LENDING CLUB CASE STUDY

Rajashree Roy

Pritam Mukherjee

DATA LOADING & UNDERSTANDING

- Using Pandas library, we have loaded loan csv data in Pandas Data Frame.
- Understand the data and its structure using various functions like Head, describe and info.

```
# Reading loan csv file and observing the data.
raw_df = pd.read_csv("loan.csv")
raw_df.head(3)
```

✓ 0.6s

```
C:\Users\z004a8ac\AppData\Local\Temp\ipykernel_19032\727734610.py:2: DtypeWarn
  raw_df = pd.read_csv("loan.csv")
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	i
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	

3 rows × 111 columns

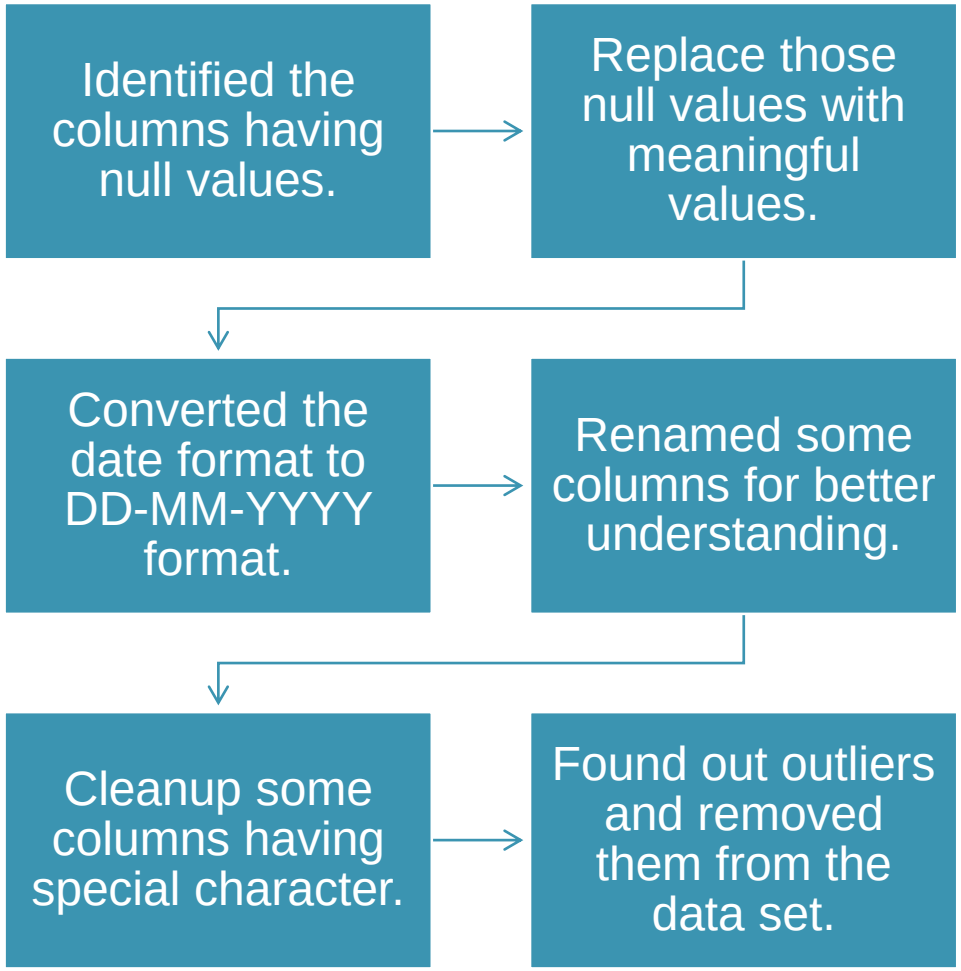
```
# Observing the data
raw_df.describe()
```

✓ 0.2s

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000

8 rows × 87 columns

DATA CLEANING AND MANIPULATION

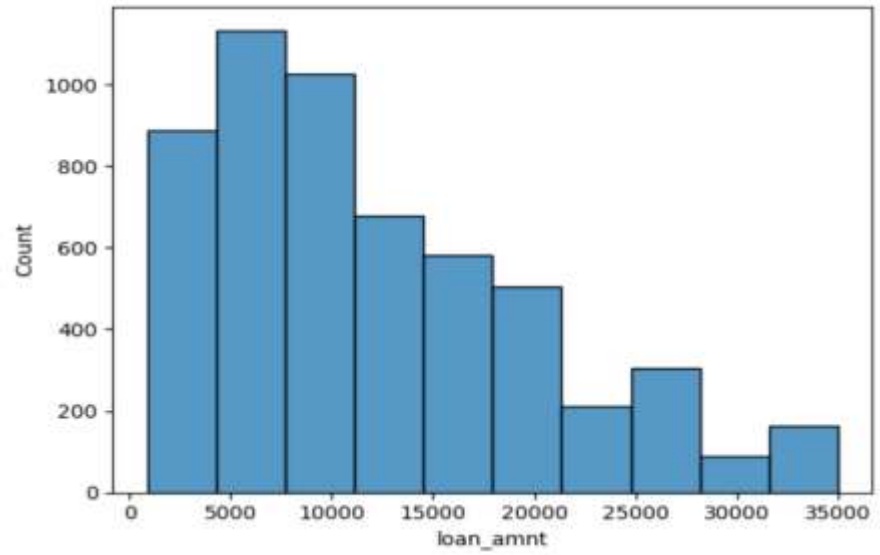


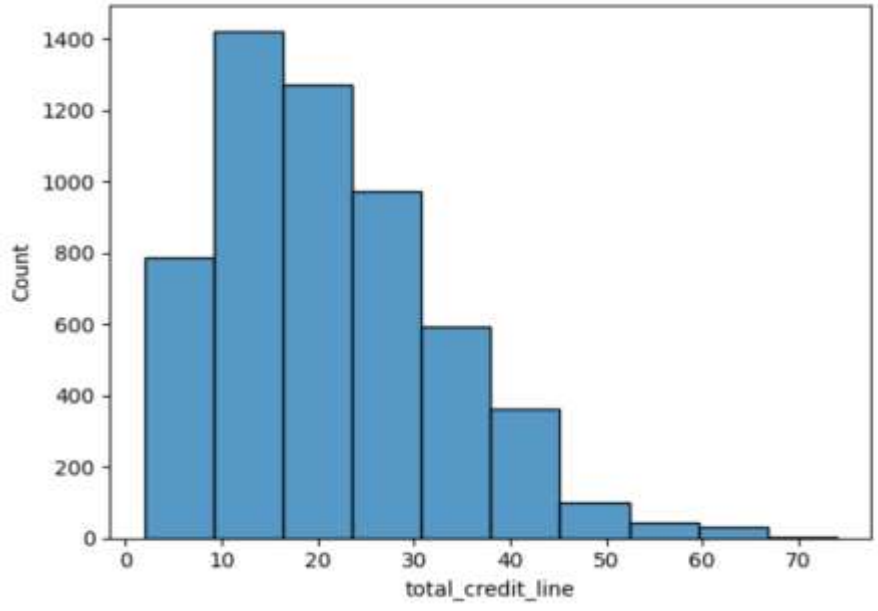
U N I V A R I A T E A N A L Y S I S

Understanding effect of numerical variables - loan amount and total credit line using histogram plot, to identify the defaulter's trends.

