DATA610 – Introduction to Data Mining
Group Project 2 – Cluster Analysis

Prepared For:
Dr. Rasitha Jayasekare

Prepared By - Group 6:
Anthony Orso
Oliver O'Keefe
Walter Miller

Date:
April 30, 2022

Table of Contents

I. Introduction and summary of project

The teams within the DATA610 class have been presented a set of data to analyze using clustering techniques. The goal is to use the practices that the students are learning in the class to analyze the datasets and complete the goals outlined below.

The dataset represents Facebook consumer engagement data for 10 Thai beauty and fashion retailers. The data contains information regarding user engagement with specific post types and timeframes. Additional knowledge could be helpful along the way, such as the fact the Facebook Live was launched to the general public in 2016. This information leads to expectations that video may become more impactful after this time. This data contains predictor variables, but no response variables. As such, it is considered unlabeled data and we have a goal to use unsupervised techniques to cluster the samples. Once clustering is complete, the team will perform post processing to understand why the specific observations were grouped together as well as differences across the clusters.

II. Approach:

In this report, the team will be analyzing a single dataset, FBOnlineSales, to identify patterns in the data. We will utilize clustering techniques that we are learning to determine how the dataset can be analyzed in an unsupervised manner.

The structure of this document is broken into several sections listed below culminating in the final interpretation and discussion.

1. Data Descriptions:
   In this phase we are getting to know the data that we will be working with. Understanding the predictors is key to this section. You will see the data analyzed both numerically and visually to help build this understanding.

2. Preprocessing of data:
   In this phase of the project, we are preparing the data for work in our models. We decide on the types of the variables we will be able to use use, as well as which ones are categorical and numerical. Outliers are identified but not removed from the dataset (base on the instructions).

3. Applying Cluster Analysis Techniques:
   In this phase of the project, we are applying multiple cluster analysis techniques to the dataset.

4. Select Cluster Technique to Analyze:
   In this phase, we will select a single cluster analysis technique from the ones performed in the previous phase.

5. Conduct post analysis:
   In the last phase of the project, the team will provide overall analysis of the selected clustering technique. We will be looking for:
   - Any interesting patterns?
   - Any interesting groupings?
   - Similarities within each group?
   - Differences between groups?

   This section will contain written analysis, as well as visual analysis (using ggplot2) to summarize the findings from the clustering output.

III. Data Descriptions

A. Numerical Descriptions

Dataset source:

https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand

Dataset Variables:

The first step in any data project is to understand the data that you are working with. The flowing list represents the set of variables in the FBOnlineSalesData dataset.

```
 [1] "status_type"      "status_published" "num_reactions"    "num_comments"     "num_shares"
 [6] "num_likes"        "num_loves"        "num_wows"         "num_hahas"        "num_sads"
[11] "num_angrys"       "status_type_num"
```

Determine the types of the variables:

In addition to understanding the available variables, it is critical to determine the types of these variables. In many cases, the actual types of the variables are not correct when imported and automatically recognized by many tools. In these cases, it is important to convert the types appropriately to continue to use these variables.

In the FBOnlineSalesData dataset most of the variables were imported as integers, but, two of the variables were imported as strings. Analyzing the data, it is clear that one of the variables is categorical. In particular, the status_type field needed to be identified as a categorical (factor) in the data with 4 levels (link, photo, status, video).

Additionally, it is interesting that the status_published variable is a date and time value that is combined in a single variable. We will consider separating this into multiple variables through this process.

Final Structure of data:

Once all of the variables are defined correctly in the dataset, the following output from our tool confirms our conversions have all been done correctly.

```
'data.frame':     2000 obs. of  11 variables:
 $ status_type     : Factor w/ 4 levels "link","photo",..: 2 2 2 1 4 4 2 2 2 2 ...
 $ status_published: chr  "2/21/2018 3:05" "6/23/2017 9:24" "3/8/2016 2:43" "5/22/2015 8:21" ...
 $ num_reactions   : int  92 1 139 1931 272 8 1271 53 14 10 ...
 $ num_comments    : int  14 0 4 26 339 0 94 3 0 1 ...
 $ num_shares      : int  1 0 0 0 158 0 9 0 0 0 ...
 $ num_likes       : int  89 1 138 1931 156 8 1249 51 14 9 ...
 $ num_loves       : int  3 0 0 0 112 0 11 2 0 0 ...
 $ num_wows        : int  0 0 1 0 3 0 8 0 0 0 ...
 $ num_hahas       : int  0 0 0 0 1 0 2 0 0 0 ...
 $ num_sads        : int  0 0 0 0 0 0 1 0 0 1 ...
 $ num_angrys      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Numerical Description Summary:

After initial variable discovery, and conversion of the variables to the appropriate types, the following table represents our variable analysis.

Variable	Type	Description
status_type	factor	Represents the type of post. “link”, ”photo”, ”status”, ”video”

status_published	character	Represents the date and time together in one field in a string format.
num_reactions	integer	Represents the total number of reactions to a post. This represents the sum of likes, loves, wows, hahas, sads, and angrys, below.
num_comments	integer	Represents the number of comments posted about a post.
num_shares	integer	Represents the number of times that a post has been shared.
num_likes	integer	Represents the number of "like" reactions that a post received.
num_loves	integer	Represents the number of "love" reactions that a post received.
num_wows	integer	Represents the number of "wow" reactions that a post received.
num_hahas	integer	Represents the number of "haha" reactions that a post received.
num_sads	integer	Represents the number of "sad" reactions that a post received.
num_angrys	integer	Represents the number of "angry" reactions that a post received.

Numerical Analysis:

Once the variables are understood and appropriately "typed", it is time to move on to numerical analysis of the data to learn more about the data. Numerical analysis will focus on the quantitative and qualitative variables, working to understand the data itself. We will be seeking to understand numerical properties of the data, such as mean, standard deviation, max, min, frequency(categorical), and more. First, we will focus on the numerical representations of the dataset, and then we will transition to a visual analysis of this dataset.

A sample of the first few rows of the data looks like:

```
  status_type    status_published num_reactions num_comments num_shares num_likes num_loves num_wows
1       photo   2/21/2018 3:05            92           14          1        89         3        0
2       photo   6/23/2017 9:24             1            0          0         1         0        0
3       photo    3/8/2016 2:43           139            4          0       138         0        1
4        link   5/22/2015 8:21          1931           26          0      1931         0        0
5       video  11/26/2017 8:01           272          339        158       156       112        3
6       video  11/10/2017 4:58             8            0          0         8         0        0

  num_hahas num_sads num_angrys
1         0        0          0
2         0        0          0
3         0        0          0
4         0        0          0
5         1        0          0
6         0        0          0
```

The following output from our tool gives a good overview of the numerical properties of the dataset.

```
 status_type   status_published   num_reactions     num_comments       num_shares
  link  :  19   Length:2000        Min.   :   0.0   Min.   :    0.0   Min.   :   0.00
  photo :1213   Class :character   1st Qu.:  18.0   1st Qu.:    0.0   1st Qu.:   0.00
  status: 103   Mode  :character   Median :  58.0   Median :    4.0   Median :   0.00
  video : 665                      Mean   : 236.5   Mean   :  200.3   Mean   :  40.94
                                   3rd Qu.: 217.0   3rd Qu.:   23.0   3rd Qu.:   5.00
                                   Max.   :4410.0   Max.   :10960.0   Max.   :1412.00

   num_likes         num_loves         num_wows         num_hahas          num_sads
 Min.   :   0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.000   Min.   : 0.00
```

```
 1st Qu.:  17.75   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.: 0.00
 Median :  55.50   Median :  0.00   Median :  0.000   Median :  0.000   Median : 0.00
 Mean   : 221.29   Mean   : 12.78   Mean   :  1.357   Mean   :  0.732   Mean   : 0.21
 3rd Qu.: 177.00   3rd Qu.:  3.00   3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.: 0.00
 Max.   :4315.00   Max.   :460.00   Max.   :278.000   Max.   :102.000   Max.   :23.00

   num_angrys
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 0.000
 Mean   : 0.119
 3rd Qu.: 0.000
 Max.   :31.000


