INDIVIDUAL-LEVEL BASED CLUSTERING ON ARTICLE READERSHIP DATA

WHAT ARE THE INTERESTS OF INDIVIDUALS READING ARTICLES ON MONDAQ?

ABSTRACT

The purpose of this research is to analyse the article readership data of different users who visit and read articles on mondaq.com website in order to derive an approach for grouping those users into clusters based on their shared interests.

This gives them a very good idea of who their customers are and making it easier to send tailored insights to the right individual at the right time.

INTRODUCTION AND LITERATURE REVIEW

- Reading behaviour
- Behavioural patterns
- Customer Segmentation
- User Interest Patterns

Customer Segmentation

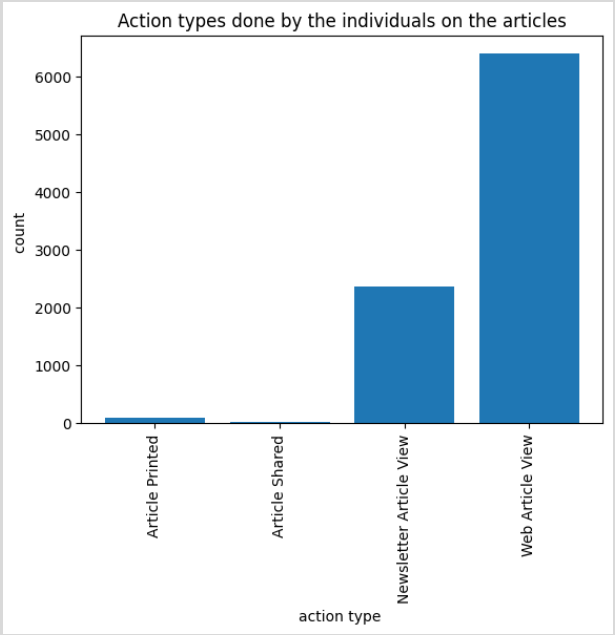
DATA SOURCE

- **Primary data** - action_date, action_month, article_id, article_country_id, article_country, article_primary_topic_id, article_primary_topic, action_type, individual_id, individual_country_id, individual_country, company_id, master_company_id, master_company_industry_id, master_company_industry

- **Articles topics data** - article_id, topic_id, topic_desc

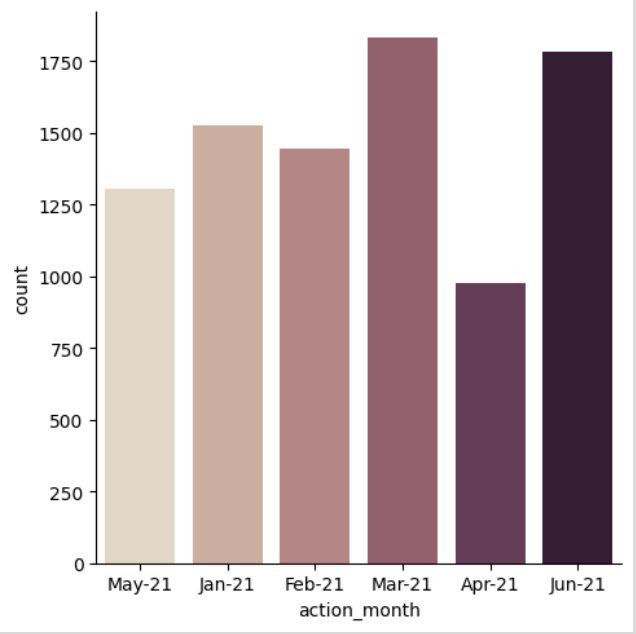
- **Newsletter subscription data** - individual_id, sub_topic_desc, action_type