1. People taking loan in a range of Rs.5000 - Rs.10,000 are prone to default.

2. People having 10-20 outstanding credit lines are likely to default.



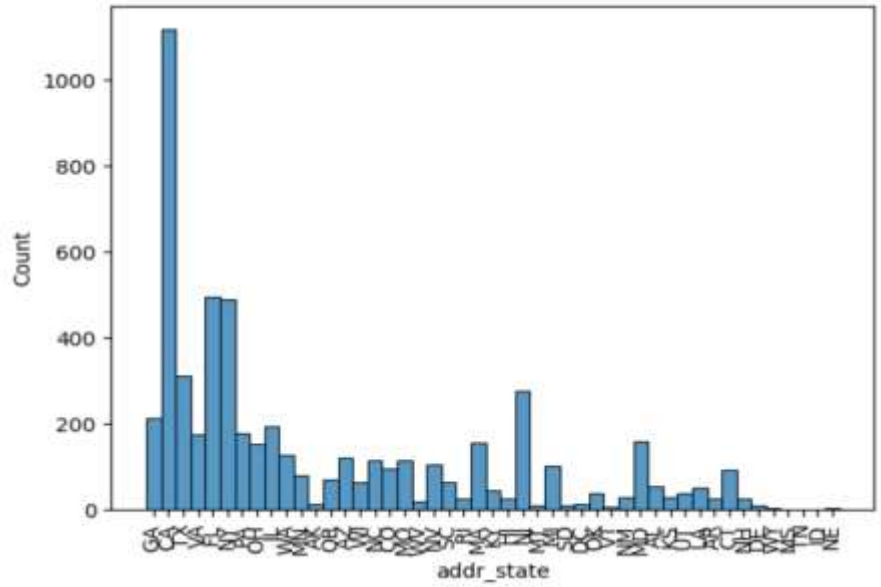


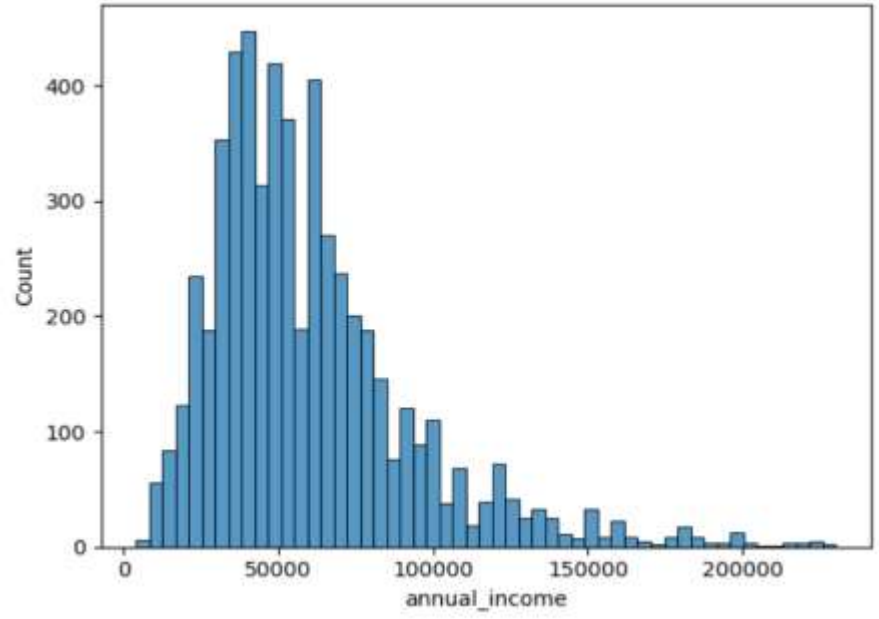
UNIVARIATE ANALYSIS

Understanding effect of important variables – address state and annual income using histogram plot, to identify the defaulter's trends.

1. Maximum default cases are coming from CA state

2. Maximum default cases are for people with annual income between 40000 to 80000.



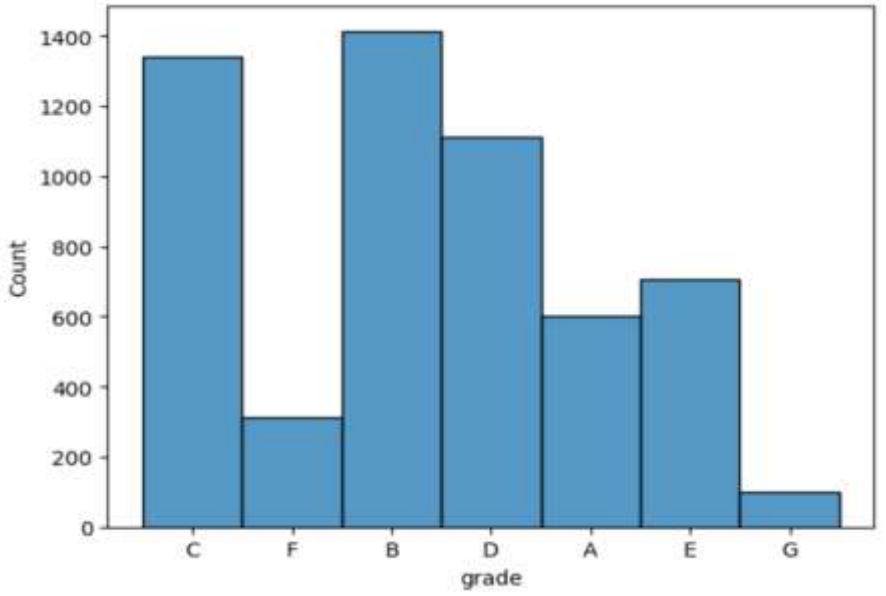


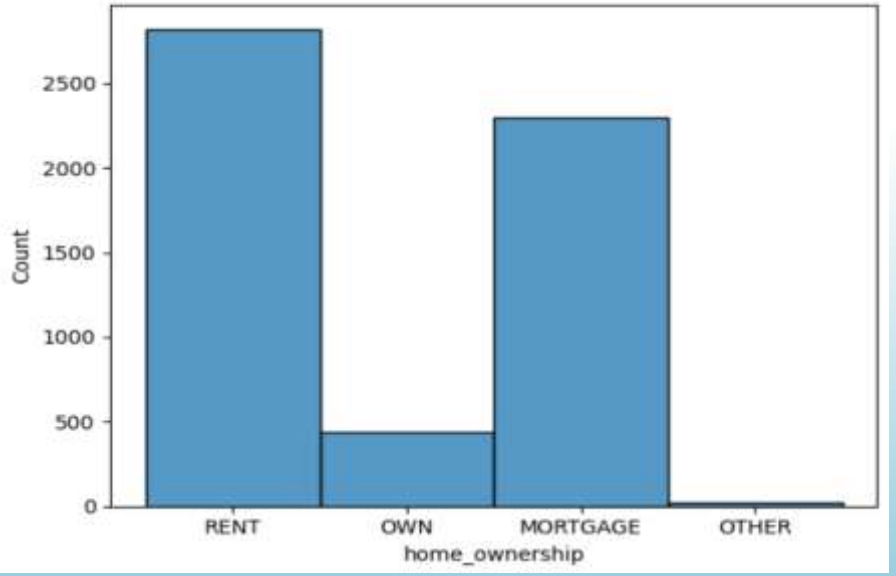
