Introduction:

The data set is a Home Equity Line of Credit data set that provides information at time of loan, as well as whether the loan defaulted (BAD=1) or repaid (BAD=0).

There are two target variables TARGET\_BAD\_FLAG and TARGET\_LOSS\_AMT. We can make 2 prediction out of this dataset. One if the loan would be defaulted or not and second the loan amount that would be defaulted.

**Descriptive Statistics:**

The Dataset shows 1189 out of 5960 had bad loan- about that about 19% of the people defaulted on their loan. The min defaulted loan amount is 224 and the max defaulted loan amount is 78987. the maximum amount looks like an outlier compare to the 25%,50% and 75%.

There are twelve independent variables and two dependent variables.

The information on the dataset shows how many columns have missing data along with the data types which would be useful in determining the type of handling.

Missing values:

While many columns have missing values, the column TARGET\_LOSS\_AMT has the most missing showing that there are about 20% of the loans that are not repaid and 80% repaid.

Column TARGET\_BAD\_FLAG and LOAN do not have missing values.

DEBTINC has the most missing values. Further analysis using graphs would be applied to determine how to impute missing values.

**Categorical Variables:**

There are two categorical variables Job and Reason.

Job category 'Other' has the most loan and also the most bad loans. But we do not have a clear idea of what the Other could be. The next job that is risky is ProfExec.

Imputing missing Job is done using a value “Missing” and further values 1 through 7 is used to encode class value to numeric values.

Reason variable shows that that there are more loans for debt consolidation compared to home improvements.

Imputing missing Reason is done using a value “DebtCon” because DebtCon is the most common reason. Further values 1 and 2 is used to encode class value to numeric values.

Numeric Variables:

* The variable DEROG and DELINQ have the most variability in values. However, the boxplot shows that most of the data is between 0 and 1.
* DELINQ - As the number of Delinquent credit lines and Derogatory remarks increased, there are more bad loans. Missing value is imputed with 0.
* DEROG: Most of the data is between 0-1 that means there are not many with derogatory remarks. The missing values are imputed grouped by mean of the reason.
* CLNO column has a good distribution of data and the missing values are imputed using the mean of the reason.
* NINQ column has most values between 0 and 5 and the missing values are imputed using the mean of the reason.
* CLAGE: highly populated between 100 - 200. Outlier need to be handled. missing values are imputed using the mean of the reason.
* VALUE: Data is between 10,000- 200,000. Values over 200,000 and 400,000 are a few. Outliers beyond 350,000 needs to be handled. The missing value is imputed using the mean of the reason.
* YOJ: extremely right skewed. Data needs to be handled.
* LOAN: loan amount between 10000 and 20000 is the most populated. There are a few outliers that is skewing the distribution right. Those will need to be handled.
* MORTDUE: most populated between 50,000 - 100,000; data past 250,000 can be dropped.
* DEBTINC: Debt to income ratio has most values populated between 20- 40 with a few outliers that are skewing the distribution a lot. Outliers need to be handled. Missing value is imputed using the mean of the reason,