The Exploration of Habitable Exoplanets using Data Mining Algorithms and Data Manipulation

Hrithik Pai[1], Srideep Dornala[1], Aly Nathoo[1], Sarina Mayya[1], Winnifred Regan[1], Ojasw Upadhyay[1], Shashank Karthik Rajan[1], Araav Diwan[1], Prachi Soni[1], and Robert Downing[1]

[1]Aspiring Scholars Directed Research Program, 46307 Warm Springs Blvd, Fremont, CA 94539

The NASA Exoplanet Archive is a dataset that is an extraction from the total sets of data from the Keck, Kepler, TESS, and Gaia observations, where observations show that the observed stellar objects have been determined to possess one or more planets. It is continually updated as more and more exoplanets, or planets outside our own solar system, are discovered and documented. Our first objective was to see how many of these entries were duplicates, which would bring the total number of entries we would work with from 29,283 to 4,259. In previous research, this dataset was filtered by determining which of these exoplanets are inside their Circumstellar Habitable-Zone (CHZ), commonly defined as the range of distance from a host star such that a planet may contain liquid water, a key requirement for life as we know it. However, this calculation was done only for exoplanets with M-type host stars. Over the course of our research, we were able to expand this calculation of the CHZ to exoplanets with host stars of all spectral types. We performed more in-depth investigation of planets with G, K, and M types stars by comparing them to planets in the Planetary Habitable Labratory (PHL) exoplanet dataset to see how many similarities there are. The PHL catalog used its own set of criteria to define those planets in it as habitable. Using this method, we determined that there were 3 exoplanets with M-type host stars, 0 exoplanets with a G-type host star, and 1 exoplanet with a K-type host star.

Habitable Zone | Kth Nearest Neighbor | Duplicates | Planet Radius | Luminosity

Correspondence: ***robert.downing@asdrp.org***

Introduction

Background. The question of whether life exists beyond Earth is a question that has been asked throughout history, and right now we have the tools to begin answering it. People began with theories like the Drake equation, which is a probabilistic argument intended to estimate the number of active, communicating extraterrestrial civilizations in our galaxy. Now, after over a decade, the NASA Exoplanet Archive, which currently holds over 29000 confirmed exoplanets, has been and is still being updated as more and more exoplanets are discovered and documented. This is done by three main instruments, the Keck, Kepler, and Gaia telescopes, as well as many other telescopes, using many methods including the radial velocity (RV) technique, gravitational microlensing, most reliably transits, and most recently direct imaging (1). This source dataset will act as our input data as we sort through the various attributes of each entry. Discovering the habitability of the exoplanets outside our solar system would tell us much about the evolution and nature of other planetary and celestial bodies besides our own. It would change our perception of our place in the universe, in the same way the Copernican revolution did, giving us insight to just how expansive the world around us truly is and just how much potential it harbors.

Planet Habitability. Arguably one of the most important contributors to habitability is the planet's ability to sustain liquid water, meaning it is inside its HZ where the planetary temperature is just right to be livable and kept stable. At a pressure of 1 atmosphere, water is liquid only across the temperature range of 0 Celsius to 100 Celsius. However it is worth noting that at higher pressures, water can remain liquid over a larger range of temperatures. At lower pressures the temperature range for liquid water is smaller, and below a pressure of 0.006 atmospheres, no liquid water can exist; it is all either solid (ice) or gaseous (water vapour). Whether the planet is far away or close enough from its host star is part of what determines whether its in its HZ, and therefore whether water will be liquid. After the formation of a solar system, changes in the star's interior means that it becomes brighter and hotter. Therefore, both the inner and outer boundaries of the HZ move outwards with time. The continuous habitable zone (CHZ) is defined as the overlap between habitable zones at two different (widely-separated) times, and represents the region where water can remain liquid over timescales long enough for life to form and evolve. Previously, using the planet's semi-minor and semi-major axes (shortest and longest radii of an elliptical orbit, respectively), we were able to determine which exoplanets in the NASA Exoplanet Archive stay within their host star's CHZ and can support liquid water.

Planets also need an atmosphere that can protect the surface from harmful radiation, which is determine by various things, one of which is the parent star type. This is their classification based on their spectral characteristics, primarily their absolute magnitude, or the measure of how bright the star appears at a standard distance of 10 parsecs. The electromagnetic radiation from the star is analyzed by splitting it with a prism or diffraction grating into a spectrum exhibiting the rainbow of colors interspersed with spectral lines. Each line

indicates a particular chemical element or molecule, with the line strength indicating the abundance of that element. The range of colors also represents their surface temperatures in that atmosphere due to Wien's law (which states that the peak emission of light from an object goes as the inverse of temperature).

