Part 1: Data Exploration and Evaluation

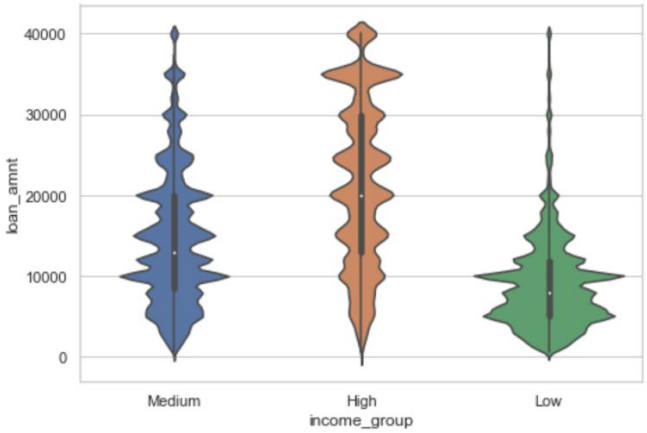
1. Loan & Income

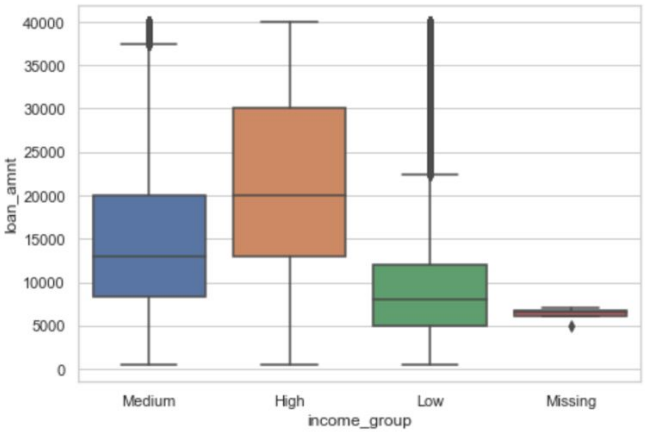
Based on the 20th and 80th percentiles of annual income, we will break income into 3 groups:

- High income group: annual income >= 100000
- Medium income group: 42000 <= annual income < 100000
- Low-income group: annual income < 42000

Summary:

1. From the violin plot below, we can see that the low-income group tends to borrow small loans and people with high (>= 100k) income tend to borrow larger loans.
2. From the boxplot below, it's clear that the median loan amount differs among the income groups. We also include a category where income is missing. There are 4 samples where income is missing. In the modeling section, we will use the median of income to impute the missing values so we can avoid the effect of outliers.
3. The 99.9th percentile of income is 600k but the max is 110M. Though it's possible that a person who makes 110M per year borrows a small loan from Lending Club, this chance is small. It could be that the currency is different, more info on the currency would be helpful. For this exercise, we can cap the income at 600k.



