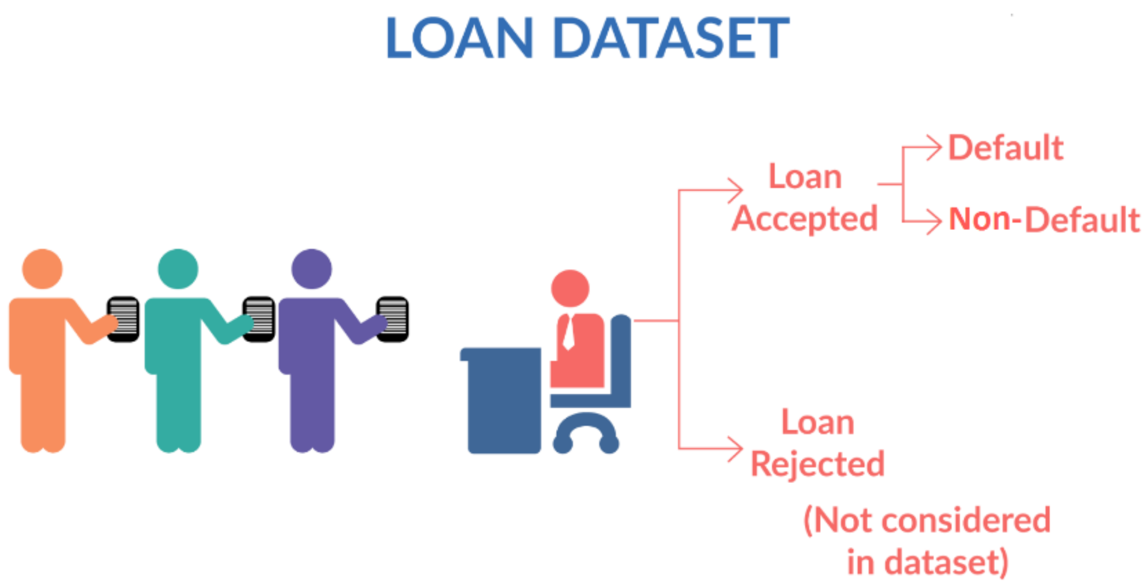
LENDING CLUB CASE STUDY

By Varun Shenoy

PROBLEM STATEMENT



AIM

To identify patterns which indicate if a person is likely to default, which may be used for taking meaningful actions.

APPROACH

Importing data → Exploratory data analysis → Cleaning data → Analyzing data to solve the aim → Conclusion / Recommendations

- Importing Libraries
- Making a data frame
- Understand the rows and columns
- Strategizing an approach
- Deriving stats and info of the data
- Removing unwanted rows columns
- Imputing data, converting data types
- Removal of outliers

- Univariate analysis
- Segmented analysis
- Bivariate analysis
- Visualising analysis outputs
- Observations and conclusions
- recommendations

