Lending Club Case Study

Group Members:

1. Sancheet Patil
2. Abhishek Sa

The Problem

Company

Lending Club is the largest online loan market-place which is facilitating different types of loans for Borrowers can easily avail through a fast online interface.

Context

Lending Club wants to understand the **main factors** behind loan default, i.e. the **driver variables** which strongly indicates of defaulter.

The company can utilise this knowledge for its portfolio and risk assessment.

Problem statement

As a data scientist working for Lending Club need to analyze the dataset containing information about past loan applicants using EDA to understand how *consumer attributes* and *loan attributes* *in*fluence the tendency of default

Analysis Approach

- Drop columns with NA values, all random values..
- Convert values to proper data-type as required.

- Analyze variables against segments of other variables
- Create derived variables

Publish insights and observations

Clean Data

Univariate Analysis

Segmented Univariate Analysis

Bivariate Analysis

Summarize Results

- Check distributions and frequencies of various numerical and categorical variables
- Create derived variables

- Do correlation analysis Check how two variables affect the value.

Data Preprocessing:

1. Approach the columnar way, i.e. if more than 50% of the data is not present then we can safely assume the column is not so useful here.
2. Convert few of the object/string data type into integer/real which can be continuous (operation-worthy) in nature.
3. Convert string dates into datetime type to derive day, month and year.
4. The outliers can impact our analyzing, trial of different ways to manage them, we have tried to cap/floor them as per distribution.
5. Imputation performed for the data which are not collected but are important, usage of descriptive (mean/median/mode) to fill those.

Univariate and Segmented univariate Analysis:

Each parameter driving the outcome

Analysis - Understanding Grades

0 - Paid off
1 - Charged off

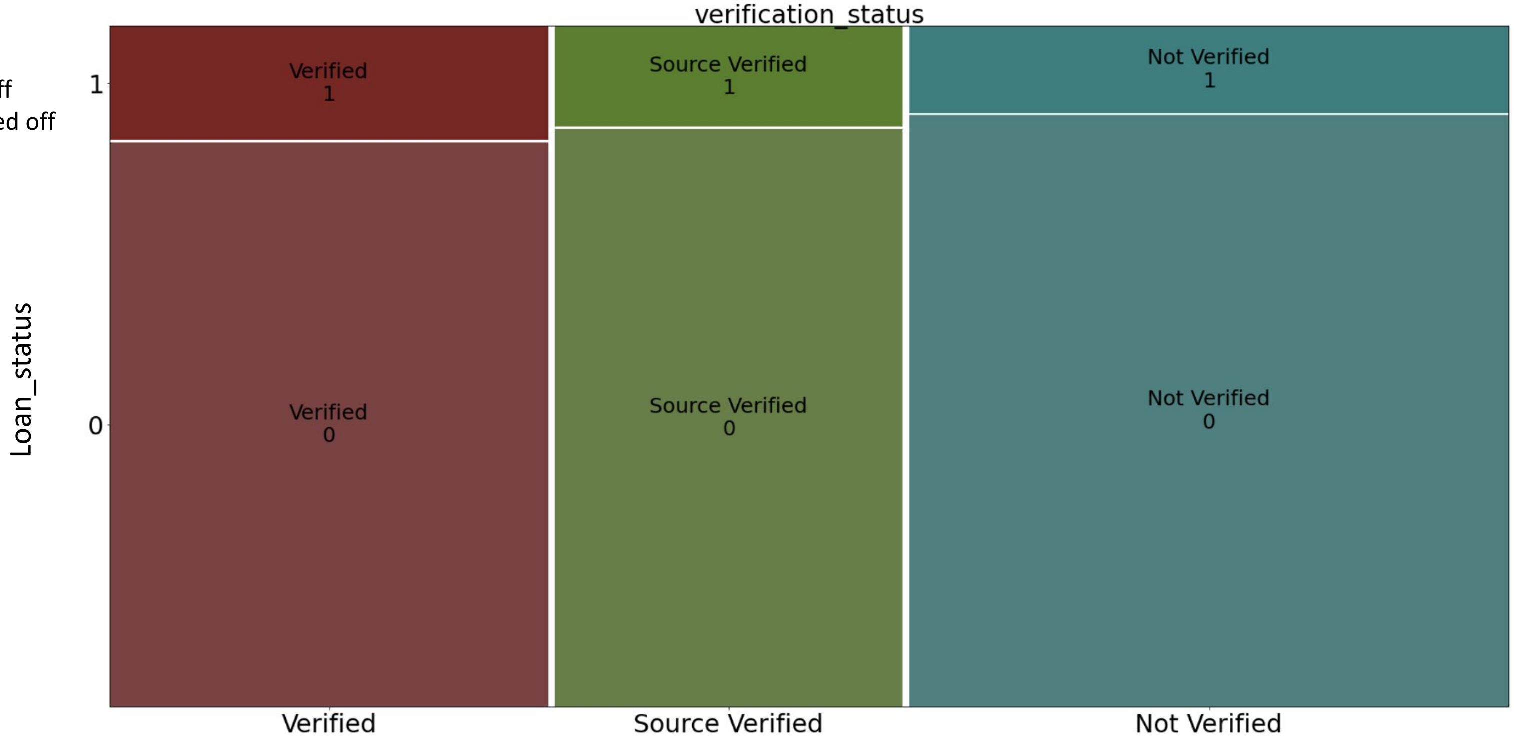
Plot description:
The width of the columns is proportional to the number of observations on horizontal.
The vertical length of the bars is proportional to the number of observations in the second variable within each level of the first variable.

