



**POLITECNICO**  
MILANO 1863

Filippo Sergenti

Gabriele Rolleri

Giovanni Zhang



*“Due cose riempiono l’animo di ammirazione e venerazione sempre nuova e crescente, quanto più spesso e più a lungo la riflessione si occupa di esse: il cielo stellato sopra di me, e la legge morale dentro di me”*

*Immanuel Kant - 1788 - Critica della Ragion Pratica*

## Introduzione

Viviamo in un'era d'oro di scoperta e caratterizzazione di esopianeti.

Dal suo lancio nel 2009, la missione Kepler ha scoperto oltre 4000 pianeti extrasolari, consentendo studi dettagliati delle loro proprietà fisiche e orbitali. In alcuni casi, Kepler è riuscito a determinarne persino le composizioni, fornendo così importanti indizi sulla formazione e l'evoluzione dei sistemi planetari.

## Astrobiologia & Cosmologia

Chi siamo? Da dove veniamo? Queste sono domande che accompagnano l'uomo dai suoi primi passi su questa terra.

La scelta di questa tesina è nata non solo dal fascino di queste due domande ma anche dalla possibilità di dagli risposta in modo rigoroso e scientifico.

Ovviamente sono domande aperte e pensiamo che lo rimarranno per qualche altro milione di anni, seppur un po' scoraggiati, cerchiamo di aggregare tutto il materiale disponibile sull'argomento studiandolo attraverso gli strumenti statistici appresi.

## Che cosa è un esopianeta?

Un pianeta extrasolare o esopianeta è un pianeta non appartenente al sistema solare, orbitante cioè attorno a una stella diversa dal Sole.

Pienamente confermata solo nel 1995, l'esistenza di pianeti extrasolari fu per lungo tempo ritenuta più che plausibile. La prima ipotesi dell'esistenza di questi corpi celesti fu formulata da Isaac Newton nel 1713.

Al 29 giugno 2021 risultano conosciuti 4700 esopianeti in 3472 sistemi planetari diversi; 2487 è il numero di pianeti candidati mentre 209 sono ancora in attesa di conferma.

La scoperta degli esopianeti è resa possibile da metodi di osservazione indiretta. A causa dei limiti delle tecniche di osservazione attuali, la maggior parte dei pianeti individuati sono giganti gassosi come Giove e, solo in misura minore, pianeti rocciosi massicci del tipo Super Terra (da 2 a 10 masse terrestri).

Le migliaia di esopianeti disponibili per la caratterizzazione consentono studi statistici robusti della popolazione totale dei pianeti. Le dimensioni, le masse, le composizioni e le dinamiche orbitali di questi pianeti ci danno indizi sulla loro formazione

## Presentazione del Campione

Nella seguente tabella elenchiamo tutte le variabili prese in considerazione per il nostro studio. Nelle pagine successive tratteremo in modo estensivo ogni parametro, ne studieremo la distribuzione e la sua importanza per determinare l'abitabilità di un pianeta.

I dati sono accessibili liberamente al seguente indirizzo: [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu)

<b>Planet Name</b> (pl_name)	Nome del pianeta più usato in letteratura
<b>Number of Stars</b> (sy_snum)	Numero di stelle nel sistema planetario
<b>Number of Planets</b> (sy_pnum)	Numero di pianeti nel sistema planetario
<b>Discovery Year</b> (disc_year)	Anno in cui è stato scoperto il pianeta
<b>Discovery Method</b> (discoverymethod)	Metodo con cui il pianeta è stato scoperto per la prima volta
<b>Orbital Period</b> (pl_orbper)	Tempo impiegato dal pianeta per compiere un'orbita completa attorno alla stella o al sistema ospite
<b>Orbit Semi-Major Axis</b> (pl_orbsmax)	Il raggio più lungo di un'orbita ellittica
<b>Planet Radius</b> (pl_rade)	Segmento che congiunge il centro del pianeta alla sua superficie. Misurata in unità di raggio della Terra
<b>Planet Mass</b> (pl_bmasse)	Migliore stima della massa del pianeta disponibile
<b>Eccentricity</b> (pl_orbeccen)	Quantità di cui l'orbita del pianeta devia da un cerchio perfetto
<b>Insolation Flux</b> (pl_insol)	Il flusso di insolazione è una misura della radiazione stellare incidente. È espresso in unità relative a quelle misurate per la Terra.
<b>Equilibrium Temperature</b> (pl_eqt)	La temperatura di equilibrio del pianeta come modellata da un corpo nero riscaldato solo dalla sua stella ospite.
<b>Stellar Effective Temperature</b> (st_teff)	Temperatura della stella modellata da un corpo nero che emette la stessa quantità totale di radiazioni elettromagnetiche
<b>Stellar Radius</b> (st_rad)	Segmento che congiunge il centro della stella alla sua superficie.
<b>Stellar Mass</b> (st_mass)	Quantità di materia contenuta nella stella, misurata in unità di massa del Sole
<b>Right ascension</b> (ra)	Ascensione retta del sistema planetario, misurata in gradi decimali
<b>Declination</b> (dec)	Declinazione del sistema planetario, misurata in gradi decimali

## Parte I: Statistica Descrittiva

Si vuole iniziare con un'indagine esplorativa del dataset a disposizione per dedurne i caratteri generali e andamenti particolari.

### Presenza di Bias

La maggior parte delle scoperte riguardano per lo più giganti gassosi che orbitano attorno alle loro stelle a breve distanza. Questo tipo di pianeti, chiamati **gioviani caldi**, influiscono notevolmente sulla velocità radiale delle loro stelle e transitano di frequente davanti a esse, facilitando la loro individuazione: per **effetto di selezione** vi è una chiara supremazia quantitativa di tali pianeti rispetto agli altri. Con il migliorare degli strumenti si sta invertendo la tendenza; è quindi evidente che la prevalenza di corpi tellurici simili alla Terra è superiore a quella dei pianeti giganti.

La frazione di pianeti più piccoli è in costante crescita, soprattutto grazie alla missione Kepler, che consente già di definire un abbozzo di classificazione esoplanetaria in base alle dimensioni.

Gli astronomi si domandano perché molti pianeti extrasolari giganti gassosi di grandi dimensioni si trovino molto vicini alla loro stella, rispetto a quelli del nostro sistema solare. Per esempio, T Bootis ha un pianeta quattro volte più grande di Giove a meno di un quarto della distanza Terra-Sole. HD 114762 ha un pianeta undici volte più grande di Giove, a meno di mezza UA. Una possibile risposta è che i metodi di ricerca odierni favoriscano l'individuazione di questo tipo di sistemi: un grande pianeta posto a piccola distanza amplifica le oscillazioni della stella, ed esse sono facilmente individuabili tramite l'effetto Doppler. Un pianeta più piccolo, a distanza più grande, provoca oscillazioni molto più piccole e difficili da vedere.

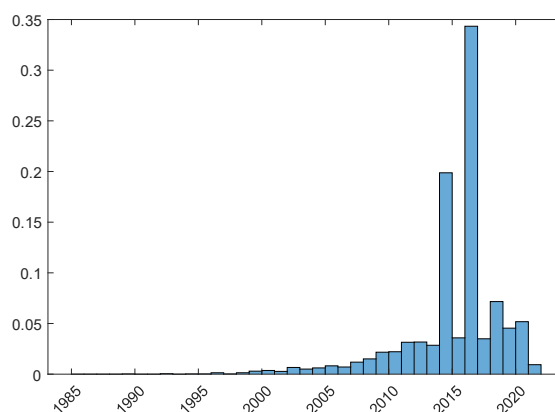
Un'altra spiegazione è che i pianeti si siano formati a distanze maggiori, per poi muoversi verso l'interno a causa delle reciproche interazioni gravitazionali. Tale modello è stato chiamato **modello dei gioviani saltellanti**, nome che rende bene l'idea.



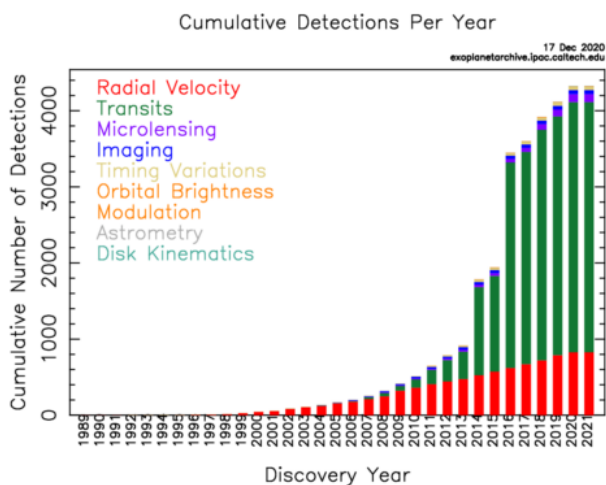
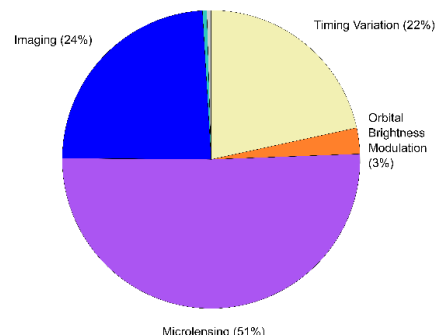
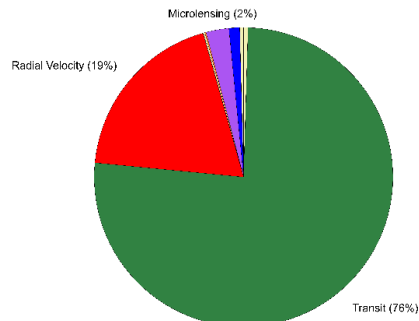
# 1) Anno di Scoperta

L'interesse scientifico sugli esopianeti è cresciuto sempre più a partire dal 1992, anno della prima scoperta confermata (PSR B1257+12).

