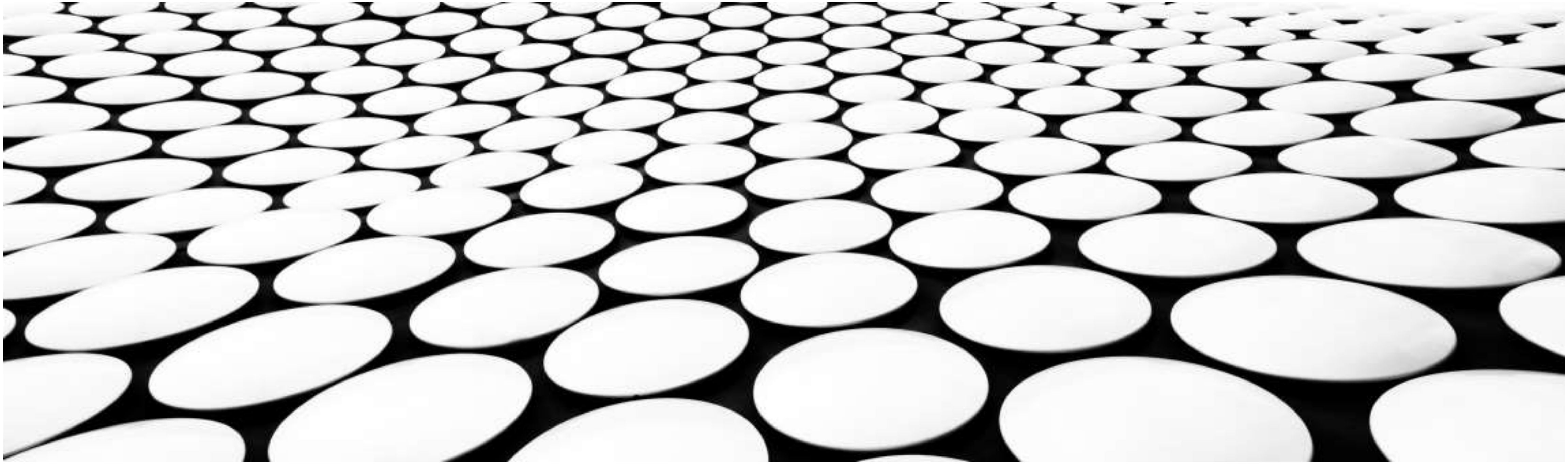
LENDING CLUB CASE STUDY

APURV AGGARWAL

AREEESHA ANJUM



PROBLEM STATEMENT

- **Domain:** Risk Analytics in Banking & Financial Services
- **Business Objective:** Minimize the risk of losing money while lending to customers.
- **Aim:** Identify driver factors behind loan default using EDA.
- **Need:** When the borrower refuses or runs away with the money owed, it causes the largest loss to lenders. Through the driver factors, we can identify risky loan applicants, reducing the amount of credit loss.

DATA OVERVIEW

- Number of records – 39717
- Number of columns – 111
- Number of null columns – 54
- Number of columns with the same value – 9
- Numerical Columns – 27
- Categorical Columns - 21

