QUI SE RESSEMBLE S'ASSEMBLE :

UN ALGORITHME DE SEGMENTATION

SEGMENTATION ET PREDICTION DES CLASSES DES UTILISATEURS DU SITE DE RENCONTRE OKCUPID SELON LEURS RESSEMBLANCES

Master 2 Modélisations Statistiques, Économiques et Financières

Data Mining

Fancello Marie Clara, Germini Eva, Gutfreund Eloïse

Table des matières

INTRODUCTION

Selon l'article *How Couples met* de Katharina Buchholz, analysant les lieux et contextes des rencontres des couples hétérosexuels, les rencontres sur internet connaissent une croissance exponentielle au XXIe siècle. En effet, en 2017, alors que 11% des couples hétérosexuels se rencontrent sur leur lieu de travail, 39% se rencontrent sur internet. Cette augmentation importante (2% en 1995 contre 39% en 2017) s'explique notamment par l'essor accru des sites de rencontres *(Buchholz, 2020)*.

Dans cette étude, nous analysons les profils des utilisateurs de Okcupid, une application de rencontre sur mobile née en 2004 aux Etats-Unis. La base de données, récupérée du site Kaggle, regroupe les utilisateurs de Okcupid et donne des informations sur leurs caractéristiques personnelles comme l'âge, le sexe et leur profile (différentes descriptions sont associées).

Ainsi, partant de l'hypothèse que les personnes avec des caractéristiques similaires seraient davantage compatibles, l'objectif de cette analyse est de regrouper les utilisateurs selon leurs ressemblances. Nous donnerons ainsi, un groupe à chaque individu. Nous prédirons également le groupe de chaque nouvel utilisateur par un algorithme de scoring, l'objectif étant de lui proposer de matcher avec des personnes de son groupe pairs.

Dans un premier temps, nous analysons la base de données et procédons aux transformations nécessaires des variables. Dans une deuxième partie, nous utilisons une Analyse en Composantes Principales (ACM) et un algorithme de machine learning non supervisé (k-means) afin de regrouper les individus selon leurs similarités. Enfin, dans un troisième développement, nous effectuons un algorithme Random Forest afin de prédire les groupes de chaque nouvel utilisateur.

I) ANALYSES ET TRANSFORMATIONS DE LA BASE DE DONNEES

A. Premières analyses de la base de données

La base de données est composée de 31 variables et de 59 946 observations. Aucun doublon n'est enregistré dans la base de données, chaque observation correspond donc à un individu distinct. 3 variables sont de type numérique et 28 variables sont de type caractère. Les variables numériques correspondent à l'âge, la taille et le salaire de l'individu. Les variables de type caractère donnent des informations personnelles (sexe, situation amoureuse, orientation sexuelle …). 10 variables notées "essay" sont des champs de textes ouverts dans l'application. Elles répondent aux caractéristiques suivantes:

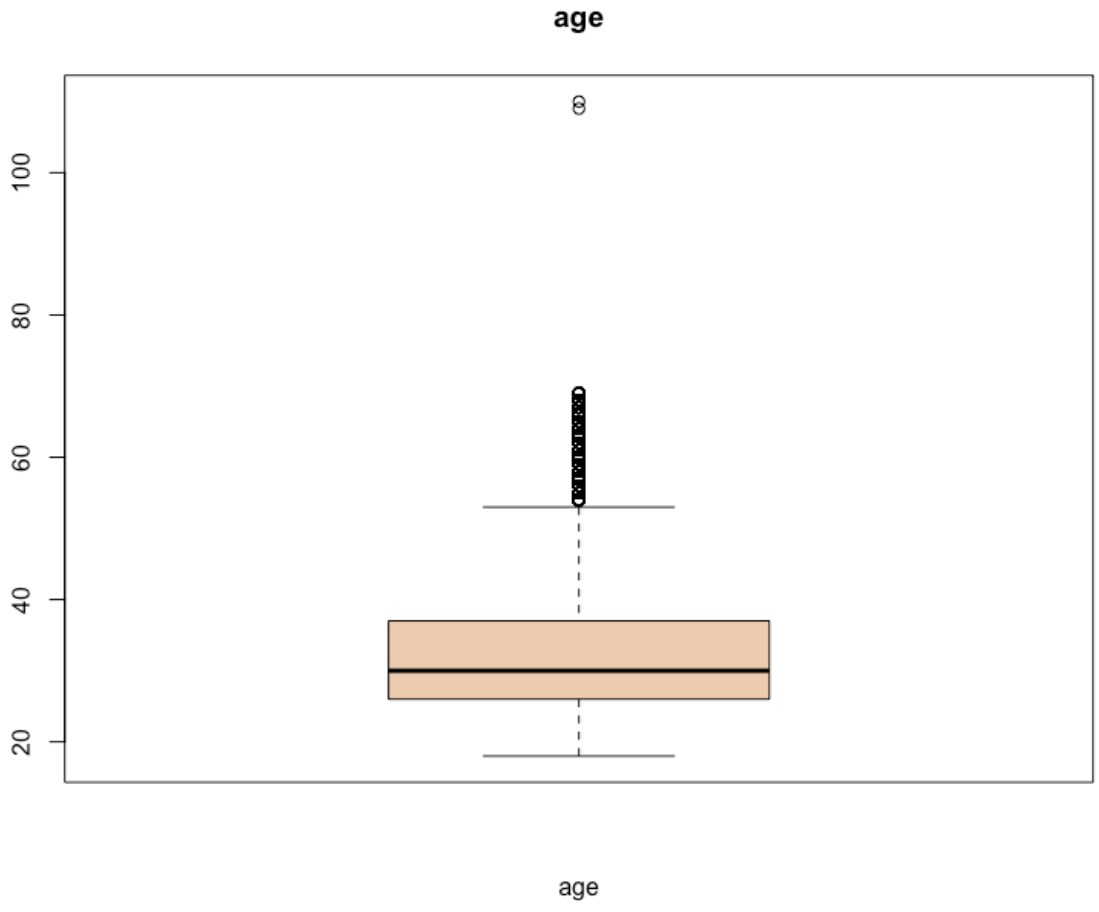
- "essay0": description générale de l'individu
- "essay1": résumé descriptif de l'individu
- "essay2": aspirations personnelles
- "essay3": traits de caractères
- "essay4": passions
- "essay5": hobbies
- "essay6": un moment/une journée parfaite
- "essay7": besoins
- "essay8": chose la plus privée sur elle même que la personne souhaite communiquer
- "essay9": attentes amoureuses

Nous analysons ensuite les valeurs manquantes dans la base de données. L'âge, le status amoureux, le sexe, l'orientation sexuelle, le revenu, la dernière connexion et la localisation ne présentent aucune valeur manquante. Si l'on analyse dans un premier temps les variables continues, seule la variable "height" présente 0.005% de valeurs manquantes. Ces valeurs manquantes sont de type MCAR (Missing Completely At Random) car elles sont totalement indépendantes des variables observables et des paramètres extérieurs et se produisent donc complètement au hasard. Concernant les variables catégorielles, les valeurs manquantes sont de type MNAR (Missing Not Completely At Random) car la valeur de la variable manquante est liée à la raison pour laquelle elle manque, et pour cause, l'individu n'a pas répondu à cette

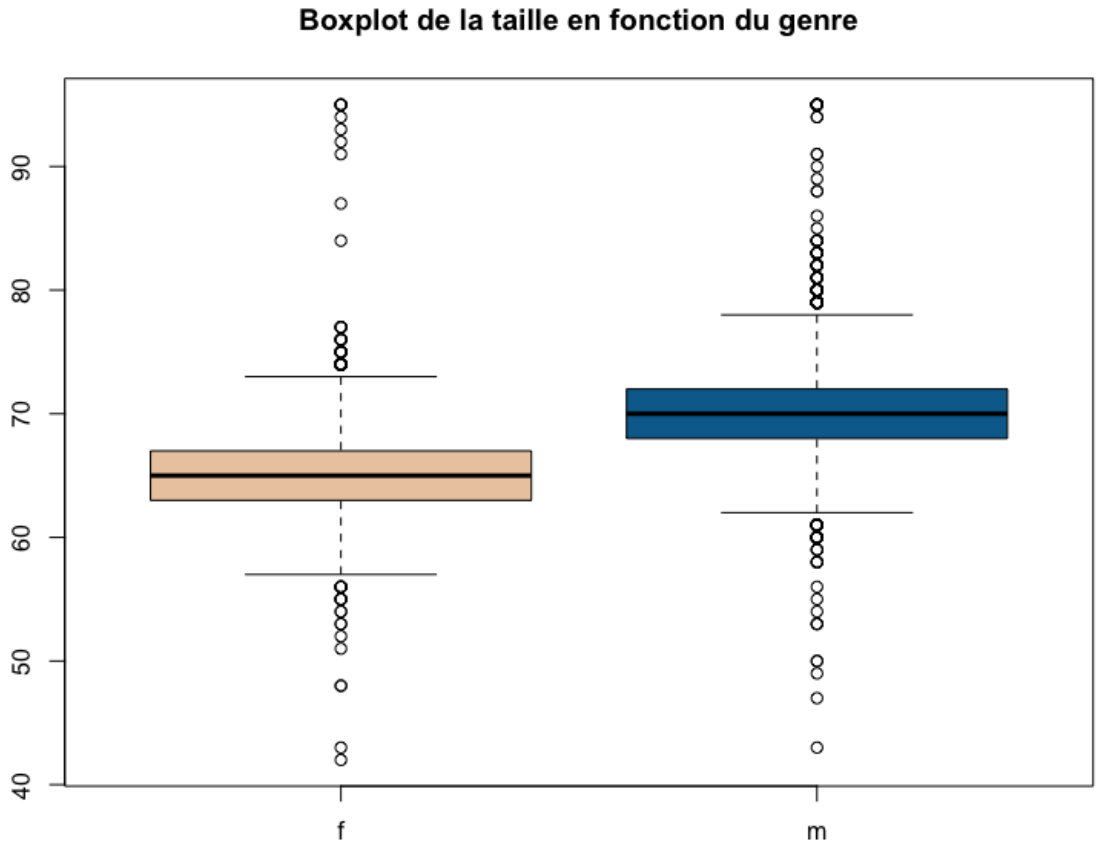
question. A titre d'exemples, environ 41% et 11% des utilisateurs n'ont respectivement pas donné d'information sur leur régime alimentaire ("diet") et leur études ("education").

B. Analyses univariées et gestion des valeurs manquantes et aberrantes des variables continues

La variable de l'âge notée "age" ne présente aucune valeur manquante. L'âge minimum est de 18 ans. L'âge maximum est supérieur à 100 ans. En moyenne, les individus ont 32 ans. 75% de la population a plus de 26 ans et la médiane est de 30 ans. La distribution comporte plusieurs valeurs extrêmes, certaines étant des valeurs envisageables pour des personnes sur des sites de rencontre, certaines semblant être des erreurs (valeurs supérieures à 100 ans). Nous traitons donc les valeurs extrêmes supérieures à 100 ans par une technique d'imputation par la médiane.



En ce qui concerne la taille des individus, nous ne connaissons pas l'unité utilisée au sein de la base de données, car non renseignée sur le site Kaggle. La médiane et la moyenne sont situées à 68, ce qui nous indique que sa distribution est centrée. Cependant, cette variable présente des valeurs aberrantes, ainsi que 0.005% de valeurs manquantes *(Graphique 1)*. Ces données concernent des hommes et des femmes, avec une proportion plus faible de femmes (60% d'hommes contre 40% de femmes). Sans surprise, les femmes sont en moyenne plus petites que les hommes. 50% des hommes ont une taille supérieure à 70 contre 65 pour les femmes.



Il est donc injustifié de traiter les valeurs manquantes et aberrantes de la même manière pour les hommes et les femmes. Nous traitons donc les valeurs aberrantes en imputant par la médiane de chaque catégorie : 70,4 pour les hommes et 65.1 pour les femmes.

Finalement, nous supprimons la variable de revenu nommée “income” qui n’est presque jamais renseignée. Effectivement, elle est imputée par -1 dans la base de données *(Graphique 2)*.

C. Analyses univariées et gestion des valeurs manquantes des variables catégorielles

Pour l’ensemble des variables catégorielles présentant des valeurs manquantes, nous créons une modalité “didnt_answer” qui correspond au fait que l’individu n’a pas répondu à la question. Également, pour chaque variable un compte des modalités est présenté en annexe *(Tableau 2 à 17)*.

Les variables “status” (situation amoureuse), étant composée d’environ 93% des individus en situation amoureuse “single” *(Tableau 2)*, et “state” comprenant plus de 99,8% des individus résidant en Californie ne permettent pas de distinguer les individus. Nous décidons donc de les exclure de l’analyse. La variable “location” étant un détail de la variable “state” et ayant un nombre de modalités trop important, nous la supprimons.

Les variables “edu” donnant le niveau d’étude, “sign” donnant le signe astrologique et “job” donnant l’emploi des individus sont composées de l’information et d’un commentaire. Pour chaque individu, nous gardons l’information et supprimons le commentaire.

Les variables “body_type” correspondant à l’apparence physique et “job” correspondant à l’emploi occupé de l’individu sont également exclues de l’analyse car ces variables présentent un grand nombre de modalités *(Tableau 3 et 4)*.

Les variables “sex” (genre de l’individu) et “orientation” (orientation sexuelle) ne présentent pas de valeurs manquantes et ne sont pas transformées. Ainsi, la base de données contient environ 40% de femmes et 60% d'hommes ainsi que 86% de personnes hétérosexuelles *(Tableau 5 et 6)*.

Les variables “drinks”, “drug” et “smokes” correspondant respectivement à la consommation d’alcool, de drogues et le fait de fumer sont transformées en variables catégorielles. Pour la

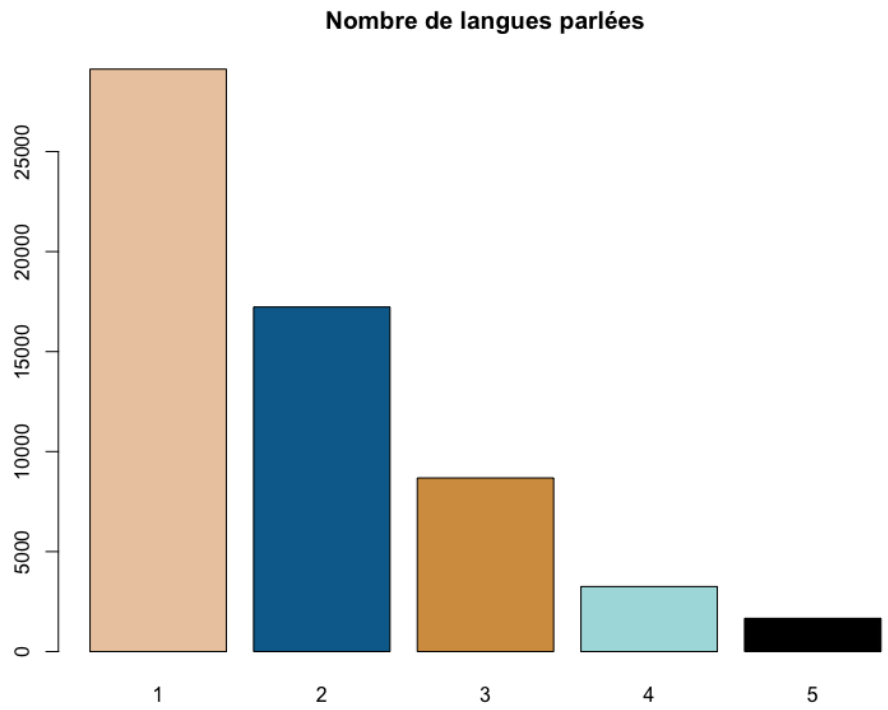
variable “drinks”, les individus ayant la modalité “not at all” sont considérés comme des personnes ne consommant jamais d’alcool (modalité “no”), les autres comme des personnes ayant au moins une consommation peu fréquente d’alcool (modalité “yes”). Ainsi, seulement 5,45% des personnes ne consomment jamais d’alcool *(Tableau 7)*. La variable “drug” classe les personnes ayant répondues “never” dans la modalité “no” (62,93% des individus), les autres sont placées dans “yes” (environ 14% des utilisateurs) *(Tableau 8)*. La variable “smokes” classe les personnes ayant répondues “no” comme étant les personnes non fumeuses (73,23% des observations), les autres sont considérées comme fumeuses (modalité “yes”) (17,58% des utilisateurs) *(Tableau 9)*.

La variable “edu” étant composée de 12 modalités et ayant la modalité “graduated” représentant 65,69% des individus, nous la traitons. Ainsi, la variable “graduated” prend la modalité “yes” si la personne est diplômée et “no” si la personne n’est pas diplômée (23,26%) *(Tableau 10)*.

La variable “pets” donnant l’information si l’individu aime ou a un chien et/ou un chat est séparée en deux variables: une première nommée “dog” prenant la modalité “dog_friendly” si l’individu aime ou a un chien (62,45% des individus), “not_dog_friendly” si ce n’est pas le cas (4,32% des utilisateurs) *(Tableau 11)* et une seconde “cat” prenant la modalité “cat_friendly” si l’individu aime ou a un chat (47,75% des individus), “not_cat_friendly” sinon (19,02% des utilisateurs) *(Tableau 12)*.

