GRAMENER CASE STUDY

RISK ANALYSIS OF LOAN APPLICATIONS

Group Members:

Ranip Hore
Sanchari Gautam
Amith Pradhan
Pratik Nath

Case Study Overview

- **Context**:

  The company 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. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Once the application is accepted and the loan is sanctioned then a borrower will re-pay the amount in monthly installments completely or can default leading to credit loss.

- **Problem Description**:

  Two types of risks are associated with the bank's decision:

  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

  The analysis is to be done to find measures, so that the company does not end up in credit loss.

- **Methodology**:

  - After data cleaning and manipulation, **Exploratory Data Analysis** is implemented to portrait risky loan applicants.
  - Univariate, bivariate and multivariate analysis are implemented on the previously approved loan applications' dataset, in order to bring the driving factors of credit loss in the nexus.
  - The analyzed dataset is then presented by data visualization techniques.

Data Understanding And Manipulation

- The data provided contains the information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- A lot of columns are having NaN values as well as values such as 0. We need to drop these columns during data cleansing. Also for few columns like '*installment*','*delinq_2yrs*' etc. there are outliers since there is a noticeable gap between 75% values and max value. We will analyse the outliers later during univariate analysis.

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	r
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000	
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200	
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219	
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000	
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000	
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000	
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000	
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000	

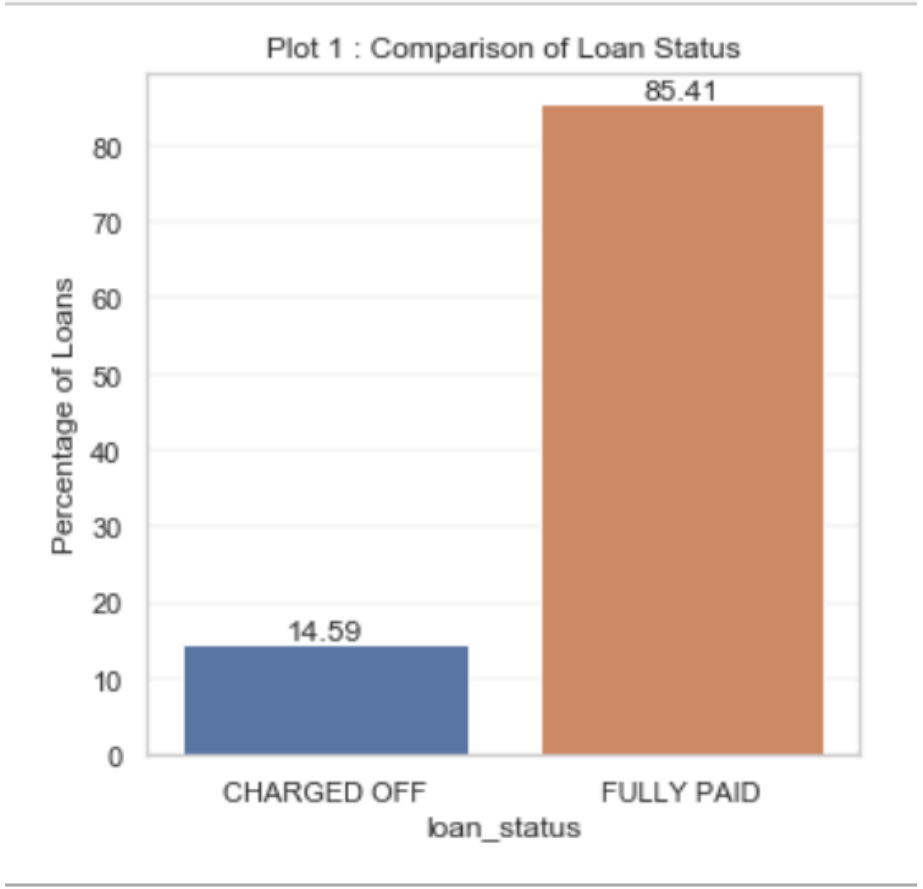
```
loan_data.loc[:,['issue_d','earliest_cr_line','last_pymnt_d','next_pymnt_d','last_credit_pull_d']].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 5 columns):
issue_d              39717 non-null object
earliest_cr_line     39717 non-null object
last_pymnt_d         39646 non-null object
next_pymnt_d         1140 non-null object
last_credit_pull_d   39715 non-null object
dtypes: object(5)
memory usage: 1.5+ MB
```

- Following 5 fields are not in standard date format [*Issue_d , earliest_cr_line , last_pymnt_d , next_pymnt_d & last_credit_pull_d*]. Hence, they are converted to appropriate date format.
- Some columns like *Loan Desc, Loan Title, URL* have heavy text format and are not necessary in data analysis. Hence we remove those columns.
- The columns *int_rate* and *revol_util* are in string format due to the presence of % symbol and the column *term* is in character format due to presence of months. All these are converted to float and integer format.
- The columns having more than 90% null value and NA values are deleted.
- All characters in all the cells are converted to uppercase in order to avoid any case sensitive issue.
- All the rows having more than 5 missing values are deleted.
- All columns having only one unique value are also dropped.

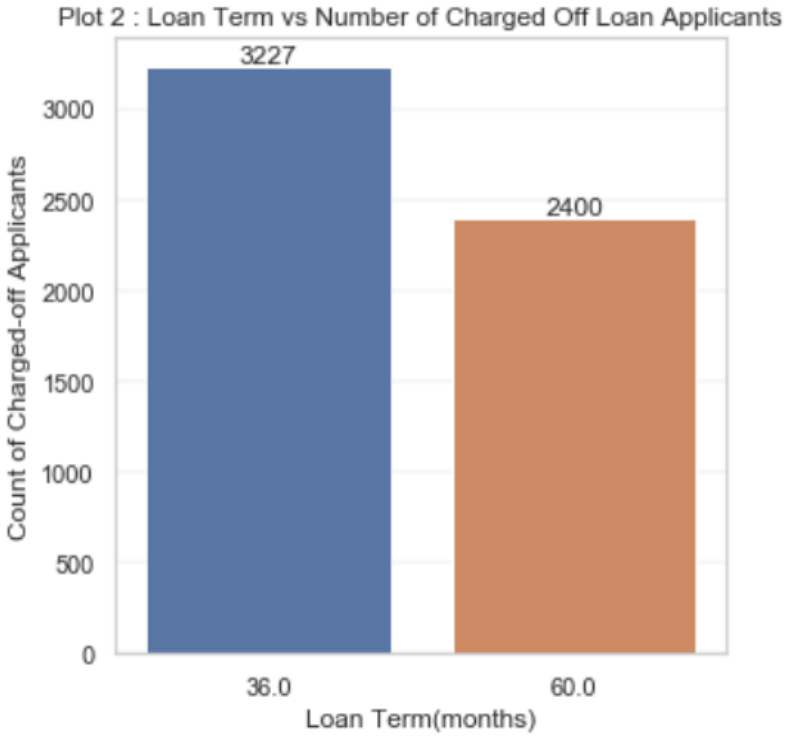
Business Backlog

- A separate data set (loan_status) is constructed from the entire dataset containing only **Charged Off** and **Fully Paid** as *Loan Status*.
- The Charged Off loan applicants are those, whose loan sanction has made a credit loss for the company as they are the defaulters. Where as, the Fully Paid loan applicants are those who have paid the full amount in a stipulated time.
- The rows having *Loan Status* with **Current** values have been discarded here because, no prediction can be made from their data.
- The percentages of Charged off and Fully Paid are computed on the total of the separately created data set.
- It can be observed from the graph that **14.59% of loan applicants have been analysed as defaulter** and have made credit loss to the company. This should be analysed further in order to find out the driving factors.



Analysis – 1

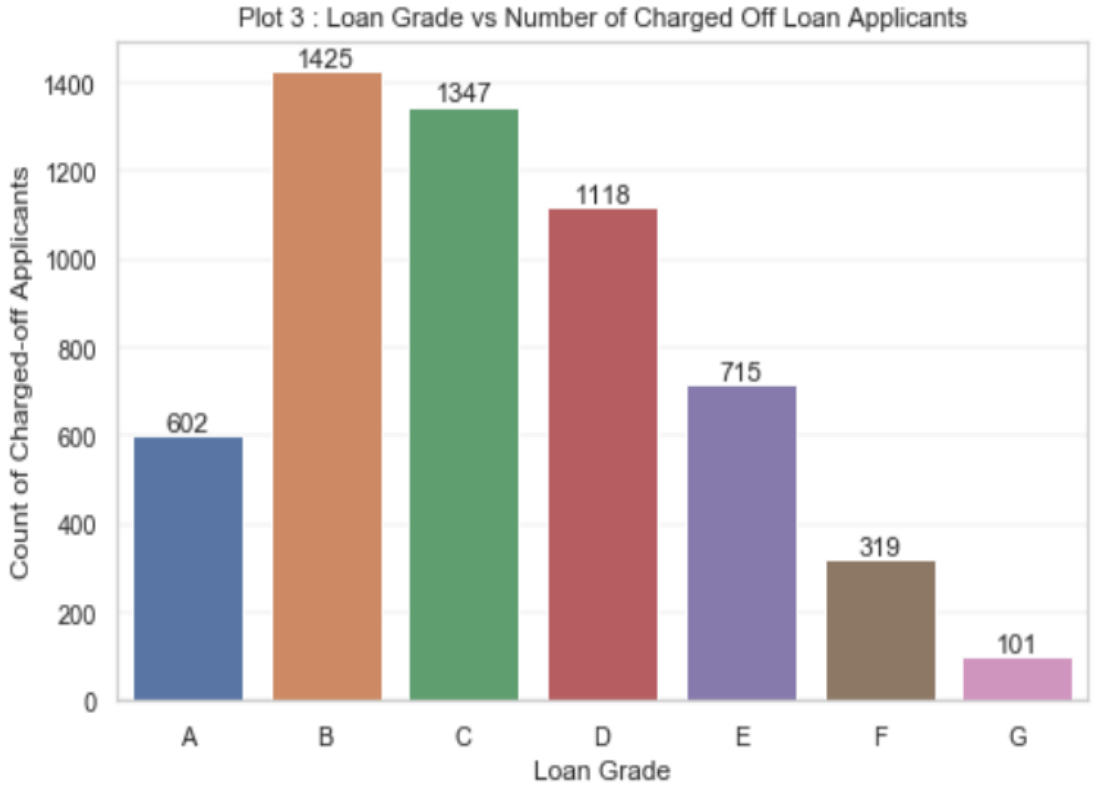
Loan Term vs. Charged Off applicants



Insight:
Number of charged-off loans for 3 years is higher than that of 5 years.

Analysis – 2

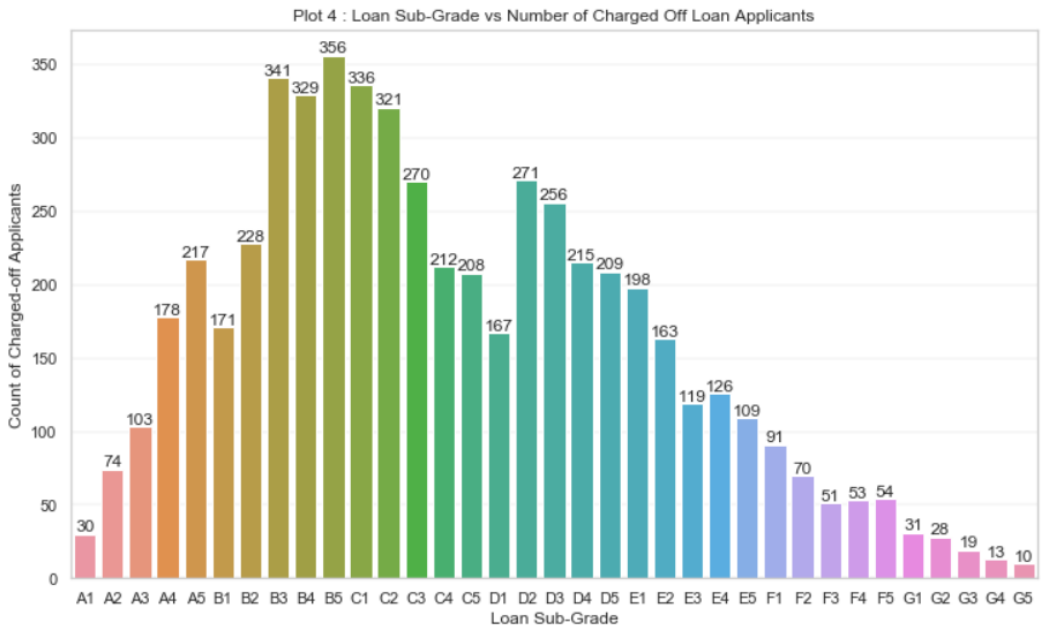
Loan Grade vs. Charged Off Applicants



Insight:
Loan Grades B,C and D contribute to the majority number of charged-off loans.

