New York Bike Sharing Dataset

Jaymeen Gandhi

11/15/2019

## R Markdown

#Our project is based on the dataset of the new york bikes, this business has gained a lot of popularity in past few years as it decreases the traffic and is environment friendly.

require(tidyverse)

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

require(dplyr)  
  
require(tm)

## Loading required package: tm

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

bikes= read.csv("D:/Into data mining, visualization/nyc/nyc.csv")  
bikes %>% head(10)

## X Trip.Duration Date Start.Time Stop.Time Start.Station.ID  
## 1 0 376 10/1/2015 12:16:26 AM 12:22:42 AM 3212  
## 2 1 739 10/1/2015 12:27:12 AM 12:39:32 AM 3207  
## 3 2 2714 10/1/2015 12:32:46 AM 1:18:01 AM 3193  
## 4 3 275 10/1/2015 12:34:31 AM 12:39:06 AM 3199  
## 5 4 561 10/1/2015 12:40:12 AM 12:49:33 AM 3183  
## 6 5 365 10/1/2015 12:41:46 AM 12:47:51 AM 3198  
## 7 6 139 10/1/2015 12:43:44 AM 12:46:03 AM 3206  
## 8 7 1299 10/1/2015 1:10:10 AM 1:31:50 AM 3197  
## 9 8 647 10/1/2015 2:01:36 AM 2:12:24 AM 3213  
## 10 9 233 10/1/2015 4:43:33 AM 4:47:27 AM 3194  
## Start.Station.Name Start.Station.Latitude Start.Station.Longitude  
## 1 Christ Hospital 40.73479 -74.05044  
## 2 Oakland Ave 40.73760 -74.05248  
## 3 Lincoln Park 40.72461 -74.07841  
## 4 Newport Pkwy 40.72874 -74.03211  
## 5 Exchange Place 40.71625 -74.03346  
## 6 Heights Elevator 40.74872 -74.04044  
## 7 Hilltop 40.73117 -74.05757  
## 8 North St 40.75256 -74.04472  
## 9 Van Vorst Park 40.71849 -74.04773  
## 10 McGinley Square 40.72534 -74.06762  
## End.Station.ID End.Station.Name End.Station.Latitude  
## 1 3207 Oakland Ave 40.73760  
## 2 3212 Christ Hospital 40.73479  
## 3 3193 Lincoln Park 40.72461  
## 4 3187 Warren St 40.72112  
## 5 3192 Liberty Light Rail 40.71124  
## 6 3215 Central Ave 40.74673  
## 7 3195 Sip Ave 40.73074  
## 8 3215 Central Ave 40.74673  
## 9 3190 Garfield Ave Station 40.71047  
## 10 3195 Sip Ave 40.73074  
## End.Station.Longitude Bike.ID User.Type Birth.Year Gender  
## 1 -74.05248 24470 Subscriber 1960 1  
## 2 -74.05044 24481 Subscriber 1960 1  
## 3 -74.07841 24628 Subscriber 1983 1  
## 4 -74.03805 24613 Subscriber 1975 1  
## 5 -74.05570 24668 Customer 1984 0  
## 6 -74.04925 24644 Customer 1984 0  
## 7 -74.06378 24482 Subscriber 1988 1  
## 8 -74.04925 24550 Customer 1984 0  
## 9 -74.07004 24650 Subscriber 1988 1  
## 10 -74.06378 24584 Subscriber 1978 2  
## Trip\_Duration\_in\_min  
## 1 6  
## 2 12  
## 3 45  
## 4 5  
## 5 9  
## 6 6  
## 7 2  
## 8 22  
## 9 11  
## 10 4

#The data used is New york bike sharing, sourced from Kaggle consisting of 735502 rows of information containing of various attributes of collected daily from 2015 through 2017. Our dataset comprises of 17 variables out of which we will be focusing on a few important variables, namely time duration, age and user type.The variables are explained below:  
  
