Analyse de données tennis 2018

ANALYSE PORTANT SUR LE TOP 50 ATP

Samir CHERGUI | STA101 | 3 Mars 2019

le cnam
Conservatoire national
des arts et métiers

Dans le cadre de ce projet, nous allons analyser un jeu de données que j'ai conçu moi-même via Microsoft Excel, ce jeu de données met en évidence certaines caractéristiques des cinquante meilleurs joueurs du monde de tennis en 2018.
Les données proviennent principalement du site de l'ATP : Atptour.com

C'est un jeu de données avec 50 individus pour 18 variables :

- Les individus représentant les joueurs pris individuellement classés par ordre alphabétique dans la première colonne.
- Les variables qui sont de nature différente, on aura 13 variables de nature quantitative et 5 variables de nature qualitative.

Certaines variables que l'on montrera auront été catégorisés afin d'améliorer la lisibilité de celles-ci, cela a été rendu possible par le fait que ces variables prennent des valeurs peu différentes quels que soient les individus.

L'objectif de cette étude sera de savoir si certaines variables ont un impact significatif sur le résultat final c'est-à-dire le classement en fin de saison.

Nous effectuerons une Analyse en Composante Principale (ACP) ou une Analyse Factorielle des Correspondances (AFC) pour essayer de résoudre ce problème à l'aide respectivement des variables quantitatives et qualitatives.

Aussi, nous chercherons à savoir si compte tenu des spécificités de chacun, il serait possible de regrouper les individus (les joueurs) dans des classes relativement homogènes via la méthode hiérarchique de Ward, nous analyserons enfin si le partionnement obtenu reflète la réalité observée.

L'ensemble du projet se déroulera avec l'aide du logiciel R et du package FactoMineR

Le plan qui sera utilisée dans le cadre de ce projet sera le suivant :

-L'analyse de quelques variables quantitatives et qualitatives indépendamment les unes des autres à l'aide d'histogrammes et de courbes selon les caractéristiques de chacune d'entre elles

-Une ACP ou AFC (selon la nature de la variable) pour déterminer quelles variables ont un impact significatif sur le jeu de données et quels sont les individus ayant le poids le plus important dans le jeu de données.

-Une typologie des individus pour pouvoir mettre en avant lesquels se ressemblent et par voie de conséquence lesquels diffèrent.

Première partie : Description des individus et des variables

On va dans un premier temps définir clairement les variables du jeu de données en précisant leur nature (quantitative ou qualitative)

Player Name correspond aux noms des individus qui seront pris en compte dans l'analyse de données, il s'agit tout simplement du nom des joueurs qui sont représentés par les différentes variables.

Age (Years) représente l'âge du joueur en années en 2018.
Variable quantitative

Professional_Experience(in_years) représente le nombre d'année d'expérience du joueur en tant que professionnel.
Variable quantitative.

End_year_ATP_points_earned(end_2018) représente le nombre de points ATP engrangés par chaque joueur au cours de l'année 2018 jusqu'à sa fin.
Ces points ATP sont gagnés en remportant des matches au cours de tournois estampillés ATP (Association of Tennis Player) ; Le nombre de points remportés varie en fonction du nombre de matchs gagnés au cours d'un même tournoi et en fonction de l'importance de celui-ci.
Variable quantitative.

Tournaments_won(in_2018) représente le nombre de tournoi estampillé ATP gagné par chaque joueur au cours de l'année 2018.
Variable quantitative.

Continent représente le continent d'origine de l'individu.
Variable qualitative.

Laterality représente l'information si le joueur est R-H ou L-H (Droitier ou Gaucher).
Variable qualitative.

End_Year_Earnings (million of dollars) représente le nombre de millions de dollars américains remportés au cours de la saison 2018.
Variable quantitative.

Percentage_of_matches_won représente le pourcentage de matchs remportés au cours de la saison 2018 par chaque joueur.
Variable qualitative.

Height (Cm) représente la taille de chaque joueur en centimètres.
Variable quantitative.

Weight (Kg) représente le poids de chaque joueur en kilogrammes.
Variable quantitative.

Number_Of_Aces représente le nombre d'aces réalisés par chaque joueur.
Variable quantitative.

Dans le tennis, un ace représente un point gagné par le serveur lors du premier coup de raquette sans que le receveur ait réussi à toucher la balle.

Double_Faults représente le nombre de double fautes réalisés par chaque joueur.
Une double-faute est lorsque le serveur fait deux erreurs consécutives au moment de servir.
Variable quantitative.

Total_Service_points_won(in_percentages) représente en pourcentages le nombre de points gagnés par le joueur au moment de servir.
Dans le tennis professionnel, il est considéré comme un avantage de servir par rapport à recevoir d'où le fait que ce taux sera systématiquement supérieur au pourcentage du nombre de points gagnés par le receveur.
Variable quantitative.

Total_Receive_points_won(in_percentages) représente en pourcentages le nombre de points remportés par le receveur lorsqu'il renvoie le service adversaire.
Dans le monde professionnel, il est considéré comme un désavantage de recevoir d'oû le fait que ce taux sera systématiquement inférieur au pourcentage du nombre de points gagnés par le serveur.
Variable quantitative.

Total_Points_won(in_percentages) représente le pourcentages le nombre de points gagnés par le joueur.
Il peut être obtenu en faisant une moyenne pondérée des deux variables précédentes.
La moyenne doit être pondéré par le nombre de points joués au service et au retour par chaque individu.
Variable quantitative.

Total_matches_played représente le nombre de matchs joué par chaque joueur.
Variable quantitative.

Tennis_Racket_Brand représente la marque de raquette utilisée par chaque joueur.
Variable qualitative.

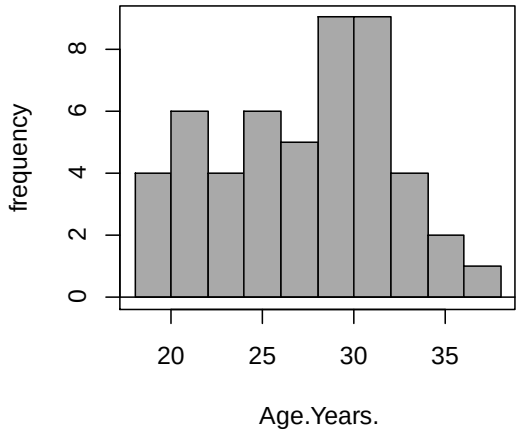
Outfit_brand représente la marque de vêtements du joueur.
Variable qualitative.

Deuxième partie : Analyse des variables dans les cadres univariées et bivariées

A/Analyse des variables univariées

Commençons par regarder l'âge des individus :

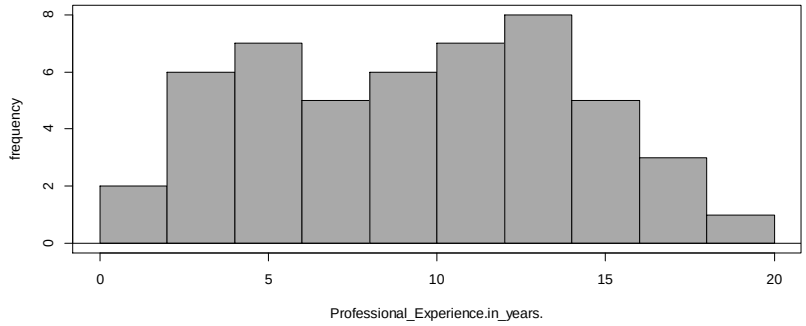
En voici son histogramme



On peut remarquer d'emblée que tous les joueurs du top 50 ont un âge compris entre 18 et 38 ans, cela démontre clairement qu'il y a une forte corrélation entre jeunesse c'est-à-dire fraicheur physique et performances de haut niveau.
Effectivement, après 32 ans, les joueurs sont de moins en moins nombreux à être performant.
Seuls trois joueurs du top 50 ont plus de 32 ans.
De l'autre côté, il est vraisemblablement très difficile d'être performant avant 18 ans car aucun joueur du top 50 n'a moins de 18 ans (Ici, une précision essentielle est à clarifier : Il n'y a pas d'âge minimum pour accéder au classement de tennis professionnel, la non-présence de joueurs de moins de 18 ans n'est pas due à une éventuelle règle qui imposerait ceux-ci d'avoir 18 ans.)
Enfin, on peut observer que 18 joueurs du top 50 ont un âge compris entre 28 et 32 ans ce qui nous amène à penser qu'il pourrait exister une sorte de juste milieu entre âge et expérience professionnelle, nous reviendrons là-dessus ultérieurement avec l'analyse de la matrice variance-covariance des variables quantitatives.
En utilisant l'onglet statistiques, résumé et statistiques descriptives, nous avons les données suivantes pour cette variable :
Moyenne=27,66 ; médiane = 28,5
Ces valeurs sont tout à fait cohérentes compte tenu de l'histogramme que l'on possède car 25 individus ont moins de 28 ans.
Enfin, le coefficient de variation empirique (égal à l'écart-type divisé par la moyenne) est de 0,17 ce qui montre une variabilité relativement élevée (supérieure à 0,15).

