Lending Club Case Study
Insights of EDA

Guozao Meng, Jiajing Gu
13-05-2023

1. Data Cleaning Methodology

Raw data has several quality issues. Before analyzing it, we implemented cleaning method.

- **Removing Columns**
  - Useless data: Columns related to human behavior — Removed 27 columns
  - >50% missing values — Removed 57 columns
  - Only 1 values — Removed 8 columns
- **Dropping Rows**
  - In “loan_status” column, value is “Current” — Dropped 1140 rows
  - Drop duplicates — No duplicates to drop

1. Data Cleaning Methodology

To be completed

Raw data has several quality issues. Before analyzing it, we implemented cleaning method.

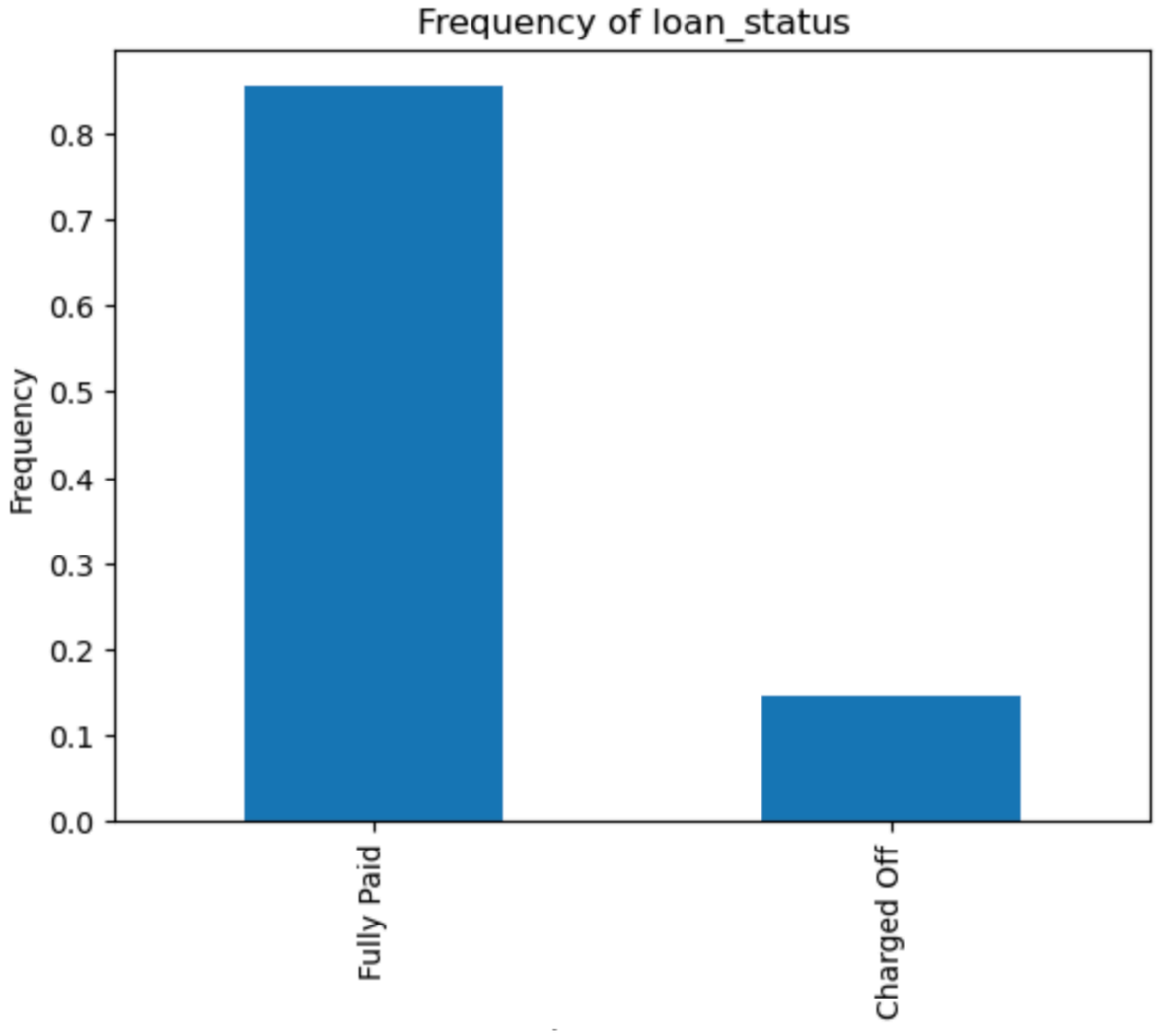
- **Impute Null Data**
  - Impute with “Unknown”

    Imputed 2.67% data
    - For “emp_length”, the correct value is unknown, so “ Unknown” is the best imputation value
  - Impute with “0.0”

    Imputed 1.81% data
    - For “pub_rec_bankruptcies”, the majority value is 0.0

2. Data Analyze and Insights

2.1. Most clients are willing to fully pay the loan.

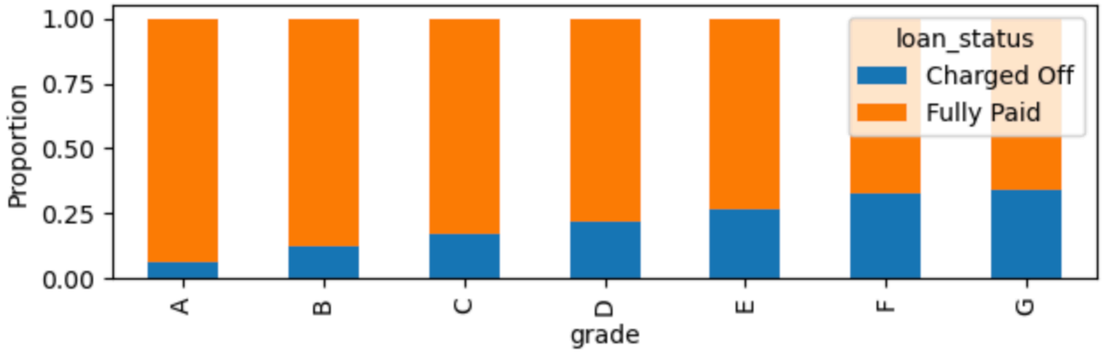


- In historical data
  - Fully paid percentage is 85.4%
  - Charged off percentage is 14.6%

