

MSBA 305

Data Processing Framework

Presented to:

Dr. Elie Nasr

Presented by:

Karen Bassil

Kenny Barza

Emily El Hajj

**Introduction**

On November 10, 2020, Xbox Series X was launched worldwide. This next generation of Xbox decreased the overall load time and increased the performance and visual qualities of games. Two days later, PlayStation 5 was released with somewhat same improvement features of reducing greatly the load time while also enhancing the visual graphics. We, as the data analyst team of PlayStation, want to explore if our console launch was more successful than Xbox’s. To be able to do so, we will extract tweets from Twitter by searching for the two words: “PS5” and “Xbox Series X”, then store the tweet texts and its information into a data frame in Python. As well, we will extract tweets based on usernames of ‘PlayStation’ and ‘Xbox’. Then after cleaning all our data, we will create a table in MySQL and start inserting some of the cleaned data. Consequently, we will begin our analysis by applying basic statistical measures such as the methods of central tendencies and dispersion, the frequencies, and others.

**Database Engineering**

First things first, we used library tweepy in order to access twitter api and retrieve Data. We first retrieved 2000 tweets, using a keyword, in our case PS5, store them in a list, and finally store them in a data frame. After cleaning the data (Next Paragraph), we made sure to divide the data into two data frames. One is dedicated to information on the users. In this data frame, we included things like Followers, Friends, Name\_User, Statuses\_Count, Verified. In addition to that, we made sure to create a new column called User\_id, that uniquely identify every User that we get. The second data frame includes information about the tweet itself. You can expect to find columns such as Date, Likes, RTs, Source, Length, Tweets\_Cleaned, Subjectivity, Polarity, Sentiment. There is also a column called id which uniquely identify a tweet. Also, we made sure to include User\_id so that we can link the two data frame together. With the two data frame ready, we can now focus on creating our Database. As a first step, we connected Python to mysql using mysql.connector.connect(). We then created a cursor that will be used to create the database alongside this query. Finally we created a new database using this line of code *mycursor.execute("CREATE DATABASE Project3").* To create a table inside this database, We made sure that we connect to this exact database using mysql.connector.connect(). After that we create a new cursor and create our first table which contains information about the users. Here is the line of code used. *my\_cursor.execute("CREATE TABLE ps5\_users (User\_id INT ,Followers INT, Friends INT, Name\_User TEXT, Verified INT,Statuses\_Count INT, Primary key(User\_id))")*. Finally in order to insert automatically the rows into the table, we created a function that will enable us to insert a records. We did a for loop over this function and we were able to insert automatically the records. We did the same thing for the second dataframe, only difference here that we linked it first table with with use of a foreign key (User\_id). Here is the line of code for the creation of the second table. *my\_cursor.execute("CREATE TABLE tweets2 (id BIGINT ,user\_id INT,Date DATETIME,Likes INT,RTs INT, Source TEXT, length INT, Tweets\_Cleaned TEXT, Subjectivity FLOAT, Polarity FLOAT, Sentiment TEXT, Primary key(id),FOREIGN KEY (User\_id) REFERENCES Project3.ps5\_users(User\_id))")* . This whole process is repeated another time for the Xbox tweets. In that case, the name of the Database is Project5.

Finally, we Decided to retrieve tweets from Xbox and PS5 account. In this case, we used usertimeline() and we did not split our data into two data frames. In fact, we stored these data into a separate Database called Project6. We provided a database Scheme for more clarity purpose. We will inser them before section where we use the queries

1. **Data Cleaning**

We identified first that the column location has a lot of missing values. Also, those wrote the location did not specify a legit location. As a result, we decided to drop it completely. We made remove duplicate rows and to reindex our dataframe. Additionally, we cleaned further the data by removing the @mentions, the hashtag symbols, and the links by creating a function called cleanTxt where we used the re.sub statement. Furthermore, another function was formulated using the re.compile command to be able to remove all emoticons, symbols & pictographs, transport & map symbols, flags (iOS), Chinese char, dingbats, and many more. We also decided to add more features as discussed later on. We also cleaned the sources column and ensured that we only have the most important categories. We grouped the less important ones under the names of other. This was done using a list that contains the important elements and a for loop that loops through the data column source, and if the name of the source is not in the list, then we replace this name by Others.

1. **Analyzing tweets of users**

Before we start, for each question answered, we will be adding the query that is used as well as an explanation of it on the end of the report. Each query will be clearly referred to which question it was. We chose this method just for aesthetics.

* 1. **Sentiment Analysis**

1) What are people saying about both consoles? Is what they are saying positive or negative?

To begin our analysis, we created two columns that will include the polarity and subjectivity scores for each tweet. The polarity is a float value that ranges between -1 and 1, where the closer to 1, the greater the text presents a positive sentiment, and the closer to -1, the greater the text presents a negative sentiment. Also, the subjectivity is a float that ranges between 0 and 1 and indicates whether the text is a personal opinion, a judgement, or a fact. The greater the subjectivity, the more likely the tweet represents a belief.

