Lending Club Case Study

- Bhanu Prathap Velampati

Problem Statement

Goal is to identify certain patterns, which helps to execute final decision on the loan application to be accepted or rejected

- Identify pattern who falls under 'Fully Paid' category, but low to moderate chances that he/she may fall in charged off category in future

- Identify pattern who falls under ''Charged Off' category, still can accept their loan applications

Execution Summary

- Data exploration
- Data cleaning
- Derived variables
- Univariate Analysis
- Univariate Analysis & Results
- Bivariate Analysis & Results
- Multivariate Analysis
- Final observation

Data Exploration

- Go through all Data Frame variables and understand meaning of the variables and their usage

- Shortlisted below list of variables which are important for analysis

```
loan_df.columns

Index(['int_rate', 'title', 'loan_status', 'funded_amnt_inv', 'funded_amnt',
       'verification_status', 'home_ownership', 'issue_d', 'dti', 'purpose',
       'annual_inc', 'sub_grade', 'grade', 'term', 'loan_amnt',
       'pub_rec_bankruptcies', 'emp_length', 'pub_rec', 'addr_state',
       'revol_util'],
      dtype='object')
```

- Delete other variables which are not so important for analysis

Data Cleaning

- Identified all null values in the short-listed data frame and taken appropriate action by filling standby values

  - Ex.,

    - loan_df['pub_rec_bankruptcies'].isna().sum()

    - loan_df.pub_rec_bankruptcies.fillna('unknown',inplace=True)

- Convert float/object datatype into integer type which is used for analysis

- Remove special characters if any in the variable data

  - Ex.,

    - loan_df['emp_length']=loan_df.emp_length.astype('str').str.extractall('(\d+)').unstack().fillna(' ').sum(axis=1).astype(int)

Cont...

- Remove outliers using the help of boxplot or filtering data using conditions

```python
sns.boxplot(x='annual_inc',data=loan_df)
plt.show()
```
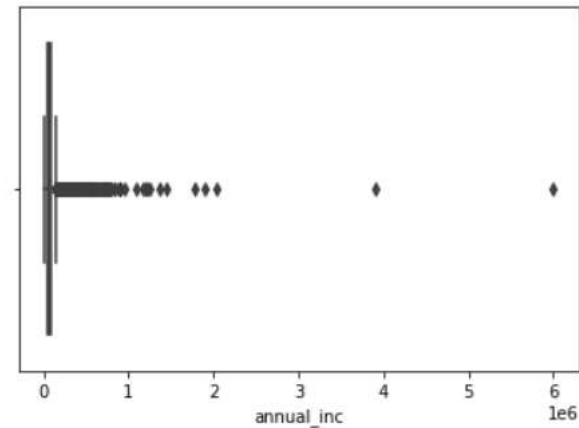


```python
loan_df.annual_inc.describe()
```

```
count    3.971700e+04
mean     6.896893e+04
std      6.379377e+04
min      4.000000e+03
25%      4.040400e+04
50%      5.900000e+04
75%      8.230000e+04
max      6.000000e+06
Name: annual_inc, dtype: float64
```

```python
loan_df=loan_df[~(loan_df.annual_inc > 3*10**6)]
```

Derived Variables

- By default, provided some variables cannot be used for analysis because of too many individual parameters and many records

- Create new categorical variables which can be used for analysis

  - Ex.,

    - loan_df['revol_util_cat'] = pd.cut(loan_df['revol_util_int'], [0, 20, 30, 50, 70, 100], labels=['0-20', '20-30', '30-50', '50-70', '70 +'])

    - loan_df['int_rate_cat']=pd.cut(loan_df['int_rate_int'], [0, 5, 10, 15, 25,30], labels=['0-5', '5-10', '10-15', '15-20','20 &More'])

Univariate Analysis

- Univariate analysis can disclose data in the dataset in the form of range of values

- Executed univariate analysis on all shortlisted variables using countplot, barchart, histplot etc

- Based on the plots on all shorted variable, my conclusion is interest rate and loan purpose plays plays important role in further analyses

