Lending Club dataset-Milestone- Report

1. Introduction

1.1. Problem Statement

Lending club is a leading financial organization which provide personal loans, auto refinancing loans, business loans, and medical financing. To make greatest profit from the loan how to make decision for the application of the loan lenders is crucial. Financial condition of the lender is the important element to be considered. In this project we will concentrate on the analysis of the status when people apply the loan to find the pattern of high quality of loan lenders.

1.2. Client

1. Small loan providers such as Lending club, SoFi, Lendingtree, Guidetolenders, LendingPoint, Prosper and soon.
2. Banks which provide personal and business loan like Bank of America, Chase.
3. Credit card providers like Capital one, America Express.

1.3. Data resources

a. https://www.kaggle.com/jayrav13/unemployment-by-county-us
b. https://www.gaslampmedia.com/download-zip-code-latitude-longitude-city-state-county-csv/
c. https://github.com/liyepeng/Spring-Board-Data-Science-Track/blob/master/Lending%20Club%20project/Three_state_unemploy.csv
d. https://www.lendingclub.com/info/download-data.action

2. Data Wrangling

Part I.

Goal: Combine the unemployment rate by county dataset and zip code by county dataset to get a dataset containing unemployment information in different zip code area

Data cleaning:

1. Dataset ‘a’ is about unemployment rate in counties of USA in 2015. Cleaning job:
    a. Changing states name to abbreviation
    b. Adding missing states unemployment information (3 states) from dataset ‘c’.

2. Dataset ‘c’ is the information of counties and corresponding zip codes. And
    a. Keeping 50 states information. This data set contains 62 states among them there are associate states, military base.
    b. Making the counties' name match in both dataset ‘a’ and dataset ‘c’. Some county names are different from that in dataset ‘a’. For example: Key is in dataset ‘a’ and value is in dataset ‘c’.

VA = OH.replace({'county' : {'Bristol City' : 'Bristol', 'CharlesCity' : 'Charles City', 'IsleofWight' : 'Isle Of Wight', 'JamesCity' : 'James City', 'KingGeorge' : 'King George', 'KingWilliam' : 'King William', 'KingandQueen' : 'King And Queen', 'NewKent' : 'New Kent', 'PrinceEdward': 'Prince Edward', 'PrinceGeorge' : 'Prince George', 'PrinceWilliam' : 'Prince William', 'Radford City' : 'Radford', 'Salem City' : 'Salem'}})

3. The last step is to combine dataset 'a' and dataset 'b' together. One county may has several zip code (XXX00 format) and one zip code may correspond with several county. Final format of the file I got:

state	county	zip_code	Rate
NY	Suffolk	500	4.80833
NY	Suffolk	6300	4.80833
NY	Suffolk	11700	4.80833
NY	Suffolk	11800	4.80833
NY	Suffolk	11900	4.80833

Part II.