Using the functions .mean() and .std(), we were able to calculate the average and the standard deviation of the polarity and subjectivity of the tweets. Although the mean of polarity for Xbox is higher than that of PS5, we cannot conclude that fans are talking about Xbox console in a more positive way than of PS5’s because as we can see that the subjectivities are as well low and that indicates that users are expressing a fact rather than an opinion about whether they like the product or not. And this will be backed up by the fact that in the next paragraphs we will determine that PlayStation is facing a problem in its supply and logistic chain, so users were nagging about the fact that they were not able to purchase the new PS5 and not complaining about the product features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Polarity** | | **Subjectivity** | |
|  | **Mean** | **Std** | **Mean** | **Std** |
| **PS5** | 0.0339 | 0.2867 | 0.3014 | 0.3307 |
| **Xbox Series X** | 0.0913 | 0.3094 | 0.3395 | 0.3219 |

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Xbox

PS5

Chart, box and whisker chart

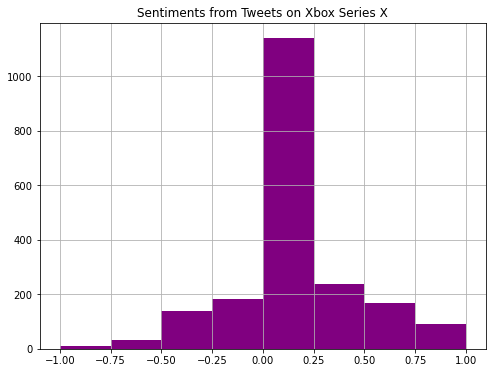
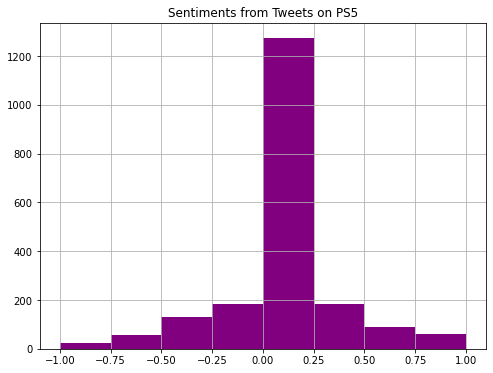
Description automatically generatedChart, box and whisker chart

Description automatically generated

Xbox

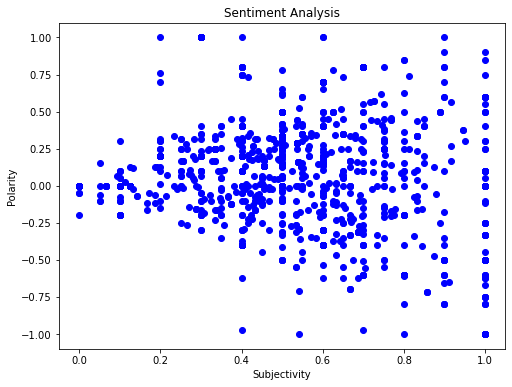
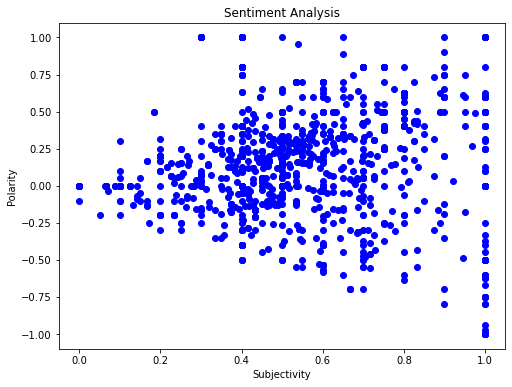
PS5

To further study the sentiments of tweets, we built two histograms for each of the products. We can visualize that both sentiments follow a somewhat normal distribution centered at the [0, 0.25] bin. Both graphs show high similarity from the range -1 to 0, however, the number of tweets for PS5 is slightly lower than for Xbox but this is only due to the fact that the PlayStation company was experiencing some struggles with their supplying process.

The below scatter plots explain the relationship between polarity and subjectivity for both products. In both cases, the variation in polarity increases as subjectivity increases. The higher the subjectivity score, the more the tweet is influenced by the user’s personal opinion, so the more likely that the absolute value of the polarity is closer to one.

PS5 Xbox Series X

* 1. **Tweets per minute**

2) Which users are the most excited about the product, PS5 fans or Xbox fans?

Analyzing the number of tweets per minute for each product will allow us to understand the magnitude of interaction on twitter and better understand which product is dominating and outperforming the other. As per the below table, the PS5 has an extremely higher number of tweets as compared to Xbox Series X, so we can safely and certainely conclude that users are excited more about the lauch of the PlayStation 5

|  |  |
| --- | --- |
|  | **Tweets per minute** |
| **PS5** | 41.58 |
| **Xbox Series X** | 3.80 |

* 1. **Word Analysis**

3) What insights can we extract from the tweets? How can we act upon it?

In this section, we will dig deeper into the use of words. After omitting Stop Words which are the common keywords in a language such as ‘is’, ‘a’, ‘the’, ‘who’, …, we were able to extract the most frequent used words between users.

As per the below figures, we can clearly visualize that ‘amazon’ is the most found word in tweets related to PS5. Plus, we can identify based on the word cloud that stock and restock are also extensively used. We can perhaps conclude that users are facing a problem regarding the stocking of PS5 specifically at Amazon stores. And based on this insight, PlayStation company must take action to satisfy its clients’ needs by supplying more PS5’s to Amazon or encouraging customers to shop directly from the PlayStation online store. This is extremely true if we relate it to what have been said on the news. In fact, some people- Scalpers – are dozen of ps5 only to resell them at higher price. This is one of the causes that is leading to this shortage of stocks and ps5 should act on in two ways. First, they should make marketing awareness related to this issue and advise people not to buy from scalpers. They should also find a way to meet the demand.

