Loan Lateness Prediction – Final Report – Capstone 2

1. Problem Statement

Statement: Predicting which loans that are likely to be late

The aim of this project is to find the best model that predicts which loans are most likely to be late.

The major clients to benefit from this project would be mostly financial institutions.

By being able to flag loans that will most likely be late, institutions can properly budget for those losses or even provide an early warning to the consumer.

2. The Dataset

2.1. Where was the data found?

The dataset being used is actual loan data from LendingClub which is a loan issuing institution. The dataset weas found on Kaggle.(https://www.kaggle.com/wendykan/lending-club-loan-data)

2.2. What does the dataset contain?

The dataset contains the following:

1. *Issued Loan data to everyday consumers from 2007-2017.*

The data contains 2.26million observations and 145 features. This is a significant amount of data which will be more than enough for modelling.

The features contain information regarding the Loanee's financial status which include employment information, debt-to-income ratio, bankruptcies, previously missed payments etc.

2. *Description of each feature*

This will aid with features that may not be easily comprehendible from the feature name.

2.3. How was the data cleaned?

2.3.1. Null Values

The following steps were taken to deal with missing values:

1. Split the data into training/testing sets (25% test size). This was done to fill Nan values using aggregations from the training set
2. Features that had more than 75% with Null observations were dropped.
3. Null values in sparse features (ex bankruptcies) with rare integers(0-2) were filled with 0.
4. Null values in features with mostly integers were filled with the mean rounded to the nearest integer.
5. Null values in features with mostly floats were filled with the mean.
6. Null values in features containing strings were filled with 'Not Given'.

2.3.2. Datatypes

The following steps were taken to deal with datatypes:

1. DateTime columns were of type Object so those were changed to timestamps.
2. Integer and Floats were changed from 64-bit to 32-bit to save on storage space.

2.3.3. Grouping Employment Titles

The training set had **411,581** unique employment titles. This will be disastrous when being modelled. The following steps were taken to reduce the number of unique titles:

1. Change all values to lower case.
2. Strip all titles of empty space.

These two steps reduced that to **331,569.** This is a reduction of 80,000 unique titles. To further reduce this, **12** lists of common employment keywords were generated to suit the following groupings:

A) Executive (eg CEO, CFO, Owner, President)
B) Assistant
C) Senior (Manager, Supervisor, Leader)
D) Skilled (Engineer, Programmer, Lawyer)
E) Technical (Mechanic, Builder)
F) Health (Nurse, doctor, Pharmacy)
G) Business (Finance, Accounting, Legal)
H) Administrative (Clerk, Office)
I) Education (Teacher, Professor)
J) Low skill (Waiter, Bartender)
K) Federal (Officer, Army, Airforce)
L) Other titles

2.3.4. Grouping Target Variable

The target variable had 9 unique entries. The requirement is to loans as either good or bad.

Good Loans are those labelled:

- 'Fully Paid',
- 'Current'
- 'Does not meet the credit policy. Status:Fully Paid'

Bad Loans are those labelled:

- 'Default'
- 'Late (31-120 days)'
- 'In Grace Period'
- 'Late (16-30 days)'
- 'Charged Off'

Now the target variable is either classified as good or bad. These were the steps taken to clean the data in preparation for feature engineering and modelling.

3. EDA

Given that this dataset contains 145 features, it would be extremely time consuming analyzing each feature. This section was structured by analyzing significant features while asking a series of questions and answering them with visuals. Details will be provided regarding the following:

- Loan amounts
- Terms
- Interest rates
- Installments
- Grades and subgrades
- Employment Analysis
- Home Ownership
- Annual income
- Loan Status
- Purpose of Loan
- Location Analysis of Loans
- How have features changed over time?
- Are there features with significant correlations?

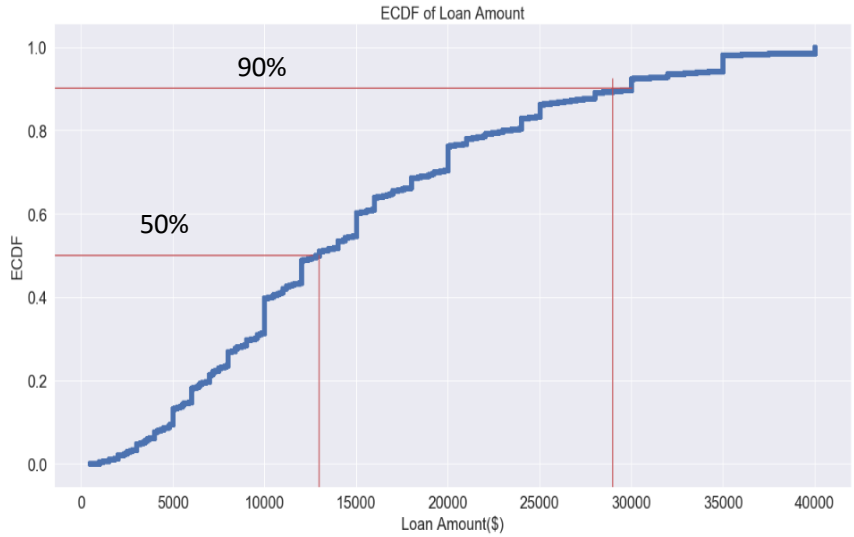
3.1. Loan Amount

A) What does the distribution of Loans look like?

- The distribution is skewed to the right
- The most common loan is $10,000

The next section will show the plot of the ECDF we can see further insights into loan distribution

B) What does ECDF of Loan Values look like?



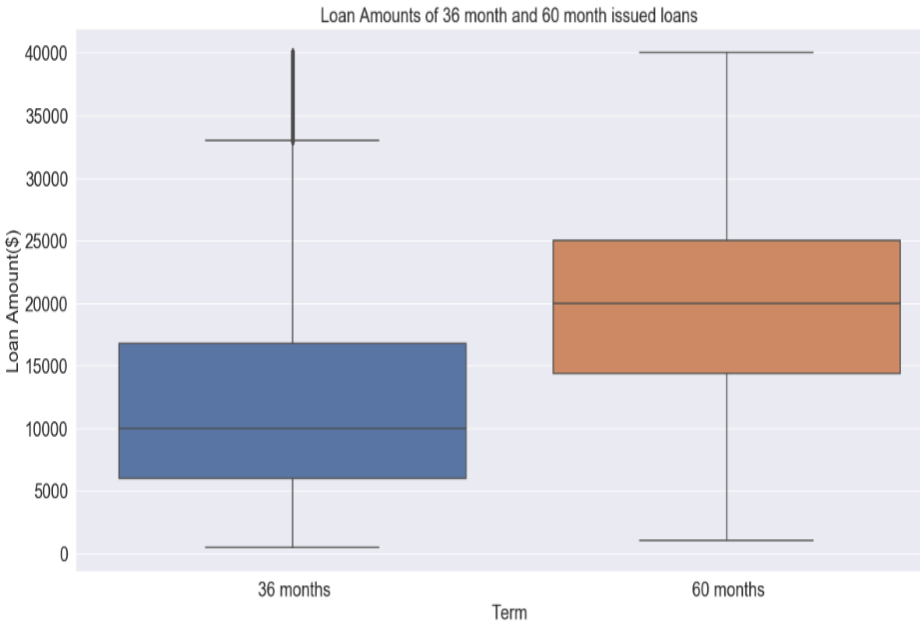
- 50% of Loans are less than $13,000
- 90% of loans are less than $29,000
- The largest loan is at $40,000 with the lowest loan at $500.
- There are spikes in loans at round values such as 10,000 ,20,000, 25,000. The biggest spike occurs at $10,000

3.2. Term of the Loans

A) How many loans are 36 months and how many are 60 months?

- 71% of issued loans are 36 months
- 29% of issued loans are 60 months

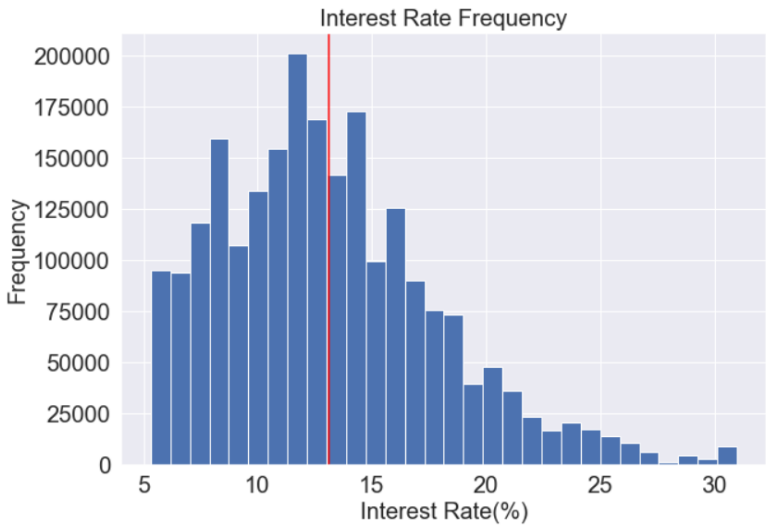
B) How do Loan amounts differ between the two terms?



- Mean for 60 months: $20,000
- Mean for 36 months: $10,000
- Both terms have an equal range of loan values
- 36 month IQR: $6,000 - $17,000
- 60 month IQR:$15,000 - $25,000
- 36 month loans has a few outliers outside the 1.5 IQR whisker.

3.3. Interest Rates

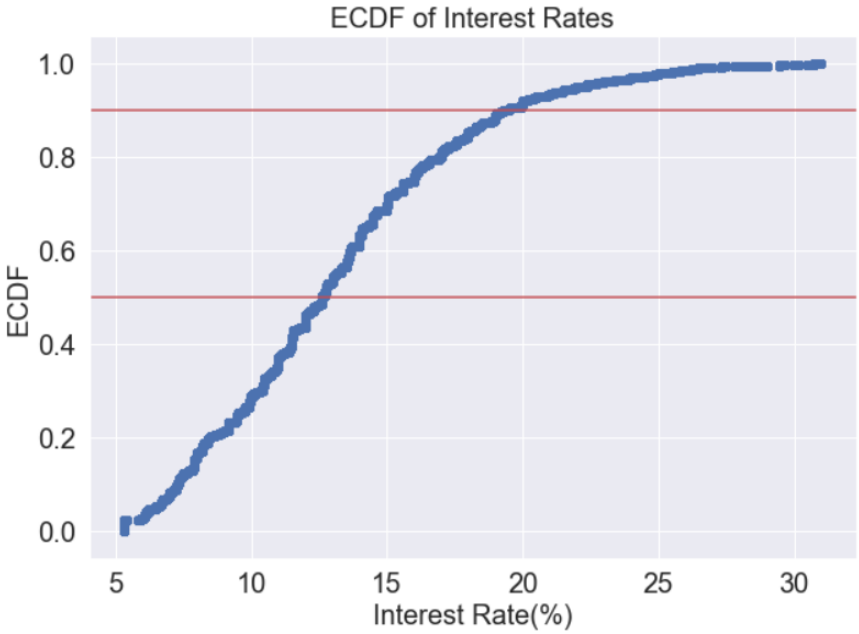
A) What does the distribution of interest rates look like?



- The most common interest is at 12-13%
- The interest rate distribution tails to the right, similar to loan amounts.

