MN50752 DATA MINING & MACHINE LEARNING COURSEWORK 2023-24

by Ritwik Singh

Word Count: 2745

MSc Business Analytics
SCHOOL OF MANAGEMENT
University of Bath

Contents

1. Introduction

This research conducts a comprehensive analysis of user behaviour data obtained from the social media platform of Company Z. The principal objective is to identify observable trends and subgroups among the user population. Company Z hopes to learn vital details about posting behaviours, content preferences, and user engagement through finding these patterns. Strategic decision-making involving tailored content delivery, concentrated advertising campaigns, and platform feature improvements could be influenced by such findings. By means of this analysis, Company Z hopes to improve its knowledge of its user base and, in turn, build more substantial relationships with customers and accelerate platform growth. This study presents the methods used to perform the clustering analysis, explains the results, and offers practical suggestions for making the most of these insights.

2. Exploratory Data Analysis

The exploratory data analysis (EDA) was conducted to gain insights into the behavioural data obtained from Company Z's social media platform. The dataset consists of 2307 observations and 13 variables/features in total.

2.1 Boxplots of the Behavioural Variables: The primary goal of boxplot creation is to provide a visual representation of the data distribution and highlight important features such central tendency, variability, and the existence of outliers for each behavioural parameter.

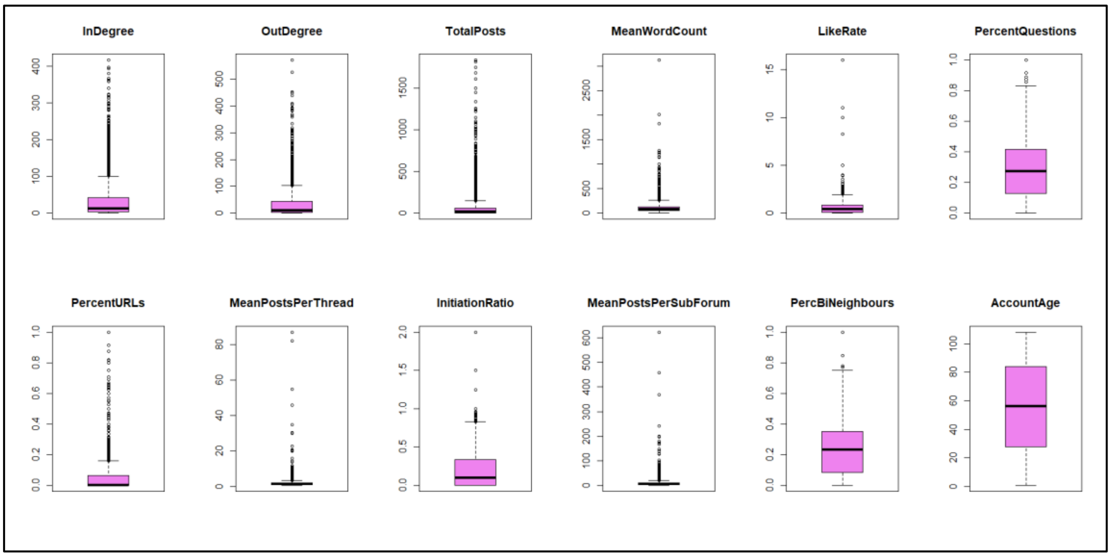


Fig 1. Exploring the Distribution of Behavioural Variables

1. **InDegree:** Most users have few incoming interactions, with only a few outliers having a high InDegree, possibly influencers.
2. **OutDegree:** Like InDegree, most users have low OutDegree values, with a few highly active users engaging with many others.
3. **TotalPosts:** Majority of users post infrequently, while a small number of 'power users' contribute significantly to the content.

4. **MeanWordCount:** Users typically keep posts short, with a few outliers consistently writing longer posts.
5. **LikeRate:** Most users receive few likes per post, with some outliers having exceptionally well-liked content.
6. **PercentQuestions:** Asking questions is common, with some users primarily focusing on question-based content.
7. **PercentURLs:** Most users post URLs infrequently, while some include URLs more often.
8. **MeanPostsPerThread:** Users generally make few posts in each thread, with some heavily engaging in specific threads.
9. **InitiationRatio:** Most users start few threads relative to their overall posting, with some starting many threads.
10. **MeanPostsPerSubForum:** Users post limitedly in each subforum, with some showing high activity in specific areas.
11. **PercBiNeighbours:** Moderate bidirectional interaction is common, with some users having strong reciprocal communication networks.
12. **AccountAge:** Account ages are evenly distributed, with some long-standing members. The high account age could be inactive users.

