

Exploration - RT

Bryce Q

February 24, 2018

```
set.seed(72) #setting seed right away

tweets <- read.csv('tweets.csv', header = TRUE)
users <- read.csv('users.csv', header = TRUE)
```


NOTE

This is currently a work in progress. The goal is focused on exploration and I try to throw time into it and try new things as I'm available to with grad school. All below this is subject to change.

Also note: this is purely exploratory. None of this is meant to be taken politically, the goal is just to explore a fun data set and try to find some interesting things. I tried to keep any of my biases external to the analysis when possible to focus on a more clear cut analysis. Additionally, I did make some assumptions around this data set being legitimate in what it claims to describe (all these tweets being from malicious accounts) but did not do a thorough analysis of how they came these conclusions.

This data set is flawed as it does not compare American 'Trolls' and other groups who were trying to influence the election. It also singularly focuses on the Russian 'troll' tweets, ignoring who the population of twitter was acting as a whole relative to the election. Ideally we would have some control group of other accounts (not related to these accounts) trying to influence the election who we could compare but as this is more a fun analysis than serious political work, we will move forward with the available data set and some strong cautions around the external validity of this analysis.

```
head(tweets, 5)
```

user_id 															
<dbl>															
1868981054															
2571870453															
1710804738															
2584152521															
1768259989															

5 rows | 1-1 of 16 columns

```
head(users, 5)
```

	id 
	<dbl>
	18710816
	100345056
	247165706
	249538861
	449689677

5 rows | 1-1 of 14 columns

First I wanted to look at the users and tweets. Everything seems pretty self explanatory based off the column names and data types.

```
#results are hidden in RMD to shorten document length

describe(tweets)

summary(tweets)
```

Notice we do have missing data around userID, createdAt, favoriteCount, tweetId, retweeted_status_ID and in_reply_to_status_id. I'm going to assume if there's missing for the reply/favorite items, then there was not favorite or it was not a reply and its a valid NA. As for the userID data... maybe the users were deleted since the tweets were created? Or maybe they just aren't included in this data set for whatever reason (Private, etc). I'm not a big twitter user or expert so it may be something relevant to look into.

Additionally, the createdAt data is needed since we're focusing on doing a time analysis. I'm going to take a quick look at the ones missing time.

```
notime <- tweets %>% filter(is.na(tweets$created_at))

unique(as.character(notime$user_key))
```

```
## [1] "luke_jones13"      "scottgohard"      "warfareww"        "blacknewsoutlet"
## [5] "todayinsyria"     "jenn_abrams"      "blk_voice"         "redlanews"
```

```
notime$text
```

```
## [1]
## 174986 Levels: ...
```

We can see its 9 different users but oddly enough, no text of the tweets exist. I'm guessing these have since been deleted. I checked a few of the accounts and they've been suspended (no surprise) so we are going to filter these ones out for sake of keeping our analysis clean.

```
remove(notime) #keeping it clean
tweets <- tweets %>% filter(!is.na(tweets$created_at))
```

A few other house keeping cleaning items...

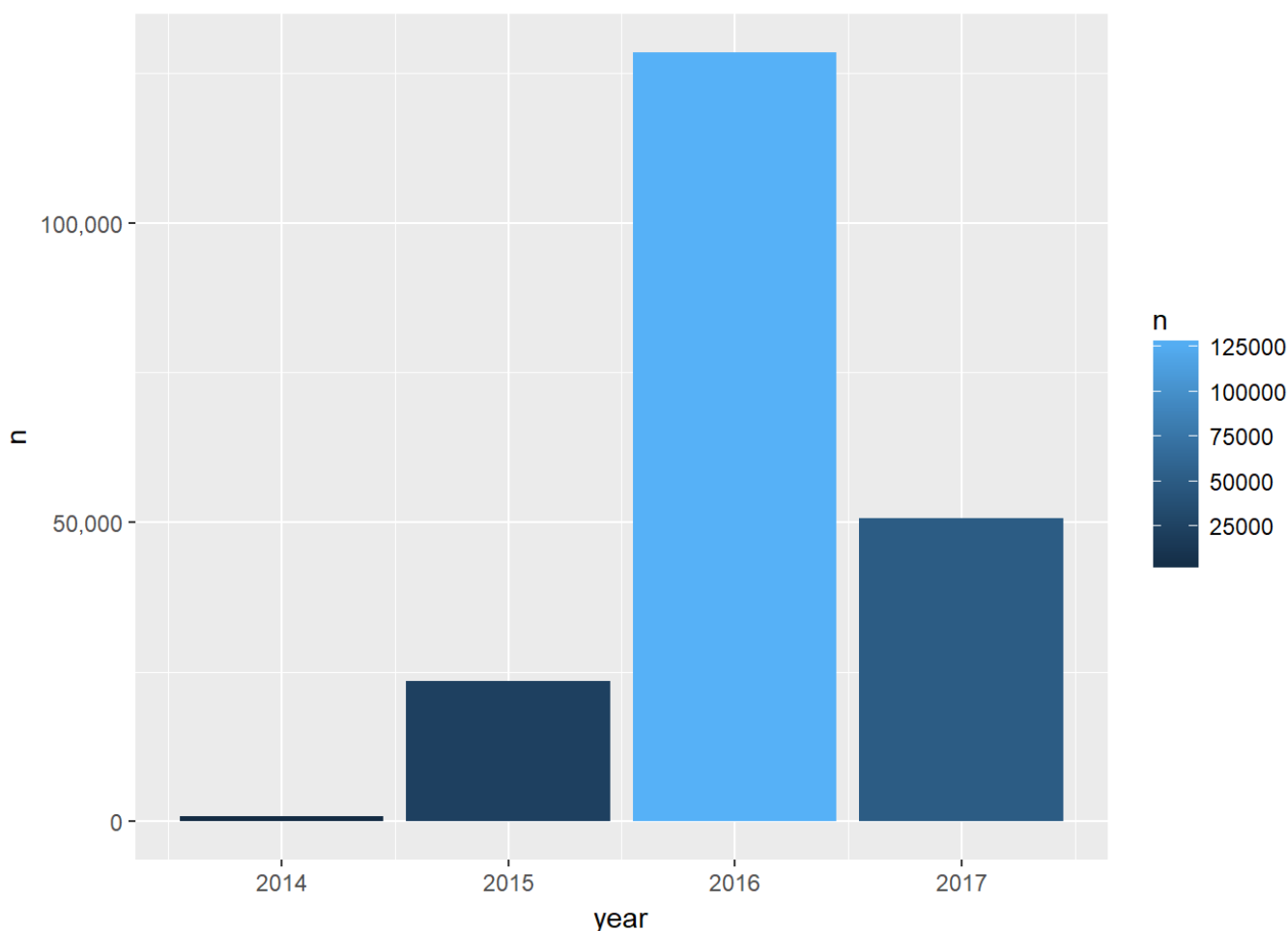
```
tweets$favorite_count[is.na(tweets$favorite_count)] <- 0  
tweets$retweet_count[is.na(tweets$retweet_count)] <- 0
```

We also need to clean the dates and times.

```
tweets$created_str <- as.Date(tweets$created_str, '%Y-%m-%d %H:%M:%S')  
tweets$month_year <- floor_date(tweets$created_str, "month")  
tweets$year <- year(tweets$created_str)
```

