

Domaine d'application

Rapport écrit par:

Samuel DARMALINGON, Enzo LERICHE

BUT Science des données FI EMS

2024-04-07



IUT de Paris - Rives de Seine
Université Paris Cité

Introduction	3
Question 1	3
Importation des données	4
Gestion des bases de données	4
Comparaison age et performance	7
La meilleure performance des athlètes en fonction de leur age selon la distance parcourue	11
Chez les Femmes	11
Chez les hommes	12
Comparaison entre les hommes et les femmes	13
Question 2	15
Comparaison des modèles sur la performance des joueurs en fonction de leur age	15
L'analyse du meilleur modèle	22
Question 3	23
Normalité Modèle linéaire	23
Normalité Modèle polynomiale	24
Calcul des différents critères	25
Question 4	28
De manière globale	28
Pour les hommes	29
Pour les femmes.....	29
Question 5	29
Importation et nettoyage des données	29
Représentation des données	30
Représentation graphique des 3 modèles	30
Échantillonnage	31
De la base de données globale	31
Des deux groupes séparés	32
Calcul des différents critères	33
Âge ou les points sont maximaux	34
Conclusion	35
Question bonus 1	35
Visuellement.....	35
Graphiquement	36
Critères.....	37
Conclusion	38
Question bonus 2	38
Choix de la question choisi.....	38
Réponse CHATGPT	38
Conclusion par rapport a nos propres résultats	40

Introduction

Le projet vise à explorer la relation entre l'âge et la performance, à travers une analyse approfondie de données réelles. Cette étude est importante pour comprendre comment le vieillissement affecte les performances sportives humaines. Les données fournies couvrent plusieurs épreuves d'athlétisme pour hommes et femmes, ainsi que les performances de joueurs d'échecs, offrant un terrain riche pour examiner l'impact de l'âge sur la performance.

Les données d'athlétisme comprennent des informations détaillées sur les athlètes, leurs performances dans différentes disciplines, leurs âges au moment des compétitions, et plus encore. Ces données permettent de construire un modèle précis de la relation âge-performance en filtrant les performances aberrantes ou manquantes et en discrétisant l'âge des athlètes de manière appropriée, que ce soit par année entière ou par trimestre.

Le projet ajoute une dimension intéressante avec l'inclusion des données sur les joueurs d'échecs. Bien que les échecs soient une discipline cognitive plutôt que physique, l'analyse de la performance en fonction de l'âge dans ce domaine peut offrir des résultats intéressants sur l'effet du vieillissement.

Pour mener à bien cette analyse, le projet propose d'ajuster trois modèles différents aux données : un modèle linéaire, un modèle polynomial et l'équation de Moore. Chaque modèle vise à capturer la nature de la relation entre l'âge et la performance, avec l'équation de Moore nécessitant une approche de régression non linéaire. Les modèles seront évalués à l'aide de plusieurs critères, notamment la somme des résidus carrés (RSS), le coefficient de détermination ajusté (R^2 ajusté), l'AIC corrigé et le BIC, pour déterminer lequel fournit le meilleur ajustement aux données.

Ce travail académique ne se contente pas de traiter les données et d'ajuster des modèles, mais il invite également à une réflexion critique sur les résultats obtenus, notamment en déterminant l'âge auquel la performance atteint son maximum pour différentes épreuves. Ce pic de performance est analysé à travers les différents modèles pour voir si un âge optimal pour la performance existe, quelles que soient les disciplines.

Enfin, le projet encourage l'exploration de nouvelles équations et la comparaison de leurs performances avec celles des modèles déjà établis, ainsi qu'une réflexion sur l'utilisation d'outils comme ChatGPT dans le processus de recherche et d'analyse.

Ce projet nous permet d'appliquer des compétences en statistiques, en programmation R, et en analyse critique pour explorer une question fondamentale : comment l'âge influence-t-il la performance humaine, et peut-on identifier des tendances à travers différentes disciplines ?

Question 1

Importation des données

```
setwd("C:/Users/User/OneDrive/Documents/BUT/2eme annee/domaine dapplication/P  
rojet semestre 4 relation âge-performance-20240126")  
  
resultats_femmes <- read.csv("resultats_femmes.csv", sep=";", encoding="latin  
1")  
  
resultats_hommes <- read.csv("resultats_hommes.csv", sep=";", encoding="latin  
1")  
  
resultats_joueurs_echecs <- read.csv2("resultats_joueurs_echecs.txt", sep = "  
;")
```

Nous avons analysé les résultats sportifs pour les femmes et les hommes, ainsi que les performances de joueurs d'échecs. Et donc avant de faire des analyses statistiques sur ce données, nous avons d'abord effectuer une gestion des bases de données.

Gestion des bases de données

Nous avons commencé par faire un nettoyage de ces bases

```
dim(resultats_femmes)  
## [1] 122319    12  
dim(resultats_hommes)  
## [1] 132569    12  
dim(resultats_joueurs_echecs)  
## [1] 139119     7
```

La base de données des résultats sportifs des femmes comporte 122 319 observations et 12 variables.

La base de données des résultats sportifs des hommes comporte 132 569 observations et 12 variables.

La base de données des résultats des joueurs d'échecs comporte 139 119 observations et 12 variables.

Puis, nous avons observé si ces bases comportaient des valeurs manquantes.

```
sum(is.na(resultats_femmes))  
## [1] 122347  
sum(is.na(resultats_hommes))  
## [1] 132621  
sum(is.na(resultats_joueurs_echecs))  
## [1] 11910
```

Et donc nous avons constaté :

- chez les femmes, il y avait 122 347 manquantes.
- chez les hommes, il y avait 132 621 manquantes.
- chez les joueurs d'échecs, il y avait 11 910 manquantes.

Puis nous avons constaté qu'il y avait une variable qui comportait que des valeurs manquantes, *Var.7* dans les résultats sportifs des hommes et des femmes. Donc, nous avons décidé de la supprimer cette variable. Donc nous nous retrouvons avec 11 variables chez les résultats sportifs des hommes et des femmes.

```
resultats_femmes <- subset(resultats_femmes, select = -Var.7)
resultats_hommes <- subset(resultats_hommes, select = -Var.7)
```

Puis, après avoir examiné la variable "*dis*", nous avons remarqué que chez les résultats sportifs des hommes et des femmes, cette variable comportait la discipline athlétique et sa catégorie dans une même variable. Donc nous avons décidé de séparer ces deux caractéristiques et placer chacune dans une variable soit la variable

- *Sport*
- *Discipline*

Dans la poursuite de notre analyse, dans les bases des résultats sportifs des hommes et des femmes, nous nous intéressons à la variable *Mark* qui exprime la durée que les sportifs ont effectué durant le parcours dans chaque discipline.

Ainsi, nous avons remarqué que certaines observations, il y avait un "h" à la fin de la durée du parcours. Donc, nous avons remplacé ces "h" par des 0 par manipuler par la suite ces données.

```
#FEMMES
resultats_femmes$Mark <- gsub("h",0, resultats_femmes$Mark)

#HOMMES
resultats_hommes$Mark <- gsub("h",0, resultats_hommes$Mark)
```

Par la suite, nous avons fait:

- D'abord, une conversion de ces données, si elles étaient de la forme "##:##", de rajouter "00:" devant pour mettre de la forme "h:min:s"
- Une conversion des données, si elles étaient de la forme "min:sec.1/100s", de les convertir en secondes
- Et puis, pour ensuite placer ces données dans une nouvelle variable nommé *temps_en_secondes*.

```
#FEMMES#
for (i in 1:nrow(resultats_femmes)) {
  mark <- resultats_femmes$Mark[i]
  if (grepl("^\\d{2}:\\d{2}$", mark)) {
    # Si la variable est composée de ##:##, ajouter :00 et copier
    resultats_femmes$Mark[i] <- paste0("00:",mark)
```

```

}
}
#HOMMES#
for (i in 1:nrow(resultats_hommes)) {
  mark <- resultats_hommes$Mark[i]
  if (grepl("^\\d{2}:\\d{2}$", mark)) {
    # Si la variable est composée de ##:##, ajouter :00 et copier
    resultats_hommes$Mark[i] <- paste0("00:", mark)
  }
}
convertir_temps <- function(temps) {
  if(grepl("\\.", temps)) { # Si le temps est en min:sec.1/100s
    temps_split <- strsplit(temps, "\\.")
    minutes <- as.numeric(substring(temps_split[[1]][1], 1, nchar(temps_split[[1]][1])-3))
    secondes <- as.numeric(substring(temps_split[[1]][1], nchar(temps_split[[1]][1])-1, nchar(temps_split[[1]][1])))
    centiemes <- as.numeric(temps_split[[1]][2])
    temps_en_secondes <- minutes * 60 + secondes + centiemes / 100
  } else { # Si le temps est en h:min:sec
    temps_split <- strsplit(temps, ":")
    heures <- as.numeric(temps_split[[1]][1])
    minutes <- as.numeric(temps_split[[1]][2])
    secondes <- as.numeric(temps_split[[1]][3])
    temps_en_secondes <- heures * 3600 + minutes * 60 + secondes
  }
}
# Appliquer la fonction a la colonne "marque" de votre dataframe
#FEMMES#
resultats_femmes$temps_en_secondes <- sapply(resultats_femmes$Mark, convertir_temps)

resultats_femmes$vitesse <- resultats_femmes$distance/resultats_femmes$temps_en_secondes

sum(is.na(resultats_femmes$temps_en_secondes))
## [1] 0

#HOMMES#
resultats_hommes$temps_en_secondes <- sapply(resultats_hommes$Mark, convertir_temps)
resultats_hommes$vitesse <- resultats_hommes$distance/resultats_hommes$temps_en_secondes

```

Ensuite, nous nous sommes focalisés sur la date de naissance des athlètes et la date à laquelle ils ont passé les disciplines, en changeant le format des dates pour ensuite calculer l'âge à laquelle ils ont passé les épreuves

```

# calcul de l'age le jour de la performance
resultats_femmes$age_compet=as.numeric(round((resultats_femmes$Date-resultats

```

```

_femmes$DOB)/365,2))
resultats_femmes$age_entier=as.numeric(round(resultats_femmes$age_compet,0))
#calcul de l'age entier

# calcul de l'age le jour de la performance
resultats_hommes$age_compet=as.numeric(round((resultats_hommes$Date-resultats
_hommes$DOB)/365,2))
resultats_hommes$age_entier=as.numeric(round(resultats_hommes$age_compet,0))

```

Ensuite, nous avons supprimé les données aberrantes, c'est-à-dire les personnes qui étaient âgés de plus de 100 ans.

```

#FEMMES
resultats_femmes <- subset(resultats_femmes, age_entier <=100)
resultats_femmes <- subset(resultats_femmes, age_compet <=100)

#HOMMES
resultats_hommes <- subset(resultats_hommes, age_entier <=100)
resultats_hommes <- subset(resultats_hommes, age_compet <=100)

```

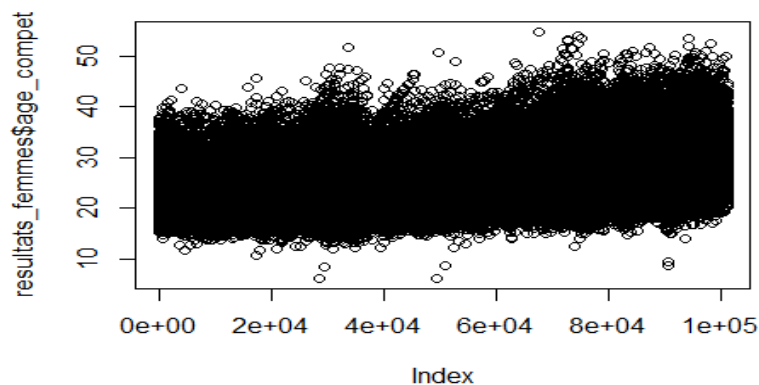
Comparaison Age et performance

Notre étude se penche sur la relation entre l'âge et la vitesse des athlètes, deux éléments clés dans l'évaluation de la performance sportive. Il est reconnu que l'âge joue un rôle significatif sur les performances physiques, avec un âge "idéal" où les athlètes sont au sommet de leurs capacités, suivi d'une phase de déclin. L'objectif de notre analyse est de vérifier si les données soutiennent cette idée d'une corrélation entre l'âge et la vitesse et d'identifier à quel âge les athlètes réalisent leurs meilleures performances. À travers l'utilisation de visualisations et de modèles statistiques, nous explorons une variété d'âges et de performances pour déterminer cette relation.

```

#FEMMES#
plot(resultats_femmes$age_compet)

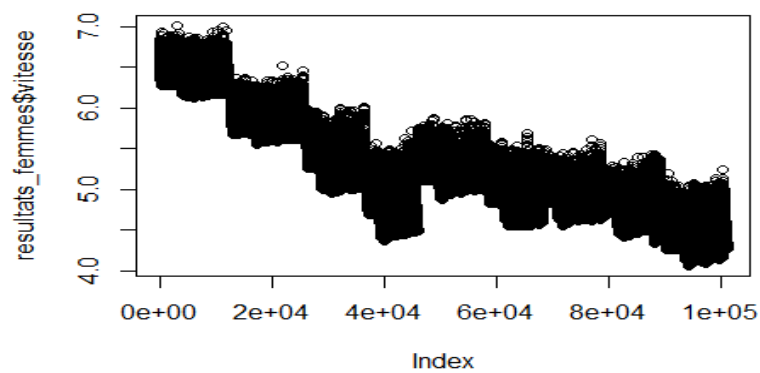
```



```

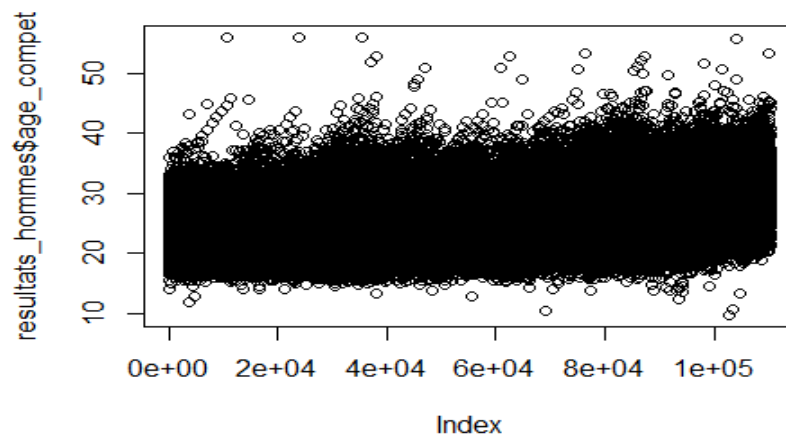
plot(resultats_femmes$vitesse)