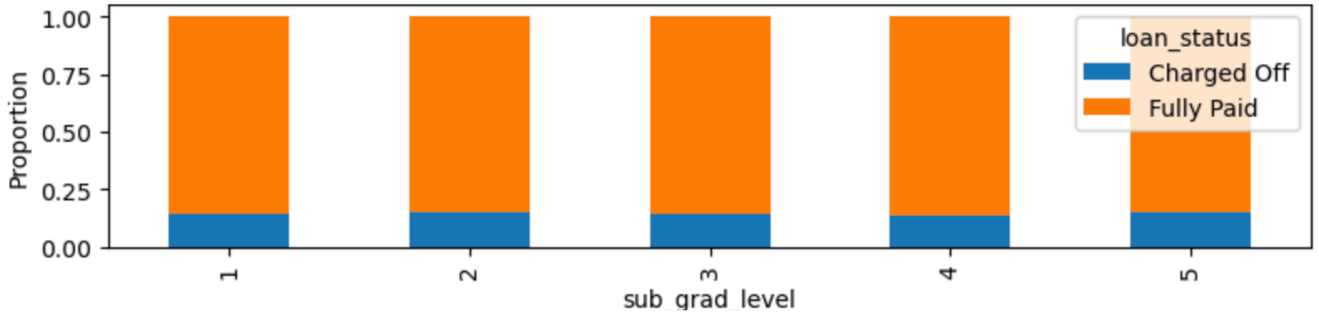
In order to identify the clients more willing to charge off and control the risks, we need to find features that related to charged off clients.

2. Data Analyze and Insights

2.2.1. The possibility of charged off is related with grade



	Charged Off	Fully Paid	Total
A	0.06	1.0	1.06
B	0.122	1.0	1.122
C	0.172	1.0	1.172
D	0.22	1.0	1.22
E	0.268	1.0	1.268
F	0.327	1.0	1.327
G	0.338	1.0	1.338

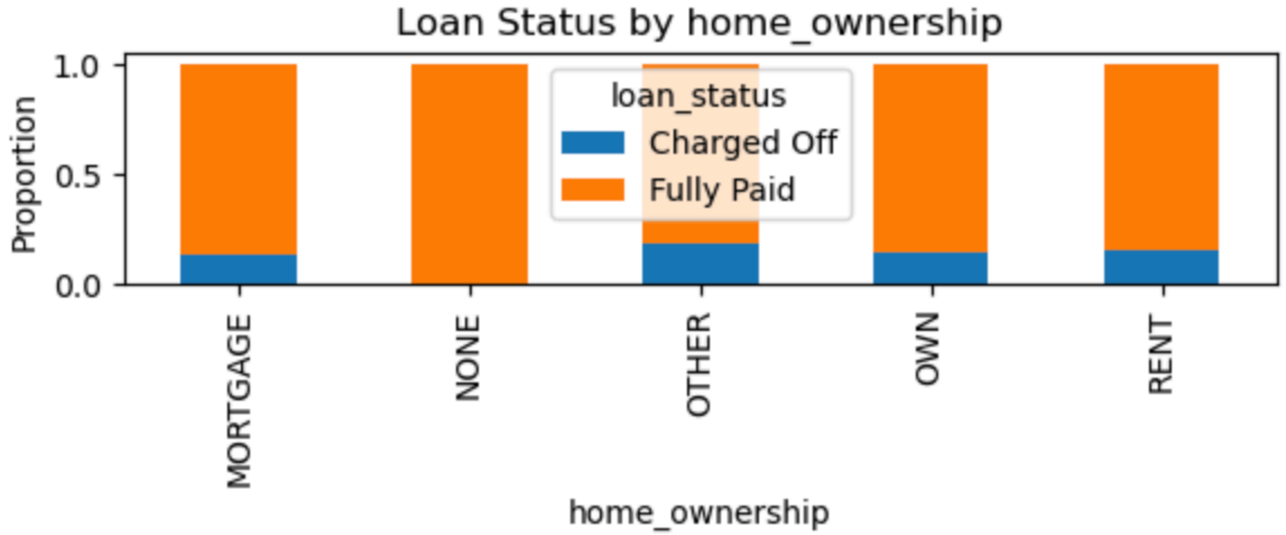


	Charged Off	Fully Paid	Total
1	0.145	1.0	1.145
2	0.151	1.0	1.151
3	0.145	1.0	1.145
4	0.14	1.0	1.14
5	0.148	1.0	1.148

The possibility of borrowers' charging off is G > F > E > D > C > B > A. Whereas the sub-grade classification doesn't have a significant influence on the possibility.

