MSDS 422
Juhwi Kim

Introduction

To accurately predict whether a Home Equity Loan is default or not, a relevant dataset containing accurate information and characteristics of borrowers needs to be examined. Once we have explored the dataset, our task will be to create machine learning models that leverage these key indicators to predict whether a loan will default.

Dataset

For this project, we used Home Equity Line of Credit (HMEQ) dataset. A home equity loan is a loan where the borrower uses the equity of his or her home as the underlying collateral. This data set reports characteristics and delinquency information for 5,960 home equity loans.

Descriptive Statistics

```
description= df.describe()
description
```

	TARGET_BAD_FLAG	TARGET_LOSS_AMT	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
count	5960.000000	1189.000000	5960.000000	5442.000000	5848.000000	5445.000000	5252.000000	5380.000000	5652.000000	5450.000000	5738.000000	4693.000000
mean	0.199497	13414.576955	18607.969799	73760.817200	101776.048741	8.922268	0.254570	0.449442	179.766275	1.186055	21.296096	33.779915
std	0.399656	10839.455965	11207.480417	44457.609458	57385.775334	7.573982	0.846047	1.127266	85.810092	1.728675	10.138933	8.601746
min	0.000000	224.000000	1100.000000	2063.000000	8000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.524499
25%	0.000000	5639.000000	11100.000000	46276.000000	66075.500000	3.000000	0.000000	0.000000	115.116702	0.000000	15.000000	29.140031
50%	0.000000	11003.000000	16300.000000	65019.000000	89235.500000	7.000000	0.000000	0.000000	173.466667	1.000000	20.000000	34.818262
75%	0.000000	17634.000000	23300.000000	91488.000000	119824.250000	13.000000	0.000000	0.000000	231.562278	2.000000	26.000000	39.003141
max	1.000000	78987.000000	89900.000000	399550.000000	855909.000000	41.000000	10.000000	15.000000	1168.233561	17.000000	71.000000	203.312149

Following the examining an overview of numerical variables' mean and standard deviations, we found that many variables such as TARGET_LOSS_AMT, LOAN, and MORTDUE that may have outliers with more than 2 standard deviations above the mean values.

Description of Good Loans (No Default)

```
Good_df.describe()
```

	TARGET_BAD_FLAG	TARGET_LOSS_AMT	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
count	4771.0	0.0	4771.000000	4359.000000	4764.000000	4321.000000	4150.000000	4263.000000	4541.000000	4336.000000	4602.000000	4290.000000
mean	0.0	NaN	19028.107315	74829.249055	102595.921018	9.154941	0.134217	0.245133	187.002355	1.032749	21.317036	33.253129
std	0.0	NaN	11115.758554	43584.993587	52748.392952	7.676033	0.514490	0.674124	84.465217	1.531322	9.682601	6.947482
min	0.0	NaN	1700.000000	2619.000000	8000.000000	0.000000	0.000000	0.000000	0.486711	0.000000	0.000000	0.720295
25%	0.0	NaN	11700.000000	47484.000000	67297.750000	3.000000	0.000000	0.000000	120.219885	0.000000	15.000000	28.905127
50%	0.0	NaN	16900.000000	66839.000000	90659.000000	7.000000	0.000000	0.000000	180.415787	1.000000	20.000000	34.541671
75%	0.0	NaN	23500.000000	93068.000000	120615.500000	13.000000	0.000000	0.000000	240.157802	2.000000	26.000000	38.739077
max	0.0	NaN	89900.000000	371003.000000	471827.000000	36.000000	6.000000	5.000000	649.747104	11.000000	56.000000	45.569843

Description of Bad Loans (Default)

MSDS 422
Juhwi Kim

```
Bad_df.describe()
```

	TARGET_BAD_FLAG	TARGET_LOSS_AMT	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
count	1189.0	1189.000000	1189.000000	1083.000000	1084.000000	1124.000000	1102.000000	1117.000000	1111.000000	1114.000000	1136.000000	403.000000
mean	1.0	13414.576955	16922.119428	69460.452973	98172.846227	8.027802	0.707804	1.229185	150.190183	1.782765	21.211268	39.387645
std	0.0	10839.455965	11418.455152	47588.194467	74339.822506	7.100735	1.468381	1.902961	84.952286	2.246976	11.812981	17.723586
min	1.0	224.000000	1100.000000	2063.000000	8800.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.524499
25%	1.0	5639.000000	9200.000000	39946.500000	59368.250000	2.000000	0.000000	0.000000	96.033333	0.000000	13.000000	32.383046
50%	1.0	11003.000000	14900.000000	60279.000000	82000.000000	6.000000	0.000000	0.000000	132.866667	1.000000	20.000000	38.079762
75%	1.0	17634.000000	21700.000000	85864.500000	116000.000000	12.000000	1.000000	2.000000	193.283333	3.000000	28.000000	43.285990
max	1.0	78987.000000	77400.000000	399550.000000	855909.000000	41.000000	10.000000	15.000000	1168.233561	17.000000	71.000000	203.312149

