# Google Data Analytics: Cyclistic Bike Share Project

Haika Ngowi

10/19/2021

## Data Wrangling Changelog

*Data Range: 2020-04-01 to 2021-31-03

*Total rows 3,489,749

*Observations on the data: We have 10552 rows with negative trip durations

*122,175 rows with missing starting station names and ID

*1,291 stations with two different start station ID.

*We can ignore station data with trip duration <=0,rows with missing start station name, missing station ID. Focus on using start station name to perform aggregate functions on date, start_station_name, member casual and rideable type.

*Cleaning reduced rows to 3,343,689

```
dir("Data",full.names = T)
```

```
## character(0)
```

```
# Upload Data for analysis

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

```
##
## Combine 12 dataframes into 1 dataframe.
##
bike_rides <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
bike_rides <- janitor::remove_empty(bike_rides, which = c("cols"))
bike_rides <- janitor::remove_empty(bike_rides, which = c("rows"))
```

```
##
## Convert Data/Time stamp to Date/Time
## Convert start and ended at date format to hms format
```

```
bike_rides$Ymd <- as.Date(bike_rides$started_at)
bike_rides$started_at <-
  lubridate::ymd_hms(bike_rides$started_at)
bike_rides$ended_at <- lubridate::ymd_hms(bike_rides$ended_at)

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

```
# calculate time difference in hours and mins
bike_rides$Hours <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("hours"))

bike_rides$Minutes <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("mins"))

df <- bike_rides %>% filter(Hours >0) %>% drop_na() %>%
  select(-ride_id,-end_station_name,-end_station_id,
         -end_station_name)
```

```
#Calculate trip duration
bike_rides$Hours <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("hours"))
bike_rides$Minutes <- difftime(bike_rides$ended_at,bike_rides$started_at,units = c("mins"))
bikerides1 <- bike_rides %>% filter(Minutes >0) %>% drop_na()
```

```
#Create summary dataframe
bikerides2 <- bike_rides %>% group_by(Weekly =
                                        floor_date(Ymd,"week"),start_hour) %>%
  summarise(
    Minutes = sum(Minutes),
    Mean = mean(Minutes),
    Median = median(Minutes),
    Max = max(Minutes),
    Min = min(Minutes),
    Count =n()
  ) %>% ungroup()
```

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

#Plots of Rides by Date ####Summary Stats: Counts *Summary of Hourly Counts

```
summary(bikerides2$Count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0   426.5  1568.0  2754.3  3903.5 15763.0
```