2. Data Analyze and Insights

2.2.2. The borrower with `NONE` homeownership are more likely to fully pay the loan

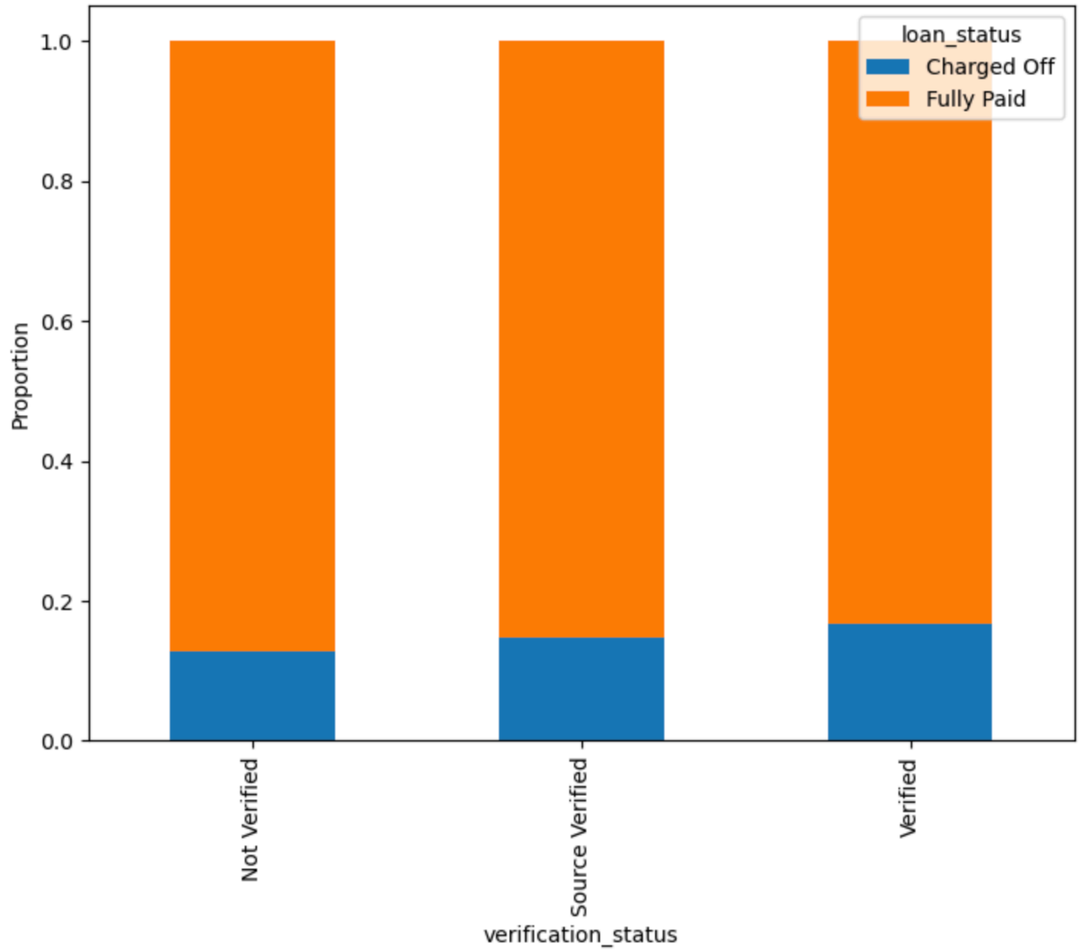


	Charged Off	Fully Paid	Total
MORTGAGE	0.137	1.0	1.137
NONE	0.0	1.0	1.0
OTHER	0.184	1.0	1.184
OWN	0.149	1.0	1.149
RENT	0.154	1.0	1.154

Other types of home ownership don't have significant differences in loaning paying status.

2. Data Analyze and Insights

2.2.3. The income verification status has influence in loaning status

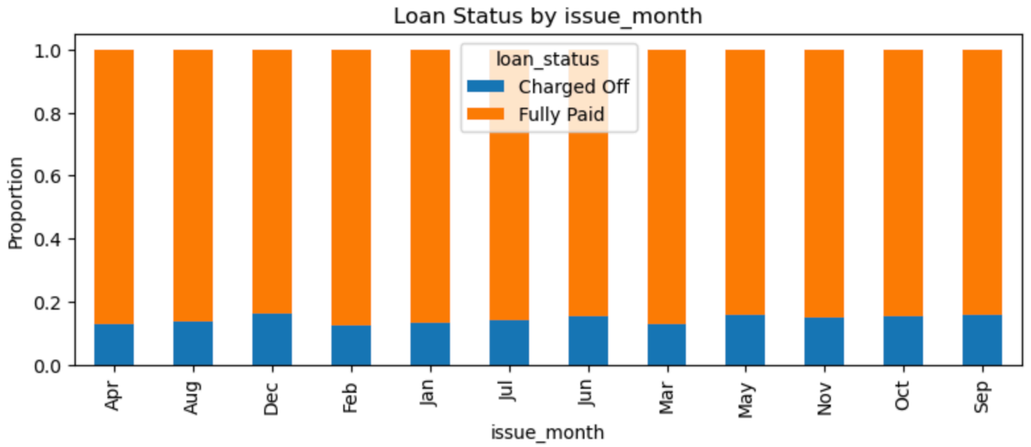


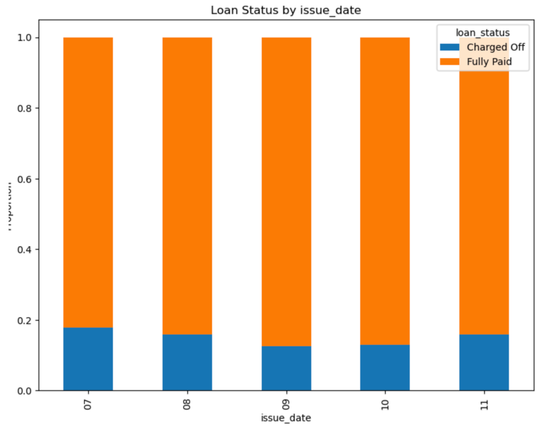
	Charged Off	Fully Paid	Total
Not Verified	0.128	1.0	1.128
Source Verified	0.148	1.0	1.148
Verified	0.168	1.0	1.168

The possibility of borrowers' charging off is "Verified"> "Source Verified" > "Not Verified".

