Continuous Habitable Zones: Discovering the True Alien Earth*

An Exploratory Analysis of Exoplanet and their Host Stars

Dingding Wang

01 May 2022

Abstract

The search for habitable worlds beyond earth is at the forefront of astronomy. As astronomical observation technology improves in the last decades, the detection of exoplanets has undergone extraodinary growth, but scientists are in lack of time and resources to describe all of them. When an exoplanet is identified using a given detection method, scientists try to determine its basic properties including what its composition, temperature, whether it contains atmosphere, and particularly, whether it is an Earth-like habitable planet. The discovery of numerous exoplanet systems containing diverse populations of planets orbiting very close to their host stars challenges the planet formation theories based on the solar system. This paper proposes a general model that uses statistical techniques to explain the relationship between exoplanets detected as of April 2022 and the major features of their host stars. The main goal is to establish a mathematical relationship between a set of variables to better describe the physical characteristics of the the planet itself and further predict its habitability.

Keywords: Habitable zone, Exoplanet, Linear regression, Cross-sectional data, Orbital properties, Dynamical evolution

Contents

*Code and data are available at: https://github.com/estherxwang/EXOPLANET-Analysis.

1 Introduction

An extrasolar planet (or exoplanet) is a planet that orbits a star other than the sun and therefore belongs to an extrasolar planetary system. For a long time in human history, the only known planetary system is our own solar system, which greatly limits the research on "alien earth". Since the very first discovery of the exoplanet orbiting a sun-like star in 1995 winning the 2019 Nobel Prize in Physics, the number of known exoplanets has exploded, fueling the field's rapid development and making it one of the most active frontiers in astronomy today. Over the past decade, NASA's Kepler satellite has revolutionized research by hunting for thousands of exoplanets using various transiting method. According to NASA, scientists have found over 4,000 exoplanets since the first discovery in 1995, and NASA's Kepler Space Telescope which launched in 2009 detected more than half of those planets. The scientific objective of the Kepler Mission is to explore the diverse planetary systems through surveying a large sample of stars to determine the percentage of Earth-like planets and larger planets in or near the habitable zones of various stars.

It has long been a dream for worldwide astronomers to find the first truly "alien Earth". In Figure 1, we can see the location of all known exoplanets that could potentially harbor life in the night sky by the Planetary Habitability Laboratory at the University of Puerto Rico at Arecibo(Wenz (2016)). Although we know of more than 4,000 planets out there in our galaxy and the list keeps growing, only a small handful of all those planets out there are considered habitable. Many of the Kepler planets are clustered as the top left corner of the map due to the spacecraft looked only at one particular portion of the sky. "The tremendous growth in the number of Earth-size candidates tells us that we're honing in on the planets Kepler was designed to detect: those that are not only Earth-size, but also are potentially habitable," said Natalie Batalha, Kepler deputy science team leader. "The more data we collect, the keener our eye for finding the smallest planets out at longer orbital periods."(NASA (2001)) Some exoplanets are too big to have a solid surface, and others are either too close in to support life, or too distant not to freeze over. To be qualified as a potential candidate for habitable planet, it must be relatively small in size and has an orbit within its host star's habitable zone. A habitable zone is the area around a star in which the surrounding earth-like planets are neither too hot nor too cold for the existence of liquid water on the surface and possibly support life. Imagine the position of Earth was where Pluto(the ninth planet from the Sun) is, the Sun would be barely visible and the extremely low temperature would lead to the freeze of Earth's ocean and much of its atmosphere.

As telescope technology advances, other factors will also be taken into account, such as the composition of the planet's atmosphere and how active its host star is. In the search for life, other similarities with Earth have become more pronounced. Many rocky planets have been detected within the size of Earth: a point conducive to life. Based on our observations in the solar system, large gas planets such as Jupiter seem unlikely to provide habitable conditions. Due to the difficulty of detecting Earth-sized planets orbiting sun-like stars in wide orbits, most of the exoplanets detected are orbiting red dwarf. For comparison, red dwarfs include the smallest of the stars which weigh between 7.5% and 50% the mass of the sun. However, these red dwarfs have a potentially deadly habit, especially when they're young: powerful flares tend to erupt from their surfaces with a certain frequency. These could sterilise planets in close orbits that are just beginning to support life, which a blow to a possible life(Brennan (2021)).

Detecting an exoplanet is an extremely difficult task as they don't emit any electromagnetic radiation of their own and are completely obscured by their extremely bright host star, meaning that ordinary telescope observation techniques are not appliable to exoplanets. Consequently, a variety of advanced techniques such as astrometry, gravitational lensing, photometry, pulsar timing, radial velocity and spectroscopy are used. The various tools are applied accordingly to the main goal of detecting the properties of exoplanets on their own star systems. To analyze the habitability of exoplanets, it is necessary to further build statistical models to explain the origin, formation and migration of these objects. For instance, a paper in 2008 (Martínez-

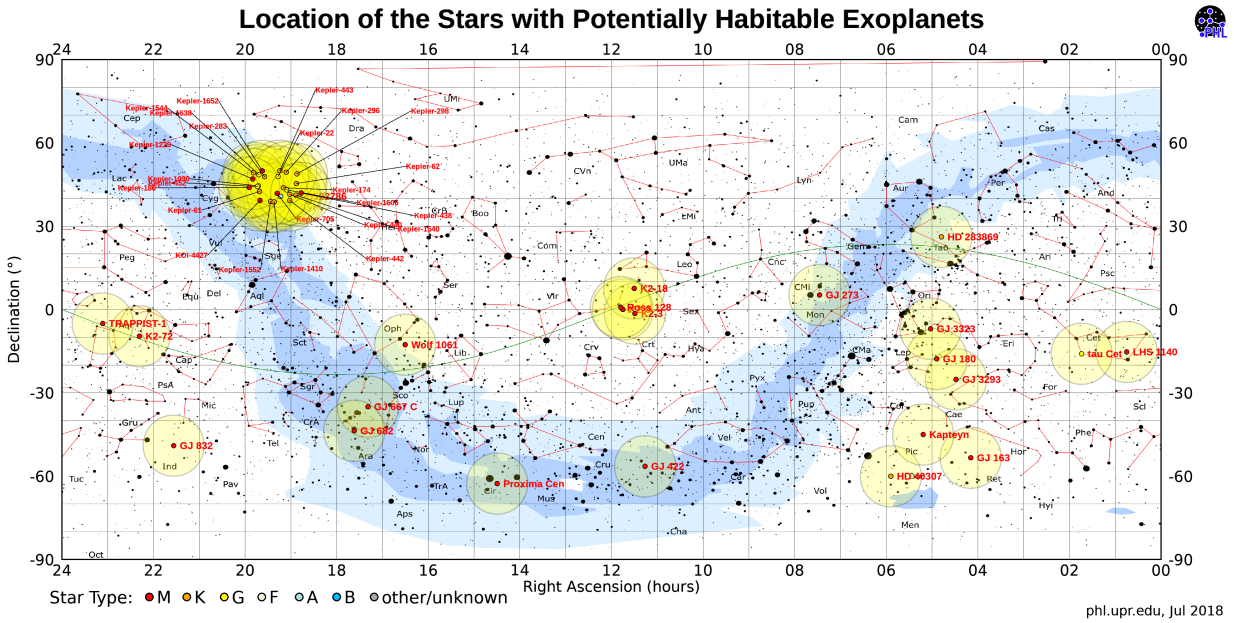

Figure 1: Habitable Exoplanet in the Night Sky

Gómez and Babu (2009)) provided analysis on cross-sectional data of exoplanets detected prior to date with linear regression techniques, aiming to analyze the relationship between host stars and orbiting exoplanets.

The dataset was obtained from the exoplanet catalog created in February 1995 (TEAM (2022)). The catalog aims to facilitate the progress of Exoplanetology studies through updating associated data from recent dectections of exoplanets by various institutes. With the collected data file, analysis are conducted to identify dynamical data of confirmed exoplanets. The raw data was loaded in, cleaned and analyzed the data using `R` (R Core Team 2020), `dplyr` (Wickham et al. 2021), `tidyr` (Wickham 2021) packages. For visualizations, figures and tables were created with `dplyr` (Wickham et al. 2021), `corrgram` (Wright 2021), `visdat` (Tierney 2017), `gridExtra` (Auguie 2017) `ggplot2` (Wickham 2016) and `kableExtra` (Zhu 2021). Finally, models were built with `stats` package, from base `R` (R Core Team 2020). In this paper, we seek to investigate the properties of exoplanets and propose a general model to explain the relationship between exoplanets detected as of April 2022 and the major features of their host stars using statistical techniques. For instance, if the mass of the orbiting planet is strongly determined by the mass of host star. First, we will be looking at the set of properties of exoplanets such as mass and radius. Then, we will proceed by building a linear mixed model with the mass of exoplanet as the response variable with other properties of the host star as predictors.

