AM04 Workshop Report 2

Study Group 2 - Clemens Scherf, Edoardo Ferri, Hamish Thomas, Linli Ding, Yifan Yang

# 

# Introduction

In the world of data science, one of the important techniques used that serves a range of industries is clustering. By segmentation, we can spot the pattern hidden behind a large number of observations. Such analysis can also assist a company in designing marketing strategies and providing better user experience for its customers. In this report, the viewing information of BBC’s iPlayer in January is analysed using clustering tools to study the behaviour of customers across different segments. After cleaning, exploring and enriching the dataset, we used different clustering techniques (k-means, partitioning around medoids (PAM), H-clustering) accompanied by elbow plots and silhouette analysis to find out the optimum number of clusters and their characteristics. This allows us to gain a comprehensive view of BBC’s iPlayer customers and the content they watch, which can help us make predictions for watching in February.

# Methods

The dataset used in this report is extracted from the BBC iPlayer database. The dataset captures 10,000 random viewers’ behavioural information in January. The whole list of variables and their descriptions are stored in *Appendix I*. Since the variables in the raw dataset give little explanatory power, ICE (inspect, clean and explore) steps are conducted to make sure the dataset is valid for analysis, with the following steps:

1. Quantitative variables are added to enrich the dataset, i.e., number of shows and total time watched; proportionate usage based on weekday/weekends, time of the day, and genre.
2. “Average percentage of program viewed” is added. This might be useful because for some segments of customers, they are completely dedicated to a single show/genre. For example, sports program viewers. Therefore, this can potentially differentiate customers.
3. Log transformation is used for total\_time to mitigate the impact of data skewness.
4. Infrequent users are removed, since the data are not fully representative of these customers’ behaviour, thus we treat them as outliers which may potentially influence our analysis.

Then we performed clustering analysis for the cleaned data set following steps:

1. Use k-means for classification and check the centres and PCA.
2. Perform elbow analysis and silhouette analysis to determine the best range of values for k.
3. Try PAM and H-clustering for the chosen value of k, compare them with k-means, and choose the best clustering method based on their result.
4. Use subsampling to check the robustness of our chosen model.

# Results and Discussion

## **Data exploration**

The boxplots and histograms (Appendix II) show us that the distributions for both variables (noShows & total\_Time) are very positively skewed, which means most customers view few shows and spend little time on BBC iPlayer. The boxplots highlight some huge positive outliers, which may capture rare viewers who are extremely dedicated to the platform. The effect of these outliers on our results depends on the clustering method. For example, in k-means, these outliers **can** affect the results by shifting the cluster centres.

### **Correlation between variables**

Firstly, there is a perfect negative correlation between ‘weekend’ and ‘weekday’ (Appendix III), as the higher the percentage of a user's watchtime that occurs on the weekend, the lower the watchtime percentage will be during the weekdays (and vice versa). The same reasons can be applied to strong negative correlations between the different parameters for the time of day (‘Evening’, ‘Day’ & ‘Afternoon’). Additionally, there is a strong positive correlation between ‘noShows’ and ‘total\_Time’ - this is because the more different shows that a user watches, the longer we expect their total watch time to be. This however is not a perfect correlation, as there will be people who switch between shows very quickly, trying to find one they really like. For simplicity, we are only using the “weekend” and the “total\_time” variable. We can also see some correlations between different show genres - for example there are notable negative correlations between genres like 'Sport' and 'Drama', as well as 'Factual' and 'Drama'. We assume that, as people who like watching drama are less likely to watch sports and documentaries.

### **Plausible range of “k”**

|  |  |
| --- | --- |
| Figure 1 Elbow plot | Figure 2 Silhouette analysis |

For the elbow plot, SSE did not sterilise within k = 10, thus we will look for points in the middle to ensure that we can have a relatively low SSE while not overfitting the data. In Silhouette analysis, we will be looking for points with high silhouette width but still keep in mind the problem of overfitting. Overall, the best possible values are located for k from 4 to 7.

We can see from the centre graph where k=2 that these clusters have a significant difference in their preferred genres. The day-viewers (more likely to be the kids and the retired) like "Children", "Factual", "Learning", "Music", "News" and "Sports" contents, while the evening-viewers (more likely to be employed adults) prefer "Comedy" and "Drama". The latter are more likely to complete the show and more dedicated to BBC iPlayer (higher total time).