Standard Deviation mapping of the quantitative variables

num_reactions  num_comments    num_shares     num_likes     num_loves      num_wows     num_hahas
      477.777       702.759       124.039       465.090        38.255         9.614         4.241

     num_sads    num_angrys
        1.143         0.865
```

From this data, we can make a few observations:

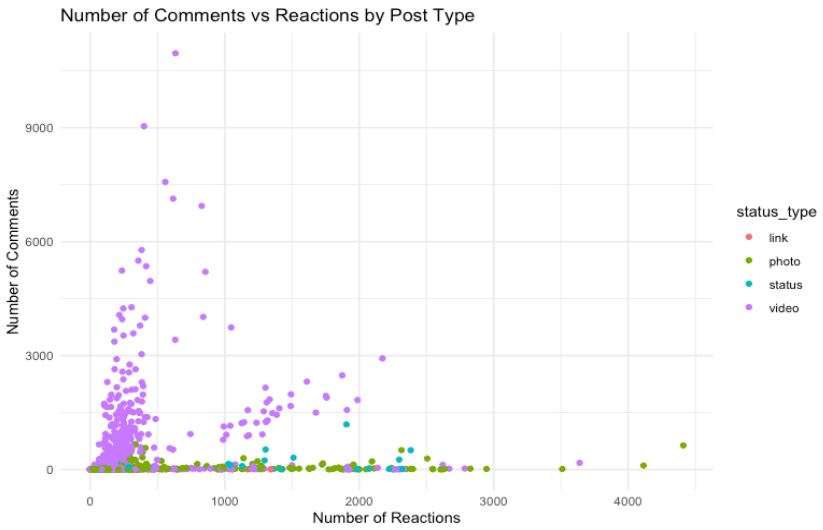
- The num_reactions is the sum of the different types of reactions (num_likes, num_loves, num_wows, num_hahas, num_sads, num_angrys).
- Positive reactions are the most common types of reactions with "likes" being the most common type of reaction and "loves" being second.
- Negative reactions of "sads" and "angrys" are the least commonly used reactions in the dataset.
- Photos are by far the most common type of posts in the dataset, with 1213 of the 2000 observations.
- The status_published variable is a character string that contains both a date and time. It will likely be advantageous to separate these for them to have meaning.

B. Visual Descriptions

Once basic numerical analysis is complete, it is important to move on to analyze the variables and their relationships visually. We will create visual representations of the data and use several multivariate graphs to explore the data.

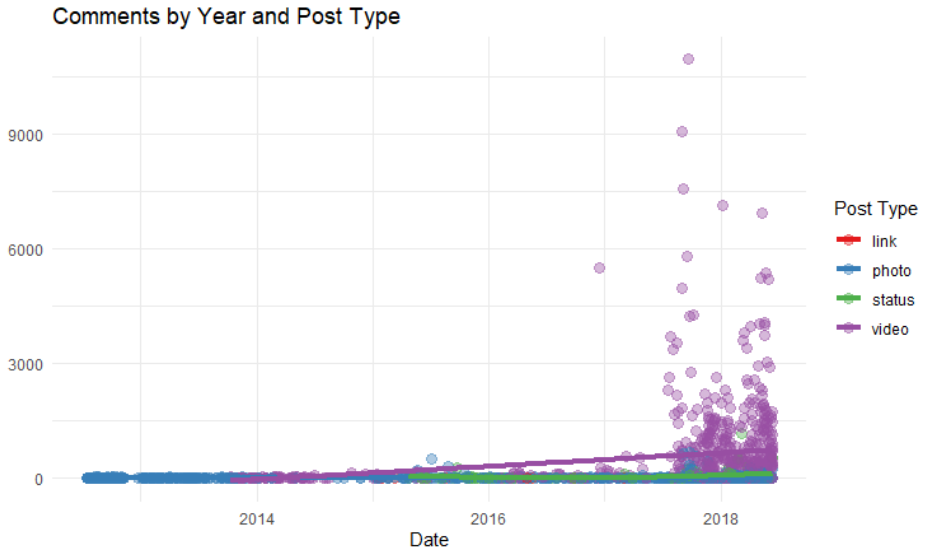
Number of comments vs. Number of reactions by Post Type:

We want to understand if the type of post (status_type) drives different levels of comments or reactions.



From this graph, it is clear that videos drive consistently more comments than any other type of posts. It seems like a few of the photo type posts also drive many reactions as well. Overall, from this visual representation it appears that videos and photos draw more comments and reactions than link and status type posts.

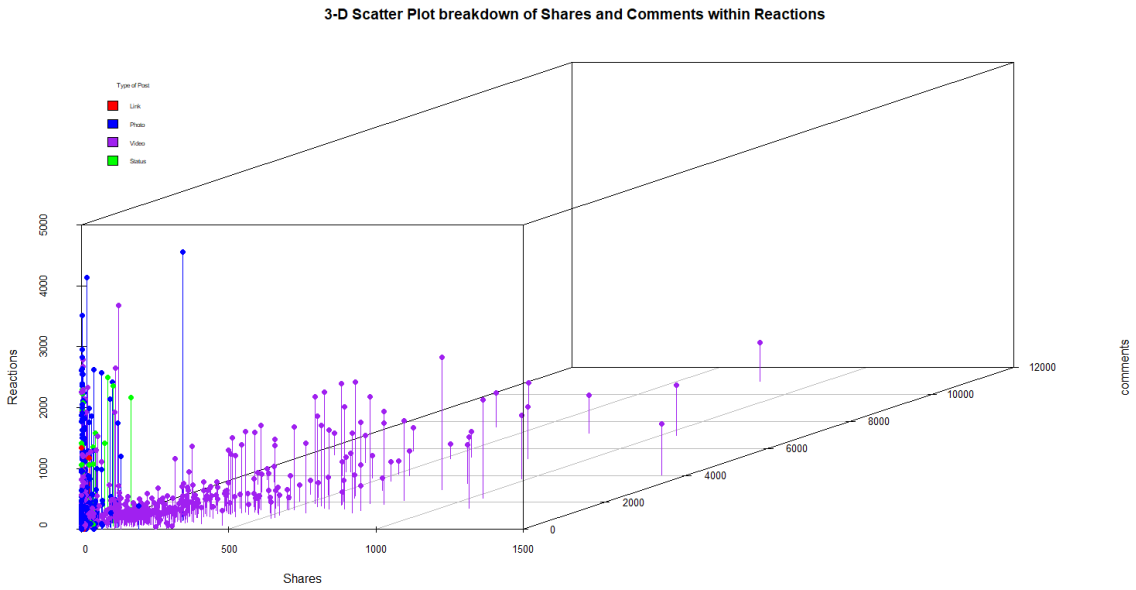
How people comment over time is a worthwhile measurement to observe. It can be argued that a comment is the most valuable form of engagement with a post on Facebook as it opens opportunities for continuous discussions. With the way Facebook shows content every comment under a post has a small chance of being shown to all your friends, the more times you comment the more times this might happen.