The remaining part of the paper is divided as follows: Data Section explains where our data comes from and gives a general idea of the variables present in the dataset. First we visualized key explanatory variables with histograms, scatterplots and boxplots, as well as the correlation plot which depicts the correlation between all the possible pairs of values in a table. Section Model shows the model that explains how mass of exoplanet is affected by different factors. We then checked the assumptions such that all variables to be multivariate normal, and obtained a summary table explaining the coefficients. Section Results explains the key findings of the data analysis and model. We used test and training datasets with a frequentist confusion matrix to measure the accuracy of our model and test the predictive property. Section Discussion expands on what is found and why it is important. Section Appendix contains supplementary graphs that support the arguments in the discussion section, as well as a comprehensive description for astronomy terminology used in this paper for readers with little relevant background in astronomy.

2 Data

2.1 Data Source and Collection

`Exoplanet.eu` is a constantly updated exoplanet encyclopedia combining interactive visualizations with detailed data on all known exoplanets. The catalog was established in February 1995 and developed and maintained by the exoplanet TEAM. It is updated daily, and up to the date of this report on April 27, 2022, it contains 5014 confirmed exoplanets. The catalog is a working tool that provides all the latest detections and data announced by professional astronomers to help advance the progress of exoplanetology. Given the heterogeneity of observational papers, a uniform catalog is unlikely to be built. Therefore, researchers are finally willing to make quantitative and scientific use of the catalogue, so that they can make their own judgments and tests on the possibility of data. Using `R` (R Core Team 2020), `tidyverse` (Wickham et al. 2019), `tidyr` (Wickham 2021) and `dplyr` (Wickham et al. 2021), I cleaned and extracted the necessary data to complete an exploratory analysis and modelling.

Among the 98 variables in the raw dataset, there are two groups of data: Planet data and Stellar data. Planet data are the latest exoplanet data taken from latest published papers and conferences and first-hand updated data on 9 professional websites, including Anglo-Australian Planet Search, California & Carnegie Planet Search, Geneva Extrasolar Planet Search Programmes, Transatlantic Exoplanet Survey, University of Texas - Dept. of Astronomy, HAT and HATS, WASP, NASA Exoplanet Archive and Kepler. The main resource - NASA Exoplanet Archive serves photometric time-series data from surveys that aim to discover transiting exoplanets, such as the Kepler Mission and CoRoT(NASA (2021)). The stellar data including multiple key properties of the host stars of exoplanets are taken from Simbad or from professional papers on exoplanets. The basic physical characteristics of a star include age, position, mass, radius, metallicity and temperature.

2.2 Important Variables

The conditions for a planet to be habitable are uncertain. Determining the conditions for a planet to be inhospitable, however, is relatively easy. To choose our variables, we browsed papers of recent year studies on exoplanet properties and habitable zones.The habitable zone for a given star describes the range of circumstellar distances from the host star within which an orbiting planet could have liquid water on the surface (Kane and Gelino (2013)). The two most important variables of exoplanets are mass and orbital distance, and we also need to take into account key properties of the host stars.

First, the mass and radius of a planet can determine whether it can support atmosphere. Y. Alibert has investigated the mass-radius relationship for a planet to be habitable, and the results show that for planets in the range of super-Earth mass (1–12 M), the radius of the planet with similar composition to Earth varies in a range of 1.7-2.2 R (Alibert (2013)). If the planet is low in mass and too small, it would not have sufficient gravity to hold atmosphere, and the atmosphere will be stripped away. Without atmosphere, the prospects for life are dim and complex life may can not evolve. As a results, scientists are concentrating on searching for massive planets with a critical lower limit for a planet to be habitable.

Second, the orbital distance plays a key role in the determination of possibility of a planet to support life. If the planet is either too close or too far to its host star, it would be either tremendously hot or cold. As the most important factor required for life to exist and develop, the possible presence of liquid water is believed to be necessary for habitability of exoplanets. On one hand, if the planet is too hot, the molecules would travel too fast generating too much energy, leading to very different chemical conditions from that on Earth and no resource to support human lifes; on the other hand, if the planet is too cold, it would be impossible to have liquid water which is essential for life. The upper and lower limits of orbital period boundaries define habitable zone with just the sufficient temperature to support life.

Finally, the classifications of stellar objects according to their feathers (e.g., magnitude, color) dates back to the 19th century (Pichara and Protopapas (2013)). A series of papers demonstrate that the properties of host star such as age, metallicity and temperature have great effect on the evolution of the habitable zone. The result shows that metallicity strongly affects the duration of the habitable zone and the distance from

the host star with maximum duration. In our general model for predicting mass of exoplanet, we will take these properties into account.

2.2.1 Planet Parameters

X..name. name of Exoplanet

mass. mass of the planet (unit: MJupiter)

radius. radius of the planet (unit: RJupiter)

eccentricity. eccentricity of the planet orbit from 0, circular orbit, to almost 1, very elongated orbit

discovered. year of discovery at the time of acceptance of a paper

orbital_period. orbital period of the planet in days

detection_type. method of detection of planet

2.2.2 Stellar parameters

star_mass. star mass in solar units (Msun)

star_radius. star radius in solar units (Rsun)

ra. Right Ascension

dec. Declination

RA (right ascension) and DEC (declination) are to the sky what longitude and latitude are to the surface of the Earth. RA corresponds to east/west direction (like longitude), while Dec measures north/south directions, like latitude.

star_distance. distance of the star to the observer (unit: pc)

star_metallicity. decimal logarithm of the massive elements (« metals ») to hydrogen ratio in solar units (i.e. Log [(metals/H)star/(metals/H)Sun])

star_age : stellar age (Gy)

star_teff : effective stellar temperature

2.3 Data Cleaning

The dataset from exoplanet catalog was mostly pre-cleaned for publication. In the data cleaning process, we first analyzed the distribution of missing data, and then converted some key variables including detection, planet status, star mass and radius, discovery year and metallicity into grouping factors for better visualizations and further studies. The criteria for grouping the mass and radius of exoplanet is to compare it with 1, since the original data is in Jupiter mass and radius. Similarly, the mass and radius of the host star is in the unit of that of the sun.

2.4 Missing Data

The dataset originally contains 98 variables with a lot of N/A values. Based on Figure 2, we have many missing values in mass, eccentricity and star age. For these missing values, we are interested in the systematic about how our data are missing. The mechanism is important since it affects how much the missing data bias the results, leading a big impact on what is a reasonable approach to dealing with the missing data. Hence it is important for us to take it into account in deciding the approach to choose (Grace-Martin et al. (2021)). The following are the brief definitions of these assumptions:

- Missing at Random (MAR): missing for reasons related to completely observed variables in the data set.

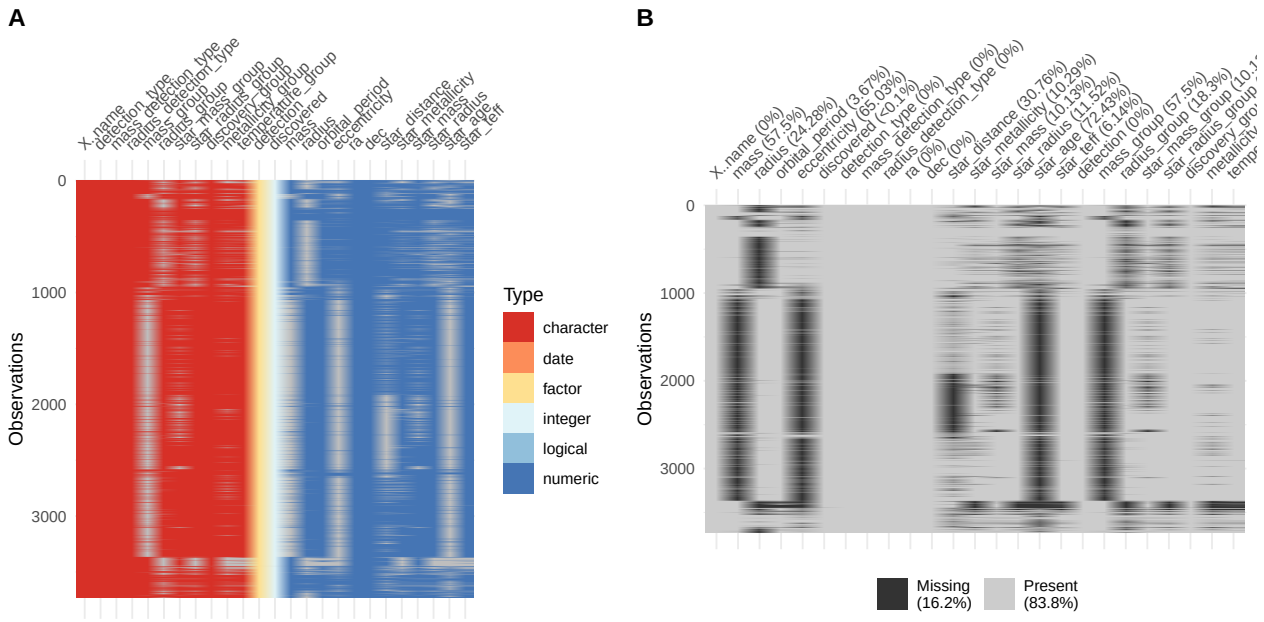

Figure 2: Missing Values in the Dataset

- Missing Completely at Random (MCAR): the propensity for a data point to be missing is completely random.
- Non-ignorable missing data: the propensity for a data point to be missing is not related to the missing data, but related to some of the observed data.
- Outliers treated as missing data.
- The assumption of an ignorable response mechanism.

