IUT de Paris - Rives de Seine
Université de Paris

RAPPORT STATISTIQUE

Analyse de données, Reporting et Datavisualisation

08/06/2022

Hawa Camara | Ruth Jeyaranjan

INTRODUCTION

Dans le cadre d'un projet invoquant **l'analyse de données, le reporting et la datavisualisation**, nos encadrants Imad HAMRI et Virginie AZOULAI nous ont fourni une base de données sportives des compétitions d'athlétisme de 2019. Ces dernières sont aussi accessibles sur le site de la Fédération Française d'Athlétisme. Elle regroupe l'ensemble des clubs faisant pratiquer l'athlétisme et organise chaque année les championnats de France.

Il existe plusieurs disciplines dans ce sport, et parmi celles proposées, notre équipe de 4 étudiants a fait le choix de traiter la base de données qui regroupe **toutes les courses de marathon de l'année 2019**. Originaire de la Grèce antique, le marathon est une épreuve de course à pied de fond, sur une distance de 42,195km. Cette distance a été fixée en 1921.

Ce rapport statistique fait suite à un rapport informatique, dont la principale fonction a été de regrouper l'ensemble des opérations réalisées sur la base de données, à savoir le nettoyage et le traitement des données. Cette première phase est utile pour la **création d'une base de données d'analyse** qui doit répondre aux besoins du **plan d'analyse statistique**. Nous procédons à présent à l'analyse des données, afin de répondre à notre **problématique**:

Des facteurs socio-démographiques, météorologiques et géographiques influencent-ils la performance des coureurs?

BASE DE DONNEES

La base de données comporte **88 288 observations**, soit le nombre de participations aux marathons en 2019. Associés à ces observations, plusieurs caractéristiques ont été relevées dans la population, d'autres sont issus des besoins de la problématique. Elles sont au nombre de 27, toutes confondues, certaines donnent des informations sur l'athlète, d'autres sur le marathon.

- Temps : temps de course sans seconde et avec seconde
-Heures : heure numérique
- Athlete : nom de l'athlète
- Record personnel : record personnel ou non
- Classement
- Club : nom du club
- Nationalite
- Sexe
- Année de naissance
- Categorie d'age : tranche d'âge soumis à la réglementaition

CATEGORIE	CODE	AGE
Masters	VE	>35 ANS
Seniors	SE	>23 ans
Espoirs	ES	>20 ans

- Age
- Tranche d'age
- Appartenance à un club
-Departement du club
- Ligue : region à laquelle le club est rattaché
- Date du Marathon: record personnel ou non
- Mois
- Ville du Marathon : nom du club
- Pays
- Température: en celcius
- Humidité
- Pression
- Precipitation

Cette présentation des données n'exhaustive, l'ensemble des données que contient notre base de données n'ayant pas servi à l'analyse statistique.

HYPOTHÈSES

Avant d'exécuter notre plan et de procéder à des analyses concrètes afin de répondre à la problématique, nous avons établi des **hypothèses** à la réponse. Elles sont peut-être basées sur des **idées reçues**, des **préjugés** ou une simple naïveté de notre part, mais voici notre réponse : oui, il existe un lien entre les facteurs socio-démographiques, météorologiques et géographiques et la performance des coureurs.

D'abord le **sexe** est un facteur évident, puisque les hommes ont de meilleures performances physiques que les femmes.

Ensuite, **l'âge** : à priori, les jeunes réalisent de meilleures performances que les personnes plus âgées.

Nous pensons aussi que **le club** auquel le coureur appartient, voire son **appartenance à un club** influencent sa performance: un coureur rallié à un club sera bien mieux entrainé. Encore mieux si son club a une renommée.

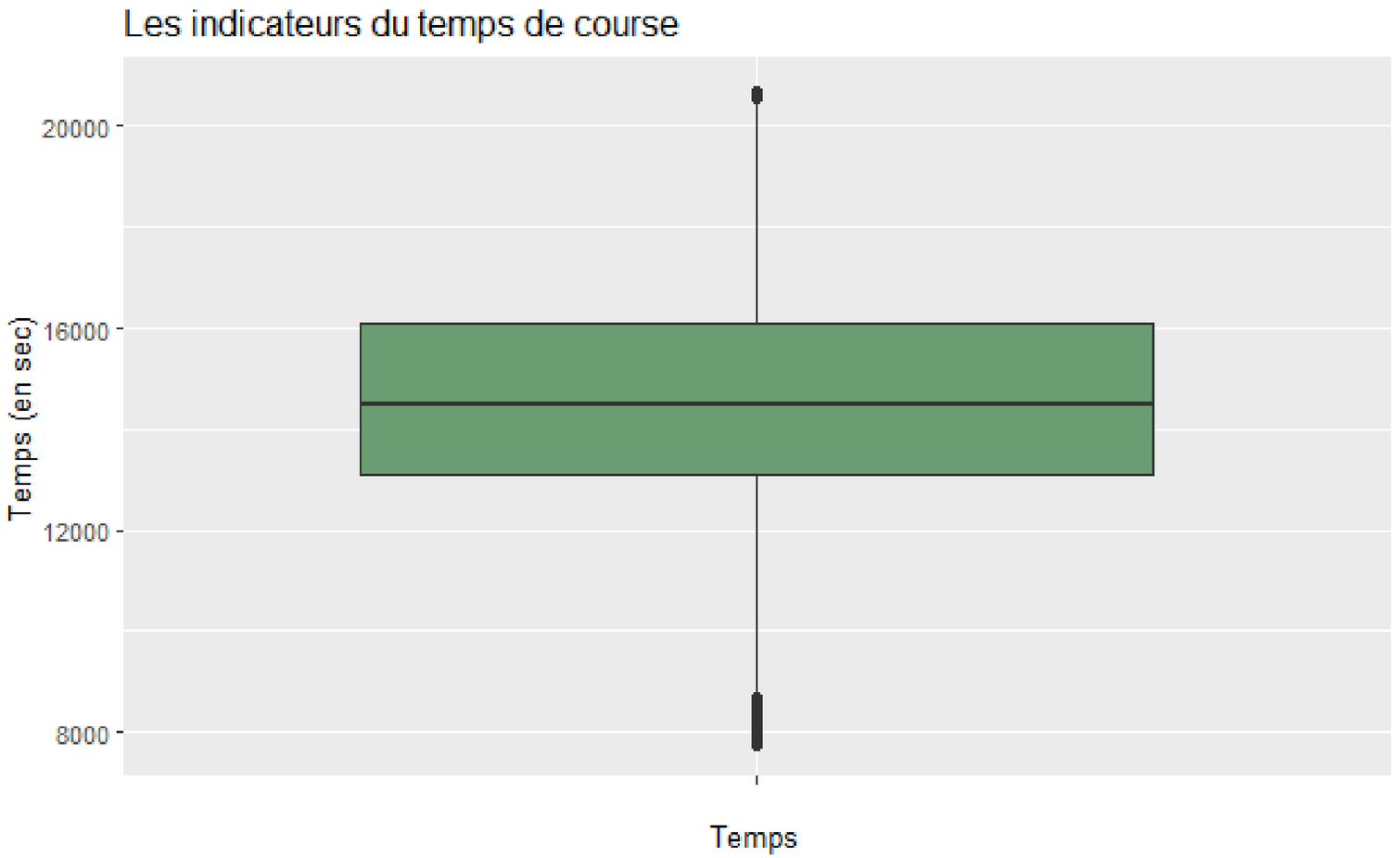
Quant à la **météo**, nous estimons qu'un vent très fort risque de freiner les coureurs; ou au contraire, de les pousser dans leur élan !

Une **température** élevée les essoufflerait plus vite et les fatiguerait. L'humidité ne devrait pas déteindre sur la course nos athlètes, d'ailleurs la fraicheur qu'elle apporte stimule peut-être leurs muscles.

Enfin, les facteurs géographiques restent encore à déterminer, mais nous pensons étroitement que le relief des villes de compétitions a un impact sur la performance des coureurs.

PLAN D'ANALYSE STATISTIQUE

Le temps/ la performance

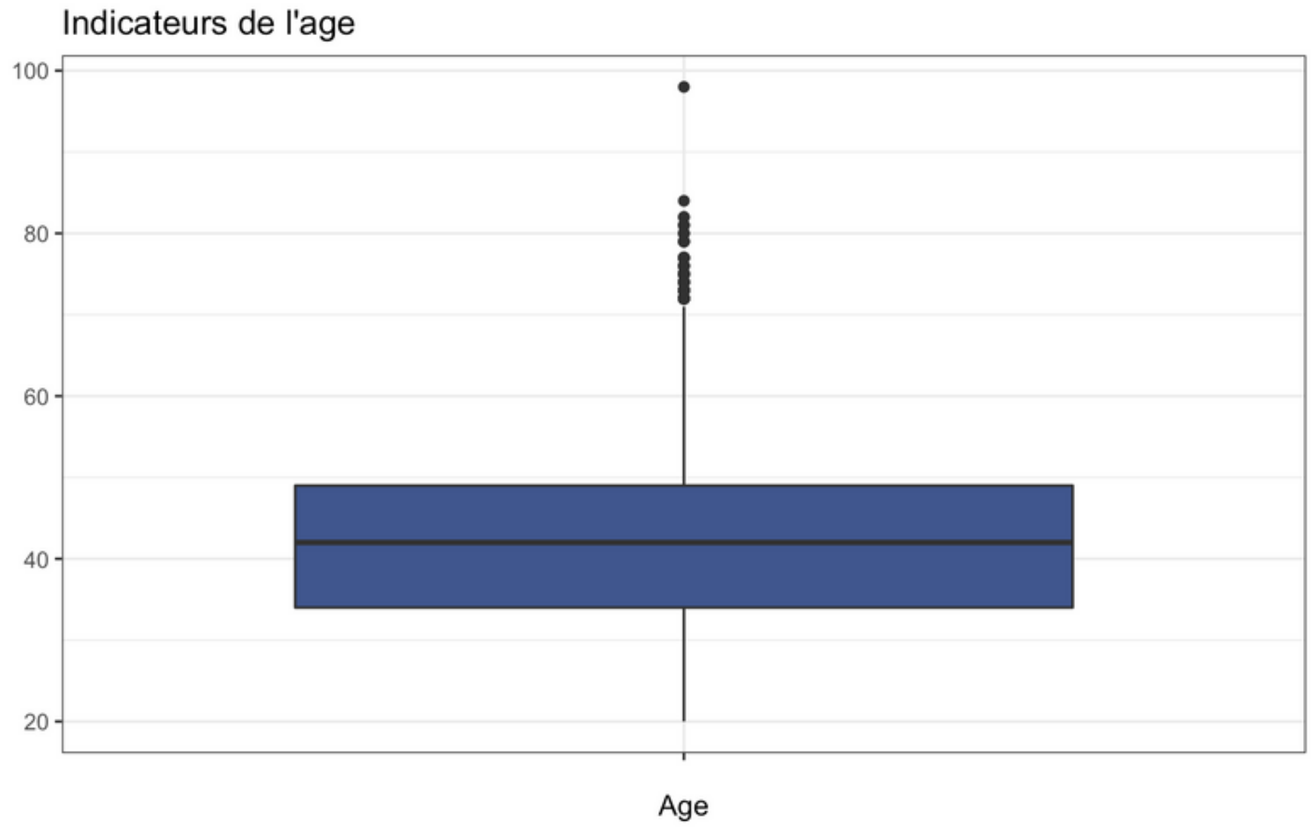


Indicateurs	Temps
Min.	2h09min
1er Qu.	3h38min
Mediane	4h01min
Moyenne	4h03min
3eme Qu.	4h28min
Max.	5h45min