Analysis outcome: Better grades have less chances of default/charged off.

Starting from A to G (A is a better grade than G)

Analysis - Understanding Verification Status

0 - Paid off
1 - Charged off

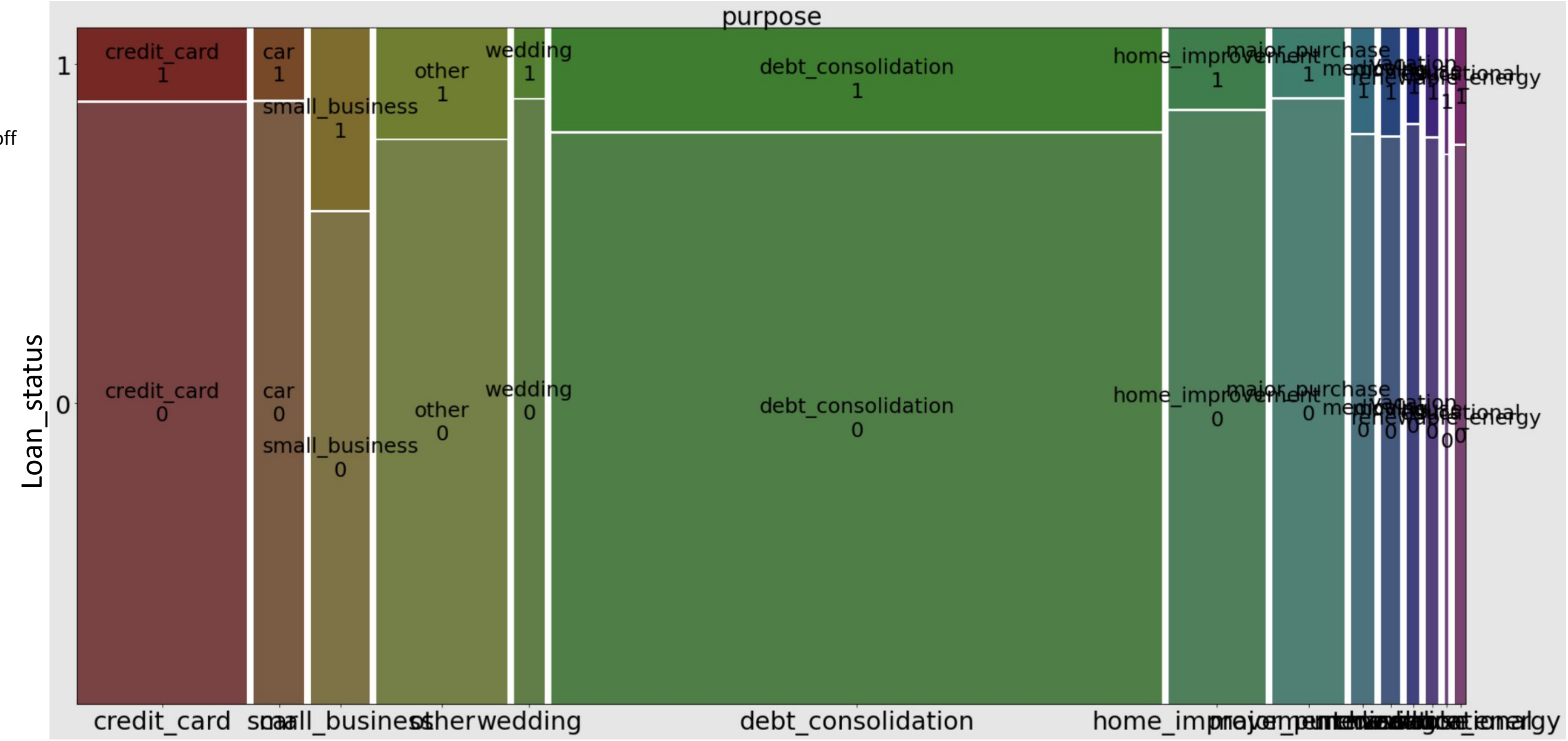


Plot description:
The width of the columns is proportional to the number of observations on horizontal.
The vertical length of the bars is proportional to the number of observations in the second variable within each level of the first variable.

