Lending Club Case Study

Group members:
Prabhjot kaur chahal
Karnatakam Akshatha

The Problem

Company

- Lending Club 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.

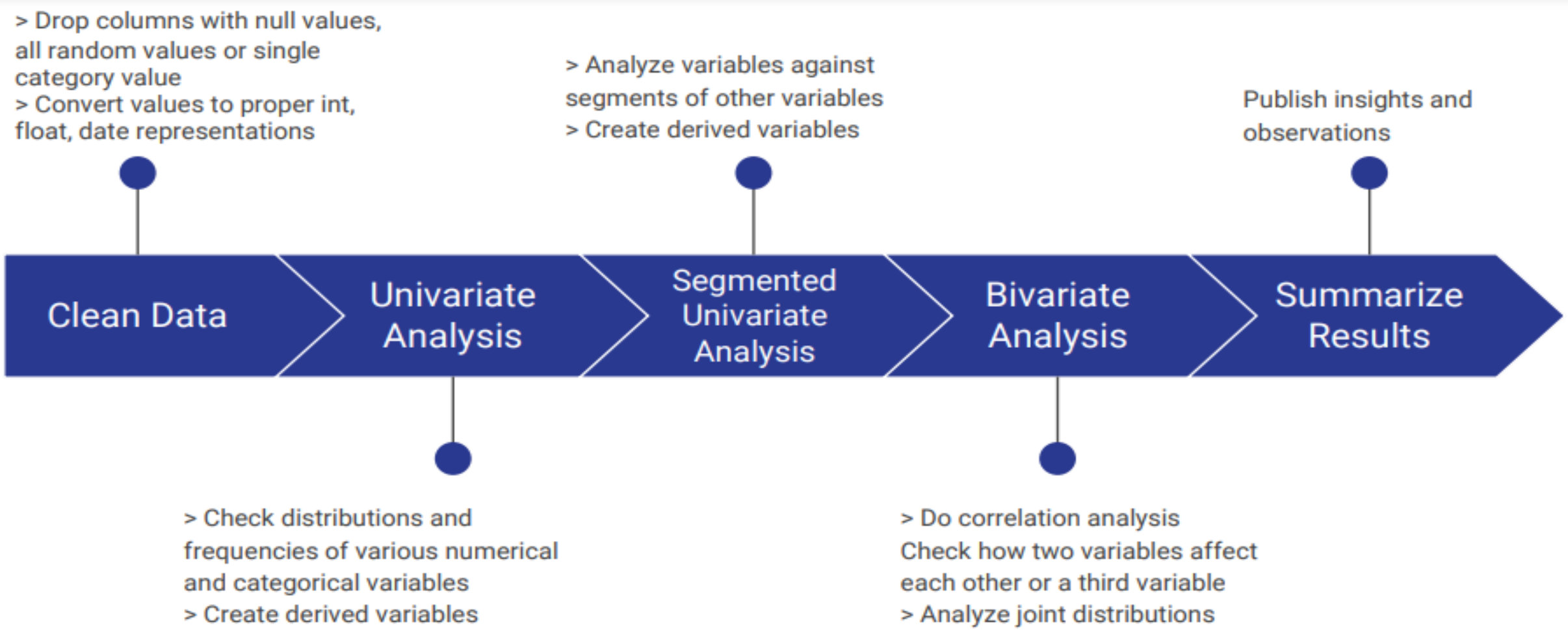
Context

- Lending Club wants to understand the driving factors behind loan default, i.e. the driver variables which are strong indicators of default.
- The company can utilize this knowledge for its portfolio and risk assessment.

Problem statement

- As a data scientist working for Lending Club analyze the dataset containing information about past loan applicants using EDA to understand how consumer attributes and loan attributes influence the tendency of default

Approach:


> Drop columns with null values, all random values or single category value
> Convert values to proper int, float, date representations
> Analyze variables against segments of other variables
> Create derived variables
Publish insights and observations
Clean Data
Univariate Analysis
Segmented Univariate Analysis
Bivariate Analysis
Summarize Results
> Check distributions and frequencies of various numerical and categorical variables
> Create derived variables
> Do correlation analysis
Check how two variables affect each other or a third variable
> Analyze joint distributions

Problem Solving Methodology

Data Understanding And Sourcing:

1. Loan Data Set contains the complete loan data for all loans issued through the time period 2007 to 2011.
2. Data Dictionary describes the meaning of each variable used in Loan Data Set.

Data summary and cleaning:

1. DataFrame Shape Initially (Observations, Variables)- (39717, 111)
2. Data-types: float64(74), int64(13), object(24)
3. Data cleaning-

- Deleting unnecessary columns and rows -We dropped the columns having more than 30% missing values. Most of the columns has 100% Null values,so can be droped safely. There are some columns where NULL value rate is 2%, 33%, 65%, 93% and 97%. 93% and 97%, such columns can be dropped too. It is necessary to put a threshold rate. Looking into the criticality of the columns and NULL values rate, it is okay to consider the rate as 30%. So any Columns having NaN or NULL values greater than or equal to 30% can be safely dropped.

  Therefore, Threshold Dropped Percentage Rate = 30%

- Filtered data- Insignificant Columns dropped like: 1. id 2. member_id 3. verification_status 4. pymnt_plan 5. url 6. zip_code 7. initial_list_status 8. collections_12_mths_ex_med 9. policy_code 10. application_type 11. acc_now_delinq 12. chargeoff_within_12_mths 13. delinq_amnt 14. tax_liens 15. title 16. total_rec_prncp 17. total_rec_int 18. total_rec_late_fee 19. recoveries 20. collection_recovery_fee 21. last_pymnt_amnt 22. revol_bal

  Dropping Record where home_ownership = NONE, duplicates if any, Removing Outliers for Column = annual_inc

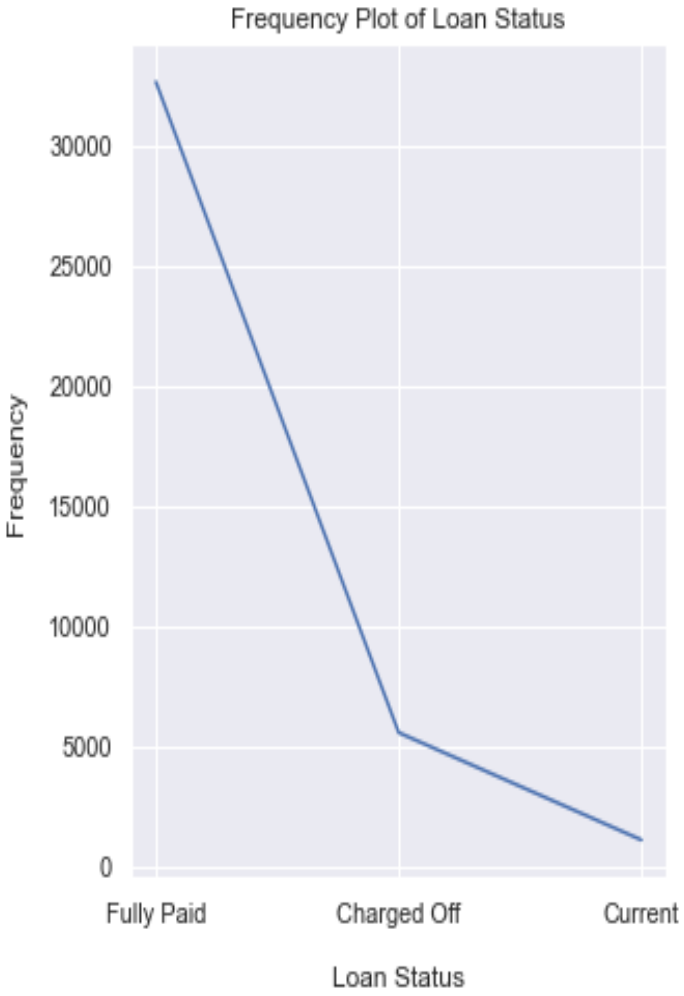
- Removing extra characters- has leading space or has repetative characters %.
- Fixing invalid values- For better Readability: Renaming Columns = int_rate => int_rate_percent
- Fixing columns and rows- Convert incorrect data-types: int_rate_percent: object => float64

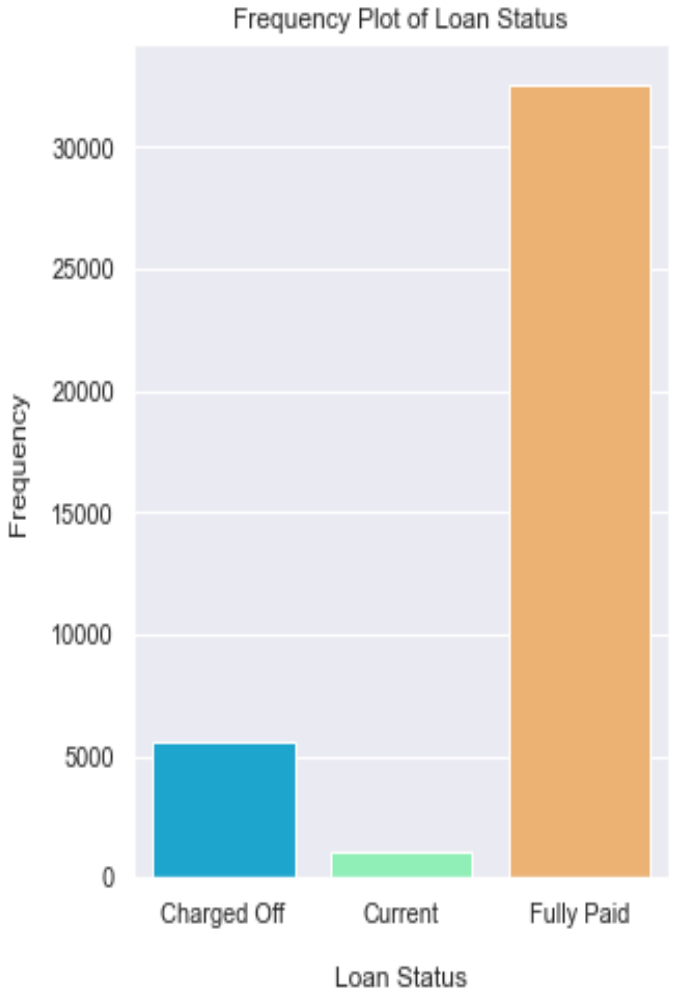
4. DERIVED METRICS :

   Deriving new metric "issue_yr" from existing variable "issue_d"

Uni-variate analysis : Unordered categorical variables -

1. loan_status



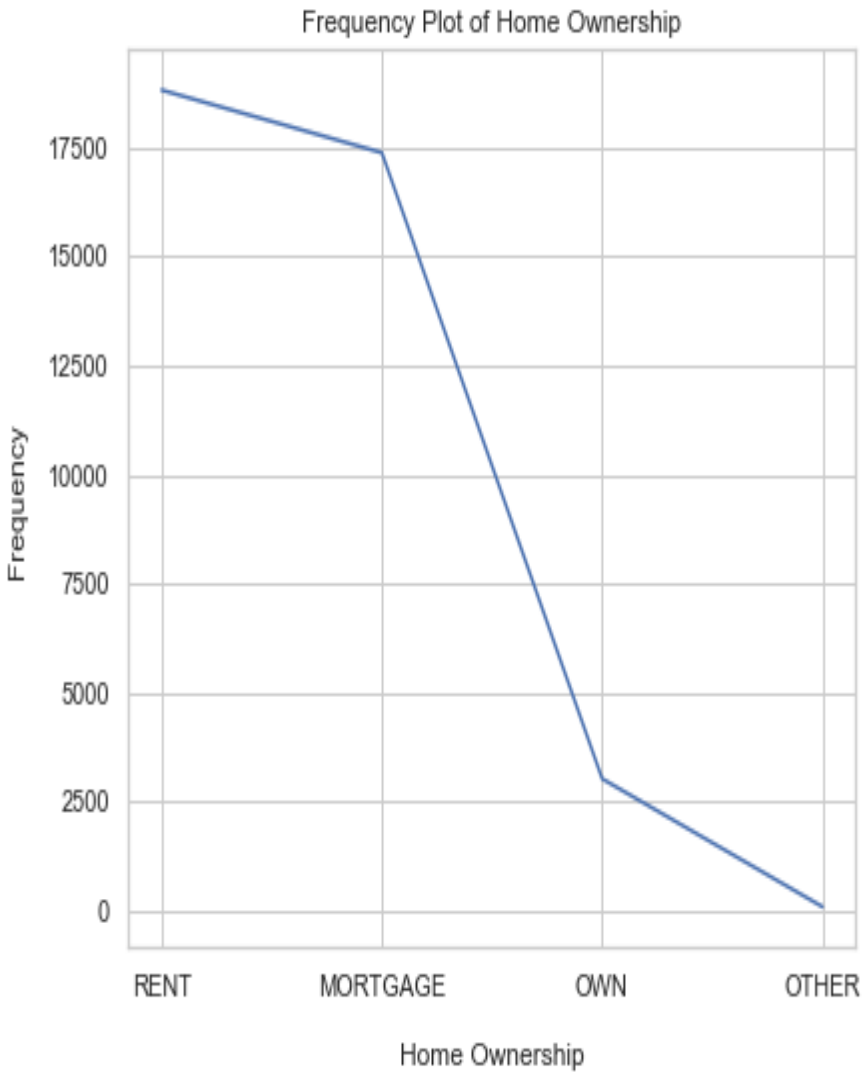


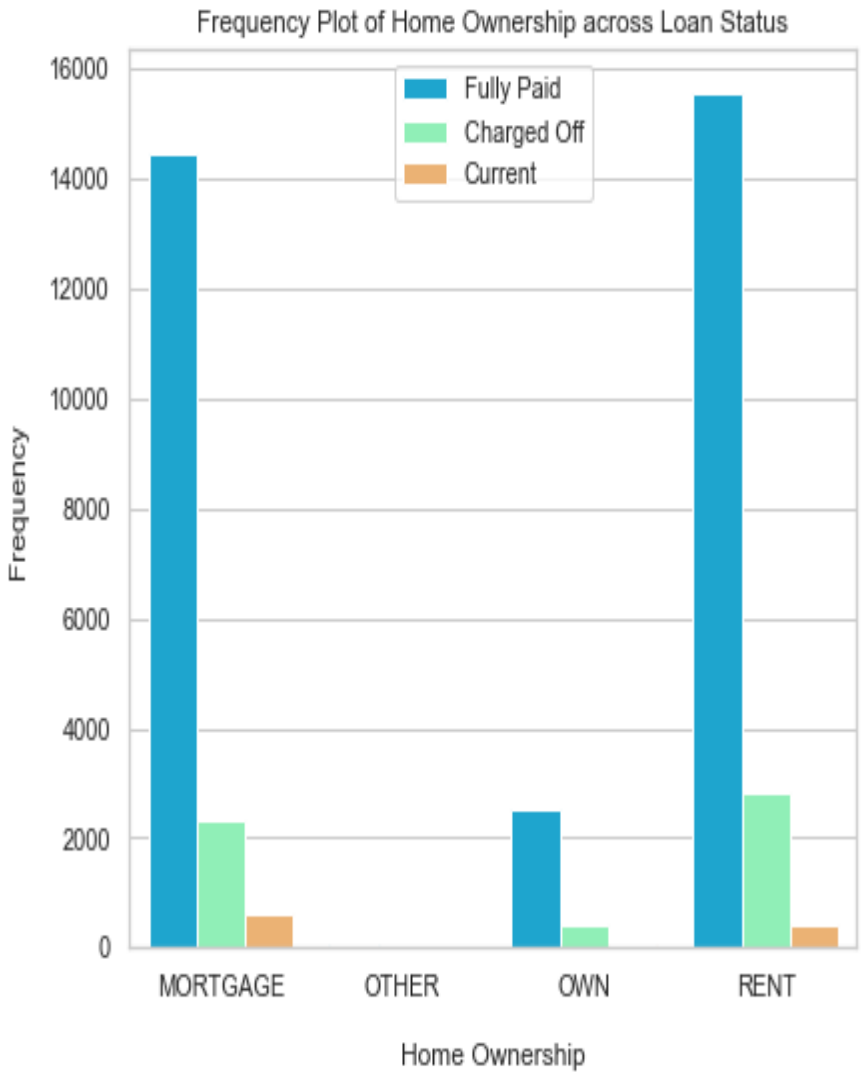
INSIGHTS :

•Around 5000 loans are charged off.

