Mini-Projet: Exploratory Data Analysis

Aymeric Le Riboter
Tanguy Ducrocq

22 avril 2025

1 Introduction

Pour ce mini-projet d'Exploration de Données (EDA), nous avons choisi d'étudier les matchs de volley du tournoi international de la VNL (Volleyball Nations League). L'objectif de ce projet est d'analyser les performances des joueurs et des équipes sur plusieurs saisons du tournoi (2021 à 2023). À partir de ces données, nous chercherons à identifier des patterns, des tendances, et des facteurs influençant les résultats des matchs.

2 Présentation du dataset

Nous avons récupéré le dataset du tournoi international de la VNL sur Kaggle, sur trois saisons successives, de 2021 à 2023. Ce dataset se concentre exclusivement sur les joueurs masculins. Il se compose de deux fichiers :

- df_mens_indv_21_23.csv : contient les performances individuelles des joueurs masculins par année et par match.
- df_mens_rosters_21_23.csv : présente les compositions des équipes par année.

Amelioration du dataset

Le dataset que nous avons récupéré contient une grande quantité d'informations détaillées sur les statistiques des joueurs par match, telles que le nombre de smashs, de blocs, de réceptions, etc., ce qui nous permet de réaliser des analyses approfondies. Cependant, une information importante manquait dans ces données : le resultat du match. Grace au site Volleyball World on a pu récupérer les résultats manquants et compléter notre dataset. Ce site est le site officiel qui diffuse et répertori les résultats de championnats de volley du monde entier, ainsi que des compétitions internationales.
Grâce à une API, nous avons pu récupérer les données du site sous format JSON en faisant du webscrapping. Nous avons ainsi téléchargé ces fichiers pour les trois saisons qui nous intéressent, le but maintenant est de croiser les sources correctement pour avoir un dataset complet pour notre etude. Le site Volleyball World site ayant aussi été utilisé pour récupérer le dataset de Kaggle ainsi cela était plus simple pour croiser les sources.

Extraction des informations des fichiers JSON

Les informations qu'on veut récupérer du JSON sont les suivantes :
- Date du match
- Les équipes (TeamA, TeamB)
- Les scores des 2 équipes
- Le total des points de chaque équipe
- Les scores de chaque set avec un format explicit : (ex. : 25-18, 21-25...)

Une fois les données des matchs bien structurées, nous les avons fusionner avec les performances individuelles des joueurs. Nous avons relié les 2 dataframes en utilisant date du match et les noms équipes. Comme les deux datasets venaient de la même source, les noms des équipes (TeamA, TeamB) étaient les mêmes.

Vérification et validation des données

Nous avons fait un nettoyage des données en supprimant les colonnes inutiles et éliminant les lignes contenant des valeurs manquantes. Pour faciliter l'analyse, toutes les valeurs ont été converties en format numérique, en s'assurant qu'aucune donnée erronée ne vienne perturber les résultats.
A la fin de toutes ces étapes on fini avec un unique fichier avec toutes les informations nécessaire pour la suite.

3 Questions choisi

Les questions que nous allons traiter pour ce sujet sont :

1 Quels sont les joueurs les plus performant ?
2 La taille d'un joueur influence-t-elle le choix de son poste sur le terrain ?
3 Quels sont les facteurs clés qui déterminent la victoire d'une équipe ?

Question n°1 : Quels sont les joueurs les plus performant ?

1

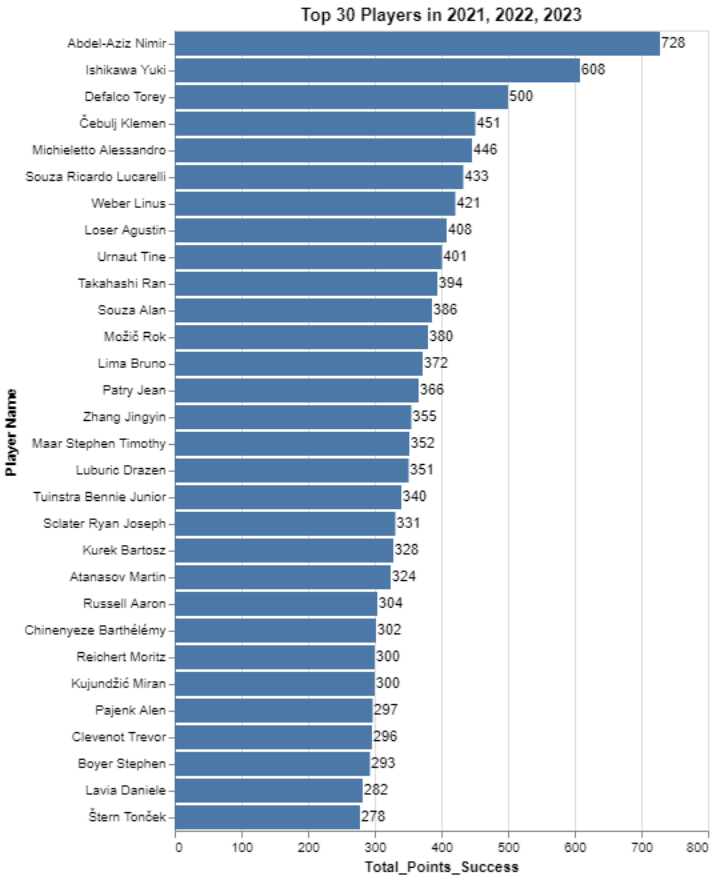

FIGURE 1 – Classement des joueurs selon le volume total de points marqués (2021–2023)

