INSA Rouen Normandie

Génie Mathématique 3ème année

Projet MSRO

Février 2024

# Analyse de données
# Étude des drafts de la NBA


Yassir ANKOUDY

# Contents

**1 Contexte** **2**

**2 Présentation du jeu de données** **3**

2.1 Nombre d'individus, années et choix de draft . . . . . 3

2.2 Variables mesurées . . . . . 3

2.3 Personnalités Notables . . . . . 4

2.4 Utilisation du Jeu de Données . . . . . 4

**3 Choix des variables et d'une plage de données** **5**

3.1 Code R . . . . . 5

**4 Analyse des variables selectionnées** **7**

4.1 Résumé des données . . . . . 7

4.2 Boîtes à moustaches . . . . . 8

**5 ACP : Analyse en composantes principales** **10**

5.1 Composantes principales . . . . . 10

5.2 Projection . . . . . 11

5.3 Résultats de l'ACP . . . . . 12

5.4 Interprétation des clusters . . . . . 12

**6 Cercle de corrélations** **13**

**7 Conclusion** **15**

# 1 Contexte

La draft de la NBA est un événement annuel où les équipes de la National Basketball Association (NBA) sélectionnent de nouveaux joueurs éligibles pour rejoindre la ligue. C'est un mécanisme central pour introduire des talents universitaires et internationaux dans la ligue et est crucial pour maintenir un équilibre concurrentiel parmi les équipes.

Le système de draft est conçu de manière à favoriser l'équité entre les équipes : celles ayant les moins bons bilans lors de la saison précédente obtiennent les chances les plus élevées de choisir en premier, grâce à une loterie. Ce processus vise à donner aux équipes en difficulté l'opportunité de s'améliorer plus rapidement en ajoutant les meilleurs jeunes talents à leur effectif.

Les joueurs éligibles pour le draft incluent généralement des stars du basket universitaire, des joueurs internationaux et d'autres qui se déclarent admissibles. Être sélectionné dans le draft, surtout dans les premiers tours, est souvent considéré comme un honneur et une opportunité significative, car cela peut mener à des contrats professionnels lucratifs et une carrière dans la ligue la plus prestigieuse de basket-ball.

La performance des joueurs lors du combine de la NBA, qui comprend une série de tests physiques, d'exercices de basket-ball et d'entretiens, peut affecter leur classement dans le draft. Les équipes utilisent les données du combine pour évaluer la condition physique, l'athlétisme et la préparation mentale des joueurs. Les jeunes talents sont évalués sur une multitude de critères, allant de leurs performances en match aux statistiques détaillées, en passant par des mesures physiques et athlétiques prises lors des combines – des évaluations pré-draft qui comprennent des tests de compétences, des entrevues, et divers exercices physiques et médicaux.

L'importance des drafts pour les franchises de la NBA ne peut être sous-estimée. Ils permettent aux équipes de se renforcer en ajoutant de jeunes joueurs prometteurs à leurs effectifs, joueurs qui peuvent devenir des superstars ou des contributeurs clés pour l'avenir. Les décisions prises lors des drafts peuvent façonner le destin d'une franchise pour de nombreuses années. Ainsi, une analyse minutieuse et complète des prospects est cruciale.

Le jeu de données dont nous disposons offre une perspective unique sur les prospects de draft en fournissant des informations détaillées sur leurs attributs physiques et athlétiques. Ces données peuvent inclure la taille sans chaussures, la taille avec chaussures, l'envergure des bras, le poids, le pourcentage de graisse corporelle, la détente verticale, la rapidité, l'agilité, et bien d'autres mesures. L'analyse de ces données peut aider à évaluer le potentiel athlétique des joueurs, ce qui est un facteur clé dans les décisions de draft.

La combinaison de l'analyse statistique des performances passées des joueurs et des données physiques offre une approche holistique de l'évaluation des talents, permettant aux équipes de la NBA de prendre des décisions éclairées et stratégiques pendant la draft. Cette analyse peut aussi être utilisée pour prédire le succès futur des joueurs dans la ligue, ce qui est d'une valeur inestimable pour les franchises qui cherchent à optimiser leur sélection de draft.