Analysis outcome: Verified guy is defaulting, unverified population is paying off the loan, it can be taken as a factor that our bank verification has a flaw, even the source verification is in the mid.

Analysis - Purpose of the loan

0 - Paid off
1 - Charged off

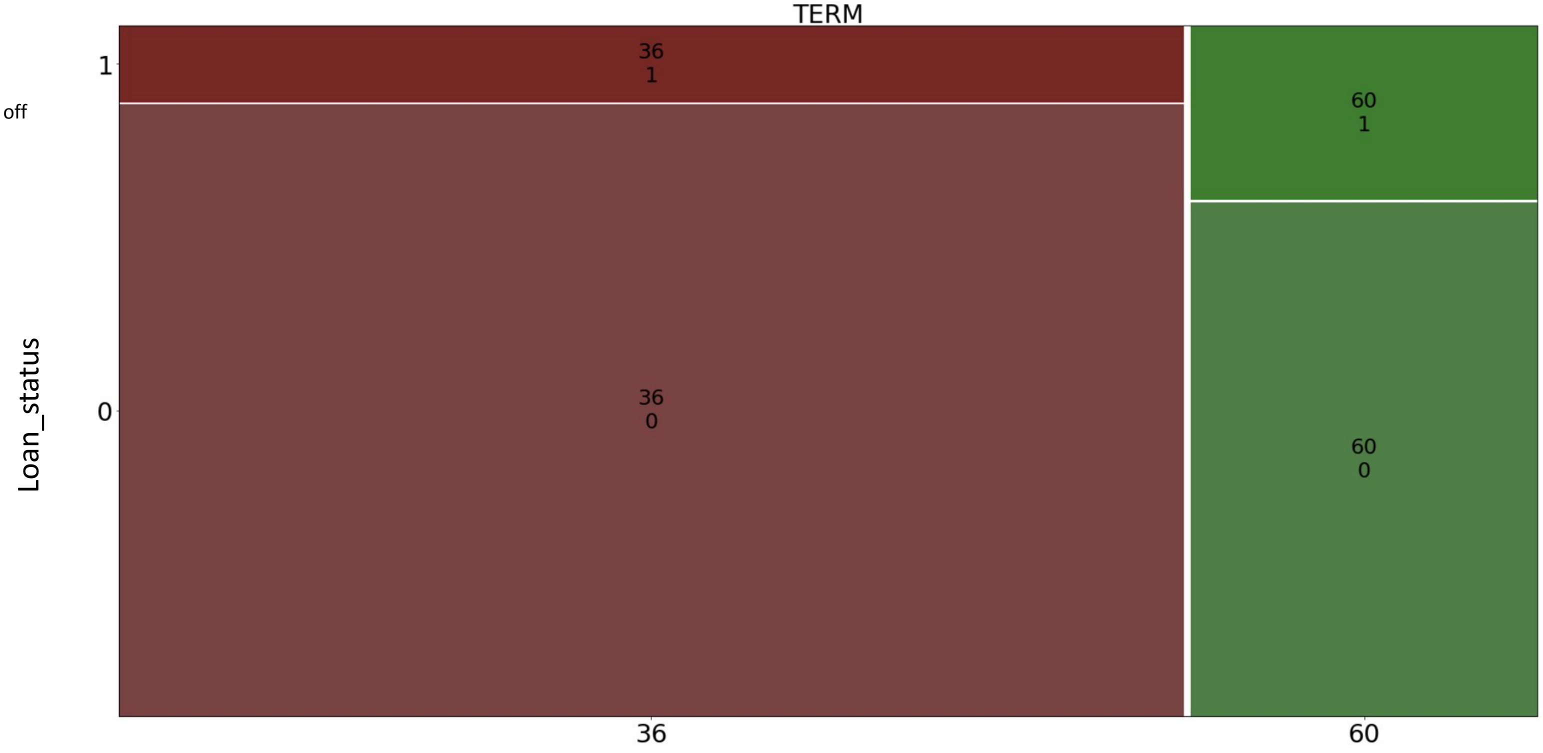


Analysis outcome: Though most of the population has taken loan for “debt consolidation”, but small business loan has seen the most defaulting rate. Major purpose, credit_card, car wedding and home_imprvement are safe bets.

