POLITECNICO
MILANO 1863

Giovanni Zhang Filippo Sergenti Gabriele Rolleri

EXOPLANETS

statistical analysis of planetary habitability

“Two things fill the mind with ever new and increasing admiration and awe
the more often and steadily we reflect upon them:
the starry heavens above me and the moral law within me”

Immanuel Kant - 1788 - Critique of Practical Reason

Introduction

We live in a golden age of exoplanet discovery and characterization.

Since its launch in 2009, the Kepler mission has discovered over 4,000 exoplanets, enabling detailed studies of their physical and orbital properties. In some cases, Kepler could even determine their compositions, thus providing essential clues to the formation and evolution of planetary systems.

Astrobiology & Cosmology

Who are we? Where do we come from? Are we alone in the universe?

We have been asking these questions since the beginning of time.

The choice of this paper came not only from the fascination with these three questions but also from the possibility of addressing them rigorously and scientifically.

For obvious reasons, these questions are still open, and we authors think they will remain so for a few million more years; although a little discouraged, we try to aggregate all the available material on the subject validating it through sophisticated statistical tools.

What’s an exoplanet anyway?

An exoplanet is a planet not belonging to the solar system: every planet orbiting around a star other than the Sun.

Fully confirmed only in 1995, the existence of exoplanets was for a long time considered more than plausible. The first hypothesis of the existence of these celestial bodies was formulated by Isaac Newton in 1713 in the "General Scholium" that concludes his Principia.

As of June 29th 2021, 4700 exoplanets are known in 3472 different planetary systems; 2487 is the number of candidate planets, while 209 are still awaiting confirmation.

The discovery of exoplanets is made possible only by indirect observation methods.

Due to the limitations of current observational techniques, most of the planets detected are gas giants like Jupiter and, to a lesser extent, massive rocky planets of the Super-Earth type (2 to 10 Earth masses). The thousands of exoplanets available for characterization allow us to conduct robust statistical studies. The sizes, masses, compositions, and orbital dynamics of these planets give us clues to their formation

Table of Contents

Dataset

In this table are listed all the variables considered for our study.

In the following pages, we will extensively treat each parameter, studying its statistical distribution, the scientific phenomenon behind it, how it is measured, and its importance in determining a planet's habitability.

Data is freely accessible at the following webpage: exoplanetarchive.ipac.caltech.edu

Planet Name (pl_name)	Name of the planet most used in literature
Number of Stars (sy_snum)	Number of stars in planetary system
Number of Planets (sy_pnum)	Number of planets in planetary system
Year of Discovery (disc_year)	Year the planet was discovered
Discovery Method (discoverymethod)	Method by which the planet was first discovered
Orbital Period (pl_orbper)	Time taken by the planet to complete an orbit around the star or host system
Orbit Semi-Major Axis (pl_orbsmax)	The longest radius of an elliptical orbit
Planet Radius (pl_rade)	Segment that connects the center of the planet to its surface. Measured in units of the Earth's radius
Planet Mass (pl_bmasse)	Best available estimate of the planet mass
Eccentricity (pl_orbeccen)	Amount by which the planet's orbit deviates from a perfect circle
Insolation Flux (pl_insol)	Measure of the incident stellar radiation. It is expressed in units relative to those measured for Earth.
Equilibrium Temperature (pl_eqt)	Temperature of the planet as modeled by a black body heated only by its host star
Stellar Effective Temperature (st_teff)	Temperature of the star modeled by a black body that emits the same total amount of electromagnetic radiation
Stellar Radius (st_rad)	Segment that connects the center of the star to its surface.
Stellar Mass (st_mass)	Amount of matter contained in the star, measured in units of the Sun's mass
Right Ascension (ra)	Right ascension of the planetary system, measured in decimal degrees
Declination (dec)	Declination of the planetary system, measured in decimal degrees

Part I: Descriptive Statistics

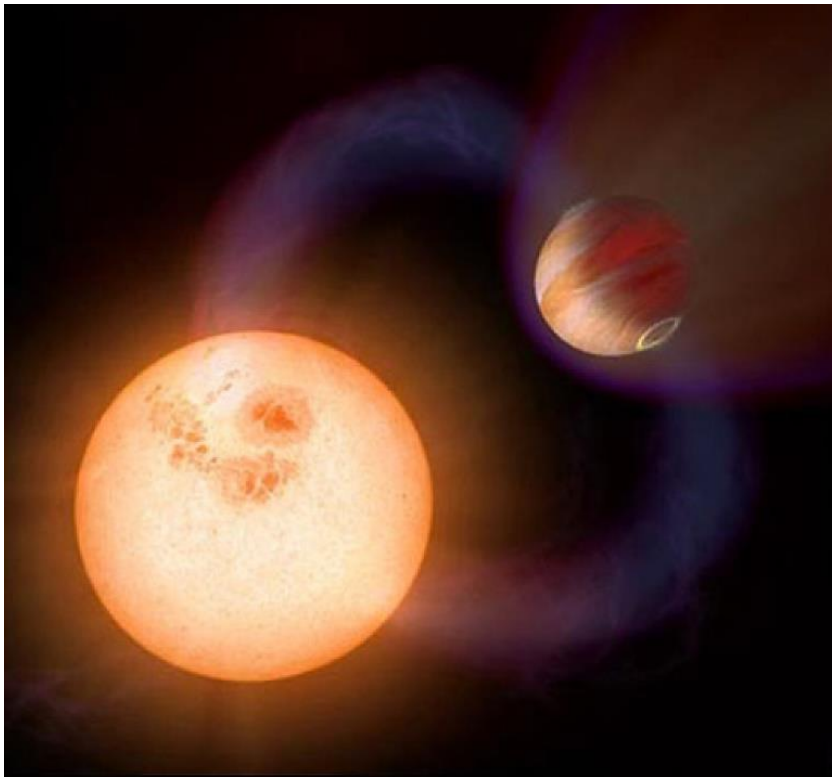
We want to start with an exploratory analysis of the available dataset to deduce its general characteristics and trends.

Dataset Bias

Astronomers wonder why many large gas giant exoplanets are close to their star compared to those in our solar system. For example, *T Bootis* has a planet four times the size of Jupiter at less than a quarter of the Earth-Sun distance. *HD 114762* has a planet eleven times the size of Jupiter, at less than half AU. One possible answer is that today's search methods favor the detection of these kinds of systems: **a large planet placed at a small distance amplifies the star's oscillations, and they are easily detected by the Doppler effect**. At a greater distance, a smaller planet causes much smaller and harder-to-see oscillations. Another explanation is that the planets formed at greater distances then move inward due to each other's gravitational interactions. This model has been called **"the jumping Jupiters model."**

As just said, most of the discoveries involve gas giants orbiting their stars at short distances. These types of planets, called **"hot Jupiters,"** significantly affect the radial velocity of their stars and frequently transit in front of them, facilitating their detection. Due to the selection effect, such planets have clear quantitative supremacy over the others in our dataset. However, with the improvement of our tools & instruments, the trend is reversing; it is, therefore, clear that the prevalence of Earth-like telluric bodies is higher than that of giant planets.

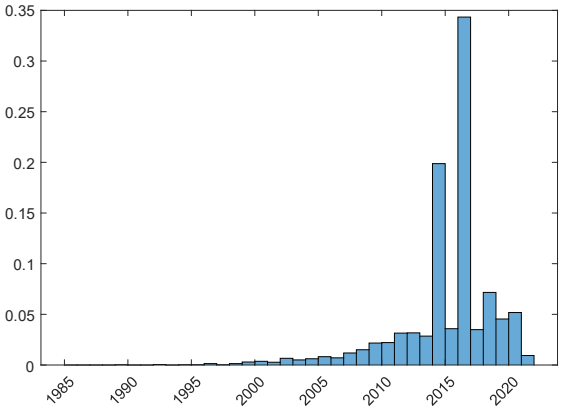
The fraction of smaller planets is constantly growing, mainly thanks to the Kepler mission, which already allows us to define an outline of an exoplanetary classification based on size.



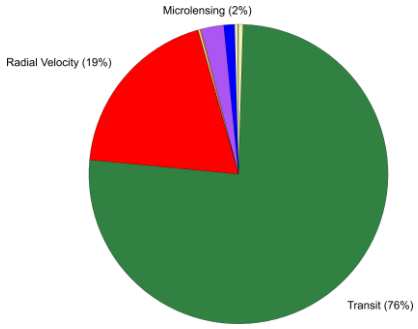
1) Discovery Year

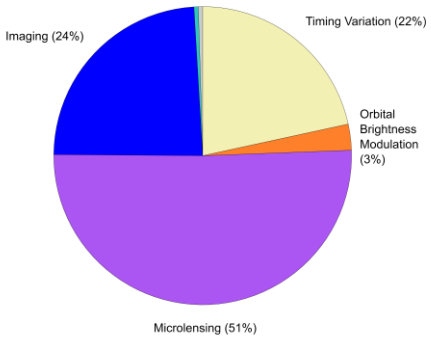
Scientific interest in exoplanets has grown increasingly since 1992, the year of the first confirmed discovery (PSR B1257 + 12).

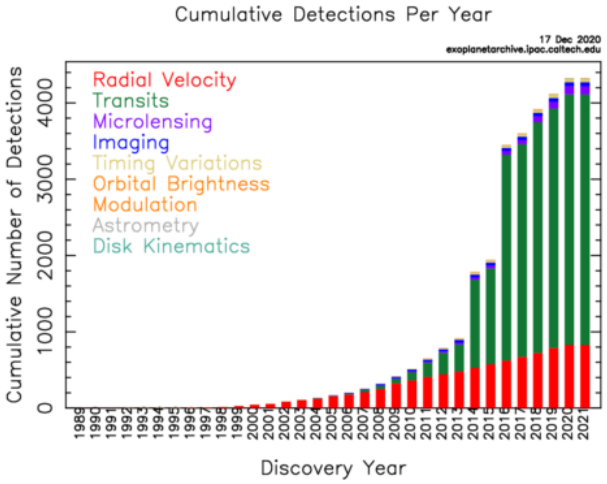
Initially, the pace of discoveries was very slow, but since the 2000s it has experienced a real surge, going from 20 planets discovered in 2000, to 189 in 2011, to almost 3500 in 2016.



