Lending Loan Case Study

GROUP MEMBERS

ASTHA AGARWAL
VINAY PRASAD K

AGENDA

- Overview- Problem Statement
- Methodology
- Data Transformation – Cleaning and imputation
- Exploratory Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
- Summary

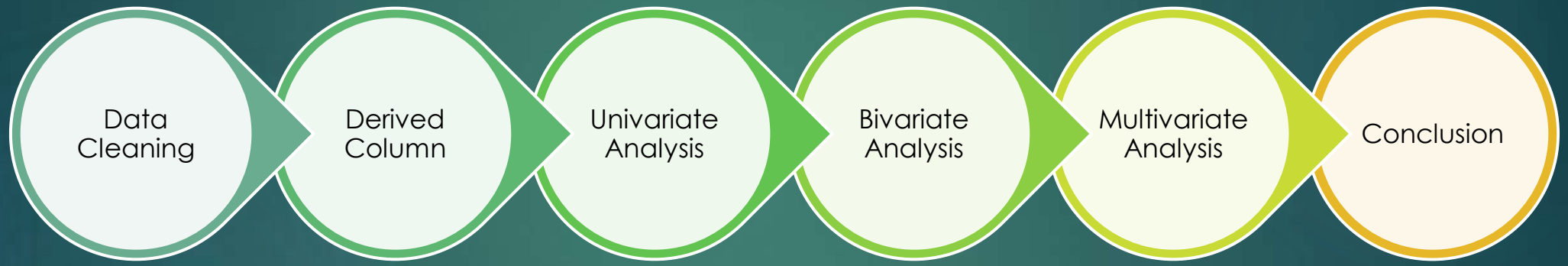
Overview

About Company : Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Problem Statement: Analyze data to identify

Driving factors behind loan default, i.e. the variables which are strong indicators of default.	Consumer attributes and loan attributes that influence the tendency of default.

Methodology



Data Cleaning and Data Transformation

Data Cleaning

1. Dropping fields with all NA values
2. Remove fields not adding valuable meaning to problem statement from domain standpoint
3. Remove post approval features (features which will not impact the defaulting nature of borrower)

