Russian Tweets

Alexandra DeKinder, ad3540

1/25/2019

R Markdown

Reading in and aggregating the csv files. This is clearly not the most elegant way; however, it worked for me at the time.

```
Rtweets<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Tweet
s/IRAhandle_tweets_1.csv")
Rtweets2<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_2.csv")
Rtweets3<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_3.csv")
Rtweets4<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_4.csv")
Rtweets5<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_5.csv")
Rtweets6<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_6.csv")
Rtweets7<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_7.csv")
Rtweets8<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_8.csv")
Rtweets9<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twee
ts/IRAhandle_tweets_9.csv")
Rtweets10<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twe
ets/IRAhandle_tweets_10.csv")
Rtweets11<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twe
ets/IRAhandle_tweets_11.csv")
Rtweets12<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twe
ets/IRAhandle_tweets_12.csv")
Rtweets13<-read.csv("/Users/alexandradekinder/Desktop/Data Science and PP/Russian Twe
ets/IRAhandle_tweets_13.csv")



Full_Rtweets<-rbind(Rtweets,Rtweets2,Rtweets3,Rtweets4,Rtweets5,Rtweets6,Rtweets7,Rtw
eets8,Rtweets9,Rtweets10,Rtweets11,Rtweets12,Rtweets13)
```

Due to the focus of our analysis being on how these tweets affected sentiment during the election, I will subset the full data to only focus on tweets in English.

```
Eng_Tweets<-Full_Rtweets[Full_Rtweets$language=="English",]
```

My graphics and analysis will focus on frequency of tweets by day or hour so I need to standardize the publish date of the tweets into a more useful format. I will put these into a new column called “NewDateTime”.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

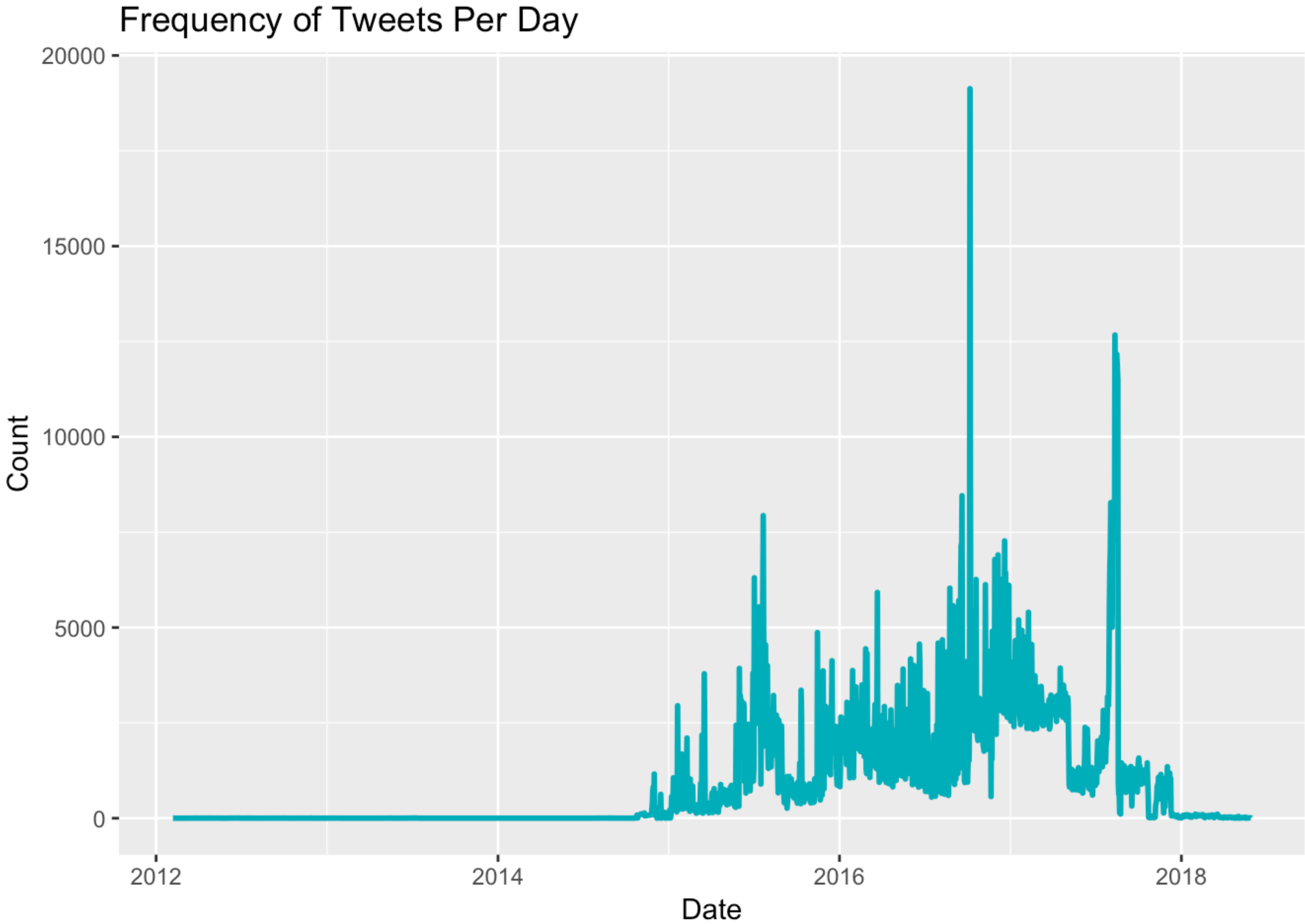
```
Eng_Tweets$NewDateTime <- as.POSIXlt(strptime(Eng_Tweets$publish_date, '%m/%d/%Y %H:%
M'))
```

