Kepler

Objects of Interest

Brita von Bartenwerffer
January/February 2021

Kepler Objects of Interest

Brita von Bartenwerffer, January/February 2021

Table of Contents

Foreword

This analysis has been performed in conjunction with the course "Data-Driven Astronomy" offered by the University of Sydney. The aim was to solidify the concepts taught during the course as well as to get more of a feel for astronomical datasets.

The test as well as the code used in the analysis and machine learning is wholly my own as the course used different datasets and -formats. The course module on exoplanets was based on SQL while I'm using Python and its pandas library and the machine learning module worked with data from the Sloan Digital Sky Survey, dealing with galaxies.

This paper makes no pretence of being a scientific work. However, to the best of my knowledge all astronomical facts and conclusions are accurate if slightly outdated since newer missions like the Transiting Exoplanet Survey Satellite (TESS) have come online.

This paper as well as the dataset used and the complete Jupyter notebook are available here: https://github.com/britavb/kepler_objects_of_interest.

I'm happy for anyone to browse the notebook and get inspiration for their own but **please DO NOT reuse my work as your own**. How would you feel if you were asked to peer-review someone's efforts and found yourself looking at your own work?! I did, once, and it wasn't fun.

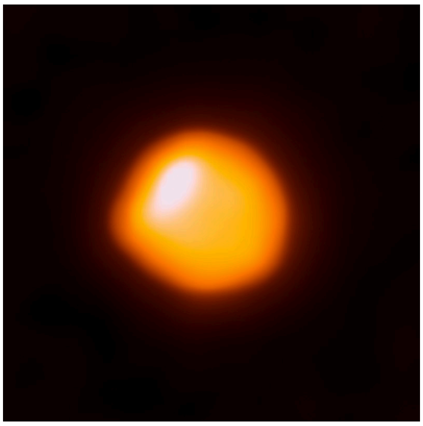
Please be aware that the analysis part is plot-heavy to avoid presenting confusing tables with many rows and columns.

Introduction

Hunting for Exoplanets

The Sun is a star.

This may seem pretty obvious today but for thousands of years people had little idea what the bright pinpricks of lights in the night sky were. Even today with much more sophisticated technology available the biggest stars still appear as little more than fuzzy discs in visible light so it wasn't easy to figure out that they actually *are* like the Sun. It wasn't until the 17th century that scientists began to consider seriously that the sun and stars might be alike, with the latter being incredibly far away.



Red Supergiant Betelgeuse
Credits: ALMA (ESO/NAOJ/NRAO)/E. Gorman/P. Kervella

One of the hot topics in astronomy today is the question "Are We Alone?". Is there intelligent life out there in the universe? Can there even *be* life as we know it elsewhere?

When the first science fiction novels came out, planets around other stars were only a theoretical concept. None had ever been observed. After all, if telescopes can barely make out details of the stars how could they detect a planet around it?

In 1992 the first two exoplanets were discovered, orbiting a pulsar of all things. A pulsar is a dead stellar remnant left behind by a supernova of a star not massive enough to form a black hole. Any planets of that system should have been destroyed by the enormous forces during the supernova event. Scientists found it hard to believe that a pulsar could have planets at all. However, the observations were confirmed and later a third planet was detected. Today it's believed that the planets actually formed after the event from the cloud of dust left over by the explosion.

It took another three and a half years for a planet to be discovered and confirmed in orbit around a Sun-like star. The star in question, 51 Pegasi, has a mass of 1.11 times the mass of the Sun and is older and slightly bigger.

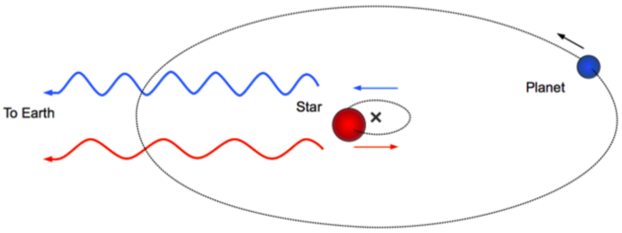
By January 22, 2021 the number of planets around other stars has risen to a staggering 4331 but only a handful were observed directly. This is hardly surprising. The biggest planet in our solar system is Jupyter with a diameter of 139,820 km but this is dwarfed by the Sun's diameter of 1,392,700 km. Stars are also billions of times more luminous that any planet. It would be fiendishly hard to spot even a huge, above Jupyter sized planet in another planetary system against the glare of its host star.

Over the years many methods have been used to find exoplanets, with more or less success. The two main successful methods are:

Detection Method: Radial Velocity

We think of Earth and all other planets of the solar system as orbiting the Sun. However, in physics things are more complicated. While the Sun's gravity keeps Earth in its orbit, Earth's gravity effects the Sun as well. This is known as the two-body problem where two bodies orbit round their combined centre of mass, the barycentre. If there were no other planets in the solar system, the Earth-Sun's barycentre would lie around 30,000 km from the centre of the Sun, well inside the Sun. Both the Sun and Earth would orbit around this barycentre, making the Sun wobble slightly. Having eight more planets in the system makes this a lot more complicated. In general, every body in the solar system orbits around the centre of their combined mass. This is true for moons as well. While they don't orbit around the Sun they and their planets orbit around their barycentre as well. In this case the planet's gravitational attraction is much higher than the Sun's.

The same orbital mechanics work for other planetary systems. They, too, orbit around their combined centre of mass. This effect cannot be observed directly with telescopes because the shift is too small to see over the distances involved. As technology evolved it became possible to observe the minute shift in the star's light as it moved towards and away from us using spectrometers. While the light emitted by the star does not change the light waves will be stretched ("red-shifted") or compressed ("blue-shifted") through the motion.



Credits: The Pennsylvania State University

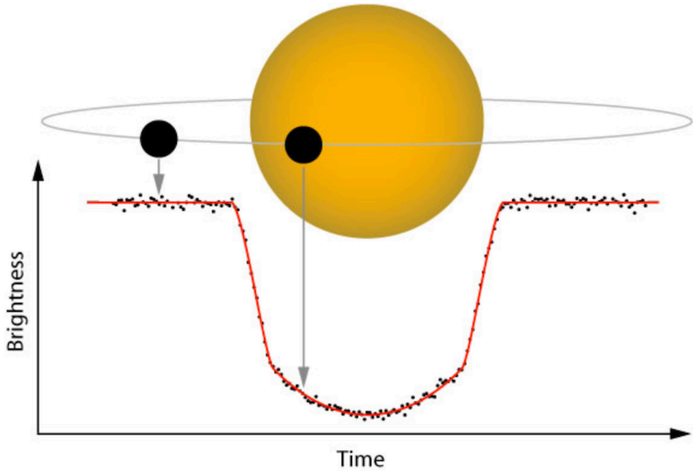
This method only works if we see the planetary system edge-on, otherwise the star will not be moving away or towards us but up and down or left and right and there are no changes in the light's spectrum. In the early days of hunting for exoplanets it was the only way to detect them. The star's "wobble" is bigger if the exoplanet is large because it exerts more of a force onto its star. This explains why almost all exoplanets found in early days were Jupiter-sized and above.

Radial velocity can be used to determine the planet's mass but not its size. A planet could be dense and rocky like Earth or less dense and gaseous like Jupiter and Saturn and still have the same mass but be very different in size.

Detection Method: Transit Photometry

When a planet passes directly in front of its star this is called a transit. Transits can be observed in our own solar system from Earth when either Mercury or Venus passes in front of the Sun. When they do so the Sun's light is dimmed by a very tiny amount, e.g. 0.001% for Venus. And if an alien was observing our solar system the Sun's light would only be dimmed by around 1% by Jupiter.

Although this amount is tiny it can be measured accurately when the planetary system is seen more or less edge-on (otherwise there is no transit). It can be inclined somewhat to the plane of the solar system but the planet still has to be able to pass in front of the star. The best measurements are obtained when the planets pass along the equator of the star though. By measuring every half hour or less a light curve of the transit can be constructed.



Credits: Unknown

The size of the orbiting body can be calculated from the vertex of the light curve. On the other hand, the mass cannot be calculated because the amount of light dimmed has no relation to either the mass or the density of the planet.

By repeated observations over a long time frame it is also possible to determine the size of its orbit from the amount of time that passes between two transits, which again gives us the distance from the star. The same goes for the radial velocity method of course.

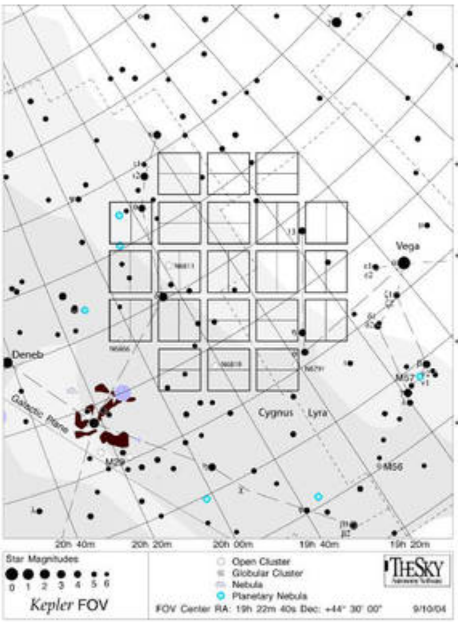
The transit method is much more successful than the radial velocity method and had overtaken the number of exoplanets detected by 2012 because it is by far the easier method to employ. This is where the Kepler mission comes in which used photometry to find exoplanets.

The Kepler Mission

Kepler was launched in early Spring of 2009 and started observations in May of the same year. It was planned to be in service for three and a half years. The mission was extended several times until the spacecraft finally ran out of fuel at the end of October 2018 after nine and a half years. The observatory was decommissioned two weeks later but the data it provided is still analysed today, including follow-up observations for exoplanet candidates.

Kepler was stationed in a heliocentric orbit (i.e. orbiting the Sun), trailing Earth in its orbit at about 151 million km distance.

Initially Kepler's field of view was fixed to a portion of the constellation Cygnus which had been decided on prior to launch. This constellation is easily visible from Earth Northern Hemisphere during summer and autumn and a favourite of (amateur) astronomers because there are *a lot* of stars there. It also hosts the first black hole ever discovered: Cygnus X1. If you are a stargazer, the field of view is located between the bright stars Vega to the East and Deneb to the West.



