The first thing that I did to tackle performing EDA on the given dataset is to load all the relevant libraries that will aid me in this process. The second step that I did was to read the dataset into a pandas data frame to be able to utilise the dataset in a much more convenient method. I then applied the info function on the dataset to understand the types of columns that we are dealing with.

My next step was to find out all the properties of each column in the dataset that includes the mean, median, the standard deviation. Which lead me to understand the way by which each column was distributed. I then singled out all the columns present in the data frame that contained string values and proceeded to fill them in with integers. This is because a lot of functions are not able to work that well with string data types. I utilised label encode for one of the columns but for the other columns I used other tricks. One column was labelled a string but in all actuality the column only had one string so I changed all of the value of the string to the mean value of the column.

Then I found out the correlation matrix of the given dataset. This led me to find out that a lot of the columns in the dataset have got a correlation of greater than 0.9. This means that there were several columns present in the dataset whose rows performed very similarly to one another. (This means that if the value in one column goes up then the value of the other column goes up.)

I then tried plotting out the values that were determined by the principal component. This led me to a graph that was just a massive giant circle through which very little could be inferred. I then decided to preprocess the dataset further this is so as to get a better result when applying the PCA. So, I proceeded with removing all the string columns from the dataset. I then proceeded to remove all integer columns this is because of the fact that a lot of the data present in the dataset was in the form of floating-point numbers and having integers as well would have made it a bit more difficult to deal with.

I then removed all the floating point columns that didn't have more than 10 distinct variables in the column that is because of the fact that if there are only 10 distinct variables in a dataset that contains 2000 rows that means that the column is acting similar to a classification field with only a few distinct columns. This could cause my data to be skewed

I then proceeded to find out all the columns that had more than 40 columns with which its correlation was greater than 0.9. I proceeded to add the names of all these columns to a list and remove all of these columns from the dataset. I then recalculated the value of PCA for two components and visualized how the graph would look like. I then saw that the issue of not being able to see distinguish between any of the data was still present. So to tackle that I utilised the z-score function to find out the z score of each column present in the dataset. This is so then I could proceed with outlier detection. After proceeding with the outlier detection I then retried implementing PCA. This time the dots on the graph are more able to be distinguished compare to the other graphs before.