# Lending Club Case Study

ML C63

Tejas Kapasi

# Content Topics

1. Problem Statement
2. Assumptions
3. EDA Analysis Approach
4. Observations and Driving Factors Highlight
5. Summarization by visual Plots
6. Conclusion
7. Acknowledgement and Thanks
8. Github Link

# Problem Statement



Identifying and mitigating the risk of loan default by analyzing key predictive indicators to reduce credit loss in reputed consumer finance online loan marketplace company.

Identifying key Consumer and Loan attributes which can impact the company to lend a loan to consumers which are risky.

# Assumptions

1. Data given is reliable and true sanity has been done before, DTI and Revolution Util ratio is made available accurately.
2. AIM for the study is to provide the key driving factors only not the rules that needs to be applied, so, Objective is not to provide rightful actions in this case study.
3. It is assumed that this same approach can be extended for the larger data set so, Volume given vs volume future can be huge. Notebook is expected to handle huge volume.
4. External libraries are permissible to use
5. Amounts are in USD

# Analysis Approach

1. **Data Collection:** Gather historical loan performance data, including borrower demographics, credit scores, loan amounts, interest rates, and repayment histories.Consumer attributes and Loan attributes.

2. **Data Cleaning and handling:** Preprocess the data to handle missing values, outliers, and inconsistencies to be corrected, Standardisation, Filtering the data

3. **Exploratory Data Analysis (EDA):** Conduct a thorough EDA to understand patterns and trends, using statistical summaries and visualization techniques. Univariate, Bivariate , Multivariate Analysis

4. **Data Visualisation:** Identify and select the most relevant features that contribute to loan defaults using correlation analysis and feature importance methods and display them in the plots.

5. **Insights and Recommendations:** Interpret the results to identify key risk indicators and provide driving factors to default.

# Data Collection and Loading

- “Loan” data is provided with Consumer Attributes and Loan Attributes.
- Data to be loaded into notebook with the use of libraries.
- Libraries are used for reading data is Pandas
- Shape of the data is 39777 raws by 111 Columns.

# Data Cleaning

- Column
  - Remove entirely Empty Columns - ***Total 57 columns found***
  - Remove not required columns, Read Data Dictionary for that.
  - Remove columns that has single value only
  - Split the columns to make it more meaningful (Loan Issue Date to Month and Year)
- Rows
  - Remove Summary raws if any
  - Remove Current Loan Data
  - Remove >50% columns missing on the raw

# Handling Missing Values and Standardising

- Find Missing % in the columns and treat them to either impute them or drop those, If > 60% Removed.
- Impute the columns -
  - Emp_Length with **'Self Employed'** where Empty found.
  - Revol_util is imputed with **mean()**
- Filter rows which are not required in the analysis (**Loan Status = Current**)
- Checking some inconsistencies in funded_amount, loan_amount, funded_amt_Investor.
- Standardising Values - Changing and correcting data type (Examples below)
  - Removing **%** from Interest Rate
  - Creating numeric from term (Object)
  - Object to Int wherever its required
  - Funded_amount to int
  - Float to Int wherever required

# Handling Outliers and Derived Metrics

- **Outliers**
  - Annual Income - Removed outliers ranging <0.1 and >0.99 quantile.
- **Derived Metrics** ( Additional Variables to Analyse)
  - Loan Amount to Income Ratio
  - Loan Status Indicator ( Numeric Variable)
  - Merging Verification Status
- **Segmented variables**
  - Interest Rate Groups
  - Annual Income Groups

# Attributes Categorization



## Categorical Features (Dimensions)

1. **Consumer Attributes**
    a. Grade
    b. Sub_grade
    c. Emp_length
    d. Home_ownership
    e. Purpose
    f. Addr_state
2. **Loan Attributes**
    a. Verification_status
    b. Issue_d
    c. Loan_status
    d. *Month
    e. *Annual_inc_group
    f. *int_rate_group

## Quantitative / Continuous Features (Facts)

1. Loan_amnt
2. Funded_amnt
3. Funded_amnt_inv
4. Term
5. Int_rate
6. Installment
7. Annual_inc
8. Dti
9. Revol_util
10. Total_pymnt
11. Total_pymnt_inv
12. Total_rec_prncp
13. Last_pymnt_amnt
14. *Amnt_to_inc_ratio
15. *Loan_status_indicator
16. *year

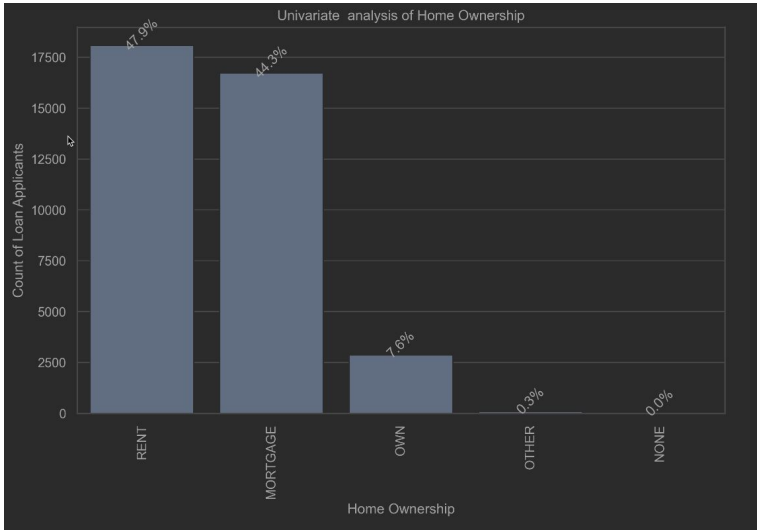
** Derived metrics created for analysis*

