Twitter Analysis

Part-A:

1.

Before decompressing the file:

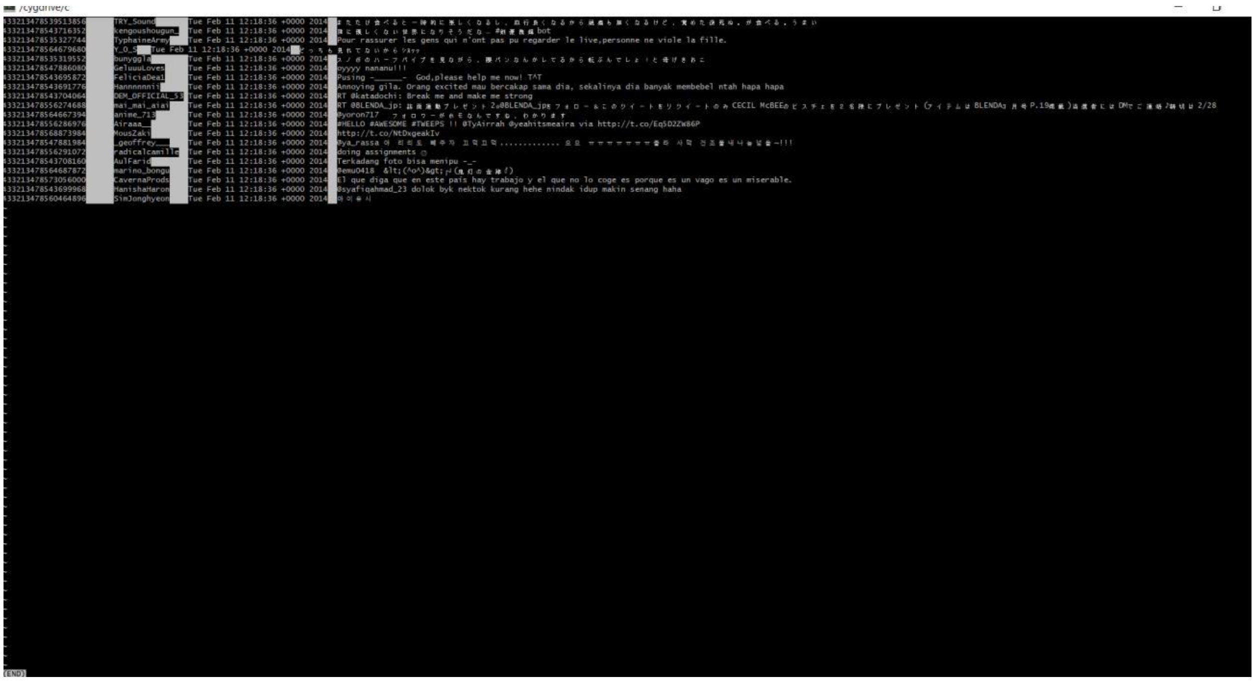
```
The file is 992Mb long.

$ ls -lh Twitter_Data_1.gz
-rwxrwx---+ 1 Administrators None 992M Oct 12 20:51 Twitter_Data_1.gz
```

After decompressing, the file got 2.2Gb long as it can be seen below in the code.

```
$ ls -lh Twitter_Data_1
-rwxr-x--- 1 Utkarsh None 2.2G Oct 18 12:54 Twitter_Data_1
```

2.

As in the picture below, tab is used to separate each column so delimiter is tab which indicates that there are four columns in total.



```
433213478539513856	TRY_Sound	Tue Feb 11 12:18:36 +0000 2014	またたび食べると一時的に楽しくなるし、血行良くなるから頭痛も無くなるけど、覚めた後死ぬ。が食べる。うまい
433213478543716352	kengoushougun_	Tue Feb 11 12:18:36 +0000 2014	[illegible]に優しくない世界になりそうだな… #[illegible] bot
433213478535327744	TyphaineArmy	Tue Feb 11 12:18:36 +0000 2014	Pour rassurer les gens qui n'ont pas pu regarder le live,personne ne viole la fille.
433213478564679680	Y_O_S	Tue Feb 11 12:18:36 +0000 2014	どっちも見れてないから ｼｽｯｯ
433213478535319552	bunyggla	Tue Feb 11 12:18:36 +0000 2014	スノボのハーフパイプを見ながら、腰パンなんかしてるから転ぶんでしょ！と毒づきおこ
433213478547886080	GeluuuLoves	Tue Feb 11 12:18:36 +0000 2014	oyyyy nananu!!!
433213478543695872	FeliciaDeal	Tue Feb 11 12:18:36 +0000 2014	Pusing -______-  God,please help me now! T^T
433213478543691776	Hannnnnnnii	Tue Feb 11 12:18:36 +0000 2014	Annoying gila. Orang excited mau bercakap sama dia, sekalinya dia banyak membebel ntah hapa hapa
433213478543704064	DEM_OFFICIAL_53	Tue Feb 11 12:18:36 +0000 2014	RT @katadochi: Break me and make me strong
433213478556274688	mai_mai_aiai	Tue Feb 11 12:18:36 +0000 2014	RT @BLENDA_jp: [illegible]プレゼント2☆@BLENDA_jpをフォロー＆このツイートをリツイートのみ CECIL McBEEのビスチェを2名様にプレゼント(アイテムはBLENDA3月号P.19掲載)当選者にはDMでご連絡♪締切は2/28
433213478564667394	anime_713	Tue Feb 11 12:18:36 +0000 2014	@yoron717  フォロワーが[illegible]なんですね、わかります
433213478556286976	Airaaa__	Tue Feb 11 12:18:36 +0000 2014	#HELLO #AWESOME #TWEEPS !! @TyAirrah @yeahitsmeaira via http://t.co/Eq5D2ZW86P
433213478568873984	MousZaki	Tue Feb 11 12:18:36 +0000 2014	http://t.co/NtDxgeakIv
433213478547881984	_geoffrey___	Tue Feb 11 12:18:36 +0000 2014	@ya_rassa [illegible]............. ㅇㅇ ㅜㅜㅜㅜㅜㅜㅜ[illegible]-!!!
433213478556291072	radicalcamille	Tue Feb 11 12:18:36 +0000 2014	doing assignments ☺
433213478543708160	AulFarid	Tue Feb 11 12:18:36 +0000 2014	Terkadang foto bisa menipu -_-
433213478564687872	marino_bongu	Tue Feb 11 12:18:36 +0000 2014	@emu0418  &lt;(^o^)&gt;↵([illegible])
433213478573056000	CavernaProds	Tue Feb 11 12:18:36 +0000 2014	El que diga que en este pais hay trabajo y el que no lo coge es porque es un vago es un miserable.
433213478543699968	HanishaHaron	Tue Feb 11 12:18:36 +0000 2014	@syafiqahmad_23 dolok byk nektok kurang hehe nindak idup makin senang haha
433213478560464896	SimJonghyeon	Tue Feb 11 12:18:36 +0000 2014	[illegible]
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
(END)
```

```
Utkarsh@DESKTOP-SIIQBHG ~
$ wc -l Twitter_Data_1
15089920 Twitter_Data_1

Utkarsh@DESKTOP-SIIQBHG ~
$ wc Twitter_Data_1
```

```
  15089920  267552778 2271087104 Twitter_Data_1

Utkarsh@DESKTOP-SIIQBHG ~
$ wc -L Twitter_Data_1
629 Twitter_Data_1
```

```
Answer(ii)
Utkarsh@DESKTOP-SIIQBHG ~
$ awk -F ' ' '{print NF; exit}' Twitter_Data_1
9

Answer is 4.

Utkarsh@DESKTOP-SIIQBHG ~
$ awk '{ if (NF > max) max = NF } END { print max }' Twitter_Data_1 78
```

3.

As in the file, Column-1 has all the unique values to identify each user which make it Unique identifier.

While Column-2 shows containing all the Users name so column-2 would be usernames.

Column-3 indicates the Time and date of the tweet in full format. So, it has the calendar containing dates and times when the tweet is made.

Column-4 has the actual tweets, so all the tweets are shown by this column.