THE DATA- EDA 1

- Data has 39717 rows and 111 columns

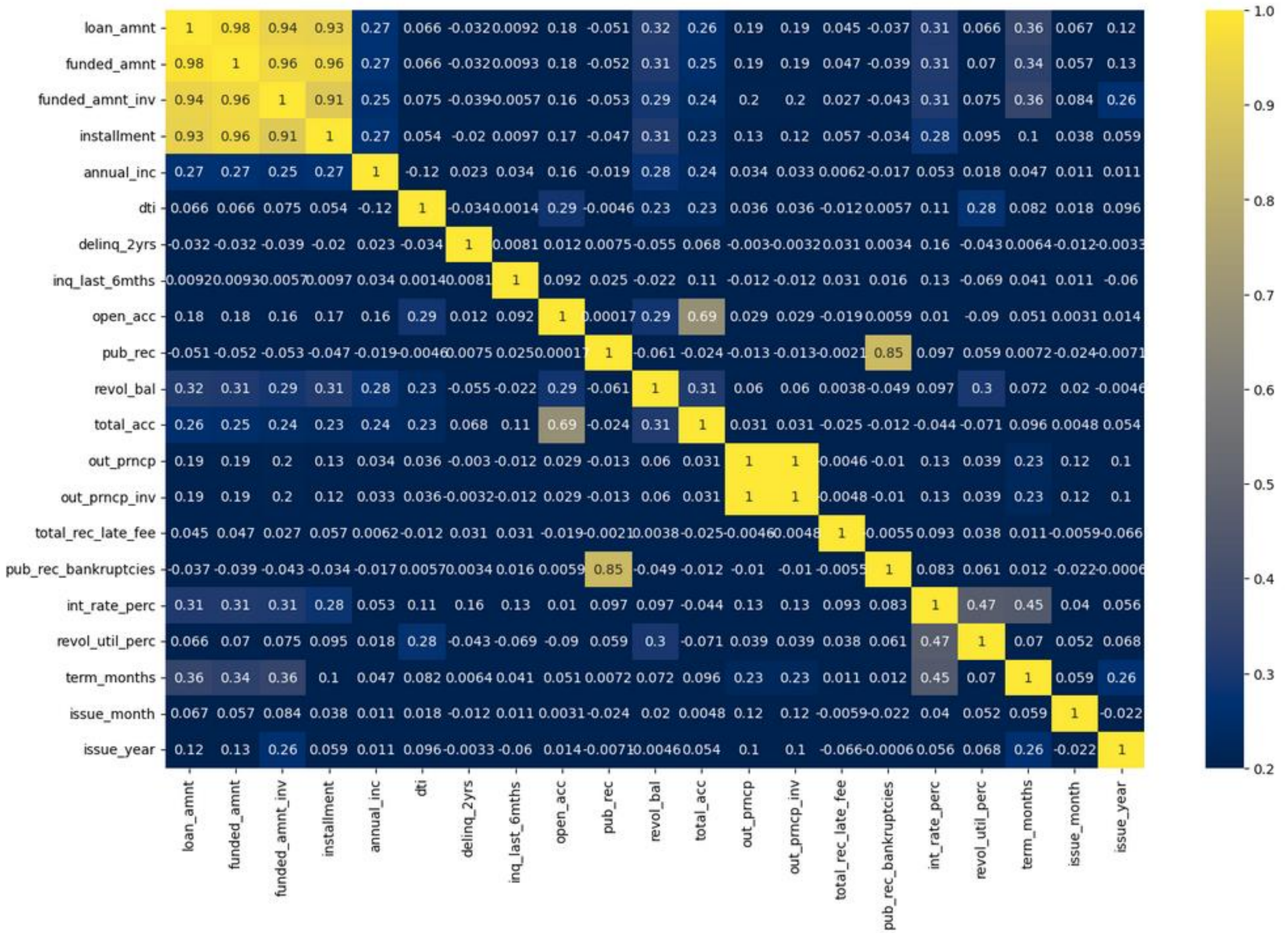
- Preliminary data analysis and observations was indicative that not all data are meaningful and many columns are blanks (NAN) or single entry columns (0,f, Individual) etc.

- Data quality checks: 54 blank columns were observed of 111 columns

  - Functions used for cleaning: df.dropna() and df.drop() used

  1. Blank columns removed
  2. Single entry columns were identified and removed
  3. Loan logging columns like id url zip etc were removed these colums add little insights to solving the aim

THE DATA- EDA 2



- Issue date column was split to month and year
- Many columns were stripped of % symbols and string entries to get data in proper data types
- Additional columns were dropped by checking its correlation with each other and neutral correlated columns and similar columns were further removed
- Correlation plot screenshot attached
- Strong and weak correlated data are studied using fully paid loan status
- Data types are converted into numerical and object data wherever applicable by using pd.to_numeric()