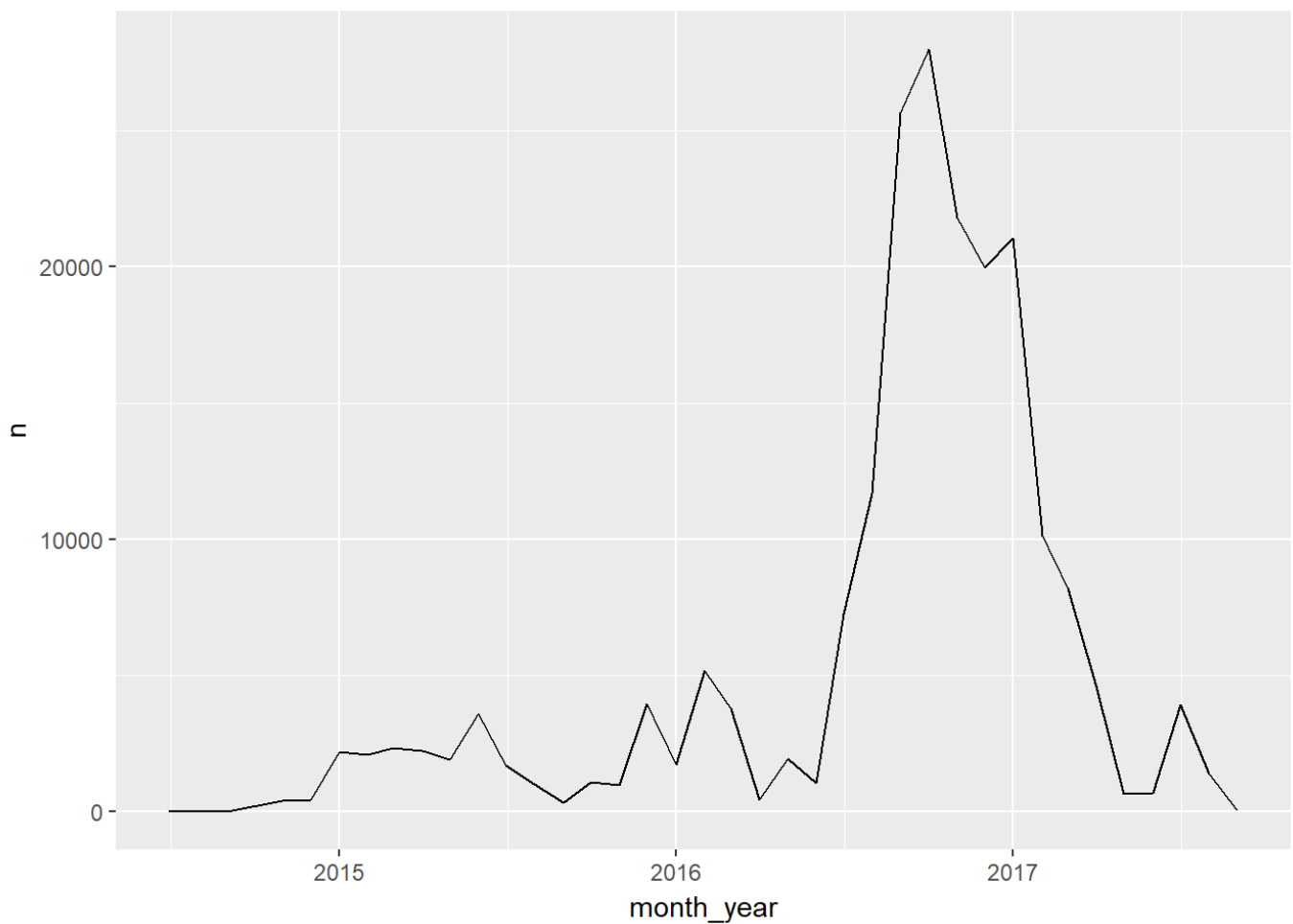
Initial Analysis

```
tweets_year <- tweets %>% group_by(year) %>% tally()  
ggplot(tweets_year) + geom_bar(stat='identity', aes(year,n, fill = n)) + scale_y_c  
ontinuous(labels = comma)
```



First, let's look at the volume of tweets over time. 2016 is the most popular year, followed by 2017 (though many of these accounts started to get shutdown in 2017 I believe towards the tail end, resulting in decreased activity).

```
remove(tweets_year)  
tweet_month <- tweets %>% group_by(month_year) %>% tally()  
ggplot(tweet_month) + geom_line(aes(month_year,n))
```



We can see a ton of activity occurring right around election time which isn't surprising, but also carrying forward the months after election into 2017 then rapidly dropping off.

Next thing we should do is look at some of the top and bottom tweets to try to get a little more domain knowledge on what's happening.

```
#Let's examine the most popular tweets
top_20_favorite <- arrange(tweets, favorite_count) %>% tail(20) %>%
  select(user_key,created_str,month_year,retweet_count,favorite_count,text,hashtags
,mentions)
top_20_favorite
```

user_key <fctr>
ten_gop
crystal1johnson
ten_gop
gloed_up
southlonestar
pamela_moore13
crystal1johnson
gloed_up
ten_gop
ten_gop

1-10 of 20 rows | 1-1 of 8 columns

Previous **1** 2 Next

Interesting! Off the bat we have all over the board stuff; a lot of Hillary related items, quite a few tweets related to black lives masters well as Donald Trump Interesting thing here is you see a lot of users have multiple tweets in the top 20; ten_GOP alone has 6 of the top 20 favorited tweets.

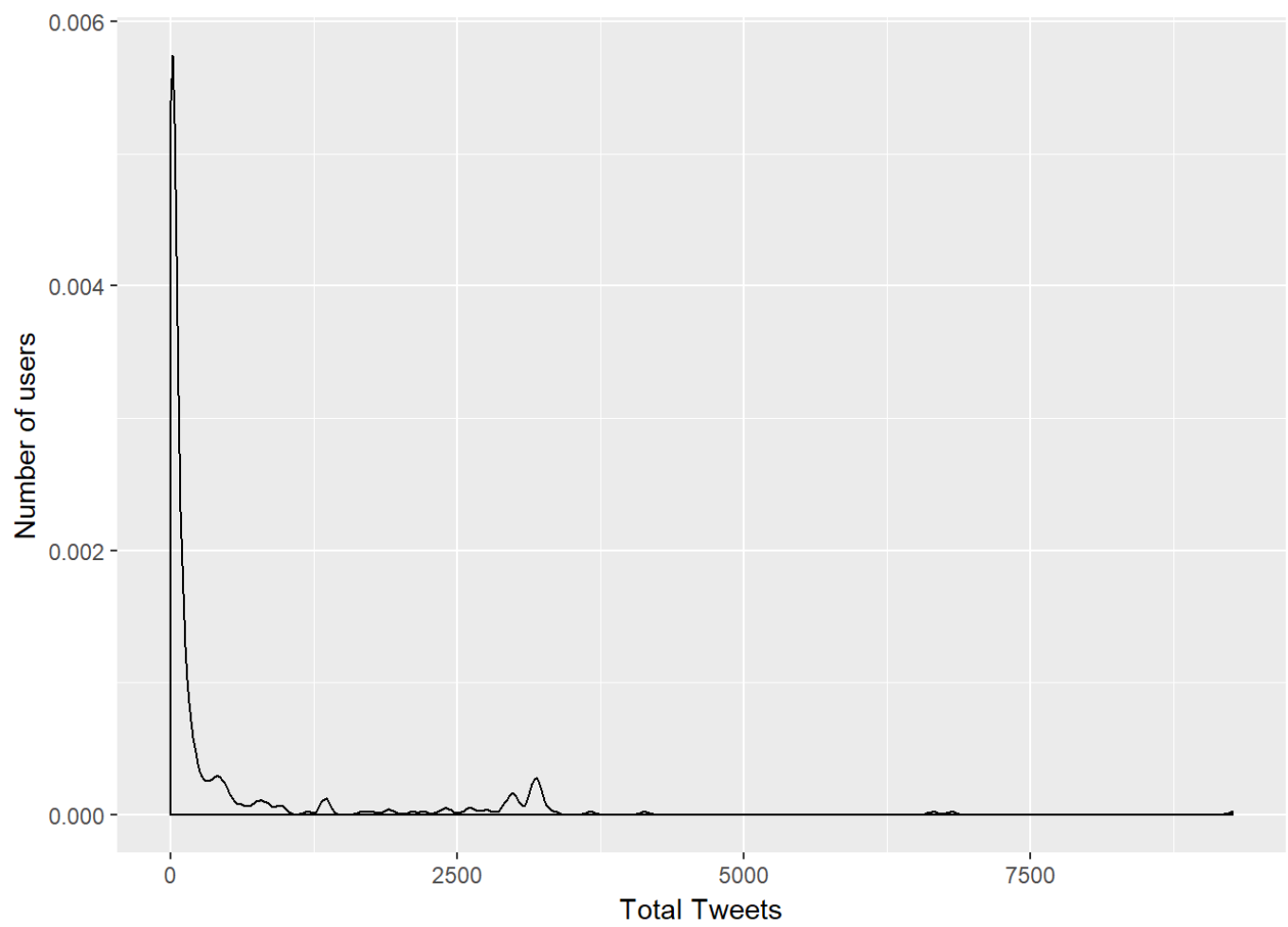
Speaking of this, lets look at users who have the most likes/re-tweets. This will be the join where we do summaries to get the sum of favorites/re-tweets/etc

```
top_user <- left_join(tweets,users,by=c('user_key'='name'))
```

