Test Assignment

Clustering Algorithms

Yulia Bezginova

Email: *ybezginova2021@gmail.com*
Solution at Github: https://github.com/ybezginova2016/UL_01_HierarchicalClustering

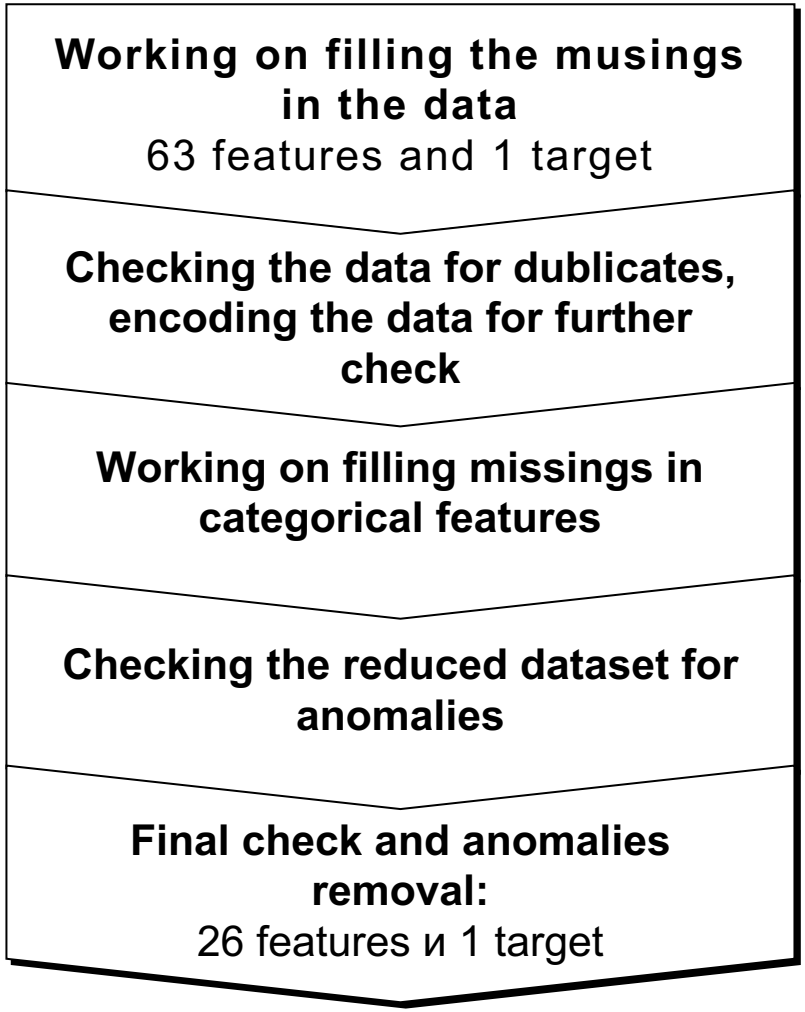
Test Assignment

1 - Data Preprocessing

Data source:
https://drive.google.com/file/d/1i9QwMZ63qYVlxxde1kB9PufeST4xByVQ/view

Task: https://sm2foundation.notion.site/348bfd44703f402787611be7328dc704

Data Preprocessing



- Initial given dataset contained 64 columns and 15403 observations
- There were not explicit dublicates in the dataset
- As a result, the feature matrix was reduced to 29 features, 1 target and 15401 observations
- Checking each column in the dataset for the unique values
- <10% missings were safely removed
- Some of the categories were filled out
- Some of the features were excluded from the analysis due to a high level of noise and missing values

Check class for a disbalance

25 features and 1 class

- Уменьшили размер матрицы признаков с 121 до 38 входных параметров

Class disbalance

- There are only 2 classes in the dataset

```
545988074 15358
647859449    40
```

Clustering is to be done via unsupervised learning (the class is excluded from the algorithm's input

As a result of data preprocessing, 39 columns out of 64 were dropped, since they had 100% of missings values; 5 columns contained 50% of missings, and were filled out by the respective values represented in the other filled 50% of the column.