```

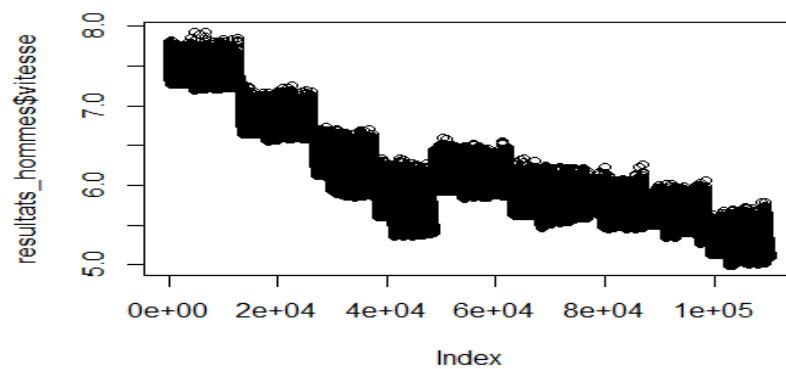


#HOMMES#

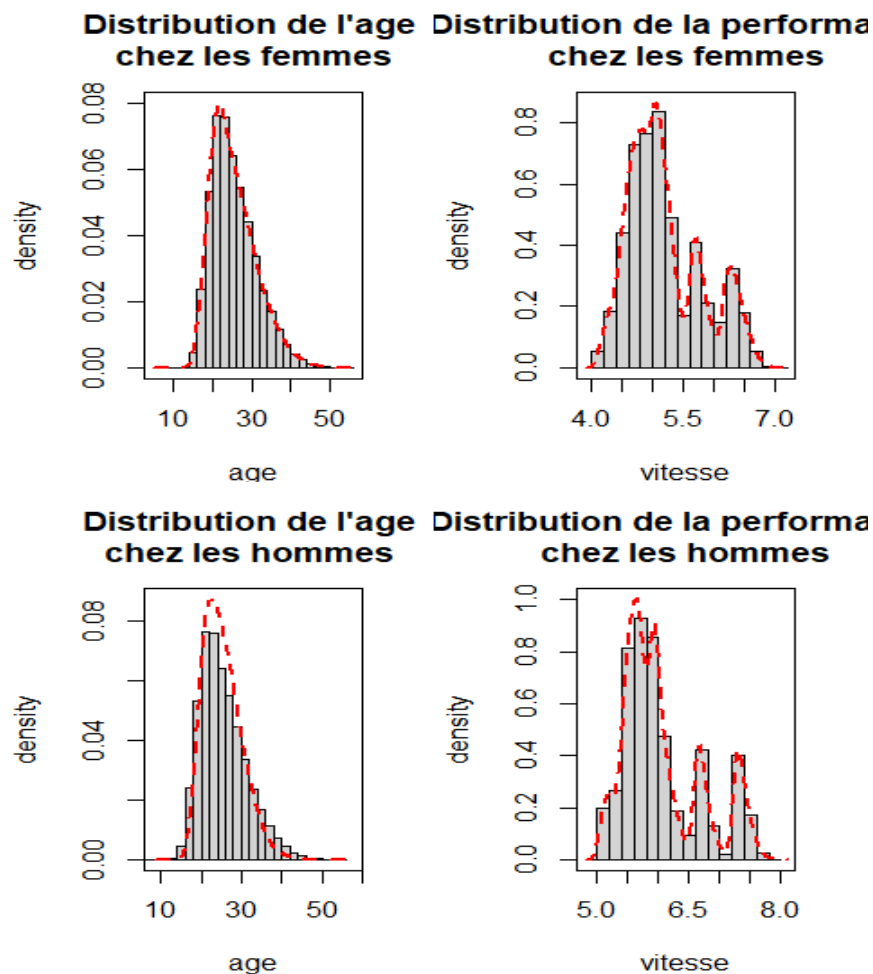
```
plot(resultats_hommes$age_compet)
```



```
plot(resultats_hommes$vitesse)
```



Nous avons d'abord observé la tendance de l'âge et la vitesse des athlètes sous la forme de nuage de points chez les hommes et les femmes.

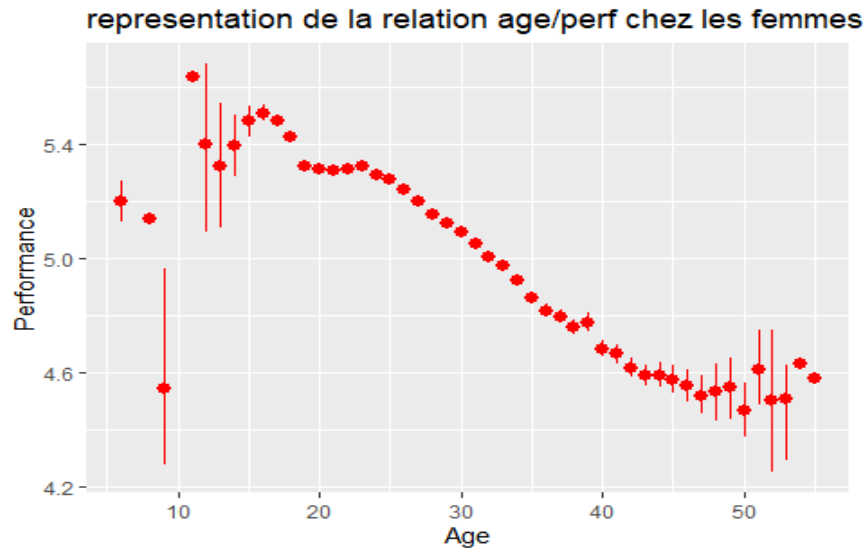


Notre étude a scruté la distribution de l'âge et de la performance parmi les compétiteurs des deux genres.

L'analyse révèle chez les femmes une concentration marquée des âges autour de la vingtaine, point après lequel la fréquence commence à décliner progressivement. Concernant leur performance, nous notons une densité de vitesse avoisinant les 5 m/s, densité qui tend à s'effriter à mesure que la vitesse s'accroît.

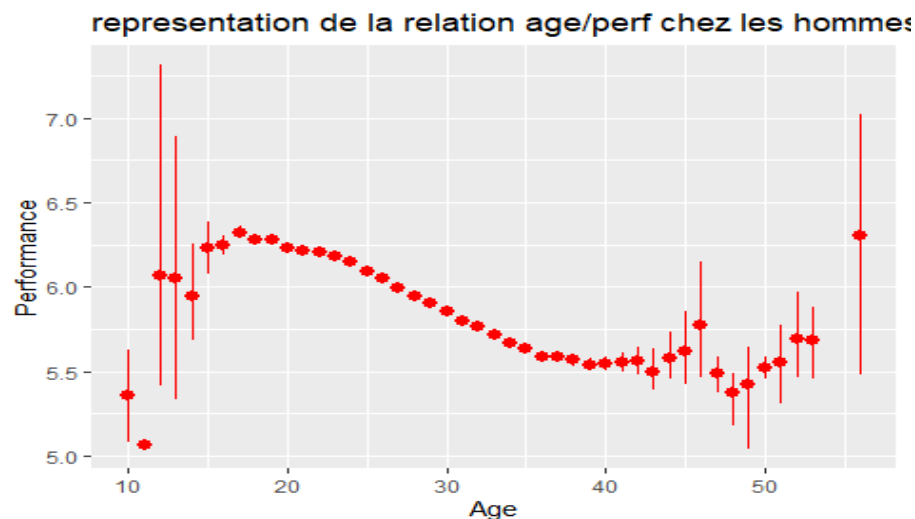
Chez les hommes, l'analyse révèle une concentration marquée des âges autour de la vingtaine, point après lequel la fréquence commence à décliner progressivement. Concernant leur performance, nous notons une densité de vitesse avoisinant les 5.5 m/s, densité qui tend à s'effriter à mesure que la vitesse s'accroît.

Par conséquent, nous avons pu analyser la performance des athlètes en fonction de leur âge chez les hommes et les femmes.



Ce graphique met en évidence l'impact de l'âge sur la performance sportive féminine, associant l'âge des athlètes (axe horizontal) à leur performance (axe vertical), exprimée en vitesse. Les points rouges traduisent la performance moyenne à chaque âge, tandis que les barres verticales rouges représentent l'intervalle de confiance, illustrant la variabilité des performances pour un âge donné.

On observe une tendance générale à la baisse de la performance avec l'âge, commençant par des performances optimales dans la vingtaine qui déclinent ensuite progressivement. La concentration des points et l'amplitude des intervalles de confiance tendent à réduire avec l'avancement en âge, suggérant une possible diminution des données disponibles ou une variabilité accrue des performances parmi les athlètes plus âgées.



Le graphique montre la relation entre l'âge et la performance, chez les hommes. Sur l'axe horizontal, l'âge est affiché, et sur l'axe vertical, nous avons la performance sportive.

Les points rouges représentent la moyenne des performances à différents âges, et les barres d'erreur indiquent la variabilité ou l'intervalle de confiance autour de ces moyennes. On

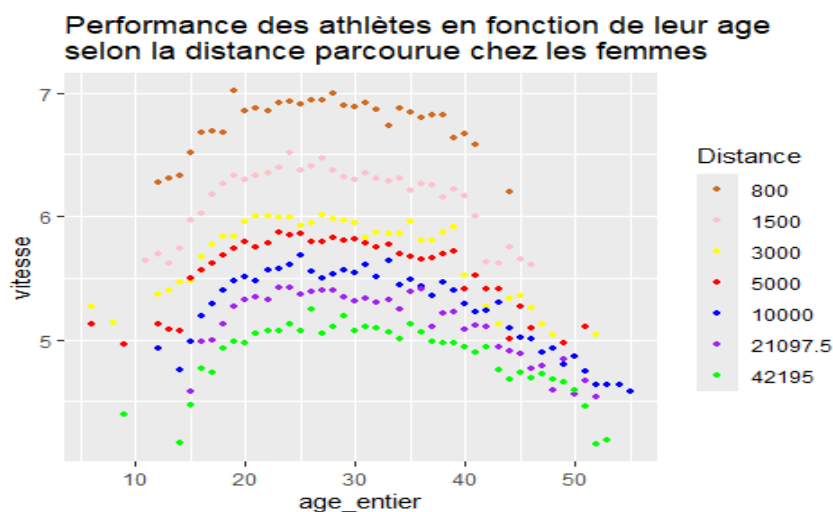
observe une tendance générale décroissante : les performances commencent relativement élevées chez les plus jeunes, atteignent un plateau pendant la vingtaine et la trentaine, puis semblent diminuer à partir de la quarantaine.

Les barres d'erreur sont particulièrement longues pour les tranches d'âge les plus jeunes et les plus âgées, ce qui pourrait indiquer une plus grande variabilité des performances dans ces groupes ou un nombre moins élevé de données, rendant l'estimation moins précise. Cela suggère que la performance sportive masculine pourrait être optimale durant les années intermédiaires et que l'âge pourrait être un facteur influençant cette performance.

La meilleure performance des athlètes en fonction de leur âge selon la distance parcourue

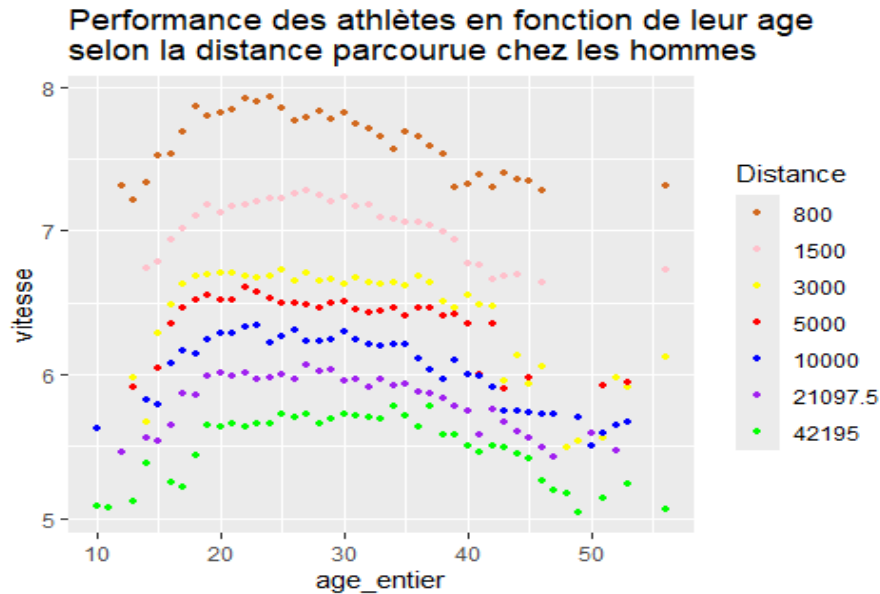
Notre analyse s'est concentrée sur l'évaluation de la performance maximale des athlètes en fonction de leur âge, et ce pour différentes distances parcourues, chez les deux sexes. Nous avons examiné comment la vitesse maximale varie avec l'âge pour chaque sexe, en prenant en compte des distances variées allant du 800m au marathon complet. Cela nous a permis de déterminer des modèles de performance qui révèlent à quel âge les athlètes sont susceptibles d'atteindre leur pic de vitesse, et comment ce pic varie selon la longueur de la course.

Chez les Femmes



Ce graphique met en lumière l'impact de l'âge sur la vitesse chez les athlètes féminines sur des épreuves allant du sprint au marathon. Chaque point représente la vitesse moyenne pour une tranche d'âge et une distance spécifique, avec des couleurs variées indiquant différentes longueurs de course. Les tendances observées indiquent une baisse générale de la vitesse avec l'âge et une performance optimale souvent atteinte dans la vingtaine ou la trentaine. Les distances plus courtes comme le 800 et le 1500 mètres montrent des vitesses plus rapides qui commencent à décroître plus tôt, tandis que les épreuves plus longues comme le semi-marathon et le marathon révèlent une plus grande variabilité et une capacité à maintenir des performances stables sur un éventail d'âges plus large. En conclusion, le graphique illustre un déclin progressif de la performance avec l'augmentation de la distance et de l'âge, mais avec une persistance remarquable de la compétence sur les longues distances.