•More than 30000 loans are fully paid.

2. home ownership-

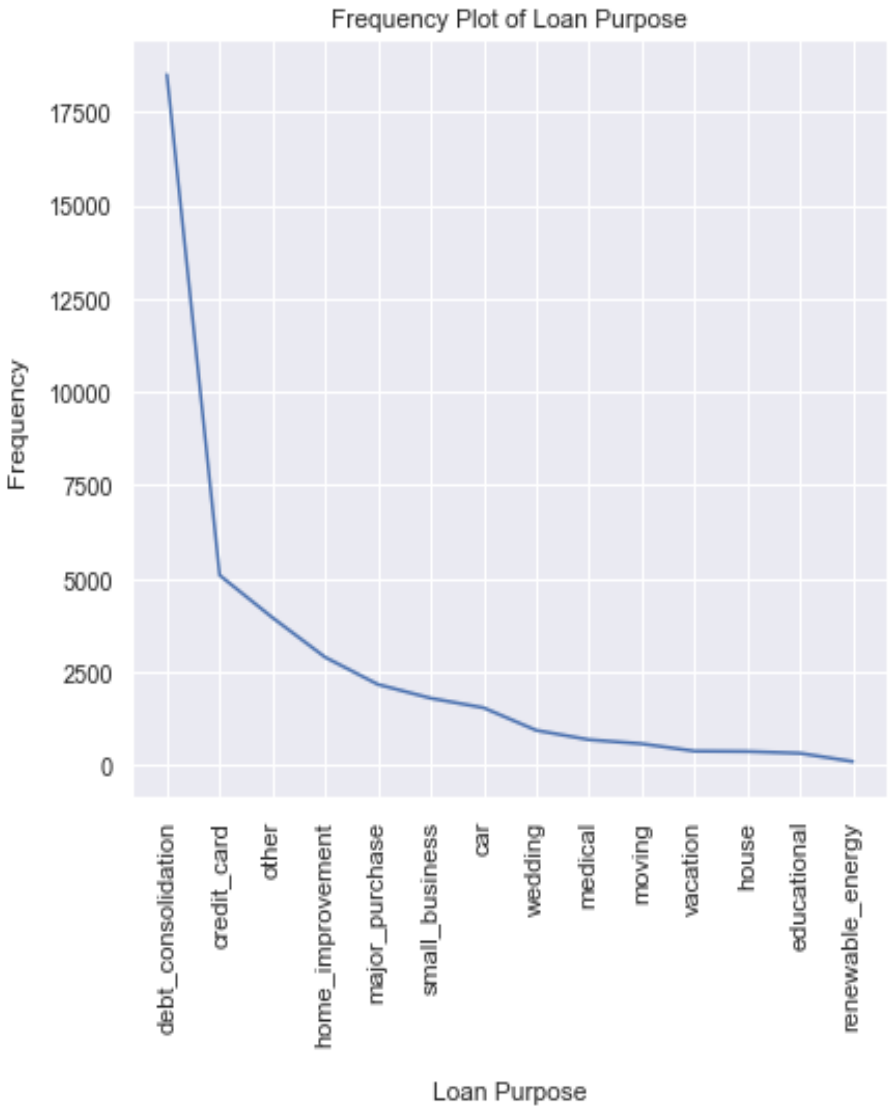


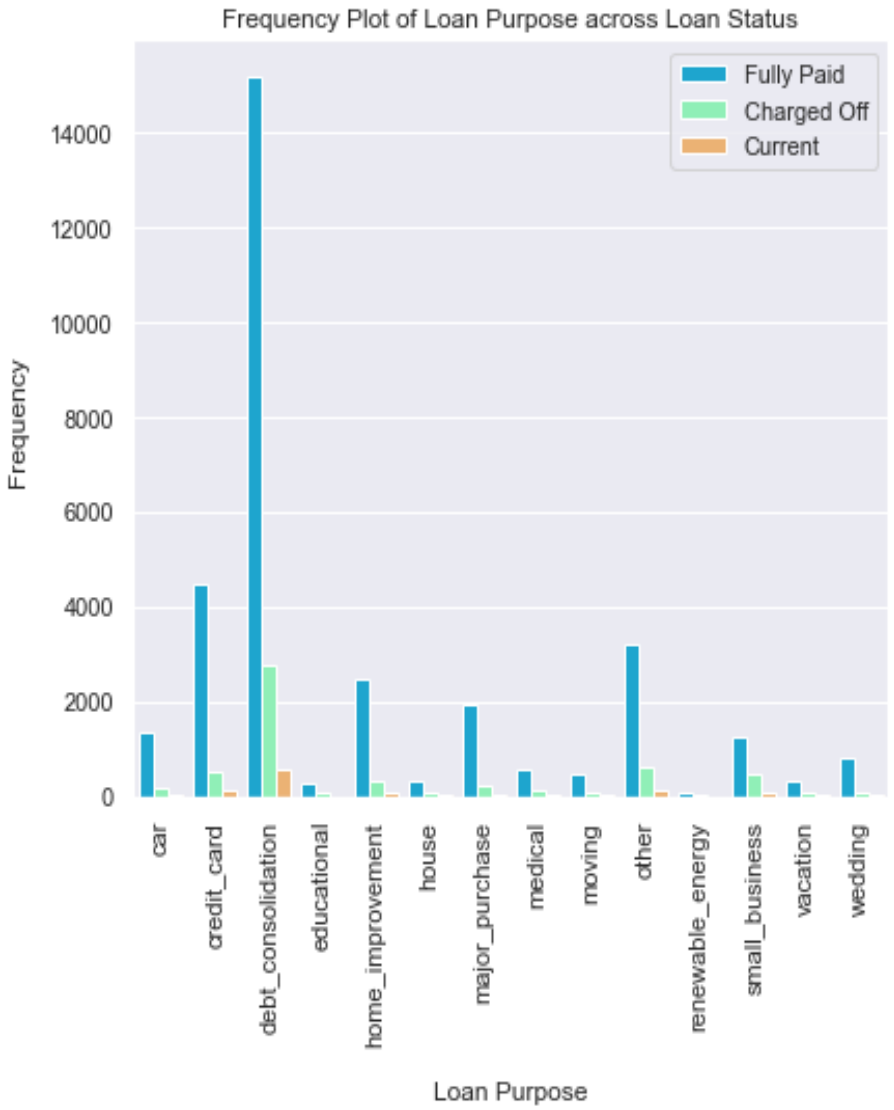


INSIGHTS:

•Frequency of MORTGAGE and RENT home ownership is high for Charged Off.

3. purpose -





INSIGHTS :

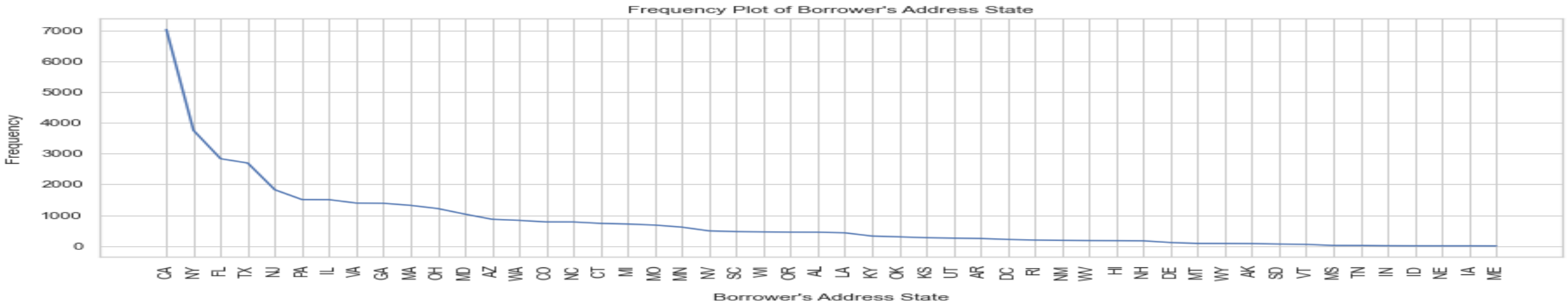
•Top 3 loan purpose of Charged off in descending order:

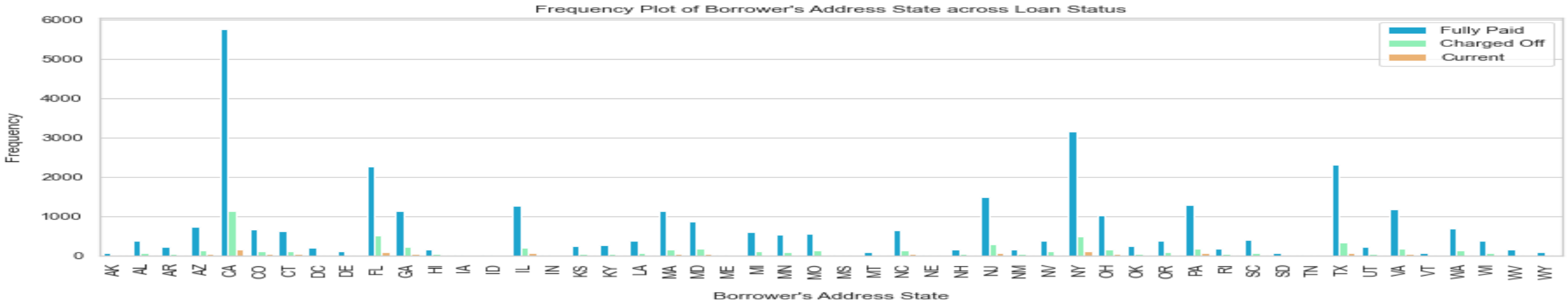
•Maximum no. of loans accepted for - debt_consolidation purpose.

•credit_card

•other

4. addr_state –



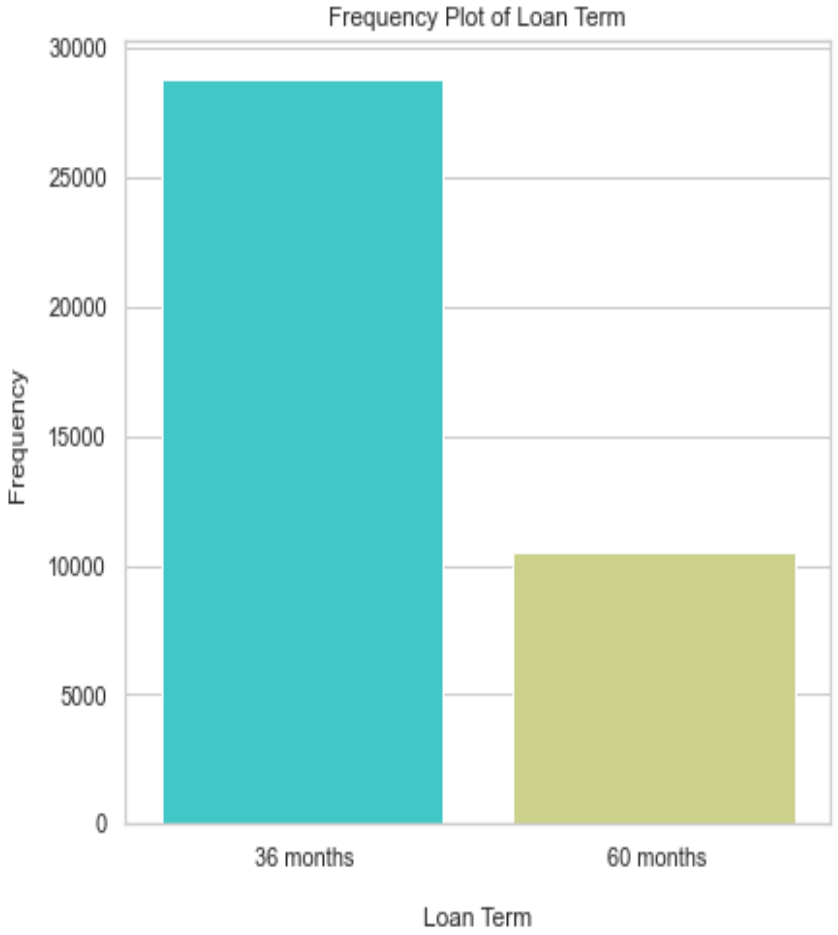


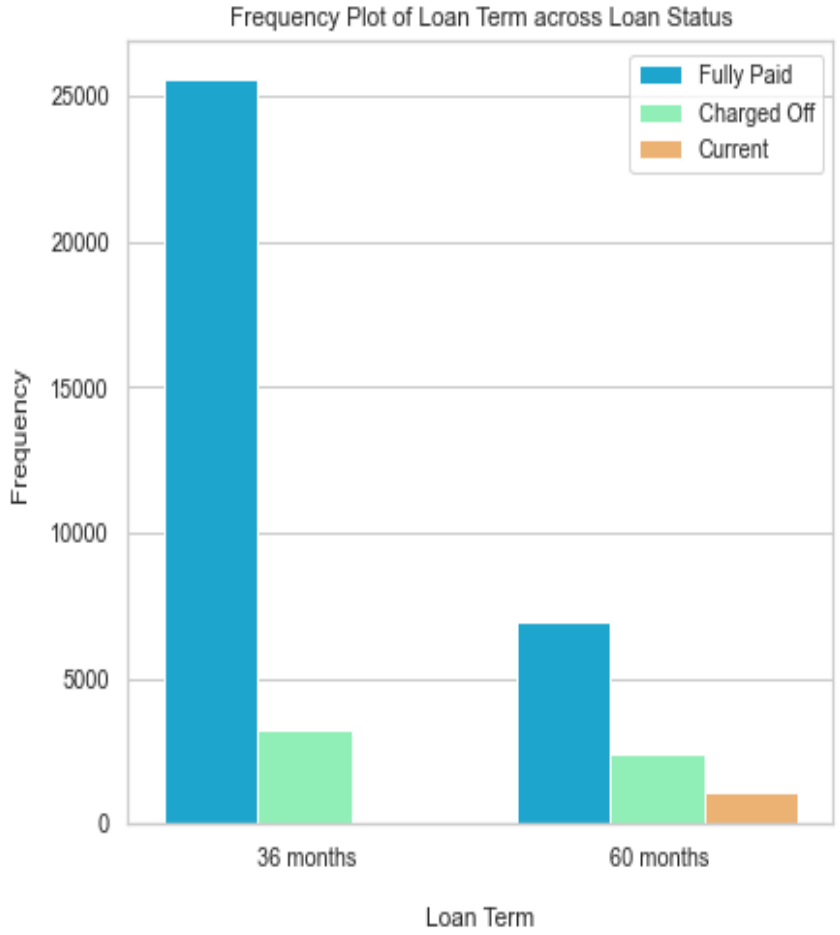
Insights

- Maximum no. of loans are applied by people from CA.
- The top 3 Address States of Charged off are : CA, NY, FL

UNIVARIATE ANALYSIS : Ordered Categorical Variables

1. term

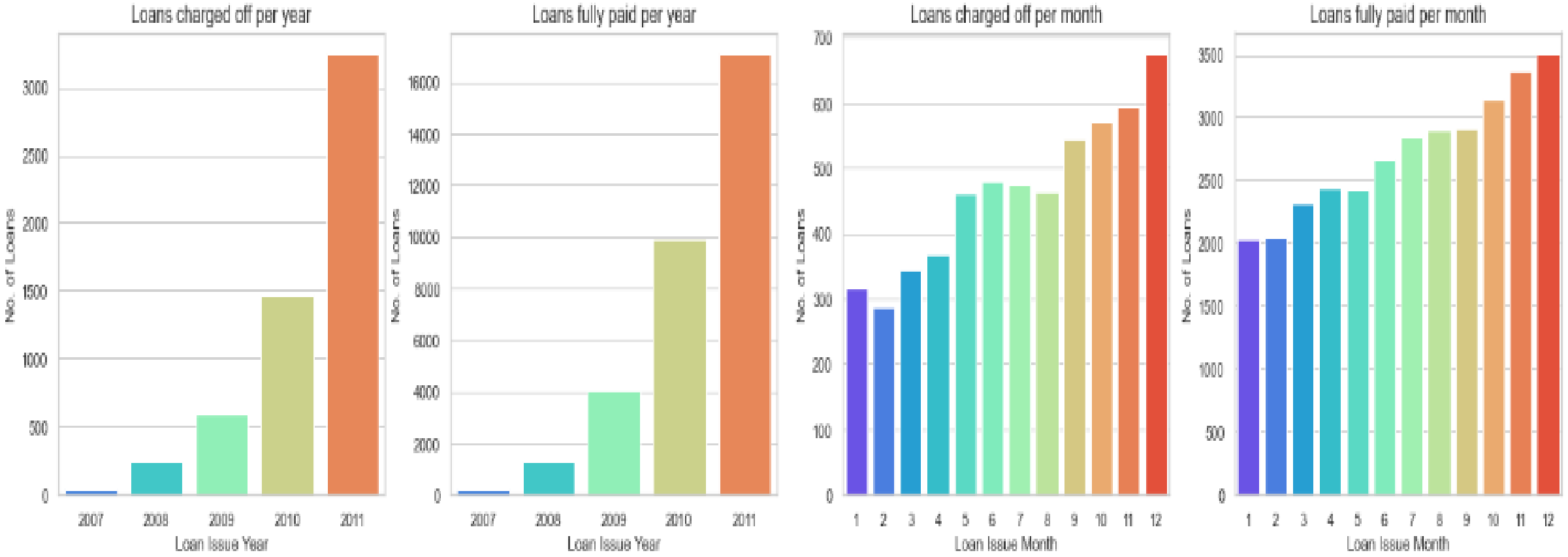




INSIGHTS :

•Total no. of loan applications are very high for 36 month term as compared to 60 month term.

