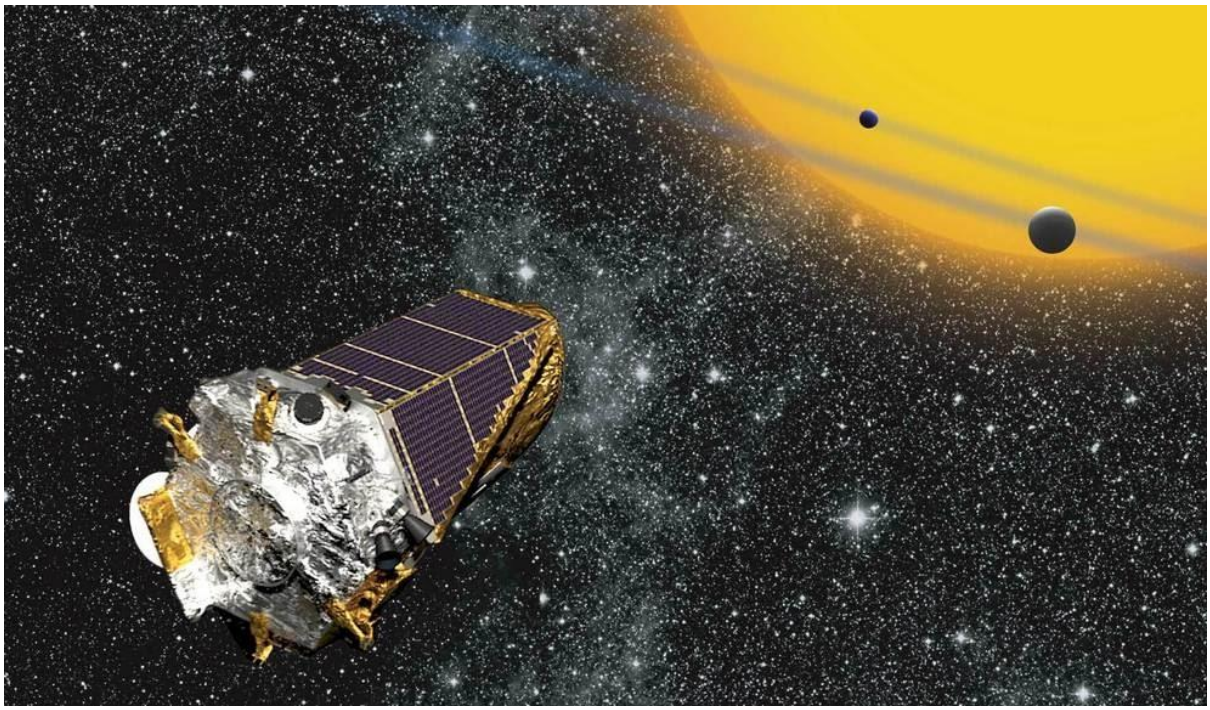
MISSIONE SPAZIALE KEPLER

Studio statistico sulla classificazione degli esopianeti



Politecnico di Milano
Corso di laurea: Ingegneria Fisica
Corso di Statistica - prof. Toigo Alessandro
Anno Accademico: 2020/2021
Relatori:

- Abbate Federica;
- Montanari Matteo;
- Rigolizzo Giulia.

Introduzione

Essendo appassionati di fisica e soprattutto di spazio, abbiamo deciso di sviluppare la nostra tesina analizzando un argomento affine al nostro percorso di studi.

Il progetto è incentrato sullo studio del dataset reso disponibile dalla NASA che concerne l'accurata analisi della sonda Kepler di corpi celesti per verificare la possibilità di classificarli come "esopianeti".

Per esopianeta si intende un pianeta al di fuori del nostro sistema solare il quale può orbitare attorno ad altre stelle o errare nello spazio ("rogue planet") slegato da qualsiasi rapporto con una singola stella.

Il progetto della NASA ha avuto come obiettivo l'invio della sonda spaziale Kepler nel marzo del 2009 per ottenere una maggiore mole di informazioni riguardo i corpi celesti esterni al nostro sistema solare. La sonda, una volta osservato un corpo, lo ha classificato in una prima istanza etichettandolo come possibile esopianeta o meno.

Successivamente, a terra, i dati sono stati ulteriormente analizzati per verificare con sicurezza le decisioni della sonda che potevano essere ribadite o ribaltate, a seconda del risultato dei test nei laboratori della NASA.

Come sono stati rilevati gli esopianeti? Kepler ha utilizzato il *metodo dei transiti:* un fotometro ha monitorato costantemente la luminosità di più di 145000 stelle cercando periodiche diminuzioni della stessa causate da esopianeti che hanno transitato di fronte alla loro stella.

Nel dataset sono anche presenti delle bandierine binarie ("flag") che possono assumere i valori 1: affermativo e 0: negativo che identificano eventuali errori delle misurazioni di Kepler. Perciò, è evidente che nel caso in cui almeno una bandierina fosse 1, l'analisi a terra doveva essere molto attenta e poteva ribaltare la decisione della sonda che poteva essere disturbata da "rumori".

Dopo aver analizzato i motivi della classificazione di alcuni corpi celesti come falsi positivi, ci siamo concentrati sullo studio dei pianeti della fascia abitabile, anche chiamata "Goldilock's zone". Ovvero, la distanza dalla stella per cui le condizioni termiche, né troppo calde né troppo fredde, rendono teoricamente possibile per un pianeta mantenere acqua liquida sulla superficie, essenziale per lo sviluppo di vita. Abbiamo quindi calcolato la fascia di abitabilità di ogni stella con le informazioni fornite dal dataset.

La nostra relazione potrebbe essere d'ispirazione per un'ulteriore missione spaziale futura con una sonda che visiti i pianeti da noi individuati e verifichi la presenza delle caratteristiche essenziali per lo sviluppo della vita sul pianeta.

Presentazione del dataset

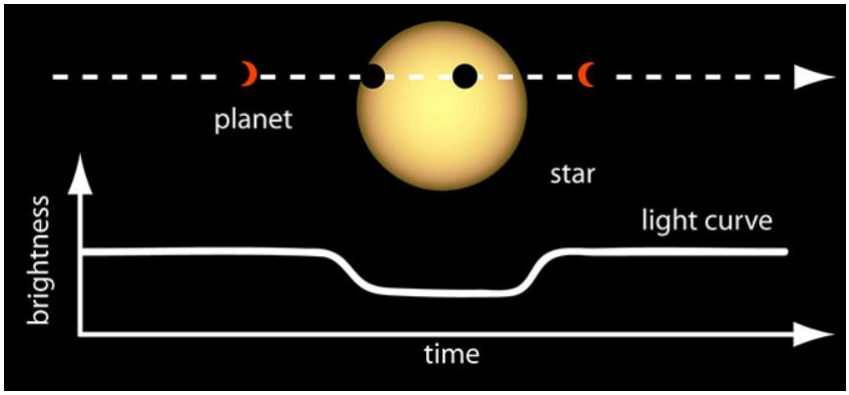
Il dataset è composto da un totale di 4500 righe e 10 colonne. Le righe sono state selezionate in maniera casuale da oltre 9500 proposte dal "NASA Exoplanet Archive", il database che l'agenzia ha dedicato allo studio degli esopianeti. Ciascuna riga è associata ad un corpo celeste.

Riportiamo qui di seguito le informazioni per ogni colonna:

- *K_DISPOSITION:* Designa la decisione presa dalla sonda Kepler per il KOI, Kepler Object of Interest. I valori sono FALSE POSITIVE e CANDIDATE:
    - CANDIDATE: il corpo supera tutti i test effettuati da Kepler e attende di essere confermato da ulteriori studi a terra;
    - FALSE POSITIVE: il corpo celeste non supera i test effettuati da Kepler.

- *DISPOSITION:* Riguarda i dati dell'archivio a terra. I valori assegnati sono CANDIDATE, FALSE POSITIVE e CONFIRMED:
    - CANDIDATE: il corpo celeste è candidato ad essere esopianeta ma bisogna attendere ulteriori ricerche;
    - FALSE POSITIVE: il corpo celeste non è un esopianeta;
    - CONFIRMED: il corpo celeste è un esopianeta.

- *NT_FLAG:* La curva di luce del KOI non è coerente con quella di un pianeta in transito di fronte ad una stella. Ciò può accadere per errori strumentali o perché quella osservata è una stella variabile, la cui luminosità cambia notevolmente nel tempo;



- *SE_FLAG:* La variazione di luminosità della stella non è causata da un pianeta in transito ma da un'eclissi di stella binaria. Una stella binaria è un sistema formato da due stelle che orbitano intorno al loro comune centro di massa;

- *CO_FLAG:* Il segnale ha subito un'interferenza da una stella vicina che l'ha compromesso;

- *EM_FLAG:* Il KOI condivide lo stesso periodo e lo stesso momento di transito di un altro oggetto. Potrebbe essere stato causato da problemi strumentali;

- *KOI_RAD:* Il raggio del pianeta normalizzato per il raggio della Terra ($R_{terra} = 6.371\ km$);

- *STELLAR_TEMP:* La temperatura della stella misurata in Kelvin;

- *STELLAR_RAD:* Il raggio della stella normalizzato per il raggio solare ($R_{sole} = 696.340\ km$);

- *DISTANCE:* Metà del semiasse maggiore dell'ellisse che definisce l'orbita di un pianeta. È in ottima approssimazione la distanza pianeta-stella. La misura è effettuata in unità astronomiche.