In order to make graphing time series easier, I will aggregate the data into a frequency table by day.

```
#Assigning a day to each observation
Eng_Tweets$tweets_per_day<-as.Date(cut(Eng_Tweets$NewDateTime,breaks = "day"))

#Creating a count column and creating a frequency table
count<-rep.int(1,2116867)
time.df<-data.frame(Eng_Tweets$tweets_per_day,count)
tweet_count<-aggregate(time.df$count, by=list(time.df$Eng_Tweets.tweets_per_day), sum
)
colnames(tweet_count)<-c("Date","Count")
```

We can now create the graphics to help with analysis.

```
library(ggplot2)

#Basic line plot
ggplot(data = tweet_count, aes(x = Date, y =Count )) +
  geom_line(color = "#00AFBB", size = 1)+labs(title = "Frequency of Tweets Per Day")
```

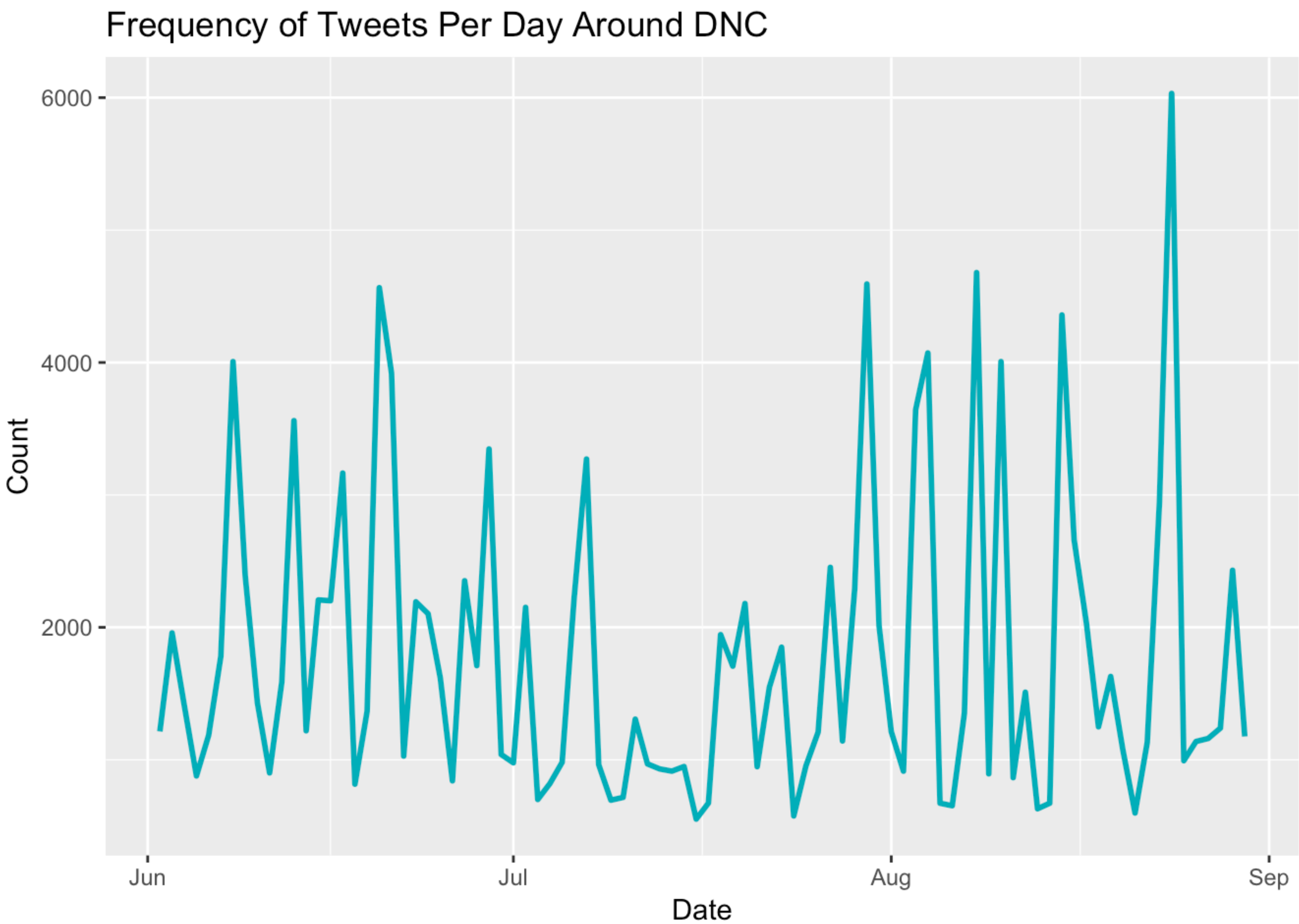



```
#Days with highest activity
MaxTweets<-tweet_count[order(tweet_count$Count,decreasing = TRUE),]
head(MaxTweets,n=10)
```

```
##            Date Count
## 788  2016-10-06 19128
## 1098 2017-08-12 12669
## 1102 2017-08-16 12161
## 1103 2017-08-17 11831
## 1104 2017-08-18 11550
## 1100 2017-08-14 11364
## 1101 2017-08-15 11118
## 1099 2017-08-13 10438
## 789  2016-10-07  8652
## 771  2016-09-19  8455
```