Chart

Description automatically generatedText

Description automatically generated

Regarding the most common used word in tweets related to Xbox Series X, we can also see from the below figures that shockingly the most common word found in the tweets is ‘PS5’ which indicates that people might be comparing the Xbox series x to the PlayStation 5. As a result, the launch of PlayStation 5 and its features are being recognized widely and people are talking about them a lot which indicates that the PS5 is in general a successful product or at least more successful than of Xbox’s. A lot of people are as well talking about the Microsoft flight simulator. This will be an exclusive title for the xbox which a positive sign for xbox since they do lack a lot of exclusive titles. Ps5 should not worry about this issue as they do have numerous exclusive titles, only playable on the new console

A picture containing chart

Description automatically generatedText

Description automatically generated

Further, and to be able to understand what users are tweeting about, we built a network word plot that shows the interconnections of words used in tweets.

In the below network figure, clearly there is a connection between the words ‘amazon’, ‘stock’, ‘restock’, ‘sold’, …, and another connection of words related to ‘got’, ‘one’, ‘get’, ‘trying’, so this supports our previous assumption which concluded that a shortage in PS5 supply is being presented at Amazon online stores where people are trying to get and purchase the PS5 machine.

Diagram

Description automatically generated

Additionally, the word network of Xbox Series X shows that users are mainly waiting for the launch of the Microsoft flight simulator in summer 2021. Also, on the lower left we can identify that some are having a giveaway for the Xbox Series X, and this might show that the Sales of Xbox are not high or at least are not as they should be so people and stores are posting giveaways to boost the shopping activity of Xbox’s.

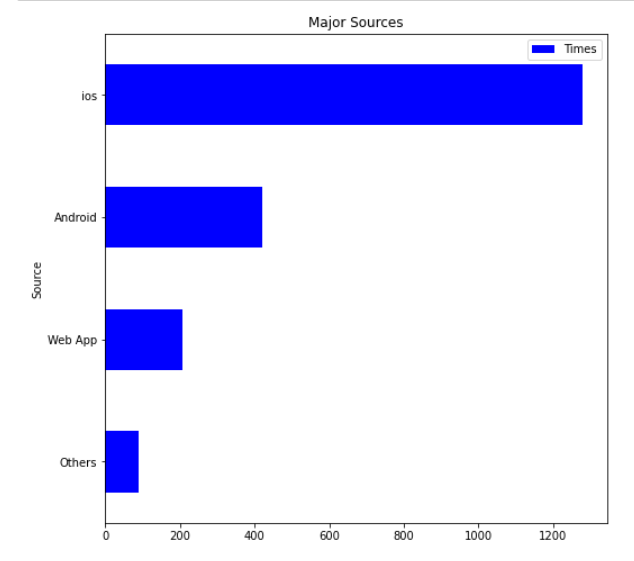
Timeline

Description automatically generated

**4. Sources**

4) From what sources People are tweeting?

In this section, we will be analyzing from which software people are tweeting the most about the PS5. In fact, we can clearly see that most of them are ios users (about 1300 tweets are from apple devices that uses ios).

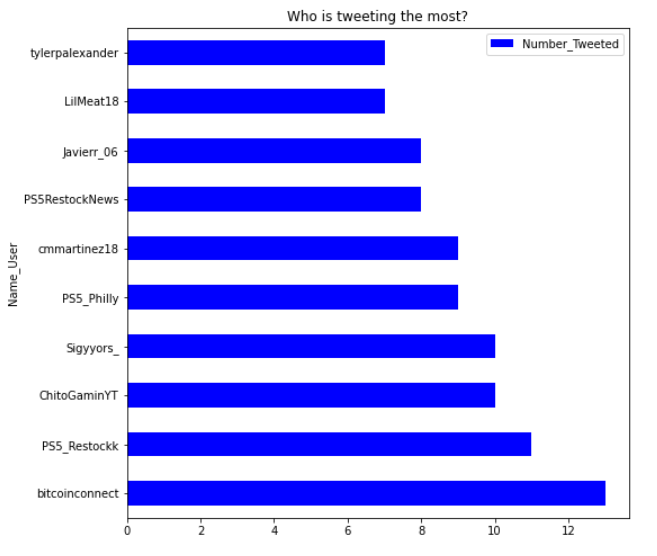


This is important and can lead to multiple ideas. For instance, if PS5 thought about introducing any of its exclusive title (ex: AstroBot) to mobile devices, it could negotiate with Apple and test the market there before launching it to other devices. That way, if the game did not succeed in the Appstore, it will surely not succeed on Android.

**5. Sources**

5) Who are the users that are tweeting the most?

In this section, we will find who are the users that are tweeting the most. This can lead to interesting marketing decision later on. Bitcoinconnect tweeted the most with 13 tweets. PS5\_Restockk comes next and he tweeted 11 tweets. The important insight to take from this is that these users are showing a lot of excitement towards the ps5 and exposing them to ads related to ps5 content might be a good idea.

