MISM 6210
Information Visuals and Dashboards for Business
Team Outliers

Research Problem

Lending Club is a peer-to-peer Lending company based in the US. Investors invest money through Lending Club, and this money is passed onto the borrowers. When borrowers pay their loans back, the capital plus the interest passes back to the investors. It is a win for everybody as they can get typically lower loan rates and higher investor returns. We chose to study Lending Club business model as this could help Lending Club improve its risk management and reduce its losses due to defaults. The model could then be applied to new loan applications to predict the likelihood of defaulting.

Our business problem is to determine how loans are distributed and to build a model to predict bad loans.

It is important to understand why loans default and how to avoid it. Our business solution is for the Loan Manager at Lending Club, and the analysis would help her/him to understand how to distribute loans better. For example, higher interest rate loan at lower annual income increases the chances of default. There are existing work in this business problem, which have been referenced in the Appendix.

Data and methodology

The source of the dataset is Lending Club. The data collected is from May 2012 to Feb 2013. The data contains 50,000 rows under 56 columns such as loan amount, interest rates, term, recovered principal, home ownership status, loan status, purpose, verification status etc. It contains various attributes which are important to predict the likelihood of defaulting.

This dataset is useful in identifying trends or patterns in loan amount, interest rate, and borrower characteristics (such as employment status, income, and home ownership).
Additionally, the inclusion of loan is bad helps to identify which loans ultimately resulted in defaults or other issues, allowing for a more detailed analysis of the factors that contribute to bad loans.

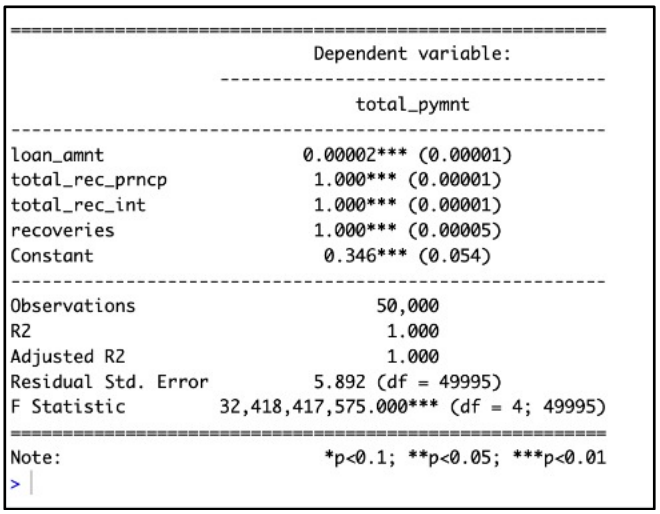
We have used the following methodology in R for cleaning of the dataset:

1. Removal of duplicate values: Duplicate values were removed from the dataset using the "unique" function to ensure that each observation was unique.
2. Removal of null values: Null values were removed from the dataset using the "complete.cases" function.
3. Checking the data structure: The "str" function was used to check the datatypes of the columns in the dataset.

Result from regression model:

1. We have used linear regression using 'Total Payment' as Y variable and 'Loan Amount', 'Total Recovered Principal', 'Total Recovered Interest' and 'Recoveries' as X variables.

The output of linear regression is as follows:



	Dependent variable:
	total_pymnt
loan_amnt	0.00002*** (0.00001)
total_rec_prncp	1.000*** (0.00001)
total_rec_int	1.000*** (0.00001)
recoveries	1.000*** (0.00005)
Constant	0.346*** (0.054)
Observations	50,000
R2	1.000
Adjusted R2	1.000
Residual Std. Error	5.892 (df = 49995)
F Statistic	32,418,417,575.000*** (df = 4; 49995)
Note:	*p<0.1; **p<0.05; ***p<0.01

From regression, it is analysed that all the mentioned variables are significant with respect to 'Total Payment'.

2. We have used logistic regression on ‘Loan is Bad’ variable as Y variable and ‘Loan Amount’, ‘Term’, ‘Interest Rate’, ‘Grade’, ‘Home Ownership’, ‘Annual Income’, ‘Verification Status’, ‘Purpose’, ‘DTI’, ‘Delinquency in Last 2 years’, ‘Total Payment’, ‘Total Recovered Interest’ and ‘Recoveries’ as X variables.

The output of logistic regression is as follows:

	Dependent variable:
	loan_is_bad
loan_amnt	0.001*** (0.00002)
term	-0.257*** (0.009)
int_rate	0.146*** (0.031)
gradeB	0.100 (0.178)
gradeC	0.017 (0.267)
gradeD	0.326 (0.350)
gradeE	0.709 (0.439)
gradeF	0.845 (0.521)
gradeG	2.615*** (0.654)
home_ownershipNONE	-2.688 (2.190)
home_ownershipOTHER	0.289 (1.051)
home_ownershipOWN	0.013 (0.119)
home_ownershipRENT	0.113 (0.070)
annual_inc	-0.00000*** (0.00000)
verification_statusSource Verified	-0.031 (0.093)
verification_statusVerified	0.122 (0.080)
purposecredit_card	0.099 (0.300)
purposedebt_consolidation	0.212 (0.295)
purposehome_improvement	0.417 (0.318)
purposehouse	0.539 (0.475)
purposemajor_purchase	-0.275 (0.369)
purposemedical	-0.307 (0.410)
purposemoving	-0.514 (0.463)
purposeother	0.093 (0.314)
purposerenewable_energy	-1.226 (1.289)
purposesmall_business	0.084 (0.383)
purposevacation	0.123 (0.421)
purposewedding	0.192 (0.395)
dti	0.022*** (0.004)
delinq_2yrs	0.061 (0.041)
total_pymnt	-0.001*** (0.00002)
total_rec_int	0.0004*** (0.00003)
recoveries	2.060 (9.081)
Constant	4.491*** (0.518)
Observations	50,000
Log Likelihood	-4,378.036
Akaike Inf. Crit.	8,824.072
Note:	*p<0.1; **p<0.05; ***p<0.01

From above, the significant predictors to ‘Loan is Bad’ variable are ‘Loan Amount’, ‘Term’, ‘Interest Rates’, ‘Grade G’, ‘Annual Income’, ‘DTI’, ‘Total payment’ and ‘Total Recovered Interest’.

This model can be used to predict loan defaults.

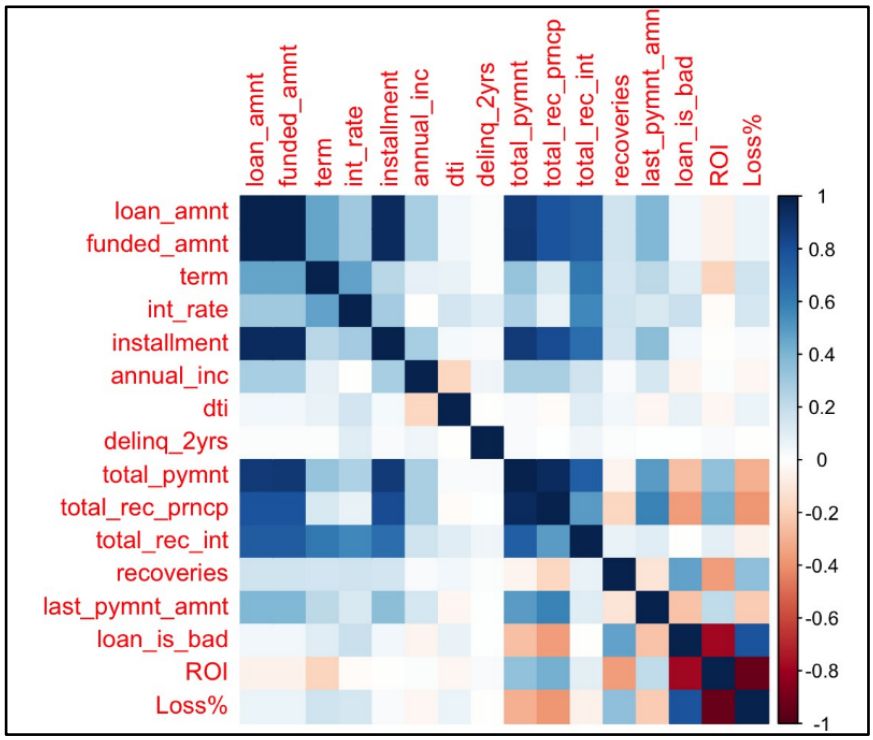
Following metrics are used to evaluate the performance of the model:

Accuracy	Precision	Recall	F1 Score
0.98328	0.9924395	0.9103098	0.9496021

Here the F1 score of 0.94 indicates that the logistic regression model has a high level of accuracy in predicting whether a loan is bad or not.

Correlation Matrix:

The following matrix shows correlation amongst different variables. It effectively shows the variables having significant relation to 'Loan is Bad'.



Based on the correlation matrix, following are the significant correlations with loan_is_bad variable:
ROI (-0.787754553), Loss% (0.781825147), Recoveries (0.478681409) and Total Recovered Principal (-0.364047715).
In the context of loans, the term "recoveries" refers to the amount of money that lenders are able to recover from borrowers who have defaulted on their loans.

Exploratory Data Analysis

1. Verification status: Most of the loans granted are when the status is either verified or source verified. The source will give more information about the borrower. We have identified that 'Verified' represents the highest category of loan amount where loans were granted. A verified source is less likely to default than a non-verified source.

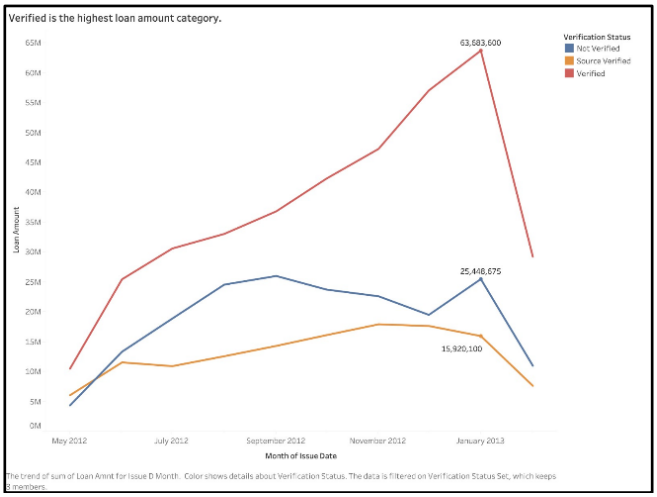


Exhibit 1: Loan Amount by Verification Status

2. Loan Amount by Grade: It gives the distribution of grade wise loans by loan amount. Grade B has the highest loan amounts and Grade G has the lowest loan amounts. This information could also be used to determine interest rates for different loan grades. From Exhibit 3, it is visible that Grade G has the highest interest rates and Grade A has the lowest interest rates. Higher grade (e.g. A) indicates lower risk, which translates to lower interest rates, while lower grade (e.g. G) indicate higher risk, which translates to higher interest rates.

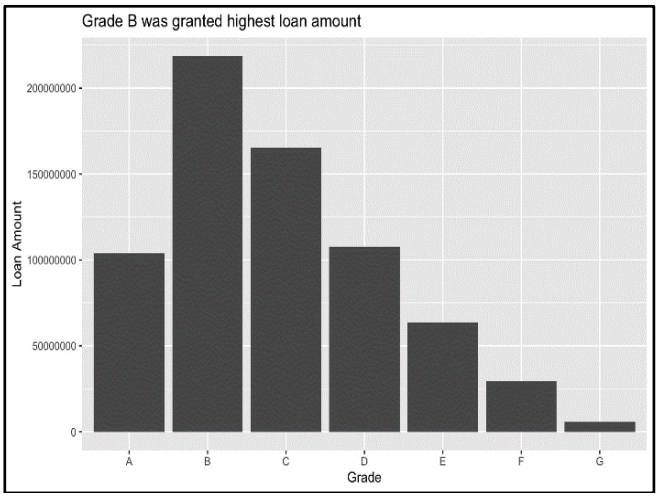


Exhibit 2: Loan Amount by Grade of Loan

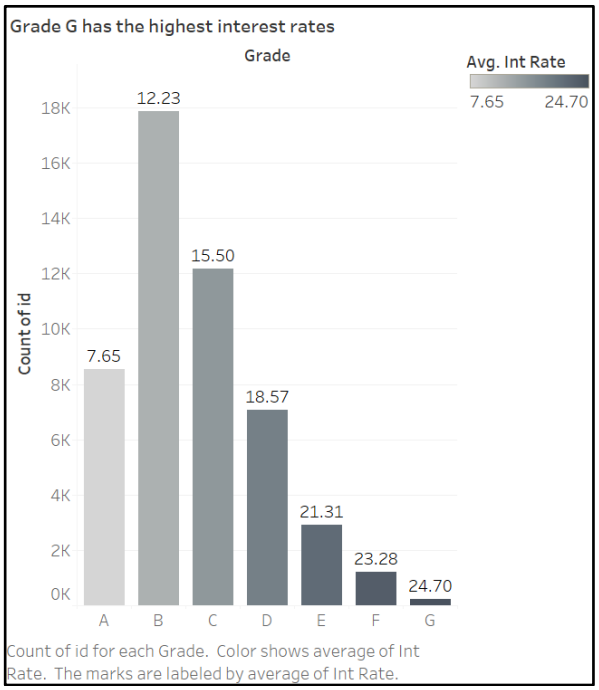


Exhibit 3: Count and Interest Rates across Grade of Loan

3. DTI by Grade: DTI (debt-to-income) ratio is a measure of a borrower's debt compared to their income, and it is calculated by dividing the borrower's monthly debt payments by their monthly gross income. Lenders typically prefer borrowers with a lower DTI ratio because they are considered less risky. Exhibit 4 shows, as the DTI ratio increases the Grade decreases. This is true as with higher DTI the creditworthiness of borrower decreases as it makes it difficult for them to make their loan payments. Grade A has the lowest DTI ratio.

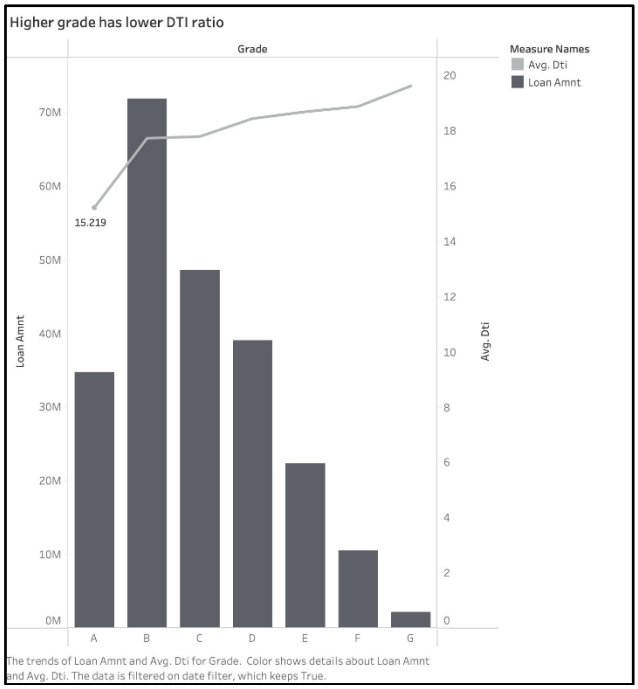


Exhibit 4: Loan Amount and DTI by Grades of Loan

4. Instalment amounts by Grade: The instalment amount for a loan may vary depending on factors such as the loan amount, repayment period, and credit grade. We see from chart below, that the instalment amount follows the same distribution as the loan amount by grade with Grade B having the highest instalment amount.