1) FACTEURS SOCIO-DEMOGRAPHIQUES

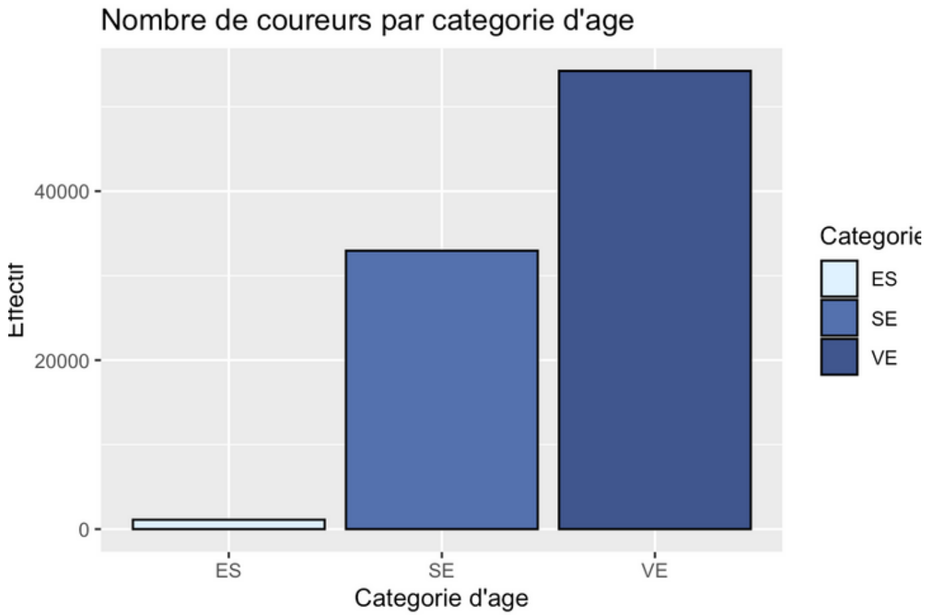
L'Age

Indicateurs de l'âge



Indicateurs	Age
Min.	20 ans
1er Qu.	34 ans
Mediane	42 ans
Moyenne	42 ans
3eme Qu.	49 ans
Max.	98 ans

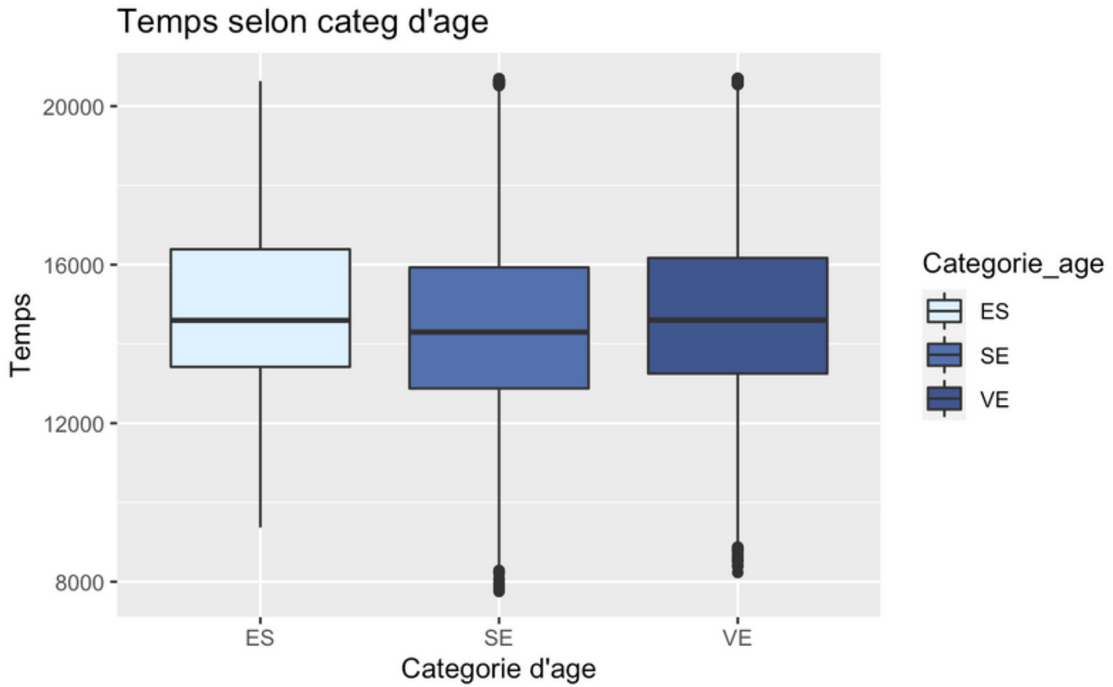
Catégories d'âge



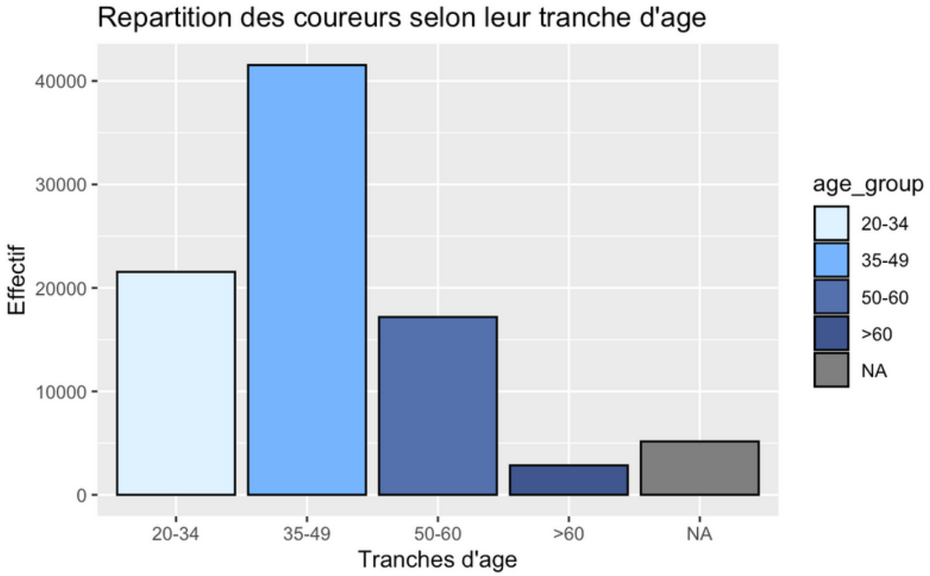
Interprétation : Dans le monde de l'athlétisme, il existe plusieurs catégories d'âge, appliquées par la réglementation de la fédération internationale de l'athlétisme et dans lesquelles sont aisément placés les coureurs.
Ici les coureurs font tous issus des catégories suivantes: ES, SE ou VE. Le graphique ci-contre nous montre la répartition des coureurs selon la catégorie d'âge, et il s'avère que les coureurs âgés entre 20 et 23 ans ont une participation minime par rapport aux coureurs âgés de plus de 24 ans. En effet la population des Masters s'élève à 54227 coureurs.

Interprétation : Le graphique ci-contre nous permet de comparer les temps de course moyens pour chacune des catégories d'âge. Nous pouvons voir qu'il y a des différences de valeurs, que peut nous confirmer le test statistique anova: celui-ci nous indique une p-valeur inférieure à 2e-16. Cela signifie qu'il y a bien une différence significative entre les différents groupes.

Cependant, ce que ous montrons ici ne nous permet pas de valider notre hypothèse: l'âge influe sur le temps de course. La construction de nos propres tranches d'âge est nécessaire.



Tranches d'âge

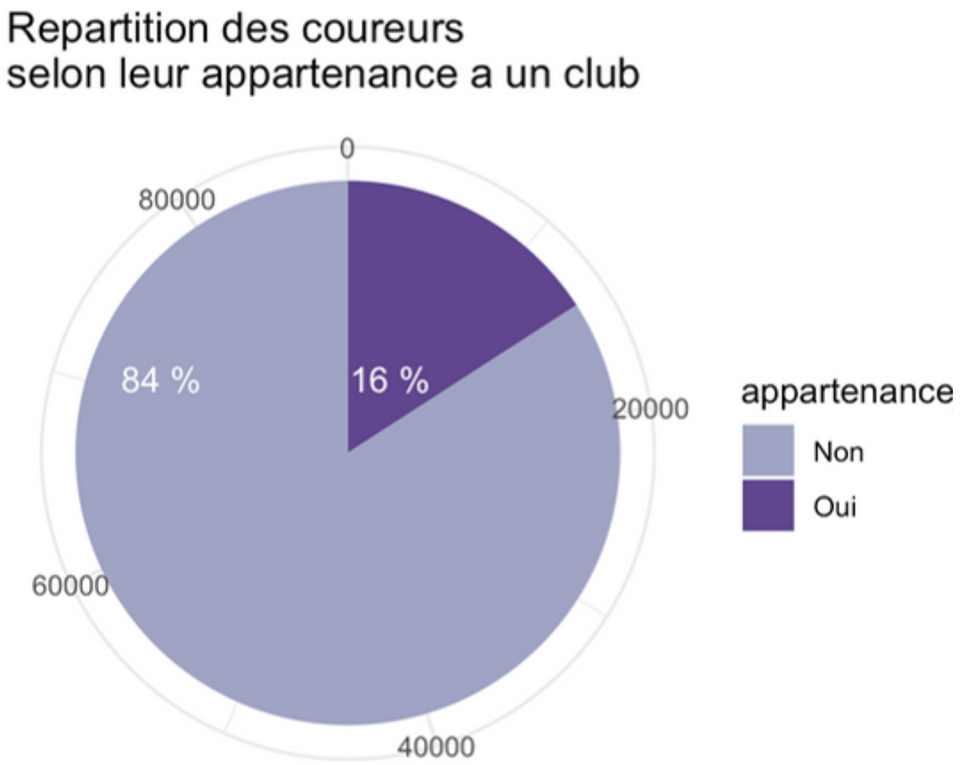


Interprétation : Cette nouvelle répartition des coureurs selon leur tranche d'âge dessine mieux le profil de notre population. Le groupe le plus grand en nombre reste celui des 35-49 ans. Cette fois-ci la plus petite est celle des plus de 60 ans.

