Au-delà des Courts : Exploration des Données des Meilleurs Joueurs de l'Association of Tennis Professionals (ATP) et de la Women's Tennis Association (WTA)

Projet de Web Scraping en Python, Second Semestre

Faculté des Sciences Économiques de Rennes 1

Sommaire

I. Introduction

- Contexte et objectifs du travail

Dans le cadre de notre projet de Web Scraping en Python, mené au cours du second semestre, nous nous sommes plongés dans une exploration approfondie des données des meilleurs joueurs de l'Association of Tennis Professionals (ATP) et de la Women's Tennis Association (WTA). Sous la direction de Joeffrey Drouard, notre objectif était de rassembler des informations exhaustives sur les profils des joueurs de tennis de haut niveau, notamment leurs noms, dates de naissance, classements ATP/WTA, performances en tournois, caractéristiques physiques et bien plus encore. Nous cherchions à comprendre en profondeur les tendances, les variations et les caractéristiques distinctives des joueurs de tennis élite, en utilisant des techniques de Web Scraping pour collecter et analyser les données pertinentes. Notre étude, intitulée "Au-delà des Courts : Exploration des Données des Meilleurs Joueurs de l'ATP et de la WTA", visait à fournir un aperçu statistique approfondi du monde du tennis professionnel, allant au-delà des simples performances sur le terrain pour explorer les différents aspects qui influent sur la carrière et le succès des athlètes.

- Présentation de la démarche adoptée

Notre projet d'exploration des données des joueurs de tennis élite s'est déroulé en trois phases distinctes, chacune étant essentielle pour atteindre nos objectifs. La première étape a été la collecte des données, où nous avons utilisé diverses techniques de web scraping pour extraire des informations détaillées à partir de deux sources fiables telles que les sites officiels des tournois et les classements ATP/WTA. Cette phase a nécessité une attention particulière à la sélection des données pertinentes et à l'optimisation des scripts de scraping pour assurer une collecte précise et efficace.

Une fois les données collectées, nous avons entamé la phase de nettoyage, où nous avons travaillé à préparer les données pour l'analyse. Cela comprenait des tâches telles que l'élimination des doublons, le traitement des valeurs manquantes, la normalisation des formats et la vérification de la cohérence des données. Cette étape était cruciale pour garantir la qualité et la fiabilité de nos données avant de passer à l'analyse proprement dite.

Enfin, la troisième phase a été celle de l'analyse des données, où nous avons exploré en profondeur les profils et les performances des joueurs de tennis élite. À l'aide de techniques de visualisation et de statistiques descriptives, nous avons identifié des tendances, des corrélations et des insights pertinents sur les facteurs qui influent sur la réussite des athlètes au plus haut niveau. Cette phase nous a permis de tirer des conclusions significatives et de fournir des perspectives précieuses sur le paysage du tennis professionnel.

II. Méthodologie

- Description des données utilisées

Nous avons utilisé des sources de données fiables et officielles pour notre projet d'exploration des données des joueurs de tennis élite. Pour les données masculines, nous avons extrait les informations à partir du site officiel de l'Association of Tennis Professionals (ATP) à l'adresse suivante : https://www.atptour.com/en/rankings/singles. Sur ce site, nous avons collecté les données en parcourant les fiches individuelles des joueurs, ce qui nous a permis d'obtenir des détails précis sur leurs performances, leurs classements, leurs titres et d'autres informations pertinentes pour notre analyse.

Quant aux données féminines, elles ont été obtenues à partir du site officiel de la Women's Tennis Association (WTA) via le lien : https://www.wtatennis.com/rankings/singles. De manière similaire à la méthode utilisée pour les joueurs masculins, nous avons récupéré les données en explorant les fiches individuelles des joueuses, ce qui nous a fourni une mine d'informations sur leurs classements, leurs résultats récents, leurs titres et d'autres caractéristiques pertinentes pour notre étude.

Hommes	
Nom	Classement ATP
Victoires/Défaites (Carrière)	Prix en argent
Pays	Âge
Poids	Taille
Année de début professionnel	Victoires/Défaites (Cette année)
Femmes	
Nom	Date de naissance
Âge	Taille
Préférence de main	Origine
Classement	Cash prize
Titres en simple cette année	Victoires cette année

Les données comprennent des informations détaillées sur les joueurs de tennis masculins et féminins. Pour les hommes, les données incluent le nom, le classement ATP, les victoires/défaites en carrière, le prix en argent, le pays d'origine, l'âge, le poids, la taille, l'année de début professionnel et les victoires/défaites de l'année en cours. Pour les femmes, les données comprennent le nom, la date de naissance, l'âge, la taille, la préférence de main, l'origine, le classement, les gains en argent, les titres en simple cette année et les victoires cette année. Ces informations offrent une perspective complète des performances et des caractéristiques des joueurs et joueuses de tennis, utiles pour l'analyse et la modélisation dans le domaine du tennis professionnel.

- Explication des outils et techniques employés

Dans notre projet, nous avons principalement utilisé Selenium pour collecter les données à partir des sites de l'ATP et de la WTA. La principale raison derrière l'utilisation exclusive de Selenium était la complexité des sites, où les éléments étaient générés dynamiquement à l'aide de JavaScript. En raison de cette particularité, Selenium s'est avéré être la seule solution viable pour automatiser l'interaction avec le navigateur et extraire les données requises de manière fiable et efficace.

Nous avons évité l'utilisation de BeautifulSoup et des expressions régulières (regex) dans ce projet en raison de cette complexité. BeautifulSoup est plus adapté à l'analyse de contenu statique, tandis que les expressions régulières sont efficaces pour extraire des motifs spécifiques de données textuelles. Cependant, étant donné que les sites de l'ATP et de la WTA utilisent des éléments dynamiques générés par JavaScript, ces méthodes ne seraient pas adéquates pour extraire les informations souhaitées. De plus, la nature dynamique des pages aurait rendu difficile l'utilisation de techniques basées sur le contenu statique en html.