Year Wise and Month Wise Distribution of charged-off and fully paid loans

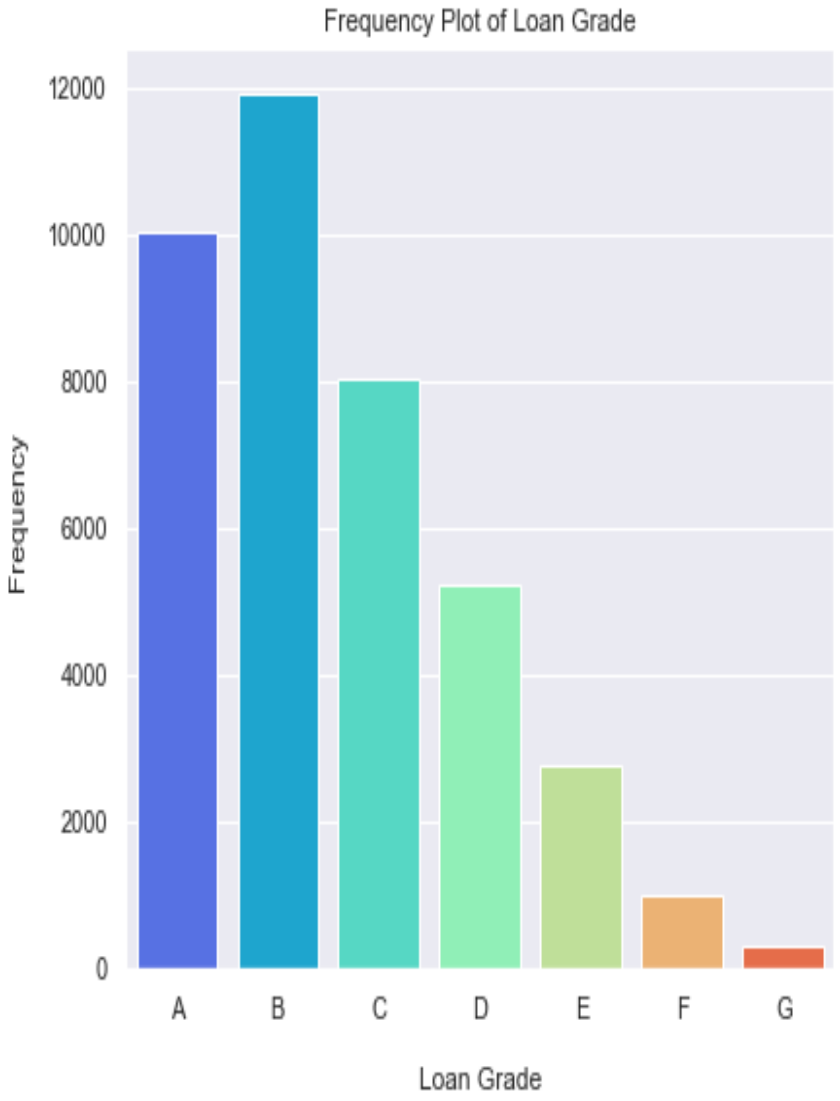


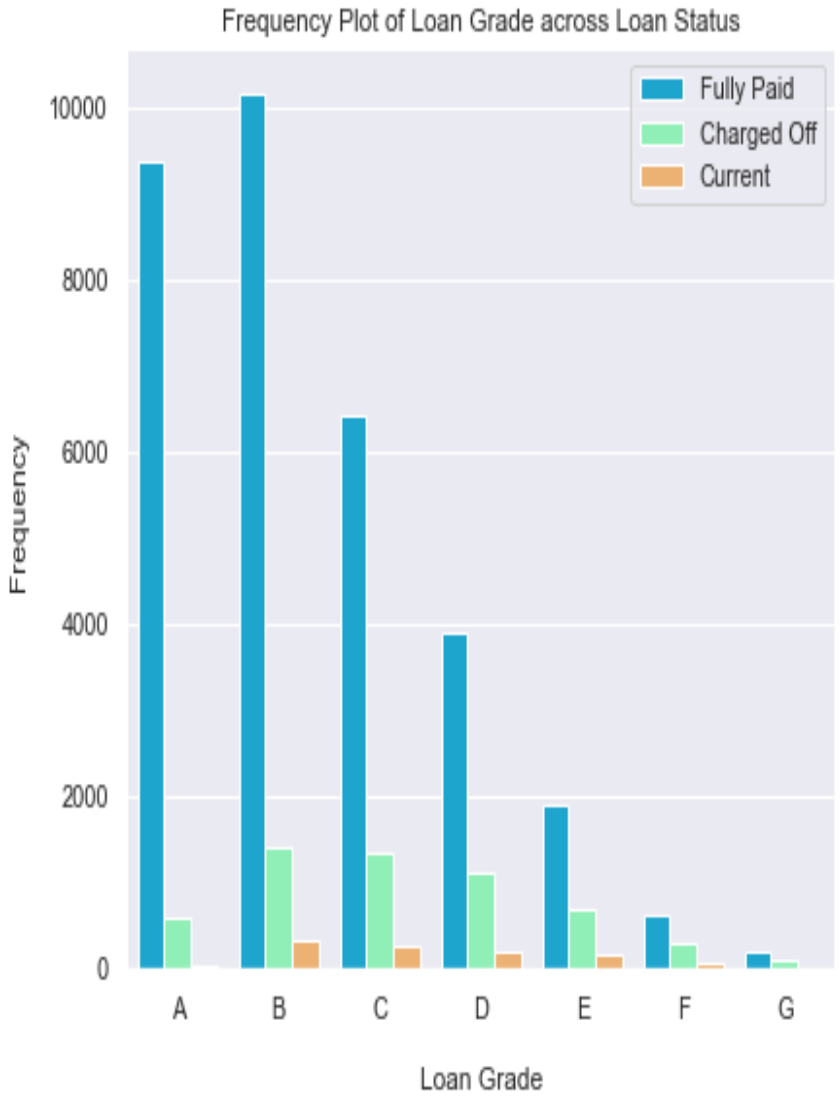
INSIGHTS:

No. of loans, fully paid and charged off are increasing every year. They are highest in the year 2011. This shows a very positive trend for Lending Club as the requirement of loans are increasing by each year.

The month-wise trend shows that most loans are fully paid as well as charged off as the year comes to an end, maximum in the month of December clearly depicting the significance of end of the year.

2. grade –





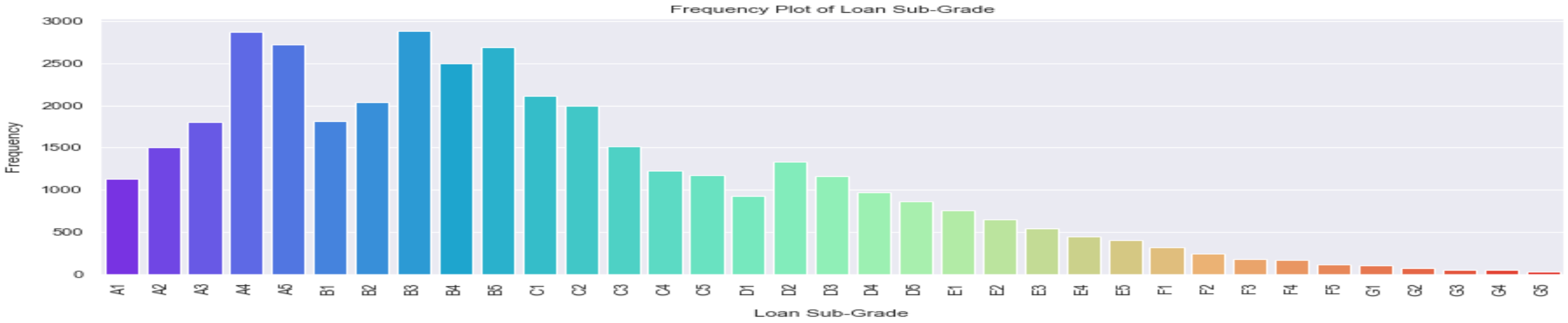
INSIGHTS

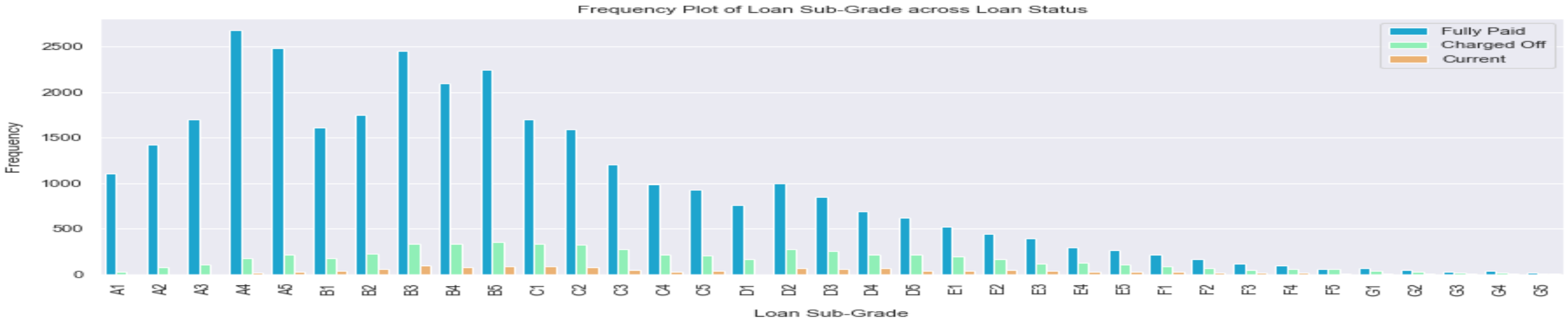
•Grade A and B loans are safe.

•Grade E, F, G loans are less safe as compared to others.

•Most of the Charged off applicants belong to Grade B, C and D.

3. sub_grade-





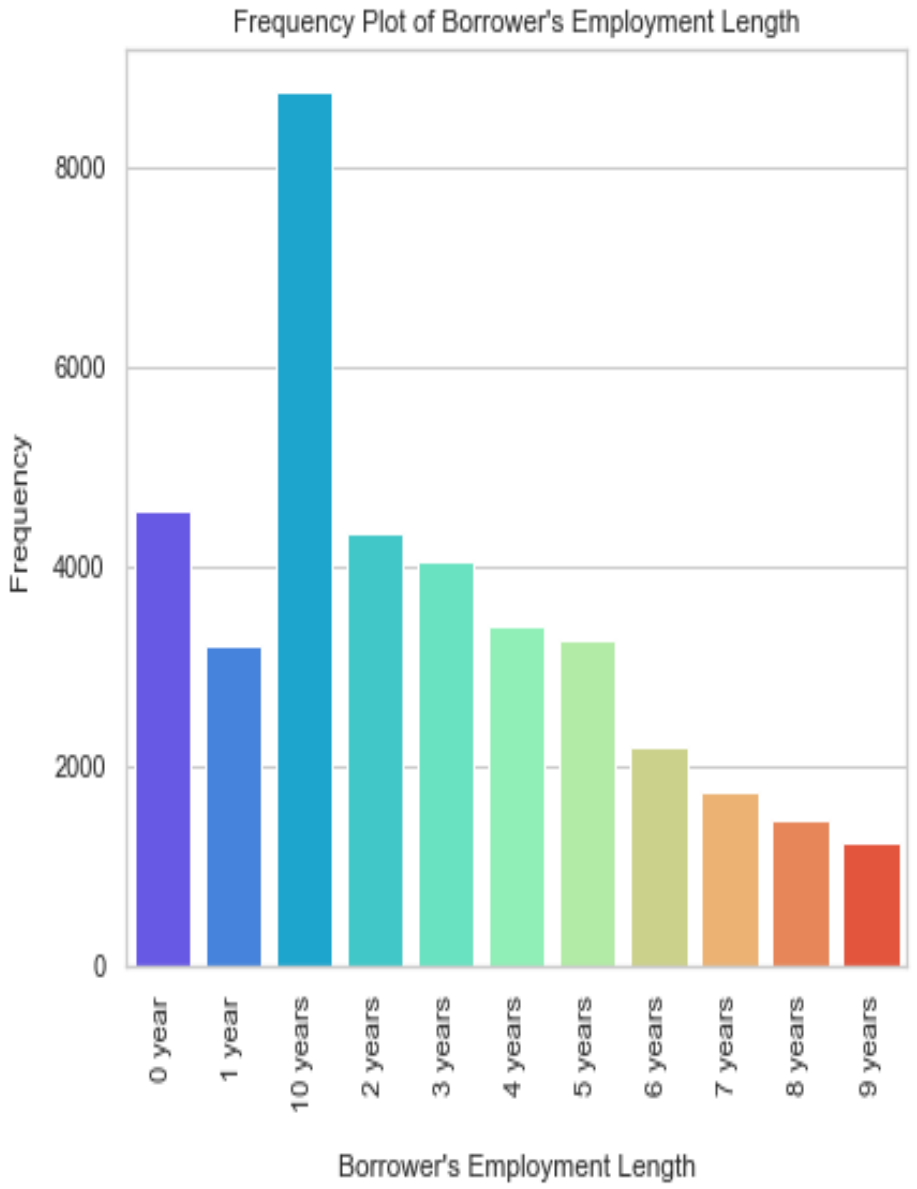
INSIGHTS

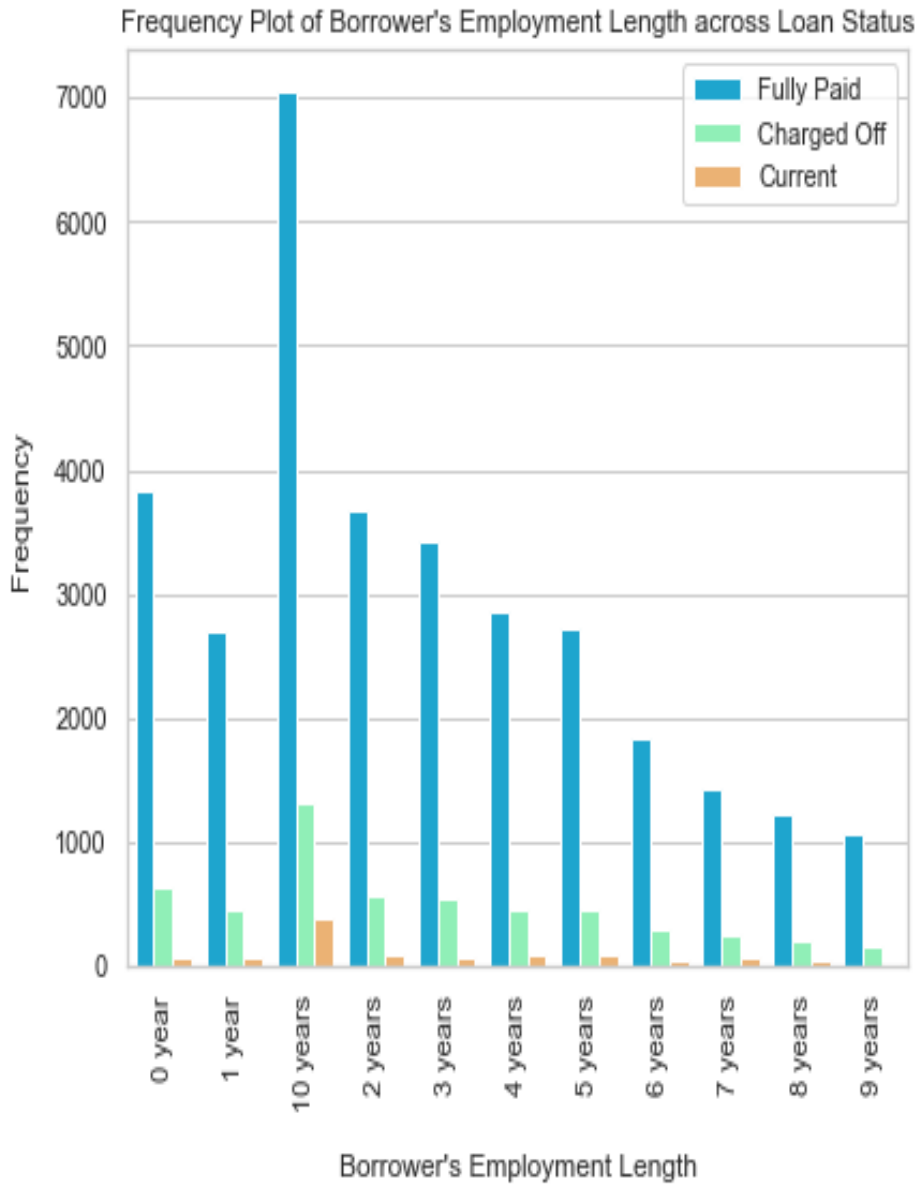
•Grade A and B loans are safe. Within these the sub-grades A4 and B3 have the highest number of loan applicants.