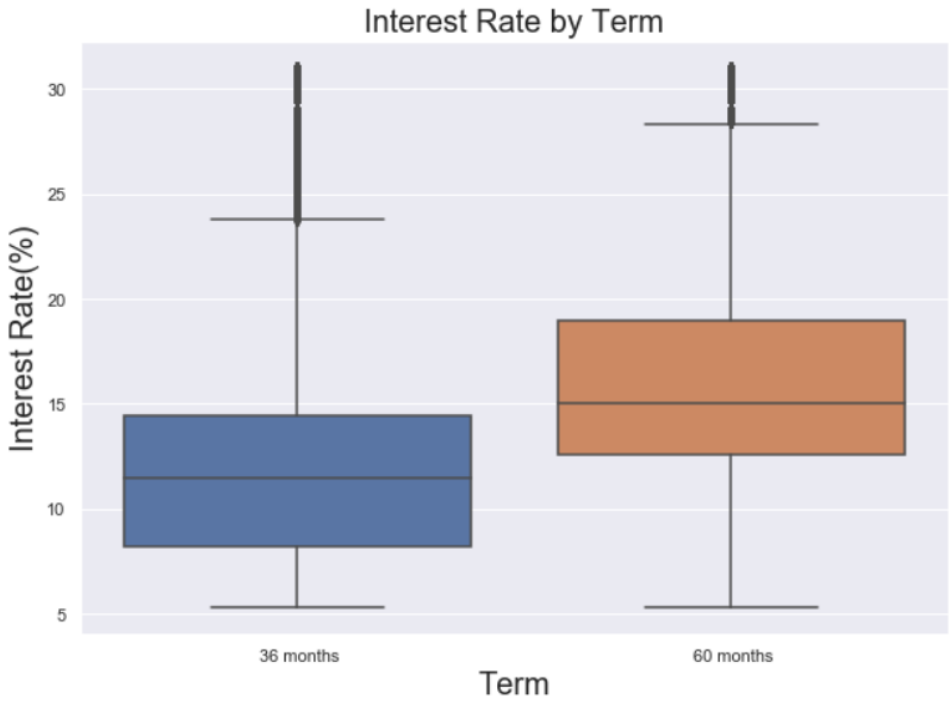
The next section will show the plot of the ECDF we can see further insights into interest rate distribution

B) What does ECDF of interest rates look like?



- 50% of Loans are issued are ≤13%
- 90% of Loans are issued are ≤20%
- The maximum interest rate is 33%
- The minimum interest rate occurs at 5%

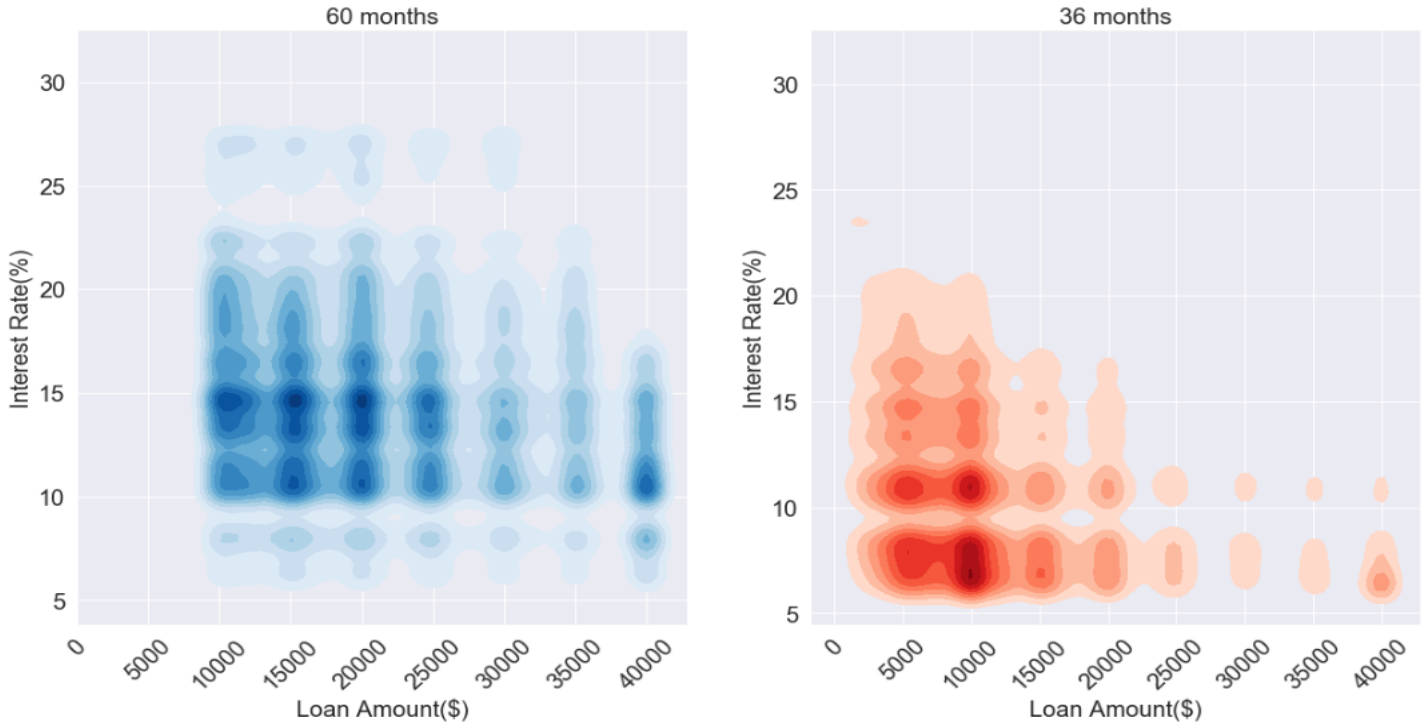
C) How does the term of the loan affect interest rate?



- **36 months**
  - The IQR is between 7 and 14%.
  - The whisker edges are between 5% and 23%
  - The range is between 5% and 33%
- **60 months**
  - The IQR is between 13% and 18%.
  - The whisker edges are between 5% and 27%
  - The range is between 5% and 33%

In order to see a more complete picture of the loans, the loan amount and loan terms will be integrated in the next question.

D) How do loan amounts, IR's and term change with respect to each other?

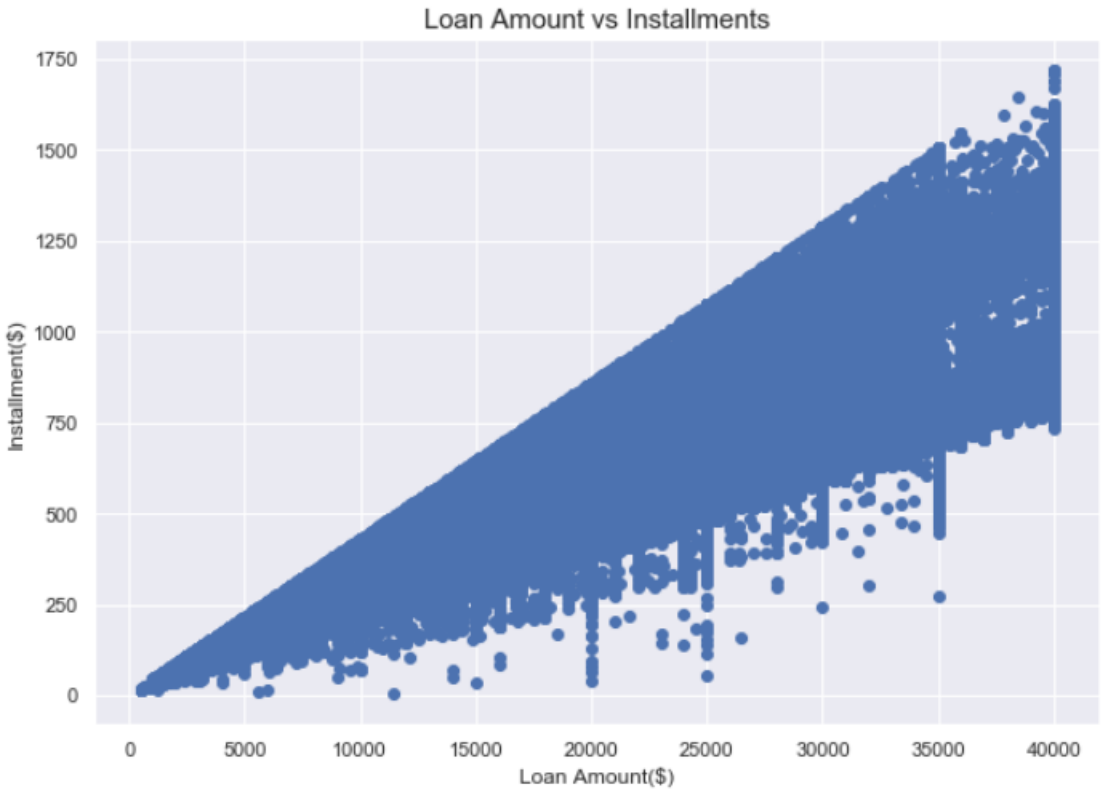


- **36 months**
  - The highest concentration of loans is between $8,000 & $12,000 with interest rates between 5% and 12%
  - Loans start to significantly decrease after $20,000.
  - The mean loan is ~$12,750 with a std of $8,550 and a mean IR of 12%.
- **60 months**
  - $9,000 seems to be where loans start getting issued at this term. Mostly at 10-15% Interest rate.
  - The higher the loan value, the lower the range of the interest rate.
  - The highest concentration of loans are between 10,000 & 20,000
  - The highest concentration of interest rates is between 10% and 17%.
  - The mean loan is ~$20,740 with a std of $8090 and a mean IR of 15.9%.

In general, the higher the loan the lower the interest rate. A $10,000 loan over 36 months tends to have a lower interest rate than a $10,000 loan over 60 months.

3.4. Installments

A) How are Installments and Loan values related?



