Mining the Data of the Universe:

Extrasolar Planet Exploration and the Militarization of Space

Lindsay Sheppard

Utica College

**Introduction**

On August 9, 2018, the Vice President of the United States of America declared that a branch of the United States Armed Forces called the Space Force would be fully implemented by the year 2020 (Insinna, 2018). The branch will seek to “secure and extend American dominance” by conducting missions focused on situational advantages, battle management, nuclear command support, lift and range operations, communications, positional and navigational optimization, missile warnings, and potentially space debris cleaning and asteroid defense (Hebden, 2019; United States, n.d.). To assist the Space Force with budgeting and militarization efforts, understanding funding allocation that will maximize Space Force presence and actualize their goals is vital.

Identifying locations where alien life and possible foreign threats may be encountered, as well as engineering space vessels with the ability to maximize speed for travel to distant locations, are top priorities. The rapid discovery of extrasolar planets (exoplanets) may provide insight into and advance current human knowledge about planetary habitability, potential space colonization, and space travel. In fact, a month after the Space Force announcement, a congressionally mandated report authored by the National Academies of Sciences, Engineering, and Medicine, urged NASA to accelerate its study of exoplanets by following a proposed “Exoplanet Science Strategy” that specifies research priorities and makes recommendations on investment allocation (National, 2018).

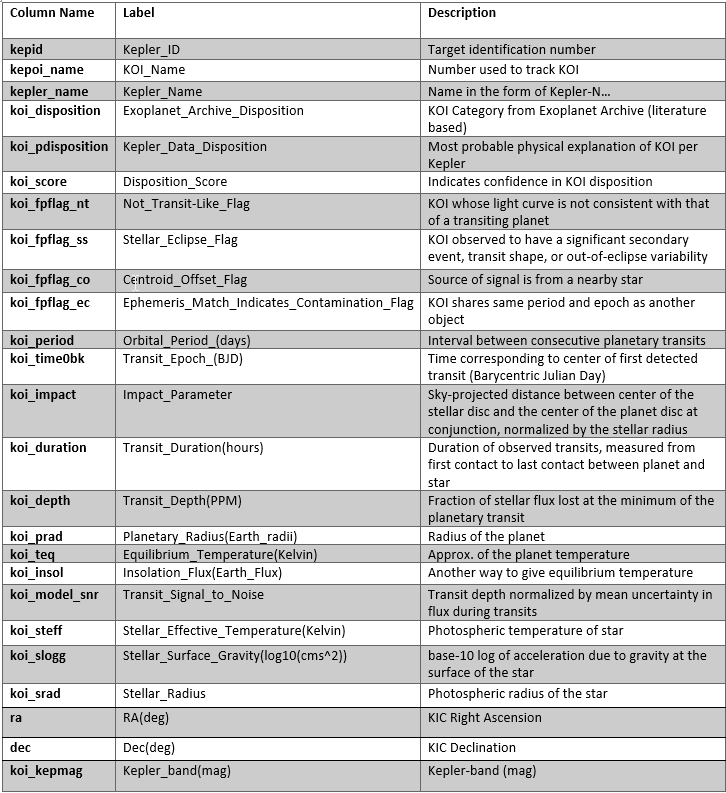
These historical achievements indicate the urgent necessity of exoplanet discovery. Initial funding should be directed toward the areas of most promise as well as those that direct space exploration further. Data mining techniques that have the power and accuracy to identify exoplanet patterns in NASA’s Kepler mission data will be utilized with increasing frequency as they provide crucial insights into space beyond current ground-based capabilities. The Kepler spacecraft amassed an enormous amount of data during its nine-year endeavor. Upon retirement in October 2018, over 500,000 stars and 2,600 exoplanets were discovered, some of which may be habitable (NASA, 2018).