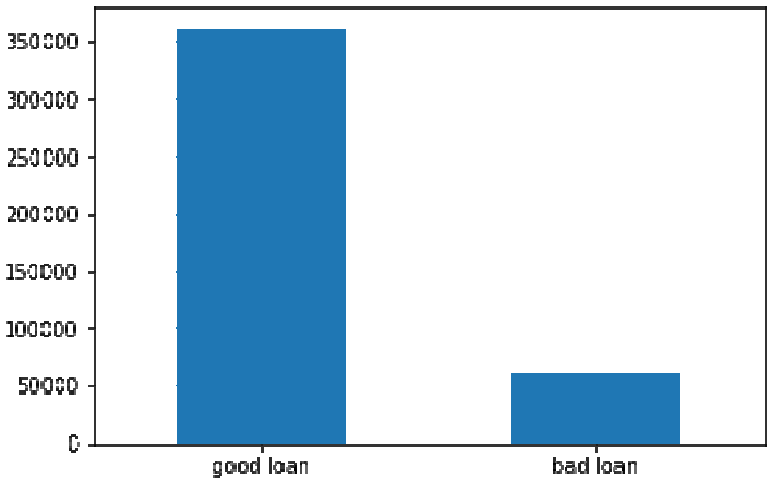
Data wrangling for lending club dataset

a. Filling blank with 'NaN'. Because the tree based algorithm doesn't care the missing values I keep these values in original status.
b. Deleting blank columns, deleting post loan and hard ship variables.
c. My target variable is loan_status and try to keep the variables it help to make the decision for the loan.
d. Data format wrangling: get rid of '%', change the strings to lower case, check if there are duplicate records.
e. Data engineering: There are more than 100000 categories for emp_title variable. I choose top ten titles and create 10 dummy variables so it can be used in EDA and model building.
f. Concatenating unemployment rate in different zip code file with lending club data set on zip code. My aim is to check if the unemployment rate has connection with loan status.
g. There are some outliers in the part of the numerical variables. The way to deal with them is depending on what kind of machine learning algorithms I plan to use. The following is the numerical variables have more than ten percentage of outliers:

```
Column delinq_2yrs has 87087 outliers
Column pub_rec has 74415 outliers
Column tot_coll_amt has 66407 outliers
Column num_accts_ever_120_pd has 102993 outliers
Column pub_rec_bankruptcies has 50916 outliers
```

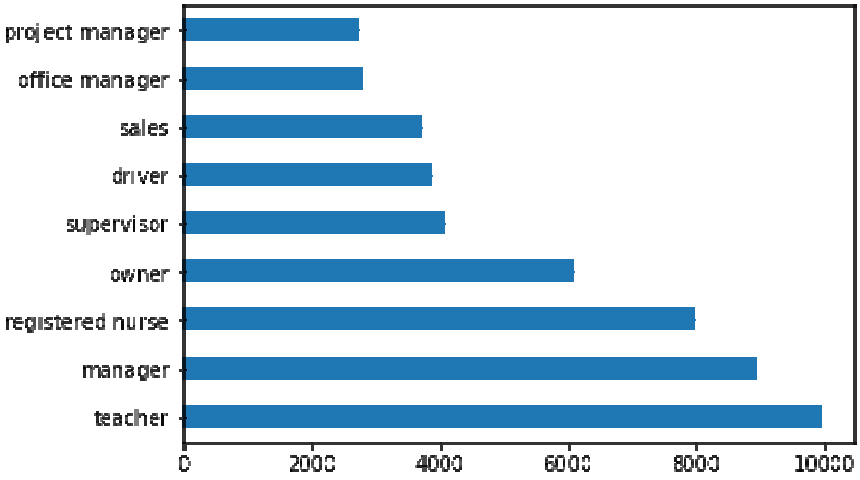
3. EDA and Data storytelling

Question 1. How many good loans and bad loans are there?

