Team Bike

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```r
library(ggplot2)
library(readr)
library(lubridate)
```

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

## The following object is masked from 'package:base':
##
##     date
```

```r
library(chron)
```

```
##
## Attaching package: 'chron'

## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

```r
#library(Rtsne)

randomsample <- read_csv("data/randomsample.csv")
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   tripduration = col_integer(),
##   starttime = col_character(),
##   stoptime = col_character(),
##   start.station.id = col_integer(),
##   start.station.name = col_character(),
##   start.station.latitude = col_double(),
##   start.station.longitude = col_double(),
##   end.station.id = col_integer(),
##   end.station.name = col_character(),
##   end.station.latitude = col_double(),
##   end.station.longitude = col_double(),
##   bikeid = col_integer(),
##   usertype = col_character(),
##   birth.year = col_integer(),
##   gender = col_integer()
## )
```

```
# Parsing all the start times into one format
mdy <- mdy_hms(randomsample$starttime)
```

```
## Warning: 285131 failed to parse.
```

```
ymd <- ymd_hms(randomsample$starttime)
```

```
## Warning: 232836 failed to parse.
```

```
f1 <- mdy_hm(randomsample$starttime)
```

```
## Warning: 482033 failed to parse.
```

```
mdy[is.na(mdy)] <- ymd[is.na(mdy)] # some dates are ambiguous, here we give
randomsample$starttime <- mdy
randomsample$starttime[is.na(randomsample$starttime)] <- f1[is.na(randomsample$starttime)]

# Parsing all the end times into one format
mdy <- mdy_hms(randomsample$stoptime)
```

```
## Warning: 285131 failed to parse.
```

```
ymd <- ymd_hms(randomsample$stoptime)
```

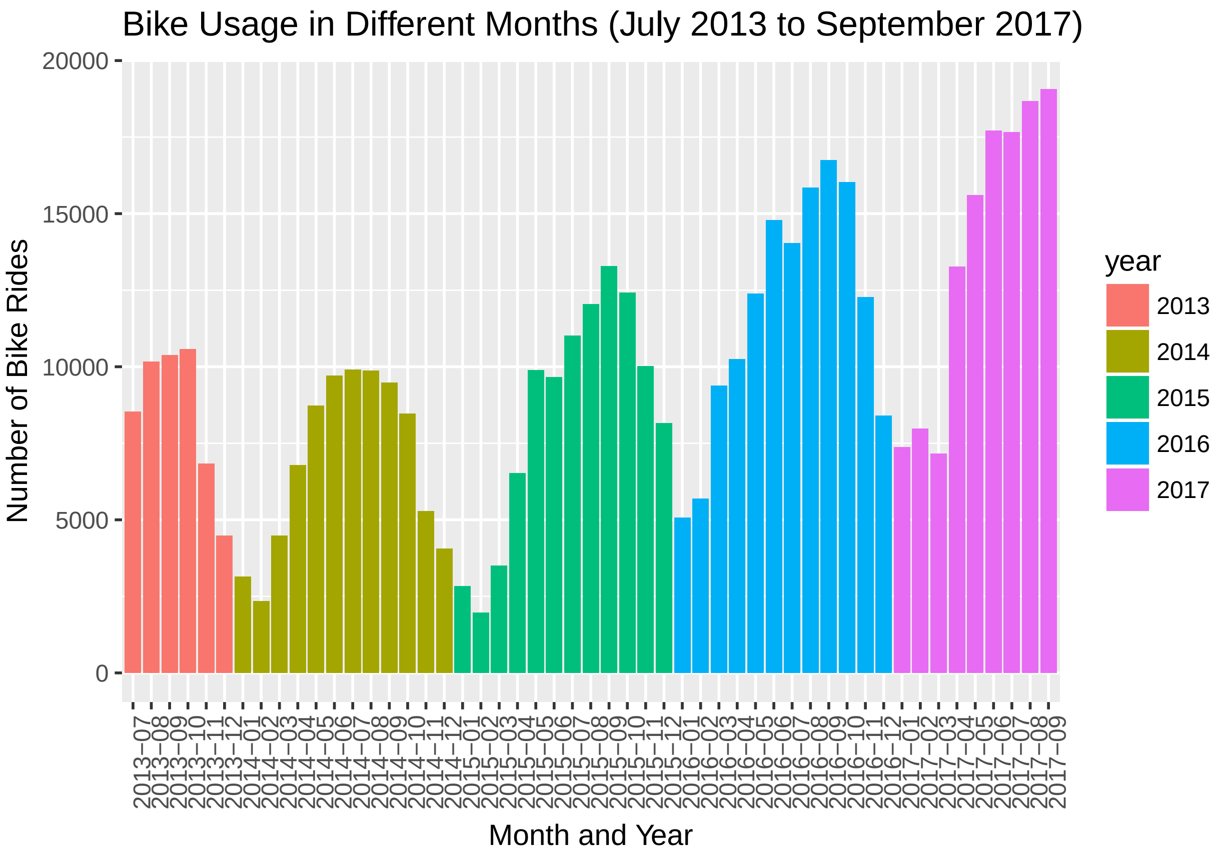
```
## Warning: 232836 failed to parse.
```

