Lending Club Case Study

Problem statement

- We would be analysing the data collected from a consumer finance company regarding the details of people for which loans were approved in the time period of 2007 to 2011.
- There are situations where people fully pay the loan, are in the process of paying (current), or did not complete the payment on time (default).
- The defaulted applicants cause a major loss in this sector, and so the aim of this study is to analyse the dataset and identify the factors that are responsible that could cause people to default.
- There could be various driving factors to this, likely related to the annual income, total amount to be paid, work experience etc.
- We will be doing univariant and bivariant analysis on the data to come up with relevant information that could be useful for the company for future risk assessment.

I. Data Cleaning

- The contents of the dataset needs to be cleaned up to add/update relevant column names, remove columns which won't be useful in the analysis and having the dataset in the format that would be useful to do proper analysis on.
- Based on the type of data seen from the file, the below fields are being removed, as it doesn't provide any useful info related to our problem statement:
  - **url**: irrelevant for analysis
  - **desc**: irrelevant for analysis
  - **pymnt_plan**: same for all applicants
  - **initial_list_status**: same for all applicants
  - **collections_12_mths_ex_med** - **total_il_high_credit_limit (**rest all columns**)**: these are either same for all or are NA
- We also remove the columns that have more than 20,000 records missing data.
- Some columns are updated with rounding the decimal points to 2 characters.
- It is checked if there are any duplicate records. If so, they will be removed.
- After these steps are completed, we are left with a cleaner, smaller dataset.

II. <u>Univariant Analysis</u>

- The dataset contains various types of columns that could be falling into either of the 2 variables:
    - Ordered: The set of columns that have a particular ordering. Some examples are:
        - emp_length
        - grade
        - sub_grade
    - Unordered: Ones that do not have a particular order to be classified in. Some examples are:
        - home_ownership
        - verification_status
        - purpose

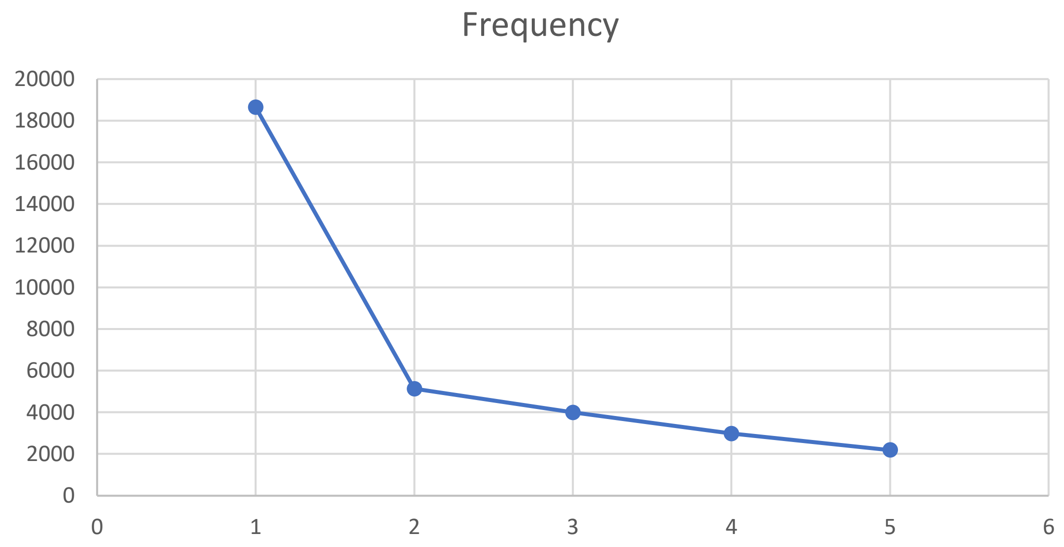
- Apart from these, we also have quantitative variables here. For example:
    - loan_amnt
    - int_rate