Pour le scrapping des données des joueurs (homme), nous avons utilisé une combinaison de Selenium et de proxies pour collecter les données des joueurs de tennis à partir du site de l'ATP. Tout d'abord, nous avons importé les modules nécessaires, y compris requests pour tester la validité des proxies, et les bibliothèques de Selenium pour l'automatisation du navigateur. Nous avons également défini le chemin du chromedriver pour le contrôle du navigateur Chrome. Ensuite, nous avons utilisé la bibliothèque pandas pour importer la liste des URL des joueurs depuis un fichier Python, et avons défini une fonction pour tester la validité des proxies en utilisant le site https://free-proxy-list.net/.

Dans la boucle principale, nous avons itéré à travers chaque URL de joueur et avons tenté de récupérer les données en utilisant les proxies disponibles. Pour chaque itération, nous avons sélectionné un proxy valide parmi la liste des proxies testés avec succès. Nous avons initialisé un navigateur Chrome avec les options de proxy, puis chargé la page de l'URL du joueur. Après avoir attendu 2 secondes pour permettre le chargement complet de la page, nous avons extrait les données du joueur en utilisant les sélecteurs CSS appropriés avec Selenium.

Une fois les données extraites, nous les avons écrites dans un fichier CSV en utilisant la bibliothèque csv. En cas d'erreur lors du processus de scraping, nous avons capturé l'exception et avons essayé avec un autre proxy jusqu'à ce qu'une page soit correctement traitée. Si aucun proxy n'était disponible ou si une autre erreur survenait, un message approprié était affiché. Enfin, nous avons fermé le navigateur une fois le processus terminé pour nettoyer les ressources système utilisées. Cette approche nous a permis de collecter efficacement les données des joueurs de tennis tout en contournant les défis liés à la détection des robots et à l'utilisation de proxies pour assurer une collecte de données fluide et fiable.

Pour le scraping des données des joueuses de tennis (femme), nous avons aussi utilisé Selenium en combinaison avec le navigateur Chrome pour automatiser le processus de navigation sur les pages web et l'extraction des informations. Tout d'abord, nous avons défini une fonction `scraper_informations_joueuses()` qui prend en entrée une liste d'URL des joueuses. À l'intérieur de cette fonction, nous avons initialisé le navigateur Chrome et ouvert un fichier CSV pour enregistrer les données extraites.

Ensuite, nous avons bouclé à travers chaque URL des joueuses et utilisé les méthodes de Selenium pour extraire les informations pertinentes. Pour chaque joueuse, nous avons extrait son nom, sa date de naissance, son âge, sa taille, sa préférence de main, son origine, son classement, le montant total de ses gains en prix, les titres en simple de cette année, et les victoires de cette année. Nous avons utilisé les sélecteurs CSS et XPath appropriés pour cibler les éléments spécifiques sur la page.

Une fois les données extraites pour une joueuse, nous les avons écrites dans le fichier CSV en utilisant la bibliothèque csv. En cas d'erreur lors du scraping d'une information spécifique, nous avons capturé l'exception correspondante et affiché un message d'erreur approprié. Enfin, une fois le scraping terminé pour toutes les joueuses, nous avons fermé le navigateur pour nettoyer les ressources système utilisées.

IV. Analyse des données

- Interprétation des graphiques et des tendances observées

Graphique 1 : ATP Nombre de joueurs ayant commencé leur carrière professionnelle chaque année

Ce graphique représente le nombre de joueurs de tennis masculins qui ont commencé leur carrière professionnelle chaque année, selon les données de l'Association of Tennis Professionals (ATP). On observe une tendance relativement stable de nouveaux joueurs entrant sur le circuit professionnel entre 2002 et 2007, avec un nombre qui ne dépasse jamais quatre par an. Cependant, à partir de 2008 jusqu'en 2010, il y a eu une augmentation significative, culminant en 2010 avec plus de 10 nouveaux joueurs. Cette hausse pourrait indiquer un intérêt croissant pour le tennis professionnel ou des changements dans les critères de qualification pour l'ATP. Après 2010, le nombre de nouveaux joueurs par an fluctue, avec une autre pointe remarquable en 2016. Le graphique montre également une baisse notable à partir de 2018, atteignant son point le plus bas en 2023. Cette diminution pourrait être due à divers facteurs, tels que des modifications des règlements de l'ATP, des barrières économiques plus élevées pour entrer dans le professionnalisme, ou des événements mondiaux ayant un impact sur le sport.

Graphique 2 : ATP Nombre de victoires / défaites des joueurs cette année

Le graphique 2 illustre le nombre de victoires et défaites des joueurs de l'ATP pour l'année en cours. Sans surprise, les joueurs qui dominent le haut du classement se distinguent au début du graphique, démontrant un nombre élevé de victoires par rapport aux défaites. Cette représentation confirme la corrélation attendue entre le classement des joueurs et leurs performances sur le court. Une mention spéciale est accordée à Jannik Sinner, qui affiche une forme exceptionnelle en ce début d'année 2024. Sa position sur le graphique suggère un nombre impressionnant de victoires, ce qui pourrait être indicatif d'une progression remarquable et d'une saison potentiellement très réussie. Le déclin progressif dans le nombre de victoires à mesure que l'on avance sur l'axe horizontal met en lumière la compétitivité intense au sein du circuit, avec des joueurs luttant pour améliorer leur classement et leur performance globale. Il est également pertinent de noter que cette représentation visuelle fournit un instantané dynamique qui pourrait évoluer à mesure que la saison avance et que les joueurs participent à plus de tournois.