#Trip.Duration= The entire trip duration is given in seconds.  
#Date= The date when the cycle was issued for transportation.  
#Start.Time= The Time at which the journey began.  
#Stop.Time= The time at which the journey finished.  
#Start.Station.ID=The Station at which the journey began.  
#Start.Station.Name= The name of the station.  
#Start.Station.Latitude=The co-ordinates of the station where the journey started.  
#Start.Station.Longitude=The co-ordinates of the station where the journey started.  
#End.Station.ID=The Station at which the journey ended.  
#End.Station.Name= The name of the station where the journey ended.  
#End.Station.Latitude=The co-ordinates of the station where the journey ended.  
#End.Station.Longitude=The co-ordinates of the station where the journey ended.  
#Bike.ID=The unqiue identifier that can be used to identify the bike.  
#User.Type=The type of user can beidentified using this variable.  
#Birth.Year= The user's age can be determined through this variable.  
#Trip\_Duration\_in\_min=The duration is given in minutes.

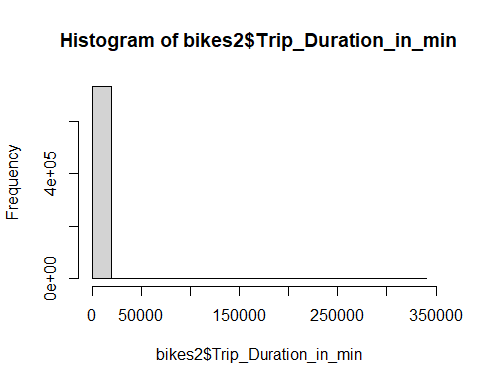
##install.packages("tm")  
#view(bikes)

##Lets create a new variable age   
bikes2=bikes %>%mutate(age = 2019 - Birth.Year)

summary(bikes2)

## X Trip.Duration Date Start.Time   
## Min. : 0 Min. : 61 Length:735502 Length:735502   
## 1st Qu.: 4484 1st Qu.: 246 Class :character Class :character   
## Median : 9156 Median : 383 Mode :character Mode :character   
## Mean :10582 Mean : 934   
## 3rd Qu.:15003 3rd Qu.: 652   
## Max. :34148 Max. :20260211   
## Stop.Time Start.Station.ID Start.Station.Name Start.Station.Latitude  
## Length:735502 Min. :3183 Length:735502 Min. :40.69   
## Class :character 1st Qu.:3186 Class :character 1st Qu.:40.72   
## Mode :character Median :3202 Mode :character Median :40.72   
## Mean :3207 Mean :40.72   
## 3rd Qu.:3211 3rd Qu.:40.73   
## Max. :3426 Max. :40.75   
## Start.Station.Longitude End.Station.ID End.Station.Name End.Station.Latitude  
## Min. :-74.10 Min. : 147 Length:735502 Min. : 0.00   
## 1st Qu.:-74.05 1st Qu.:3186 Class :character 1st Qu.:40.72   
## Median :-74.04 Median :3199 Mode :character Median :40.72   
## Mean :-74.05 Mean :3203 Mean :40.72   
## 3rd Qu.:-74.04 3rd Qu.:3211 3rd Qu.:40.73   
## Max. :-74.03 Max. :3442 Max. :40.80   
## End.Station.Longitude Bike.ID User.Type Birth.Year   
## Min. :-74.10 Min. :14552 Length:735502 Min. :1900   
## 1st Qu.:-74.05 1st Qu.:24486 Class :character 1st Qu.:1975   
## Median :-74.04 Median :24602 Mode :character Median :1982   
## Mean :-74.05 Mean :24914 Mean :1980   
## 3rd Qu.:-74.04 3rd Qu.:24711 3rd Qu.:1986   
## Max. : 0.00 Max. :29296 Max. :2000   
## Gender Trip\_Duration\_in\_min age   
## Min. :0.000 Min. : 1.0 Min. : 19.00   
## 1st Qu.:1.000 1st Qu.: 4.0 1st Qu.: 33.00   
## Median :1.000 Median : 6.0 Median : 37.00   
## Mean :1.124 Mean : 15.6 Mean : 39.46   
## 3rd Qu.:1.000 3rd Qu.: 11.0 3rd Qu.: 44.00   
## Max. :2.000 Max. :337670.0 Max. :119.00

