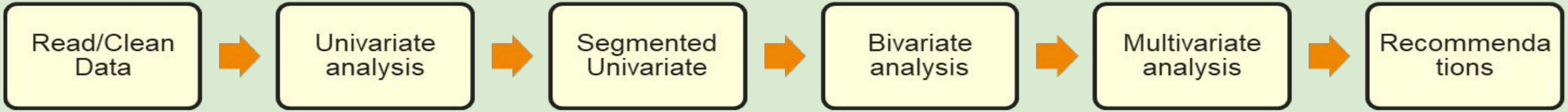
Lending club case study

Group members :

1. Raj Hujigal
2. Rakesh Bhadra

Methodology



Problem Statement

The company wants to understand the driving factors and the variables which are strong indicators of loan default, so that the company can utilise this knowledge for its portfolio and risk assessment.

The data that has been provided are :

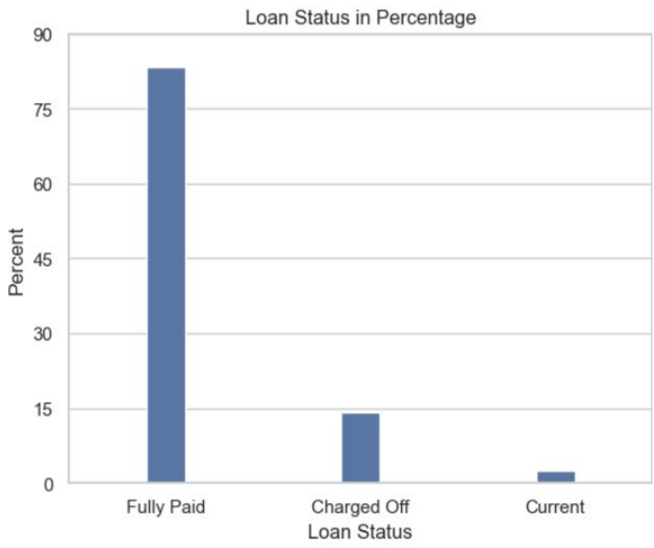
1) Loan Data
2) Loan Data Dictionary

Read Data & Clean Data

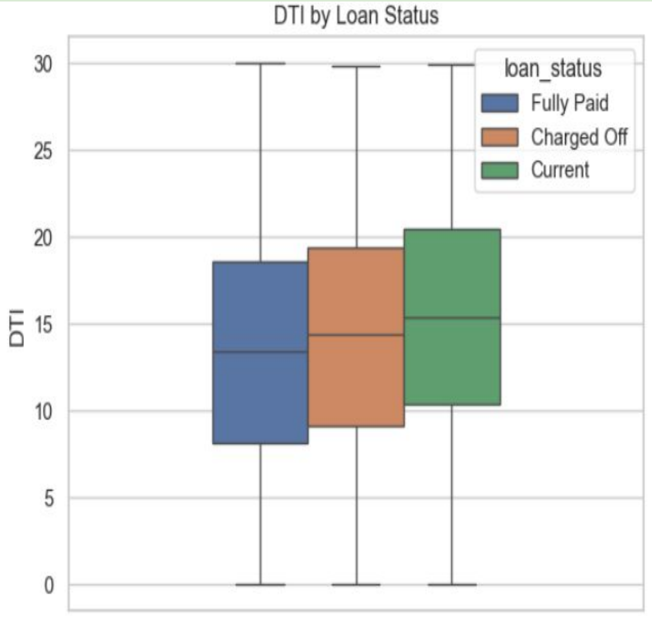
1. Read data from the loan data file in a dataframe.
2. Check for null only value columns and drop such columns.
3. Replace white space in columns if any with nan, so that those columns can be cleaned.
4. Drop all columns where there are 40 percent of null values in each column.
5. Check for the columns whose values needs special treatment like removing certain symbols or a specific text and clean such columns.
6. Check if all columns have proper data type, and if required convert the columns needed to the required data type.
7. Fill "na" values with any specific string if required for special columns. eg: Desc column with "Unknown" for "na" values.
8. Now for all the columns which still have "na" values , replace them with median for int and float data type columns, and mode for string and object data type columns.
9. Check for "na" values and make sure there are no such values.
10. Drop if any duplicate rows from the dataframe.
11. Remove rows for columns which have outliers. Since it is financial and loan data, we have removed outliers only in specific columns like loan amount , int rate and annual income.
12. Now the data is ready for analysis and generate any required data frames using various conditions.

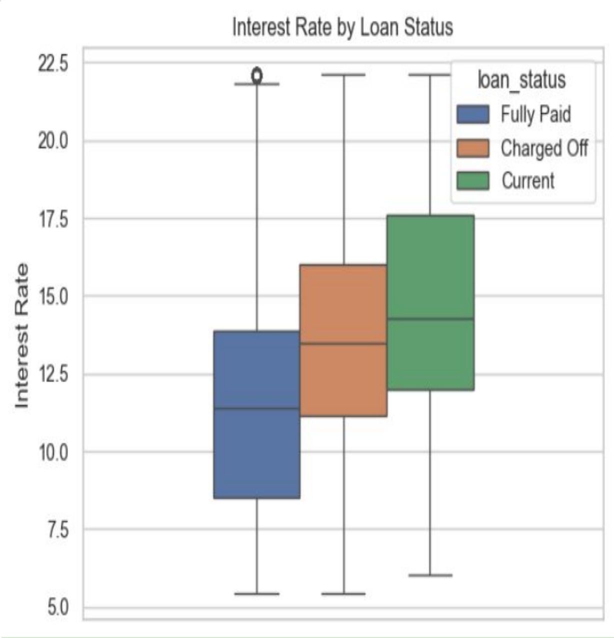
Univariate Analysis

1. About 83% of the loans are fully paid , 14.5 % are charged-off-loans and 2.5% are current loans.
2. The median for Current loans is greater than Charged Off loans, which in turn has value greater than Fully Paid loans.
3. The median Interest Rate for Current loans is greater than Charged off loans, which in turn has value greater than Fully-Paid loans.
4. We can generate plots graphs specific to charged-off-loans too to draw more inferences.



