Lihua Li

January 19, 2018

# Readin csv file

raw\_df <- read.csv("../conversion\_data.csv")

# Handle abnormal data

summary(raw\_df)

## country age new\_user source   
## China : 76602 Min. : 17.00 Min. :0.0000 Ads : 88740   
## Germany: 13056 1st Qu.: 24.00 1st Qu.:0.0000 Direct: 72420   
## UK : 48450 Median : 30.00 Median :1.0000 Seo :155040   
## US :178092 Mean : 30.57 Mean :0.6855   
## 3rd Qu.: 36.00 3rd Qu.:1.0000   
## Max. :123.00 Max. :1.0000   
## total\_pages\_visited converted   
## Min. : 1.000 Min. :0.00000   
## 1st Qu.: 2.000 1st Qu.:0.00000   
## Median : 4.000 Median :0.00000   
## Mean : 4.873 Mean :0.03226   
## 3rd Qu.: 7.000 3rd Qu.:0.00000   
## Max. :29.000 Max. :1.00000

Max age 123 seems not practical. Need a little investigation here.

df\_outlier <- subset(raw\_df, age >= 100)  
new\_df <- subset(raw\_df, age < 100)  
summary(new\_df)

## country age new\_user source   
## China : 76602 Min. :17.00 Min. :0.0000 Ads : 88739   
## Germany: 13055 1st Qu.:24.00 1st Qu.:0.0000 Direct: 72420   
## UK : 48449 Median :30.00 Median :1.0000 Seo :155039   
## US :178092 Mean :30.57 Mean :0.6855   
## 3rd Qu.:36.00 3rd Qu.:1.0000   
## Max. :79.00 Max. :1.0000   
## total\_pages\_visited converted   
## Min. : 1.000 Min. :0.00000   
## 1st Qu.: 2.000 1st Qu.:0.00000   
## Median : 4.000 Median :0.00000   
## Mean : 4.873 Mean :0.03225   
## 3rd Qu.: 7.000 3rd Qu.:0.00000   
## Max. :29.000 Max. :1.00000

Now max age being 79 is making more sense.

# Study regions and its relationship to conversion rates

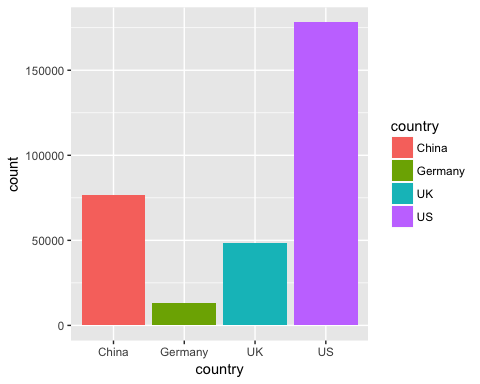
##   
## Attaching package: 'dplyr'

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

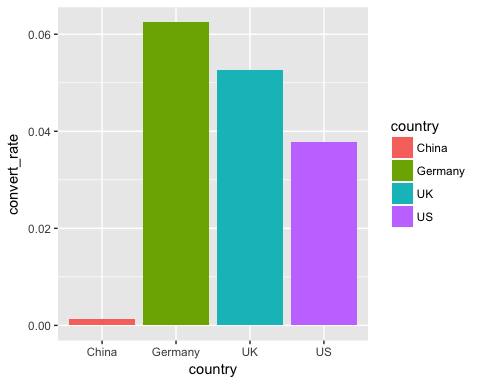
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Warning: package 'ggplot2' was built under R version 3.2.5

country\_count <- new\_df %>% group\_by(country) %>% summarise(count = n())  
country\_conversion <- new\_df %>% group\_by(country) %>% summarise(convert\_rate = mean(converted))  
  
ggplot(data = country\_count, aes(x = country, y = count))+ geom\_bar(stat = "identity", aes(fill = country))