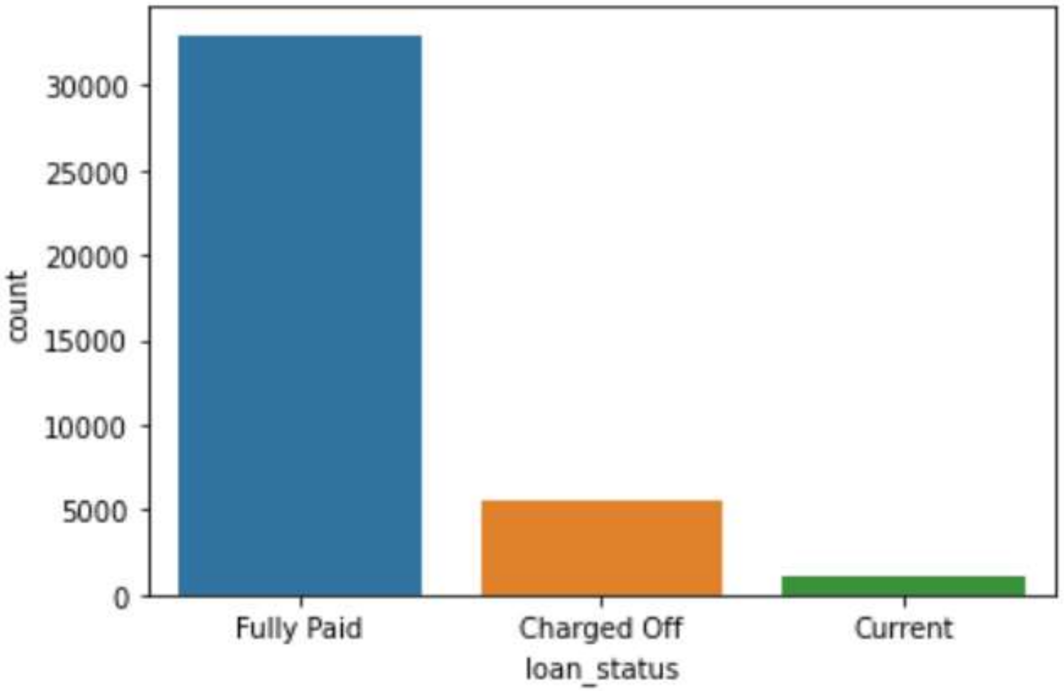
STEP 1: DATA CLEANING

- We checked and removed null columns and columns with the same value, 63 columns were found.
- We checked for duplicate rows, 0 records were found.
- We discarded columns with more than 50% missing values, These columns **[mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d’]** were discarded.
- We did not choose to impute missing values in columns less than 50% missing values to preserve data integrity as we have enough data for analysis.
- We did not consider redundant columns like **[‘funded_amnt’, ‘funded_amnt_inv’].**
- Converted the columns to correct datatype wherever required. Example: **‘issue_d’** was converted to datetime.
- We also divided the columns into numerical and categorical classes for better analysis.

STEP 2: UNIVARIATE ANALYSIS (CATEGORICAL)

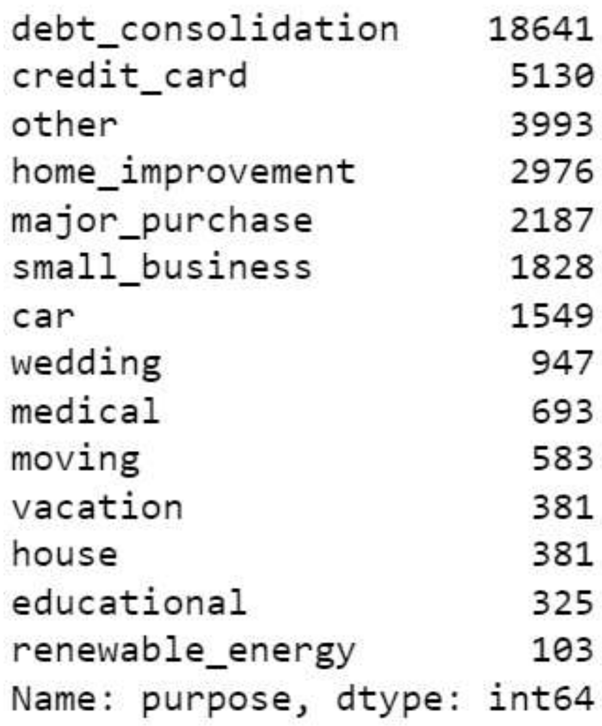
loan_status

- Target variable. All the analysis will be made comparatively to this column.
- Imbalanced column since ‘**Charged Off**’ only corresponds to **14.16** % in the entire dataset.



purpose

- Purpose of loan.
- About **59.85 %** of loans were taken to repay existing debts.

