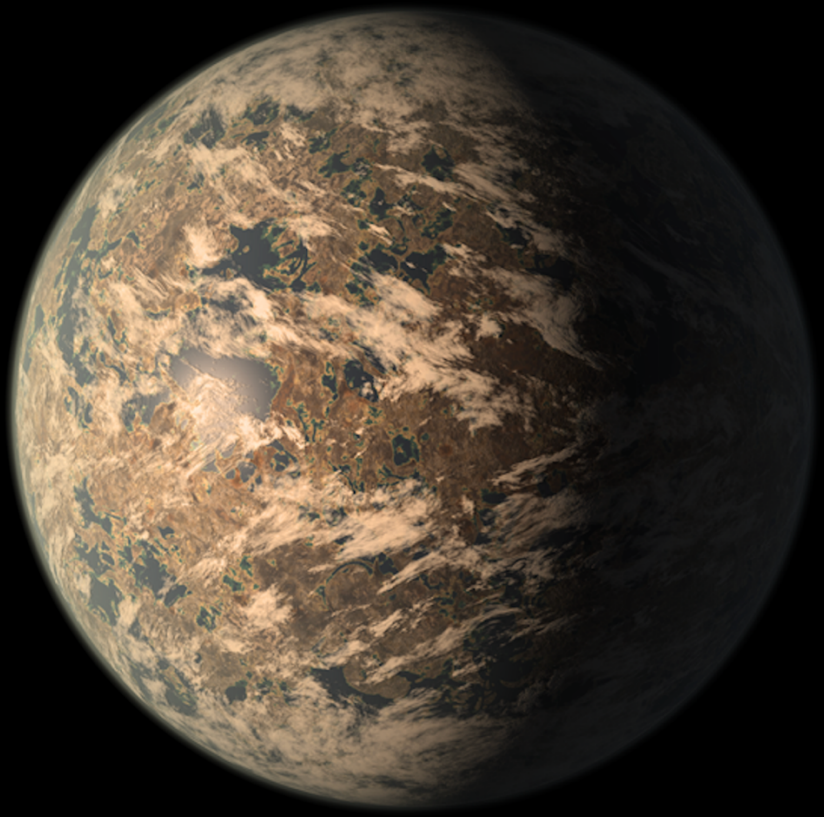
EXOPLANET CLUSTERING WITH K-MEANS

DEPLOYING THE UNSUPERVISED LEARNING ALGORITHM TO FIND EARTH-II CANDIDATES



Shane J. Robinson

Northwestern University
633 Clark Street, Evanston, IL 60208

Abstract

One of the primary challenges in the search for extraterrestrial signals is knowing where to look in the first place. We cannot scan the entire celestial sphere – we must choose where to aim our efforts. As such, utilizing a tool with which we can narrow our search field is essential to focusing the search. Enter, machine learning. What's proposed in these pages is the deployment of unsupervised learning to group clusters of like exoplanets, finding Earth's place within those clusters, and in so doing, determining which exoplanets are most likely to be promising Earth II candidates. By extension, given our sole data point, the most-like-Earth candidates could be considered the most likely to produce those elusive extraterrestrial signals.

1 Introduction

With continuously updated data on thousands of confirmed exoplanets and their star systems, we now have the requisite information to begin building a ranking or grading system for exoplanets based on their potential similarity to the Earth. This system could then be used to order the search for Earth-like planets. Professional and amateur astronomers could better coordinate efforts, having a more optimal guide for where to search for signals and where to devote further study.

The primary objective here is to create a prioritization framework for searching for Earth II and extraterrestrial signals by clustering exoplanets based on their potential habitability and similarity to Earth. Leveraging a comprehensive dataset of exoplanetary properties, a three-step methodology will be employed. The hope is that the results demonstrate the efficacy of utilizing an unsupervised clustering model - specifically the k-means algorithm - to automatically classify exoplanets, establishing an extensible framework to be built upon with more/better parameters, and providing valuable insights for prioritizing regions in the search for extraterrestrial signals by professional and amateur astronomers, and potential habitable worlds during astronomical research and space exploration endeavors.