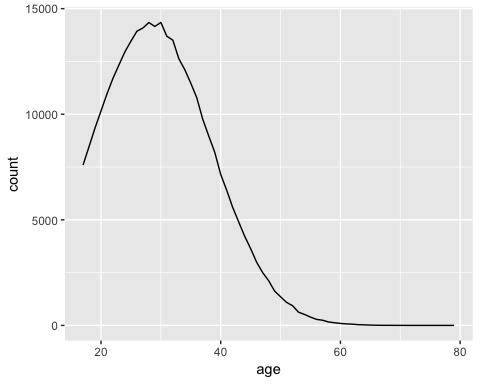


ggplot(data = country\_conversion, aes(x = country, y = convert\_rate))+ geom\_bar(stat = "identity", aes(fill = country))

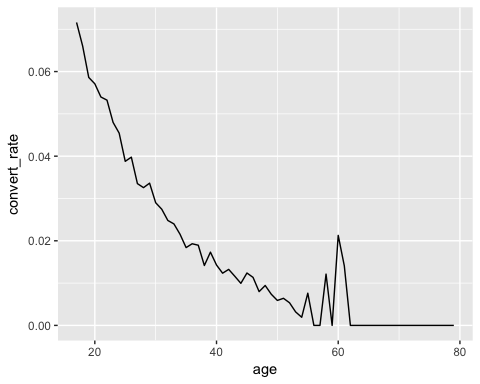
 Quick two findings: 1. There are a lot of users coming from China, but China has the lowest conversion rate. (Marketing team should investigate see if the website/product is culturally fit with Chinese) 2. Germany has the highest conversion rate, however, there is not many users coming from Germany. (Marketing team should broader their market to Germany)

# Study relationship between age and conversion rates

age\_count <- new\_df %>% group\_by(age) %>% summarise(count=n())  
age\_conversion <- new\_df %>% group\_by(age) %>% summarise(convert\_rate = mean(converted))  
  
ggplot(data = age\_count, aes(x = age, y = count)) + geom\_line(stat = 'identity', position = 'identity')

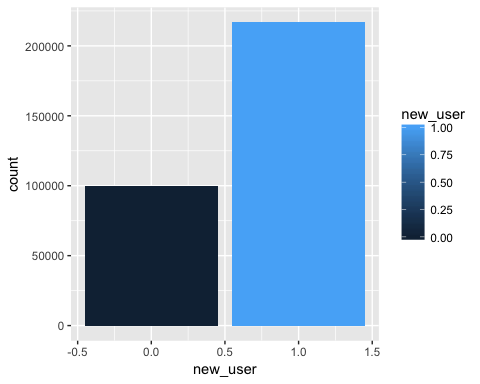


ggplot(data = age\_conversion, aes(x = age, y = convert\_rate)) + geom\_line(stat = 'identity', position = 'identity')

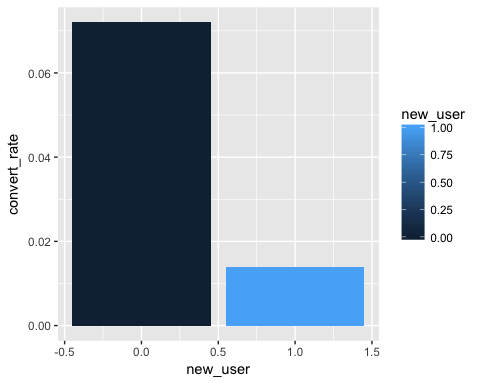
 The conversion rate is declining as age increases, and there is not many users above age 60, hence the bump around age 60 is simply due to noise.

# Study relationship between conversion rate and whether this user is repeat or not

newuser\_count <- new\_df %>% group\_by(new\_user) %>% summarise(count=n())  
newuser\_conversion <- new\_df %>% group\_by(new\_user) %>% summarise(convert\_rate = mean(converted))  
  
ggplot(data = newuser\_count, aes(x = new\_user, y = count)) + geom\_bar(stat = 'identity', aes(fill = new\_user))

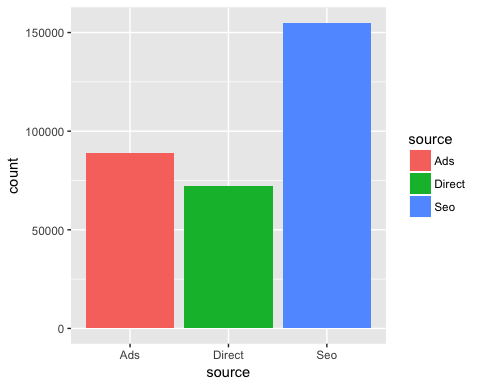


ggplot(data = newuser\_conversion, aes(x = new\_user, y = convert\_rate)) + geom\_bar(stat = 'identity', aes(fill = new\_user))