Per quanto riguarda i calcoli matematici, per ricavare la fascia di abitabilità abbiamo considerato la luminosità del Sole come: $L_{Sole} = 3.832 \cdot 10^{26}\ W$ e utilizzato le seguenti formule:

$$D_{stella} = 1.77\sqrt{\frac{L_{stella}}{L_{Sole}}} \quad d_{stella} = 0.75\sqrt{\frac{L_{stella}}{L_{Sole}}}$$

Per D si intende la massima distanza dalla stella e d la minima in modo tale che la differenza tra le due identifichi l'ampiezza della corona circolare abitabile.

Invece, per la luminosità della stella $L_{stella} = 4\pi\ \sigma_{S-B}\ R_{stella}^2\ T_{stella}^4$ dove $\sigma_{S-B} = 5.670367 \cdot 10^{-8}\frac{W}{m^2k^4}$

A questo punto combinando le due formule si ottiene:

$$D_{stella} = 1.77\sqrt{\frac{4\pi\ \sigma_{S-B}\ R_{stella}^2\ T_{stella}^4}{L_{Sole}}} = 1.77 \cdot 2\ R_{stella}\ T_{stella}^2\sqrt{\frac{\pi\ \sigma_{S-B}}{L_{Sole}}}$$

$$d_{stella} = 0.75\sqrt{\frac{4\pi\ \sigma_{S-B}\ R_{stella}^2\ T_{stella}^4}{L_{Sole}}} = 0.75 \cdot 2\ R_{stella}\ T_{stella}^2\sqrt{\frac{\pi\ \sigma_{S-B}}{L_{Sole}}}$$

Statistica descrittiva

La prima parte della nostra ricerca consiste nello studiare il dataset nel suo complesso. La sonda Kepler opera un primo discernimento tra falsi positivi e candidati ad essere esopianeti. Gli oggetti identificati vengono ulteriormente studiati tramite numerosi test di compatibilità con l'obiettivo di stilare una lista di esopianeti confermati.

1. Studio di *Disposition* e *K_disposition*

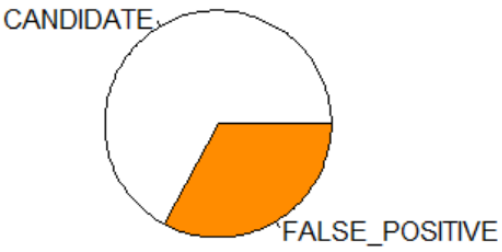
In questa prima tabella osserviamo il numero di esopianeti confermati, di falsi positivi e di quelli che richiedono ulteriori studi.

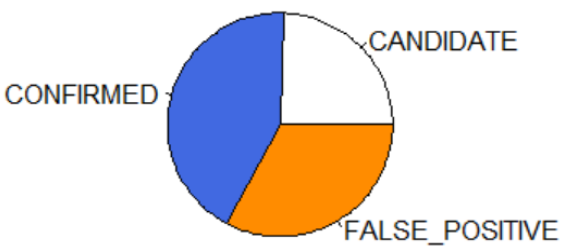
In prima analisi Kepler indica il 67% dei corpi celesti come "candidate" di cui il 63% diventa "confirmed" mentre la restante parte rimane "candidate", ad esclusione di un solo corpo celeste che viene decretato come "false positive". Dunque, più della metà dei candidati viene, in seguito, riconosciuto come esopianeta e quindi aggiunto al catalogo ufficiale degli esopianeti curato dalla NASA.

Per quanto riguarda i falsi positivi, che sono inizialmente il 33% dei corpi celesti, solo 4 vengono ritenuti in seconda analisi degli esopianeti confermati. Dunque, possiamo affermare che la valutazione della sonda, nonostante sia solo un iniziale risultato, è molto puntuale.

```
> table(K_DISPOSITION)
K_DISPOSITION
     CANDIDATE FALSE_POSITIVE
          3021           1479
> table(DISPOSITION)
DISPOSITION
     CANDIDATE      CONFIRMED FALSE_POSITIVE
          1098           1926           1476
> table(K_DISPOSITION, DISPOSITION)
                DISPOSITION
K_DISPOSITION    CANDIDATE CONFIRMED FALSE_POSITIVE
  CANDIDATE           1098      1922              1
  FALSE_POSITIVE         0         4           1475
```

Riportiamo di seguito i grafici a torta relativi alla tabella precedente:





2. Studio delle flag

Consideriamo in un primo momento i soli falsi positivi e studiamone le caratteristiche. Per quale motivo Kepler registra degli oggetti che si rivelano essere dei falsi positivi fin da una prima analisi?

Innanzitutto, notiamo che le quattro flag sono un campione numeroso con distribuzione binomiale. Per la legge dei grandi numeri vale che la frequenza empirica dei successi converge alla probabilità teorica all'aumentare della numerosità delle prove. Quindi, possiamo calcolare qual è la probabilità che un falso positivo sia causato da una determinata flag calcolando la media sulle singole colonne ed emerge che la flag più frequente è la "co_flag" che indica che il segnale di tali oggetti è falsato dalla presenza di stelle vicine.

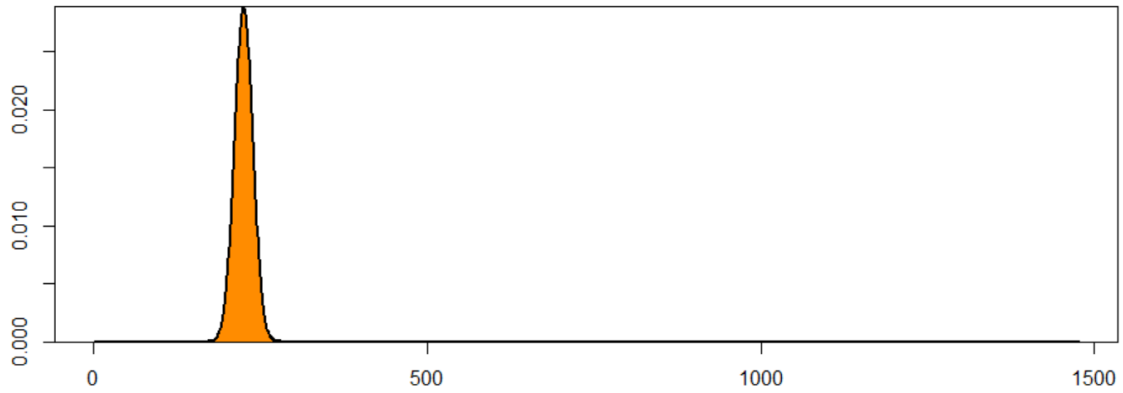
```
> mean(falsi.positivi$NT_FLAG)
[1] 0.152439
> mean(falsi.positivi$SE_FLAG)
[1] 0.5149051
> mean(falsi.positivi$CO_FLAG)
[1] 0.6084011
> mean(falsi.positivi$EM_FLAG)
[1] 0.3414634
```

Numerosità popolazione FALSE_POSITIVE

```
> dim(falsi.positivi)
[1] 1476    4
```

È inoltre possibile approssimare la distribuzione binomiale con una densità gaussiana data la numerosità del campione. Riportiamo a titolo esemplificativo il grafico della distribuzione binomiale (in arancione) della flag "nt_flag". Aggiungendo la curva della distribuzione normale (in nero) notiamo che le due distribuzioni sono perfettamente combacianti.

