Université du Québec à Chicoutimi

RAPPORT DU PROJET DE SESSION

Traitement d’un ensemble de données portant sur les espèces de champignons Agaricus et Lepiota

Présenté à

M. Julien Maitre

Dans le cadre du cours

**8INF418**

Forage de données

Réalisé par

Pierre-Marc Laforest

Marc-André Émond  
EMOM07089105

Le Vendredi 26 Avril 2019

Table des matières

[INTRODUCTION 3](#_Toc6766854)

[Description de l’ensemble de données 4](#_Toc6766855)

[Prétraitement des données 5](#_Toc6766856)

[Analyse des données 6](#_Toc6766857)

[Corrélation 10](#_Toc6766858)

[Corrélation symétrique vs Corrélation asymétrique 11](#_Toc6766859)

[BIBLIOGRAPHIE 13](#_Toc6766860)

# INTRODUCTION

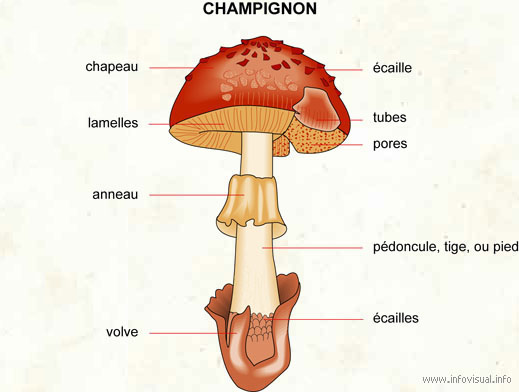
Après un rude hiver québécois, c’est l’arrivée du Printemps. La neige fond, les arbres et les fleurs bourgeonnent, de nouvelles couleurs et odeurs font leurs arrivées. Les champignons poussent. Certains adeptes de cuisine ou aventurier du camping en forêt vont s’adonner à la cueillette de champignons sauvages. En effet, c’est une activité de plus en plus populaire. En revanche, il faut être extrêmement prudent, car certains champignons sont très toxiques, voire mortels. Statistiquement, le centre antipoison du Québec reçoit environ entre 400 et 500 appels à la suite d’ingestion de champignon.

Les mycologues estiment qu’il existe environ 2824 espèces de champignons. De plus, avec les changements climatiques de plus en plus en vogue, de nouvelles espèces de champignon font leur apparition chaque année. Le rapport qui suit vise à comprendre si un ou plusieurs attributs peuvent aider à déterminer si un champignon est comestible ou s’il est toxique et dangereux.

Le rapport qui suit sera divisé en quatre sections: description et statistique sur l’ensemble de données, classification et comparaison de différents modèles

# Description de l’ensemble de données

L’ensemble de données qui a été utilisé pour ce travail a été tiré du site internet UCI Machine Learning. Depuis l’adresse suivante : https://archive.ics.uci.edu/ml/datasets/Mushroom. Les données hypothétiques de 8124 champignons sont présentes dans le dataset. Ces champignons font partie de 23 espèces dans la famille des Lepiota et des Agaricus. Les classes de cet ensemble sont la comestibilité du champignon et comporte deux valeurs e et p : e, désignant comestible et p désignant non-comestible. L’ensemble de données comporte aussi des valeurs manquantes. Les données sont tirées du guide nord-américain des champignons de 1981. Le guide indique clairement qu’aucun attribut n’est assez discriminatoire pour classifier chaque champignon.



Au total, l’ensemble de données comporte 8149 entrées et 22 attributs. Pour mieux comprendre ces attributs, la figure 1 a été ajoutée au travail. Voici la liste des attributs de ces données.