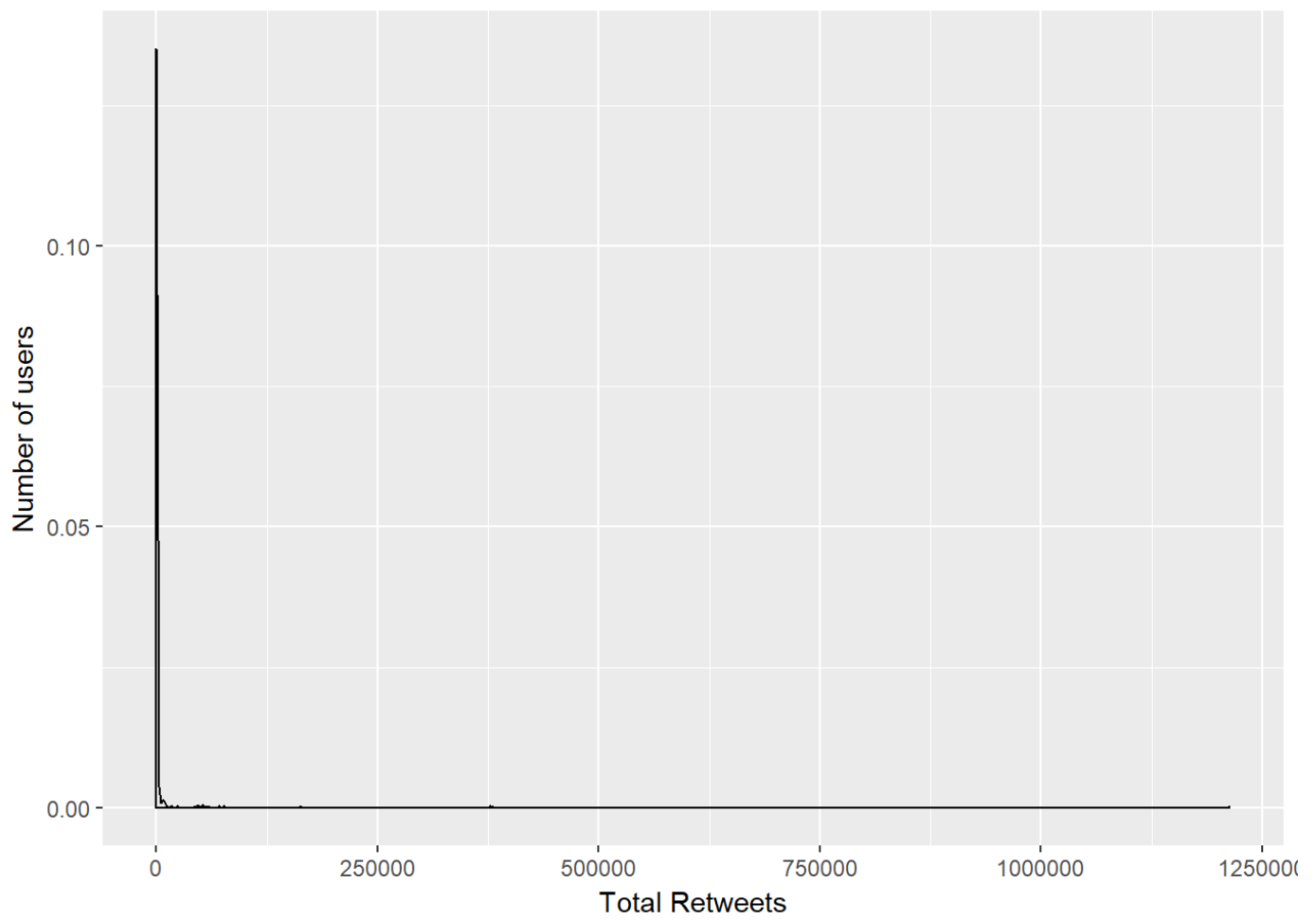
```
## Warning: Column `user_key`/`name` joining factors with different levels,
## coercing to character vector
```

```
top_user <- top_user %>% group_by(user_key) %>%
  summarise(total_fav = sum(favorite_count),
            total_rt = sum(retweet_count),
            total_tweets = n()
  )

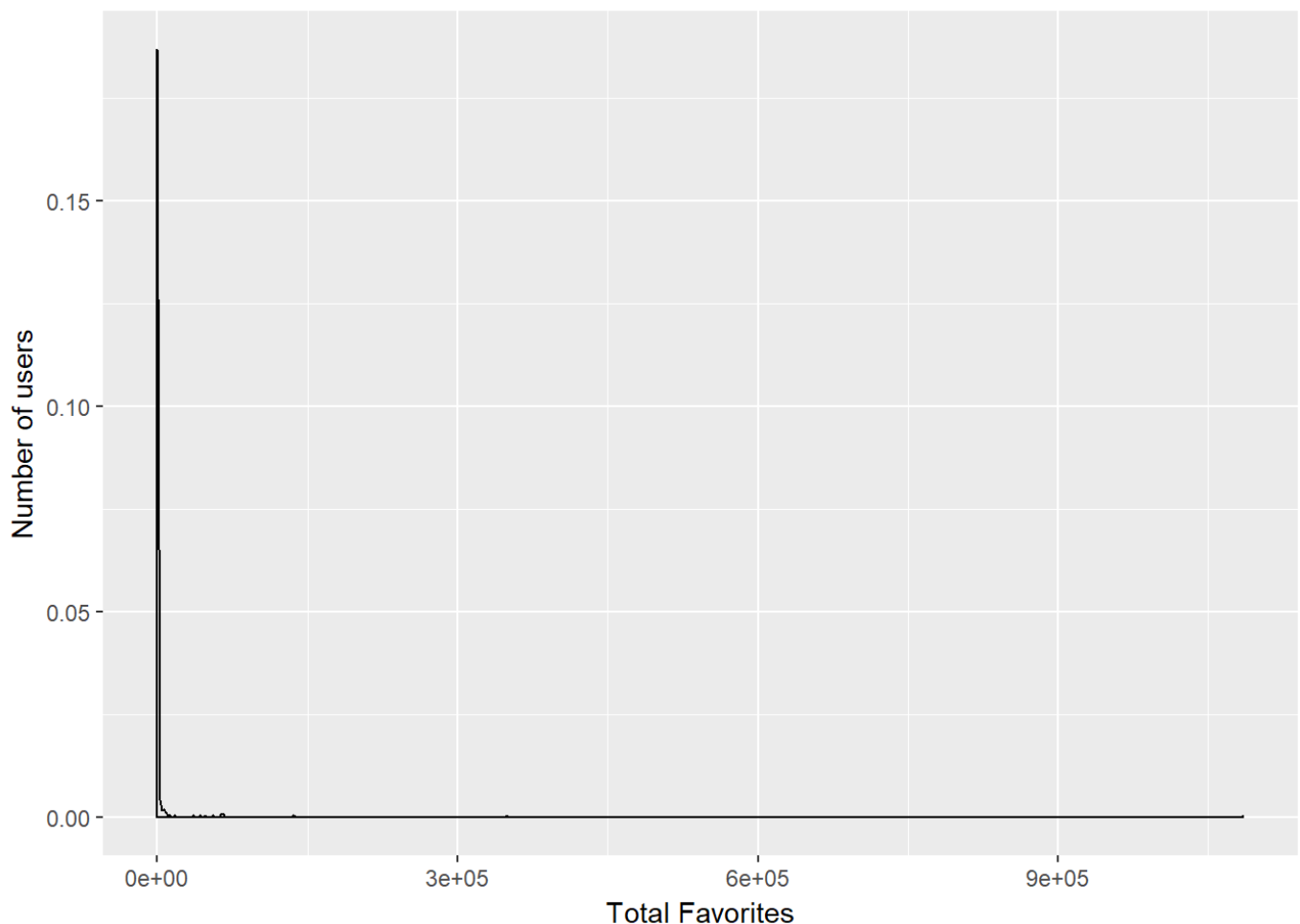
ggplot(top_user) + geom_density(aes(total_tweets)) + xlab('Total Tweets') + ylab('
Number of users')
```



```
ggplot(top_user) + geom_density(aes(total_rt)) + xlab('Total Retweets') + ylab('Number of users')
```



```
ggplot(top_user) + geom_density(aes(total_fav)) + xlab('Total Favorites') + ylab('Number of users')
```



The above density plots are pretty ugly but it shows us something important; most of the users were relatively unknown and there were outliers to ones which were very successful. We'll see users with 1000s of tweets and not a single re-tweet or favorite; this mimics the actual usage of twitter where there is tons of traffic and a few actual very popular tweets.

