group_by(member_casual,weekly = floor_date(start_date,unit = "week"),start_hour) %>%
  summarise(
    minutes = sum(trip_duration_miniuts),
    median = median(trip_duration_miniuts),
    mean = mean(trip_duration_miniuts),
    max = max(trip_duration_miniuts),
    min = min(trip_duration_miniuts),
    trip_counts = n()
  ) %>% ungroup()
```

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

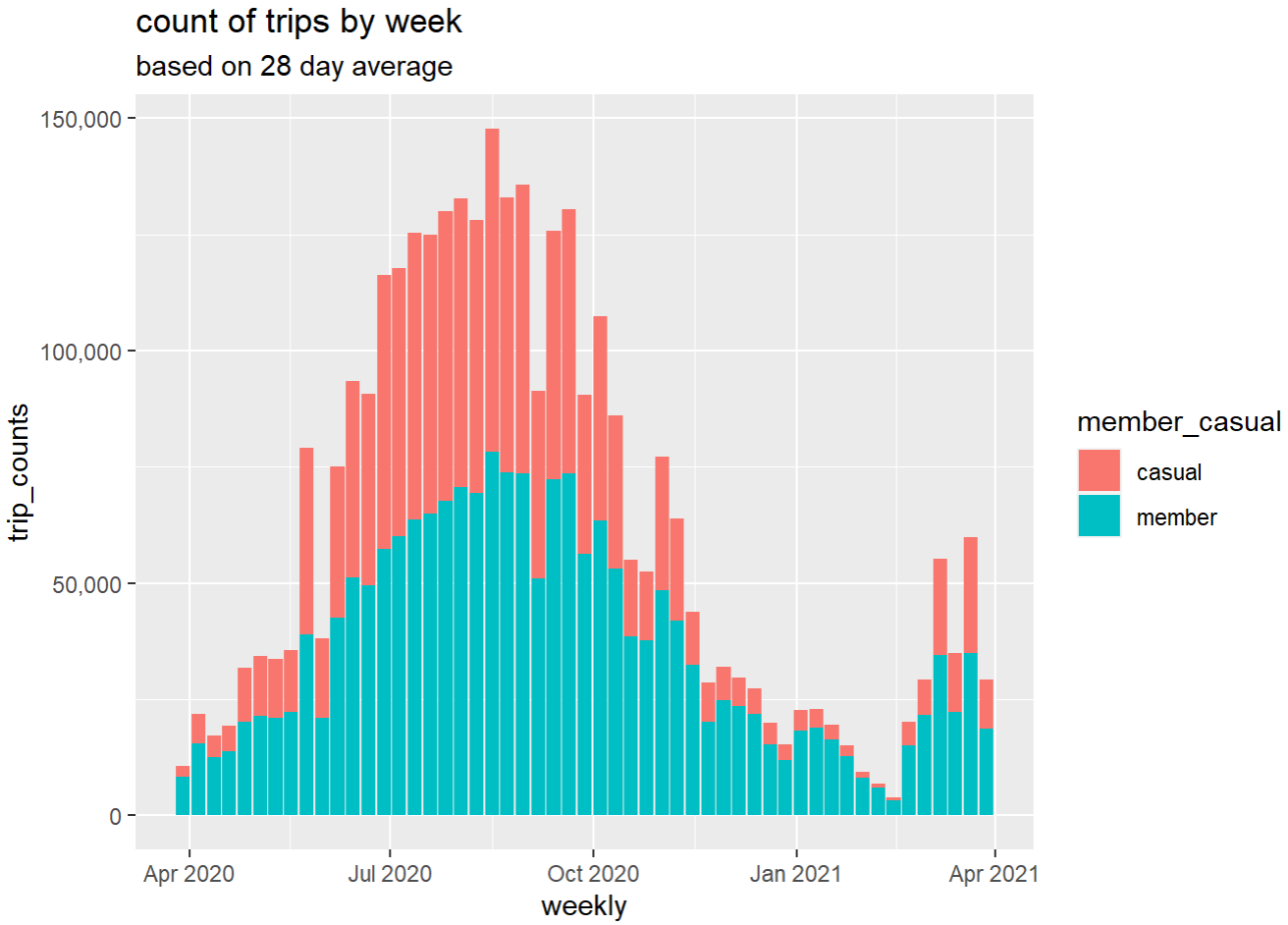
```
summary(bike_rides_stats$trip_counts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2     138     634    1296    1926    8468
```

```
xtabs(bike_rides_stats$trip_counts~bike_rides_stats$start_hour)
```

```
## bike_rides_stats$start_hour
##      0      1      2      3      4      5      6      7      8      9     10
##  31584  18934  10282   5743   6304  21346  66233 112317 127743 120707 142226
##     11     12     13     14     15     16     17     18     19     20     21
## 185084 220783 227991 234020 254047 289600 336804 297349 214654 138916  90410
##     22     23
##  71831  53723
```