Analysons maintenant l'expérience professionnelle des individus :

Ici également nous utiliserons un histogramme :

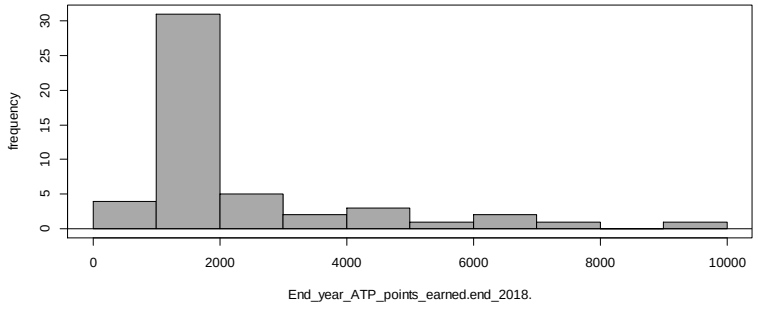


Ici, on observe que les joueurs du top 50 ont une expérience comprise entre 0 et 20 années.
On observe aussi une certaine hétérogénéité dans la mesure où les individus sont bien répartis sur l'histogramme et qu'il n'y a pas un nombre d'années d'expérience typique à acquérir pour faire partie du top 50.
En effet, 24 individus ont plus de 10 ans d'expériences, 3 individus ont 10 ans d'expérience et enfin 23 individus ont moins de 10 ans d'expérience.
D'ailleurs, le logiciel confirme bien l'analyse car la médiane du nombre d'années d'expérience professionnelle est bien de 10 années avec une moyenne de 9,84 années.
Cela nous amène à penser que cette variable a peu d'impact dans le classement de tennis dans le cadre des 50 meilleurs joueurs du monde, nous vérifierons cette hypothèse lors de l'ACP.
Par ailleurs, on a un coefficient de variation empirique égal à 48% ce qui montre une forte variabilité entre les individus et confirme leur forte dispersion.

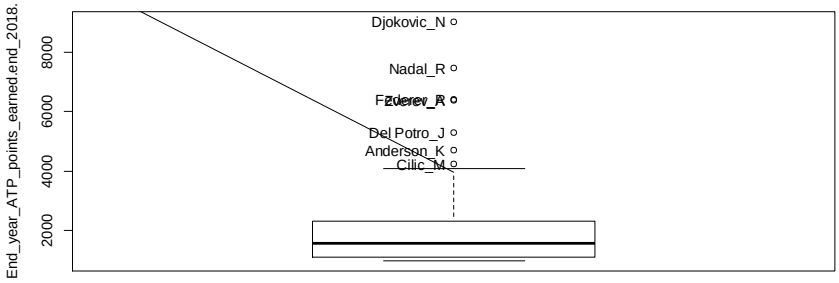
Analyse du nombre de points ATP pour l'ensemble des individus.

A titre d'information, le nombre de points ATP détermine le classement de l'individu, plus le nombre de points est élevé, mieux l'individu sera classé.
Dans le cadre de l'analyse de cette variable, on peut essayer d'utiliser un histogramme

Voici l'histogramme :



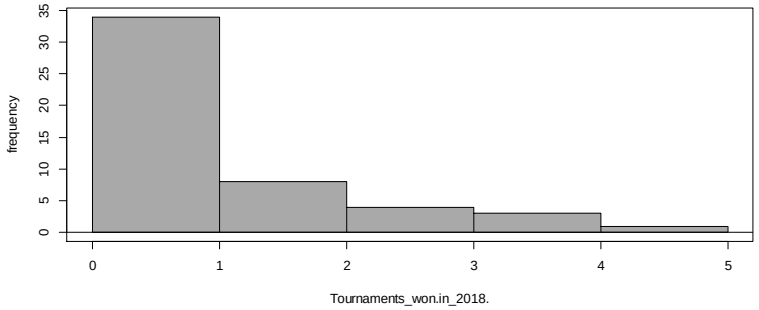
Ici, on constate que l'histogramme n'est pas du tout approprié car on observe une forte homogénéité des individus.
En effet, 30 individus soit 60% des joueurs ont un nombre de points compris entre 1000 et 2000.
Pour appréhender de façon pertinente cette variable, nous utiliserons donc une boite de dispersion.
La voici :



Grâce à ce graphique, on peut constater d'importantes inégalités entre les joueurs notamment entre les 7 meilleurs classés et le reste.
Ici, on a une moyenne de points ATP de 2256 entre les différents joueurs or on a un écart-type égal à 1847,8 ce qui montre une forte variabilité entre les individus
(Il est indispensable de comparer l'écart-type avec la moyenne pour essayer d'en interpréter une variabilité forte ou faible). On peut utiliser le coefficient de variation empirique qui est égal à l'écart-type divisé par la moyenne, ici il est égal à 0,82 ce qui met en évidence une très forte variabilité.
La boite de dispersion montre bien que certaines valeurs sont trois voire plus de quatre fois supérieure à la moyenne (le maximum étant égal à 9045).

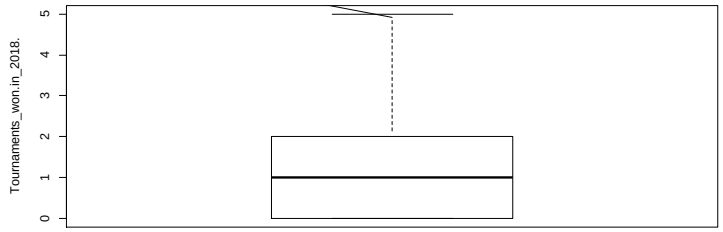
Analyse du nombre de tournois remportés

L'histogramme ci-contre montre le nombre de tournois remportés associés aux effectifs :



On voit que 68% (34 individus) ont gagnés un tournoi ou moins au cours de l'année 2018 et que le reste 32% (16 individus) en ont gagnés 2 ou plus (jusqu'à cinq) ce qui démontre

là aussi de fortes inégalités, inégalités qui se confirment lorsque l'on observe un coefficient de variation supérieur à 1.
La boite de dispersion confirme cette tendance :



Grâce à celle-ci, nous pouvons constater que 75% des joueurs ont gagnés deux tournois au plus et que 25% de ceux-ci n'ont gagnés aucun tournoi ce qui ne fait que démontrer davantage la forte variabilité entre les joueurs.
Enfin, sur les 59 tournois remportés par le top 50, 27 ont été remportés par le top 10(soit près de 46% des tournois remportés par 20% des individus) et 19 par le top 5 (soit 32% des tournois gagnés par 10% des effectifs), on s'aperçoit donc qu'à l'instar des points ATP gagnés, les échantillons sont de plus en plus volatiles au fur et à mesure que l'on prend en compte des joueurs de mieux en mieux classés).

Analyse de la variable continent

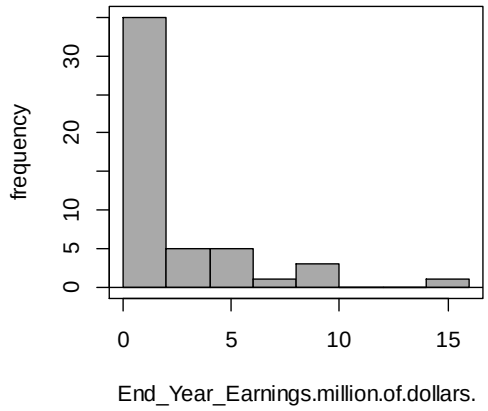
Cette variable étant qualitative, on se contentera d'énumérer les fréquences
Africa = 4%
Asia = 8%
Europe = 62%
North-America = 12%
Oceania = 8%
South-America = 6%

Analyse de la variable Laterality

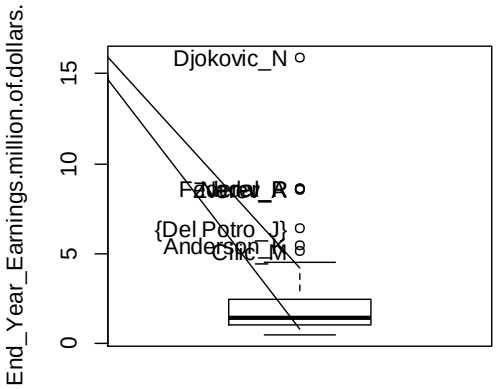
Seuls 10% des joueurs sont gauchers les autres (90%) sont droitiers

Analyse de la variable End_Year_Earnings (million of dollars)

Essayons d'analyser cette variable à partir de l'histogramme :



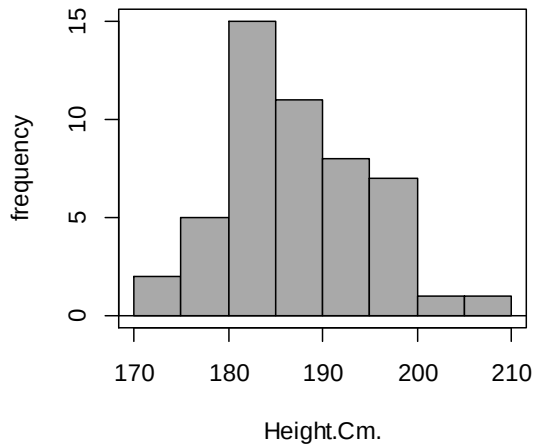
A l'instar du nombre de points ATP, on constate que les individus sont homogènes dans la mesure où 70% d'entre eux ont obtenu un revenu relativement similaire en 2018 soit entre 0 et 2 millions de dollars, seuls 30% d'entre eux ont un revenu supérieur à 2 millions de dollars.
Regardons la boite de dispersion pour plus de détails :



On voit en effet que les inégalités sont criantes dans la mesure où l'on a une moyenne de gains égale à 2,55 millions de dollars remportés mais que la médiane est de 1,465 millions, le coefficient de variation étant de 1,385, les revenus entre joueurs sont très volatiles. Moins de 25% des joueurs ont eu un revenu supérieur à la moyenne et ces inégalités explosent lorsque l'on compare les joueurs les mieux classés :
Les 2è, 3è et 4è remporte 8 millions chacun quand le mieux classé remporte 15,9 millions soit près du double de ses poursuivants directs ou soit plus de 30 fois le revenu du moins bien classé.

