Portfolio-Exam Part I «MADS-MMS»

Tom Wüsten

December 9, 2021

1 Data Acquisition and Initial Data Analysis – 10 points

Obtain the dataset [1] from the UCI Machine Learning Repository.
Conduct a brief initial analysis of the raw dataset (henceforth called Dataset A).

1.1 (2 points) What do the rows of the dataset represent?

Every row represent a facebook post. This post can be a video, photo, link or a status. Every post got various attributes like status_published , num_reactions or num_likes.

1.2 (2 points) How many different instances does the dataset contain?

The dataset contains 7050 instances.

1.3 (2 points) How many attributes (columns) are in the dataset?

The dataset contains 12 attributes.

1.4 (4 points) What is the standard deviation of the feature num_likes?

For the calculation I used the sample standard deviation from the libaray pandas[2] as following:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

The result of the sample standard deviation for the feature num_likes is 449.472.

2 k-Means Clustering on the Plain Data – 40 points

Begin the analysis using the k-means approach.

2.1 (5 points) Which features of the dataset do not suggest themselves as features analysis? For each of these features, briefly state why you exclude them.

The _id has no value because it's an id and every instances has a different id. The status_id is a categorical feature that describes the kind of the post. The k-means approach doesn't allow categorical features. It would be possible to transform this feature via one hot encoding. Then we would have 3 additional features. But in this categorical format we can't use the status_type. The status_published describes the time and day when the post was posted. This time format is not usable for the k-means approach. It would be possible to change the type of the feature via feature engineering approaches.
In summary the first three columns status_id, status_type and status_published aren't interesting for the k-means clustering.

2.2 (15 points) For the next tasks, restrict Dataset A to the following features: num_reactions, num_comments, num_share, num_likes, num_loves, num_wows, num_hahas.
We will call this Dataset B On B compute k-means clusterings of the dataset using different choices for k: 2, 3, . . . , 10.
Use a seed of 1 to make the experiments reproducible. For each k compute the silhouette coefficient and plot it against k in a diagram. Interpret the diagram!

The calclutated silhouette coefficient for different choices of k: 2, 3, . . . , 10 looks like that:

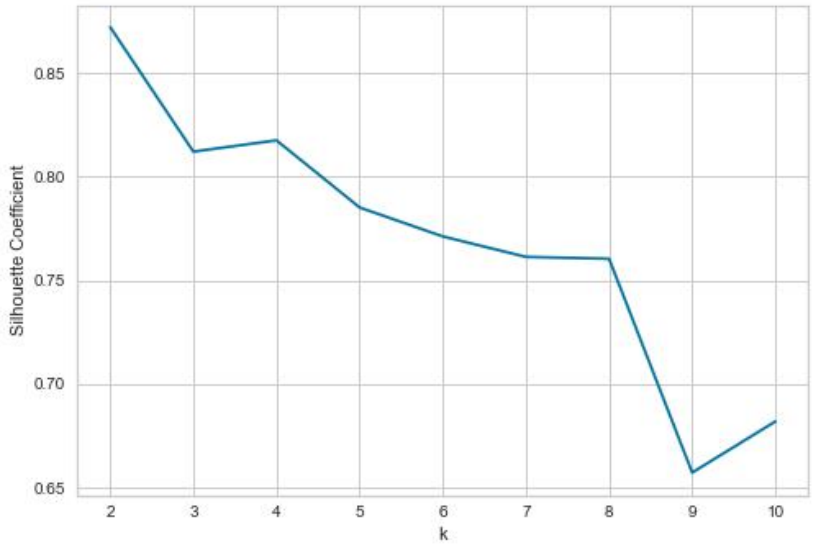

Figure 1: silhouette coefficient for different k

The figure 1 shows that the silhouette coefficient decrease if the cluster size k increase. The best silhouette coefficient is at k=2. In general is the structure of cluster for small k good. In the lecture we defined that a silhouette coefficent sc>0.7 has a good structure. That means that until k=8 the structure of clusters are good.

2.3 (10 points) Create a silhouette plot for the k with the highest silhouette coefficient in the previous experiment. Interpret the diagram!