#as per the summary we can observe that there are some outliers present in the data, but there are no missing values and hence we have to eliminate the outliers to make sure that it does not affect the prediction.  
  
hist(bikes2$Trip\_Duration\_in\_min)



#On reviewing the data, we find that the least usage of citybike is 1 min on 10/01/2015 to 10/05/2015 andm highest usage of citybike is 337670 on 05/17/2016. The total usage of citybike that have been mined so far is 11452258 as of the last record of the dataset which is 03/31/2017.A summary of the key statistics of the data is given below:

summary(bikes)

## X Trip.Duration Date Start.Time   
## Min. : 0 Min. : 61 Length:735502 Length:735502   
## 1st Qu.: 4484 1st Qu.: 246 Class :character Class :character   
## Median : 9156 Median : 383 Mode :character Mode :character   
## Mean :10582 Mean : 934   
## 3rd Qu.:15003 3rd Qu.: 652   
## Max. :34148 Max. :20260211   
## Stop.Time Start.Station.ID Start.Station.Name Start.Station.Latitude  
## Length:735502 Min. :3183 Length:735502 Min. :40.69   
## Class :character 1st Qu.:3186 Class :character 1st Qu.:40.72   
## Mode :character Median :3202 Mode :character Median :40.72   
## Mean :3207 Mean :40.72   
## 3rd Qu.:3211 3rd Qu.:40.73   
## Max. :3426 Max. :40.75   
## Start.Station.Longitude End.Station.ID End.Station.Name End.Station.Latitude  
## Min. :-74.10 Min. : 147 Length:735502 Min. : 0.00   
## 1st Qu.:-74.05 1st Qu.:3186 Class :character 1st Qu.:40.72   
## Median :-74.04 Median :3199 Mode :character Median :40.72   
## Mean :-74.05 Mean :3203 Mean :40.72   
## 3rd Qu.:-74.04 3rd Qu.:3211 3rd Qu.:40.73   
## Max. :-74.03 Max. :3442 Max. :40.80   
## End.Station.Longitude Bike.ID User.Type Birth.Year   
## Min. :-74.10 Min. :14552 Length:735502 Min. :1900   
## 1st Qu.:-74.05 1st Qu.:24486 Class :character 1st Qu.:1975   
## Median :-74.04 Median :24602 Mode :character Median :1982   
## Mean :-74.05 Mean :24914 Mean :1980   
## 3rd Qu.:-74.04 3rd Qu.:24711 3rd Qu.:1986   
## Max. : 0.00 Max. :29296 Max. :2000   
## Gender Trip\_Duration\_in\_min  
## Min. :0.000 Min. : 1.0   
## 1st Qu.:1.000 1st Qu.: 4.0   
## Median :1.000 Median : 6.0   
## Mean :1.124 Mean : 15.6   
## 3rd Qu.:1.000 3rd Qu.: 11.0   
## Max. :2.000 Max. :337670.0

#top 10 data showing least usage of citybike

bikes%>%group\_by(`Start.Time`)%>%arrange(Trip\_Duration\_in\_min)%>%head(10)