We can see that there is alot of variancy in the activity level of the tweets, also it is clear that while 2016 was active leading up to and around the election, the activity of the accounts did not stop. In fact we can see in the table that many of the most active days were during the late summer of 2017.

```r
#Graphics for subset of tweets from June through August of 2016
SubTweets <- subset(tweet_count, Date > as.Date("2016-06-01") & Date < as.Date("2016-
08-31"))
ggplot(data = SubTweets, aes(x = Date, y =Count ))+
  geom_line(color = "#00AFBB", size = 1)+labs(title = "Frequency of Tweets Per Day Ar
ound DNC")
```




```r
SubMaxTweets<-SubTweets[order(SubTweets$Count,decreasing = TRUE),]
head(SubMaxTweets,n=10)
```

```
##           Date Count
## 745 2016-08-24  6032
## 729 2016-08-08  4679
## 720 2016-07-30  4593
## 680 2016-06-20  4565
## 736 2016-08-15  4359
## 725 2016-08-04  4072
## 668 2016-06-08  4007
## 731 2016-08-10  4007
## 681 2016-06-21  3921
## 724 2016-08-03  3643
```

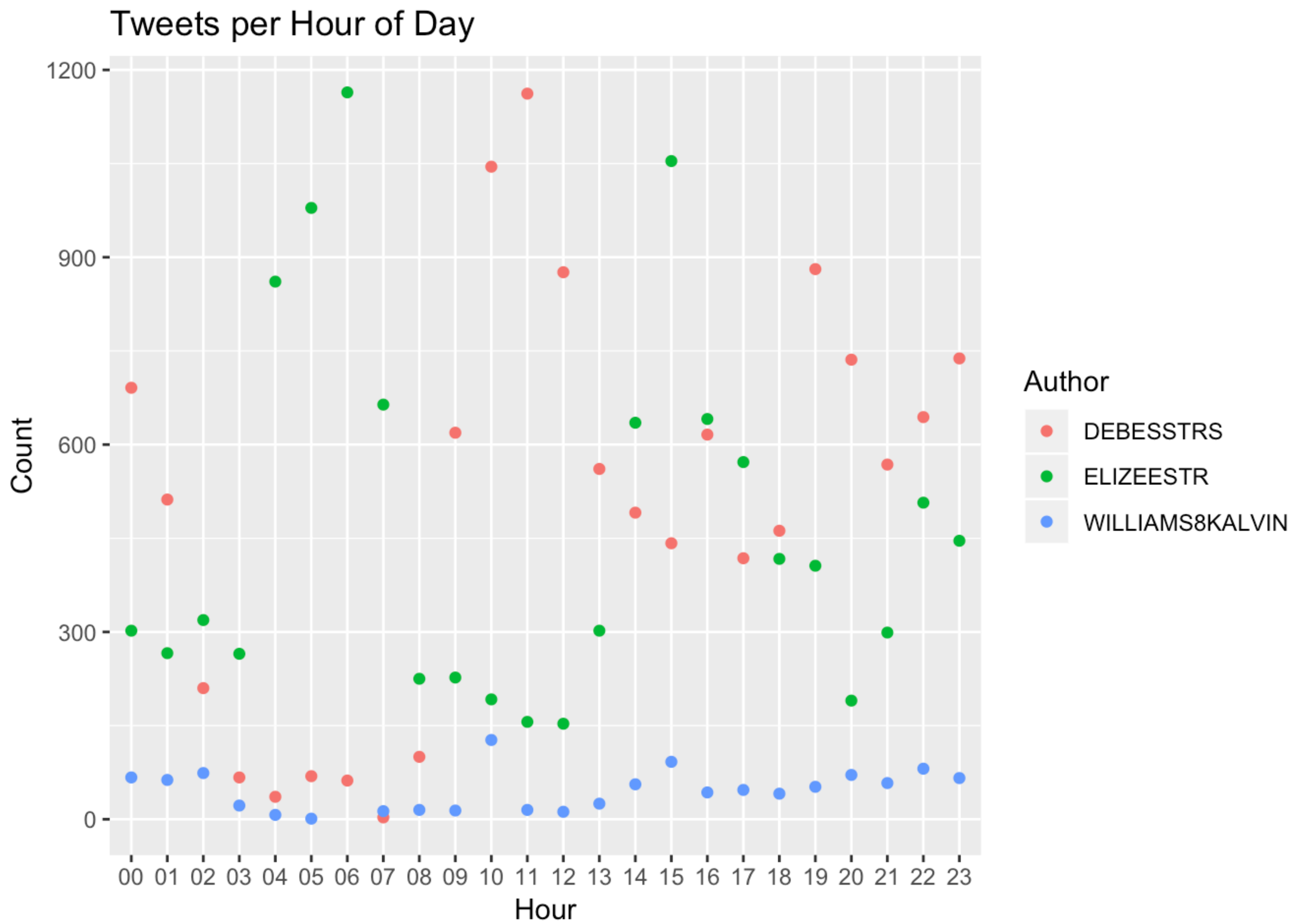
These graphics represent the activity during the 3 month span of 2016 that included the WikiLeaks and Democratic National Convention. The time series plot indicates quite a lot of back and forth activity. There was a relative lull in mid July; however, there was a steady rise through August. We can also see the 10 most active days in the table. The DNC was July 25-28 and we can that there was a rise in twitter activity around the end of July; however, July 30th was the only day to make the top 10 most active days list.

Now I will examine the accounts with the highest tweets per minute numbers to see if there is any pattern to the time of day that these accounts are tweeting.

