Para empezar el proyecto me metí en la base de datos de exoplanetas de la NASA y estuve mirando que tipos de datos había para organizar el tipo de gráficos que iba a poder hacer. Con esto claro, estuve mirando que era necesario para sacar los datos. Resulta que hay un paquete, llamado pyvo, que simplifica mucho las queries a este tipo de bases de datos.

Empecé intentando sacar todos los datos del tirón, pero resulta que con más de 29000 filas y 300 columnas, nunca llegué a completar un query antes de que el ordenador me diera un problema. Decidí por lo tanto centrarme directamente en ciertas columnas que eran las que me interesaban. Al final la query quedó así:

import pyvo as vo  
import pandas as pd  
tap\_service = vo.dal.TAPService(" https://exoplanetarchive.ipac.caltech.edu/TAP")  
tap\_results = tap\_service.search("""SELECT pl\_name, pl\_controv\_flag, sy\_snum, sy\_pnum, cb\_flag, discoverymethod, disc\_locale, rv\_flag, pul\_flag, ptv\_flag,  
 tran\_flag, ast\_flag, obm\_flag, micro\_flag, etv\_flag, ima\_flag, dkin\_flag, pl\_orbper, pl\_orbsmax, pl\_rade, pl\_radj,   
 pl\_masse, pl\_massj, pl\_msinie, pl\_msinij, pl\_orbeccen, pl\_tranmid, st\_teff, st\_rad, st\_mass, st\_age, sy\_dist, rowupdate FROM PS""")  
data = pd.DataFrame(tap\_results)  
data.to\_csv("Datos\_exoplanetas.csv")  
print("Done")

Lo que hace cada columna viene explicado aquí (<https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html> ), pero también he añadido una tabla a streamlit con las definiciones traducidas para las columnas utilizadas.

|  |  |
| --- | --- |
| **ast\_flag** | Bandera indicando si la estrella se ha detectado usando astrometría. |
| **cb\_flag** | Bandera indicando si el planeta orbita un sistema binario |
| **disc\_locale** | Indica si el planeta se ha detectado desde la Tierra o desde el espacio |
| **discoverymethod** | Indica con que método se descubrió el planeta |
| **sy\_snum** | Número de estrellas en el sistema |
| **dkin\_flag** | Indica si el planeta se ha detectado mediante cinemática de discos |
| **etv\_flag** | Indica si el planeta se ha detectado mediante variaciones en el tiempo de eclipse |
| **ima\_flag** | Indica si el planeta fue detectado viéndolo directamente |
| **obm\_flag** | Indica si el planeta se detecto mediante variaciones en el brillo orbital |
| **pl\_masse** | La masa del planeta como múltiplo de la masa de la Tierra |
| **pl\_massj** | La masa del planeta como múltiplo de la masa de la Tierra |
| **pl\_msinie** | Masa mínima del planeta en masas de la Tierra |
| **pl\_msinij** | Masa mínima del planeta en masas de Júpiter |
| **pl\_orbeccen** | Excentricidad de la órbita |
| **pl\_orbper** | Periodo orbital del planeta |
| **pl\_orbsmax** | El semieje mayor de la órbita |
| **pl\_rade** | Radio del planeta en radios de la Tierra |
| **pl\_radj** | Radio del planeta en radios de Júpiter |
| **pl\_tranmid** | El tiempo que tarda el planeta en cruzar la estrella de un lado a otro visto desde la Tierra |
| **ptv\_flag** | Indica si el planeta se detecto mediantes variaciones de tiempo de pulsación. |
| **pul\_flag** | Indica si el planeta se detecto mediante variaciones de tiempo de pulsar |
| **rv\_flag** | Bandera indicando si la estrella exhibe variaciones de velocidad radial debido al planeta |
| **st\_age** | La edad de la estrella |
| **st\_mass** | La masa de la estrella en masas solares |
| **st\_rad** | El radio de la estrella, medido en radios solares |
| **st\_teff** | La temperatura de la estrella en Kelvin |
| **sy\_dist** | Distancia al sistema en parsecs |
| **sy\_pnum** | Número de planetas en el sistema |
| **tran\_flag** | Bandera indicando si el planeta se ha detectado mediante tránsito primario |

Con la query hecha, la pasé a un csv para no tener que hacerla cada vez que ejecutaba el programa.

# Organizar datos

Resulta que de esas 29000 entradas, no todas estaban cercioradas y había muchas repeticiones. Para evitar datos que sean dudosas, empecé quitando todas las entradas cuya existencia estuviera disputada en la literatura (con la bandera pl\_controv\_flag). Una vez hecho esto, hice un aggregate usando el nombre del planeta para juntar los datos de varios experimentos, evitando así perder datos. Dependiendo del tipo de dato, el aggregate realizaba una operación u otra. En la siguiente tabla se muestran las operaciones realizadas:

| **Columna** | **operación** |
| --- | --- |
| **ast\_flag** | first |
| **cb\_flag** | max |
| **disc\_locale** | first |
| **discoverymethod** | first |
| **sy\_snum** | max |
| **dkin\_flag** | max |
| **etv\_flag** | max |
| **ima\_flag** | max |
| **obm\_flag** | max |
| **pl\_masse** | mean |
| **pl\_massj** | mean |
| **pl\_msinie** | mean |
| **pl\_msinij** | mean |
| **pl\_orbeccen** | max |
| **pl\_orbper** | max |
| **pl\_orbsmax** | max |
| **pl\_rade** | mean |
| **pl\_radj** | mean |
| **pl\_tranmid** | mean |
| **ptv\_flag** | max |
| **pul\_flag** | max |
| **rv\_flag** | max |
| **st\_age** | mean |
| **st\_mass** | mean |
| **st\_rad** | mean |
| **st\_teff** | mean |
| **sy\_dist** | mean |
| **sy\_pnum** | max |
| **tran\_flag** | max |

Una vez hecho esto, el dataframe estaba listo para ser analizado.

# Analizando los datos

El análisis de datos se llevó a cabo en un notebook, y luego se paso a un streamlit para poder presentarlo. Los gráficos se realizaron tanto en matplotlib como en plotly.

Las gráficas realizadas fueron las siguientes:

* Scatter plots:
  + Masa del planeta contra periodo orbital organizado por método de descubrimiento
  + Masa del planeta contra semieje mayor organizado por método de descubrimiento
  + Masa del planeta contra periodo orbital organizado por emplazamiento de descubrimiento
  + Masa del planeta contra semieje mayor organizado por emplazamiento de descubrimiento
  + Júpiters calientes organizados por método de descubrimiento
  + Júpiters calientes
  + Periodo orbital contra distancia
* Bar plots:
  + Planetas descubiertos con cada método
  + Número de estrellas por sistema

Con los gráficos realizados, se hizo un streamlit para poder presentarlo.

# Streamlit

El streamlit está dividido en 4 secciones:

1. Introducción – En la que se discute brevemente de que va el proyecto y la motivación.
2. Métodos de detección – En la que se explican los distintos métodos de detección con animaciones donde fue posible encontrarlas para que el lector pueda saber de que va la cosa.
3. Datos – En la que se explica como se llevo a cabo el análisis de datos.
4. Visualización – En la que se presentan los distintos gráficos y se explica brevemente su significado.