Stellar luminosity is estimated from the apparent bolometric magnitude of a star, which is related the apparent visual magnitude through the equation

$$M = V + BC \quad \textbf{(1)}$$

where BC is the theoretical bolometric correction (2). This is a concept we use later when calculating the CHZ for different stellar types, as the most important source of energy in determining the planet's surface conditions, or whether it is in its HZ, is the insolation from the host star. The response of a planet's atmosphere depends on the amount of incident energy (the stellar luminosity) and the spectrum of the incoming radiation (determined by the composition and st_teff (A)) (3).

The estimated mass of a planet is also critical to habitability, because planets much smaller than 1 M (mass of the earth) may lose heavier gases to space, in addition to H and He. Or, plate tectonics might not occur, which will cause excess CO_2 in the atmosphere and will not be able to create oxygen. This would make the planet uninhabitable because plate tectonics is key factor in stabilizing Earth's long-term climate (4). On the other hand, planets larger than about 10 M are considered likely to capture nebular gas during their formation and evolve into gas- or ice-giants. Planets between 2 and 10 Earth masses (so-called "super-Earths") are expected to have tectonic behavior different from Earth's (5). Plate-tectonics, slow movement of rigid lithospheric plates over the underlying mantle of Earth, have been proposed as a necessary condition for life. This knowledge of planetary mass, consequently density as well, combined with orbital characteristics and spectral type has allowed us to determine whether a planet lies inside of its CHZ and investigate more specific qualities of planets based on spectral types by comparing these values in the NASA catalog to that of the PHL dataset.

Methods

Managing Duplicates. Firstly, we decided to address an issue of data management, specifically different entries for the same planet. Although there are 29,283 entries in the NASA Exoplanet Archive Planetary Systems database as of the January 11, 2021 gamma release, there are only 4,331 unique confirmed planets. Most planets have more than one entry, each from a different publication, and thus each entry for a planet has different and sometimes contradictory data. To combat this, we decided to determine a method to remove all but one entry for each unique planet in order to use the most accurate information for our analysis. Operating under the assumption that the most recent publications and observations of planets are the most accurate due to higher definition instrumentation and improved technology, we decided to eliminate all entries except the one with the most recent publication date.

Within the NASA dataset, entries of planets with the same name (the same string value in the pl_name (A) column) appear consecutively to one another. However, the way in which these entries are ordered is seemingly random. In order to systematically remove the duplicates while preserving the most recent observation, we decided that our algorithm must first order the dataset from most to least recent, and then remove all but the first entry for the planet. We decided that the best indicator for how recent the observation would be the pl_pubdate (A) column, indicating the date of the publication of the planet parameters with an entry. Additionally, in order to prevent loss of data, we decided that the removed duplicates should be moved to an additional table to ensure that they are not completely thrown out. With these considerations in mind, the code could be developed. In order to facilitate the handling of a large dataset such as the NASA one, the best option for us was to employ Python's pandas© library and its methods to manipulate the csv files.

The first step in our deduplication process was to import the csv file containing the data as an object of the pandas data frame class. Data frames are the primary method of storing data within pandas. Data frames are easily manipulated, and data can be easily stored in rows and columns within them, making them an optimal object to use in order to manipulate the NASA dataset. We also initialized the other two dataframes that would hold the duplicates, and the deduped data.

Since the deduplication algorithm depends upon the value in the pl_pubdate column, data entries with no value or values that cannot be ordered as dates cannot be used. Thus, these entries must be dropped. Firstly, the source of the errors must be accounted for for debugging and documentation purposes. The program checks if there are any null values in the pl_pubdate and prints out how many nulls there are. Next, all entries with null values in the pl_pubdate column are dropped, and the size of the data frame is printed before and after to see the new size of the dataset.

Besides entries with null pl_pubdate values, there are some entries with bad data that cannot be used in any way as dates that can be ordered. Indexing through each value, if the length of the date is invalid or if it is formatted incorrectly, the entry with that date is dropped and relocated to the bad data frame.