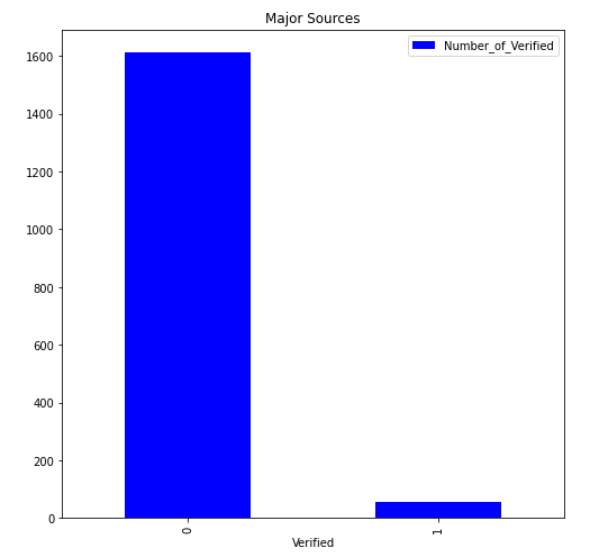
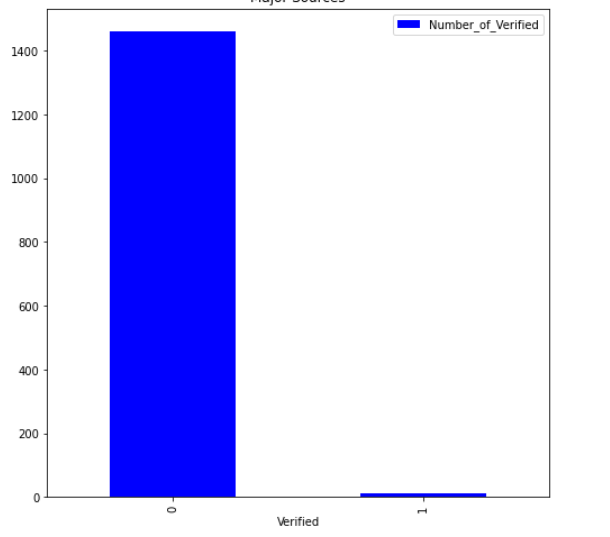


**6. Final metrics on who is doing better**

6) How many verified users there are?

In this section, we will be looking at the number of verified users. Verified users are in fact well known users and their voice can reach to multiple people. If for instance a well-known person is talking about a brand, this could indirectly help the brand. In our case, we could see that more verified people are talking about Xbox which could prove to be a negative thing Ps5. To fix this, PS5 can pay verified users in order to encourage them to talk about PS5 more often.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Verified | Not verified | Proportion |
| Ps5 | 11 | 1460 | 0.0075 |
| Xbox | 56 | 1612 | 0.034 |



Finally, we are going to displaying some quick measurements relatively useful. These measurements can have a link with how “popular” the users are. We can see that people who tweeted about Xbox have on average more like, Retweets and friends which is helping indirectly Xbox

|  |  |  |  |
| --- | --- | --- | --- |
| Brand Mean | Likes | RTs | Friends |
| PS5 | 2.1965 | 0.229 | 489.215 |
| Xbox | 18.695 | 1.590 | 838.56 |

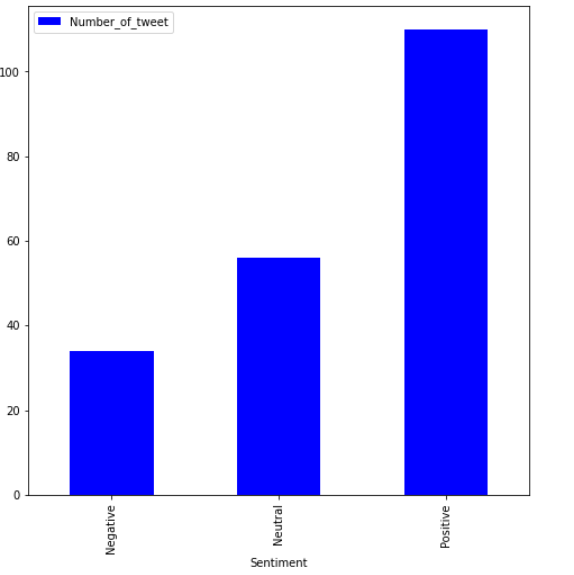
1. **Analyzing Tweets of PS5 Account**

In the consequent sections we will analyze the tweets of the accounts of PS5 and Xbox while generating insights that will benefit the PlayStation company to better utilize the Twitter Platform and retain additional customers. Our aim behind these sections is to allow PlayStation to improve its tweets.

* 1. **Sentiment Analysis**

7) What is the tone used by the ps5\_acc?

By building a histogram, we were able to detect that the social media team of PS5 was able to maintain a positive attitude in most of their tweets. However, few texts presented a negative tone thus, they should always keep on monitoring them to identify the ones that really result in a harmful outcome and decrease customer satisfaction. Note that Twitter Polarity and Sentiment are calculated based on word analyses where it would assign a low polarity score if the tweet includes negative words, but this does not always mean that the text is bad. For example, a tweet from the PS5 account is as follows: “Poor Colt, your weapon is jammed! In Deathloop, the DualSense controller’s adaptive triggers can block a press to each”, so we can clearly see that based on the word “Poor”, the tweet was identified as negative sentiment while in fact it is not. By displaying all the tweets using the following function : data.loc[ data[‘Sentiment’] == ’Negative’, ['Tweets\_Cleaned']], we were able to conclude that none of the extracted tweets indeed presents a negative sentiment.

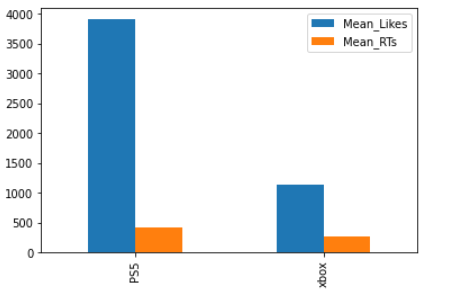


**2. Comparison Between Accounts**

8) Which account is doing better?

We will be comparing the performance of the PlayStation account tweet against the Xbox account. We can clearly see that on average, PS5 get a lot more likes to their tweets compared to Xbox. Also, the number of Retweets is also better in the case of the PS5. We can say that PS5 are already on the right track. But we can further improve our performance, in order to distance ourself more from our competitors, before they catch up. That is what we are going to do in the next section.