•Initial Univariate Analysis says, of the Grade B, C and D, most of the Charged off applicants belong to Sub Grades:

1. Grade B => B3, B5, B4 2. Grade C => C3, C4, C5 3. Grade D => D3, D4, D5

4. emp_length-



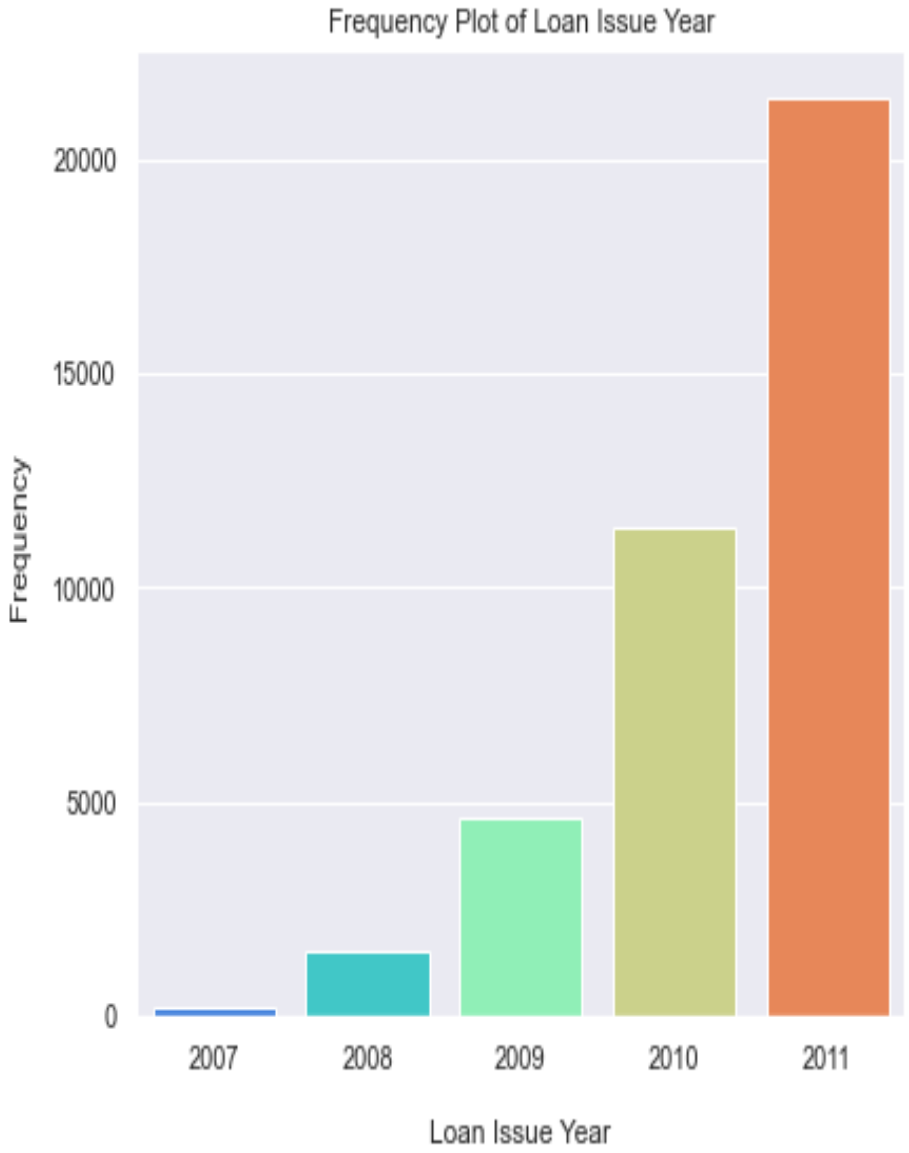


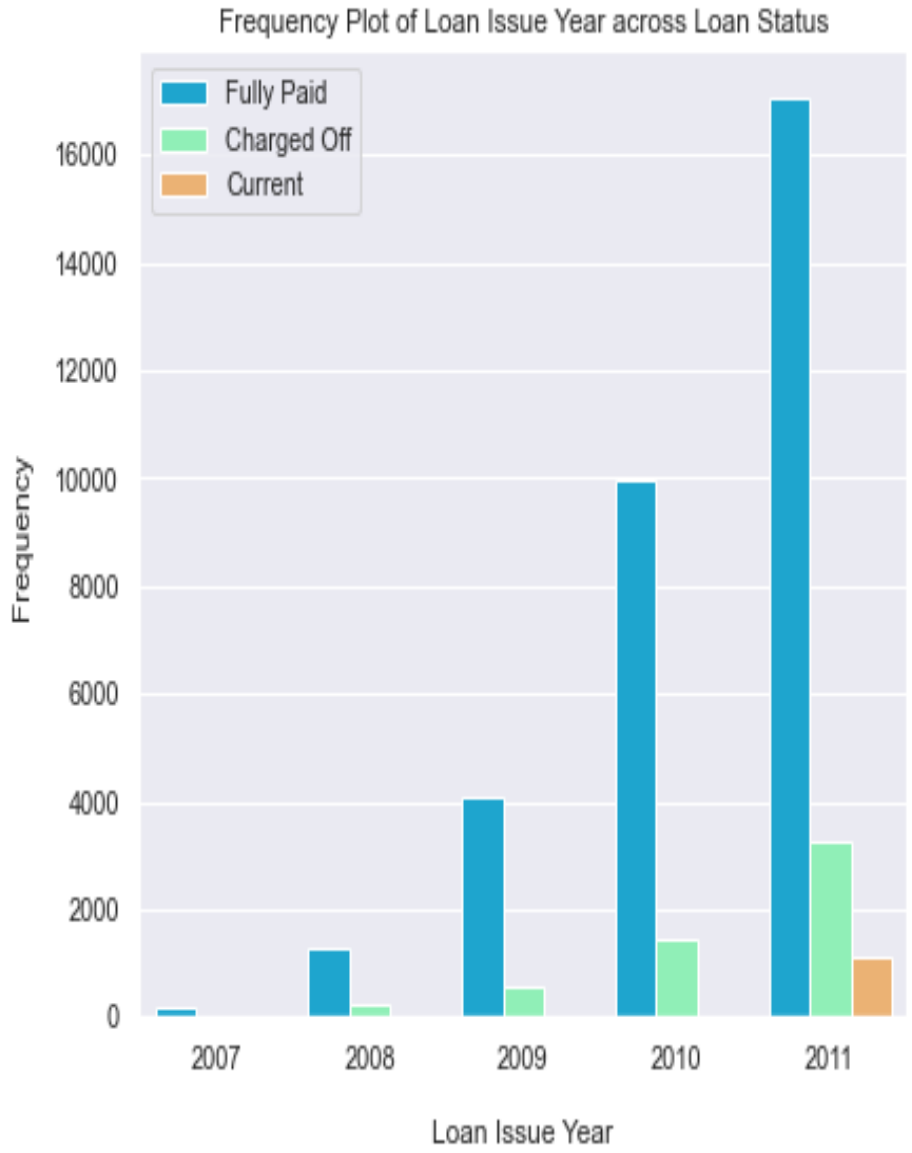
INSIGHTS

•Maximum number of loans are charged off for people having 10+ years of experience.

•or < 1 year

5. issue_yr -





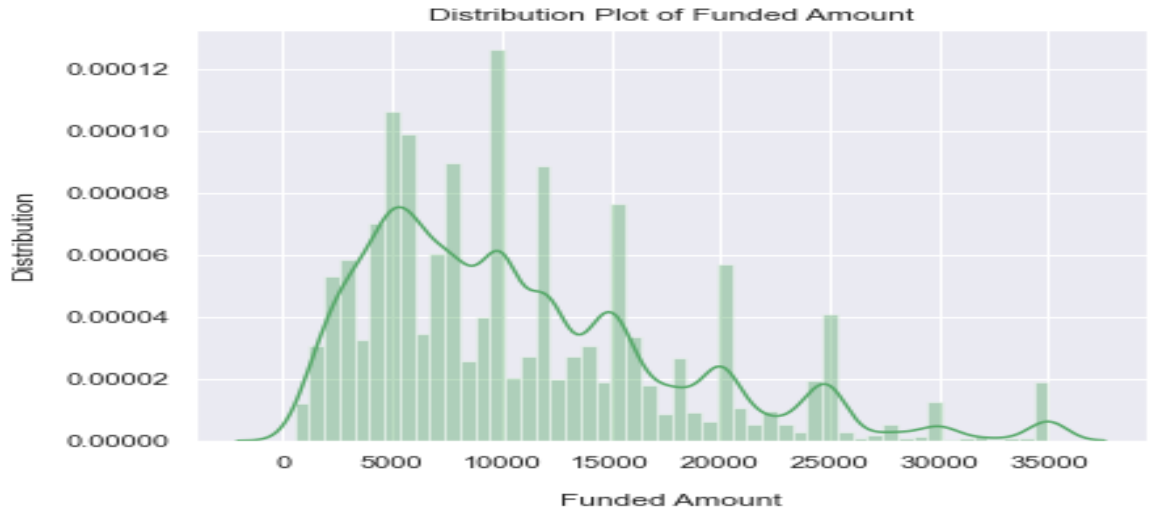
INSIGHTS

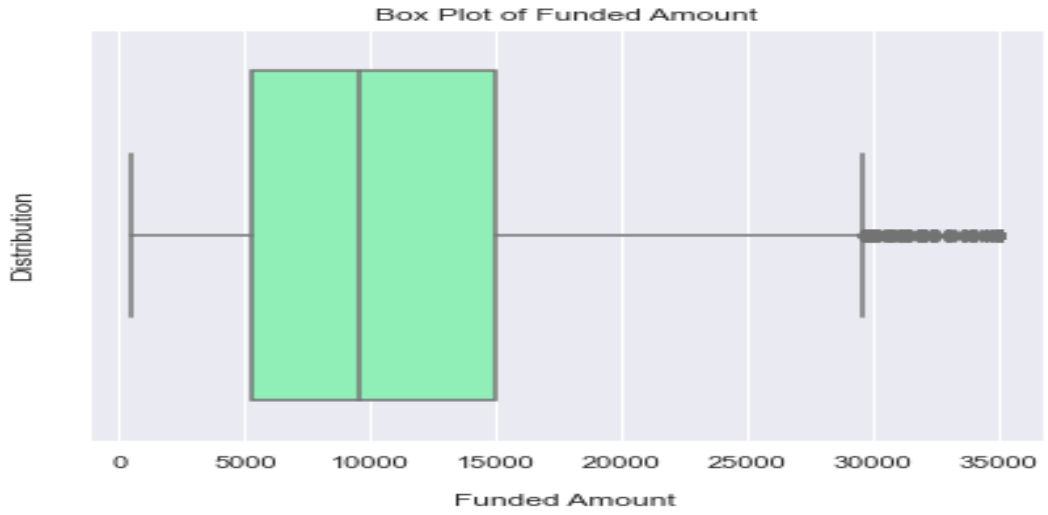
•Avg Interest Rate is increasing for charged off and fully paid loans year-wise.

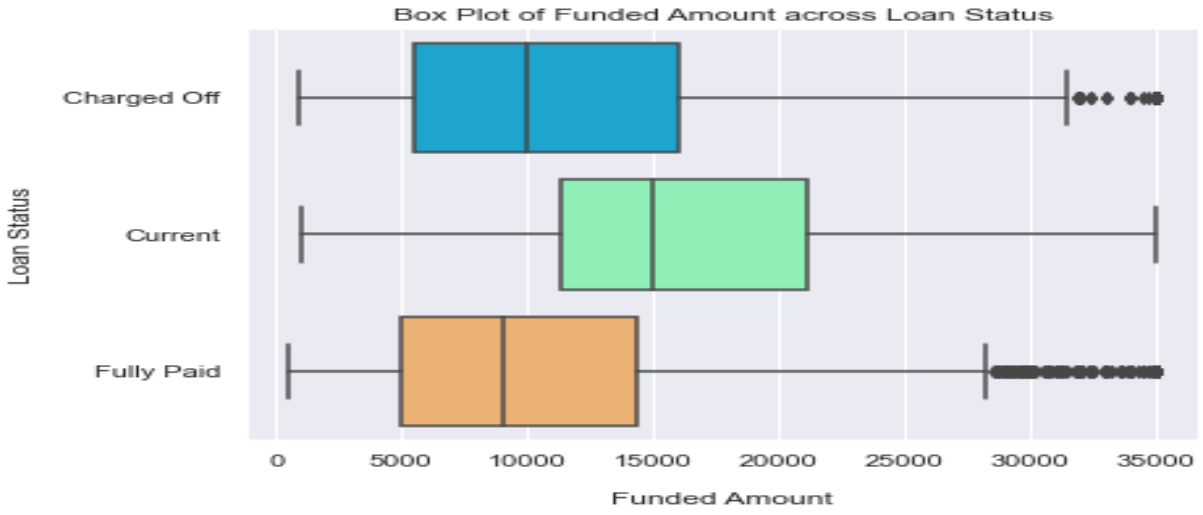
•Charged Off Applicants are more in the Loan issued year 2011.