- As shows in the images, interest rates increased as high as 24% and good number of "Charged Off" records exists
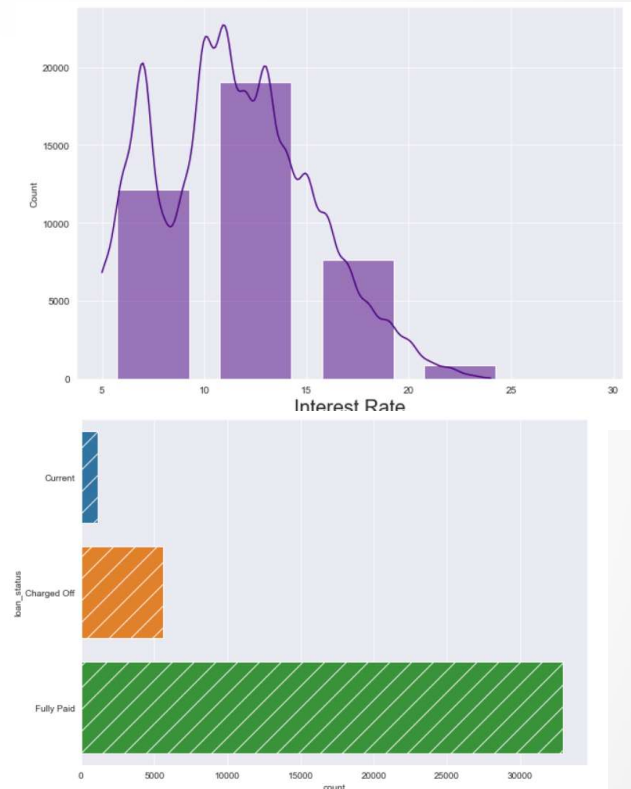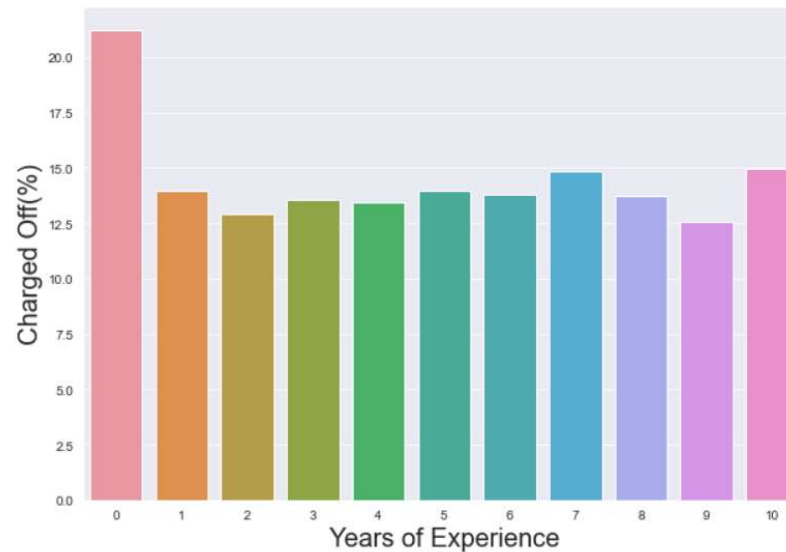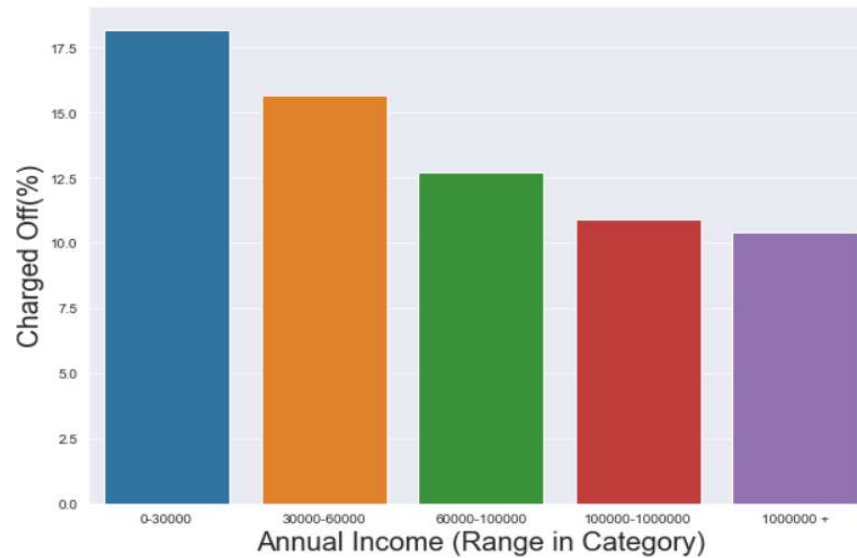
cont..