En regroupant les catégories d'âge ES et SE, et en décomposant la catégorie d'âge VE , nous accompagnons l'âge avec les conditions physiques qu'il amène avec le temps; entre jeunesse et vieillesse.

Interprétation : Ainsi, ce graphique nous montre clairement un lien significatif entre l'âge des coureurs et la performance. En effet, nous constatons que plus les coureurs sont agés, plus le temps de course moyen est élevé! Donc naturellement, les coureurs les plus jeunes ont de meilleurs performances. Néanmoins, nous ne pouvons ignorer le facteur qu'est **l'expérience** : encore une fois, et bien qu'en plus grand nombre, les coureurs agés de 35 à 49 ans réalisent le meilleur temps ce course moyen de **4 heures**. La tranche d'âge des 20-34 ans n'est pas loin de ce score. A l'inverse, les coureurs les plus agés détiennent le pire temps de course moyen: **4 heures 21 minutes**.

L'appartenance à un club



Interprétation : Le graphique ci-contre nous donne la répartition des coureurs selon leur appartenance à un club ou non. Ainsi, la part de coureurs professionnels, du moins inscrits à un clubs n'est que de **16%,** soit **14 010** ! Alors que les coureurs "libres", donc ceux n'appartenant à aucun club, représentent **84% la population.**

Une fois de plus, nous pouvons affirmer qu'un coureur ne doit pas nécessairement adhérer à un club d'athlétisme: un brassard, une appétence pour la course à pied (ou pas) et le goût du challenge suffit.

L'appartenance à un club et la performance

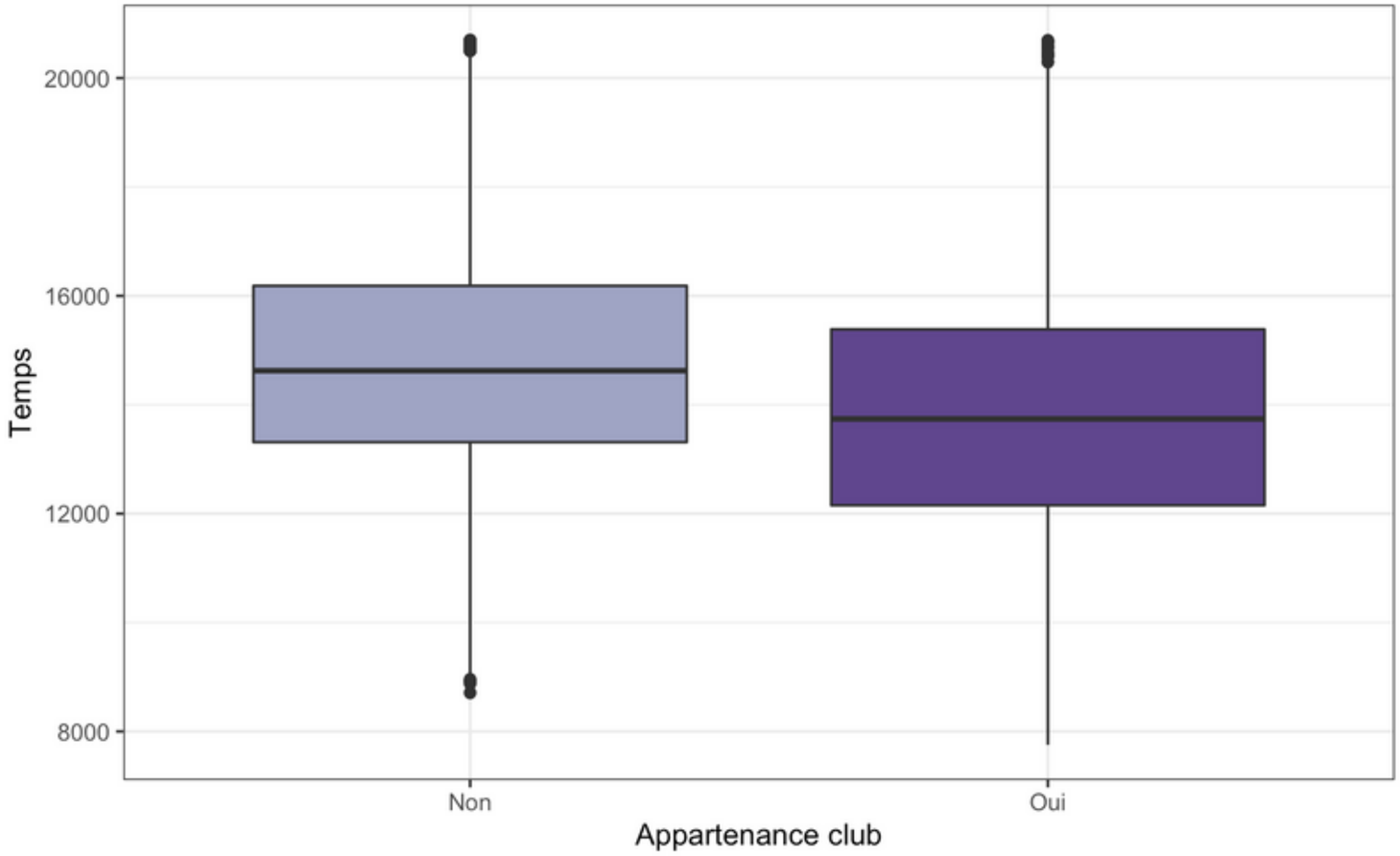
Interprétation : L'opposition des indicateurs du temps de course des coureurs pour chacun des deux groupes (adhérents et non-adhérents) nous permet de voir que l'appartenance à un club ou non a bien une influence sur la performance d'un coureur.

En effet: le temps de course moyen d'un coureur professionnel est de

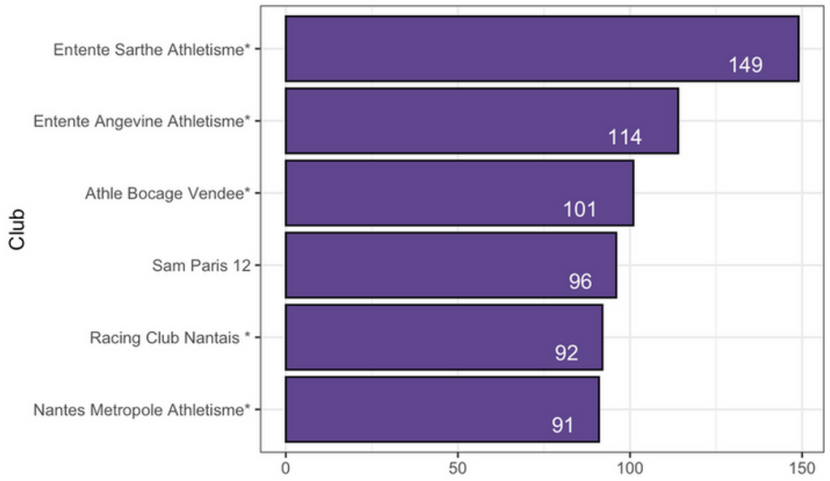
3heures 49minutes 55 secondes,

Tandis que celui d'un coureur non-professionnel est de

4 heures 5 minutes 56 secondes



Les six clubs les plus populaires



Interprétation : C'est donc le club **Entente Sarthe Athletisme** qui compte le plus grand nombre de coureurs de marathon professionnels en son club, avec 149 adhérents. Est-ce que cela fait-il de lui le meilleur club ? A l'inverse, le club Nantes Metropole

La base de données réunit 1418 clubs d'athlétisme en France.

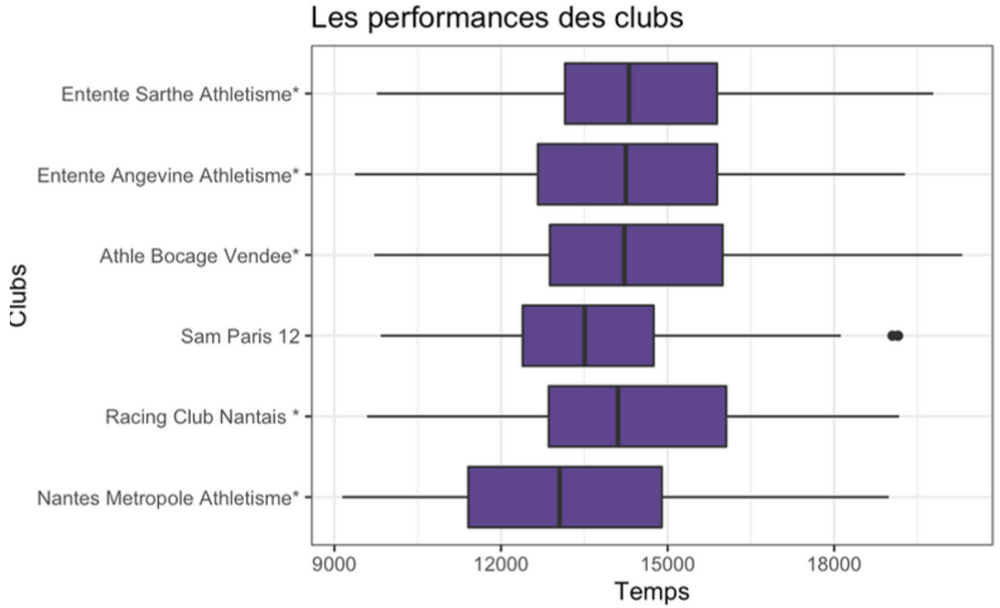
Interprétation : Le graphique ci-contre nous permet de comparer tous les indicateurs du temps en fonction des clubs. Ainsi nous tenons le **meilleur club**, soit le club auquel le coureur ayant réalisé le **meilleur temps de l'année** appartient: **Nantes Métropole Athlétisme !**

À vrai dire s'il mérite réellement ce titre, c'est parce qu'il réalise aussi le **meilleur temps moyen**: **3heures, 37 minutes, 42 secondes**