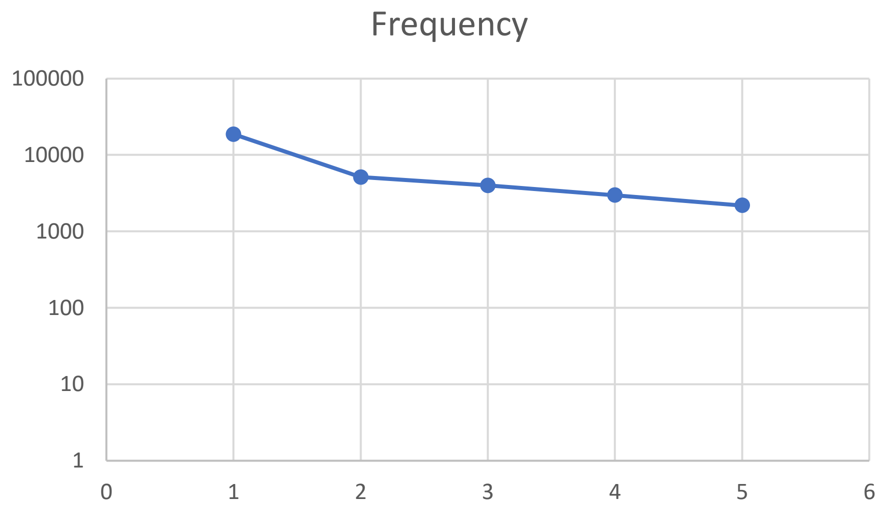
Let’s try to see the frequency plot for some of the unordered sets

Row Labels	Count of home_ownership
MORTGAGE	17659
NONE	3
OTHER	98
OWN	3058
RENT	18899
Grand Total	**39717**

The first plot shows the frequency vs rank scatter plot based on ranking the home ownership of more counts.

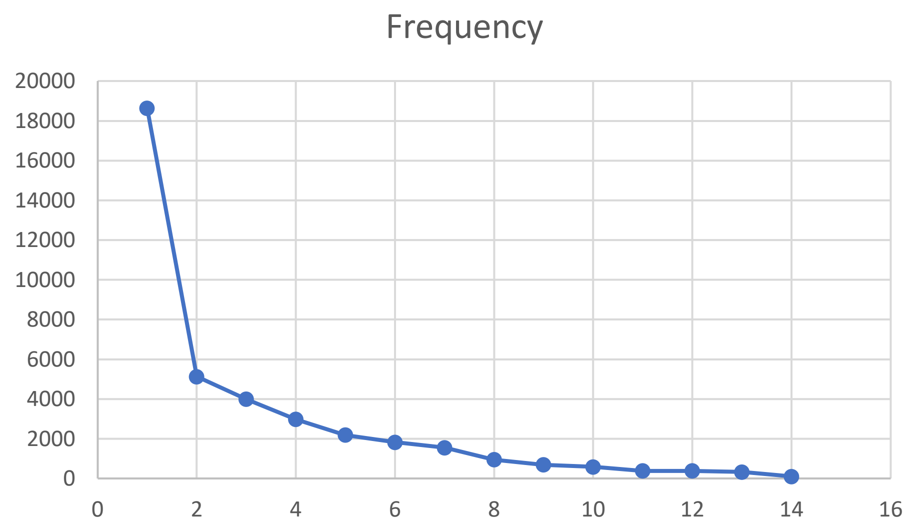
The second plot shows the same distribution but with Y axis in log scale. This shows us that the curve is turning to be a straight line.