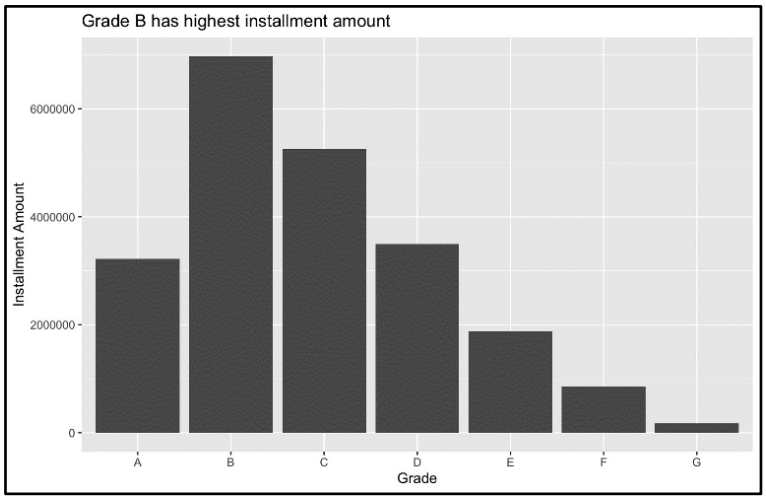


Exhibit 5: Instalment Amount by Grade of Loan

5. Recoveries by Verification Status: As mentioned earlier, the term "recoveries" refers to the amount of money that lenders are able to recover from borrowers who have defaulted on their loans. Recoveries were highest for 'Verified' type of loan status. Overall, higher recoveries for the 'Verified' type of loan status indicates that these loans are associated with lower default risk and may be a safer investment for lenders. Hence Lending Club should grant verified loans as it has recoveries twice as that of loans that have a non verified status.

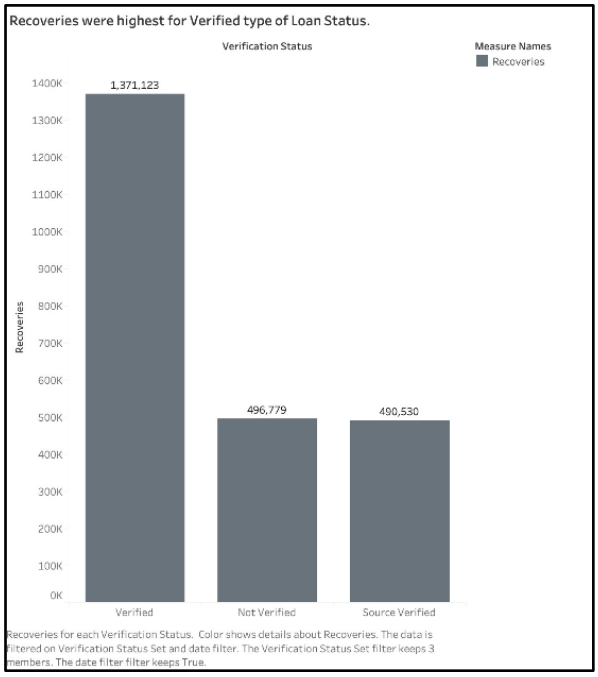


Exhibit 6: Recoveries by Verification Status

6. Loan Amount by Home Ownership: Exhibit 7 plots loan granted by home ownership. These can be any type of loans such as car, debt consolidation, credit card etc. Mortgage is distributed with the highest loan amount followed by rented property and least amount of loan granted to owned property. The bar graph in Exhibit 7 shows loans granted with mortgage as home ownership results in lowest loss percentage.

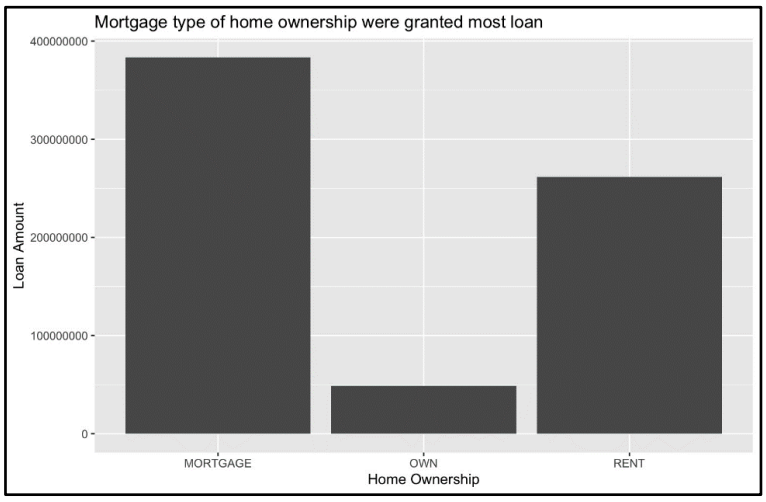


Exhibit 7 : Loan Amount by Home Ownership Type

7. Delinquencies by home ownership: Although mortgage ownership has the lowest loss percentage, it has the highest frequency of delinquencies measured within the last 2 years, followed by rented and least delinquencies in owned property. Loss is measured by actual default, and delinquencies measure delay in payments. To address the problem of loans turning into default, loan manager should assess borrowers' credit history and past delinquencies if any especially for those having mortgage type of home ownership.

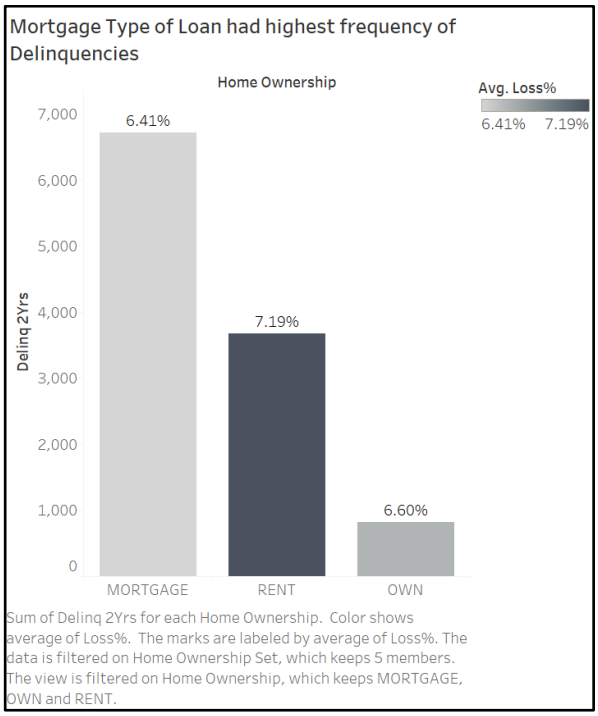


Exhibit 8: Delinquency (in 2 Years) and Loss % by Home Ownership

8. Home ownership versus ROI: Loan with owned type of home ownership has the highest ROI when compared with mortgage or rented property. Loan manager should consider approving more loans of borrowers with owned property.

