
# D.U. Data Analyst

---

## Rapport de projet
## Rédaction de rapport d'analyse

---

**Auteur :** Nicolas Tsakalos
**Enseignant :** Matthieu Cisel
Août 2023

# 1 Introduction

Les étoiles qui constellent notre ciel nocturne ont toujours captivé l'humanité, suscitant une curiosité profonde quant à la possibilité d'autres mondes en dehors de notre système solaire. Au cours des dernières décennies, les avancées technologiques ont permis de transformer cette curiosité en une quête scientifique, aboutissant à la découverte et à l'étude des exoplanètes, des planètes qui orbitent autour d'étoiles autres que notre Soleil. Le recensement minutieux de ces exoplanètes, leurs propriétés et leurs étoiles hôtes a ouvert une fenêtre sur la diversité incroyable de systèmes planétaires dans l'univers.

Le jeu de données exhaustif provenant du site Kaggle, produit par la NASA et rassemblant les découvertes exoplanétaires effectuées jusqu'en 2021 ainsi que des caractéristiques associées aux planètes et à leurs étoiles, offre une occasion unique d'explorer et d'analyser ces nouveaux mondes. Cette étude vise à dévoiler les tendances et les relations cachées au sein de ces données, en mettant en évidence les propriétés qui pourraient être étroitement corrélées.

L'importance de cette recherche réside dans la compréhension accrue de l'incroyable variété des systèmes planétaires. L'analyse de ces exoplanètes révèle des informations cruciales sur les processus de formation et d'évolution des planètes, ainsi que sur les environnements et les conditions qui peuvent permettre l'émergence de la vie. De plus, l'étude de l'influence de la méthode de détection des planètes sur leurs propriétés pourrait apporter des éclaircissements sur la fiabilité de ces méthodes, essentielles à l'interprétation des résultats des missions d'observation et de détection.

Dans cette perspective, cette étude vise à répondre à la question fondamentale : dans le vaste éventail de données sur les exoplanètes, quelles sont les caractéristiques qui se recoupent et comment la méthode de détection de la planète peut-elle influencer les propriétés observées ? En particulier, on cherchera à déterminer si la méthode de détection et le type spectral de l'étoile hôte ont une influence fondamentale sur la masse des exoplanètes détectées Cette exploration approfondie non seulement révélera des liens entre les grandeurs mesurées, mais permettra également de savoir quelle méthode privilégier et quelle type spectral d'étoile pointer afin d'augmenter les possibilités de trouver des exoplanètes dont les caractéristiques se rapprochent le plus de la Terre, seule planète actuellement dont on sait qu'elle héberge diverse forme de vie.

# 2 Méthodologie

Avant d'entreprendre des analyses plus poussées, on effectue une exploration initiale des données. Cela comprend la familiarisation avec les différentes colonnes du jeu de données, l'identification des types de variables (quantitatives ou catégoriques) et la détection d'éventuelles valeurs manquantes ou aberrantes.

Pour identifier les relations potentielles entre les différentes propriétés des exoplanètes et de leurs étoiles, on effectue une analyse des corrélations entre variable quantitative. On tracera le diagramme de corrélation entre ces différentes variables, puis on effectuera les cas échéants des régression linéaires dont on représentera les courbes correspondantes. Cela permettra de mettre en évidence les associations positives ou négatives entre les variables et éventuellement d'établir des relations entre elles. Cette étape permettra de déterminer si certaines caractéristiques sont intimement liées.

Étant donné que les exoplanètes sont détectées à l'aide de diverses méthodes, il est essentiel de catégoriser les données en fonction de la méthode de détection spécifique utilisée. Cela permettra d'analyser comment la méthode de détection pourrait influencer les propriétés observées des planètes. On ne prendra en considération que les principales méthodes de détection, celles qui ont permis de détecter le plus d'exoplanètes, et on se concentrera sur les propriétés des exoplanètes les plus élémentaires.

Une fois les données catégorisées en fonction des méthodes de détection, on comparera les propriétés des exoplanètes et de leurs étoiles entre les différents groupes. On utilisera des analyses statistiques appropriées, telles que des tests ANOVA, pour déterminer s'il existe des différences significatives dans les caractéristiques des planètes détectées par différentes méthodes.

Les résultats obtenus permettront de déterminer quelles méthodes de détection sélectionner selon les objectifs désirés, notamment pour développer la recherche de planète semblables à la Terre, pouvant potentiellement abriter des formes de vie extra terrestre.

# 3 Résultats

Le jeu de données utilisé provient de la plateforme web Kaggle, qui fournit des jeux de données aux scientifiques pour leurs besoins dans la réalisation de projet en analyse de données. Il est disponible à l'adresse suivante `https://www.kaggle.com/datasets/shivamb/all-exoplanets-dataset/download?datasetVersionNumber=1`. Il a été produit grâce aux archives des exoplanètes de la NASA `http://exoplanetarchive.ipac.caltech.edu/`.

