HW0-Fangzhou Song

Fangzhou Song

2019楠<9e>1閺<88><88>19閺<83>

### Package loading

rm(list = ls())  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.4.4

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'readr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.4

### Read data

data=read\_csv("C:/Users/ArkSong/Desktop/GWU/Stat 6240-Statistical Data Mining/Assignments/HW0/events\_log.csv",col\_names=TRUE)

### Data processing

glimpse(data)

## Observations: 400,165  
## Variables: 9  
## $ uuid <chr> "00000736167c507e8ec225bd9e71f9e5", "00000c69f...  
## $ timestamp <dbl> 2.016030e+13, 2.016031e+13, 2.016030e+13, 2.01...  
## $ session\_id <chr> "78245c2c3fba013a", "c559c3be98dca8a4", "760bf...  
## $ group <chr> "b", "a", "a", "a", "a", "a", "a", "b", "a", "...  
## $ action <chr> "searchResultPage", "searchResultPage", "check...  
## $ checkin <int> NA, NA, 30, 60, 30, 180, 240, NA, 180, 150, NA...  
## $ page\_id <chr> "cbeb66d1bc1f1bc2", "eb658e8722aad674", "f99a9...  
## $ n\_results <int> 5, 10, NA, NA, NA, NA, NA, 15, NA, NA, 20, NA,...  
## $ result\_position <int> NA, NA, NA, 10, NA, NA, NA, NA, 1, 1, NA, 1, N...

#### Question 1

1.What is their daily overall clickthrough rate? How does it vary between the groups?

Add variables year, month, day

data\_ymd=mutate(data,  
 year=timestamp %/% 1e+10,  
 month=(timestamp %/% 1e+8) %% 100 ,  
 day=(timestamp %/% 1e+6) %% 100)

Calculate daily overall clickthrough rate

data1=data\_ymd %>%  
 select(year,month,day,group,session\_id,timestamp,action) %>%  
 filter(action=="searchResultPage" |action=="visitPage" ) %>%  
 arrange(day,session\_id,timestamp) %>%  
 group\_by(year,month,day,session\_id) %>%  
 mutate(  
 action\_lag=lag(action)  
 ) %>%  
 filter(  
 action=="searchResultPage" | (action=="visitPage" & action\_lag=="searchResultPage")  
 )   
  
data1 %>%  
 group\_by(year,month,day) %>%  
 summarise(  
 visitPage\_count=sum(action=="visitPage"),  
 searchResultPage\_count=sum(action=="searchResultPage"),  
 cr=visitPage\_count/searchResultPage\_count  
 )

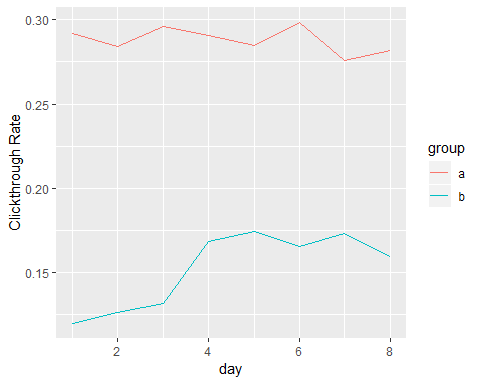
## # A tibble: 8 x 6  
## # Groups: year, month [?]  
## year month day visitPage\_count searchResultPage\_count cr  
## <dbl> <dbl> <dbl> <int> <int> <dbl>  
## 1 2016 3.00 1.00 4364 18374 0.238  
## 2 2016 3.00 2.00 4476 18902 0.237  
## 3 2016 3.00 3.00 4704 19159 0.246  
## 4 2016 3.00 4.00 4189 16675 0.251  
## 5 2016 3.00 5.00 3251 13204 0.246  
## 6 2016 3.00 6.00 3678 14612 0.252  
## 7 2016 3.00 7.00 4598 19011 0.242  
## 8 2016 3.00 8.00 3932 16297 0.241

Between groups

(data1\_group=data1 %>%  
 group\_by(year,month,day,group) %>%  
 summarise(  
 visitPage\_count=sum(action=="visitPage"),  
 searchResultPage\_count=sum(action=="searchResultPage"),  
 cr=visitPage\_count/searchResultPage\_count  
 )   
)