It is worth noting that there are two valid formats for the dates. Some dates have only a year and month entered, and some had days as well. So first we check the length of the date as a string. If it is less than 8 characters, we concatenate a string to it, giving it a day (first day of the month). From there, we use a try and except method to see if the string can be converted to a datetime object. If it can't, it is dropped.

With all the unusable data dropped, the data can finally be sorted and deduplicated.

Listing 1. Part of the Python algorithm that removes duplicate entries

```
# Converting dates to datetime objects
```

```
data['pl_pubdate'] = pd.to_datetime(data['
    pl_pubdate'])

# Sorting the data
sortedData = data.sort_values(by=['pl_name', '
    pl_pubdate'], ascending= [1,0], inplace=False,
     na_position='last')

# Store the duplicates before removing them
duplicates = pd.DataFrame(columns=list(data.
    columns))
mask = sortedData.duplicated(subset='pl_name',
    keep='first')
df_keep = sortedData.loc[~mask]
duplicates = duplicates.append(sortedData.loc[mask
    ])

# Dropping the duplicates
deDupedData = sortedData.drop_duplicates(subset =
    'pl_name', keep='first', inplace=False,
    ignore_index=True)
```

As shown in the above code, we employed Python datetime objects to order the entries by observation publication date. Every pl_pubdate value in each entry still remaining in the data frame is converted to a datetime object. Next, the entries are sorted firstly by their name, and secondly by their value in the pl_pubdate column in ascending and descending order respectively. Thus, all observations of the same planet are kept together, and the *dropDuplicates* method can be invoked on the data to keep the first value, which has the highest pl_pubdate value and is therefore the most recent. The duplicates are first stored in the duplicates data frame, and are then dropped, completing the deduplication process. The size of the dataset before and after is compared with the size of the duplicates data frame, and finally, the deduplicated data is returned.

Out of the 29283 initial entries: 14 entries were discarded due to bad or unusable data. 268 entries were discarded due to null entries in the pl_pubdate column. 24742 entries were discarded for being duplicates.

A More Applicable Habitable Zone. From there, we checked the habitability of an exoplanet with the calculation of the CHZ, the Circumstellar Habitable Zone, around each star. The inner boundary of the CHZ is limited to where liquid water evaporates. On the other hand, the outer boundary of CHZ is created by the "maximum greenhouse limit", the point at which CO2 begins to condense into the atmosphere (6). When the two boundaries are combined, the CHZ is the region around a star where liquid water can exist. The distance at which the CHZ exists depends on the star type. For example, the circumstellar zone around an F-type star occurs farther out than that for our sun, a G type star. Likewise, the CHZ for K and M type stars are closer to the star because K and M stars are colder than G type stars. We followed the method stated by Tom E. Morris, which uses the star's luminosity to calculate the inner and outer boundaries of the CHZ for that star (7). Using his method, we eliminated exoplanets that do not fall within those bounds. Luminosity is a measure of how much energy a star emits per unit of time. Luminosity and temperature are positively correlated, so the CHZ for a star with a high luminosity will not be as close compared to a star with a low luminosity.

The first step to calculating the CHZ was treating the dataset. The star's luminosity is given in the dataset, but it is measured in units of solar luminosities. Solar luminosity is the brightness of our sun, which is $3.828\text{x}10^{28}$ watts. We converted the star's luminosity column to watts by multiplying each observation by the brightness of our sun. Morris's approach uses apparent magnitude to calculate the star's bolometric luminosity. Apparent magnitude is a measure of the star's brightness observed from earth. The values for apparent magnitude are highly inaccurate because it is an observed value, not a calculated one (8). To keep the data accurate, NASA removed the column from their dataset, and we had to use another way to calculate the star's bolometric magnitude. To determine that value, we used the formula

$$Bolometric\_Luminosity = \frac{Stellar\_Luminosity}{4\pi(Distance)^2} \quad \textbf{(2)}$$

Using the calculated bolometric luminosity, we converted that value to bolometric magnitude using the bolometric correction constant. The final step in our algorithm was finding absolute luminosity, which we did using the formula stated by Morris. Once we had that value, we found the inner and outer boundaries of the CHZ and eliminated planets that were farther than the outer boundary or closer than the inner boundary.