1. la forme du chapeau

2. la texture du chapeau

3. la couleur du chapeau

4. la présence de défauts

5. l’odeur

6. l’attachement des lamelles

7. la distance des lamelles

8. la grosseur des lamelles

9. la couleur des lamelles

10. la forme du pied

11. les racines du pied

12. la surface du pied au-dessus de l’anneau

13. la surface du pied en dessous de l’anneau

14. la couleur du pied au-dessus de l’anneau

15. la couleur du pied en dessous de l’anneau

16. le type de volve

17. la couleur du volve

18. le nombre d’anneaux

19. le type d’anneaux

20. la couleur des spores

21. population

22. habitat

Tel que mentionné précédemment le 23e attribut est en fait la comestibilité du champignon e pour comestible et p pour non-comestible. Le but de ce travail sera donc de classifier, à l’aide de modèles d’apprentissages, les champignons selon leur comestibilité.

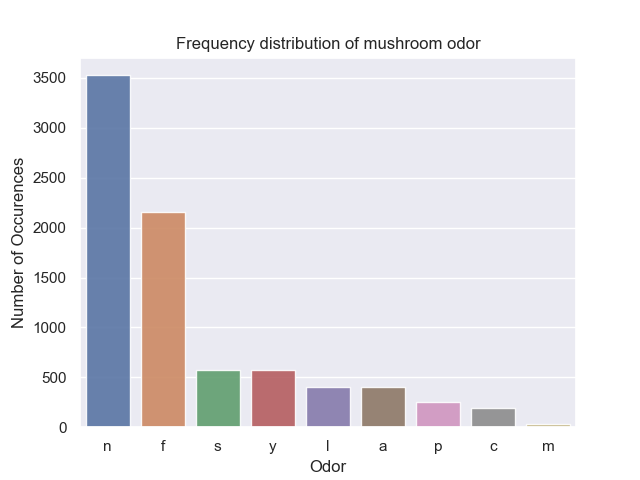
# Prétraitement des données

Le prétraitement des données a été très facile à effectuer. Il n’a fallu que créer un en-tête. arff pour rendre le fichier lisible dans le logiciel weka. Par la suite une fois son importation réussit dans le logiciel, il a été possible d’enregistrer le fichier dans le format désiré. Bien que l’ensemble comporte des valeurs manquantes (dénotée par ? dans les instances) celles-ci n’ont pas été enlevées de prime abord. De plus, un fichier Excel a été créé pour une meilleure visibilité des données.

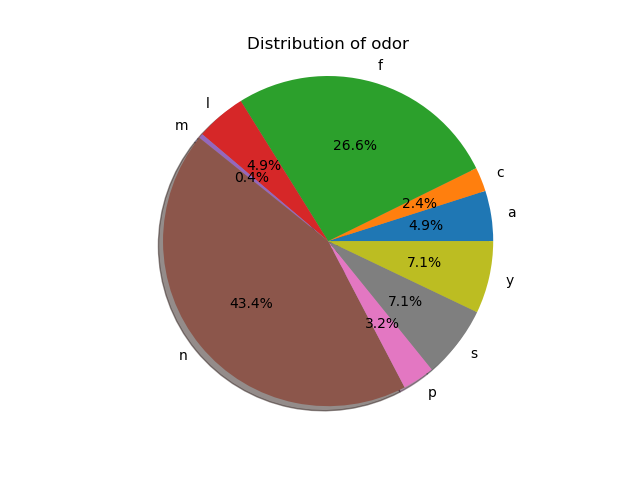
# Analyse des données

Tout d’abord, plusieurs fonctions ont été implémentées dans Python pour être en mesure de récolter le plus de statistique possible. Premièrement, a fait ressortir le nombre d’instances dans chaque classe. Il y a 4208 champignons comestibles et 3916 champignons non comestibles et toxiques, pour un total de 8124 champignons. Par la suite, d’autres statistiques ont été ressorties en fonction des règles de classifications. En effet, on peut remarquer dans le fichier agaricus-lepiota.names que l’odeur est l’un des attributs utilisés pour classifier les champignons non comestibles. Des statistiques ont donc été ressorties avec cet attribut. De plus, l’article intitulé « Handling Categorical Data in Python » de M. Manish Pathak a été d’une grande d’aide dans l’extraction générale de statistique.