```
f1 <- mdy_hm(randomsample$stoptime)
```

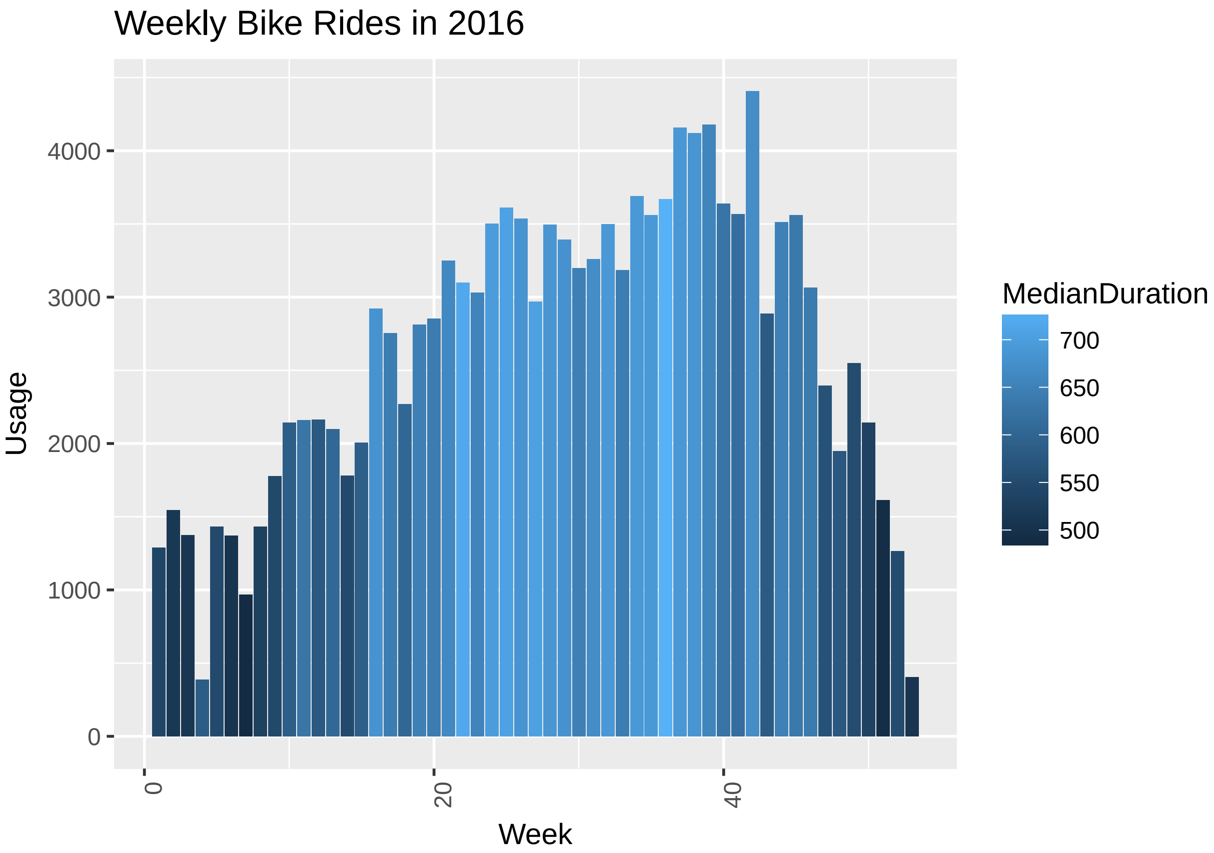
```
## Warning: 482033 failed to parse.
```