## # A tibble: 16 x 7  
## # Groups: year, month, day [?]  
## year month day group visitPage\_count searchResultPage\_count cr  
## <dbl> <dbl> <dbl> <chr> <int> <int> <dbl>  
## 1 2016 3.00 1.00 a 3671 12582 0.292  
## 2 2016 3.00 1.00 b 693 5792 0.120  
## 3 2016 3.00 2.00 a 3757 13209 0.284  
## 4 2016 3.00 2.00 b 719 5693 0.126  
## 5 2016 3.00 3.00 a 3930 13280 0.296  
## 6 2016 3.00 3.00 b 774 5879 0.132  
## 7 2016 3.00 4.00 a 3283 11298 0.291  
## 8 2016 3.00 4.00 b 906 5377 0.168  
## 9 2016 3.00 5.00 a 2451 8612 0.285  
## 10 2016 3.00 5.00 b 800 4592 0.174  
## 11 2016 3.00 6.00 a 2827 9469 0.299  
## 12 2016 3.00 6.00 b 851 5143 0.165  
## 13 2016 3.00 7.00 a 3506 12699 0.276  
## 14 2016 3.00 7.00 b 1092 6312 0.173  
## 15 2016 3.00 8.00 a 3074 10907 0.282  
## 16 2016 3.00 8.00 b 858 5390 0.159

ggplot(data=data1\_group)+  
 geom\_line(mapping = aes(x=day,y=cr,color=group))+  
 ylab("Clickthrough Rate")



#### Qusetion 2

2.Which results do people tend to try first? How does it change day-to-day?

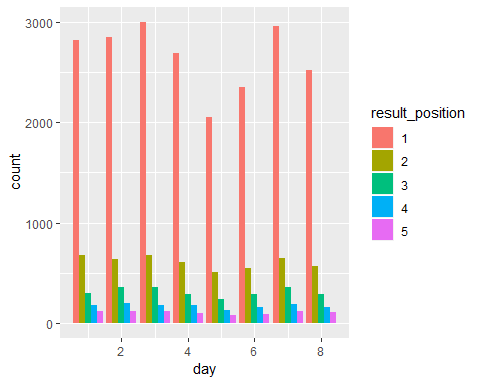
Calculate top 10 click results

head(data\_ymd %>%  
 filter(action=="visitPage") %>%  
 count(result\_position),10)

## # A tibble: 10 x 2  
## result\_position n  
## <int> <int>  
## 1 1 21263  
## 2 2 4858  
## 3 3 2482  
## 4 4 1358  
## 5 5 847  
## 6 6 613  
## 7 7 436  
## 8 8 293  
## 9 9 237  
## 10 10 216

Top 5 Change day-to-day

data\_ymd %>%  
 mutate(result\_position=factor(result\_position)) %>%  
 filter(action=="visitPage",result\_position %in% c(1:5)) %>%  
 group\_by(year,month,day) %>%  
 ggplot()+  
 geom\_bar(mapping = aes(x=day,fill=result\_position),position = "dodge")



#### Question 3

3.What is their daily overall zero results rate? How does it vary between the groups?

Daily overall zero results rate

data\_ymd %>%  
 filter(action=="searchResultPage") %>%  
 group\_by(year,month,day) %>%  
 summarise(  
 zero\_result=sum(n\_results==0),  
 total=n(),  
 zero\_rate=zero\_result/total  
 )

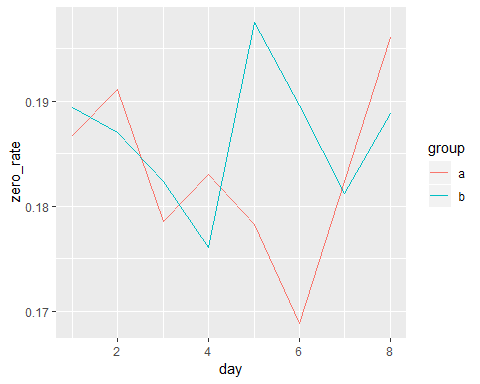
## # A tibble: 8 x 6  
## # Groups: year, month [?]  
## year month day zero\_result total zero\_rate  
## <dbl> <dbl> <dbl> <int> <int> <dbl>  
## 1 2016 3.00 1.00 3447 18374 0.188  
## 2 2016 3.00 2.00 3589 18902 0.190  
## 3 2016 3.00 3.00 3443 19159 0.180  
## 4 2016 3.00 4.00 3015 16675 0.181  
## 5 2016 3.00 5.00 2442 13204 0.185  
## 6 2016 3.00 6.00 2574 14612 0.176  
## 7 2016 3.00 7.00 3460 19011 0.182  
## 8 2016 3.00 8.00 3157 16297 0.194