Note: Celestial North beyond the top left corner
Credits: NASA Ames Research Center

Kepler was a low-cost mission that was build mostly from off-the-shelf rather than bespoke components. Whether this is the reason why one of the reactions wheels needed to keep the space observatory stable broke after three years remains unclear. There was a spare that could be used so the mission was not jeopardized. It was actually extended beyond the mission's planned duration of three and a half years in order to gain more observation time. Unfortunately only half a year later a second reaction wheel broke. This time there was no spare and it could not be repaired remotely. The mission came to an end in May 2013 almost exactly four years after first light. Later that year a secondary mission called K2 was approved and Kepler continued to collect loads of data on exoplanets but its field of view was no longer fixed to the Cygnus constellation.

During its total lifetime of nine and a half years Kepler observed over 530,000 stars and had detected 2,662 confirmed exoplanets by the end of its secondary mission. The number of confirmed planets is subject to change due to follow-up observations made.

The Dataset

When the primary mission ended after four years a catalogue containing over 9564 objects was compiled. These objects were now known as "Kepler Objects of Interest (KOIs)". Scientists analysing them had identified all of them as exoplanet candidates from the many, many images taken by the space observatory during its mission.

In August 2018 something like half of the objects had been weeded out as False Positives by follow-up observations or cross-referencing with other datasets, leaving the other half either as candidates or confirmed exoplanets. Since then there don't seem to have been any updates to the catalogue.

The cumulative dataset can be viewed and downloaded for further analysis from the NASA Exoplanet Archive: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=cumulative and is the data basis this analysis works with.

"The cumulative KOI table gathers information from the individual KOI activity tables that describe the current results of different searches of the Kepler light curves. The intent of the cumulative table is to provide the most accurate dispositions and stellar and planetary information for all KOIs in one place. All the information in this table has provenance in other KOI activity tables."
(https://exoplanetarchive.ipac.caltech.edu/docs/PurposeOfKOITable.html#cumulative)

Data Analysis

Stars of Interest

Masses and Sizes

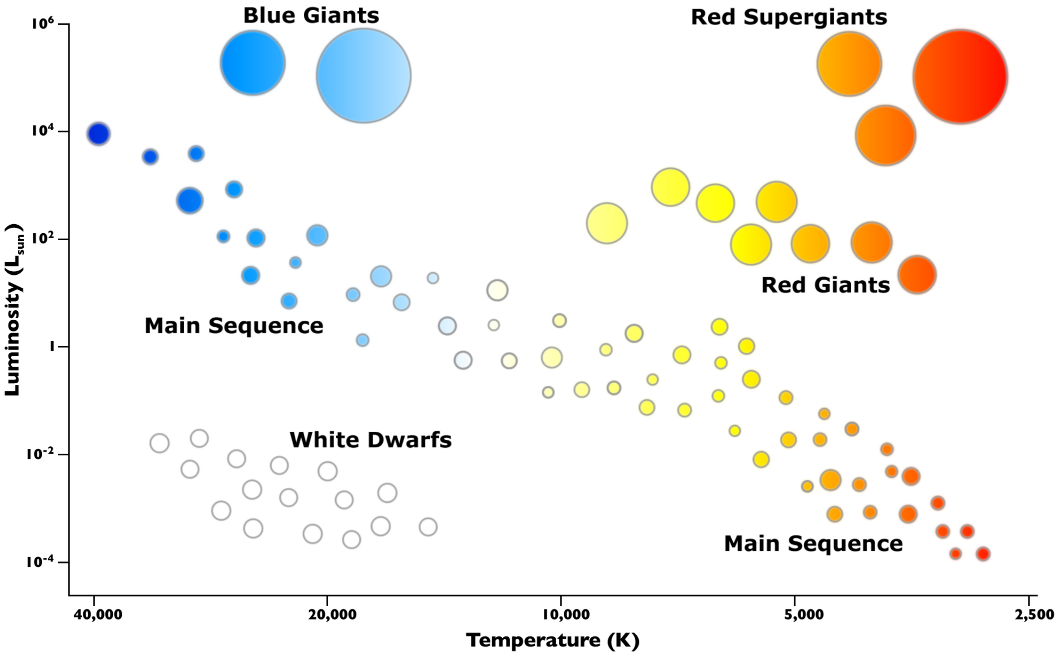
The catalogue contains 8214 stars around which one or more exoplanet candidates or Kepler Objects of Interest (KOIs) were found. Their masses range from 0.1 to 3.7 times the Sun's mass ($M_{\odot}$) with mean and median values being near the mass of our Sun. Deneb, the brightest star in the constellation Cygnus although not in Kepler's field of view, has a mass of approx. 19 $M_{\odot}$. It is a blue supergiant so it seems likely that higher mass stars either don't have planets in the first place or have ejected or consumed them when they left the main sequence and became giant stars.

Our Sun will begin to swell up, cool and become a red giant in about 5 billion years. So far the jury is out on how Earth will be affected but there's little doubt that both Mercury and Venus will be consumed by the swelling Sun. Earth might also find itself consumed or its orbit might slowly drifts outward. The Sun will not explode as a supernova because it doesn't have enough mass. Once it has ejected its outer layers if will contract again and become a white dwarf, a small, hot, dense stellar remnant that will slowly cool over billions of years.

The sizes of the stars in the catalogue are much more diverse, ranging from 0.1 solar radii ($R_{\odot}$) to just over 229. Here, Deneb has a radius of 209 $R_{\odot}$ and we already know it is a supergiant that likely has no planets of its own. Let's have look at all stars with $R_{\odot}$ above 100 and see if they are giants in their own right and can be neglected for this analysis.

Radius (Rsol)	Mass (Msol)	Temp (Kelvin)	Exoplanet Status
229.908000	3.426000	3681.000000	FALSE POSITIVE
180.013000	3.735000	3486.000000	FALSE POSITIVE
162.725000	1.075000	3198.000000	FALSE POSITIVE
152.969000	1.110000	3287.000000	CANDIDATE
151.184000	1.436000	3297.000000	CANDIDATE
150.091000	3.686000	3711.000000	FALSE POSITIVE
138.056000	1.291000	3420.000000	FALSE POSITIVE
116.965000	0.935000	3306.000000	FALSE POSITIVE
101.451000	1.679000	3598.000000	FALSE POSITIVE

Without exception all temperatures are below 4000 K which denotes a cool star. For comparison, the Sun has a photospheric temperature of 5772 K. Looking at the Hertzsprung-Russell-Diagram below tells us that these stars are either much smaller than the Sun, which is obviously not the case here, or have entered their red giant-phase and are less likely to (still) have exoplanets. As the table above shows there are no confirmed exoplanets for either of them.



Credits: Las Cumbres Observatory

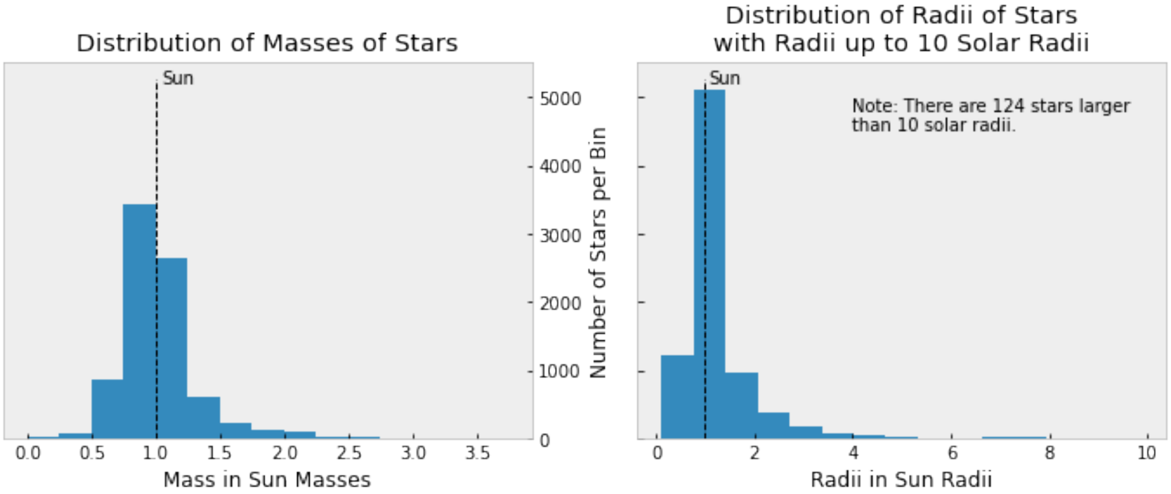
Another possibility is that these stars are still in their infant phases as they, too, are still bloated until each contracts, starts hydrogen fusion in its core and becomes a star on the main sequence. However, this is unlikely to be the explanation here as they can generally only be observed in the infrared because they are still surrounded by the discs of material they collapsed from. Kepler was solely observing in the visible spectrum. On top of that Kepler's field of view was chosen because it did not have many star-forming regions in it.

It should also be noted that observation of planets around giants is more difficult because the planets' orbits are by nature bigger and so might not have been observed once during the four-year window Kepler was actively staring at them.

For plotting purposes all stars with a radius above 10 M☉ will be ignored when the radius is shown in the plot. Out of these 124 stars 106 have been classified as false positives, only one is listed as confirmed and the rest are candidates.

