Projet Shiny - Profilage des épisodes de GOT

Abstract—L'objectif de ce projet est de réaliser une application avec la bibliothèque $Shiny$ sur $R\ Studio$ à travers les données de $Game\ Of\ Thrones$. $Game\ Of\ Thrones$ est une série américaine contenant 73 épisodes en 8 saisons. Nous avons utilisé une base de données contenant l'indice de l'épisode (numéro d'épisode-numéro de saison), son titre, un indice de popularité, un indice de noblesse et le nombre total de mort dans chaque épisode.

1 Modélisation de la base des données

Nous allons détailler la partie Data engineering qui va nous permettre le profilage des épisodes de GOT. En effet, on va combiner les bases de données des TP GOT et on va s'appuyer sur une notre base trouvée sur la plateforme $Kaggle$ qui contient pour chaque personnage sa popularité (entre 0 et 1) et sa noblesse (1 si le personnage est noble et 0 sinon). Après la jointure des bases de données et les calculs d'agrégations on a pu modéliser les critères de profilages suivants :

- **Indice de popularité :**c'est un indice propre à chaque épisode et on l'obtient en calculant la moyenne pondérée des durée de présence de chaque personne par leur indice de popularité c.a.d; pour un épisode donnée N, soit $t_{i,N}$ la durée de présence du personnage i et p_i sa popularité alors :

$$Popularity - Index_N = (\sum_i t_{i,N} * p_i) / \sum_i t_{i,N} \quad (1)$$

- **Indice de noblesse :**C'est un indice propre à chaque épisode et c'est exactement la somme des personnages nobles présents dans l'épisode.
- **Indice de mort :**c'est la somme des personnages morts lors de l'épisode.

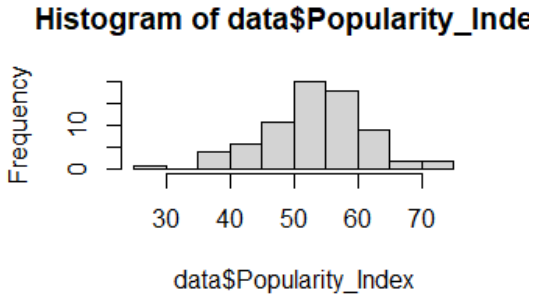

Fig. 1: ***Histogramme de l'indice de popularité***

Les indices ont été transformé pour avoir une échelle entre 0 et 100 pour mieux visualiser les histogramme (les transformations sur l'échelle des indices ne change pas la forme de la distribution). En effet, les indices suivent tous une loi normale bien que l'indice de mort soit un peu faussé à droite mais la distribution des trois indices reste généralement très satisfaisante.

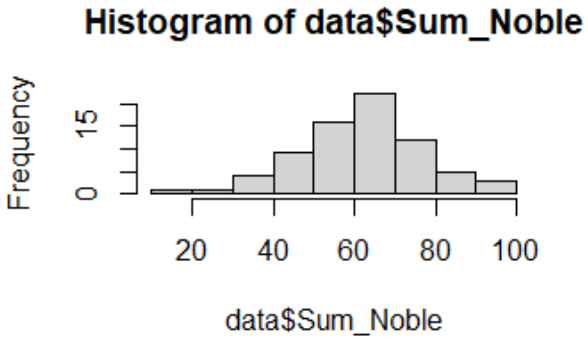

Fig. 2: ***Histogramme de l'indice de noblesse***

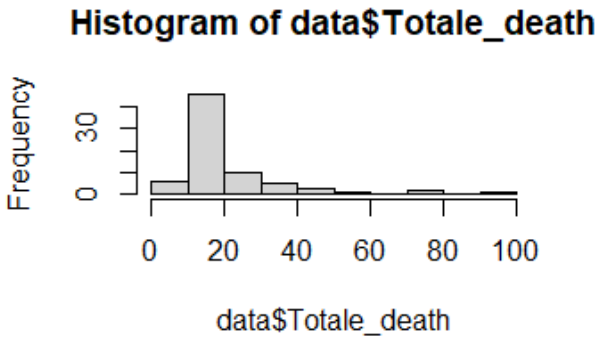

Fig. 3: ***Histogramme de l'indice de mort***

On va ensuite effectuer des transformations standard sur chaque indice afin de mettre toutes les variables dans la même échelle, cette transformation est la plus adéquate pour l'algorithme $k - means$. Par conséquent on retranche la moyenne et on divise sur l'écart-type sur chaque variable du problème.

2 CLUSTERING DES ÉPISODES

On va utiliser l'algorithme $k-means$ pour le profilage des épisodes, et on va se baser sur la méthode $Elbow$ et la moyenne du score $Silhouette$ principalement pour déterminer le nombre de profil le plus adéquat à notre problème.