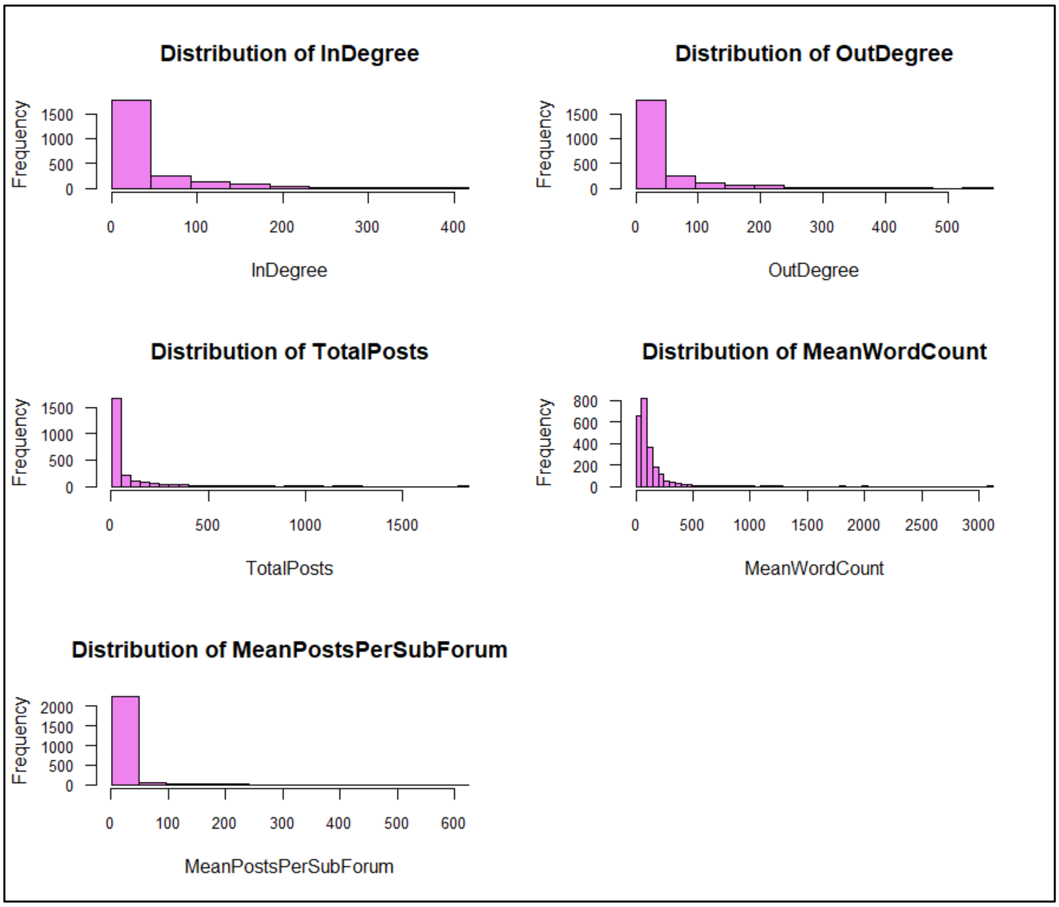


Fig 3. Histograms of some Behavioural Variables

1. **InDegree Histogram:** Reflects the total number of unique network neighbors interacting with a user. Skewed right, suggesting most users have few interacting neighbors.

2. **OutDegree Histogram:** Represents the total number of unique network neighbors receiving posts from a user. Like InDegree, skewed right, indicating most users have few receivers.
3. **TotalPosts Histogram:** Indicates the total number of posts per user. Right-skewed, revealing a right-skewed pattern indicative of few highly active users and many with lower activity levels. The long tail signifies outliers who contribute significantly more posts.
4. **MeanWordCount Histogram:** Depicts the mean average word count per user's post. Right-skewed, suggesting most posts have relatively fewer words.
5. **MeanPostsPerSubForum Histogram:** Illustrates the average number of posts per subforum per user. Right-skewed, implying most users are less active in most subforums.

2.2 Summary Statistics: Important numerical properties for every variable are provided by using summary(z_df). For example, the variable "InDegree" has a wide range: its mean is 36.32, its maximum is 416, and its minimum is 0. Comparable examinations are conducted for "OutDegree," "TotalPosts," and "MeanWordCount," exposing their distribution patterns and key characteristics. Metrics such as "LikeRate," "PercentQuestions," and "MeanPostsPerThread" provide important information about how users interact. In order to understand numerical properties and facilitate additional analysis, this summary serves as an important resource.

2.3 Missing Values: Missing values in every column are found and quantified using is.na() and colSums(), providing information about data completeness and possible quality problems. Analysing the distribution and existence of missing data improves data integrity and analysis precision by providing information for later data management techniques. In this dataset, **no missing data** is found, indicating no need for data imputation or removal of observations with missing values.

2.4 Pair Plotting: A thorough visual description of the relationships between variables is provided by the pair plot. Variable distributions are displayed by diagonal histograms, and the correlations between variable pairs are shown by off-diagonal scatterplots. These graphs aid in locating patterns and correlations within the data.