Pour commencer on a décidé d'étudier les joueurs ayant marqués les plus de points sur les 3 saisons. Pour cela on a représenté les 30 joueurs ayant marqués le plus de points sur les 3 années (2021-2023). Cette visualisation permet ainsi d'avoir un premier aperçu global des meilleurs joueurs en termes d'efficacité offensive sur la période étudiée. On peut remarquer un écart assez net entre le premier entre les 3 premiers, ensuite l'écart entre les joueurs se réduit.

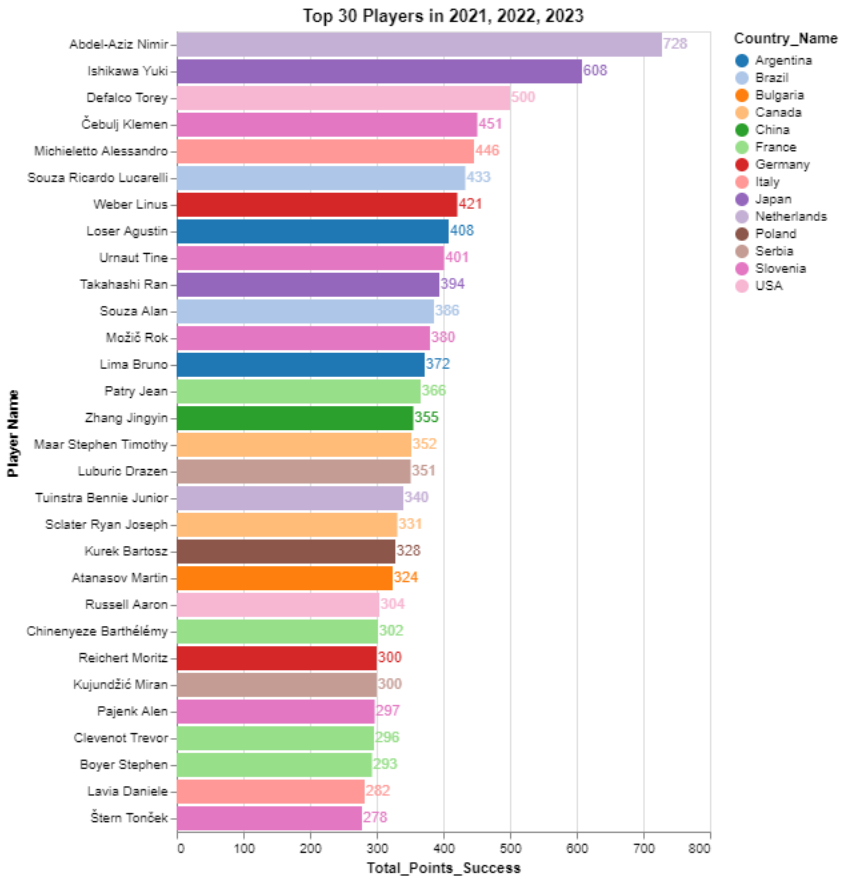

FIGURE 2 – Classement des joueurs selon le volume total de points marqués en fonction du pays (2021–2023)

Cet histogramme présente les 30 joueurs les plus performants en termes de points marqués avec succès sur les années 2021, 2022 et 2023. Par rapport à la représentation précédentes, on a rajouté le pays d'origine des joueurs pour faciliter l'analyse, cela permet aussi d'étudier la répartition des joueurs qui marquent beaucoup de points à travers le monde.
On observe qu'Abdel-Aziz Nimir (Pays-Bas) domine le classement avec un total de 728 points, suivi par Ishikawa Yuki (Japon) avec 608 points.
Ensuite, les écarts se resserrent avec des joueurs comme Defalco Torey (États-Unis) à 500 points et Čebulj Klemen (Slovénie) à 451 points.
La diversité des pays représentés met en évidence un haut niveau de performance réparti à l'échelle mondiale. On note aussi une forte présence de joueurs issus d'équipes européennes et sud-américaines comme l'Italie, le Brésil et l'Argentine.

