FEEDBACK ON MAS2008 ASSIGNMENT 3
(EXOPLANETS DATA ANALYSIS)

Marks (out of a total of 100) were assigned as follows:

- For each of the functions that you were asked to define:
  - One mark was deducted if there was no docstring.
  - Marks were deducted if the name of the function or the arguments were not as specified, or if column names in the various dataframes were not as specified. This includes trivial things like capitalisation, use of underscores instead of spaces, and so on, because this kind of issue is a common source of error when you try to make your code interact with other people's code.
- There were 4 marks for `cap_masses`. One mark was deducted if the function created a new dataframe instead of modifying the existing one. Some people used methods that converted missing masses (indicated by `NaN`) to 500; one mark was also deducted for that.
- There were 10 marks for `overall_map()`, corresponding closely to the points listed in Section 3.2 of the brief.
- There were 4 marks for explaining the meaning of the two circled regions. For these you needed to mention that the points in the red circle were mostly recorded by the Kepler space telescope in its original mission, and that the blue circle is in the direction of the centre of the galaxy, with most observations contributed by OGLE, KMTNet and MOA.
- There were 10 marks for `analyse_systems()`. Most of the marks lost were for minor failures to comply with the specification.
- There were 12 marks for `analyse_facilities()`. Most of the marks lost were for minor failures to comply with the specification.
- There were 3 marks for `show_facility()`. For full marks, either the size or colour of the markers had to reflect information about the planet itself, rather than information about the relevant facility.
- There were 3 marks for adding the `colour` column to the dataframe. A mark was deducted for calling the column `color` or `colours`.
- There were 14 marks for `show_insolation()`. Some people tried to split the dataset in two by looking for points where `pl_eqt` was less than the median, or where `pl_insol` was less than the median. These are equivalent to cutting along a vertical or horizontal line, which is clearly not correct. Some people lost a mark because they did not return the pair `(df1, df2)`.
- There were 14 marks for the first exploration question that you answered, and 13 marks each for the other two.

Most of the above points were marked automatically, and you will see messages about the marking at the top of the notebook that is returned to you. In some cases you will see a message in orange and also a more specific message in red covering essentially the same point; this is normal. Some orange messages are harmless and you will see that no marks have been deducted because of them.

The most important facilities

Here are some elements of the best answers.

- A list of the most important facilities, typically 10 of them. This should include the fact that Kepler and K2 are different missions of the same satellite, and that some facility names refer to networks of several ground-based telescopes.
- Background information about the different facilities. For ground-based telescopes, this could include the geographic location, date of construction, institutional affiliation, type of telescope and presence of special equipment like high-precision spectrographs. For space telescopes, this could include the launch date and the type of orbit around the earth or the sun.

- The number of records contributed by each facility, and the (much lower) number of distinct planets corresponding to these. (Marks were sometimes deducted for confusion between these numbers.)
- A clear and efficient summary of the detection methods used: almost all records use transit photometry, except for records from Keck and La Silla which use the radial velocity method, and OGLE, KMTNet and MOA which use microlensing. There are also a few records from Kepler that use methods other than transit photometry.
- An efficient graphical representation of the mix of methods (which should not consist of 10 or 20 separate plots).
- Some discussion of how the different methods work and their advantages and disadvantages.
- An efficient tabular or (preferably) graphical representation of the discovery dates and publication dates of the records from each facility. Ths could include some discussion of the precise meaning of these dates, and the places where details were published (from the `pl_refname` column), including a link to the NASA Exoplanet Archive.

Fields of view of different facilities

Here are some elements of the best answers.