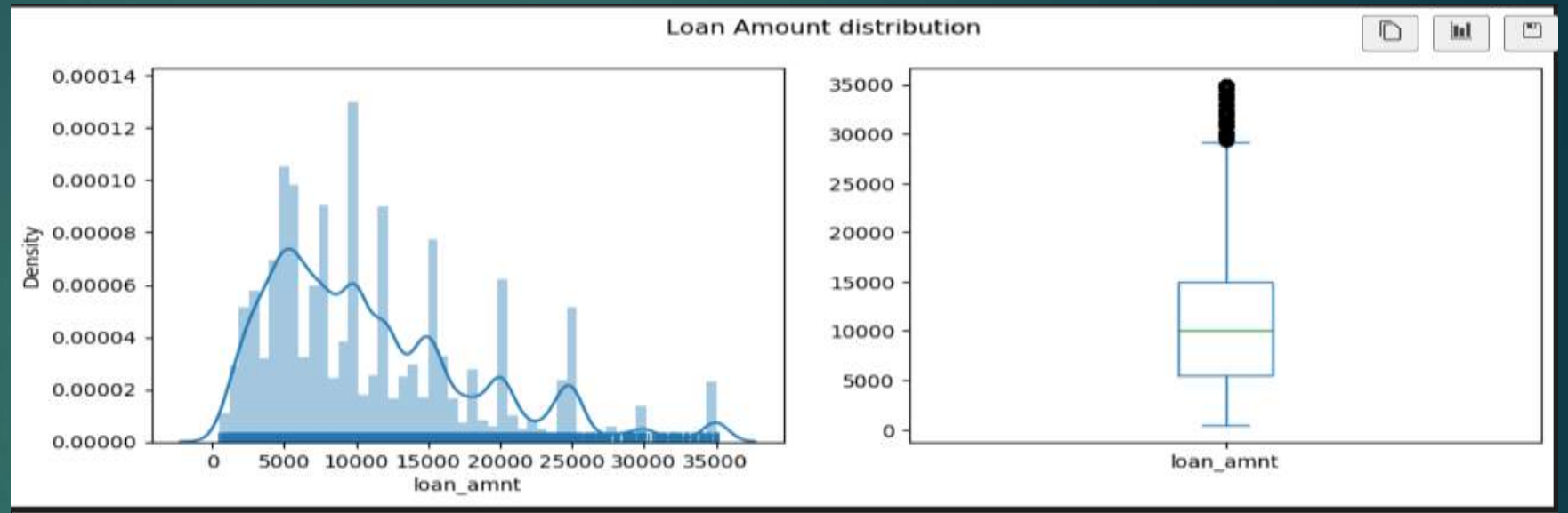
Data Transformations Techniques

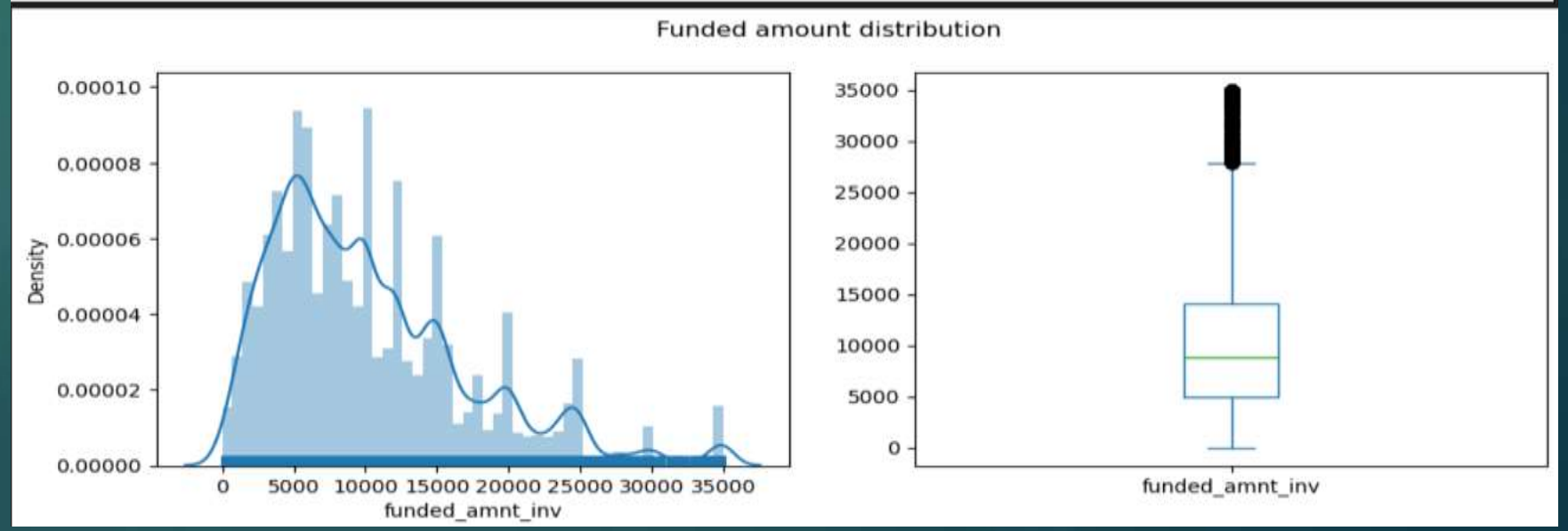
1. Analyze content of categorical columns and numerical columns and clean if needed
2. Covert values to proper int ,float and date representations
3. Check for outliers

Univariate Analysis (Numerical Features)

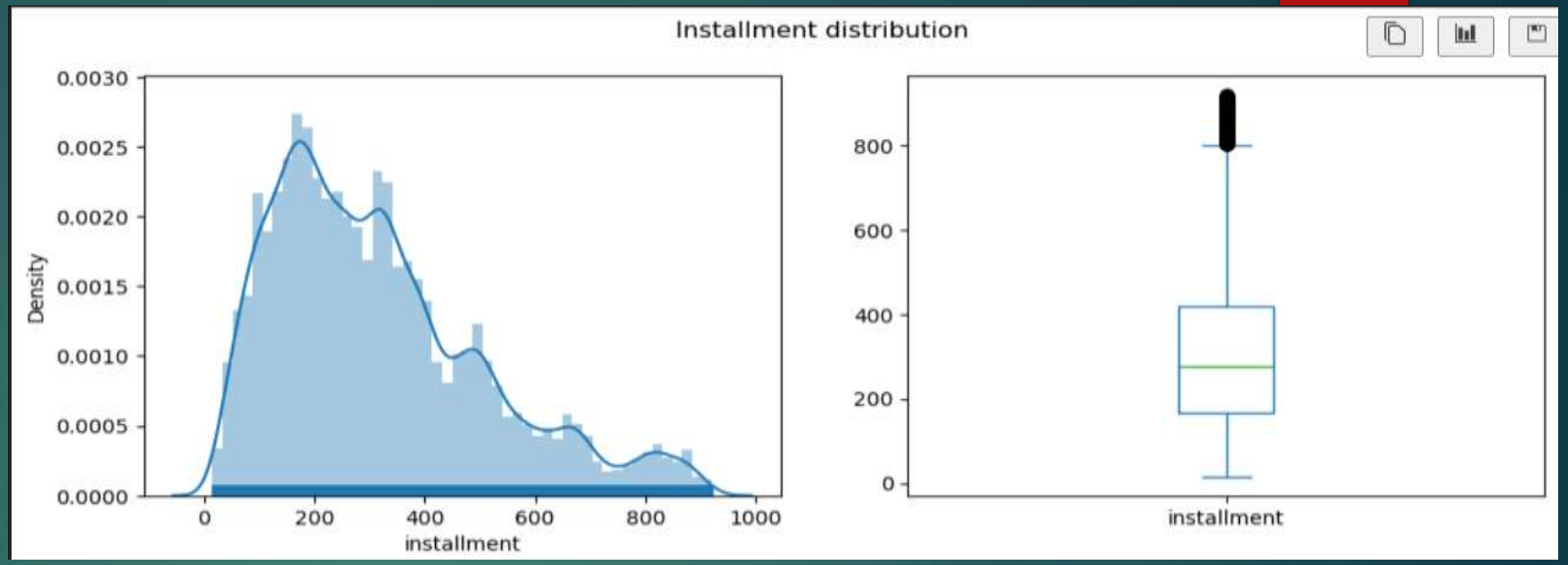
Key Observations:

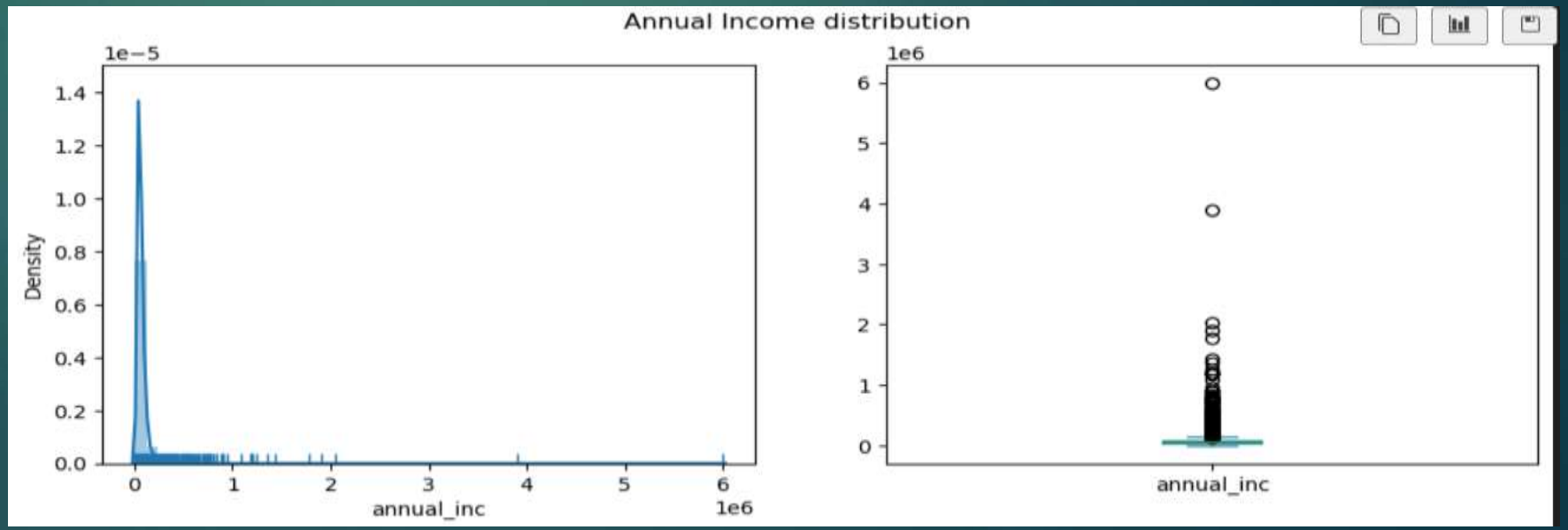
1. Most people have taken loan of around 10K as median and mean both are around same value. Maximum people have taken a loan of less than 25K as it is the 95th percentile of data distribution, after that only few people have taken loan in range of 25K to 35K





Univariate Analysis (Numerical Features)

