Analysis of Secrecy VS Security in social context

##### Software tools used : R,Tableau,Text mining tools.

#### The objective is to analyse the public sentiment about Privacy and Secrecy, and find a demacration between privacy and secrecy in a social context. For this study we consider four major incidents

###### 1. "Ashley Madison incident"

###### 2. "Celebrity photo hack incident"

###### 3. "Edward snowden incident"

###### 4. "FBI VS APPLE encryption incident"

## Data Extraction:

##### In order to perform analysis we extracted social media data from Twitter(tweets) using the Crimson hexagon platform. This platform has lot of data(millions of tweets) so we took a sample of tweets for each of these incidents and the sampling is done based on certain keywords, incident triggered date. For example in case of Ashley Madison incident keywords used are "The Ashely Madison Hack" OR "Ashley Madison" OR site: "ashleymadison.com" and for each of the four different incidents, four data samples are obtained based on the 15-day slots based on the event triggered date (before 15 days, first 15 days, after 16-30 days,31-45 days).

##### The below table shows the dimensions of samples for each of the incidents.

## Data Description:

##### Once we obatin four samples for each incident we can combine them and make it one sample,which has data for 60-days.The data extracted from Crimson hexagon has several useful variables for analysis, and the schema is below.

##### "GUID","Date","URL", "Contents","Author","Name","Country","State/Region","City/Urban Area","Category","Source","Klout Score","Gender","Posts","Followers" and "Following" are the variables/columns in the data.

##### Among all of these variables we are more interested in the "Date","Contents" (Raw tweets),"Country","Category"(Sentiment: Positive/Negative/Neutral),"Gender" variables and the "Category" variable is coded by the 'Crimson Hexagon' platform.

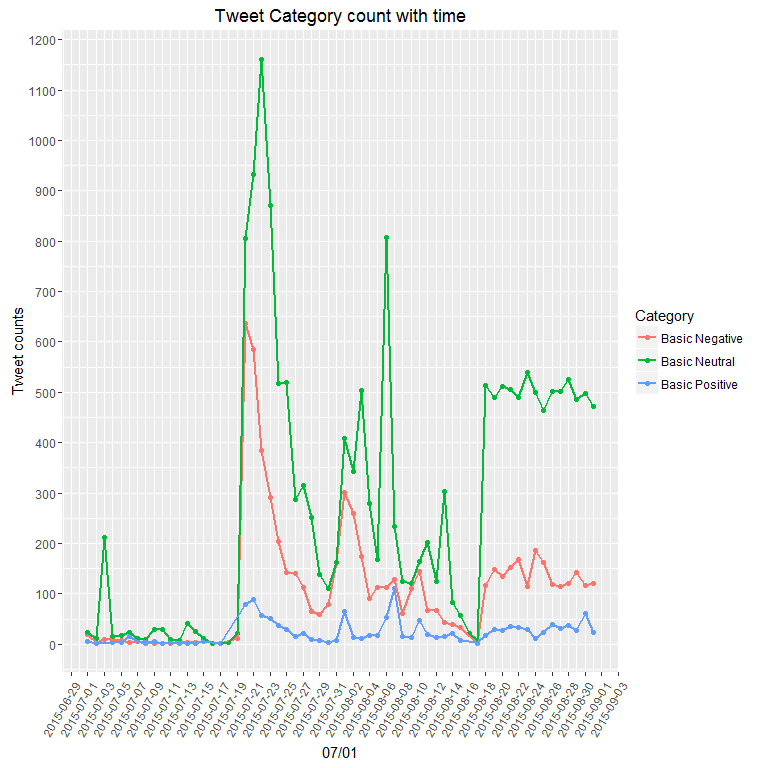
## Data Exploration:

##### By now we have the raw data and it has several variables ,one of them is "Contents" which refers the actual tweet.We know that each tweet has several words and is there any way to associate the words of tweet to the sentiment(Poistive/Negative/Neutral).Below are the plots that can dymistify some of the patterns.

##### We consider Ashley madison incident for the plots below ans the same procedure can be repliacted for other incidents.

