Université de Paris

BMW BERLIN MARATHON

Etablissement du profil type du coureur au marathon de Berlin

Farhi Léa-Louise

Caglayan Mustafa

Sadallah Islem

Remond Alexis

Enseignant : Imad HAMRI

IUT de Paris

2020-2021

Sommaire :

Université de Paris

1. Introduction et contextualisation

1.1 Contextualisation

Le marathon est une épreuve sportive individuelle de course à pied. Le marathon de Berlin s'organise tous les ans depuis 1974, généralement à la fin du mois de septembre. Les participants doivent courir le long d'un parcours de 42 km dans les rues de Berlin. C'est l'un des plus rapides : les 6 derniers records du monde de marathon ont été établis à Berlin. Il fait partie du World Marathon Majors, compétition regroupant six marathons majeurs (New York, Chicago, Boston, Berlin, Tokyo et Londres).

1.2 Objectif et méthode d'organisation

L'optimisation de la performance est une problématique très importante dans le domaine du sport. Pour essayer d'y contribuer, nous avons fait le choix d'étudier à l'aide de nos variables d'intérêts, les caractéristiques du coureur qui maximise la performance. En d'autres termes : Quel est le profil type du bon coureur, ce profil est-il toujours le même ?
Pour répondre à ces questions nous avions à notre disposition des logiciels statistiques comme R et des logiciels de data visualisation comme Tableau. Nous étions une équipe de 4 personnes et chacun avait un rôle bien précis. Une personne s'occupait de la partie organisation, une autre plus de la partie rédaction, puis les deux derniers de la partie technique. Néanmoins, chacun restait très polyvalent, et il n'était pas rare que l'on s'aide entre nous. Nous avons fait le choix de tester à chaque fois ce que nous faisions (méthode agile), rien n'était fait sans la validation de toute l'équipe. C'est pourquoi nous nous réunissions toutes les semaines sur zoom, pour faire un compte rendu de notre avancement. De plus, nous avons choisi de commenter notre script R de manière rigoureuse pour pouvoir le présenter en annexe (Page 17).
Nous avons séparé cette analyse en 3 parties, nous analyserons d'abord l'âge-type permettant de courir le plus vite, puis nous étudierons la provenance des bons coureurs et nous terminerons par l'analyse du renouvellement d'une année à l'autre des bons coureurs.

2. Présentation de la base de données

2.1 La base de données et les variables d'intérêts

Nous avions à notre disposition une base de données contenant les résultats de 10 marathons de Berlin consécutifs, de 2001 à 2010, ce qui regroupe les données de plus de 307 031 participants. Nous avions accès pour chacun d'eux au nom, au sexe et à l'âge. L'âge limite d'inscription au marathon est de 18 ans, c'est pourquoi nous avons retiré les données qui dépassent cet âge. Aussi, le temps maximum à partir duquel les résultats ne sont plus comptabilisés est de 6h15min, c'est pourquoi nous avons retiré les données ne respectant pas cette condition. Ensuite, pour enrichir notre base de données nous sommes allés chercher les données sur la nationalité des participants de 2005 à 2010 (197 789 participants) sur les archives de bmw-berlin-marathon.com[1], via une méthode de scraping à l'aide de Python (beautifulsoup). Aussi grâce à cette méthode, 17 034 de nos données manquantes sur l'âge ont pu être complétées.

Par ailleurs, nous avons choisi 5 variables d'intérêts. Il y a l'âge, l'année, la nationalité, le résultat et le sexe. Ceci va nous permettre de faire des analyses et de pouvoir conclure par la suite sur le profil type du bon coureur.

Pour l'âge, suite aux premières analyses, nous avons pu remarquer que l'âge moyen était de 42 ans, l'âge le plus petit est de 19 ans et l'âge maximum est de 80 ans. De plus, nous remarquons que ce sont principalement des hommes qui y participent, par exemple en 2008 ils étaient 80%. En 2008, ce fut aussi l'année ou il eut le à plus de participation avec 35 573 personnes. Au niveau de la nationalité des participants il y a 95% d'européens avec principalement des Allemands. Pour finir, en ce qui concerne la principale variable d'intérêt, le résultat, nous nous occuperons de la présenter dans la partie suivante.