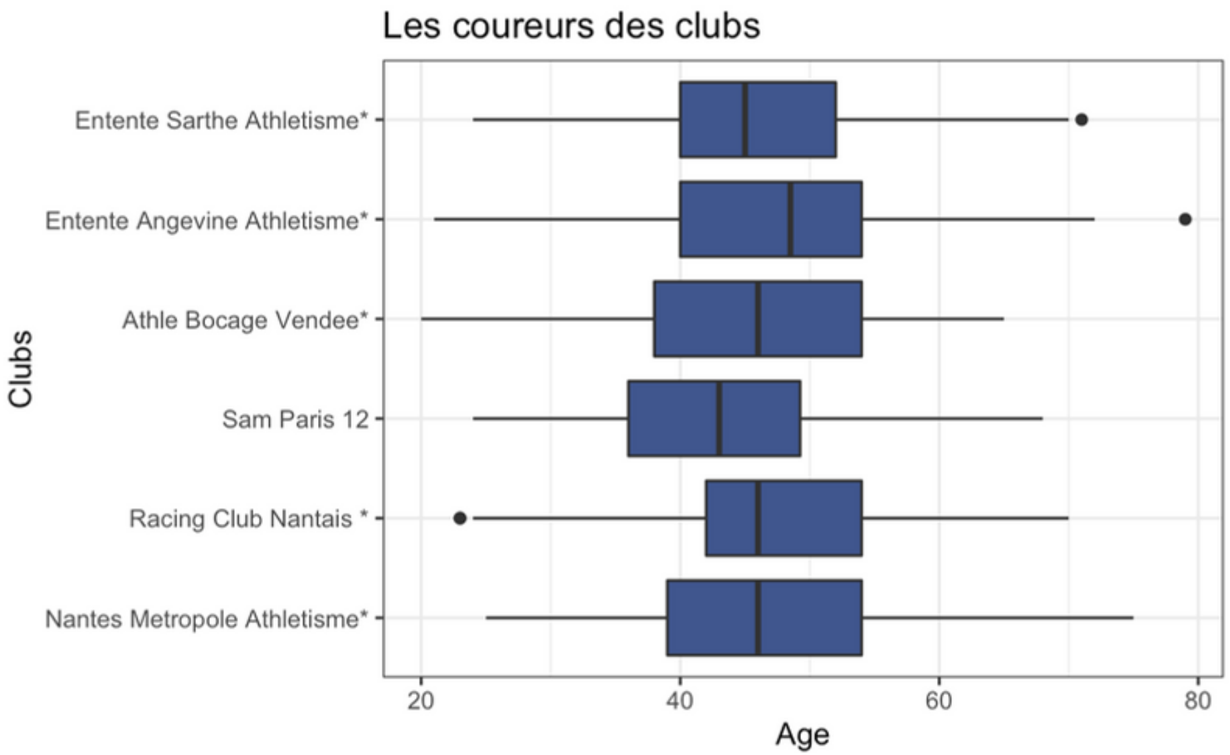
Nous pouvons aussi décerner un titre au club **Sam Paris 12**, qui détient une moyenne de temps de: **3heures, 48 minutes, 24 secondes**

En revanche, les 4 autres clubs semblent avoir la même moyenne soit autour de 14 300 secondes, soit : 3heures, 58 minutes, 20 secondes

Finalement, les moyennes varient seulement d'une dizaine de minutes !



Les coureurs des clubs

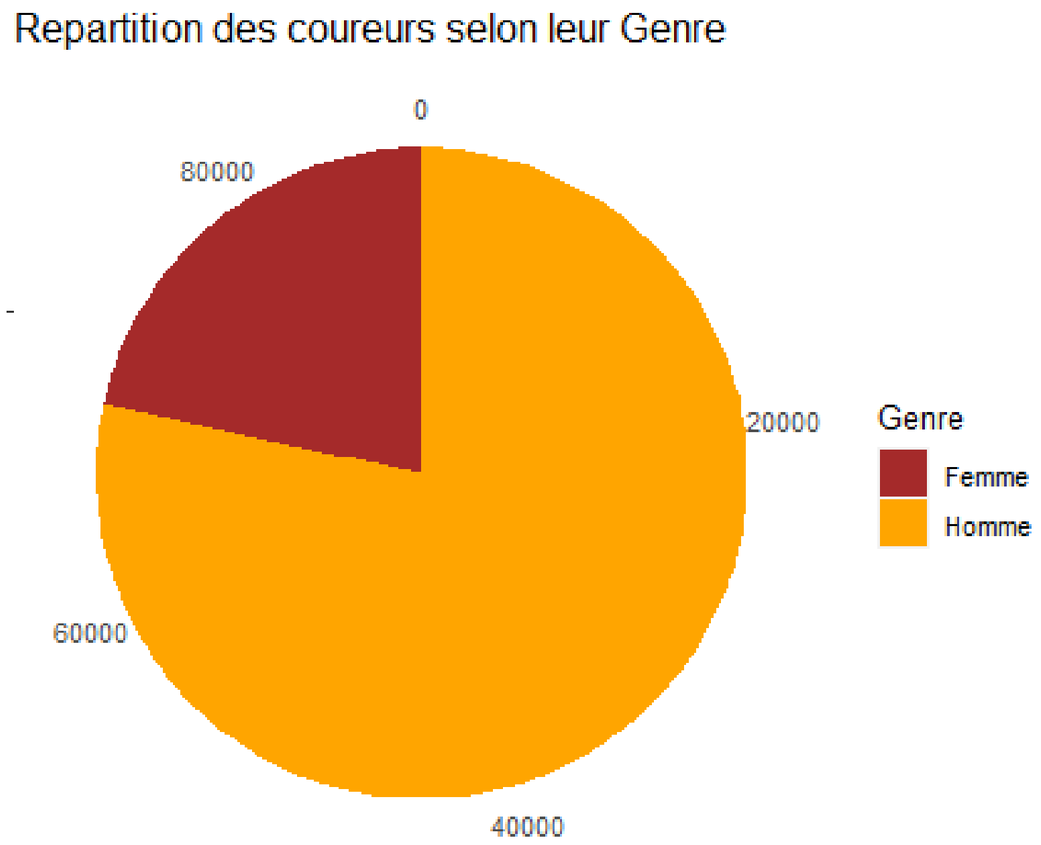


Interprétation : Il reste encore à déterminer, et ce toujours dans le cadre de l'étude statistique, si la performance des coureurs dépend réellement de leur club, ou bien d'un autre facteur dissimulé (de confusion), comme **l'âge**.

Effectivement, il est possible qu'un club détienne un meilleur temps moyen non pas du fait qu'il ait les **meilleurs coachs**, mais plutôt du fait du **jeune âge de l'ensemble de ses adhérents**. Par ailleurs nous avons précédemment vu que **l'âge** d'un coureur a une **influence** sur sa performance, cette hypothèse ayant été appuyée par des tests statistiques.

C'est pourquoi afin de lever cette interrogation, nous avons calculé les **indicateurs de l'âge des coureurs pour chaque club**: le graphique ci-contre nous permet de les comparer. Ainsi, le club de **Nantes Métropole Athlétisme**, bien qu'ayant le meilleur temps de course moyen, ne détient pas la plus faible moyenne d'âge: c'est le club **Sam Paris 12** avec une moyenne de **43 ans**.

Le genre

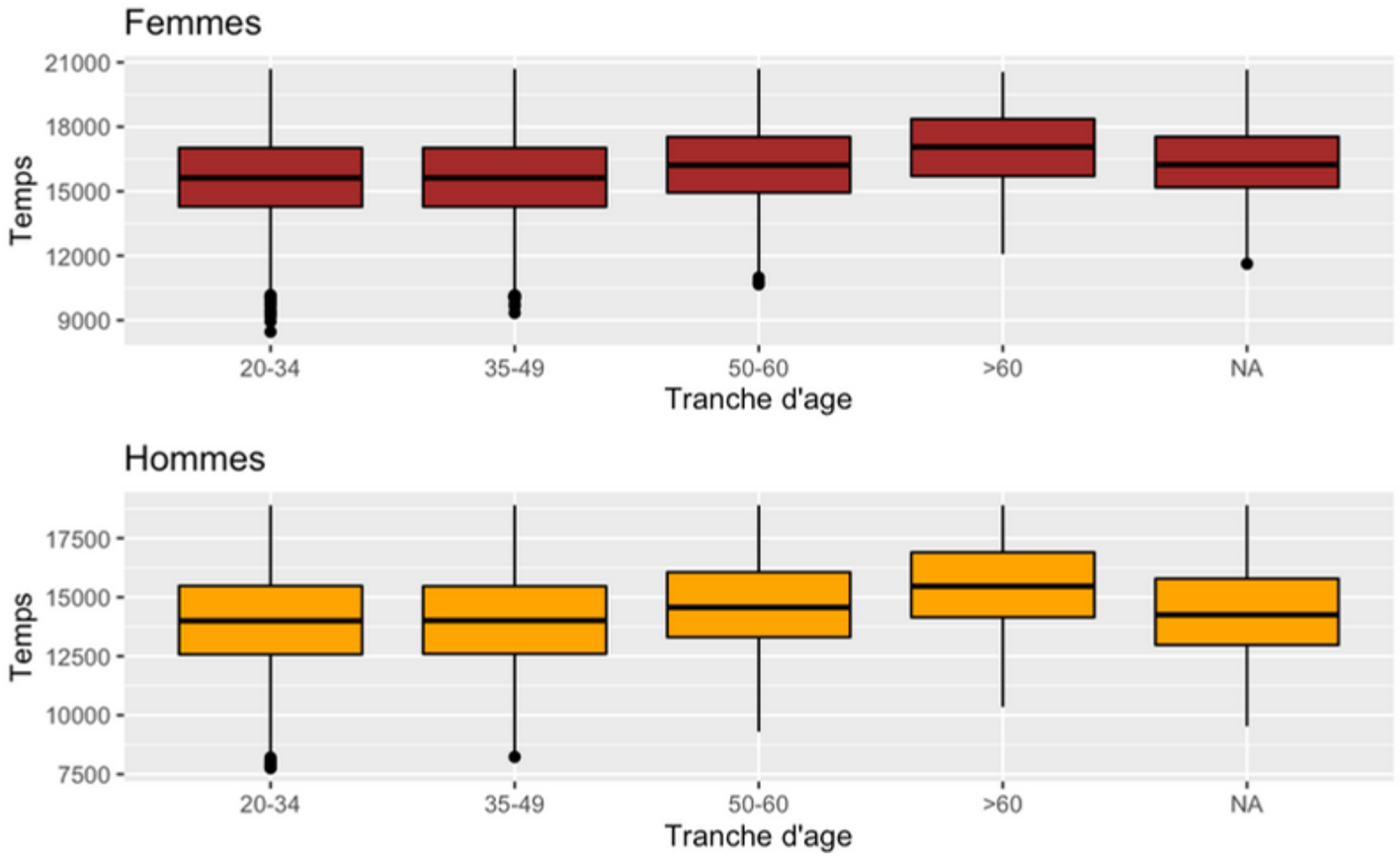


Interprétation : On constate directement que les hommes participent beaucoup plus aux marathons que les femmes . Il y'a environ **4 fois plus d'hommes.**

Les différences de physiques entre les deux genres rendraient absurdes une comparaison en terme de performance.

C'est pourquoi nous analyserons non pas les performances des coureurs en fonction de leur sexe, mais plutôt l'influence des différents facteurs socio-démographiques sur la performance, chez les hommes et chez les femmes.