```
mdy[is.na(mdy)] <- ymd[is.na(mdy)]
randomsample$stoptime <- mdy
randomsample$stoptime[is.na(randomsample$stoptime)] <- f1[is.na(randomsample$stoptime)]

# Flooring the Hour
randomsample$starttime <- floor_date(randomsample$starttime,unit = "hour")
# Trying to Find Day of the Week
randomsample <- randomsample %>%
  mutate(WeekDay=wday(as.Date(starttime), label=TRUE)) %>%
  mutate(WorkingDay=ifelse((WeekDay=="Mon"|WeekDay=="Tues"|WeekDay=="Wed"|WeekDay=="Thurs"|WeekDay=="Fr

# Trying to find the holidays - Not quite working. Need to do this manually!
randomsample <- randomsample %>%
  mutate(holiday=is.weekend(starttime))



# Joining the weather data and the random sample
combined <- randomsample

# Number of Stations Every Year
rs <- randomsample %>%
  mutate(year=year(starttime),monyear=substr(starttime,1,7)) %>%
  group_by(year,usertype,monyear) %>%
  summarise(stations=n_distinct(start.station.id),rides=n())
combined$WorkingDay[combined$WorkingDay=="Yes"]<- "Weekday"
combined$WorkingDay[combined$WorkingDay=="No"]<- "Weekend"
```

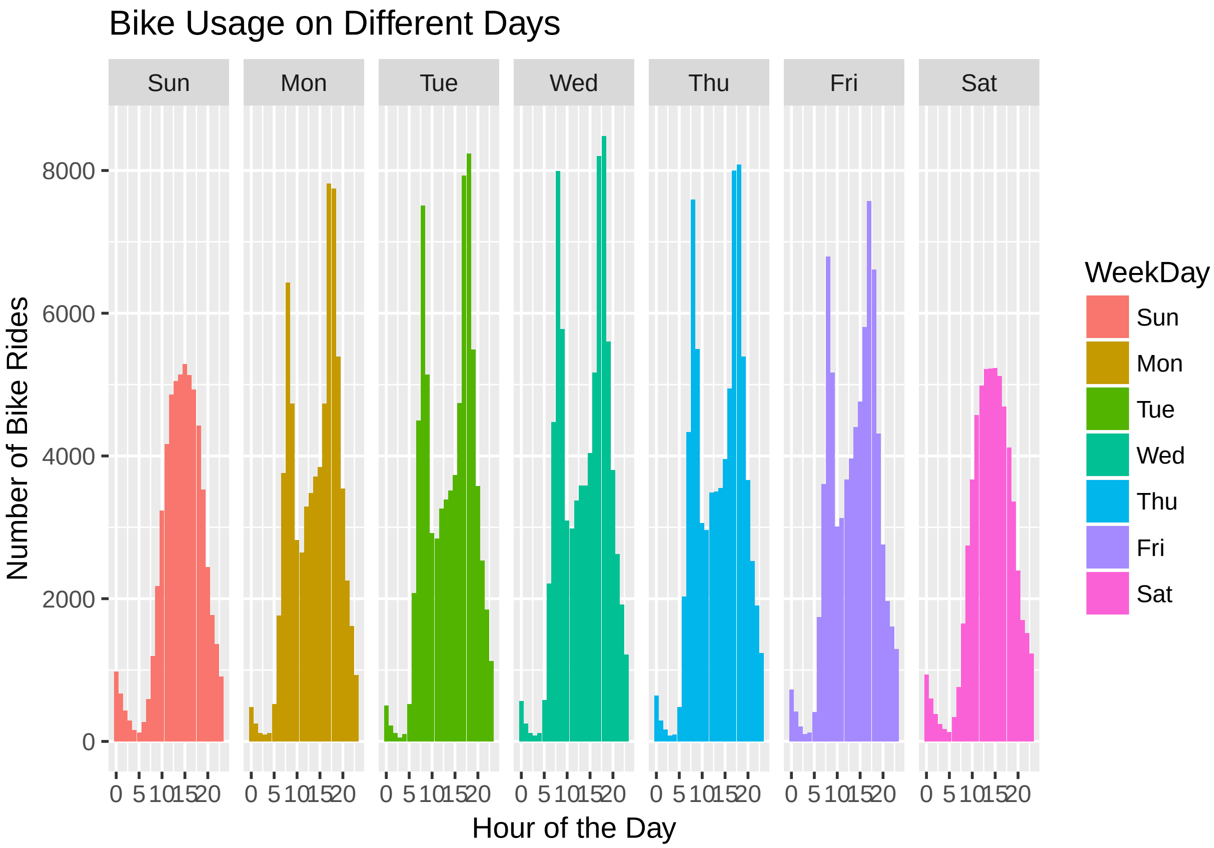
```
# Exploratory Analysis -
combined1 <- combined%>%
  mutate(year=year(starttime),month=month(starttime),monyear=substr(starttime,1,7)) %>%
  group_by(monyear,year)%>%
  summarise(usage=n())

combined1$year <- as.factor(combined1$year)