```
bike_rides_stats %>% ggplot()+
  geom_col(aes(x=weekly,y=trip_counts,fill=member_casual))+
  scale_y_continuous(labels = comma)+labs(title = "count of trips by week",
                                          subtitle = "based on 28 day average")
```



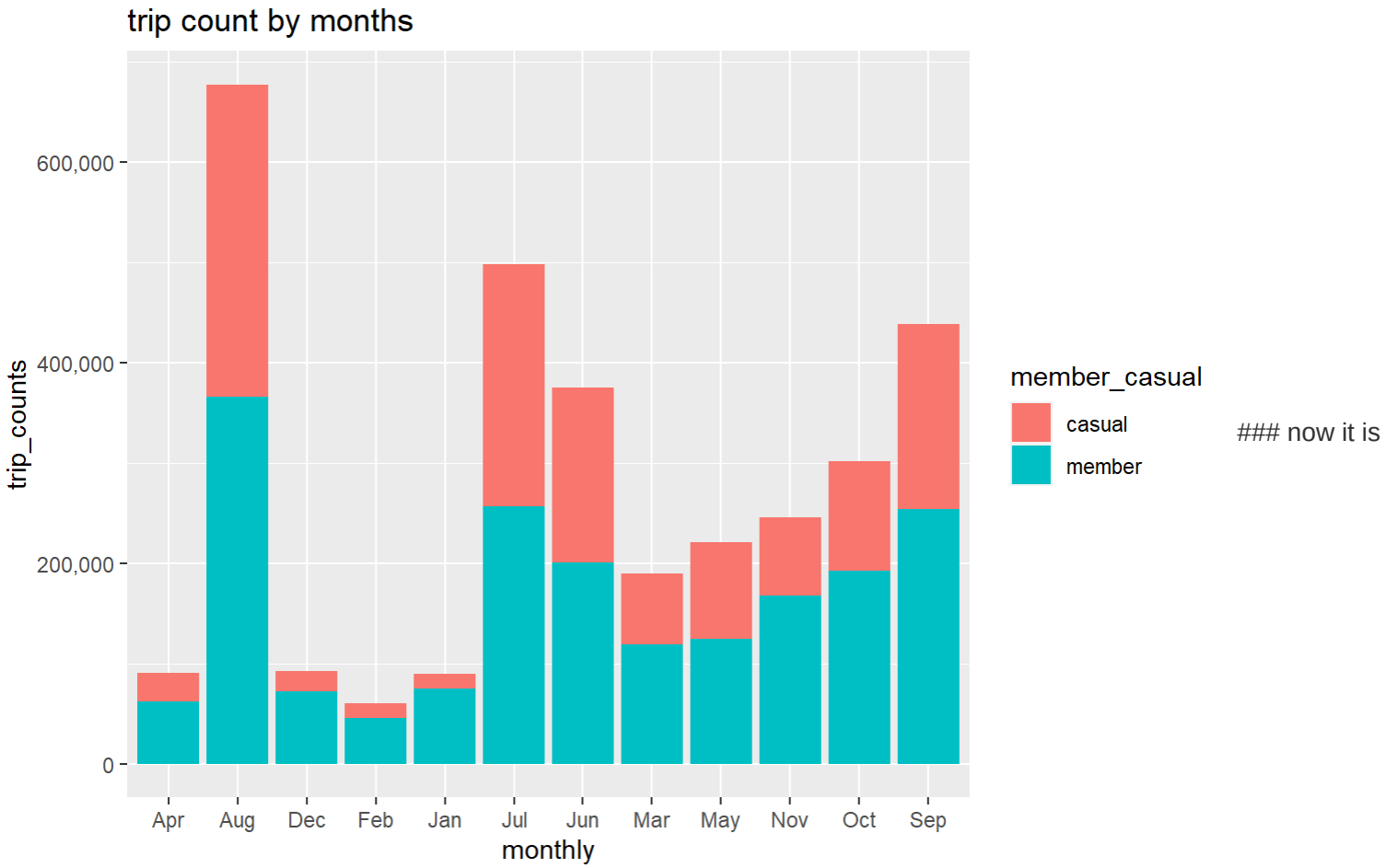
this graph shows that most of the year their is a big diffrenc betwwen anual and casual members but this is not the case from the mid of June to October which is mostly summer time lest investigate more by each month

adding rides by month

```
bike_rides_stats$monthly<-
  month.abb[lubridate::month(bike_rides_stats$weekly)]
```

trip count by months

```
bike_rides_stats %>% ggplot()+
     geom_col(aes(x=monthly,y=trip_counts,fill=member_casual))+
     scale_y_continuous(labels = comma)+
   labs(title = "trip count by months")
```



now it is more clear most of the rides done by the casual members are in summer time from June to September

CREATING SUMMARY STATISCTICS DATA (for bike types)

```
bike_rides_stats_types<- df %>%
     group_by(weekly = floor_date(start_date,unit = "week"),rideable_type,member_casual) %>%
     summarise(
       minutes = sum(trip_duration_miniuts),
       median = median(trip_duration_miniuts),
       mean = mean(trip_duration_miniuts),
       max = max(trip_duration_miniuts),
       min = min(trip_duration_miniuts),
```

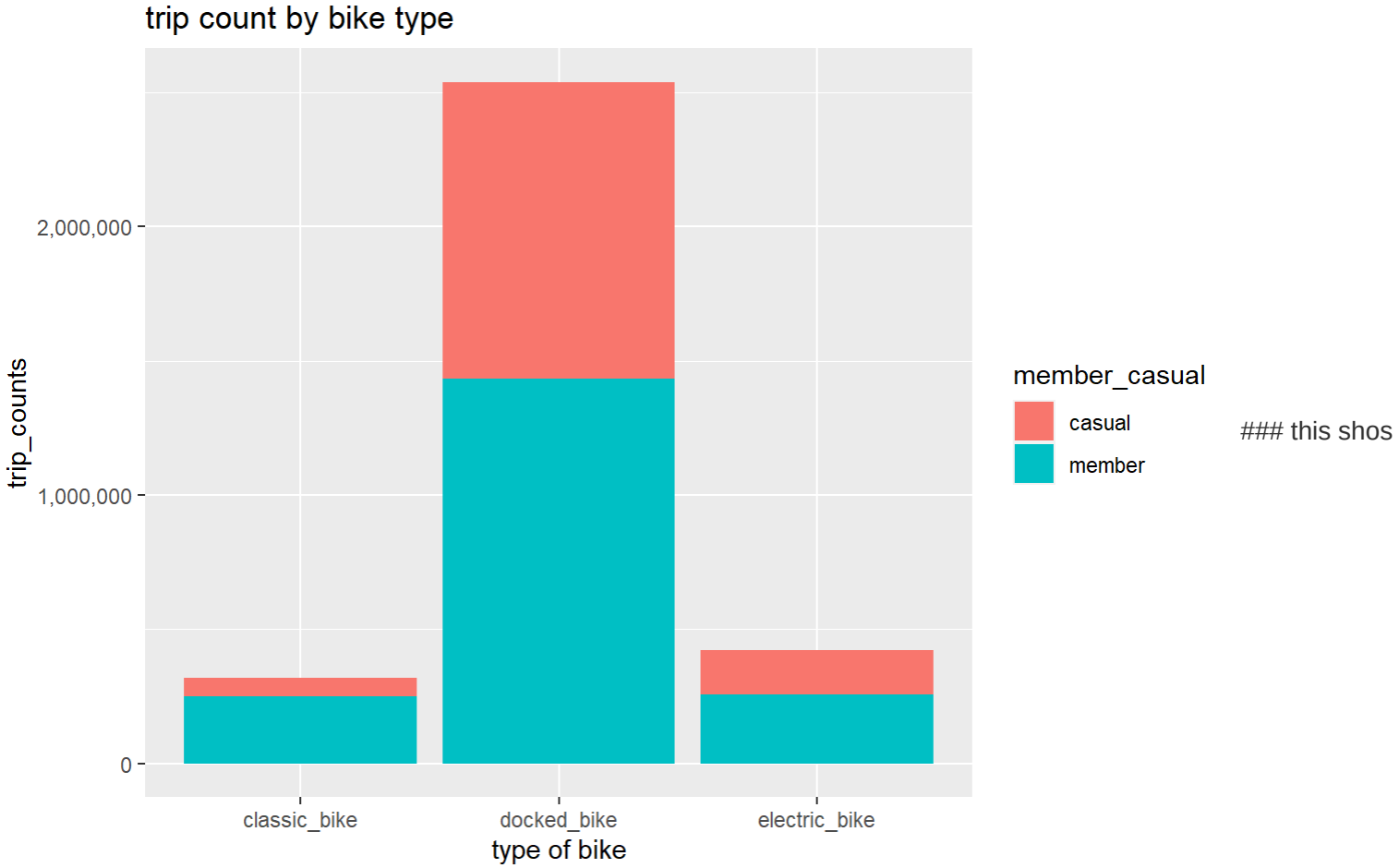
```
      trip_counts = n()
    ) %>% ungroup()
```

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

trip count vs bike type