- Compared to borrowers that did not default on their loan, defaulted loans include extreme outlier with their home value at $855K and Debt to Income Ratio seems to be much higher among borrowers defaulted on their loan.

Categorical Variables against Target Variables:

```
Class =  REASON
REASON
DebtCon    3928
HomeImp    1780
Name: REASON, dtype: int64
Bad Flag REASON
DebtCon    0.189664
HomeImp    0.222472
Name: TARGET_BAD_FLAG, dtype: float64
 .................
Loss Amount REASON
DebtCon    16005.163758
HomeImp     8388.090909
Name: TARGET_LOSS_AMT, dtype: float64
 ===============
```

- Compared to taking a home equity loan to use towards home improvement, Debt Consolidation results in higher amount of unpaid loan ($8388 vs. 16005).

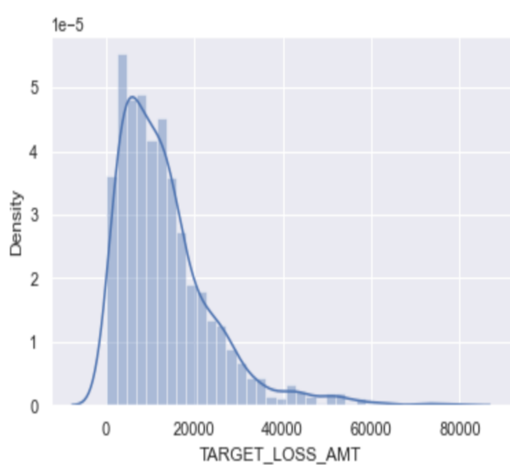
```
 Class =  JOB
JOB
Mgr          767
Office       948
Other       2388
ProfExe     1276
Sales        109
Self         193
Name: JOB, dtype: int64
Bad Flag JOB
Mgr         0.233377
Office      0.131857
Other       0.231993
ProfExe     0.166144
Sales       0.348624
```

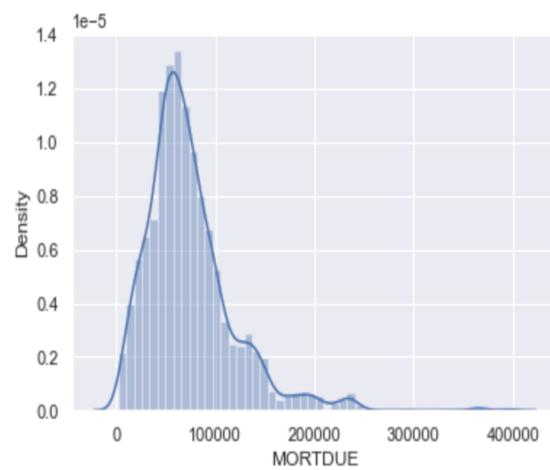
```
Self       0.300518
Name: TARGET_BAD_FLAG, dtype: float64
 .................
Loss Amount JOB
Mgr        14141.536313
Office     13475.304000
Other      11570.102888
ProfExe    14660.966981
Sales      16421.447368
Self       22232.362069
Name: TARGET_LOSS_AMT, dtype: float64
 ===============
```

- Compared to other jobs, people working in Sales showed higher probabilities of defaulting their home equity loans followed by self-employed borrowers. Comparing the amount of unpaid loans, self-employed borrowers show higher unpaid loan amount.