# Graph to See the Monthly Usage
ggplot(combined1,aes(x=monyear,y=usage,fill=year))+geom_bar(stat="identity")+ theme(axis.text.x = eleme
```



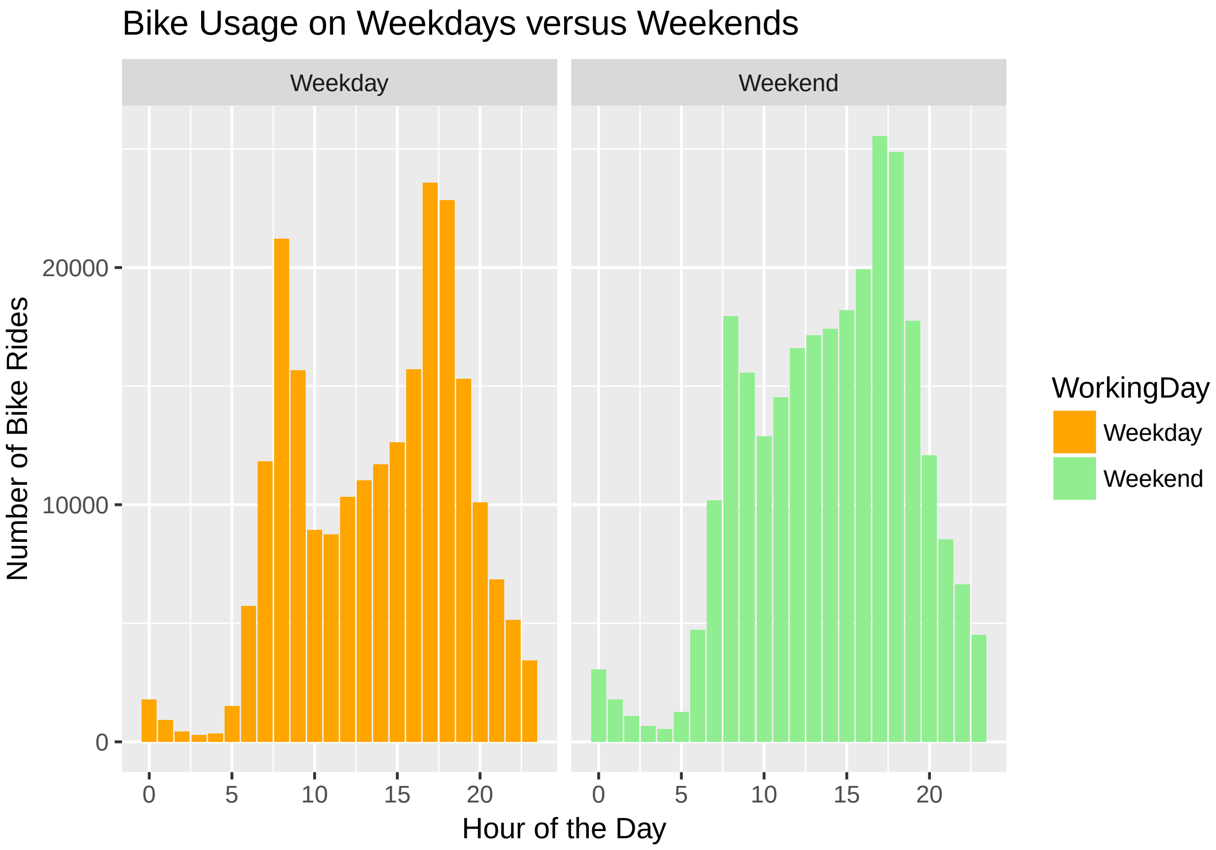
```
# Wrangling the data
combined2 <-  combined%>%
  mutate(year=year(starttime),week=week(starttime),monyear=substr(starttime,1,7)) %>%
  group_by(week,year)%>%
  summarise(usage=n(),MedianDuration=median(tripduration)) %>%
  mutate(wy=paste(week,"-",year)) %>%
  filter(year==2016) %>%
  arrange(year)