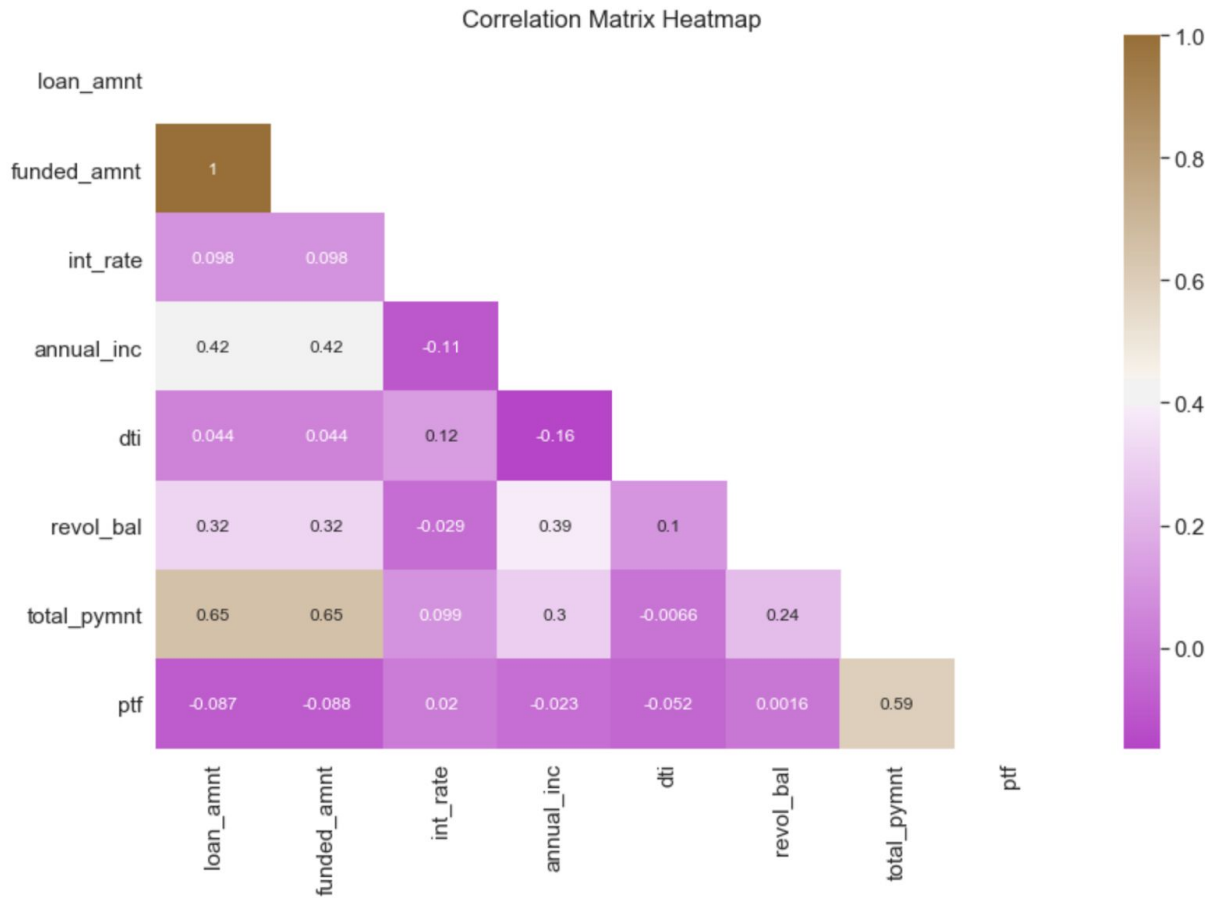


2. Correlation Heatmap

Summary:

- The funded amount and loan amount are highly correlated. For modeling, we probably only need to use ***funded amount*** in the model as that's what borrowers need to pay off.
- Total payment has a correlation coefficient of 0.65 with the funded amount.

- Most variables are not correlated with each other, we don't have to worry about multi-colinearity here.