2) Discovery Method



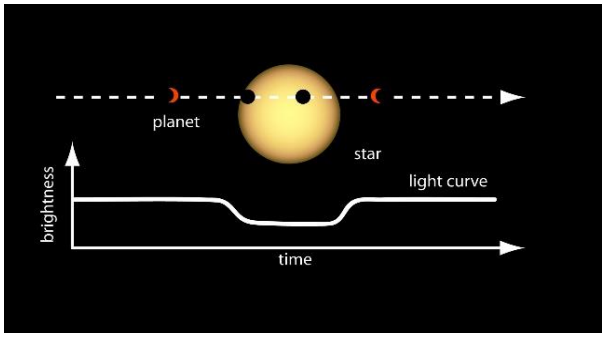




Discovery Method is a categorical variable that describes the process by which the planet was discovered.

The two prevalent methods are Transit and Radial Velocity, which account for about 95% of all new discoveries.

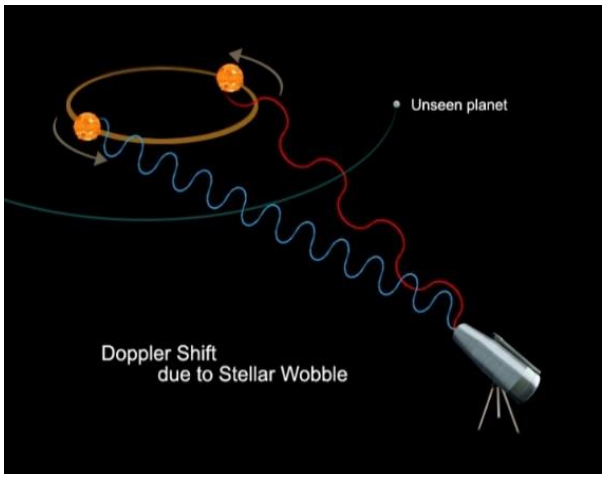
Transit was a real revolution for the field of exoplanetology since it considerably increased the number of planets discovered yearly.



Transit

Most recent and promising method.

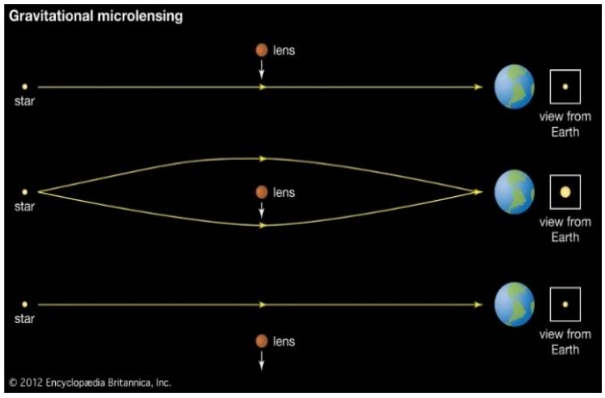
It occurs when a planet passes between a star and its observer. We can record a decrease in brightness during the event.



Radial Velocity

It is possible to detect a planet's presence by observing the speed at which its star moves relative to the Earth (known as imbalances in the star's spectral line).

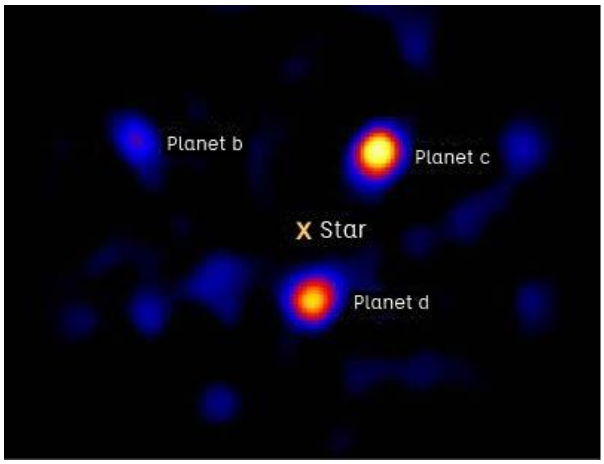
This method can easily locate planets very close to their star. However, it does not work for planets with long orbital periods since the more frequent it happens, the more we are sure the planet exist



Microlensing

Gravitational fields of two celestial bodies can cooperate to focus the light of a distant star. However, perfect alignment happens rarely, and the effect is minimal.

This method makes it possible to discover planets of comparable mass to Earth using today's technologies.

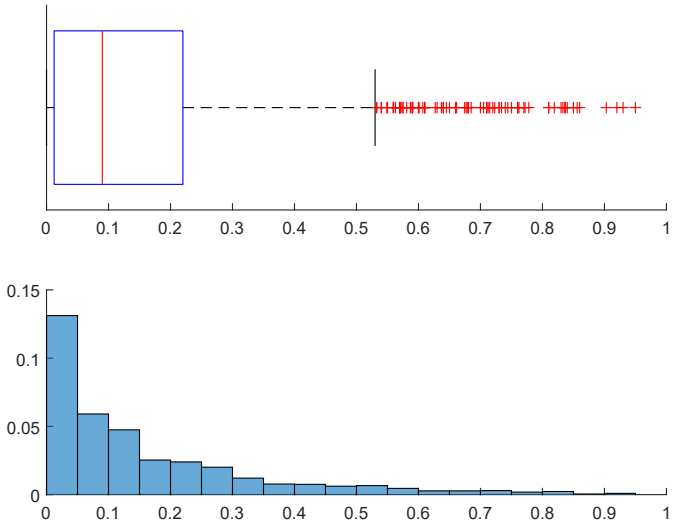


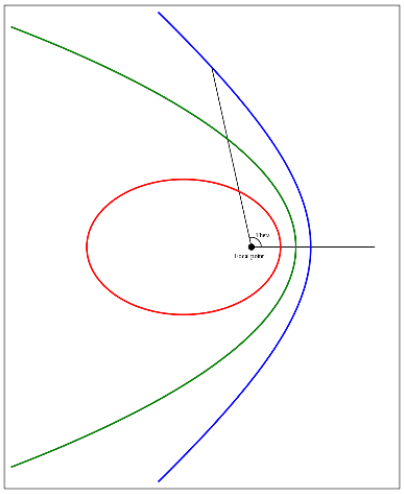
Direct Imaging

Direct imaging uses infrared wavelengths to observe planets.

This method works for planets very far from their stars but does not allow astronomers to measure their mass directly. Studying the star's spectrum & brightness can reveal informations about temperature, composition, and diameter of an exoplanet

3) Eccentricity





This dimensionless parameter indicates how much the orbit deviates from a perfect circle.

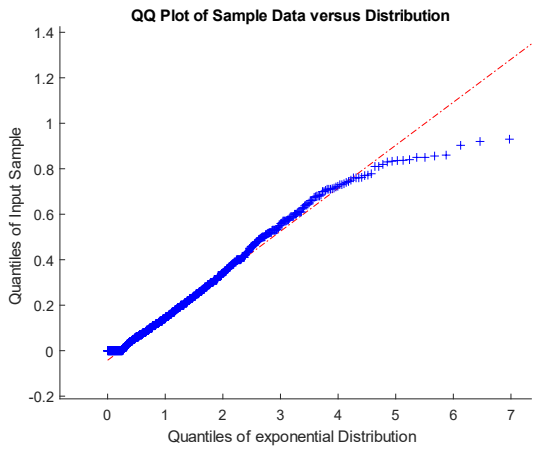
0 indicates a circular orbit, values **between 0 and 1** indicate an elliptical orbit, **1** indicates a parabolic escape orbit, and **greater than 1** indicates a hyperbole (free floating planet).

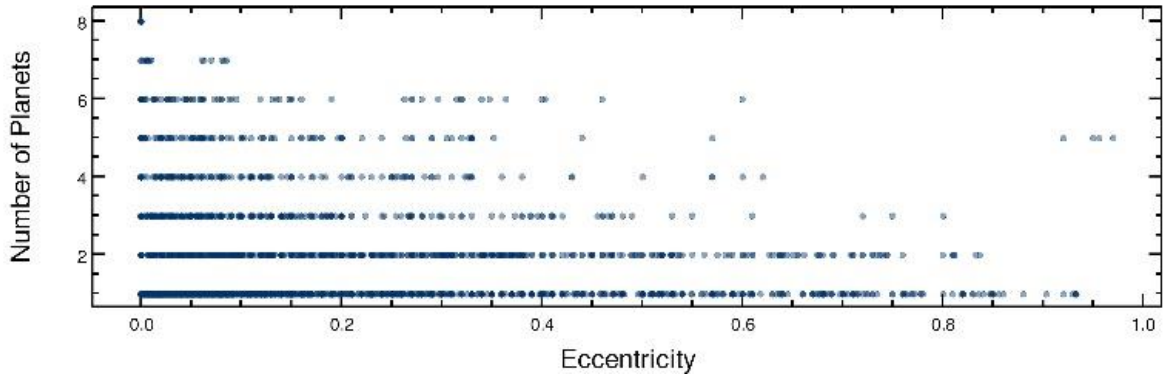
A planet with a stable orbit cannot have eccentricity greater than or equal to 1.

The Earth's orbit is almost circular, with eccentricity less than **0.02**.

We can see how the distribution is strongly shifted to the left with a very long right whisker and a modest amount of outliers beyond that. The median informs us that **more than half of the planets have eccentricity <0.1**, and the histogram suggests that the sample may follow an exponential distribution. We then proceed to make a QQ-plot for this distribution.

The center of the sample adheres very well to the line; while the high number of null values creates a slight deviation in the left tail, the right tail also deviates from the line quite significantly. Ultimately, the distribution is not perfectly exponential but follows its characteristics.





The data collected have surprised most researchers**: 90% of the eccentricities are greater than those of the planets in the solar system.**

These facts have been widely interpreted to indicate that the Solar System is an atypical member of the overall population of planetary systems.