Due to the exponential growth of astronomical data, the number of missing data also grows and it is necessary to apply statistical inference to take full advantage of data. In the literature, multiple interpolation is the standard method to deal with missing data. For example, the missing data can be filled using Monte Carlo approaches where each missing value is drawn from a distribution determined from the training set (Takahashi (2017)). Although the catalog did not give specific reasons for missing data, we choose to not omit all the missing values since the method would dramatically reduce the size of the training set.

2.5 Data Visualizations

Figure 3 A shows the distribution of all host stars of the exoplanet in the dataset on sky map coordinates, with color indicating if the mass of star is greater than mass of the sun and size indicating the radius. We can notice that there is an overdensity area in (290,50), and this is due to some telescopes point at a specific area of sky when observing exoplanets. Figure 3 B is a zoomed in of the area with 7 of most habitable exoplanets detected by Kepler marked. Among the stars, Kepler-22b is the most famous super-Earth which lies about 600 light-years away from us. It was also the very first Kepler planet found in the habitable zone orbiting its host star with the size about 2.4 times of Earth. The orbital period is about 290, which is pretty similar to that of Earth (Howell and Harvey (2022)).

2.5.1 Detection per year

Several methods for detecting exoplanets have been developed: Astrometry, Direct Imaging, Transit Observations, Microlensing, and Radial Velocity. Figure 4 shows the distribution and trend of detection types of all the exoplanets in the catalog by the discovery year. We can observe that after year 2010, Primary

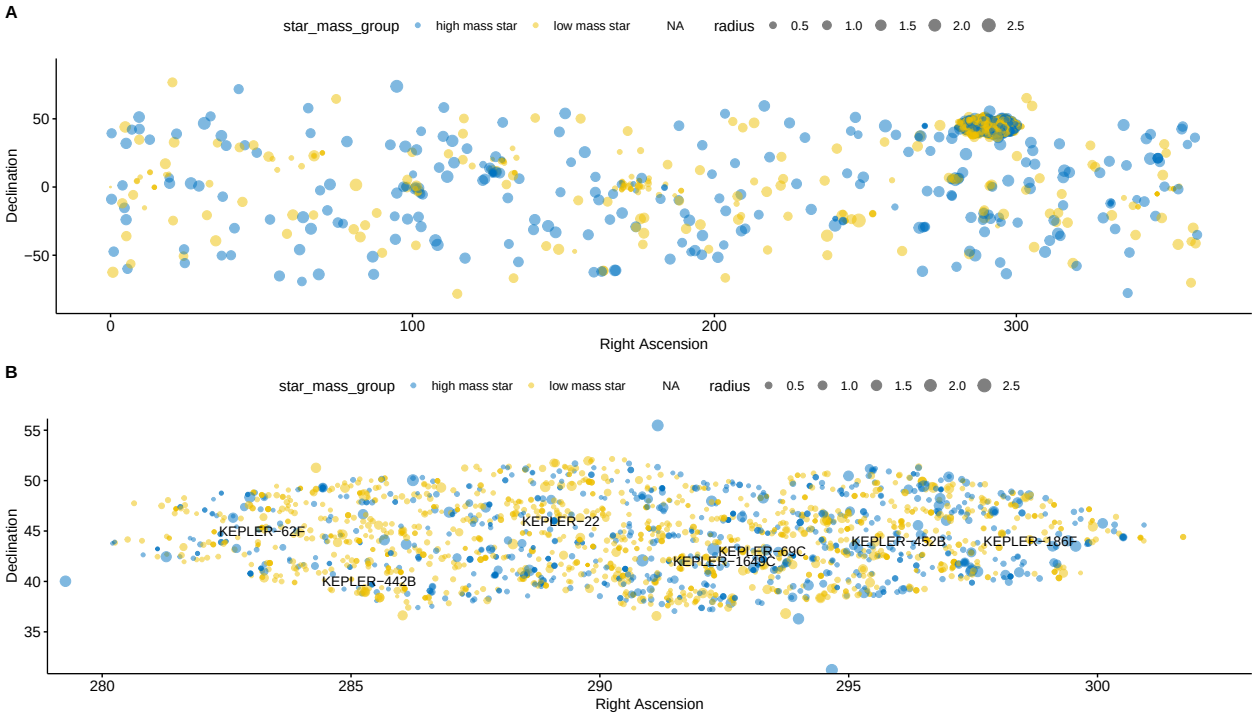

Figure 3: Sky Map of Exoplanet on the Milky Way

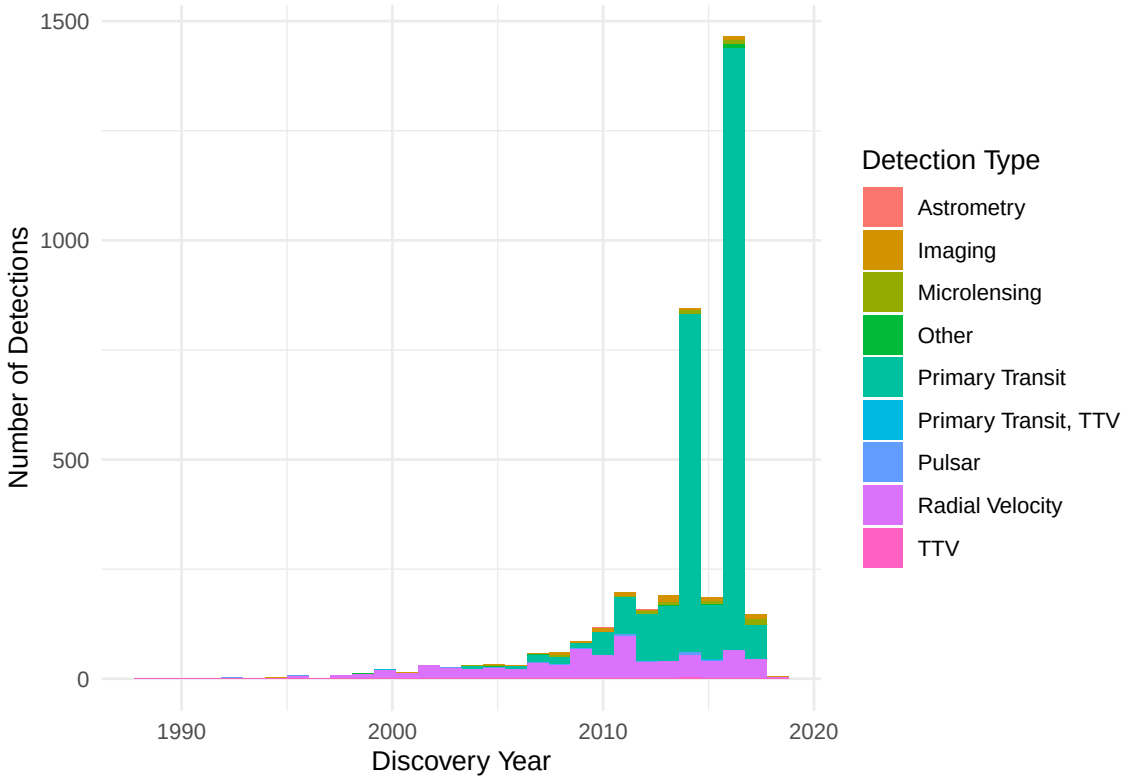

Figure 4: Detection Type by Year

Transit has become the domain method of detection, and Radial Velocity is the second most popular. The primary transit method is a photometric method which detects the presence of one or more exoplanets in orbit around a star indirectly. The science behind this is that a transit occurs when a planet passes between the observer and its host star. For instance, within our solar system, transits can be observed from Earth when the planets in between us and the sun (eg. Mercury and Venus) travel right in between(Brennan (2022)). While both the radial velocity and transit methods rely on the detection of variations in light from the star, transit methods is still currently the most effective and sensitive method for detecting exoplanets.