UNIVARIATE ANALYSIS : Quantiatative or continuous variables –

1. funded_amt







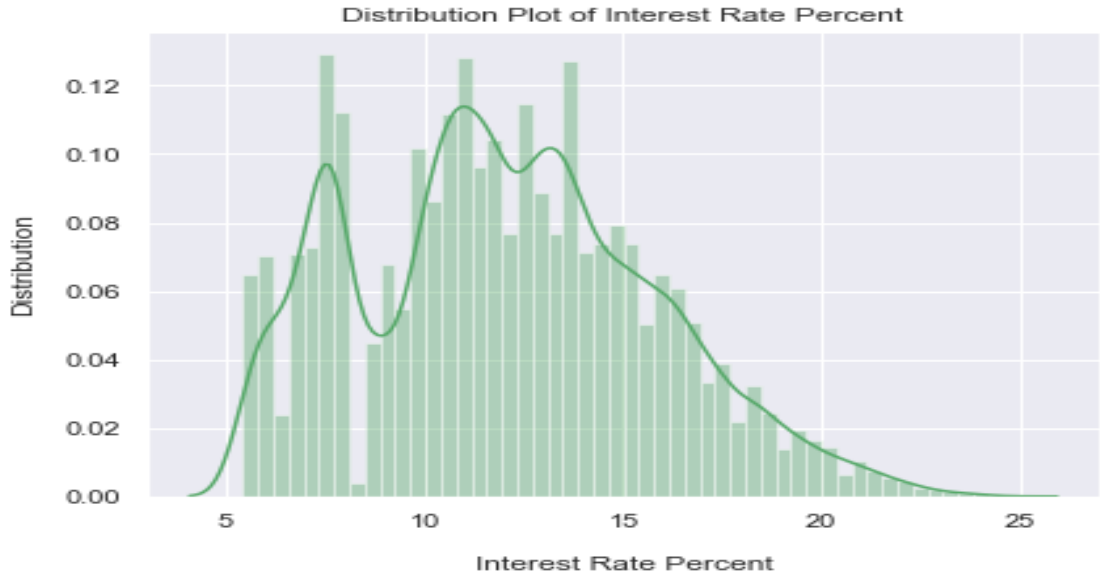
INSIGHTS

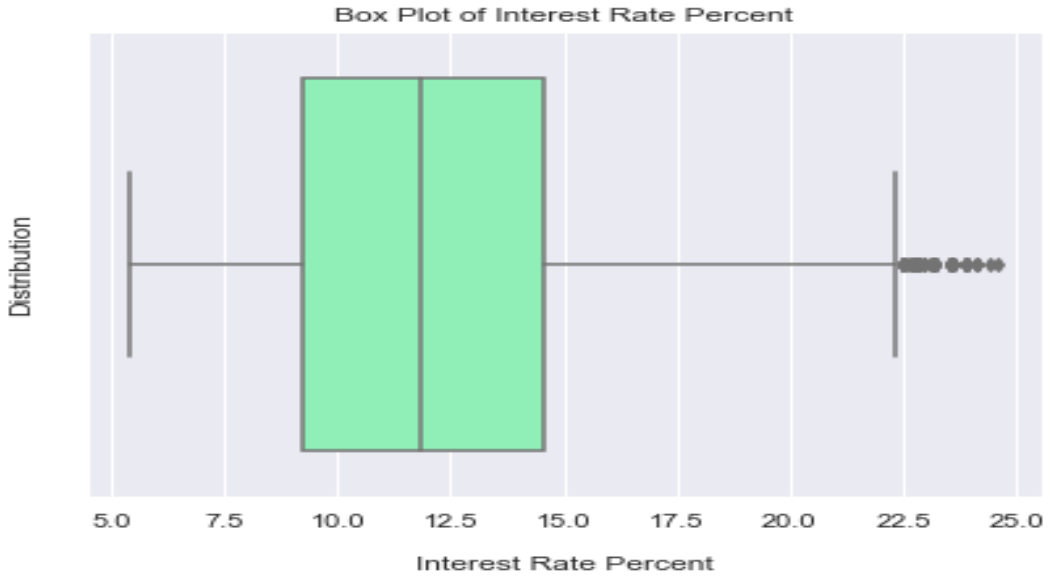
Measure of Central Tendency

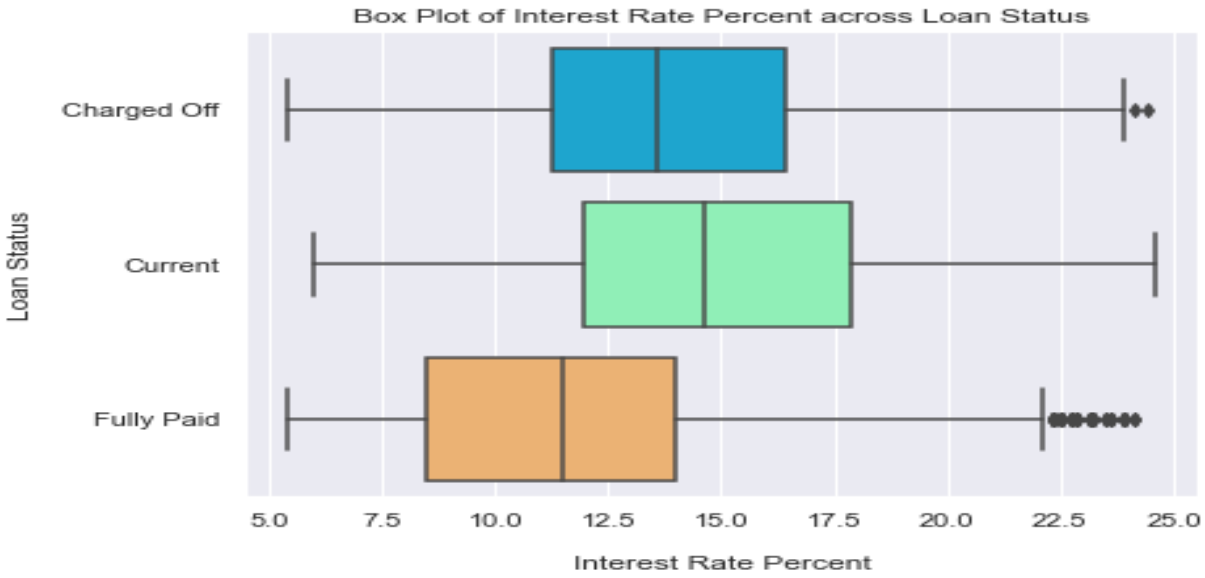
•From the description of funded_amnt, we can observe that the mean (10866.3585) > median (9600.0000). This shows the data is positively skewed or right skewed which could be due to outliers. Henceforth, Median as the measure of central tendency would be the right choice for the respective distribution. Also, the reason being, median is not affected by outliers.

•Charged Off Applicants has a central tendency of Funded Amount as 10,000 with minimum 25th percentile value as 5500 and maximum 75th percentile value as 16000.

2. int_rate_percent





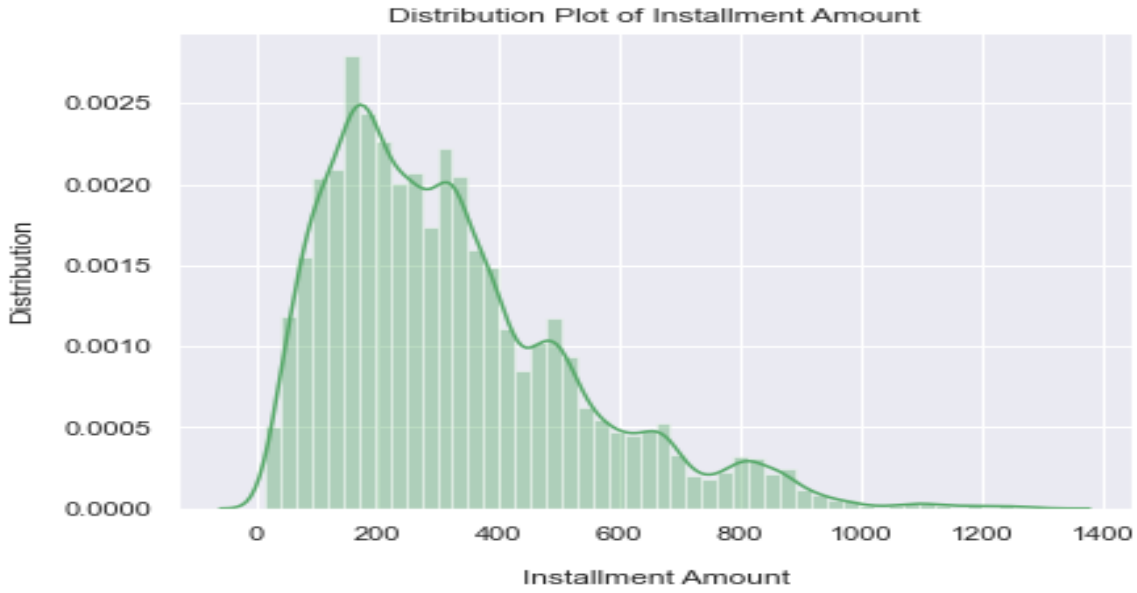


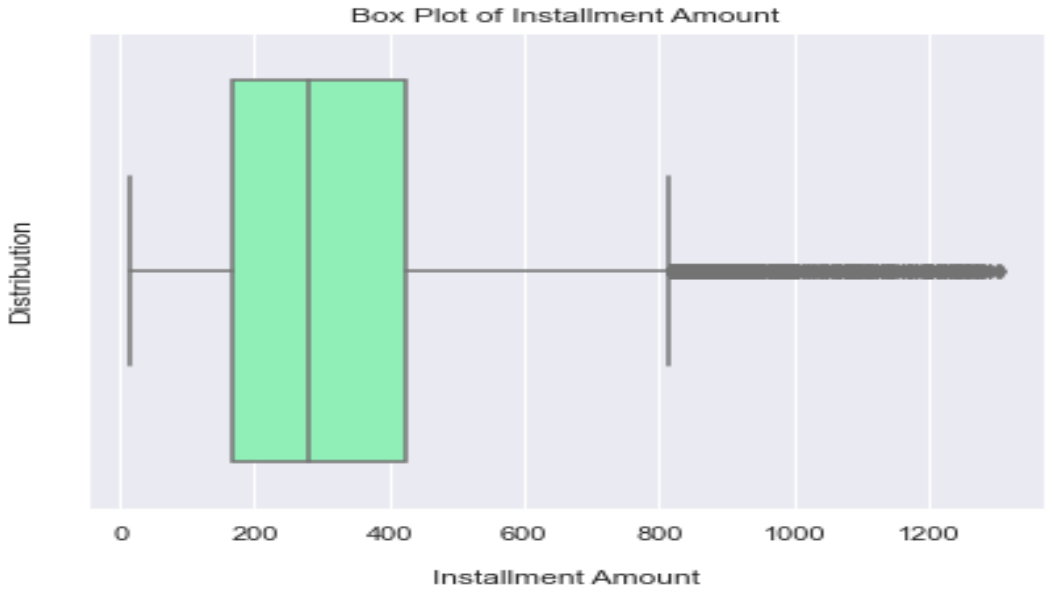
INSIGHTS

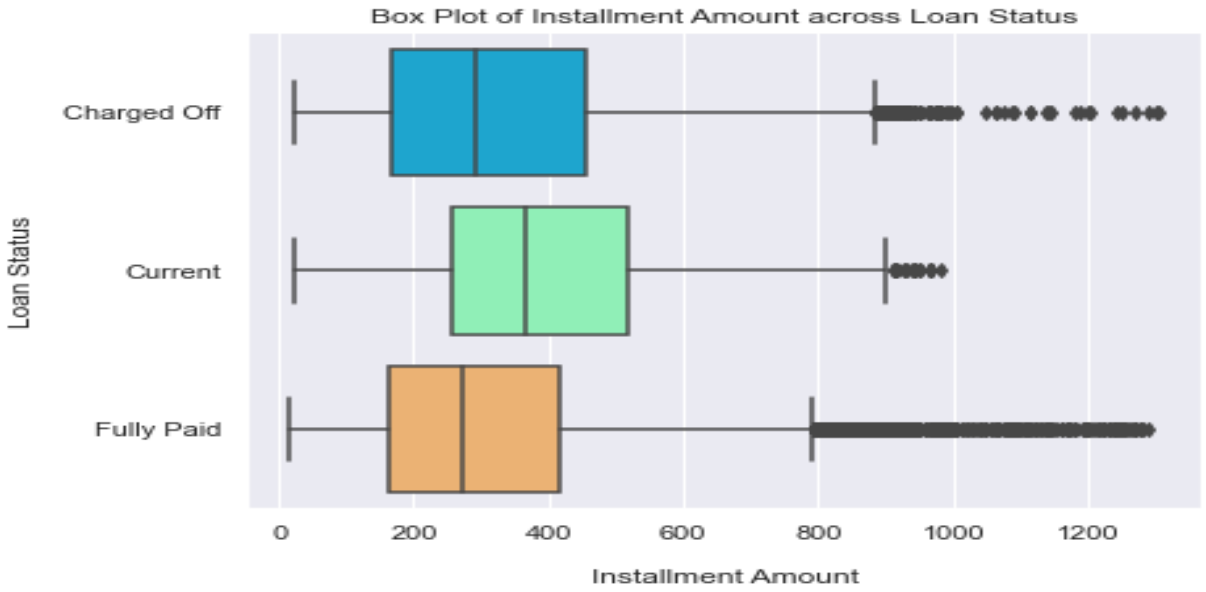
Measure of Central Tendency

Charged Off Applicants has a central tendency of Interest Rate as 13.57% with minimum 25th percentile value as 11.28% and maximum 75th percentile value as 16.40%.

3. installment





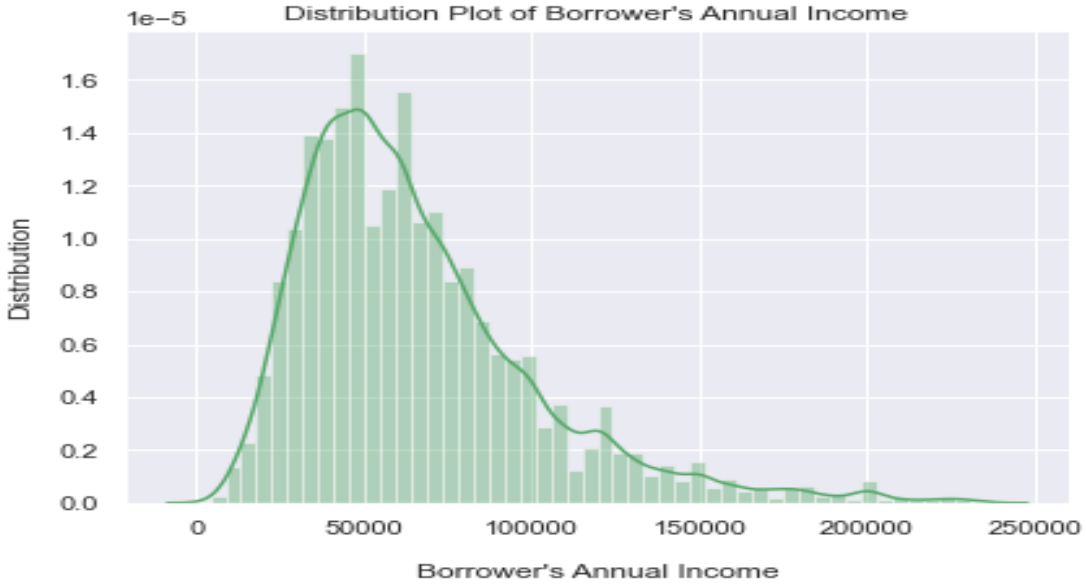


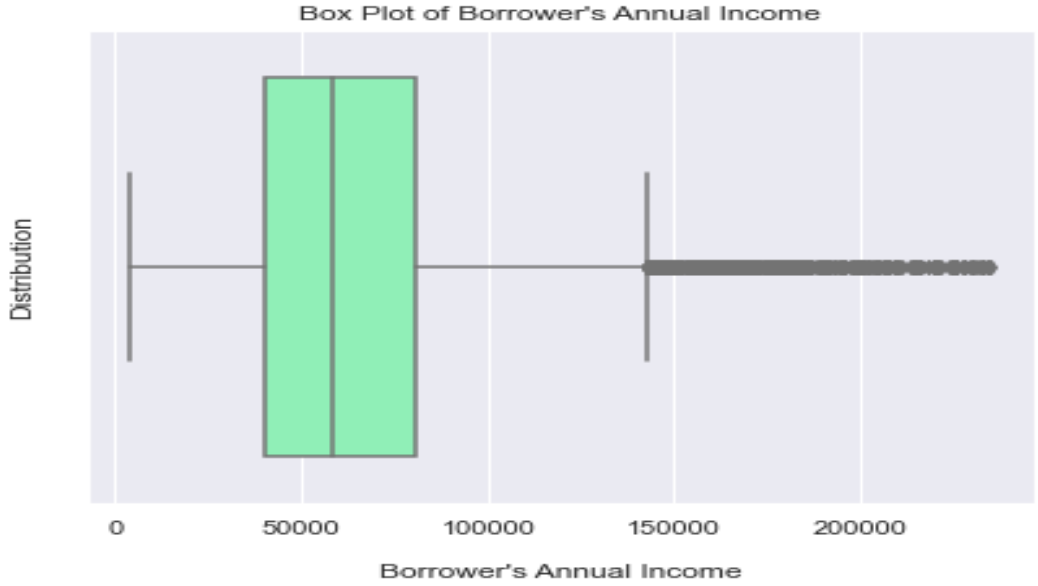
INSIGHTS

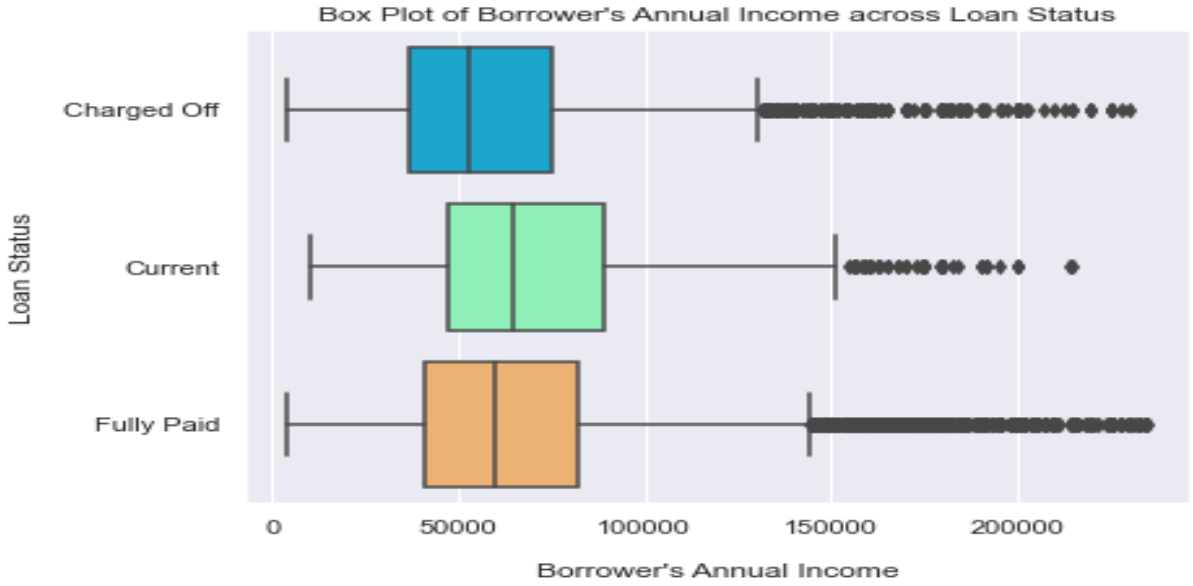
Measure of Central Tendency

Charged Off Applicants have a central tendency of installment as 292.04 with minimum 25th percentile value as 168.45 and maximum 75th percentile value as 454.38.

