

Projet Analyse de données exploratoire
Etude de la base de données *Onlinenewspopularity*

Léa BRASSEUR & Benjamin CHAUDET

Table des matières

1	Introduction	2
1.1	Présentation	2
1.2	Modifications	2
2	Statistiques descriptives à une variable	3
2.1	Nombre de partages	3
2.2	Nombre de liens	4
2.3	Nombre d'images	4
2.4	Nombre de vidéos	5
2.5	Jours de publication	5
2.6	Thèmes	6
3	Statistiques descriptives à deux variables	7
3.1	Jours de publication	7
3.2	Thèmes	7
3.3	Articles populaires vs impopulaires	8
4	Corrélation	11
5	Conclusion	11

1 Introduction

1.1 Présentation

Dans un premier temps, nous pouvons voir que la base de données porte sur les articles publiés sur le site **Mashable**. Elle contient **39644** articles et a **61** variables d'études. Dans les **61** variables nous pouvons constater qu'il y a seulement **1** variable qualitative et **60** variables quantitatives. Les variables quantitatives sont composées de **27** variables codées en binaire (entre 0 et 1 ou -1 et 0).

Après vérification, nous avons pu constater qu'aucunes données n'étaient manquantes dans la base de données.

Nous avons décidé d'étudier certaines caractéristiques des articles telles que le nombre de liens, d'images ou de vidéos. Notre but par la suite a été de relier toutes ces études à la variable **shares** représentant le nombre de partages de chaque article. Nous voulions voir ce qui pouvait inciter les lecteurs à partager un article.

1.2 Modifications

Nous avons fait quelques modifications sur la base de données. Nous avons retiré les variables **url** et **timedelta** qui sont non-prédictives.

Puis, nous avons regardé la répartition du nombre de mots dans les articles. Nous avons remarqué qu'il y avait des observations avec aucun mot et nous avons décidé de les supprimer de la base de donnée car non pertinentes.

TABLE 1 – Nombre de mots

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	246	409	546.5147	716	8474

Nous allons ensuite réunir les variables portant sur les jours pour en avoir une seule au lieu de 7. En effet, les 7 variables concernant les jours étaient codées en binaire nous les avons réunis dans le but d'avoir une seule variable qualitative. De même pour les thèmes des articles, passant de 6 variables à une seule qualitative. Pour la variable weekend nous l'avons modifié pour passer d'une variable binaire à une variable qualitative.

Nous avons créé une nouvelle base de données en enlevant les variables binaires vu précédemment et en les remplaçant par les variables qualitatives **weekday** et **channel** que nous venons de modifier. Voici un exemple des nouvelles variables :

TABLE 2 – Sommaire

shares	is_weekend	channel	weekday
593	Semaine	Entertainment	Lundi
711	Semaine	Business	Lundi
1500	Semaine	Business	Lundi
1200	Semaine	Entertainment	Lundi
505	Semaine	Tech	Lundi
855	Semaine	Tech	Lundi

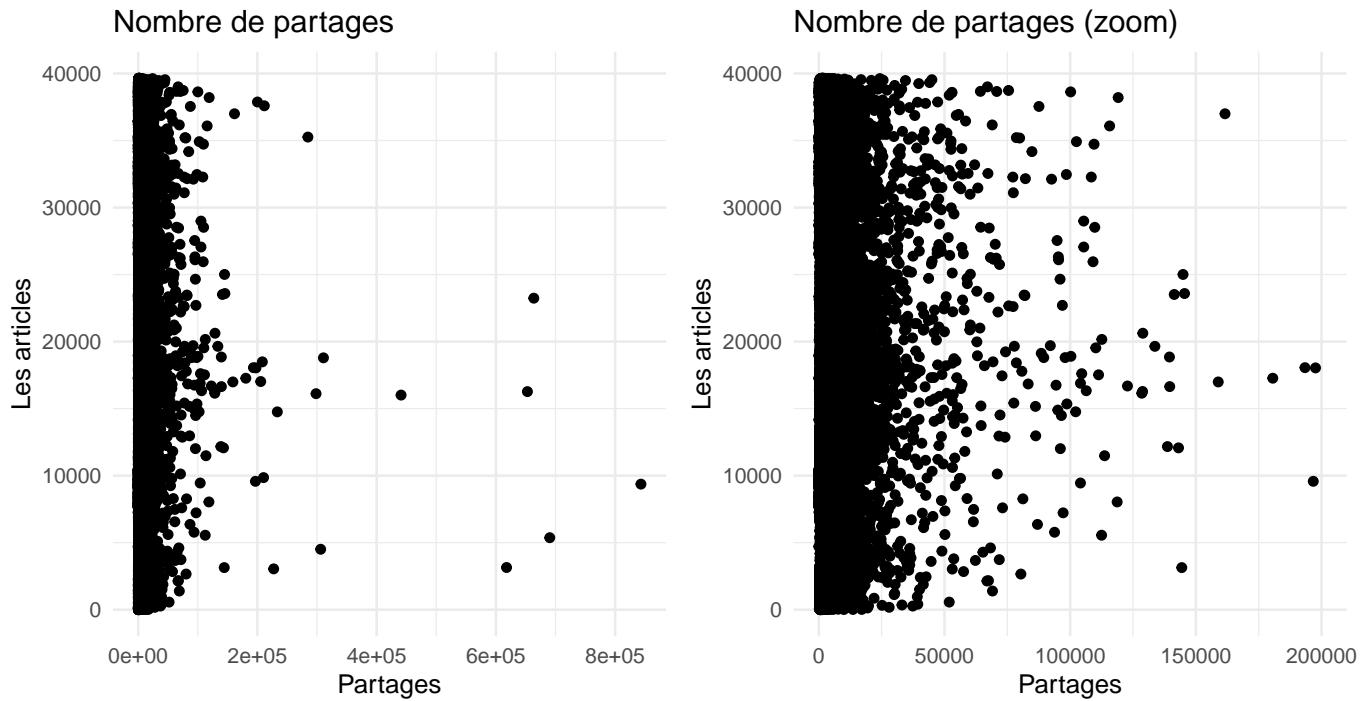
2 Statistiques descriptives à une variable