2 Related Work

According to a team of researchers at NASA's Goddard Institute for Space Studies (GISS), the potential habitability of exoplanets is often usually determined by categorizing them based on their nominal equilibrium temperature, assuming they have an Earth-like planetary albedo (Del Genio et al 2017). This means considering how much sunlight a planet reflects into space, like Earth's reflective properties; and calculating the equilibrium temperature involves the luminosity of the host star (which determines the energy received by the planet) and the planet's distance from it (which influences how this energy is distributed). By assuming an Earth-like albedo, scientists can estimate whether conditions on an exoplanet might fall within the range of suitability for liquid water and, thus, potentially habitable environments. The team at GISS contend, however, that albedo and equilibrium temperature are insufficient indicators of habitability because exoplanet albedos can be very different and surface temperature exceeds equilibrium temperature by the amount of the atmosphere's greenhouse effect. They used a computer program called the GISS ROCKE-3D GCM to determine if more useful predictions for Earth-like planets than equilibrium temperature with information that will be available for every

known exoplanet. Equilibrium temperature, also called effective temperature, is an estimate of the temperature a planet would have if it were a perfect black body (an object that absorbs all incoming radiation and emits radiation according to its temperature) and in thermal equilibrium with the energy it receives from its parent star. Their use of the GISS ROCKE-3D GCM coupled exoplanets' atmospheres to a dynamic ocean, which is necessary for a realistic assessment of the habitability of a planet with oceans. As a large ensemble of simulations that vary every possible external parameter to represent all possible habitable planets doesn't yet exist, they instead used 29 simulations already conducted with their ROCKE-3D GCM / dynamic ocean coupling. Some of their predictions were for popularly discussed exoplanets like Kepler-186 f – unlikely to be habitable, and TRAPPIST-1 e – habitable.

With the vast amount of data now available for celestial objects from a multitude of satellite imaging (across the spectrum of light), much work has been done to apply machine learning methods to astronomical data for exoplanet detection and classification, spectroscopy, and morphological reproduction of star systems and galaxies. In a 2010 submission to the Royal Astronomical Society, Manda Banerji and team trained an artificial neural network on a subset of previously classified objects, then tested whether their machine learning algorithm reproduced those classifications for the test set (Banerji et al 2010). They found that the neural network's success in matching the human classifications greatly depended on the set of chosen input parameters for the algorithm; and after fine-tuning the parameters, the neural network, it was able to reproduce the human classification over 90% of the time.

A white paper published by IEEE introduced astroML, machine learning for astrophysics (VanderPlas et al 2012). The authors submit that in the coming years, the volume of astronomical data will reach the petabytes level, and that astronomy and physics students haven't traditionally been trained to handle big data. Their astroML initiative bases itself in Python's Scikit-learn library and builds a compendium of machine learning tools to address the needs of future astronomy and physics students and researchers. It appears that in the years since their paper and the wave of ever more celestial object data, students and researchers have had the requisite training to incorporate the wrangling of big data. Perhaps their astroML was a part of, or influenced, that training.

Transit spectroscopy has been a powerful tool for discerning the chemical compositions of the atmospheres of extrasolar planets. Using data from the Kepler catalog, a team of researchers trained artificial neural networks (ANNs), with carefully selected features accounting for the differing sensitivity of the Transiting Exoplanet Survey Satellite (TESS) and were able to predict the most likely short-period transiting (a relatively brief transit period – when the planet passes in front of the star from our perspective) exoplanets to be accompanied by additional transiting objects (Lam and Kipping, 2017). They predict that their trained ANN would have discovery of additional transiter yield improved by a factor of two, and that it would enable follow-up strategy for surveys to target additional planets, thereby improving the yield of habitable exoplanet candidates.

