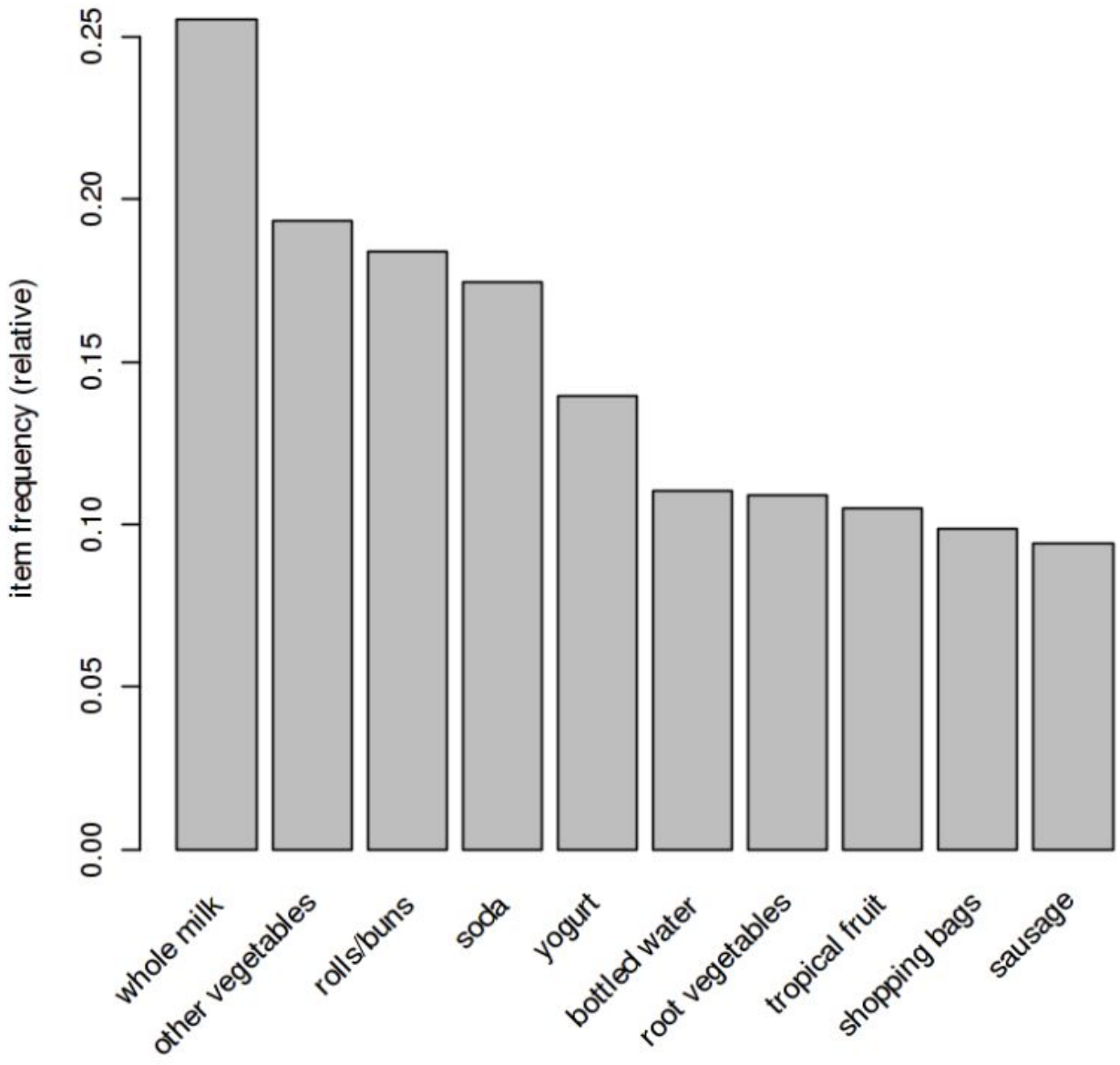
4DATA - Graded Exercise

Le dataset “Groceries” contient 1 mois d’achat réalisé dans un magasin. Dans ce répertoire on trouve 9835 transaction, sur une sélection de 169 catégories.

PARTIE 1 : Analyse du dataset Groceries

Chaque transaction contient en moyenne 4 à 5 articles, ce qui donne donc une densité de 2.6%. En regardant la densité de chaque produit on peut voir lesquels sont les plus consommés.