La variable “offsprings” représentant le fait d’avoir ou souhaitant avoir des enfants permet de créer une nouvelle variable catégorielle “kids_friendly” avec la modalité “kid_friendly” si l’individu a ou souhaite avoir des enfants, représentant 32.71% de la population et “not_kid_friendly” sinon, représentant 7,97% de la population *(Tableau 13)*.

La variable “speaks” représentant les langues parlées permet de créer une nouvelle variable nommée “number_languages” comptant le nombre de langues parlées. Sans surprise, le nombre de langues parlées est décroissanPlus de 15 000 individus parlent deux langues.



La variable “ethnicity” donnant l’ethnicité des utilisateurs contient majoritairement l’ethnie “white”. Les individus pouvant spécifier plusieurs ethnies, nous créons une variable catégorielle “white”. 63,19% des individus ont spécifié cette ethnie, 27,33% des personnes ont renseigné une autre ethnie et 9,48% n’ont pas répondu *(Tableau 14)*.

La variable “diet” indiquant le régime alimentaire des utilisateurs permet de créer une variable “cat_diet” prenant les modalités suivantes:

- “anything” si l’individu a précisé qu’il n’avait pas de régime alimentaire spécifique (environ 47% des individus)
- “vegetarian/vegan” si l’individu précise un régime végétarien ou végan (environ 9% des individus)
- “other” représentant d’autres régimes alimentaires (environ 3% des individus)
- “didnt_answer” si l’individu n’a pas spécifié de régime alimentaire (environ 41% des individus) *(Tableau 15)*.

La variable “religion” donnant des précisions sur les pratiques religieuses des individus est traitée afin de créer la variable “religious” ayant la modalité “religious” si l’individu est

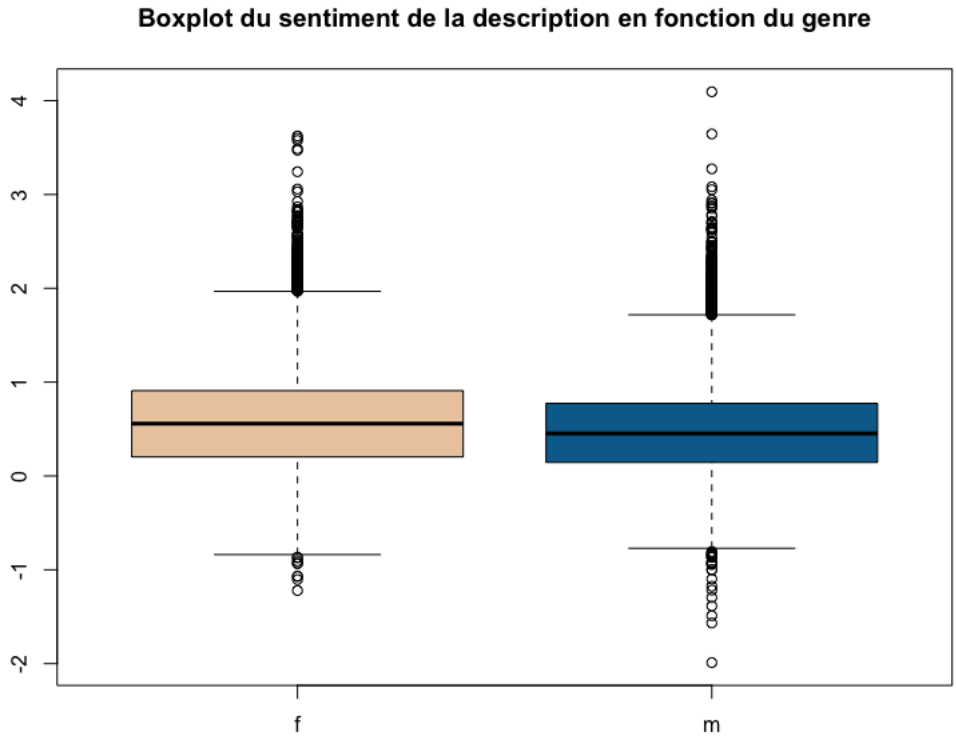
religieux “not_religious” sinon. Nous considérons que les personnes agnostiques et les personnes mentionnant “laugh_religion “ ne sont pas religieux. Environ 41% des individus sont religieux contre 26% qui ne le sont pas *(Tableau 16)*.

Enfin, la variable concernant le signe astrologique est composée de 14 modalités (13 signes astrologiques et “didnt_answer” pour les utilisateurs n’ayant pas répondu). A titre d'exemples, 7,30% des individus sont lions et 6.90% sont scorpions. 18,44% des individus n’ont pas répondu à cette question *(Tableau 17)*.

D. Analyses univariées et gestion des valeurs manquantes des variables textuelles

Les variables textuelles présentes dans la base de données sont préfixées par “essay”.
La première variable que nous traitons est “essay0” correspondant à la description générale de l’individu. Nous utilisons la librairie “sentimentr” afin de savoir si l’individu se décrit de manière positive, négative ou neutre. Elle permet de calculer la polarité d’un texte en anglais. Cette librairie utilise un dictionnaire de mots qui attribue une valeur à chaque mot. S’il y a des négations, elle pondère cette polarité en l’inversant. Elle prend également en compte les amplificateurs (mots qui augmentent l’impact d’un mot polarisé), les désamplificateurs (mots qui réduisent l’impact d’un mot polarisé) ainsi que les conjonctions adversatives (mots qui annulent la clause précédente contenant un mot polarisé) dans le calcul du score de sentiment. Nous obtenons alors une variable continue se nommant “description_sentiment” indiquant la polarité de chaque description des individus. Pour l’interprétation, nous considérons que si la polarité est supérieure à 0 alors l’utilisateur se décrit de manière positive, si elle est inférieure à 0 alors l’individu se décrit de manière négative et si elle vaut 0 alors l’individu se décrit de manière neutre *(Rinker, 2018)*.

En moyenne, la polarisation est d’environ 0,54. L’individu se présentant avec le sentiment le plus négatif à une polarité d’environ -1,99, tandis que l’utilisateur se présentant avec le sentiment le plus positif à une polarité d’environ 4,09. L’écart entre les personnes se présentant de manière positive et négative est donc important *(Tableau 18)*. Nous analysons le sentiment des descriptions par genre :



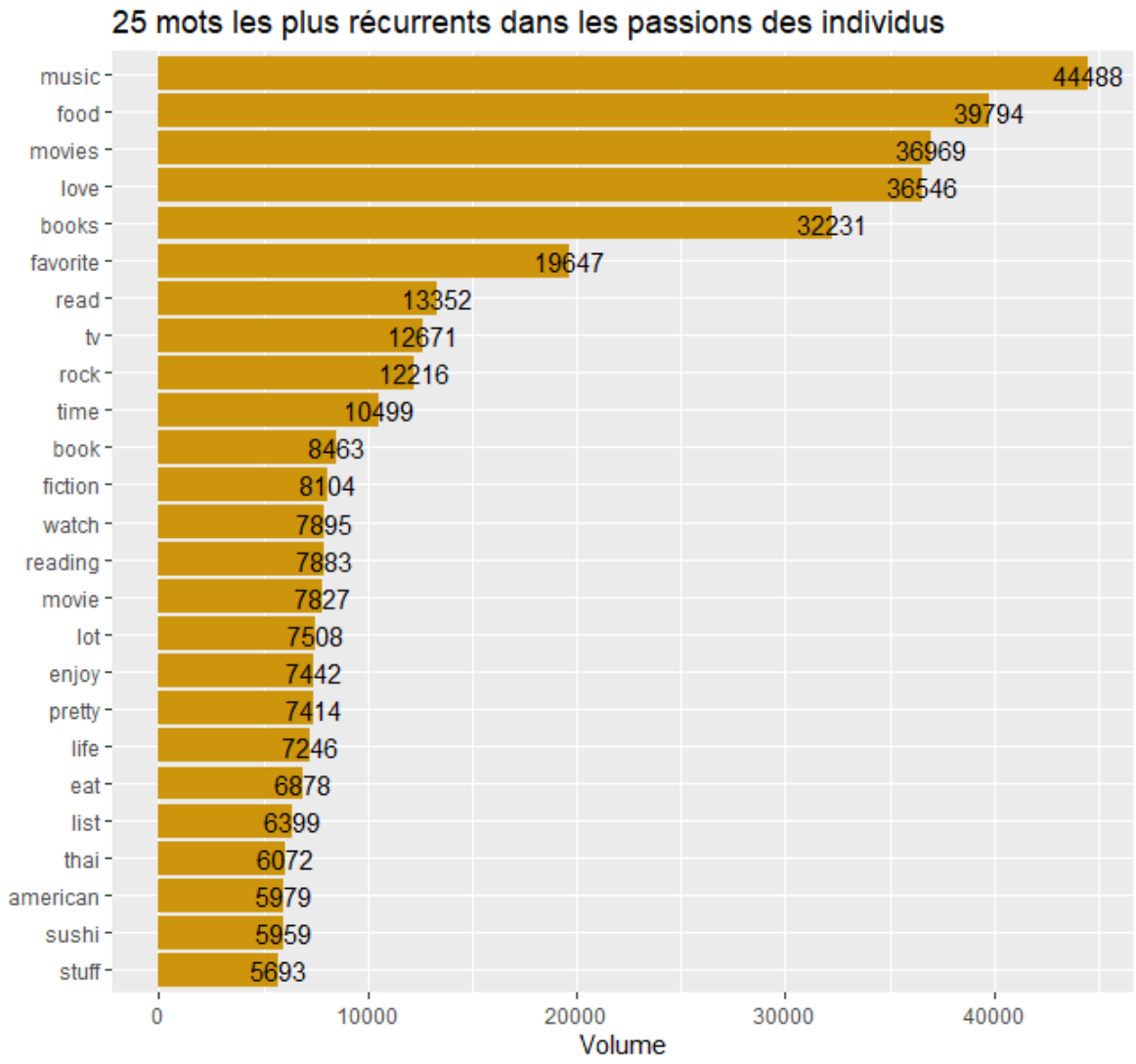
Globalement, les femmes se décrivent d'une manière plus positive que les hommes, cependant, il faut noter que l'écart de genre est dérisoire. Les personnes avec le sentiment le plus positif et le sentiment le plus négatif sont des hommes. Pour les femmes, le minimum se situe entre -1 et -2: les femmes ne se décrivent donc jamais de manière fortement négative.

Pour les variables "essay" allant de 1 à 9, nous comptons le nombre de mots les plus occurrents. Dans cet objectif, nous traitons le texte. Dans un premier temps, les contractions sont remplacées par leurs extensions. Par exemple, "won't" est remplacé par "will not". Dans un second temps, tous les caractères spéciaux sont supprimés des textes et les lettres sont passées en minuscules. Pour finir, les "stop words" sont retirés. Ce sont des mots qui n'apportent aucune information quant à la nature du texte. Un exemple de stop words peut être "the".

La variable "essay1" correspondant au résumé descriptif de l'individu n'est pas traitée car elle reprend la variable "essay0" déjà traitée précédemment. Ayant un faible nombre d'occurrences par mot, les variables "essay2" (aspirations des individus), "essay6" (description d'un moment parfait), "essay8" (chose la plus privée sur elle-même que la personne souhaite communiquer) et "essay9" (attentes amoureuses) sont également non

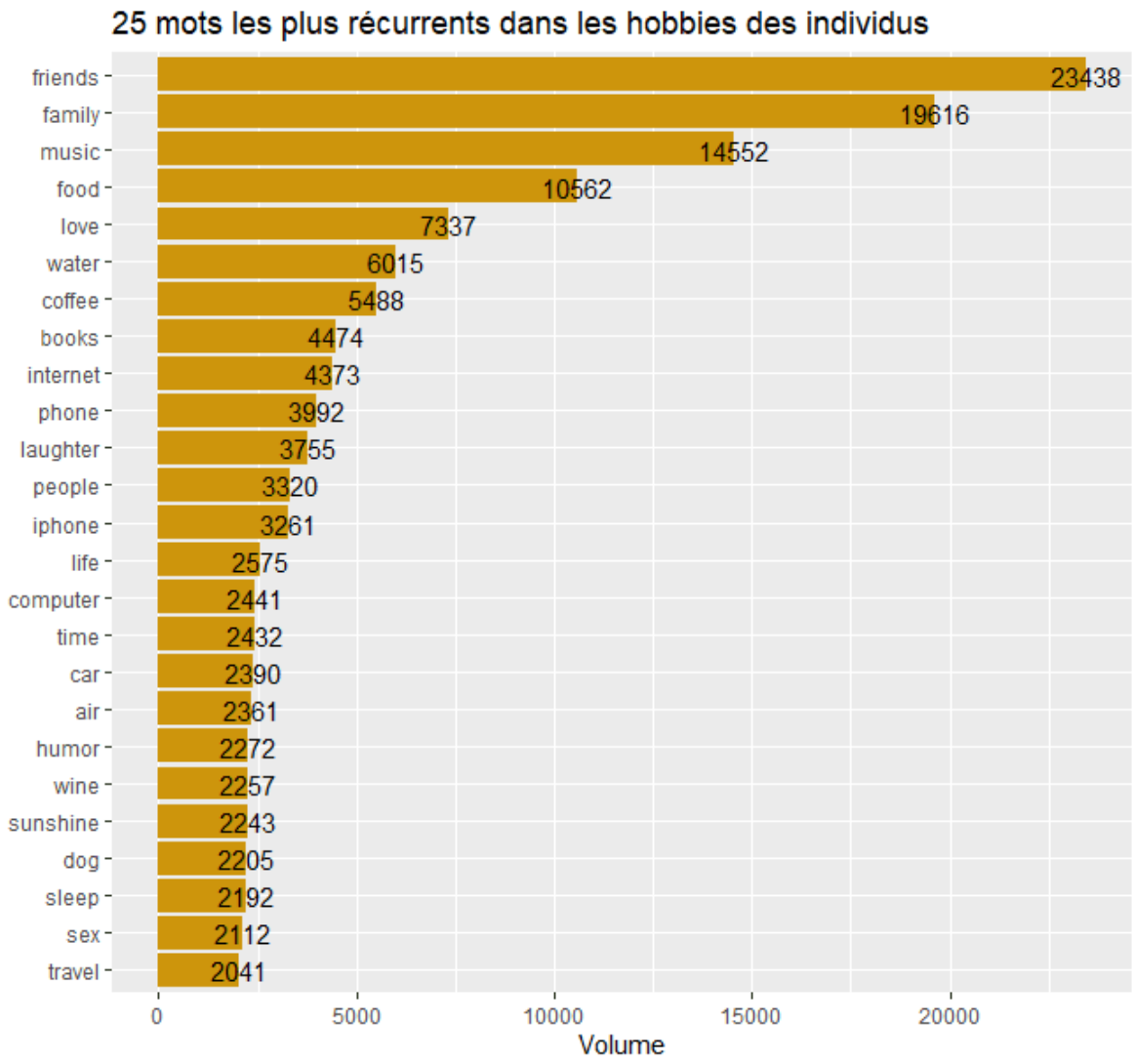
utilisées *(Graphique 3, 4, 5 et 6)*. La variable “essay3”, correspondant aux traits de caractères de l’individu étant subjective, nous ne l’analysons pas. Enfin, nous ne gardons pas la variable “essay7”, correspondant au besoin des individus, car son mot le plus représenté “friends” est repris par la variable “essay5” *(Graphique 7)*.

La variable “essay4” permet de constater que 4 passions sont récurrentes chez les utilisateurs : la musique, la nourriture, les films et les livres.



Nous créons donc quatre variables binaires : “loves_music”, “loves_movies”, “loves_food” et “loves_book”. Si l’individu a mentionné une de ces passions alors la valeur 1 lui est attribuée 0 sinon. 54,38%, 46,79%, 49,55% et 42,73% des individus aiment respectivement la musique, les films, la nourriture et les livres *(Tableau 19, 20, 21 et 22)*.

Pour la variable “essay5” (hobbies), deux mots sont récurrents : “friends” et “family”. Nous créons donc deux variables binaires : “friends” et “family”. Si l’individu a mentionné un de ces hobbies alors la valeur 1 lui est attribuée 0 sinon. 39,02% et 31,99% des individus mentionnent respectivement leurs amis et leur famille *(Tableau 23 et 24)*.