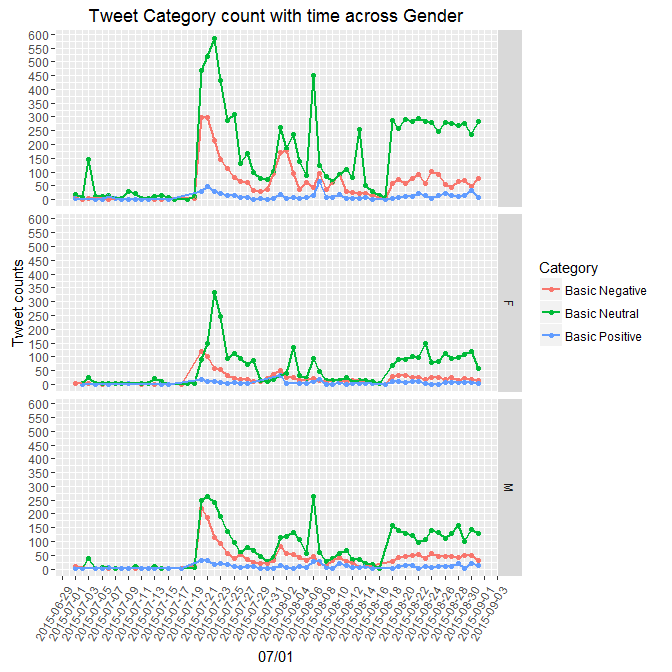
##### plot for the tweet counts for different categories for the 60-days.

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



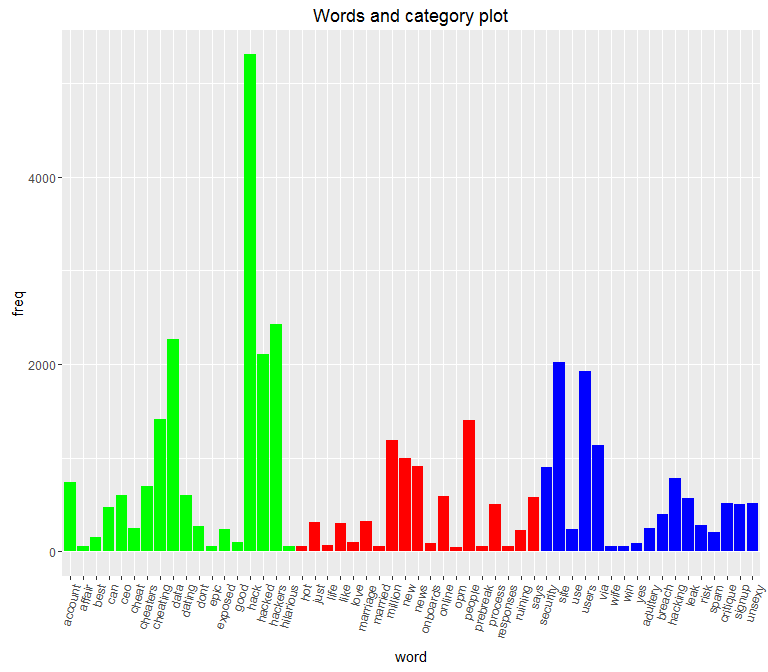
##### We can do similar plot as above for different "Gender"

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



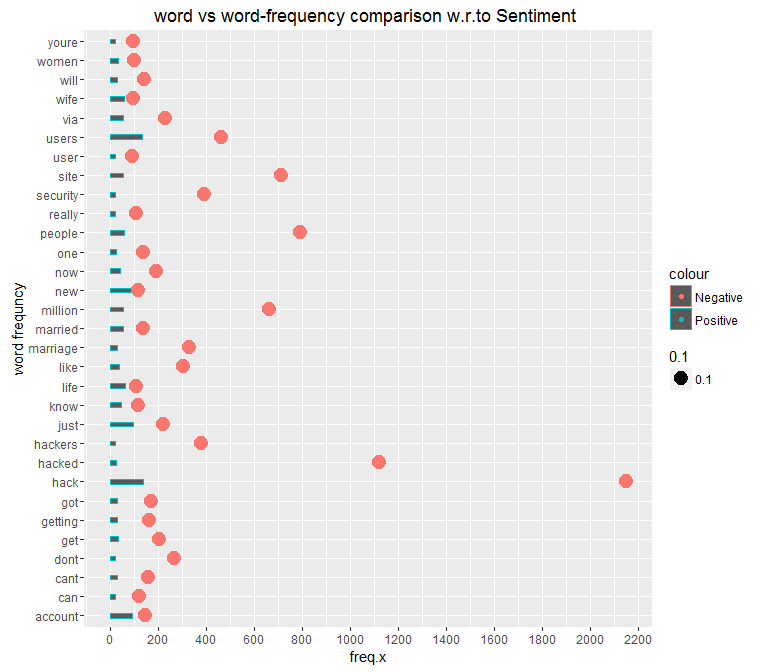
##### In the above plot we can see three different subplots one for each Gender 'F','M' and missing Gender.

