July High DAU Invesitgation

This report analyze the reason of July high DAU by looking new/current/returning users and their retention rate and review our current social decaying model to see if we missed anything in the model or this is an accidental event.

Methodology: Group A: 1/1/2020-3/8/2020 (Before Covid19) Group B: 3/9/2020-6/13/2020 (Initial Social Distancing Data) Group C: 7/1/2020-7/18/2020 (July Data)

The final report can be found [here](https://confluenator.kingsisle.com:8443/display/WIZ/Investigation+of+July+2020+High+User+Volume).

DAU Plot

#Packages  
library(dplyr)

##   
## 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

library(ggplot2)  
library(sqldf)

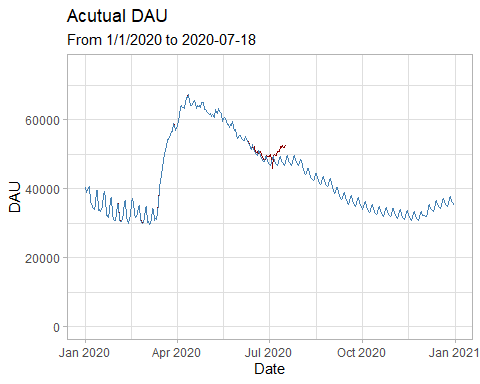
## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

data<-read.csv('./July High DAU Investigation Data.csv')  
data$date<-as.Date(data$date)  
#Update model version date  
first\_model\_date<-"2020-06-13"  
second\_model\_date<-"2020-07-18"  
#This is exactly data in the latest version  
all\_model\_data<-data%>%  
 mutate(actual=ifelse(date>=as.Date("2020-01-01")&date<=second\_model\_date,actual,NA),  
 current=ifelse(date>=as.Date("2020-01-01")&date<=second\_model\_date,current,NA),  
 new=ifelse(date>=as.Date("2020-01-01")&date<=second\_model\_date,new,NA),  
 returning=ifelse(date>=as.Date("2020-01-01")&date<=second\_model\_date,returning,NA))%>%  
 arrange(date)  
ggplot(all\_model\_data,aes(x=date))+  
 geom\_line(aes(y=actual),color='darkred')+  
 geom\_line(aes(y=social\_distancing),color='steelblue')+  
 ggtitle("Acutual DAU", subtitle=paste("From 1/1/2020 to ",second\_model\_date,sep=''))+  
 xlab("Date")+ylab("DAU")+ylim(0,75000)+  
 theme\_light()

## Warning: Removed 166 row(s) containing missing values (geom\_path).



There is an increase starting from early July

group\_a<-all\_model\_data%>%  
 filter(date<='2020-03-08')%>%  
 arrange(date)  
group\_b<-all\_model\_data%>%  
 filter(date>='2020-03-09',date<=first\_model\_date)%>%  
 arrange(date)  
group\_c<-all\_model\_data%>%  
 filter(date>='2020-07-01',date<=second\_model\_date)%>%  
 arrange(date)

User Retention Rate T Test: Group B vs Group C When P value is lower than 0.05, two groups retentions are significantly different

t.test(group\_b$current\_r,group\_c$current\_r)

##   
## Welch Two Sample t-test  
##   
## data: group\_b$current\_r and group\_c$current\_r  
## t = -4.3653, df = 29.325, p-value = 0.0001444  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.005331969 -0.001930865  
## sample estimates:  
## mean of x mean of y   
## 0.9332435 0.9368749

t.test(group\_b$new\_r,group\_c$new\_r)

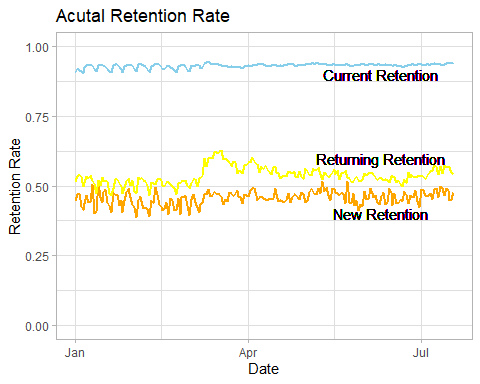
##   
## Welch Two Sample t-test  
##   
## data: group\_b$new\_r and group\_c$new\_r  
## t = -2.6889, df = 27.226, p-value = 0.01209  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.019651954 -0.002644611  
## sample estimates:  
## mean of x mean of y   
## 0.4632590 0.4744073

t.test(group\_b$returning\_r,group\_c$returning\_r)

##   
## Welch Two Sample t-test  
##   
## data: group\_b$returning\_r and group\_c$returning\_r  
## t = 0.31133, df = 38.576, p-value = 0.7572  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.008040857 0.010965230  
## sample estimates:  
## mean of x mean of y   
## 0.5548779 0.5534157

Current users and new users retention rate in July is significantly higher than March to June Retention Plot

#Plot Retention  
ggplot(data=all\_model\_data%>%filter(!is.na(actual)))+  
 geom\_line(aes(x=date,y=current\_r),color='sky blue',size=1)+  
 geom\_line(aes(x=date,y=new\_r),color='orange',size=1)+  
 geom\_line(aes(x=date,y=returning\_r),color='yellow',size=1)+  
 ggtitle("Acutal Retention Rate") + xlab("Date") + ylab("Retention Rate")+  
 geom\_text(aes(x=as.Date('2020-06-10'), y = 0.9, label = "Current Retention"))+  
 geom\_text(aes(x=as.Date('2020-06-10'), y = 0.6, label = "Returning Retention"))+  
 geom\_text(aes(x=as.Date('2020-06-10'), y = 0.4, label = "New Retention"))+  
 ylim(0,1)+  
 theme\_light()