# Graph to See the Usage in 2016
ggplot(combined2,aes(x=week,y=usage,fill=MedianDuration))+geom_bar(stat="identity")+theme(axis.text.x =
```



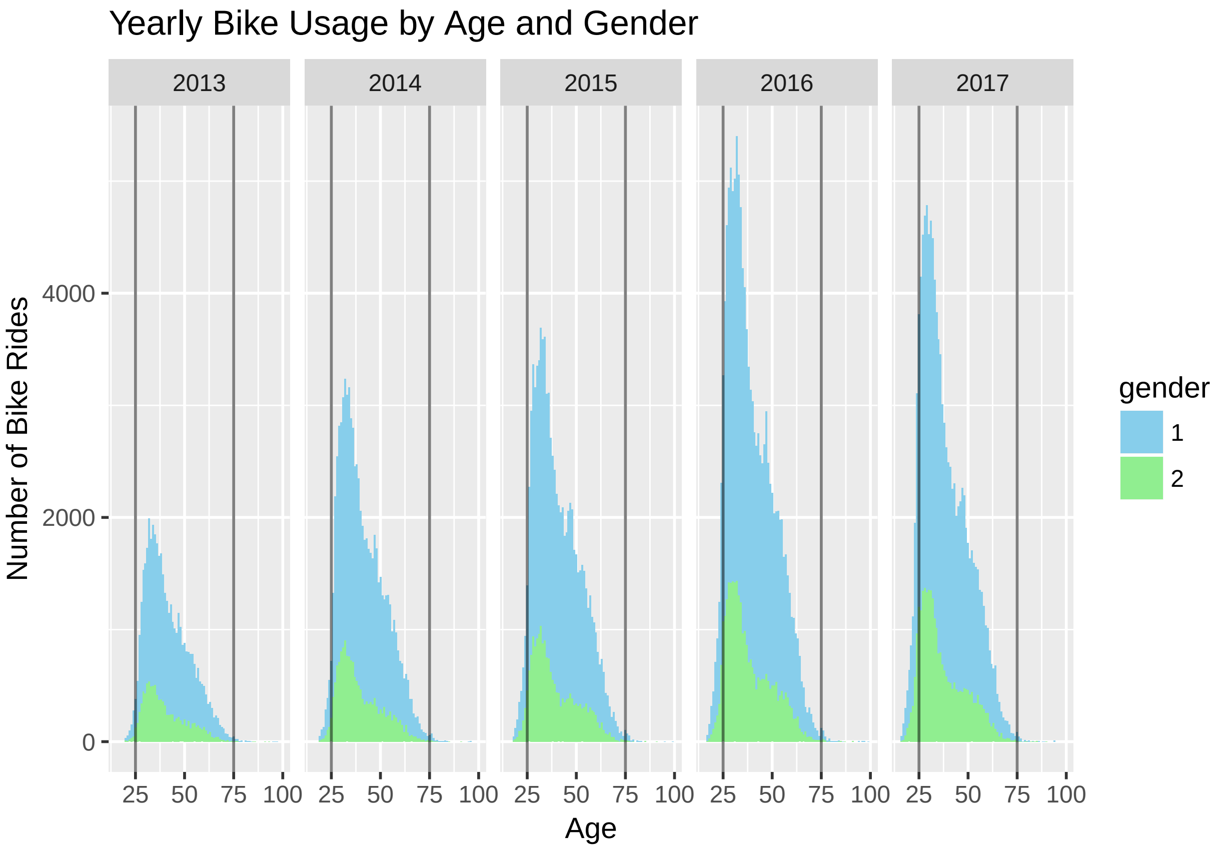
```
combined4 <- combined%>%
  mutate(year=year(starttime),month=month(starttime),hour=hour(starttime)) %>%
  filter(!is.na(usertype))%>%
  group_by(hour,gender,usertype,WeekDay)%>%
  summarise(usage=n(),MedianDuration=median(tripduration))
# The Number of Rides for every hour - Will make binary variables for TSNE Clustering then.