THE DATA- EDA 3

- The final data frame consisted of 39717 and 25 columns

#	Column	Non-Null Count	Dtype
0	loan_amnt	39717 non-null	int64
1	funded_amnt_inv	39717 non-null	float64
2	installment	39717 non-null	float64
3	grade	39717 non-null	object
4	sub_grade	39717 non-null	object
5	emp_title	37258 non-null	object
6	emp_length	38642 non-null	object
7	home_ownership	39717 non-null	object
8	annual_inc	39717 non-null	float64
9	verification_status	39717 non-null	object
10	loan_status	39717 non-null	object
11	purpose	39717 non-null	object
12	title	39706 non-null	object
13	addr_state	39717 non-null	object
14	dti	39717 non-null	float64
15	delinq_2yrs	39717 non-null	int64
16	open_acc	39717 non-null	int64
17	pub_rec	39717 non-null	int64
18	revol_bal	39717 non-null	int64
19	total_acc	39717 non-null	int64
20	pub_rec_bankruptcies	39020 non-null	float64
21	int_rate_perc	39717 non-null	float64
22	term_months	39717 non-null	int64
23	issue_month	39717 non-null	int64
24	issue_year	39717 non-null	int64

STATISTICAL ANALYSIS-UNIVARIATE ANALYSIS OF CATEGORICAL DATA

For all loans

```
## understanding behaviour of different categorical valriables for all loans

categorical_cols=["grade",'sub_grade',"emp_title","emp_length","home_ownership","verification_status","loan_status","purpose
for val in categorical_cols:
    #print(filtered_loan_df[val].value_counts(dropna=False))
    print("Maximum people have a loan of",val,filtered_loan_df[val].value_counts().index[0],filtered_loan_df[val].value_cour
```

```
Maximum people have a loan of grade B 12020 i.e 30.26 % of all loans
Maximum people have a loan of sub_grade B3 2917 i.e 7.34 % of all loans
Maximum people have a loan of emp_title US Army 134 i.e 0.34 % of all loans
Maximum people have a loan of emp_length 10+ years 8879 i.e 22.36 % of all loans
Maximum people have a loan of home_ownership RENT 18899 i.e 47.58 % of all loans
Maximum people have a loan of verification_status Not Verified 16921 i.e 42.6 % of all loans
Maximum people have a loan of loan_status Fully Paid 32950 i.e 82.96 % of all loans
Maximum people have a loan of purpose debt_consolidation 18641 i.e 46.93 % of all loans
Maximum people have a loan of title Debt Consolidation 2184 i.e 5.5 % of all loans
Maximum people have a loan of addr_state CA 7099 i.e 17.87 % of all loans
```

For defaulted loans

```
Top 2 categories of defaults for grade are
B    1425
C    1347
Name: grade, dtype: int64
i.e 49.26% which accounts to total defaults under grade

Top 2 categories of defaults for sub_grade are
B5    356
B3    341
Name: sub_grade, dtype: int64
i.e 12.39% which accounts to total defaults under sub_grade

Top 2 categories of defaults for emp_title are
Bank of America    20
US Army            18
Name: emp_title, dtype: int64
i.e 0.68% which accounts to total defaults under emp_title

Top 2 categories of defaults for emp_length are
10+ years    1331
< 1 year      639
Name: emp_length, dtype: int64
i.e 35.01% which accounts to total defaults under emp_length

Top 2 categories of defaults for home_ownership are
RENT        2839
MORTGAGE    2327
Name: home_ownership, dtype: int64
i.e 91.81% which accounts to total defaults under home_ownership

Top 2 categories of defaults for verification_status are
Not Verified    2142
Verified        2051
Name: verification_status, dtype: int64
i.e 74.52% which accounts to total defaults under verification_status

Top 2 categories of defaults for purpose are
debt_consolidation    2767
other                  633
Name: purpose, dtype: int64
i.e 60.42% which accounts to total defaults under purpose

Top 2 categories of defaults for title are
Debt Consolidation         305
Debt Consolidation Loan    274
Name: title, dtype: int64
i.e 10.29% which accounts to total defaults under title

Top 2 categories of defaults for addr_state are
CA    1125
FL     504
Name: addr_state, dtype: int64
i.e 28.95% which accounts to total defaults under addr_state
```

