Projet : Classification / Analyse d'exoplanètes (NASA Exoplanet Archive)

Date limite de rendu des livrables : 05/06/2025 à 9h00

Objectifs

Utiliser un jeu de données contenant les propriétés de milliers d'exoplanètes pour répondre aux questions numérotées. Vous devrez rendre un notebook. Voici vos objectifs :

- **Explorer** les corrélations physiques entre variables
- **Classer** les exoplanètes en fonction de leurs caractéristiques
- **Prévoir** si une exoplanète est potentiellement habitable ou non

---

Évaluation

Les **soutenances** auront lieu le **jeudi 5 juin**.

Notation : 12 points qualité du code (6pts avancement + 4pts qualité de l'algorithmique + 2pts commentaires) + 8 points soutenance (2pts qualité des slides et de l'argumentation + 2pts travail d'équipe + 4pts qualité des graphiques).

Bonus : un ou plusieurs algorithmes de machine learning ont été mis en œuvre de manière pertinente. Leurs objectifs peuvent être de compléter des valeurs manquantes et / ou d'effectuer des tâches de classification.

Données : NASA Exoplanet Archive

La **NASA Exoplanet Archive** publie une base de données publique contenant :

- Données tabulaires sur > 5000 exoplanètes confirmées.
- Colonnes typiques :
    - pl_name (nom)
    - pl_bmassj (masse, en unité de Jupiter)
    - pl_radj (rayon, en unité de Jupiter)
    - pl_orbper (période orbitale)
    - pl_eqt (température d'équilibre)
    - pl_discmethod (méthode de détection : transit, vitesse radiale, etc.)
    - st_teff, st_rad, st_mass (paramètres de l'étoile hôte)
    - sy_dist (distance en parsecs)
    - …
- Disponible au format **CSV** ou via API.

🔗 Accès aux données :

- Téléchargement direct : **ici**
- Pour éventuellement compléter : NASA Exoplanet Archive (Filtered Table): https://exoplanetarchive.ipac.caltech.edu/docs/data.html

---

Premiers traitements / analyses à mener

1. Charger le fichier avec read_csv(). Vous devrez utiliser l'option skip_rows (voir la documentation de pandas) puis appliquer la fonction .head()
2. Extraire la liste du nom des colonnes du dataset.
3. Quelle est la taille du dataset ?
4. Renommer les colonnes pour qu'elles aient des noms plus explicites, si besoin.
5. Afficher le taux de remplissage de chaque colonne.
6. Repérer où sont les valeurs nulles et proposer une explication. Est-ce que ces valeurs n'existent pas ou est-ce qu'elle n'ont pas été remplies ?

Quelques comptes

7. Combien y a-t-il de planètes différentes découvertes ?
8. Combien y a-t-il de systèmes doubles ? (c'est-à-dire de systèmes de planètes orbitant autour de deux étoiles)
9. Y a-t-il des doublons ? Vous pourrez utiliser la fonction duplicated() pour répondre à cette question.

Analyse univariée et bivariée

10. Tracer la distribution de chaque indicateur (fonction histplot() de seaborn)
11. Tracer un boxplot pour chaque indicateur (fonction boxplot() de seaborn)
12. Calculer la matrice de corrélation et afficher là avec la fonction heatmap() de seaborn. Y a-t-il corrélation entre certaines variables ?

Vérification de la troisième loi de Kepler

D'après la troisième loi de Kepler, on a

$$\frac{T^2}{a^3} = \frac{4\pi^2}{G(M+m)}$$

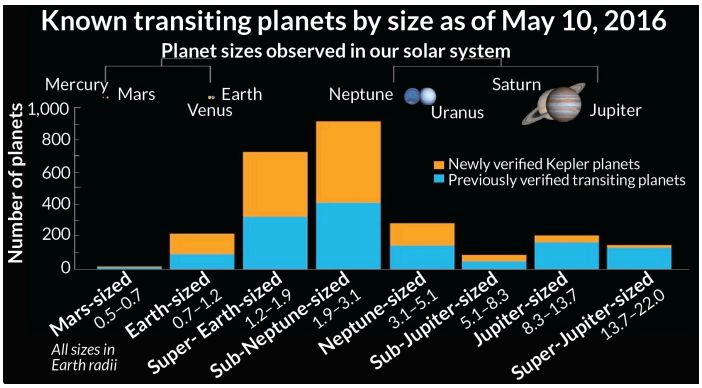
avec :