The figure 2 shows the silhouette plot of KMeans clustering for 7050 samples in 2 centers. In this figure you can see that the most data points are in cluster 0. The cluster 0 has 6909 data points and the cluster 1 has 141 data points detected. That means that 98% of all data points are in the cluster 0. The average slihouette score of the cluster is 0.8721. This score is really high but it's not good because the clustering detected 98% of the data points in one cluster. The value of this clustering is low because almost all data points are on one cluster. Also the silhouette score of the cluster 1 is 0.6 that is relatively low.

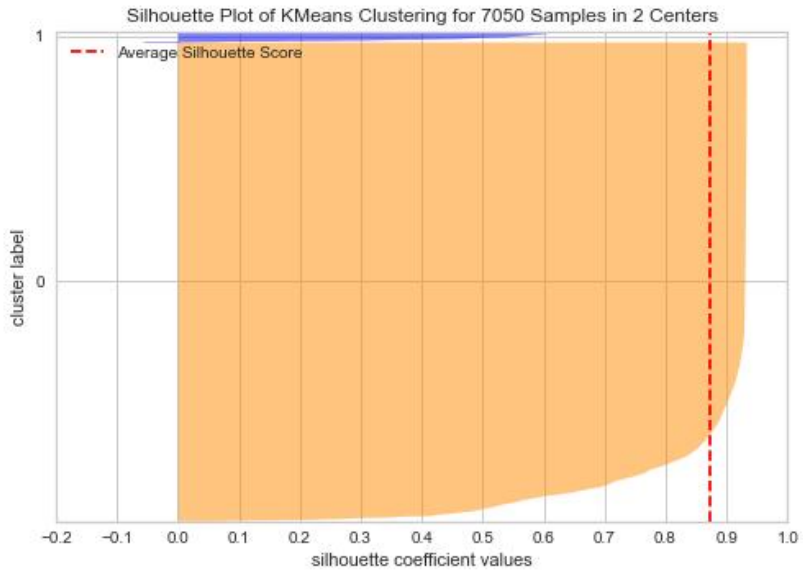

Figure 2: Silhouette Plot of KMeans Clustering for 7050 Samples in 2 centers

2.4 (10 points) For the same k, create a plot of the data where you use only the two features num_reactions and num_likes as the axes. Use color to distinguish instances from different clusters. Also highlight the cluster centroids of the k-means clustering. Interpret the diagram, considering only the above two features. Is there a clear clustering structure visible?

The figure 3 shows a KMeans clustering with k=2 and the features num_reactions and num_likes as the axes. The orange points are in the cluster 0 and the blue points are in the cluster 1. The centroids are marked with a red star. It's not a good clustering structure visible because the most data points lie on top of each other. This both feature have correlation and there aren't good to show a clear structure.