Les diagrammes ci-dessous représentent la distribution des champignons par odeur.



**Distribution des odeurs des champignons sur un diagramme circulaire**



En regardant attentivement les deux diagrammes ci-dessus montrant la distribution des odeurs. On remarque, en ordre décroissant, que 43,4 % des champignons ne sentent rien de particulier (n), 26.6 % des champignons ont une odeur infecte (f), 7,1% sentent le poisson (y), également 7,1% ont une odeur épicée (s), 4,9% sentent l’amande (a), également 4,9% sentent l’anise (l), 3,2 % ont une odeur piquante (p), 2,5% possèdent une odeur de créosote (c) et finalement seulement 0,4% des champignons ont une odeur de moisi (m).

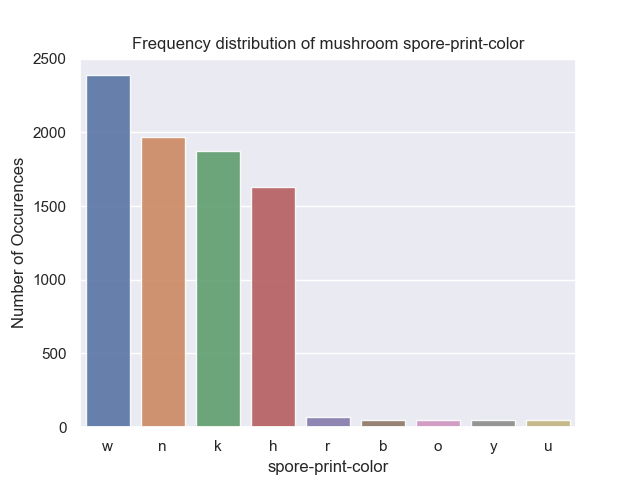
Le tableau ci-dessous montre le nombre de champignons dans chaque classe selon leur odeur.

**Nombre de champignons de chaque classe en fonction de leur odeur.**

|  |  |  |
| --- | --- | --- |
| **Odeurs** | **Nombre de Champignons comestibles** | **Champignon non comestible** |
| Amande (a) | 400 | 0 |
| Anise (l) | 400 | 0 |
| Creosote (c) | 0 | 192 |
| Poisson (y) | 0 | 576 |
| Infect (f) | 0 | 2160 |
| Moisi (m) | 0 | 36 |
| Aucune odeur (n) | 3408 | 120 |
| Piquant (p) | 0 | 256 |
| Épicé (s) | 0 | 576 |

En regardant attentivement le tableau, on peut remarquer au premier coup d’œil, la première règle de classification, En effet, les champignons comestibles ont une odeur d’amande d’anise ou ils n’ont aucune odeur alors que les champignons toxiques sont répertoriés à travers toutes les autres odeurs.

Une autre règle utilisée pour classifier les champignons toxiques est la couleur des spores. La règle indique que si les spores du champignon sont de couleur verte, ce dernier est classé comme toxique dans 99,41% des cas. Le diagramme ci-dessous montre la distribution des champignons selon la couleur de leur spore.

