Data cleaning procedures using NumPy

* The dataset contains 10000 rows and 14 columns.
* Calculating the mean for each column to figure out how many columns contain numeric and object data types in our dataset to perform our analysis ,When calculating the mean for each column if a get a value it means that the column contains numeric data if no value or a NAN is returned it means that the column contains object data type or categorical data that is why no value or NAN is returned, performing this operation has confirmed that there are eight columns with NAN values or no value returned this means that these columns contain object data type or filled with categorical data and six columns contain numeric data.
* Splitting Text data and numeric data for better cleaning and processing data, to perform a split we should know the index values of columns containing numeric data and string data, hence we use np.argwhere() function.
* Of the 14 columns in the dataset column numbers 2,4,6,9,10,11,12,13 contain string values and the remaining columns contain numeric data.

Working with string data

* In the first column in string data “issue date” the value “-15” was stripped since all the data in this column contained this value and the month names where converted to their equivalent numbers.
* The next manipulation was performed on the “loan status” column, the values in this column was converted to dummy variables 0 and 1 with keeping worst case scenario in mind, here “0” means status is bad and “1” means status is good, the values like “blank space”, “charged off”, “default” and “late (31-120days)” is considered bad status that is “0” and the rest of the values are assumed good status that is “1”
* The blank spaces in the “Term” column were replaced with the value 60 referring to a duration of 60 months term duration.
* Then taking the Grade column, in this column the blank values was replaced with 2nd highest repeating value in the column, just to not overpopulate the highest occurring value in the dataset
* The missing values in the “sub-grade” column was replaced with the worst possible case in the data.
* The values in the “verification status” was replaced with dummy variables, the values like “blank space” and “not verified” has been treated as “0” and the rest as “1”.
* Deleted the “Url” column.

Working with numeric data

* Next we begin manipulating the numeric data, we start by looking for null values in the dataset, we do not get any null value because we filled the null values with a filler value before.
* The “Id” column is a unique representation of customers and fortunately there are no null values in this column.
* Then we calculated the statistics that is minimum maximum and mean for all the numeric columns as they might be helpful in filling the filler value we added before.
* Filling “funded amount” column with a worst case possible number.
* The remaining numeric columns was replaced with appropriate values and all missing values was addressed
* Finally both string data and numeric data was merged together for analysis.