2.5.2 Plots for exoplanet

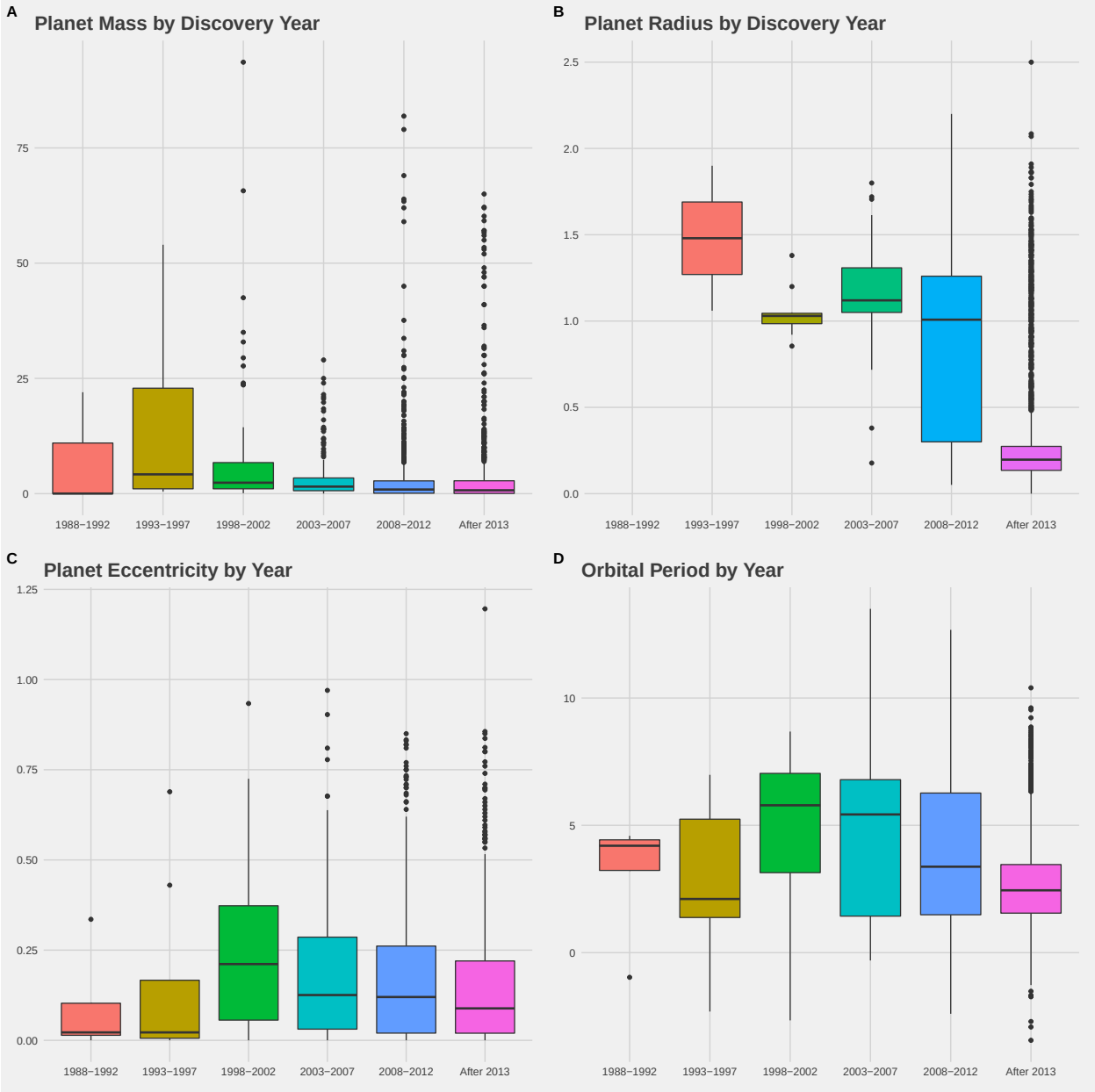

Figure 5: Mass, Radius, Eccentricity and Orbital Period by Year

Figure 5 shows the distribution of four important planet parameters by the discovery year. We can notice that a trend in decrease of the average radius of the planets discovered through years. The average radius

of late 20th century was about 6 times compared with detection in the recent years. The reason is that the as the widely use of Doppler technique for detections, it is most suitable for detecting massive and large planets orbiting close to the host star. That's because the host star wobbles more when it has a larger planet nearby, resulting in a larger, easier to detect spectral shift. With the better instruments in the recent years, astronomers are able to measure the radius better and detect those exoplanets smaller in size.

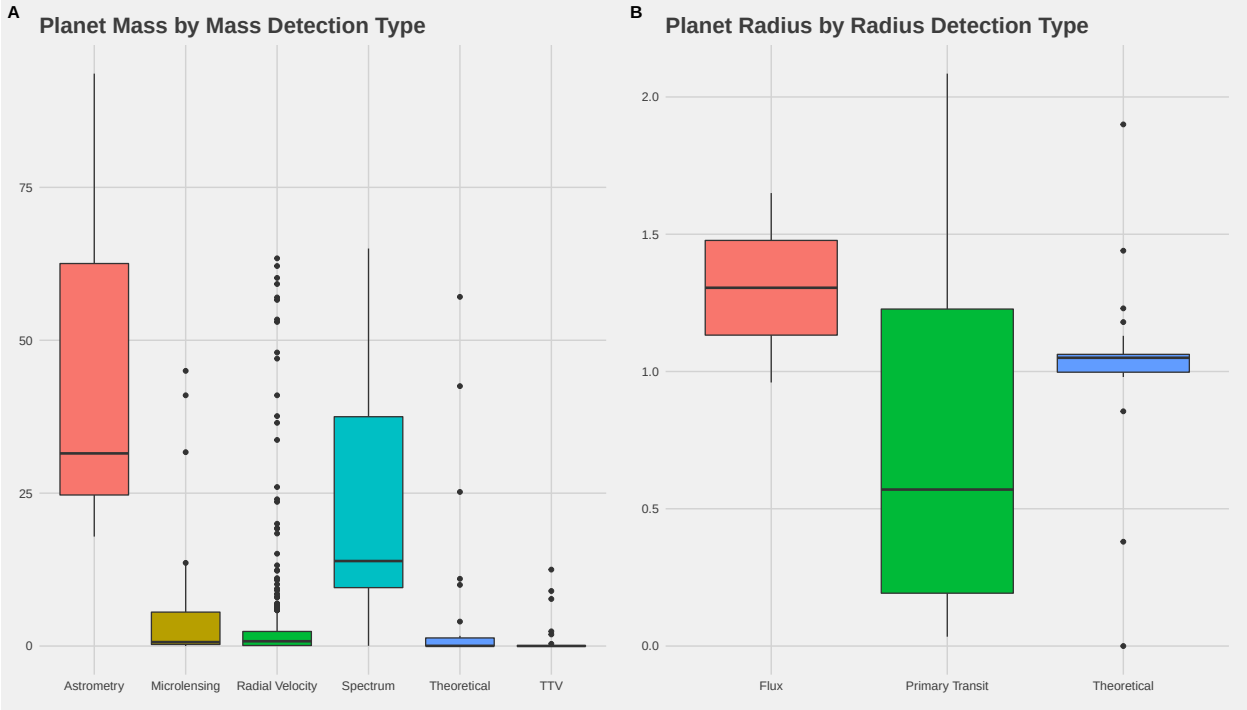

Figure 6: Mass, Radius Detection Methods by Year

Figure 6 shows the distribution of mass and radius of exoplanets by different detection methods. We can see that the average mass for planets detected using Astrometry is the highest. The planets whose radius was detected using Flux had the highest average radius. Those with the radius detected by Primary Transit had the widest range.

2.5.3 Plots for host star

Figure 7 shows the distribution of four important stellar variables of the host star by detection year. We can see from Plot A and B that the trends in mass and radius are similar. Over the past four decades, there have been extensive studies of low-mass stars with very little metal in the Milky Way's halo (Beers, Preston, and Shectman (1985)). Based on Plot C and D, more metal poor stars and cooler stars are detected in the recent decade with the advanced detection tools. An important aspect for habitability is that the host star must exist long enough for life to possibly evolve on the planet. Since cooler stars have longer lifespans than hotter stars, it is more likely for them to develop ecological niches.

2.5.4 Correlation Matrix Diagram

Figure 8 shows the correlation matrix diagram of the important variables we select to be applied to the model. Each of the cells in the table shows the linear correlation between two variables. -1 indicates a perfectly negative linear correlation between two variables, 0 indicates no linear correlation between two variables, and 1 indicates a perfectly positive linear correlation between two variables. From the corrgram we see that the mass and radius of host star is the strongest linear correlation (0.70), while the correlation between mass and radius of exoplanet is very low (0.15). In addition, the planet has a negative correlation to the host star's mass.