Inizialmente il ritmo delle scoperte è stato molto lento, ma a partire dagli anni 2000 ha conosciuto una vera e propria impennata, passando dai 20 pianeti scoperti al 2000, ai 189 al 2011, ai quasi 3500 al 2016.

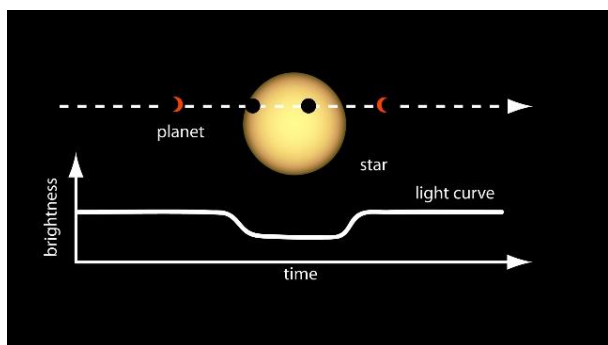


# 2) Metodo di Scoperta



Il Metodo di scoperta è una variabile categorica che descrive il processo con cui è stato scoperto il pianeta. Sono prevalenti 2 metodi: Transit e Radial Velocity che assieme contano circa il 95% del totale.

Transit è stata una vera e propria rivoluzione per il campo dell'esoplanetologia aumentando considerevolmente il numero di pianeti scoperti per singolo anno.



## Transit

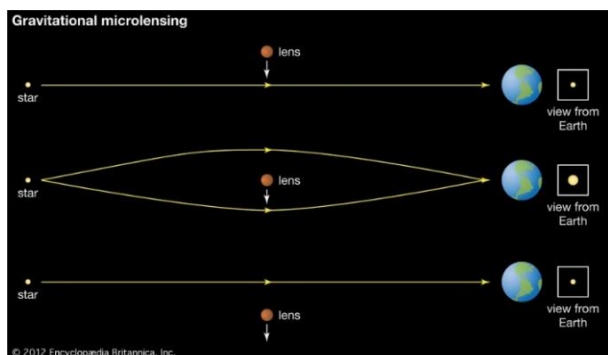
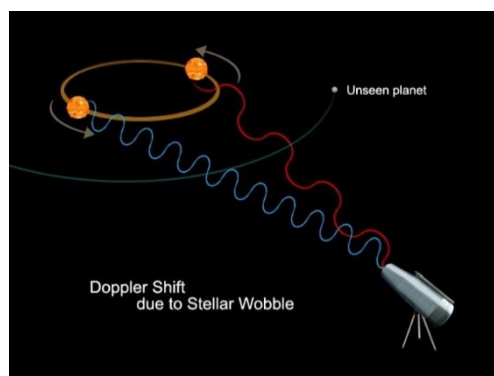
Metodo più recente e promettente.

Quando un pianeta transita di fronte ad una stella se ne registra una diminuzione nella sua luminosità.

## Radial Velocity

Le variazioni nella velocità con cui la stella si avvicina o si allontana dalla Terra possono far dedurre la presenza di un pianeta a causa di sbilanciamenti della linea spettrale della stella (Effetto Doppler).

Questo metodo è in grado di individuare facilmente pianeti molto vicini alla loro stella, non funziona per pianeti con lunghi periodi orbitali.



## Microlensing

I campi gravitazionali di due corpi celesti possono cooperare per focalizzare la luce di una stella lontana.

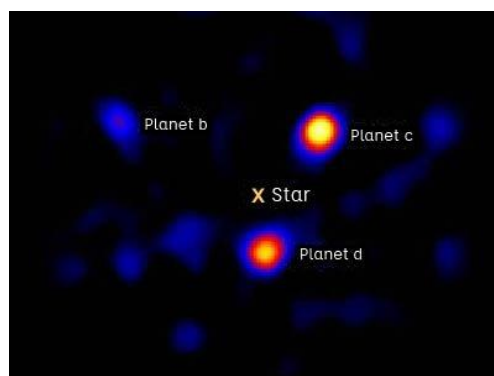
Un allineamento perfetto capita molto di rado e l'effetto è molto piccolo.

Questo metodo permette di scoprire pianeti di massa equiparabile a quella terrestre usando le tecnologie oggi disponibili

## Direct Imaging

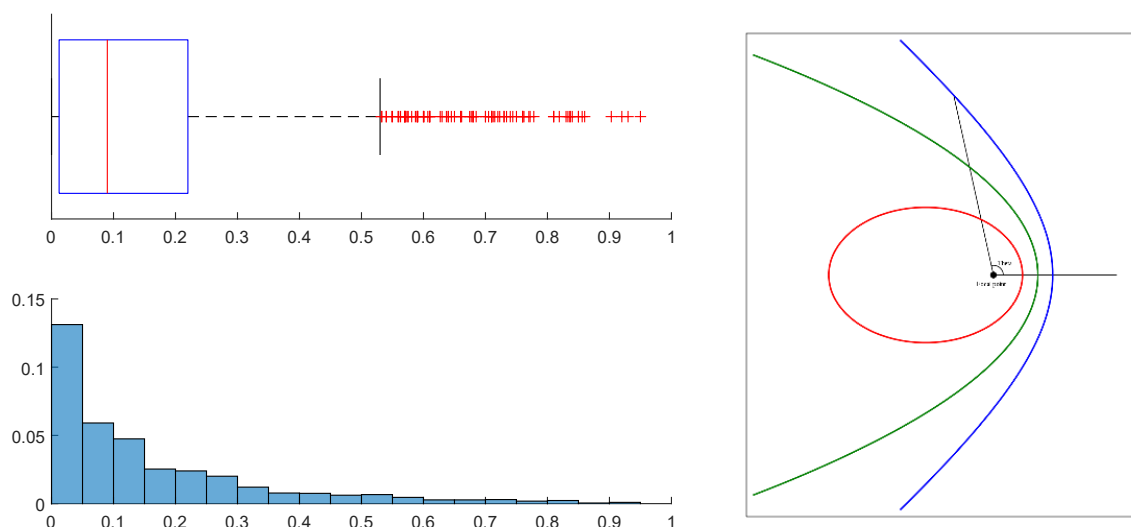
L'imaging diretto utilizza lunghezze d'onda infrarosse per osservare i pianeti.

Questo metodo funziona per i pianeti che sono molto lontani dalle loro stelle, ma non consente agli astronomi di misurare direttamente la massa di un pianeta. Si possono utilizzare lo spettro e la luminosità per ottenere informazioni sulla temperatura e sul diametro.





### 3) Eccentricità



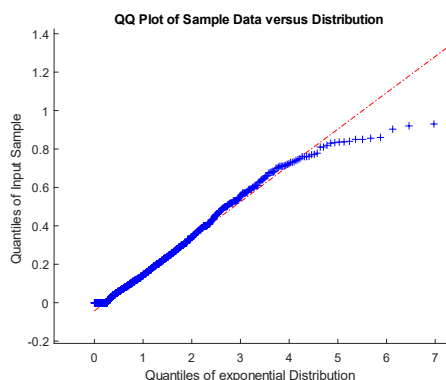
Parametro adimensionale che indica quanto l'orbita devia da un cerchio perfetto.

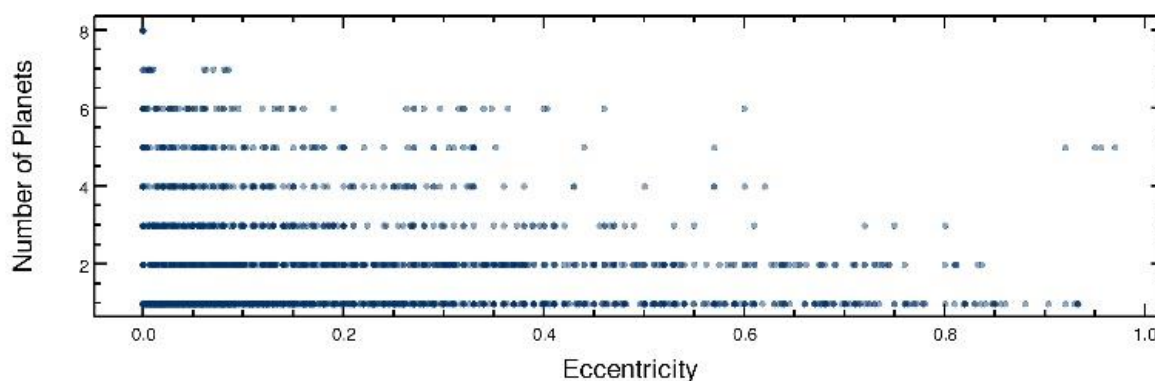
Un valore di 0 indica un'orbita circolare, i valori tra 0 e 1 formano un'orbita ellittica, 1 è un'orbita di fuga parabolica, se maggiore di 1 descrive un'iperbole.