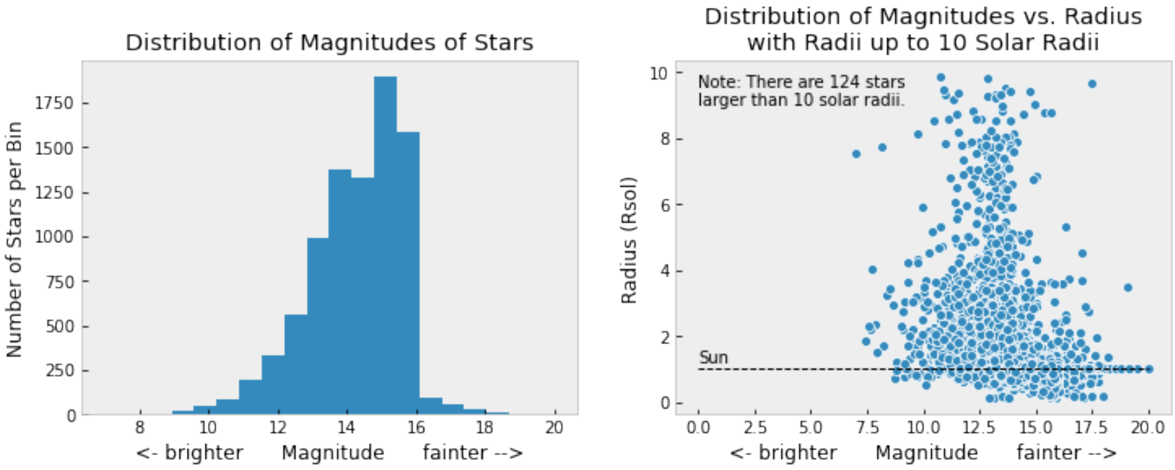
Masses, Sizes and Magnitudes Plotted

Here we can see the distribution of masses and radii of the stars in the catalogue. The vast majority of them are comparable to our Sun, with masses slightly below and radii slightly above Sun's values.



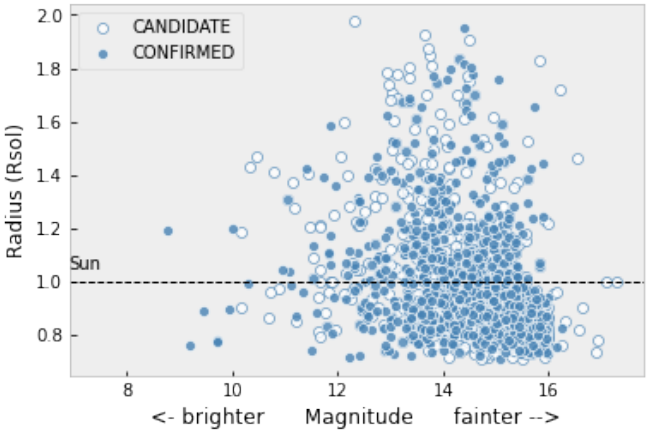
Another characteristic of stars is their visual or apparent magnitude. It measures how bright the star appears to the observer. If the distance to the star is known the absolute magnitude of the star can be calculated. This dataset does not include distances but the plots can tell us a bit about this.

Magnitudes are measured in reverse order: the lower the number the brighter the object. Magnitudes with negative values are also possible: These are very bright indeed! The full Moon has an apparent magnitude of -12.4. The lowest magnitude visible with the naked eye is 6. Any values higher (=fainter) than this require at least binoculars to see.



These plots show the magnitudes measured during the mission by the Kepler space telescope rather than those in the Kepler Input Catalogue. The distribution is screwed to the left with a mean and median around 14.4. None of the stars is visible with the naked eye and only very few might be visible with binoculars.

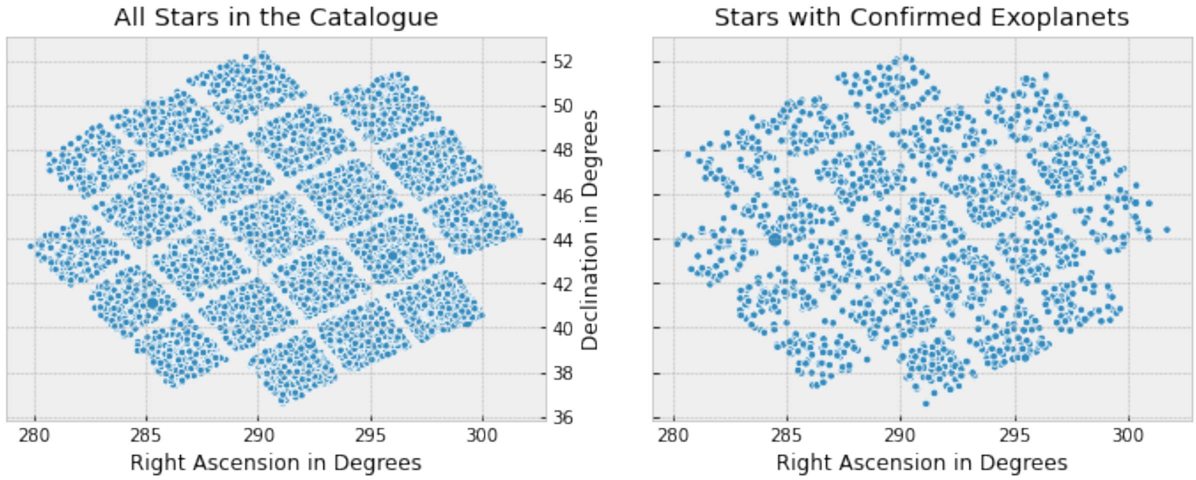
Plotting the magnitudes against the radii of the stars shows a wide spread up to four times the size of the Sun. This gives us an idea of distances: Stars with the same size on the left of the bulge are nearer to Earth while those on the right side are farthest away. It is more difficult to detect exoplanets around smaller stars that are also farther away but in our case there is no bias introduced as both confirmed exoplanets and candidates seem fair evenly distributed for Sun-like stars:



However, there seems to be some sort of bias in the data because there is a clear cut at 16 magnitudes but it's hard to say whether this is by accident or by design. We can ignore this for further analysis though as the visual magnitude does not provide any additional value. A Sun-like star four light-years away can have just as many planets as the Sun does. It's just much fainter to our eyes. If there were alien eyes around there they'd see us just as faint!

Distribution in Field of View

All the stars are fairly evenly distributed across Kepler's field of view. The same goes for stars with confirmed exoplanets. This supports the conclusion that exoplanets are a common phenomenon throughout at least our galaxy if not the whole observable universe. There certainly doesn't seem to be any obvious reason why planets should only be able to form in certain parts of the galaxy but it was the Kepler mission that actually confirmed this view.



Stars and Their Exoplanets

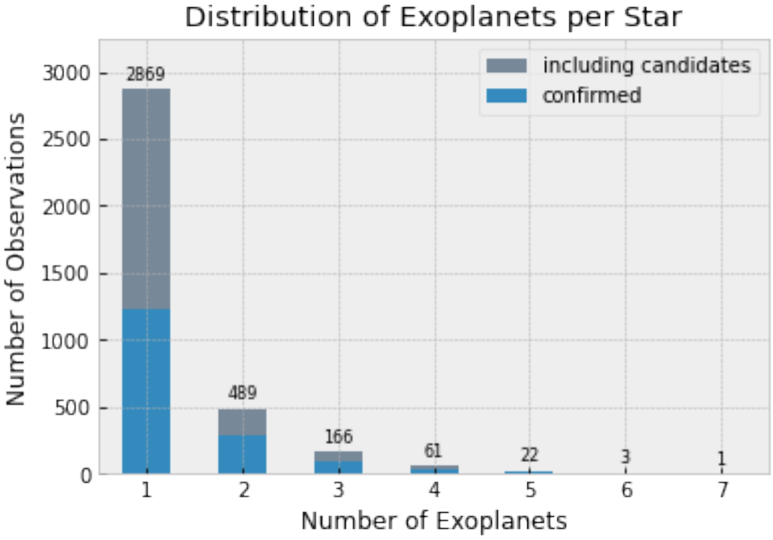
The dataset was last vetted in August 2018 and it's not quite clear whether scientist are still working with it and we may expect updates at some date in the near future.

Count of Exoplanets per Star

Most stars in the survey have only one confirmed exoplanet. The number of planets per star does not change if all candidates are also included which, of course, may or may not turn out to be false positives.

As shown in the plot there are still a lot of exoplanet candidates so it's hard to infer any facts from this. Another problem is that the Kepler mission only went on for four years so there are some limitations:

1. There wasn't enough time to observe most exoplanets with an orbital period longer than four years as these will not have crossed in front of their parent stars at all.
2. A single planet that passed twice in front in its star may in fact have been two different planets with very similar light curves, e.g. a smaller one near the star and a bigger one further out. This is especially true for the candidates.
3. Small planets that orbit further out are very hard to spot so may have been missed when analysing the light curves taken by Kepler.



More observations will be needed to come to any conclusions how many planets per star are typical.

Distances from the Star

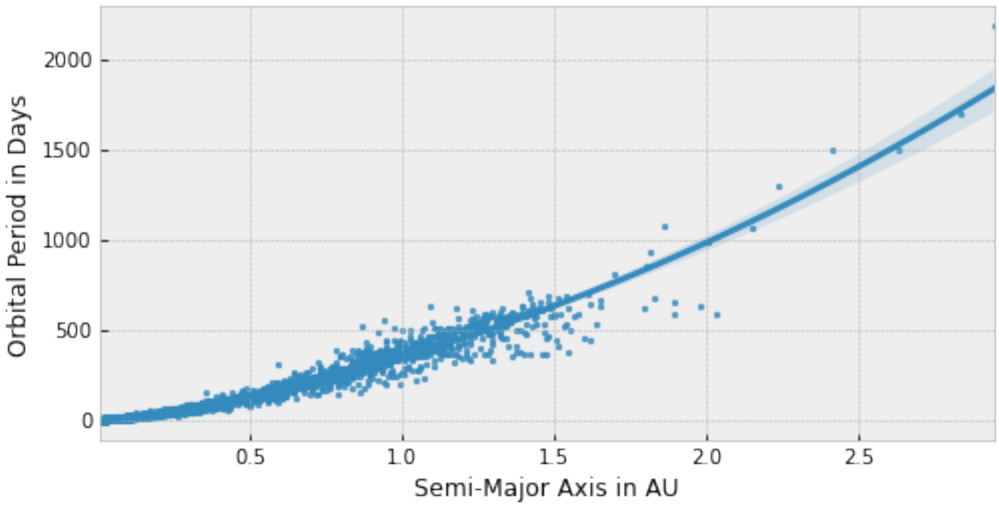
Another interesting property to look at is the semi-major axis of the planet's orbit. The semi-major axis is the distance between the centre and the widest point of an ellipse. In case of a circular orbit around a star it is the actual distance between the star and the planet. In this dataset the eccentricity of the planet's orbit is zero for 9201 of them; the small rest are unknown. So we can take it that the semi-major axis given in the dataset represents the star-planet distance.