By using a scatterplot with best-fit- lines, we can observe an overall trend in general interaction across all forms of posts over time. Most notably is the sharp increase in 2017-2018 over video posts. While we cannot necessarily attribute any one thing that would spur the userbase to comment more we can assume the reason there are more

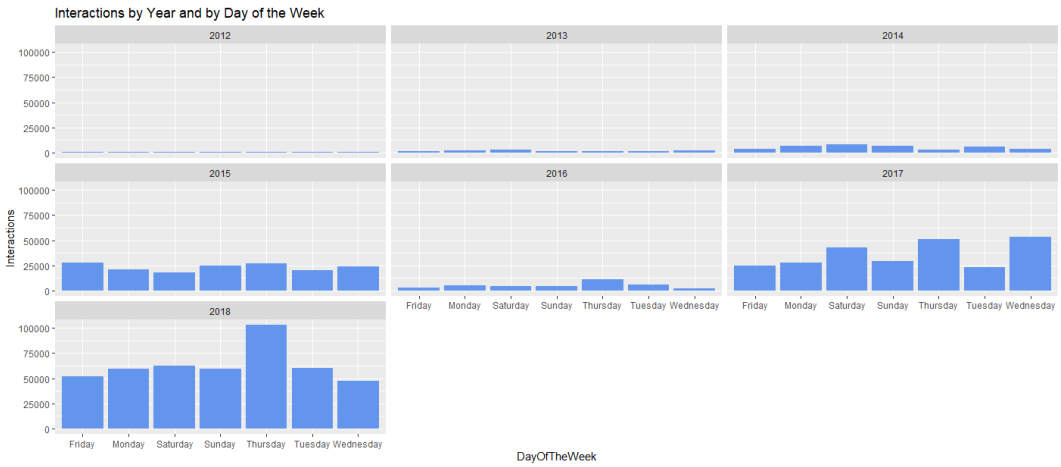
comments under video posts is because this type of post is clearly favored by Facebook's algorithms.

It is also important to look at general interactions on the same posts. Posts that had a lot of comments would not necessarily have a lot of shares. Or posts with a lot of shares would not have an equal number of reactions.



By mapping each individual post to a 3-D scatterplot of reactions, shares, and comments we can disprove some common bias held towards engagement. First, as stated above, a high number of one type of engagement does not guarantee comparable results in other forms of engagement. Second, certain types of post are inherently linked to certain types of interactions. Video posts excel in receiving shares and comments, but, on average, are beaten out in reactions by all other types of posts. This shows evidence to a pattern of behavior based on the type of post. We see meaningful engagement outside of what the preferred type of post is according to Facebook's algorithm.

Finally, it would help us to understand when people are using Facebook. This is a measure across all types of interactions (shares, comments, and reactions) divided by what day of the week they were posted and what year they were posted.



Based on the data above, posts are more likely to garner interactions from Wednesday – Saturday. Thursday stands out as a particularly high example, though, for reasons we cannot fully understand. We can also notice a general trend of increased post interaction throughout the years.

IV. Preprocessing of Data

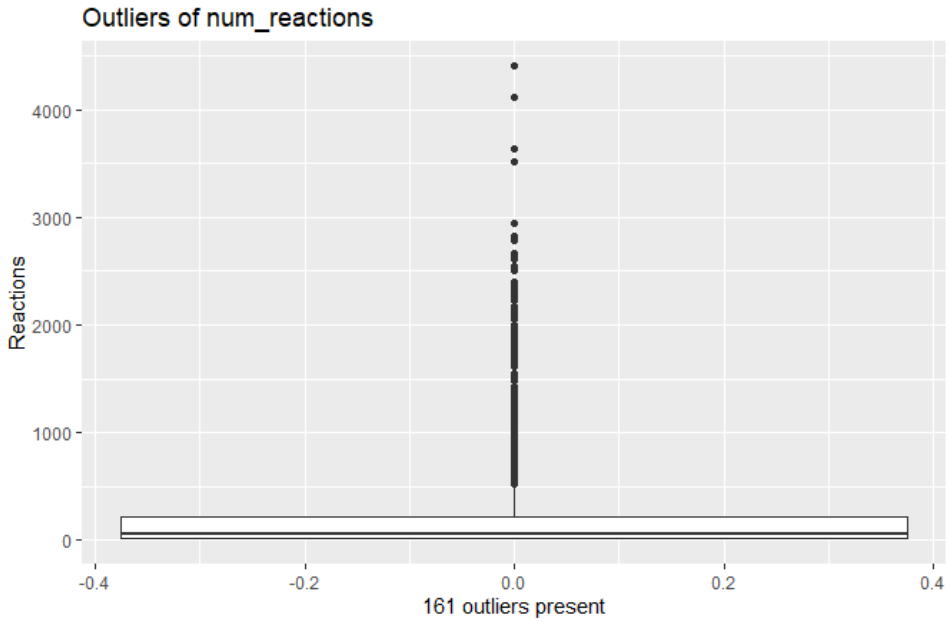
A. Missing value identification