Analysis – 3

Loan Subgrade vs Charged Loan Applicants

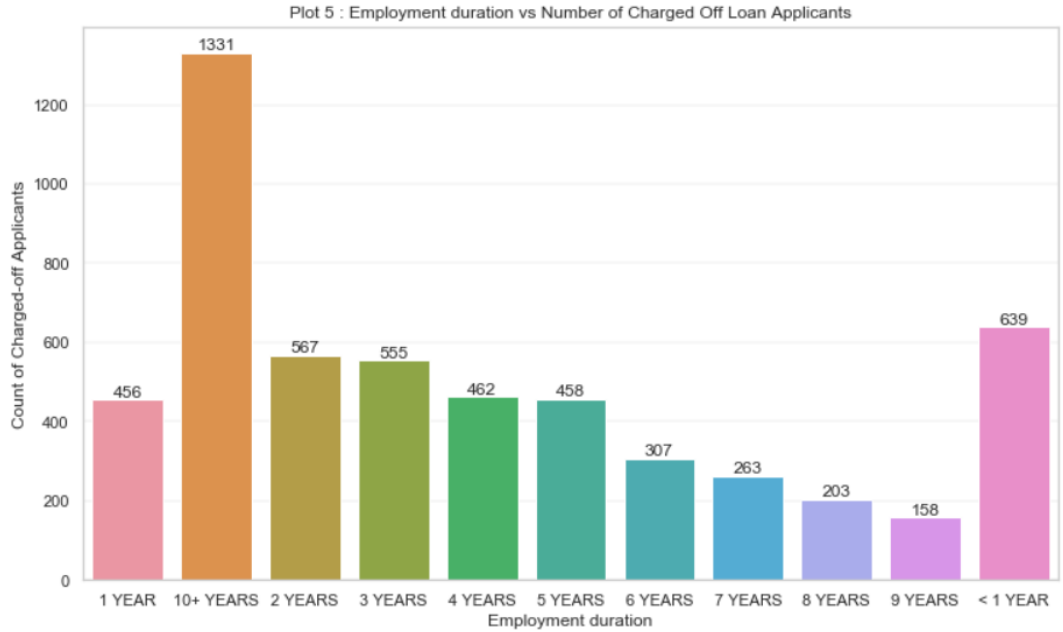


Insight:

The highest number of charged-off loans are in B3~C3 and also D2~D5 sub-grades.

Analysis – 4

Employment Duration vs Charged Loan Applicants

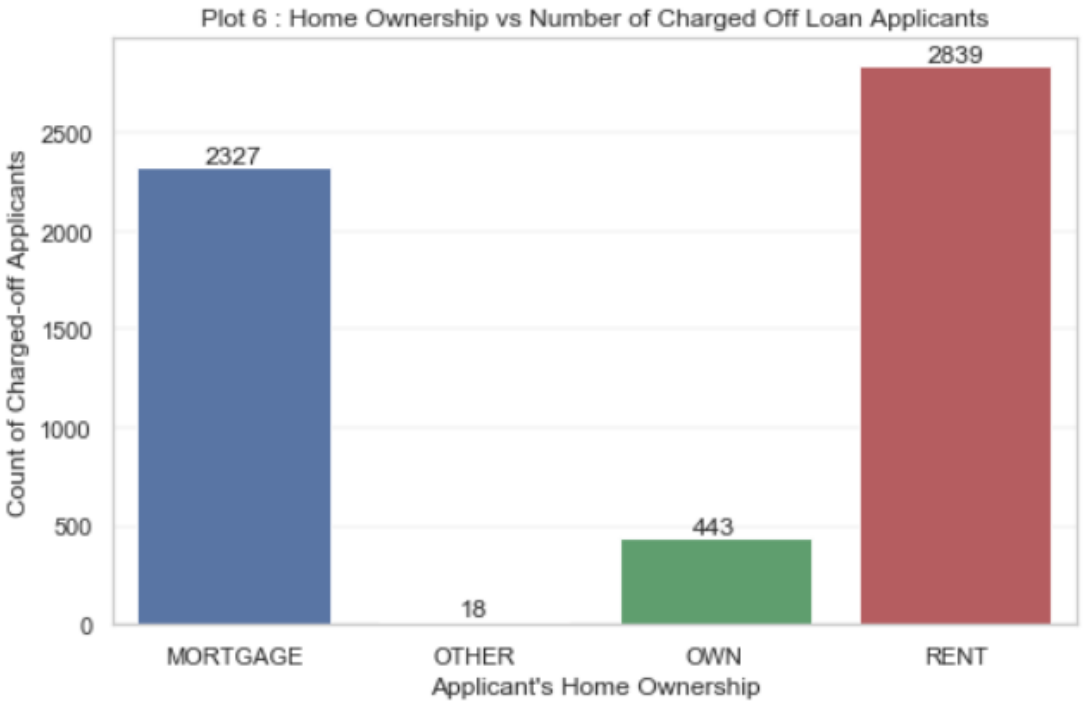


Insight:

Charged-off loans has a decreasing trend with respect to employee employment tenure. However charged-off loans for employment duration less than 1 year and beyond 10 years are considerably high.

Analysis – 5

Home Ownership vs Charged Loan Applicants

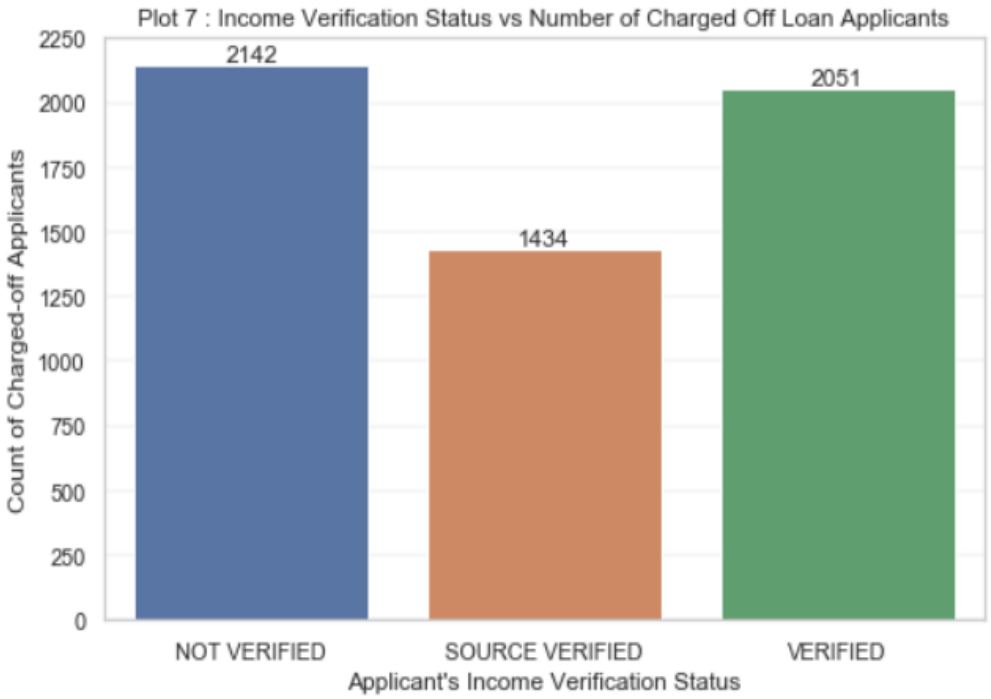


Insight:

Applicant's having Rented or Mortgaged accomodation have higher cases of charged-off loans.

Analysis – 6

Verification Status vs Charged Loan Applicants

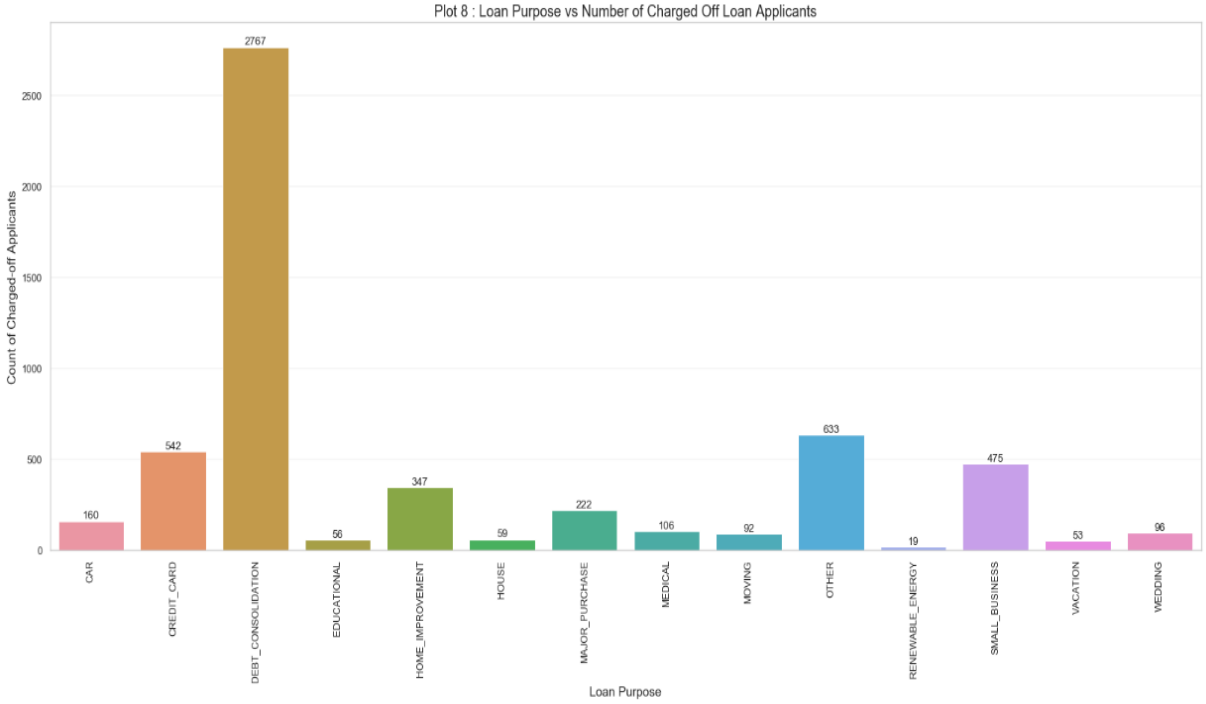


Insight:

Number of charged-off loans for applicants with not verified income is more than that of verifed and source-verified.

Analysis – 7

Loan Purpose vs Charged Loan Applicants

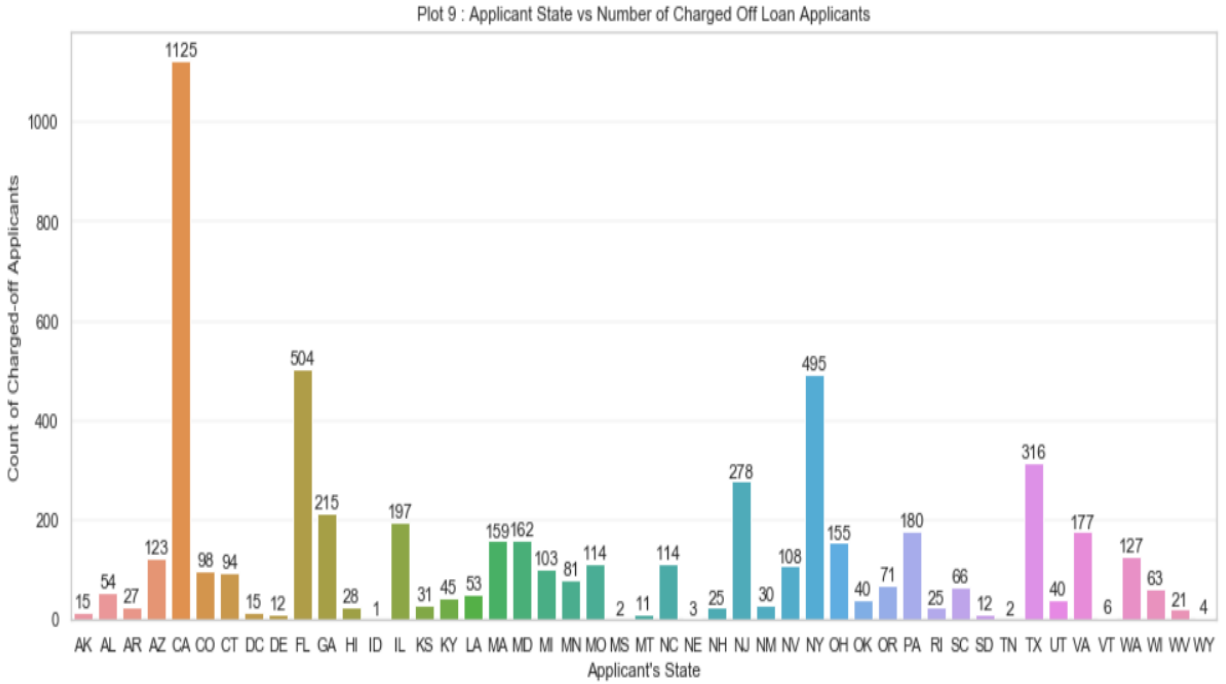


Insight:

Number of charged-off loans is highest for debt consolidation. Also a high number of applicant's for 'Others' category signifies that the data collection method is not adequate, company needs to incorporate more number of purpose's which applicant can select during loan application.