$$\mathcal{B}(n,p) \approx N(np, np(1-p))$$

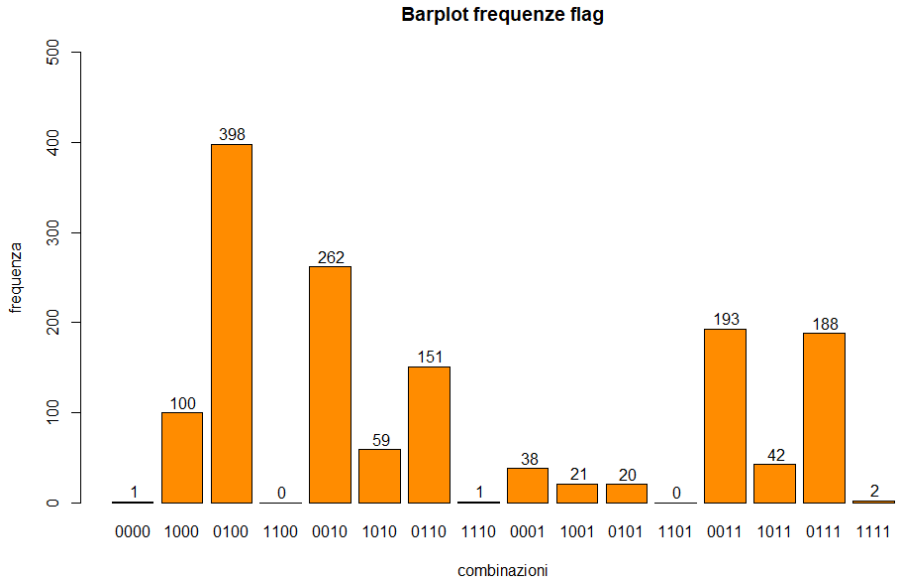


Poiché è possibile che un solo corpo celeste presenti più di una flag, studiamo la frequenza di tutte le combinazioni possibili.

```
> freq.flag
   nt_flag se_flag co_flag em_flag Freq
1        0       0       0       0    1
2        1       0       0       0  100
3        0       1       0       0  398
4        1       1       0       0    0
5        0       0       1       0  262
6        1       0       1       0   59
7        0       1       1       0  151
8        1       1       1       0    1
9        0       0       0       1   38
10       1       0       0       1   21
11       0       1       0       1   20
12       1       1       0       1    0
13       0       0       1       1  193
14       1       0       1       1   42
15       0       1       1       1  188
16       1       1       1       1    2
```

Mentre in prima analisi la “co_flag” sembrava la flag più frequente (con media più alta), scopriamo ora che è spesso associata alle altre flag, dato che da sola (combinazione 0100) causa circa il 18% dei falsi positivi. L’errore più comune è invece causato dalla sola “se_flag”, che rappresenta il 27% del totale.

Di seguito riportiamo il barplot che mostra le frequenze delle flag:



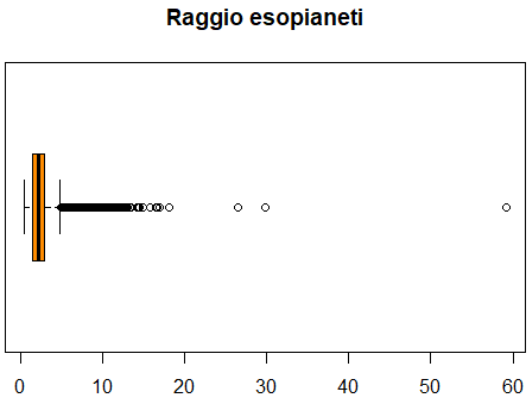
Quindi la maggior parte degli errori di Kepler è esclusivamente causato da una eclissi di una stella binaria che viene scambiata per un esopianeta in transito davanti alla stella.

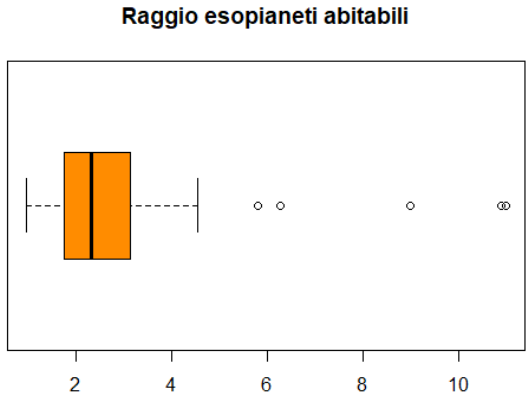
3. Studio delle caratteristiche degli esopianeti e delle stelle

Nell’ambito della statistica descrittiva qui di seguito tratteremo i soli esopianeti confermati tali dall’analisi a terra. In particolare, l'obiettivo è quello di analizzare diversi parametri confrontando i grafici ottenuti per gli esopianeti “CONFIRMED” e per quelli tra questi che rientrano anche nella fascia abitabile. Sarà possibile così iniziare a cogliere delle relazioni tra i parametri selezionati e l’abitabilità dei KOI, argomento che sarà poi oggetto di uno studio approfondito successivamente. I parametri selezionati per questa prima analisi qualitativa sono: il raggio del KOI, la distanza del KOI dalla stella, il raggio e la temperatura della stella. Saranno riportati due box-plot e due summary: il primo relativo a tutti gli esopianeti, il secondo riguardante i soli esopianeti nella zona abitabile.

Raggio esopianeta:

Innanzitutto, occorre far presente che nella colonna KOI_RAD del dataset è riportato il raggio del KOI normalizzato per il raggio della terra.






```
> summary(esopianeti$KOI_RAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.500   1.500   2.110   2.641   2.840  59.190
> summary(abitabili$KOI_RAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.970   1.770   2.325   2.858   3.095  10.990
> var(esopianeti$KOI_RAD)
[1] 6.694899
> var(abitabili$KOI_RAD)
[1] 4.206707
```

Osservando il boxplot relativo al raggio degli esopianeti si nota un numero di outliers molto elevato e il baffo a destra più lungo, asimmetria confermata anche dagli indici di posizione (Mean>Median).

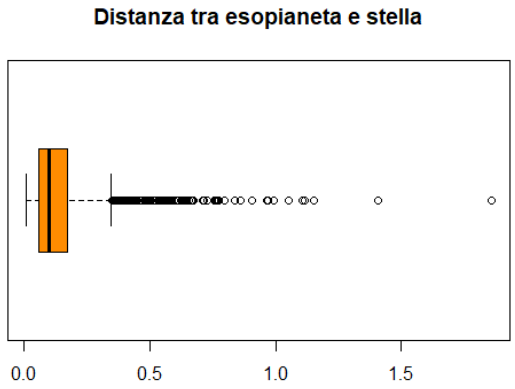
Le stesse osservazioni valgono per il grafico degli esopianeti abitabili.

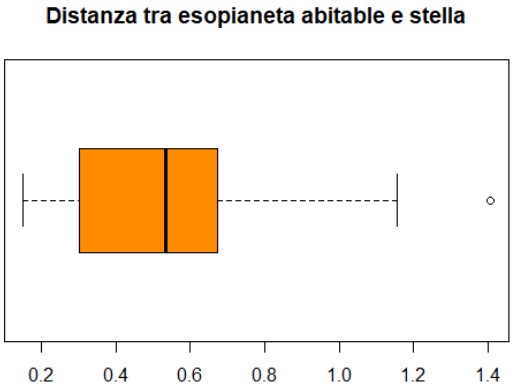
Il fatto che il raggio sia normalizzato con quello terrestre è un importante spunto di riflessione.

In particolare, si osserva che il massimo nel caso degli esopianeti abitabili si riduce in maniera drastica (di circa 48 unità), sebbene le medie rimangano molto simili. Il minimo si avvicina invece notevolmente all'unità e la varianza campionaria diminuisce. Quindi è evidente che i pianeti nella zona abitabile abbiano un range minore. Infatti, non ci aspetteremo dimensioni né troppo minori né troppo maggiori di quelle della Terra per i Goldiloc's planets.

Allora sorge spontaneo un quesito: è vero che il raggio è un utile elemento di confronto per presupporre l'abitabilità di un pianeta? Questo spunto verrà approfondito nella sezione di regressione.

Distanza esopianeta-stella:





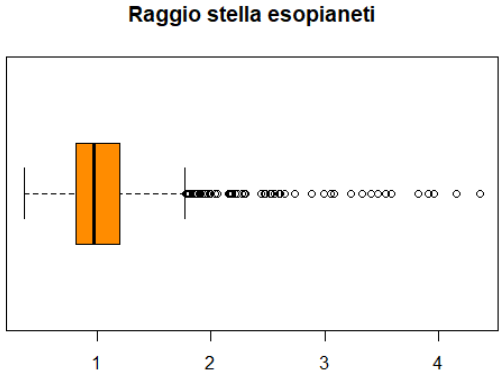

```
> summary(esopianeti$DISTANCE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00890 0.05750 0.09925 0.14507 0.17220 1.86000
> summary(abitabili$DISTANCE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1510  0.3029  0.5340  0.5387  0.6721  1.4062
> var(esopianeti$DISTANCE)
[1] 0.02218776
> var(abitabili$DISTANCE)
[1] 0.08315792
```

Considerando che i dati sono in unità astronomiche, dove un'unità nasce come valore medio della distanza tra Terra e Sole, è evidente che la distanza esopianeta-stella per i pianeti abitabili si avvicina molto più a quella Terra-Sole rispetto a tutti gli esopianeti.

Inoltre, la presenza di un solo outlier dimostra che i valori delle distanze dei pianeti abitabili sono molto più concentrati attorno alla mediana.

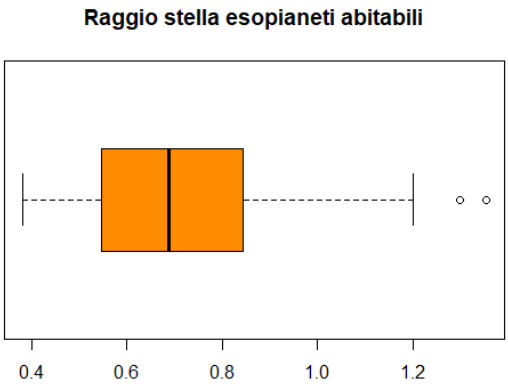
A questo punto riteniamo necessario approfondire lo studio di due grandezze della stella (il raggio e la temperatura) fondamentali nel determinare la fascia di abitabilità per un esopianeta.