# 2 Présentation du jeu de données

Le jeu de données analysé dans ce rapport comprend des mesures collectées lors des Drafts de la NBA, qui couvrent les années 2009 à 2017. Ces informations ont été obtenues grâce aux données fournies par DraftExpress, accessibles publiquement via un répertoire GitHub et data.world.

## 2.1 Nombre d'individus, années et choix de draft

Le dataset contient les mesures de 517 joueurs uniques, fournissant un échantillon substantiel pour l'analyse statistique. Les 3 premières colonnes correspondent au nom et prénom du joueur, l'année durant laquelle il a été drafté et le numéro du choix de draft. Comme expliquée dans le contexte, la "draft pick" en NBA, ou le choix de draft, est une position assignée à une équipe. Les équipes choisissent à tour de rôle des joueurs éligibles. Les choix de draft sont déterminés par un mélange de loterie pour les premiers choix et les bilans de la saison précédente pour le reste de la sélection. Les équipes utilisent les choix de draft pour sélectionner de nouveaux talents dans l'espoir de bâtir ou de renforcer leur effectif pour les saisons à venir. La position de draft est critique, car elle détermine l'ordre dans lequel les équipes peuvent choisir les joueurs. Les premiers choix sont particulièrement précieux car ils offrent l'opportunité de sélectionner les joueurs les plus talentueux et les plus prometteurs disponibles.

## 2.2 Variables mesurées

Plusieurs variables quantitatives sont incluses dans le jeu de données, offrant des détails sur les attributs physiques et les capacités athlétiques des joueurs :

- **Height (No Shoes)** - Taille (sans chaussures)
- **Height (With Shoes)** - Taille (avec chaussures)
- **Wingspan** - Envergure
- **Standing reach** - Atteinte en position debout
- **Vertical (Max)** - Détente verticale (max)
- **Vertical (Max Reach)** - Détente verticale maximale (atteinte)
- **Vertical (No Step)** - Détente verticale (sans élan)
- **Vertical (No Step Reach)** - Détente verticale sans élan (atteinte)
- **Weight** - Poids

- **Body Fat** - Pourcentage de graisse corporelle
- **Hand (Length)** - Longueur de la main
- **Hand (Width)** - Largeur de la main
- **Bench** - Banc de musculation (nombre de répétitions au développé-couché)
- **Agility** - Agilité
- **Sprint** - Sprint

Ces mesures clés sont utilisées pour évaluer les prospects de la draft et peuvent influencer de manière significative les décisions des équipes de la NBA.

## 2.3 Personnalités Notables

Le jeu de données comprend également plusieurs joueurs qui ont par la suite fait leurs preuves en NBA, ce qui ajoute une valeur historique et analytique au dataset, on peut citer Stephen Curry, Klay Thompson, Jimmy Butler et bien d'autres.

Figure 1: Stephen Curry et Klay Thompson champions NBA en 2021/2022

## 2.4 Utilisation du Jeu de Données

Les données sont particulièrement utiles pour des analyses avancées, telles que l'Analyse en Composantes Principales (ACP) et la modélisation prédictive, pour aider à prédire la réussite des joueurs en NBA basée sur leurs attributs mesurés avant leur draft.

# 3 Choix des variables et d'une plage de données

Pour l'analyse des données du combine de la draft NBA, nous avons opéré une sélection rigoureuse des variables afin d'optimiser la pertinence et la qualité de notre étude.

Les variables non quantitatives telles que *Year*, *X*, et *Draft Pick* ont été exclues de l'analyse, celles-ci ne contribuant pas significativement à l'évaluation des performances physiques des joueurs.

De même, les mesures *Hand Length* et *Hand Width* ont été omises en raison du nombre insuffisant de données, ce qui pourrait conduire à des interprétations erronées ou à un biais dans les résultats de l'Analyse en Composantes Principales (ACP).