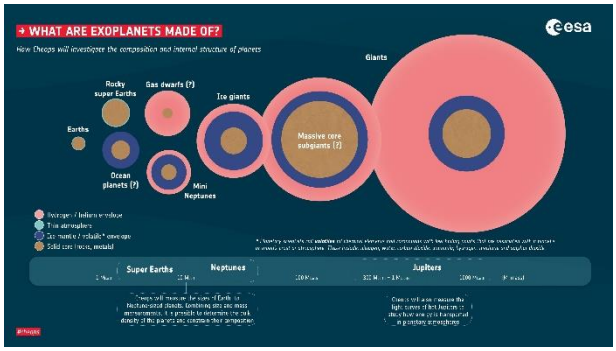
Furthermore, orbital and rotational stability is essential for the celestial body to be habitable. The greater the orbital eccentricity, the greater the temperature fluctuation on the planet's surface. The planet must also have moderate seasons and a reasonable day-night cycle

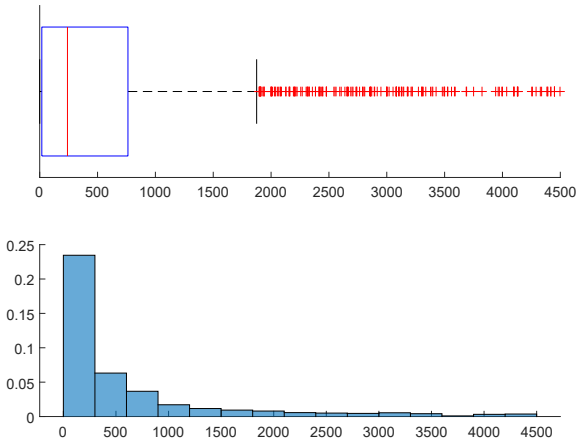
Despite adapting, living organisms cannot withstand excessive temperature variations, especially if they reach the boiling and freezing point of the planet's **main biotic solvent** (on Earth, it is liquid water).

We report a strong anti-correlation of orbital eccentricity with **multiplicity** (number of planets in the system). If low eccentricities actually favor high multiplicities, habitability may be more common in systems with a larger number of planets.

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA’s
0.0000	0.0125	0.0900	0.1525	0.2200	0.9500	2780

4) Mass





The distribution of the planets' masses is very similar to that of eccentricities but with values even flattened towards zero and a series of even more numerous right-hand outliers. The same tests carried out previously reveal again that the distribution is far from the exponential one. As for the eccentricity, we note that the mass entry-point is missing for a large part of our sampl**e** (57.5%, 2520 out of 4383)

Low-mass planets would be bad candidates for life for the following reasons:

Their *gravity would be lower and their atmosphere less dense*. The molecules that make up life have a much higher probability of reaching escape velocity and being ejected into space by the propulsion of the solar wind or by a collision.

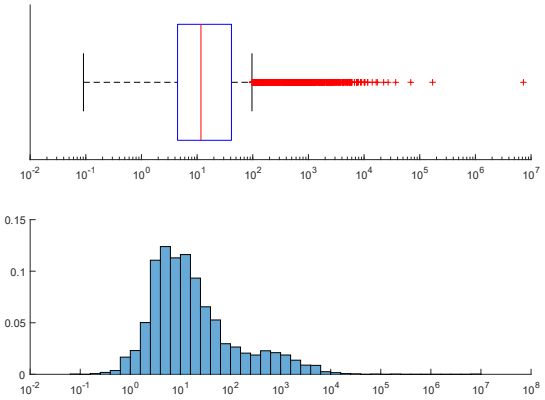
A thin atmosphere meas less protection against high-frequency radiation and meteorites, do not have enough matter for the initial biochemistry, are not thermally insulated enough, and have low thermal conductivity across their surface.

Furthermore, smaller planets have a smaller diameter and, therefore, *greater surface-to-volume* ratio than larger planets. Such bodies lose energy much more rapidly after their formation and have little geological activity. They do not have volcanoes, earthquakes, or tectonic activity that provide the surface with elements favorable to life and the atmosphere with molecules capable of regulating the temperature (such as carbon dioxide).

Our Earth is large enough for its gravitational force to hold back its atmosphere and for its liquid core to remain active and hot, thus generating geological activity on the surface. In addition, the disintegration of radioactive elements in the planet's heart is another heat resource. On the other hand, Mars is nearly inactive and has lost most of its atmosphere. **Therefore, it is conceivable that the minimum mass of a planet that can be habitable lies between that of Mars and Earth.**

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
0.02	18.30	240.00	744.13	762.79	17668.17	2520

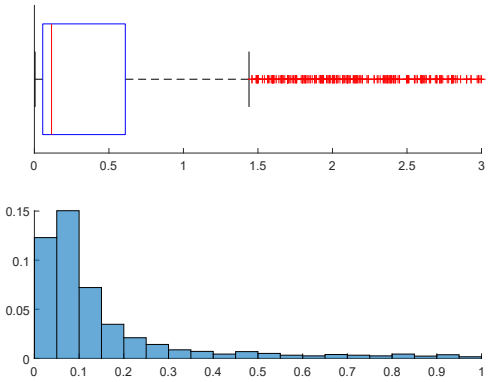
5) Orbital Period

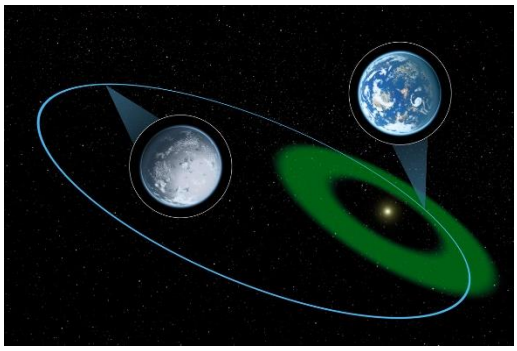


Given the range of data, varying from less than one day to more than 7 million days, it was necessary to introduce a log scale. Unfortunately, this choice makes the probability (density) on the Y-axis unusable. In any case, the distribution has two maximums, one higher than the other. The second maximum includes most of the right-hand outliers. This high number of outliers and extreme outliers make the average roughly 200 times greater than the median.

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
0	4	12	2048	41	7300000	147

6) Major Semi-Axis

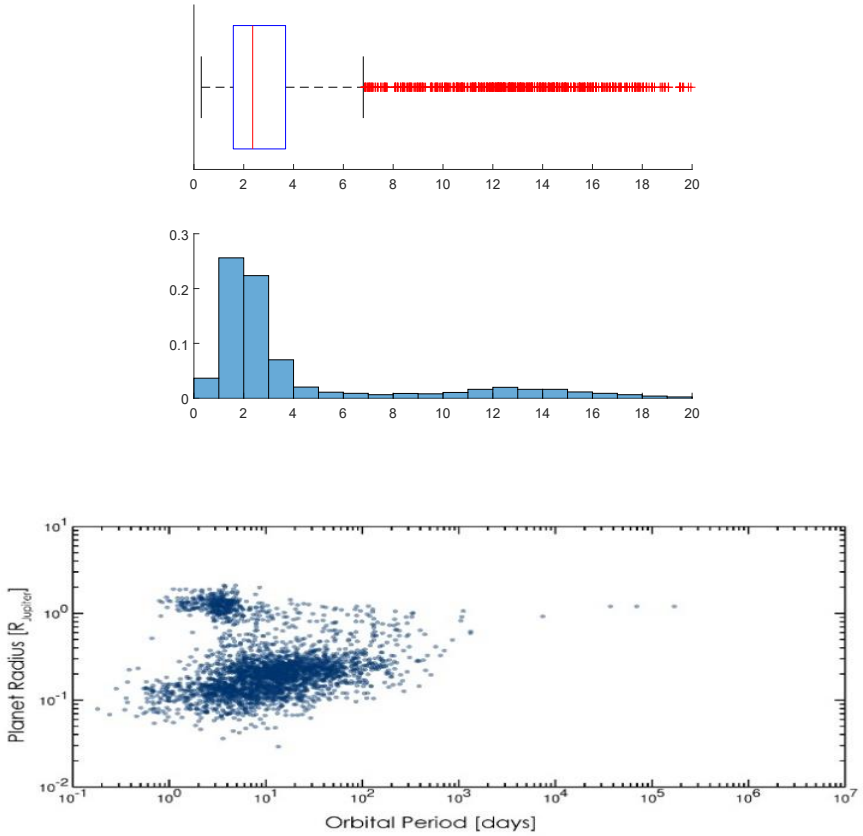




As the positions of the 1at & 2nd quartiles suggest, the distribution of the semi-major axis is dense towards zero with a long tail and outliers thousands of times larger than the average.

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
0.004	0.056	0.115	8.620	0.608	3500.00	1749

7) Radius



The Fulton gap, photoevaporation valley, or **"Sub Jovian Desert"** is an observed scarcity of planets with radii between 1.5 and 2 times the radius of the Earth.

This bimodality in the exoplanet population was first observed in 2013. It was noted as a possible confirmation of an emerging hypothesis that photoevaporation could drive atmospheric mass loss on nearby, low-mass planets.

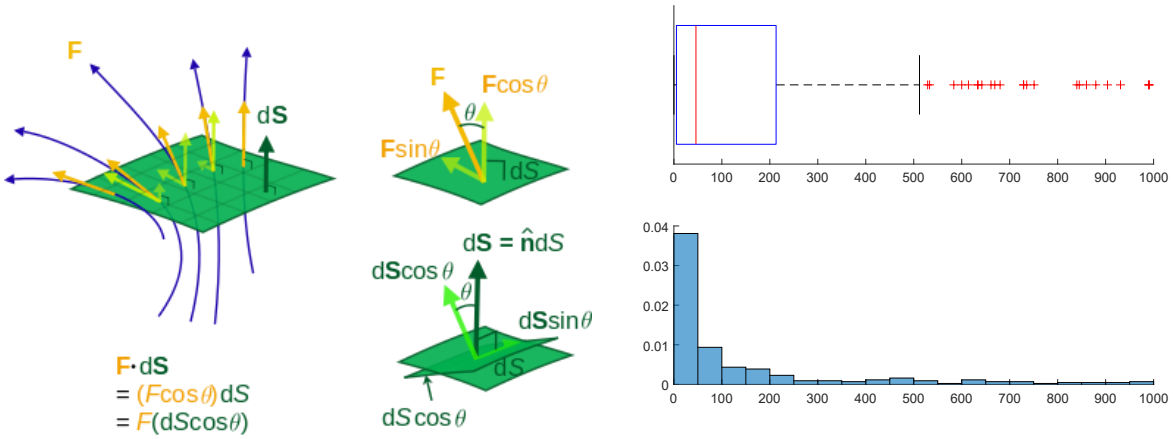
This phenomenon would lead to a population of bare rock cores with smaller radii at small separations from their parent stars and planets with thick envelopes dominated by hydrogen and helium with larger radii at greater separations.

The term **"photoevaporation"** refers to the process by which the atoms or molecules are torn away from an accumulation (a planet's atmosphere, circumstellar disk, or nebula) by high-energy photons emitted by a star.

Despite the implication of the word "gap," this does not represent a radius range completely absent from the observed population of exoplanets but rather a range that appears to be relatively rare. Consequently, 'valley' is often used instead of 'gap.'

