# Table of Contents

Executive Summary – ---------------------------------------------------------------------------------- 2

Statement of Scope – ---------------------------------------------------------------------------------- 2

Project Schedule – --------------------------------------------------------------------------------------- 3

Data Preparation – --------------------------------------------------------------------------------------- 6

Data Access- --------------------------------------------------------------------------------------6

Data Consolidation – --------------------------------------------------------------------------- 7

Data Dictionary – -------------------------------------------------------------------------------- 7

Data Cleaning – ---------------------------------------------------------------------------------- 7

Analysis – --------------------------------------------------------------------------------------------------- 8

Exploratory Data Analysis – ------------------------------------------------------------------- 8

Topic Modeling – --------------------------------------------------------------------------------- 16

Sentiment Analysis – --------------------------------------------------------------------------- 17

Conclusion – ----------------------------------------------------------------------------------------------- 19

Appendices – --------------------------------------------------------------------------------------------- 21

# Executive Summary

Fear or shame has kept many people from reporting crimes of sexual harassment or assault. Recently, the “#MeToo” movement has encouraged victims to file complaints to authorities and to gain support from other victims over social media after sharing their traumatic experiences. This movement has had a positive impact on the lives of many victims who otherwise would not have shared their story and who by sharing it have contributed to bringing the assailants to justice. We feel that it is important to understand the different sentiments associated with the movement on social media as well as explore the popularity of the movement. To do this, we will use R and Python to scrape tweets from the web and explore their popularity through data visualization as well as their polarity by performing sentiment analysis on them.

# Statement of Scope

Our goal is to study the sentiment behind the #MeToo movement on twitter to learn about the different attitudes and levels of support amongst the public. The movement was extremely popular over the past year and has had widespread support from a large portion of the population. Through analyzing the tweets associated with this movement, we hope to get a better understanding of the public support or resistance to it.

**Project Objectives -**

1) Web Scraping of Twitter data using the “#MeToo” hashtag

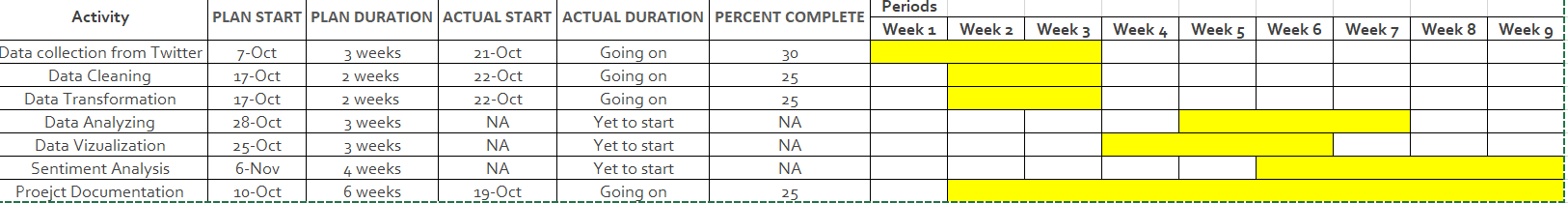
2) Descriptive data exploration and visualization of the tweets

3) Text mining and sentiment analysis across the tweets

**Variables –**

The dataset will consist of tweets gathered from Twitter. The variables will consist of the text portion of the tweet itself as well as information regarding the tweet such as date, number of retweets, and favorites. The sentiment of the tweet will also be added as a variable once the dataset is cleaned and analyzed for sentiment.

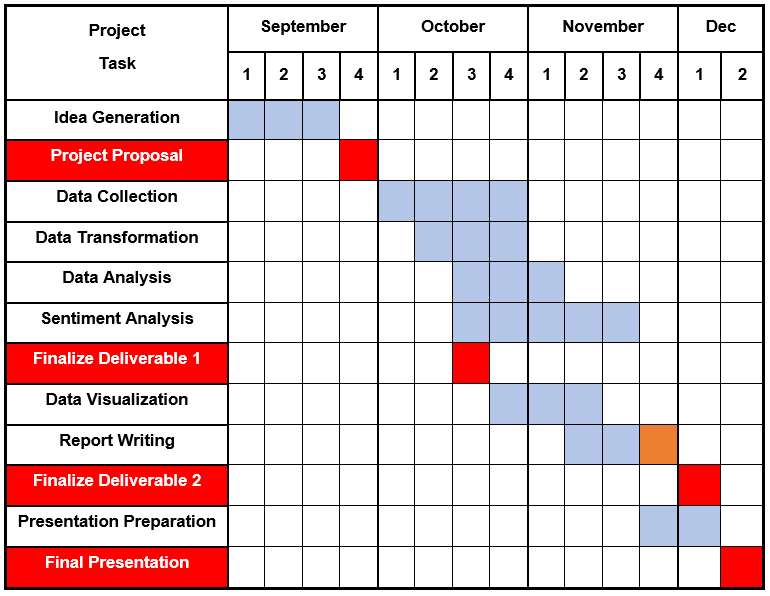
# Project Schedule

 All of the project timelines and their iterations are included in this section. The initial timeline shown in figure 1 was a broad estimate of completion dates. A few weeks into the project, we found that it was necessary to adjust completion dates and include more detailed steps for each task.

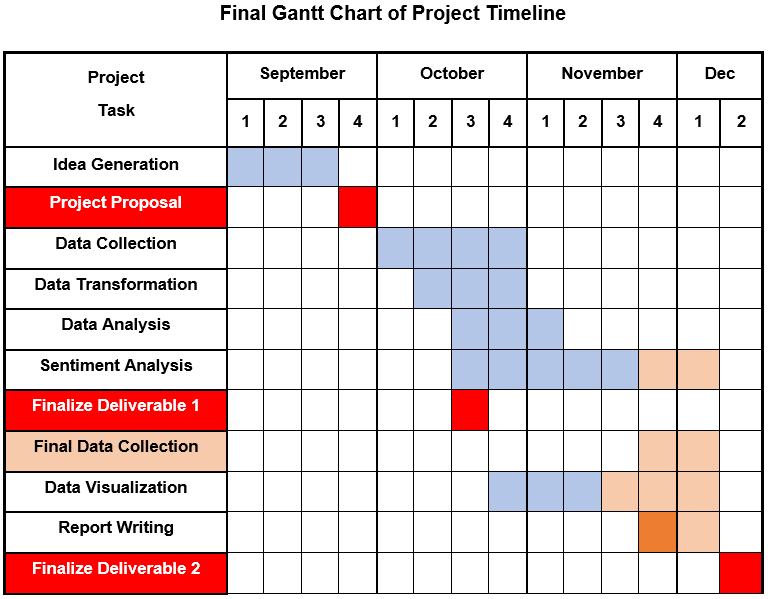
**Figure 1. Initial Project Timeline.**

After the edits and updates were made, we created the timeline shown in Figure 2. At this time, we had also established clear work assignments that each of us would focus on while also overlooking and helping with each other’s work as well. As general rules; Harish and Gopi were given the task of scraping tweets from twitter and preparing the data for analysis. Luis and Sai Teja focused on the visualization of the data and documenting the results and process for the report. The whole team played a part in the text analysis potion of the project.