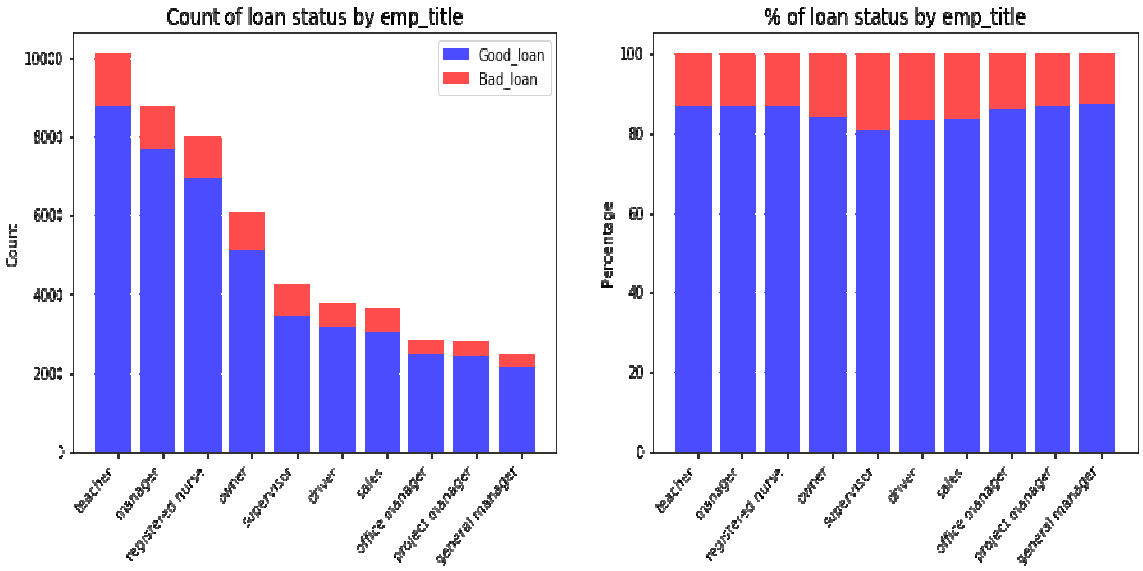


Bad loan's percentage is about 20% in 2015. 2015 is eight years after the economic depression since 2007. 20% bad loan percentage means among 5 lenders there will be one lenders in 'Charged Off', 'Late (16-30 days)', 'Late (31-120 days)' or 'Default' status. This rate is relatively high.

Question2. The relationship between loan status and employment title

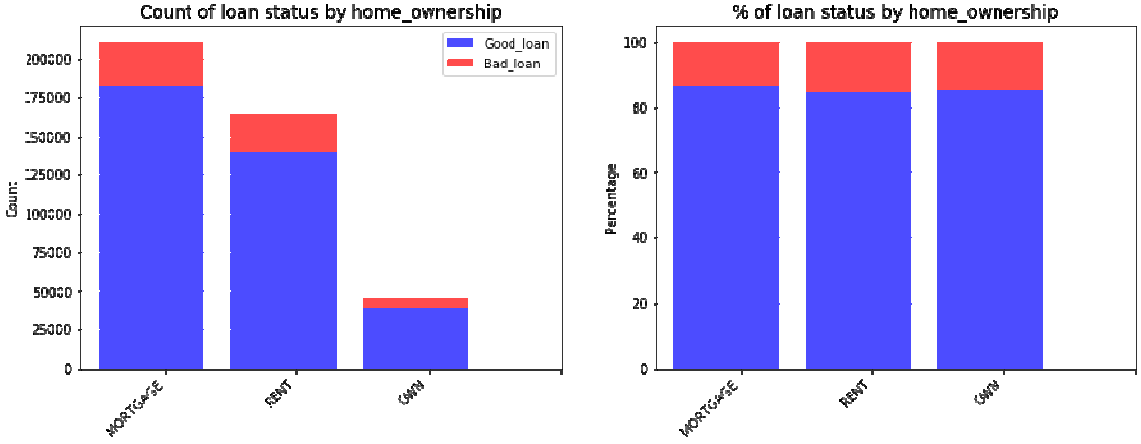


Teacher rank first in the top ten occupations of lenders. The loan count for the teacher is more than three times of that for project manager. It shows teacher are in bad financial condition though it is a respectable and import career to the society.