##### In a similar way we can explore the tweet counts across "Country","Region","City" etc. And we can move on to actual words in the tweets and what interesting things they reveal!



##### In the plot we have top-25 most frequent words for each category i.e: Green refers Positive , Red refers Negative, Blue refers Neutral sentiment. By observing the plot we can likely to infer the sentiment of a particular tweet based on the words it has!

##### We can formulate the question such that, is it possible to get the sentiment of a tweet based on the occurance of a particular word?, the plot below compares the frequencies of some of the common words that are present in both Positive/Negative sentiment tweets.



##### From the plot above Red refers Negative,Blue refers Positive and we can infer the sentiment based on the frequency ,for example consider the word "hack" and we see that it has bigger frequncy in Negative context than Postive context.

## Coding the Topic(Secret/Secuirty):

##### To this point our focus is on analyzing sentiment of the tweets and this research purpose is to code evrey tweet with a Topic (Secret/Security/Other), so we create a new variable "Topic".since we have tweets of multiple thousands to be coded it's better to model/train using Machine Learning algorithm for the labelling process on a chunk of tweets(1000 sampled tweets) which are handcoded(manual Topic coding).And let the trained model predict the "Topic" for the rest of the dataset.

##### For the Topic prediction we used SVM(support vector machines) algorithm,it is a supervised machine learning algorithm and input to this algorithm will be words from the tweets and output will be "Topic" and technical details are below.

### SVM Model Training:

##### "In order to train Machine Learning model for predicting the relevant "TOPIC" we make use of "Contents" column in the dataset i.e; raw tweets. And the tweets from the users are untidy for analysis like they include special characters/symbols,un-wanted spaces,case discrepancies etc. as well as the tweets need to be split into one gram words to form a 'Document Term Matrix' for modeling purposes. The 'Document Term Matrix' is a matrix where rows represent the individual tweets and columns are one gram words with word frequency as matrix values. This task of creating 'Document Term Matrix' is achieved with the "tm"(text mining) package in R and it has several built in functions for data cleansing like tm\_tolower(),tm\_stripwhitespaces() etc. There is a point to be noted here, 'Document Term Matrix' is a sparse matrix because the columns are words ,there will be thousands of words from tweets and the rows are for individual tweets , a single tweet can't have all the words(column names) so most of the matrix will be filled with '0' word frequencies. To handle the sparsity in the data we can use 'removeSparseTerms()' a R function by specifying a limiting value.And as a final step we need to column bind the target variable 'Topic' to the 'Document Term Matrix' so that we have a cleaned dataset for modeling task with column 'Topic' as 'dependent variable' and remaining columns as 'independent variables'".

##### SVM is a state of the art algorithm for prediction and it performed well on interms of accuracy.It has achieved ~80% accuracy in the predicting Topic for the Ashley Madison incident, and ~95% accuracy for the Celebrity photo hack incident.Topic coding Work is in progress for Edward snowden and FBI-APPLE encryption incidents.