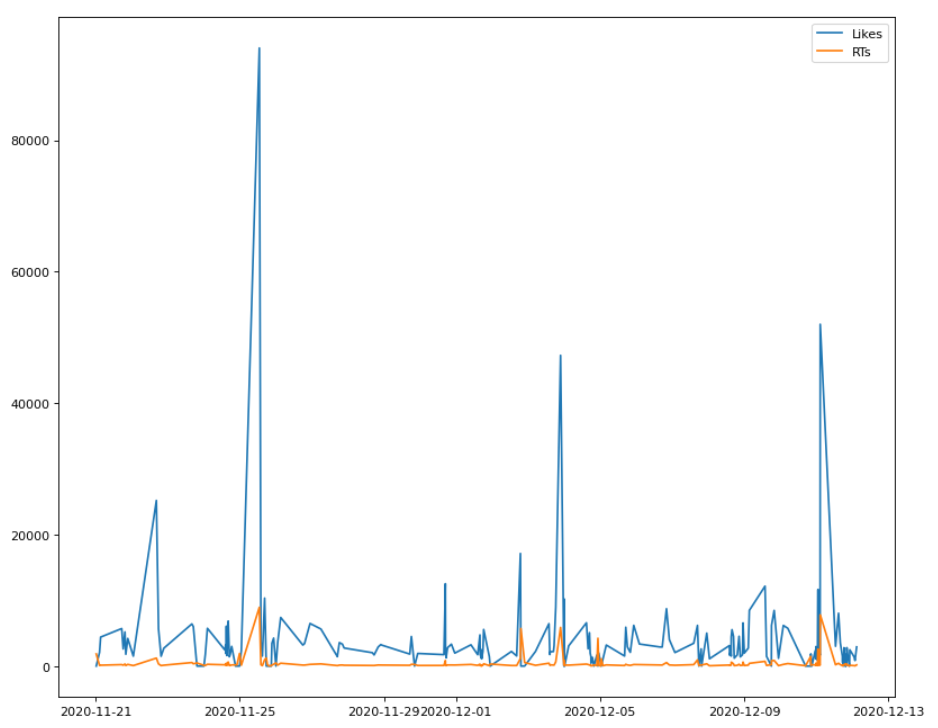
|  |  |  |
| --- | --- | --- |
| Mean  Brand Name | Likes | RTs |
| PS5 | 3913.315 | 424.755 |
| Xbox | 1129.080 | 265.890 |



**4. Tweet Time Series**

9) When are we receiving the most likes and Retweets?

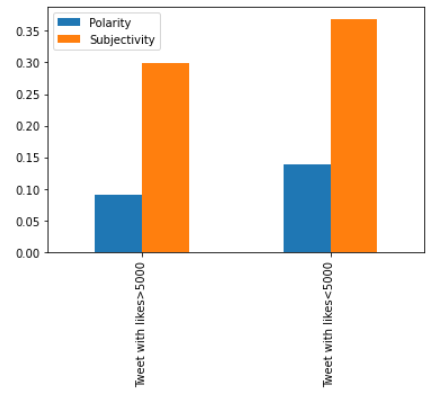
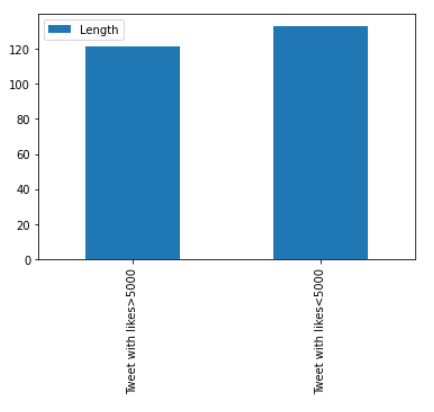
When plotting the likes and Retweets as the time goes, we 3 important things. First, whenever we have a peak in likes, we have a peak in retweet which obvious, since people who are liking the tweet might as well retweet it. Second, we can identify 4 massive peaks. Third, these peaks are spaced at 4-7 days each, meaning that for this time span, we can expect a from PS5 to go viral. It would be interesting to analyze what these tweets have in common. That way, PS5 can use the same characteristics in order to maximize their chance of earning more likes and Retweets. That is what we are going to do next.



**5. Analyzing tweets that have 5000+ likes**

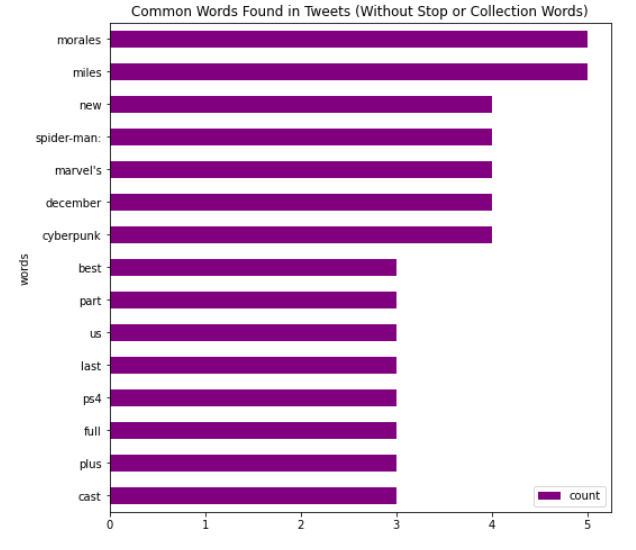
10) What is the optimal length, polarity and Subjectivity of a tweet?

When comparing tweets that got more than 5000 likes and tweets that got less, we can clearly see that the average length of the viral tweets is smaller. This is and interesting finding. It goes to show that people nowadays are always busy, and they barely have time to browse twitter. That is why they will be more receptive to shorter tweets. Also, viral tweet tends to have a polarity close to 0, meaning they are neutral tweet. Also, their subjectivity is low, equal to approximately 0.28. As a result, it is recommended that PS5 uses neutral tone and relatively short tweet to maximize their chance of getting more likes



11) What words are being used in the tweets that went viral?