2.2 Présentation de la principale variable d'intérêt : Résultats

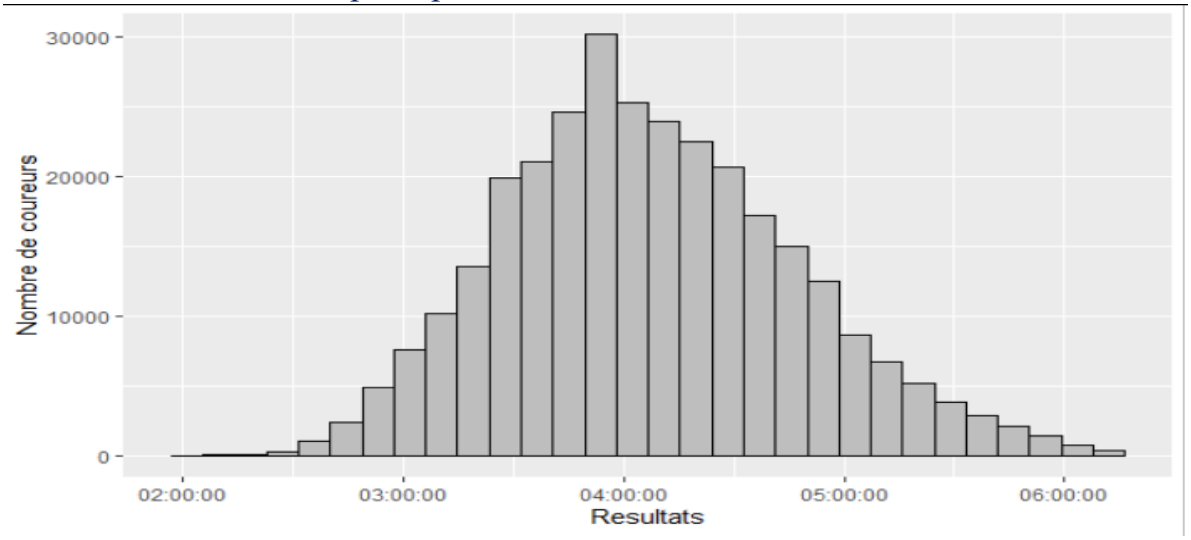

Figure 1 : Histogramme des résultats

[1] https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive

Pour analyser et étudier au mieux la variable résultat nous avons choisi de faire un diagramme en barre. Ce diagramme est rangé en classes d'amplitudes de 8 minutes. Nous pouvons voir que le temps moyen est de 4:07:14. Le temps minimum est de 02:03:59 et le temps maximum de 6:15:00. Aussi la plus grande partie des coureurs font un temps peu inférieur à 4:00:00. Sur le diagramme, on peut distinguer une forte ressemblance de la distribution à une gaussienne. C'est pour cela que nous avons donc réalisé un test de normalité dans le but de vérifier l'hypothèse de normalité sur la répartition du résultat des coureurs.

Test de normalité

Pour vérifier notre hypothèse de normalité sur la répartition du résultat des coureurs, nous avons utilisé la méthode visuelle du QQplot. Le QQ plot (ou quantile-quantile plot) établit la corrélation entre un échantillon donné et la distribution normale. Une ligne de référence de 45 degrés est également tracée. Dans un QQ plot, chaque observation est tracée sous la forme d'un point unique. Si les données sont normales, les points doivent former une ligne droite. Voici le QQplot associé à la variable Résultats :

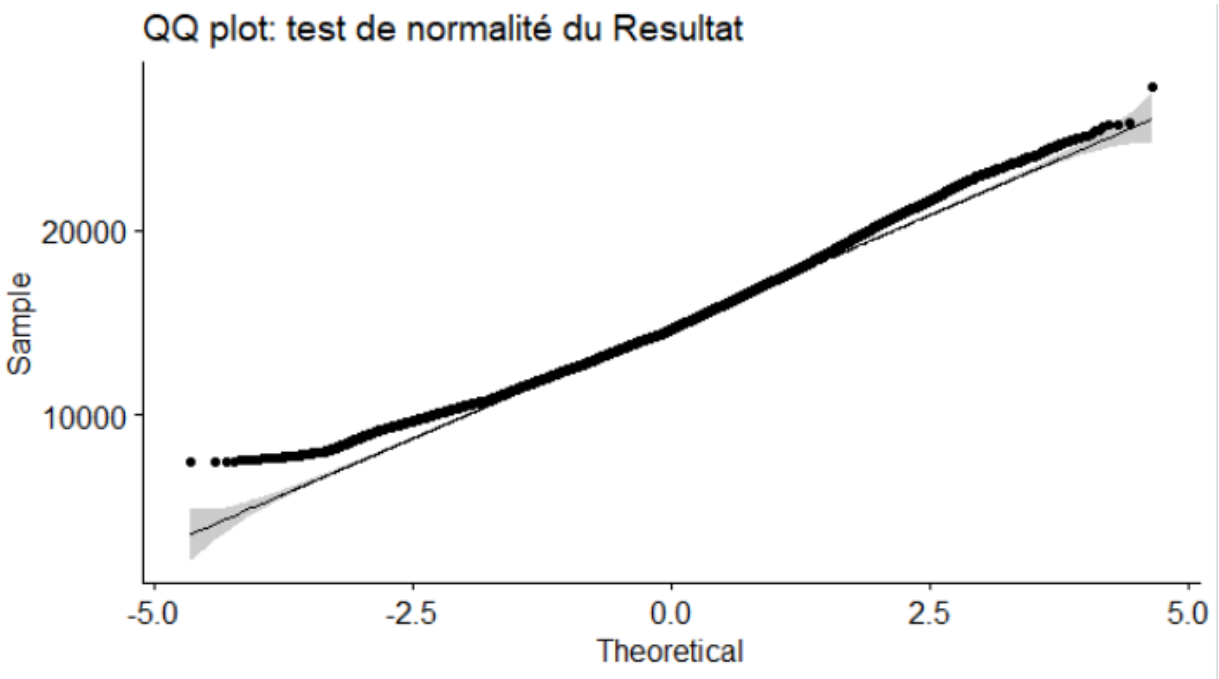

Figure 2 : Test de normalité du résultat

Comme tous les points se situent approximativement le long de cette ligne de référence, nous pouvons confirmer notre hypothèse de normalité. Ainsi la répartition de notre variable résultat s'approche bien de la répartition d'une loi normale.

3. Etude sur l'âge du "bon coureur"

L'âge est une variable qui a beaucoup d'impact sur la performance. En effet, avec l'âge qui évolue, il y'a l'état physique souvent qui se dégrade. Nous allons ici étudier l'importance qu'a cette variable sur la performance du coureur, puis nous déterminerons l'âge idéal pour un coureur.

