SAE – Collecte automatisée de données web

Membres du groupe :

Antoine MAURIN

Rafaël SANZ

Année universitaire : 2026 – 2027

Introduction

Dans le cadre de notre module de Web Scraping, nous avons souhaité orienter notre projet vers un domaine qui nous plait à tous les deux : le sport. Lors de la phase de réflexion, le choix du rugby s'est imposé comme une évidence.

Étant un fervent supporter de l'**Union Bordeaux Bègles (UBB)**, j'ai suggéré de concentrer nos efforts sur ce club. Ce choix est motivé par deux facteurs principaux :

1. **L'intérêt personnel :** Travailler sur un sujet que l'on maîtrise permet une meilleure critique et analyse des données récoltées.
2. **La richesse des données :** Un club de Top 14 génère une quantité importante de statistiques (performances des joueurs, historique des matchs, effectifs) qui se prêtent parfaitement à la création d'un tableau de bord.

L'objectif de ce rapport est de détailler notre démarche technique, de la collecte des données à leur visualisation dans un outil de Business Intelligence.

Liste des sites web choisis

Pour réaliser ce projet, nous nous sommes concentrés sur le site officiel du club.

- **Site principal :** https://www.ubbrugby.com/
- **Joueurs :** https://www.ubbrugby.com/equipes/equipe-premiere/effectif.html/...
- **Classement :** https://www.ubbrugby.com/equipes/equipe-premiere/classement.html
- **Résultats :** https://www.ubbrugby.com/equipes/equipe-premiere/calendrier-resultats.html

Légalité et Éthique (Robots.txt)

Avant de lancer le moindre script de récupération de données, nous nous sommes assurés de respecter la politique du site web en matière de scraping.

Nous avons consulté le fichier robots.txt situé à la racine du site (https://www.ubbrugby.com/robots.txt).

- Allow: / : L'accès est autorisé depuis la racine du site.

Aucune restriction majeure ne bloquant l'accès aux pages de statistiques ou d'effectif, nous avons pu procéder à l'élaboration de nos scripts en toute légalité.

Analyse de la récolte et du traitement

Initialement, nous avions commencé la récolte en récupérant tous les liens de toutes les pages des joueurs pour faire le scrap. Ça fonctionnait, cependant, nous avons repris la collecte via le lien parent

de toutes ces pages pour n'avoir qu'un seul lien dans notre code. Notre code a été fait de sorte à garder les éléments jugés en amont comme pertinents : "player_id", "name", "position", "height_cm", "weight_kg", "age", "nationality", "since_year", "caps", "matches", "tries", "points", "url". Les données que nous avons pu collecter étaient donc déjà relativement propres. Ainsi, il n'était donc pas difficile d'utiliser ces données pour le tableau de bord car elles ne nécessitaient pas un gros retraitement. Concernant les résultats de match et les classements, le principe a été le même.

Conclusion suite à l'analyse

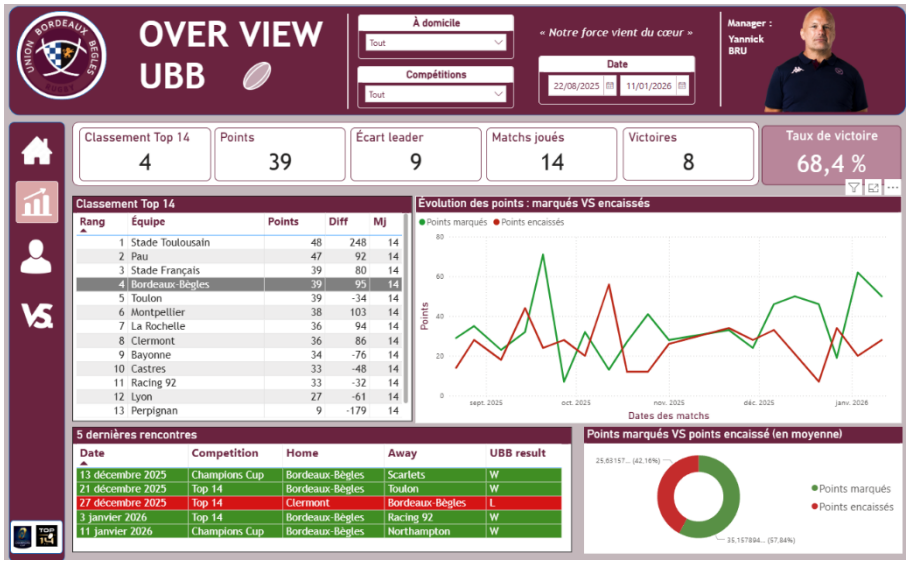
Le produit fini est un rapport PowerBI interactif composé de 4 onglets, permettant une navigation du général au particulier fluide et simplifiée par le volet de gauche contenant des icones cliquables dans une thématique visuelle aux couleurs du club.

Onglet 1 : Contexte du projet

Cette page d'accueil présente le cadre de l'étude, la méthodologie employée et les sources. Elle sert d'introduction à l'utilisateur du tableau de bord.

Onglet 2 : Statistiques globales du club

Cet onglet offre une vue macroscopique de la saison de l'UBB. Il permet d'analyser les performances collectives de l'équipe (classement, points marqués/encaissés, ratio de victoires).

