Final_Proj

Danni Fu

2017/12/15

I. Introduction

Nowadays, as rapid growth of usage of internet, social media plays a larger and larger role on the daily life of the public. The social media can reflect different cities' atmosphere of holidays. Bloggers with a lot of followers can affect the costumers' shopping decisions. Then, the another role Twitter can play in Christmas is social marketing, since it can released a new infographic which outlines the opportunities on Twitter for marketers ahead of the holiday season.

In order to figure out what Twitter users, are talking about Christmas, all the tweets contains hashtag about Christmas and the geographical location info are collected, researches about regions that has the majority celebrate christmas and related analysis would be useful for commercial purposes.

II. Dataset Summary

In this project, all the data come from Twitter. The data of this project includes tweets mentioning christmas collecting in 30 mins at around 1 pm in Eastern stardard time..

```
# set up connection
#requestURL <- "https://api.twitter.com/oauth/request_token"
#accessURL <- "https://api.twitter.com/oauth/access_token"
#authURL <- "https://api.twitter.com/oauth/authorize"
#api_key <-     "GGD6XUROrlweJRDBCPfeoFCke"
#api_secret <- "LWuS1jwNcjir4wFbYT9tvQ1hmwpqn9Ko1VdxwhrEEWHEUICkcm"
#access_token <- "940364005654360064-ENMUkOgVIvxgkwtWDIMwX1QpLTjPSqp"
#access_token_secret <- "1tbt9o6HXsPRzFIfeCAuRv9kE7hdF2Y954R1QtedtK4Nx"
######### Prepare for streamR
#my_oauth <- OAuthFactory$new(consumerKey = api_key, consumerSecret = api_secret,
#requestURL = requestURL, accessURL = accessURL, authURL = authURL)
#my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl"))
#save(my_oauth, file = "my_oauth.Rdata")
#load("my_oauth.Rdata")

######### Prepare for twitteR
#setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
```

```
#load("my_oauth.Rdata")
#
#  filterStream( file="Tweet.json", track="christmas",
#      locations=c(-74,40,-73,41), timeout=1800, oauth=my_oauth )
```

```
tweets_raw.df <- parseTweets("Tweet.json",verbose = FALSE)
```

```
## Warning in readLines(tweets, encoding = "UTF-8"): incomplete final line
## found on 'Tweet.json'
```

```
keep <- c("text","lang","listed_count","geo_enabled","statuses_count","followers_count",
          "favourites_count","friends_count","time_zone","country_code","full_name",
          "place_lat","place_lon")
```

```
tweets.df <- tweets_raw.df[,keep];


write.csv(tweets.df,"tweets_df.csv")
```

After searching tweets for 30 mins, cleaning it and obtaining locations' information, I get a data frames with all the information including listed counts,statuse count, followers count, favourites count,time zone, country code, city, states, longitude, latitude.