Graphique 3 : ATP Performance de tennis par pays

Le graphique 3 présente une vue d'ensemble des performances des joueurs de tennis par pays, en se basant sur le ratio entre le nombre de victoires/défaites et les gains en argent pour l'année en cours. Encore une fois, nous voyons que Jannik Sinner surpasse nettement ses concurrents, avec une performance exceptionnelle qui le place en tête non seulement en nombre de victoires mais

aussi en termes de prix en argent gagnés. Son point est nettement plus élevé sur l'axe vertical, ce qui implique qu'il a généré des gains significatifs par rapport à ses pairs.

En outre, le graphique indique des joueurs comme Novak Djokovic, Daniil Medvedev, Alexander Zverev et Carlos Alcaraz, tous situés dans la zone supérieure du graphique, reflétant leur succès tant sur le plan sportif que financier. Les couleurs représentent les différents pays, permettant de distinguer facilement les performances par nation. La position élevée de certains joueurs comme Taylor Fritz et Alex de Minaur souligne également les réussites notables des États-Unis et de l'Australie dans le tennis professionnel cette année.

Ce type de visualisation aide à mettre en évidence la corrélation entre la performance sur le court et la rentabilité financière des joueurs, tout en fournissant un aperçu de la répartition du talent à travers les différentes nations. Elle illustre l'importance de la réussite dans les tournois majeurs, qui se traduit directement par des récompenses financières élevées, et comment certains joueurs, comme Sinner, peuvent dominer la saison non seulement en termes de victoires mais aussi en attirant des revenus considérables.

Graphique 4 : ATP Distribution Age, Poids et taille des joueurs

Le graphique 4 présente les boîtes à moustaches pour la distribution de l'âge, du poids et de la taille des joueurs de l'Association of Tennis Professionals (ATP). Ces graphiques sont particulièrement utiles pour comprendre la répartition et la dispersion des données.

En ce qui concerne l'âge, on remarque que la médiane, indiquée par la ligne orange à l'intérieur de la boîte, se situe un peu au-dessus de 25 ans. Le premier quartile (Q1) se trouve aux alentours de 22 ans, ce qui signifie que 25 % des joueurs sont plus jeunes que cet âge. Le troisième quartile (Q3), quant à lui, se situe juste en dessous de 30 ans, indiquant que 75 % des joueurs ont moins de 30 ans. Les points au-dessus de la moustache supérieure représentent des valeurs aberrantes, montrant qu'il existe des joueurs qui sont bien au-dessus de l'âge médian de l'ATP, certains ayant plus de 35 ans (Novak Djokovic).

Pour le poids des joueurs, la médiane se situe près de 80 kg. Le premier quartile est autour de 75 kg, ce qui implique que 25 % des joueurs pèsent moins que cela. Le troisième quartile est près de 85 kg, ce qui signifie que 75 % des joueurs ont un poids inférieur à ce chiffre. La distribution semble assez symétrique sans valeurs aberrantes significatives, ce qui suggère une certaine homogénéité dans le poids des joueurs de tennis professionnels.

Enfin, pour la taille, la médiane est légèrement en dessous de 1,85 m. Le premier quartile est d'environ 1,80 m et le troisième quartile est d'environ 1,90 m. La présence d'une valeur aberrante indique l'existence d'un ou plusieurs joueurs dont la taille est bien inférieure à celle du premier quartile.

Dans l'ensemble, ces graphiques montrent que bien qu'il y ait une variété dans les caractéristiques physiques des joueurs de tennis professionnels, il existe une tendance centrale autour de laquelle la majorité des joueurs se regroupe. Les valeurs aberrantes dans les trois

catégories indiquent également que des joueurs ayant des attributs physiques moins typiques font partie du circuit ATP.

Graphique 5 : WTA Répartition des préférences de main des joueuses

Le graphique 5 illustre la répartition des préférences de main chez les joueuses de la Women's Tennis Association (WTA). On constate une nette prédominance des joueuses droitières par rapport aux gauchères. Avec une majorité écrasante de joueuses ayant une préférence pour la main droite, cela reflète une tendance commune dans la population générale où la droitierie est plus fréquente que la gaucherie.

Le nombre de joueuses gauchères est bien plus faible, ce qui est représentatif des statistiques mondiales sur la latéralité; seulement une petite fraction de la population est naturellement gauchère. Cette différence peut introduire une dynamique intéressante dans le jeu, car les joueuses gauchères sont souvent considérées comme ayant un avantage tactique du fait de la rareté et donc de l'inhabitualité de leur style de jeu pour la plupart des adversaires.

Cette distribution peut également avoir des implications pour l'entraînement et la préparation des joueuses, car affronter une gauchère peut nécessiter une stratégie différente. En conclusion, le graphique souligne l'asymétrie significative entre les joueuses droitières et gauchères dans le tennis professionnel féminin.

Graphique 6 : WTA Répartition des nationalités en fonction des titres et du cash prize

Le graphique 6 illustre la répartition des victoires/titres et des prix en argent accumulés par les joueuses de tennis de la Women's Tennis Association (WTA) pour les 20 nations principales. Sur l'axe horizontal, nous avons le prix en argent sur une échelle logarithmique, ce qui permet de mieux visualiser la distribution des gains qui s'étendent sur de grandes étendues de valeurs. L'axe vertical montre le total des matches gagnés.

Chaque cercle représente une joueuse, avec sa couleur indiquant son pays d'origine et la taille du cercle reflétant le nombre total de matches qu'elle a gagnés. On peut observer une diversité de nationalités parmi les joueuses les plus performantes, avec une présence notable de joueuses de l'Europe de l'Est, comme illustré par les cercles en provenance de Biélorussie, de Russie et d'Ukraine.