## 3.1 Exploration des données

L'exploration des données permet d'obtenir un premier aperçu des caractéristiques du jeu de données recensant les exoplanètes découvertes par la NASA jusqu'en 2021, ainsi que les propriétés associées à ces planètes et à leurs étoiles. Le jeu de données est composé de 4575 lignes et 23 colonnes. Chacune des lignes correspond à une exoplanète et chacune de ces colonnes représente une variable spécifique. Voici quelques éléments saillants issus de l'exploration des données :

Les variables étudiées dans le jeu de données sont regroupés dans le tableau 1. Elles recouvrent une grande diversité de grandeurs qui vont des propriétés orbitales des planètes (comme la période orbitale et l'excentricité) aux caractéristiques des étoiles hôtes (température effective, rayon, masse, métallicité, etc.). De plus, des informations contextuelles telles que la méthode de découverte, l'année de découverte et le dispositif de découverte sont également incluses. Certaines variables présentent des valeurs manquantes.

| Variables |
|---|
| No. |
| Planet Name |
| Planet Host |
| Num Stars |
| Num Planets |
| Discovery Method |
| Discovery Year |
| Discovery Facility |
| Orbital Period Days |
| Orbit Semi-Major Axis |
| Mass |
| Eccentricity |
| Insolation Flux |
| Equilibrium Temperature |
| Spectral Type |
| Stellar Effective Temperature |
| Stellar Radius |
| Stellar Mass |
| Stellar Metallicity |
| Stellar Metallicity Ratio |
| Stellar Surface Gravity |
| Distance |
| Gaia Magnitude |

TABLE 1 – Variables du jeu de données

Parmi les propriétés des planètes, c'est le cas de la période orbitale, du demi grand axe de l'orbite elliptique, de la masse, de l'excentricité, du flux lumineux et de la température d'équilibre. On concentrera l'étude sur les variables relatives aux propriétés des planètes.

Toutes les variables relatives aux propriétés de l'étoile hôte contiennent des valeurs manquantes dont le nombre varie d'une colonne à l'autre.

La Figure 1 représente l'évolution du nombre d'exoplanètes découvertes chaque année depuis 1989.

La Figure 2 représente le nombre d'exoplanètes détectées selon la méthode de détection.

La Table 2 classe les méthodes de détection par ordre décroissant du nombre d'exoplanète détectées.

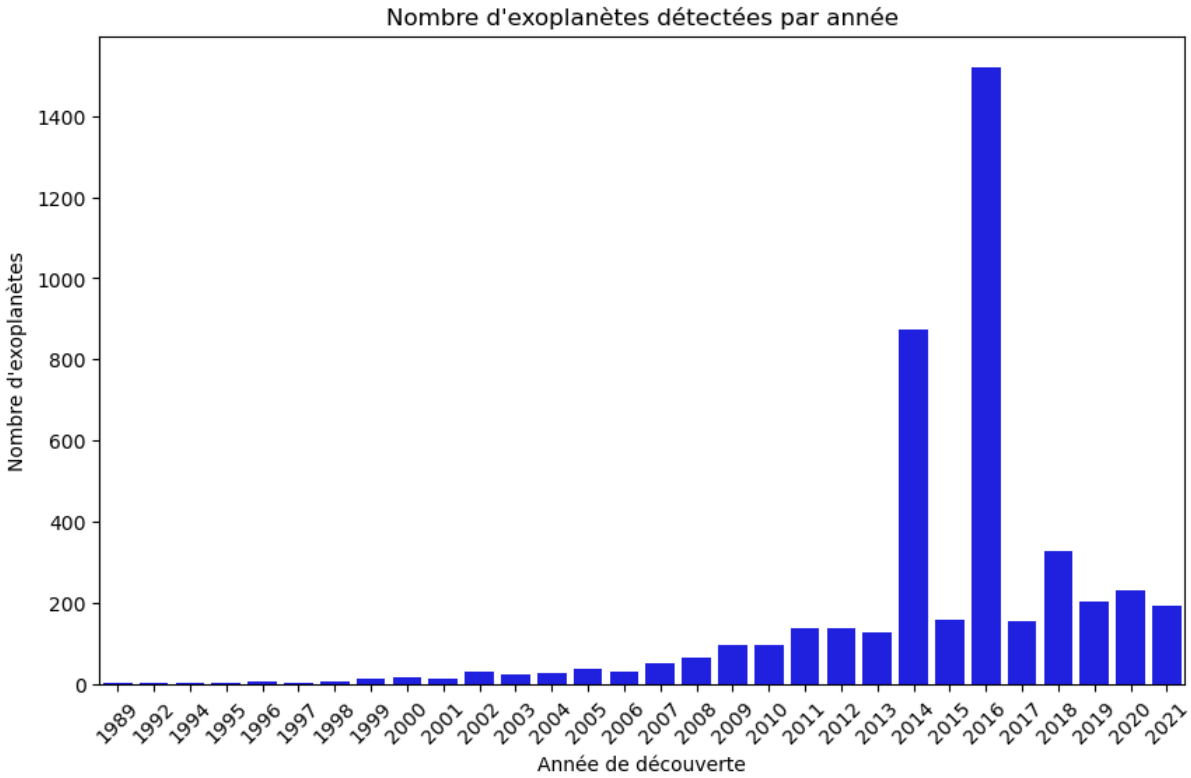


FIGURE 1 – Évolution du nombre d'exoplanètes découvertes depuis 1989

| Discovery Method | |
|---|---|
| Transit | 3444 |
| Radial Velocity | 899 |
| Microlensing | 120 |
| Imaging | 54 |
| Transit Timing Variations | 22 |
| Eclipse Timing Variations | 16 |
| Orbital Brightness Modulation | 9 |
| Pulsar Timing | 7 |
| Pulsation Timing Variations | 2 |
| Astrometry | 1 |
| Disk Kinematics | 1 |