U N I V A R I A T E A N A L Y S I S

Understanding effect of categorical variables – grade and home ownership using histogram plot, to identify the defaulter's trends.

1. People in grade B,C,D are more likely to default.

2. People staying in rent or having Mortgage are more likely to default.



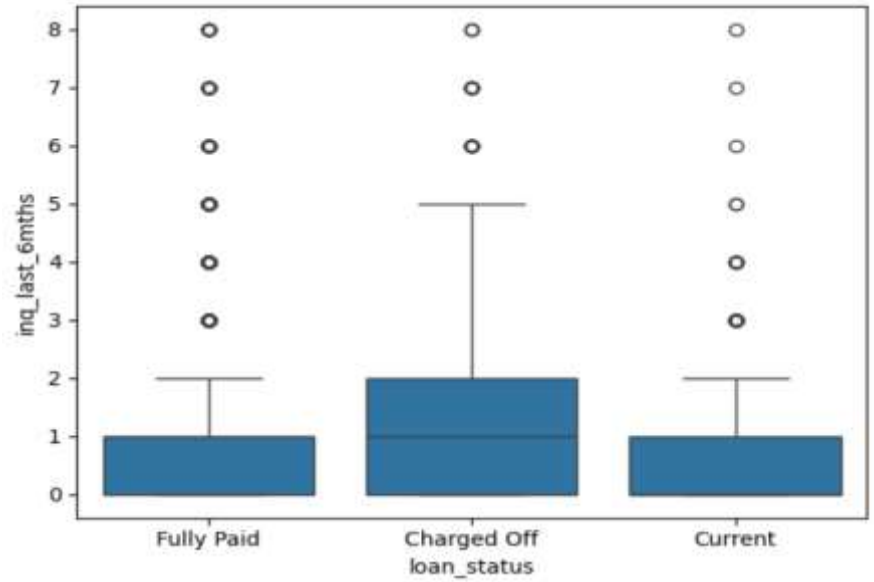


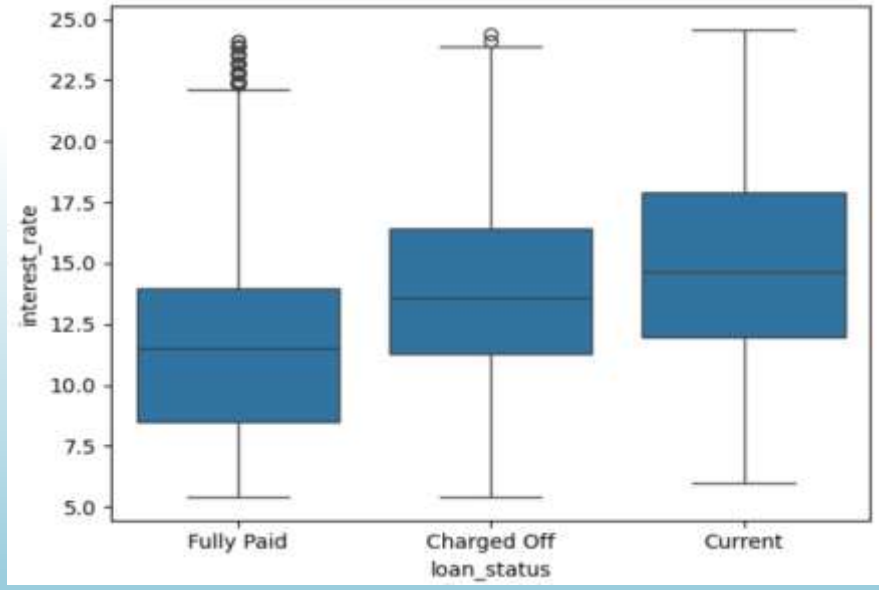
SEGMENTED UNIVARIATE ANALYSIS

Understanding effect of variables – inq_last_6mths and interest_rate with loan_status using boxplot, to identify the defaulter's trends.