3 Data

The data used for the clustering models comes from the NASA Exoplanet Archive, which is an online astronomical exoplanet and stellar catalog and data service that collates and cross-correlates astronomical data and information on exoplanets and their host stars. It also provides tools to work with the data. Within the dataset are stellar parameters (positions, magnitudes, temperatures, etc.), exoplanet parameters (masses, orbital periods and eccentricities, etc.), and discovery accreditation and information. The archive was created to aid in the search for and characterization of exoplanets and their host stars; it is a collaborative effort between NASA and Caltech, which hosts the dataset for public availability on its web site: https://exoplanetarchive.ipac.caltech.edu. Within the downloadable dataset there are a total of 34,892 records and 40 columns (features/parameters). The data was cleaned to account for mass nulls; administrative details were dropped, such as discovery year, date of last update, and release date; and the data was filtered down to include only target parameters, chosen based on domain knowledge, that prioritize planetary characteristics most likely to predict similarity to Earth: planet radius and mass (for the strength of its gravity), orbital eccentricity (an indicator of its orbital stability), and equilibrium temperature.

4 Methods

After data preparation and feature engineering, including filtering the dataset for chosen parameters and adding Earth's relevant characteristics to the set, a process of normalization (also called standardization) was followed to ensure the factors used to make predictions were of the same scale. This is important because it helps the algorithm learn more effectively and prevents some features from having too much influence due to their extremely large or small size.

Following normalization came the principal component analysis (PCA). This mathematical technique used for dimensionality reduction is typically most valuable when dealing with datasets containing many features or variables. While the cleaning and engineering process reduced the dataset to four target features, PCA was still an important feature to employ as it allowed for noise reduction, easier visualization, model simplification, and more feature engineering: new features (or, principal components) were created that describe the dataset, and those components were used for plotting the k-means clusters. These created components can sometimes reveal hidden patterns in the data.

After the PCA, scree and silhouette analyses were conducted. The scree plot is a technique used to determine the number of principal components to retain. It is a method for deciding how many components capture the most important information in the data. The silhouette analysis is a technique used to evaluate the quality of clustering in unsupervised machine learning. It helps determine how well-defined and distinct the clusters are within the data, the optimal number of clusters the algorithm should create, and visualizes the cluster quality. Once the scree and silhouette analyses indicated the appropriate number of principal components and number of clusters, the k-means algorithm was deployed.

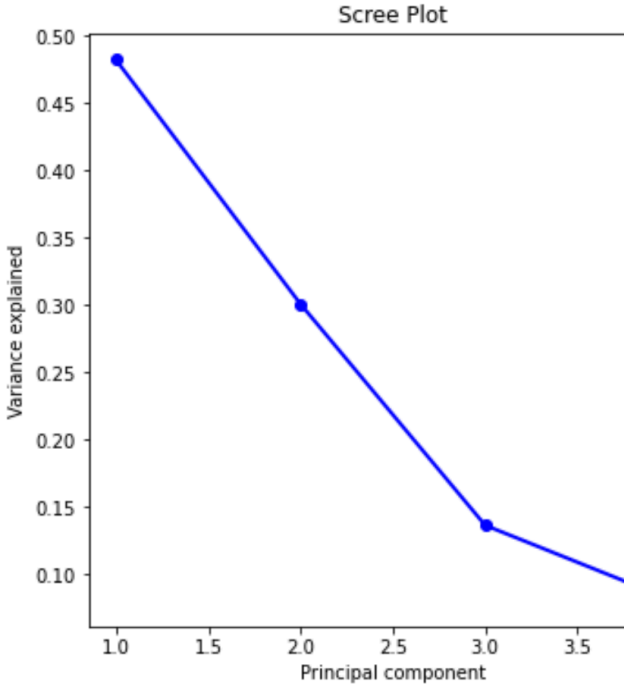
K-means is a popular clustering algorithm used in machine learning and data mining to partition data into K clusters based on similarity or distance. While it is often referred to as "k-means

clustering," it does involve an optimization process. The k-means optimization algorithm works as follows: 1) initialization: randomly select K data points from the dataset as initial cluster centroids, 2) assignment: assign each data point to the nearest centroid, creating K clusters, 3) update: recalculate the centroids of the K clusters based on the mean of the data points within each cluster, and 4) iteration: repeat steps 2 and 3 until convergence or a maximum number of iterations is reached.