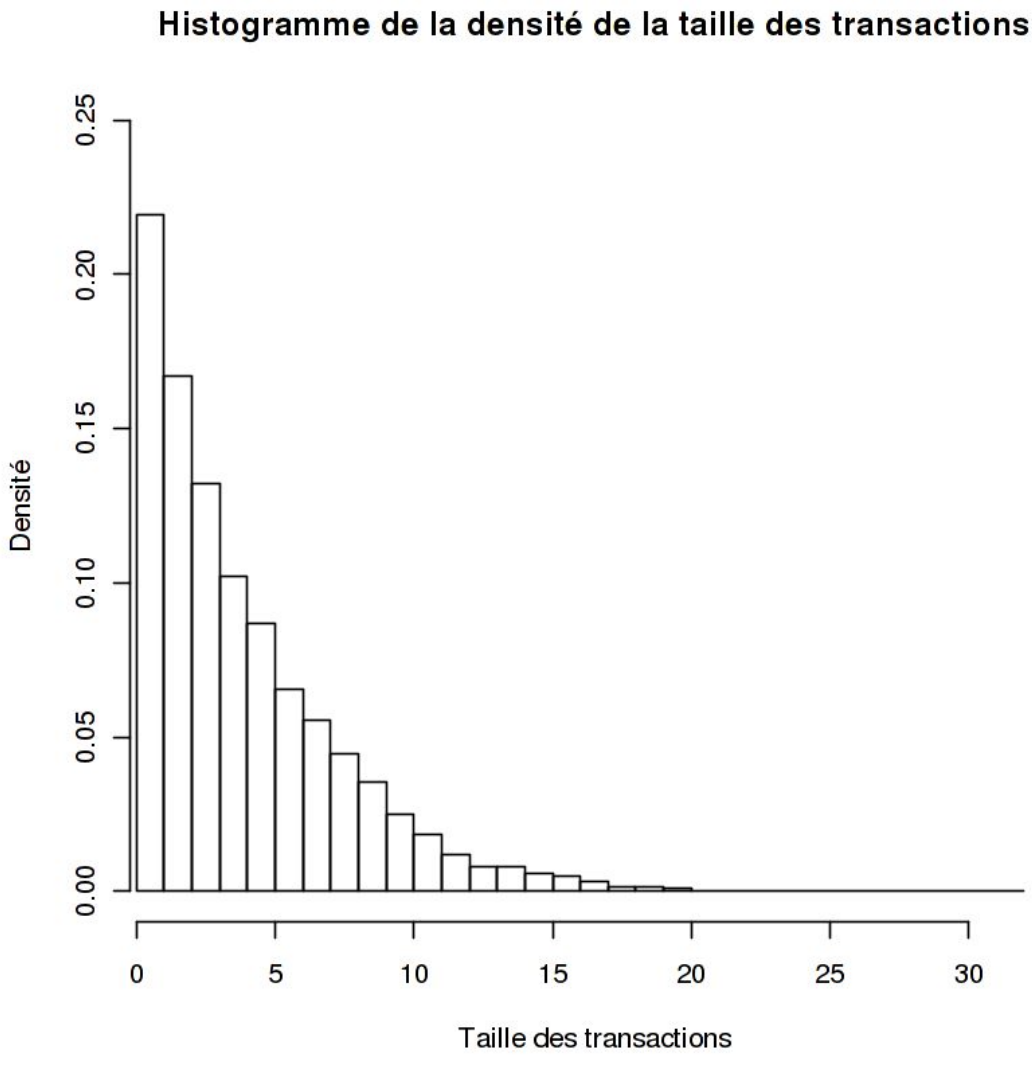


On peut ainsi voir que les 5 articles les plus consommés sont : le lait entier, les légumes, les petits pains, les sodas et le yaourt, dans le graphique précédent leur densité. Voici les valeurs de la densité pour les différents articles :

