**Análisis Exploratorio de Datos de Exoplanetas**

**Pregunta a responder**

El proyecto se propone investigar cómo las propiedades de las estrellas anfitrionas condicionan las características de sus exoplanetas. En particular se busca responder si el tipo espectral, la masa o el tamaño de la estrella influyen en el tamaño y la masa de los planetas, y con qué frecuencia aparecen planetas gigantes en diferentes clases estelares.

**Generalidades del dataset**

1. **Exoplanet Encyclopaedia — catálogo completo (CSV)**

**Enlace:** <https://exoplanet.eu/catalog/csv/>

**Qué contiene:** descarga íntegra del catálogo (una fila por planeta confirmado; opción “Download as CSV”).

**Para qué nos sirve:** base con parámetros planetarios y estelares para explorar relaciones estrella–planeta.

1. **NASA Exoplanet Archive — Planetary Systems Composite Parameters (PSCompPars)**

**Enlace (CSV vía TAP):** [https://exoplanetarchive.ipac.caltech.edu/TAP/sync?query=select%20\*%20from%20pscomppars&format=csv](https://exoplanetarchive.ipac.caltech.edu/TAP/sync?query=select%20*%20from%20pscomppars&format=csv)

**Qué contiene:** tabla compuesta (una fila por planeta) con parámetros del sistema, estrella y planeta.

**Para qué nos sirve:** referencia oficial para validar/extender resultados y futuras integraciones.

1. **Île‑de‑France Smart Services — mirror del catálogo Exoplanet.eu**

**Enlace:** <https://data.smartidf.services/explore/dataset/exoplaneteu_catalog-1/download/?format=csv>

**Qué contiene:** copia sincronizada del catálogo de la Exoplanet Encyclopaedia (mismas columnas/registros).

**Para qué nos sirve:** alternativa estable cuando la descarga directa falla; con este mirror generamos el CSV del EDA.

**Fuente principal:** Usaremos el NASA Exoplanet Archive – Planetary Systems Composite Parameters (PSCompPars), una tabla compuesta que reúne parámetros del sistema, de la estrella y del planeta en una fila por planeta confirmado, pensada para análisis estadísticos de poblaciones. Cuando falta un parámetro en la referencia por defecto, la tabla selecciona automáticamente valores de otras referencias disponibles (lo que maximiza cobertura, aunque no garantiza coherencia interna a nivel de sistema).

* Granularidad: 1 fila = 1 planeta confirmado.
* Número de filas y columnas: 7 782 planetas confirmados y 98 variables numéricas y categóricas.
* Principales variables: mass (masa del planeta, en masas de Júpiter), radius (radio del planeta, en radios de Júpiter), star\_mass (masa estelar, en masas solares), star\_radius (radio estelar, en radios solares), semi\_major\_axis (semieje mayor orbital).
* Fuente: catálogo público de la Exoplanet Encyclopaedia, similar a la tabla PSCompPars del NASA Exoplanet Archive.
* Valores faltantes: alrededor del 53 % de los planetas carece de masa, el 34 % carece de radio, el 47,8 % no tiene tipo espectral; no se detectaron nombres duplicados.
* Columnas (ejemplos):
  + Identificadores y sistema: pl\_name, hostname, número de planetas y de estrellas.
  + Parámetros del planeta: masas y radios en distintas unidades, período orbital, semieje mayor, excentricidad, inclinación, densidad, temperatura de equilibrio, indicadores de detección (tránsito, VR, etc.).
  + Parámetros de la estrella: masa, radio, temperatura efectiva, metalicidad, luminosidad, tipo espectral; además de astrometría (distancia, paralaje, movimiento propio).
  + Metadatos de descubrimiento: método, año, referencia, instalación/instrumento.
  + Tipos de dato: numéricos continuos (masas, radios, periodos), categóricos (tipo espectral, método), fechas/años y banderas. [Archivo de Exoplanetas de NASA](https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html?utm_source=chatgpt.com)

**Origen y recolección.**

* Descarga reproducible vía TAP/API (ADQL o endpoints REST), p. ej., select \* from pscomppars en CSV/JSON; el archivo mantiene recuentos oficiales y documentación pública de columnas.

