EDA

1. See if there are any rows that have too much missing value

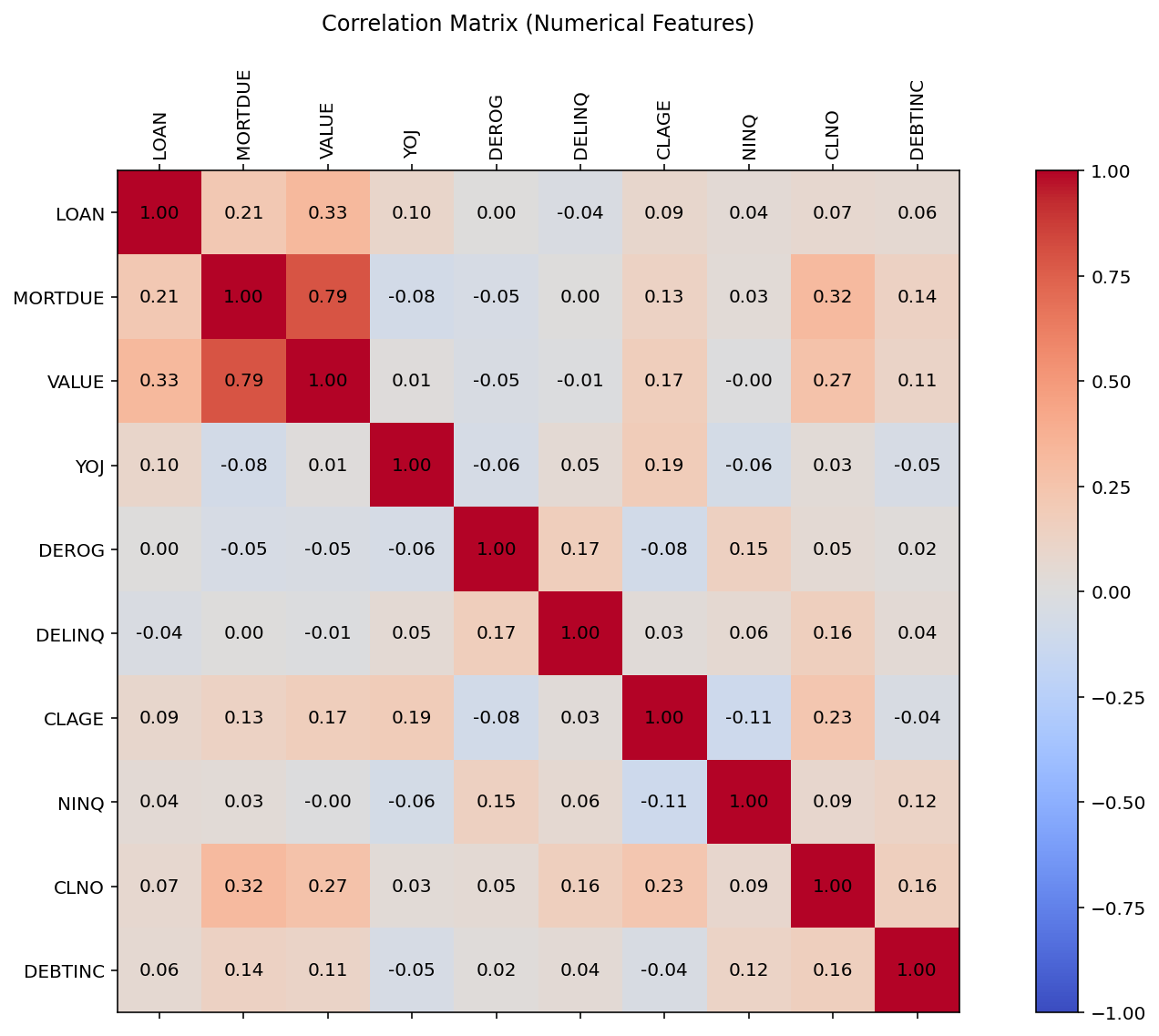
A graph of a number of missing values

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| **Num Missing** | **Count** | **Percentage** |
| 0 | 3364 | 56.44 |
| 1 | 1589 | 26.66 |
| 2 | 449 | 7.53 |
| 3 | 219 | 3.67 |
| 4 | 64 | 1.07 |
| 5 | 83 | 1.39 |
| 6 | 66 | 1.11 |
| 7 | 25 | 0.42 |
| 8 | 39 | 0.65 |
| 9 | 49 | 0.82 |
| 10 | 11 | 0.18 |
| 11 | 2 | 0.03 |

Not too much data has more than 5 missing variables (less than 5%, 4.6% 275 rows, we drop these rows to improve overall data quality/improve model reliability while still retain most / majority of the overall dataset)

1. From correlation, we can see that also see that:

* MORTDUE and Value has very high correlation (0.79)
* Loan & Value has moderate correlation (0.33)
* Make a new var (kind of like adjusted Loan to Value ratio 🡪 (Loan + Mortdue) / Value since Mortdue shows existing mortgage obligations (loan + mortdue would show the total obligation / total debt that a person will have if the loan are approved)
* LTV ratio: higher usually riskier
* 

1. see if the missing value have any meaning?



It may have meaningful missingness, so we check with chi-square test

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Chi-Square Statistic** | **p-value** | **Missingness Informative?** |
| DEBTINC | 1,700.89 | 0.00000.E+00 | **Yes** |
| DEROG | 54.2161 | 1.79605E-13 | **Yes** |
| DELINQ | 49.4445 | 2.04059E-12 | **Yes** |
| LoanValueRatio | 46.5162 | 9.0865E-12 | **Yes** |
| NINQ | 27.8847 | 1.28765E-07 | **Yes** |
| YOJ | 10.1527 | 0.00144093 | **Yes** |
| JOB | 9.10627 | 0.00254735 | **Yes** |
| REASON | 2.97622 | 0.0844966 | **No** |
| CLNO | 1.05477 | 0.304412 | **No** |
| CLAGE | 0.776983 | 0.378065 | **No** |

For p-value <0.05 there is a the missingness is informative

1. Add missing indicator variable for each variable that have missing value (except for REASON , CLNO and CLAGE)

* For every variable that has missing value, a new column will show whether it has missing value or not (0 if not, 1 if yes)
* This will help the model to know that the missing value have meaning before we impute the missing value

1. The missing value will be Imputed as below

|  |  |
| --- | --- |
| **Feature** | **Imputed Method** |
| LTV Ratio | As a ratio, imputed with very large number (9999) since it indicates higher risk (as seen as the figure above where the missing value for VALUE and DEBTINC has a higher BAD rate) |
| DEBTINC |
| DEROG | Imputed with mode since the distribution is highly skewed to the right |
| DELINQ |
| NINQ |
| CLAGE | Imputed with median since it is well-distributed |
| CLNO |
| YOJ |
| REASON | Categorical data, Imputed with “Not Provided” |
| JOB | Categorical data, Imputed with “Other” (mode) |

The process of imputing & scaling will be in the pipeline during modeling part to avoid data leakage

Modelling part

1. Split train test
2. One hot encoding (for categorical – ‘REASON’ & ‘JOB’)
3. Robust Scaling (train data only) – use robust scaling since we imputed very high number for LTV Ratio & DEBTINC so it doesn’t be too skewed
4. Run the model CV & Test (using threshold & best parameters)