In this dataset the semi-major axis is given in astronomical units (AU) with 1 AU being the equivalent of the mean distance between the Sun and Earth: 149.6 million km. In this dataset the highest value is 44.98 AU which is about 10% further from the star than Pluto is from the Sun (or 50% further from the Sun than Neptune is if you want to stick to regular planets). This is an outlier in the dataset as can be seen when querying for planets with a semi-major axis larger than 2.5 AU.

koi_period	koi_sma	koi_prad	koi_disposition	koi_srad	koi_smass	koi_period_years
1500.14068	2.6340	18.35	CANDIDATE	2.283	1.083	4.109974
1693.66362	2.8364	10.23	CANDIDATE	1.021	1.059	4.640174
2190.70104	2.9456	3.12	CANDIDATE	0.736	0.709	6.001921
129995.7784	44.9892	2.99	CANDIDATE	0.726	0.717	356.152818

This exoplanet candidate is the only outlier in the dataset with such a large semi-major axis. By itself it isn't particularly remarkable. Having a radius of just under three times Earth's radius it lies firmly inside the bulk of observed and confirmed exoplanets. It orbits around K01174.01 but is the only planet observed to do so. With both radius and mass being 0.7 times the Sun's values the planet is probably comparable to our ice giants, Uranus and Neptune - or a small star in orbit around a bigger one. The distances involved certainly support the possibility of it being a small stellar object; it is questionable whether a smallish star's gravity is sufficient to keep a smallish ice giant in orbit this far out. For now though I will simply ignore this one planet in order to keep the plots reasonably readable.

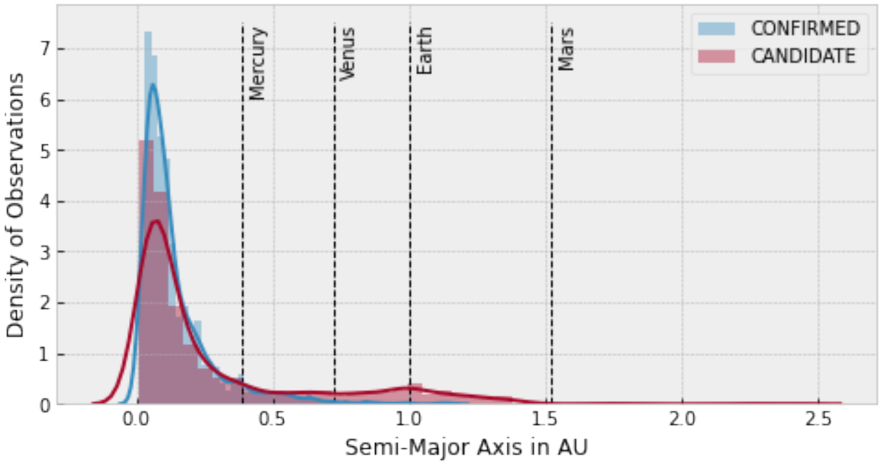
Plotting orbital period against the semi-major axis doesn't show any more extreme outliers when filtering out this one planet.



It has produced a very busy plot though! It does seem clear enough that exoplanets in general follow the same rules as planets in our solar system do: The farther from the Sun, the longer it takes to orbit the Sun. After all, the same laws of physics apply everywhere in the universe, not just in our solar system.

It's interesting that in this case binominal regression fits the data better than simple linear regression does as is visible by the upward curving line. We don't have enough data to go any further than this since Kepler's primary mission ended after four years or under 1500 days. It will have seen very few planets with orbital periods above that. They might not even have crossed in front of their parent stars once during these four years. Remember that Jupiter takes 12 years to orbit the Sun!

Looking at planets' orbits we are mainly interested how these are distributed for planets around Sun-like stars in order to make some inference about habitability. G stars like the Sun typically have a mass of 0.8 to 1.04 times the Sun's mass. Here our outlier planet gets filtered out because it's orbiting a K class star (0.45 - 0.8 $M_{\odot}$).



As the plot shows the vast majority of planets around Sun-line stars have orbits that lie even closer in than Mercury's orbit. From this it can be inferred that the vast majority of them is not habitable since their surface temperatures will be very high. Many of them will be tidally looked or very nearly so. It might be that a couple of them do have atmospheres that help distribute the heat between day and night side but this close to the star it seems unlikely. The stellar wind will have stripped away all but the densest atmosphere long ago.

As far as exoplanet candidates are concerned there's also a small but marked peak around an Earth-like orbit. This might be a bias introduced by the mission's goal, which was to look for Earth-like planets around Earth-like stars.

Exoplanets or KOIs

The dataset contains 9546 Kepler Objects of Interest out of which 2358 are confirmed exoplanets. Candidates still have to be vetted by follow-up observations while false positives can be grouped into four main categories as defined by NASA:

- _nt: A KOI whose light curve is not consistent with that of a transiting planet. This includes, but is not limited to, instrumental artifacts, non-eclipsing variable stars, and spurious (very low SNR) detections.
- _ss: A KOI that is observed to have a significant secondary event, transit shape, or out-of-eclipse variability, which indicates that the transit-like event is most likely caused by an eclipsing binary. However, self-luminous, hot Jupiters with a visible secondary eclipse will also have this flag set.
- _co: The source of the signal is from a nearby star.
- _ec: The KOI shares the same period and epoch as another object and is judged to be the result of flux contamination in the aperture or electronic crosstalk.

The first two also turn up in around one fifth of rows for confirmed exoplanets but there is no indication given whether these are simply leftovers from early vetting stats before follow-up observations.

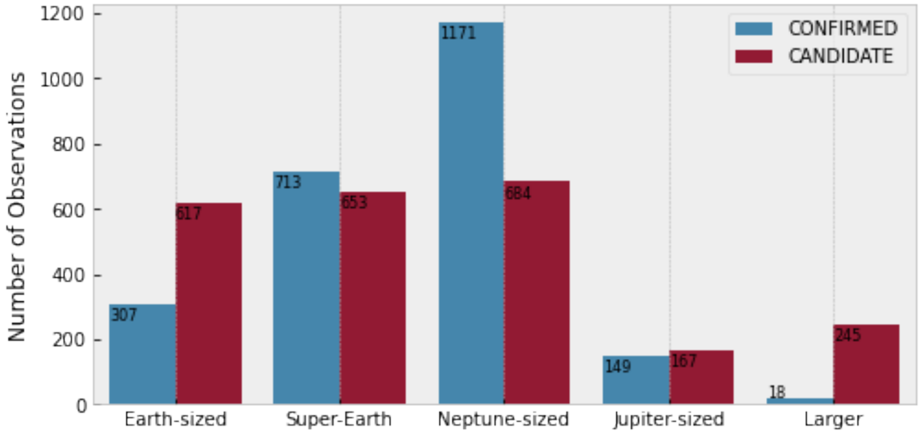
Placing Exoplanets in Categories

There are many different scales for planet sizes to be found on the internet, issued by various organisations. NASA is using the following five categories for exoplanets on their Exoplanet Exploration website, depending on their radius.

Category	In Earth's radii
Earth-sized	< 1.25 times
Super-Earth	1.25 to < 2 times
Neptune-sized	2 to <6 times
Jupiter-sized	6 to < 15 times
Larger	15 times and above

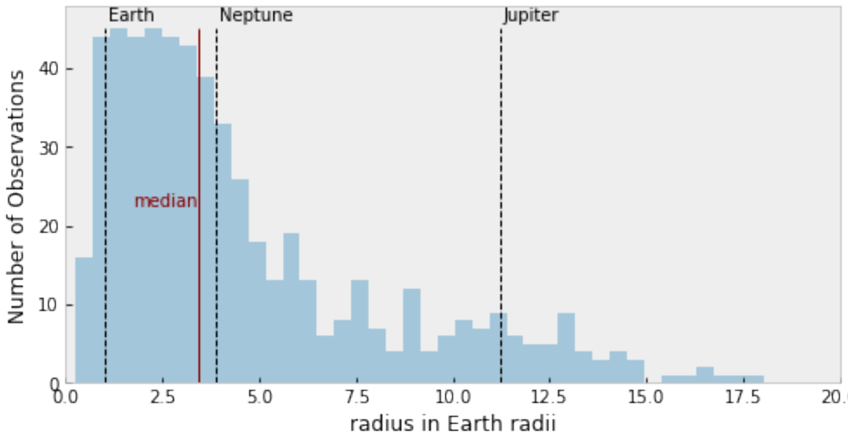
This will also be used for the analysis although the results per category are broader than in more detailed scales you might find on the internet. The radius by itself doesn't give any indication of the composition of the planets. Without knowing the masses there's no way to distinguish between Super-Earths and mini-Neptunes we often read about in the press.

The most surprising result of the Kepler mission was that there are a lot less Jupiter-size planets or larger than expected:



Before Kepler it was believed that most exoplanets were "hot Jupiters", Jupiter-sized stars that orbit very close to their host stars and are accordingly hot.

