# ML C54
# Lending case study

Srikanth Sadhanala

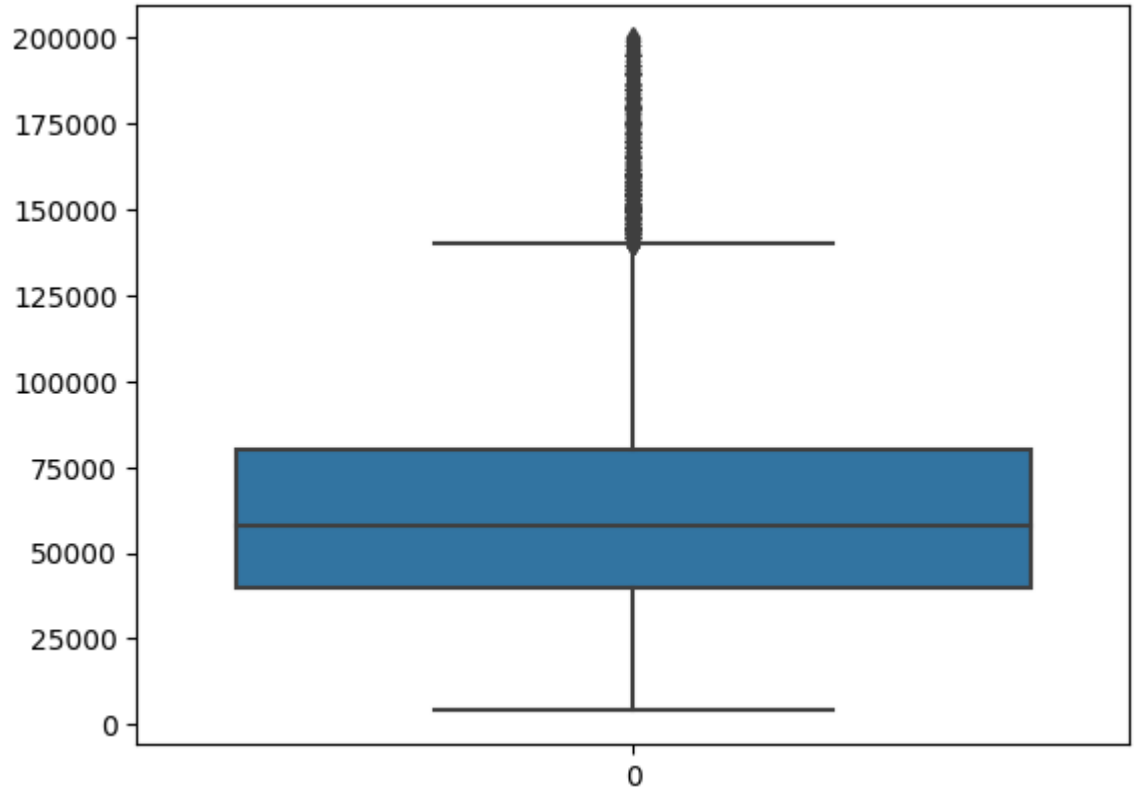
Srinivas Vaskuri

# Data cleaning and Manipulation

- Dataset has 39717 rows and 111 columns.
- Identified the columns which has all NA values and deleted them from the dataset.
- Identified the columns which have only one value. For ex : **application_type** – INDIVIDUAL. Deleted these columns also from the dataset.
- **int_rate** column has % included in every value. Removed the % and converted the type of the column to integer.
- **term** has ‘months’ in the values. Removed ‘months’ and converted the column to integer.
- **id** and **member_id** has all unique values and will not be used in the analysis. Deleted these columns from the dataset.
- There are columns which has data related to the customer behavior after the loan was disbursed. For ex : **revol_bal** - Total credit revolving balance, **total_rec_late_fee** - Late fees received to date
- Removed ‘years’ and special characters from **emp_length** column. Replaces < 1 to 0 and 10+ to 10. Converted this column to integer.
- All the records where loan_status is Current are removes, since these loan doesn’t have a conclusion yet.