Chez les hommes



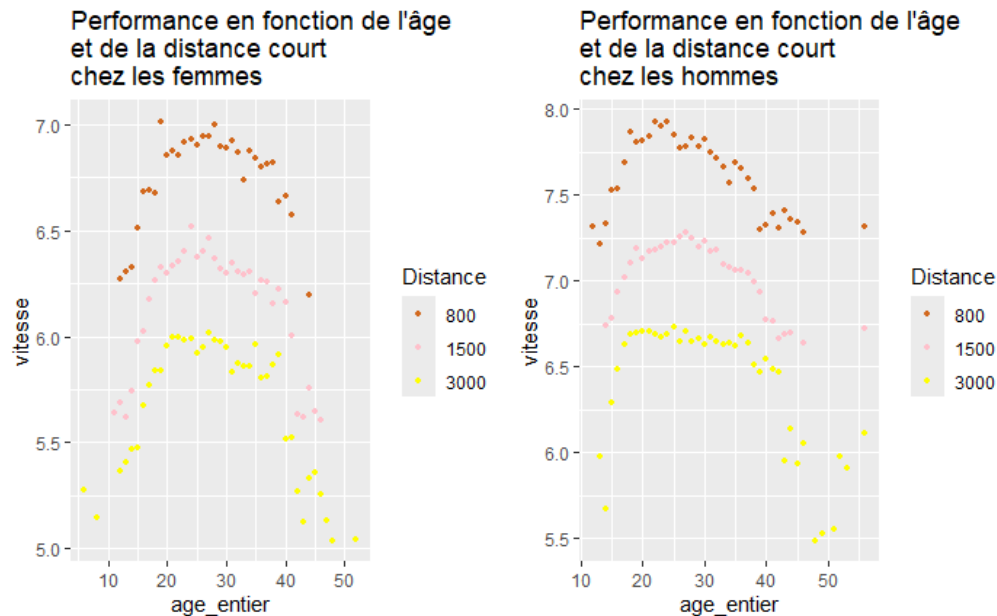
Ce graphique révèle la relation entre l'âge et la vitesse des coureurs masculins sur diverses distances de course. Les groupes de points démontrent les vitesses moyennes pour chaque distance et groupe d'âge, la couleur des points distinguant chaque épreuve.

Les points marron indiquent que les athlètes atteignent des vitesses maximales sur des distances courtes comme le 800 mètres, avec une baisse notable après la vingtaine. À mesure que la distance augmente, représentée par des couleurs changeantes, les vitesses moyennes baissent, une tendance attendue avec l'augmentation de la distance.

Une diminution de la vitesse avec l'avancée en âge est visible pour toutes les distances. Cependant, cette tendance est plus douce pour les marathons, représentés par des points vert foncé et violets, suggérant une capacité à maintenir une performance plus stable avec l'âge sur les longues distances.

La dispersion des points illustre une performance variée qui implique que l'impact de l'âge sur la vitesse diffère selon la distance. Ainsi, ce graphique offre une vue d'ensemble sur l'impact de l'âge sur la vitesse masculine, des sprints aux courses d'endurance.

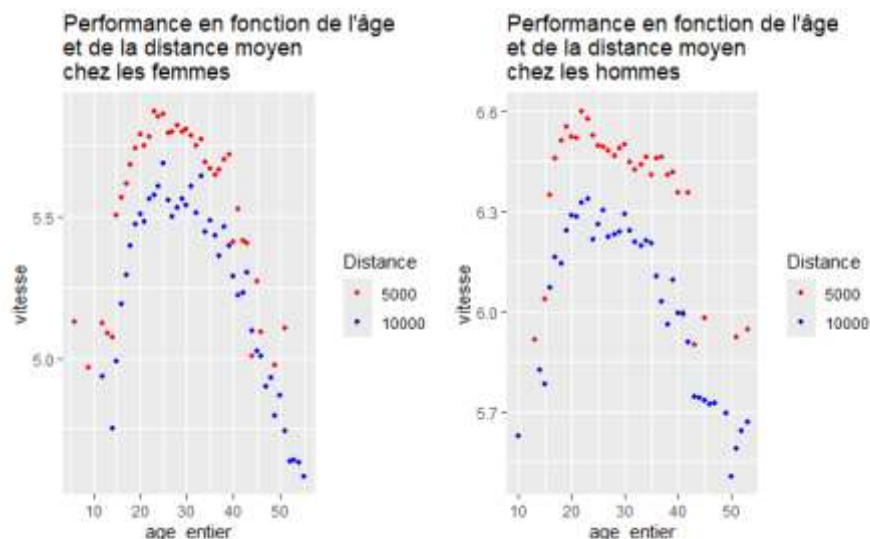
Comparaison entre les hommes et les femmes



Ces graphiques comparent la performance athlétique par âge et par distance, différenciant les résultats entre les hommes et les femmes pour les courses considérées comme courtes. Cette comparaison apporte un éclairage sur les différences de performance liées au sexe et à la distance, nous permettant de décrypter la relation entre l'âge et la performance en athlétisme.

Le graphique des femmes montre des vitesses de pointe atteintes principalement dans la vingtaine pour les 800 mètres, tandis que pour les 1500 et 3000 mètres, la performance décroît plus rapidement avec l'âge. En revanche, le graphique des hommes indique que les performances pour les 800 mètres sont généralement supérieures à celles des femmes, mais avec une régression plus marquée après le pic.

En bref, ces visualisations ne se contentent pas d'illustrer comment la performance varie avec l'âge, mais elles mettent également en avant des variations entre les hommes et les femmes sur des distances similaires, soulignant l'importance de prendre en compte le sexe et la distance dans l'évaluation des capacités athlétiques.

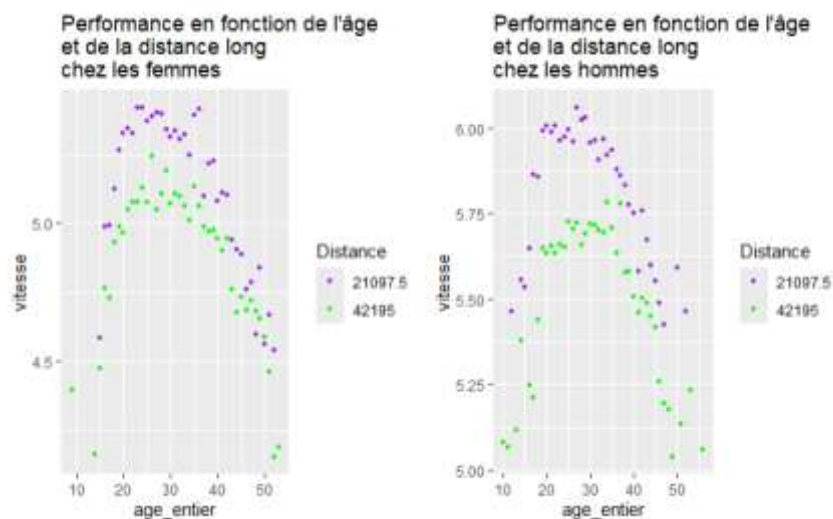


Ces graphiques portent un regard sur les performances athlétiques liées à l'âge dans les épreuves de moyenne distance, le 5000 mètres et le 10000 mètres, qui sont plus axées sur l'endurance que sur la vitesse brute. Ces distances révèlent comment endurance et performance évoluent chez les athlètes au fil des années.

Pour les femmes, on constate que les performances pour les 5000 mètres et les 10000 mètres, marquées respectivement en rouge et en bleu, atteignent leur apogée entre la vingtaine et la trentaine.

Du côté des hommes, la performance maximale est également observée dans la jeunesse, avec une baisse graduelle avec l'âge. La dispersion des points pour les 10000 mètres est plus faible que chez les femmes, ce qui pourrait signifier une moindre variabilité de performance ou une fréquence plus basse de participation à cette distance.

Pour les deux sexes, ces analyses mettent en évidence un schéma commun : les meilleures performances dans les courses de moyenne distance sont généralement atteintes plus tôt dans la carrière des athlètes, suivies d'un déclin plus ou moins prononcé après la trentaine.



Les graphiques que nous examinons maintenant illustrent comment l'âge affecte la vitesse sur les longues distances du semi-marathon et du marathon pour les athlètes des deux sexes. L'âge des coureurs est placé sur l'axe horizontal, tandis que leur vitesse apparaît sur l'axe vertical.

Les couleurs des points sur les graphiques sont utilisées pour différencier les distances : le violet pour le semi-marathon de 21,097.5 mètres et le vert pour le marathon de 42,195 mètres. Ces couleurs aident à distinguer clairement les performances pour chacune de ces épreuves d'endurance.

Ces graphiques offrent une vue d'ensemble des tendances de performance selon l'âge pour ces longues distances, mettant en évidence les périodes de pic de performance, qui varient selon le sexe des coureurs et la distance parcourue.

Question 2

Comparaison des modèles sur la performance des joueurs en fonction de leur âge

Une fonction a d'abord été créée, la fonction des **Moyennes des Moindres au Carré**, (*MMC*) afin d'établir un modèle choisi entre la performance en fonction de l'âge sur les observations et ainsi observer sa tendance sur les différents parcours que les athlètes effectuent chez les hommes et les femmes.

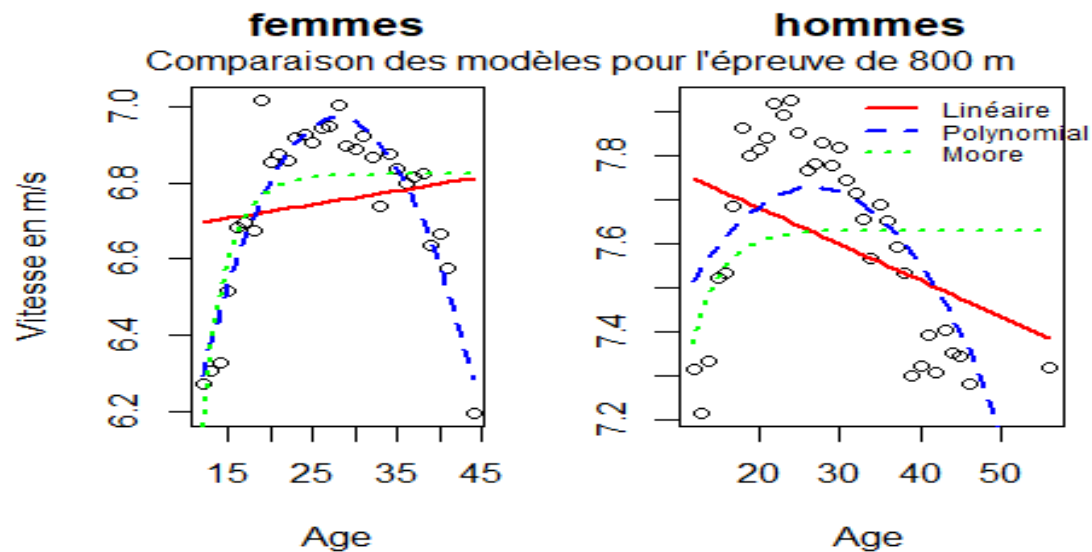
Nous avons analysé différents modèles :

- Le **Modèle Linéaire**
- Le **Modèle Polynomial**
- Le **Modèle de Moore**

```
# Définition de la fonction de modèle linéaire
modele_lineaire <- function(age, parametre) {
  return(parametre[1] * age + parametre[2]) # Modèle linéaire :  $y = a \cdot x + b$ 
}

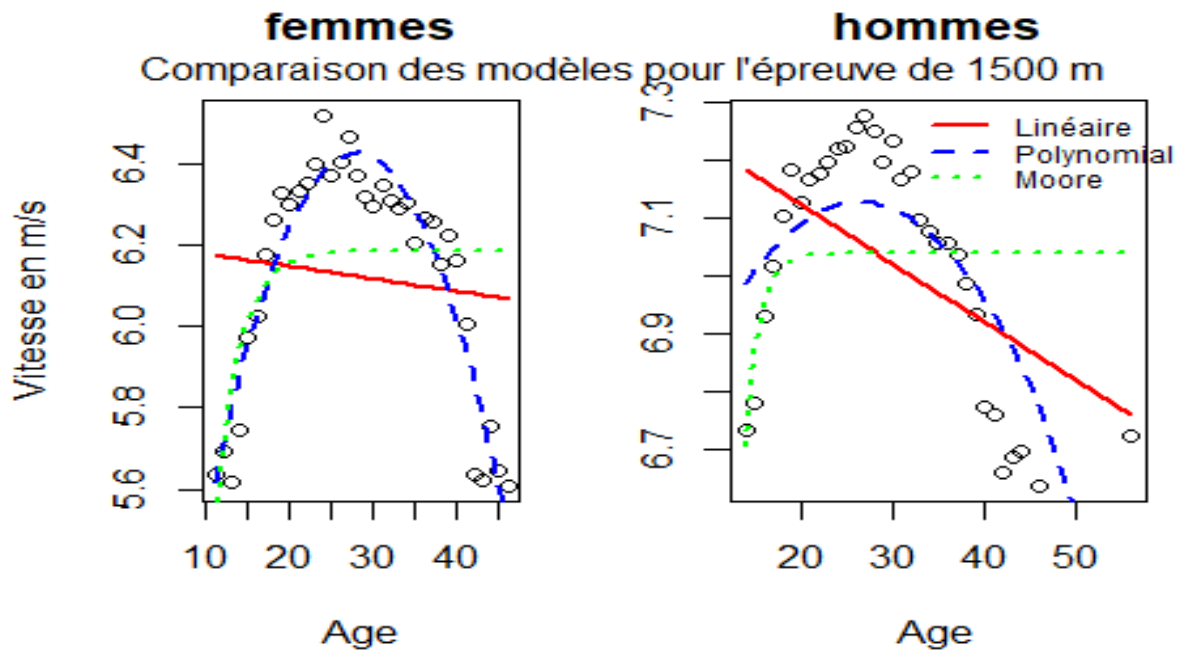
# Définition de la fonction de modèle polynomiale
modele_polynomial <- function(age, parametre) {
  return(parametre[1] + parametre[2] * age + parametre[3] * age^2) # Modèle polynomial d'ordre 2
}

# Définition de la fonction de modèle de Moore
modele_moore <- function(x,p){
  p[1]*(1-exp(-p[2]*x)) + p[3]*(1-exp(p[4]*x))
}
```