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
0.296	1.585	2.360	4.287	3.680	77.342	1024

8) Insolation Flux



This coefficient measures the total amount of electromagnetic radiation incident on the planet's surface calculated under the **black body hypothesis**.
Although the temperature is constant in the data reported, it is essential to note that **all stars have variations in brightness**, and the amplitude of these fluctuations is very different from one star to another.

Most stars are relatively stable, but a significant minority are not and often have both drops and sudden increases in brightness. As a result, orbiting bodies receive abrupt fluctuations in radiated energy. *These stars are, therefore, bad candidates to host planets capable of allowing life* as the strong variations in energy flux (and thus temperature) negatively impact organisms' survival.

For example, living things adapted to a particular temperature domain would likely have problems surviving major temperature changes. Furthermore, variations in brightness are generally accompanied by the *emission of massive doses of gamma rays* and X-rays, radiation that could be lethal to biological organisms.

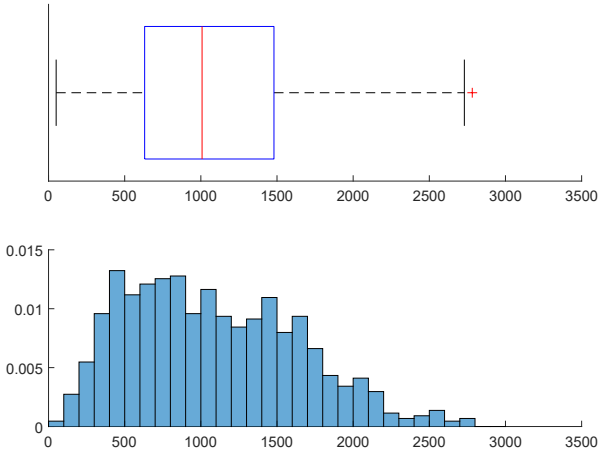
The *atmosphere* of the planets can attenuate such effects (a 100% increase in stellar brightness does not necessarily imply a 100% increase in the planet's temperature). However, it is also possible that such planets will not be able to hold their atmospheres due to the strong incident radiations.

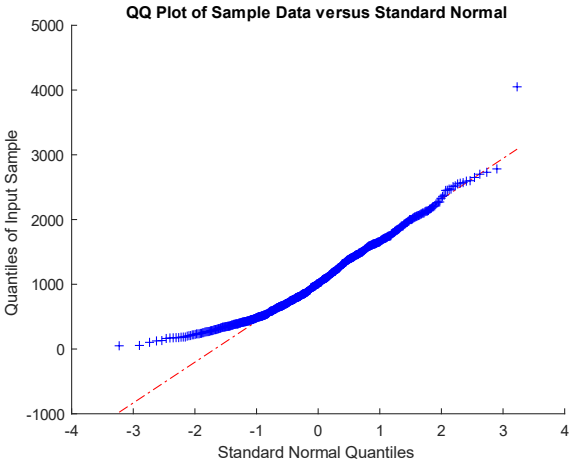
The Sun does not have this type of variation: during the solar cycle the difference between the minimum and maximum brightness is around 0.1%

The distribution is noticeably skewed, with a tail pronounced to the right. Confirming this is also the fact that the mean is larger than the median. **In the whole dataset this is the parameter with the least number of entries** (92.5% missing).

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
0.06	5.40	46.00	428.19	213.00	44900.00	4054

9) Equilibrium Temperature





The planetary equilibrium temperature is the theoretical temperature that a planet would reach if it were a **black body**, heated only by its star.

In this model, the presence of an **atmosphere** (and therefore a possible greenhouse effect) is not considered and the temperature is attributed to an idealized surface of the planet.

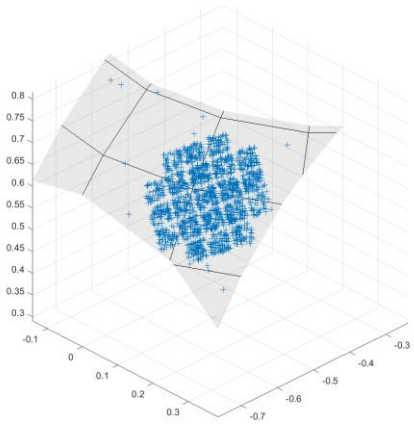
It is the most similar Gaussian distribution on our dataset. The average, being more sensitive to the data on the extremes, is slightly greater than the median, which justifies the hint of a tail on the right.

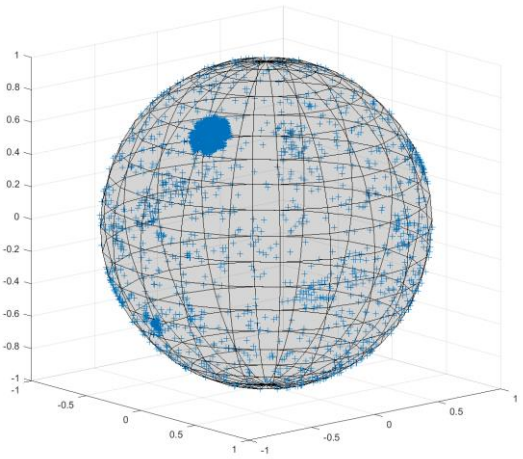
The planet's surface temperature plays a crucial role in the search for habitable exoplanets, but there are currently no direct measurements available. Many physical processes affect a planet's surface temperature distribution. However, the dominant influence is an energy balance between the stellar radiation input and the radiative surface heat loss. With the assumptions of a uniform planetary surface temperature, no filtering of incoming radiation, and black body emission, the only variables are the stellar brightness and the exoplanet's radial distance from the star.

Some exoplanets also revealed the presence of **high-speed winds** on the surface with peaks of 14,000 km / h. These winds keep the temperature of these planets constant over the entire surface with minimal variations.

Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA’s
50	631	1008	1082	1480	0.4050	3579

10) Spatial Distribution

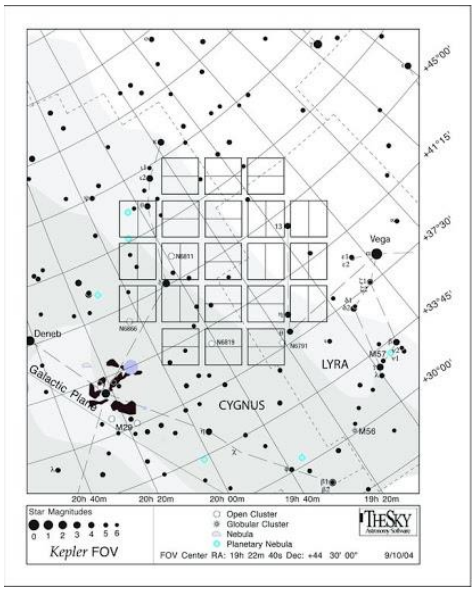


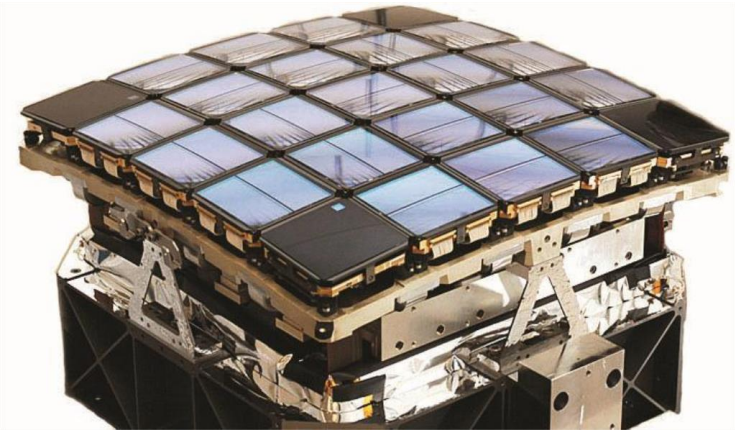


The distribution of exoplanets relative to the earth is entirely random. However, there is a very significant bias in the dataset. The pattern in the first figure is not accidental: it represents the photometer of the Kepler Space Telescope.

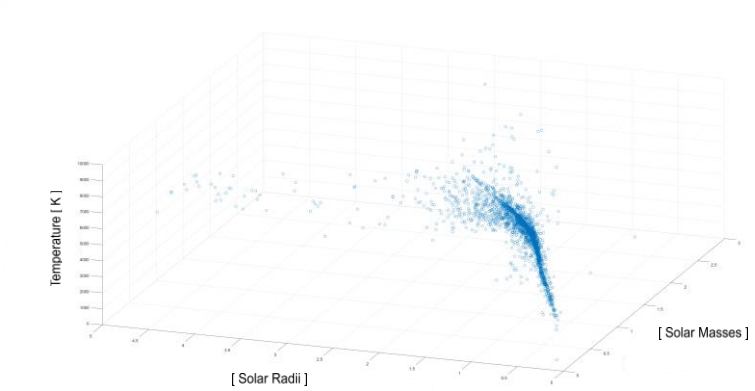
The Kepler Space Telescope was specifically designed to monitor a portion of the Milky Way region to find out how many of the billions of stars in our galaxy could potentially host habitable planets.

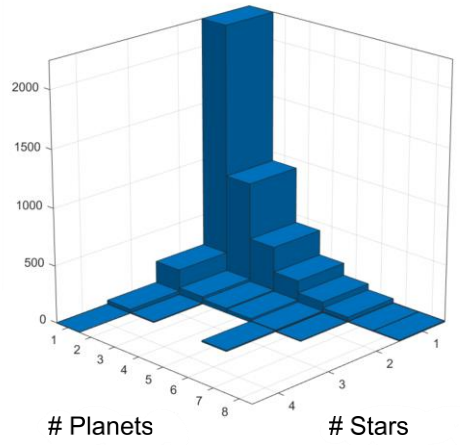
The photometer constantly monitors the brightness of more than 145,000 main sequence stars near the constellations of Cygnus, Lyra, and Dragon in its fixed field of view.