To iterate this let's do a really simple calculation to find the most popular accounts by activity.

```
top_user$total_act <- top_user$total_fav + top_user$total_rt
top_user$act_ratio <- top_user$total_act / top_user$total_tweets
top_user <- arrange(top_user, act_ratio)
tail(top_user, 5)
```

user_key

<chr>

pamela_moore13

trayneshacole

ten_gop

crystal1johnson

williams_diana_

5 rows | 1-1 of 6 columns

In the above we count an 'activity' as a re-tweet or a favorite. We take the total number of 'activity' divided by the total tweets to get the activity per tweet ratio. We can see a few of those same accounts we saw in our top most favorited tweets are here with the highest activity per tweet ratio.


```
length(top_user$user_key[top_user$act_ratio == 0])
```

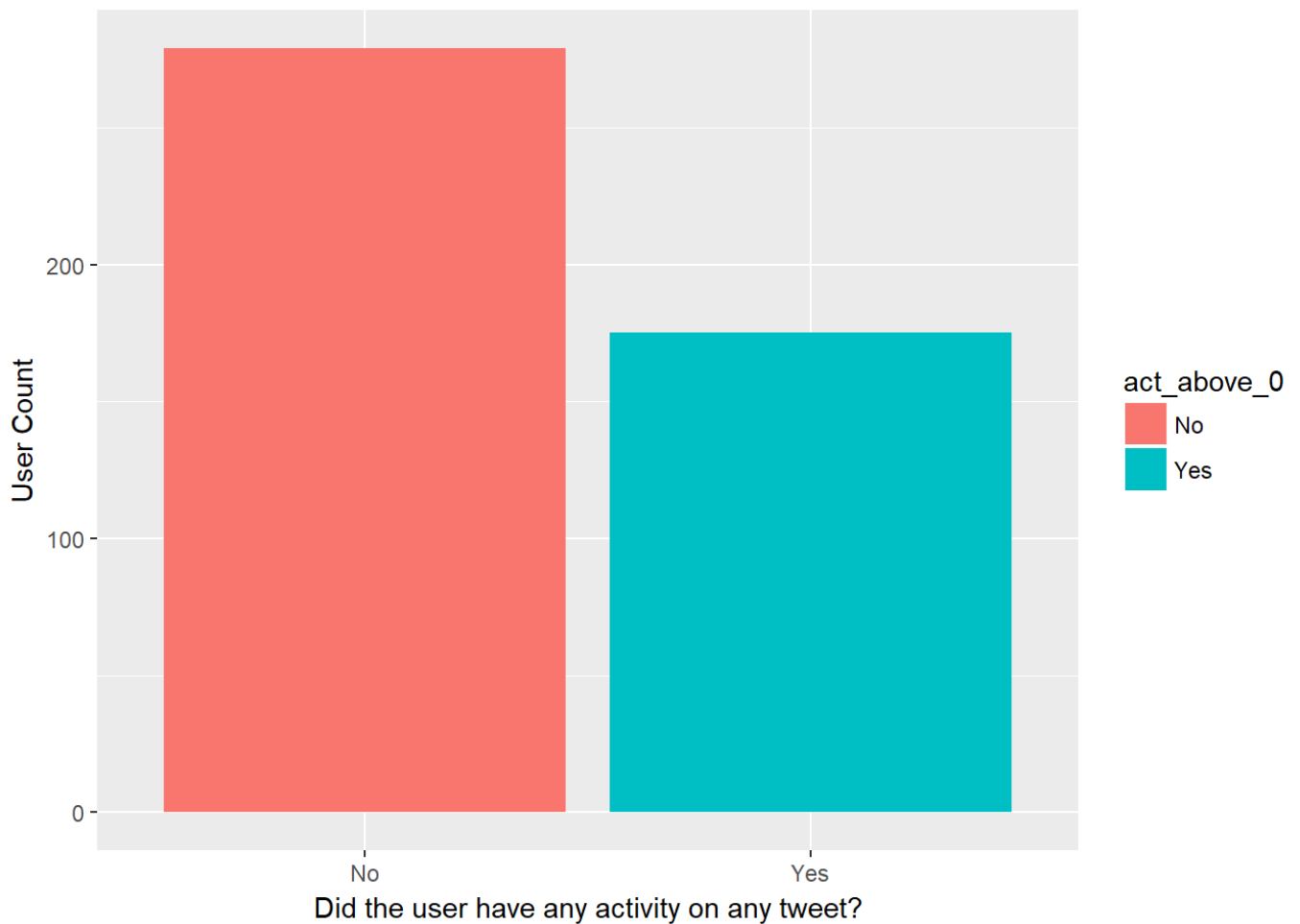
```
## [1] 279
```

```
sum(top_user$total_tweets[top_user$act_ratio == 0])
```

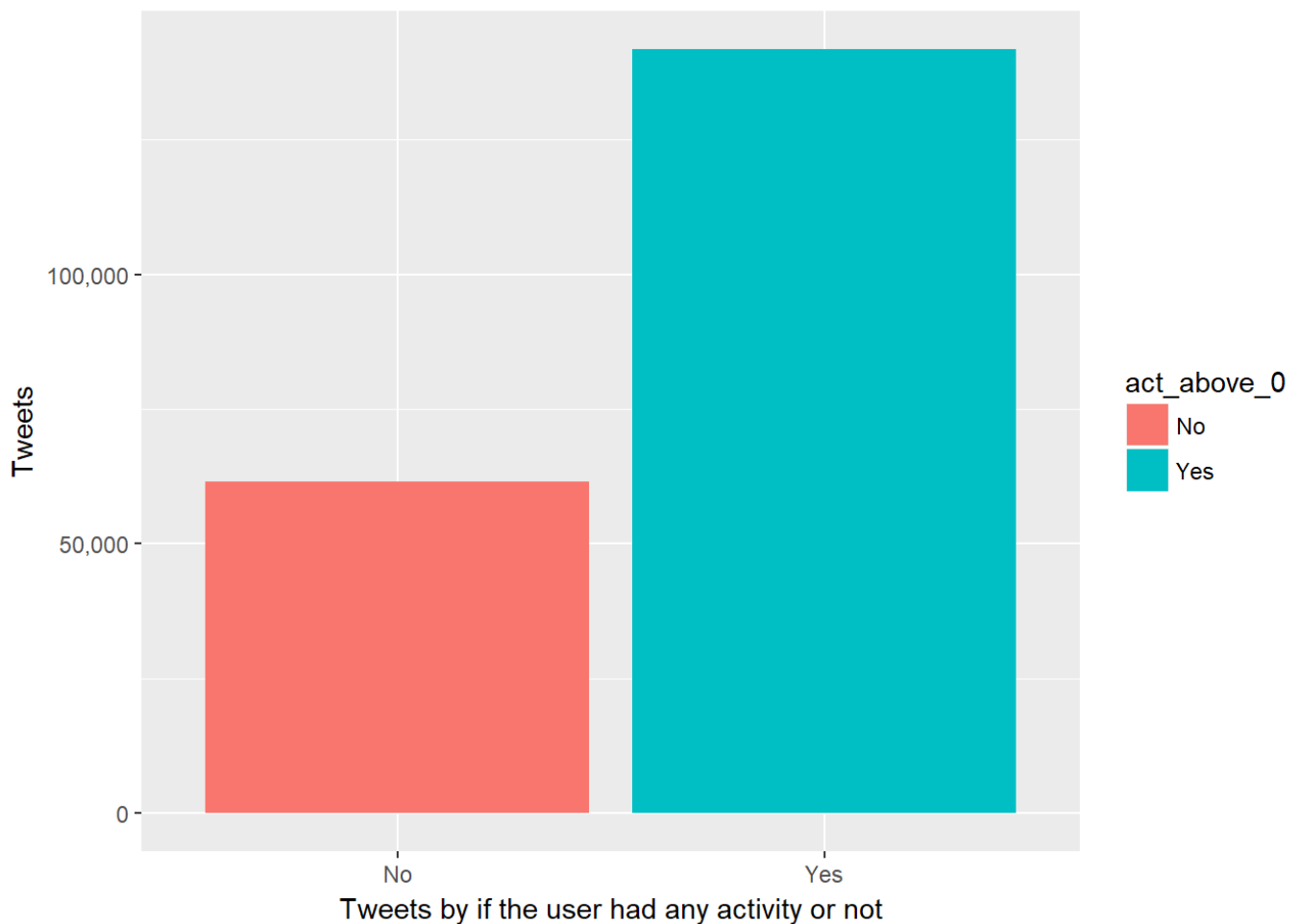
```
## [1] 61645
```

```
top_user$act_above_0 <- ifelse(top_user$act_ratio == 0, 'No', 'Yes')
```

```
ggplot(top_user) + geom_bar(aes(act_above_0, fill=act_above_0)) + xlab('Did the use  
r have any activity on any tweet?') + ylab('User Count')
```



```
ggplot(top_user) + geom_bar(stat='identity', aes(act_above_0, total_tweets, fill=act_  
above_0)) + xlab('Tweets by if the user had any activity or not') + ylab('Tweets')  
+ scale_y_continuous(labels = comma)
```



Additionally, we can see there were 279 users who had NO activity whatsoever, and still created over 61,645 tweets! That's a little shy of a third of all the tweets coming from users who never had ANY activity whatsoever. Still though, that means more than 2/3s the tweet came from accounts that were active with other twitter accounts in some way or another.

Topic Modeling

So let's get to the core of this analysis: Topic modeling! I thought it would be interesting to run latent Dirichlet allocation on this data set, which finds the underlying topics within the entire corpus of tweets, and then allows us to label tweets specifically by their topic. This is interesting since it helps us understand what topics were being discussed in what frequency relative to what was occurring during that timeline. The results may not be too surprising but it would help us understand the topics that were emphasized to influence the election relative to what was going on in the greater political atmosphere.