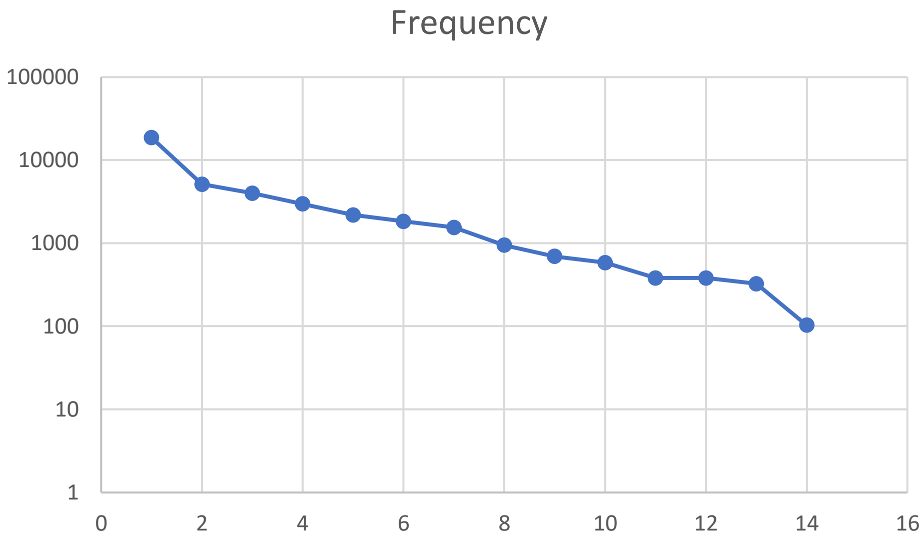




Row Labels	Count of purpose
debt_consolidation	18641
credit_card	5130
other	3993
home_improvement	2976
major_purchase	2187
small_business	1828
car	1549
wedding	947
medical	693
moving	583
vacation	381
house	381
educational	325
renewable_energy	103
Grand Total	**39717**

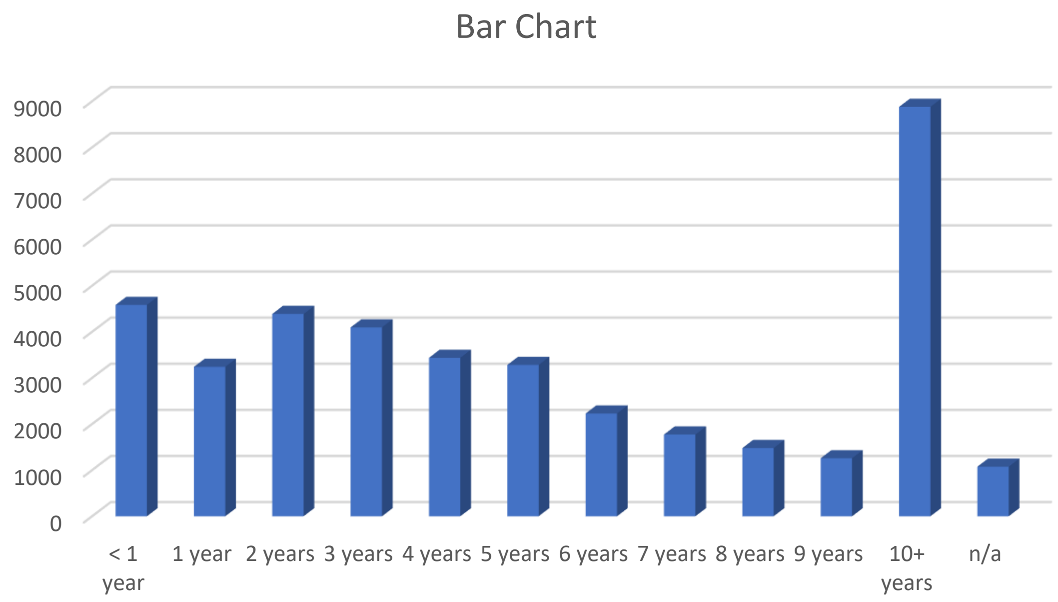
Similar to the previous example, we can see that the curves are turning more into a straight line as we move into a log scale