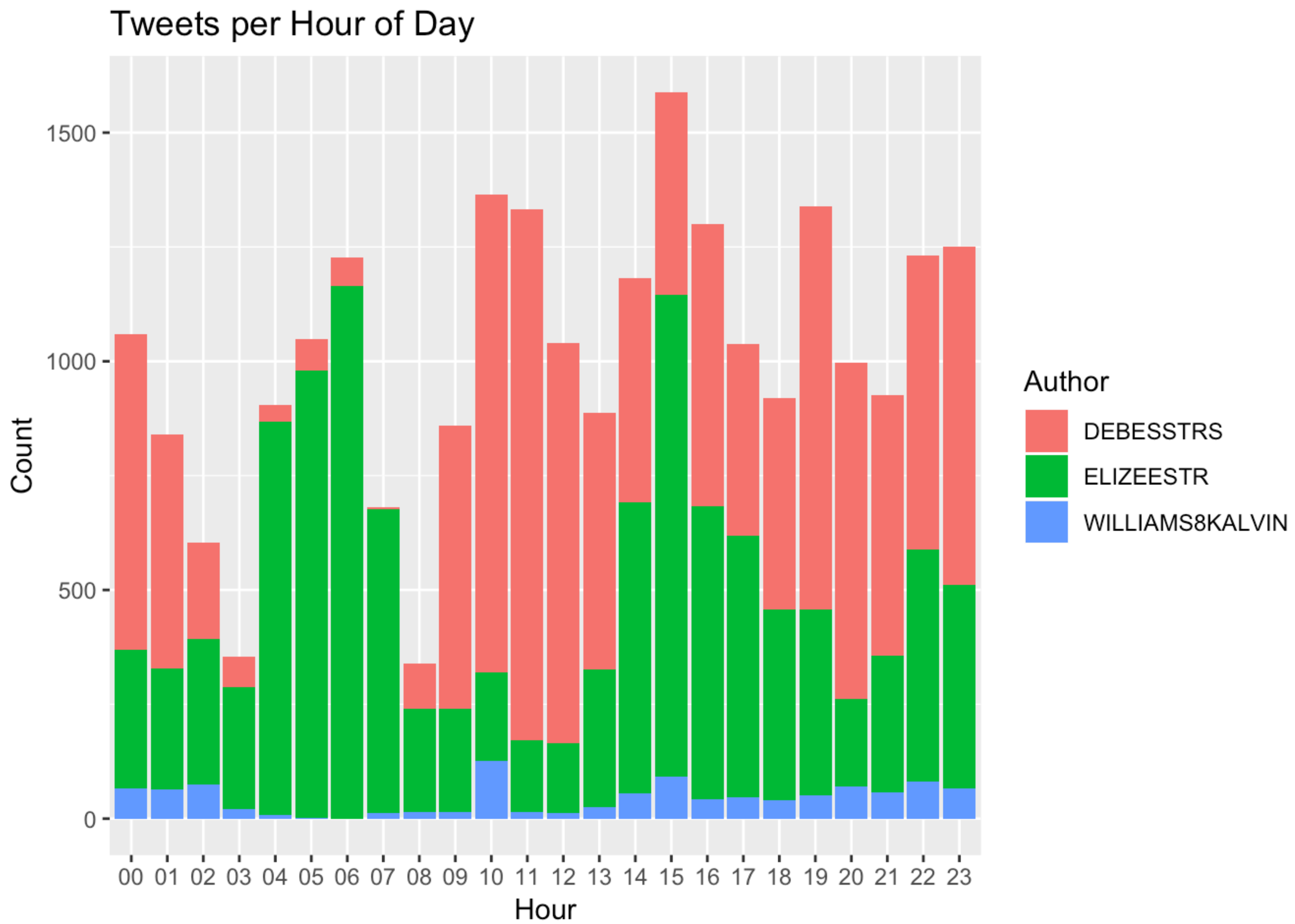
I will focus on the authors: WILLIAMS8KALVIN, ELIZEESTR, and DEBESSTRS

```
####WILLIAMS8KALVIN####


Will_tweets<-Full_Rtweets[Full_Rtweets$author == "WILLIAMS8KALVIN",]

#Creating count column
Will_tweets$Count<-rep(1,1062)

#Formatting date
Will_tweets$NewDateTime <- as.character(Will_tweets$publish_date)

#Creating tweets per hour column

Will_tweets$hour<-format(as.POSIXct(strptime(Will_tweets$publish_date,"%m/%d/%Y %H:%M
",tz="")) ,
                         format = "%H")
Will_time.df<-data.frame(Will_tweets$hour,Will_tweets$Count)
Will_tweet_hour_count<-aggregate(Will_time.df$Will_tweets.Count, by=list(Will_time.df
$Will_tweets.hour), sum)
colnames(Will_tweet_hour_count)<-c("Hour","Count")


####ELIZEESTR####

Eliz_tweets<-Full_Rtweets[Full_Rtweets$author == "ELIZEESTR",]

#Creating count column
Eliz_tweets$Count<-rep(1,length(Eliz_tweets$author))

#Creating tweets per hour column

Eliz_tweets$Hour<-format(as.POSIXct(strptime(Eliz_tweets$publish_date,"%m/%d/%Y %H:%M
",tz="")) ,
                         format = "%H")
Eliz_time.df<-data.frame(Eliz_tweets$Hour,Eliz_tweets$Count)
Eliz_tweet_hour_count<-aggregate(Eliz_time.df$Eliz_tweets.Count, by=list(Eliz_time.df
$Eliz_tweets.Hour), sum)
colnames(Eliz_tweet_hour_count)<-c("Hour","Count")
```