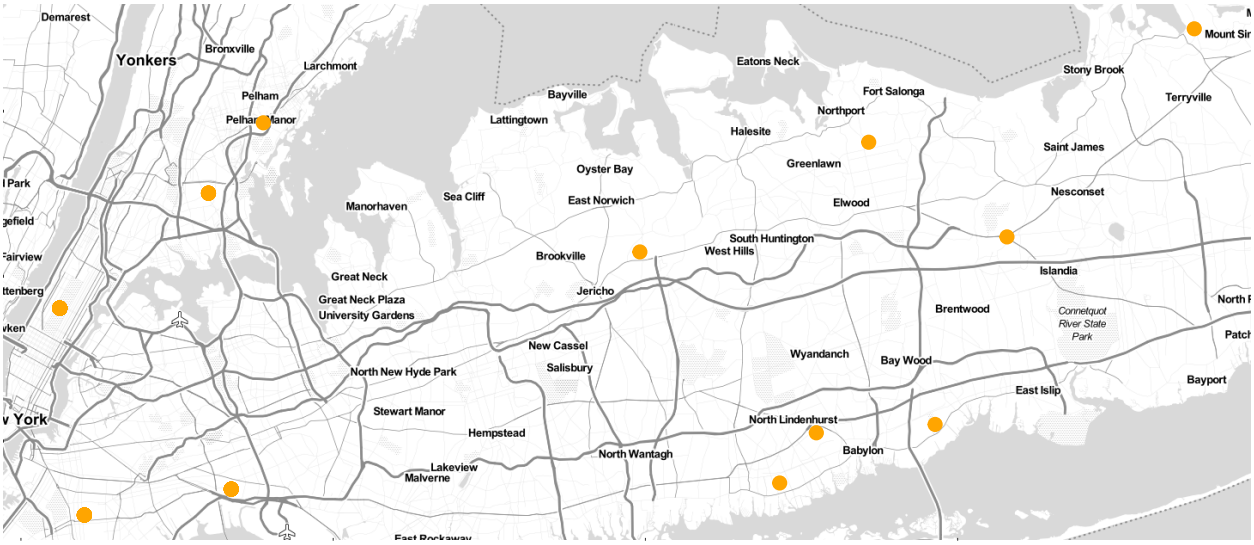
III. Mapping

To have a general of views of the distribution of those tweets especially in different cities in US, we use mapping to see which cities might be my insteret of studies.

```
# install.packages("ggmap")
library(ggmap)

tweets.df <- tweets.df[tweets.df$lang=="en",]
tweets.df <- tweets.df[tweets.df$country_code=="US",]
tweets.df <- tweets.df[tweets.df$geo_enabled==TRUE,]
tweets.df <- tweets.df[tweets.df$place_lat >25 & tweets.df$place_lat <50 &
                         tweets.df$place_lon > -125 & tweets.df$place_lon< -66,]

write.csv(tweets.df,"cleaned_tweets_df.csv")


#filter with place
ny <- data.frame(filter(tweets.df, grepl(', NY', full_name)))
dc <- data.frame(filter(tweets.df, grepl(', DC', full_name)))
il <- data.frame(filter(tweets.df, grepl(', IL', full_name)))
ca <- data.frame(filter(tweets.df, grepl(', CA', full_name)))
ma <- data.frame(filter(tweets.df, grepl(', MA', full_name)))
```

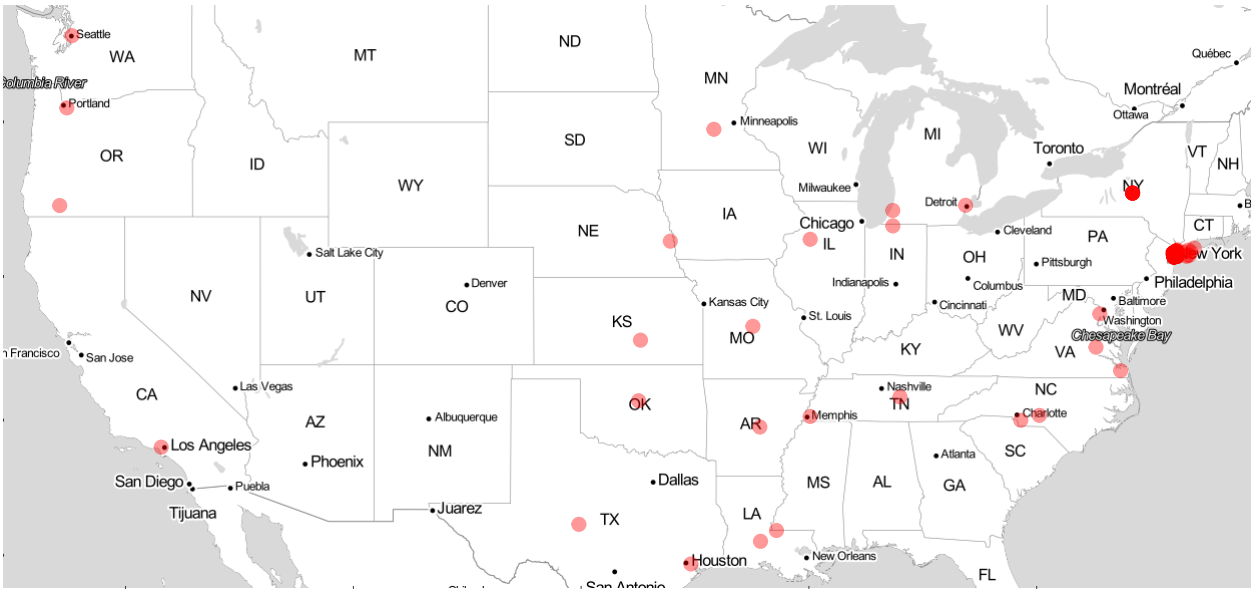
```
qmplot(place_lon, place_lat, data = ny, colour = I('orange'), size = I(2),
       mapcolor = "bw", main="NY")
```

NY




```
qmplot(place_lon, place_lat, data = tweets.df, colour = I('red'), size = I(2), alpha=I(0.4),
       mapcolor = "bw",main="USA")
```

USA



According to the map of the christmas tweets, we can see most of the tweets are mainly from New York, which are the cities for bussiness, fashion and so on. So the next step is to have a closer study of New York.