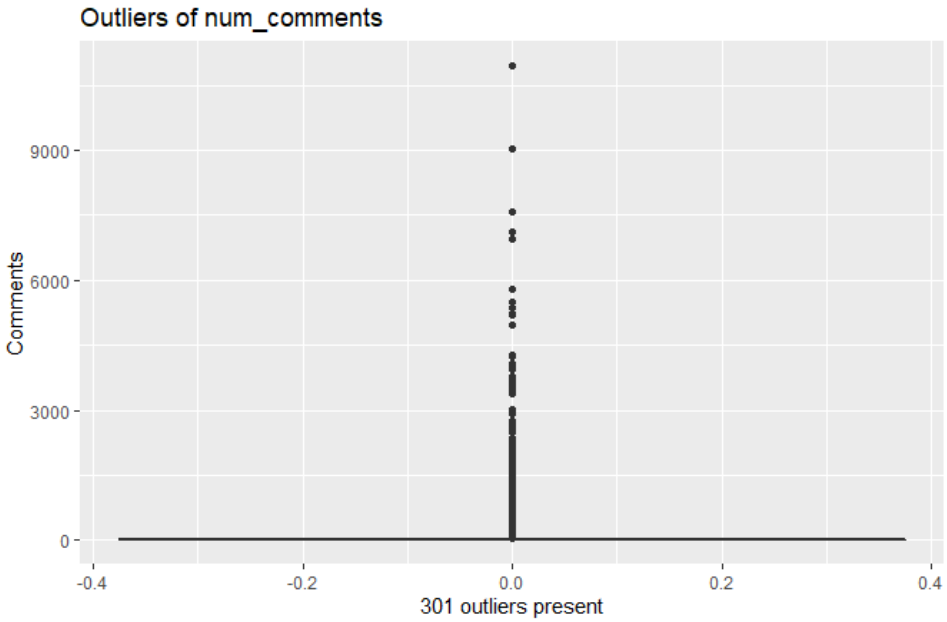
After using our tool analyzing the 2000 samples in our FBOnlineSalesData subset, we have determined that there are no missing values among our 11 variables.

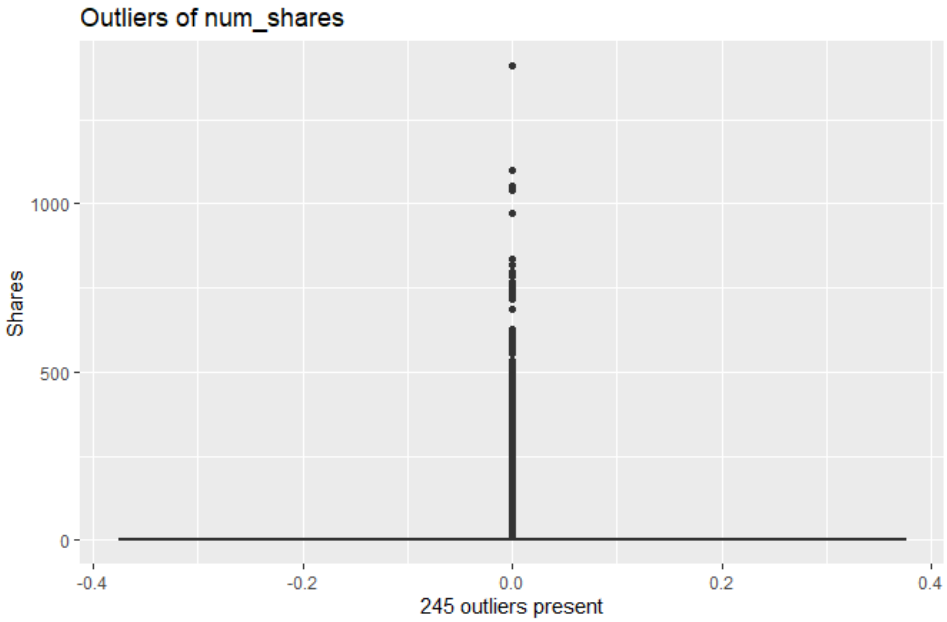
B. Outlier identification (Not removal)

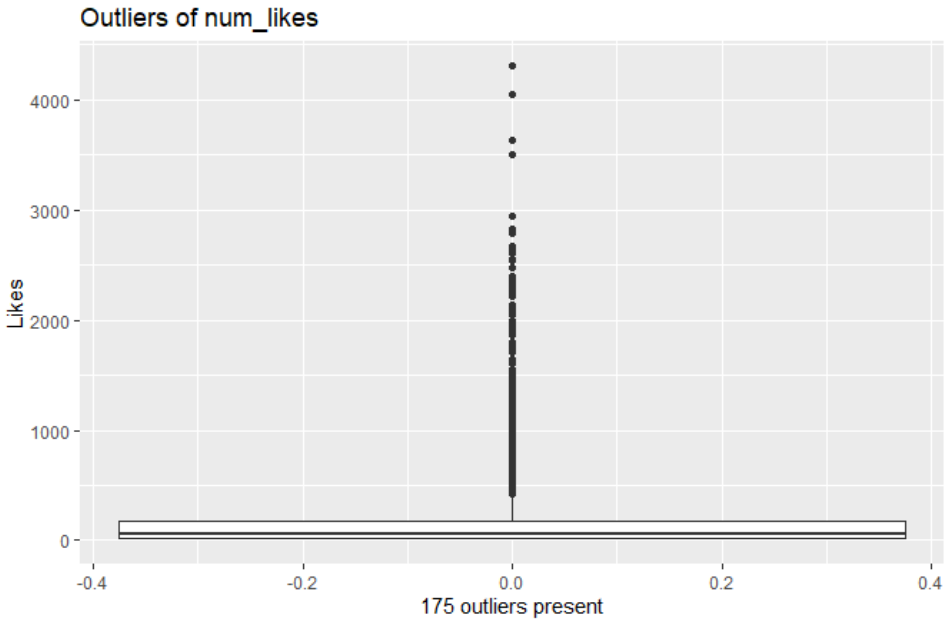
Typically speaking, when working with big data you want to remove any outliers from your data set and work with what can be considered the most "consistent" data available. But the nature of social media means that often the outliers are the desired observations from our data. It could be used as an indicator of "viral" content depending on the circumstances.

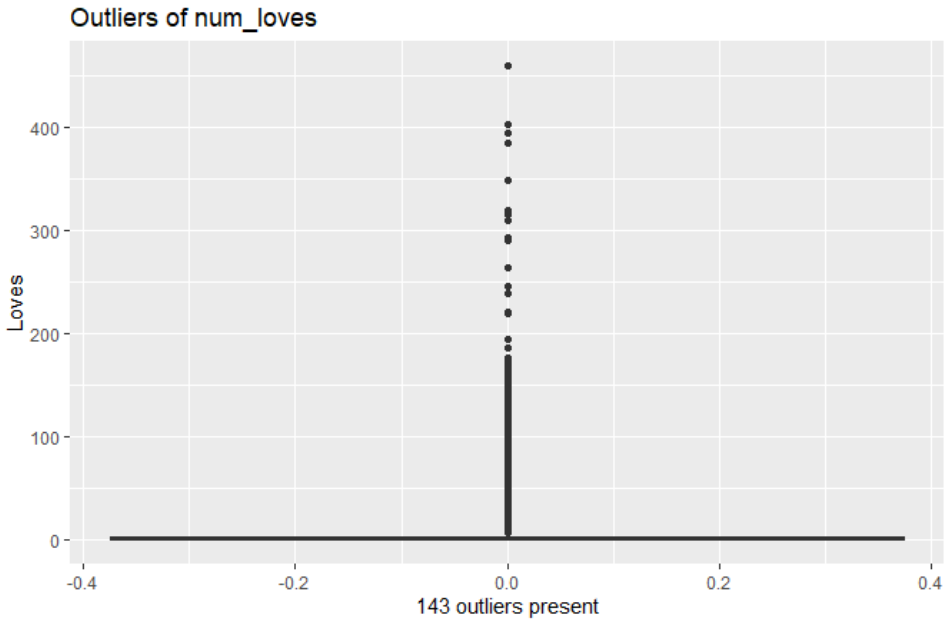
With this in mind, we will be reviewing the outliers of those variables that warrant such review but we will not be removing them for the purposes of modeling.

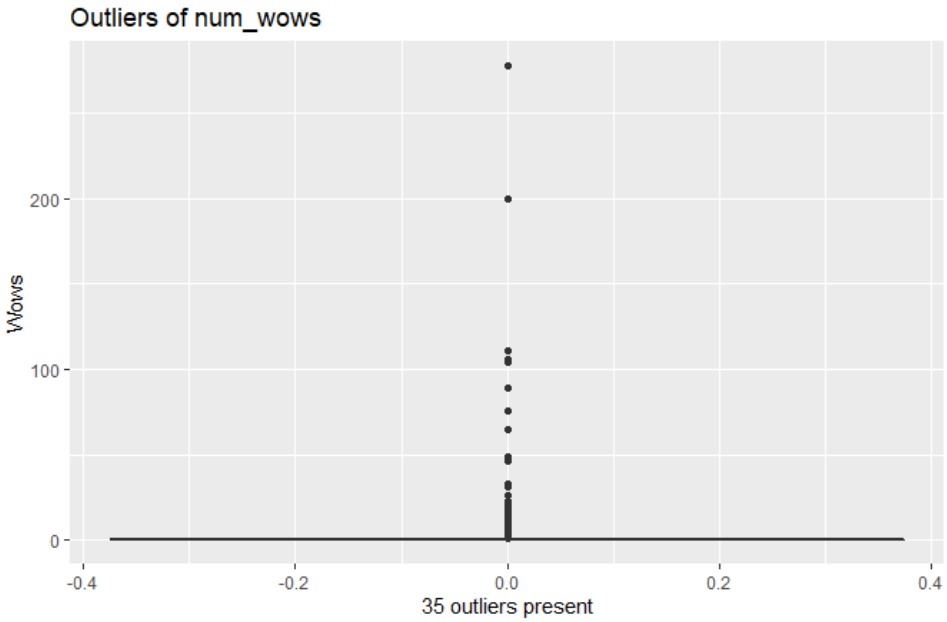


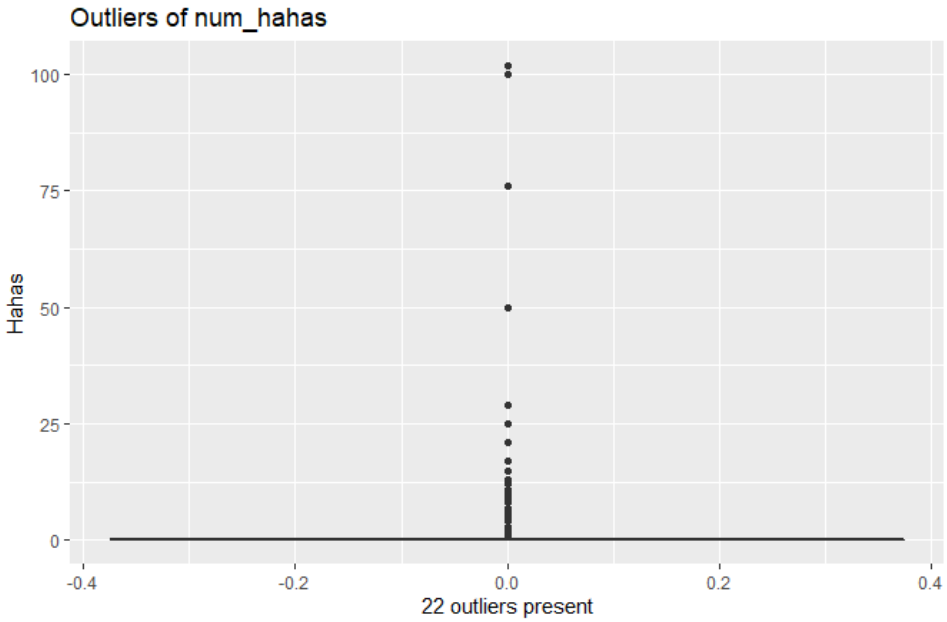

Outliers of num_comments
9000
6000
3000
0
Comments
-0.4
-0.2
0.0
0.2
0.4
301 outliers present

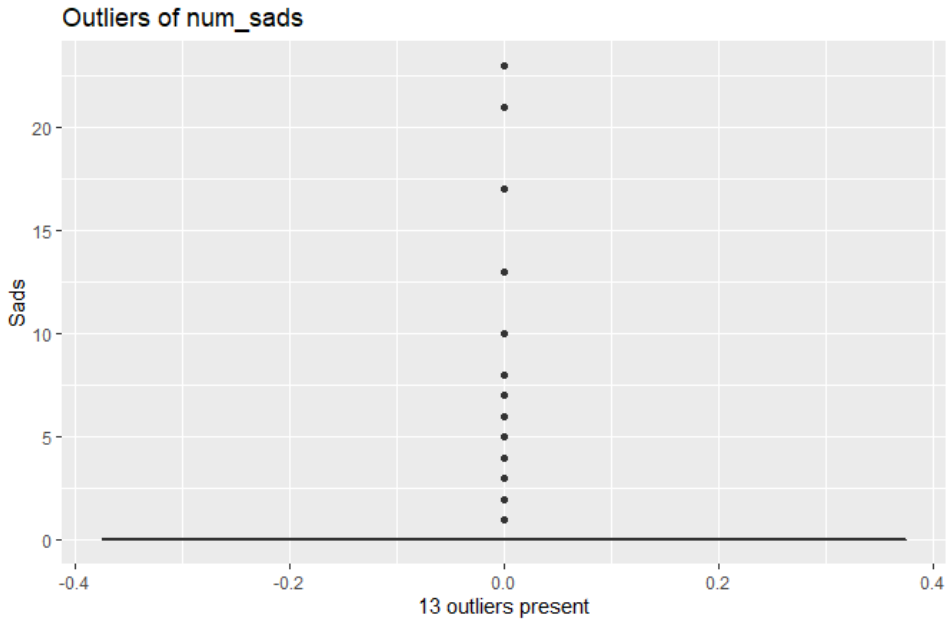

Outliers of num_shares
1000
500
0
Shares
-0.4
-0.2
0.0
0.2
0.4
245 outliers present

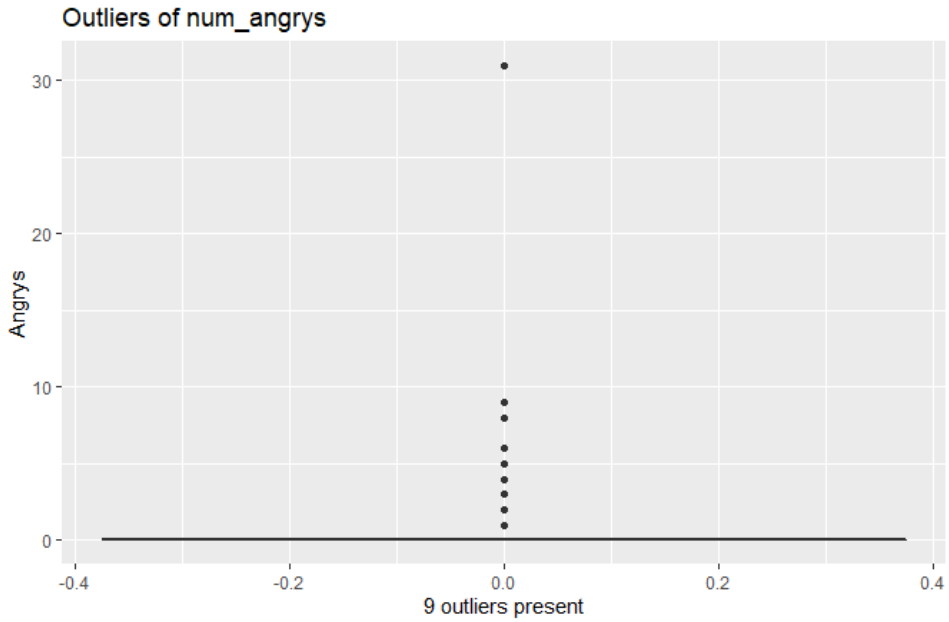

Outliers of num_likes
Likes
4000
3000
2000
1000
0
-0.4
-0.2
0.0
0.2
0.4
175 outliers present


Outliers of num_loves
Loves
400
300
200
100
0
-0.4
-0.2
0.0
0.2
0.4
143 outliers present


Outliers of num_wows
Wows
200
100
0
-0.4
-0.2
0.0
0.2
0.4
35 outliers present


Outliers of num_hahas
Hahas
100
75
50
25
0
-0.4
-0.2
0.0
0.2
0.4
22 outliers present





C. Data expansion and manipulation

As part of the preprocessing, our team decided it was important to be able to post analyze our clusters with knowledge of the date, time, month, year, and day of the week. We believe that there are time-based events (elections, holidays, feature releases, etc.) that might affect the reaction that people have to Facebook posts. As such, we decided to split the status_published field into these various additional fields to be used as we analyze our clustering results. After this transformation, our new set of columns in our data is:

