Projet de visualisation de données : Le Chocolat

Réalisé par Barbillon Héléna, Cordebar Benjamin, Hornberger Théo et Ménestrel Victor

Introduction

1. Présentation de la base de données

La base de données que nous avons utilisée est "Chocolate Ratings" disponible sur kaggle. Cette base de données provient du site web "Flavors of Cacao", et recense les résultats de dégustations de chocolat de la Manhattan's Chocolate Society depuis 2007. C'est un groupe de personnes se réunissant pour goûter, débattre sur les qualités gustatives et donner une note à des tablettes de chocolat.

Ces données comprennent des informations sur chaque chocolat évalué, tels que le fabricant, sa localisation, la date d'évaluation, l'origine des fèves de cacao, le pourcentage de cacao, les ingrédients, les caractéristiques notables des barres de chocolat et la note finale donnée.

2. Choix de la base de données

Nous avons choisi cette base de données pour plusieurs raisons. Tout d'abord, nous voulions une base de données en lien avec l'alimentation, comme le chocolat ou le vin. Cette base de données était très bien fournie et son aspect très détaillé dans ses caractéristiques, en plus d'avoir une note pour chaque chocolat, nous a convaincu. Les différents types de données mis à disposition (chiffres, localisation, textes) nous ont permis de construire des représentations graphiques diverses et originales.

3. Outils utilisés

Nous avons réalisé ce projet en python, avec plusieurs librairies de visualisations de données comme matplotlib et plotly.

4. Objectifs visés

Nous nous sommes d'abord intéressés au contexte des chocolats évalués par la Manhattan's Chocolate Society, en cherchant à savoir d'où ils venaient. Nous nous sommes ensuite penchés sur les caractéristiques gustatives des chocolats et avons cherché un lien entre la qualité d'une barre de chocolat caractérisée par sa note, et ses ingrédients.

I. Pré-traitements

Nous avons commencé par vérifier s'il y avait des données manquantes ou des doublons. Il n'y avait pas de doublons, mais le champ 'ingrédients' était vide pour 87 lignes (sur 2530 entrées). Étant donné la proportion très faible de ces

lignes par rapport à l'ensemble des données, nous avons décidé de les supprimer lors des visualisations concernant ce champ. D'autres traitements ont été réalisés sur la base de données, notamment sur la teneur en cacao des chocolats ('cocoa_percent'), pour lesquels nous avons supprimé le symbole "%" qui était présent après chaque chiffre. Des traitements ont également été réalisés sur la catégorie 'ingredients' (partie III) et 'caractéristiques_notables' (partie IV), ils sont détaillés plus loin dans ce compte rendu.

II. Contexte des dégustations : d'où viennent les chocolats ?

1. Origine de la matière première : les fèves de cacao

Nous avons visualisé les pays producteurs de fèves de cacao qui sont représentés par la valeur "country_of_bean_origin" avec un histogramme pour voir les différences de quantités, et un diagramme circulaire en donut pour visualiser les proportions.

La problématique principale était d'avoir un rendu lisible, puisqu'il y a beaucoup de pays différents. C'est pour cela que dans l'histogramme ne sont affichés que les pays mentionnés plus de 20 fois. En revanche dans le diagramme circulaire, nous avons laissé tous les pays affichés pour pouvoir les voir dans la légende sur la droite, et décocher les plus gros producteurs pour laisser plus de place aux autres, même si c'est moins lisible.

Nous avons ensuite visualisé ces données sur une carte du monde. Le choix de projection "natural earth" était celui le plus facilement compréhensible pour ces données. Le paramètre "location_mode" a permis de donner en entrée les noms des pays sans avoir à renseigner leurs coordonnées. On voit ainsi aisément que les fèves de cacao des chocolats évalués étaient produites en Amérique du Sud, en Afrique et en Asie.

Cependant, il ne faut pas généraliser cette conclusion au niveau mondial, puisque ces données ne caractérisent que l'échantillon de barres de chocolats goûtées par la Manhattan's Chocolate Society.

2. Localisation des fabricants de chocolat

Nous avons répété le processus précédent pour connaître les pays où sont localisés les fabricants de chocolat, et le résultat est le plus flagrant sur le diagramme circulaire : les chocolats évalués sont principalement des marques nord-américaines. Si on remet les données dans leur contexte, il semble logique que les chocolats évalués soient des marques américaines, puisque ces données sont basées sur des observations d'un groupe de testeurs de chocolats basé à Manhattan.

