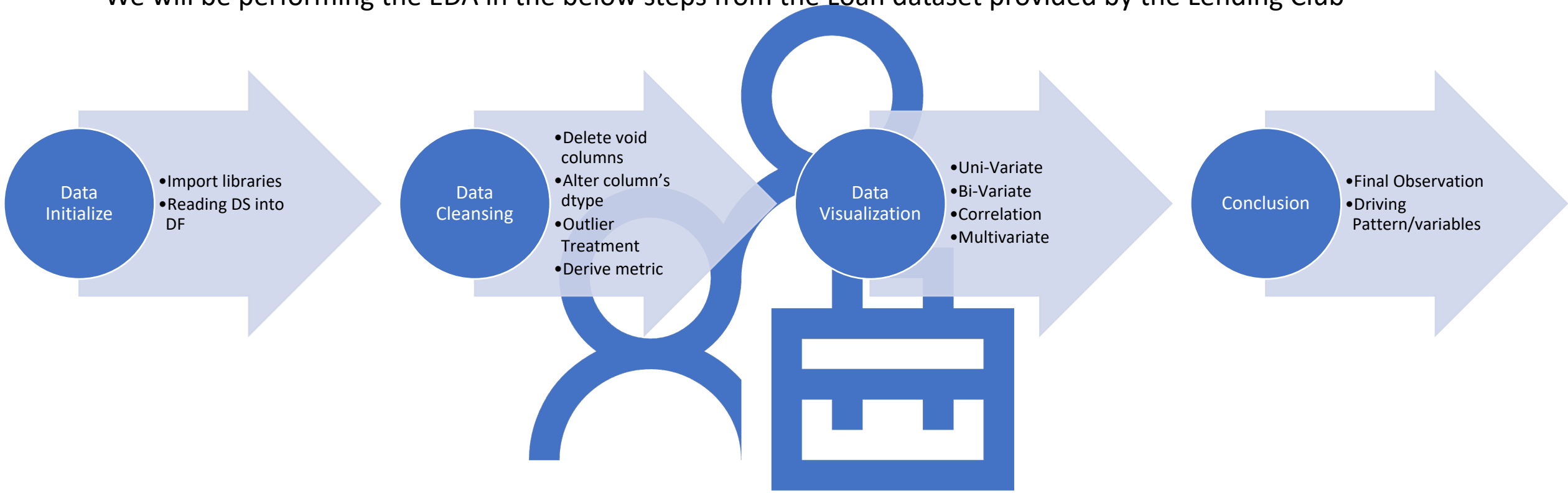
Lending Club Case Study

Exploratory Data Analysis

Srinivasa Reddy M
Shalini K

Process Of EDA

We will be performing the EDA in the below steps from the Loan dataset provided by the Lending Club



Data Initialize

Importing libraries
Numpy
Pandas
Matplotlib.pyplot
Seaborn

We read the loan dataset(loans.csv into a dataframe using above libraries) into **df**

Data Cleansing

1

In this process we remove/drop the columns which we consider void to our EDA

2

We alter the variable/column dtype to appropriate dtype by doing some transformation if needed

3

We do the outlier treatment if need on our scope

4

We derive some extra metrics as part to analyse the patterns

Columns Removed/dropped

We dropped the columns based on below criterias

- Columns that have only Nulls
- Columns that have single values
- Columns like id, member ID, and also encoded/masked data columns

Altering Dtypes

We modified the dtype for below columns to appropriate

- To Date type
  issue_d, earliest_cr_line, last_pymnt_d, and last_credit_pull_d
- To Numeric
  int_rate to float
  term to int
  emp_length to float
  revol_util to float

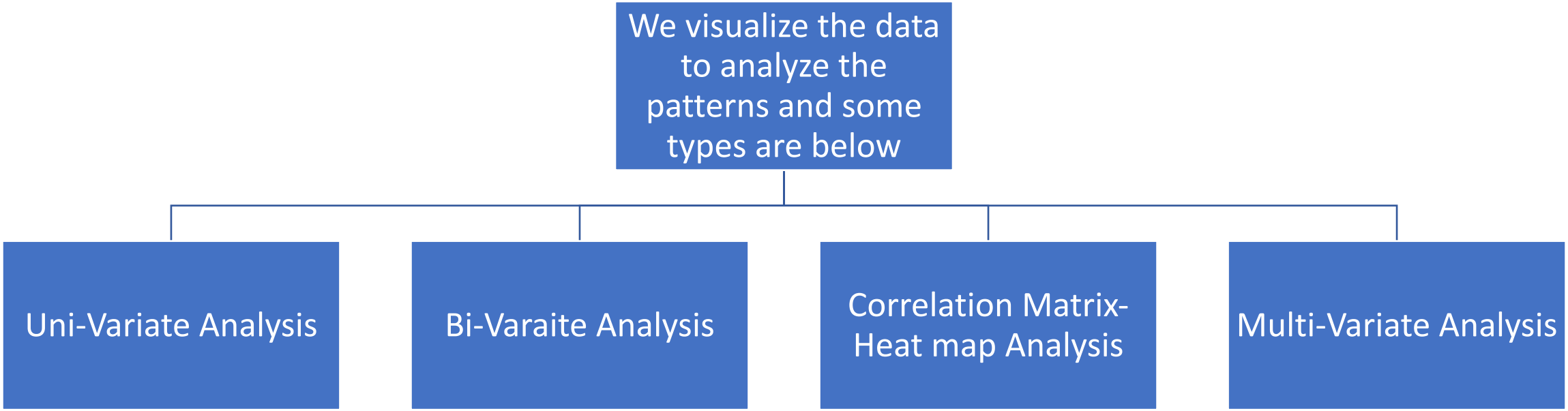
Derived Metrics

- Derived Closing date based on issue date and term of loan
- Derived year, and month from the loan issue date

Outlier Treatment

- Annual Income – had to remove outliers
- Debt to Income ratio dti – No impact
- Loan Amount – No impact
- Funded Amount – No impact