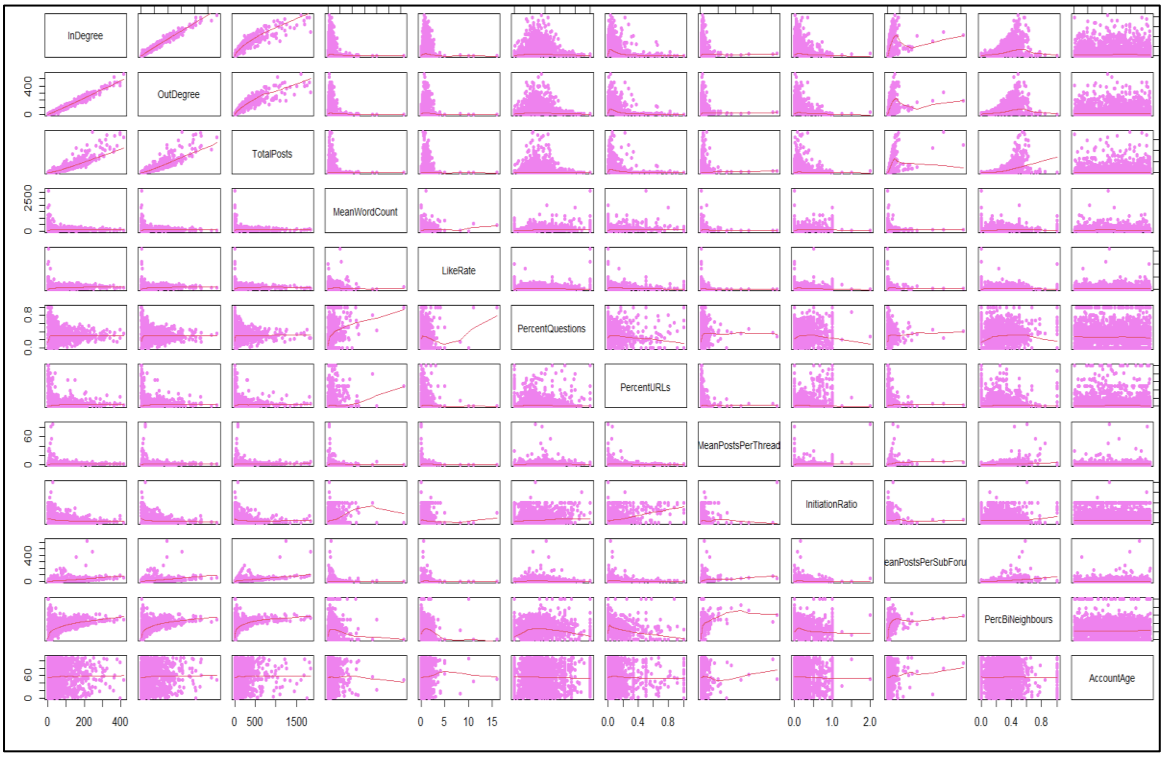


Fig 2. Pair plot Matrix of Behavioural Variables

Upon analysis of the pair plot:

- Variables like 'InDegree' and 'OutDegree' exhibit a positive correlation, indicated by an upward trend in their scatterplot.
- 'AccountAge' demonstrates a relatively uniform distribution across its range, as depicted in the histogram.
- Some variables, such as 'PercentQuestions' and 'PercentURLs', show scatterplots with less apparent trends, suggesting weaker correlations with other variables.
- While certain relationships appear linear, others lack a clear pattern, making it challenging to categorize them as exponential or sigmoidal.

2.5 User Engagement from Different Account Age Groups:

The scatter plot reveals insightful patterns:

Younger accounts (1 to 5 years) exhibit a wide range of post counts, typically lower than older accounts.

As account age surpasses 5 years, fewer accounts demonstrate very high post counts. While there's a tendency for older accounts to have more posts, the relationship isn't strictly linear, indicating considerable variability across all ages.

Notably, a cluster of 1-year-old accounts displays a relatively high number of posts, suggesting heightened activity during the initial year.

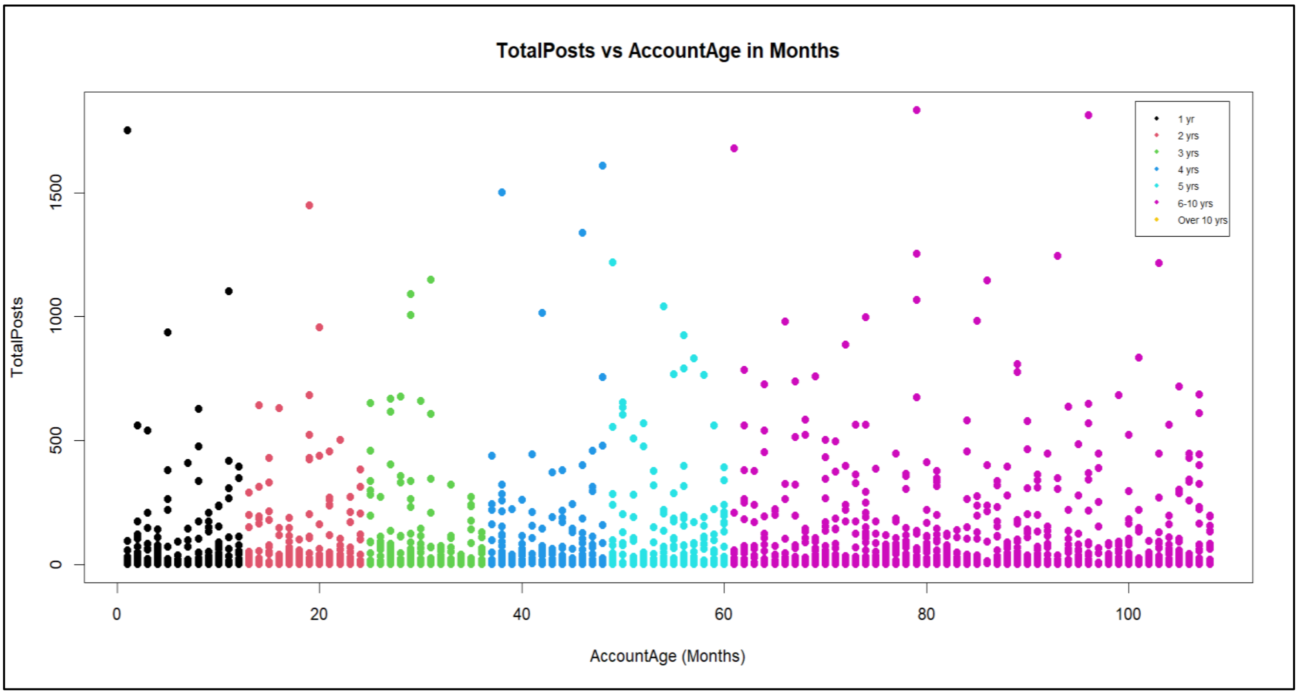
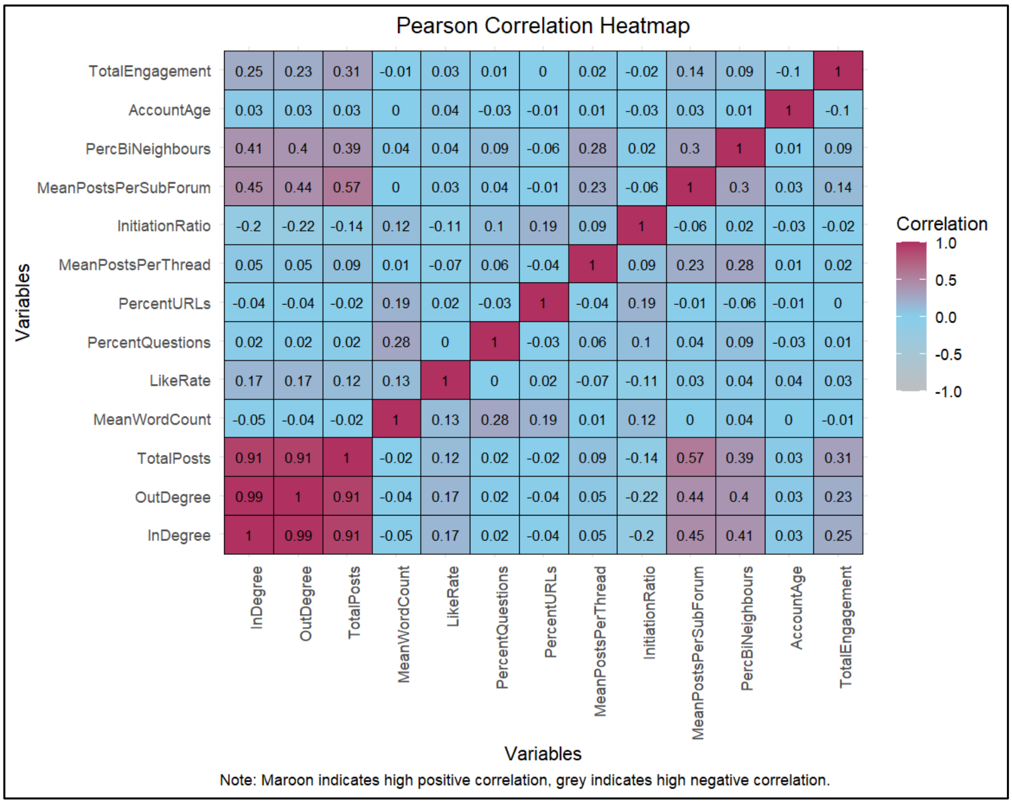