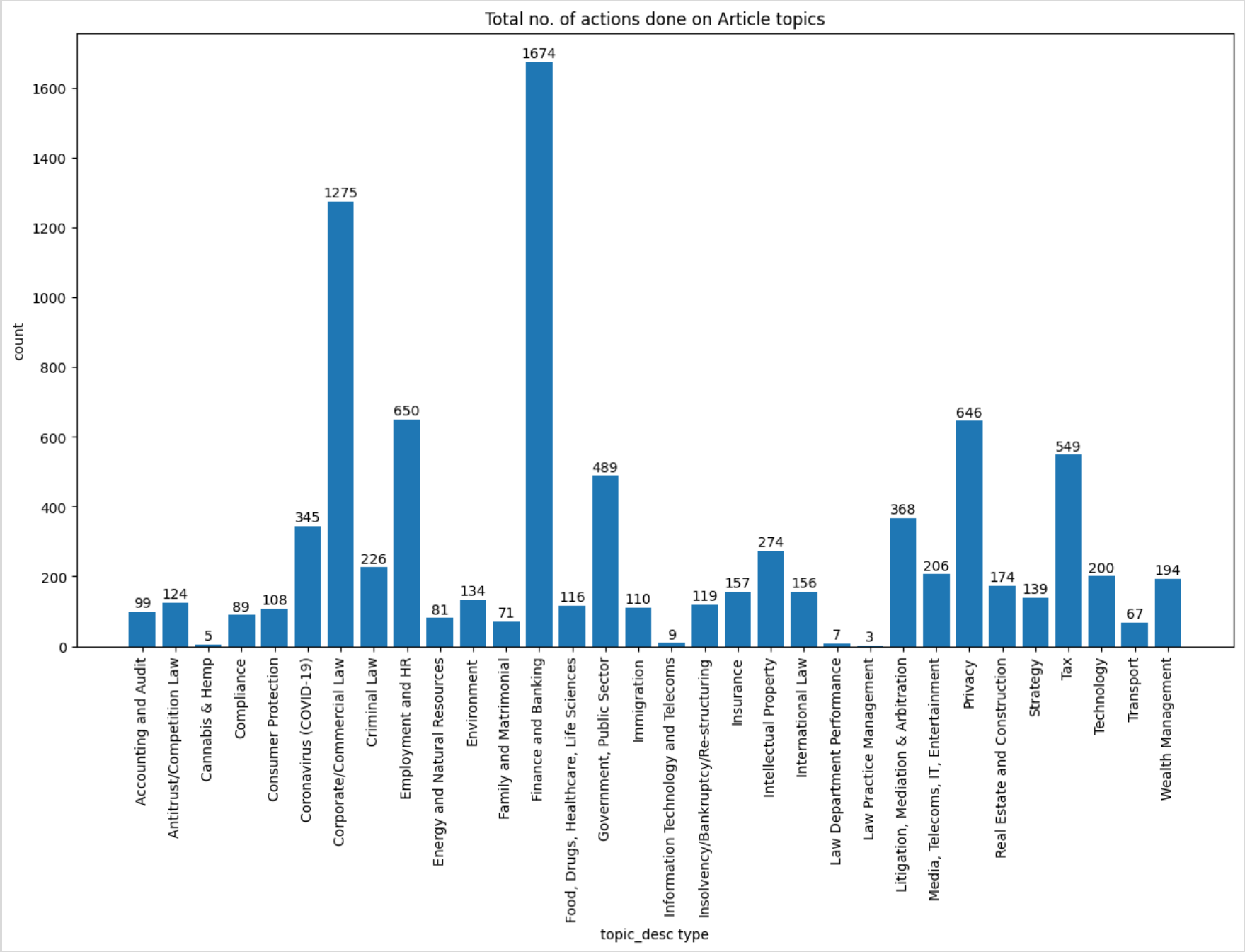
EXPLORATORY DATA ANALYSIS



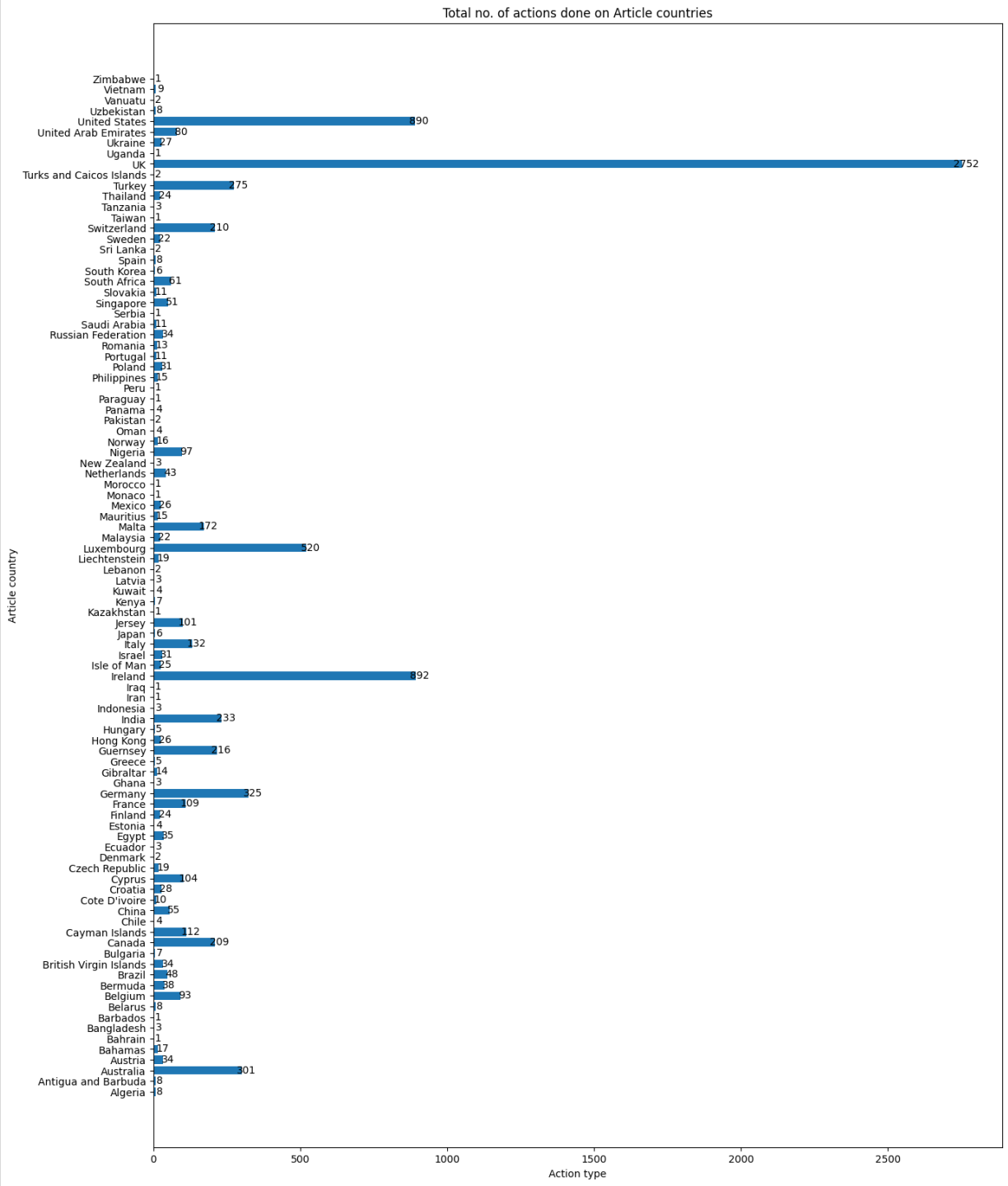
Action type

Action month

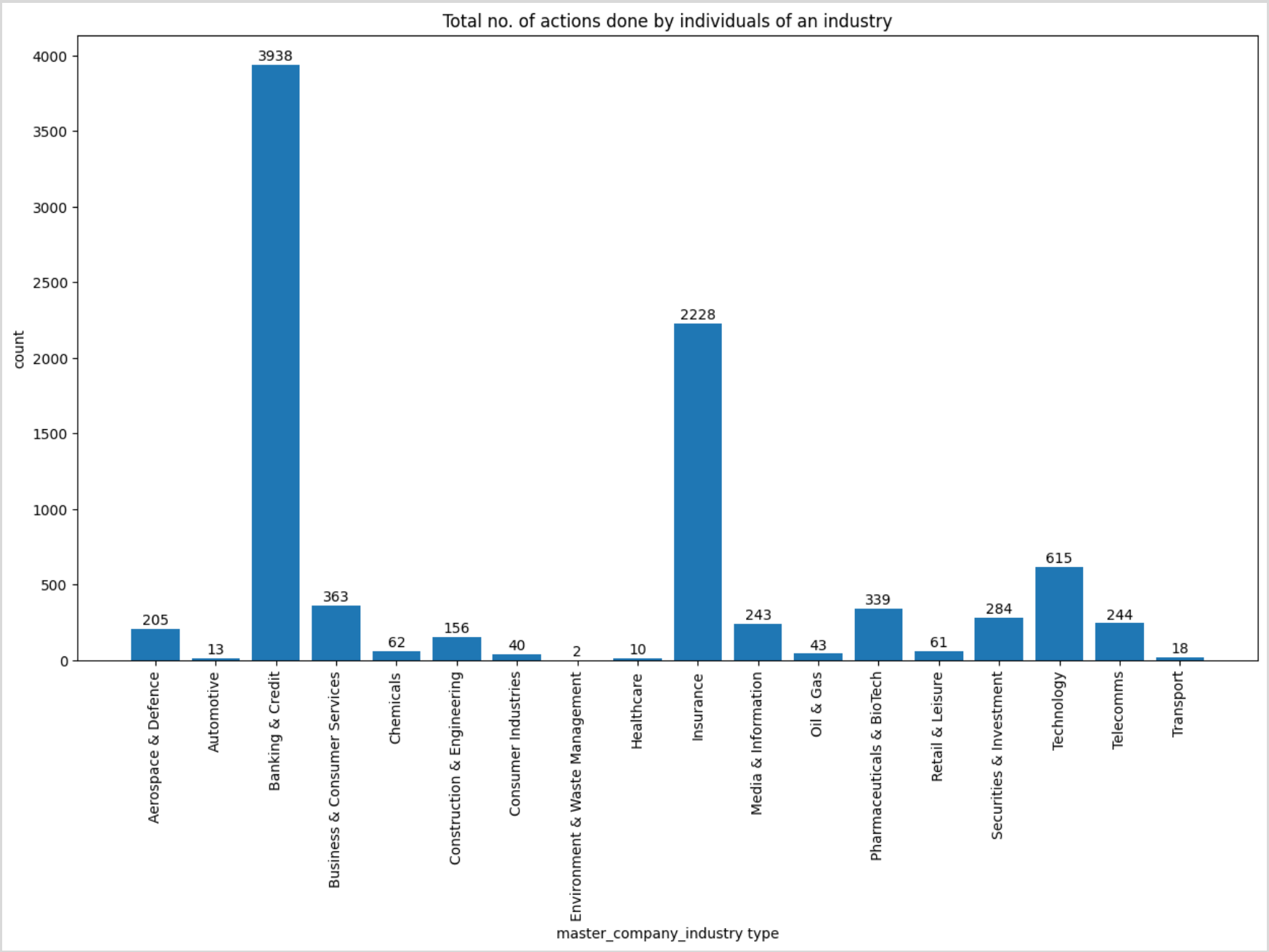