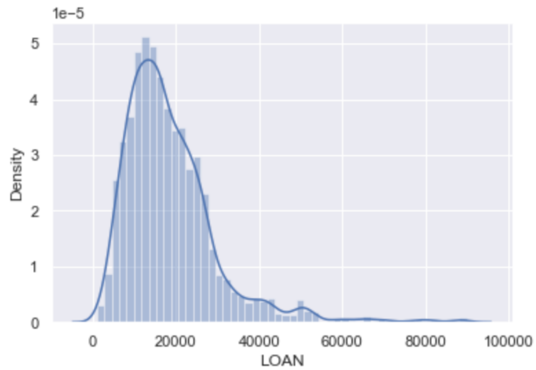
Distribution Visualization

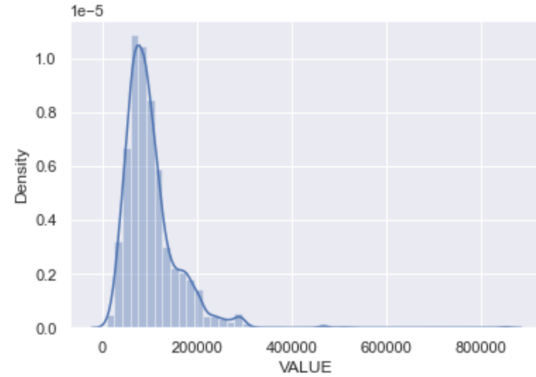
To understand dataset better, I have evaluated distribution charts for all numerical variables for skewedness and visualize outliers.