Raggio stella:



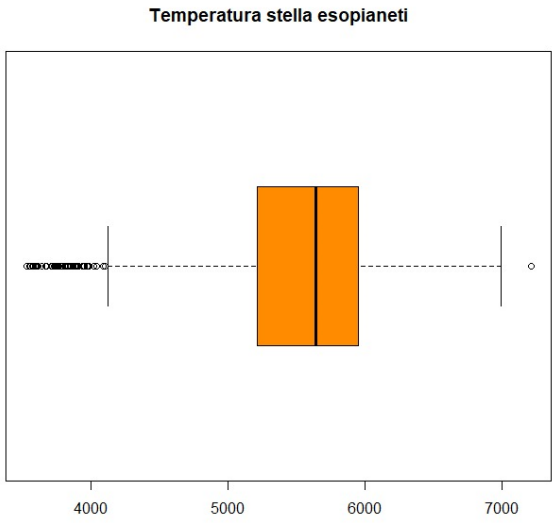

```
> summary(esopianeti$STELLAR_RAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.369   0.818   0.975   1.058   1.202   4.361
> summary(abitabili$STELLAR_RAD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3830  0.5497  0.6880  0.7214  0.8407  1.3540
> var(esopianeti$STELLAR_RAD)
[1] 0.1765831
> var(abitabili$STELLAR_RAD)
[1] 0.04549764
```

Dalla distribuzione dei dati nel primo boxplot si riscontra una forte asimmetria, in particolare si nota che è presente una grande quantità di outliers. La coda di destra è più lunga di quella di sinistra, evidenza che ci viene confermata dall'osservazione degli indici di posizione (media> mediana).



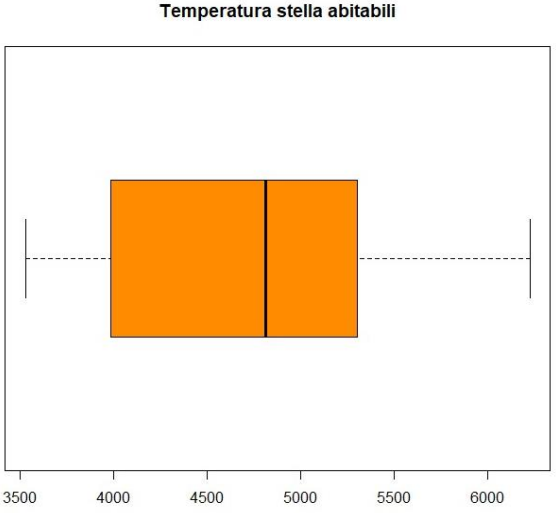
La maggior parte degli esopianeti della fascia abitabile ha una stella con un raggio inferiore rispetto alla media degli esopianeti confermati.
Notiamo che il raggio del Sole è più simile alla media di tutti gli esopianeti piuttosto che alla media degli abitabili.
Cosa che ci fa supporre, come poi verificheremo, che il raggio solare non è un parametro di confronto rilevante per l'abitabilità.

Temperatura stella:




```
> summary(esopianeti$STELLAR_TEMP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3526    5214    5640    5511    5951    7216
> summary(abitabili$STELLAR_TEMP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3526    3992    4810    4721    5296    6226
> var(esopianeti$STELLAR_TEMP)
[1] 400200.9
> var(abitabili$STELLAR_TEMP)
[1] 551902.7
```

Anche per la temperatura della stella è presente un'evidente asimmetria in entrambi i boxplot.
Si nota inoltre che i grafici presentano il baffo a destra più lungo, come riportato anche dagli indici di posizione (media>mediana).



Mentre a destra del box plot degli esopianeti è presente un unico outlier, a sinistra ce ne sono in abbondanza. Invece, non sono presenti outliers nel boxplot degli abitabili, a dimostrazione della forte concentrazione dei dati intorno alla mediana.

Risalta subito l'elevato valore della varianza in entrambi i casi dovuto ad un range ampio ed a dei valori dell'ordine di 10^3.

Inferenza statistica - test d'ipotesi

Proseguiamo lo sviluppo della tesina con l'inferenza statistica, in cui cercheremo le risposte ai quesiti sorti in precedenza.

Faremo dapprima un'analisi più approfondita sulle flag tramite test d'ipotesi e intervalli di confidenza. Successivamente, confronteremo le caratteristiche dei pianeti di Goldilock con quelle della Terra.

Analisi delle flag

In statistica descrittiva abbiamo osservato che le flag più frequenti sono la "se_flag" e la "co_flag", con una stima puntuale della loro frequenza rispettivamente pari $\widehat{p_1}$=0,515 e $\widehat{p_2}$=0,608. Nonostante la frequenza campionaria sia uno stimatore non distorto della probabilità, vogliamo studiare più approfonditamente la loro probabilità realizzando dei test d'ipotesi su popolazioni binomiali.

Vogliamo stabilire se i dati che abbiamo a disposizione ci permettono di affermare con certezza che:

- *"La probabilità di riscontrare "se_flag" è maggiore del 50%"*
- *"La probabilità di riscontrare "co_flag" è maggiore del 60%"*
- *"La probabilità di riscontrare "co_flag" è maggiore di "se_flag""*

1. La probabilità di riscontrare "se_flag" è maggiore del 50%?

Facciamo un test per un campione di Bernoulli numeroso con le seguenti ipotesi:

H_0	H_1	*"Rifiuto H_0 se"*	*Statistica test*
$p = 0.50$	$p > 0.50$	$Z_0 > z_{1-\alpha}$	$Z_0 = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$

```
> prop.test(length(se.flag), 1476, p=0.5, alternative='greater')

        1-sample proportions test with continuity correction

data:  length(se.flag) out of 1476, null probability 0.5
X-squared = 1.2527, df = 1, p-value = 0.1315
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4931615 1.0000000
sample estimates:
        p
0.5149051
```

Questo test con livello di confidenza al 95% mostra un p-value alto (>5%) che ci porta a non rifiutare Ho e quindi ad affermare che la probabilità è del 50% con una conclusione debole. Quindi, con i dati che possediamo, considerando che li abbiamo inizialmente ridotti, non possiamo affermare che più del 50% degli errori è causato da un'eclisse di stella binaria.

Realizziamo inoltre il seguente intervallo di confidenza bilatero:

```
95 percent confidence interval:
 0.4890646 0.5406673
```

2. La probabilità di riscontrare "co_flag" è maggiore del 60%?

Ripetiamo la stessa tipologia di test con la "co_flag", utilizzando le seguenti ipotesi

H_0	H_1	"Rifiuto H_0 se"	Statistica test
$p = 0.60$	$p > 0.60$	$Z_0 > z_{1-\alpha}$	$Z_0 = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$

```
> prop.test(length(co.flag), 1476, p=0.6, alternative='greater')

        1-sample proportions test with continuity correction

data:  length(co.flag) out of 1476, null probability 0.6
X-squared = 0.39976, df = 1, p-value = 0.2636
alternative hypothesis: true p is greater than 0.6
95 percent confidence interval:
 0.5869818 1.0000000
sample estimates:
        p
0.6084011
```

Anche in questo caso il p-value è maggiore di 0,05 quindi è da ritenere valida l'ipotesi Ho con ancora maggiore certezza essendo questo p-value più alto di quello del test precedente. Questo significa che la frequenza della "co_flag" è del 60%.

L'intervallo di confidenza è il seguente:

```
95 percent confidence interval:
 0.5829075 0.6333242
```

3. La probabilità di riscontrare "co_flag" è maggiore di riscontrare "se_flag"?

Visto il ribaltamento delle stime effettuate nella statistica descrittiva avvenuto nei due punti precedenti, ci chiediamo ora se possiamo comunque affermare che la flag più frequente è la "co_flag".

Facciamo quindi un test per la differenza tra medie di due campioni Bernoulliani indipendenti, ponendo come ipotesi alternativa che la "se_flag" è meno frequente della "co_flag".

$$p_x = frequenza\ "se\_flag"$$
$$p_y = frequenza\ "co\_flag"$$

$$\hat{P} = \frac{m\bar{X}_m + n\bar{Y}_n}{m+n}$$

H_0	H_1	"$Rifiuto\ H_0\ se$"	$Statistica\ test$
$p_x - p_y = 0$	$p_x - p_y < 0$	$Z_0 < z_{1-\alpha}$	$Z_0 = \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{m}+\frac{1}{n}\right)}}$

```
> prop.test(x=c(length(se.flag),length(co.flag)), n=c(1476, 1476), alternative="less" )

        2-sample test for equality of proportions with continuity correction

data:  c(length(se.flag), length(co.flag)) out of c(1476, 1476)
X-squared = 25.825, df = 1, p-value = 1.869e-07
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.06290914
sample estimates:
   prop 1    prop 2
0.5149051 0.6084011
```

Con un p-value così basso (<5%) possiamo, con una conclusione forte, rifiutare H_0 e accettare H_1, quindi dire con certezza che la flag che causa (o contribuisce a causare) il maggior numero di falsi positivi rimane la "co_flag", come avevamo già ipotizzato nella statistica descrittiva.

