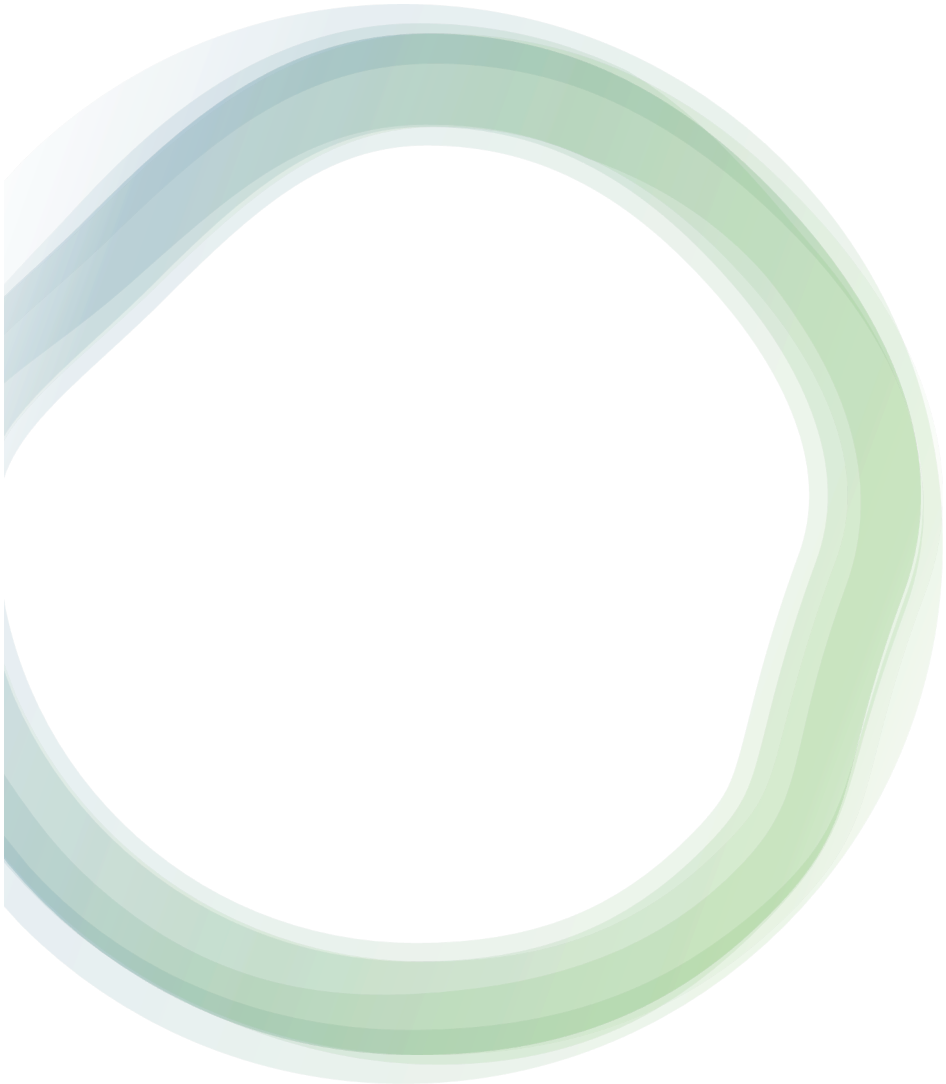
# Lending Club Case Study
# EDA

Naresh Padiyar



- **Problem Statement**
- **Data Summary**
- **Data Cleaning**
- **Data conversions and Derived Columns**
- **Dropping / Imputing the Rows / Outliers**
- **Univariate Analysis**
- **Bivariate Analysis**
- **Multivariate Analysis**
- **Correlations**
- **Summary**

# Problem Statement

## Problem

You work for a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. 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

## Objective

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.

## Constraints

When a person applies for a loan, there are two types of decisions that could be taken by the company:

- **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:
  - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
  - **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
  - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
- **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.).

# Dataset Summary

- ✓ The **Loan.csv** file contains **111 columns and 39717 Rows**.
- ✓ The dataset contains **Loan Attributes and Customer Attributes**.
- ✓ The **Data_Dictionary.xls** file provided along with the dataset, has a description of each column of the **Loan.csv** file and was used to make a decision if the column would contribute to the EDA analysis.

# Data Cleaning

- ✓ There are **54** columns which have all rows containing NA/Null values. These columns were removed/dropped.
- ✓ Columns **id, member_id and url** have unique value in every row so these 3 columns were also dropped.
- ✓ Columns **desc, title and emp_title** contain descriptions so these columns were dropped as they won't contribute to the EDA analysis.
- ✓ Columns **pymnt_plan, initial_list_status, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt and tax_liens** have only 1 unique value and hence were dropped as they won't contribute to the EDA analysis.
- ✓ Columns **mths_since_last_delinq, mths_since_last_record, and next_pymnt_d** were dropped as they were having more than **50%** NA/Null Rows.
- ✓ Columns **delinq_2yrs, earliest_cr_line, last_pymnt_amnt, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_credit_pull_d** were dropped as these columns either represent behavioural data or post loan approval data, hence these columns were dropped as they won't contribute to the EDA analysis.

# Data Conversion and Derived Columns

- ✓ Additional string value in **term** column has been trimmed and column has been converted to int data type.
- ✓ Additional string value in **int_rate** column has been trimmed and column has been converted to int data type.
- ✓ Columns **loan_amnt** and **funded_amnt** have been converted to int data type.
- ✓ Column **issue_d** converted to datetime data type.
- ✓ Created derived columns **issue_month** and **issue_year** from **issue_d** column.
- ✓ Created derived columns **loan_amnt_bucket, annual_income_bucket, int_rate_bucket, dti_bucket, installment_bucket and funded_amnt_inv_bucket** from columns **loan_amnt, annual_income, int_rate, dti, installment and funded_amnt_inv** respectively for **bucketing** purpose.

# Dropping / Imputing Rows / Outliers

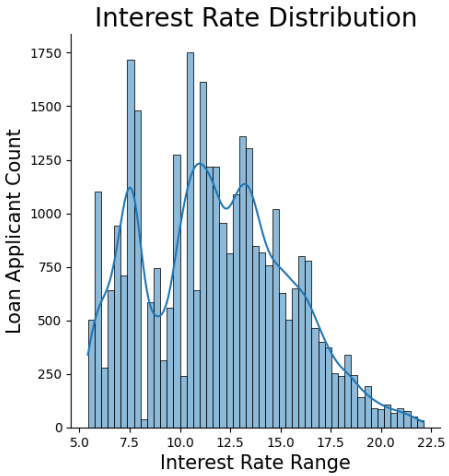
- ✓ Since the goal of the EDA is to identify who is likely to default and this can only be said in case of either fully paid or charged off loans, hence we dropped all rows where **loan_status** is equal to "Current"

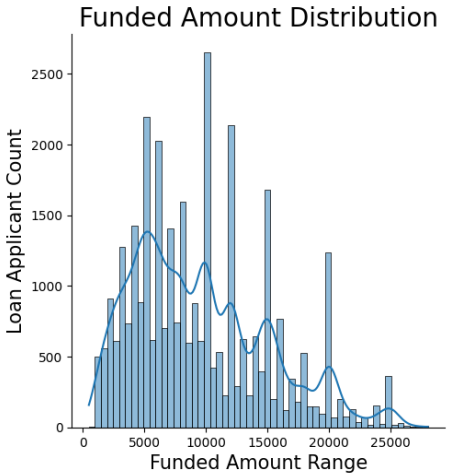
- ✓ **verification_status** column has 3 categories **Not Verified, Verified and Source Verified**. **Verified and Source Verified** both mean the same thing, hence we have replaced the value **Source Verified to Verified**. This leaves us with only 2 categories in the **verification_status** column, i.e. **Not Verified and Verified**

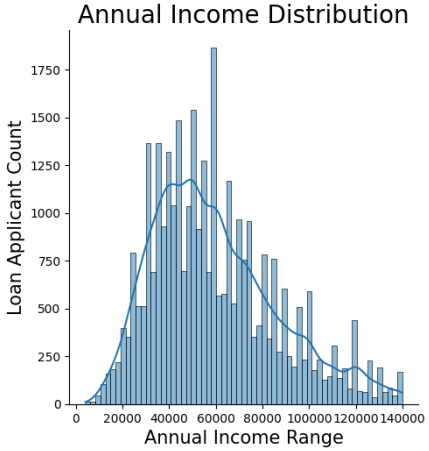
- ✓ Columns **emp_length and pub_rec_bankruptcies** both contains 2.67% and 1.80% of rows as null respectively. We have dropped the rows where the value for these columns was NA/Null.