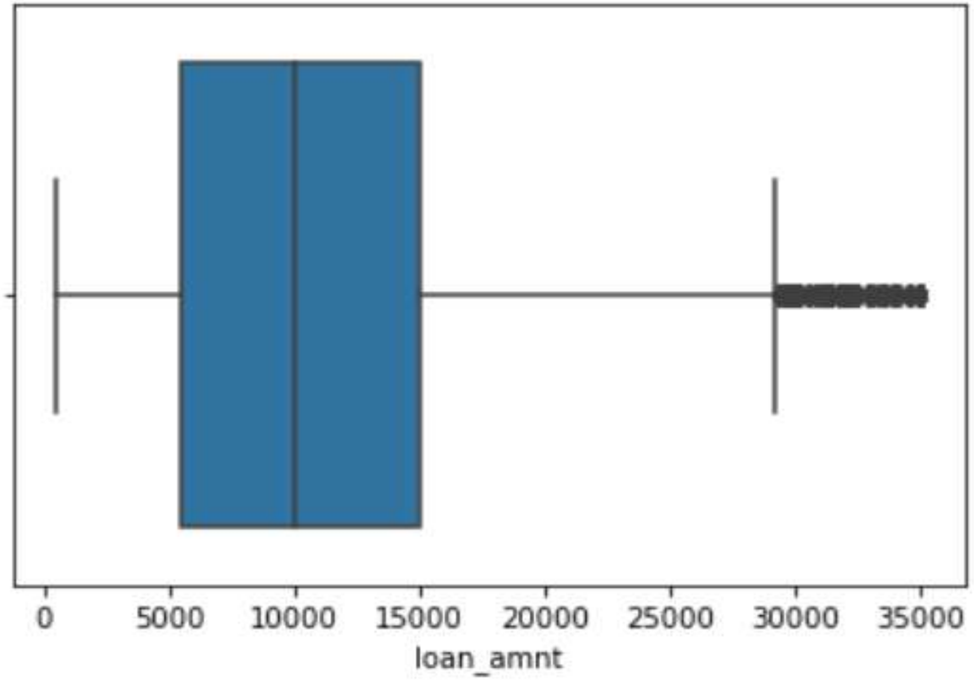


```
debt_consolidation    18641
credit_card            5130
other                  3993
home_improvement       2976
major_purchase         2187
small_business         1828
car                    1549
wedding                 947
medical                 693
moving                  583
vacation                381
house                   381
educational             325
renewable_energy        103
Name: purpose, dtype: int64
```

STEP 3: UNIVARIATE ANALYSIS (NUMERICAL)

