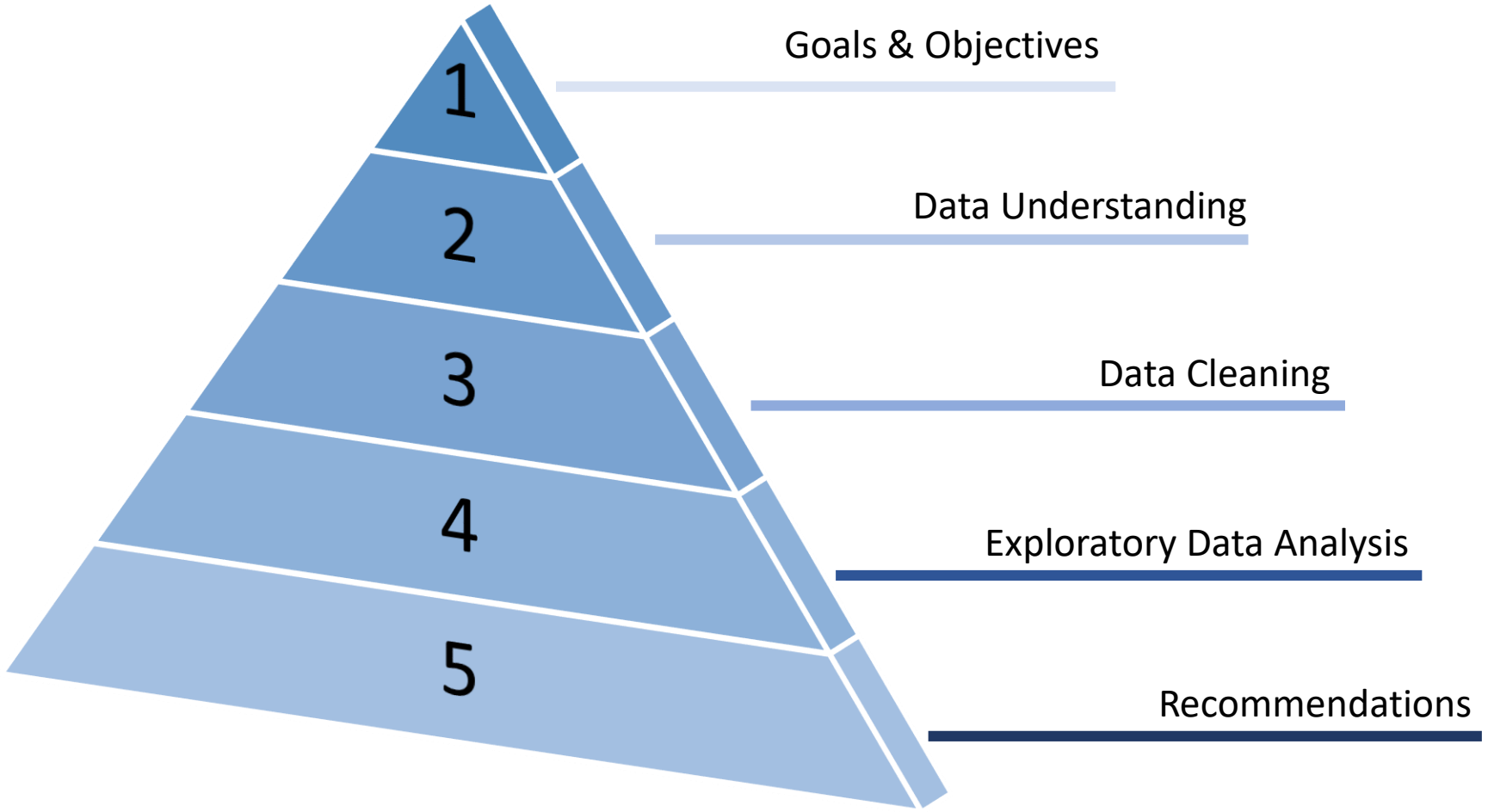
Lending Club Case Study

From : Vibhu Dutt Raturi


1
2
3
4
5
Goals & Objectives
Data Understanding
Data Cleaning
Exploratory Data Analysis
Recommendations

Goals

- Identify risk & predictors of default using Exploratory Data Analysis (EDA) in this case study.

Objectives

- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.
- To utilize this analysis in portfolio and risk assessment.
- Lessen the false positives

Dataset

Lets deep dive into the data provided and analyze it.

Number of Rows (records) – 39717
Number of Columns (features) – 111

Few Observations:

NA values – Out of 111 columns 54 columns had all NA values.

Missing Value Percentage – There were 3 columns which had more than 60% of missing values.

Customer Behavior Features – The customer behavior variables are not available at the time of loan application and we identified 21 such columns.

Duplicate Records – No duplicate records in this dataset.

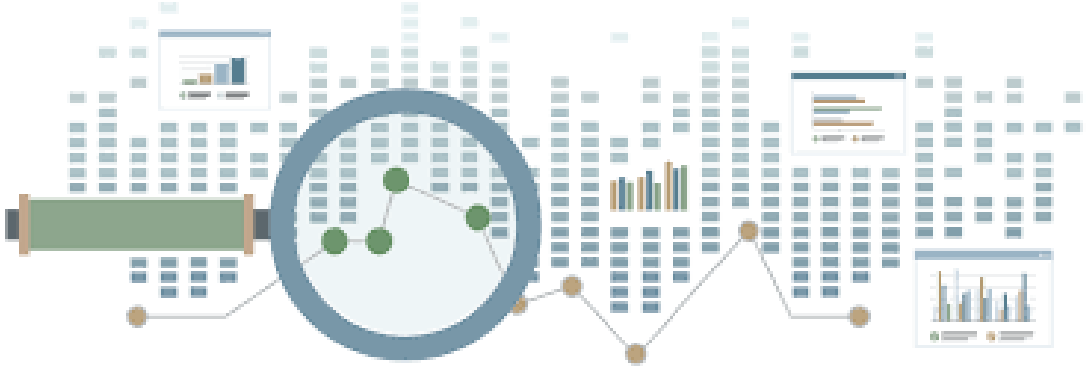
Unique Values – There were 8 columns which had only 1 unique value throughout the its rows.

Cleaning Process

- Removed all the columns having all NA values
- Removed columns having more than 60% of missingness
- Removed customer behavior features as they wont be useful in our analysis
- Removed columns which had only 1 unique value throughout
- After removing columns which were not useful we were left with 21 columns from 111 initially
- “term” column had extra whitespaces which had to be removed
- Outlier Detection – In our analysis we used boxplots to check the distribution of numerical features and observed outliers. There are many transformation techniques like log, exp, sqrt to overcome and one of the most famous method is quantile method to set a threshold percentile (eg 0.95) and check whether outliers fall beyond that and can be removed.

Cleaning Process

- “int_rate” column which is the interest rate of that particular record had values in percentage (eg 12.4%, 10.5%). We removed the percentage sign to make it easy for visualization
- “issue_d” column which had date (month-year format) when the loan was funded (eg Dec-11) we extracted month and year and created two new columns of month and year for visualization
- We also created categories for some features
  - “addr_state” : Had state abbreviations we divided them among 5 regions of the US
  - “annual_inc” : Had annual income of borrowers we divided them in 4 categories (Low, Medium, High & Very High Income)
  - “dti” , “loan_amnt” , “funded_amnt_inv” : DTI is the Debt-to-Income Ratio, Loan amount is the amount applied by the borrower, Funded amount is the amount funded by the investor of the borrower. We created buckets/bins for these 3 columns and mapped the borrowers accordingly (eg 5-10k , 10-15k or 0-5% , 5-10%)