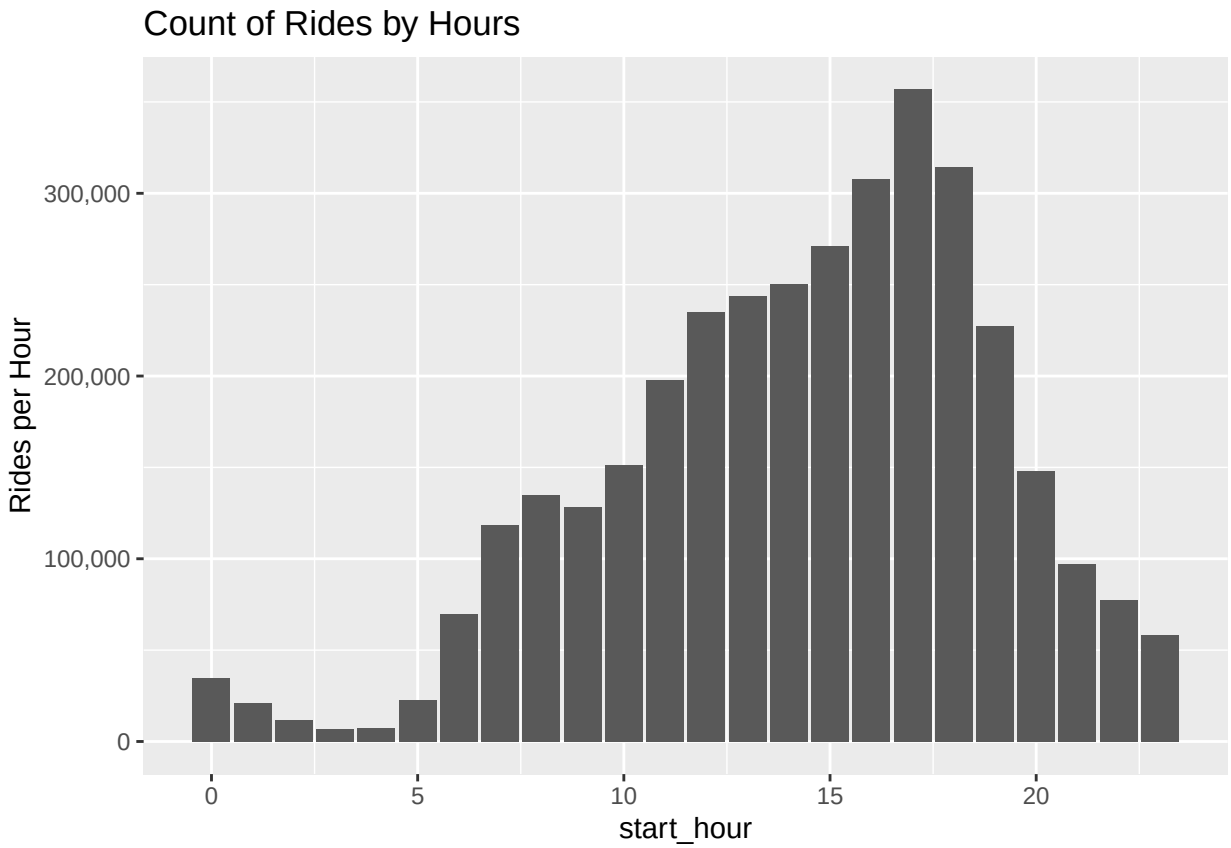
*Count of Rides Per Hour

```
xtabs(bikerides2$Count~bikerides2$start_hour)
```

```
## bikerides2$start_hour
##      0      1      2      3      4      5      6      7      8      9     10
##  34623  20863  11372   6465   7032  22752  69390 118074 134962 128198 151391
##     11     12     13     14     15     16     17     18     19     20     21
## 197430 235215 243406 250150 271140 307515 356796 314352 227396 148181  97034
##     22     23
##  77685  58326
```

```
bikerides2 %>% ggplot() + geom_col(aes(x=start_hour,y=Count)) +
  scale_y_continuous(labels = comma) +
  labs(title = "Count of Rides by Hours",
```