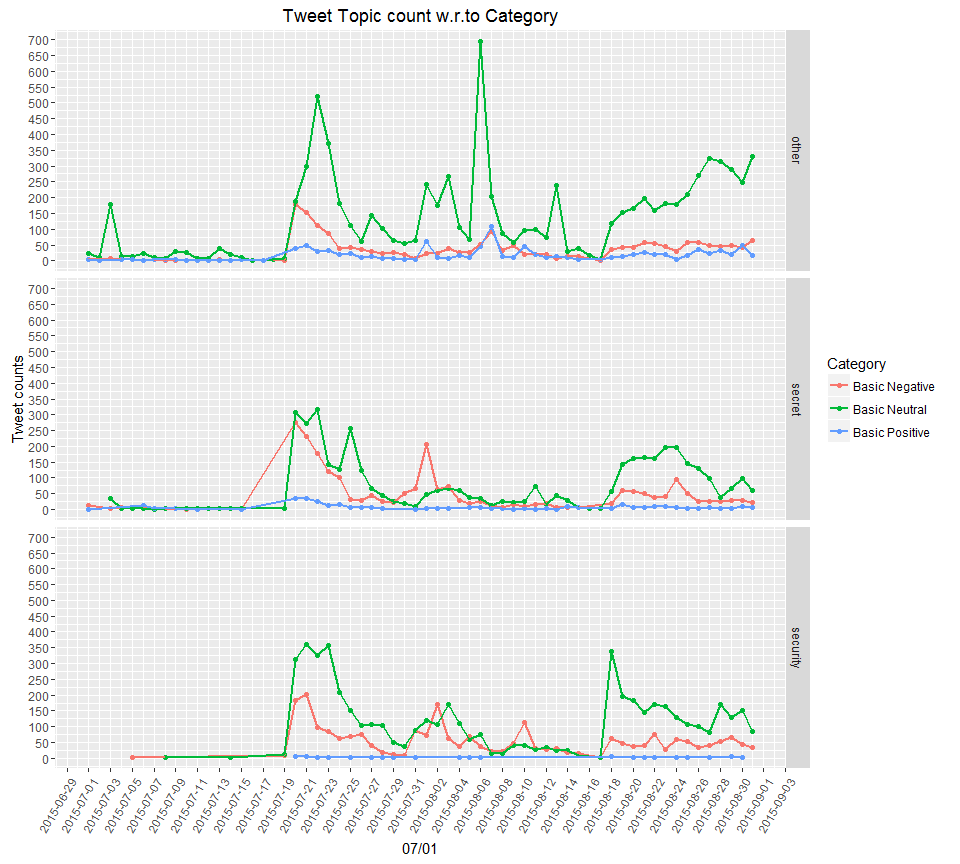
##### Now we have Topic coded data sets for Ashley Madison ,Celebrity photo hack incidents, so we can explore the data in the research direction i.e: analysing the Topic w.r.to Several other variables like Date,Category,Gender etc.

## Exploration of Topic coded data:

##### Lets explore the with newly created Topic variable by using plots.

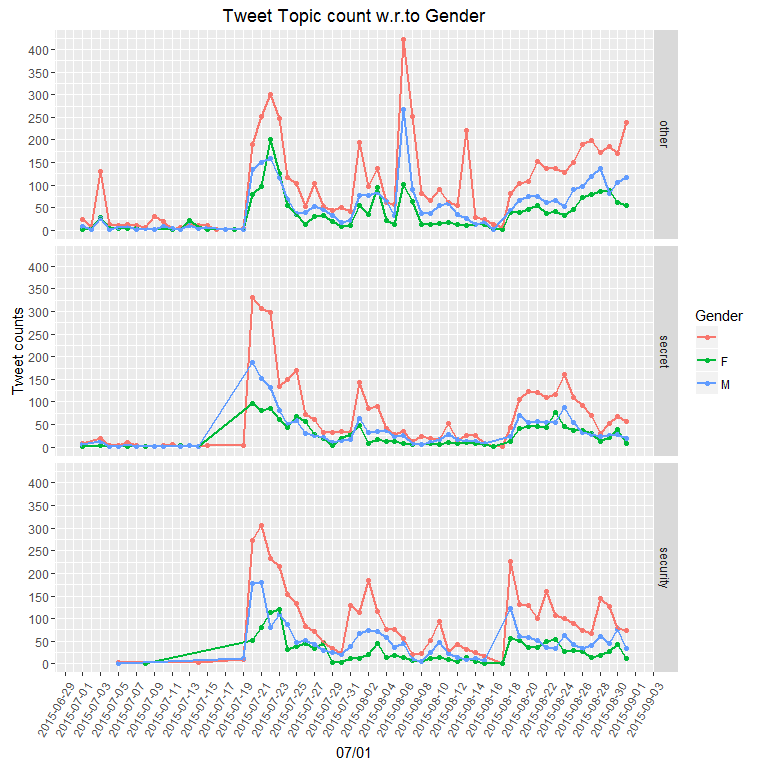
##### Plot for tweet counts for Topic w.r.to Category for the 60-days.

## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



##### Similarly we can obtain the Tweet counts for the Topic w.r.to Gender as below

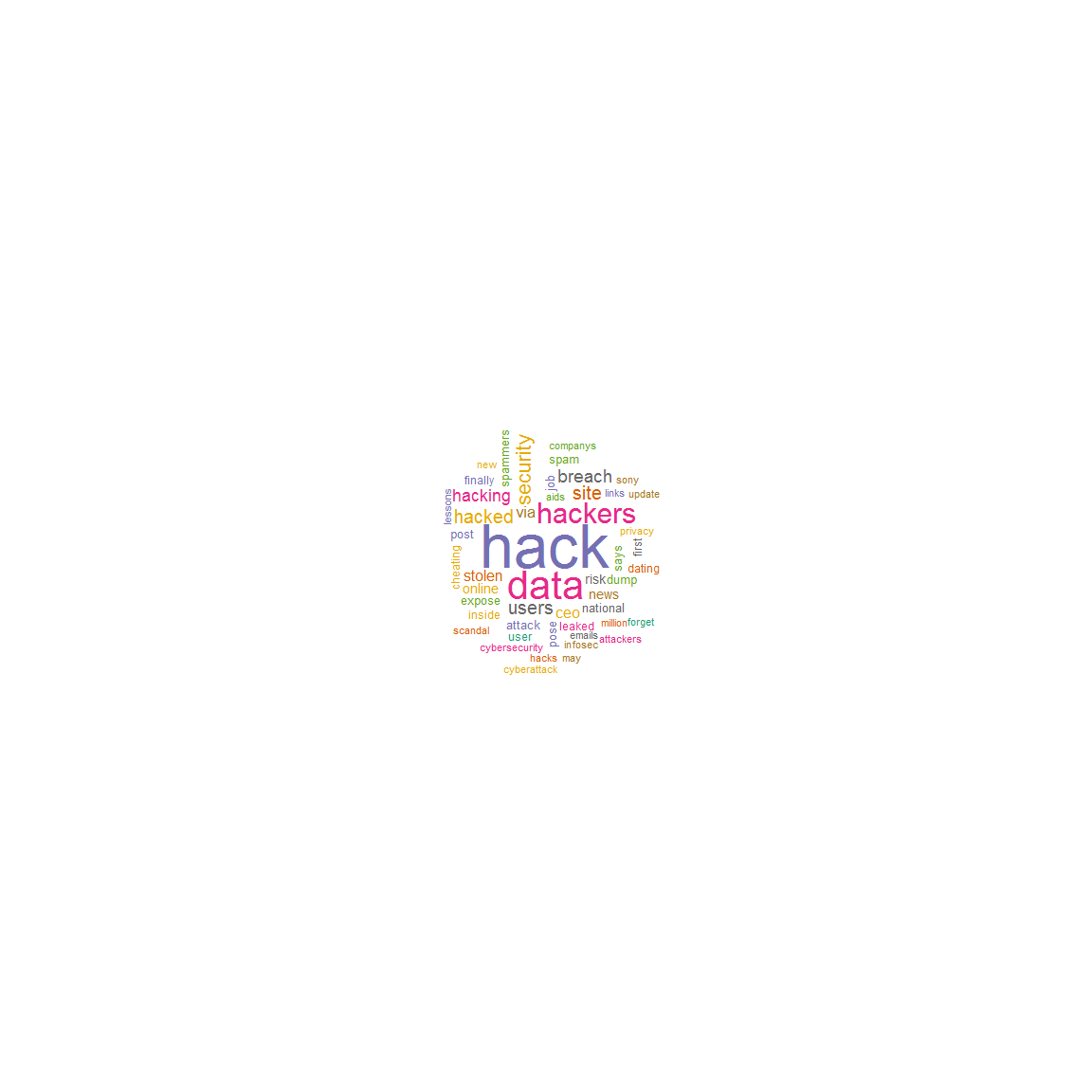
## Scale for 'x' is already present. Adding another scale for 'x', which  
## will replace the existing scale.