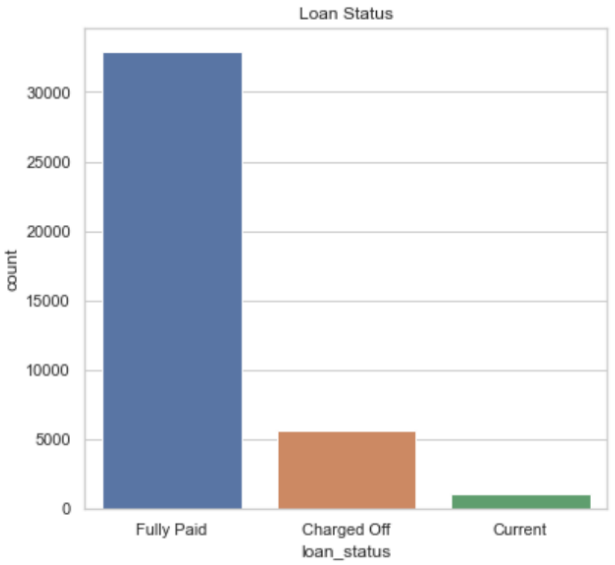


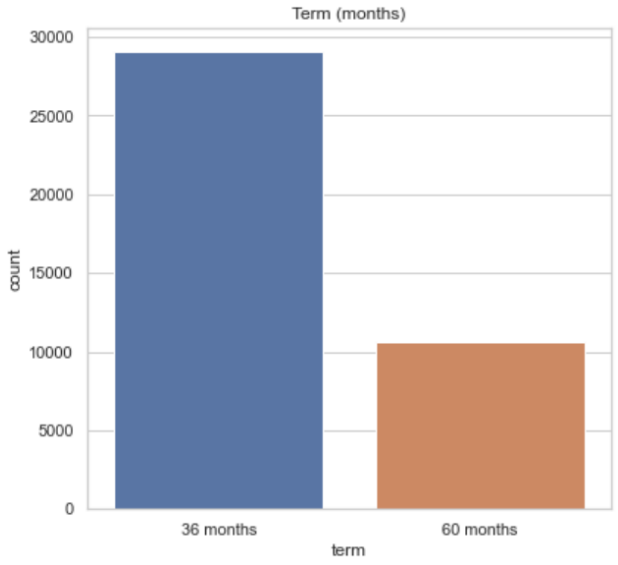
Data Analysis

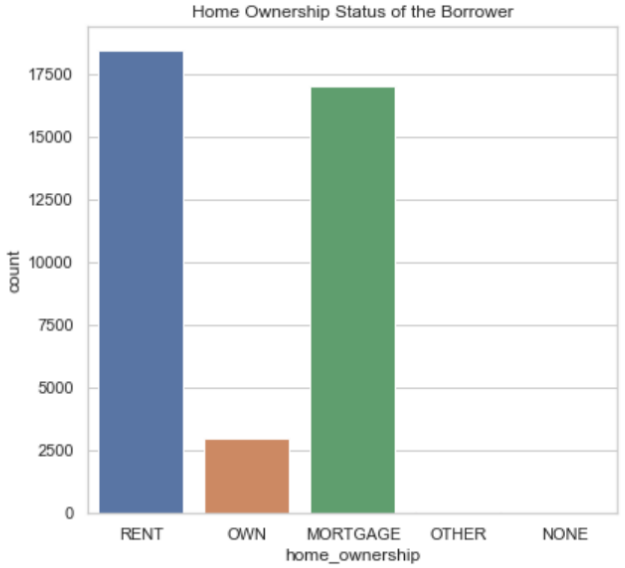
- Loan Status – There are 3 types of loans
  - Fully Paid : Completed
  - Charged Off : Defaulted
  - Current : Ongoing loan

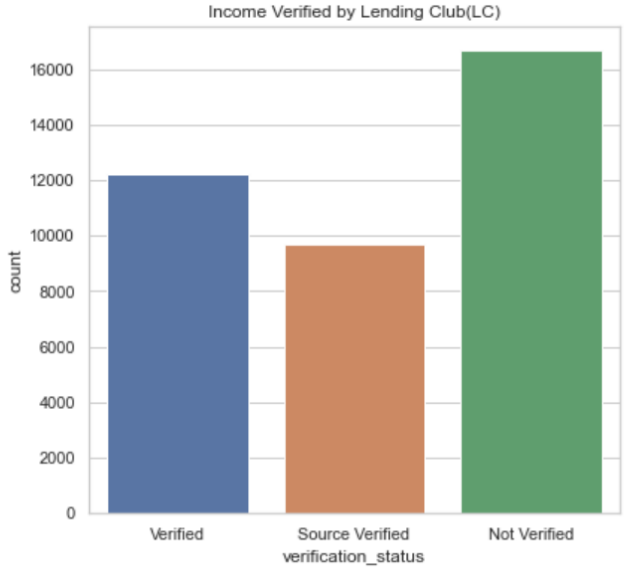
- Term is the tenure of the loans and there are only 2 terms 36 months and 60 months

- Home Ownership have 5 categories out of which Rent and Mortgage category have most borrowers
- Most number of loans are not verified which means the annual income is not verified of the borrowers which becomes a risk factor.



