Lending Club Case Study

HTTPS://GITHUB.COM/ELCOMMSELVA/LENDINGCLUBCASESTUDY.GIT

Rajamanickam K
raja_anr7@outlook.com
Selvakumar Karuppannan
elcommselva@gmail.com

Lending Club Case Study

Problem Statement, How & What

Business Problem:

Identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

How:

Understand how “consumer attributes” and “loan attributes” influence the tendency of default

- Univariate and segmented univariate analysis

- Business-driven, type-driven and data-driven metrics are created for the important variables and utilized for analysis

- Bivariate analysis

- Appropriate plots

What:

The lending club wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

The analysis shall successfully identify at least the 5 important driver variables (i.e. variables which are strong indicators of default).

Lending Club Case Study

Approach

- Data import
- Data understanding

- Data Cleaning/Data Quality & Fix
- Data Content Analysis

- Summary of EDA
- Conclusion

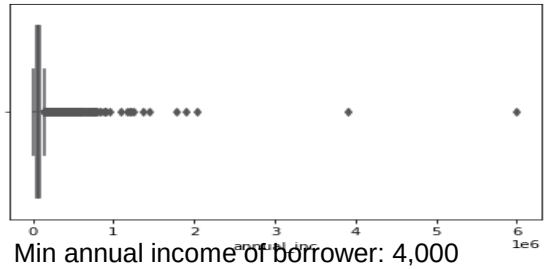
Lending Club Case Study

Data Overview – Observations on the Data Quality

Data Description:
The Data contains the complete loan data for all loans issued through the time period 2007 to 2011. And the data dictionary which describes the meaning of these variables.

View of data before cleaning

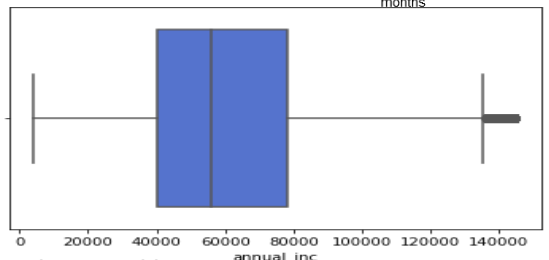
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_past
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	
39714	90395	90390	5000	5000	1325.0	36 months	8.07%	156.84	A	A4	...	NaN	
39715	90376	89243	5000	5000	650.0	36 months	7.43%	155.38	A	A2	...	NaN	
39716	87023	86999	7500	7500	800.0	36 months	13.75%	255.43	E	E2	...	NaN	



Min annual income of borrower: 4,000
Max annual income of borrower: 60,00,000

View of data after cleaning

	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_length	home_ownership	...	loan_status	purpos
0	5000	5000	4975.00	36 months	10.65	162.87	B	B2	10	RENT	...	Fully Paid	credit_ca
1	2500	2500	2500.00	60 months	15.27	59.83	C	C4	1	RENT	...	Charged Off	c
2	2400	2400	2400.00	36 months	15.96	84.33	C	C5	10	RENT	...	Fully Paid	small_busines
39713	8500	8500	875.00	36 months	10.28	275.38	C	C1	3	RENT	...	Fully Paid	credit_ca
39714	5000	5000	1325.00	36 months	8.07	156.84	A	A4	1	MORTGAGE	...	Fully Paid	debt_consolidati
39716	7500	7500	800.00	36 months	13.75	255.43	E	E2	1	OWN	...	Fully Paid	debt_consolidati



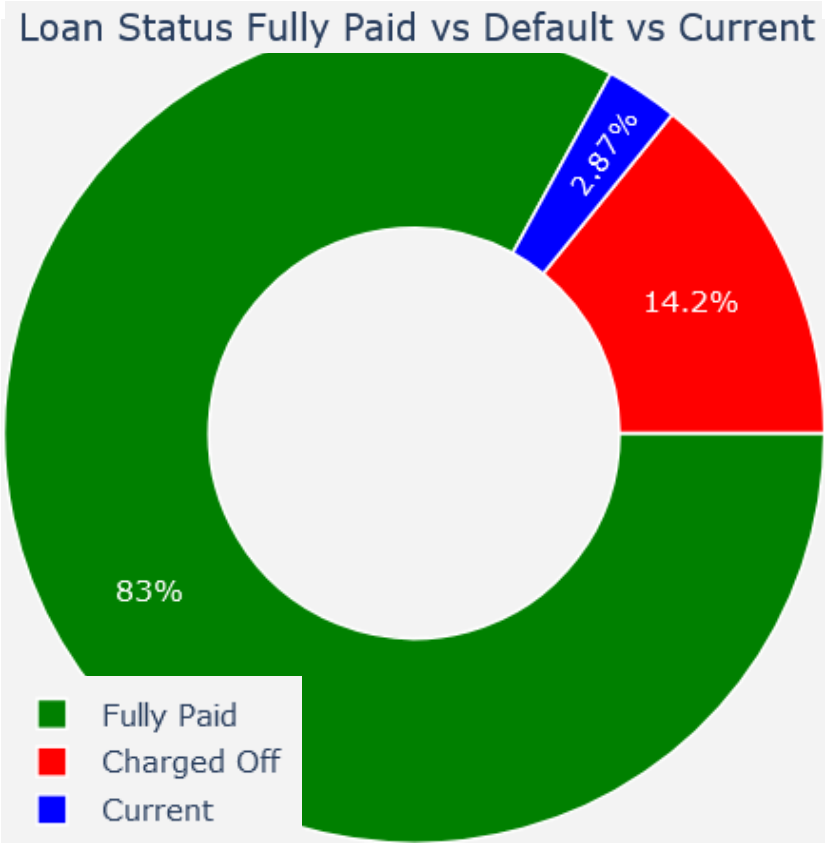
Min annual income of borrower: 4,000
Max annual income of borrower: 1,45,000

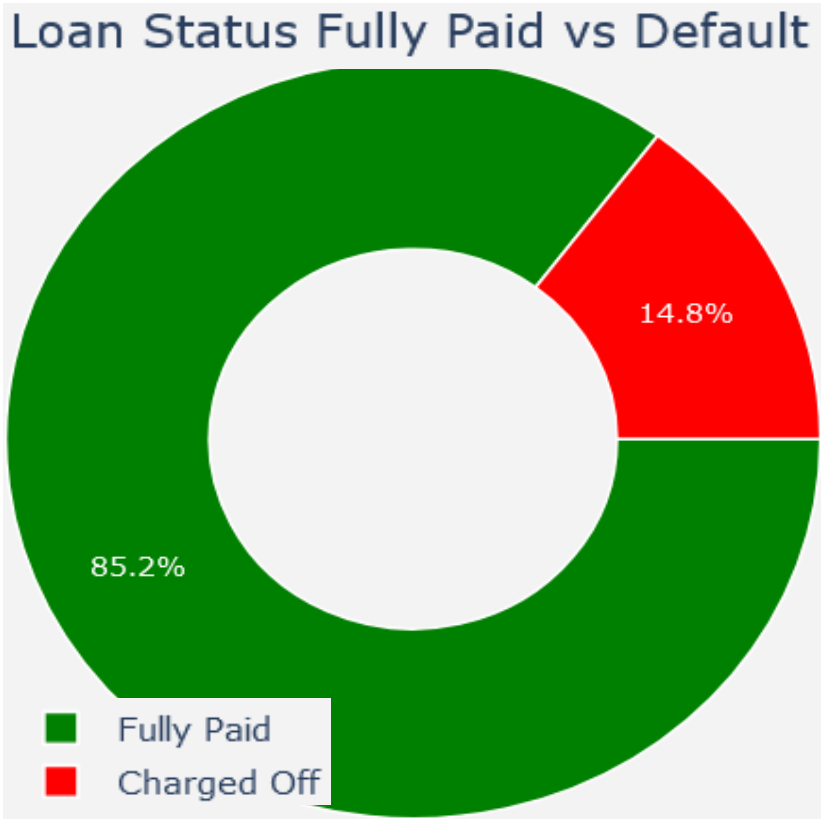
Dataset	Rows	Columns
Before Cleaning	39717	111
After Cleaning	36815	23