3.1 Nuage de points et couloirs de performance

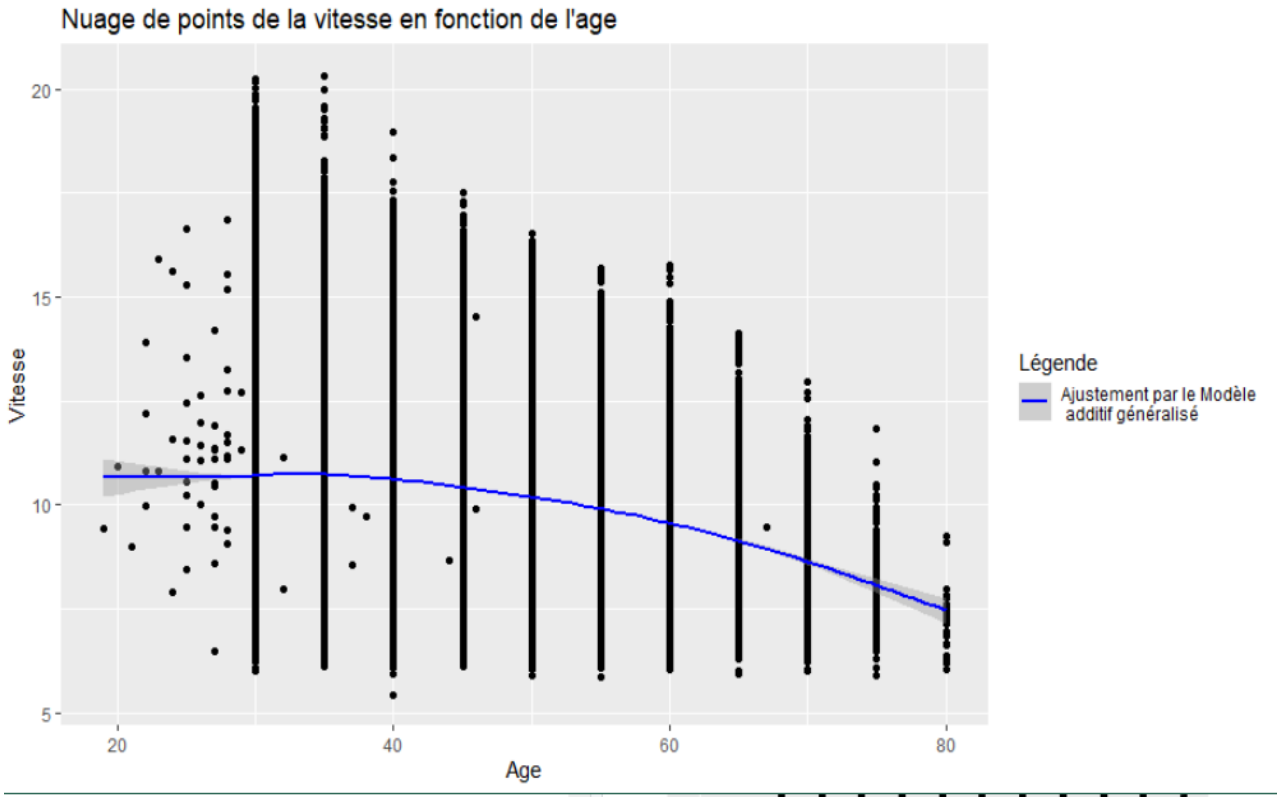

Figure 3 : Nuage de point de la vitesse en fonction de l'âge

Nous avons représenté sur ce nuage de points, la vitesse (en km/h) de chaque coureur en fonction de son âge. Plus on vieillit, moins on a tendance à faire des bons temps. C'est ce que dit l'ajustement de la tendance, qui a une allure décroissante. On remarque aussi que les meilleurs temps se situent autour des 32 ans. Néanmoins, ce nuage de points n'est pas très adapté. D'abord, car la vitesse des coureurs est très disparate, rendant la tendance moins significative. Aussi, car la répartition de l'âge des coureurs comporte des discontinuités, visibles par les longues bandes noires formées par les points. On voit ici une des limites de notre base de données.
C'est pourquoi nous avons choisi par la suite d'autres méthodes de représentation de ces deux variables.