loan_status	
Fully Paid	83.343776
Charged Off	14.089956
Current	2.566268

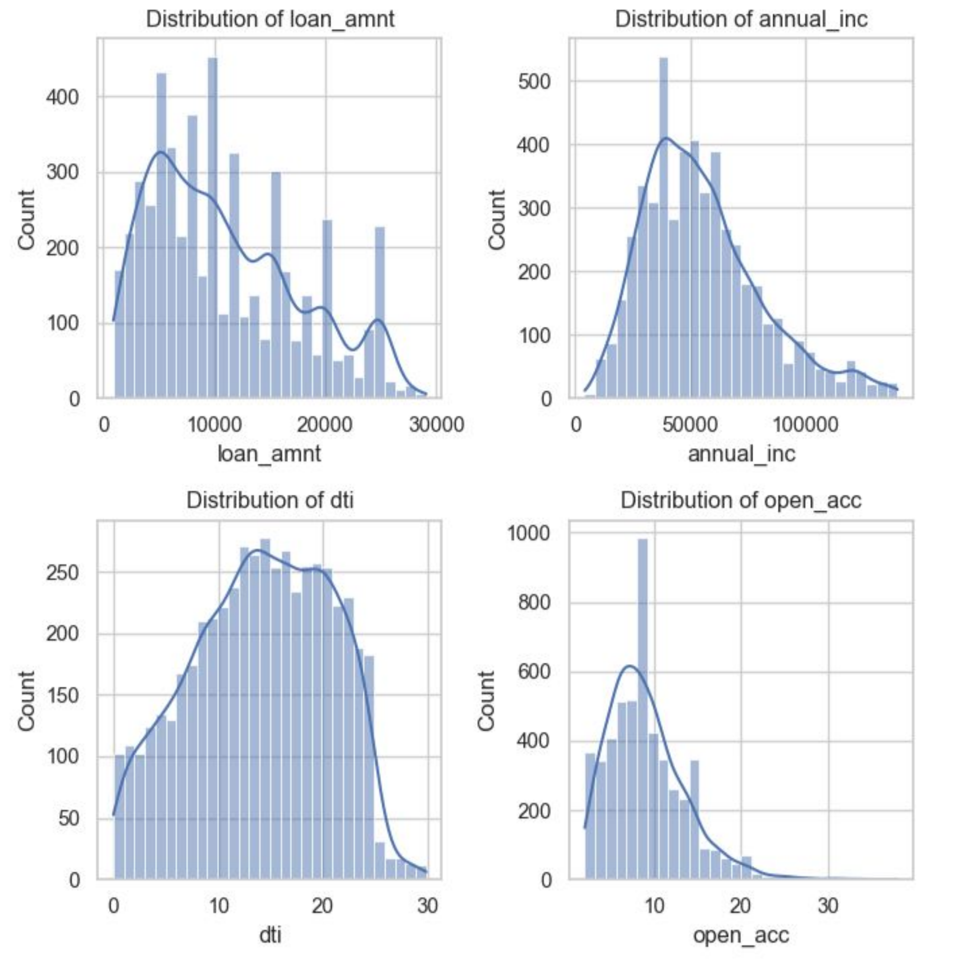




Univariate Analysis for Numerical Columns - Charged off Loans

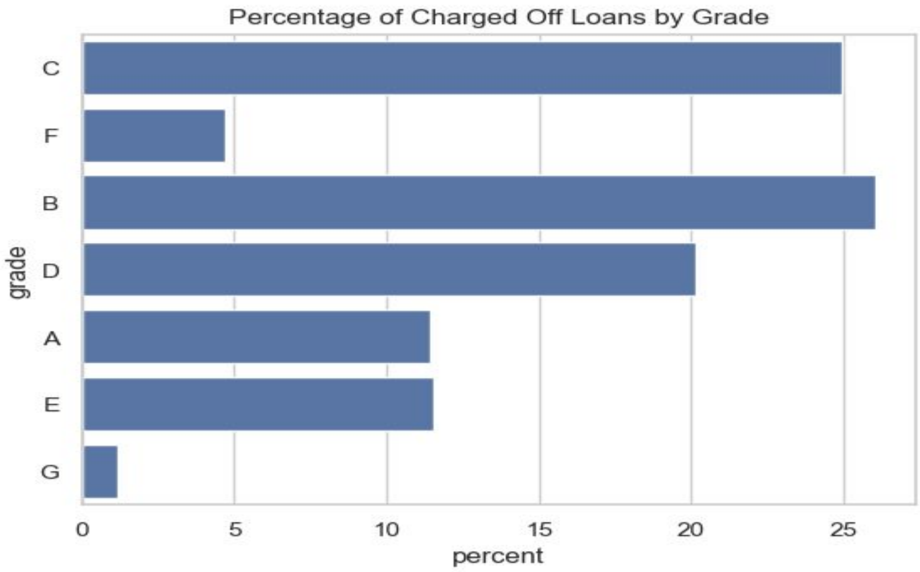
1. The loan amounts are right-skewed, with the most common charged off loan amounts between 5000 and 15000.

2. The dti is in the range of 7 to 24 for charged off loans for most borrowers.

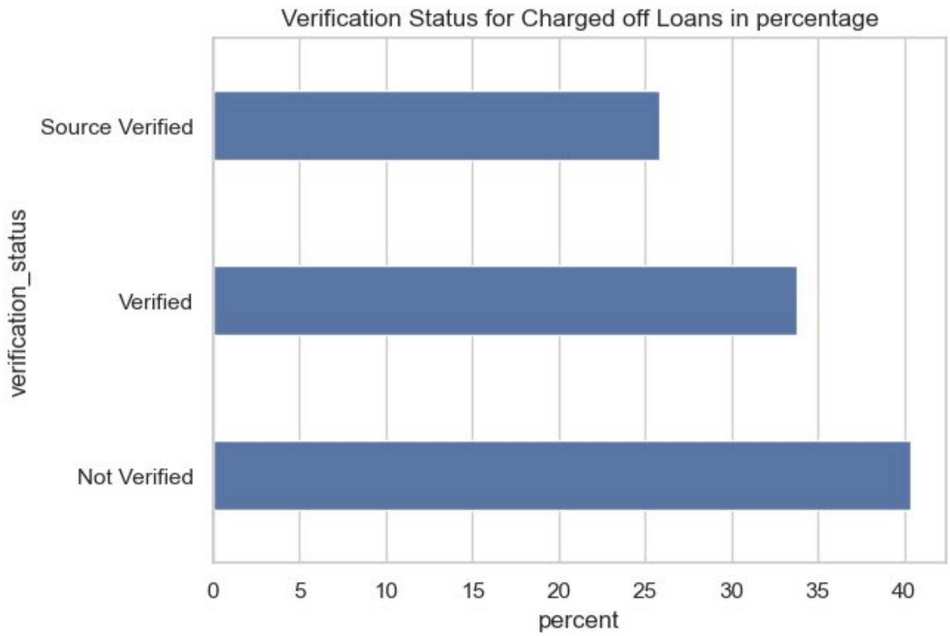


Univariate Analysis for Categorical Columns - Charged off Loans

1. Grades - B (26%), C (24.9%), D (20.14%) contain about 70% of the defaulted loans.
2. About 40% of the defaulted loans are not verified.