TABLE 2 – Méthodes de détection

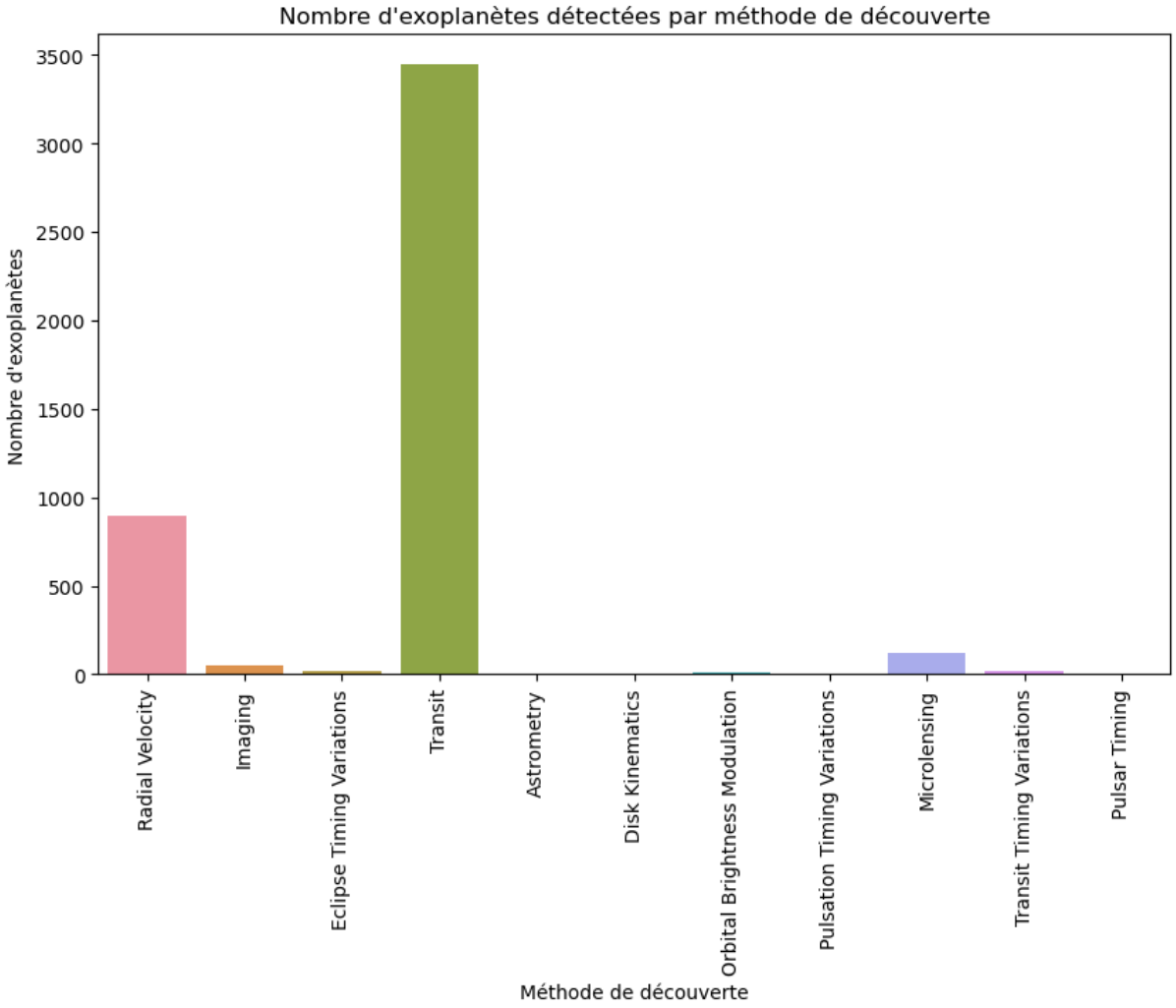


FIGURE 2 – Nombre d'exoplanètes détéctées selon la méthode de détection

## 3.2 Analyse des corrélations

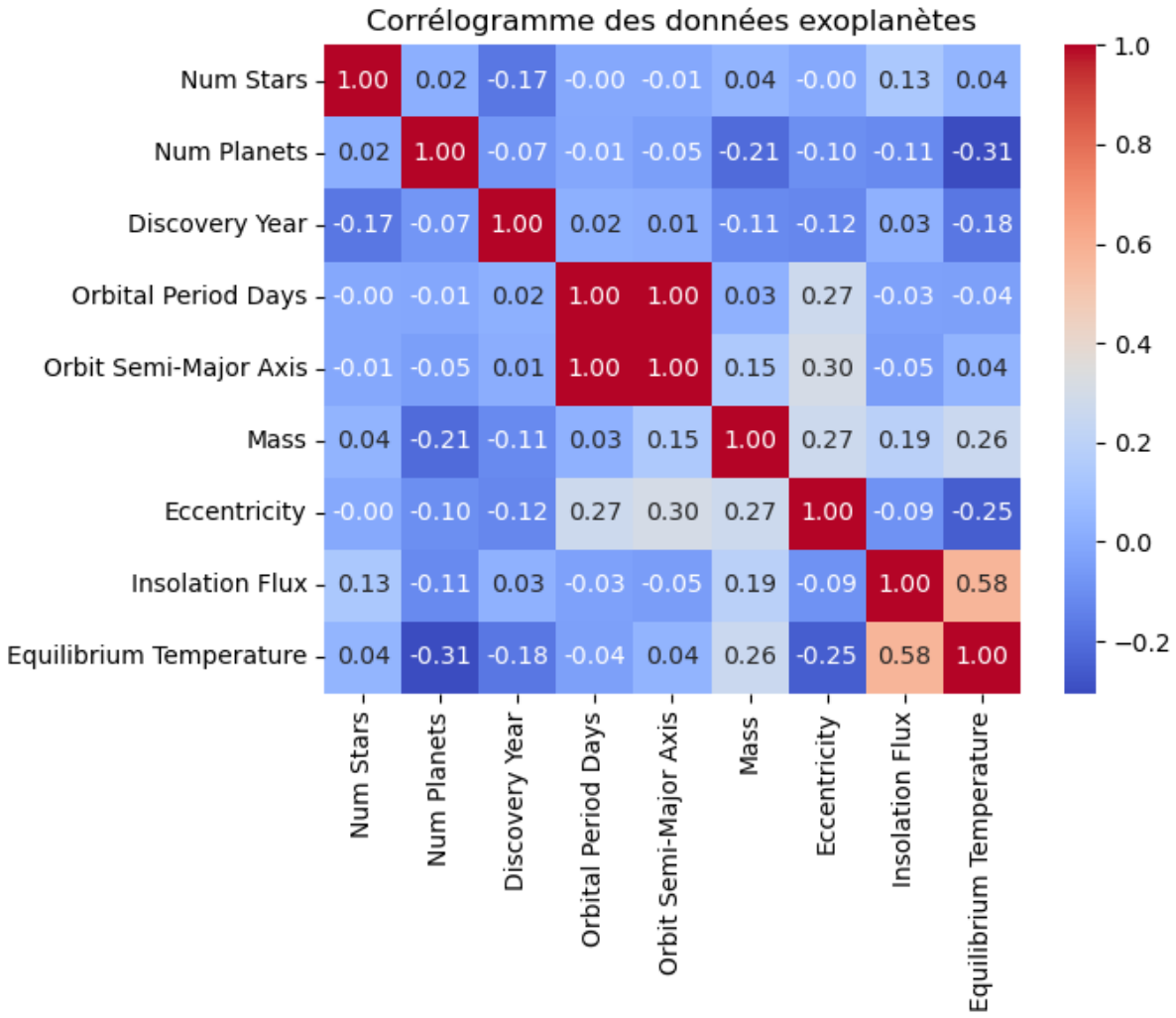


FIGURE 3 – Corrélogramme des données sur les exoplanètes

La Figure 3 représente les corrélations entre les différentes variables quantitatives du jeu de données. Les variables portent uniquement sur les propriétés des planètes. On a volontairement écarté les propriétés des étoiles hôtes dont les données ne sont jamais complètes.

Les variables représentant la période orbitale de la planète en jours et le demi grand axe de l'orbite de révolution sont parfaitement corrélées.

Les variables représentant le flux lumineux et la température d'équilibre sont également fortement corrélées.

La Figure 4 représente la régression linéaire du logarithme de la période orbitale de révolution de la planète en jours en fonction du logarithme du demi grand axe de l'orbite de révolution.

La Figure 5 représente la régression linéaire du logarithme de la température d'équilibre en fonction du logarithme du flux lumineux.