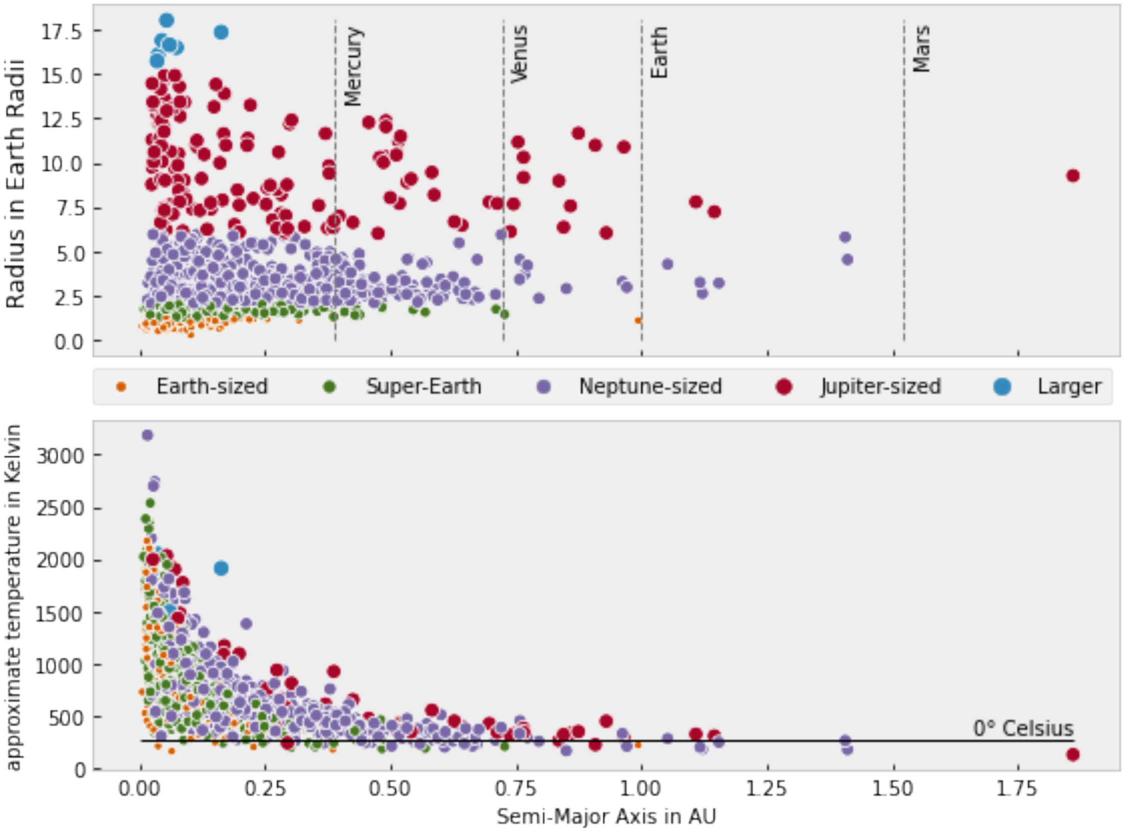
Kepler changed our view completely. It now became clear that Jupiter-sized stars and larger are actually quite rare. This becomes even clearer when looking at a histogram of planets with a radius of up to 20 Earth radii that were actually confirmed to be planets (candidates have been filtered out here). This threshold was chosen to keep the plot from becoming too compressed. There are only 10 confirmed planets with a larger radius.



50 % of confirmed exoplanets are smaller than Neptune. Of them roughly half are bigger than 2 Earth radii, having earned them the name 'Super-Earth' or 'mini-

Neptune', depending on their size, mass and density. Unfortunately for this analysis the density can't be calculated as the planets' masses are unknown. This can only be obtained by follow-up observations using the radial velocity method from which the mass can then be calculated. While a number of exoplanets have been confirmed by using this method the results were not included in the dataset.

It might be expected that Neptune-sized planets and bigger orbit quite far out from their host star but this is actually not the distribution we see in the Kepler data. In fact all observed and confirmed planets larger than 15 and smaller than 20 Earth radii orbit very close to their stars. The same applies to any planets larger than this but these have been filtered out to make the plot less crowded.

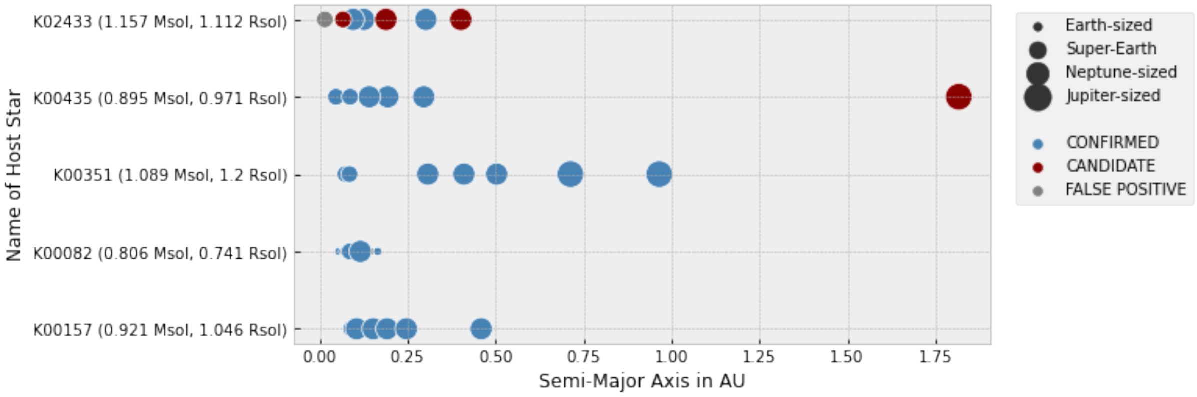


Both very large stars and Earth-sized planets orbit much farther in than even Mercury does but interestingly this does not actually mean that all of them are also very hot. The equilibrium temperature of the Earth is 255 K and average or surface temperature is 288 K. This is a very small margin of -18° to +15°C. It is not possible to extrapolate from the equilibrium temperature to the surface temperature. For example, Venus has an equilibrium temperature of approximately 260 K, but a surface temperature of 740 K due to its run-away greenhouse effect.

Provided every confirmed exoplanet has an atmosphere but none have a run-away greenhouse effect there are only 21 planets to be found in the dataset within an equilibrium temperature range between 240 and 270 K (+- 15 degrees of Earth's equilibrium temperature).

Multi-Planetary Systems

As seen above only a handful of planetary systems in the dataset have six or seven exoplanets where most or all have not turned out to be false positives. These planets are not spread out to many astronomical units; most don't reach even beyond Earth's orbit. The system around star 'K00082' is so compact that the plot barely shows all five planets! This is amazing since their stars are comparable to the Sun's radius und mass so given our own solar system we would expect something along the same lines.



Data Cleaning

Around 25% of all Kepler Objects of Interest are currently marked as candidates for exoplanets. So far only three of them have been marked as 'light curve not consistent with a planet transit' or having a significant secondary event, pointing to a binary eclipse of two stars rather than a star and a planet.

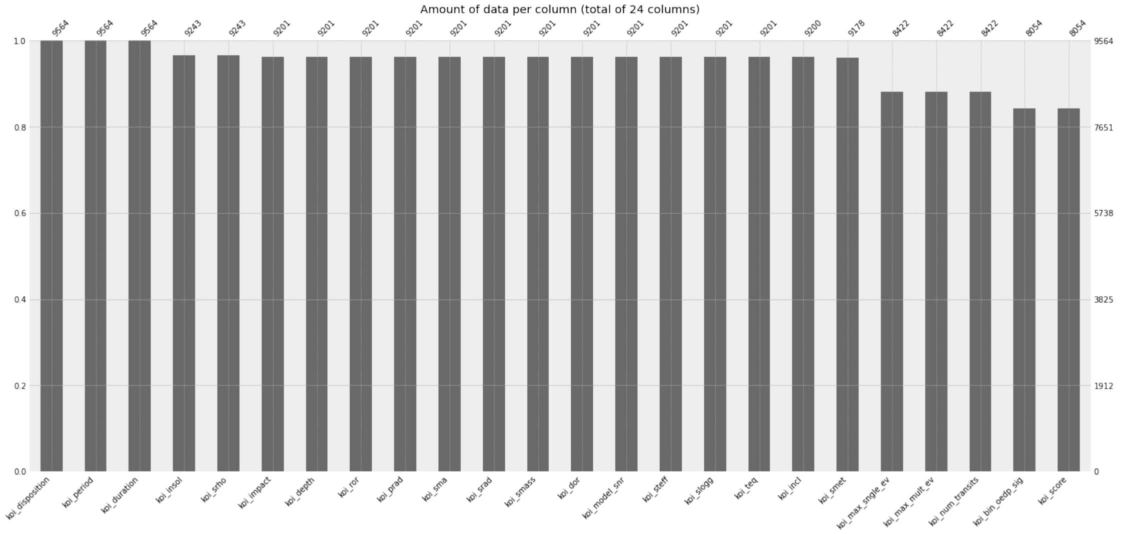
While follow-up missions are definitely needed there are enough features in the dataset to try and predict how many of them are likely to be exoplanets based on the remaining 75% that were already classified.

Reducing Number of Dataset Columns

Before any machine learning techniques can be applied, the data must be cleaned of inconsistent, inconsequential and missing values. Our analysis above did not show any (obvious) inconsistent data, just outliers that are physically possible and should be kept in the data. There are quite a few columns that don't provide any value to the analysis and machine learning algorithms for different reasons. These should be dropped from the dataframe altogether before visualizing and dealing with missing data:

- Entirely empty columns
- Columns with exactly one unique value
- Categorical data like the names of stars and confirmed exoplanets and their ids.
- Descriptive data, e.g. methods used to obtain the data plus all columns that are subordinate to them (e.g. Limb Darkening Model Names and the coefficients).
- Descriptive data pertaining to the potential host star. This includes the sky coordinates and its visual magnitude at different wavelengths. These values are not actual physical properties of the stars.
- Vetting statistics were not observed but added after analysis, about two thirds of them for false positives.

Doing so leaves us with 24 columns which is still a lot for any machine algorithm to work with. Nonetheless we can now visualize the amount of data we have for each of the columns.



That doesn't look bad. Almost every column has less than 10% of data missing.