Analysis – 8

Residence State vs Charged Loan Applicants

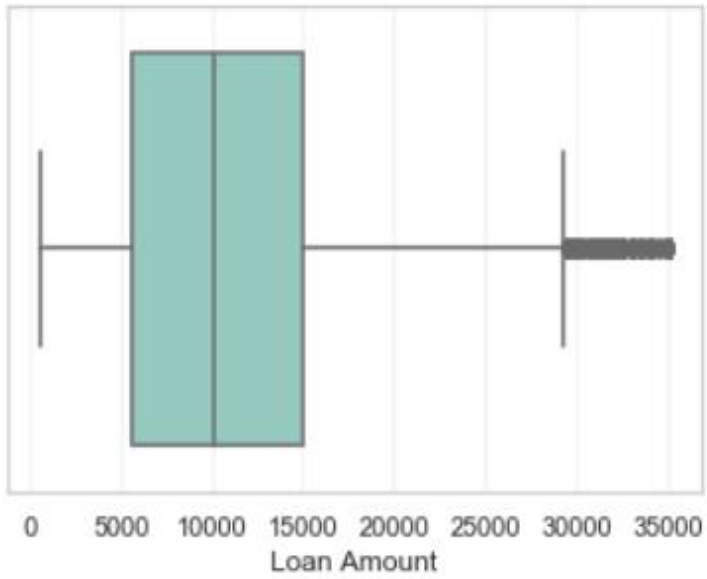


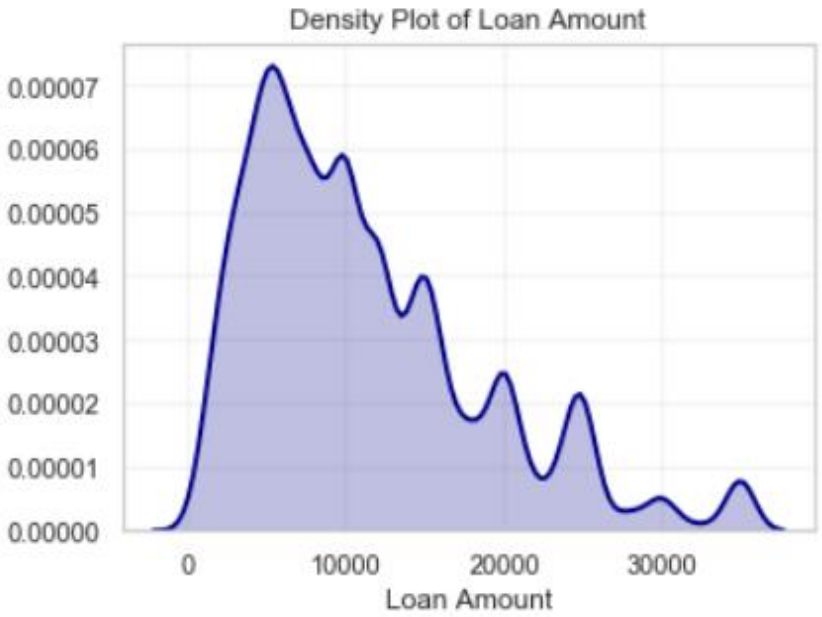
Insight:

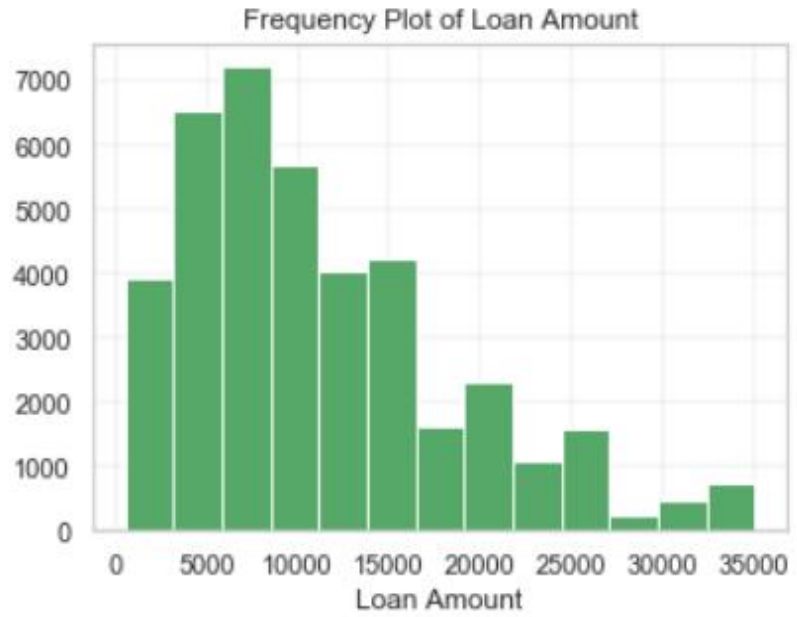
California, Florida, New-York, Texas, New-Jersey are the states having very high number of defaulters.

Analysis – 9

Univariate Analysis : Loan Amount





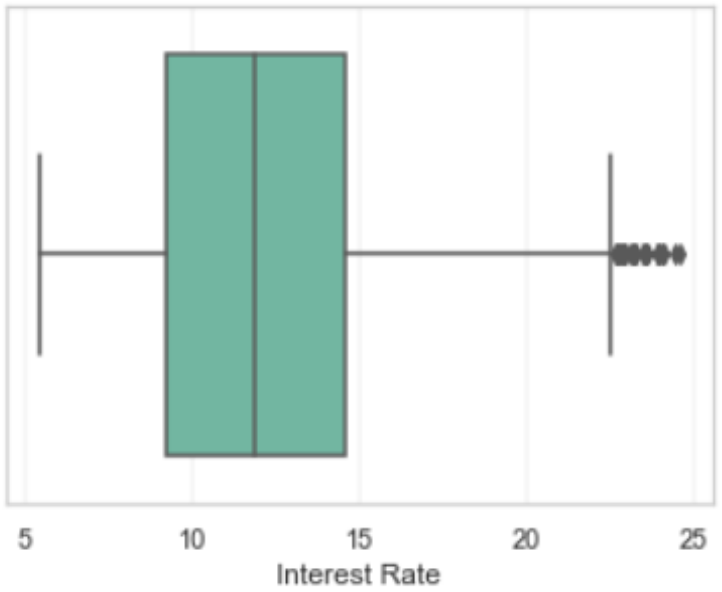


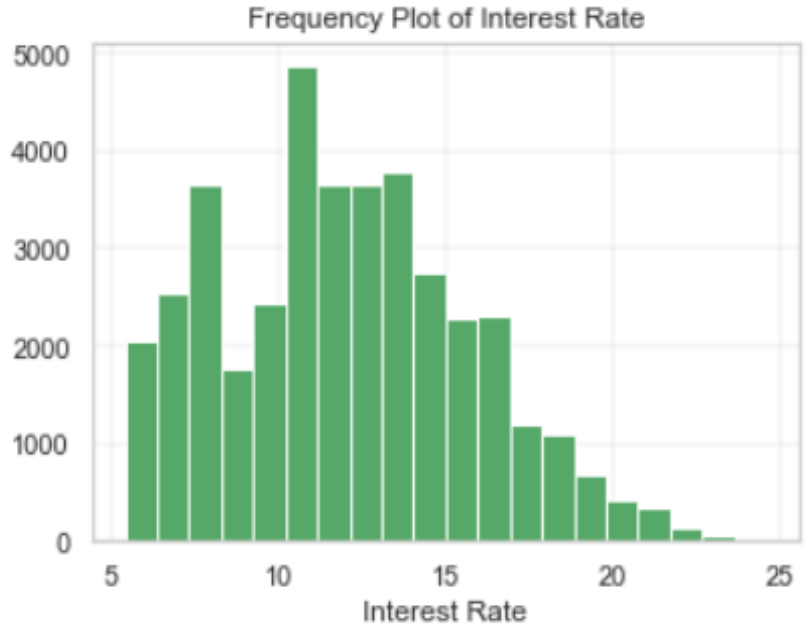
Insight:

From the above box-plot and the other distributions, it is evident that loan amount is approximately a Gaussian distribution and has outliers at the far end.

Analysis – 10

Univariate Analysis : Interest Rate



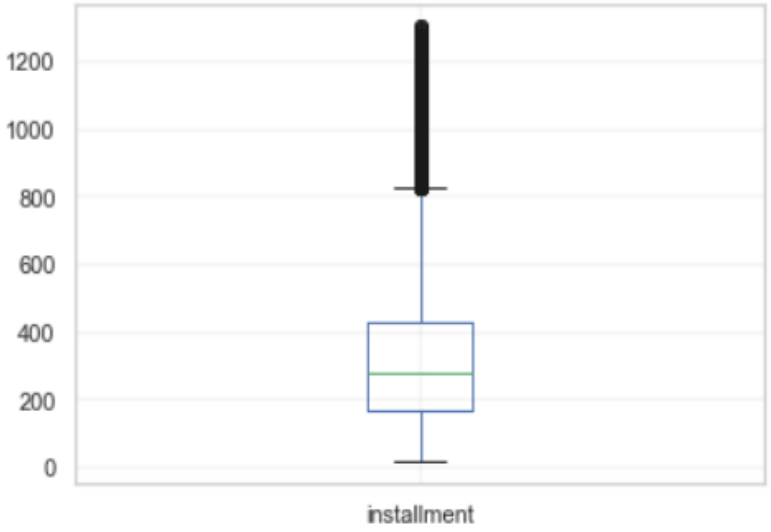


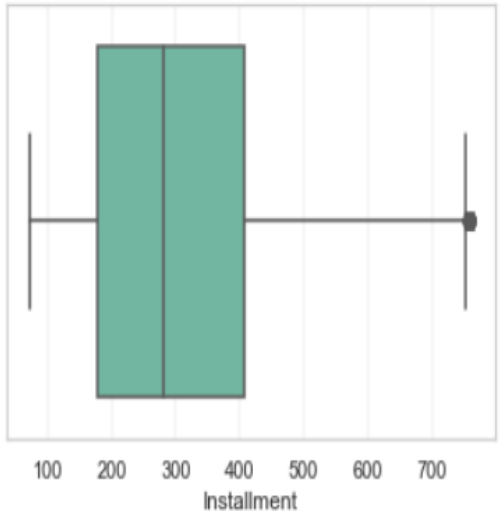
Insight:

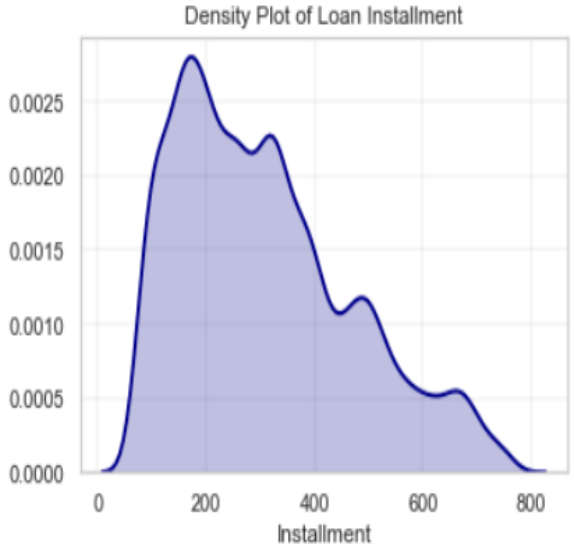
From the box-plot and the other distributions, we can see that the count of loans with interest rate spikes between 7-8% and again between 11-13%.

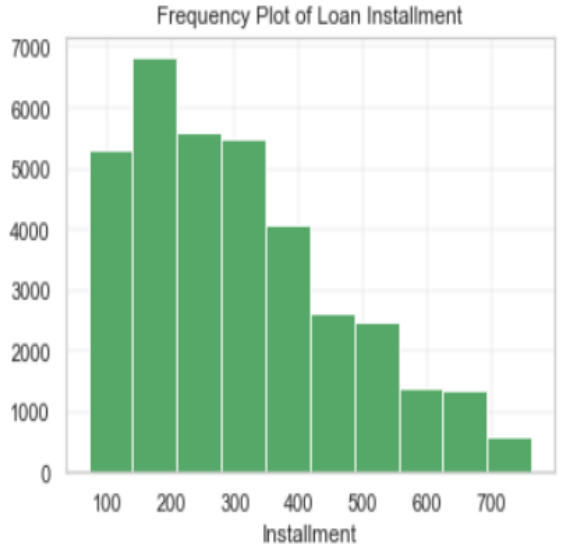
Analysis – 11

Univariate Analysis : Loan Instalment







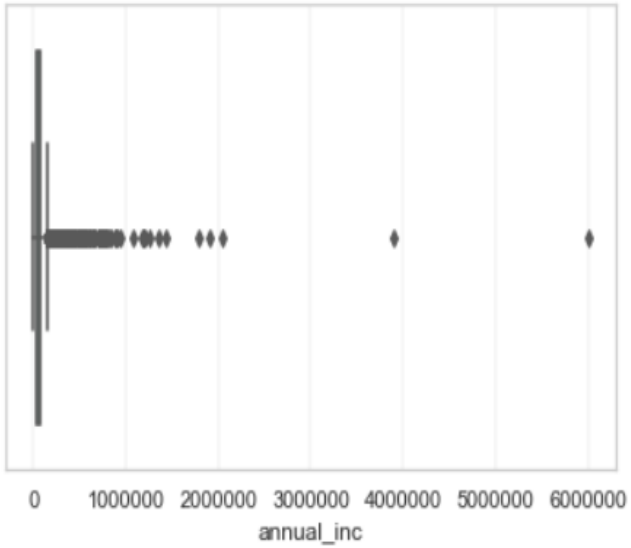


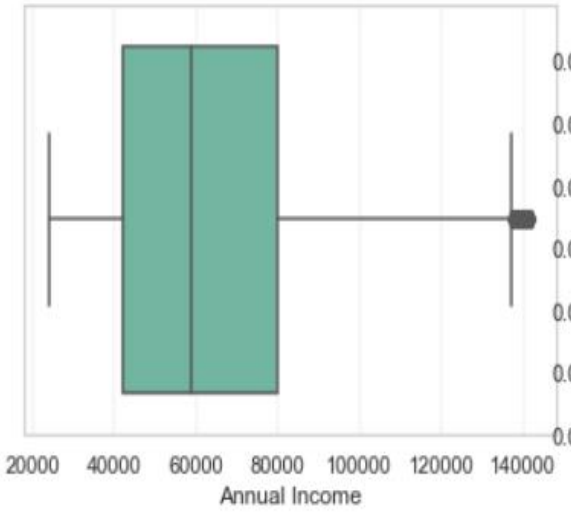
Insight:

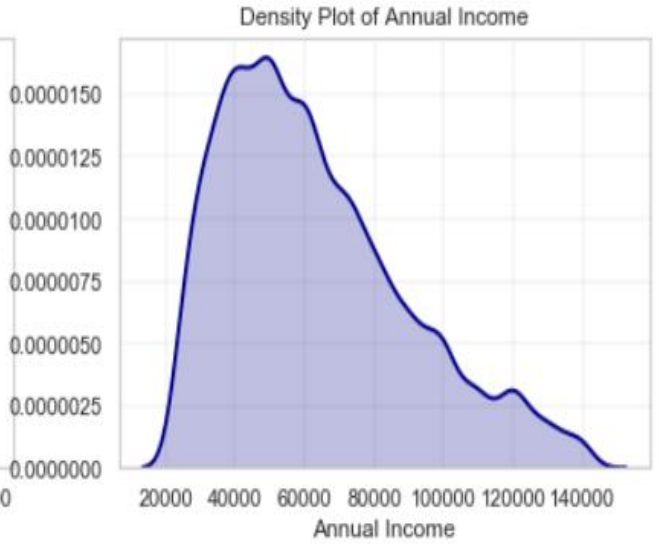
From the left plot, it is clear that the dataset has a considerable high outliers as evident from the mean and median.
Hence, we remove the outliers outside 5% and 95% quartile.
From the box plots, it can be observed that the majority of installment amount is less than 400$.