Note: Due to the nature we already discussed, the topics focused on may not represent the actual influence. It is very apparent based of re-tweets and favorites the quality of the tweet matters more than the quantity. That being said, let's see if topic modeling results make any sense.

```

#creating a corpus of tweets
tweet_doc <- Corpus(VectorSource(as.character(tweets$text)))

#removing punctuation
tweet_doc <- tm_map(tweet_doc, removePunctuation)
#removing numbers too
tweet_doc <- tm_map(tweet_doc, removeNumbers)
#putting it all to lower case
tweet_doc <- tm_map(tweet_doc, content_transformer(tolower))
#removing stop words with no specific topic
tweet_doc <- tm_map(tweet_doc, removeWords, stopwords("english"))
#remove any irrelevant white space
tweet_doc <- tm_map(tweet_doc, stripWhitespace)

```

The above is just preparing to run topic models. We are stripping out all the things we don't care about; punctuation, numbers, stop words, white space etc. Note I did not stem the words in this case as I wanted to see the full outputs of different words due to the nature of things like twitter handles (real vs realDonaldTrump is an important distinction, etc).

```

set.seed(72)
tfm <- DocumentTermMatrix(tweet_doc)

#removing sparse items in the matrix
tfm<-removeSparseTerms(tfm, sparse=0.999) #since these are stemmed tweets i'm expecting to have a lot of sparseness
#setting the bar low for non-sparse items

#removing sparse rows
ui = unique(tfm$i)
tfm = tfm[ui,]

print(dim(tfm))

```

```
## [1] 191101 1333
```

Above is our term frequency matrix. We use this term frequency matrix to run LDA on to find our topics. Note that I removed all rows without terms in the term frequency matrix. Additionally, I removed the really scarce terms to hopefully decrease the size of the matrix and increase the speed at which this runs. Removing scarce terms means we can't generalize to all tweets (about 9% will drop out) but given twitter is not traditional and pure format for words since it can include pictures/images and a lot of non-traditional text (emoji's etc) I think this was an effective way to focus on the tweets that were most legible and relevant in regards to their text context.

```
set.seed(72)
results <- LDA(tfm, k = 12, method = "Gibbs")

w= 8
thresh = 0.01 #was .015
Terms12_nostem <- terms(results, w,thresh)

Terms12_nostem #this one is actually pretty solid
```

```
##      Topic 1      Topic 2      Topic 3      Topic 4 Topic 5
## [1,] "one"      "post"      "vote"      "now"      "say"
## [2,] "just"     "new"      "httpsâ<U+0080>"      "get"      "dont"
## [3,] "get"      "says"     "america"    "love"     "see"
## [4,] "midnight" "cnn"      "maga"      "good"     "can"
## [5,] "got"      "conservatexian" "tcot"      "right"    "like"
## [6,] "ðŸ<U+0098>"      "news"      "great"      "much"     "know"
## [7,] "said"     "trumps"   "realdonaldtrump" "life"     "want"
## [8,] "thats"    "via"      "pjnet"      "youre"    "think"
##      Topic 6      Topic 7      Topic 8      Topic 9      Topic 10
## [1,] "president" "never"    "amp"        "women"      "work"
## [2,] "years"     "media"    "httpsâ<U+0080>"      "black"      "muslim"
## [3,] "obama"     "will"     "stop"       "man"        "real"
## [4,] "time"      "american" "die"        "httpstcâ<U+0080>" "take"
## [5,] "first"     "people"   "please"     "white"      "back"
## [6,] "today"     "show"     "httpstcoâ<U+0080>" "house"      "country"
## [7,] "day"       "support"  "httpstâ<U+0080>" "police"     "world"
## [8,] "every"     "election" "must"       "blicqer"    "going"
##      Topic 11      Topic 12
## [1,] "trump"        "clinton"
## [2,] "video"        "bill"
## [3,] "watch"        "hillary"
## [4,] "says"         "debate"
## [5,] "politics"     "dnc"
## [6,] "live"         "campaign"
## [7,] "donald"       "fbi"
## [8,] "supporters"   "clintons"
```

After trying different thresholds and sizes, 12 seems to give us the clearest set of topics. Not every topic is perfect but I think we have a few ones that really stick out!

“Get/midnight/said”

```
Terms12_nostem[1:8]
```

```
## [1] "one"      "just"      "get"      "midnight" "got"      "ðŸ<U+0098>"
## [7] "said"     "thats"
```

```
top = c()
top[1] = 'Get/Midnight/Said'
```

This was the most generic of the topics discovered. There was not much value in it, but there is consistency

around things like get/got, said, etc. The one interesting point is the word midnight. As much as I want to try to summarize each topic, for this one I'm just going to list a few of the words as a title as at this point I'm having a hard time deriving any insights from it.

"Liberal Media relative to conservatives/Trump"

```
Terms12_nostem[9:16]
```

```
## [1] "post"          "new"           "says"          "cnn"
## [5] "conservatexian" "news"          "trumps"        "via"
```

```
top[2] = 'Liberal Media relative to conservatives/Trump'
```

This topic specific references a few liberal media sources (CNN/Washington post) but also along side a popular conservative twitter handle (@conservatexian) as well as Trumps (which would also be Trump's in this case due to how we cleaned our words). My guess is this is a discussion of liberal media sources from other perspectives (most likely conservative ones that support Trump).

"Make America Great Again"

```
Terms12_nostem[17:24]
```

```
## [1] "vote"          "httpsâ&U+0080>" "america"        "maga"
## [5] "tcot"          "great"          "realdonaldtrump" "pjnet"
```

```
top[3] = 'MAGA'
```

This topic is relatively simple; focused on trump slogans like #MAGA and including Trump's actual twitter handle as well as PJnet which is the 'patriotic journalist network' set up for voicing conservative opinions on twitter. This a pro-Trump hash tag topic.

"General Positive Words"

```
Terms12_nostem[25:32]
```

```
## [1] "now"  "get"  "love" "good" "right" "much" "life" "youre"
```

```
top[4] = 'General Positive Words'
```

This topic is generally focused on positive words (love, good, right, life).

"General Action Words"

```
Terms12_nostem[33:40]
```

```
## [1] "say"  "dont" "see"  "can"  "like" "know" "want" "think"
```

```
top[5] = 'General Action Words'
```

This topic is generally focused on action words (say, see, like, want, think).

“Obama & Time”

```
Terms12_nostem[41:48]
```

```
## [1] "president" "years"      "obama"      "time"      "first"      "today"
## [7] "day"        "every"
```

```
top[6] = 'Obama & Time'
```

This topic is focused on Obama (president, Obama) and also has a large constraint of time related words (today, day, year, time).

“American people & the election”

```
Terms12_nostem[49:56]
```

```
## [1] "never"      "media"      "will"      "american"  "people"    "show"
## [7] "support"    "election"
```

```
top[7] = 'American people & the election'
```

This topic is focused on the American people, as well as the election. Show/support/election are definitely action words used by a side to promote their candidate. That being said the interesting one here is never and media. My initial thoughts are its referring to the media negatively, and saying the American people will decide the election but that may be over playing my own thoughts so I'll keep the topic at American people & the election.

“Imperative call for action”

```
Terms12_nostem[57:64]
```

```
## [1] "amp"          "httpsâ&U+0080>" "stop"        "die"         "please"
## [6] "httpstcoâ&U+0080>" "httpstâ&U+0080>" "must"
```

```
top[8] = 'Imperative call for action'
```

Another really interesting one, this topic uses very strong words (stop, die, must) that when combined with please make me think its imperative tweets, often request for action. With the combination of die in there, this may be a topic strongly related to either police brutality or overseas military conflict based off my initial readings of the tweets so far.

“Race & Gender”

```
Terms12_nostem[65:72]
```

```
## [1] "women"      "black"      "man"        "httpstcâ&U+0080>" "white"      "house"
## [7] "police"     "blicqer"
```

```
top[9] = 'Race & Gender'
```

This topic is definitely related to race and gender in America. In addition to race (black/white) we also see police and @blicqer (popular black news twitter handle).

“Global topics and nationalism”

```
Terms12_nostem[73:80]
```

```
## [1] "work"      "muslim"    "real"      "take"      "back"      "country"   "world"
## [8] "going"
```

```
top[10] = 'Global topics and nationalis'
```

This is a global topic, definitely focused on American foreign relations and nationalism. Muslim and world are definitely global words. Work, take back and real all sound vaguely nationalistic as well.