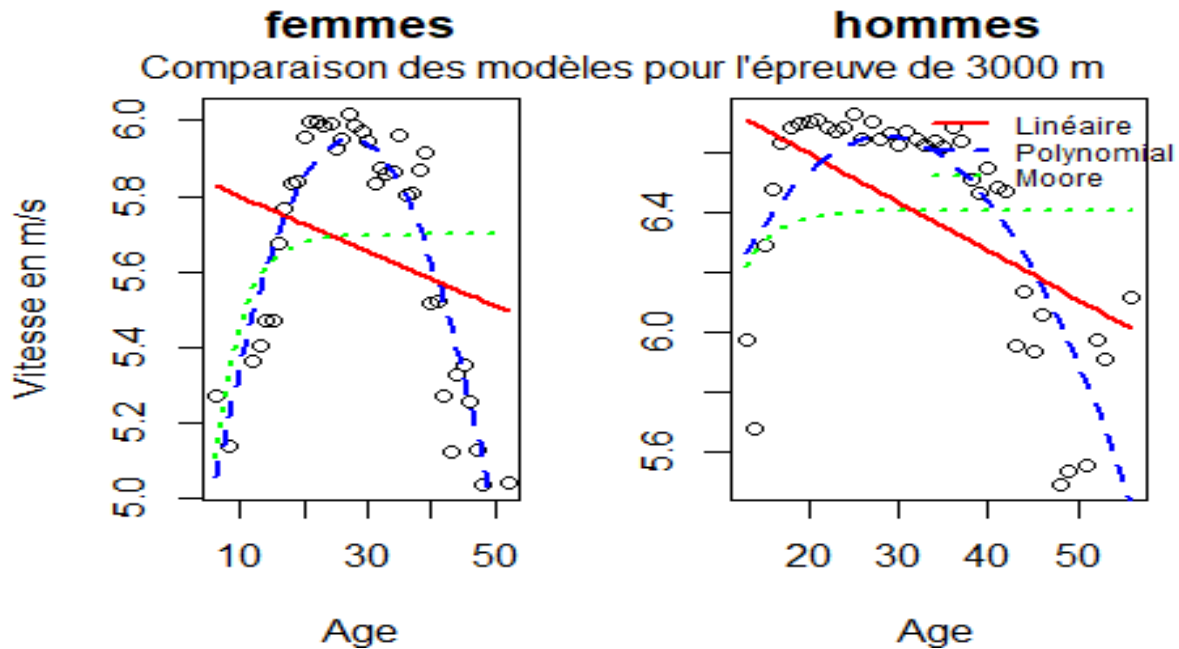
Sur 800m :

Les graphiques illustrent la performance athlétique sur la distance de 800 mètres en fonction de l'âge, avec une distinction entre les données pour les femmes et les hommes. On observe que pour les deux genres, le modèle polynomial, illustré par la ligne bleue en pointillés, épouse le mieux la distribution des points individuels, suggérant une performance qui augmente jusqu'à un certain âge avant de décliner. Cela contraste avec le modèle linéaire, représenté par la ligne rouge, qui montre une tendance constamment décroissante pour les hommes et croissante pour les femmes, ne capturant pas l'apogée de performance visible dans les données. Le modèle de Moore, en ligne verte pointillée, semble également moins adapté, en particulier pour les données masculines où il surévalue la performance aux âges extrêmes. Globalement, le modèle polynomial semble offrir la représentation la plus fidèle de l'effet de l'âge sur la performance sur 800 mètres, révélant un sommet de performance dans la tranche d'âge des 20-30 ans pour les deux genres avant une diminution avec l'avancement en âge.



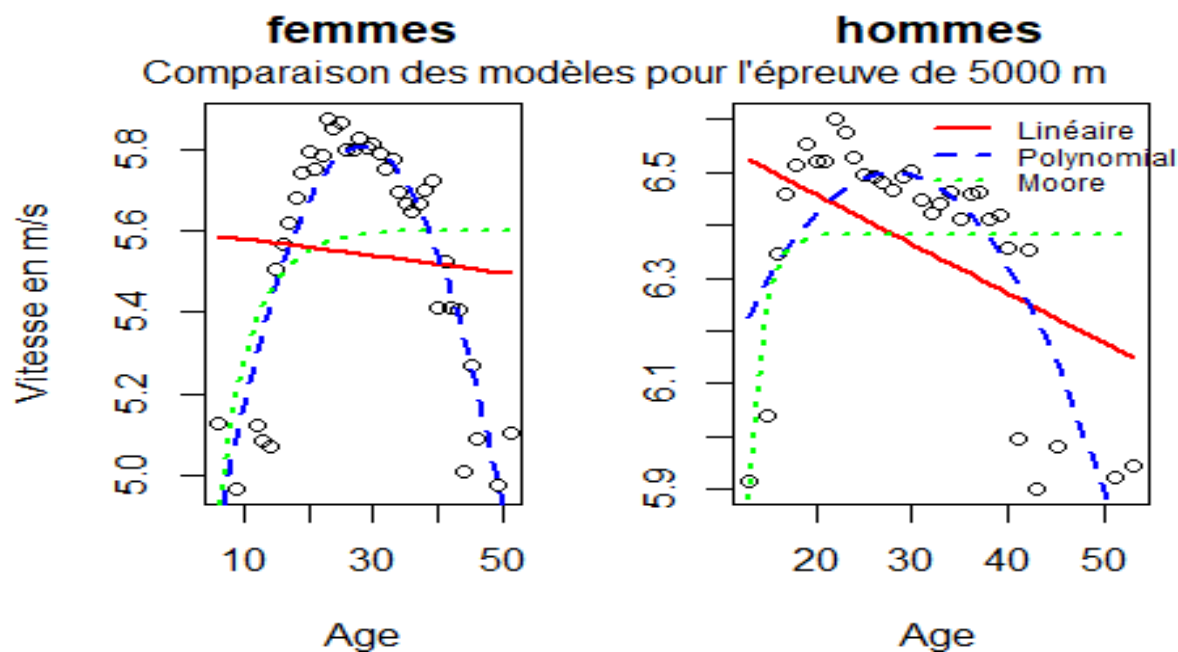
Sur 1500m :

Dans ces graphiques, nous comparons les performances sur 1500 mètres pour les femmes et les hommes à travers trois modèles différents. Pour les femmes, le modèle polynomial semble mieux suivre la tendance des données, atteignant un pic avant de diminuer avec l'âge, tandis que le modèle linéaire suggère une performance stable et le modèle de Moore montre une surévaluation de la performance pour les sportifs âgées. Pour les hommes, le modèle polynomial indique également une hausse et une baisse de la performance avec l'âge, mais la tendance décroissante est plus prononcée dans le modèle linéaire. Le modèle de Moore, bien qu'il suive une courbe similaire au modèle polynomial, semble moins précis, en particulier dans les tranches d'âge moyennes et plus élevées. Ces observations suggèrent que le modèle polynomial capte mieux les variations de performance en fonction de l'âge pour les épreuves de 1500 mètres.



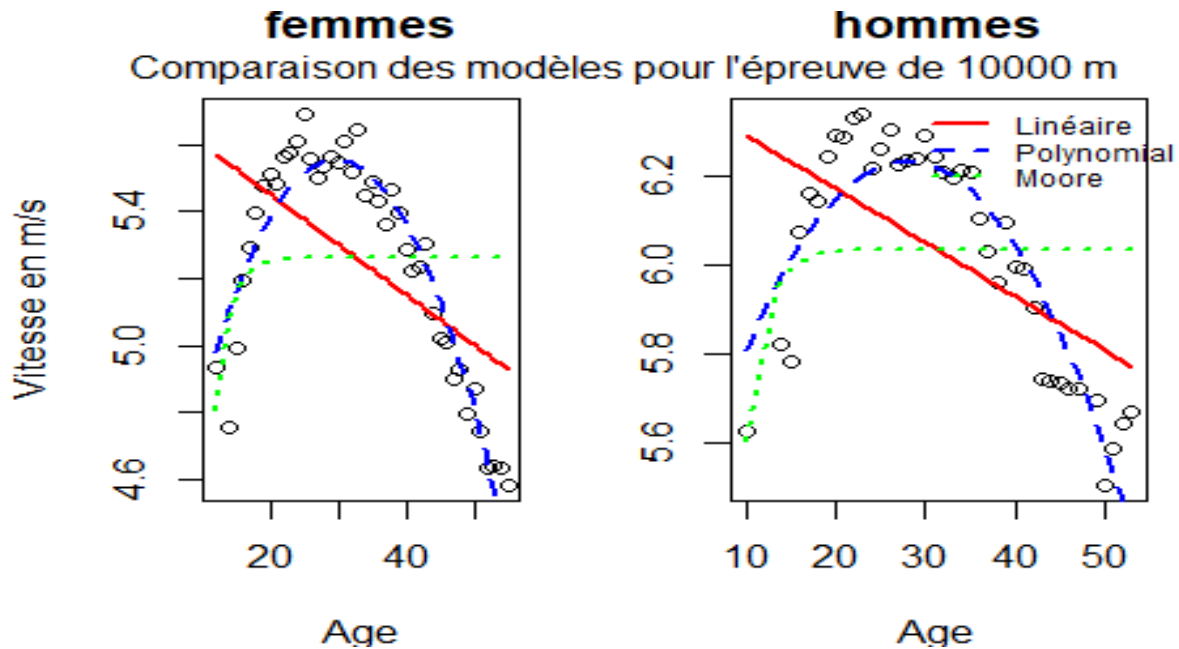
Sur 3000m :

Les graphiques présentés illustrent la comparaison des performances sur la distance de 3000 mètres pour les femmes et les hommes en utilisant trois modèles statistiques différents : linéaire, polynomial et Moore. Chez les femmes, le modèle polynomial suit une courbe qui atteint un sommet avant de redescendre, indiquant une augmentation puis une diminution de la vitesse avec l'âge. Le modèle linéaire suggère une légère dégradation constante de la performance, alors que le modèle de Moore s'aplatit avec l'âge. Pour les hommes, le modèle polynomial démontre une diminution plus prononcée après le pic de performance comparé au modèle linéaire qui affiche une baisse constante sur toute la gamme d'âge. Ces tendances suggèrent que le modèle polynomial est plus à même de capturer la réalité de la variation de performance en fonction de l'âge pour les épreuves de 3000 mètres.



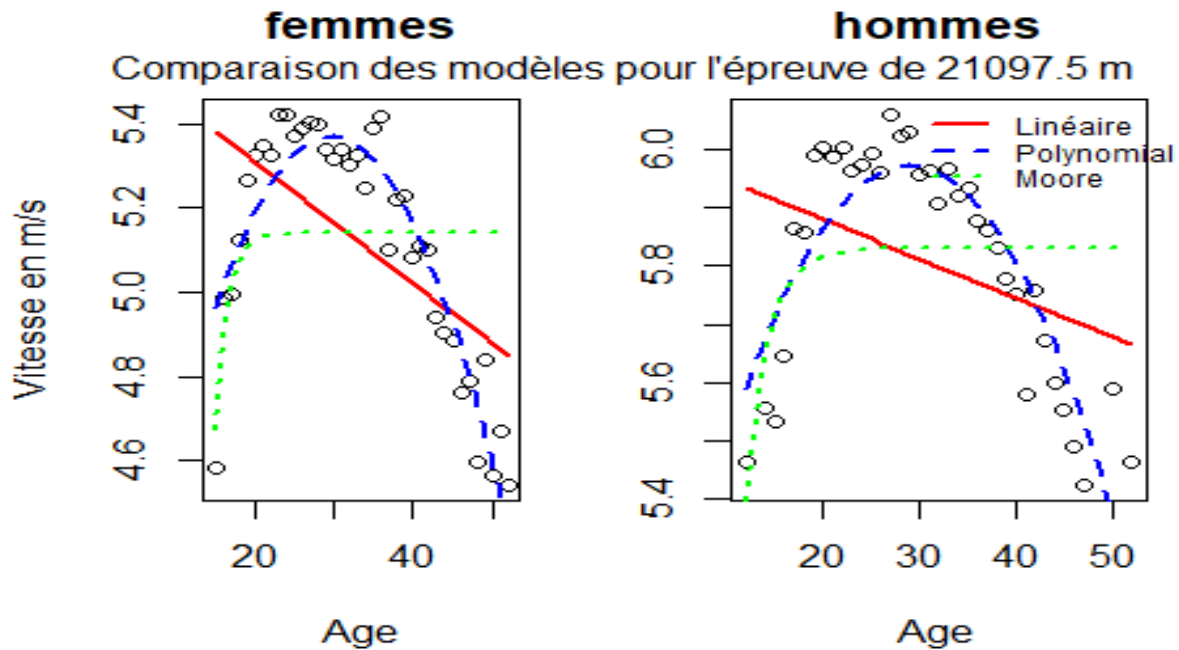
Sur 5000m :

Ces graphiques comparent les performances sur la distance de 5000 mètres pour les catégories féminines et masculines à travers différents modèles statistiques. Chez les femmes, la courbe du modèle polynomial montre un pic de performance avant de décliner, traduisant une hausse puis une baisse de la vitesse avec l'âge. Le modèle linéaire indique une décroissance relativement constante de la vitesse, tandis que le modèle de Moore semble aplatir la performance dans la deuxième partie de la courbe d'âge avant de décliner fortement. Chez les hommes, le modèle polynomial capture également le pic de performance avant de montrer un déclin, en harmonie avec les données observées. Le modèle linéaire présente une pente décroissante constante, et le modèle de Moore affiche une stabilité après une forte augmentation. Comme pour les autres distances, le modèle polynomial paraît le plus adapté pour décrire la variation de performance avec l'âge sur 5000 mètres, car il suit de plus près la tendance des données réelles observées.



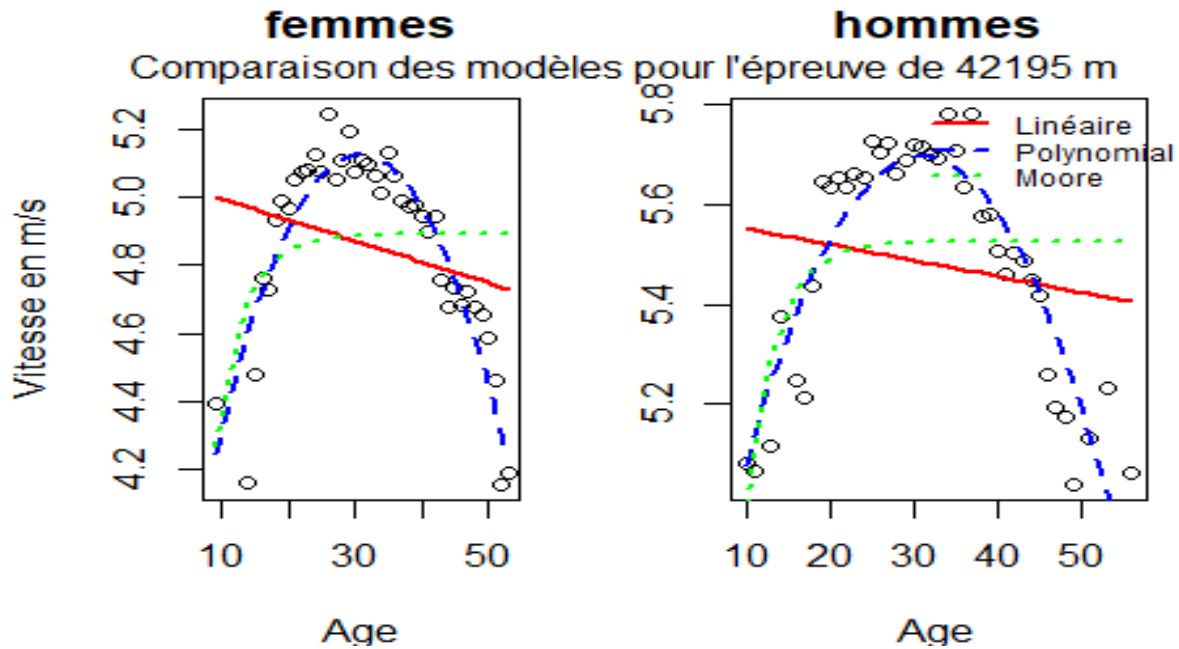
Sur 10000m :

Sur ce graphique, nous observons les performances sur la distance de 10000 mètres, comparées pour les femmes et les hommes à travers différents modèles. Chez les femmes, la courbe du modèle polynomial culmine avant de chuter, reflétant une augmentation puis une diminution de la vitesse avec l'âge. Le modèle linéaire suggère une baisse progressive de la vitesse, tandis que le modèle de Moore surévalue les performances dans la tranche d'âge plus vieille. Chez les hommes, les résultats sont similaires avec le modèle polynomial dépeignant un pic de performance avant de descendre, ce qui semble correspondre aux données observées. Le modèle linéaire décrit une décroissance constante et le modèle de Moore indique une stabilité. Le modèle polynomial semble être le plus adéquat pour suivre la tendance des données réelles sur cette distance, démontrant sa capacité à représenter la variation de performance en fonction de l'âge.