Article topics



Article countries



Individual master company industry

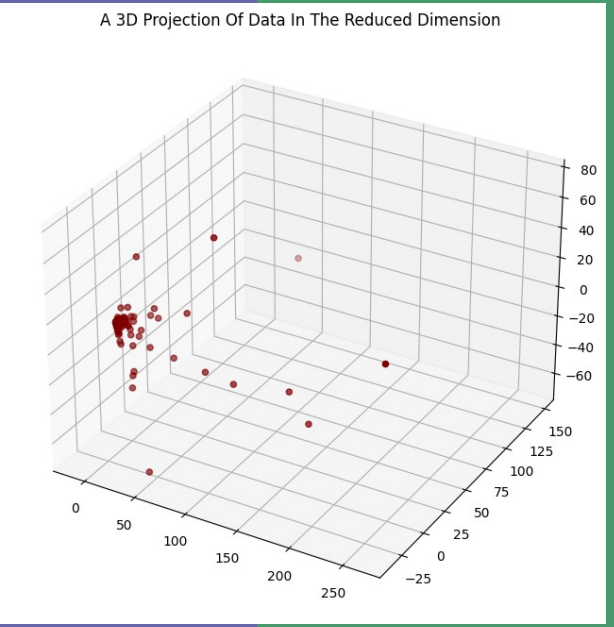
FEATURE ENGINEERING