Insights:
Dataset had redundancies, null values, missing values and outliers. And they were removed or treated using appropriate methods for the analysis

Lending Club Case Study

Data Content Analysis – Univariate



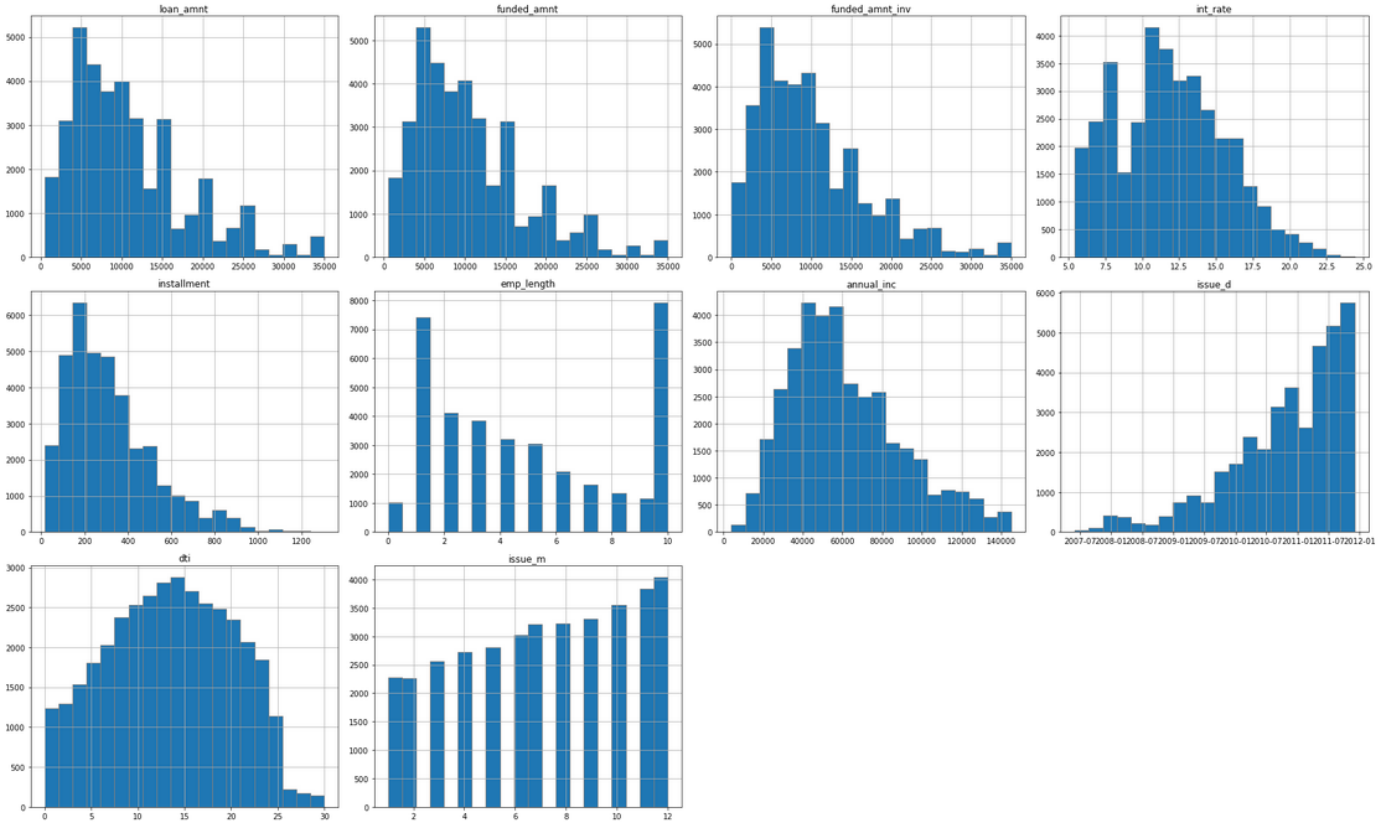


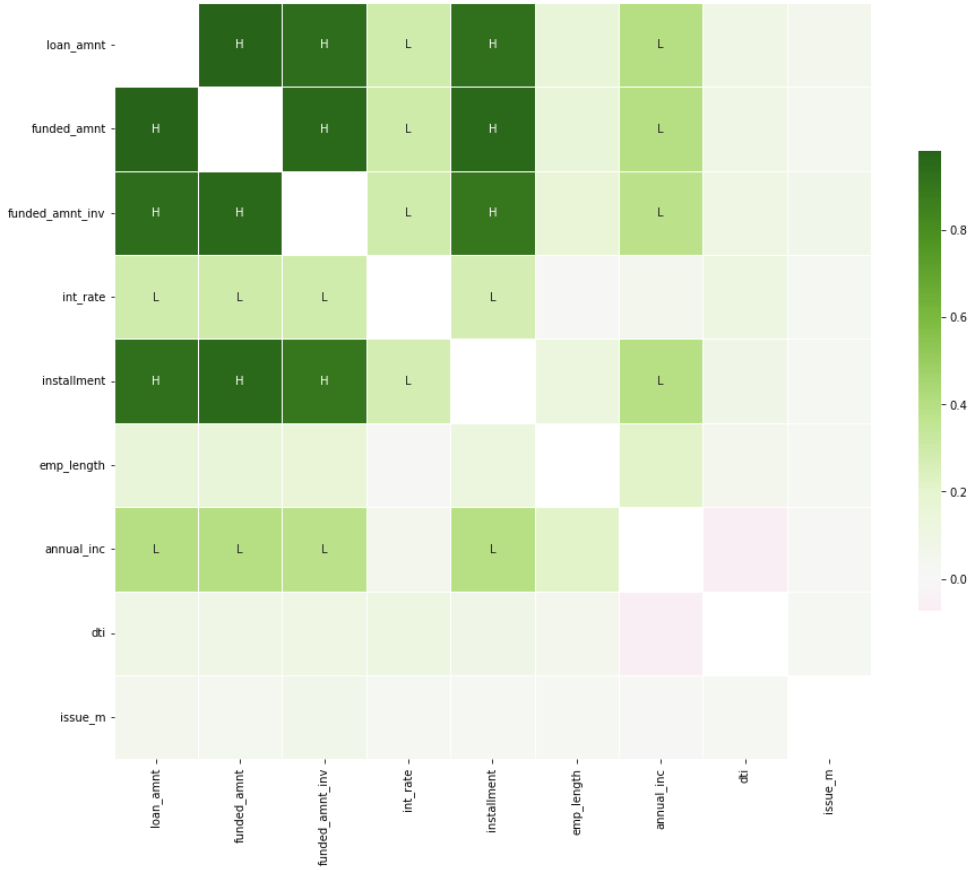
Insights:

14% of the loans lended were charged-off or defaulted in the past. Loans that are in progress ‘current’ are not considered for further analysis.