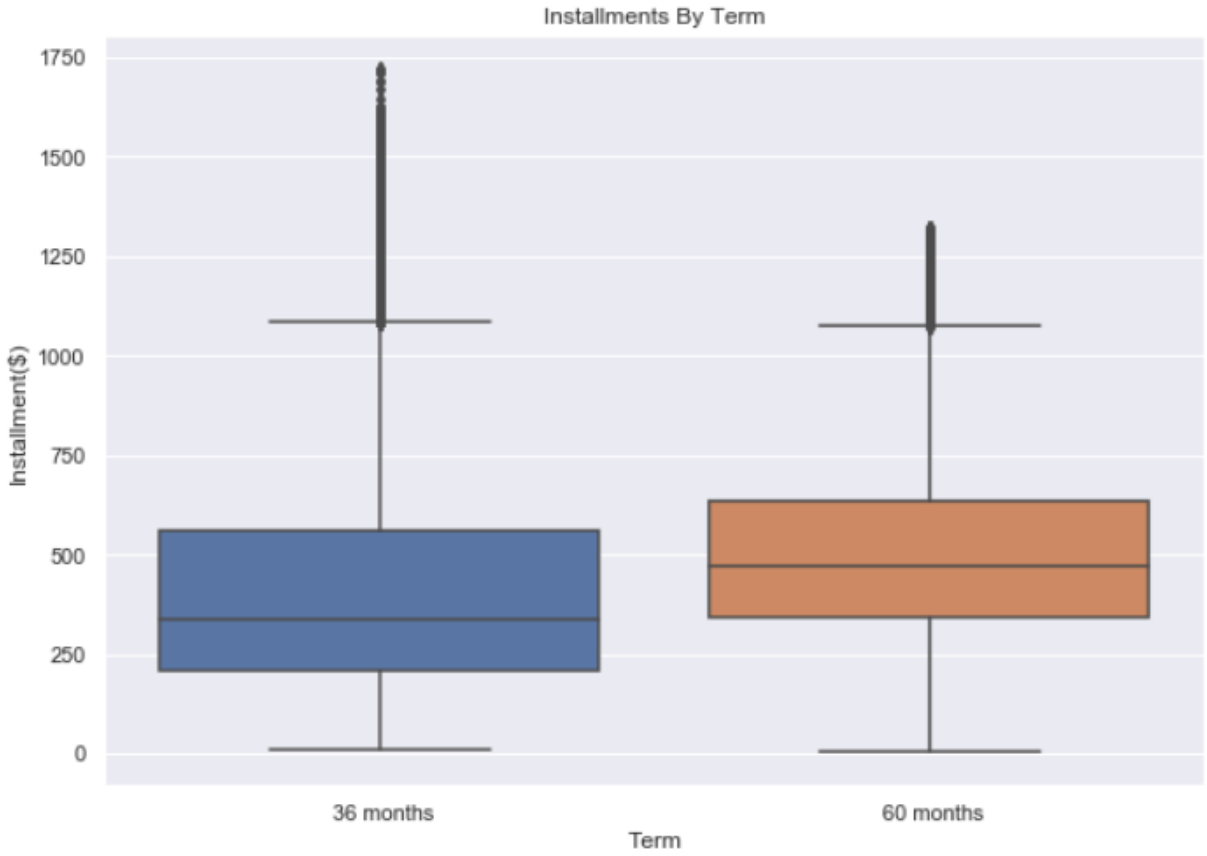
Observations:

- There is a linearly increasing ceiling for installments which indicates a direct relationship between installments and loan amounts.
- As the loan value increase the variability of the installment value increases but does it increase by the same proportion?

  Example:

- For a $10,000 loan, the maximum installment is around $500 and the minimum installment is $100. A $400 range.
- For a $40,000 loan, the maximum installment is around $1,750 and the minimum installment is $750. A $1,000 range.
- The loan value increased by 4x but the range only increased by 2.5x.

B) How do Installments differ by Term?

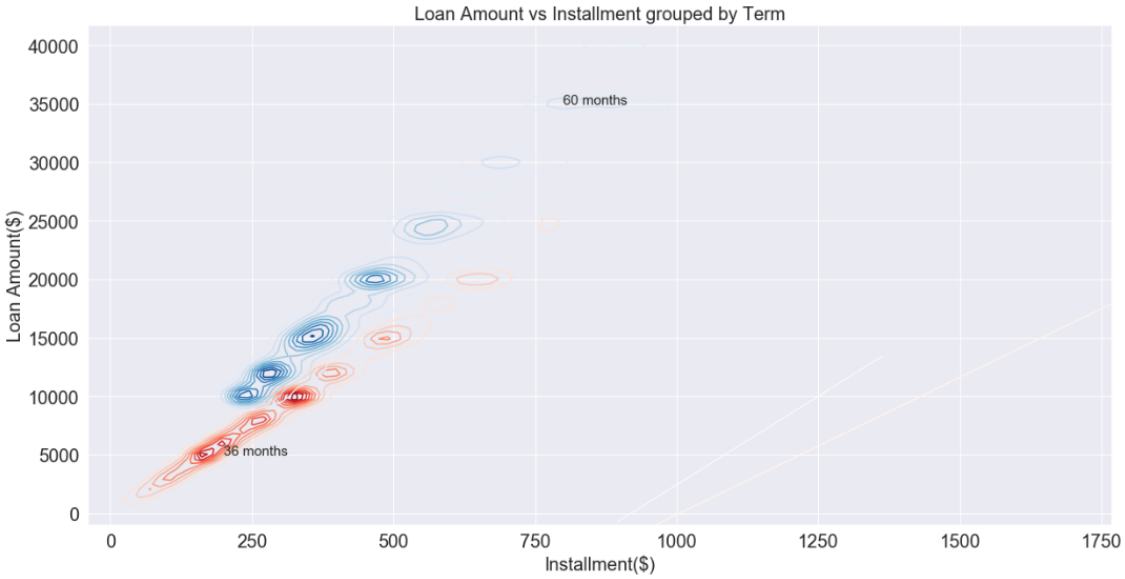


- 36 months has a range of installments between 0 and $1750
- 60 months has a range of installments between 0 and $1300
- 36 months has an IQR between $230 and $550.
- 60 months has an IQR between $300 and $650.

36 months has a wider range of installments than 60 months but the on average 60 months pay more every installment. Let's integrate loan amount in the above graph.

In order to see a more complete picture of the loans, installment and term will be integrated in the next question.

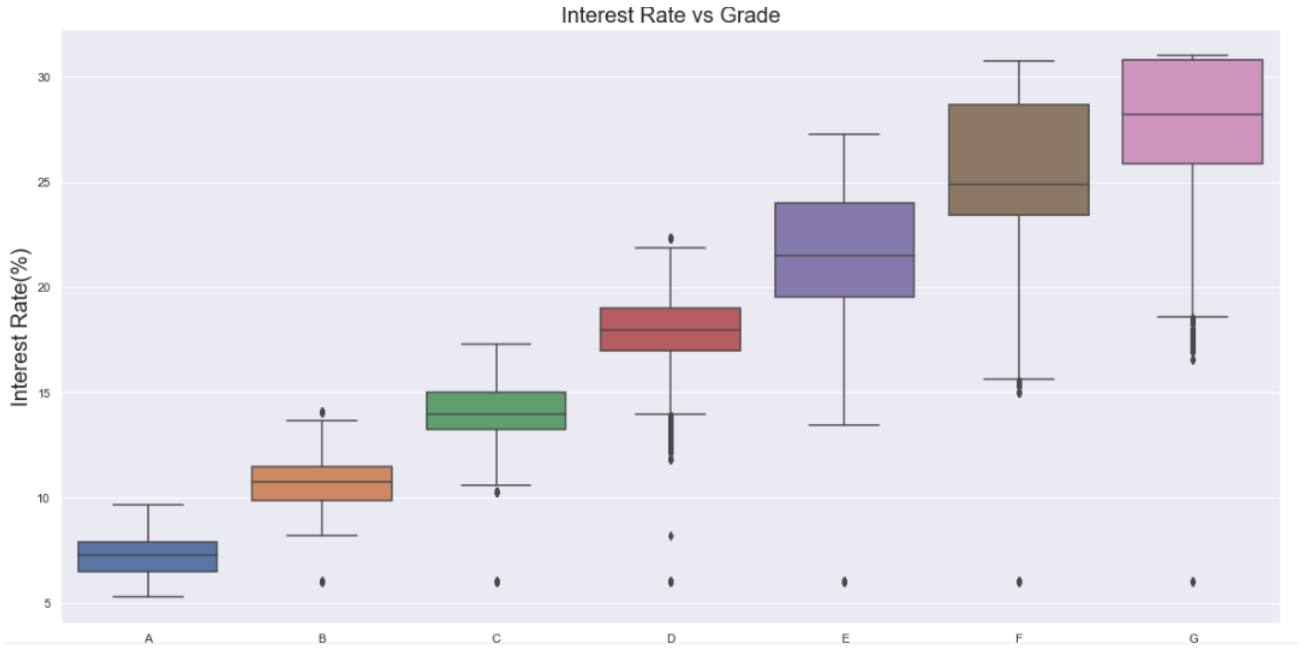
C) How do Installments differ by Term and loan amount?



For the same loan amount, 36 months have higher installments than 60 months as can be seen above.

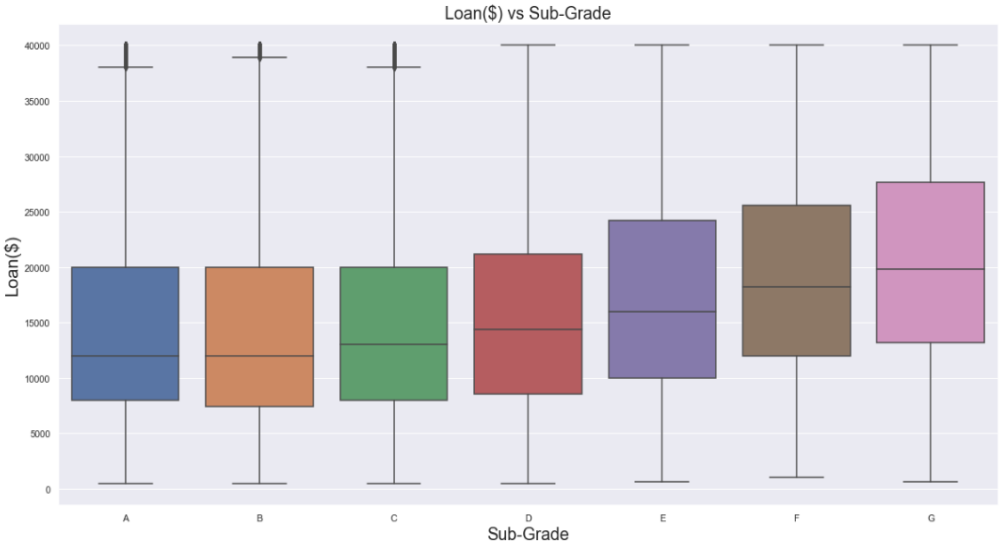
3.5. Grade and Subgrade

A) How does interest rate change with grade of the loans?