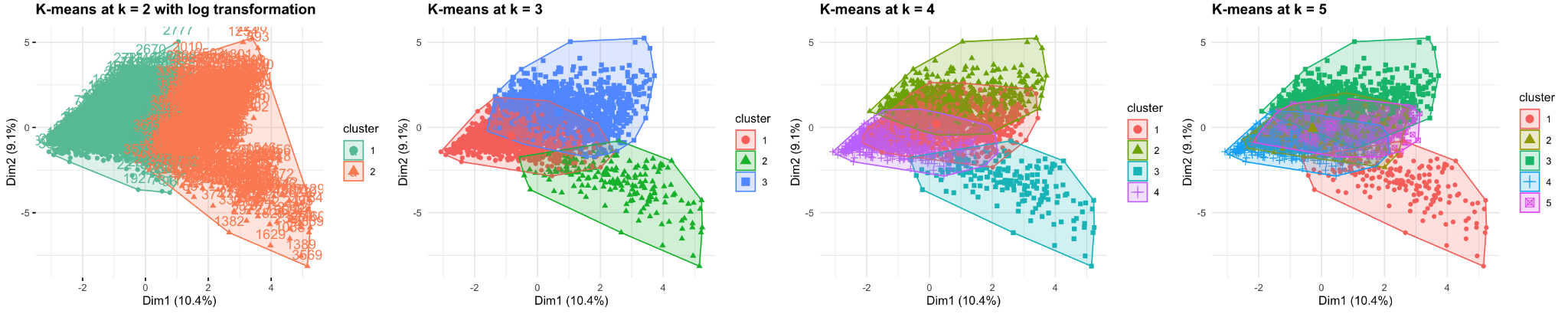


Figure 3 PCA plots for k ranged from 2 to 5

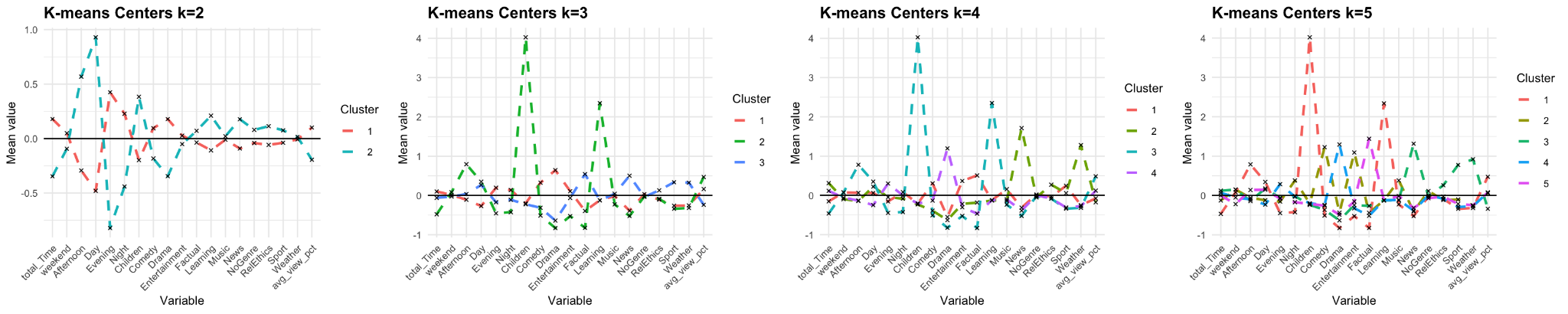


Figure 4 Centre plots for k ranged from 2 to 5

The above PCA plots show that increasing the number of clusters increases the overlap between clusters. When k = 3, the range of each cluster is so big that some points are not strongly related to the others within the same cluster. There is still room to subset the clusters. And according to the silhouette method, the distance between clusters is not as high as that in k=2 or k=4. When k=5, cluster 2 and 5 largely overlap (thus can be very similar) and it shows a sign of overfitting. Therefore, k=4 seems most reasonable among the four plots. Although there is still overlap present, the clusters are (relatively) more independent and the sizes of the clusters is more reasonable.

### Chart, histogram Description automatically generated**Comparing k-means, PAM and H-clust to see which is optimal**

Figure 6 Centre plots at k = 4 for PAM, k-means and H-clustering

Figure 5 PCA plots at k = 4 for PAM, k-means and H-clustering

From the cluster centers plots above we can see that k-means method is better at differentiating customers who watch BBC iPlayer by time of the day than by different genres. The other two methods, PAM and H-clust are better at differentiating customers by the different genres of shows they watch. For example, the spikes for “Entertainment” and “Sport” in PAM clearly show two distinct customer segments. However, it fails to accurately identify the “children” cluster. H-clust seems to have blurred a few segments together in the blue cluster, since it has 5 spikes of genre in total. It is reasonable to assume that one segment only views 1-2 genres, 5 genres are too many. Therefore, the optimal method is k-means since it relatively accurately captures most of the different characteristics to differentiate between customer segments.