```
trip_counts_by_bike_type<-xtabs(bike_rides_stats_types$trip_counts~bike_rides_stats_types$rideable_type)

  bike_rides_stats_types  %>% ggplot()+
    geom_col(aes(x=rideable_type,y=trip_counts,fill=member_casual))+
    scale_y_continuous(labels = comma,
    )+labs(title = "trip count by bike type",x="type of bike")
```

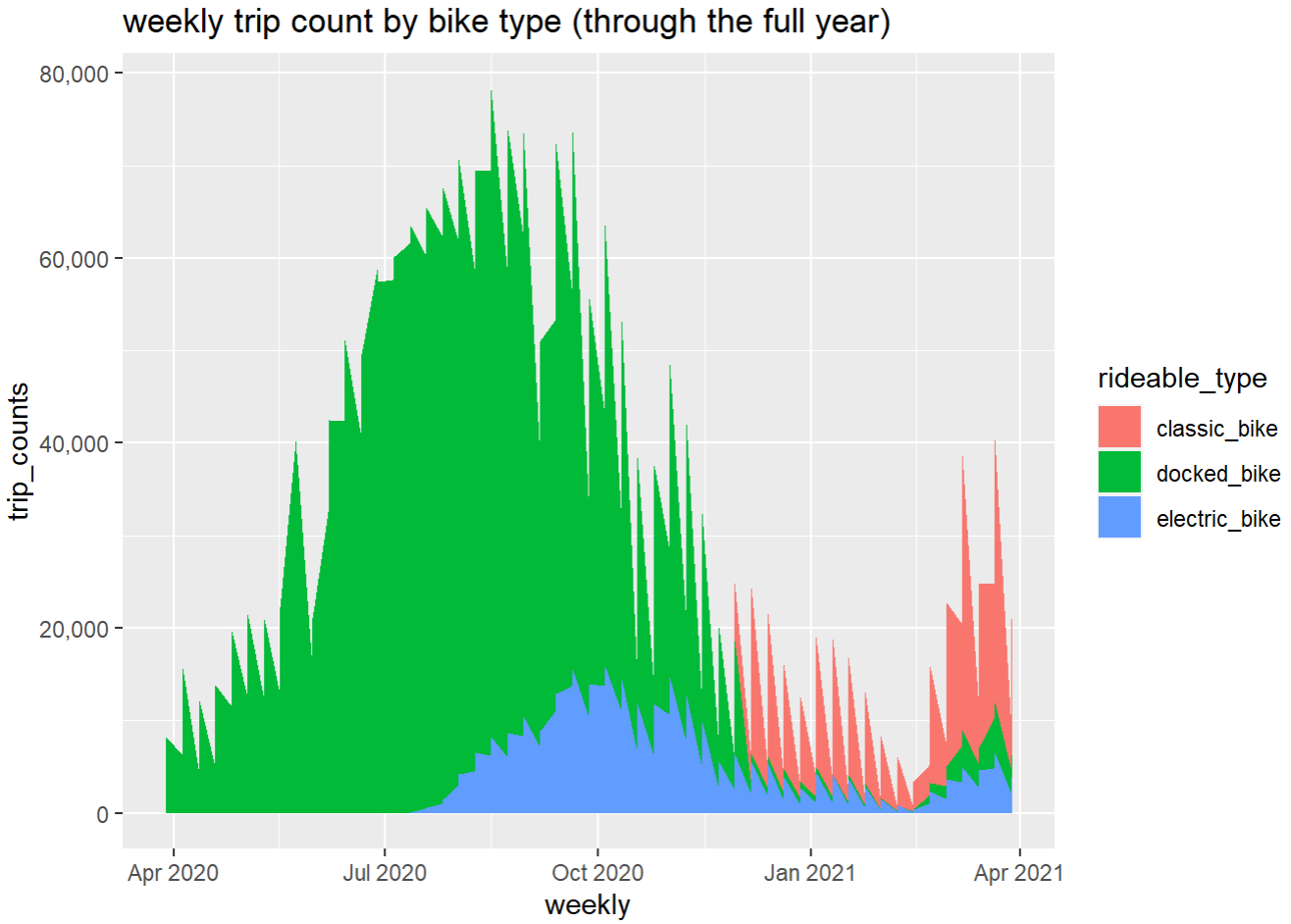


that most of casual rides is done by (docked bikes) this type of bike is prefered when the user of the bike dont have certain stops this tells up more about the way of ussage of the bikes by casual members

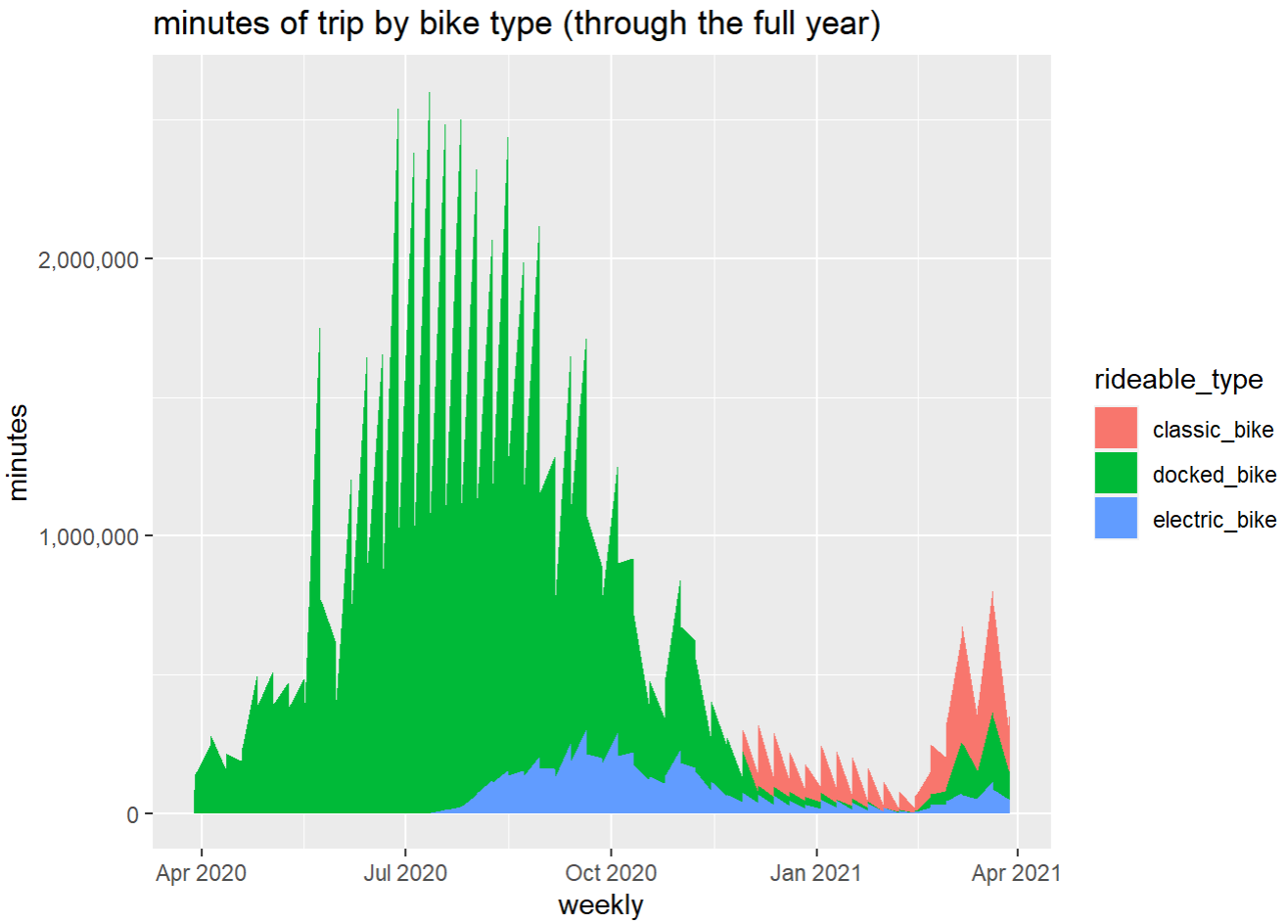
this also shows that member users also used docked bike more than classic and electric this high demand for docked bikes will need an increase in this type