On our Gantt Chart, Thanksgiving break was represented with an orange color. It fell on the last week of November which is one week before the Deliverable 2 submission deadline. In order to avoid any difficulties, we planned to begin working on the final report two weeks before thanksgiving break so that minimal work would be left to do during the break. While working on the report, we also decided which parts each of us would present.

****

**Figure 2. Gantt Chart of Project Timeline (Version 2**

** Figure 3. Gantt Chart of Project Timeline (Version 3)**

The final project Gantt chart (Figure 3) shows the final schedule actually followed to finish the project on time. The light orange color signifies the changes to the second iteration of the schedule (see figure 2). The project deadlines were changed which gave the team some extra time to perform the analysis. The main change was the collection of additional data in the later stages of the project due to small issues with the initial data collected. The final data collection required the data preparation process that was done on the first dataset to be performed once again on the new data. This did not take long however as the general process of collecting, cleaning, and analyzing the data had already been performed on the preliminary dataset. Overall the major change was the addition of time spent on the project in the later parts of the project. Additionally, the presentation portion of the project was cancelled and so was removed from the final chart.

The schedule was followed relatively well and the team was diligent in completing their assigned tasks by the dates which had been agreed upon. Only slight delays happened which came about due to conflicts with other work and exams that team members had in other classes and some slight problems with the initial data. Delays due to other projects and deadlines are to be expected, therefore, the only major change we would make to the project schedule would be the inclusion of one to two weeks of “buffer time” to allow for the delays which would likely happen. This way, the project can stay as close to the planned schedule, even with circumstances outside of our immediate control.

# Data Preparation

**Data Access -**

We have extracted the data from Twitter using the API keys. Harish has used his unique consumer keys and access tokens to authenticate the handler in Twitter and create an API (for data access and data cleaning code see Appendix A).

Once the API is created, the tweets are extracted on the #MeToo hashtag from 10/01/2019 to 10/22/2019. This gives us a list of the following 5 variables:

* Created\_Time
* Tweet\_ID
* Tweet\_Text
* Favorite\_Count
* Retweet\_Count

**Data Consolidation –**

Once the tweets were extracted, we converted the tweets into a data frame using the Pandas library in Python. Pandas is a library that provides many functions in python that can be used for data consolidation and data mining. We have allowed for 1000 characters per each tweet as the maximum width of the Tweet variable. Using these tools, the data was structured form a large block of text into columns and rows for ease of analysis.

**Data Cleaning –**

From the Natural Language Tooling Kit (nltk), we have imported packages that are useful in tokenizing, stemming and lemmatizing and that have stopwords dictionary. Using those packages, the tweets were converted to lower case letters, numeric digits 0 to 9 and all special characters such as ‘@’, ‘!’, were removed.

Tokenization – Tokenization is the process of breaking the text into smaller pieces called tokens. This is accomplished by first removing redundant words called stopwords that carry no meaning. Apart from the default set of stopwords that nltk corpus provides, we have added additional words such as ‘like’, ‘share’, ‘moment’.

Lemmatization - The aim of lemmatization, is to reduce inflectional forms to a common base form. Lemmatization does not simply remove inflections, instead it uses lexical knowledge bases to get the correct base forms of words. From the nltk corpus, we imported word tokenizer and wordNetLemmatizer to lemmatize the tweets.

Part-of-speech tagging (pos) aims to assign parts of speech to each word of a given text (such as nouns, verbs, and adjectives) based on its definition and its context. We have converted the tweets to a list format. We mapped the pos tag sent to the list of tweets to complete the tagging.

**Data Reduction –**

The “reduction” of data in the case of this project was done during the collection of data from Twitter. Although there are other time periods available for usage, we only scraped tweets in English from a specific time period and which included the #MeToo hashtag. In addition to this, we removed retweets from the analysis.

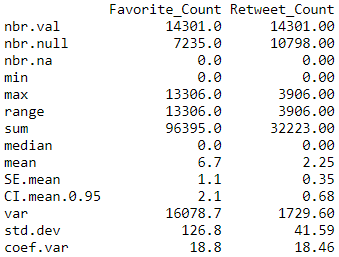
**Data Dictionary** –

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Data Type** | **Description** |
| Created\_Time | Datetime | The time of the tweet |
| ID | Numeric | A unique ID for each tweet |
| Tweet\_text | Text | The content of the tweet |
| favorite\_count | Numeric | Count of favorites for the tweet |
| retweet\_count | Numeric | Count of retweets for the tweet |

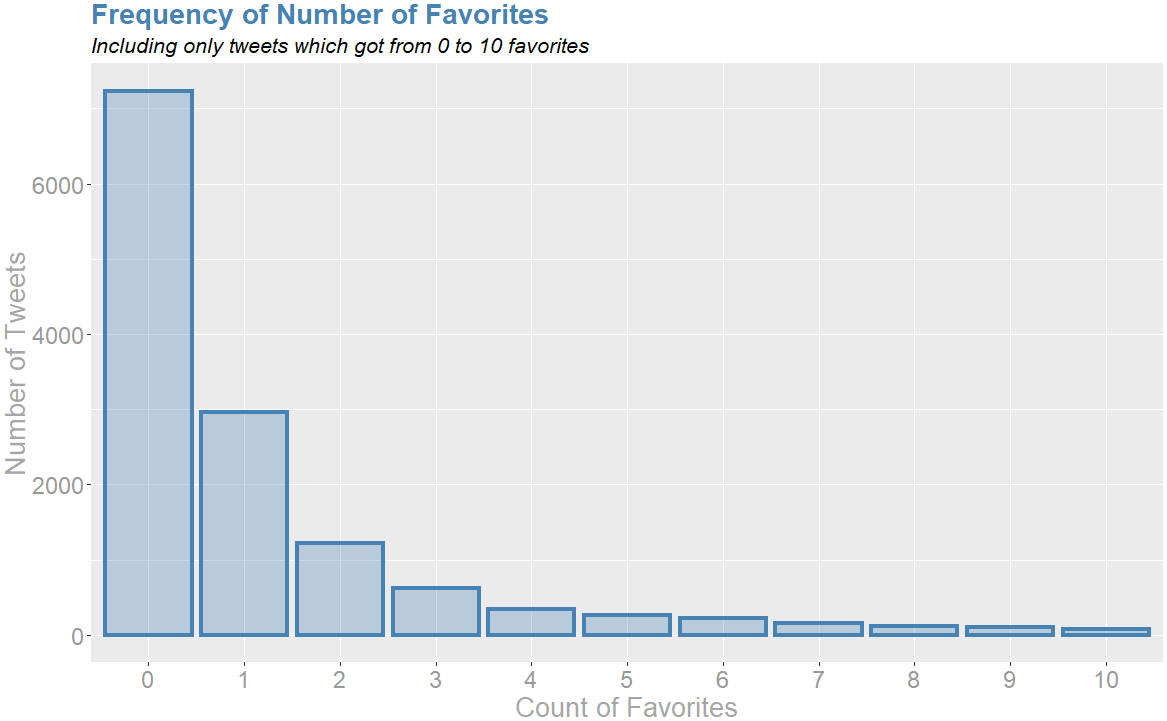
# Data Analysis

The process of analysis began with general exploratory data analysis. After this was conducted, the contents of the tweets were studied using named entity extraction and finally sentiment analysis.