NASA’s Washington-based director of the astrophysics division stated that there are “exciting discoveries lurking…in (the) Kepler data, waiting for the right tool… to unearth them” (NASA, 2017). The Space Force will benefit from classification models for candidate, confirmed, or non-exoplanets, based on factors such as planetary radii, orbital periods, and surface temperatures, from which newly discovered exoplanets found by Kepler’s replacement, Transiting Exoplanet Survey Satellite (TESS), may be rapidly classified. Additionally, a sufficient way to group exoplanets based on known criteria or perhaps location within space is necessary for a similar reason. Specific areas within habitability zones may be identified to increase space-based intelligence and advance American space technologies (United States, n.d.). Model development will allow Space Force to quickly target areas of high interest for regulation, observation, or potential colonization (Pelton, 2018).

**Method of Analysis**

The current analysis focused on finding answers relevant to Space Force planning, such as determining which stars are most likely to host planets, where potentially habitable exoplanets are located, whether a forest model could predict exoplanet candidates with accuracy, and what features, if any, categorize planets. The NASA Exoplanet Science Institute (NESI) houses the Kepler data, which is available for download from their website (NASA, 2019). Several Kepler data tables amalgamated to produce the cumulative data, which was updated on September 27, 2018 and features 9,564 observations made by the Kepler telescope during its operational period. There are 50 total variables: 38 quantitative and 12 categorical. Key columns are *kepoi\_name* (KOI\_Name), which identifies a Kepler Object of Interest (KOI), *koi\_disposition* (Exoplanet\_Archive\_Disposition), which provides a literature-based disposition of candidate, false positive, or confirmed for each KOI, and *koi\_pdisposition* (Kepler\_Data\_Disposition), which assigns candidate or false positive status to each KOI as assigned by the Kepler telescope.

Remaining columns refer to either planetary or stellar features, such as radii, orbit, and temperature, or contain error measurements for quantitative variables. The error measurements contained a large amount of null values; they were excluded. Brief metadata is presented tabularly in Figure 1; extensive definitions can be found on the NESI website (NASA, 2019).



**Figure 1:** Preview of Metadata (NASA, 2019)

Alteryx Designer version 20818.3.5.52487 cleaned the dataset and performed the data mining techniques. R version 3.5.2 created graphs and generated results where Alteryx was limited. Kepler IDs 8669092, 5872150, and 4078157 were initially coded as false positives; NASA determined recently that these KOIs are candidates and were adjusted appropriately. Filtering the Exoplanet\_Archive\_Disposition column determined that of 2,293 KOIs marked “confirmed” by the literature, Kepler calculations considered 2,248 candidates and 45 false positives. The remainder of the planets with either “candidate” (n= 2,253) or “false positive” (n = 27) status in the Exoplanet\_Archive\_Disposition column were marked the same in the Kepler\_Data\_Disposition column (Table 1), and there is little discrepancy between the two disposition columns.

**Table 1: Exoplanet\_Archive\_Disposition vs. Kepler\_Data\_Disposition**







Review of the field summary (Figure 2) suggested there were several variables with missing data. Since the Kepler name field only contains names for approximately 24% of the KOIs, that column was removed. Additionally, Disposition\_Score contained quite a few missing variables but is a column that determines the likelihood of a planets’ disposition. It remained in the data for exploratory and clustering purposes but was excluded during classification. A total of 363 records were missing a large amount of values in certain predictor variables like transit depth, planetary radius, and others. These were removed prior to building a forest model as the classification technique using Alteryx could not handle null values.