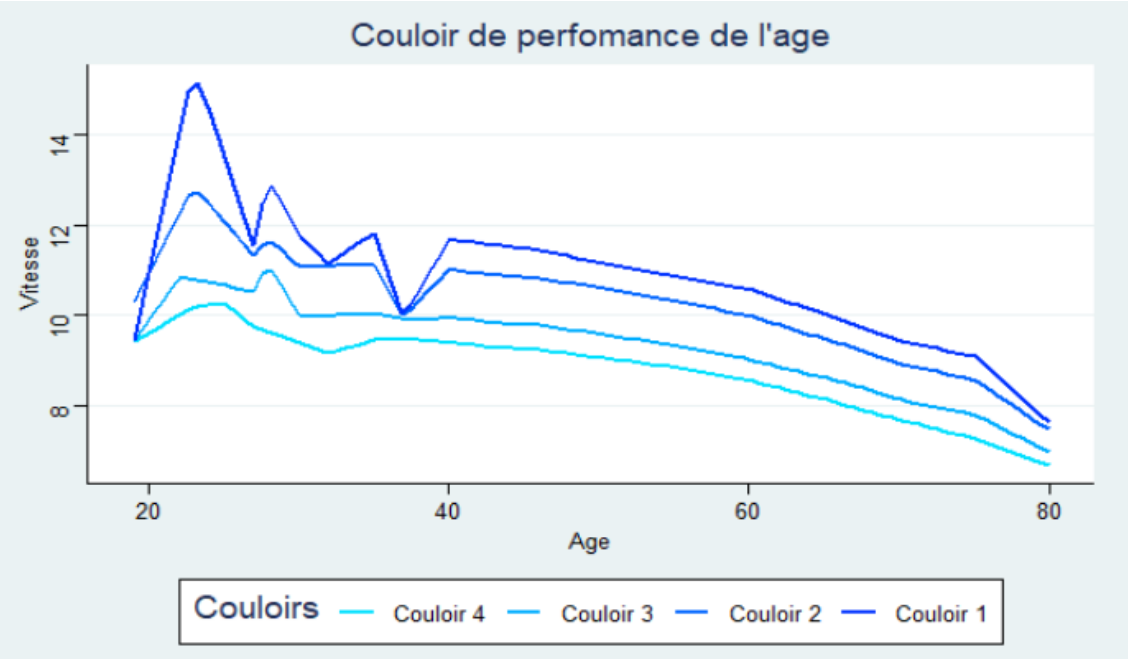

Figure 4 : Couloir de performance de l'âge

Sur ce couloir des performances on utilise deux variables, l'âge et la vitesse (en km/h). Cela nous permet de pouvoir se rendre compte de l'âge pour lequel les participants sont les plus performants, et de mieux capter les différentes tendances par classe de résultats.
Avec l'âge qui augmente, les courbes de couloir de performance diminuent. Mais aussi l'écart entre les couloirs diminue, notamment entre les couloirs 1 et 2 et entre les couloirs 3 et 4. Nous pouvons voir une démarcation entre les 50% premières et 50% dernières. Elle est presque similaire au fil des années.
On peut remarquer que sur le couloir 1 se situe en moyenne les 25% meilleurs coureurs. Sur le couloir 4 se situent les 25% moins performants.
Nous pouvons voir un grand écart entre le couloir 1 et le couloir 2 à 25 ans. La vitesse moyenne est de 20 km/h dans le couloir 1 contre 13 km/h dans le couloir 2. Soit un écart de 7 km/h entre les 2 premiers couloirs. Puis avec l'âge qui avance cet écart-là diminue voire n'existe plus, à 37 ans le couloir 1 et 2 se frôle.
On conclut donc que quel que soit le couloir, la tendance est décroissante, mais que plus les coureurs sont jeunes, plus les écarts entre les coureurs est grand. On en déduit que, l'âge où un coureur peut le plus se démarquer de la normale au niveau de sa performance est entre 20 et 40 ans.

3.2 Méthode de construction

Pour tracer le nuage de points de la vitesse en fonction de l'âge, nous avons fait le calcul de la vitesse grâce à la distance parcourue de 42 km. On a ainsi obtenu la vitesse de chaque coureur en km/h et tracé le nuage de point sur R en fonction de l'âge grâce au package ggplot2. L'ajustement de la tendance s'est quant à lui fait grâce à la procédure geom_smooth() qui utilise la méthode gam (Modèle additif généralisé). Cette méthode n'est pas la plus adaptée, nous avons d'abord opté pour la méthode du modèle de Moore et le modèle IMAP1 (Integrative

model of age-performance)[2] qui sont des modèles plus adaptés pour une étude sur l'âge. Malheureusement, comme nous pouvons le voir sur le nuage de points, il y a de grandes discontinuités dans la base de données pour l'âge des coureurs. Ainsi, les algorithmes d'optimisation de paramètres que nous avons mis en œuvre n'ont pas aboutis à des ajustements représentatifs.

Le couloir de performance quant à lui a été tracé grâce à la méthode geom_quantile() du package ggplot2. Il permet de faire une régression par quantile de la vitesse, en utilisant le modèle "rqss" (Additive Quantile Regression Smoothing). Ces couloirs sont très utilisés pour comparer l'âge avec une autre variable, c'est notamment la méthode utilisée pour comparer l'âge à la croissance ou au poids.

Aussi pour les deux graphiques nous avons fait le choix de retirer les participants en dessus de 18 ans, le marathon étant réservé aux plus de 18 ans. À cela s'ajoute, les valeurs aberrantes pouvant être observées chez les plus de 80 ans que nous avons retirées lors du nettoyage de la base de données.

[2] https://www.nature.com/articles/s41598-018-36707-3

4. Etude sur la nationalité du bon coureur

4.1 Provenance des coureurs de la base de données

Nous avons choisi ici de présenter d'abord la provenance de nos coureurs sans s'occuper de leurs résultats.

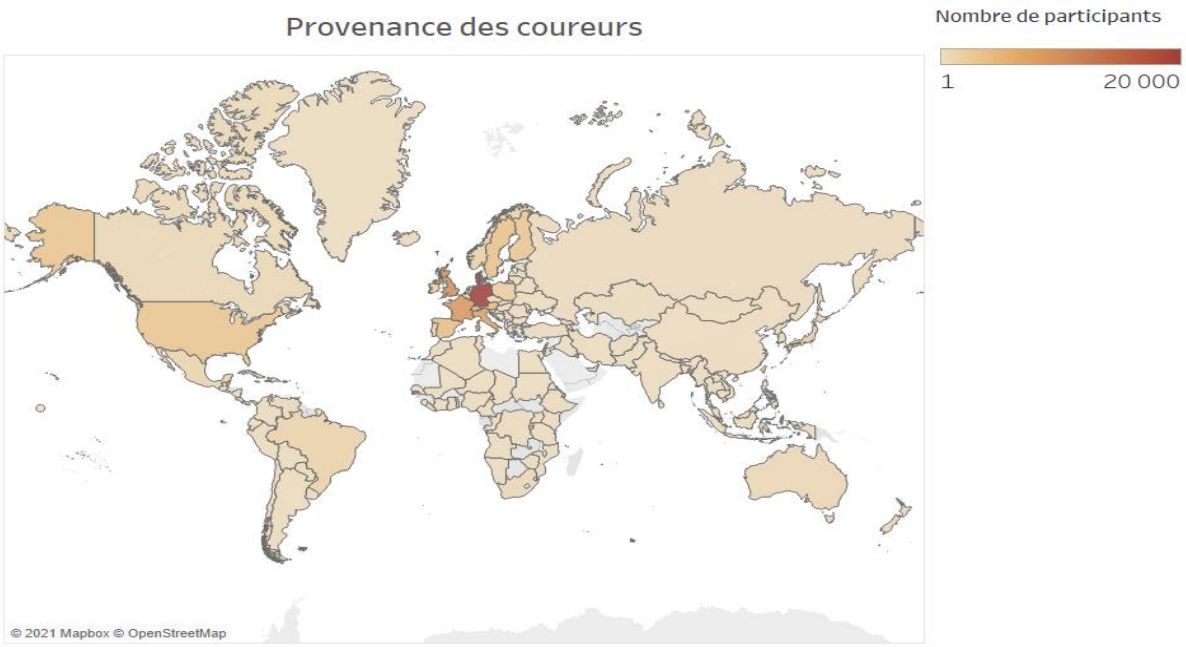

Figure 5 : Provenance des coureurs dans le monde

Ici nous avons une carte du monde qui montre la répartition des coureur(e)s participant au marathon de Berlin de 2005 à 2010. Plus le pays se rapproche du rouge, plus il y a de participants qui courent ce marathon. On peut remarquer la forte présence des participants du marathon dans le continent Européen. Et à l'inverse, la proportion de coureurs dans les autres continents est très faible durant ces 5 années.