Les deux méthodes donnent les résultats suivants :

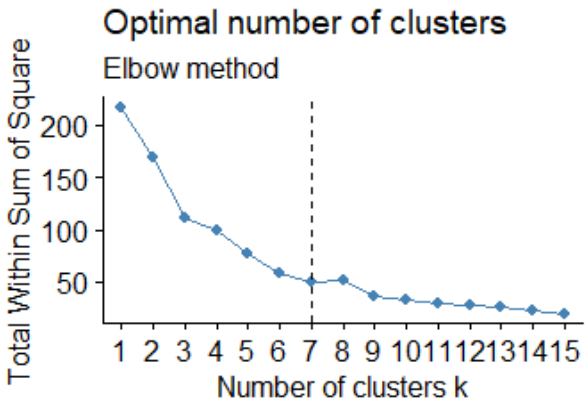

Fig. 4: ***Méthode*** *Elbow* ***pour le choix de K***

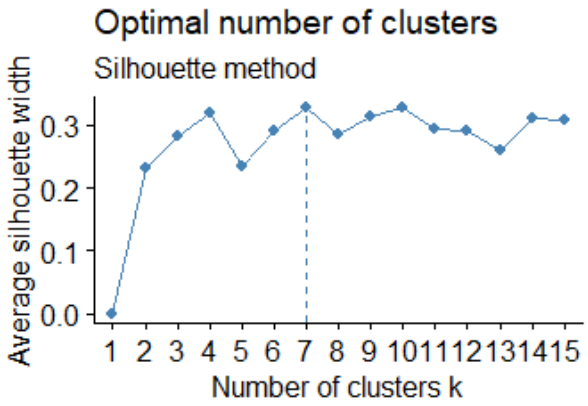

Fig. 5: ***Méthode*** *Silhouette* ***pour le choix de K***

A présent on sait le nombre de profil que nous allons choisir, nous allons voir rapidement quelques statistiques de chaque profil avec un outil dashboard $Power\ Bi$.

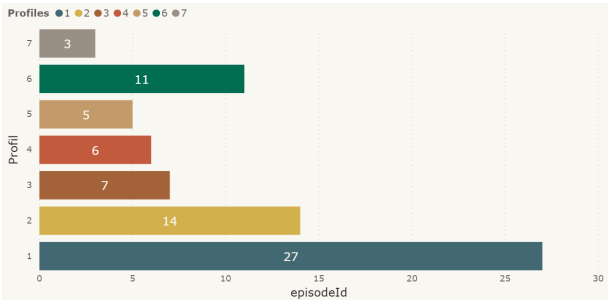

Fig. 6: ***Population de chaque Profil***

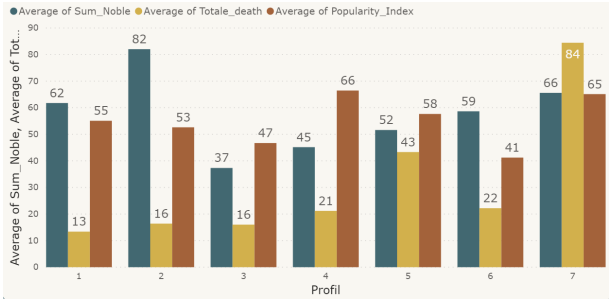

Fig. 7: ***Moyenne des Indices par Profil***

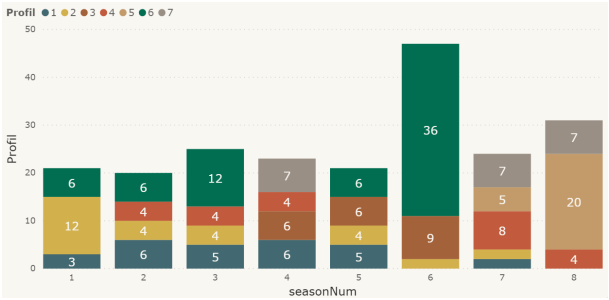

Fig. 8: ***Répartition des profils selon les saisons***

On voit bien que les indices sont répartis différemment selon le profil ce qui permet de conclure que le profilage est satisfaisant vu le manque des critères et des bases de données. Ensuite, en se basant sur la documentation de GOT, la saison 4, 7 et 8 ce sont des saisons où on a beaucoup de guerres et si on fait la liaison avec le profilage on voit bien que le profil 7 n'est présent que dans les saisons 4,7 et 8. En effet, ce profil est caractérisé par un nombre de mort très élevé par rapport aux autres profils ce qui justifie bien la méthodologie utilisée.

Nous avons décidé de donner un surnom à chaque profil selon ses caractéristiques :

- **Profil 1:** Les nobles populaires
- **Profil 2:** Les celebrites de GOT
- **Profil 3:** On peut tous mourir
- **Profil 4:** Let's survive
- **Profil 5:** Les vedettes des chateaux
- **Profil 6:** Richesse,Popularite Vs LA MORT
- **Profil 7:** Le deuil

3 CONCLUSION

On va exploiter les résultats de ce profilage pour construire une SIMPLE application avec $shiny$ qui permet de sélectionner un épisode puis elle indique les épisodes les plus similaires selon les critères que nous avons utilisés.