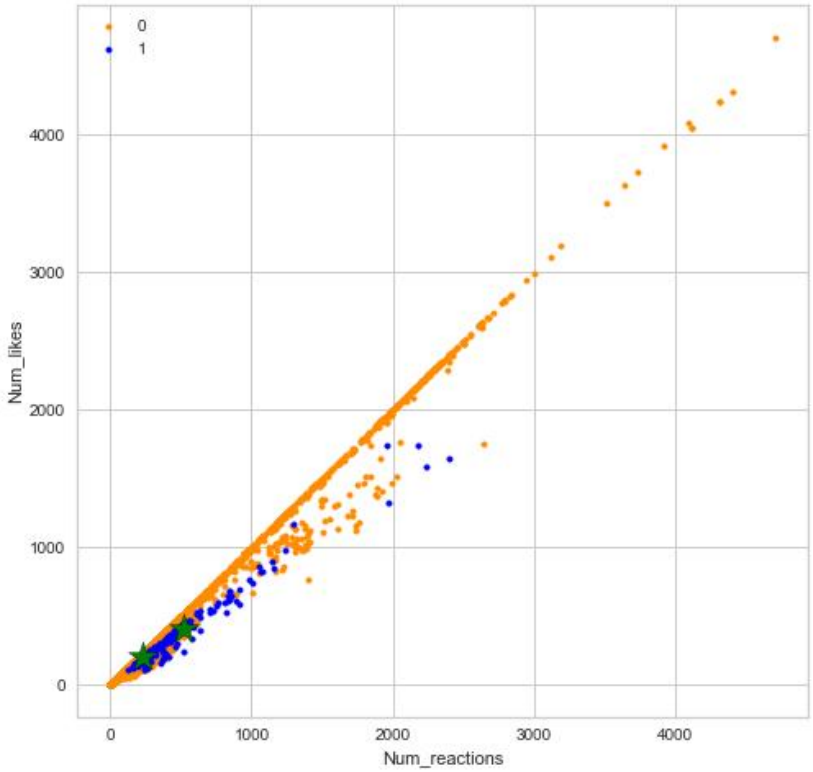

Figure 3: KMeans Clustering for 2 clusters

3 Scaling and Feature Selection – 50 points

In these next experiments, we preprocess and restrict Dataset B further through scaling and using variance as a criterion for feature selection. Particularly, use the class sklearn.feature_selection.VarianceThreshold.

3.1 (10 points) Describe in your own words, what the class VarianceThreshold is used for and explain why looking at a feature's variance is meaningful.

The class VarianceThreshold drops column where the Variance is under a specific chosen threshold. Variance shows the variability in a distribution. In context of features it shows us how much spread has this feature. If it has a variance of 0 then the information is not meaningful. But if the variance is high then those feature provides more information. Variance Threshold contains to the field of feature selection. Feature selection describes the process of choosing the most important features while trying to retain as much information as possible.

3.2 (5 points) The features of the data set are in different ranges. To be able to compare by variance we should scale the data first. Which is the better choice for the variance threshold method Min-Max-Scaling or the Standard-Scaler (z-score transformation)? Explain your answer.

StandardScaler performs the task of Standardization. This is necessary when the data set is used in different scales.

3.3 (10 points) Use Min-Max-Scaling on Dataset B to yield Dataset C and rerun the above experiments on C(k-means clusterings for k "2, 3, . . . , 10, silhouette plot for the best k, plot of clustered data and centroids). How does that compare to the previous experiments?.

After the Min-Max-Scaling the highest silhouette coefficient is found for k = 3. In the figure 4 we can see that best k are for 2 and 3. In comparison to figure 1, the silhouette coefficient does not decrease rapidly with increasing K, but drops sharply at k=4. Furthermore, the silhouette coefficient increases again at K=5, and then drops to the lowest silhouette coefficient. From k=6, a slight increase can be seen. It can also be seen that the average silhouette score is lower than in figure 1.
The figure 5 now shows the silhouette plot for KMeans clustering for 7050 samples with 3 centers. Compared to Figure 2, the number of k has increased. By applying the min-max scaler, another cluster could now be found. However, it is still true that most of the data points were assigned to cluster 0. Cluster 0 has 92.2%, cluster 1 has 5.3% and cluster 2 has 2.5% of the data points. Like in the first experiment the most data points are in cluster 0. The average silhouette score also decreased compared to the experiment before. This is due to the fact that cluster 2 also has negative silhouette scores.
Figure 6 shows a KMeans clustering for k = 3 with the features num_reactions and num_likes as axes. The centroid for cluster 0 is again at the same position but the centroid of cluster 1 moved to 0.4 (x and y). But by using the Min-Max-Scaler a much better cluster structure could be created. Cluster 0 is as in experiment 1 in the range 0 to 0.2 (normalized) or 1000 (for x and y). But cluster 1 is now starting from 0.2 and covers all data points that are linear and greater then 0.2. Thus, a clear cluster structure is already visible. Furthermore, cluster 2 also covers data points that are not completely linear. But there are in the range from 0 to 0.2 (x and y) many superimposed points, which show a worse cluster structure. That's why I additionally plotted for k=2 in figure 7, since the average silhouette score was a bit lower. This plot shows a better cluster structure, because almost all points larger than 0.2 (x and y) can be assigned to a cluster. In summary, the min-max scaler resulted a lower average silhouette score compared to experiment 1, but a much better cluster structure.