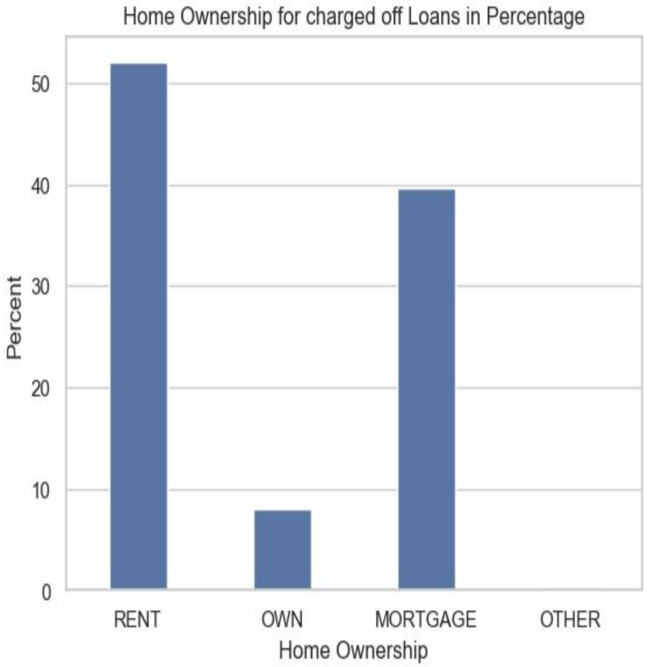

```
Percentage of Charged Off Loans Count by Grade
grade
B     26.044084
C     24.941995
D     20.146945
E     11.562258
A     11.426914
F      4.698376
G      1.179428
```



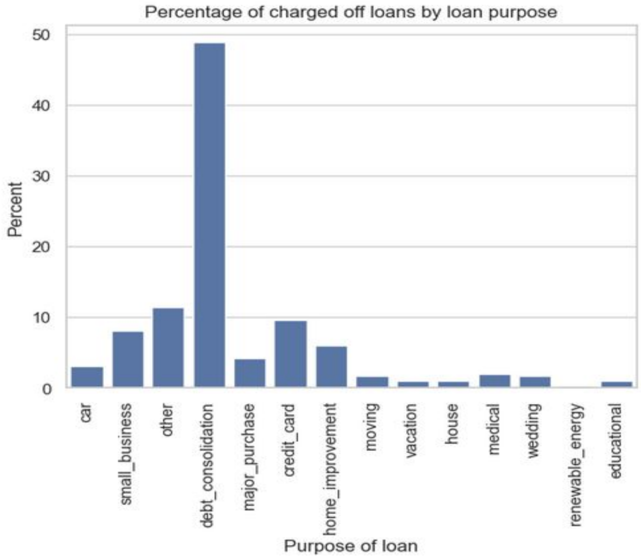

```
Verification Status for Charged off loans in Percentage
verification_status
Not Verified       40.390565
Verified           33.778036
Source Verified    25.831400
```

Univariate Analysis for Categorical Columns - Charged off Loans (Continued.)

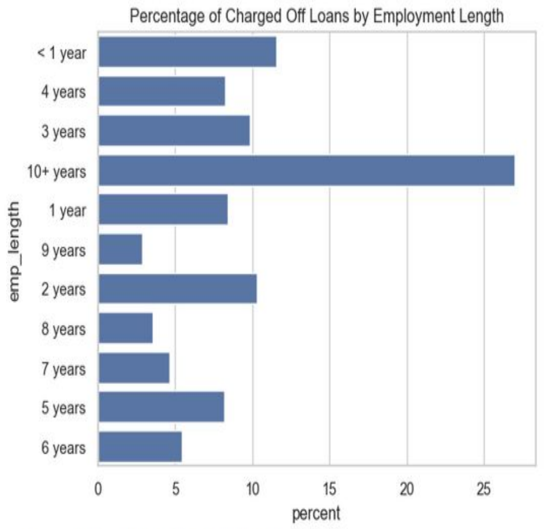
1. Most of the defaulted loans have home ownership as Rent - 52 percent followed by Mortgage - 39.5 percent.
2. The loan purpose of most of the defaulted loans is debt consolidation (48.8 percent) followed by other which is (11.4 percent).
3. Most of the charged off loans have employee length of 10+ years, followed by < 1 year and then 2 years.




```
home_ownership
RENT        52.010828
MORTGAGE    39.597834
OWN          8.062645
OTHER        0.328693
```



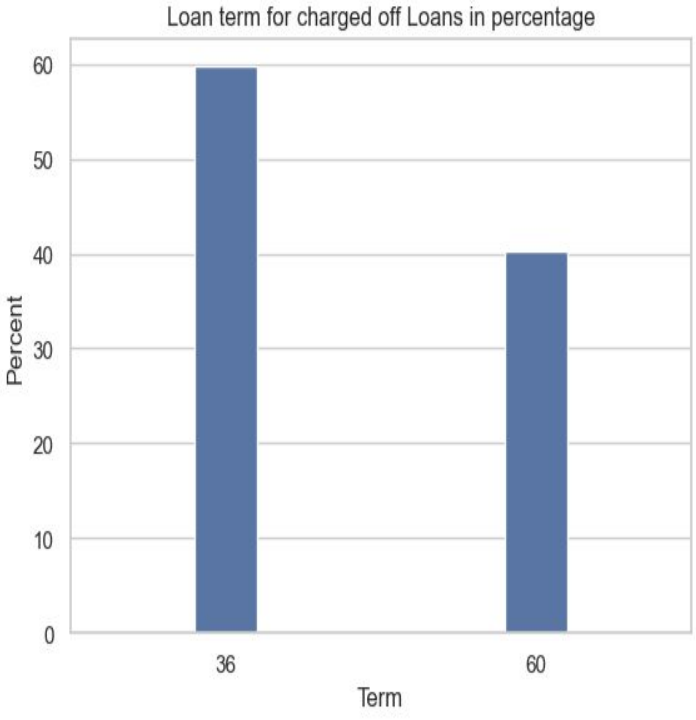

```
purpose
debt_consolidation    48.820572
other                 11.407579
credit_card            9.648105
small_business         8.043310
home_improvement       6.051817
major_purchase         4.156999
car                    3.054911
medical                1.972158
moving                 1.740139
wedding                1.740139
educational            1.044084
vacation               1.005414
house                  0.986079
renewable_energy       0.328693
```



