

# Citi Bike Usage in NYC

Espresso

24 06 2021

**MAT381E**

## **PROJECT REPORT**

### **Team Espresso**

- Oğuzhan Seleker
- Nilay Nacak
- Betül Kul

### **PROJECT INTRODUCTION**

- 2020 Covid-19 Pandemic has changed the habits and social behaviors of society in many ways. Cycling, the healthiest way of travel, is among the industries that was affected by the Covid-19 lock down. During the pandemic, social distancing was promoted throughout the world, especially in major cities such as New York.
- Visualizing the rental bike usage retrieved from Citi Bike database will enable the analysis of the effects of pandemic and lock down for the bike industry.
- It is also possible to make deductions about the user base depending on the change in the average age of the users over time, observe how the speed of cycling changes according to age groups, have information about the gender of the people who cycle frequently. In addition, mapping the most occupied and popular bike destinations of the city based on the inferences about the frequencies of stations.

## CITI BIKE INTRODUCTION

- New York City's bike share system, Citi Bike is the largest system in the USA and an essential part of the city's transportation network.
- Citi Bike consists of 20,000 bikes that are locked into a network of docking stations throughout the city and 1,300 stations across Manhattan, Brooklyn, Queens, the Bronx and Jersey City.
- The bikes can be unlocked from one station and returned to any other station in the system, making them ideal for one-way trips.
- People use bike share to commute to work or school, run errands, get to appointments or social engagements, and more.

### **How its used?**

Citi Bike is available for use 24 hours/day, 7 days/week, 365 days/year. Customers either become an Annual Member or buy a short-term pass through the Citi Bike app. Then it is possible to find an available bike nearby, and get a ride code or use your member key to unlock it. There is no limit for taking any rides once your pass or membership is active. The customer can return the rented bike to any station, and wait for the green light on the dock to make sure it's locked.

## PROJECT AIM

### **The CitiBike NYC Project aim :**

- To analyze Citi Bike users according to the personal information such as gender, age and customer type and gain a better understanding of the target customer profile of Citi Bike.
- To analyze the demographic nature of the with the calculation of the busiest bus stations.
- To show how people's bicycle usage was affected after the COVID-19 pandemic was announced.

## New York Covid-19 Pandemic

- The first case of COVID-19 in the USA, New York during the pandemic was confirmed on March 1, 2020.
- The state quickly became an epicenter of the pandemic.
- New York had the highest number of confirmed cases of any state from the start of U.S. outbreak until July 22.
- Approximately half of the state's cases have been in New York City, where nearly half the state's population lives.

## DATA SOURCE

Citi Bike has an immense transportation data that is harvested from the Citi Bike app. Their database is updated with thousand of new logs every single day. The large amounts of data is collected according to the NYCBS Data Use Policy by the company and is open to public on the company website. The bike sharing service has invited developers, engineers, statisticians, artists, academics to analyse, visualize and manipulate the NYC bike share data without any consequence.

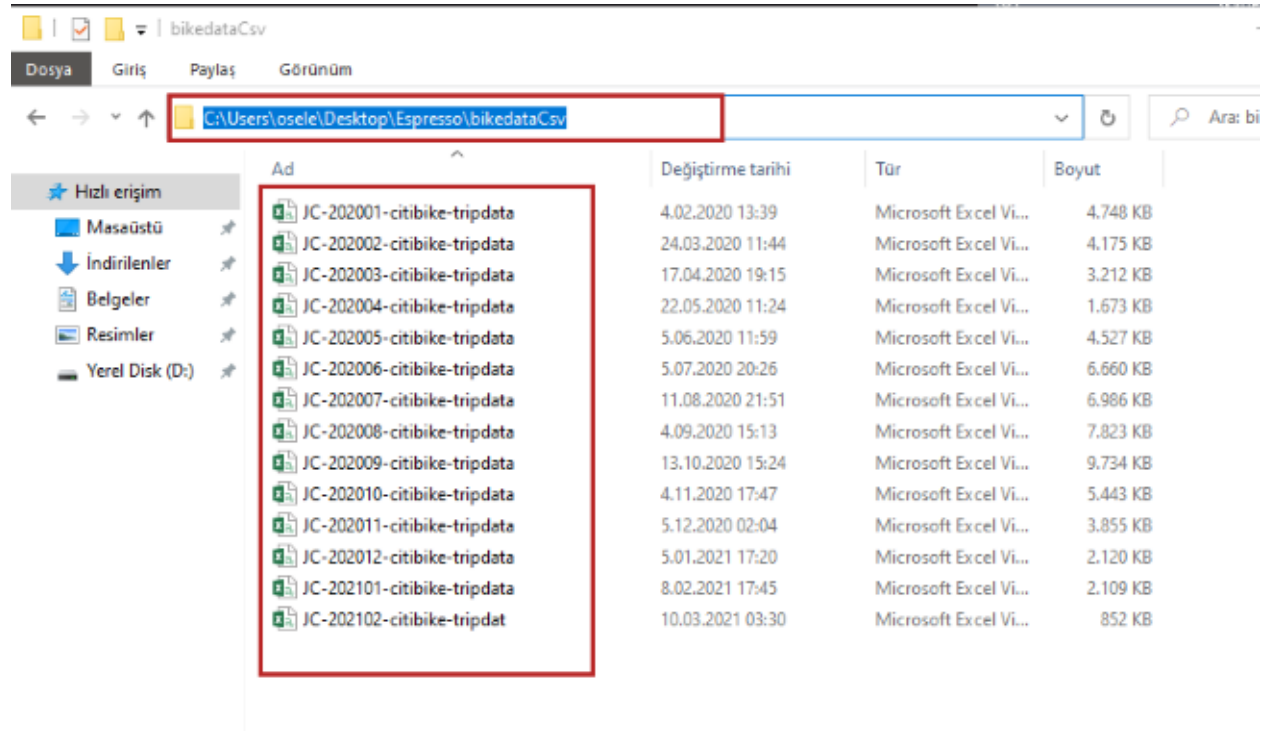
## LINK TO DATA

<https://s3.amazonaws.com/tripdata/index.html>

## TIDYING THE PROJECT DATA

Citi bike shares the application data in the format of .csv month by month. So, in order to examine from January 2020 to May 2021, it was necessary to combine the monthly .csv files into a single .csv file.

1-) For this, press Windows+R keys respectively. Type cmd in the Run window that opens. On the terminal screen that opens, to the directory where our .csv files are located.