Observations

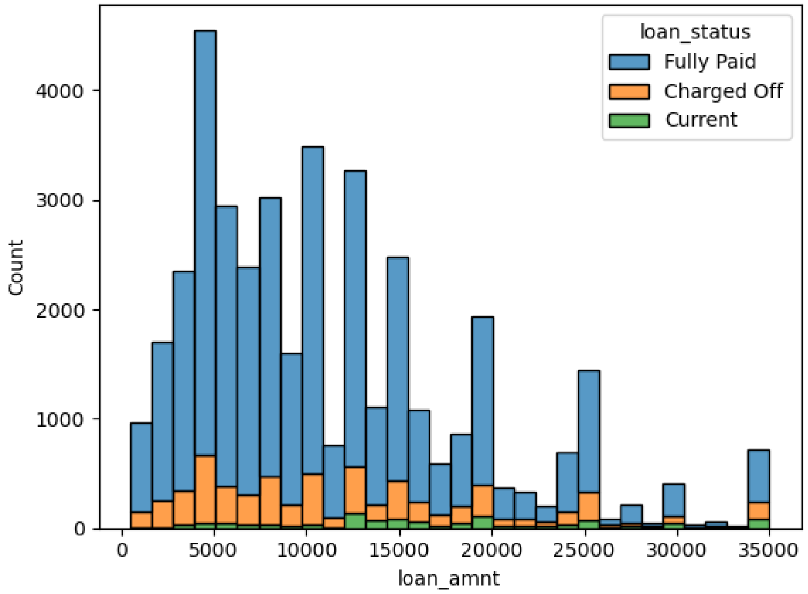
1. Source of income of around 42.6% of borrowers are not verified by LC.This is a huge number of people and hence chances of default can be reduced with proper verification.
2. Professionals with 10+ years work experience make 22.3% of the borrowers
3. Maximum loans are availed from or in State of CA will be studied with other select categorical variables for defaultSurprisingly Grade B and C observes maximum number of defaults ~50%
4. Better income source verification can reduce the default strongly
5. Majority of defaults 90% are by borrowers who have a home mortgage or are on Rent

STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

- Loan amount with loan status histogram plot- univaraite analysis

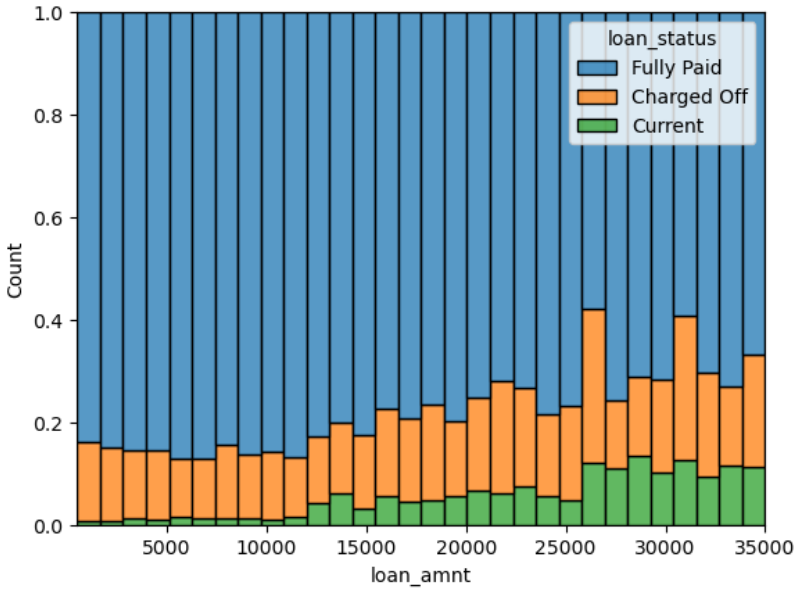
```
#understand defaulters at different loan bands
sns.histplot(data=filtered_loan_df,bins=30,x="loan_amnt",hue="loan_status",multiple='stack')
```

