Projet du module analyse de données

Niveau: Master I - Data Science

Date de soumission: ce projet est à rendre au plus tard le 19 juillet 2025 à minuit

---

> ***NB***: Une attention particulière sera portée à la clarté et à l'exactitude des réponses
> - [x] Le notebook doit être nommé de la manière suivante: **Prénom_Nom_ISI_M1_Projet.ipynb**
> - [x] Les **explications** doivent être écrites dans des cellules pour **commentaire (markdown)**
> - [x] Le **code** doit être **écrit et exécuté** dans des cellules pour **code**
> - [x] **Pas de réponse générée par une intelligence artificielle**
> - [x] **Le non-respect d'une de ces conditions entraîne une perte de points**
> - [x] Votre notebook correctement nommé, sera envoyé à **openacademy.education@gmail.com**

Description de la problèmatique (gestion des compagnies aériennes)

L'objectif est non seulement d'assurer la qualité et la fiabilité des données en appliquant un prétraitement rigoureux (**gestion des valeurs manquantes, détection et correction des valeurs aberrantes, mise à l'échelle commune des variables**, etc.), mais aussi d'analyser ce jeu de données de 336 776 vols pour extraire des indicateurs de performance clés sur les opérations aériennes. L'étude se concentrera principalement sur la ponctualité, en analysant les retards au départ et à l'arrivée pour identifier les tendances générales. Voici une description succincte du dataset:

- **id** : identifiant unique du vol.
- **year** : année du vol.
- **month** : mois du vol.
- **day** : jour du vol.
- **dep_time** : heure réelle de départ.
- **sched_dep_time** : heure de départ prévue.
- **dep_delay** : retard au départ en minutes.
- **arr_time** : heure réelle d'arrivée.
- **sched_arr_time** : heure d'arrivée prévue.
- **arr_delay** : retard à l'arrivée en minutes.
- **carrier** : code du transporteur aérien.
- **flight** : numéro de vol.

- **tailnum** : identifiant unique de l'aéronef utilisé pour le vol.
- **origin** : aéroport de départ.
- **dest** : aéroport de destination.
- **air_time** : durée du vol en minutes.
- **distance** : distance du vol en miles.
- **hour** : heure de départ prévue (heure).
- **minute** : heure de départ prévue (minute).
- **time_hour** : heure prévue de départ au format datetime.
- **name** : nom de la compagnie aérienne.

1. On considère que les valeurs manquantes sont de type **MNAR** (Missing Not At Random) pour les variables numériques, et de type **MCAR** (Missing Completely At Random) pour les variables qualitatives. Sélectionnez les méthodes d'imputation adéquates, puis appliquez-les correctement
2. Remplacer les valeurs aberrantes de façon à ce que les distributions de données ne soient pas modifiées
3. Donner la répartition des vols par compagnie aérienne. Quelle compagnie a le plus grand/ petit nombre de vols ?
4. Donner la répartition du nombre de vols retardés au depart par mois
5. Donner la répartition du nombre de vols retardés à l'arrivée par mois
6. Quelles sont les retards au cours de la journée ?
7. Quelle est la compagnie ayant accusé le plus de retard ?
8. Visualisez la répartition des retards moyens en fonction des différents modèles d'avions (tailnum). Quels modèles semblent être les plus sujets aux retards ?
9. Mettre les variables à une échelle commune si nécessaire, **sans modifier les distributions** de données.
10. Encoder la variable **origin** en sélectionnant une méthode adéquate
11. Sauvegarder toutes vos modifications

Bonne chance!