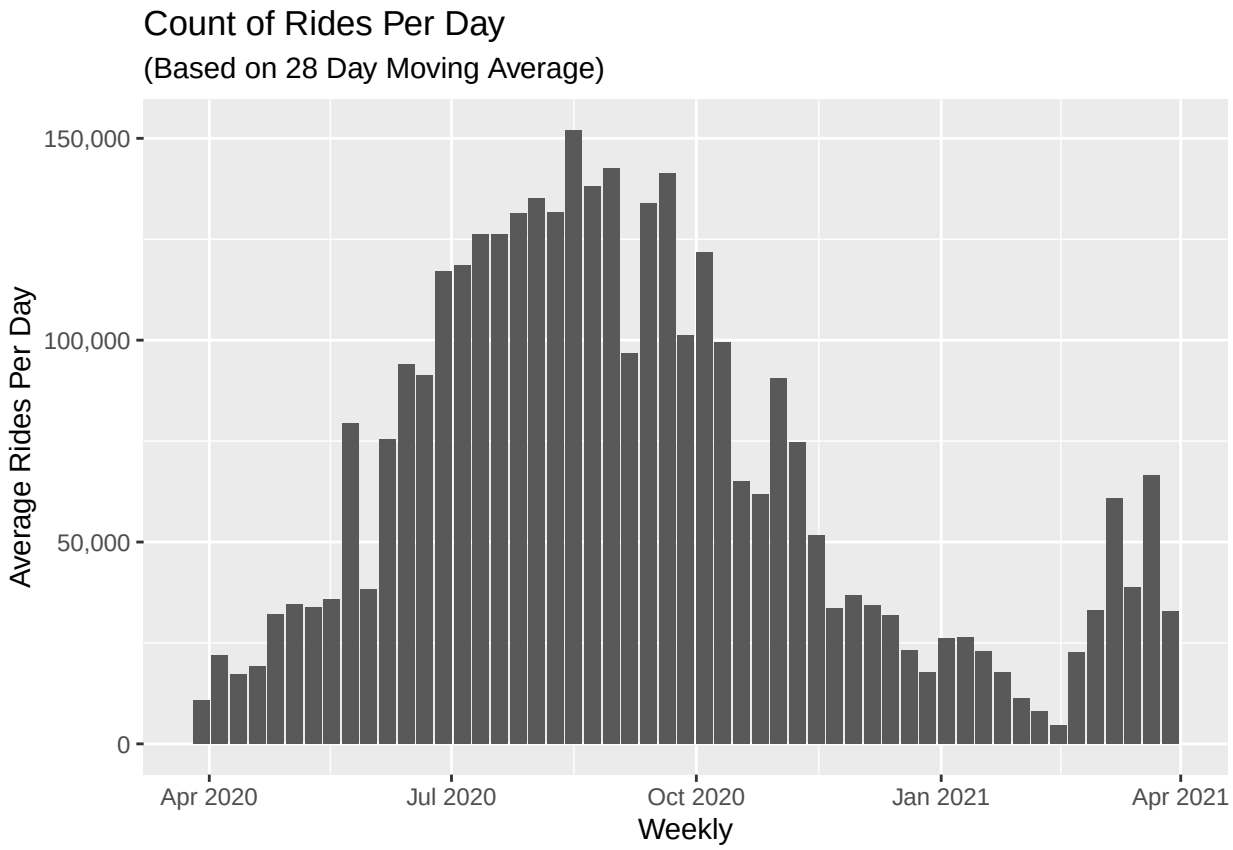
```
      y="Rides per Hour")
```




*Count of Rides Per Day

```
bikerides2$Monthly <- lubridate::month(bikerides2$Weekly)

bikerides2 %>% ggplot() + geom_col(aes(x=Weekly,y=Count)) +
  scale_y_continuous(labels = comma) +
  labs(title = "Count of Rides Per Day",
       subtitle = "(Based on 28 Day Moving Average)",
       y="Average Rides Per Day")
```




## Count of Rides by Bike Type

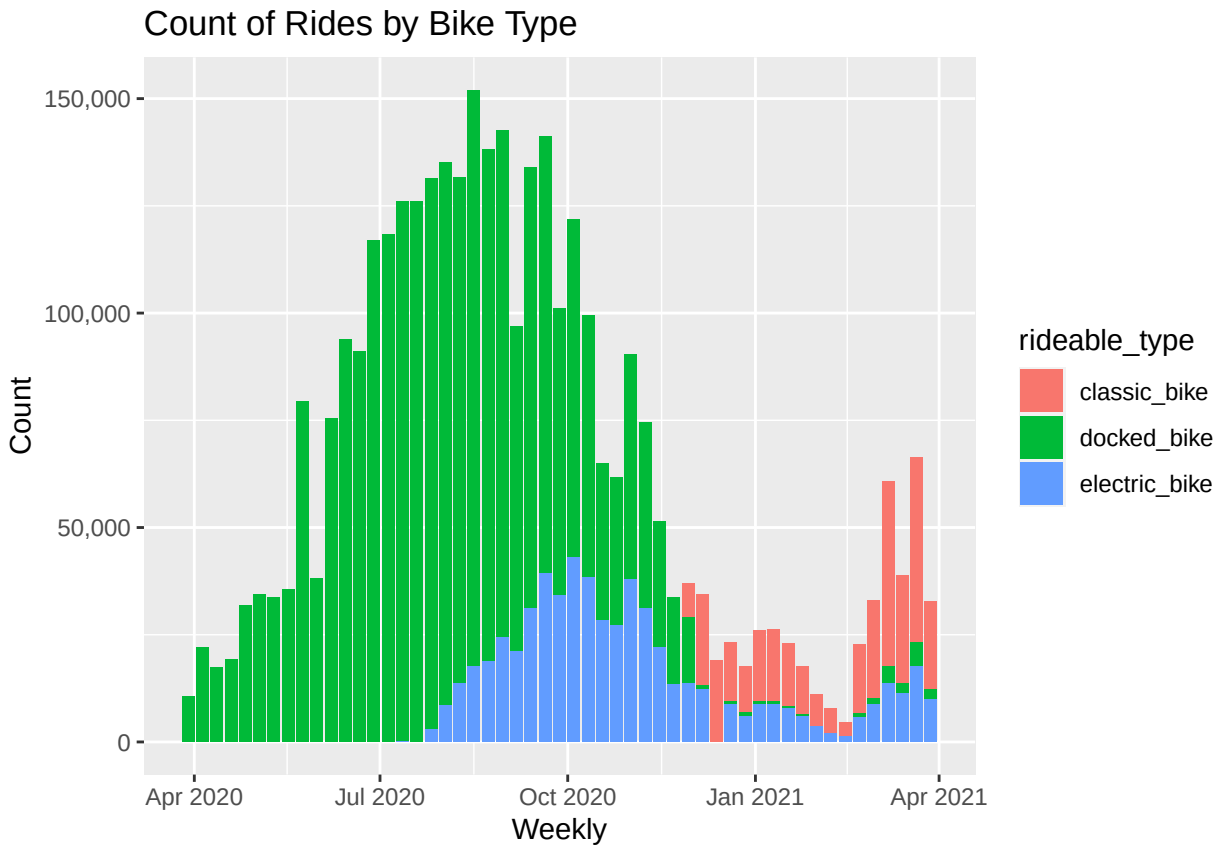
```
biketype <- bike_rides %>% group_by(member_casual,rideable_type,Weekly =
floor_date(Ymd,"week")) %>%
  summarize(
    Minutes = sum(Minutes),
    Mean = mean(Minutes),
    Median = median(Minutes),
    Max = max(Minutes),
    Min = min(Minutes),
    Count =n()
  ) %>% ungroup()
```

