**Case Study Project for EDA(Exploratory Data Analysis**)

**Case Study 1:**

1. **Female users tend to have significantly larger friend networks than Male users.**

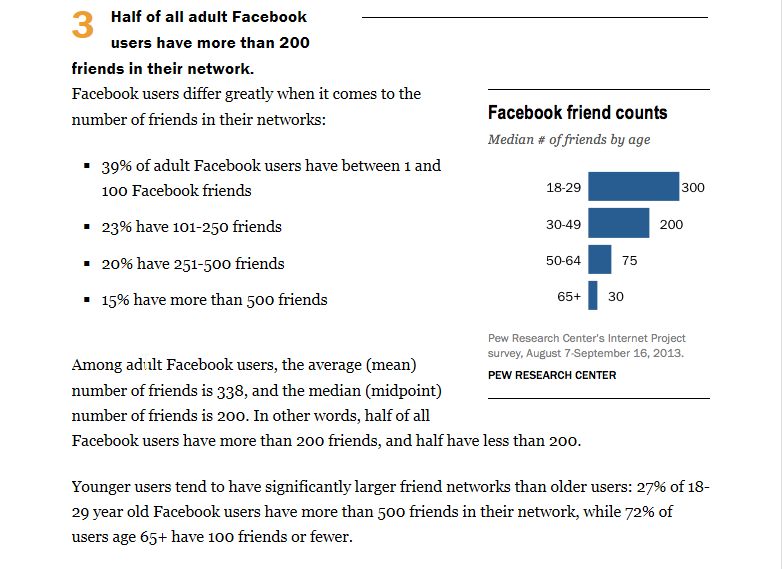
**Case Study 2:**

1. **How many times of the more friends does the average female users have than the male users?**

**Case Study 3:**

1. **Users accumulate more friends over years using the years they joined.**

**Case Study 4: From Pew Research**

****

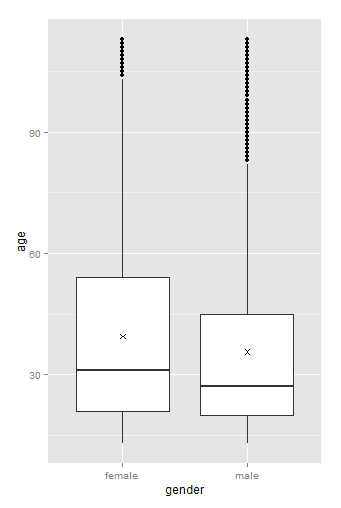
**Case Study 1:**

Created a box plot for each gender with the function of y as mean.

ggplot(aes(x = gender, y = age),

data = subset(pf, !is.na(gender))) + geom\_boxplot() +

stat\_summary(fun.y =mean , geom = 'point' , shape = 4)



Where x shows the average of each gender count since we used the shape = 4.

*Quite interesting observation I found is ,*

***Male users are younger compared to female users.***

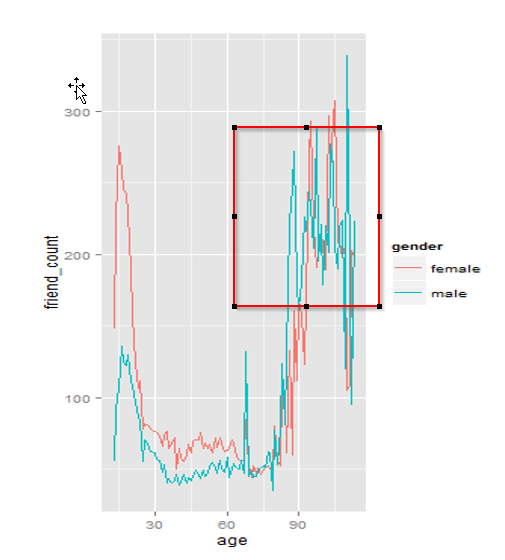
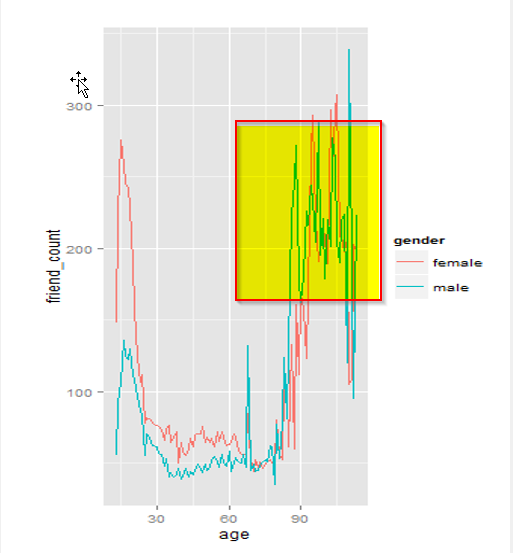
But this plot doesn’t capture the substantial difference in friend count.

