REPORT

DMA18 — Assignment 2, Michael Froehlich

1. RESEARCH QUESTION

Describe what it is you want to ask of this dataset.

It would be interesting to know whether the number of reviews, the average text length and the votes received for reviews cluster into groups. From *Yelp's* point of view identifying reviewer's in specific clusters could be of particular interest.

Cluster with the following features could be of interest. Note that the list presented is not exhaustive and a clustering based on these features may be interesting for other use case as well.

# reviews	**# text length**	**# votes**	**potential use case**
HIGH	-	HIGH	Reviewers from this cluster are particularly valuable for Yelp. Therefore, keeping them on the platform should be a priority.
LOW	LOW	HIGH	Reviewers from this cluster receive high ratings, but do not write many reviews. Yelp could incentivize them to write more reviews and improve their text length.
HIGH	HIGH	LOW	Reviewers from this cluster are highly active, but do not write good reviews. In the age of 140-character messages on Twitter, one can hypothesis whether shorter reviews would lead to better ratings.
HIGH	-	LOW	Reviewers in this cluster are active, but do not produce good reviews. Targeted advertisement and help on how to write good reviews could convert them to valuable members on the platform.
LOW	-	LOW	Reviewers in this cluster are neither active contributors not do they produce helpful review. Hence, Yelp should not use too many resources on them.

2. DATASET

Describe the *yelp_reviews.csv* to *yelp_reviewers.csv* transformation.

All features required for the creation and analysis of the suggested clustering are already present in the provided *yelp_reviewers.csv.* Therefore, the following section provides only a theoretical description of how the selected features would be transformed into from the original *yelp_reviews.csv.* from assignment 01 to the *yelp_reviewers.csv.*

2.1 FEATURES SELECTED

Briefly describe the features from the dataset you choose (at least 1 features should be from a q16)

Each instance in the dataset is specific to one reviewer and has the following features.

Feature	Description
q3	Number of reviews.
q15	Average review text length (in non-space characters).
q16j	Average number of feedback votes.

q3 — Number of Reviews

Getting the average number of reviews per reviewer from the original dataset can be achieved by grouping the dataset by `user_id` and counting the size of the rows per user.
This transformation corresponds to question **3A** of assignment 01.

q15 — Average Review Text Length

Calculating the average text length per reviewer without spaces is a multi-step process. Starting from the original dataset the following transformations are needed.

1. Remove the spaces from all reviews texts by replacing them with an empty string
2. Add a column `text_length` per review, representing the number of chars of the review text.
3. Group the dataset by `user_id` while taking the mean value of the grouped `text_length`.

This transformation corresponds to question **3L** of assignment 01.

q16j — Average Number of Feedback Votes

Calculating the average number of feedback votes takes multiple steps as well.

1. Create a new column `total_votes` which contains the sum of `the cool_votes`, `useful_votes` and `funny_votes`.
2. Group the dataset by `user_id` while taking the mean value of the grouped `total_votes`.

3. METHODS

Briefly describe the methods you used (such as k-means) and why you used them

K-means — being the only clustering technique discussed as of now — is also used to investigate this clustering. In order to create a meaningful clustering, the following steps were executed.

3.1 PRE-PROCESSING

Initially all non-relevant features were dropped from the dataset and a random set of 10299 instances was sampled from the dataset.
The selected instances were then normalized to avoid emphasizing the weight of any of the features in the clustering.

3.2 SILHOUTTE SCORE

With the normalized dataset the silhouette score was calculated for each k in [2,8]. The score achieved were

k	Silhouette Score
2	0.5656276370850483
3	0.5406516023674773
4	0.49849950955122563
5	0.49794042764313323
6	0.47582333657491593
7	0.4417290185334103
8	0.4001700149347306

Given the shallow decline of the silhouette score — and the fact that a 2-clustering would be less interesting with regards to the selected features — it was decided to use k-means with k=3 to cluster the samples.