2

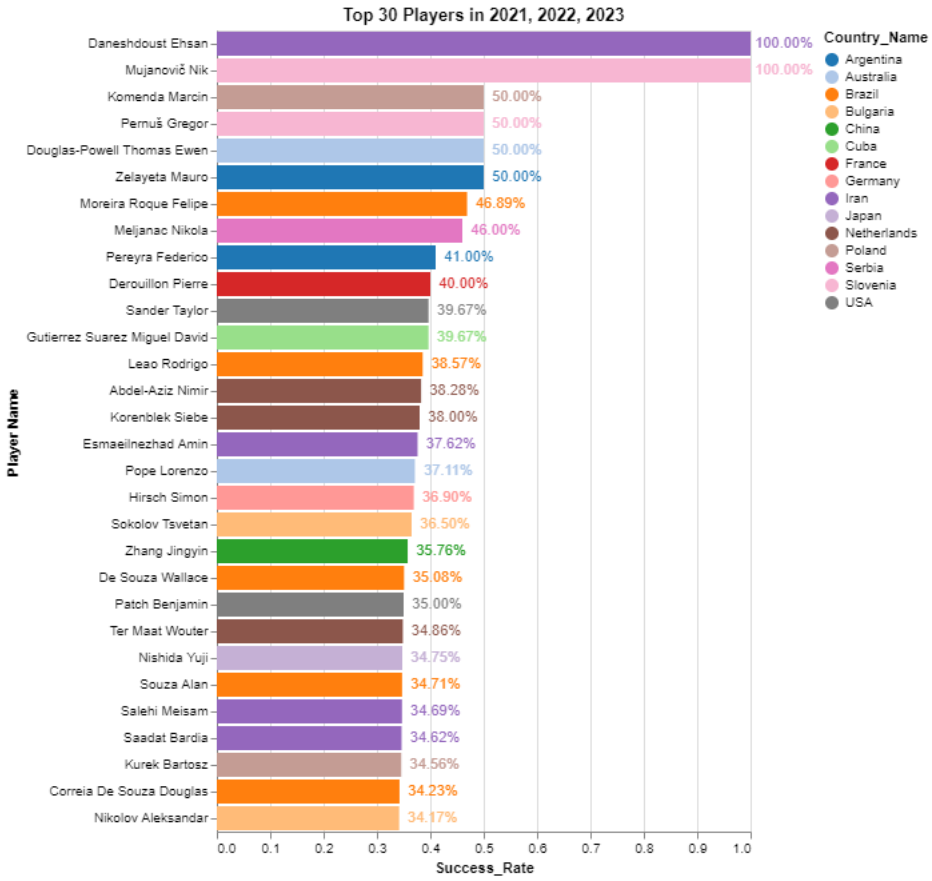

FIGURE 3 – Les joueurs les plus efficaces : classement par taux de réussite (2021–2023)

Ce deuxième histogramme propose une autre perspective de la performance des joueurs, cette fois-ci à travers leur taux de réussite (Success_Rate), c'est-à-dire la proportion de points marqués par rapport aux tentatives. Contrairement au graphique précédent basé sur le volume total de points, celui-ci met en avant l'efficacité des joueurs.
On remarque que Daneshdoust Ehsan (Iran) et Mujanović Nik (Serbie) affichent un taux de réussite parfait de 100 %, ce qui peut indiquer une très grande efficacité, bien que cela soit probablement lié à un nombre limité de tentatives (il suffit d'une seule tentative reussie pour avoir 100 % de reussite). Plusieurs joueurs comme Komenda Marcin (Pologne), Douglas-Powell Thomas Ewen (Australie) ou encore Zelayeta Mauro (Argentine) affichent un taux de 50 %, témoignant d'une efficacité solide.
À l'inverse, certains joueurs très présents dans le premier graphique, comme Abdel-Aziz Nimir (Pays-Bas), n'atteignent ici que 38,28 %, ce qui suggère une performance plus volumineuse qu'efficace. Cette visualisation complète donc la précédente en soulignant que les meilleurs scoreurs ne sont pas toujours les plus efficaces, et met en lumière des joueurs souvent moins visibles, mais très performants en proportion.