Missing Values

Two columns do have more than 10% data missing:

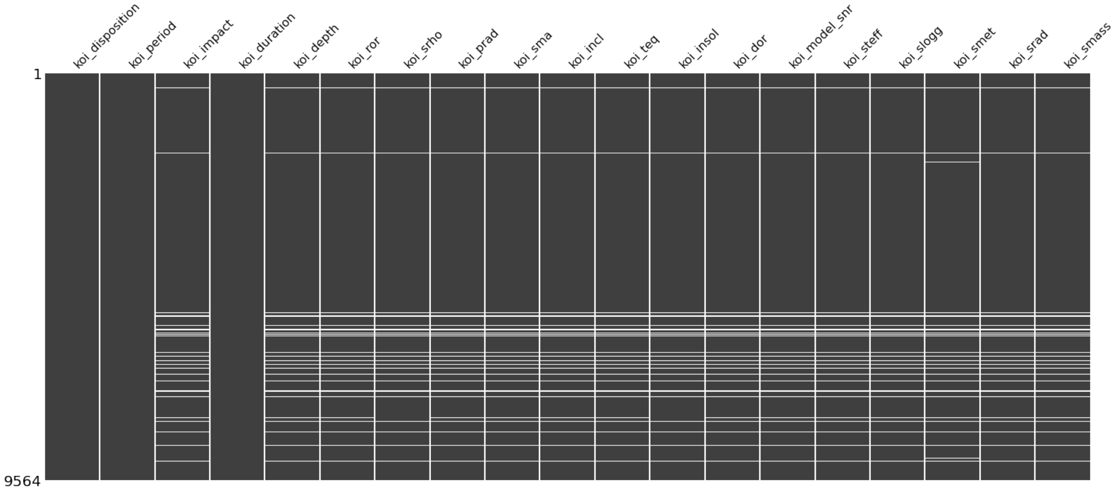
- koi_bin_oedp_sig which doesn't have a description in the data model provided by NASA! Neither is Google able to find one so the definition is probably not hidden on any other website. This can be safely dropped.
- koi_score is "a value between 0 and 1 that indicates the confidence in the KOI disposition." and is therefore part of the vetting statistics already dropped. For some reason it's not listed together with the vetting stats.

The second-least populated columns are

- koi_num_transits which describes the number of expected or only partially observed transits. Originally this data referred to candidates so it seems safe to drop the columns as well.
- koi_max_sngle_ev and koi_max_mult_ev are both calculated values of maximum single/multiple events. While they may well be important for predictions it will be difficult to infer the missing values. There are 1142 rows with missing data and 8421 unique values for each of the columns which makes it difficult if not impossible to impute values. The range is also really huge and the difference between mean and median is not negligible either. To avoid confusion for the machine learning model these columns should be dropped as well.

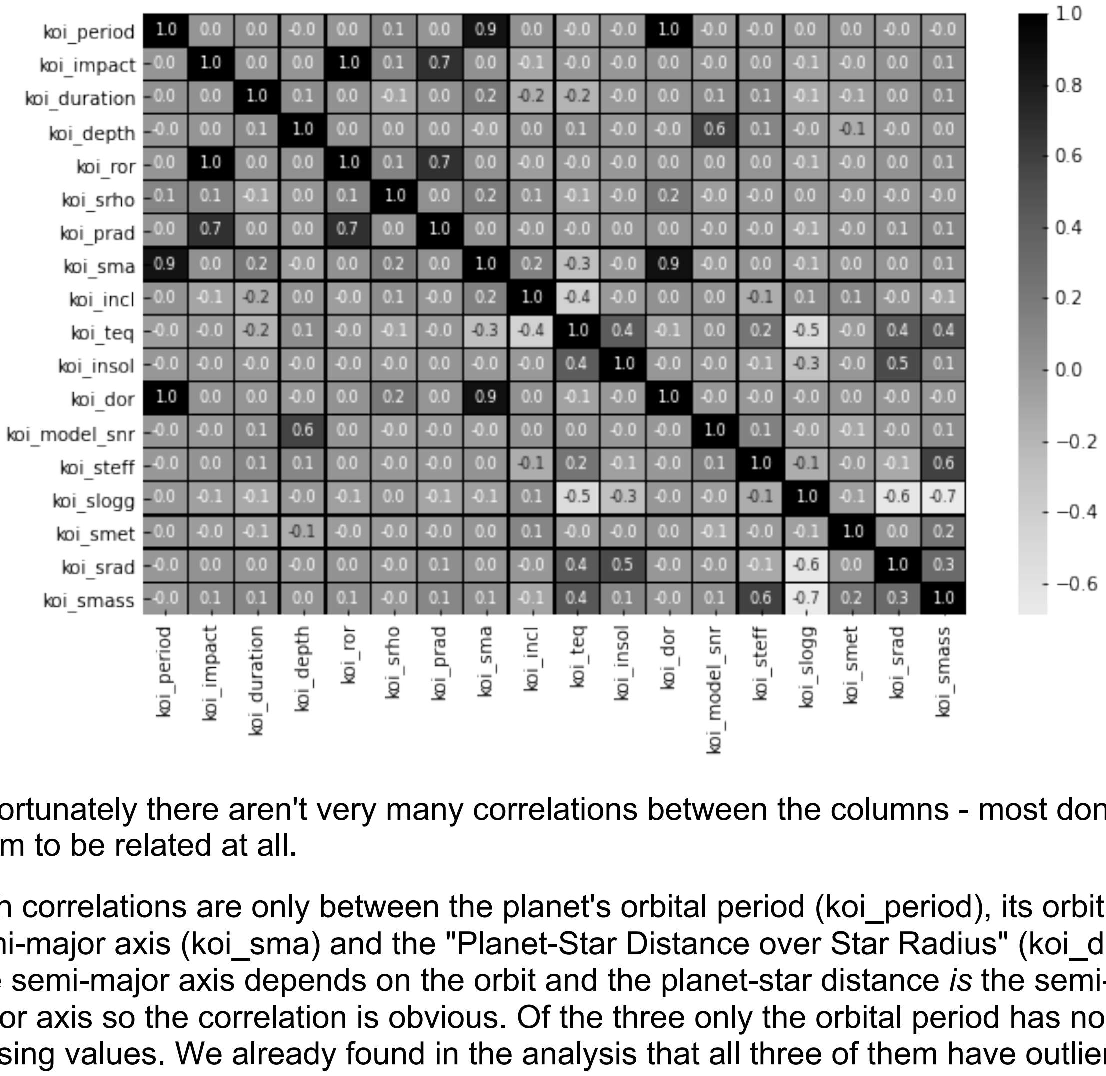
17 of the 20 columns still left have a smaller amount of missing data. It looks as if the amount of missing data for all of them is the same. This is basically confirmed by the missingness matrix (white lines are missing values) and it also shows that mostly the same rows across columns are affected.

Overall there are now 635 rows where data is missing in at least one column. This corresponds to 6.6% of the data.



We will have to have a look at the candidates data to decide whether to delete just the rows with missing data or drop the columns entirely.

Correlations in the Dataset



Unfortunately there aren't very many correlations between the columns - most don't seem to be related at all.

High correlations are only between the planet's orbital period (koi_period), its orbit's semi-major axis (koi_sma) and the "Planet-Star Distance over Star Radius" (koi_dor). The semi-major axis depends on the orbit and the planet-star distance *is* the semi-major axis so the correlation is obvious. Of the three only the orbital period has no missing values. We already found in the analysis that all three of them have outliers and a non-linear distribution so it doesn't seem safe to impute values for the missing data.

There are a handful of moderate positive and negative correlations, most notably between Impact Parameter (koi_impact), Planet-Star Radius Ratio (koi_ror) and the exoplanet's radius (koi_prad). All three of them have the same number of missing values. All three suffer from the same outliers the orbital period does.

Other than that there is a moderate negative correlation between the star's mass (koi_smass), its radius (koi_srad), gravity (koi_slogg) and the planet's equilibrium temperature. There's also a positive correlation to the planet's incoming radiation (koi_insol) and the effective temperature of the star (koi_steff). Again all of them have the same number of missing values and the same problem with outliers.

My conclusion is to accept the loss of 6.6% of the data rather than screwing the results by imputing values that are too complex to impute. Dropping the rows leaves us with 8929 Kepler Objects of Interest to use for machine learning.

Otherwise it looks very much as if the machine learning algorithms will have to figure out the underlying patters by themselves! I certainly can't see any trends that might help.

Machine Learning

Why use machine learning at all on such a comparatively small dataset? After data cleaning there are only just above 2250 candidates left and there are thousands of astronomers around. So why not simply ask the astronomy community to make observations on one star and its candidates each? Well, the most important reason is that you can't use any small telescope equipped with the right instruments because of Earth's atmospheric turbulence, also called "seeing". This turbulence is what makes stars shimmer or twinkle as we look up at them, making any tiny change either in radial velocity or luminosity very hard to detect. There are a number of high-end telescopes now that use adaptive optics to account for this. Telescopes in space, be it in orbit around Earth or sitting in a Lagrange point, don't have this problem because there's very little or no atmosphere hindering their view. However, there are also fewer of them. Observation time on both is highly sought after and there is fierce competition. The application process is quite complicated, takes a long time and even then it's more likely to have your application declined then accepted. Looking at exoplanet candidates more or less haphazardly is not going to make even the first round of applications!

Machine learning can help astronomers to determine the most interesting candidates by finding patterns in the data a human would be hard put to find. Humans are exceptionally good at finding visual patterns even when there are in fact none but we are dealing with rows and rows of dry numbers here that would be rather difficult to visualize in any meaningful way. A computer, on the other hand, is ideally suited to finding patterns in numbers. Using different techniques the number of promising candidates can be whittled down to a more manageable number. These could then be cross-referenced with catalogues from other missions, like TESS for exoplanets or GAIA for stars. * see PostScript

The feature we want to predict is a categorical value consisting of two possible categories: 'CONFIRMED' and 'FALSE POSITIVE' which could also be described as a Boolean value of True for confirmed exoplanets and false for false alarms.

Since the Kepler data is labelled accordingly for 75% of the data we only use supervised learning techniques here. There are several machine learning algorithms that support categorical or discrete values as the target feature. Due to limited resources of computational power I have chosen four of them:

- logistic regression
- k nearest neighbour
- decision tree
- random forest

All have medium to high accuracy and, except random forests, transparency. All of these will be compared and evaluated using the metrics provided by the scikit learn library.

First it is necessary to split the dataset into features and training and test data. The later comprises all rows in the dataset that are not marked as 'CANDIDATES'.

The feature to predict is in the column 'koi_disposition'. All others features will be used for prediction.

```
The labelled data contains 6919 rows.
The unknown data for which prediction will be made later contains 2258 rows.
```

The next step is to split the labelled data into training and test sets.

The training set is used, as the name suggests, to train the model and should be the same for each of the four methods.

The test set will be use to evaluate the accuracy and other metrics of the algorithms' results.