Feature Scaling -

- Robust Scaling

Dimensionality Reduction technique -

- Principle Component Analysis

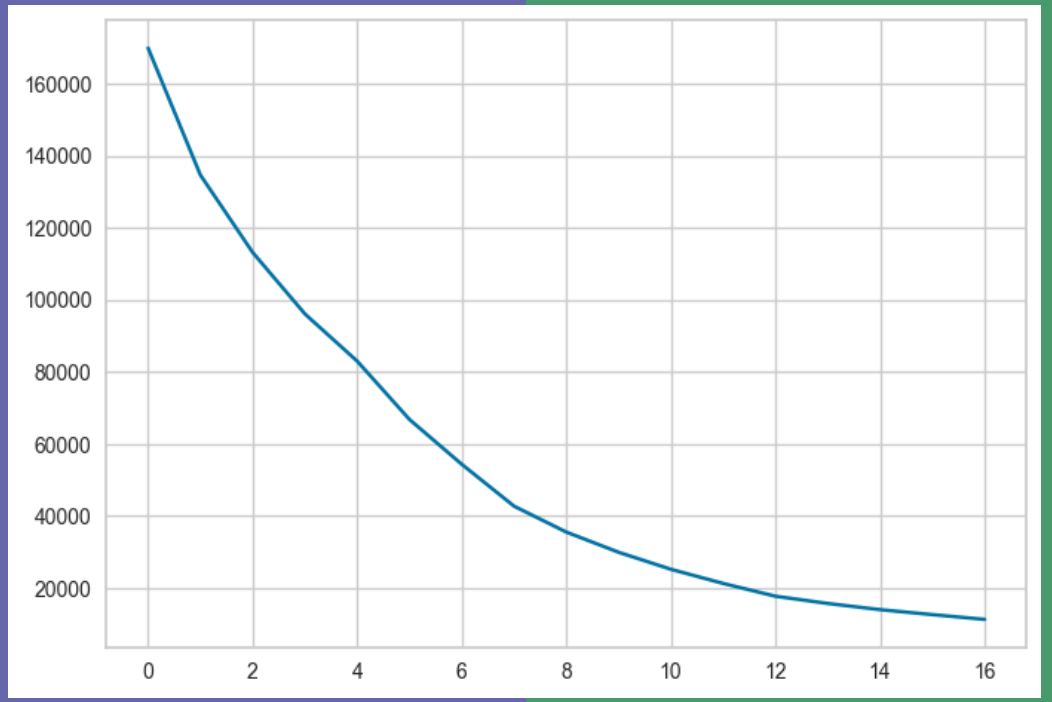


PCA

METHODS