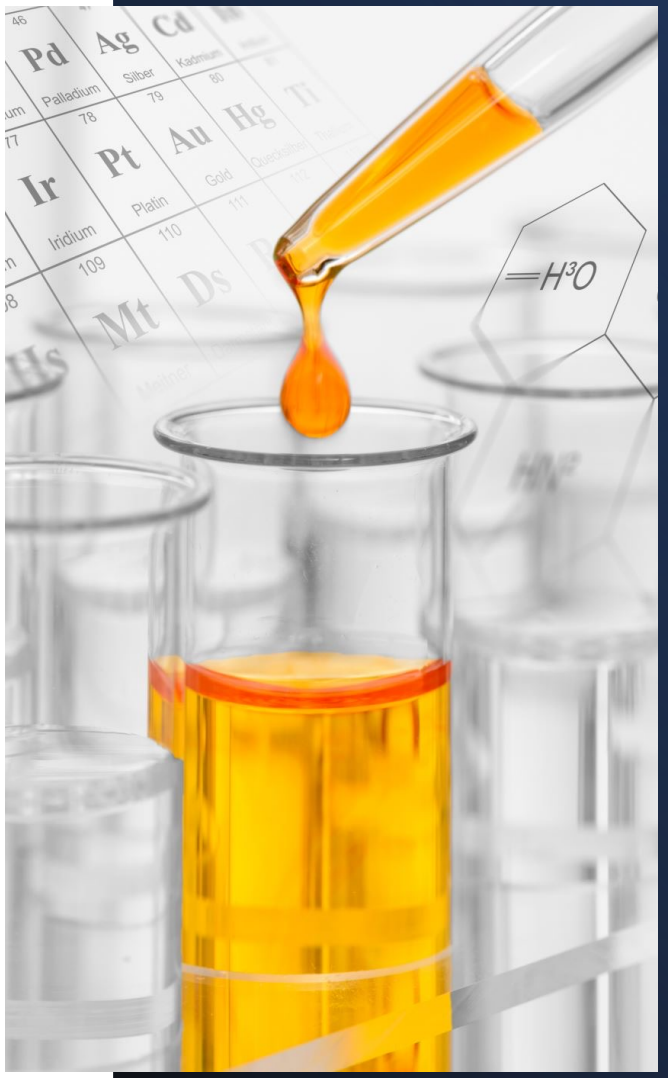
Data Visualization



Uni-Varaite Analysis

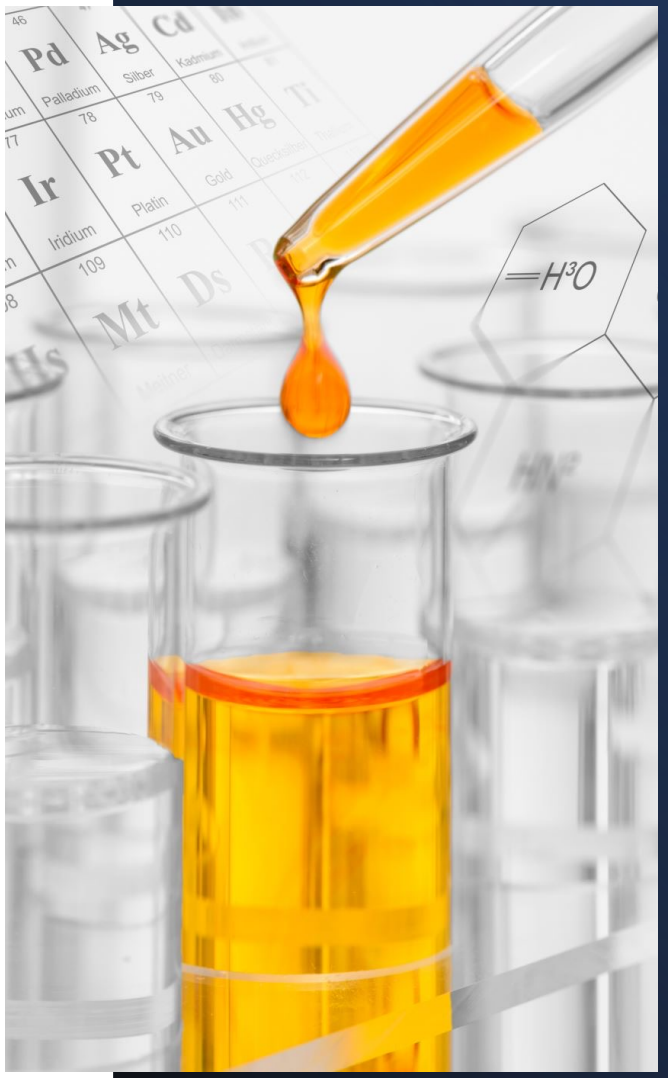
As part of uni-variate analysis, we mainly consider the dataset related to charged off or defaulters. We do it two parts

- Dealing with Variables directly
- Binning the variables



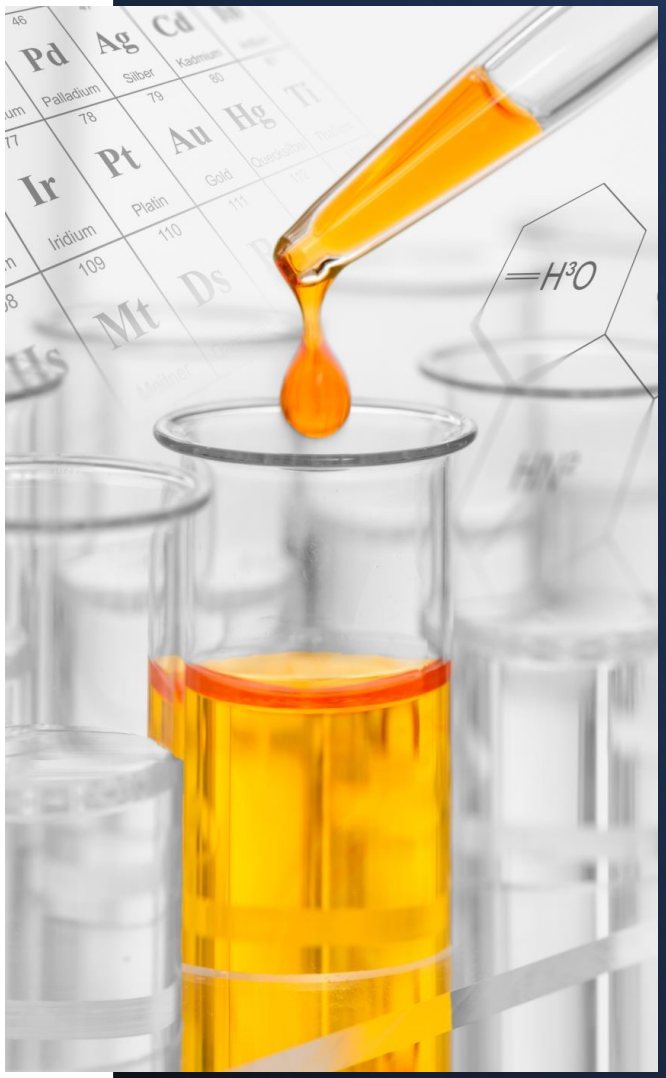
Variables in scope for Uni-Variate No Bins

- ✓ Loan Status
- ✓ Grade & sub grade
- ✓ Employee title
- ✓ Home ownership
- ✓ Purpose of loan
- ✓ State of residence
- ✓ Employment length
- ✓ Term of loan
- ✓ Recoveries
- ✓ Public record of derogatory and bankruptcies
- ✓ Inquiry in last 6 months
- ✓ Closing date
- ✓ Month and year of loan issue date
- ✓ Total payment<funded amount (segmented)
- ✓ Credit inquiry data > last payment date (segmented)