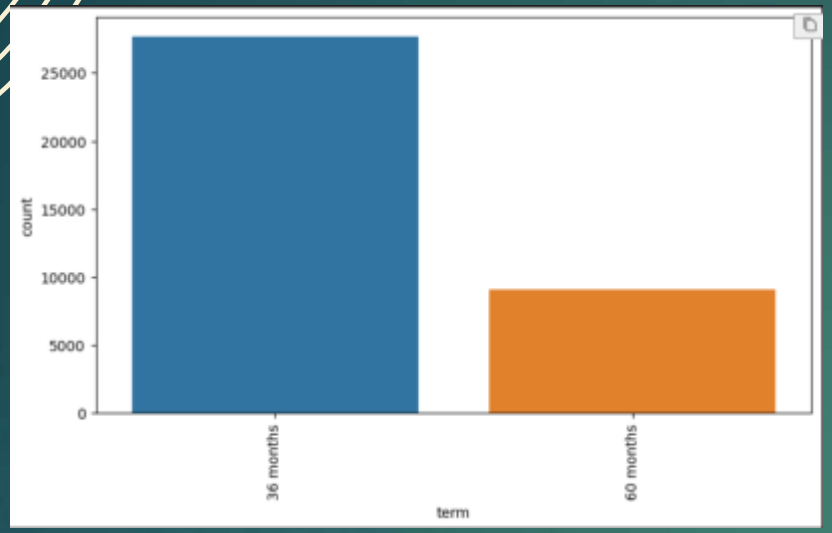




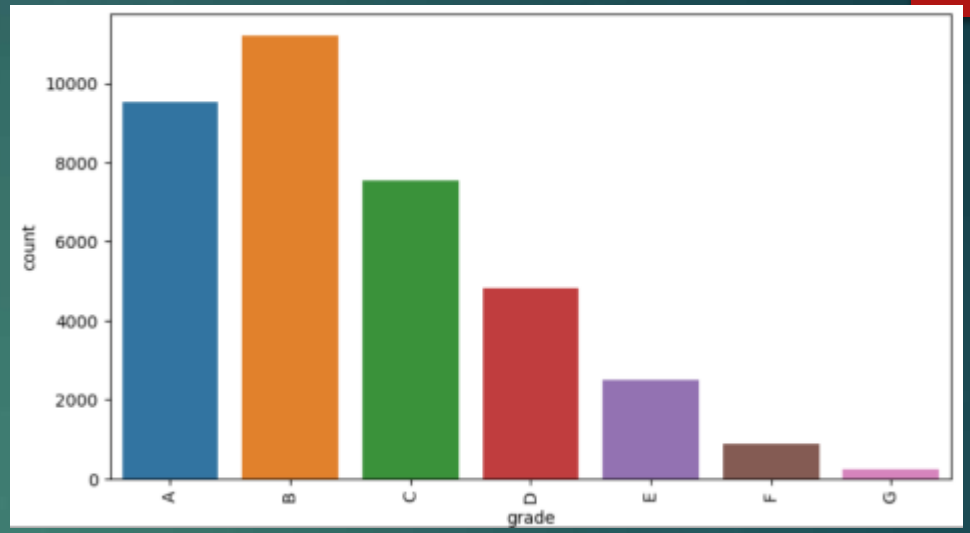
Key Observations:

1. **Installment** distribution is static after 921 of value, 99 percentile is at 921 and for 1 percentile people installment is more than that. As per analysis , this is for people who applied for a loan of more than 25000. We can drop the outliers for installment
2. There are a lot of outliers in **annual income** field, 99 percentile customers have annual income below 235 K and rest are outliers and we can skip from data

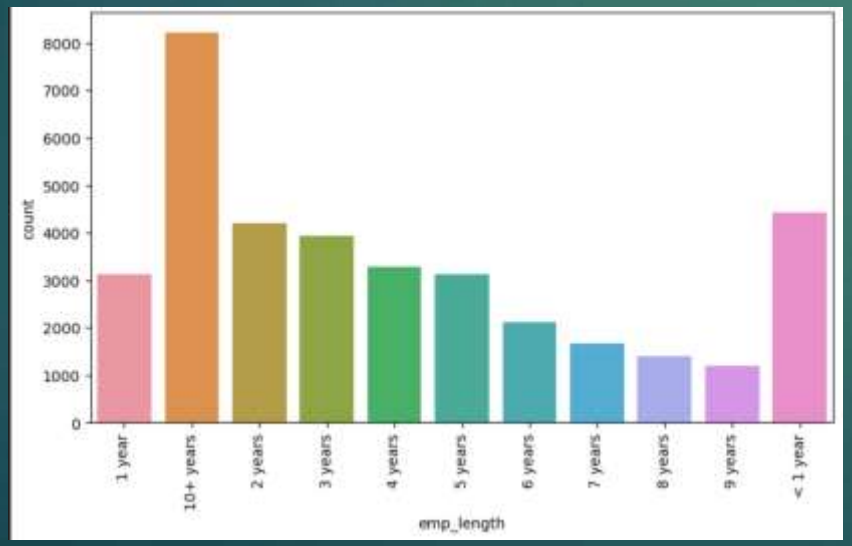
Univariate Analysis (Categorical Features)