En outre, les individus présentant des valeurs manquantes (NA) ont été retirés de l'ensemble de données. Cette décision a été prise car notre connaissance actuelle ne nous permet pas encore de déterminer une méthode de remplacement des données adéquate sans risquer d'introduire des distorsions.

La période choisie pour notre analyse s'étend de 2009 à 2013, correspondant à ce qui est souvent considéré comme l'âge d'or de la NBA au XXIe siècle. Cette ère est marquée par l'émergence de plusieurs figures emblématiques qui sont aujourd'hui des légendes du sport, comme celles que nous avons pu citer plus tôt. Le choix de cette plage temporelle spécifique nous permet non seulement de concentrer notre analyse sur un segment historique notable mais aussi d'observer l'évolution et l'impact de joueurs de renom qui ont influencé le jeu à leur manière.

## 3.1 Code R

### Modification du jeu de données initial (data)

Voici le code R que nous avons utilisé pour modifier le jeu de données, nous avons aussi modifié les noms des variables en français pour une meilleure compréhension.

```r
data <- read.csv(file = "nba_draft_combine_all_years.csv", header = TRUE,
    sep = ',', dec = ',', stringsAsFactors = TRUE)
names(data)
str(data$Year)
data$Year <- as.numeric(as.character(data$Year))
data <- subset(data, Year >= 2009 & Year <= 2013)
donnees <- subset(data, select = -c(`Hand..Length.`, `Hand..Width.`, `X`,`
    Draft.pick` , `Year`))
donnees <- na.omit(donnees)
donnees_num <- donnees[sapply(donnees, is.numeric)]

names(donnees_num)[names(donnees_num) == "Height..No.Shoes."] <- "hauteur_
    sans_chaussures"
names(donnees_num)[names(donnees_num) == "Height..With.Shoes."] <- "hauteur
    _avec_chaussures"
names(donnees_num)[names(donnees_num) == "Wingspan"] <- "envergure"
names(donnees_num)[names(donnees_num) == "Standing.reach"] <- "atteinte_
    debout"
```

```r
names(donnees_num)[names(donnees_num) == "Vertical..Max."] <- "saut_vertical_max"
names(donnees_num)[names(donnees_num) == "Vertical..Max.Reach."] <- "atteinte_max_saut"
names(donnees_num)[names(donnees_num) == "Vertical..No.Step."] <- "saut_sans_elan"
names(donnees_num)[names(donnees_num) == "Vertical..No.Step.Reach."] <- "atteinte_sans_elan"
names(donnees_num)[names(donnees_num) == "Weight"] <- "poids"
names(donnees_num)[names(donnees_num) == "Body.Fat"] <- "pourcentage_graisse_corp"
names(donnees_num)[names(donnees_num) == "Bench"] <- "repetitions_developpe_couche"
names(donnees_num)[names(donnees_num) == "Agility"] <- "agilite"
names(donnees_num)[names(donnees_num) == "Sprint"] <- "temps_sprint"
```

# 4 Analyse des variables selectionnées

## 4.1 Résumé des données

Pour commencer notre analyse, il est logique de commencer par ce qu'il y a de plus simple mais aussi de plus important: la moyenne, l'écart-type, et les quartiles.

Voici le code qui nous permettra de générer le tableau contenant ces informations.

```r
summary(donnees_num)
statistiques <- data.frame(
  Moyenne = sapply(donnees_num, mean, na.rm = TRUE),
  EcartType = sapply(donnees_num, sd, na.rm = TRUE),
  Q1 = sapply(donnees_num, quantile, probs = 0.25, na.rm = TRUE),
  Median = sapply(donnees_num, median, na.rm = TRUE),
  Q3 = sapply(donnees_num, quantile, probs = 0.75, na.rm = TRUE)
)
print(statistiques)
```