Lending Club Case Study

Data Content Analysis – Univariate & Correlation





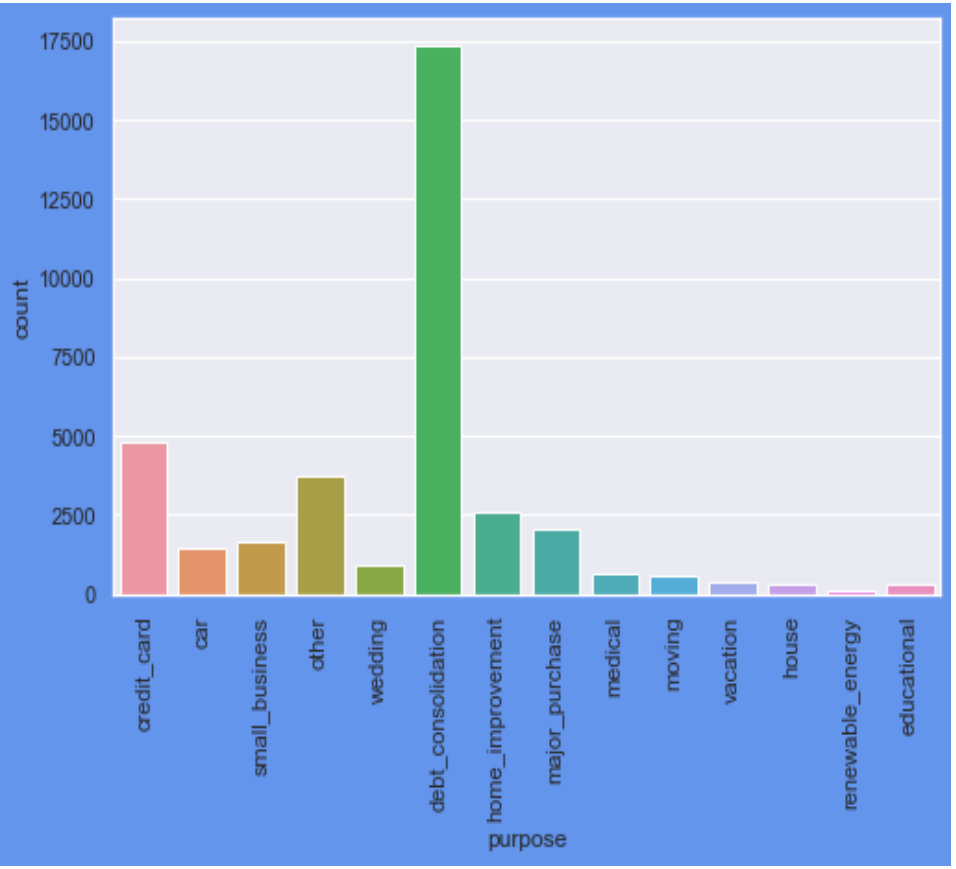
Insights:

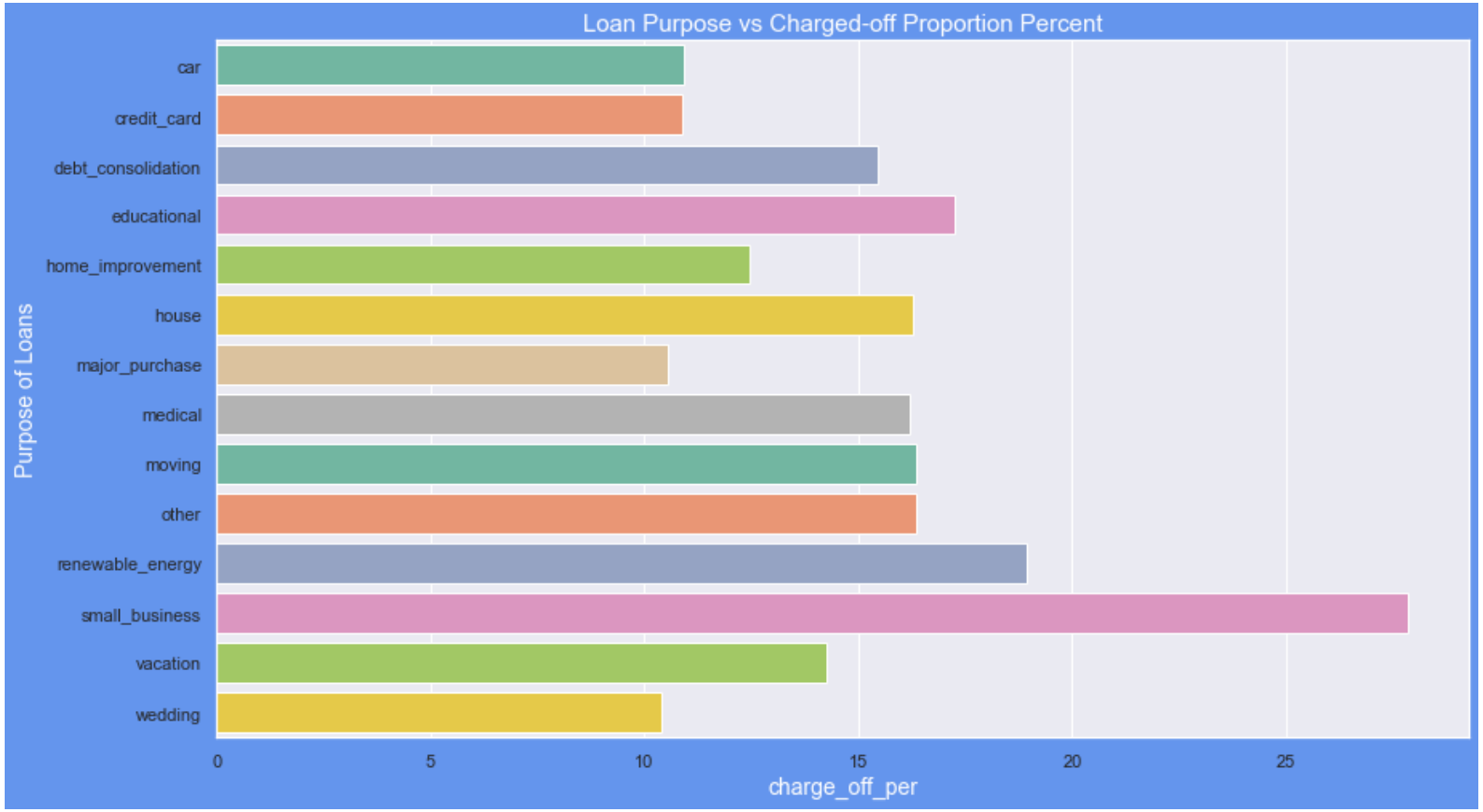
loan amount, funded amount, funded amount_inv are highly corelated. Further analysis was based on loan_amnt