Les joueuses provenant de grandes villes telles que New York et Moscou semblent avoir un nombre de victoires élevé, indiqué par les grands cercles, et également des gains en prix d'argent considérables. On note également la présence de joueuses de pays moins traditionnels dans le tennis de haut niveau, comme la Chine ou l'Argentine, ce qui indique une mondialisation croissante du sport.

Ce graphique est particulièrement utile pour évaluer comment les performances sur le court se traduisent en termes financiers et pour observer la distribution globale du succès dans le tennis

féminin. Il met en évidence que, si les joueuses des nations traditionnellement fortes dans le tennis ont tendance à accumuler plus de gains, il existe une variété de pays représentés, suggérant une concurrence et un talent mondialement diversifiés.

Graphique 7 : Comparaison ATP / WTA Histogramme de l'âge moyen et répartition des nationalités dans le top 100

Le graphique 7 présente une comparaison entre les joueurs de l'Association of Tennis Professionals (ATP) et les joueuses de la Women's Tennis Association (WTA) en termes d'âge moyen et de répartition des nationalités dans le top 100.

Dans la première partie de l'histogramme, on observe que l'âge moyen des joueurs ATP est légèrement inférieur à celui des joueuses WTA, ce qui peut refléter des différences dans les parcours de carrière ou dans la longévité des joueurs dans chaque circuit.

La répartition des nationalités pour le WTA est illustrée par une camembert, où l'on peut constater une grande diversité, avec une dominance notable des joueuses de Moscou, Russie, représentant près d'un tiers des joueuses. D'autres villes comme Minsk, Dallas et Nanjing ont également une représentation significative, montrant que le tennis féminin professionnel bénéficie d'un talent issu d'une variété de contextes géographiques. De plus le marqueur "-" en orange dans le camembert de la répartition des nationalités pour les joueuses WTA représente les joueuses de nationalité russe. Cette notation spéciale a été utilisée parce que leur statut de nationalité a été officiellement retiré par la fédération à la suite de l'invasion russe en Ukraine. Ainsi, bien que ces joueuses soient de nationalité russe, elles sont répertoriées sans une désignation de pays spécifique pour refléter les sanctions sportives internationales imposées à la Russie. Cette situation souligne les répercussions politiques sur le sport et comment des événements mondiaux peuvent influencer la représentation des athlètes dans les compétitions internationales.

Pour l'ATP, la répartition des nationalités est présentée dans un deuxième diagramme en camembert. Les États-Unis mènent avec une part conséquente des joueurs, suivis par la France et d'autres pays européens tels que l'Allemagne et l'Espagne. Cette représentation souligne l'influence de certaines nations qui ont traditionnellement une forte présence dans le tennis masculin.

En somme, ces visualisations offrent un aperçu de la démographie et de la distribution géographique des joueurs et joueuses de tennis de haut niveau, révélant à la fois les similitudes et les distinctions entre les circuits masculin et féminin.

V. Difficultés rencontrées

- Identification des obstacles et des problèmes rencontrés durant l'étude

Dans la section consacrée aux difficultés rencontrées lors de notre étude, l'une des premières embûches auxquelles nous avons été confrontés était l'erreur de protocole 403 que nous avons rencontrée lors de l'utilisation de la bibliothèque requests pour récupérer les données à partir du site de l'ATP. L'erreur 403, également connue sous le nom de "Forbidden", est un code d'état HTTP qui indique que le serveur a compris la requête de l'utilisateur, mais refuse de la traiter. Cela peut se produire pour diverses raisons, notamment lorsque l'utilisateur n'est pas autorisé à accéder à la ressource demandée. Dans notre cas, cette erreur a été déclenchée par des restrictions d'accès imposées par le serveur, ce qui nous a empêchés d'accéder aux données souhaitées.

Un autre problème majeur auquel nous avons été confrontés a été la détection par le site que nous agissions comme un robot, ce qui a entraîné la présentation de défis supplémentaires sous forme de résolutions de captcha. Cette détection automatisée des robots vise à protéger les sites web contre le scraping automatisé et à garantir une expérience utilisateur optimale. Cependant, pour nous en tant que chercheurs de données, cela a posé des difficultés considérables, car la résolution des captchas manuellement pouvait s'avérer fastidieuse et chronophage. Cette situation a ralenti notre processus de collecte de données et a nécessité une attention particulière pour trouver des solutions permettant de contourner ces obstacles tout en respectant les politiques et les règlements du site.

- Propositions de solutions ou d'alternatives pour surmonter ces difficultés

Après avoir identifié l'erreur 403 comme un obstacle majeur dans notre processus de collecte de données, nous avons entrepris de résoudre ce problème en utilisant une approche différente. Nous avons découvert qu'en modifiant les en-têtes de nos requêtes HTTP pour inclure un "User-Agent" simulant un navigateur web standard, nous pourrions contourner les restrictions d'accès imposées par le serveur. Le "User-Agent" est un en-tête HTTP qui permet au serveur de reconnaître le type de navigateur web utilisé par l'utilisateur. En ajoutant un User-Agent dans nos requêtes, nous avons pu simuler le comportement d'un navigateur web conventionnel, ce qui a permis d'éviter les blocages et les erreurs de protocole 403. Dans notre cas, nous avons utilisé un User-Agent qui imitait les paramètres d'un navigateur Chrome. En plus de l'erreur 403, nous avons également pris en compte les différents "status code" HTTP qui pourraient être renvoyés par le serveur en vérifiant si la requête avait réussi grâce au "status code" 200. Cette approche nous a permis de surmonter efficacement les obstacles initiaux rencontrés lors de notre collecte de données, nous permettant ainsi de poursuivre notre étude de manière fluide et productive.