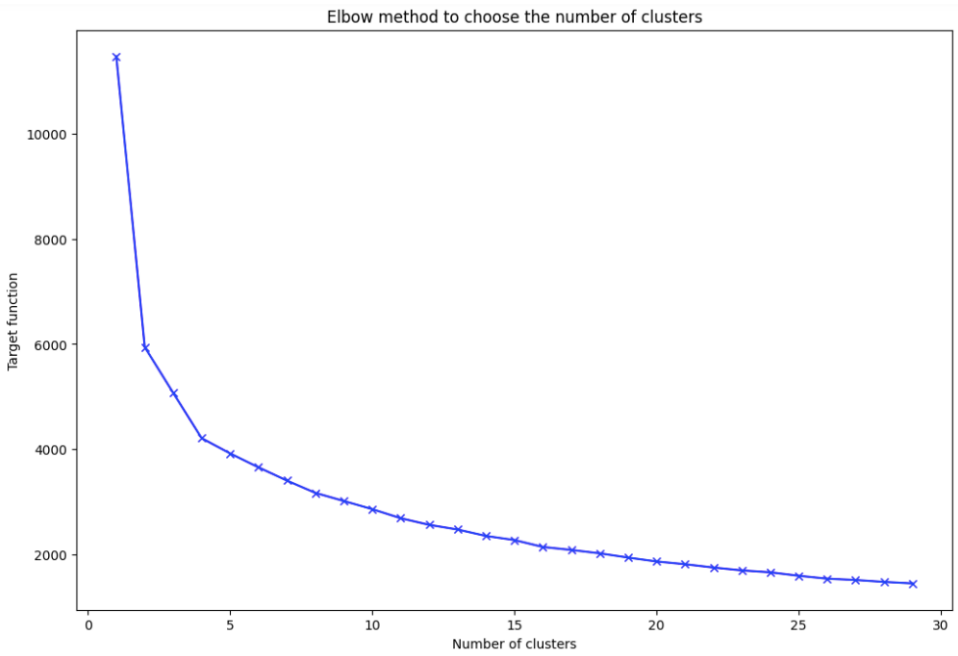
data_clean.csv was recoded for further clustering

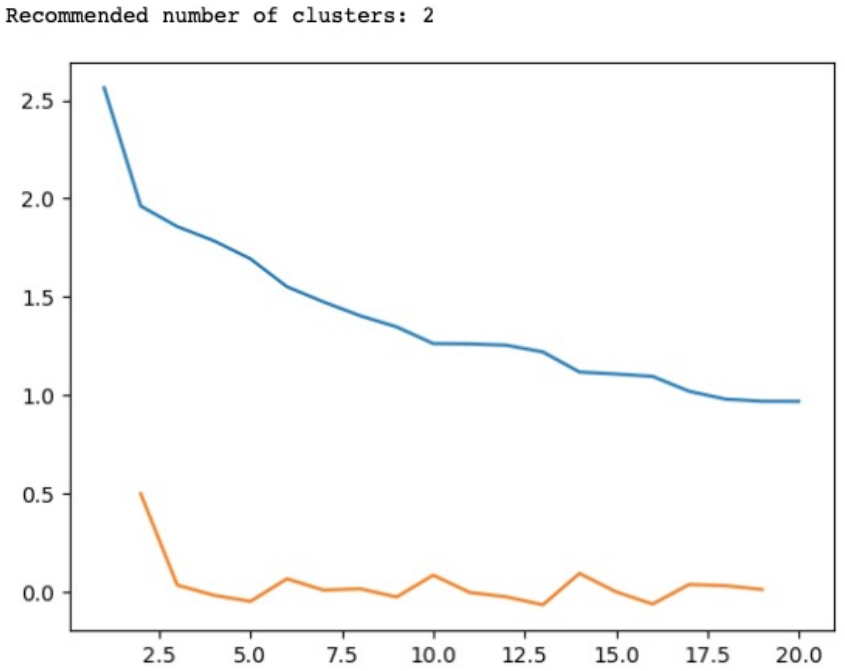
Test Assignment

2 - Clustering

- k-means algorithm
- Agglomerative Hierarchical Clustering
- clustering algorithm evaluation

k-means clustering: Elbow method





Elbow method

- 2, 5, 7, 10, 30 clusters to try
- 2 clusters were advised by the Elbow method

Clustering Algorithm Evaluation

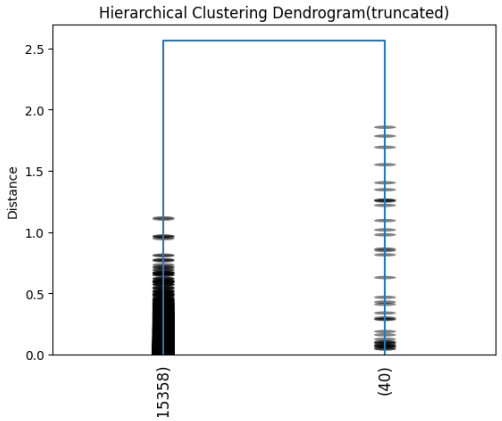
- Homogeneity metric
- Completeness metric
- V-Measure

K-Means basic clustering algorithm was built for 2, 5, 10, 15 and 30 clusters.