2.1 Nombre de partages

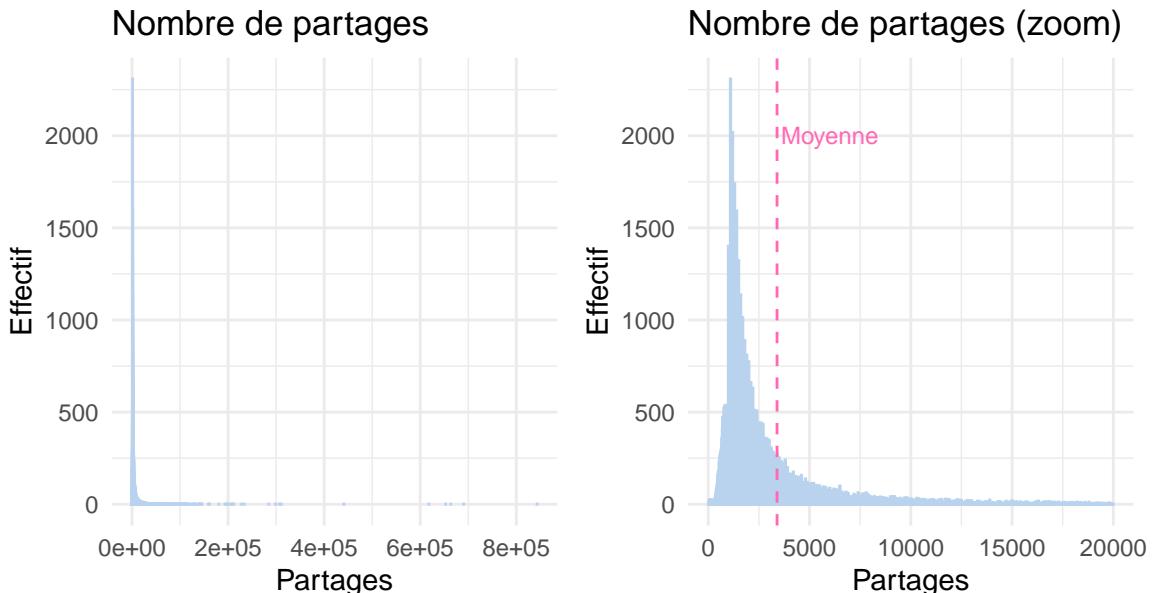
Nous avons analysé la variable portant sur le partage des différents articles.

TABLE 3 – Nombre de partages

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	946	1400	3395.38	2800	843300



La plupart des partages sont entre 1 et 100 000. Il y en a très peu au-delà de 200 000. Nous pouvons voir que 5 articles sortent du lot et ont été beaucoup plus partagés que les autres (plus de 600 000 partages).



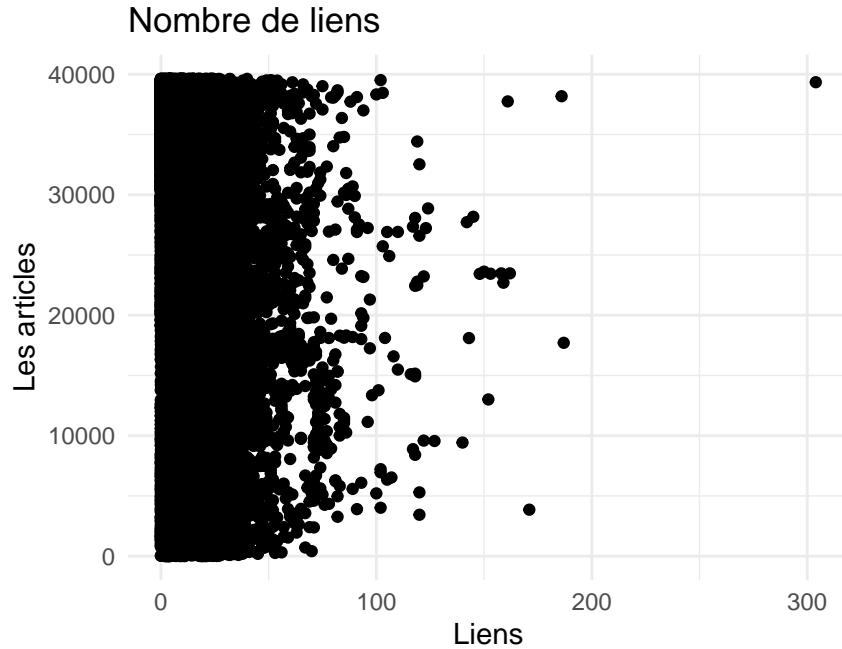
Les articles de la base de données sont partagés en moyenne **3395** fois. Mais la moitié des articles sont partagés moins de **1400** fois. La différence entre la moyenne et la médiane est due au fait que les valeurs extrêmes (les articles vu précédemment, ayant énormément de partages) augmentent fortement la moyenne.