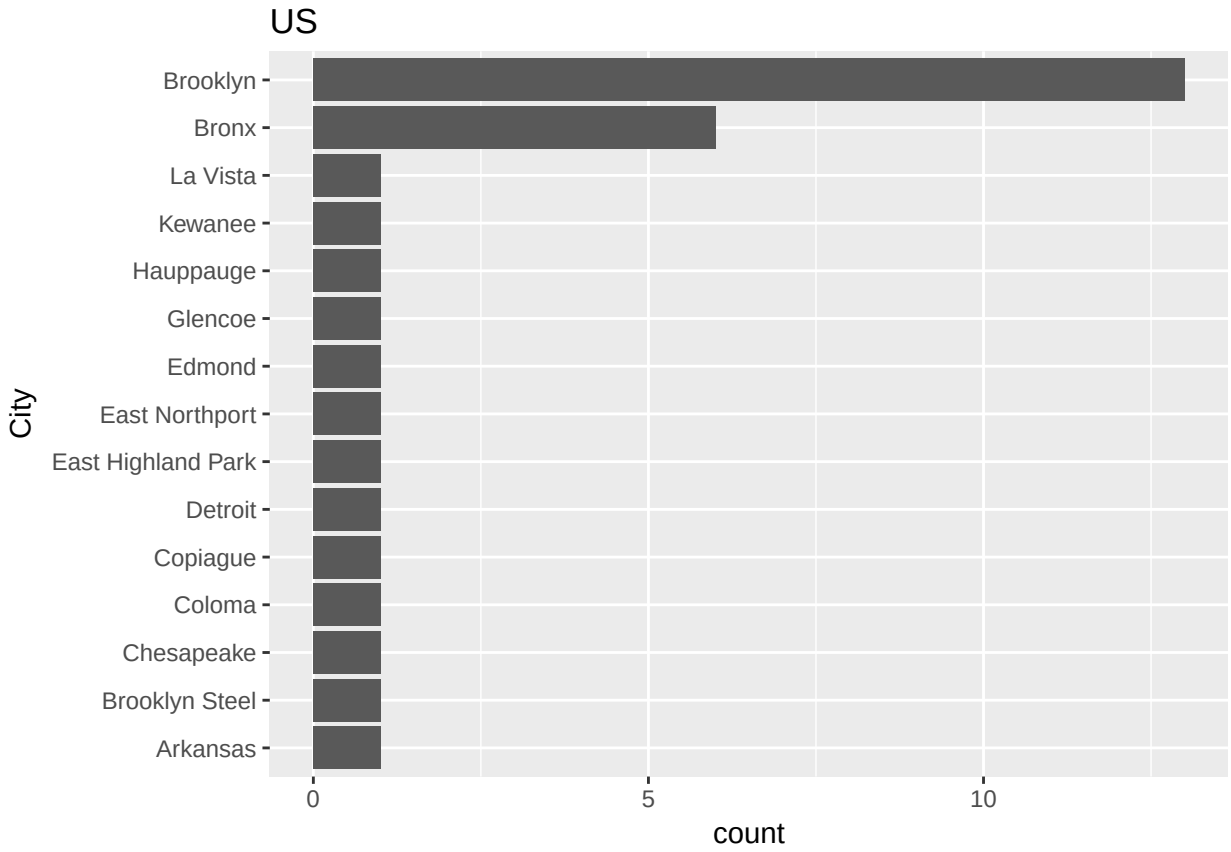
IV. Statistical Analysis

Having the targeted states, Some question below will be explored: 1. Whether location of the cities affect the popularity of tweets. 2. What is significant relationship between each pair of different information countable factors of the users, like followers, favorites, lists that uses belong to and so on.

```
ny[,"state"] <- rep("NY",nrow(ny));
ca[,"state"] <- rep("CA",nrow(ca));
ma[,"state"] <- rep("MA",nrow(ma));
```

```
#get city name
ny$full_name <- as.character(ny$full_name);
ma$full_name <- as.character(ma$full_name);
ca$full_name <- as.character(ca$full_name);
tweets.df$full_name <- as.character(tweets.df$full_name);

get_city <- function(x){
  city_name <- c()
  name <- strsplit(x$full_name,",")
  for(i in 1:nrow(x)){
    city_name <- c(city_name,name[[i]][1])
  }
  return(city_name)
}

#transform to factor
ny[,"city"] <- factor(get_city(ny));
#ma[,"city"] <- factor(get_city(ma));
#ca[,"city"] <- factor(get_city(ca));
tweets.df[,"city"] <- factor(get_city(tweets.df))