There is no hard and fast rule how the labelled data should be split into the two sets. With having quite a lot of data we can afford to use 80% for training and 20% for testing. It is imported to keep the ratio of confirmed and false positives to avoid introducing a bias towards false positives as there are twice as many as confirmed exoplanets. I'm also using a seed here to keep the results the same for every run. Otherwise the split would occur differently every time the code is run.

I'm not using techniques like k-fold or cross-validation here, partly because of limited resources, partly because I'm not really proficient in them.

```
training data:                          testing data:
FALSE POSITIVE    3649                  FALSE POSITIVE    913
CONFIRMED         1886                  CONFIRMED         471
```

Models in Action

Logistic Regression

Logistic regression is typically used for machine learning problems where the outcome is binary, i.e. one of two classes: True/False, Yes/No or - as in this case - confirmed or false positive. The correlation between the outcome and the input variables can't be linear; otherwise linear regression should be used. As we have seen in chapter Distances from the Star this is not a linear relationship although they depend on each other. We would find that the same is true for other of the independent variables.

The algorithm estimates the probability of either outcome event occurring based on a combination of independent variables and predicts the one with the higher likelihood. If we could plot an n-dimensional depiction of all data points we would see a sinuous plane dividing all data points into our two categories.

This is the result when using logistic regression on our problem:

```
Accuracy Score:
===============
0.838150289017341

PREDICTIONS COUNT
=================
overall predictions for CONFIRMED: 631
overall predictions for FALSE POSITIVE: 753

Confusion Matrix (actual-predicted):
====================================
CONFIRMED - CONFIRMED 439
CONFIRMED - FALSE POSITIVE 32
FALSE POSITIVE - CONFIRMED 192
FALSE POSITIVE - FALSE POSITIVE 721
```

This model has an accuracy score of 0.84, meaning that 84% of the test set classifications were predicted correctly.

The (prettified) confusion matrix shows that almost 200 candidates have been predicted to be confirmed exoplanets when they should have been classified as false positives.

Tweaking the solver or allowing the algorithm more iterations didn't change the results.

K Nearest Neighbour

When using K nearest neighbour the algorithm does it best to place similar data points near to each other. The outcome is the category most of the data points in the vicinity belong to. This can be used for n-nominal problems, i.e. problems with more than two outcome possibilities, because there can be any number of groups of data points in different categories as long as these don't overlap.

This algorithm works by being lazy. Instead of learning from the training data set immediately it only takes action when presented with the test set! Then it searches for similarities between one data point in the test set and all data points in the training set, clusters those and decides which ones the test data lies closes to. Then it takes the next data point and does the same until it reaches the end. Since it doesn't really learn from the training data it is likely to be less accurate when the data is diverse or small.

```
Accuracy Score:
===============
0.8179190751445087

PREDICTIONS COUNT
=================
overall predictions for CONFIRMED: 551
overall predictions for FALSE POSITIVE: 833

Confusion Matrix (actual-predicted):
====================================
CONFIRMED - CONFIRMED 385
CONFIRMED - FALSE POSITIVE 86
FALSE POSITIVE - CONFIRMED 166
FALSE POSITIVE - FALSE POSITIVE 747

n_neighbours: 65
```

This model has an even lower accuracy score of only 82% correct predictions.

Again a high ratio of candidates was predicted to be confirmed exoplanets that are, in fact, false positives. There are also almost three times the number of confirmed exoplanets that were predicted to be false positives compared to logistic regression.

For this algorithm the best number for neighbours was chosen by iterating through all values between 1 and 20 and determining which had the best accuracy score.

Decision Tree

A decision tree is basically a family tree put upside down. It starts from a root node and splits the training data into two sets according to rules it derives on its own. Each set is then split into two again unless the outcome is already uniform for all data points in one set. This is repeated until the results don't change by further splitting. You can see the schematics of the decision tree used here in the appendix.

The outcome variable for decision trees can be n-nominal as long as there are no ambiguities. Every data point has to belong to one category, and one only. Every independent input variable can be used once or several times, or not at all. This is determined by the algorithm and can't be controlled through input parameters. It is possible to keep the leaf nodes, i.e. the nodes that have a unique outcome, down to a small number to keep the algorithm from going overboard. However, it is mostly better to let the algorithm determine the number for itself to get the best possible accuracy.

```
Accuracy Score:
===============
0.9183526011560693

PREDICTIONS COUNT
=================
overall predictions for CONFIRMED: 472
overall predictions for FALSE POSITIVE: 912

Confusion Matrix (actual-predicted):
====================================
CONFIRMED - CONFIRMED 415
CONFIRMED - FALSE POSITIVE 56
FALSE POSITIVE - CONFIRMED 57
FALSE POSITIVE - FALSE POSITIVE 856

max leaf nodes: 80
```

The decision tree has been able to predict 92% of the labels correctly.

Here, the ration between overall predictions for both labels is basically the same as in the test set (1/3 CONFIRMED, 2/3 FALSE POSITIVE). This has not been the case with the other two algorithms, especially logistic regression which tended towards 50-50.

The amount of false predictions for both labels is almost the same, just below 60.

It doesn't seem possible that choosing even more leaf nodes for the tree would tweak the results to an even better outcome. As with the previous model the best number was reached by iterating through a range of values, this time ranging from just 5 to 100.

Random Forest

Decision trees have the disadvantage that they tend to overfit the data and then perform poorly on any data not seen before, in our case the candidates. To avoid this it is recommended to use a random forest which takes many decision trees and compares them to get the best possible accuracy.

It uses random subsets of the training data for this. The randomness of the subsets ensures that each data point will be used at least twice. Once all decision trees have been build the algorithm takes the majority vote on each data point across all trees and uses this as its prediction. Random Forests work best if they are allowed to build many decision trees. In Python the default value is 10 but it is better to use another algorithm to find the best number. In our case the number chosen was 85.

The main disadvantage of a random forest is that it takes a much longer time and might not get significantly better results than a single decision tree that already had good accuracy not only with the test data but also with data it hasn't seen before which is actually the case here.

```
Accuracy Score:
===============
0.9328034682080925

PREDICTIONS COUNT
=================
overall predictions for CONFIRMED: 470
overall predictions for FALSE POSITIVE: 914

Confusion Matrix (actual-predicted):
====================================
CONFIRMED - CONFIRMED 424
CONFIRMED - FALSE POSITIVE 47
FALSE POSITIVE - CONFIRMED 46
FALSE POSITIVE - FALSE POSITIVE 867

n_estimator: 85
```

This model has the best accuracy score yet at 93% correct predictions. This came at a price though. Including finding the best number for the total number of decision trees of the random forest it took over a minute to run compared to just 1.5 seconds for using just one decision tree with the best amount of leaf nodes.

The amount of false predictions for both labels is almost identical again and only a little lower than for just the one decision tree.

For finding the best number of total trees the iteration used values between 5 and 200.

Comparison and Evaluation

Visual Comparison of Predictions of Known Outcomes

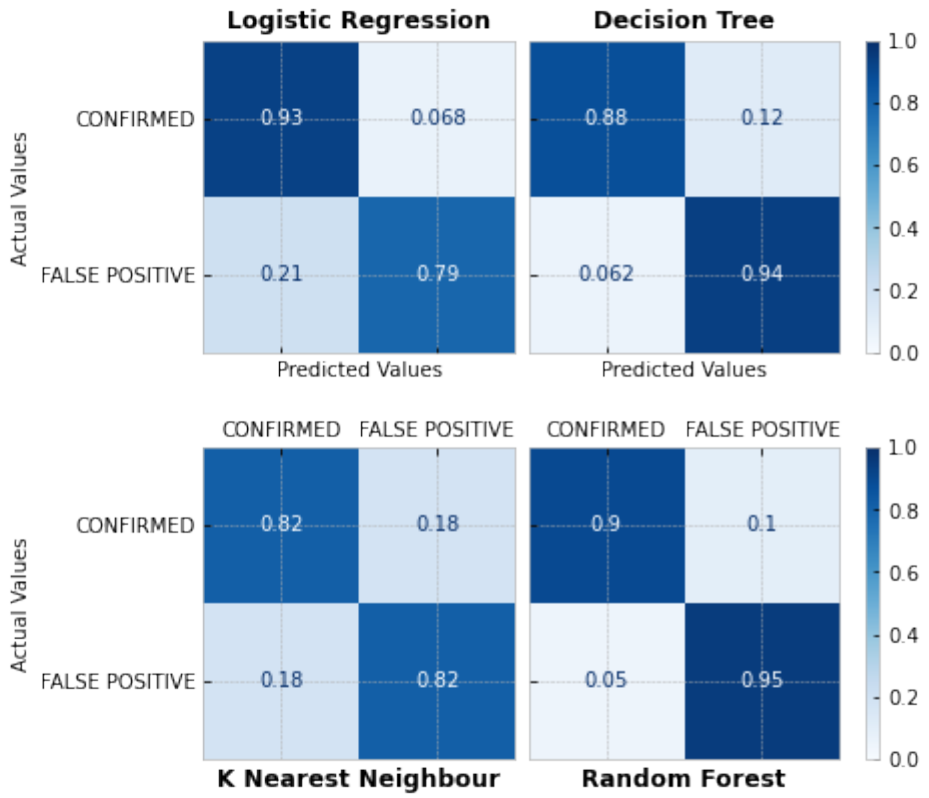
Above we were looking at the accuracy score of each model and the number of overall false predictions for each model in turn. This made direct comparison somewhat difficult.

To get a feel for the models it's much easier looking at a visual representation of how many of the predictions on the test set were actually true and how many were not.

The following confusion matrix uses ratios in each row rather than actual values. This ensures that the scale is the same for every matrix and makes comparisons much easier.

Although the accuracy score for logistic regression was rather low at 84% the predictions for confirmed exoplanets were better than for every other model used. Only 7% were predicted as false positives while the others are much higher at 9.6 to 18%.

On the other hand random forest was able to make the best predictions regarding false positives with less than 5% predicted as confirmed while both logistic regression and k nearest neighbour have an error rate of around 20%.