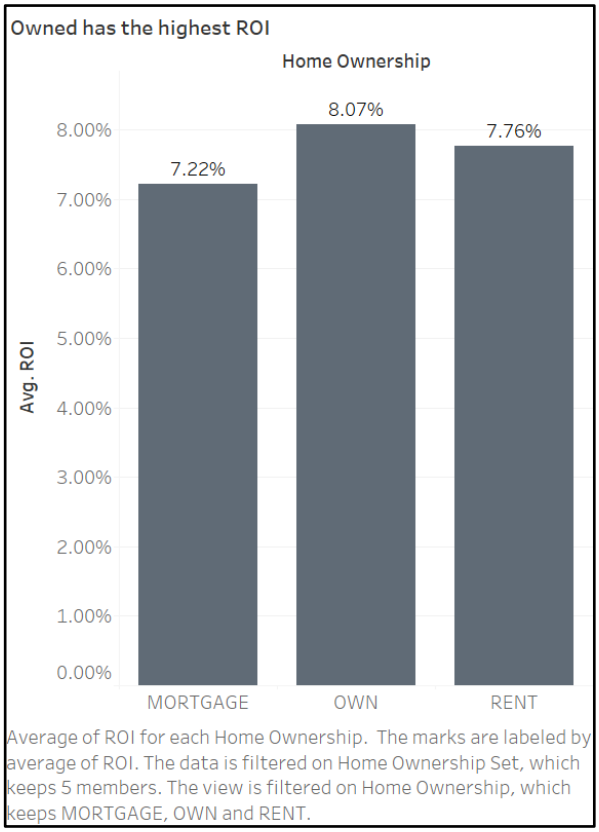


Exhibit 9: Average ROI by Home Ownership Type

9. Principal Amount recovered as compared to Loan Amount: Debt consolidation (for all time periods) has the highest amount of recovered principal followed by credit card, home improvement and small business. Suitable measures such as restructuring debt, negotiate the debt for lower settlement amount can be considered to recover principal amounts.

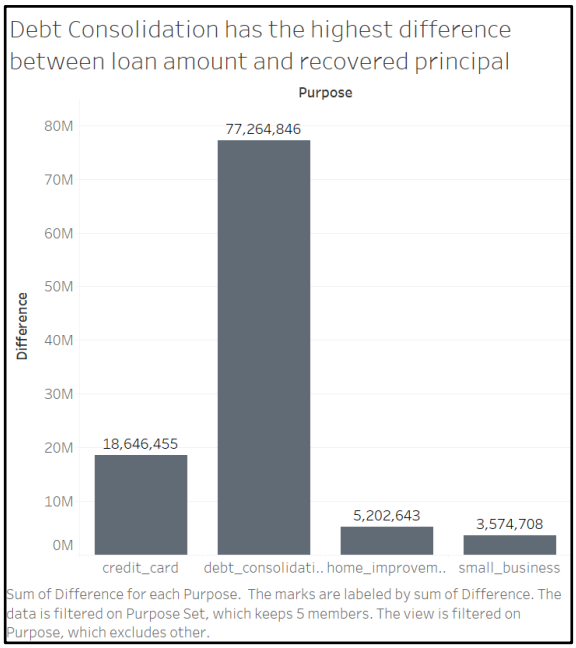


Exhibit 10: Principal minus total loan by Purpose of Loan

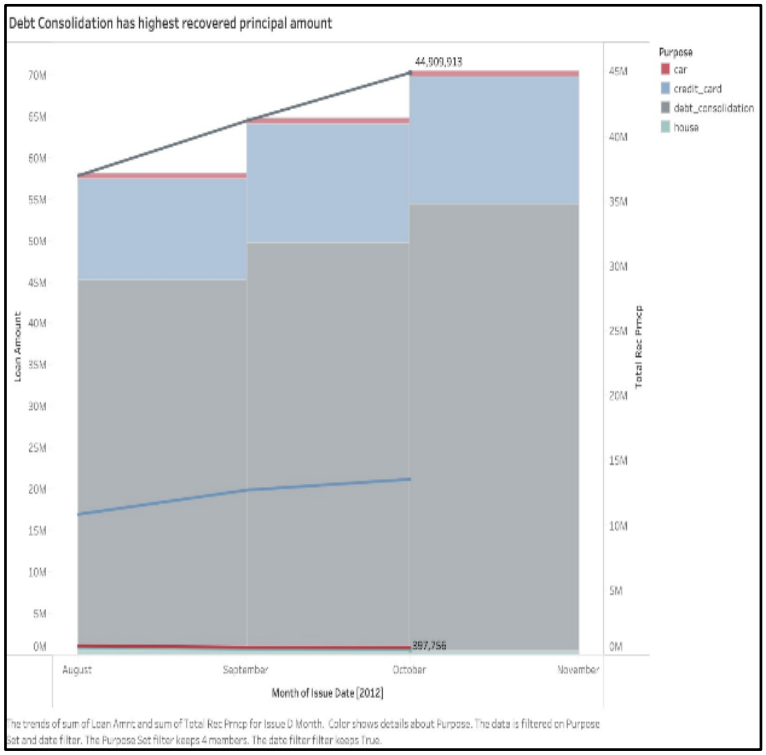


Exhibit 11: Loan Amount and Recovered Principal by Purpose of Loan over August - October

10. Loan Type versus ROI: Credit card loans have the highest ROI of ~11% when compared to loans for debt consolidation (~8%), car (~7%), and house (~1%) etc.

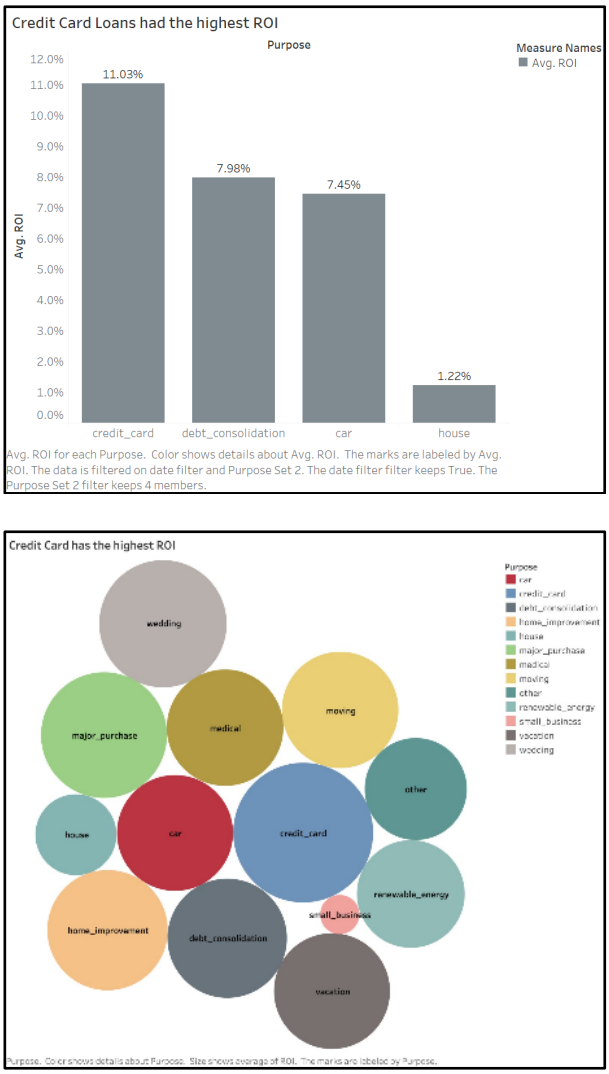


Exhibit 12: Average ROI by Purpose of Loan

11. State – wise ROI and DTI: Texas has the one of the highest ROI (~11%) and high DTI (18.22). High ROI means more returns in that state but high DTI ratio means it is over leveraged. This means it is riskier to give loans in that state anymore.