2-) Copy the directory; example: “C:/Users/PcName/Desktop/Esspresso/bikeDataCsv” and paste the “C:/Users/PcName/Desktop/Esspresso/bikeDataCsv” onto the command line, then press enter.

```
C:\WINDOWS\system32\cmd.exe
```

```
Microsoft Windows [Version 10.0.19042.1052]
(c) Microsoft Corporation. Tüm hakları saklıdır.

C:\Users\osele>cd C:\Users\osele\Desktop\Esspresso\bikedataCsv_
```

```
C:\WINDOWS\system32\cmd.exe
```

```
Microsoft Windows [Version 10.0.19042.1052]
(c) Microsoft Corporation. Tüm hakları saklıdır.

C:\Users\osele>cd C:\Users\osele\Desktop\Esspresso\bikedataCsv

C:\Users\osele\Desktop\Esspresso\bikedataCsv>
```

3-) After going to the directory, write "copy \*.csv all\_data.csv" onto the command line, then press Enter.

```
Microsoft Windows [Version 10.0.19042.1052]
(c) Microsoft Corporation. Tüm hakları saklıdır.

C:\Users\osele>cd C:\Users\osele\Desktop\Esspresso\bikedataCsv

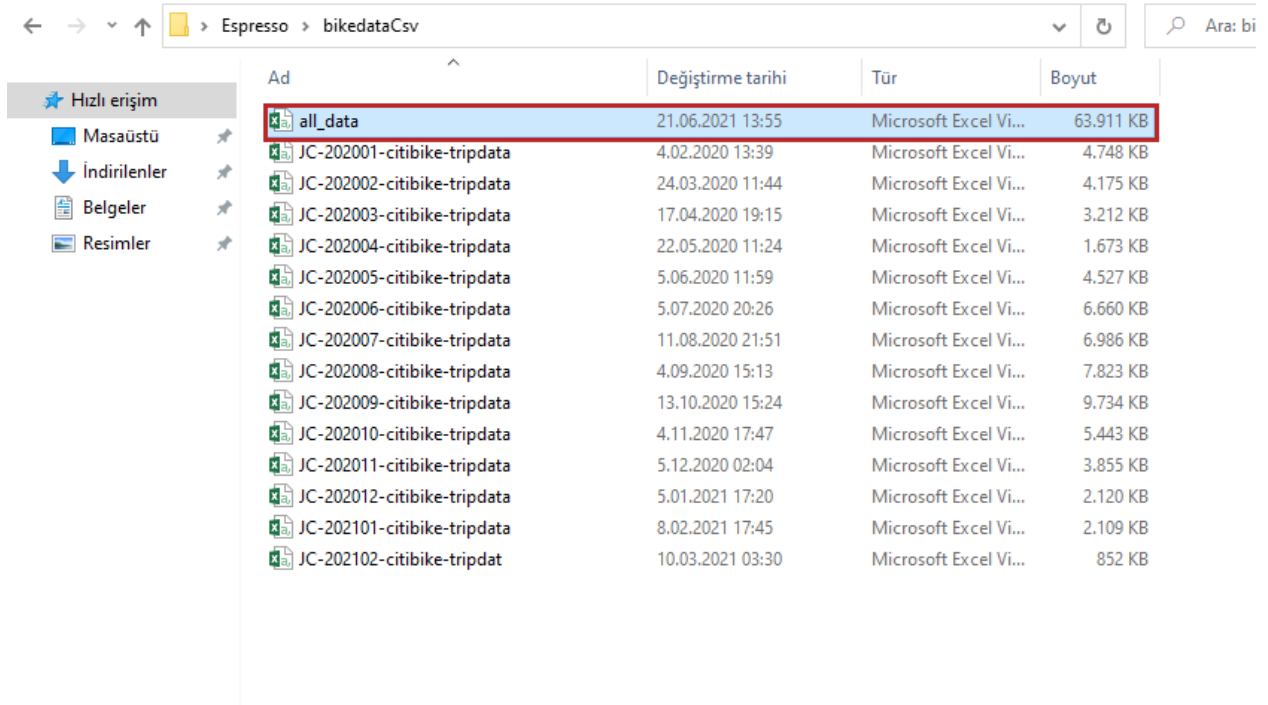
C:\Users\osele\Desktop\Esspresso\bikedataCsv>copy *.csv all_data.csv
```

```
C:\Users\osele>cd C:\Users\osele\Desktop\Esspresso\bikedataCsv

C:\Users\osele\Desktop\Esspresso\bikedataCsv>copy *.csv all_data.csv
JC-202001-citibike-tripdata.csv
JC-202002-citibike-tripdata.csv
JC-202003-citibike-tripdata.csv
JC-202004-citibike-tripdata.csv
JC-202005-citibike-tripdata.csv
JC-202006-citibike-tripdata.csv
JC-202007-citibike-tripdata.csv
JC-202008-citibike-tripdata.csv
JC-202009-citibike-tripdata.csv
JC-202010-citibike-tripdata.csv
JC-202011-citibike-tripdata.csv
JC-202012-citibike-tripdata.csv
JC-202101-citibike-tripdata.csv
JC-202102-citibike-tripdat.csv
1 file(s) copied.

C:\Users\osele\Desktop\Esspresso\bikedataCsv>_
```

4-) Find the new “all\_data.csv” file under the directory that it was created.



The screenshot shows a Windows File Explorer window with the address bar set to 'Esspresso > bikedataCsv'. The left sidebar shows the 'Hızlı erişim' (Quick access) pane with 'Masaüstü' (Desktop), 'İndirilenler' (Downloads), 'Belgeler' (Documents), and 'Resimler' (Pictures). The main pane displays a list of files with columns for 'Ad' (Name), 'Değiştirme tarihi' (Date modified), 'Tür' (Type), and 'Boyut' (Size). The file 'all\_data' is highlighted with a red box.

| Ad                          | Değiştirme tarihi | Tür                   | Boyut     |
|-----------------------------|-------------------|-----------------------|-----------|
| all_data                    | 21.06.2021 13:55  | Microsoft Excel Vi... | 63.911 KB |
| JC-202001-citibike-tripdata | 4.02.2020 13:39   | Microsoft Excel Vi... | 4.748 KB  |
| JC-202002-citibike-tripdata | 24.03.2020 11:44  | Microsoft Excel Vi... | 4.175 KB  |
| JC-202003-citibike-tripdata | 17.04.2020 19:15  | Microsoft Excel Vi... | 3.212 KB  |
| JC-202004-citibike-tripdata | 22.05.2020 11:24  | Microsoft Excel Vi... | 1.673 KB  |
| JC-202005-citibike-tripdata | 5.06.2020 11:59   | Microsoft Excel Vi... | 4.527 KB  |
| JC-202006-citibike-tripdata | 5.07.2020 20:26   | Microsoft Excel Vi... | 6.660 KB  |
| JC-202007-citibike-tripdata | 11.08.2020 21:51  | Microsoft Excel Vi... | 6.986 KB  |
| JC-202008-citibike-tripdata | 4.09.2020 15:13   | Microsoft Excel Vi... | 7.823 KB  |
| JC-202009-citibike-tripdata | 13.10.2020 15:24  | Microsoft Excel Vi... | 9.734 KB  |
| JC-202010-citibike-tripdata | 4.11.2020 17:47   | Microsoft Excel Vi... | 5.443 KB  |
| JC-202011-citibike-tripdata | 5.12.2020 02:04   | Microsoft Excel Vi... | 3.855 KB  |
| JC-202012-citibike-tripdata | 5.01.2021 17:20   | Microsoft Excel Vi... | 2.120 KB  |
| JC-202101-citibike-tripdata | 8.02.2021 17:45   | Microsoft Excel Vi... | 2.109 KB  |
| JC-202102-citibike-tripdat  | 10.03.2021 03:30  | Microsoft Excel Vi... | 852 KB    |

## ABOUT THE PROJECT DATA

The data frame is made out of **353176 lines** and **15 columns** that are type “character” .This means that the NYC Bike Usage Data will analyze more than 350k bike rides.

