Readme.pdf

For our project we needed to decide on a question to answer which needed to be interesting and captivating to the reader. We firstly navigated to Kaggle, A well-known website for popular and coherent datasets for a diverse variety of data science projects and analysis. Our initial group members were made up of computer scientists and we were interested in topics such as: Space, Cyber Security, Astronomy and other related topics. With lots of discussion and deliberation we chose to base our project on exoplanets. We had difficulty in deciding on the question, but we thought habitability would be an interesting pick.

From Kaggle we decided to use the open exoplanet catalogue version 2 located at (https://www.kaggle.com/mrisdal/open-exoplanet-catalogue/version/2) we made a Kaggle account and downloaded the .csv file and then loaded it into Jupiter notebook named "Exoplanets.ipynb

In order to reproduce what we initially created, we used the data we loaded in from the csv into the notebook and condensed the data we wanted to initially analyse into a data-frame. We wanted to look at well-known theory in physics and astronomy and relate this to habitability. We did this by loading planet orbit period; planet Jupiter mass and planet and discovery method to produce well-structured and visually pleasing graphs to clearly illustrate the key concepts, I then displayed this data frame and noticed that the way we coloured the scatter plot gave me a range of errors. This was because we couldn't colour the scatter plot in relation to strings, it only took integers as input. Therefore, I created a function which converted these string values to integer values based on the discovery method. For example: If the planet was discovered by "Transit" my function would overwrite the data frame to have the integer value 2 in place of "Transit" I did this for all Strings.

After investigating each theoretical concept and applying it to metallicity I realized that we couldn't plot most data because a lot of the planetary mass column was missing entries. Therefore, some of the most important theoretical concepts couldn't be considered. This was a big problem. We were told by - Amaury Triaud located - in the School of Physics and Astronomy that his friend maintains the dataset we were looking at or at least a very similar version to it. From this we then decided to try and find a more up to date version without the missing entries. After some digging we found a less edited version with more raw data. This is located at (https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets)

From this new data we can now begin reconstructing the notebook.

You can first open Exoplanets.ipynb and run the first few cells to produce the table which displays the data frame named "mass Period" which is later used to plot the mass against the orbital period of the planet. (both axes logged with base 10.) this should then produce a scatter plot of clear clusters when you run the next couple of cells. This is called Hot Jupiter's graph and clearly indicates groupings of planets and from these groupings you can see if a planet is rocky; a hot Jupiter; cold Jupiter etc.

when running these cells, it should read the "oec.csv" this should exactly match the data inside of the Jupyter Notebook when displayed.

"Hot Jupiter’s are a class of gas giant exoplanets that are inferred to be physically similar to Jupiter but that have very short orbital periods. The close proximity to their stars and high surface-atmosphere temperatures resulted in the moniker 'hot Jupiter’s'" - Top left cluster - hot Jupiter’s, Top right are cool Jupiter’s and bottom cluster are "super-earths"

From this information and the theory behind hot Jupiter’s we can draw basic conclusions and determine if this property contributes to habitability or not. we can see that a significant number of planets fall into a cluster of planets that are similar to that of Jupiter and probably won’t be a good candidate for a habitable planet.

We can therefore potentially rule them out of our candidates depending on whether or not they fall into each cluster. this can significantly help us determine with confidence; the habitability.

If you run the following couple of cells you will see another table displayed, this will later be plotted. Where the x axis is the planet orbit max and the y axis is the orbital period. This later demonstrates Kepler’s third law.

In astronomy, Kepler's laws of planetary motion are three scientific laws describing the motion of planets around the Sun.

* The orbit of a planet is an ellipse with the Sun at one of the two foci.

* A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.

* The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

Although Kepler’s third law doesn't help us much with habitability, in fact doesn't really help us with determining planetary habitability at all. it gives us confidence in our work and correctness so far.

If we are able to demonstrate Kepler’s third law, then we know we are on track and are following correct relationships of our data to then determine which planets could be a potential candidate for mailability.

We then plotted Eccentricity against the log(Orbital Period (days)).

This describes the "objects" distance from its sun.

this is very helpful in planetary habitability as it relates to the habitable zone commonly known as the "goldilocks zone." The "Goldilocks zone" is a zone where the distance between the "Object" (planet) and the sun is in a zone; such that the planet is "not too hot" or "not too cold" hence the common name "goldilocks zone" in relation to the book; named "goldilocks and the three bears" written by Robert Southey.

if the planet is on the "goldilocks zone" then the planet has the right temperature for liquid water to exist on the planet. If the planet has potential for liquid water, the planet then has the chance of supporting life. for life to exist the planet must contain liquid water.

So, depending on whether the planets are within the "goldilocks zone" contributes greatly to whether that planet can be habitable.

After the previous plot we then decided to plot orbital period against the Jupiter radius. Which produced similar clustering to that of the hot Jupiter's graph talked about previously in this report.

Similarly, we plotted more graphs which demonstrate different concepts in which we talk about in the notebook.

After analyzing the graphs, I wrote a basic simulation. At this point I took to Wikipedia to determine how many planets humans have discovered that have potential to be habitable. If you navigate to

(https://en.wikipedia.org/wiki/List_of_potentially_habitable_exoplanets) and scroll down to the "List of exoplanets in optimistic habitable zone" you will see that there are 30 rows. Therefore, there are 30 planets we have discovered that are in the "optimistic habitable zone." If you run the cell with my basic simulation, you will see that per 100 million planets we discover roughly 700,000 planets should be in the optimistic habitable zone (roughly.)

You should see the following description:
Above I did a basic simulation with 100 Million planets and took our calculated probability which gives us about 700 thousand planets being habitable for every 100 Million planets we discover. This may seem like a lot but ultimately, each planet is light years away; so, us humans are unlikely able to make it our new home. The optimistic goldilocks zone doesn't really conclusively say whether or not a planet is habitable thus we must consider other characteristics we have been analyzing throughout this project. For example, a large rocky planet with water may be in the habitable zone, which is good in all, however this planet may be much larger than earth and its gravitational pull may be too strong for humans to survive. There are many other factors to consider too.

From here if you open the "Clustering.ipynd" notebook and run all the cells you can reproduce the clustering algorithm we created. This will display a graph with clear coloured clustering's and also a table of our final potential candidates for habitability. These are our final results that you see where we draw conclusions with some analysis in markdown.