```
[1]  "status_type"   "Date"          "Time"          "Year"          "Month"         "DayOfTheWeek"
[7]  "num_reactions" "num_comments"  "num_shares"    "num_likes"     "num_loves"     "num_wows"
[13] "num_hahas"     "num_sads"      "num_angrys"
```

The first few rows of the data now look like:

	status_type	Date	Time	Year	Month	DayOfTheWeek	num_reactions	num_comments	num_shares
1	photo	2018-02-21	3:05	2018	February	Wednesday	92	14	1
2	photo	2017-06-23	9:24	2017	June	Friday	1	0	0
3	photo	2016-03-08	2:43	2016	March	Tuesday	139	4	0
4	link	2015-05-22	8:21	2015	May	Friday	1931	26	0
5	video	2017-11-26	8:01	2017	November	Sunday	272	339	158
6	video	2017-11-10	4:58	2017	November	Friday	8	0	0

	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
1	89	3	0	0	0	0
2	1	0	0	0	0	0
3	138	0	1	0	0	0
4	1931	0	0	0	0	0
5	156	112	3	1	0	0
6	8	0	0	0	0	0

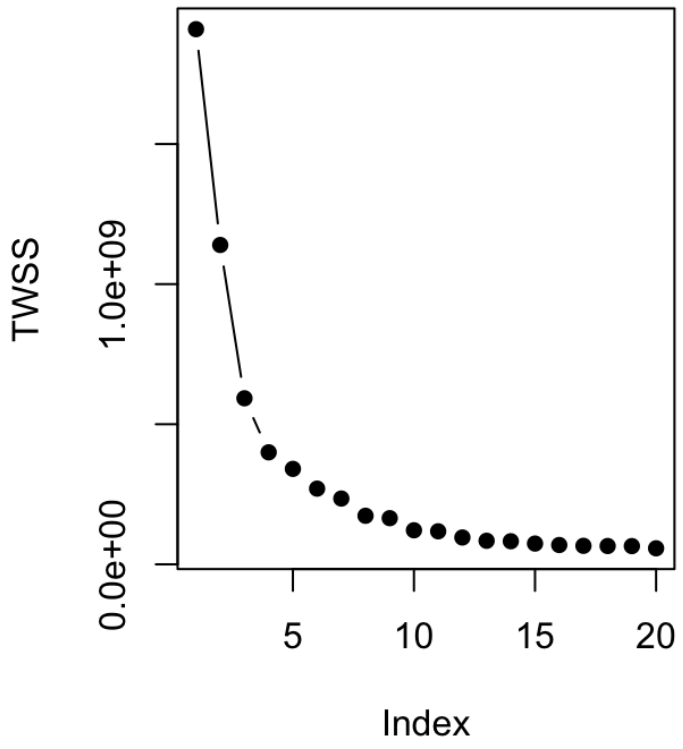
V. Applying Clustering Analysis Techniques

A. Introduction

We will perform both k-means clustering and hierarchical clustering on the dataset. To evaluate the optimal number of clusters in both algorithms, we will analyze gap statistics, silhouette statistics, a dendrogram, and a total within sum of squares plot.

B. K-Means Clustering

K-means clustering uses squared distances to group together data points based on proximity to the nearest centroid, which is the average value in a cluster. A data point merges with the cluster that has the closest centroid. This process is repeated over and over until a specified value of k clusters are created. The statistical measure for determining the optimal number of clusters that is unique to k-means clustering is the total within sum of squares (TWSS) value. This value is obtained by taking the squared values of all data points' distances to their clusters' centroid. The goal is to minimize the TWSS without having so many clusters that there is a diminishing return on insight drawn from the clustering. The TWSS plot generated from our data set is below.

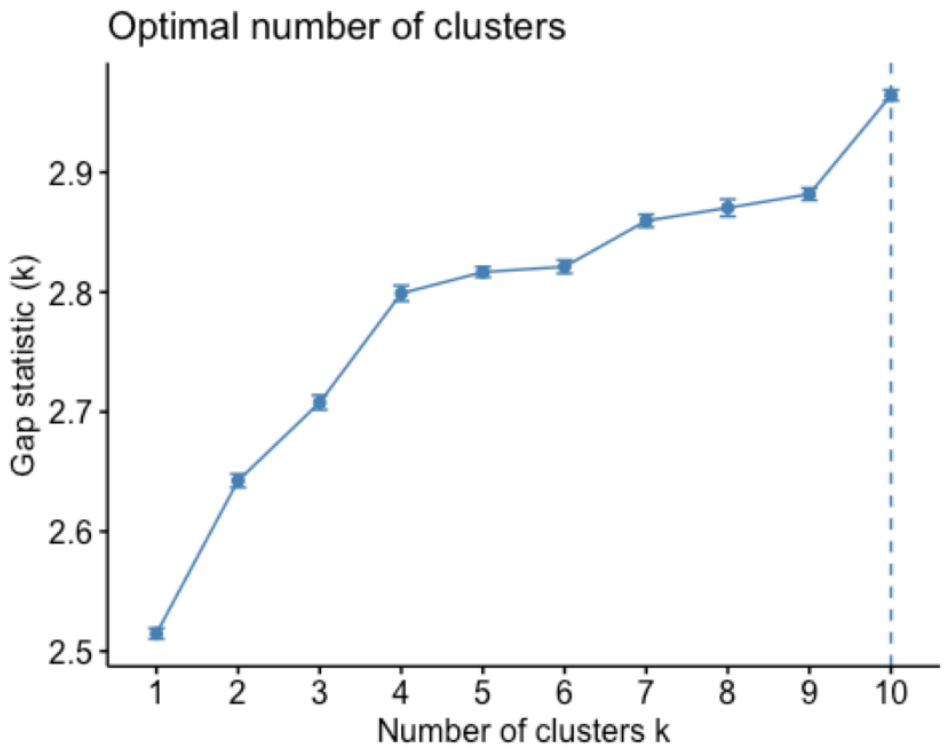


The ideal number of clusters occurs at the “elbow” of the graph, or where the decrease in TWSS begins to level off after measurable drops. Based on the TWSS plot, we have determined that the ideal value for k is 3. However, we will also want to calculate the average silhouette statistics and gap statistics.

The silhouette statistic is a measure of the proximity of an individual data point to other values in its cluster, as well its distance from observations outside the cluster. The silhouette coefficient ranges from -1 to 1. Each data point has a silhouette statistic, so the goal is to choose a number of clusters where the average of all silhouette coefficients is as close to as 1 possible. This can be executed through creating a for loop and iterating through the candidate values of k. The results of the for loop are below.