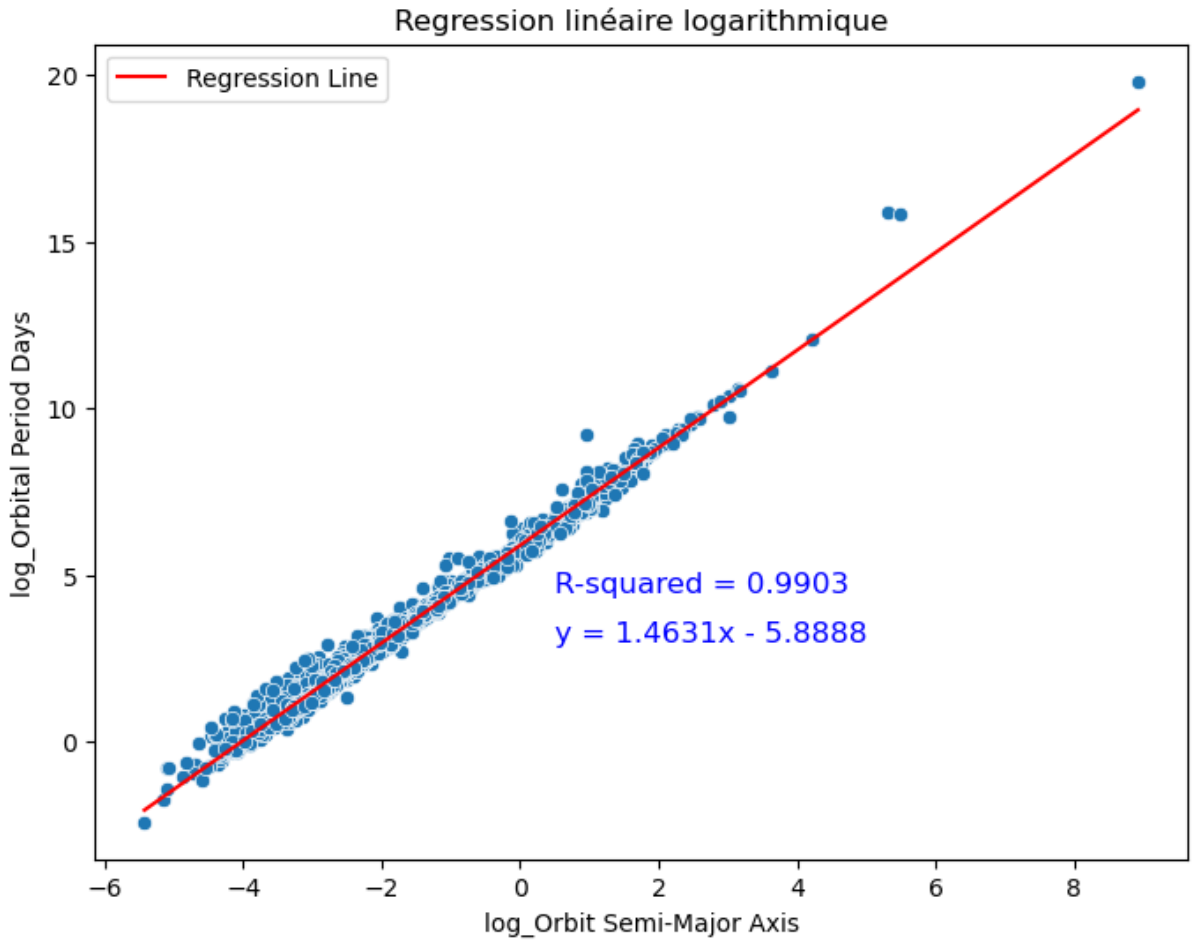


FIGURE 4 – Modélisation du logarithme de la période orbitale en fonction du logarithme du demi grand axe

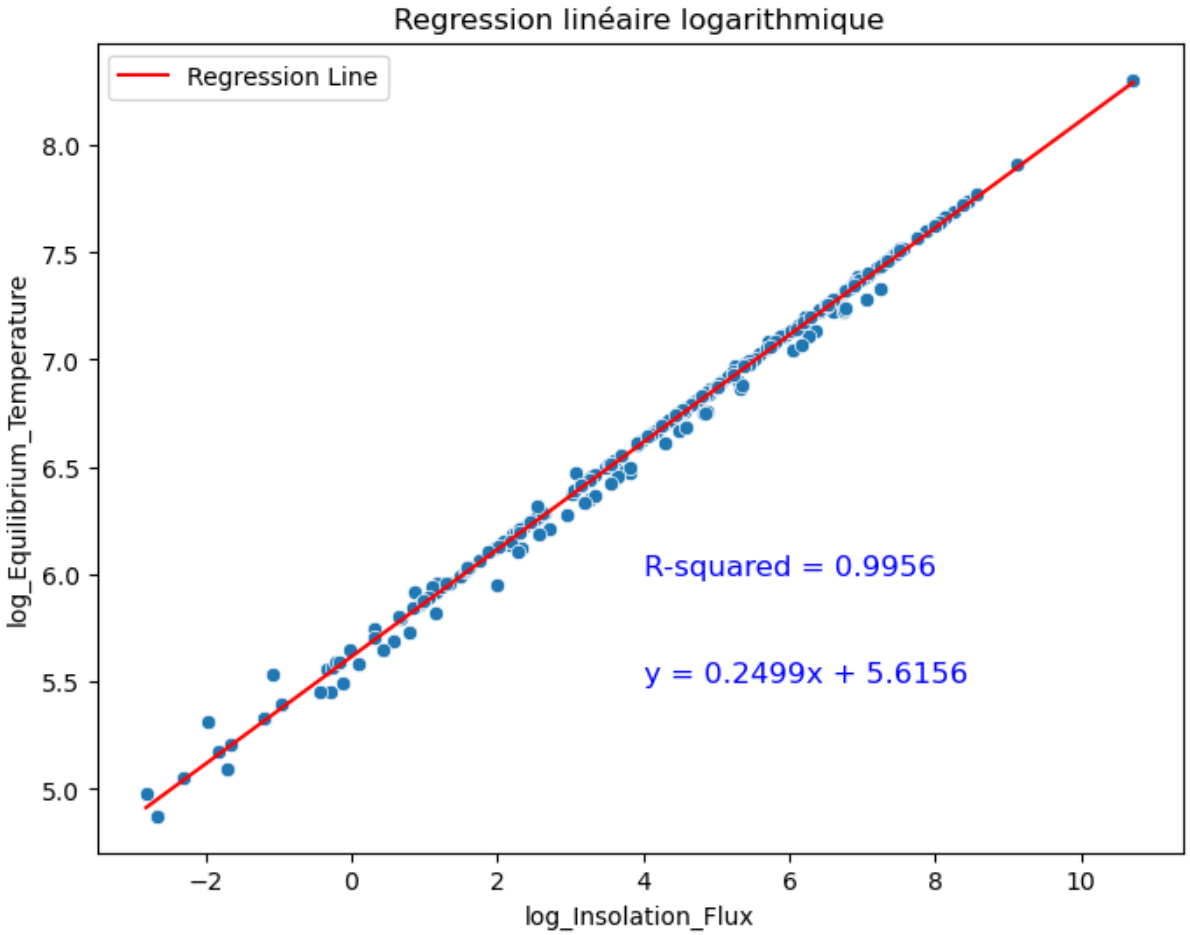


FIGURE 5 – Modélisation du logarithme de la température d'équilibre en fonction du logarithme du flux lumineux

### 3.3 Influence des méthodes de détection sur les propriétés des exoplanètes

La méthode de détection étant une variable catégorique, on réalise sur la Figure 6 des diagrammes en moustache pour chacune des quatre principales méthodes de détection des différentes propriétés mesurées des planètes (période orbitale, demi grand axe, masse, excentricité et température d'équilibre).