# EDA - Univariate Analysis

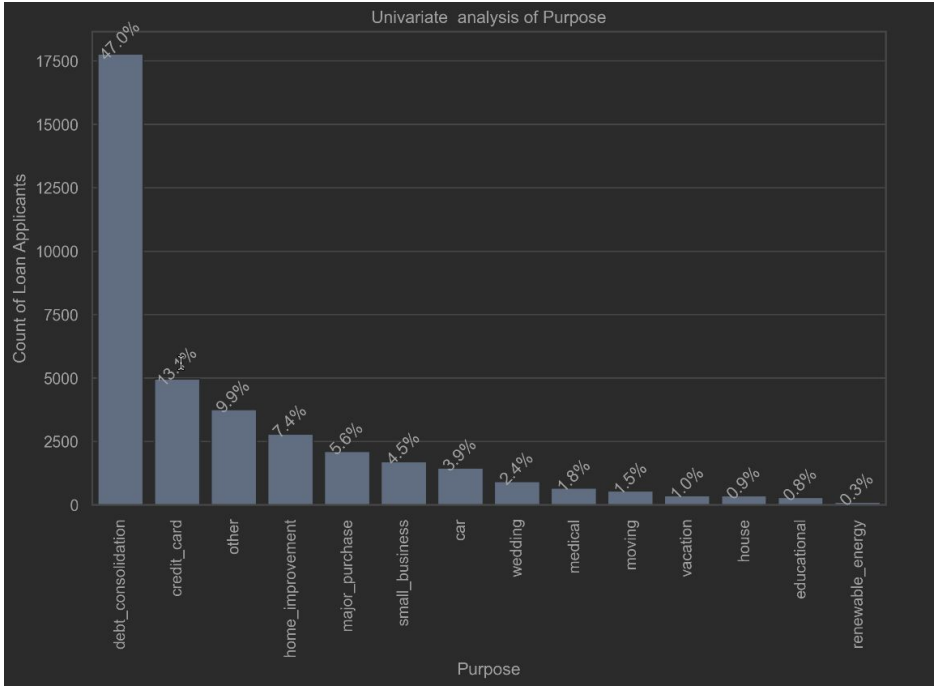


- Ordered Categorical Features Analysis
  - 53% loan issued in the Year 2011 -
  - Maximum loan are issued in end of the year especially in the month of December, November and October.
  - **Employment length** shows, people with 10+ years has taken **22%** loan and loan is issued them is highest in the number.
  - We can see maximum loan is issued for **36 months** compare to 60 months term.
  - **Grade B** has given highest loan then A and then C ( 30%, 26% and 20%)
  - Also, We can see **A4, B3, A5** are top subgrade who got highest number of times loan.
- Un-Ordered Categorical Feature Analysis
  - **57%** of Loans are verified and 43% are not verified.
  - State **CA** = California has the highest number of loan given. **NY and FL** followed second and 3rd.
  - **14.5%** loans are charged off, Defaulted.
  - **Debt_Consolidation** has been the top purpose of loans that are issued. Then "Credit card" and for "other" purpose.
  - **Rented and Mortgage** is the top category of people who has availed maximum loans. (48% and 44% respectively)
- Quantitative Feature analysis
  - Loan Amount to Income Ratio is Concentrated on lower side, and Mean is **18%** which is suggesting overall good nature.
  - Installment Amount Avg Concentrated around **200 to 400** USD
  - Loan Amount and Funded Amount is almost same, and Loan Amount range is **5000 to 15000 USD.**
  - **53% loans** are issued in the year 2011 whereas only 0.6% in the year 2007. 30% in 2010.
  - Year end months are commonly highest % of loans are issued. 11% , 10.4%, 9.7% are for Dec, Nov, Oct respectively.
- Segmented Univariate analysis
  - Income ranging between **50K-100K USD** are given highest loan
  - **10-15% Interest rate** is there for maximum loans, lowest loans issued on **20-25%**

# Univariate Plots

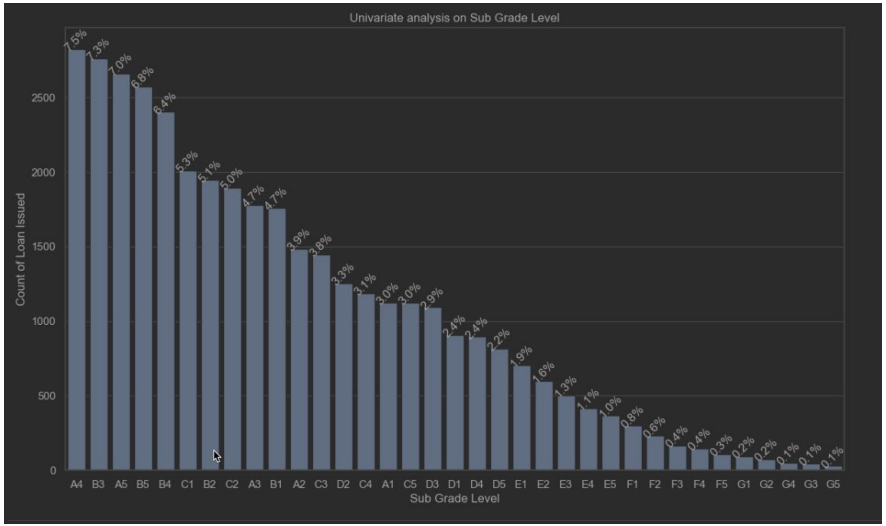


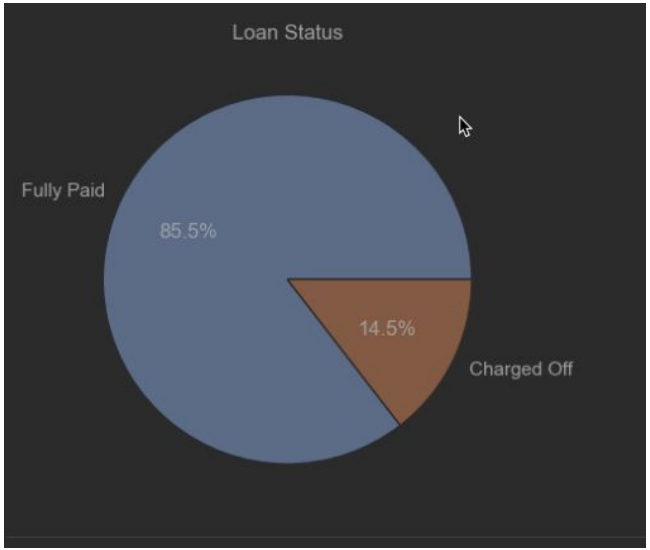

We can see Rent and Mortgage are the primary values who are availing more loans.




We can see The Majority of the purpose to take the loans are debt_consolidations and to pay Credit_card bill then for other purposes.