So I created the same plot as a function of y as median.

ggplot(aes(x = age, y = friend\_count),

data = subset(pf, !is.na(gender))) +

geom\_line(aes(color = gender), stat = 'summary' , fun.y =median )

***Observation:***

This plot gave me an idea as median count for female friends count are larger than male users.

*At the age 70 both the male and female users have more or less the same size of friend count regardless of the gender.*

*Gender difference is larger at the young age.*

Since the older age (yellow highlighted area) has noisy distribution, I thought of reproduce the plot using the summary of data with Dplyr package.

**Case Study 3:**

Created new data frame called pf.fc\_by\_age\_gender.

pf.fc\_by\_age\_gender <- pf %>%

group\_by(age, gender) %>%

summarise(mean\_friend\_count = mean(friend\_count),

median\_friend\_count = median(friend\_count),

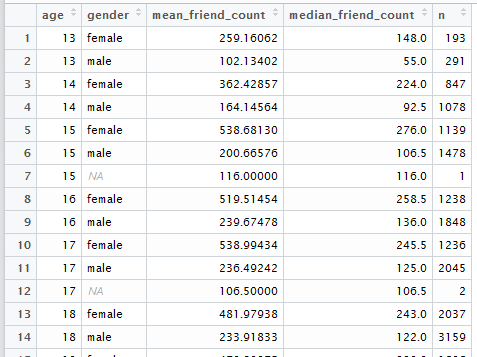
n= n()) %>%

ungroup() %>%

arrange(age)

How many times of the more friends does the average female users have than the male users?

Right now our data has long format with repeating ages.



I need to reshape the data by removing duplicates using reshape2 package.

Created new data frame using dcast pf.fc\_by\_age\_gender.wide

pf.fc\_by\_age\_gender.wide <- dcast(pf.fc\_by\_age\_gender,

age ~ gender ,

value.var = 'median\_friend\_count')

head(pf.fc\_by\_age\_gender.wide)

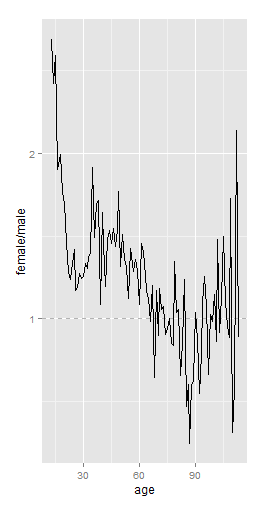
To find the ration plot,

ggplot(aes(x = age, y = female / male),

data = pf.fc\_by\_age\_gender.wide) +

geom\_line() +

geom\_hline(yintercept = 1, alpha = 0.3 , linetype = 2)



***I can see very young users than the median female users has over 2 and ½ times as many friends as the median of male users.***

I am able to visualize now with the condition of age with the relationship of friend’s count and gender.

**Case Study 3:**

Users accumulate more friends over years using the years they joined.

One way to explore is age , gender and the alpha variable or

Tenure is the good way to use.

pf$year\_joined <- floor(2014 - pf$tenure/365)

pf$year\_joined.bucket <- cut(pf$year\_joined,

c(2004,2009,2011,2012,2014))

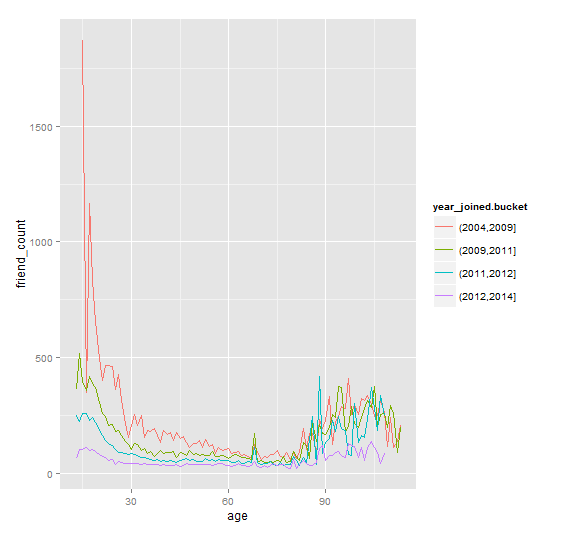
At this point New column with the name year\_joined has been added to pf data frame.