- An efficient graphical representation of the fields of view of the most important facilities. This should not consist of 10 or 20 separate full-sized plots. The most interesting facilities for this question are Kepler, K2, CoRoT, OGLE, KMTNet and MOA; these can be shown as six subplots of a single figure, with the Milky Way shown for context, and extraneous axes etc removed to reduce clutter. Then you can have a single plot showing the facilities that observe the northern hemisphere, and another for the southern hemisphere.
- If you do have separate plots for different facilities, they should all use `set_xlim()` and `set_ylim()` to show the whole sky, otherwise they cannot be compared in a meaningful way.
- The observations for Kepler lie in a small fixed patch. The best answers include an explanation for this in terms of the design of the telescope and the mission, including the nature of its orbit around the sun (not the earth), the need to exclude sunlight from the telescope, and the desire to find planets at the same distance from the galactic centre as the earth.
- The observations for K2 lie close to the plane of the ecliptic. The best answers include an explanation of why this is so, and a plot of the ecliptic along with the K2 observations.
- The observations for CoRoT lie in a two small fixed patches, opposite each other on the sky. The best answers include an explanation for this.
- The observations for OGLE, KMTNet and MOA are in the direction of the galactic centre, except that OGLE also has a small patch in the direction of the Magellanic Clouds. The best answers include an explanation for this in terms of the requirements of the microlensing method.
- The phrase "observe the southern hemisphere" should be defined intelligently after observing the data; a hard cutoff at declination zero is not appropriate. An optimal answer would also point out that the single high-northern-hemisphere record attributed to La Silla is a mistake, and should have been attributed to the Lick Observatory instead.
- The best answers include information about the geographical location of ground-based telescopes compared with their fields of view.

Temperatures of stars and planets

Most students drew a scatter plot of `st_teff` against `pl_eqt` with a regression line, and calculated the correlation coefficient. This is OK as a start, but substantially more was needed for a high mark. The best answers were organised in terms of the formula

$$T_{\text{eq}} = T_{\text{star}}\sqrt{\frac{R}{2a}}(1-A)^{1/4}$$

which can be found on the Wikipedia page about planetary equilibrium temperature, for example. Here T_{eq}, T_{star}, R and a correspond to the columns `pl_eqt`, `st_teff`, `st_rad` and `pl_orbsmax`. The number A is the albedo of the planet, which is not given in our dataframe. One can plot T_{eq} against $T_{\text{star}}\sqrt{R/(2a)}$ and see that there is a clear relationship, which can be quantified by regression regression. Some other answers investigated the influence of pl_orbsmax in less precise ways, which was also good. Some answers included various tests of statistical significance.

Orbital period and semi-major axis

Here are some elements of the best answers.

- Scatter plots of pl_orbper against pl_orbsmax, one for the full dataset and the other for the subset of planets with pl_orbper less than 100 days. It is best to use logarithmic scales on both axes, especially for the first plot.
- Discussion of the extreme outliers, especially the planet COCONUTS-2b with its orbital period of 4×10^8 days.
- Discussion of the direct imaging method, and the reason why it finds planets with very large mass, orbital period and semi-major axis.
- Discussion of Kepler's law, giving the formula $T = 2\pi G^{-1} a^{3/2} M^{-1/2}$, with T, a and M corresponding to the columns pl_orbper, pl_orbsmax and st_mass, and G being the gravitational constant.
- Discussion of the fact that most stars in our dataframe have masses fairly similar to that of the sun, so it is not too bad to treat M as a constant in the above formula.
- Discussion of the fact that the rule $T = Ca^{3/2}$ gives a straight line on a log-log plot.
- Plots of T against $a^{3/2}$ and $a^{3/2} M^{-1/2}$, either for the full dataset or for the subset of planets with pl_orbper less than 100 days, preferably with logarithmic scales. Regression lines, correlation coefficients and tests of statistical significance.

Comparison of methods

Here are some elements of the best answers.

- A list of the 11 different methods, with the number of records and the number of distinct planets found by each.
- A shorter list of the methods contributing at least 100 records.
- A pivot table showing the averages of all numerical columns against the different methods.
- Explanations of how the methods work, with discussion of what kinds of planets they are best or worst at finding, with reference to the values in the pivot table.
- Discussion of the kinds of telescopes, satellites, spectrographs or other equipment that are typically used with each method.
- Observations (with numbers) such as:
  - Transit method planets are typically close to their stars and orbit quickly with low eccentricity.
  - Radial velocity method planets are relatively close to the solar system but further from their host stars and orbit more slowly, with quite high eccentricity.
  - Microlensing method planets are very far from the solar system and also have eccentric orbits.
  - Imaging method planets have very large masses, orbital periods and semi-major axes.
  - Pulsar timing method planets are detected by radio telescopes rather than optical telescopes, so there are no records for their visual magnitude.