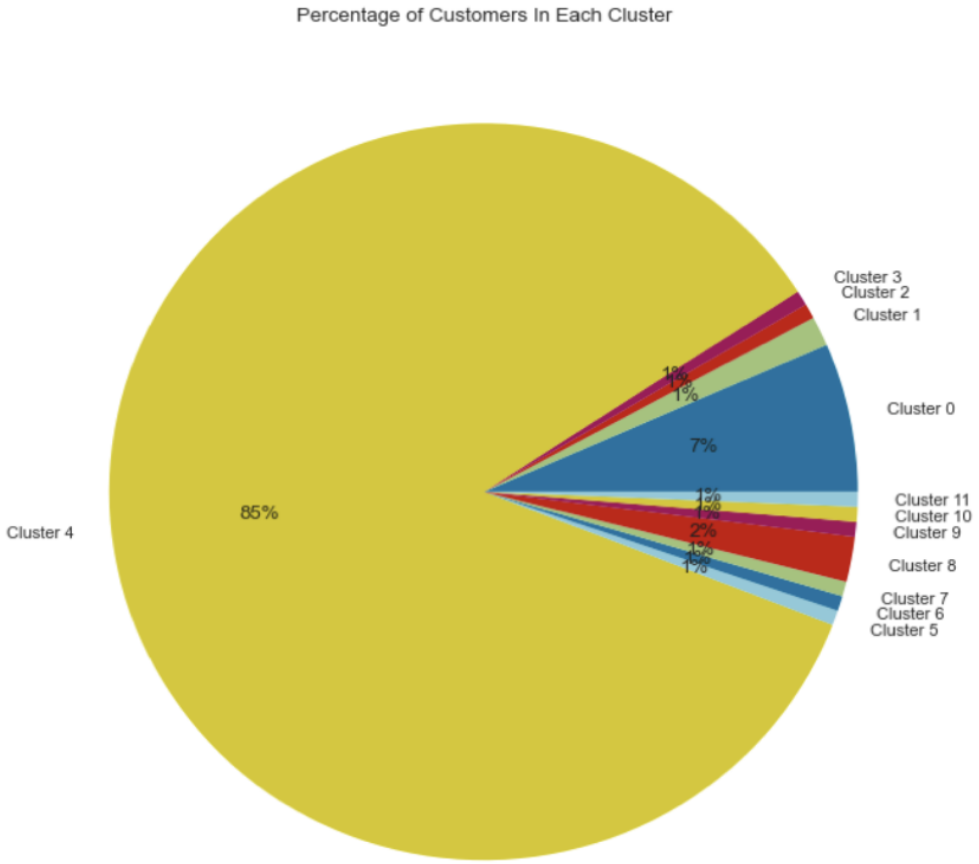
Clustering Techniques

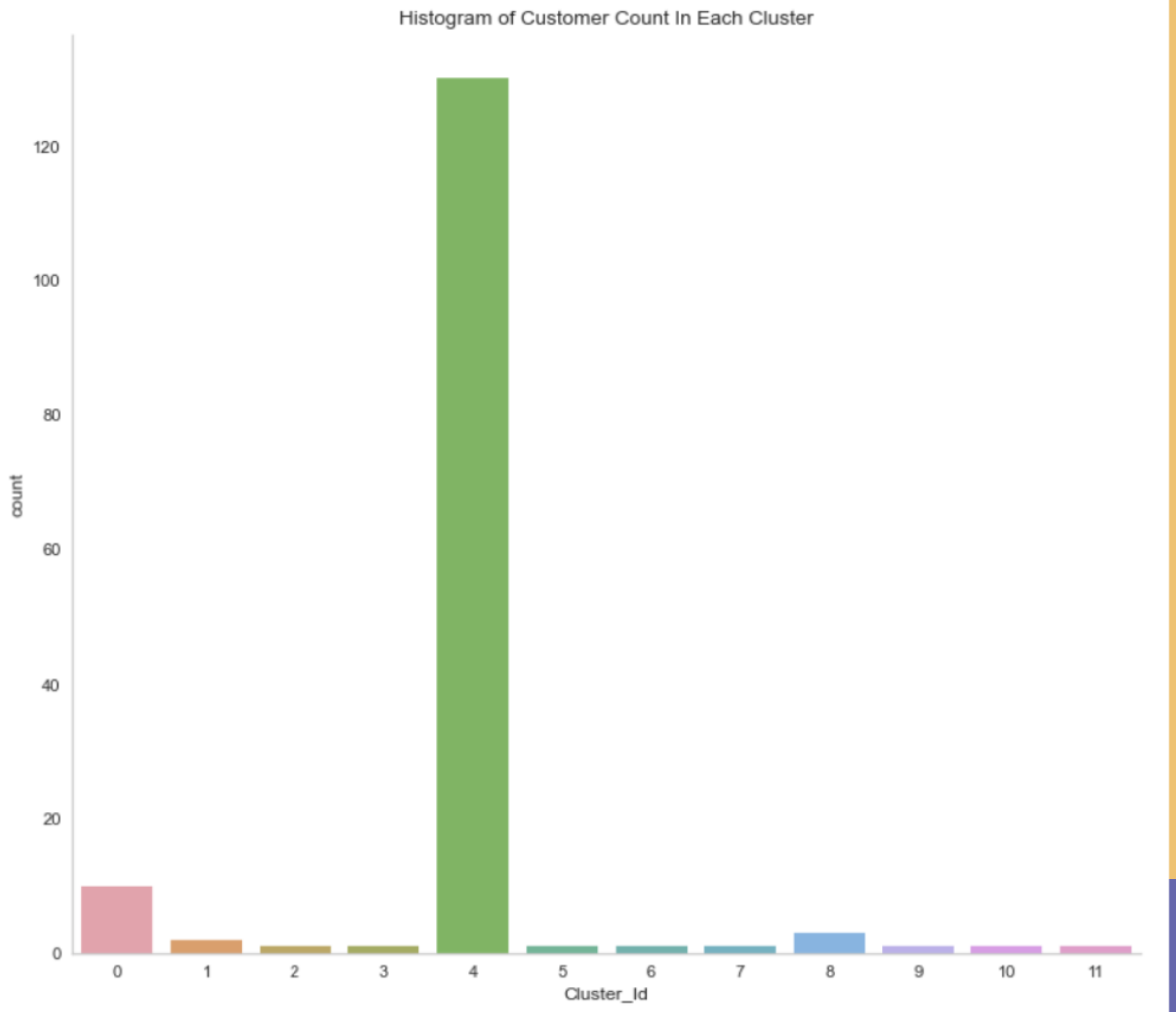
- Kmeans
- DBSCAN
- Minibatch Kmeans



K Elbow method for 12 clusters