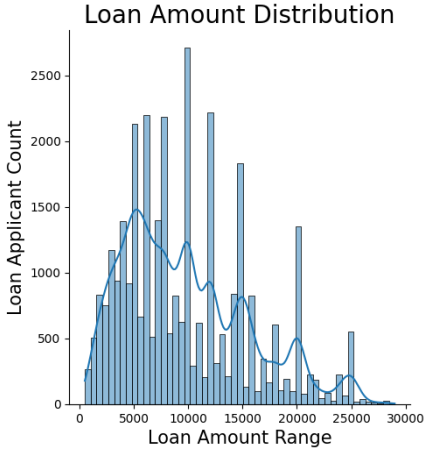
- ✓ Outliers exits for numeric data **loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment and annual_inc.** We have treated the outliers in these columns using the using quantile mechanism.

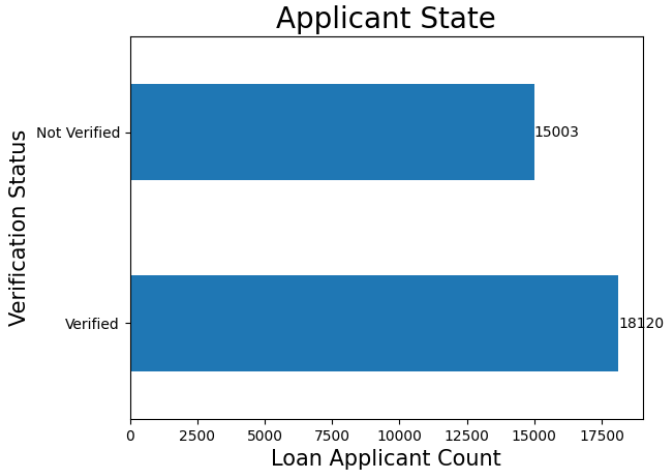
# Univariate Analysis

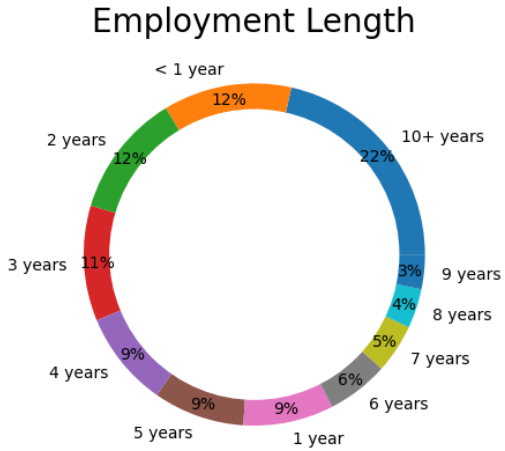

Interest Rate Distribution
Loan Applicant Count
Interest Rate Range
0
250
500
750
1000
1250
1500
1750
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

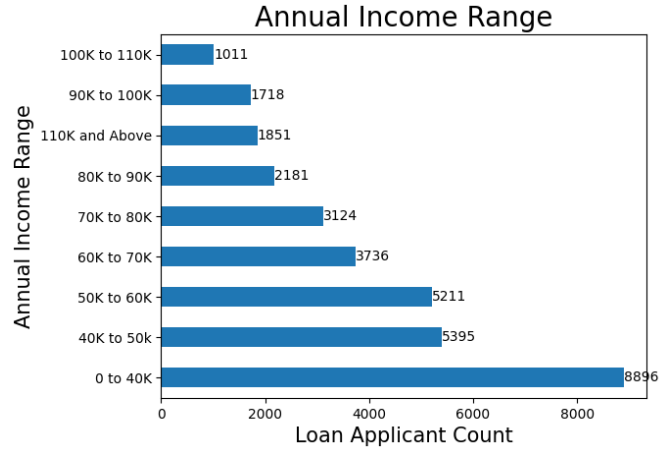

Funded Amount Distribution
Loan Applicant Count
Funded Amount Range
0
500
1000
1500
2000
2500
0
5000
10000
15000
20000
25000

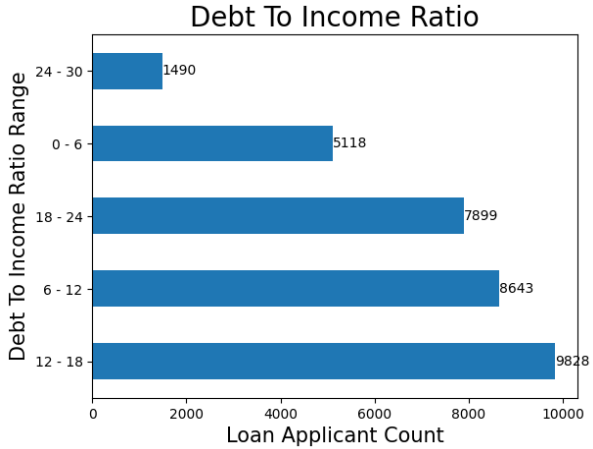

Annual Income Distribution
Loan Applicant Count
Annual Income Range
0
250
500
750
1000
1250
1500
1750
0
20000
40000
60000
80000
100000
120000
140000

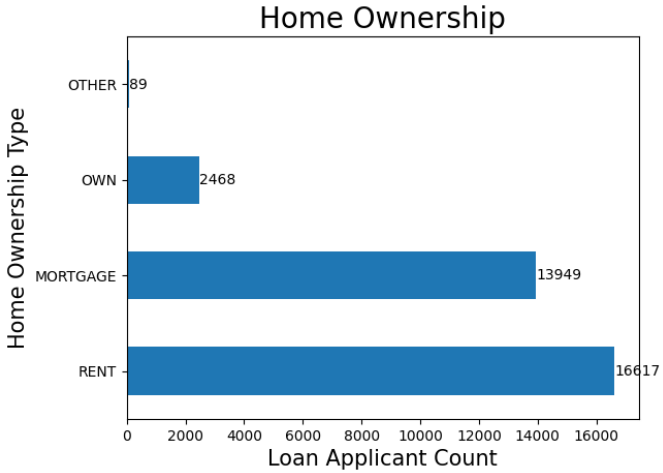

Loan Amount Distribution
Loan Applicant Count
Loan Amount Range
0
500
1000
1500
2000
2500
0
5000
10000
15000
20000
25000
30000

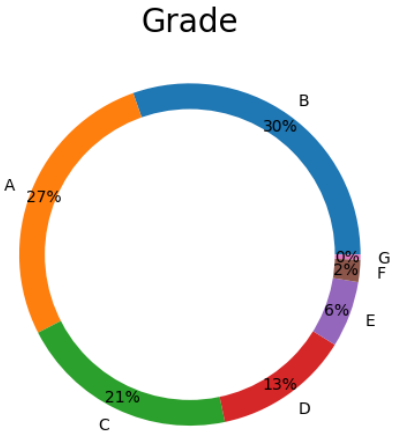

Applicant State
Verification Status
Not Verified
Verified
15003
18120
0
2500
5000
7500
10000
12500
15000
17500
Loan Applicant Count

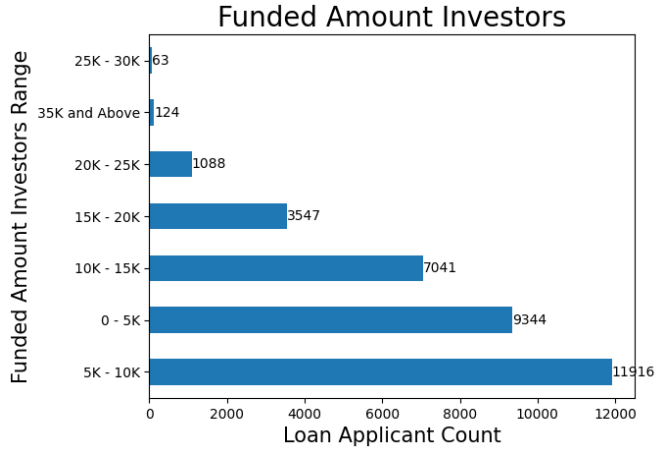

Employment Length
< 1 year
12%
10+ years
22%
2 years
12%
3 years
11%
9 years
3%
8 years
4%
7 years
5%
6 years
6%
1 year
9%
5 years
9%
4 years
9%