Voici le tableau obtenu:

| | Moyenne | EcartType | Q1 | Median | Q3 |
|---|---|---|---|---|---|
| hauteur_sans_chaussures | 77.638393 | 3.2723600 | 75.4375 | 77.875 | 80.2500 |
| hauteur_avec_chaussures | 78.921875 | 3.2727355 | 76.9375 | 79.000 | 81.3125 |
| envergure | 82.352009 | 3.9824118 | 79.4375 | 82.500 | 85.3125 |
| atteinte_debout | 103.251116 | 4.8135353 | 100.0000 | 103.500 | 106.6250 |
| saut_vertical_max | 34.906250 | 3.4439986 | 32.5000 | 35.000 | 37.5000 |
| atteinte_max_saut | 138.157366 | 4.1452146 | 135.5000 | 138.500 | 141.5000 |
| saut_sans_elan | 29.397321 | 2.8396128 | 27.5000 | 29.500 | 31.5000 |
| atteinte_sans_elan | 132.648438 | 4.5476436 | 129.5000 | 133.000 | 135.6250 |
| poids | 217.205357 | 24.4971448 | 198.0000 | 216.000 | 234.0000 |
| pourcentage_graisse_corp | 7.150893 | 2.2509189 | 5.5000 | 6.750 | 8.2250 |
| repetitions_developpe_couche | 10.754464 | 5.0270548 | 7.0000 | 10.000 | 15.0000 |
| agilite | 11.306429 | 0.5679463 | 10.9375 | 11.205 | 11.6625 |
| temps_sprint | 3.298036 | 0.1310348 | 3.1975 | 3.280 | 3.3800 |

Figure 2: Résumé des données avec moyenne, écart-type et quartiles

L'accent est mis ici sur **l'Écart Type** qui mesure la dispersion des valeurs autour de la moyenne pour chaque variable. Un écart type élevé indique une grande variabilité des mesures, suggérant une diversité significative dans la caractéristique mesurée parmi les joueurs. Inversement, un écart type faible implique que les valeurs sont plus uniformément réparties autour de la moyenne, indiquant moins de variabilité. Par exemple, la variable *hauteur_sans_chaussures* a un écart type de 3.27, ce qui est relativement faible, reflétant une homogénéité dans la taille des joueurs sans chaussures. En contraste, la variable *poids* présente un écart type beaucoup plus élevé de 24.49, mettant en évidence une variabilité significative dans le poids des joueurs.

Ce niveau de dispersion est essentiel pour comprendre l'éventail des capacités et attributs physiques des joueurs, permettant aux équipes de la NBA de mieux évaluer et comparer les prospects. Une analyse détaillée des écarts types peut aider à identifier les caractéristiques physiques qui diffèrent le plus parmi les joueurs, ce qui peut être un facteur déterminant dans les décisions de draft.

## 4.2 Boîtes à moustaches

Les boîtes à moustaches, ou diagrammes en boîte, sont un outil statistique graphique qui permet de visualiser la distribution d'une série de données. Chaque boîte à moustaches représente cinq mesures de position : le minimum, le premier quartile (Q1), la médiane, le troisième quartile (Q3), et le maximum. Les "moustaches" s'étendent du premier au troisième quartile, couvrant l'intervalle interquartile (IIQ), et offrent une vue sur la variabilité et la dispersion des données. Les points situés en dehors des moustaches sont souvent considérés comme des valeurs aberrantes.

Voici le code R nous permettant de générer ces boîtes.

```r
par(mfrow=c(4,4))
for(i in 1:ncol(donnees_num)) {
  boxplot(donnees_num[,i], main = names(donnees_num)[i])
}
```