Fig 4. Total Posts vs Account Age Distribution

This visualization facilitates the analysis of user engagement dynamics over time, offering valuable insights for community management and marketing strategies.

2.6 **Pearson Correlation Coefficient Heatmap**: Correlation does not mean causation, to understand causality an in-depth analysis is essential. The Pearson correlation coefficients between a set of variables are shown visually in this Pearson Correlation Heatmap. The linear connection between two variables is measured by the Pearson correlation coefficient, which has a range of -1 to 1. One signifies an ideal positive correlation, minus one indicates an ideal negative correlation, and zero indicates no correlation.

Insights regarding user activity intensity in relation to account longevity are provided by the "TotalEngagement" statistic, which is calculated by dividing TotalPosts by AccountAge. This information helps to inform personalised plans and targeted interventions.



Variables	InDegree	OutDegree	TotalPosts	MeanWordCount	LikeRate	PercentQuestions	PercentURLs	MeanPostsPerThread	InitiationRatio	MeanPostsPerSubForum	PercBiNeighbours	AccountAge	TotalEngagement
TotalEngagement	0.25	0.23	0.31	-0.01	0.03	0.01	0	0.02	-0.02	0.14	0.09	-0.1	1
AccountAge	0.03	0.03	0.03	0	0.04	-0.03	-0.01	0.01	-0.03	0.03	0.01	1	-0.1
PercBiNeighbours	0.41	0.4	0.39	0.04	0.04	0.09	-0.06	0.28	0.02	0.3	1	0.01	0.09
MeanPostsPerSubForum	0.45	0.44	0.57	0	0.03	0.04	-0.01	0.23	-0.06	1	0.3	0.03	0.14
InitiationRatio	-0.2	-0.22	-0.14	0.12	-0.11	0.1	0.19	0.09	1	-0.06	0.02	-0.03	-0.02
MeanPostsPerThread	0.05	0.05	0.09	0.01	-0.07	0.06	-0.04	1	0.09	0.23	0.28	0.01	0.02
PercentURLs	-0.04	-0.04	-0.02	0.19	0.02	-0.03	1	-0.04	0.19	-0.01	-0.06	-0.01	0
PercentQuestions	0.02	0.02	0.02	0.28	0	1	-0.03	0.06	0.1	0.04	0.09	-0.03	0.01
LikeRate	0.17	0.17	0.12	0.13	1	0	0.02	-0.07	-0.11	0.03	0.04	0.04	0.03
MeanWordCount	-0.05	-0.04	-0.02	1	0.13	0.28	0.19	0.01	0.12	0	0.04	0	-0.01
TotalPosts	0.91	0.91	1	-0.02	0.12	0.02	-0.02	0.09	-0.14	0.57	0.39	0.03	0.31
OutDegree	0.99	1	0.91	-0.04	0.17	0.02	-0.04	0.05	-0.22	0.44	0.4	0.03	0.23
InDegree	1	0.99	0.91	-0.05	0.17	0.02	-0.04	0.05	-0.2	0.45	0.41	0.03	0.25

Note: Maroon indicates high positive correlation, grey indicates high negative correlation.

Fig 5. Pearson Correlation Coefficient Heatmap

Strong Positive Correlation: As one variable increases, the other tends to increase.
Strong Negative Correlation: As one variable increases, the other tends to decrease.
Weaker Correlations: Positive or negative.