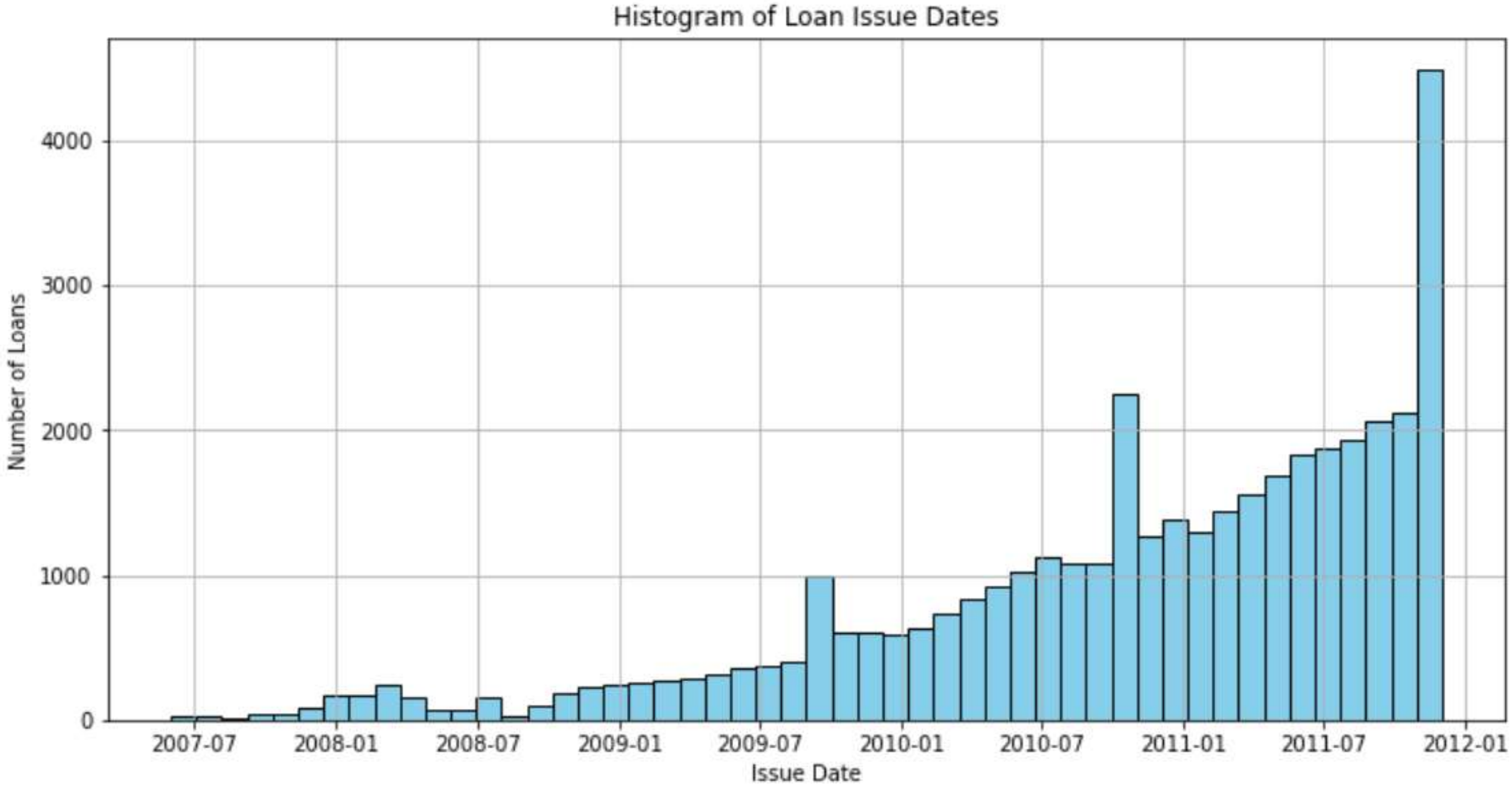
loan_amnt

- Amount requested by the borrower.
- **1613** outliers were found based on the **percentile** method. We checked the outliers, they seemed legit.



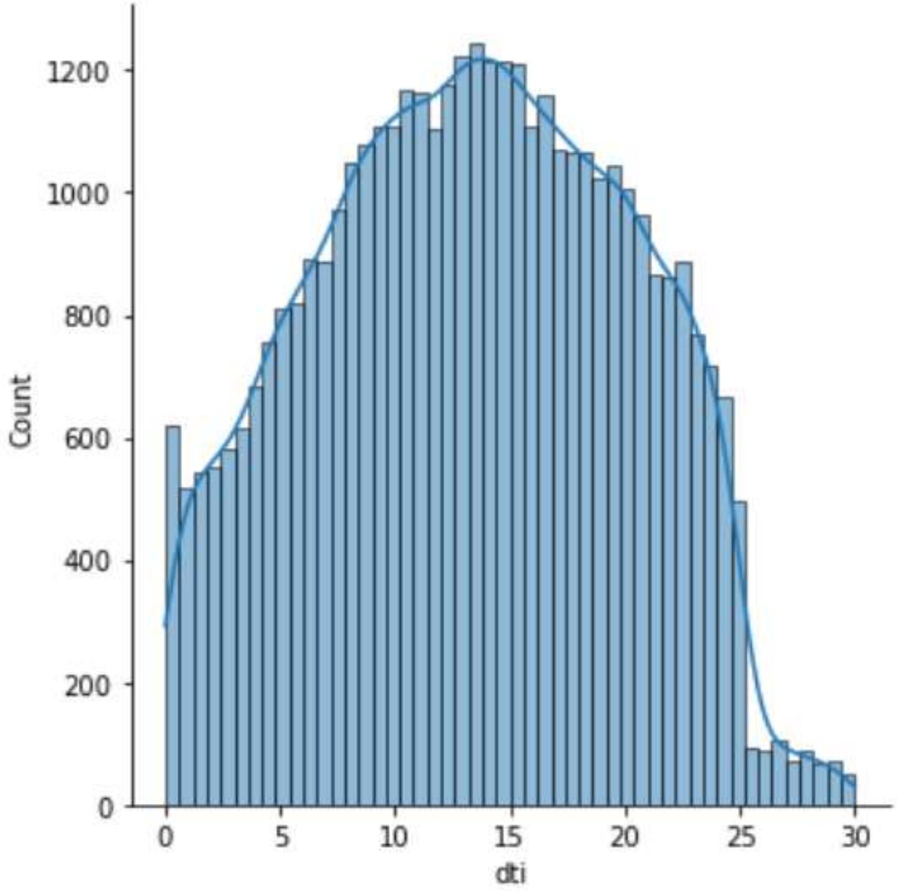
issue_d

- Month on which loan was funded.
- The number of loans funded **increases** by the year, hence marking the growth of the company.