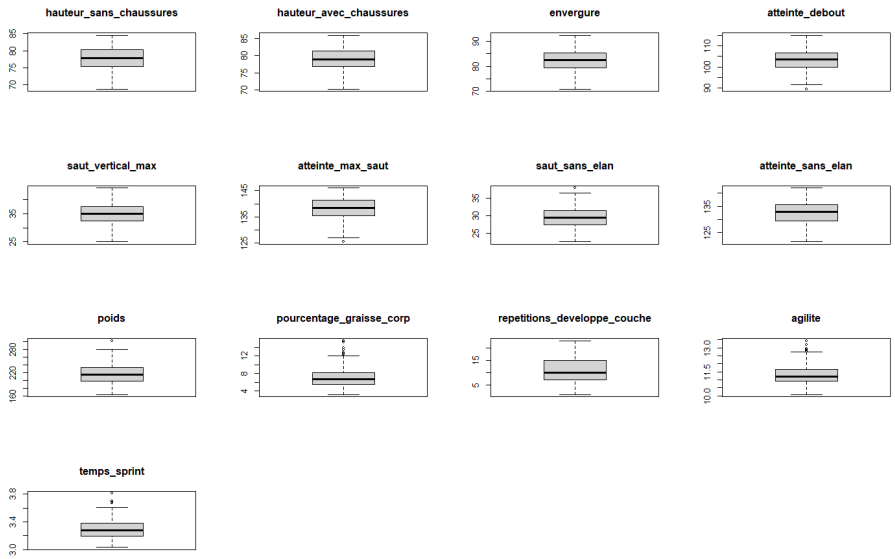


Figure 3: Boîtes à moustache

L'analyse des boîtes à moustaches pour les mesures physiques et athlétiques des joueurs

nous montre les caractéristiques suivantes :

- La **hauteur sans chaussures** et la **hauteur avec chaussures** montrent une distribution assez symétrique autour de la médiane, indiquant une similitude dans les mesures de la taille des joueurs.
- L'**envergure** et l'**atteinte debout** présentent également des distributions symétriques, ce qui suggère une variabilité modérée et une absence de valeurs aberrantes significatives.
- Le **saut vertical maximal** et le **saut sans élan** révèlent une variabilité plus grande, comme en témoignent les plus larges intervalles interquartiles.
- Le **poids** et le **pourcentage de graisse corporelle** montrent des distributions avec des valeurs aberrantes, indiquant une hétérogénéité notable dans la composition corporelle des joueurs.
- Les **répétitions au développé couché** et l'**agilité** semblent avoir une variabilité relativement faible, tandis que le **temps de sprint** montre une distribution plus étendue, suggérant des différences marquées dans la vitesse des joueurs.

Ces observations nous permettent de comprendre que les variables dans lesquelles ont retrouve le plus de "différences" entre les individus ne sont pas celles qui concernent le physique comme la hauteur, l'envergure, ou encore l'atteinte debout, mais plutôt celles qui concernent les aptitudes physiques des joueurs, comme le saut, l'agilité etc..., mais aussi celles qui concernent l'entretien corporel, comme le poids et le pourcentage de graisse. Ces boîtes à moustaches rendent compte de la diversité des profils au niveau de la qualité du joueur et non simplement sur ses aptitudes innées (des capacités qu'il a obtenu par naissance), et ceci est très intéressant car cela nous montre que les drafts NBA ne sont pas uniquement basées sur la taille comme on pourrait le penser, mais plutôt sur le sérieux, et le travail qu'effectue le joueur pour s'améliorer.

# 5 ACP : Analyse en composantes principales

L'Analyse en Composantes Principales (ACP) a été réalisée sur notre jeu de données afin d'identifier les structures sous-jacentes et de réduire la dimensionnalité tout en conservant un maximum d'information. Cette technique statistique transforme les variables corrélées en un nombre réduit de composantes indépendantes qui expliquent une proportion substantielle de la variabilité dans les données.

## 5.1 Composantes principales

Le *screen plot* est un outil visuel utilisé pour déterminer le nombre de composantes principales à retenir dans une analyse en composantes principales (ACP). La figure ci-dessous illustre le pourcentage de variance expliquée par chaque composante principale dans notre jeu de données.