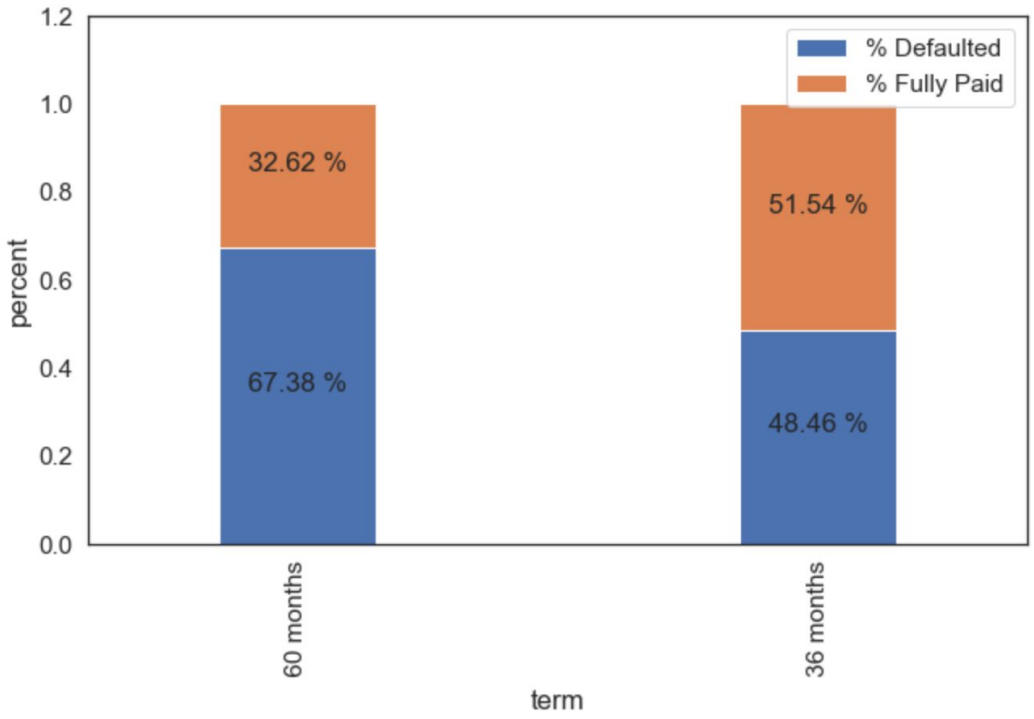


Part 2: Business Analysis

1) What percentage of loans has been fully paid?

If our goal is to evaluate whether the 36-month term loans would make for a good investment. We should probably remove loans with either less or more than a 36-month term. As shown in the bar plot below, the rate of default differs by almost 20% between 36-month and 60-month term loans. Therefore, I made the decision to remove loans with a 60-month term. This action will remove 650k samples from our dataset. In actual work, I'd consult with the product owner before making this decision.

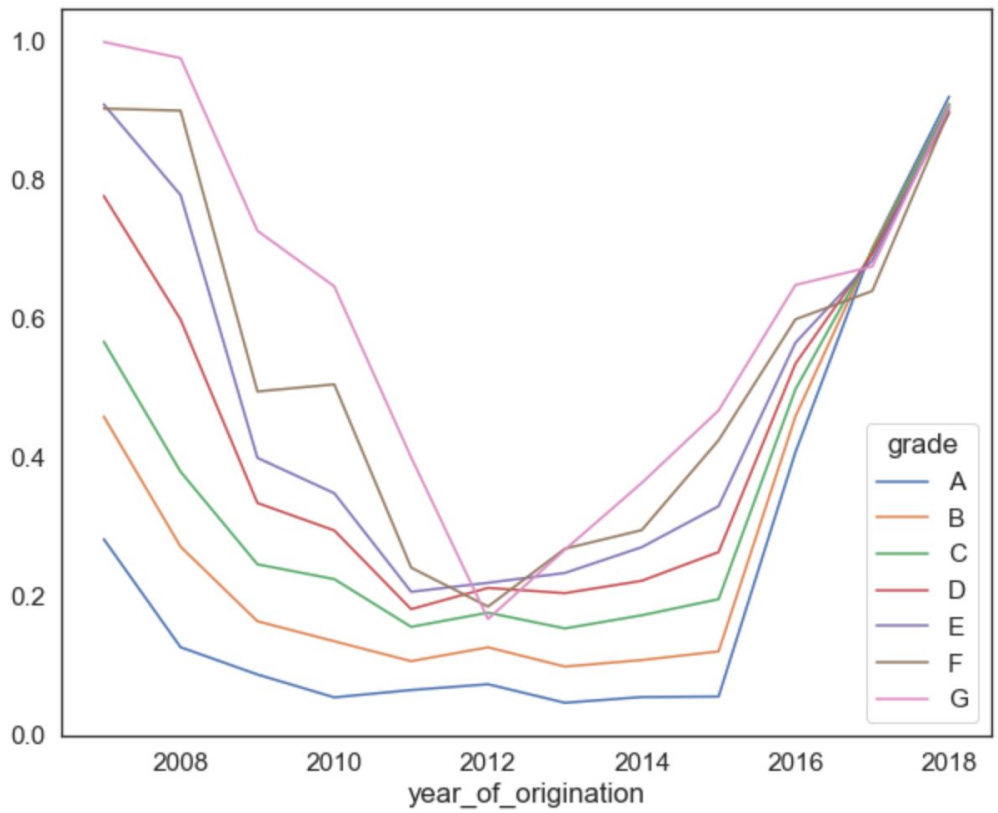
Answer: 51.54% of 36-month term loans are fully paid.