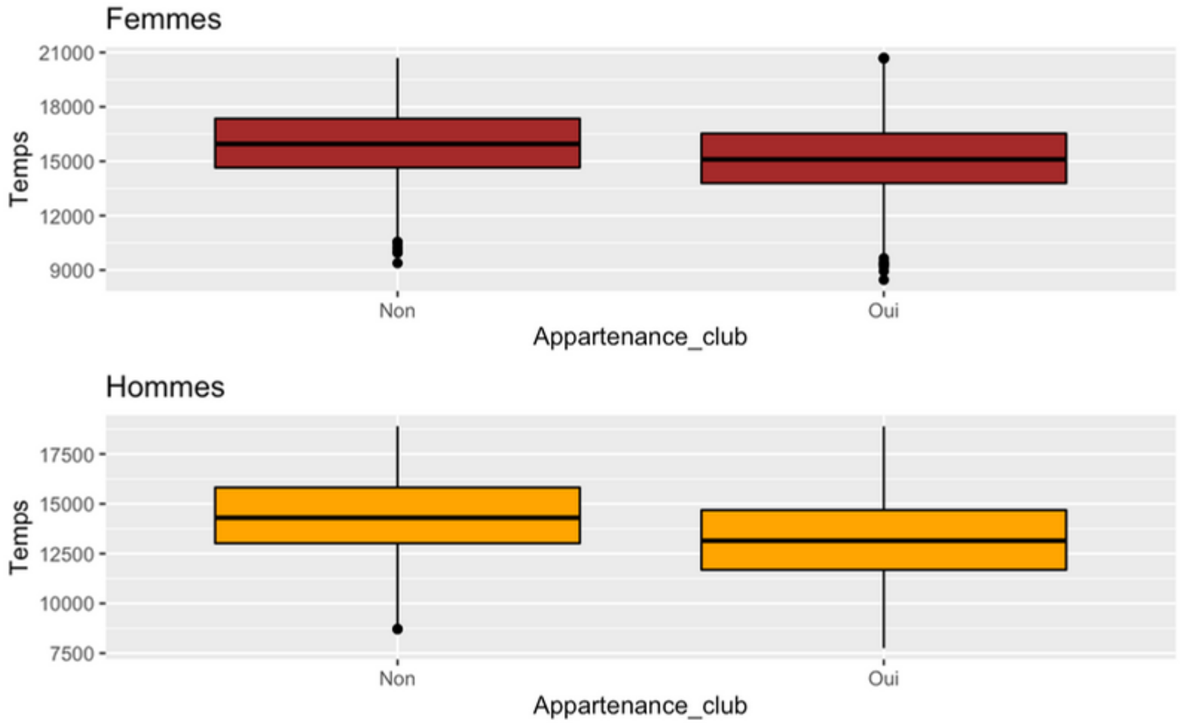
Âge et performances en fonction du genre



Interprétation : **Les hommes** comme les femmes suivent à peu près le même schéma ; c'est-à-dire plus on est avancé en âge plus en met de temps à finir le marathon et donc moins on est performant.

Cependant, on remarque que les performances sont quasiment les mêmes **entre 20 ans et 49ans** .

Le genre et l'appartenance au Club



Interprétation : Comme vu précédemment les personnes **appartenant à un club** sont **moins nombreuses** et réalise de **meilleurs performances**

2. FACTEURS MÉTÉOROLOGIQUES

Variables

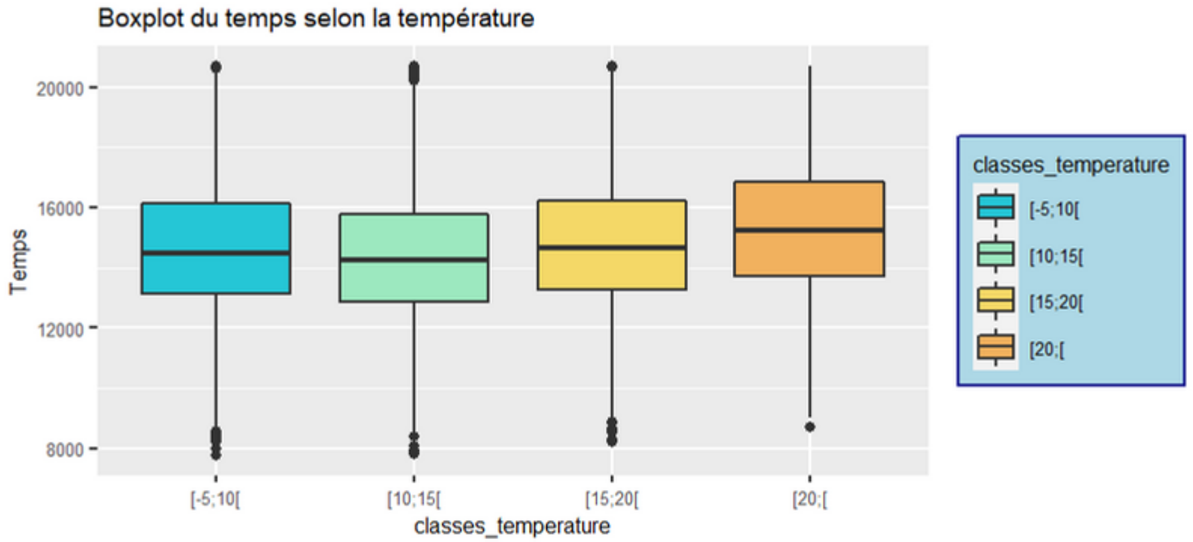
classes_temperature *variable quantitative continue*	**[-5;5[**	**[5;15[**	**[15;25[**	**[25;35[**
effectifs	2461	63635	22104	63

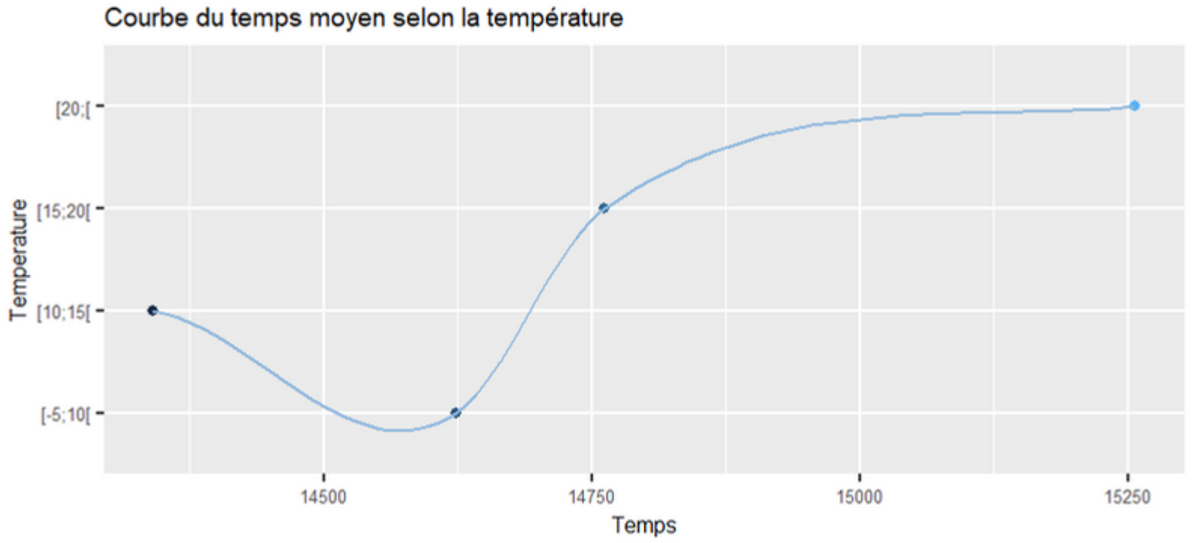
temps *variable quantitative continue*	**Min**	**Median**	**Mean**	**Max**
valeurs	-5.00	10.49	11.48	30.00

classes_humidité : catégorielle
température : quantitative continue
humidité : quantitative continue
saison : qualitative nominale
mois : qualitative ordinale

Informations : les données relatives à la météorologie ont étés récoltés manuellement sur https://tcktcktck.org

Les températures



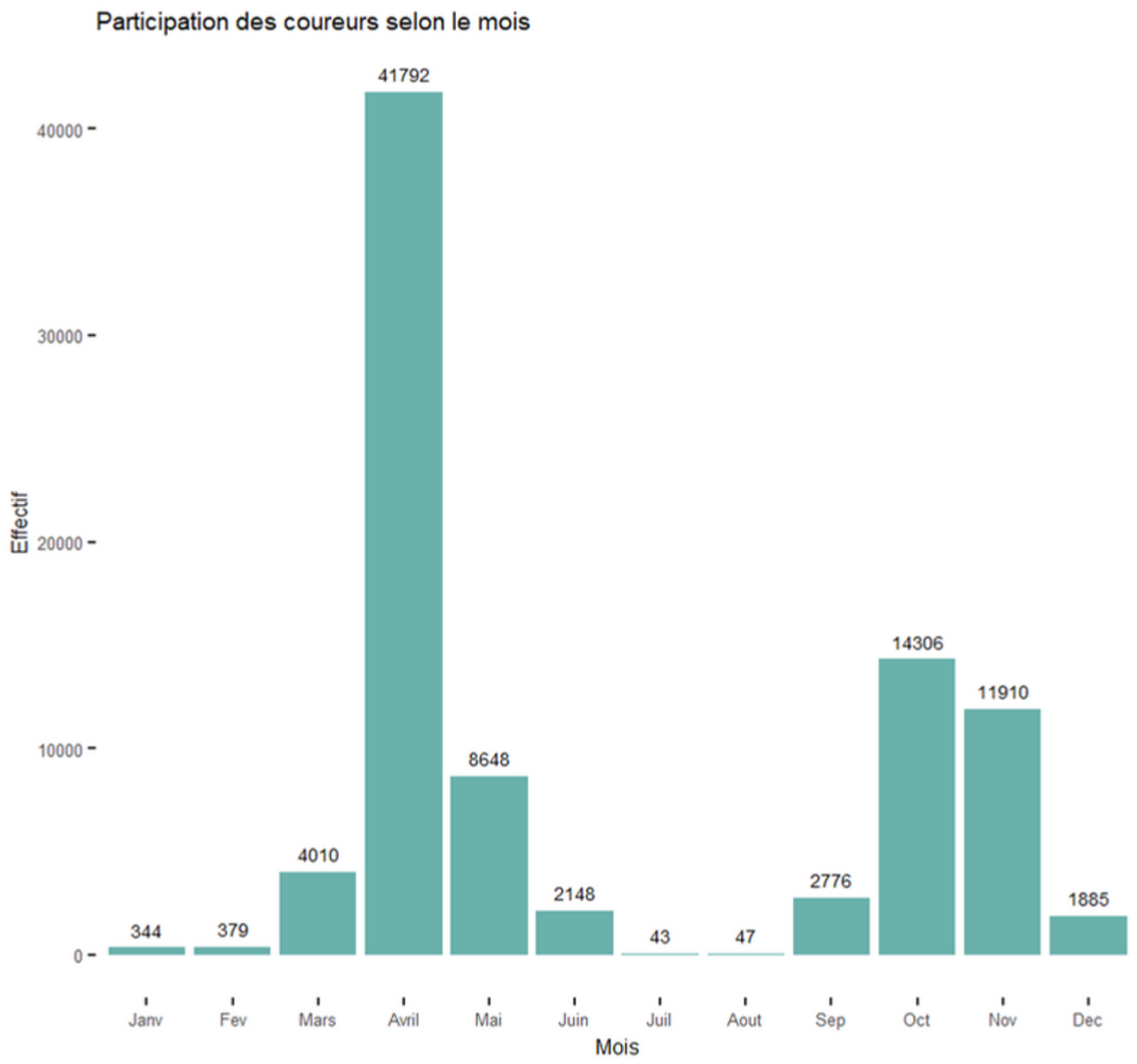


Interprétation : Entre -5 et 10°C, le temps de course moyen est de 14624 secondes soit 4:03:44 heures. Entre 10 et 15°C, il est à 14 341 secondes, soit 3:59:01. Entre 15 et 20°C, on enregistre 14 762 secondes, soit 4:06:02. Et à plus de 20°C, on est à 15 257 secondes, soit 4:14:17.