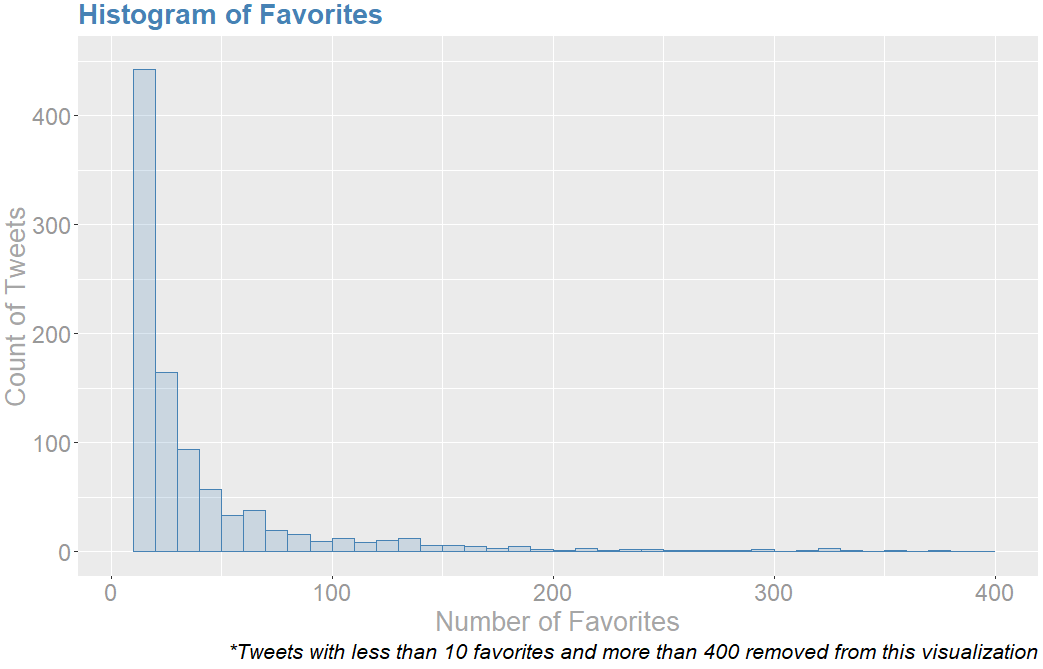
**Exploratory Data Analysis –**

Figure 3. Shows the descriptive statistics for the numeric values of our dataset. Although we had only two numeric variables, they provide us with valuable insight regarding the activity of people on twitter. For example, based on our sample of data, it is much more popular to favorite tweets than to retweet them. Also, the values for both retweets and favorites seem to have high variance and standard deviation (see appendix B for Exploratory Data Analysis code).

**Figure 4. Descriptive Statistics of Numeric Variables.**

To further explore the trends in the numeric variables, data visualizations were created using the ggplot2 package in R and seaborn package in Python. Figure 5 shows the number of tweets which had between 0 and 10 favorites.

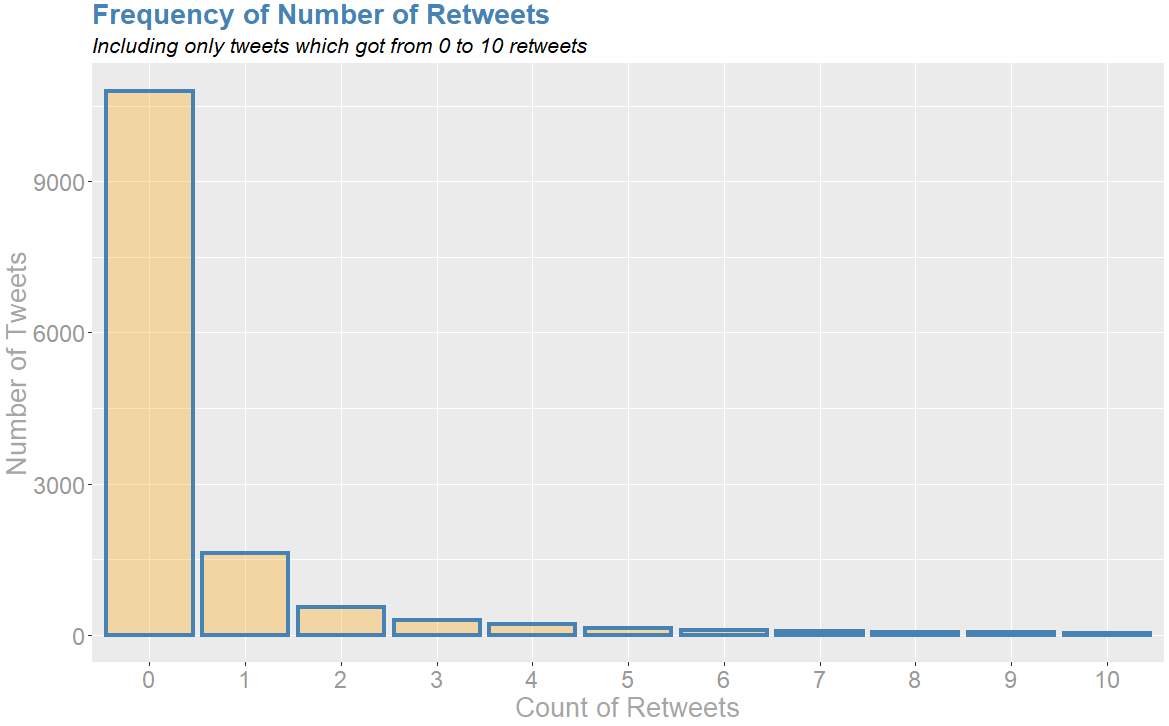
**Figure 5. Number of Tweets with Count of Favorites Between 0 and 10.**

This range encompassed the majority of observations (13,312) and within this group a majority of tweets had 0 favorites (7,235). This means that out of all of the tweets in our dataset, about half of them had no favorites at all. The previous graph merely shows the tweets with 0 – 10 retweets, the rest of the retweets are visualized in Figure 6. These were separated given that the range between the values was so large (0 – 13,306), it was difficult to accurately see the distribution of the smaller values if they were all on the same graph.