****

**Figure 2:**Field summary showing variables with and percentage of missing values (red)

Exploratory Data Analysis

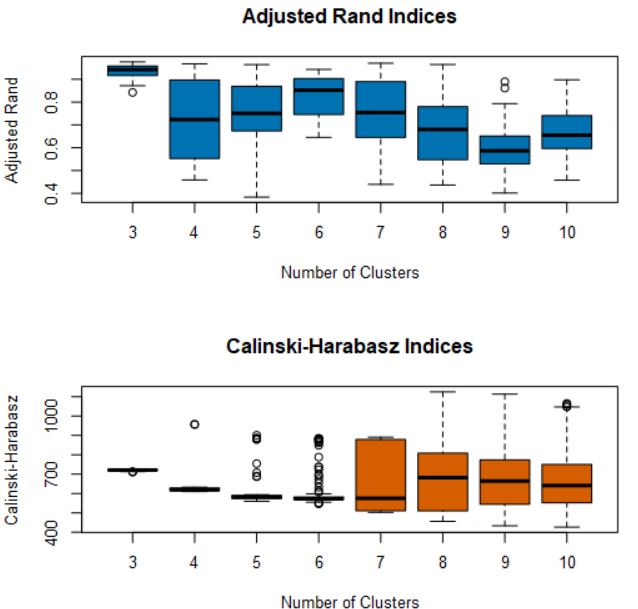
To determine the difference between binary and non-binary stars, the column *Stellar\_Eclipse\_Flag*, which marks binary stars with a 1 and non-binary stars with a 0, was visualized with a plot using ggplot2 in RStudio using *Exoplanet\_Archive\_Disposition* as reference. Binary stars are defined as star systems that consist of two stars which orbit around a common barycenter (Binary, n.d.). The locations of potentially habitable exoplanets were graphed in a similar manner. Planetary habitability is dependent upon large mass, orbital period, planetary radius, and temperature conditions. Earth-like conditions are generally sought, as Earth is the only planet known to harbor life. Therefore, candidates for habitability were candidate or confirmed exoplanets with a disposition score of at least 0.5, an orbital period between 100 and 500 days, a planetary radius between half to two times Earth’s radius, and a temperature between 200- and 400-degrees Kelvin to account for unknown stellar-planetary distance.

Classification

To build an exoplanet classification model, a random forest of 500 trees was selected as historically, forest models have performed well with similar data. It is highly accurate, runs efficiently, and is not prone to overfitting (Breiman & Cutler, n.d.). Eighty percent of the data was used to train the model, the other 20% was withheld for validation. The variable Exoplanet\_Archive\_Disposition was chosen as the target variable. The chosen predictor variables included: *Not\_Transit-Like\_Flag, Stellar\_Eclipse\_Flag, Centroid\_Offset\_Flag, Ephemeris\_Match\_Indicates\_Contamination\_Flag, Orbital\_Period, Transit\_Epoch\_(BJD), Impact\_Parameter, Transit\_Duration(hours), Transit\_Depth(PPM), Planetary\_Radius(Earth\_radii), Equilibrium\_Temperature(Kelvin), Insolation\_Flux(Earth\_Flux), Transit\_Signal\_to\_Noise, Stellar\_Effective\_Temperature (Kelvin), Stellar\_Surface\_Gravity(log10(cms^2), Stellar\_Radius, RA(deg), and Dec(deg).*

Cluster Analysis

In the unsupervised approach to planet categorization using K-means clustering, dimensionality was not reduced prior to analysis because other studies on similar data did not demonstrate this to be necessary (Bazzo et al., 2019; Torretta, 2019). However, Disposition\_Score was used to ensure that the labels assigned to each KOI were legitimate; any planet with a score less than 0.5 was excluded from the analysis. Categorical features were also excluded as K-means utilized Euclidean distances. K-means was chosen specifically as it is one of the most popular clustering algorithms and generally chosen first when attempting to identify dataset structure (Dabbura, 2018). The variables *Orbital\_Period\_(days), Transit\_Epoch\_(BJD), Impact\_Parameter, Transit\_Duration(hours), Transit\_Depth(PPM), Planetary\_Radius(Earth\_radii),* and *Equilibrium\_Temperature(Kelvin)* were included in the cluster analysis. Standardizing the fields using a z-score method reduced outliers to which K-means is sensitive. Prior to clustering, examination of the Adjusted Rand and Calinski-Harabasz indices (Figure 3) indicated that three clusters may be the appropriate starting number. This was determined using a minimum and maximum clustering of three and ten, respectively, and seven starting seeds.