## # A tibble: 10 x 18  
## # Groups: Start.Time [10]  
## X Trip.Duration Date Start.Time Stop.Time Start.Station.ID  
## <int> <int> <chr> <chr> <chr> <int>  
## 1 48 88 10/1~ 7:33:05 AM 7:34:33 ~ 3211  
## 2 452 84 10/1~ 8:08:35 PM 8:10:00 ~ 3183  
## 3 797 73 10/3~ 12:54:15 ~ 12:55:29~ 3185  
## 4 974 80 10/3~ 11:20:24 ~ 11:21:44~ 3209  
## 5 1063 88 10/4~ 10:24:17 ~ 10:25:46~ 3213  
## 6 1085 82 10/4~ 10:52:04 ~ 10:53:27~ 3199  
## 7 1181 80 10/4~ 12:53:06 ~ 12:54:27~ 3185  
## 8 1186 62 10/4~ 1:05:27 PM 1:06:30 ~ 3214  
## 9 1722 85 10/5~ 8:25:19 AM 8:26:45 ~ 3199  
## 10 2010 86 10/5~ 5:31:49 PM 5:33:15 ~ 3184  
## # ... with 12 more variables: Start.Station.Name <chr>,  
## # Start.Station.Latitude <dbl>, Start.Station.Longitude <dbl>,  
## # End.Station.ID <int>, End.Station.Name <chr>, End.Station.Latitude <dbl>,  
## # End.Station.Longitude <dbl>, Bike.ID <int>, User.Type <chr>,  
## # Birth.Year <int>, Gender <int>, Trip\_Duration\_in\_min <int>

#top 10 data showing highest usage of citybike

bikes%>%group\_by(`Start.Time`)%>%arrange(desc(Trip\_Duration\_in\_min))%>%head(10)

## # A tibble: 10 x 18  
## # Groups: Start.Time [6]  
## X Trip.Duration Date Start.Time Stop.Time Start.Station.ID  
## <int> <int> <chr> <chr> <chr> <int>  
## 1 2869 20260211 9/26~ 4:20:59 AM 4:11:10 ~ 3197  
## 2 8903 16329808 3/22~ 7:02:10 AM 7:05:38 ~ 3215  
## 3 8903 16329808 3/22~ 7:02:10 AM 7:05:38 ~ 3215  
## 4 5237 6065936 12/1~ 9:04:35 PM 2:03:32 ~ 3190  
## 5 5237 6065936 12/1~ 9:04:35 PM 2:03:32 ~ 3190  
## 6 13908 5366099 11/2~ 1:49:07 PM 4:24:07 ~ 3192  
## 7 13908 5366099 11/2~ 1:49:07 PM 4:24:07 ~ 3192  
## 8 18510 4826890 11/2~ 5:38:36 PM 2:26:46 ~ 3280  
## 9 18510 4826890 11/2~ 5:38:36 PM 2:26:46 ~ 3280  
## 10 3233 2104123 2/12~ 7:27:56 AM 3:56:40 ~ 3214  
## # ... with 12 more variables: Start.Station.Name <chr>,  
## # Start.Station.Latitude <dbl>, Start.Station.Longitude <dbl>,  
## # End.Station.ID <int>, End.Station.Name <chr>, End.Station.Latitude <dbl>,  
## # End.Station.Longitude <dbl>, Bike.ID <int>, User.Type <chr>,  
## # Birth.Year <int>, Gender <int>, Trip\_Duration\_in\_min <int>

#total usage of citybikes in minutes from 10/1/2015 to 03/31/2017

sum(bikes$Trip\_Duration\_in\_min)

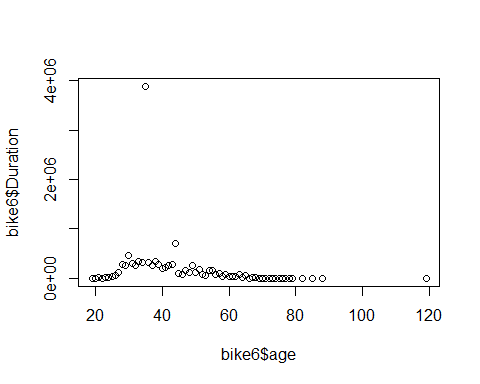
## [1] 11452258

##How does the age affect the trip duration?

bike6=bikes2%>%group\_by(age)%>% summarise(sum(Trip\_Duration\_in\_min))  
colnames(bike6) <- c("age","Duration")  
bike6 %>% arrange(desc(Duration)) %>% head(10)

## # A tibble: 10 x 2  
## age Duration  
## <dbl> <int>  
## 1 35 3885562  
## 2 44 709562  
## 3 30 457577  
## 4 38 348796  
## 5 33 331954  
## 6 36 328710  
## 7 34 323070  
## 8 31 299726  
## 9 39 288213  
## 10 43 283838

plot(bike6$age,bike6$Duration)



##This will allow the company to target the customers according to the age groups.