No other correlation observed on the numerical variables

Lending Club Case Study

Data Content Analysis – Bivariate



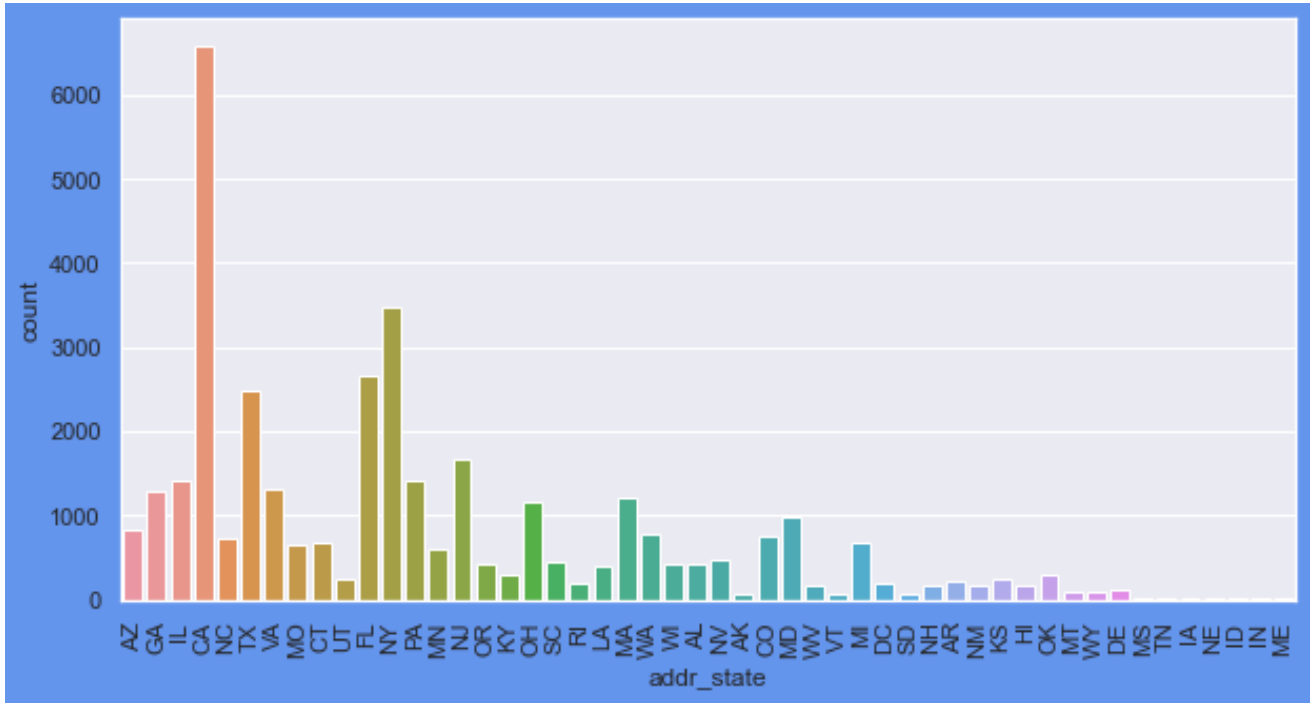


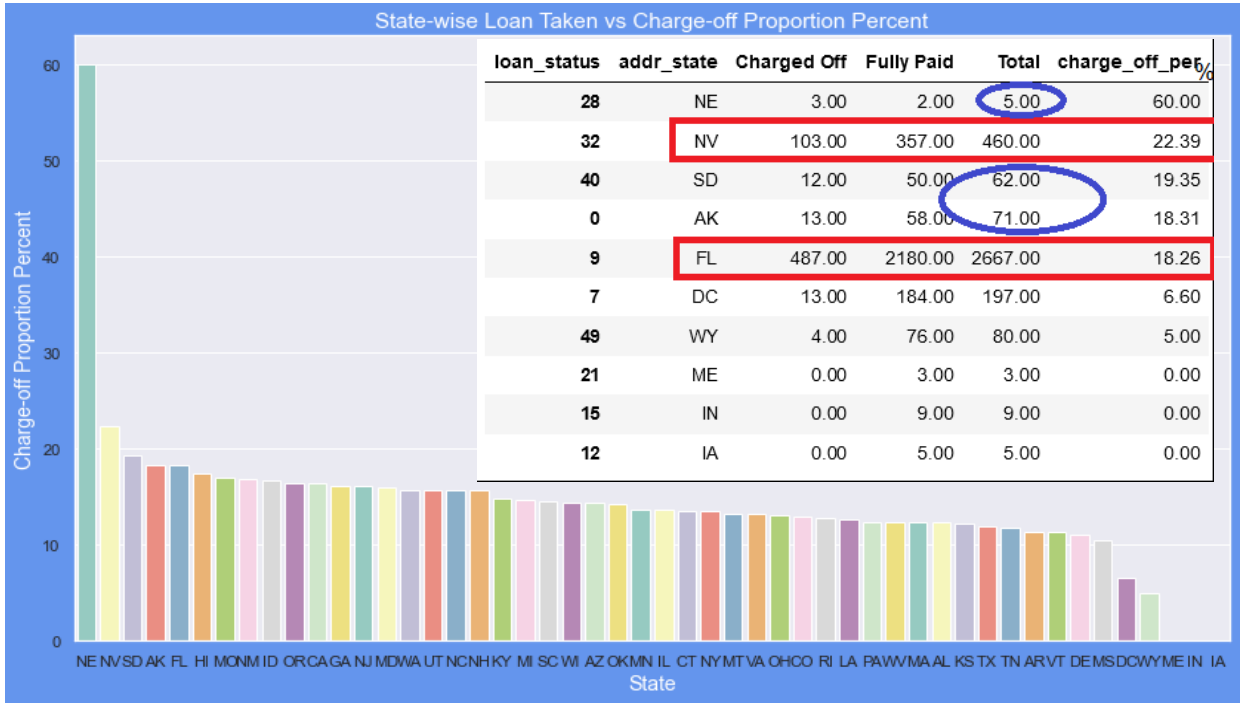
Insights:

Loans procured for debt consolidation purpose is the most loan lended. However the higher default/charged-off loans were taken for small_business

Lending Club Case Study

Data Content Analysis – Bivariate





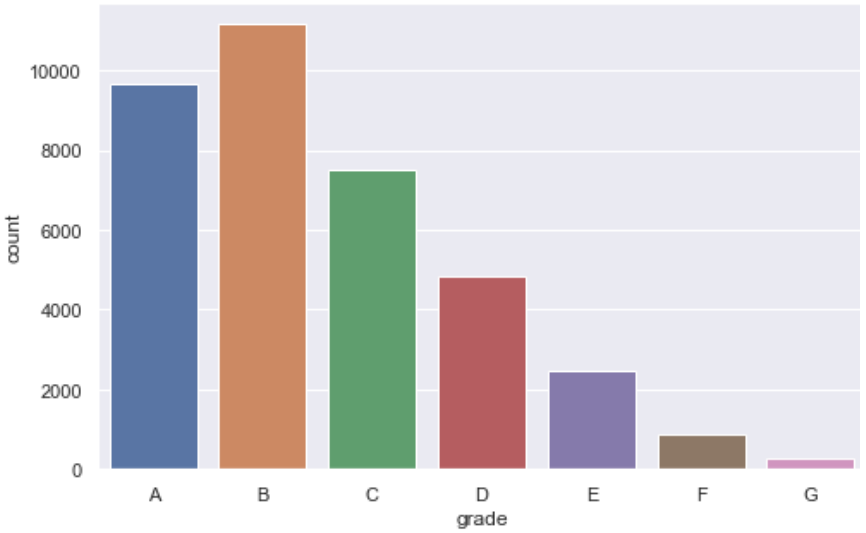
loan_status	addr_state	Charged Off	Fully Paid	Total	charge_off_per%
28	NE	3.00	2.00	5.00	60.00
32	NV	103.00	357.00	460.00	22.39
40	SD	12.00	50.00	62.00	19.35
0	AK	13.00	58.00	71.00	18.31
9	FL	487.00	2180.00	2667.00	18.26
7	DC	13.00	184.00	197.00	6.60
49	WY	4.00	76.00	80.00	5.00
21	ME	0.00	3.00	3.00	0.00
15	IN	0.00	9.00	9.00	0.00
12	IA	0.00	5.00	5.00	0.00

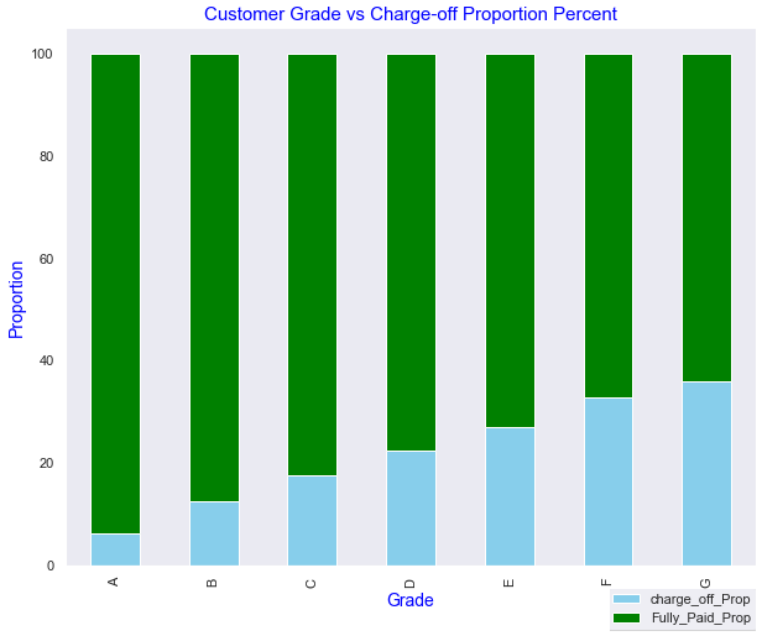
Insights:

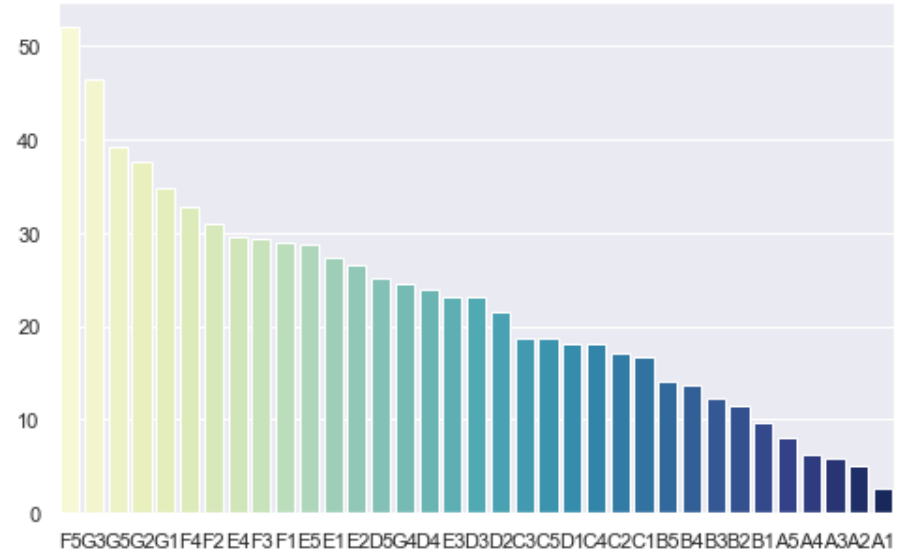
Most Loans procured were from CA state. NE state has very high chances of charged off but number of applications are too low to make any decisions. FL & NV states shows good number of charged offs in good number of applications.

Lending Club Case Study

Data Content Analysis – Bivariate





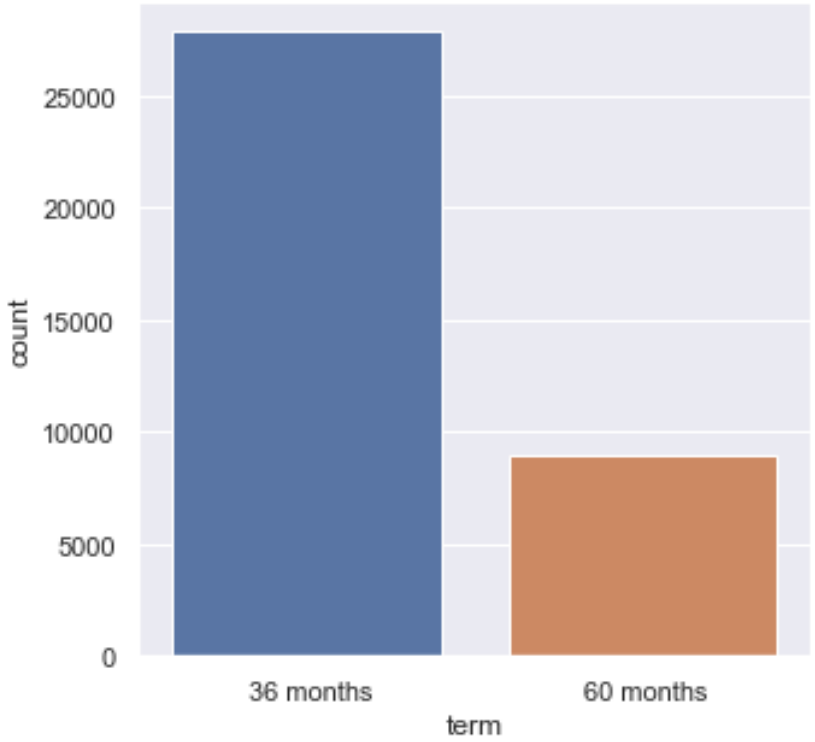


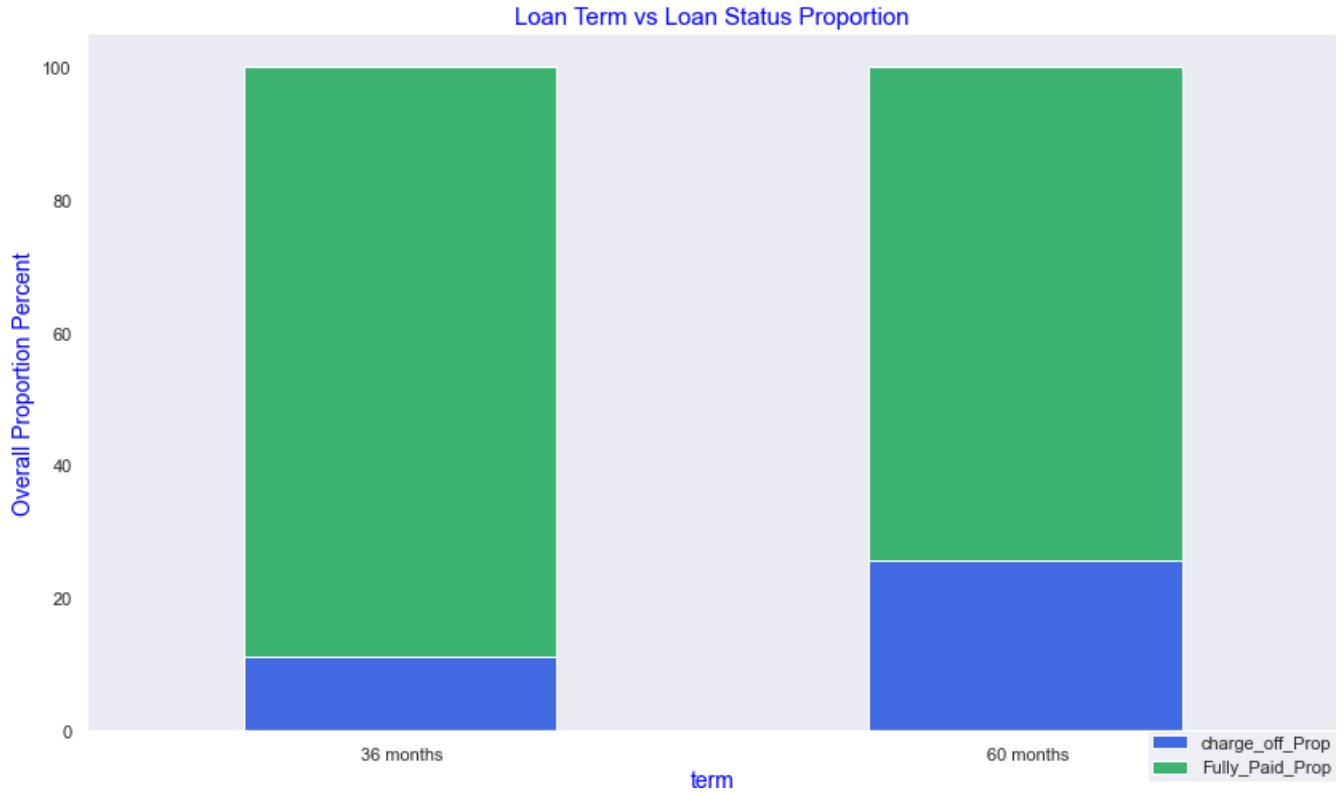
Insights:

Most borrowers are categorized under Grade B followed by Grade A & C. Customers with Grade F & G are high risk customers and there are very high chances of charge off. Customers with SubGrade F5 are high risk customers and are prone to default a loan followed by G3,G2

