
Apex
En busca del más allá…
Proyecto final Data Science
Nicolas Rodriguez
Comisión 39990

Índice

Presentación

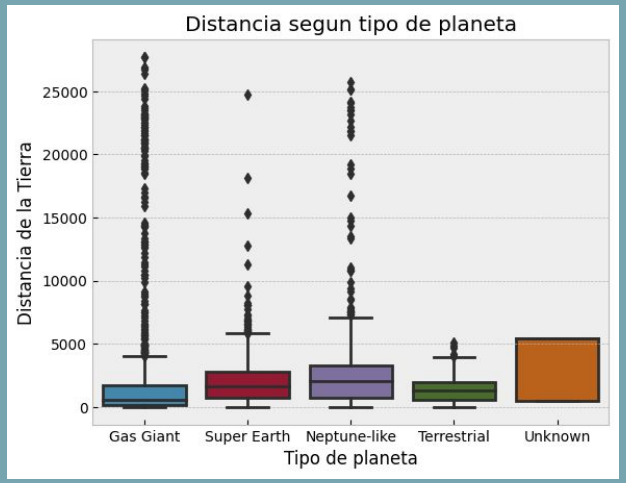
El Planetario Galileo Galilei está haciendo una recopilación de datos en base al tiempo transcurrido sobre el hallazgo de planetas con el fin de realizar una presentación para exponer a las nuevas generaciones y atraer al mundo de la astronomía a aquellos interesados.

Nos contrataron para realizar una reseña de todos los datos recopilados con visualizaciones que ayuden a realizar esta presentación para el público.

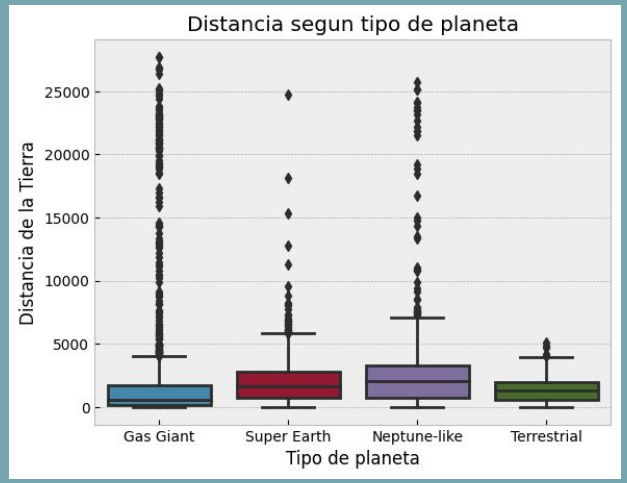
Se nos ha proporcionado un dataset en formato CSV el cual contiene la información acumulada de hallazgos de 1992 a 2023.

Análisis de Datos Nulos

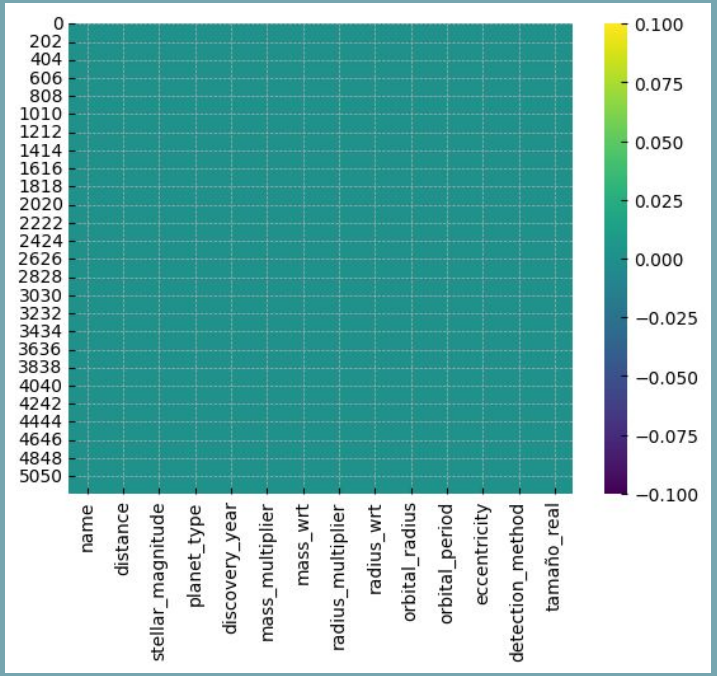
Al explorar los datos nos encontramos con la necesidad de realizar un Data Wrangling.
Para ello realizamos lo siguiente:



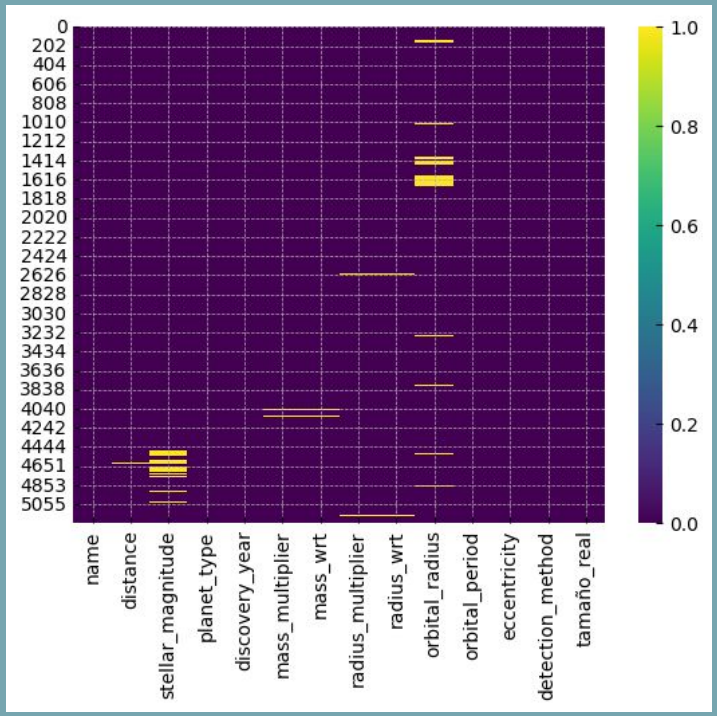
Eliminamos 5 filas de datos incompletos



Análisis de Datos Nulos



Aquellos datos vacíos los reemplazamos por la moda.



Ahora si podemos continuar con nuestros objetivos.

Composición

Nuestro dataset está compuesto por 13 columnas, posteriormente crearemos dos más sintéticas a partir de otras columnas ya existentes

Antes:

	name	distance	stellar_magnitude	planet_type	discovery_year	mass_multiplier	mass_wrt	radius_multiplier	radius_wrt	orbital_radius	orbital_period	eccentricity	detection_method
11 Comae Berenices b	304.0	4.72307	Gas Giant	2007	19.4	Jupiter	1.08	Jupiter	1.29	0.892539	0.23	Radial Velocity

Después:

	name	distance	stellar_magnitude	planet_type	discovery_year	mass_multiplier	mass_wrt	radius_multiplier	radius_wrt	orbital_radius	orbital_period	eccentricity	detection_method	MasaReal	RadioReal
11 Comae Berenices b	304.0	4.72307	Gas Giant	2007	19.4	Jupiter	1.08	Jupiter	1.29	0.892539	0.23	Radial Velocity	36821.2	36821.2

La obtención de estas variables se da gracias a la multiplicación de los multiplicadores de masa y radio con su correlación de tamaño del planeta Júpiter o la Tierra.

Composición

0	name	5245	non-null	object
1	distance	5245	non-null	int64
2	stellar_magnitude	5245	non-null	int64
3	planet_type	5245	non-null	int64
4	discovery_year	5245	non-null	int64
5	mass_multiplier	5245	non-null	int64
6	mass_wrt	5245	non-null	object
7	radius_multiplier	5245	non-null	int64
8	radius_wrt	5245	non-null	object
9	orbital_radius	5245	non-null	int64
10	orbital_period	5245	non-null	int64
11	eccentricity	5245	non-null	int64
12	detection_method	5245	non-null	object
13	MasaReal	5245	non-null	float64
14	RadioReal	5245	non-null	float64

Nuestros objetivos:

Según los tipos de planetas, ¿Cuántos hay de cada uno?

¿Qué relación encuentras entre los datos de los planetas?

¿Cómo los encontramos?

¿Hubo un aumento de hallazgos a medida que pasaba el tiempo?

Nuestros objetivos:

Según los tipos de planetas, ¿Cuántos hay de cada uno?