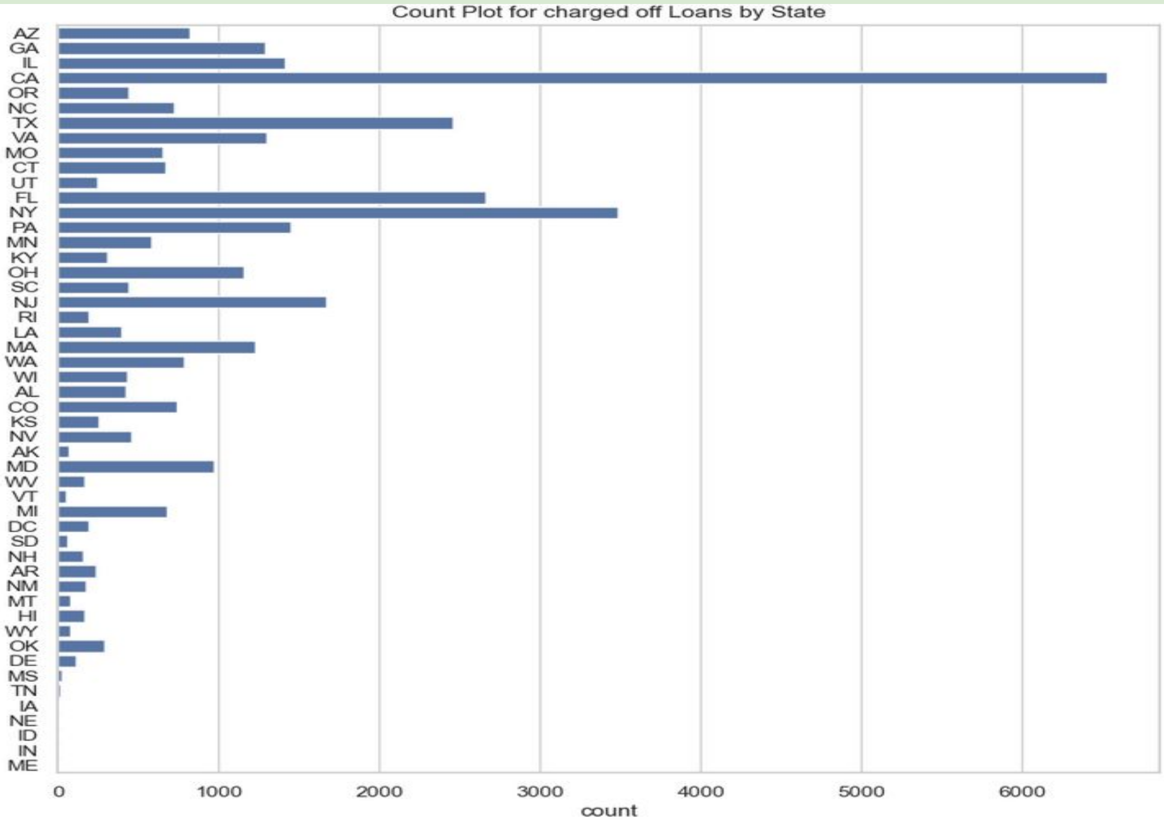

```
Percentage of Charged Off Loans Count by Grade
emp_length
10+ years    26.991493
< 1 year     11.581593
2 years      10.324826
3 years       9.822119
1 year        8.430008
4 years       8.236659
5 years       8.159319
6 years       5.452436
7 years       4.621036
8 years       3.538283
9 years       2.842227
```

Univariate Analysis for Categorical Columns (Continued..)

1. Most of the charged off loan have originated in CA state, followed by NY, FL and TX.
2. Almost 60% charged off loans have 36 months term, followed by 40% for 60 months term.