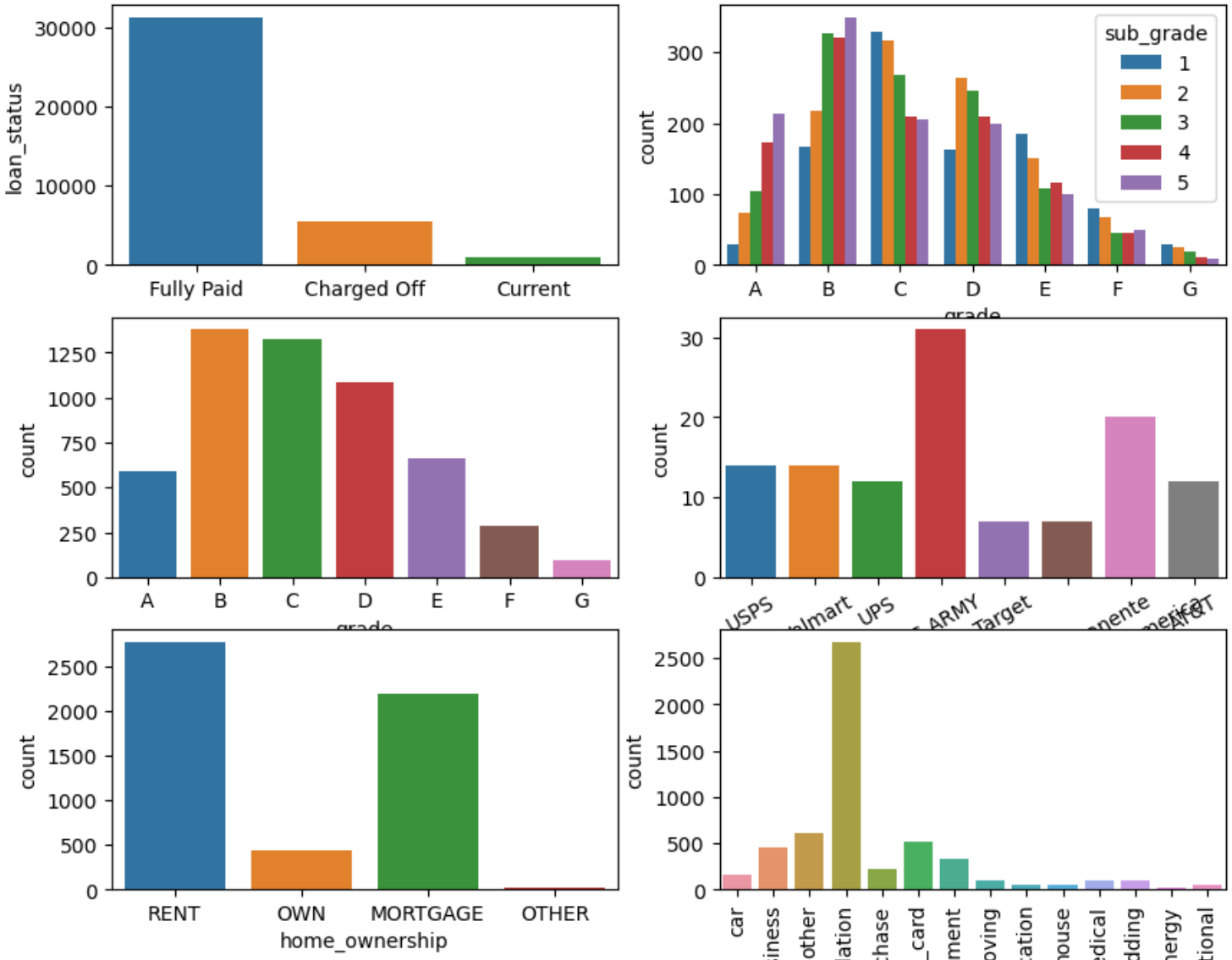


Variables in scope for Uni-Variate Bins

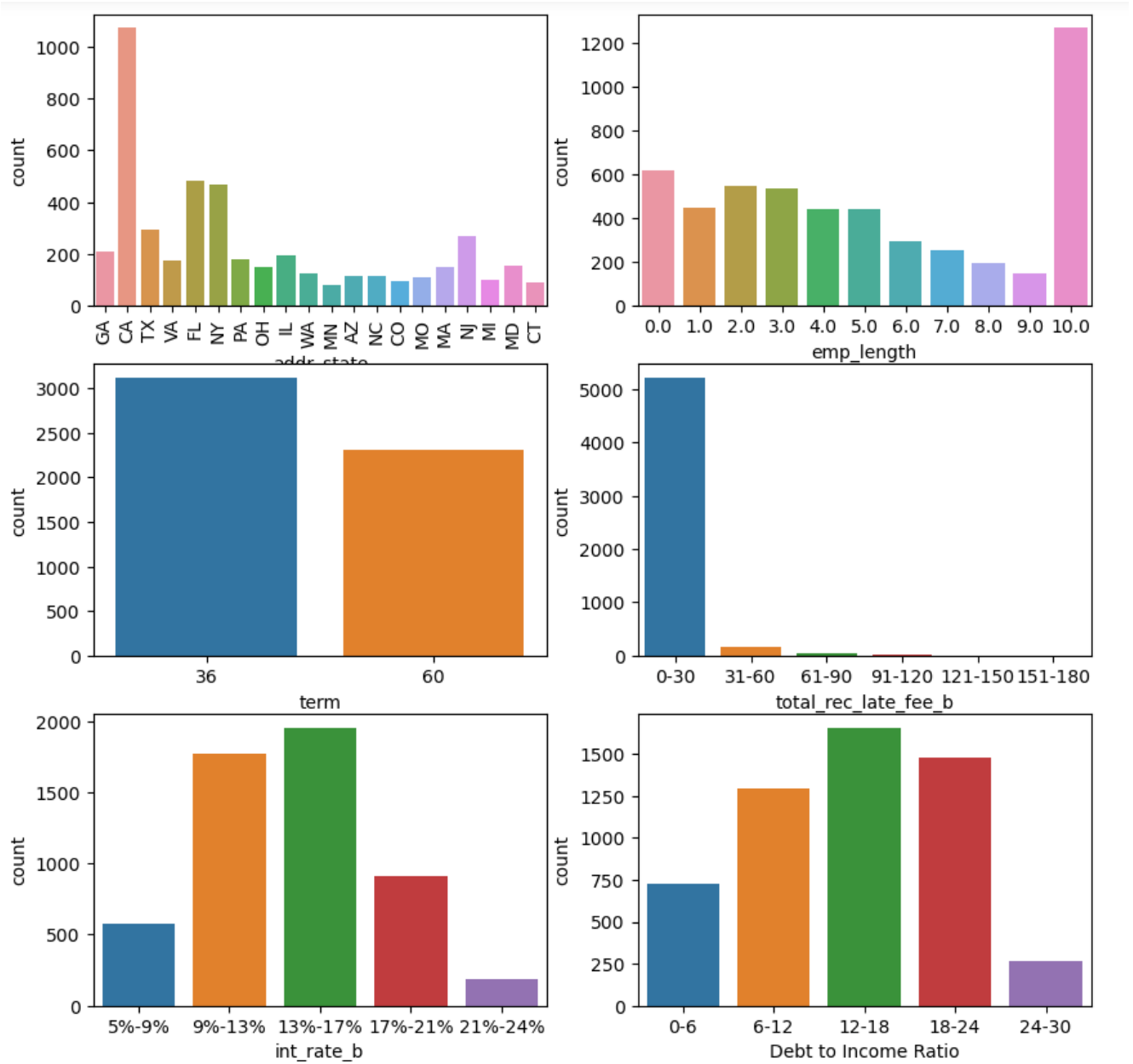
- Interest Rate
- Open Accounts
- Total Accounts
- Annual Income
- Total late recovery Fees
- Debt to Income ratio (dti)
- Installments
- Loan amounts
- Funded Amount



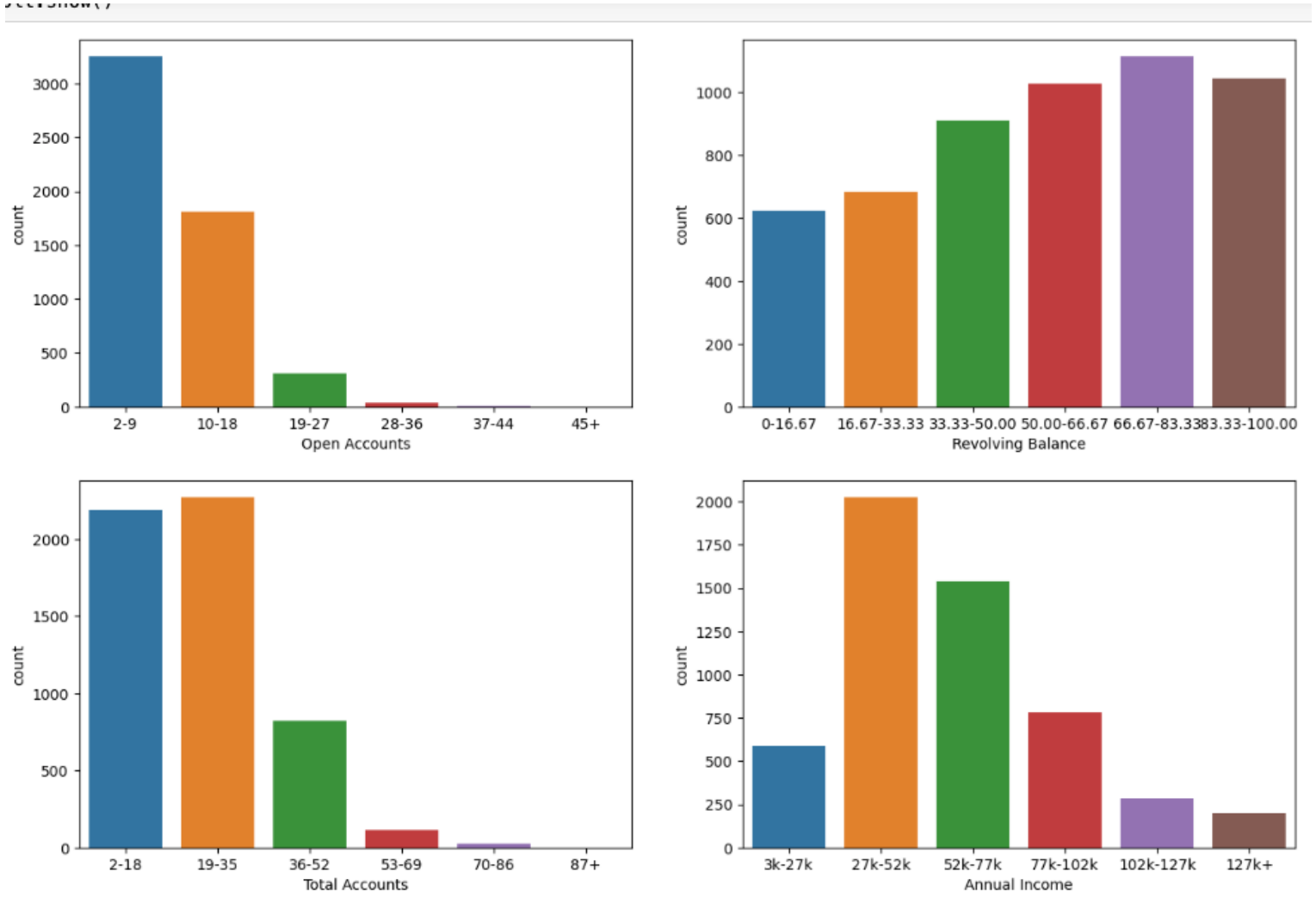
Uni-variate for loan status, grade, subgrade, title, home ownership, purpose