# Data cleaning and Manipulation

- Removed the outliers based on the annual_inc. Removed the record with annual_inc higher than 127000 based on the box plot.

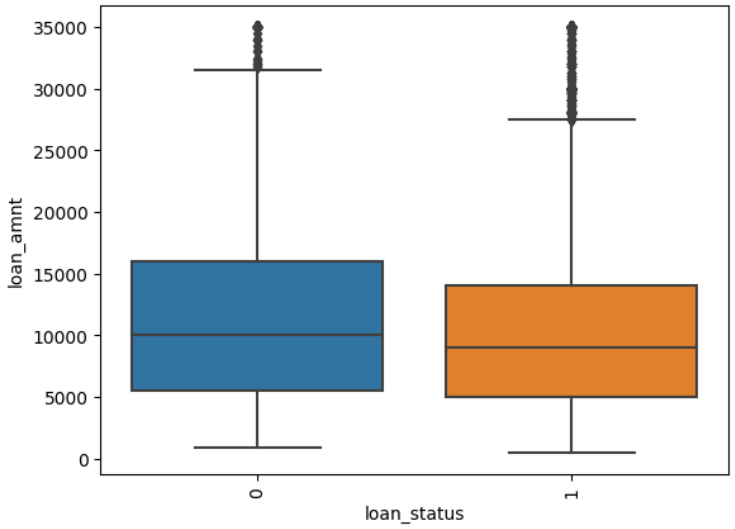


# Analysis

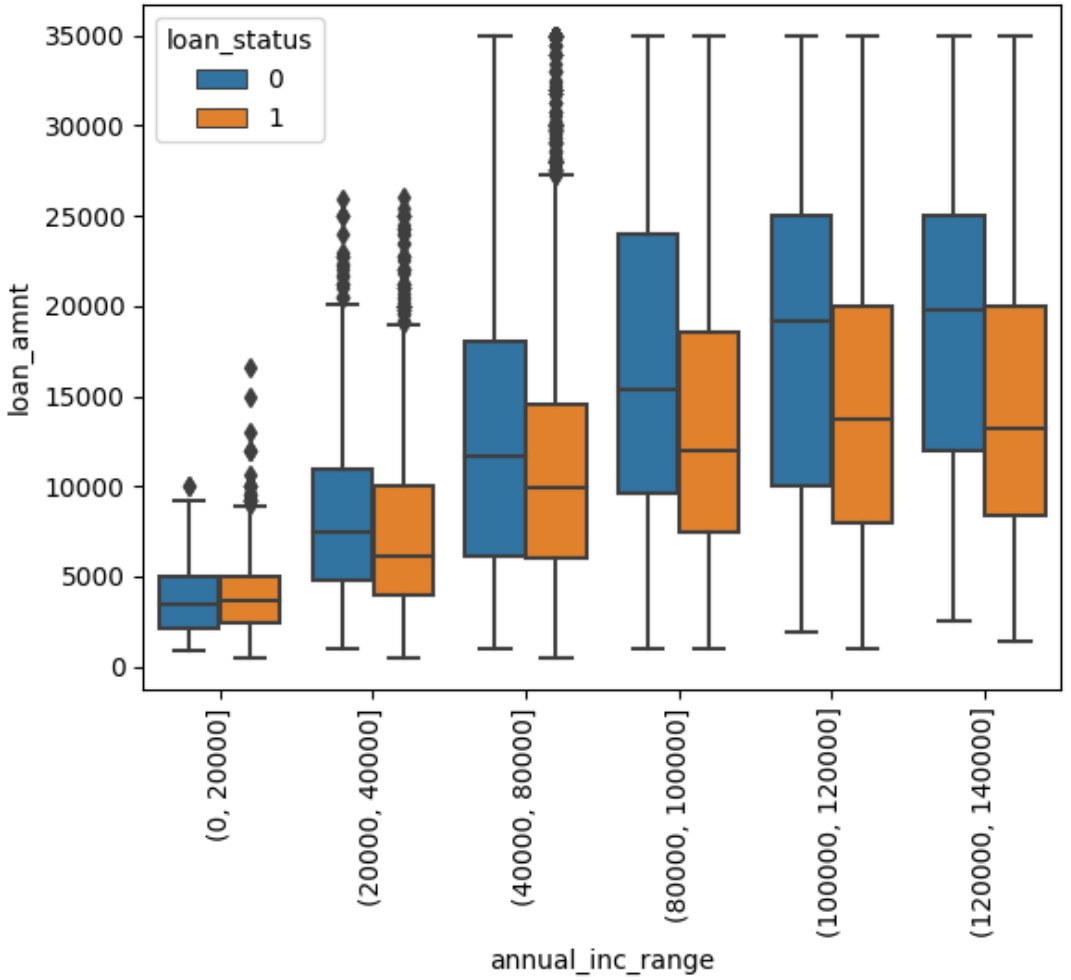
- After cleaning the columns, we  are left with the below columns