dti

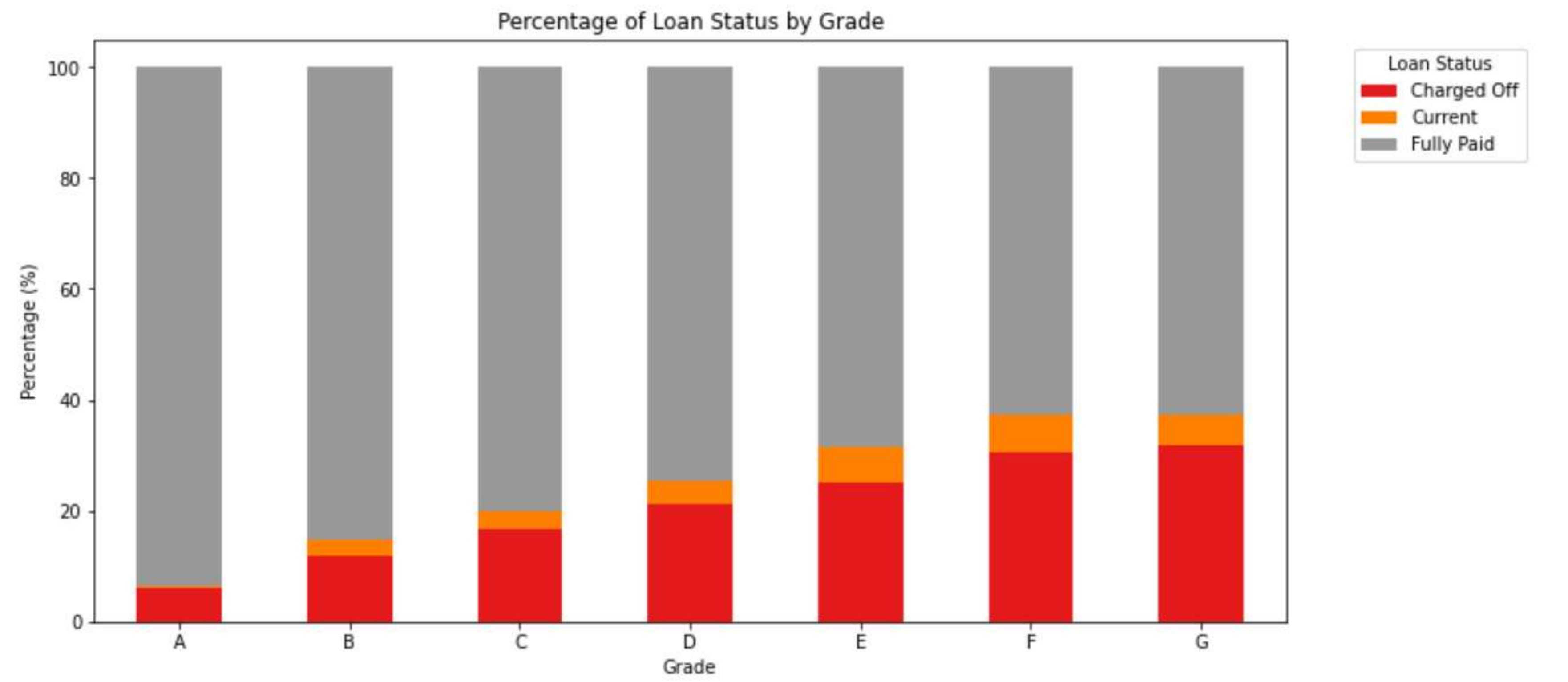
- debt-to-income ratio.
- The maximum value for ‘dti’ is 30%, which means beyond this value loans are simply rejected.



STEP 4: BIVARIATE ANALYSIS (CATEGORICAL)