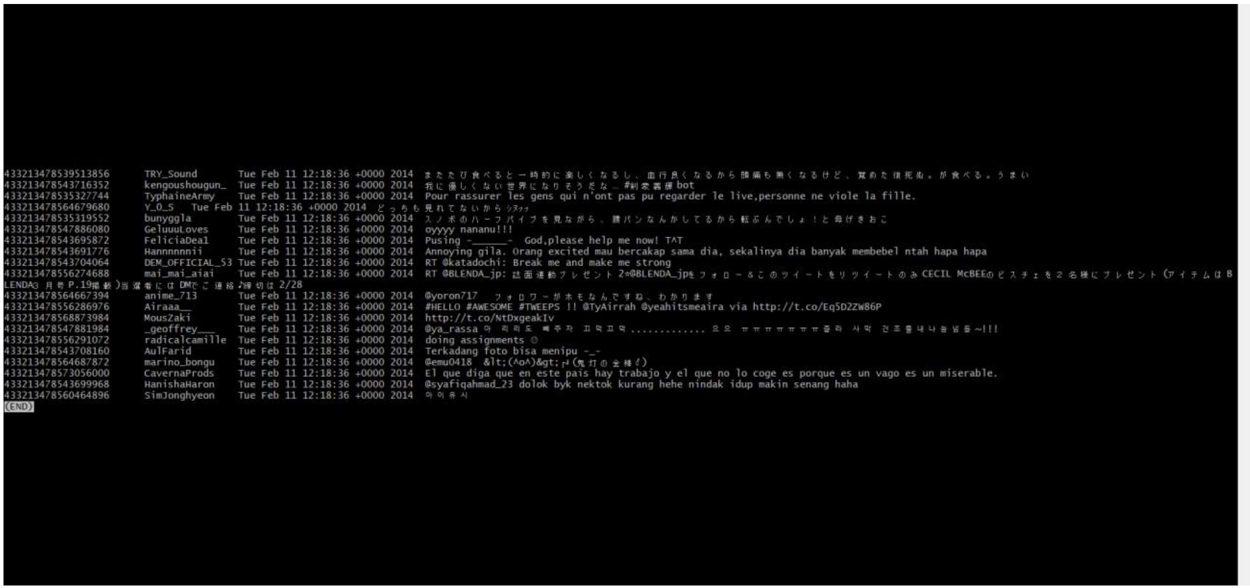
4.

There are 15089920 tweets in the entire file.

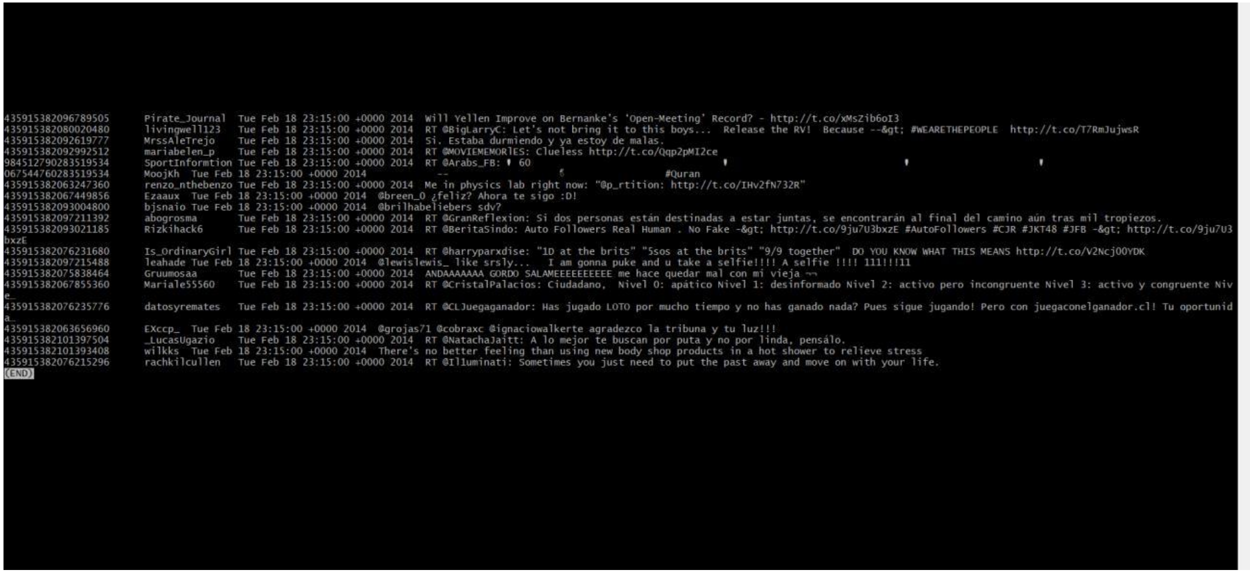
```
Utkarsh@DESKTOP-SIIQBHG /cygdrive/c/cygwin64/home/Utkarsh
$ wc -l Twitter_Data_1
15089920 Twitter_Data_1
```

5.

As the picture below is the header and tail of the file which has time and date information too. Looking at that image, it seems that the date range is 11th February 2014 to 18th February 2014.



```
433213478539513856      TRY_Sound       Tue Feb 11 12:18:36 +0000 2014  またたび食べると一時的に楽しくなるし、血行良くなるから頭痛も無くなるけど、覚めた後死ぬ。が食べる。うまい
433213478543716352      kengoushougun_  Tue Feb 11 12:18:36 +0000 2014  我に優しくない世界になりそうだな…#剣豪義輝bot
433213478535327744      TyphaineArmy    Tue Feb 11 12:18:36 +0000 2014  Pour rassurer les gens qui n'ont pas pu regarder le live,personne ne viole la fille.
433213478564679680      Y_0_S   Tue Feb 11 12:18:36 +0000 2014  どっちも見れてないからｼﾇｯｯ
433213478535319552      bunyggla        Tue Feb 11 12:18:36 +0000 2014  スノボのハーフパイプを見ながら、腰パンなんかしてるから転ぶんでしょ！と母げきおこ
433213478547886080      GeluuuLoves     Tue Feb 11 12:18:36 +0000 2014  oyyyy nananu!!!
433213478543695872      FeliciaDeal     Tue Feb 11 12:18:36 +0000 2014  Pusing -______-  God,please help me now! T^T
433213478543691776      Hannnnnnii      Tue Feb 11 12:18:36 +0000 2014  Annoying gila. Orang excited mau bercakap sama dia, sekalinya dia banyak membebel ntah hapa hapa
433213478543704064      DEM_OFFICIAL_53 Tue Feb 11 12:18:36 +0000 2014  RT @katadochi: Break me and make me strong
433213478556274688      mai_mai_aiai    Tue Feb 11 12:18:36 +0000 2014  RT @BLENDA_jp: 誌面連動プレゼント 2☆@BLENDA_jpをフォロー＆このツイートをリツイートのみ CECIL McBEEのビスチェを２名様にプレゼント（アイテムはB
LENDA3月号P.19掲載）当選者にはDMでご連絡♪締切は 2/28
433213478564667394      anime_713       Tue Feb 11 12:18:36 +0000 2014  @yoron717  フォロワーがホモなんですね、わかります
433213478556286976      Airaaa__        Tue Feb 11 12:18:36 +0000 2014  #HELLO #AWESOME #TWEEPS !! @TyAirrah @yeahitsmeaira via http://t.co/Eq5D2ZW86P
433213478568873984      MousZaki        Tue Feb 11 12:18:36 +0000 2014  http://t.co/NtDxgeakIv
433213478547881984      _geoffrey___    Tue Feb 11 12:18:36 +0000 2014  @ya_rassa 야 리리도 빼주자 끄덕끄덕............. 으으 ㅠㅠㅠㅠㅠㅠㅠ줄라 사막 건조를내나는넘들~!!!
433213478556291072      radicalcamille  Tue Feb 11 12:18:36 +0000 2014  doing assignments ☺
433213478543708160      AulFarid        Tue Feb 11 12:18:36 +0000 2014  Terkadang foto bisa menipu -_-
433213478564687872      marino_bongu    Tue Feb 11 12:18:36 +0000 2014  @emu0418  &lt;(^o^)&gt;┌┘(鬼灯の金棒♪)
433213478573056000      CavernaProds    Tue Feb 11 12:18:36 +0000 2014  El que diga que en este pais hay trabajo y el que no lo coge es porque es un vago es un miserable.
433213478543699968      HanishaHaron    Tue Feb 11 12:18:36 +0000 2014  @syafiqahmad_23 dolok byk nektok kurang hehe nindak idup makin senang haha
433213478560464896      SimJonghyeon    Tue Feb 11 12:18:36 +0000 2014  아이유시
(END)
```



```
435915382096789505      Pirate_Journal  Tue Feb 18 23:15:00 +0000 2014  Will Yellen Improve on Bernanke's 'Open-Meeting' Record? - http://t.co/xMsZib6oI3
435915382080020480      livingwell123   Tue Feb 18 23:15:00 +0000 2014  RT @BigLarryC: Let's not bring it to this boys...  Release the RV!  Because --&gt; #WEARETHEPEOPLE  http://t.co/T7RmJujwsR
435915382092619777      MrssAleTrejo    Tue Feb 18 23:15:00 +0000 2014  Si. Estaba durmiendo y ya estoy de malas.
435915382092992512      mariabelen_p    Tue Feb 18 23:15:00 +0000 2014  RT @MOVIEMEMORIES: Clueless http://t.co/Qqp2pMI2ce
984512790283519534      SportInformtion Tue Feb 18 23:15:00 +0000 2014  RT @Arabs_FB: ▮ 60
067544760283519534      MoojKh  Tue Feb 18 23:15:00 +0000 2014          --                                   #Quran
435915382063247360      renzo_nthebenzo Tue Feb 18 23:15:00 +0000 2014  Me in physics lab right now: "@p_rtition: http://t.co/IHv2fN732R"
435915382067449856      Ezaaux  Tue Feb 18 23:15:00 +0000 2014  @breen_O ¿feliz? Ahora te sigo :D!
435915382093004800      bjsnaio Tue Feb 18 23:15:00 +0000 2014  @brilhabeliebers sdv?
435915382097211392      abogrosma       Tue Feb 18 23:15:00 +0000 2014  RT @GranReflexion: Si dos personas están destinadas a estar juntas, se encontrarán al final del camino aún tras mil tropiezos.
435915382093021185      Rizkihack6      Tue Feb 18 23:15:00 +0000 2014  RT @BeritaSindo: Auto Followers Real Human . No Fake -&gt; http://t.co/9ju7U3bxzE #AutoFollowers #CJR #JKT48 #JFB -&gt; http://t.co/9ju7U3
bxzE
435915382076231680      Is_OrdinaryGirl Tue Feb 18 23:15:00 +0000 2014  RT @harryparxdise: "1D at the brits" "5sos at the brits" "9/9 together"  DO YOU KNOW WHAT THIS MEANS http://t.co/V2Ncj0OYDK
435915382097215488      leahade Tue Feb 18 23:15:00 +0000 2014  @lewislewis_ like srsly...   I am gonna puke and u take a selfie!!!! A selfie !!!! 111!!!11
435915382075838464      Gruumosaa       Tue Feb 18 23:15:00 +0000 2014  ANDAAAAAAA GORDO SALAMEEEEEEEEEE me hace quedar mal con mi vieja ¬¬
435915382067855360      Mariale55560    Tue Feb 18 23:15:00 +0000 2014  RT @CristalPalacios: Ciudadano,  Nivel 0: apático Nivel 1: desinformado Nivel 2: activo pero incongruente Nivel 3: activo y congruente Niv
e_
435915382076235776      datosyremates   Tue Feb 18 23:15:00 +0000 2014  RT @CLJuegaganador: Has jugado LOTO por mucho tiempo y no has ganado nada? Pues sigue jugando! Pero con juegaconelganador.cl! Tu oportunid
a_
435915382063656960      EXccp_  Tue Feb 18 23:15:00 +0000 2014  @grojas71 @cobraxc @ignaciowalkerte agradezco la tribuna y tu luz!!!
435915382101397504      _LucasUgazio    Tue Feb 18 23:15:00 +0000 2014  RT @NatachaJaitt: A lo mejor te buscan por puta y no por linda, pensálo.
435915382101393408      wilkks  Tue Feb 18 23:15:00 +0000 2014  There's no better feeling than using new body shop products in a hot shower to relieve stress
435915382076215296      rachkilcullen   Tue Feb 18 23:15:00 +0000 2014  RT @Illuminati: Sometimes you just need to put the past away and move on with your life.
(END)
```