Taking the significant population into observation

Analysis - Term of the loan

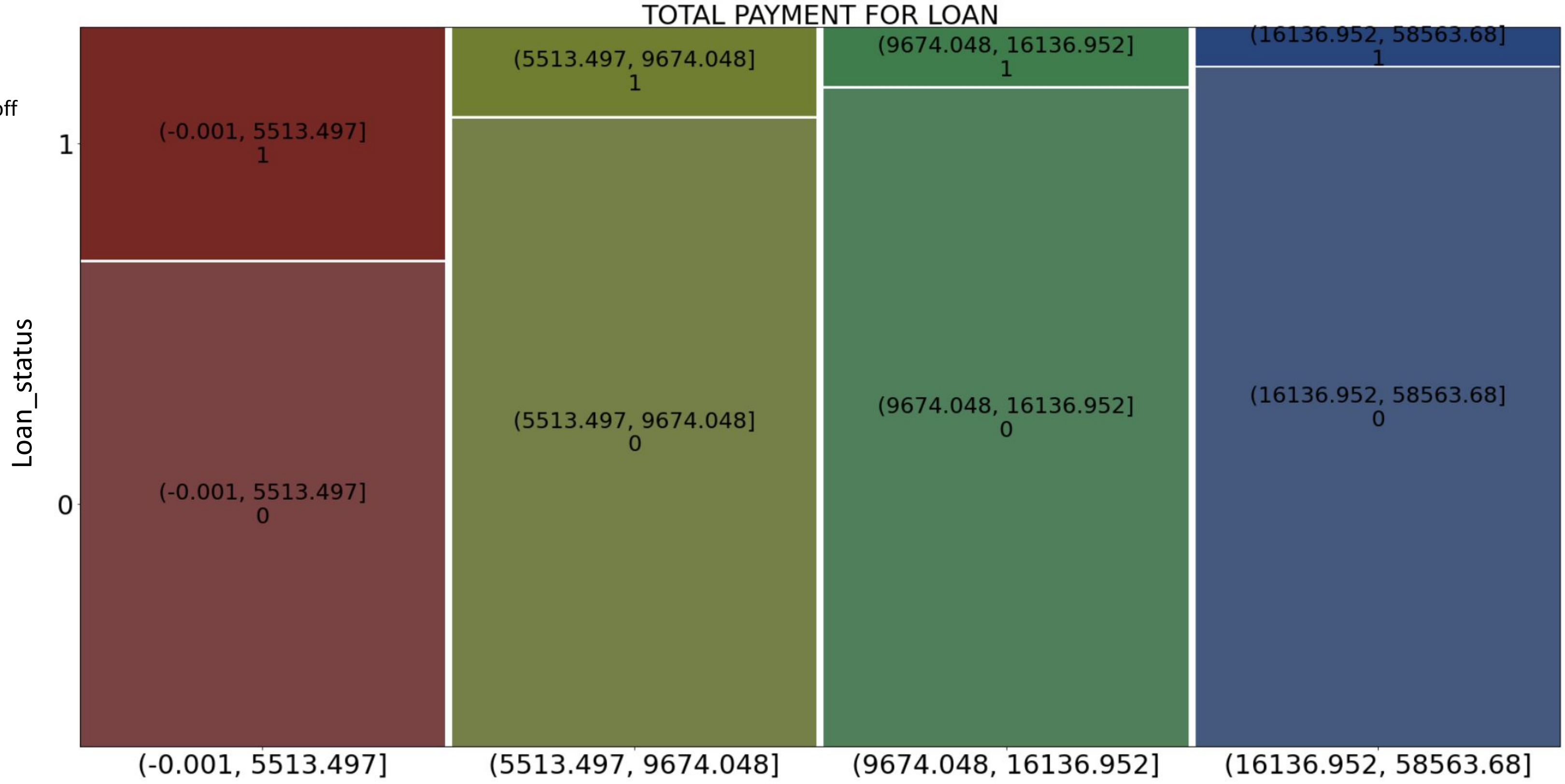
0 - Paid off
1 - Charged off



Analysis outcome: Majority of people have gone for 36 month tenure and have paid off much more and defaulted less. Otherwise the 60 year term shows the default level increasing and paid off coming down in ratio. Higher tenure can give an increase to the default.

Analysis - Total Payment by the loan applicant.