<Axes: xlabel='loan_amnt', ylabel='Count'>



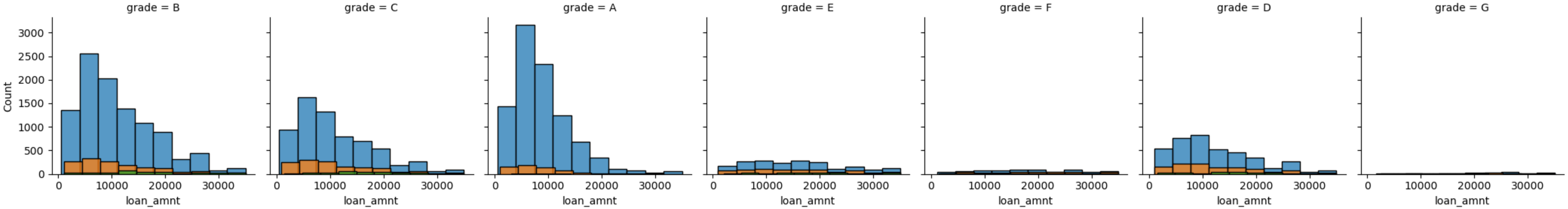

```
#to understand percentage of defaulters
sns.histplot(data=filtered_loan_df,bins=30,x="loan_amnt",hue="loan_status",multiple='fill')
```