### The contents of the NYC Citi Bike Data :

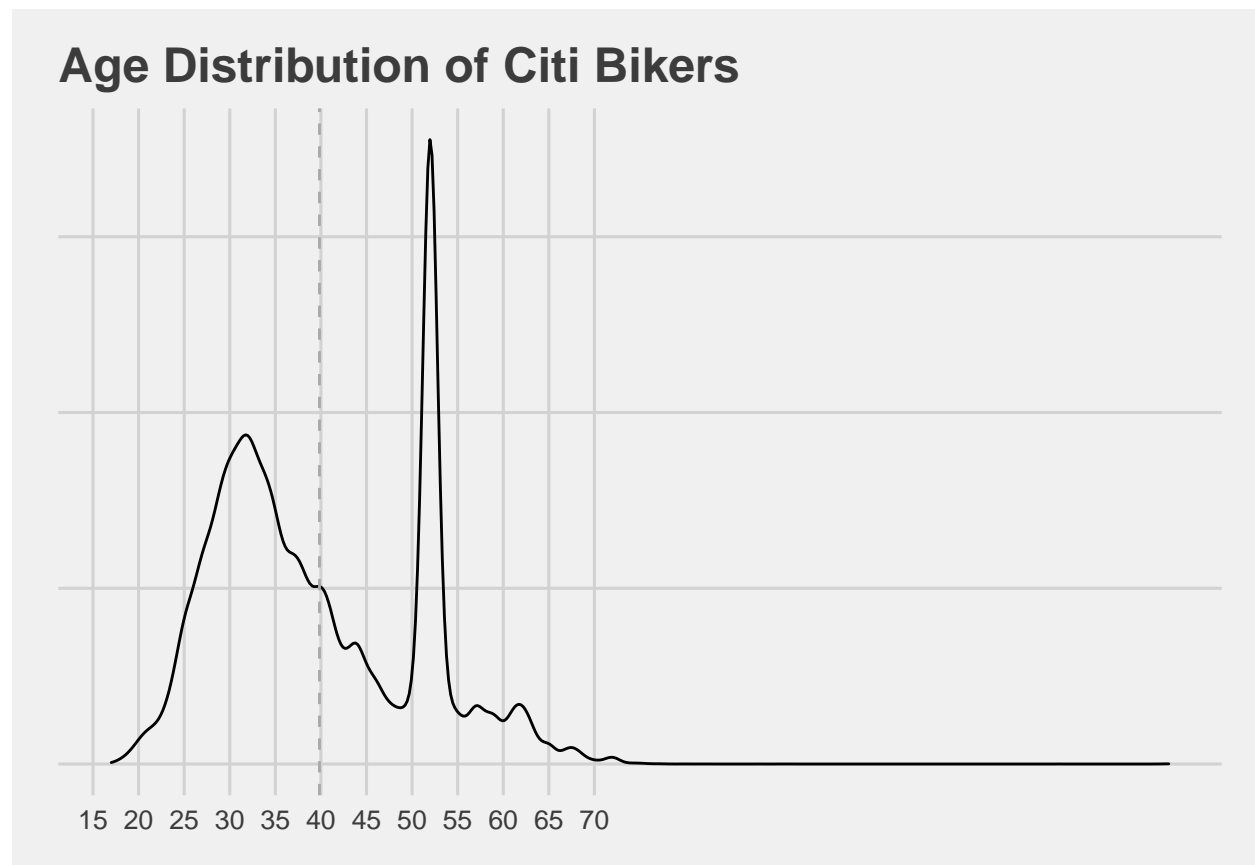
- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

## Plotting the Age Distribution of Citi Bikers

- The project data “all\_data” is stored as the variable name “citi” with read.csv() function of R.
- Birth Year column of the citi is *character* type. So in order to calculate the age of the customers, we turn it to numeric value using the as.numeric function.

```
library(ggplot2)
library(dplyr)
library(ggthemes)
citi <- read.csv("all_data.csv")
citi$birth.year <- as.numeric(citi$birth.year)
citi$age <- 2021-citi$birth.year
citi %>% ggplot(aes(x=age)) + geom_density()+theme_fivethirtyeight()+
  theme(axis.text.y=element_blank())+ggtitle("Age Distribution of Citi Bikers")+
  geom_vline(xintercept=median(citi$age,na.rm=T),linetype="dashed",col="dark grey")
```