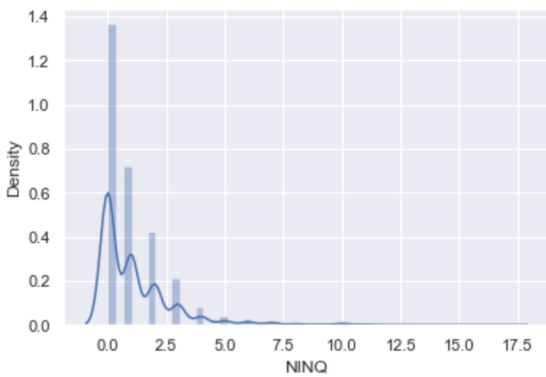


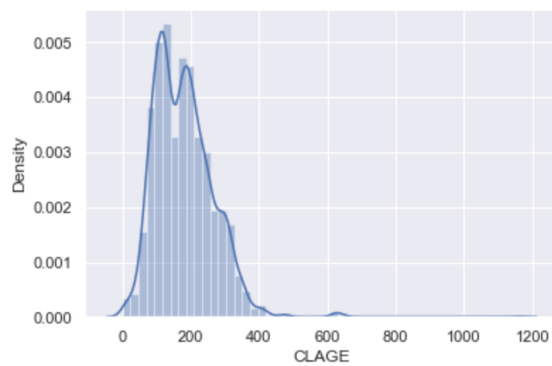


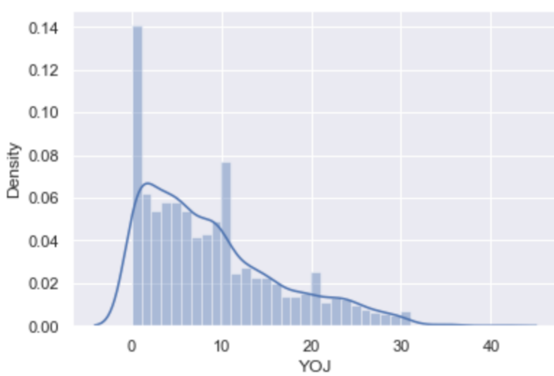
MSDS 422
Juhwi Kim

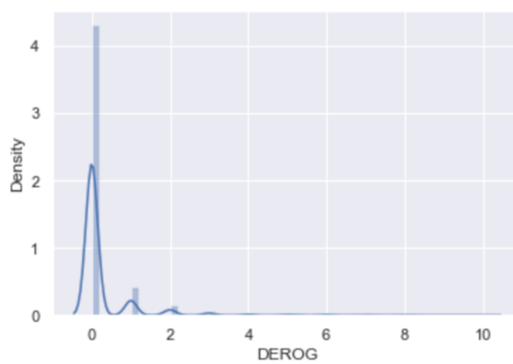


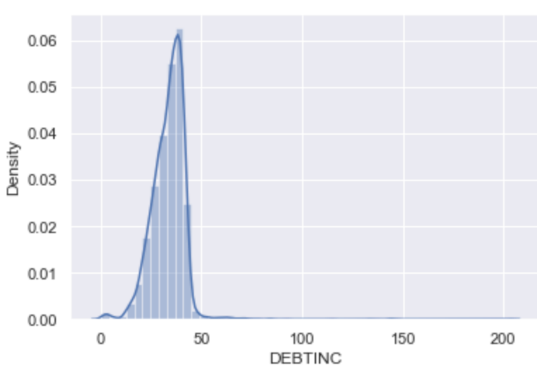


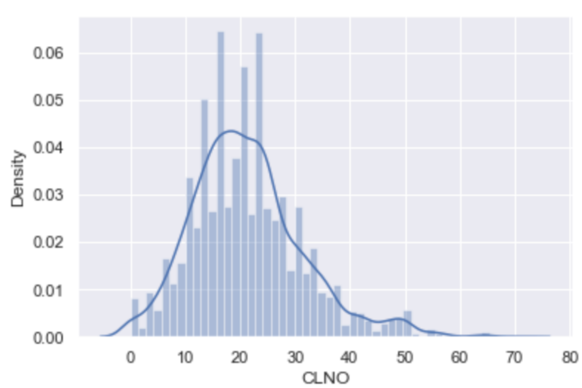












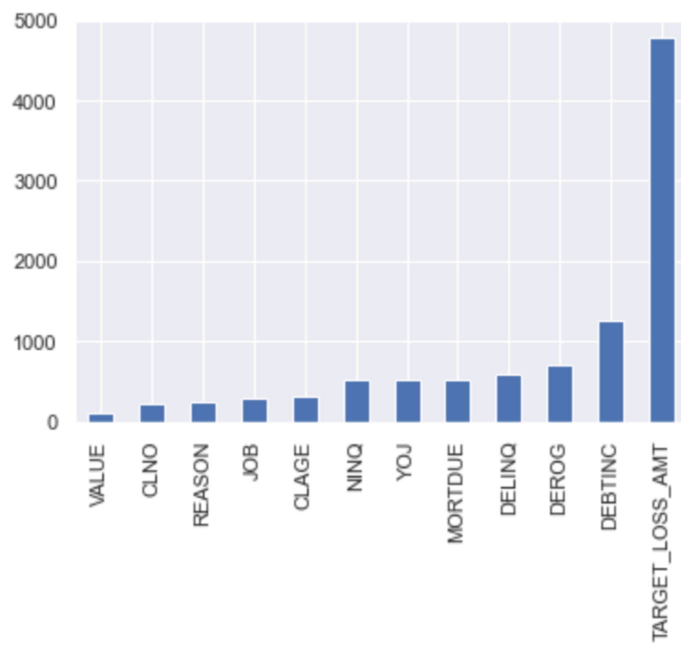
MSDS 422
Juhwi Kim

Through distribution charting, we can observe that majority of variables show left-sided skewness. Only CLNO, the number of credit lines a borrower has, exhibits close to normally distributed form.

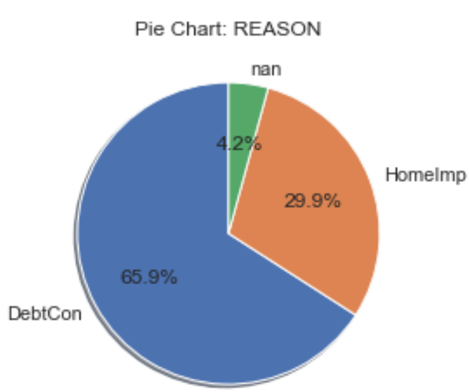
Missing Values

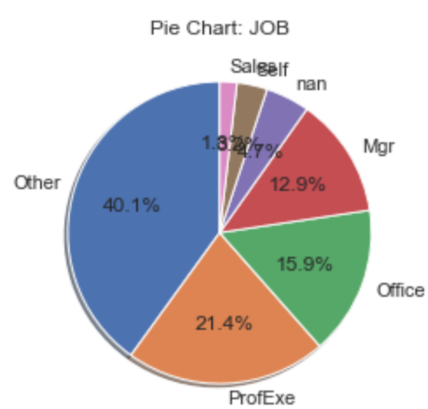
Overall, 12 out of 14 variables contain missing values and the number of missing values is listed below:

	Variables	Missing Values
0	TARGET_BAD_FLAG	0
1	TARGET_LOSS_AMT	4771
2	LOAN	0
3	MORTDUE	518
4	VALUE	112
5	REASON	252
6	JOB	279
7	YOJ	515
8	DEROG	708
9	DELINQ	580
10	CLAGE	308
11	NINQ	510
12	CLNO	222
13	DEBTINC	1267



- TARGET_LOSS_AMT has the highest number of missing values; however, it is due to corresponding loans have not been defaulted. Only when loan is defaulted, it will be flagged and result in amount of unpaid loan gets logged. For this variable, it will make sense to note it as 0.