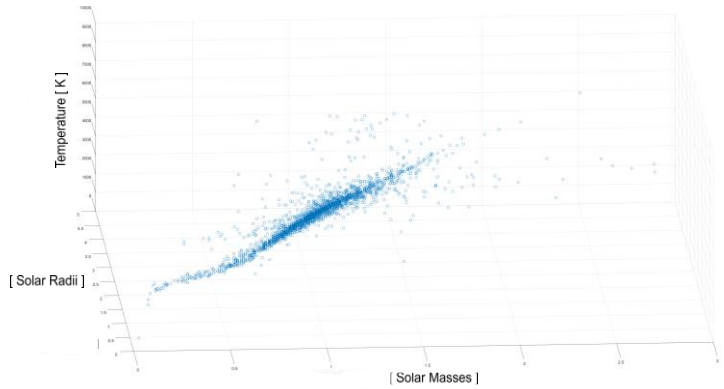




11) Stellar Parameters







The last three parameters concern the stars' mass, radius, and temperature.

The scatterplot of these three shows a strong correlation, so we tried to find a multiple linear regression model. Unfortunately, the model performed poorly even by removing the outliers and including quadratic and cubic terms for the two predictors (mass and radius).

Analyzing the components of planetary systems, as expected, we note that **one-star systems** are the most common.

However, it is essential to note that the number of single-planet systems is clearly due to a strong bias related to the nature of the research and methodology.

It cannot be assumed that there are predominantly single-planet systems in nature.

Part II: Hypothesis Testing

Following the review of the fundamental parameters that characterize exoplanets, we want to subject these elements to a study that allows us to deduce a model that can predict their temperature. To this end, we want to support two assertions through hypothesis tests:

1) The average temperatures do not depend on the number of stars in the system
2) Discovery methods generated a strong bias in the data pool

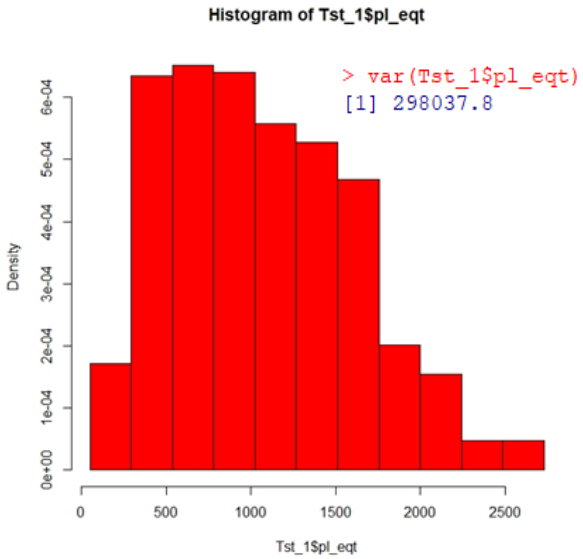
We proceed by dividing the temperatures by the number of stars into categories.

Temperatures of planets with a **single star**:

T_st1 692 data points

Classes $1 + \log_2(692) = [10.43] = 11$

```
> summary(Tst_1$pl_eqt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   50.0   636.5  1001.5  1069.8  1460.0  2730.0
```

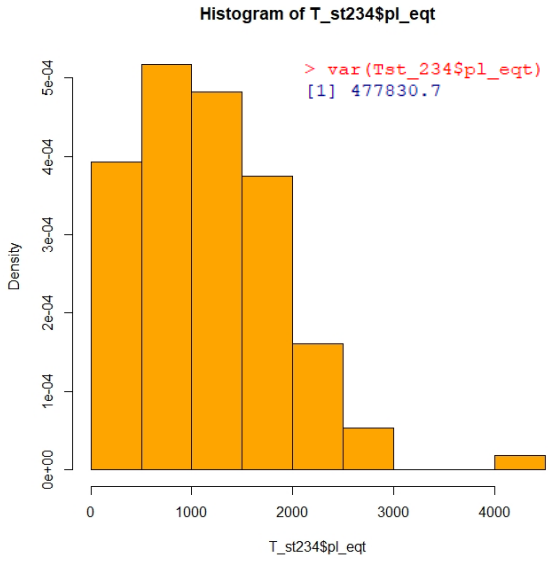


Temperatures of planets with **more than two stars**:

T_st234 112 data points

Classes $1 + \log_2(112) = [8.8] = 9$

```
> summary(Tst_234$pl_eqt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  125.0   562.8  1102.5  1158.3  1585.5  4050.0
```



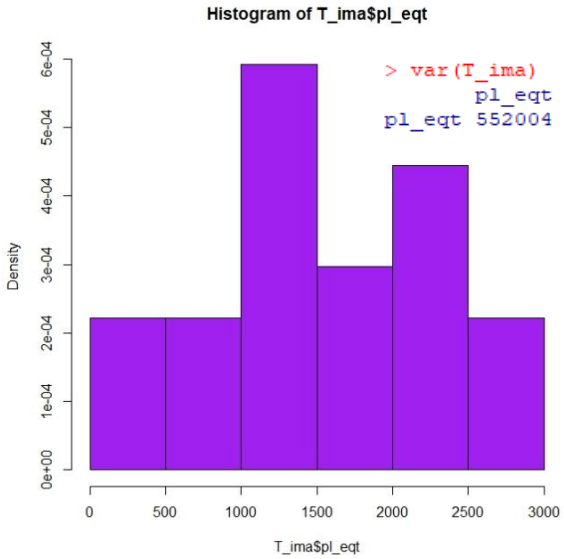
We also divide temperatures into categories by method of discovery.

Planetary temperatures measured by **imaging**

T_ima 27 data points

Classes 1 + $\log_2(27)$ = [5.75] = 6

```
> summary(T_ima$pl_eqt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    323    1100    1500    1558    2208    2700
```

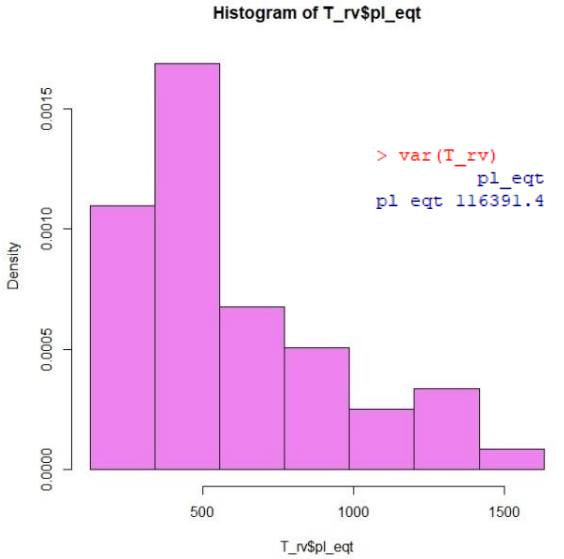


Planetary temperatures measured by **radial velocity**

T_ rv 55 data points

Classes 1+ $\log_2(55)$ = [6.78] = 7

```
> summary(T_rv$pl_eqt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  125.0   350.5   468.0   571.7   741.5  1632.0
```

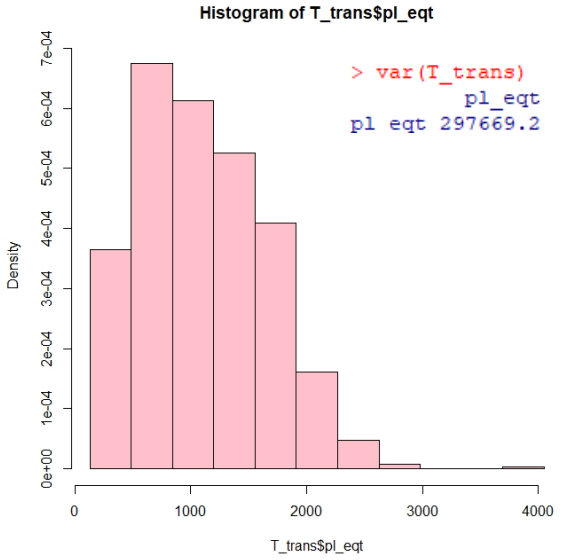


Planetary temperatures measured by **transit**

T_trans 715 data points

Classes 1 + $\log_2(715)$ = [10.48] = 11

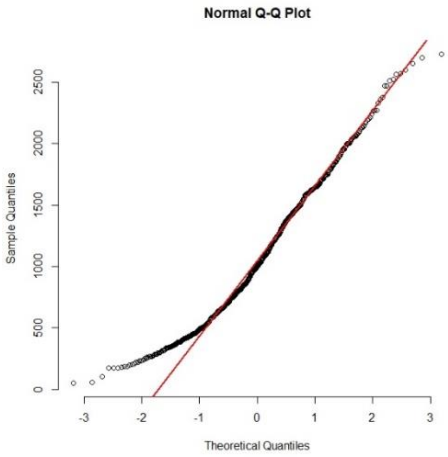
```
> summary(T_trans$pl_eqt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    131     673    1053    1109    1498    4050
```



Normality analysis

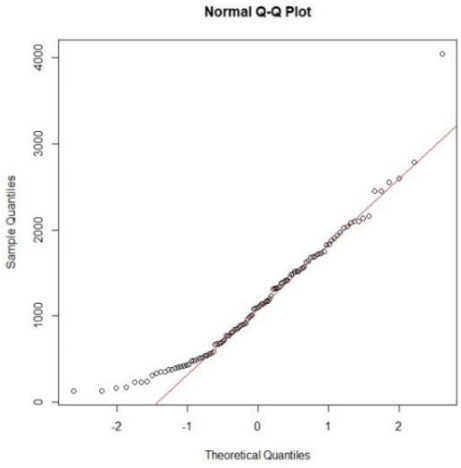
Following the subdivision of temperatures according to the categories described, we proceed by studying the gaussianity of the relative distributions.