Un pianeta avente orbita stabile non può avere eccentricità maggiore o uguale a 1. L'orbita della Terra è pressoché circolare, essendo l'eccentricità inferiore a 0,02.

Possiamo notare come la distribuzione sia fortemente spostata a sinistra con un baffo destro molto lungo e una modesta quantità di outlier oltre tale baffo, la mediana ci informa che più di metà dei pianeti hanno eccentricità  $< 0.1$ , l'istogramma suggerisce che il campione possa seguire una distribuzione di tipo esponenziale, procediamo quindi a fare un qqplot rispetto a tale distribuzione.

Il centro del campione aderisce molto bene alla retta, mentre l'elevato numero di valori nulli crea un leggero scostamento nella coda sinistra, anche la coda destra si discosta dalla retta in maniera abbastanza importante. In definitiva la distribuzione non è perfettamente esponenziale ma ne ricalca le caratteristiche.





I dati raccolti sulla eccentricità dei pianeti extrasolari hanno sorpreso la maggior parte dei ricercatori: 90% delle eccentricità sono maggiori di quelli dei pianeti nel sistema solare.

Questi fatti sono stati ampiamente interpretati per indicare che il Sistema Solare è un membro atipico della popolazione complessiva dei sistemi planetari.

Inoltre, la stabilità orbitale e di rotazione è essenziale affinché il corpo celeste sia abitabile. Maggiore è l'eccentricità orbitale, più grande è la fluttuazione della temperatura sulla superficie del pianeta.

È fondamentale che il pianeta abbia stagioni moderate, ugualmente Il ciclo dì-notte non deve essere troppo lungo.

Nonostante si adattino, gli organismi viventi non sono in grado di sopportare eccessive variazioni, specialmente se esse raggiungono talvolta il punto di ebollizione e il punto di fusione del principale **solvente biotico** del pianeta (sulla Terra, l'acqua).

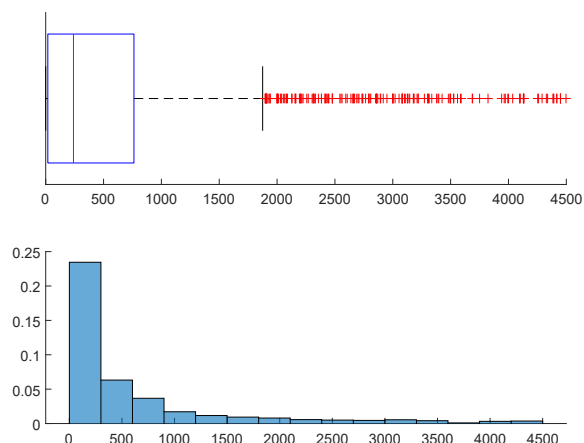
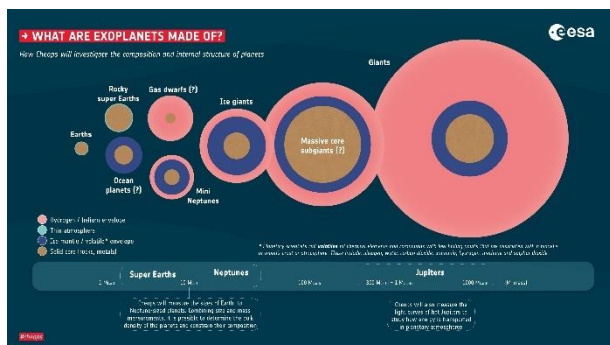
Riportiamo una forte anti-correlazione dell'eccentricità orbitale con la molteplicità (numero di pianeti nel sistema). Se basse eccentricità favoriscono effettivamente molteplicità elevate, l'abitabilità può essere più comune nei sistemi con un numero maggiore di pianeti.

Questi dati potrebbero presentare un bias, infatti una forte eccentricità aumenta l'oscillazione della stella e quindi facilita l'individuazione del pianeta.

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



## 4) Massa



La distribuzione delle masse dei pianeti è molto simile a quello dell'eccentricità ma con valori ancora più schiacciati verso lo zero e una serie di outlier destri ancora più numerosi. Gli stessi test effettuati in precedenza rivelano nuovamente che la distribuzione è lontana da quella esponenziale. Come per l'eccentricità notiamo che per un buon numero di pianeti il valore della massa è mancante (2520 su 4383).

I pianeti con una massa scarsa sarebbero dei cattivi candidati ad ospitare la vita per due ragioni. Innanzitutto, la loro gravità risulterebbe più bassa e la loro atmosfera meno densa. Le molecole che costituiscono la vita hanno una probabilità molto più elevata di raggiungere la velocità di fuga e di essere espulse nello spazio per la propulsione del vento solare o per una collisione.

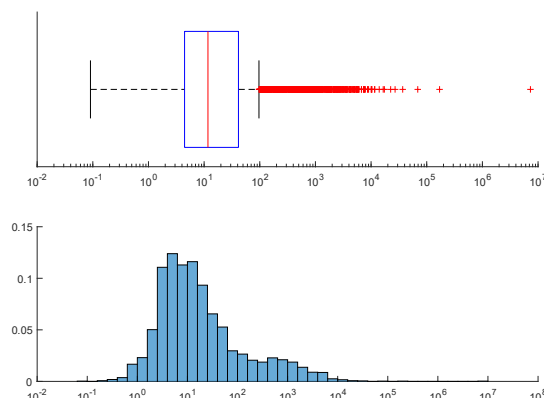
I pianeti la cui atmosfera non è spessa hanno meno protezione contro le radiazioni ad alta frequenza e meteoriti, non disporranno di sufficiente materia per la biochimica iniziale, non sono abbastanza isolati termicamente e hanno una bassa conducibilità termica attraverso la loro superficie.

Inoltre, i pianeti più piccoli hanno un diametro minore e dunque maggiore rapporto superficie-volume rispetto ai pianeti di maggiori dimensioni. Tali corpi tendono a perdere energia molto più rapidamente dopo la loro formazione ed hanno dunque scarsa attività geologica. Non presentano vulcani, terremoti né attività tettonica che forniscano alla superficie elementi favorevoli alla vita e all'atmosfera molecole in grado di regolare la temperatura (come il biossido di carbonio).

La nostra Terra è sufficientemente grande perché la sua forza gravitazionale trattienga la sua atmosfera e perché il suo nucleo liquido continui a restare attivo e caldo, generando così un'attività geologica sulla superficie (la disintegrazione di elementi radioattivi nel cuore del pianeta è un'altra risorsa di calore). Marte, al contrario, è pressoché inattivo ed ha perso la maggior parte della sua atmosfera. Pertanto, è ipotizzabile che la massa minima di un pianeta che possa risultare abitabile si situi tra quella di Marte e quella della Terra.

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

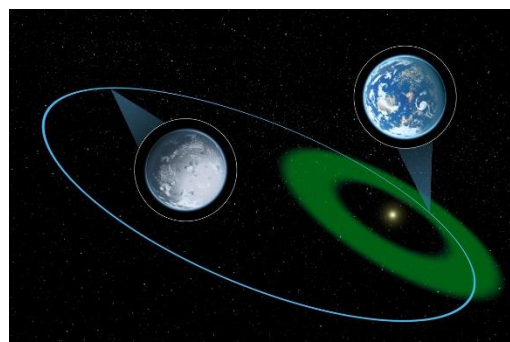
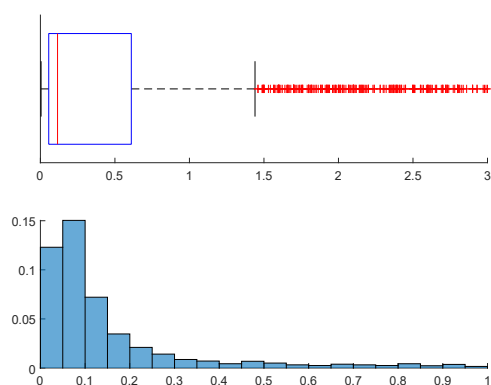
## 5) Periodo Orbitale



Dato l'enorme range dei dati, che variano da meno di un giorno a più di 7 milioni di giorni, è stato necessario introdurre una scala logaritmica, questo ovviamente rende inutilizzabili le informazioni riguardo alla probabilità riferita all'area delle barre. In ogni caso la distribuzione ha due massimi di cui uno nettamente più alto dell'altro, il secondo massimo comprende la maggior parte degli outlier destri. Questo alto numero di outlier e degli outlier estremi fanno sì che la media sia  $\approx 200$  volte maggiore della mediana.