On observe que quand les températures dépassent 15°C le temps moyen est plus élevé. Les **températures idéales pour un marathon sont celles froides**, et plus particulièrement entre 10 et 15°C.

Test ANOVA : Pour étudier une variable catégorielle et une quantitative, on utilise le test ANOVA à un facteur, qui a pour but de comparer les moyennes d'échantillons.
La p-valeur obtenue est **<2e-16**, ce qui signifie que **le temps et la température sont très significativement liées.**

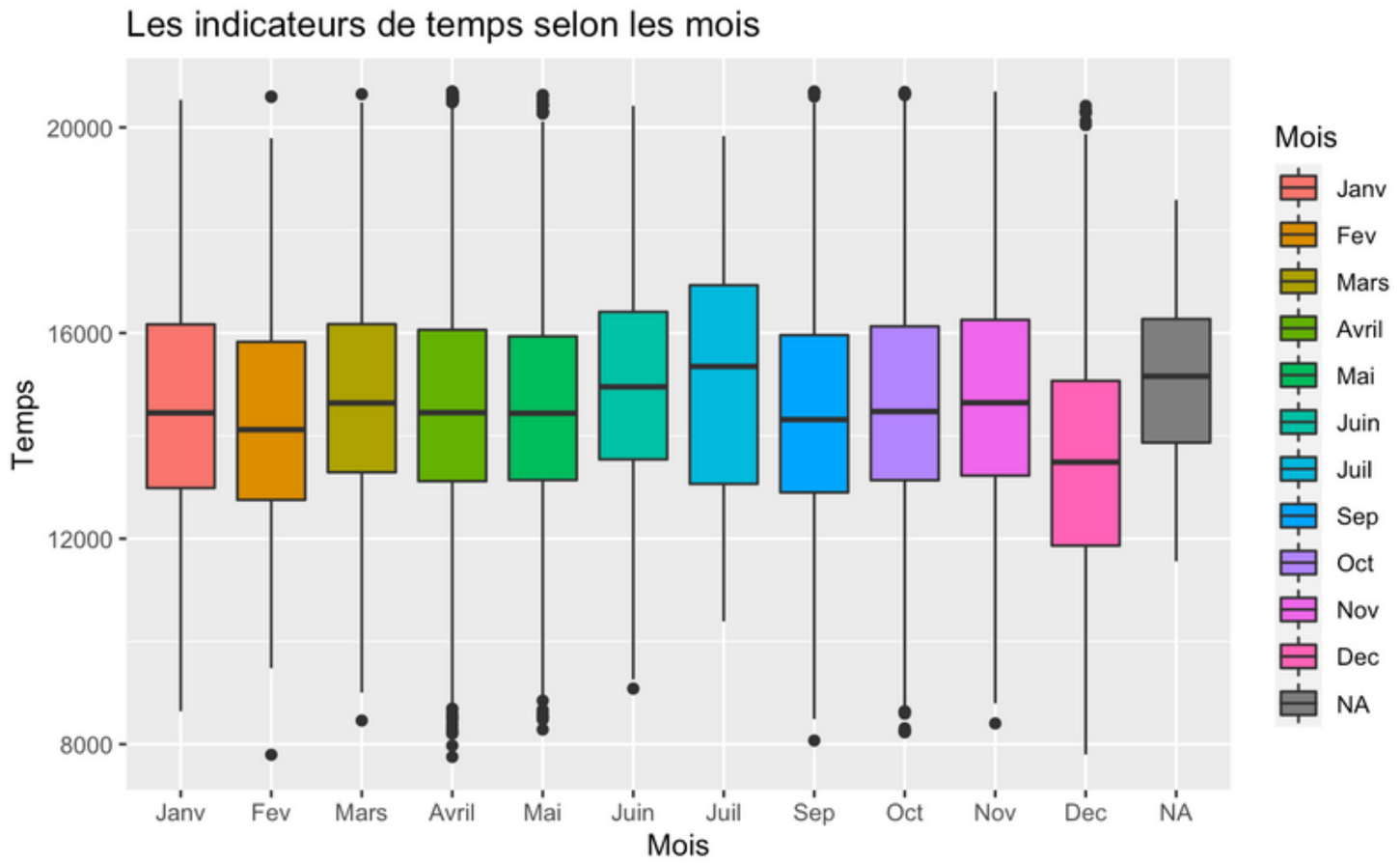
Les mois



Interprétation : On constate que la participation des coureurs atteint **un point culminant de 41 792 coureurs au mois d'avril.** Les mois d'octobre et de novembre ont respectivement, 14306 et 11910 participations.

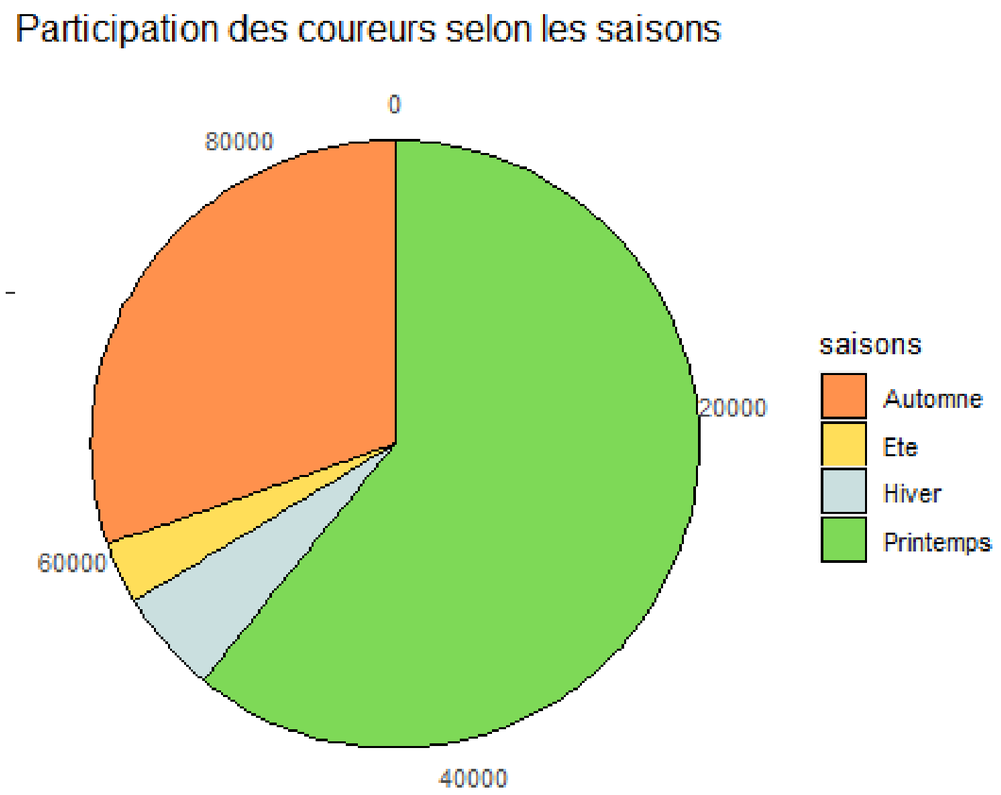
En revanche, les mois de juillet et d'août correspondent aux mois où le moins de coureurs participent, avec 43 et 47 participations chacun.

Les mois et le temps



Interprétation : En se focalisant sur les moyennes, on remarque que les mois où les meilleurs temps ont été effectués en **décembre, février et août**.
Cependant **la répartition est très inégalitaire entre les mois**. En effet pendant le mois d'**avril**, sont recensés **47 172 marathoniens soit 47,33 % de la population** tandis qu'en d**écembre n'ont été enregistré que 1885 coureurs soit 2,21 %**

Les saisons



Interprétation : Avec environ 50 % des coureurs en avril, il est évident que la participation soit la plus élevée se déroule **au printemps avec 52 588 participants**. Suivi de **22 216 participations en automne**.

En revanche, en **été et en hiver, il y a le moins de participation**. Cela est sûrement dû au fait que les températures de ces saisons sont jugées trop fortes pour l'été ou trop faible en hiver.

3. FACTEURS GÉOGRAPHIQUES

Variables

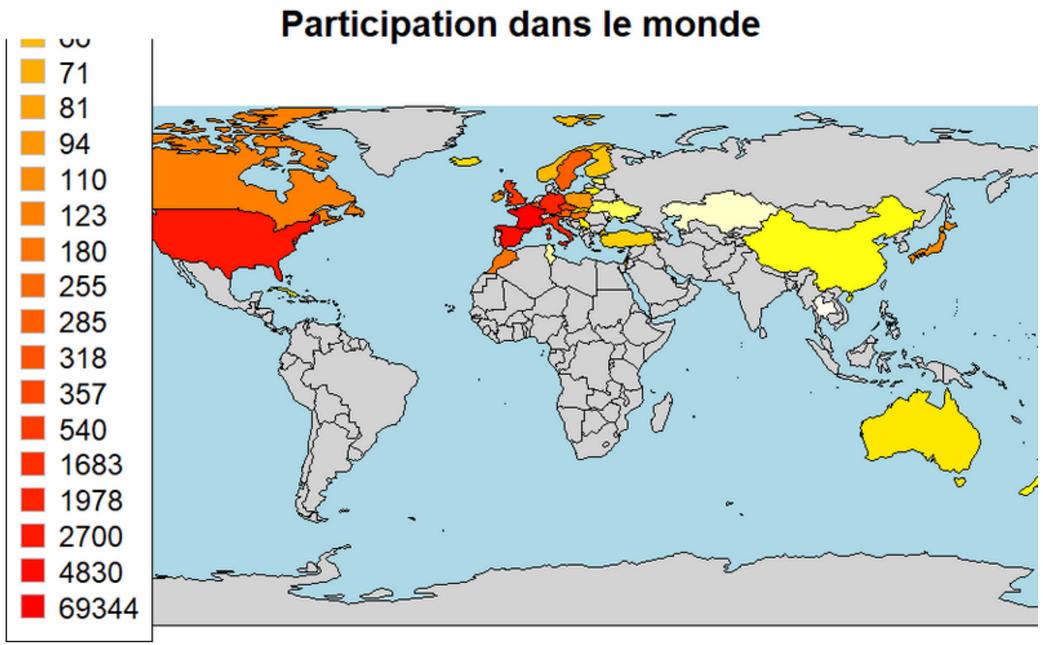
temps *variable quantitative continue*	**Min**	**Median**	**Mean**	**Max**
valeurs	-5.00	10.49	11.48	30.00