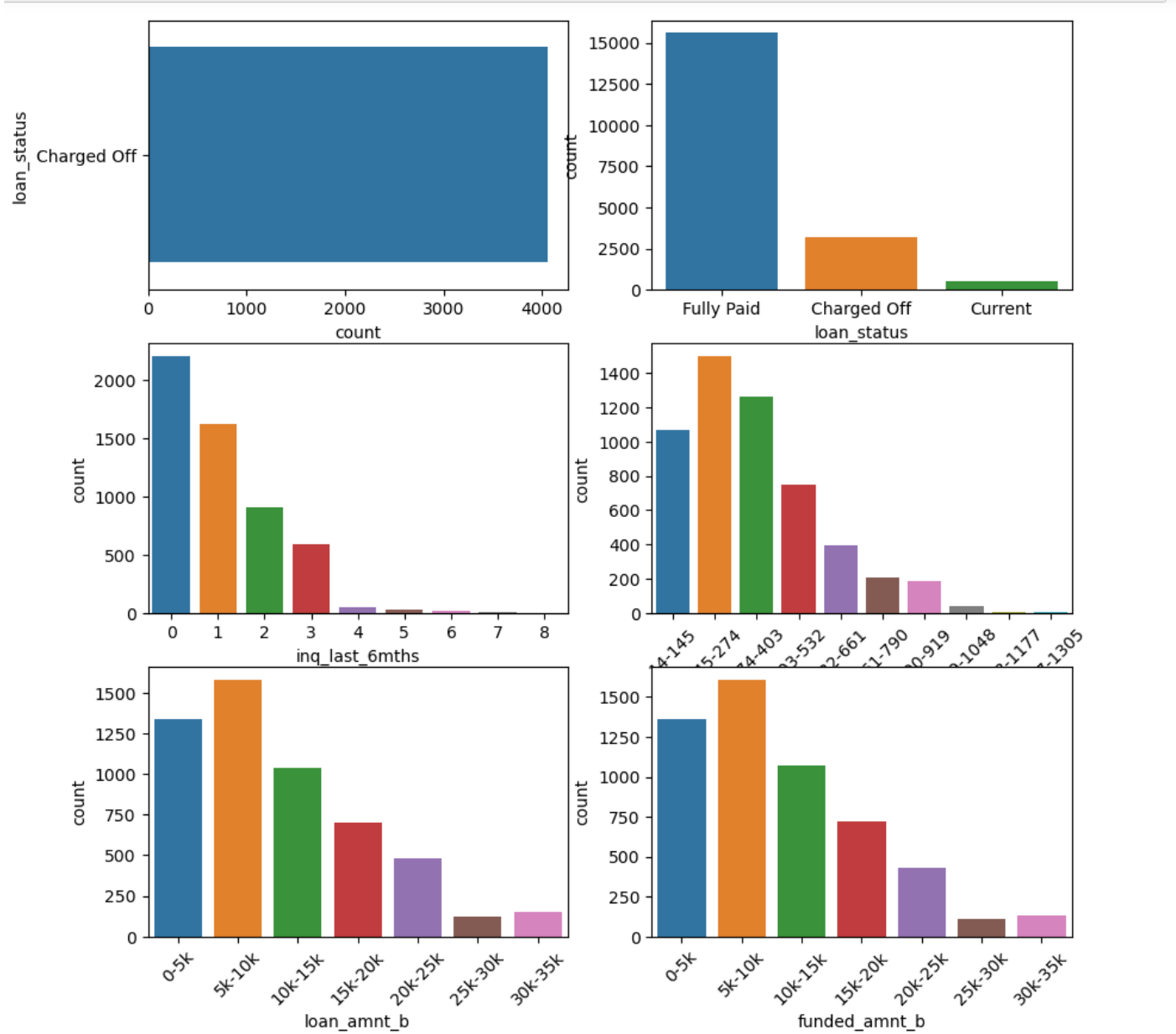
Uni-variate for state, emp_length, term, total recovery fee bins, interest rate bins, dti bins



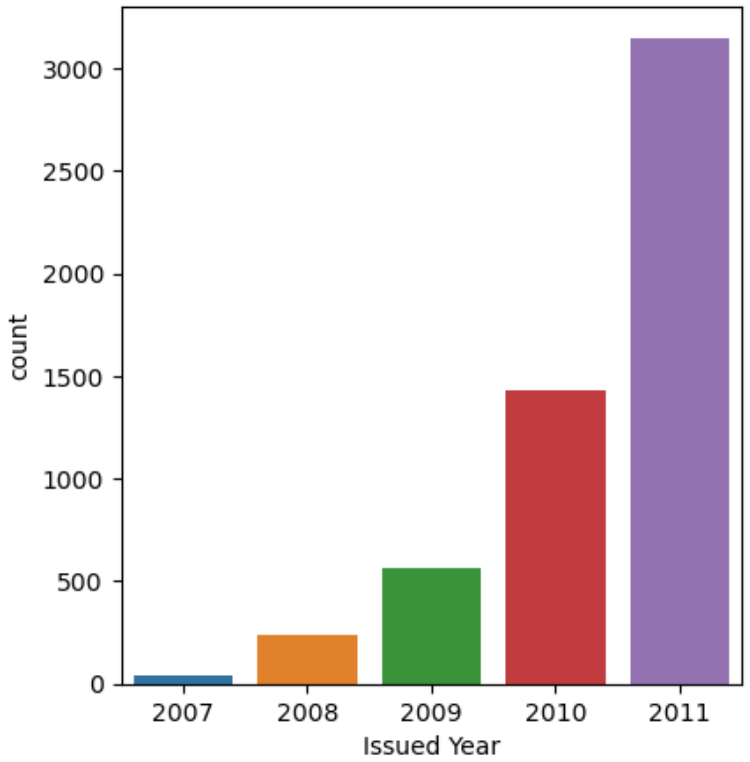
Uni-Variate analysis for open account bins, revolving util ratio bins, total account bins, annual income bins