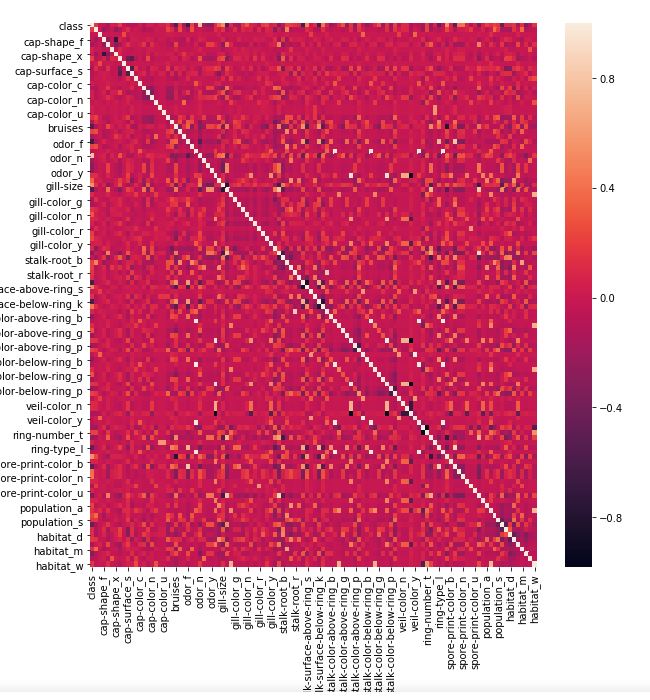


**Le tableau ci-dessous décompte le nombre de champignons de chaque classe selon la couleur des spores.**

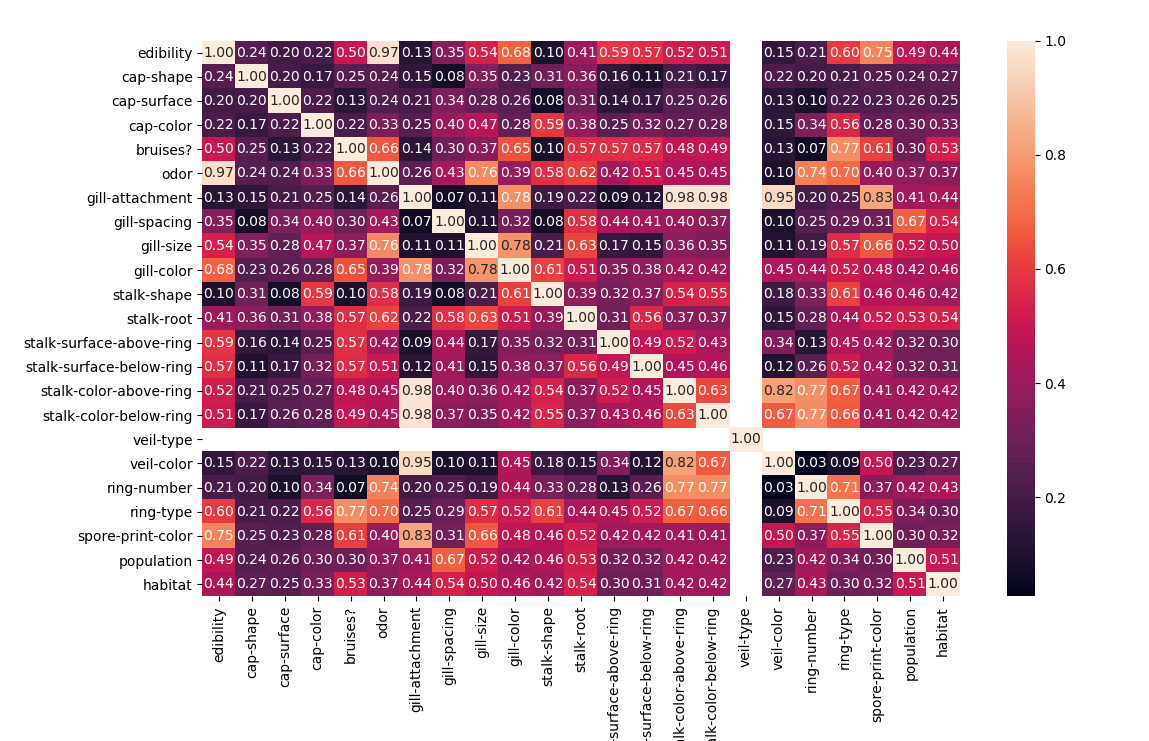
|  |  |  |
| --- | --- | --- |
| **Couleur des spores** | **Nombre de Champignons comestibles** | **Nombre de Champignon non comestible** |
| Vert | 0 | 72 |
| Brun | 1744 | 224 |
| Noir | 1648 | 224 |
| Chamois | 48 | 0 |
| Chocolat | 48 | 1584 |
| Orange | 48 | 0 |
| Mauve | 48 | 0 |
| Blanc | 576 | 1812 |
| Jaune | 48 | 0 |