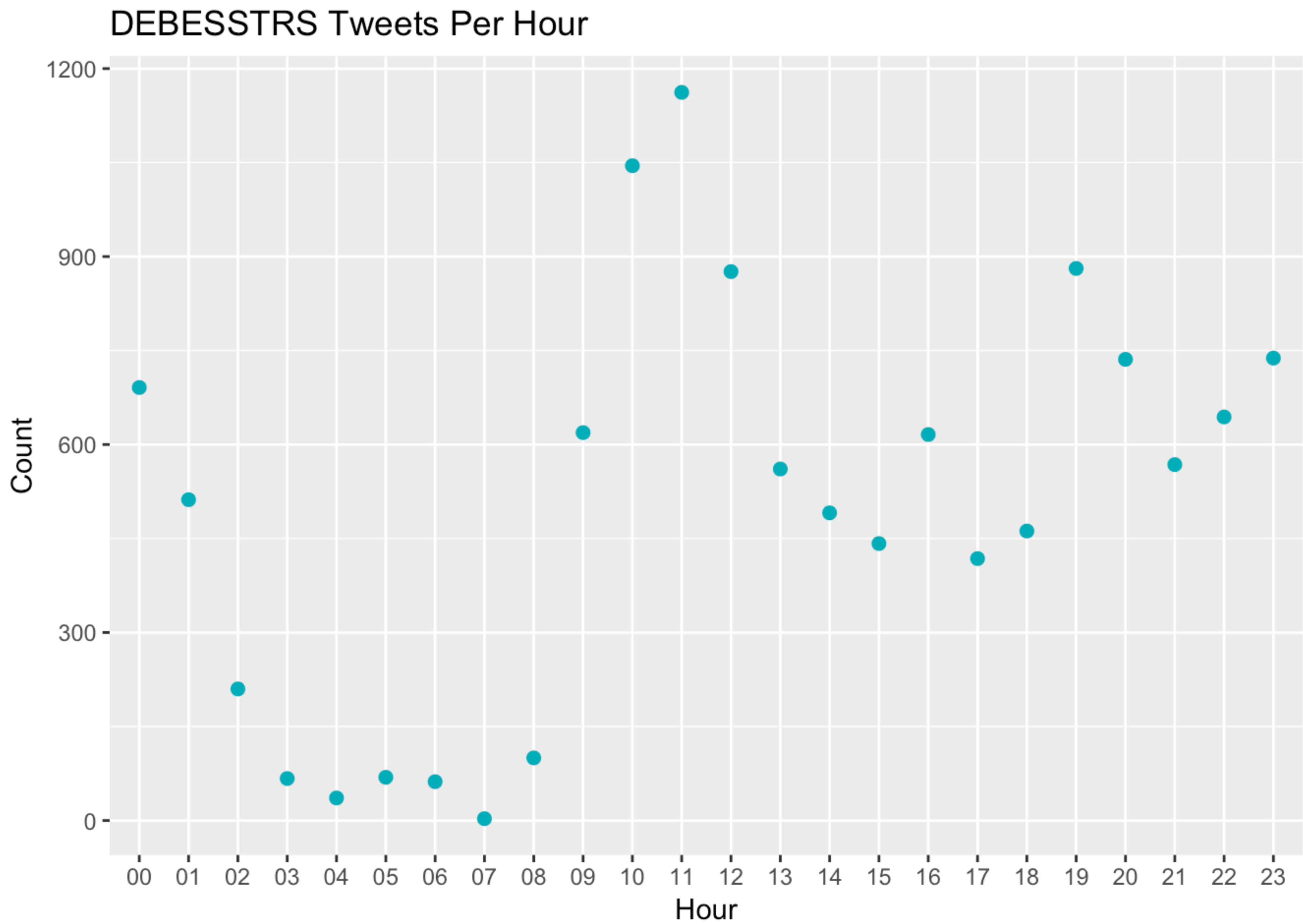
```
####DEBESSTRS####


Deb_tweets<-Full_Rtweets[Full_Rtweets$author == "DEBESSTRS",]

#Creating count column
Deb_tweets$Count<-rep(1,length(Deb_tweets$author))

#Creating tweets per hour column

Deb_tweets$Hour<-format(as.POSIXct(strptime(Deb_tweets$publish_date,"%m/%d/%Y %H:%M",
tz="")) ,
                         format = "%H")
Deb_time.df<-data.frame(Deb_tweets$Hour,Deb_tweets$Count)
Deb_tweet_hour_count<-aggregate(Deb_time.df$Deb_tweets.Count, by=list(Deb_time.df$Deb
_tweets.Hour), sum)
colnames(Deb_tweet_hour_count)<-c("Hour","Count")

####GRAPHICS####

#First I need to reshape the data to make plotting multiple series on one plot easier

library(reshape2)

melt_hour_tweets<- melt(list(DEBESSTRS=Deb_tweet_hour_count,ELIZEESTR=Eliz_tweet_hour
_count,
                             WILLIAMS8KALVIN=Will_tweet_hour_count), id.vars="Hour")
colnames(melt_hour_tweets)<-c("Hour","variable","Count","Author")

#Now I can make my graphs

ggplot(melt_hour_tweets, aes(Hour,Count, color = Author))+geom_point()+labs(title = "
Tweets per Hour of Day")
```