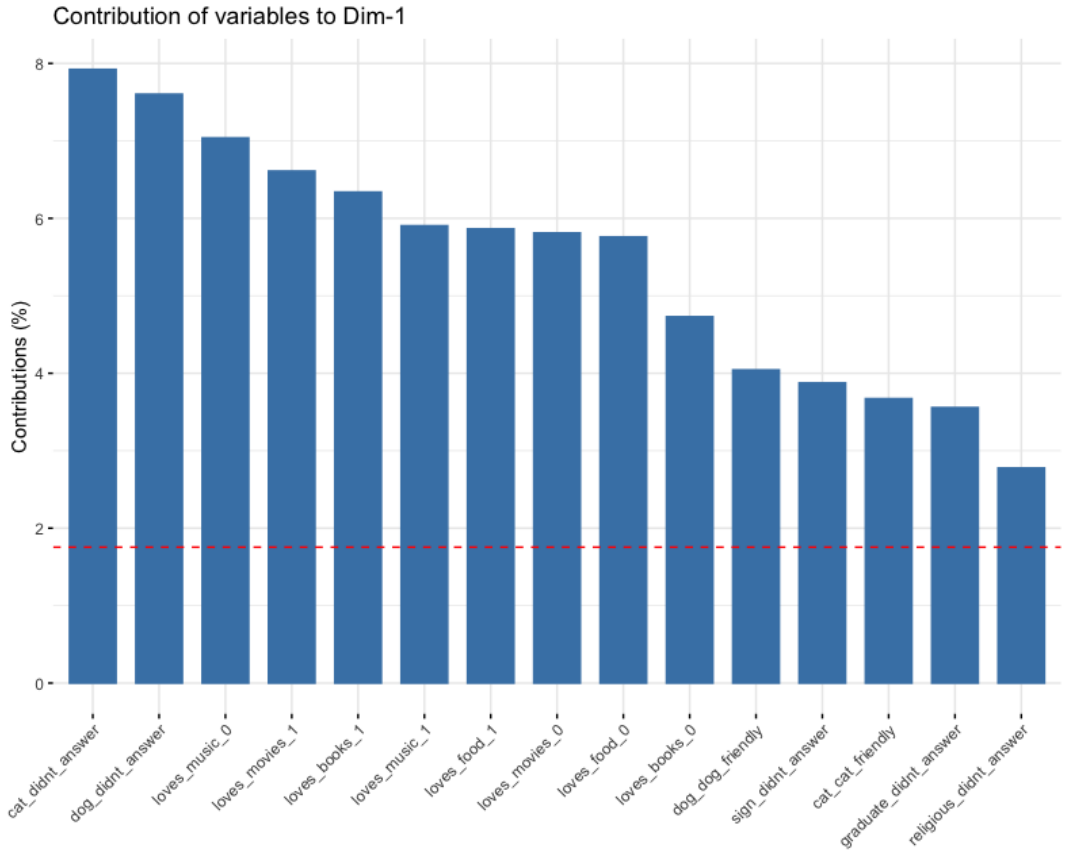


II) SEGMENTATION DES UTILISATEURS SELON LEURS RESSEMBLANCES

A. Transformation des variables catégorielles en variables continues

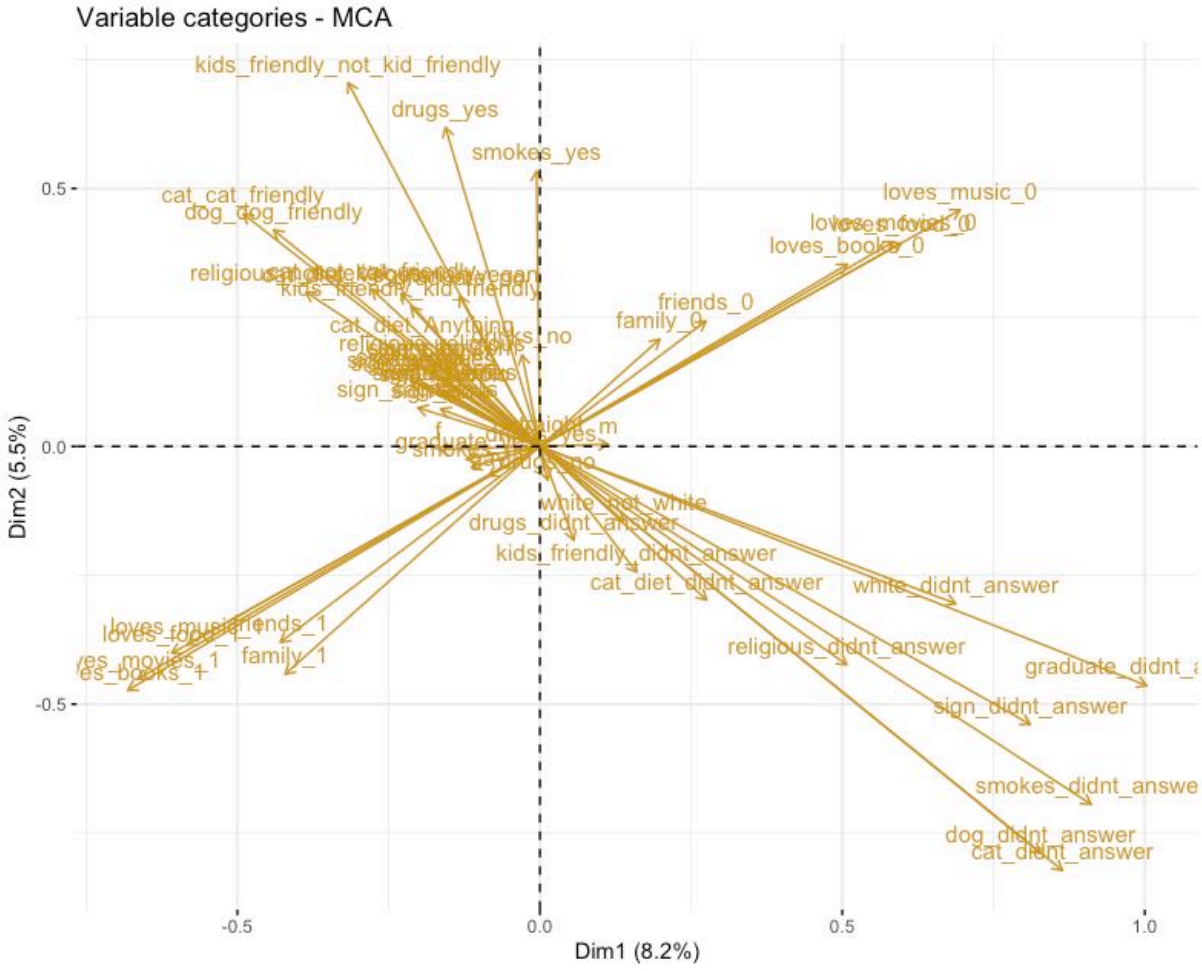
L'Analyse en Composantes Multiples (ACM) permet d'une part d'obtenir des variables continues à partir de variables catégorielles mais également d'analyser la position des individus quant-aux variables qualitatives.

Afin de représenter les individus sur les axes, l'ACM projette sur chaque axe un condensé d'information de chaque modalité. L'inertie mesure la dispersion entre les nuages de points. Chaque axe principal maximise cette inertie expliquée. Ainsi, le premier axe principal a une inertie expliquée de 8,24%, supérieure à l'inertie expliquée du second axe principal 5,49%. Afin d'analyser la position des individus sur les axes, nous pouvons observer par axe les modalités ayant le plus d'importance.



Les modalités “cat_didnt_answer” et “dog_didnt_answer”, étant le fait de ne pas avoir mentionné son avis sur les animaux de compagnie, sont les modalités ayant le plus d’importance sur la dimension 1.

Nous pouvons également analyser la position des variables sur les deux premiers axes :



La partie en bas à droite est constituée de toutes les modalités “l’individu n’a pas répondu”. La partie en haut à gauche présente les variables décrivant les personnes fumant, aimant les chats et les chiens, ou encore prenant des drogues. La partie en bas à gauche décrit les personnes pour qui la famille et les amis sont importants, et qui ont mentionné le fait d’aimer la musique, les livres, la nourriture et les films. La partie en haut à droite représente les individus ayant les caractéristiques opposées à ces derniers.

L’inertie expliquée par chaque dimension étant décroissante avec l’ajout de dimensions, nous ne conservons que 28 dimensions, représentant 80% d’inertie expliquée. Nous utilisons ces dimensions afin de créer une segmentation par l’algorithme des k-means et pour la prédiction des clusters par l’algorithme Random Forest.

B. Analyse des corrélations entre les dimensions

Dans cette partie, nous analysons les corrélations entre les variables continues et les dimensions formées précédemment grâce à l’ACM. Effectivement, lorsque deux variables sont corrélées, cela signifie qu’elles apportent la même information. Il n’est donc pas judicieux d’utiliser deux variables corrélées dans un algorithme de machine Learning. Plus le coefficient de corrélation est proche de 1, plus les variables sont corrélées positivement. A contrario, plus le coefficient de corrélation est proche de -1, plus les variables sont corrélées négativement. Finalement, aucune variable ne sont corrélées *(Graphique 8)*.

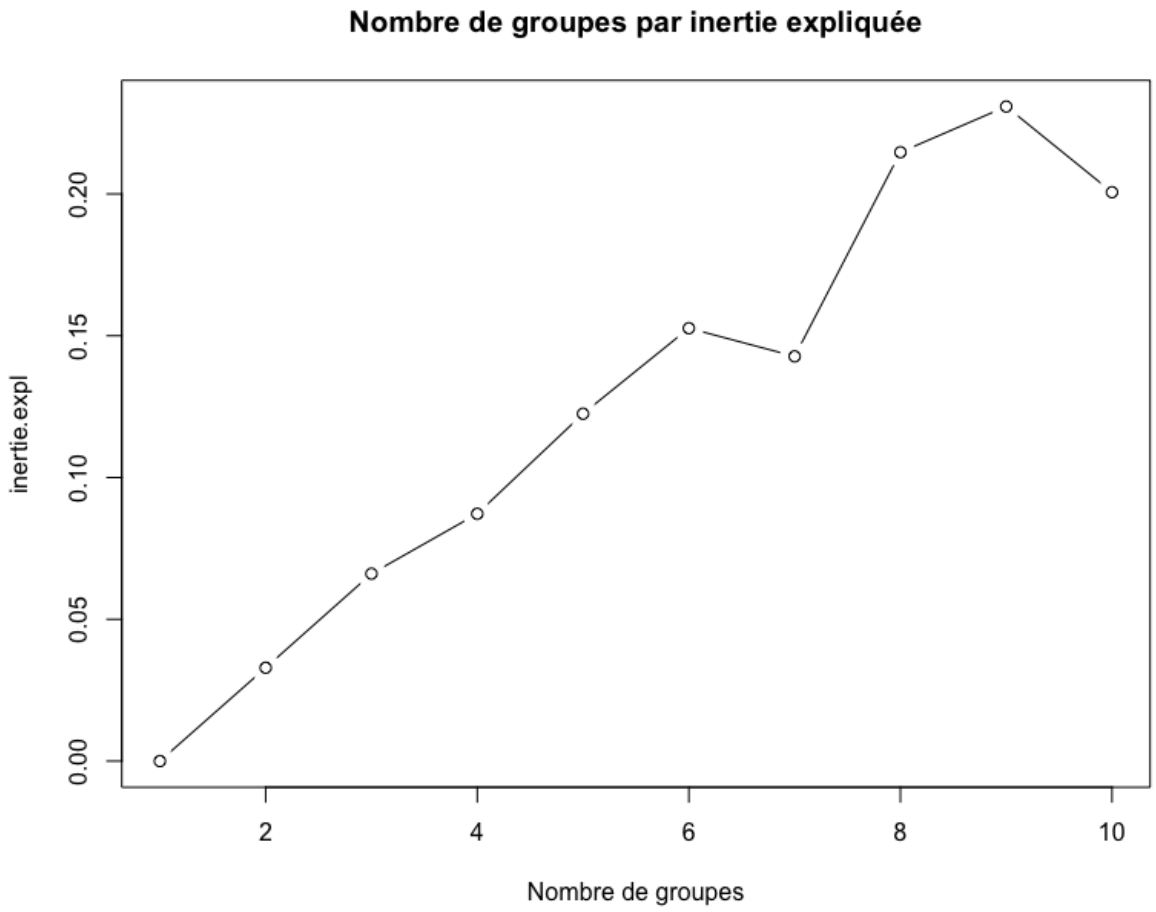
C. Segmentation des utilisateurs par l’algorithme des k-means

L’algorithme des k-means permet d'analyser une base de données, caractérisée par un ensemble de variables, afin de regrouper les données homogènes en clusters. Afin de créer les groupes d’individus, la méthode des k-means définit k centroïdes (centres de gravité de chaque nuage de point). Grâce à la minimisation de la distance euclidienne entre ces deux points, à chaque itération, l’observation est associée à son centroid le plus proche. Cette étape est répétée jusqu'à obtenir des groupes d’individus. Les distances inter-classe et intra-classe entre les clusters sont donc respectivement maximisées et minimisées.

La première étape consiste à sélectionner le nombre de clusters noté k. Afin de sélectionner le bon nombre de clusters, nous observons l’inertie expliquée. L’inertie expliquée consiste à diviser l’inertie inter-classe par l’inertie intra-classe. L’inertie inter-classe est une mesure de la séparation des classes, elle doit donc être maximisée afin que chaque classe diffère des

autres. L'inertie intra-classe doit être minimisée afin de garantir la similarité des observations au sein de chaque cluster.

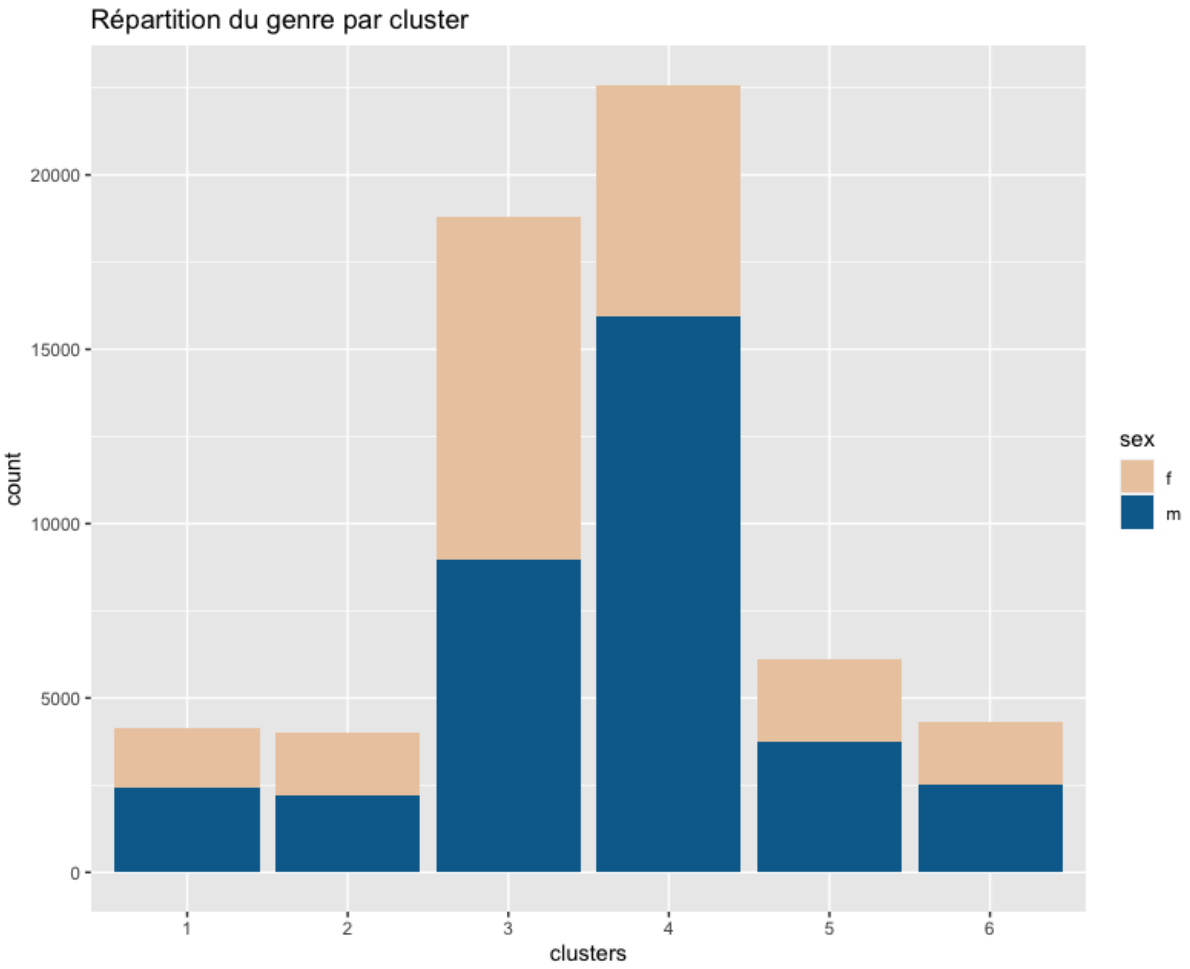
Nous observons ainsi l'inertie expliquée en fonction du nombre de clusters sur le graphique ci-dessous :



A partir de 6 clusters, l'inertie expliquée est de 14.1%. Nous pourrions avoir une inertie plus importante en prenant 9 clusters, mais le gain d'inertie comparativement à la complexité d'interprétation étant négligeable, nous sélectionnons 6 clusters. Nous obtenons donc 6 clusters regroupant respectivement : 4306, 4023, 3573, 23861, 16387 et 7796 individus.

Passons à présent à l'analyse descriptive des clusters. Dans le corps du texte, nous analysons cinq variables qui permettent de discriminer les clusters. Le croisement des classes avec les autres variables sont présentés en annexe *(Graphique 9 à 10)*.