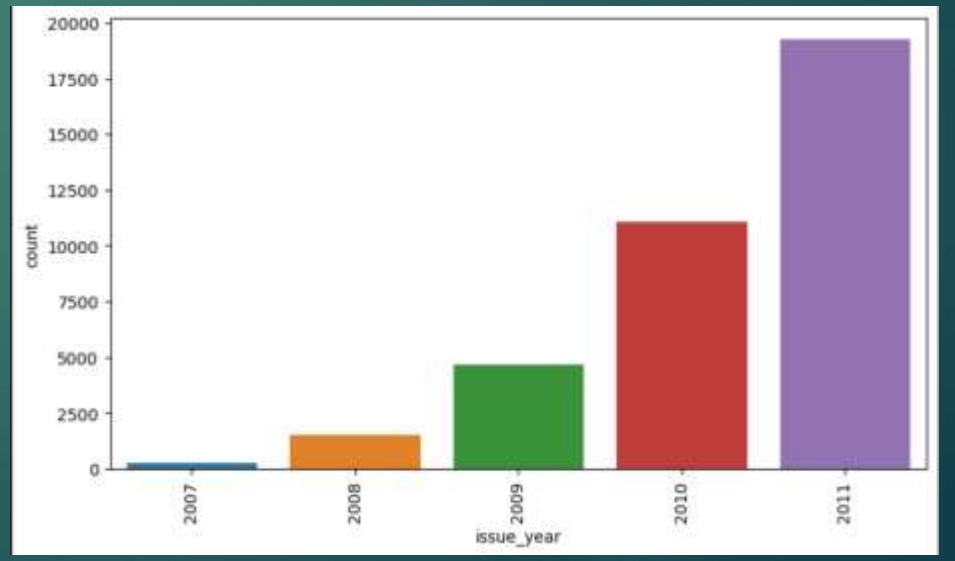
Distribution of applicants over term



Distribution of applicants over grade



Distribution of applicants over emp_length

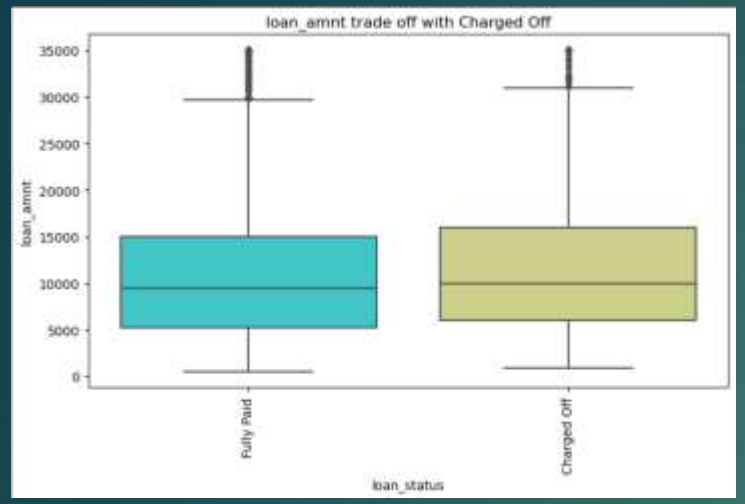


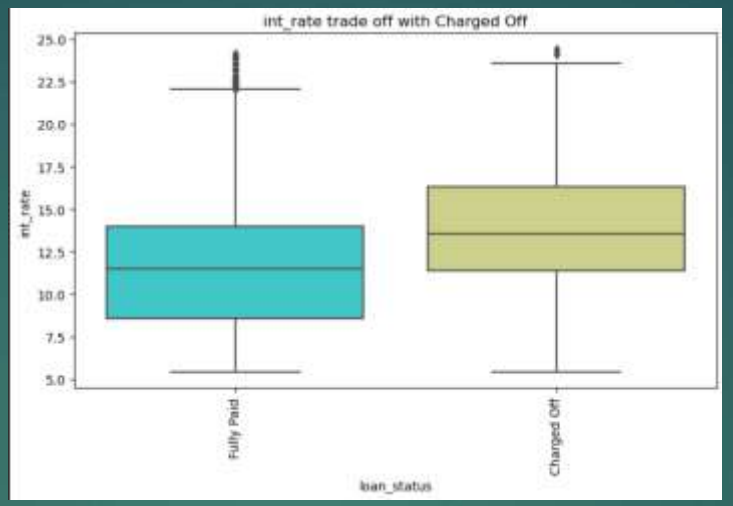
Distribution of applicants over loan issued year

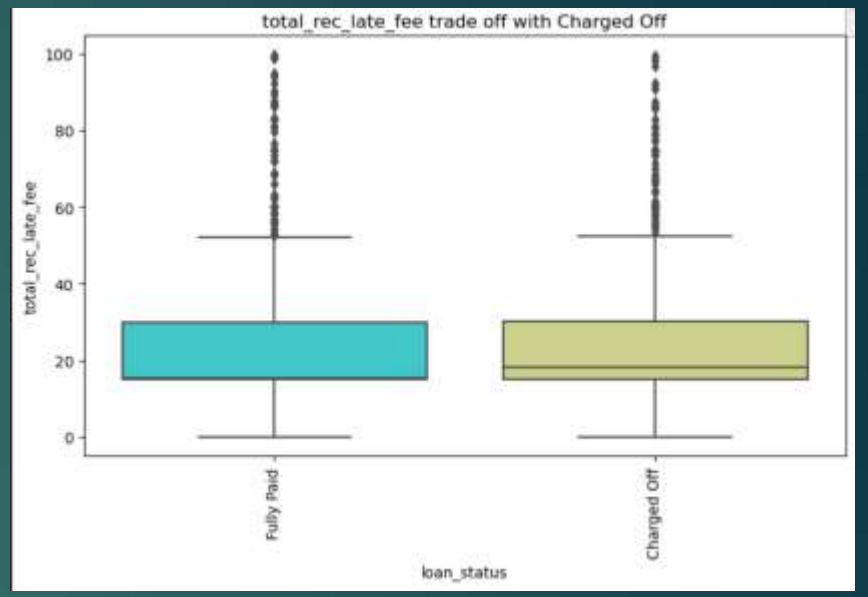
Key Observations

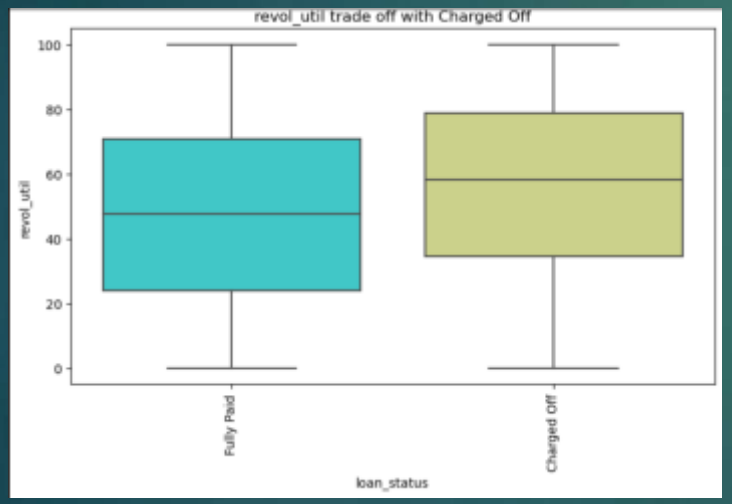
1. Maximum number of loans are taken by people who have been employed for more than 10 years
2. People who have either rented out property or have mortgaged their home have high chances of taking a loan
3. Most of the loans were issues in 2011,and number of loan applications are increasing every year
4. Most of the people generally take loans towards the end of the year