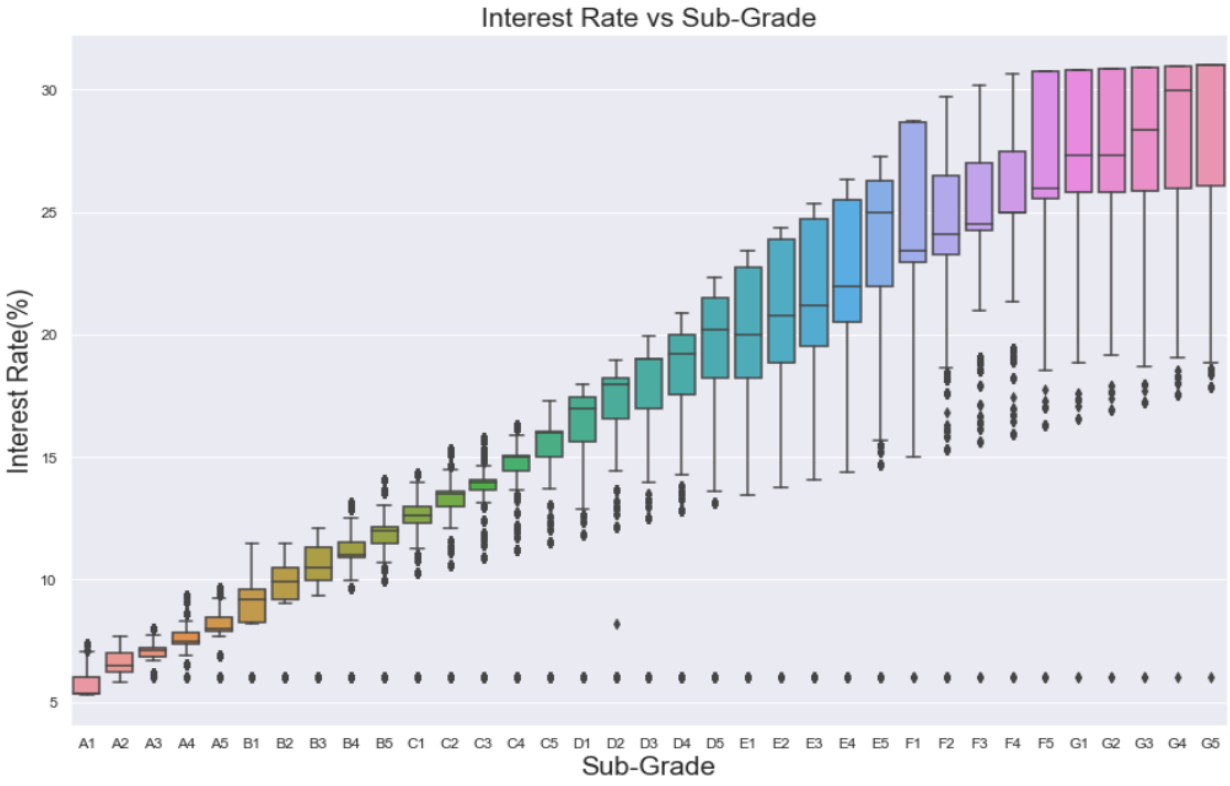
- Interest rate is highly dependant on the grade assigned.
- There exists a few outliers for each grade at very low interest rates.

- **How do loans change with grade of the loans?**



Grade and loan amount start having a correlation starting from grade D.

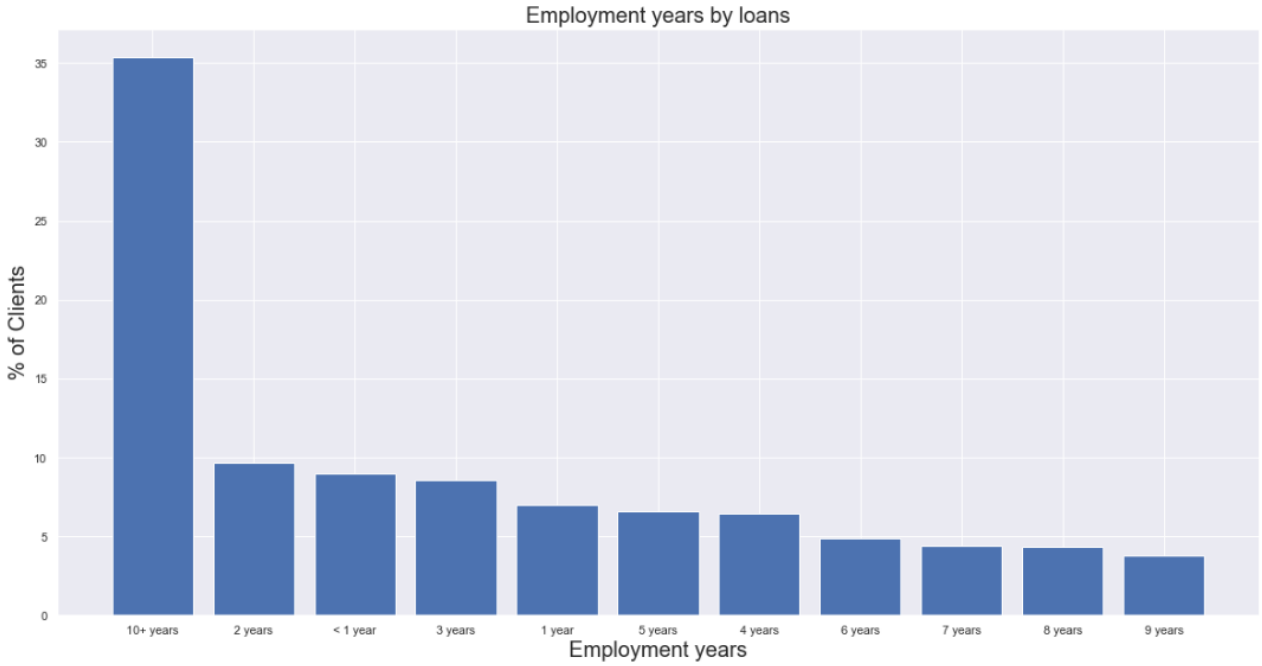
C) How does interest rate change with subgrade?



As subgrade deteriorates, not only does the interest rate increase, but the range increases as well. Example: A B4 loan has a range between 10-13%. An E4 loan has a range between 14-26%. This shows a huge increase in range and variability.

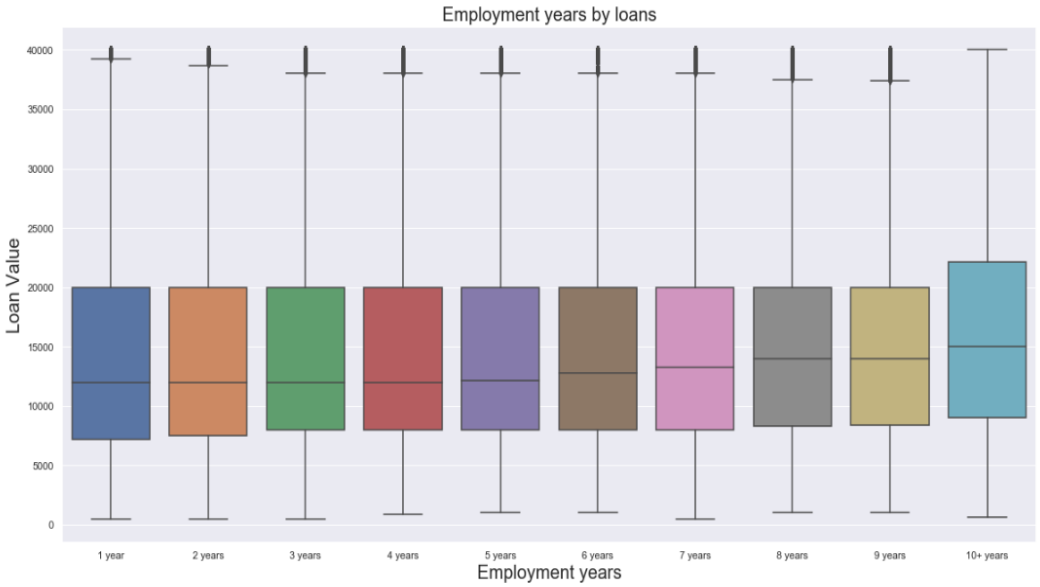
3.6. Employment Information

A) How long have the loan holders been in employment?



- 35% of Loanees have been working for 10+ years
- 65% have been working less than 10 years as seen in the above graph.

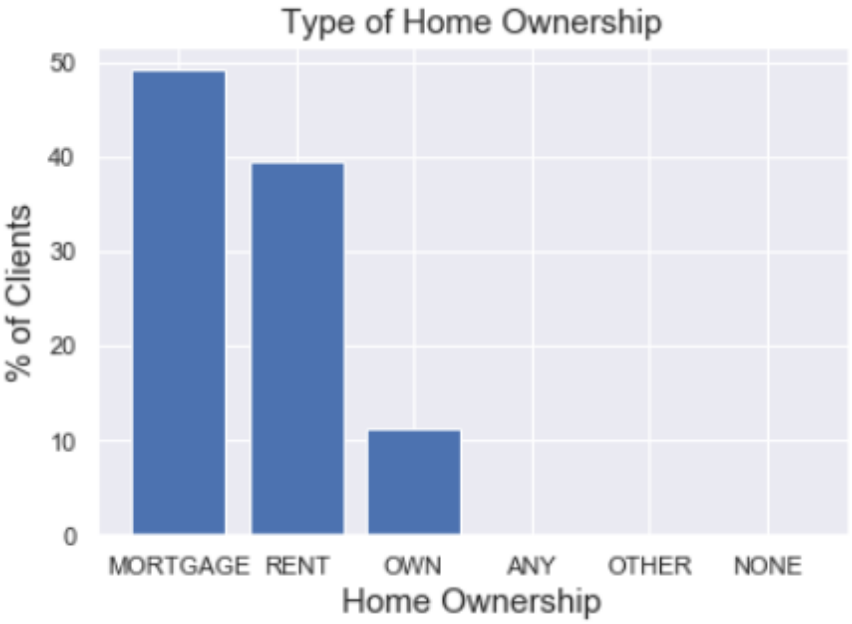
B) Do loan values change based on employment years?



- The IQR for those employed for 10+ years is between $9,000 and $23,000.
- The IQR for those working for less than 10 years is between USD8,000 and USD 20,000.

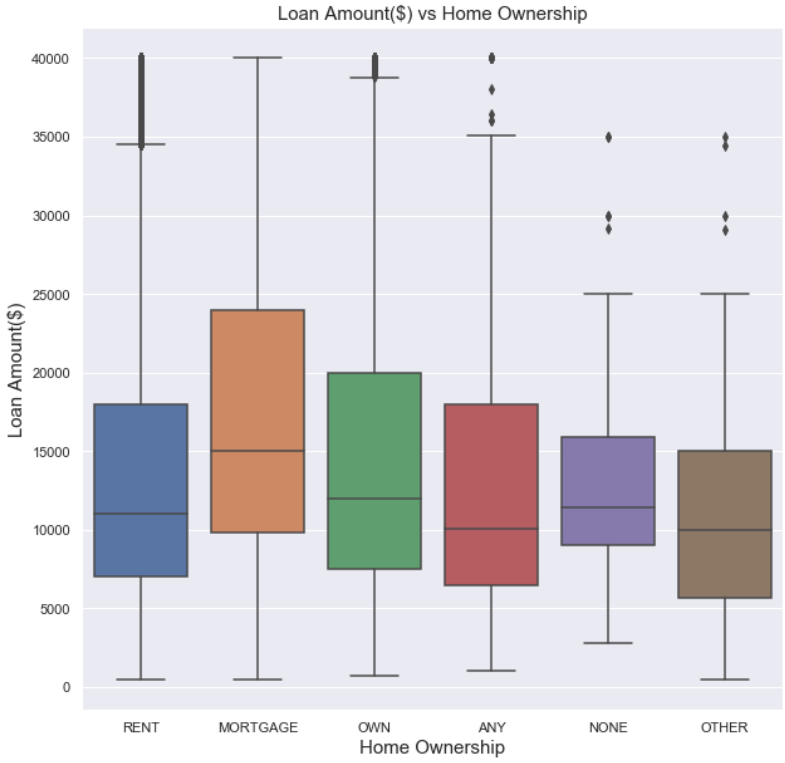
3.7. Home Ownership

A) What kind of home ownership do clients have?



- 50% of clients have a mortgage
- 40% of clients rent a property.
- 9% already own a property.
- 1% are unidentified.