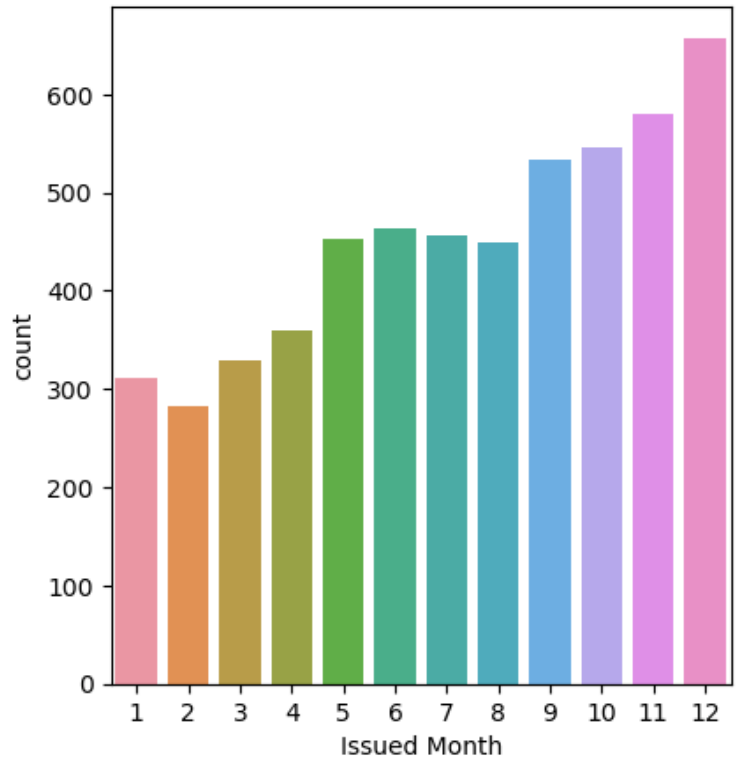


Uni-variate analysis for loan status, recoveries,inquiry in last 6 months, installment bins, loan amount bins, funded amount bins



Uni-variate analysis for issued month and year derived from issued date





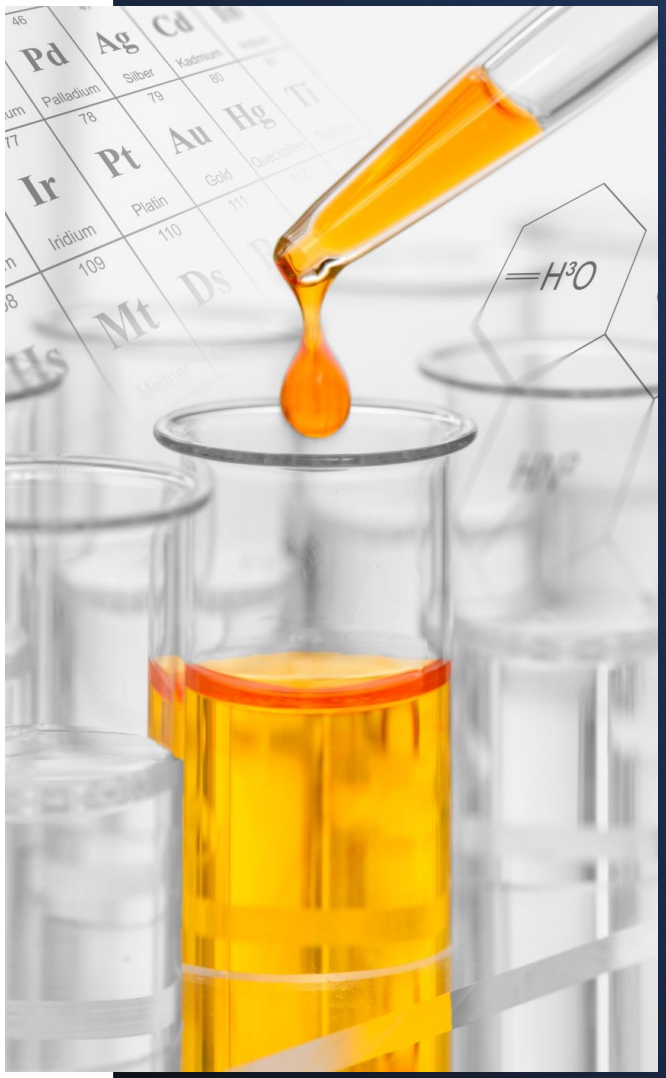
Variables in scope for Bi-variate Analysis

We are doing bi-variate analysis for Annual income and Loan amount with below listed variables

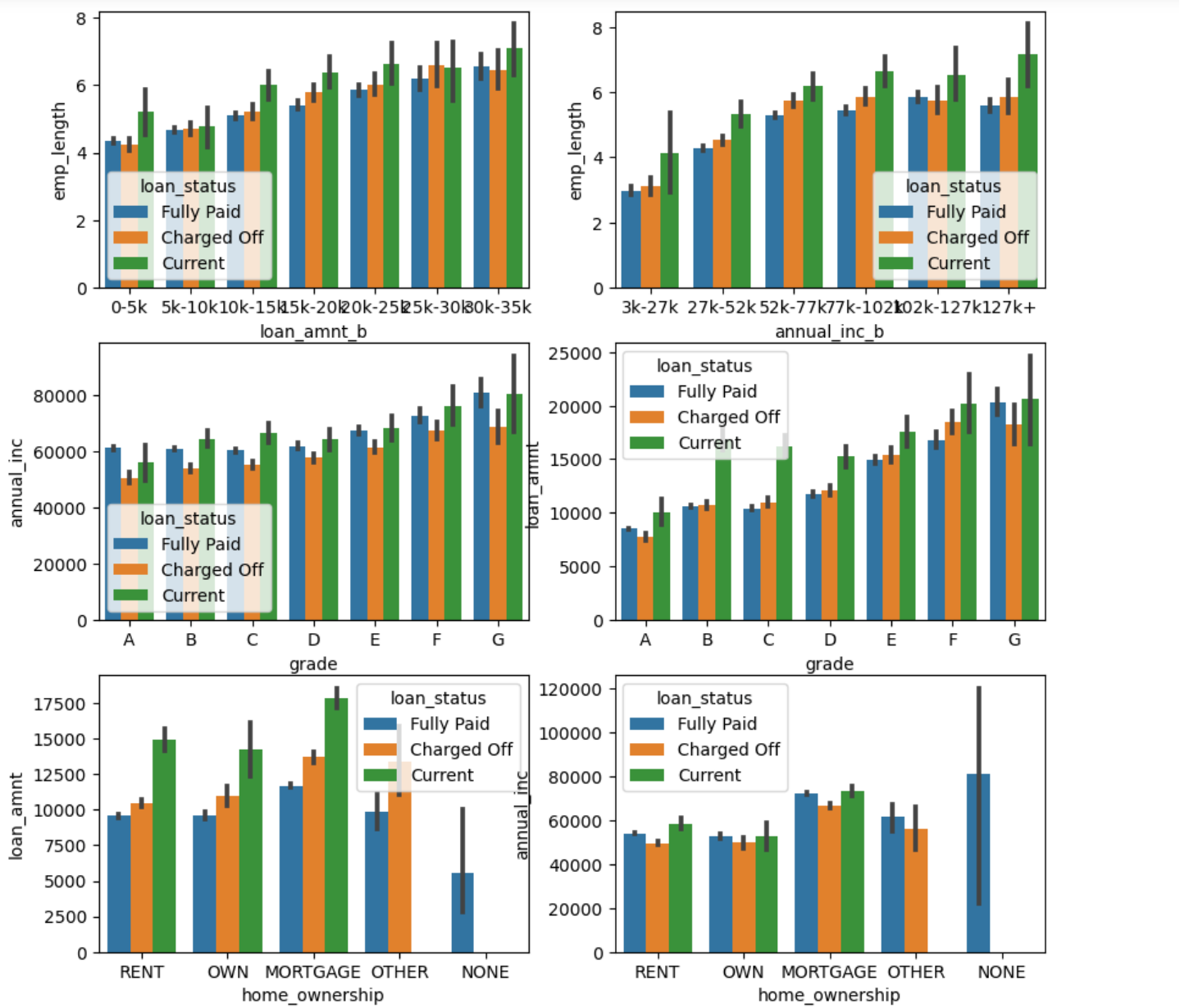
- Employee length
- Grade
- Home ownership
- Verification Status
- Purpose