#combine
total <- ny
total$lang <- as.character(total$lang)
tweets.df$lang <- as.character(tweets.df$lang)
total_us <- filter(tweets.df,lang=="en")
total <- filter(total,lang=="en")
# tweets number for each city in each state
count <- summary(ny$city)[1:15];ny_city_count <- as.data.frame(count)
#count <- summary(ma$city)[1:15];ma_city_count <- as.data.frame(count)
#count <- summary(ca$city)[1:15];ca_city_count <- as.data.frame(count)
count <- summary(tweets.df$city)[1:15];us_city_count <- as.data.frame(count)
```

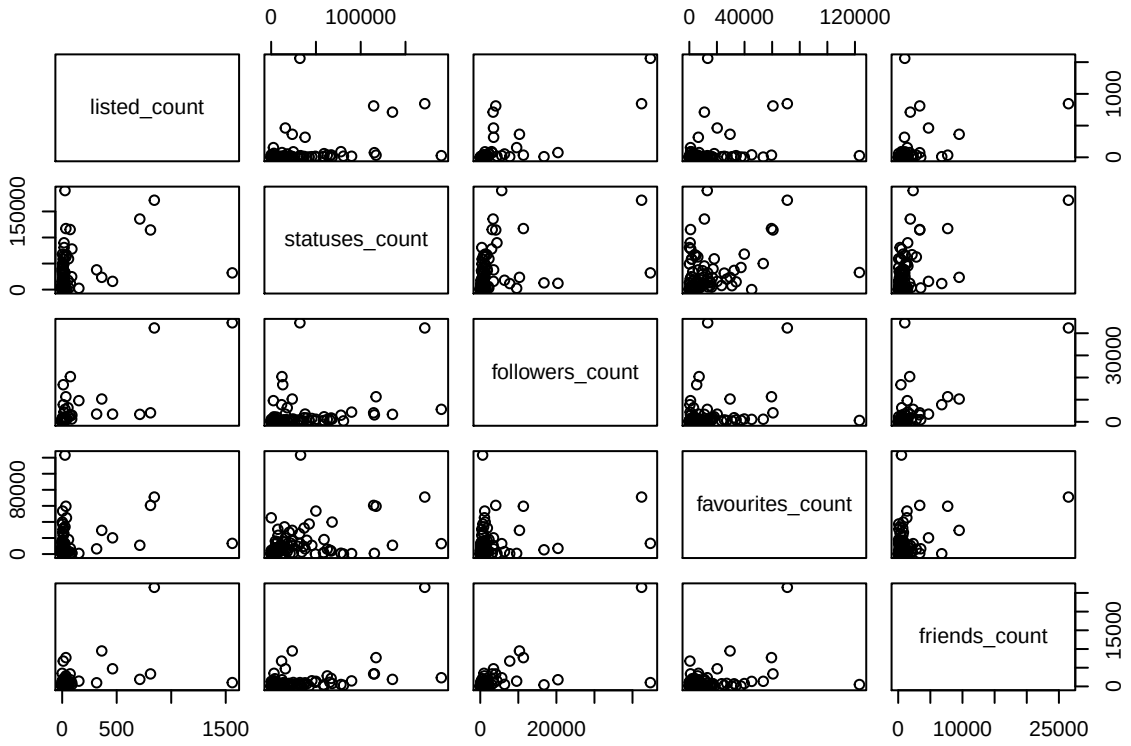
```
#plot
uscityplot <- ggplot(us_city_count,aes(reorder(rownames(us_city_count),count),count))+
  geom_bar(stat = "identity")+coord_flip()+xlab("City")+ggtitle("US")
uscityplot
```



According to plots of US, Brooklyn is the cities where most of those christmas tweets are from. Based on the plots of each interested states, the city which is the fashion center of the states tend to have the highest tweets counts, for example, New York is Manhattan.