# Univariate Plots



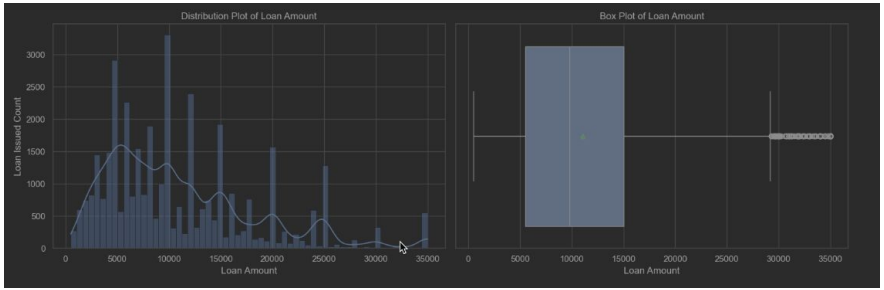



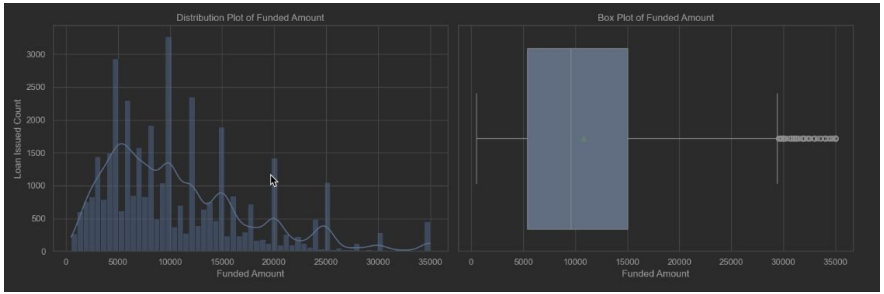

A4,B3, A5, B5 are top subgrade which received more loans. Lending club issued more loans to these sub grades.

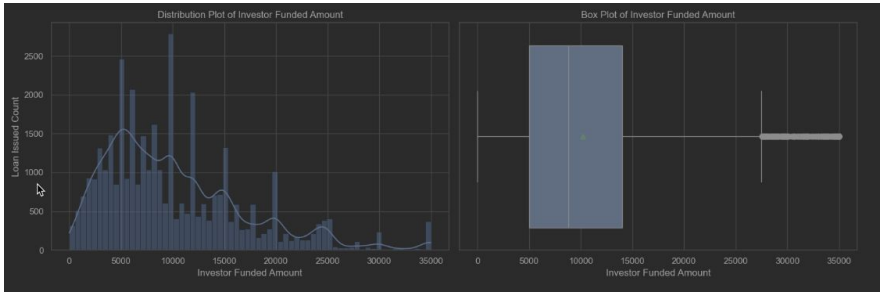
14.5% loans are charged off, Defaulted.

# Univariate Plots





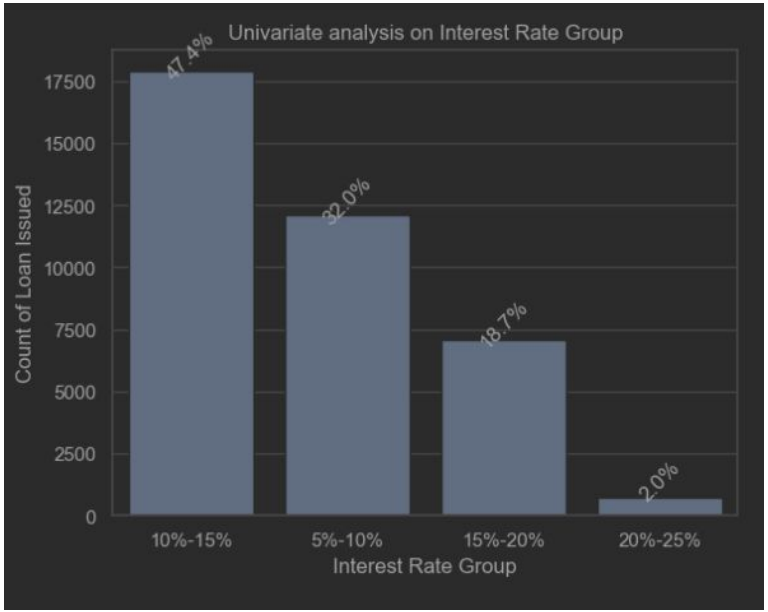



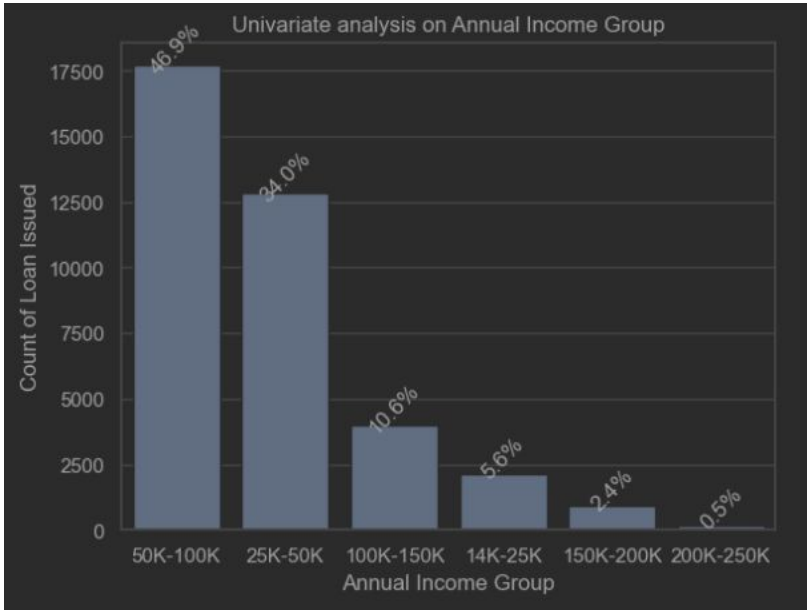

## Observation:

Loan amount, Funded amount and Investor funded amount concentrated on center equally.

Funded and Investor funded amount are almost granted as borrower's requested loan amount.

# Univariate Plots




10-15% rate of interest is the highest number of times loans are issued.