## Results of the Age Distribution of Citi Bikers Plot



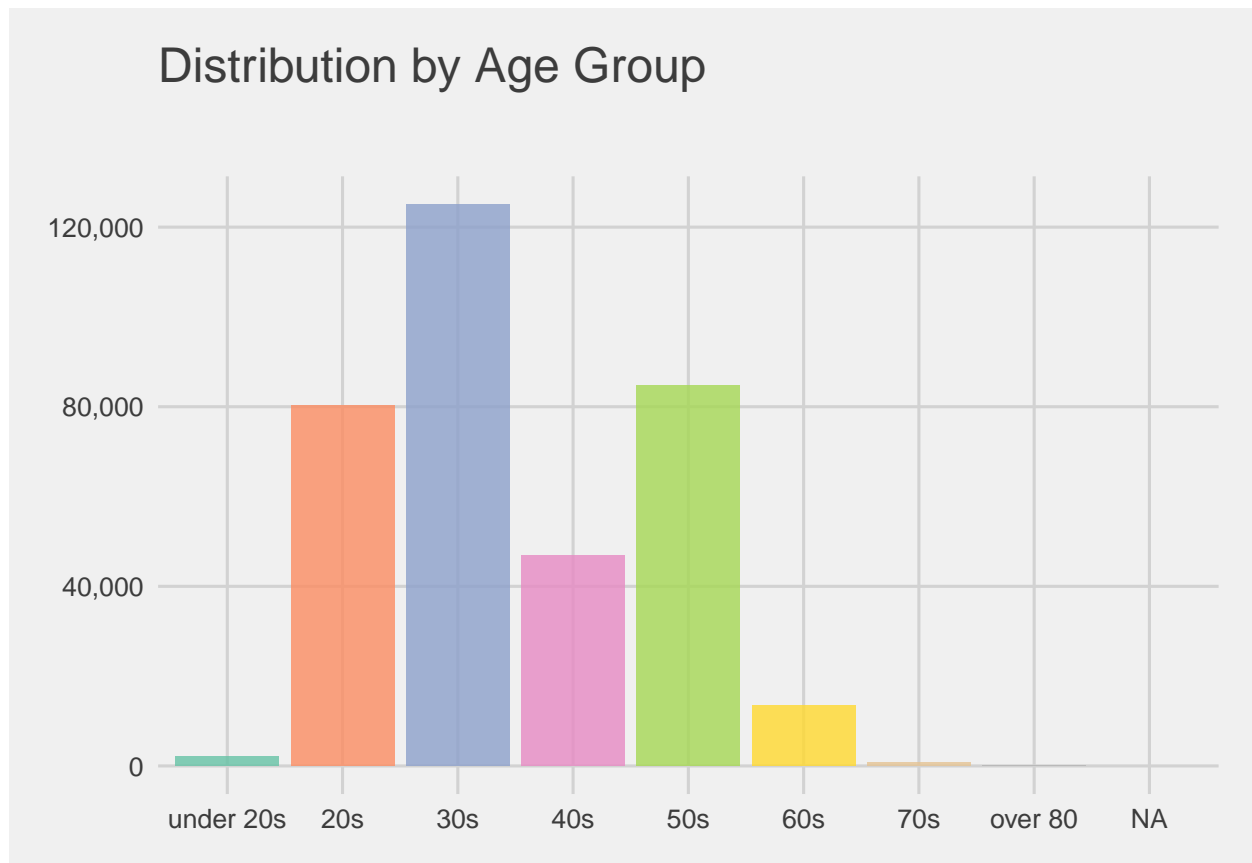
Average Citi Bike user age is close to 40. According to the graph ages between 50-55 are among the most frequently bike rented age interval. It is also easy to see that ages from 30 to 35 prefer Citi Bike more than others.

## Plotting Age Groups of Citi Bikers

After creating an age column for **citi**, we created the age groups starting from 20 to 90 labeled as 'under 20s' to '90s' separated by decade.

```
library(ggplot2)
library(dplyr)
library(ggthemes)
citi <- read.csv("all_data.csv")
citi$birth.year <- as.numeric(citi$birth.year)
citi$age <- 2021-citi$birth.year
citi$age_group <- cut(citi$age,breaks=c(0,20,30,40,50,60,70,80,90),
                      ,labels=c("under 20s","20s","30s","40s","50s","60s","70s","over 80"))
citi %>% ggplot(aes(x=age_group,fill=age_group)) + geom_bar(alpha=.8)+
  theme_fivethirtyeight()+scale_fill_brewer(palette="Set2")+
  theme(legend.position="false")+ggtitle(expression(atop("Distribution by Age Group")))+
  scale_y_continuous(labels=comma)
```

## Results of Age Groups of Citi Bikers Plot



Citi Bike has customer from every age group up until the 70s. People who are in their 30s are the most frequent Citi Bike users of NYC. People in their 50s and 20s have a similar user base around 80k.



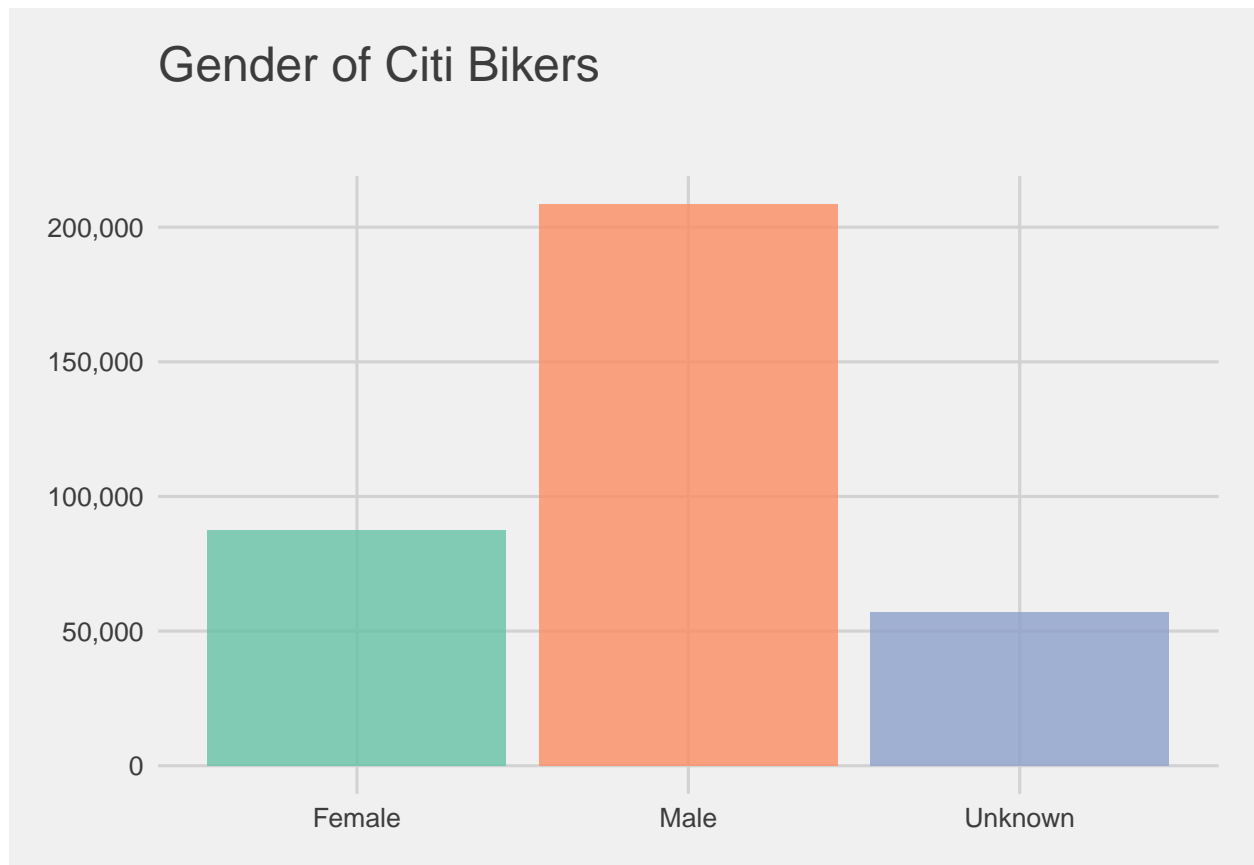
## Plotting Gender Distribution of Citi Bikers

- The gender value of the **citi** data frame consists three values from 0 to 2.
- In order to prepare the data for plotting, using ifelse statament the values are changed according to their equivalent string values.

```
library(ggplot2)
library(dplyr)
library(ggthemes)
citi <- read.csv("all_data.csv")
citi$gender <- ifelse(citi$gender==1,"Male",ifelse(citi$gender==2,"Female","Unknown"))

citi %>% ggplot(aes(x=gender,fill=gender)) + geom_bar(alpha=.8) +
  theme_fivethirtyeight() + scale_fill_brewer(palette="Set2")+
  theme(legend.position="none")
+ggtitle(expression(atop("Gender of Citi Bikers")))+scale_y_continuous(labels=comma)
```

## Results of Gender Distribution Plot



Male users create most of the Citi Biker base , more than twice of the female bikers. Some of the logs do'nt have gender information because they are not subscribers.