3. Importations du fabricant principal des chocolats évalués : les États-Unis

Nous avons ensuite voulu visualiser sur une carte le trajet effectué par les chocolats. Dans un souci de simplicité (et de faisabilité), nous avons représenté les trajets entre le principal fabricant de chocolat, les États-Unis, et les pays de production des fèves de cacao correspondants.
L'outil line_geo utilisé pour créer des cartes permet de tracer des itinéraires, et non des liaisons entre deux points. Nous avons donc considéré les données comme des itinéraires entre le pays producteur et les États-Unis, avec toujours un aller-retour aux USA avant de passer au pays producteur suivant. Nous avons choisi la projection "orthographic" pour son côté interactif, et pour bien voir que les chocolats sont produits dans beaucoup de pays différents et éloignés du pays du fabricant de chocolat.

III. Caractère gustatif du chocolat

1. Répartition des notes

Nous avons affiché la répartition des notes afin de mieux comprendre dans quel contexte nous nous situons. Nous avons utilisé un diagramme en boîte à moustaches pour avoir accès à des éléments statistiques. Les notes sont comprises entre 1 et 4, et sont généralement plutôt bonnes.

2. Lien entre la note attribuée et les ingrédients

Pour pouvoir réaliser des visualisations en rapport avec les ingrédients présents dans chaque chocolat, nous avons modifié la base de données en y ajoutant plusieurs colonnes : une indiquant le nombre d'ingrédients présent dans chaque chocolat, et une colonne pour chaque ingrédient, étant à 0 si l'ingrédient n'est pas présent dans le chocolat, et à 1 s'il l'est.
Nous avons donc ajouté : *Beans*, *Sugar*, *Sweetener other than sugar*, *Cocoa Butter*, *Vanilla*, *Lecithin*, *Salt* et enfin *Num_Ingredient*.
Nous avons représenté la répartition des notes en fonction du nombre d'ingrédients avec des diagrammes en boîte, ainsi qu'avec un graphique où on peut voir la note minimale, maximale, médiane et moyenne. Nous avons également fait une corrélation entre ces derniers ce qui nous permet de dire que le nombre d'ingrédients ne semble pas avoir un impact significatif sur la qualité du chocolat.

3. Lien entre la note attribuée et la teneur en cacao

Nous nous sommes ensuite demandés si la teneur en cacao avait un impact sur la qualité du chocolat. Nous avons donc visualisé la note moyenne donnée en fonction de la teneur en cacao des chocolats, en ajoutant la note minimale et maximale pour avoir un peu plus de contexte. On obtient un

graphique assez irrégulier et difficile à interpréter pour les faibles pourcentages de cacao. Pour essayer de mieux comprendre ce graphique, nous avons produit un diagramme en violon de la teneur en cacao des chocolats, ainsi on voit que la moitié des concentrations en cacao des chocolats est comprise entre 70% (q1) et 74% (q3). On voit par ailleurs que peu de chocolats de moins de 50% ou plus de 90% de cacao ont été testés. On peut conclure, avec prudence, que les chocolats de plus de 90% de cacao sont généralement moins appréciés, mais qu'il y a peu de lien entre la qualité d'un chocolat et sa teneur en cacao.

4. Lien entre le goût du chocolat et sa teneur en cacao

Nous allons maintenant nous pencher sur la teneur en cacao et les différentes caractéristiques des chocolats quant à leur influence sur la qualité du chocolat.

Tout d'abord, nous avons affiché le pourcentage moyen de cacao associé à chaque caractéristique. Afin d'obtenir des résultats plus pertinents, nous choisissons de traiter les caractéristiques ayant au moins été référencées 10 fois. Avec les données révélées par ce graphique, nous en avons conclu que les caractéristiques gustatives ne sont pas liées à la teneur en cacao. Cependant, aux extrêmes, on remarque une différence de comportement. Celle-ci est facilement explicable lorsque l'on observe que ces extrêmes sont liés à l'amertume et à la teneur en sucre du chocolat. En effet, un chocolat très fort en cacao (>80%) sera plus souvent décrit comme amer alors qu'un chocolat très faible en cacao (<60%) sera plus souvent décrit comme très sucré, car l'ingrédient y est présent en plus grande quantité.

IV. Qualifier le chocolat

1. Caractéristiques gustatives des chocolats

A quoi pensent les testeurs quand ils dégustent du chocolat ?