Sur 21097.5m :

Le graphique illustre la comparaison des modèles de performance sur une distance de 21097.5 mètres (semi-marathon) pour les femmes et les hommes. Pour les deux genres, le modèle polynomial montre une performance qui augmente avec l'âge jusqu'à un certain point avant de décliner, ce qui suggère une performance optimale à un âge moyen spécifique. Le modèle linéaire indique une baisse régulière de la performance avec l'âge. Le modèle de Moore présente un aplatissement des performances pour les tranches d'âge plus vieille. Cette visualisation des données suggère que le modèle polynomial capture le mieux la nature des performances athlétiques sur des distances semi-marathon, en reflétant un pic de performance avant une diminution attendue avec l'âge avancé.



Sur 42195m :

Sur ces graphiques, nous comparons les performances des athlètes sur la distance marathon (42195 mètres) à travers trois modèles différents. Pour les femmes comme pour les hommes, le modèle polynomial suit une courbe qui monte jusqu'à un point culminant avant de redescendre, suggérant un âge de performance optimale. Le modèle linéaire montre une dégradation constante de la performance avec l'âge. Quant au modèle de Moore, il présente une performance initialement surévaluée pour les plus jeunes avant de s'aplatir, ce qui semble moins représentatif des données observées. Ces tendances mettent en évidence que le modèle polynomial est probablement le plus adapté pour refléter la réalité des performances sur un marathon, avec un pic de performance avant un déclin lié à l'âge.

L'analyse du meilleur modèle

Nous avons approfondi l'analyse de la relation entre l'âge et la performance. L'accent est mis sur l'ajustement de trois modèles mathématiques - linéaire, polynomial et celui de Moore - à des ensembles de données sur des performances sportives. Nous utilisons ici des méthodes statistiques pour déterminer quel modèle représente le mieux les données. Cet ajustement des modèles est crucial pour comprendre les tendances dans les performances liées à l'âge et potentiellement pour prédire les performances futures. Le calcul des résidus, le coefficient de détermination ajusté (R^2 ajusté), l'AICc et le BIC nous aideront à évaluer la précision des modèles. Cela mettra en évidence l'importance d'une sélection rigoureuse du modèle pour interpréter correctement les données et tirer des conclusions fiables sur la relation âge-performance.

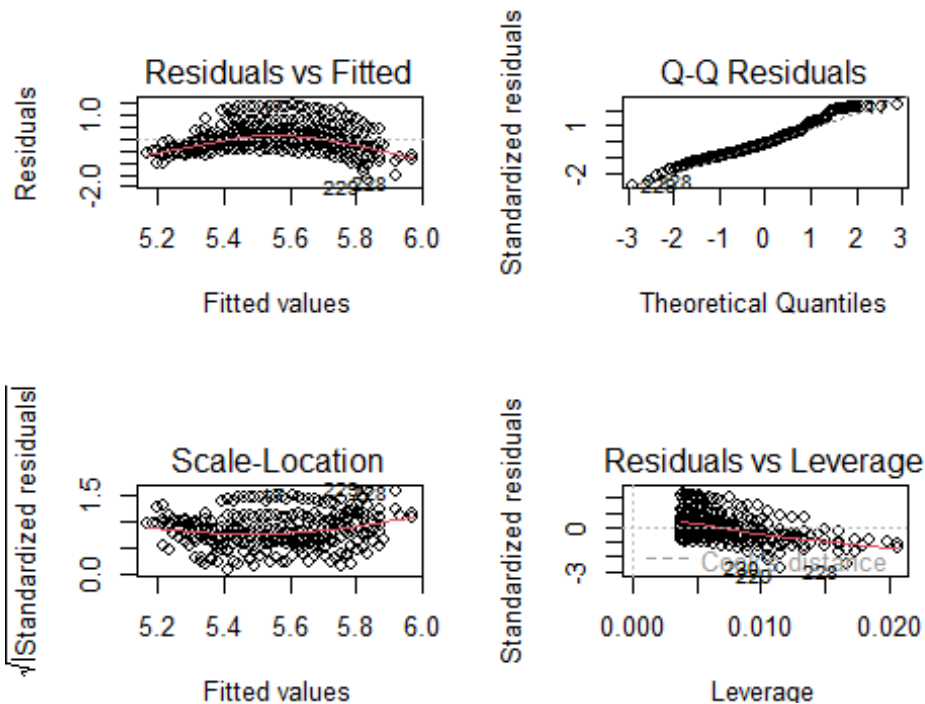
Dans cette étape, nous allons nous pencher sur l'examen des résidus des différents modèles pour les catégories hommes et femmes. L'objectif est de déterminer si les modèles sélectionnés reflètent adéquatement les données réelles. Pour ce faire, l'analyse de la normalité des résidus sera notre point de départ ; cela nous permettra de comprendre si les écarts entre les performances prédites par nos modèles et les performances réelles observées suivent une distribution normale. Cette normalité est un indicateur clé de la fiabilité de nos modèles et de leur pertinence pour représenter la relation entre l'âge et la performance athlétique. Si les

résidus sont normalement distribués, cela renforce notre confiance dans les modèles utilisés et justifie leur application pour les interprétations ultérieures.

Question 3

Normalité Modèle linéaire

```
par(mfrow=c(2,2))  
plot(lm(meilleures_performances_femmes$vitesse~ meilleures_performances_femmes$age_entier))
```



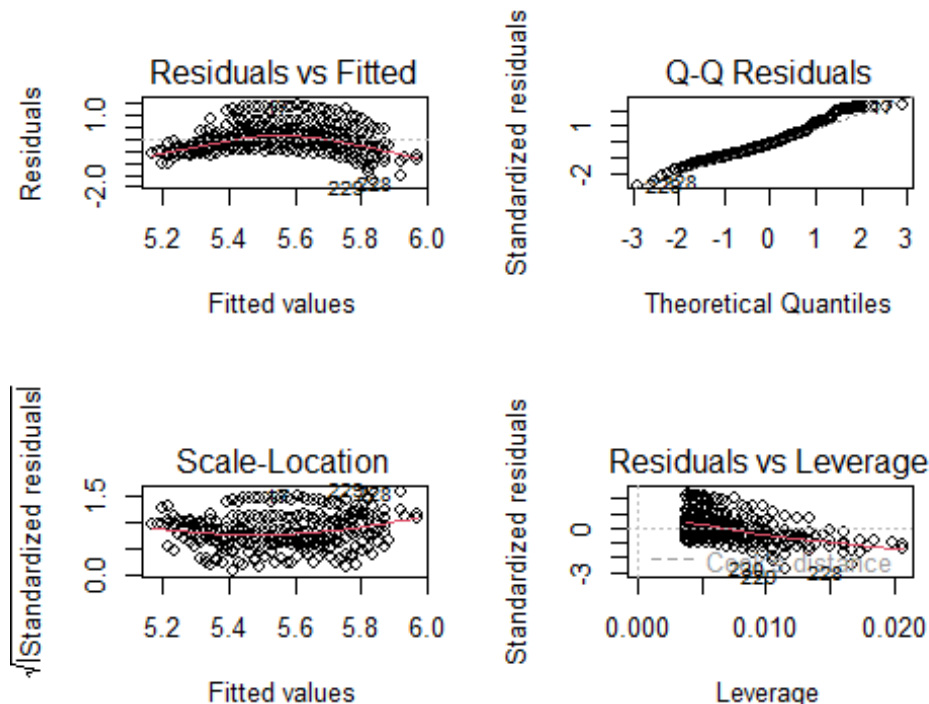
Ces graphiques fournissent une évaluation complète du modèle linéaire à travers différentes perspectives :

- **Résidus vs Valeurs Ajustées** : Ce graphique révèle l'existence de motifs ou de tendances non capturés par le modèle, indiquant une possible inadéquation entre le modèle et les données.
- **Graphique Q-Q des Résidus** : Il compare la distribution des résidus avec une distribution normale théorique, permettant d'évaluer si les erreurs suivent une distribution normale, une hypothèse clé de la régression linéaire.
- **Échelle-Localisation (ou Spread-Location)** : Ce graphique vérifie l'homoscédasticité, c'est-à-dire si la dispersion des résidus est constante à travers les valeurs ajustées, une autre hypothèse importante de la régression linéaire.
- **Résidus vs Levier (ou Influence)** : Il identifie les observations ayant une influence disproportionnée sur l'ajustement du modèle, mettant en lumière les points de données potentiellement problématiques.

L'analyse combinée de ces graphiques est cruciale pour confirmer la validité du modèle linéaire et la fiabilité de ses prédictions, en s'assurant que les hypothèses fondamentales de la régression linéaire sont satisfaites.

Normalité Modèle polynomiale

```
par(mfrow=c(2,2))
plot(lm(meilleures_performances_femmes$vitesse~ I(meilleures_performances_femmes$age_entier)^2))
```



Les quatre graphiques de diagnostic examinent la validité d'un modèle polynomiale :

- **Résidus vs Ajustements** : montre si les erreurs de prédiction se dispersent de manière aléatoire, ce qui est attendu d'un bon modèle.
- **Q-Q des résidus** : vérifie si les résidus suivent une distribution normale, une hypothèse clé de la régression.
- **Échelle-Localisation** : évalue si la variance des erreurs est constante (homoscédasticité), autre signe d'un ajustement approprié.
- **Résidus vs Levier** : identifie les points ayant une influence disproportionnée sur le modèle, qui pourraient fausser les résultats.

Ensemble, ils fournissent une image complète du modèle polynomial pour les données analysées.

Calcul des différents critères

Après avoir vérifié les autres modèles, nous avons étudié l'AIC, le BIC, le R2 ajusté et la somme des résidus de chaque modèle afin de voir quel est le meilleur modèle.

```
# Fonction pour calculer et retourner un dataframe de résultats
compute_results <- function(data, linear_result, polynomial_result, moore_res
ult, distance, gender) {
  n <- length(data$age_entier) # Nombre d'observations

  # Calculs pour Le modèle linéaire
  RSS_linear <- sum(residuals(lm(data$vitesse ~ data$age_entier))^2)
  k_linear <- 2 # Nombre de paramètres pour Le modèle linéaire
  R2_linear <- linear_result$R2
  R2_ajuste_linear <- 1 - ((1 - R2_linear) * (n - 1)) / (n - k_linear - 1)
  AIC_linear <- n * log(RSS_linear/n) + 2 * k_linear + (2*k_linear*(k_linear
+ 1)) / (n - k_linear - 1)
  BIC_linear <- n * log(RSS_linear/n) + k_linear * log(n)

  # Calculs pour Le modèle polynomial
  RSS_poly <- sum(residuals(lm(data$vitesse ~ I(data$age_entier)^2))^2)
  k_poly <- 3 # Nombre de paramètres pour Le modèle polynomial
  R2_poly <- polynomial_result$R2
  R2_ajuste_poly <- 1 - ((1 - R2_poly) * (n - 1)) / (n - k_poly - 1)
  AIC_poly <- n * log(RSS_poly/n) + 2 * k_poly + (2*k_poly*(k_poly + 1)) / (n
- k_poly - 1)
  BIC_poly <- n * log(RSS_poly/n) + k_poly * log(n)

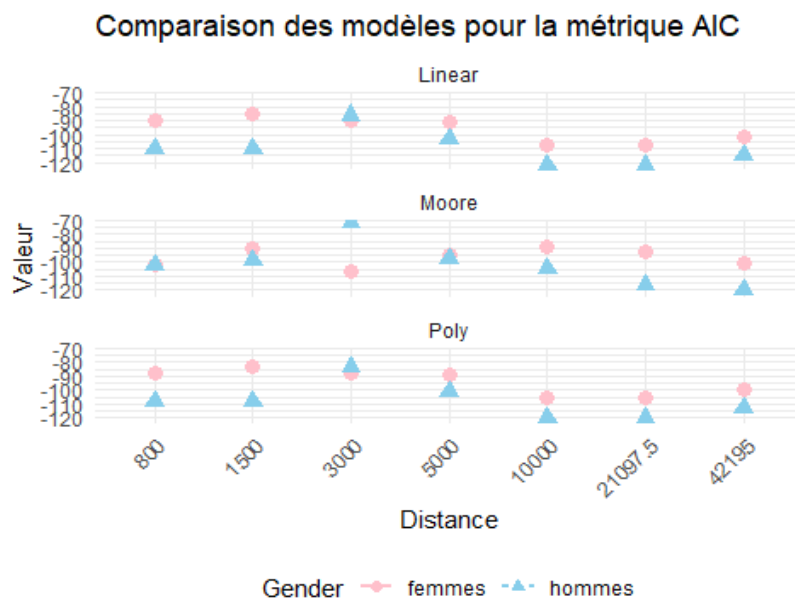
  # Calculs pour Le modèle de Moore
  RSS_moore <- moore_result$objectif
  k_moore <- 4 # Nombre de paramètres pour Le modèle de Moore
  R2_moore <- moore_result$R2
  R2_ajuste_moore <- 1 - ((1 - R2_moore) * (n - 1)) / (n - k_moore - 1)
  AIC_moore <- n * log(RSS_moore/n) + 2 * k_moore + (2*k_moore*(k_moore + 1))
/ (n - k_moore - 1)
  BIC_moore <- n * log(RSS_moore/n) + k_moore * log(n)

  # Création du dataframe de résultats
  results <- data.frame(
```

```

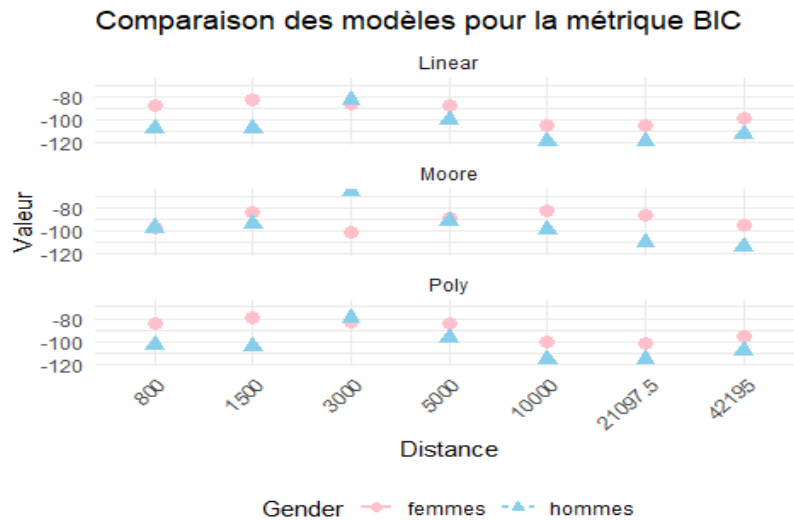
Distance = distance,
Gender = gender,
RSS_Linear = RSS_linear,
R2_Linear = R2_ajuste_linear,
AIC_Linear = AIC_linear,
BIC_Linear = BIC_linear,
RSS_Poly = RSS_poly,
R2_Poly = R2_ajuste_poly,
AIC_Poly = AIC_poly,
BIC_Poly = BIC_poly,
RSS_Moore = RSS_moore,
R2_Moore = R2_ajuste_moore,
AIC_Moore = AIC_moore,
BIC_Moore = BIC_moore
)
return(results)
}