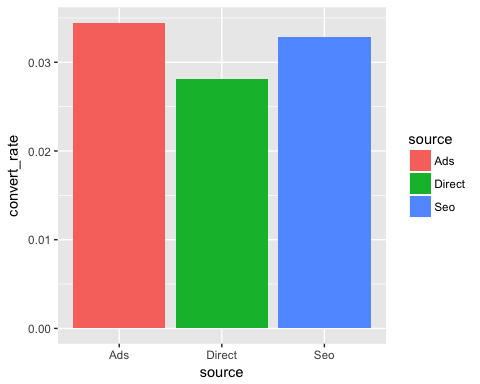
 The number of new users coming to this site is more than twice of repeat users, which is a sign of this product being not sticky enough. Since new users have much lower conversion rate than repeat users, marketing team should try to run promotions to existing users for bringing them back, rather than advertizing to new users. On the other hand, the team should survey new users to see the reasons of why they do not want to convert.

# Study relationship between conversion rate and source

source\_count <- new\_df %>% group\_by(source) %>% summarise(count = n())  
source\_conversion <- new\_df %>% group\_by(source) %>% summarise(convert\_rate = mean(converted))  
  
ggplot(data = source\_count, aes(x = source, y = count)) + geom\_bar(stat = 'identity', aes(fill = source))

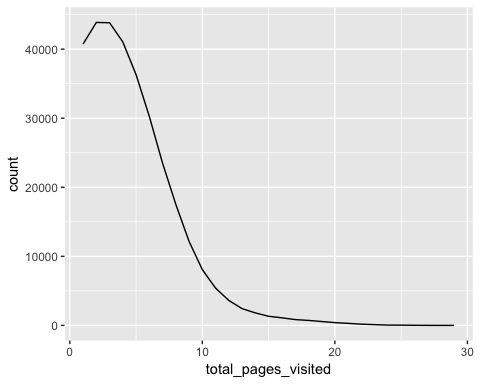


ggplot(data = source\_conversion, aes(x = source, y = convert\_rate)) + geom\_bar(stat = 'identity', aes(fill = source))

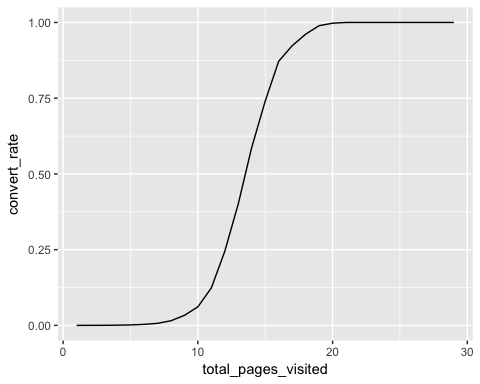
 Ads has the highest conversion rate even though it does not have the highest visiting rates.

# Study relationship between conversion rate and total\_pages\_visited

pages\_count <- new\_df %>% group\_by(total\_pages\_visited) %>% summarise(count=n())  
pages\_conversion <- new\_df %>% group\_by(total\_pages\_visited) %>% summarise(convert\_rate = mean(converted))  
  
ggplot(data = pages\_count, aes(x = total\_pages\_visited, y = count)) + geom\_line(stat = 'identity', position = 'identity')



ggplot(data = pages\_conversion, aes(x = total\_pages\_visited, y = convert\_rate)) + geom\_line(stat = 'identity', position = 'identity')

 The amount of ppl who visited above 10 pages dramatically decreases, however, once they visited more than 10 pages, the conversion rate is much higher and almost guaranteed to be converted if they have visited above 20 pages. This may imply that those consumers who are more interested in the product are more likely to spend more time on the website. But this could also be a sign that the website may fail to emphasize the strength of the product at the first few pages.