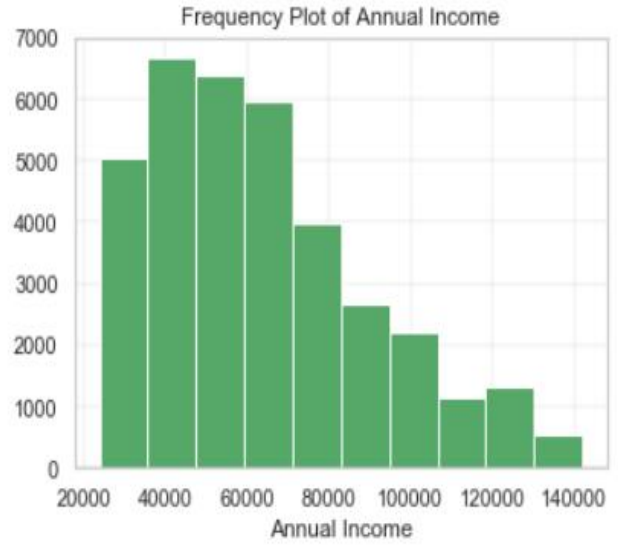
Analysis – 12

Univariate Analysis : Annual Income









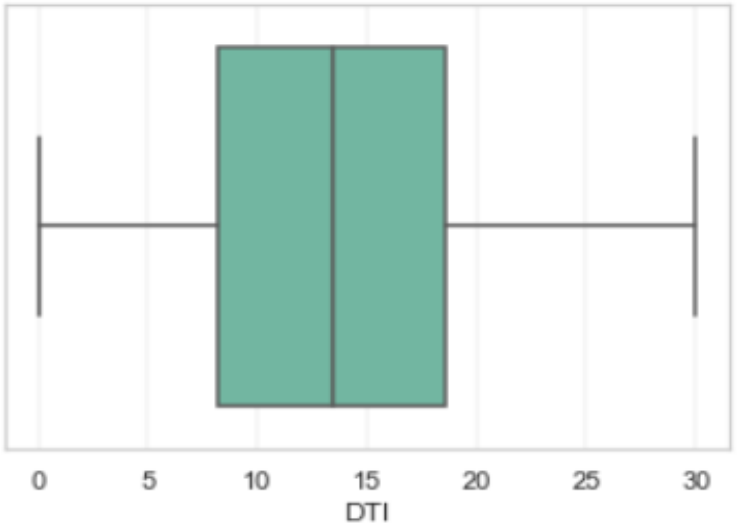
Insight:

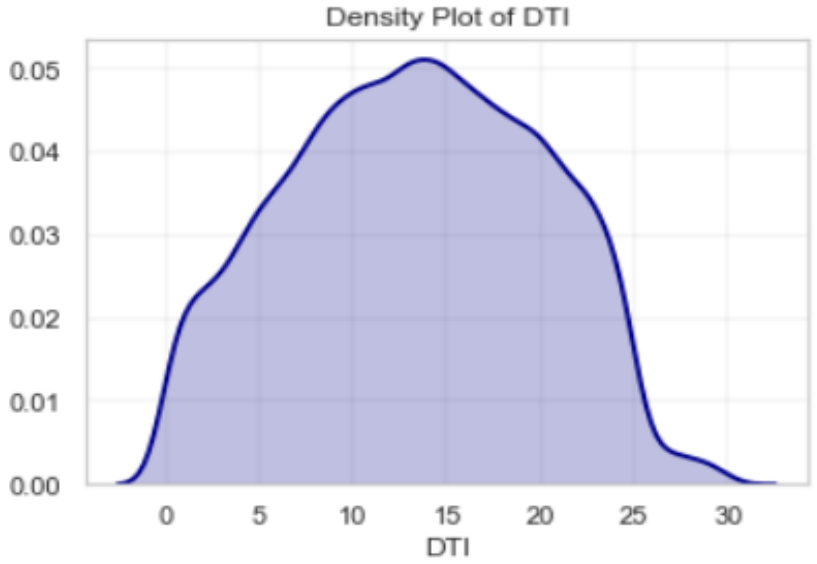
From the left plot, It can be seen from the box-plot that there are significant ouliers on the higher side.Lets restrict the upper bound using 95 percentile.

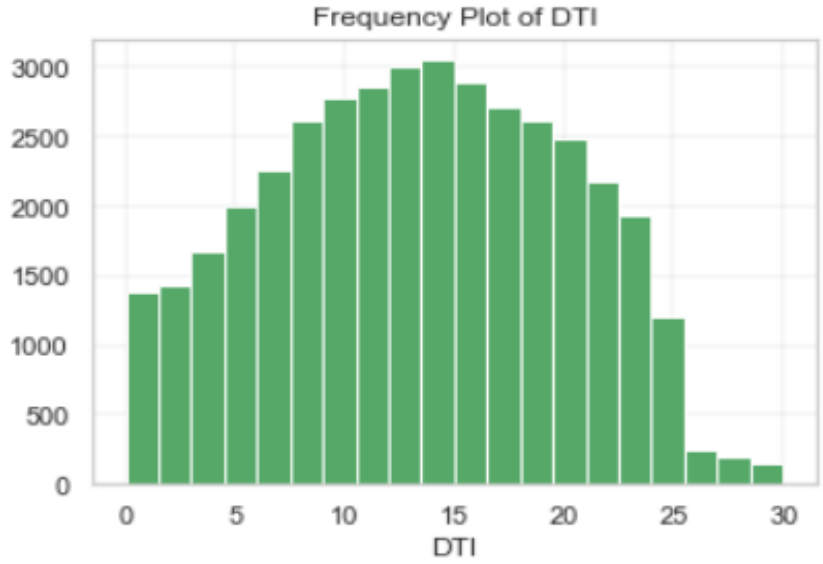
From the box plots, it can be observed that the majority of loan applicants have income in the range of 40,000$-80,000$.

Analysis – 13

Univariate Analysis : DTI Rate





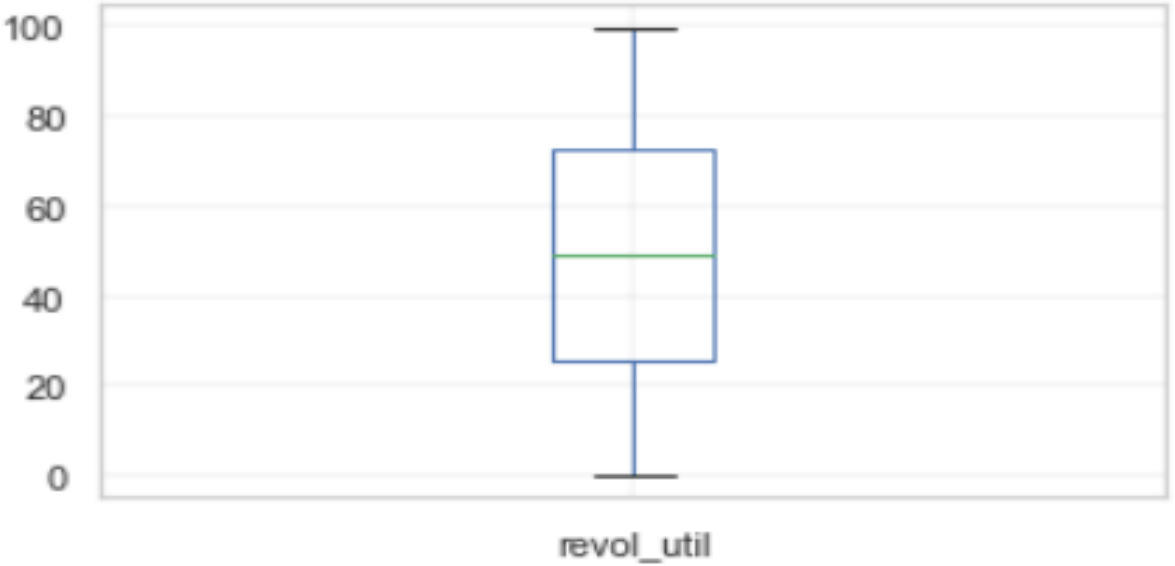


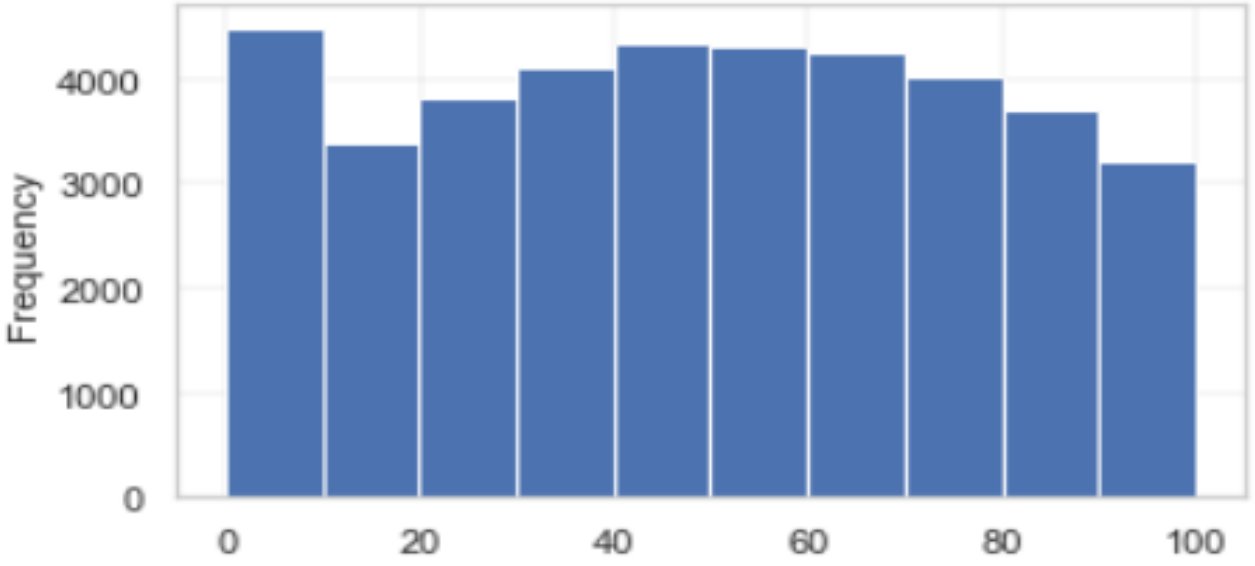
Insight:

From the box-plot and the other distributions, it can be observed that DTI is almost normally distributed with majority between 13-14%.

Analysis – 14

Univariate Analysis : Revolving Utilization Rate





Insight:

From the box-plot and the other distributions, it can be observed that the majority of loan applicants have revolving rate 0-10. For more than 10 revolving rate, the frequency of the loan applicants takes a normal distribution.