```
bike_rides_stats_types %>% ggplot()+
```

```
  geom_area(aes(x=weekly,y=trip_counts,fill=rideable_type))+
  scale_y_continuous(labels = comma)+
  labs(title = "weekly trip count by bike type (through the full year)")
```




```
bike_rides_stats_types %>% ggplot()+
  geom_area(aes(x=weekly,y=minutes,fill=rideable_type))+
  scale_y_continuous(labels = comma)+
  labs(title = "minutes of trip by bike type (through the full year)")
```



this graph shoes that the (docked bikes) also dominates the need of the users not only by trip count number but also by its distribution through the year and the houres it is used for relative to electric and classic bikes

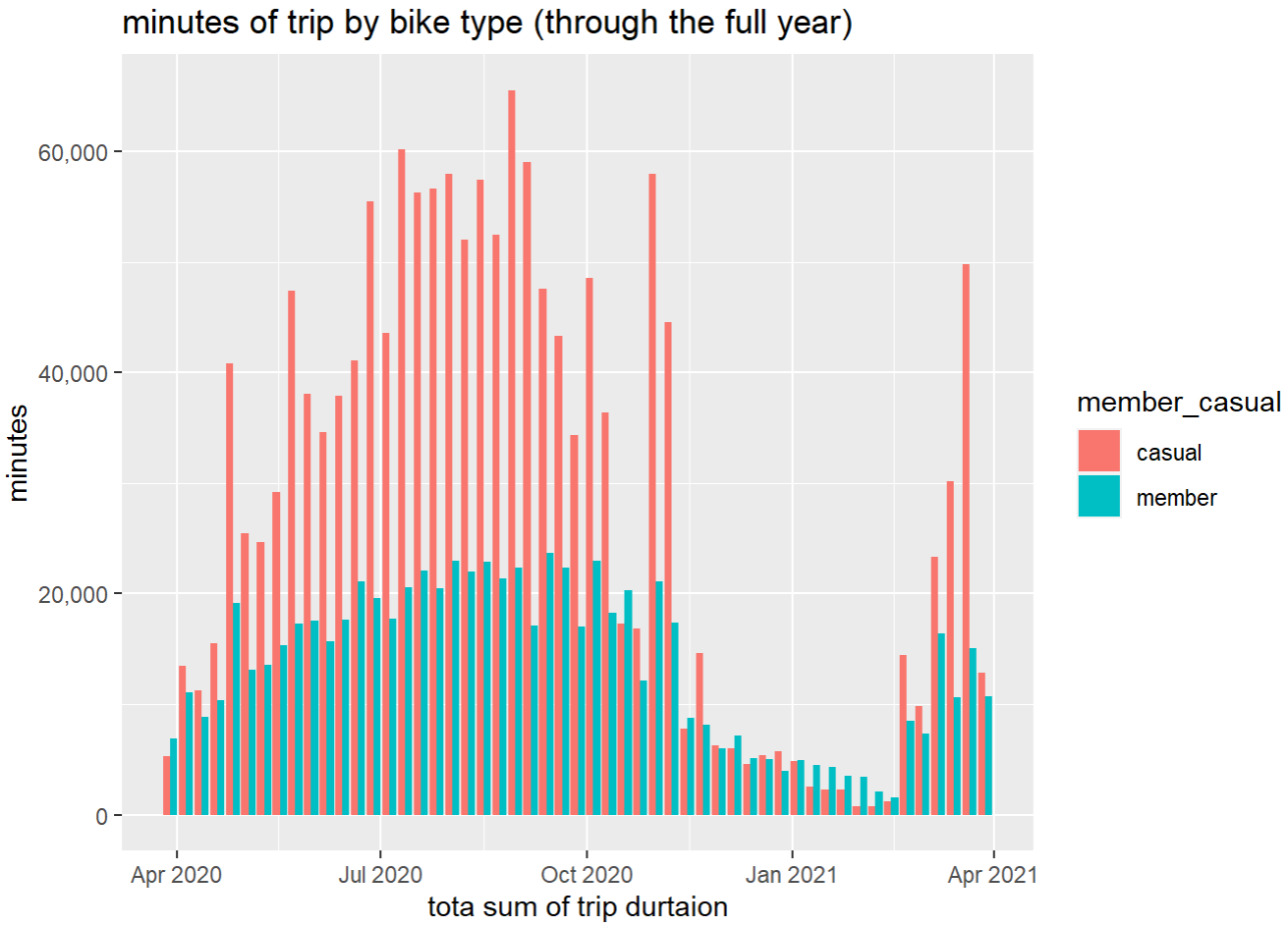
CREATING SUMMARY STATISCTICS DATA (for member vs casual)