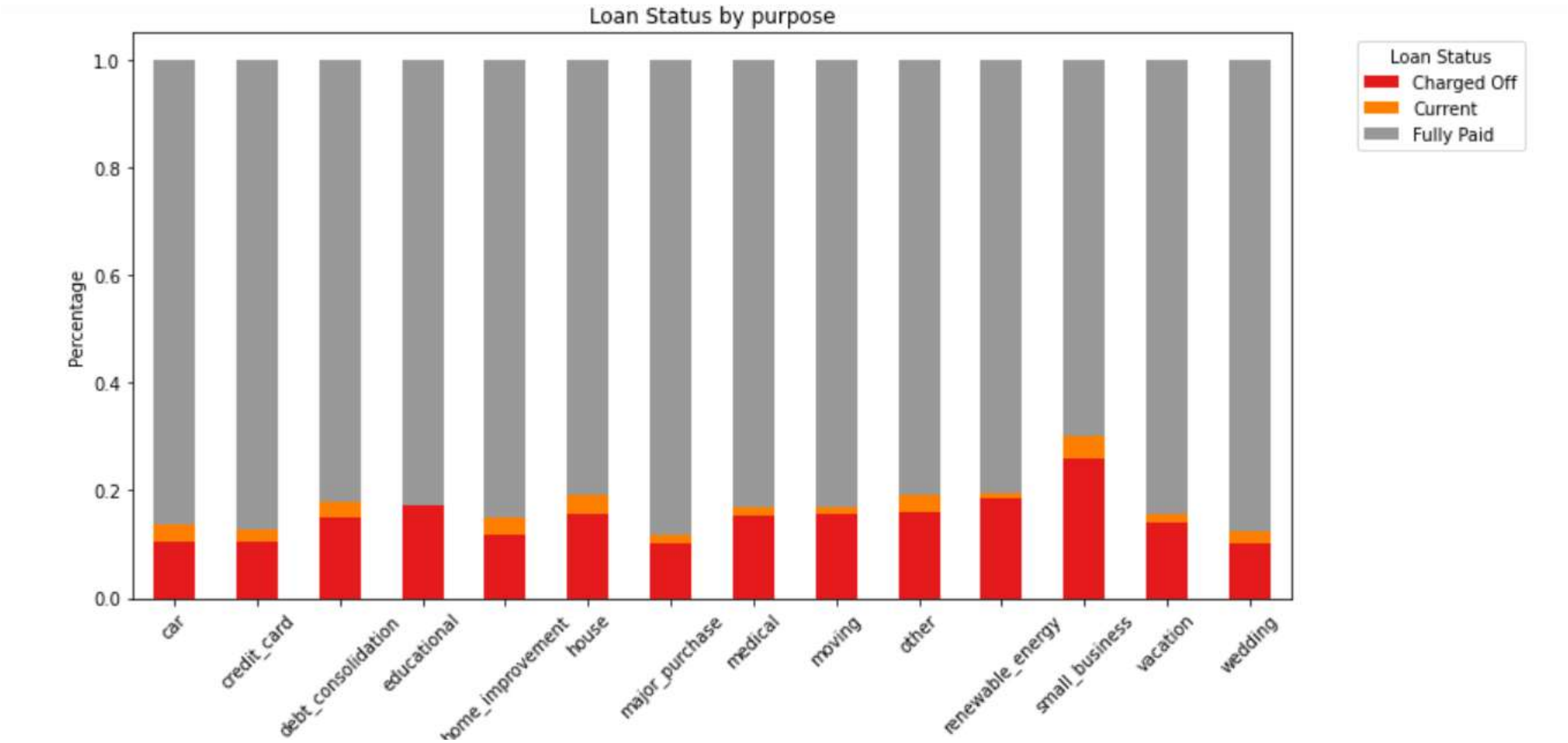
grade

- Assigned loan grade.
- As the ‘**grade**’ decreases, the ‘**Charged Off**’ rate increases.