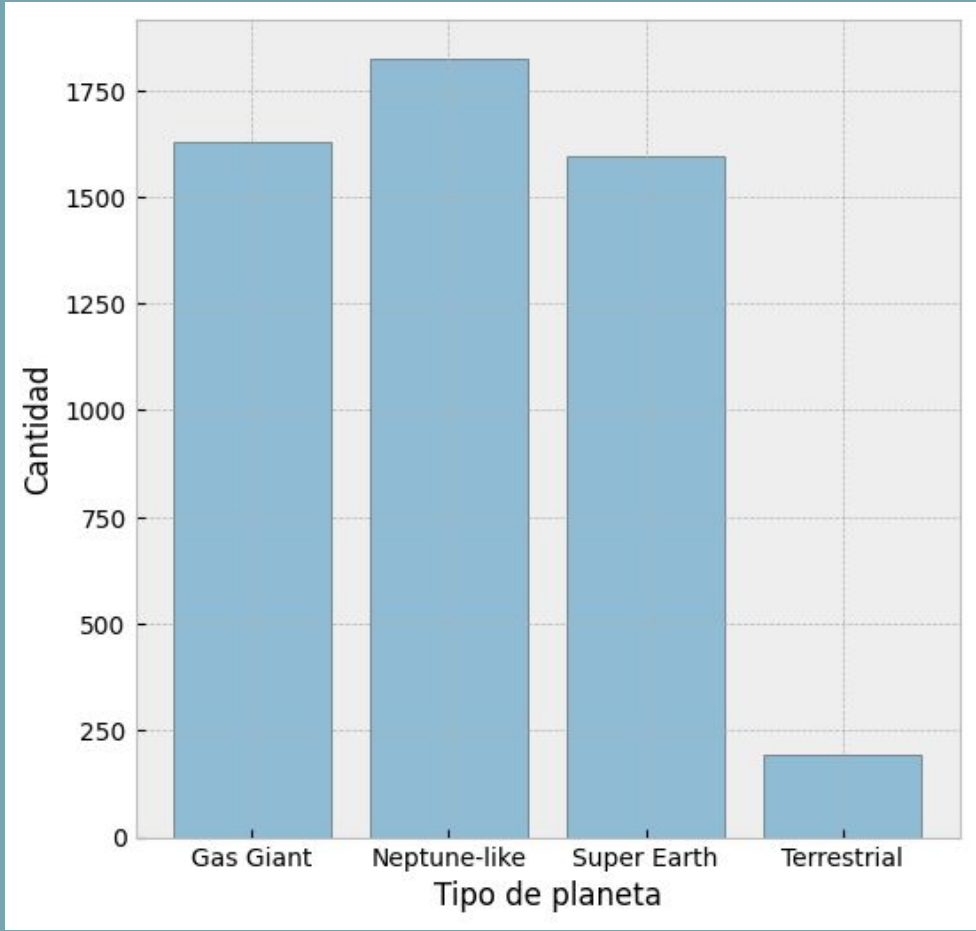
Para ellos filtramos por los tipos de planetas y se realizó la sumatoria de todos los hallazgos en el paso del tiempo. Nos encontramos con cuatro grupos:

Gigantes gaseosos(Gas Giant) 1630 planetas

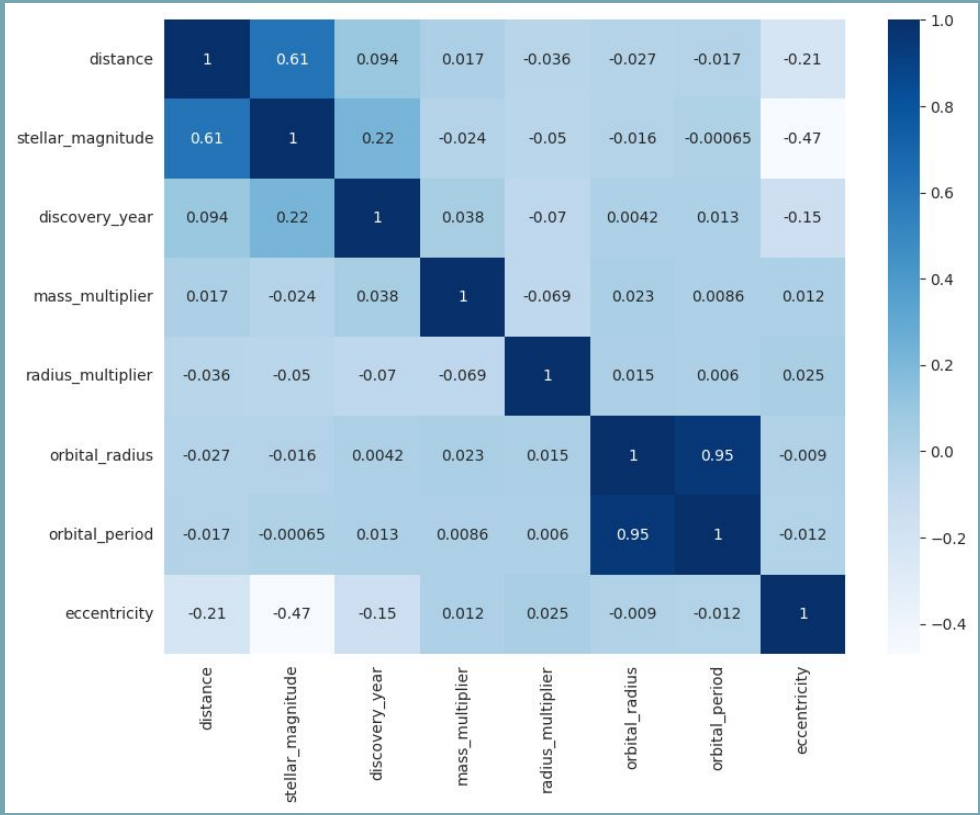
Tipo Neptuno (Neptune-like) 1825 planetas

Super Tierras (Super Earth) 1595 planetas

Terrestres (Terrestrial) 195 planetas



Nuestros objetivos:



¿Qué relación encuentras entre los datos de los planetas?

Primero veamos la relación entre variables según el mapa de correlación, donde el color más intenso es el que muestra más compatibilidad.