2. Data Analyze and Insights

2.2.4. The charging off possibility has no relation with the month whereas is influenced by the date.





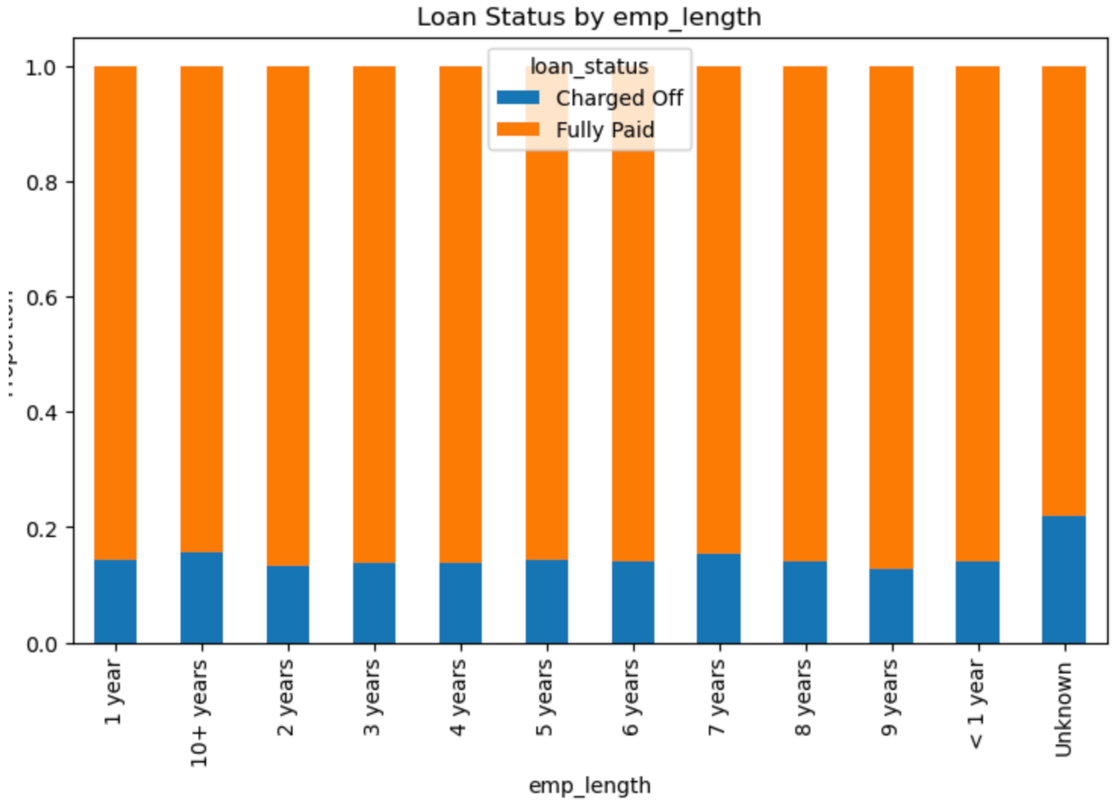
	Charged Off	Fully Paid	Total
07	0.179	1.0	1.179
08	0.158	1.0	1.158
09	0.126	1.0	1.126
10	0.129	1.0	1.129
11	0.159	1.0	1.159

	Charged Off	Fully Paid	Total
Apr	0.131	1.0	1.131
Aug	0.138	1.0	1.138
Dec	0.161	1.0	1.161
Feb	0.123	1.0	1.123
Jan	0.135	1.0	1.135
Jul	0.143	1.0	1.143
Jun	0.152	1.0	1.152
Mar	0.129	1.0	1.129
May	0.16	1.0	1.16
Nov	0.149	1.0	1.149
Oct	0.154	1.0	1.154
Sep	0.156	1.0	1.156

The borrowers whose loan was funded on the 7th of each month have the most possibility to charge off.

2. Data Analyze and Insights

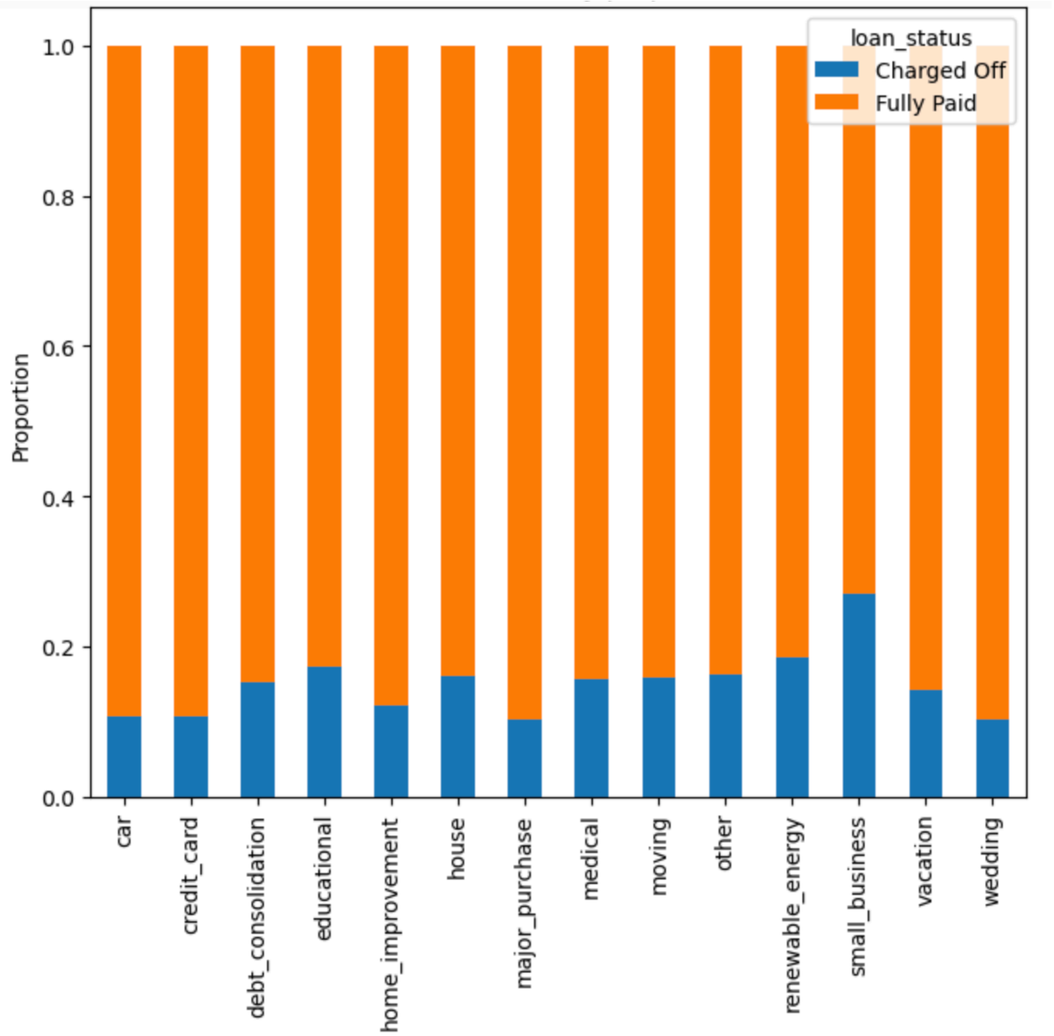
2.2.5. Employment length in years doesn’t influence the loan status



	Charged Off	Fully Paid	Total
1 year	0.144	1.0	1.144
10+ years	0.157	1.0	1.157
2 years	0.132	1.0	1.132
3 years	0.138	1.0	1.138
4 years	0.138	1.0	1.138
5 years	0.143	1.0	1.143
6 years	0.142	1.0	1.142
7 years	0.154	1.0	1.154
8 years	0.141	1.0	1.141
9 years	0.129	1.0	1.129
< 1 year	0.142	1.0	1.142
Unknown	0.221	1.0	1.221