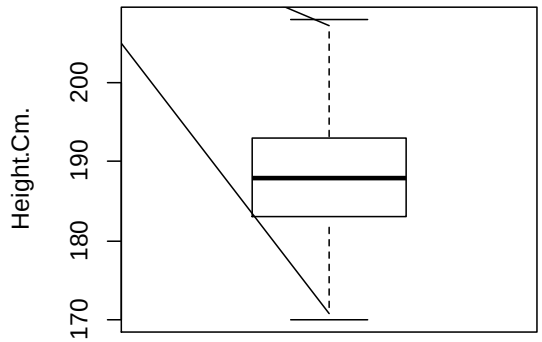
Analyse de la variable Height(cm)

Intéressons-nous à la taille des individus, commençons par voir si l'histogramme est pertinent :



Grace à l'histogramme, nous pouvons voir que la majeure partie des individus ont une taille comprise entre 180 et 200cm, seuls 9 individus soit 18% des effectifs ont une taille « extrême » c'est-à-dire soit moins de 180 cm soit plus de 200cm.
Cela peut nous amener à penser que la taille d'un joueur peut avoir un impact significatif sur ses performances car sinon l'histogramme nous montrerait des fréquences bien plus régulière quelle que soit la taille. La moyenne de 187,98 cm est comparable à la médiane d e188 cm ce qui montre une faible variabilité, faible variabilité confirmée par un coefficient de variation empirique très faible (3,8%).

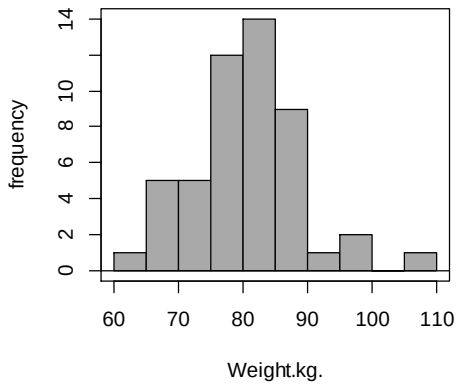
La boite de dispersion montre clairement que les individus sont plutôt peu dispersés :



Analyse de la variable Weight (Kg)

Les individus ont un poids moyen de 81,72 kilogrammes, le plus léger à un poids de 64 kilogrammes quand le plus lourd a un poids de 108 kilogrammes.
La médiane étant de 81,5 kilogrammes, elle est à peu près égale à la moyenne.

Cela peut nous amener à penser que le poids des individus varie faiblement d'autant que le coefficient de variation est de 10%.
Analysons cela de plus près à l'aide d'un histogramme :

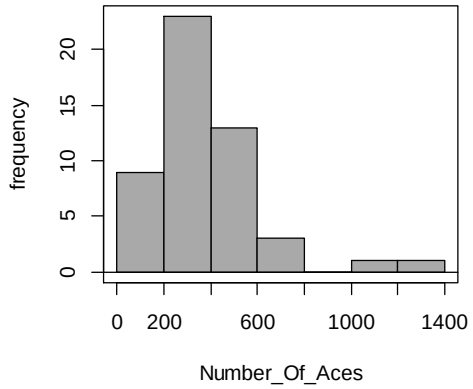


Ici, on voit que les valeurs extrêmes sont peu représentés dans la mesure où 45 individus soit 90% des effectifs ont un poids compris entre 65 et 90 kilogrammes.

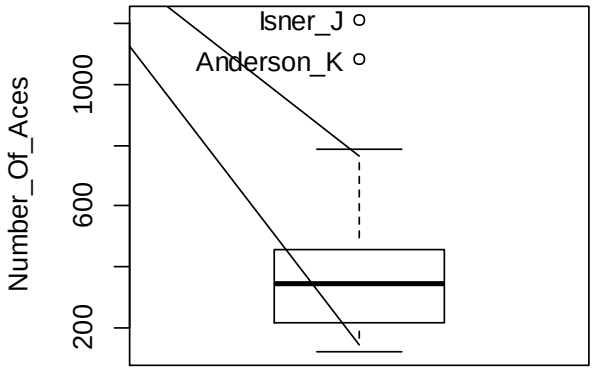
Analyse de la variable Number_Of_Aces

Le nombre de Aces réalisés par chaque individu est en moyenne de 379 pour l'année 2018.
La médiane étant de 346 aces, on peut se mettre à penser que la variable étudiée à une variabilité relativement faible mais le coefficient de variation étant de 58,9% montre que ce n'est pas le cas.
Généralement, l'intuition nous dit que dans ces cas de situation, il existe de nombreuses « valeurs extrêmes » voire aberrante.

Vérifions l'histogramme pour en savoir davantage :

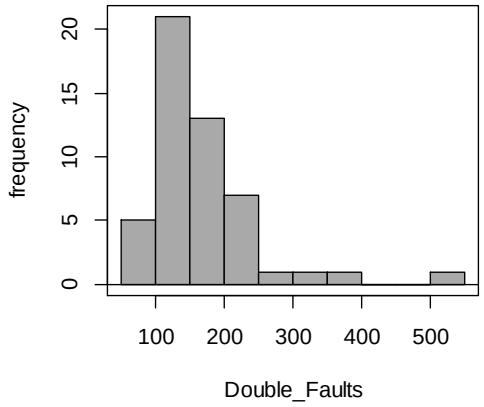


L'histogramme indique bien que même si la plupart des joueurs ont un nombre d'ace réalisés relativement homogène, il existe un faible nombre d'individus qui ont un nombre d'aces bien plus important que la normale, cela peut être mieux représenté par une boite de dispersion.



On voit bien grâce à la boite qu'il existe une forte variabilité avec deux individus ayant fait beaucoup d'aces. Cependant, lorsque l'on regarde de plus près, on voit qu'il y a un plus grand nombre d'individus qui ont une réalisés une faible quantité d'aces en 2018 soit environ 9 joueurs.
Cette pondération permet d'avoir une moyenne relativement stable par rapport à la médiane malgré la forte variabilité de la variable.

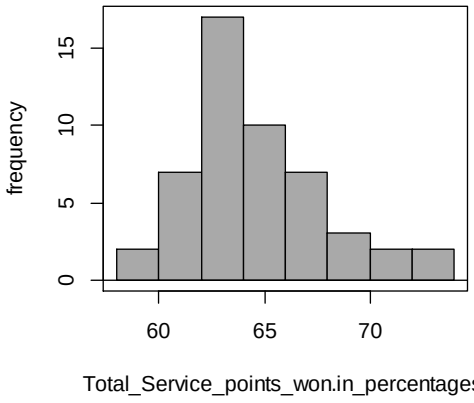
Analyse de la variable Double_Faults



Selon l'histogramme ci-dessus, la majeure partie des joueurs ont fait entre 64 et 250 doubles fautes pour l'année 2018 ce qui explique que la moyenne est relativement proche de la médiane.
En effet, s'il existe des valeurs très élevés, elles sont peu nombreuses et donc ont un poids moins important.
Le nombre de double fautes réalisés par chaque individu est en moyenne de 167 pour l'année 2018 pour une médiane égale à 150.
Le coefficient de variation est de 49 ,6% ce qui montre une forte variabilité entre les joueurs.

Analyse de la variable Total_Service_points_won(in_percentages)

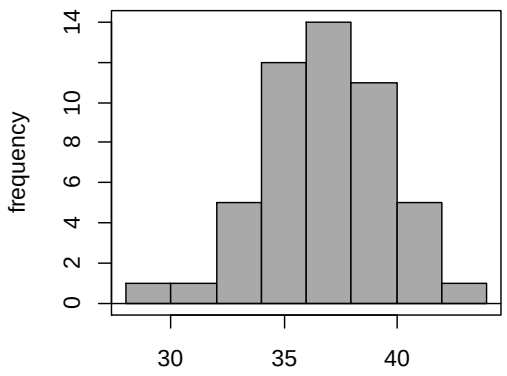
Ici, nous allons comparer l'efficacité au service des individus,
Voici l'historique de cette variable afin de pouvoir appréhender les différences entre les joueurs :



On voit via l'histogramme que tout d'abord l'ensemble des joueurs remportent plus de 59% des points au service ce qui montre que systématiquement, le joueur est favorisé lorsqu'il doit servir au lieu de recevoir.
Aussi, la moyenne des individus étant de 65%, elle est relativement proche de la médiane étant elle de 64% de points gagnés au service.
Relativement seulement car il est à souligner ici que l'écart-type ici est très faible (environ 3,25%). Le coefficient de variation est de 5% ce qui démontre une faible variabilité malgré des individus ayant des valeurs élevées.