Observaciones de relaciones:

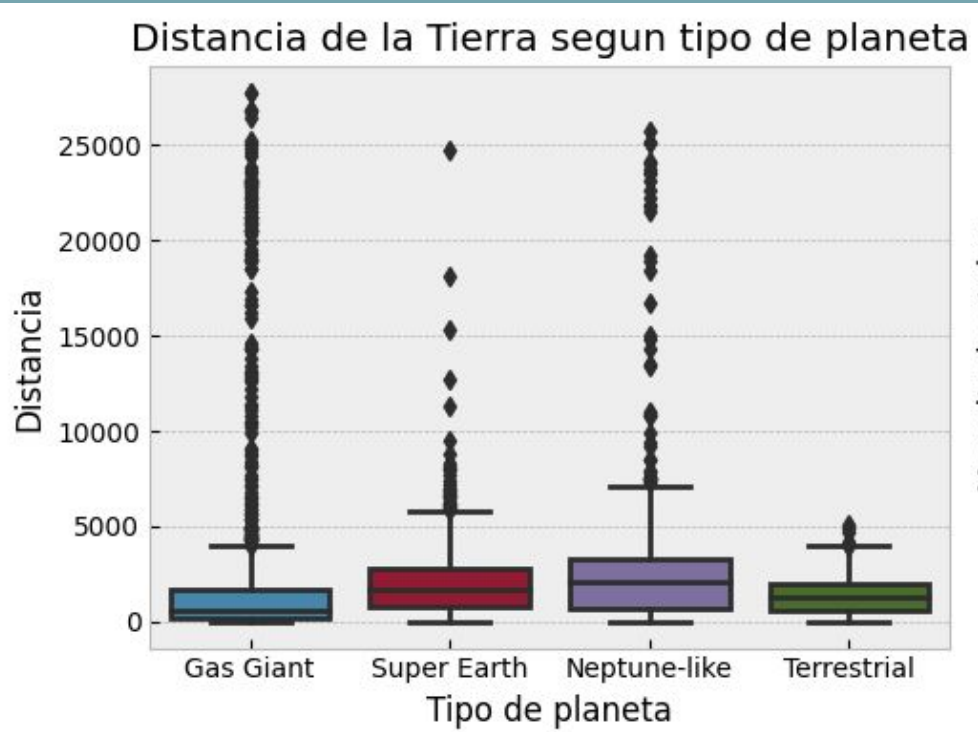
-Magnitud estallar-Distancia

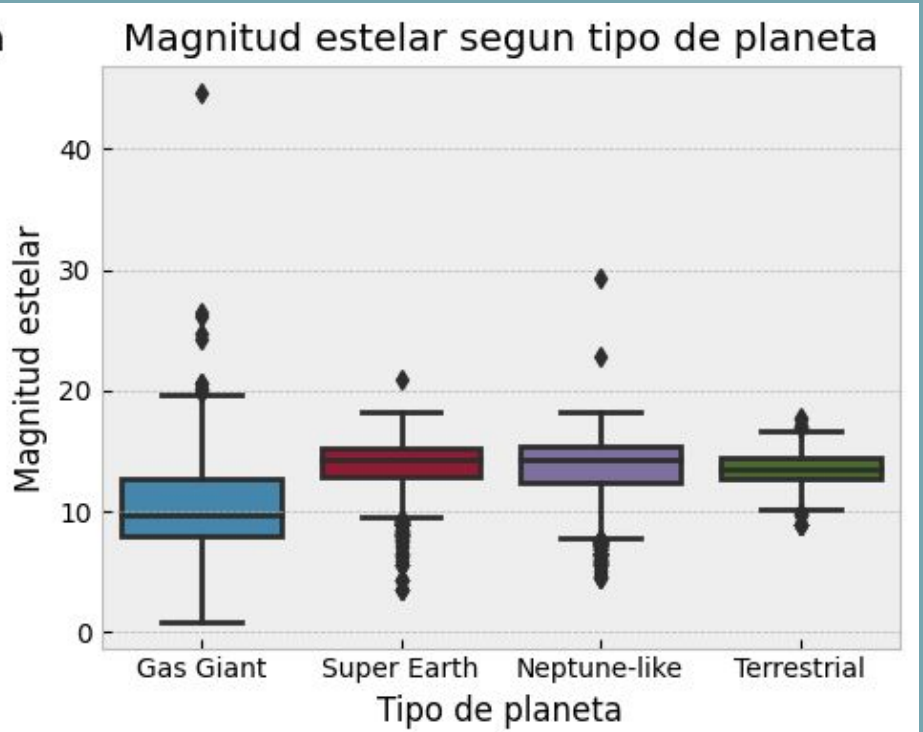
-Radio Orbital-Periodo Orbital

Nuestros objetivos:

¿Qué relación encuentras entre los datos de los planetas?