2. Data Analyze and Insights

2.2.6. Some borrowing purposes have a significant high possibility of charging of



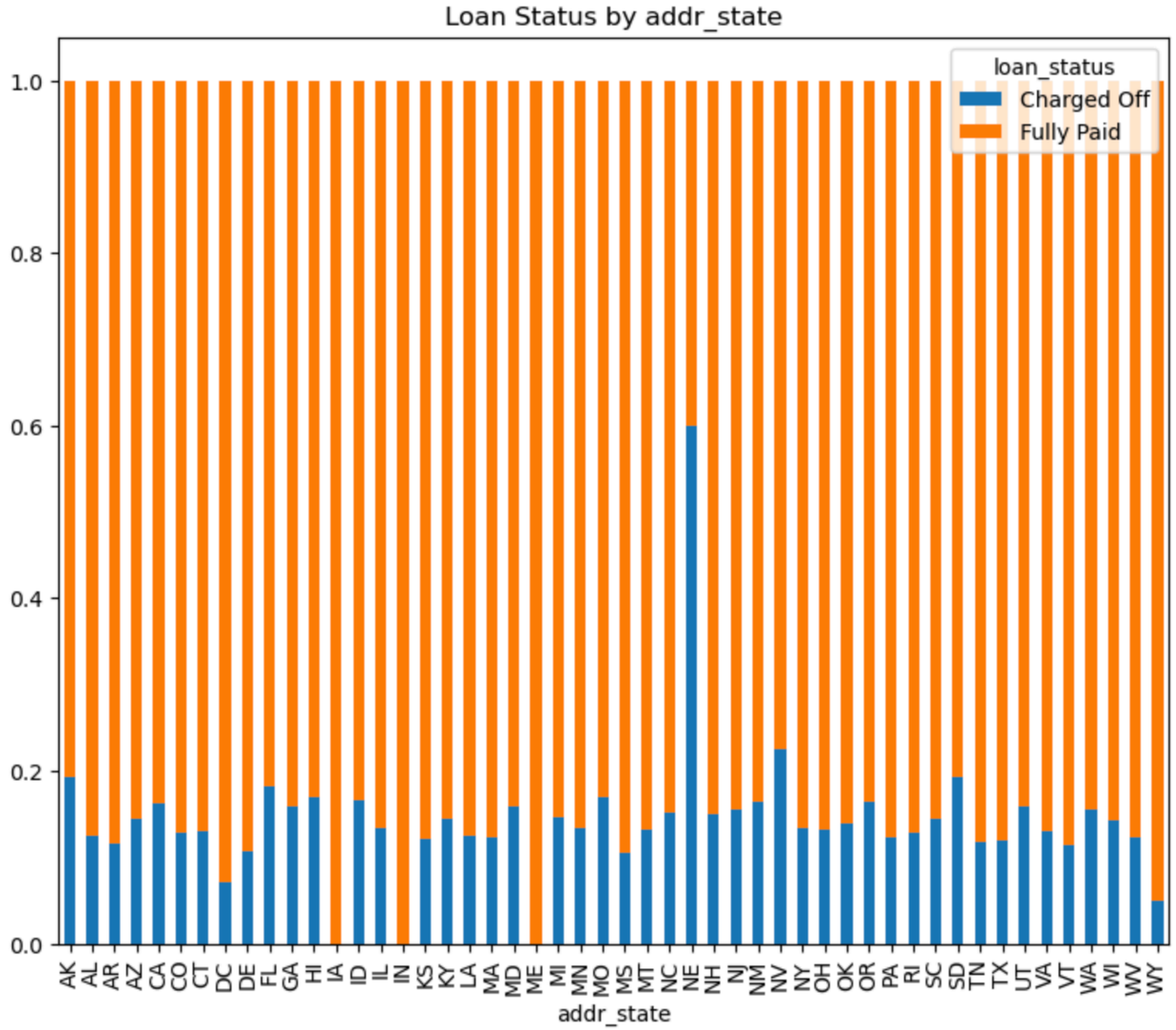
purpose

	Charged Off	Fully Paid	Total
car	0.107	1.0	1.107
credit_card	0.108	1.0	1.108
debt_consolidation	0.153	1.0	1.153
educational	0.172	1.0	1.172
home_improvement	0.121	1.0	1.121
house	0.161	1.0	1.161
major_purchase	0.103	1.0	1.103
medical	0.156	1.0	1.156
moving	0.16	1.0	1.16
other	0.164	1.0	1.164
renewable_energy	0.186	1.0	1.186
small_business	0.271	1.0	1.271
vacation	0.141	1.0	1.141
wedding	0.104	1.0	1.104

"Small_business" has much higher possibility to charge off compare to other purpose.

2. Data Analyze and Insights

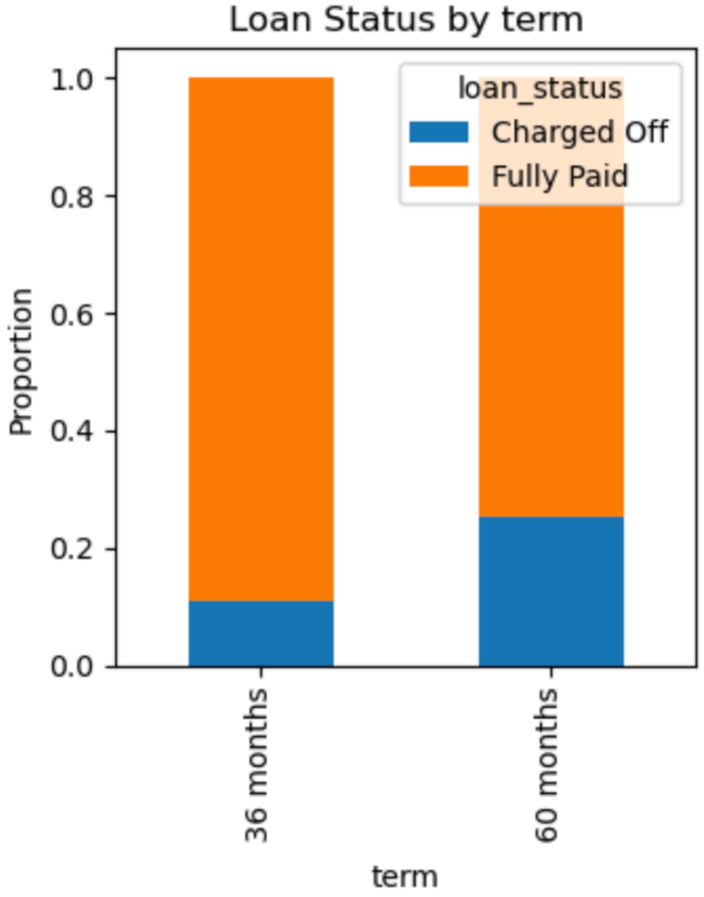
2.2.7. The borrowers from some states are more willing to fully pay, whereas some have high risk of charge off