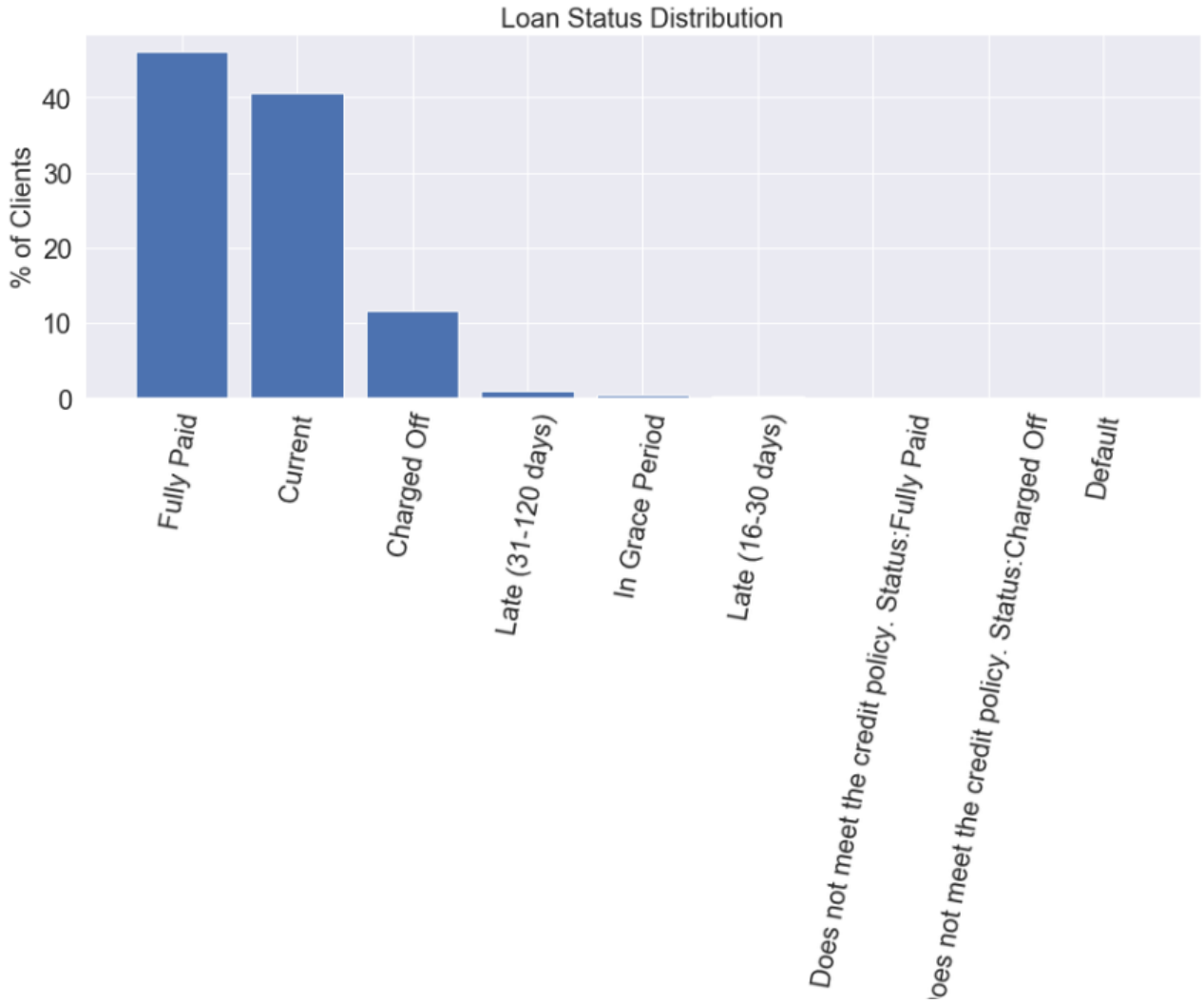
B) How does loan amount change with homeownership?



- Mortgage holders tend to hold higher valued loans and have the widest variability with a mean of $15,000.
- Owners seem to have a similar mean to renters.
- Owners have higher variability than renters.

3.8. Loan Status Analysis

A) How are loan statuses distributed?

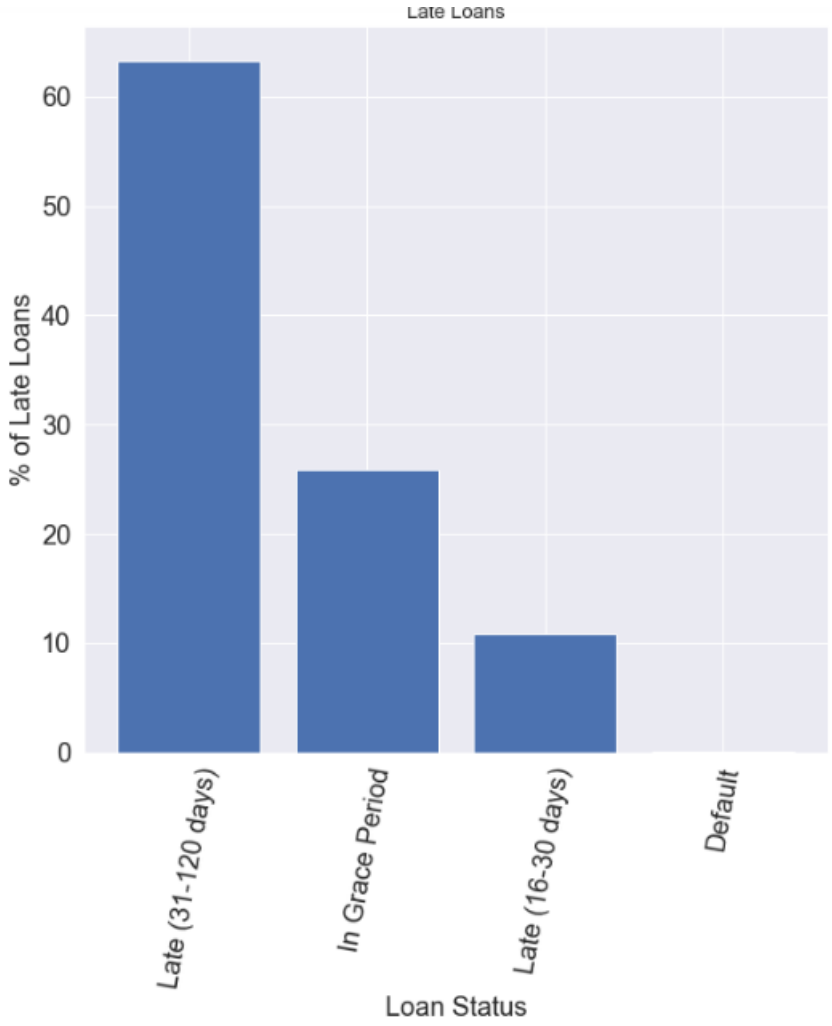


- 80-85% of loans are in good shape (Fully Paid or Current)
- 10-12% of loans are charged off
- Cannot make an accurate estimation on prediction of bad loans.

B) How are bad loan statuses distributed?

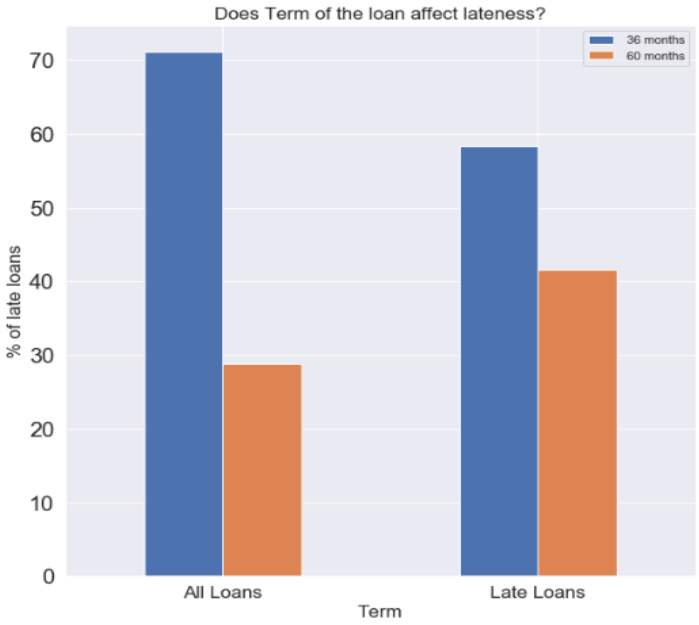
Bad Loans were calculated to be 1.53% of total loans and are classified as the following:

1. ***In Grace Period***
2. ***Late (16-30 days)***
3. ***Late (31-120 days)***
4. ***Default (120+ days)***



- 63-65% of late loans are between 31-120 days late.
- 20-25% of late loans are in grace period (1-16 days)
- 10-12% of late loans are 16-30days late.
- Default loans are negligible.

C) Has term distribution changed between Good loans and late loans?



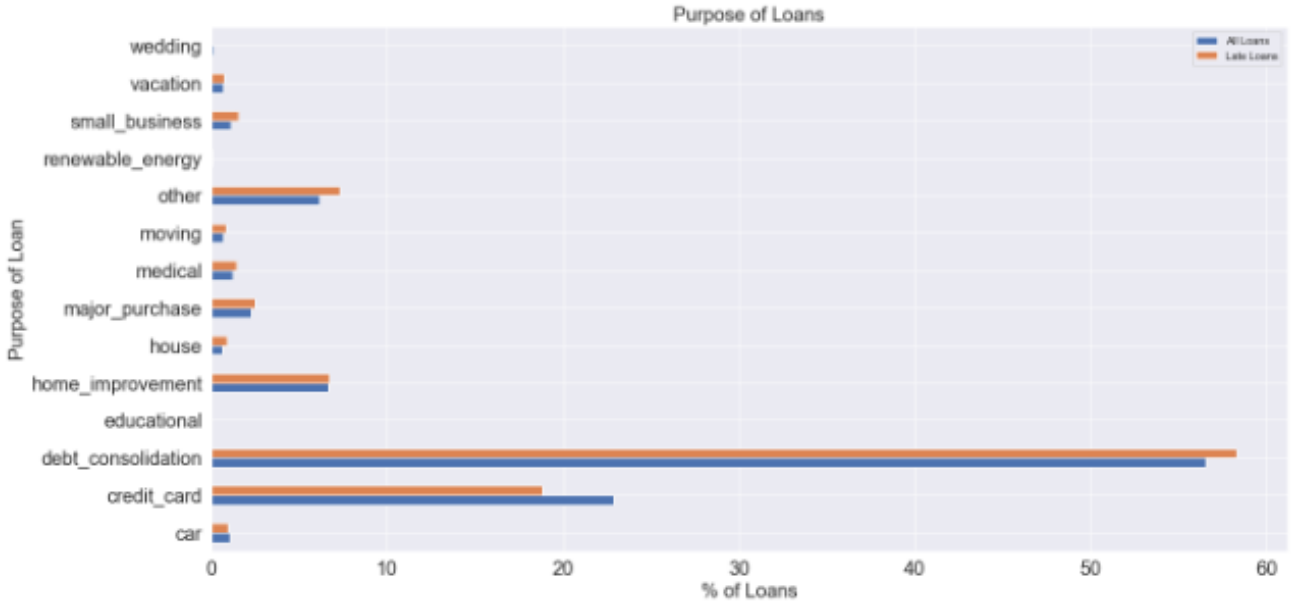
- 60 months loans made up 28% of good loans but made up 42% of late loans.
- 36 months loans made up 72% of good loans but made up 57% of late loans.
- The term of the loan does indeed affect the lateness of the loan.