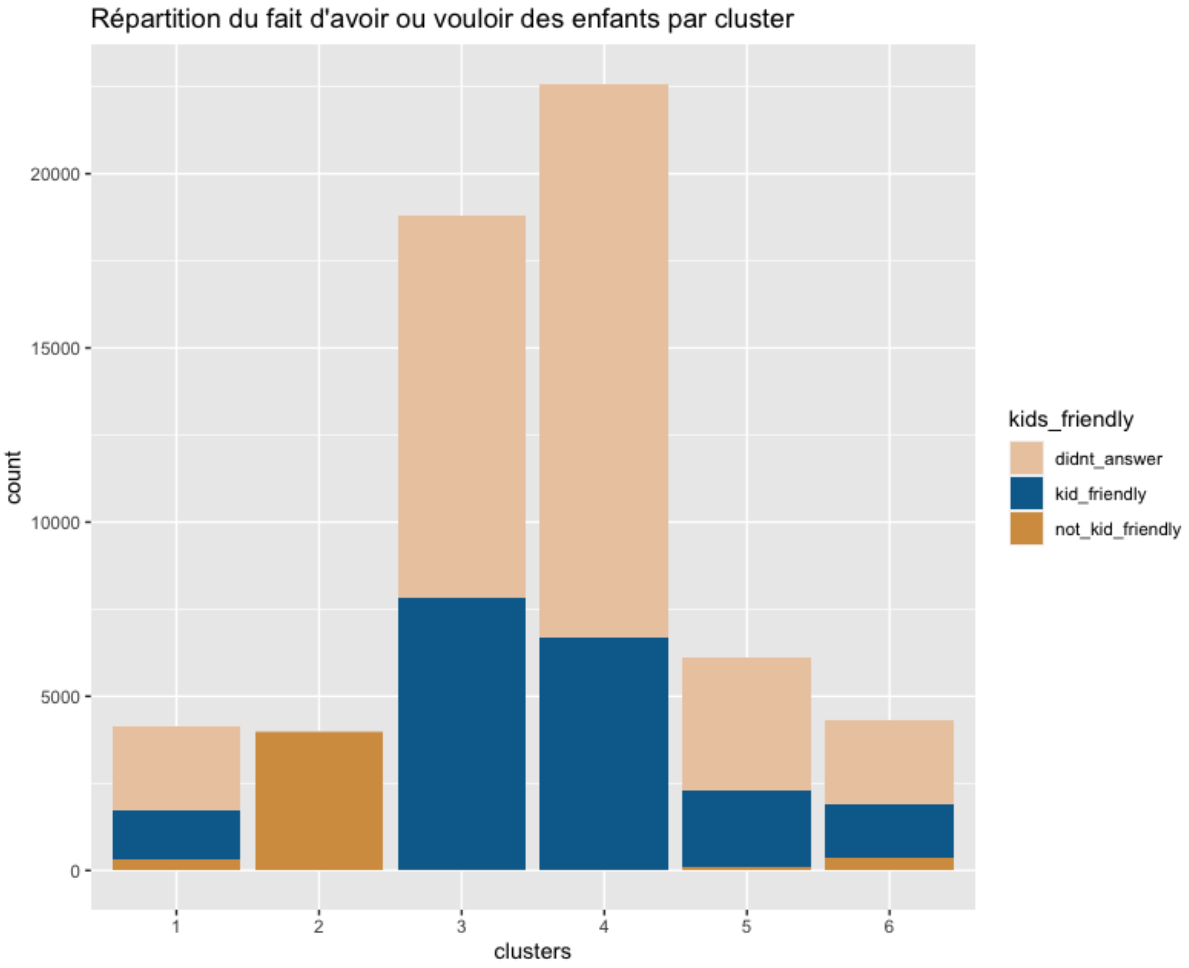
Nous observons d'abord la distribution du genre par cluster. Notre but est d'avoir environ autant d'hommes que de femmes dans chaque cluster, ayant une majorité de personnes hétérosexuelles.



Il y a une légère majorité de femmes dans le cluster 3, mais cette règle est globalement bien respectée dans tous les clusters.

On observe à présent une variable particulièrement discriminante, "kid_friendly", correspondant au fait d'avoir des enfants ou de vouloir des enfants

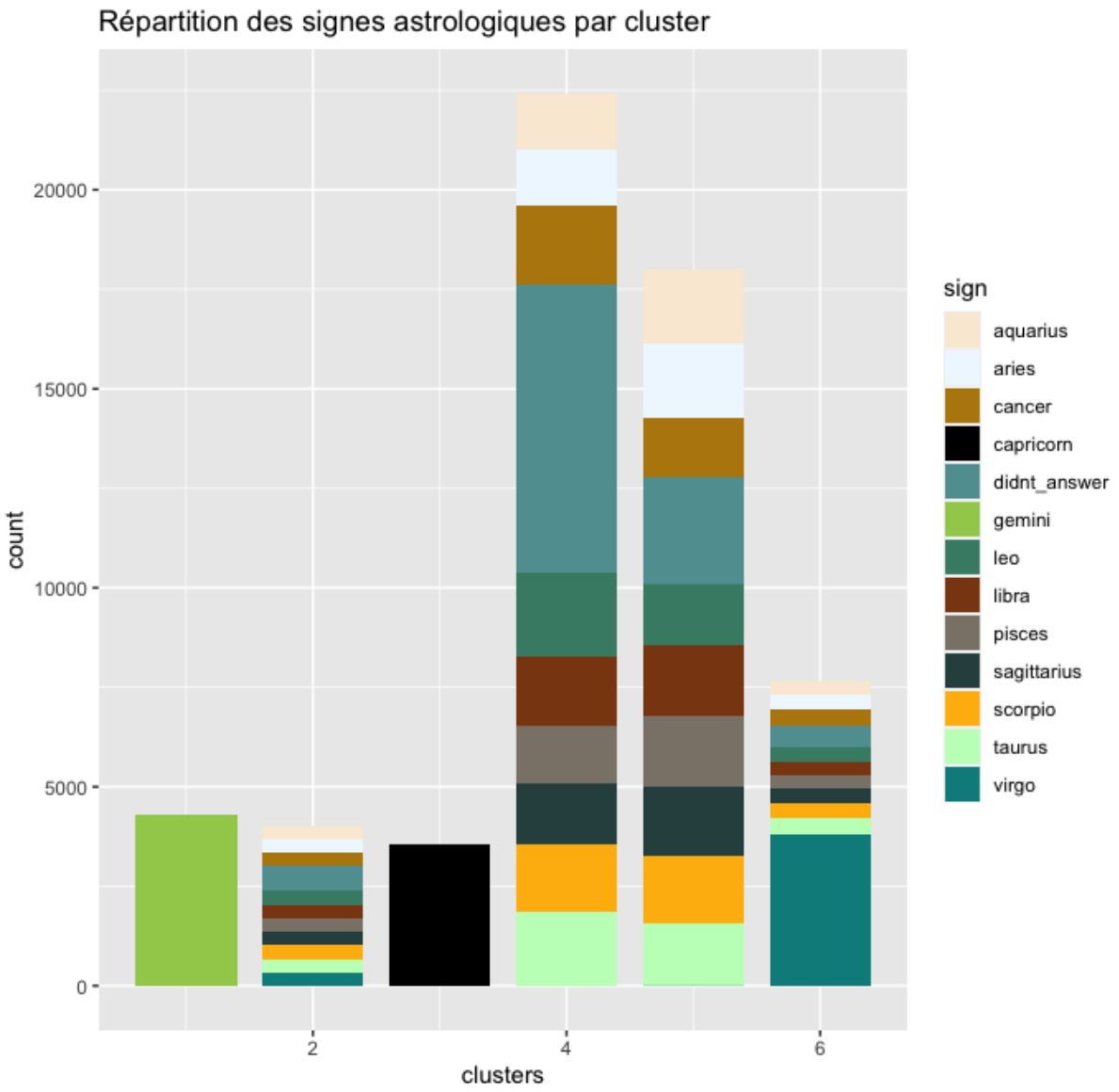
:



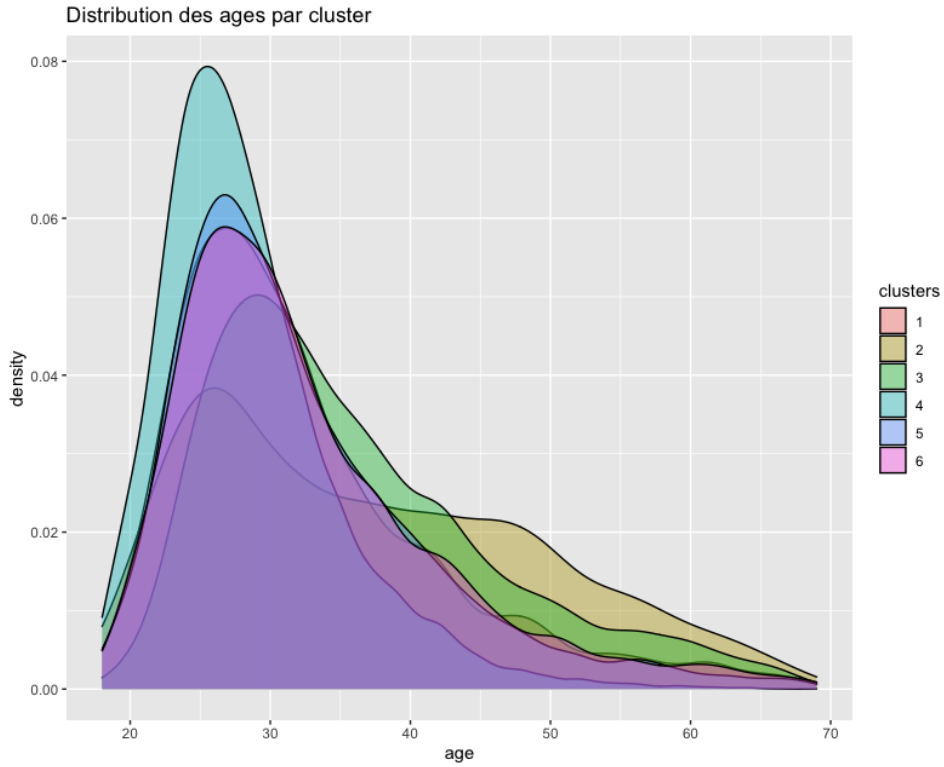
Le cluster 2 est uniquement constitué de personnes ne désirant pas avoir d'enfants, tandis que les clusters 3 et 4 sont constitués de personnes n'ayant pas répondu ou ayant exprimé ce désir. Finalement, les clusters 1, 5 et 6 sont constitués d'une minorité de personnes ayant des enfants ou désirant avoir des enfants.

Une autre variable particulièrement discriminante est liée au signe astrologique de l'individu :

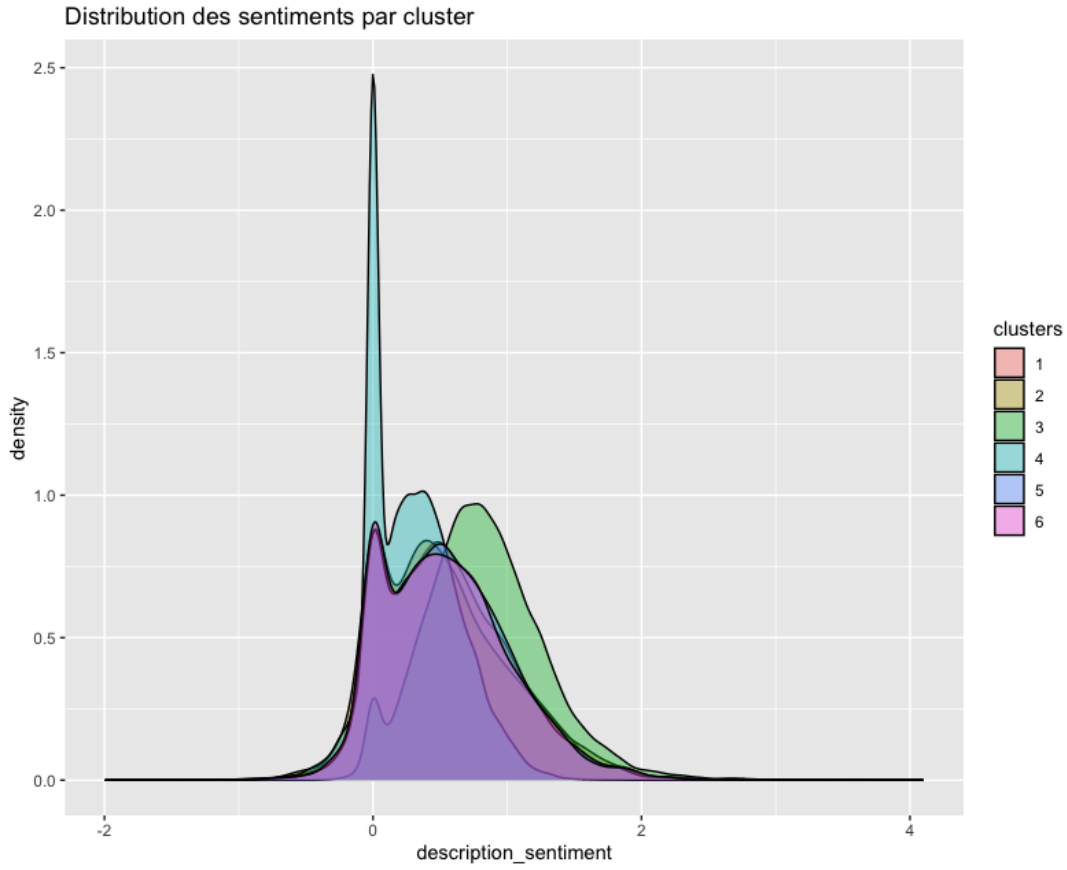
Le cluster 1 est uniquement constitué de gémeaux et le cluster 3 de capricorne. Le cluster 6 est constitué en majorité de vierges.



Nous observons également la distribution des âges dans chaque cluster sur le graphique situé à la page suivante. Les utilisateurs ont majoritairement entre 20 et 35 ans. Le cluster 4 est le cluster regroupant le plus de personnes jeunes et le cluster 2 regroupe le plus de personnes ayant plus de 45 ans.



Finalement, nous observons que les personnes du clusters 4 donnent une description majoritairement neutre. Par extension, les personnes du cluster 3 se décrivent globalement positivement.

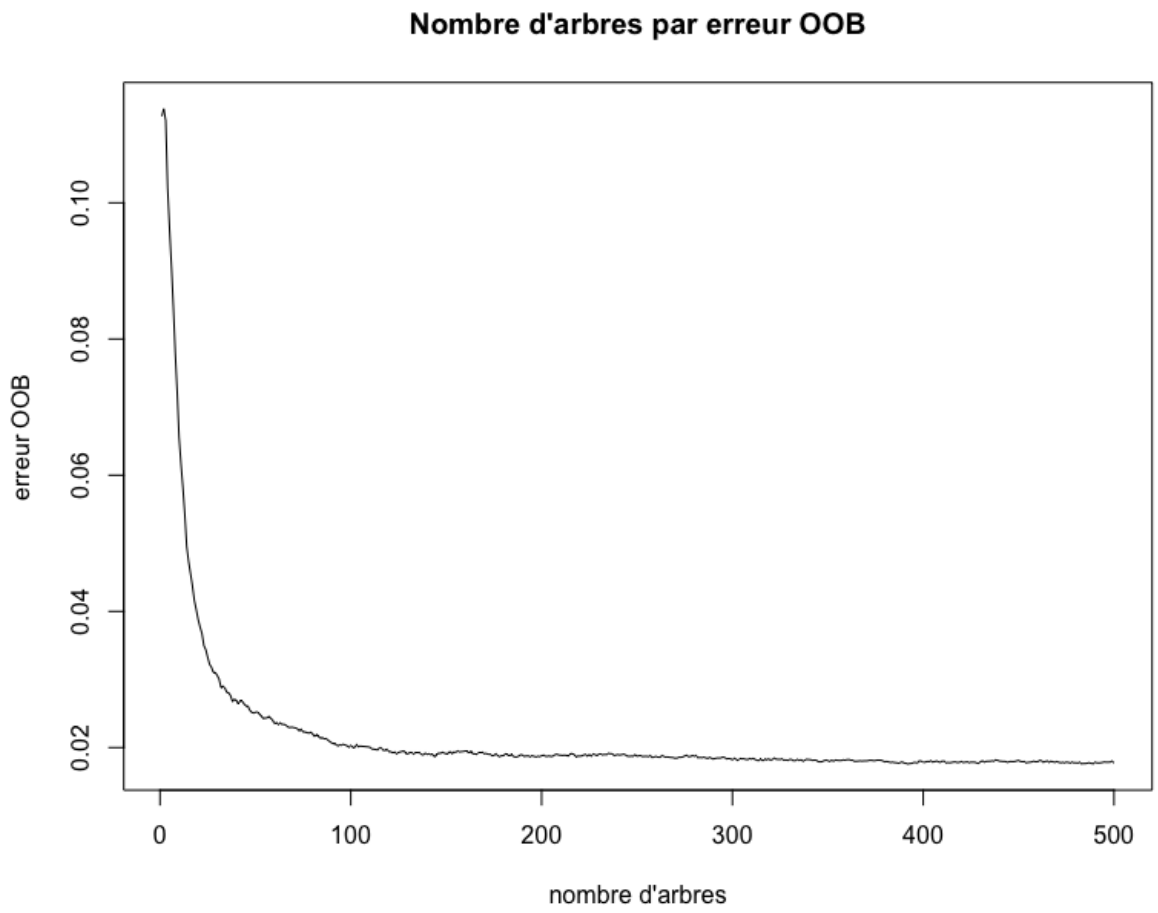


D. L'algorithme Random Forest: prédiction des clusters

Afin de prédire l'appartenance à chaque cluster, nous utilisons finalement un modèle de classification Random Forest. Le principe de la Random Forest est de créer aléatoirement plusieurs arbres de classification entraînés sur des sous-ensembles de données avec des sous-ensembles aléatoires de variables. Les arbres de décision créent une hiérarchie de classement des observations selon différents critères. Pour chaque échantillon bootstrap, nous obtenons les estimateurs ainsi que leur espérance et leur variance. Lorsque nous agrégeons ces estimations, le biais n'est pas modifié cependant, la variance baisse car elle est divisée par le nombre d'échantillons bootstrap. La variance étant plus faible, le modèle est donc généralisable sur un nouveau jeu de données. La moyennisation des arbres permet donc de réduire la variance et d'améliorer les performances prédictives par rapport à un arbre de classification simple.