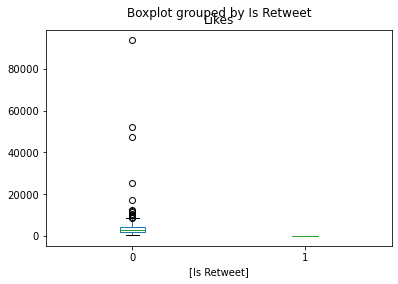
The tweets that went viral (5000+ Likes), include words like “spider-man”, “miles”, “cyberpunk”, all of which are games that were or will be released on the new console. It seems that people are interested in these games and as a result, these tweets are getting more likes. Interestingly enough, we can see that some of these tweets have the word PS4 in it. This ensure that some people are still using their PS4 and are happy to see that some of these games will also be launched on the old console. As a recommendation, PS5 should use more often the name of famous game in order for people to be excited. Also, since people are engaged with tweets related to cyberpunk as an example, PS5 can launch new accessories with cyberpunk theme (PS5 controller with Cyberpunk theme). It would surely sell based on what we have seen as results.



1. **Variables Association**

Both a boxplot and a histogram are computed to determine if there is an association between the number of likes and the tweets that are actually a retweet from another account. In other words, we can see that the retweets done by PlaySatation are not getting any likes. Consequently, the PlayStation team can either improve the retweets by targeting the tweets for which its followers have interest in or stop retweeting and focus on enhancing their own tweets.

**A picture containing logo

Description automatically generated**

Chart, box and whisker chart

Description automatically generatedAnother association to test is whether the length of tweets impact the total number of likes that a tweet will get. Let us first indicate the mean and standard deviation of the length of tweets by computing the .mean() and .std() commands while also visualizing them using a boxplot.

We can see that the majority of the tweets are of length greater than 130 characters, where in fact the mean is equal to 131 characters and the standard deviation is 19.23 which means that in general the tweets vary from 110 to 150 characters on average.

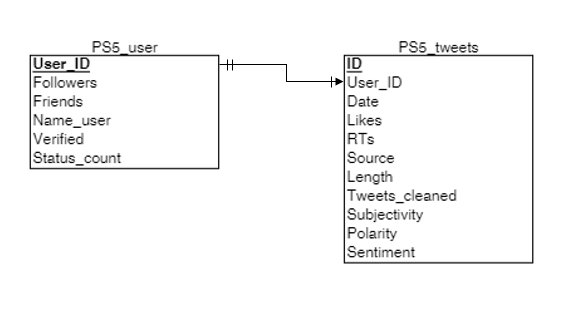
Moreover, a scatterplot was constructed, however, we cannot conclude clearly if there is a relation between both variables, the Likes and len (length of tweets). Thus, we calculated the Pearson correlation by using the following query “pearsonr(data['Likes'], data['len'])[0]” and we got an output of -0.227. The latter indicates that holding everything constant, if the length of a tweet increases by one unit, the total number of likes will decrease by 0.227. We can then recommend PlayStation to aim at reducing the length of their tweets which will allow their users to rapidly read and engage with their texts.