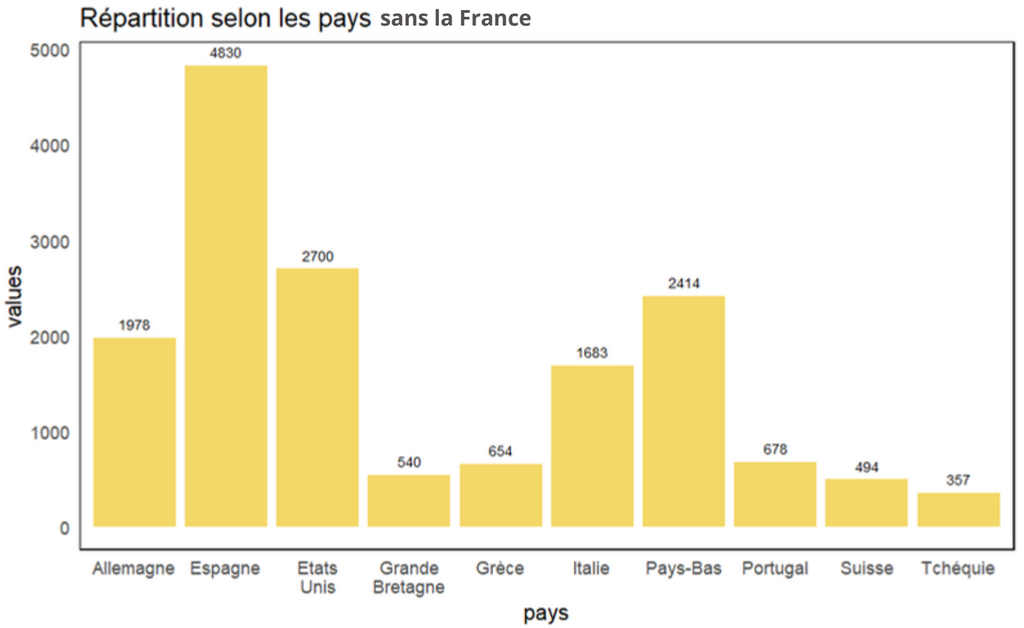
⊕

pays : qualitative nominale

régions : qualitative nominale

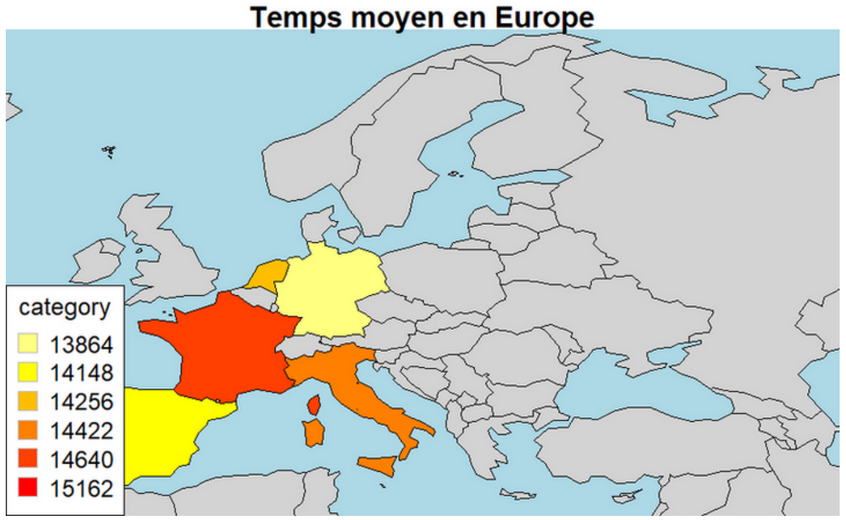
Les pays

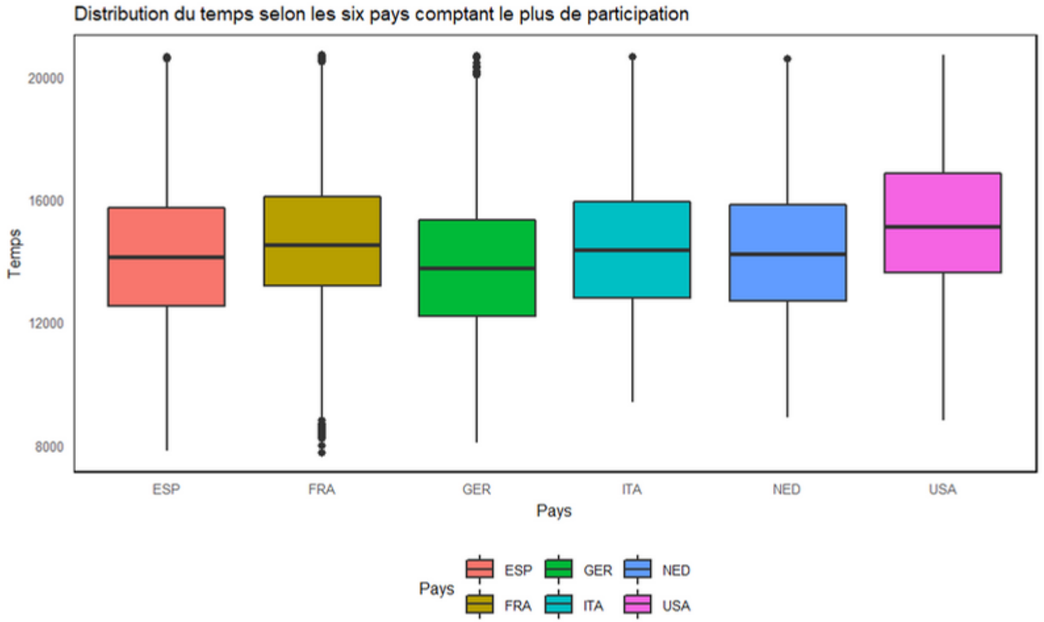


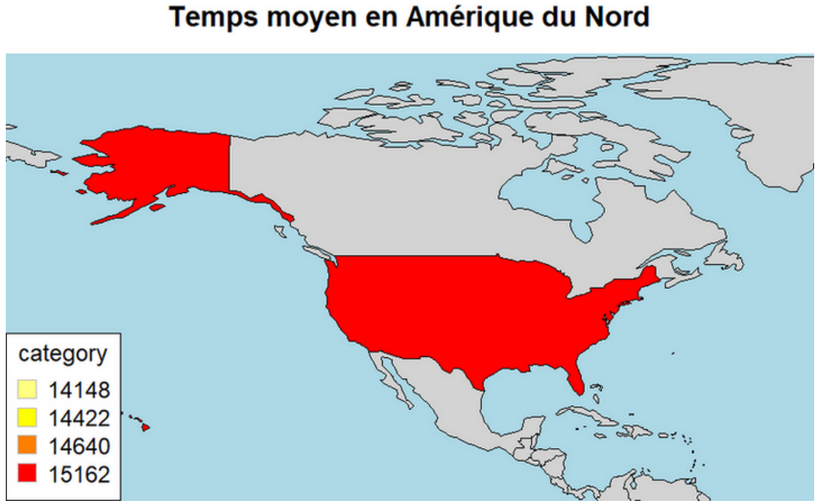


Interprétation : A part en France où il y a eu 69344 participations, les pays où les coureurs ont le plus participé sont l'**Espagne, les Etats-Unis, les Pays-Bas et l'Allemagne**, avec respectivement, 4830, 2700, 2414 et 1976.
En effet, les régions du monde cumulant le plus de participations sont l'**Amérique du Nord et l'Europe**. A l'inverse, la participation est faible en Asie et en Océanie. Les régions où il n'y a presque pas de participation sont représentés en gris sur la carte, à savoir l'Amérique du Sud, l'Afrique et une partie de l'Asie.

