[SLIDE 2]

Tout d'abord, quelques mots sur le RGPD et son application à ce projet.

Le RGPD résulte de la transposition en droit français du paquet européen de protection des données de 2016 et il impose un contrôle *a posteriori* avec une obligation de documentation à la charge du responsable du traitement des données. Il concerne les données personnelles contenues dans un fichier, que ce fichier soit électronique ou non, et garantit certains droits à la personne physique dont les données sont collectées ; il interdit également certains traitements (comme la collecte de données de santé ou d'appartenance religieuse) et impose au responsable du traitement de déclarer les violations de ces règles à la CNIL. Les données personnelles doivent en outre être collectées et conservées selon certaines règles assez précises : elles doivent l'être de manière licite, loyale et transparente (ce qui implique le consentement de la personne physique), pour des finalités déterminées, explicites et légitimes (dont la personne physique doit être informée), les données doivent être adéquates, pertinentes et limitées, exactes et tenues à jour, et conservées de manière sécurisée pour une durée limitée.

[SLIDE 3]

Dans le cadre de notre étude, les seules données personnelles présentes sont l'adresse email de la personne ayant renseigné les données dans la base de notre site internet, et que nous avons supprimées de notre analyse, ce qui nous permet de nous affranchir de la limitation de durée de conservation sur ces données puisqu'elles sont totalement anonymisées et ne peuvent donc plus être identifiées avec « des moyens raisonnables ». Nous n'aurons pas besoin non plus de les mettre à jour, sauf pour les besoins statistiques de notre étude. Attention toutefois, ces règles ne s'appliquent qu'aux données anonymisées utilisées pour la suite de cette analyse, et pas aux données du site internet qui elles, restent soumises de plein droit au RGPD.

[SLIDE 4]

Nous allons donc maintenant passer à la présentation des différents éléments de cette analyse de données, depuis le téléchargement, le nettoyage et l'exploration jusqu'aux analyses uni-et multi-variées, en passant par la réduction de dimension via l'ACP, et ANOVA et une de ses alternatives non-paramétriques.

[SLIDE 5]

Le fichier de données initial était de grande taille, avec plus de 320,000 lignes et 129 colonnes, et comportait des erreurs de formatage sur le séparateur du csv. Il était en outre d'une qualité médiocre, avec plus de 83% de valeurs manquantes et 20 colonnes entièrement vides ; toutes ces caractéristiques ralentissaient considérablement l'import et alourdissaient inutilement l'usage mémoire du fichier, donc après un échantillonnage initial, on a procédé à un certain nombre d'optimisations pour pouvoir importer le fichier sans time-out, comme par exemple définir les colonnes à garder et leur type de données via des dictionnaires json. Les différentes techniques d'optimisation utilisées pour l'import nous ont permis de diviser l'usage mémoire presque par 3.

[SLIDE 6]

Une fois les données importées, leur exploration nous a permis d'identifier le code produit comme clé primaire. Nous n'avons pas trouvé de doublons dans le fichier, la suppression des "mauvaises" lignes à l'import ayant également supprimé 23 doublons dans la colonne ['code'] produit qui étaient des NULLs dus à des lignes mal formées et décalées vers la droite.

Afin de conserver suffisamment d'information pour pouvoir caractériser nos produits de façon satisfaisante et efficiente, nous avons choisi dans un premier temps de ne conserver que les lignes correspondant à des produits ayant un nom et un label PNNS 1 et 2 renseigné (c'est-à-dire non vide et non inconnu), ces identifiants ayant été choisis car ils étaient unitaires et présentaient peu d'erreurs

de RegEx. Ces premiers retraitements nous laissaient avec plus de 68,000 lignes, mais avec plus des 2/3 des colonnes vides à plus de 80%. On a donc fixé un cut-off à 53% de valeurs vides, ce qui nous permettait de conserver la colonne fiber_100g dans l'analyse, étant une donnée nutritionnelle importante. Tous ces retraitements ont été synthétises dans une fonction nommée load_food_data.