Income ranging between 50K-100K USD are given highest loan

# EDA - Bivariate Analysis Observations



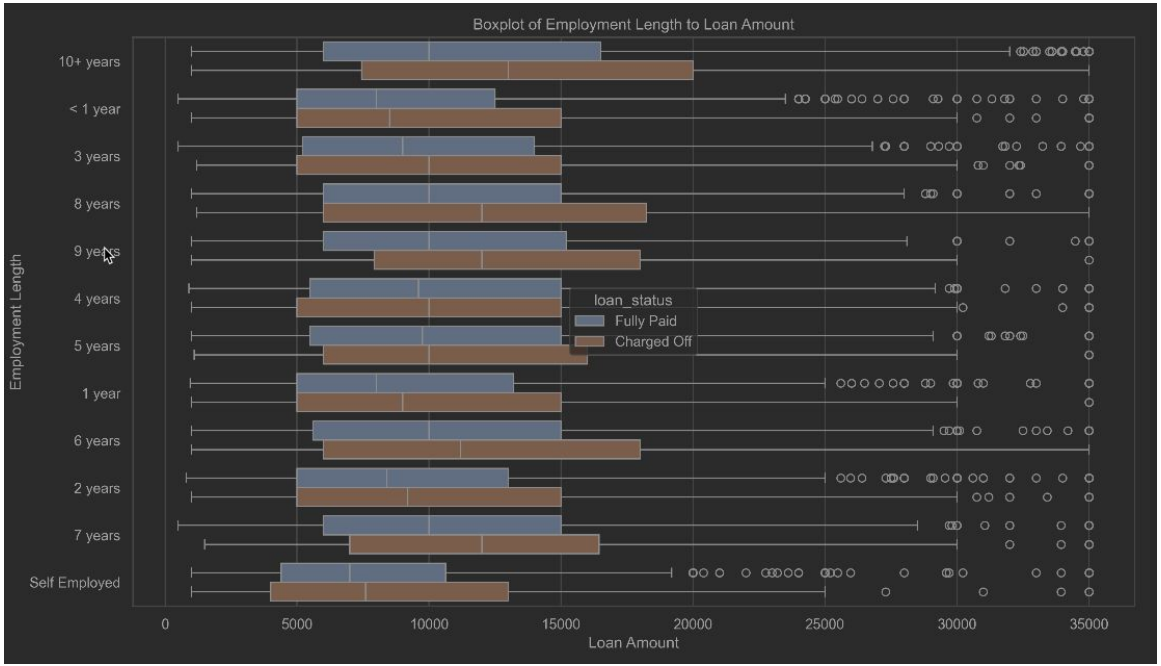
- To measure Loan Status along with the other categorical and numerical variables.
- Heat Map to check correlation for numeric features
- Box Plot with hue of “Loan Status” to see combination of Categorical and Numerical Features
- Count Plot with hue of “Loan Status” to see combination more clearly.

***Observations:***

- Loan Amount and Interest Rate Group can’t be compare with respect to Loan Status as there’s no big difference.
- Loan Amount, Fund Amount and Investor Fund Amount are positively correlate by 1
- Total Received Principle and Total Payment received correlating with Loan Amount by 0.84 which is positive correlate
- Employment **length 10+ years** got **more loan amount** and got high number of times default
- Employment **years 7,8,9 years median** are almost same which is **second highest** default
- UT state is having **Q3 at 25K** loan Amount. **75% UT defaulted at 25K** , then AK at 23K, then WY
- On Average, Many states got **median** of default **near to 10K**

# Bivariate Analysis Plots





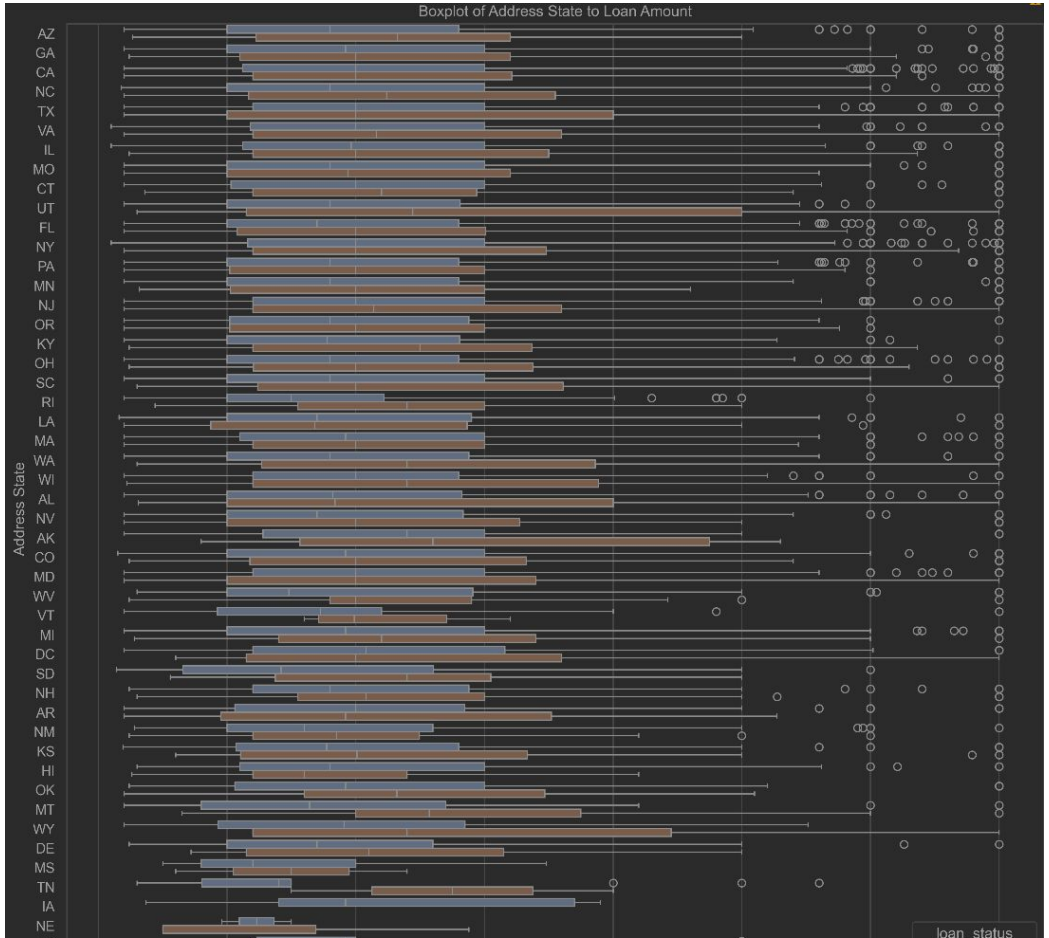

**Observation:**

Employment length 10+ years got more loan amount and got highest number of times defaulted