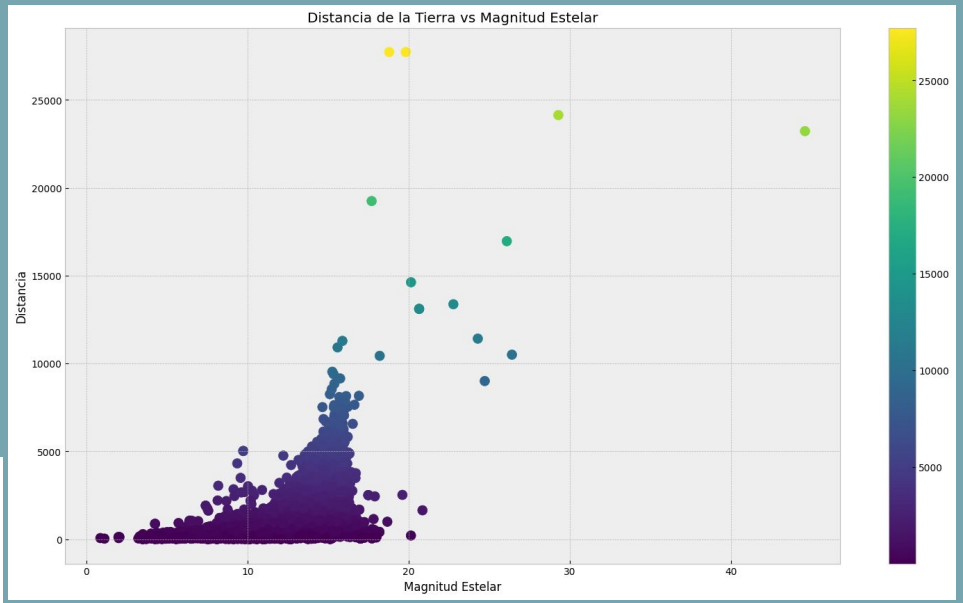
Los gigantes gaseosos (Gas Giant) son los que predominan en magnitud estelar, lo que significa que tienen un mayor espectro luminoso lo que nos permite hallarlos con más facilidad y lejanía, siendo así, también son los que tienen mayor distancia con respecto a la Tierra.

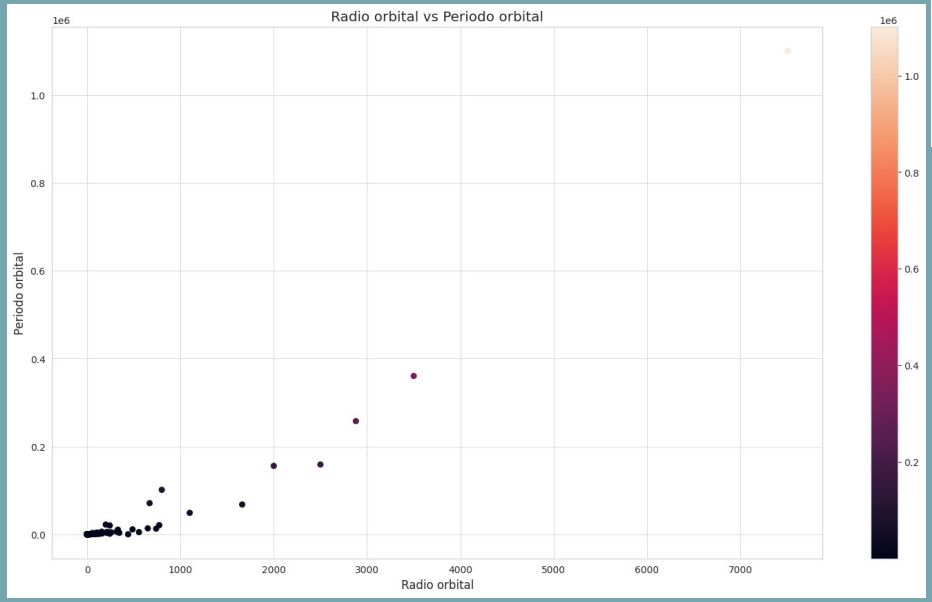




Nuestros objetivos:

Esto nos deja ver una tendencia creciente entre la distancia y la magnitud estelar en la mayoría de los planetas.