**Tipos de dato**

El conjunto incluye **84 columnas numéricas** y **14 categóricas**. Las variables numéricas abarcan masas, radios, periodos, semiejes y otros parámetros físicos; las categóricas incluyen nombres (name), tipo espectral de la estrella (star\_sp\_type) y método de descubrimiento.

Se derivaron dos variables auxiliares para este análisis:

* **spec\_class**: primera letra del tipo espectral agrupada en clases A, F, G, K, M y O; otros valores se agrupan en “Other”.
* **is\_giant**: indicador binario que marca un planeta como gigante si su masa ≥ 0,3 M\_J o su radio ≥ 0,6 R\_J.

**Estadística descriptiva**

Las principales variables numéricas presentan las siguientes estadísticas resumidas:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Count | Mínimo | Mediana | Media | Máximo |
| Masa del planeta (mass) | 3 663 | 1.42e‑09 M\_J | 2.50 M\_J | 14.94 M\_J | 78.27 M\_J |
| Radio del planeta (radius) | 5 153 | 0.00107 R\_J | 0.236 R\_J | 0.542 R\_J | 18.80 R\_J |
| Masa estelar (star\_mass) | 6 488 | 0.011 M\_⊙ | 0.94 M\_⊙ | 0.932 M\_⊙ | 23.47 M\_⊙ |
| Radio estelar (star\_radius) | 6 044 | 1.43e‑05 R\_⊙ | 0.95 R\_⊙ | 1.49 R\_⊙ | 91.40 R\_⊙ |
| Semieje mayor (semi\_major\_axis) | 5 427 | 0.00087 AU | 0.1201 AU | 196.10 AU | 80 679 AU |

Estas cifras revelan la gran dispersión de las masas planetarias (de microplanetas hasta gigantes de 78 M\_J) y el amplio rango de órbitas, desde planetas ultra‑cercanos (0,00087 AU) hasta cuerpos extremadamente distantes.

M⊙: masa solar (unidad para comparar masas estelares); ≈ 1,988 × 10³⁰ kg.

R⊙: radio solar (unidad para tamaños estelares); la IAU adoptó valores nominales de referencia.

L⊙: luminosidad solar (unidad para potencias radiadas); con valor nominal recomendado por la IAU.

AU (au): unidad astronómica (longitud); exactamente 149 597 870 700 m (IAU 2012, Res. B2).

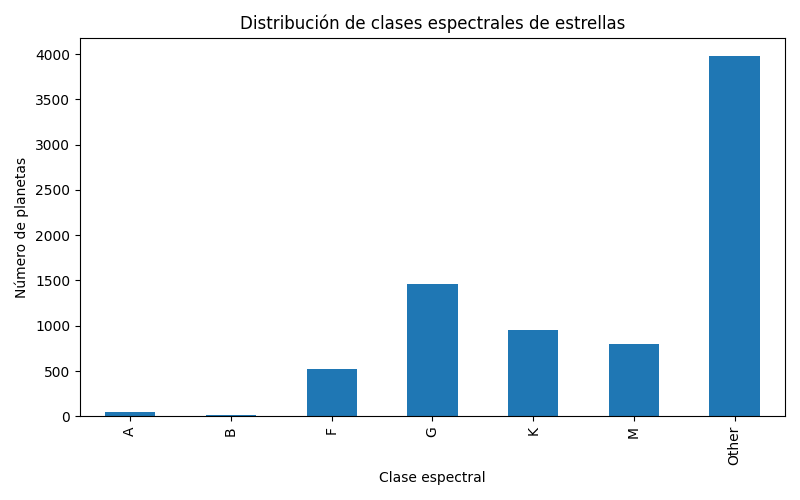
R\_J: radio de Júpiter (unidad para radios planetarios); referencia típica ≈ 69 911 km (radio medio).

M\_J: masa de Júpiter (unidad para masas planetarias); ≈ 1,898 × 10²⁷ kg (JPL/SSD).