##Where are more parking station required?

bike2\_test<-bikes2%>%group\_by(`Start.Station.Name`)%>%count('Start Station Name') %>% arrange(desc(n))%>% head(10)  
bike2\_test<-bike2\_test[,-2]  
bike2\_test

## # A tibble: 10 x 2  
## # Groups: Start.Station.Name [10]  
## Start.Station.Name n  
## <chr> <int>  
## 1 Grove St PATH 86155  
## 2 Exchange Place 56282  
## 3 Sip Ave 49545  
## 4 Hamilton Park 47610  
## 5 Newport PATH 40024  
## 6 Brunswick St 29337  
## 7 Essex Light Rail 28459  
## 8 Newark Ave 28245  
## 9 Van Vorst Park 26882  
## 10 Newport Pkwy 26154

bikes2\_t=bikes2%>%group\_by(`End.Station.Name`)%>%count('End Station Name') %>% arrange(desc(n))%>% head(10)  
bikes2\_t<-bikes2\_t[,-2]  
bikes2\_t

## # A tibble: 10 x 2  
## # Groups: End.Station.Name [10]  
## End.Station.Name n  
## <chr> <int>  
## 1 Grove St PATH 114148  
## 2 Exchange Place 65154  
## 3 Hamilton Park 46802  
## 4 Sip Ave 46543  
## 5 Newport PATH 39789  
## 6 Essex Light Rail 30433  
## 7 Newport Pkwy 26272  
## 8 Newark Ave 25490  
## 9 Brunswick St 25484  
## 10 Warren St 25473

##So we can interpret that Grove Street and Exchange Place are the busiest places for the cycles.So we need to install more bike parking station in these two places.

##Which Gender has more cycle usage?

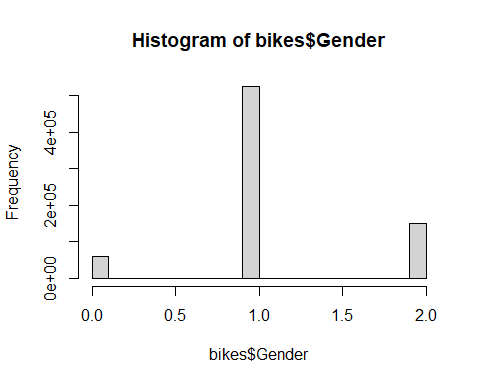
bikes%>%filter(Gender=='1')%>%summarise(mean\_male\_duration=mean(Trip\_Duration\_in\_min))

## mean\_male\_duration  
## 1 11.68941

bikes%>%filter(Gender=='2')%>%summarise(mean\_female\_duration=mean(Trip\_Duration\_in\_min))