1 star



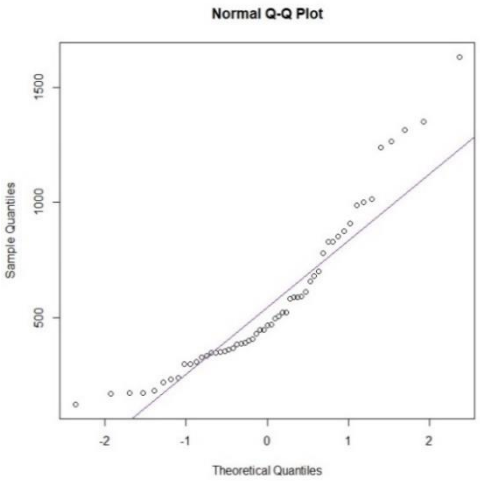

```
> shapiro.test(Tst_1$pl_eqt)

        Shapiro-Wilk normality test

data:  Tst_1$pl_eqt
W = 0.97015, p-value = 1.138e-10
```

2+ stars



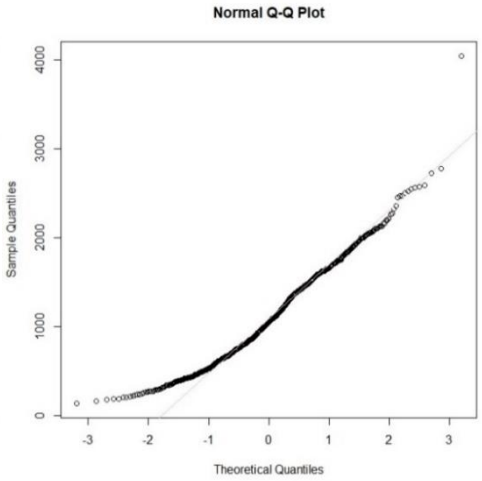

```
> shapiro.test(Tst_234$pl_eqt)

        Shapiro-Wilk normality test

data:  Tst_234$pl_eqt
W = 0.94522, p-value = 0.0001716
```

Radial Velocity



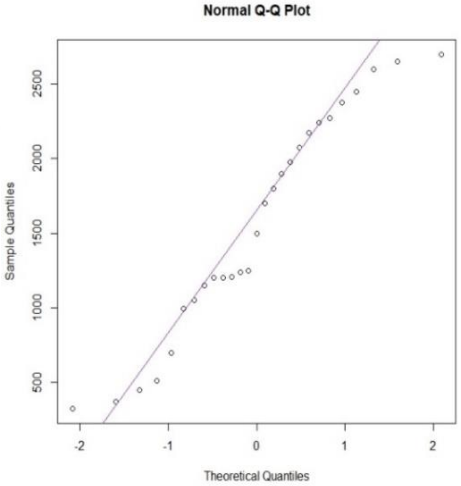

```
> shapiro.test(T_rv$pl_eqt)

        Shapiro-Wilk normality test

data:  T_rv$pl_eqt
W = 0.89347, p-value = 0.0001487
```

Transit




```
> shapiro.test(T_trans$pl_eqt)

        Shapiro-Wilk normality test

data:  T_trans$pl_eqt
W = 0.9676, p-value = 1.761e-11
```

Imaging




```
> shapiro.test(T_ima$pl_eqt)

        Shapiro-Wilk normality test

data:  T_ima$pl_eqt
W = 0.94428, p-value = 0.1554
```

Although the number of samples influenced the power of the Shapiro-Wilks test, making it very susceptible to the slightest variations from the quantile line, the QQ-plots push us to affirm the alternative hypothesis of non-gaussianity. The exception is **T_ima** which, with a low number of n = 27, has a p-value of 0.15554. In this case, there is no strong evidence to reject the null hypothesis, so the normality of its distribution is accepted.

We can assume that the sample means have an approximately normal distribution thanks to the central limit theorem since the sample is large enough (n > 30).

Number of stars

The purpose of the test is to support the idea that the average of temperatures is not correlated with the number of stars belonging to the system. To this end, a bilateral Z-test is performed on two numerous samples.

$$Z_0 = \frac{\overline{X_m} - \overline{Y_n} - \delta_0}{\sqrt{\frac{(S_x)^2}{m} + \frac{(S_y)^2}{n}}} \quad con \quad \delta_0 = 0$$

“Are the average temperatures of planets with a single star in their planetary system different from planets with multiple stars?”

$$\boldsymbol{H_0} : \mu_{(T\_st1)} = \mu_{(T\_st234)}$$

$$\boldsymbol{H_1} : \mu_{(T\_st1)} \neq \mu_{(T\_st234)}$$

$$\boldsymbol{RC} : |\boldsymbol{Z_0}| > \boldsymbol{Z_{1-\alpha/2}}$$

```
> z.test(Tst_1,Tst_234,mu=0,alternative="two.sided",sigma.x=545.92, sigma.y=691.25)

        Two-sample z-Test

data:  Tst_1 and Tst_234
z = -1.2913, p-value = 0.1966
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -222.82486   45.82568
sample estimates:
mean of x mean of y
 1069.786  1158.286
```

The test has a p-value of 0.1966. There is no strong evidence to reject the null hypothesis. It is concluded that the average temperatures of the planets are not dependent on the number of stars belonging to the system.

Discovery method

We now proceed with the definition of a second bilateral Z-test that verifies the difference in mean temperature detected according to the methodology used. The comparison takes place between the methods of Transit and Radial Velocity, on which the TLC is applicable.

"Do detection methods influence the average temperature detected with a strong bias?"

Transit VS Radial Velocity

$$H_0 : \mu_{(T\_rv)} = \mu_{(T\_trans)}$$
$$H_1 : \mu_{(T\_rv)} \neq \mu_{(T\_trans)}$$
$$RC : |Z_0| > Z_{1-\alpha/2}$$

```
> z.test(T_rv,T_trans,alternative="two.sided",mu=0,sigma.x=341.16, sigma.y=545.59,conf.level=0.95)

        Two-sample z-Test

data:  T_rv and T_trans
z = -10.677, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -635.9185 -438.6521
sample estimates:
mean of x mean of y
 571.7455 1109.0308
```

The extremely low p-value allows us to have strong evidence against the null hypothesis. Therefore, the alternative hypothesis is assumed to be true.

Radial Velocity VS Imaging

The subsequent Z-test between **Radial Velocity** and **Imaging** assumes the normality of the sample mean of the data taken from the Imaging although it has a rather borderline number n = 27 < 30. This assumption is valid since it has been verified with a Shapiro-Wilks test that its normality is not refutable. The same goes for the unknown variance, which can be estimated with good precision with:

$$S^2{}_y = \frac{\sum_{i=0}^{n}(x_i - \overline{x})}{n-1}$$

$$H_0 : \mu_{(T\_rv)} = \mu_{(T\_ima)}$$

$$H_1 : \mu_{(T\_rv)} \neq \mu_{(T\_ima)}$$

$$RC : |Z_0| > Z_{1-\alpha/2}$$

```
>  z.test(T_rv,T_ima,alternative="two.sided",mu=0,sigma.x=341.16, sigma.y=742.96,conf.level=0.95)

        Two-sample z-Test

data:  T_rv and T_ima
z = -6.5643, p-value = 5.229e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1280.3462  -691.5703
sample estimates:
mean of x mean of y
 571.7455 1557.7037
```

The result obtained, as could already be assumed, has an extremely low p-value, giving strong evidence to reject the null hypothesis and accept that alternative. We can again say that the means are different.

Transit VS Imaging

In this last test, all the assumptions made in the previous point are held.

$$H_0 : \mu_{(T\_trans)} = \mu_{(T\_ima)}$$

$$H_1 : \mu_{(T\_trans)} \neq \mu_{(T\_ima)}$$

$$RC : |Z_0| > Z_{1-\alpha/2}$$

```
> z.test(T_trans,T_ima,alternative="two.sided",mu=0,sigma.x=545.59, sigma.y=742.96,conf.level=0.95)

        Two-sample z-Test

data:  T_trans and T_ima
z = -3.1065, p-value = 0.001893
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -731.7529 -165.5929
sample estimates:
mean of x mean of y
 1109.031  1557.704
```

The test has a low p-value. The null hypothesis is rejected, and the alternative is accepted.

In conclusion, it can be said that the average of temperatures varies strongly according to the methods of detection of the planets. We therefore affirm the presence in the pool of a strong bias in favor of the method used, favoring the presence of data relating to the “Transit” discovery method (the most numerous).

Rover Survival Temperature

A fundamental role in the study of extraterrestrial planets is played by small albeit sophisticated robots, capable of operating under extreme conditions to provide necessary data to study the universe: **rovers**.

We now want to study the probability that a rover can survive on a desired exoplanet.

To this end, we considered one of the most famous if not the most resistant rovers, capable of withstanding up to extreme temperatures of 1300 °C**: Perseverance**.

Due to the nature of **pl_eqt** (planet's equilibrium temperature), a parameter in our possession, it is assumed that all the planets are black bodies at constant temperature.

We then proceed by carrying out a sample proportion test. The sample size **n = 804 > 30** allows us to find a precise $\widehat{\boldsymbol{p}}$ estimator of the expected value of the Bernoulli variable.

It is assumed that each random variable is independent and identically distributed.

$$p_i = \begin{cases} 0 & \text{if } T_i \geq 1573\ K \\ 1 & \text{if } T_i < 1573\ K \end{cases}$$

"The probability that Perseverance withstands the planet's temperature is less than 79%"

$$Z_0 = \frac{\overline{X_n} - \mathrm{p}_0}{\sqrt{\mathrm{p}_0(1-\mathrm{p}_0)}}\sqrt{n}$$

$$\boldsymbol{H_0} : \boldsymbol{p} \geq \mathrm{p}_0 = 0.79$$

$$\boldsymbol{H_1} : \boldsymbol{p} < \mathrm{p}_0$$

$$\boldsymbol{RC} : |\boldsymbol{Z_0}| < -\boldsymbol{Z_{(1-\alpha)}}$$

```
> prop.test(x=635,n=804,correct="FALSE",p=0.79,alternative="less")

        1-sample proportions test without continuity correction

data:  635 out of 804, null probability 0.79
X-squared = 0.00019193, df = 1, p-value = 0.4945
alternative hypothesis: true p is less than 0.79
95 percent confidence interval:
 0.0000000 0.8124454
sample estimates:
      p
0.789801
```

Given the **p-value = 0.49 > 0.05**, there is insufficient evidence to reject the null hypothesis and accept the alternative hypothesis. Therefore, it is concluded that the probability that the rover can survive the temperature of any planet is greater than 79%.

We construct a 95% confidence interval to obtain a range of possible values for $\hat{p}$ using the following formula:

$$IC_{\gamma}(p): \ \hat{p} \pm z_{\left(1+\frac{\gamma}{2}\right)} \frac{\hat{p}\,(1-\hat{p})}{n}$$

$$\hat{p} = \frac{n_{(T<1300)}}{n} = 0.789801 \quad , \qquad \gamma = 0.95$$

```
> binom.confint(sum(T_true), length(T_true), conf.level = 0.95, method = 'all')
          method   x   n     mean     lower     upper
1  agresti-coull 635 804 0.789801 0.7602586 0.8165872
2     asymptotic 635 804 0.789801 0.7616370 0.8179650
3          bayes 635 804 0.789441 0.7611218 0.8173567
4        cloglog 635 804 0.789801 0.7599754 0.8163776
5          exact 635 804 0.789801 0.7599767 0.8174823
6          logit 635 804 0.789801 0.7602552 0.8165840
7         probit 635 804 0.789801 0.7605466 0.8168432
8        profile 635 804 0.789801 0.7607411 0.8170168
9            lrt 635 804 0.789801 0.7607494 0.8170122
10     prop.test 635 804 0.789801 0.7596430 0.8171422
11        wilson 635 804 0.789801 0.7602922 0.8165537
```

Taking as a reference the result obtained from the prop test, we build the interval:

$$IC_{(0.95)}: (\ 0.7596430\ , 0.8171422\ )$$

Placed as the center $\hat{p} = 0.79$ the interval straddles between the regions supported by the two hypotheses. Precisely the right region $\hat{p} \geq 0.79$ belongs to the null hypothesis, while $\hat{p} < 0.79$ belongs to the region of the alternative hypothesis

Part III: Multivariable regression

The goal of our regression will be to determine a model for predicting the planet's surface temperature, using predictors based on parameters of the planetary system of interest.