Listing 2. Python code calculating absolute luminosity, inner and outer boundaries, and removing planets that are not within the boundaries

```
abs_lum = 10**((bolometric_mag -4.72)/-2.5)

inner_boundary = np.sqrt(abs_lum/1.1)
outer_boundary = np.sqrt(abs_lum/0.53)

exo['pl_orbsmax'] = pd.to_numeric(exo['pl_orbsmax'
    ])
for i in range (len(exo['pl_orbsmax'])):
  axis = exo['pl_orbsmax'][i]
  if (axis < inner_boundary[i]) | (axis >
      outer_boundary[i]):
    exo = exo.drop([i])
exo = exo.reset_index(drop = True)
exo['pl_name']
```

To test the validity of our method, we applied our algorithm to the PHL (9). To keep the data consistent, we used the data from the NASA dataset, but the planets on the PHL website. PHL derived a list of 60 exoplanets that are potentially habitable by measuring their similarities with Earth. Theoretically, when we apply our CHZ algorithm to the PHL dataset, we should see that almost all of the planets are within the inner and outer bounds of the host star's CHZ.

An Extended Ruleset. Once we successfully removed duplicate records and elimated planets outside the CHZ, our third approach followed the Kth Nearest Neighbor [KNN] Machine Learning algorithm. We creating an algorithm to find exoplanets with similar attributes to the exoplanets listed in the PHL Habitable Exoplanets Catalog. This catalog consists of all potentially habitable exoplanets and is updated

regularly. We decided to split the catalog by the stellar type of the star that each planet orbits. Stellar type is vital to determining whether the specific attributes actually allow for the planet to have things such as atmospheric water vapor. For example, a planet may be far enough away to be considered as habitable for one stellar type, but would not be enough for a different stellar type.

Once we split the catalog by stellar type, we decided to compare planets based on their size. To do this, we compared exoplanets to habitable exoplanets that had a similar planetary radius.

Listing 3. Python code that calculates the best matching PHL planet for each NASA planet based on planet radius

```
for i in range (len(m_type['pl_rade'])):
  current_dist = np.abs(phl_m_type['pl_rade'][0]-
      m_type['pl_rade'][i])
    if current_dist < 20:
      closest_distance_away = current_dist
      index = 0

  current_dist = np.abs(phl_m_type['pl_rade'][1]-
      m_type['pl_rade'][i])
    if current_dist < 20:
      if closest_distance_away > current_dist:
        closest_distance_away = current_dist
        index = 1

  current_dist = np.abs(phl_m_type['pl_rade'][2]-
      m_type['pl_rade'][i])
    if current_dist < 20:
      if closest_distance_away > current_dist:
        closest_distance_away = current_dist
        index = 2
```

Comparing planets based on their size allows us to have higher confidence that the attributes of the habitable exoplanet apply to other exoplanets. For example, a planet with a radius of 0.5 earth radii might be habitable, but probably won't have similar attributes to another habitable planet with a radius of 5 earth radii.

To do this, we created different programs for each stellar type. In each program, we created two data frames using Python's Pandas Library. These data frames either contained potentially habitable planets from the PHL Habitable Exoplanets catalog, or the deduplicated exoplanets from the NASA Planetary System table. Some planet records contain missing values and attributes, so we needed to either calculate the missing values or drop them entirely.

Two of the attributes we are unable to calculate, pl_orbeccen (A) and pl_rade (A), so we had to drop all planetary records without these attributes.

We were however able to calculate two other attributes. These are pl_dens (A) and pl_orbper (A). For planetary density, we were able to calculate it using planetary radius and planetary mass. For planetary orbital periods, we were able to calculate it using two different methods. One was using a planet's orbital semi-major axis. Another used the ratio of stellar radius to orbital semi-major axis, and stellar radius. If we were unable to calculate density and orbital period for a planet using these methods, we simply dropped the record from the data frames.

Once we had all the necessary variables calculated, we looped through each planet remaining in the data frame, and compared it to the PHL planet that was most similar in size. We found the planet that was most similar in size by comparing both the planets' planetary radii.

Our criteria to be considered potentially habitable are as follows:

A planet must have a density value between 4 and 7. This makes sure that a planet is rocky, which is as of now, the only type of planet proven to be able to support life. A planet must not have a controversial flag, which is a flag that questions whether a planet actually exists. A planet must be discovered using the transit discovery method, which is one the most reliable discovery methods. A planet must have an orbital eccentricity less than 0.2. The closer a planet is to 0, the more circular its orbit is. If a planet's orbit is not circular, it could mean that it could be outside of its host star's habitable zone for a part of its orbit. Finally, a planet must have an orbital period that is less than 50 earth days from the planetary orbit of the planet to which we are comparing it.