6.

There is total 6260301 unique users in the file which contains users appears only at once. This does not count the users who have their name more than once.

But after counting those users appear more than once, there are 8977904 unique users in total.

For total unique user (not multiple times included)

```
Utkarsh@DESKTOP-SIIQBHG ~
$ cut -f 2 Twitter_Data_1 | sort | uniq -u | wc -l
6260301

Utkarsh@DESKTOP-SIIQBHG ~
$ cut -f 2 Twitter_Data_1 | sort | uniq -c | wc -l 8977904
```

7.

The first time, “Donald Trump” was mentioned in tweet in the file was at 12:28:36 +0000 on 11th February 2014. As seen below, the tweet was “`Be interesting to see the detail on this one: BBC News - Donald Trump loses offshore wind farm challenge` http://t.co/qAcG…” from `@aedan_smith.`

```
Utkarsh@DESKTOP-SIIQBHG ~
$ cut -f 1,2,3,4 Twitter_Data_1 | sort | grep "Donald Trump" | head -5
433215995134476289      Maddog4U_1st    Tue Feb 11 12:28:36 +0000 2014  RT
@aedan_smith: Be interesting to see the detail on this one:  BBC News - Donald Trump loses
offshore wind farm challenge http://t.co/qAcG…
433220734702612480      ScottishPleb    Tue Feb 11 12:47:26 +0000 2014  RT
@havantacluOTMP: Donald Trump loses legal challenge to windfarm near his Scottish golf
resort http://t.co/3OQPw7hpA4 via @guardian
433222676652781568      liliannberg     Tue Feb 11 12:55:09 +0000 2014  RT
@TheScotsman: Donald Trump loses Aberdeen Bay wind farm legal challenge:
http://t.co/Cu232VeNnd
433229555319316480      carsinogenic    Tue Feb 11 13:22:29 +0000 2014  RT @BusinessGreen:
Breaking: Donald Trump trumped over offshore wind farm challenge http://t.co/qyI2FL5sRn
433230067037392896      Noord_Holland_  Tue Feb 11 13:24:31 +0000 2014  Donald Trump
vangt bot in windmolenzaak http://t.co/XyWlqei0UW #nholland #noordholland #haarlem
```

8.

No, we have not captured all the references to Donald. Because Donald may be any person’s name. For example: Donald Trump. Some people might want to find “Donald Trump” by different way, maybe ‘Donald’ or ‘Trump’ or without ignoring case. It means we might face problem in case someone finds the name by typing in lower or upper case such as Donald, trump, DONALD, TRUMP etc. In those cases, they cannot find the person Donald whom they originally refer to.

[We might try -i which ignores the case so that it searches for all possible different strings of the same name]

Part-B:

1.

The term ‘Obama’ appears exactly 11999 times in the file.

```
Utkarsh@DESKTOP-SIIQBHG ~
$ grep -o 'Obama' Twitter_Data_1 | wc -l 11999
```

2.

The following command lets you find the timestamps for all the tweets about Obama and store it into the text file.

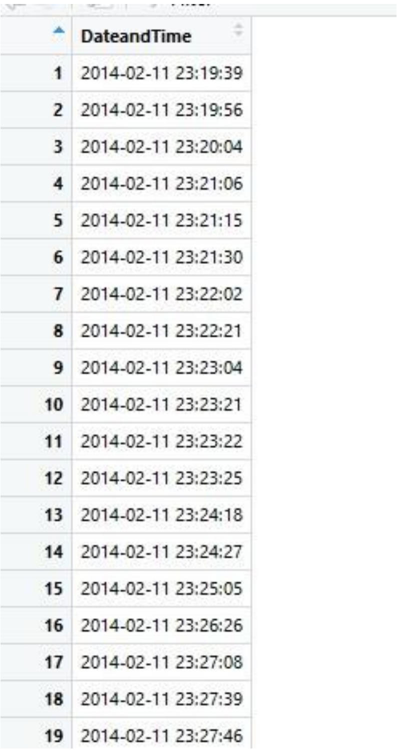
```
Utkarsh@DESKTOP-SIIQBHG /cygdrive/c/cygwin64/home/Utkarsh
$ grep -i 'Obama' Twitter_Data_1 | cut -f 3 > obamatime.txt
```

You may convert the txt file into csv file [plus adding the column name=” DateandTime”] and read it into R using following commands which would lead you to the output in the picture below.

obamafile=read.csv("obamatime.csv”, head=T) View(obamafile) obamafile$DateandTime<-

strptime(obamafile$DateandTime,"%a %b %d %H:%M:%S %z %Y",tz = "")

View(obamafile)



	DateandTime
1	2014-02-11 23:19:39
2	2014-02-11 23:19:56
3	2014-02-11 23:20:04
4	2014-02-11 23:21:06
5	2014-02-11 23:21:15
6	2014-02-11 23:21:30
7	2014-02-11 23:22:02
8	2014-02-11 23:22:21
9	2014-02-11 23:23:04
10	2014-02-11 23:23:21
11	2014-02-11 23:23:22
12	2014-02-11 23:23:25
13	2014-02-11 23:24:18
14	2014-02-11 23:24:27
15	2014-02-11 23:25:05
16	2014-02-11 23:26:26
17	2014-02-11 23:27:08
18	2014-02-11 23:27:39
19	2014-02-11 23:27:46

3. Histogram function:

```
> bins<- c(0,2,4,6,8,10)
> hist(obamafile$DateandTime, col="darkgreen", ylim=c(0,10), main ="HISTOGRAM
of number of discussions about Obama over time", xlab="Time",col = "blue", br
eaks=bins,, las=2, cex.lab = 1.3)
```

4. The pattern has an unusual shape shape before 15th February but after that it can be seen in the histogram in R by the code above that it is really greatly distributed with discussion get more as time grows.