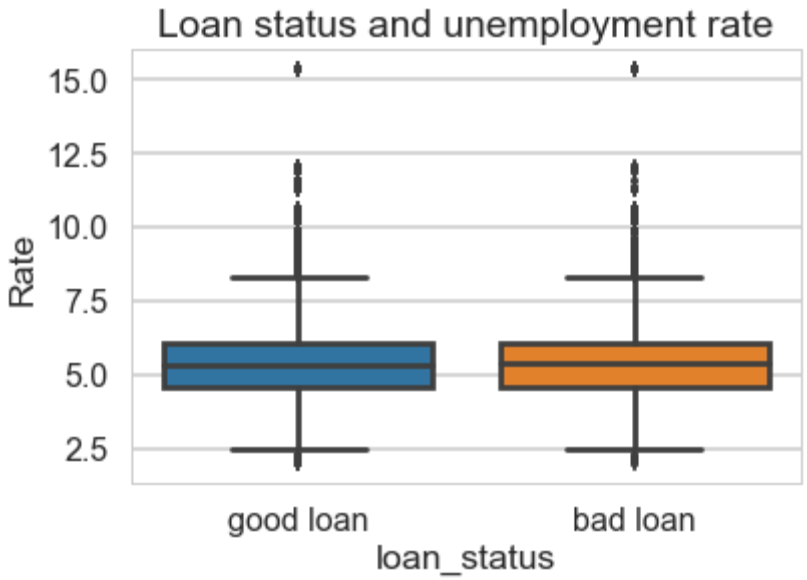
Teacher rank first in the top ten occupation of borrows while they seem to have lowest bad loan percentage. Teacher's salary is low and they have great demand of loan. But they also have better credit to pay off the debt on time than other occupations. Supervisor seems to have the highest bad loan percentage and the reason is unknown.

Question3. The relationship between loan status and home ownership

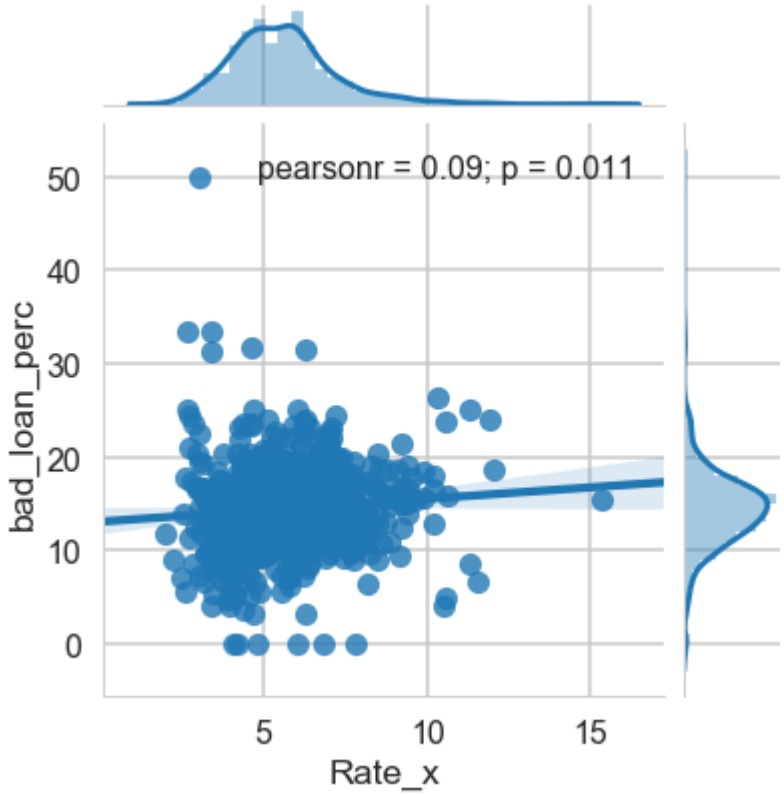


The lenders who are in mortgage have the largest loan count but they have the lowest bad loan percentage compare to the home renters or home owners. So this group of people are the valuable customers for lending organization.