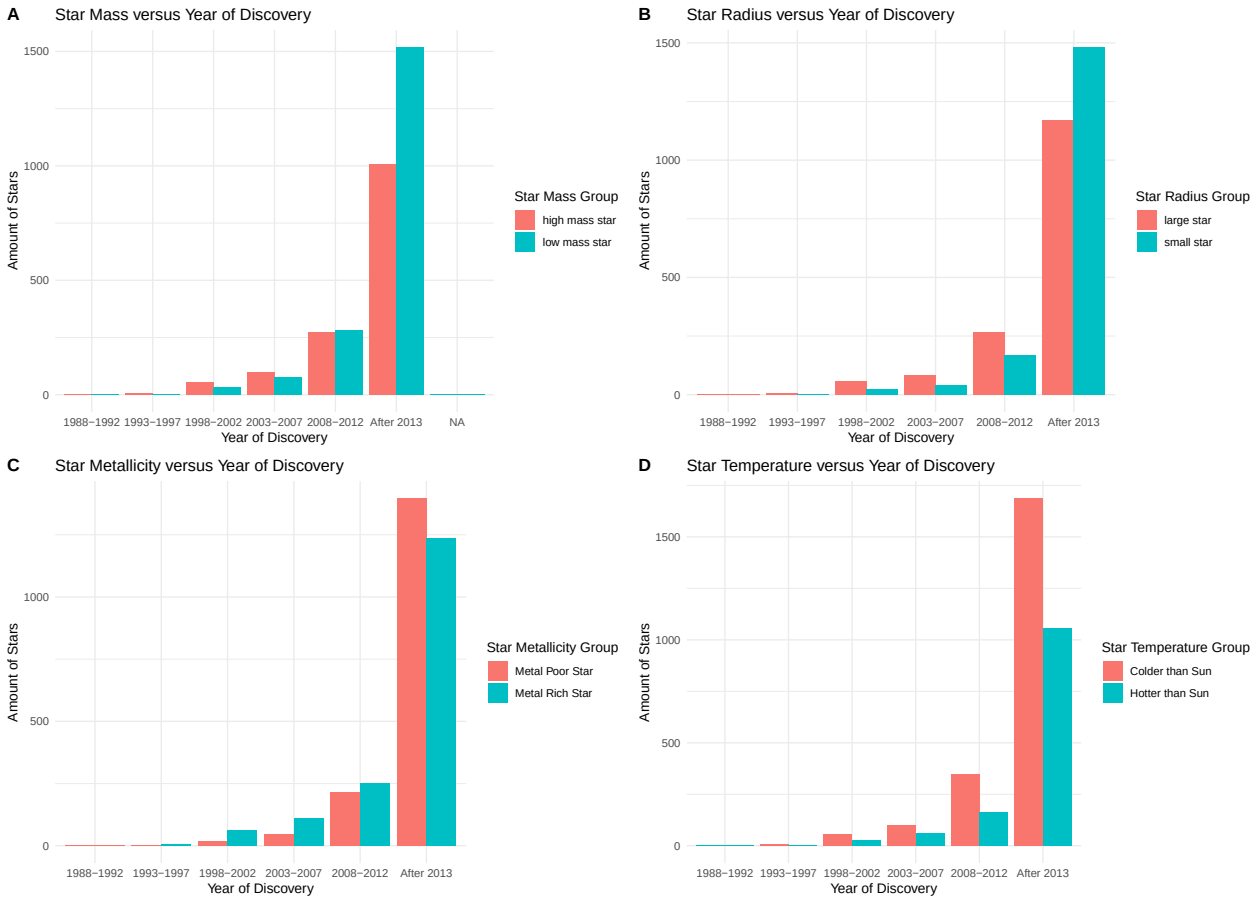

Figure 7: Mass, Radius, Metallicity and Temperature by Year

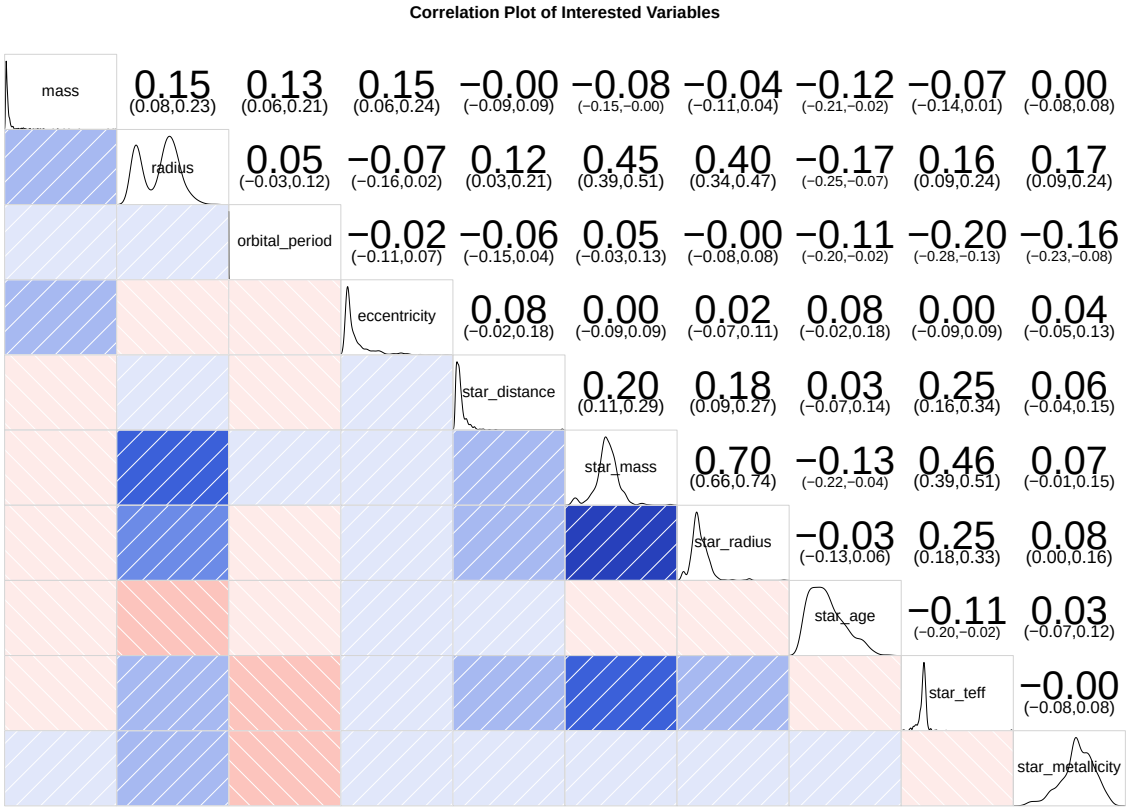

Figure 8: Correlation Matrix Diagram

3 Model

3.1 General Linear Regression Model

We start with the full linear regression model describe by the equation:

$$M_P = \beta_0 + \beta_1 MS + \beta_2 RS + \beta_3 DS + \beta_4 ME + \beta_5 TS + \epsilon_i$$

where the respondent variable Mp is the mass of exoplanet. We choose the mass as the respondent variable because the mass of an exoplanet can determine whether it is possible for it to support an atmosphere. On the one hand, if the mass is too low, low gravity would lead to no force to hold an atmosphere. Life is not possible to exist and evolve without atmosphere. The β_0 is the intercept, $\beta_1 \sim \beta_5$ are the coefficient estimates of the effect of mass, radius, distance, metallicity, temperature of the host star on the mass of exoplanet, and ϵ_i is the unknown error, the part of Y the regression model is unable to explain. This model gives a general idea of the relationship between mass of exoplanet and key features of its host star. However, it needs improvements in some discrepancies detected in the basic assumptions.

The result of the linear model indicates that two of the variables: distance between planet and star and temperature of the star has coefficients approximate to 0, which means the linear relationship between them and the response variable is very low. To check that whether these two variables are needed, we perform a series of model comparison tests between the model with and without the two variables. The null hypothesis of the test is that the model without these two variables explains the data as well as the model with them. Since the p-value > 0.05, we fail to reject our null hypothesis and we can remove the two predictors from our model in further analysis. The reduced model is:

$$M_P = \beta_0 + \beta_1 MS + \beta_2 RS + \beta_3 ME + \epsilon_i$$

3.1.1 Assumption check

Linear regression makes several assumptions about the data, which includes :

- Linearity of the data: The relationship between the predictor variable (x) and the response variable (y) is assumed to be linear.
- Normality of residuals: The residual errors are assumed to be normally distributed.
- Homogeneity of residuals variance: The residuals are assumed to have a constant variance.
- Independence of residuals error terms.

All these assumptions can be checked by visualizing the residual errors with diagnostic plots.

The diagnostic plots in (Figure 9) show residuals in four different ways:

1. Residuals vs Fitted. The plot is used to check the linear relationship assumptions. We can see there is a horizontal line without distinct patterns is an indication for a linear relationship, which is good.
2. Normal Q-Q. The plot is used to examine whether the residuals are normally distributed. Since the residuals points follow the straight dashed line, this is also satisfied.
3. Scale-Location (or Spread-Location). The plot is used to check the homogeneity of variance of the residuals. We can see that the horizontal line with equally spread points is a good indication of homoscedasticity.
4. Residuals vs Leverage. The plot is used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis.

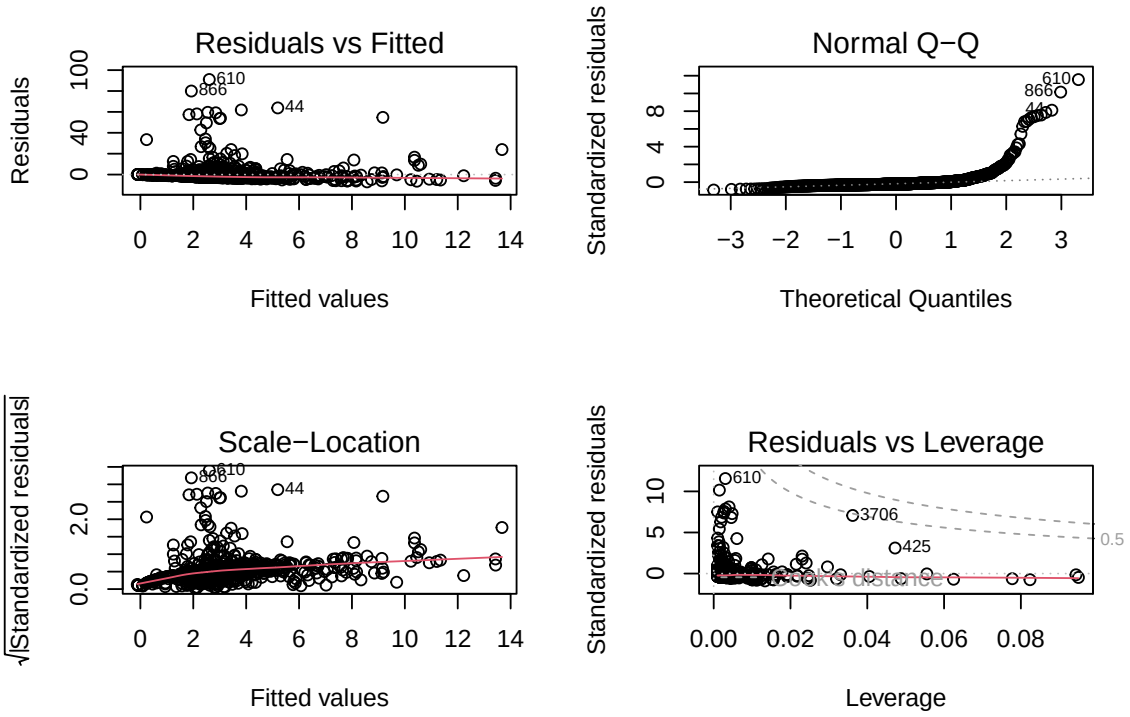

Figure 9: Regression Model Diagnostics

3.2 Logistic Regression Model: Binary Dependent Variables

Our logistic regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 MS + \beta_2 RS + \beta_3 ME + (1|discovered) + \epsilon_i$$