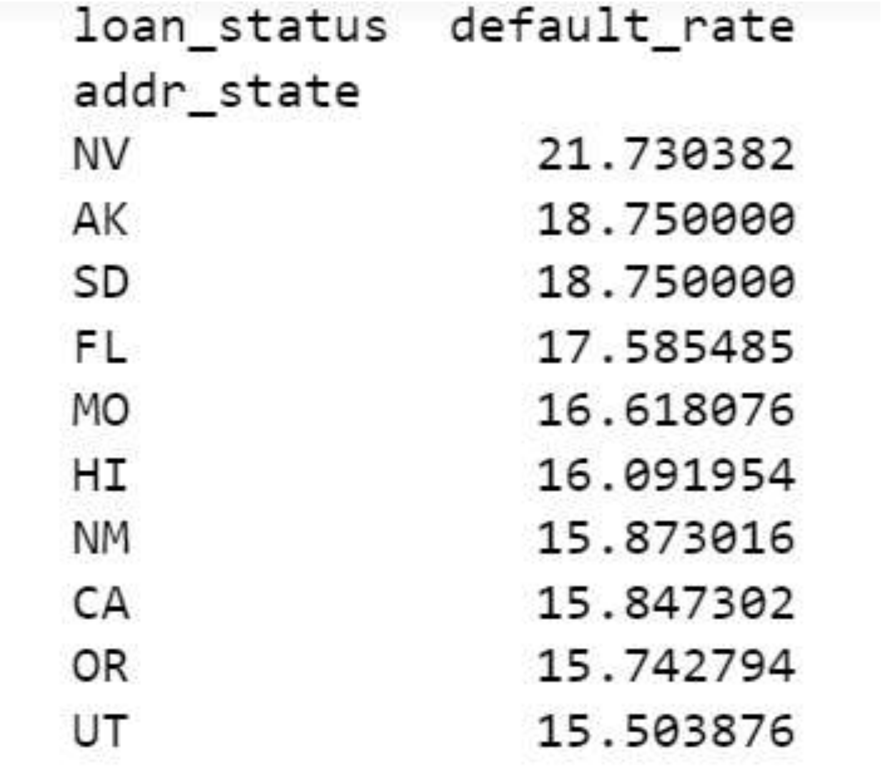
purpose

- We can see '**small_business**' are significantly more probable to '**Charged Off**'.



addr_state

- **‘NV’, ‘AK’, ‘SD’, and ‘FL’** have particularly high ‘**Charged Off**’ rates than others.

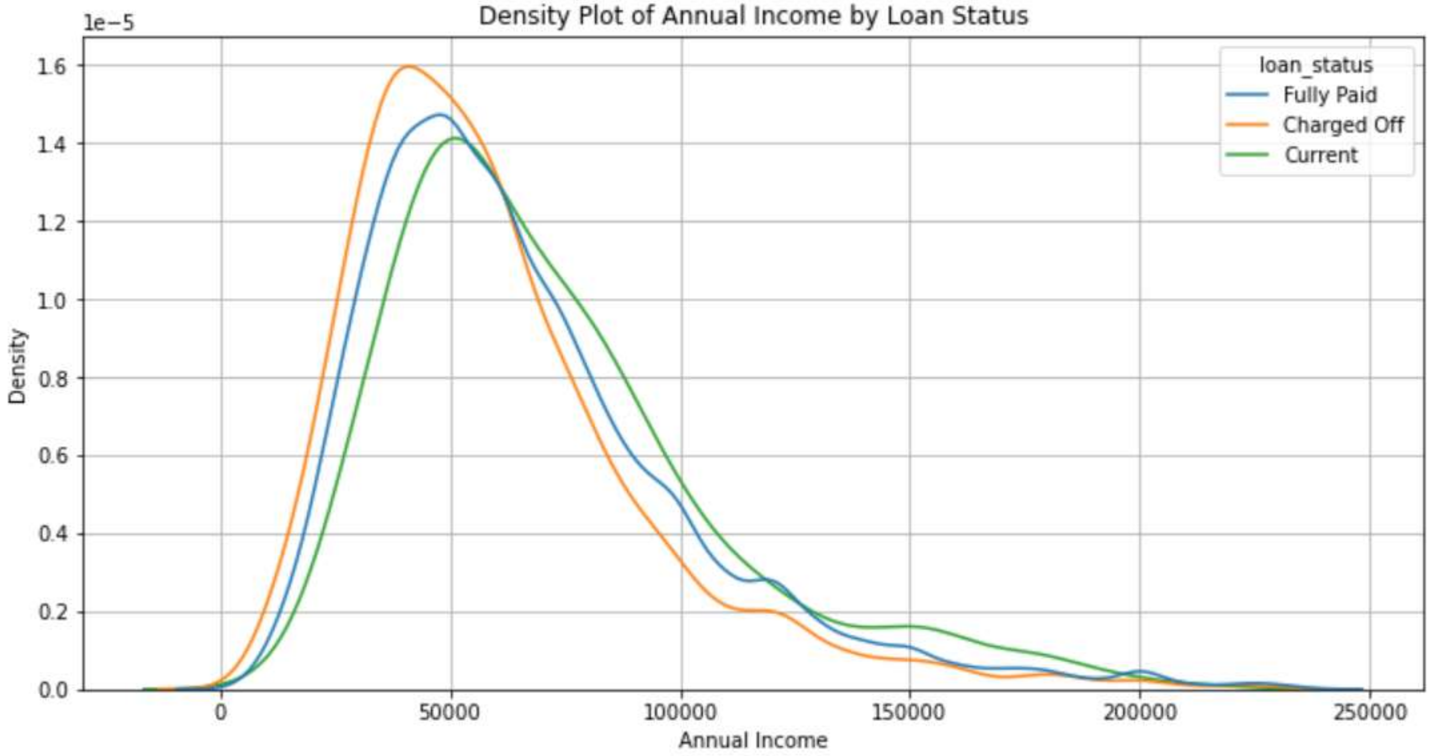


```
loan_status  default_rate
addr_state
NV              21.730382
AK              18.750000
SD              18.750000
FL              17.585485
MO              16.618076
HI              16.091954
NM              15.873016
CA              15.847302
OR              15.742794
UT              15.503876
```