7,8,9 years median are almost same which is second highest defaulted

# Bivariate Analysis Plots






**Observation:**

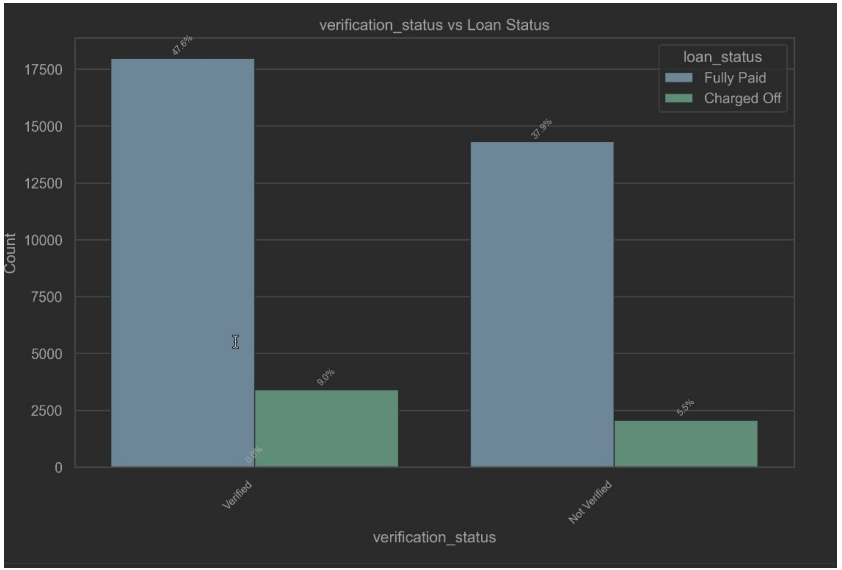
UT state is having Q3 at 25K loan Amount. 75% UT defaulted at 25K

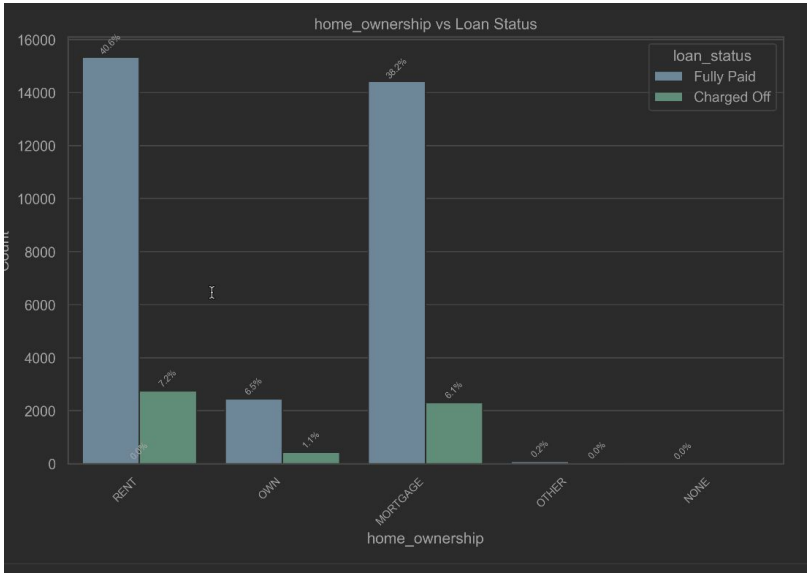
AK at 23K then WY is highest defaulted

On Average, Many states got median of default near to 10K where it gets defaulted.

# Bivariate Analysis Plots





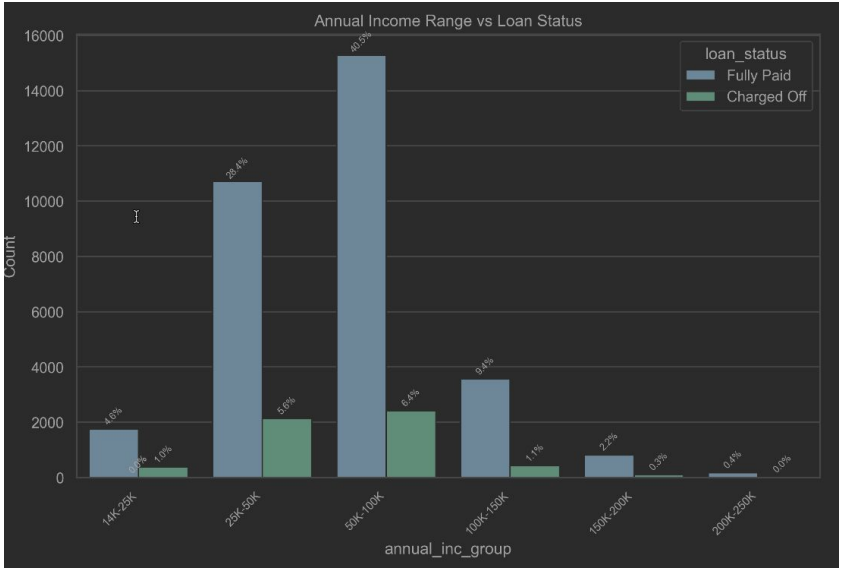



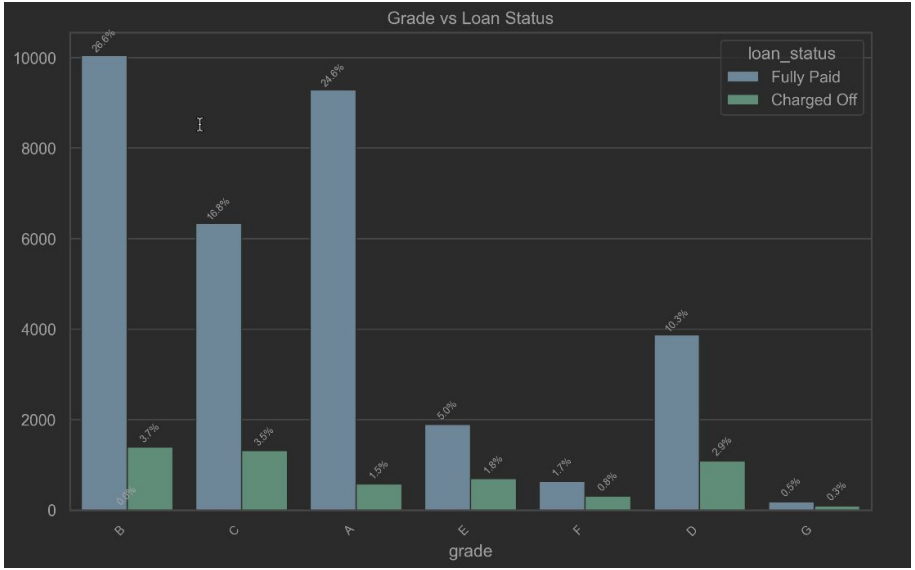

9% Defaulted loan were verified.