MSDS 422
Juhwi Kim

- Job and Reason are two object/categorical variables. While large portion of the reason for taking the loan is to consolidate debt, there could be other reasons beside debt consolidation or home improvement to take the loan; therefore, imputing missing values with mode may not truly reflect the data or possibly skew the result when there are only two options (DebtCons vs. HomeImp). Therefore, for Reason, I will proceed with imputing missing values by labeling them "Missing".
- For Jobs, the large portion of the jobs fall under "Other". Since Other is overarching and broader categorical label, I will proceed with imputing missing values using Mode which will transform NaN values into "Other".

- Following imputing missing values, I have applied one hot encoding to transform categorical variables into numerical variables:
    - For each reason and job category, if a borrower falls under the category, it will be indicated by 1. Otherwise, the value will be 0.

```
df["REASON_DEBTCON"] = (df.IMP_REASON.isin( ["DebtCon"] ) + 0 )
df["REASON_HOMEIMP"] = (df.IMP_REASON.isin( ["HomeImp"] ) + 0 )
df["REASON_MISSING"] = (df.IMP_REASON.isin( ["Missing"] ) + 0 )
```

```
df["JOB_Mgr"] = (df.IMP_JOB.isin( ["Mgr"] ) + 0 )
df["JOB_Office"] = (df.IMP_JOB.isin( ["Office"] ) + 0 )
df["JOB_Other"] = (df.IMP_JOB.isin( ["Other"] ) + 0 )
df["JOB_ProfExe"] = (df.IMP_JOB.isin( ["ProfExe"] ) + 0 )
df["JOB_Sales"] = (df.IMP_JOB.isin( ["Sales"] ) + 0 )
df["JOB_Self"] = (df.IMP_JOB.isin( ["Self"] ) + 0 )
```

- The missing values in the numerical variables: 'LOAN', 'MORTDUE', 'VALUE', 'YOJ', 'DEROG', 'DELINQ', 'CLAGE', 'NINQ', 'CLNO', 'DEBTINC' have been imputed using the column's median value.

Following the imputation of missing values and transformation of categorical variables into numerical variables, the dataset has been transformed to following format. The updated variables along with the first five rows of dataset is listed below:

MSDS 422
Juhwi Kim

	0	1	2	3	4
TARGET_BAD_FLAG	1.000000	1.000000	1.000000	1.000000	0.000000
LOAN	1100.000000	1300.000000	1500.000000	1500.000000	1700.000000
IMP_TARGET_LOSS_AMT	641.000000	1109.000000	767.000000	1425.000000	0.000000
REASON_DEBTCON	0.000000	0.000000	0.000000	0.000000	0.000000
REASON_HOMEIMP	1.000000	1.000000	1.000000	0.000000	1.000000
REASON_MISSING	0.000000	0.000000	0.000000	1.000000	0.000000
JOB_Mgr	0.000000	0.000000	0.000000	0.000000	0.000000
JOB_Office	0.000000	0.000000	0.000000	0.000000	1.000000
JOB_Other	1.000000	1.000000	1.000000	1.000000	0.000000
JOB_ProfExe	0.000000	0.000000	0.000000	0.000000	0.000000
JOB_Sales	0.000000	0.000000	0.000000	0.000000	0.000000
JOB_Self	0.000000	0.000000	0.000000	0.000000	0.000000
M_MORTDUE	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_MORTDUE	25860.000000	70053.000000	13500.000000	65019.000000	97800.000000
M_VALUE	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_VALUE	39025.000000	68400.000000	16700.000000	89235.500000	112000.000000
M_YOJ	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_YOJ	10.500000	7.000000	4.000000	7.000000	3.000000
M_DEROG	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_DEROG	0.000000	0.000000	0.000000	0.000000	0.000000
M_DELINQ	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_DELINQ	0.000000	2.000000	0.000000	0.000000	0.000000
M_CLAGE	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_CLAGE	94.366667	121.833333	149.466667	173.466667	93.333333
M_NINQ	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_NINQ	1.000000	0.000000	1.000000	0.000000	0.000000
M_CLNO	0.000000	0.000000	0.000000	1.000000	0.000000
IMP_CLNO	9.000000	14.000000	10.000000	20.000000	14.000000
M_DEBTINC	1.000000	1.000000	1.000000	1.000000	1.000000
IMP_DEBTINC	34.818262	34.818262	34.818262	34.818262	34.818262

Upon successful data preparation, initial correlation test against target variables has been executed, which points to the debt-to-income ratio, number of delinquencies and derogatory may be more closely correlated with loan defaulting and subsequent unpaid loans.