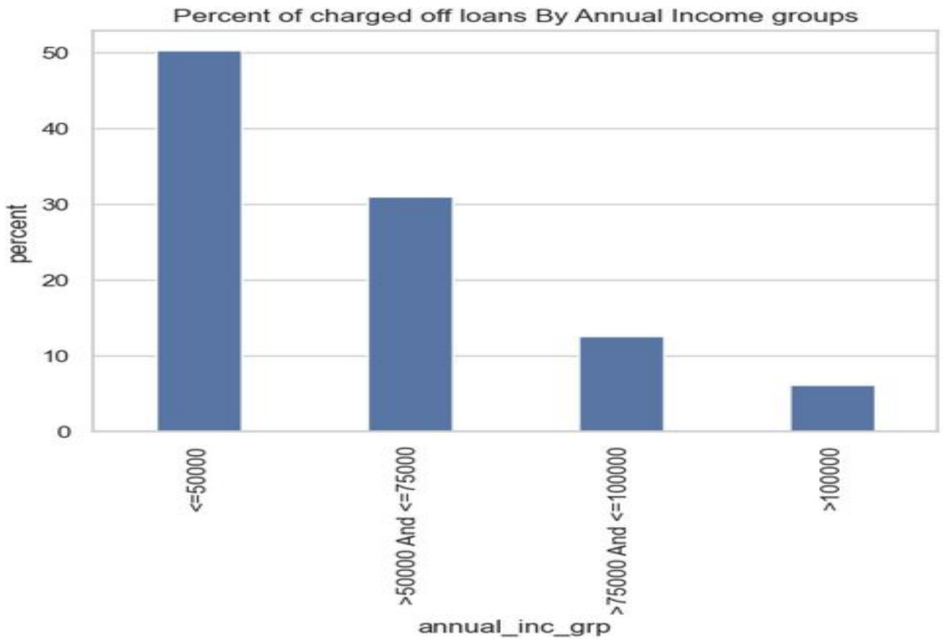


term	
36	59.74478
60	40.25522

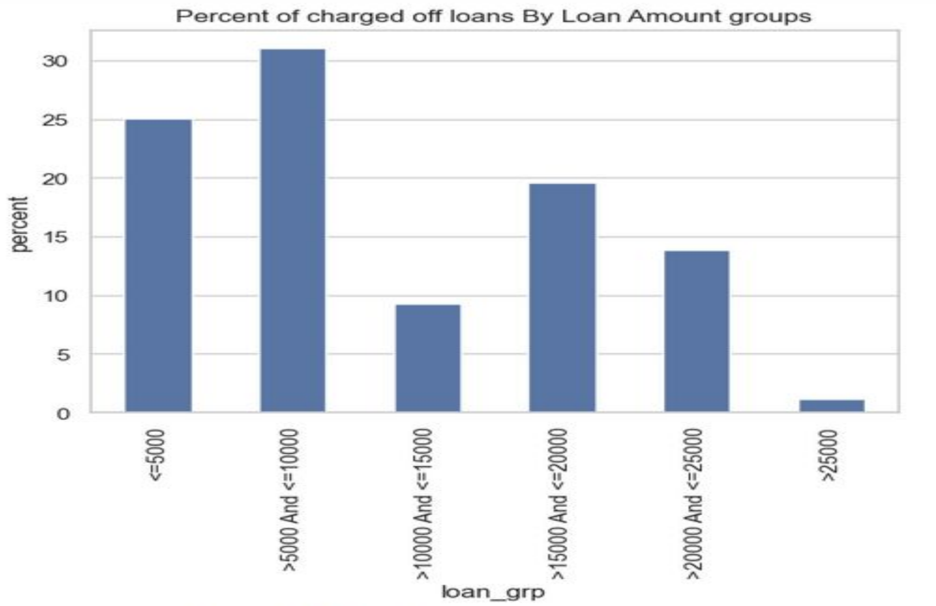


Segmented Univariate Analysis of Loan, Annual Income

1. About 50% of charged off loans have Annual Income <= 50000 and about 31% of charged off loans have Annual Income > 50000 And <= 75000.
2. About 56% of the loans have loan amount <=10000. 31% of the which have loan amount > 5000 And <=10000, followed by 25% where loan amount <= 5000.



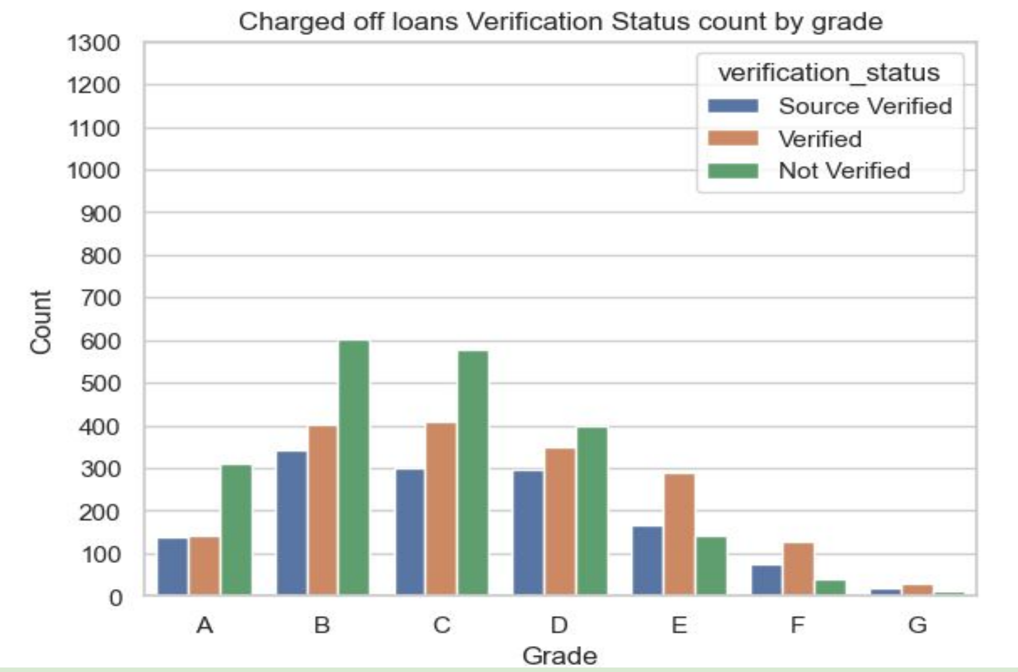
	annual_inc_grp	Count	Percentage
0	<=50000	2603	50.328693
1	>100000	318	6.148492
2	>50000 And <=75000	1604	31.013148
3	>75000 And <=100000	647	12.509667

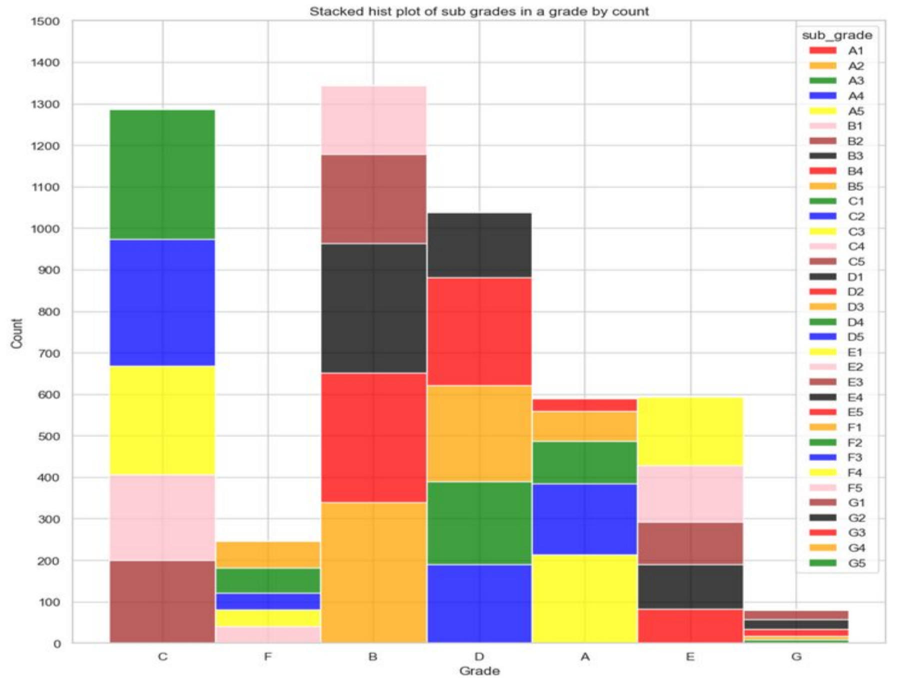


	loan_grp	Count	Percentage
0	<=5000	1298	25.096674
1	>10000 And <=15000	1015	19.624903
2	>15000 And <=20000	715	13.824439
3	>20000 And <=25000	478	9.242073
4	>25000	61	1.179428
5	>5000 And <=10000	1605	31.032483