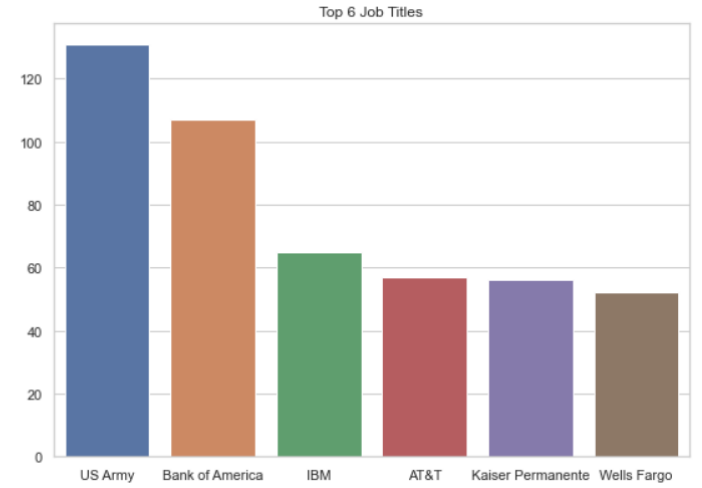


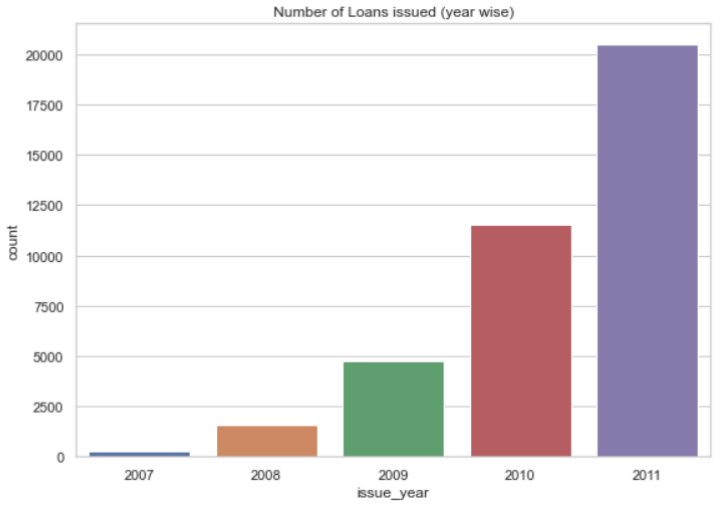


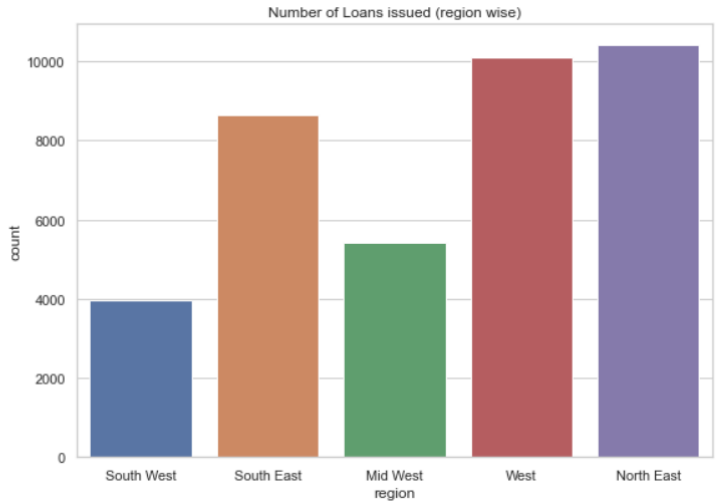


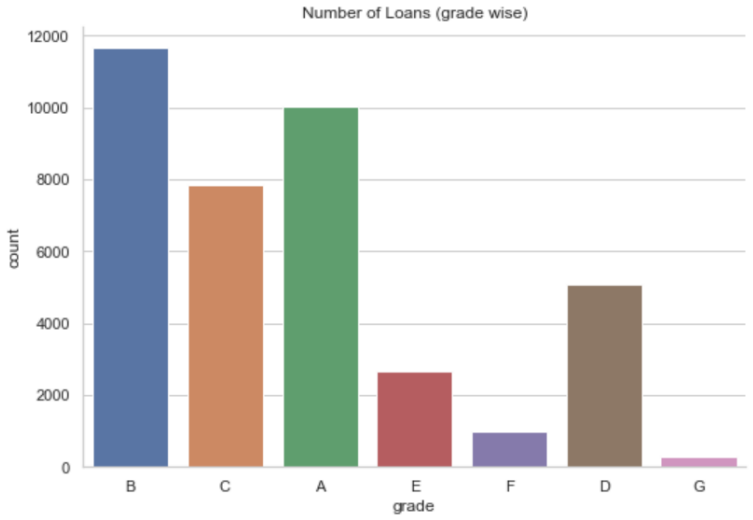
Data Analysis

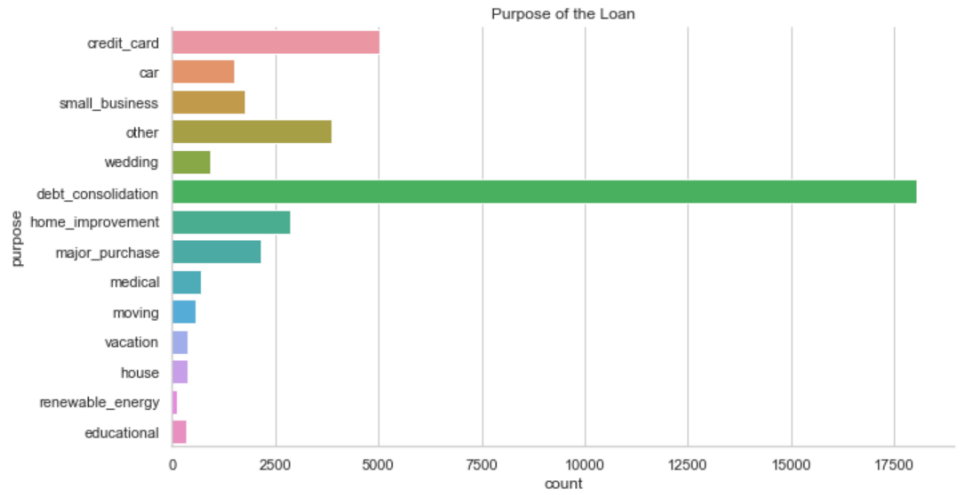
- Job Title – Gives us an insight of the top 6 Job title's which applied for a loan and we can see US Army employees have maximum number of loans
- Loan issued from 2007-2011 sees a drastic increase every year
- North East & West region have maximum number of loans issued.
- Grade B borrowers followed by Grade A borrowers have maximum loans issued.

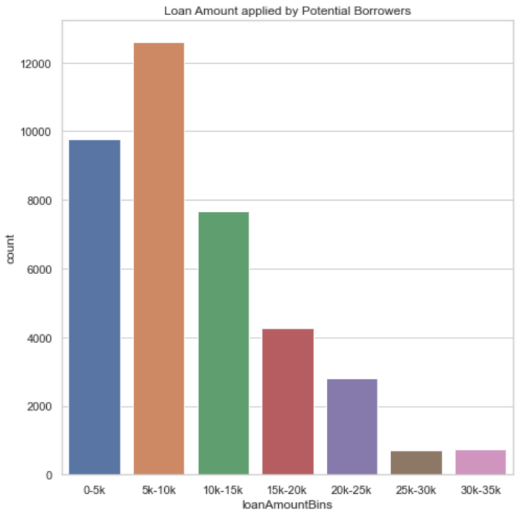


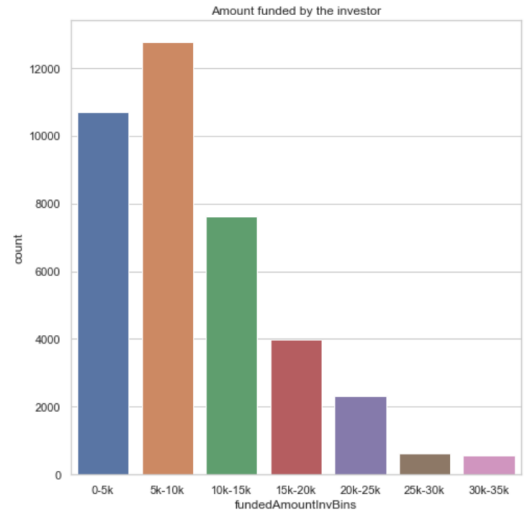






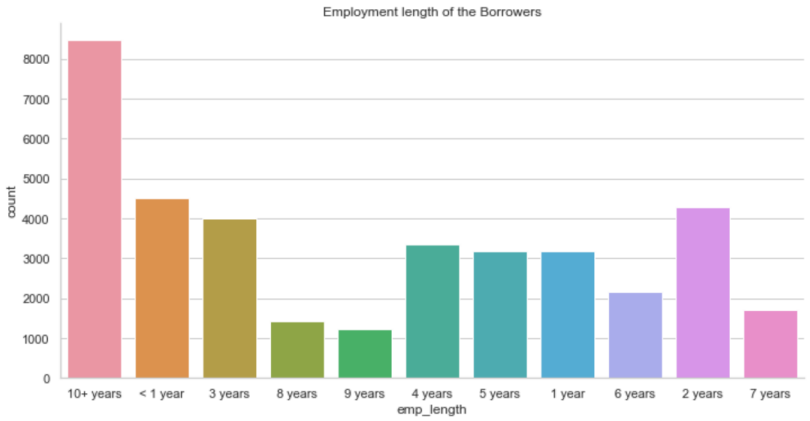


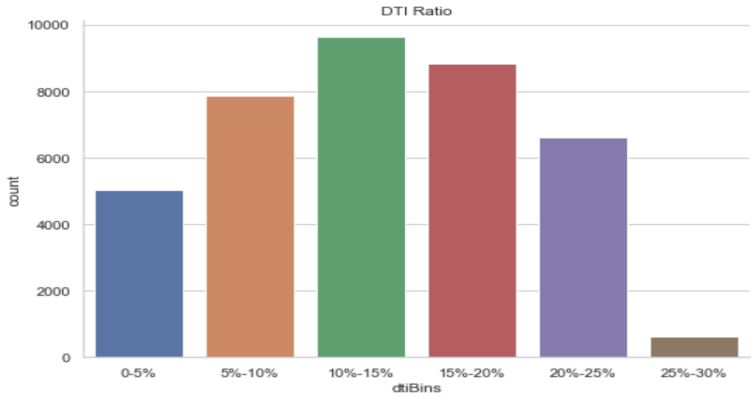


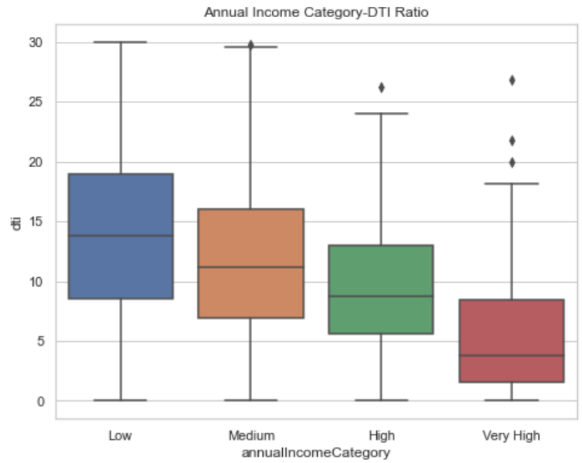


Data Analysis

- 1st Visual Purpose of the Loan – As we can see maximum number of loans are under 'debt consolidation' purpose and then followed by 'credit card'.
- 2nd Visual tells us about the Loan amount applied by the borrowers and we can see that count of maximum amount was between 5k-10k then followed by 0-5k range and so on.
- 3rd Visual tells us the Amount funded by the investor to the borrower and here as well maximum lie's between 5k-10k range. Now if we look closely at these visuals we can see that loan amount in 0-5k bar increases in the funded amount by investor graph, which means the borrower applied for a bigger Loan amount but got an amount in the 0-5k range lower to what applied. Similarly, in 15k-20k and 20k-25k range Loan(Fully Paid & Charged Off) was applied by the borrowers but received an amount lower to it which we can see in Amount funded by the investors graph.







Data Analysis

- Employment Length – 10 plus years of experience have most number of loans followed by borrowers with less than 1 year of experience which is little interesting.

- DTI Buckets – Most of the borrowers fall in 10-15% DTI and 25-30% being the least.