Between groups

(data3=data\_ymd %>%  
 filter(action=="searchResultPage") %>%  
 group\_by(year,month,day,group) %>%  
 summarise(  
 zero\_result=sum(n\_results==0),  
 total=n(),  
 zero\_rate=zero\_result/total  
 )  
)

## # A tibble: 16 x 7  
## # Groups: year, month, day [?]  
## year month day group zero\_result total zero\_rate  
## <dbl> <dbl> <dbl> <chr> <int> <int> <dbl>  
## 1 2016 3.00 1.00 a 2350 12582 0.187  
## 2 2016 3.00 1.00 b 1097 5792 0.189  
## 3 2016 3.00 2.00 a 2524 13209 0.191  
## 4 2016 3.00 2.00 b 1065 5693 0.187  
## 5 2016 3.00 3.00 a 2371 13280 0.179  
## 6 2016 3.00 3.00 b 1072 5879 0.182  
## 7 2016 3.00 4.00 a 2068 11298 0.183  
## 8 2016 3.00 4.00 b 947 5377 0.176  
## 9 2016 3.00 5.00 a 1535 8612 0.178  
## 10 2016 3.00 5.00 b 907 4592 0.198  
## 11 2016 3.00 6.00 a 1599 9469 0.169  
## 12 2016 3.00 6.00 b 975 5143 0.190  
## 13 2016 3.00 7.00 a 2316 12699 0.182  
## 14 2016 3.00 7.00 b 1144 6312 0.181  
## 15 2016 3.00 8.00 a 2139 10907 0.196  
## 16 2016 3.00 8.00 b 1018 5390 0.189

data3 %>%  
 ggplot()+  
 geom\_line(mapping = aes(x=day,y=zero\_rate,color=group))



#### Question 4

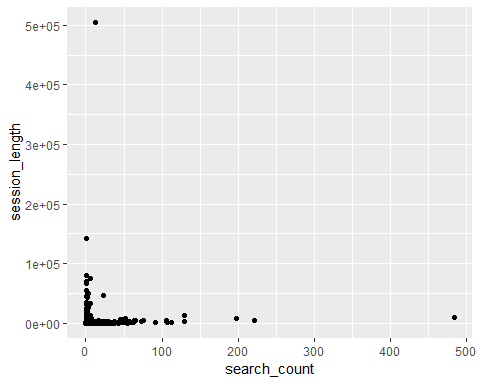
4.Let session length be approximately the time between the first event and the last event in a session. Choose a variable from the dataset and describe its relationship to session length.

Change timestamp into second unit

(data4=data %>%  
 mutate(  
 time=timestamp %% 100 + #second  
 ((timestamp %/% 100) %% 100)\*60 + #minute  
 ((timestamp %/% 10000) %% 100)\*3600+ #hour  
 ((timestamp %/% 1000000) %% 100)\*3600\*24 #day  
 ) %>%  
 arrange(session\_id,time) %>%  
 group\_by(session\_id) %>%  
 summarise(  
 session\_length=last(time)-first(time),  
 search\_count=sum(action=="searchResultPage")  
 )  
)

## # A tibble: 68,028 x 3  
## session\_id session\_length search\_count  
## <chr> <dbl> <int>  
## 1 0000cbcb67c19c45 0 1  
## 2 0001382e027b2ea4 303 1  
## 3 0001e8bb90445cb2 435 1  
## 4 000216cf18ae1ab1 58.0 6  
## 5 000527f711d50dfc 0 1  
## 6 00064fe774048046 43.0 2  
## 7 00071a2cf97168df 0 1  
## 8 0007582fe23d51e6 0 1  
## 9 0007b7f6b575feb6 339 1  
## 10 00086b6ff8156928 0 1  
## # ... with 68,018 more rows

ggplot(data=data4,mapping = aes(x=search\_count,y=session\_length))+  
 geom\_point()



data4 %>%  
 filter(session\_length < 1000 & session\_length!=0,search\_count < 50) %>%  
 ggplot(mapping = aes(x=search\_count,y=session\_length))+  
 geom\_point(alpha=0.2)+  
 geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