7.2% loans are charged off when Home Ownership is Rent and then 6.1% when Mortgage.

# Bivariate Analysis Port






Income group ranging from 50K-100K and 25K-50K are charged off highest no of times. 6.4% and 5.6% respectively.

B, C and D grade has the maximum charged off loans.

# EDA - Bivariate Analysis Summary

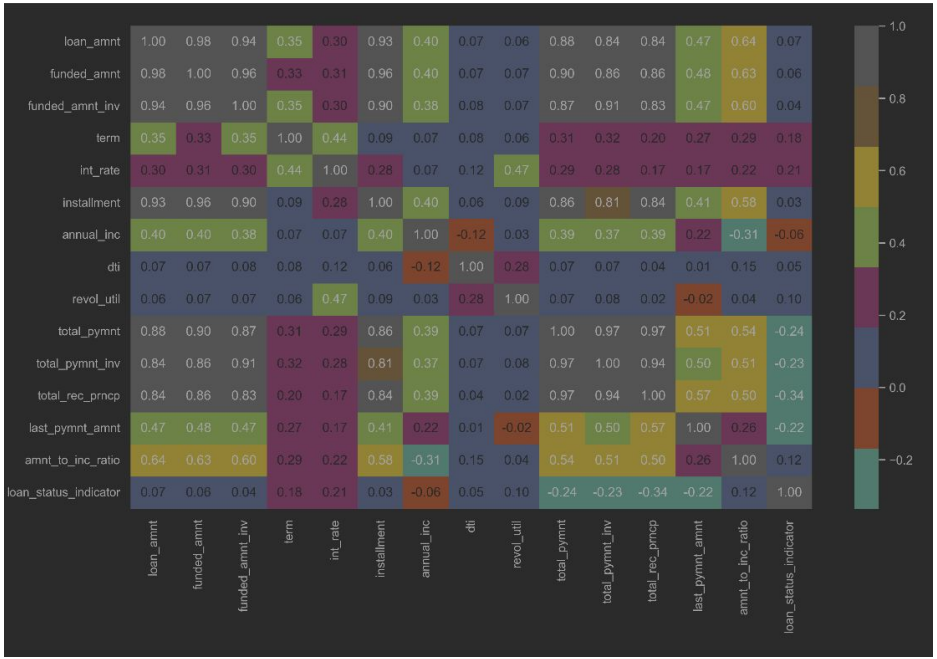


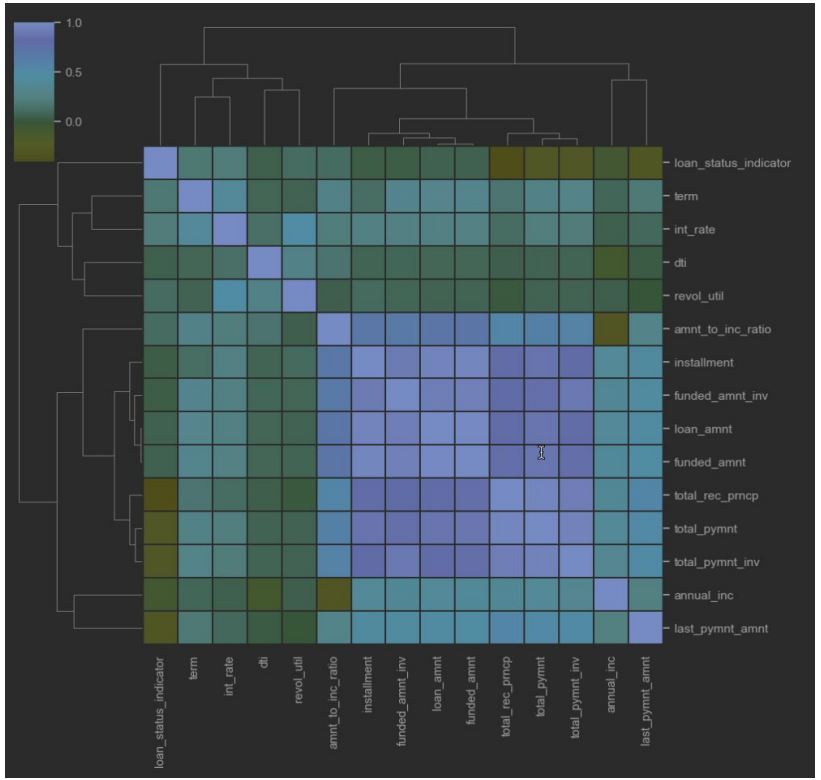
1. B, C and D grade has the highest charged off.
2. Loan issued for **36 Months** are the **highest 8.2% charged off**, Also Fully paid ratio is 67.1 % for 36 Months term loan.
3. Loans are **charged off 7%** of times when **rate of interest** is between **10-15%,** followed by **4.6%** when ROI is 15-20%
4. Income group ranging from **50K-100K and 25K-50K** are charged off highest no of times. **6.4% and 5.6%** respectively.
5. Applicants with **10+ years** and then **<1 Years** are Defaulted more number of times compared to others. **3.5%** and 1.6%
6. December and Novembers Charged Off numbers are slightly higher compare to other months.
7. In Year, 2011 Charged Off were the highest 8.5% significantly higher than other years.
8. **7.2% Charged off** are there when Home Ownership is **Rent** and then **6.1% when Mortgage**.
9. **9%** Defaulted loan were **verified**
10. 7.2% Loans defaulted were for **Debt_Consolidation**, And the 1.6% for Other Purpose.
11. CA, NY and FL states got highest number of default loan.