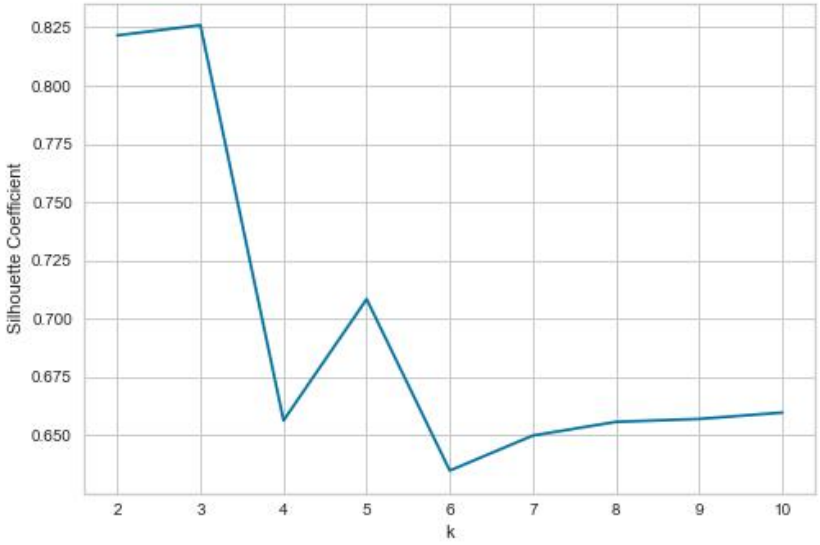

Figure 4: silhouette coefficient for different k

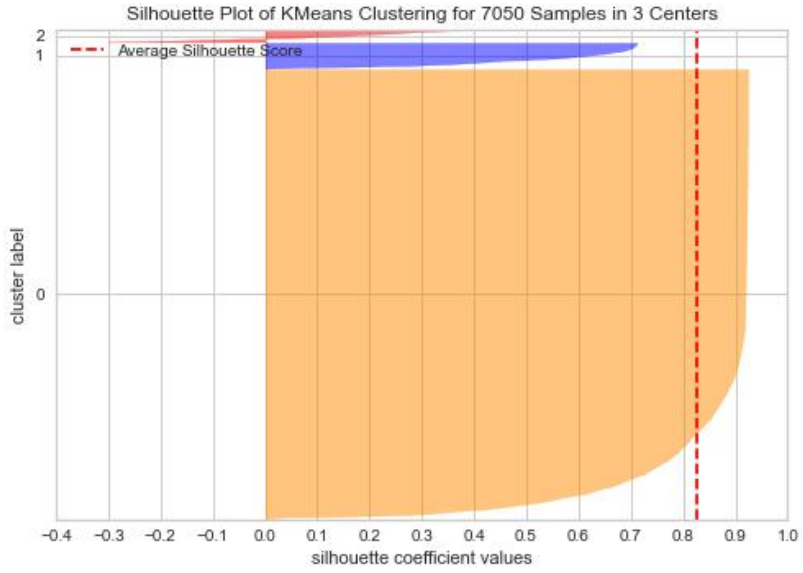

Figure 5: Silhouette Plot of KMeans Clustering for 7050 Samples in 3 centers

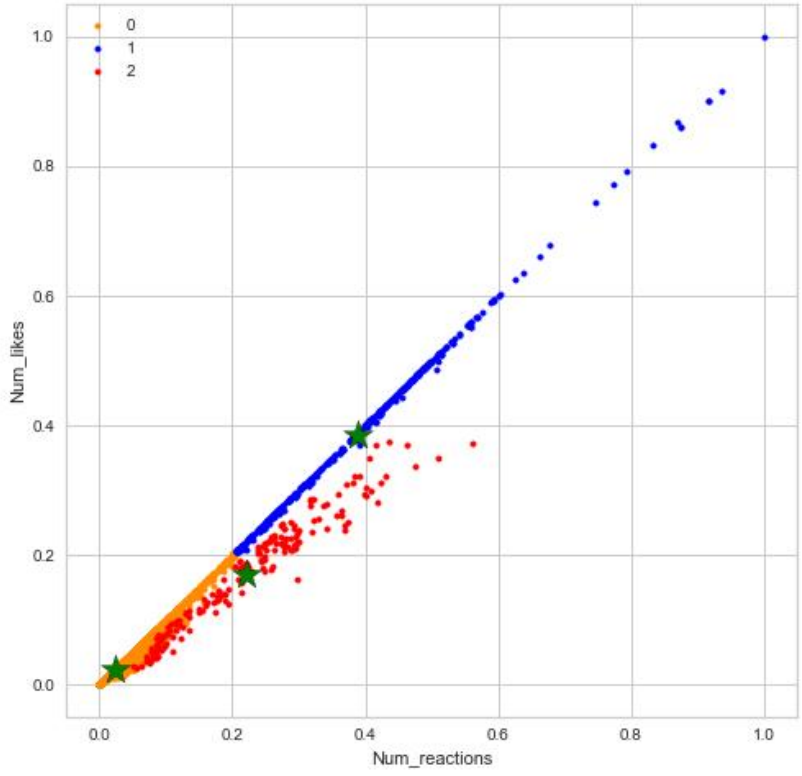

Figure 6: KMeans Clustering for 3 clusters

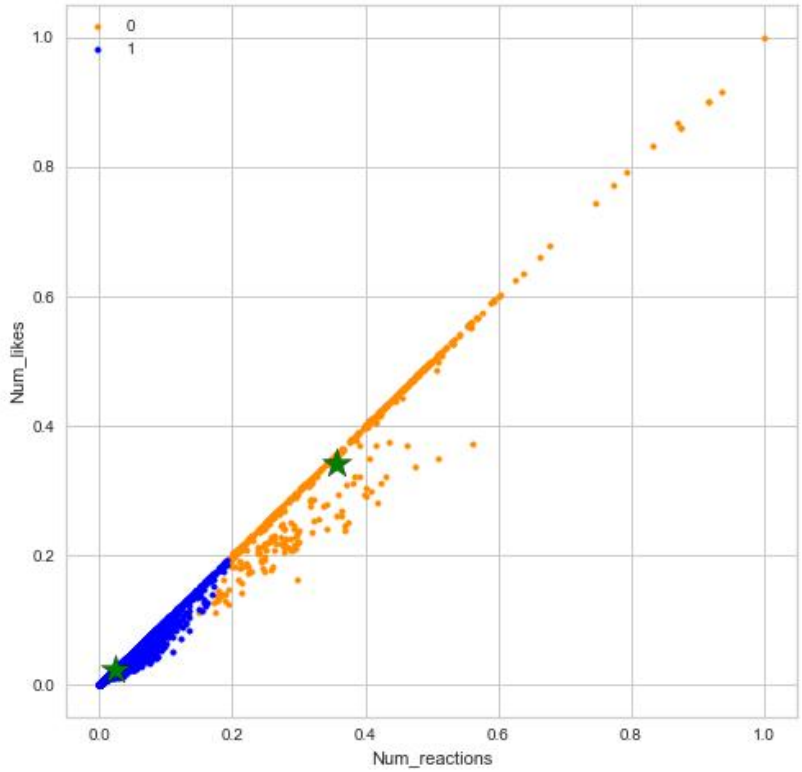

Figure 7: KMeans Clustering for 2 clusters

3.4 (10 points) Use the variance threshold method for feature selection on the scaled data to restrict the features further. Set the variance threshold to 0.005. Which features does the resulting dataset (Dataset d) include, what is their variance (computed on the scaled features)?

After the calculation of the variance threshold only the columns num_likes and num_reactions had a higher variance then the threshold. The variance of num_likes is 0.0091 and of num_reactions is 0.0096. The dataset d contains now only the features num_likes and num_reactions.

3.5 (5 points) Repeat the experiments from above on Dataset D (k-means clusterings for k " 2, 3, . . . , 10, silhouette plot for the best k, plot of clustered data and centroids). Compare the outcome to the previous experiments.

The figure 8 is the data set d used to show the silhouette coefficients for different k. It can be seen again as in the figure 1 that as k increases, the silhouette coefficient decreases. For k=2 the highest silhouette coefficient has been reached. Compared to figure 4, a steady decrease of the Silhouette coefficient with increasing k can be seen. The course is similar to the diagram from figure 1.
In the Figure 9 shows a silhouette plot for a KMeans clustering for 7050 samples of 2 centers. Figure 9 is very similar to figure 1, the only difference being that Cluster 1 has more data points assigned to it.The silhouette average score of this experiment is 0.874. In comparison, experiment 1 had a silhouette average score of 0.872 and experiment 2 a score of 0.826. The last experiment leads to the best result in terms of silhouette averages score due to the variance threshold.

Figure 10 shows the clustered data with the two centroids. This figure shows the best cluster structure of all, because a clear boundary is evident. There are also no superimposed points as in the previous experiments. Furthermore, as in the previous experiments, most of the data points are in cluster 0. In this experiment, 93% of the data are in cluster 0 and 7% are in cluster 1. From the figure, it can be seen that the use of the variance threshold has found out the features that have the highest significance. Furthermore, it can be seen that the other features have distorted the significance in the previous experiments.