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

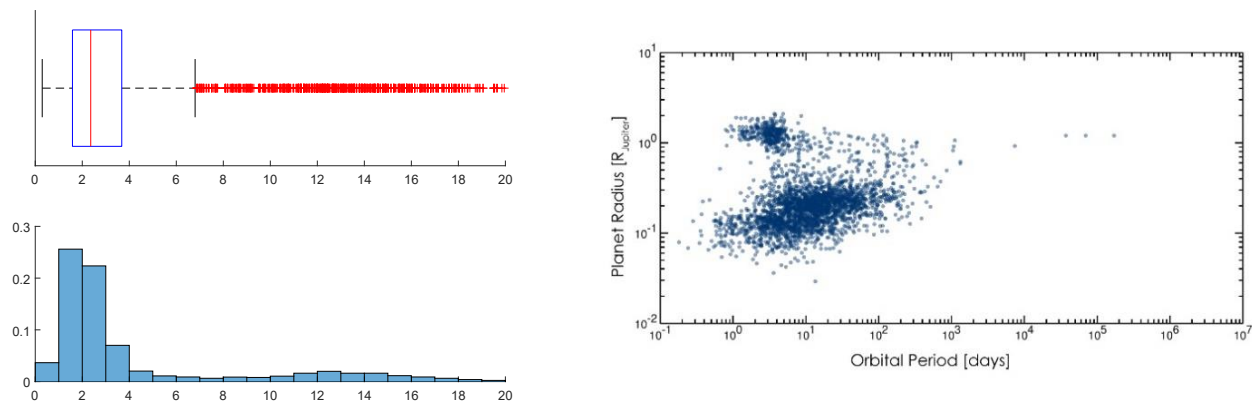
## 6) Semiasse Maggiore



Come ci suggeriscono le posizioni del primo e secondo quartile la distribuzione del semiasse maggiore torna ad essere molto densa verso lo zero con una coda molto lunga e outlier estremi migliaia di volte più grandi della media

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

## 7) Raggio



La Fulton gap, photoevaporation valley, o “**Deserto sub Gioviano**” è una scarsità osservata di pianeti con raggi tra 1,5 e 2 volte il raggio della Terra, probabilmente a causa della perdita di massa causata dalla fotoevaporazione.

Una bimodalità nella popolazione degli esopianeti Kepler è stata osservata per la prima volta nel 2013, ed è stata notata come possibile conferma di un'ipotesi emergente secondo cui la fotoevaporazione potrebbe guidare la perdita di massa atmosferica su pianeti vicini e di piccola massa.

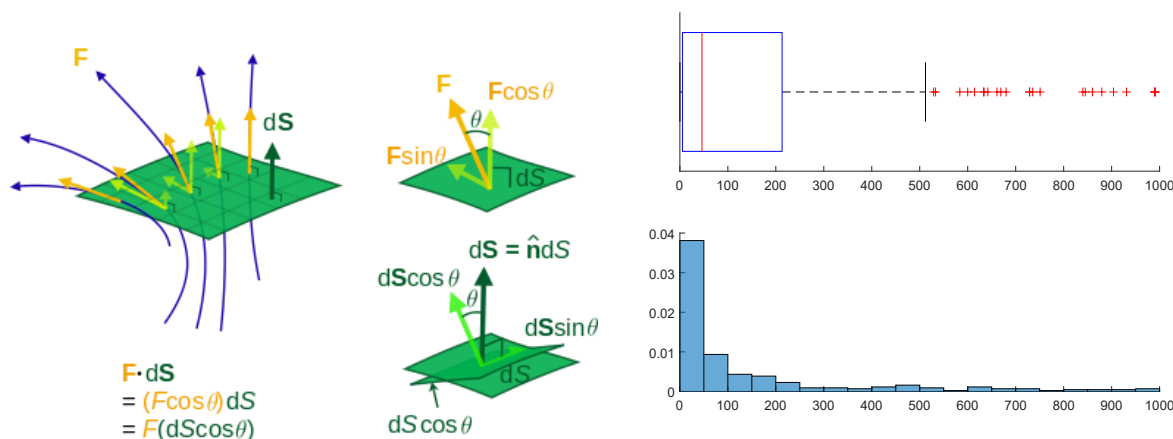
Ciò porterebbe a una popolazione di nuclei rocciosi nudi con raggi più piccoli a piccole separazioni dalle loro stelle madri e pianeti con involucri spessi dominati da idrogeno ed elio con raggi più grandi a separazioni maggiori.

Il termine **fotoevaporazione** designa il processo mediante il quale gli atomi o le molecole di un gas vengono strappati via da un accumulo ( un'atmosfera planetaria, un disco circumstellare o una nebulosa) da parte dei fotoni ad alta energia emessi da una stella.

Nonostante l'implicazione della parola "gap", questa non rappresenta in realtà una gamma di raggi completamente assente dalla popolazione di esopianeti osservata, ma piuttosto una gamma di raggi che sembra essere relativamente rara. Di conseguenza, 'valle' è spesso usato al posto di '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) Flusso Radiante



È una misura della quantità totale di radiazione elettromagnetica incidente sulla superficie del pianeta, viene calcolata sotto ipotesi di corpo nero.

Anche se nei dati riportati la temperatura è costante, è importante far notare che tutte le stelle conoscono delle variazioni di luminosità, l'ampiezza di tali fluttuazioni è molto diversa da una stella all'altra.

La maggior parte delle stelle sono relativamente stabili, ma una significativa minoranza tra esse è variabile e presenta spesso sia cali che aumenti improvvisi di luminosità, di conseguenza, la quantità di energia irradiata che i corpi orbitanti ricevono subisce delle brusche variazioni. Queste stelle sono dunque delle cattive candidate ad ospitare pianeti in grado di permettere la vita in quanto le forti variazioni di flusso energetico hanno un impatto negativo sulla sopravvivenza degli organismi.

Per esempio, esseri viventi adattati ad un dominio di temperatura particolare avrebbero probabilmente problemi a sopravvivere ad importanti variazioni di temperatura. Inoltre, le variazioni di luminosità sono generalmente accompagnate dall'emissione di dosi massicce di raggi gamma e di raggi X, radiazioni che potrebbero essere letali.

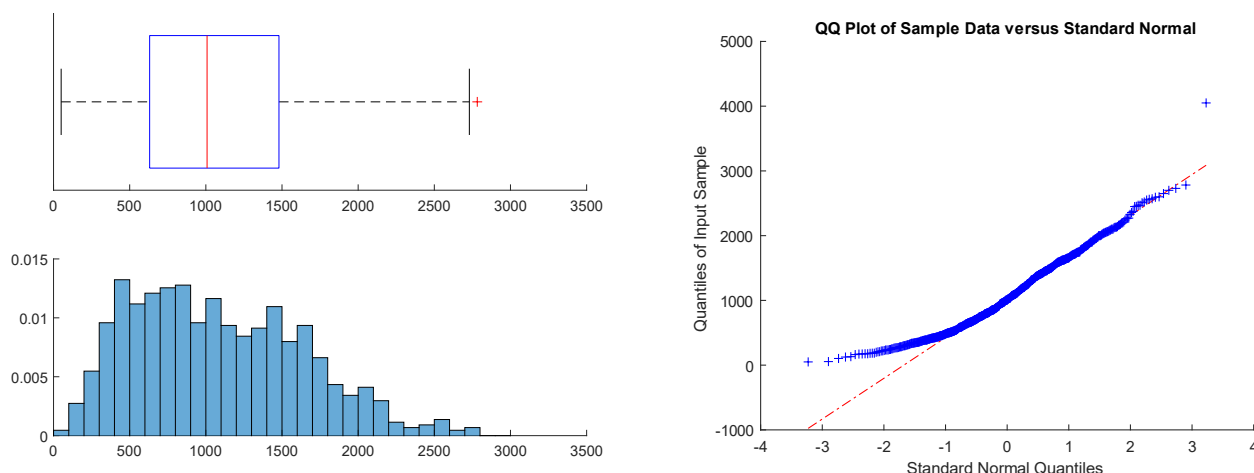
L'atmosfera dei pianeti è in grado di attenuare tali effetti (un aumento del 100 % della luminosità stellare non implica necessariamente un aumento del 100 % della temperatura del pianeta), ma è ugualmente possibile che tali pianeti non siano in grado di trattenere la loro atmosfera perché le forti radiazioni incidenti a ripetizione potrebbero disperderla.

Il Sole non conosce questo tipo di variazioni: nel corso del ciclo solare lo scarto tra la luminosità minima e massima si aggira attorno allo 0,1 %

La distribuzione è notevolmente asimmetrica, coda molto pronunciata a destra, conferma di ciò è anche la media molto più grande della mediana. In tutto il campione, è il parametro con meno riscontri fra i pianeti, solo 3 centinaia di valori sono presenti.

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

## 9) Temperatura d'equilibrio



La temperatura di equilibrio planetaria è la temperatura teorica che raggiungerebbe un pianeta se fosse un corpo nero, riscaldato soltanto dalla propria stella. In questo modello, non è considerata l'eventuale presenza di un'atmosfera (e quindi dell'effetto serra che potrebbe interessarla) e la temperatura calcolata è attribuita ad una superficie idealizzata del pianeta.

Non può essere calcolata direttamente, si deriva principalmente dagli altri parametri.