To investigate the relationship between each pair of information factors, the scatter matrix plot and the corrlation tables are made to see any sigificant simple regression.

```
pairs(~listed_count+statuses_count+followers_count+favourites_count+friends_count, data=total_us)
```




```
cor(total[,c(3,5,6,7,8)])
```

```
##                  listed_count statuses_count followers_count
## listed_count        1.0000000     0.35261278      0.74166810
## statuses_count      0.3526128     1.00000000      0.07411436
## followers_count     0.7416681     0.07411436      1.00000000
## favourites_count    0.1897448     0.11444395      0.01930868
## friends_count       0.3333560     0.23892411      0.20245442
##                  favourites_count friends_count
## listed_count           0.18974480     0.3333560
## statuses_count         0.11444395     0.2389241
## followers_count        0.01930868     0.2024544
## favourites_count       1.00000000     0.1545734
## friends_count          0.15457339     1.0000000
```

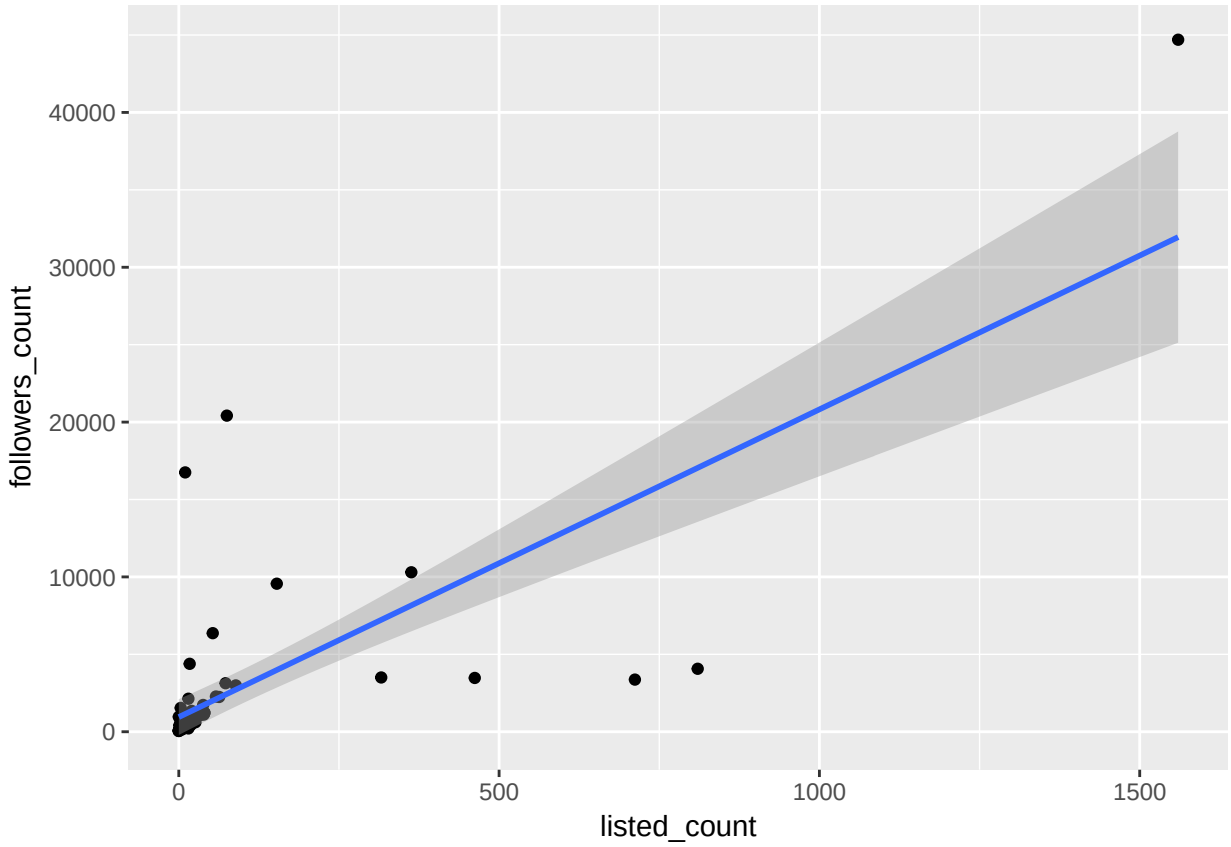
Based on R result, the followers_count tends to have a correlation with the list_count, and the change of one can affect another one noticeably. So the next step is to have a model to study those two variables.

```
model_fl <- lm(followers_count~listed_count,data=total_us)
summary(model_fl)
```

```
##
## Call:
## lm(formula = followers_count ~ listed_count, data = total_us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16514.1   -881.6   -690.2   -123.9  20954.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    975.18     461.81   2.112   0.0373 *
```

```
## listed_count    24.21       2.09  11.581   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4385 on 97 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.576
## F-statistic: 134.1 on 1 and 97 DF,  p-value: < 2.2e-16
```