STEP 5: BIVARIATE ANALYSIS (NUMERICAL)

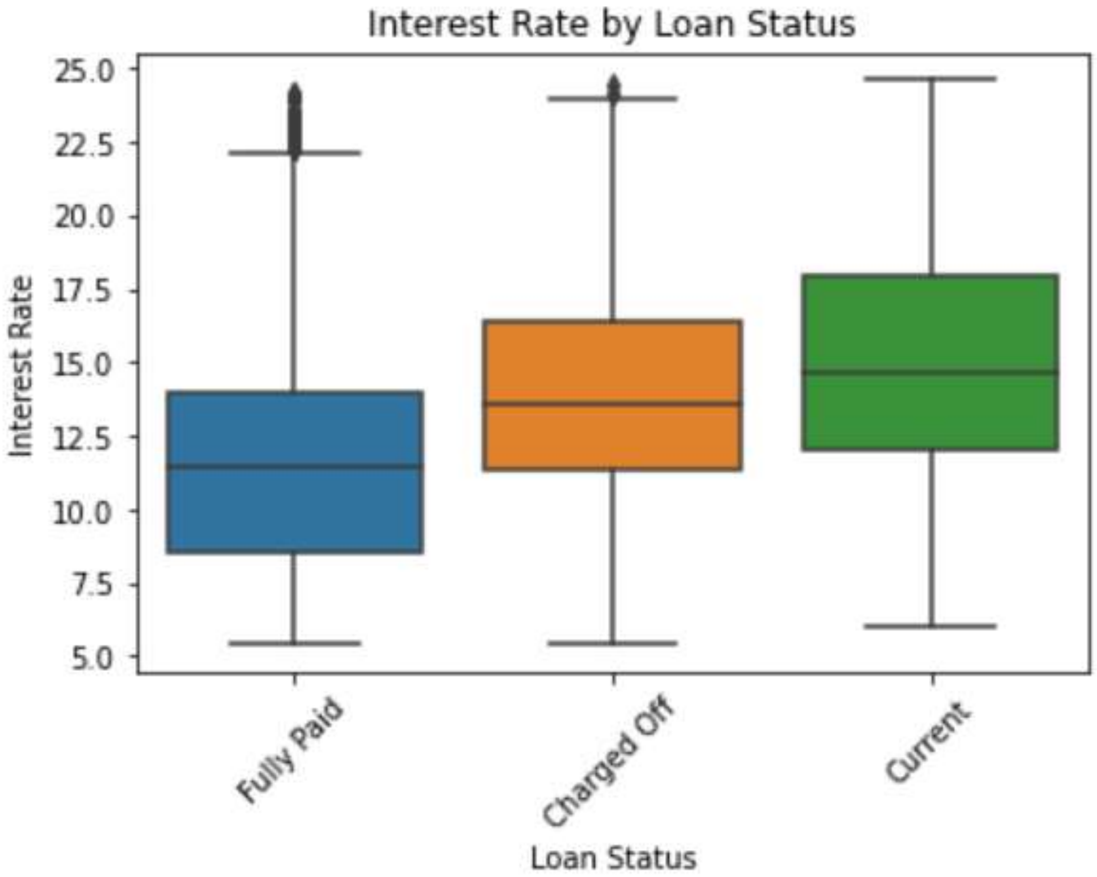
annual_inc

- Annual income of the borrower.
- People with **lower** incomes are more likely to be ‘**Charged Off**’.



int_rate

- Interest rate.
- Above **13% ‘Charged Off’** rate is more probable.



CONCLUSIONS

- grade
- purpose
- addr_state
- annual_inc
- int_rate

These are top-5 key factors responsible for influencing ‘Charged Off’ rates.