Ivan Lytovka, 500861433
Daniyil Kotov, 500877422

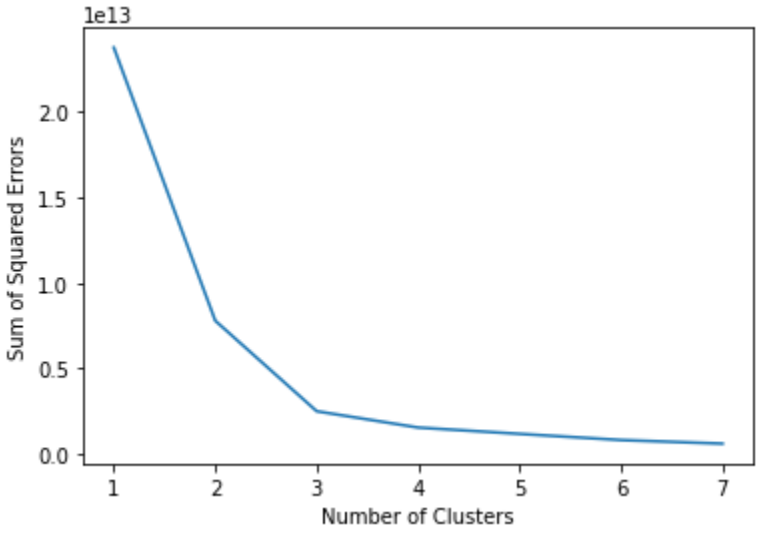
Background

For the data classification task in assignment we will use the Dry Bean data set from UCI Machine Learning Repository [1]. The rationale to use this data set is because it has an acceptably large number of records: 13611 data entries and 17 attributes in each record. This data is distributed into 7 classes.

For the association analysis task, we will use a different dataset Congressional Voting Records Data Set [2]. This dataset contains categorical data that's suitable for Apriori rule generation. Each record has a class (republican or democrat) representing a particular senator's attribution to a political party. In addition, it provides records on how a senator voted for 16 different initiatives (yea or nay). By utilizing this dataset, one might establish voting patterns, which makes this dataset interesting.