Logistic Regression brings great benefits when building model for the data since it does not make all the assumptions that Linear Regression does. For instance, it does not require a linear relationship between the dependent and independent variable. In addition, the cleaned dataset consists of over 3000 observations, which is a considerably large sample size in order for the Logistic Regression to perform well. The output of this logistic model gives the probability of whether or not the planet is in higher mass group(i.e. greater than Jupiter mass). The response in the model is the probability that a planet in in the higher mass group. A generalized model is used with random effect variable as discovery year as the response variable is binary. The predictor variables are the mass, radius and metallicity of the host star.

3.3 Verification of the Linear Regression Assumptions

After setting up the model, we check the assumptions that whether a star is in higher or lower mass group is binary and also all observations are independent of each other. - Linearity: The model passes all the Ramsey tests for linearity

- Multicollineality: There is no correlation among the independent variables
- Normality: The value of the Jarque-Bera statistic concludes that the residuals are normally distributed.
- Homogeneity: The model already includes the effect of a dummy variable.

A stepwise AIC was performed to check there is no multicollinearity in the model, and also helps remove the highly correlated variables. With Figure 10, we checked that the predicted values of the response variable in the model is binary. The figure shows that their is an "S" shape from 0 to 1, indicating all the possible

values are between 0 and 1 and probabilities does not increase linearly. We do not use the residual plots as for linear regression model because they are not helpful in the binary response variable case.

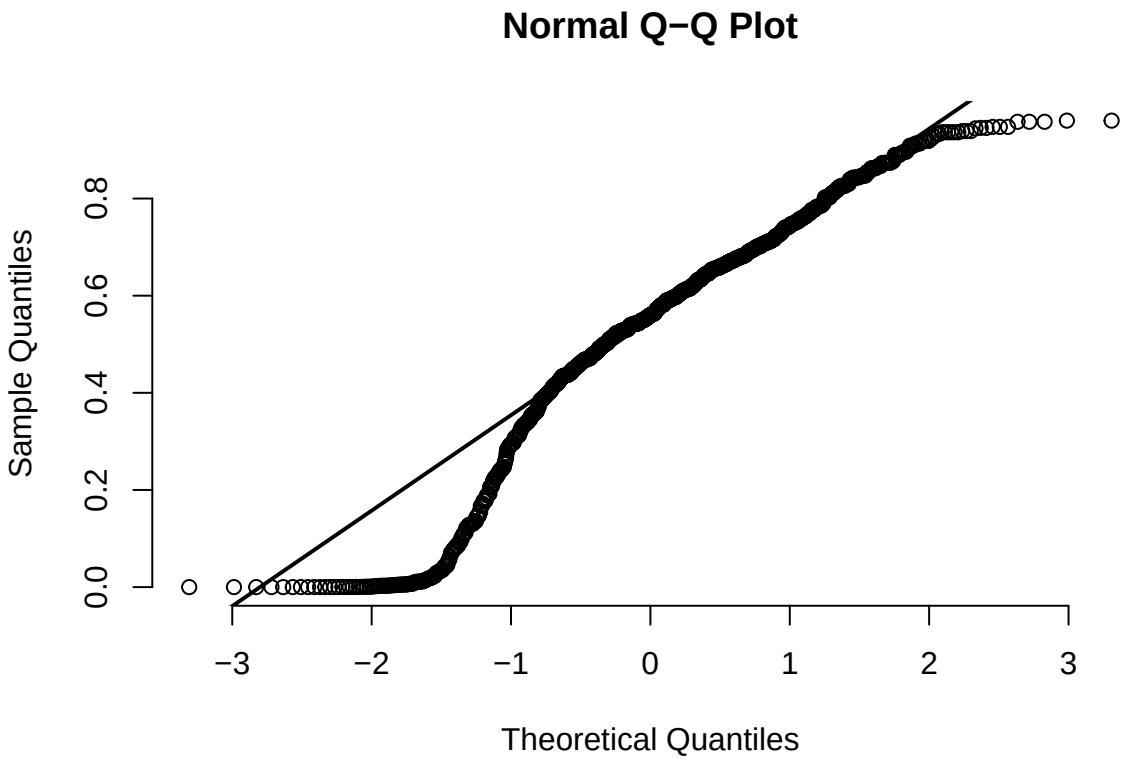

Figure 10: Binary Fitted Value Probabilities of Two Models

4 Results

Table 1: Number of detections and Proportions

Mass Group of Planet	Number of Detections	Percentage
high mass planet	776	0.21
low mass planet	810	0.22

Table 1 shows the number of detection of high and low mass exoplanet and their percentage in the dataset. We can notice that the proportion of each group is very similar, with low mass groups slightly more than high mass group.

4.1 Linear Model

$$M_P = -0.30 + 2.61MS + 0.19RS - 0.65ME$$

Table 2: Estimates of the Model Coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2959	0.7257	-0.4077	0.6836	-1.7198	1.1281
star_mass	2.6120	0.6765	3.8613	0.0001	1.2847	3.9393
star_radius	0.1857	0.0533	3.4842	0.0005	0.0811	0.2903
star_metallicity	-0.6484	1.1405	-0.5685	0.5698	-2.8863	1.5895

Table 2 shows the estimates of the model coefficients of the linear model. We can see that the mass and radius of the host star are very significant indicators of the mass of exoplanet, while metallicity does not have a significant p-value. This may be due to the limitation of linear regression model.

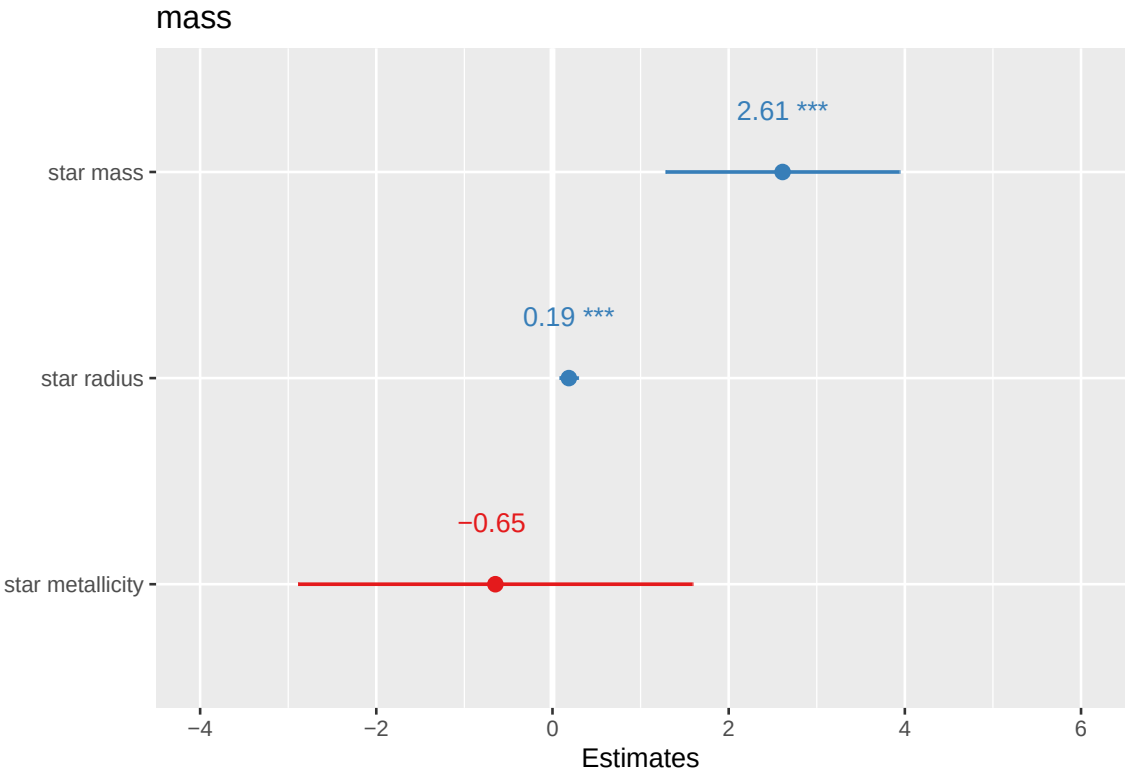

Figure 11: Estimates of P-Values of Variables in the Model

Figure 11 visualizes the p-values of three indicators.