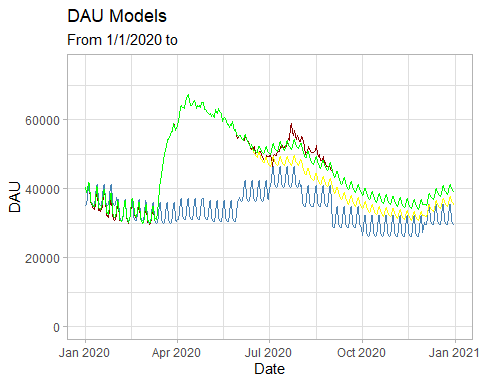


first\_model\_data<-data%>%  
 mutate(actual=ifelse(date>=as.Date("2020-01-01")&date<=first\_model\_date,actual,NA),  
 current=ifelse(date>=as.Date("2020-01-01")&date<=first\_model\_date,current,NA),  
 new=ifelse(date>=as.Date("2020-01-01")&date<=first\_model\_date,new,NA),  
 returning=ifelse(date>=as.Date("2020-01-01")&date<=first\_model\_date,returning,NA))%>%  
 arrange(date)  
first\_model\_na<-first\_model\_data%>%filter(!is.na(actual))  
cr<-mean(first\_model\_na$current\_r)  
nr<-mean(first\_model\_na$new\_r)  
rr<-mean(first\_model\_na$returning\_r)  
cr\_july<-mean(group\_c$current\_r)  
nr\_july<-mean(group\_c$new\_r)

result\_1<-first\_model\_data %>%  
 mutate(current\_new\_1=ifelse(date<=first\_model\_date,current,lag(current)\*cr+lag(new)\*nr+lag(returning)\*rr),  
 new\_new\_1=ifelse(date<=first\_model\_date,new,lag(new)-20.74),  
 return\_new\_1=ifelse(date<=first\_model\_date,returning,lag(returning)-24.547)  
 )  
start\_row<-nrow(first\_model\_data%>%filter(!is.na(actual)))+2  
end\_row<-nrow(first\_model\_data)  
for (i in start\_row:end\_row){  
 result\_1[i,15]=result\_1[i-1,15]-20.74  
 result\_1[i,16]=result\_1[i-1,16]-24.547  
 result\_2<-result\_1  
}  
result\_3<-result\_2%>%  
 mutate(new\_new\_1=ifelse(new\_new\_1>=new\_linear,new\_new\_1,new\_linear),  
 return\_new\_1=ifelse(return\_new\_1>=return\_log,return\_new\_1,return\_log),  
 current\_new\_2=current\_new\_1)  
for (i in start\_row:end\_row){  
 result\_3[i,14]=result\_3[i-1,14]\*cr+result\_3[i-1,15]\*nr+result\_3[i-1,16]\*rr  
 result\_3[i,17]=result\_3[i-1,17]\*cr\_july+result\_3[i-1,15]\*nr\_july+result\_3[i-1,16]\*rr  
 result\_4<-result\_3  
}  
#Current\_new\_1 is based on Covid19 first model high retention rate and Current\_new\_2 is based on July high retention rate

result\_5<-result\_4%>%  
 mutate(actual\_dau=current+new+returning,  
 december\_model\_dau=current\_log+new\_linear+return\_log,  
 june\_model\_dau=current\_new\_1+new\_new\_1+return\_new\_1,  
 july\_model\_dau=current\_new\_2+new\_new\_1+return\_new\_1)%>%  
 arrange(date)%>%  
 select(actual\_dau,december\_model\_dau,june\_model\_dau,july\_model\_dau)  
final\_result<-cbind(data,result\_5)  
ggplot(final\_result,aes(x=date))+  
 geom\_line(aes(y=actual),color='darkred')+  
 geom\_line(aes(y=december\_model\_dau),color='steelblue')+  
 geom\_line(aes(y=june\_model\_dau),color='yellow')+  
 geom\_line(aes(y=july\_model\_dau),color='green')+  
 ggtitle("DAU Models", subtitle="From 1/1/2020 to ")+  
 xlab("Date")+ylab("DAU")+ylim(0,75000)+  
 theme\_light()

## Warning: Removed 122 row(s) containing missing values (geom\_path).

 In July, current users and new users retention rate went up to push DAU up then DAU follows this high retention decaying line till the end of August.

covid<-read.csv('./covid.csv')  
covid$date<-as.Date(covid$date)  
data\_covid<-sqldf(  
 '  
 select data.date,data.current\_r,data.new\_r,daily\_case  
 from data  
 left join covid on data.date=covid.date  
 '  
 )  
cor.test(data\_covid$daily\_case,data\_covid$current\_r)

##   
## Pearson's product-moment correlation  
##   
## data: data\_covid$daily\_case and data\_covid$current\_r  
## t = 4.1309, df = 220, p-value = 5.134e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1416245 0.3863200  
## sample estimates:  
## cor   
## 0.2682945

cor.test(data\_covid$daily\_case,data\_covid$new\_r)

##   
## Pearson's product-moment correlation  
##   
## data: data\_covid$daily\_case and data\_covid$new\_r  
## t = 3.7595, df = 220, p-value = 0.0002182  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1178363 0.3655443  
## sample estimates:  
## cor   
## 0.2456972

There is a small correlation (>0.2) between current users/new users retention rate and daily Covid 19 cases