2.2 Nombre de liens

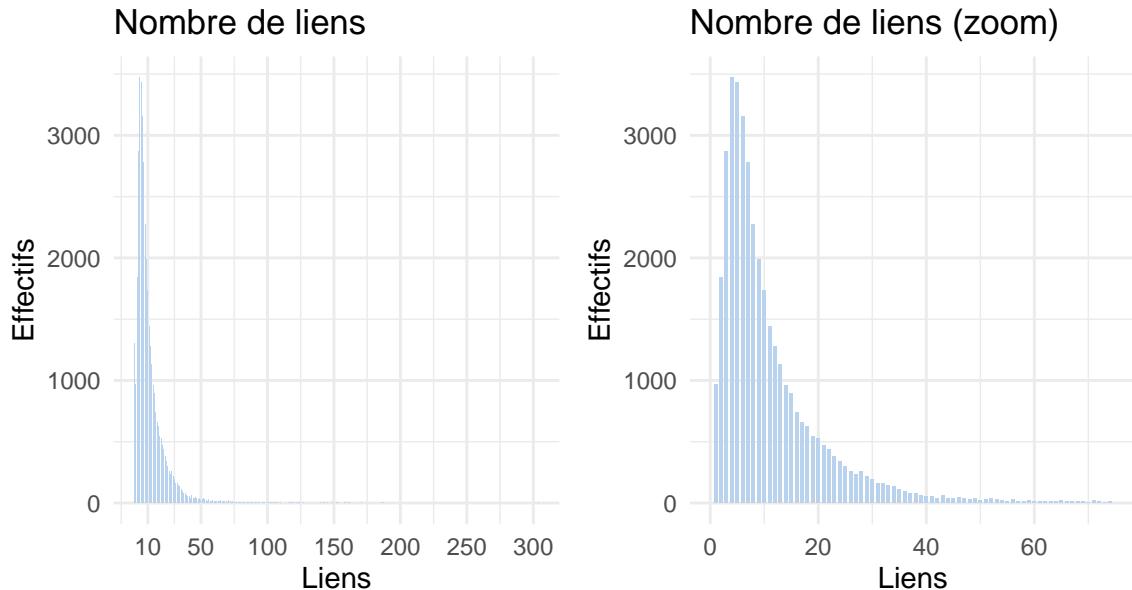
Puis nous avons regardés le nombre de liens dans les articles.

TABLE 4 – Nombre de liens

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	4	8	10.88369	14	304



Nous voyons qu'il y a peu d'articles avec plus de 100 liens. Et que la plupart ont entre 0 et 50 liens.



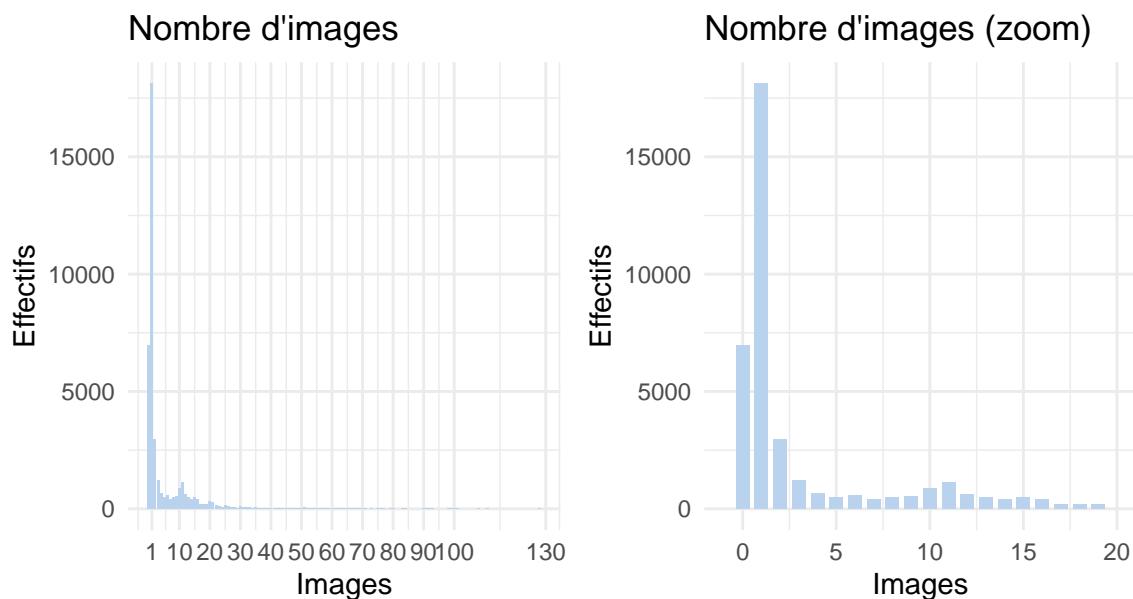
Grâce à ces graphiques et au sommaire nous voyons qu'en moyenne les articles ont entre 0 et 10 liens.

2.3 Nombre d'images

Nous avons ensuite étudié les images présentes dans les articles.