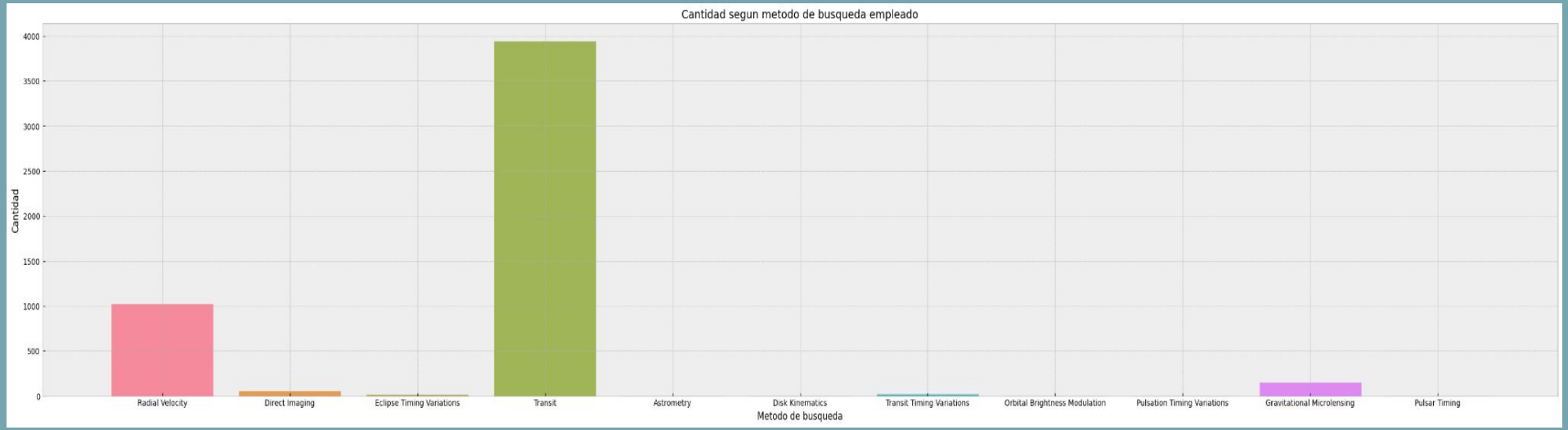


Aquí se evidencia la relación entre estas dos variables. Esto se debe ya que al aumentar el radio orbital, el planeta va a requerir un periodo orbital mayor para recorrer su órbita alrededor de su estrella

Nuestros objetivos:

¿Cómo los encontramos?

Analizando los datos podemos observar, como de los métodos, el de tránsito es el que más planetas ha capturado seguido del método de velocidad radial. El resto de los métodos ha capturado considerablemente muchos menos.

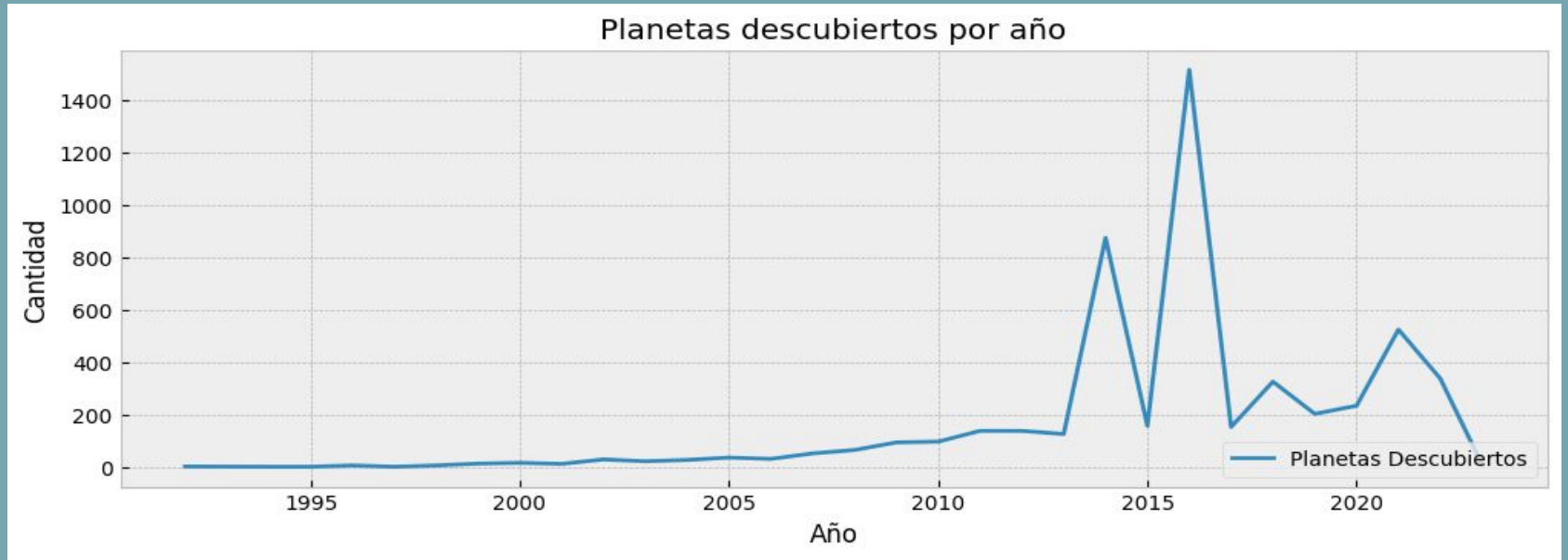


Nuestros objetivos:

¿Hubo un aumento de hallazgos a medida que pasaba el tiempo?

Podemos observar un alcance mayor al pasar los años en cuestión de distancia. Es decir, con el avance tecnológico de los años, hemos logrado "ir cada vez más lejos" para poder encontrar más planetas, eso lo observamos al comparar los descubrimientos al pasar los años con respecto a la distancia (ver último gráfico).