Question4. The relationship of zip code, unemployment rate and bad loan percentage

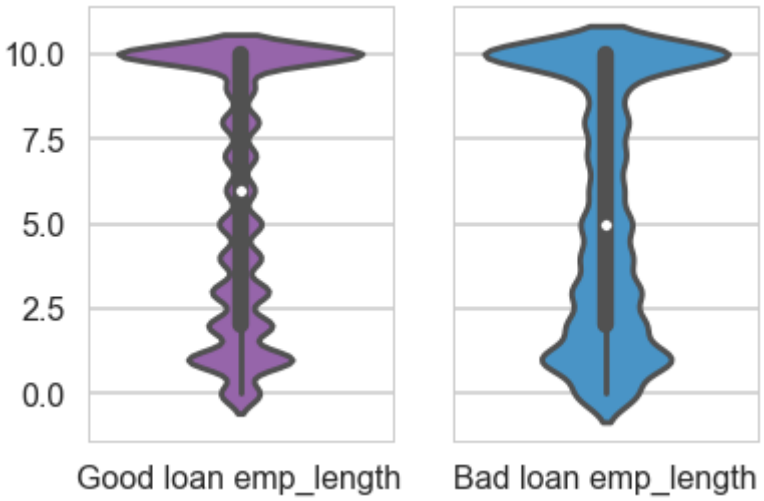


Good loan lenders seem to have the close median unemployment rate with bad loan lenders.



To investigate the same zip code area bad loan percentage has a positive relationship with unemployment rate. The higher the unemployment rate the higher the bad loan percentage.

Question5. The relationship of employment length and bad loan percentage



Average employment length of good loan lenders is higher than bad loan lenders. It is intuitively right for the lender who has longer employment length would be in good financial situation.