RESULTS

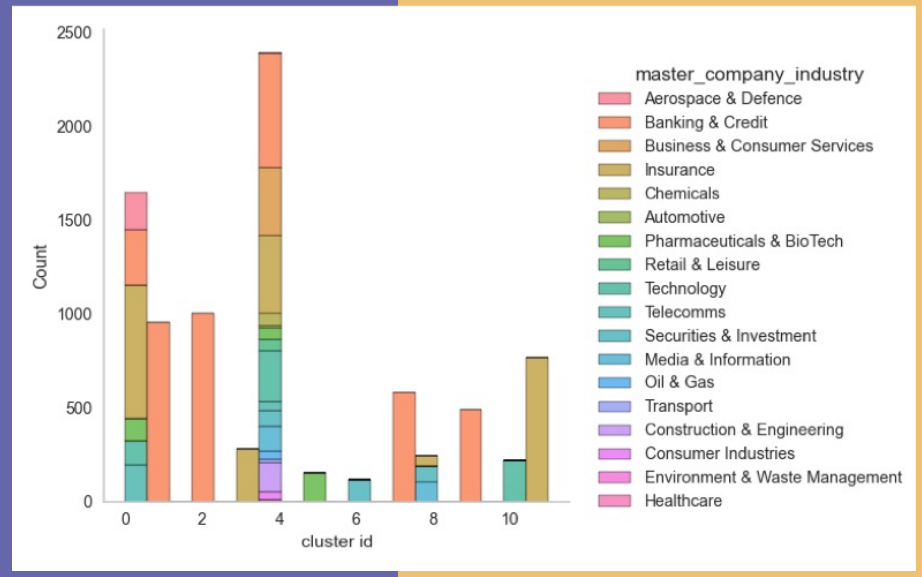




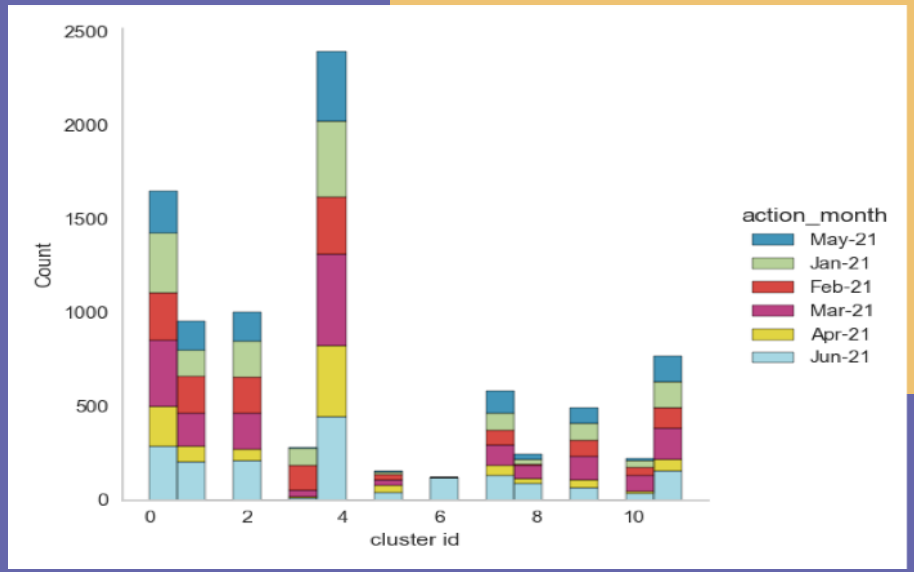
Pie chart and histogram for 12 optimal no of clusters