Bivariate Analysis - Charged off Loans

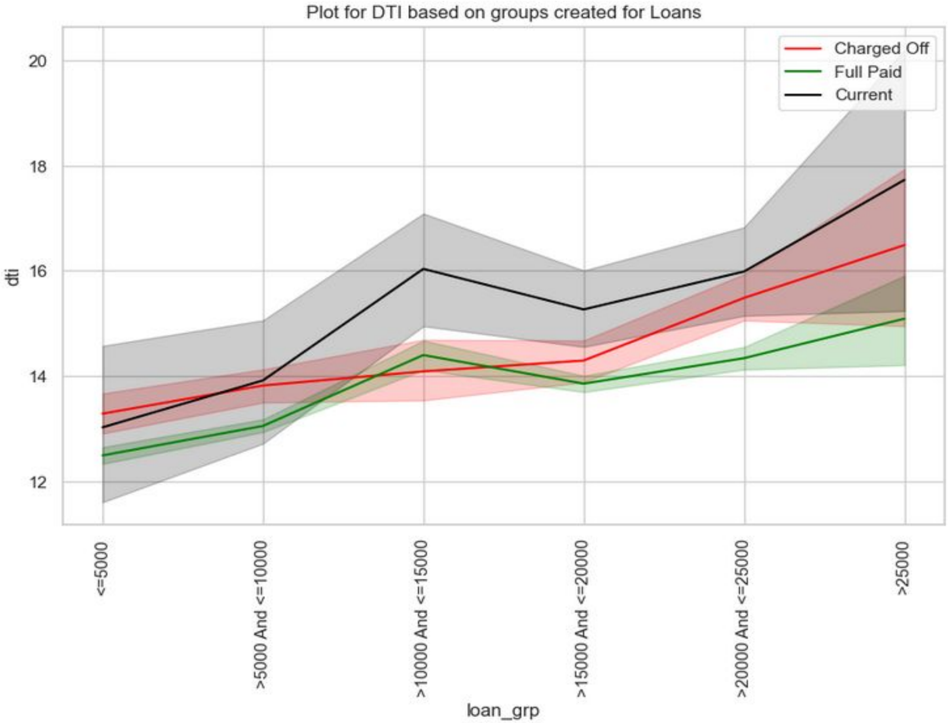
1. Most of the B Grade default loans are not verified , followed by C and D grade loans.
2. Verified defaulted loans are almost same for B and C grade loans.
3. Source verified defaulted loans is almost same for C and D grade loans.
4. There are more B5 subgrade loans, followed by B3, B4, B2, B1 in B grade loans.
5. Similarly there are more C1 subgrade loans followed by C2, C3,C4, C5 in C grade loans.
6. Similarly there are more D2 subgrade loans followed by D3,D4, D5, D1 in D grade loans.

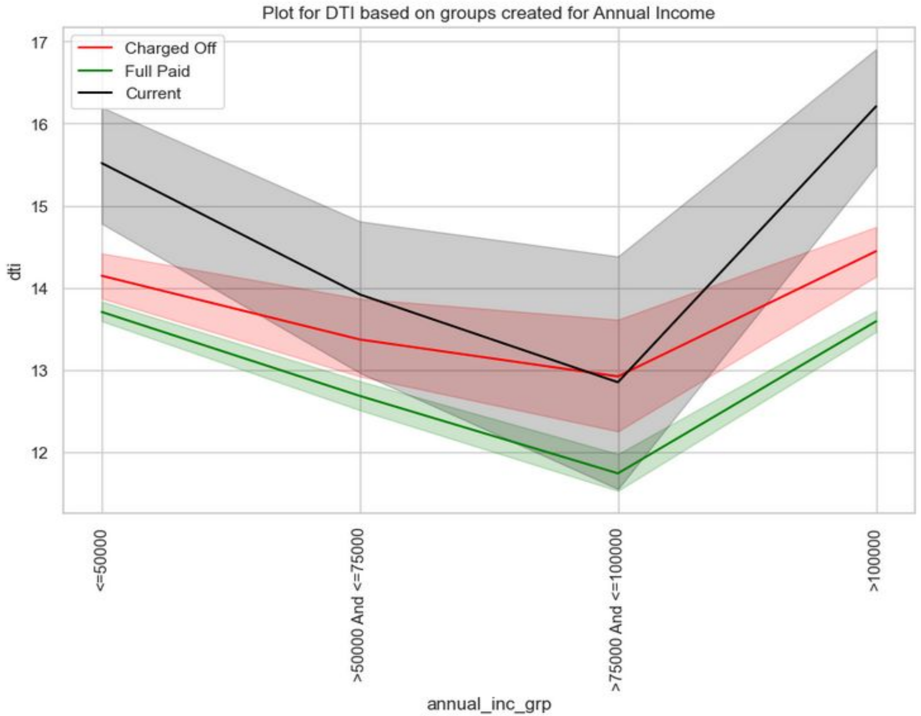




Bivariate Analysis

1. Based on the plot for DTI vs Loan Group by Loan Status, certain current loans greater than 5000 have higher DTI compared to the charged off loans and might default, and need to be evaluated further using other factors.
2. Based on the plot for DTI vs Annual Income Group by Loan Status, certain current loans might default when compared to charged off loans , and need to be evaluated further using other factors.