To begin with, all predictors that could have some correlation with the planet's temperature were included, which is why the parameters on the planet's position in the sky were immediately excluded. It will be noted that the request to have a value for each column immediately restricts the number of usable samples, which is always around a hundred.

```
>> mdl = fitlm(P,'pl_eqt~sy_snum+sy_pnum+discoverymethod+disc_year+pl_orbper+pl_orbsmax+pl_rade+pl_bmasse+pl_orbeccen+pl_insol+st_teff+st_mass+st_rad')
Warning: Regression design matrix is rank deficient to within machine precision.
> In classreg.regr/CompactTermsRegression/checkDesignRank (line 35)
In LinearModel.fit (line 1048)
In fitlm (line 121)

mdl =


Linear regression model:
    pl_eqt ~ 1 + sy_snum + sy_pnum + discoverymethod + disc_year + pl_orbper + pl_orbsmax + pl_rade + pl_bmasse + pl_orbeccen + pl_insol + st_teff + st_rad + st_mass

Estimated Coefficients:
                                                Estimate        SE        tStat       pValue
                                                ________    ________    ________    ________

    (Intercept)                                    17102       19922     0.85848     0.39279
    sy_snum                                       -28.899       44.68    -0.6468     0.51932
    sy_pnum                                       -11.392      17.247    -0.66053     0.51051
    discoverymethod_Disk Kinematics                     0           0         NaN         NaN
    discoverymethod_Eclipse Timing Variations           0           0         NaN         NaN
    discoverymethod_Imaging                             0           0         NaN         NaN
    discoverymethod_Microlensing                        0           0         NaN         NaN
    discoverymethod_Orbital Brightness Modulation       0           0         NaN         NaN
    discoverymethod_Pulsar Timing                       0           0         NaN         NaN
    discoverymethod_Pulsation Timing Variations         0           0         NaN         NaN
    discoverymethod_Radial Velocity               -55.244      81.145    -0.6808     0.49765
    discoverymethod_Transit                             0           0         NaN         NaN
    discoverymethod_Transit Timing Variations           0           0         NaN         NaN
    discoverymethod_Timing Variation                    0           0         NaN         NaN
    disc_year                                     -8.5026      9.8499    -0.86321     0.39019
    pl_orbper                                      2.1272     0.32892      6.4672   4.3033e-09
    pl_orbsmax                                    -1983.5      235.82      -8.411   4.0688e-13
    pl_rade                                        19.546      5.0304      3.8856    0.0001888
    pl_bmasse                                   -0.026747    0.031683    -0.84421      0.40067
    pl_orbeccen                                   -67.234      213.51    -0.31491      0.75352
    pl_insol                                    0.035805   0.0058585      6.1117   2.1653e-08
    st_teff                                      0.14914    0.066207      2.2527     0.026579
    st_rad                                        218.33      105.32       2.073     0.040879
    st_mass                                       170.98      295.03     0.57952     0.56361


Number of observations: 119, Error degrees of freedom: 105
Root Mean Squared Error: 200
R-squared: 0.899,  Adjusted R-Squared: 0.887
F-statistic vs. constant model: 72, p-value = 3.85e-46
```

We then proceed to perform the multiple linear fit and analyze the results:
The first model that includes the totality of predictors has apparent problems. All categorical predictors regarding the discovery method do not have enough entry points to give significant results (hence the NaN). We note, however, that the low value of the F statistic assures us of the good overall significance of the model. We then remove the categorical predictor "**disc_year**" and perform the fit again. The second model is improved, but we note that the p-values for many predictors are very high.

The p-value provided by MATLAB are, in fact, related to the following test:

$$H_0 : \beta_k = 0$$

$$H_1 : \beta_k \neq 0$$

```
>> mdl = fitlm(P,'pl_eqt~sy_snum+sy_pnum+disc_year+pl_orbper+pl_orbsmax+pl_rade+pl_bmasse+pl_orbeccen+pl_insol+st_teff+st_mass+st_rad')

mdl =


Linear regression model:
    pl_eqt ~ 1 + sy_snum + sy_pnum + disc_year + pl_orbper + pl_orbsmax + pl_rade + pl_bmasse + pl_orbeccen + pl_insol + st_teff + st_rad + st_mass

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    _______    ________    _________

    (Intercept)        16530       19853     0.83262      0.40693
    sy_snum          -26.498      44.427    -0.59643      0.55216
    sy_pnum          -14.186       16.71    -0.84899       0.3978
    disc_year        -8.2121      9.8157    -0.83663      0.40469
    pl_orbper          2.124     0.32805      6.4745    3.0367e-09
    pl_orbsmax       -1983.4      235.22     -8.4322    1.8449e-13
    pl_rade            20.05      4.9631      4.0398    0.00010149
    pl_bmasse      -0.026192    0.031592    -0.82905      0.40894
    pl_orbeccen      -72.055      212.85    -0.33853      0.73563
    pl_insol        0.036141    0.005823      6.2066    1.0704e-08
    st_teff          0.14529    0.065797      2.2081      0.029391
    st_rad            223.84      104.74      2.1371      0.034891
    st_mass           165.08      294.16     0.56118      0.57586


Number of observations: 119, Error degrees of freedom: 106
Root Mean Squared Error: 200
R-squared: 0.899,  Adjusted R-Squared: 0.887
F-statistic vs. constant model: 78.4, p-value = 5.33e-47
```

We then proceed iteratively to remove the predictor with the highest p-value (> 10%), the result is a series of predictors with very low p-values (< 10^{-5}).

Remarkably, with this process we also removed the predictor referred to the number of stars, confirming what in the hypothesis test seemed likely.

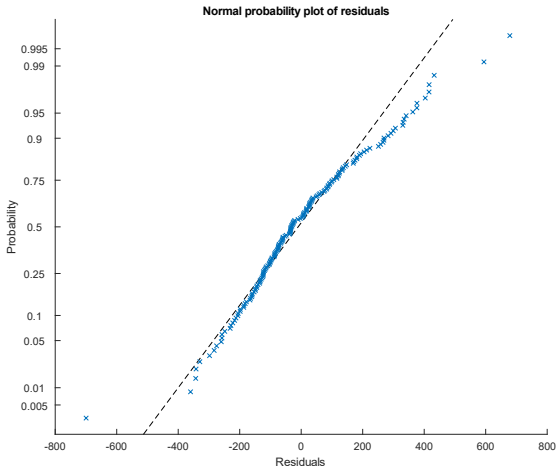
The only high p-value is that of the intercept, in fact we see that a test at the significance level $\alpha = 0.05$ would lead us to reject $H_0 : \beta_0 = 0$ for very little, not being satisfied with this we leave open the possibility that the intercept is nothing.

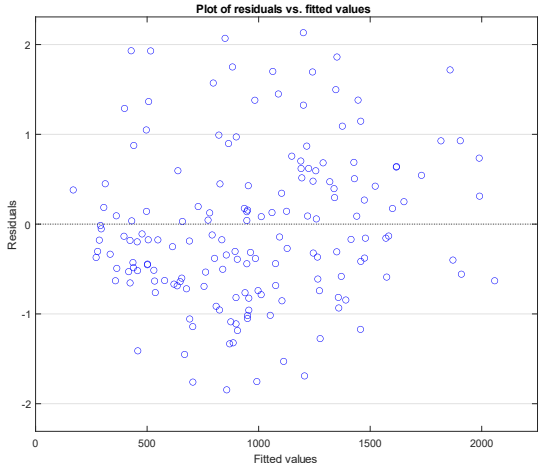
```
>> mdl = fitlm(P,'pl_eqt~pl_orbper+pl_orbsmax+pl_rade+pl_insol+st_teff+st_rad')

mdl =


Linear regression model:
    pl_eqt ~ 1 + pl_orbper + pl_orbsmax + pl_rade + pl_insol + st_teff + st_rad

Estimated Coefficients:
                    Estimate        SE          tStat       pValue
                    ________    _________    ________    __________

    (Intercept)      -179.71       89.943     -1.998       0.047319
    pl_orbper           1.97      0.22762      8.655      3.749e-15
    pl_orbsmax       -1856.1       164.79    -11.264      2.6035e-22
    pl_rade            22.601       3.3732     6.7003     2.9649e-10
    pl_insol         0.037687    0.0048757     7.7295     9.1881e-13
    st_teff           0.17831     0.025765     6.9207     8.9092e-11
    st_rad             239.04       55.477     4.3088     2.7794e-05


Number of observations: 176, Error degrees of freedom: 169
Root Mean Squared Error: 195
R-squared: 0.885,  Adjusted R-Squared: 0.881
F-statistic vs. constant model: 216, p-value = 1.37e-76
```

The coefficient of determination has not changed much from the first model and at the moment it stands at a good value of $\boldsymbol{R^2 = 0.885}$ the number of predictors is very low compared to the number of observations, so we are sure not to have exceeded into overfitting, this is confirmed by the always good value of

$$\boldsymbol{R^2_{adj}} = 0.881$$

Let's analyze the distribution of residues:

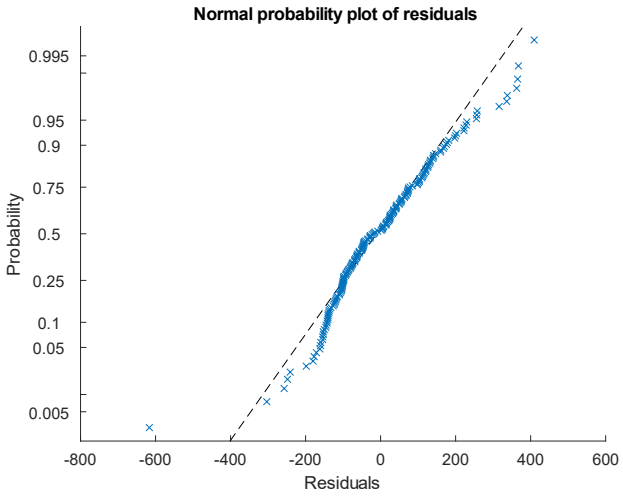


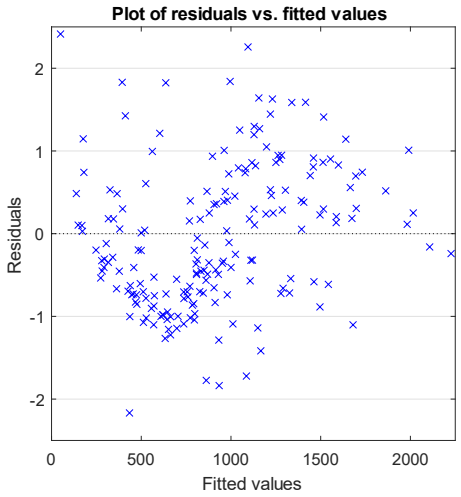


Although the graph of the residuals on fitted values shows a homoskedasticity at the limit of the acceptable, the QQ-plot of the residuals forces us to review the model.