Let's try to see the frequency plot for some of the ordered sets

Row Labels	Count of emp_length
10+ years	8879
< 1 year	4583
2 years	4388
3 years	4095
4 years	3436
5 years	3282
1 year	3240
6 years	2229
7 years	1773
8 years	1479
9 years	1258
n/a	1075
Grand Total	**39717**



This analysis shows us that the people in the employment length of more than 10 years are more likely to opt for the loan than the other categories

The different descriptive statistics for Quantitative variables are done

- Mean
- Median
- Interquartile difference

loan_amnt
5000
2500
2400
10000
3000
5000
7000
3000
5600
5375
6500
12000
9000
3000
10000
1000

Mean	**Median**	**SD**	**Quantities**	**Values**
11219.44	10000	7456.671	25th percentile	5500
			50th percentile	10000
			75th percentile	15000
			100th percentile	35000

Here, we can see that the quartiles are a good way to understand the spread. The loan amount would be a big factor to consider in this analysis. In order to get the proper set of details, it would be good to analyse the data between the 25th percentile and the 75th percentile.

III. Segmented Univariant Analysis

Grade	Charged Off	Current	Fully Paid	Grand Total
A	602	40	9443	10085
B	1425	345	10250	12020
C	1347	264	6487	8098
D	1118	222	3967	5307
E	715	179	1948	2842
F	319	73	657	1049
G	101	17	198	316
Grand Total	**5627**	**1140**	**32950**	**39717**

Experience	Charged Off	Current	Fully Paid	Grand Total
< 1 year	639	75	3869	4583
1 year	456	71	2713	3240
10+ years	1331	391	7157	8879
2 years	567	97	3724	4388
3 years	555	83	3457	4095
4 years	462	94	2880	3436
5 years	458	88	2736	3282
6 years	307	61	1861	2229
7 years	263	62	1448	1773
8 years	203	44	1232	1479
9 years	158	32	1068	1258
n/a	228	42	805	1075
Grand Total	**5627**	**1140**	**32950**	**39717**

Home ownership	Charged Off	Current	Fully Paid	Grand Total
MORTGAGE	2327	638	14694	17659
NONE			3	3
OTHER	18		80	98
OWN	443	83	2532	3058
RENT	2839	419	15641	18899
Grand Total	**5627**	**1140**	**32950**	**39717**

The analysis has been done for various factors against the loan status. The major ones that could be pointed out here are Home ownership and Work Experience. It seems like people living on Rent, as well as people having more than 10 years of exp have more chances of defaulting.

IV. Bivariant Analysis

Row Labels	Average of annual_inc
car	**61842.04167**
Charged Off	54560.0245
Current	58038.82
Fully Paid	62854.20285
credit_card	**70439.14779**
Charged Off	64052.04421
Current	75787.81456
Fully Paid	71088.17733
debt_consolidation	**67322.05922**
Charged Off	61665.68666
Current	74824.57218
Fully Paid	68058.2386
educational	**53471.37409**
Charged Off	51711.80357
Fully Paid	53837.67874
home_improvement	**89736.78495**
Charged Off	77190.18882
Current	96555.15406
Fully Paid	91186.55297
house	**76772.28388**
Charged Off	71540.46847
Current	77157.14286
Fully Paid	77756.9887
major_purchase	**66391.5229**
Charged Off	56707.54523
Current	59994.16757
Fully Paid	67629.35755
medical	**68252.86377**
Charged Off	57261.35849
Current	111103.08
Fully Paid	69384.85849
moving	**61801.5783**
Charged Off	55533.76087
Current	47465.28571
Fully Paid	63200.32469
other	**63147.25236**
Charged Off	58676.59368
Current	72583.74844
Fully Paid	63649.12595
renewable_energy	**77490.00612**
Charged Off	59240.82263
Current	109000
Fully Paid	81287.89157
small_business	**75062.51516**
Charged Off	67556.26162
Current	71143.43243
Fully Paid	78076.96594
vacation	**59218.93346**
Charged Off	52452.5283
Current	65000
Fully Paid	60224.9368
wedding	**68663.28147**
Charged Off	66634.81542
Current	79228.11905
Fully Paid	68630.59611
Grand Total	**68968.92638**

Different set of analysis have been conducted and analysed.