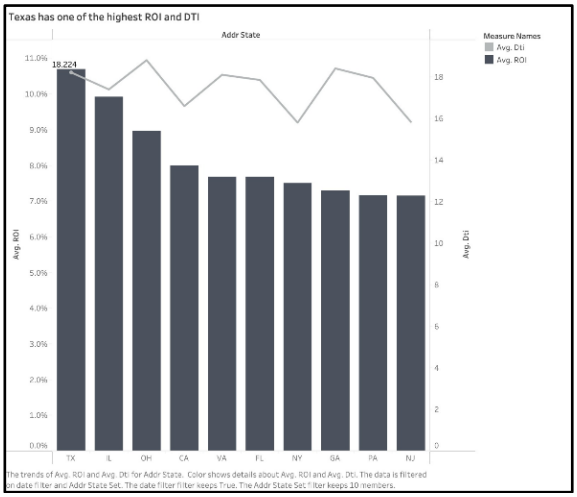


Exhibit 13: Average ROI and Average DTI by State

12. Interest rate by term: Interest rate is higher (>15%) for higher loan term (60 months). Interest rates for loans are influenced by several factors, including the loan term, loan amount, creditworthiness of the borrower, and prevailing market conditions. Some lenders may charge higher interest rates for longer-term loans, such as personal loans, due to the increased risk of default over a longer period of time. The longer the loan term, the more opportunity there is for unexpected events or financial difficulties to arise, making it more challenging for the borrower to repay the loan.

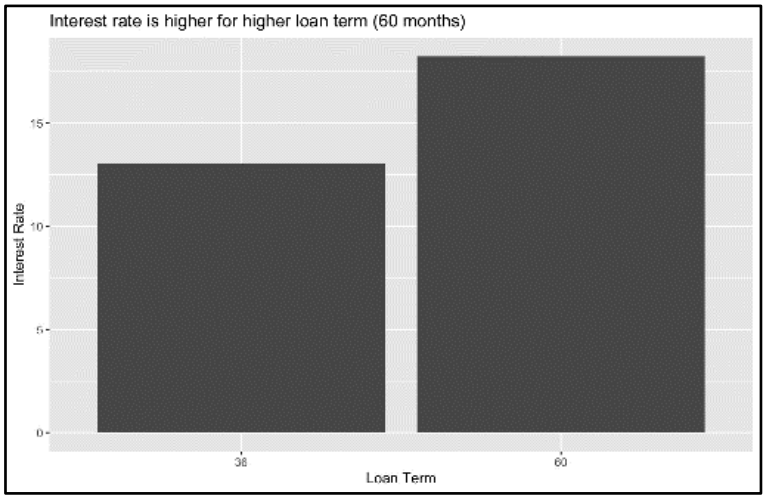


Exhibit 14:Average Interest Rate by Loan Term

13. Loan distribution over Loan Term: We see that more loan amount was distributed for 36 month loan term. Lenders may prefer to offer loans with shorter repayment terms in order to reduce their risk exposure. Shorter-term loans are generally considered to be less risky because the borrower is required to make larger monthly payments, which can make it easier for the lender to recover their losses if the borrower defaults.
Borrowers may prefer to take out loans with shorter repayment terms in order to pay off the debt more quickly and minimize the amount of interest they pay over the life of the loan.

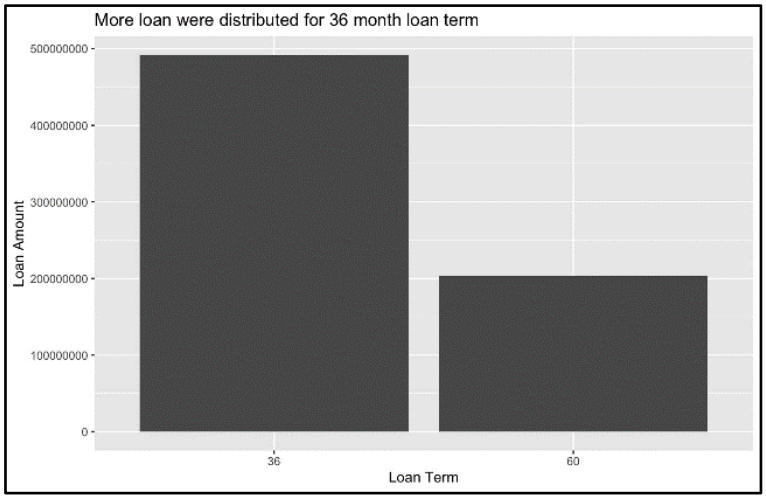


Exhibit 15: Loan Amount by Loan Term

14. Bad Loan Amounts (and %): We see that amount of good loans are ~ 600 M and bad loans are ~ 100 M. % of good loans comes out to be 83% of total loan amount in the dataset and % of bad loans comes out to be 17% of total loan amount in the dataset.

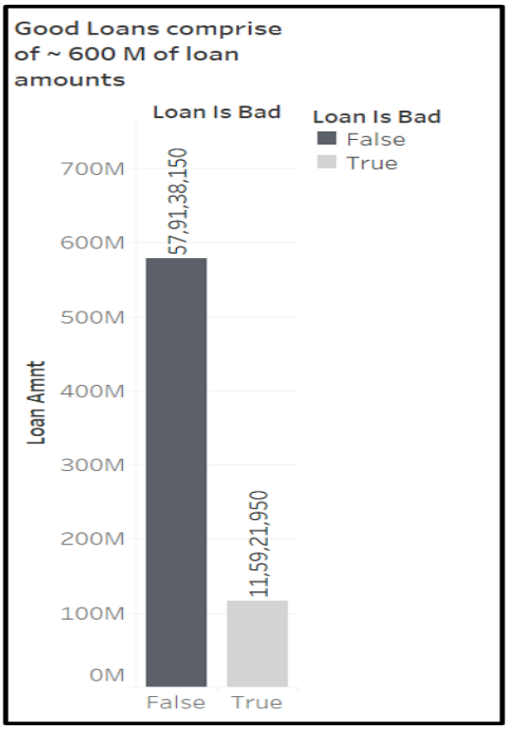


Exhibit 16: Good Loans comprise of ~ 600 M loan amounts

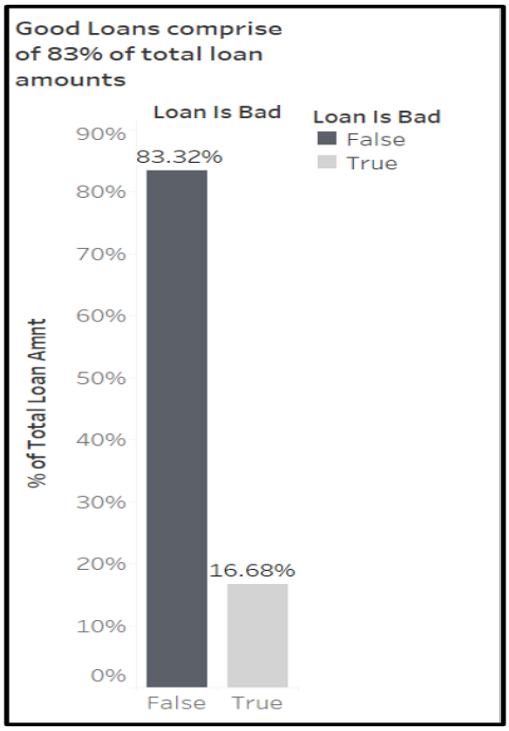


Exhibit 17: Good Loans comprise of 83% of total loan amount

Final Visualizations

Below are the final visualizations predicting trends of bad loans. We have created a dashboard with few selected visualisations. However, we have analysed all the charts in detail.

