Tools and technologies used

My visualization contains a NASA exoplanet dashboard which visualizes intriguing attributes about exoplanets and their host stars. To complete this visualization, I used Python 3.9 along with the `Vega-Altair`[1] library from PyPI. To process the data, I parsed the CSV file using the `Pandas`[2] library from PyPI. For pre-processing, the initial 96 rows were only for documentation which was disregarded for data visualization. Furthermore, I filtered the data by dropping rows if it had empty values in the specific columns that I needed for my visualization using the `dropna(subset=[])` function or if it used a discovery method for the exoplanet that was not mentioned more than once in the entire dataset (to reduce the cluttering of categorical values with a low frequency).

Dataset

The dataset used in this visualization is a **table dataset** from the ”Planetary Systems Table” on the official `NASA Exoplanet Archive`[3]. Each row represents an item in this static dataset containing an exoplanet discovery ranging from 1995 to 2023 along with various columns for attributes about the exoplanet and its host star. Initially, the imported CSV contained 35258 rows with 92 columns which was not practical for a visualization. After pre-processing, the dataset contained 2591 rows and 13 columns that I chose after multiple weeks of finding good relationships between them:

1. **pl_name** for the name assigned to the discovered exoplanet which is a key attribute and its attribute type is categorical (nominal). This is only shown in the tooltip.
2. **sy_snum** for the number of stars in the planetary system and its attribute type is quantitative (discrete).
3. **sy_pnum** for the number of planets in the planetary system) and its attribute type is quantitative (discrete).
4. **discoverymethod** for the method of discovery and its attribute type is categorical (nominal).
5. **disc_year** for the year of discovery and its attribute type is quantitative (discrete).
6. **pl_controv_flag** representing if the discovery was controversial in published literature and its attribute type is categorical (nominal).
7. **pl_orbper** for the orbital period, its attribute type is quantitative (continuous) and is derived as its mean in visualizations.
8. **pl_bmasse** for the mass of the exoplanet and its attribute type is quantitative (continuous).
9. **pl_orbeccen** for eccentricity i.e. deviation of the orbit of the exoplanet from a perfect circle and its attribute type is quantitative (continuous). It is derived as its mean in visualizations.
10. **st_teff** for the temperature of an exoplanet’s star and its attribute type is quantitative (continuous).
11. **st_mett** for the metallic content on an exoplanet’s star, its attribute type is quantitative (continuous) and is derived as its mean in visualizations.
12. **sy_dist** for the distance to the exoplanet from earth and its attribute type is quantitative (continuous). It is also used to derive extra data to get the median of the highest 10% of distances each year to draw a blue trend line in the first plot.

13. **sy_kmag** for the brightness of the exoplanet's star and its attribute type is quantitative (continuous).

This dataset is extremely complex because of the vast nature of space and the universe. These 12 attributes (excluding pl_name) in each row have complex relationships between each other based on aspects of physics and astronomy. With over 2500 discoveries to plot, the visualization of this dataset is difficult because of its complexity from the volume of rows, complicated structure of relationships between elements and visual complexity required to plot points effectively without losing meaning. However, the cost of doing so comes with the benefit of understanding more about the universe we all share, which has an invaluable benefit to humanity and hence is why it is necessary to do so. From trial and error, techniques such as jittering to prevent over-plotting actually made visualizations worse because of the high number of points. As many astronomers and astrophysicists have stated[4], embracing the "craziness" of data from space and integrating any over-plotting effectively to portray an underlying message is the best approach.

Tasks (*action* and target)

Firstly, row one of my visualization shows the distance of exoplanet discoveries (controversial or non-controversial) over time. This allows the viewer to *consume* the recorded distribution of discoveries and *discover* trends across many years. This task is also supported in the second row of my visualization which displays three different plots to aid the viewer in discovering the patterns and symmetry of data from exoplanets and their host stars. Secondly, the third row of my visualization shows the mean metallic content of exoplanet's stars found over time. This supports the previous task of discovering trends, but also promotes *searching* for the years when dips occur in metallic content to *browse* features between row one and row three to see if these dips correspond to controversies in a similar structure. Finally, row four displays two plots to show the relationship between the deviation of an exoplanet's orbit and its host star, as well as relationships between the host star's brightness and temperature. In the first plot, viewers are *presented* trends (like row one) in exoplanet's orbits. The second plot allows the user to *query* the most common density to *summarize* the brightness and temperature from exoplanet's stars.

Encoding channels and idioms (with justification)

Idioms and encoding channels were chosen carefully to maximize their strengths and implemented based on the data relationship[5]. Row one in my visualization uses a strip plot as the idiom which is used because it supports the "craziness", as mentioned before, from the complexity of the data. Over-plotting is not an issue here because it is used as a tool to visualize the distribution of points. The encoding of color for the controversial status of discoveries was used because it is linked to row three later on and it requires a powerful visual cue for categorical attributes (in this case, a boolean). Furthermore, the encoding of shape is suitable for the discovery method because it is good for categorical data, such as ours that has few classes to minimize look-up time. Row two uses a scatterplot as well as a horizontal and vertical bar chart. The use of these idioms are justified because of they work well for showing the correlation between exoplanet distance from earth and the brightness of their host star, the magnitude of brightness ranges from their host star compared to the number of planets in their planetary system, as well as the ranking of this to their mass. Row three uses a surplus/deficit filled line as an idiom because this shows deviation over time efficiently and subsequently color is encoded to boost the encoding of position already showing this deviation. Row 4 uses a connected scatterplot to show relationships between the number of stars in a planetary system with the deviation of an exoplanet's orbit. This idiom is justified because it accurately shows the correction between each other. I encoded the orbital period as size and this is justified because screen real-estate is freely available here; it's important to note that I didn't encode size until this graph because encoding size on graphs with many points affects visibility and meaning, but it works here since the number of points are limited. Finally, I used a heatmap as an idiom to clearly show the density of discoveries in relation to the brightness and temperature of an exoplanet's star. I encoded brightness since it effectively displays the grids that are more common to occur, which is the main goal.

Novelty of the visualization

This visualization is novel because this data has not been popularly visualized in this way before. The relationship between exoplanets and their stars is displayed while meeting the requirements of using more than one idiom/using an idiom that is unique with the use of multiple encoding channels and using multiple facets to create juxtaposed views using a multiform approach for the dashboard. The implementation of this dashboard is highly complex due to the use of a wide variety of idioms and encoding channels while also maintaining function over form. Each idiom is chosen carefully to bring a new sense of meaning to each row in the dashboard. Furthermore, the specific encoding channels used are deliberately chosen to display the underlying meaning behind the data more effectively with a good mix of explanatory and exploratory visualizations.

Critical analysis of the strengths and weaknesses

In terms of weaknesses, the potential combination of multiple data sets from other space agencies (not only NASA) could bring new insights that were not yet covered. Furthermore, it could be argued that more trend lines (such as the blue one in row one) could be used to exemplify the trends in each plot - this was decided against because of distracting viewers from the underlying data, but this is up to personal opinion. In terms of strengths, this visualization effectively uses various idioms and encoding channels in a way that elegantly shows relationships between exoplanets and their host stars i..e function over form. Furthermore, all rows are visible with minimal scrolling on a web page and are displayed in a coherent manner such that it feels like a story being told as the viewer sees it from top to bottom, with an ”eyes beat memory”[6] approach; each row beforehand has an impact on the row afterward. These aspects make the visualization exceed the normal baseline compared to regular visualizations.

References:

[1] Vega-Altair on PyPI

[2] Pandas on PyPI

[3] Planetary Systems Table on the NASA Exoplanet Archive (Dataset)

[4] Section 8.2 - ”Astronomical Data Volume and Diversity” in Visualization in Astrophysics

[5] Visual vocabulary poster by the Financial Times

[6] Chapter 6 - ”Rules of Thumb” in Visualization Design and Analysis