Lending Club Case Study

Data Content Analysis – Bivariate



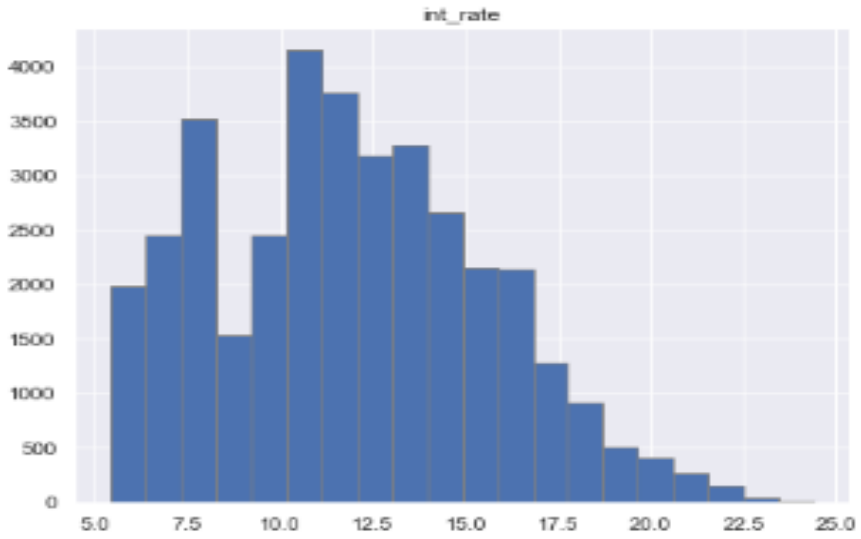


Insights:

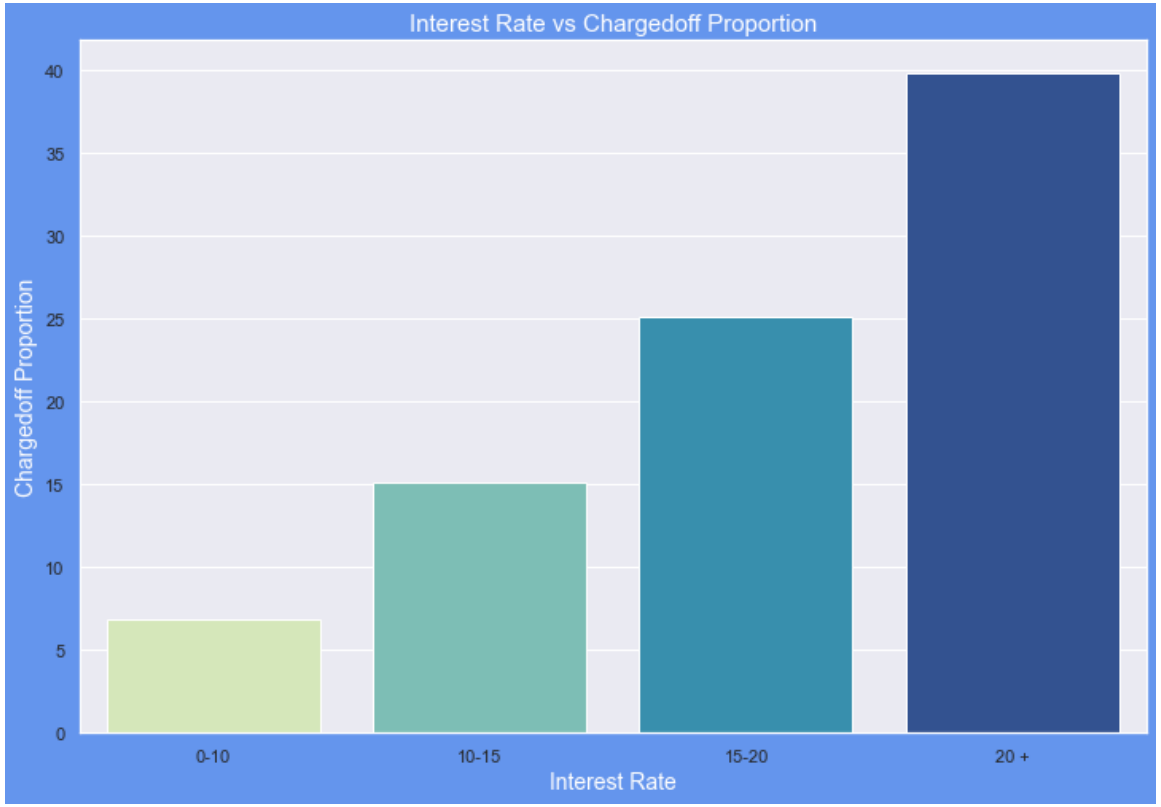
36 months seems to be preferred loan term for most borrowers. Long loan tenure has high Charge-off/default rates

Lending Club Case Study

Data Content Analysis – Bivariate



From the data, we infer that there are many distinct interest rates. Hence to check the influence of interest rates on Loan Status we grouped the interest rate



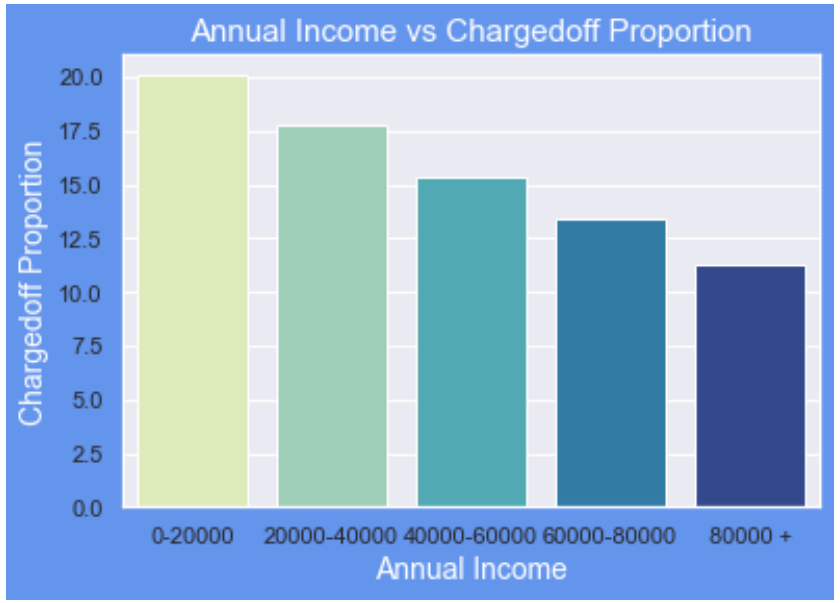
Insights:

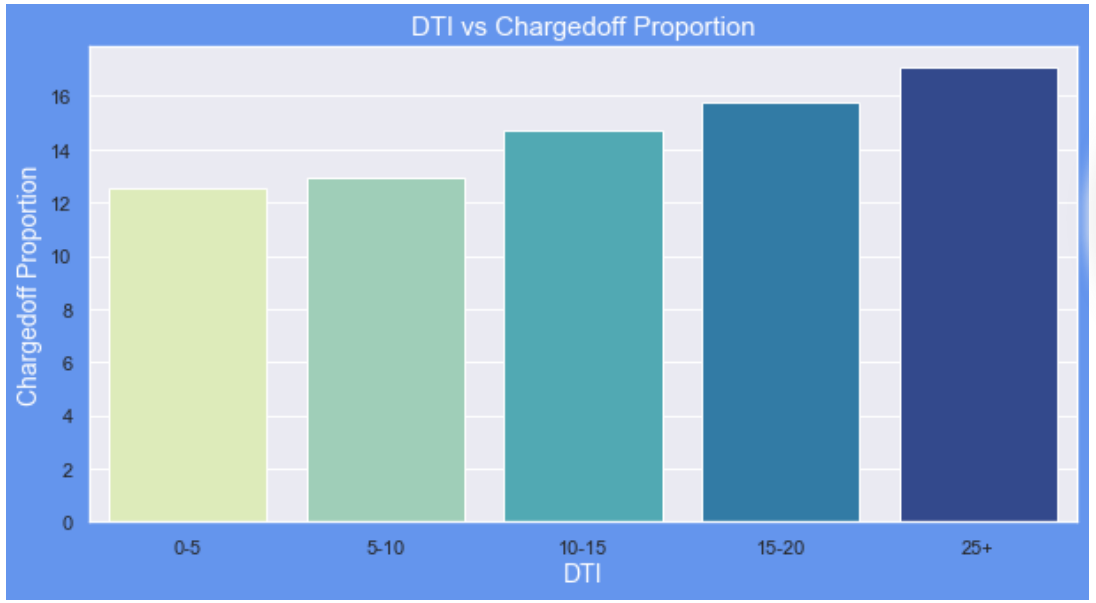
- interest rate less than 10% has very less chances of charged off. Interest rates are starting from minimum 5 %