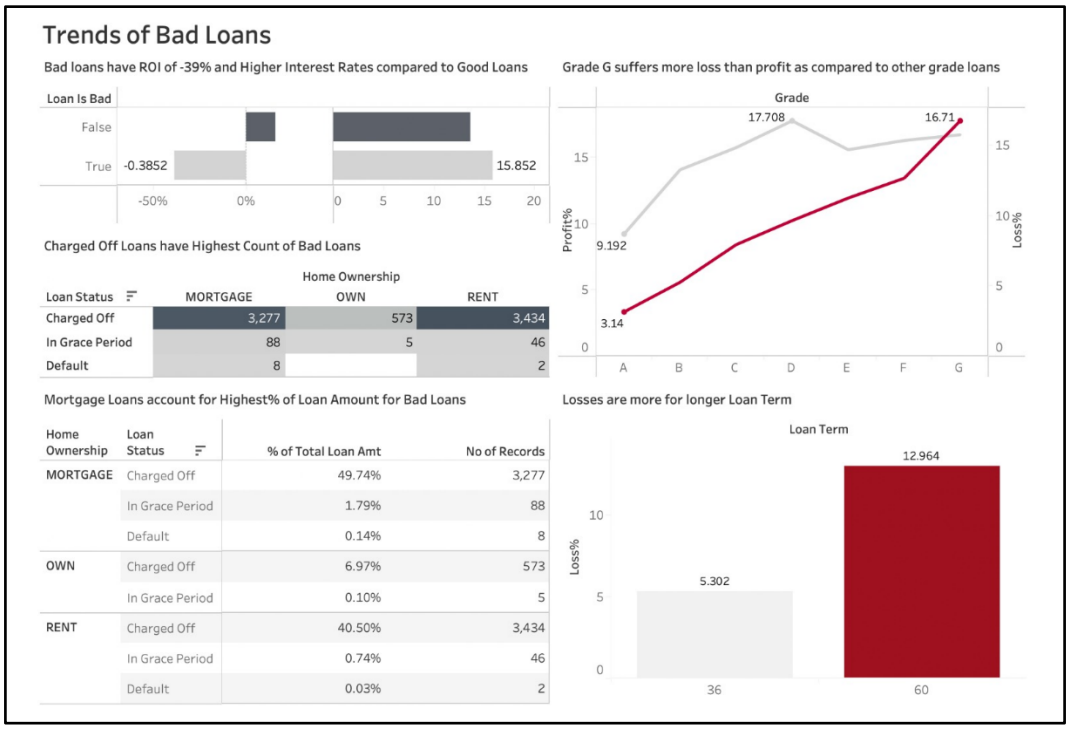


Exhibit 18: Trends of Bad Loans

1. Bad loans ROI and Interest Rates: A bad loan is typically a loan where the borrower has a higher risk of defaulting on the loan due to factors such as a poor credit score, high levels of debt, or a history of missed payments. We see that bad loans have a low ROI of -39% compared to good loans with a ROI of 16%. Also, we see that higher interest rates are associated with bad loans with an average interest rate of 15.85% compared to good loans with an average interest rate of 13.65%. In the second chart below, we see that bad loans are associated with cases with lower annual income and high interest rates. Lenders may charge higher interest rates for bad loans as a way to compensate for the increased risk of default. By charging a higher interest rate, the lender is able to earn a greater return on the loan if the borrower is able to make their payments as agreed. This can help offset the higher risk of default and reduce the lender's potential losses if the borrower is unable to repay the loan.

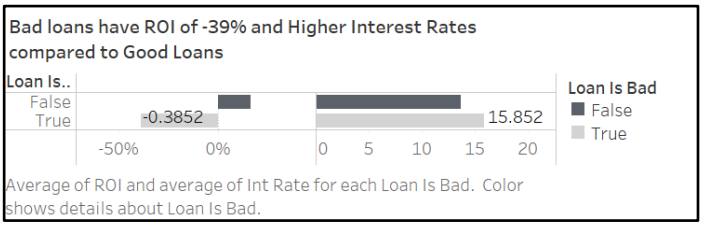


Exhibit 19: Average ROI (left) and Interest Rate (right) of Loan is Bad

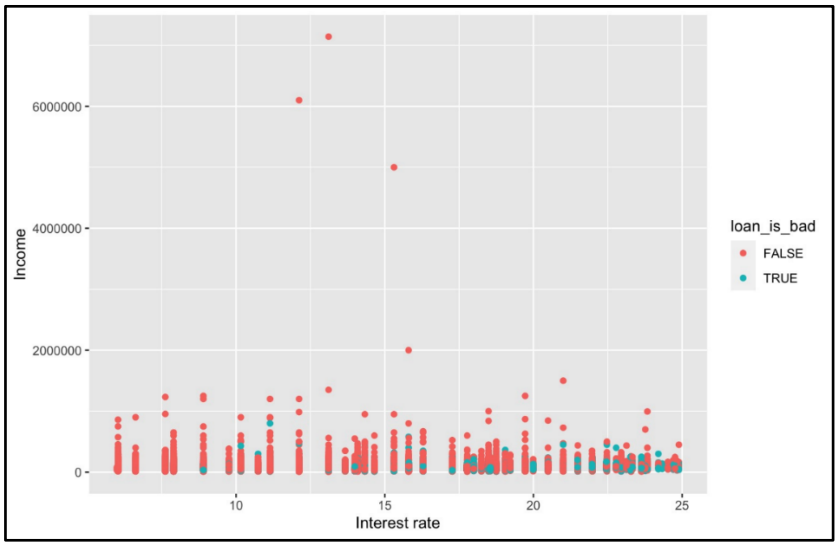


Exhibit 20: Income and Interest Rate by Loan is Bad (as color)

2. Grade wise Profit% and Loss%: Loans with lower grades (i.e. those considered to be higher risk) have a higher likelihood of default and may experience higher losses for lenders. Loans with higher grades (i.e. those considered to be lower risk) are generally expected to have lower default rates and may experience higher profits for lenders. Although Grade D and G are considered to offer higher average profit % compared to Grade A, B and C loans, we see that Grade G suffers more loss than profit compared to other grade loans. Lenders should be vary of offering such lower grade loans as they can result in negative ROI. It's important to carefully evaluate the risks and potential returns of any investment, including loans, and to consider factors such as diversification and risk management in order to mitigate potential losses.