And also for below pairs too

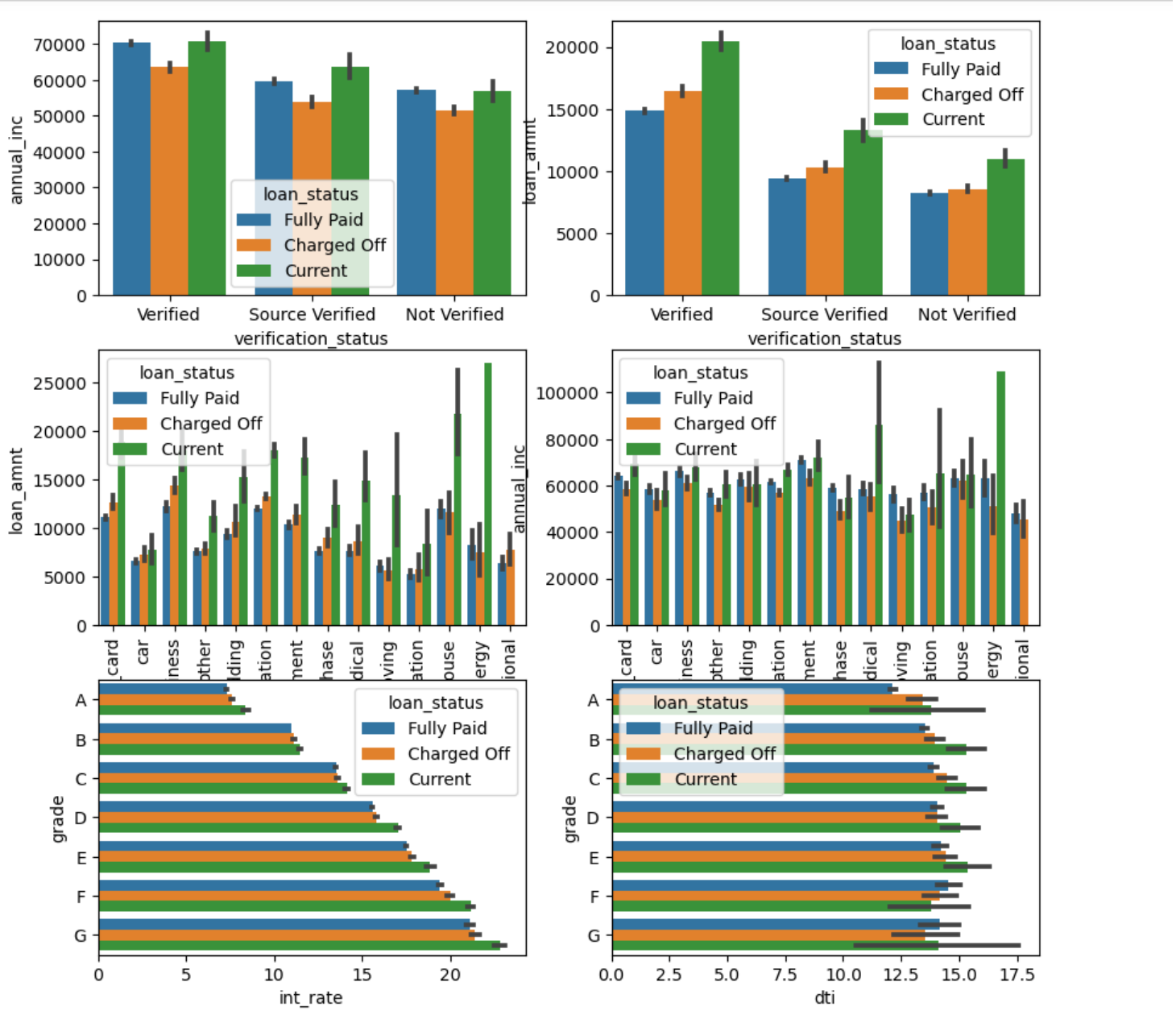
- Interest rate vs Grade
- Dti vs Grade



Bi Variate for emp_length, grade and home ownership

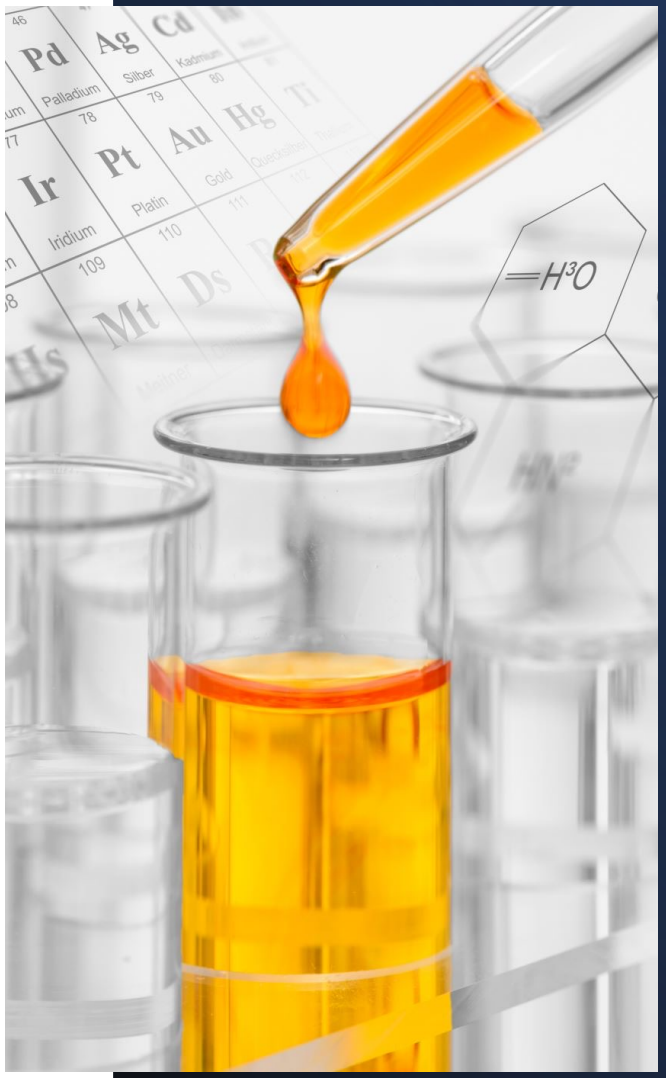


Bi Variate for verification_status, purpose and int_rate vs grade, grade vs dti

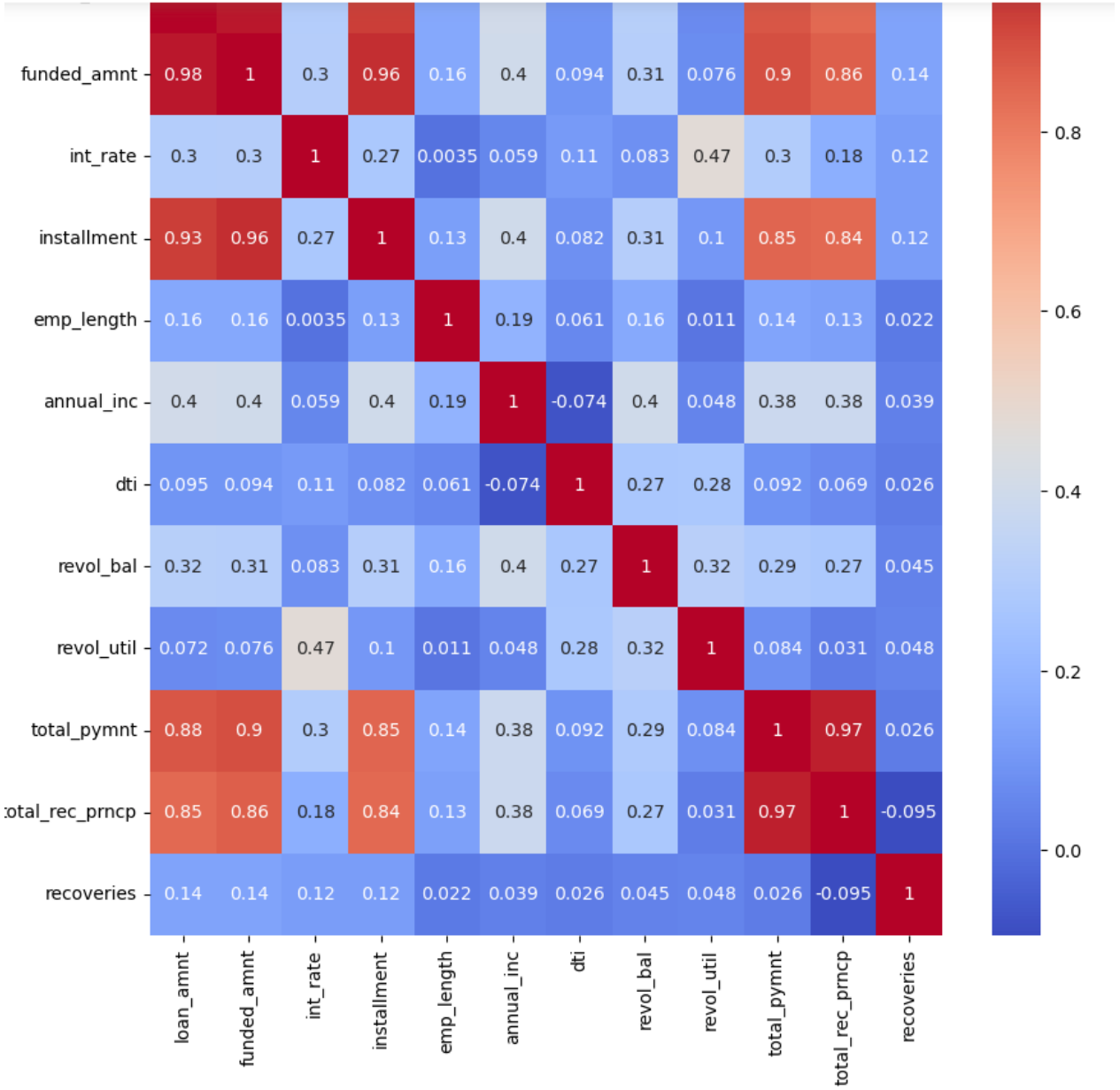


Variables in scope for correlation – Heat map analysis