**Figure 6. Distribution of Count of Favorites Greater Than 10.**

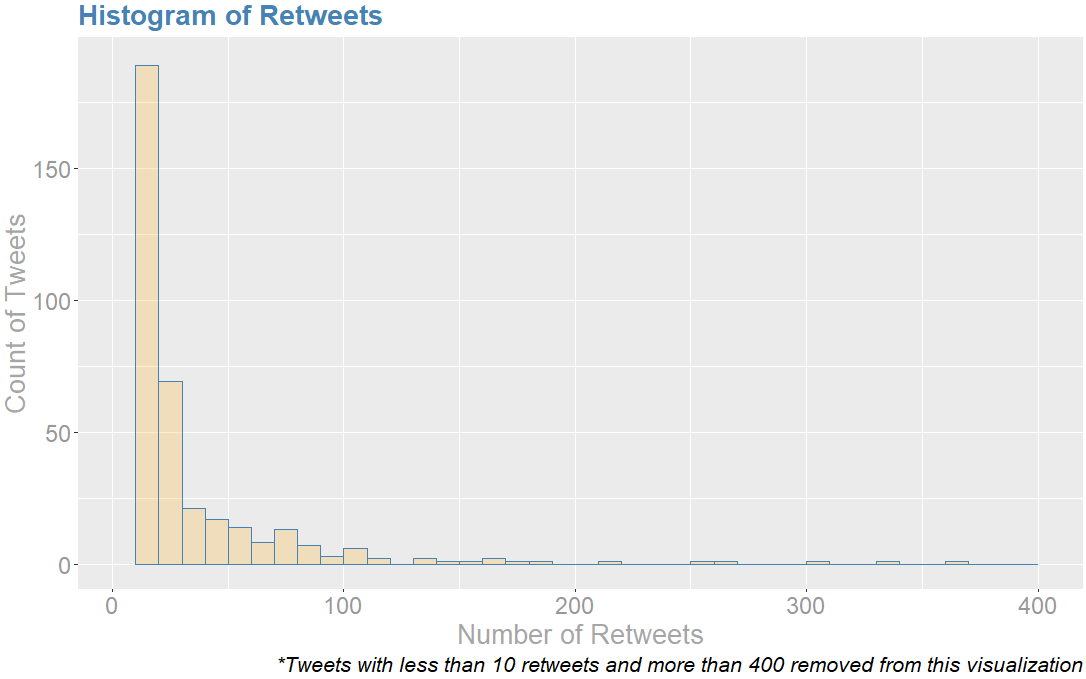
The histogram shows a similar trend as before; the number tweets drops off quickly as the number of favorites increases. Overall the distribution of the favorites variable is highly right skewed and leptokurtic with a majority of values concentrated on none to a very low number of favorites.

In a similar way, we now explore the number of times the tweets in the dataset were retweeted. Figure 7 shows the number of tweets with ten or less retweets.

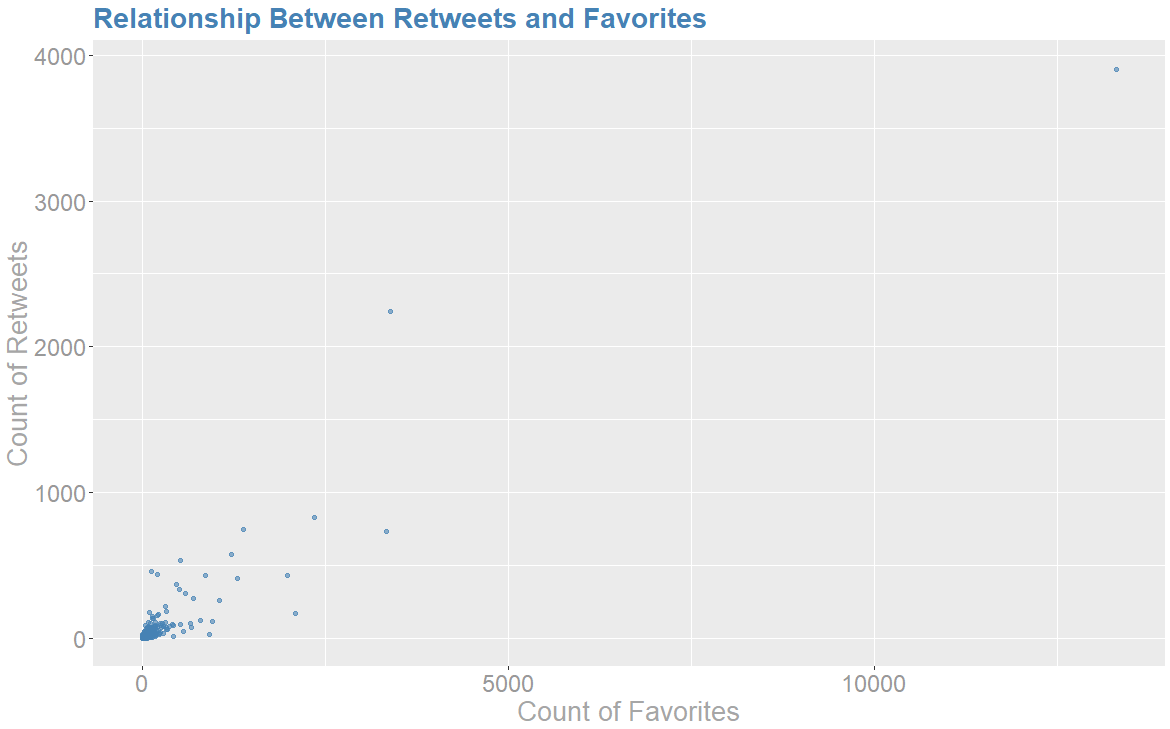
 **Figure 7. Number of Tweets with Count of Retweets Between 0 and 1.**