We therefore propose other models with squared predictors, most of which contribute to an increase in the coefficient of determination but none of them solve the problem of heteroskedasticity of the data

```
>> mdl = fitlm(P,'pl_eqt~pl_orbper+pl_orbsmax+pl_insol+st_rad+pl_orbper^2+pl_insol^2+st_rad^2')

mdl =


Linear regression model:
    pl_eqt ~ 1 + pl_orbper + pl_orbsmax + pl_insol + st_rad + pl_orbper^2 + pl_insol^2 + st_rad^2

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                    ________    ________    ________    ____________

    (Intercept)       198.13      34.631      5.7212    4.0322e-08
    pl_orbper         6.4081     0.47601      13.462    1.7413e-29
    pl_orbsmax       -3536.4      215.65     -16.399    2.6756e-38
    pl_insol         0.25728    0.016554      15.543    9.5684e-36
    st_rad              1344      70.709      19.008    6.4901e-46
    pl_orbper^2     -0.0006433  6.5937e-05    -9.7563    1.6758e-18
    pl_insol^2      -4.3767e-06  3.589e-07    -12.195    1.1333e-25
    st_rad^2           -351.62       30.94    -11.365    3.3945e-23


Number of observations: 199, Error degrees of freedom: 191
Root Mean Squared Error: 140
R-squared: 0.939,  Adjusted R-Squared: 0.937
F-statistic vs. constant model: 422, p-value = 1.68e-112
```





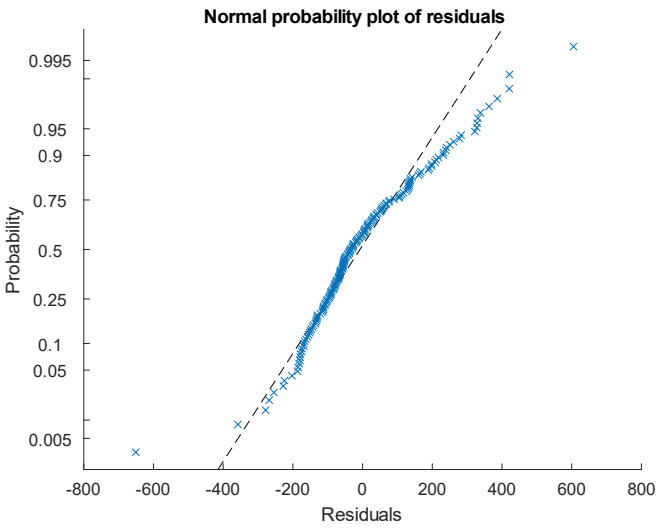
This model was obtained by including all predictors and their squares and proceeding as before with the removal of the least significant predictors in order of decreasing p-values.

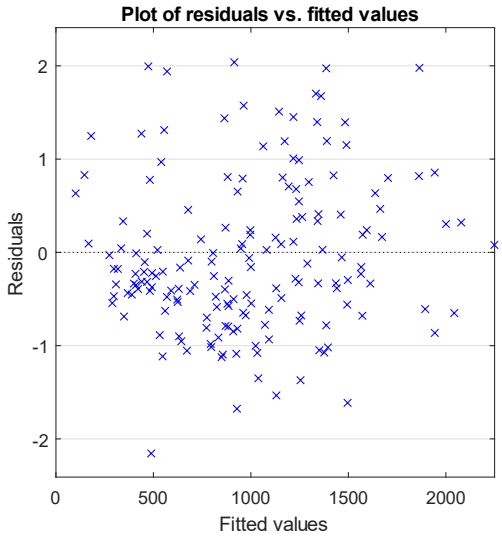
The value of R^2 of this model has increased, settling at $\boldsymbol{R^2 = 0.939}$ nevertheless the adherence of the data to the quantiles in the QQ-plot has worsened and also the residual plot shows a strongly heteroskedastic trend, with a hint of parable.

Again, we note that the model's coefficient of determination is very good, but the residuals are distinctly heteroskedastic.

In an attempt to find a model that respected the assumption of homoskedasticity of the residues, we tried several models: models with simple interactions, square or cubic interactions, logarithms, and exponentials, but none gave the desired results. Although homoskedastic, we consider the first model the best we could find.

We emphasize that the model is far from perfect, however we do not have all the parameters that could affect the temperature, such as the composition of the planet or the composition of the star, so we believe the model is quite satisfactory for our purposes.





```
>> mdl = fitlm(P,'pl_eqt~pl_orbper+pl_orbsmax+pl_rade+pl_insol+st_teff+st_rad+pl_orbper^2')

mdl =


Linear regression model:
    pl_eqt ~ 1 + pl_orbper + pl_orbsmax + pl_rade + pl_insol + st_teff + st_rad + pl_orbper^2

Estimated Coefficients:
                    Estimate          SE         tStat       pValue
                   ___________    _________    _______    __________

    (Intercept)        -155.99       76.361    -2.0428       0.04264
    pl_orbper           7.0109      0.64627     10.848    4.0681e-21
    pl_orbsmax         -3877.6        284.1    -13.649     5.084e-29
    pl_rade             15.289       2.9984     5.0991    9.1274e-07
    pl_insol          0.034031    0.0041606     8.1794    6.7066e-14
    st_teff            0.18616      0.02188     8.5083    9.3824e-15
    st_rad              333.65       48.468     6.8838    1.1075e-10
    pl_orbper^2     -0.00073712    9.0186e-05    -8.1733    6.9553e-14


Number of observations: 176, Error degrees of freedom: 168
Root Mean Squared Error: 166
R-squared: 0.918,  Adjusted R-Squared: 0.914
F-statistic vs. constant model: 267, p-value = 1.41e-87
```

```
>> [ypred,yci] = predict(mdl,Pnew)

ypred =

   1.0e+04 *

    0.0553
    0.0211
   -0.0024
   -0.0373
    0.0215
    0.4667
    2.5801
    6.3212


yci =

   1.0e+04 *

    0.0471    0.0635
    0.0092    0.0330
   -0.0171    0.0122
   -0.0557   -0.0188
   -0.0306    0.0737
    0.2694    0.6639
    1.7945    3.3658
    4.5644    8.0781
```

To test the model, we try to make predictions. In particular, we tried to predict the temperature of the planets of the solar system based on their parameters using the first model.

MATLAB returns both point estimates for temperatures and 95% confidence intervals. The outputs are ordered in increasing order by the distance from the Sun. The results are nothing short of terrible. The only planet whose temperature falls within the confidence intervals is Mercury, whose temperature is:

$$471K < 600K < 635K$$

for all the other planets the confidence intervals are completely wrong, even giving negative results (note that temperatures are given in Kelvin).

We can hypothesize that the low temperatures of the solar system do not agree with the pool of very hot planets that we used to create the regression model and that therefore our model is not able to accurately predict such atypical members with respect to our dataset.

Conclusions

The dataset chosen, albeit very large, allows a limited number of strong deductions. The large bias presented at the beginning leaves no room for comparisons with the only planets we know for direct measurements (our solar system).

The habitability of exoplanets is much more complex than originally thought and less prone to statistical analysis. Nevertheless, these results are essential for all astronomical and astrophysical studies.

Without an analysis of this kind, the intrinsic bias in discovering exoplanets is not evident. The scientist in front of this dataset could draw erroneous conclusions about the actual population of exoplanets.

In conclusion, we are proud to have read, understood, and verified all the major publications regarding the properties of exoplanets from the raw data. Peer review is a critical activity for the integrity of human knowledge and ultimately leads to better science for everyone.

Additional Resources

Datasets:

- NASA Exoplanet Archive
- NASA Planetary Fact Sheet
- PHL's Habitable Exoplanets Catalogue
- PHL's Exoplanets Catalog
- PHL's Data of Potentially Habitable Worlds

Software:

- Matlab
- R Studio

Acknowledgements:

- Terrestrial Planet Formation in Exoplanetary Systems (2008) **M. J. Fogg**
- Statistical properties of exoplanets-Planets and Astrobiology (2018-2019) **G. Vladilo**
- Surface Temperatures of Exoplanets (2015) **Weisfeiler, Turcotte, Kellogg**
- The Masses and Orbital Dynamics of Exoplanets (2016) **Weiss, Marcy**
- HD10647 and the Distribution of Exoplanet Properties (2004) **J.P. Beaulieu, et Al.**
- Finding the Mass of an Exoplanet (2020) **J. Chatelain, J. Jones, R. Sketch**
- The sub-Jovian desert of exoplanets (2019) **Gyula M. Szabó, Szilárd Kálmán**
- Beyond the exoplanet mass-radius relation (2019) **Ulmer-Moll et Al.**
- Exoplanet orbital eccentricity (2015) **Mary A. Limbach, Edwin L. Turner**
- Studio delle condizioni fisiche favorevoli alla fotosintesi (2020) **R.M. Ienco**
- Missione Spaziale Kepler (2021) **F. Abbate, M. Montanari, G. Rigolizzo**
- Properties of Extrasolar Planets – **University of Hawaii**
- Direct Imaging - **Las Cumbres Observatory**
- Calculating Exoplanet Properties - **Simon Fraser University**