3. Unsupervised Learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning (ML) algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

Unsupervised learning's ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation and image recognition.[4]

3.1 **Elbow Plot:** In the plot shown below, the y-axis is labelled "Sum of Squared Distances", and it represents the within-cluster sum of squares (WCSS), which is a measure of the variance within each cluster.
Each point on the graph represents the WCSS for a specific number of clusters. Starting with one cluster, the WCSS is at its highest. As the number of clusters increases, the WCSS decreases because the data points are closer to the centroids of their respective clusters.

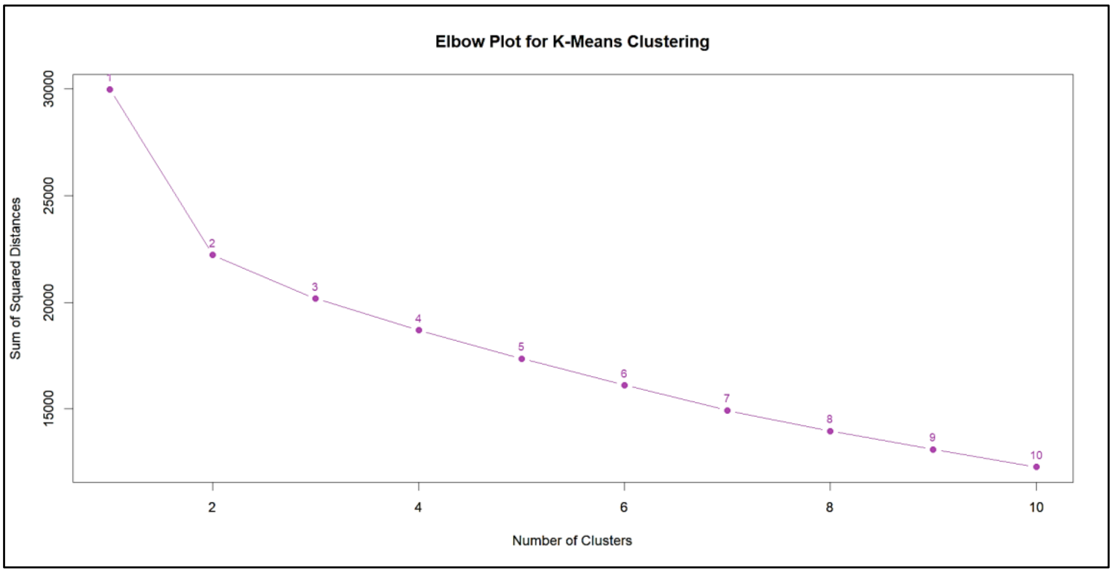


Fig 6. Elbow Plot for K-Means Clustering

The plot exhibits an "elbow" shape, marking the point where the rate of decrease notably changes, typically indicating the optimal number of clusters. This "elbow" signifies that adding more clusters beyond 2 may not significantly improve data modelling, suggesting 2 as the potential optimal number of clusters for the dataset.

3.2 **K-Means Clustering:** To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.[2]

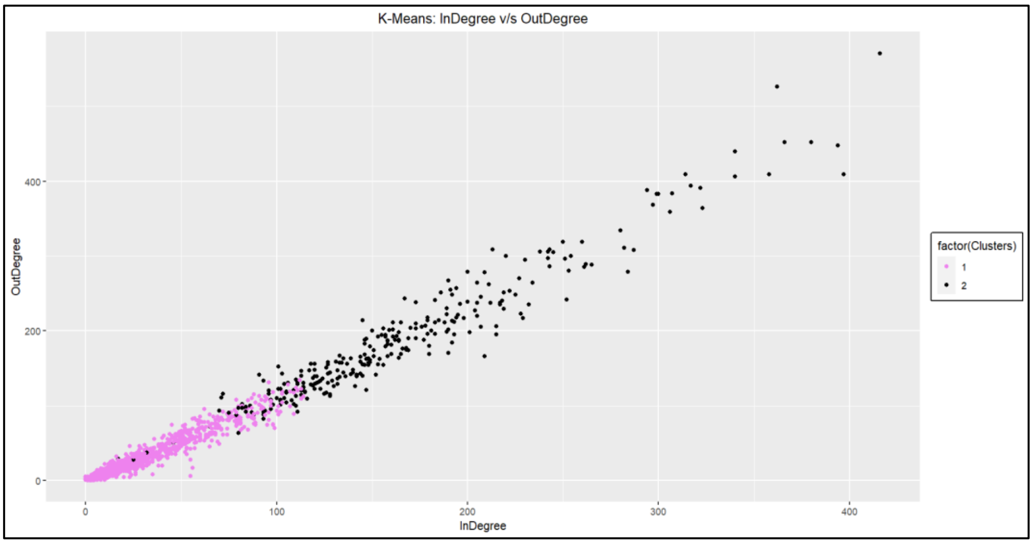


Fig 7. K-Means Clustering: InDegree v/s OutDegree

This scatter plot reveals insights into the network's structure and the relationship between user accounts' InDegree and OutDegree:

- **Clustering:** The data is clustered into two distinct groups on the basis of elbow, suggesting two distinct clusters fitting on the data.
- **Correlation:** There's a positive correlation between the number of interactions received (InDegree) and those initiated (OutDegree), especially in the higher-degree cluster.
- **Outliers:** Some user accounts deviate significantly from the cluster patterns, potentially indicating unique activity or influence within the network.
- **Dense Region:** The lower-degree cluster shows consistent patterns with balanced interaction levels.