0 - Paid off
1 - Charged off

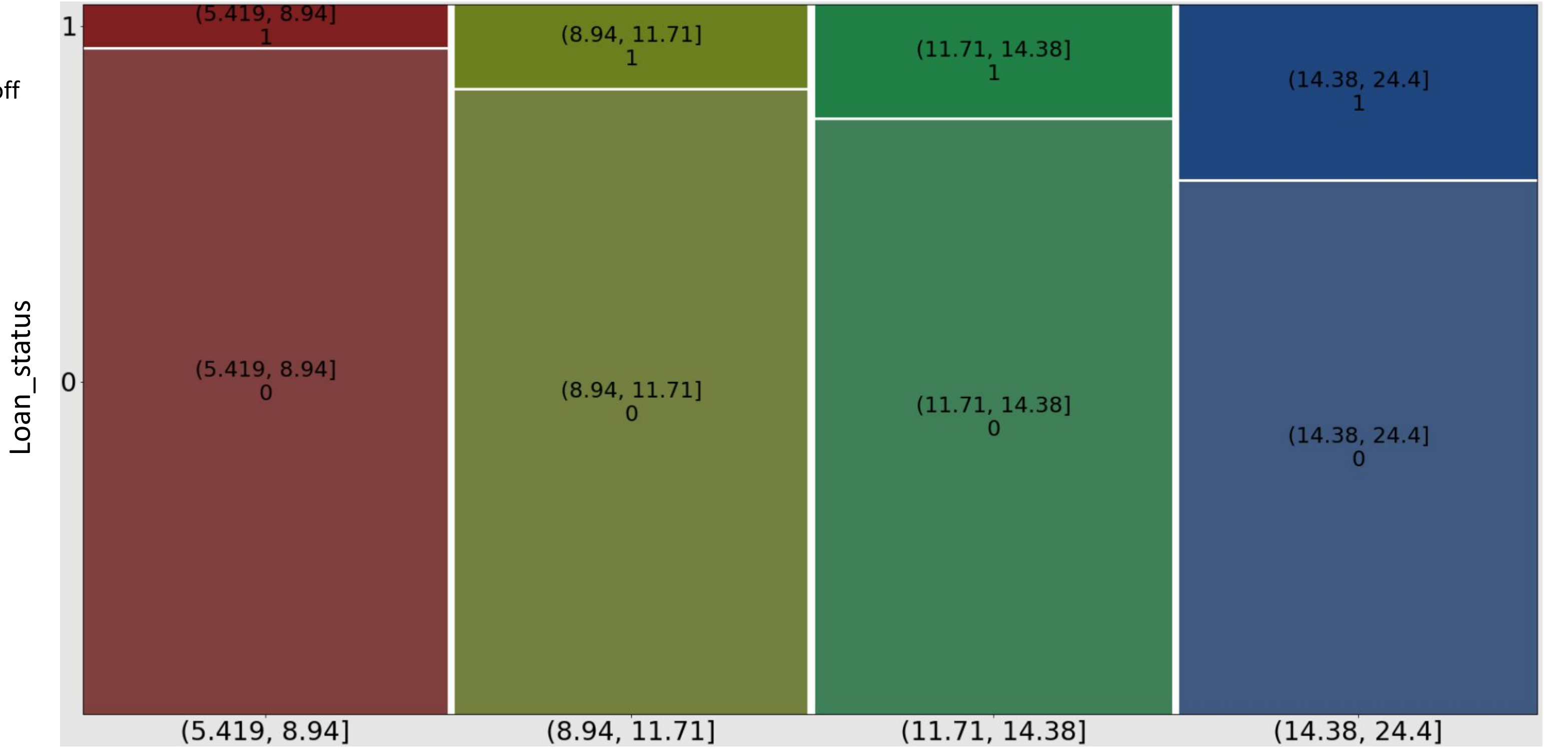


Analysis Outcome: Over an isometric distribution, the exceeding total payment is actually reducing the default rate. The default mostly exists for smaller loans as can be observed.

For spread out data, bins of equal distribution was used (quantiles)