This is an example of the details related to the purpose of the loan and how that affects the loan status, based on average annual income of the person as well.

Row Labels	Charged Off	Current	Fully Paid	Grand Total
< 1 year	**639**	**75**	**3869**	**4583**
MORTGAGE	177	39	1085	1301
NONE			2	2
OTHER	2		19	21
OWN	52	4	288	344
RENT	408	32	2475	2915
1 year	**456**	**71**	**2713**	**3240**
MORTGAGE	137	30	806	973
OTHER	4		11	15
OWN	28	4	186	218
RENT	287	37	1710	2034
10+ years	**1331**	**391**	**7157**	**8879**
MORTGAGE	753	265	4535	5553
OTHER	5		14	19
OWN	99	24	633	756
RENT	474	102	1975	2551
2 years	**567**	**97**	**3724**	**4388**
MORTGAGE	174	40	1255	1469
OTHER	2		8	10
OWN	43	4	236	283
RENT	348	53	2225	2626
3 years	**555**	**83**	**3457**	**4095**
MORTGAGE	202	35	1334	1571
OTHER	3		7	10
OWN	38	7	208	253
RENT	312	41	1908	2261
4 years	**462**	**94**	**2880**	**3436**
MORTGAGE	176	43	1183	1402
OTHER			7	7
OWN	28	9	204	241
RENT	258	42	1486	1786
5 years	**458**	**88**	**2736**	**3282**
MORTGAGE	197	49	1238	1484
NONE			1	1
OTHER	1		4	5
OWN	36	7	190	233
RENT	224	32	1303	1559
6 years	**307**	**61**	**1861**	**2229**
MORTGAGE	131	36	925	1092
OTHER			4	4
OWN	27	3	131	161
RENT	149	22	801	972
7 years	**263**	**62**	**1448**	**1773**
MORTGAGE	117	34	733	884
OTHER	1		2	3
OWN	22	4	114	140
RENT	123	24	599	746
8 years	**203**	**44**	**1232**	**1479**
MORTGAGE	102	23	656	781
OTHER			4	4
OWN	13	4	103	120
RENT	88	17	469	574
9 years	**158**	**32**	**1068**	**1258**
MORTGAGE	79	23	600	702
OWN	14	1	77	92
RENT	65	8	391	464
n/a	**228**	**42**	**805**	**1075**
MORTGAGE	82	21	344	447
OWN	43	12	162	217
RENT	103	9	299	411
Grand Total	**5627**	**1140**	**32950**	**39717**

This is an example of the analysis conducted based on the years of experience of the person that has an impact on the loan status, considering home ownership also as a factor.

V. Observations & Conclusions

Based on the analysis conducted, below are the conclusions that we could come up with:

- It seems that **higher interest rates** (e.g., 10.00%, 10.25%, etc.) have a higher frequency of borrowers in the 'Charged Off' category compared to 'Fully Paid'.
- It appears that borrowers with **higher employment lengths** (e.g., >10 years) have a higher frequency in the 'Charged Off' category compared to 'Fully Paid'.
- It appears that borrowers who are **renting** (RENT) have a higher frequency in the 'Charged Off' category compared to 'Fully Paid', while borrowers with a mortgage (MORTGAGE) have a higher frequency in the 'Fully Paid' category.
- It seems that borrowers who took loans for **purposes such as credit card, debt consolidation, and small business** have higher frequencies in the 'Charged Off' category compared to 'Fully Paid'.
- So to conclude, it is safe to assume that the above factors need to be prioritized while approving a loan, like trying to find the home ownership status and the employment length of the person and then using that to understand the principal amount and interest rates that will be used.