Document explicatif de la démarche

J'ai décidé de travailler sur ces courses car non seulement le triathlon est une discipline qui m'intéresse beaucoup mais aussi car pour la saison 2023 en particulier les informations de courses étaient très complètes et regroupées de façon cohérente sur un même site (https://www.coachcox.co.uk).

Récupération des données :

Résultats des courses IronMan Nord América 2023 :

J'ai noté dans le fichier Excel "Races2023" les différentes informations sur les courses (nom, localisation, lien web vers les résultats, etc.).
Le site ne permettait pas d'exporter directement les résultats en format CSV, mais les données étaient présentes dans le code source de la page. Je les ai donc récupérées directement à partir du code source en utilisant les bibliothèques Python BeautifulSoup et JSON. Je les ai ensuite transformées en DataFrame et j'ai créé un dictionnaire qui regroupe tous les DataFrame.

Pour chaque course, j'ai récupéré les données qui me semblaient pertinentes pour l'analyse :
- Temps moyen
- Écart interquartile (IQR)
- Écart hommes-femmes
- Taux de finishers
- Répartition hommes-femmes
- ...

Données GPS :

Les fichiers GPX des courses étaient disponibles et téléchargeables sur les pages web des résultats. Je les ai donc directement téléchargés et enregistrés dans le dossier « GPX2023 ». Deux fichiers GPX n'étaient pas disponibles pour 2023, j'ai donc pris ceux de 2022 en partant du principe que les trajets n'étaient pas ou très peu modifiés d'une année sur l'autre (le profil de la course ne varie pas considérablement). Cela peut néanmoins constituer un biais dans l'analyse.

Les fichiers GPX ne contenaient pas les informations d'élévation, ce qui pose un problème pour l'analyse car le dénivelé est un facteur de performance important. J'ai donc estimé le dénivelé de chaque course en requêtant l'altitude de chaque point GPS via l'API « open-Elevation ». Les requêtes API étaient longues, donc j'ai enregistré les résultats dans des fichiers Excel que j'importais ensuite.

Les données étaient bruitées, j'ai donc appliqué un filtre passe-bas en ajustant les paramètres de manière empirique par essai-erreur.

J'ai vérifié mes données de dénivelé avec celles disponibles sur internet. Les résultats semblent cohérents, avec quelques différences, mais les profils de course sont respectés.

J'ai calculé les distances à partir des coordonnées GPS en utilisant la méthode de Haversine. Cela m'a permis de tracer les profils d'élévation des courses : Altitude vs Distance.

Données météo :

À partir de la date des courses et des coordonnées GPS du départ, j'ai obtenu des données météo via l'API « OpenWeatherMap ». J'ai fait confiance aux données de l'API sans vérifier, en raison de contraintes de temps.
Les données récoltées sont les suivantes :

- Température
- Pression
- Humidité
- Vent
- Précipitation
- ...

Toutes ces données ont alimenté le DataFrame « Races_info », qui regroupe ainsi toutes les informations pour toutes les courses de l'année 2023.

Analyse de corrélation

J'ai réalisé une analyse de corrélation sur les données de « Races_info », en resserrant progressivement l'analyse. On y remarque des corrélations importantes entre le dénivelé, l'humidité, et les écarts hommes, ainsi que les IQR.

C'est sous cet angle que je suis parti pour l'analyse.
J'ai effectué mon analyse sur ces paramètres

J'ai également remarqué que, sur une course, le temps moyen des femmes est plus faible que celui des hommes. J'ai effectué des analyses complémentaires pour essayer d'expliquer ce phénomène contre-intuitif.

Limites :

Il y a clairement des limites dans cette analyse. La principale réside dans le fait que seulement 10 courses sont étudiées, ce qui peut facilement introduire des erreurs dans les tendances observées, celles-ci servant de base à l'analyse. Avec un peu plus de temps, j'aurais récolté plus de données avant d'émettre des hypothèses, notamment dans le cadre de résultats allant parfois à l'encontre de ce que l'on trouve dans la littérature. Par exemple, statistiques sur plusieurs années, plusieurs régions….

Sur l'hypothèse des types de vélo utilisés, j'aimerais beaucoup disposer des données sur les types de vélo utilisés par les participants afin de pouvoir vérifier cette hypothèse.