Caratteristiche dei pianeti abitabili

Ci vogliamo interrogare su quali caratteristiche abbiano in particolare gli esopianeti nella fascia abitabile, il loro rapporto con la corrispondente stella e se vi è una correlazione con i dati del nostro sistema solare.

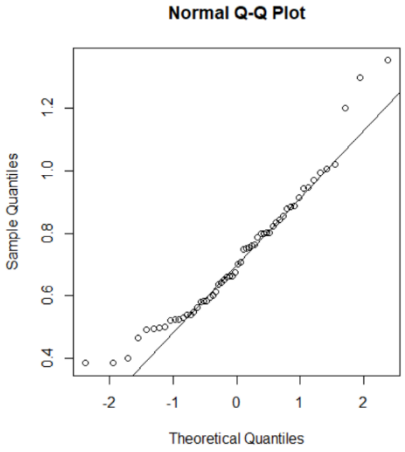
Infatti, ci siamo posti tre domande che ci hanno incuriosito sin da principio, ovvero, se di fatto gli esopianeti nella fascia abitabile appartengano ad un sistema molto simile a quello terrestre:

- *"Gli esopianeti nella fascia abitabile hanno lo stesso raggio della Terra?"*
- *"La stella di riferimento attorno alla quale orbitano ha lo stesso raggio del Sole?"*
- *"La stella di riferimento attorno alla quale orbitano ha la stessa temperatura del Sole?"*

Un riscontro positivo a queste domande ci porterebbe a pensare che i criteri di abitabilità siano facilmente deducibili dalle condizioni terrestri, altrimenti, saremmo disposti a pensare che molti altri fattori condizionano l'abitabilità di un pianeta e che devono essere studiati mediante un'ulteriore missione spaziale.

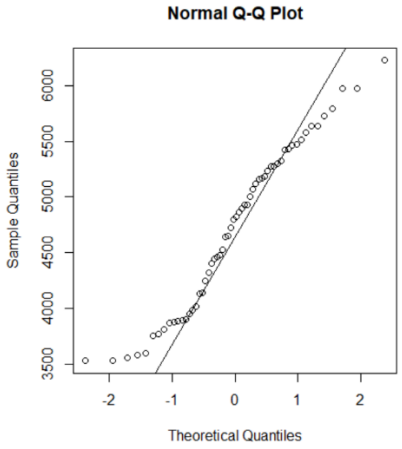
Innanzitutto, studiamo l'eventuale gaussianità delle popolazioni sulle quali ci concentriamo per effettuare i test d'ipotesi. Qui di seguito i QQ-plot e i relativi p-value dei test di Shapiro-Wilk delle popolazioni da analizzare.

Raggio della stella di riferimento



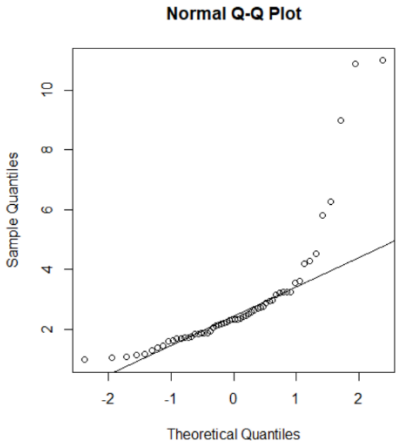

```
        Shapiro-Wilk normality test

data:  abitabili$STELLAR_RAD
W = 0.94788, p-value = 0.01455
```

Temperatura della stella di riferimento




```
        Shapiro-Wilk normality test

data:  abitabili$STELLAR_TEMP
W = 0.95664, p-value = 0.03704
```

Raggio esopianeta




```
        Shapiro-Wilk normality test

data:  abitabili$KOI_RAD
W = 0.67177, p-value = 4.112e-10
```

Notiamo che nei tre casi sia il test di Shapiro-Wilk che il normal QQ-plot confermano la non gaussianità delle due popolazioni.

In particolare, riscontriamo un p-value molto basso nel terzo caso e due più alti ma comunque minori del 5%.

Ora controlliamo la numerosità del campione per verificare se sono soddisfatte le ipotesi del Teorema del Limite Centrale.

```
> length(abitabili$DISTANCE)
[1] 58
```

A questo punto, data la quantità della popolazione (≥30) applichiamo il TLC che ci assicura, seppur approssimando, la gaussianità della media campionaria: $\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

Ora, procediamo con l'impostazione e lo svolgimento dei test d'ipotesi a varianza incognita per un campione numeroso, tutti e tre effettuati con un livello di significatività $\alpha = 0.95$. Ricordiamo che sia il raggio della stella che quello dell'esopianeta sono normalizzati rispettivamente per il raggio del Sole e quello della Terra.

1. *"Gli esopianeti nella fascia abitabile hanno lo stesso raggio della Terra?"*

H_0	H_1	"*Rifiuto* H_0 *se*"	*Statistisca test*
$\mu = 1$	$\mu \neq 1$	$\lvert Z_0 \rvert > z_{1-\frac{\alpha}{2}}$	$Z_0 = \frac{\bar{\bar{X}}_n - \mu_0}{S_n}\sqrt{n}$

```
> z.test(abitabili$KOI_RAD,sigma.x=sd(abitabili$KOI_RAD),mu=1)

        One-sample z-Test

data:  abitabili$KOI_RAD
z = 6.8994, p-value = 5.221e-12
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 2.330260 3.385947
sample estimates:
mean of x
 2.858103
```

2. *"La stella di riferimento attorno alla quale orbitano ha lo stesso raggio del Sole?"*

H_0	H_1	"*Rifiuto* H_0 *se*"	*Statistisca test*
$\mu = 1$	$\mu \neq 1$	$\lvert Z_0 \rvert > z_{1-\frac{\alpha}{2}}$	$Z_0 = \frac{\bar{\bar{X}}_n - \mu_0}{S_n}\sqrt{n}$

```
> z.test(abitabili$STELLAR_RAD,sigma.x=sd(abitabili$STELLAR_RAD),mu=1)

        One-sample z-Test

data:  abitabili$STELLAR_RAD
z = -9.9479, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.6664848 0.7762738
sample estimates:
mean of x
0.7213793
```

3. *"La stella di riferimento attorno alla quale orbitano ha la stessa temperatura del Sole?"*

H_0	H_1	"*Rifiuto* H_0 *se*"	*Statistisca test*
$\mu = 5778\ K$	$\mu \neq 5778\ K$	$\lvert Z_0 \rvert > z_{1-\frac{\alpha}{2}}$	$Z_0 = \frac{\bar{\bar{X}}_n - \mu_0}{S_n}\sqrt{n}$

```
> z.test(abitabili$STELLAR_TEMP,sigma.x=sd(abitabili$STELLAR_TEMP),mu=5778)

        One-sample z-Test

data:  abitabili$STELLAR_TEMP
z = -10.84, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 5778
95 percent confidence interval:
 4529.379 4911.759
sample estimates:
mean of x
 4720.569
```

Guardando i tre p-value, essendo valori bassissimi, è quindi possibile rifiutare le ipotesi nulle (conclusioni forti) a favore di quella alternative. Perciò, grazie a questi test, è possibile capire che non bisogna supporre che un pianeta per rientrare nella fascia abitabile debba avere caratteristiche molto simili a quelle del nostro sistema solare. Ciò, è probabilmente dovuto anche al fatto che concorrono numerosissimi parametri per classificare un pianeta come abitabile. Ad esempio, l'atmosfera, la presenza o meno di acqua e la temperatura del pianeta. Abbiamo quindi raggiunto il nostro obiettivo, quello di mostrare la necessità di effettuare ulteriori missioni spaziali per lo studio di questi esopianeti al fine ultimo di concludere la loro eventuale disponibilità a sviluppare vita.

Regressione lineare

Regressione lineare multipla:

Distanza in funzione della temperatura e del raggio della stella

Innanzitutto, abbiamo pensato di definire un modello di regressione multipla per studiare la dipendenza della distanza esopianeta-stella dal raggio e dalla temperatura della stella.

Ci limitiamo a costruire un modello solo per gli esopianeti che rientrano nella fascia abitabile, perché come vedremo in seguito solo per i pianeti Goldilock è evidente una relazione tra i parametri.