The primary goal of the k-means optimization process is to minimize the sum of squared distances (also known as the "within-cluster sum of squares" or "inertia") between each data point and its corresponding centroid. This optimization process aims to find the best placement of cluster centroids, which in turn minimizes the overall variance within each cluster.

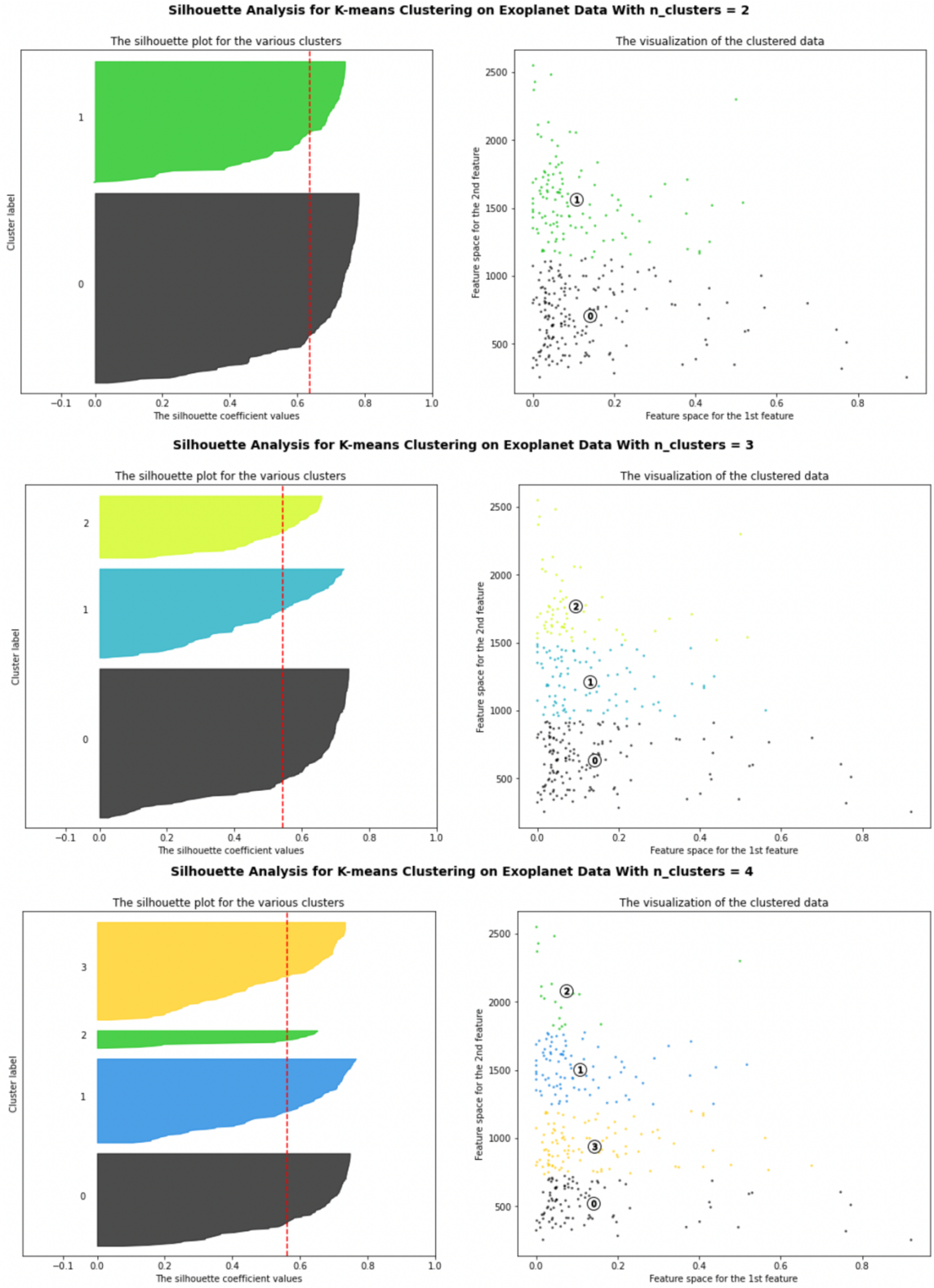
5 Results