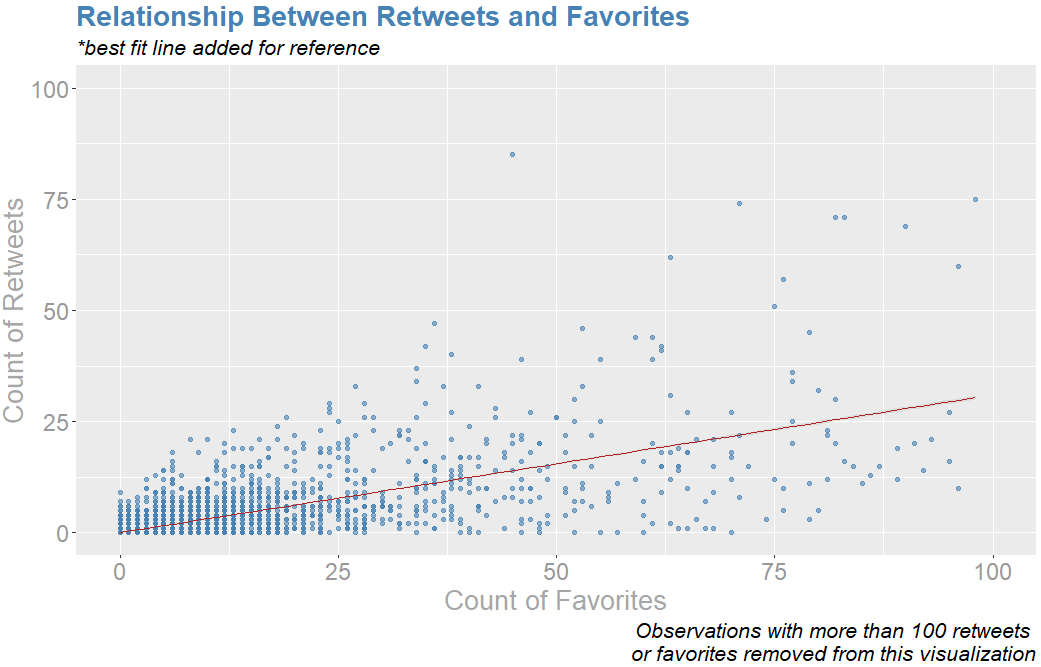
Once again we see that the majority of tweets are centered around the lower values with 10,798 having zero retweets. This means that the distribution drops off even faster for retweets than for favorites, with only 2,514 out of 14,301 having at least one retweet (about 18%).

The distribution of the tweets with greater than 10 retweets is explored in Figure 8 and shows similar trends as the ones we have seen so far.

** Figure 8. Distribution of Number of Retweets with Retweets Over 10**

The graphs show a heavily right skewed and leptokurtic distribution, with the number of tweets with many retweets being sparse. Given how popular the #MeToo movement was at its onset, it is slightly surprising to see so many tweets which include this hashtag receive little to no support.

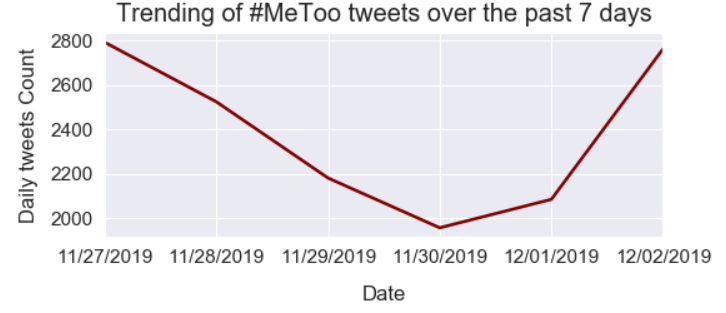
The next visualization (Figure 9) explores the relationship between favorites and retweets. First, the scatterplot shows that there are several outliers which are very far from the rest of the data. There is especially one which stands far from all of the others. While this plot is helpful to visualize the outliers, it makes it difficult to see the rest of the data points. Given that the data is centered around the lower values, we created a separate scatterplot of these two variables on a subset of values which had a relatively low number of retweets and favorites (Figure 10). This plot more clearly visualizes the positive correlation ****between these two variables.

** Figure 9. Scatterplot of Favorites and Retweets**

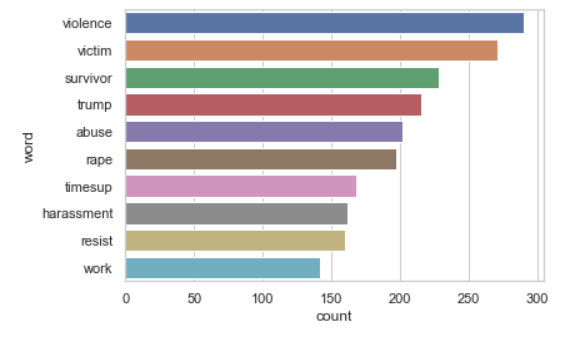
**Figure 10. Scatterplot of Tweets with less than 100 Retweets of Favorites.**

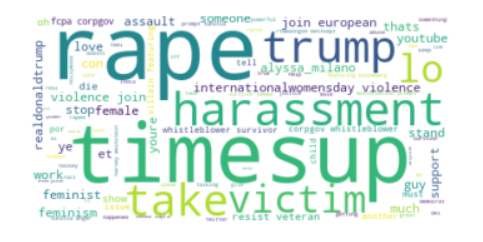
From both of these scatterplots we see a positive correlation between the two variables with favorites increasing more quickly than retweets.

The next visualization is of the dates of the tweets. Figure 11 shows the number of distinct tweets with #MeToo in the period from 11/27 to 12/02. The plot shows that the number of tweets peaked on Wednesday, fell low on weekends and rose again in the weekdays.

** Figure 11. Number of Tweets From 11/27 to 12/02**

After exploring the numerical and date variables in our data, we performed some exploratory analysis on the text portion of the dataset. First, we aggregated the words and counted the frequency of all the words then plotted the top 10 most frequently occurring words in the tweets (Figure 12).

 **Figure 12. Most Frequently Occurring Words.**

After this, using the word count corpus, we generated the text of word clouds on a white background using bilinear interpolation.

**Figure 13. Wordcloud for Selected Tweets.**

It is expected that the words “rape”, “times up”, “harassment”, and “victim” appear in the cloud. However, the fact that the word ‘trump’ is appearing is quite interesting given that this is not a subject that he commonly talks about.

**Topic modeling –**

Topic modeling is a text clustering technique designed to determine categories or topics of text. The modeling technique searches through all the documents to find matching patterns of terms and groups them together into topics. We have used Latent Dirichlet Allocation (LDA) type topic modeling technique. LDA is a natural language processing technique that creates topics based on the co-occurrence of words in documents. If a set of words appear more frequently together in a set of documents, then that denotes a topic. Each document (or tweet in our example) can contain multiple topics. Some topics are represented less in a document than others depending on word frequency. From sklearn module, we have imported LatentDirichletAllocation CountVectorizer packages and created a document-term matrix with 5 components and a random state of 35 (full data analysis code can be found in Appendix C). We then identified the top 10 words in each component. This technique was useful to see which were the predominant conversations or topics which were being spoken of with most popularity. Figure 14 shows the results of the topic modeling.