### **Interpretation of clusters**

Following our extensive quantitative analysis using various algorithms, we can now identify 4 distinct customer clusters. The first cluster involves viewers who consume comedy and drama genres during the evenings, thus most likely being working professionals. The second cluster we identified, revolves around daytime viewers of sports and news. The third cluster involves afternoon consumption of factual and children shows, so we can conclude that this cluster revolves around children coming home from school and turning on their TV. Our fourth and last cluster involves night time viewers of entertainment and sports, most likely adults trying to unwind from a long day at work. With this information, we can now put together a better selection of shows for each customer segment at different times of the day, and thereby more successfully fulfill BBCs mission of providing entertainment, education and information to the public.

# Conclusion

Our chosen k value is k=4 - this has been selected based on the combination of results from the elbow chart and the silhouette method. A value of k=4 gives us the most consistently good result in our analysis and also ensures that we are **not** over- or under-fitting our data. Notably, a k value of 4 gives us clusters that have considerably less overlap (and therefore greater differentiation) than k=5 - and therefore k=4 gives us greater explanatory power and interpretability. Overall, the best method is the k-means method (with k=4). This is because the k-means method accurately defines most clusters - the PAM method for example, fails to identify the ‘children’ cluster.

However, our conclusions and results are sensitive to the following assumptions:

* Each viewer only belongs to one single cluster. However, people’s behaviour is complex, could possess features of multiple clusters, and may change over time.
* The variables in this report are the only crucial parameters for clustering. For example at high k, we may observe very similar clusters and decide that this is not an optimal representation of the population. However, in real life there may be other attributes that differentiate these two clusters, which means that in fact they are not overfitting the data.

It is important to understand that there may be further variables that can heavily impact the results. We can check the robustness of our results by analysing subsamples. If we split the data into a training and testing dataset, we can perform clustering analysis on both sets and see if the characteristics of clusters are similar.

Finally, the information from this cluster analysis can be used to help us build a recommendation system for iPlayer (and/or other streaming services). By using our clusters and our conclusions relating to user preferences we can:

* Recommend general materials and genres (sport, drama, etc.) to people within the appropriate clusters, creating better recommendations and improving user experience.
* Target recommendations and push notifications during times when we expect users (within given clusters) to be active.
* Make future predictions about viewership of certain genres which allow for better tailored services - produce new drama shows if there is a considerable upward trend in the number of people who are watching drama (within the drama cluster).

# Appendix

## **Appendix I**

The table below shows the whole list of variables and their description used in the clustering methods.

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| user\_id | a unique identifier for the viewer |
| program\_id and series\_id | these identify the program and the series that the program belongs to |
| genre | the programme’s genre (e.g., drama, factual, news, sport, comedy, etc) |
| start\_date\_time | the streaming start date/time of the event |
| Streaming id | a unique identifier per streaming event |
| prog\_duration\_min | the program duration in minutes |
| time\_viewed\_min | how long the customer watched the program in minutes |
| duration\_more\_30s | equals 1 if the program duration is more than 30 seconds, equals 0 otherwise |
| time\_viewed\_more\_5s | equals 1 if time\_viewed is more than 5 seconds, equals 0 otherwise |
| percentage\_program\_viewed | percentage of the program viewed |
| watched\_more\_60\_percent | equals 1 if more than 60% of the program is watched, equals 0 otherwise |
| month, day, hour, weekend | timing of the viewing |
| time\_of\_day | equals “Night” if the viewing occurs between 22 and 6am, "Day" if it occurs between 6AM and 14, “Afternoon” if the it occurs between 14 and 17, “Evening” otherwise |

## 

## **Appendix II**

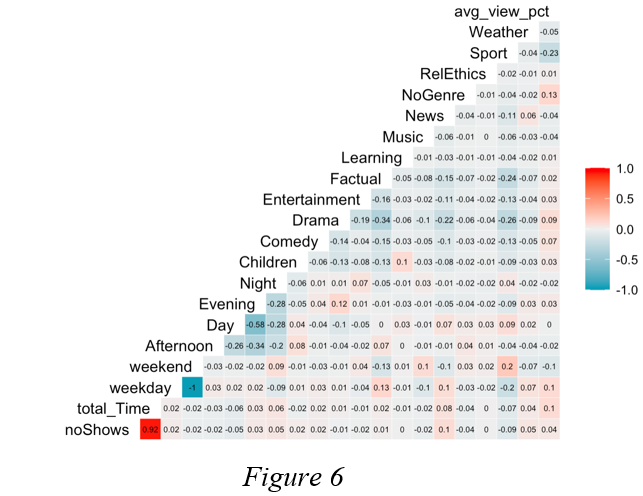
Boxplots and histograms for no\_of\_shows and total\_time

|  |  |
| --- | --- |
|  |  |
|  |  |

## 

## **Appendix III**

Correlation matrix for the variables



## 

## **Appendix IV**

PCA plots for k-means at k = 2 when using and not using log transformation for total\_time

## 

|  |  |
| --- | --- |
|  |  |