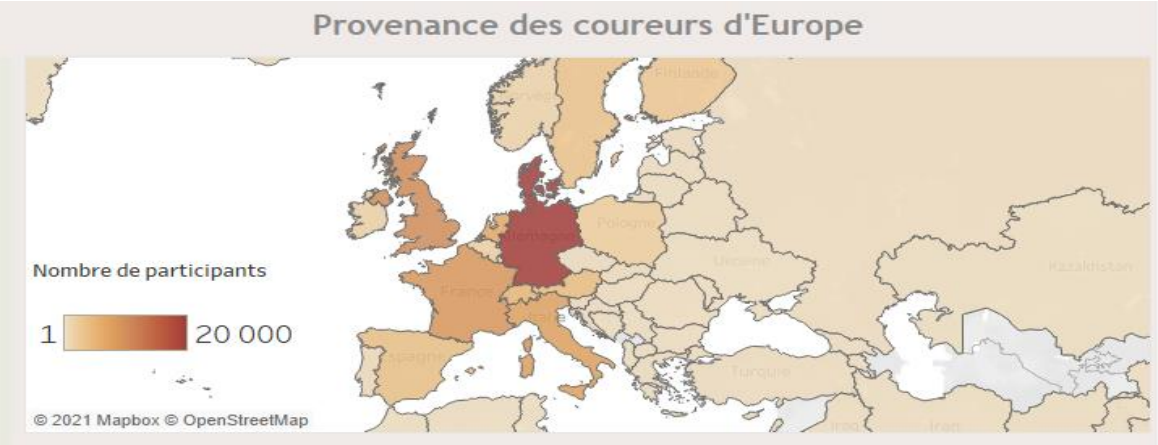

Figure 6 : Provenance des coureurs d'Europe

En centrant notre analyse sur l'Europe, on distingue que ce sont les Allemands qui ont participés le plus durant ces cinq années au marathon de Berlin. Les autres pays européens qui ont beaucoup de coureurs qui participent au marathon de Berlin sont la France, l'Italie et l'Angleterre.

4.2 Analyse sur la provenance des bons coureurs en fonction du Résultats

Après avoir défini la provenance générale des coureurs, dans cette partie nous nous concentrons sur la nationalité des coureurs en fonction de leur performance. Nous allons essayer de déterminer la provenance des meilleurs coureurs.

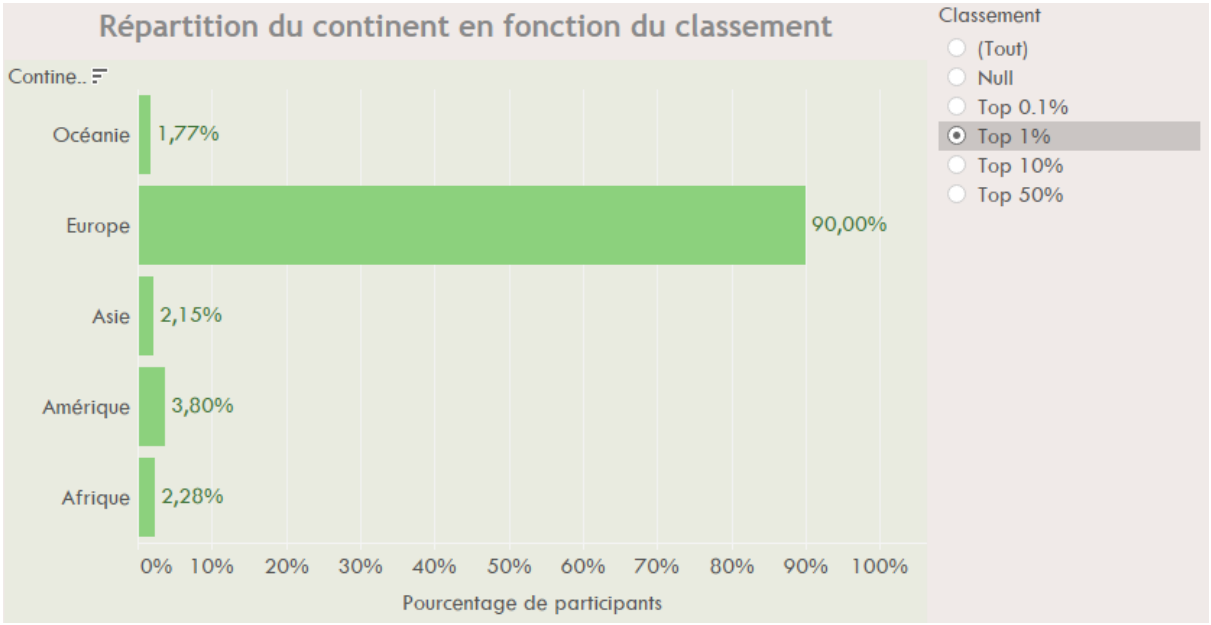

Figure 7 : Diagramme en barres sur la répartition du continent en fonction du classement dans le top 1%

Ce graphique nous montre que le top 1%, c'est à dire à peu près les 3000 meilleurs coureurs toutes années confondues, est majoritairement Européen. De plus, il compte 3.8% d'Américains et 2.28% d'Africains. On voit donc que la majorité d'Européens observés dans la base de données est toujours conservée au top 1% des meilleurs coureurs.