Face au défi des résolutions de captcha, nous avons envisagé deux approches potentielles pour résoudre ce problème. La première option consistait à recourir à des services de résolution de captcha payants, souvent basés sur des API spécialisées qui peuvent automatiser le processus de résolution. Cependant, cette solution aurait nécessité un investissement financier supplémentaire, ce qui n'était pas idéal pour notre projet. Par conséquent, nous avons opté pour une deuxième approche, qui consistait à utiliser des proxies pour masquer notre adresse IP et éviter la détection en tant que robot par le site. Pour obtenir une liste régulièrement mise à jour de proxies, nous avons utilisé le site https://free-proxy-list.net, qui met à jour ses listes de proxies toutes les 10 minutes. En utilisant ces proxies, nous avons pu diversifier nos adresses IP et ainsi contourner les mécanismes de détection de robots du site, permettant une collecte de données plus fluide et efficace. Cette approche nous a permis de surmonter avec succès le défi des résolutions de captcha sans avoir à recourir à des solutions payantes, démontrant ainsi notre capacité à trouver des solutions créatives et économiques aux obstacles rencontrés lors de notre étude.

VI. Conclusion

- Récapitulation des principales conclusions de l'étude

Ce dossier intitulé "Au-delà des Courts : Exploration des Données des Meilleurs Joueurs de l'Association of Tennis Professionals (ATP) et de la Women's Tennis Association (WTA)" a permis d'approfondir la compréhension des tendances et des caractéristiques qui définissent les carrières des joueurs et joueuses de tennis professionnels. Grâce à une méthodologie rigoureuse de web scraping en Python, des informations détaillées ont été recueillies, nettoyées et analysées, offrant un panorama des profils, des performances et des aspects financiers du tennis de haut niveau.

Les analyses révèlent une diversité de profils physiques et une prédominance de joueurs commençant leur carrière professionnelle dans des périodes précises, pointant vers une variabilité dans l'intérêt pour le tennis professionnel ainsi que dans les opportunités ou les barrières à l'entrée du circuit. De plus, l'étude a mis en lumière la corrélation entre la performance des joueurs et leur succès financier, avec des figures marquantes telles que Jannik Sinner qui se démarquent tant sur le court qu'en termes de gains.

Le projet a également abordé la complexité de la collecte de données dans des environnements web dynamiques, faisant preuve d'ingéniosité face aux défis techniques comme les erreurs de protocole et les mécanismes anti-robot. Les solutions adoptées, telles que l'usage de proxies et de User-Agents spécifiques, ont prouvé leur efficacité dans l'extraction de données précises et fiables, ouvrant la voie à des études futures qui pourraient explorer d'autres facettes du tennis professionnel ou appliquer des techniques similaires à différents domaines.

Les conclusions et insights tirés de ce travail soulignent l'importance de la data science dans le sport, permettant non seulement d'analyser la performance mais aussi de comprendre les impacts sociétaux et économiques sur le tennis. Les outils et techniques développés durant cette étude

constituent une base solide pour de futures recherches et pourraient être étendus pour inclure des données longitudinales ou des modèles prédictifs, enrichissant ainsi la compréhension du tennis professionnel et de son évolution.

- Perspectives et recommandations pour de futures recherches

Pour les recherches futures, il y a plusieurs avenues passionnantes à explorer pour enrichir et étendre les connaissances acquises dans cette étude sur l'ATP et la WTA :

1. Élargir la portée des données : Alors que cette étude se concentrait sur les 100 meilleurs joueurs et joueuses de chaque circuit, il serait bénéfique d'inclure des joueurs classés au-delà du top 100. Cela offrirait une image plus complète du paysage du tennis professionnel, y compris des insights sur les joueurs émergents et sur la profondeur de la compétition à différents niveaux.

2. Exploration de données historiques : La collecte de données sur plusieurs années permettrait d'analyser les tendances historiques et l'évolution du jeu, des prix en argent, et des carrières des joueurs. Cela pourrait également inclure l'étude de l'impact des changements de règles, des innovations technologiques et de l'économie globale sur le tennis.

3. Analyse comparative : Il serait intéressant de comparer les données de l'ATP et de la WTA avec celles d'autres sports professionnels pour comprendre les singularités du tennis en termes de diversité, de parité financière et de dynamiques de carrière.

4. Analyse démographique plus approfondie : Poursuivre l'exploration des données démographiques, telles que l'âge, la nationalité et les attributs physiques, en corrélation avec les performances et les blessures, pourrait fournir des informations précieuses sur la gestion de carrière et la prévention des blessures.

5. Suivi de la performance des joueurs : Développer des modèles pour suivre la progression des joueurs au fil du temps, en analysant les données de performance match par match, pourrait offrir des prédictions sur les trajectoires de carrière et les futurs champions.