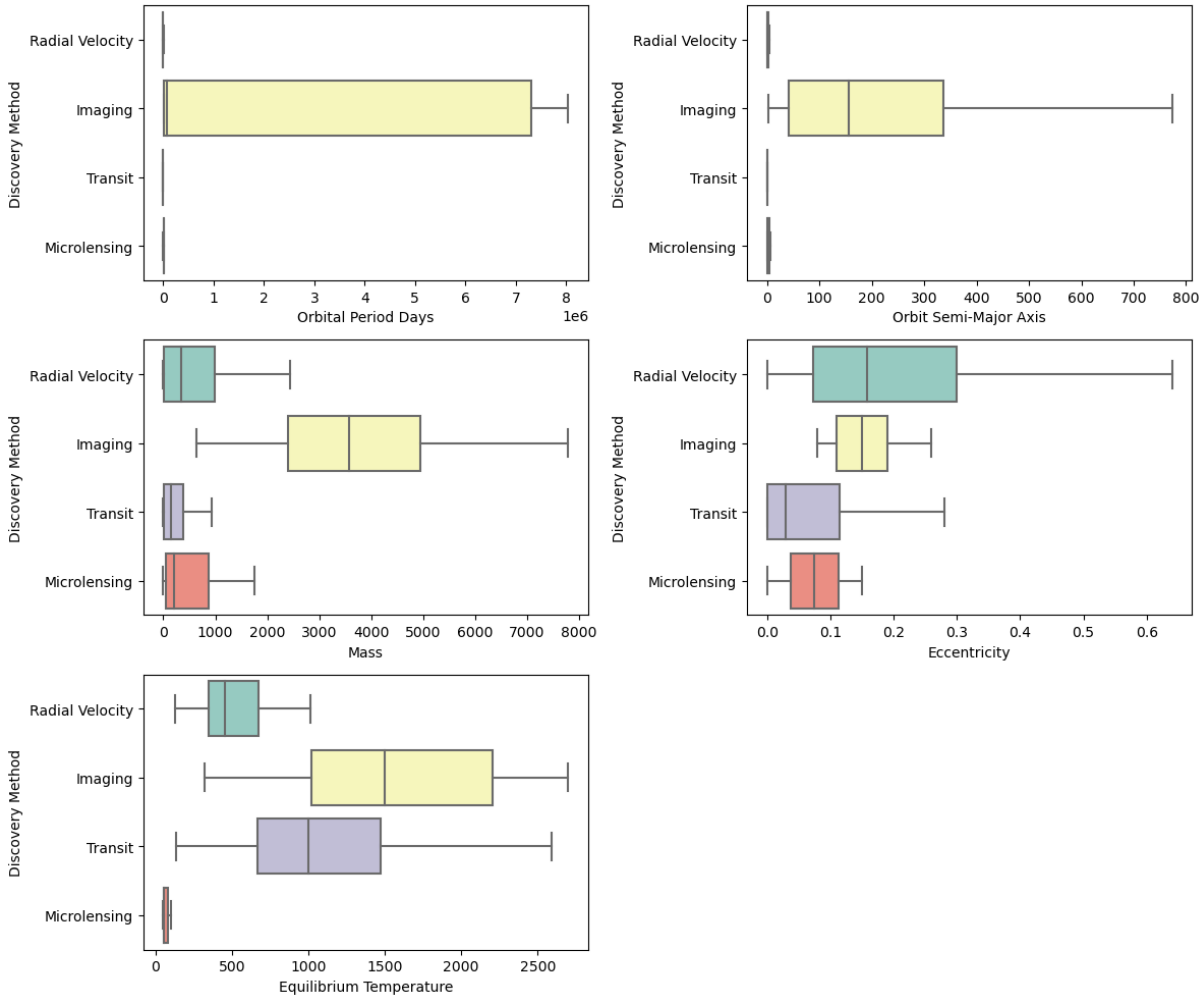


FIGURE 6 – Diagrammes en moustaches des principales propriétés des planètes pour chacune des méthodes de détection

Les diagrammes en moustache obtenus à la Figure 6 font apparaître des valeurs moyenne différentes des propriétés des exoplanètes selon la méthode de détection employée. Il convient donc d'affiner ce résultat préliminaire à l'aide d'un test statistique.

On va effectuer le test d'ANOVA pour comparer les masses des exoplanètes entre les groupes 'Transit', 'Imaging', 'Radial Velocity' et 'Microlensing' qui sont les 4 principales méthodes de détection. Pour cela on fait l'hypothèse que les données sont distribuées normalement et que pour chaque groupes les variances sont égales.