**Figure 14. Top 10 Word in Each of the Topics**

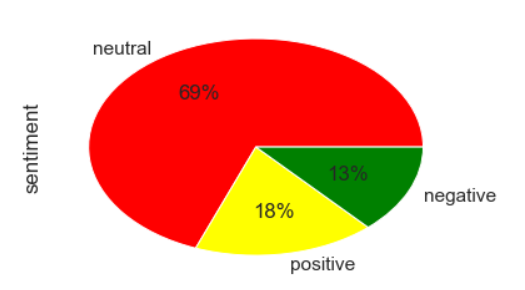
If we study the most commonly cited words in the five topics, it is clear that the #MeToo hashtag has become highly political. For example, the first two topics mention politicians such as Michael Bloomberg (“Bloomberg”), Gordon Sondland (“Sondland”), and Donald Trump (“realdonaldtrump”). The tweets in these topics tend to be accusatory and critical of these politicians and less in support of the victims of crimes.

In contrast, the final three topics are more about people sharing their stories and finding support. For example, the terms “survivor”, “free”, “support”, and “resist” are common in these topics, indicating positive feelings. Negative topics that are likely related to speaking out against the abuse are also found, such as, “harassment”, “predator”, “victim”, and “violence”. The fact that these two types of terms are found in the same topics may indicate that victims are sharing stories and speaking out while also talking about positive and encouraging things.

**Sentiment Analysis –**

We have used the text blob package in python to do the sentiment analysis. Textblob allows users to split the tweets into different sets of words. Then it analyses the words and identifies the polarity of the sentiment. For this deliverable, we decided to categorize the sentiment into 3 levels ‘Positive’, ‘Negative’ and ‘Neutral’.

Figure 14 shows the percentage of tweets falling within each of the categories, with the maximum being Neutral (69%) followed by Positive (18%) and finally Negative (13%).



**Figure 14: Pie Chart of Distribution of the Number of Tweets by Sentiment.**

From the above analysis, we can see that the #MeToo movement on Twitter has created a mixed opinion in the public. 69% of the population has a neutral opinion on it which means that there are no strong positive or negative words within the tweet which allow it to be classified. This is interesting as we should expect a topic like this to include highly emotional words which are easy to classify. An 18% positive sentiment shows that a good percentage of people are mentioning the movement in a positive manner. This could be that they are speaking out in support of victims or sharing their own stories with a positive tone. The 13% negative sentiment is difficult to interpret, some of the tweets in here may be a victim sharing their story with words of anger or fear, which is understandable given the sensitivity of the topic. However, some tweets in this category may also be of those which see the movement itself as negative.

# Conclusions

**Discussion of Results –**

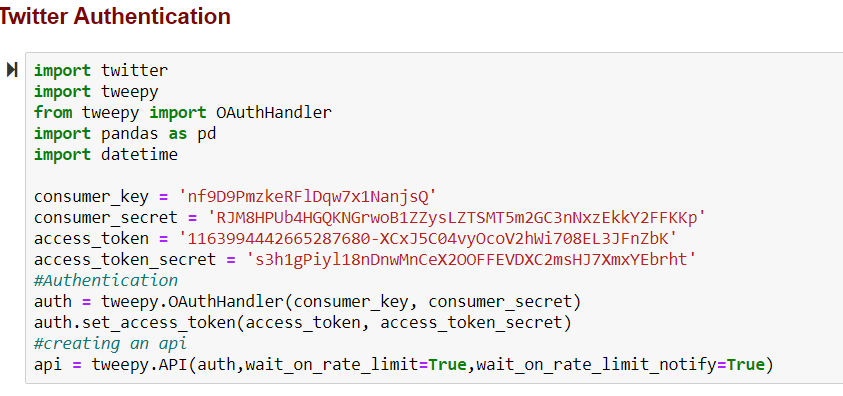
The #MeToo movement started off as a viral and popular phenomenon, when it began, it was generally intended for good and provided many victims with much needed support. From our analysis, we find that while it is still a popular hashtag and topic on twitter, the majority of tweets which include it are not getting much support via retweets or favorites. This low relative interest in these tweets could signify that the #MeToo movement has either lost some of its momentum or that the hashtag is not always being used in the manner which was originally intended. For example, tweets which make jokes about a sports team loosing with an inappropriately included “#MeToo” would be included in this dataset but are unlikely to gain much support from the general community and are not related to the movement itself. Similarly, many have complained about the movement and called it unfair given the impact that it could potentially have if someone were to be falsely accused. Because of this, the #MeToo movement and the hashtag in particular has apparently lost some of its impact and significance.

**Expansion and Future Studies –**

Based on our text sentiment analysis, it was found that 18% of tweets had a positive sentiment and 13% of tweets had a negative sentiment. We also found the popularity of these tweets based on their favorites and retweets. What is not answered in this analysis is the distribution of those sentiments based on demographics. (Though Twitter does not provide this data) it would be worth exploring if positive sentiments favored females or males and by how much it was favored. Additionally, the age of the person tweeting would provide further insights. For example, it is possible that college aged men have a positive sentiment while mothers of young males have a negative sentiment towards the movement because they fear their son may be falsely accused of sexual harassment. In October 2018 a mother’s tweet went viral after she posted a photo of her son stating that he “respects women” and that “he won’t go on solo dates due to the current climate of false sexual accusations by radical feminists”. This tweet did not include the #MeToo hashtag, but rather, included the hashtag, #HimToo. This hashtag was then used by many others in speaking out against false rape accusations. Analyzing this hashtag as well and comparing it to the #MeToo movement may be an interesting way to compare our own analysis to another one in terms of both sentiment and popularity.