TABLE 5 – Nombre d'images

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1	1	4.544143	4	128



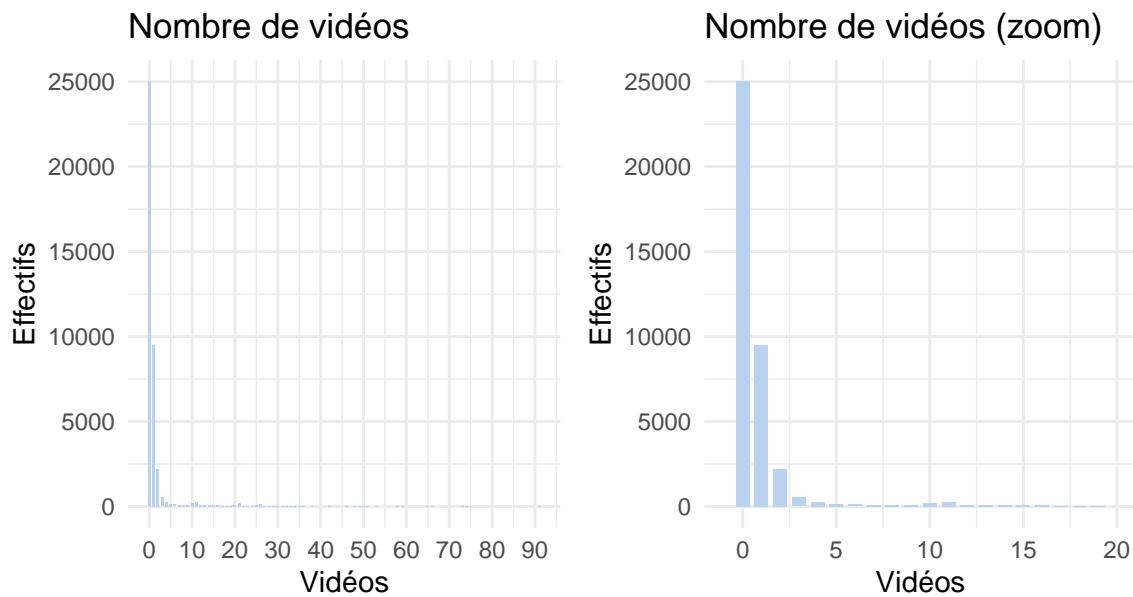
La grande majorité des articles n'ont qu'une seule image.

2.4 Nombre de vidéos

Puis, nous avons observé les vidéos.

TABLE 6 – Nombre de vidéos

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	1.249874	1	91

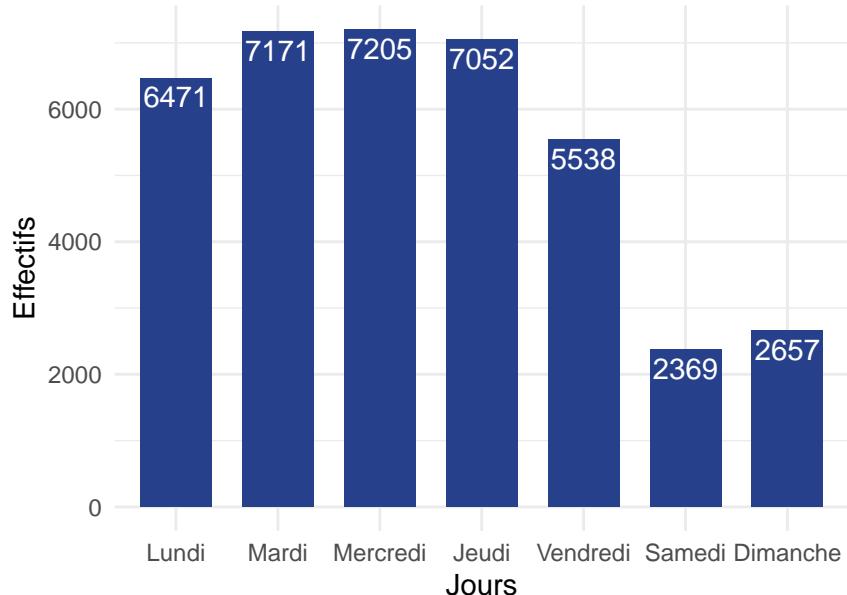


La majorité des articles n'ont pas de vidéos.

2.5 Jours de publication

Grâce à la nouvelle base de données nous pouvons comparer les jours de publication des articles.