```
bike_rides_stats_member_casual <- df %>%
  group_by(member_casual,start_hour,week_days,weekly = floor_date(start_date,unit = "week")
           ) %>%
  summarise(
    minutes = sum(trip_duration_miniuts),
    median = median(trip_duration_miniuts),
    mean = mean(trip_duration_miniuts),
    max = max(trip_duration_miniuts),
    min = min(trip_duration_miniuts),
    trip_counts = n()
  ) %>% ungroup()
```

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

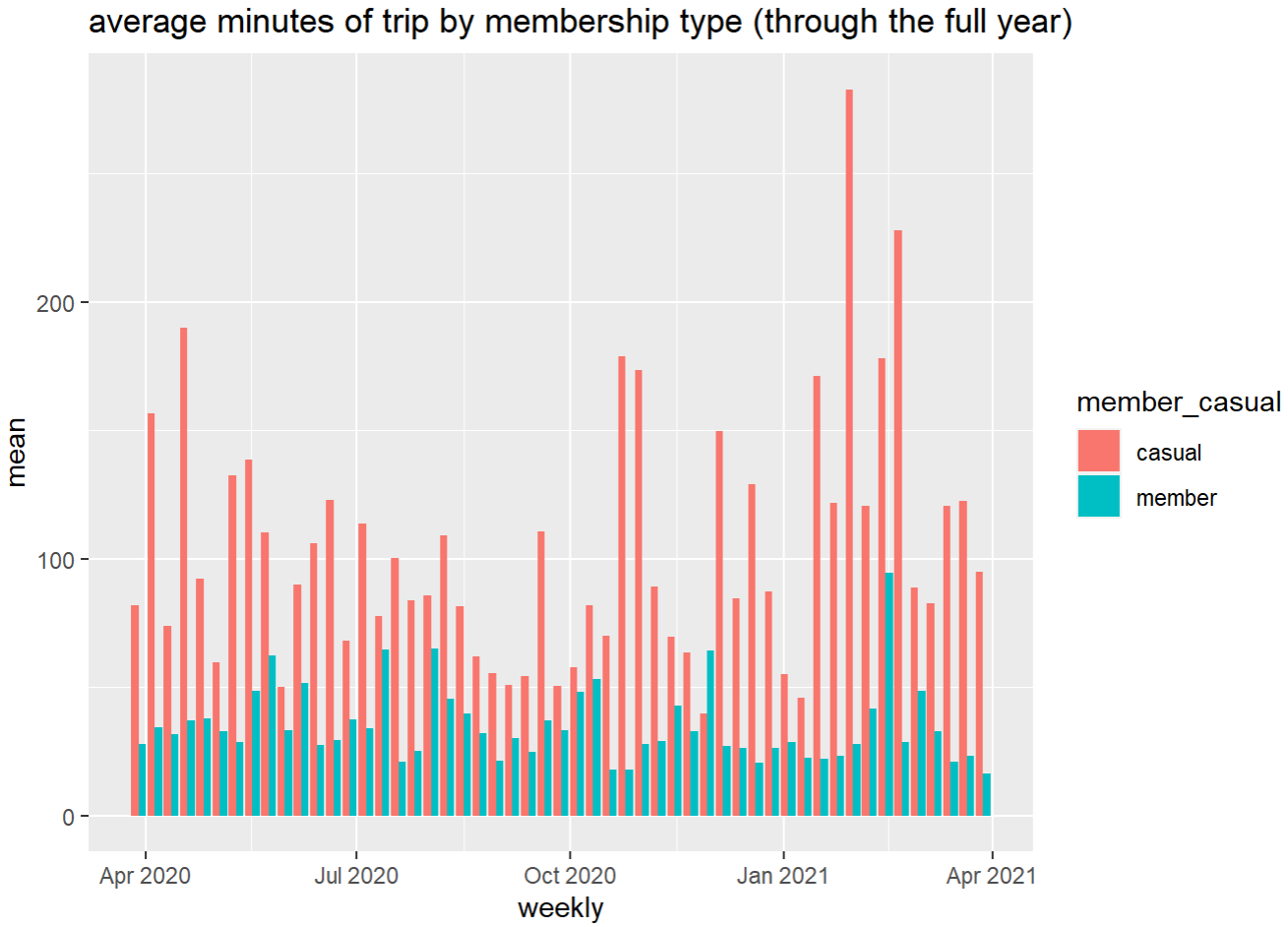
comparing member vs casual by ride time (total minuits) and (through the full year)

```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=weekly,y=minutes,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "minutes of trip by bike type (through the full year)",x="tota sum of trip durt
aion")
```



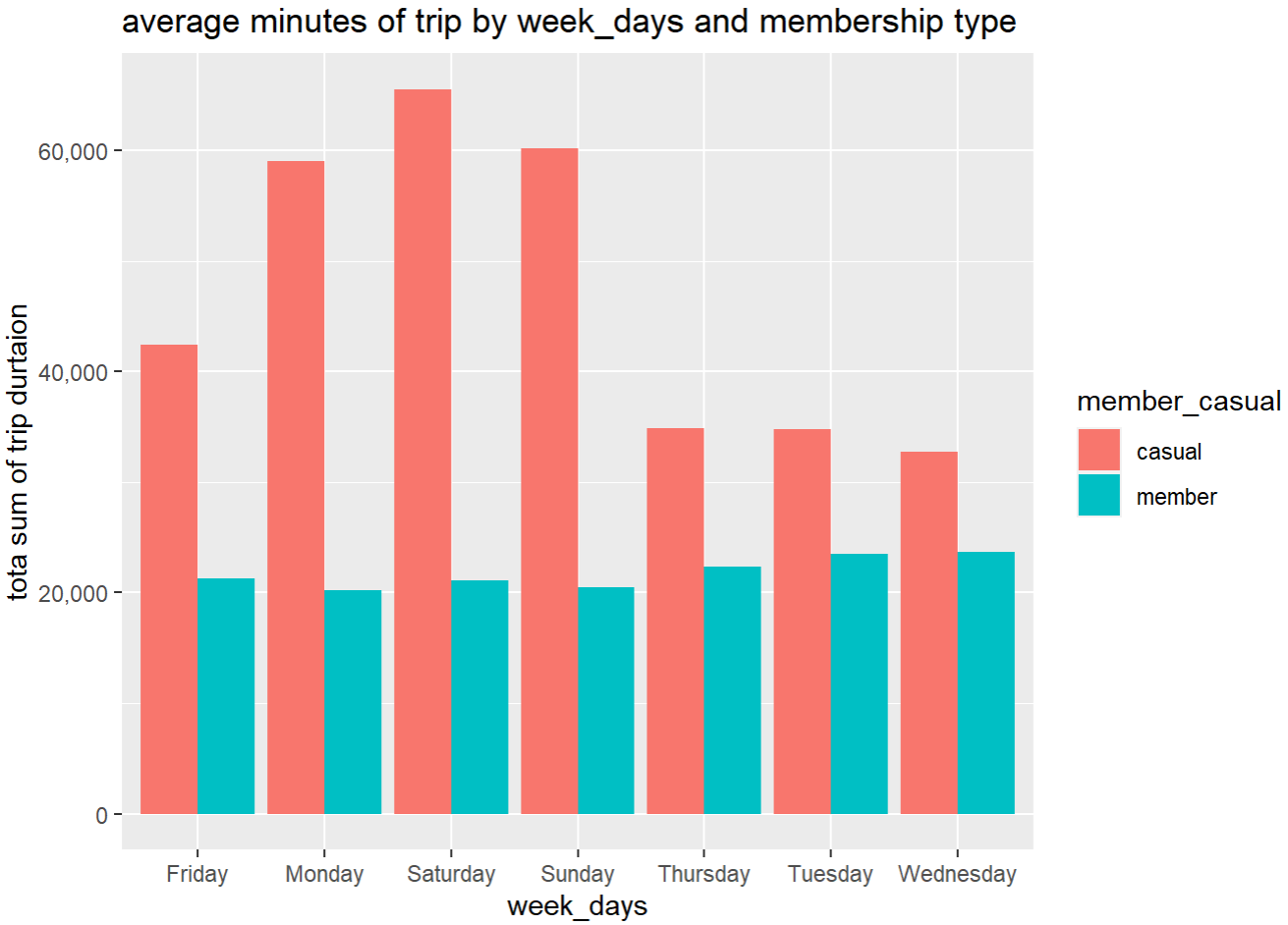
comparing member vs casual by ride time (average minutes) and (through the full year)