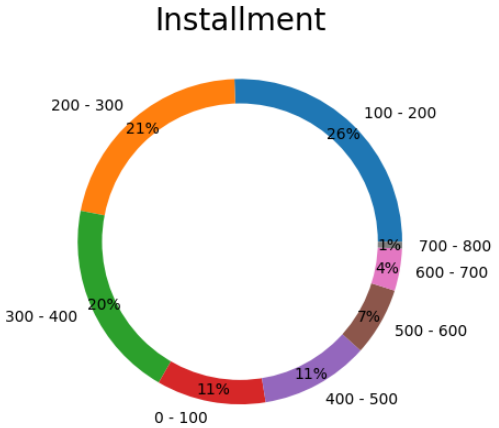

Annual Income Range
Annual Income Range
100K to 110K
1011
90K to 100K
1718
110K and Above
1851
80K to 90K
2181
70K to 80K
3124
60K to 70K
3736
50K to 60K
5211
40K to 50k
5395
0 to 40K
8896
0
2000
4000
6000
8000
Loan Applicant Count

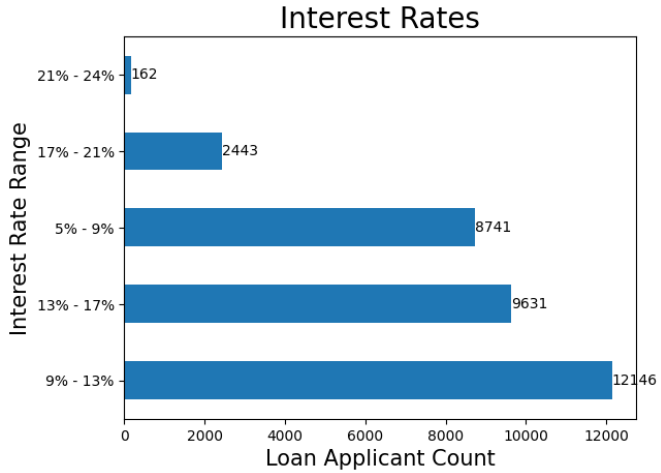

Debt To Income Ratio
Debt To Income Ratio Range
24 - 30
1490
0 - 6
5118
18 - 24
7899
6 - 12
8643
12 - 18
9828
0
2000
4000
6000
8000
10000
Loan Applicant Count

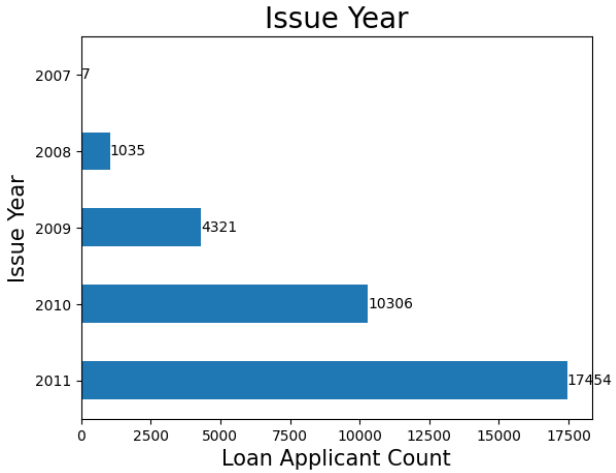

Home Ownership
Home Ownership Type
OTHER
OWN
MORTGAGE
RENT
89
2468
13949
16617
0
2000
4000
6000
8000
10000
12000
14000
16000
Loan Applicant Count

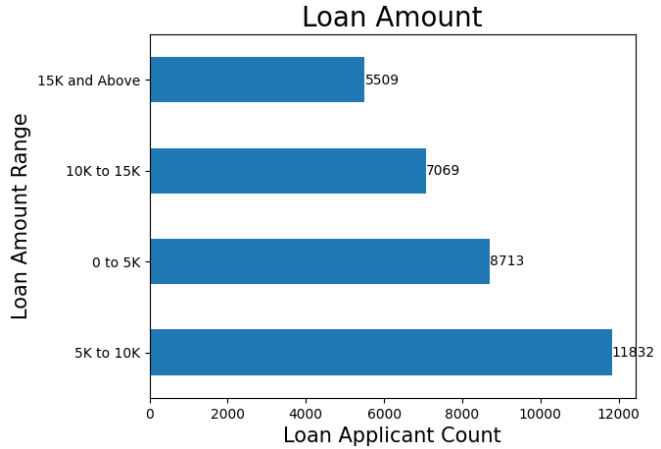

Grade
B
30%
A
27%
G
0%
F
2%
E
6%
D
13%
C
21%

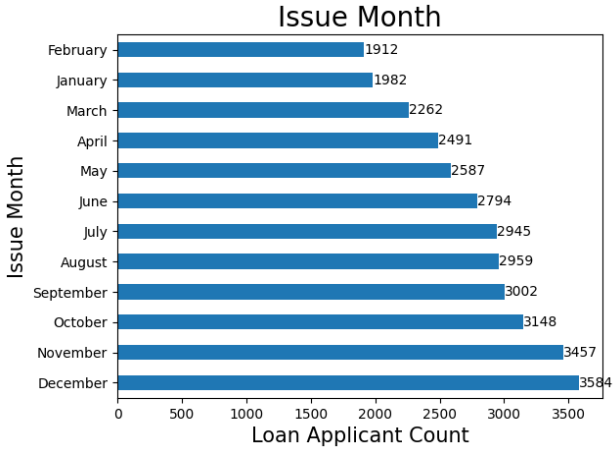

Funded Amount Investors
Funded Amount Investors Range
25K - 30K
35K and Above
20K - 25K
15K - 20K
10K - 15K
0 - 5K
5K - 10K
63
124
1088
3547
7041
9344
11916
0
2000
4000
6000
8000
10000
12000
Loan Applicant Count

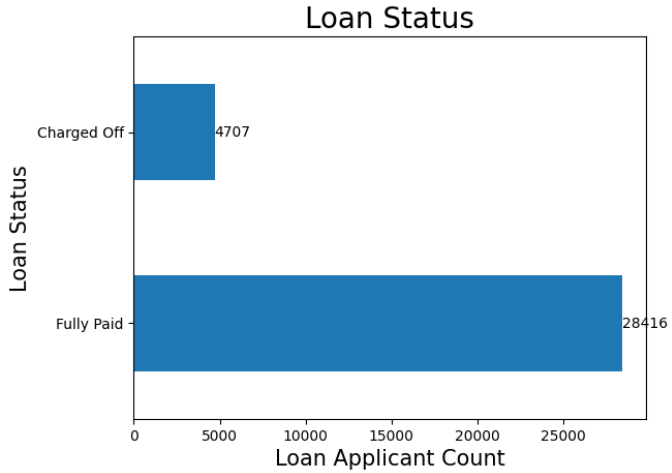

Installment
100 - 200
26%
200 - 300
21%
700 - 800
1%
600 - 700
4%
300 - 400
20%
500 - 600
7%
400 - 500
11%
0 - 100
11%

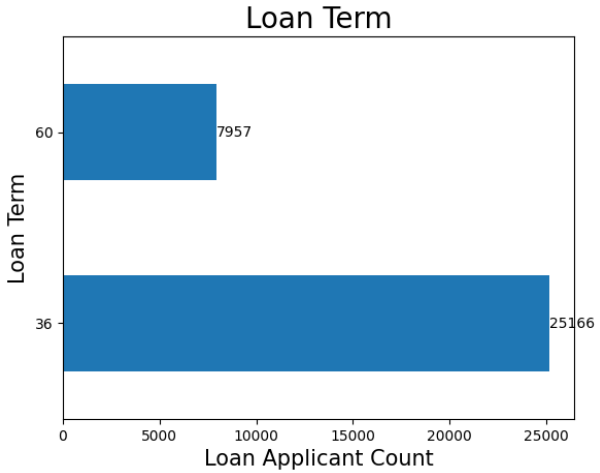

Interest Rates
Interest Rate Range
21% - 24%
17% - 21%
5% - 9%
13% - 17%
9% - 13%
162
2443
8741
9631
12146
0
2000
4000
6000
8000
10000
12000
Loan Applicant Count

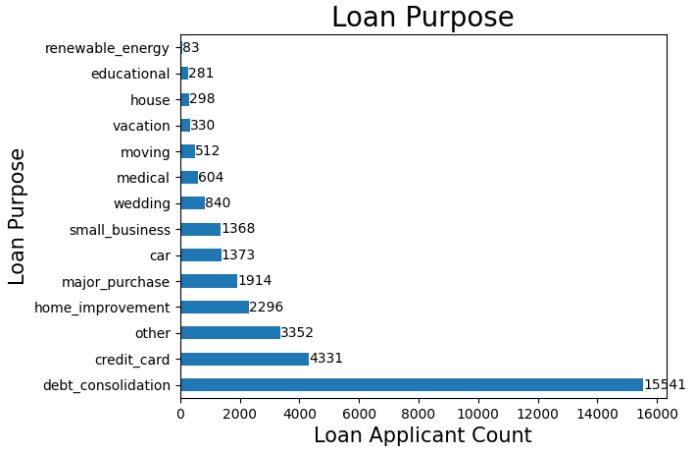

Issue Year
Issue Year
2007
2008
2009
2010
2011
7
1035
4321
10306
17454
0
2500
5000
7500
10000
12500
15000
17500
Loan Applicant Count