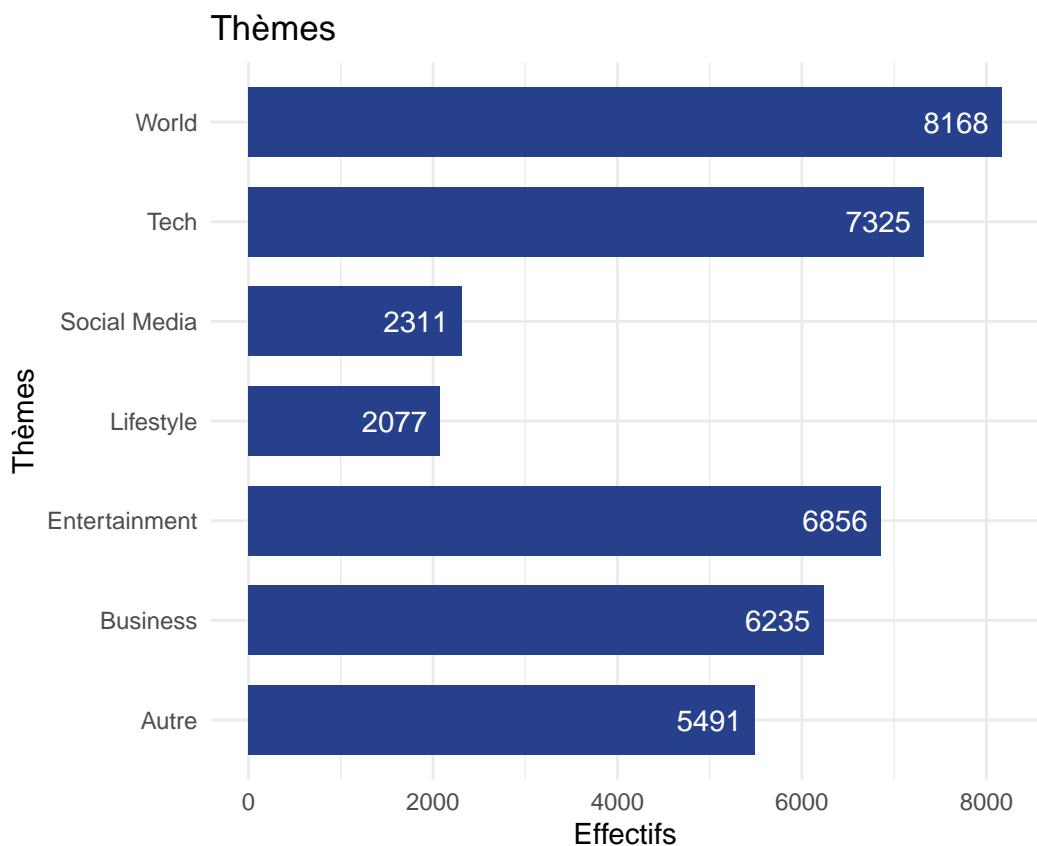
Jours de publication



Nous pouvons voir que les articles sont le plus souvent publiés entre le *lundi* et le *jeudi*. Nous constatons une légère baisse le *vendredi* au niveau des publications. Baisse qui se poursuit le week-end.

2.6 Thèmes

Nous pouvons également regarder les thèmes des articles.



Peu d'articles sur les thèmes *lifestyle* et *social media* sont publiés par rapport aux autres thèmes qui sont à peu près équivalents. Le thème le plus présent est *world*.

3 Statistiques descriptives à deux variables

Dans cette partie nous allons relier notre variable principale **shares** aux autres variables.

3.1 Jours de publication

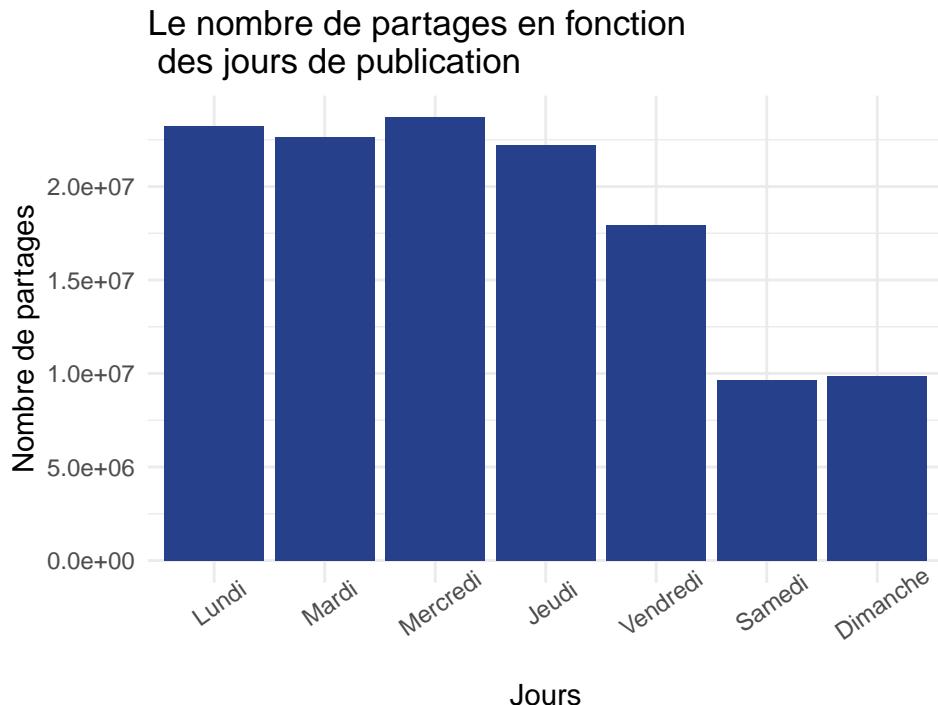


TABLE 7 – Jours et partages

	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Jours	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Moyenne des partages	3589	3155	3286	3144	3233	4055	3707
Effectif total par jour	6471	7171	7205	7052	5538	2369	2657

Nous pouvons constater grâce au tableau qu'en moyenne les articles les plus partagés sont publiés le *samedi* et le *dimanche*.

La moyenne des partages est **3355**, nous constatons donc que la moyenne pour *samedi* et *dimanche* est au-dessus.