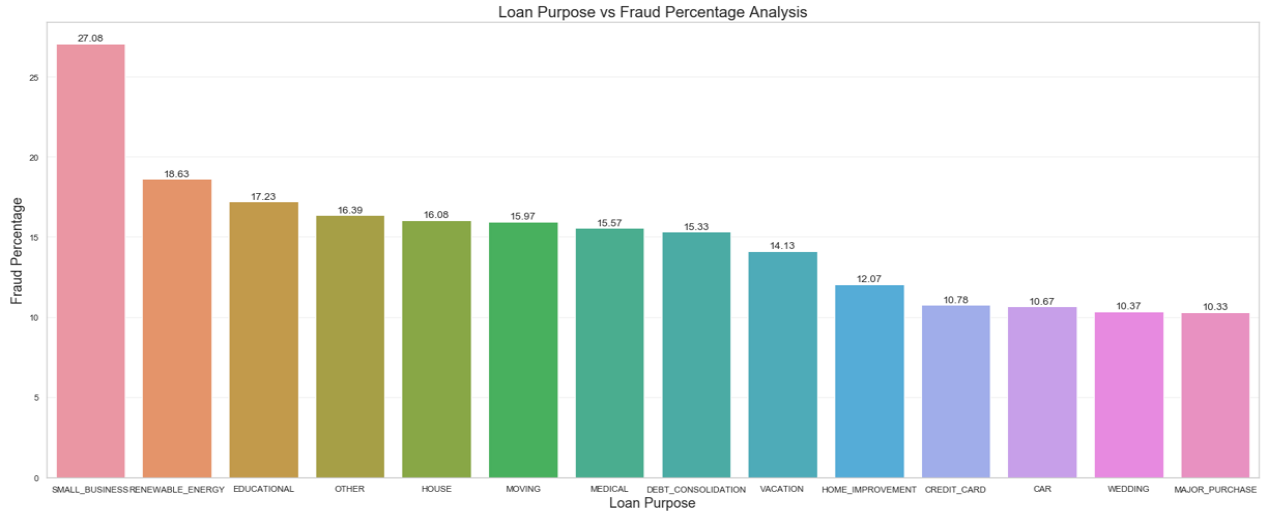
Analysis – 15

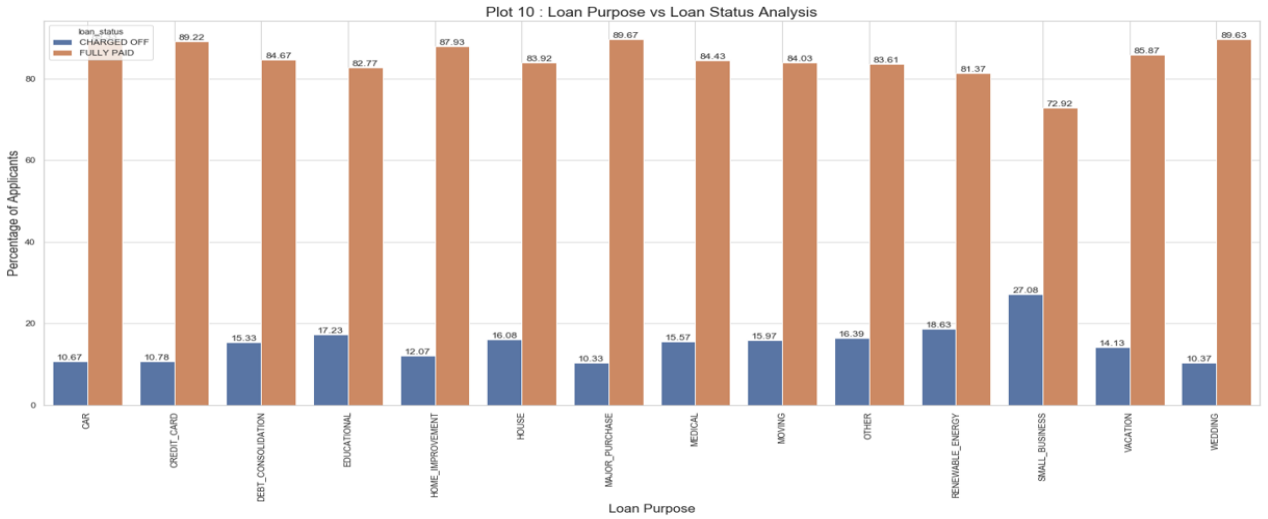
Bivariate Analysis : Loan Purpose vs Loan Status

:

	purpose	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE
11	SMALL_BUSINESS	475	1279	1754	27.080958
10	RENEWABLE_ENERGY	19	83	102	18.627451
3	EDUCATIONAL	56	269	325	17.230769
9	OTHER	633	3229	3862	16.390471
5	HOUSE	59	308	367	16.076294
8	MOVING	92	484	576	15.972222
7	MEDICAL	106	575	681	15.565345
2	DEBT_CONSOLIDATION	2767	15287	18054	15.326243
12	VACATION	53	322	375	14.133333
4	HOME_IMPROVEMENT	347	2528	2875	12.069565
1	CREDIT_CARD	542	4485	5027	10.781778
0	CAR	160	1339	1499	10.673783
13	WEDDING	96	830	926	10.367171
6	MAJOR_PURCHASE	222	1928	2150	10.325581

	purpose	loan_status	Percentage of Applicants
0	CAR	CHARGED OFF	10.67
1	CAR	FULLY PAID	89.33
2	CREDIT_CARD	CHARGED OFF	10.78
3	CREDIT_CARD	FULLY PAID	89.22
4	DEBT_CONSOLIDATION	CHARGED OFF	15.33





Insight:

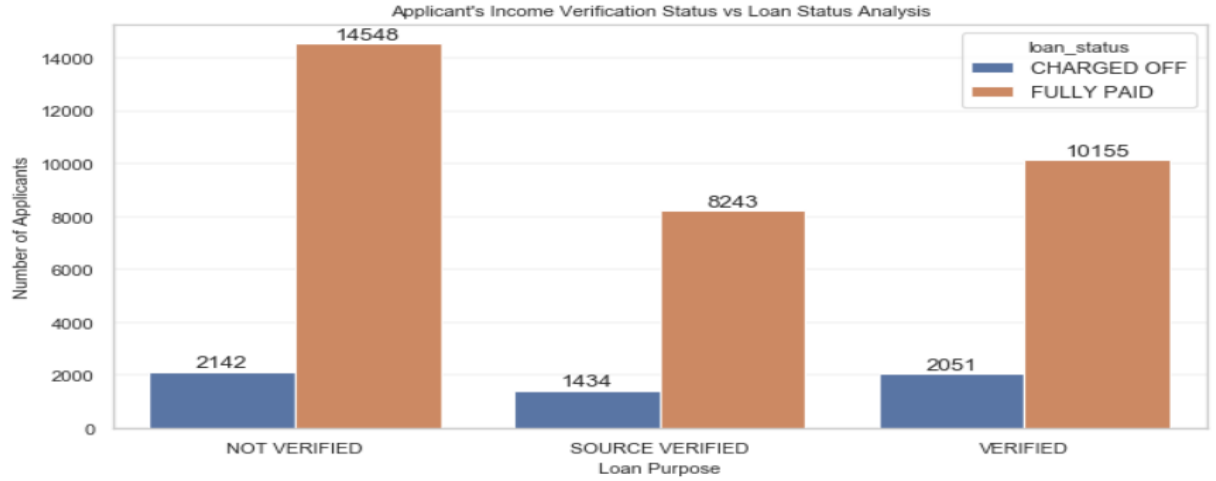
The maximum defaulter percentage is for Small Business Purpose. But the number of applicants is the highest for Debt Consolidation.

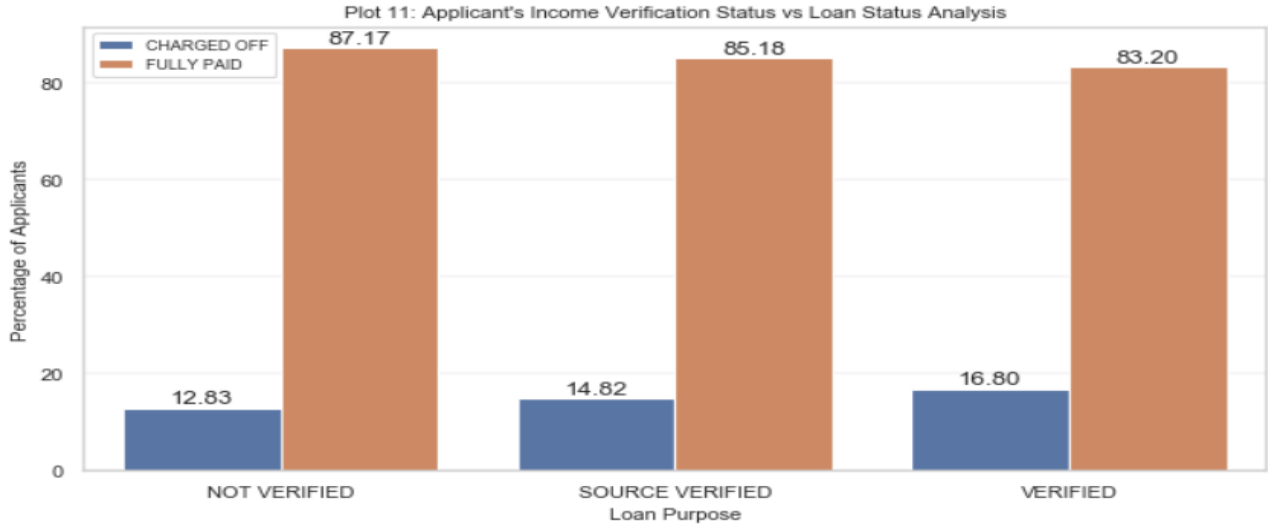
Analysis – 16

Bivariate Analysis : Income Verification vs Loan Status

	verification_status	loan_status	Percentage of Applicants
0	NOT VERIFIED	CHARGED OFF	12.83
1	NOT VERIFIED	FULLY PAID	87.17
2	SOURCE VERIFIED	CHARGED OFF	14.82
3	SOURCE VERIFIED	FULLY PAID	85.18
4	VERIFIED	CHARGED OFF	16.80
5	VERIFIED	FULLY PAID	83.20

	verification_status	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE
2	VERIFIED	2051	10155	12206	16.803212
1	SOURCE VERIFIED	1434	8243	9677	14.818642
0	NOT VERIFIED	2142	14548	16690	12.834032





Insight:

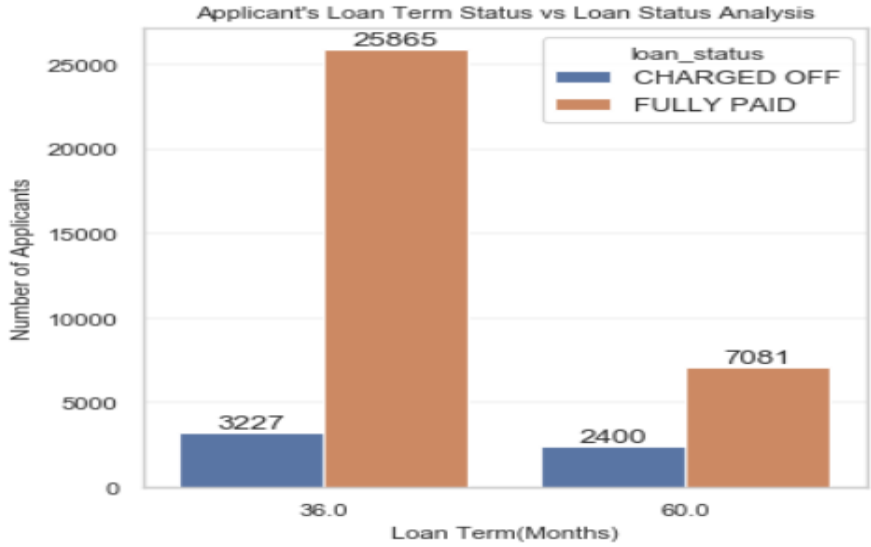
The company has incorporated a system to review the income source of the loan applicant. From the table it is clear that the highest number of applications received not verified. However, from the plot 11, it is evident that applications that have the income verified or source verified have a higher chance of leading to credit loss. LC must appraise and analyze the system implemented for verification.

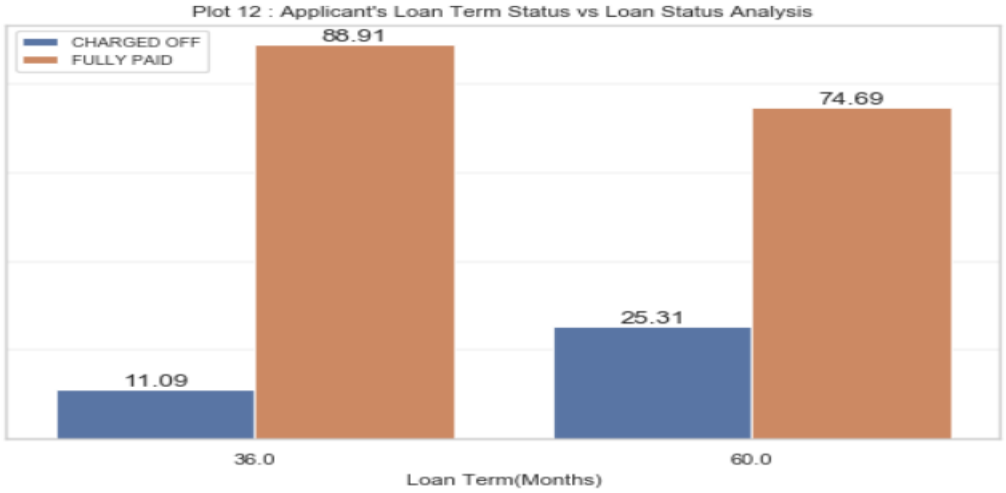
Analysis – 17

Bivariate Analysis : Loan Term vs Loan Status

	term	loan_status	Number of Applicants
0	36.0	CHARGED OFF	3227
1	36.0	FULLY PAID	25865
2	60.0	CHARGED OFF	2400
3	60.0	FULLY PAID	7081

	term	loan_status	Percentage of Applicants
0	36.0	CHARGED OFF	11.09
1	36.0	FULLY PAID	88.91
2	60.0	CHARGED OFF	25.31
3	60.0	FULLY PAID	74.69





Insight:

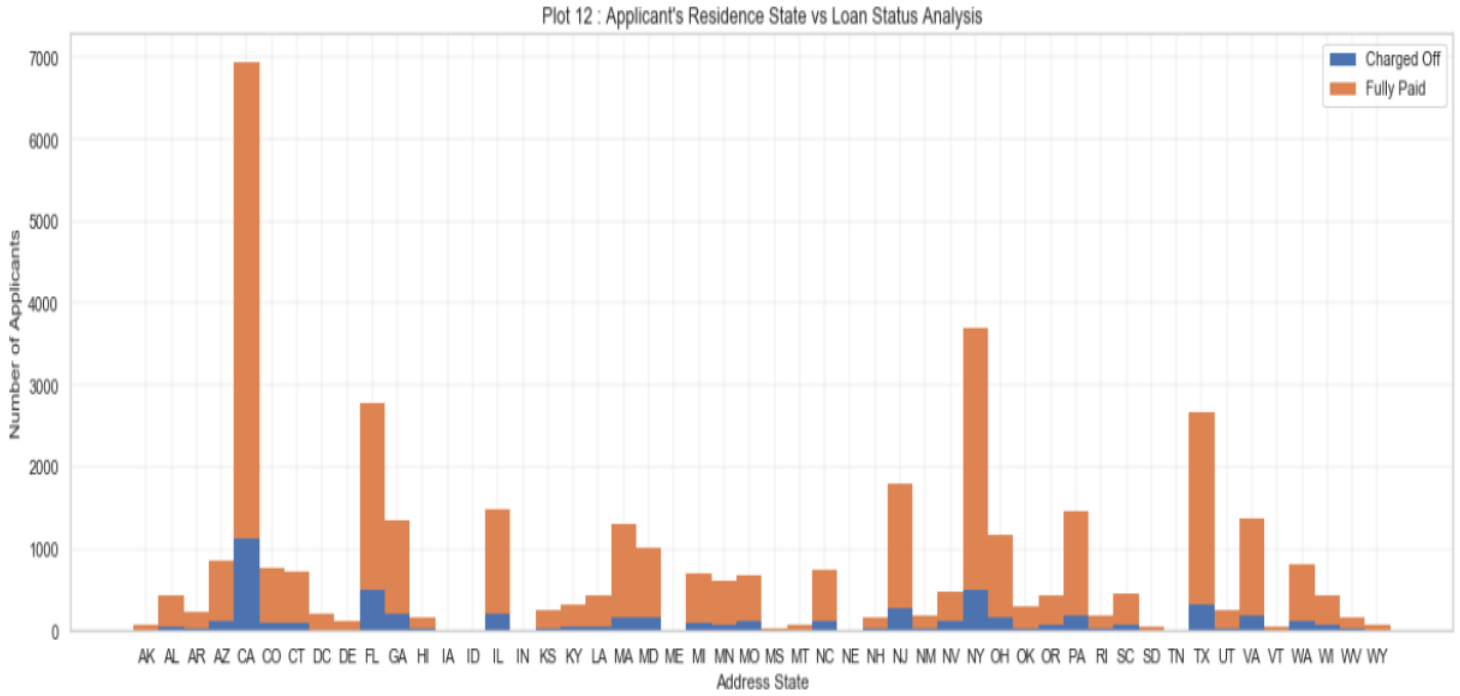
Majority of loans are issue for a duration of 36 months. However, from the plot it is clear that a loans issued for a term of 60 months has a significantly higher chance of resulting in credit loss.

Analysis – 18

Bivariate Analysis : Residence State vs Loan Status

	addr_state	loan_status	Number of Applicants
0	AK	CHARGED OFF	15
1	AK	FULLY PAID	63
2	AL	CHARGED OFF	54
3	AL	FULLY PAID	381
4	AR	CHARGED OFF	27

	addr_state	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE
32	NV	108.0	371.0	479.0	22.546973
40	SD	12.0	50.0	62.0	19.354839
0	AK	15.0	63.0	78.0	19.230769
9	FL	504.0	2277.0	2781.0	18.122977
24	MO	114.0	556.0	670.0	17.014925



Insight:

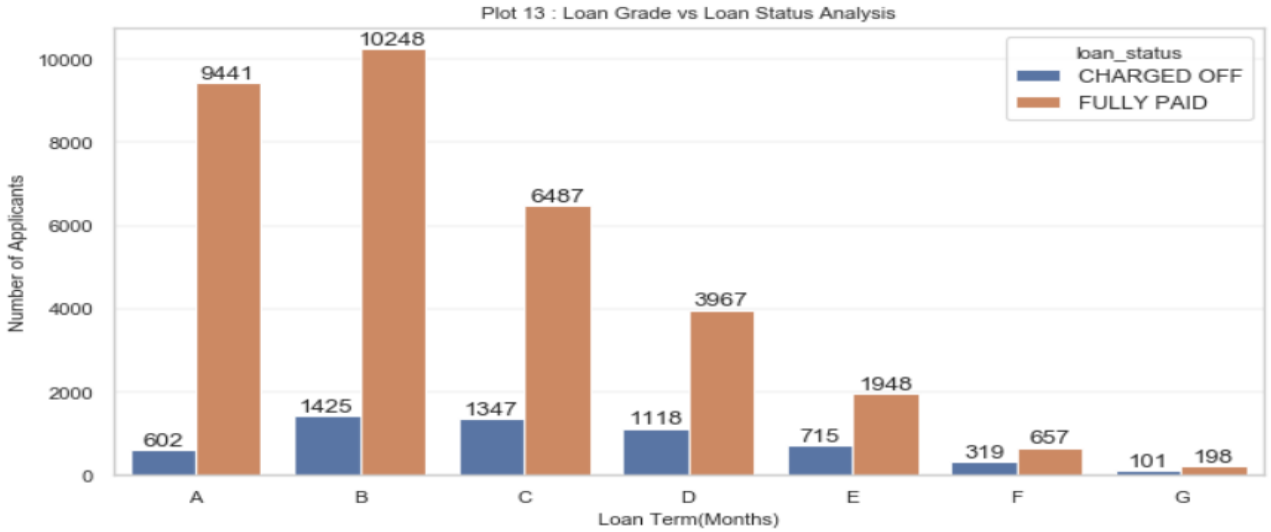
From the table, it is evident that borrowers belonging to Nevraska contribute to the highest default percentage(**22.5%**). The highest number of loan application are received from Canada, Florida, New-York and Texas. Their respective default rates are 16.1%,13.4%,18.1% and 11.9%.

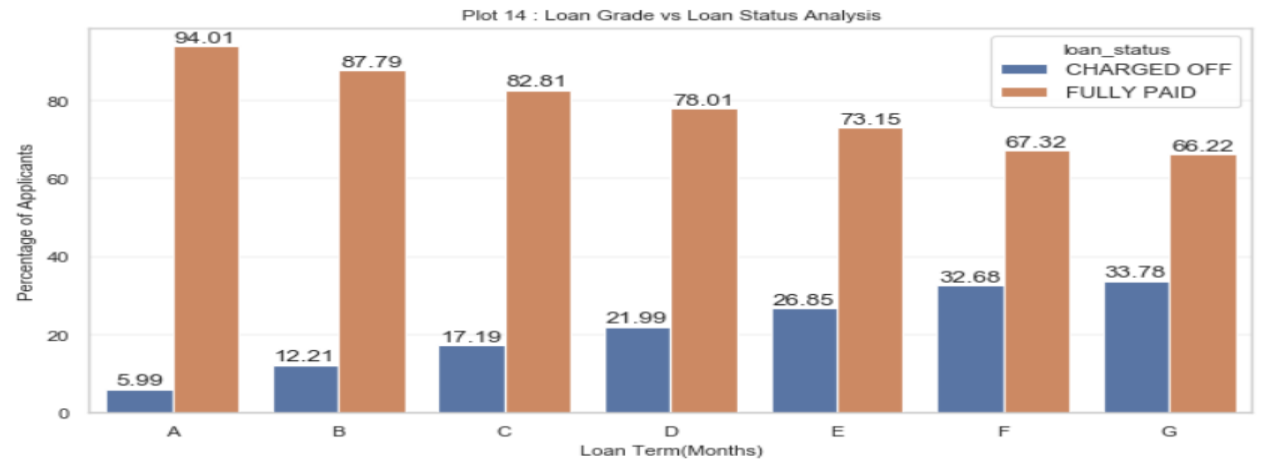
Analysis – 19

Bivariate Analysis : Loan Grade vs Loan Status

	grade	loan_status	Number of Applicants
0	A	CHARGED OFF	602
1	A	FULLY PAID	9441
2	B	CHARGED OFF	1425
3	B	FULLY PAID	10248
4	C	CHARGED OFF	1347
5	C	FULLY PAID	6487
6	D	CHARGED OFF	1118
7	D	FULLY PAID	3967
8	E	CHARGED OFF	715
9	E	FULLY PAID	1948
10	F	CHARGED OFF	319
11	F	FULLY PAID	657
12	G	CHARGED OFF	101
13	G	FULLY PAID	198

	grade	loan_status	Percentage of Applicants
0	A	CHARGED OFF	5.99
1	A	FULLY PAID	94.01
2	B	CHARGED OFF	12.21
3	B	FULLY PAID	87.79
4	C	CHARGED OFF	17.19
5	C	FULLY PAID	82.81
6	D	CHARGED OFF	21.99
7	D	FULLY PAID	78.01
8	E	CHARGED OFF	26.85
9	E	FULLY PAID	73.15
10	F	CHARGED OFF	32.68
11	F	FULLY PAID	67.32
12	G	CHARGED OFF	33.78
13	G	FULLY PAID	66.22





Insight:

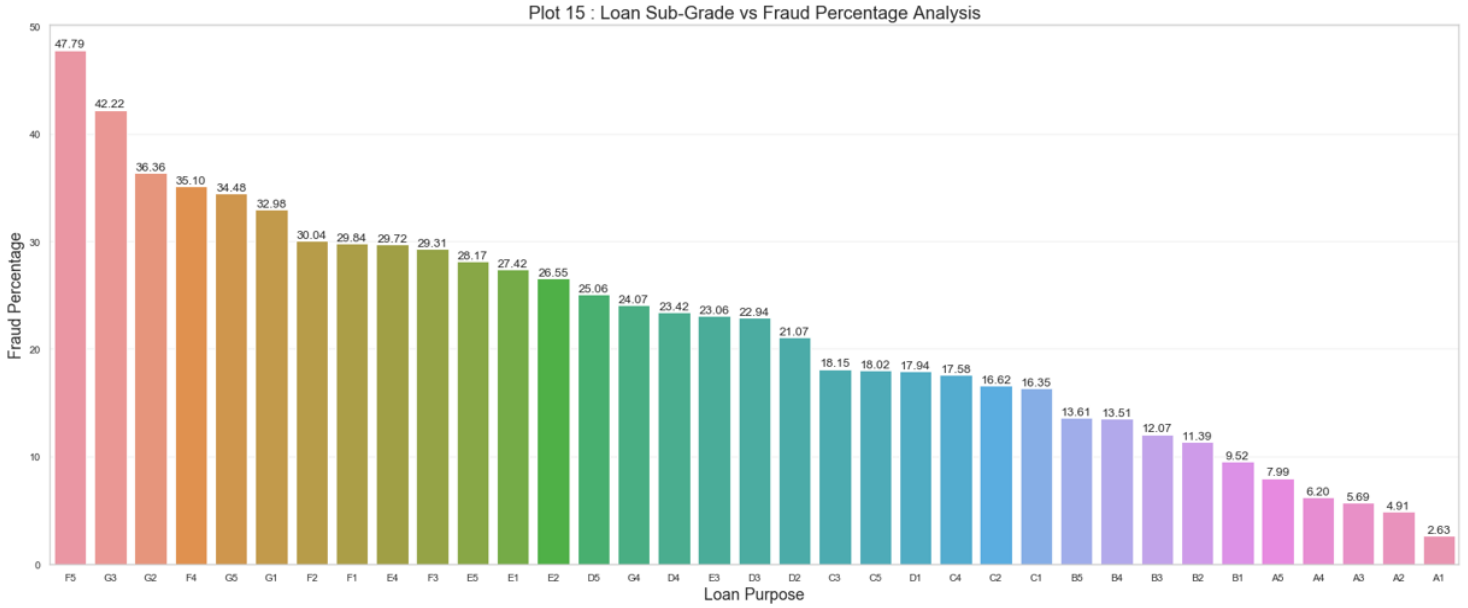
From Plot 13, it can be observed that a high frequency of loans belongs to grade A,B,C and D. However when we look at the percentage of default(Plot 14), we could see that there is a clear trend of increase in the percentage of defaulters from grade A to G.(With A having lowest percentage of default and G having highest **33.78%**)