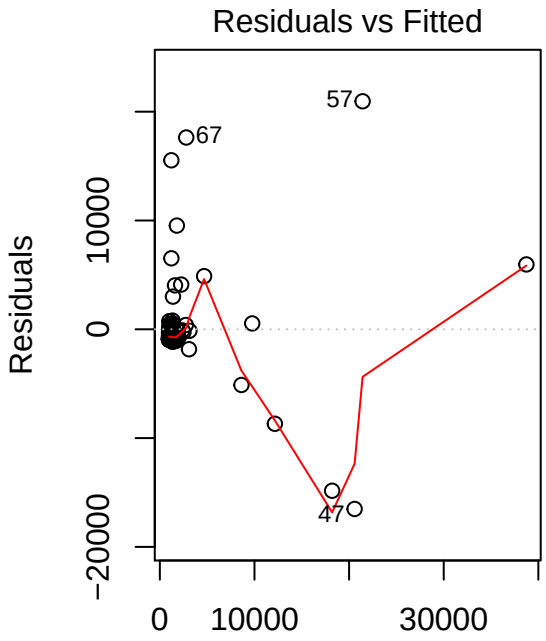
```
ggplot(total,aes(x=listed_count,y=followers_count))+geom_point()+geom_smooth(method = "lm")
```

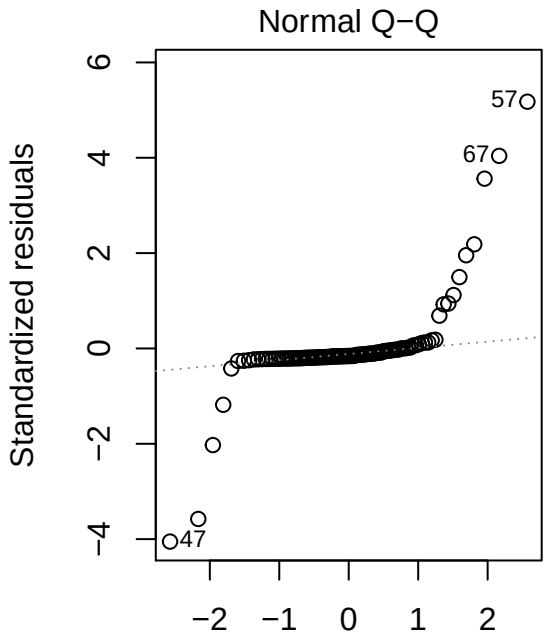



```
ggplot
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <environment: namespace:ggplot2>
```

```
#Analysis of regression:
par(mfrow=c(1,2))
plot(model_fl,1)
plot(model_fl,2)
```