Analyse de la variable Total_Receive_Points_won (in_percentages)

A la différence de la variable précédente, on voit qu'ici tous les joueurs ont un pourcentage de points gagnés au retour inférieur à 44% ce qui démontre que recevoir est un handicap pour le joueur, voici l'histogramme représentant la variable :

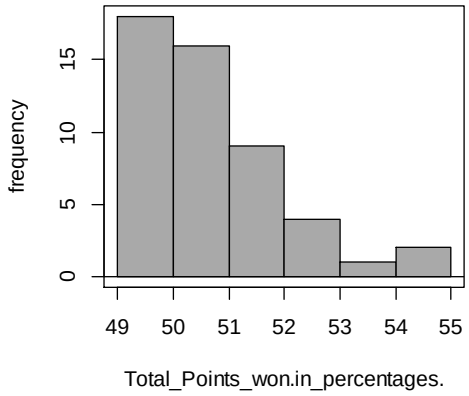


Ici, on peut constater que la plupart des joueurs ont un taux de réussite à la relance compris entre 32 et 40% ceci est en concordance avec la moyenne et la médiane de la variable respectivement égales à 37,22% et 37%.
Enfin le coefficient de variation égal à 7,5% confirme la faible variabilité des individus malgré une étendue importante égale à 15 points.

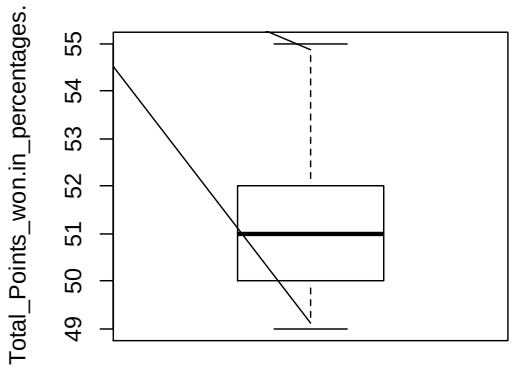
Analyse de la variable Total_Points_won(in_percentages)

A première vue on aurait pu penser que cette variable n'est que le résultat de la moyenne entre le taux de points gagnés au service et à la relance (seuls ces deux options sont disponibles pour un joueur de tennis).
Ce n'est pas vraiment le cas dans la mesure où une pondération est indispensable pour avoir des valeurs reflétant la réalité or il ne m'a pas été possible d'obtenir le nombre de points joué au service ou à la relance pour chaque individu (obtenir ces informations n'aurait pas été impossible mais cela aurait pris des dizaines voire une centaine d'heure, il aurait fallu éplucher chaque match de chacun des 50 individus puis faire une compilation des résultats).

Finalement, observons l'histogramme lié à cette variable :



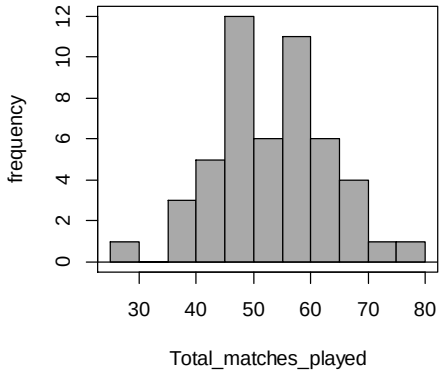
Déjà, on peut s'apercevoir qu'à la différence des deux précédentes variables nous n'avons pas de valeurs extrêmes de part et d'autre mais seulement avec des valeurs « élevés ».
En effet, le taux de points gagnés par les joueurs est assez homogène car la différence entre le troisième quartile et le premier quartile est égal à la différence entre le dernier quartile et le troisième quartile.
Aussi, la médiane est égale à la moyenne (51%) et le coefficient de variation est très faible (égal à 2,8%).
La boite de dispersion peut nous aider à nous rendre compte des inégalités liées au taux de points gagnés :



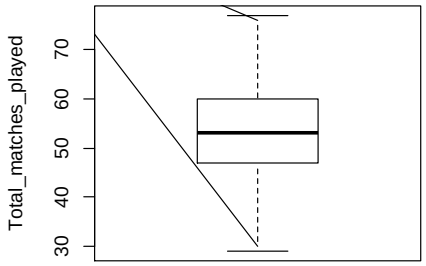
On voit qu'effectivement les individus sont concentrés vers le bas ce qui démontre une concentration des individus vers les variables les plus faibles.
On a donc quelques joueurs qui ont des valeurs très élevés en taux de points gagnés.
On observe bien que certains ont remportés 53 voire 55% de leurs points joués.
Ce sont des valeurs très élevés car on voit que l'écart-type n'est que de 1,44 points et donc la valeur maximale est par exemple trois fois plus éloignée que l'écart-type, on peut donc la considérer comme étant élevé.

Analyse de la variable Total_matchs_Played

Regardons l'histogramme pour appréhender cette variable :



On voit que la grande majorité de joueurs ont joué entre 40 et 70 matches au cours de l'année et plus particulièrement on a 29 joueurs qui ont joué entre 45 et 60 matches pendant cette saison 2018.
Aussi, on constate une variabilité importante avec quelques valeurs extrêmes répertoriés : un joueur a joué 29 matches quand un autre en a joué 77 soit presque trois fois plus.
La médiane (53) est comparable à la moyenne (53,58) cependant on a un écart-type élevé par rapport à la moyenne (égal à 9,81) ce qui met en évidence un coefficient de corrélation élevé supérieur à 15% (18,3%).
Essayons de mieux observer les différences grâce à la boite de dispersion :



On voit ici que la boite de dispersion est bien étendue ce qui confirme une variabilité élevée.

Analyse de la variable Tennis_Racket_Brand

On a pour cette variable qualitative les informations suivantes :

5 joueurs utilisent une Babolat (10%)
1 joueur utilise une Dunlop (2%)
12 joueurs utilisent une Head (24%)
3 joueurs utilisent une Prince (6%)
5 joueurs utilisent une Technifibre (10%)
19 joueurs utilisent une Wilson (38%)
5 joueurs utilisent une Yonex (10%)

Analyse de la variable Outfit_Brand

On a pour cette variable qualitative représentant la marque des vêtements des joueurs les informations suivantes :

Adidas : 8 joueurs
Asics : 4 joueurs
Diadora : 1 joueur
Fila : 6 joueurs
Hydrogen : 2 joueurs
Joma : 1 joueur
Lacoste : 4 joueurs
Le Coq Sportif : 1 joueur
Lotto : 6 joueurs
New Balance : 1 joueur
Nike : 12 joueurs
Sergio Tacchini : 2 joueurs
Uniqlo :2 joueurs

B/ Analyse des variables bivariées

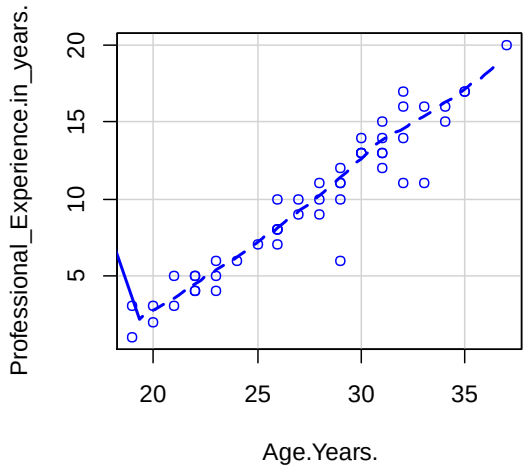
Pour analyser les rapports entre variables deux à deux on commence par regarder la matrice de variance covariance.
Il est à noter que seules les variables quantitatives seront prises en compte dans cette analyse.

Dans le cadre de cette analyse, nous utiliserons là encore le package FactoMineR du logiciel R. Après avoir chargé le jeu de données, nous allons dans l'onglet statistiques, on choisit résumé puis matrice des corrélations en sélectionnant toutes les variables quantitatives.
On prendra en compte le coefficient de Pearson pour analyser les corrélations entre les variables.

On prendra en compte les covariances supérieures à 0,70 en valeur absolue uniquement. La variance est systématiquement égale à 1 ce qui est logique car il existe un lien parfait entre deux variables identiques.

Pour l'âge des joueurs, on observe une très forte corrélation (+0,95%) avec la variable expérience professionnelle en années ce qui démontre que plus un joueur est âgé et plus son expérience l'est aussi.

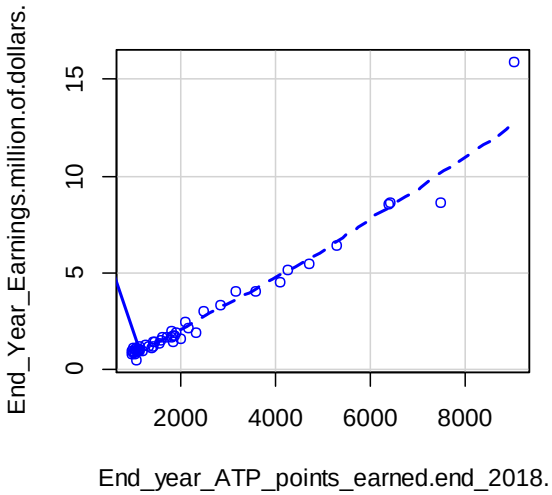
Voici un nuage de points correspondant à la corrélation de ces deux variables :



Le graphique ci-dessus montre clairement une forte tendance linéaire positive qui pourrait facilement être représentée par une droite.