Procediamo con la regressione:

```
> regr<-lm(abitabili$DISTANCE~abitabili$STELLAR_TEMP + abitabili$STELLAR_RAD)
> summary(regr)

Call:
lm(formula = abitabili$DISTANCE ~ abitabili$STELLAR_TEMP + abitabili$STELLAR_RAD)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18483 -0.07962 -0.03809  0.06255  0.35530

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -9.781e-01  1.475e-01  -6.634 1.50e-08 ***
abitabili$STELLAR_TEMP  2.817e-04  5.607e-05   5.025 5.66e-06 ***
abitabili$STELLAR_RAD   2.590e-01  1.953e-01   1.326     0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.125 on 55 degrees of freedom
Multiple R-squared:  0.8186,    Adjusted R-squared:  0.8121
F-statistic: 124.1 on 2 and 55 DF,  p-value: < 2.2e-16
```

Dall’output di R si osserva un modello globalmente significativo, infatti il p-value dell’F-test è circa zero e l’R^2_{adj} di 81.21% che indica che il modello spiega una percentuale elevata dei dati raccolti.

Tuttavia, se si osservano i p-value dei test relativi ai diversi predittori è evidente che il raggio della stella ha un p-value del 19% a dimostrazione della scarsa significatività.

Al fine di migliorare il modello proposto è quindi necessario eliminare tale regressore.

Regressione lineare semplice:

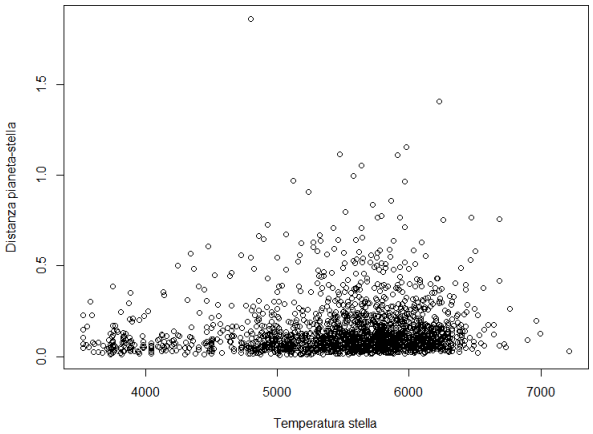
Distanza pianeta-stella in funzione della temperatura al quadrato

Come primo caso di studio della regressione lineare semplice ci chiediamo se la distanza tra un pianeta e la stella intorno a cui orbita dipenda dalla temperatura di quest'ultima al quadrato. Prima ancora di procedere con l'analisi dei dati ci aspettiamo che la relazione non sussista in generale per tutti gli esopianeti.

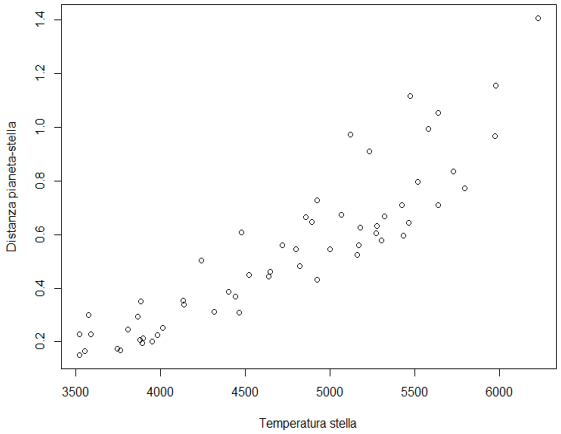
Realizziamo dunque i plot per la distanza in funzione della temperatura delle seguenti popolazioni:

- Esopianeti confermati;
- Esopianeti nella fascia abitabile.

Esopianeti



Esopianeti nella fascia abitabile

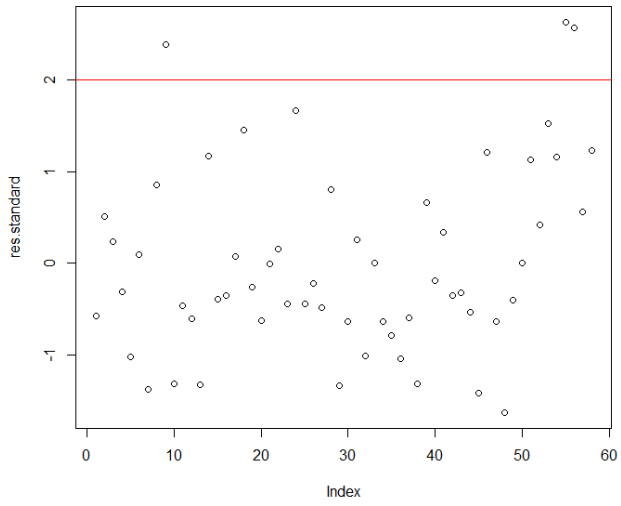


Notiamo subito che, considerando gli esopianeti, il grafico mostra una distribuzione completamente casuale dei dati. Quindi, è confermato che non c'è alcuna relazione tra le due grandezze se consideriamo una categoria così grande. Invece, per i soli pianeti nella zona abitabile abbiamo evidenza di una correlazione tra le due grandezze. Ricordando che nella formula utilizzata per il calcolo della fascia di abitabilità compare la temperatura al quadrato, procediamo quindi con il modello di regressione:

```
> regr=lm(abitabili$DISTANCE~I(abitabili$STELLAR_TEMP^2))
> summary(regr)

Call:
lm(formula = abitabili$DISTANCE ~ I(abitabili$STELLAR_TEMP^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.19289 -0.07349 -0.03339  0.06491  0.31180

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -3.136e-01  5.423e-02  -5.784 3.41e-07 ***
I(abitabili$STELLAR_TEMP^2)  3.734e-08  2.272e-09  16.433  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1206 on 56 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.8252
F-statistic:   270 on 1 and 56 DF,  p-value: < 2.2e-16
```

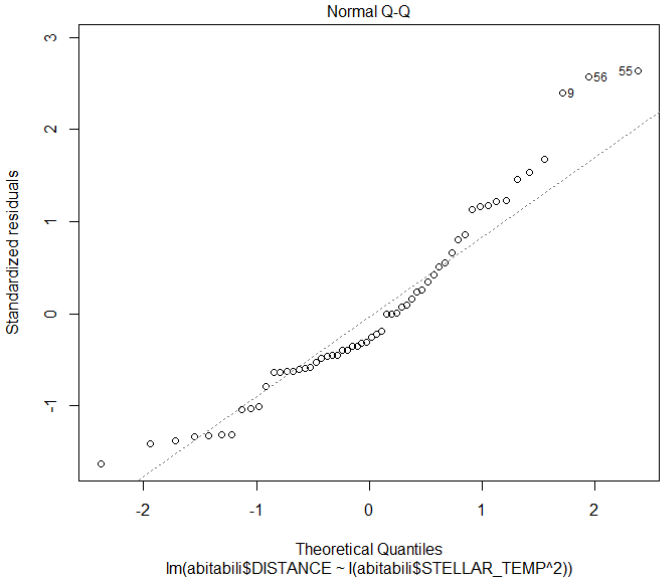
Dalla summary della regressione notiamo che il predittore della temperatura è significativo con un p-value bassissimo. Inoltre, il coefficiente di determinazione R^2 indica che più dell'82% dei dati è rappresentato efficacemente.

Ora che abbiamo dimostrato la validità del modello, passiamo alla verifica delle ipotesi sugli errori. Controlliamo quindi l'omoschedasticità dei residui realizzandone uno scatterplot.



Notiamo che i residui sono distribuiti in modo del tutto casuale e che 55 su 58, quindi circa il 95% di essi, si trova tra -2 e 2, come evidenziato anche dalla retta rossa sovrapposta al grafico. Possiamo dunque affermare che i residui sono indipendenti.

Proseguiamo realizzando il QQ-plot dei residui standardizzati per studiarne la gaussianità.



Notiamo però che i punti non si allineano bene alla QQ-line quindi non possiamo affermare che i residui hanno distribuzione normale.

Questa osservazione è confermata anche dal test di Shapiro-Wilk che presenta un p-value basso.

```
> shapiro.test(res.standard)

        Shapiro-Wilk normality test

data:  res.standard
W = 0.94157, p-value = 0.007593
```

Questo modello non può essere considerato valido perché non sono verificate le ipotesi di gaussianità dei residui. Procediamo con un altro modello di regressione.

Temperatura della stella in funzione del suo raggio

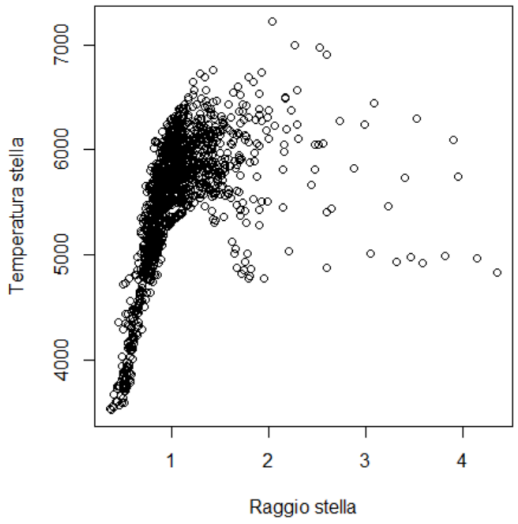
Logicamente, ci si aspetta che le stelle attorno alle quali orbitano i corpi celesti abbiano temperatura maggiore in funzione dell'aumento del proprio raggio.