“NE” has a VERY HIGH risk for charging off.
“IA”, “IN”, and “ME” are relatively safe place to loan to borrowers.

2. Data Analyze and Insights

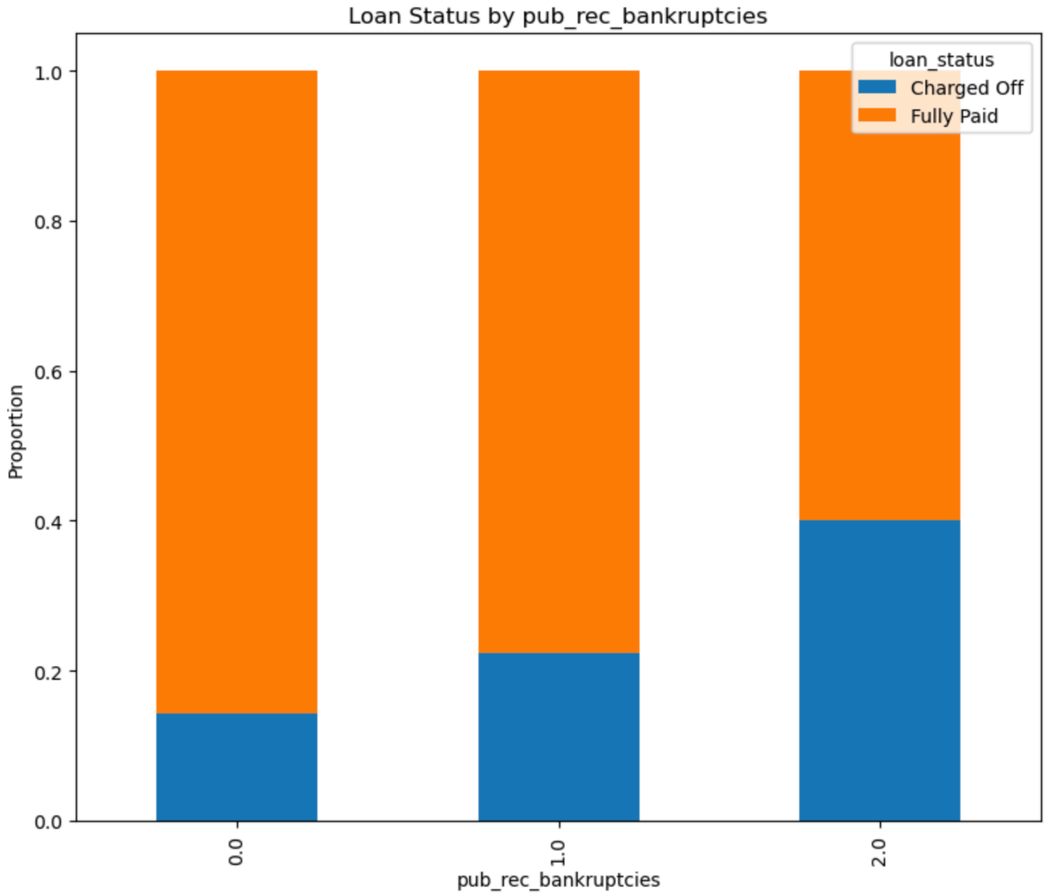
2.2.8. Loans with longer term are more likely to charge off