1. People who enquired more in last 6mnths, are prone to default .

2. Chance to default is more if interest rate increase.

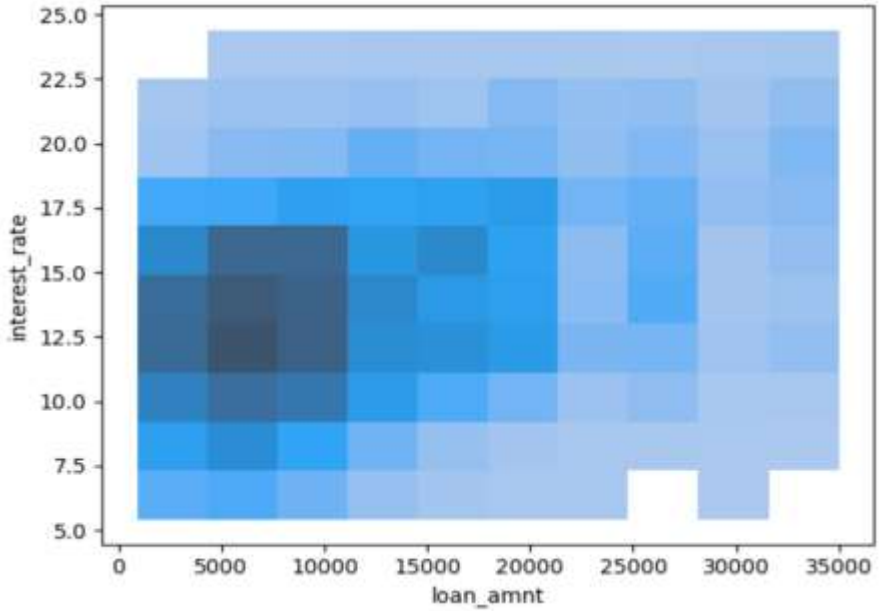


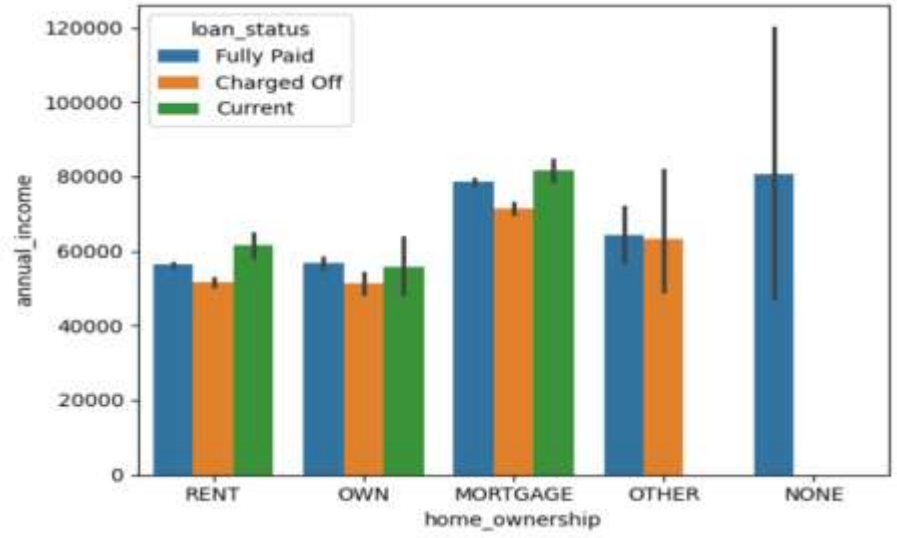


BIVARIATE ANALYSIS

Understanding effect of two variables – interest_rate with loan_amnt and annual_income with home_ownership, to identify the defaulter's trends.