[SLIDE 7]

Sur les 29 colonnes restantes, 10 sont en fait peu utiles pour notre analyse, étant des colonnes de texte descriptif multivaluées et difficiles à exploiter, que nous avons donc également supprimées. Parmi les colonnes conservées, nous avons choisi la variable catégorielle du nutriscore comme cible de notre analyse, les autres colonnes représentant des variables numériques discrètes (comme le nombre d'additifs ou le nombre de produits dérives d'huile de palme contenus dans un produit), ou des variables numériques continues (comme le taux de protéines ou de sel pour 100g de produit).

[SLIDE 8]

Nous avons alors retraité les valeurs manquantes en appliquant plusieurs stratégies :

- Tout d'abord, on a bien évidemment supprimé les lignes où notre variable-cible, le nutriscore, était absente, ce qui nous a laissés avec quelque 50,000 lignes ;
- Partout où c'était possible, la logique métier a été prépondérante, qu'il s'agisse des imputations ou des mises à zéro. On a par exemple supprimé les lignes où la variable energy_100g était manquante, car pour ces lignes toutes les variables qui permettent de la calculer (protein_100g, carbohydrates_100g et fat_100g) étaient aussi manquantes, on a mis à zéro la variable saturated_fat_100g si fat_100g valait aussi zéro, on a mis à zéro les fibres si cette variable était absente pour les produits d'origine

animale qui n'en contiennent pas, on a remplacé les carbohydrates_100g manquants par la somme de sugar_100g et fiber_100g, pour ne citer que quelques exemples ;

- Et les valeurs manquantes restantes ont été renseignées avec l'IterativeImputer de Scikit-Learn.

[SLIDE 9]

Concernant les outliers, on a corrigé les aberrations mathématiques, là encore de façon prépondérante avec des connaissances métier là où c'était possible. Par exemple, on a supprimé les lignes là où le poids total des macronutriments par 100g excédait 100g, on a retiré les lignes où les calories déclarées excédaient celles de la graisse pure ou celles calculées à partir des macronutriments de plus de 20% (qui est la marge admise par la FDA), on a remplacé les valeurs négatives par leur valeur absolue après inspection de plusieurs échantillons, pour ne citer que quelques exemples, et on a synthétisé ces retraitements dans une fonction food_data_extrim.

[SLIDE 10]

Suite à ces retraitements, tous les autres outliers, que nous voyons sur les graphiques sur ce slide, ont été considérés comme des valeurs atypiques et non aberrantes, et ont donc tous été conserves dans l'analyse.

[SLIDE 11]

Enfin, afin de procéder à des analyses plus fines et sous des angles plus variés que le simple nutriscore, nous avons ajouté plusieurs variables à partir des noms des produits, à savoir des catégories produits allégés, produits biologiques, produits contenant de l'huile de palme ou un de ses dérivés et produits dont la consommation quotidienne est recommandée. Nous verrons le résultat de ces différentes analyses au point 2.5, mais tout d'abord…

[SLIDE 12]

… quelques analyses univariées. Les produits à mauvais nutriscore représentent plus de la moitié de nos références, tout comme les catégories Salty snacks, Sugary snacks, Fat and sauces et Composite fonds. Nous avons donc à notre disposition une base de données qui recense majoritairement des produits ultra-transformés, ce que nous appelons UPF (Ultra-Processed Food).

[SLIDE 13]

Si l'on croise ces deux types de catégories, on s'aperçoit en effet que les catégories Sugary snacks, Salty snacks et Fats and sauces sont composées exclusivement ou presque exclusivement de produits à mauvais nutriscore, et que les catégories Cereals and potatoes et Fruits and vegetables sont très largement dominées par des produits à bon nutriscore, et qui peuvent donc être consommes quotidiennement.

[SLIDE 14]

Afin de déterminer quels outils d'analyse nous avions à notre disposition sur ce jeu de données, nous avons procédé au test de Kolmogorov-Smirnov sur toutes nos variables quantitatives continues, ce qui nous a permis de noter qu'aucune d'entre elles ne suit une distribution Gaussienne et nous a donc conduits à utiliser essentiellement des tests non-paramétriques.

[SLIDE 15]

Nous avons donc utilisé le Rho de Spearman pour caractériser les corrélations entre nos différentes paires de variables, ce qui nous a amenés à constater un lien fort entre carbohydrates_100g d'une part et sugars_100g et fiber_100g d'autre part, ce qui est en fait normal car d'un point de vue diététique et nutritionnel, cette première variable est en fait la somme des deux autres. On

voit par ailleurs que le contenu en sucre d'un produit semble être une fonction décroissante de son contenu en protéines et en sel, que le sel et le sodium ont un Rho de 1 (l'un étant encore une fois un multiple de l'autre, 2.5 fois pour être exact), ou que le contenu en graisses saturées est une fonction croissante du contenu en graisses.