Chart, scatter chart

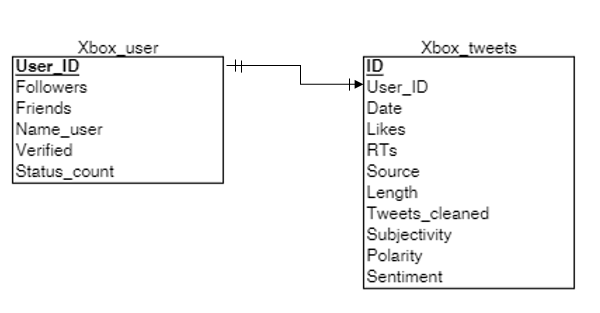
Description automatically generated

**Databases**

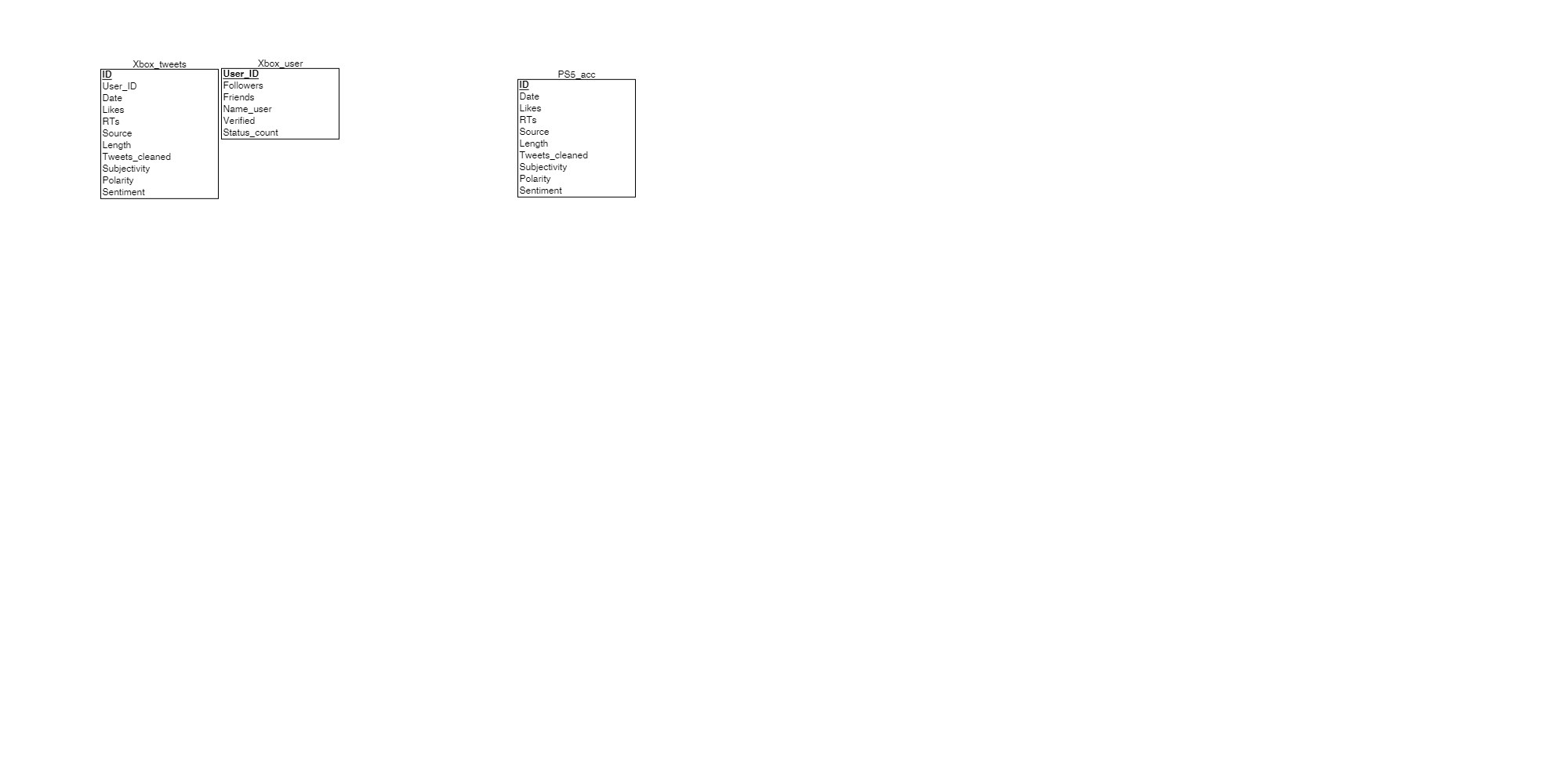
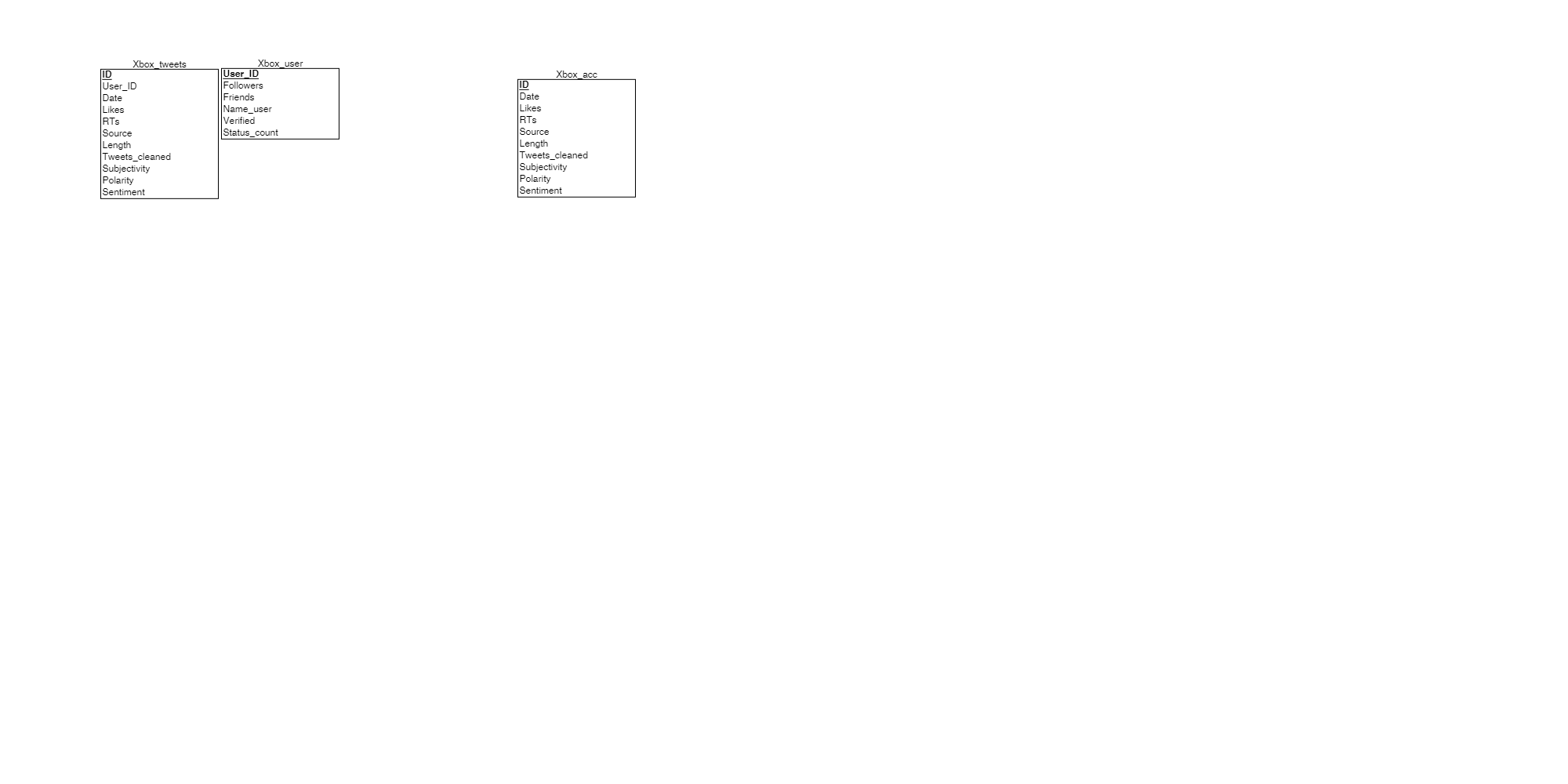
Project 3 Database