5. Looking at the histogram, I think the highest day having most mentioning of Obama is 16th February 2014 and as in the histogram for every hour, it seems like this day is the busiest day on twitter for

Obama. It can be deduced that this day has the higher amount of data distributed in the graph and it keeps growing as the day progresses. The histogram is made based on the function underneath.

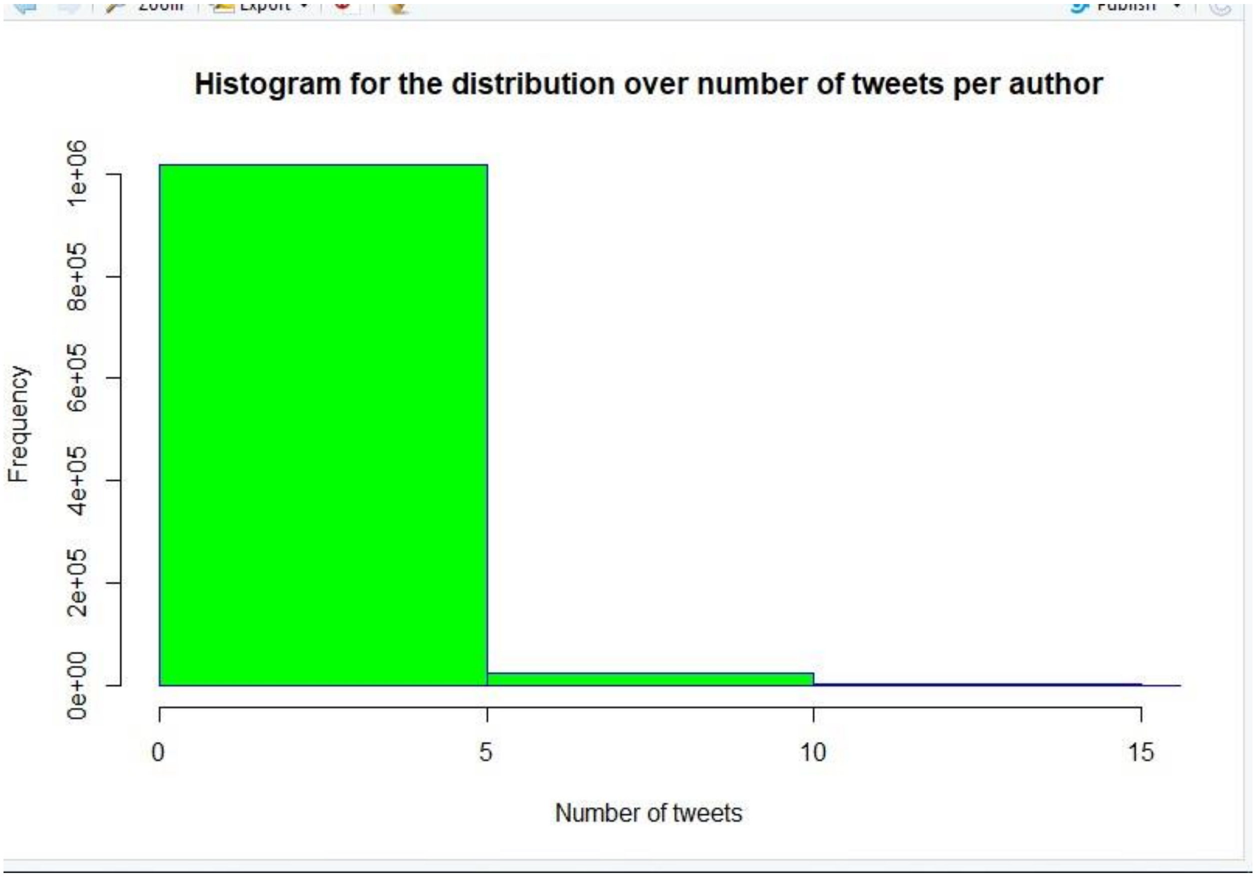
```
> bins<- c(0,2,4,6,8,10)
> hist(obamafile$DateandTime, col="darkgreen", ylim=c(0,30), main ="HISTOGRAM
of number of discussions about Obama over time", xlab="Time[Every hour]",col =
"blue", breaks=bins,, las=4, cex.lab = 0.3)
```

6.

The command to import data into file is as follow.

```
Utkarsh@DESKTOP-SIIQBHG ~
$ cut -f 2 Twitter_Data_1 | sort | uniq -c > QuestionB6.txt
```

```
> hist(tweets,
      main="Histogram for the distribution over number of tweets per author",
+       xlab="Number of tweets",
+       border="blue",
+       col="green",
+       breaks=10
+       xlim=c(0,15))
```



Part-C: [Additional Research Study]

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016.

Note that the Winter and Summer Games were held in the same year up until 1992. After that, they staggered them such that Winter Games occur on a four-year cycle starting with 1994, then Summer in 1996, then Winter in 1998, and so on.

The data is about biodata of athletes participated in Olympic games which is in the file. The columns are given as in the dataset which includes their name, id number, gender, age, height, weight, team they played for, season of the event, host city, name of the sport, year when event happened and the medal type.

The purpose of this file is to find out the clusters of three bronze, silver and gold medals. The clustering will be to see how athlete's height and weight relationships are (ratio of those two shows the physical estimated fitness of an athlete) are throughout all Olympics distributed in three clusters of bronze, silver, and gold medals.

Another analysis is done based on the art competition happened in all the Olympics. The purpose is to show which country has won what kind of and how many medals overall.

Analysis of Art competition medal:

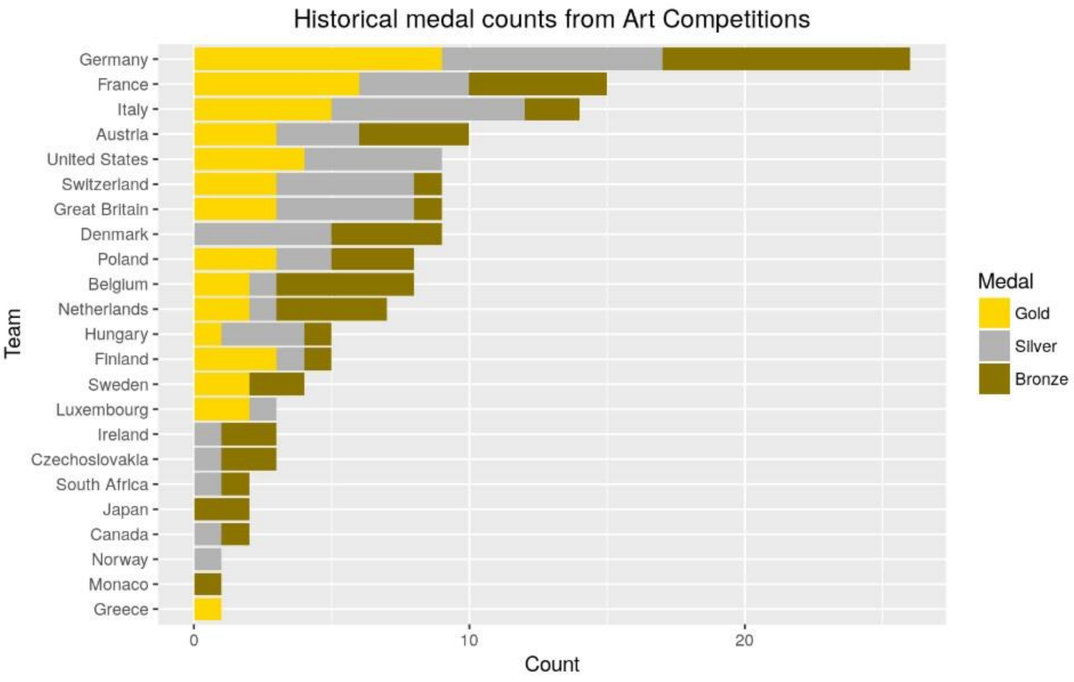
Using clustering to determine who won the most art medal in all the Olympics. There might be some libraries like plotly etc required to use some functions. The plot I used to show the result is ggplot.