If a planet passes all of these criteria, we append the planet to a new DataFrame containing all the habitable exoplanets we found using the method.

Results

Our deduplication process gave a final output in the form of a new DataFrame. While the size of the most recent version of the NASA Exoplanet Archive is 29367, this number is changing as more exoplanets are discovered. Still, after applying the deduplication method to this set, the result gives 4302 values. This dataframe now has only the most recent unique exoplanets.

As a result of the CHZ algorithm, about half of the dataset was eliminated. We determined that 1954 exoplanets lie in the CHZ zone of their host star. To further analyze our data, we found the number of exoplanets from the NASA dataset that were within the bounds, farther than the outer bound, and closer than the inner bound (Figure 1).

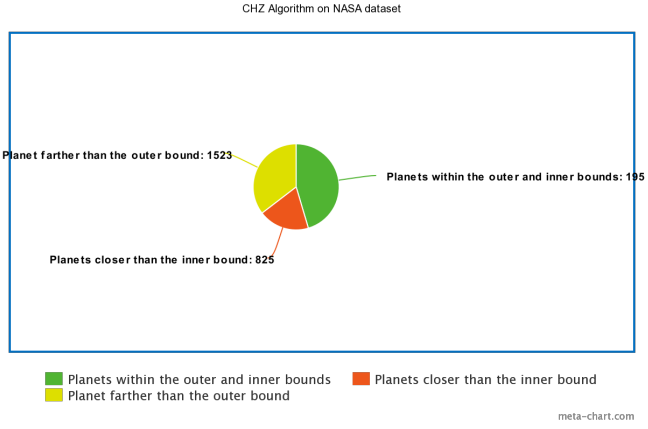

Fig. 1. Pie chart showing CHZ algorithm results on the NASA dataset

As mentioned earlier, we also tested our algorithm on the PHL database to see how accurate our algorithm was. When we ran our code on the PHL dataset, we found that our code

worked for about 88% of the planets (Figure 2). We see that 53 planets were within the inner and outer bounds, seven planets were not.

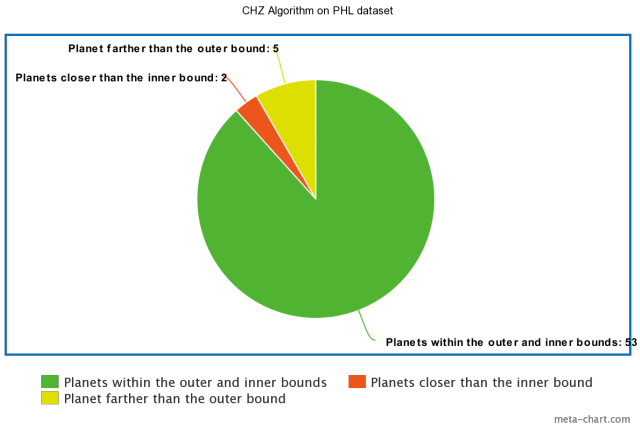

Fig. 2. Pie chart showing CHZ algorithm results on the PHL dataset

As for our Kth Nearest Neighbor algorithm, in which we compare the PHL dataset to the NASA dataset, we ended up with 4 total planets that were considered habitable. When we followed the process listed in the Extended Ruleset section, we ended up with one remaining K-type planet that was considered potentially habitable. This planet is GJ 143 b, which has an orbital period of 35.6 days, and planet radius of 2.61 Earth radii. We ended up with zero G-type planets that were considered potentially habitable. Finally, we ended up with three M-type planets that are considered potentially habitable. These planets are L 98-59 c (Orbital Period: 3.6 days, Planet Radius: 1.35 Earth radii), LTT 3780 b (Orbital Period: 0.8 days, Planet Radius: 1.33 Earth radii), and TOI-1266 b (Orbital Period: 10.9 days, Planet Radius: 2.37 Earth radii).

Discussion

The first approach we took with this project was to remove the duplicate planet entries from the database. It was considered a duplicate if, naturally, the name was repeated, and we decided which ones to keep based on the publication date. However, when sorting based on this criteria, there is the possibility that a duplicate entry had both the same name and the same publication date, especially since many entries were specified only by year and month, not day.

As a result, it is possible some entries were chosen by random selection. In the future, this process can be fine-tuned to sort not just by publication date, but also by whether it has a controversial flag and the credibility of its discover method, labeled by pl_controvflag (A) and discorverymethod (A), respectively. Originally, we decided that the most recent recording of an exoplanet would mean it is most likely the most credible. This remains true, but it would be more definite if we looked directly at any controversy it has and checked if any duplicate entries of the same planet were discovered using a more trustworthy detection method.