<Axes: xlabel='loan_amnt', ylabel='Count'>



STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

Grade segmented analysis



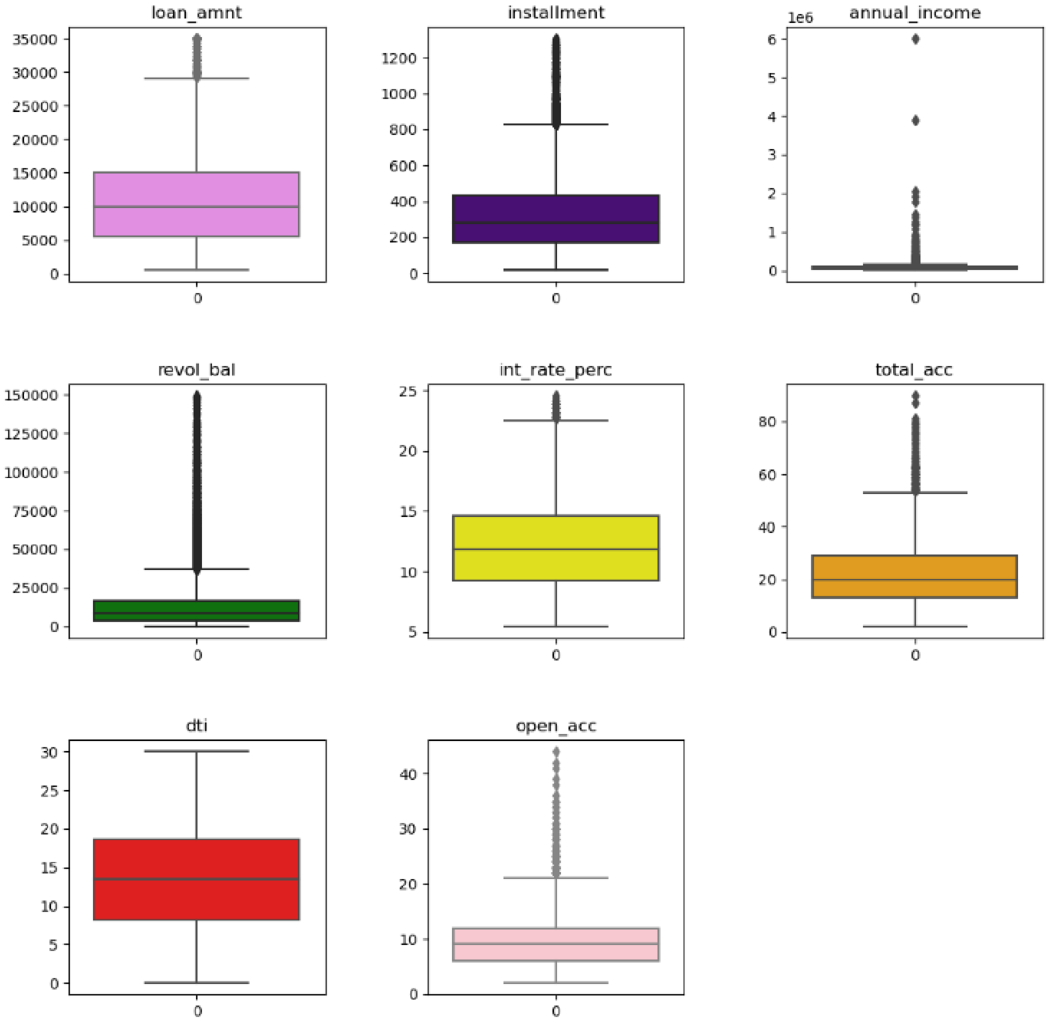
Observations

1. After Grade D the % of defaults increase in almost all bands

2. The histogram shows that defaults are comparatively higher above 15000. This may not be representative as 75% of data lies below 15000 for loan amount

3. In current scenario high amount borrowers are more and hence caution advised

ANALYZING OUTLIERS



Observations

1. High fluctuations are observed for Annual income the data is described using .describe() and studied for all loans v/s defaults to understand the behavior of annual income

All loans

	loan_amnt	funded_amnt_inv	installment	annual_inc \
count	39717.000000	39717.000000	39717.000000	3.971700e+04
mean	11219.443815	10397.448868	324.561922	6.896893e+04
std	7456.670694	7128.450439	208.874874	6.379377e+04
min	500.000000	0.000000	15.690000	4.000000e+03
25%	5500.000000	5000.000000	167.020000	4.040400e+04
50%	10000.000000	8975.000000	280.220000	5.900000e+04
75%	15000.000000	14400.000000	430.780000	8.230000e+04
max	35000.000000	35000.000000	1305.190000	6.000000e+06

defaulted loans

	loan_amnt	funded_amnt_inv	installment	annual_inc
count	5627.000000	5627.000000	5627.00000	5.627000e+03
mean	12104.385108	10864.521324	336.17500	6.242730e+04
std	8085.732038	7661.750540	217.05184	4.777601e+04
min	900.000000	0.000000	22.79000	4.080000e+03
25%	5600.000000	5000.000000	168.55500	3.700000e+04
50%	10000.000000	9401.209477	293.87000	5.300000e+04
75%	16500.000000	15000.000000	457.84000	7.500000e+04
max	35000.000000	35000.000000	1305.19000	1.250000e+06

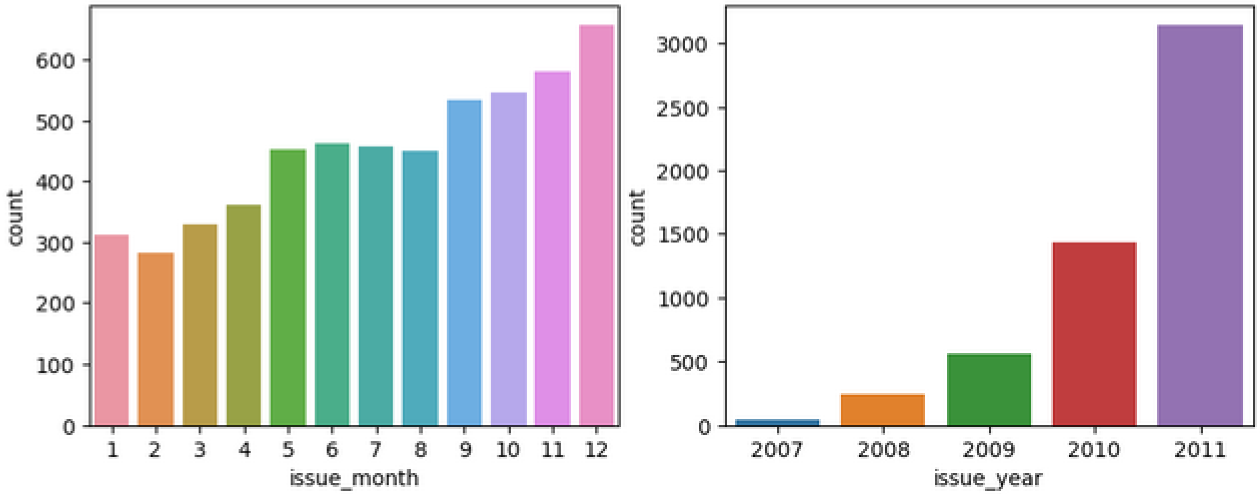
2. 95^{th} percentile is used to remove outliers from annual income and new filtered data set is formed

STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

- Univariate analysis on month and year

```
plt.figure(figsize=(10,8))
plt.subplot(221)
sns.countplot(x='issue_month', data=filtered_loan_df[filtered_loan_df['loan_status']=='Charged Off'])
plt.subplot(222)
sns.countplot(x='issue_year', data=filtered_loan_df[filtered_loan_df['loan_status']=='Charged Off'])
```

```
<Axes: xlabel='issue_year', ylabel='count'>
```

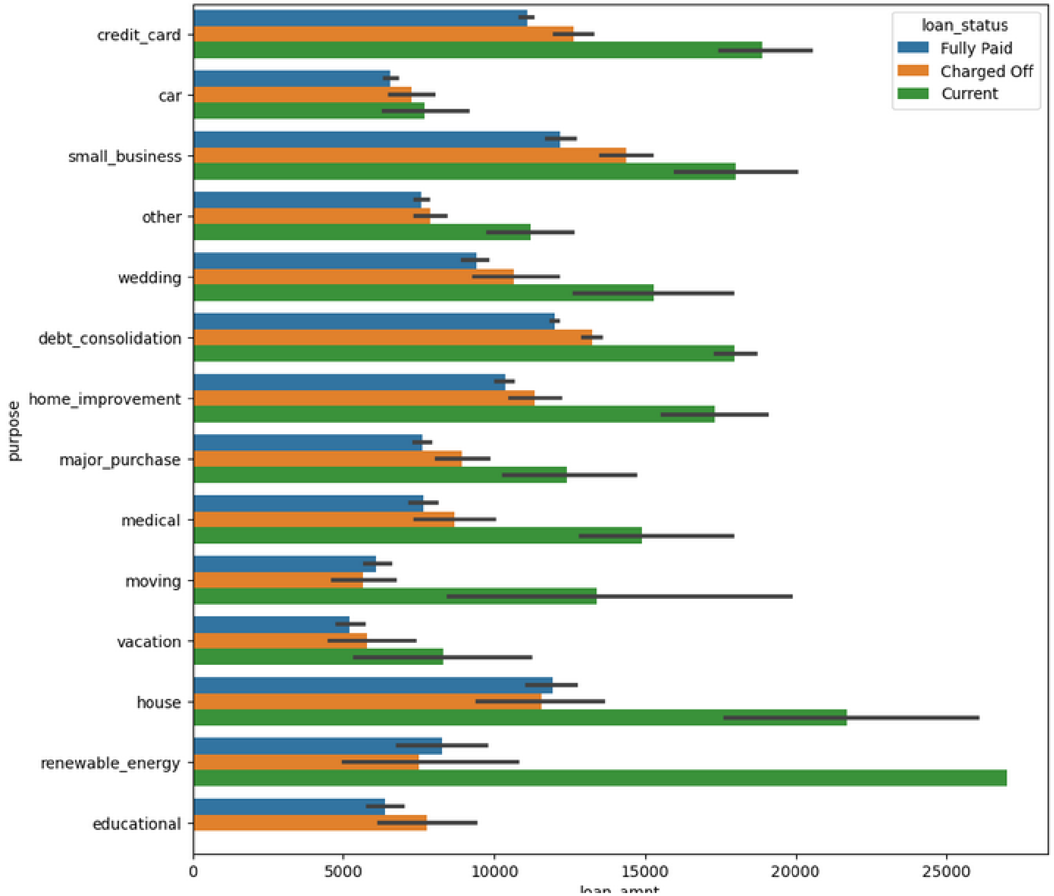


<u>Observations</u>

1. Highest loan availed was in 2011 and highest loan availing quarter was Q4 (in all years)
2. Loan applicants are increasing YoY almost in exponential manner

STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

- Understanding purpose and defaults with loan amount



- Understanding employee tenure and default
- Groupby() used for employee length and loan status

loan_status	Charged Off	Current	Fully Paid
emp_length			
9 years	150	31	1004
8 years	194	40	1151
7 years	252	58	1392
6 years	295	57	1781
5 years	441	81	2607
4 years	443	90	2759
1 year	449	66	2598
3 years	537	76	3293
2 years	548	91	3557
< 1 year	617	69	3714
10+ years	1270	354	6623

STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

- Pivot table of median value of all entries to gain better insights
- used with aggfunc of np.median

loan_status	annual_inc	delinq_2yrs	dti	funded_amnt_inv	installment	int_rate_perc	issue_month	issue_year	loan_amnt	open_acc	pub_rec	pub_rec_bankruptcies	revol_bal	term_months	total_acc
Charged Off	51996.0	0	14.40	9000.0	286.99	13.49	8	2011	10000	8	0	0.0	8926	36	19
Fully Paid	57000.0	0	13.43	8200.0	267.74	11.49	7	2011	9000	8	0	0.0	8418	36	20

- Pivot table for understanding sub-grades wrt defaults
- used with aggfunc of stats. mode

grade	A	B	C	D	E	F	G
loan_status							
Charged Off	([A5], [213])	([B5], [349])	([C1], [328])	([D2], [264])	([E1], [185])	([F1], [80])	([G1], [30])
Current	([A5], [26])	([B3], [89])	([C1], [78])	([D4], [61])	([E2], [39])	([F1], [18])	([G1], [9])
Fully Paid	([A4], [2579])	([B3], [2332])	([C1], [1637])	([D2], [959])	([E1], [490])	([F1], [196])	([G1], [58])

STATISTICAL ANALYSIS AND VISUALIZATION-
ON NUMERICAL DATA (SEGMENTED UNIVARIATE AND BIVARIATE)

- Defaulters for term and interest

Pivot table used with agg func as np.median()

term_months	36	60
loan_status		
Charged Off	12.53	15.99
Current	NaN	14.27
Fully Paid	10.75	14.17

- Highest defaulters for Grouped interest rates and installment
- Pd.cut and Pivot table used with agg func as np.median()

int_rate_perc	5-10%	10-15%	15-20%	20-25%
loan_status				
Charged Off	210.325	268.52	347.98	534.235
Fully Paid	224.630	274.48	347.79	539.840

OBSERVATIONS FROM SLIDE 12-15

1. People are likely to avail loans and default when purpose of loan is small business, credit card hence these claims need to verified and evaluated thoroughly

2. Those who are less than a year of work experience are likely to default. this is as expected

3. people with more that 10 years of experience are likely to default this needs to be investigated by LC

4. if interest rate and term of loan is more the default percentage is more

5. if interest rate and installment time is more the default percentage is more

CONCLUSIONS

Categories which can increase risk of default by borrowers are

1. median annual income less than ~52000
2. Home ownership- Rented or Mortgaged
3. Employment tenure is less than a year or more than 10 years
4. Loan application for amount more than 15000
5. Purpose of loan being for Debt consolidation, small business, credit card
6. Higher loan tenure borrowers i.e 60 months and higher installments have a higher default percentage
7. Most defaulting states are CA and FL can attract marginally higher intrest rates
8. Grade of D E F have a very high default count
9. Dti ratio of above ~14
10. Source not verified is the biggest cause in 74% of all defaults