**Visualización**

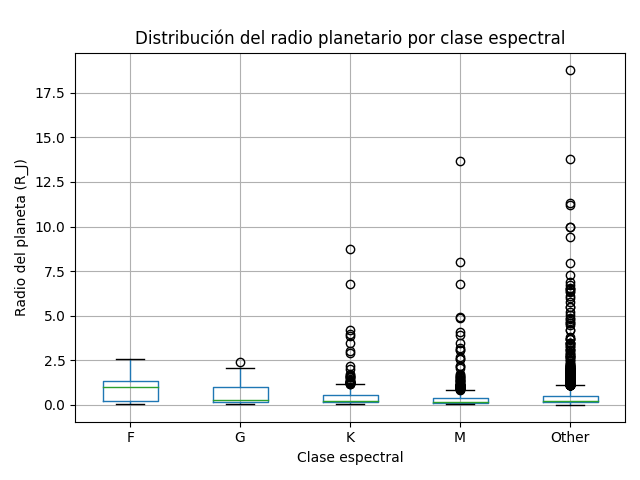
**Distribución de clases espectrales**

Las clases estelares **G** y **K** son las más comunes, seguidas de **M** y **F**. Casi la mitad de los registros corresponden a estrellas sin clasificación específica o poco usual (“Other”).



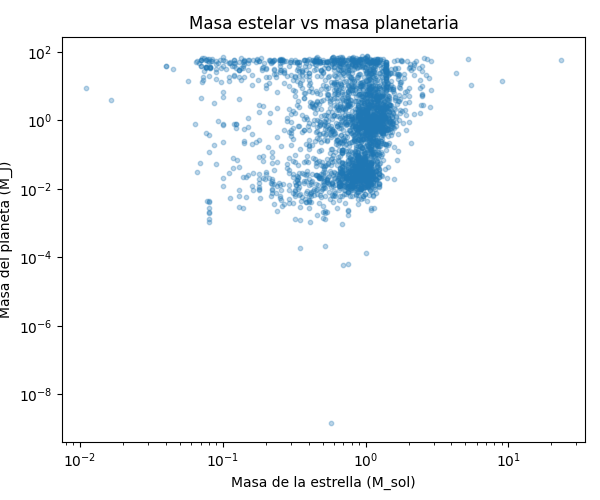
**Tamaño planetario según clase**

Un diagrama de caja muestra que las estrellas de tipo **F** y **G** albergan planetas más grandes (medianas ~1,0 y 0,26 R\_J), mientras que las de tipo **K** y **M** hospedan planetas más pequeños (~0,22 y 0,18 R\_J).



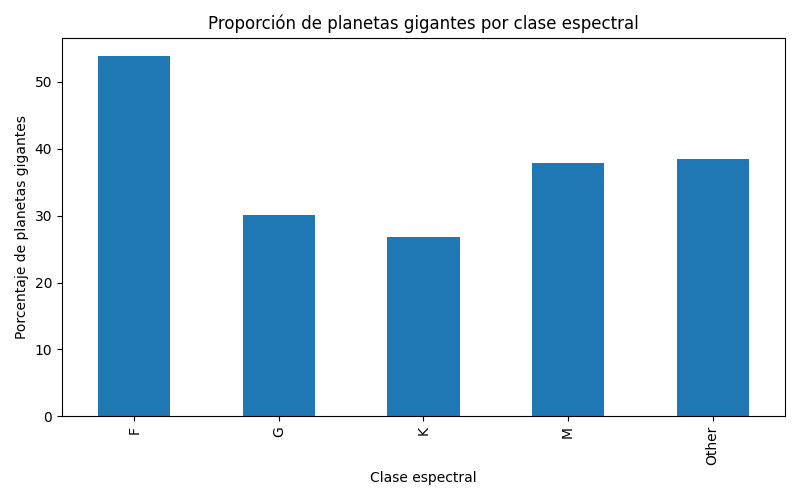
**Relación masa estelar–masa planetaria**

El diagrama de dispersión revela que la correlación entre masa estelar y masa planetaria es baja (coeficiente ~−0,05). Las estrellas de mayor masa no muestran tendencia clara a albergar planetas más masivos.



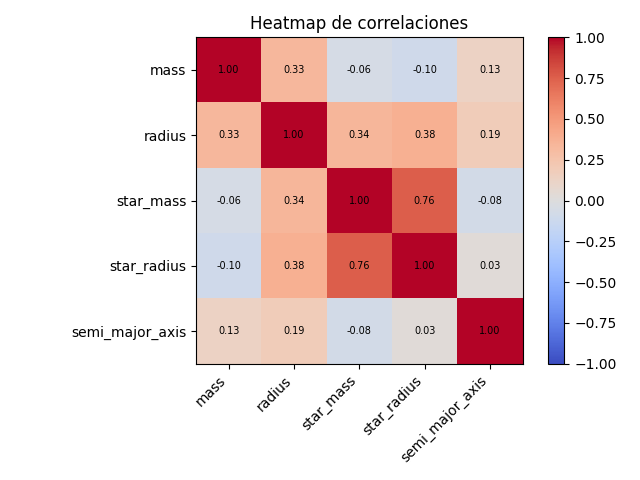
**Frecuencia de planetas gigantes**

Las estrellas **F** presentan la mayor fracción de planetas gigantes (~54 %), mientras que **G** y **K** tienen proporciones menores (~30 % y 27 %). Las enanas **M** también muestran una fracción notable de gigantes (~38 %).