3

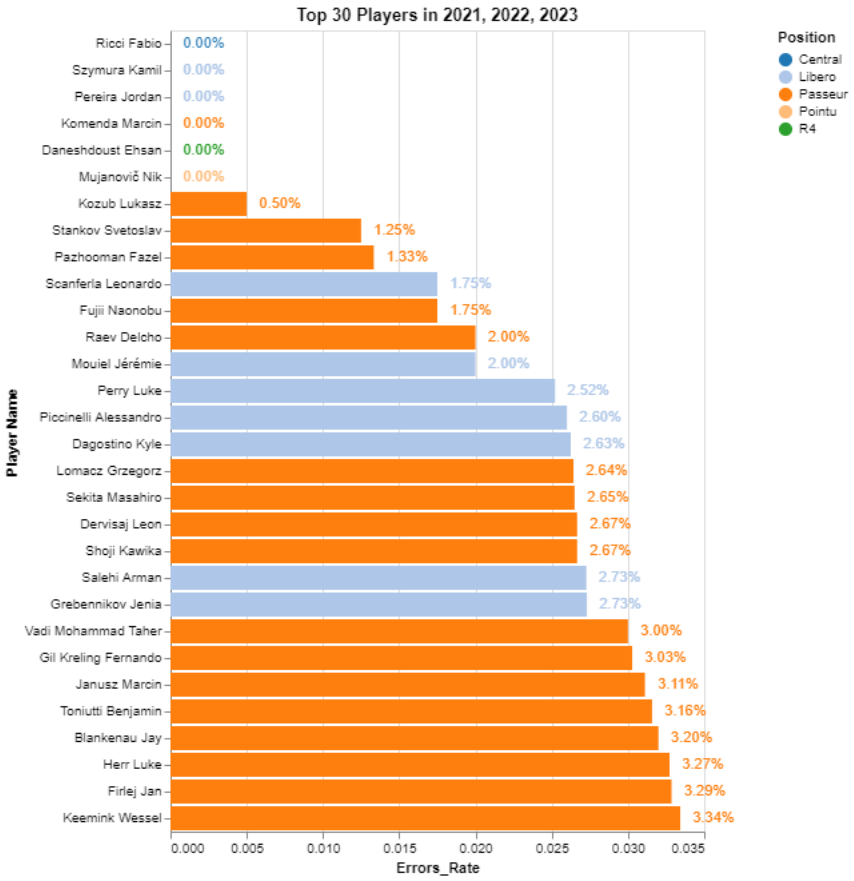

FIGURE 4 – Fiabilité des joueurs : classement par taux d'erreurs (2021–2023)

Ce troisième histogramme se concentre sur le taux d'erreurs (Errors_Rate) des joueurs les plus actifs entre 2021 et 2023, offrant une autre facette importante de leur performance : la fiabilité. Plus ce taux est faible, plus le joueur est considéré comme régulier et sûr dans ses actions.
On remarque que plusieurs joueurs, comme Ricci Fabio, Szymura Kamil ou encore Daneshdoust Ehsan, affichent un taux d'erreurs de 0 %, ce qui est remarquable. Cela dit, tout comme pour les taux de réussite parfaits dans le graphique précédent, ces chiffres peuvent être liés à une participation plus limitée. (Comme avant, 1 defense ou attaque reussie suffit pour avoir 0% d'erreur)
Keemink Wessel présente le taux d'erreurs le plus élevé avec 3,34 %, suivi de près par Fijel Jan et Herr Luke. Ce sont principalement des joueurs occupant le poste de passeur (orange), une position naturellement plus exposée aux fautes techniques et stratégiques.
Ce graphique met également en évidence la différence de profils selon les positions. Par exemple, les liberos (bleu clair), spécialistes de la réception et de la défense, tendent à avoir des taux d'erreurs plus bas, ce qui est cohérent avec leur rôle de defenseur.
Ainsi, cette visualisation complète les précédentes en apportant une lecture plus qualitative, et montre que la performance ne se résume pas uniquement à l'efficacité ou au volume, mais aussi à la capacité à minimiser les fautes.

Question n°2 : La taille d'un joueur influence-t-elle le choix de son poste sur le terrain ?