INDIVIDUALS

- Cluster 0 – Consists of 10 individuals who have interests in **32** topics of **61** region based articles and are employees of **6** master company industries (**subscribers and non-subscribers**).
- Cluster 1 - Consists of 2 individuals who have interests in **29** topics of **72** region based articles and are employees of **Banking and Credit** master company industry (non-subscriber).
- Cluster 2 - Consists of 1 individual who has interests in **29** topics of **51** region based articles and is working in a **Banking and Credit** master company industry (**non-subscriber**).
- Cluster 3 - Consists of 1 individual who has interests in **16** topics of **9** region based articles and is working in a **Insurance** master company industry (subscriber).
- Cluster 4 - Consists of 130 individuals who have interests in **29** topics of **58** region based articles and are employees of **18** master company industries (subscribers and non-subscribers).



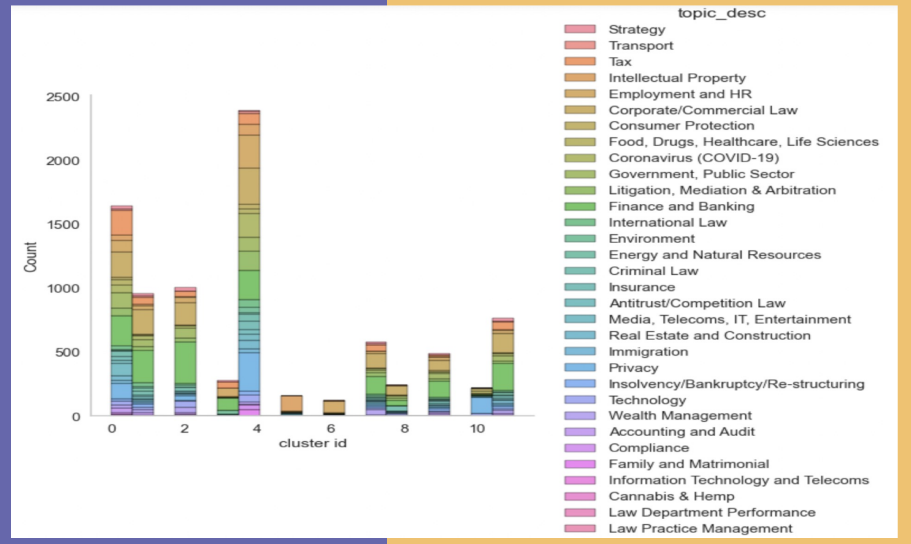
Master company industry vs Clusters



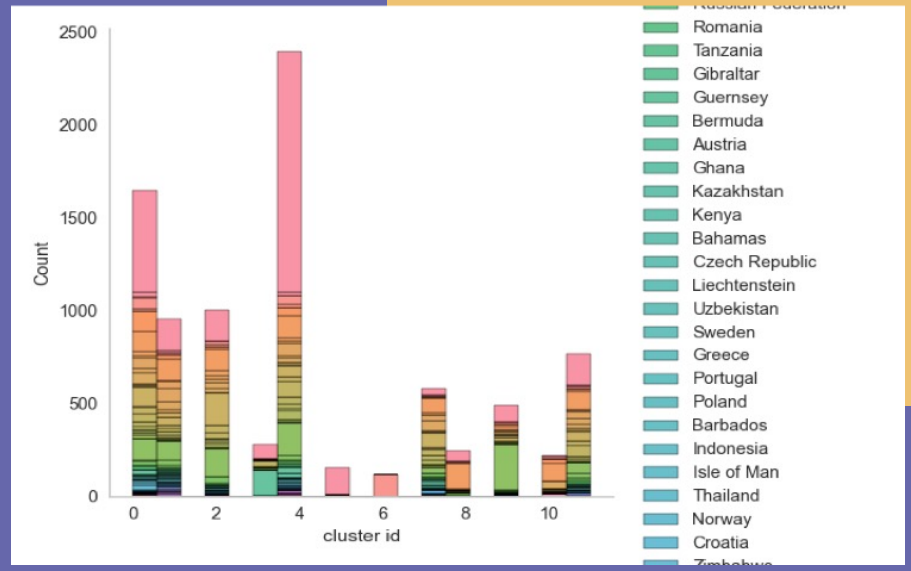
Action month vs Clusters

- Cluster 5 - Consists of 1 individual who has interests in **10** topics of **Germany and UK** region based articles and is working in a **Pharmaceuticals and Biotech** master company industry (subscriber).
- Cluster 6 - Consists of 1 individual who has interests in **9** topics of **Australian** region based articles and is working in a Securities and Investment master company industry (subscriber [June month]).

