Data Wrangling

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics

The original dataset contains 1321864 entries and 151 features

```
loans_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1321864 entries, 0 to 103547
Columns: 151 entries, id to settlement_term
dtypes: float64(112), object(39)
memory usage: 1.5+ GB
```

The steps to be performed in Data Cleaning can be divided into three categories

1. Columns to Retain

- Drop the columns which have more than 30% missing values.
- Drop the columns which are not applicable to the analysis (first pass of the dataset). The features like **id** and **url** are the information about LC which we don't need for the analysis are dropped from the dataset.
- Drop the columns that have only one unique value. If all the values in the column are same, then that feature would not be useful for the analysis.

2. Rows to retain

- Drop the rows which has null values in each column

3. Handling Missing Values

Missing values can lead to wrong prediction or classification and can also cause a high bias for any given model being used. The missing values are filled differently for categorical and numerical columns.

- For numeric attributes, replace the missing or null values with the mean of the respective column the missing value belongs to
- For categorical attributes, replace the missing or null values with the mode of the respective column (most frequent value) the missing value belongs to

```
In [38]: filtered_numeric_loans_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1321847 entries, 0 to 103545
Data columns (total 66 columns):
```

After performing data wrangling techniques, the dataset size is reduced to 1321847 and 66 columns