```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=weekly,y=mean,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "average minutes of trip by membership type (through the full year)")
```



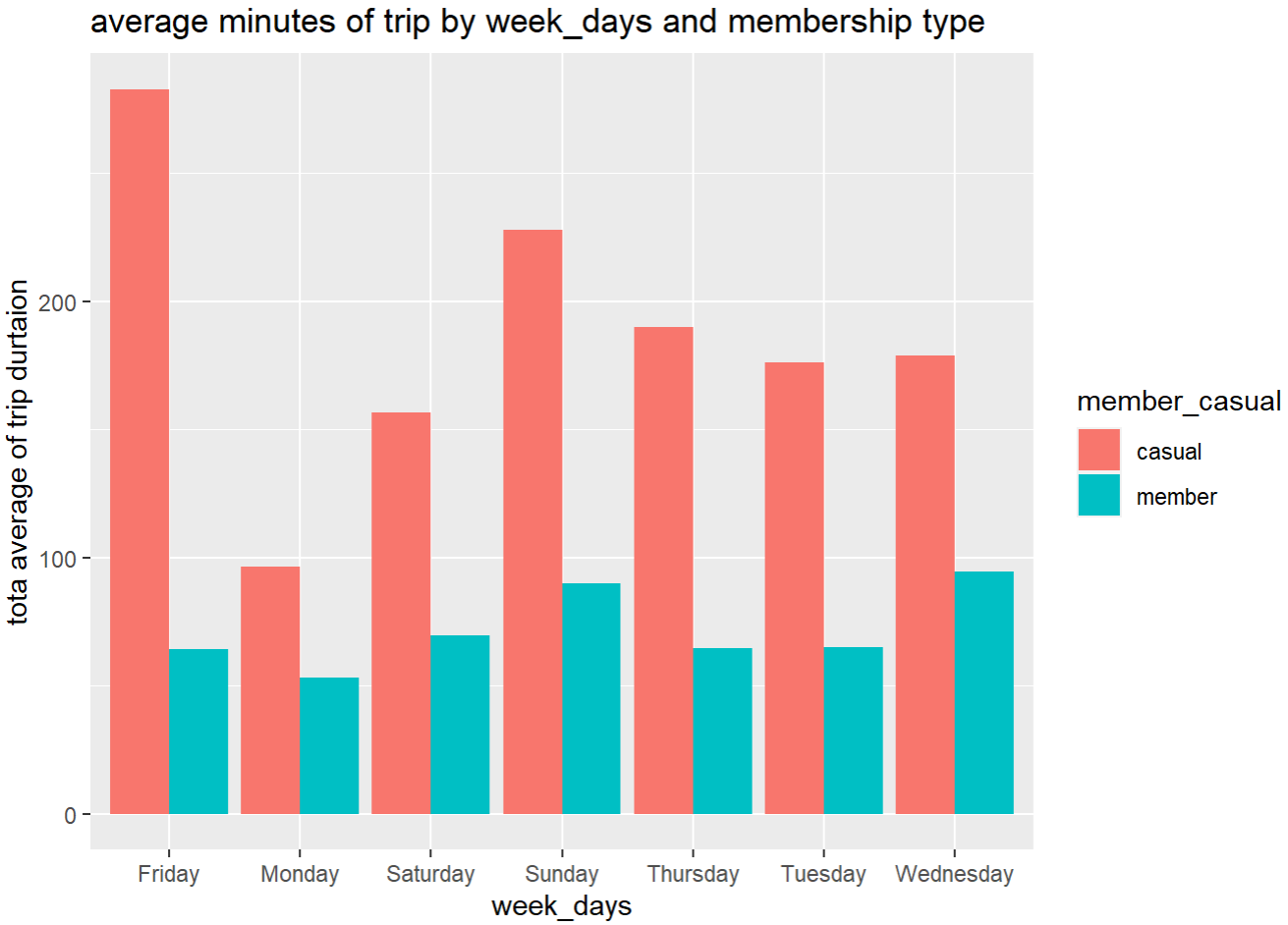
comparing member vs casual by ride time (total minuits) and (by weekd days)

