

Summary of Findings

Univariate exploration and wrangling efforts:

1. Completed loans and Current loans constitute the largest shares of LoanStatus.
2. Most borrowers are employed and are in the mid-range (\$25,000-49,999) to upper-range (\$100,000+) income brackets.
3. 92% of borrowers also have a verifiable source of income.
4. Most people also have 0 to 1 recommendations.
5. 'Other' and 'Professional' constitute the largest share of occupations but they are vague, given how detailed the other occupations are. Occupation will thus not be useful for analysis, and also because it has too many categories.
6. The BorrowerRate is generally right-skewed and multimodal.
7. After cleaning, the DebtToIncomeRatio distribution was skewed to the right, with a general smooth slope and some unusual gaps. Dtl is concentrated around 0.2, which is an acceptable ratio because the lower the Dtl the better. Dtl's higher than 0.5 are considered unsafe by lenders and they appear as outliers in the boxplot distribution.

Wrangling:

1. I combined redundant bins for LoanStatus, EmploymentStatus and IncomeRange.
2. I changed IncomeRange into an ordered categorical type.
3. I changed the column name of Listing Category (numeric) to one without a column space i.e. ListingCategory
4. I also replaced the numeric values in ListingCategory with the actual descriptive categories.
5. I dropped extreme outliers for DebtToIncomeRatio with values beyond 1, as they also made the distribution extremely right-skewed.
6. Some of the listings had suspiciously low BorrowerRates given that they were high risk and in the lower ranges of credit scores. I dropped these outliers.

Bivariate Exploration:

1. DebtToIncomeRatio has a positive correlation with BorrowerRate because a high Dtl lowers creditworthiness and leads to higher borrower rates due to lenders seeing the listing as risky.

2. Completed Loan listings have a generally lower median BorrowerRate, indicating that lower interest rates increase likelihood of loan clearance.
3. The Recommendation distributions for LoanStatus are all thin slices at similar levels as the values are concentrated around 0 to 1. However, it's interesting to note that high outliers, including max values, are concentrated at Completed, showing that listings with more recommendations are more likely to have the loans cleared.
4. Most borrowers with Completed or Current loan status have an income of at least \$25,000.
5. Borrowers with a Completed or Current loan status typically have a verifiable income source.

Multivariate Exploration:

1. There is a positive relationship between DebtToIncomeRatio and BorrowerRate, and Current and Completed LoanStatus have high concentrations of values around the lower regions of both numeric variables, while Cancelled, Defaulted and PastDue have very low concentrations in the lower regions of both numeric variables. Thus borrowers with healthy (lower) Dtl and lower borrower rates are more likely to clear their loans.
2. There is not much of a significant relationship between LoanStatus, IncomeRange and BorrowerRate. However, higher income brackets tend to have lower Dtl than the rest.
3. There also is not much difference between box plots for the relationship between LoanStatus, BorrowerRate and EmploymentStatus, but employed and retired tend to have lower median rates.