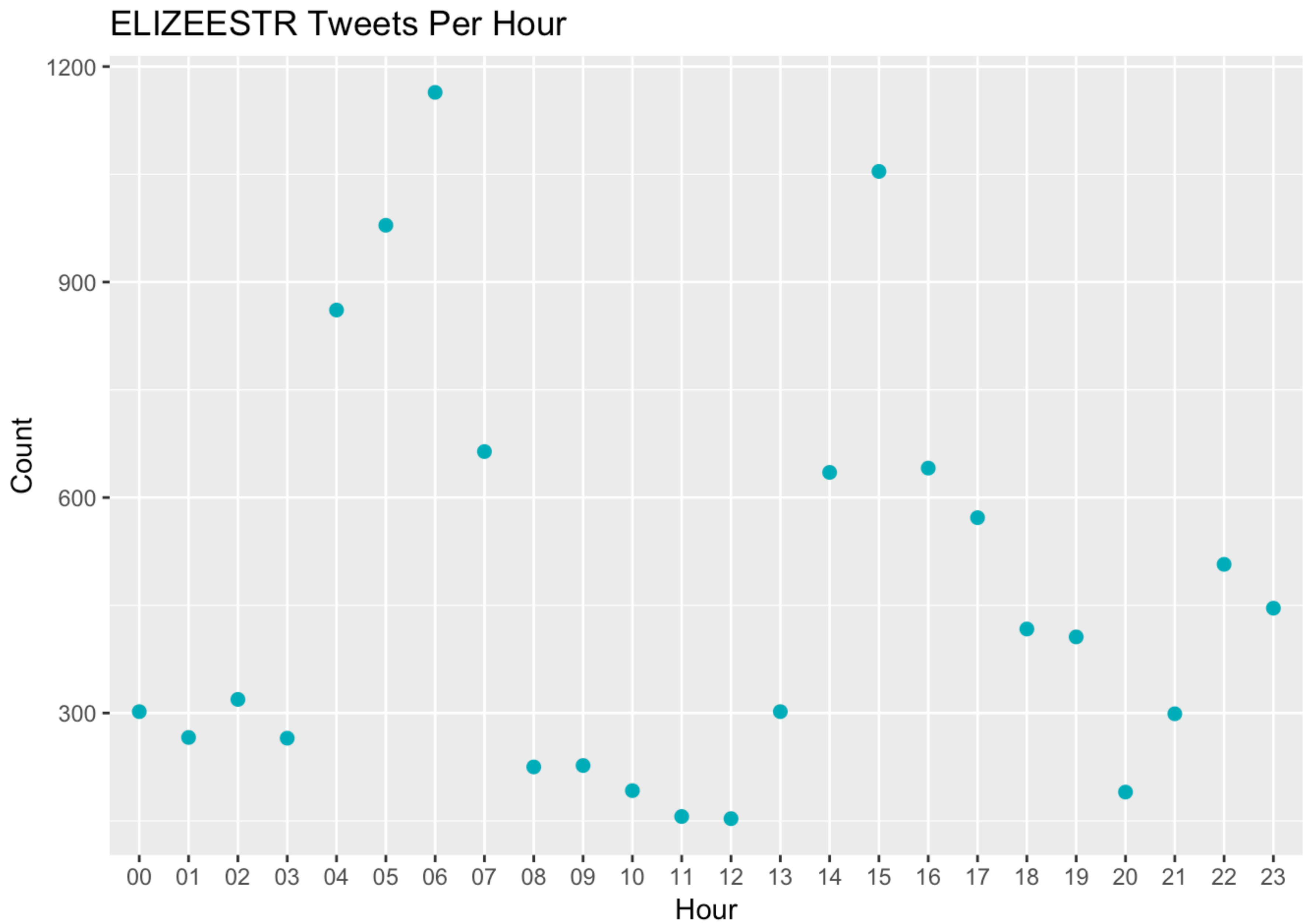

```
ggplot(melt_hour_tweets,aes(x=Hour, y= Count,fill=Author))+geom_bar(stat = "identity"
)+labs(title = "Tweets per Hour of Day")
```