ggplot(combined4,aes(x=hour,y=usage,fill=WeekDay))+geom_bar(stat="identity")+facet_grid(~WeekDay)+xlab(
```



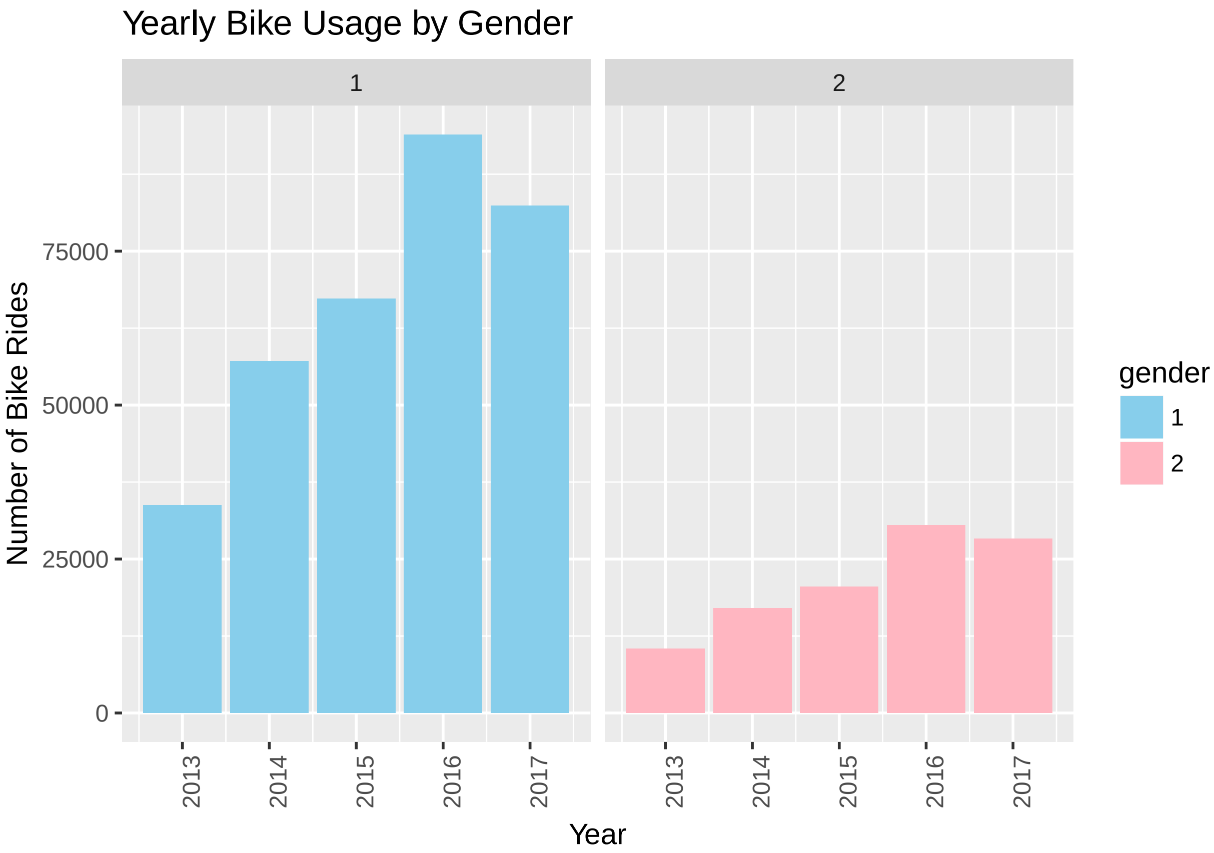
```
combined7 <- combined%>%
  mutate(year=year(starttime),month=month(starttime),hour=hour(starttime)) %>%
  filter(!is.na(usertype))%>%
  group_by(hour,gender,usertype,WorkingDay)%>%
  summarise(usage=n(),MedianDuration=median(tripduration))
# The Number of Rides for every hour - Will make binary variables for TSNE Clustering then.

ggplot(combined7,aes(x=hour,y=usage,fill=WorkingDay))+geom_bar(stat="identity")+facet_grid(~WorkingDay)
```



```
# Making Different Levels for the Time of the Day
combined4 <- combined4 %>%
  mutate(EarlyMorning=ifelse(hour=="3"|hour=="4"|hour=="5"|hour=="6",1,0)) %>%
  mutate(Commuting=ifelse((hour=="7"|hour=="8"|hour=="9"),1,0)) %>%
  mutate(DayTime=ifelse((hour=="10"|hour=="11"|hour=="12"|hour=="13"|hour=="14"|hour=="15"),1,0)) %>%
  mutate(Evening=ifelse((hour=="16"|hour=="17"|hour=="18"|hour=="19"|hour=="20"),1,0)) %>%
  mutate(Night=ifelse((hour=="21"|hour=="22"|hour=="23"|hour=="0"|hour=="1"|hour=="2"),1,0))



#
combined5 <- combined%>%
    filter(!is.na(birth.year)) %>%
  mutate(year=year(starttime),month=month(starttime),monyear=substr(starttime,1,7), age = 2017-birth.yea
  group_by(monyear,year,age,usertype,gender)%>%
  summarise(usage=n())%>%
  filter(age<100,gender!=0)
combined5$gender <- as.factor(combined5$gender)

ggplot(combined5,aes(x=age,y=usage,fill=gender))+geom_bar(stat="identity")+facet_grid(~year)+ggtitle("Y
```