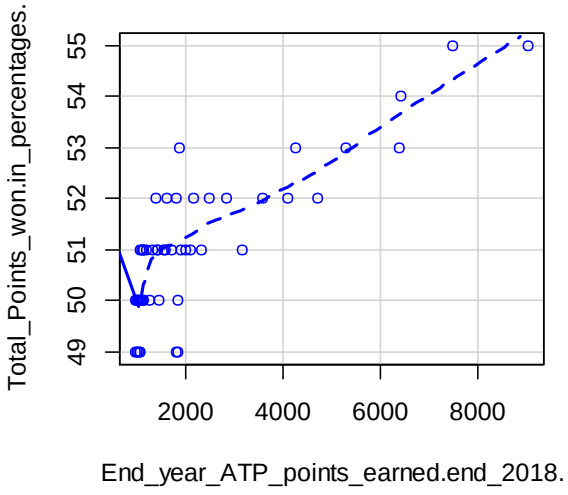
On n'observe pas de corrélation significative entre la variable Double_Faults et les autres car elles sont toutes inférieures à 0.3 en valeur absolue.

La variable End_year_ATP_points_earned(end_2018) est très corrélée de façon positive (+0,977) à la variable End_Year_Earnings (million.of.dollars) ce qui montre que mieux le joueur est classé et plus ses revenus le sont également et inversement plus ses revenus sont importants et plus sont classement est élevé.
En voici la représentation graphique :



On voit ici également qu'il serait possible de représenter le nuage de points à l'aide d'une droite même si on voit que l'écart-type augmente au fur et à mesure que les valeurs augmentent.

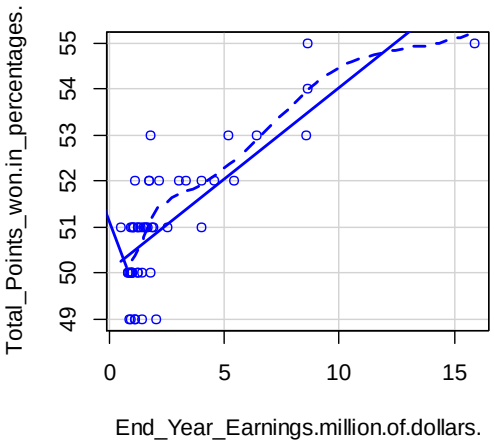
Cette même variable est également corrélée au taux de points gagnés (+0,823) et au nombre de tournois gagnés (+0,759) cela confirme le fait que plus le taux de points remportés ou le nombre de tournois remportés est élevé et plus le nombre de points ATP le sera aussi.

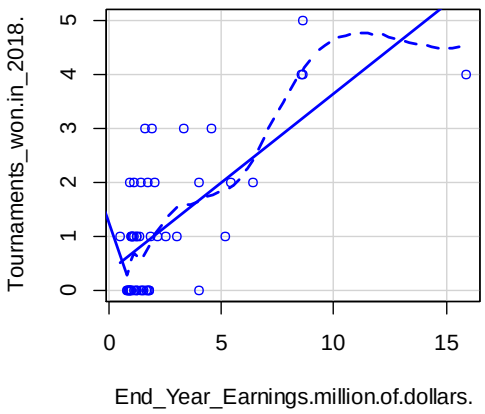


Sur ce graphique, on peut constater qu'il est tout à fait possible de représenter le nuage de points à l'aide d'une droite néanmoins, on voit que par rapport aux graphiques

précédents qu'un coefficient de corrélation relativement faible apporte une description moins précise des variables dans la mesure où l'on a une somme des écarts-type plus élevé.

Pour la variable des gains monétaires annuels, on constate qu'elle est corrélée positivement au taux de points remportés en pourcentages et au nombre de tournois gagnés (+0,78 et +0,70).

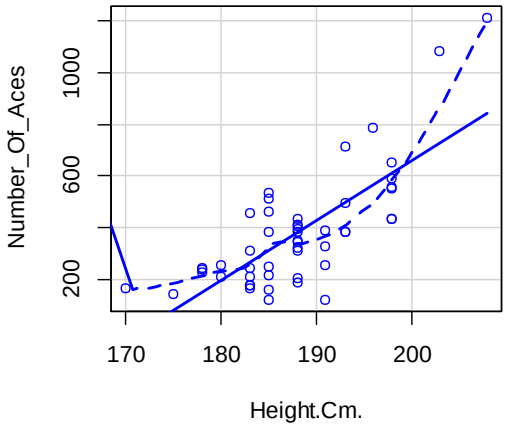




On peut voir que ces deux graphiques nous montrent une tendance claire dans la mesure où à chaque fois les deux variables croient ensemble positivement. Cependant on voit qu'il est difficile de représenter l'ensemble des points grâce à une droite ; d'ailleurs, la droite des moindres carrés met en évidence un écart-type beaucoup plus important que lorsque l'on a comparé l'âge et l'expérience des joueurs. Cela est du au fait que les coefficients de corrélation ici sont plus faible qu'auparavant.

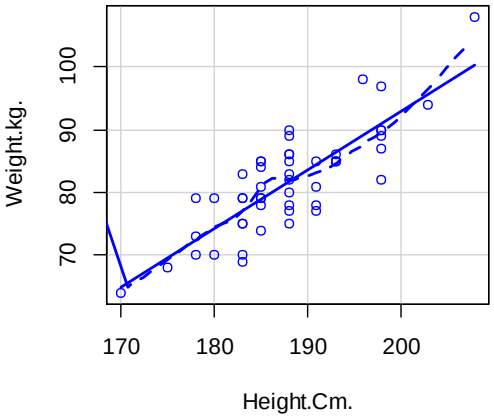
Analysons maintenant la variable taille en cm :

Grâce à la matrice des variances-covariances, on constate que cette variable est corrélée significativement avec la variable nombre d'aces (+0,76) et la variable poids(+0,83), cela nous dit que plus un joueur est grand et plus le nombre d'aces qu'il effectue est important :



Le graphique ci-dessus démontre bien la corrélation positive entre ces deux variables malgré une forte variabilité due aux individus extrêmes.

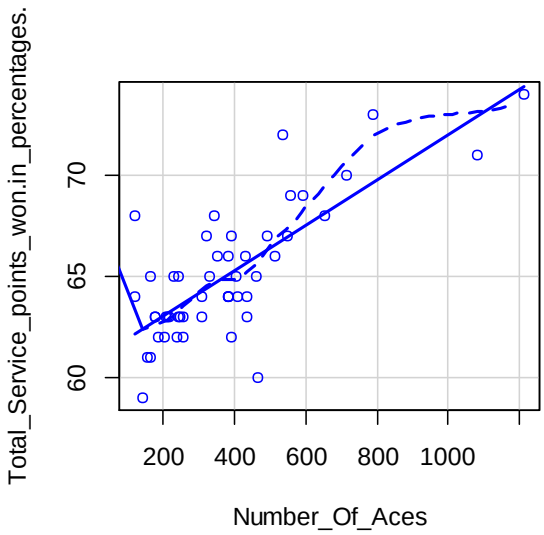
On voit aussi que généralement, plus un individu est grand et plus son poids l'est aussi :



Ici, on voit que ces deux variables sont nettement mieux corrélées car en effet les écarts-type sont plus faibles et le coefficient de corrélation est plus important.

Intéressons-nous maintenant à la variable du nombre d'aces :

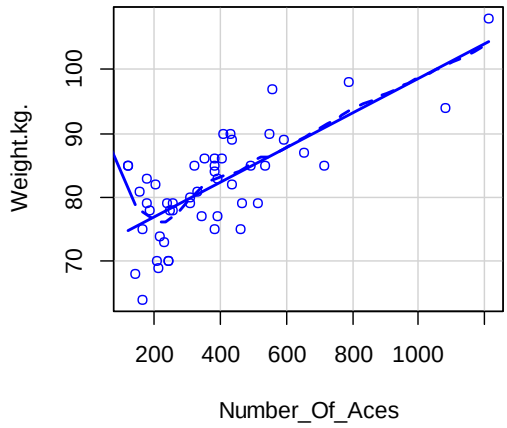
Logiquement, on observe une corrélation positive entre le nombre d'aces effectués et le nombre total de points gagnés au service, cela est logique dans la mesure où un point gagné au service permet d'augmenter directement son taux de points joués au service.
Le coefficient de corrélation pour ces deux valeurs est égal à 0,77.



On voit grâce à ce graphique que la variabilité est plus élevée pour les valeurs fortes.

Une dernière remarque pouvant être faite dans le cadre de l'étude de l'analyse bivariée est le fait que les coefficients de variation négative sont existants mais faibles en valeur absolue inférieur à (-0,60 en valeur absolue), par conséquent elles n'ont pas été décrites dans le cadre de cet exposé

Aussi, on voit que le nombre d'aces est corrélé au poids du joueur : on a vu auparavant que le nombre d'aces est corrélé positivement avec la taille du joueur et que la taille du joueur est également corrélée positivement avec son poids.
Finalement, on peut observer une sorte de transitivité dans la mesure où lorsque deux variables sont très corrélées, et que l'une de ces deux variables est corrélée avec une autre variable notée X alors la deuxième variable est corrélée avec X.



Le graphique ci-dessus montre bien la corrélation positive entre les variables malgré la forte variabilité.
Le reste de la matrice de variance-covariance ne montre pas de corrélation supérieure à 0,7.
On peut néanmoins observer une corrélation de 0,63 entre le taux de points gagnés et le nombre de tournois gagnés ainsi que qu'un même coefficient de corrélation entre le taux de points gagnés et le taux de points gagnés au service.