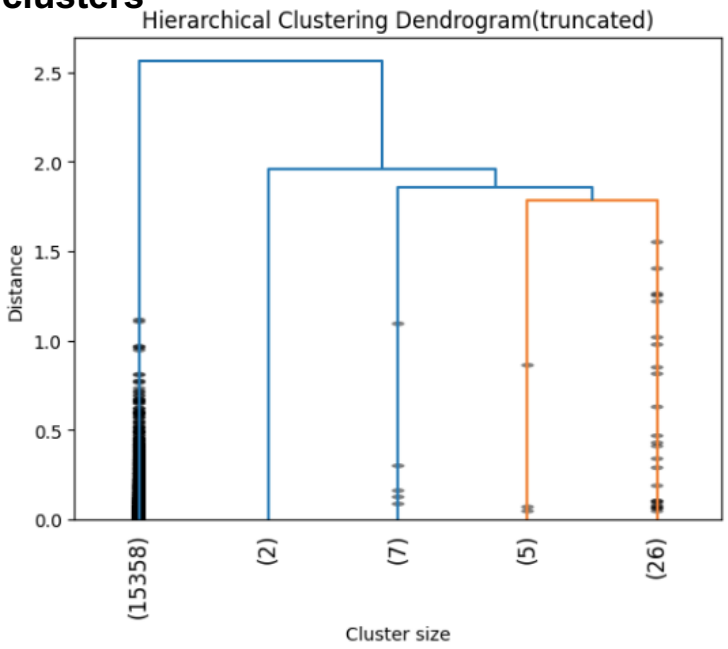
From homogeneity, completeness and V-measure we can conclude that 2 clusters are the most precise, since two-clustering algo shows the least homogeneity.