A questo punto, generiamo i grafici per lo studio di questa ipotizzata relazione per:

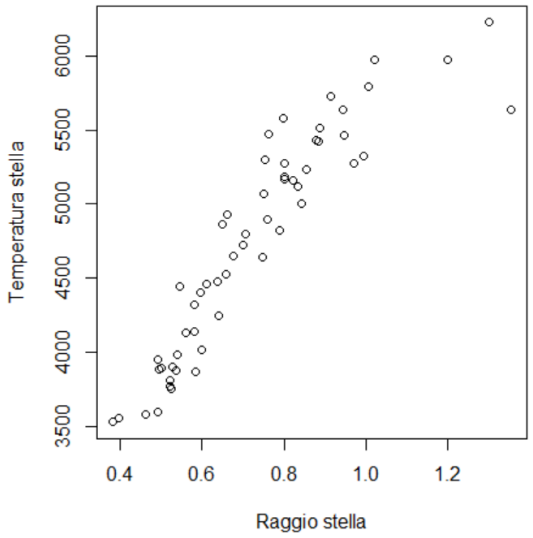
- Esopianeti confermati;
- Esopianeti nella fascia abitabile.

Ci aspettiamo che nel primo caso la correlazione sia casuale, mentre nel secondo c'è una dipendenza forte dato che le condizioni per l'abitabilità sono molto stringenti.

Esopianeti



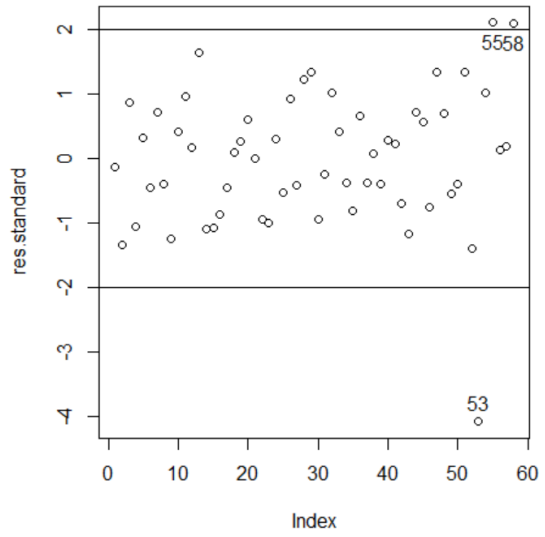
Esopianeti nella fascia abitabile



Notiamo subito una netta conferma delle ipotesi prima avanzate. A questo punto, proseguiamo costruendo il modello di regressione per i soli pianeti abitabili.

```
> regr<-lm(abitabili$STELLAR_TEMP~abitabili$STELLAR_RAD)
> summary(regr)

Call:
lm(formula = abitabili$STELLAR_TEMP ~ abitabili$STELLAR_RAD)

Residuals:
     Min       1Q   Median       3Q      Max
-1106.30  -195.43    24.84   204.02   620.42

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              2415.2      139.1   17.36   <2e-16 ***
abitabili$STELLAR_RAD    3195.8      185.0   17.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 298 on 56 degrees of freedom
Multiple R-squared:  0.842,     Adjusted R-squared:  0.8391
F-statistic: 298.3 on 1 and 56 DF,  p-value: < 2.2e-16
```

Vediamo dalla summary effettuata che il valore del coefficiente di variabilità spiegata è superiore, seppur di poco, a 0.81 che nel nostro caso significa che il modello spiega correttamente l’84,2% dei valori. Inoltre, è evidente dai tre asterischi e dal p-value dell’intercetta e del coefficiente angolare che sono entrambi significativi (ovvero diversi da zero) e che descrivono molto bene il modello.



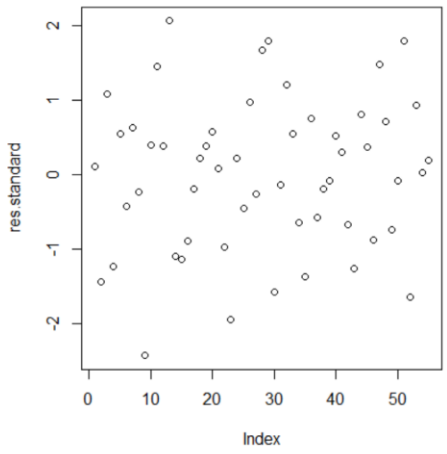
Riportiamo anche lo scatterplot dei residui standardizzati che ci porta ad identificare ed eliminare gli outlier, ovvero quelli presenti sopra il valore 2 sotto il -2 nelle ordinate mediante la funzione identify.

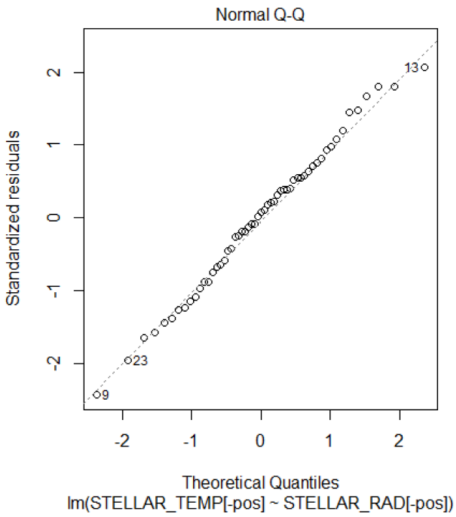
A questo punto, proviamo a rilanciare il modello di regressione modificato per controllare se risulta migliore, come d'altronde ci aspettiamo.

```
regr<-lm(abitabili$STELLAR_TEMP[-pos]~abitabili$STELLAR_RAD[-pos])
> summary(regr)

Call:
lm(formula = abitabili$STELLAR_TEMP[-pos] ~ abitabili$STELLAR_RAD[-pos])

Residuals:
    Min      1Q  Median      3Q     Max
-508.17 -161.74   18.11  136.38  470.69

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2209.4      114.7   19.26   <2e-16 ***
abitabili$STELLAR_RAD[-pos]    3483.3      156.0   22.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 230.1 on 53 degrees of freedom
Multiple R-squared:  0.9039,    Adjusted R-squared:  0.9021
F-statistic: 498.4 on 1 and 53 DF,  p-value: < 2.2e-16
```

Notiamo con grande sollievo che il valore R^2 è pari a 0.9039, decidiamo a questo punto di proseguire l'analisi studiando questo secondo modello di regressione, essendo più preciso del primo.

Verifichiamo le ipotesi di gaussianità del modello:






```
> shapiro.test(regr$residuals)

        Shapiro-Wilk normality test

data:  regr$residuals
W = 0.98906, p-value = 0.8968
```

Generando il Normal QQ-Plot e lo scatterplot dei residui standardizzati notiamo la forte gaussianità di quest'ultimi, come evidenziato dall'alto p-value dello Shapiro-Wilk Test. Inoltre, non si presenta un pattern preciso individuabile nello scatterplot, quindi assumiamo l'omoschedasticità dei residui e perciò il modello è accettabile.

Arrivati a questo punto, vogliamo effettuare una previsione utilizzando questo modello. Infatti, impostando come raggio il valore 1 cioè pari al raggio solare, ci aspettiamo che la temperatura predetta sia appunto quella della nostra stella che ricordiamo essere pari a 5778K.

Inoltre, dal database della NASA scegliamo due pianeti nella fascia abitabile in modo da verificare che le temperature delle rispettive stelle attorno alla quale orbitano siano predette correttamente.

Pianeta	Stella di riferimento	Temperatura stella (K)	Raggio stella (AU)
Terra	Sole	5778	1
Teergarden's Star b	Teergarden's Star	2637	0.127
Kepler-452b	Kepler-452	5757	1.11

```
> newdata<-data.frame(STELLAR_RAD=c(1,0.127,1.11))
> predict(regr, newdata, interval = "prediction")
       fit      lwr      upr
1 5692.678 5218.038 6167.318
2 2651.798 2151.852 3151.744
3 6075.836 5593.378 6558.294
```

Notiamo con soddisfazione una complessiva correttezza nella previsione dei valori. In particolare, la seconda stella, avente raggio molto piccolo, ha una temperatura più accuratamente predetta di quella con raggio maggiore. Questo particolare era intuibile perché la media dei raggi delle stelle di riferimento dei pianeti abitabili è minore di 1. Quindi, il modello è stato costruito utilizzando perlopiù stelle piccole e ha difficoltà a predire valori elevati.