- DTI ratio with respect to annual income category and we can see with increase in income, DTI ratio decreases which is a good sign for loan approval.

Charged Off (Defaulted) Loans Proportion

Charged Off Proportion (Purpose of the Loan)

loan_status	purpose	Charged Off	Fully Paid	Total	chargedOffProportion
11	small_business	475	1279	1754	27.08
10	renewable_energy	19	83	102	18.63
3	educational	56	269	325	17.23
9	other	633	3232	3865	16.38
5	house	59	308	367	16.08
8	moving	92	484	576	15.97
7	medical	106	575	681	15.57
2	debt_consolidation	2767	15288	18055	15.33
12	vacation	53	322	375	14.13
4	home_improvement	347	2528	2875	12.07
1	credit_card	542	4485	5027	10.78
0	car	160	1339	1499	10.67
13	wedding	96	830	926	10.37
6	major_purchase	222	1928	2150	10.33

Charged Off Proportion (Annual Income category)

loan_status	annualIncomeCategory	Charged Off	Fully Paid	total	chargedOffProportion
1	Low	5035	28061	33096	15.21
3	Very High	21	150	171	12.28
2	Medium	532	4379	4911	10.83
0	High	39	360	399	9.77

Key Observations

- Purpose – Small Business has the highest charged off loans with 27.08 % and if we look closely debt consolidation has maximum number of loans with 15.33% charged off loans.

- Annual Income – Low Income has the maximum percentage 15.21% for charged off loans & interestingly followed by Very High income 12.28%

Charged Off (Defaulted) Loans Proportion

Charged Off Proportion (States)

loan_status	addr_state	Charged Off	Fully Paid	Total	chargedOffProportion
28	NE	3.00	2.00	5.00	60.00
32	NV	108.00	371.00	479.00	22.55
40	SD	12.00	50.00	62.00	19.35
0	AK	15.00	63.00	78.00	19.23
9	FL	504.00	2277.00	2781.00	18.12
24	MO	114.00	556.00	670.00	17.01
11	HI	28.00	138.00	166.00	16.87
13	ID	1.00	5.00	6.00	16.67
31	NM	30.00	153.00	183.00	16.39
36	OR	71.00	364.00	435.00	16.32
4	CA	1125.00	5824.00	6949.00	16.19
43	UT	40.00	212.00	252.00	15.87
20	MD	162.00	861.00	1023.00	15.84
10	GA	215.00	1144.00	1359.00	15.82
30	NJ	278.00	1512.00	1790.00	15.53
46	WA	127.00	691.00	818.00	15.53

Charged Off Proportion (Region)

loan_status	region	Charged Off	Fully Paid	Total	chargedOffProportion
4	West	1628	8484	10112	16.10
2	South East	1307	7351	8658	15.10
0	Mid West	759	4652	5411	14.03
1	North East	1424	8994	10418	13.67
3	South West	509	3469	3978	12.80

Key Observations

- Region – West region has maximum charged off loans with 16.10% and South West region being the lowest.

- States – NE(Nebraska) has the highest Charged Off Proportion but it is because we have very less applicants(only 5) from that state. NV(Nevada) ,FL(Florida) and CA(California) are the states which has substantial number of loans given and also the charged off proportion is high which is 22.55%, 18.12% and 16.19% respectively.

Charged Off (Defaulted) Loans Proportion

Charged Off Proportion (Sub-Grades)

	loan_status	sub_grade	Charged Off	Fully Paid	Total	chargedOffProportion
29		F5	54	59	113	47.79
32		G3	19	26	45	42.22
31		G2	28	49	77	36.36
28		F4	53	98	151	35.10
34		G5	10	19	29	34.48
30		G1	31	63	94	32.98
26		F2	70	163	233	30.04
25		F1	91	214	305	29.84
23		E4	126	298	424	29.72
27		F3	51	123	174	29.31
24		E5	109	278	387	28.17
20		E1	198	524	722	27.42
21		E2	163	451	614	26.55
19		D5	209	625	834	25.06
33		G4	13	41	54	24.07
18		D4	215	703	918	23.42
22		E3	119	397	516	23.06
17		D3	256	860	1116	22.94
16		D2	271	1015	1286	21.07

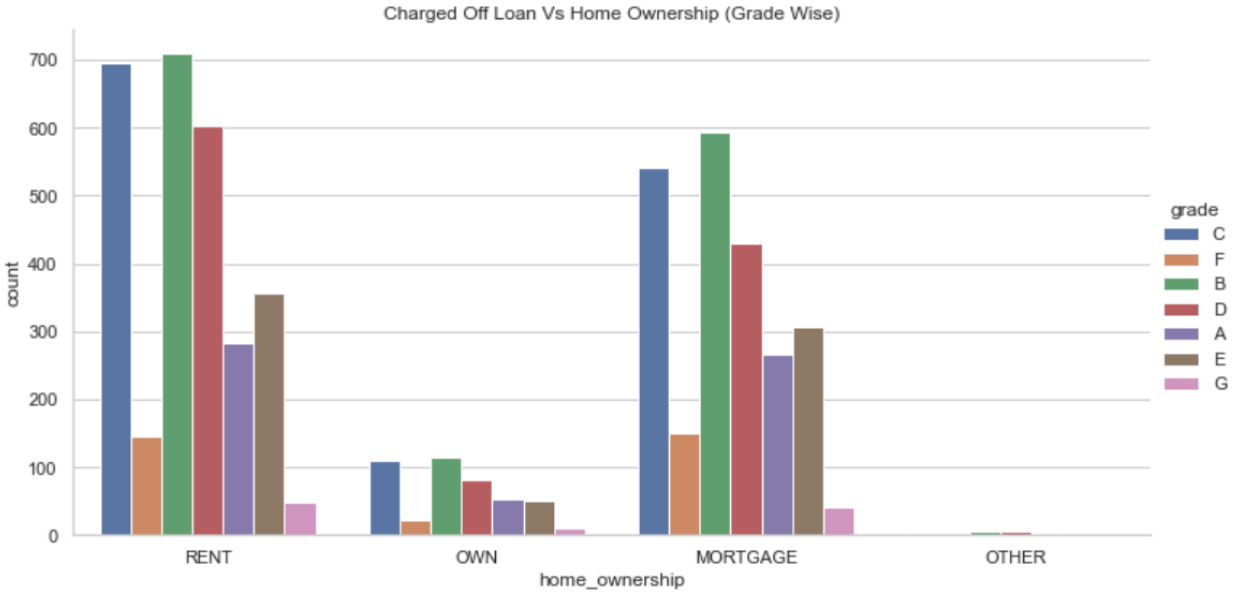
Charged Off Proportion (Grades)

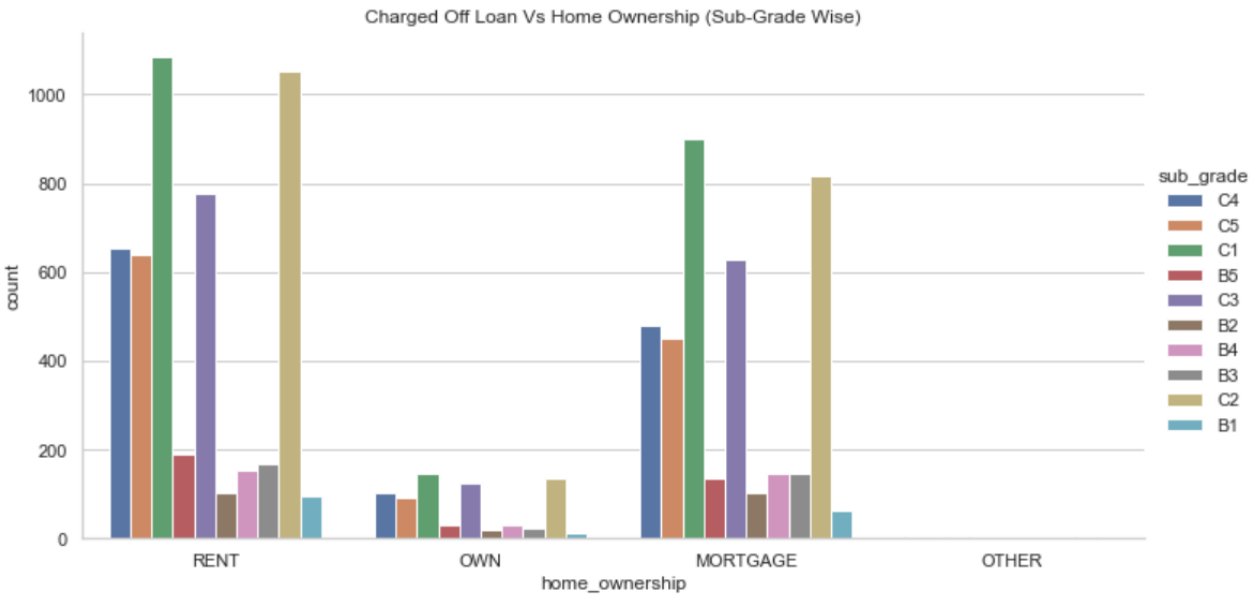
	loan_status	grade	Charged Off	Fully Paid	Total	chargedOffProportion
0		A	602	9443	10045	5.99
1		B	1425	10250	11675	12.21
2		C	1347	6487	7834	17.19
3		D	1118	3967	5085	21.99
4		E	715	1948	2663	26.85
5		F	319	657	976	32.68
6		G	101	198	299	33.78

