## Tidying the data

Now that we’ve identified several errors in the data set, we need to fix them before we continue with our analysis. Let’s review the issues:

There are missing values in some variables. Based on the importance of the variables, we will decide on the method to use.

Looking at the distributions of the data, we noticed that ApplicantIncome and LoanAmount have outliers.

Fixing outliers can be tricky. It’s hard to tell if they were caused by measurement error, errors while recording, or if the outliers are real anomalies. If we decide to remove records, we have to document the reason behind this decision.

In this data set, we will assume that missing values are systematic because the missing data are coming in certain variables in a random manner. Also, we note that missing values are on both numerical and categorical data, therefore, we will be creating different functions to handle these scenarios. These functions help in imputing missing values with plausible data values. These values are inferred from a distribution that is designed for each missing data point.

**Type of Variables:**

1. Input variable (Predictor): Gender, Married, Education, Self\_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History
2. Output variable (Target): Loan\_Status

**Variable category:**

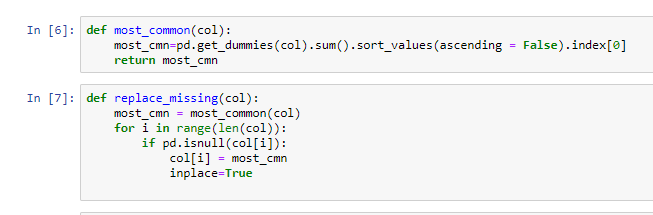
1. Categorical variables: Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status
2. Continuous variables: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term

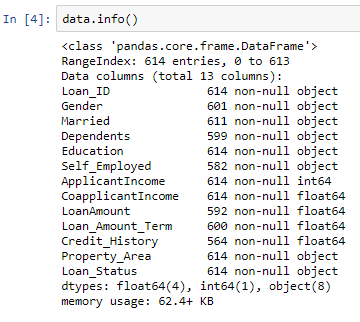
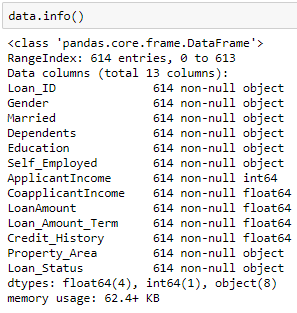
**Data Pre-processing Steps:**

1. Handling missing values
2. Creation of required variables
3. Replacing the data-values

**Handling Missing Values:**

1. Defining function to fetch “most\_common” value of the feature
2. Defining function to “replace\_missing” values with most\_common/mean values
3. Calling “replace\_missing” function for all features with null values by passing the feature name to the function.



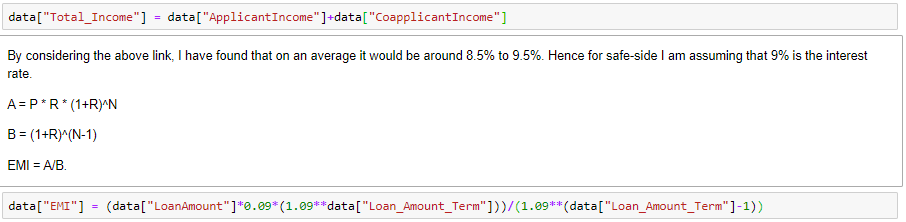
**Before Handling Null Values After Handling Null Values**

**Creation of required variables:**

1. Creating “Total\_Income” variable by adding “ApplicantIncome” to “CoapplicantIncome”
2. Calculating the “EMI” variable

By considering the above link, I have found that on an average it would be around 8.5% to 9.5%. Hence for the safe-side I am assuming that 9% is the interest rate.

* A = P \* R \* (1+R)^N
* B = (1+R)^(N-1)
* EMI = A/B



**Replacing the data-values:**

1. “Dependents” variable has some data-values marked as “**3+**”, so changing them as value “**3**”.