1. Loan amount between 5K-10K with interest rate between 11%-16% has high chance to default
2. If annual income is above 50K and home ownership is Mortgage, chance of making defaults are high.



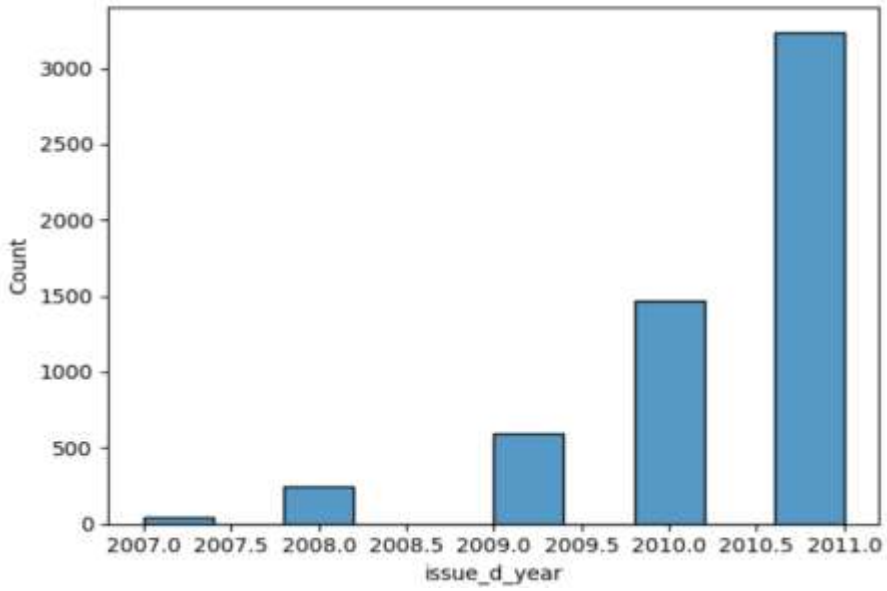


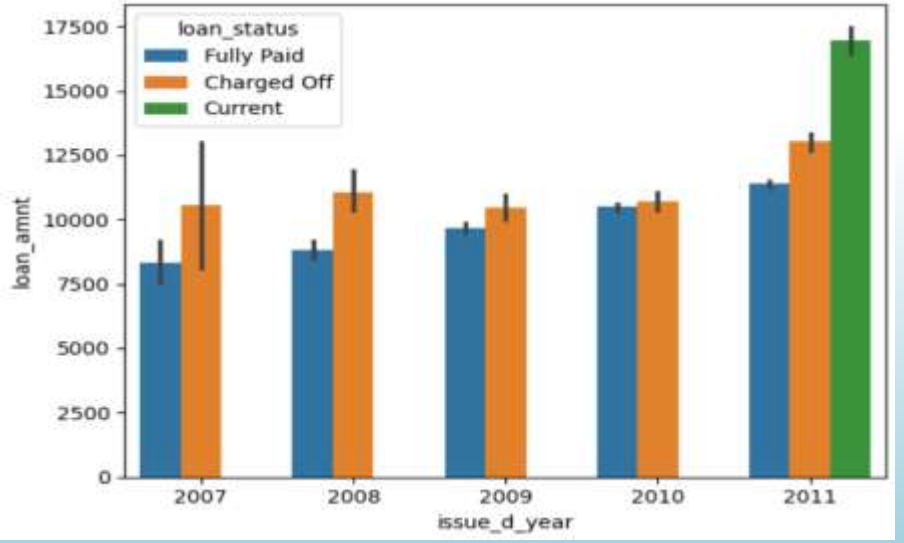
DERIVED METRICS

Understanding effect of derived variable – extracted year value from issue_d columns and used that column to identify some pattern.

1. Defaulter's count increased year-wise and most default happened in 2011

2. Defaulters increased as issued loan amount increased every year.





C O N C L U S I O N

From our analysis we can see below are few highly impactful variables for defaulters –

1. **Type of home ownership with respect to the customer's annual income can have effect on charge off.**
2. **Loan amount with certain interest rate can predict probable defaulters.**
3. **No of enquiries made by customers can be a good indicator for charged off customers.**
4. **Loan taken from particular state can also be a strong factor indicating chance of defaulting.**
5. **People with more outstanding credit line are more prone to default.**

THANK
YOU!