Osservazione:

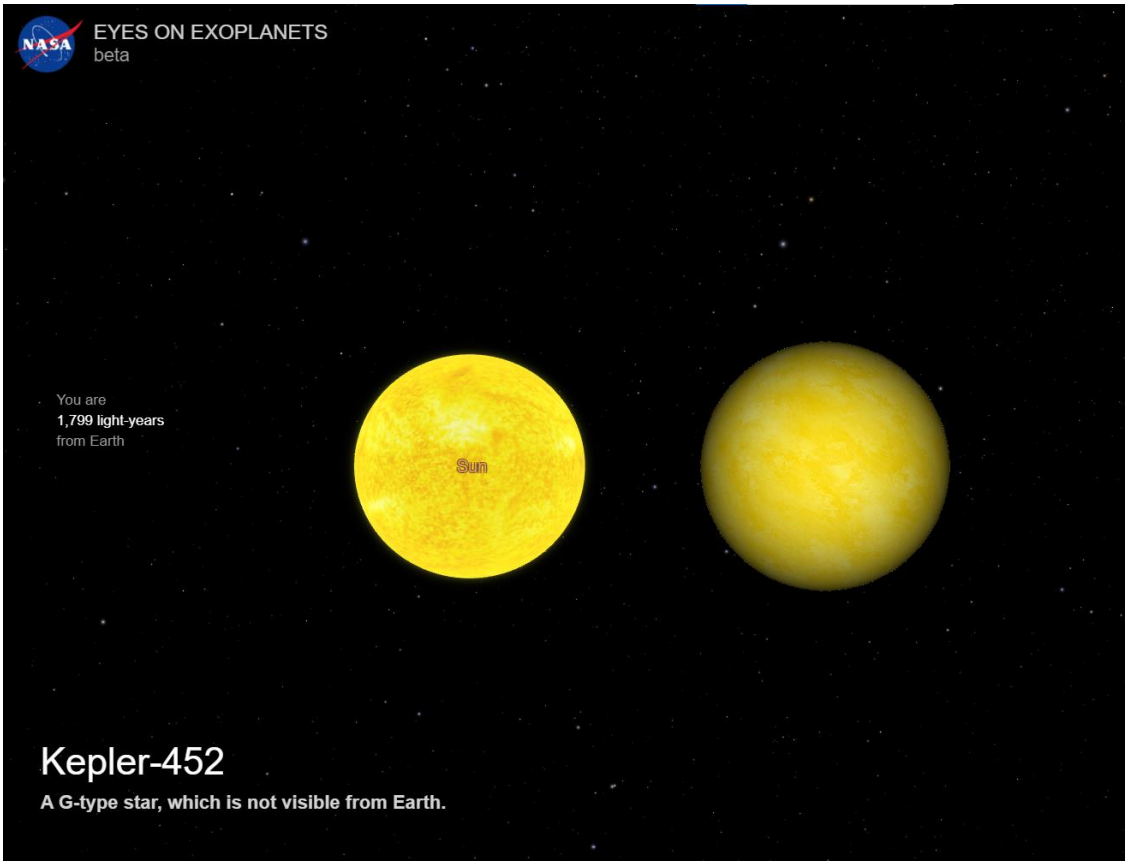
Vogliamo infine porre l'attenzione su un ulteriore modello creato per prevedere il raggio della stella attorno alla quale orbitano gli esopianeti nella fascia abitabile mediante la temperatura della stessa. Il modello è stato scartato in quanto meno intuitivo del modello prima analizzato.

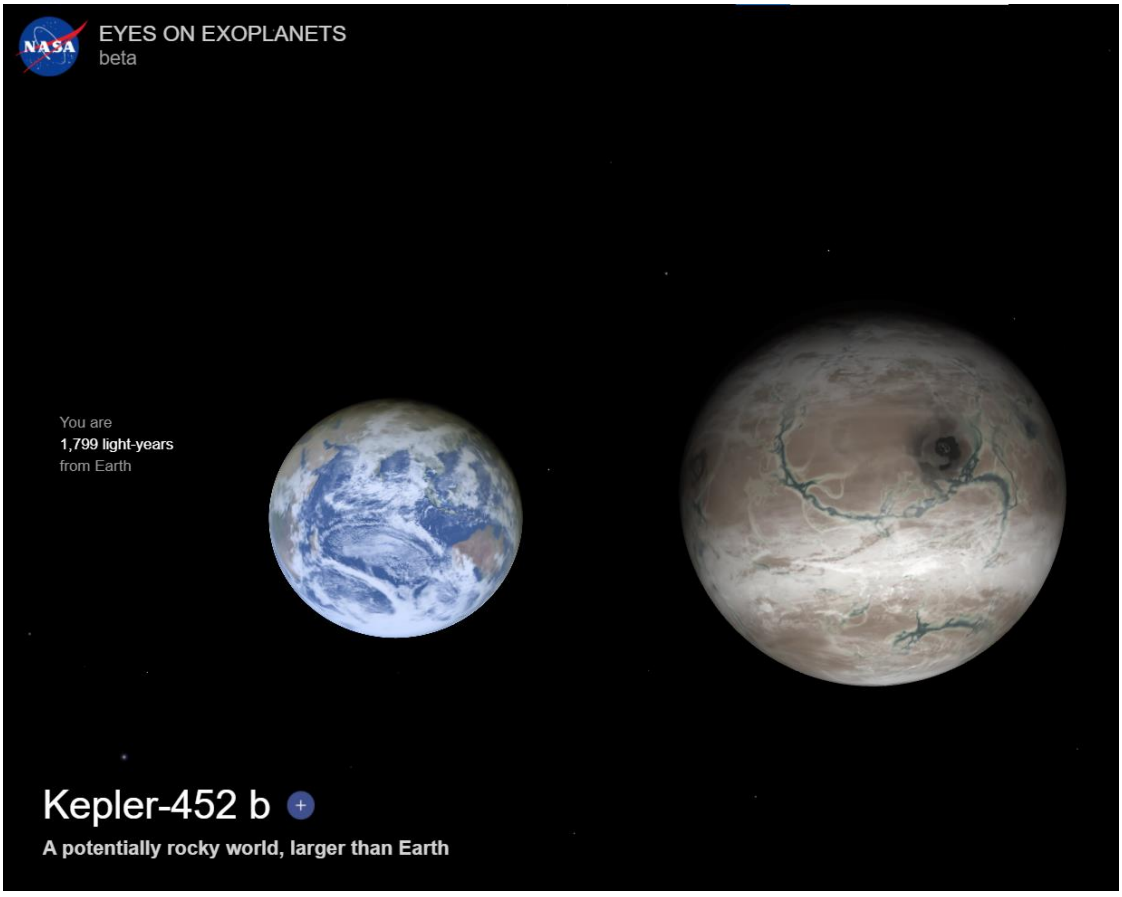
```
> regr<-lm(abitabili$STELLAR_RAD~abitabili$STELLAR_TEMP)
> summary(regr)

Call:
lm(formula = abitabili$STELLAR_RAD ~ abitabili$STELLAR_TEMP)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15687 -0.03798 -0.01311  0.04082  0.39145

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)            -5.223e-01  7.288e-02  -7.167 1.84e-09 ***
abitabili$STELLAR_TEMP  2.634e-04  1.525e-05  17.272  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08555 on 56 degrees of freedom
Multiple R-squared:  0.842,     Adjusted R-squared:  0.8391
F-statistic: 298.3 on 1 and 56 DF,  p-value: < 2.2e-16
```

Un valore all'inizio destabilizzante risalta immediatamente: l'intercetta negativa della retta di regressione.

Dapprima, pare impossibile in quanto vorrebbe dire che con un raggio tendente a 0 mi trovo in una situazione di temperatura negativa.

Ma osservando bene, il valore negativo è dell'ordine di 10^{-1}. Ciò significa che è comunque molto vicino all'origine del sistema di riferimento ma non passa per essa in quanto la retta di regressione interpola al meglio i dati nel nostro dataset e quindi è di fatto un'approssimazione non perfetta della legge fisica.





Immagini che comparano il sistema Kepler-452 a quello solare.

Software utilizzati

Per la realizzazione del progetto abbiamo utilizzato come software:

- R: grazie al quale abbiamo generato i grafici riportati nella relazione e calcolato tutte le informazioni numeriche che ci hanno permesso un corretto svolgimento dell'esperienza;
- EXCEL: ci ha permesso di effettuare una prima scrematura dei dati, essendo in parte incompleti in quanto non in tutte le righe vi erano tabulati i dati. Inoltre, il dataset scaricato dal sito della NASA è appunto un file .xlxs che poi è stato convertito in file .txt per trasportare i dati su R. Tramite Excel abbiamo potuto identificare facilmente i pianeti nella fascia abitabile. Infatti, abbiamo generato due colonne aggiuntive in cui abbiamo calcolato il raggio esterno ed interno della zona abitabile. Infine, ci siamo appoggiati ad un'ulteriore colonna booleana che ha segnalato l'appartenenza o meno dell'esopianeta a tale fascia.
- Mathca.io: Un ottimo semplificatore del linguaggio LATEX per la scrittura di formule matematiche che ci ha permesso di esprimere con semplicità le relazioni tra luminosità di una stella e distanza stella-esopianeta.

Bibliografia e sitografia

- Link dataset sito NASA:
    - https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi
- Spiegazione dati nel dataset:
    - https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html
- Calcolo della zona abitabile:
    - https://exoplanetarchive.ipac.caltech.edu/docs/poet_calculations.html
- Lista dei pianeti nella fascia abitabile riconosciuti dalla NASA :
    - https://en.wikipedia.org/wiki/List_of_potentially_habitable_exoplanets