4.2 Logistic Model

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 2.93 - 2.02MS - 0.33RS - 1.07ME$$

Table 3: Estimates of the Model Coefficients

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.9334	0.3088	9.4997	0.0000	2.3448	3.5558
star_mass	-2.0180	0.3199	-6.3089	0.0000	-2.6529	-1.3986
star_radius	-0.3302	0.0730	-4.5220	0.0000	-0.4905	-0.2044
star_metallicity	-1.0668	0.3441	-3.1001	0.0019	-1.7475	-0.3972
1 | discoveredTRUE	NA	NA	NA	NA	NA	NA

Table 3 shows the estimates of the model coefficients of the logistic model. We can see that all three predictors of the host star are very significant indicators of the mass of exoplanet.

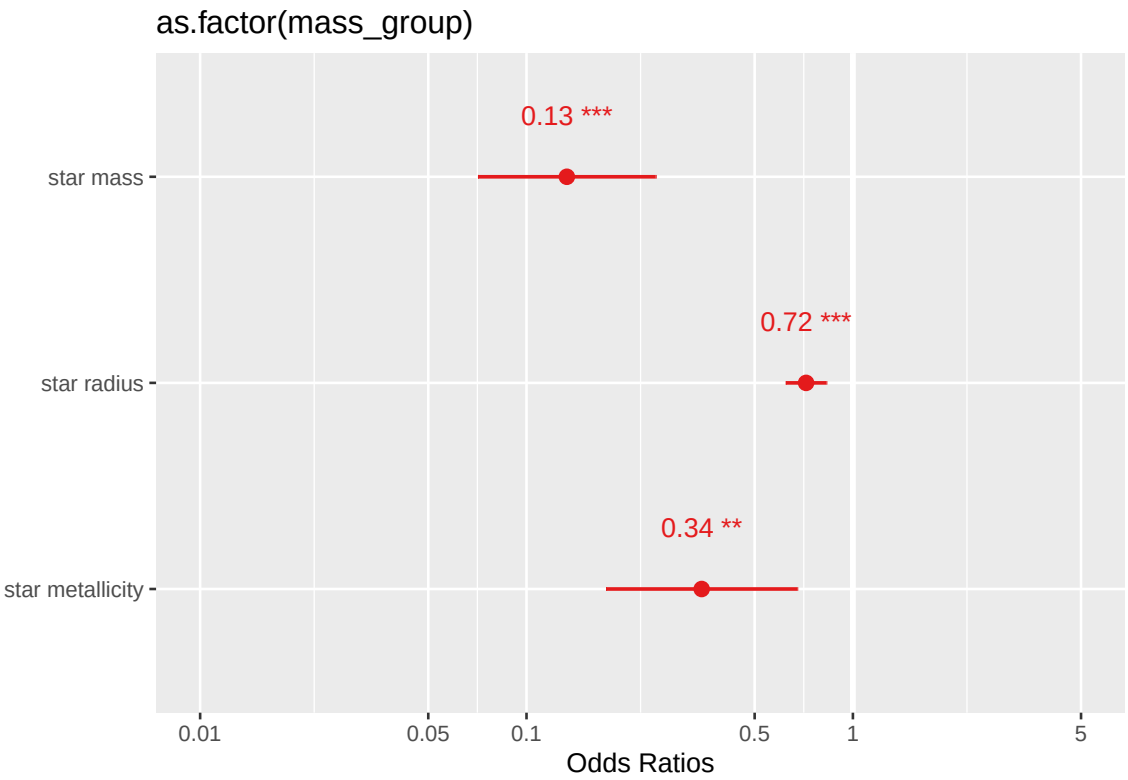

Figure 12: Estimates of P-Values of Variables in the Model

Figure 12 visualizes the p-values of three indicators.

4.3 Fit & evaluate models with training dataset

Table 4: Training and Testing Model Predictions Compared to True Observations

Prediction	Truth	Truth.
	High Mass Planet	Low Mass Planet
High Mass Planet	73	46
Low Mass Planet	32	110

(Table 4) shows the summary table of how well our model performance based on the training and testing datasets compared withthe true observations. The table shows both the number of predicted values for each

mass group and the true value for detection. The aim is to analyze whether our model can be generally applied to the model for all exoplanets. We see that the proportion of correct predictions is 0.7, with the proportion of correctly predicted that the planet is in high mass group is 0.69 and the proportion of correctly predicted that the planet is in low mass group is 0.70. This indicates that our model is likely to have the same accuracy when predicting whether the planet is in high or low mass group. Overall, the performance of our model is good but can be improved in the future.

5 Discussion:

5.1 Interpretation of Results & Key Findings

For the final logistic model, the estimated coefficients for T_assignment was positive and the p-value was very small (i.e. < 0.05), which indicated that the mass of exoplanet is significantly affected by the three predictors. Also, the performance of our final model is good when applying to test dataset. For the model for Language, although the p-value of metallicity was larger than 0.05, its estimate for T_assignment was positive and the variable turns our to be significant in logistic model. Thus, I could conclude that mass of exoplanet is increases with increase in mass and radius of host star, as well as decrease in metallicity of host star.

Our model shows that the three factors of the host star that best predict the mass of the exoplanet are mass, radius and metallicity. In 2018, a research on planetary mass and stellar radius relationship (Jiang and Zhu (2018)) also shows that there is a significant relationship between the stellar radius and the mass of the exoplanets. In addition, (Thorngren et al. (2016)) showed that the accumulation of heavy elements in the giant planets is negatively correlated with the metallicity of the host star, suggesting that the mass of the giant planets decreases with the increasing metallicity of the star. These literature supports the coefficients in our model.

5.2 Limitations

Based on our extensive statistical analysis, we proposed a logistic model to predict the whether mass of an exoplanet is greater or smaller than mass of Jupiter. In this work, we are assuming that the mass is the quantitative variable which can be representative for the whole physical characteristic of habitability of the exoplanet. However, this is not completely true. Whether an exoplanet is able to maintain liquid water on its surface depends on the complex interplay of features of the planet, host star and planetary system during the planet's life cycle. While planet habitability depends primarily on the type of star and mass of itself, many other factors also influence habitability (Meadows and Barnes (1970)). As a result, more qualitative information should be taken into account for a more comprehensive model.

Missing data has been a common but challenging issue in this study, which may lead to biased or inefficient inferences. In the Data Section, we analyzed and visualized the missing data. Although it is common to have large amount of missing data in astronomy datasets due to limitation of technology. Multiple imputation is one of the modern techniques for handling missing data and has very broad applications. In this work, due to limit of time we did not apply any bayesian inference to take full advantage of the dataset, so the model may not be applicable to all the exoplanet data.

During data cleaning, we introduced several grouping factors of parameters including mass, radius, metallicity and temperature. Generally, the criteria for grouping those factors is to compare with 1 as the unit is either Jupiter or the sun. Since the high and low mass and radius groups are classified simply by comparing that of Jupiter or Sun, the result may not be very accurate. However, there is no general applicable criteria for classifying whether the mass of an exoplanet greater than a specific value should be identified as high mass star, it is still acceptable to compare it with that of Jupiter at current stage of studies.

5.3 Future Steps

In the future, Bayesian inference including prior construction, posterior computation, model comparison can be applied for missing data to take full advantage of the exoplanet catalog. With the growth of the explanet

catalog, mixed models can be applied and updated accordingly. We could also include more predictor variables in the full model such as component and density of exoplanet to build a more comprehensive model. The Kepler team has been updating striter definition of what constitutes a habitable zone in their new catalog to take the warming effects of planetary atmospheres into account, which may lead to the change in orbital periods of a star. Finally, studies focus on the confirmed list of candidate habitable exoplanet should be conducted to research on the correlation between their properties, which saves time for future studies on exoplanet and finding true alien earth.

6 Appendix

6.1 Glossary of Astronomy Terms (alphabetical order)**

Exoplanet. An exoplanet is any planet beyond our solar system. Most orbit other stars, but free-floating exoplanets, called rogue planets, orbit the galactic center and are untethered to any star.

Habitability. Habitability is a quality of being good enough to live in.

Habitable Zone. The distance from a star at which liquid water could exist on orbiting planets' surfaces. Habitable zones are also known as Goldilocks' zones, where conditions might be just right – neither too hot nor too cold – for life.

Host Star. The star around which a particular planet, brown dwarf, or lesser object revolves. Also known as the central star or primary.

Metallicity. The fraction of heavier elements is called "metallicity", and it provides one way for astronomers to measure when a particular star or nebula formed. A low metallicity star must have formed a long time ago, while a higher metallicity star is of more recent vintage.

Orbital Period. The orbital period (also revolution period) is the amount of time a given astronomical object takes to complete one orbit around another object. In astronomy, it usually applies to planets or asteroids orbiting the Sun, moons orbiting planets, exoplanets orbiting other stars, or binary stars.