### Summary of Bike Types

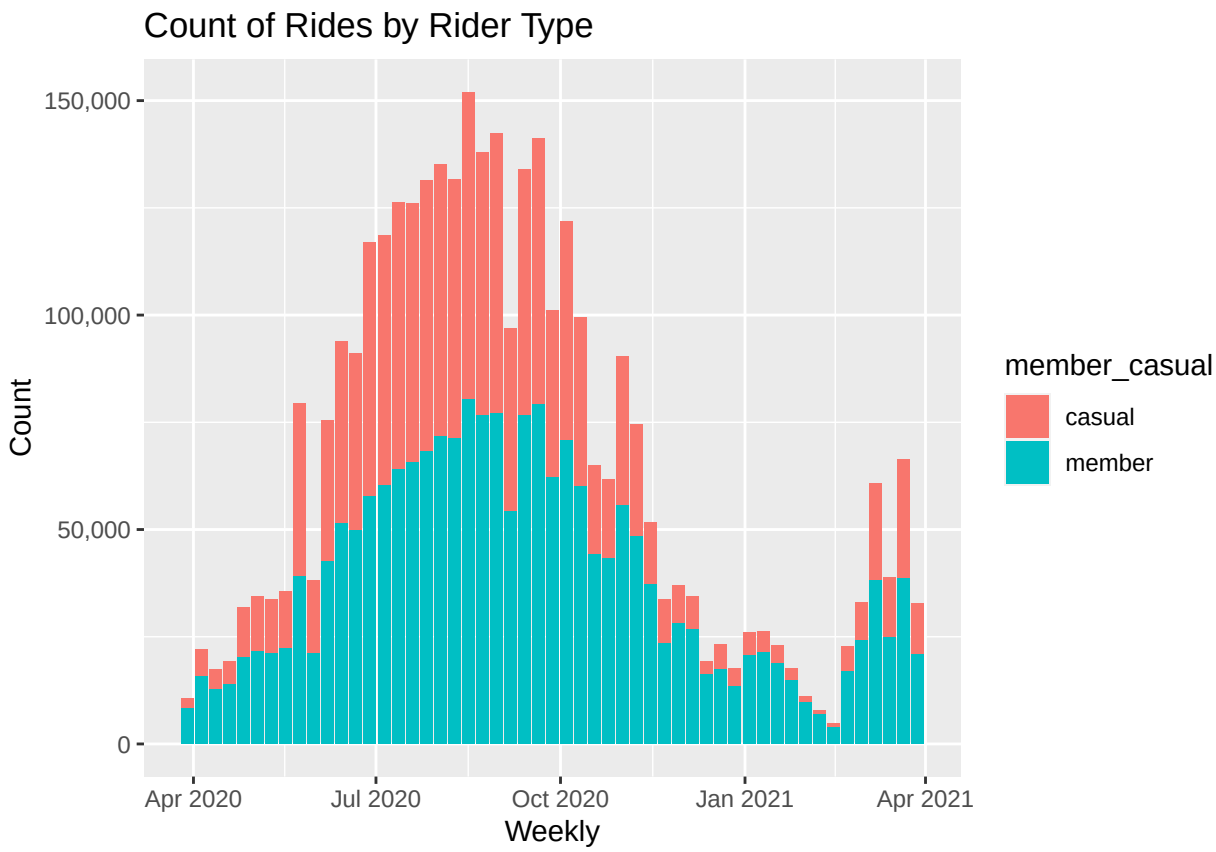
```
## `summarise()` has grouped output by 'member_casual', 'rideable_type'. You can override using the `.g
bikestype <- biketype %>% filter(Minutes >0,Mean >0, Median >0, Max >0, Min >0) %>% drop_na()
```