```



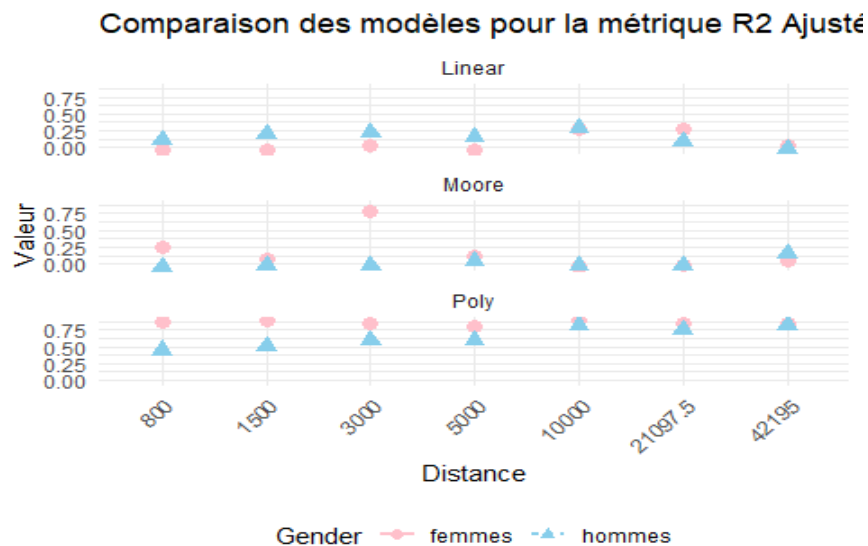
La méthode AIC :

Le graphique met en lumière la qualité de trois modèles statistiques pour différentes distances de course, en se basant sur le Critère d'Information d'Akaike corrigé (AICc). Des valeurs AICc inférieures signalent un meilleur modèle. Les résultats, distingués par sexe et modèle, montrent que le modèle polynomial surpasse souvent les autres, avec des AICc plus bas pour presque toutes les distances, ce qui indique qu'il s'adapte mieux aux données de performance athlétique liées à l'âge.



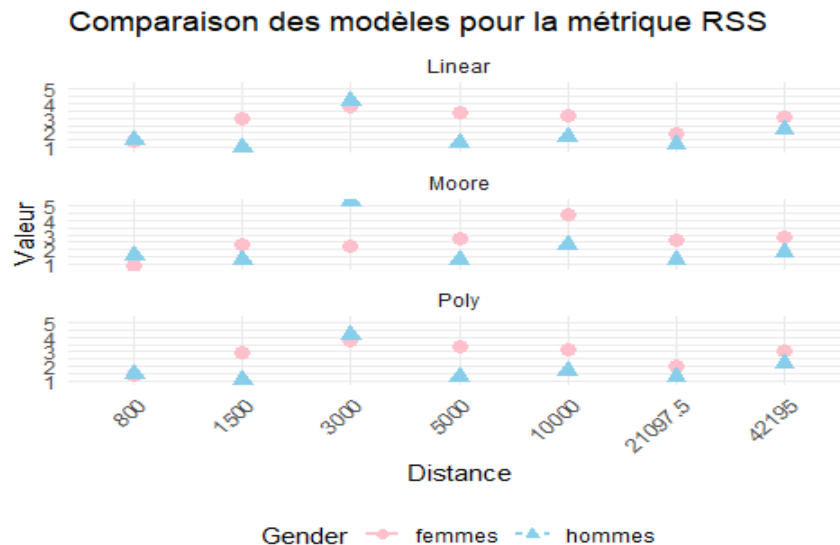
La méthode BIC :

Le graphique évalue l'efficacité de trois modèles statistiques « linéaire, polynomial, et Moore » à l'aide du BIC pour des distances variées en athlétisme, pour hommes et femmes. Un BIC plus bas signifie un meilleur modèle, et le modèle polynomial se distingue comme le plus performant, suggérant une meilleure adéquation globale des données.



Le R2 Ajusté :

Le graphique compare l'efficacité des modèles linéaire, Moore, et polynomial en utilisant le R2 ajusté, qui évalue la variance expliquée en ajustant le nombre de prédicteurs, pour des distances allant de 800 à 42195 mètres chez les athlètes féminins et masculins. Le modèle polynomial se démarque en affichant les valeurs de R2 ajusté les plus élevées, suggérant qu'il explique le mieux la variabilité des performances en fonction de l'âge. Cette tendance se vérifie pour les deux sexes sur l'ensemble des distances, indiquant que le modèle polynomial est probablement le plus adapté pour comprendre la relation entre l'âge et la vitesse dans les données analysées.



Le RSS :

Le graphique évalue l'efficacité des modèles linéaire, Moore et polynomial à travers le RSS, qui mesure la proportion de variance expliquée par chaque modèle, ajustée pour le nombre de prédicteurs. Séparées par sexe et distance, les valeurs du RSS révèlent la précision de chaque modèle dans l'explication de la performance athlétique par rapport à l'âge. Le modèle polynomial se distingue, avec les scores de RSS les plus bas, indiquant sa capacité supérieure à modéliser la performance à travers l'âge, malgré sa complexité plus grande, ce qui en fait un outil prometteur pour une analyse plus approfondie de la relation entre l'âge et la performance athlétique.

Question 4

Nous explorons la question 4, qui se penche sur le lien entre l'âge et les performances sportives, révélant une tendance générale en "U" inversé. Cette observation nous amène à chercher l'âge auquel les athlètes atteignent leur apogée de performance. L'enjeu est de voir si cet âge "optimal" est constant à travers diverses disciplines. Pour ce faire, nous appliquons la méthode MMC et utilisons la fonction `optimize` de R pour estimer cet âge pour différentes compétitions, ce qui nous permettra de comprendre à quel âge les performances sont maximales dans chaque type d'épreuve.

De manière globale

```
table(results$age)
## 6  9 10 11 12 13 14 15 25 26 27 28 29 30 31 44 46 51 52 53 55 56
## 2  1  2  1  3  2  1  1  1  1  2  5  2  2  2  2  1  1  4  2  1  3
```

Le tableau montre les âges de pic de performance pour différents modèles et disciplines sportives. Il n'existe pas d'âge optimal de performance maximale valable pour toutes les disciplines, mais l'âge de 28 ans se distingue en étant le plus souvent associé à des performances optimales, suggérant que cet âge pourrait être particulièrement propice pour exceller dans plusieurs sports.

Pour les hommes

```
table(results_h$age)
## 10 12 13 14 25 26 27 28 29 31 52 53 56
##  2  2  2  1  1  1  1  2  2  1  2  1  3
```

Pour les hommes, l'analyse des âges où la performance est optimale révèle une distribution assez équilibrée sur une large gamme d'âges.

Pour les femmes

```
table(results_f$age)
##  6  9 11 12 15 27 28 30 31 44 46 51 52 53 55
##  2  1  1  1  1  1  3  2  1  2  1  1  2  1  1
```

Pour les femmes, l'analyse des âges où la performance est optimale révèle une distribution assez équilibrée sur une large gamme d'âges.

Question 5

Nous abordons l'exercice 5 en intégrant les données de "resultats_joueurs_echecs.csv", qui détaillent les performances des joueurs d'échecs par âge et nombre de points. Notre but est de tracer la courbe montrant la relation entre l'âge des joueurs et leurs points, en calculant d'abord les âges, puis en adoptant les méthodes de visualisation utilisées précédemment.

Importation et nettoyage des données

```
resultats_joueurs_echecs <- read.csv2("resultats_joueurs_echecs.txt", sep = ";" )
```

Le code pour le jeu de données "resultats_joueurs_echecs" réalise le nettoyage suivant :

- Élimination des lignes avec des données manquantes pour garantir l'intégrité de l'analyse.
- Extraction de l'année de naissance de la date de naissance et conversion en chiffres pour faciliter les calculs.
- Calcul de l'âge en soustrayant l'année de naissance de l'année de la performance, et stockage dans une nouvelle colonne.
- Correction des données de performance en retirant les espaces superflus et conversion en format numérique.
- Suppression des enregistrements où l'âge calculé est négatif pour éviter les incohérences.

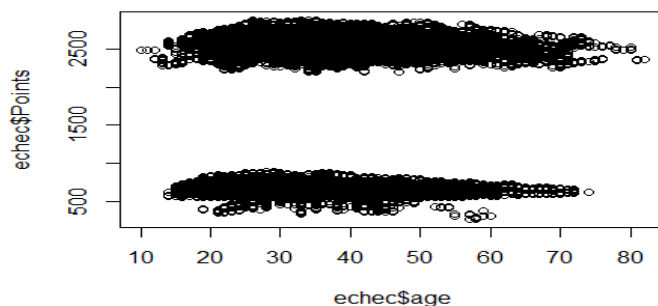
On réalise un *summary* pour vérifier une dernière fois l'absence de données aberrantes.

```
summary(echec)
```

```
##      Année      Mois      Numéro      Nom
## Min.   :1843   Length:125883   Min.    :  1.00   Length:125883
## 1st Qu.:1913   Class :character   1st Qu.: 18.00   Class :character
## Median :1951   Mode  :character   Median : 41.00   Mode  :character
## Mean   :1945                      Mean   : 43.86
## 3rd Qu.:1978                      3rd Qu.: 67.00
## Max.   :2004                      Max.   :100.00
##      Points      Date.naissance      Date.mort      année.naissance
## Min.    : 267   Length:125883   Length:125883   Min.    :1788
## 1st Qu.: 647   Class :character   Class :character   1st Qu.:1877
## Median :2422   Mode  :character   Mode  :character   Median :1913
## Mean    :1688                      Mean    :1909
## 3rd Qu.:2565                      3rd Qu.:1944
## Max.    :2885                      Max.    :1990
##      age
## Min.    :10.0
## 1st Qu.:28.0
## Median :34.0
## Mean    :35.9
## 3rd Qu.:42.0
## Max.    :82.0
```

Représentation des données

```
echecs <- select(echec, age, Points)
plot(echec$age, echec$Points)
```

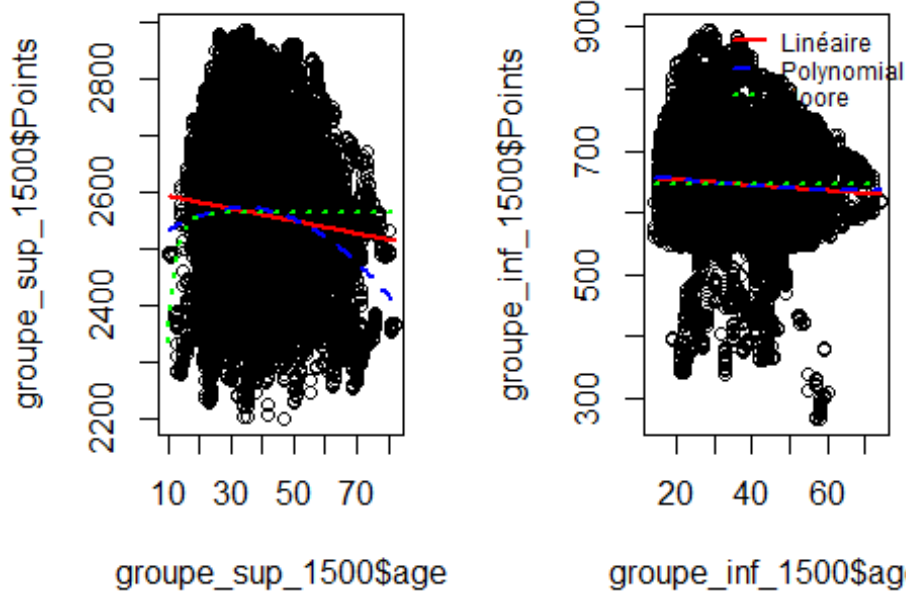


Le graphique illustre clairement la présence de deux groupes distincts lorsqu'on observe la relation entre l'âge et les points chez les joueurs d'échecs. Ces deux groupes pourraient indiquer des tendances ou des caractéristiques spécifiques au sein de la population étudiée. Il pourrait être intéressant de diviser la base de données pour analyser chaque groupe séparément.

Représentation graphique des 3 modèles

Echec

Comparaison des modèles



Ces deux graphiques illustrent la performance des joueurs d'échecs selon leur âge, séparés en deux catégories basées sur leur score : au-dessus et en dessous de 1500 points. Le premier graphique concerne les joueurs ayant un score supérieur à 1500 points, où l'on constate que les modèles linéaire, polynomial, et de Moore affichent des tendances similaires, indiquant une performance relativement constante avec l'âge. Toutefois, le modèle polynomial révèle une baisse subtile de la performance avec le vieillissement, suggérant que l'expérience ne se traduit pas nécessairement par une amélioration chez les joueurs d'élite.