```
ggplot(combined5,aes(x=year,y=usage,fill=gender))+geom_bar(stat="identity")+theme(axis.text.x = element_
```



```
# run Rtsne with default parameters - tsne
```

Rationale for Choosing the weather data variables:

The variables relating to the weather data chosen are air temperature in Fahrenheit(tmpf), Dew Point Temperature in Fahrenheit (dwpf), Relative Humidity in percentage (relh), Wind Speed in knots (sknt), Pressure altimeter in inches (alti), and visibility in miles (vsby). We chose these variables because each of these variables can have an impact on a biker's decision to choose or not to choose to bike and can significantly affect bike usage. We chose these variables to build a preliminary model and make predictions based off the model. The other variables like wind direction in degrees from north, sky level coverages, sky level altitudes, and wind gust had some missing observations. We believe that these variables don't affect bike usage much and not including them would not affect the model.

Exploratory Data Analysis

Based on a preliminary exploratory analysis the following insights were found which could be useful for Citi Bikes:

1) It was found that bike usage for both females and males has increased each year since the initiation of this bike sharing project in May 2013 and since the data became available in July 2013. After the missing observations for gender were removed, it was found that bikes are used predominantly by males. We believe this trend is unexpected and needs to be corrected. We posit that female prefer walking, which is why the bike usage for females is not very high as compared to bike usage for males. The 2017 data is only available till September, which is why when compared to 2016 it shows a decline.

2) It was found that bike usage has increased the most for the 25-50 years age group. The bike usage for the age group of 50-75 has increased significantly too, while the usage for the age group less than 25 increased too. The bike usage for the age group of 75 years or older has not increased much. Based on this information, we believe that the target market for Citi Bikes is 25 to 50 year olds.

3) In the year 2016 bike usage peaked from week 21 to week 44. This corresponds to the dates from May 16 to October 31. This is because bike usage is generally higher during the summer and fall months. It was also found that the median bike ride is longer during these months as the weather does not act as impediment for bikers. The highest median bike ride was in week 36 (12 minutes and 9 seconds), while the lowest was in week 7 (8 minutes and 8 seconds). Hence weather does play a part in biker's decision whether to bike or not. However, it remains to be seen which of the six weather predictors play the most important role in a user's decision to bike or not.

4) Based on the first graph, it can be seen that the overall popularity of this bike sharing system has surged significantly each year. It has increased each year since its inception. The number of bike rides continue to grow, but the number of bike rides in the last 4 months of the years are higher in 2013 than in 2014, albeit not by a big amount. This may have been because of the weather conditions. The average temperature in September 2013 was 67.2 degree Farenheit, 2 degrees higher than the average temperature in September 2014, which could have translated in the marginally lower number of bike rides in in September 2014 than in September 2013. Similarly, the average temperature in December 2013 was 2 degrees lower than the average temperature in December 2014. We hypothesize that weather conditions could be the reason for this slight decrease in bike rentals. However, it is still not quite clear about why the number of bike rides rented in October and November 2013 were marginally lower than in October and November 2014.

5) The bike usage is significantly higher during the weekdays than during weekends. This could be because the bikes are predominantly used by commuters commuting to work on weekdays. This drop is the highest during the commuting hours from 7am to 9am and from 4pm to 7pm, which makes us believe that commuters account for a majority of the bike rides in New York City. Based on this information, Citi Bikes could shift its major focus to commuters instead of any other group. This could be undertaking or promoting campaigns to bike to work, which would increase their annual memberships.

It is important to note that the exploratory analysis is based on a sample of 500,000 observations, which is 1% of the total population. This sample is representative of the population because of its large size.