The following was observed from the given data:

	Average Loan Amount	***Average Interest Rate(%)***
Good Loans	***$14,940***	***12.7%***
Late Loans	***$15,700***	***15.7%***

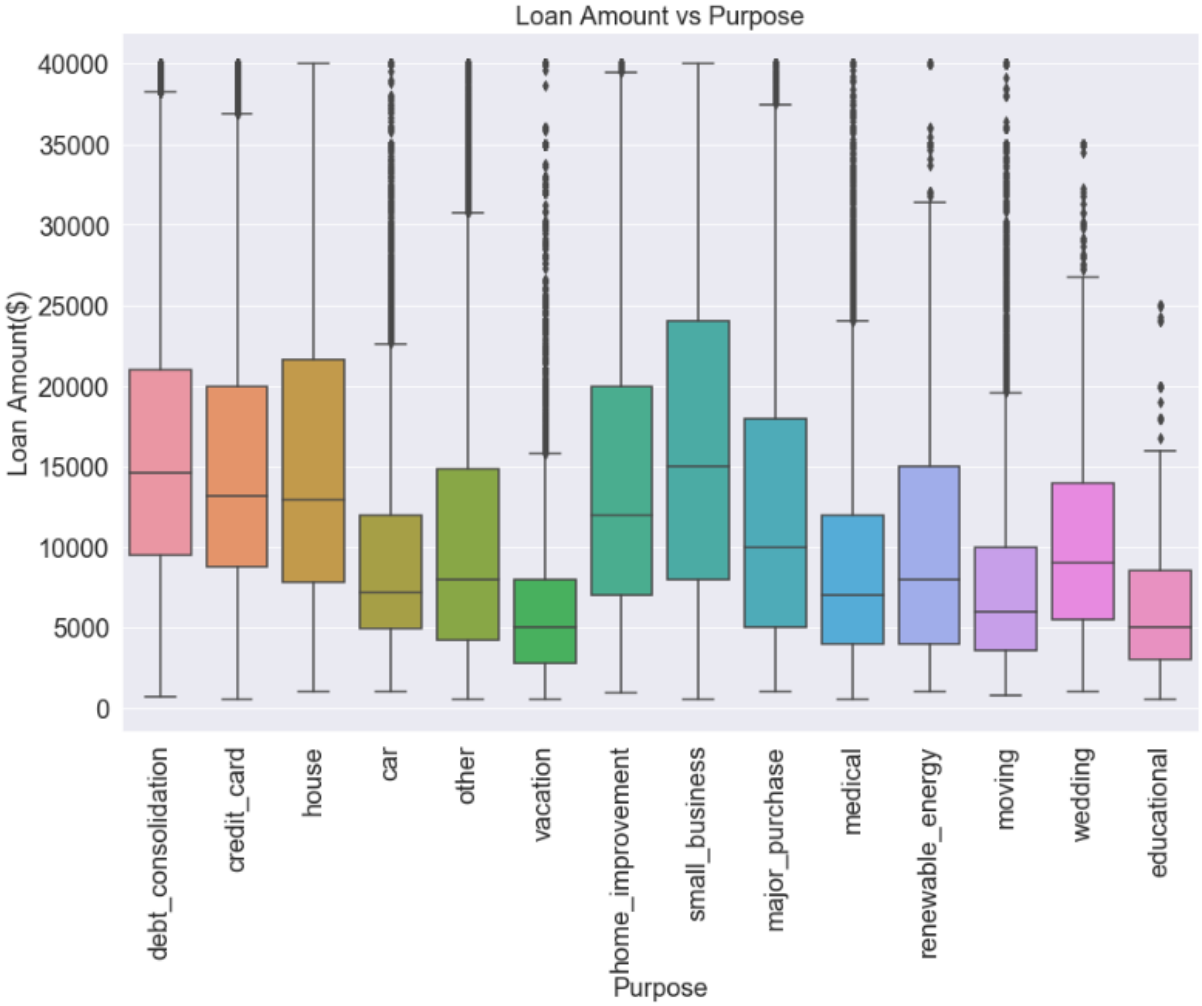
3.9. Loan Purpose Analysis

A) What purposes are used for the loans?



- 55-60% of loans are for debt consolidation.
- 17-23% of loans are used to pay Credit card
- 7-8% of loans are used for home improvement purposes
- There is a decrease in loans for Credit card from 23% for all loans to 17% late loans.

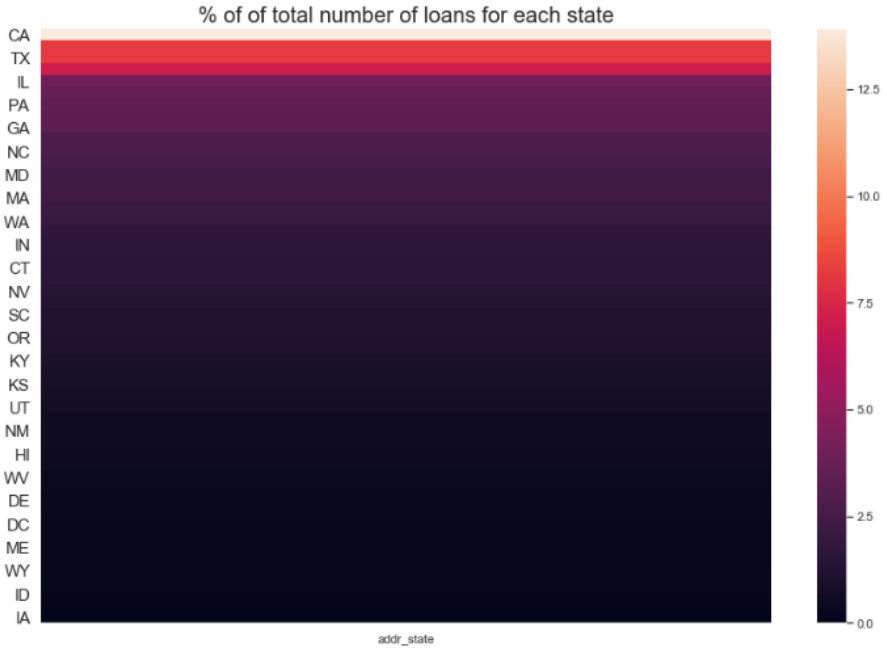
B) How does the purpose change the loan amount?



- Loan amounts is highest for small businesses.
- Loan amounts is lowest for vacations.

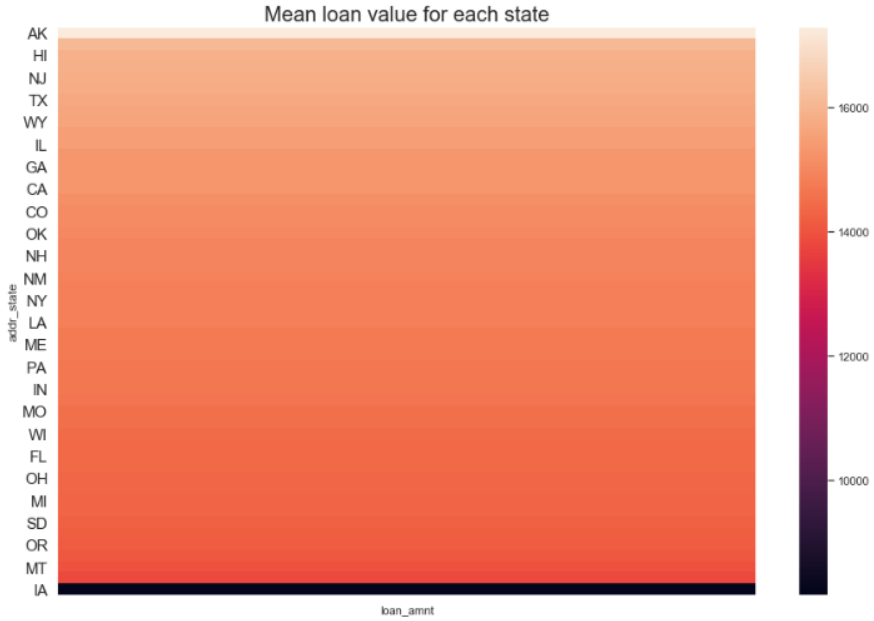
3.10. Location Analysis

A) How do number of loans vary by state?



- California has the most amount of issued loans(13%).
- Iowa has the lowest amount of issued loans at close to 0%

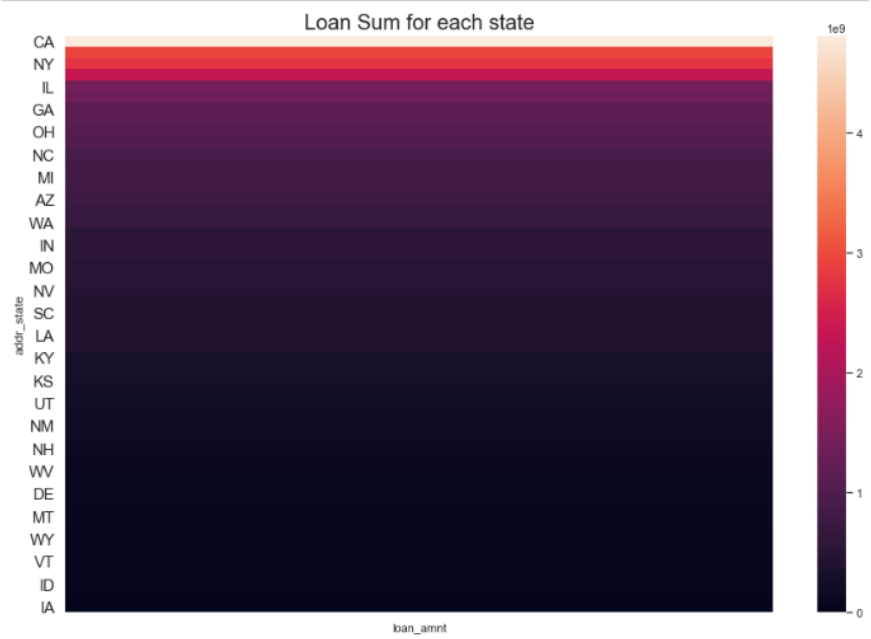
B) How does the mean loan value vary by state?



- California has the 8th highest mean loan value.
- Alaska has the highest mean loan

In order to combine the effect of loan value and number of loans, what does the sum of all loans for each state look like?