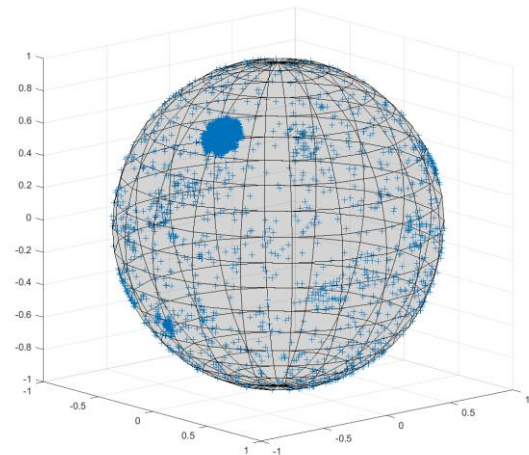
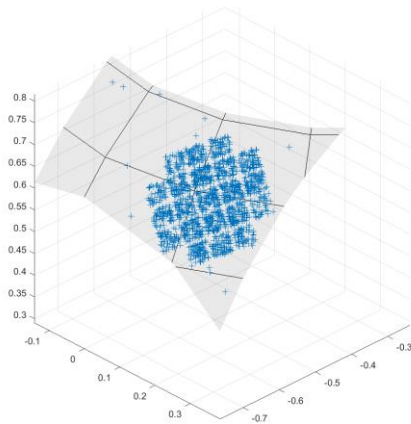
E' la distribuzione più similmente gaussiana del campione, la media, parametro più sensibile ai dati sugli estremi, è di poco maggiore della mediana, ciò giustifica l'accento di coda a destra.

Nella ricerca di esopianeti abitabili, la temperatura superficiale del pianeta gioca un ruolo cruciale. Sfortunatamente, al momento non sono disponibili misurazioni dirette. Molti processi fisici influenzano la distribuzione della temperatura superficiale di un pianeta. Tuttavia, l'influenza dominante è un bilancio energetico tra l'input di radiazione stellare e la perdita di calore superficiale radiativa. Con le ulteriori ipotesi di una temperatura superficiale planetaria uniforme, nessun filtraggio della radiazione in arrivo e l'emissione del corpo nero, le uniche variabili sono la luminosità stellare e la distanza radiale dell'esopianeta dalla stella.

Analisi di alcuni pianeti extrasolari inoltre hanno rivelato la presenza di venti molto veloci sulla superficie con punte di 14000 km/h. Questi venti mantengono la temperatura di questi pianeti costante su tutta la superficie con escursioni termiche molto ridotte.

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

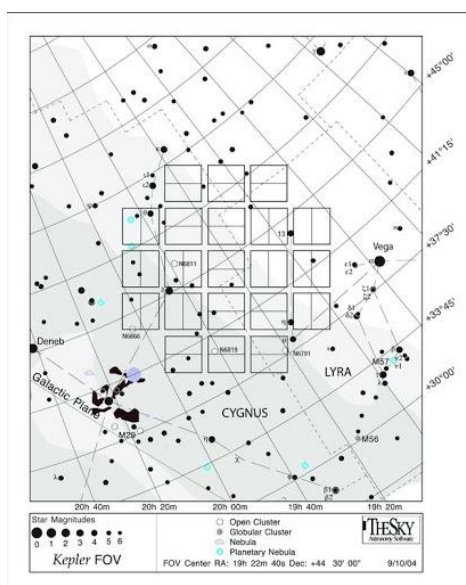
## 10) Distribuzione spaziale



La distribuzione degli esopianeti rispetto alla terra è abbastanza casuale. Tuttavia, nel dataset vi è un bias veramente marcato. Il pattern visibile nella prima figura non è casuale: rappresenta il fotometro del telescopio spaziale Kepler.

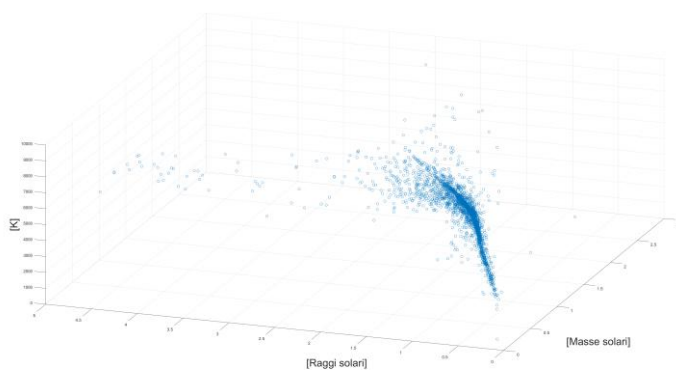
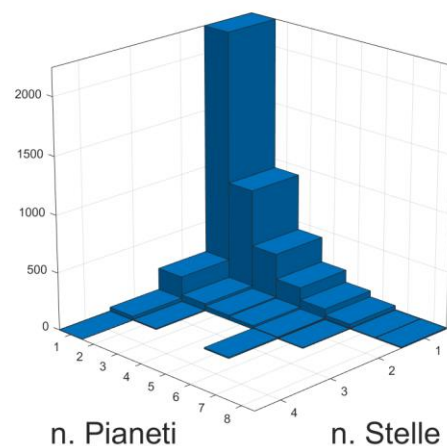
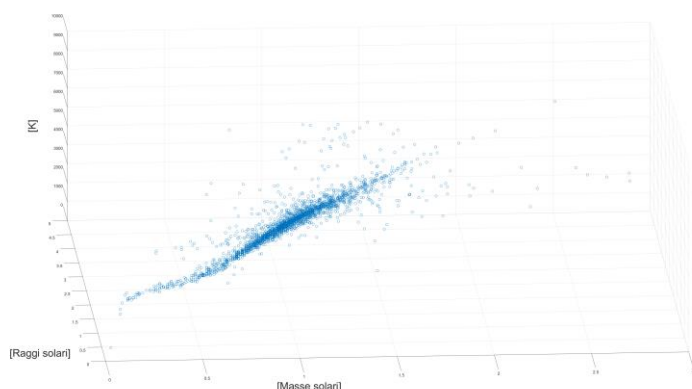
Il telescopio Kepler è stato specificatamente progettato per monitorare una porzione della regione della Via Lattea e scoprire quante delle miliardi di stelle della nostra galassia posseggano pianeti potenzialmente abitabili.

Per fare ciò, un fotometro ha monitorato costantemente la luminosità di più di 145 000 stelle della sequenza principale nel suo campo di vista fissato, presso le costellazioni del Cigno, della Lira e del Drago.





## 11) Parametri Stellari



Gli ultimi tre parametri riguardano la stella del sistema, la sua massa, il suo raggio e la sua temperatura.

Lo scatterplot di queste tre grandezze mostra una forte correlazione, abbiamo quindi provato a trovare un modello di regressione lineare multipla. Il modello risultava scadente anche rimuovendo gli outlier e includendo termini quadratici e cubici per i due predittori massa e raggio.

Analizzando le componenti dei sistemi planetari notiamo che, come ci si può aspettare, i sistemi ad una stella sono i più comuni.

Tuttavia è importante notare che la numerosità dei pianeti singoli è chiaramente dovuta a un forte bias legato alla natura della ricerca e della metodologia.

Non si può assumere che in natura vi siano prevalentemente sistemi con un singolo pianeta.

## Parte II: Test D'ipotesi

In seguito alla rassegna dei parametri fondamentali che caratterizzano gli esopianeti, si vuole sottoporre tali elementi a uno studio che ci permetta di dedurre un modello che possa predire la loro temperatura. A tale fine si vogliono supportare attraverso i test d'ipotesi due asserzioni:

- 1) La media delle temperature non dipende dal numero di stelle del sistema di cui fa parte
- 2) I metodi di scoperta hanno generato un forte bias nel pool dei dati

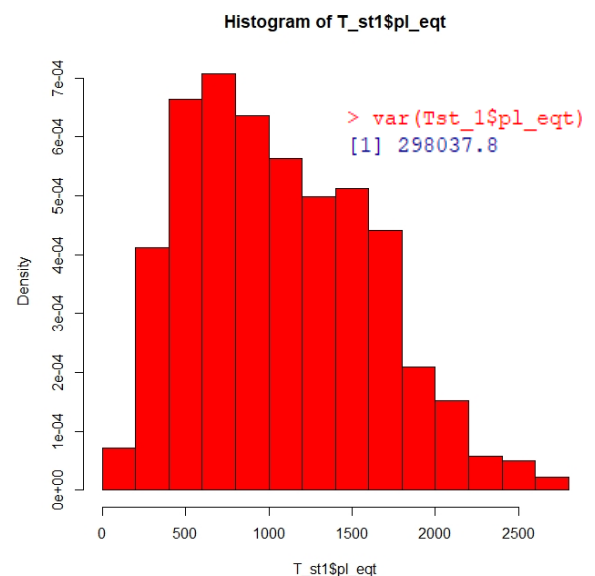
Si procede, dunque, suddividiamo in categorie minori le temperature per numero di stelle.