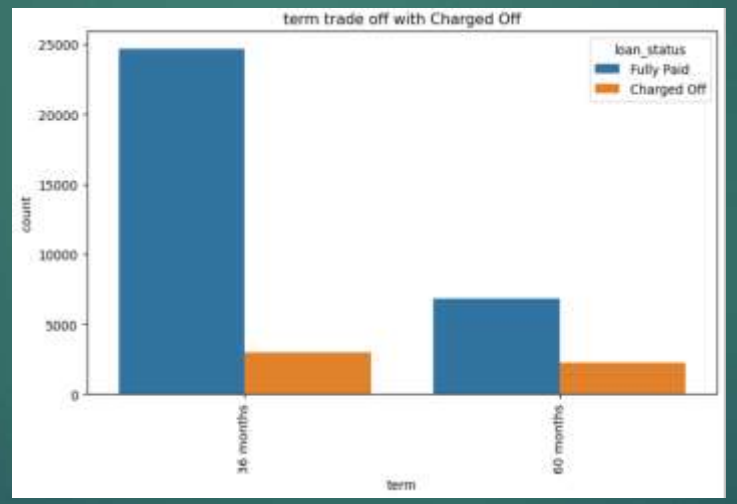
Bivariate Analysis





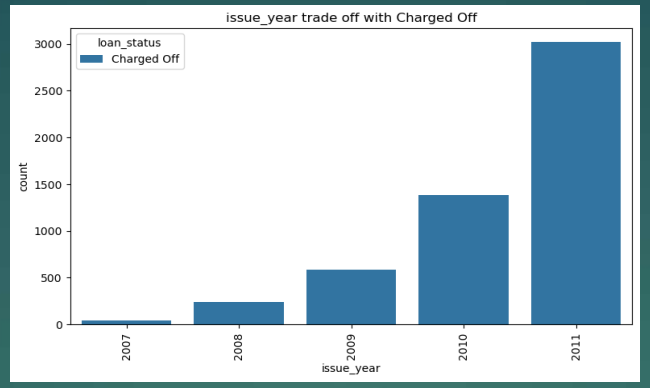


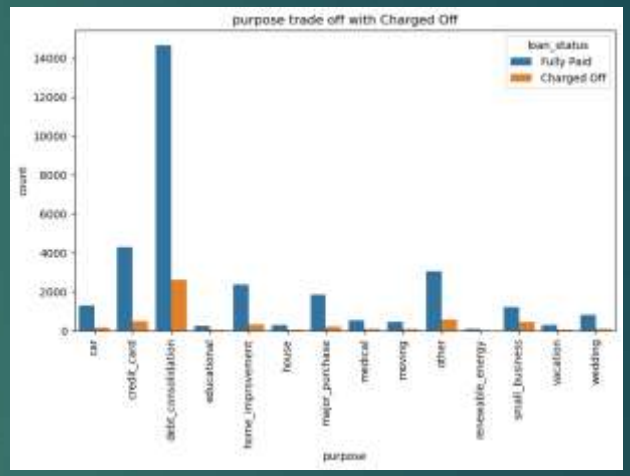


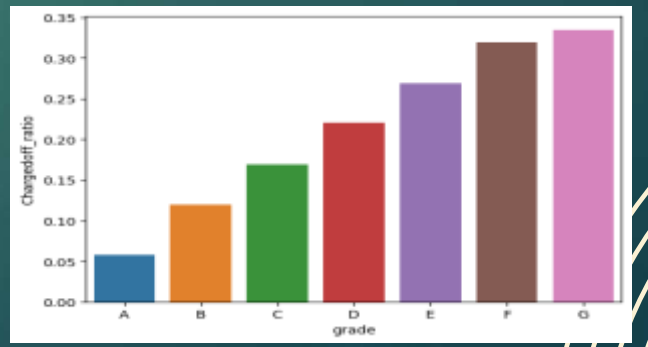


Key Observations:

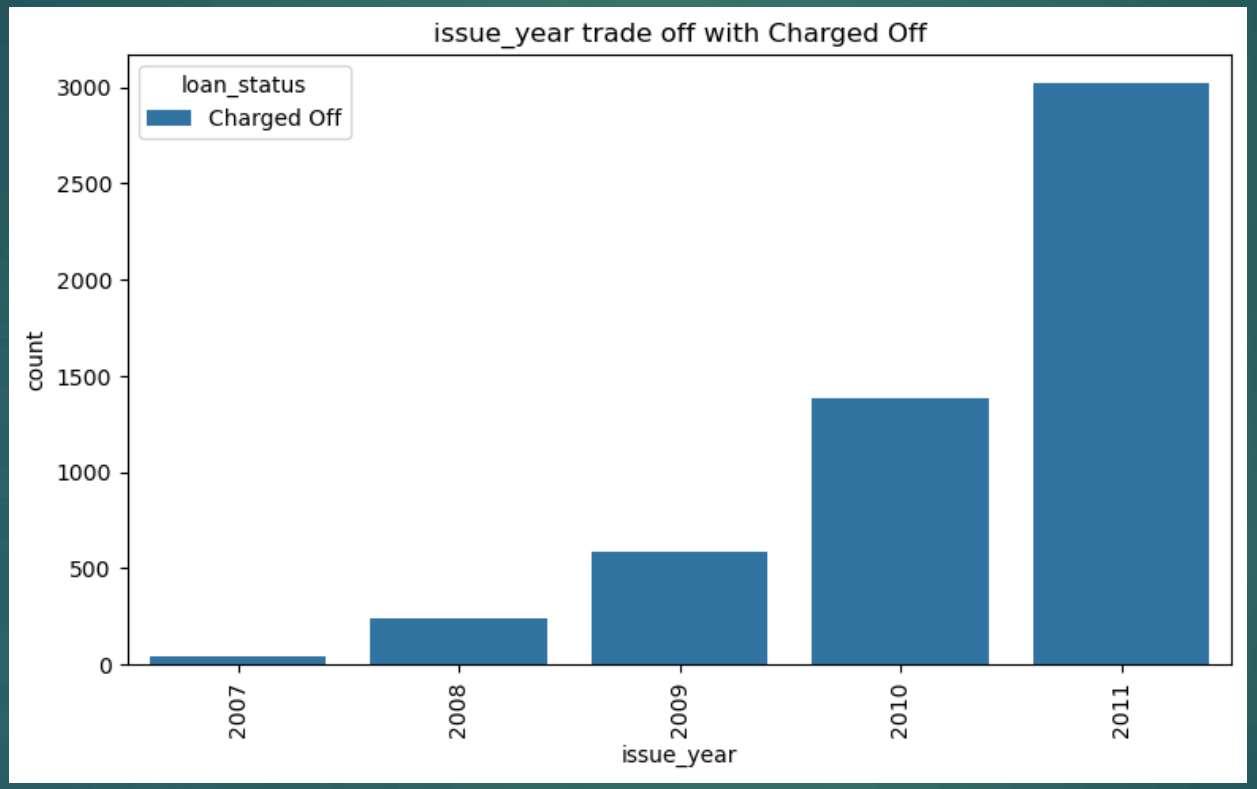
1. Charged Off loans have higher amounts than Fully Paid ones.
2. This Chart clearly portrays, higher the interest rate more are the chance of Defaulting the loan.
3. People paying more late fees are more likely to default
4. The amount of credit the borrower is using relative to all available revolving credit is higher for people marked Charged Off
5. Proportion of Charged Off to Total loan applications is more in case of 60 months term







Key Observations:
1.Number of applicants defaulting is increasing year over year linearly
2.Maximum cases of charged off are those for which applications are not verified, and it makes sense and thus it is important that verification should be done for applicants to reduce defaults
3.Small business applicants are more likely to go charged off, proportion for Charged off to total is high
4.Chances of charged off is increasing with grade moving from "A" towards "G"
(Segmentation Univariate Analysis)

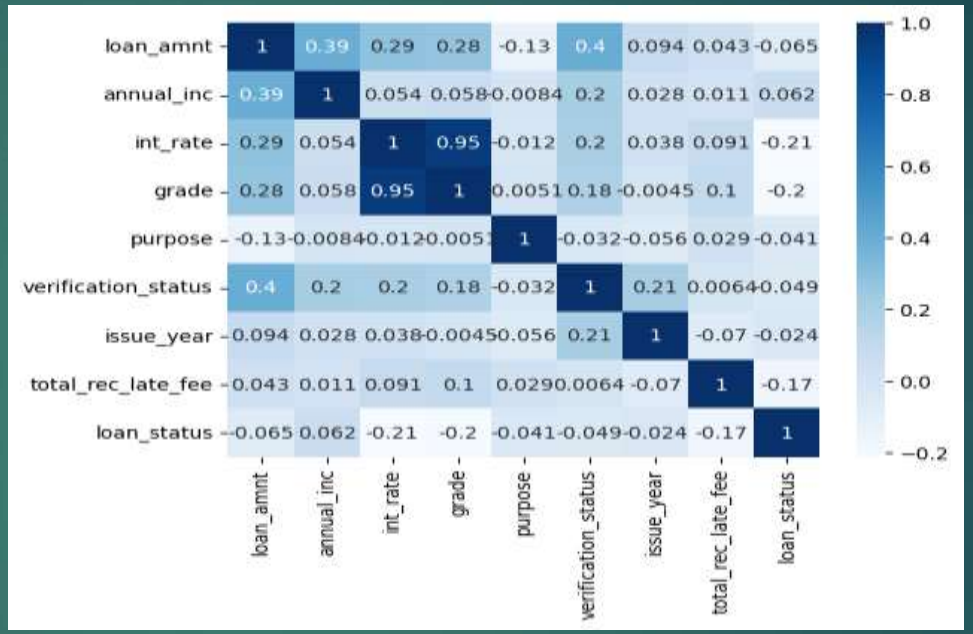


Observation: Number of applicants defaulting is increasing year over year linearly