3.3 Hierarchical Clustering: Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps:

- Identify the two clusters that are closest together, and
- Merge the two most similar clusters. This iterative process continues until all the clusters are merged.[1]

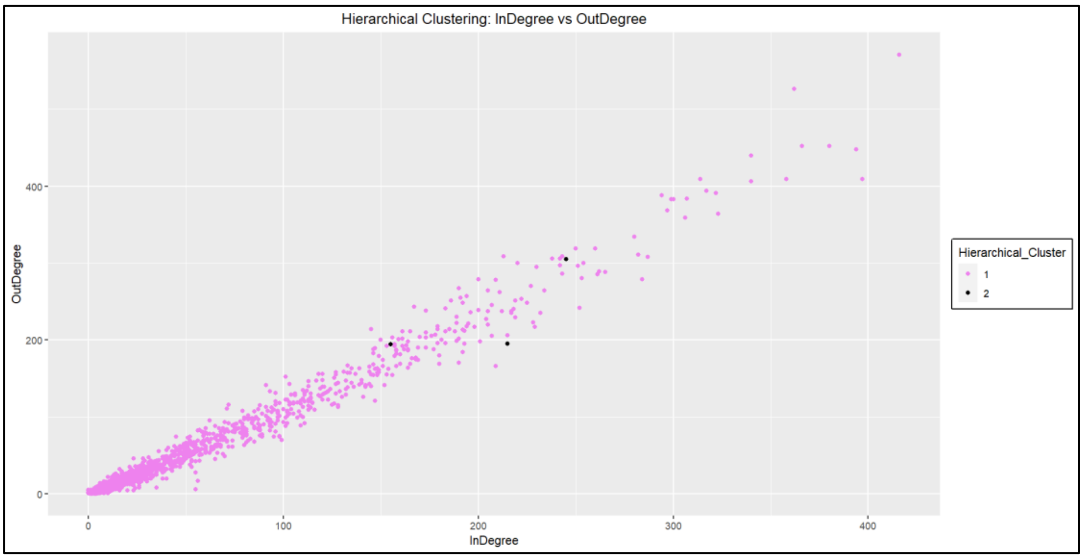


Fig 8. Hierarchical Clustering: InDegree v/s OutDegree

Most data points are coloured pink and form a dense cloud at the lower left corner, indicating many user accounts with low InDegree and OutDegree values. A few pink points are spread towards higher values of InDegree and OutDegree. The black points are fewer and are scattered among the pink points with generally higher InDegree and/or OutDegree values, suggesting these user accounts might have a different role or characteristic within the network compared to the majority pink cluster.

3.4 Cluster Analysis: The K-Means clustering plot (Fig. 7) exhibits a clear two-cluster arrangement of user accounts, indicating a distinct clustering. As a result, this grouping has been chosen for further analysis. Two distinct groups are visible in the scatter plot of InDegree versus OutDegree, which may indicate distinct patterns or attributes among user accounts. This selection is consistent with the results of the Elbow Plot, which suggested that two clusters would be the best option for this dataset. The analysis indicates a positive correlation between the number of interactions received (InDegree) and those initiated (OutDegree), especially evident in the higher-degree cluster. Furthermore, while most user accounts exhibit similar interaction levels (as depicted by the dense pink cluster), a minority of accounts, represented by scattered black points, demonstrate notably higher interaction values. These outliers could signify unique activity or influence within the network. Overall, the decision to employ K-Means clustering with two clusters is well-supported by the data's structure and distribution. This approach will provide valuable insights into user engagement patterns and facilitate further exploration of the dataset's characteristics and underlying dynamics.

3.5 **Centre of Variables for the Two Clusters:** The x-axis illustrates the value of various variables, with bars extending from a central axis at '0'.
Upon closer examination:
The variable "TotalPosts" demonstrates a higher central value in cluster '2' compared to cluster '1'.
"PercentURLs" exhibits a more central value in cluster '1' than in cluster '2'.
"PercentQuestions" showcases a higher central value in cluster '2'.
This trend continues across all variables, highlighting the differing central values between the two clusters.
These distinctions suggest that the clusters encapsulate distinct user behavioural patterns or characteristics. For instance, the higher value of "TotalPosts" in cluster '2' implies greater activity in posting content, whereas the elevated "PercentQuestions" in cluster '1' suggests a propensity for asking questions.

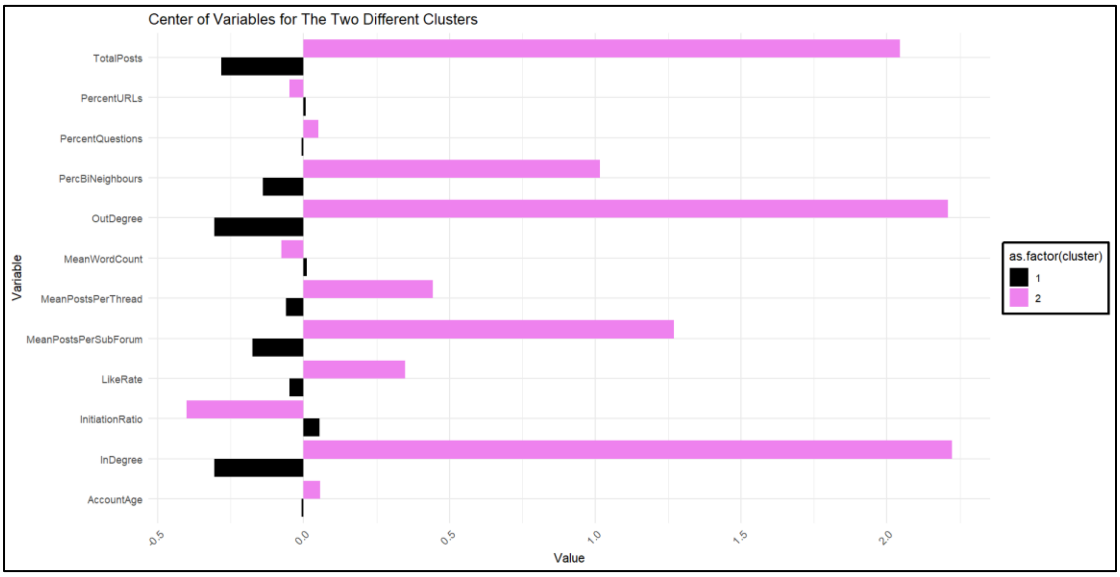


Fig 9: Centre of Variables for the Two Clusters

Cluster 2 comprises highly active users who engage widely by posting frequently, interacting with many others, and participating in various discussions. In contrast, Cluster 1 users, with older accounts, prioritize quantity, exhibiting a higher mean word count and potentially greater influence in initiating conversations and receiving likes.