Afin de procéder à l'algorithme Random Forest, nous divisons premièrement notre base de données en deux échantillons : 60% et 40% des observations sont respectivement alloués à l'entraînement et au test du modèle.

La première étape de la Random Forest consiste à déterminer le nombre d'arbres dans la forêt. Pour ce faire, nous observons l'erreur OOB par rapport au nombre d'arbres. L'erreur OOB décrivant l'erreur "Out-of-Bag" correspond à l'estimation acceptable du taux d'erreur théorique obtenu grâce au bootstrapping.



Nous sélectionnons le nombre d'arbres à partir duquel l'erreur OOB se stabilise,et choisissons ainsi 400 arbres.

Nous procédons par la suite à la sélection du nombre optimal de "mtry" qui définit le nombre de variables échantillonnées aléatoirement comme candidates à chaque division. Il existe deux méthodes pour sélectionner le "mtry" : le faire varier entre 0 et 10, ou garder la valeur par défaut de R, et faire un modèle en divisant cette valeur par défaut, et un autre en la doublant. Nous sélectionnons cette seconde option car elle permet de voir si l'estimation est améliorée en réduisant ou agrandissant cette valeur par défaut. Le nombre de "mtry" minimisant l'estimation par OOB du taux d'erreur est 4.

Sur le jeu de données d'entraînement, nous obtenons la matrice de confusion ci-dessous, ainsi qu'une accuracy de 94.39%, ce qui signifie que dans 5,61% des cas les individus sont mal classés :

	1	2	3	4	5	6	class.error
1	2630	0	0	0	0	0	0.000000000
2	0	2405	0	7	0	5	0.004964832
3	0	0	2095	0	0	0	0.000000000
4	0	0	0	12750	737	18	0.055905220
5	0	1	0	1057	9642	58	0.103736754
6	0	5	0	95	62	4400	0.035510741

Les clusters 1 et 3 sont parfaitement estimés. Avec un taux d'erreur OOB de 10%, le cluster 5 possède le plus d'erreurs : 1057 et 58 personnes ayant le label 5 sont respectivement classées dans le cluster 4 et 6 par erreur.

Sur l'échantillon test, l'accuracy est de 94.67%, ce qui signifie que dans moins de 6% des cas les individus sont mal classés. Nous obtenons la matrice de confusion suivantes :

Prediction	1	2	3	4	5	6
1	1678	2	0	0	0	0
2	0	1593	0	0	0	8
3	0	0	1478	0	0	0
4	0	6	0	8478	676	58
5	0	0	0	441	6514	31
6	0	8	0	9	40	2959

Les individus appartenant à la classe 3 sont toujours bien classés. Deux individus provenant du cluster 1 sont classés dans le cluster 2. 8 individus du cluster 2 sont classés dans le cluster 6. Le cluster 5 possède un grand nombre d'erreurs : 441 et 31 personnes ayant le label 5 sont respectivement classées dans le cluster 4 et 6 par erreur. Les résultats sont fortement semblables à ceux observés lors de l'entraînement du modèle. Les résultats détaillés des performances prédictives sont présentés en annexe *(Tableau 25)*.

L’accuracy du test fixée à 94.67%, étant supérieure à celle du train (94.39%), indique l’absence de sur-apprentissage, signifiant que ce modèle a une bonne capacité d’ajustement en plus d’un très bon pouvoir prédictif.

CONCLUSION :

L’augmentation des rencontres sur internet entraînée par l’essor des applications de rencontre et des réseaux sociaux constitue une nouvelle demande des utilisateurs. En effet, l’enjeu principal pour les sites de rencontre comme Okcupid est de connaître et comprendre les utilisateurs afin de répondre à leurs besoins.

Dans cette étude, nous utilisons une base de données du site Kaggle, regroupant les utilisateurs de Okcupid. Dans un premier temps, cette base de données est analysée afin de connaître les caractéristiques des individus.

Également, partant de l’hypothèse que les personnes avec des caractéristiques similaires seraient davantage compatibles, nous segmentons cette population en 6 groupes d’individus homogènes par l'algorithme des k-means. A titre d’exemple, le fait de vouloir et/ou d’avoir des enfants constitue un axe de similarité intéressant entre les utilisateurs. Ce regroupement permet de proposer à un individu de matcher avec une personne de son groupe parent. Les nouveaux utilisateurs se voient également attribuer un groupe grâce à l’algorithme de Random Forest.

Cette analyse est cependant discutable car il est courant que certaines personnes matchent avec des personnes qui ne leur ressemblent pas nécessairement. Il serait donc intéressant d’obtenir les données concernant les paires de match afin de proposer un algorithme de recommandation.

SITOGRAPHIE

BUCHHOLZ, Katharina. «*How Couples met*». Statistica, 13/02/2020
• Chart: How Couples Met | Statista

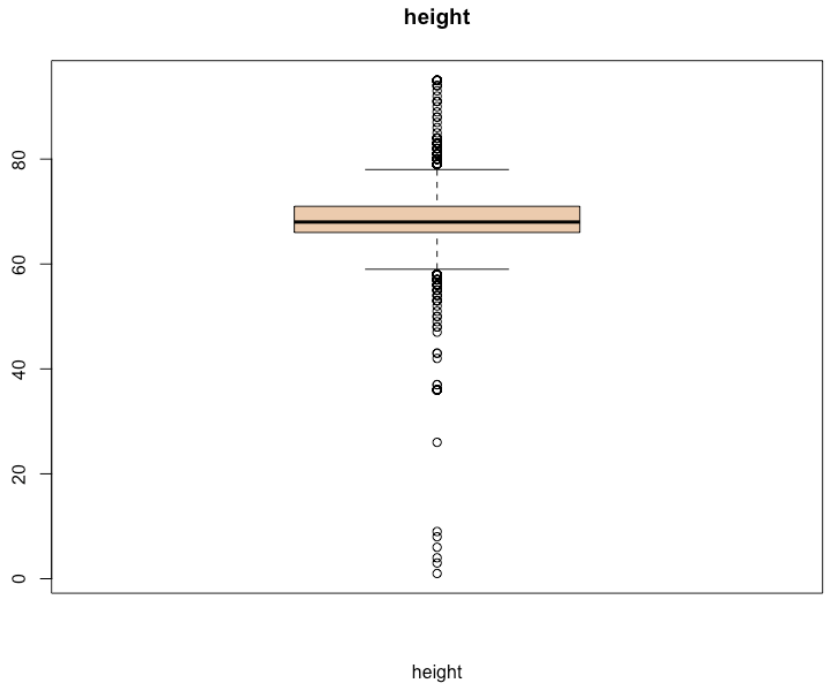
RINKER, Tyler. "sentimentr". Github, 20/10/2018
https://github.com/trinker/sentimentr#examples

ANNEXE

Tableau 1: *Pourcentage de valeurs manquantes par variable*

age	0%	job	13.68%
status	0%	last_online	0%
sex	0%	location	0%
orientation	0%	offspring	59.32%
body_type	8.83%	pets	33.23%
diet	40.69%	religion	33.74%
drinks	4.98%	sign	18.44%
drugs	23.48%	smokes	9.19%
education	11.06%	speaks	0.008%
ethnicity	9.48%	essay0	9.15%
height	0.005%	essay1	12.53%
income	0%	essay2	16.08%
essay3	19.14%	essay4	17.58%
essay5	18.10%	essay6	22.97%
essay7	20.77	essay8	32.07%
essay9	21.02%		

Graphique 1: *Boxplot de la taille des individus*



Graphique 2: *Boxplot des revenus des individus*

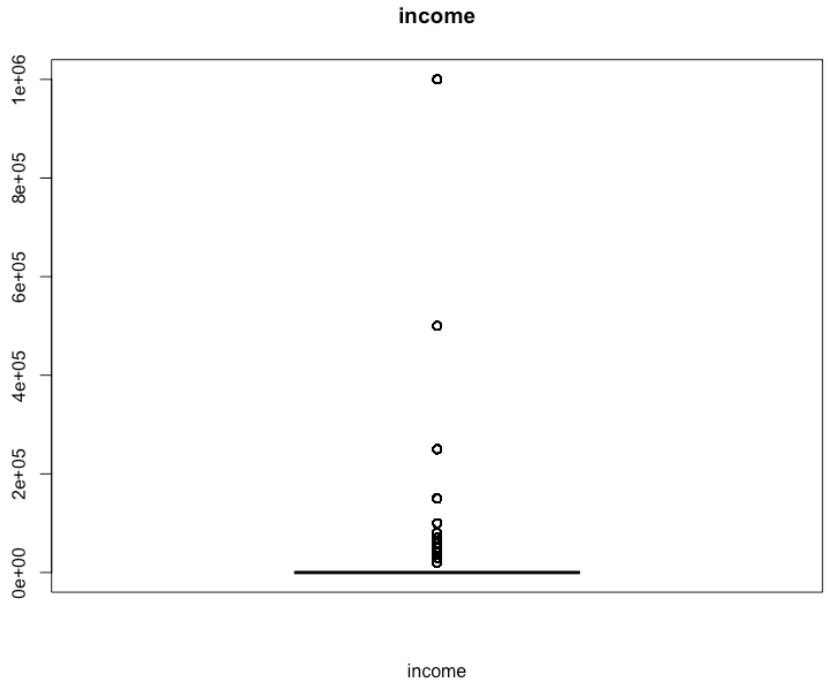


Tableau 2: *Comptage des modalités pour la variable "status"*

available	3,11%
married	0,52%
seeing someone	3,44%
single	92,91%
unknown	0,02%

Tableau 3: *Comptage des modalités pour la variable "body_type"*

a little extra	4,39%
athletic	19,72%
average	24,44%
curvy	6,55%
fit	21,20%
full figured	1,68%
jacked	0,70%
overweight	0,74%
rather not say	0,33%
skinny	2,96%
thin	7,86%
used up	0,59%
didnt_answer	8,83%

Tableau 4: *Comptage des modalités pour la variable "job"*

artistic / musical / writer	7,40%
banking / financial / real estate	3,78%
clerical / administrative	1,34%
computer / hardware / software	7,86%
construction / craftsmanship	1,70%
didnt_answer	13,68%
education / academia	5,86%
entertainment / media	3,75%
executive / management	3,96%
hospitality / travel	2,28%
law / legal services	2,30%
medicine / health	6,14%
military	0,34%
other	12,66%
political / government	1,18%
rather not say	0,73%
retired	0,42%
sales / marketing / biz dev	7,32%

science / tech / engineering	8,09%
student	8,14%
transportation	0,61%
unemployed	0,46%

Tableau 5: *Comptage des modalités pour la variable "sex"*

f	40,23%
m	59,77%

Tableau 6: *Comptage des modalités pour la variable "orientation"*

bisexual	4,62%
gay	9,30%
straight	86,09%

Tableau 7: *Comptage des modalités pour la variable "drinks"*

didnt_answer	4,98%
no	5,45%
yes	89,57%

Tableau 8: *Comptage des modalités pour la variable "drug"*

didnt_answer	23,49%
no	62,93%
yes	13,58%

Tableau 9: *Comptage des modalités pour la variable "smokes"*

didnt_answer	9,19%
no	73,23%
yes	17,58%

Tableau 10: *Comptage des modalités pour la variable "graduated"*

didnt_answer	11,06%
no	23,26%
yes	65,69%

Tableau 11: *Comptage des modalités pour la variable "dog"*

didnt_answer	33,23%

dog_friendly	62,45%
not_dog_friendly	4,32%

Tableau 12: *Comptage des modalités pour la variable "cat"*

cat_friendly	47,75%
didnt_answer	33,23%
not_cat_friendly	19,02%

Tableau 13: *Comptage des modalités pour la variable "kids_friendly"*

didnt_answer	59,32%
kid_friendly	32,71%
not_kid_friendly	7,97%

Tableau 14: *Comptage des modalités pour la variable "white"*

didnt_answer	9,48%
not_white	27,33%
white	63,19%

Tableau 15: *Comptage des modalités pour la variable "cat_diet"*

Anything	46,51%
didnt_answer	40,69%
other	3,31%
Vegetarian/vegan	9,49%

Tableau 16: *Comptage des modalités pour la variable "religious"*

didnt_answer	33,74%
not_religious	25,54%
religious	40,72%

Tableau 17: *Comptage des modalités pour la variable "sign"*

aquarius	6,55%
aries	6,65%
cancer	7,02%
capricorn	5,96%
didnt_answer	18,44%
gemini	7,19%
leo	7,30%
libra	7,02%
pisces	6,58%

sagittarius	6,58%
scorpio	6,90%
taurus	6,91%
virgo	6,91%

Tableau 18: *Statistiques descriptives de la variable "description_sentiment"*

Min	Q1	Médiane	Moyenne	Q3	Max
-1.9894	0.1637	0.4917	0.5358	0.8286	4.0947

Tableau 19: *Comptage des modalités pour la variable "loves_music"*

0	45,62%
1	54,38%

Tableau 20: *Comptage des modalités pour la variable "loves_movies"*

0	53,21%
1	46,79%

Tableau 21: *Comptage des modalités pour la variable "loves_food"*

0	50,45%
1	49,55%

Tableau 22: *Comptage des modalités pour la variable "loves_book"*

0	57,27%
1	42,73%

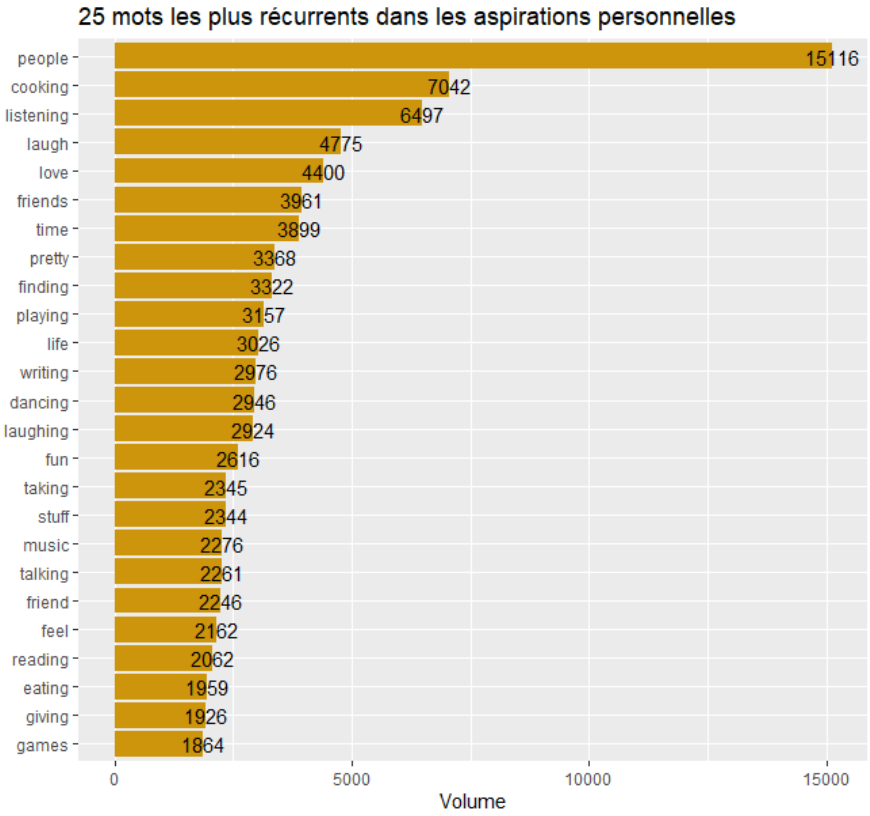
Tableau 23: *Comptage des modalités pour la variable "friend"*