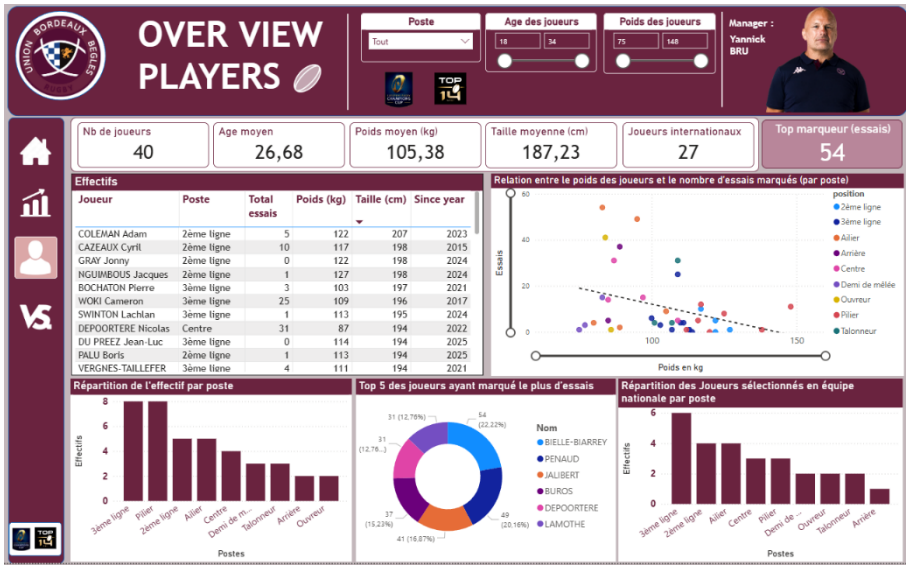


Onglet 3 : Statistiques individuelles (Joueurs)

C'est le cœur de notre base de données. Cet onglet permet d'explorer l'effectif avec des fonctionnalités de filtrage avancées :

- Filtrage par joueur spécifique.
- Segmentation par tranche d'âge.

- Analyse par poste ou nationalité.



Onglet 4 : Comparateur de joueurs

Pour aller plus loin dans l'analyse technique, nous avons créé un outil de comparaison directe (Head-to-Head). Il permet de sélectionner deux joueurs et de confronter leurs statistiques côte à côte, ce qui est particulièrement utile pour comparer des profils évoluant au même poste.

Résultat et interprétation :

L'aboutissement de ce projet est un rapport PowerBI interactif de quatre pages qui valorise les données structurées extraites du site officiel, dont la propreté initiale a limité le besoin de retraitement. L'analyse propose une navigation du général au particulier : elle débute par une vue macroscopique des performances du club (classement, ratio de victoires), avant d'explorer le cœur de l'effectif via des statistiques individuelles filtrables par âge ou poste. Pour affiner l'interprétation technique, nous avons intégré un comparateur "Head-to-Head" permettant de confronter graphiquement les profils de deux

joueurs. Ce tableau de bord valide notre chaîne de collecte en transformant les données brutes en un outil visuel et cohérent.

Parties individualisée/responsabilisée

Analyse et extraction

Antoine MAURIN : Je me suis occupé de cette partie en suivant différents tutos pour comprendre le fonctionnement du scraping en python et pour m'initier sur un cas simple avant de faire le scrap pour notre cas. https://www.youtube.com/watch?v=GjKQ6V_ViQE. J'avais pu partager les liens des vidéos utilisées à Rafaël afin que lui aussi puisse en faire. Rafaël a beaucoup aidé notamment dans l'identification des balises dans le code source. Ainsi, j'ai fait la base du scrap, et le code a été adapté en fonction de ce que nous voulions récupérer, notamment pour les images des joueurs que nous avons scrap après la construction du tableau de bord (c'est pour ça que nous n'avons pas juste rajouter dans le fichier de base).

Traitement et export

Rafaël SANZ : Je me suis chargé de faire tout le traitement des données pour que nous puissions faire notre tableau de bord sans difficultés liées à la gouvernance de celles-ci. Avec tout ce que nous avons pu faire en cours depuis le début du BUT, mon attention c'est tout de suite porté sur les formats dates qui sont très souvent dans le mauvais format. Cela n'a pas fait exception dans notre projet. Grâce à ce recul que j'ai tout de suite su prendre fasse aux données que nous avions pu extraire, j'ai instantanément pu corriger les problèmes futur potentiels.

Visualisation et Interprétation

Cette partie à été répartie entre nous deux en théorie. Cependant, il était compliqué de faire en sorte de se partager notre tableau de bord powerBI sans voir la version payante. Ainsi, Rafaël s'en est occupé (dans la réalisation). Quant à la réflexion, elle, s'est bien faite à deux sur le contenu qu'il y allait y avoir dans notre dashboard. Pour la thème et l'ambiance, Rafaël avait vu un tableau de bord réalisé par un autre sur un club de football qui lui a servi d'inspiration. Afin de ne pas être inutile sur cette partie, Antoine s'est occuper d'aller scraper les éléments manquants (images des joueurs pour la page de comparaison) et chercher les icones pour les mettre à disposition de Rafaël pour qu'il n'est pas à perdre de temps car l'élaboration du Tableau de bord a été très chronophage (beaucoup plus que prévu) car Rafaël est très perfectionniste.

Pour conclure sur le projet en général, nous sommes tous les deux d'accord pour dire qu'il était très intéressant et très complet. Nous avons pu apprendre énormément de choses qui pourront nous être utiles à l'avenir. Notre tableau de bord vient tout récupérer dans les fichiers disponibles sur notre Github, cependant, nous pourrions encore améliorer la pérennité du projet en faisant en sorte que lorsque que l'on scrap, les csv créés aillent directement dans un répertoire de notre repo ce qui n'est pas encore le cas pour le moment.