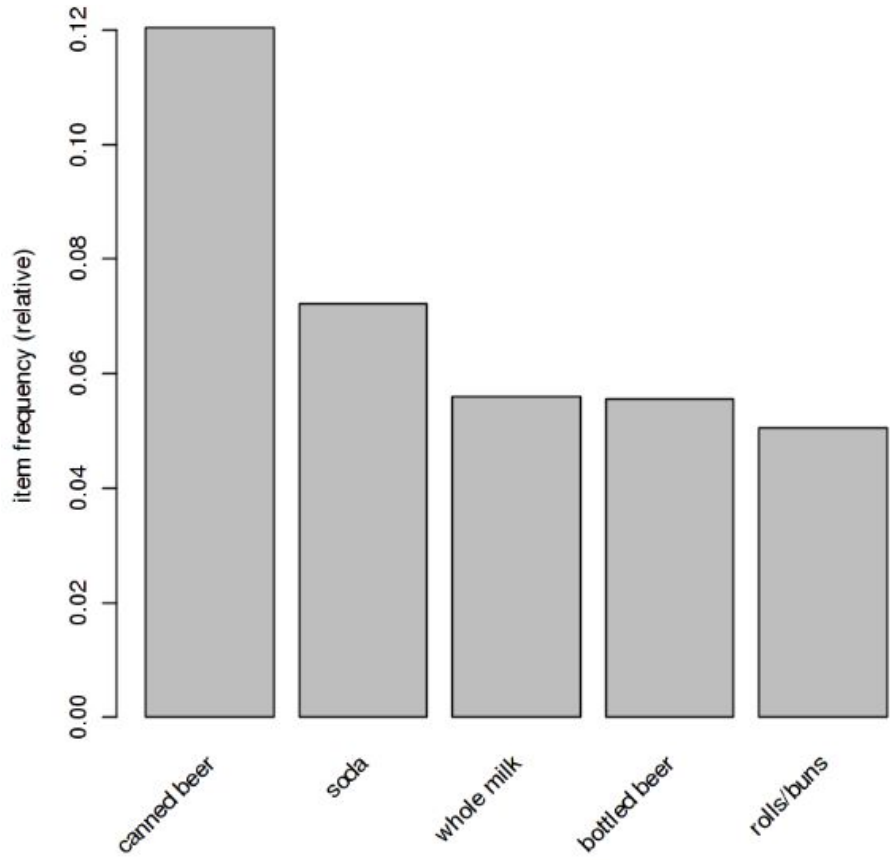
whole milk	0.255516014234875
other vegetables	0.193492628368073
rolls/buns	0.183934926283681
soda	0.174377224199288
yogurt	0.139501779359431
bottled water	0.110523640061007
root vegetables	0.108998474834774
tropical fruit	0.10493136756482
shopping bags	0.0985256736146416
sausage	0.0939501779359431

On peut donc voir que le lait est présent dans plus d'un quart des transactions.

La tailles des différentes transactions est également très variable, elles vont de 1 article jusqu'à 32.



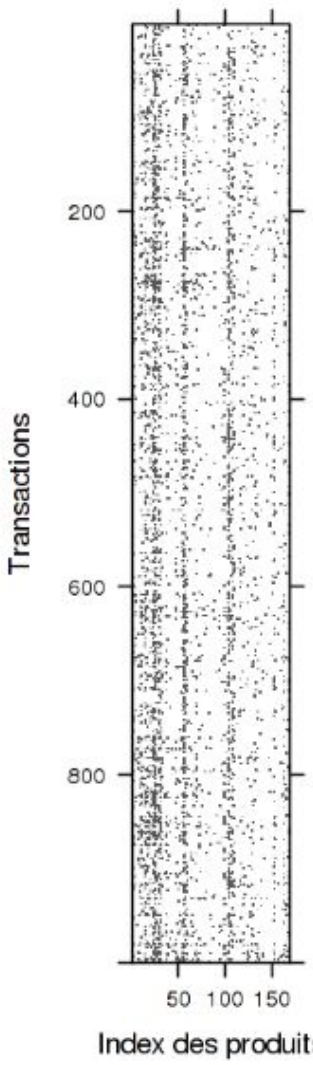
On peut donc voir que les personnes favorisent les transactions avec peu d'articles. Par exemple, il y a 1643 transactions avec 2 articles et 545 transactions avec 7 articles. On peut également noter que les articles les plus achetés dépendent du nombre d'articles. L'exemple le plus marquant à lieu avec les transactions de 1 article, alors que, de manière général l'article le plus consommé est le lait entier. Pour les transactions à un article on trouve le résultat suivant :



Le produit le plus consommé dans les transactions à 1 article sont les cannettes de bières.

Il existe des moyens de visualiser l'ensemble de notre dataset, pour plus de visibilité nous allons prendre un échantillon de 1000 transactions :

Représentation d'un échantillon du dataset Groceries



On distingue 3 "lignes" qui se tracent. Si on regarde les fréquences des éléments autour de ces index on retrouve ainsi les items qui sont les plus fréquent dans le dataset.