```
## [1] "# of clusters = 2 --> 0.800577276945945"
## [1] "# of clusters = 3 --> 0.808940561996679"
## [1] "# of clusters = 4 --> 0.794379140539534"
## [1] "# of clusters = 5 --> 0.660234996308003"
## [1] "# of clusters = 6 --> 0.65990628964415"
## [1] "# of clusters = 7 --> 0.670441079814997"
## [1] "# of clusters = 8 --> 0.683649776809241"
## [1] "# of clusters = 9 --> 0.6265379619871"
## [1] "# of clusters = 10 --> 0.663838103756207"
```

Based on these values, the optimal number of clusters is three, which matches our TWSS plot. Next, we will evaluate the gap statistic, which is a more sophisticated measure that requires boostrapping to resample.

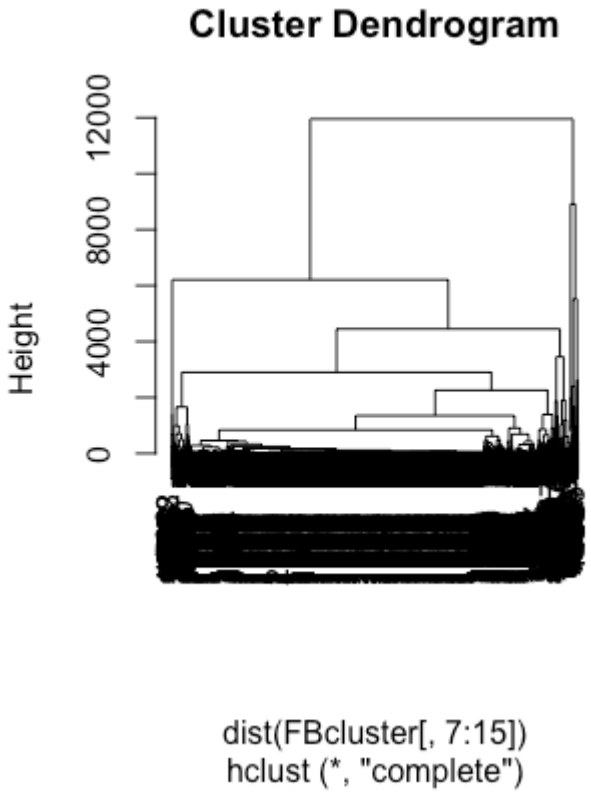


Multiple for loops were executed, and we began with 50 bootstraps and finished with 200 to ensure the accuracy of the plot did not change. Unfortunately, the optimal number of

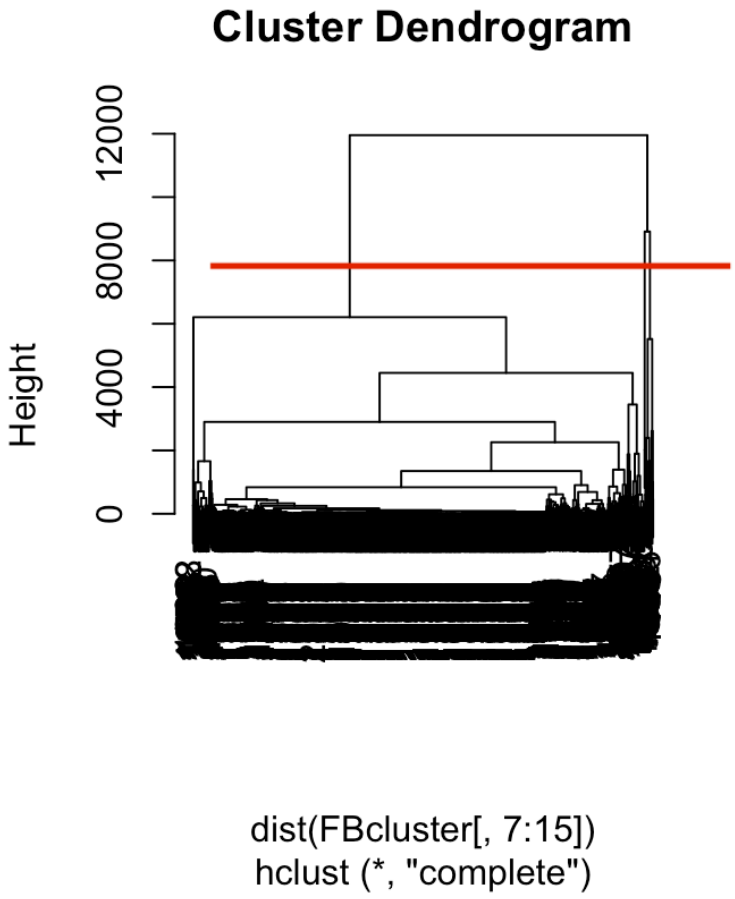
clusters within the given range occurred at the highest value of 10, so this plot is not useful for determining the optimal k-value.

C. Hierarchical Clustering

Hierarchical clustering is a cluster analysis algorithm that does not require committing to a certain number of clusters, as seen with k-means clustering. Hierarchical clustering can be either top-down or bottom-up. In bottom-up clustering, every value is treated as its own cluster, and they are slowly merged based on likeness until each data point is in one cluster. Top-down clustering involves doing the exact opposite. This algorithm calculates the Euclidian distance between data points to determines homogeneity among values through a process known as linkage. There are four types of linkage: single, complete, average, and centroid. Single linkage identifies data points in two clusters that are closest to each other to merge into a new cluster. Complete linkage, on the other hand, measures the distance between the furthest apart data points in two clusters. In average linkage, the average distances between every data point in two clusters are computed when linking. The output of hierarchical clustering is a dendrogram. The dendrogram for our dataset is below.



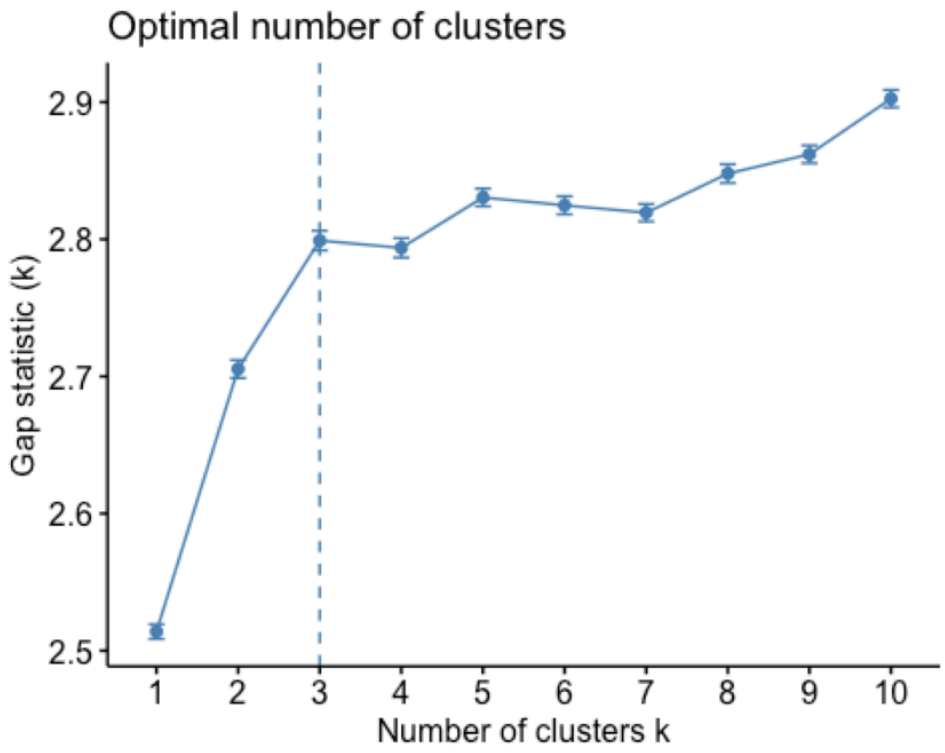
As you can see, the algorithm used complete linkage. The optimal number of clusters is observed when you draw a horizontal line through the dendrogram where the largest gap in height occurs between two distinct levels of clustering.



The red line touches three clusters, so the dendrogram suggests dividing the data out into three clusters. The average silhouette coefficients generated from 10 levels of hierarchical clustering were also observed.

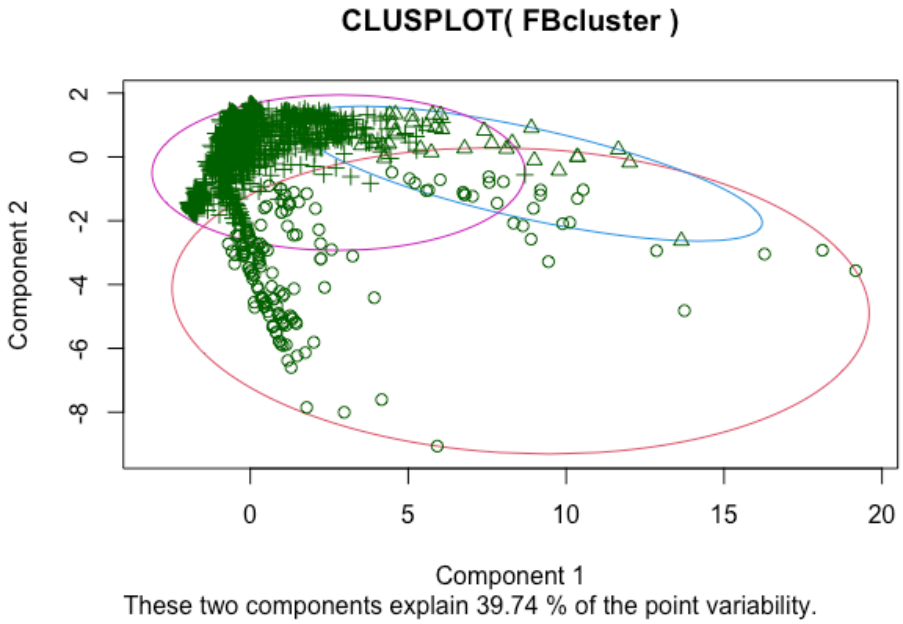
```
## [1] "# of clusters = 2 --> 0.837001678752563"
## [1] "# of clusters = 3 --> 0.824088343991942"
## [1] "# of clusters = 4 --> 0.812058463961017"
## [1] "# of clusters = 5 --> 0.778883790204313"
## [1] "# of clusters = 6 --> 0.798640843380416"
## [1] "# of clusters = 7 --> 0.795465559919775"
## [1] "# of clusters = 8 --> 0.762741744375292"
## [1] "# of clusters = 9 --> 0.762183958701071"
## [1] "# of clusters = 10 --> 0.75487960952403"
```