- Continuous : annual_inc, installment, int_rate, loan_amnt,
- Categorical : loan_status, purpose, grade, sub_grade, term, verification_status, addr_state, emp_length, home_ownership
- Derived **annual_inc_range** column from **annual_inc** to segregate the data based on the range.

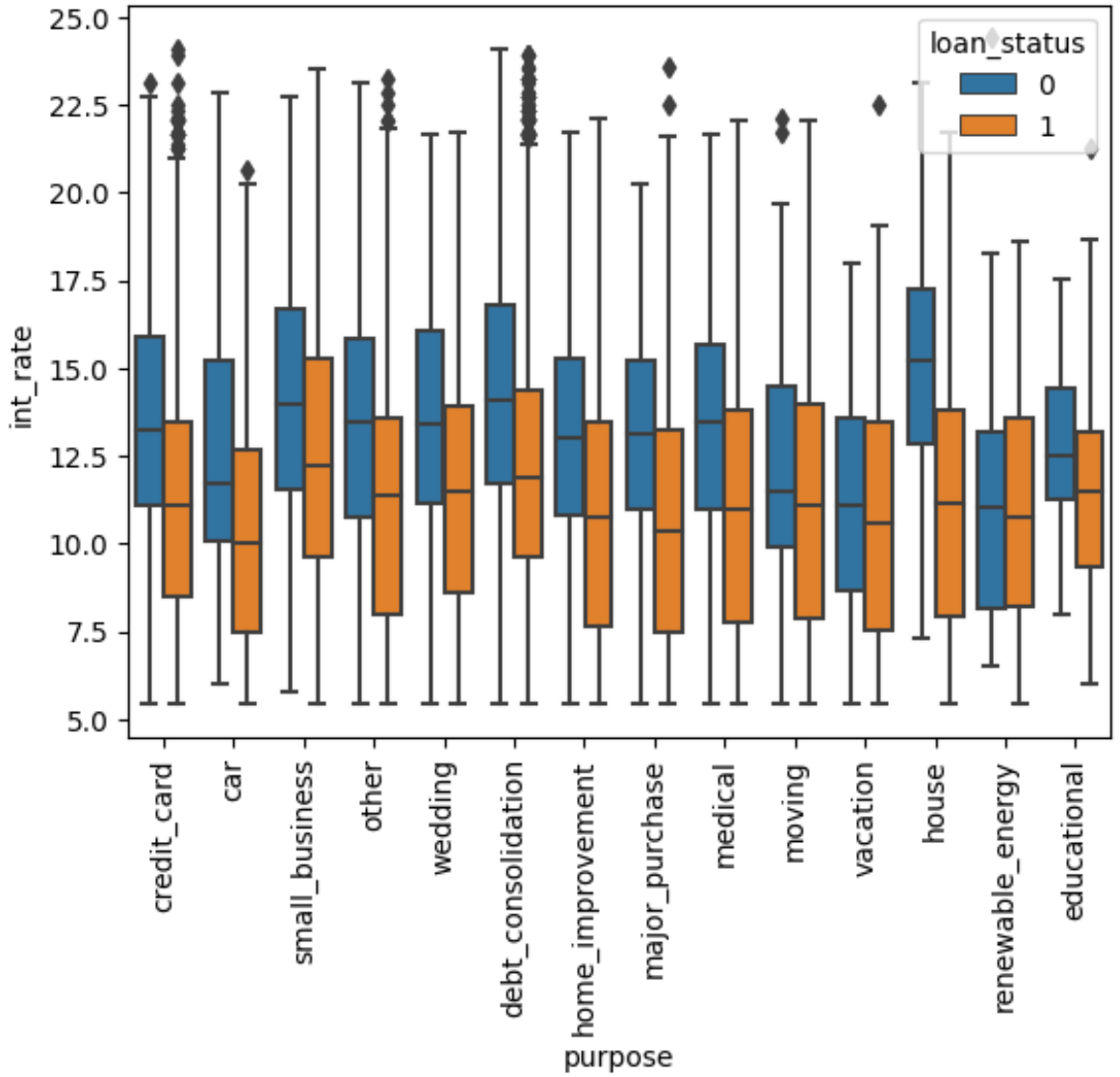
- Our target variable is loan_status.
- We need to check how loan_status changes and which are the driver variables