3.2 Thèmes

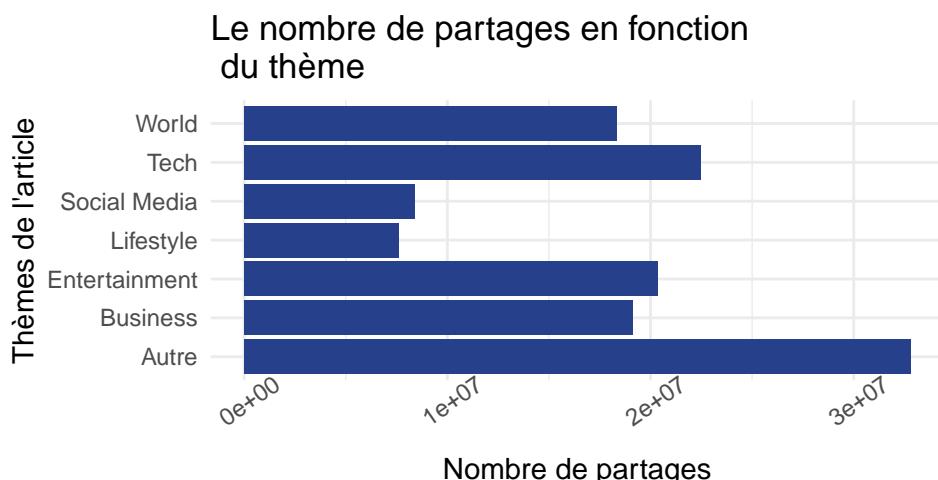


TABLE 8 – Thèmes et partages

	Lifestyle	Entertainment	Business	Social Media	Tech	World	Autre
Moyenne des partages	3657	2971	3065	3632	3069	2244	5970
Effectif total par thème	2077	6856	6235	2311	7325	8168	5491

Le tableau nous permet de voir qu'en moyenne les catégories d'articles les plus partagés, en dehors de la catégorie autre, sont *lifestyle* et *social media*. Alors que leurs effectifs sont minoritaires dans la base de données. Au contraire, les articles sur le thème world sont plus nombreux mais moins partagés.

La moyenne des partages toutes catégories confondues est **3355**, nous constatons donc que la moyenne pour lifestyle et social media est au-dessus, contrairement aux autres moyennes qui sont en-dessous.

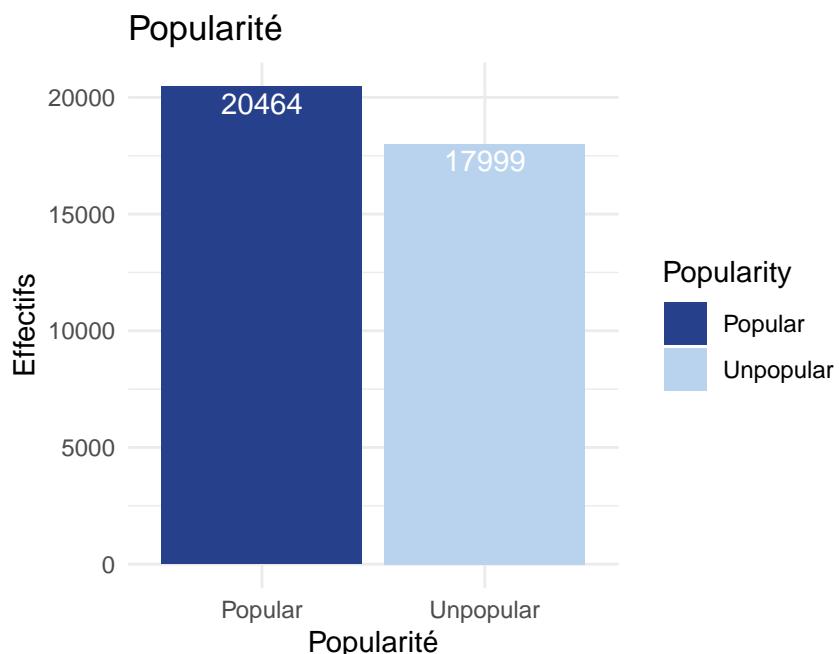
3.3 Articles populaires vs impopulaires

TABLE 9 – Sommaire

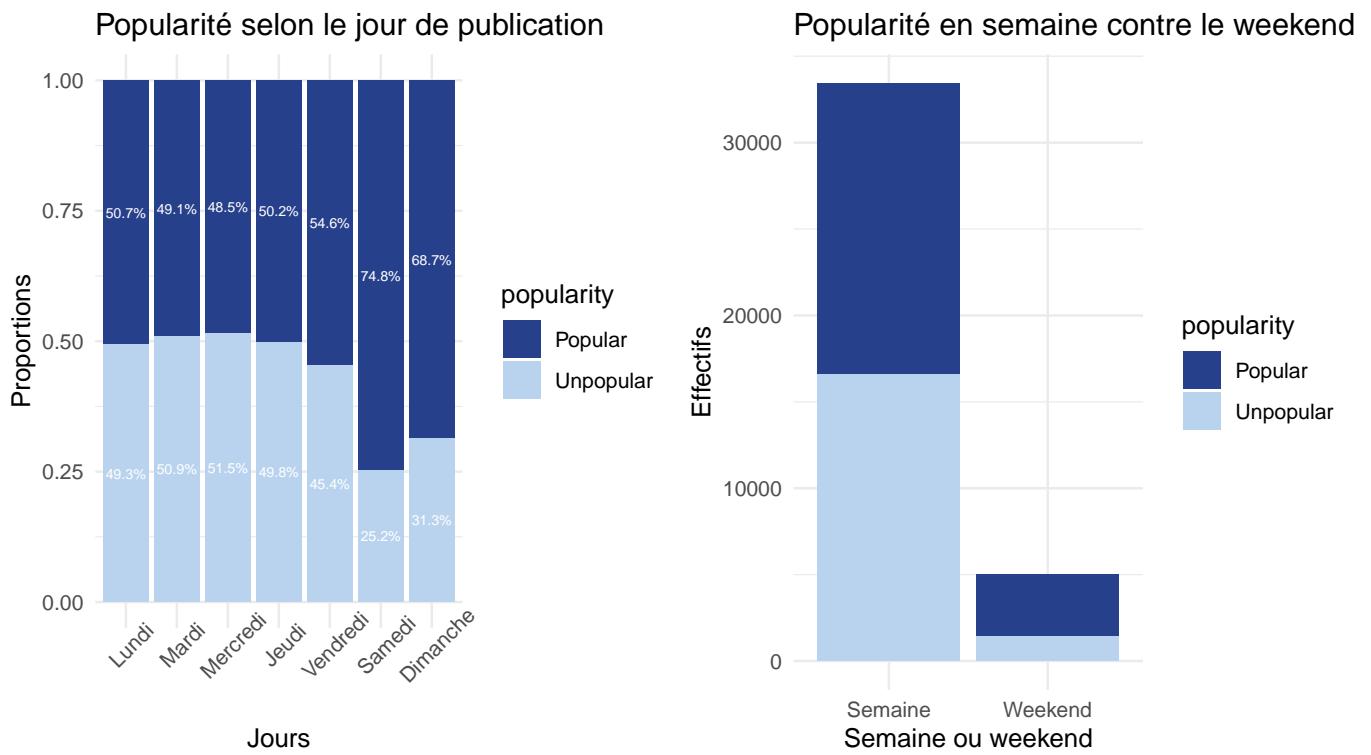
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	945	1400	3355.36	2700	843300