The code is following where I got the sport(game) to art using filter followed by removing the empty values. And I compared that game with all of the countries and plotted the graph as follow.

```
medalcounting <- art %>% filter(! is.na(Medal))%>%

group_by(Team,                    Medal)                    %>%

summarize(Count=length(Medal))
```

```
medalart <- medalcounting %>%   group_by (Team) %>%

summarize(Total=sum(Count)) %>%   arrange(Total) %>%   select(Team)

medalcounting $Team <- factor(medalcounting$Team, levels=medalart$Team)
```

```
plotting ggplot(medalcounting, aes(x=Team, y=Count, fill=Medal))

+geom_col() +coord_flip()

+scale_fill_manual(values=c("gold1","gray70","gold4")) + ggtitle("Historical medal counts from Art ")
+  theme(plot.title = element_text(hjust = 0.5))
```



Out of around 50 nations that participated in the Art Competitions, fewer than half won a medal, and over a third of all medals were awarded to artists representing just three countries: Germany, France, and Italy.

It is remarkable that Germany won the most medals in the Art Competitions between 1912 and 1948, considering that Germany was not invited to participate in 3 of the 7 Olympics during this period (they were banned from the 1920, 1924, and 1948 Olympics due to post-war politics). However, Germany made up for these absences with an especially strong showing at the 1936 Berlin Olympics, the Nazi Olympics, in which they won around 40% of the medals in the Art Competitions and about 60% of all the Art Competition medals in the country’s history.

It can be concluded that Germany is the most successful team in Art competition held throughout all Olympics which is why they have won most medals than any other country.

Analysis based on healthy and fit athletes’ ratio of winning medals:

The amount of fat is the critical measurement. A good indicator of how much fat you carry is the body mass index (BMI). BMI is the ratio of height and weight. BMI can indicate whether you are overweight or not which is very important in any sport. As an athlete, it is necessarily to control their weight and BMI too in order to stay fit and active. Here, the plot of height vs weight is generated which would indicate how many medals are won as BMI (the ratio) goes up for athletes. Higher the BMI is, greater the fitness would be. In other words, it is about finding the growth of medals with respect to Athletes’ fitness level.

These three medals are three clusters, and the ratio of height and weight is clustered in three parts to show how much fit athletes by numbers (with higher ratio) can bring more worthy medals. There will be three clusters for Gold, Silver and Bronze medals as last analysis which are in different colours.

Code to read file and data manipulation followed by K-means clustering method to find the cluster regions to plot the graphs in the end.

```
###############################code################################

df <- read.csv("athlete_events.csv")


#View(df)

#str(df)

#summary(df)

#Load data into the dataset

#Assign blank value as NA


df[df == “”] <- NA



#Remove the NA values using na.omit()




df <- na.omit(df)


#View(df)

#str(df)

#summary(df)
```

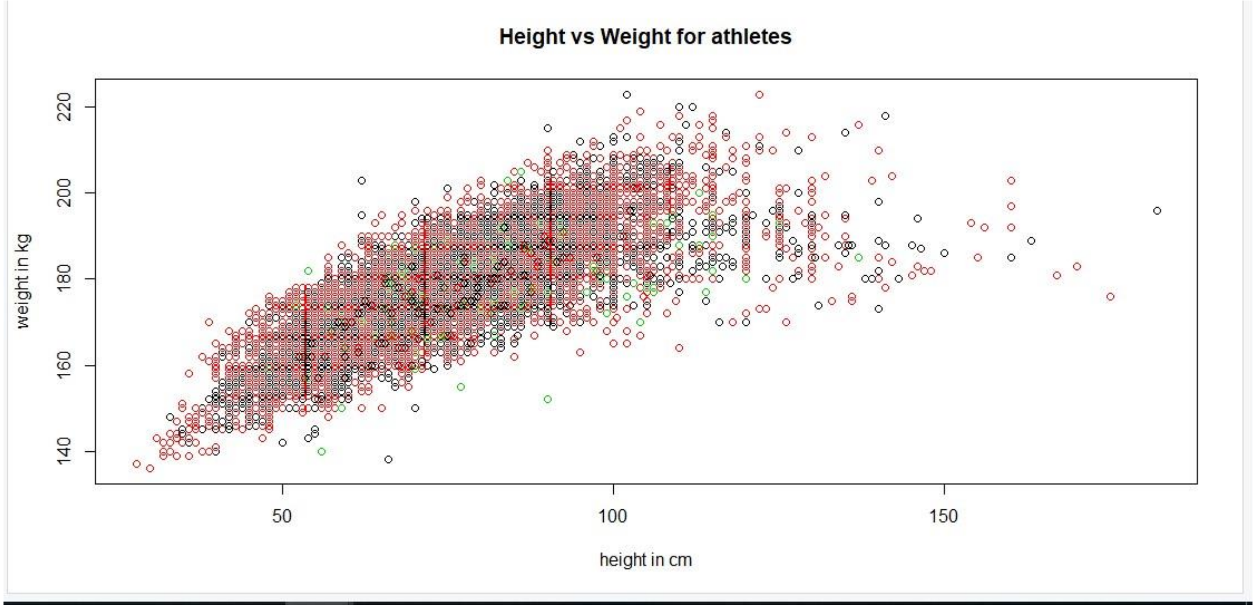
```
> answer=read.csv("athlete_events.csv")
> View(answer)

> answer[answer==""]<-NA
> answer<-na.omit(answer)
> View(answer)

> rownames(answer)<-NULL
> View(answer)

> answer.features=answer
> answer.features$ID<-NULL
> answer.features$Name<-NULL
> answer.features$Sex<-NULL
> answer.features$Team<-NULL
> answer.features$Season<-NULL
> answer.features$City<-NULL
> answer.features$Sport<-NULL
> answer.features$Medal<-NULL
> answer.features$Event<-NULL
> answer.features$Games<-NULL >
View(answer.features)
```