1

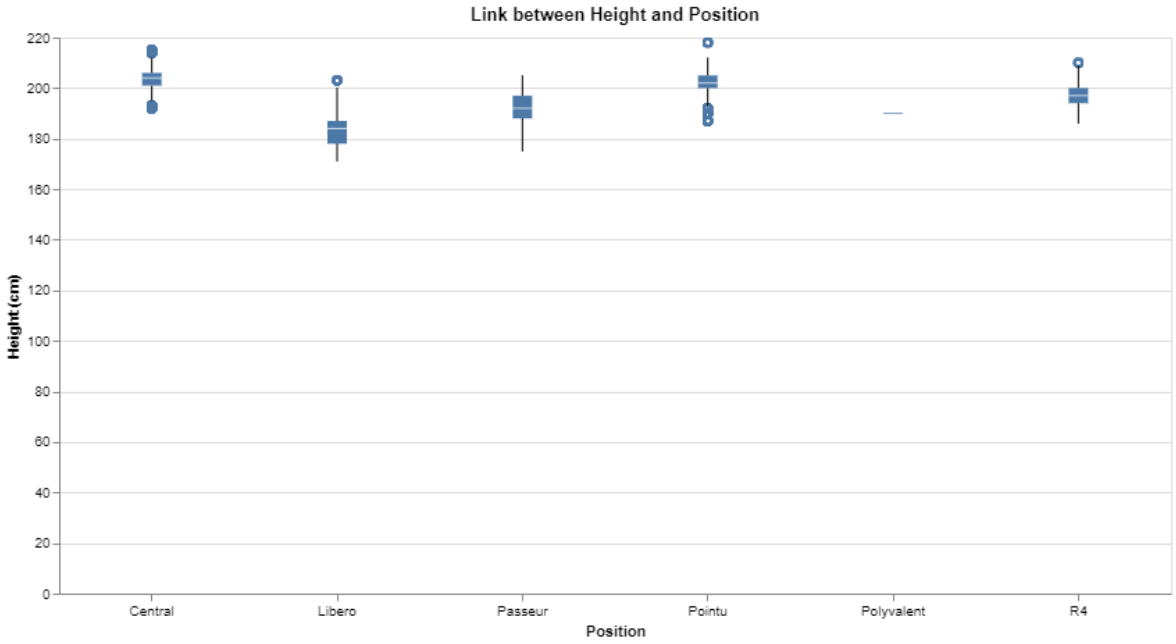

FIGURE 5 – Distribution des tailles en fonction des postes occupés

Le boxplot montre la relation entre la taille des joueurs (en centimètres) et leur position sur le terrain.
On observe que les joueurs occupant les postes de libero et de passeur sont en moyenne plus petits que ceux des autres postes.
Cette différence s'explique par leurs rôles spécifiques au sein de l'équipe. En effet, ces postes sont moins impliqués dans le jeu au filet : le libero a une fonction purement défensive, tandis que le passeur ne sert qu'à la construction du jeu, il participe peu à l'attaque et ne réalise généralement que des blocs.

2

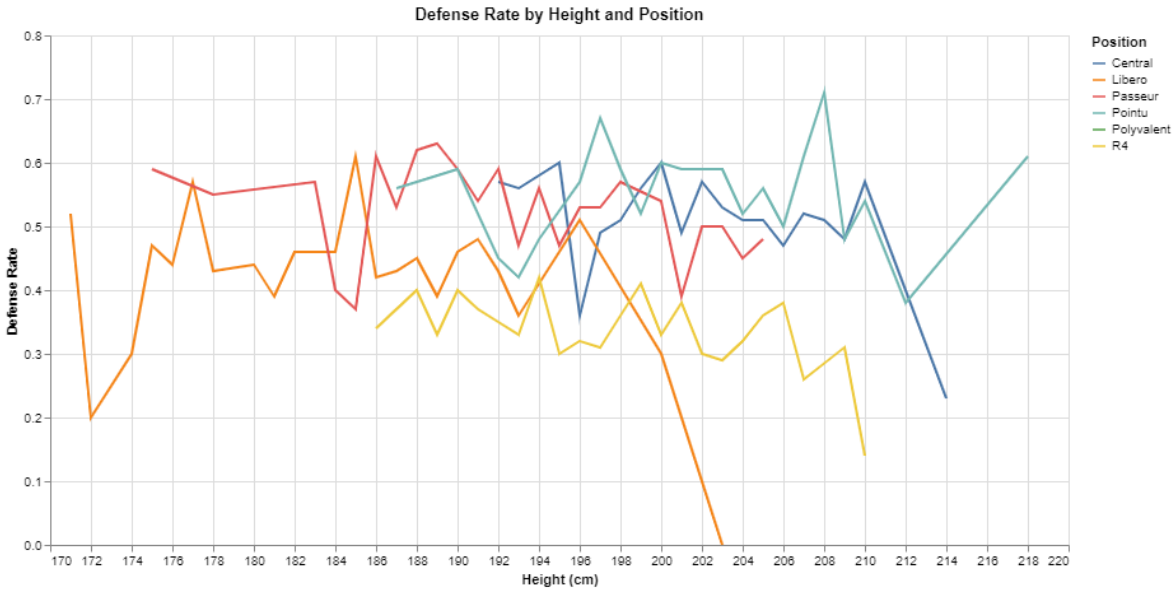

FIGURE 6 – Évolution du taux de défense en fonction de la taille et de la position

On observe que les liberos (courbe orange) ont généralement un taux de défense plus élevé, en cohérence avec leur rôle principalement défensif dans l'équipe. Ce taux semble relativement stable quelle que soit leur taille.
Les passeurs (courbe rouge) présentent également un bon taux de défense, ce qui s'explique par leur implication dans la récupération et la distribution du jeu.
À l'inverse, les joueurs plus grands, notamment les Centraux (courbe bleue) et les Pointus (courbe verte), ont un taux de défense plus variable et parfois plus faible. Cela s'explique par leur rôle offensif, les rendant moins impliqués dans la défense de fond de terrain.