Les 6 pays comptant le plus de participations et le temps

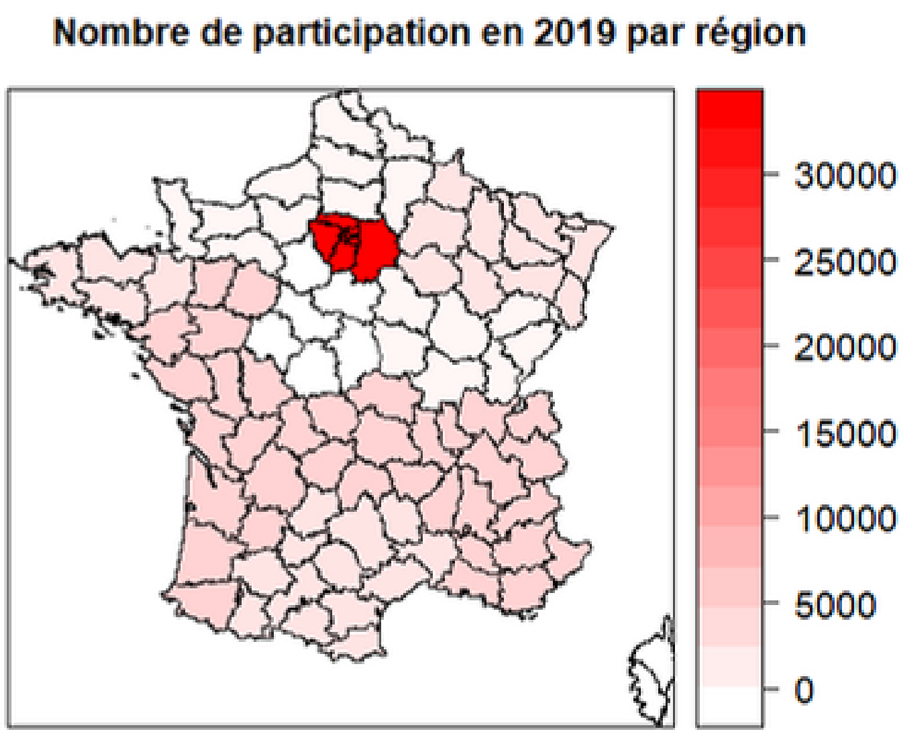


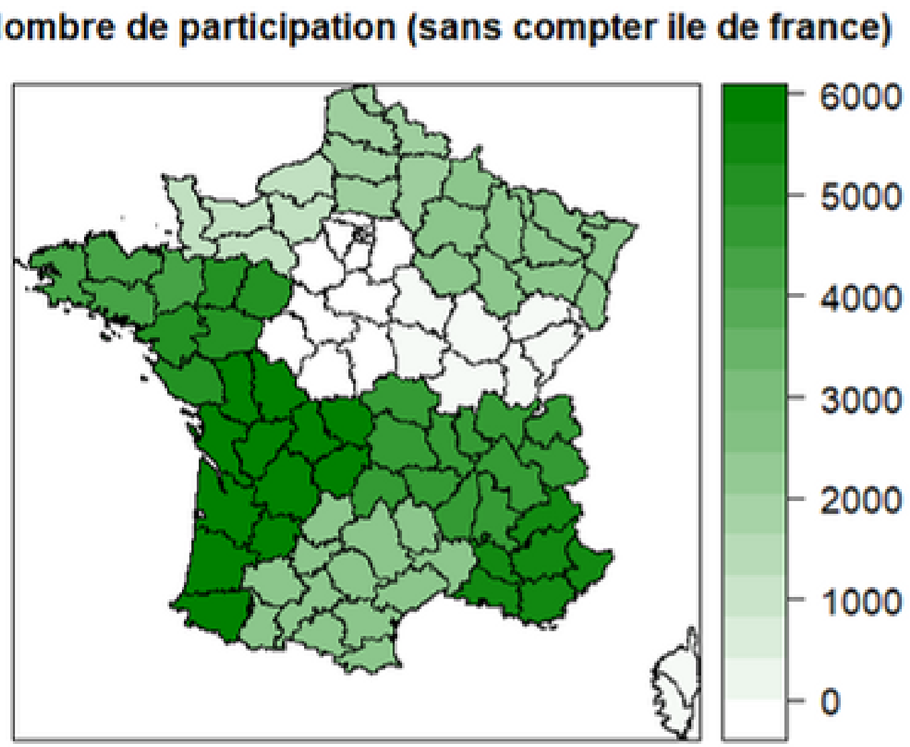




Interprétation : Dans la carte du monde que l'on voit à droite, les 6 pays cumulant le plus de participation sont représentés. En jaune, sont marqués les pays avec une **durée de course moyenne faible**. Ces pays sont **l'Espagne et l'Allemagne avec respectivement**, **13864 et 14148 secondes**. A l'inverse, la moyenne de temps est plus élevé aux Etats-Unis et en France avec 15162 et 14640 secondes.

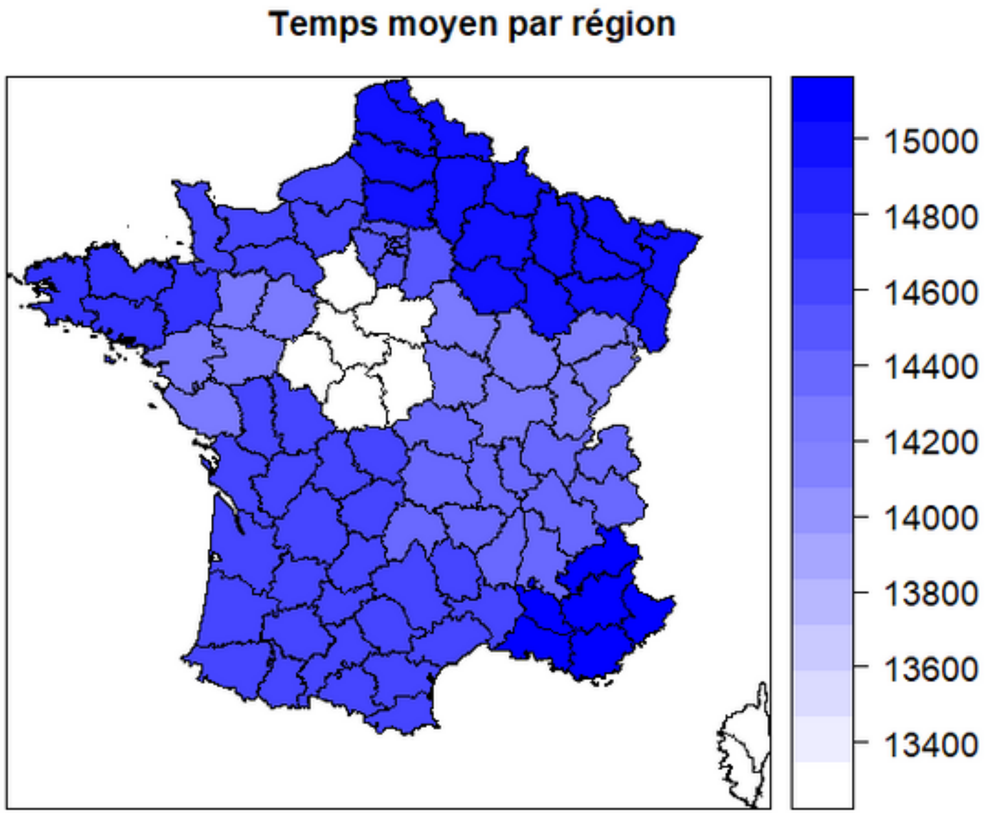
Les régions de France





Test ANOVA : Le nombre de participation est **très élevé en Ile-de-France avec 31725 participations**. Hors Ile-de-France, les régions où il y a le plus de participants sont **"Provence-Alpes-Côte-d'Azur", "Pays de la Loire", "Rhone-Alpes", "Aquitaine" et "Bretagne"** avec 5474, 5153, 4800, 4621 et 4221 participations.

Les régions et le temps



Interprétation : Les meilleurs temps se situent dans le centre est, l'ouest de la France et la Corse.

CONCLUSION

Des facteurs socio-démographiques, météorologiques et géographiques influencent-ils la performance des coureurs?

Nous devons naturellement conclure en apportant une réponse à notre problématique, mais aussi retrouver nos hypothèses et les rejeter si elles s'avèrent être fausses.

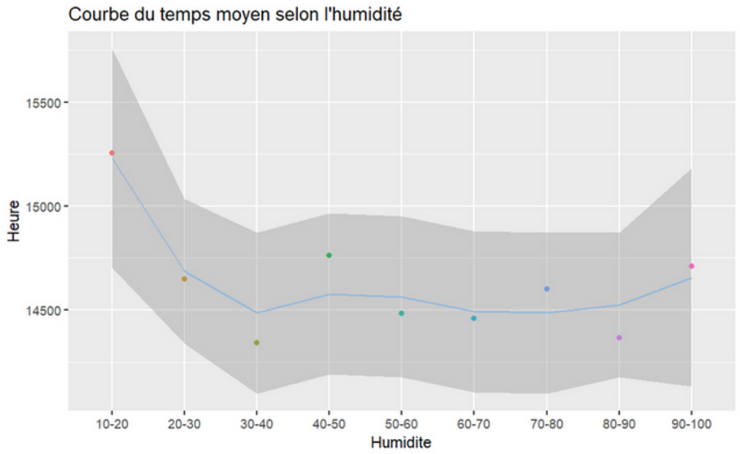
Tout au long de l'étude, nous avons pu constater que **des facteurs socio-démographiques, météorologiques et géographiques influençaient la performance des coureurs**. Ainsi, les hypothèses établies avant le lancement du plan d'analyse statistique peut être confirmées; à quelques détails près.

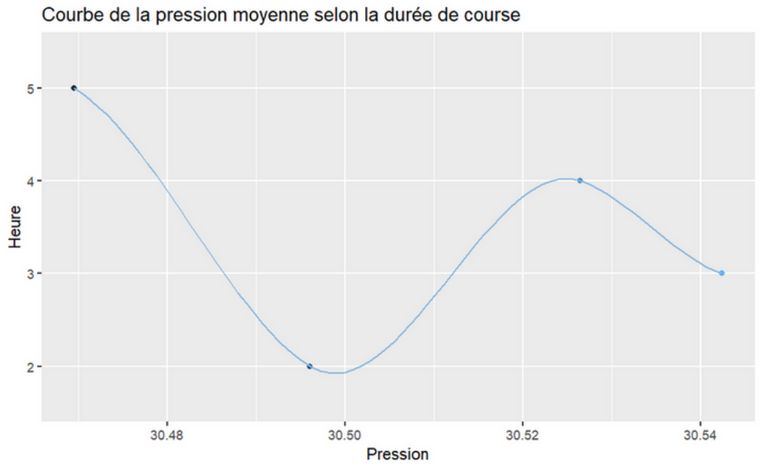
Au niveau des **facteurs sociodémographiques**, les coureurs les plus jeunes enregistrent de meilleurs performances (catégorie SE, entre 20 et 40 ans). Tout comme faire partie d'un club influence positivement la performance d'un coureur, et le nom du club quant à sa réputation.

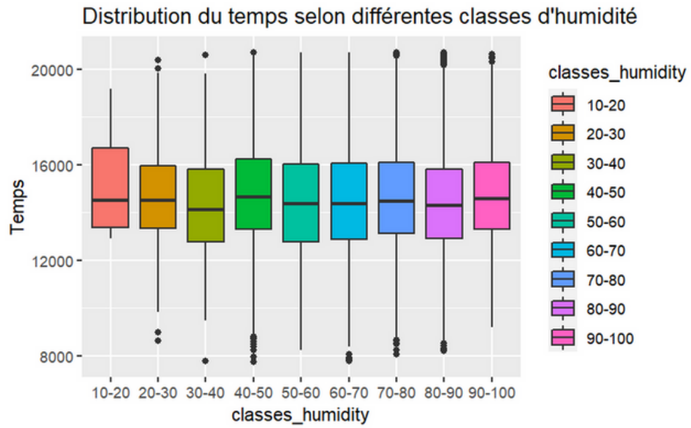
Au niveau des **facteurs météorologiques**, nous pouvons conclure que les températures froides sont optimales pour de meilleurs performances, et les meilleurs mois sont le mois de décembre, de février et d'août.

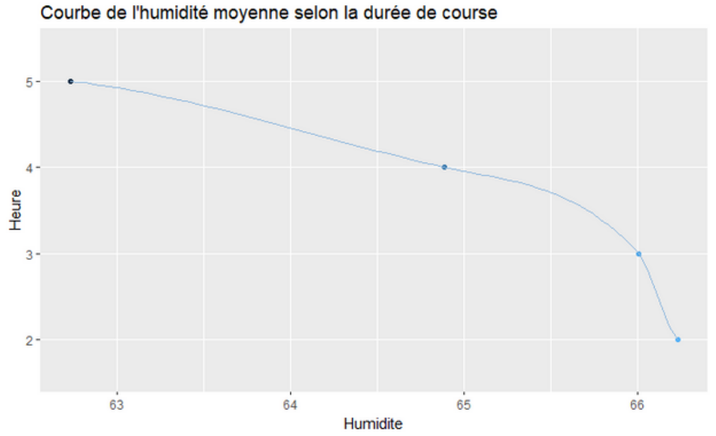
Au niveau des **facteurs géographiques**, les français enregistrent de meilleurs résultats en Espagne et en Allemagne. A l'échelle de la France, les meilleurs performances sont observés dans le centre-est.

Annexe


Courbe du temps moyen selon l'humidité
Heure
15500
15000
14500
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Humidite


Courbe de la pression moyenne selon la durée de course
Heure
5
4
3
2
30.48
30.50
30.52
30.54
Pression


Distribution du temps selon différentes classes d'humidité
Temps
20000
16000
12000
8000
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
classes_humidity
classes_humidity
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100


Courbe de l'humidité moyenne selon la durée de course
Heure
5
4
3
2
63
64
65
66
Humidite