La première ligne apparaît dans les 50 premiers items.
Si l'on regarde de plus près on voit que 3 des 5 produits les plus consommé se trouve à cet endroit:

other vegetables	0.193492628368073
packaged fruit/vegeta...	0.0130147432638536
whole milk	0.255516014234875
yogurt	0.139501779359431

Une deuxième tendance se dessine un peu après le 50ème index, il s'agit du troisième élément le plus acheté: **rolls/buns:** 0.183934926283681.

La troisième ligne apparaît un peu après le produit avec l'index 100. C'est la que se trouve l'élément manquant du top 5, le soda:

bottled water	0.110523640061007
soda	0.174377224199288

PARTIE 2 : Créer des règles d'association

Dans cette partie nous allons analyser les données des transactions en utilisant des règles d'association pour identifier les éléments qui sont fréquemment associés dans un ensemble de données.

Pour ce faire nous allons utiliser l'algorithme *Apriori* pour générer nos règles d'associations.

Une règle d'association comporte trois principaux paramètres :

- Le support qui correspond à la probabilité qu'un groupe d'objet soit dans la même transaction.
- La confidence qui correspond à la probabilité que la règle d'association s'avère être vraie.
- Le Lift qui correspond à la dépendance que deux éléments ont. Un lift de 1 veut dire que les deux éléments sont totalement indépendants (la présence d'un élément n'influence pas la présence de l'autre élément). Plus un lift est grand (supérieur à 1), plus les deux éléments sont dépendants (la présence de l'un **augmente** la probabilité que l'autre élément soit également présent). Plus le lift est faible (inférieur à 1), plus les deux éléments sont incompatibles (la présence de l'un **diminue** la probabilité que l'autre élément soit également présent).

Voici les 10 règles dont le lift est le plus élevé avec au moins 0,9% de support et 50% de confidence dans le dataset *Groceries*.

	lhs		rhs	support	confidence	lift	count
[1]	{citrus fruit,root vegetables}	=>	{other vegetables}	0.010371124	0.5862069	3.029608	102
[2]	{tropical fruit,root vegetables}	=>	{other vegetables}	0.012302999	0.5845411	3.020999	121
[3]	{root vegetables,rolls/buns}	=>	{other vegetables}	0.012201322	0.5020921	2.594890	120
[4]	{root vegetables,yogurt}	=>	{other vegetables}	0.012913066	0.5000000	2.584078	127
[5]	{butter,yogurt}	=>	{whole milk}	0.009354347	0.6388889	2.500387	92
[6]	{curd,yogurt}	=>	{whole milk}	0.010066090	0.5823529	2.279125	99
[7]	{other vegetables,curd}	=>	{whole milk}	0.009862735	0.5739645	2.246296	97
[8]	{other vegetables,butter}	=>	{whole milk}	0.011489578	0.5736041	2.244885	113
[9]	{tropical fruit,root vegetables}	=>	{whole milk}	0.011997966	0.5700483	2.230969	118
[10]	{root vegetables,yogurt}	=>	{whole milk}	0.014539908	0.5629921	2.203354	143

Par exemple pour la règle numéro 5, un client à 0.93% de chance de prendre du beurre, du yaourt et du lait entier, mais à 63,89% de prendre du lait entier si il prend du beurre et du yaourt.

Pour un support de 0,1% et une confidence de 10%, voici la règle avec le support le plus élevé :

```
     lhs                rhs                support    confidence lift     count
[1]  {}              => {whole milk}       0.25551601 0.2555160  1.000000 2513
```

Avec cette règle on peut voir qu'un client à 25,5% de chance d'acheter du lait.

Voici la règle avec la confidence la plus haute :

```
    lhs             rhs          support     confidence lift     count
[1] {rice,sugar} => {whole milk} 0.001220132 1          3.913649 12
```

Cette règle nous montre que quand un client prend du riz et du sucre il prend aussi du lait.

Voici la règle avec le lift le plus élevé :

```
    lhs                             rhs      support     confidence lift     count
[1] {bottled beer,red/blush wine} => {liquor} 0.001931876 0.3958333  35.71579 19
```

Cette règle nous montre que les personnes qui achète de la bière et du vin ont fortement tendance à acheter de l'alcool.

Et pour finir voici la règle avec le lift le plus bas :

```
    lhs                           rhs          support     confidence lift      count
[1] {canned beer,shopping bags} => {whole milk} 0.001220132 0.1071429  0.4193195 12
```

Ici on peut voir que les personnes qui achète de la bière et des sacs ont tendance à ne pas prendre de lait.