Ce graphique a été obtenu avec le code R suivant:

```r
barplot(res.acp$eig[,2], names.arg = 1:nrow(res.acp$eig), main = "
    Pourcentage de variance expliqu e", xlab = "Axes", ylab = "Pourcentage
     de variance")
print(paste("Nombre de composantes principales s lectionn es : ", res.acp
    $ncp))
```

Voici le graphique obtenu après éxécution:

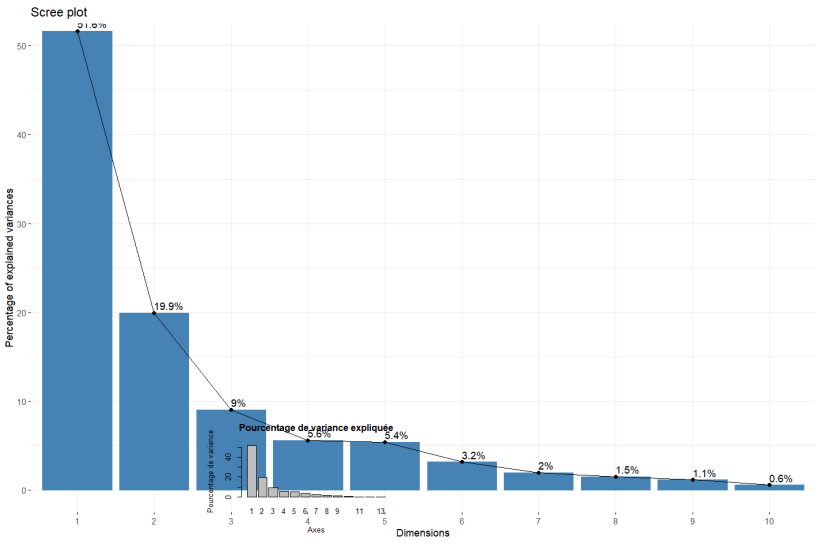


Figure 4: Diagramme des composantes principales

La première composante principale (Dim1) domine le graphique, expliquant 51.7% de la variance totale, ce qui met en évidence son importance dans la capture des informations

contenues dans le jeu de données. La deuxième composante principale (Dim2) explique 19.9% de la variance, contribuant ainsi à un total cumulatif de plus de 70% avec la première composante.

Les composantes suivantes présentent une contribution décroissante à la variance totale, comme illustré par les barres successivement plus petites sur le graphique. Cette tendance décroissante indique que les premières composantes retiennent la majorité des informations significatives. La recherche d'un point d'inflexion, ou 'coude', nous permet de constater que les gains marginaux en termes de variance expliquée deviennent minimes après la deuxième composante. Cela justifie notre choix de se concentrer sur les deux premières composantes pour l'analyse subséquente.

En résumé, le graphique suggère que Dim1 et Dim2 fournissent une représentation fidèle et simplifiée de la structure sous-jacente des données, capturant l'essentiel des variations observées parmi les mesures des joueurs.

## 5.2 Projection

Voici la projection des individus obtenue avec le code R suivant.

```r
res.acp <- PCA(donnees_num, scale.unit = TRUE, graph = FALSE)

fviz_pca_ind(res.acp, title = "Projection des individus (ACP)")

fviz_eig(res.acp, addlabels = TRUE, ylim = c(0, 50))
```