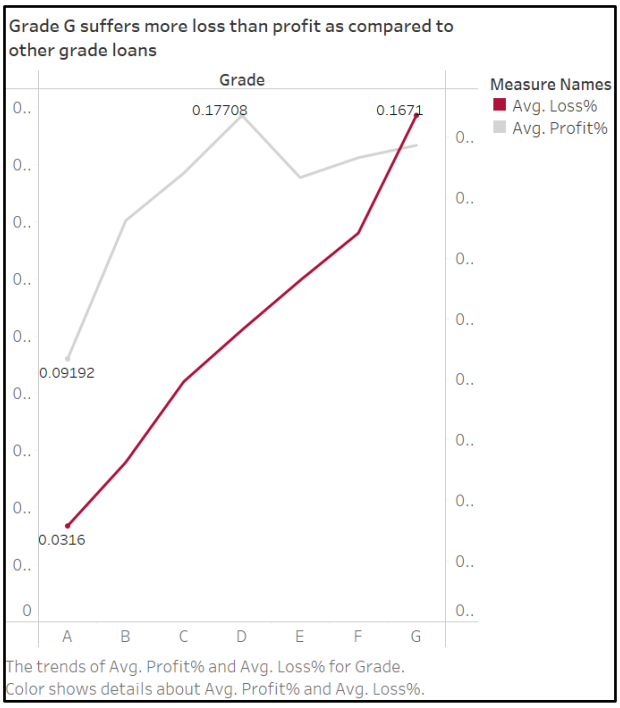


Exhibit 21: Average Profit % and Loss % by Grade of Loan

3. Bad Loan Count by Loan Status: Charged off loans are loans that the lender has written off as uncollectible and removed from their books as a loss. These loans are considered to be the most severe form of delinquency and indicate that the borrower has not made payments on the loan for an extended period of time.

It's important for lenders to carefully manage their loan portfolios and to employ effective risk management strategies in order to minimize the number of bad loans and reduce the risk of charge offs.

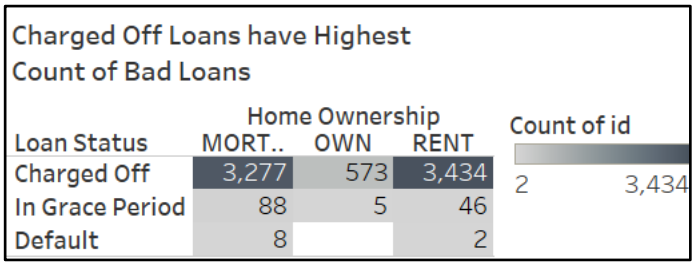
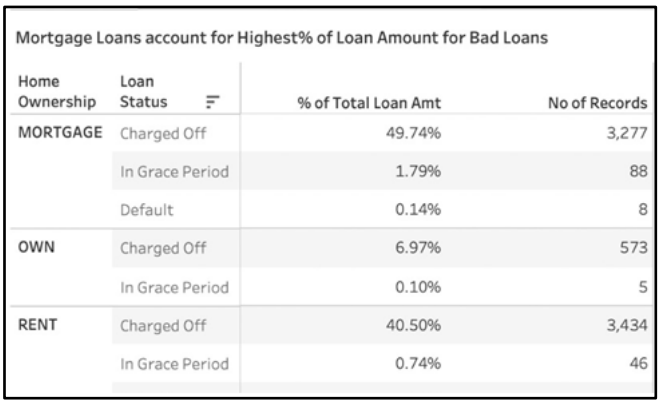


Exhibit 22: Count of loans by Home Ownership and Loan Status

4. Home Ownership vs Bad Loans: It's worth noting that the percentage of bad loan amount accounted for by mortgage property owners is the highest when compared to own and rented property owners. As was mentioned earlier, mortgage loans have highest delinquencies resulting in higher charged off loans. And since charged off loans are the highest % of bad loan amount, mortgage result in high % of bad loan amounts. Main reason behind this is that mortgage loans are often large and long-term commitments, increasing the default risk of the borrower.



Home Ownership	Loan Status	% of Total Loan Amt	No of Records
MORTGAGE	Charged Off	49.74%	3,277
	In Grace Period	1.79%	88
	Default	0.14%	8
OWN	Charged Off	6.97%	573
	In Grace Period	0.10%	5
RENT	Charged Off	40.50%	3,434
	In Grace Period	0.74%	46

Exhibit 23: Loan Amount % (of total) and Count of loans by Home Ownership and Loan Status

5. Term wise losses: In general, longer loan terms can lead to higher total interest costs and thus higher overall losses for borrowers. This is because with longer loan terms, borrowers have more time to accrue interest, which can add up significantly over time. We see from chart below, that longer loan term (60 months) has a high loss% of 13% and shorter loan term (36 months) have less loss% of ~ 5.3%. Lenders should be vary of the borrower's ability to pay the loan over long period of term as the interest rate burden also increases on the borrower.

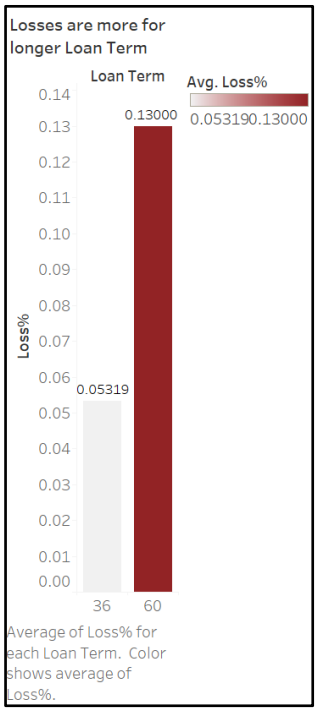


Exhibit 24: Average Loss % over Loan Term

6. DTI for Bad Loans: It can be seen from the graph that loan is considered bad at higher DTI. Debt to income ratio decides how much is the borrower indebted to take on loans. High debt to income ratio increases the chances of the borrower defaulting. Lenders should consider such parameter while deciding on giving loans. DTI > 20 is considered worst, categorizing majority of loans as bad.

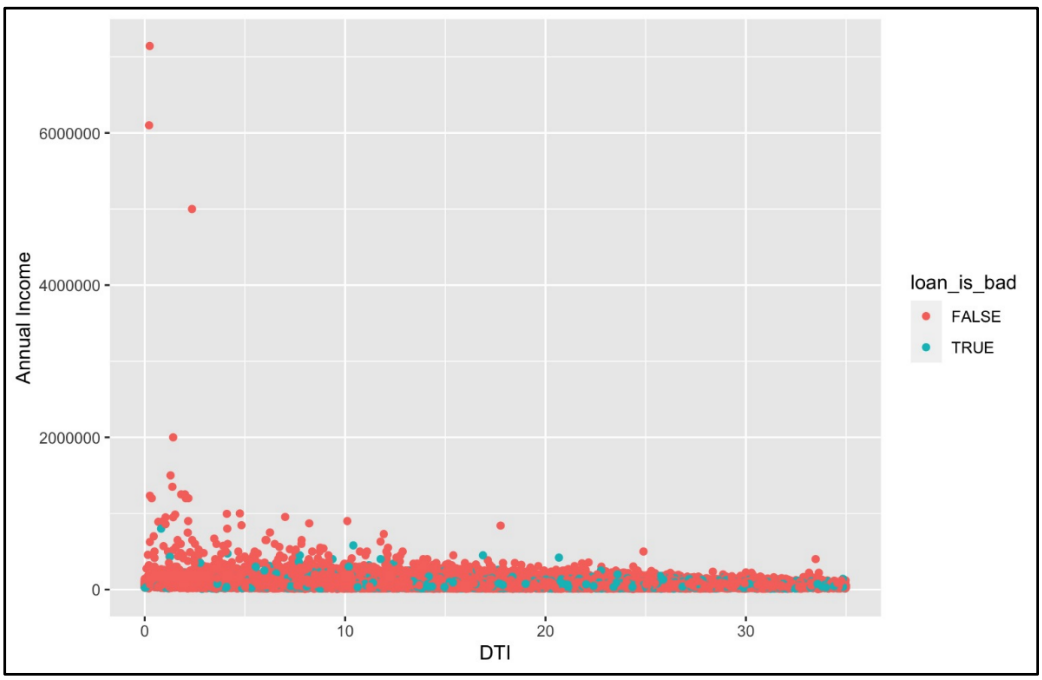


Exhibit 25: Annual Income and DTI by Loan is Bad (in color)

7. Purpose wise Bad Loans: When we filter the dataset based on bad loans, we see that debt consolidation and credit card have the highest amount of charged off loans and hence highest amount of bad loans. This is because they are typically unsecured loans, meaning that there is no collateral backing the loan.