La médiane des partages est de 1400.

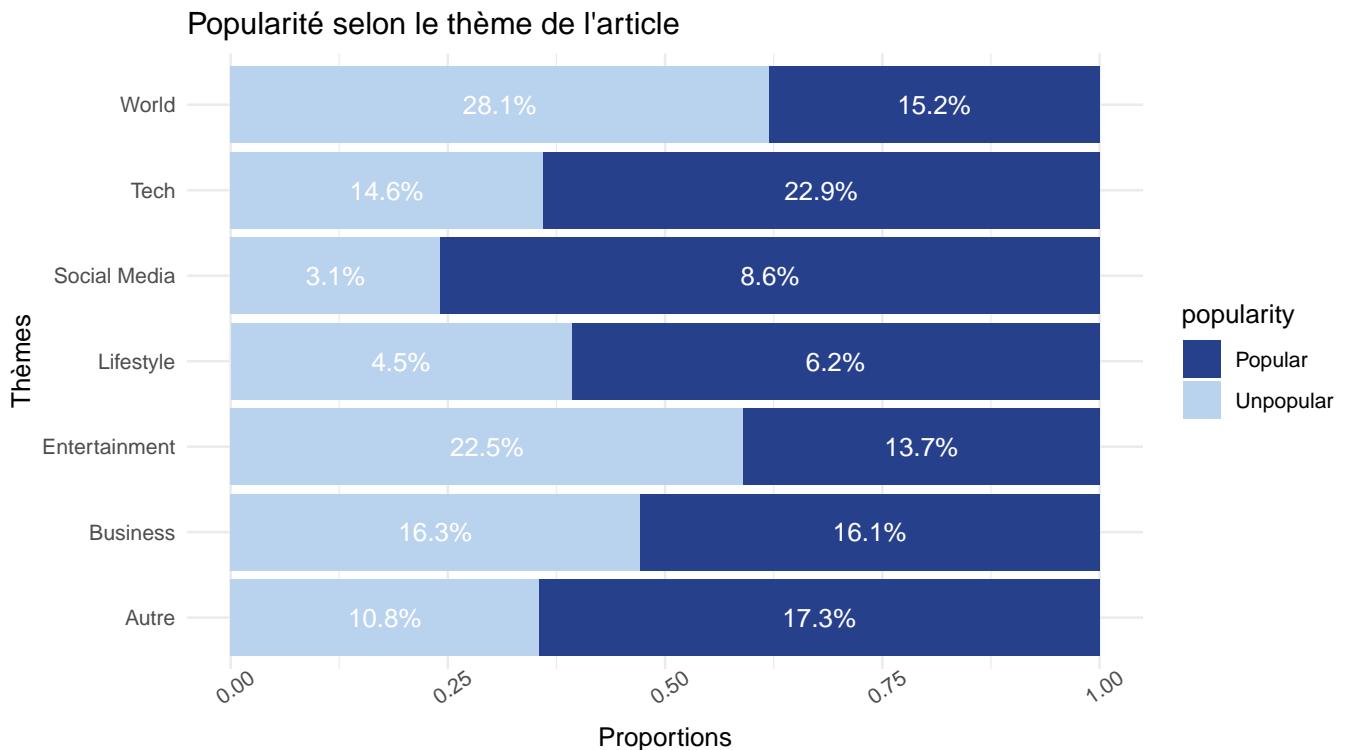
Nous allons créer une variable **popularity** qui prend pour valeur *popular* si l'article est partagé 1400 fois ou plus (le nombre médian de partages) et pour valeur *unpopular* si l'article est partagé moins de 1400 fois.



Comme dans la description nous trouvons 20464 articles populaires et 17999 articles impopulaires.



- 13.1 % des articles sont publiés le *weekend* dont 71.6 % sont populaires (respectivement 74.8 % d'articles populaires pour le *samedi* et 68.7 % pour le *dimanche*).
- 86.9 % des articles sont publiés la semaine dont 50.4 % sont populaires.