3.6 Silhouette Analysis: The graph provided below is a plot of the silhouette score against the number of clusters for two different clustering methods: K-Means and Hierarchical clustering. The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

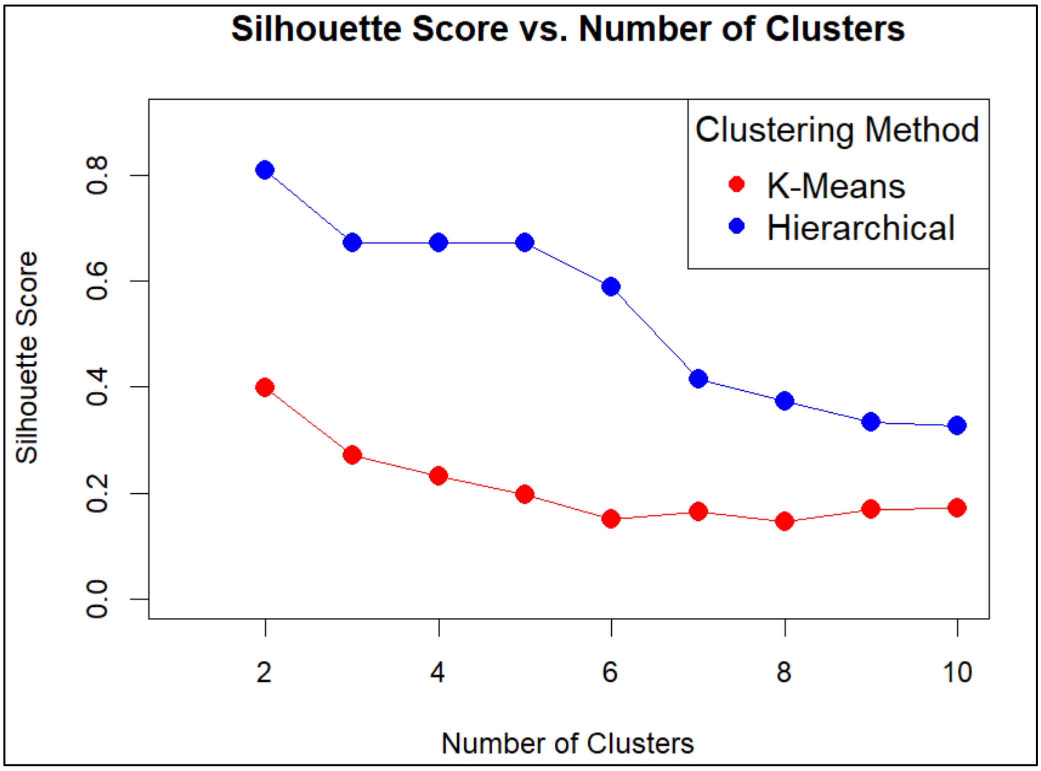


Fig 10: Silhouette Score vs Number of Clusters

From the graph, we can observe the following:
The silhouette scores for both K-Means (red line) and Hierarchical (blue line) clustering methods generally decrease as the number of clusters increases. This suggests that both methods find more cohesive and separated clusters when the number of clusters is smaller.