Key Observations

- Grades – G(33.78%) & F (32.68%) Grades have the highest Charged off percentage
- Sub-Grades – Further inside F5 (47.79%) & G3(42.22%) sub grades impact the most

Data Analysis

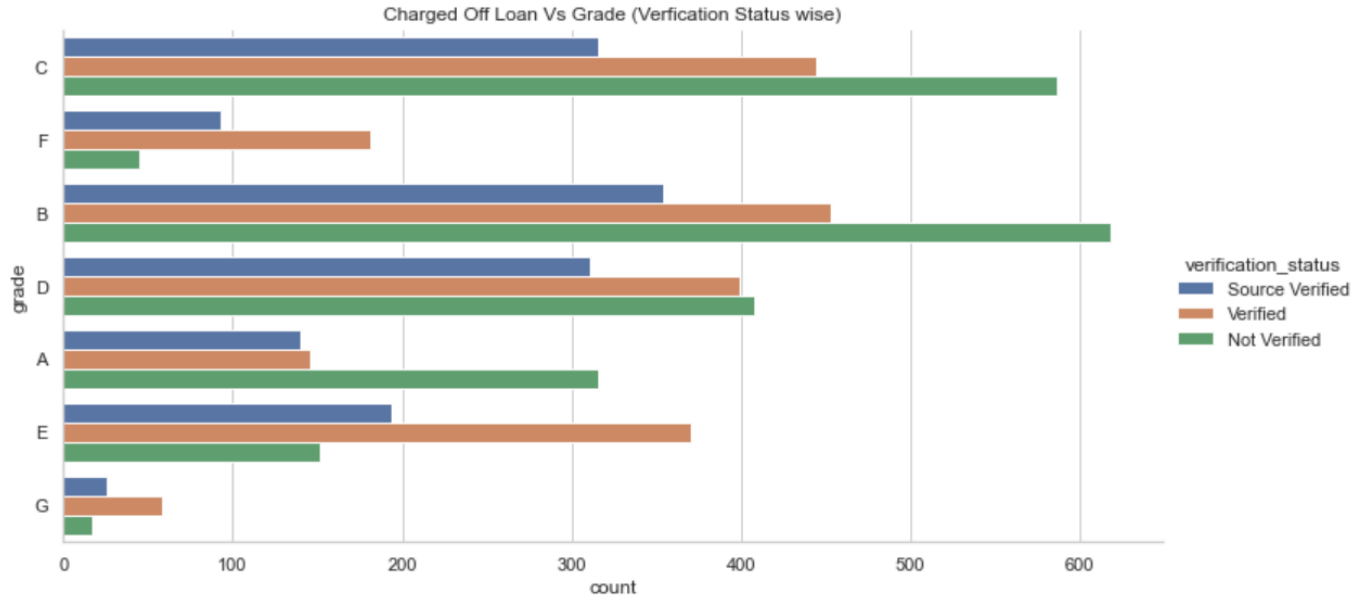




Key Observations

- 1st Visual : Most Charged off loans come from Borrowers with Rent and then Mortgage as their Home ownership. If a Borrower has Home Ownership as Rent & has a B Grade score(Green bar) then they are most likely to default on their loan.
- Within these home ownership category we can see B & C Grade are the ones which Default maximum followed by other Grades.
- 2nd Visual : Further into B & C Grades, if a Borrower has home ownership as Rent and falls under Grade C and within it has a sub grade of C1 followed by C2 are more likely to default on a loan.

Data Analysis

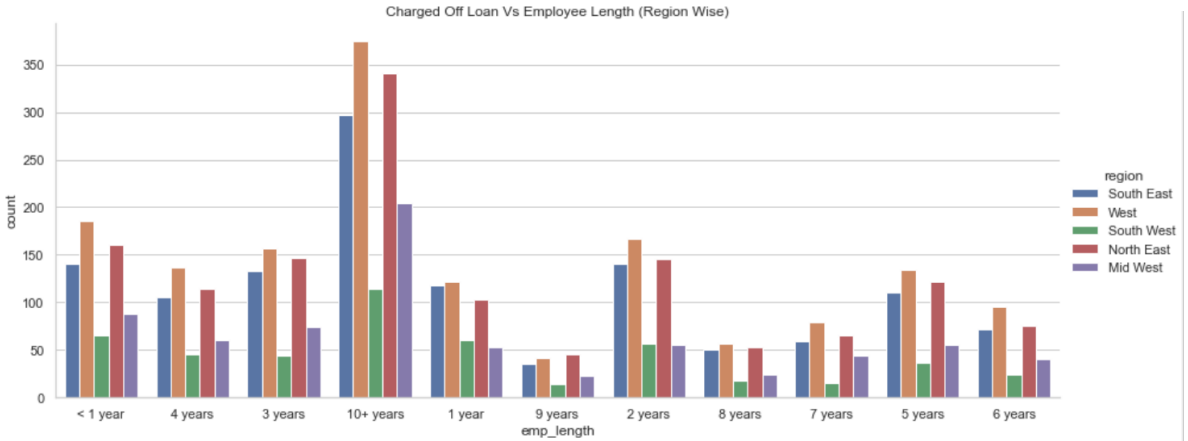


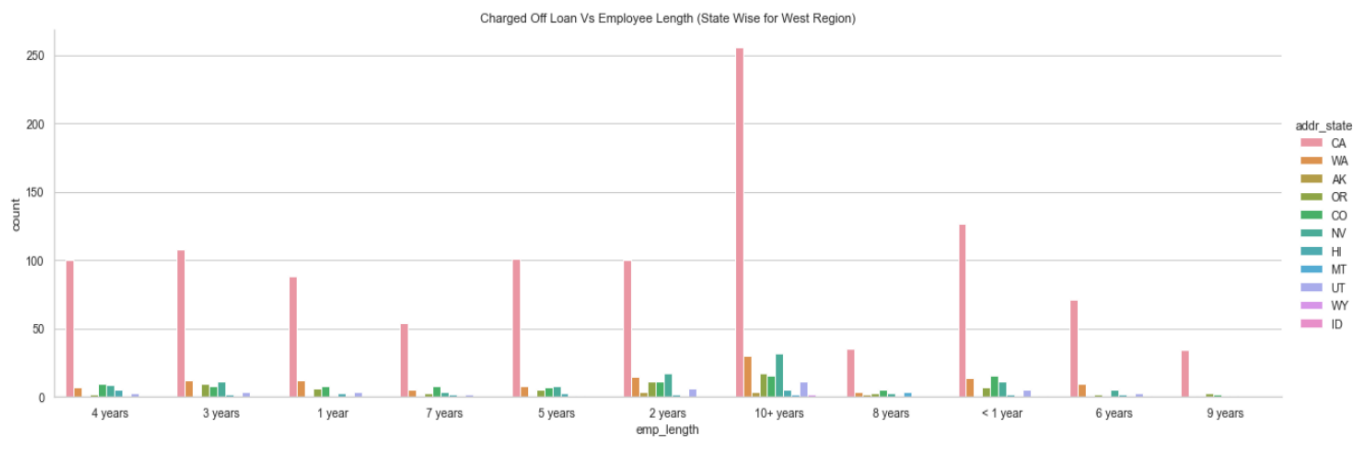
Key Observations

- 1st Visual : We have seen in previous few visuals that B & C Grades have been at top for Charged off Loan and if we look at this visual it somewhat tells us that the loans defaulted are maximum when the verification status is Not Verified.

- Risk of defaulting a loan increases when a Borrower has a grade of B or C and it is not verified before approving the loan.