“Trump communication to supporters”

```
Terms12_nostem[81:88]
```

```
## [1] "trump"      "video"      "watch"      "says"      "politics"
## [6] "live"       "donald"     "supporters"
```

```
top[11] = 'Trump communication to supporters'
```

This topic encompasses Trump, then a lot of words related to communication (live, watch, video, says) and his supporters. My guess is these tweets are about Trump addressing supporters, at rallies or via another form of communication.

“Clinton’s campaign and FBI”

```
Terms12_nostem[89:96]
```

```
## [1] "clinton"    "bill"      "hillary"    "debate"    "dnc"      "campaign"
## [7] "fbi"        "clintons"
```

```
top[12] = "Clinton's campaign and FBI"
```

This is massively interesting to me. Hillary or any form of Hillary Clinton where only found in a single topic while Trump was separate from his twitter handle (Hillary’s handle was very absent compared to Trump’s). Additionally, this is a very strong collection of words and was the easiest to interpret.

Now that we have our topics lets do some further analysis!

Let’s look at volume by topic, but even more relevant I think would be looking at how the topics adjusted over time and what topics were being pushed during whats points in the election. Before that though, I’m going to add in the topics to the original tweet data set so we can manipulate the data with the full set of features.

```

#find the tweets we used
index_tweets <- data.frame(id = as.numeric(tfm$dimnames$Docs), used = 'yes')

#create counting index in tweets
tweets_used <- tweets %>% mutate(id = seq.int(nrow(tweets)))

#merging index to tweets used. any NA tweets were not used in our analysis and will
be filtered out
tweets_used <- left_join(tweets_used, index_tweets, by='id')
tweets_used <- tweets_used %>% filter(used == 'yes')

```

First I narrowed down my list of tweets to just those that were used to build our term-frequency matrix (ie: those that stayed on topic). This was 191k out of 208k so it was over 90% of the total tweets.

```

tweets_used$tweet_top <- topics(results, 1)

top_df <- data.frame(index = seq.int(12), topic = top)

tweets_used <- left_join(tweets_used, top_df, by=c('tweet_top'='index'))

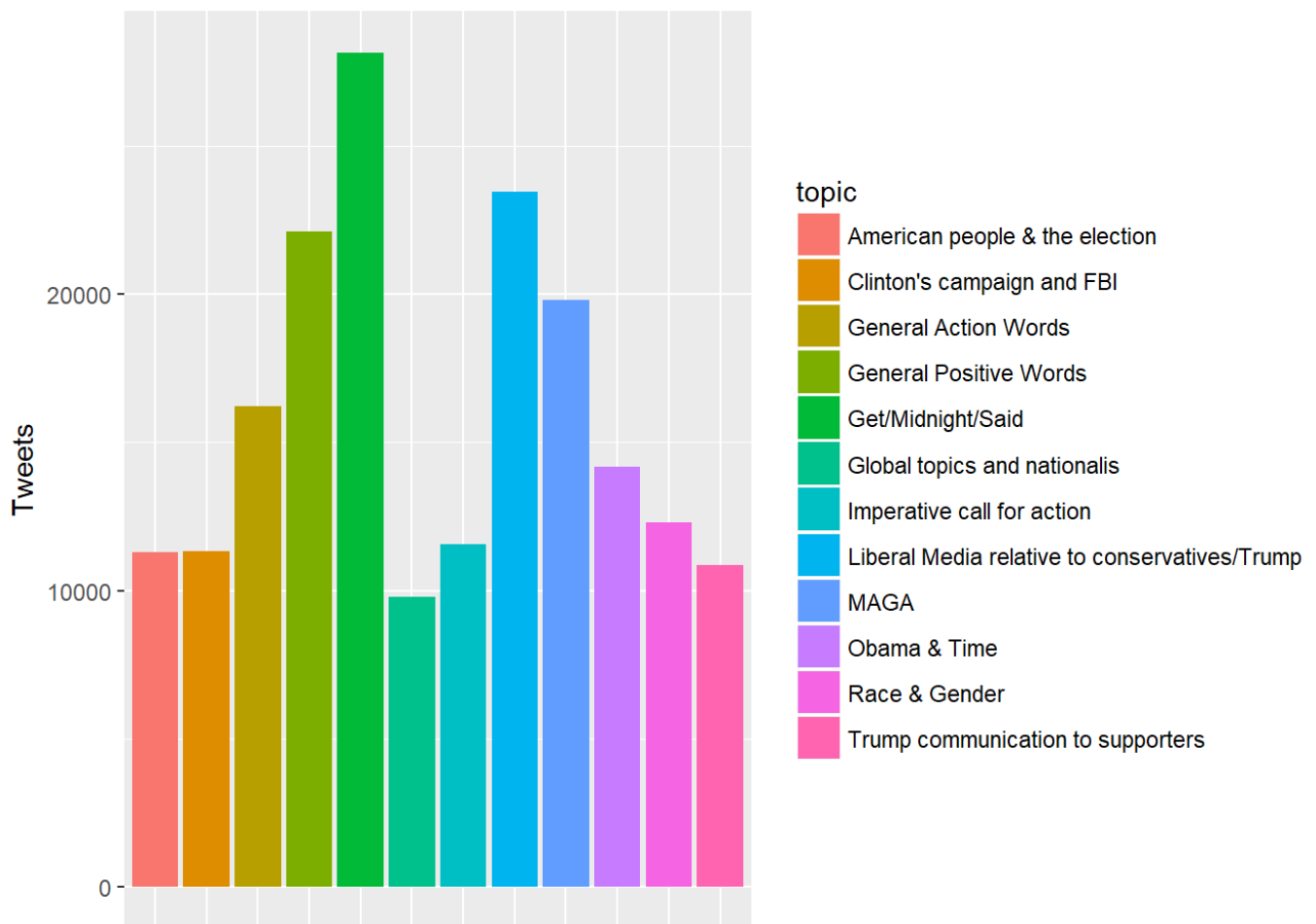
```

The above gives us the topic for each document out of 191,101 elements. We then merge that in with each corresponding tweet and can see the primary topic of each tweet.

```

ggplot(tweets_used) + geom_bar(aes(topic, fill = topic)) + theme(axis.text.x=elem
ent_blank(),
                                                                    axis.ticks.x=el
ement_blank(),
                                                                    axis.title.x=el
ement_blank()) + ylab('Tweets')

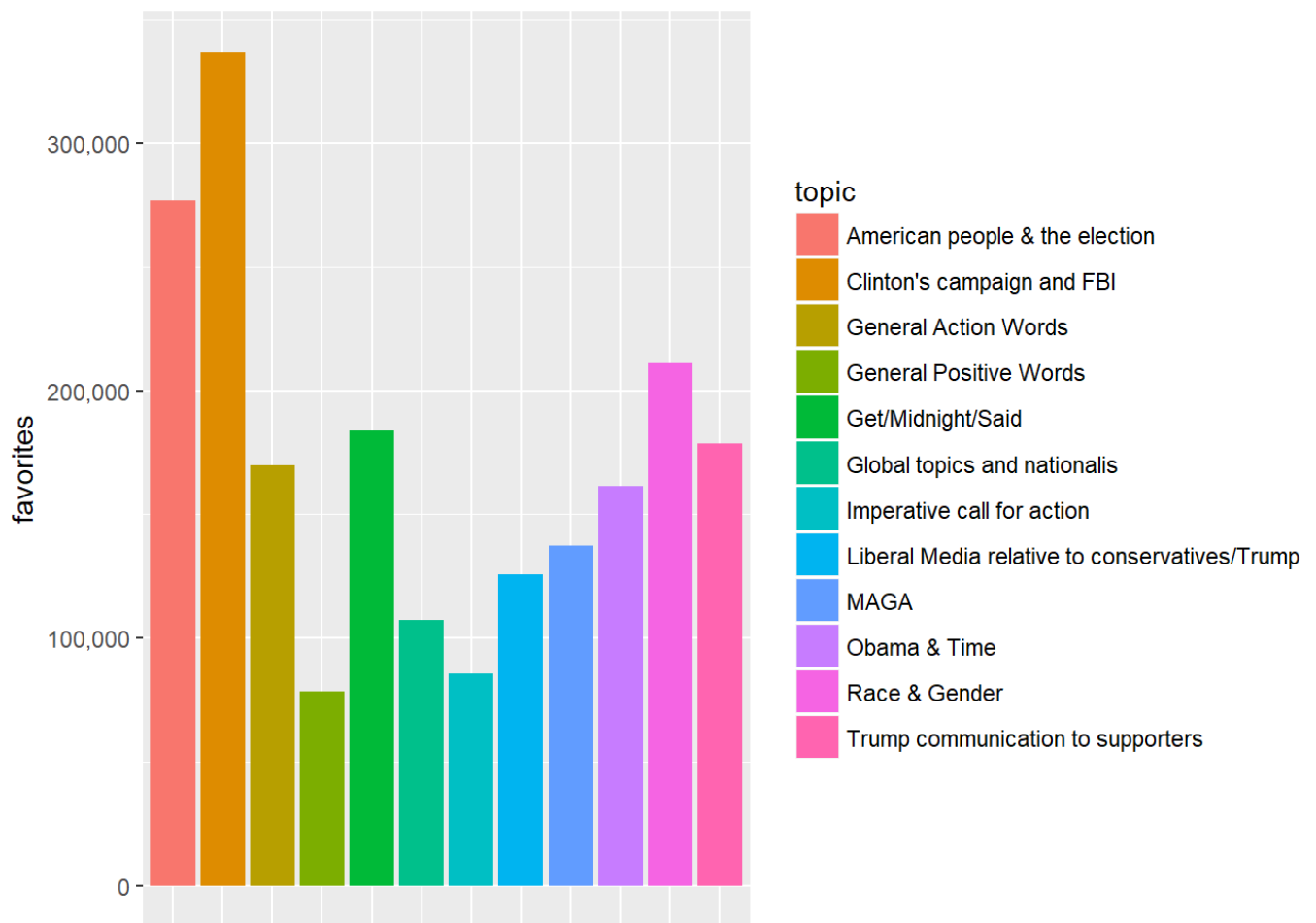
```