0	60,98%
1	39,02%

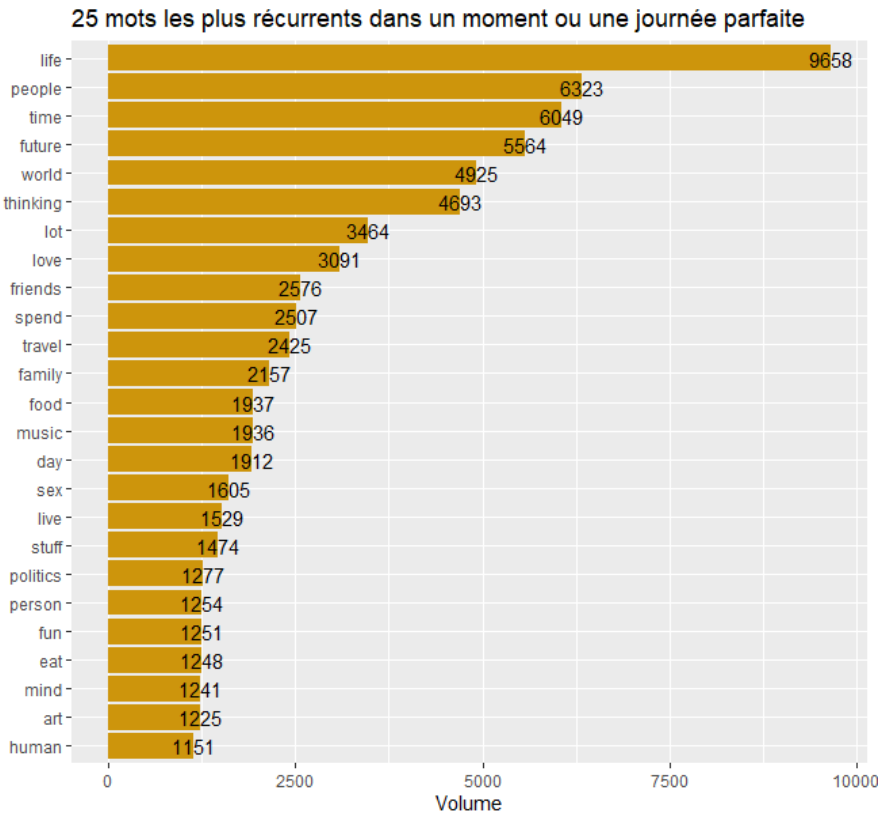
Tableau 24: *Comptage des modalités pour la variable "family"*

0	68,01%
1	31,99%

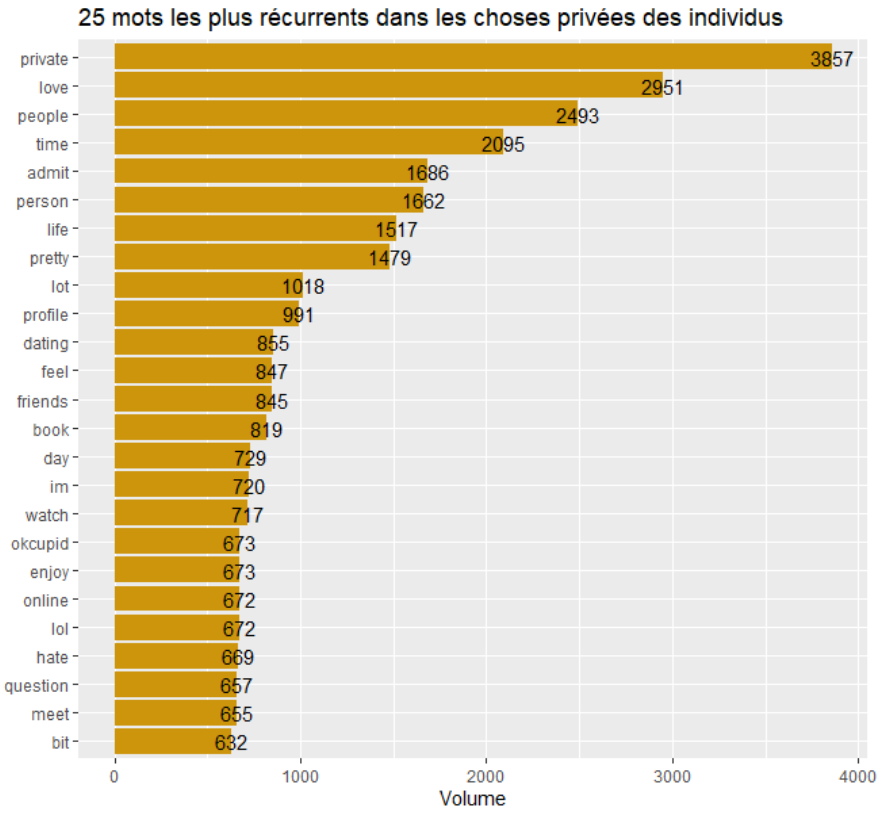
Graphique 3: *Nombre d'occurrences des mots au sein de la variable "essay2"*



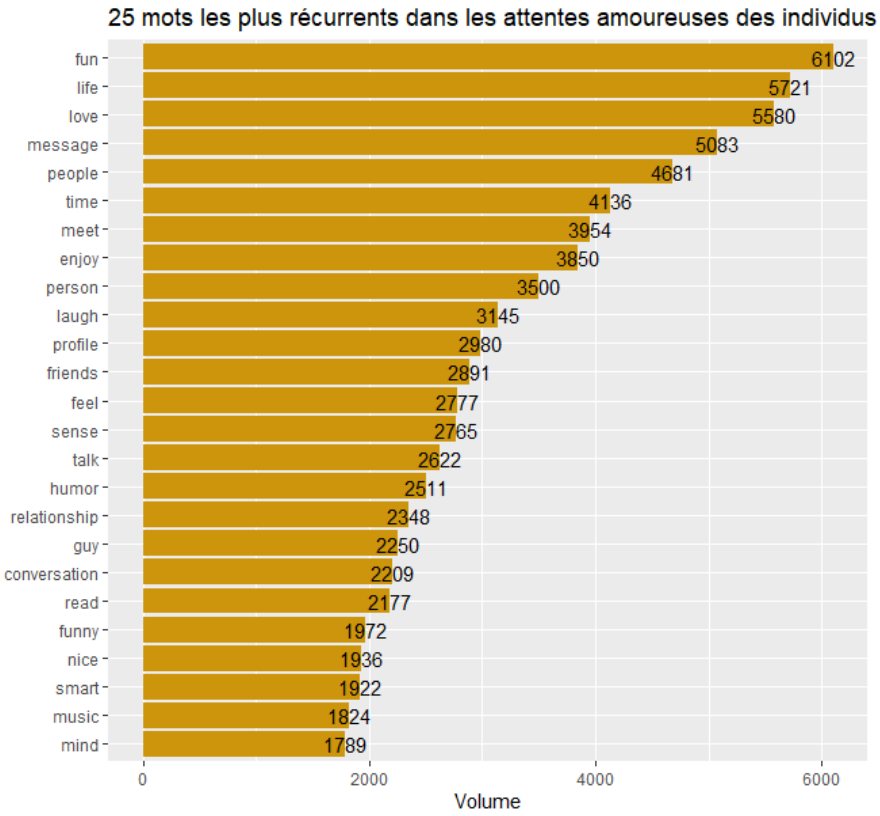
Graphique 4: *Nombre d'occurrences des mots au sein de la variable "essay6"*



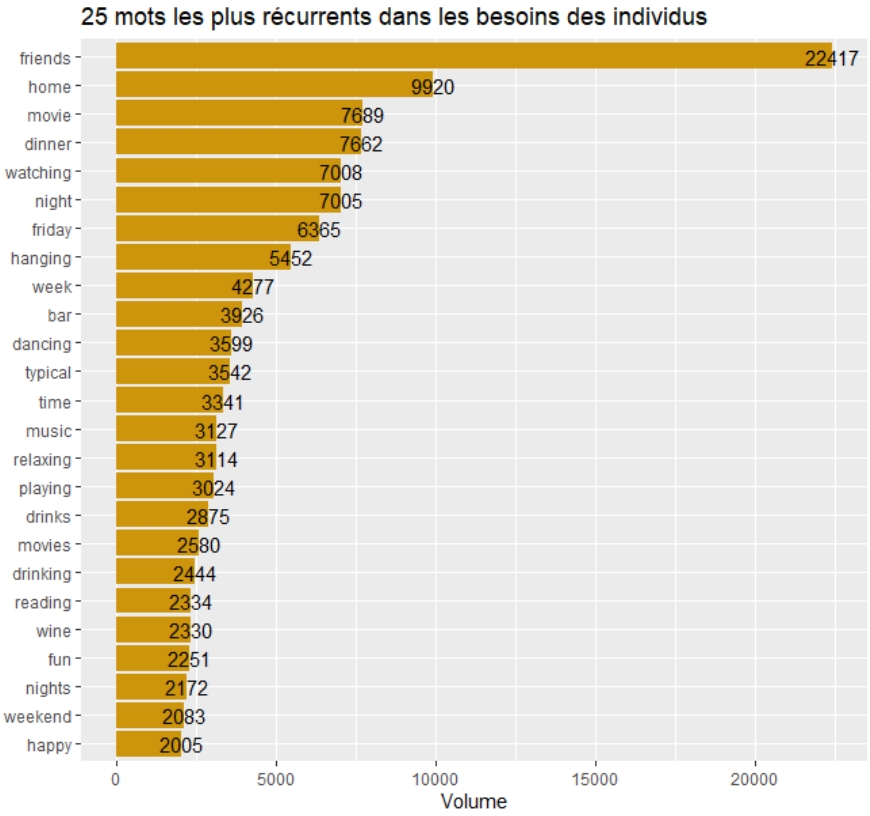
Graphique 5: *Nombre d'occurrences des mots au sein de la variable "essay8"*



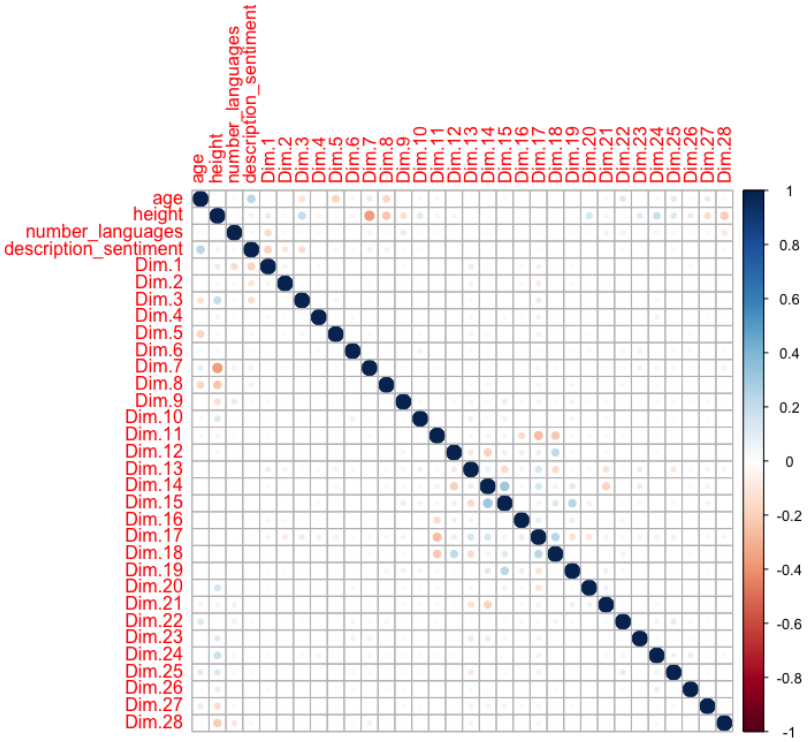
Graphique 6: *Nombre d'occurrences des mots au sein de la variable "essay9"*



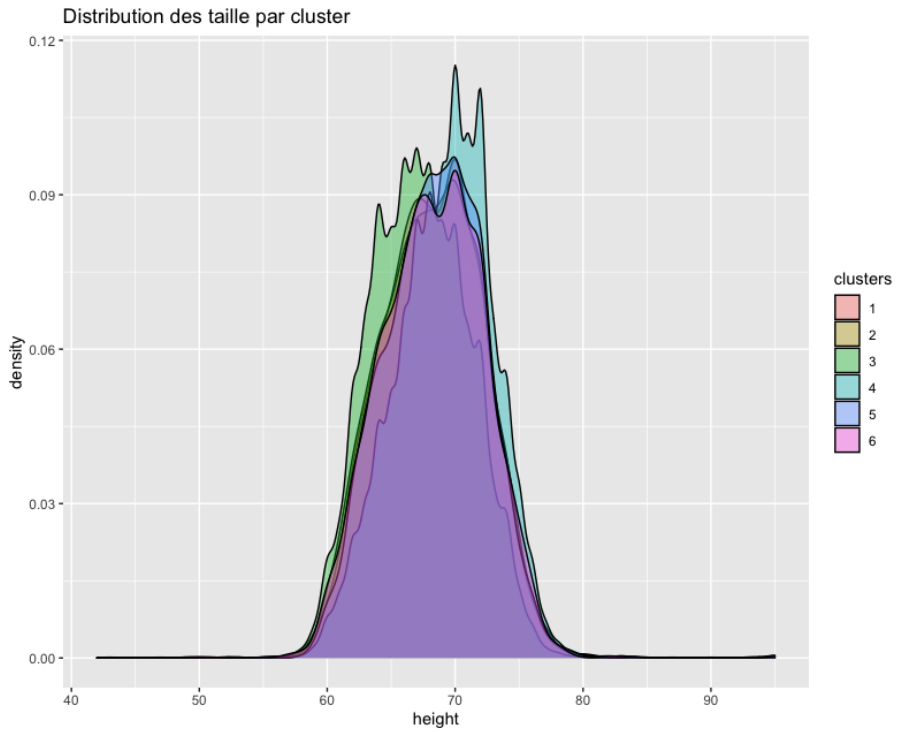
Graphique 7: *Nombre d'occurrences des mots au sein de la variable "essay7"*



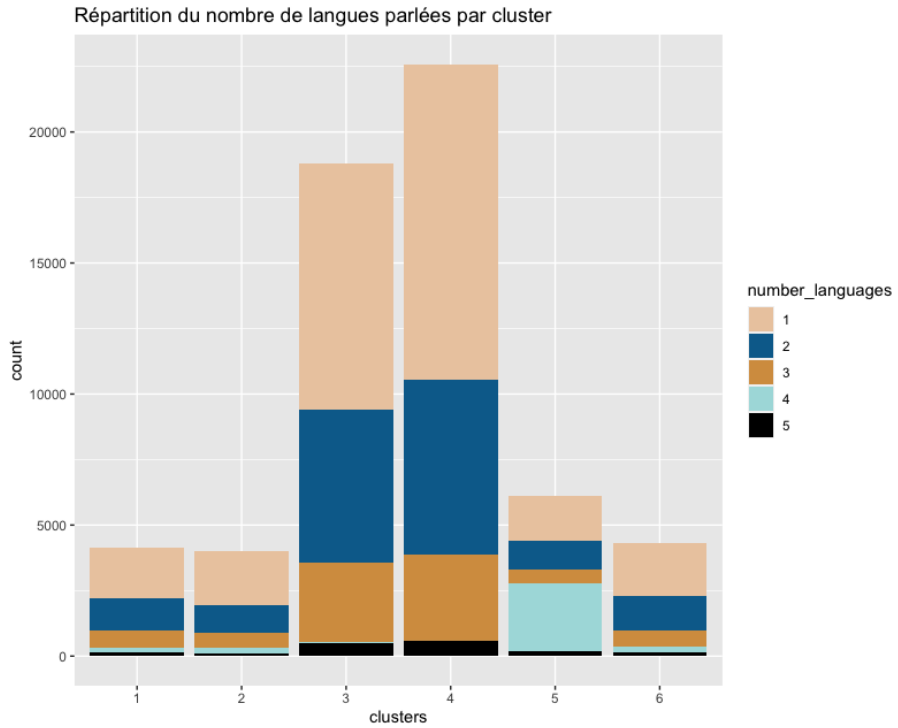
Graphique 8: *Heatmap des corrélations entre les variables continues et les dimensions*



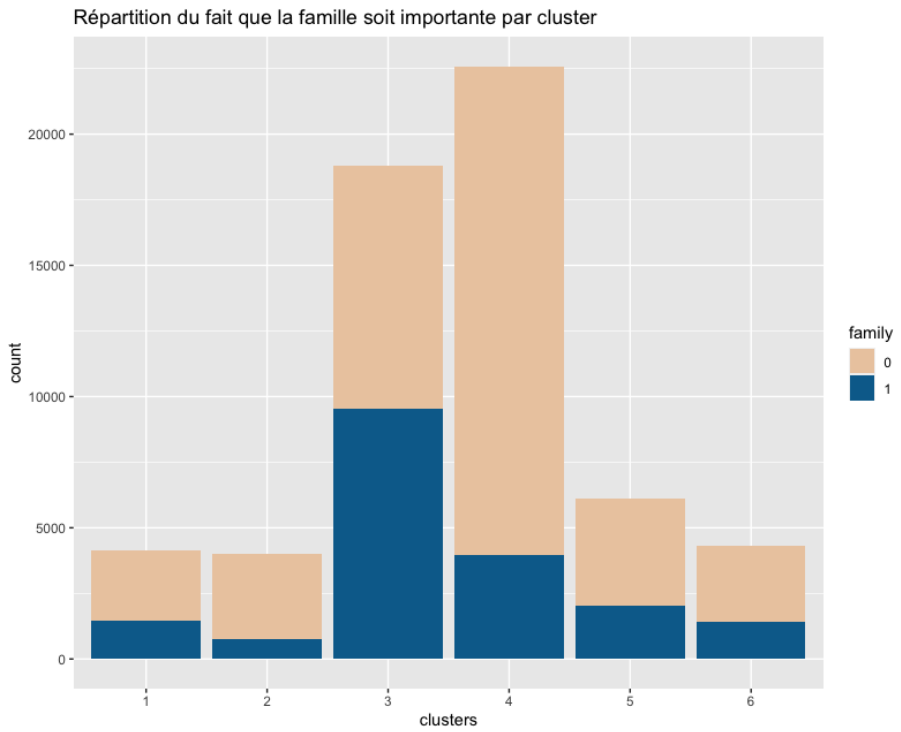
Graphique 9: *Distribution des tailles par cluster*



