**3. Create a documentation file explaining what you understood after applying EDA.**

**Basic Commands to know about dataset:**

data.shape: - We have total 5839 rows and 42 columns in this dataset.

data.columns: - We have below mentioned columns name in dataset

['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N','O', 'P', 'Q', 'R', 'X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8','X9', 'Y1', 'Y2', 'Y3', 'Y4', 'Y5', 'Y6', 'Y7', 'Y8', 'Y9', 'Z1', 'Z2','Z4', 'Z5', 'Z6', 'Class']

data.info(): - Total 37 features has float values, 4 int and 1 object.

data.isnull().sum().sum():- There is no null value in dataset.

data.describe():- To know statistical information about dataset.

**Class distribution in dataset: -** After plot pie chart, we can check that dataset has very few records w.r.t ‘D’ class. Because of this imbalance, our model accuracy will affect.

**Check correlation between features: -** After plotting heatmap of correlation, we can see some of features are highly correlated with each other. So, we will drop one of them.

**Check which features are highly correlated with target feature: -** After plotting box plot of highly correlated features, still are outliers for some classes. It means we cannot consider these features only. We have to go for dimension reduction for this.

**Check Pair plot: -** In pair plot, we can see, all datapoints of each class are overlapping. For this type of dataset we need to use a robust classifier like random forest, decision tree.