For the algorithm that calculates the Circumstellar Habitable Zone for each star and determines which planets lie within the bounds, our algorithm was mostly accurate, which was confirmed by the PHL dataset. One addition attribute we can consider in the future is the Greenhouse Gas effect. As a result of the atmosphere, gases get trapped inside the atmosphere, making the planet hotter than what was calculated. Since the planet is hotter than expected, the CHZ for that star can be farther out. If this is accounted for, we should expect more planets to fall within the CHZ zone, and increase our accuracy.

As for the third step of this project, we compared exoplanets of the NASA Exoplanet Archive to that of the PHL dataset. The main purpose of this was to see how many exoplanets from the NASA Exoplanet Archive fit the criteria of habitability set forth by the PHL dataset. However many G-type stars were eliminated in the process. We suspect that it was caused by missing values. Since our algorithm identifies any planets that have a missing value not habitable, this could be the primary reason for the low number of G-type habitable planets. One technique we can use to avoid this would be to calculate the statistical mean as a method of replacing missing values in a way that has a negligible effect on the overall distribution of values. By doing this, the accuracy of our algorithm will increase since we are not eliminating as many planets that have missing values.

Conclusion

Exoplanets, and generally celestial bodies beyond Earth, have long been a source of interest for research. One of the most commonly cited sources is one of the original papers to define the HZ (4). As I already discussed in my previously, it was a fundamental starting point for deducing whether or not liquid water can be supported on the planet's surface. However, there is far more that could influence habitability than just a range of distance from a planet's host star, which has been the focus for this project. After removing the duplicates from the dataset that we were working with, we calculated how many were within the CHZ of their host star, then finally comparing the orbital period as well as the density of the planets in our dataset to that of the PHL dataset. However, this only scratches the surface of applying what we know about habitability to the NASA Exoplanet Archive. For example, a later study that used general-circulation climate models (GCM) and energy-balance climate models (EBM) found that long-term climate stability, and as a result the potential to support liquid water, was dependent primarily on the average stellar flux received over an entire orbit, not the length of the time spent within the HZ (10). This methodology, which came to be know as the mean flux approximation, was then given limits for planets with very eccentric (elliptical) orbits (11). While we currently have been able to apply the well-known HZ calculations to our dataset for all stellar types, this is a future avenue that can be investigated to further narrow down which planets may have a suitable climate and temperature for life as we know it.

A paper by Tyler D. Robinson is a perfect example of how the published literature has set up the foundation for a project like ours. After an in depth discussion of observational tech-

niques and approaches to constraining habitability, it concluded that transit spectroscopy, secondary eclipse observations, high contrast imaging, and overall reflected light observations all have the potential to reveal key planetary properties related to habitability (12). Given that our data does contain many different photometric measurements for the planet entries, such research would be a very good place to continue our data mining. Not that we have taken into account many geometric attributes, HZ fundamentals, as well as some data mining, the mean flux approximation methodology or photometric measurements are just a few of the various viable paths to continue down.

Association learning is yet another tool this data mining process can make good use of. After the first detection of water in the atmosphere of an HZ planet, K2-18b, the characteristics of this planet were analyzed and converted into a rule that was applied to the NASA Exoplanet Archive (13). However this was just one planet. While K2-18b is the only planet orbiting a star outside the solar system (or "exoplanet") known to have both water and temperatures that could support life, according to NASA, it isn't the only planet with special characteristics worth investigating. For example, a study that explored the rotation history of the planet GJ 581d that is assumed to have composition similar to that of the terrestrial planets of the Solar system determined that the planet is indeed potentially habitable due to high improbability that the planet could have reached 1:1 resonance (14). Such a planet, whose system is in the NASA Exoplanet Archive, could be a source of information regarding how many other planets hold a similar degree of habitability based on their planetary characteristics.

A key component in this research was our ability to compare values for the exoplanet entries either to another dataset, or to use it for our own calculations. However, the use of the desired attribute wasn't always possible, because these values were often missing. Planetary characteristics such as eccentricity, semi minor axis, or radius were often not filled in, and could therefore be a way to improve and expand the algorithm we built. If we figure out a way to calculate eccentricity, particularly from values that we have more of (so not semi-minor axis, for example) we would have more planets to work with, and potentially more in the output.