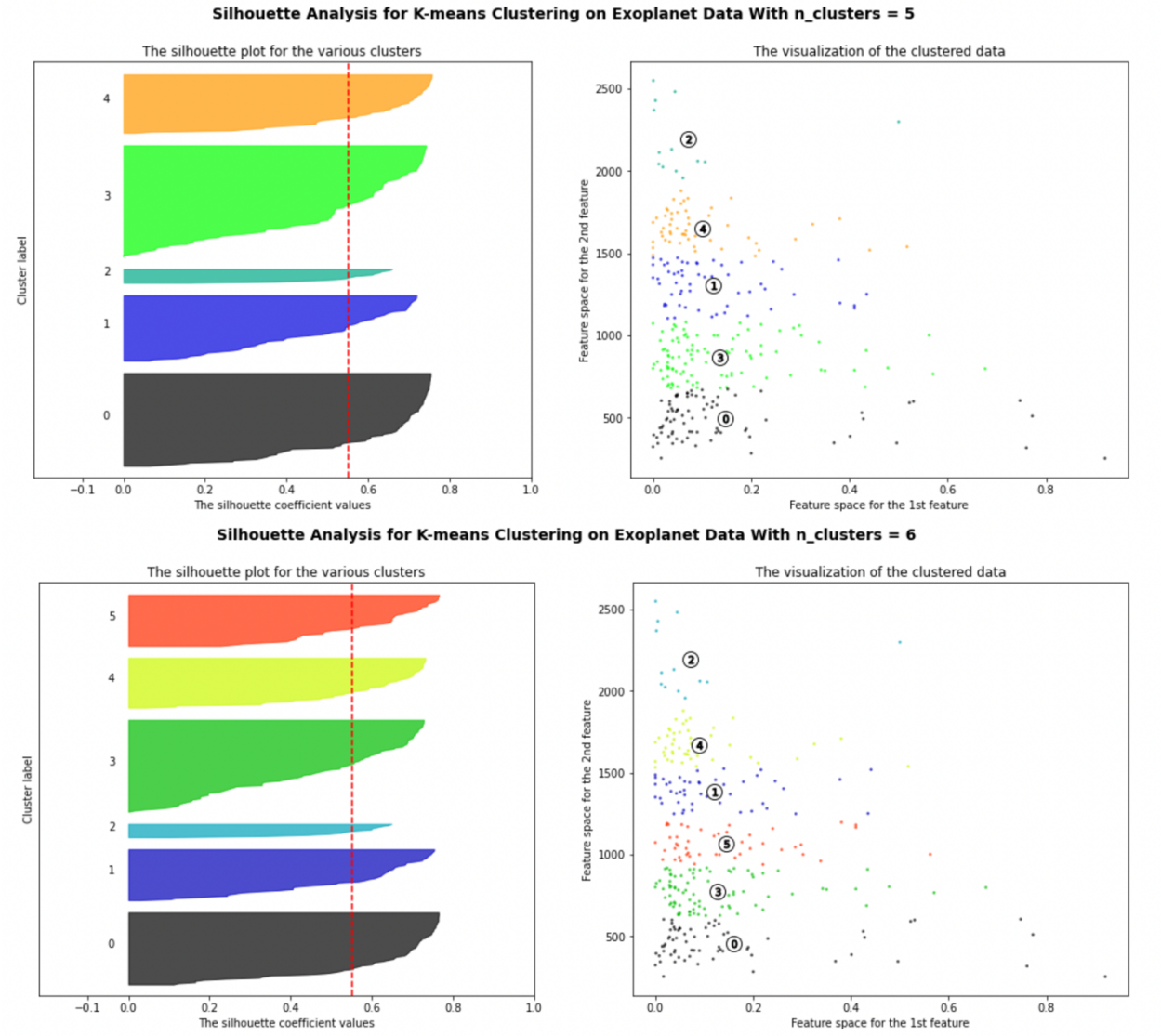
5.1. Scree Plot



The scree plot indicated that the first principal component explained ~48% of the variance in the data, and the second principal component explained ~30%. With the first two principal components explaining nearly 80% of the data's variance, and the remaining components all explaining under 15%, this was a clear determinant for using the first two principal components for clustering.