```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=week_days,y=minutes,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "average minutes of trip by week_days and membership type",y="tota sum of trip durtaion")
```



comparing member vs casual by ride time (average minuits) and (by weekd days)

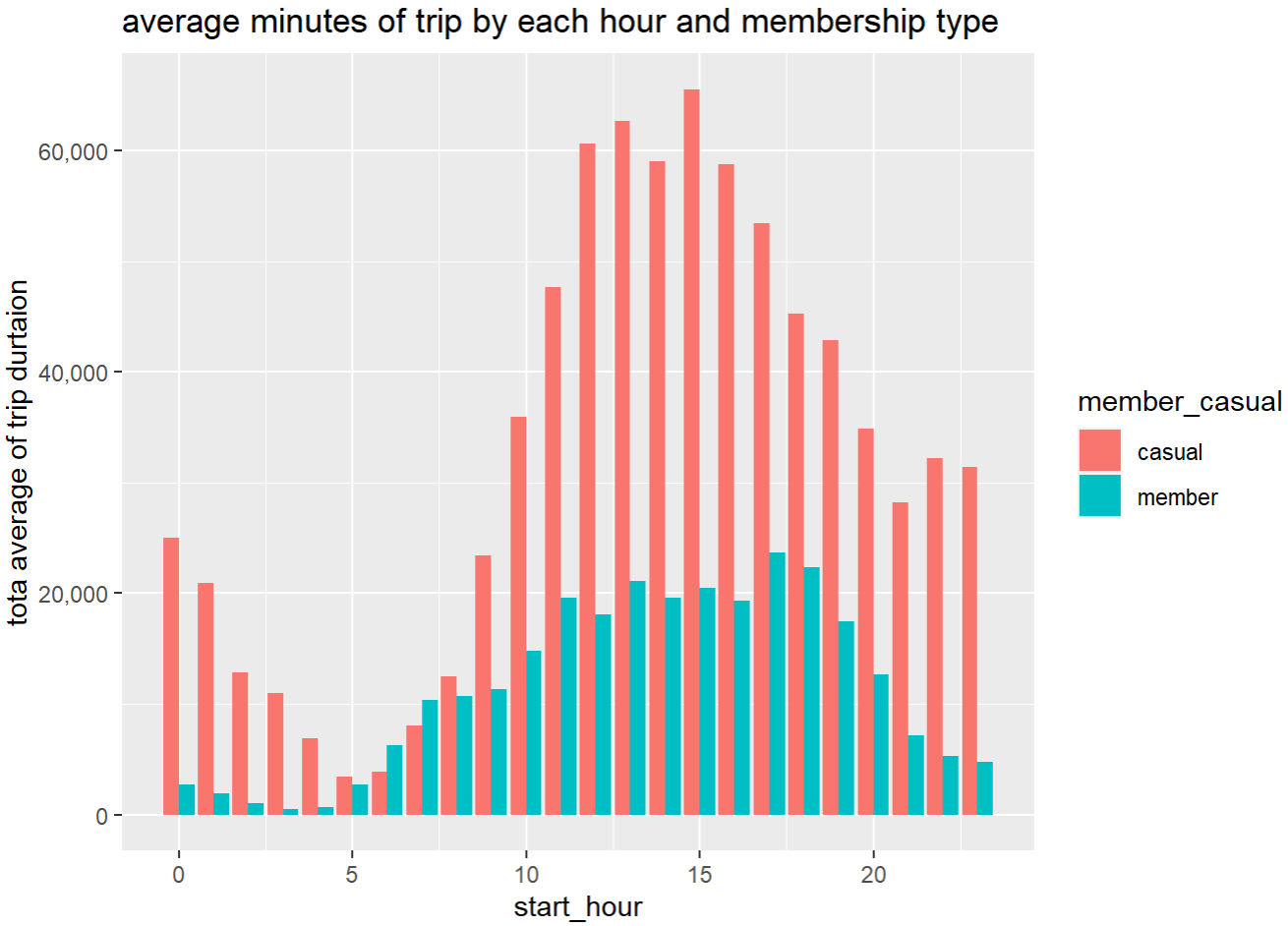
```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=week_days,y=mean,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "average minutes of trip by week_days and membership type",y="tota average of tr
ip durtaion")
```



comparing member vs casual by ride time (total minuits) and (by each hour)

```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=start_hour,y=minutes,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "average minutes of trip by each hour and membership type",y="tota average of tr
ip durtaion")
```

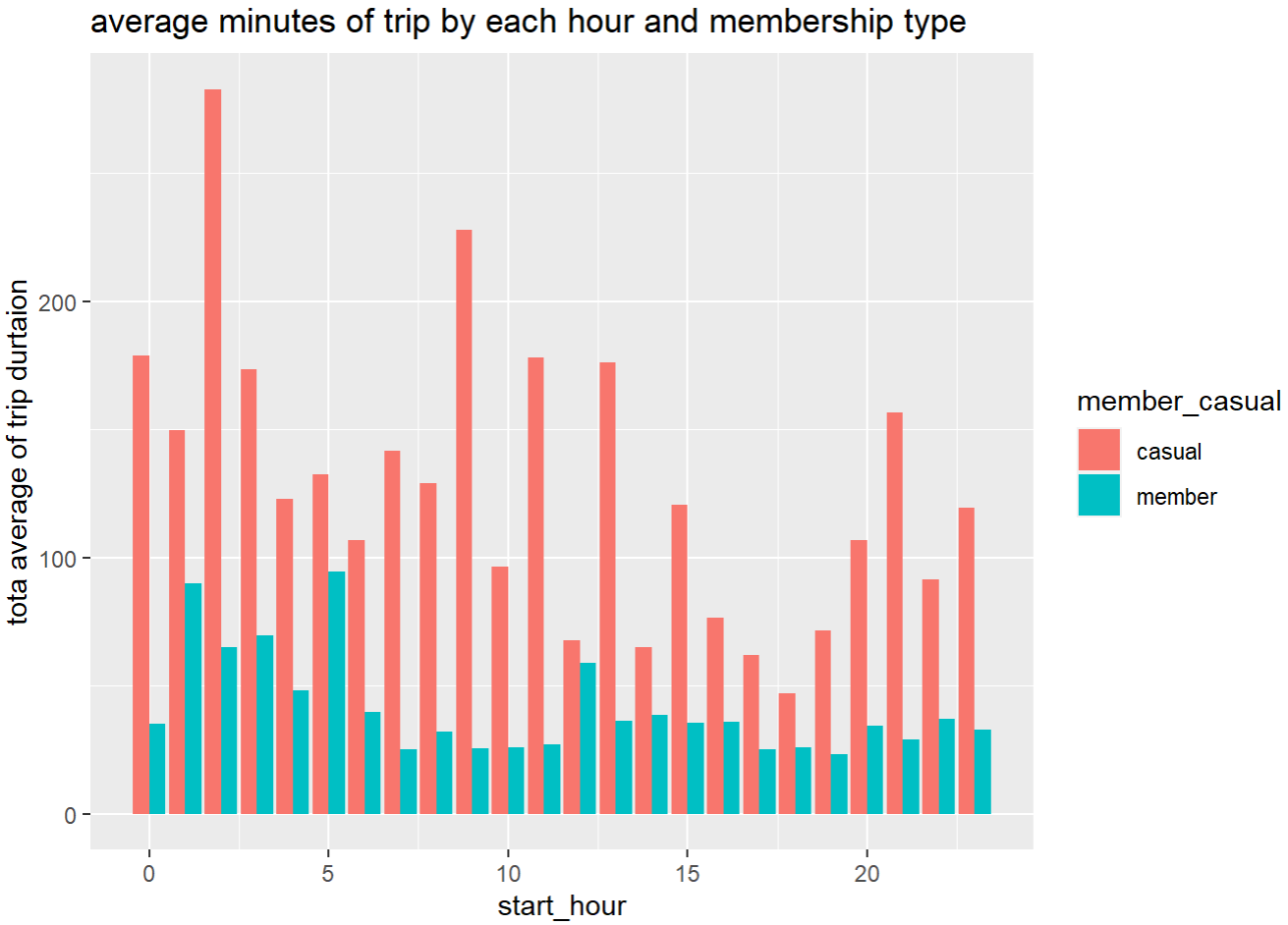
###



comparing member vs casual by ride time (average minuits) and (by each hour)