Analysis - Interest rate

0 - Paid off
1 - Charged off



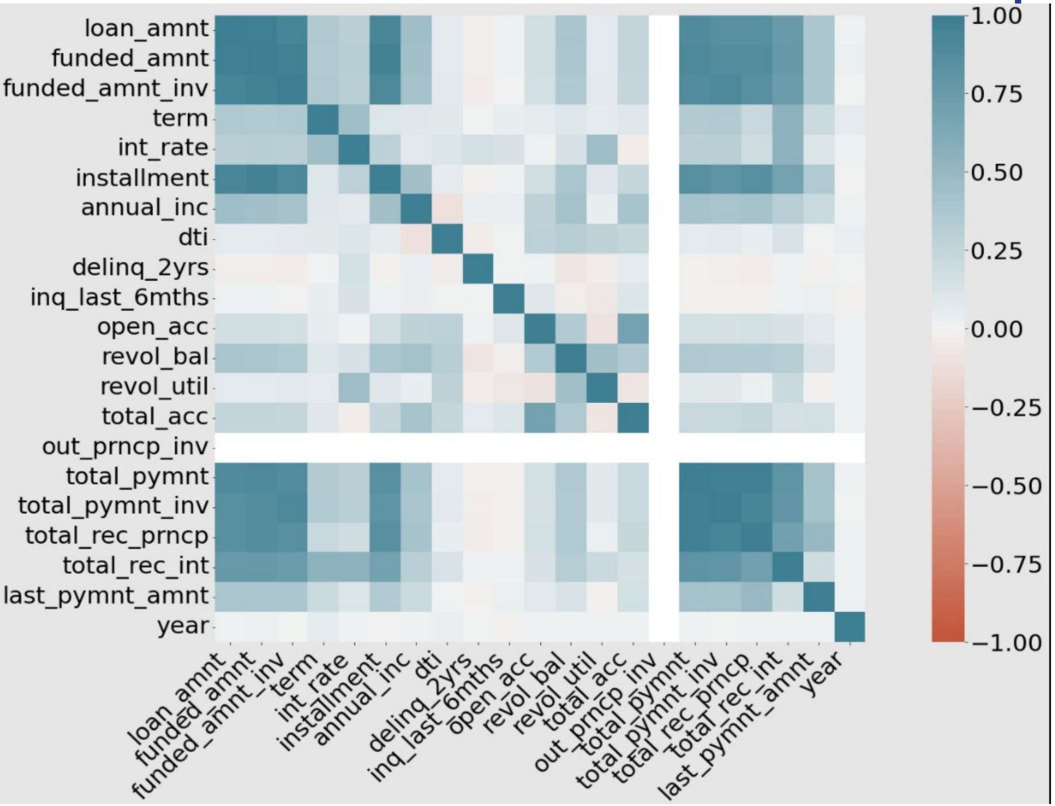
Analysis Outcome: Interest rate is directly proportional to the default rate as interest rate increases the defaulter rate also increase significantly.

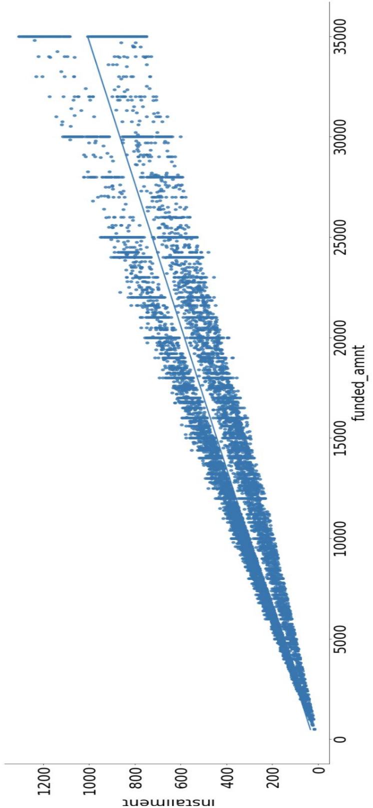
For spread out data, bins of equal distribution was used (quantiles)

Bivariate Analysis:

Two parameters impacting the loan status

Correlation: heat map and scatter plot





Analysis Outcome: Installment is highly correlated with the funded amount.

Open_acc is correlated to the total_access, both sort of explains the line of credit.

revol_util is correlated to int_rate that signifies the minimal int_rate for good revolving credit

Only works for the numeric data type, not for categories

A scatter plot can't be done between a categorical and a continuous variable. Hence we try a pivot table and layering with grade and subgrade.

Bivariate between Categorical and Numeric types

A scatter plot can't be done between a categorical and a contnuous variable. Hence we try a pivot table and layering with grade and subgrade.

Count of loan_status	Column Labels												
	(3999.999, 40000.0]		(3999.999, 40000.0] Total	(40000.0, 58868.0]		(40000.0, 58868.0] Total	(58868.0, 82000.0]		(58868.0, 82000.0] Total	(82000.0, 116000.0]		(82000.0, 116000.0] Total	Grand Total
Row Labels	0	1		0	1		0	1		0	1		
A	2312	238	2550	2385	174	2559	2470	116	2586	2276	74	2350	10045
B	2584	498	3082	2513	369	2882	2596	322	2918	2557	236	2793	11675
C	1672	458	2130	1670	352	2022	1545	314	1859	1600	223	1823	7834
D	940	334	1274	974	285	1259	1013	277	1290	1040	222	1262	5085
E	335	156	491	457	183	640	493	193	686	663	183	846	2663
F	88	46	134	121	64	185	173	109	282	275	100	375	976
G	16	21	37	29	15	44	49	30	79	104	35	139	299
Grand Total	**7947**	**1751**	**9698**	**8149**	**1442**	**9591**	**8339**	**1361**	**9700**	**8515**	**1073**	**9588**	**38577**

From grade A to G, people with increasing income may or may not defaulting less/more. In this situation we need a ratio.