3

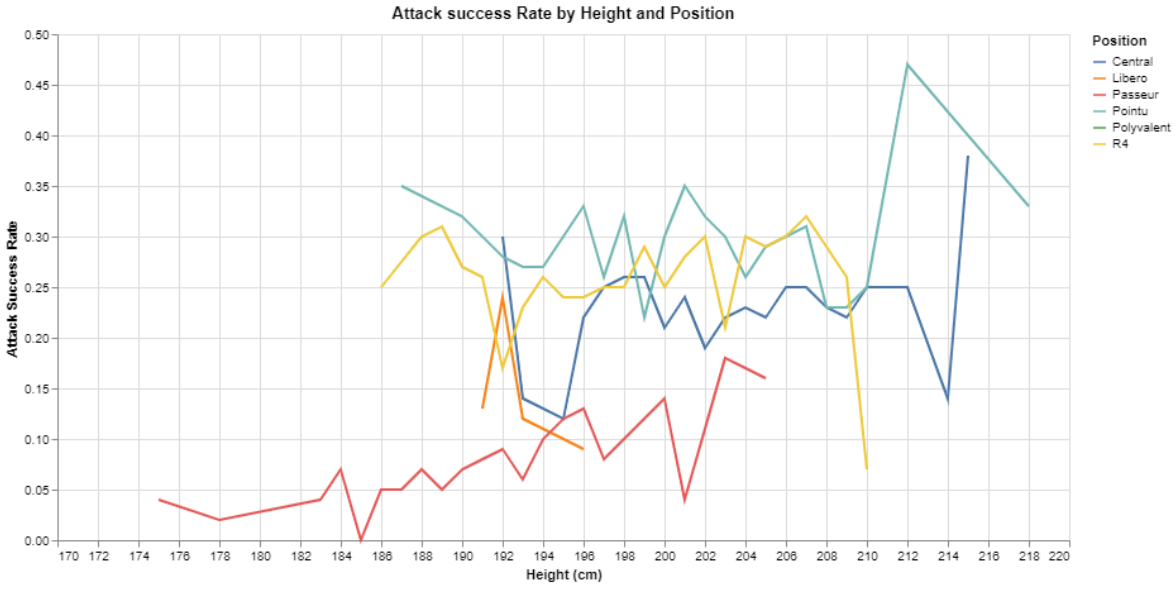

FIGURE 7 – Influence de la taille et de la position sur le taux de réussite en attaque

On observe une tendance nette où les joueurs de plus petite taille, notamment les liberos (courbe rouge), ont un taux de réussite en attaque extrêmement faible. Cela s'explique par le fait que les liberos n'ont pas un rôle offensif et ne sont pas (ou alors tres rare) autorisés à attaquer au filet.
De leur côté, les passeurs (courbe orange) montrent également un taux de réussite assez bas, ce qui est logique puisqu'ils se concentrent principalement sur la distribution du jeu plutôt que sur l'attaque.
À l'inverse, les Centraux (courbe bleue), les Pointus (courbe verte) et les Polyvalents (courbe jaune) affichent des taux de réussite plus élevés. Cela confirme que les postes impliqués dans l'attaque sont généralement occupés par des joueurs plus grands, leur permettant d'être plus efficaces au filet.
On remarque aussi une augmentation du taux de réussite à mesure que la taille des joueurs augmente, en particulier pour les Pointus et les Centraux, qui atteignent les meilleures performances.

Question n°3 : Quels sont les facteurs clés qui déterminent la victoire d'une équipe ?

1

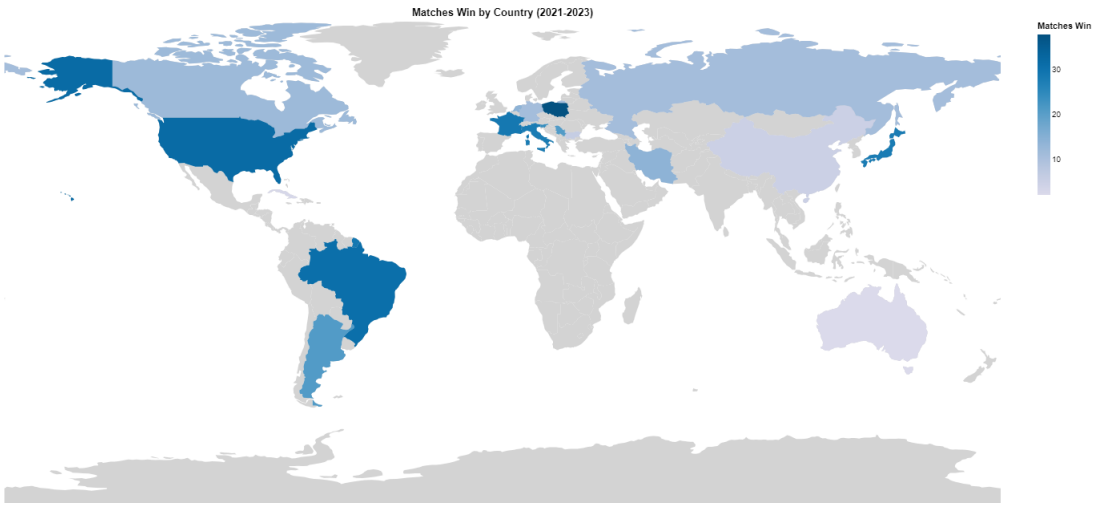

FIGURE 8 – Carte du nombres de matchs de VNL gagnés

Sur cette carte du monde on peut voir toutes les équipes qui ont participé au tournois sur les années 2021-2023. Cela nous permet déjà d'identifier l'équipe qui a gagné le plus de matchs et celle qui en a gagné le moins.
On remarque une forte présence des pays américains et européens, qui obtiennent de bons résultats, tandis que les pays africains sont totalement absents. Par ailleurs, seuls quelques pays d'Asie et d'Océanie, comme la Chine, le Japon et l'Australie, participent.