##### There is something interesting, we can have a look at the words which are most frequent in the tweets that are coded with Topic as "Secret","Security"

##### word cloud for Topic "Security"

## hack data hackers security hacked   
## 3420 2093 1390 858 733   
## site users breach hacking via   
## 712 707 700 598 527   
## stolen ceo online news risk   
## 482 465 452 378 342   
## job expose national post says   
## 328 321 319 315 312   
## spam attack pose spammers dump   
## 307 305 303 292 277   
## leaked finally cheating dating inside   
## 270 268 267 255 254   
## user first cybersecurity infosec scandal   
## 247 242 210 208 208   
## privacy million hacks update emails   
## 188 186 185 182 178   
## links new may companys cyberattack   
## 177 174 173 171 163   
## attackers sony forget aids lessons   
## 161 161 158 157 155

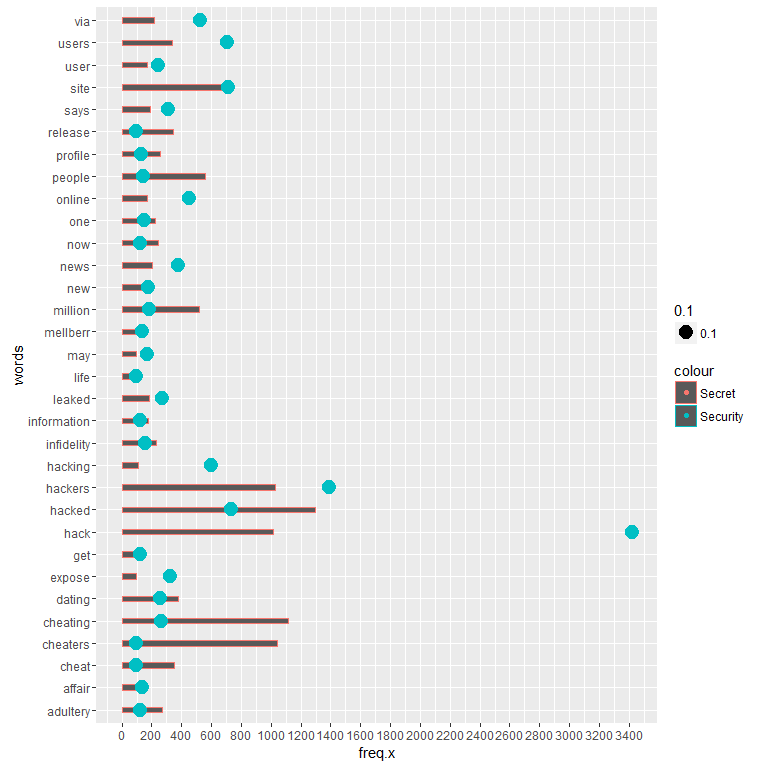


##### word cloud for Topic "Secret"

## hacked cheating cheaters hackers hack site   
## 1296 1117 1041 1027 1014 679   
## people account million list josh dating   
## 556 523 521 416 396 377   
## info duggar cheat release users paid   
## 361 355 354 341 336 279   
## exposed names adultery man will profile   
## 277 274 268 268 265 259   
## just now infidelity one married threaten   
## 250 241 230 226 223 220   
## via news wife used know adult   
## 216 205 201 195 192 188   
## says gets erase leak leaked star   
## 188 187 184 180 180 179   
## dont spouses information accounts got online   
## 177 177 174 171 171 168   
## new user   
## 167 167



##### We can compare the words between Topics "Secret","Security" as below



##### In the above plot Blue refers Security ,Red refers Secret.And we can compare the frequencies for word "adultery" it has bigger frequency for Secret context and lower frequency for Security context.Where as the word "hack" has the frequencies vice versa to the "aultery".

##### We can do much deeper analysis by creating several new variables so that we can infer more about the data. At the end Full data set has the additional columns such as "Topic"(the predicted column using the SVM model),"Hashtags", "Urls", "At\_tags", "About\_Privacy" and "is\_RT" are created for explicitly for data exploration purposes.

##### some more visulations are in the below pdf these are self explanatory.