4. annual_inc





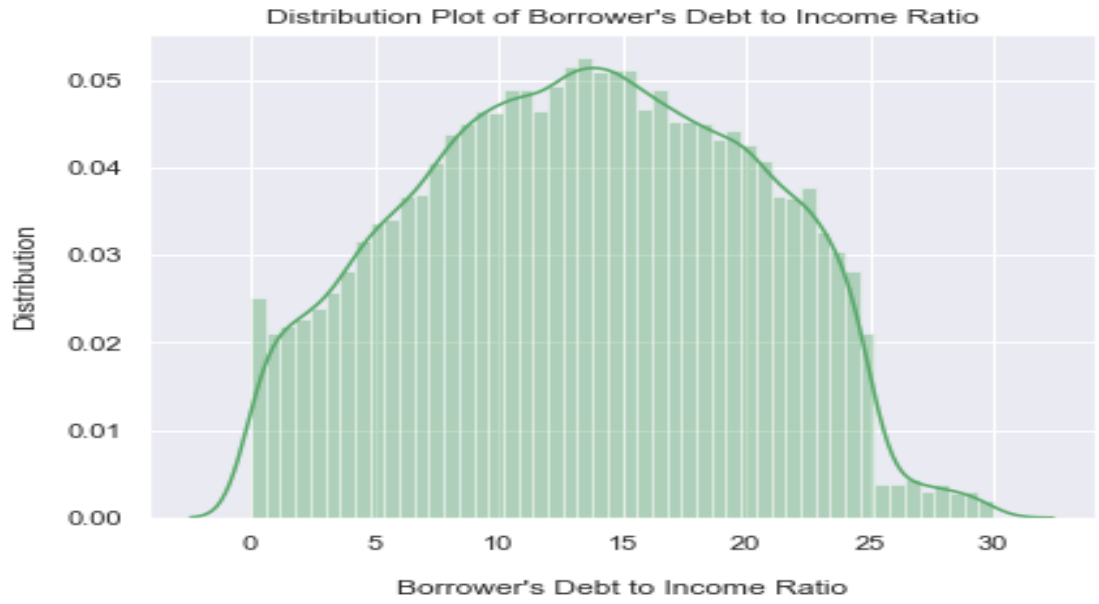


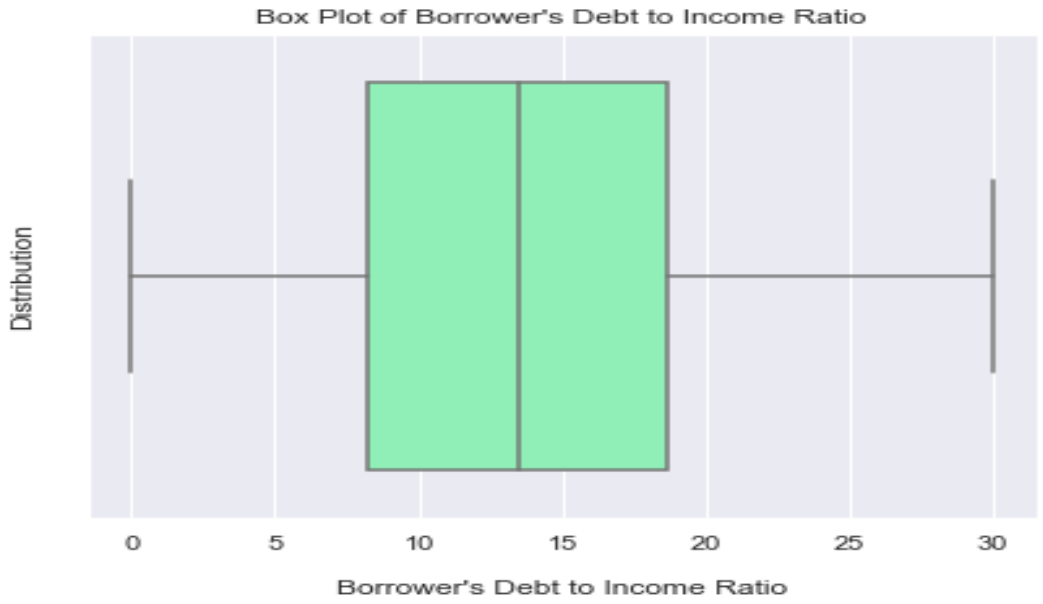
INSIGHTS

Measure of Central Tendency

Charged Off Applicants has a central tendency of Annual Income as 52,800 with minimum 25th percentile value as 37,000 and maximum 75th percentile value as 74,879.

5. dti





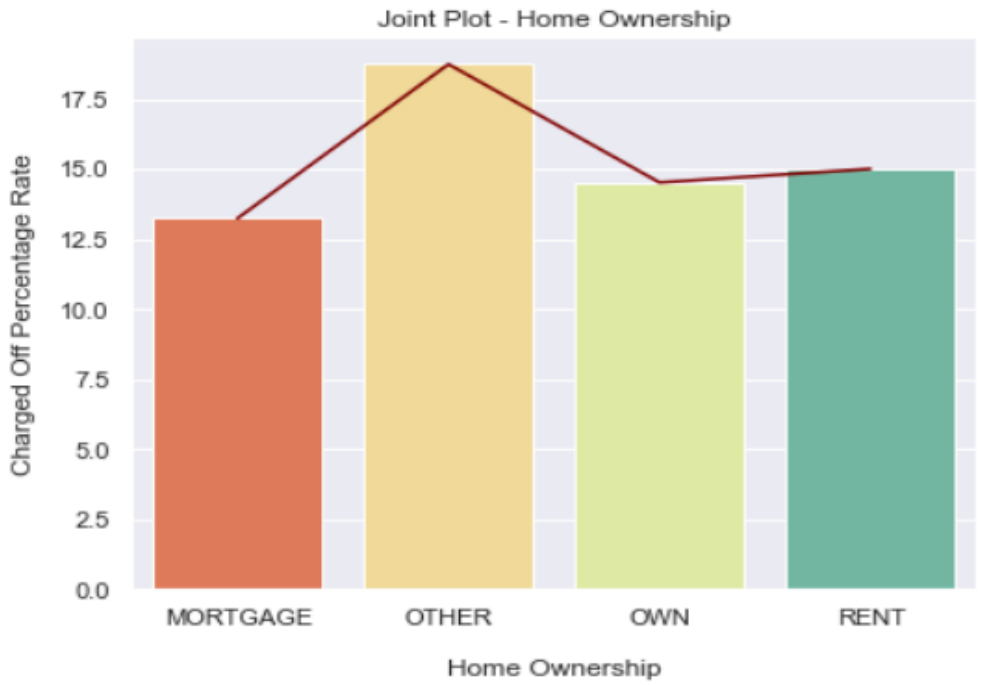
INSIGHTS

Measure of Central Tendency

Charged Off Applicants has a central tendency of dti (Debt-to-Income) ratio as 14.34 with minimum 25th percentile value as 9.13 and maximum 75th percentile value as 19.31.

BIVARIATE ANALYSIS: Categorical Variables

1. home_ownership against Charged Off Percentage Rate

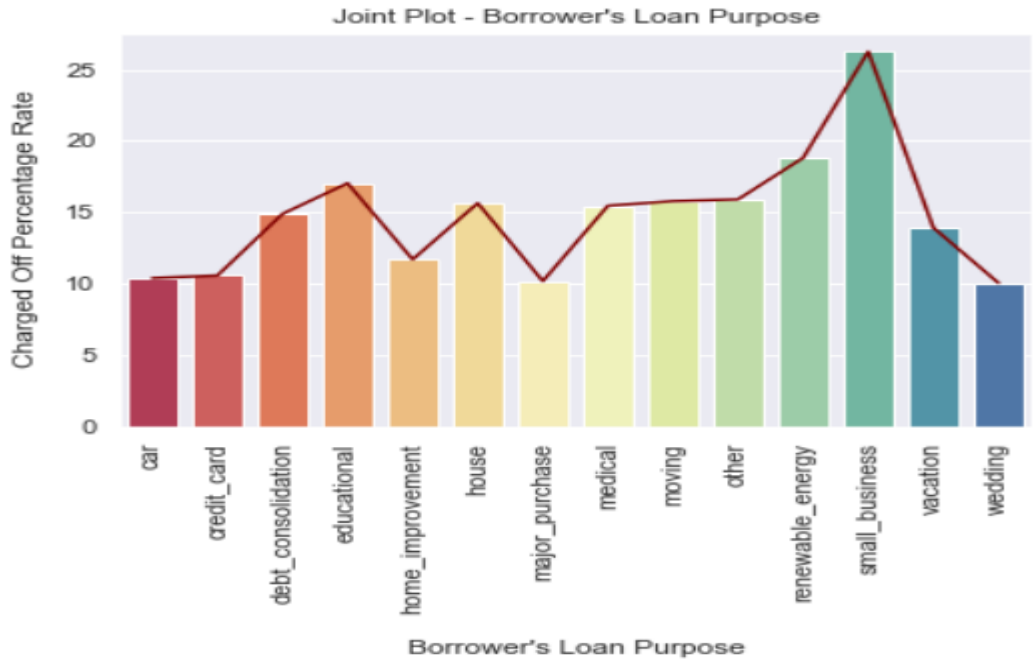


INSIGHTS

Charged Off Rate - Top 3 Borrower's Home Ownership are:

- OTHER = 18.75%
- RENT = 15.01%
- OWN = 14.53%

2. purpose against Charged Off Percentage Rate



INSIGHTS

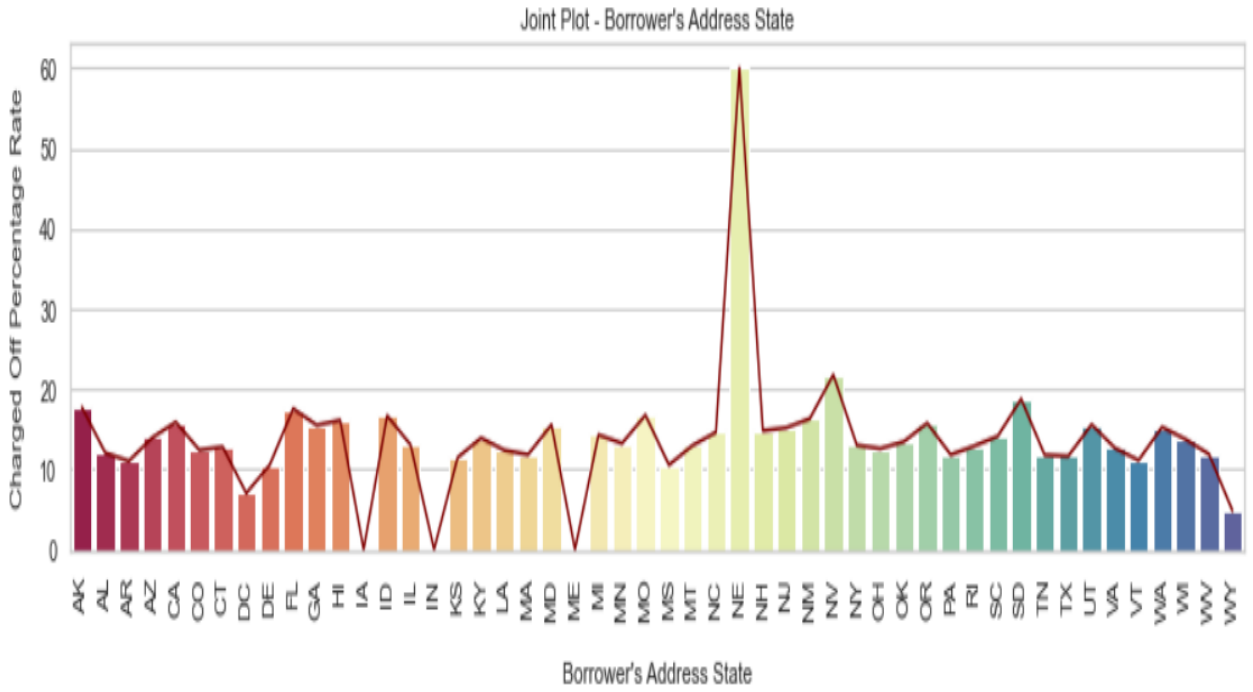
Charged Off Rate - Top 3 Borrower's Purpose:

•small_business = 26.27%

•renewable_energy = 18.81%

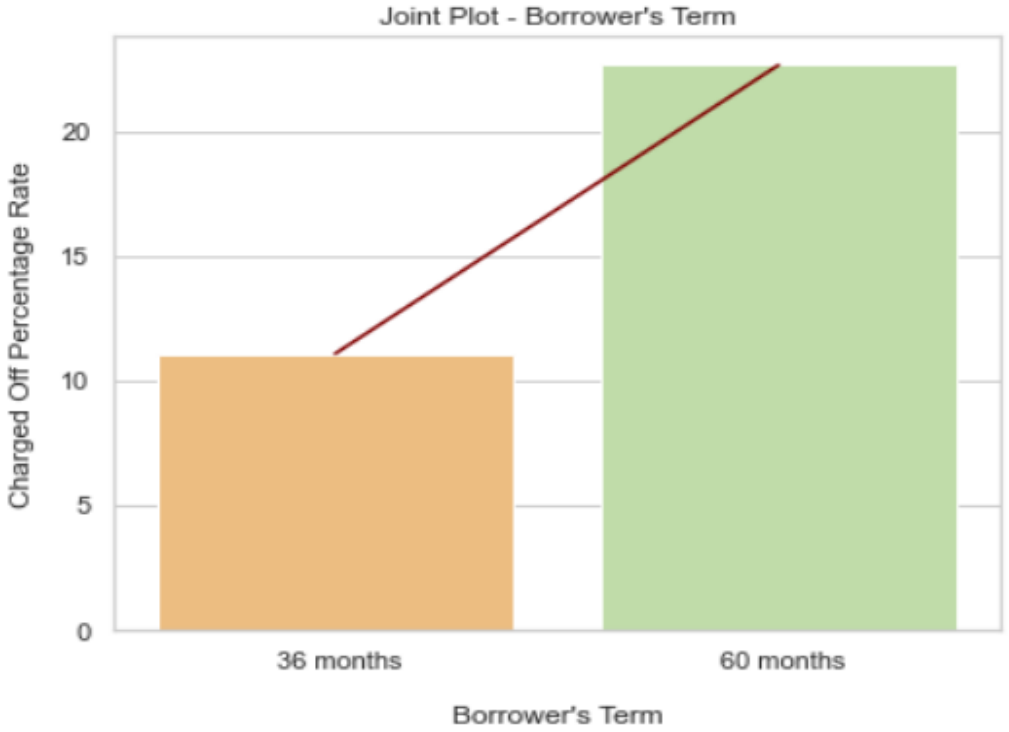
•educational = 17.03%

3. addr_state against Charged Off Percentage Rate



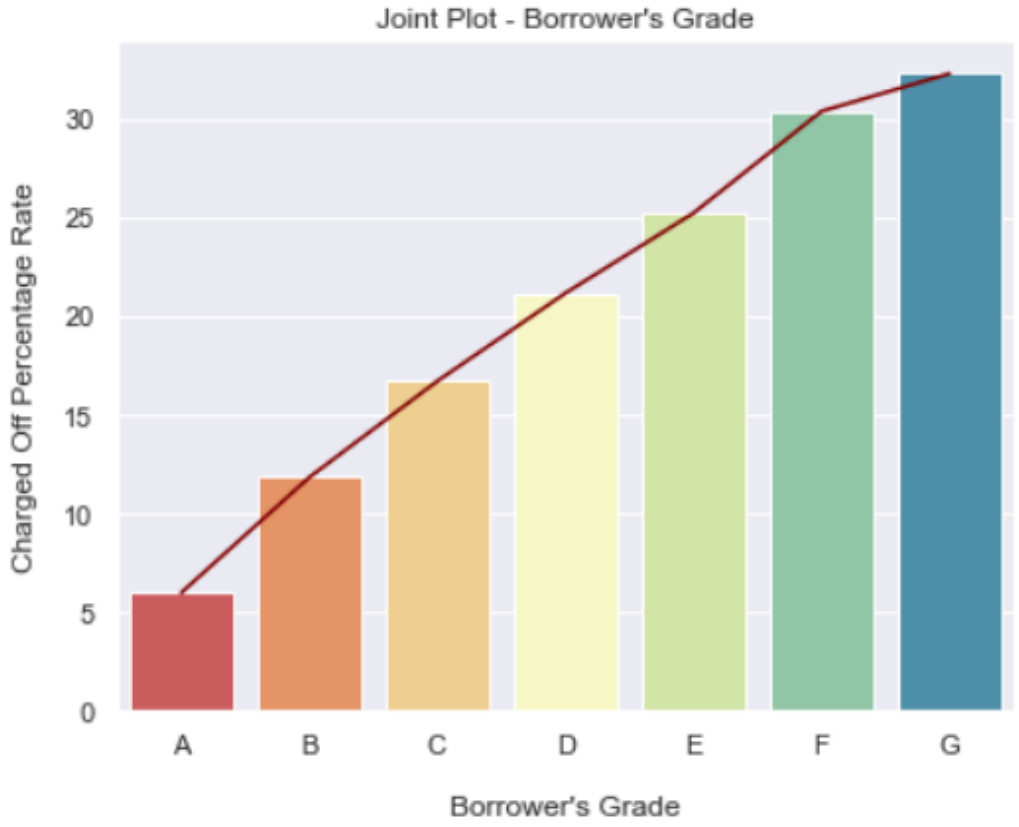
INSIGHTS
Charged Off Rate - Top 5 Borrower's Address State:
NE (Nebraska) = 60.00%
NV (Nevada) = 21.75%
SD (South Dakota) = 18.75%
AK (Alaska) = 17.72%
FL (Florida) = 17.54%
*NE has a total of 5 loans out of which 3 were charged-off.

4. term against Charged Off Percentage Rate



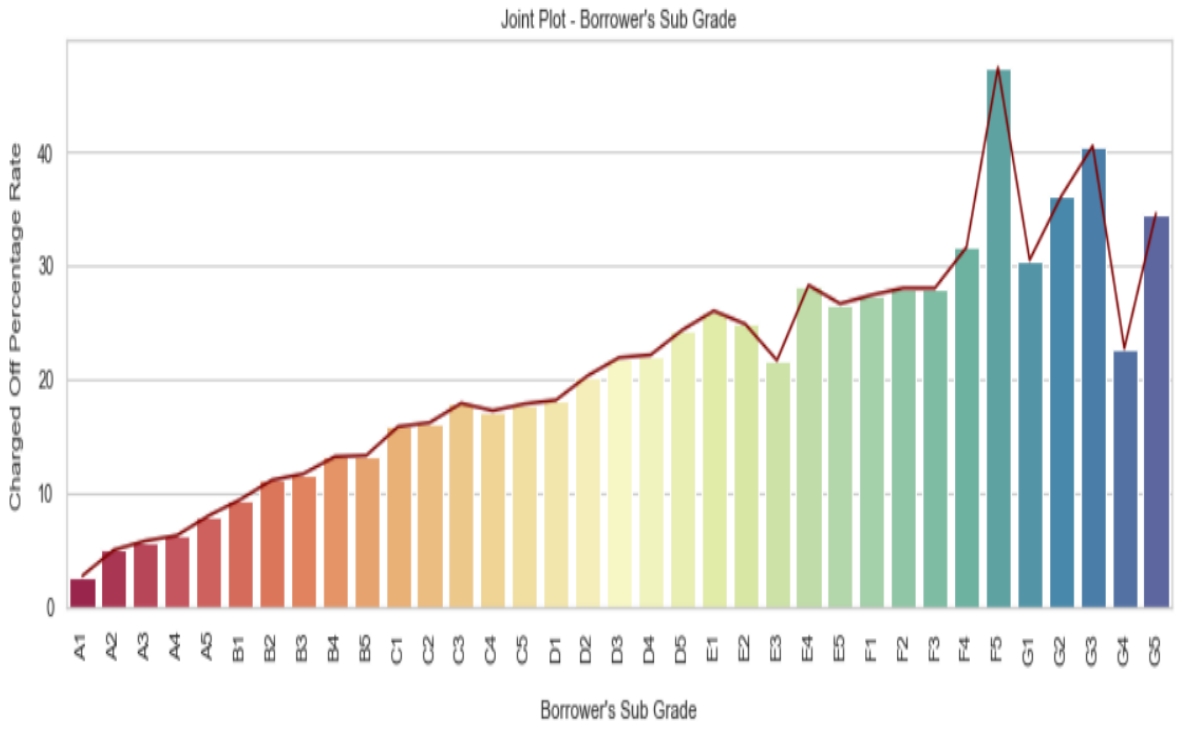
INSIGHTS
Charged Off Rate - Borrower's Top Loan Term:
60 months = 22.70% % of loans getting charged-off for 60 month term i.e. 22.70% is higher as compared to 36 month term.

5. grade against Charged Off Percentage Rate



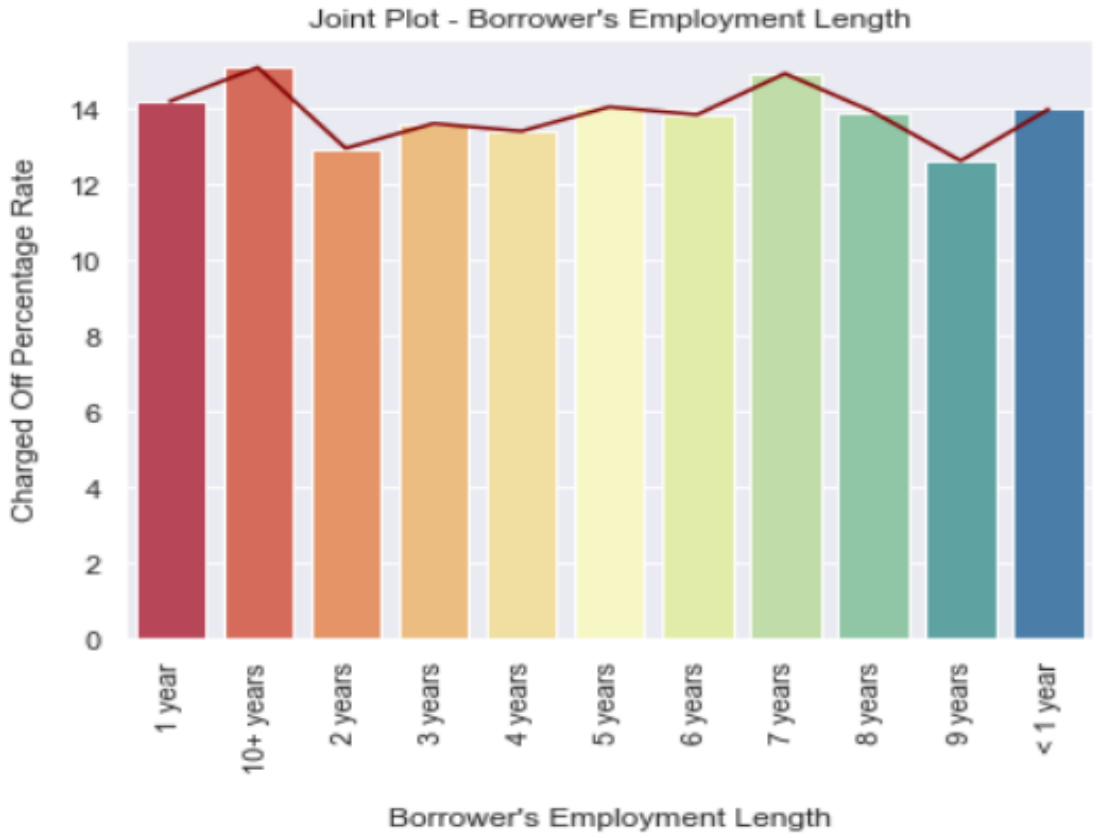
INSIGHTS
Charged Off Rate - Grade: As the Grade increases, Charged Off Rate increases. Top Order:
1. G 2. F 3. E 4. D 5. C 6. B 7. A

6. sub_grade against Charged Off Percentage Rate



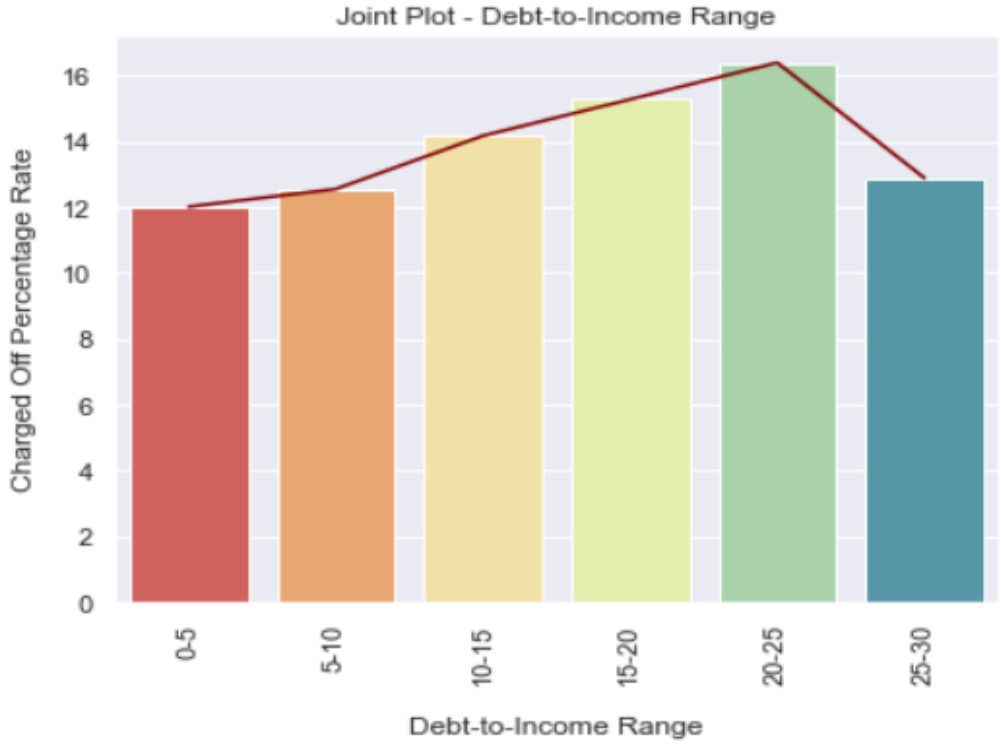
INSIGHTS
Charged Off Rate - Sub Grade: As the Grade and the Sub Grades increases, Charged Off Rate increases. Top Order:
1. G 2. F 3. E 4. D 5. C 6. B 7. A

7. emp_length against Charged Off Percentage Rate



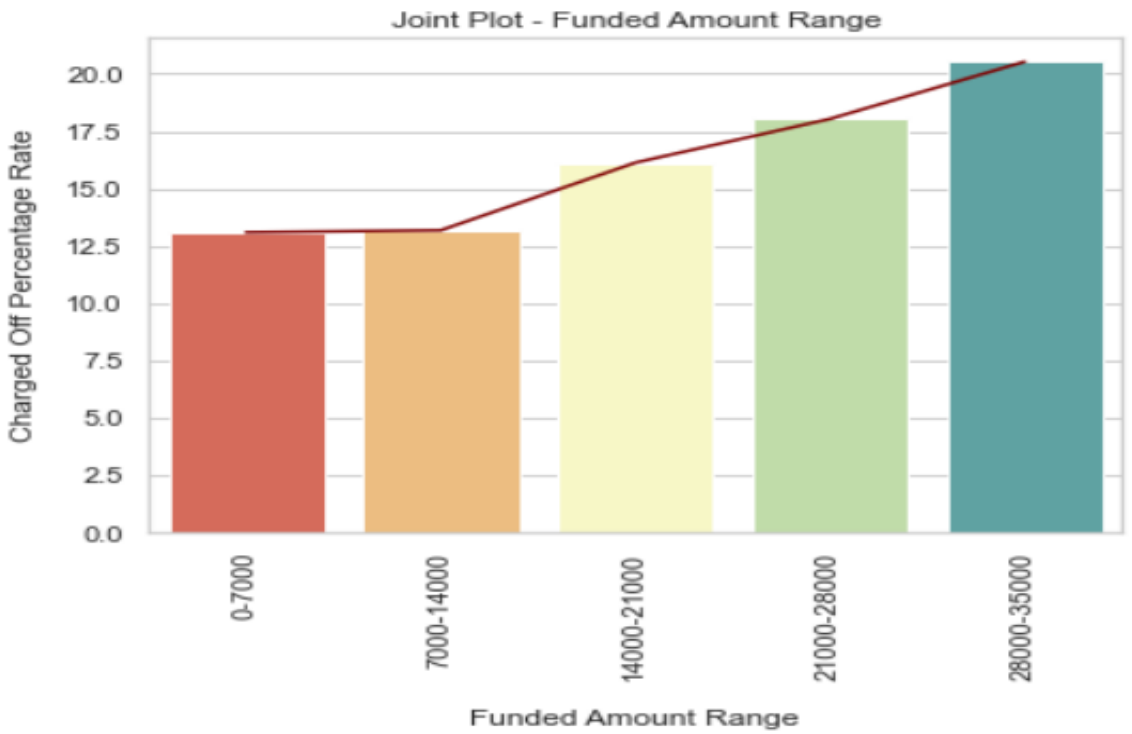
INSIGHTS
Top 3 employment length belonging to Charged Off category : 1. 1 year 2. 0 year (< 1 year) 3. 10 Years (> 10 years)

8. dti against Charged Off Percentage Rate



INSIGHTS
As dti (Debt-to-Income) value increases, Charged Off Rate increases
Exception is in range 25-30. This is because the data of population in this range is not high.

9.funded_amnt against Charged Off Percentage Rate



INSIGHTS

As the Loan or Funded Amount increases, Charge Off Rate Increases as well.