- Temperature dei pianeti con una **singola** stella:

**T\_st1** 692 dati

**Classi**  $1 + \log_2(692) = \lceil 10.43 \rceil = 11$

```
> 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
```

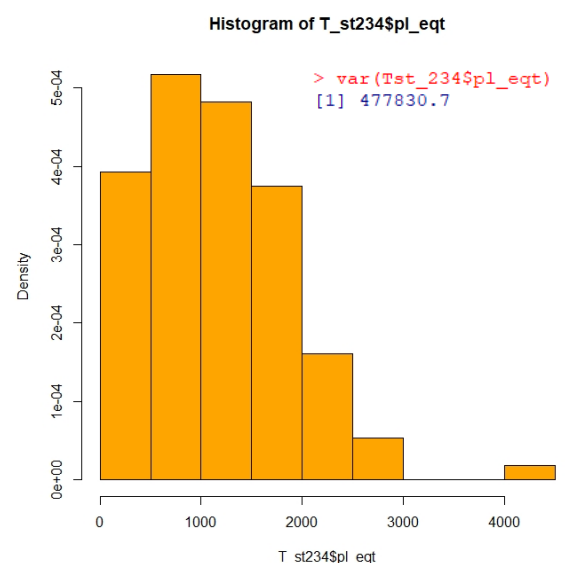


- Temperature dei pianeti con **più di due** stelle:

**T\_st234** 112 dati

**Classi**  $1 + \log_2(112) = \lceil 8.8 \rceil = 9$

```
> 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
```



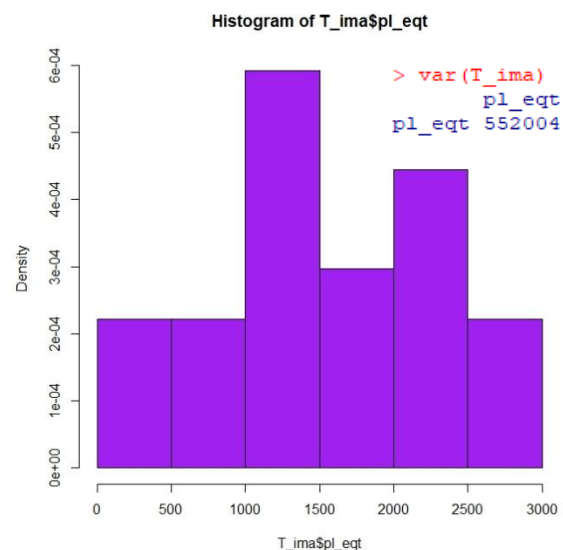
Suddividiamo in categorie minori le temperature anche per metodo di scoperta.

- Temperature dei pianeti rilevati tramite **imaging**

**T\_ima** 27 dati

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

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

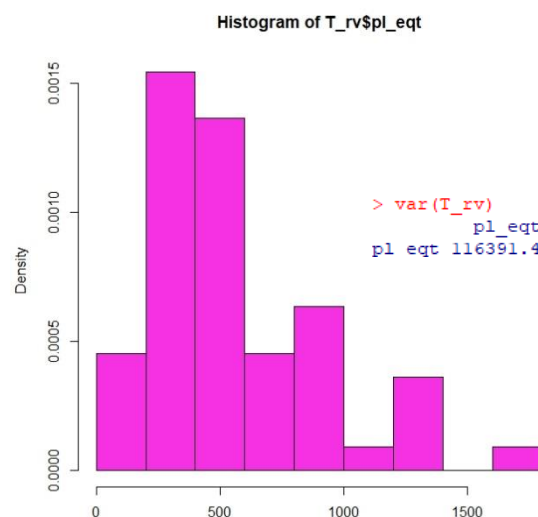


- Temperature dei pianeti rilevati tramite **radial velocity**

**T\_rv** 55 dati

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

```
> 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
```

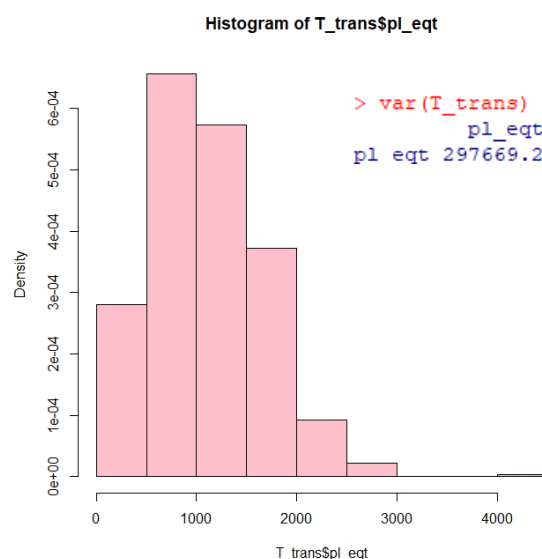


- Temperature dei pianeti rilevati tramite **transit**

**T\_trans** 715 dati

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

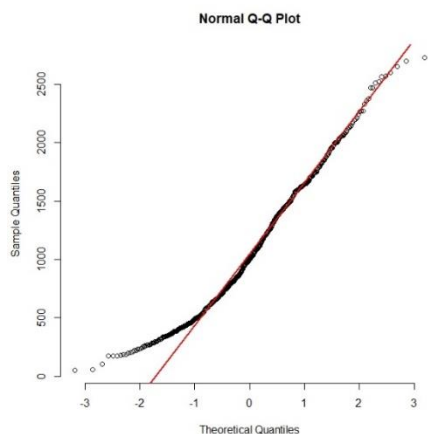
```
> summary(T_trans$pl_eqt)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   131    673   1053   1109   1498   4050
```



## Analisi normalità dei campioni

In seguito alla suddivisione delle temperature in base alle categorie descritte, si studia la gaussianità delle relative distribuzioni.

### Temperatura 1 stella

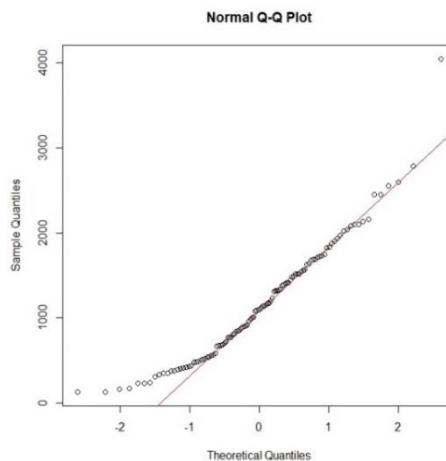


```
> shapiro.test(Tst_1$pl_eqt)

Shapiro-Wilk normality test

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

### Temperatura 2+ stelle

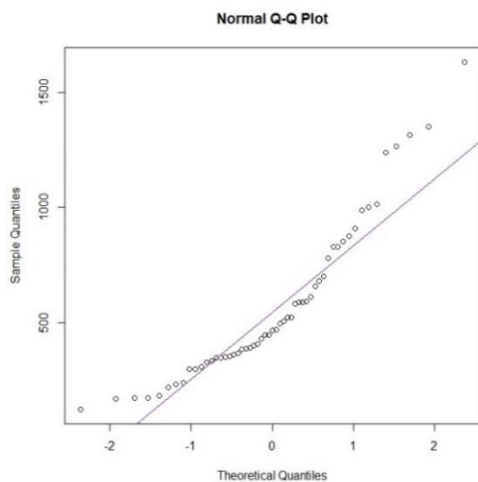


```
> shapiro.test(Tst_234$pl_eqt)

Shapiro-Wilk normality test

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

### Radial Velocity

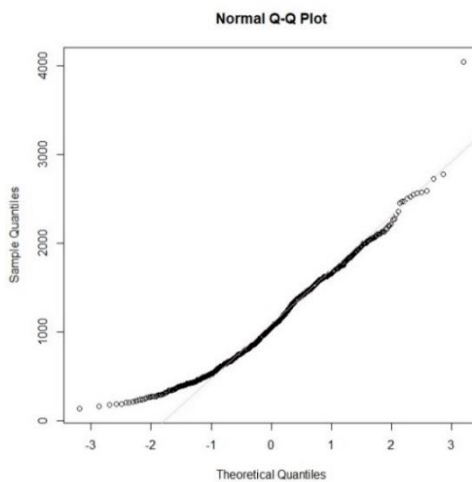


```
> shapiro.test(T_rv$pl_eqt)

Shapiro-Wilk normality test

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

### Transit

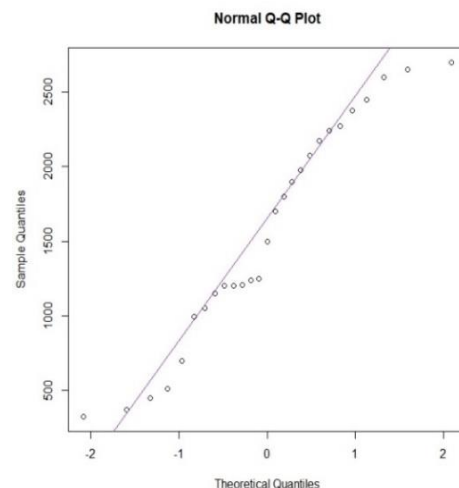


```
> shapiro.test(T_trans$pl_eqt)

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
```

Sebbene la numerosità dei campioni abbia influenzato la potenza del test di Shapiro-Wilks, rendendolo molto suscettibile alle minime variazioni dalla retta dei quantili, i qq plot ci spingono ad affermare l'ipotesi alternativa di non gaussianità. Caso d'eccezione spetta a **T\_ima** che con una numerosità bassa di  $n=27$ , presenta un p-value di 0.1554. In questo caso non vi è forte evidenza per rifiutare l'ipotesi nulla, dunque si accetta la normalità della sua distribuzione.

Per il **teorema del limite centrale** in presenza di campioni numerosi ( $n>30$ ) possiamo assumere che le medie campionarie abbiano approssimativamente distribuzione normale.

## Numero di stelle:

Lo scopo del test è di supportare l'idea che la media delle temperature non sia correlata dal numero di stelle appartenenti al sistema.

A tal fine si effettua un Z-test su due campioni numerosi di tipo bilaterale.

$$Z_0 = \frac{\bar{X}_m - \bar{Y}_n - \delta_0}{\sqrt{\frac{(S_x)^2}{m} + \frac{(S_y)^2}{n}}} \quad \text{con } \delta_0 = 0$$

**“La media delle temperature dei pianeti con una singola stella è diversa da quella di pianeti con più stelle?”**

$$H_0: \mu_{(T_{st1})} = \mu_{(T_{st234})}$$

$$H_1: \mu_{(T_{st1})} \neq \mu_{(T_{st234})}$$

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