Our original thought process was to utilize Keplerian Mechanics to determine some of these missing values. We were attempting to calculate Mean anomaly which would allow us to find Eccentricity anomaly and then True anomaly. These values could be used in the formula

$$R = \frac{a(1-e*e)}{1+e*cos(x)} \quad \textbf{(3)}$$

where we would be able to find missing eccentricity values for roughly 600 planets. Unfortunately, there always seemed to be one variable missing from the dataset, indispensable for our calculation process. We hope to discover an alternative method to fill this gap in the near future.

ACKNOWLEDGEMENTS
We would like to thank Professor Robert Downing for his valuable aid and support throughout our research. In addition, we would like to thank NASA and the California Institute of Technology for access to the exoplanet dataset. The authors gratefully acknowledge the Aspiring Scholars Directed Research Program and sponsors for providing us with the opportunity to conduct this research.

Bibliography

1. James Kasting, W Traub, A Roberge, A Leger, A Schwartz, A Wooten, A Vosteen, A Lo, A Brack, A Tanner, et al. Exoplanet characterization and the search for life. *arXiv preprint arXiv:0911.2936*, 2009.
2. Bradley W Carroll and Dale A Ostlie. *An Introduction to Modern Astrophysics*. Addison-Wesley, New York, 1996.
3. Patrick A Young. Stellar composition, structure, and evolution: Impact on habitability. In *Handbook of Exoplanets*, pages 2959–2980. Springer International Publishing, 2018.
4. James F Kasting, Daniel P Whitmire, and Ray T Reynolds. Habitable zones around main sequence stars. *Icarus*, 101(1):108–128, 1993.
5. Diana Valencia, Richard J O'connell, and Dimitar D Sasselov. Inevitability of plate tectonics on super-earths. *The Astrophysical Journal Letters*, 670(1):L45, 2007.
6. Jacob Haqq-Misra, Ravi Kumar Kopparapu, Natasha E Batalha, Chester E Harman, and James F Kasting. Limit cycles can reduce the width of the habitable zone. *The Astrophysical Journal*, 827(2):120, 2016.
7. Tom E. Morris. Calculating the habitable zone, Oct 2010.
8. SE Schröder, L Kaper, HJGLM Lamers, and AGA Brown. On the hipparcos parallaxes of o stars. *Astronomy & Astrophysics*, 428(1):149–157, 2004.
9. The habitable exoplanets catalog.
10. Darren M Williams and David Pollard. Earth-like worlds on eccentric orbits: excursions beyond the habitable zone. *International Journal of Astrobiology*, 1(1):61, 2002.
11. Emeline Bolmont, Anne-Sophie Libert, Jeremy Leconte, and Franck Selsis. Habitability of planets on eccentric orbits: Limits of the mean flux approximation. *Astronomy & Astrophysics*, 591:A106, 2016.
12. Tyler D Robinson. Characterizing exoplanet habitability. *arXiv preprint arXiv:1701.05205*, 2017.
13. Alan Chen, Hrithik Pai, Aaryan Rustagi, et al. An application of data mining to search for potentially habitable exoplanets. *ASDRP Communications*, 3:8–12, 2020.
14. Valeri V Makarov, Ciprian Berghea, and Michael Efroimsky. Dynamical evolution and spin–orbit resonances of potentially habitable exoplanets: the case of gj 581d. *The Astrophysical Journal*, 761(2):83, 2012.

Supplementary Note A: Attribute Names and Definitions

1. st_teff: Stellar Effective Temperature (temperature of the star as modeled by a black body emitting the same total amount of electromagnetic radiation)
2. pl_name: Planet Name (planet name most commonly used in the literature)
3. pl_pubdate: Discovery Publication Date (date of the publication of the given planet parameter set)
4. pl_orbeccen: Eccentricity (amount by which the orbit of the planet deviates from a perfect circle)
5. pl_rade: Stellar Radius (length of a line segment from the center of the star to its surface, measured in units of radius of the Sun)
6. pl_dens: Planet Density (amount of mass per unit of volume of the planet)
7. pl_orbper: Orbital Period [days] (time the planet takes to make a complete orbit around the host star or system)
8. pl_controvflag: Controversial Flag (flag indicating whether the confirmation status of a planet has been questioned in the published literature)
9. discoverymethod: Discovery Method (method by which the planet was first identified)
10.