4. RESULTS

Tell the story of how you investigated the question (plots are always nice) and the conclusions you drew.

Using the methodology described above we generated the following clusters.

Cluster	# Instances	Center
C0	9263	**q3**: 0.00453321, **q15**: 0.00730632, **q16j**: 0.00314
C1	955	**q3**: 0.02464726, **q15**: 0.00806536, **q16j**: 0.00525886
C3	81	**q3**: 0.00873684, **q15**: 0.01823072, **q16j**: 0.07021569

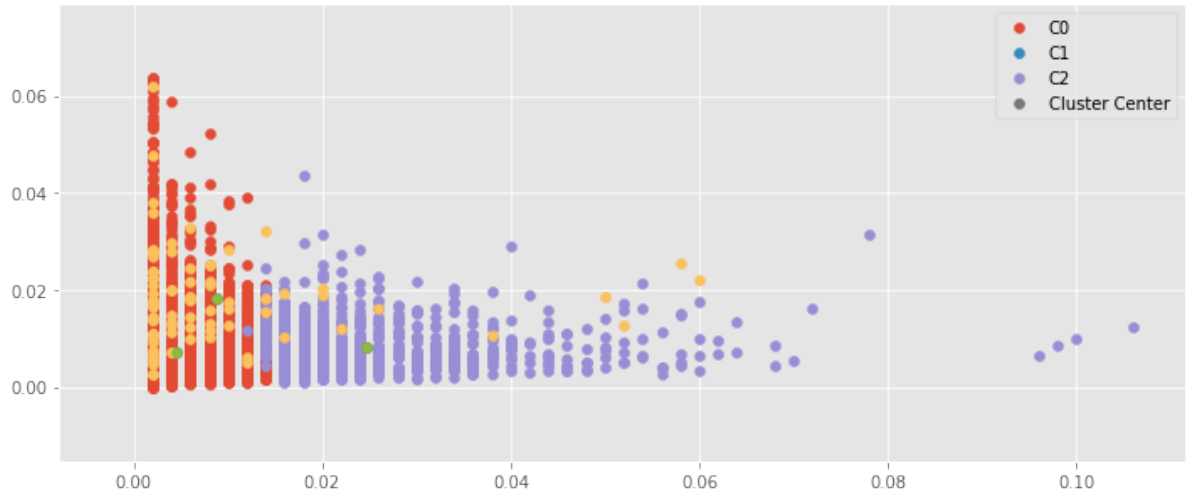

Figure 1: Plot of the resulting clustering. (x: number of reviews, y: average text length)