Count of loan_status	Column Labels																
	(3999.999, 40000.0]		(3999.999, 40000.0] Total	(40000.0, 58868.0]		(40000.0, 58868.0] Total	(58868.0, 82000.0]		(58868.0, 82000.0] Total	(82000.0, 116000.0]		(82000.0, 116000.0] Total	Grand Total	Default rate			
Row Labels	0	1		0	1		0	1		0	1			CO:PO-Q1	CO:PO-Q2	CO:PO-Q3	CO:PO-Q4
A	**2312**	**238**	**2550**	**2385**	**174**	**2559**	**2470**	**116**	**2586**	**2276**	**74**	**2350**	**10045**	0.042145594			0.0035587
A1	250	11	261	281	10	291	298	8	306	280	1	281	1139	0.090909091	0.0343643	0.0261438	0.0140449
A2	340	34	374	376	25	401	367	10	377	351	5	356	1508	0.083333333	0.0623441	0.0265252	0.0437018
A3	429	39	468	470	31	501	436	16	452	372	17	389	1810	0.083443709	0.0618762	0.0353982	0.0337423
A4	692	63	755	656	52	708	717	41	758	630	22	652	2873	0.13150289	0.0734463	0.0540897	0.0431548
A5	601	91	692	602	56	658	652	41	693	643	29	672	2715	0.161583387	0.0851064	0.0591631	0.084497
B	**2584**	**498**	**3082**	**2513**	**369**	**2882**	**2596**	**322**	**2918**	**2557**	**236**	**2793**	**11675**	0.126415094	0.1280361	0.1103496	0.0598958
B1	463	67	530	404	42	446	398	39	437	361	23	384	1797	0.150735294	0.0941704	0.0892449	0.0956341
B2	462	82	544	449	58	507	427	42	469	435	46	481	2001	0.159440559	0.1143984	0.0895522	0.0786517
B3	601	114	715	615	79	694	612	92	704	656	56	712	2825	0.176661264	0.1138329	0.1306818	0.0964052
B4	508	109	617	488	84	572	559	77	636	553	59	612	2437	0.186390533	0.1468531	0.1210692	0.0860927
B5	550	126	676	557	106	663	600	72	672	552	52	604	2615	0.215023474	0.1598793	0.1071429	0.1223258
C	**1672**	**458**	**2130**	**1670**	**352**	**2022**	**1545**	**314**	**1859**	**1600**	**223**	**1823**	**7834**	0.217625899	0.1740851	0.168908	0.1209016
C1	435	121	556	446	86	532	409	70	479	429	59	488	2055	0.225378788	0.1616541	0.1461378	0.1123348
C2	409	119	528	402	82	484	396	69	465	403	51	454	1931	0.223376623	0.1694215	0.1483871	0.1347305
C3	299	86	385	329	74	403	301	65	366	289	45	334	1488	0.212962963	0.1836228	0.1775956	0.1227437
C4	255	69	324	262	50	312	234	59	293	243	34	277	1206	0.18694362	0.1602564	0.2013652	0.1259259
C5	274	63	337	231	60	291	205	51	256	236	34	270	1154	0.262166405	0.2061856	0.1992188	0.1759113
D	**940**	**334**	**1274**	**974**	**285**	**1259**	**1013**	**277**	**1290**	**1040**	**222**	**1262**	**5085**	0.231617647	0.2263701	0.2147287	0.1724138
D1	209	63	272	194	34	228	193	35	228	168	35	203	931	0.256329114	0.1491228	0.1535088	0.1420118
D2	235	81	316	238	77	315	252	65	317	290	48	338	1286	0.249134948	0.2444444	0.2050473	0.1954887
D3	217	72	289	221	71	292	208	61	269	214	52	266	1116	0.336492891	0.2431507	0.2267658	0.1689498
D4	140	71	211	171	50	221	210	57	267	182	37	219	918	0.252688172	0.2262443	0.2134831	0.2118644
D5	139	47	186	150	53	203	150	59	209	186	50	236	834	0.317718941	0.2610837	0.2822967	0.2163121
E	**335**	**156**	**491**	**457**	**183**	**640**	**493**	**193**	**686**	**663**	**183**	**846**	**2663**	0.300653595	0.2859375	0.2813411	0.241206
E1	107	46	153	122	53	175	144	51	195	151	48	199	722	0.269230769	0.3028571	0.2615385	0.1891892
E2	95	35	130	101	44	145	105	49	154	150	35	185	614	0.35106383	0.3034483	0.3181818	0.1604938
E3	61	33	94	109	29	138	91	31	122	136	26	162	516	0.348484848	0.2101449	0.2540984	0.2519084
E4	43	23	66	69	33	102	88	37	125	98	33	131	424	0.395833333	0.3235294	0.296	0.2426036
E5	29	19	48	56	24	80	65	25	90	128	41	169	387	0.343283582	0.3	0.2777778	0.2666667
F	**88**	**46**	**134**	**121**	**64**	**185**	**173**	**109**	**282**	**275**	**100**	**375**	**976**	0.304347826	0.3459459	0.3865248	0.2689076
F1	32	14	46	41	17	58	54	28	82	87	32	119	305	0.428571429	0.2931034	0.3414634	0.2183908
F2	20	15	35	35	17	52	40	19	59	68	19	87	233	0.263157895	0.3269231	0.3220339	0.2105263
F3	14	5	19	17	11	28	32	19	51	60	16	76	174	0.227272727	0.3928571	0.372549	0.3953488
F4	17	5	22	21	13	34	34	18	52	26	17	43	151	0.583333333	0.3823529	0.3461538	0.32
F5	5	7	12	7	6	13	13	25	38	34	16	50	113	0.567567568	0.4615385	0.6578947	0.2517986
G	**16**	**21**	**37**	**29**	**15**	**44**	**49**	**30**	**79**	**104**	**35**	**139**	**299**	0.5	0.3409091	0.3797468	0.2272727
G1	8	8	16	10	4	14	11	9	20	34	10	44	94	0.6	0.2857143	0.45	0.3225806
G2	4	6	10	8	5	13	16	7	23	21	10	31	77	0.666666667	0.3846154	0.3043478	0.3333333
G3	1	2	3	5	3	8	8	8	16	12	6	18	45	0.8	0.375	0.5	0.1333333
G4	1	4	5	5	3	8	9	2	11	26	4	30	54	0.333333333	0.375	0.1818182	0.3125
G5	2	1	3	1		1	5	4	9	11	5	16	29	0.180552691	0	0.4444444	0.1119107
Grand Total	**7947**	**1751**	**9698**	**8149**	**1442**	**9591**	**8339**	**1361**	**9700**	**8515**	**1073**	**9588**	**38577**	#DIV/0!			#DIV/0!