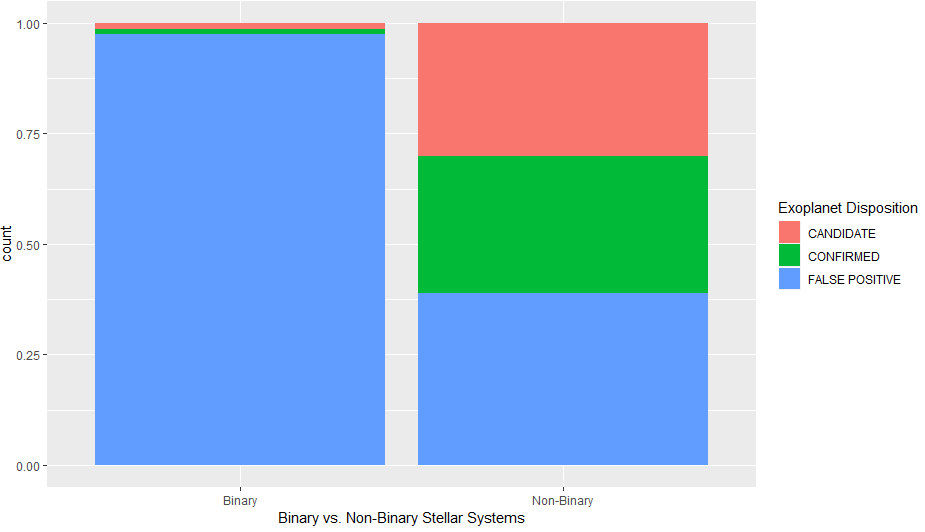


**Figure 3:** Indices for cluster determination

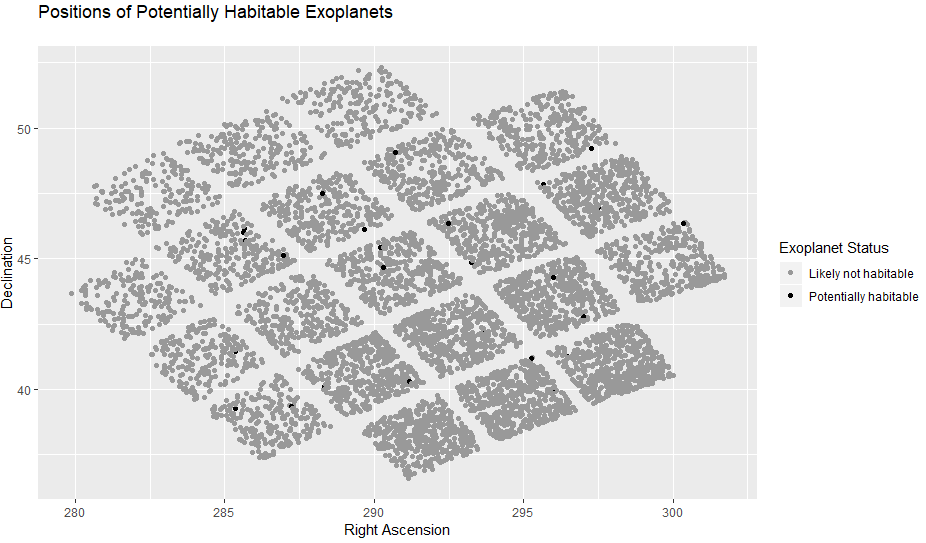
**Results**

Exploratory Data Analysis

Visualization of binary versus non-binary stars (Figure 4) determined non-binary stars are more likely to host potential exoplanets. Whether binary stars are incapable of hosting exoplanets, or if the detection of their planets is more difficult to achieve than it is for non-binary stars, is uncertain (NASA, 2019; Planetary, n.d.). The plot of exoplanet positions (Figure 5) suggested that Earth-like planets are dispersed randomly in space; no particular area appears to include more habitable planets than others.