Dans notre base de données, on a pu observer que chaque chocolat goûté avait une description de ses caractéristiques gustatives les plus remarquables. Dans ce sens, on peut se demander quelles sont les caractéristiques les plus courantes et donc à quoi pensent généralement les gens quand ils dégustent du chocolat.

Pour répondre à cette question, nous devons tout d'abord prétraiter les données. En effet, il faut transformer la chaîne de caractères des caractéristiques, représentées comme ceci : "sweet, fat, nut", en une liste : [sweet, fat, nut]. Cela nous permettra alors de pouvoir compter ces caractéristiques.

Pour ce faire, différentes représentations graphiques s'offrent à nous. Il est tout à fait possible de les représenter à l'aide d'un histogramme comme fait précédemment pour le goût et l'amertume du chocolat. Cependant, étant donné que nous voulons voir les caractéristiques les plus présentes, mais aussi en voir un nombre important, une façon plus "créative" est la représentation en nuage de mots. À l'aide de la librairie WordCloud, cela se fait plutôt bien, et nous

avons même décidé d'y ajouter la forme et les couleurs de fèves de cacao, afin de rendre la représentation plus agréable à regarder.

Finalement, cette représentation nous permet facilement de comprendre quelles sont les caractéristiques gustatives les plus présentes dans le chocolat, et donc à quoi pensent les testeurs quand ils dégustent du chocolat.

2. Caractéristiques gustatives des chocolats les mieux notés et les moins bien notés

Quelles sont les caractéristiques gustatives préférées lors de la dégustation du chocolat ? Et celles moins appréciées ?

La question précédente nous permet de savoir quelles sont les caractéristiques gustatives les plus fréquentes, mais cela ne nous apporte aucune information quant à la qualité de ces caractéristiques. En effet, une caractéristique souvent présente ne veut pas dire que celle-ci est une caractéristique particulièrement appréciée, ou à l'inverse peu appréciée. Nous pouvons donc nous poser les questions suivantes : quelles sont les caractéristiques les plus appréciées lors de la dégustation de chocolat, et à contrario celles qui sont le plus détestées ?

Pour répondre à cette question, nous avons fait une démarche très similaire à la question précédente. En effet, la représentation sous un nuage de mots paraît tout aussi pertinente ici. Cependant, il était important de sélectionner un nombre minimal d'apparitions de la caractéristique. En effet, une caractéristique qui n'apparaît que pour un seul chocolat est peu pertinente. Nous avons alors décidé que la limite d'apparition serait de 3, afin de garder les caractéristiques qui sont suffisamment rares (qualifiant peut-être des caractères exceptionnels) sans pour autant être uniques. De plus, nous avons joué sur les couleurs afin d'appuyer les idées de bonnes et mauvaises caractéristiques.

Grâce à ces deux représentations, nous voyons facilement les caractéristiques les plus agréables et les plus détestables.

Conclusion

En résumé, cette analyse des données sur les dégustations de chocolat nous offre une vue d'ensemble intéressante. Nous avons d'abord regardé d'où venaient les chocolats, en examinant d'où provenaient les fèves de cacao, et en avons déduit que les fèves venaient principalement d'Amérique latine, d'Afrique et d'Asie, et que le principal fabricant de chocolat était les Etats-Unis, parmi les chocolats de notre base de données.

Ensuite, nous nous sommes penchés sur le goût du chocolat, et nous avons constaté que la quantité d'ingrédients n'influence pas la qualité du chocolat, tout comme sa teneur en cacao dans une certaine mesure.

Enfin, nous avons examiné les caractéristiques gustatives des chocolats les mieux et les moins bien notés, ce qui nous a permis de voir ce qui fait un bon (ou un mauvais) chocolat selon les dégustateurs.

Cependant, ces données sont à resituer dans leur contexte, puisque la base de données est basée sur les dégustations d'un regroupement d'amateurs de chocolat de Manhattan. Ces données ne sont pas toujours complètes et exhaustives, et on peut supposer que les résultats d'une dégustation auraient tendance à varier selon la culture, le mode de vie, les préférences de chacun et d'autres aspects sociaux-culturels. Nous avons aussi peu d'informations sur la façon dont ont été récoltées les données, et sur le caractère objectif des dégustations.

Nous restons donc prudents dans nos déductions quant à ce qui caractérise un bon chocolat, et vous laissons juger par vous-même.