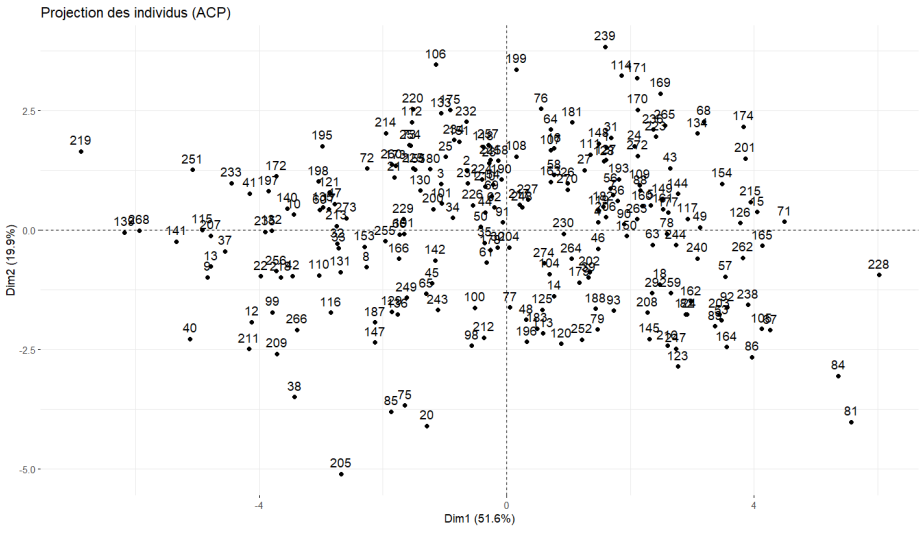


Figure 5: ACP

## 5.3 Résultats de l'ACP

La projection des individus sur les deux premières composantes principales nous donne un aperçu des tendances et des regroupements potentiels parmi les joueurs. La première composante principale, peut être interprétée comme un facteur reflétant une combinaison de caractéristiques physiques et athlétiques. La deuxième composante principale, pourrait capturer d'autres aspects de la performance ou des attributs physiques qui ne sont pas pris en compte par la première composante.

## 5.4 Interprétation des clusters

Dans la figure ci-dessus, nous observons une dispersion significative des joueurs le long de la première composante, suggérant qu'elle pourrait représenter une mesure globale de la capacité athlétique ou une caractéristique physique prédominante comme la taille ou la portée du joueur. La répartition plus modérée des joueurs le long de la deuxième composante suggère qu'elle pourrait être associée à des capacités spécifiques, telles que l'agilité ou la puissance.

Bien que les données ne montrent pas de regroupements clairement définis, il existe une concentration d'individus autour de l'origine, ce qui peut indiquer un profil moyen des joueurs de draft. Les points qui sont situés plus loin de l'origine peuvent représenter des joueurs exceptionnels ayant des capacités uniques ou des caractéristiques physiques extrêmes.

Cette analyse préliminaire de l'ACP offre des informations précieuses pour les stratégies de recrutement, soulignant l'importance de considérer une gamme de caractéristiques lors de l'évaluation des prospects de la NBA Draft. L'étude suggère également un potentiel pour des analyses plus approfondies, telles que l'intégration d'autres variables ou l'utilisation de méthodes de clustering avancées pour définir des typologies de joueurs plus détaillées.

# 6 Cercle de corrélations

Le cercle des corrélations, visualisé dans le contexte d'une Analyse en Composantes Principales (ACP), permet d'évaluer la contribution des différentes mesures athlétiques et physiques des joueurs aux deux premières composantes principales identifiées par l'ACP. Cette analyse révèle les relations linéaires entre les variables originales et les composantes principales.

Nous avons pu générer le cercle avec le code R suivan:

```
fviz_pca_var(res.acp, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,
             title = "Cercle des corr lations (ACP)")
```

Nous avons obtenu le résultat suivant:

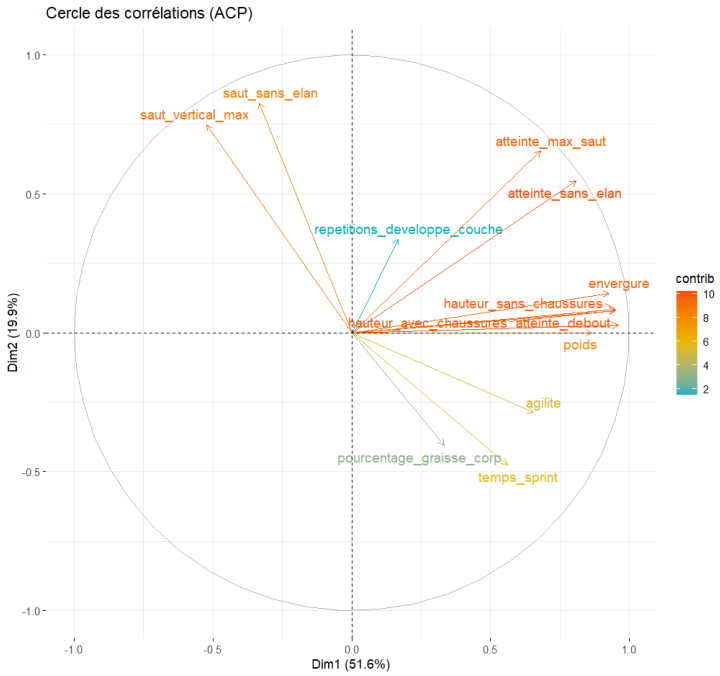


Figure 6: Cercle de corrélations

Dans notre cercle des corrélations, nous pouvons voir que les variables telles que *atteinte_sans_elan* et *atteinte_max_saut* et *saut_sans_elan* sont fortement corrélées au premier axe, suggérant que cette composante principale est fortement influencée par les capacités de saut et d'atteinte verticale des joueurs. Ceci est pertinent pour la draft NBA car ces attributs sont essentiels pour de nombreuses actions en jeu comme les dunks et les contres.

D'autre part, des mesures telles que *hauteur_sans_chaussures*, *hauteur_avec_chaussures* et *envergure* montrent une corrélation positive avec le second axe , ce qui indique que cette composante pourrait être associée à la taille et l'allonge des joueurs, des facteurs tout aussi cruciaux dans le basketball moderne. La taille et l'envergure peuvent influencer le choix des joueurs pendant la draft car elles sont des indicateurs clés du potentiel d'un joueur à défendre contre des adversaires et à réussir des tirs par-dessus des défenseurs.

Les variables ayant une faible corrélation avec les deux premières composantes, comme *pourcentage_graisse_corp* et *temps_sprint*, suggèrent que ces aspects, bien qu'importants, ne sont pas les facteurs prédominants de différenciation dans cette analyse. Toutefois, cela ne diminue pas leur importance potentielle dans l'évaluation globale des capacités d'un joueur.

En résumé, le cercle des corrélations fournit une vue d'ensemble des caractéristiques qui pourraient influencer la sélection d'un joueur lors de la draft NBA. Les équipes peuvent utiliser ces informations pour prendre des décisions stratégiques et ciblées lors de la sélection des joueurs, en se concentrant sur les attributs qui correspondent le mieux à leurs besoins et à leur style de jeu.

# 7 Conclusion

À travers l'analyse menée, nous avons pu extraire des informations significatives sur les mesures physiques et athlétiques des prospects de la draft NBA en utilisant une Analyse en Composantes Principales (ACP). Cette étude a non seulement révélé les attributs prépondérants qui caractérisent les joueurs entrant dans la ligue mais a également permis d'identifier des certaines données qui peuvent éclairer les décisions de recrutement des équipes.

Les deux premières composantes principales ont démontré leur capacité à résumer efficacement la variance des données, mettant en lumière des aspects fondamentaux des performances athlétiques. Le cercle des corrélations a aidé à interpréter ces composantes, soulignant l'importance des capacités de saut et de la stature physique dans la distinction des joueurs.

Ce projet souligne la valeur de l'ACP comme outil dans la prédiction du potentiel des joueurs et dans l'élaboration des stratégies de sélection des équipes. Les perspectives pour des recherches futures incluent l'application de techniques de clustering avancées et l'intégration d'autres formes de données, comme les statistiques en match et les données biométriques, pour améliorer encore la précision des évaluations.

En conclusion, l'intégration des méthodes d'analyse de données dans le processus de sélection de la draft NBA offre une opportunité précieuse pour les équipes de maximiser leur retour sur investissement dans la recherche de nouveaux talents.