On peut facilement remarquer à l’aide de la première ligne du tableau qu’aucun champignon comestible ne possède de spore de couleur verte. On remarque aussi rapidement que la plupart des champignons comestibles possèdent des spores de couleur brune, noir ou blanc, alors que la majorité des champignons classés non comestible ont des spores de couleurs chocolat ou blanc. Il est aussi intéressant de remarquer qu’aucun champignon non comestible n’a de spore de couleur chamois, orange, mauve ou jaune.

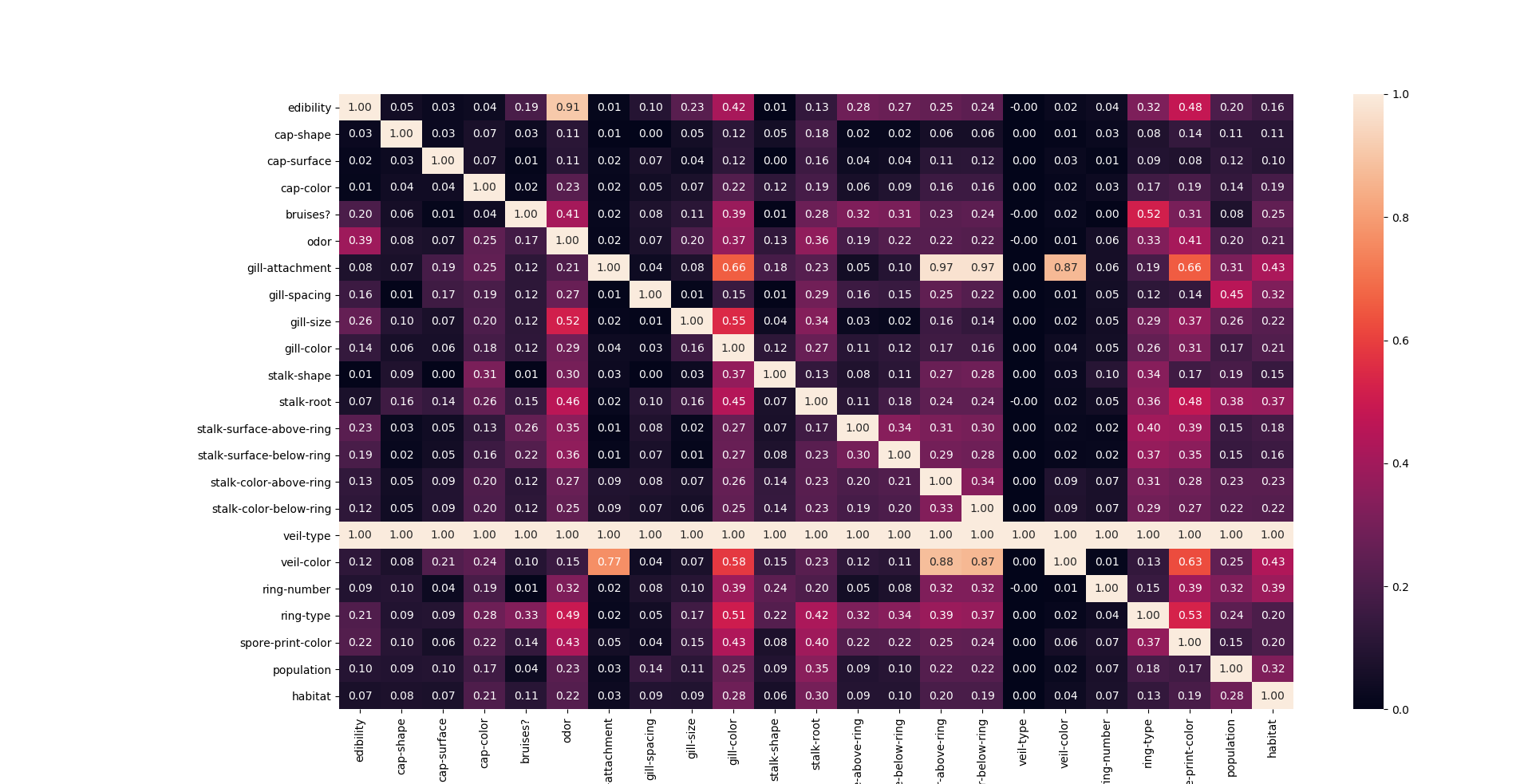
# Corrélation

Pour trouver la corrélation entre des attributs catégoriques et nominaux, on doit utiliser des règles différentes que celle vue en cours. C’est en lisant un article de l’ingénieur en algorithme Shaked Zychlinski sur le site internet Towards Data Science que nous avons pu comprendre les algorithmes utilisés pour faire la corrélation entre des valeurs catégoriques. De plus, par pur hasard, il se trouve que le dataset utilisé par l’auteur dans l’article est le même que le nôtre. L’auteur explique d’abord pourquoi il n’est pas optimal d’utiliser la corrélation de Pearson avec des données catégoriques. Normalement, un moyen de prendre en charge la corrélation avec des données catégoriques est d’encoder les données avec la méthode one-hot encoding. Cependant, dans le cas présent, cet algorithme d’encodage convertit les 22 attributs en 112 attributs. Cependant, en essayant de construire une heat-map de corrélation avec les attributs encodés, on obtient la heat-map suivante, tirée de l’article :

Comme on peut facilement le remarquer, il est difficile de faire ressortir les corrélations d’un tel graphique. Comme l’indique l’auteur de l’article, on a donc besoin de quelque chose pour mesurer l’association entre deux valeurs catégoriques. On utilise donc Cramer’s V qui est basé sur le « Pearson's chi-squared test ». Cet algorithme retourne une valeur entre 0 et 1. Une valeur proche de 0 correspond à aucune