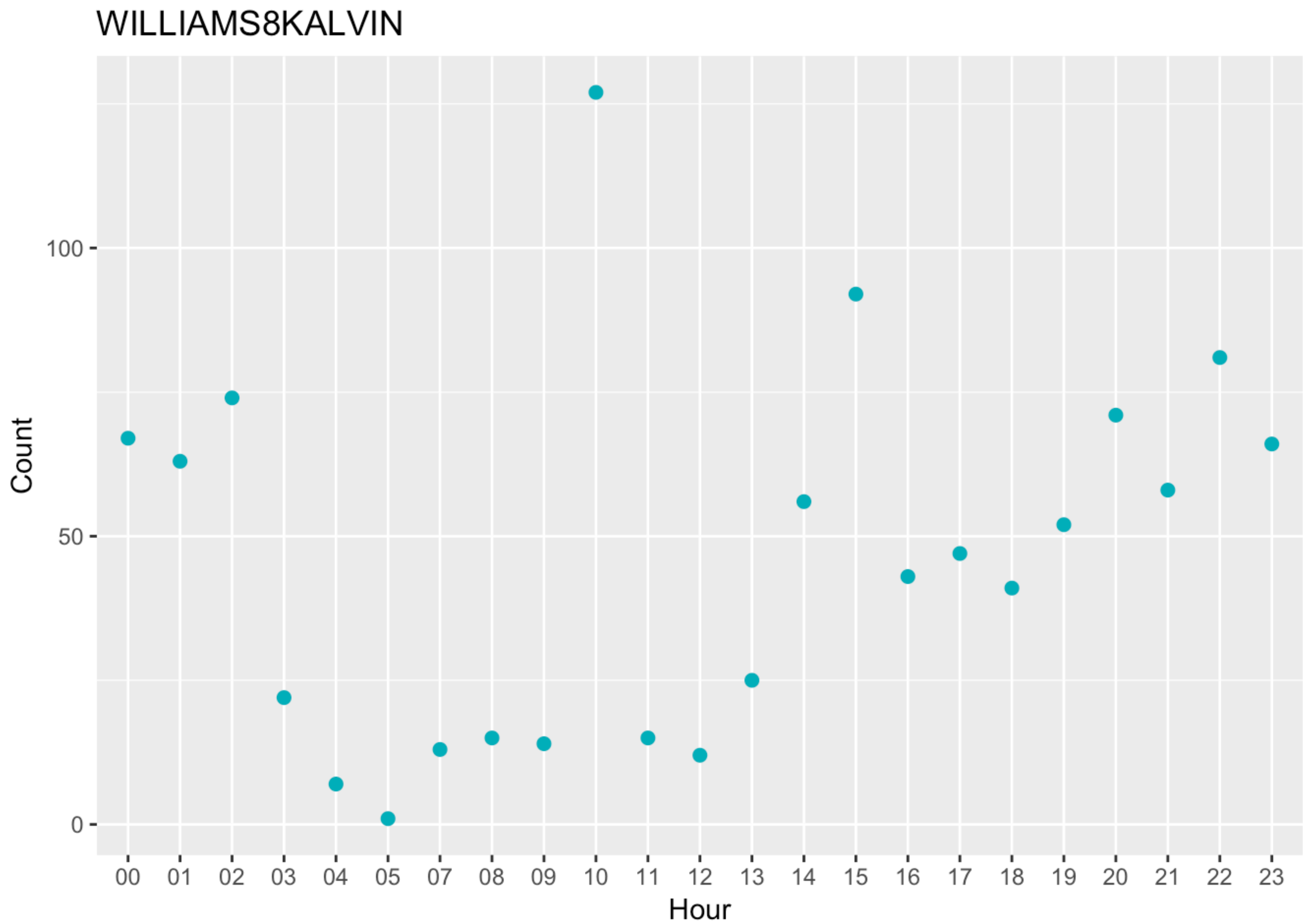

```
ggplot(data = Deb_tweet_hour_count, aes(x = Hour, y =Count))+
  geom_point(color = "#00AFBB", size = 2)+labs(title = "DEBESSTRS Tweets Per Hour")
```




```
ggplot(data = Eliz_tweet_hour_count, aes(x = Hour, y =Count ))+
  geom_point(color = "#00AFBB", size = 2)+labs(title = "ELIZEESTR Tweets Per Hour")
```




```
ggplot(data = Will_tweet_hour_count, aes(x = Hour, y =Count ))+
  geom_point(color = "#00AFBB", size = 2)+labs(title = "WILLIAMS8KALVIN")
```



Based on the above graphs, it appears there could be more activity from these accounts in the afternoon, early-evening hours.

Since there appears to be some similarities between two of these accounts I will now see if the full dataset has any similar patterns.

```
#Creating tweets per hour column

#Full_Rtweets$Newhour<-format(as.POSIXct(strptime(Full_Rtweets$publish_date,"%m/%d/%Y
%H:%M",tz="")),
                          #format = "%H")
#Full_time.df<-data.frame(Full_Rtweets$Newhour,Full_Rtweets$Count)
#Full_tweet_hour_count<-aggregate(Full_time.df$Full_Rtweets.Count,
                                 #by=list(Full_time.df$Full_Rtweets.Newhour), sum)
#colnames(Full_tweet_hour_count)<-c("Hour","Count")


#Graphic

#ggplot(Full_tweet_hour_count,aes(x=Hour, y= Count,fill=Hour))+geom_bar(stat = "ident
ity")+labs(title = "Tweets per Hour of Day")
```

The graph of the full dataset coroborates the pattern found when just a few accounts were analyzed. The exact cause of the heightened activity in the afternoon/evening would need furhter analysis; however, it is clear that these times are more active.