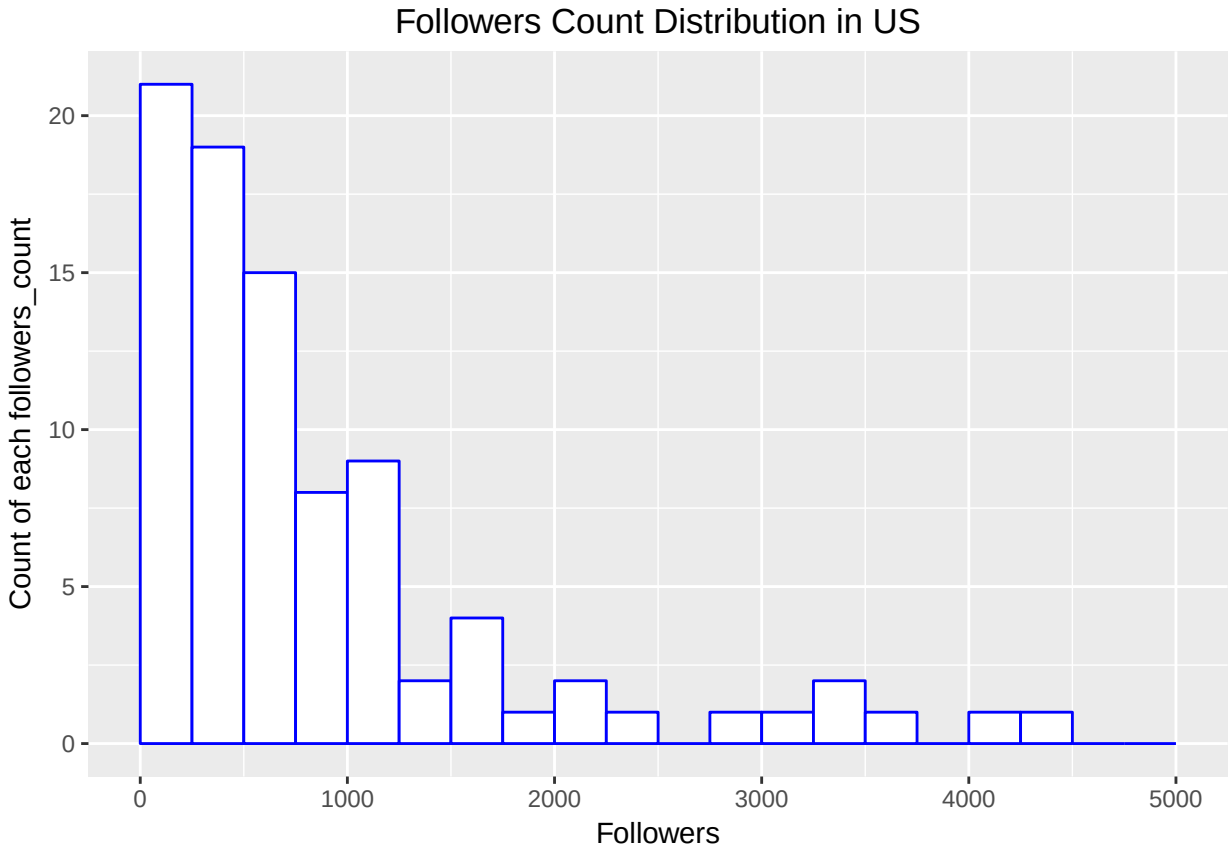


The listed_count is the number of public lists that this user is a member of. Based on the p-value of the model of listed_count and followers_count, the simple linear regression is significant. Considering the potential outliers and high leverage point, there might be randonness in the residual plot. We can see that the more public lists that this user is a memeber of, more follower the user have. As list increase by 1 , the follower count will increase by approximately 109.

Then, to further study the social media impact in different cities, the follower count distribution is computated.