*Count by Bike Type (Total by Week)

```
ggplot(bikestype) +
  geom_col(aes(x=Weekly,y=Count,fill=rideable_type)) +
  scale_y_continuous(labels = comma) +
  labs(title="Count of Rides by Bike Type")
```

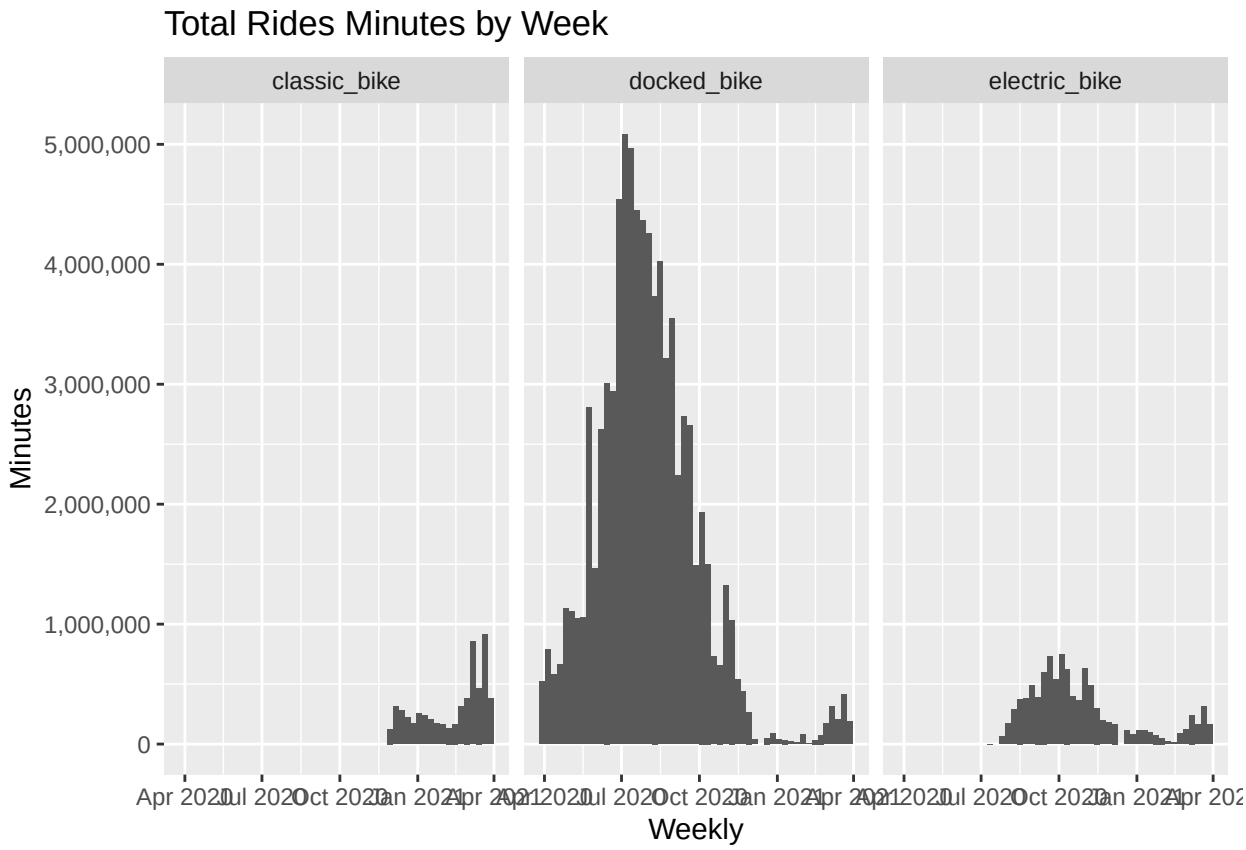



```
ggplot(bikestype) +
  geom_col(aes(x=Weekly,y=Count,fill=member_casual)) +
  scale_y_continuous(labels = comma) +
  labs(title="Count of Rides by Rider Type")
```




