Note explicative

R5.C.06
Exploitation de bases de données

Rendu le 18 11 2025, soutenu par :

M. DUBOIS Thibault

M. TERRADE Richard

3D2

Contexte et questions de recherche

Contexte, question

Nous nous sommes mis dans la peau de jeunes entrepreneurs, créant leur propre boutique de sport à l'échelle nationale. Certaines régions (proche de Bretagne) seront privilégiées car le siège social se situera à Rennes. Sachant que les boutiques ne peuvent couvrir tous les sports, la question est la suivante :

Quels sont les sports les plus pratiqués près de la Bretagne et dans lesquels investir ?

Le modèle créé consiste en l'analyse des données sur les fédérations sportives, présentes publiquement et gratuitement sur data.gouv. Depuis ces données nous pouvons les traiter, les nettoyer et les homogénéiser afin de connaître les sports les plus pratiqués selon différents critères : age, genre, région… etc

Pourquoi ce contexte ?

Sachant que le sport est un domaine qui nous intéresse tous deux, (Thibault et moi), nous avons naturellement choisi ce sujet à la fois concret et intéressant.

Choix du modèle et justification

Langage de programmation, format du modèle

Comme demandé, le modèle est un simple outil d'analyse statistique simple, sans intégration d'IA. Pour ce faire, nous avons utilisé python avec sa librairie pratique streamlit, permettant de créer une interface web local pour notamment afficher les graphiques voulus.

Remarque : Pour utiliser ce projet (hébergé sur github), il faut avoir un minimum de connaissances en informatique. Nous sommes conscients qu'il eût été mieux de proposer un outil clé en main et plus facile à utiliser pour le grand public, mais nous n'avons pu le faire par manque de temps. C'est une perspective d'amélioration

Nos graphes prennent surtout la forme de diagrammes en barres (horizontaux et verticaux) afin de visualiser facilement les écarts de proportions. Aussi nous avons un diagramme circulaire affichant des proportions utiles, comme les genres.

Méthodologie et traitement des données

Récupération initiale des données

L'objectif était de trouver un fichier répertoriant les sports et leurs effectifs, par genre et par région également. Or il est toujours difficile de trouver parfaitement ce que l'on cherche. Afin d'avoir des données statistiques riches (pertinentes et nombreuses), nous avons été chercher le fichier au format CSV sur le site du gouvernement. Nous y avons trouvé un fichier presque parfait (lic-data-2022.csv) : il recense les effectifs des clubs sportifs par genre et région.

Deux problèmes se sont alors soulevés :

- Le fichier est volumineux (presque 1M de lignes), nous ne pouvons nous permettre de le charger fréquemment
- Le fichier contient seulement le nom des CLUBS sportifs (fédérations) donc il faudra trouver un moyen d'inférer le nom de leurs sports.

C'est donc pour cela que nous avons eu besoin d'un nettoyage de données.

Nettoyage des données

Pour résoudre les deux problèmes énumérés ci-dessus, nous avons utilisé python et ses fonctions pratiques pour la lecture de fichier. Aussi, lors du projet de SAE 3, nos camarades en FI ont développé un outil généraliste pour transformer et valider le format des colonnes des fichiers CSV à notre guise. Nous avons repris et adapté cet outil pour répondre à notre besoin.

Voici une liste succincte des étapes de nettoyage :

- Supprimer les colonnes inutilisées (Code commune, Code QPV, Statut géo..)
- Renommer les colonnes pour homogénéiser leur nom (federation, region…)
- Ajouter des colonnes calculées pratiques (h_count et f_count pour compter les effectifs par genre)
- Ajouter la colonne sport déduite de la colonne
- Valider les données de chaque colonne

Spécifications

Avant de nettoyer les données, nous utilisons le programme create_standardized_file.py afin de créer un fichier de règles JSON permettant de lier chaque fédération à un sport. Cela permet par la suite de créer la colonne sport en fonction de la fédération lors du nettoyage.

Il est possible que le fichier JSON initialement créé contienne des coquilles dans le nom des sports. Un second passage est nécessaire.

Résultats et interprétations

Visualisation des graphiques

Une fois le fichier CSV nettoyé et le serveur streamlit lancé, nous pouvons consulter les graphes générés afin d'analyser les données.

Rappel de la question principale :

Quels sont les sports les plus pratiqués près de la Bretagne et dans lesquels investir ?

Voici les sports les plus pratiqués en Bretagne, Normandie et Pays de la Loire :

- Football (424k)
- Tennis (138k)
- Equitation (120k)
- Basket (119k)

Interprétation et réponse

Pour lancer notre activité, **nous devons donc nous focaliser sur ces quatres sports**. Egalement, nous pouvons mettre en vente des produits multisports en légère quantité, car nous ne pouvons ignorer les licenciés des écoles (UGSEL et UNSS > 300k) qui représentent une grande part du marché. Ces licenciés au lycée ou plus jeunes sont d'autant plus importants car leur portefeuille moyen est plus important que celui des étudiants (les parents paient 299€/an/enfant alors qu'une licence universitaire coûte 20-40€/an).

Limites et pistes d'amélioration

Le modèle traite déjà des données à l'échelle de la France, donc l'augmentation de l'échelle n'est pas une limite.

En revanche, l'outil est un peu dans un état primaire et n'est pas facilement utilisable pour tout utilisateur. Cela pourrait être amélioré.

Egalement, le modèle fonctionne uniquement avec un format de fichier CSV bien précis, celui de lic-data-2022.csv. Si nous voulons nous adapter aux fichiers des années ultérieures pour être à jour, il faut que leur structure reste la même. Sinon le modèle est sensible à la moindre altération.