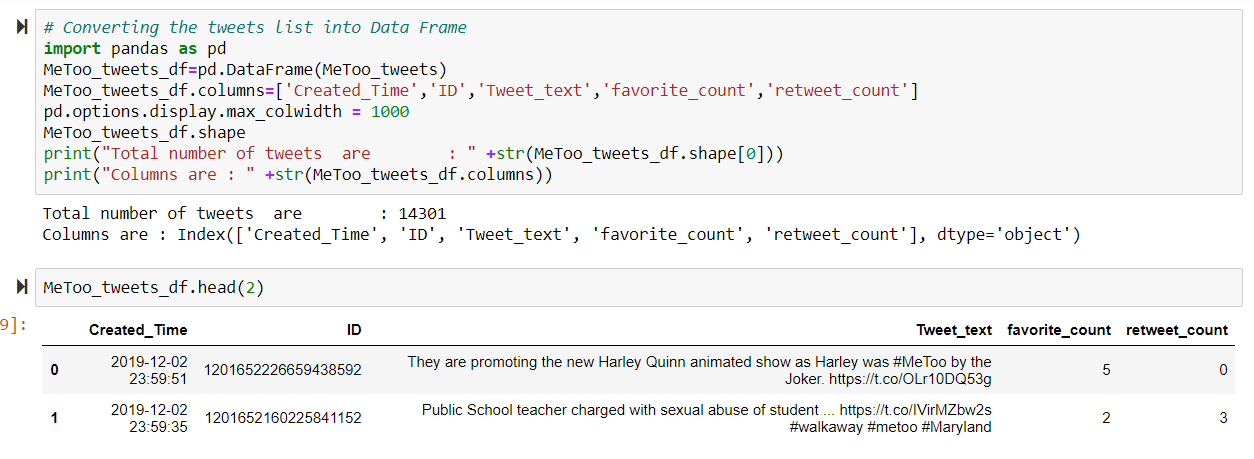
# Appendix A: Data Access and Cleaning

**Twitter Authentication –**

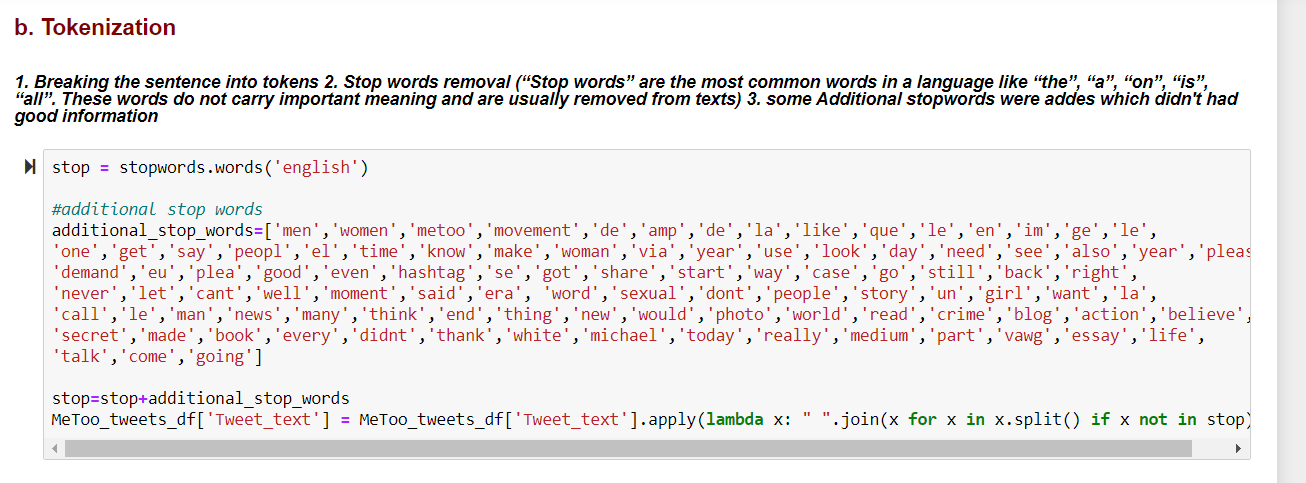


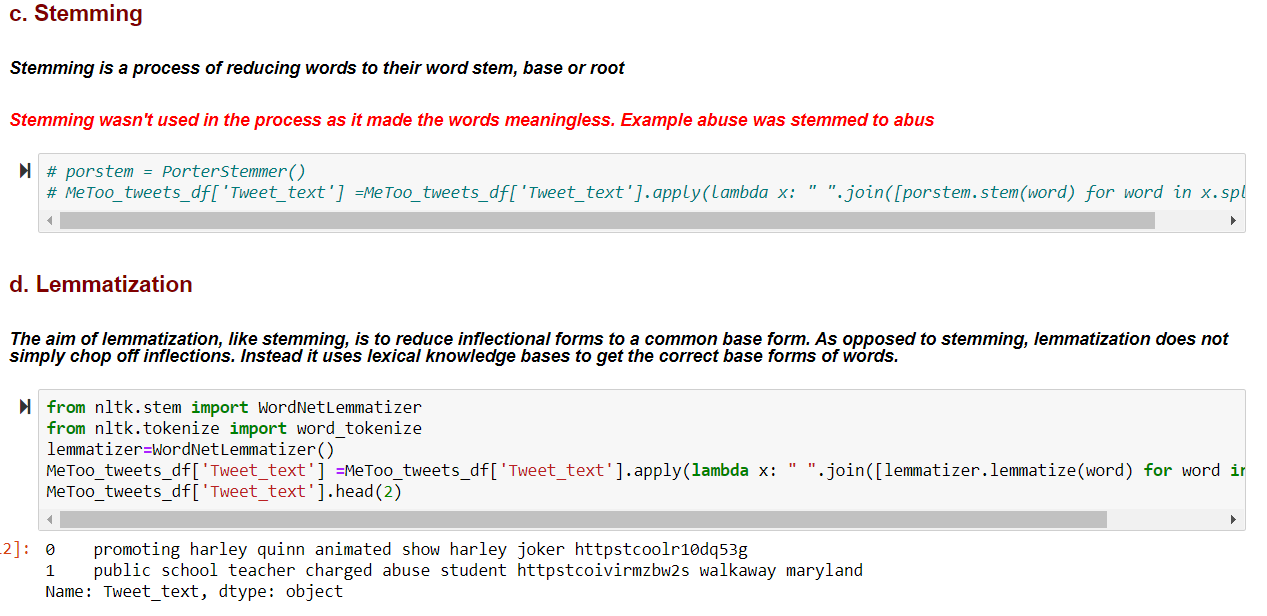
**Extracting Tweets –** 

**Converting Scraped Text to Data frame -**



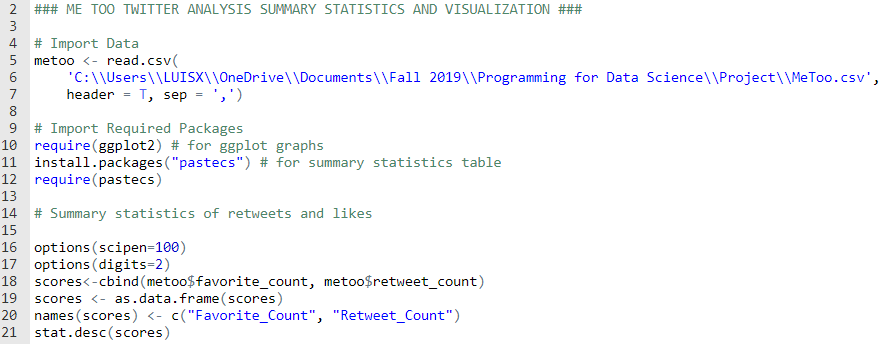