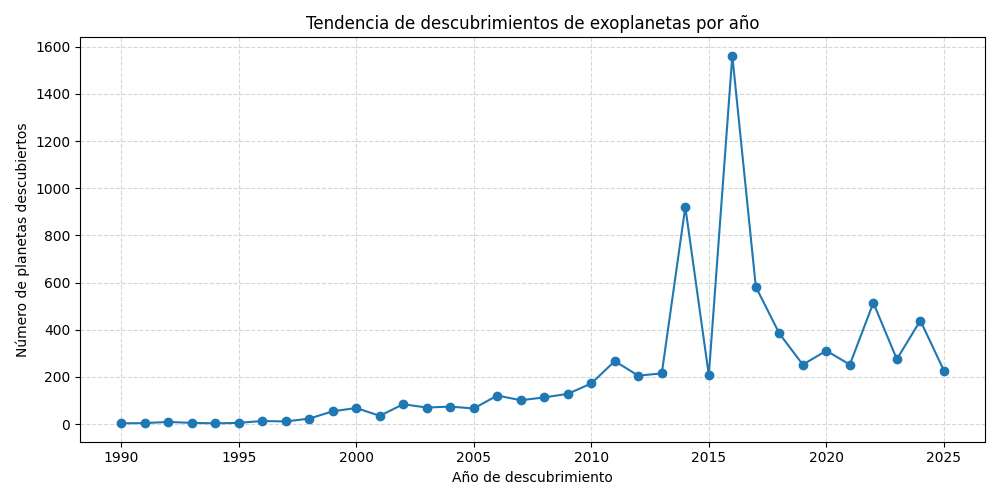
**Correlaciones**

El mapa de calor indica que el radio planetario se correlaciona moderadamente con el radio estelar (0,38) y la masa estelar (0,34). La masa del planeta se correlaciona débilmente con el radio y el semieje mayor.



**Tendencia temporal**

El número de descubrimientos de exoplanetas aumenta bruscamente entre 2014 y 2016 y se estabiliza alrededor de 400–500 descubrimientos anuales en los últimos años.



**Interacciones**

* **Masa y radio estelar vs. variables planetarias:** El radio del planeta muestra una relación moderada con el radio estelar y la masa estelar, indicando que estrellas más grandes tienden a albergar planetas de mayor tamaño. La correlación de masas es muy débil, sugiriendo que la masa de la estrella no determina directamente la masa del planeta.
* **Clase espectral y distribución de tamaños:** Las clases **F** y **G** destacan por tener mayor proporción de planetas gigantes y radios planetarios más altos. Las clases **K** y **M** concentran supertierras y mini‑neptunos, aunque todavía presentan algunos gigantes.
* **Frecuencia de gigantes:** La comparación de frecuencias confirma que las enanas rojas (tipo **M**) tienen menos gigantes gaseosos en porcentaje, pero no los excluyen, matizando la hipótesis de que estas estrellas no albergan Júpiteres calientes.

**Sumarización**

El análisis exploratorio de este catálogo de exoplanetas muestra que la distribución de tamaños y masas planetarias depende fuertemente del tipo espectral y del tamaño de la estrella anfitriona, mientras que la masa estelar se correlaciona poco con la masa planetaria. Las estrellas de tipo **F** y **G** albergan una mayor proporción de planetas gigantes, en contraste con las clases **K** y **M**. El alto porcentaje de valores faltantes sugiere la necesidad de imputación y depuración antes de modelar. Como siguientes pasos se recomienda profundizar en modelos predictivos que incorporen metalicidad y edad estelar, y comparar los resultados con la tabla **PSCompPars** del NASA Exoplanet Archive para validar la robustez de estos hallazgos.