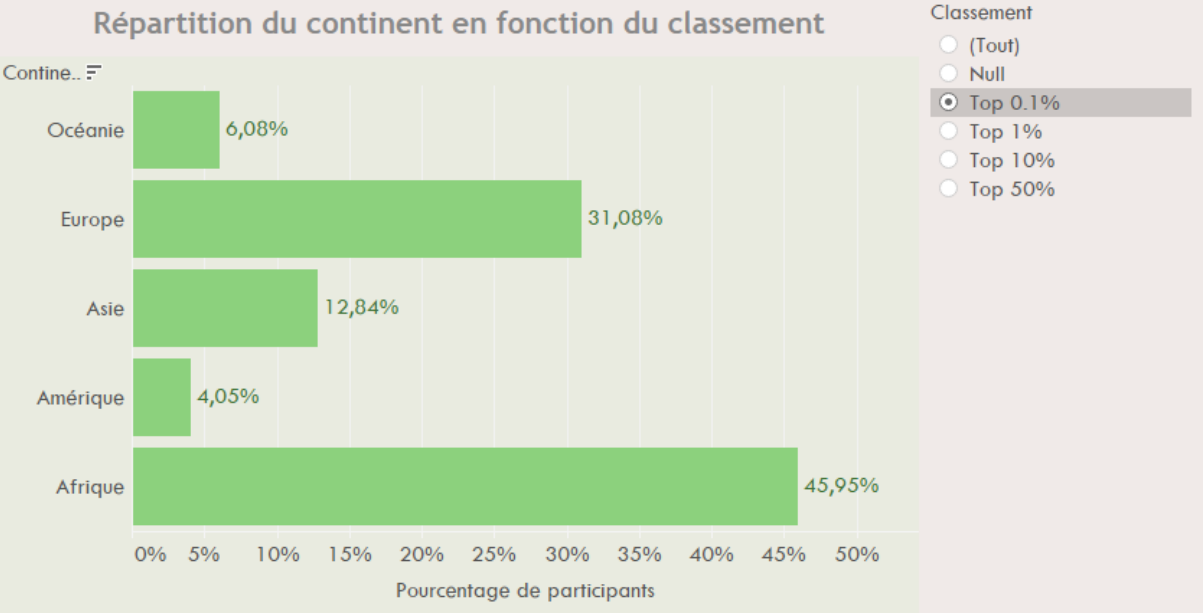

Figure 8 : Diagramme en barres sur la répartition du continent en fonction du classement dans le top 0,1%

La différence s'observe au niveau du top 0.1%, c'est-à-dire, à peu près les 300 meilleurs coureurs toutes années confondues. On voit que les Européens ne sont plus que 31%, malgré une surreprésentation dans la base d'origine. En effet, ce sont les Africains qui sont les plus nombreux au niveau des 300 meilleurs, ils sont 46%. On conclut donc qu'au niveau de la nationalité des coureurs, on n'observe pas de différence pour les 3000 premiers. En revanche, pour les 300 premiers, les Africains ont tendance à être plus nombreux.

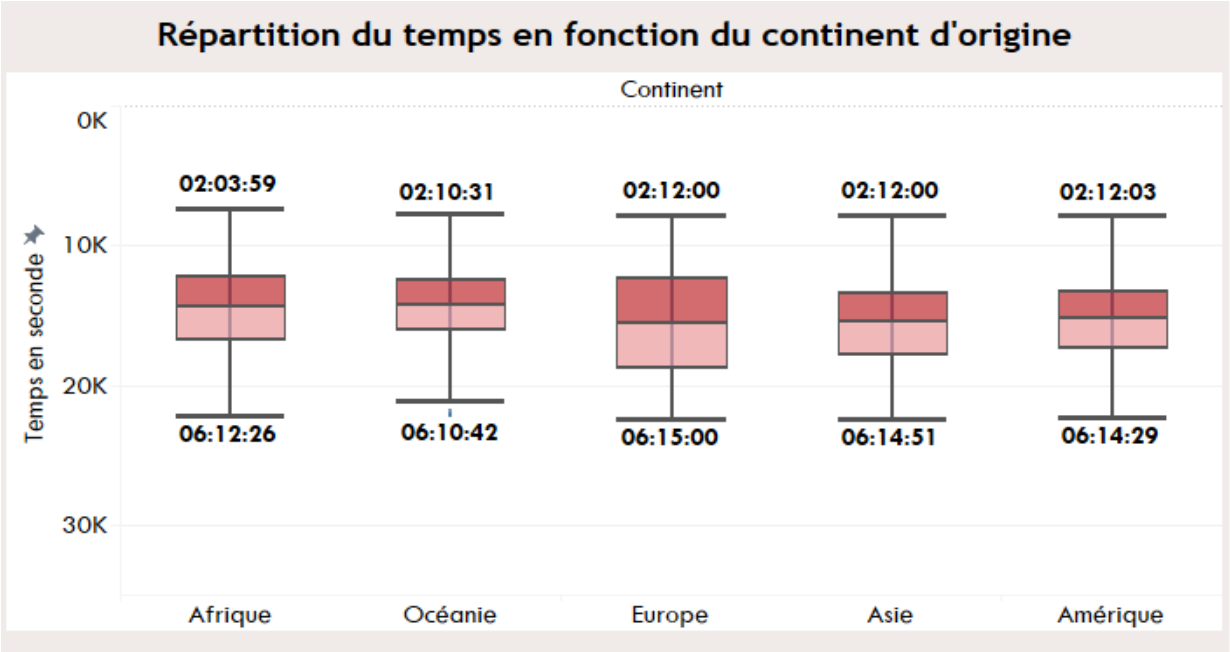

Figure 9 : Boîtes à moustache sur la répartition du temps en fonction du continent d'origine

Les boîtes à moustaches ci- dessus montrent la répartition du temps en fonction du continent d'origine. On remarque que l'Afrique est le continent avec les meilleurs résultats dont le minimum est 02 :03 :59 et le maximum de 06 :12 :26. L'Europe est le continent avec le pire résultat, qui est de 07 :42 :28. On peut voir aussi que l'Océanie à une médiane avec un temps de 03 :57 :23, très proche de l'Afrique dont le temps médian est de 03 :59 :39, qui fait de l'Océanie le deuxième meilleur continent. Ainsi les différences observées entre Européens et Africains dans nos diagrammes en barre se confirment sur ces boites à moustache.