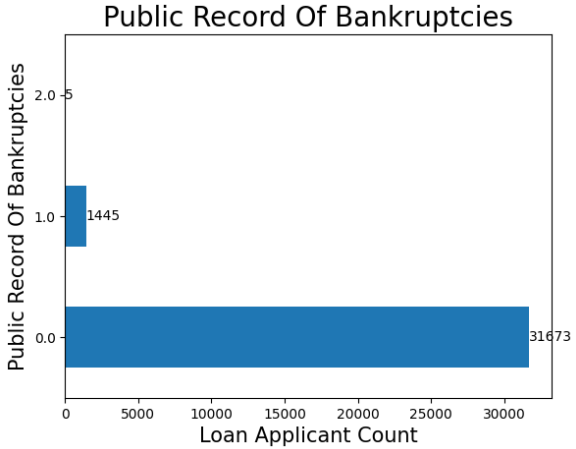

Loan Amount
Loan Amount Range
15K and Above
10K to 15K
0 to 5K
5K to 10K
5509
7069
8713
11832
0
2000
4000
6000
8000
10000
12000
Loan Applicant Count


Issue Month
Issue Month
February
January
March
April
May
June
July
August
September
October
November
December
1912
1982
2262
2491
2587
2794
2945
2959
3002
3148
3457
3584
0
500
1000
1500
2000
2500
3000
3500
Loan Applicant Count


Loan Status
Charged Off
Fully Paid
4707
28416
Loan Status
0
5000
10000
15000
20000
25000
Loan Applicant Count


Loan Term
60
36
7957
25166
Loan Term
0
5000
10000
15000
20000
25000
Loan Applicant Count

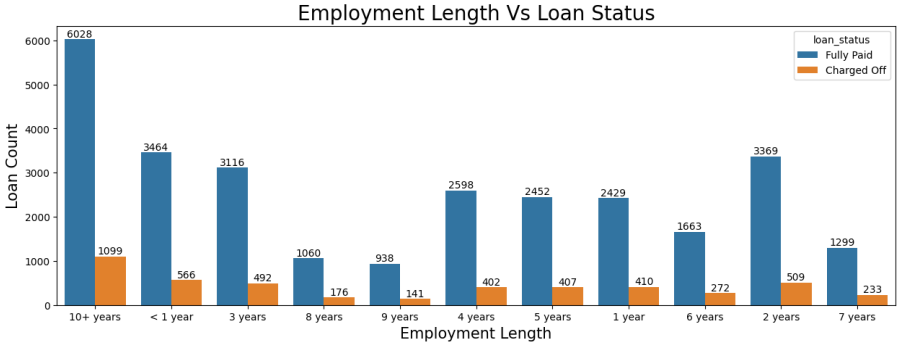

Loan Purpose
renewable_energy
educational
house
vacation
moving
medical
wedding
small_business
car
major_purchase
home_improvement
other
credit_card
debt_consolidation
83
281
298
330
512
604
840
1368
1373
1914
2296
3352
4331
15541
Loan Purpose
0
2000
4000
6000
8000
10000
12000
14000
16000
Loan Applicant Count

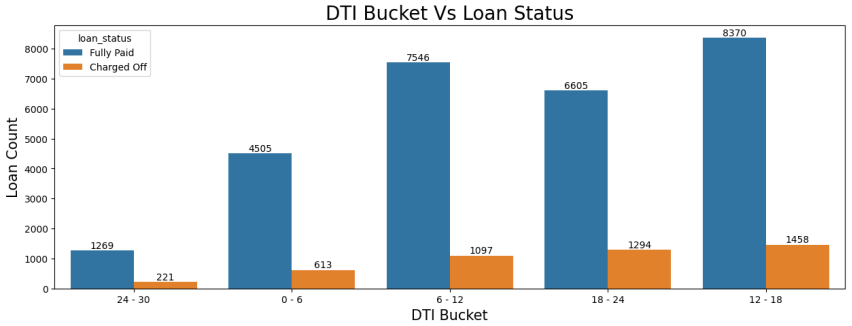

Public Record Of Bankruptcies
2.0
1.0
0.0
5
1445
31673
Public Record Of Bankruptcies
0
5000
10000
15000
20000
25000
30000
Loan Applicant Count

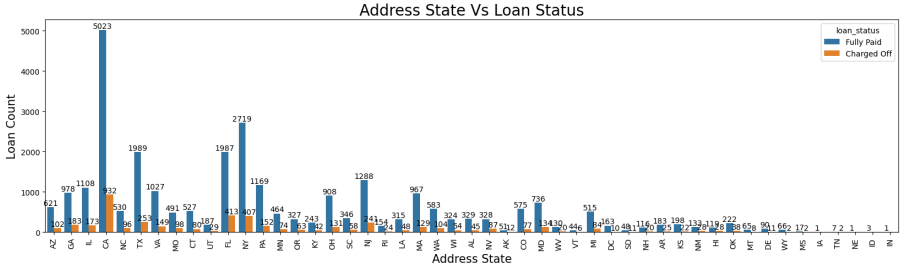
# Conclusion from Univariate Analysis

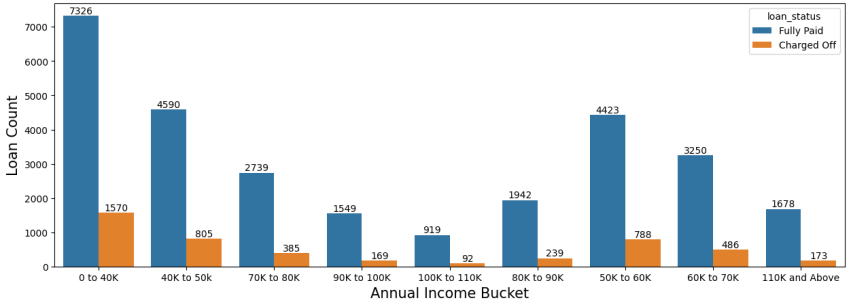
- ✓ 1. The annual income of most of the loan applicants is between 40k-75k and the average annual income is 59840.
- ✓ 2. Most of the loan amount applied was in the range of 5K to 14K and maximum loan amount applied was ~27K.
- ✓ 3. The funded amount for most of the loan applicants is between the range of 5K to 13K and maximum funded amount is ~25K.
- ✓ 4. Most of the loan applicants have interest rate between range of 8% - 14%. The average rate of interest is 11.7%.
- ✓ 5. Majority of loan applicants are living either on Rent or on Mortgage.
- ✓ 6. The purpose of most of the loan applicants is debt_consolidation.
- ✓ 7. Majority of the loan applicants have status as Fully Paid.
- ✓ 8. Highest number of loan applicants are from CA state.
- ✓ 9. Majority of the loan applicants have 10+ years of experience..
- ✓ 10. Majority of the loan applicants belong to Grade B.
- ✓ 11. Majority of the loan applicants will payment installment between 100-200.
- ✓ 12. Majority of the investors have committed for an amount between 5K to 10K.
- ✓ 13. Majority of the loan applicants belong to Verified status.
- ✓ 14. Majority of the loans were funded in the month of December, followed by November and then October.
- ✓ 15. Majority of loans funded were issued in the year 2011.
- ✓ 16. Majority of the Loans have a term of 36 months.
- ✓ 17. Annual Income of majority of the loan applicants is below 40K.
- ✓ 18. Public Record of Bankruptcies for majority of the loan applicants is 0.
- ✓ 19. Majority of the loan applicants have taken loan amount between 5K to 10k.
- ✓ 20. Majority of the loan applicants have debt to income ratio between the range of 12 – 18.