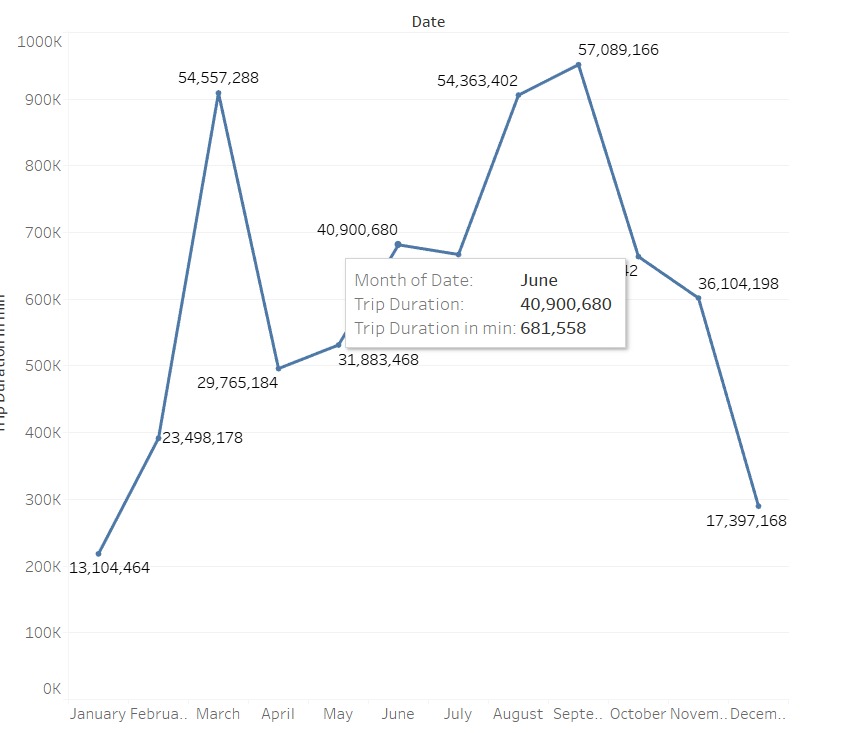
## mean\_female\_duration  
## 1 11.73786

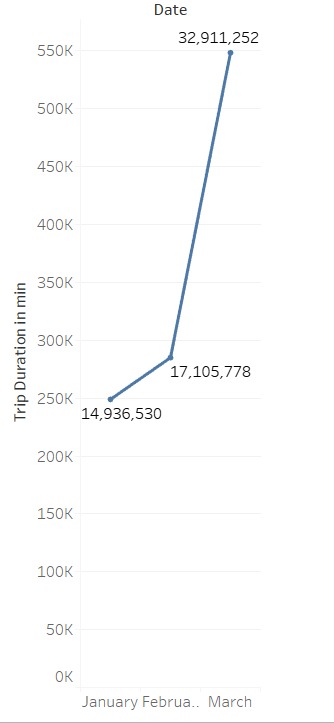
hist(bikes$Gender)

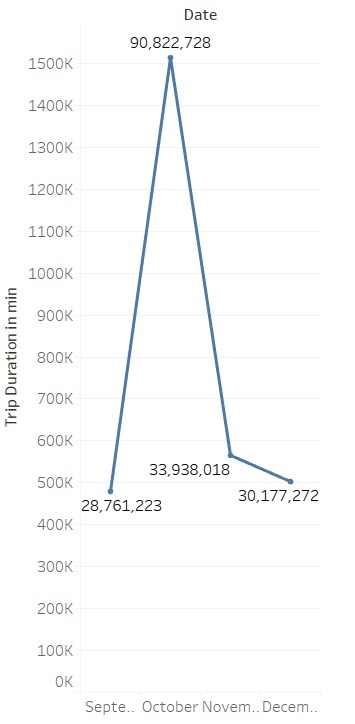


## Here 1 represents the male population and 2 represents the female population. So we can interpret that the bike company should target female consumers more than the male consumers.

##Which year had the most cycle usage?







##So we can observe that the highest useage of cycles was in the year 2016.

##What are the chances that the suscriber will turn up into a customer?

bikes %>% filter(`User.Type`=='Subscriber')%>%summarise(mean(Trip\_Duration\_in\_min))

## mean(Trip\_Duration\_in\_min)  
## 1 11.70939

bikes %>% filter(`User.Type`=='Customer')%>%summarise(mean(Trip\_Duration\_in\_min))

## mean(Trip\_Duration\_in\_min)  
## 1 71.67265

##So we can interpret that a suscriber who will travel more than 12 mins in a particular period will turn up into a customer.

# This report helps to understand the dataset how various demographics and geographic locations affect the usage of cycles throughout the year. Establishing various types of relationships between attributes. At the end achiving the scope about which type of customers should be targetted by the company.