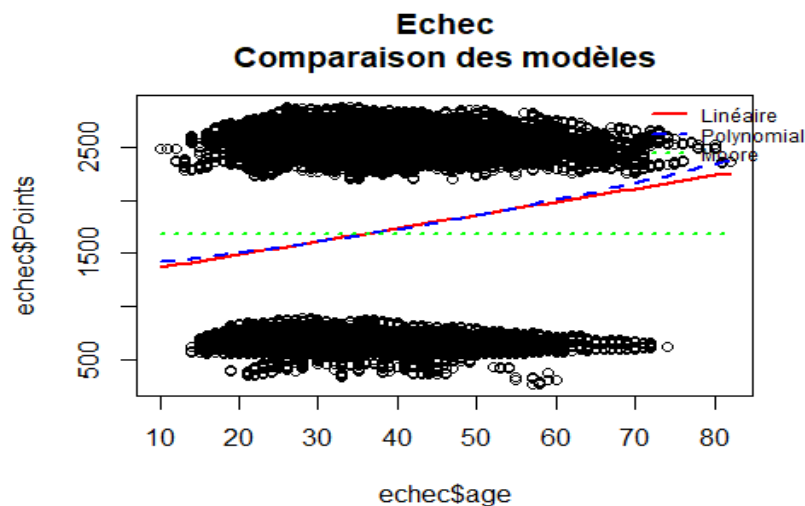
Le second graphique, représentant les joueurs avec moins de 1500 points, montre une plus grande variabilité des performances, particulièrement chez les joueurs plus âgés, où quelques scores extrêmement bas émergent. Malgré cela, les modèles montrent une tendance générale de stabilité entre l'âge et la performance, indiquant que, pour les joueurs ayant des scores inférieurs, l'âge n'influe pas significativement sur leur capacité à jouer.

Echantillonnage

Pour affiner notre analyse, nous avons envisagé de créer un échantillon réduit pour observer si, avec moins d'observations, des différences significatives entre les modèles pourraient émerger.

De la base de données globale

```
echecs <- echecs %>%
  sample_n(500)
```

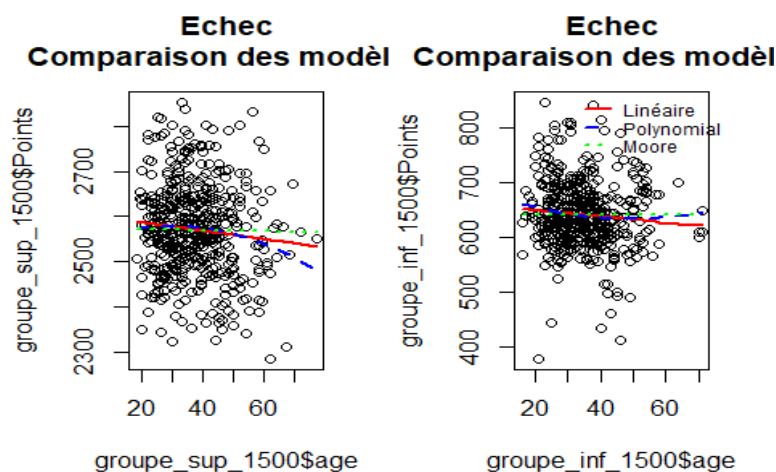


Sur l'échantillon réduit de la base de données, il apparaît clairement qu'aucun modèle ne se distingue particulièrement. Les deux groupes identifiés restent trop séparés pour qu'un modèle unique puisse décrire précisément les deux tendances.

Des deux groupes séparés

Nous allons maintenant procéder à l'échantillonnage au sein des deux groupes identifiés dans la base de données. Cette étape nous permettra d'examiner de plus près les données et de déterminer si des tendances plus claires ressortent lorsque nous nous concentrons sur des sous-ensembles spécifiques.

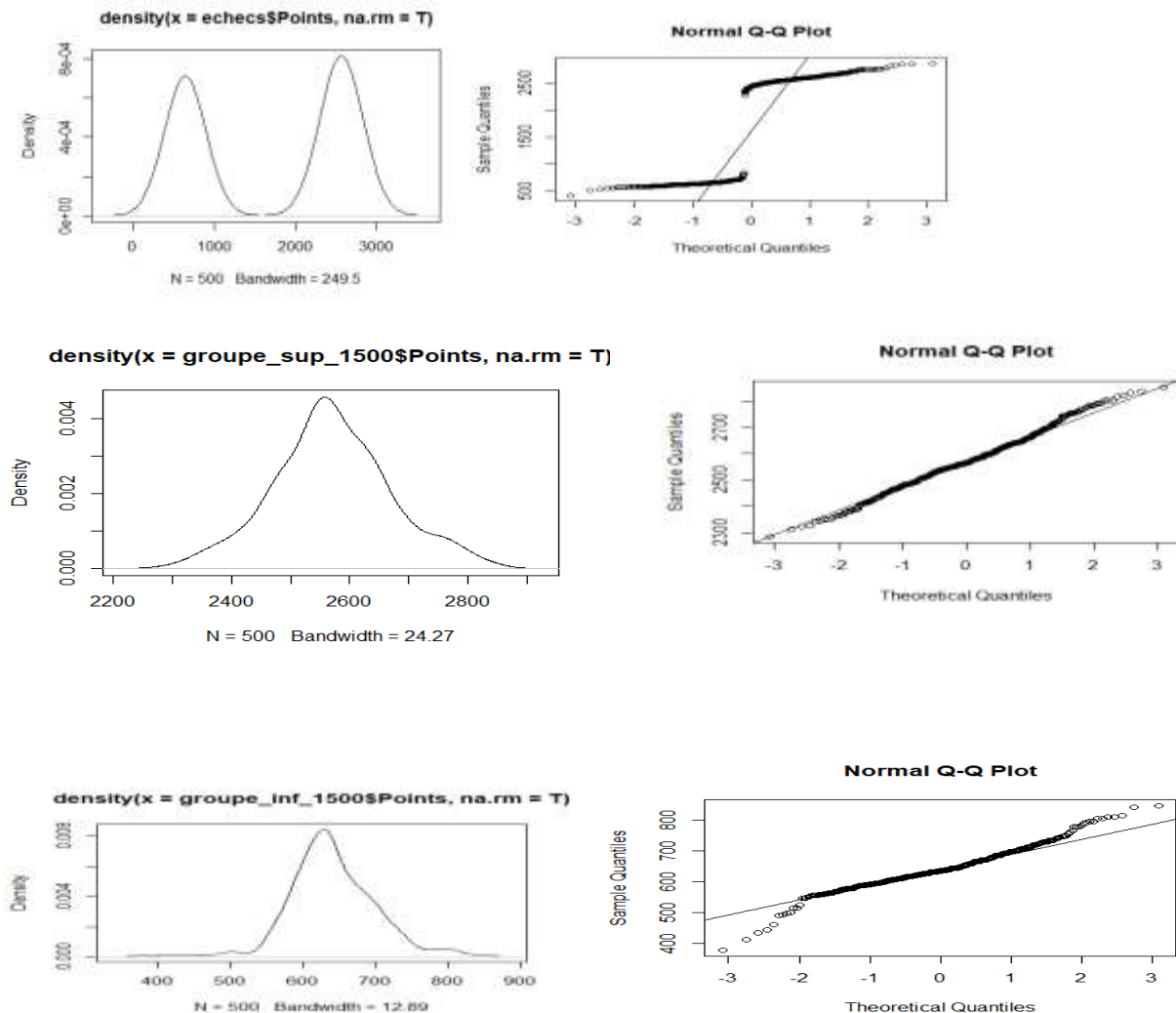
```
groupe_sup_1500 <- groupe_sup_1500 %>%
  sample_n(500)
groupe_inf_1500 <- groupe_inf_1500 %>%
  sample_n(500)
```



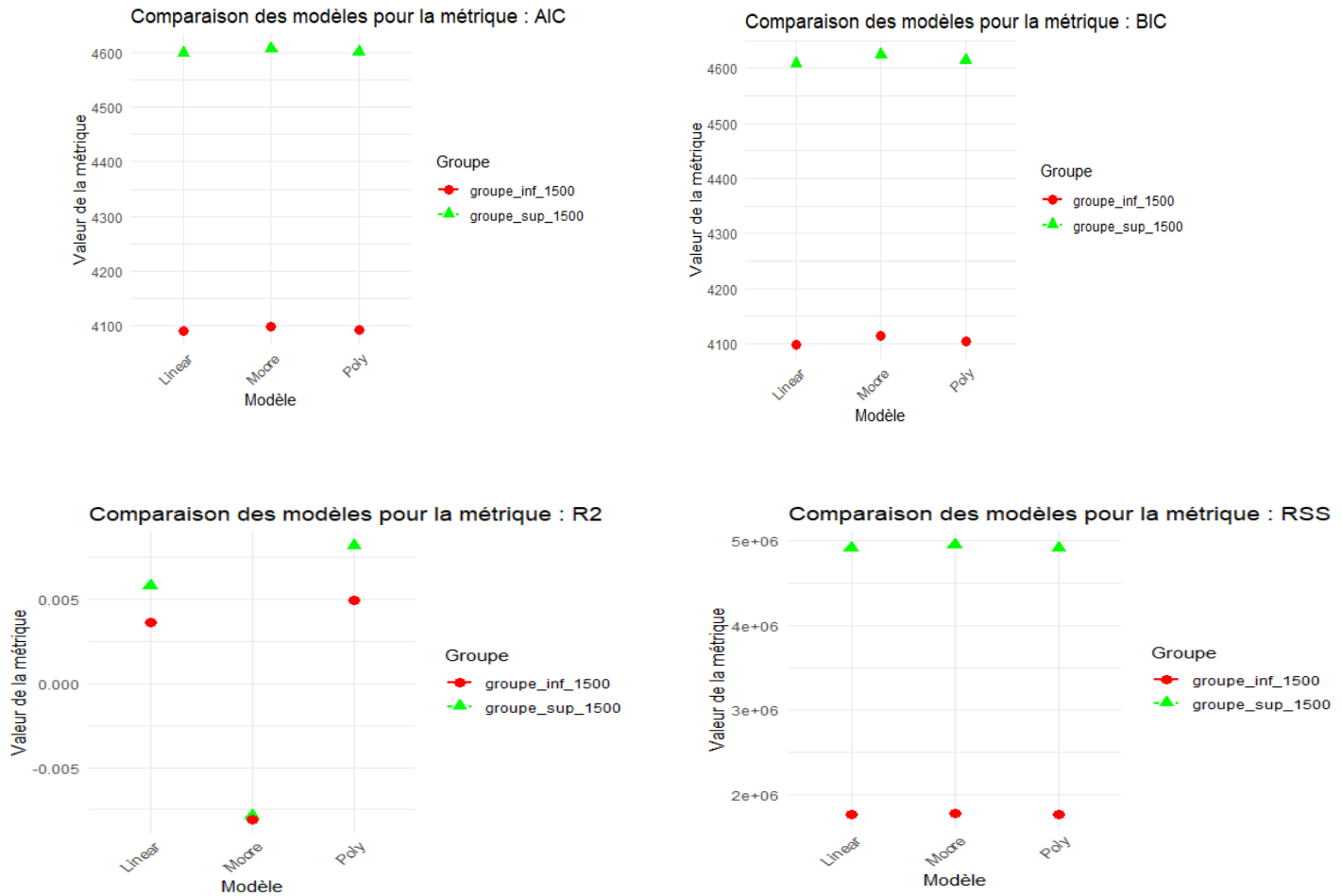
Les deux graphiques, illustrant les groupes au-dessus et en dessous de 1500 points, montrent que les modèles linéaire, polynomial et de Moore ne révèlent pas de différences notables en

termes de tendance par rapport aux observations précédentes. L'échantillonnage n'a pas changé la trajectoire de chaque modèle, qui demeure semblable à celle observée précédemment, ne permettant pas de distinguer nettement un modèle comme meilleur aux autres.

Calcul des différents critères



Les données de échecs ne suivent pas une distribution normale, mais la distribution est normale ou s'approche de la normalité pour les jeux de données supérieur et inférieur à 1500 points. On utilisera donc uniquement les jeux de données inférieur ou supérieur pour le calcul des différents critères.



L'analyse des quatre graphiques montre que les modèles linéaire, polynomial et de Moore ont des performances relativement similaires sur les deux groupes, ceux avec plus et ceux avec moins de 1500 points. Le modèle polynomial semble avoir un léger avantage, bien que cette différence soit minime et ne représente pas un changement significatif par rapport aux autres modèles.

Âge ou les points sont maximaux

##	jeux	modele	Points_maximum	age_maximum
## 1	groupe_sup_1500	modele_lineaire	2588.0049	18
## 2	groupe_inf_1500	modele_lineaire	651.5045	16
## 3	groupe_sup_1500	modele_polynomial	2577.2802	30
## 4	groupe_inf_1500	modele_polynomial	661.7558	16
## 5	groupe_sup_1500	modele_moore	2570.3705	71
## 6	groupe_inf_1500	modele_moore	641.2147	63

On calcule ici l'âge qui correspond aux points maximums obtenu dans la fonction MMC. On remarque que l'âge n'est jamais le même, sauf pour le groupe inférieur à 1500 on remarque que l'âge 14 revient 2 fois. On peut donc conclure que des joueurs qui gagnent peu de points ont tendance à atteindre leur performance maximale à un jeune âge de 14 ans. Pour les joueurs qui gagnent beaucoup de points on ne peut pas vraiment définir d'âge.

Conclusion

En résumé, aucun modèle ne se démarque nettement des autres selon ces critères. La relation âge-points ne révèle pas de tendance ou de forme particulière qui permettrait de privilégier un modèle spécifique. On a simplement pu découvrir que les joueurs qui gagnent peu de points ont tendance à atteindre leur performance maximale à 14 ans.

Question bonus 1

L'objectif de la question bonus est d'évaluer et de comparer les modèles linéaire, polynomial, et de Moore déjà examinés avec une nouvelle équation proposée, en utilisant des paramètres ajustables $(\alpha_0, \alpha_r, \beta_0, \beta_r, t_d)$ pour modéliser la performance $P(t)$ en fonction de l'âge t . Nous cherchons à identifier le modèle qui s'ajuste le mieux et à justifier notre choix.

Notre approche se décompose en plusieurs phases : tout d'abord, nous décortiquerons l'équation pour comprendre ses composantes. Puis, nous appliquerons le nouveau modèle à nos données pour en évaluer graphiquement le comportement. Enfin, nous recourrons à divers critères pour effectuer une comparaison statistique entre les modèles.