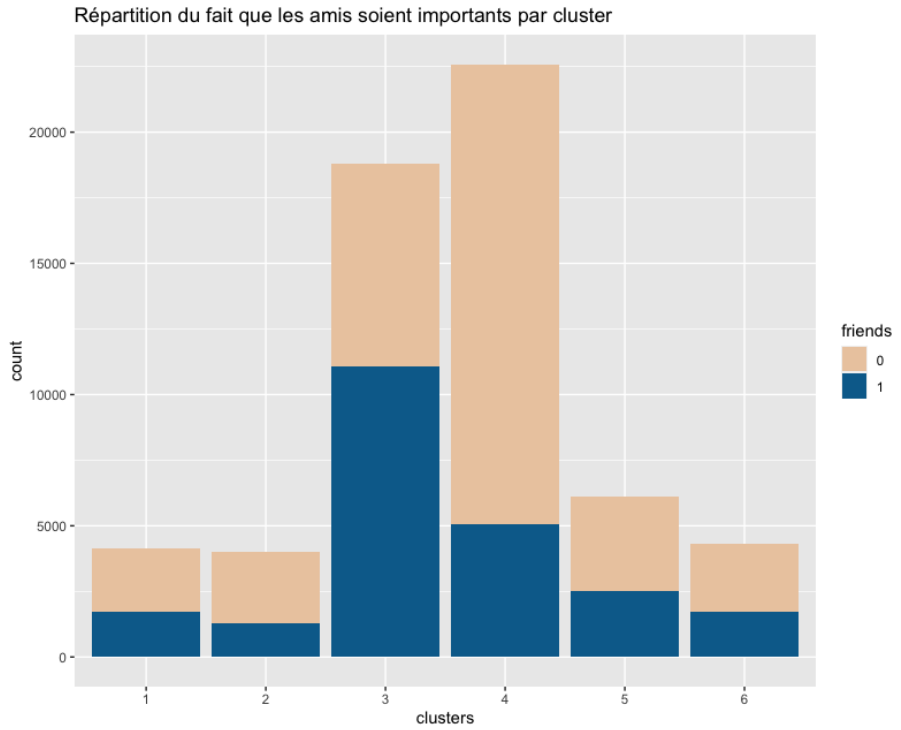
Graphique 10: *Répartition du nombre de langues parlées par cluster*



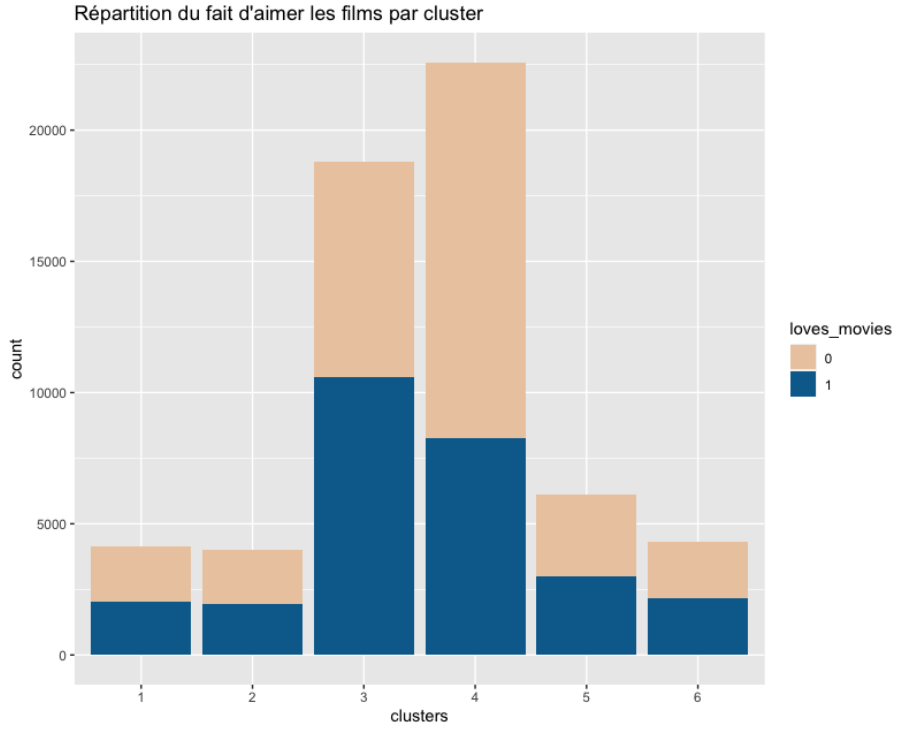
Graphique 11: *Répartition du fait que la famille soit importante par cluster*



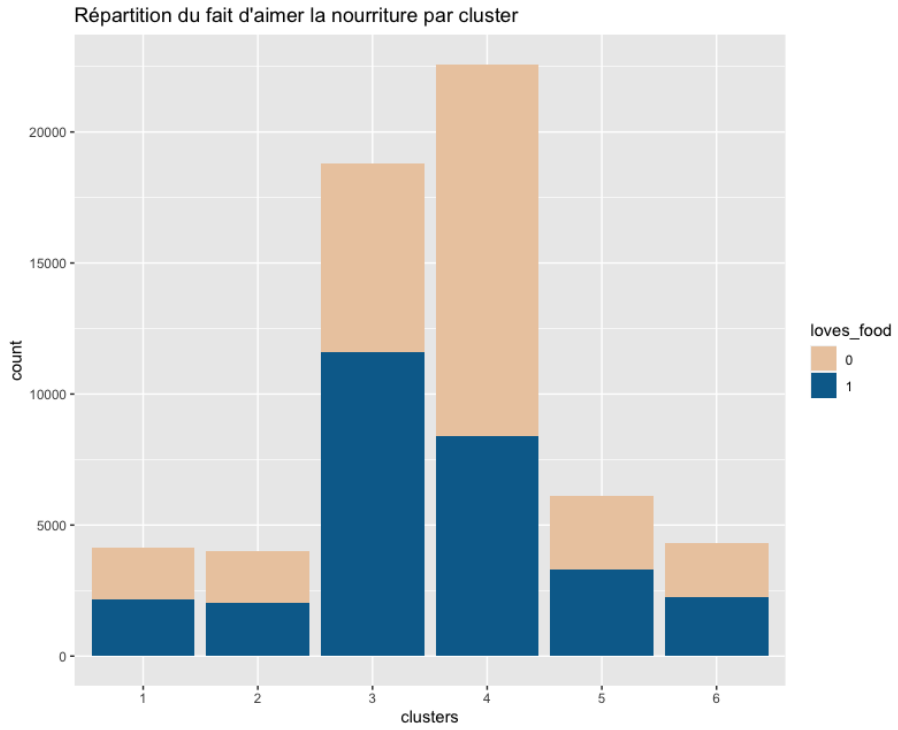
Graphique 12: *Répartition du fait que les amis soient importants par cluster*



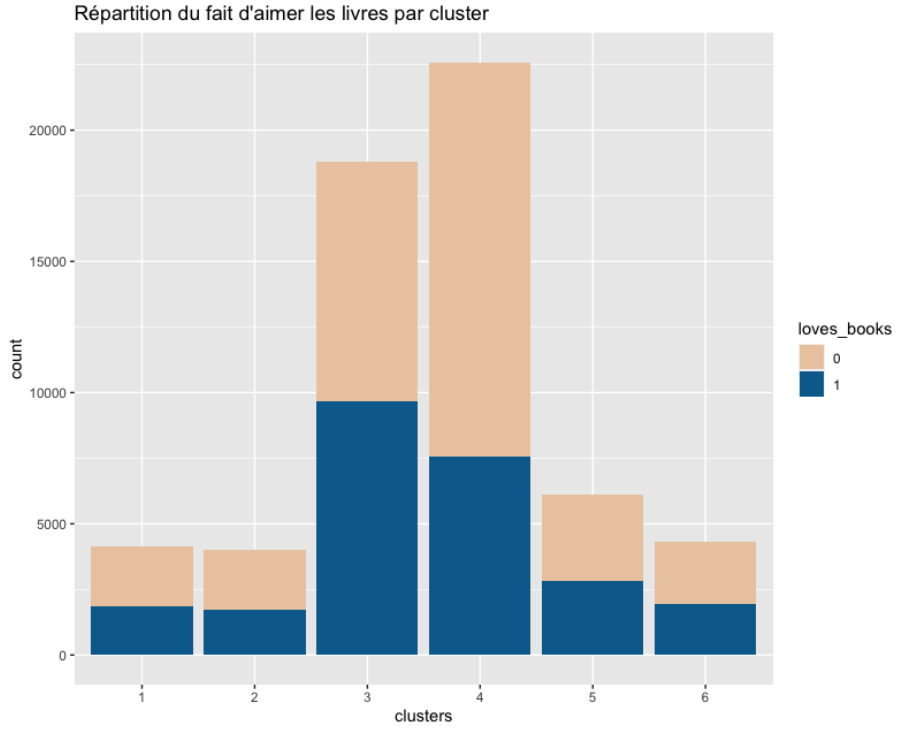
Graphique 13: *Répartition du fait d'aimer les films par cluster*



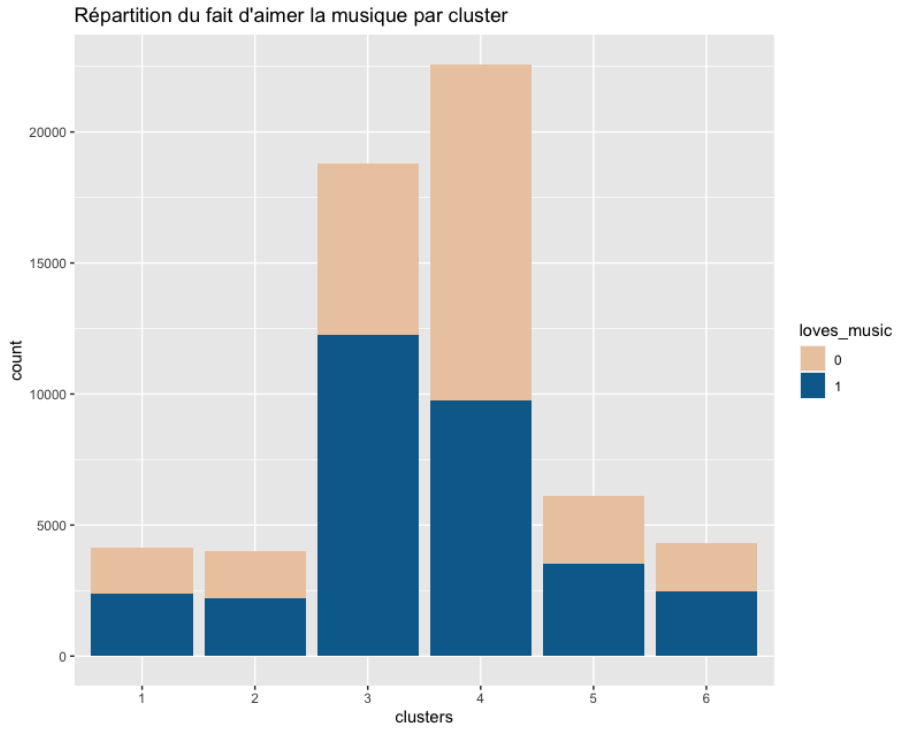
Graphique 14: *Répartition du fait d'aimer la nourriture par cluster*



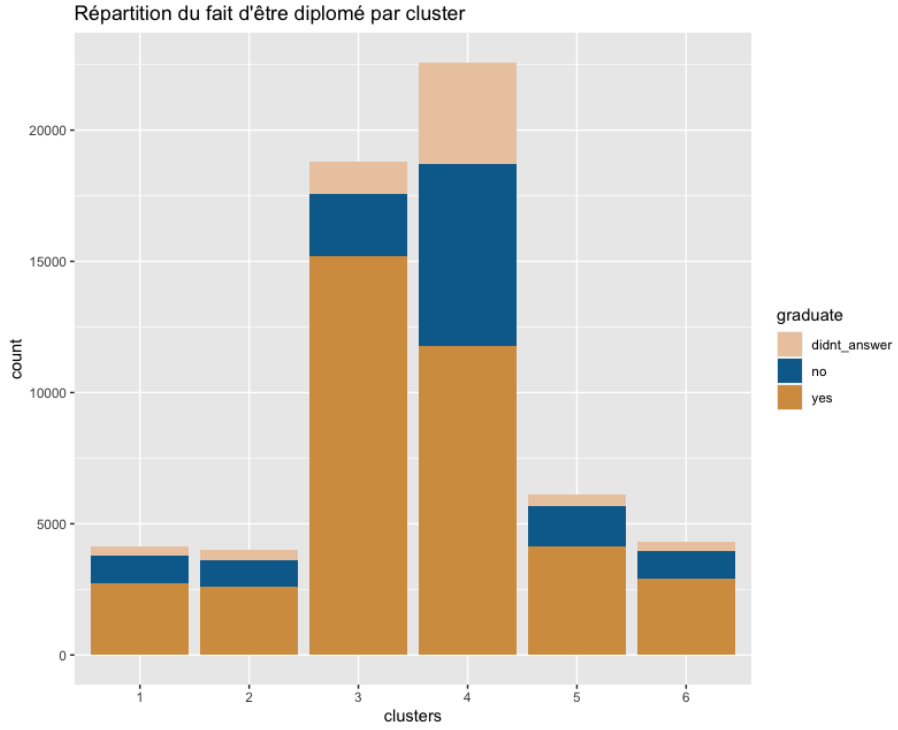
Graphique 15: *Répartition du fait d'aimer les livres par cluster*



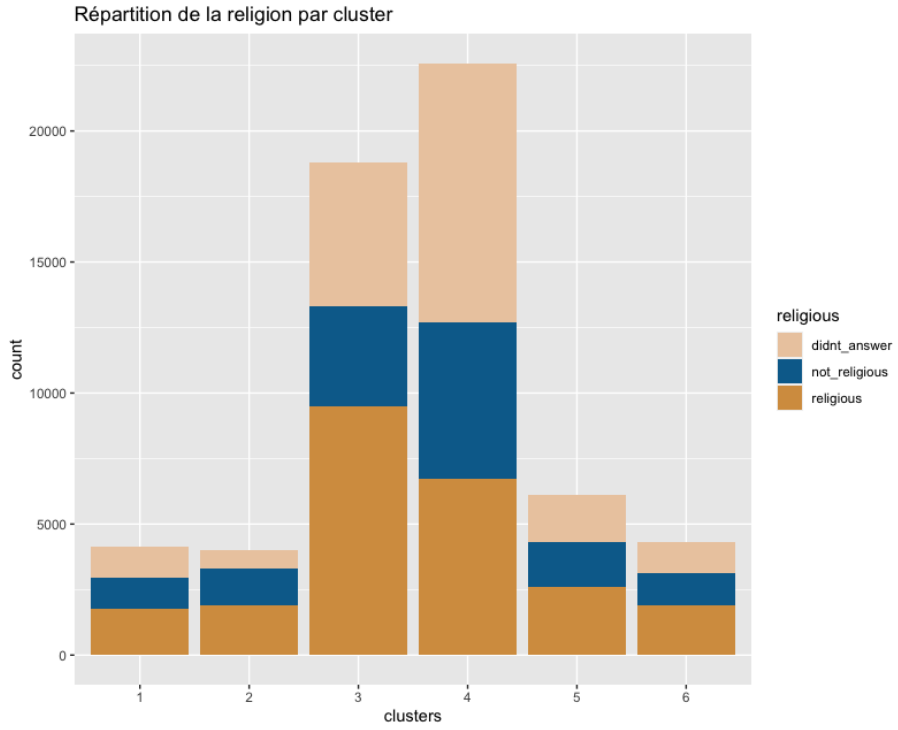
Graphique 16: *Répartition du fait d'aimer la musique par cluster*



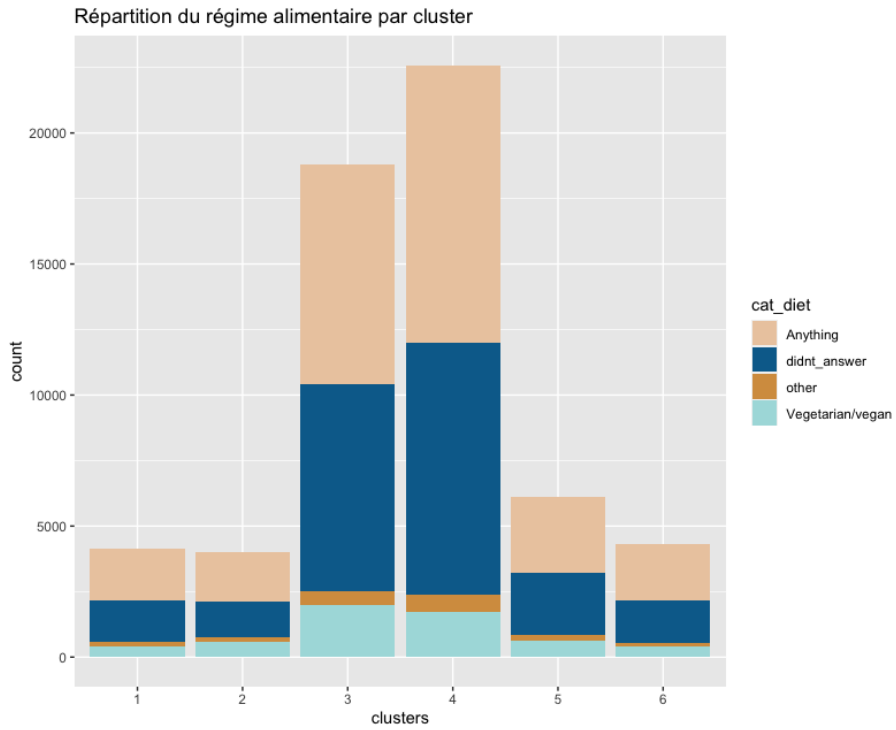
Graphique 17: *Répartition du fait d'être diplômé par cluster*