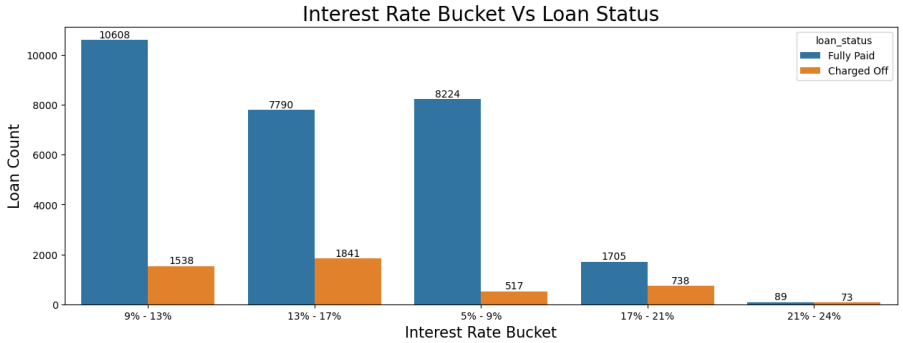
# Bivariate Analysis

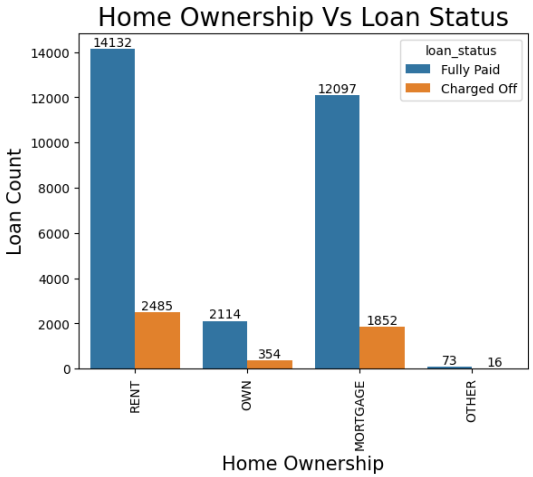

Employment Length Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Employment Length
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
6028
1099
3464
566
3116
492
1060
176
938
141
2598
402
2452
407
2429
410
1663
272
3369
509
1299
233

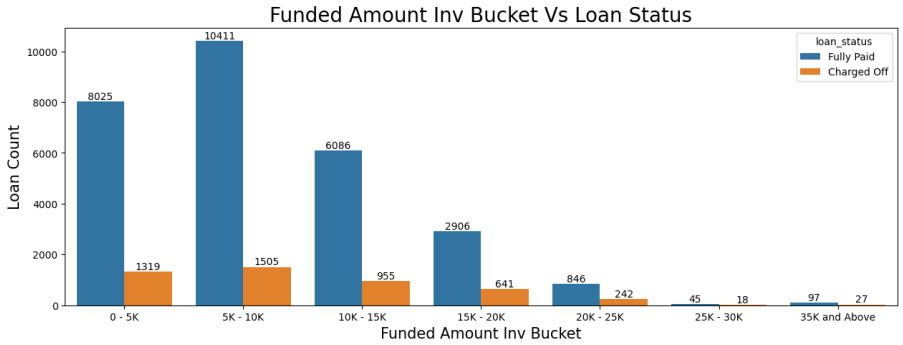

DTI Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
DTI Bucket
24 - 30
0 - 6
6 - 12
18 - 24
12 - 18
1269
221
4505
613
7546
1097
6605
1294
8370
1458

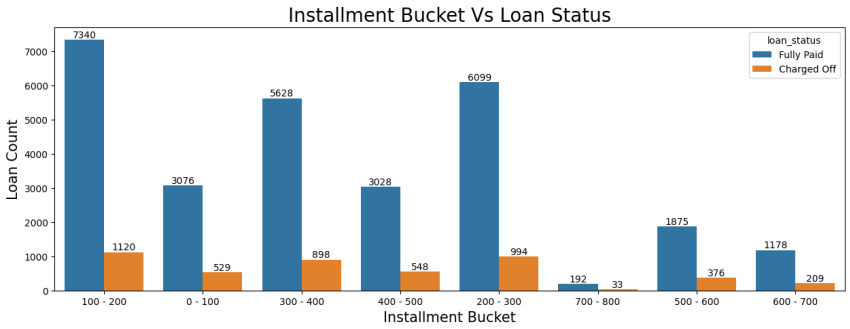

Address State Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Address State

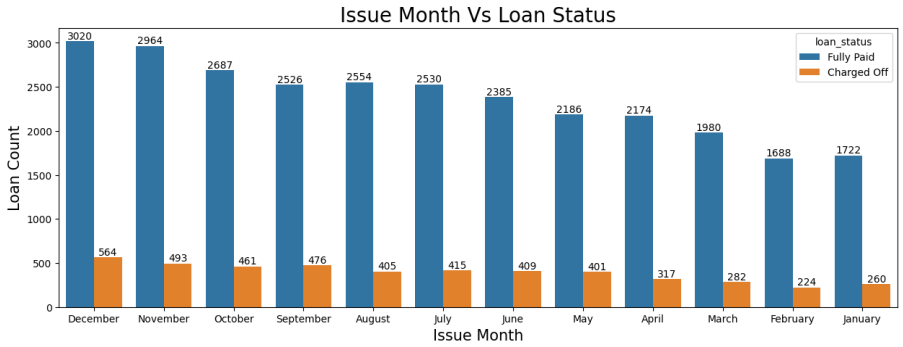

Annual Income Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Annual Income Bucket
0 to 40K
40K to 50k
70K to 80K
90K to 100K
100K to 110K
80K to 90K
50K to 60K
60K to 70K
110K and Above
7326
1570
4590
805
2739
385
1549
169
919
92
1942
239
4423
788
3250
486
1678
173

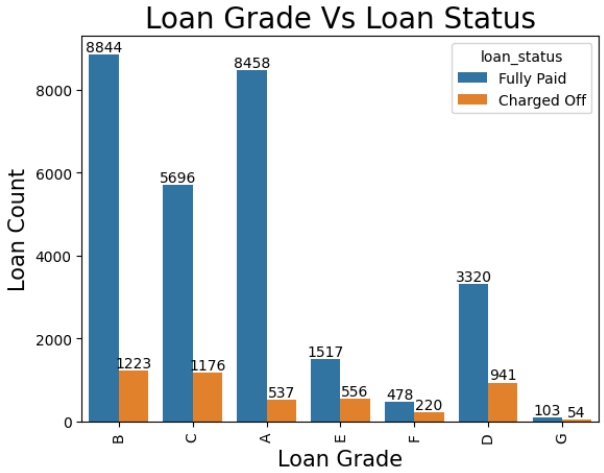

Interest Rate Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Interest Rate Bucket
9% - 13%
13% - 17%
5% - 9%
17% - 21%
21% - 24%
10608
1538
7790
1841
8224
517
1705
738
89
73

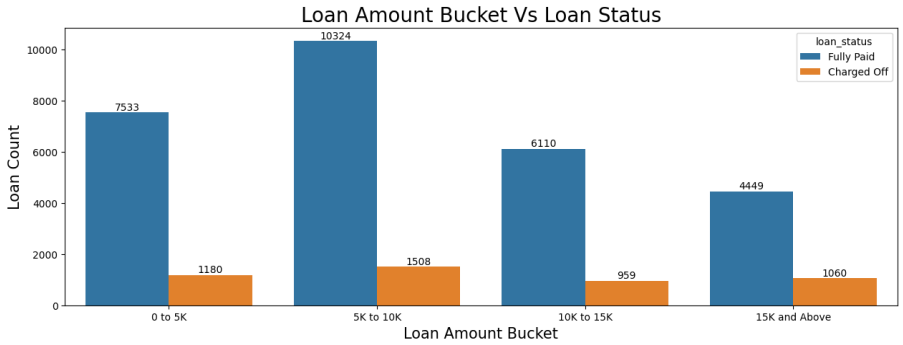

Home Ownership Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Home Ownership
RENT
OWN
MORTGAGE
OTHER
14132
2485
2114
354
12097
1852
73
16

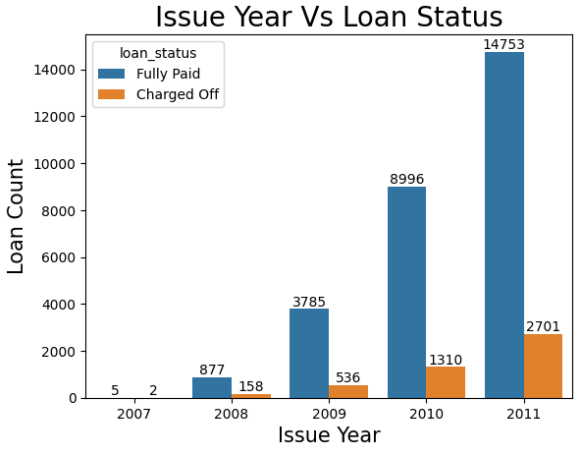

Funded Amount Inv Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Funded Amount Inv Bucket
0 - 5K
5K - 10K
10K - 15K
15K - 20K
20K - 25K
25K - 30K
35K and Above
8025
1319
10411
1505
6086
955
2906
641
846
242
45
18
97
27