2.2 When bucketed by year of origination and grade, which cohort has the highest rate of defaults? Here you may assume that any loan which was not fully paid had "defaulted".

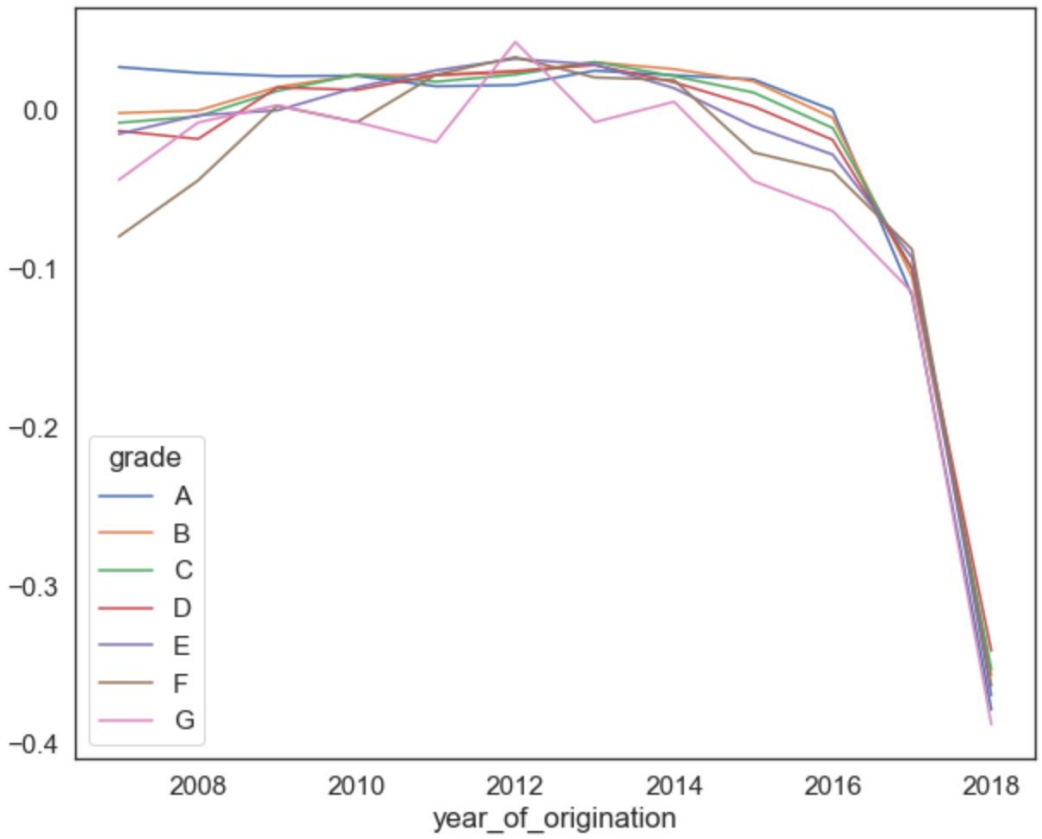
Answer:

- 2007 Grade G loans have the highest rate of defaults
- The high default rate in 2018 is likely due to our assumption that any loan which hasn't been fully paid is "defaulted"



2.3 When bucketed by year of origination and grade, what annualized rate of return have these loans generated on average?

- The rate of return is slightly above 0% (before 2017 for grade A-F. Grade G loans have a mostly negative rate of return across the years, indicating high risk.
- In 2017 and 2018, the rate of return has significantly dropped to below -10% likely due to many loans are still being paid off.