6.2 Datasheet for Dataset

6.2.1 Motivation

1. **For what purpose was the dataset created?**

This catalog is a working tool providing all the latest detections and data announced by professional astronomers, useful to facilitate progress in exoplanetology. Given the heterogeneity of observational papers, a uniform catalog (with uniform degree of credibility of planets) is impossible. Therefore, ultimately, researchers willing to make a quantitative, scientific, use of the catalog can make their own judgement on the likelihood of data and detections.

2. **Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)**?

This dataset was created by the portal exoplanet.eu of The Extrasolar Planets Encyclopaedia, and edited by Dingding Wang of the University of Toronto.

3. **What support was needed to make this dataset?**

The project was not funded.

4. **Any other comments?**

No.

6.2.2 Composition

1. **What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)?**

Each row of the main dataset is an exoplanet, and contains the information about that specific planet.

2. **How many instances are there in total (of each type, if appropriate)?**

There are about 3732 instances in the original datset.

3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

The dataset contains all the up to date data of exoplanet.

4. **What data does each instance consist of?**

Each instance consists of stellar and planetary parameters.

5. **Is there a label or target associated with each instance?**

Yes, there is a special name assigned with each planet.

6. **Is any information missing from individual instances?**

There is a lot of missing information due to the technical difficulty in collecting astronomial data.

7. **Are relationships between individual instances made explicit (e.g. users' movie ratings, social network links)?**

Yes.

8. **Are there recommended data splits (e.g. training, development/validation, testing)?**

No.

9. **Are there any errors, sources of noise, or redundancies in the dataset?**

No information.

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)?**

It is taken from latest published papers or professional preprints and conferences and first-hand updated data on professional websites.

11. **Does the dataset contain data that might be considered confidential (e.g. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

No.

12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

No.

13. **Does the dataset relate to people?**

No.

14. **Does the dataset identify any subpopulations (e.g. by age, gender)?**

No.

15. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

No.

16. **Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

No.

17. **Any other comments?**

No.

6.2.3 Collection

1. **How was the data associated with each instance acquired?**

Through latest published papers or professional preprints and conferences and first-hand updated data on professional websites.

2. **What mechanisms or procedures were used to collect the data (e.g. hardware apparatus or sensor, manual human curation, software program, software API)?**

The basic criterion is the mass limit: 60 Jupiter mass.

3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?**

N/A

4. **Who was involved in the data collection process (e.g. students, crowdworkers, contractors) and how were they compensated (e.g. how much were crowdworkers paid)?**

Researchers, Astronomers. Pay structure is unknown.

5. **Over what timeframe was the data collected?**

February 1995 to March 2018.

7. **Were any ethical review processes conducted (e.g. by an institutional review board)?**

N/A

8. **Does the dataset relate to people?**

No.

9. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g. websites)?**

N/A

10. **Were the individuals in question notified about the data collection?**

N/A

11. **Did the individuals in question consent to the collection and use of their data?**

N/A

12. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

N/A

13. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. a data protection impact analysis) been conducted?**

N/A

14. **Any other comments?**

N/A

6.2.4 Preprocessing / Cleaning / Labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g. discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

Yes, the data was pre-cleaned but missing values were not adjusted.

2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g. to support unanticipated future uses)?**

No.

3. **Is the software used to preprocess/clean/label the instances available?**

No.

4. **Any other comments?**

No.

6.2.5 Uses

1. **Has the dataset been used for any tasks already?**

Yes, relevant research and theory works have been conducted in the past years.

2. **Is there a repository that links to any or all papers or systems that use the dataset?**

Yes. LINK: http://exoplanet.eu/research/

3. **What (other) tasks could the dataset be used for?**

It could also be used for all kinds of research related to exoplanet.

4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

N/A

5. **Are there tasks for which the dataset should not be used?**

N/A

6. **Any other comments?**

N/A

6.2.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g. company, institution, organization) on behalf of which the dataset was created?**

No, the dataset is only available on the exoplanet.eu website.

2. **How will the dataset be distributed (e.g. tarball on website, API, GitHub)?**

N/A

3. **When will the dataset be distributed?**

The dataset is already available.

4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

N/A

5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

N/A

6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

N/A

7. **Any other comments?**

N/A

6.2.7 Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**

exoplanet Team.

2. **How can the owner/curator/manager of the dataset be contacted (e.g. email address)?**

Email address: `vo.exoplanet@obspm.fr`

3. **Is there an erratum?**

No, there is no erratum.

4. **Will the dataset be updated (e.g. to correct labeling errors, add new instances, delete instances)?**

No, the dataset is final.

5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g. were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

No, there is no limit.

6. **Will older versions of the dataset continue to be supported/hosted/maintained?**

No, the dataset is updated daily

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

No.

8. **Any other comments?**

No.

References

Alibert, Y. 2013. "On the Radius of Habitable Planets." *Astronomy &Amp; Astrophysics*. EDP Sciences. https://www.aanda.org/articles/aa/full_html/2014/01/aa22293-13/aa22293-13.html.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Beers, T. C., G. W. Preston, and S. A. Shectman. 1985. "A Search for Stars of Very Low Metal Abundance. i." *NASA/ADS*. https://ui.adsabs.harvard.edu/abs/1985AJ.....90.2089B/abstract.

Brennan, Pat. 2021. "The Habitable Zone." *NASA*. NASA. https://exoplanets.nasa.gov/search-for-life/habitable-zone/.

———. 2022. "What's a Transit? – Exoplanet Exploration: Planets Beyond Our Solar System." *NASA*. NASA. https://exoplanets.nasa.gov/faq/31/whats-a-transit/#:~:text=Most%20known%20exoplanets%20have%20been,between%20us%20and%20the%20Sun.

Grace-Martin, Karen, Martins Ahmed says, Martins Ahmed, Harold Gomes says, Harold Gomes, Jeremy Taylor says, Jeremy Taylor, Karen says, and Karen. 2021. *The Analysis Factor*. https://www.theanalysisfactor.com/mar-and-mcar-missing-data/.

Howell, Elizabeth, and Ailsa Harvey. 2022. "The 10 Most Earth-Like Exoplanets." *Space.com*. Space. https://www.space.com/30172-six-most-earth-like-alien-planets.html.

Jiang, Jonathan H., and Sheldon Zhu. 2018. "A Planetary Mass and Stellar Radius Relationship for Exoplanets Orbiting Red Giants." *Research Notes of the AAS*. IOP Publishing. https://iopscience.iop.org/article/10.3847/2515-5172/aae48c.

Kane, Stephen, and Dawn Gelino. 2013. "Title: The Habitable Zone and Extreme Planetary Orbits - Arxiv.org." https://arxiv.org/pdf/1205.2429.pdf.

Martínez-Gómez, E., and G. Babu. 2009. "A Statistical Model for the Relation Between Exoplanets and Their Host Stars," August.

Meadows, Victoria S., and Rory K. Barnes. 1970. "Factors Affecting Exoplanet Habitability." *NASA/ADS*. https://ui.adsabs.harvard.edu/abs/2018haex.bookE..57M/abstract#:~:text=Processes%20which%20can%20modify%20a,minor%20bodies%3B%20and%20galactic%20phenomena.

NASA. 2001. "NASA's Kepler Mission Confirms Its First Planet in Habitable Zone of Sun-Like Star." *NASA*. NASA. https://www.nasa.gov/mission_pages/kepler/news/kepscicon-briefing.html#:~:text=%22The%20tremendous%20growth%20in%20the,University%20in%20San%20Jose%2C%20Calif.

———. 2021. "NASA Exoplanet Archive." *NASA*. NASA. https://exoplanetarchive.ipac.caltech.edu/docs/intro.html.

Pichara, Karim, and Pavlos Protopapas. 2013. "AUTOMATIC CLASSIFICATION OF VARIABLE STARS IN CATALOGS WITH MISSING DATA." *The Astrophysical Journal*. IOP Publishing. https://iopscience.iop.org/article/10.1088/0004-637X/777/2/83#apj484836r23.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Takahashi, Masayoshi. 2017. *Data Science Journal*. Ubiquity Press. https://datascience.codata.org/articles/10.5334/dsj-2017-037/#5-traditional-methods-of-handling-missing-data.

TEAM, exoplanet. 2022. "The Extrasolar Planets Encyclopaedia." *Exoplanet.eu*. http://exoplanet.eu/.

Thorngren, Daniel P., Jonathan J. Fortney, Ruth A. Murray-Clay, and Eric D. Lopez. 2016. "THE MASS–METALLICITY RELATION FOR GIANT PLANETS." *The Astrophysical Journal*. IOP Publishing. https://iopscience.iop.org/article/10.3847/0004-637X/831/1/64.

Tierney, Nicholas. 2017. "Visdat: Visualising Whole Data Frames." *JOSS* 2 (16): 355. https://doi.org/10.21105/joss.00355.

Wenz. 2016. "This Map Shows Where in the Sky You Might Find Habitable Exoplanets." *Astronomy.com*. Astronomy. https://astronomy.com/news/2016/05/this-map-shows-where-in-the-sky-you-might-find-habitable-exoplanets.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Tidyr: Tidy Messy Data*.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data*

Manipulation.
Wright, Kevin. 2021. *Corrgram: Plot a Correlogram*. https://kwstat.github.io/corrgram/.
Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.