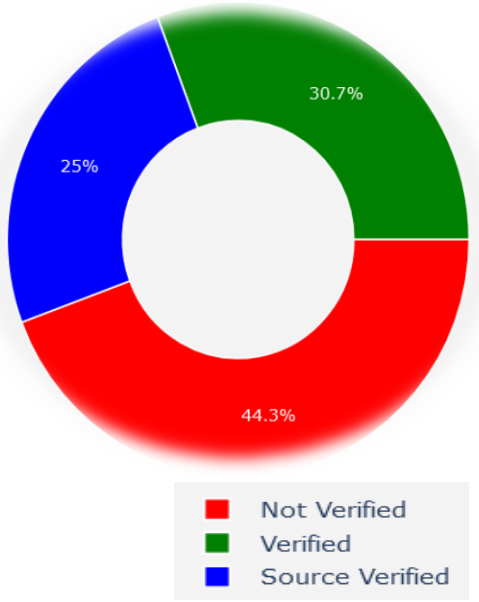
- interest rate more than 15% has good chance of charged off as compared to other category intrest rates

Lending Club Case Study

Data Content Analysis – Bivariate



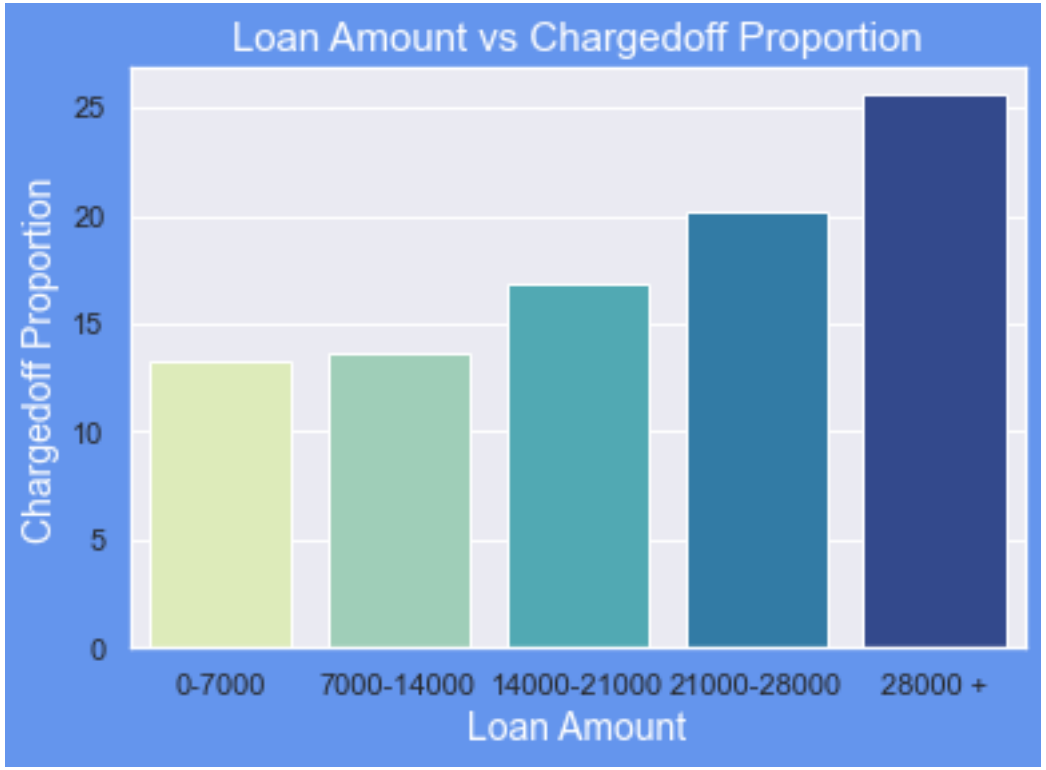


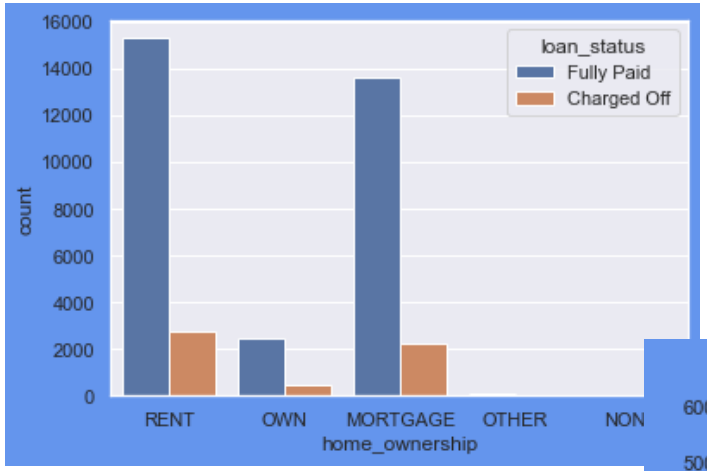


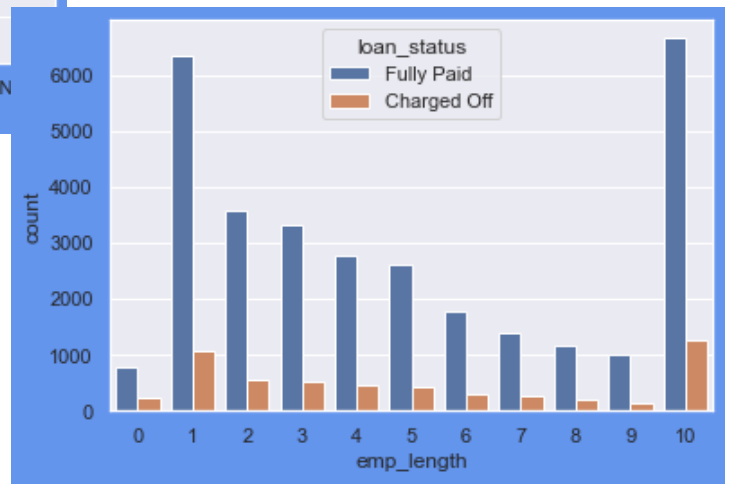
- **Income range 0-20,000 has high chances of charged off. Notice that with increase in annual income charged off proportion got decreased**
- **DTI (debt to Income) is directly proportional to the Charge off. The more debt to income, the more chance of Defaulting.**
- **Many sources are Not verified**
- **Suggestion: More Verification can to be done by the lending club**