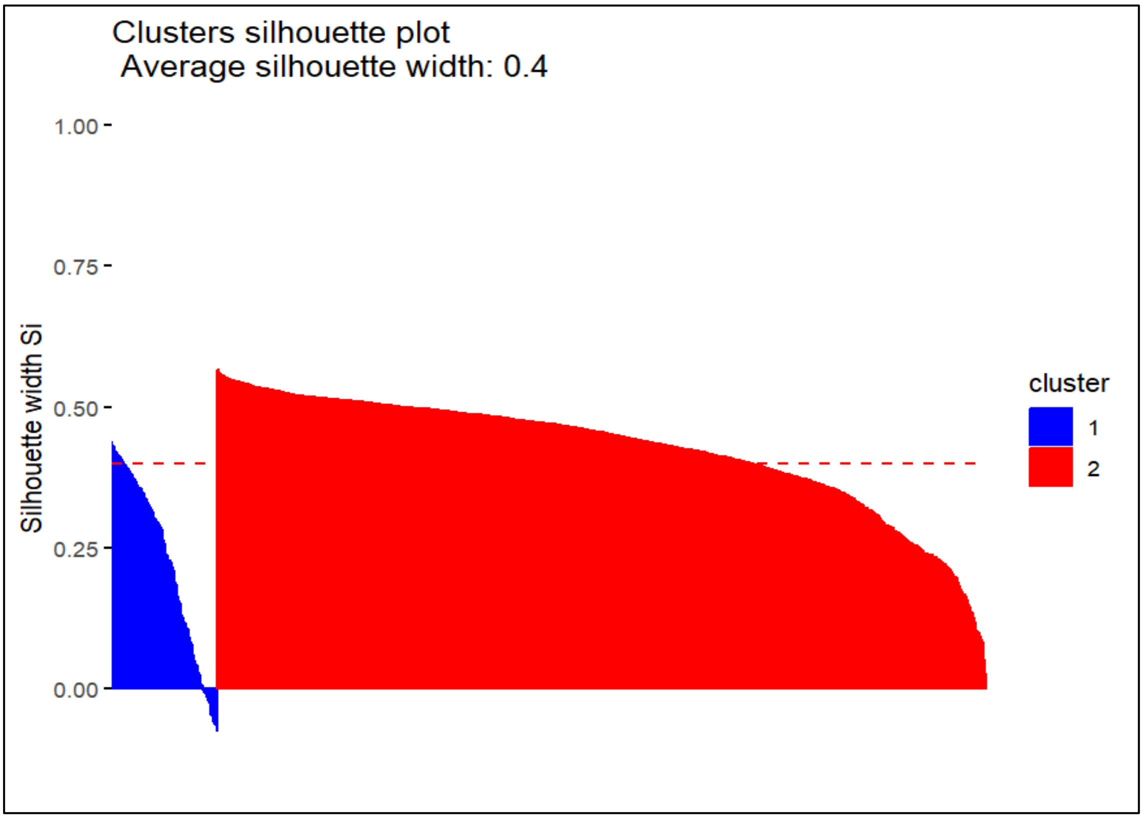


Fig 11. Cluster Silhouette Plot

Although Hierarchical clustering demonstrates superior performance in terms of silhouette scores, the decision to opt for K-Means over Hierarchical clustering is motivated by practical considerations. Despite its slightly lower silhouette scores, K-Means produces two distinctly separated clusters (as shown in Fig 7), aligning with the desired outcome for this specific analysis. Therefore, while Hierarchical clustering may offer theoretically better cluster separation, the clear distinction between two clusters achieved by K-Means makes it the preferred choice for this scenario.

4. Supervised Learning

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labelled data sets to train algorithms that to classify data or predict outcomes accurately.

As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. It can be used to build highly accurate machine learning models.[3]

Two models have been developed i.e. Random Forest (RF) and k-Nearest Neighbors (KNN). To choose the appropriate model among the two, their performance needs to be analysed based on the below metrics.

4.1 Random Forest:

- **Confusion Matrix:** The confusion matrix for Random Forest shows:

Prediction	Cluster	
	1	2
1	600	4
2	6	82

Table 1. Confusion Matrix for Random Forest

- 600 instances of class 1 were correctly classified.
- 82 instances of class 2 were correctly classified.
- There were 6 misclassifications of class 1 as class 2.
- There were 4 misclassifications of class 2 as class 1.

- **Accuracy:** The accuracy of Random Forest is **98.55%**.

4.2 k-Nearest Neighbors (KNN):

- **Confusion Matrix:** The confusion matrix for k-Nearest Neighbors shows:

Prediction	Cluster	
	1	2
1	597	5
2	9	81

Table 2. Confusion Matrix for k-Nearest Neighbors

- 597 instances of class 1 were correctly classified.
- 81 instances of class 2 were correctly classified.
- There were 9 misclassifications of class 1 as class 2.
- There were 5 misclassifications of class 2 as class 1.

- **Accuracy:** The accuracy of for k-Nearest Neighbors is **97.98%**.

4.3 Performance Metrics Comparison:

Metric	Random Forest	KNN
Precision	99.34%	98.51%
Recall	98.84%	98.51%
F1-Score	99.09%	98.51%

Table 3. Performance Metrics Comparison: Random Forest vs. KNN

4.4 Conclusion:

- Both Random Forest and k-Nearest Neighbors demonstrate high accuracy rates, with Random Forest slightly outperforming KNN.
- Random Forest exhibits a higher accuracy (98.55%) compared to KNN (97.98%).
- Considering the higher accuracy and the nature of the problem, where misclassifications can have significant implications, **Random Forest seems to be the preferred choice**.
- Both Random Forest and KNN models exhibit high precision, recall, and F1-score, indicating their effectiveness in classifying instances. Random Forest

slightly **outperforms** KNN across these metrics, but other factors such as model complexity and interpretability should be considered in model selection.

- Moreover, Random Forest provides robustness to overfitting, handles high-dimensional data well, and requires minimal tuning of hyperparameters, making it suitable for this classification task.

Therefore, based on the analysis of performance metrics, Random Forest is recommended as the model of choice for this classification problem.

4.5 Recommendations: Leveraging insights from dataset variables is essential to increase user engagement and draw in a larger user base. Analysing high-engagement content metrics like "LikeRate," "TotalPosts," and "MeanWordCount" can reveal user-resonant patterns. Using variables like "PercentQuestions" and "PercentURLs," one may create personalised content recommendations that are tailored to the user's tastes and past interactions, thus increasing user happiness. Putting tactics like focused onboarding procedures, community-building activities, and gamification components into practice encourages engagement and builds a feeling of community. Providing incentives for user contributions and recognising significant users based on metrics like "InDegree" and "OutDegree" can help to increase engagement. By regularly refining the platform's user experience, as guided by metrics like "TotalPosts" and "AccountAge," a smooth and delightful user journey is guaranteed. Expanding the user base is facilitated by customising marketing campaigns based on insights from factors such as "PercBiNeighbours". By defining precise KPIs for user engagement measurements, the client may proactively monitor results and adjust plans, leading to long-term success and improvement.

5. References:

1) Bock, T., 2018. What is Hierarchical Clustering? | Displayr.com [Online]. Displayr. Available from: https://www.displayr.com/what-is-hierarchical-clustering/.
2) Garbade, M., 2018. Understanding K-means Clustering in Machine Learning [Online]. Towards Data Science. Towards Data Science. Available from: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.
3) IBM, 2023a. What is supervised learning? [Online]. IBM. Available from: https://www.ibm.com/topics/supervised-learning.
4) IBM, 2023b. What is Unsupervised Learning? | IBM [Online]. www.ibm.com. Available from: https://www.ibm.com/topics/unsupervised-learning.