D'après la Table 3, les exoplanètes découvertes par "Imaging" ont une masse estimée de 3875.7880 avec un intervalle de confiance entre 3547.681 et 4203.895. Les exoplanètes découvertes par "Microlensing" ont une masse estimée de 724.1637 avec un intervalle de confiance entre 512.371 et 935.956. Les exoplanètes découvertes par "Radial Velocity" ont une masse estimée de 866.9462 avec un intervalle de confiance entre 789.568 et 944.325. Les exoplanètes découvertes par "Transit" ont une masse estimée de 367.1953 avec un intervalle de confiance entre 289.251 et 445.140. Ces résultats indiquent que les groupes de méthodes de découverte présentent des différences statistiquement significatives en termes de masse moyenne des exoplanètes. Les valeurs p sont toutes $< 0{,}05$, ce qui suggère que les différences observées

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| C(Q("Discovery Method"))[Imaging] | 3875.7880 | 167.301 | 23.167 | 0.000 | 3547.681 | 4203.895 |
| C(Q("Discovery Method"))[Microlensing] | 724.1637 | 107.992 | 6.706 | 0.000 | 512.371 | 935.956 |
| C(Q("Discovery Method"))[Radial Velocity] | 866.9462 | 39.455 | 21.973 | 0.000 | 789.568 | 944.325 |
| C(Q("Discovery Method"))[Transit] | 367.1953 | 39.744 | 9.239 | 0.000 | 289.251 | 445.140 |

TABLE 3 – Table d'ANOVA avec statistiques inférentielles de la masse des exoplanètes pour chaque groupe de méthode de détection

sont significatives.

En conclusion, le test d'ANOVA indique qu'il existe des différences significatives entre les masses moyennes des exoplanètes en fonction de la méthode de découverte utilisée.

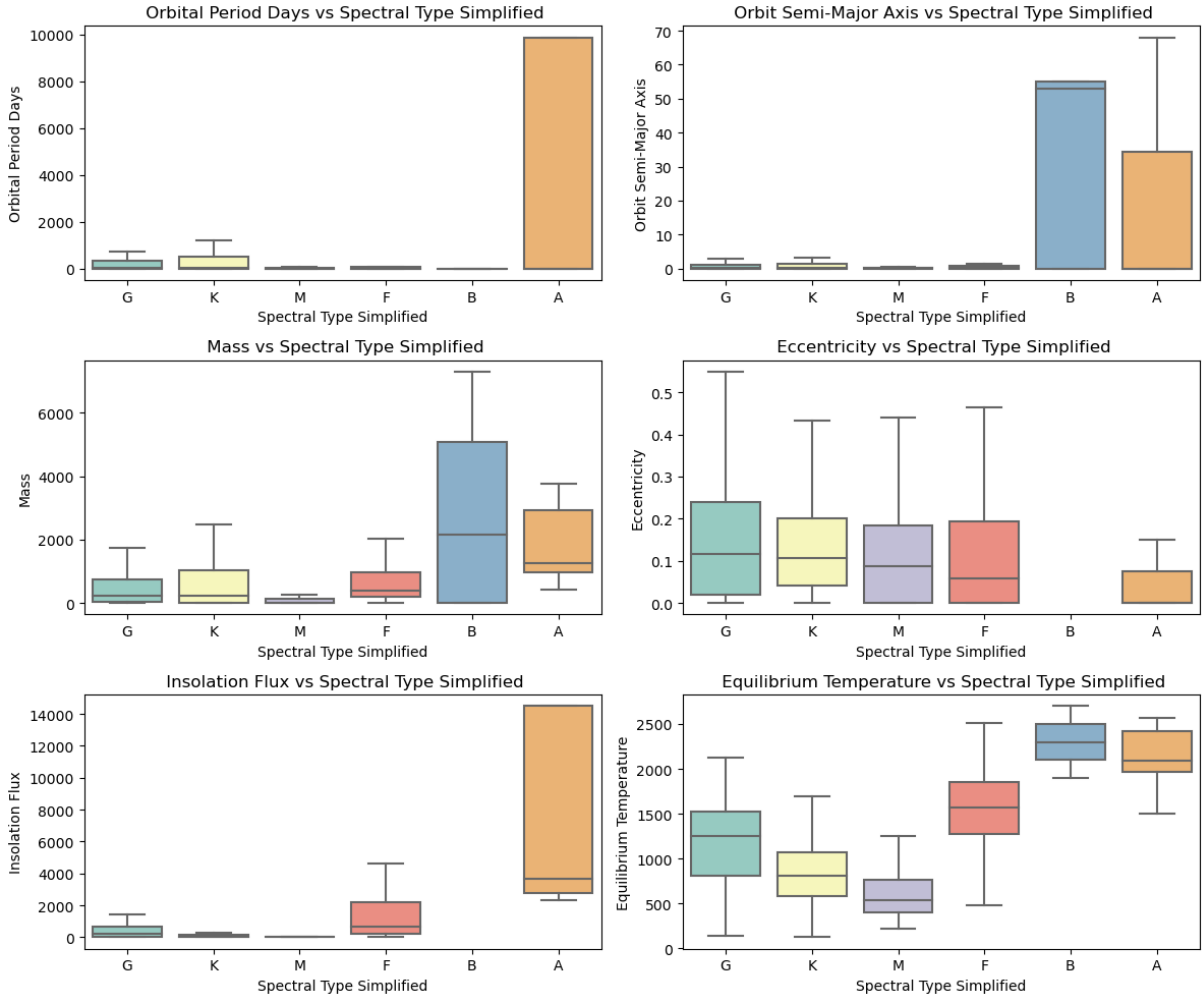


FIGURE 7 – Propriétés des exoplanètes en fonction de la classe de leur(s) étoile(s) hôte(s)

La Figure 7 représente les diagrammes en moustache des différentes propriétés des exoplanètes pour chacun des types spectraux des étoiles hôtes.

On constate des que la masse moyenne des exoplanètes détectées diffèrent selon le type spectral de leur étoile hôte.