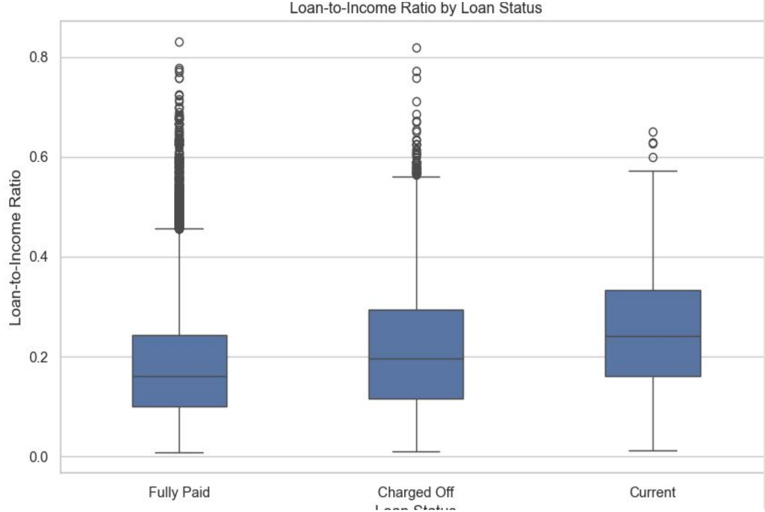
Bivariate Analysis - Derived Metrics

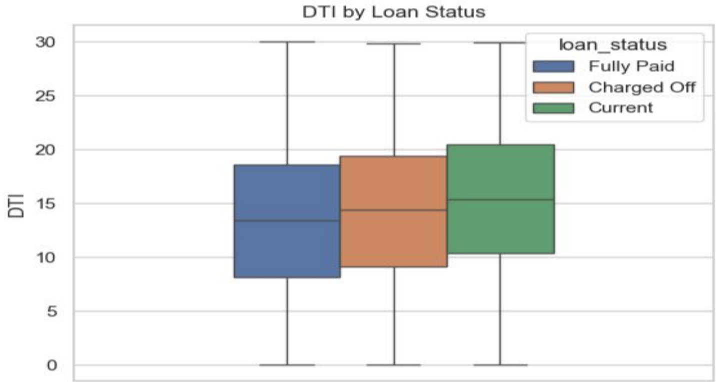
1. The DTI for Current loans is greater than Charged Off loans, which in turn is greater than Fully Paid loans.
   DTI for Current Loans > Charged off Loans > Fully-Paid Loans

2. The median loan-to-income ratio for borrowers whose loans were charged-off is around 20%. The distribution is spread out with some borrowers having a ratio as high as 60%, suggesting that higher loan-to-income ratios correlate with a higher risk of default.

3. A loan-to-income ratio greater than 40% is associated with a 25.2% default rate (as shown by the proportion of "Charged Off" in the data). This suggests that when loan amounts represent a large portion of the borrower’s income, the risk of default increases significantly

```
Default Rate for Borrowers with Loan-to-Income Ratio > 40%:
loan_status
Fully Paid     0.681485
Charged Off    0.250000
Current        0.068515
```





Multivariate Analysis of all Loan status

Do a group by on multiple columns with aggregation on multiple columns like int rate , loan amount , annual income and dti for Below is the sample data for charged-0ff loans and current loans

			int_rate		loan_amnt		annual_inc		dti	
loan_status	verification_status	emp_length	median	count	median	count	median	count	median	count
Charged Off	Not Verified	1 year	12.835	204	7500.0	204	45000.0	204	13.575	204
		10+ years	12.195	500	7350.0	500	51000.0	500	14.105	500
		2 years	12.710	226	7000.0	226	42000.0	226	14.655	226
		3 years	13.450	208	7800.0	208	45000.0	208	15.090	208
		4 years	12.870	174	7337.5	174	49600.0	174	14.440	174
		5 years	13.110	171	8000.0	171	50000.0	171	15.020	171
		6 years	12.690	105	7750.0	105	47000.0	105	15.600	105
		7 years	12.895	98	10000.0	98	54000.0	98	16.355	98
		8 years	12.490	75	8000.0	75	50000.0	75	13.490	75
		9 years	12.870	61	9800.0	61	53000.0	61	13.280	61
		< 1 year	12.530	267	6825.0	267	39000.0	267	13.100	267
	Source Verified	1 year	13.295	116	8000.0	116	42000.0	116	14.155	116
		10+ years	13.800	310	10000.0	310	59464.0	310	13.275	310
		2 years	14.270	147	8000.0	147	40000.0	147	12.870	147
		3 years	14.270	149	8000.0	149	50000.0	149	13.140	149
		4 years	14.460	109	9000.0	109	45000.0	109	12.480	109
		5 years	13.800	108	8000.0	108	47700.0	108	13.095	108
		6 years	14.960	77	10000.0	77	55000.0	77	14.030	77
		7 years	13.230	63	9600.0	63	45000.0	63	14.740	63
		8 years	13.645	38	12000.0	38	55154.0	38	16.115	38
		9 years	14.925	38	9800.0	38	54500.0	38	14.160	38
		< 1 year	13.490	181	7000.0	181	40000.0	181	11.900	181
	Verified	1 year	14.050	116	14675.0	116	51000.0	116	15.130	116
		10+ years	14.270	586	15850.0	586	60000.0	586	16.140	586
		2 years	14.420	161	12375.0	161	53000.0	161	14.060	161
		3 years	14.110	151	12000.0	151	55000.0	151	14.310	151
		4 years	15.050	143	13000.0	143	54500.0	143	13.710	143
		5 years	14.170	143	15000.0	143	59000.0	143	17.000	143
		6 years	14.270	100	15125.0	100	60000.0	100	15.445	100
		7 years	14.530	78	15000.0	78	50000.0	78	15.755	78
		8 years	14.220	70	15000.0	70	58865.0	70	16.685	70
		9 years	13.700	48	14750.0	48	61500.0	48	16.975	48
		< 1 year	13.800	151	13750.0	151	50960.0	151	15.940	151
Current	Not Verified	1 year	13.490	20	12000.0	20	56000.0	20	16.305	20
		10+ years	12.555	78	12000.0	78	54600.0	78	15.860	78
		2 years	12.690	19	10000.0	19	53640.0	19	16.830	19
		3 years	13.490	19	12000.0	19	52000.0	19	11.400	19
		4 years	13.025	20	10937.5	20	46500.0	20	15.005	20
		5 years	12.990	12	12000.0	12	51500.0	12	14.410	12
		6 years	10.990	11	12000.0	11	57000.0	11	13.720	11
		7 years	12.840	14	12000.0	14	49572.0	14	14.840	14
		8 years	15.270	19	12000.0	19	56004.0	19	14.350	19
		9 years	15.990	7	13000.0	7	58000.0	7	15.330	7
		< 1 year	17.380	4	8362.5	4	31500.0	4	11.235	4
	Source Verified	1 year	15.990	23	12400.0	23	56000.0	23	13.340	23
		10+ years	13.490	74	14000.0	74	73000.0	74	13.745	74
		2 years	16.490	23	13650.0	23	48000.0	23	12.620	23
		3 years	16.240	24	12000.0	24	46200.0	24	10.735	24
		4 years	15.960	25	12000.0	25	51000.0	25	10.470	25
		5 years	14.650	23	12000.0	23	53100.0	23	10.770	23
		6 years	15.615	14	7637.5	14	48500.0	14	12.300	14
		7 years	17.490	15	16000.0	15	56004.0	15	13.090	15
		8 years	15.990	5	14000.0	5	48000.0	5	15.170	5
		9 years	15.490	8	13500.0	8	74700.0	8	16.130	8
		< 1 year	15.990	33	12000.0	33	50000.0	33	12.000	33
	Verified	1 year	15.030	18	15600.0	18	57500.0	18	17.140	18
		10+ years	14.270	188	20000.0	188	64295.0	188	16.840	188
		2 years	16.770	41	18550.0	41	55000.0	41	18.040	41
		3 years	15.390	22	18350.0	22	61500.0	22	18.190	22
		4 years	15.230	39	20000.0	39	60000.0	39	16.200	39
		5 years	13.880	38	20000.0	38	61050.0	38	17.335	38
		6 years	17.080	28	16650.0	28	65000.0	28	14.760	28
		7 years	14.460	28	19975.0	28	60000.0	28	14.270	28
		8 years	16.770	11	19000.0	11	56500.0	11	20.730	11
		9 years	12.205	12	16800.0	12	64450.0	12	18.705	12
		< 1 year	16.490	27	16000.0	27	68000.0	27	18.140	27