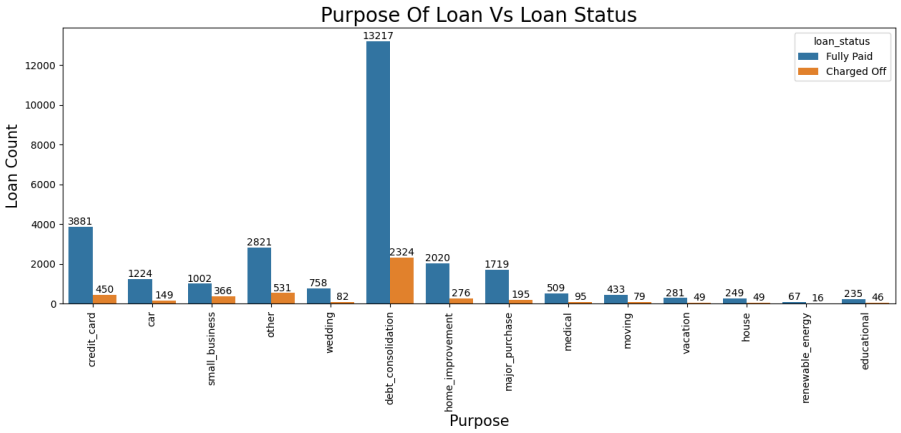

Installment Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Installment Bucket
100 - 200
0 - 100
300 - 400
400 - 500
200 - 300
700 - 800
500 - 600
600 - 700
7340
1120
3076
529
5628
898
3028
548
6099
994
192
33
1875
376
1178
209

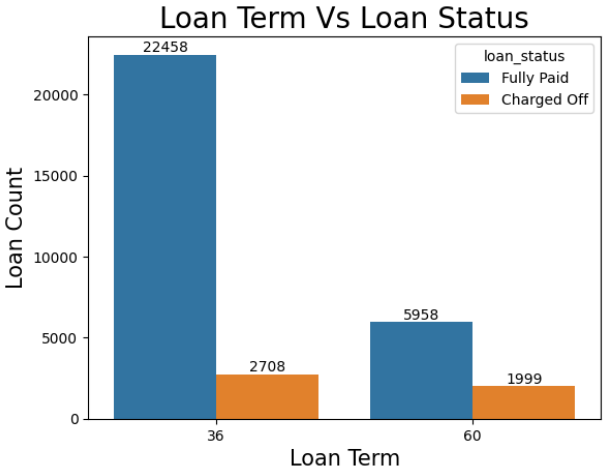

Issue Month Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Issue Month
December
November
October
September
August
July
June
May
April
March
February
January
3020
564
2964
493
2687
461
2526
476
2554
405
2530
415
2385
409
2186
401
2174
317
1980
282
1688
224
1722
260

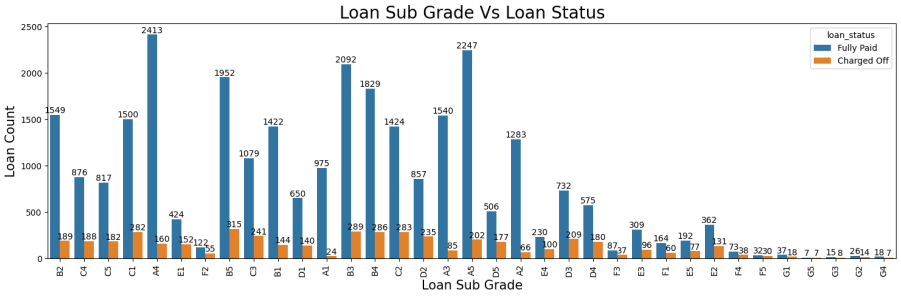

Loan Grade Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Loan Grade
B
C
A
E
F
D
G
8844
1223
5696
1176
8458
537
1517
556
478
220
3320
941
103
54

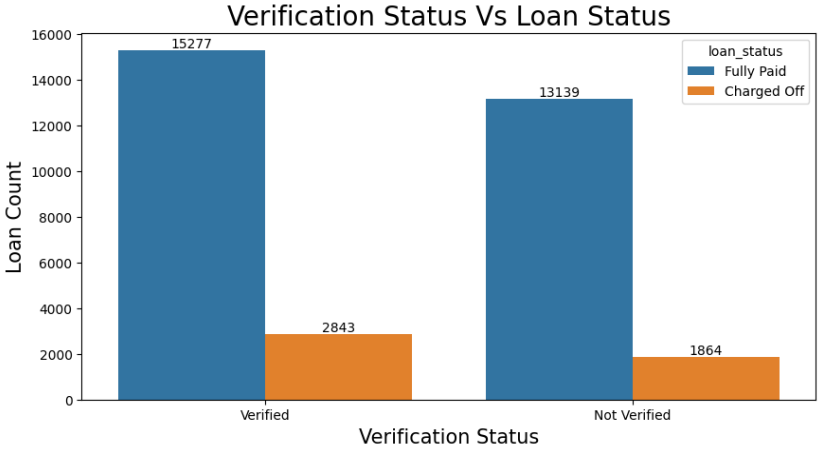

Loan Amount Bucket Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Loan Amount Bucket
0 to 5K
5K to 10K
10K to 15K
15K and Above
7533
1180
10324
1508
6110
959
4449
1060


Issue Year Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Issue Year
2007
2008
2009
2010
2011
5
2
877
158
3785
536
8996
1310
14753
2701


Purpose Of Loan Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
3881
450
1224
149
1002
366
2821
531
758
82
13217
2324
2020
276
1719
195
509
95
433
79
281
49
249
49
67
16
235
46


Loan Term Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Loan Term
22458
2708
5958
1999
36
60

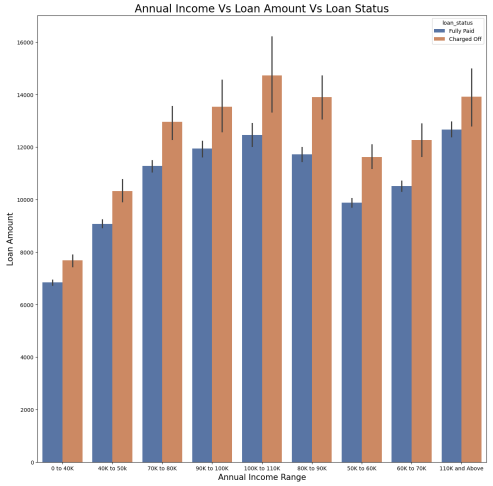

Loan Sub Grade Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Loan Sub Grade

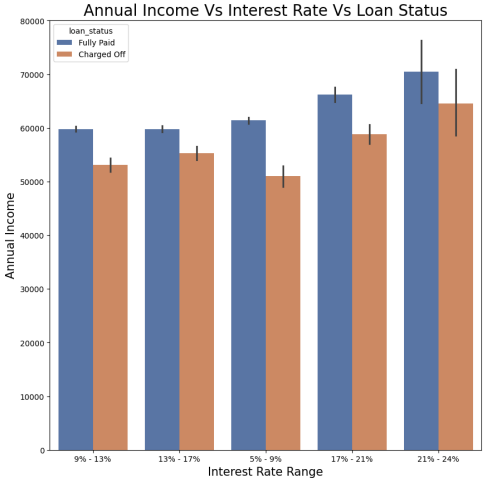

Verification Status Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Count
Verification Status
15277
2843
13139
1864
Verified
Not Verified