association alors qu’une valeur proche de 1 indique une forte corrélation entre les deux variables. L’algorithme Cramer’s V est déjà implémenté et est disponible dans la librairie dython. Lorsque l’on applique l’algorithme au dataset, on obtient une heat-map beaucoup plus visuelle. La corrélation entre l’odeur et la classe est évidente avec une valeur de 0.97.



# Corrélation symétrique vs Corrélation asymétrique

Un problème survient toutefois lors de l’utilisation de l’algorithme de corrélation catégorique Cramer’s V. En effet, puisque l’algorithme est symétrique, on perd de l’information au niveau des données. En regardant la heat-map, on voie que si on connait la valeur de l’odeur, on peut déterminer la classe. Et vice-versa. Or, le fait de connaitre la classe d’un champignon ne garantit pas de connaitre sa senteur. Pour être en mesure de ne pas perdre cette information, il faut utiliser un algorithme asymétrique. L’algorithme Theil’s U aussi connu sous le nom du coefficient incertain permet de calculer les coefficients en prenant en compte que connaitre la valeur de x garanties y mais pas le contraire. En appliquant cet algorithme lors de la construction d’une nouvelle matrice de corrélation donne le graphique ci-dessous. En regardant attentivement la nouvelle matrice, le nouveau coefficient calculé offre des informations plus réalistes de la situation. En effet, avec un coefficient de 0.91, on peut voir que si on connait l’odeur d’un champignon, il y a de fortes chances de trouver sa classe. En revanche, le coefficient de 0,39 indique que connaitre la classe d’un champignon ne garantit pas de trouver son odeur.

# BIBLIOGRAPHIE

Pathak Manish, « Handling Categorical Data in Python » Le 22 mai 2018, Data Camp, URL: <https://www.datacamp.com/community/tutorials/categorical-data>

Zychlinski Shaked, « The Search for Categorical Correlation » Le 23 février 2018, Toward Data Science, URL: <https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9>