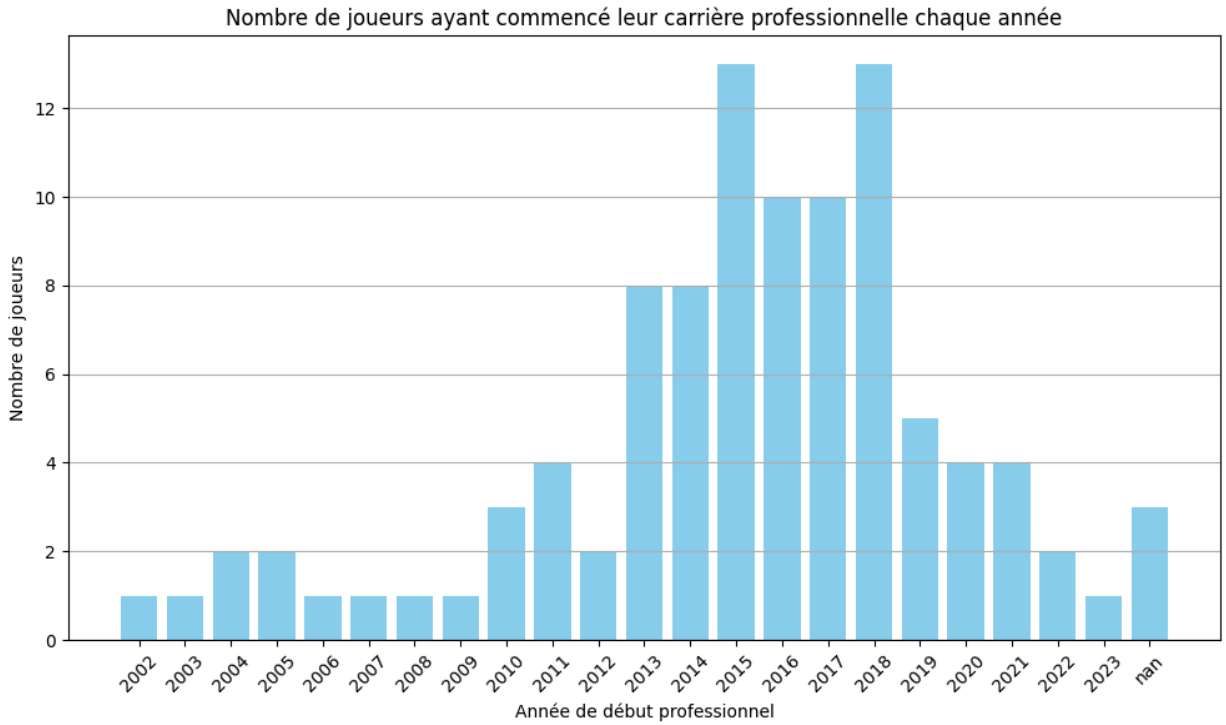
6. Impact socio-politique : Comme mis en évidence par la notation "-" pour les joueuses russes, les événements géopolitiques ont un impact sur le sport. Une étude plus approfondie de ces impacts pourrait offrir des perspectives sur la relation entre le sport et la politique mondiale.

7. Utilisation d'API et de techniques de scraping avancées : Les sites web modernes utilisent souvent des API pour charger les données, et explorer ces API pourrait rendre le scraping plus efficace. De plus, l'application de techniques avancées de scraping, y compris l'apprentissage automatique pour la classification des captchas, pourrait surmonter certains obstacles techniques rencontrés.