Based on the reported average silhouette coefficients using a hierarchical clustering algorithm, two clusters should be selected. The gap statistic plot, however, suggested that three clusters might be more appropriate.



D. Selection of One Clustering Technique

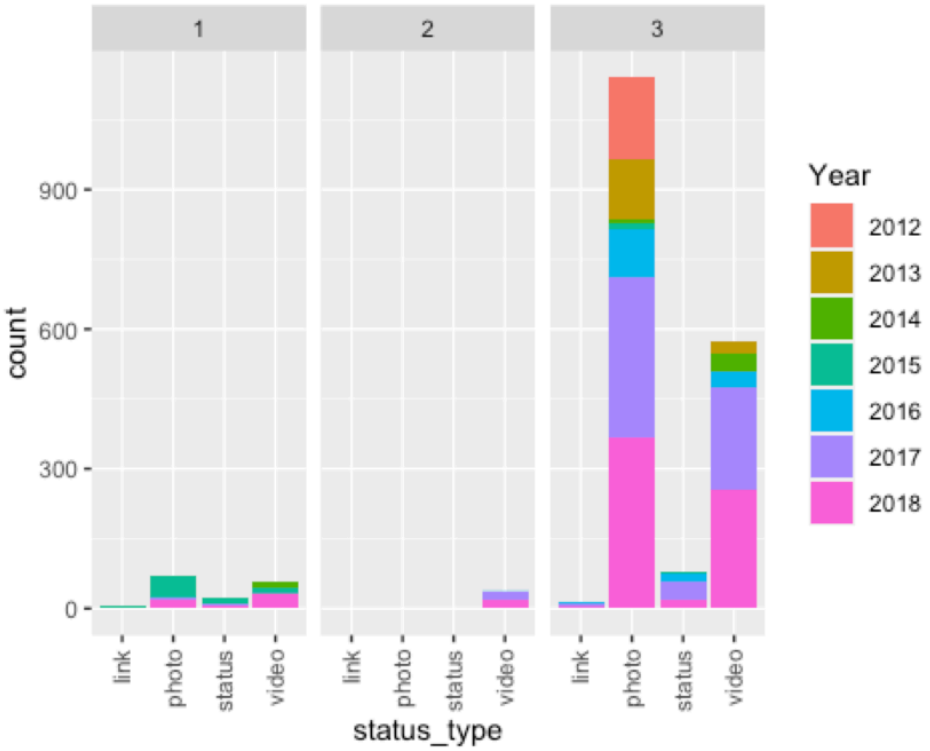
Because three was the most observed optimal number of clusters, we will settle on three clusters. When applying this number of clusters to both algorithms, between 90 to 95% of the outputs are stored in just one cluster. To ensure we have distinct clusters, we will opt for the k-means algorithm, as it has fewer results in its largest cluster. The k-means cluster plot is below.



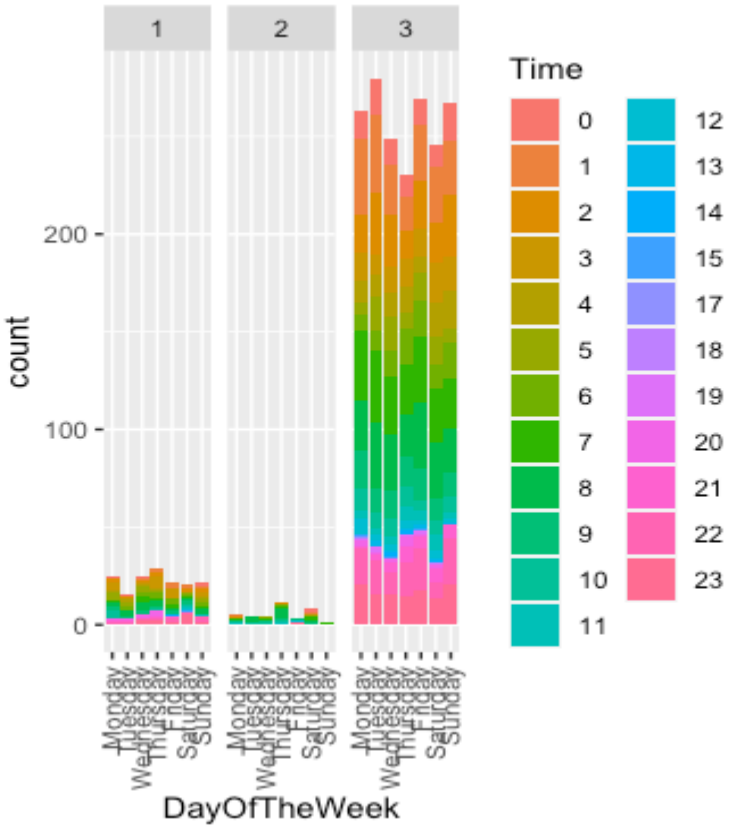
VI. Post analysis of selected clustering technique

To understand the patterns among our clusters, we will begin by plotting out the distribution of values among various categorical variables. Because these data points were collected during a time period in which Facebook Live was created, special attention will be paid to the Year and status_type variables. In addition, our team member, Anthony, is a subject matter expert on social media marketing. According to his work in the marketing industry, social media has distinct patterns in the morning, afternoon, evening, and late night. Because this data was collected before the pandemic, social media engagement was usually highest in the morning during peoples' commutes to work and at night when they were unwinding before bed. Therefore, special attention will also be paid to the Time variable, which is in military time.

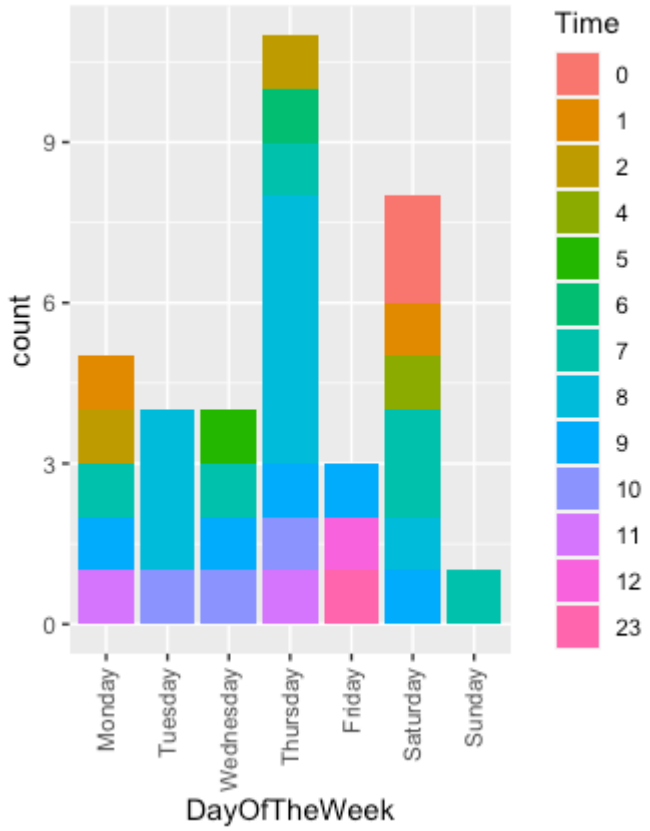
In addition, Anthony explained to our team that comments and shares are indicators of greater engagement in social media marketing, so we hypothesize that if a cluster reports higher amounts of shares and comments than other clusters, this is indicative that the cluster will be defined as the high-performing cluster. Reactions are very passive indications of customer engagement, but shares and comments boost a post's performance and reach through Facebook algorithms. Thus, we also plan to map out reactions, comments, and shares with categorical variable fills and a cluster facet wrap.



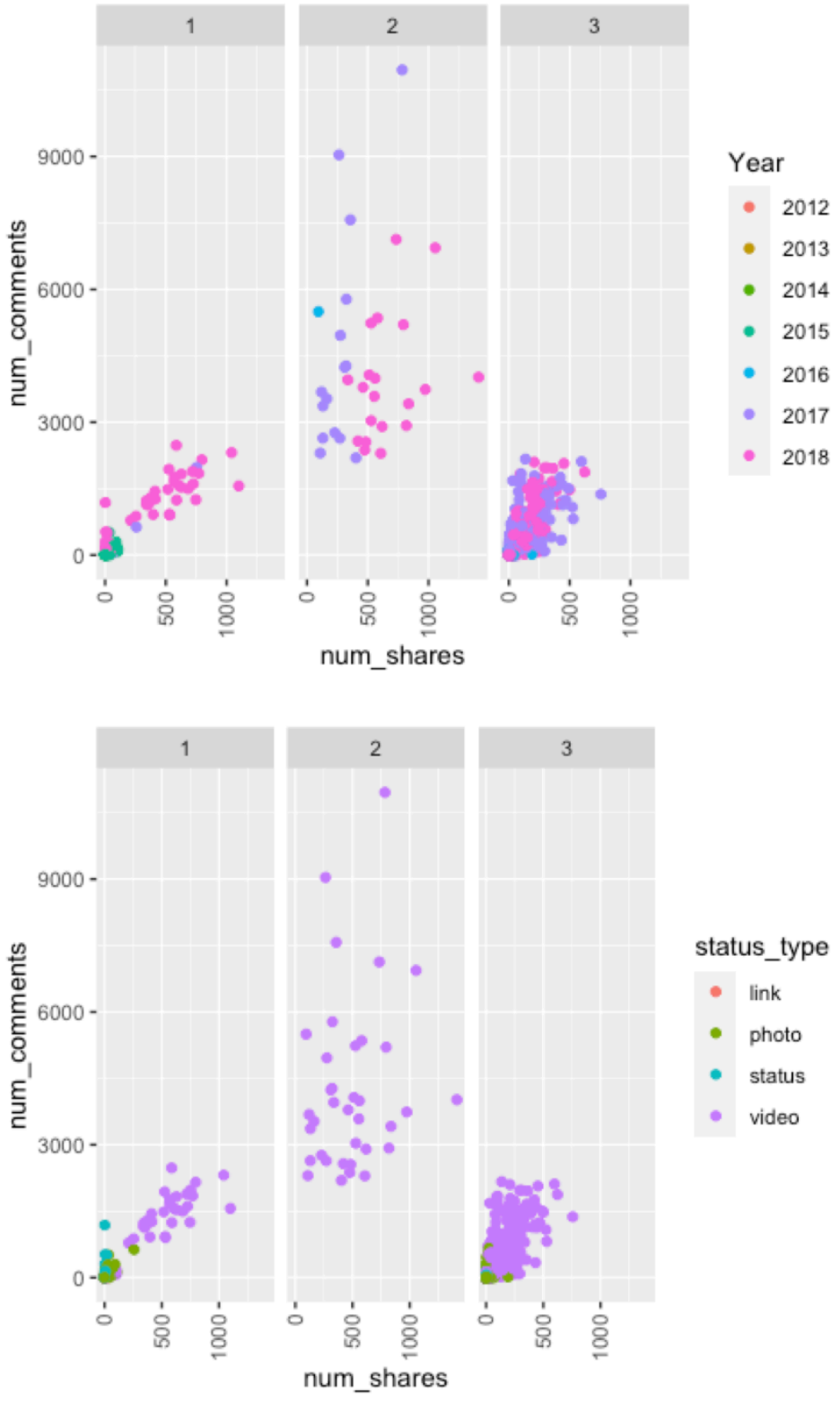
Based on the above graph, our second cluster is distinct in that it contains videos posted only between 2016 and 2018, which coincides with the launch of Facebook Live. Live video usually produces higher levels of audience engagement, so we expect Cluster 2 to be the highest performing grouping. Cluster 3 is somewhat of a kitchen sink category, as it contains posts from every year and status type, although it clearly favors photos. We will need to drill into the quantitative variables to see what makes Cluster 3 unique. Cluster 2's data points date back to 2014 and are as recent as 2018. The number of videos and photos are nearly the same, so more quantitative analysis is needed to understand what makes it distinct.