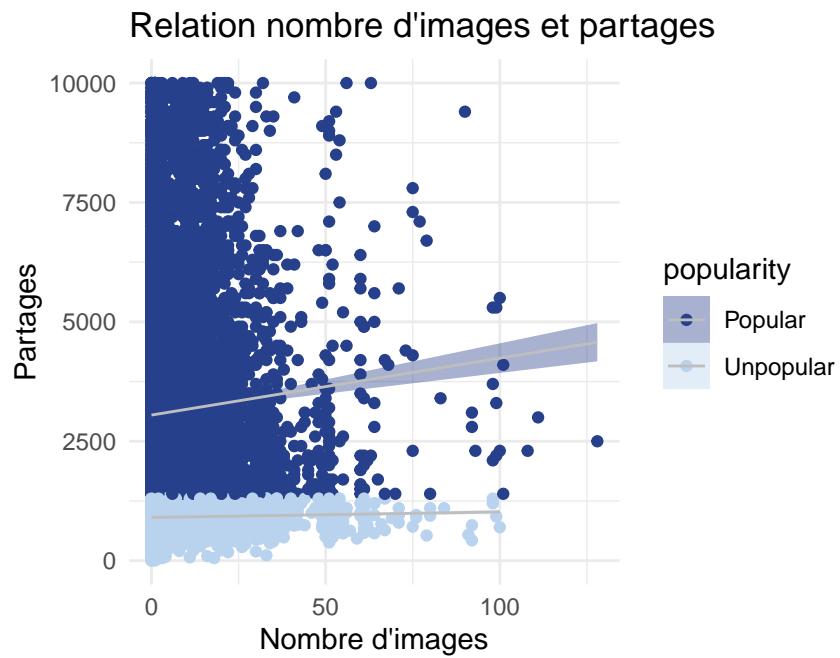
En proportion, le thème *social media* est le plus populaire et le thème *world* est le moins populaire.

- 6 % des articles sont du thème *social media*, dont 75.9 % sont populaires.
- 21.2 % des articles sont du thème *world*, dont seulement 38.1 % sont populaires.

3.3.1 Images, vidéos et liens

Nous nous intéressons maintenant à l'influence du nombre d'images sur les partages. Nous allons regarder en particulier comment cela influe en fonction d'un article populaire ou non.

Nous nous restreignons ici à 10 000 partages maximum ce qui représente 95% de nos observations, car les articles grandement partagés influencent beaucoup la relation.



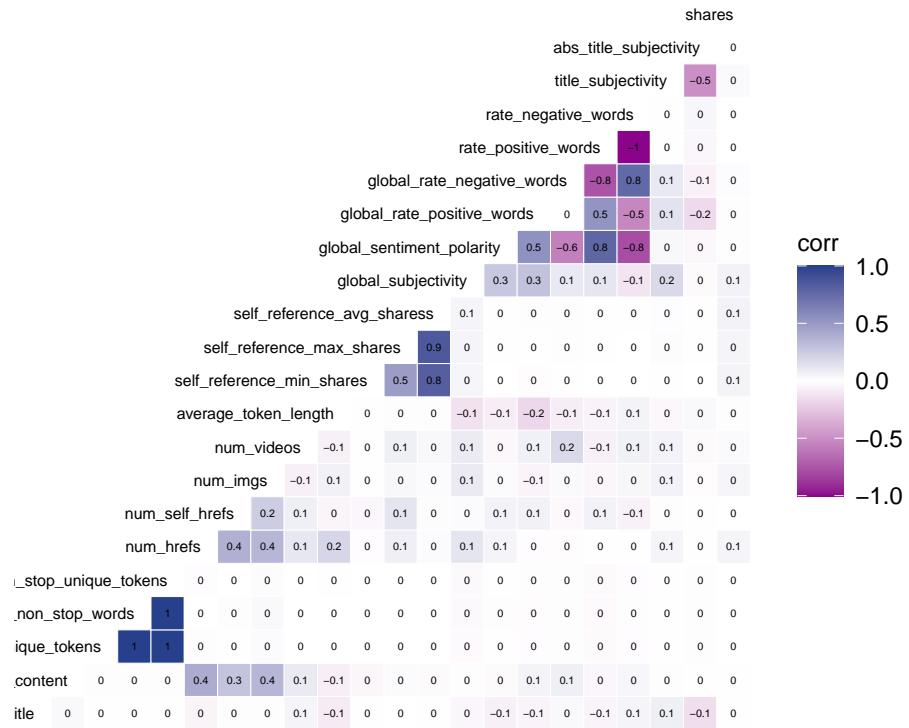
Nous observons qu'il y a une relation positive entre le nombre de partages et le nombres d'images pour les articles populaires donc partagés plus de 1400 fois. Cependant, pour les articles impopulaires la relation n'est pas concluante.

Les conclusions sont identiques pour les relations entre les variables **partage** et **vidéo** puis **partage** et **lien**. Une relation positive se distingue pour les articles populaires mais pas pour les impopulaires.

En revanche, comme vu précédemment dans la partie sur les statistiques à une variable, la plupart des articles ont peu d'images, de vidéos et de liens donc ici les valeurs extrêmes influencent fortement la relation.

4 Corrélation

Nous retirons toutes les variables que nous ne jugeons pas nécessaires telles que les variables en rapport avec les mots clés ou encore les variables de polarité.



Cette matrice de corrélation pourra nous être utile dans l'ACP et dans l'afc qui seront faites prochainement. Même si nous voyons que très peu de variables sont corrélées entre elles.

5 Conclusion

Pour conclure, nous pouvons dire que les partages dépendent de plusieurs variables. Par exemple, les articles sur certains thèmes tels que *social media* et *lifestyle* sont plus propices à être partagés. Les articles publiés le week-end ont aussi plus de chance d'être relayés par les lecteurs.

Nous avons également vu que le contenu de l'article (les images, liens et vidéos) peut, dans une certaine mesure, influencer les partages. Nous pouvons faire l'hypothèse que les articles ont tendance à être plus partagés quand leur contenu est plus diversifié et qu'ils ne se limitent pas seulement à du texte.