Data Analysis



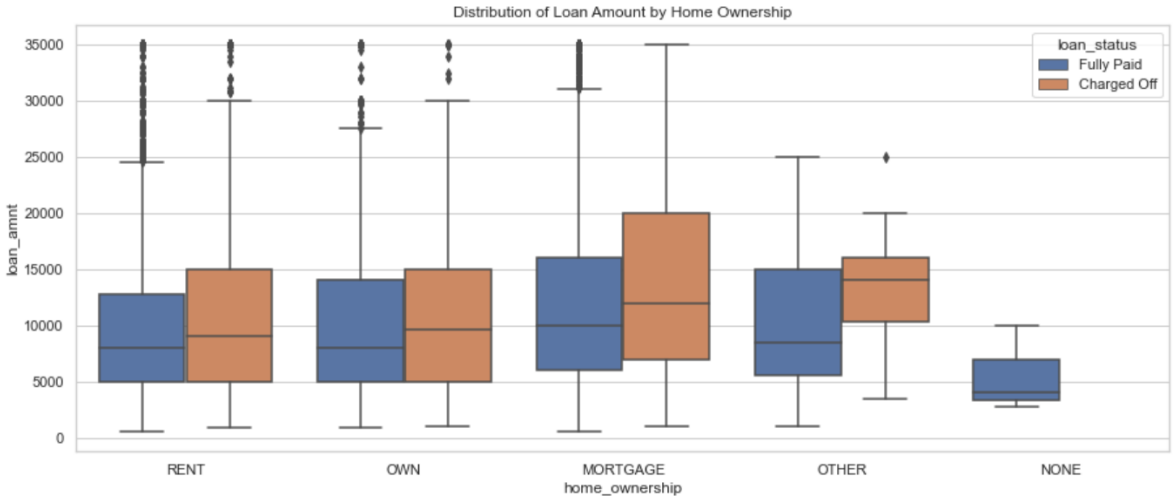


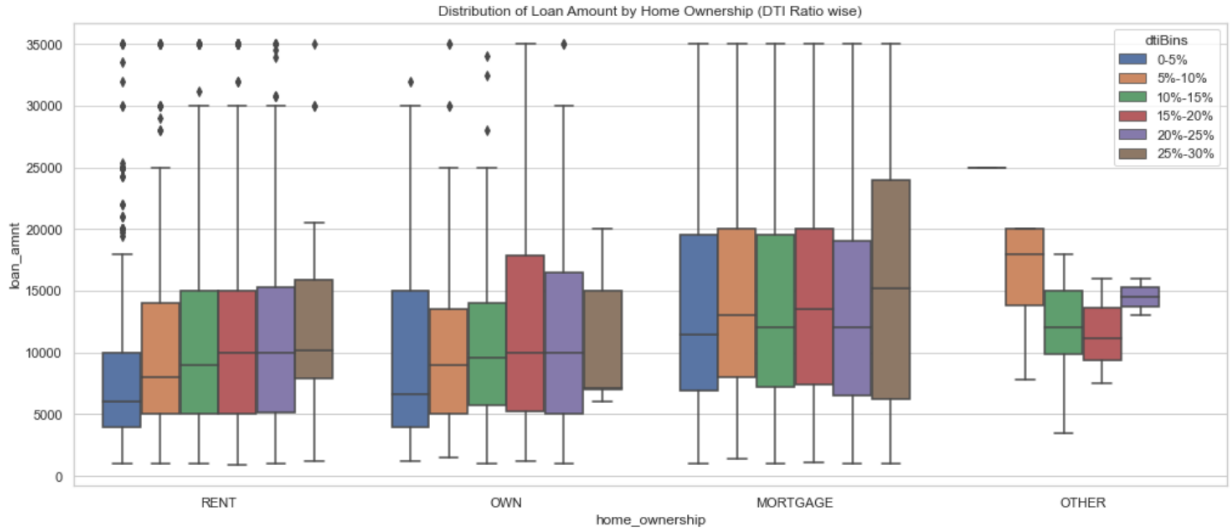
Key Observations

- 1st Visual : We can see almost a same pattern across the categories of employee length. Orange bar(West), Red bar(North East), Blue bar(South East), Purple bar(Mid West) and then Green bar(South West) have almost and same pattern or shape throughout. If a Borrower is from West region and has 10 plus years of experience they are more likely to default on a loan. Across all employee length 'West' region has more risk involved while giving a loan.

- 2nd Visual : CA(California) is the one which has maximum number of defaults, though it is also because the number of loans are maximum in CA with a 16.18% of charged off proportion of the state. If a Borrower is from West region specially from CA with employment length of 10+ years is more likely to Default on a loan. CA state has maximum number of Defaults throughout employment length years and has more risk involved while giving a loan.

Data Analysis



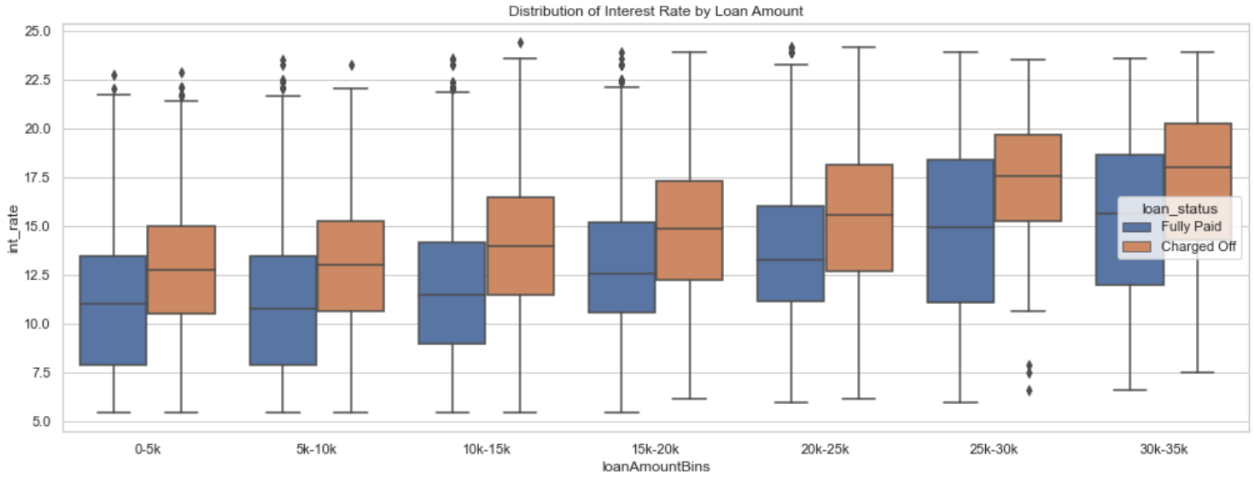


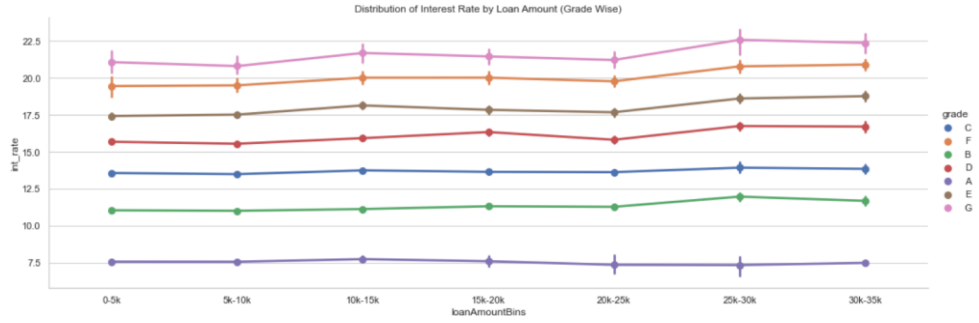
Key Observations

- 1st Visual : Distribution of the loan amount applied by the Borrowers with respect to their home ownership status. Distribution of loan amount for Mortgage when compared to others categories is more. Borrowers with Mortgage have more risk involved

- 2nd Visual : Now once we got to know that Mortgage from Home Ownership has bigger distribution of charged off loans we went further inside to check DTI Ratio's of these Borrowers. Borrowers with Home ownership status as Mortgage and DTI Ratio falling between 25%-30% have a greater risk involved for a loan.