**More Plots in Notebook*

# Multivariate Plots



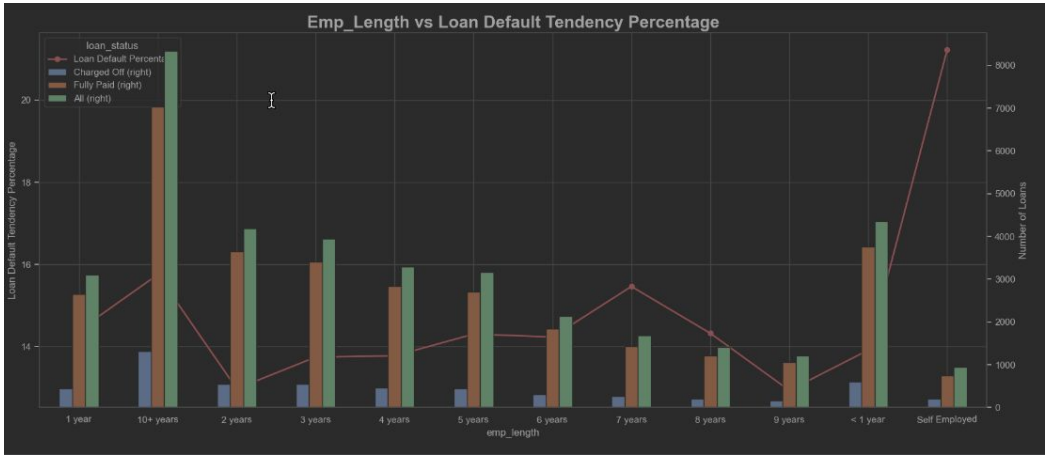




Loan amount and Funded amount and Funded Investor Amount are positively correlated

Annual Income and Amount to Income Ratio Negatively Correlated

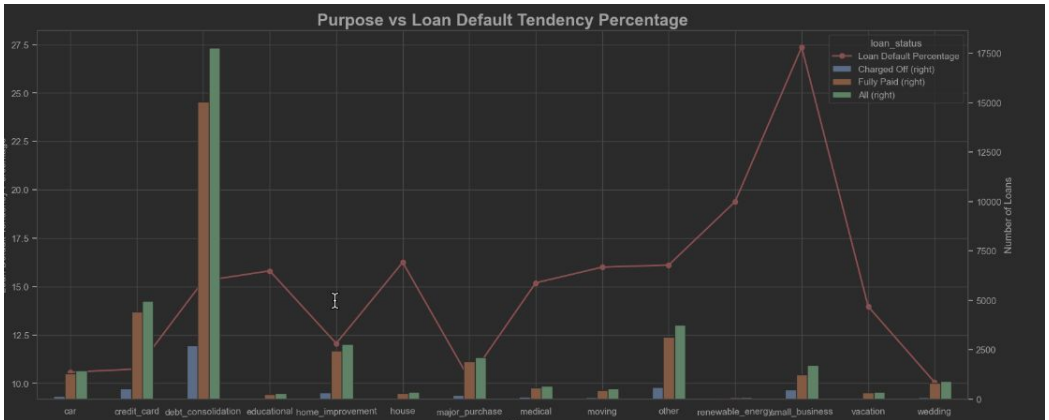
# Multivariate Plots






**Observation:**

Self Employed has the highest tendency to default, then 10+ and 7 Years employed borrowers comes highest

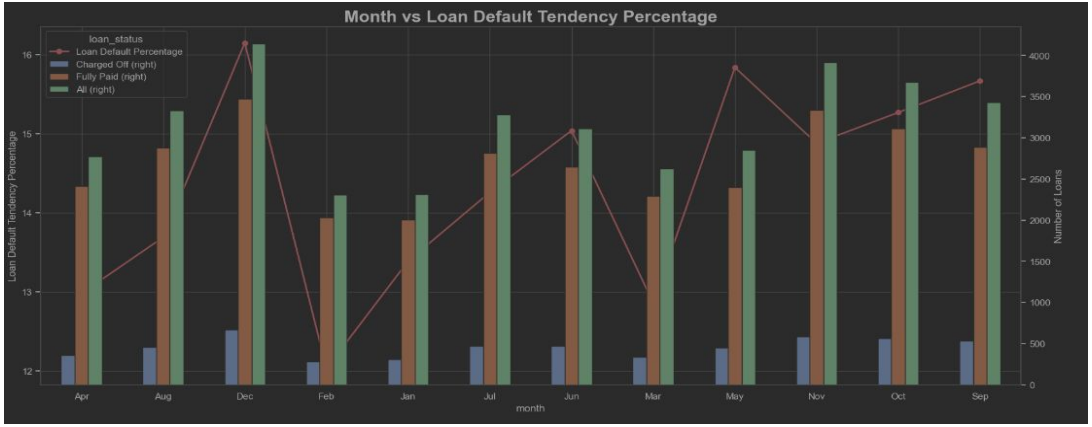



**Observation:**

Small Business has the more tendency to default and then Renewable Energy and then Other/House.

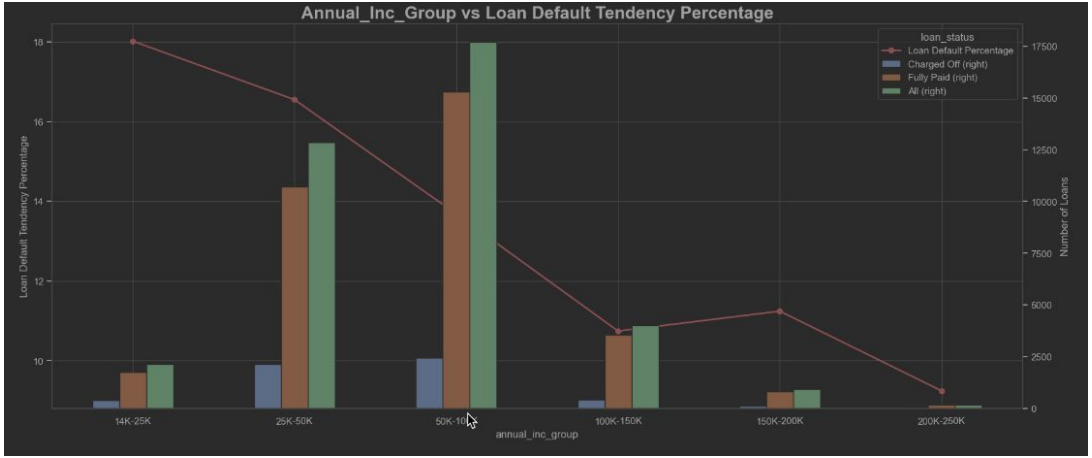
# EDA Multivariate Plots






**Observation:**

Dec, September and May months higher default rate is noticed.