Lending Club Case Study

Data Content Analysis – Bivariate





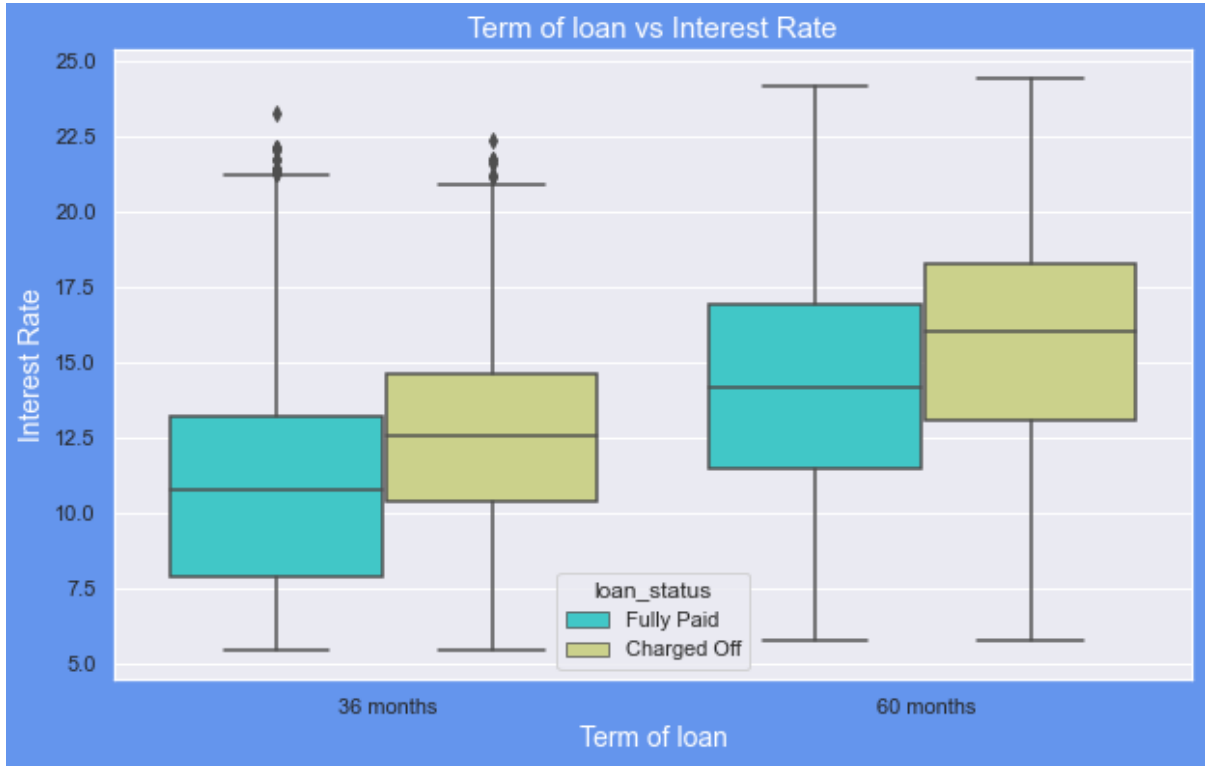


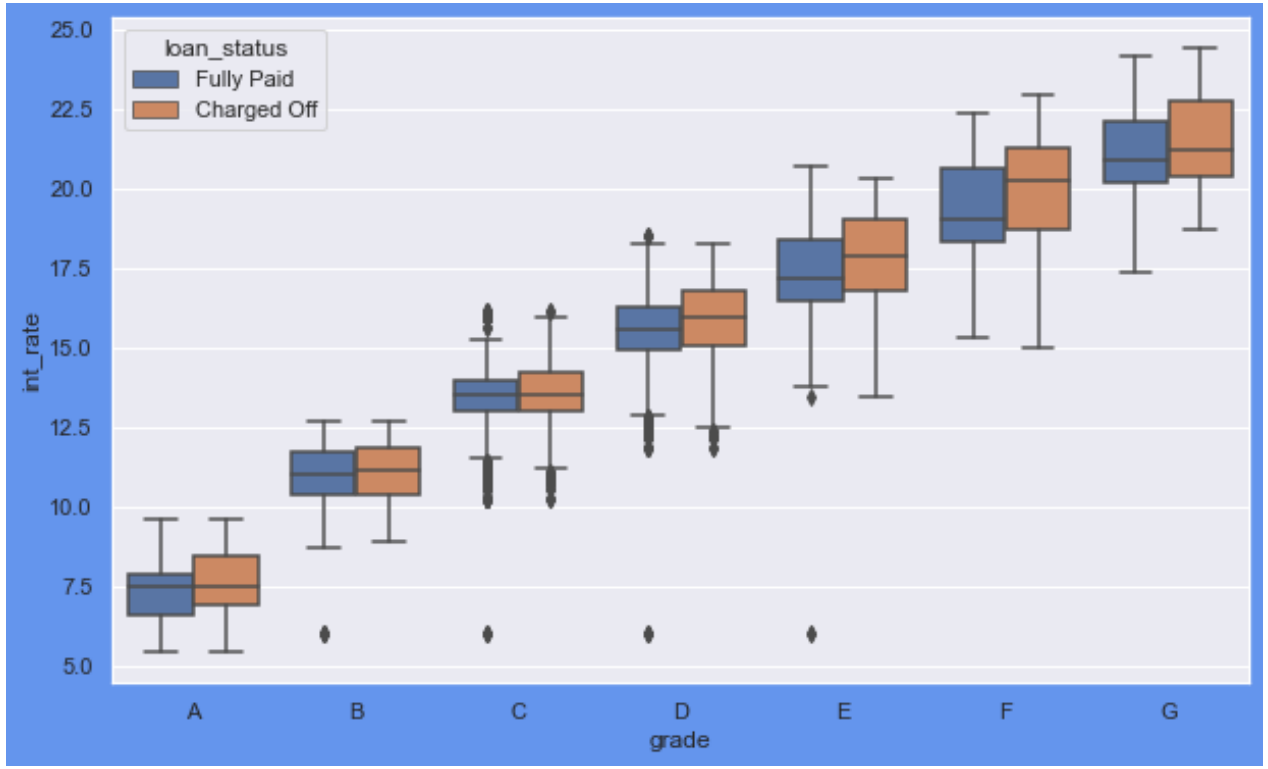
Insights:

- **Higher Loan amount has more chances of charged off**
- **Loan amount having 28000+$ has high chances of charged off**
- **Home ownership and employment length has less influence on defaulting**

Lending Club Case Study

Data Content Analysis – Bivariate





Insights:

It is clear that median interest rate is higher for 60 months loan term and grades like F & G

Most of the loans issued for longer term had higher interest rates and hence has high chances of default.

Lending Club Case Study

Summary

- There were 111 columns and many columns with Null values, redundant values and have been removed for analysis.
- There were categorical types and few have been type casted to required format for analysis.
- Outlier was treated using IQR method on annual_inc (annual income) has been done.
- Formatting of Data and time on issue_d
- No missing values were found in the Target variable **loan_status**
- Data Dictionary was used for the variable meaning. So no separate renaming of variables was done.

loan_amnt, funded amount and fundded amount_inv are highly correlated. So loan_amnt was used for the analysis

Below are the key insights from the EDA:

1. Persons who has taken loan for small business are more prone towards charge-off/default.
2. NV & FL states shows good number of charged offs/defaulters in good number of applications
3. Customers with Grade F & G are high risk customers and there are very high chances of charge off
4. Customers with SubGrade F5 are high risk customers and are prone to default a loan followed by G3,G2 Customers with SubGrade G and F are high risk customers are prone to default.
5. Longer the loan tenure higher the chance of Charge-off/default
6. Interest rate more than 15% has good chance of charged off
7. Income range 0-20,000 has high chances of charged off
8. Loan amount having 28000+$ has high chances of charged off
9. The DTI is directly proportional to the Charge off. The more debt to income, the more chance of Defaulting.

Lending Club Case Study

Conclusion

Below are the strong indicators of Charge off or defaulting based on the EDA on the dataset provided:

1. 'purpose' of the loan - loans procured for "small business" are more prone towards chargeoff/default
2. 'int_rate' interest rate - loans with higher interest rate in combination with 'grade' like F & G (or) - loans with higher interest rate in combination longer duration of term are likely to default
3. 'grade' Grading of the borrower - Customers with Grade F & G are high risk customers and there are very high chances of charge off
4. 'dti' Debt to Income - Debt to Income ratio is directly proporsational to the Charged off or default
5. 'addr_state' State - People from geographical regions state: NV & FL are likely to default more