Part 3: Modeling

Framework

We want to consider the fact that loan default might behave differently across years. To build a high-quality model, it's important to set up a robust validation framework. For this exercise we will do the following:

- Split the data into training (data through 2017) and test (data of 2018)
- For the training data, apply a fixed start incremental window validation strategy:
    - Train model on data of 2007-2014 and validate on 2015
    - Train model on data of 2007-2015 and validate on 2016
    - Train model on data of 2007-2016 and validate on 2017
- To evaluate the model, we will look at the weighted avg. precision, recall, and F1-score across the validation sets
- Once we are satisfied with the model performance, train a model on data of 2007-2017 and make predictions on 2018 data and evaluate the model

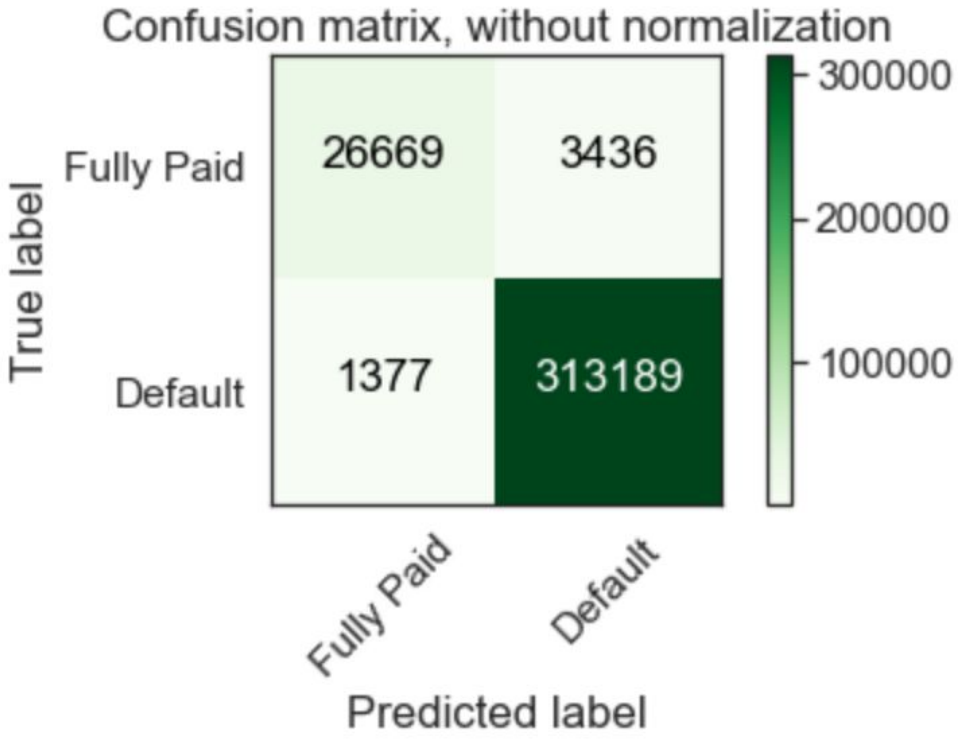
Validation Results

- Precision: The model is capturing default loans with high precision - out of all defaulted loans, the model predicts 98.9% of them correctly
- Recall: the recall is a bit low as compared to precision. After inspection, it turns the 2016 validation set only has a recall of 0.52.
- F1 score: across 2015-2017, the F1 score is pretty reasonable at 0.85
- Based on the validation performance, the model is effective but needs improvement on recall

Prediction

The model has an almost perfect precision and recall on the 2018 data. However, it should be noted that the data is highly skewed toward the default class - 90% of the 2018 data belong to the default class. As mentioned before, this is likely due to us assuming "current" loan as "default" which is probably not valid for loans of recent years.

	precision	recall	f1-score	support
0	0.95	0.89	0.92	30105
1	0.99	1.00	0.99	314566
micro avg	0.99	0.99	0.99	344671
macro avg	0.97	0.94	0.95	344671
weighted avg	0.99	0.99	0.99	344671



Part 4: Next steps

- For the data quality part, we should understand better why there are outliers in income, is it due to an inconsistent currency?
- Throughout the exercise, we assume any current loans as defaulted. This assumption is flawed as we don't have enough information on loans of more recent years. This assumption causes the distribution of data to be different between recent years and previous years. e.g., rate of return. To be conservative, it'd be reasonable to have a more strict definition of defaulted loans.
- For the business analysis and modeling part, I removed 60-month term loans. This drops 650k samples from the dataset. Whether this is the right interpretation of the question merits more context and information.
- For modeling, it'd be reasonable to look at years with low precision/recall (e.g., 2016) and understand the reasons behind.
- More feature engineering on the existing variables.
- More EDA on other variables outside of the requirements of this exercise.