```
> 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
```

Il test presenta un p-value di 0.1966. Non vi sono evidenze forti per rifiutare l'ipotesi nulla. Si conclude che la media delle temperature dei pianeti non è dipendente dal numero di stelle appartenenti al sistema.

## Metodo di scoperta:

Si procede ora con il definire un secondo Z-test di tipo bilaterale che verifichi la differenza di temperatura media rilevata in base alla metodologia utilizzata. Il confronto avviene fra i metodi di **Transit** e **Radial Velocity**, sulle quali è applicabile il TLC.

**“I metodi di rilevazione influenzano con un forte bias la temperatura media rilevata?”**

$$H_0: \mu(T_{rv}) = \mu(T_{trans})$$

$$H_1: \mu(T_{rv}) \neq \mu(T_{trans})$$

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

```
> 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
```

Il p-value estremamente basso ci permette di avere forte evidenza contro l'ipotesi nulla. Dunque, si assume vera l'ipotesi alternativa per cui le medie non coincidono.

Il successivo Z-test tra **Radial Velocity** e **Imaging** assume la normalità della media campionaria dei dati tratti dall'Imaging sebbene essa presenti una numerosità piuttosto borderline **n=27<30**. Tale assunzione è valida poiché si è verificato con un test di Shapiro-Wilks che la sua normalità non è rifiutabile. Lo stesso discorso vale per la varianza ignota, stimabile con una buona precisione con

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

$$H_0: \mu(T_{rv}) = \mu(T_{ima})$$

$$H_1: \mu(T_{rv}) \neq \mu(T_{ima})$$

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

```
> 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
```



Il risultato che si ottiene, come si poteva già presupporre, presenta un p-value estremamente basso, dando forti evidenze per rifiutare l'ipotesi nulla e accettare quella alternativa. Possiamo nuovamente affermare che le media siano diverse.

In quest'ultimo test fra **Transit** e **Imaging** si tengono tutte le supposizioni fatte nel punto precedente.

$$H_0: \mu_{(T_{trans})} = \mu_{(T_{ima})}$$

$$H_1: \mu_{(T_{trans})} \neq \mu_{(T_{ima})}$$

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

```
> 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
```

Il test presenta p-value basso. Si rifiuta l'ipotesi nulla e si accetta quella alternativa.

In conclusione si può affermare che la media delle temperature varia fortemente in base ai metodi di rilevazione dei pianeti. Affermiamo quindi la presenza nel pool di un forte bias in favore del metodo utilizzato, favorendo la presenza dei dati relativi al metodo di scoperta "Transit" (il più numeroso).

## Temperatura sopravvivenza rover:

Un ruolo fondamentale nello studio dei pianeti extraterrestri viene ricoperto da piccoli ma sofisticati robot, capaci di operare sotto condizioni estreme per fornire un grande quantitativo di dati necessari allo studio dell'universo: **i rover**.

Si vuole ora studiare la probabilità che un rover possa sopravvivere sugli esopianeti. A tal fine, si considera come modello uno dei più famosi se non più resistenti rover, in grado di resistere fino a temperature estreme di **1300°C** : Perseverance. Per la natura di **pl\_eqt**, parametro in nostro possesso, si suppone che tutti i pianeti siano dei corpi neri a temperatura costante.

Procediamo dunque effettuando un test di proporzione del campione. La numerosità del campione  $n=804 > 30$ , ci permette di trovare uno stimatore  $\hat{p}$  preciso del valore atteso della variabile Bernoulliana. Si assume che ciascuna v.a. sia i.i.d e abbia valore "1" per  $T < 1573K$  o "0" per  $T \geq 1573K$ .

**"La probabilità che Perseverance resista alla temperatura del pianeta è inferiore al 79%"**

$$Z_0 = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$$

$$H_0: p \geq p_0 = 0.79$$

$$H_1: p < p_0$$

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

```
> 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
```

Dato il  $p\text{-value}=0.49>0.05$ , non vi sono prove sufficienti per rifiutare l'ipotesi nulla e accettare l'ipotesi alternativa. Dunque si conclude che la probabilità che il rover possa sopravvivere alla temperatura di un pianeta qualsiasi è maggiore del 79%.

Costruiamo un intervallo di confidenza al 95% per ottenere un range di valori possibili per  $p$ . Si utilizza il valore stimato

$$\hat{p} = \frac{n(T < 1300)}{n} = 0.789801$$

```
> 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
```

$$\text{per costruire un IC del 95\%: } \hat{p} \pm z_{\left(1+\frac{\gamma}{2}\right)} \frac{\hat{p}(1-\hat{p})}{n}$$

Prendendo come riferimento il risultato ottenuto da `prop.test`, costruiamo l'intervallo

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

Posto come centro  $p=0.79$ , l'intervallo si pone a cavallo delle due regioni supportate dalle due ipotesi. Precisamente la regione destra  $p \geq 0.79$  appartiene all'ipotesi nulla, mentre  $p < 0.79$  appartiene alla regione dell'ipotesi alternativa

## Parte III: Regressione

Obbiettivo della nostra regressione sarà quello di determinare un modello per la predizione della temperatura superficiale del pianeta, in base a predittori basati su parametri del sistema planetario d'interesse.

Per iniziare sono stati inclusi tutti i predittori che potessero avere una qualche correlazione con la temperatura del pianeta, per questo sono stati esclusi da subito i parametri sulla posizione del pianeta nel cielo. Si noterà che già da subito la richiesta di avere un valore per ogni colonna restringe di molto il numero di sample utilizzabili, che si attesta sempre attorno al centinaio.

```
>> 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_rad+st_mass')
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

```

Procediamo quindi a ed effettuare il fit lineare multiplo e analizziamo i risultati.

Il primo modello che include la totalità dei predittori ha dei problemi evidenti, tutti i predittori categorici riguardanti il metodo di scoperta sono troppo poco numerosi per dare risultati significativi e tutti i loro valori sono N.A. . Notiamo comunque che il basso valore della statistica F ci assicura la buona significatività generale del modello. Procediamo quindi a rimuovere il predittore categorico “disc\_year” ed effettuiamo nuovamente il fit. Il secondo modello è decisamente migliorato ma notiamo che i pValue per molti predittori sono molto alti. I pValue forniti da MATLAB sono infatti relativi a questo 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

```

Procediamo quindi iterativamente a rimuovere il predittore con i pValue più alto  $> 10\%$ , il risultato finale è una serie di predittori con pValue molto bassi  $< 10^{-5}$ . Notevolmente con questo processo abbiamo anche rimosso il predittore riferito al numero di stelle, confermando ciò che nel test di ipotesi sembrava probabile.

L'unico pValue alto è quello dell'intercetta, vediamo infatti che un test al livello di significatività  $\alpha = 0.05$  ci porterebbe a rifiutare  $H_0: \beta_0 = 0$  per molto poco, non essendo soddisfatti di ciò lasciamo aperta la possibilità che l'intercetta sia nulla.

```
>> 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

```

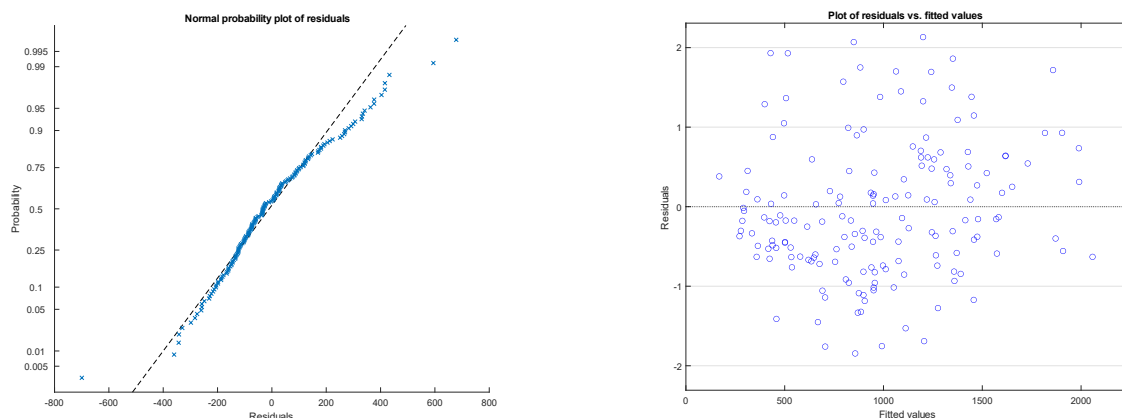
Number of observations: 176, Error degrees of freedom: 169
Root Mean Squared Error: 195
R-squared: 0.885, Adjusted R-Squared: 0.881
F-statistic vs. constant model: 216, p-value = 1.37e-76

```

Il coefficiente di determinazione non è variato molto dal primo modello e al momento si attesta ad un buon valore di  $R^2 = 0.885$  il numero di predittori è molto basso rispetto al numero di osservazioni quindi siamo sicuri non aver sfiorato nell'overfitting, ciò è confermato dal valore sempre buono di

$$R_{adj}^2 = 0.881$$

Analizziamo quindi la distribuzione dei residui:



Sebbene il grafico dei residui sui fitted values mostri una omoschedasticità al limite dell'accettabile, il QQ-plot dei residui ci obbliga a rivedere il modello.

Proponiamo quindi degli altri modelli con predittori elevati al quadrato, la maggior parte contribuiscono ad un aumento del coefficiente di determinazione ma nessuno di essi risolve il problema dell'eteroschedasticità dei dati