k-means clustering: dendrogram

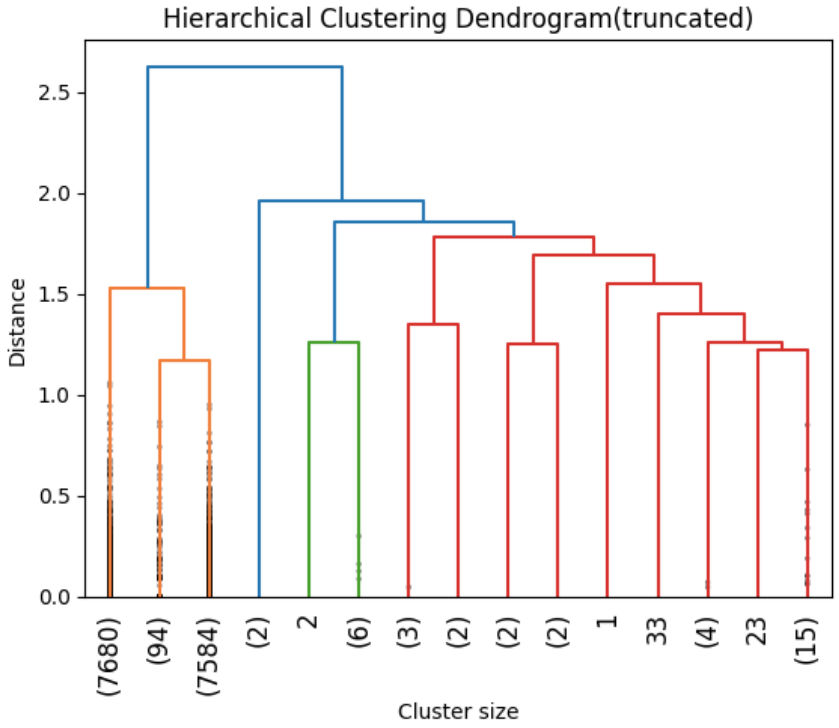
2 clusters



5 clusters



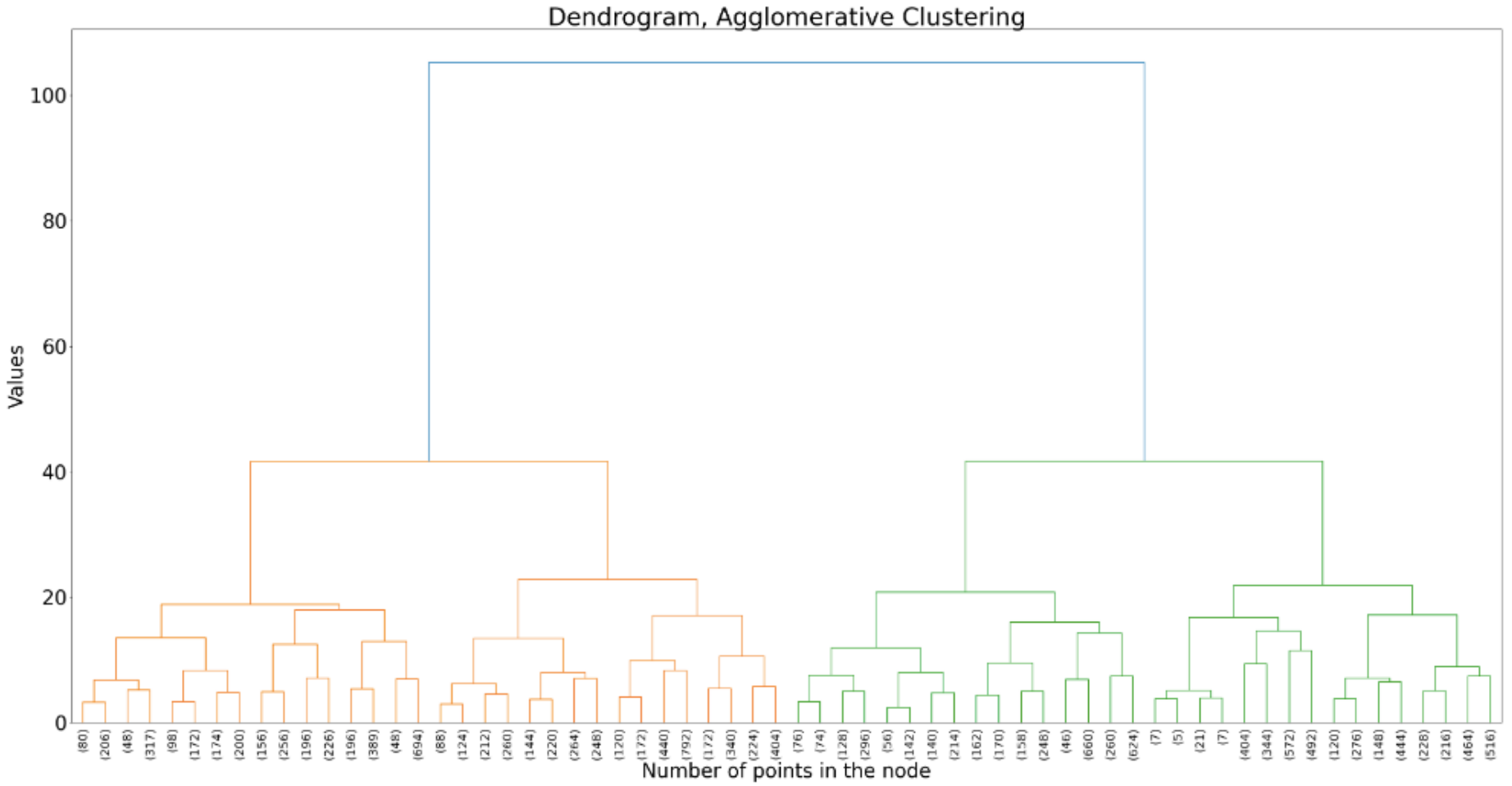
15 clusters



After training k-means algorithms (using Euclidian distance), dendrograms show that the dataset
well divided into 2 clusters.

Agglomerative Hierarchical Clustering

```
from sklearn.cluster import AgglomerativeClustering
```



Even though we set in the agglomerative hierarchical clustering algorithm, the dendrogram shows that the dataset still nicely divided only into 2 clusters.

Test Assignment

Thank you

Yulia Bezginova

Email: *ybezginova2021@gmail.com*
Solution at Github: https://github.com/ybezginova2016/UL_01_HierarchicalClustering