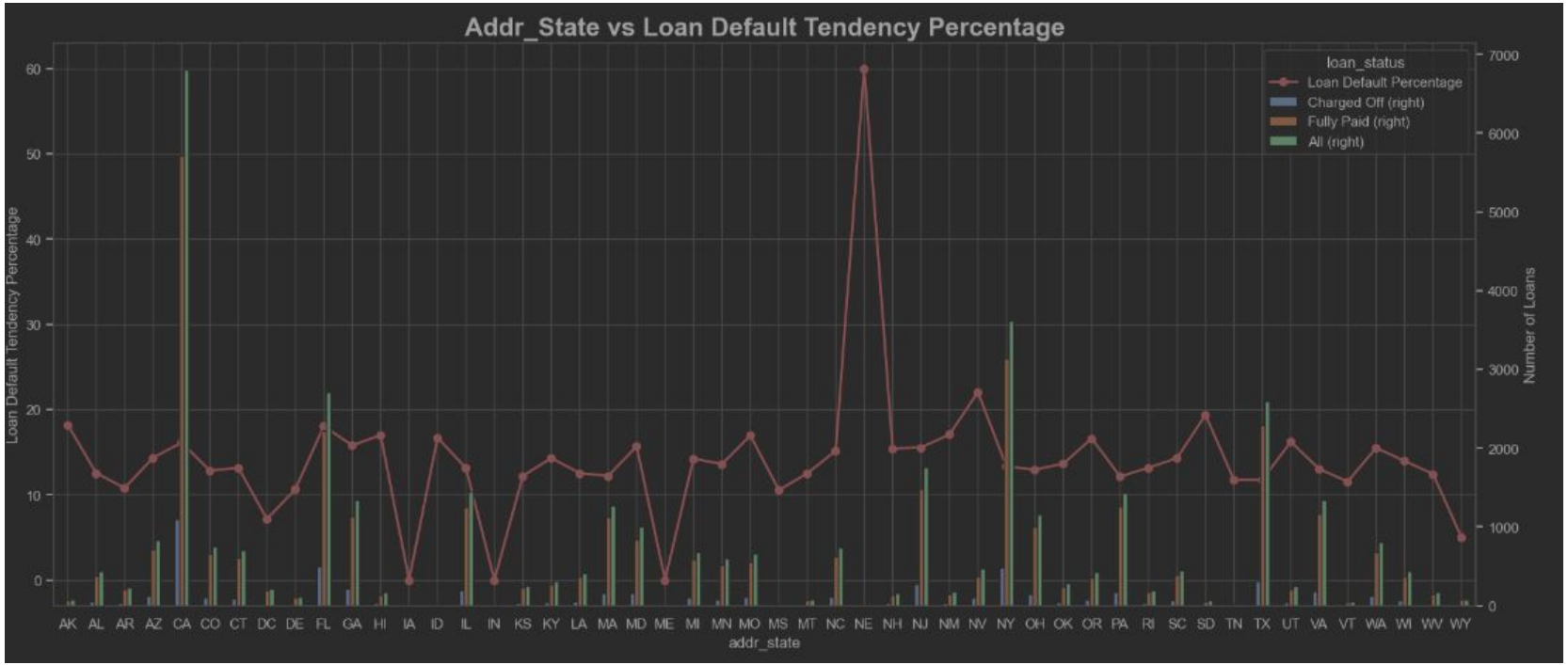



**Observation:**

Borrowers with $4K-50K income group are noticed to be defaulted more.

# Address State vs Count vs Loan Default Tendency %






Borrowers in the state with Defaulted %, NE 60%, NV ~22%, SD 20%, AK

# EDA - Multivariate Analysis Summary



- Heatmap and Pairplot are easier to perform multivariate analysis
- Crosstab table generation and plotting bar chart and line chart with hue of Loan Status Indicator.

**Analysis:**

1. Fund Amount, Investor Fund Amount are **positively correlated** with Loan Amount
2. Interestingly, after receiving 10K USD Principle in general there are fewer defaulters, Ranging Interest Rate between 10-20%
3. Loan Amount vs Amount to Income Ratio is **negatively** correlated.
4. **Dec, September and May** months has higher default tendency rate **near to 16%**
5. **Term 60** Months has the default rate higher **near to 25%**
6. **NE - 60%**, NV ~22%, SD-20%, AK-18% states where borrowers defaulted maximum
7. Interestingly Verified borrowers got defaulted more times than Non Verified.
8. **14K-50K USD** income group has the more tendency to default by **16-18%**
9. With **higher Rate of interest** tendency to default is high **near to 40%**
10. **Small Business** has the more **tendency (27.5%)** to default and then Renewable Energy and then Other/House.
11. **Home Ownership (18.5%)** with Other and Own and Rent has more tendency to default.
12. **Self Employed** has the highest **tendency (23%)** to default, then 10+ and 7 Years.
13. Sub Grade **F5(49%)**, G3, G5 has the highest tendency to default to Compare to A4,A5, B3,B5 where total defaulters are more because total loans issued are also more.
14. In Grade A  to Grade G (34%) borrowers to get defaulter is increased in tendency.

# Conclusion and Recommendation

As per the EDA analysis we clearly identified the key factor that is driving factor for the loan default. Below key factors to be considered while issuing the loan.

## Consumer Attributes

1. Address State
2. Grade and Subgrade
3. Purpose
4. Employment Length
5. Home Ownership

## Loan Attributes

1. Loan Amount to Income Ratio
2. Interest Rate
3. Term
4. Verification Status
5. Issued Month

While issuing loan above factors to be considered carefully, take the advantage of these analysis to be able to carefully issue a loan.

# Acknowledgements

*Credits to Upgrad teaching faculties, professional expert, coaches and buddy.*

Thank you