Visual Analytics on Exoplanet Dataset

Riccardo Viviano

October 28, 2024

1 Introduction

In the current report we explore a dataset of exoplanets sourced from NASA's Exoplanet Catalog https://science.nasa.gov/exoplanets/exoplanet-catalog/. The dataset was filtered and visualized to highlight key characteristics of the exoplanets, such as their mass, radius, discovery date, distance from their star, star type, distance from earth. planet type, and discovery method. The goal is to provide insights into the distribution and relationships between these features using various visualization techniques:

Scatterplot:
used to visualize the relationship between the exoplanets' features. Each point represents an exoplanet, and the scatterplot allows for the identification of any correlation between their characteristcs. The scatterplot also enables the differentiation between different types of planets by using color-coding.

Parallel coordinate:
used to explore the relationships between multiple features simultaneously. Each feature is represented by a vertical axis, and each exoplanet is represented by a line that intersects these axes according to its feature values.

Bubbleplot:
designed and implemented in order to be highly customizable and flexible in displaying the data.

Column bars:
employed to visualize categorical data such as the number of exoplanets discovered using the different techniques. It has been implemented considering the number of features that are interchanchable along the axis.

The primary users of these visualizations are researchers in the field of exoplanet discovery. These researchers are focused on understanding the relationship between the characteristics of the exoplanets the characteristics of their stars and the methods used to discover them. By analyzing these visualizations, they can identify patterns that suggest which discovery techniques are most effective for certain types of exoplanets, which type of planet forms around which type of star, etc. The code is available on https://github.com/VivianoRiccardo/VA-project with a live demo on https://visual-analytics.netlify.app/

2 Related Works

The discovery of new exoplanets is a challenging endeavor that numerous studies have sought to advance by proposing innovative techniques aimed at improving detection methods. [3] describes an exhaustive list of techniques used to detect exoplanets with indirect methods such as radial velocity method, transmit photometry method, microlensing method and direct imaging. In [4] they analyze the existing discovery methods plotting graphs in relation to the type of the technique used, the revolution period of the exoplanet and its mass using the NASA's Exoplanet Catalog. In [1] they presented a data visualization tool for astronomers offering different plotting views where we took inspiration from. The proposed website, however, lacks of a multiple plots display which we integrated in our demo.

3 Dataset

The Data used is a subset of the Nasa Exoplanet Archive [1] directly taken from https://science.nasa.gov/exoplanets/exoplanet-catalog/ through a custom script used to scrap the website.

3.1 Data Preprocessing

In the Data we deal with we get through the following features:

- **Planet Radius**
- **Planet Mass**
- **Discovery Date**
- **Distance from own star**
- **Planet Type**
- **Discovery Method**
- **Distance from Earth**
- **Star Type**
- **Solar System's Number of Planets**

The dataset used in this analysis focuses on planets discovered after 2011. Additionally, any entries lacking key feature information, as previously outlined, were excluded, resulting in the removal of 664 planets. The planets were further categorized into four types based on their characteristics: *Earth-like*, *Super-Earth*, *Neptune-like*, and *Gas giant*. The *Unknown* planet type, consisting of three exoplanets, was also excluded from the analysis. Furthermore, the planetary radius and mass were normalized relative to Jupiter's radius and mass, and recorded as radius (J) and mass (J). The distance from the host star was expressed in astronomical units, while the distance from the Earth was expressed in Parsec. Star types depends on the temperature expressed in kelvin (K). Below is presented a table with the features expressed in the demo project:

Radius (J)	Planet Type	Star Distance (AU)	Mass (J)	Discovery Date	Discovery Method	Earth Distance (Parsec)	N.Planets-Solar System	Star Type
Number	Gas Giant Neptune-Like Super Earth Terrestrial	Number	Number	2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024	Transit Radial Velocity Microlensing Imaging Eclipse Timing Variations Pulsar Timing Pulsation Timing Variations Disk Kinematics Orbital Brightness Modulation	Number	Number	(Y) 80–500 K (T) 500–1,500 K (M) 2,400–3,700 K (K) 3,700–5,200 K (G) 5,200–6,000 K (F) 6,000–7,500 K (A) 7,500–10,000 K (B) 10,000–30,000 K (O) 30,000–50,000 K

Table 1: Project Features

4 Visualization Techniques

4.1 Scatter Plot

The scatter plot showed in the picture below is computed using a technique called multidimensional scaling (MDS). MDS is a statistical technique widely used for dimensionality reduction, which simplifies the visualization of complex datasets. Specifically, MDS allows for the representation of high-dimensional data in a lower-dimensional space (commonly 2 or 3 dimensions) while preserving the relationships between the original data points as accurately as possible. It does so by converting dissimilarities between pairs of objects into distances between points in the reduced-dimensional space. This facilitates easier interpretation and visualization, particularly in situations where the original data may consist of numerous variables that are difficult to analyze simultaneously.

In this case, we apply MDS to our dataset, where the key features include the type of the planet, its radius, mass, and the distance of the planet from its host star. The aim of using MDS in this context is to transform the complex relationships between the planets—based on these features—into a two-dimensional scatter plot, where the distance between points reflects the degree of dissimilarity between the corresponding exoplanets. Planets that appear closer together in the plot have more similar attributes, while those farther apart are more dissimilar in terms of their type, size, mass, or proximity to their star.

Our MDS is inspired by [2] where a force direct approach technique is used to compute the distance matrix trying to minimize a loss. The loss function used here is stress (or a form of strain), which is essentially a squared error loss between the pairwise distances in the original space (target) and the corresponding distances in the reduced space (predicted). The distance matrix for the MDS technique has been computed following the method in [6], where the distance for non categorical data is computed as the euclidian distance and for the categorical data the distance is computed introducing a planet type value scale showed below as well as a star type value scale:

Type 1	Type 2	Distance
Gas Giant	Gas Giant	0
Gas Giant	Neptune-Like	0.33
Gas Giant	Super-Earth	0.66
Gas Giant	Terrestrial	1
Neptune-Like	Neptune-Like	0
Neptune-Like	Super-Earth	0.33
Neptune-Like	Terrestrial	0.66
Super-Earth	Super-Earth	0
Super-Earth	Terrestrial	0.33
Terrestrial	Terrestrial	0

Table 2: Distances between planet types.

Type 1	Type 2	Distance
(Y) 80–500 K	(Y) 80–500 K	0
(Y) 80–500 K	(T) 500–1,500 K	0.125
(Y) 80–500 K	(M) 2,400–3,700 K	0.250
(Y) 80–500 K	(K) 3,700–5,200 K	0.375
(Y) 80–500 K	(G) 5,200–6,000 K	0.5
(Y) 80–500 K	(F) 6,000–7,500 K	0.625
(Y) 80–500 K	(A) 7,500–10,000 K	0.75
(Y) 80–500 K	(B) 10,000–30,000 K	0.875
(Y) 80–500 K	(O) 30,000–50,000 K	1
(T) 500–1,500 K	(T) 500–1,500 K	0
(T) 500–1,500 K	(M) 2,400–3,700 K	0.125
(T) 500–1,500 K	(K) 3,700–5,200 K	0.250
(T) 500–1,500 K	(G) 5,200–6,000 K	0.375
(T) 500–1,500 K	(F) 6,000–7,500 K	0.5
(T) 500–1,500 K	(A) 7,500–10,000 K	0.625
(T) 500–1,500 K	(B) 10,000–30,000 K	0.75
(T) 500–1,500 K	(O) 30,000–50,000 K	0.875
(M) 2,400–3,700 K	(M) 2,400–3,700 K	0
(M) 2,400–3,700 K	(K) 3,700–5,200 K	0.125
(M) 2,400–3,700 K	(G) 5,200–6,000 K	0.25
(M) 2,400–3,700 K	(F) 6,000–7,500 K	0.375
(M) 2,400–3,700 K	(A) 7,500–10,000 K	0.5
(M) 2,400–3,700 K	(B) 10,000–30,000 K	0.625
(M) 2,400–3,700 K	(O) 30,000–50,000 K	0.75
(K) 3,700–5,200 K	(K) 3,700–5,200 K	0
(K) 3,700–5,200 K	(G) 5,200–6,000 K	0.125
(K) 3,700–5,200 K	(F) 6,000–7,500 K	0.25
(K) 3,700–5,200 K	(A) 7,500–10,000 K	0.375
(K) 3,700–5,200 K	(B) 10,000–30,000 K	0.5
(K) 3,700–5,200 K	(O) 30,000–50,000 K	0.625
(G) 5,200–6,000 K	(G) 5,200–6,000 K	0
(G) 5,200–6,000 K	(F) 6,000–7,500 K	0.125
(G) 5,200–6,000 K	(A) 7,500–10,000 K	0.25
(G) 5,200–6,000 K	(B) 10,000–30,000 K	0.375
(G) 5,200–6,000 K	(O) 30,000–50,000 K	0.5
(F) 6,000–7,500 K	(F) 6,000–7,500 K	0
(F) 6,000–7,500 K	(A) 7,500–10,000 K	0.125
(F) 6,000–7,500 K	(B) 10,000–30,000 K	0.25
(F) 6,000–7,500 K	(O) 30,000–50,000 K	0.375
(A) 7,500–10,000 K	(A) 7,500–10,000 K	0
(A) 7,500–10,000 K	(B) 10,000–30,000 K	0.125
(A) 7,500–10,000 K	(O) 30,000–50,000 K	0.25
(B) 10,000–30,000 K	(B) 10,000–30,000 K	0
(B) 10,000–30,000 K	(O) 30,000–50,000 K	0.125
(O) 30,000–50,000 K	(O) 30,000–50,000 K	0

Table 3: Distances between star types.

Neptune-like planets and Jupiter-like planets (also called in general gas giants) are both composed mainly of gases and ices, and they tend to have thick atmospheres made up mostly of hydrogen and helium. While Neptune is smaller and classified as an "ice giant" due to its higher concentration of ices (such as water, ammonia, and methane), both Neptune and Jupiter lack a solid surface like terrestrial planets and for these reason they are considered more similar all the planet type features. Terrestrial

planets instead are rocky, with solid surfaces and much smaller sizes compared to gas and ice giants. Super-Earths are a class of planets larger than Earth but smaller than Neptune, and their composition can vary—they can be rocky like terrestrial planets or have thick atmospheres, making them a hybrid between terrestrial and Neptune-like planets. Regarding the star types the difference is set according to the temperature of the stars. Each adjacient star type differs of 0.125 in increasing order. Before of computing the distance matrix a min max normalization has been applied to non categorical data to be in a range [0,1]

In the scatterplot is possible to run on the front-end the MDS technique allowing the selection and deselection of the features that must be taken into consideration during the computation through some checkboxs. Moreover, since the MDS is computed with a force direct approach, the value of the hyperparameters (learning rate and number of steps) can be decided for the computation. The data points are colored according to either the planet type, the discovery method or the star type. This selection is allowed through the select input on the top-right of the plot. The data can be clicked to be highlighted (the border will be colored by red) and it's possible to zoom in and zoom out in the plot. Also, the plot can be rescaled removing the labels that do not want to show up.

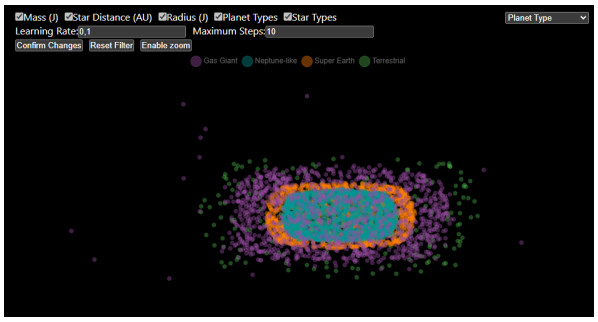

Figure 1: Scatter Plot

4.2 Parallel Coordinates

Parallel coordinates is an effective technique for visualizing complex, multidimensional datasets by arranging multiple parallel axes within a two-dimensional space. For our exoplanet dataset, each axis corresponds to a specific attribute, such as the planet's radius, mass, distance from its star, discovery date, discovery method, earth's distance, star type, or planet type (e.g., gas giant, Neptune-like, super-Earth, or terrestrial). The lines that pass through these axes represent individual data points or groups, such as planets of a specific type or discovered by a particular method, connecting values across each feature.

However, a common issue with this type of visualization is over-plotting, where many overlapping lines can obscure the ability to recognize trends, patterns, or key insights. To mitigate this, the chart incorporates a "brushing" function, enabling users to focus on and highlight specific data lines by interacting with the relevant axes, filtering out unnecessary details.

This plot is fundamental since the filter highlights also the data point plotted in the other charts

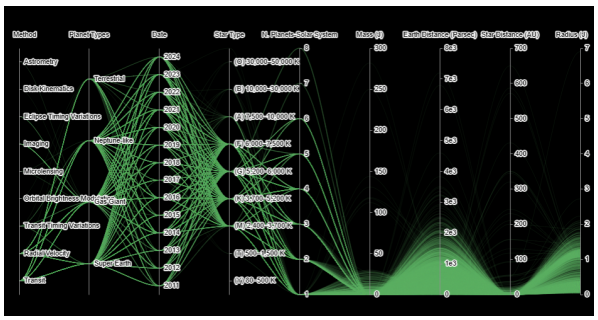

Figure 2: Parallel Coordinates

4.3 Bubble Plot

The demonstration features an interactive bubble plot where the color of each data point corresponds to a specific label based on the selected legend. Users can choose to categorize the data points by either the type of planet—options include gas giant, Neptune-like, terrestrial, or super-Earth—or by the type of star that the planet orbits, with nine distinct labels available for star classification. Alternatively, users may select a legend based on the discovery method employed for each planet.

The axes of the plot are interchangeable, allowing users to dynamically configure the x-axis and y-axis to represent various features, including planetary radius, planetary mass, distance from the host star, distance from Earth, and the number of planets discovered within the solar system. The size of the data points is also customizable, offering four different size categories based on the selected feature, which can be adjusted according to a specified range.

Additionally, the plot includes functionality for zooming in and out, which can be activated via a dedicated zoom button. A timeline is integrated at the bottom of the interface, spanning the years 2011 to 2024. When the play button is activated, the timeline progresses year by year, visually illustrating the data points relevant to each corresponding year.

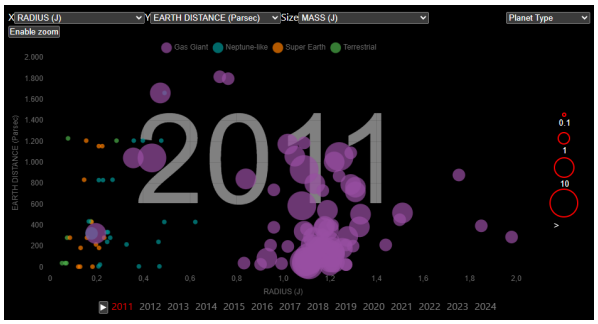

Figure 3: Bubble Plot

4.4 Histogram

The demonstration presents a histogram in which the colors of the data points correspond to specific labels determined by the user-selected legend. (planet type, star type, discovery method)

The x-axis of the histogram can be dynamically adjusted to display various features based on user preferences. Possible features include the radius of the planets, their mass, the distance of the planet from its host star, the distance from Earth, the number of planets discovered within the same solar system, and the date of discovery.

The y-axis can represent either the number of data points corresponding to each label or the percentage of data points relative to the total for each label. Additionally, users have the capability to zoom in and out of specific areas of the histogram by utilizing a designated zoom button, enhancing the analytical experience and allowing for detailed exploration of the data.

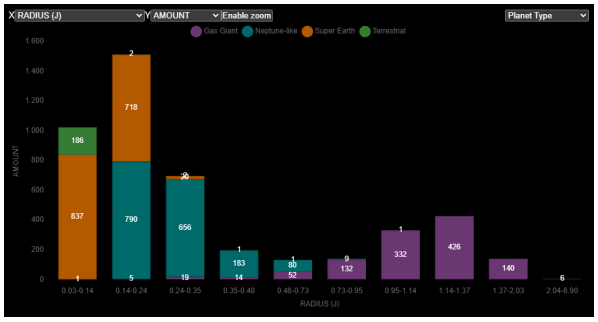

Figure 4: Histogram

5 Use Case Example

Let's assume our target user wants to know what type of discovering method is the most adequate to spot the planets with some specific features (For example high Earth Distance). In this case the first thing that the user can do is to highlight with a filter the datapoints with the highest earth distance on the parallel coordinates. The user could select a filter that goes from 3e3 to 8e3 Parsecs as showed in the figure below

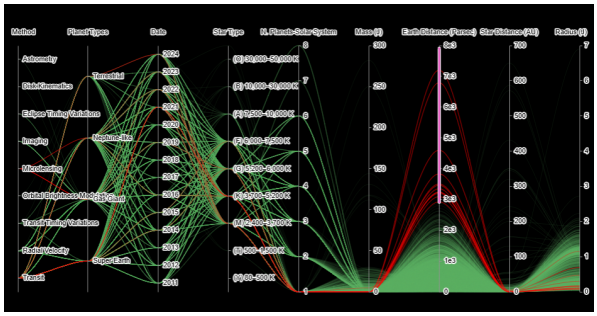

Figure 5: Use Case 1 Parallel Coordinates

According to the data highlighted seems that the discovering methods of these distant planets are only Microseling and Transit.

Since the Transit method seems to be the most exploited, the user can decide to go to the histogram chart and see how many exoplanets the Microlensing method has discovered as showed in the image below

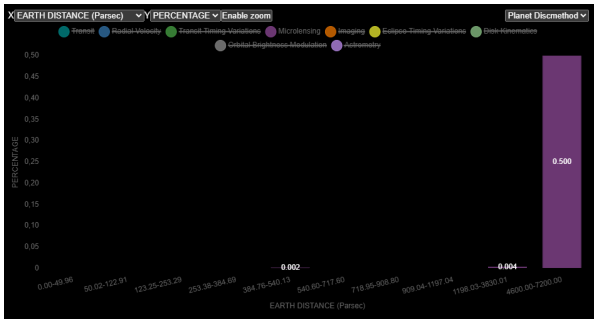

Figure 6: Use Case 1 Histogram Microlensing

In this case the user can decide to select as legend the discovering method, can decide to remove all the labels targeting all the discovering methods except for Microlensing to remove useless data on the chart and can select on the X-Axis the Earth distance and on the Y Axis the percentage. In this case is showed that 50% of the most distant planets (in the range of 4600-7600 parsecs from the Earth) have been discovered through the Microlensing method while other planets discovered with this method are in the half above the average for the earth distance. In this case seems that the microlensing method could be more adequate to discover more distant planets as already stated in [5]. However to have a confirmation we can remove the microlensing label and select the transit label as showed in the figure below:

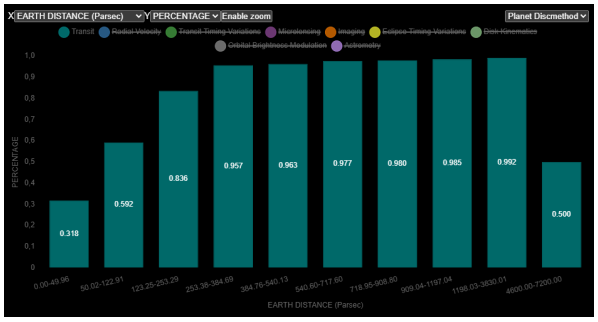

Figure 7: Use Case 1 Histogram Transit

In this case the user can see that the transit method can discover planets in every Earth Distance range, however even if the majority of discoveries have been done with the Transit method with an average of $\sim 89\%$,the amount of planets discovered in the highest range (4600-7600 parsecs) is shared with the microlensing method with only a 50% t. This can be a confirmation that since the few planet discoveries with the microlensing w.r.t. the transit the first one is more adequate in discoverying distant planets.

Also from the parallel coordinates seems that the only stars with these type of planets so far from Earth are those regarding the Stars in the temperature range 2400-7500, and seems also that these planets could have in common similar star distances and masses. The user can use the Scatter Plot to confirm or disprove this claim.

In the general plot we can see that, filtering the scatter plot according to the filters set on the parallel coordinates and setting the distance computation according to only the mass and the star distance, the points are really close togheter. Also zooming in and filtering according to the planet type we can see that these type of planets are common among Terrestrial, Neptune-Like and Super-Earth, but not among Gas-Giant planets. The user can check this also in the bubble plot setting the 2011-2024 data and filtering by Earth Distance, Star Distance and Mass on the X,Y axis and on the size parameter. Since among the exoplanets with huge mass and huge distance no one has a huge distance from Earth the user can conclude that the methods used for discovering distant exoplanets around a star will tend to have also these characteristics. This agrees with the common literature: since some of these distant planets are discovered with the transit method, for example, a planet that orbits close to its star blocks a larger percentage of the star's light (because its angular size compared to the star is larger), making the transit more detectable [5] (there is an increase in the dip, which is the decrease in the brightness of a star when a planet passes) and as the star's distance from us increases, the overall brightness we observe decreases. As a result, the small dip in brightness caused by a transit becomes harder to detect, since the light reaching us is more diffused and weaker overall. This means that with the transit method is harder to find distant planets, but as the planet is closer to its star it becomes esier.

Finally the User can conclude that the majority of distant planets with these discoverying methods will be Neptune-Like, Terrestrial and Super Earth and rarely Gas-Giant as stated previously according to the scatterplot.

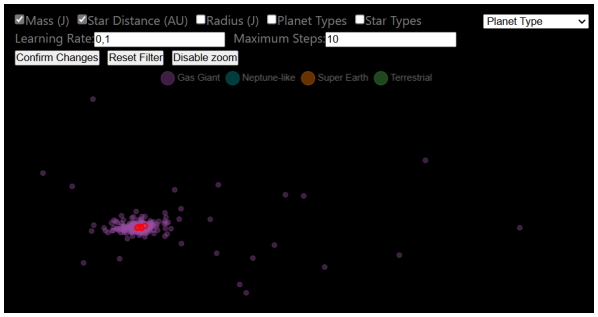

Figure 8: Use Case 1 Scatter Plot

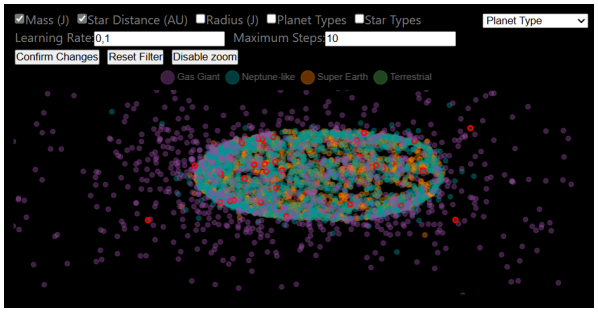

Figure 9: Use Case 1 Scatter Plot Super Earth

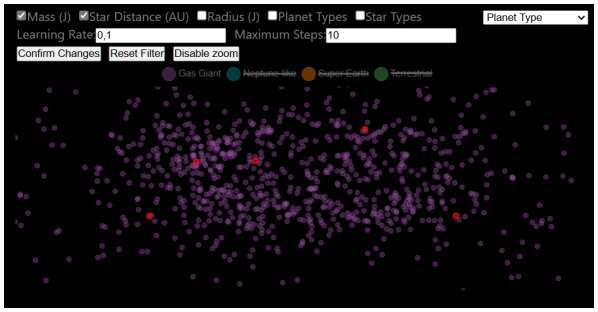

Figure 10: Use Case 1 Scatter Plot Gas Giant

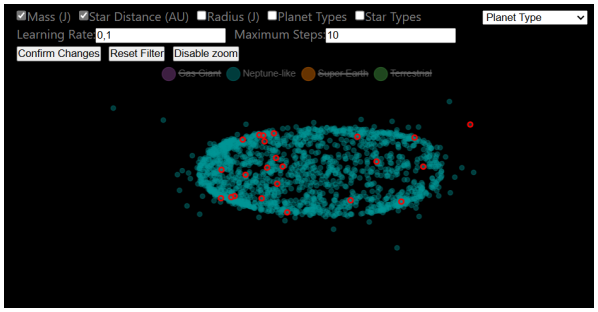

Figure 11: Use Case 1 Scatter Plot Neptune-Like

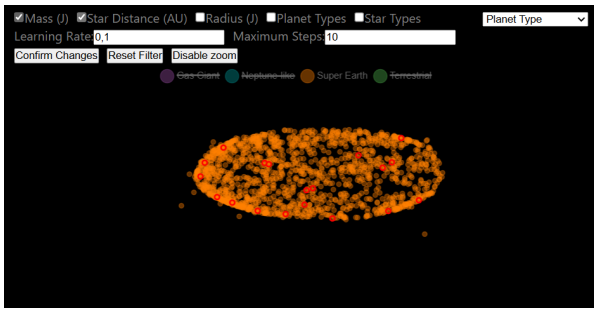

Figure 12: Use Case 1 Scatter Plot Terrestrial

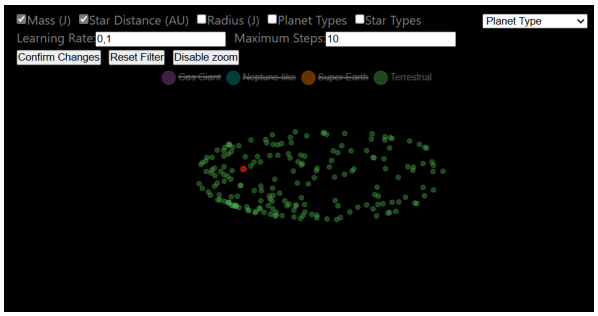

Figure 13: Use Case 1 Scatter Plot Terrestrial

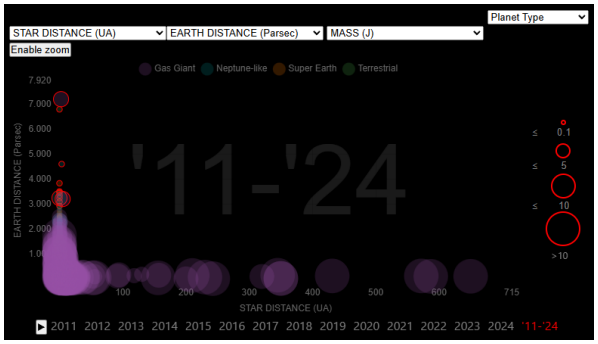

Figure 14: Use Case 1 Scatter Plot Terrestrial

References

[1] R. L. Akeson, X. Chen, D. Ciardi, M. Crane, J. Good, M. Harbut, E. Jackson, S. R. Kane, A. C. Laity, S. Leifer, M. Lynn, D. L. McElroy, M. Papin, P. Plavchan, S. V. Ramírez, R. Rey, K. von Braun, M. Wittman, M. Abajian, B. Ali, C. Beichman, A. Beekley, G. B. Berriman, S. Berukoff, G. Bryden, B. Chan, S. Groom, C. Lau, A. N. Payne, M. Regelson, M. Saucedo, M. Schmitz, J. Stauffer, P. Wyatt, and A. Zhang. The nasa exoplanet archive: Data and tools for exoplanet research. *Publications of the Astronomical Society of the Pacific*, 125(930):989–999, August 2013.

[2] Michael L Littman Nathaniel Dean Heike Hofmann Andreas Buja, Deborah F Swayne and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008.

[3] Mark J. Bentum. The search for exoplanets using ultra-long wavelength radio astronomy. In *2017 IEEE Aerospace Conference*, pages 1–7, 2017.

[4] Mark J. Bentum. Algorithms for direct radio detections of exoplanets in the neighbourhood of radiating host stars. In *2018 IEEE Aerospace Conference*, pages 1–7, 2018.

[5] Ziqi Dai, Dong Ni, Lizhuang Pan, and Yiheng Zhu. Five methods of exoplanet detection. *Journal of Physics: Conference Series*, 2012:012135, 09 2021.

[6] Dae-Hak Kim. On the clustering of huge categorical data. *Journal of the Korean Data and Information Science Society*, 21, 01 2010.