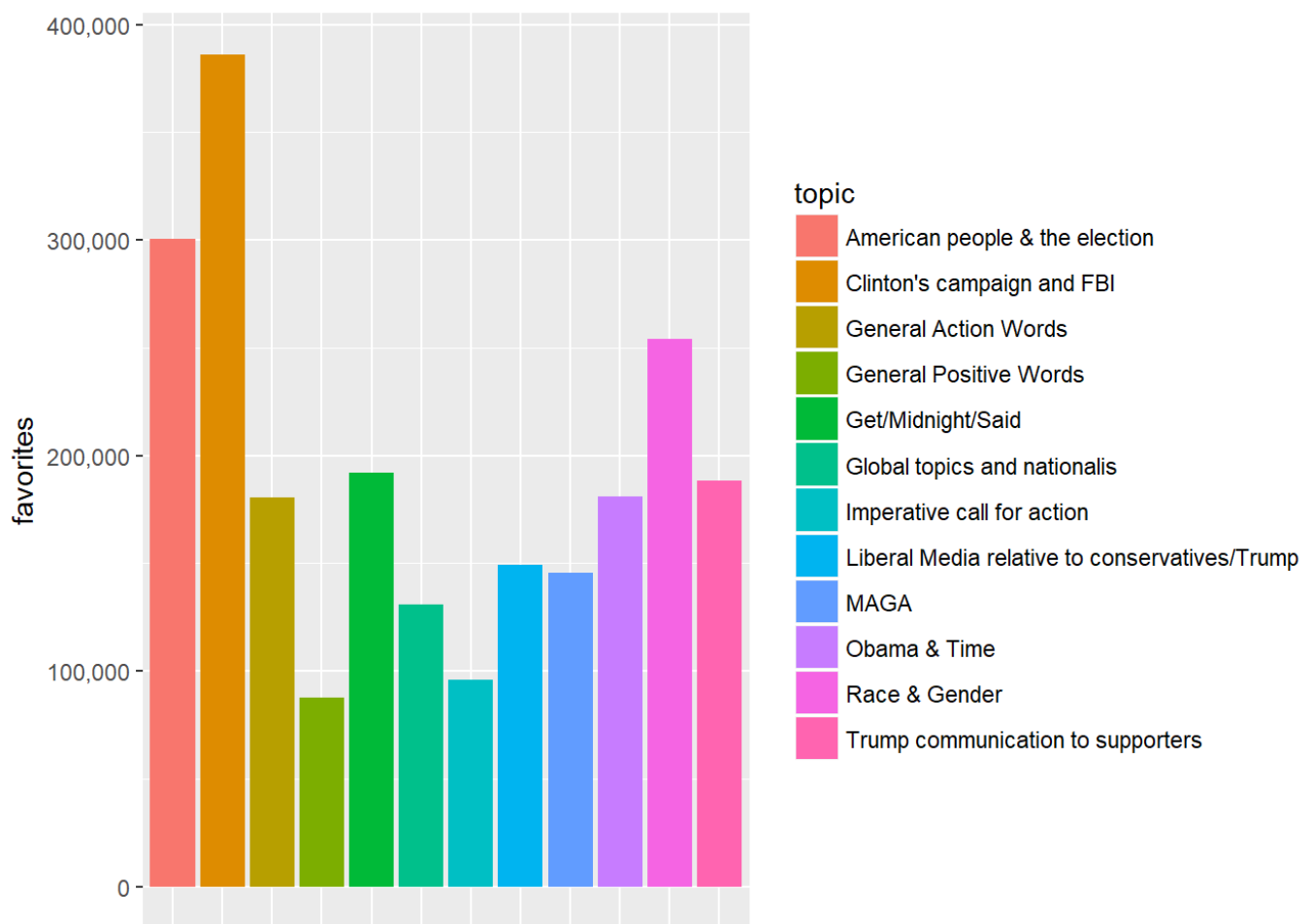
The more generic topic (general*, get/midnight/said) have a pretty high volume overall which makes sense. But additionally, we see the most popular relevant topics are 'imperative call for action' and 'liberal media relative to conservatives/Trump'. Lower via volume are topics on 'race & gender' as well as 'trump communications to supporters'. That being said, we've already established that volume of tweets is a VERY poor way to establish impact. 60k+ tweets were from users without a single interaction in on any of their tweets. Let's instead look by favorite/re-tweets.

```
fav_rt_topic <- tweets_used %>% group_by(topic) %>%
  summarise(
    favorites = sum(favorite_count),
    retweets = sum(retweet_count)
  )

ggplot(fav_rt_topic) + geom_bar(stat = 'identity', aes(topic, favorites, fill = topic)) +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.title.x=element_blank(),
        ylab('favorites') +
        scale_y_continuous(labels = comma))
```



```
ggplot(fav_rt_topic) + geom_bar(stat = 'identity', aes(topic, retweets, fill = topic)
) + theme(axis.text.x=element_blank(),
axis.ticks.x=element_blank(),
axis.title.x=element_blank()) + ylab('favorites') +
scale_y_continuous(labels = comma)
```