- *T* : période orbitale (pl_orbper en jour)
- *a* : demi-grand axe (pl_orbsmax en UA)
- *G* : constante de gravitation universelle valant 6.67 x 10^{-11} SI
- *M* : masse de l'étoile
- *m* : masse de la planète

13. La troisième loi de Kepler est-elle vérifiée dans le dataset ?

Analyse des exoplanètes par tailles

14. Tracer le graphique suivant :



Pour avoir les bons labels, il va falloir créer une nouvelle colonne "planet_type" que vous remplirez des labels "mars-sized", "earth-sized", … au préalable.

Analyse des exoplanètes par densité

15. Partout où c'est possible, calculez la densité de la planète avec la formule suivante, exprimant le rapport entre la masse de la planète et son volume :

$$\rho = \frac{M_p}{\frac{4}{3}\pi R_p^3}$$

avec :

- M_p : masse de la planète (en g)
- R_p : rayon de la planète (en cm)

Remarques :

- le nombre π s'obtient, par exemple, avec numpy.pi
- Attention, M_p et R_p doivent d'abord être converties en kg et en mètre pour le calcul de la densité !

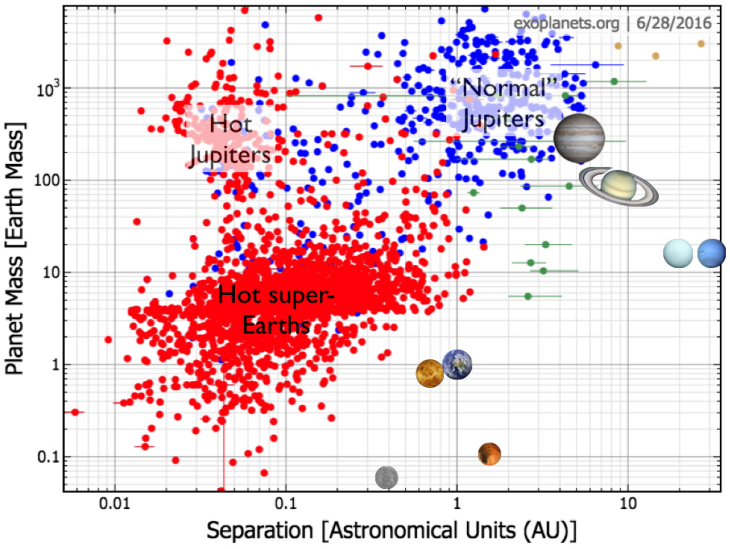
16. Remplissez une nouvelle colonne "composition" avec les labels "rocky" (rocheux) ou "gaseous" (gazeux) suivant la densité de la planète. Pour information, voici les données des planètes du Système Solaire et leurs labels :