After getting the numerical values in file to generate graph. Here is how it generates the cluster table and k-means clustering method.

```
> ans<-kmeans(answer.features,3)
> ans
K-means clustering with 3 clusters of sizes 12272, 16358, 1551

Cluster means:
       Age   Height   Weight     Year
1 24.69003 177.5059 73.81992 1974.150
2 26.03405 177.7926 73.77387 2004.423
3 24.89491 177.1380 73.01418 1924.489

Clustering vector:
   [1] 2 3 3 3 3 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2
  [32] 2 1 2 2 1 1 2 1 1 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 2 3 1 2 2
  [63] 3 3 2 2 1 3 2 2 2 2 2 1 1 2 2 1 1 2 1 2 3 3 3 2 1 2 2 2 1 2 2
  [94] 2 1 1 1 2 2 1 2 1 2 1 1 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2
 [125] 3 3 1 1 2 2 3 3 2 2 2 2 2 2 1 2 2 2 1 1 2 2 2 2 2 2 1 1 2 2 2
 [156] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 1 2 1 1 2 2 2 2 1 1 1
 [187] 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 1 3 3 1 2 1 2 2 2 3 2 2 2
 [218] 2 1 1 1 2 1 1 1 1 2 2 2 1 1 2 1 1 2 2 2 2 2 2 2 1 1 1 1 2 2 2
 [249] 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 1 2 2 1 2 2 2
 [280] 2 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 1 1 2 1 1 2 2 2 2
 [311] 2 2 2 1 2 2 2 2 2 1 1 1 1 2 2 1 1 2 2 2 2 2 2 2 1 1 1 2 3 1 2
 [342] 1 2 1 3 1 2 1 1 3 1 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 2 2 1 1 1 2
 [373] 2 2 2 2 2 1 2 1 1 2 2 1 1 1 1 2 2 1 2 1 2 2 3 1 2 2 1 2 1 3 1
 [404] 2 2 2 2 2 2 2 2 2 2 2 2 3 2 1 2 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2
 [435] 1 1 1 2 3 2 1 1 2 3 3 3 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2
 [466] 3 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 1 1
 [497] 2 2 2 1 2 1 1 2 2 1 2 2 1 1 2 2 1 1 2 2 2 2 1 2 2 1 2 2 2 1 2
 [528] 1 1 1 2 2 3 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1
 [559] 2 2 1 1 2 1 2 2 1 2 2 2 2 1 2 2 2 2 3 1 1 1 2 2 2 2 2 2 2 1 1
 [590] 1 2 2 2 2 3 1 2 1 1 1 3 1 3 1 1 1 1 1 2 2 2 2 1 2 1 2 1 2 2 3
 [621] 3 2 2 2 2 2 2 2 2 1 1 2 2 2 1 1 1 1 1 2 2 2 1 3 3 2 1 3 3 1 1
 [652] 2 2 1 1 1 1 2 2 1 2 2 2 2 1 1 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1
 [683] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 2 1 2 2 2 1 1 1 1 2 1 1 1 1 1 2
 [714] 2 2 2 1 2 2 1 2 2 1 2 1 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 2 2 1 1
 [745] 2 2 2 2 2 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 2 2 2 1 2 1 3 1 1 1
 [776] 1 3 3 2 2 2 2 1 1 1 1 1 1 1 1 2 2 2 1 2 2 2 2 2 2 2 1 1 1 1 2
 [807] 2 2 3 3 2 2 2 1 2 2 1 2 2 2 1 2 2 1 2 3 1 2 1 2 2 1 1 1 1 2 1
 [838] 1 1 1 3 2 2 3 1 2 1 1 1 2 2 1 1 2 2 1 2 2 2 2 2 1 2 2 1 1 1 1
 [869] 1 1 1 2 1 2 2 2 1 1 1 2 2 2 2 2 1 2 2 1 2 2 2 1 1 2 1 1 2 2 2
 [900] 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 1 1 1 1 1 2 1 1 1 2 1 1 2 2 2 2
 [931] 2 2 1 1 1 1 1 1 2 2 2 1 1 2 1 2 1 1 1 1 2 2 2 1 2 1 1 2 3 1 2
 [962] 1 2 2 1 1 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 1 2
 [993] 1 1 1 2 3 3 2 1
 [ reached getOption("max.print") -- omitted 29181 entries ]


Within cluster sum of squares by cluster:
[1] 5472632.4 7542927.2  690077.6
 (between_SS / total_SS =  48.8 %)


Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

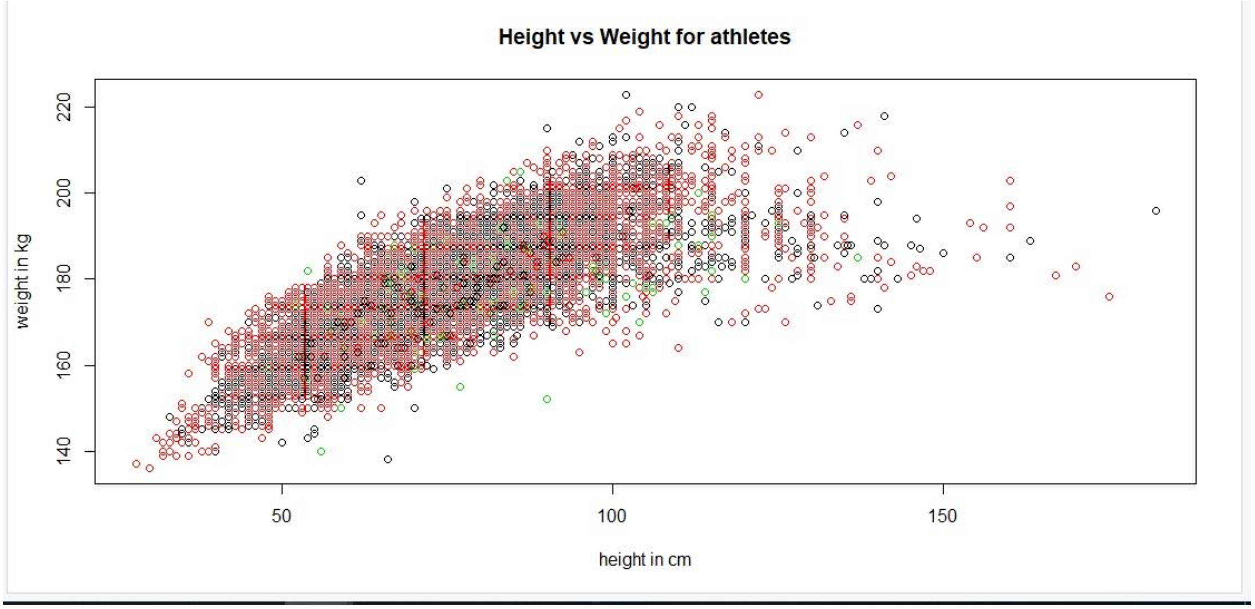
After having the clusters, we are plotting the clusters in a graph for Athletes’ height And weight as in the graph below.

```
plot(answer[c("Height","Weight")], col=ans$cluster)
```



The cluster graph must be same for the one for medal and it is as we can see in the graph below.

plot(answer[c("Height","Weight")], col=anser$medal)



However, as we can see in the graph that three of the clusters seems to be overlapping in most of the cases which makes it unclear to clarify three medals in the graph. It is actually obvious too. Because when you are plotting athlete's height vs weight data means kind of athlete to see how many and what kind of medal they won, it would absolutely make it messy. That is, we will have to find the same thing for specific place which would make it easier to see.
In the next section, we will plot the same thing but only for one place – 'LONDON'. We Will see what medals lies whereas in height vs weight graph ratio for Athletes.

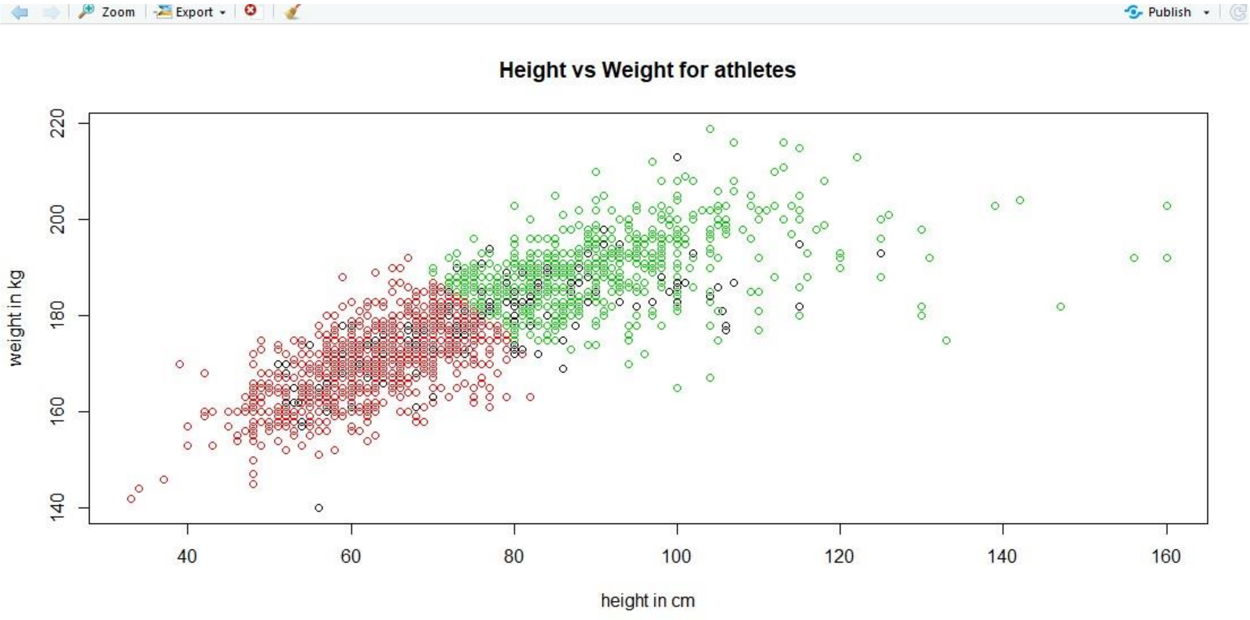
```
> dataset=read.csv("athlete_events.csv")
> View(dataset)
> dataset[dataset==""]<-NA
> dataset<-na.omit(dataset)
> rownames(dataset)<-NULL
> View(dataset)
```

```
> Londondata=subset(dataset,City=="London")
> View(Londondata)
```

```
> Londondata.features=Londondata
> Londondata.features$ID<-NULL
> Londondata.features$Name<-NULL
> Londondata.features$Sex<-NULL
> Londondata.features$Team<-NULL
> Londondata.features$Season<-NULL
> Londondata.features$Medal<-NULL
> Londondata.features$City<-NULL
> Londondata.features$Sport<-NULL
> Londondata.features$Event<-NULL
> Londondata.features$Games<-NULL
> View(Londondata.features)

> rownames(Londondata)<-NULL
> View(Londondata)
> rownames(Londondata.features)<-NULL
> View(Londondata.features)

> Londonclus<-kmeans(Londondata.features,3)
>
> Londonclus
K-means clustering with 3 clusters of sizes 316, 1138, 777

Cluster means:
       Age   Height   Weight     Year
1 25.65823 178.5095 74.38449 1931.038
2 25.56678 170.9306 63.30404 2012.000
3 26.88288 189.2780 89.33012 2012.000