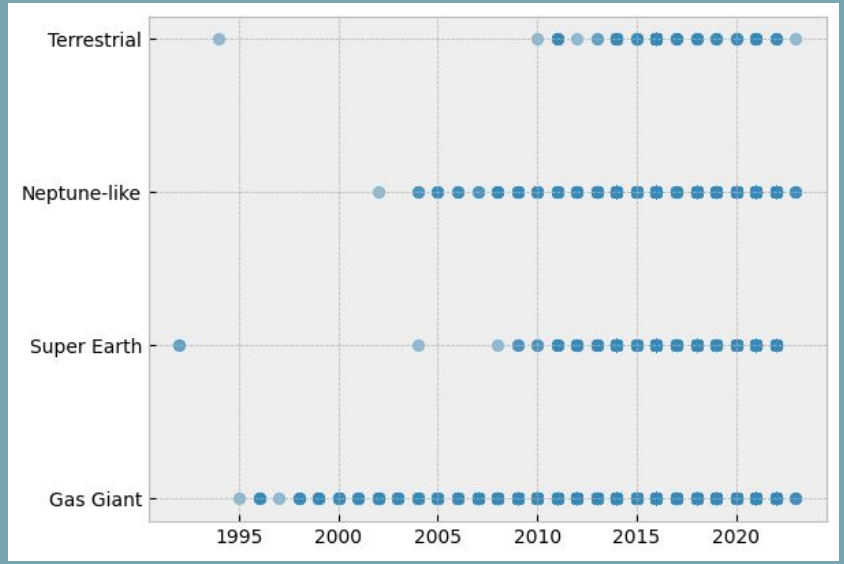
Aun así, no podemos decir que al pasar los años hemos descubierto cada vez más, esto se dio hasta el 2016/2017 llegando al pico de planetas hallados de manera creciente. En consecuencia tuvimos una baja de hallazgos con un pico leve en 2021.

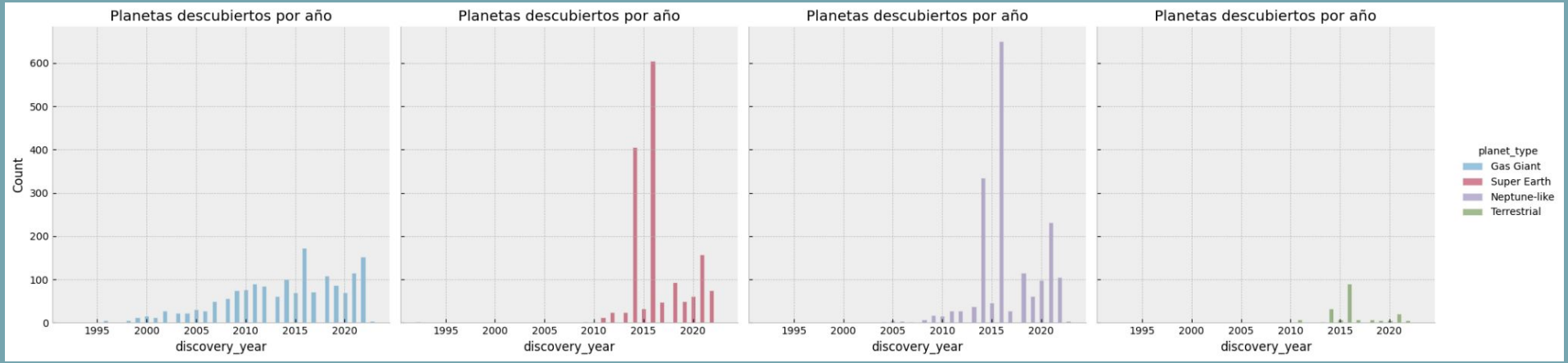


Nuestros objetivos:

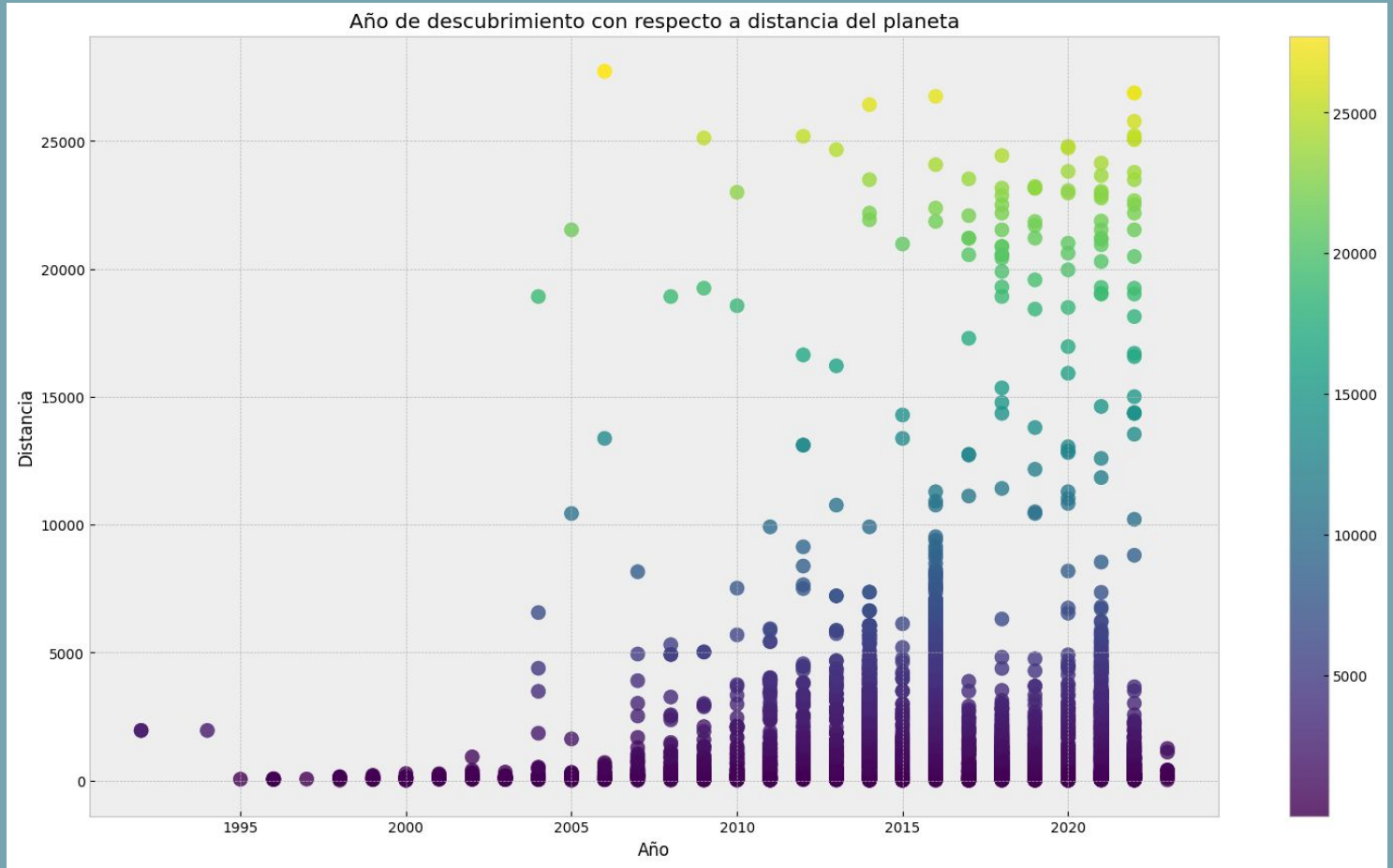
Si los vemos de manera individual por tipos de planetas, notamos un hallazgo creciente en el tiempo con respecto a los Gigantes Gaseosos.