## Mapping the Station Frequencies Based on Starting Points

Station are grouped according to their id's using `group_by` function, then added station informations; latitude, longitude, name and trip count(n).

```
library(ggplot2)
library(dplyr)
library(scales)
library(htmlwidgets)
library(leaflet)

station.info <- citi %>%
  group_by(start.station.id) %>%
  summarise(lat=as.numeric(start.station.latitude[1]),
            long=as.numeric(start.station.longitude[1]),
            name=start.station.name[1],
            n.trips=n())
mybins <- seq(0, 60000, by=10000)
mypalette <- colorBin( palette="YlOrBr", domain=station.info$n.trips
                      , na.color="transparent", bins=mybins)
```

- Map Tooltip

The tooltip enables viewing station information when hovered on the map.

```
mytext <- paste(
  "Frequency: ", station.info$n.trips, "<br/>",
  "StationName: ", station.info$name, "<br/>",
  "StationId: ", station.info$start.station.id, sep="") %>%
  lapply(htmltools::HTML)
```

- Creating & Saving the Map

“Esri.WorldImagery” provides world map and zoom on the NYC province. Then using `addCircleMarkers`, station nlocations are marked with circles according to usage frequency.

```
m <- leaflet(stationNumber) %>%
  addTiles() %>%
  setView( lat=40.73, lng = -74.04 , zoom=13) %>%
  addProviderTiles("Esri.WorldImagery") %>%
  addCircleMarkers(~long, ~lat,
                  fillColor = ~mypalette(n.trips), fillOpacity = 0.7, color="white"
                  , radius=8,stroke=FALSE,
                  label = mytext,
                  labelOptions = labelOptions( style = list("font-weight" = "normal"
                  , padding = "3px 8px"), textSize = "13px", direction = "auto")
  ) %>%
  addLegend( pal=mypalette, values=~n.trips, opacity=0.9, title = "Frequency"
            , position = "bottomright" )
saveWidget(m, file=paste0( getwd(), "/StationMapping.html"))
```

## Results of the Station Frequencies Based on Starting Points Map

Since the map has motion attributes, check out the HTML output linked below.

[Click Here to View the Map](#)

The map colors the Citi Bike stations based on the frequency of bike usage on NYC. With hovering over the station locations on the map frequency of bike usage, station name and id is shown.

## Plotting the Average Speed of Age Groups of Citi Bikers

In order to calculate speed, using **distHaversine()** function distance taken on every trip is determined.

```
library(ggplot2)
library(ggthemes)
library(dplyr)
library(scales)
library(ggmap)
library(geosphere)

citi <- read.csv("all_data.csv")
citi$birth.year <- as.numeric(citi$birth.year)
citi$age <- 2021-citi$birth.year
citi$age_group <- cut(citi$age,breaks=c(0,30,40,50,60,70,80,100)
                     ,labels=c("under 30","30s","40s","50s","60s","70s","over 80"))
citi$start.station.latitude <- as.numeric(citi$start.station.latitude)
citi$start.station.longitude <- as.numeric(citi$start.station.longitude)
citi$end.station.latitude <- as.numeric(citi$end.station.latitude)
citi$end.station.longitude <- as.numeric(citi$end.station.longitude)
citi$Distance <- distHaversine(citi[,6:7], citi[,10:11])
```

- Calculating Speed

Since trip duration is type 'character', in order to parse it to integer, **as.integer** function is used. Then average speed is calculated by dividing distance to duration. Then speed unit is turned into km/h.

```
#m/s
citi$tripduration <- as.integer(citi$tripduration)
citi$avg_speed <- citi$Distance / citi$tripduration

#converted m/s to km/h
citi$avg_speed <- citi$avg_speed * 36 / 10
```

Some Citi Bike users returned their bike to the same station they departed, therefore distance and speed of these users are equal to zero. By filtering trips with average speed greater than zero, the calculations are corrected.

A new data frame is created by taking average speed data grouped by age.