Project 5 Database



Project 6 Database



**Query Used**

We Used two libraries, sqlachemy and pymysql to do these tasks. This will allow us to retrieve data from mysql and put them as a pandas data frame. It really works like magic! **In bold, you can check the query that was used.**

1) What are people saying about both consoles? Is what they are saying positive or negative?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Polarity ,Subjectivity, Sentiment FROM Project3.ps5\_tweets’****, con=db\_connection)*

Explanation:

For this task, we needed the polarity, subjectivity and Sentiment column of the “tweets2” which is the table that contains tweets related to PS5. After doing the query, we got ourselves a data frame that contains these 3 columns. With some additional code lines mentioned in the code, we were able to get all the results mentioned above. We did the same query for Xbox and the compared both of them in the report. Only difference is that the name of the database and table.

2) Which users are the most excited about the product, PS5 fans or Xbox fans?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Date FROM Project3.ps5\_tweets Order by Date Desc'****, con=db\_connection)*

To calculate the number of tweets per minute, we only needed to extract the date column. We made sure to order them by Descendent Order so that it would be easier to calculate the difference between the last tweet and the most recent tweet. Finally, we divide the difference by the total number of tweets. (The process is a little bit more complicated than that, you can refer to the python code for more details)

3) What insights can we extract from the tweets? How can we act upon it?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Tweets\_Cleaned FROM Project3.ps5\_tweets****, con=db\_connection)*

To do some basic text analytics, we only needed the tweets from the tables. As a result, we just queried the tweets from the table plotted the wordcloud and the barplot displaying the most used words. Again, we did the same thing for xbox as well.

4) From what sources People are tweeting?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Source, Count(\*) As Times FROM Project3.ps5\_tweets Group By Source Order By Count(\*)****', con=db\_connection)*

With this query, we were able to create a data frame with source and times as columns where times represent how the number of times a source was used. The data frame obtained was then used to display the bar plot.

5) Who are the users that are tweeting the most?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Name\_User,Count(\*) As Number\_Tweeted FROM Project3.ps5\_tweets,Project3.ps5\_users Where ps5\_tweets.user\_id=ps5\_users.User\_id group by Name\_User Order By Count(\*) DESC Limit 10'****, con=db\_connection)*

Here, we needed to join the two tables. We tried to create a dataframe that contains the Name\_User as a column and the number of tweets, ordered by count(\*) and finally in descendant order while only limiting it to 10. We then directly used a bar plot to display the results

6) How many verified users there are?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project3'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Verified, Count(\*) As Number\_of\_Verified FROM Project3.ps5\_users Group by Verified****', con=db\_connection)*

Here, I was able to get a data from our table ps5\_users that has the Verified and Number\_of\_Verified as columns. Note that we used group by in order and count to have a data frame that counts how many verified and non-verified there is. After that, we used a bar plot to display the results.

7) What is the tone used by the ps5\_acc?

*db\_connection\_str = 'mysql+pymysql://root:kenken123@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Sentiment, Count(\*) As Number\_of\_tweet FROM Project6.ps5\_acc group by Sentiment order by Count(\*)****', con=db\_connection)*

In this query, we retrieved the sentiment and grouped by number\_of\_tweets. This was done on the ps5\_acc data. We used the data frame that we got in order to display the barplot

8) Which account is doing better?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df\_ps5 = pd.read\_sql(****'SELECT Likes, RTs FROM Project6.ps5\_acc'****, con=db\_connection)*

Here, we wanted to check how many likes and RTs on average each account has. We queried a data frame that contains only these two parameters. We then computed the mean of likes and Rts . We did the same thing for Xbox and compared the results in a barplot

Note that this could also be done otherwise, by using directly select mean(Likes) from…

9) When are we receiving the most likes and Retweets?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Date, Likes, RTs FROM Project6.ps5\_acc'****, con=db\_connection)*

Here, we needed Date, Likes and RTs to compute the Time Series plot. This is exactly what was done in this query. The dataframe is later used to create the plot.

10) What is the optimal length, polarity and Subjectivity of a tweet?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT \* FROM Project6.ps5\_acc Where Likes>5000'****, con=db\_connection)*

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df2 = pd.read\_sql(****'SELECT \* FROM Project6.ps5\_acc Where Likes<5000****', con=db\_connection)*

In order to compare between tweets that got higher than 5000 and tweets that got lower than 5000, we created two data frame and we made sure to insert where likes>5000 for the first data frame and where likes<5000 for the second data frame. These are later on used to compare things like the mean of likes, the mean of length etc.

11) What words are being used in the tweets that went viral?

*db\_connection\_str = 'mysql+pymysql://root:password@localhost/Project6'*

*db\_connection = create\_engine(db\_connection\_str)*

*df = pd.read\_sql(****'SELECT Tweets\_Cleaned FROM Project6.ps5\_acc Where Likes>5000****', con=db\_connection)*

Here, we retrieved only the tweets from ps5\_acc that has 5000+ likes. We then did the barplot to display what words are mostly used