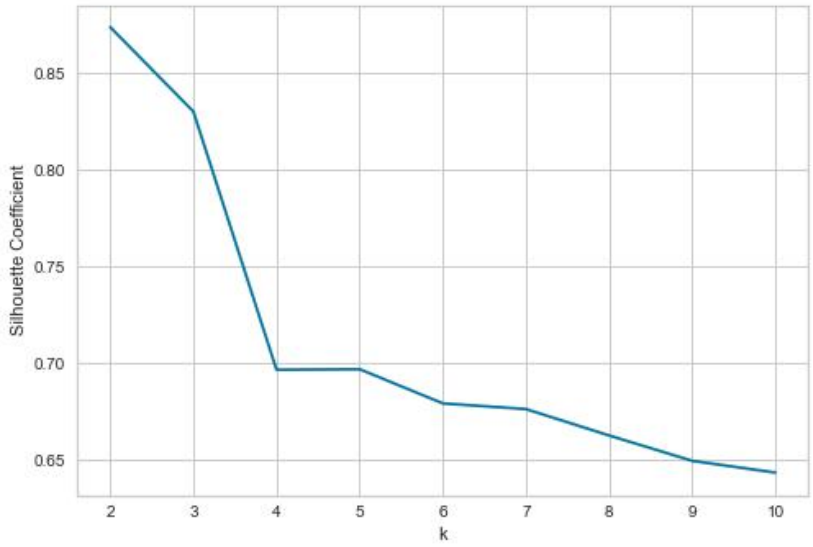

Figure 8: silhouette coefficient for different k

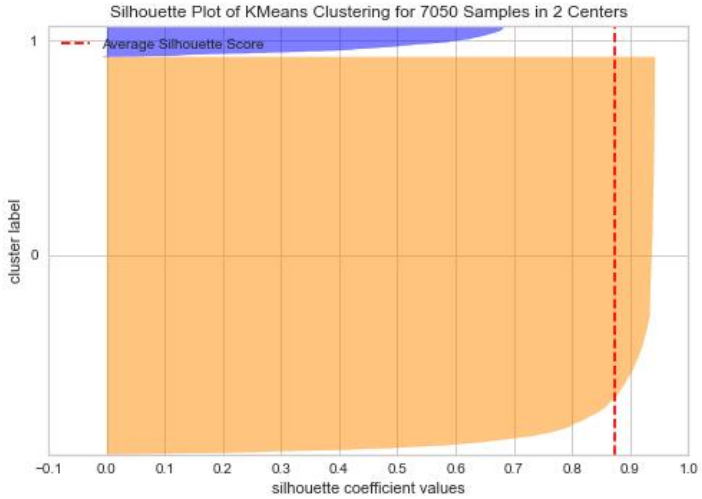

Figure 9: Silhouette Plot of KMeans Clustering for 7050 Samples in 2 centers

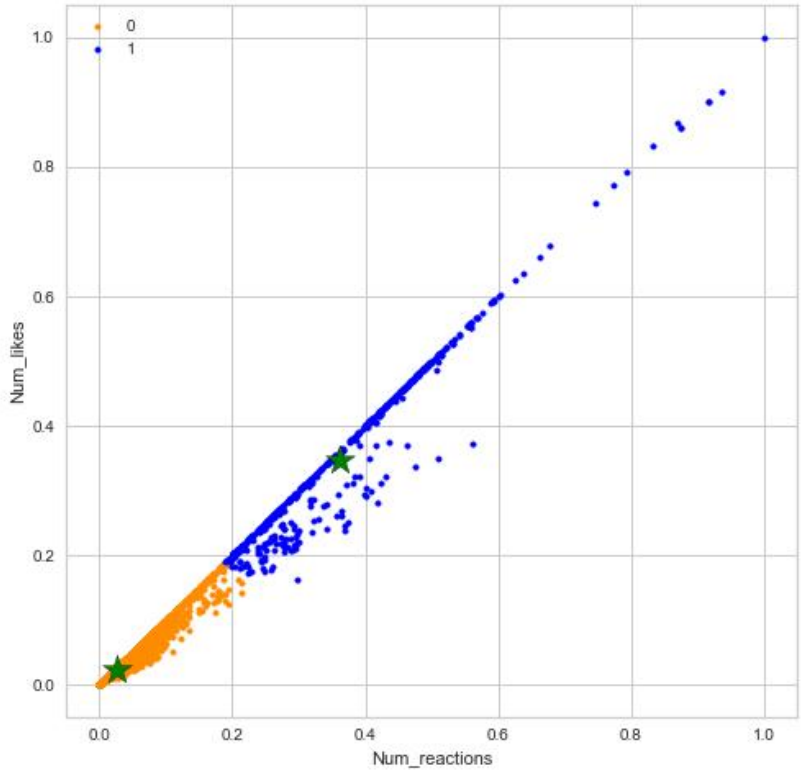

Figure 10: KMeans Clustering for 2 clusters

3.6 (10 points) Compare the feature distributions over the clusters for features num_reactions, num_likes, and num_shares. Use violin plots for the original data (the unscaled values) and interpret your results.

Figure 11 shows the violin plot for 2 clusters and the feature num_reactions. The figure shows that 75% of the data points from cluster 0 are between 0 and 174. For cluster 1, 75% of the values are in the range of 893 to 2084.5. This clearly shows that a good division of clusters has been made.
Figure 12 shows the violin plot for 2 clusters and the feature num_likes. From the figure it can be seen that 75% of the data points from cluster 0 are between 0 and 147. For cluster 1, 75% of the values are in the range of 765 to 2064.3. This clearly shows that a good division of clusters has been made.
Figure 13 shows the violin plot for 2 clusters and the feature num_share. For cluster 0, 75% of the data points are in the range between 0 and 3. In cluster 1, 75% of the data points are between 0 and 49.25. For both clusters, there are also outliers that are greater than 1500. Both clusters have very many points that are in the same range. From this data and the violin plot, no information can be obtained from this feature.
In summary, we see that the variance threshold has filtered out the feature num_likes and num_reactions as meaningful. For both features it is clear from the violin plot that clustering into two clusters is useful. A clear clustering structure can be recognized. In comparison, no clear clustering structure can be recognized from the feature num_share.