Key Observation - Based on MultiVariate Analysis

Loan Status: Charged Off

1. **Not Verified:**
   a) Borrowers with 3-7 years of employment generally had higher DTI values, with higher interest rates and loans between and loans between 5000 and 15000 will most likely default.

2. **Source Verified:**
   a) Borrowers with longer employment lengths (10+ years) received higher loan amounts (median 10000), with higher DTI. This suggests that the verification process allows lenders to extend higher risk loans.

   b) Borrowers with shorter employment (less than or equal to 1 year) had lower DTI values and lower loan amounts, but still defaulted at a higher rate.

3. **Verified:**
   a) Borrowers with verified income generally received higher loan amounts, with loans increasing to 15000 to 20000 for those employed from 1 year to 10+ years.

   b) Their DTIs were higher for all employer lengths.

   c) These verified borrowers were still charged higher interest rates for their loan amounts ranging from 10000 to 15000 in terms of median.

Key Observation - Based on MultiVariate Analysis (Continued ..)

Loan Status: Current

1. Borrowers in the "Current" loan status, particularly those with verified status, tended to have higher DTI ratios, suggesting they are likely to experience financial difficulties.

2. Borrowers in the "Current" loan status, particularly those with source verified status have higher DTI ratios and higher interest rate for 8 years employee length who may default as their annual income is less and loan amount is high.

3. Borrowers with Verified incomes, especially those employed for 1-7 years, had high loan amounts (15000 to 20000) and relatively higher interest rates. These individuals may be under financial pressure, but they have not yet defaulted.

Loan Status: Fully Paid

1. Not Verified: Borrowers in this group had lower interest rates across all employment lengths and tended to have lower DTI ratios, particularly those with shorter employment lengths. These borrowers were more likely to successfully repay their loans.

2. Source Verified and Verified: Borrowers with verified income typically received higher loan amounts, but the interest rates were relatively low. As their annual income is high, their DTI ratios remained manageable, making them less likely to default

Key Insights and Risk Factors:

Interest Rate and Loan Amount: 1.Borrowers who were verified or source verified tended to receive higher loan amounts across all employment lengths, with higher interest rates if emp length is 1 to 9 years.

Debt-to-Income Ratio:1.Borrowers with higher DTI ratios (above 15%) are at greater risk of default, especially if their income is not verified or if they have shorter employment histories.
2.Borrowers in the Current or Charged Off categories consistently had higher DTI ratios than those who successfully repaid their loans.

Employment Length: 1.Borrowers with 3-7 years of employment face a higher risk of default, particularly when their loan amounts and DTI ratios are higher. While they are offered relatively high loan amounts, they still tend to default at higher rates compared to longer-term employees (10+ years).
2.Longer employment lengths (10+ years) are generally associated with lower interest rates and more manageable DTI ratios, leading to a higher likelihood of loan repayment.

Loan To Income: 1.A loan-to-income ratio greater than 40% is associated with a 25.2% default rate in "Charged Off" loan data. This suggests that when loan amounts represent a large portion of the borrower's income, the risk of default increases significantly

Recommendations:

Stricter Terms for Borrowers with High DTI:

1. Borrowers with high DTI ratios (above 15%) should either receive smaller loan amounts or be charged higher interest rates to compensate for the increased risk.
2. Borrowers with higher DTI ratios can be given loan amounts if their annual income is high
3. Borrowers with unverified incomes and high DTI ratios should be flagged as high risk, and stricter loan approval criteria should be applied.

Employment Length Consideration:

1. Lenders should be cautious when lending to borrowers with shorter employment histories (less than 3 years) who have high DTI ratios. Offering smaller loan amounts or additional verification could reduce the risk of default.
2. Borrowers with 3-7 years of employment should be monitored more closely, as they tend to default more often when offered higher loan amounts, despite their relatively longer employment.

Income Verification:

1. Verifying income, especially for borrowers with high loan amounts and high DTI ratios, is crucial to reducing default risk.
2. Verified borrowers tend to perform better even when offered larger loans, but additional verification could help further mitigate risks.

By using these insights, lenders can improve their loan approval process, reducing the likelihood of defaults while still offering competitive loan products to qualified borrowers.