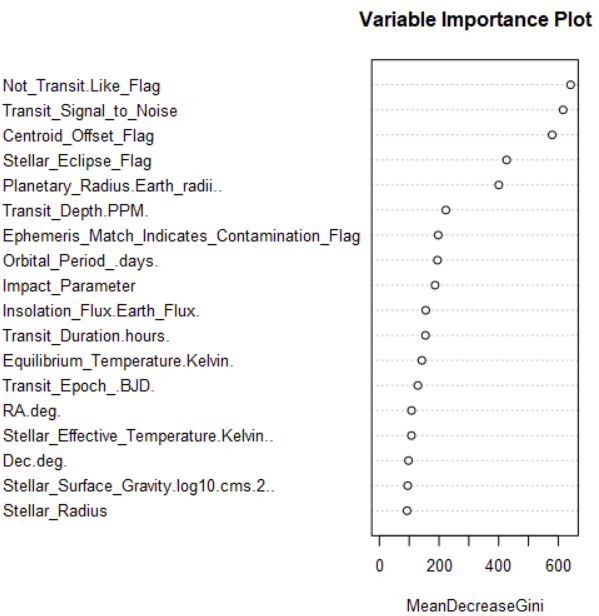
**Figure 4:** Binary vs. non-binary stars and exoplanet disposition; code adapted from Bazzo et al. (2019)



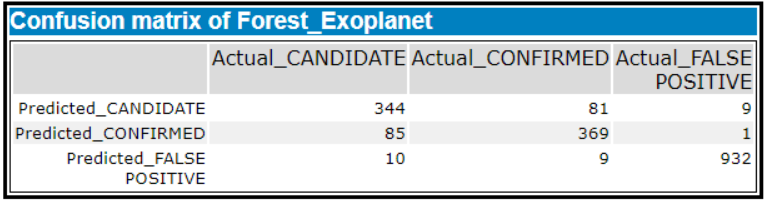
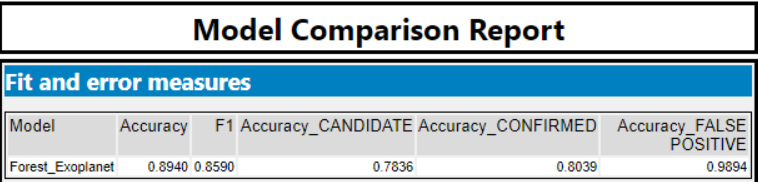
**Figure 5:** Positions of candidate and confirmed exoplanets: Earth-like versus Not Earth-like; code adapted from Bazzo et al. (2019).

Classification

The initial random forest out of the box estimate of error was 10.4%. The variable importance plot displayed the most significant variables in creating the model, which were mostly the flag, radius, and transit variables (Figure 6). The model (Figure 7) performed well, with an overall accuracy of 89.4%, but was most accurate at predicting false positives (98.9%) and least accurate at predicting exoplanet candidates (78.4%).



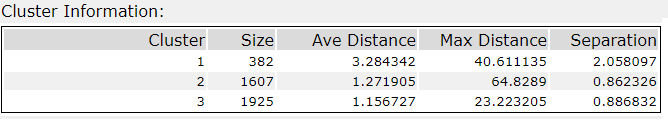
**Figure 6:** Variable Importance Plot



**Figure 7:** Random forest model results

Cluster Analysis

The K-means cluster analysis (Figure 8) provided a solution on two principal components and resulted in three clusters: cluster 1 contained 382 exoplanet candidates with an average distance of 3.28 and separation of 2.06. The other two clusters were larger, with 1607 and 1925 exoplanets, respectively, had smaller average distances (1.27 and 1.16), and smaller separations (.86 and .89).