On voit également que le taux de de points gagnés au service (+0,63) est supérieur au taux de points gagnés à la relance (+0,48) lorsque l'on analyse leur corrélation avec le nombre de points gagnés)
L'intuition nous dit que le service serait un facteur plus « décisif » que la relance, nous verrons cela plus précisément lors de l'analyse multivariée.

Troisième partie : Analyse des variables dans le cadre d'une analyse multivariée

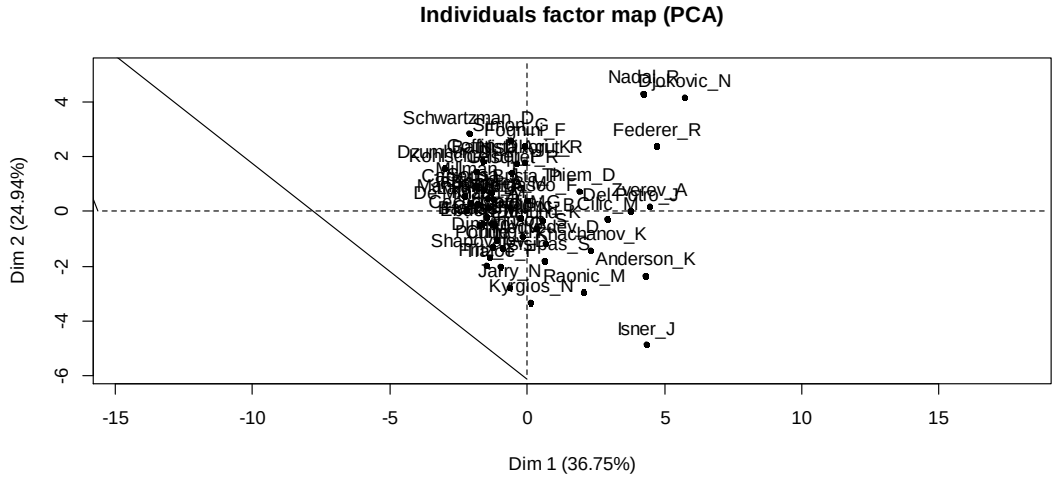
Dans le cadre de l'analyse multivariée nous allons effectuer une analyse en composante principale pour essayer de résumer l'ensemble de l'information contenue dans le jeu de donnée, cela nous permettra de savoir notamment quand est ce que les joueurs se ressemblent du point de vue des variables et de créer éventuellement ultérieurement des groupes d'individus homogènes compte tenu de l'ensemble des valeurs du jeu de données.

A/Analyse en composante principale

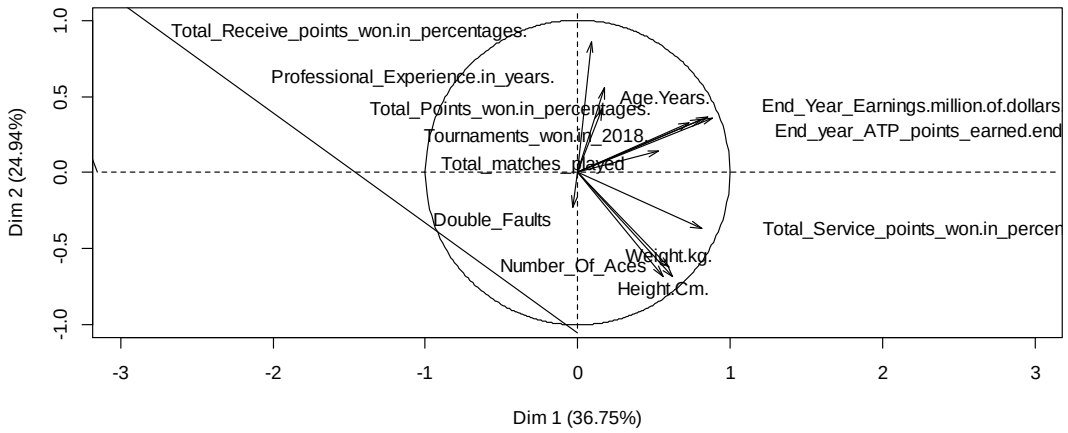
Pour faire l'ACP, nous utiliserons le package FactoMineR, après avoir chargé le jeu de données, nous cliquerons sur FactoMineR et ACP.

Nous utiliserons plusieurs graphiques dans le cadre de cette étude, le graphe des individus par exemple sera dupliqué en changeant l'échelle afin de mieux observer les différences entre eux.

En effet, regardons le fichier suivant :



Ce graphe n'est pas dénué d'intérêts dans la mesure où il permet de voir clairement les individus qui se détache et les individus qui se ressemblent cependant il est impossible de voir plus précisément les individus qui se ressemblent.
Nous commencerons donc par décrire les axes à l'aide du graphe des cercles des corrélations que voici :



Il est relativement peu lisible du fait des nombres variables quantitatives qu'il contient, cela s'explique par la colinéarité de certaines variables c'est-à-dire un coefficient de corrélation élevé entre deux variables projetées.

Par exemple, on voit que les variables End_Year_Earnings et End_Year_ATP_points_earned se confondent ce qui s'explique par le fort coefficient de corrélation que l'on a pu constater entre ces deux variables.
Le même raisonnement peut avoir lieu avec les variables Height, Weight et Number_Of_Aces.
Aussi, on peut constater que les variables susmentionnées sont relativement bien projetées dans le cercle des corrélations dans la mesure où elles sont près du bord du cercle.

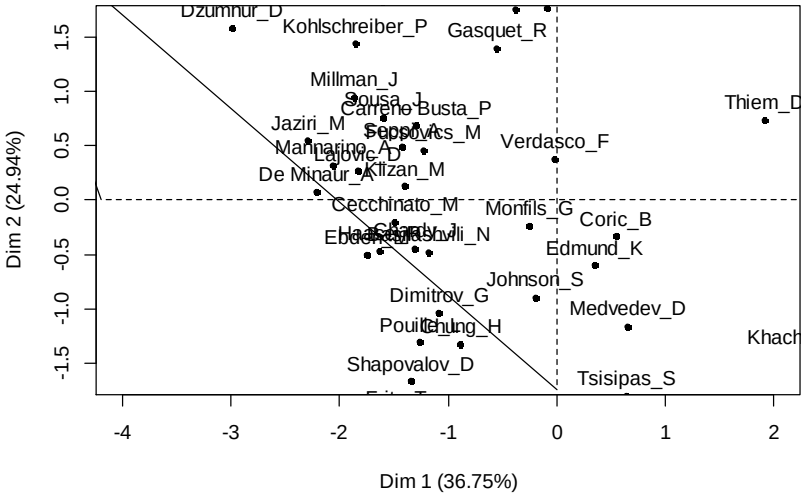
Dans le cadre de cette étude, nous nous limiterons à l'étude de trois axes car même si le premier plan nous donne 61,9% de la totalité de l'information du jeu de données (le premier plan représentant l'addition des deux premiers axes orthogonaux), le supplément d'information offert par le troisième axe est à notre sens significatif car il nous apporte 15,2% d'informations supplémentaires. D'ailleurs, le logiciel va nous inciter à sélectionner les trois premiers axes en nous donnant pas défaut les cosinus carrées et les contributions des trois premiers axes.

Essayons maintenant de mettre en évidence la construction des axes,
Commençons par l'axe 1 :

Grace au cercle des corrélations on peut voir quels sont les variables qui ont permis de construire l'axe 1, on voit que les variables End_Year_Earnings, End_Year_ATP_points_earned, Total_Service_Points_Won(in_percentages) et Tournaments_won ont de fortes valeurs sur l'axe 1.
Cela veut dire que les individus ayant de fortes valeurs sur l'axe 1 seront ceux ayant beaucoup de points ATP, ceux qui auront gagnés beaucoup d'argent, ceux ayant gagnés un grand nombre de tournois et/ou ceux qui auront un taux de service gagnant élevé.
Cela veut dire aussi que ceux qui auront de faibles valeurs en nombre de points ATP, argent gagné ou taux de points gagné au service se situeront à gauche de l'axe 1.

Si l'on regarde le graphique précédent, on voit que certains joueurs tels que Federer, Djokovic, Nadal, Zverev, Anderson, Del Potro, Cilic ou Isner ont de fortes valeurs sur l'axe 1.
Cela met en évidence le fait que ces individus se démarquent nettement des autres sur cet axe dans la mesure où leurs performances sur les variables composant l'axe 1 sont élevés.
Exemple : Djokovic est le joueur le mieux classé (plus grand nombre de points ATP), il est le joueur le mieux rémunéré en 2018, a un taux de points gagnés au service parmi les 50% meilleurs et a gagné un grand nombre de tournois.
L'agrégation de ces critères font de lui l'individu ayant la coordonnée la plus forte dans l'axe 1.
Federer, Nadal, Isner, Zverev et Anderson ont aussi des valeurs importantes sur l'axe 1 car l'agrégation de ces quatre critères mettent en avant leurs performances.