[SLIDE 16]

Dans ce contexte, une réduction de dimensions via l'ACP pourrait être pertinente. Ici nous avons limité nos résultats au 5 premiers axes, qui nous permettent de capturer 80% de la variance, seuil au-delà duquel par ailleurs la contribution des eigenvectors devient inférieure à 1, n'apportant pas significativement plus d'information. On voit bien comment notre premier axe est dominé par le contenu en graisses et en carbohydrates, qui drive la valeur calorique de nos produits, alors que le 2ème axe est dominé par les produits riches en protéines et en sel et pauvres en carbohydrates.

[SLIDE 17]

Si l'on projette à présent les nutriscores de nos différents produits sur ce premier plan factoriel, on voit bien comment les produits de bon nutriscore se positionnent à gauche de l'ordonnée le long de l'axe des abscisses et comment le nutriscore se dégrade à mesure que l'on s'éloigne de l'ordonnée vers la droite du graphique, où l'on retrouvera d'ailleurs les produits de PNNS Groups 1 tels que les Sugary snacks et les Salty snacks.

[SLIDE 18]

Nous avons ensuite essayé de mettre en lumière le rôle du sodium dans l'attribution du nutriscore. Une analyse ANOVA aurait pu être pertinente ici, mais seule une de ses 3 hypothèses de départ a pu être vérifiée. On voit bien sur les

QQ plots sur ce slide que le contenu en sodium de nos produits ne suit pas une distribution gaussienne…

[SLIDE 19]

… et le test de Levene sur les variances ne nous a pas permis de vérifier la condition d'homoscédasticité, seule l'hypothèse d'observations indépendantes étant considérée comme vérifiée en l'absence d'informations contraires. Nous avons néanmoins calculé le Êta carré qui, à 0.04 environ, nous inviterait à conclure à une faible relation entre le contenu en sodium d'un produit et son nutriscore.

[SLIDE 20]

Afin de caractériser cette relation avec un outil statistique plus adapté, nous avons donc utilisé le test de Kruskal-Wallis, qui nous permet au seuil alpha = 5% de conclure qu'au moins un de nos 5 groupes de nutriscores a un taux médian en sodium différent des autres, et le test post-hoc de Dunn nous permettra en fait de conclure que tous ont des niveaux médians différents sauf les groupes B et E.

[SLIDE 21]

Nous avons également comme on l'a dit mené des analyses sur nos engineered features en utilisant le test du Khi-2 de contingence sur les variables catégorielles ou le test de Mann-Whitney sur les paires variables catégorielles-variables numériques, et ces analyses nous ont montré que le caractère biologique d'un produit a une influence sur son contenu en huile de palme, son nutriscore, son contenu en additifs et son PNNS Group ; même si ces relations sont d'intensité faible, elles sont néanmoins statistiquement significatives.

[SLIDE 22]

Ces analyses nous ont également montré que le caractère allégé d'un produit est indépendant de son contenu en huile de palme, mais pas de son nutriscore ni de son contenu en additifs. La encore, ces relations sont d'intensité faible mais statistiquement significatives.

[SLIDE 23]

Enfin, les produits dont la consommation quotidienne est recommandée ne sont pas indépendants de leur contenu en additifs ni en huile de palme ni de leur PNNS Groups 1, là encore avec des intensités variables du très faible au très fort.

[SLIDE 24]

En conclusion, l'étude des relations entre nos différentes variables fait apparaître des relations soit mathématiques soit statistiques qui rendent possibles la mise en place d'un système d'auto-complétion des données. Cette phase requerrait un modèle de machine learning qui sort du cadre de ce projet mais qui est tout à fait réalisable. Malgré cela, ma recommandation principale serait de travailler en amont à l'amélioration de la qualité des données par d'autres moyens avant la mise en place de ce système, comme la data validation, l'interdiction de soumission de formulaires incomplets ou mieux encore, l'implication des industriels producteurs de ces produits pour renseigner la base à leur mise sur le marché.