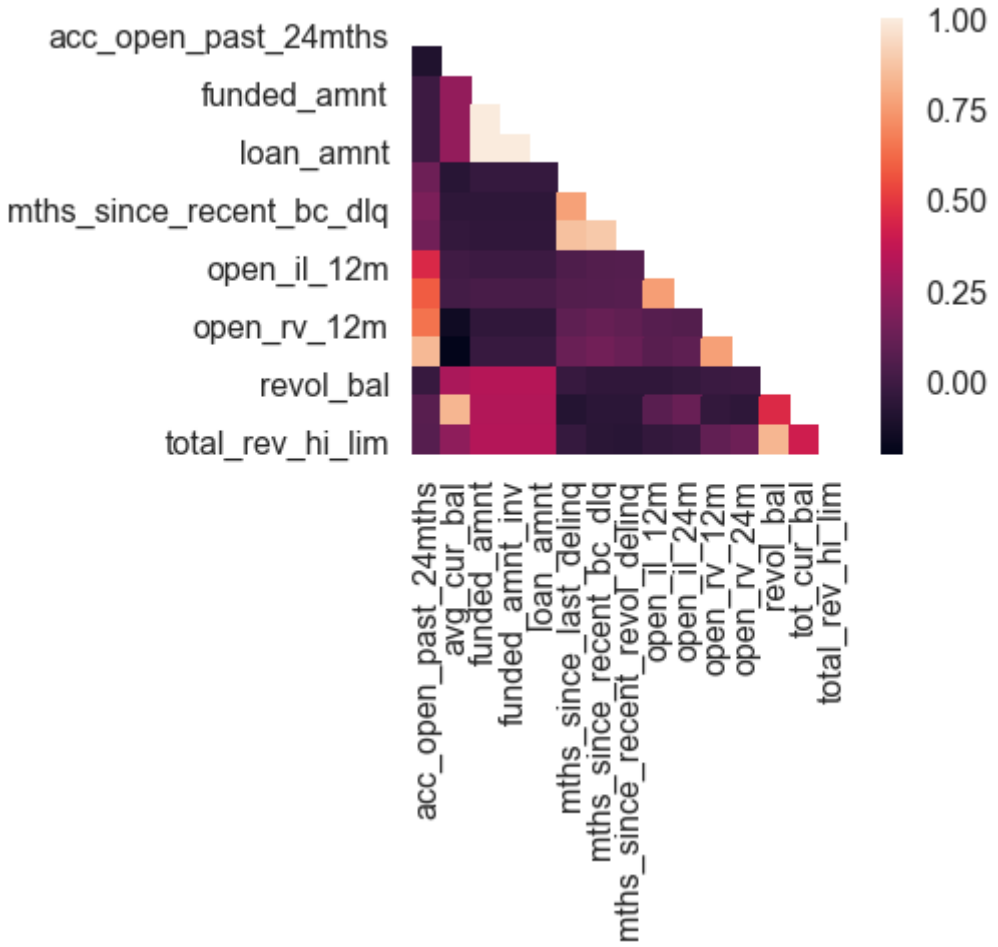
4. Statistical Analysis

a. Numeric variables correlation analysis

Highly correlated variables:

variable_0	variable_1	correlation
funded_amnt	loan_amnt	1
funded_amnt_inv	loan_amnt	0.99999
funded_amnt_inv	funded_amnt	0.99999
open_il_24m	open_il_12m	0.76055
open_rv_24m	open_rv_12m	0.76726

Hot map for part of the highly correlated variables



There are more than 30 pairs of numeric variables highly correlated (correlation coefficient is more than 0.75). Among them funded_amnt/loan_amnt, funded_amnt_inv/loan_amnt and funded_amnt_inv/funded_amnt's correlation coefficient are close to 1.

b. Chi square test for the loan status and employment title

Hypothesis

H 0 : In the population, variable 'loan_status' and variable 'emp_title' are independent.

H 1 : In the population, variable 'loan_status' and variable 'emp_title' are dependent.

Test statistics

chi_squared_stat: 269.3493768308436

Critical value: 21.0260698175

P value: 0.0

Conclusion:

We reject H 0 and consider variable 'loan_status' and variable 'emp_title are dependent. In other words loan status depends on what kind of employment title of the borrowers.

c. Chi square test for loan status and home ownership

Hypothesis

H 0: In the population, variable 'loan_status' and variable 'home_ownership' are independent.

H 1: In the population, variable 'loan_status' and variable 'home_ownership' are dependent.

Test Statistics

chi_squared_stat: 1692.8473338733406

Critical value: 11.0704976935

P value: 0.0

Conclusion:

We reject H 0 and consider variable 'loan_status' and variable 'home_ownership' are dependent. It means loan status depends on whether borrowers are renting a home or have purchased house.

d. Kruskal-Wallis H-test for the median of employment length of different loan status

The Kruskal-Wallis H-test tests the null hypothesis that the population median of all of the groups are equal. It is a non-parametric version of ANOVA. The test works on 2 or more independent samples, which may have different sizes. Note that rejecting the null hypothesis does not indicate which of the groups differs. Post-hoc comparisons between groups are required to determine which groups are different.

Hypothesis:

H 0: The population median of employment length in good loan borrowers and bad loan borrowers are equal.

H 1: The population median of employment length in good loan borrowers and bad loan borrowers are not equal.

Test statistics:

Kruskal Wallis H-test test:

H-statistic: 189435635.441

P-Value: 0.0

Conclusion:

We reject H 0 and consider the population median of employment length in good loan borrowers and bad loan borrowers are equal. The median of good loan borrowers' employment length is higher than that of bad loan borrowers.

e. Two sample t-test for the mean of unemployment rate in good loan borrowers and bad loan borrowers

Hypothesis:

H 0: The population mean of unemployment rate of good loan borrowers and bad loan borrowers are equal.

H 1: The population mean of unemployment rate of good loan borrowers and bad loan borrowers are not equal.

Test statistics:

ttest_ind: t = -9.58285 p = 9.71499e-22

Conclusions:

We reject H0 and consider the mean of unemployment rate of good loan borrowers is lower than that of the bad loan borrowers.

5. Milestone Project Report Conclusion

We have now completed data wrangling, inferential statistics and EDA on our lending club dataset. With this, we are now ready to use it for the next step of data modelling. However, these steps are repetitive process and we may have to repeat again during subsequent phases of data science.