- Loan Amount
- Funded Amount
- Interest Rate
- Installment
- Employment length
- Annual income
- Debt to Income Ratio(dti)
- Revolving Balance
- Revolving Utilization rate
- Total payment received
- Principal received to date
- Recoveries



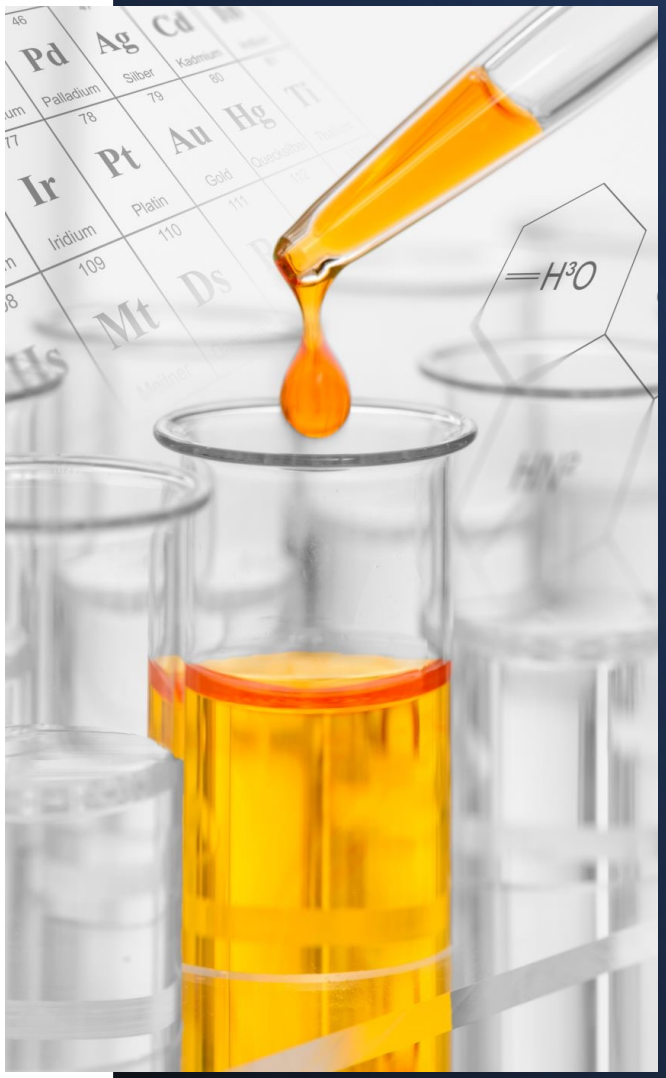
Correlation- Heat map



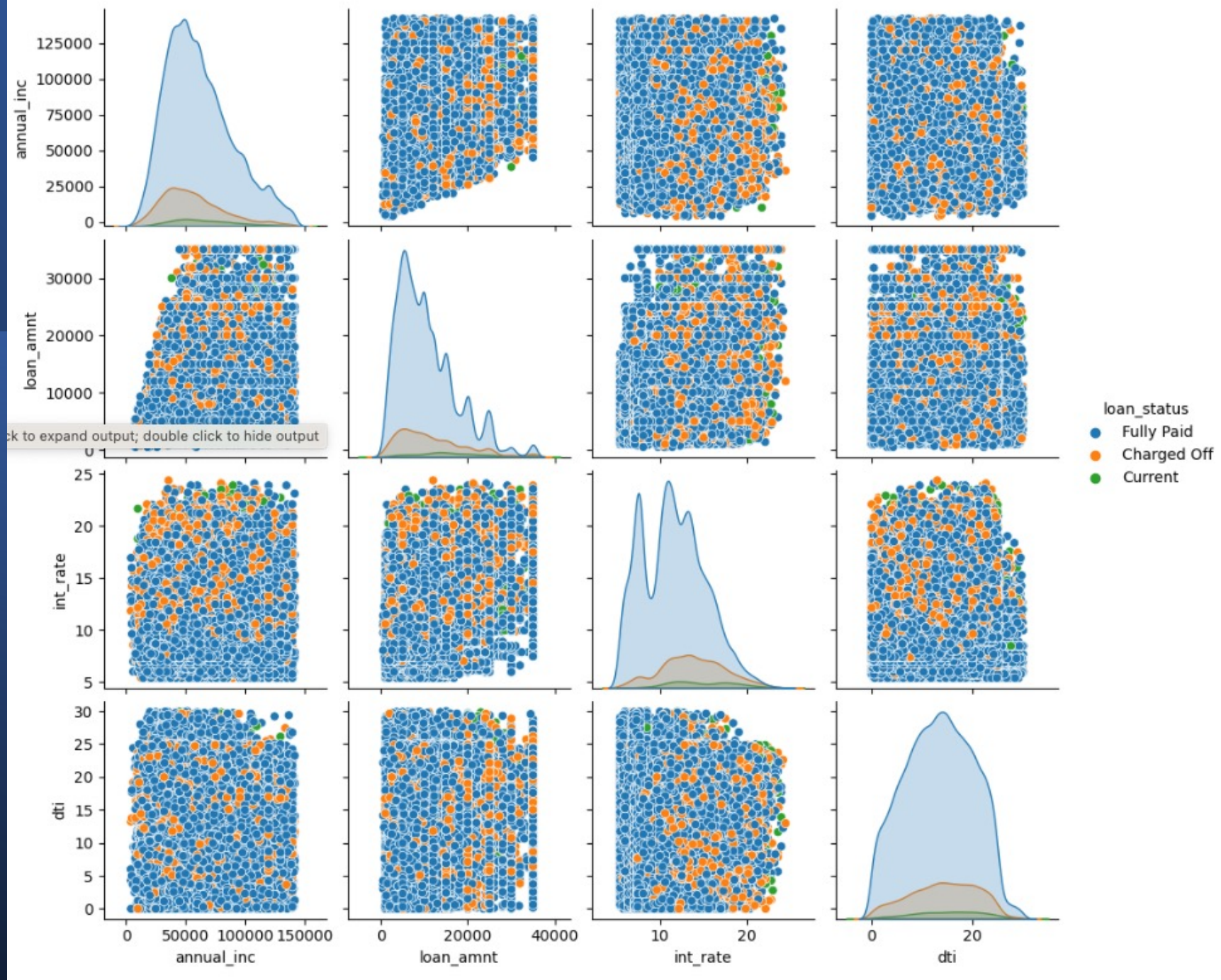
Variables in scope for multivariate analysis