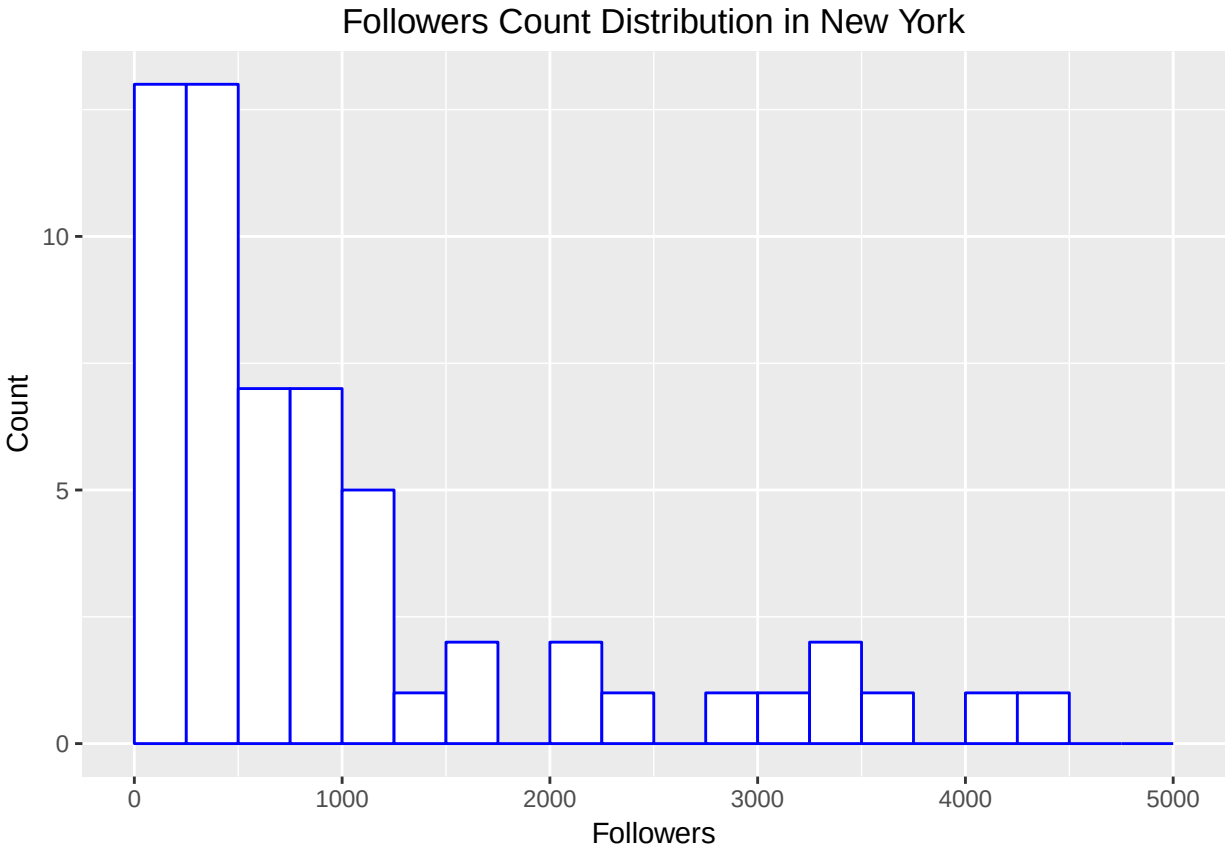
```
#add ggplot code
a <- ggplot(data=tweets.df, aes(followers_count)) +
  geom_histogram(breaks = seq(0,5000,by=250),
                 col = "blue",
                 fill = "white") +
  labs(title = "Followers Count Distribution in US") +
  labs(x="Followers", y = "Count of each followers_count")+
  theme(plot.title = element_text(hjust=0.5))
#printing the graphs
print(a)
```

```
## Warning: Removed 1554 rows containing non-finite values (stat_bin).
```



From the followers count distribution in US, we can see that most of those christmas tweets are from blogger with under 1000 followers. So we can see public use tweets to celebrate christma. More follower the blogger have more impact the tweets might bring to the public, so we can see the christmas tweets will bring a great impact to the public to celebrate the holidays.

```
#ggplot code for followers count distribution in New York
f <- ggplot(data=ny, aes(followers_count)) +
  geom_histogram(breaks = seq(0,5000,by=250),
                 col = "blue",
                 fill = "white") +
  labs(title = "Followers Count Distribution in New York") +
  labs(x="Followers", y = "Count")+
  theme(plot.title = element_text(hjust=0.5))
#printing the graphs
print(f)
```



According to the followers distribution in big and socialable states New York, we can see that New York's christmas tweets are from blogger with the most followers, which is understandable since there are most population and companies there. Twitter plays a much larger role in public's life and work. in New York.

V. Shiny

For the Shiny application, I created a navbar with all the graphs that I created earlier in different tabs, including Maps, statistic models. Descriptions of each graph are also included

All shiny codes are included in the folder called "shiny_app".

Shiny application are available through here: https://dannif.shinyapps.io/shiny/

VI. Future Improvements

The data that I get from 30 mins searching of christmas is about over ten thousands of data; however, after filtering tweets with valid location information and english languages, there is only 3000 datas left. In addition, the target is United states, the dataset is even smaller which might be not convincing enough to draw our conclusion. What's more, as the data are collected around 12pm in Eastern Stardard time, the california time is around morning, which might affect our results since people usually tweets more during lunch break than the busy morning. To avoid any bias of the study, I might collect more data by searching of christmas tag for longer and in different time periods of the day. To extend the study of social media effect, we might study other social media, like instagram, facebook and so on.