Visuellement

Voici comment est présenté l'équation :

$$P(t) = \beta_0 N_\infty \cdot e^{-\frac{\alpha_0}{\alpha_r} e^{-\alpha_r t}} \cdot \left(1 - e^{\beta_r(t-t_d)}\right)$$

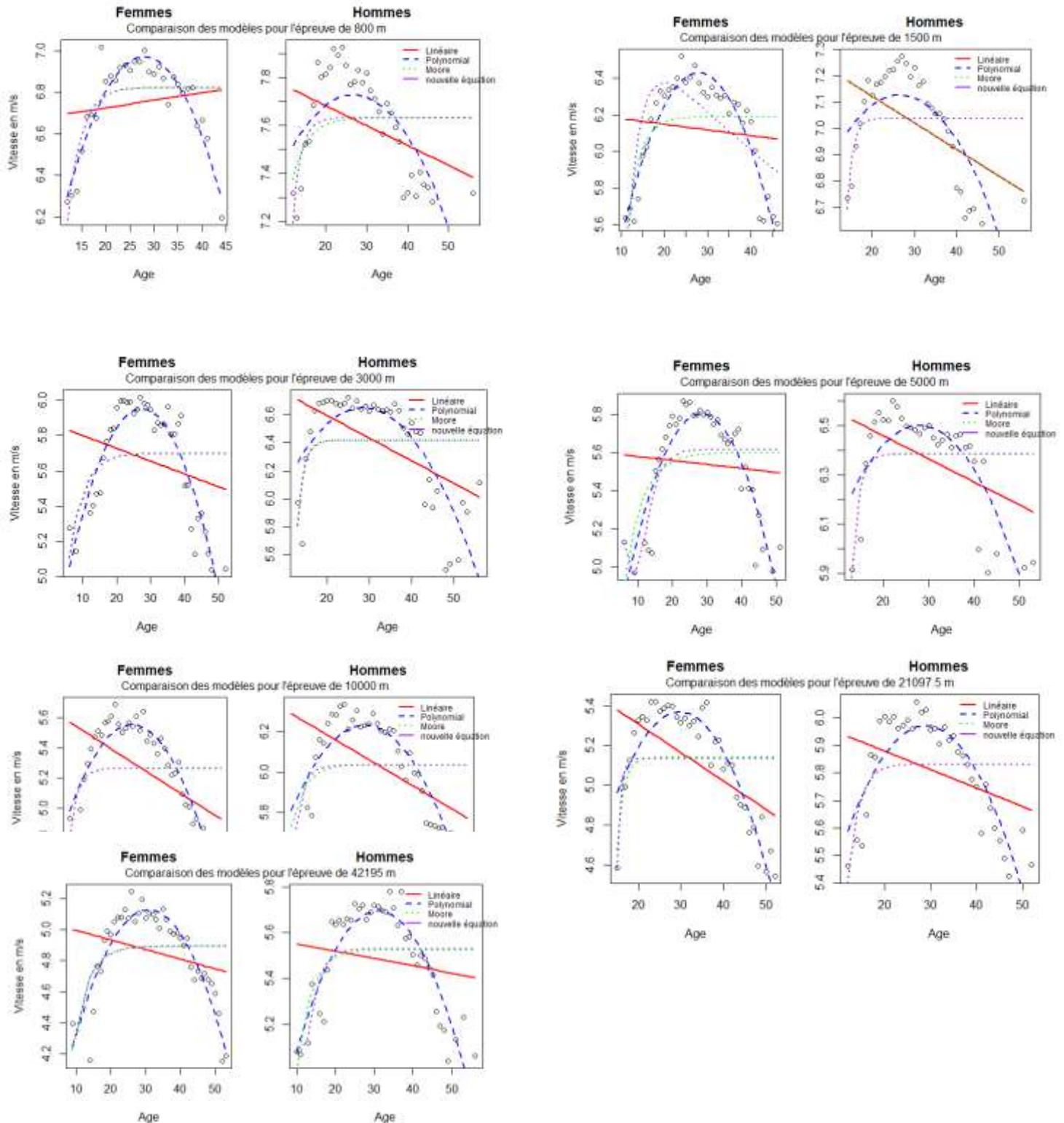
L'utilisation d'exponentielle signifie que le modèle peut représenter des variations rapides, qu'elles soient en hausse ou en baisse. Toutefois, comme nous l'avons vu avec l'équation de Moore et sa double exponentielle, bien qu'elle capture efficacement les évolutions rapides au départ, sa tendance est à s'aplatir par la suite. Il est donc probable que ce nouveau modèle présente un comportement similaire.

Nous envisageons d'implémenter le modèle en R, qui requiert l'ajustement de cinq paramètres. Cependant, concernant N_∞ , sa signification précise n'étant pas définie puisqu'il ne figure pas parmi les paramètres ajustables, nous l'avons interprété comme étant équivalent à la moyenne de la vitesse observée.

```
nouvelle_fonction <- function(t, params) {  
  beta0 <- params[3]  
  alpha0 <- params[1]  
  alpha_r <- params[2]  
  beta_r <- params[4]  
  t_d <- params[5]  
  N <- 5.955  
  P_t <- beta0 * N * exp((-alpha0 / alpha_r) * exp(-alpha_r * t)) * (1 - exp(  
beta_r * (t - t_d)))  
}
```

Graphiquement

On va maintenant appliquer notre nouveau modèle grâce à la fonction MMC. On va ensuite afficher les différents modèles sur nos différents jeux de données pour pouvoir les comparer.

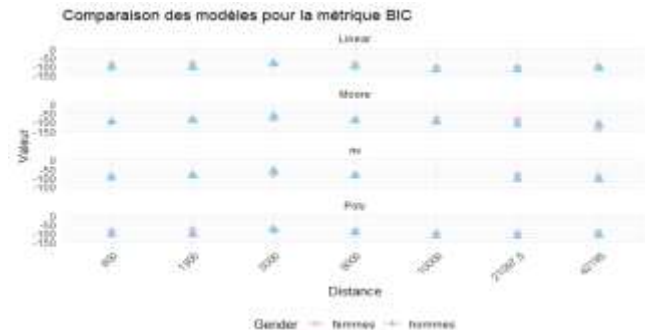


En regardant de près les graphiques pour les modèles linéaire, polynomial, et de Moore, les résultats correspondent à ceux que nous avons déjà identifiés. Le nouveau modèle quant à lui ressemble fortement au modèle de Moore.

Critères

Maintenant que nous avons comparé les modèles de façon graphique, nous allons examiner des critères numériques pour choisir le meilleur. Ces critères vont nous permettre de justifier notre choix en s'appuyant sur des données précises et des mesures de performance pour chaque modèle. Nous avons choisi les critères AICc, BIC, RSS et R^2 ajusté pour comparer les modèles statistiquement et déterminer le plus performant.

##	RSS_nv	R2_nv	AIC_nv	BIC_nv
## 1	0.7783682	0.37182902	-101.82083	-97.05089
## 2	2.1957217	0.14992435	-88.69229	-82.77470
## 3	1.9238258	0.47031226	-109.61784	-102.93815
## 4	2.7182042	0.08311929	-92.06184	-85.56221
## 5	4.4313953	-0.07510151	-86.09526	-78.91088
## 6	2.6151347	-0.03010003	-89.82328	-83.51035
## 7	2.8224920	0.01899290	-97.99974	-91.14617
## 8	1.5252533	0.00627562	-101.80890	-95.89131
## 9	1.2360501	-0.06484462	-96.16914	-90.90882
## 10	5.2377760	-0.05306526	-69.55459	-62.87490
## 11	1.3108867	0.01035581	-94.22930	-88.96899
## 12	1452.0217266	660.93430553	155.43801	162.11771
## 13	1.1969405	0.02339843	-115.01704	-108.89793
## 14	1.6930623	0.14316627	-118.95407	-112.10050



Nous constatons que les modèles de Moore et le nouveau modèle présentent parfois des valeurs aberrantes pour certains critères. Par exemple, le R^2 ajusté du nouveau modèle est supérieur à 1 ou même négatif dans certaines disciplines, ce qui est incohérent. Cette anomalie nous empêche de conclure que ce modèle est supérieur aux autres.

Conclusion

Les anomalies observées jusqu'à présent pourraient très bien résulter d'erreurs dans notre formulation du modèle, donc nous préférons opter pour le modèle polynomial. Ce dernier présente une courbe en forme de U inversé qui suit fidèlement les données et il se distingue par sa simplicité de mise en œuvre ainsi que par ses excellentes valeurs selon les critères évalués précédemment.

Question bonus 2

Nous abordons maintenant la question bonus 2, où l'exercice consiste à exploiter ChatGPT pour répondre à une question de notre choix tirée de ce TD. Par la suite, nous fournirons une analyse concise soulignant les atouts et les limites de la réponse générée par ChatGPT comparativement à celle que nous avons formulée.

Choix de la question choisie

Nous allons nous pencher sur la question 4 pour comparer les réponses fournies par CHATGPT et celles que nous pourrions donner. La question 4 explore comment l'âge affecte la performance sportive, en observant un pic de performance qui prend la forme d'un "U" inversé. Le but est de savoir si l'âge où les athlètes sont au sommet de leur forme est le même pour tous les sports et toutes les distances. Pour cela, on utilisera la fonction MMC afin de déterminer à quel âge les athlètes atteignent ce pic et si ce dernier varie selon les épreuves. L'objectif final est de déduire l'âge de performance optimale pour chaque épreuve analysée.

Réponse CHATGPT

Voici la réponse de CHATGPT non modifiée pour cette question :

```
# Initialisation du dataframe pour stocker les résultats finaux
results2 <- data.frame(sexe = character(),
                       epreuve = character(),
                       modele = character(),
                       vitessemaximum = numeric(),
                       age_maximum = numeric(),
                       stringsAsFactors = FALSE)

# Boucle sur les épreuves
for(epreuve in epreuves) {
  distance_num <- as.numeric(epreuve)
  # Données pour les femmes et les hommes
  femmes_data <- subset(meilleures_performances_femmes, distance == distance_
num)
  hommes_data <- subset(meilleures_performances_hommes, distance == distance_
num)
```

```

# Boucle sur Les modèles
for(modele_name in modeles) {
  # Obtenir la fonction de modèle
  modele_function <- get(modele_name)

  # Nombre de paramètres pour Le modèle
  nbpara <- ifelse(modele_name == "modele_lineaire", 2, ifelse(modele_name
== "modele_polynomial", 3, 4))

  # Appliquer la fonction MMC pour obtenir Le pic de performance et Les par
amètres pour Les femmes
  resultats_femmes <- MMC(femmes_data$age_entier, femmes_data$vitesse, mode
le_function, nbpara)

  # Trouver L'âge de performance maximale avec optimize pour Les femmes
  age_optim_femmes <- optimize(f = function(age) -modele_function(age, resu
ltats_femmes$parametre),
                              interval = range(femmes_data$age_entier),
                              maximum = TRUE)
  # Stocker Les résultats pour Les femmes
  results <- rbind(results, data.frame(sexe = "femmes",
                                     epreuve = epreuve,
                                     modele = modele_name,
                                     vitessemaximum = resultats_femmes$pe
ak,
                                     age_maximum = age_optim_femmes$maxim
um))
  # Appliquer la fonction MMC pour obtenir Le pic de performance et Les par
amètres pour Les hommes
  resultats_hommes <- MMC(hommes_data$age_entier, hommes_data$vitesse, mode
le_function, nbpara)

  # Trouver L'âge de performance maximale avec optimize pour Les hommes
  age_optim_hommes <- optimize(f = function(age) -modele_function(age, resu
ltats_hommes$parametre),
                              interval = range(hommes_data$age_entier),
                              maximum = TRUE)
  # Stocker Les résultats pour Les hommes
  results2 <- rbind(results2, data.frame(sexe = "hommes",
                                     epreuve = epreuve,
                                     modele = modele_name,
                                     vitessemaximum = resultats_hommes$pe
ak,
                                     age_maximum = age_optim_hommes$maxim
um))
}
}
# Afficher Les résultats
head(results2, 10)

```


##	sexe	epreuve	modele	vitessemaximum	age_maximum
## 1	femmes	800	modele_lineaire	6.811253	12.00007
## 2	hommes	800	modele_lineaire	7.749461	55.99994
## 3	femmes	800	modele_polynomial	6.970052	43.99993
## 4	hommes	800	modele_polynomial	7.728010	55.99994
## 5	femmes	800	modele_moore	6.962987	43.99993
## 6	hommes	800	modele_moore	7.631096	12.00006
## 7	femmes	1500	modele_lineaire	6.176054	45.99992
## 8	hommes	1500	modele_lineaire	7.183608	55.99995
## 9	femmes	1500	modele_polynomial	6.427188	45.99992
## 10	hommes	1500	modele_polynomial	7.128339	55.99995

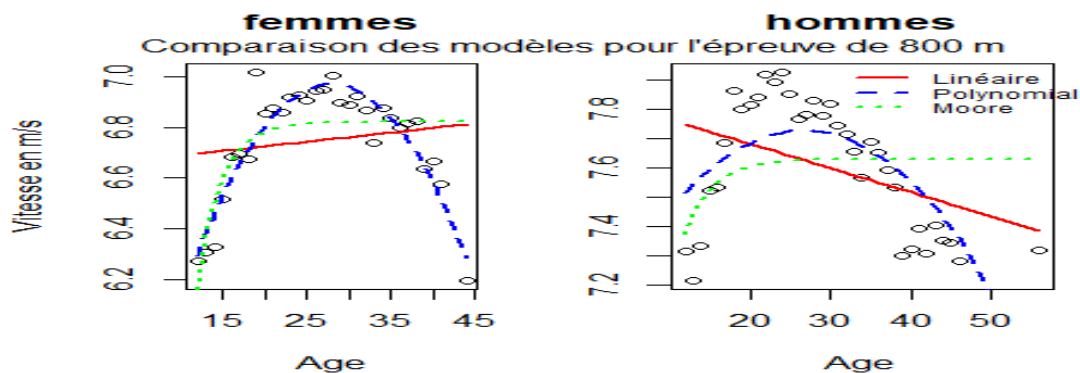
Conclusion par rapport à nos propres résultats

Pour faire une comparaison avec nos résultats, nous allons les affichés pour faciliter la comparaison.

`head(results, 10)`

##	sexe	epreuve	modele	vitessemaximum	age_maximum	age
## 1	femmes	800	modele_lineaire	6.811253	43.99993	44
## 2	hommes	800	modele_lineaire	7.749460	12.00006	12
## 3	femmes	800	modele_polynomial	6.970058	27.94931	28
## 4	hommes	800	modele_polynomial	7.728097	26.70313	27
## 5	femmes	800	modele_moore	6.822364	43.99992	44
## 6	hommes	800	modele_moore	7.631097	55.99993	56
## 7	femmes	1500	modele_lineaire	6.176053	11.00008	11
## 8	hommes	1500	modele_lineaire	7.183607	14.00005	14
## 9	femmes	1500	modele_polynomial	6.427194	27.95356	28
## 10	hommes	1500	modele_polynomial	7.128474	26.38383	26

Après avoir confronté nos résultats avec ceux générés par ChatGPT, nous avons noté des divergences significatives. Pour mieux comprendre les actions de ChatGPT, il est conseillé de revoir les graphiques, notamment celui de l'épreuve des 800 mètres.



ChatGPT semble avoir inversé les données, identifiant l'âge de la performance minimale plutôt que maximale. Malgré nos efforts pour corriger cette erreur auprès de ChatGPT, le problème persiste. Cela souligne l'utilité de ChatGPT tout en rappelant qu'il peut commettre des erreurs et avoir du mal à les rectifier. À l'inverse, notre analyse a correctement déterminé l'âge associé à la vitesse maximale.