On peut maintenant réaliser une ANOVA à deux facteurs en prenant en compte l'interaction entre la méthode de détection et le type spectrale de l'étoile autour de laquelle gravite l'exoplanète afin de connaître l'influence de ces deux variables sur la masse.

La Table 4 indique que pour la méthode de détection, il y a des variations significatives des masses

| | F | PR(>F) |
|---|---|---|
| C(Q("Discovery Method")) | 194.169910 | 8.310430e-117 |
| C(Q("Spectral Type Simplified")) | 0.043977 | 8.339498e-01 |
| C(Q("Discovery Method")) :C(Q("Spectral Type Simplified")) | 2.903119 | 9.185136e-04 |

TABLE 4 – Table ANOVA tenant compte de l'intéraction entre la méthode de détection et le type spectral de l'étoile

des exoplanètes entre les méthodes de détection ($F = 194$ et $p < 0.05$). Pour le type spectral de l'étoile, il n'y a pas de différence significative entre les masses des exoplanètes détectées ($F = 0.04$ et $p = 0.8$). L'interaction entre méthode de détection et type spectral de l'étoile est significative ($F = 2.9$) : il pourrait y avoir des différences significatives dans l'impact des méthodes de détection sur les masses des exoplanètes en fonction des types spectraux simplifiés.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Microlensing Type B | 1.737e-12 | 7.85e-13 | 2.213 | 0.027 | 1.96e-13 | 3.28e-12 |
| Radial Velocity Type B | 2.285e-12 | 1.16e-12 | 1.969 | 0.049 | 7.19e-15 | 4.56e-12 |
| Microlensing Type F | 1.665e-12 | 7.23e-13 | 2.304 | 0.021 | 2.46e-13 | 3.08e-12 |
| Transit Type K | -2024.4650 | 838.782 | -2.414 | 0.016 | -3670.898 | -378.032 |
| Transit Type M | -1938.7150 | 765.064 | -2.534 | 0.011 | -3440.449 | -436.981 |

TABLE 5 – ANOVA à deux facteurs avec intéraction entre méthode et type spectral pour lesquelles $p < 0.05$

La Table 5 détaille les résultats de l'ANOVA pour les intéractions où $p < 0.05$. Les différences observées dans les masses des exoplanètes entre ces groupes ne sont probablement pas dues au hasard.

# 4 Analyse

## 4.1 Analyse des corrélations

La Figure 4 a permis de mettre en évidence une corrélation entre le logarithme de la période orbitale des exoplanètes et le logarithme du demi grand axe de leur orbite de révolution. Le cœfficient directeur de la régression linéaire vaut 1.46. La période orbitale est donc proportionnelle à la puissance 1.46 ($\approx 1.5 = \frac{3}{2}$) du demi grand axe de l'orbite elliptique. On retrouve en réalité ici la 3$^{e}$loi de KEPLER : le carré de la période est proportionnelle au cube du demi grand axe de l'orbite elliptique, soit

$$\frac{T^2}{a^3} = \frac{GM}{4\pi^2}$$

où $M$ est la masse de l'étoile hôte.

La Figure 5 a permis de mettre en évidence une corrélation entre le logarithme de la température d'équilibre et le logarithme du flux lumineux. Dans cette régression linéaire logarithmique, le coefficient directeur vaut 0.25 avec un R-squared de 0.996. On peut en déduire que $T$ est proportionnel au flux à la puissance $\frac{1}{4}$. On retrouve en réalité la loi de Stefan Boltzmann :

$$P = \sigma \cdot T^4$$

où $P$ est la puissance rayonnée par unité de surface

## 4.2 Analyse de l'influence des méthode de détection sur les propriétés des exoplanètes

La Figure 6 et la Table 3 ont permis de montré qu'il y avait des différences significatives dans les masses des exoplanètes détectées selon la méthode de détection. La méthode des transits permet de détecter les exoplanètes les moins massives tandis que la méthode par imagerie directe permet de détecter les exoplanètes les plus massives. On privilégiera donc la méthode des transits pour détecter des planètes

peu massive, comme c'est le cas pour la Terre. Les cœfficient de la Table 5 pour l'intéraction entre méthode des transits et type spectrale K (cœff = -2024) et pour l'intéraction entre méthode des transits et type spectral M (cœff = -1938) sont négatifs et indiquent que les masses des exoplanètes détectées par la méthode des transit est significativement plus petite que pour les autres types spectrale d'étoiles. Par conséquent comme la moyenne des masses des planètes détectées est de 367 masse terrestre (Table 3) , il convient de pointer les détecteur vers des étoiles de type spectral K ou M si l'on souhaite augmenter les chances de détecter des planètes dont la masse se rapproche de celle de la Terre.

# 5 Discussion

En réalité, les exoplanètes qui pourraient potentiellement héberger des formes de vie, doivent vérifier les conditions qui ont permis le développement de la vie sur Terre, c'est à dire posséder de l'eau à l'état liquide. Pour cela, ce n'est pas uniquement la masse de la planète qui importe mais un ensemble de condition décrivant ce qu'on appelle la zone d'habitabilité de la planète. Ces conditions dépendent de différents facteurs, comme la température de l'étoile, la distance entre l'étoile et la planète, la taille de la planète, la gravité à sa surface, .... Aussi, afin de déterminer plus précisément les conditions qui permettraient de détecter les exoplanètes hébergeant probablement des formes de vies, il faudrait prendre en compte tous ces facteurs d'habitabilité et disposer d'un jeu de données qui comportent toutes ces informations.