- Cluster 7 - Consists of 1 individual who has interests in **28** topics of **45** region based articles and is working in a **Banking and Credit** master company industry (non-subscriber).
- Cluster 8 - Consists of 3 individuals who have interests in **23** topics of **12** region based articles and are employees of **3** master company industry (subscriber and non-subscriber).
- Cluster 9 - Consists of 1 individual who has interests in **28** topics of **28** region based articles and is working in a **Banking and Credit** master company industry (non-subscriber).
- Cluster 10 - Consists of 1 individual who has interests in **12** topics of **12** region based articles and is working in a **Technology** master company industry (subscriber).
- Cluster 11 - Consists of 1 individual who has interests in **28** topics of **45** region based articles and is working in a **Insurance** master company industry (non-subscriber).



Article topics vs Clusters



Article countries vs Clusters

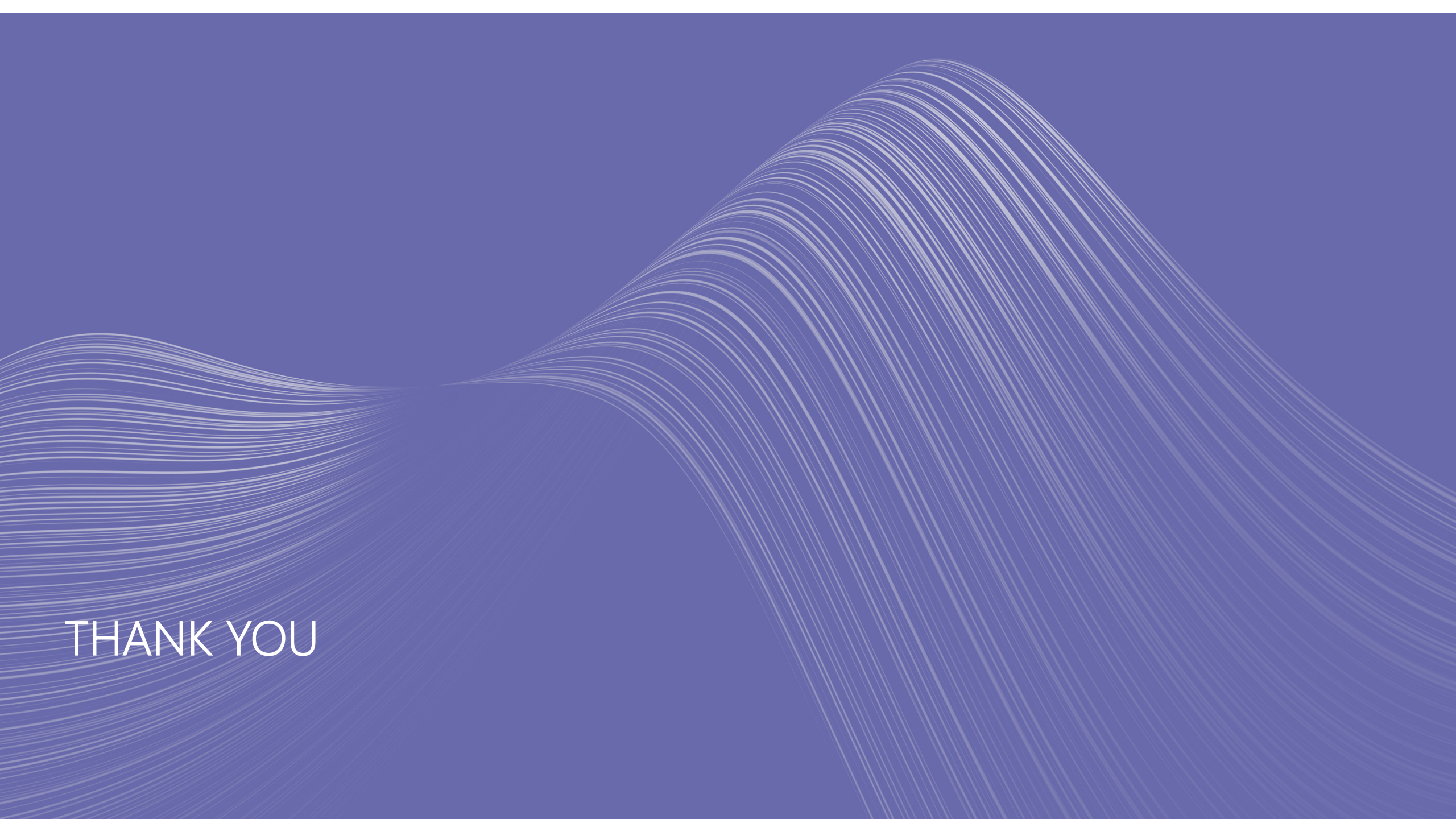
CONCLUSIONS

Silhouette Scores :

Algorithms	Scores
Kmeans	0.78
MiniBatch Kmeans	0.30
DBSCAN	0.18

FUTURE WORK

- A new attribute can be added to the dataset, shared to how many people and their usage actions (co-readership).
- Do the viewers read all the subscribed content and does that action lead to more subscription of topics and article regions.


THANK YOU