	Charged Off	Fully Paid	Total
36 months	0.111	1.0	1.111
60 months	0.253	1.0	1.253

2. Data Analyze and Insights

2.2.9. The more public record bankruptcies, the more possibility to charge off

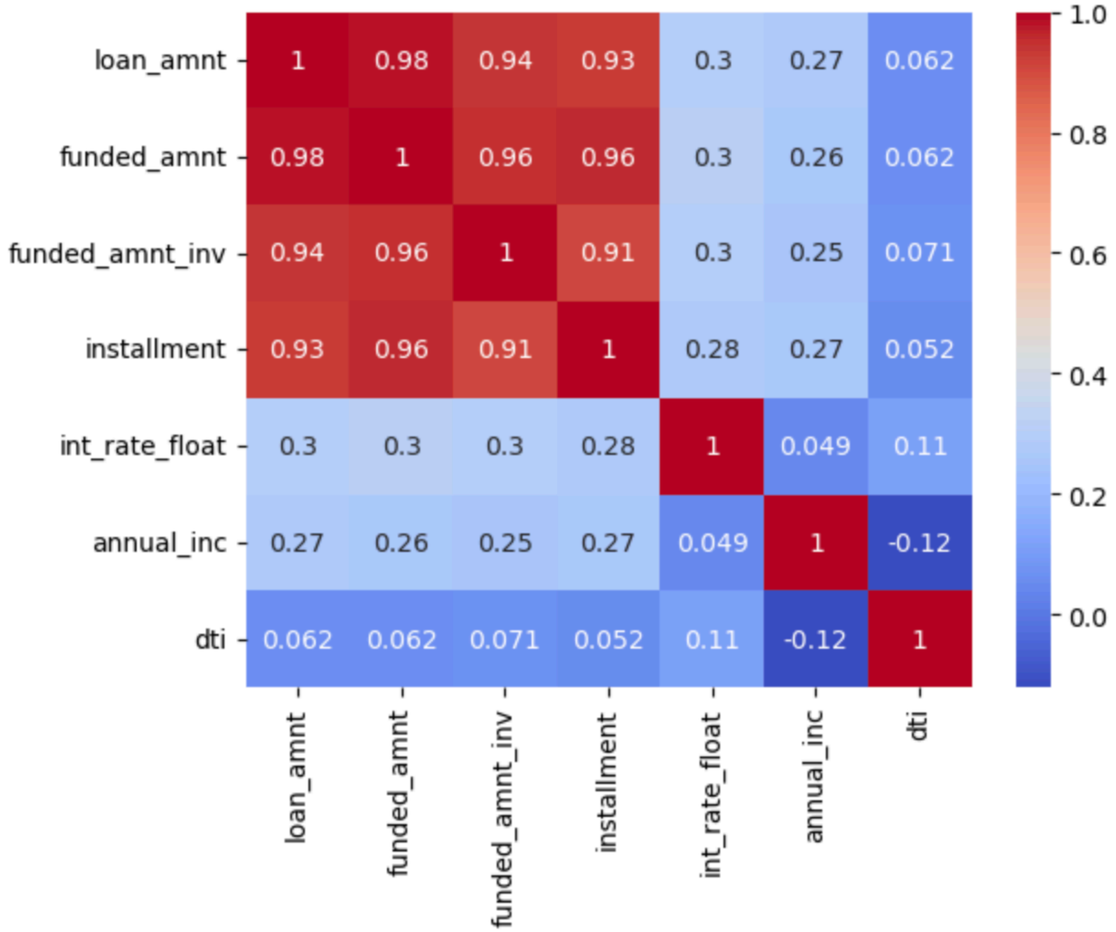


	Charged Off	Fully Paid	Total
0.0	0.142	1.0	1.142
1.0	0.224	1.0	1.224
2.0	0.4	1.0	1.4

The possibility of borrowers' charging off is:
"Public record bankruptcies of twice"
> "Public record bankruptcies of once"
> "No public record bankruptcies".

2. Data Analyze and Insights

2.3.1. Analyze the correlation between “loan_status” and all the Continuous Variables

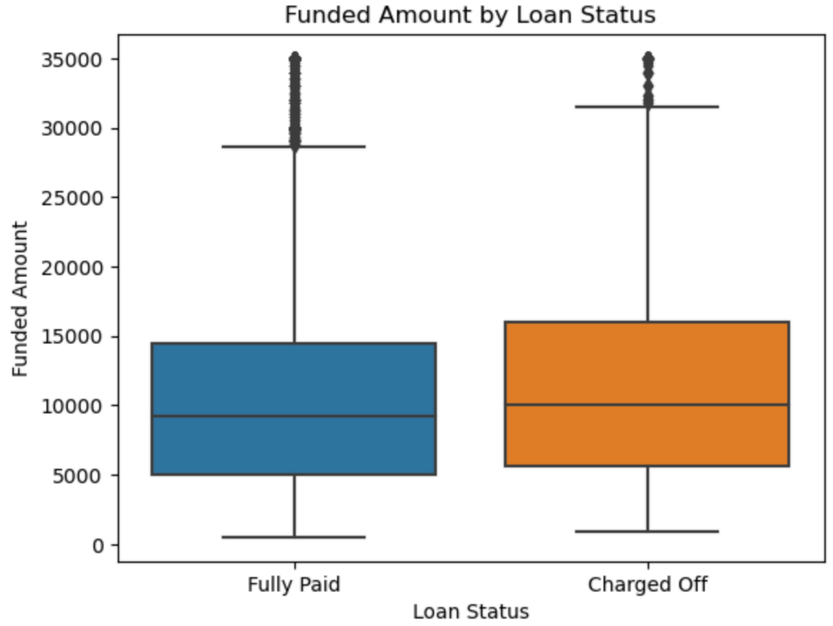


- “loan_amnt”, “funded_amnt”, “funded_amnt_inv”, and “installment” have a high correlation with each other.
- No need to analyze the data with high correlation since they contain similar features.
- Picking anyone of the 4 to analyze their influence on 'loan_status'.

let’s choose ‘funded_amnt' for the following analysis!