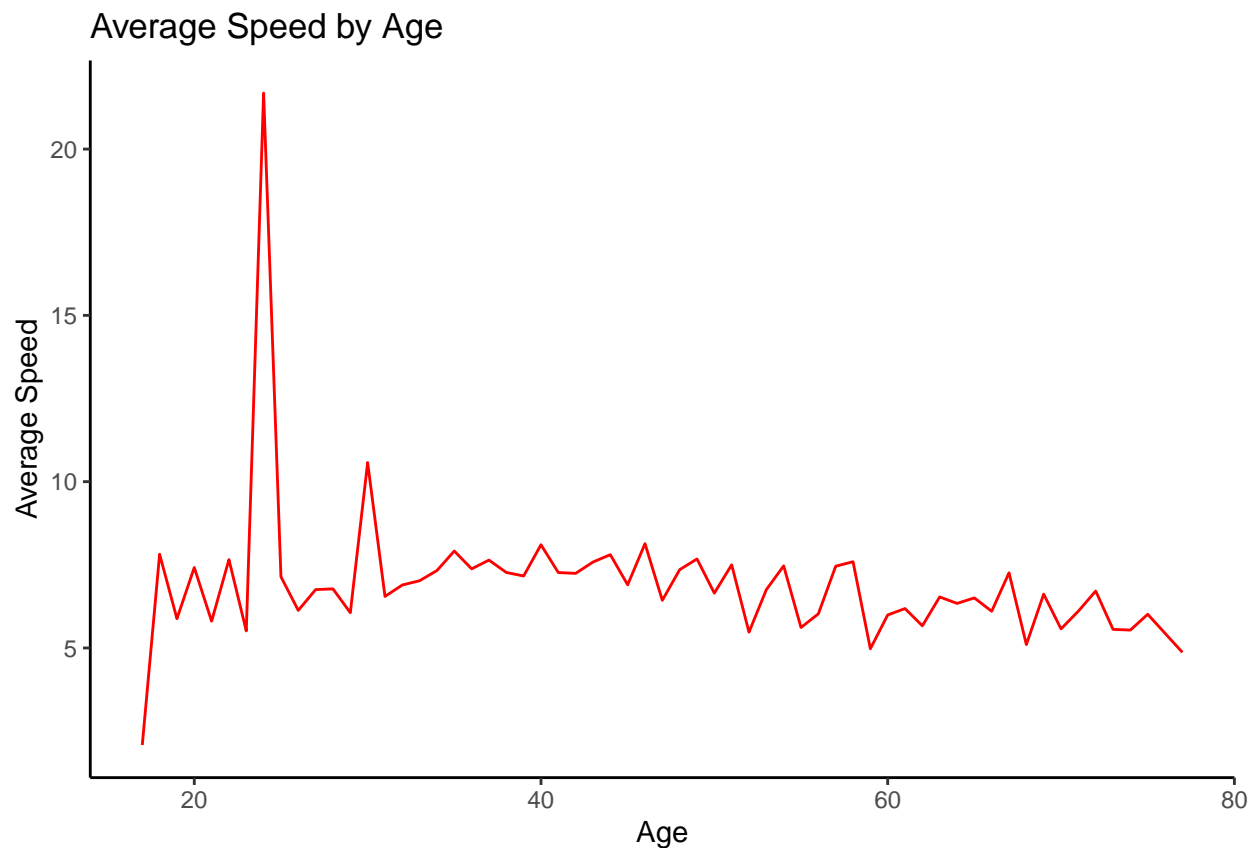
Some Citi Bike users entered illogical birth year informations, thus filtering users older than 80 was necessary.

```
#filter average speed bigger than 0
citi <- citi %>% filter(avg_speed > 0)

# average speed group by age group
average_speed_by_age <- citi %>% group_by(age) %>% summarize(mean_speed = mean(avg_speed))

#filter age smaller than 80
average_speed_by_age <- average_speed_by_age %>% filter(age < 80)
line_graph <- ggplot(average_speed_by_age) + geom_line(aes(age, mean_speed), color="red")
+ theme_classic() + labs(title="Average Speed by Age", x= "Age", y="Average Speed")
line_graph
```

## Results of the Average Speed by Age Plot



The graph shows average user speed according to age groups. Average speed is highest on people in their 30s with a huge difference from the others. Other age groups have ups and downs.

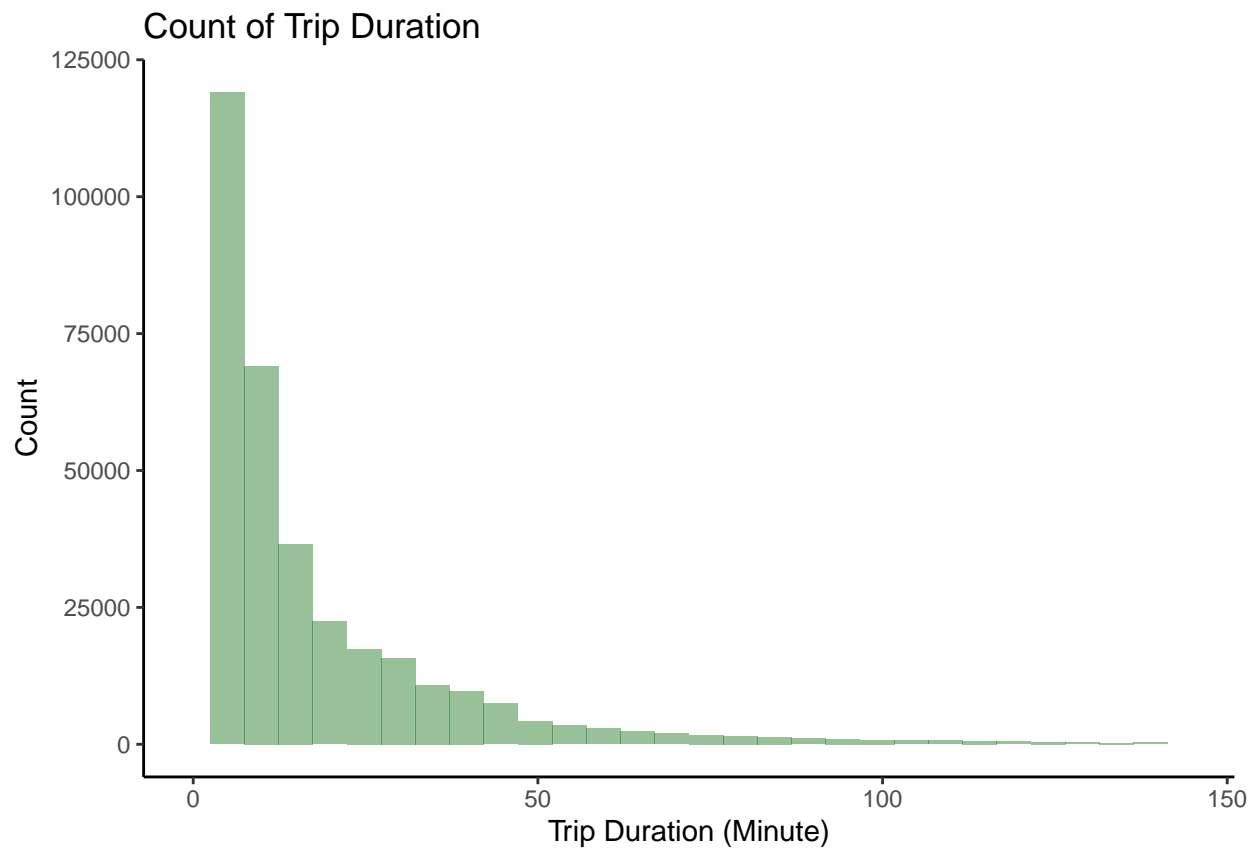
## Plotting the Trip Count of Trip Durations

Trip duration is turned into minutes then count of trips of based on trip durations with five minute time intervals are calculated.

```
library(ggplot2)
library(dplyr)

citi <- read.csv("all_data.csv")
# Trip duration data graph
citi <- mutate(citi, tripduration.min = as.numeric(tripduration)/60, na.rm=T)
x.max <- quantile(citi$tripduration.min, 0.99, na.rm=T)
ggplot(citi) + geom_histogram(aes(tripduration.min), fill="darkgreen", alpha = 0.4) +
  xlim(c(0,x.max)) + theme_classic() + labs(title="Count of Trip Duration",
                                             x= "Trip Duration (Minute)", y="Count")
```

## Results of the Trip Count of Trip Durations Plot



Most Citi Bike rides last 0 to 5 minutes. Thus Citi Bike is mostly used for short trips.