Debt consolidation loans are used to pay off multiple debts, such as credit card balances and personal loans, and consolidate them into one monthly payment. These loans can be risky for lenders because they are often taken out by individuals who are already struggling to manage their debt. Lenders should have stricter norms while granting such loans.

Credit card loans often have high-interest rates, which can make it difficult for borrowers to pay off the balances.

Hence, we see high amount of bad loans in both debt consolidation and credit card.

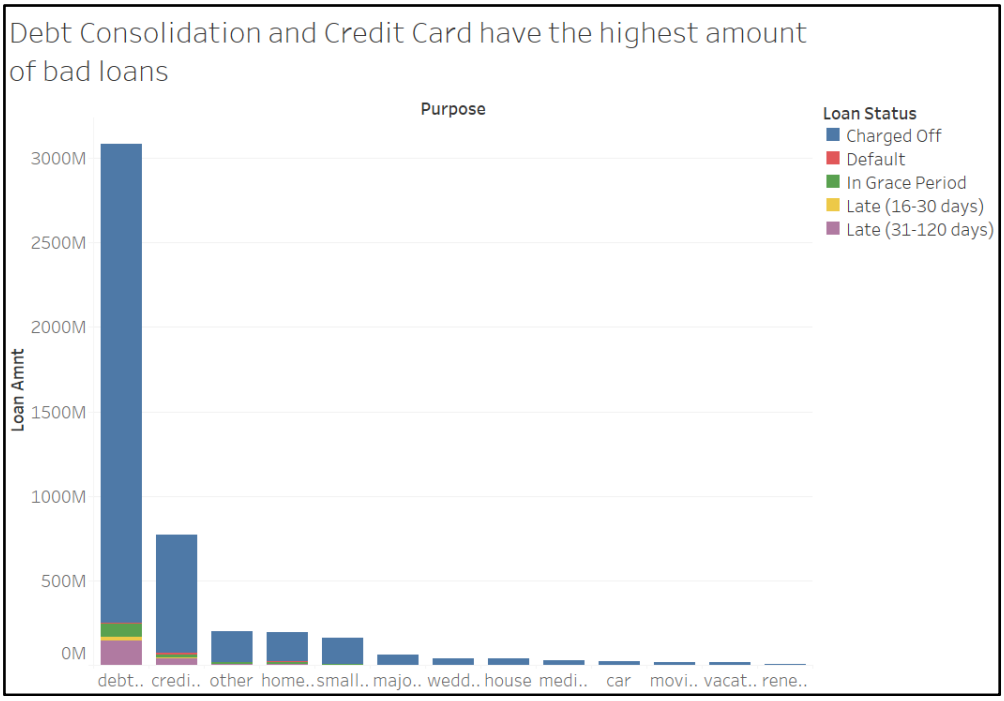


Exhibit 26: Bad Loan Amount by Purpose of Loan

Conclusion

To analyse the lending club business model, several factors were considered such as loan grades, loan amounts, interest rates, loan term, and purpose of the loan. The following takeaways can be derived from the analysis. These takeaways can help the loan manager in making a decision on which factors to grant the loan.

1. Low annual income and high interest rates are major considerations for bad loans (Exhibit:20)

2. Higher grades loans ensure higher profit% over loss % and hence higher average ROI on such loans (Exhibit: 21)

3. Good loans can have ROI touching 16% and bad loans have ROI dropping to -39% (Exhibit: 19)

4. Good loans have lower DTI (less than 20) (Exhibit: 25)

5. More than double the losses (at 13%) for longer term loans (60 months) than shorter term loans (36 months) which is at ~ 5% (Exhibit: 24)

6. 'Charged Off' accounts for maximum number of bad loans and has 'rent' and 'mortgage' as the frequent type of home ownership (Exhibit: 23)

7. 'Debt consolidation' and 'credit card' as purpose have the highest amount of bad loans (Exhibit:26) and hence stricter norms should be in place before granting such loans.

8. Large and long-term commitments such as mortgage loans increase the default risk of the borrower (Exhibit: 22)

9. Shorter-term loans are generally considered to be less risky as it minimizes the interest payment over the life of the loan (Exhibit: 15)

10. Lending Club can consider offering loans with high ROI such as credit card (11%), debt consolidation (~8%) and car (~7%) (Exhibit: 12)

11. Verified loans are more preferred with recoveries being ~2x over non-verified category (Exhibit: 6)

12. Higher grade ensures better creditworthiness, more surety of loan being granted but with lower interest rates (Exhibit: 2)

Below are several limitations in lending club loan data analysis:

1. Lack of up-to-date information: The data collected is from May 2012 to Feb 2013, which is not representative of the complete historical data. The default rate may also be higher than what is shown in the dataset since some borrowers may have stopped making payments on their loans after the data was gathered. Also, there might have been further recoveries post the data collection phase.

2. Imbalanced classes: The dataset contains a higher proportion of good loans compared to bad loans, which can impact the performance of predictive models.

3. Limited loan information: The information does not take into account multiple debts or collateral used to secure the loan. Multiple debts can lower the grade of the loan. Collateral used to secure the loan can lower the interest rates. Grades of loans or interest rates may differ for such loan types and effect of such data points has to be understood.

4. Sampling bias: Only loans that were issued and funded via the Lending Club platform are included in the dataset, thus they might not be entirely indicative of loans from other sources or the overall market.

Appendix

There has been a significant amount of existing work on similar problems. Many studies have explored the use of machine learning and statistical models to predict loan defaults, assess credit risk, and improve loan portfolio performance. Some examples of related work include:

1. *Consumer credit risk models via machine-learning algorithms - Andrew Lo*. (n.d.). Retrieved April 18, 2023, from https://alo.mit.edu/wp-content/uploads/2015/06/CRisk_final.pdf
2. *Credit risk analysis with machine learning techniques in peer-to-peer lending market www.diva-portal.org*. (n.d.). Retrieved April 18, 2023, from http://www.diva-portal.org/smash/get/diva2:1375762/FULLTEXT01.pdf
3. *Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020, September 25). Explainable Machine Learning in Credit Risk Management - Computational Economics. SpringerLink. https://doi.org/10.1007/s10614-020-10042-0*
4. Shi, S., Tse, R., Luo, W. *et al.* Machine learning-driven credit risk: a systemic review. *Neural Comput & Applic* **34**, 14327–14339 (2022). https://doi.org/10.1007/s00521-022-07472-2
5. *Yuan, Danny. Applications of Machine Learning: Consumer Credit Risk Analysis. 2015. https://dspace.mit.edu/bitstream/handle/1721.1/100614/932622145-MIT.pdf?sequence=1*