- Univariate Analysis results:
  - Below are top observations
    - loan_amnt, funded_amnt, funded_amnt_inv: Pattern in these variables is almost same, consider only funded_amnt_inv for further analysis
    - Moderate number of "Charged Off" records
    - Highest interest is around 24% which is pretty high
    - Employees who has <=1 years experience also given loans
    - Most of the loans were given for interest rates between 5-20
    - Most of the loans were allotted in the year 2010 and 2011
    - There are more number records which as revolving credit percentage more than 30 which is not good

Bivarate Analysis

- Bivariate analsys is done on two variables

    - In univariate analysis i see some patterns in interest rate and annual income which has some extreme entries like high interest rate and lowincomes.

    - Using Bivariate analysis, having plot on loan_status vs annual income, we can have some conclusions like

        - Customers getting low salaries have high chances of charged of ration

cont..

- Bivariate Analysis results:
  - Below are top observations
    - More interest rates when having less grades
    - More chances of charged off ratio when grade is less
    - Interest rate increased when funded amount increased
    - High chances of write off ratio when purpose of the loan is "Small Business"
    - If tenure is for 60months then more chances of charged of ratio
    - Less income can lead to more charged off ratio
    - Interest rate increasing year after year
    - High chances of charged off if bankrupt and pub_rec cases
    -

Multivariate Analysis

- Multivariate analysis can help us with some strong patterns to conclude some fair observations

  #1. Higher is the interst rate, more chances of applicants falling in "charged off" category

  - Ex,.
    - Annual income vs interest rate vs loan status
    - Funded amount vs interest rate vs loan status
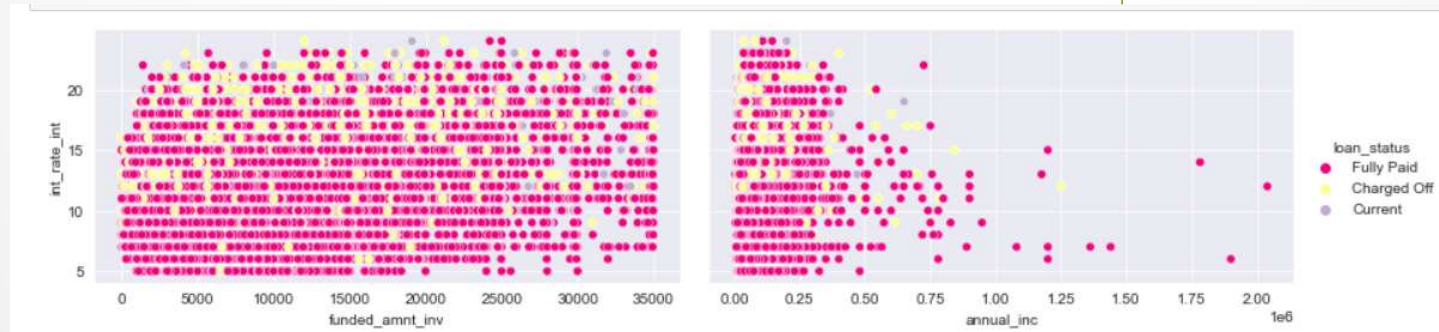
Final Observation

- Observations on applications who are already in Charged off category
    - Higher is the interest rate, more chances of applicants falling in "charged off" category
    - Higher is the interest rate when having low grades may fall in "charged off" category
    - Seeking loans for small business, other, house categories may fall in "charged Of" category
    - Up to some extent, with less income <30k and are having revol_credit score more than 30% may fall in "Charged Off" category
    - pub_rec having 1 and 2 and having court cases have high chances to fall in "Charged Off" category
    - High chances of "Charged off" ration when having more bankrupt cases
    - Less is the income and having high interest rate have high chances to fall in "Charged Off" category
- Observations on applications who are in "fully Paid" category
    - High risk to provide loans when having pub_rec as 1 and 2 and having court cases
    - High risk when revol_credit score greater than 30% and getting less income
    - High risk when interest rate is high for low-income people
    - Low risk when getting less income and having least grades