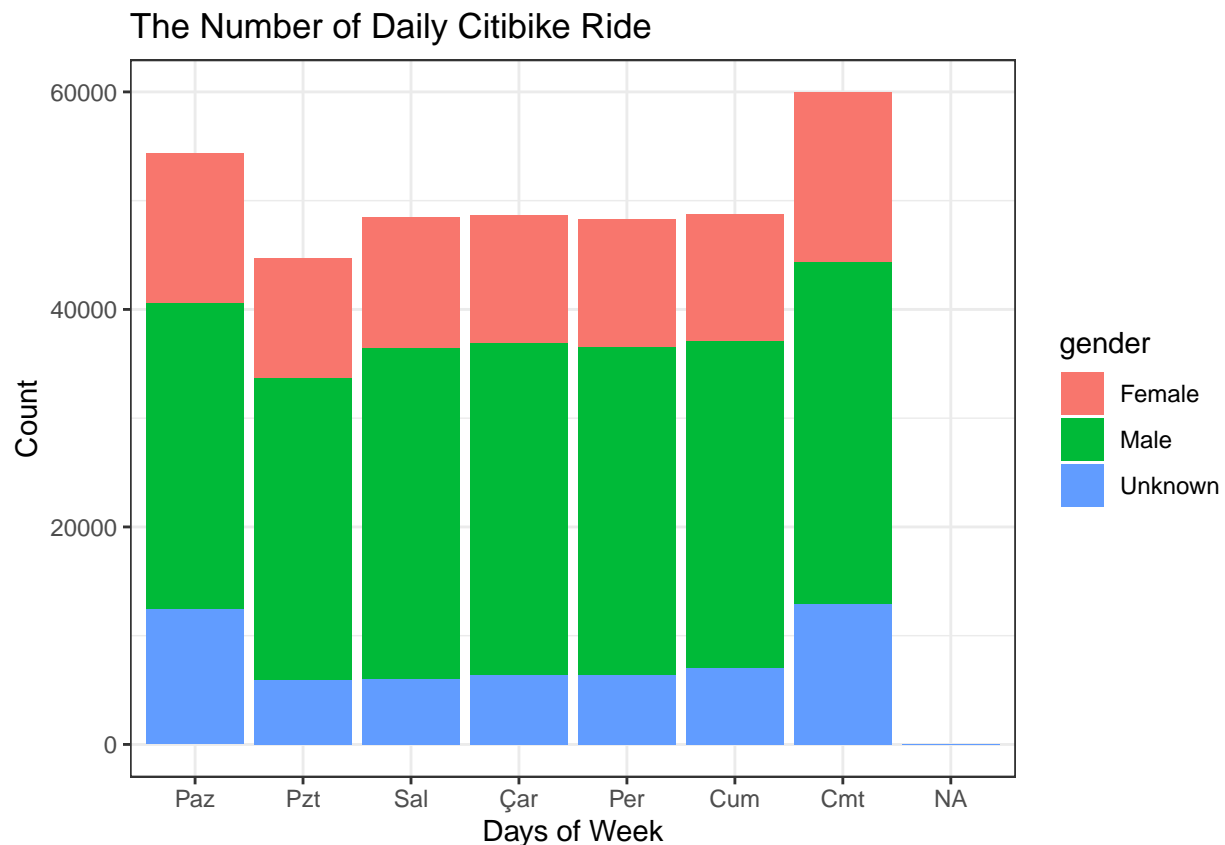
## Plotting the Occupency of Weekdays Based on Gender

Using the date information from the start time column, which day of the week the trip was made is determined.

```
library(ggplot2)
library(dplyr)
library(lubridate)

citi <- read.csv("all_data.csv")
citi$gender <- ifelse(citi$gender==1,"Male",ifelse(citi$gender==2,"Female","Unknown"))
citi$weekday <- wday(as.Date(citi$starttime), label=TRUE,)
ggplot(citi) + geom_bar(aes(x=weekday, fill=gender)) + theme_bw() + ylab("") +
  labs(title="The Number of Daily Citibike Ride", x= "Days of Week", y="Count")
```

## Results of the Occupency of Weekdays Based on Gender Plot



The number of daily citi bike ride graph shows that rental bike usage is higher on weekends according to the weekdays. It is possible to conclude that male riders are more than female users for every day of the week.

## Plotting the Effects of Pandemic

- After the Pandemic was Announced

Taking the rides made within the two months time period, from March 1 to May 1, after pandemic was first announced, 'CityPandemic' data frame is created.

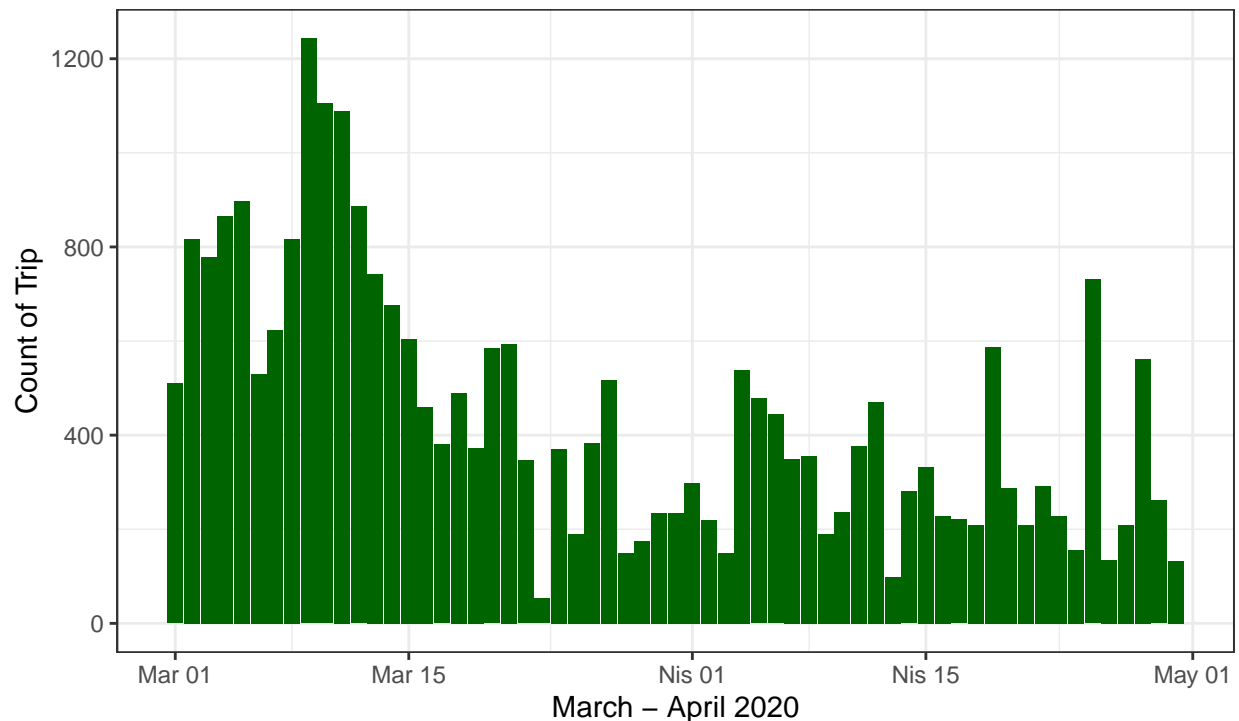
```
library(ggplot2)
library(dplyr)
library(lubridate)

citi <- read.csv("all_data.csv")
CitiPandemic <- subset(citi,(starttime >= "2020-03-01" & starttime <= "2020-04-31")
, select=c(starttime))
ggplot(CitiPandemic) + geom_bar(aes(x=as.Date(starttime)),fill="darkgreen")
+ theme_bw() + ylab("") +labs(title="The Time When Pandemic Was Declared"
, x= "March - April 2020", y=" Count of Trip "
, subtitle = "The first case of the COVID-19 pandemic in New York City
was confirmed on March 2020")
```