Multivariate Analysis

Shortlisted fields:

1. Grade
2. loan_amount
3. int_rate
4. purpose
5. Year of issued application
6. total_rec_late_fee
7. verification_status
8. annual_inc

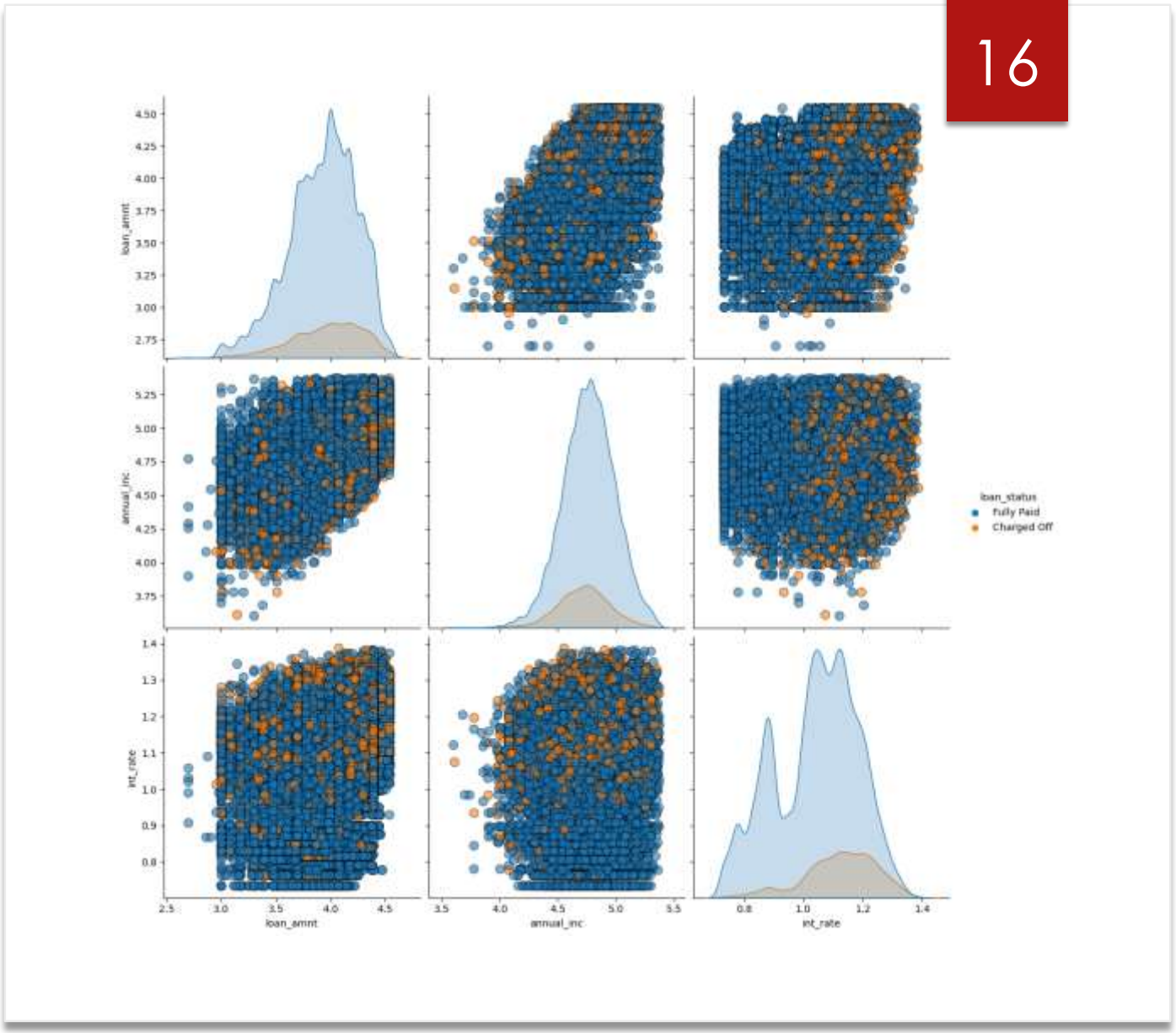


Key Observations:

1. There is high correlation between Grade and interest rate
2. There is negative correlation between all features and loan status field, i.e., loan status moved towards 0 state (Charged Off) for higher values of loan amount, annual inc, interest rate, purpose (small business), issue year, late fees, which we have already observed in previous analytical charts

Pair plot for multivariate

- Observation: Same trends can be observed with pair plots



Summary

- ✓ Higher the interest rate higher charged off ratio
- ✓ Higher the annual income higher the loan amount slightly.
- ✓ Interest rate is increasing with loan amount increase
- ✓ Number of people defaulting is increasing every year
- ✓ People paying more late fees are more likely to default., i.e., if they are regulating delaying monthly payments for loan, more likely they are undergoing monetary crunch and are likely to default
- ✓ People taking loan for small business are more likely to default
- ✓ People whose verification is pending are more likely to default

THANK YOU