EDA

Aakanksha Baid

1/13/2020

## 5c i. Read `Dating-Men.csv

Men\_DF = read.csv("Data/Dating-Men.csv")

## 5c ii. Saving data in R data format

save(Men\_DF, file = "Dating-Men.RData")

## 5c iii. Examine the first and last few lines

head(Men\_DF)

emails rating height bmi yrs\_education age days\_active  
1 1 0.6125247 71.5 29.56526 14 3 8  
2 3 -0.1009420 67.5 27.00137 18 1 74  
3 9 0.2975212 69.5 25.46970 16 3 30  
4 5 -0.0284386 69.5 26.92511 21 1 67  
5 19 0.4072408 69.5 24.01428 21 3 61  
6 0 -0.3510217 75.5 25.28222 14 2 70

tail(Men\_DF)

emails rating height bmi yrs\_education age days\_active  
1502 6 -0.4222424 71.5 26.81500 12.5 2 108  
1503 4 -0.2282498 67.5 27.00137 16.0 2 106  
1504 0 -0.1405917 67.5 22.37256 12.0 3 108  
1505 2 -0.2613251 75.5 24.04894 17.0 2 2  
1506 18 -0.1723964 69.5 25.46970 16.0 3 108  
1507 2 -0.7940676 71.5 29.56526 14.0 1 108

## 5c iv. Summary Statistics

library(psych)  
summary(Men\_DF)

emails rating height bmi   
 Min. : 0.000 Min. :-1.93967 Min. :61.00 Min. :12.37   
 1st Qu.: 1.000 1st Qu.:-0.45517 1st Qu.:69.50 1st Qu.:24.01   
 Median : 4.000 Median :-0.12909 Median :71.50 Median :25.44   
 Mean : 7.695 Mean :-0.09149 Mean :70.93 Mean :25.76   
 3rd Qu.:10.000 3rd Qu.: 0.23631 3rd Qu.:73.50 3rd Qu.:27.00   
 Max. :88.000 Max. : 2.52981 Max. :85.00 Max. :36.44   
 yrs\_education age days\_active   
 Min. : 8.00 Min. :1.000 Min. : 1.00   
 1st Qu.:14.00 1st Qu.:1.000 1st Qu.: 50.00   
 Median :16.00 Median :2.000 Median : 85.00   
 Mean :15.66 Mean :1.946 Mean : 76.68   
 3rd Qu.:17.50 3rd Qu.:3.000 3rd Qu.:108.00   
 Max. :21.00 Max. :3.000 Max. :108.00

describe(Men\_DF)

vars n mean sd median trimmed mad min max range  
emails 1 1507 7.69 10.26 4.00 5.62 4.45 0.00 88.00 88.00  
rating 2 1507 -0.09 0.54 -0.13 -0.11 0.51 -1.94 2.53 4.47  
height 3 1507 70.93 2.69 71.50 70.93 2.97 61.00 85.00 24.00  
bmi 4 1507 25.76 2.81 25.44 25.59 2.20 12.37 36.44 24.07  
yrs\_education 5 1507 15.66 2.58 16.00 15.72 2.97 8.00 21.00 13.00  
age 6 1507 1.95 0.82 2.00 1.93 1.48 1.00 3.00 2.00  
days\_active 7 1507 76.68 34.13 85.00 80.90 34.10 1.00 108.00 107.00  
 skew kurtosis se  
emails 2.97 12.54 0.26  
rating 0.38 0.61 0.01  
height 0.06 0.64 0.07  
bmi 0.53 1.41 0.07  
yrs\_education -0.47 0.79 0.07  
age 0.10 -1.51 0.02  
days\_active -0.67 -0.91 0.88

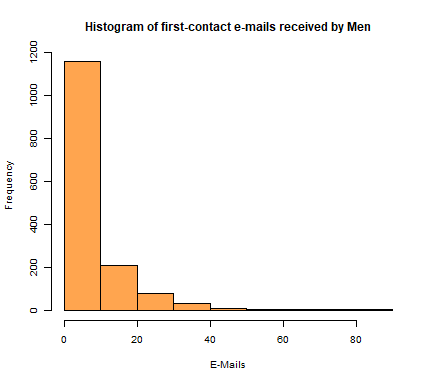
From the summary statistics, we can see that males of the dating site received between 0 and 88 first-contact e-mails during the observation period. On average, the male users received 7.69 e-mails, while the median number of e-mails received is low at 4, which indicates that the distribution of e-mails received is skewed. The first quartile is 1, meaning 25 percent of the site users received at most 1 e-mail. The third quartile is 10, or 75 percent of users received up to 10 e-mails.

The looks rating variable has both negative and positive values and is roughly centered at zero, with a mean of -0.09. This is due to the fact that original looks ratings, collected on a 1 to 10 scale, were normalized to have mean zero and variance one. The sample represents only a small part of the whole data base and is concentrated among users in a specific age range. This explains why the mean of the rating variable in our sample is not zero.

The height variable ranges from 61 inches to 85 inches. The mean hieght is 70.93 inches and median height is lower at 71.5 inches.

## 5c v. Histogram

par(cex = 0.65)   
hist(Men\_DF$emails,  
 col = "tan1",  
 main = "Histogram of first-contact e-mails received by Men",  
 xlab = "E-Mails")



The above histogram shows that the distribution of first-contact emails is right-skewed or positively skewed, meaning that there are a number of observations greater than the mode.

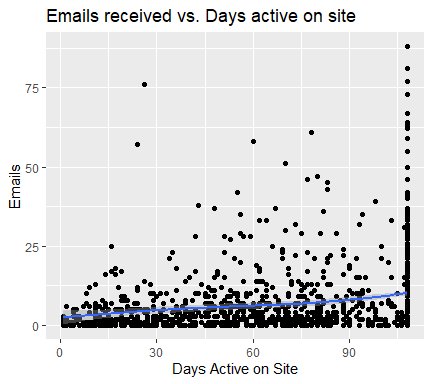
Comparing the histogram to the histogram ran on the emails received by women, we can see that both the distribution of emails for males and females are positively skewed.Though the range is higher for women, going up to 186 emails, while for men the maximum number of emails received is 88. However, the frequency of emails for males is higher in the first bin, going to nearly 1150. For females, the frequency only goes upto 650. This could be due to larger data size of males (1507) compared to females (1366).

## 5c vi. New package

install.packages("ggplot2",repos = "http://cran.us.r-project.org")

package 'ggplot2' successfully unpacked and MD5 sums checked  
  
The downloaded binary packages are in  
 C:\Users\aakan\AppData\Local\Temp\RtmpY1CzTQ\downloaded\_packages

library(ggplot2)  
  
ggplot(data=Men\_DF,mapping=aes(x=days\_active,y=emails))+  
 geom\_point()+  
 geom\_smooth() +  
 labs(title="Emails received vs. Days active on site",  
 x="Days Active on Site",y="Emails")



The above scatterplot plotted using ggplot package shows a positve correlation between emails received and days the male users are active on site.