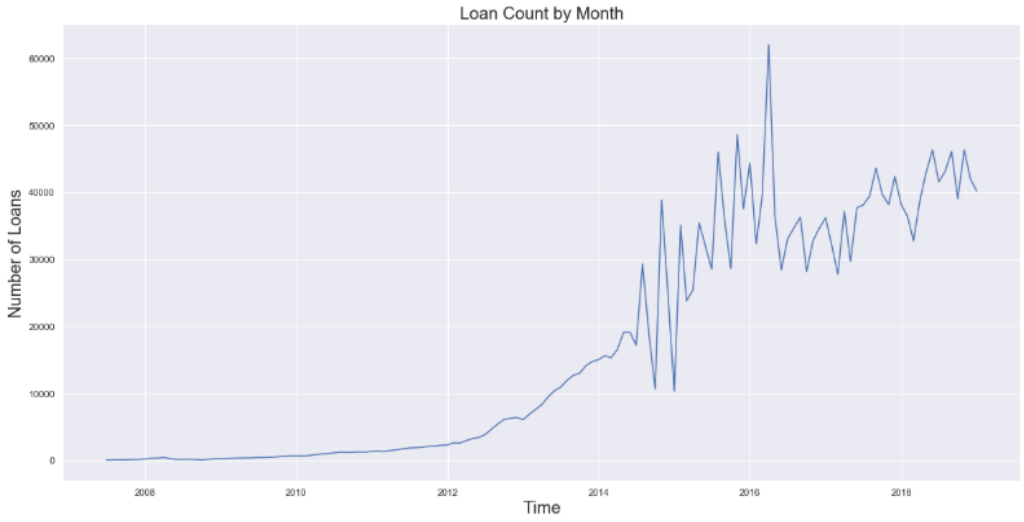
C) How does the sum of all loans vary by state?



- Like the number of loans with CA state as the #1.
- The order is mildly different.

3.11. How have features changed with time

A) How does Loan count change across the years?

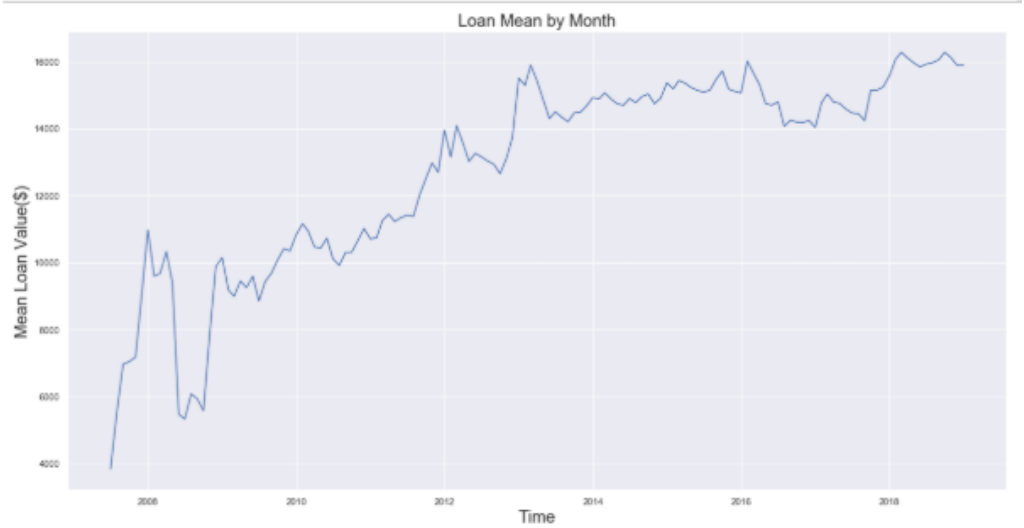


Observations

- It took 6-7 years(2007-2013) for the number of loans to go from 0 to 10,000. The increase was steady with no fluctuations month to month.
- It took 2-3 years (2013-2015) for the number of loans to go from 10,000 to 20,000.The increase was steady with no fluctuations month to month.
- It took 1-2 year(2015-2016) year to go from 20,000 to 60,000 loans. But there was severe fluctuations month to month.
- There was a peak of loans in 2016 but then decreased rapidly to 30,000 issued loans.

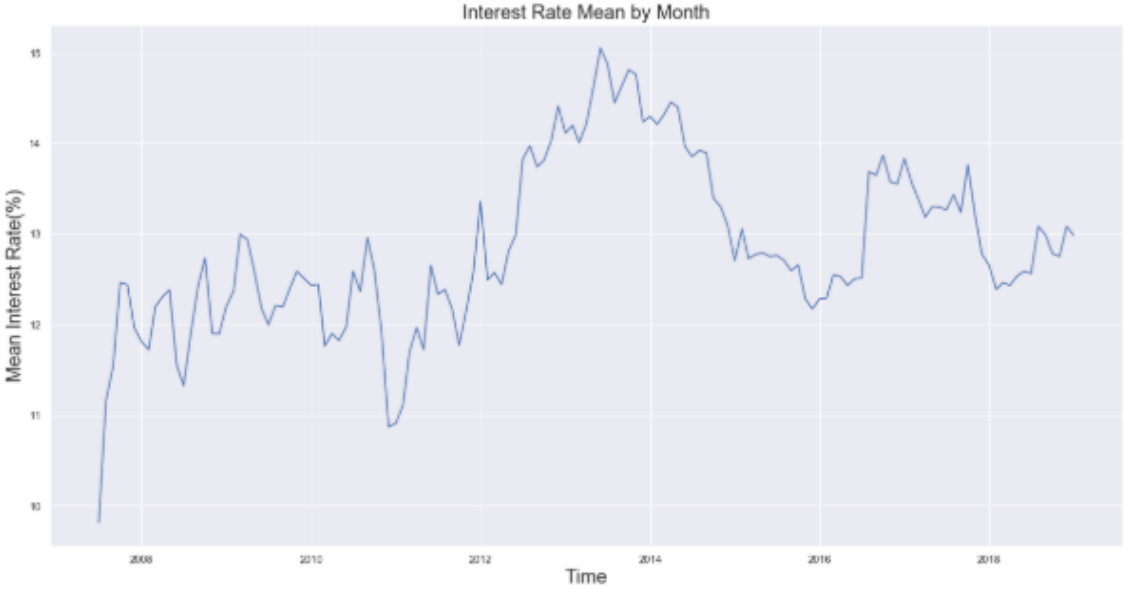
B) How does the mean loan change across the years?

.



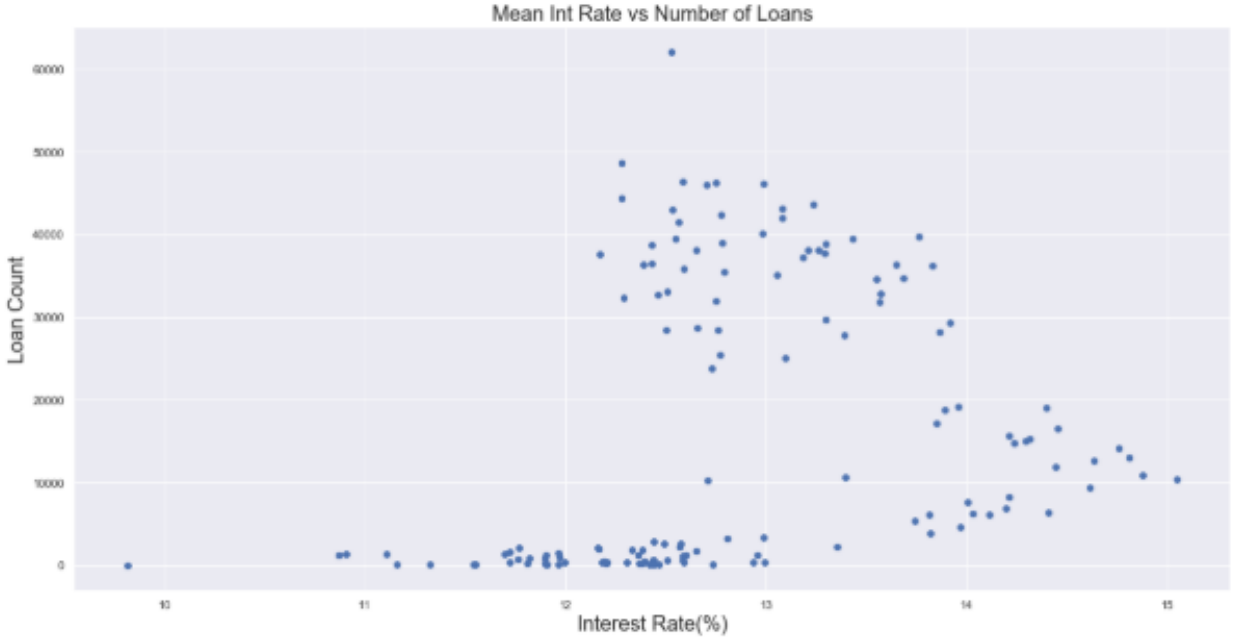
The mean loan value increased from 2007 to 2012, flattening at around $15-16,000 per loan.

C) How does the interest rate change across the years?



- Interest rates were steady between 2008 and 2012 ranging between 11-13%.
- Interest rates increased from 2012 to 2014 reaching a peak of 15%.
- They then decreased to 12% reaching a low in 2016.

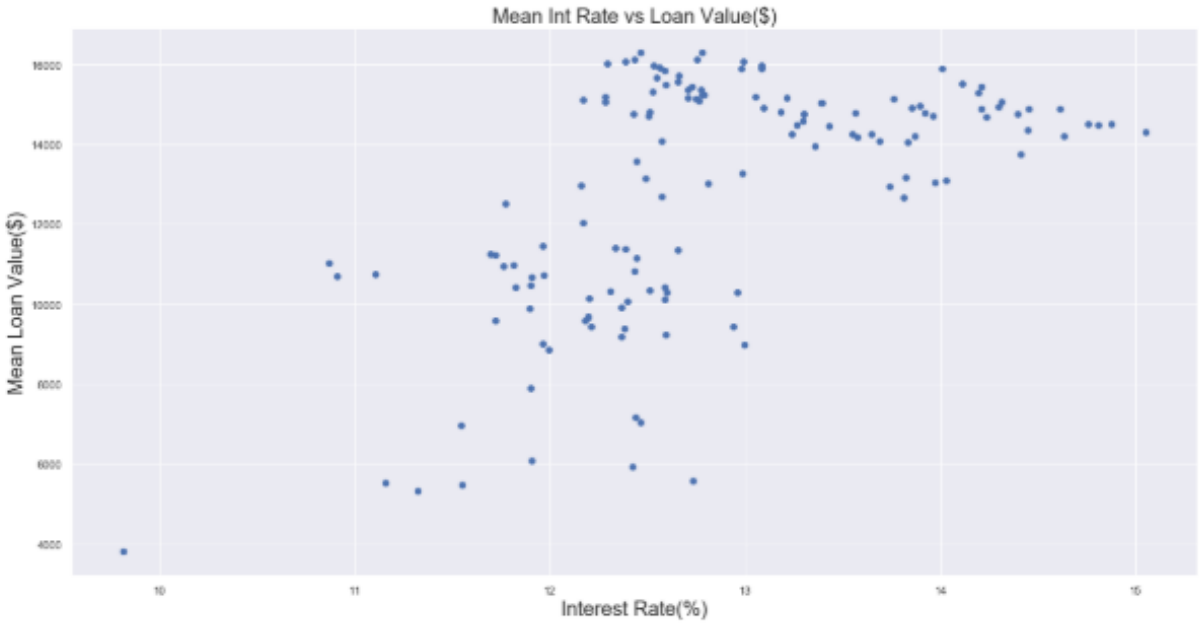
D) How does Loan count change with interest rate?



There are two observations evident.