Loan repayment status is not getting affected by the size of loan amount much




loan_amt when combined with income range shows that higher the loan amount, the defaults of loan increases. This is more relevant in the higher income groups.




The box plot clearly shows that int_rate is an important driver in loan repayment.

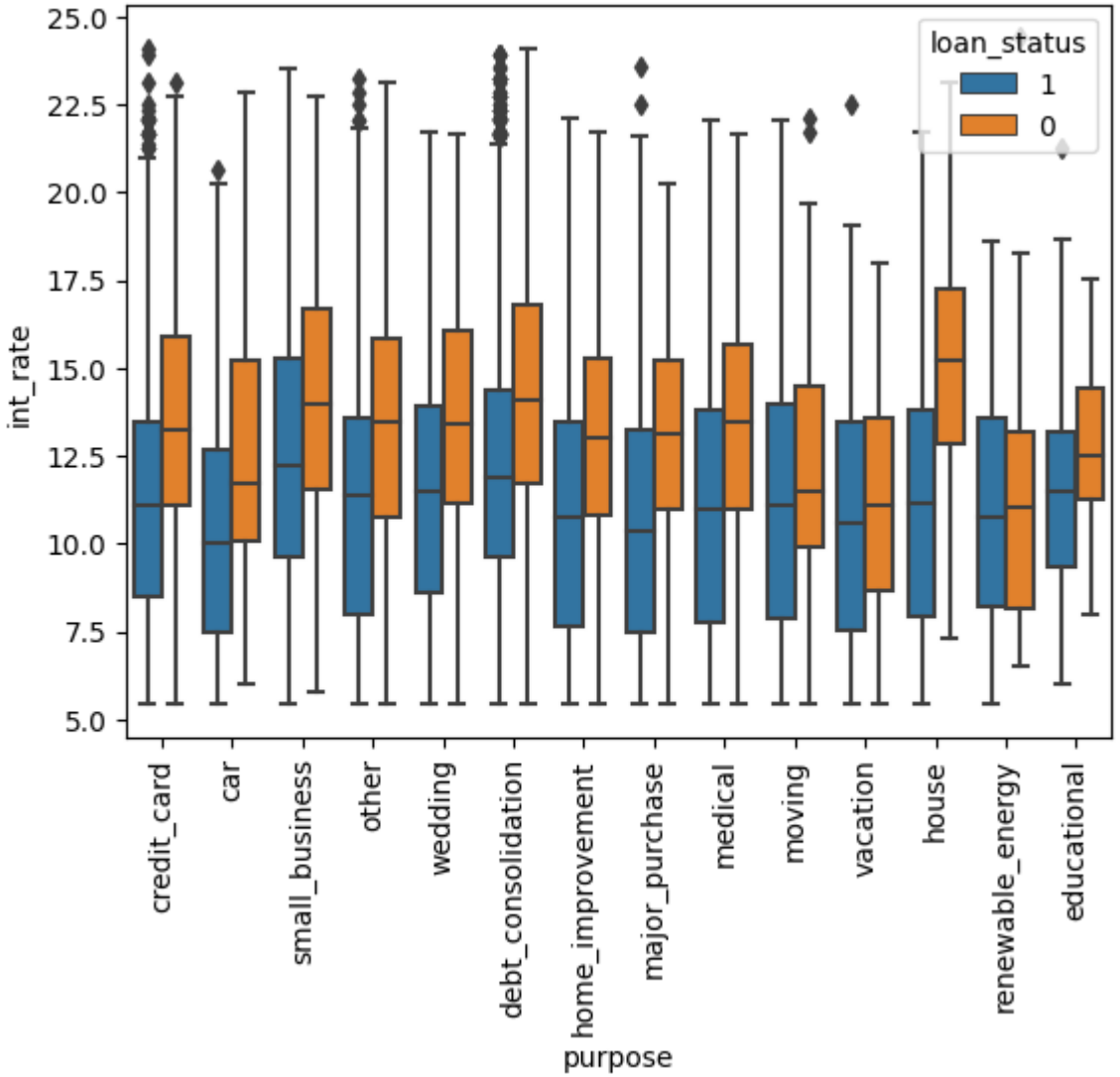
| purpose | loan_amnt | int_rate | installment | emp_length | annual_inc | loan_status |
|---|---|---|---|---|---|---|
| **small_business** | 12656.689918 | 12.851211 | 363.569391 | 4.374128 | 62437.207806 | 0.717819 |
| **renewable_energy** | 8008.333333 | 11.285914 | 225.941505 | 5.268817 | 59413.662688 | 0.806452 |
| **educational** | 6618.971061 | 11.627974 | 211.360965 | 3.340836 | 47228.169068 | 0.826367 |
| **moving** | 5879.575472 | 11.519660 | 175.395698 | 3.407547 | 51313.738094 | 0.832075 |
| **other** | 7574.856204 | 11.660507 | 225.828036 | 4.499863 | 54918.698447 | 0.835935 |
| **medical** | 7779.266348 | 11.425247 | 224.815167 | 4.827751 | 56348.509777 | 0.838915 |
| **house** | 11742.771084 | 12.012410 | 337.652801 | 4.487952 | 61678.241446 | 0.840361 |
| **debt_consolidation** | 12055.568986 | 12.328587 | 351.147284 | 4.869399 | 59201.469468 | 0.844458 |
| **vacation** | 5270.334262 | 10.790111 | 156.596267 | 5.125348 | 54850.712117 | 0.857939 |
| **home_improvement** | 10199.877301 | 11.157894 | 288.057591 | 5.548466 | 66894.148613 | 0.873211 |
| **car** | 6597.058824 | 10.477640 | 185.930427 | 4.683473 | 56659.465420 | 0.889356 |
| **credit_card** | 11139.995716 | 11.559501 | 332.570088 | 4.688303 | 61766.576894 | 0.889889 |
| **major_purchase** | 7623.854011 | 10.775755 | 223.245655 | 4.491779 | 56293.393279 | 0.893373 |
| **wedding** | 9358.803222 | 11.731853 | 279.895259 | 3.855006 | 59990.561623 | 0.896433 |

Observations based on above table

1.Repayment of loans related to **Life style** related expences like wedding, purchase, credit card, car and home_improvement is high.

2.Repayment of loans related to Small business loans, Renewable energy, Educational is very low.