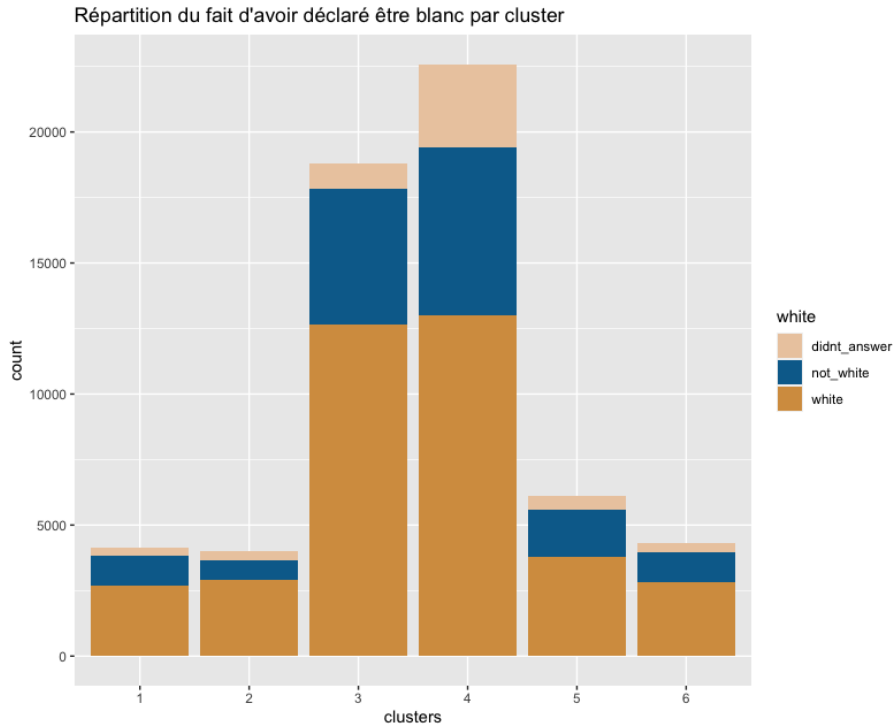
Graphique 18: *Répartition du fait d'avoir une religion par cluster*



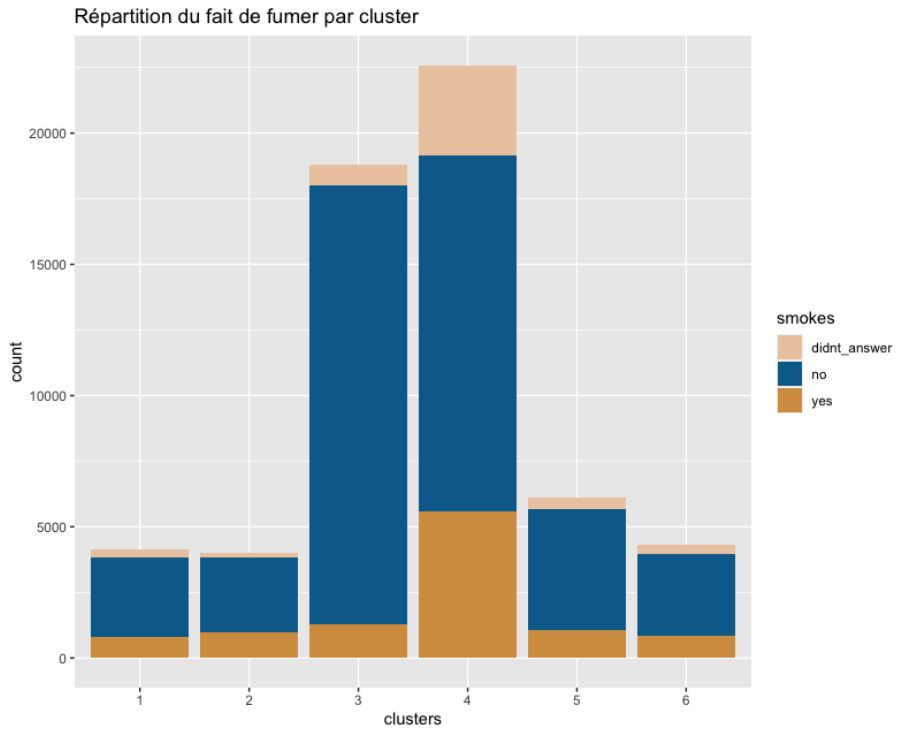
Graphique 19: *Répartition du fait d'avoir un régime alimentaire par cluster*



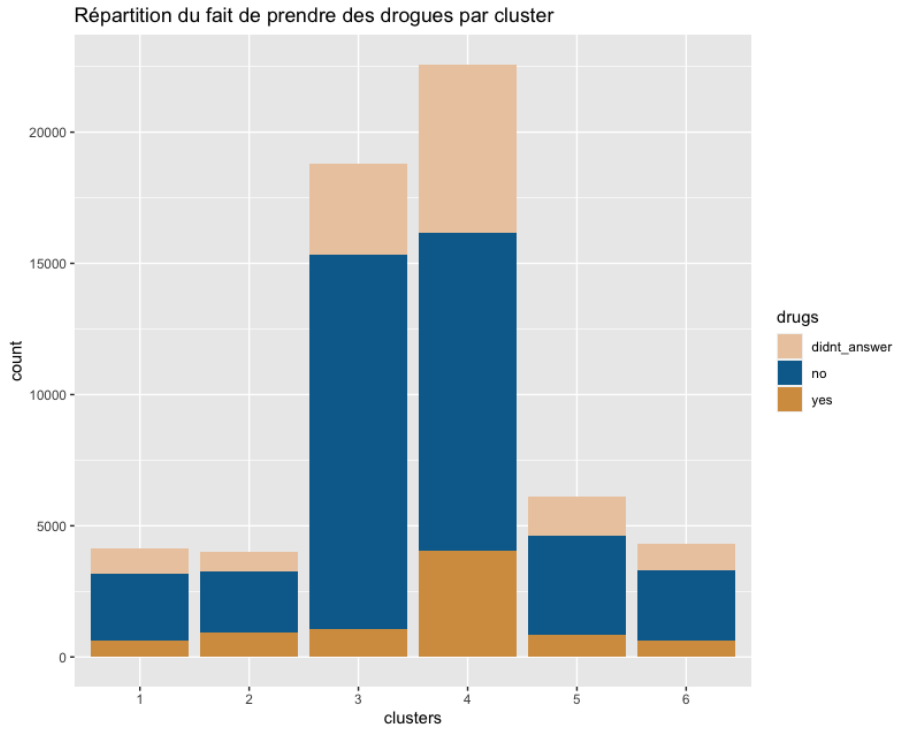
Graphique 20: *Répartition du faire d'avoir déclaré être blanc par cluster*



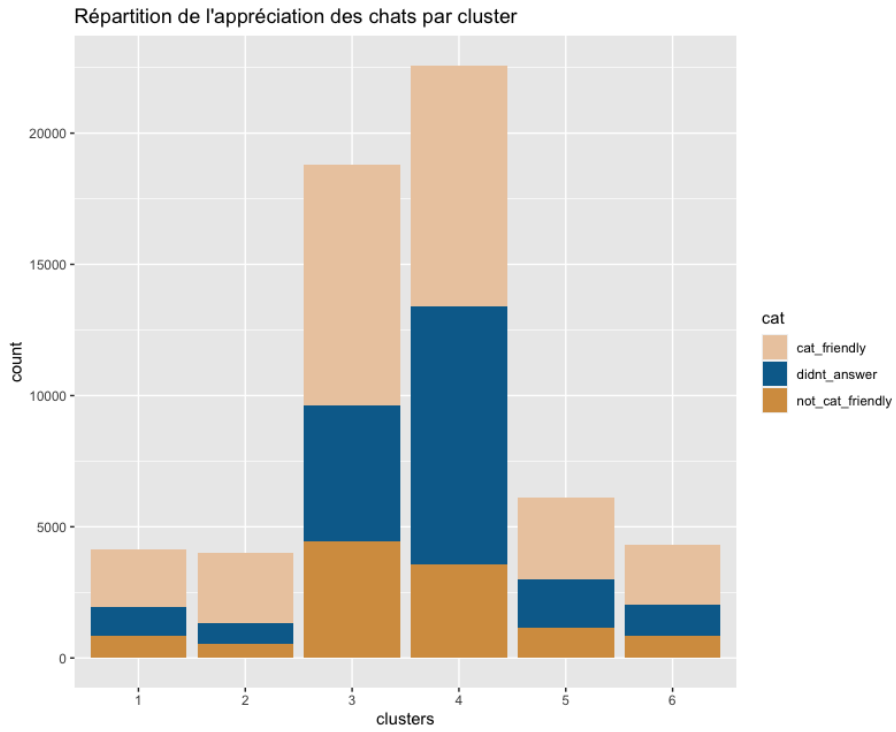
Graphique 21: *Répartition du fait de fumer par cluster*



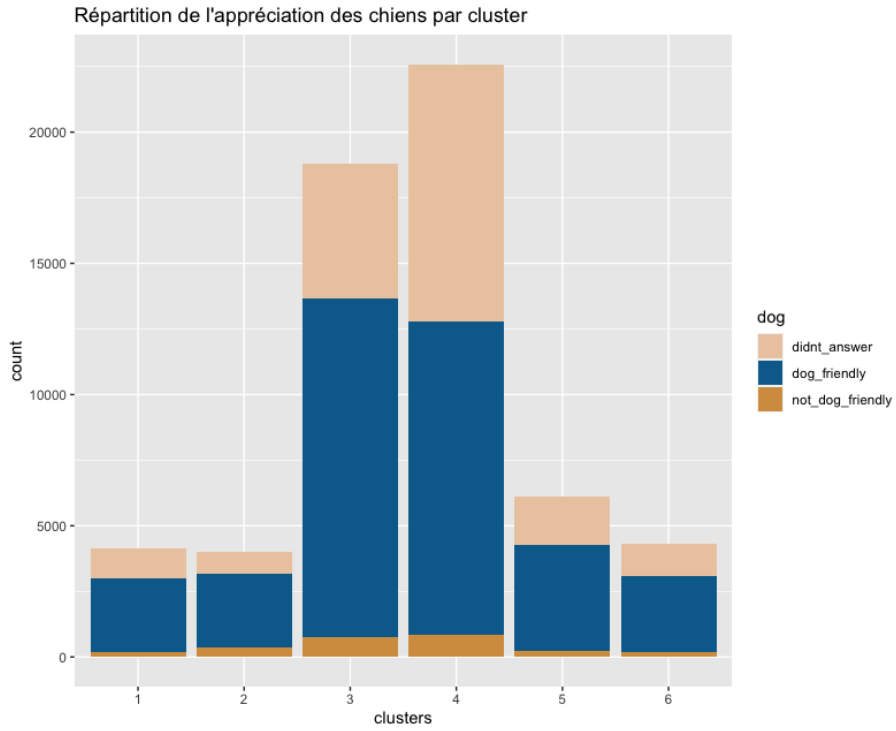
Graphique 22: *Répartition du fait de prendre des drogues par cluster*



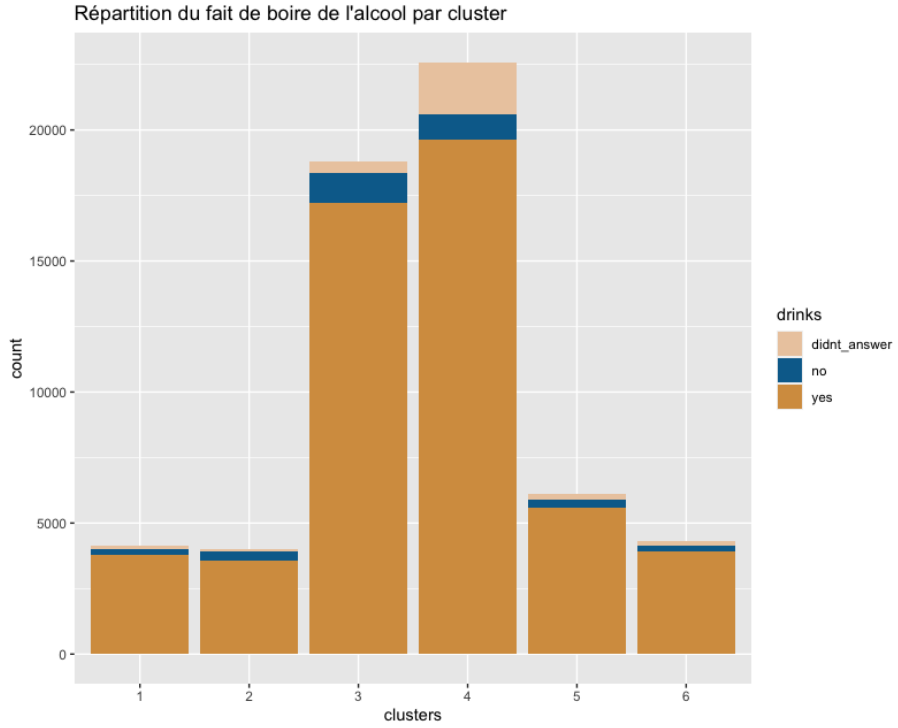
Graphique 23: *Répartition de l'appréciation des chats par cluster*



Graphique 24: *Répartition de l'appréciation des chiens par cluster*



Graphique 25: *Répartition du fait de boire de l'alcool par cluster*



Graphique 26: *Répartition de l'orientation sexuelle par cluster*

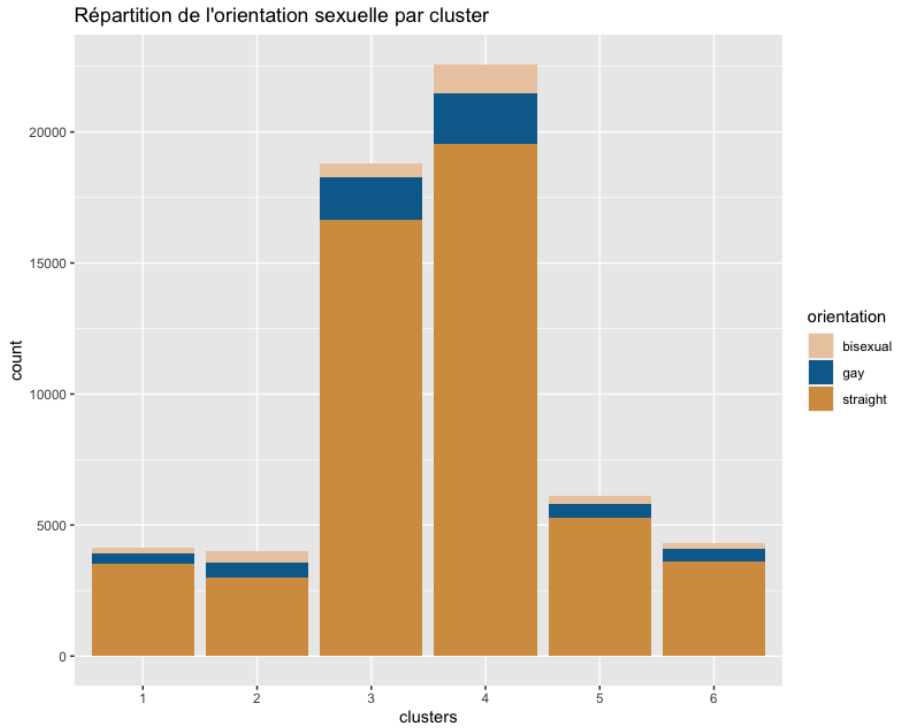


Tableau 25: *Analyse des performances prédictives de la Random Forest sur l'échantillon de test*

	Class: 1	**Class: 2**	**Class: 3**	**Class: 4**	**Class: 5**	**Class: 6**
Sensitivity	1.00000	0.99006	1.00000	0.9496	0.9010	0.9683
Specificity	0.99991	0.99964	1.00000	0.9508	0.9718	0.9973
Pos Pred Value	0.99881	0.99500	1.00000	0.9197	0.9324	0.9811
Neg Pred Value	1.00000	0.99929	1.00000	0.9695	0.9579	0.9954
Prevalence	0.06998	0.06710	0.06164	0.3723	0.3015	0.1274
Detection Rate	0.06998	0.06643	0.06164	0.3536	0.2717	0.1234
Detection Prevalence	0.07006	0.06677	0.06164	0.3844	0.2913	0.1258
Balanced Accuracy	0.99996	0.99485	1.00000	0.9502	0.9364	0.9828