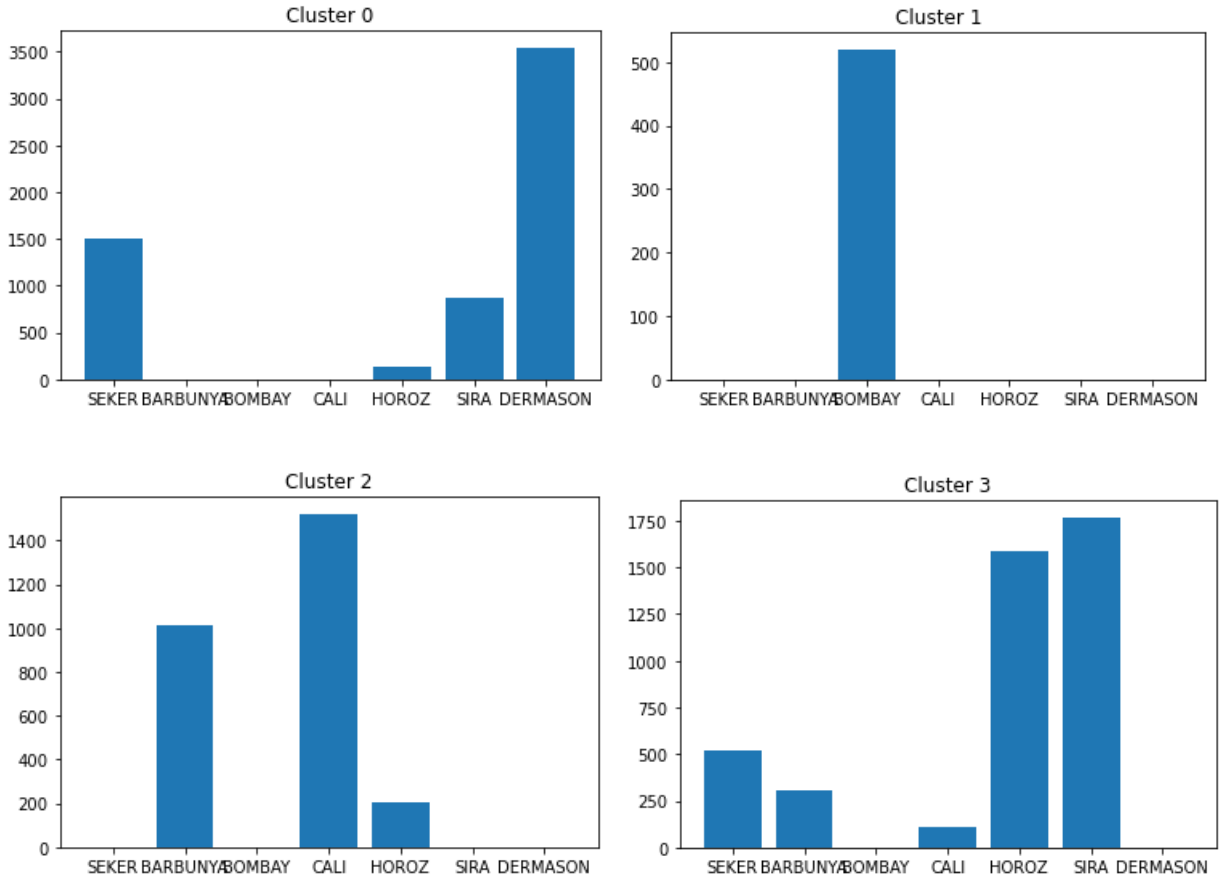
Methods & Results - Clustering

Data clustering method consists of identifying a suitable number of clusters and plotting bar charts that show attribution of record to a newly-formed cluster. First, we plot a graph to visualize kmeans inertia-to-values relationship. Upon plotting such a graph, we concluded that the optimal cluster number is 4: graph decreases linearly onwards.



Ivan Lytovka, 500861433
Daniyil Kotov, 500877422

Next, we iterate unique cluster values to plot bar charts that show distribution of actual classes inside of clusters. We also calculated each cluster's purity to get a numerical appreciation of how dominant the most common cluster was.



Ivan Lytovka, 500861433
Daniyil Kotov, 500877422

Methods & Results - Association

Since we're using apriori algorithm, we need to supply min_support and min_confidence values. We decide to consider rules where attributes are presented in at least 50% of the transactions with probability of 80%. For example, one of the following generated rules hints that democrats were more likely to vote 'nay' for the physician fee freeze bill.

We had to alter the initial dataset by applying a lambda function that appends column name to the attribute value in order to differentiate identical values in different attributes (initially each attribute was binary - y or n. We converted it to name-of-the-bill-y/n)

```
Rules:
['adoption-of-the-budget-resolution_y'] --> ['class_democrat']
Support: 0.5310344827586206 Confidence: 0.9130434782608695 Lift: 1.4875427454811918

['adoption-of-the-budget-resolution_y'] --> ['physician-fee-freeze_n']
Support: 0.503448275862069 Confidence: 0.8656126482213439 Lift: 1.5244595221711927

['aid-to-nicaraguan-contras_y'] --> ['class_democrat']
Support: 0.5011494252873563 Confidence: 0.9008264462809917 Lift: 1.4676385922555484

['class_democrat'] --> ['physician-fee-freeze_n']
Support: 0.5632183908045977 Confidence: 0.9176029962546816 Lift: 1.6160214711367875

['adoption-of-the-budget-resolution_y'] --> ['physician-fee-freeze_n', 'class_democrat']
Support: 0.503448275862069 Confidence: 0.8656126482213439 Lift: 1.5369040896991208
```

Conclusions

From the clustering analysis, we were able to conclude that the elbow method for approximating the k number of clusters is not the most precise, because our initial dataset had 7 classes instead of the 4, derived by the elbow method. This caused some of the classes to be unrepresented by their own clusters and hence harmed the purity of each cluster. However, the elbow method is a relatively easy method to estimate the amount of clusters needed in the case if you do not have empirical data to derive the needed number.

Ivan Lytovka, 500861433
Daniyil Kotov, 500877422

From the association analysis, we were able to conclude that the apriori algorithm (at least its python interpretation in the apyori library) is vulnerable to different attributes having the same range of values, which adds an additional step in the dataset preprocessing stage.

References

[1] https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset

[2] https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records