```
bike_rides_stats_member_casual %>% ggplot()+
  geom_col(aes(x=start_hour,y=mean,fill=member_casual),position = "dodge")+
  scale_y_continuous(labels = comma)+
  labs(title = "average minutes of trip by each hour and membership type",y="tota average of tr
ip durtaion")
```

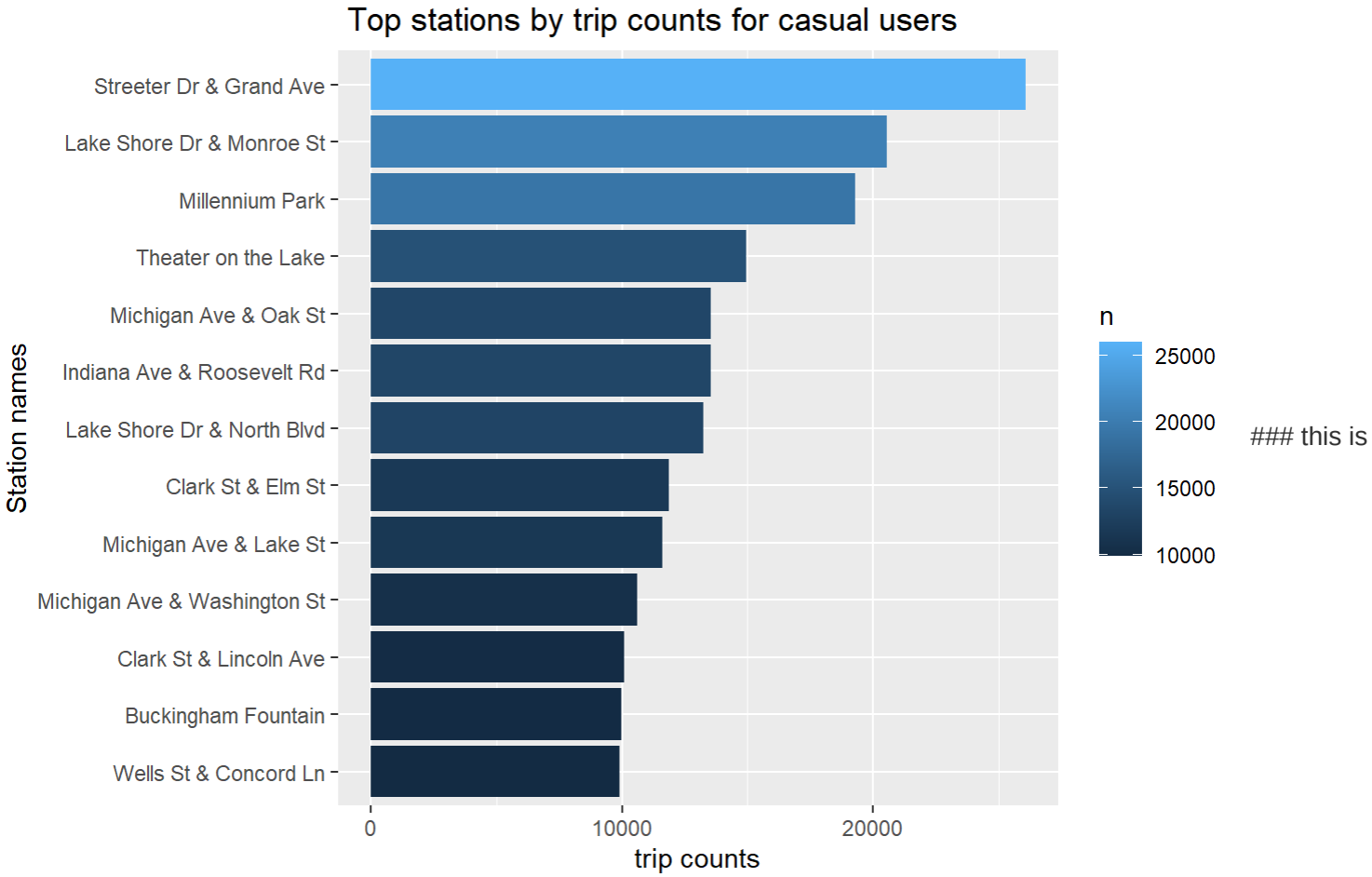
the



previus two graph shows that throught the full year,each weekday and weekends and through each hour of the day the casual users use cyclystics for alot more minuits (in total and in average) than anual users this shows a great diffrent in the type of usage between the two members it shows that casual users need a diffrent approch in their anual package membership

CREATING SUMMARY STATISCTICS DATA (for geographic location)

```
bike_rides_for_year %>% filter(member_casual=="casual")  %>% group_by(member_casual) %>%
  count(start_station_name,sort=T) %>%
  filter(n>9900)%>% ggplot()+
  geom_col(aes(y=reorder(start_station_name,n)
               ,x=n
               ,fill=n))+
  labs(title = " Top stations by trip counts for casual users",
       x="trip counts",y="Station names")
```



the start station with the most trips of casual users that marketing campain can focus more on this areas as the habit of riding a bike seems to be high in this areas and the company cyclystic has alot of users alredy

Summary of the analysis

after investigating the data we summaries that casual users doesn't use Cyclistic as a transportation way to work or on a daily basis the casual uses are more active in summer time (June,July,august,September) also they are active at weekend days (sunday and stauday) it was realized also that casuel users spend alot more minuits when riding bikes than anual ones which shows that casual users dont rent the bike for a certain task also casual users mostly rent docked bikes

top three recommendations based on my analysis

1.A new annual subscription should by hours instead of by days

2.The marketing campaign should be done before the summer to convert more casual users to subscribe

3.The marketing campaign should be focused on the top 15 areas that has active casual users

4.reinforce commuting riders by Advertisement promoting riding a bike to work (exercise and environment friendly)

cyclist bike company