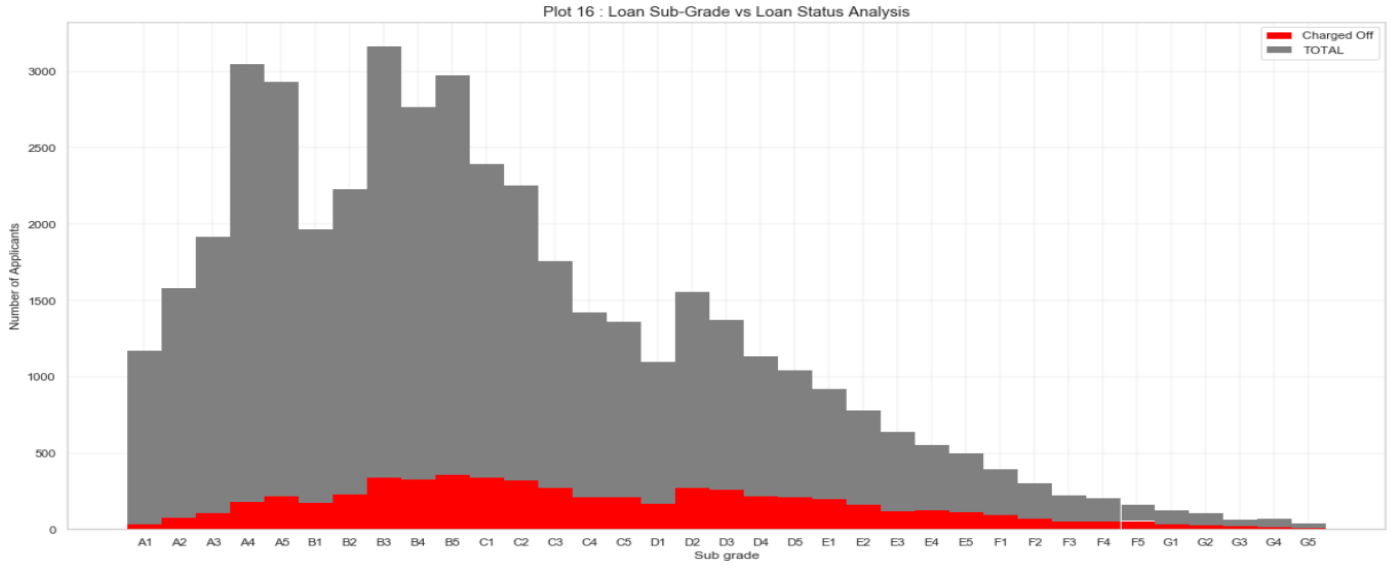
Analysis – 20

Bivariate Analysis : Loan Sub-Grade vs Loan Status

	sub_grade	loan_status	Number of Applicants
0	A1	CHARGED OFF	30
1	A1	FULLY PAID	1109
2	A2	CHARGED OFF	74
3	A2	FULLY PAID	1433
4	A3	CHARGED OFF	103

	sub_grade	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE
29	F5	54	59	113	47.787611
32	G3	19	26	45	42.222222
31	G2	28	49	77	36.363636
28	F4	53	98	151	35.099338
34	G5	10	19	29	34.482759
30	G1	31	63	94	32.978723
26	F2	70	163	233	30.042918
25	F1	91	214	305	29.836066
23	E4	126	298	424	29.716981
27	F3	51	123	174	29.310345
24	E5	109	278	387	28.165375
20	E1	198	524	722	27.423823





Insight:

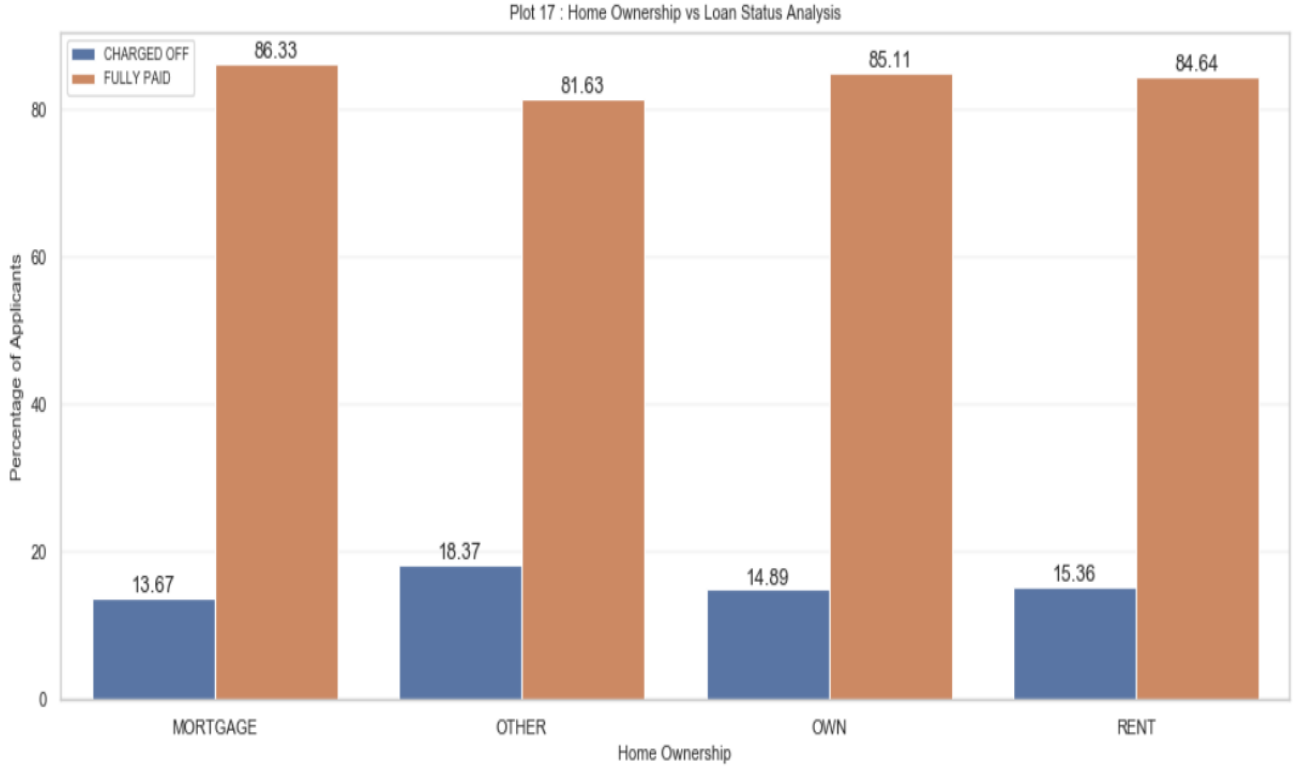
On drilling down to the sub-grade level we could observed from Plot.15 that borrowers with sub-grades E1~G5 have the highest chance of resulting in credit loss[** With F5-48% , G3-42%,G2-36% representing the top 3 likely defaulters.**]

Analysis – 21

Bivariate Analysis : Home Ownership vs Loan Status

	home_ownership	loan_status	Number of Applicants
0	MORTGAGE	CHARGED OFF	2327
1	MORTGAGE	FULLY PAID	14692
3	OTHER	CHARGED OFF	18
4	OTHER	FULLY PAID	80
5	OWN	CHARGED OFF	443
6	OWN	FULLY PAID	2532
7	RENT	CHARGED OFF	2839
8	RENT	FULLY PAID	15641

	home_ownership	loan_status	Percentage of Applicants
0	MORTGAGE	CHARGED OFF	13.67
1	MORTGAGE	FULLY PAID	86.33
2	OTHER	CHARGED OFF	18.37
3	OTHER	FULLY PAID	81.63
4	OWN	CHARGED OFF	14.89
5	OWN	FULLY PAID	85.11
6	RENT	CHARGED OFF	15.36
7	RENT	FULLY PAID	84.64



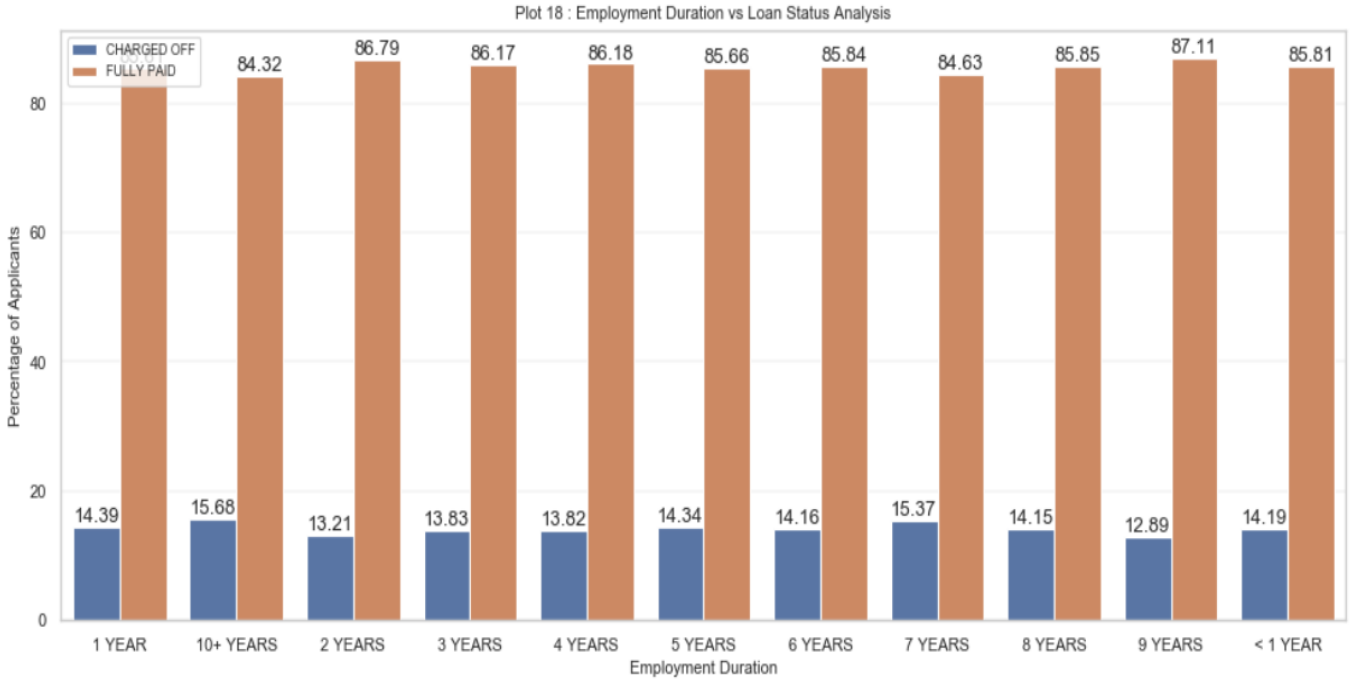
Insight:

From the above plot it is clear that applicants who states home ownership as other have higher chances of defaulting the loan payment. From the table it is also clear that majority of the loan applicants state rent or mortgage as home-ownership, from which 15.36% and 13.67% result in credit loss respectively.

Analysis – 22

Bivariate Analysis : Employment Length vs Loan Status

	emp_length	loan_status	Percentage of Applicants
0	1 YEAR	CHARGED OFF	14.39
1	1 YEAR	FULLY PAID	85.61
2	10+ YEARS	CHARGED OFF	15.68
3	10+ YEARS	FULLY PAID	84.32
4	2 YEARS	CHARGED OFF	13.21
5	2 YEARS	FULLY PAID	86.79
6	3 YEARS	CHARGED OFF	13.83
7	3 YEARS	FULLY PAID	86.17
8	4 YEARS	CHARGED OFF	13.82
9	4 YEARS	FULLY PAID	86.18
10	5 YEARS	CHARGED OFF	14.34
11	5 YEARS	FULLY PAID	85.66



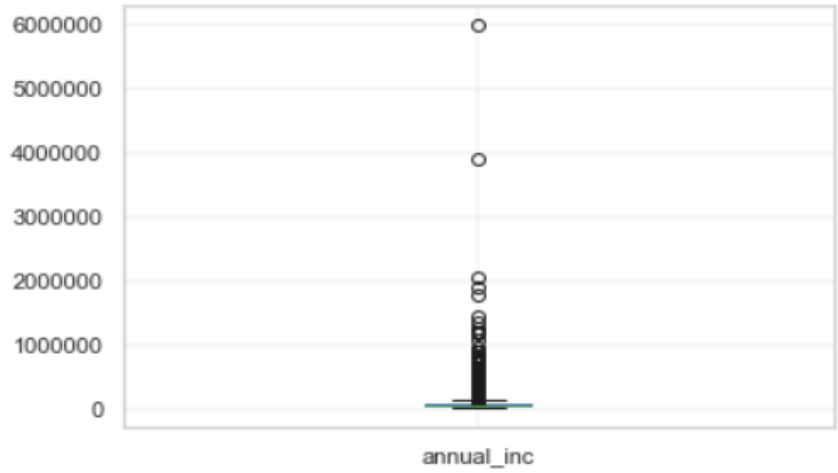
Insight:

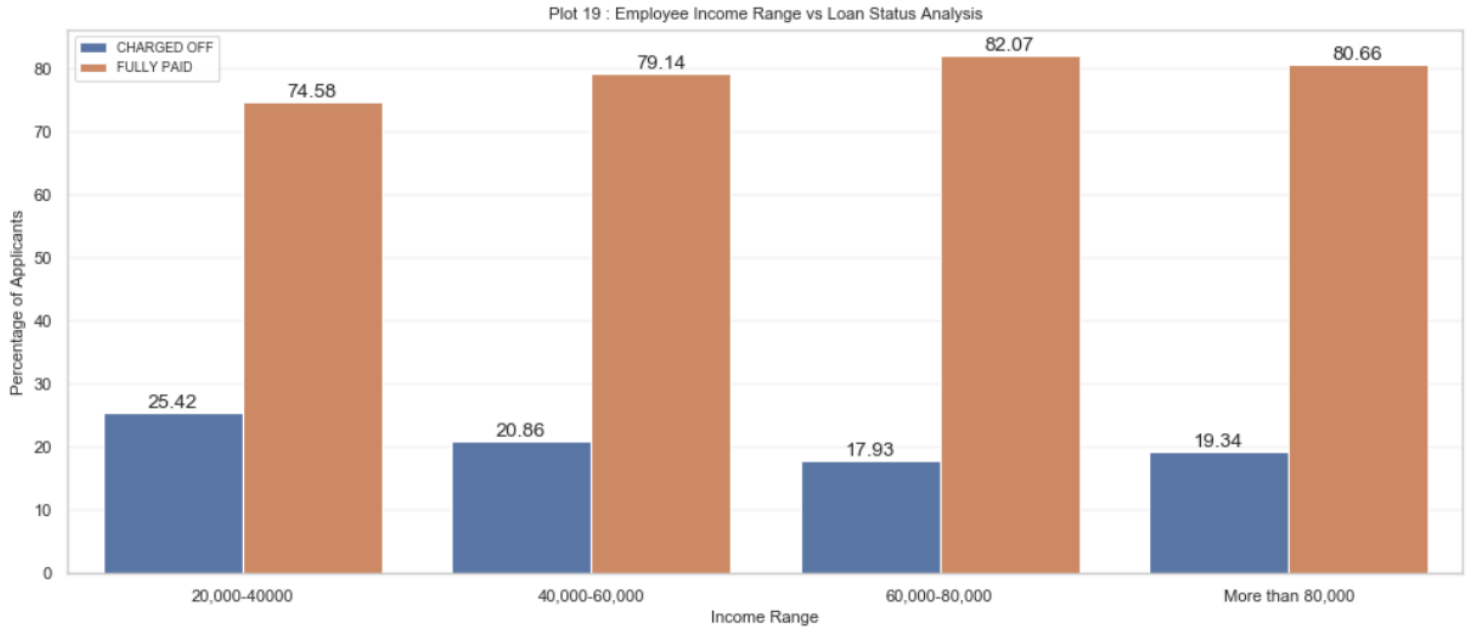
From the table it can be observed that the applicants working more than 10 years apply for loan the most number of times and are likely to default the most too. The applicants working less than 1 year is likely to default more than the rest.

Analysis – 23

Bivariate Analysis : Annual Income vs Loan Status

	Income_range	loan_status	Percentage of Applicants
0	20,000-40000	CHARGED OFF	25.42
1	20,000-40000	FULLY PAID	74.58
2	40,000-60,000	CHARGED OFF	20.86
3	40,000-60,000	FULLY PAID	79.14
4	60,000-80,000	CHARGED OFF	17.93
5	60,000-80,000	FULLY PAID	82.07
6	More than 80,000	CHARGED OFF	19.34
7	More than 80,000	FULLY PAID	80.66





Insight:

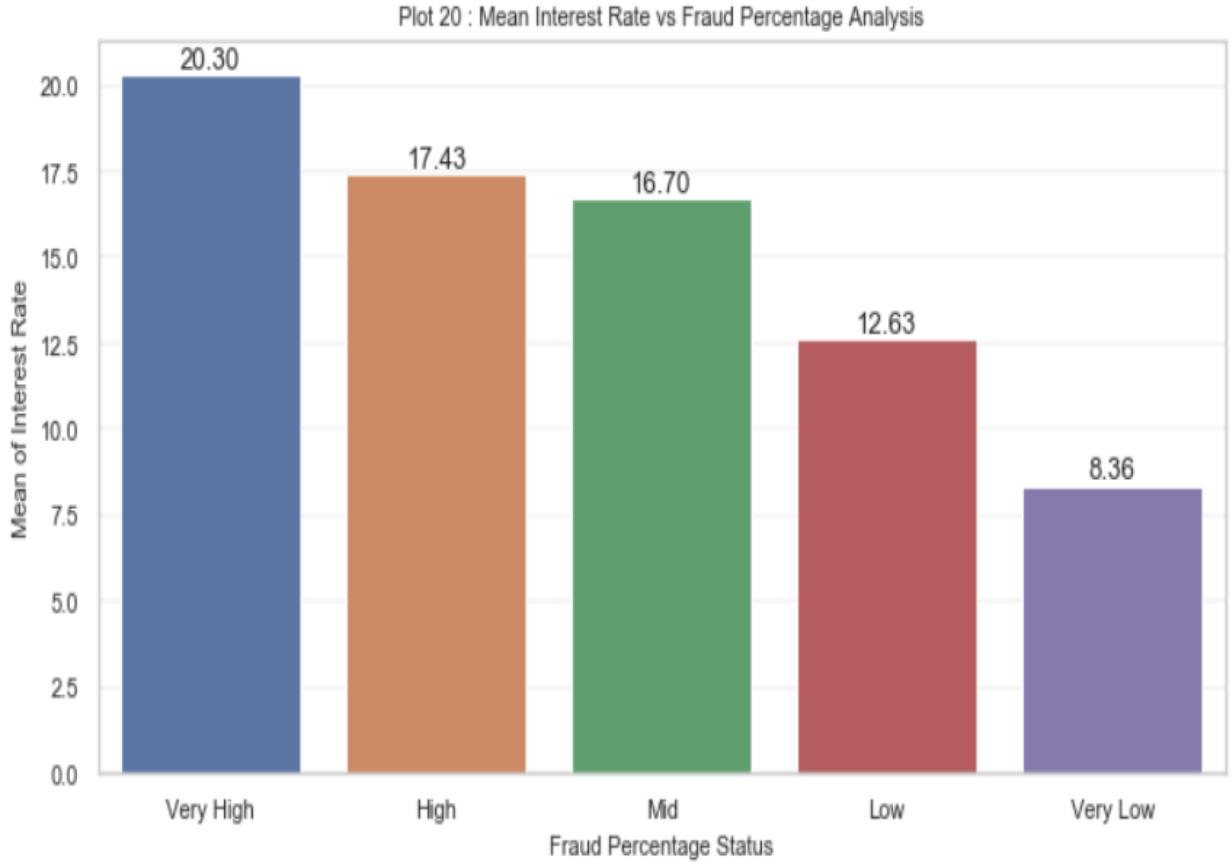
It is observed that as the annual income range increases the percentage of charged-off loans is also decreasing. Highest percentage of charged-off loans lies for income range 20000-40000$.There are some exceptions in the range of more than 80000 since we could see that the fraud percentage is slightly more than the range of 60000-80000.

Analysis – 24

Bivariate Analysis : Interest Rate vs Loan Status

	int_rate	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE	Status
351	21.64	2.0	1.0	3.0	66.666667	Very High
365	23.52	4.0	2.0	6.0	66.666667	Very High
174	13.93	2.0	1.0	3.0	66.666667	Very High
363	23.13	5.0	3.0	8.0	62.500000	Very High
353	21.74	18.0	11.0	29.0	62.068966	Very High

	Fraud Percentage Status	Mean of int_rate
3	Very High	20.296000
0	High	17.432571
2	Mid	16.703761
1	Low	12.626364
4	Very Low	8.356389



Insight:

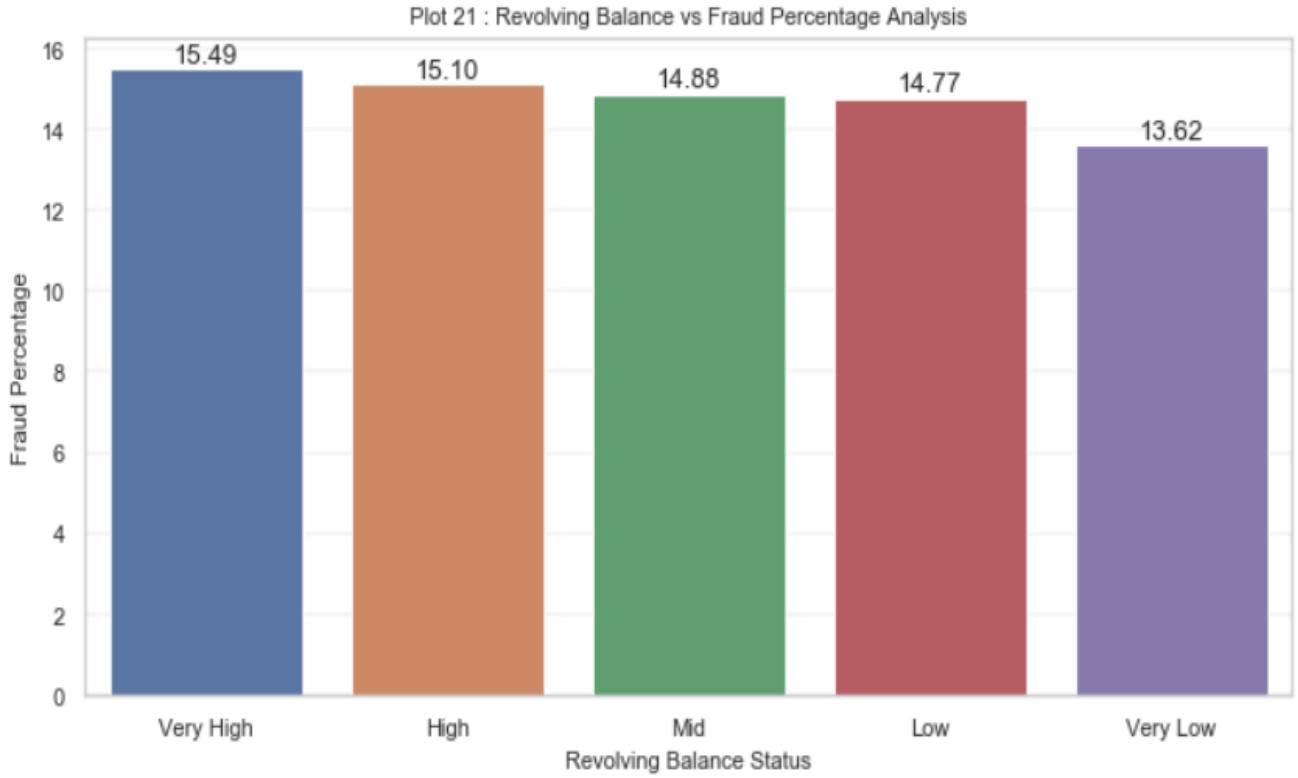
It can be inferred from the above plot that, with increase in interest rate, the percentage of fraud also increases. More likely there will be chances of charged-off loans when interest rate is high.

Analysis – 25

Bivariate Analysis : Revolving Balance vs Loan Status

	Revolving Balance Status	loan_status	Number of Applicants
0	High	CHARGED OFF	577
1	High	FULLY PAID	3243
2	Low	CHARGED OFF	1295
3	Low	FULLY PAID	7474
4	Mid	CHARGED OFF	903
5	Mid	FULLY PAID	5166
6	Very High	CHARGED OFF	1158
7	Very High	FULLY PAID	6320
8	Very Low	CHARGED OFF	1694
9	Very Low	FULLY PAID	10743

	Revolving Balance Status	CHARGED OFF	FULLY PAID	TOTAL	FRAUD PERCENTAGE
3	Very High	1158	6320	7478	15.485424
0	High	577	3243	3820	15.104712
2	Mid	903	5166	6069	14.878893
1	Low	1295	7474	8769	14.767932
4	Very Low	1694	10743	12437	13.620648



Insight:

It can be observed from the above plot, that with increasing revolving balance the likelihood of being a defaulter also increases.

Analysis – 26

Bivariate Analysis : Correlation

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths
id	1.000000	0.993529	0.120403	0.131078	0.231355	0.176314	0.053692	0.075864	0.005847	0.091638	-0.008504	-0.041139
member_id	0.993529	1.000000	0.120179	0.130098	0.241077	0.194833	0.050557	0.070687	0.006723	0.092763	-0.007994	-0.046001
loan_amnt	0.120403	0.120179	1.000000	0.981788	0.937921	0.346616	0.301204	0.932254	0.269119	0.062391	-0.031982	0.012913
funded_amnt	0.131078	0.130098	0.981788	1.000000	0.956173	0.324858	0.304870	0.958031	0.264917	0.062150	-0.031896	0.012830
funded_amnt_inv	0.231355	0.241077	0.937921	0.956173	1.000000	0.343882	0.297387	0.905460	0.252147	0.070585	-0.038213	-0.002849
term	0.176314	0.194833	0.346616	0.324858	0.343882	1.000000	0.440179	0.090407	0.043926	0.076150	0.007253	0.047689
int_rate	0.053692	0.050557	0.301204	0.304870	0.297387	0.440179	1.000000	0.277142	0.049009	0.110845	0.158459	0.133328
installment	0.075864	0.070687	0.932254	0.958031	0.905460	0.090407	0.277142	1.000000	0.267959	0.051994	-0.019785	0.010988
annual_inc	0.005847	0.006723	0.269119	0.264917	0.252147	0.043926	0.049009	0.267959	1.000000	-0.121458	0.022260	0.035517
dti	0.091638	0.092763	0.062391	0.062150	0.070585	0.076150	0.110845	0.051994	-0.121458	1.000000	-0.033369	0.002112
delinq_2yrs	-0.008504	-0.007994	-0.031982	-0.031896	-0.038213	0.007253	0.158459	-0.019785	0.022260	-0.033369	1.000000	0.008723

Insight:

Their is strong correlation between loan_amnt and installment(which is obvious). Also moderate correlation between interest rate-revol_util and loan_amnt-term exists.

Conclusion:

- We have done EDA on the loan data to find out the driving factors of credit loss of the company.
- Data Quality issues (missing value, NA values, improper format) have been addressed and outliers have been removed accordingly during analysis.
- Derived metrics have been created for annual income, interest rate, revolving balance etc. for segregating data to give meaningful insights in our analysis.

- Based on EDA analysis, following variables can be said to be influencing the loan status to large extent and are primary drivers for loan default:
    - Loan Purpose
    - Loan Term
    - Home Ownership
    - Loan Grade
    - Verification Status
    - Public Bankruptcy records
    - Interest Rate
    - Employment Term
    - Applicant's Annual Income Range

Thank You!