## Results of the Effects of Pandemic Plot

### The Time The Pandemic Was Declared

The first case of the COVID-19 pandemic in New York City  
was confirmed on March 2020



The graph show the daily Citi Bike Counts between March 1 of to May 1. After a short time when the Covid-19 pandemic confirmed in NYC, the Citi Bike usage in the city decreased immensely.

## Plotting the Effects of Pandemic

- From the Start of Pandemic to Today

In order to see the general effect of pandemic to Citi Bike rides, times period is extended to May 31.

```
library(ggplot2)
library(dplyr)
library(lubridate)

citi <- read.csv("all_data.csv")

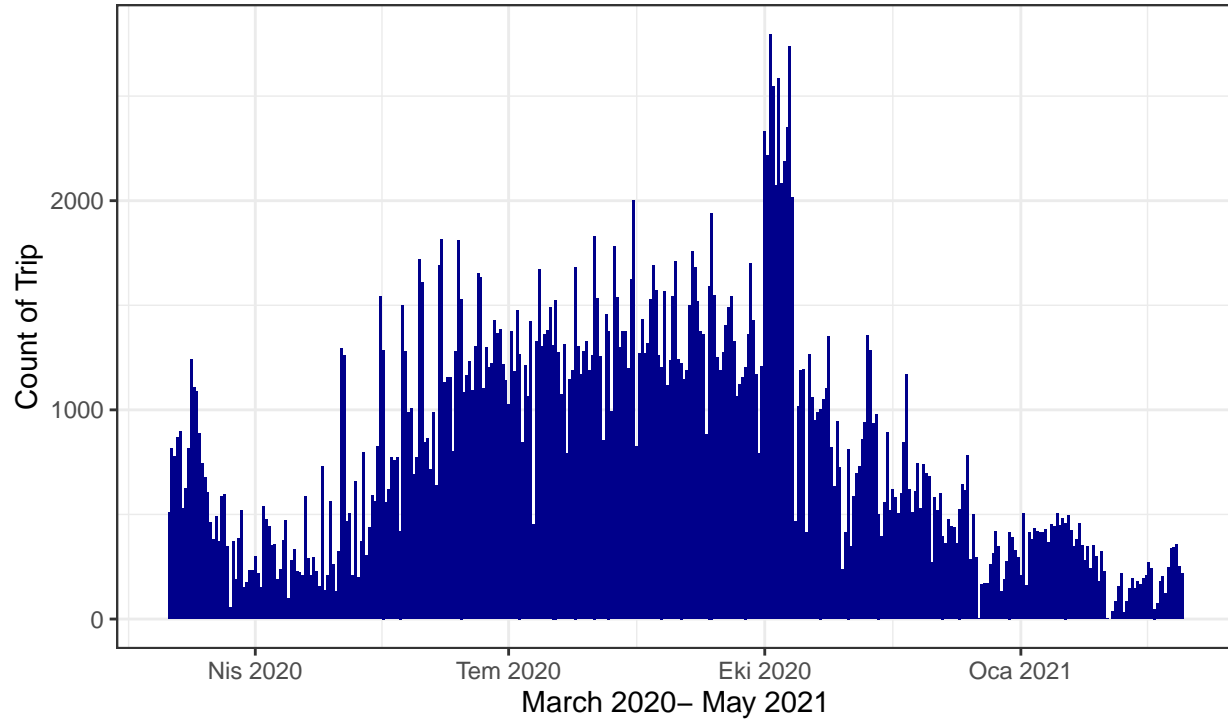
CitiAllPandemic <- subset(citi,(starttime >= "2020-03-01" & starttime <= "2021-05-31")
                          , select=c(starttime))
ggplot(CitiAllPandemic) + geom_bar(aes(x=as.Date(starttime),fill="darkblue"))
+ theme_bw() + ylab("") +labs(title="Daily Bike Usage After Covid-19 Pandemic"
                             , x= "March - April 2020", y=" Count of Trip "
                             , subtitle = "The Pandemic is continuing since March 1.
The vaccination started on
January 8, 2021 for New Yorkers.")
```



## Results of the Effects of Pandemic Plot

### Daily Bike Usage After Covid–19 Pandemic

The Pandemic is continuing since March 1. The vaccination started on January 8, 2021 for New Yorkers.



It is possible to detect that during 2020 summer, people prefer riding bikes since it enables social distancing. The reason of the sudden increase in ride counts during October 2020 could be explained with the article linked below. Citi Bike has tripled the number of bikes in their fleet until the end of September and continued to add hundreds of bikes every week until the end of October. [Article](#)

## Conclusion

As a result, in our study, we have seen the change in the prevalence of bicycle use in NYC over time and the effect of the pandemic. At the same time, we observed in the graphs that we produced that gender and age factors have a significant effect on usage. Although it was observed that bicycle use was affected at the beginning of the pandemic, in conclusion, it is determined that users preferred bicycles again and the number of Citi Bike bicycle rentals increased from time to time. According to the results and conclusion that was obtained, it could be stated that there will be an increase in the number and use of bicycles in the near future.

## References

Citi Bike's ebike rides in 2020 reach 1 million, Brooklyn Daily Eagle, 2020. Retrieved from ; <https://brooklyneagle.com/articles/2020/10/16/citi-bikes-ebike-rides-in-2020-reach-1-million/>

Covid-19 Pandemic in New York State, Wikipedia. Retrieved from ; [https://en.wikipedia.org/wiki/COVID-19\\_pandemic\\_in\\_New\\_York\\_\(state\)](https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_(state))

Citi Bike How it works?, citi bike. Retrieved from ; <https://www.citibikenyc.com/how-it-works>