We are doing multivariate analysis with hue as loan status

- Loan Amount
- Annual income
- Debt to Income Ratio(dti)
- Interest rate



MultiVariate Analysis



Conclusion

We now list out the final observations and also go through the driving variable & patterns

- Final Observations
- Driving Variables & Patterns

Final Observations

Uni-variate analysis (Segmented included)

Bi-Variate Analysis

Co-relation-Heat map analysis

Multivariate analysis

Uni-Variate Analysis

Uni-Variate Analysis: Key Insights

- Charged Off accounts represent approximately 1/7th of the combined total.
- Accounts with grades and sub-grades B(5,3,4), C(1,2,3) have a higher incidence of charge-offs.
- More charged-off accounts are associated with borrowers having B and C grades.
- Not total inference Emp_title "US Army" is linked to a higher occurrence of charge-offs.
- Borrowers with home ownership as "Rent" and "Mortgage" tend to default more frequently.
- "Debt Consolidation" purpose shows a higher rate of charge-off accounts.
- Zip_code is masked and encoded, making it irrelevant for analysis.
- Borrowers residing in California, Florida, New York, and New Jersey have a higher default rate.
- Employment length greater than 10 years is associated with a higher likelihood of default.
- Shorter loan terms (36 months) have a higher default rate compared to longer terms (60 months).
- Borrowers with interest rates between 13%-17% have a higher default rate.
- A DTI between 12-18 is linked to the highest occurrence of charge-offs.
- Income between 27K - 77K is associated with more default cases.

Uni- Variate Analysis contd..

- Total accounts between 2-35 and open accounts between 2-18 are more prone to default.
- Higher revolving utilization rates increase the likelihood of default.
- Recovery late fee between 0-30 is linked to the highest charge-off accounts.
- Accounts with recoveries exceeding 0 are more likely to default.
- Borrowers with 0 inquiries in the last 6 months tend to default more.
- Lower installments, particularly between 145-274, are associated with higher default rates.
- Loan amounts between 5k -10k have more charge-offs, followed by 0-5k.
- Funded amounts between 5k-10k have a higher default rate, followed by 0-5k.
- Accounts with no history of public bankruptcies tend to default more frequently than those with at least one bankruptcy record. This observation is supported by the fact that a higher proportion of accounts fall into the category of no bankruptcies as compared to those with one bankruptcy, and accounts with one bankruptcy exhibit a higher likelihood of default when considering the ratio of total accounts to defaulters
- "Not verified" accounts tend to default more.
- Accounts with 0 derogatory records tend to default more.
- Accounts given during the 2011 recession have a higher risk of default.
- Defaults are more common toward the end of the year, especially in December, November, and October.
- Recent credit inquiries after the last payment date increase the risk of default

Bi- Variate Analysis

1. Accounts with emp_length 6-7 and loan_amnt between 30k-35k have more defaulters.
2. Accounts with emp_length between 5-7 and annual_inc 127k+ have more defaulters.
3. accounts with G grade followed by F and annual_inc between 60k to 80k have more defaulters.
4. accounts with G grade and loan_amnt between 15K to 20k have more defaulters
5. Accounts with loan_amnt between 10k -17.5k and home ownership as mortgage has more defaulters
6. accounts with annual_inc between 40k-60k and home ownership as Mortgage has more defaulters
7. Accounts which are verified and annual_inc between 60k - 70k have more defaulters
8. Accounts which are verified and loan_amnt between 15k - 17.5k have more defaulters
9. accounts with purpose as small business, followed by debt consolidation and loan_amnt between 10k-15k have more defaulters
10. accounts with purpose as home improvement followed by debt consolidation,small business, wedding and annual_inc between 60k-65k have more defaulters
11. Accounts with grade G and int_rate more than 20 tend to have chances of defaulting
12. Accounts with grade C and dti as 12.5 have more defaulters

Multi- Variate and Correlation Analysis

- **Co-Relation Pairs**

1. loan_amnt & funded_amnt
2. loan_amnt & installment
3. funded_amnt & installment

- **Multi-Variate Analysis:**

1. As interest rate increases, we can see increase in Charged off accounts and also the loan amount
2. Those with low income and high interest rate tend to default more than others
3. As annual income increases, there's increase in loan amount too, not fully but there's a probable relation

Driving Variables & Patterns

- **Employment Length**: Accounts with employment lengths greater than 10 years are more prone to default than others.
- **Home Ownership**: Borrowers with home ownership as "Rent" and "Mortgage" are more likely to default compared to other categories.
- **Loan Purpose**: Loans with the purpose of "Debt Consolidation" tend to have higher charge-off accounts.
- **State**: Borrowers residing in states such as California, Florida, New York, and New Jersey tend to default more than those in other states.
- **Interest Rate**: Borrowers with interest rates between 13%-17% are more likely to default.
- **Debt-to-Income Ratio (DTI)**: Borrowers with DTI between 12-18 have a higher number of charged-off accounts.
- **Annual Income**: Borrowers with annual incomes between 27K - 77K tend to default more.
- **Total Accounts and Open Accounts**: Borrowers with total accounts between 2-35 and open accounts between 2-18 are more likely to default.
- **Revolving Utilization Rate**: Higher revolving utilization rates are associated with a higher likelihood of default.
- **Recovery Late Fee and Recovery Fee**: Accounts with recovery late fees and recovery fees in the range of 0-30 tend to have more charged-off accounts.
- **Grade**: Accounts with grade B and C tend to default more than others

Driving Variables & Patterns

- **Bankruptcies**: Accounts with 0 public bankruptcies tend to default more than others.
- **Verification Status**: Accounts marked as "Not Verified" tend to default more than others.
- **Derogatory Records**: Accounts with 0 derogatory records tend to default more than others.
- **Issue Date**: Accounts that defaulted more in 2011, coinciding with the recession, had a significant impact and also the same goes for loans issued towards the last months of the year.
- **Latest Credit Inquiries**: Borrowers with recent credit inquiries after their last payment date are at a higher risk of defaulting.
- **Correlation Analysis**: Variables like loan_amnt, funded_amnt, and installment have high positive correlations, indicating their influence on each other.
- **Annual Income and Interest Rate**: Borrowers with low income and high interest rates tend to default more.

Thanks

Please reach out to us in case of any questions

- **Srinivasa Reddy M**
  Email ID:
  srinivasareddymus@gmail.com
  Phone No:918186026029

- **Shalini Kushwaha**
  Email ID:
  shalini2021.delhi@gmail.com
  Phone no: 919821940292