Data Analysis

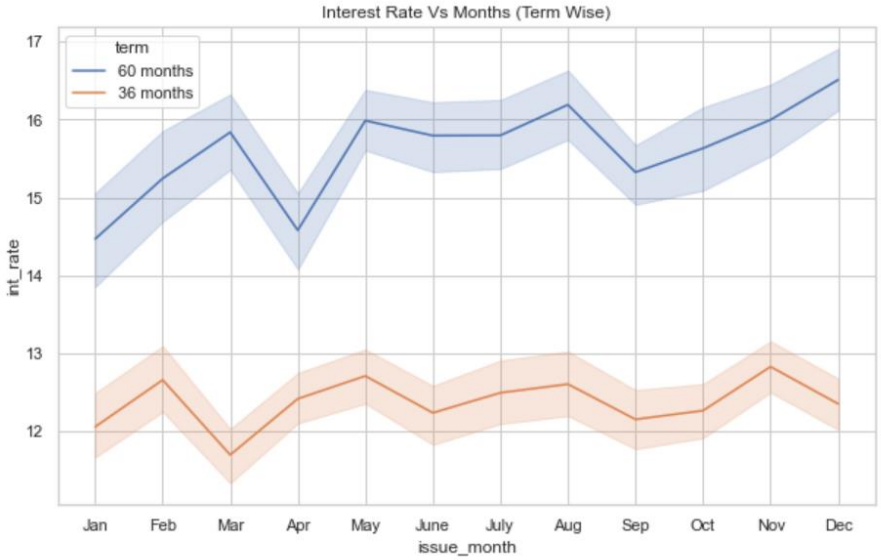


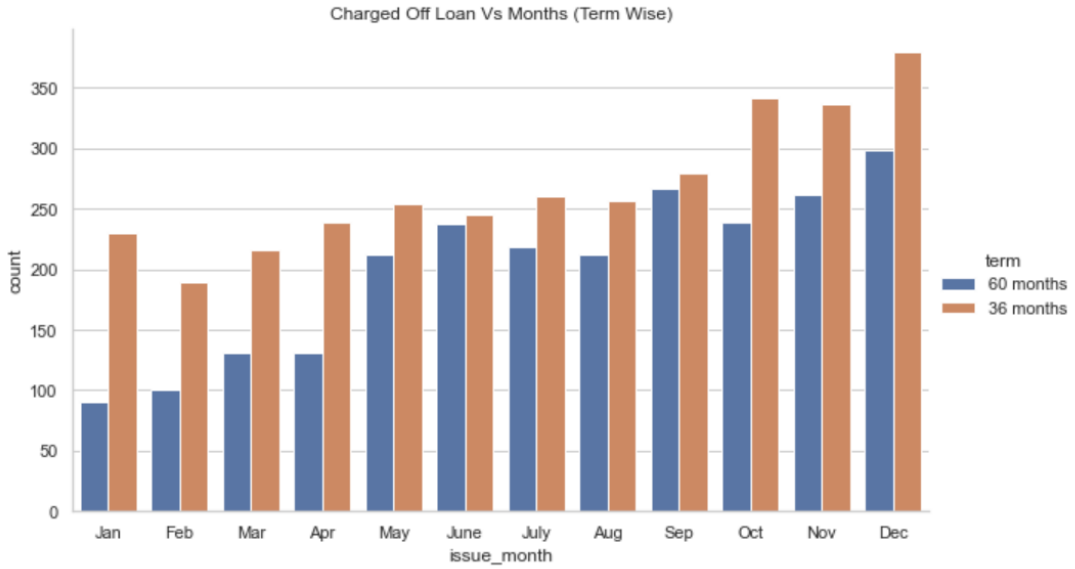


Key Observations

- 1st Visual : The interest rate of Charged Off loans is greater than Fully Paid loans across all the loan amount categories which can be a strong factor for defaulted loans. Now we wanted to check interest rates specifically for charged off loans and wanted to see if it is getting affected by the Grade score of the Borrowers.
- 2nd Visual : Interest rate increases from A to G(pink line) . G ‘s int rate is higher than 20% as it involves a lot of risk.
- 3rd Visual : We know that charged off loans have a greater interest rate through out and it is higher when a Borrower has a G Grade score. It gives us an insight that if a Borrower in G Grade and falls under low income category and applies for a loan amount of greater than 10k till 25k it has a greater risk involved.

Data Analysis



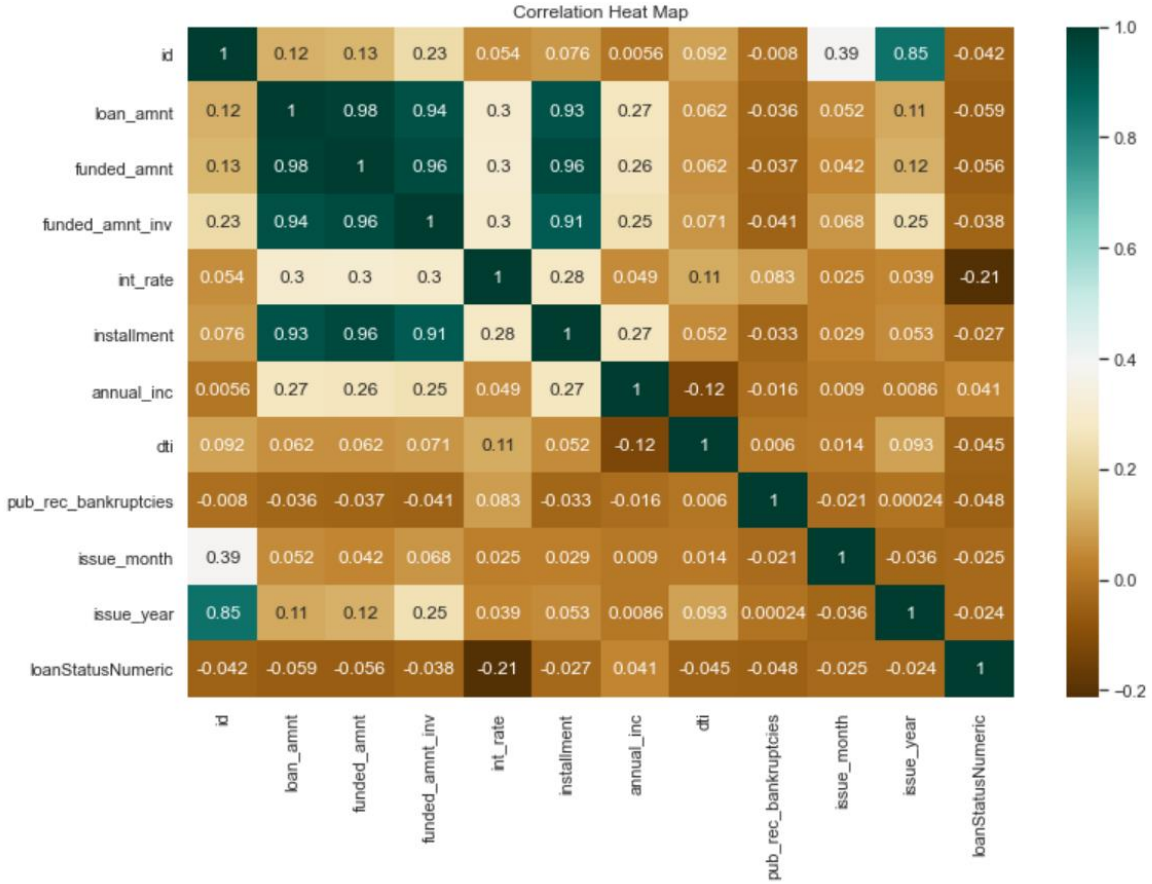


Key Observations

- 1st Visual : Gives us an insight on the interest rate with of the term of a loan, month wise and as we can see lower the duration of the term, lower is the interest on the loan. For 36 months its more or less similar throughout a year but a slight increase in interest rate for 60 months.

- 2nd Visual : Gives us an interesting insight that loans having a term of 36 months are getting charged off more than the loans having a term of 60 months even when the interest rate and duration is less.