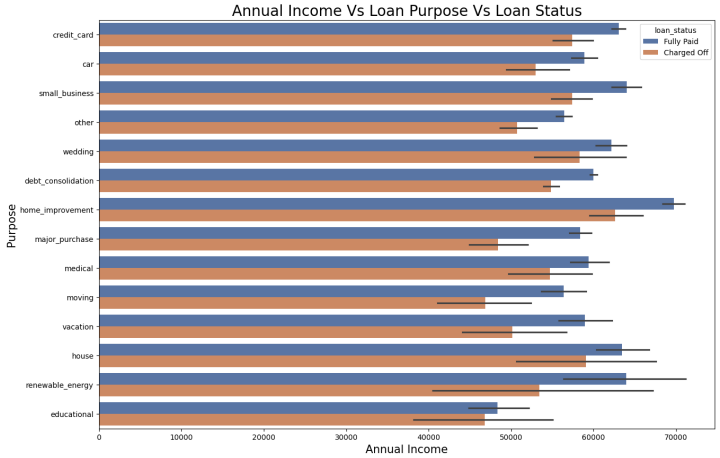
# Conclusion from Bivariate Analysis

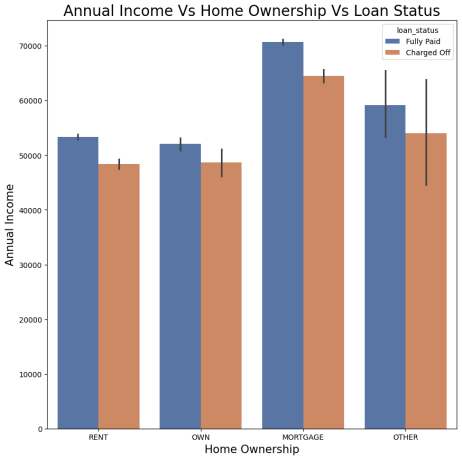
- ✓ 1. Debt Consolidation is the category where the maximum number of loans are 36-month and people have defaulted in the same category.
- ✓ 2. Loan applicants with home ownership as RENT and MORTGAGE are more likely to default.
- ✓ 3. Verified loan applicants are defaulting more than who are Not Verified.
- ✓ 4. Loan applicants from states of California (CA), Florida (FL), and New York (NY) are most likely to default.
- ✓ 5. Loan applicants belonging to Grades B, C and D contribute to the most of the Charged Off loans.
- ✓ 6. Loan applicants belonging to Sub Grades B3, B4 and B5 contribute to the most of the Charged Off loans.
- ✓ 7. Loan applicants applying for 36-month term are more likely to default than those taking loans for 60 months.
- ✓ 8. Loan applicants with more than 10 years employment length contribute the most to the charge off loans.
- ✓ 9. The number of loan applicants have steadily increased from 2007 to 2011, indicating a positive trend in upcoming years, 2011 year has the highest number of loan applicants and defaulters.
- ✓ 10. December is the most preferred month for taking loans.

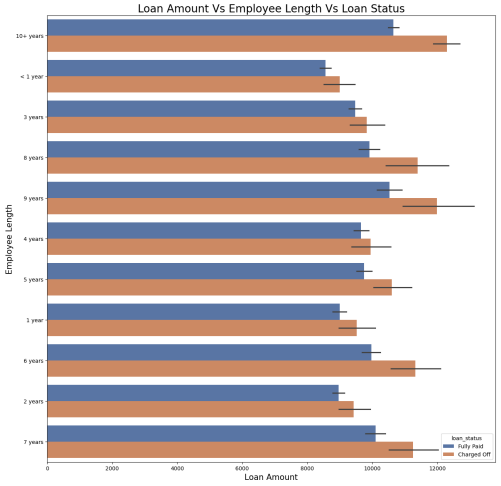
# Multivariate Analysis

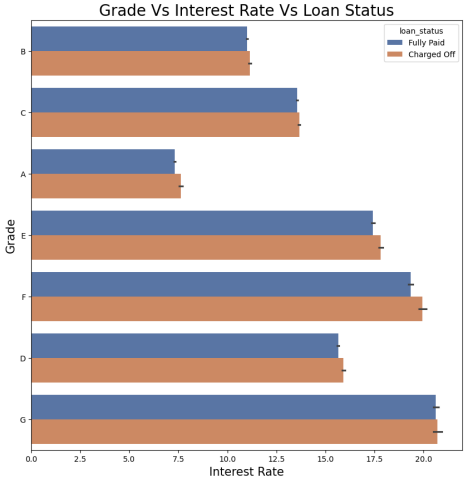

Annual Income Vs Loan Amount Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Amount
0
2000
4000
6000
8000
10000
12000
14000
16000
0 to 40K
40K to 50k
70K to 80K
90K to 100K
100K to 110K
80K to 90K
50K to 60K
60K to 70K
110K and Above
Annual Income Range

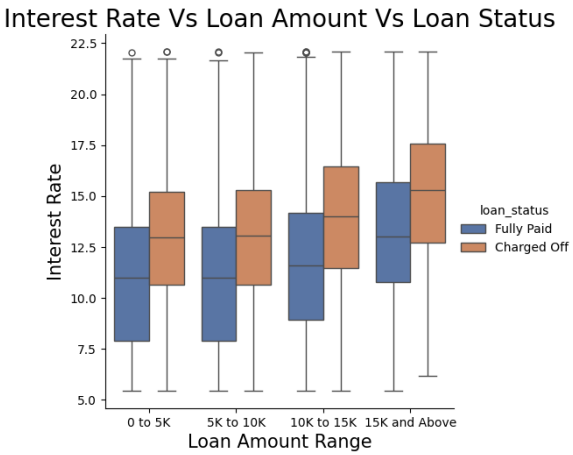

Annual Income Vs Interest Rate Vs Loan Status
loan_status
Fully Paid
Charged Off
Annual Income
0
10000
20000
30000
40000
50000
60000
70000
80000
9% - 13%
13% - 17%
5% - 9%
17% - 21%
21% - 24%
Interest Rate Range

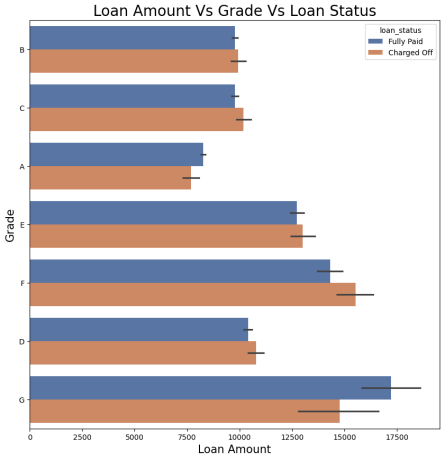

Annual Income Vs Loan Purpose Vs Loan Status
loan_status
Fully Paid
Charged Off
Purpose
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
0
10000
20000
30000
40000
50000
60000
70000
Annual Income

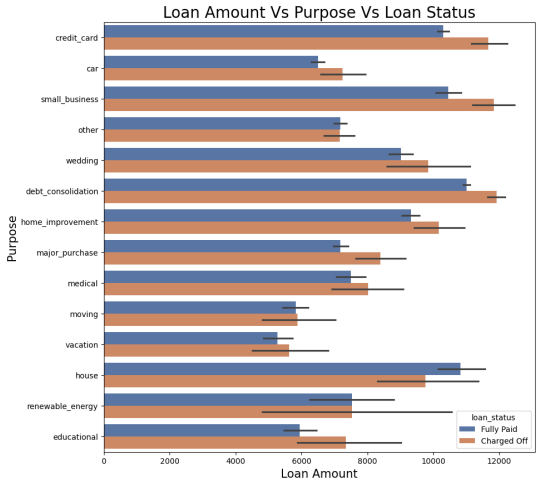

Annual Income Vs Home Ownership Vs Loan Status
loan_status
Fully Paid
Charged Off
Annual Income
0
10000
20000
30000
40000
50000
60000
70000
RENT
OWN
MORTGAGE
OTHER
Home Ownership

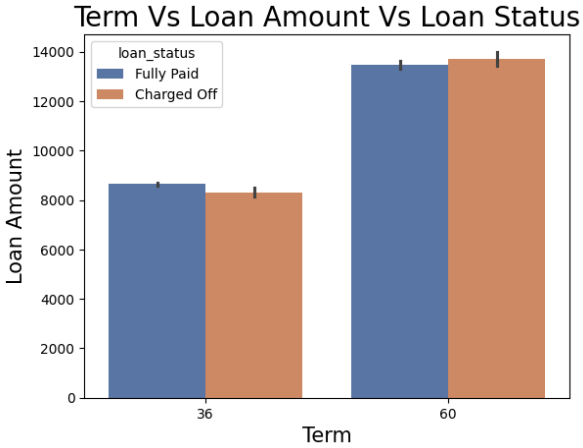

Loan Amount Vs Employee Length Vs Loan Status
Employee Length
10+ years
< 1 year
3 years
8 years
9 years
4 years
5 years
1 year
6 years
2 years
7 years
0
2000
4000
6000
8000
10000
12000
Loan Amount
loan_status
Fully Paid
Charged Off