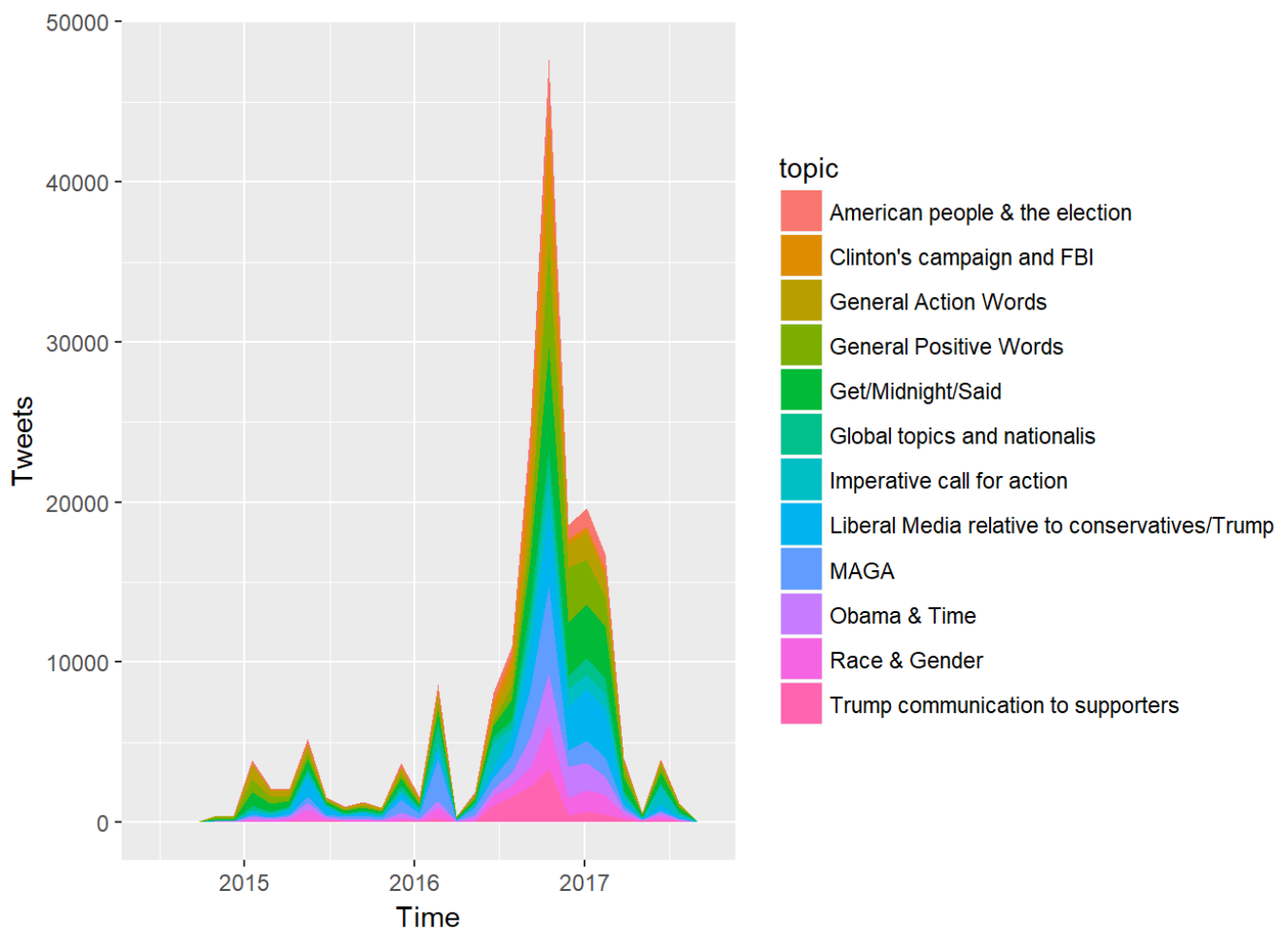


When looking by interactions, we can see that the general topics get very little interaction here but the tweets that get the most favorites and re-tweets are 'Clinton's campaign and FBI', 'American people & the election' and 'Race & Gender'. So the successful tweets were really focused on those specific topics as opposed to being generic. The Clinton/FBI tweets stand out as significantly more than any other topic.

The other important analysis to look at is what topics were being pushed over time and did the topics being discussed in these tweets adjust relative to current events.

```
ggplot(tweets_used, aes(x=month_year, fill=topic)) + geom_area(stat='bin') + xlab('Time') + ylab('Tweets')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The above is looking at what was being tweeted most frequently and when. One thing that stands out is at a high level the ratios seem fairly similar. We do see a lot more of the pinkish ('Race & gender' and 'trump communication to supporters') right around election but its hard to really tell from this high of a view. Let's take a closer look at the final 4 months of 2016 (peak election time)

```
tweets_used_6mo_elec <- tweets_used %>% filter(month_year < '2017-01-01' & month_year > '2016-06-01')

ggplot(tweets_used_6mo_elec, aes(x=topic, fill = topic)) + geom_bar() + facet_wrap(~month_year) + theme(axis.text.x=element_blank(),
axis.ticks.x=element_blank(),
axis.title.x=element_blank())
```

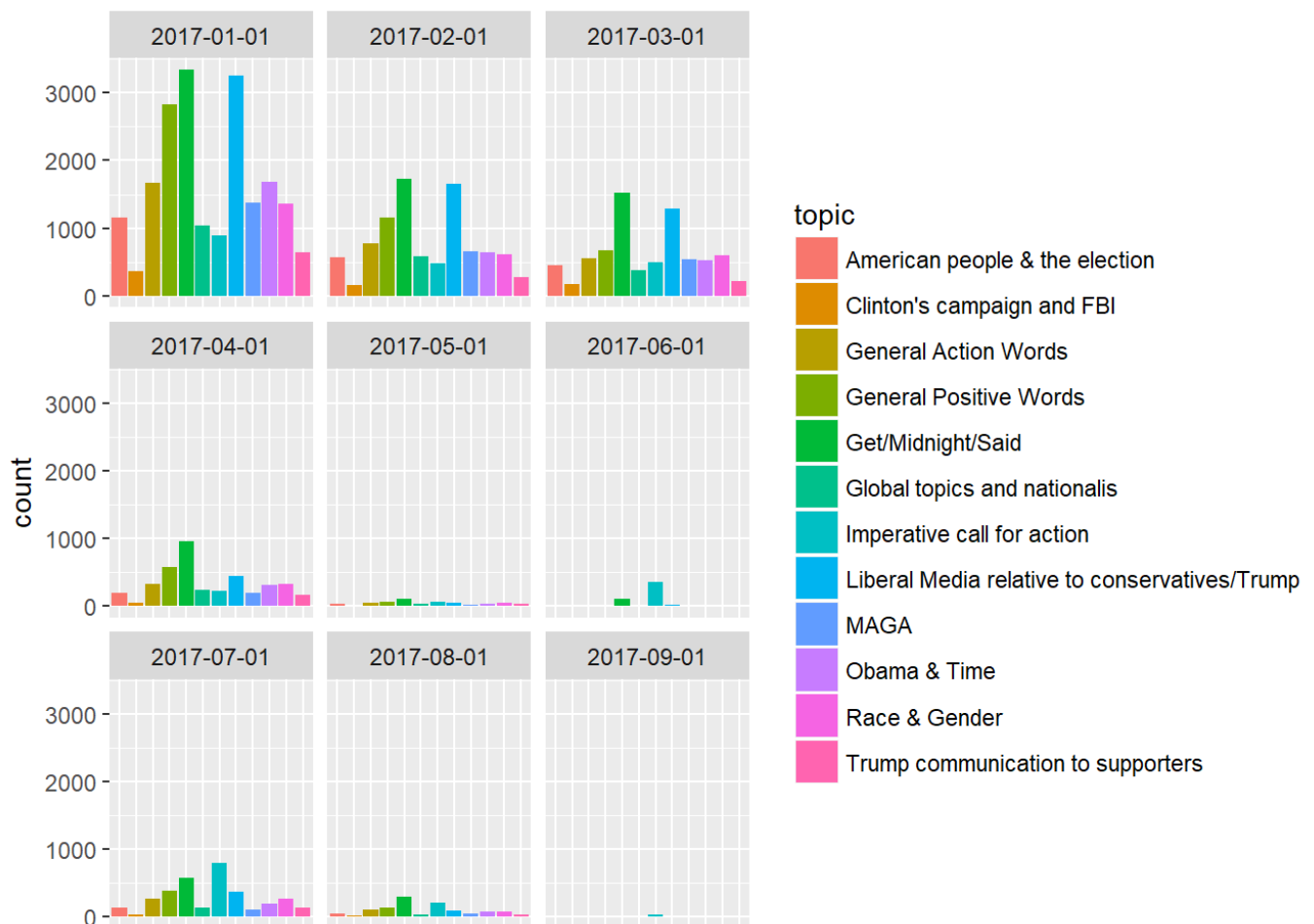


It looks like leading up to the election, specifically in September and October, there was a skyrocket in tweets related to 'Clinton's campaign and FBI' that quickly died of after. Come December, this was actually the least tweeted topic (which kind of makes sense since at that point the Clinton campaign had been defeated and was out of the picture). If we check back in the months prior to the election year, we see this wasn't really being discussed either. This topic was primarily pushed in the month directly prior to the election but had very little head before/after that.

We do also see a strong 'MAGA' push in pro trump tweets as well as 'trump communication to supporters'. Interestingly in December we see discussion around 'liberal media relative to conservatives/Trump' picking up; this does not surprise me either given the consistent tension between Trump and liberal media sources since his time in office.

```
tweets_used_6mo_2017 <- tweets_used %>% filter(month_year > '2016-12-31' & month_year < '2017-12-01')

ggplot(tweets_used_6mo_2017, aes(x=topic, fill = topic)) + geom_bar() + facet_wrap(~month_year) + theme(axis.text.x=element_blank(),
axis.ticks.x=element_blank(),
axis.title.x=element_blank())
```



If we continue to look forward into 2017, we see the tweets themselves drop off (or are shut down) by September 2017. General topics stay on top through 2017, while the next highest is 'liberal media relative to conservatives/Trump'. This aligns with December of 2016, and makes intuitive sense. People are talking less about Obama/Hillary/The Election, and the new normalized trending topic is the conflicts between Trump and some media outlets.

Conclusions

Assuming that the topics discussed here were planned and at some level of strategy (and not just a response to what was already in the media), we learned a few key things:

- Upwards of 450 twitter handles were used in this data set for influencing the election. More than half had no activity, but the majority of the tweets (~150k out of ~210k) were from accounts with atleast some type of interactions with another account. Note that it is highly like accounts were liking one another to boost stats, which is why its especially important to look at the top accounts with the most activity.
- A few top accounts had the most activity by far. These accounts ranged from many different topics and areas, not just a single political focus.
- The majority of tweets were related to general topics. That being said, these general tweets had very little success as far as favorites/re-tweets. The most interacted with (and therefore successful tweets) were related to the Clinton's and the FBI, race and gender, or the American people and the election in general.
- Time is an important factor. Obviously more tweets and focus occurred right around the election, but tweets related to Hillary Clinton and the FBI skyrocketed directly prior to the election, then quickly died off after. Following the election, the most talked about topic was the liberal media relative to

conservatives/Trump.