4.3 Méthode de construction

Pour les graphiques de cette partie, le logiciel que nous avons principalement utilisé est Tableau. Comme dit en introduction nous avons extrait les données sur la nationalité des coureurs (de 2005 à 2010) sur le site BMW Berlin Marathon à l'aide de la library BeautifulSoup[3] de Python. Ensuite, une fois les données jointes, nous avons calculé le classement des coureurs sur les années de 2005 à 2010. Aussi nous avons rangé les nationalités par continent pour obtenir une idée plus globale de la provenance des coureurs. La base de données étant grande nous avons choisi de ranger par 4 classes "exponentielle" le classement : Top 0.1%, Top 1%, Top 10%, Top 50%. Enfin, nous avons réalisé un dashboard interactif[4] permettant au mieux de visualiser ses données en fonction du résultat et du classement. Vous trouverez le dashboard interactif lié à cette en annexe.

[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[4] https://public.tableau.com/app/profile/caglayan/viz/paysss/Tableaudebord2

5. Les bons coureurs sont-ils toujours les mêmes ?

5.1 Taux de renouvellement en fonction de l'année

Maintenant que nous avons établi les caractéristiques qui ressortent le plus chez les bons coureurs. Nous allons distinguer si elles sont toujours les mêmes au cours des années. D'une année sur l'autre, les bons profils se répètent-ils ?

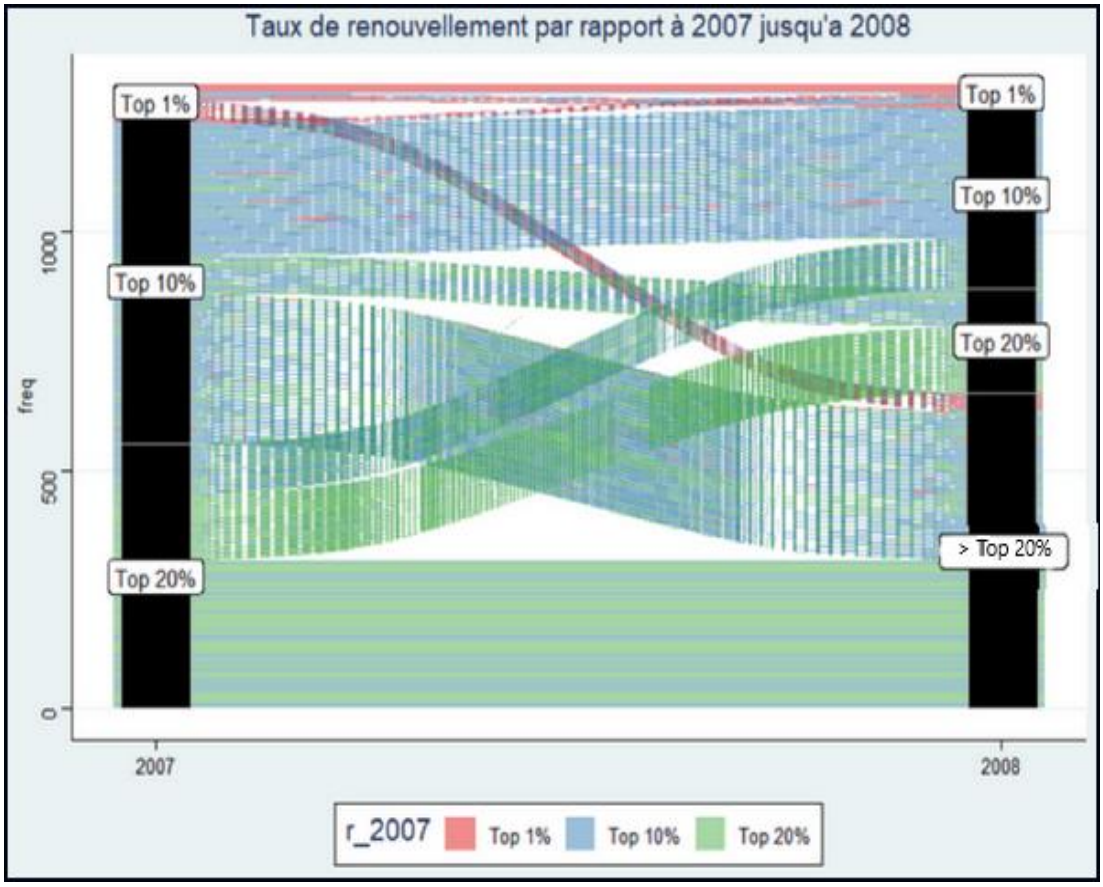

Figure 10 : Taux de renouvellement par rapport à 2007 jusqu'à 2008

Ce graphique du taux de renouvellement représente l'évolution générale des tops que nous avons voulu représenter, en fonction de l'année de 2007 à 2008. Nous avons représenté le "Top 1%" avec la couleur rouge, le "Top 10%" avec la couleur bleue et le "Top 20%" avec la couleur verte. Par exemple, on peut voir qu'environ 30% du top 1% des coureurs de 2007 sont descendus au top 20% en 2008. Enfin, nous pouvons apercevoir que 22% des participants du top 20% en 2007 sont restés dans le top 10% en 2008, alors qu'environ 45% des coureurs de 2007 sont descendus dans le top inférieur à 20% en 2008.