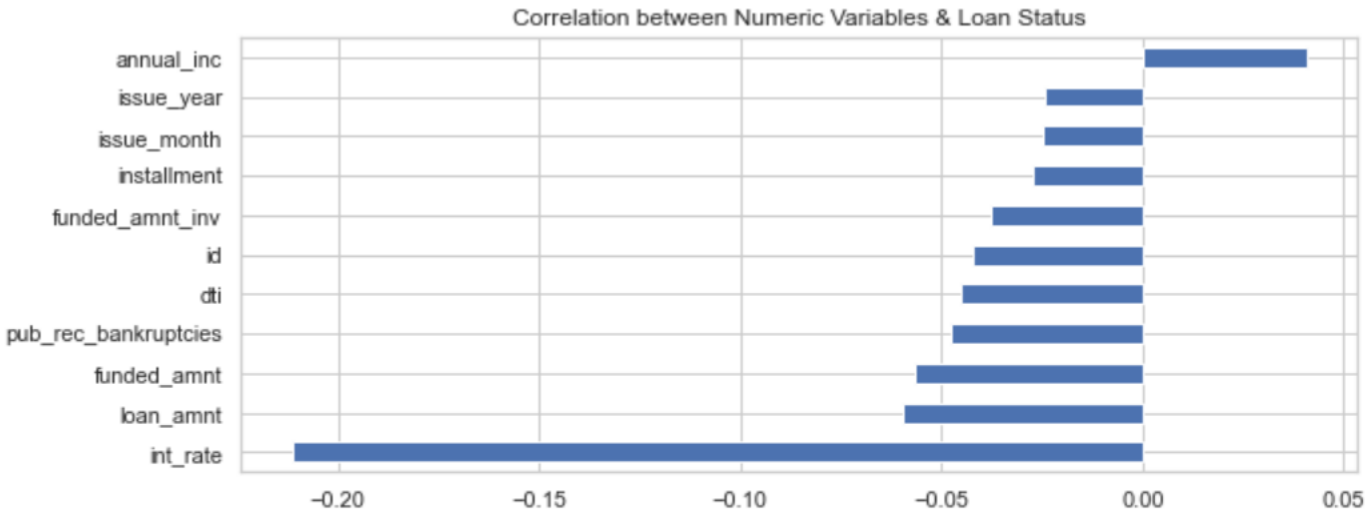
Data Analysis



Key Observations

- Visual : Heat Map gives us an insight on how numerical features of this dataset are corelated to each other where correlation ranges from negative(-1) to positive(+1) correlation. Darker the color more strong is the correlation.
- New column loanStatusNumeric created. We can see strong positive correlation between Loan Amount and the installment for it (0.93), DTI has a negative correlation with annual income (-0.12). Which means when Annual income increases the DTI ratio linked to it decreases. If DTI Ratio is more, mostly the annual income will be less and more chances of risk is involved.

Data Analysis



Key Observations

- Visual : loanStatusNumeric column has Loan Status as 1 - Fully Paid & 0 - Charged Off
- This is a correlation visual between Loan Status and the numeric features of this dataset.
- Most of the features are negatively correlated with loan status and the strongest negative correlation of loanStatusNumeric is with int_rate which means when interest rate increases for a loan, it has more risk involved and is more likely to default.
- Annual Income has a positive correlation with loan status, not strong enough but still positive.

Data Analysis

Details w.r.t Annual Income & Purpose

	annualIncomeCategory	purpose	int_rate	loan_amnt	fully_paid_loan_count	charged_off_loan_count	total_loan_count	Default/Total_loan (%)
16	High	renewable_energy	21.64	25000.00	0.00	1.00	1.00	100.00
5	High	educational	14.96	12000.00	0.00	1.00	1.00	100.00
94	Very High	wedding	13.74	15666.67	4.00	2.00	6.00	33.33
42	Low	small_business	12.78	12149.91	1023.00	416.00	1439.00	28.91
13	High	moving	11.62	17900.00	3.00	1.00	4.00	25.00

Key Observations

- Visual : We sorted the values with respect to Default Ratio. If we look at the Default Ratio and ignore the ones which has little to no data at all. We can say that Borrowers applying for a loan for the purpose of small business with an income falling in Low category are more likely to default(28.91%) on a loan and more risk is involved.

Data Analysis

Details w.r.t Annual Income & Purpose

	annualIncomeCategory	purpose	int_rate	loan_amnt	fully_paid_loan_count	charged_off_loan_count	total_loan_count	Default/Total_loan (%)
16	High	renewable_energy	21.64	25000.00	0.00	1.00	1.00	100.00
88	Very High	moving	18.08	15125.00	3.00	1.00	4.00	25.00
91	Very High	renewable_energy	15.78	15000.00	2.00	0.00	2.00	0.00
5	High	educational	14.96	12000.00	0.00	1.00	1.00	100.00
4	High	debt_consolidation	14.47	21599.65	130.00	14.00	144.00	9.72
90	Very High	other	14.20	14864.29	11.00	3.00	14.00	21.43
94	Very High	wedding	13.74	15666.67	4.00	2.00	6.00	33.33
15	High	other	13.71	14576.88	33.00	7.00	40.00	17.50
93	Very High	small_business	13.62	19410.42	10.00	2.00	12.00	16.67
78	Very High	debt_consolidation	13.59	18488.71	57.00	5.00	62.00	8.06
10	High	major_purchase	13.50	16883.70	23.00	0.00	23.00	0.00
71	Medium	small_business	13.42	18390.96	217.00	54.00	271.00	19.93
8	High	house	13.41	27625.00	7.00	1.00	8.00	12.50
77	Very High	credit_card	13.37	14921.43	13.00	1.00	14.00	7.14
83	Very High	house	13.31	27750.00	4.00	0.00	4.00	0.00

Key Observations

- Visual : We have sorted the values with respect to Interest rate and Default Ratio. If we look closely most of the categories have very less data to analyze and conclude but looking at the data we can say that Borrowers with purpose as small business with Default Ratio 19.93% and interest rate 13.42% & Medium annual income or debt consolidation with Default Ratio 9.72% and interest rate 14.47% with High annual income are more likely to default.

Data Analysis

Details w.r.t Annual Income & Purpose

	annualIncomeCategory	purpose	int_rate	loan_amnt	fully_paid_loan_count	charged_off_loan_count	total_loan_count	Default/Total_loan (%)
5	High	educational	14.96	12000.00	0.00	1.00	1.00	100.00
16	High	renewable_energy	21.64	25000.00	0.00	1.00	1.00	100.00
94	Very High	wedding	13.74	15666.67	4.00	2.00	6.00	33.33
42	Low	small_business	12.78	12149.91	1023.00	416.00	1439.00	28.91
13	High	moving	11.62	17900.00	3.00	1.00	4.00	25.00
88	Very High	moving	18.08	15125.00	3.00	1.00	4.00	25.00
58	Medium	house	12.91	17216.67	40.00	11.00	51.00	21.57
90	Very High	other	14.20	14864.29	11.00	3.00	14.00	21.43
40	Low	renewable_energy	11.18	7309.94	66.00	17.00	83.00	20.48
85	Very High	major_purchase	11.14	11090.00	8.00	2.00	10.00	20.00
71	Medium	small_business	13.42	18390.96	217.00	54.00	271.00	19.93
15	High	other	13.71	14576.88	33.00	7.00	40.00	17.50
36	Low	moving	11.56	5588.53	412.00	87.00	499.00	17.43
27	Low	educational	11.65	6219.48	240.00	50.00	290.00	17.24
35	Low	medical	11.47	7613.18	490.00	98.00	588.00	16.67

Key Observations

- Visual : We have sorted with respect to Default Ratio and wanted interest rate to start from ascending order which works category wise. Now as we have already discussed about small business, we also noticed and even when the interest rate is low ranging from (11-15%) borrowers applying for a medical with low income category(Default ratio - 16.67%) or educational loan with low & high income category(Default ratio - 17.24% and for high only 1 applicant which defaulted) also have an high Default Ratio and there is risk involved.

To conclude

- Have to be extra careful when approving a loan for a purpose of small business specially when income is in Low or Medium Category, need to be careful if the purpose is debt consolidation as well.
- Observed that Medical & Educational loan's were bit risky as well when income was in low or medium category and even when the interest rate was low(10-15%) and hence need extra attention while approval.
- West region had most percentage of charged off loans and its important to cross check all the details properly. Specially when the applicant has 10 plus years of experience and from CA(California).
- Loan approval of an applicant from California(CA), Florida(FL) or Nevada(NV) state are most likely to default and hence need to be careful and check all the details
- Grades G & F (G2,G3,F4 & F5) are strong indicators of defaulting a loan.
- Home Ownership as Rent or Mortgage needs special attention. Specially if Grade is C(C1 & C2) or B. Also observed that majorly the loans approved for Grade B & C were not verified and more likely to become a bad loan hence need to verify before any loan approval.
- Need to be extra careful when DTI Ratio of an applicant is above 25%.
- Applicant from Low Income Category with Grade G is also a bad combination specially when the loan amount applied ranges between 10k-25k.