2

Dans cette section on va étudier les facteurs qui ont influencés les performances de la meilleure équipe (celle qui a gagné le plus de matchs) et de la plus mauvaise (celle qui en a gagné le moins).

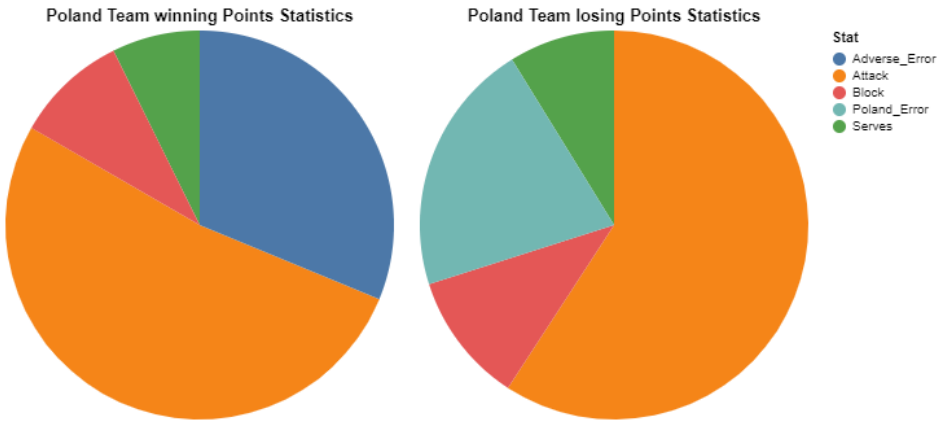

FIGURE 9 – Répartition des facteurs qui ont influencés les points gagnés et perdus par la Pologne

Nous avons tout d'abord analysé la répartition des points gagnés par la Pologne, puis celle des points qu'elle a concédés. Il en ressort que la majorité des points marqués par la Pologne provient de ses attaques. Environ 30 % des points gagnés sont dus à des erreurs adverses, ce qui suggère que les Polonais parviennent à provoquer des fautes, qu'il s'agisse de réceptions ratées ou de blocs mal maîtrisés qui sortent du terrain.
Concernant les points perdus, la répartition est relativement similaire : la Pologne concède principalement des points sur des attaques qu'elle n'a pas su défendre, ainsi que sur ses propres erreurs.

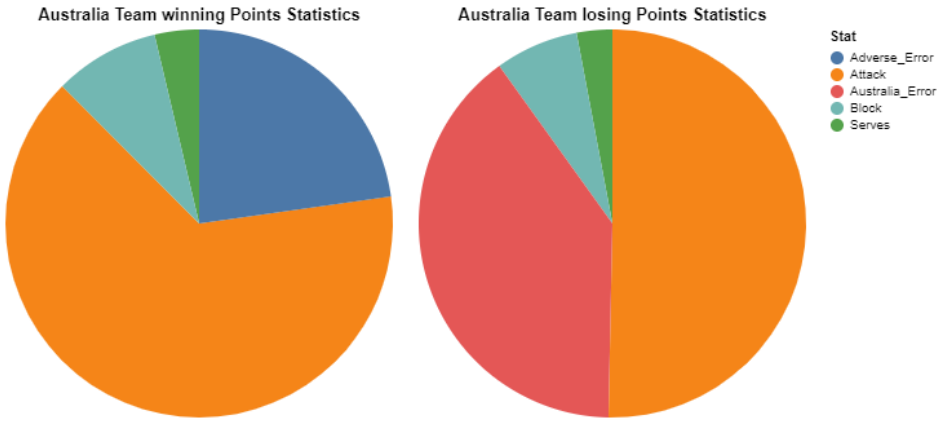

FIGURE 10 – Répartition des facteurs qui ont influencés les points gagnés et perdus par l'Australie

L'équipe ayant perdu le plus de matchs est l'Australie, ce qui permet de tirer des informations pertinentes sur les actions déterminantes pour une victoire.

La majorité des points marqués par l'Australie proviennent de son attaque, représentant près de 65 % du total. Les erreurs des adversaires comptent pour environ 20 %, tandis que les blocs et les services contribuent à hauteur de 15 %.

Cependant, l'Australie concède un nombre important de points sur ses propres erreurs, soit près de 40 %. Cela signifie que presque un point sur deux perdu aurait potentiellement pu être évité et transformé en action offensive ou défensive (attaque, bloc, etc.). Une autre part importante des points perdus provient des attaques adverses non défendues, qui représentent également plus de 50 % des points encaissés.

Enfin, les points perdus sur les services et les blocs restent marginaux, représentant à peine 10 % du total.