Cluster 3 has posts during all hours of the day, and Cluster 1 is similar. Let's zoom into Cluster 2.

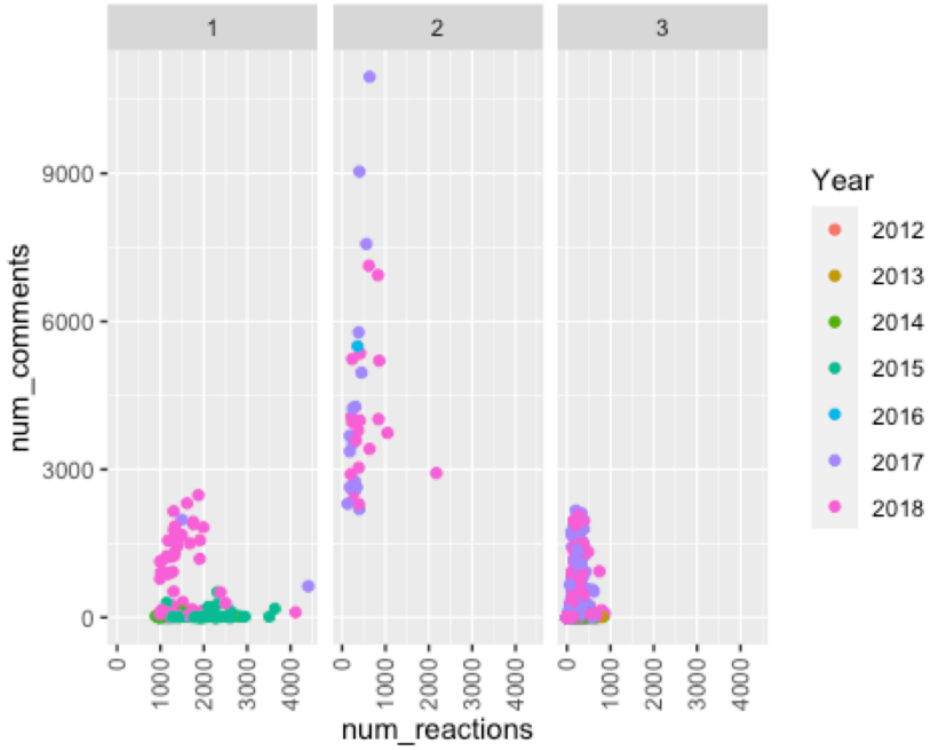


As you can see, these posts occur in the morning or late at night. It is a known fact in marketing that social media has the highest engagement between 8 a.m. and noon, and one of the best days for engagement is Thursday. All of these are evidenced in Cluster 2, which leads us to conclude that it is, indeed, the high-engagement category. In addition, we believe Cluster 3 will be the low-performing cohort and Cluster 1 will be a medium-performing group. We were able to confirm this by mapping engagement measures with a cluster wrap.

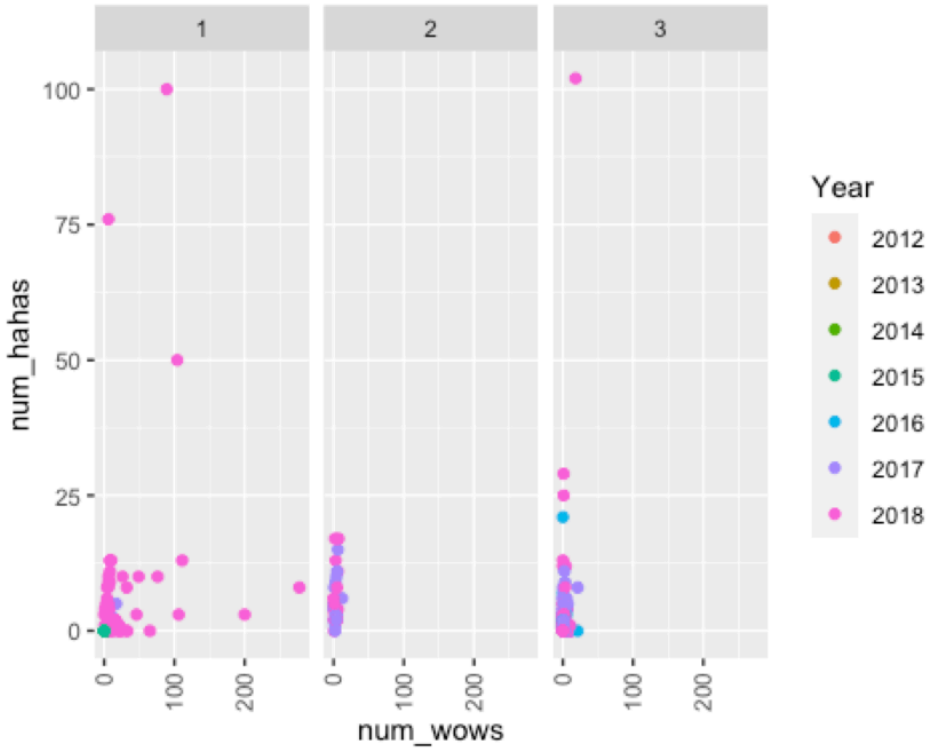


Cluster 2 has by far the greatest engagement, especially with comments. When fashion and beauty brands are hosting live sessions with product demonstrations, makeup

tutorials, and/or fashion shows, users often feel the need to be “heard.” Some brands even monitor comment sections to respond to customer questions, allowing for more intimate and personalized communication with potential customers. Overall, there is a clear association between the creation of Facebook Live and interactive video content with customer engagement.



When we map comments compared to reactions (likes, loves, etc.), color code data points based on year, and use a cluster facet wrap, we can see Cluster 3 continues to be much less engaged on both axes compared to the others. This graph does provide interesting insight, however, into Cluster 1. Cluster 1 is distinguished by higher levels of passive engagement through simple reactions (especially for posts before 2016), but its data points that occur after Facebook Live’s creation had greater levels of active customers engaging directly with brands. However, this interaction was not as strong as Cluster 2. This is further illustrated when we map out num_hahas and num_wows.



Not only does Cluster 1 again show the highest levels of passive engagement, but Cluster 2 also has the lowest counts of these reactions. It is possible the video content in Cluster 1 is not Facebook Live content, but rather a short, branded video. If we were able to drill into the types of video content, we would be able to understand more clearly these observed differences between video content performance in clusters 1 and 2.

Our k-means cluster analysis confirmed our suspicions based on Anthony's domain expertise in marketing surrounding trends in user behavior on social media sites. The highest levels of customer engagement occurred after the introduction of Facebook Live through the growth of interactive video content, as well as during peak times of the day and days of the week.