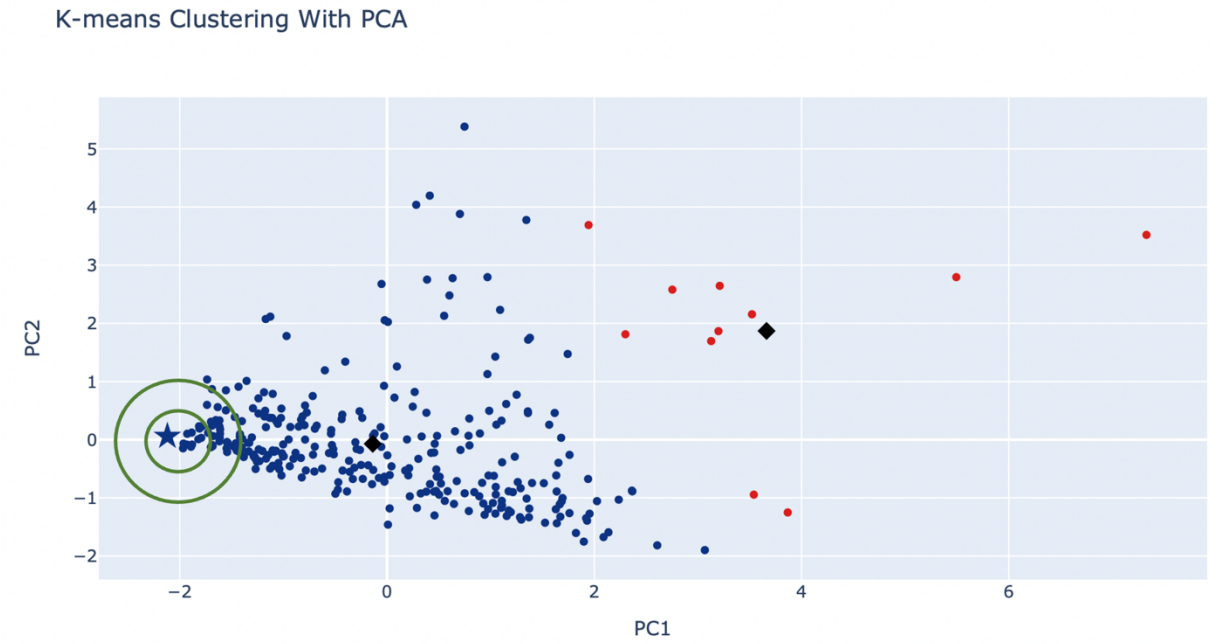
5.2. Silhouette Analysis





The silhouette analysis indicated that 2 clusters were optimal, returning the highest silhouette score of 0.638. The remaining silhouette scores were [3 clusters: 0.5429, 4 clusters: 0.5635, 5 clusters: 0.5517, 6 clusters: 0.5519]. As such, the subsequent k-means clustering was conducted with a target of 2 clusters.