Created a variable name called year\_joined based on Tenure variable and converted this variable to year\_joined\_bucket based .

ggplot(aes(x = age, y = friend\_count),

data = subset(pf, !is.na(year\_joined.bucket))) +

geom\_line(aes(color = year\_joined.bucket), stat = 'summary' , fun.y =median )



**Observation:**

The above examination across friend count and age split up by year\_joined.bucket.

My assumptions are confirmed by look at this plot.

Longer tenure has the higher rate of friends count with the exception of older age of 80 and above.

**Case Study 4:**

I tried to simulate the research results from

http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/

Since the data set are different I get the different similar results.

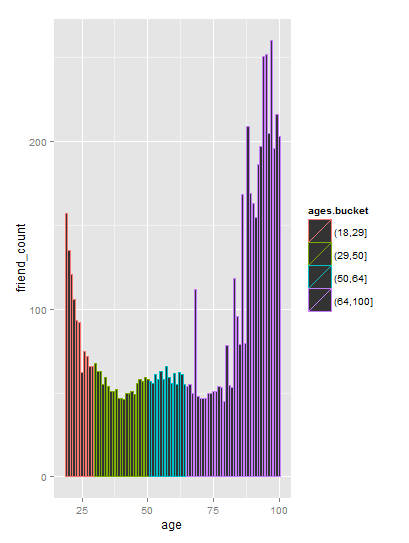
pf$ages.bucket <- cut(pf$age, breaks = c(18,29,50,64,100))

table(pf$ages.bucket ,useNA = 'ifany')

ggplot(aes(x = age, y = friend\_count),

data =subset(pf, !is.na(ages.bucket))) +

geom\_histogram(aes(color = ages.bucket), stat = 'summary' , fun.y =median)



ggplot(aes(x = ages.bucket , y = friend\_count),

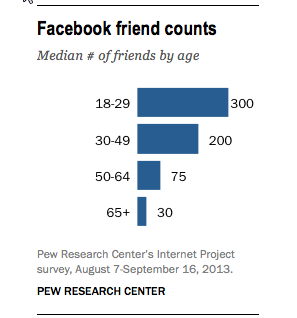
data =subset(pf, !is.na(gender))) +

geom\_histogram(aes(color = ages.bucket), stat = 'summary' ,fun.y =median)

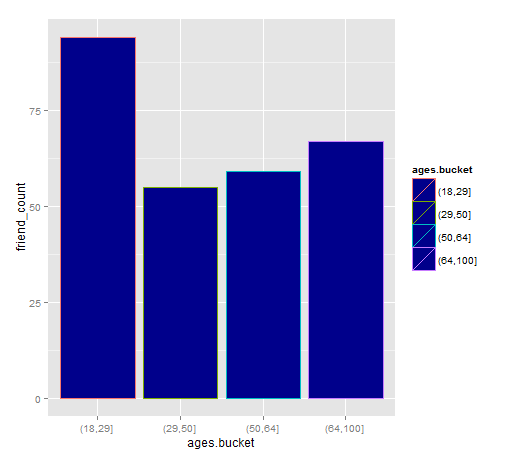
ggplot(aes(x = ages.bucket , y = friend\_count),

data =subset(pf, !is.na(pf$ages.bucket))) +

geom\_histogram(aes(color = ages.bucket), stat = 'summary' ,fill = 'darkblue' , fun.y =median)



My results:



**Pew Research Results:**

Among adult Facebook users, the average (mean) number of friends is 338, and the median (midpoint) number of friends is 200. In other words, half of all Facebook users have more than 200 friends, and half have less than 200.

**My Results:**

Among adult Facebook users, the average (mean) number of friends is 196, and the median (midpoint) number of friends is 88. In other words, half of all Facebook users have more than 88 friends, and half have less than 88.