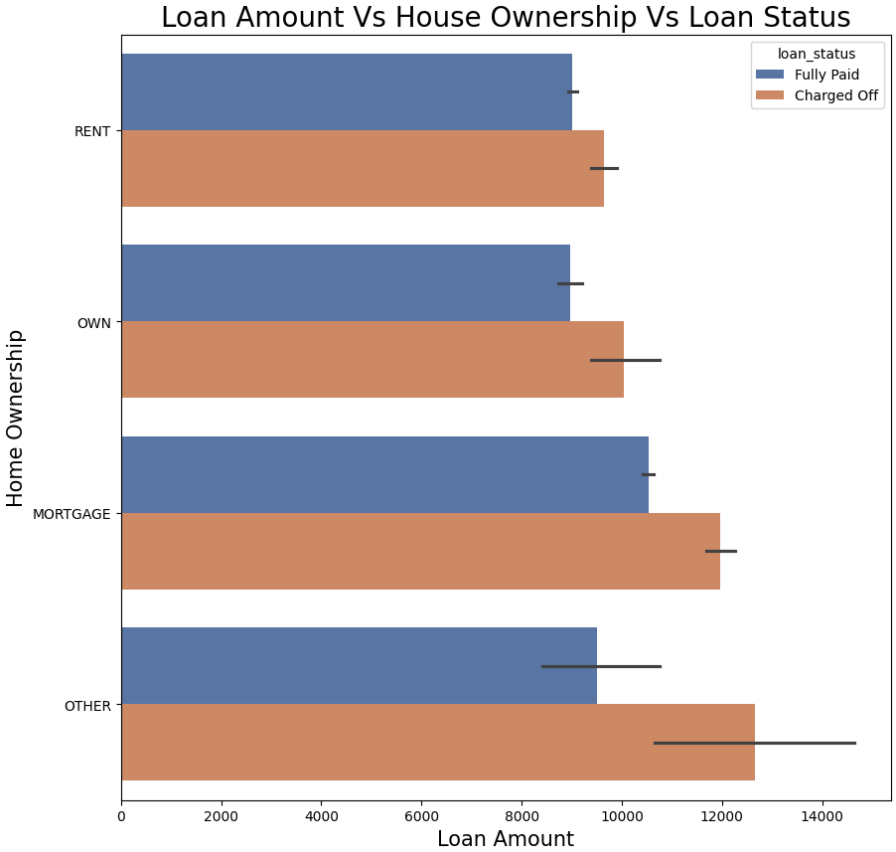

Grade Vs Interest Rate Vs Loan Status
Grade
B
C
A
E
F
D
G
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Interest Rate
loan_status
Fully Paid
Charged Off


Interest Rate Vs Loan Amount Vs Loan Status
Interest Rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
0 to 5K
5K to 10K
10K to 15K
15K and Above
Loan Amount Range
loan_status
Fully Paid
Charged Off


Loan Amount Vs Grade Vs Loan Status
Grade
B
C
A
E
F
D
G
0
2500
5000
7500
10000
12500
15000
17500
Loan Amount
loan_status
Fully Paid
Charged Off


Loan Amount Vs Purpose Vs Loan Status
credit_card
car
small_business
other
wedding
debt_consolidation
home_improvement
major_purchase
medical
moving
vacation
house
renewable_energy
educational
Purpose
loan_status
Fully Paid
Charged Off
0
2000
4000
6000
8000
10000
12000
Loan Amount


Term Vs Loan Amount Vs Loan Status
loan_status
Fully Paid
Charged Off
Loan Amount
0
2000
4000
6000
8000
10000
12000
14000
36
60
Term


Loan Amount Vs House Ownership Vs Loan Status
loan_status
Fully Paid
Charged Off
Home Ownership
RENT
OWN
MORTGAGE
OTHER
0
2000
4000
6000
8000
10000
12000
14000
Loan Amount

# Conclusions from Multivariate Analysis

The below analysis is with respect to the charged off loans. There is a more probability of defaulting when :

- ✓ 1. Though the number of loans applied and defaulted are the highest in number for "debt_consolidation". The annual income of those who applied under the "debt_consolidation" purpose isn't the highest. Applicants with higher annual income applied loans for home_improvement, house, renewable_energy and small_businesses.
- ✓ 2. The highest number of loan applicants who have defaulted have a house ownership of 'MORTGAGE' and have the highest Annual Income.
- ✓ 3. Across all income groups, the number of defaulters and the loan amount is higher is higher income groups.
- ✓ 4. The interest rate for charged off loans at a rate of 21-24% and have an income of 60-70k.
- ✓ 5. Loan applicants with loan amount of 15K and above have an interest rate between 14 - 16%.
- ✓ 6. Loan applicants with purpose of 'small_business' have taken highest loan amount of 12k.
- ✓ 7. Loan applicants whose ownership is OTHER have taken the highest loan amount between 12-14k while the applicants whose ownership is 'MORTGAGE' have taken loan amount between 12K to 14K.
- ✓ 8. For grade of F the loan amount is between 15K to 17.5K.
- ✓ 9. For applicants with employment length above 10 years, the loan amount is greater than 12K.
- ✓ 10. Applicants with Grade G have interest rate more than 20%.
- ✓ 11. As the Loan Amount increases, the Interest Rate also increases. The interest rate for charged off loans is higher than that of fully paid loans in all the loan_amount groups.
- ✓ 12. Loan applicants who have applied for long term of 60 Months have applied for higher loan amounts.

# Correlations

# Correlation Analysis

## Strong Correlation

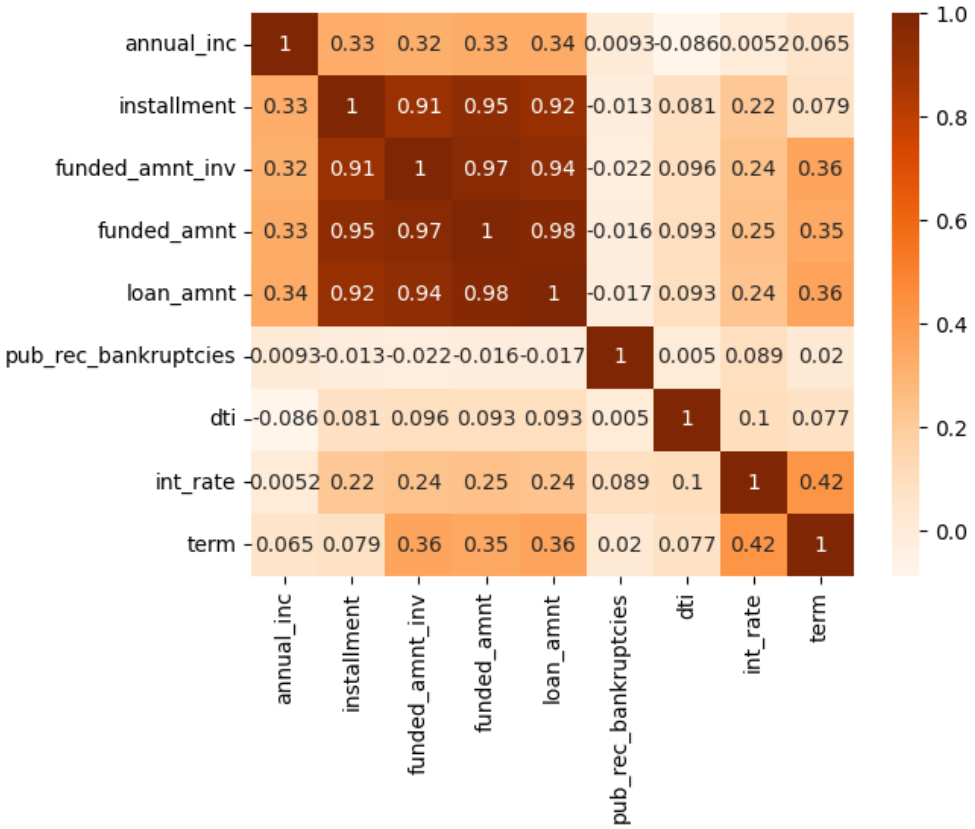
- ✓ **installment** has strong correlation with **funded_amnt, loan_amnt and funded_amnt_inv**
- ✓ **term** has a strong correlation with **int_rate**
- ✓ **annual_inc** has a strong correlation with **loan_amnt**

## Weak Correlation

- ✓ **dti** has weak correlation with most of the fields.

## Negative Correlation

- ✓ **annual_inc** has negative correlation with **dti**
- ✓ **pub_rec_bankruptcies** has a negative correlation with almost every field



# Summary

- ✓ 1. There has been a positive trend in loan applications from 2007 to 2011. The company should capitalize on the market's growth trend by meeting the increased demand.
- ✓ 2. Highest number of loan applications are seen in the month of December. Possibly due to holiday season. Company can use this trend and create new marketing strategy.
- ✓ 3. Grades B, C and D have a high possibility of defaulting and hence needs stricter risk assessment.
- ✓ 4. Sub Grades B3, B4 and B5 have a high possibility of defaulting and hence needs stricter risk assessment.
- ✓ 5. Company should carefully evaluate loan applications applied for debt consolidation purpose.
- ✓ 6. Company should evaluate the home ownership status of the loan applicant to access housing stability, and it impact on the applicant's ability to repay the loan.
- ✓ 7. It has been seen that verified applicants were the highest in loan defaulters. Company needs consider improvising on their verification process.
- ✓ 8. It has been seen that highest number of defaulter belong to CA, NY and FL states. Company should consider implementing stricter risk assessment for these states.
- ✓ 9. Loan applicants with short term, 36 months term have been seen defaulting more compared to long term, 60 months tenure. Company can think of increasing the interest rates for short term loans.