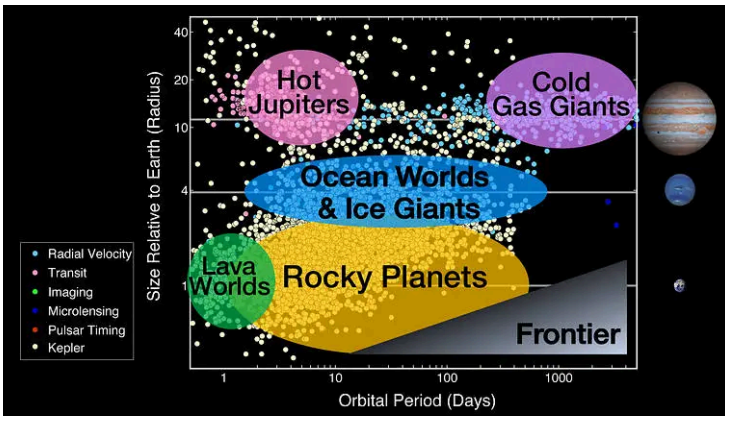
	Planète	Diamètre (km)	Rayon (R_Terre)	Masse (M_Terre)	Densité (g/cm^3)	Composition
1	Mercure	4879	0.382	0.055	**5.43**	**rocky**
2	Vénus	12103	0.949	0.815	**5.24**	**rocky**
3	Terre	12756	1.000	1.000	**5.52**	**rocky**
4	Mars	6794	0.532	0.107	**3.93**	**rocky**
5	Jupiter	142984	11.21	317.8	**1.33**	**gaseous**
6	Saturne	120536	9.433	95.16	**0.69**	**gaseous**
7	Uranus	51118	4.007	14.54	**1.27**	**gaseous**
8	Neptune	49528	3.883	17.15	**1.64**	**gaseous**

17. Quel critère utiliser pour remplir les labels sur la composition à partir des planètes du système solaire ? Vous pourrez utiliser un critère de seuil et un algorithme de classification supervisée, entraîné sur les données du système solaire, puis comparer les résultats.

Analyse des exoplanètes par rayon / masse et période orbitale

En colorisant vos graphiques à l'aide des colonnes "planet_type" et / ou "composition" (argument *hue* dans le scatterplot de seaborn), tracer des équivalents des graphiques suivant. Attention à l'échelle utilisée !





Autres analyses exploratoires possibles

Clusterisation (non supervisée)

- Utiliser K-Means pour découvrir des regroupements de planètes selon leur rayon, masse et période orbitale.

Classification selon la méthode de détection

- Entrée : masse, rayon, distance, température de l'étoile
- Sortie : pl_discmethod (méthode de détection)
- Modèles : KNN, SVM

Analyse des planètes potentiellement habitables

- Définir une zone habitable simple : température entre 200 K et 350 K, rayon entre 0.5 et 2 R_terre, etc.
- Créer une colonne binaire habitable (oui/non), puis entraîner un modèle de prédiction.

Quelques conseils

Tout au long de l'analyse, **produire des visualisations** afin de mieux comprendre les données. **Effectuer une analyse univariée** pour chaque variable intéressante, afin de synthétiser son comportement.

> L'appel à projets spécifie que l'analyse doit être simple à comprendre pour un public néophyte. **Soyez donc attentif à la lisibilité :** taille des textes, choix des couleurs, netteté suffisante, et variez les graphiques (boxplots, histogrammes, diagrammes circulaires, nuages de points…) pour illustrer au mieux votre propos.

3) **Confirmer ou infirmer les hypothèses à l'aide d'une analyse multivariée**. **Effectuer les tests statistiques appropriés** pour vérifier la significativité des résultats.

4) **Justifier votre idée d'application.** Identifier des arguments justifiant la faisabilité (ou non) de l'application à partir des données Open Food Facts.

5) **Présenter vos résultats** et **pitcher votre idée** durant la soutenance du projet.

Livrables

- Un **notebook du nettoyage** des données (non cleané, pour comprendre votre démarche).
- Un **notebook d'exploration** comportant une analyse univariée, multivariée, une réduction dimensionnelle, ainsi que les différentes questions de recherches associées (non cleané, pour comprendre votre démarche).

Soutenance (12mns présentation + 5mns questions)

Votre présentation doit contenir les éléments suivants :

- La présentation de votre idée d'application.
- La répartition des tâches (gestion d'équipe)
- Les opérations de nettoyage effectuées.
- La description et l'analyse univariée des différentes variables importantes avec les visualisations associées.
- 3 observations solidement étayées (graphes et/ou tests statistiques à l'appui au besoin) évaluant la pertinence et la faisabilité de votre application.

- La synthèse des différentes conclusions sur la faisabilité de votre projet.