You can see the higher the interest rate the probability of re-paying the loan is less. This is primarily seen in house, car and small business loans



| annual_inc_range<br>purpose | (0, 20000] | (20000, 40000] | (40000, 80000] | (80000, 100000] | (100000, 120000] | (120000, 140000] |
|---|---|---|---|---|---|---|
| **car** | 0.855072 | 0.877384 | 0.893297 | 0.900000 | 0.905882 | 0.937500 |
| **credit_card** | 0.901961 | 0.856405 | 0.892380 | 0.908065 | 0.918699 | 0.936508 |
| **debt_consolidation** | 0.810748 | 0.813657 | 0.847784 | 0.873624 | 0.886660 | 0.863636 |
| **educational** | 0.833333 | 0.806122 | 0.841121 | 0.840000 | 0.842105 | 0.500000 |
| **home_improvement** | 0.818182 | 0.848558 | 0.857827 | 0.895928 | 0.945098 | 0.936170 |
| **house** | 0.777778 | 0.788732 | 0.870787 | 0.847826 | 0.750000 | 0.875000 |
| **major_purchase** | 0.821782 | 0.858195 | 0.905095 | 0.934211 | 0.943925 | 0.888889 |
| **medical** | 0.625000 | 0.839779 | 0.829114 | 0.910448 | 0.945946 | 0.500000 |
| **moving** | 0.777778 | 0.764706 | 0.876543 | 0.829268 | 0.923077 | 1.000000 |
| **other** | 0.750000 | 0.816888 | 0.843262 | 0.896040 | 0.860465 | 0.724138 |
| **renewable_energy** | 0.600000 | 0.740741 | 0.829268 | 0.900000 | 0.888889 | 1.000000 |
| **small_business** | 0.645833 | 0.689759 | 0.703337 | 0.776000 | 0.791304 | 0.782609 |

- Educational and Medical loans in the higher income customers shows higher defaults.
- Also, renewable energy, medical and small_business loans in the lower income groups see higher defaults.

# Grade Vs Annual income

| annual_inc_range | (0, 20000] | (20000, 40000] | (40000, 80000] | (80000, 100000] |
|---|---|---|---|---|
| grade | | | | |
| A | 0.882353 | 0.909982 | 0.942948 | 0.965004 |
| B | 0.823848 | 0.840398 | 0.880444 | 0.905395 |
| C | 0.744275 | 0.790685 | 0.827514 | 0.860438 |
| D | 0.718563 | 0.740741 | 0.779600 | 0.812395 |
| E | 0.709091 | 0.678899 | 0.713505 | 0.786704 |
| F | 0.777778 | 0.637931 | 0.629386 | 0.745098 |
| G | 0.333333 | 0.441176 | 0.638655 | 0.650000 |

| annual_inc_range | (100000, 120000] | (120000, 140000] |
|---|---|---|
| grade | | |
| A | 0.973597 | 0.965517 |
| B | 0.924585 | 0.916667 |
| C | 0.875576 | 0.872093 |
| D | 0.824138 | 0.793103 |
| E | 0.801020 | 0.800000 |
| F | 0.768421 | 0.600000 |
| G | 0.727273 | 0.600000 |

Pivot table to compare grades against income range and see the loan status
**obsevation** Low loan grades for the customer with annual income less than 40k is resulting in higher defaults

# Observations

1. Repayment of loans related to **Life style** related expenses like wedding, purchase, credit card, car and **home_improvement** is high.
2. Repayment of loans related to Small business loans, Renewable energy, Educational is very low.
3. You can see the higher the interest rate the probability of re-paying the loan is less. This is primarily seen in house, car and small business loans. May be higher interest rate indicate the loans are riskier.
4. Educational and Medical loans in the higher income customers shows higher defaults.
5. Also, renewable energy, medical and **small_business** loans in the lower income groups see higher defaults.
6. Avoid low graded loans for the lowers income groups, they have very low re-payment rate.