Model Scores

For anyone interested this is a comprehensive overview of the detailed scores of each model:

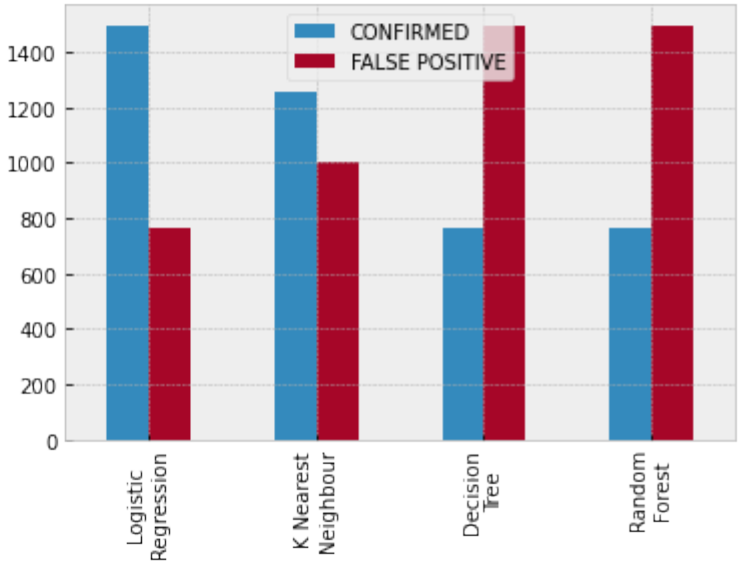
```
                                        CONFIRMED   FALSE POSITIVE
Score      Model
fscore     1. Logistic Regression        0.80          0.87
           2. K Nearest Neighbour        0.74          0.85
           3. Decision Tree              0.88          0.94
           4. Random Forest              0.90          0.95
precision  1. Logistic Regression        0.70          0.96
           2. K Nearest Neighbour        0.69          0.89
           3. Decision Tree              0.88          0.94
           4. Random Forest              0.90          0.95
recall     1. Logistic Regression        0.93          0.79
           2. K Nearest Neighbour        0.81          0.81
           3. Decision Tree              0.88          0.94
           4. Random Forest              0.90          0.95
support    1. Logistic Regression      471.00        913.00
           2. K Nearest Neighbour      471.00        913.00
```

```
3. Decision Tree            471.00        913.00
4. Random Forest            471.00        913.00
```

Model Predictions

Predicting Confirmed and False Positives

All four models have been used to make predictions about our data with unknown labels, aka candidates.



These results are rather surprising. Both logistic regression and k nearest neighbours predict the opposite ratio of the decision tree and random forest! The latter ratio corresponds to the labelled data: There are around 1/3 of confirmed exoplanets and 2/3 of false positives.

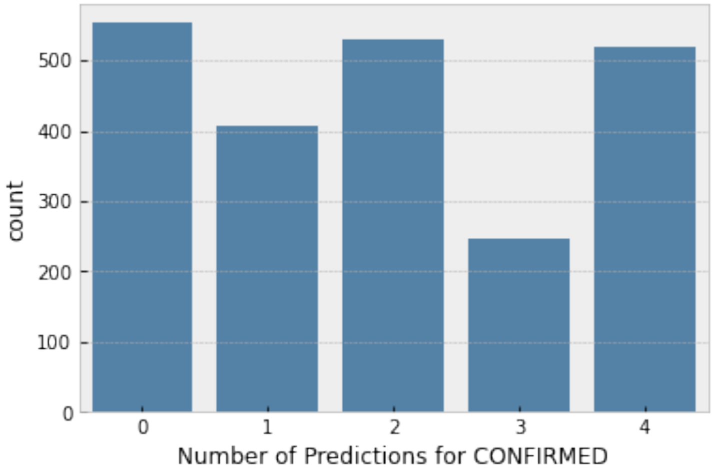
Considering that logistic regression performed best on confirmed exoplanets in the test data and random forest best on false positives this is quite contradictory. They must disagree on around 1/3 of the predictions. Checking counts they can't agree on approx. 750 confirmed and 60 false positive predictions of logistic regression.

Referring to the model scores above logistic regression has a much lower precision than recall score. While it has managed to predict most confirmed exoplanets as confirmed it also predicted 21% of false positives as confirmed. K nearest neighbour wasn't performing any better here although the spread wasn't quite as wide. Its precision score was even very slightly lower and the ratio of correctly predicted confirmed exoplanets was also lower. We can take it that neither is a good algorithm to use here!

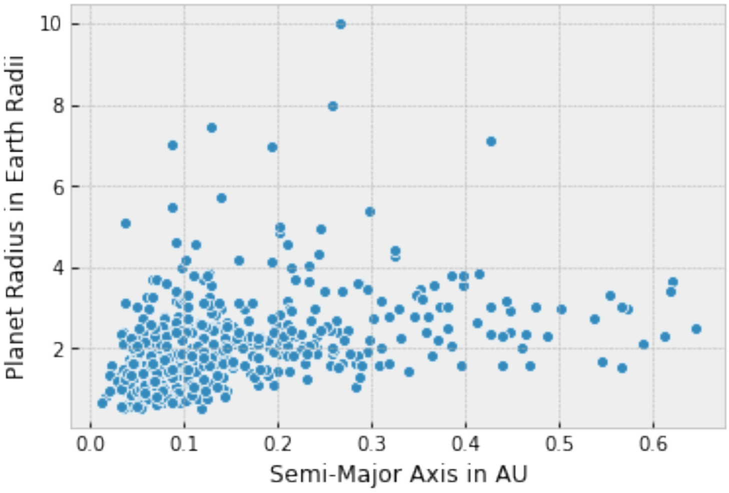
Which leaves us with the simpler decision tree and the much more complex and time-intensive random forest. Both perform almost equally well and they also agree on most predictions. Out of 2258 records in the unlabelled data, they couldn't agree on 122 predictions for confirmed exoplanets and 141 false positives (predictions by decision tree).

Cutting Down on "Confirmed" Predictions

Since the goal of machine learning here is to give astronomers an idea which of the candidates to look at next I'm not going to prefer one model over another. All models performed at least reasonably well on predicting confirmed exoplanets in the test data so I'll look next at how often each of the candidates was predicted to be an exoplanet:



This leaves us with just over 500 candidates to get a second look at, ideally using a different method if detection. Let's have a quick look at the distribution of these candidates regarding their size and orbits.



All of them orbit very near their star. Unless the star is a red dwarf this makes it unlikely that any of them are habitable so it might be hard to get telescope time for follow-up observations. Before even thinking about applying for it the host stars should be cross-referenced to observations taken with different instruments. This is beyond the scope of this work though.

Postscript: The day I finished this paper I spent part of the evening catching up on my YouTube subscriptions. One had a new video called "Discovery of a Strange New World" from The Cool Worlds Lab (link see below). They are a team of astronomers at the Department of Astronomy, Columbia University, dedicated to studying other planetary systems. In this video they have been using exactly this approach of cross-referencing for verifying exoplanet candidates. They have started out from TOIs or TESS Objects of Interest and used other star catalogues than I had in mind, from missions that measured radial velocity of stars over periods of twenty years or more, and checked whether these matched up with the transits TESS saw.

Appendix

Description of Columns of Interest

A complete overview of all columns in the dataset as well as more in depth explanations can be found here: https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

Column name	Description
koi_disposition	The category of this KOI from the Exoplanet Archive. Current values are CANDIDATE, FALSE POSITIVE, NOT DISPOSITIONED or CONFIRMED.
koi_period	The interval between consecutive planetary transits.
koi_impact	The sky-projected distance between the centre of the stellar disc and the centre of the planet disc at conjunction, normalized by the stellar radius. Corresponds to the semi-stellar axis.
koi_duration	The duration of the observed transits. Duration is measured from first contact between the planet and star until last contact.
koi_depth	The fraction of stellar flux lost at the minimum of the planetary transit.
koi_ror	The planet radius divided by the stellar radius.
koi_srho	Fitted stellar density is a direct observable from the light curve that, in the small-planet approximation, depends only on the transit's period, depth, and duration
koi_prad	The radius of the planet in Earth radii
koi_sma	Half of the long axis of the ellipse defining a planet's orbit. For a circular orbit this is the planet-star separation radius.
koi_incl	Orbital Inclination in degrees.
koi_teq	Approximation for the temperature of the planet.
koi_insol	Insolation flux is another way to give the equilibrium temperature.
koi_dor	The distance between the planet and the star at mid-transit divided by the stellar radius.
koi_model_snr	Transit depth normalized by the mean uncertainty in the flux during the transits.
koi_steff	The photospheric temperature of the star.
koi_slogg	Stellar Surface Gravity on a logarithmic scale
koi_smet	Stellar Metallicity
koi_srad	The photospheric radius of the star.
koi_smass	The mass of the star.

Astronomy Weblinks

I could not have performed the analysis and written this paper if there weren't so many sources out there to get some more in-depth information and verify that my own domain knowledge is accurate. Here are some of them for you if you'd like to dive deeper into the topic:

Gravity Assist: Exoplanet Hunting with Jon Jenkins
https://www.nasa.gov/mediacast/gravity-assist-exoplanet-hunting-with-jon-jenkins

Launch Pad Astronomy

- Lightning Talk: Discovering Extrasolar Planets with Kepler
  https://youtu.be/Q06iPD2dpXg (7 min)
- Kepler Space Telescope End of Mission
  https://youtu.be/72u03O8p2gY (approx. 8 min).
  Please be aware that the animations of exoplanets are artist's impressions.

All@Home: Exoplaneten presented by Stiftung Planetarium Berlin
https://youtu.be/e27wumRt6G4 (30 min)

The Cool Worlds Lab
http://coolworlds.astro.columbia.edu/
video on cross-referencing: https://youtu.be/RuyLMDaodlo (19 min)

Bad Astronomy Blog, written by Phil Plait
https://www.syfy.com/tags/bad-astronomy

Crash Course Astronomy with Phil Plait
https://thecrashcourse.com/courses/astronomy

eBook: Astronomy by OpenStax
https://openstax.org/details/books/astronomy
This eBook is free of charge and has over 1,000 pages in pdf-format.
It is also available for kindle and on iTunes.

Decision Tree Chart

Top --> Bottom