Clustering vector:
   [1] 1 1 1 1 3 2 2 2 3 3 2 2 3 3 2 2 2 3 3 3 2 2 3 3 3 2 1 1 3 3 3 3 3 3 2 2 2 3 3 1 2
  [72] 2 2 3 2 3 2 1 2 3 2 2 2 3 2 3 2 2 2 3 2 2 2 1 1 2 2 2 2 2 2 2 1 1 2 2 2 3 2 2 2 3
 [143] 3 3 3 1 2 2 3 1 1 1 3 2 2 1 2 2 2 2 2 2 1 1 2 2 2 3 1 3 3 3 2 3 2 2 2 2 3 1 1 3 2
 [214] 3 3 2 2 1 2 3 2 2 1 1 1 2 2 2 2 1 1 3 2 3 2 2 2 2 3 3 2 2 3 3 3 3 3 2 3 2 2 2 1 2
 [285] 2 3 3 2 2 2 2 2 2 1 2 3 2 1 1 3 3 3 2 3 2 2 3 2 2 3 3 2 3 3 1 2 3 2 2 3 2 2 2 2 2
 [356] 1 1 1 1 3 2 3 1 1 3 2 1 1 2 3 2 2 2 2 1 3 2 1 1 2 2 2 2 2 2 3 3 3 2 3 3 3 3 3 2 3
 [427] 3 2 3 1 1 3 3 2 1 3 2 2 3 2 2 3 1 2 2 1 3 2 2 2 2 2 2 1 1 2 3 2 2 2 2 2 2 3 1 1 3
 [498] 3 2 2 2 1 2 2 3 2 3 3 2 3 3 2 3 3 3 2 3 1 2 2 2 3 1 1 1 3 3 3 2 2 2 3 1 2 3 3 2 2
 [569] 2 2 2 3 3 3 2 2 2 3 2 2 2 2 2 2 2 3 3 2 3 3 2 2 2 3 3 2 1 3 2 3 2 2 1 3 1 3 3 1 3
 [640] 3 3 3 3 2 2 3 3 2 1 1 3 2 2 3 3 1 1 1 3 3 3 2 3 3 2 2 2 1 2 2 1 1 3 2 3 3 2 3 2 3
 [711] 3 2 3 2 2 3 2 3 3 3 3 1 3 3 3 2 3 2 3 3 2 1 3 1 2 3 1 2 2 3 1 2 1 2 2 2 2 2 2 2 3
 [782] 1 3 1 2 2 2 2 3 2 2 3 3 1 2 3 1 2 2 2 2 2 3 2 2 2 3 1 2 2 2 3 2 2 3 2 2 1 3 3 2 2
 [853] 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 3 2 2 2 2 3 3 3 3 3 3 2 2 2 2 2 2 2 1 2 2 2 2 2 2 3
 [924] 3 2 3 2 2 2 3 3 2 2 2 2 3 2 2 2 2 3 1 1 1 2 3 3 2 3 3 3 3 2 1 2 1 2 2 2 2 2 3 2 2
[995] 3 2 1 2 3 2
 [ reached getOption("max.print") -- omitted 1231 entries ]

Within cluster sum of squares by cluster:
[1] 206495.6 173893.7 184069.9
 (between_SS / total_SS =  79.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
"betweenss > table(Londondata$Medal, Londonclus$cluster )


          1   2   3
  Bronze  93 404 265
  Gold   131 360 262
  Silver  92 374 250




> Londdata<- subset(Londondata,Year>2000)
> View(Londdata)
> rownames(Londdata)<-NULL
> View(Londdata)
> Londdata.features=Londdata
> Londdata.features$ID<-NULL
> Londdata.features$Name<-NULL
> Londdata.features$Sex<-NULL
> Londdata.features$Team<-NULL
> Londdata.features$Season<-NULL
> Londdata.features$City<-NULL
> Londdata.features$Sport<-NULL
> Londdata.features$Medal<-NULL
> Londdata.features$Event<-NULL
> Londdata.features$Games<-NULL

> Londclus<-kmeans(Londdata.features,3)
>
> Londclus
K-means clustering with 3 clusters of sizes 407, 716, 792

Cluster means:
       Age   Height   Weight Year
1 27.43243 193.4816 97.80713 2012
2 25.15782 166.8506 58.72207 2012
3 26.26894 181.0303 75.24874 2012

Clustering vector:
   [1] 3 2 2 2 1 1 3 2 1 1 3 3 3 1 1 1 2 2 1 1 1 2 1 3 1 3 3 1 2 3 2 1 3 2 1 1 2 3 2 1 1
  [72] 3 2 3 2 2 3 3 2 3 2 2 2 3 3 3 3 3 2 3 3 1 3 3 2 3 3 1 3 3 3 3 1 1 3 3 2 2 2 2 3 2
 [143] 1 3 1 2 3 3 2 1 3 2 1 1 3 3 3 3 3 3 3 2 3 3 2 3 3 2 1 1 3 1 1 1 1 3 1 2 2 3 1 3 3
 [214] 1 3 3 1 3 1 1 3 1 2 3 3 2 3 2 1 1 2 1 3 1 1 3 2 2 3 3 2 3 2 2 2 3 2 1 3 3 1 3 2 1
 [285] 2 3 2 2 2 3 3 1 3 3 3 3 3 1 3 3 3 3 2 3 1 2 3 1 2 2 2 2 3 3 2 3 3 3 3 3 3 3 1 2 3
 [356] 3 3 3 1 3 3 3 3 3 3 3 2 3 2 3 1 3 2 3 3 2 2 3 3 2 2 1 3 3 2 2 2 2 3 1 3 2 2 2 2 3
 [427] 1 2 3 1 3 3 3 2 3 2 1 3 1 3 3 3 3 3 2 3 3 3 2 3 3 3 3 3 3 2 3 2 3 3 2 2 1 2 3 3 1
 [498] 2 1 1 2 3 3 1 2 2 3 3 1 1 2 3 3 3 1 3 3 3 3 3 2 2 2 2 1 3 3 3 3 2 2 2 3 2 2 3 1 3
 [569] 1 1 1 1 1 3 3 2 1 2 3 2 2 1 2 3 1 3 1 1 3 3 2 1 3 3 3 3 3 3 2 3 3 3 3 1 1 1 1 1 1
 [640] 2 3 1 3 3 3 3 3 3 1 1 3 3 3 2 2 3 2 1 2 3 3 3 3 1 2 2 2 2 2 3 2 3 3 3 2 3 2 3 2 3
 [711] 1 1 3 3 2 2 2 2 2 3 3 3 3 2 3 3 3 1 3 2 2 2 1 3 3 1 1 1 2 2 3 2 3 3 3 3 2 2 2 2 2
 [782] 2 3 1 2 2 2 2 3 3 3 2 2 1 2 1 3 3 1 1 1 1 2 3 2 3 3 3 3 3 3 2 2 2 2 2 2 1 3 3 3 2
 [853] 1 3 3 3 3 2 2 3 1 3 3 3 3 3 2 3 2 3 2 3 3 2 2 3 2 3 2 3 1 2 2 2 3 2 2 2 3 2 3 3 3
 [924] 2 3 3 2 2 2 2 3 3 2 2 1 3 1 3 3 3 2 1 1 3 3 3 3 3 2 2 1 3 2 2 3 3 3 3 3 3 1 2 3 2
[995] 3 1 3 3 2 3
 [ reached getOption("max.print") -- omitted 915 entries ]

Within cluster sum of squares by cluster:
[1] 85008.61 74377.87 73626.99
 (between_SS / total_SS =  71.8 %)






Available components:


[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
"betweenss
> table(Londdata$Medal, Londclus$cluster)