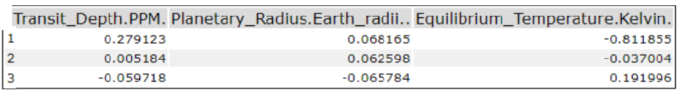
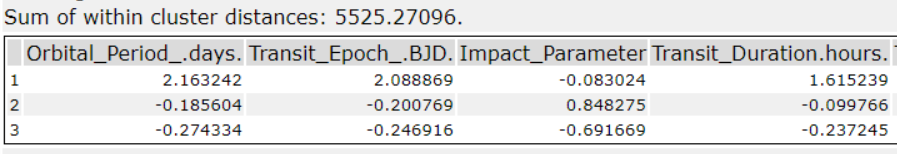


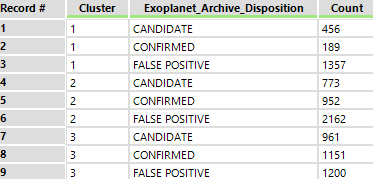
**Figure 8**: K-means cluster results

**Result Interpretation**

The binary star plot determined that multiple star systems produce less candidate exoplanets, while the plot of exoplanet locations in space suggested that exoplanets that are most Earth-like are not found in specific, targetable locations but rather randomly distributed. These results suggest that the Space Force may focus attention on non-binary systems until further research is performed on binary systems, but there do not seem to be specific locations in space where one might find them. However, estimates suggest that somewhere between 33% to more than 50% of all stellar systems are binary, and that binary systems may be able to support life depending on a variety of factors (Planetary, n.d.). Once targets are found, however, the random forest model may perform well when presented with new, unclassified data from TESS or other exploratory ventures, assisting the Space Force in maximizing time and resources. The model had a low misclassification rate around 10% and seemed to predict exoplanet categories well with little bias in any one category.

For unsupervised planet categorization, cluster analysis using K-means determined that Cluster 1 had larger “time” variables, such as *Orbital\_Period(days),* *Transit\_Epoch\_(BJD),* and *Transit\_Duration(hours)* than Cluster 3, suggesting these clusters contain objects that are somewhat opposite from each other. Cluster 1 also contained a smaller number of objects than the others, which may indicate that the variables related to those objects are unusual or are outliers and requires further examination. Cluster 2 was different from the other clusters in terms of impact parameter objects and transit variables. Cluster 3 was mostly different than the other clusters in regard to almost all variables except planetary radius. Most notably, cluster 1 contained the fewest “confirmed” planets, and cluster 2 contained the most “false positives”, yet each cluster contained a significant amount of each type (Figure 9).



  
**Figure 9**: Cluster results, comparison to known exoplanet status

**Limitations**

There is not enough data to definitely say why non-binary stars are likely to host exoplanets (NASA, 2019). More research is required in this area. Furthermore, NASA stated that the NESI dataset is cumulative and not suggested for use with analyses requiring uniform distributions. Despite its “cumulative” status, the dataset is lacking variables that other NESI datasets contain that are important predictors of exoplanet status, such as metallicity, mass, and chemistry (Planetary, n.d.). The dataset was limited to only the few features used as determinants of “Earth-like” status and the determination of potentially habitable exoplanets was rudimentary. Thus, there may be several more exoplanets that are Earth-like than shown in Figure 4.

The classification and clustering models also faced limitations. Alteryx is powerful software, yet it cannot build random forest models with missing values without separate imputation. It also cannot use more than ten initial seeds for K-means clustering and may experience long runtimes depending on the number of clusters and initial seeds chosen. R may be a better choice for these processes; in Alteryx, iterative processes are not always efficient yet required in order to optimize results. Further iterations in initial number of clusters and starting seeds may be necessary. In addition, the K-means results are not intuitive or conclusive; K-means is sensitive to outliers and non-globular clusters, so perhaps scaling the data is insufficient without further examination. Other methods may perform more efficiently and deliver more obvious or understandable results.

**Suggestions**

Establishing the difference between binary and nonbinary stars is necessary to understanding which are likely to host exoplanets. Resources may be targeted towards nonbinary stars if deemed likely candidates for planetary habitability or suitable locations for positional and navigational optimization. Research may be conducted to improve current methodologies for the study of binary stars.

Additionally, using Kepler names or KOI labels to join the KOI cumulative data with other available NASA data that includes other useful variables may allow for enhanced detection and visualization of potentially habitable exoplanets and their locations in space. Moreover, other data sets contain more information regarding planetary positions in space, which could be modeled using a 3d scatterplot to target locations of interest. Alternatively, the NESI confirmed exoplanets table alone could be utilized for location determinations and may also improve classification and clustering models; the dataset contains additional predictor variables, such as stellar flux observations, that may expand results.