2. Data Analyze and Insights

2.3.2. Less ‘loan_amnt’/‘funded_amnt'/'funded_amnt_inv'/ 'installment' indicate more willing to 'Fully paid'

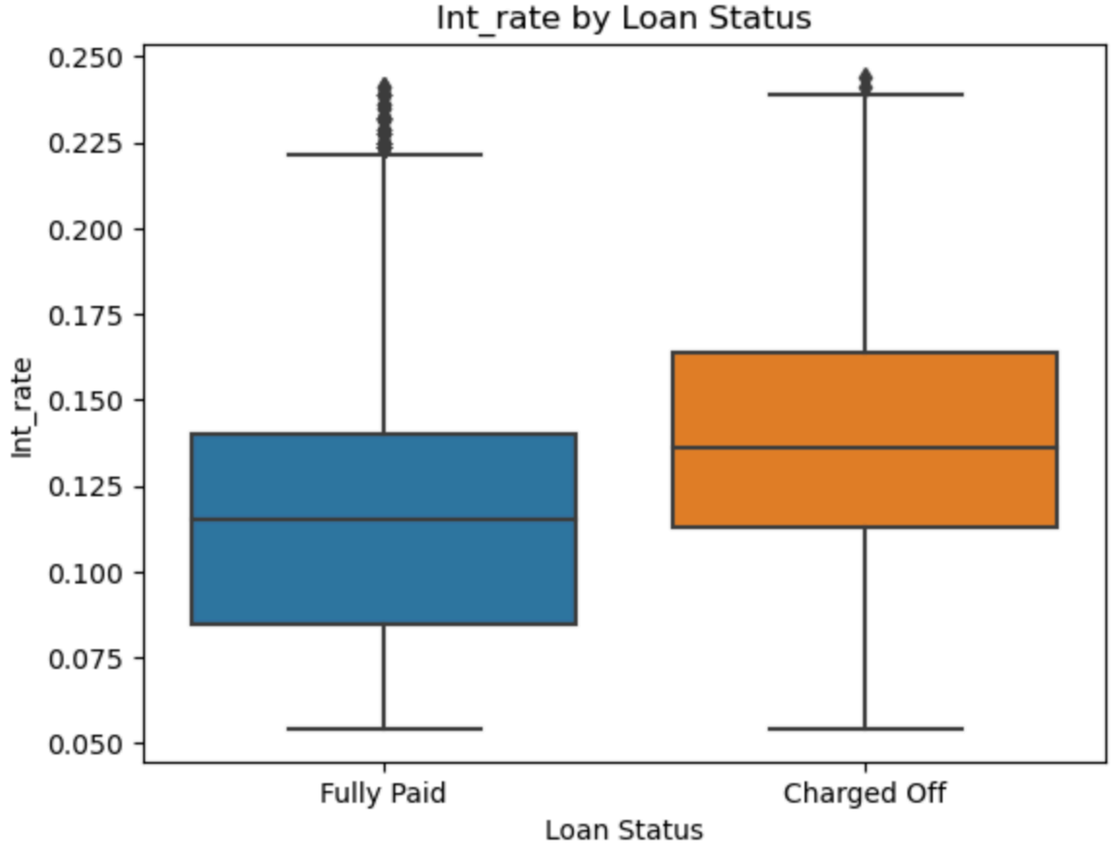


Data Type	Fully paid	Charged off	Charged off/Full paid -1
Mean	10618.5	11753.4	10.7%
3/4 quantile	14500	16000	10.3%
Medium	9200	10000	8.7%
1/4 quantile	5050	5575	10.3%

It indicates that the borrowers with more loans applied/amount committed to that loan/ amount committed by investors for that loan/monthly payment owed have higher possibility to charge off.

2. Data Analyze and Insights

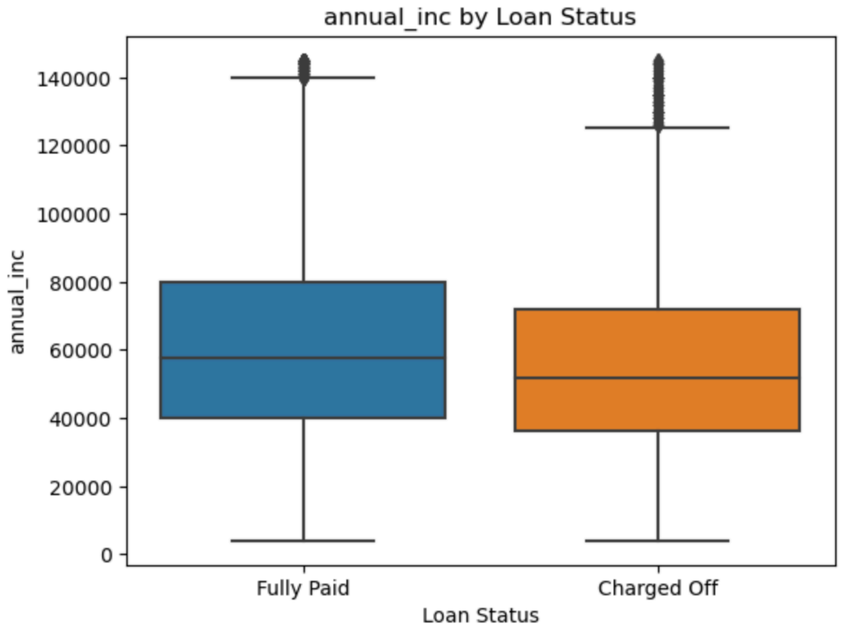
2.3.3. The higher “int_rate”, the higher possible of “Charged off”



Data Type	Fully paid	Charged off
Mean	11.6%	13.8%
3/4 quantile	14.0%	16.4%
Medium	11.5%	13.6%
1/4 quantile	8.5%	11.3%

2. Data Analyze and Insights

2.3.4. The less “annual_inc”, the more possiblility of “Charged off”

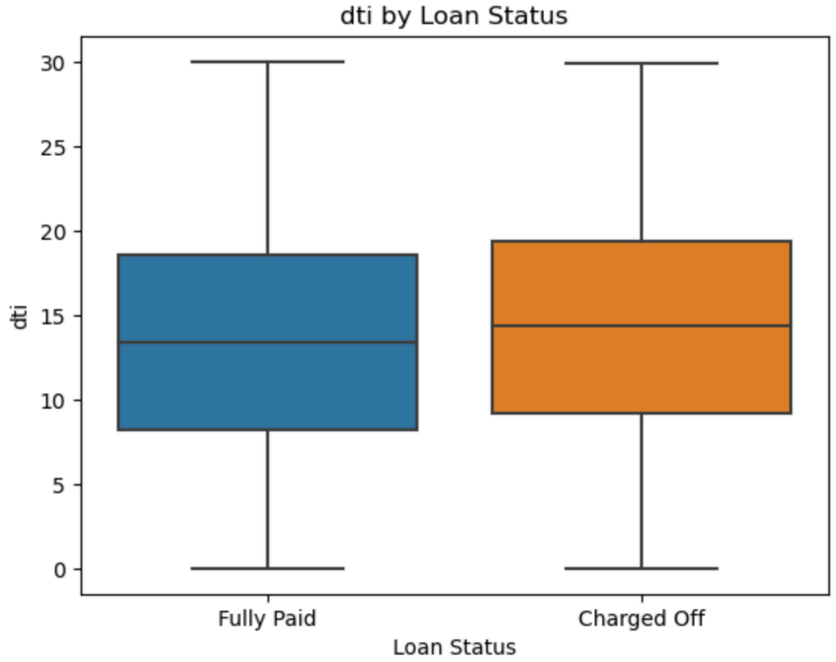


Data Type	Fully paid	Charged off	Full paid/Charged off -1
Mean	62008.7	56027.3	10.7%
3/4 quantile	78000	70374	10.8%
Medium	57000	51600	10.5%
1/4 quantile	40000	36071	10.9%

For data on borrowers’ annual income, there are many outliers that are significantly higher than others.
We used “Tukey's method” to identify outliers on the IQR and eliminate the outliers before analysis.

2. Data Analyze and Insights

2.3.5. Higher “dti” has a slightly positive influence on the possibility of “Charged off”



Data Type	Fully paid	Charged off	Charged off/Full paid -1
Mean	13.32	14.10	5.9%
3/4 quantile	18.58	19.40	4.4%
Medium	13.42	14.40	7.3%
1/4 quantile	8.19	9.18	12.1%

The ratio of total monthly debt payments to the reported monthly income (DTI) has only a small influence on the possibility of “Charged off”.

Thank you!