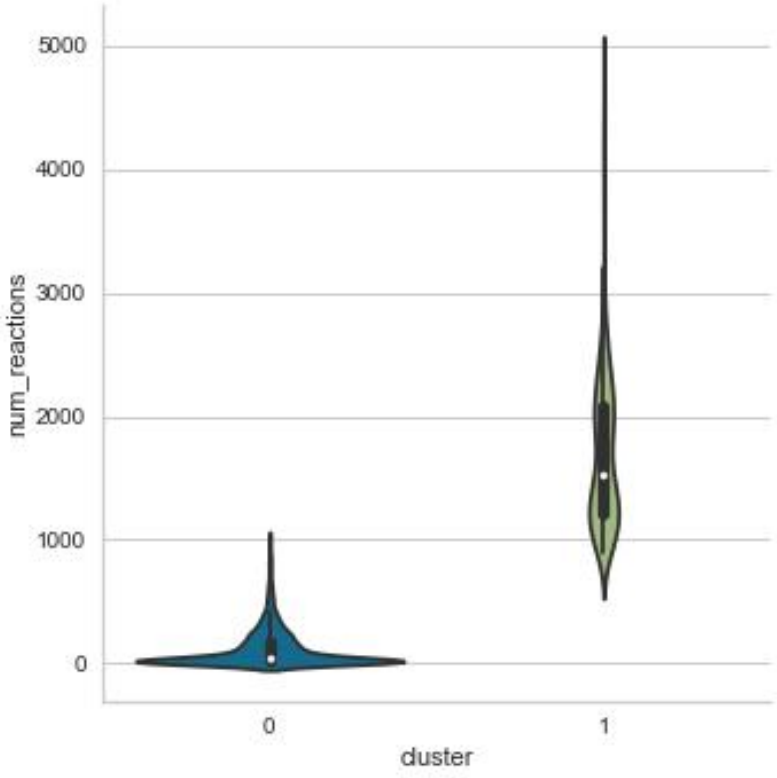

Figure 11: num_reactions distribution for 2 cluster

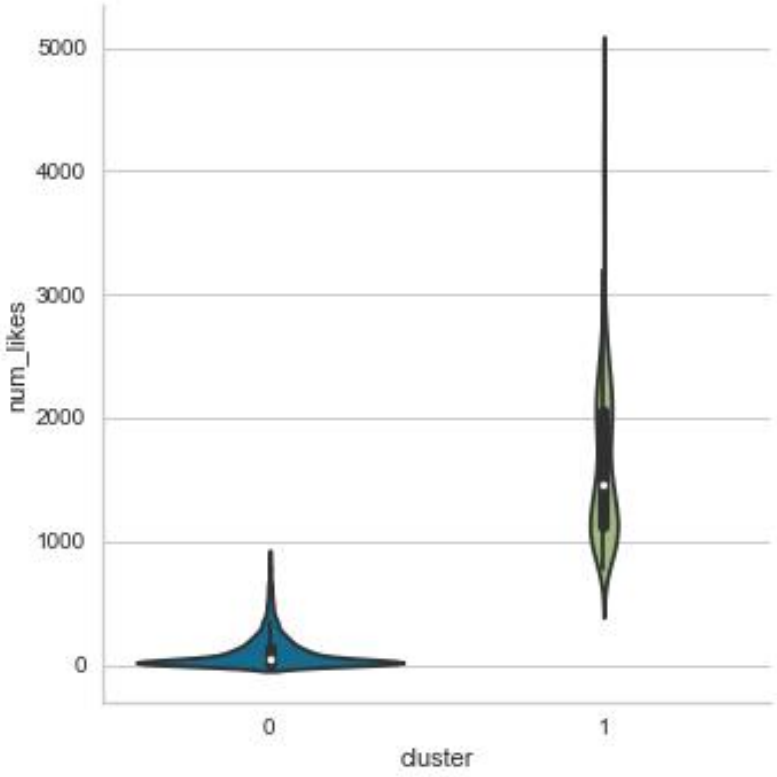

Figure 12: num_likes distribution for 2 cluster

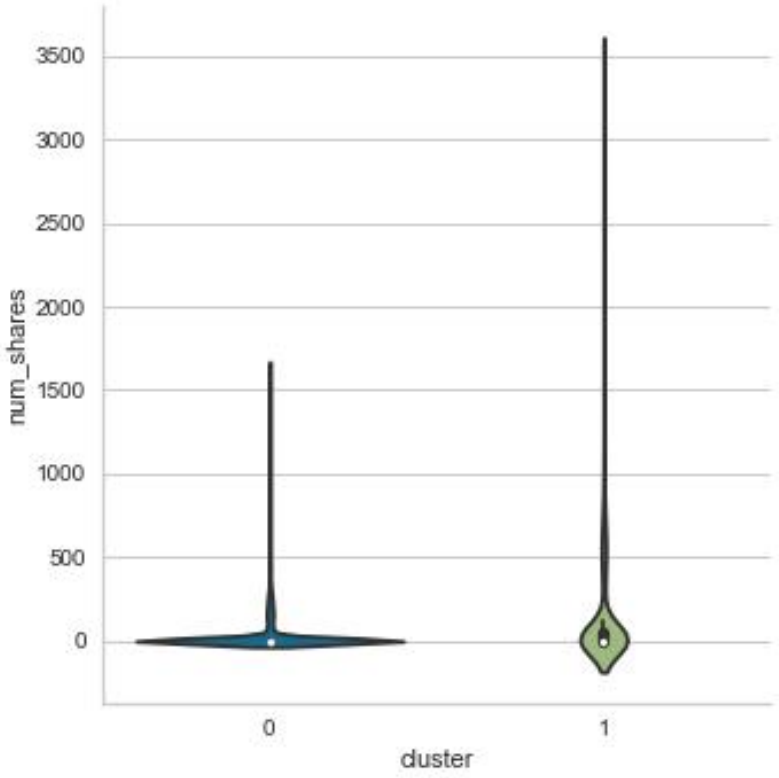

Figure 13: num_shares distribution for 2 cluster

References

[1] N. Dehouche and A. Wongkitrungrueng, "Facebook live as a direct selling channel, 2018," in Pro- ceedings of ANZMAC 2018: The 20th Conference of the Australian and New Zealand Marketing Academy. Adelaide (Australia), pp. 3–5 December, 2018. Available at https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand