To begin performing Exploratory Data Analysis (EDA) on the dataset, I started by loading all the necessary libraries to support the process. Next, I imported the dataset into a pandas DataFrame to facilitate more convenient data manipulation. I then used the `info` function to understand the types of columns present in the dataset.

Following this, I examined each column's properties, including the mean, median, and standard deviation, to get a sense of their distributions. I identified columns with string values and converted these to integers, as many functions struggle with string data types. For some columns, I used label encoding, while for others, I substituted the string values with the mean value of the column if it contained only one unique string. For the third column I saw how the data was spread and saw that a lot of the data present was heavily skewed towards a certain variable and just converted the entire column dependent on that variable. If the value in the column was that variable it was given a value 1, otherwise it was given a value of 0.

Next, I generated the correlation matrix of the dataset, which revealed that several columns had correlations greater than 0.9. This indicated that many columns were highly similar, meaning that an increase in one column often led to an increase in another and vice versa.

I attempted to visualize the data using principal component analysis (PCA), but the resulting plot was a large, indistinct circle. To improve this, I pre-processed the dataset further by removing all string columns, then removed integer columns, as the majority of the data were in floating-point format. I also discarded floating-point columns with fewer than 10 distinct values, as these resembled classification fields and could skew the data.

I then identified the names of columns with correlations greater than 0.9 with more than 40 other columns and removed these from the dataset. After recalculating PCA for two components and visualizing the results, the issue of indistinct data persisted. To address this, I applied the z-score function for outlier detection. After removing the outliers, I re-ran PCA. This time, the plot showed improved separation between data points compared to previous attempts.