```
>> 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:

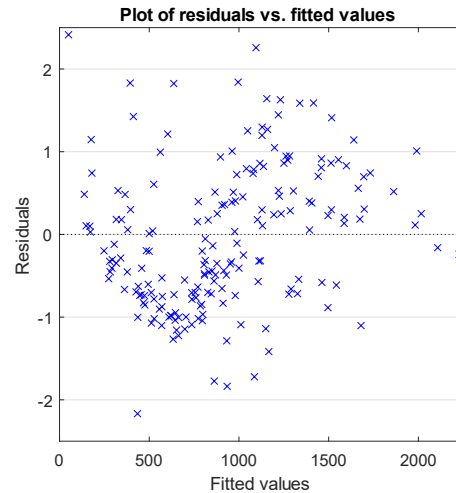
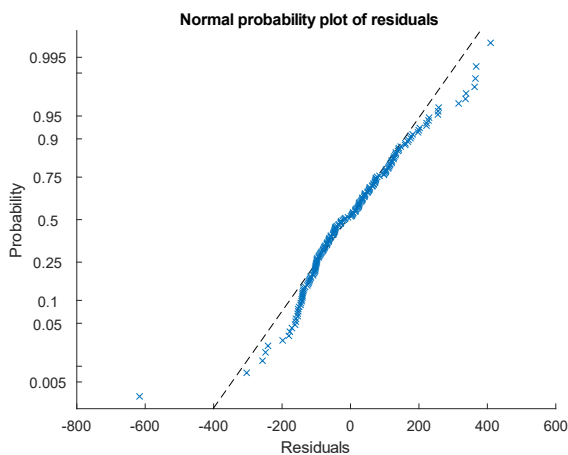
```

	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

```

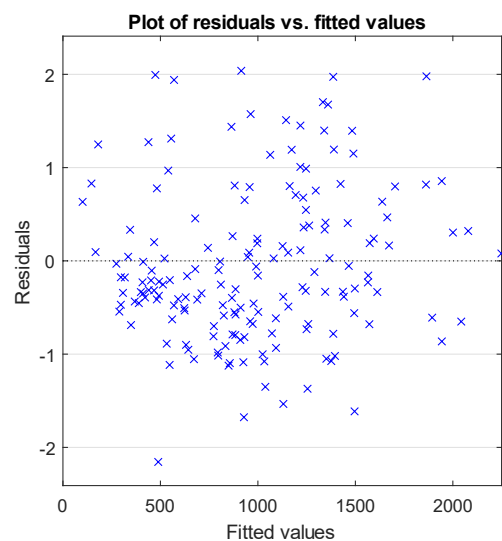
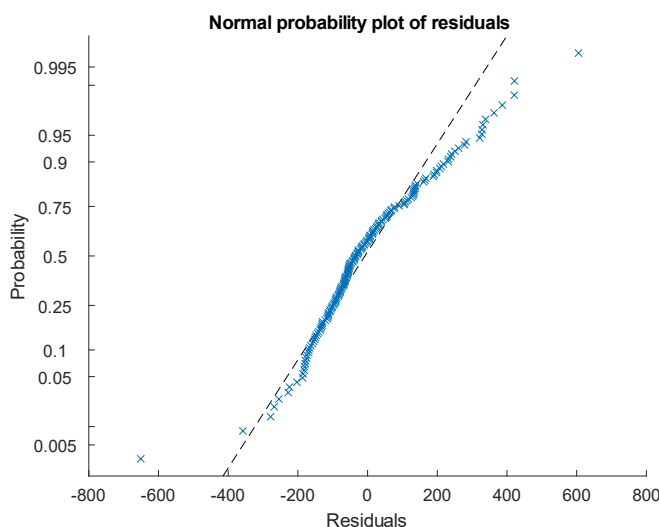


Questo modello è stato ottenuto includendo tutti i predittori e i loro quadrati e procedendo come in precedenza con la rimozione dei predittori meno significativi in ordine di pValue.

Il valore del coefficiente di determinazione di questo modello è decisamente aumentato, attestandosi a  $R^2 = 0.939$  nonostante l'aderenza dei dati ai quantili nel qq-plot è decisamente peggiorata ed anche il plot dei residui mostra un andamento fortemente eteroschedastico, con un accenno di parabola.

Nuovamente notiamo che il coefficiente di determinazione del modello è molto buono ma i residui sono distintamente eteroschedastici.

Nel tentativo di trovare un modello che rispettasse l'assunzione di omoschedasticità dei residui sono stati svariati i modelli che non hanno dato i risultati sperati, modelli con interazioni semplici, con interazioni quadrate o cubiche, con i logaritmi o gli inversi di alcuni predittori. Sono tutti risultati peggiori di quelli sopra presentati, siamo quindi obbligati a ritenere il primo modello, quello con i residui più omoschedastici, quello definitivo. Sottolineiamo comunque che il modello è ben lontano dall'essere perfetto, tuttavia non abbiamo tutti i parametri che potrebbero influenzare la temperatura, come la composizione del pianeta o la composizione della stella, riteniamo dunque il modello abbastanza soddisfacente ai nostri scopi.





```
>> 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
```

Per testare il modello proviamo a fare delle previsioni, in particolare proviamo a predire la temperatura dei pianeti del sistema solare basandoci sui loro parametri usando il primo modello ottenuto.

MATLAB restituisce sia delle stime puntuali per le temperature che degli intervalli di confidenza al 95%. Gli output sono ordinati in maniera crescente per la distanza dal Sole. I risultati sono a dir poco terribili, l'unico pianeta la cui temperatura rientra negli intervalli di confidenza è Mercurio, la cui temperatura è

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

per tutti gli altri pianeti gli intervalli di confidenza sono completamente sbagliati, fornendo addirittura risultati negativi (ricordiamo che le temperature sono in gradi Kelvin).

Possiamo fare un'ipotesi per cui le basse temperature del sistema solare non si accordino con il pool di pianeti molto caldi che abbiamo utilizzato per creare il modello di regressione e che quindi il nostro modello non sia in grado di predire accuratamente membri così atipici rispetto al nostro dataset.

## Conclusioni

Il dataset scelto seppur molto grande permette un numero limitato di deduzioni forti e il grande bias presentato all'inizio non lascia spazio a confronti con gli unici pianeti che conosciamo per misurazioni dirette. L'argomento dell'abitabilità dei pianeti risulta estremamente complesso ed è molto meno incline ad un'analisi statistica di quello che si era inizialmente pensato. Ciononostante i risultati ottenuti potrebbero essere di supporto a uno studio astronomico, senza un'analisi statistica di questo tipo infatti non risulta evidente il bias intrinseco alla scoperta degli esopianeti e lo scienziato messo davanti a questo dataset potrebbe trarre conclusioni erronee sulla popolazione effettiva di esopianeti nell'universo.

## Risorse utilizzate

- Supporto computazionale: MATLAB, R.
- Wikipedia.
- NASA Exoplanet Archive.
- Statistical properties of exoplanets-Planets and Astrobiology (2018-2019) G. Vladilo
- Properties of Extrasolar Planets ASTR 241.
- Direct Imaging, Las Cumbres Observatory.
- Surface Temperatures of Exoplanets (2015) Weisfeiler, Turcotte, Kellogg.
- The Masses and Orbital Dynamics of Exoplanets (2016) Weiss, Marcy.
- Finding the Mass of an Exoplanet - Joey Chatelain, Jeremy Jones, and Ryan Sketch
- PHL's Habitable Exoplanets Catalogue
- PHL's Exoplanets Catalog
- PHL's Data of Potentially Habitable Worlds
- Calculating Exoplanet Properties, Simon Fraser University
- Studio delle condizioni fisiche favorevoli alla fotosintesi per esopianeti nella zona abitabile (2019-2020) Riccardo Maria Ienco
- HD10647 and the Distribution of Exoplanet Properties with Semi-major Axis (2004) J.P. Beaulieu, A. Lecavelierdes Etangs, C. Terquem
- The sub-Jovian desert of exoplanets: parameter dependent boundaries and implications on planet formation (2019) Gyula M. Szabó, Szilárd Kálmán
- Beyond the exoplanet mass-radius relation (2019) Ulmer-Moll, Santos, Figueira, Brinchmann, Faria.
- Exoplanet orbital eccentricity: Multiplicity relation and the Solar System (2015) Mary Anne Limbach and Edwin L. Turner