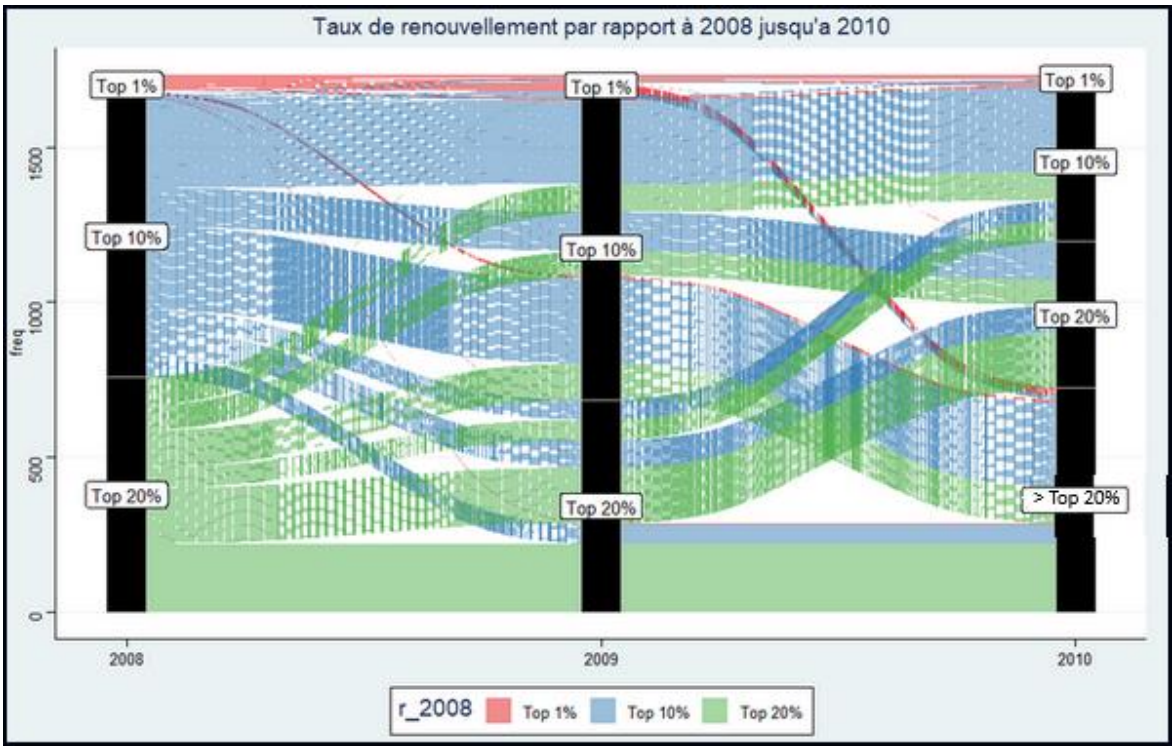

Figure 11 : Taux de renouvellement par rapport à 2008 jusqu'à 2010

Dans ce graphique du taux de renouvellement, nous avons voulu voir la progression des différents tops entre 2008-2009-2010. Nous avons représenté le "top 1%" en couleur rouge, le "top 10%" en couleur bleu et le "top 20%" en couleur verte. Nous avons décidé de garder la même couleur pour chaque top dans les trois années pour voir la réelle évolution entre 2008 et 2010. On peut remarquer qu'environ 5% du top 1% des coureurs en 2008 sont descendus au top 10% en 2009, qui sont ensuite descendus en dessous du top 20% en 2010. Mais aussi, on peut constater que 20% des top 20% des coureurs de 2008 sont passés dans le top 10% en 2009, mais seulement 10% sont restés dans le top 10% en 2010.

5.2 Méthode de construction

Ce type de diagramme alluvial est très utile pour représenter les changements au fil du temps. Il a été réalisé grâce à la documentation de Jason Cory Brunson[5]. Cette documentation utilise la fonction geom_alluvial() du package ggplot2. Le plus compliqué pour nous a été de mettre notre base de données dans un format reconnu "alluvial data". Après avoir calculé le classement de chaque observation conditionnellement à l'année, nous avons transformé notre base de données en un tableau récapitulant le classement de chaque athlète en fonction de l'année. Nous nous sommes principalement concentrés sur les années de 2007 à 2010 car ce sont les années où le taux de réinscription d'une année à l'autre est le plus élevé (5% de re-participation en moyenne). Malgré cela, nous avons fait le choix de ne pas représenter sur ce graphique ceux qui ne participaient pas d'une année à l'autre, car ils sont trop nombreux. Pour les mêmes raisons d'ergonomie graphique nous n'avons représenté les supérieurs au Top 20% que pour la dernière année.

[5] https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html

6. Conclusion

Les éléments étudiés au cours de ce rapport nous permettent de conclure sur les caractéristiques qui ressortent le plus chez les bons coureurs. D'abord, comme montré sur le couloir de performance, les meilleurs coureurs font leurs preuves entre 20 et 40 ans. Ensuite, nos données sur la nationalité ont permis de conclure que les meilleurs coureurs provenaient essentiellement du continent Africain. Enfin les coureurs qui font partie du top 1% conservent souvent leur place depuis longtemps. Malgré un top 10% très volatile, les bons coureurs sont souvent des anciens participants. Ainsi, tous ses éléments nous permettent d'avoir une idée précise sur le profil type du bon coureur.

L'une des difficultés qui nous a le plus ralentie dans ce projet, était la maîtrise du logiciel R. En effet, au début du semestre nos cours en R n'avaient pas suffisamment avancé pour nous permettre de mener des analyses sur notre projet. Nous avons dû apprendre le logiciel à l'aide de documentation en autonomie. Cela nous a permis d'avoir une maîtrise du logiciel plus complète et surtout d'avoir des connaissances sur des graphiques encore jamais tracés. Une des autres difficultés rencontrées a été le nombre de valeurs manquantes au niveau de l'âge des participants, que nous avons d'ailleurs essayé de combler avec le scraping du site de BMW Marathon Berlin. Malheureusement, cela n'a pas suffi à régler le problème de discontinuité et a empêché de tracer le modèle de Moore. Pour conclure, ce projet nous a appris à travailler en équipe et à apprendre de nouvelles compétences de manière autonome.

Tables des illustrations :

Annexes :

Script commenté permettant de tracer les graphiques :

https://github.com/MustafaCaglayan/Marathon-de-Berlin/blob/main/Script%20principal.R

Dashboard interactif sur la provenance des coureurs :

https://public.tableau.com/app/profile/caglayan/viz/paysss/Tableaudebord2

Pour en savoir plus :

Impact of Environmental Parameters on Marathon Running Performance (plos.org)

Different race pacing strategies among runners covering the 2017 Berlin

The Age-Related Performance Decline in Marathon Running: The Paradigm of the Berlin Marathon. - Abstract - Europe PMC

Different race pacing strategies among runners covering the 2017 Berlin Marathon under 3 hours and 30 minutes - PubMed (nih.gov)

Sources :

Marathon de Berlin — Wikipédia (wikipedia.org)

BMW BERLIN-MARATHON: bmw-berlin-marathon.com

Université de Paris