          1   2   3
  Bronze 134 251 284
  Gold   143 219 260
  Silver 130 246 248
```

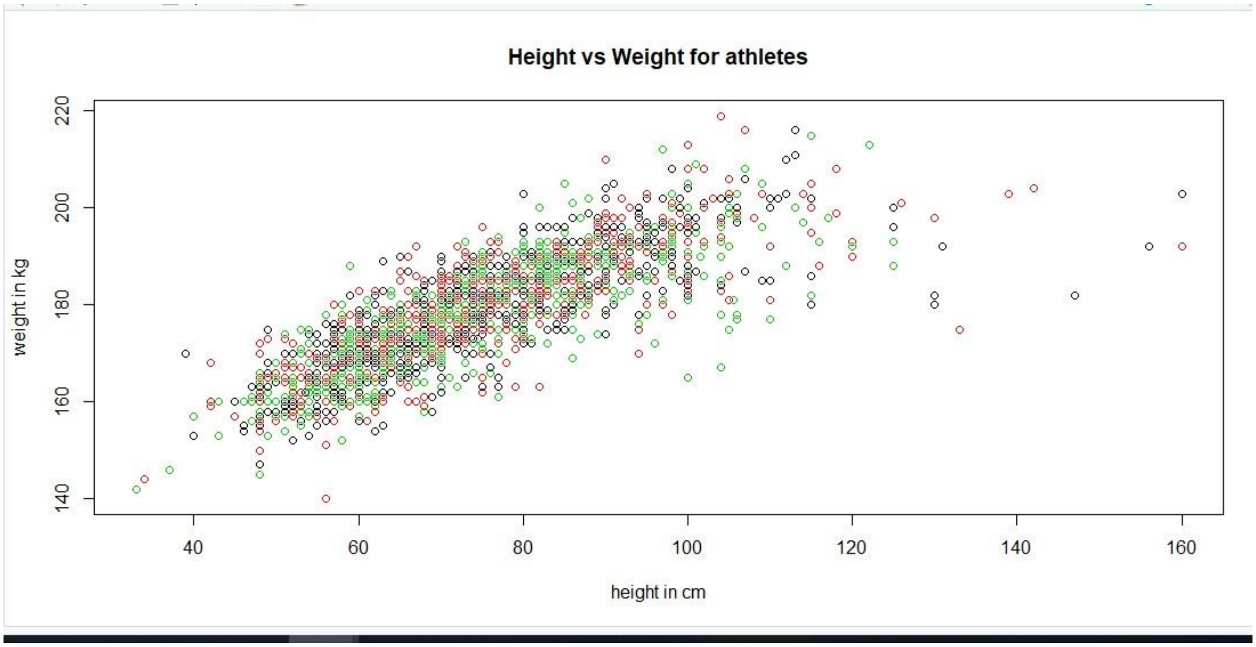
Plotting:

```
plot(Londondata[c("Height","Weight")], col=Londonclus$cluster)
```



[Bronze medals – red colour, Gold medals – Light yellow and Silver Medal-Violent blue color]

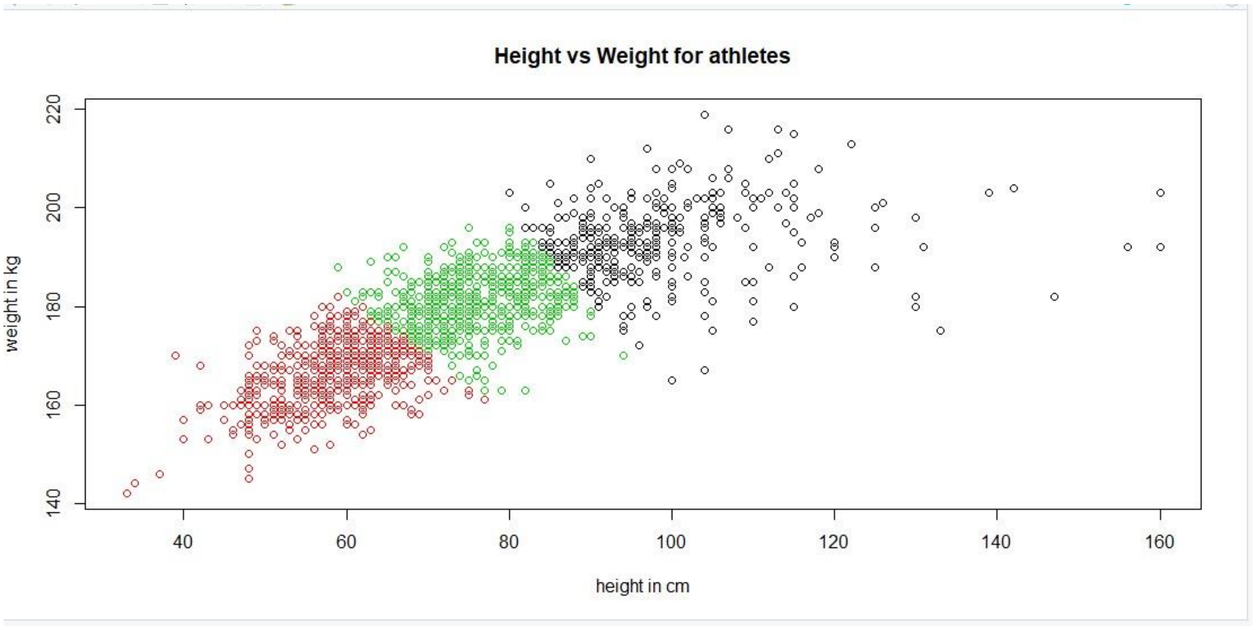
```
plot(Londondata[c("Height","Weight")], col=Londondata$Medal)
```



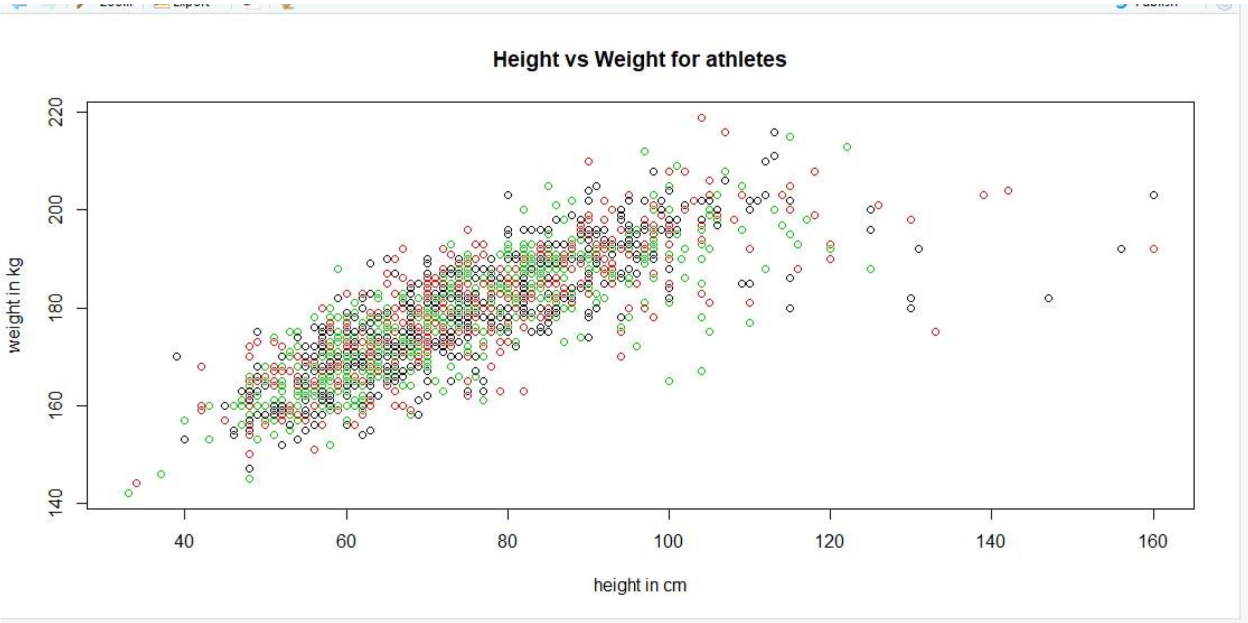
Now, it can be seen more clearly than the one in the last time's graph. We can see that whenever the ratio of height and weight gets higher than the number of Gold medals increases while Silver and Bronze significantly goes down. There are some outliers too. Some of them are just way too high which are carrying gold medal. It looks like Athletes having higher ratio of heights and weights are more successful and manage to win more gold medals.

However, we will plot the data now only for year after 2000 and only for place 'London'. It would give us more accurate results which be only for year 2012 as there is only one Olympic in London after year 2000. Hence, the next manipulation would be for London Olympic, 2012.

plot (Londdata[c("Weight","Height")], col=Londclus$cluster, xlab="height in cm", ylab="weight in kg")



```
plot (Londdata[c("Weight","Height")], col=Londdata$Medal, xlab="height in cm", ylab="weight in kg")
```



As it seems in both of the graphs above, it is pretty clear that success of winning Gold medals or worthy medals do not completely depend on the ratio of height and weight. There can be a bit less healthy and fit athletes win gold medals. But in most cases, it relies highly on the ration of height and weight. Most of the outliers have got gold medals which indicates Athletes won Gold based on their higher ratio of height and weight, but it is not true for all of the cases.

The link to the dataset is as follow.
https://drive.google.com/file/d/1mfCcgJyxxsczicl3iO5SDoHZjcQTMKVG/view?usp=sharing