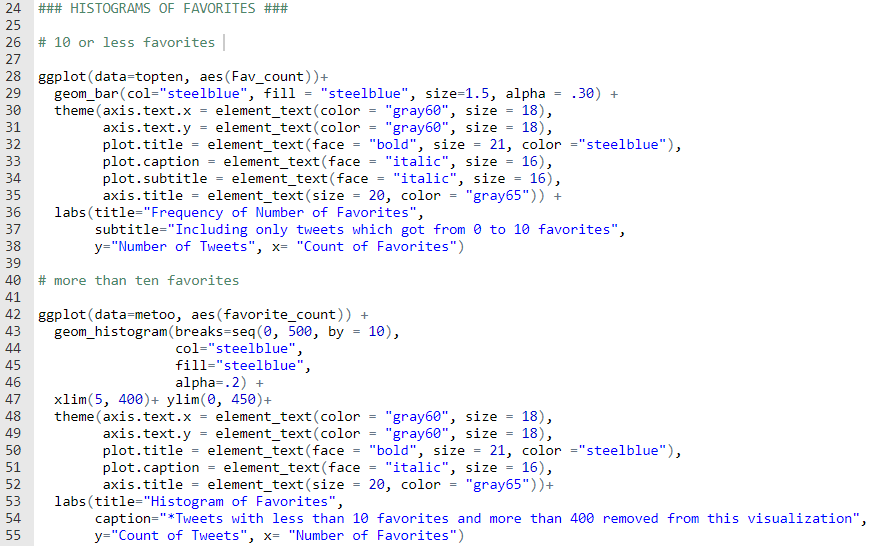


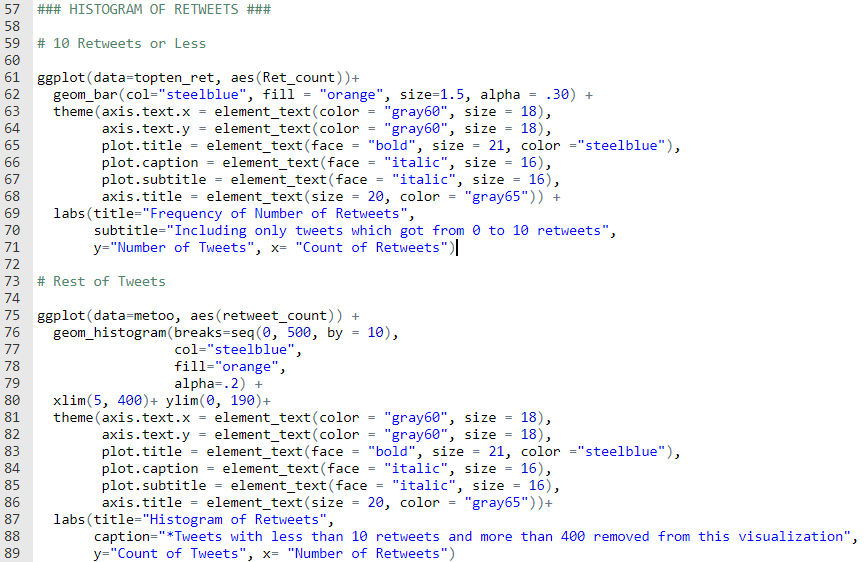


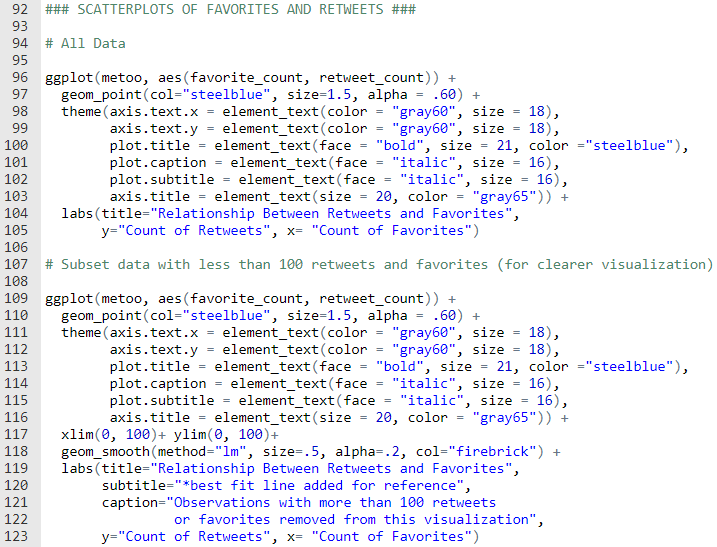


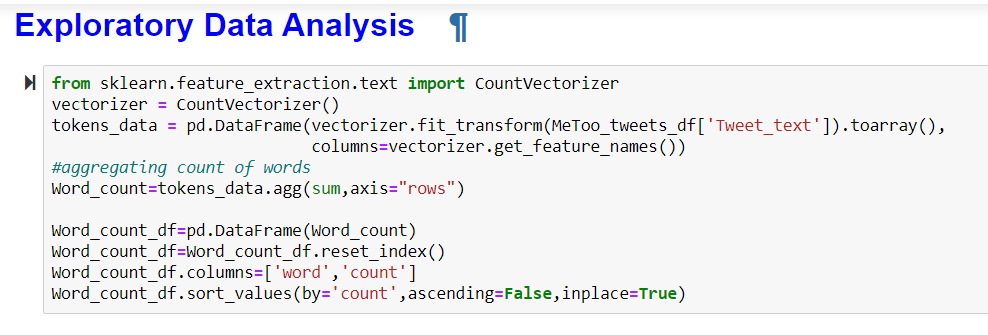
# Appendix B: Exploratory Data Analysis and Visualization



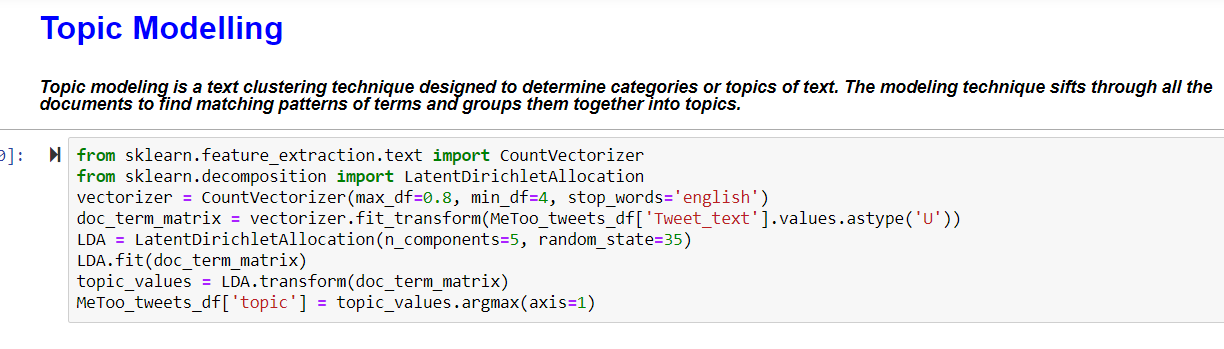








# APPENDIX C: TEXT ANALYSIS





**Sentiment Analysis –**