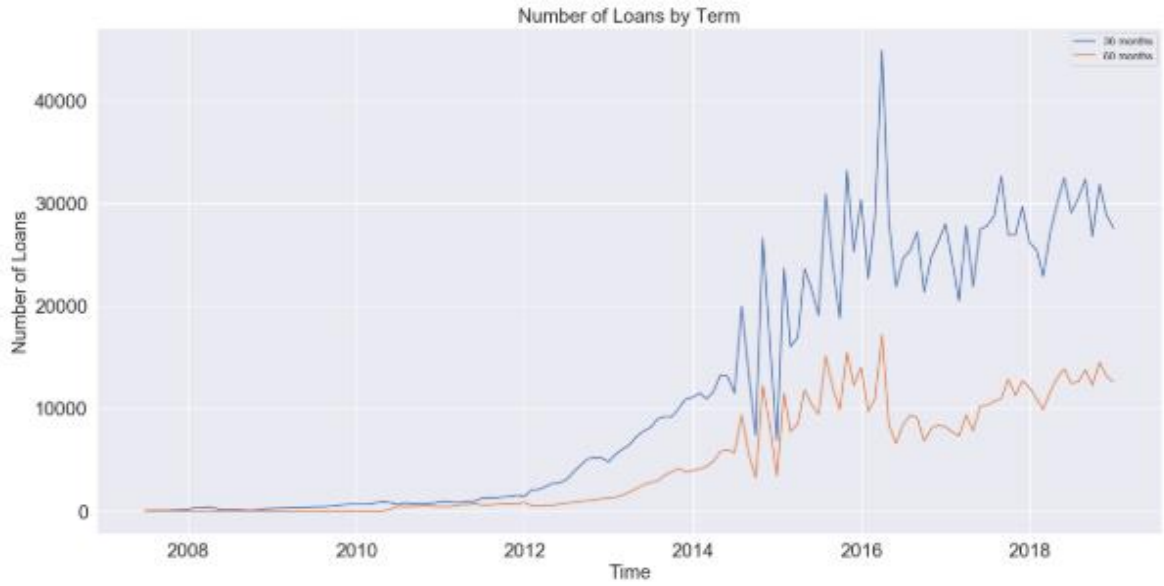
- 1st Trend: We see negative relationship from 12% to 15% for most of the months.
- 2nd Trend: We see a flat relationship between 9-13% for a very low number of loans.

E) How does mean interest rate change with Loan Value?



There is a strong positive relationship between loan value and interest rate for loans between 4,000 and 16,000.

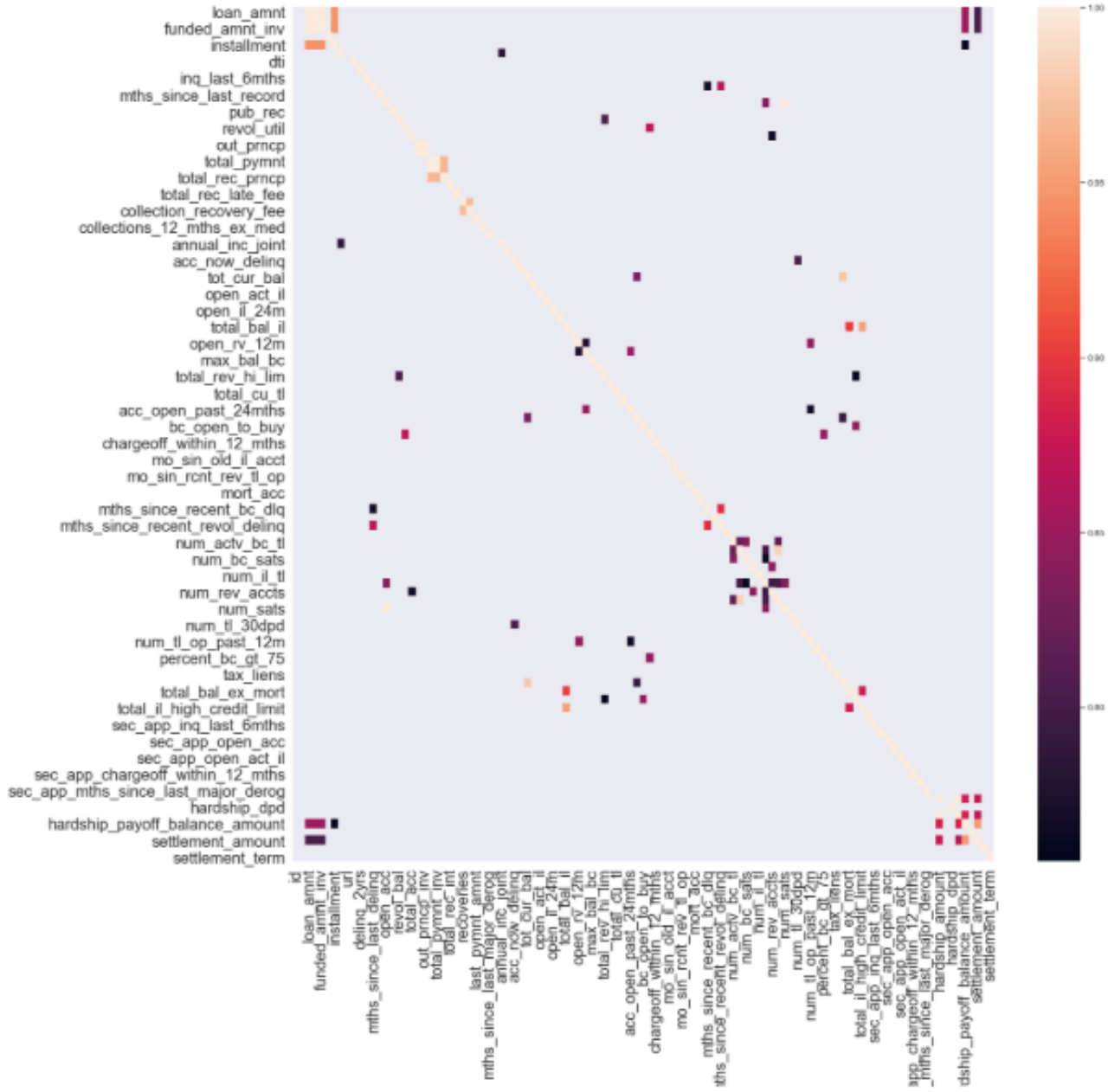
F) How has the term length change across the years?



The pattern of change for both terms of loans are similar but there is an evident difference of distribution as examined previously.

3.12. Are there significant correlations between the numeric variables in the data frame?

The aim of this question is to visualize if there are significant correlations with Pearson correlation coefficients with magnitudes greater than 0.7.



There does exist multiple features with strong correlations that could be helpful in predicting whether a loan will be bad.

4. Inferential Statistical Analysis

Four questions were asked that can aid comprehending the data some more. These questions were translated into hypothesis tests that were all done by comparing distributions.

This was done using the following steps:

1. Stating the null hypothesis
2. Combine the two datasets together
3. Randomly scramble the data so that the original order is lost.
4. Divide the scramble data into 2 portions. Each portion would be the same length as the original dataset. Except now they are from the scrambled set. These are called permutated samples.
5. Compute a permutated replicate of choice. (In this scenario the replicate will be the difference of means.
6. Compute 2,000 replicates.
7. Run the test to compute the p-value.(The probability of observing a test statistic equally or more extreme than the one observed, given that the null hypothesis is true)

4.1. Do mortgage holders have a different interest rate distribution as those who own their homes?

Null Hypothesis: Mortgage Holders and Owners have the same interest rate distribution

Observed Difference of Means	**0.38%**
P-value	**0**
Hypothesis Test	**Reject the Null Hypothesis**
Conclusion	**Mortgage Holders and Owners do not have the same interest rate distribution**

4.2. Do mortgage holders have different distributions of loans vs those who rent their homes?

Null Hypothesis: Mortgage Holders and Renters have the same loan value distribution.

Observed Difference of Means	**$3495**
P-value	**0**
Hypothesis Test	**Reject the Null Hypothesis**
Conclusion	**Mortgage Holders and Renters do not have the same loan value distribution**

4.3. Is interest rate distribution the same for all loan Status (Bad vs Good)

Null Hypothesis: Bad and Good Loans have the same interest rate distribution.

Observed Difference of Means	**3%**
P-value	**0**
Hypothesis Test	**Reject the Null Hypothesis**
Conclusion	**Bad and Good Loans do not have the same interest rate distribution**

4.4. Is loan amount distribution the same for all loan Status (Bad vs Good)?

Null Hypothesis: Bad and good Loans have the same loan amount distribution.

Observed Difference of Means	**$749**
P-value	**0**
Hypothesis Test	**Reject the Null Hypothesis**
Conclusion	**Bad and Good Loans do not have the same loan amount distribution**

5. Modelling

5.1. Models used

- Logistic Regression
- Random Forest
- XG Boost Classifier

5.2. Three Sets of Data:

- Training Set – Fit models
- Validation Set – Compare results from models
- Testing Set – Model performance based on selected model

5.3. Success Metrics

Recall	Precision
$\frac{\#\ of\ True\ Positives}{True\ Positive + False\ Negatives}$	$\frac{\#\ of\ True\ Positives}{True\ Positive + False\ Positive}$
$\frac{\#\ of\ Predicted\ Bad\ Loans\ \&\ actually\ bad}{\#\ of\ actual\ bad\ loans}$	$\frac{\#\ of\ Predicted\ Bad\ Loans\ \&\ actually\ bad}{\#\ of\ predicted\ bad\ loans}$
A recall of 1 = 100% bad loans were flagged	**A precision of 1 = No misclassification of bad loans**

- A precision closest to 1 is important
    - Misclassification can lead to flagging customers that are okay. This may be inconvenient to current customers.
- A recall greater than 0.8 is acceptable
    - This means 80% of bad loans will be flagged.

5.4. Finding the best model:

5.4.1. Logistic Regression

- Parameters to tune:
  - C – Inverse Regularization Parameter
  - Penalty – L1 or L2
- Best Model:
  - C = 1
  - Penalty = L2
- Results:
  - Precision = 0.99
  - Recall = 0.88

5.4.2. Random Forest

- Parameters to tune:

Parameter	Description	Best Params
max_depth	Depth of each tree in forest	29
max_features	Maximum features allowed to try in one tree	Sqrt
n_estimators	number of trees in the forest	56

- Results:
  - Precision = 0.99
  - Recall = 0.78

5.4.3. Random Forest

- Parameters to tune:

Parameter	Description	Best Parameters
max_depth	Depth of each tree in forest	61
max_features	Maximum features allowed to try in one tree	22
n_estimators	number of trees in the forest	771
Learning_rate	Loss function optimizer	0.019
Colsample_by_node	subsample ratio of columns for each node	0.4

- Results:
  - Precision = 0.99
  - Recall = 0.86

5.5. Winner and Test Set

Logistic Regression is the winner with a recall score of 88%.

Results on test set:

- Recall: 88%
- Precision: 99%

6. Steps Forward

1. Use a larger RAM to test on more data as the RAM used to test this data is only 8GB and could not fit the entire dataset.
2. Once a stronger RAM is used, more time is available, and more models can be tested such as SVM.
3. Once confidence is high regarding the model, the model can be slowly used with a subset of consumers in real time.
4. Results from this implementation can be used to fine tune the model such as discover more features or remove features to be able to more accurately predict.