As the forest model used several predictor variables, variable reduction techniques performed prior to forest model generation may provide improved results. Attempting other classification algorithms and comparing the results to the current model may also prove beneficial. Programming languages such as R and Python might be more efficient than Alteryx for iterative clustering techniques. However, Alteryx is better at appending cluster labels quickly after a model has been created. Using the K-means clustering method but validating it against new candidate data is recommended. Alternate techniques, such as K-medoids or neural gas, may be considered in an attempt to improve clustering results. Ultimately, the goal is to apply these techniques to the new TESS data for rapid classification, efficiently using time and resources.

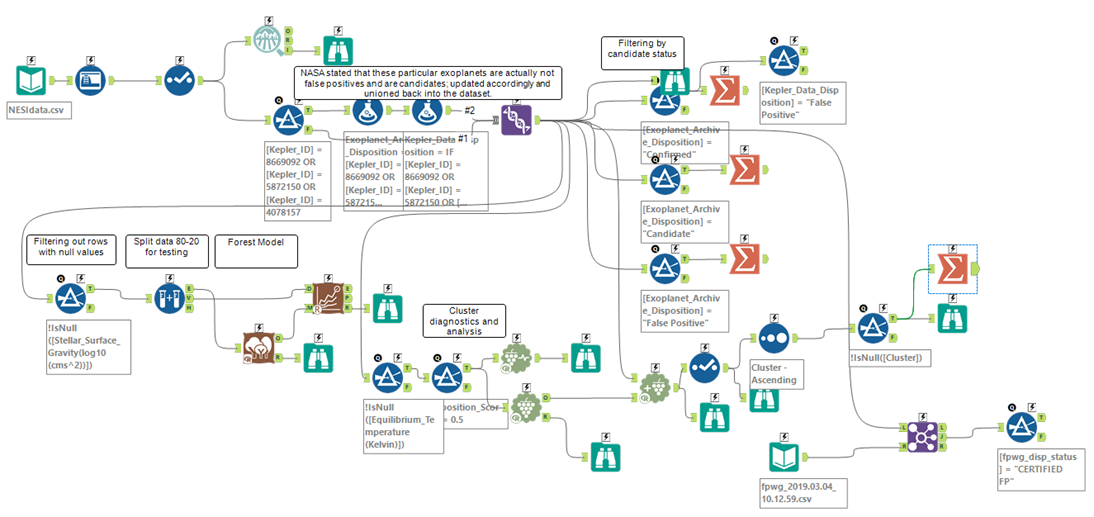
**Conclusion**

The Space Force will be able to use the results of this study as a relevant starting point for exoplanet detection. Research efforts are best directed toward improving exoplanet detection methods for binary stellar systems, as well as increasing efforts in locating promising non-binary based planetary systems. Compiling the confirmed exoplanet data into a larger resource for further study may reveal additional insights into spatial locations and potentially habitable planets. Determining locations that are prime for situational advantage, battle management, and lift and range operations are important because then the Space Force can direct research and development efforts into engineering efficient space travel.

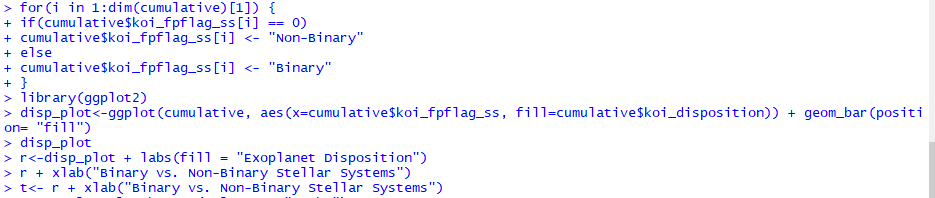
Rapid classification of new TESS-found exoplanets may be performed using the random forest model generated in this analysis initially; the model may also be adjusted as with extra variables pulled from other data sets or through pre-classification variable reduction. Clustering techniques are able to group the exoplanets in an unsupervised manner and will be useful for exploratory analysis of new exoplanets as they are found, yet K-means may not be the best approach for status determination without significant adjustment. However, the insight gained from data mining may certainly decrease time and resources spent by the Space Force and will allow for the successful militarization and potential colonization of space.