En el caso de las Súper Tierras y los tipos Neptuno tienen un espectro de hallazgo muy similar al general anterior.





Nuestros objetivos:



INSIGHTS:

- Definimos 4 tipos de planetas:
  - Gigantes Gaseosos,
  - Super Tierras,
  - Terrestres, y
  - del tipo Neptuno, siendo este último el más poseedor.
- Nos encontramos con una relación entre el paso del tiempo y la distancia de alcance para hallar nuevos planetas. Que también está acompañada de la cantidad de planetas descubiertos al paso del tiempo, gracias al avance tecnológico.
- Entre los 11 métodos de rastreo, solo 2 resaltan:
  - Tránsito (quien se lleva todo el crédito)
  - Velocidad Radial
- Al ver las características entre planetas, los Gigantes gaseosos se repiten en ser los más lejanos y con mayor magnitud estelar, siendo que estos dos datos influyen uno con el otro.
- Otra relación se da entre el radio orbital y el periodo orbital que ambos influyen entre sí, donde mayor sea uno de igual manera el otro será mayor.

Aplicación de modelos:

Probamos sobre nuestro dataset los siguientes modelos:

Random Forest:

	precision	recall	f1-score	support
0.0	0.98	1.00	0.99	471
1.0	0.96	0.99	0.97	553
2.0	0.98	0.95	0.96	498
3.0	0.98	0.83	0.90	52
accuracy			0.97	1574
macro avg	0.98	0.94	0.96	1574
weighted avg	0.97	0.97	0.97	1574

Random Forest con PCA:

	precision	recall	f1-score	support
0	0.88	0.85	0.87	345
1	0.92	0.97	0.95	367
2	0.85	0.86	0.85	300
3	0.96	0.70	0.81	37
accuracy			0.89	1049
macro avg	0.90	0.85	0.87	1049
weighted avg	0.89	0.89	0.89	1049

SVM:

	precision	recall	f1-score	support
0	0.45	0.71	0.55	418
1	0.43	0.60	0.50	460
2	0.00	0.00	0.00	386
3	0.00	0.00	0.00	48
accuracy			0.44	1312
macro avg	0.22	0.33	0.26	1312
weighted avg	0.29	0.44	0.35	1312

Cross Validation:

Utilizamos un Stratified K-Fold con 5 iteraciones sobre un Random Forest

```
Iteracion:  1  Accuracy: 0.9475691134413727
Iteracion:  2  Accuracy: 0.9532888465204957
Iteracion:  3  Accuracy: 0.9447092469018112
Iteracion:  4  Accuracy: 0.9399428026692088
Iteracion:  5  Accuracy: 0.9551954242135366
Accuracy promedio: 0.9481410867492851
```

Se implementó exitosamente la técnica de validación cruzada utilizando Stratified K-Fold para evaluar el rendimiento de nuestro modelo. Los resultados obtenidos tras cinco iteraciones demostraron una alta precisión en la clasificación, con una media de precisión (accuracy) promedio del 94.81%. Estos valores indican una eficaz capacidad de generalización y predictibilidad de los modelos desarrollados.