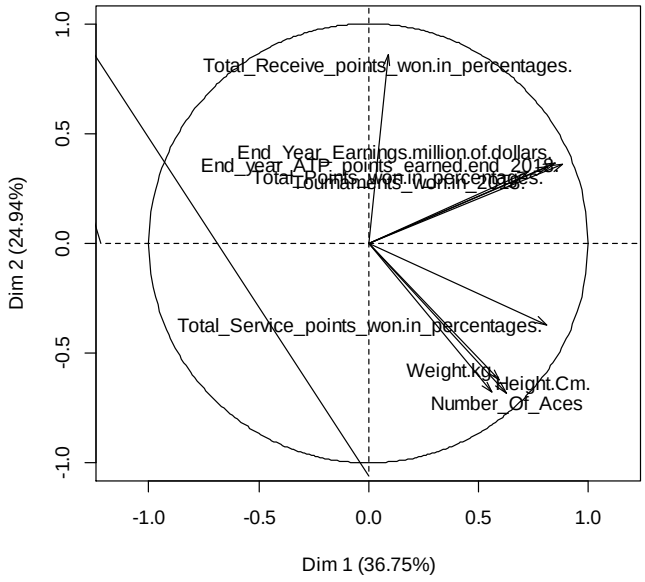
Voici le graphe des individus avec une échelle modifiée permettant de mieux visualiser les différences entre individus qui se ressemblent.



A l'inverse, les joueurs tels Dzumhur, Jaziri ou De Minaur ont des valeurs négatives sur l'axe 1 ce qui démontre que l'agrégation des quatre critères (variables retenues pour expliquer l'axe 1) donne un indice très faible relativement aux autres joueurs du top 50.

Aussi, on a des cas « moyens » tels ici que Verdasco qui a une valeur à peu près nulle sur l'axe 1 ce qui montre qu'il possède un indice moyen sur la conjonction des trois critères.

Maintenant expliquons l'axe 2, pour cela on retourne au graphe du cercle des corrélations.
On voit déjà que les corrélations sont moins fortes ce qui s'explique par le fait qu'il s'agisse du deuxième axe et que par définition l'axe 2 est un axe de variabilité moins important que l'axe 1.



Ici, on n'a pris en compte que les variables ayant un cosinus supérieur à 60% afin de mettre davantage en évidence les variables bien projetées.
On voit que l'axe 2 est expliqué par les variables Total_Receive_points_won(in_percentages) et dans une moindre mesure les variables Weight, Height, Number_Of_Aces et Total_Service_points_won(in_percentages).
Total_Receive_points_won(in_percentages) a une forte valeur sur l'axe 2 ce qui indique que les meilleurs relanceurs du top 50 seront en haut du graphe des individus et que les moins bons relanceurs seront en bas.
La pondération des variables qui construit l'axe 2 permet de montrer que les joueurs ayant une grande taille et un fort taux de relance seront attirés là où la pondération est la plus élevée.
Si l'on observe le graphe des individus, on constate en effet que les meilleurs relanceurs se situent le plus au nord de l'axe 2 plus précisément, on voit que les meilleurs relanceurs tels que Djokovic, Nadal, Schwartzman, Fognini ou Federer ont une contribution élevée sur l'axe 2 ce qui souligne leur efficacité à la relance.
A l'inverse, ceux qui ont un taux de points gagnés à la relance plus faible se situeront au sud de l'axe 2.
Pareillement, on peut voir que les joueurs ayant une taille ou un poids élevé vont être au sud de l'axe 2 car on observe sur le graphe des corrélations que les variables Weight et Height sont corrélées négativement à l'axe 2 ce qui impliquerait que ceux ayant un poids et une taille plus faible seront au nord.
Aussi, on peut regarder les valeurs calculées afin de mieux décrire les contributions sur les axes qui par la suite détermineront l'emplacement des individus.
La contribution de Height et Weight est en effet très élevée pour l'axe 2(respectivement 14,27 et 12,01).

Comme annoncé précédemment, on va analyser l'axe 3 pour capter les 15% d'informations nécessaires. Pour cet axe, nous ne pourrons pas analyser le graphique des corrélations vu précédemment car il n'est pas représenté.
Nous nous contenterons donc d'analyser les données numériques produites par FactoMineR.

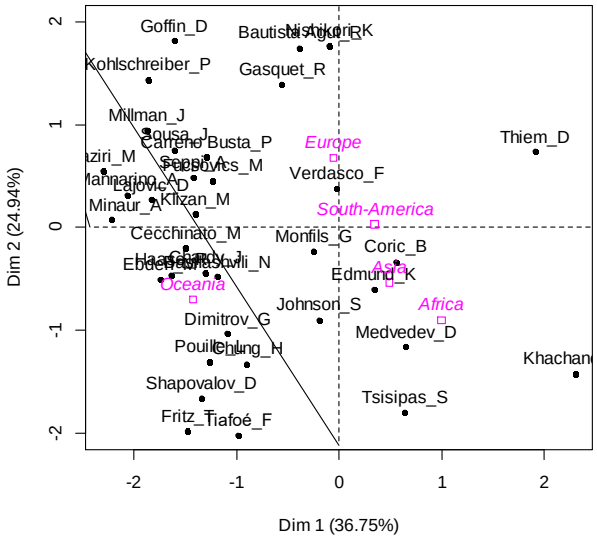
Sur l'axe 3, les variables qui contribuent le plus à l'élaboration de celui-ci sont Age (Years), Professionnal_Experience(in_Years) et Double_Faults avec respectivement des contributions égales à 32,246, 26,11 et 13,97.
Dans la mesure où leurs cosinus carrés sont relativement élevés, ils peuvent être utilisés car leur projection sera bonne notamment la variable Double_Faults qui a un cos carré égal à 0,352.
Parce qu'il existe un coefficient de corrélation important (environ 0,95) entre l'âge et l'expérience professionnelle, il est cohérent de voir ces deux variables apporter une contribution élevée ensemble.
Par conséquent, nous verrons dans l'axe 3 une séparation claire entre jeunes et moins jeunes individus.
Les individus les plus jeunes tels que De Minaur(19 ans) ou Zverev(22 ans) seront situés au sud de l'axe car ils ont des valeurs faibles sur les principales variables de contribution de l'axe 3.

A l'inverse, Federer étant le joueur le plus âgé(37 ans) aura de fortes coordonnées sur l'axe 3 et sera situé au nord.
Les joueurs moyens tels que Goffin(28 ans) ou Dimitrov(27 ans) seront situés sur le centre de l'axe 3.
Aussi, l'axe 3 mettra en évidence les joueurs adeptes des doubles fautes tels que Basilashvili qui compte tenu du grand nombre de doubles fautes qu'il a commis sera situé au sud de l'axe 3 car la variable Double_Faults a explique de façon négative l'axe 3.

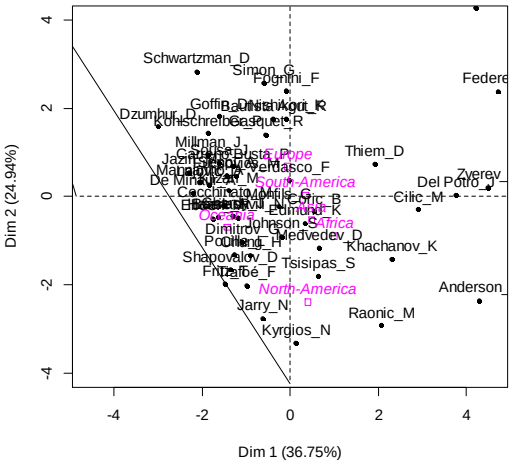
Ceci conclut notre ACP qui a permis de restituer 76,8% des informations du jeu de données en ne prenant compte que les variables quantitatives.

Maintenant, on va essayer d'exploiter les informations des variables qualitatives du jeu de données en les ajoutant en tant que variables illustratives dans le cadre de l'ACP.

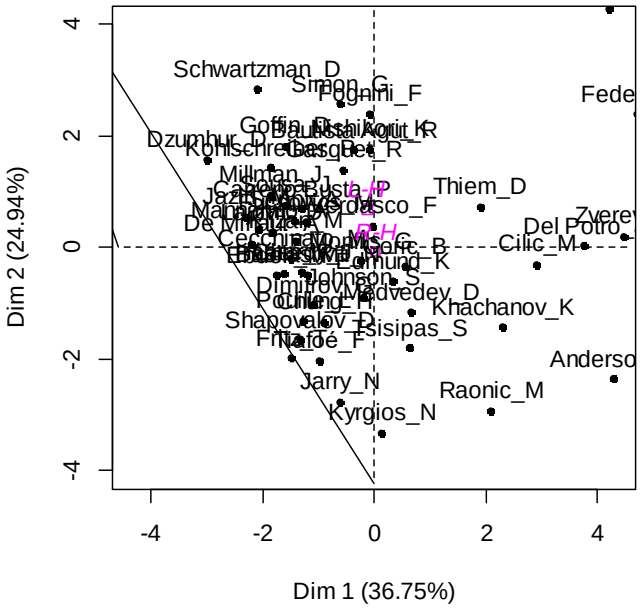
Dans un souci de lisibilité, nous ajouterons les variables qualitatives les unes après les autres.