**Appendix**

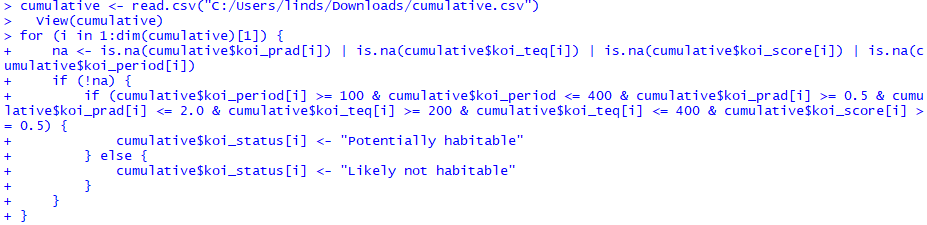
Alteryx Workflow

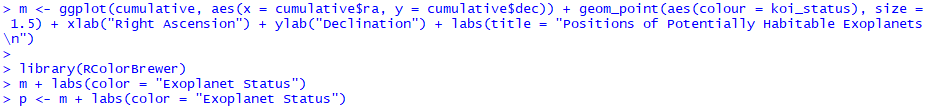
****

R Code for Figure 3



R Code for Figure 4







**References**

Bazzo, M., Choe, S.H., Yan, S., Zhang, A. (2019, February 2). Kepler exoplanet project. *GitHub.* Retrieved

from https://github.com/soochoe/Kepler-Exoplanet-Project/blob/master/README.md.

Binary star. (n.d.) In *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Binary\_star

Breiman, L. & Cutler, A. (2004, June 15). Random forests. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc\_software.htm

Dabbura, I. (2018, September 17). K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. *Towards Data Science.* Retrieved from https://towardsdatascience.com/k-means clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Hebden, K. (2019, February 20). Trump signs directive to initiate his Space Force. *Room: The Space*

*Journal*. Retrieved from https://room.eu.com/news/trump-signs-directive-to-initiate-his-space-

force

Insinna, V. (2018, December 20). Trump’s new space force to reside under department of the air force. *DefenseNews.* Retrieved from https://www.defensenews.com/space/2018/12/20/trumps-new- space-force-to-reside-under-department-of-the-air-force/

NASA. (2017, December 14). Artificial intelligence, NASA data used to discover eighth planet circling distant star. Retrieved from https://www.nasa.gov/press-release/artificial-intelligence-nasa- data-used-to-discover-eighth-planet-circling-distant-star

NASA. (2018, October 30). NASA retires Kepler space telescope. Retrieved from https://www.jpl.nasa.gov/news/news.php?feature=7272

NASA. (2019). NASA exoplanet archive [data file]. *NASA Exoplanet Science Institute.* Retrieved from

https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-

tblView?app=ExoTbls&config=cumulative

National Academies of Sciences, Engineering, Medicine. *Exoplanet Science Strategy*. Retrieved from https://www.nap.edu/download/25187

Pelton, J. (2018). Saving humanity– is space up to the job? *Room: The Space Journal*. Retrieved from https://room.eu.com/article/saving-humanity-is-space-up-to-the-job

Planetary habitability. (n.d.) In *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Planetary\_habitability

Torretta, M. (2019, February 21). Kepler exoplanets search EDA. *Kaggle.* Retrieved from

https://www.kaggle.com/marceltorretta/kepler-exoplanets-search-results-eda.

United States Space Force. (n.d.). in *Wikipedia.* Retrieved from

https://en.wikipedia.org/wiki/United\_States\_Space\_Force.