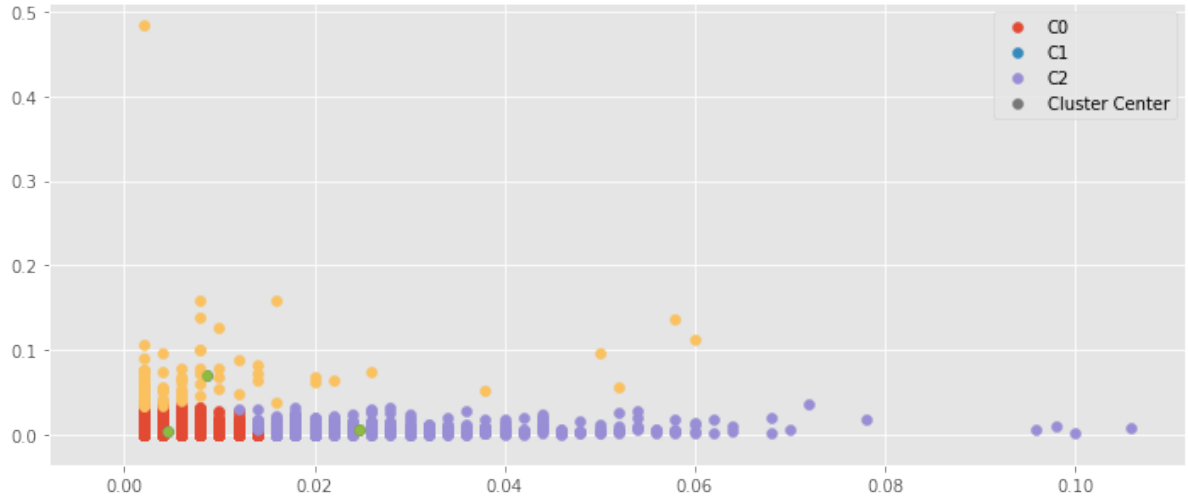

Figure 2 Plot of the resulting clustering. (x: number of reviews, y: average votes)

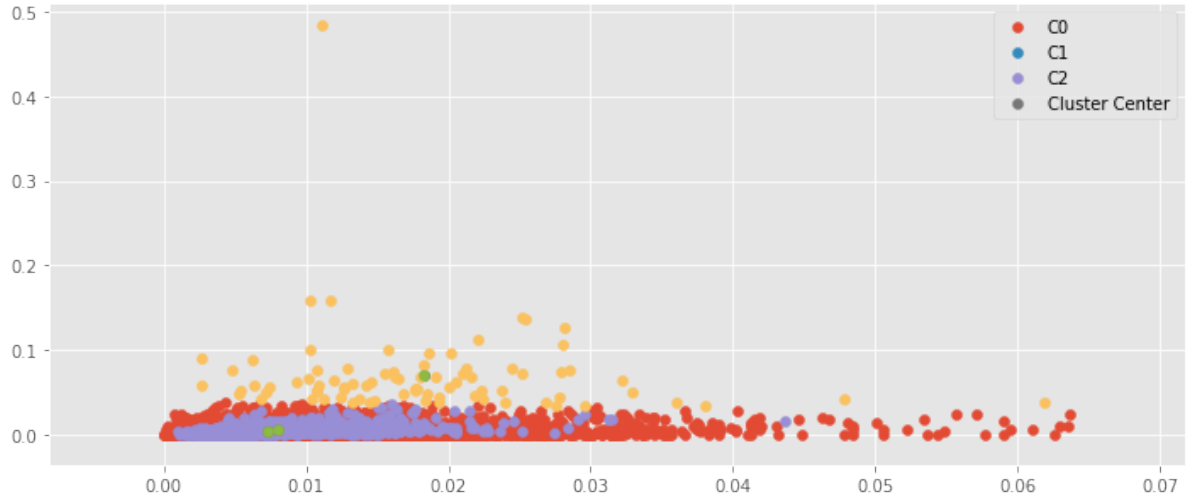

Figure 3 Plot of the resulting clustering. (x: average text length, y: average votes)

4.1 INTEPRETATION

Given the resulting clustering we have identified that cluster **C0** contains the most users from the sampled set. The cluster center is along all three features strictly worse than the centers of the other clusters, indicating that less active and less interesting users are grouped in this cluster. This interpretation is supported by the fact that the

The center of the **C1** cluster has a **q3** value which is about 5 times higher than the other cluster centers. This indicates that highly active user — those who write many reviews — are in this cluster.

The **C2** cluster seems to group users who received a remarkably high number of votes for their reviews. The **q16j** center value is 12x higher than that of any cluster and interestingly the **q15** value is more than double the value of the other clusters. This indicates that longer reviews are awarded with more votes than shorter ones.

4.2 LIMITATIONS

This approach is limited by several factors.
First, we used only a small sample (6 %) of the dataset for the clustering. Thus, a clustering considering all instance could result in a largely different clustering.
A second limitation is that the interpretations of the clusters is only considering the cluster centers, ignoring potential spread around the center of the cluster.

This clustering can therefore only be seen as an initial indication that useful clusters can indeed be created with the suggested approach. Further research on the entire dataset in necessary to confirm this assumption.