*Total Bike Rides by Week

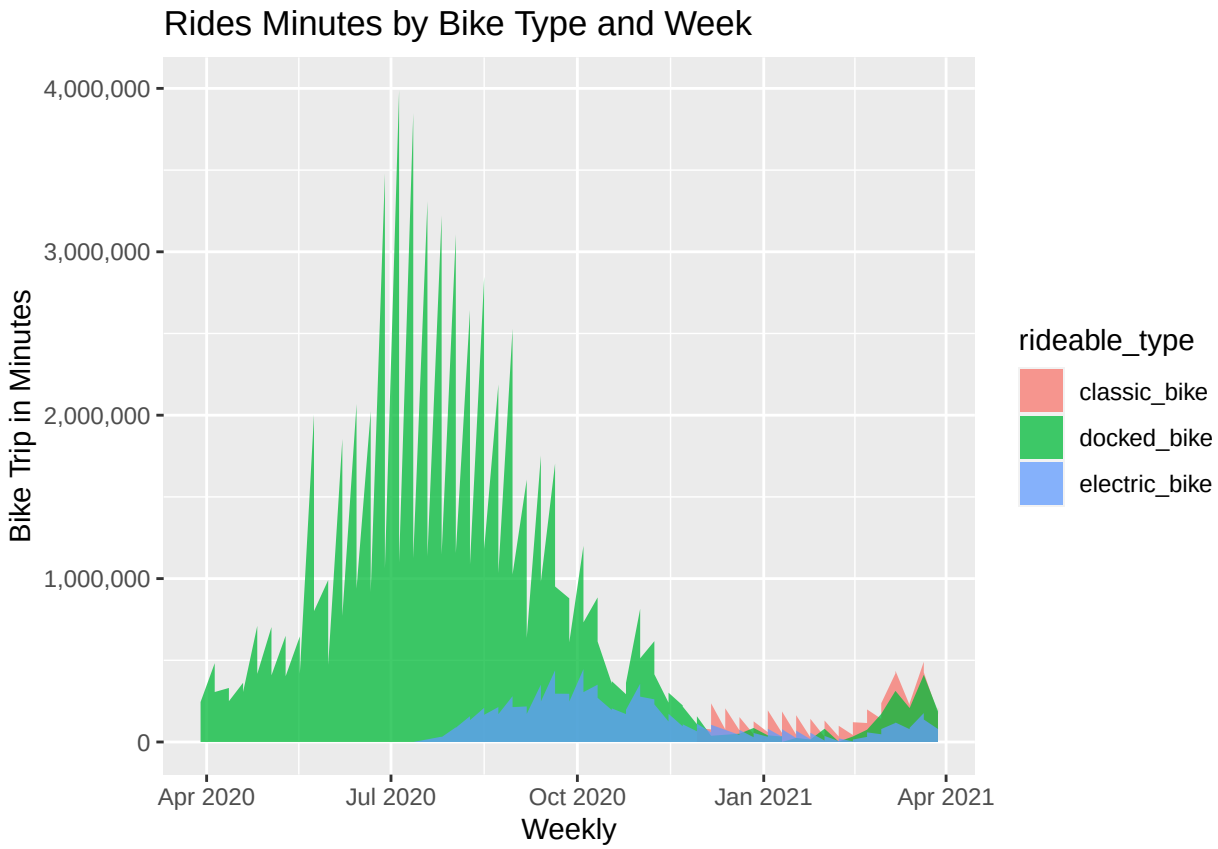
```
ggplot(bikestype) + geom_col(aes(x=Weekly,y=Minutes)) +
  scale_y_continuous(labels = comma) +
  facet_wrap(~rideable_type) +
  labs(title="Total Rides Minutes by Week")
```




*Rides By Bike Type and Week

```
ggplot(bikestype,aes(x=Weekly,y=Minutes,fill=rideable_type)) +
  geom_area(stat = "identity",position = position_dodge(),
            alpha = 0.75) +
  scale_y_continuous(labels = comma) +
  labs(title = "Rides Minutes by Bike Type and Week",
       y="Bike Trip in Minutes")
```

```
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```




Above is the comprehensive summary of the cyclistic bike share data for the twelve month period starting April 2020 to March 2021. More insights on the data are avalable in the attached analysis results as performed on SQLite and excel workbook. No further analysis can be conducted on the membership types and how it affects the number of rides since there is no enough data on the type of membership and the choice of membership per rider.