Pour la variable 'continent', on voit que chaque variable est placée au barycentre des individus prenant cette variable :
Par exemple, Africa qui est pris par seulement deux individus est placé exactement entre Anderson et Jaziri qui sont les deux joueurs « africains ».
On remarque que plus une variable est représentée par les individus et plus elle va avoir tendance à se situer à l'origine des axes :
Une grande partie des joueurs est européen et on peut voir que la variable europe est située au milieu de l'axe 1.
On peut aussi avoir des variables fortes qui impactent la localisation de la variable continent : Isner par exemple qui a une très forte valeur sur l'axe 2 impacte la variable North_America qui elle aussi prend une forte variable sur l'axe 2 :
Voici le même graphique avec une échelle différente pour nous permettre d'analyser ce fait :



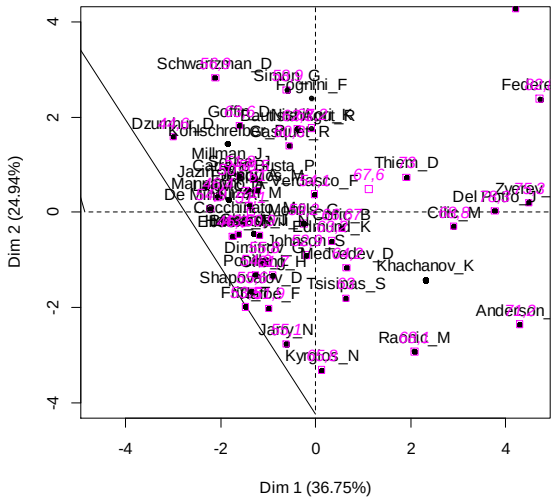
Pour la variable Laterality, on voit que dans la mesure où 94% des joueurs sont droitiers, la variable R-H (Right-Hand) se situe à l'intersection des axes 1 et 2.



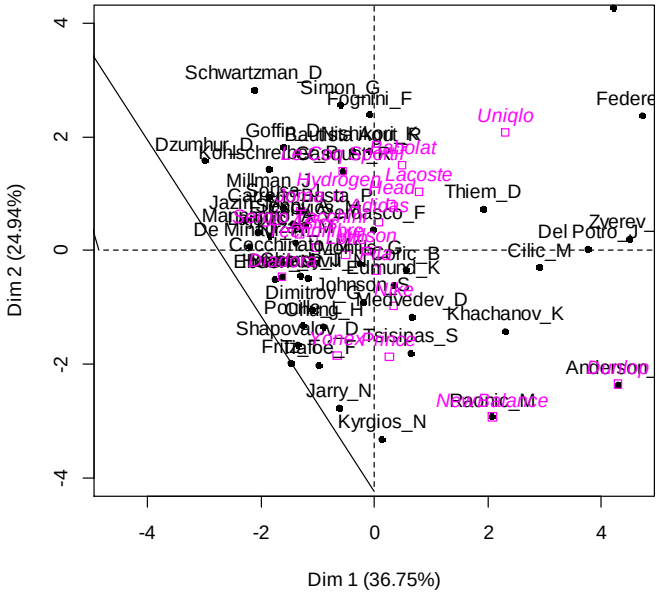
La modalité L-H (Left-Hand) est elle tirée par l'individu Nadal qui a une forte valeur sur l'axe 2.

Quand il existe un nombre de modalités égale ou presque au nombre de variables, le barycentre des modalités va se confondre avec les variables prenant cette modalité :

Nous pouvons le constater aisément via le graphique suivant mettant en avant le pourcentage de matches remportés :



Enfin voici le premier plan du jeu de données montrant la localisation des modalités des variables qualitatives Outfit-Brand et Tennis-Racket-Brand par rapport aux individus :



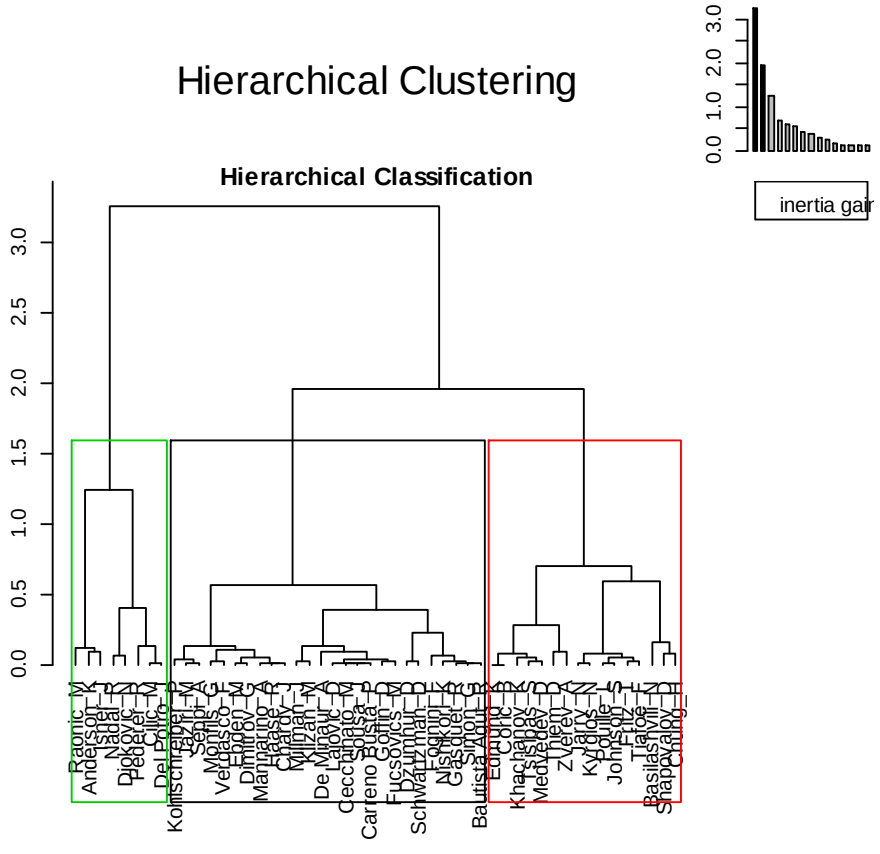
Malgré l'échelle, on peut constater que les modalités sont au barycentre des variables qui les prennent.
Uniqlo est précisément entre Federer et Nishikori, New balance est situé sur Raonic car il est le seul à avoir new balance comme marque de vêtements.

Quatrième partie : Typologie et classification des individus

L'intérêt de la classification est de mettre en évidence des liens hiérarchiques entre les individus ou de détecter un nombre de classes naturel au sein d'une population.
Pour cela, nous allons comme depuis le début utiliser le package FactoMineR du logiciel R.

Cette classification ne peut avoir lieu qu'après avoir effectué l'ACP du jeu de données.
En effet, il sera important de comparer l'information capté par les axes lors de l'ACP pour faire une comparaison avec la classification ascendante hiérarchique.

Regardons de plus près l'arbre représentant l'analyse en composante hiérarchique :



On peut déjà constater un certain nombre de classe, parce que le nombre de classe est élevé, il est nécessaire de « couper » l'arbre afin de bien mettre en évidence les ressemblances entre individus.
Le découpage qui va être pris en compte est celui proposé par le logiciel, il est égal à trois classes.
On ne prendra en compte que trois classes parce qu'à partir du quatrième saut, la perte d'inertie inter devient relativement faible.
On a 13 variables quantitatives sur le jeu de donnée par conséquent, la somme des sauts sera égale à 13.
On choisira un niveau de découpage en trois niveaux afin de prendre en compte ((3,3+2+1,3)/13) = 52%) 52% de l'ensemble de l'information du jeu de données.

Avec l'ACP, on a pu collecter environ 76% de l'information du jeu de données.
La classification nous donne donc un résumé plus grossier.

Voyons de plus près les individus qui composent chaque classe :
On remarque d'emblée que les variables ayant été à l'origine de la construction des trois premiers axes lors de l'ACP sont très bien représentées.
Effectivement on a des groupes qui rassemblent des joueurs avec des caractéristiques communes ; par exemple dans le groupe « vert », on voit que les joueurs ont de fortes coordonnées sur le premier axe et sachant que le premier axe est construit grâce à des variables de performance au service ou de gains monétaires, il est cohérent de voir Raonic, Isner et Anderson (gros serveurs) et Federer, Nadal et Djokovic (gains monétaires importants).

Par ailleurs, on voit un groupe « rouge » qui met en évidence des joueurs ayant des caractéristiques communes par rapport à l'âge (l'âge ayant une forte contribution avec l'axe 3, il est cohérent que ce groupe « rouge » soit le fruit d'un découpage entre 2 et 3 classes.
Parce que le groupe rouge est plus éloignée du groupe « vert » que le groupe « noir », il est clair que le groupe « vert » représente l'axe 3.
Le groupe « noir » représente des joueurs ayant de fortes valeurs sur l'axe 2, l'ensemble des joueurs de ce groupe a effectivement un taux points gagnés à la relance élevé or cette variable possède une forte projection sur l'axe 2.

Conclusion de l'exposé

Finalement lorsque l'on compare la hiérarchisation obtenue via l'ACP et la méthode de Ward, on voit que les joueurs sont regroupés par rapport à des variables « fictives » qui résultent de l'agrégation des variables du jeu de données initial.
Le regroupement est cohérent mais on ne retrouve pas un regroupement basé sur les rangs des joueurs (classement ATP) même si les trois meilleurs joueurs partagent la même classe.
On voit donc que l'ACP et la méthode de Ward s'utilisent assez bien pour résumer des jeux de données mais ne prennent pas en compte une potentielle hiérarchisation des variables.
En effet, avec la méthode de Ward on voit que Raonic est dans le même groupe que Nadal et Djokovic alors que Zverev (qui est dans le groupe « rouge ») a des performances largement supérieures à Raonic.
Finalement, on voit que si l'on voulait avoir des résultats en rapport avec la réalité, il aurait fallu utiliser un outil permettant de hiérarchiser les variables du jeu de données de tel sorte que l'on ait au final des résultats en rapport avec les performances des joueurs.