5.3. K-Means Clustering



The k-means algorithm was able to produce 2 distinct clusters based on the data's 2 principal components. The plotted clusters are in blue and red, with the cluster centroids represented by black diamonds, and Earth is plotted as the blue star within its cluster.

6 Discussion

The initial thinking was that the silhouette analysis would suggest more than 2 clusters would be optimal for the algorithm. This would have enabled a more classical "grading" of the exoplanets, with the cluster nearest Earth receiving an "A," the second nearest a "B," and so forth. While that did not turn out to be the case, at least, not in this iteration of clustering, there was still valuable insight and interpretation of the plotted principal components. As can be seen in the "K-means Clustering With PCA" chart, green rings were place around the blue star representing Earth. These are theoretical radii that could be used to conclude that exoplanets falling within are the most Earth-like, and the grading could begin from that point.

7 Conclusions

Clustering with k-means to cluster exoplanets shows promise. The scree and silhouette analyses were "clean," producing clear directives for the number of principal components and the number of clusters. And as NASA's Exoplanet Archive gets fleshed out, filling null values with newly acquired observational data and with newly discovered stars and planets, the models will only increase in their predictive power. With more data, perhaps even more clearly delineated clusters are possible. There is optimism that the exoplanets plotted nearest to Earth are more Earth-like than those plotted farther away, with livable gravitation and temperature and a stable orbit that holds it within the system; but more work needs to be done to further test clustering as a method

for determining such similarity, and investigation into the properties of the exoplanets deemed most similar to Earth by k-means, and those deemed most dissimilar. That acknowledged, given the result here, exoplanet clustering with k-means could be the beginning of an extensible framework for a tool that accurately filters for the best Earth II candidates.

8 Future Work

As referenced earlier, an investigation into the characteristics of the exoplanets deemed most like Earth by the algorithm - those that are within the same cluster and near Earth – needs to be conducted. From what's learned after study can inform how subsequent iterations of the clustering are constructed. The feature engineering can also be adjusted to test which configurations produce better clustering results – maybe orbital period in addition to orbital eccentricity should be included; additionally, a broader star-planet model could be constructed, incorporating exoplanets' host star characteristics into the model. Maybe the goal should be to find the most Earth/Sol-like system rather than concentrating solely on the exoplanets. And finally, k-means will need to be tested against other unsupervised learning algorithms. Perhaps the more modern extension of it, x-means, performs better. DBSCAN and its extension HDBSCAN are other alternatives for clustering, and the best performing models should be used to form the basis of unsupervised clustering to find Earth II candidates.

References

Del Genio, Anthony, Michael Way, David Amundsen, Linda Sohl, Yuka Fujii, Yuka Ebihara, Nancy Kiang, Mark Chandler, Igor Aleinov, and Maxwell Kelley. "Equilibrium temperatures and albedos of habitable earth-like planets in a coupled atmosphere-ocean GCM." *Habitable Worlds 2017: A System Science Workshop*, no. GSFC-E-DAA TN51079. 2017.

Banerji, Manda, Ofer Lahav, Chris J. Lintott, Filipe B. Abdalla, Kevin Schawinski, Steven P. Bamford, Dan Andreescu et al. "Galaxy Zoo: reproducing galaxy morphologies via machine learning." *Monthly Notices of the Royal Astronomical Society* 406, no. 1: 342-353. 2010

VanderPlas, Jacob, Andrew J. Connolly, Željko Ivezić, and Alex Gray. "Introduction to astroML: Machine learning for astrophysics." *2012 Conference on Intelligent Data Understanding*, 47-54. IEEE, 2012.

Lam, Christopher, and David M. Kipping. "Transit Clairvoyance: Enhancing TESS follow-up using artificial neural networks." *American Astronomical Society Meeting Abstracts,* vol. 229, 415-04. 2017.