Charged Off Rate - Top 3 Loan Funded Amount Range:

28000-35000 = 20.53%

21000-28000 = 18.04%

14000-21000 = 16.13%

10. int_rate_percent against Charged Off Percentage Rate

INSIGHTS

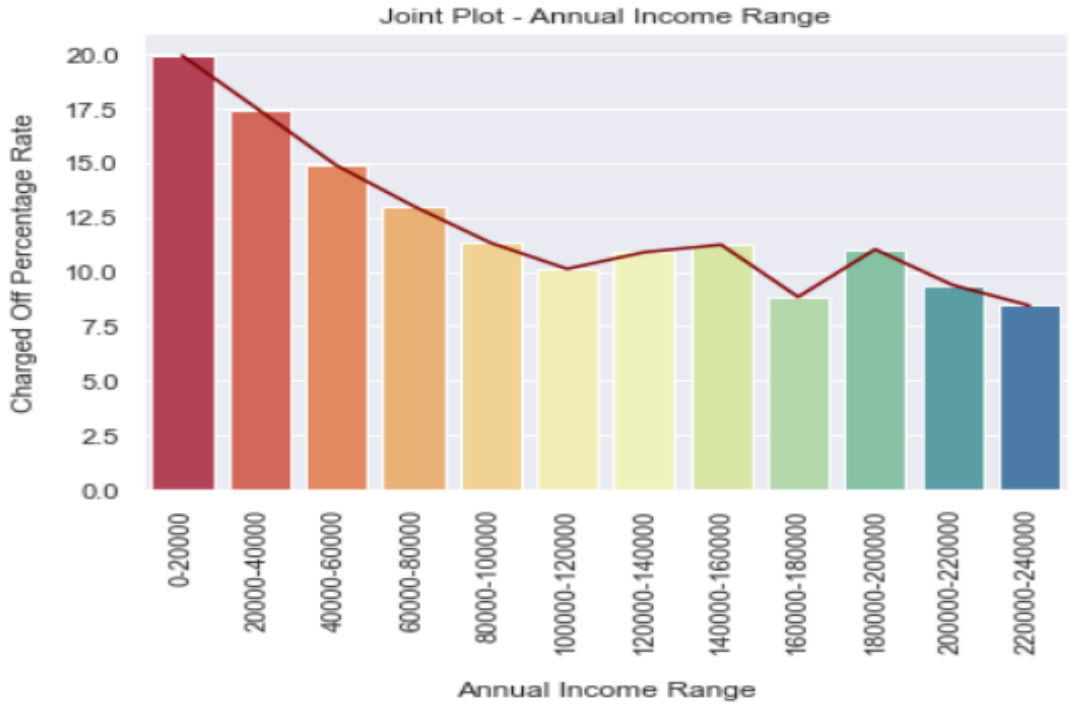
As the Interest Rate increases, Charge Off Rate Increases

Charged Off Rate - Top 2 Interest Rate Range:

20-25 = 34.44%

15-10 = 23.51%

11. annual_inc against Charged Off Percentage Rate



INSIGHTS

As the Anual Income decreases, Charge Off Rate Increases

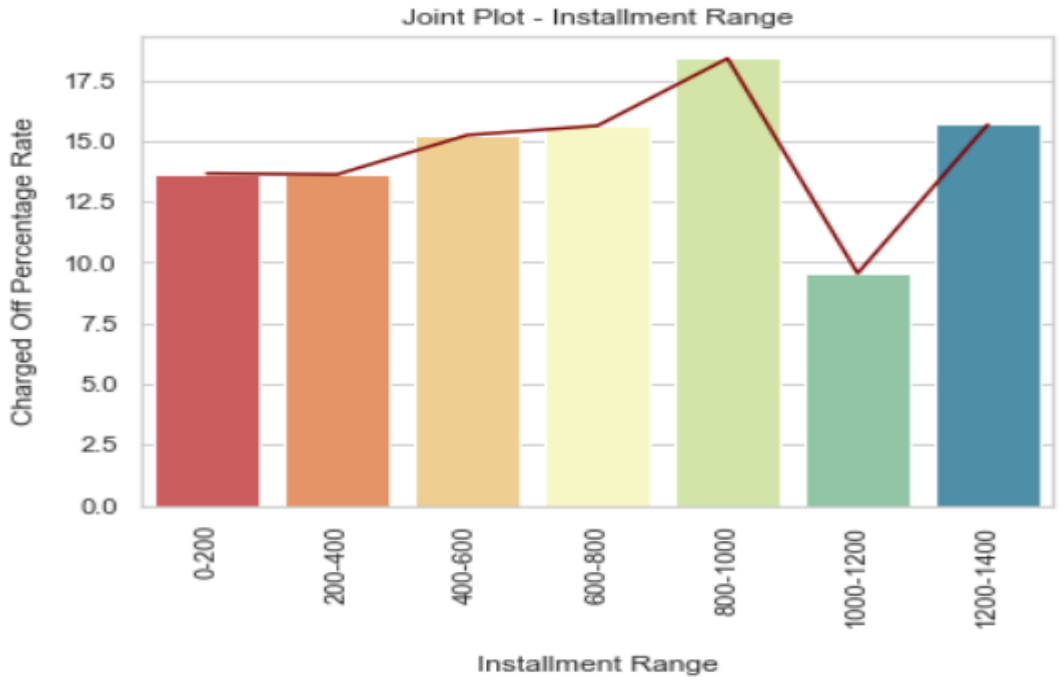
Charged Off Rate - Top 3 Annual Income Range:

0-20000 = 19.93%

20000-40000 = 17.43%

40000-60000 = 14.90%

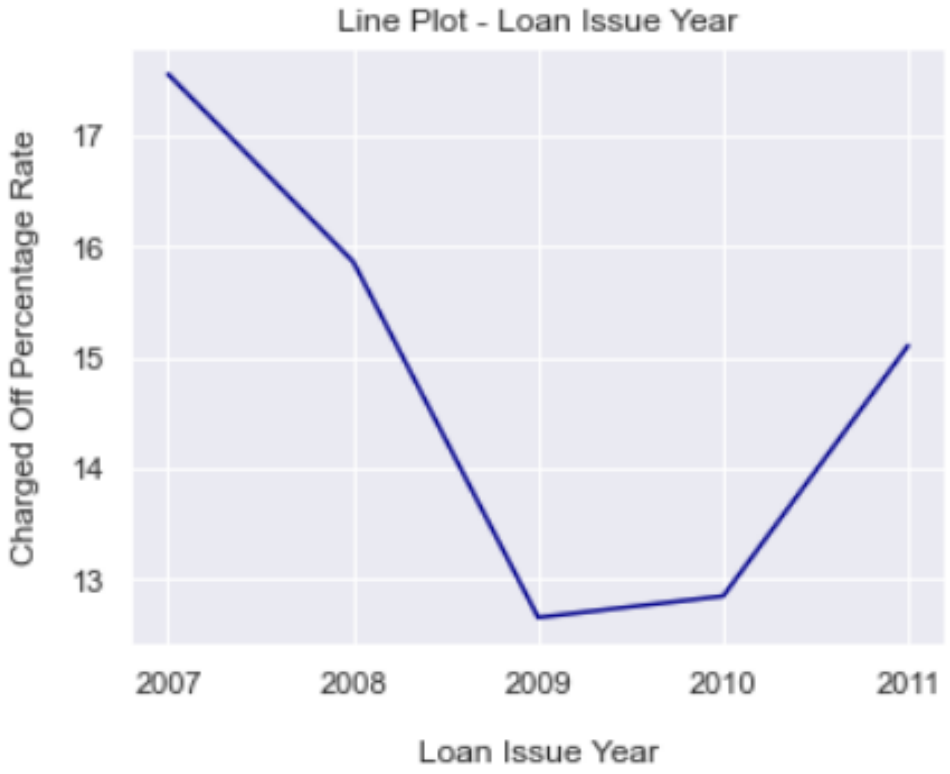
12. installment against Charged Off Percentage Rate

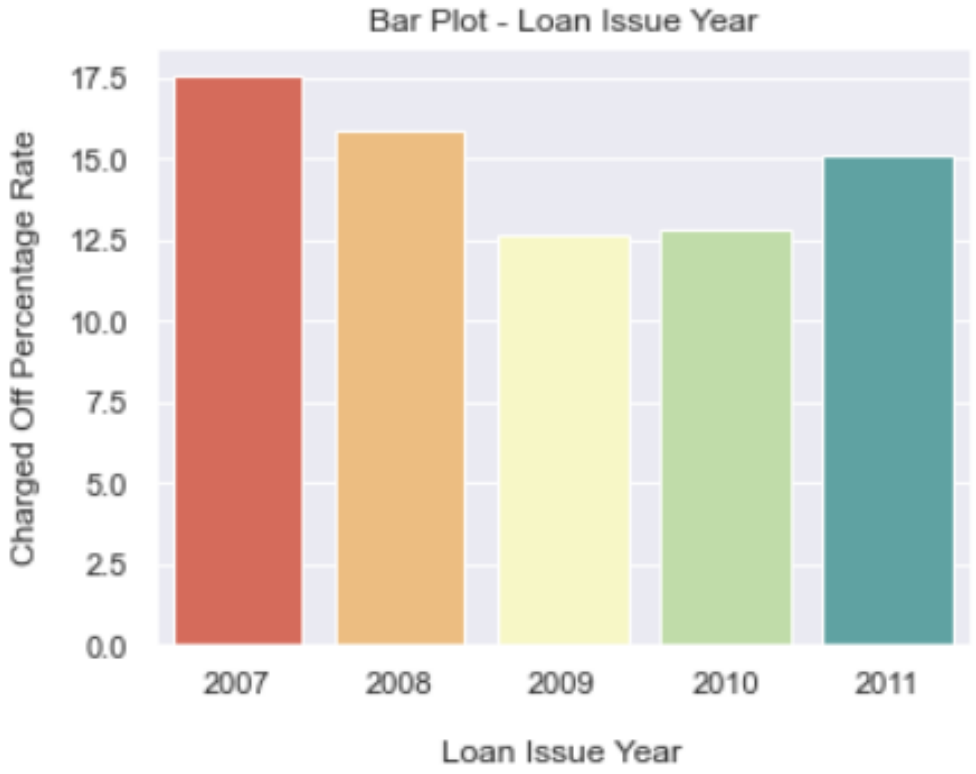


INSIGHTS

As the installment increases, Charged Off Rate increases.

13. issue_yr against Charged Off Percentage Rate

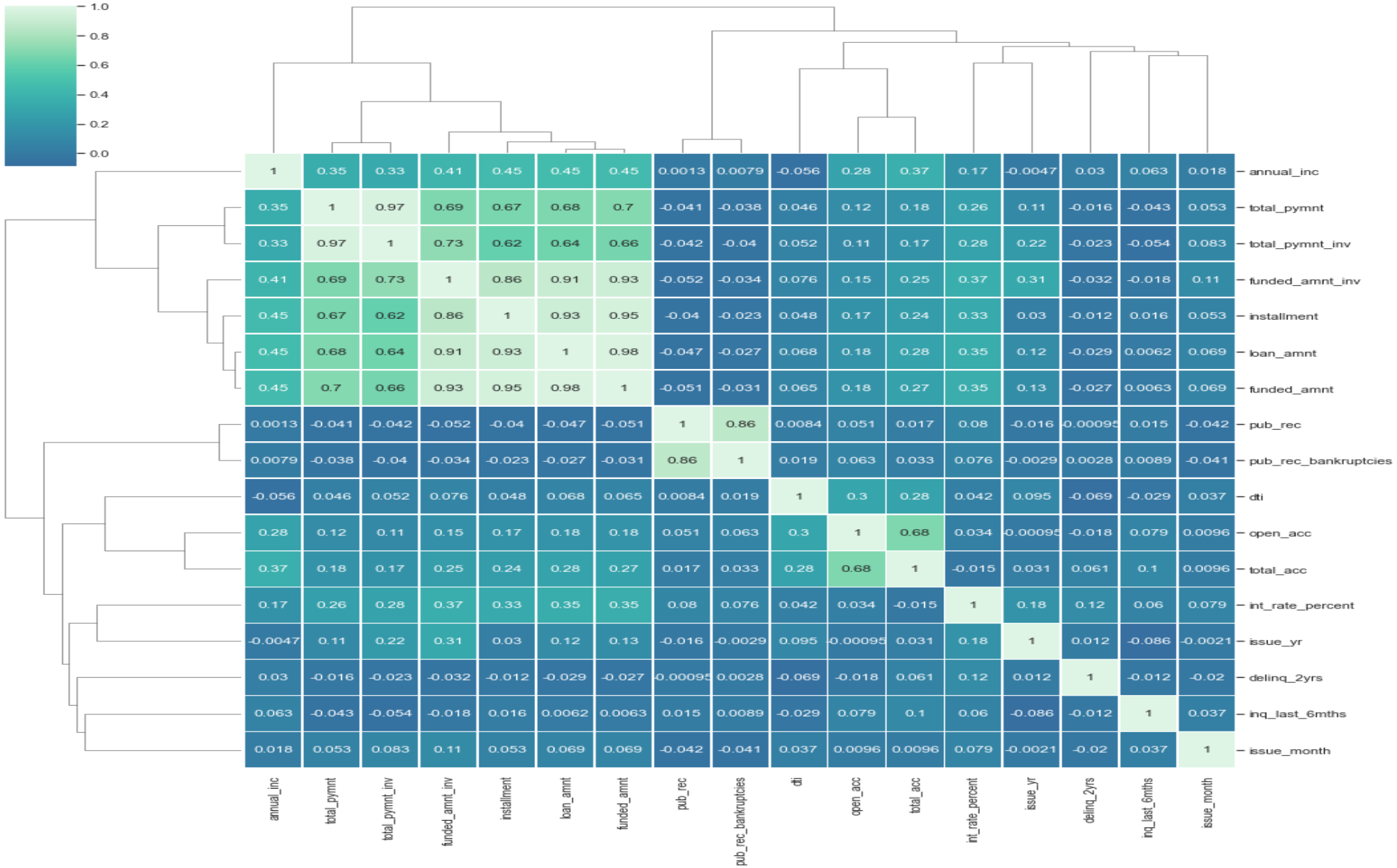




INSIGHTS

Charged off percentage reduced from year 2007 to 2009 and then gradually increased from 2009 to 2011. It is not adding any substantial value to the Charged off Rate.

Bivariate Analysis: **Continuous Variables :**Cluster Map:



INSIGHTS

- installment, funded_amnt, loan_amnt and funded_amnt_inv are highly correlated (positive) to each other. They form a cluster.
- dti (Debt-to-Income Ratio) is high correlated (positive) to issue_yr and vice-versa.
- Annual Income is negatively correlated with DTI.
- int_rate_percent is negatively correlated to total_acc.

CONCLUSION:

- Significant variables to consider for loan prediction: 1. Loan Purpose 2. Employment Length 3. Interest Rate 4. Annual Income 5. Grade & Sub-grade 6. Term 7. Address State
- In Detail:
- Loan Purpose: Charged Off Rate is high for loan purpose:
  - small_business = 26.27%
  - renewable_energy = 18.81%
  - educational = 17.03%
- Employment Length: We see that the Charged Off rate increases for peopple with less employment length. Charged Off Rate is high for employment length:
  - 0 year (< 1 year)
  - 1 year
- Interest Rate: As the Interest Rate increases, Charge Off Rate Increases Charged Off Rate is high for interest rate:
  - 20-25 = 34.44%
  - 15-20 = 23.51%
- Annual Income: So, as the Anual Income decreases, Charge Off Rate Increases Charged Off Rate is high for Annual Income:
  - 0-20000 = 19.93%
  - 20000-40000 = 17.43%
  - 40000-60000 = 14.90%

- Grade & Sub-grade: Most of the Charged off applicants belong to Grade B, C and D. Of the Grade B, C and D, most of the Charged off applicants belong to Sub Grades:
  - Grade B => B3, B5, B4
  - Grade C => C3, C4, C5
  - Grade D => D3, D4, D5 Plus, as the Grade increases, Charged Off Rate increases. Top Order:
  - G
  - F
  - E
  - D
  - C
  - B
  - A
- Term: Charge Off Rate increases as the Term increases.
- Address State: Note- NE has a total of 5 loans out of which 3 were charged-off. Charged Off Rate is high for Address State:
  - NE (Nebraska) = 60.00%
  - NV (Nevada) = 21.75%
  - SD (South Dakota) = 18.75%
  - AK (Alaska) = 17.72%
  - FL (Florida) = 17.54%

RECOMMENDATIONS:

- Lending Club should be cautious of the loans where the purpose is Small Business as the percentage of a loan being charged off is maximum. Accepting loans for the purpose of Weddings, major purchase, car and credit card is highly recommended.
- Higher the loan amount, the higher the chances of loan being charged off. Therefore Lending Company should consider accepting loans of lower amount.Hence the risk factor remains low for lending club.
- Lending Club should consider accepting more loans from applicants whose annual income is greater than 100000 as their probablity of charge off is minimum mostly.
- Lending Club should consider accepting more loans where interest rate is less than 6.7% as their probablity of charge off is minimum.
- Lending Club should consider accepting more loans of grade A and B. It should be vary of loans falling in grades E,F and G.
- Lending Club should consider accepting more loans from people who owns a house.
- Lending Club should accept more loans for the term of 36 months as the % of charged off loans is less and the no. of loan applicants are more.
- People with more number of public derogatory records are having more chance of filing abankruptcy. Lending club should make sure there are no public derogatory records for borrower.

Thank You