Comparaison Pologne Australie

Si on compare les répartitions des points marqués par la Pologne et l'Australie, on peut remarquer que les facteurs qui influencent ces points gagnés sont a peu près classés dans le même ordre, mais avec des ordres de grandeurs assez différents. Ainsi la Pologne marque prés de 30% de leur point dû à des erreurs adverses contre seulement 20% pour l'Australie donc on peut se dire qu'ils provoquent plus les erreurs adverses. Cela peut être du a de très bonnes attaques qui vont provoquer des blocks out ou des défenses loupés, ou encore de bons services qui empêchent une bonne réception. On peut par contre remarqué que l'Australie marque plus de point en attaquant.

Si on compare maintenant les points perdus par les deux équipes, on se rend compte que la proportion de points perdus par l'Australie dû à des erreurs de leur part est bien plus importante que pour la Pologne, ainsi si l'Australie est l'équipe qui a perdu le plus de matchs cela peut venir du fait qu'ils font beaucoup d'erreurs.

3

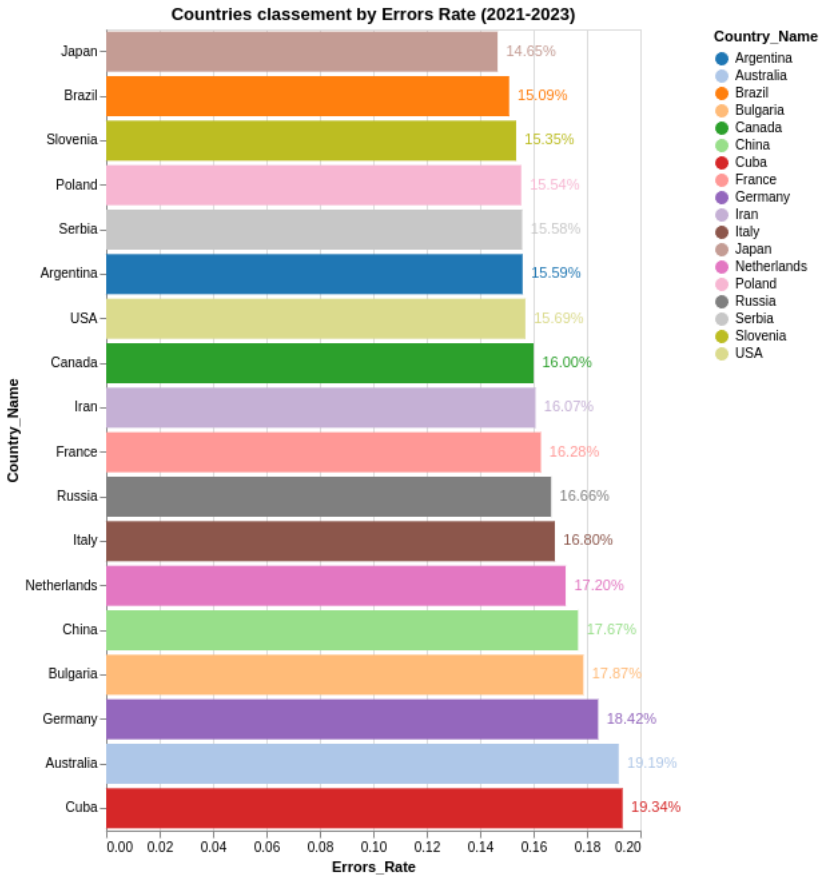

FIGURE 11 – Classement du taux d'erreurs par pays

Après avoir comparé les facteurs qui ont influencés les points gagnés et perdus par la meilleure et la pire équipe, on a pu remarqué que le facteur qui était le plus impactant sur les points perdus était les erreurs qu'une équipe fait, ce qui veut dire les points qu'elle "donne" à ses adversaires.

On a donc décidé d'étudier cette métrique pour les différentes équipes. Si on compare ce graphe avec la carte des victoires par nation, on se rend compte d'une certaine corrélation entre le taux d'erreurs d'une équipe et le nombre de matchs qu'elle a gagné.

Ainsi des équipes comme le Japon, le Brésil et la Pologne qui font peut d'erreurs sont des équipes qui comptent le plus grand nombre de matchs remportés. Et les équipes comme Cuba et l'Australie qui font beaucoup d'erreurs comptabilisent très peu de victoires.

4 Conclusion

Durant cette études on a pu voir dans un premier temps quels étaient les meilleurs joueurs en se basant sur différentes métriques : tout d'abord le nombre d'attaque réussi, puis le taux de succès des attaques, et enfin le taux d'erreurs des joueurs.

Dans un second temps on a étudier le rôle de la taille sur le poste d'un joueur. On a pu comparer par poste l'impact de la taille sur le taux de réussite des attaques et sur le taux d'erreurs commises par le joueur.

Enfin on a pu étudier les facteurs qui impacts les victoires et les défaites des équipes, pour cela on a étudier la meilleure et la pire équipe en se basant sur le nombre de victoire, puis on a réalisé une étude sur le facteur qui avait l'air d'avoir le plus d'impact sur les points perdus par une équipe (le taux d'erreur).