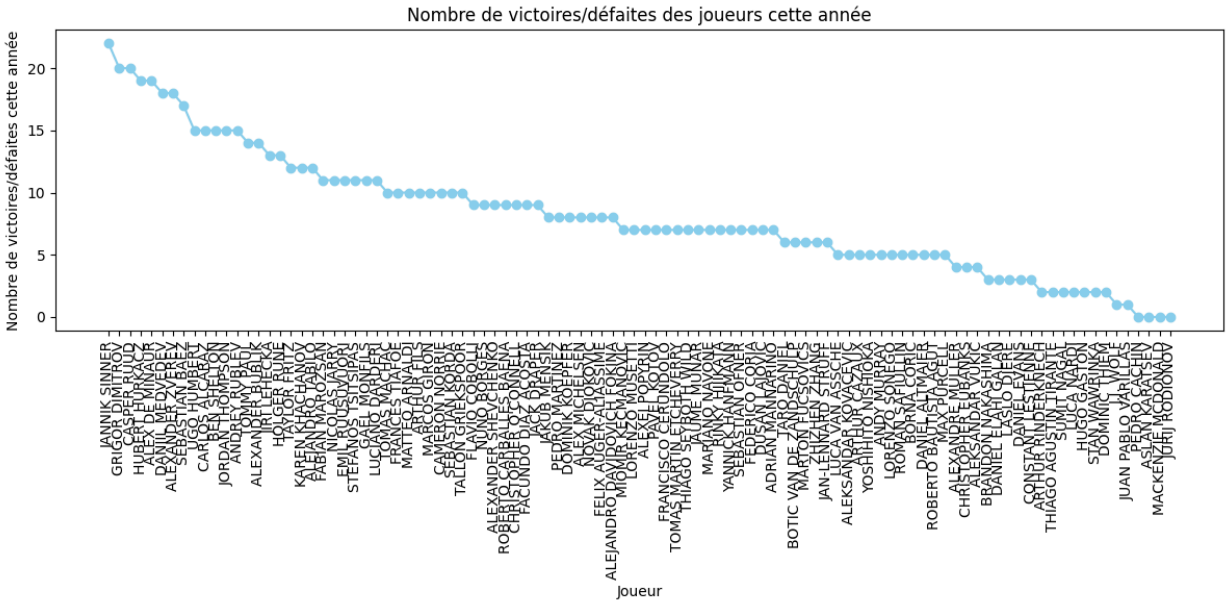
VII. Annexes

- Graphiques

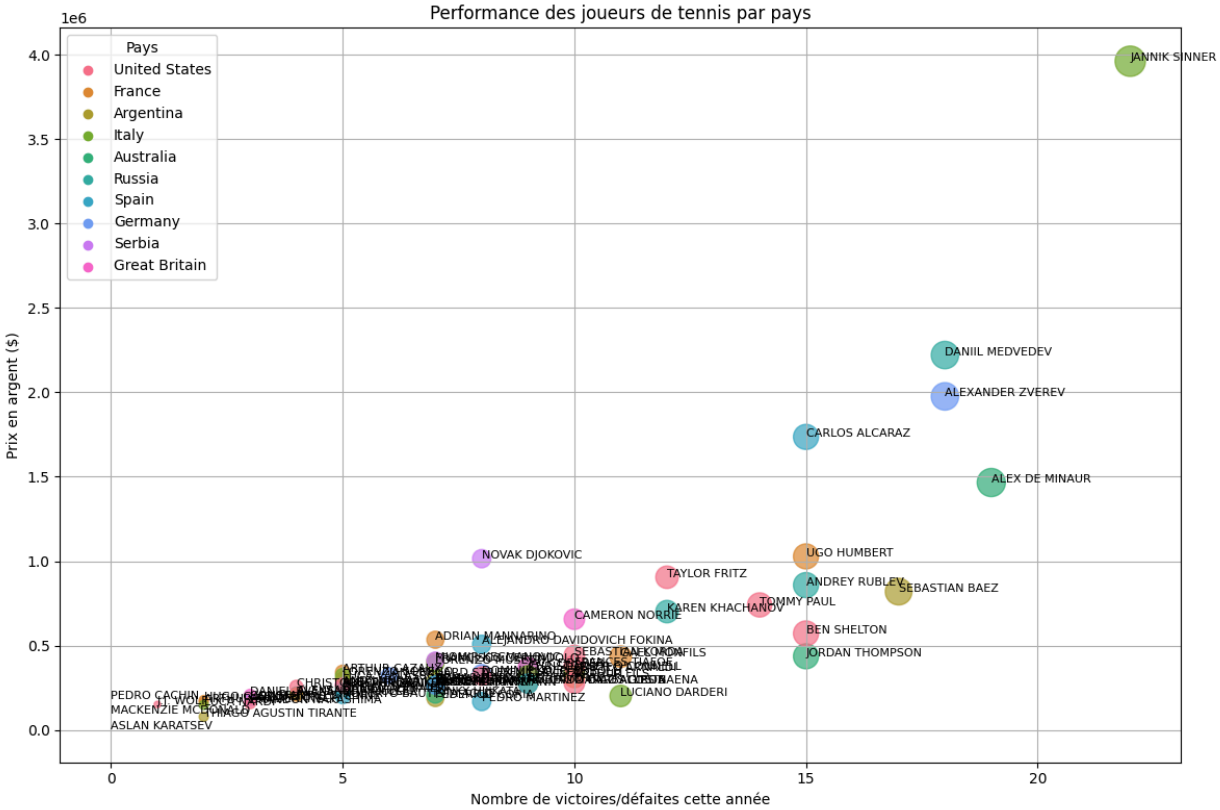
Graphique 1 : ATP Nombre de joueurs ayant commencé leur carrière professionnelle chaque année



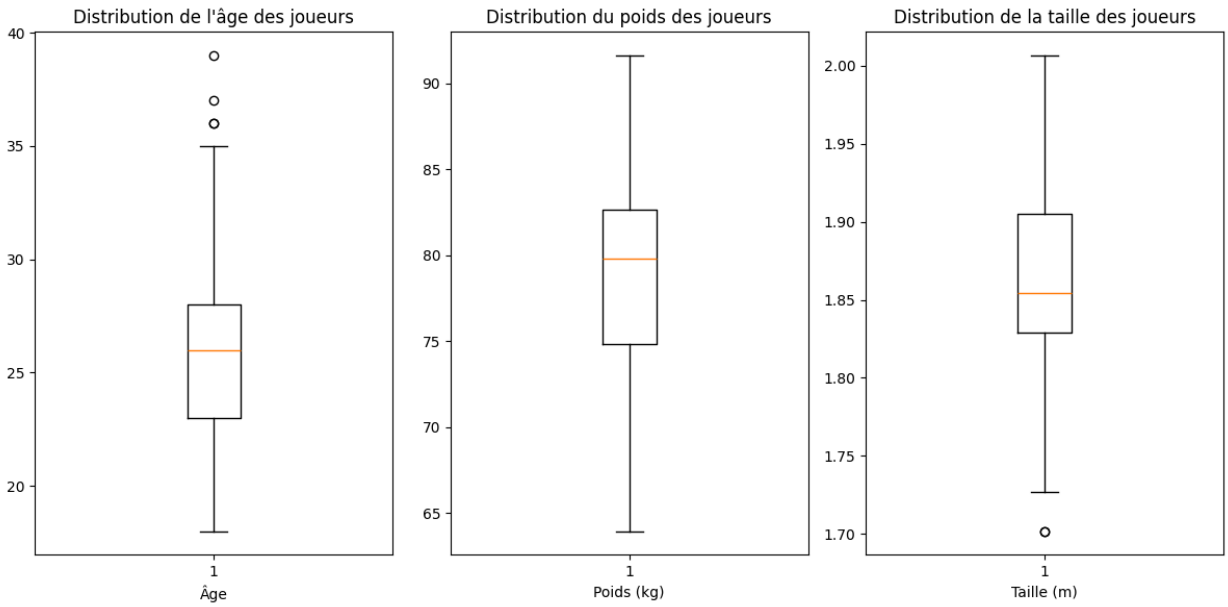
Graphique 2 : ATP Nombre de victoires / défaites des joueurs cette année



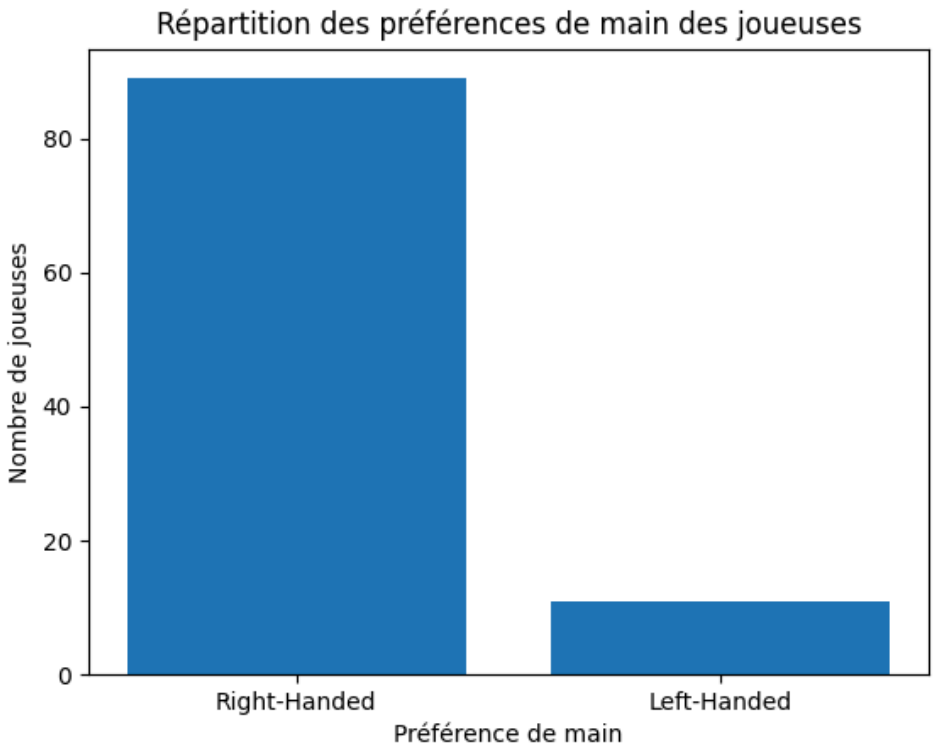
Graphique 3 : ATP Performance de tennis par pays



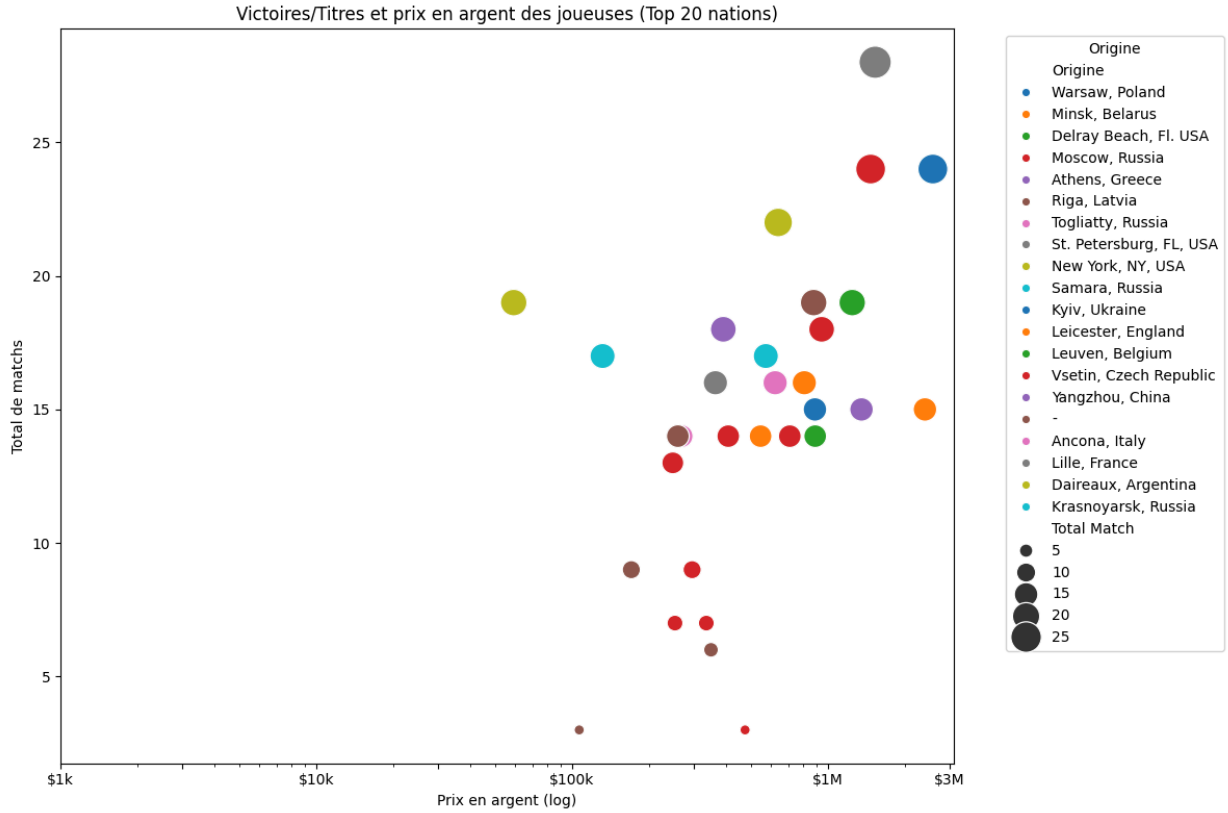
Graphique 4 : ATP Distribution Age, Poids et taille des joueurs



Graphique 5 : WTA Répartition des préférences de main des joueuses



Graphique 6 : WTA Répartition des nationalités en fonction des titres et du cash prize



Graphique 7 : Comparaison ATP / WTA Histogramme de l'âge moyen et répartition des nationalités dans le top 100