With the increasing income and top grades of the population the default rate is decreasing.

The highlight is from left to right

Bivariate between types:

Total Payment Loan bins against the interest rate bins, output the loan status

Count of loan_status	Column Labels												
	(11.71, 14.38]		(11.71, 14.38] Total	(14.38, 24.4]		(14.38, 24.4] Total	(5.419, 8.94]		(5.419, 8.94] Total	(8.94, 11.71]		(8.94, 11.71] Total	Grand Total
Row Labels	0	1		0	1		0	1		0	1		
(-0.001, 5513.497]	1521	903	2424	877	1132	2009	2459	417	2876	1661	675	2336	9645
(16136.952, 58563.68]	2528	118	2646	3449	346	3795	1026	6	1032	2129	42	2171	9644
(5513.497, 9674.048]	1887	314	2201	1245	514	1759	3036	119	3155	2283	246	2529	9644
(9674.048, 16136.952]	2143	203	2346	1691	389	2080	2702	43	2745	2313	160	2473	9644
Grand Total	**8079**	**1538**	**9617**	**7262**	**2381**	**9643**	**9223**	**585**	**9808**	**8386**	**1123**	**9509**	**38577**

- As we can see, the maximum default for the lower bin of total payment(row) at any interest rate bin.
- The least risk of defaulting at any rate of interest bracket for the second bracket total payment.
- The most default is for lowest bracket of total amount payable with the highest rate of interest.
- The least default is for the second bracket of total amount payable at the smallest bin of rate of interest.

Count of loan_status	Column Labels												
	(-0.001, 5513.497]		(-0.001, 5513.497] Total	(16136.952, 58563.68]		(16136.952, 58563.68] Total	(5513.497, 9674.048]		(5513.497, 9674.048] Total	(9674.048, 16136.952]		(9674.048, 16136.952] Total	Grand Total
Row Labels	0	1		0	1		0	1		0	1		
Not Verified	3855	1435	5290	2025	62	2087	4595	435	5030	4077	210	4287	16694
Source Verified	1746	877	2623	1851	84	1935	2267	296	2563	2379	177	2556	9677
Verified	917	815	1732	5256	366	5622	1589	462	2051	2393	408	2801	12206
Grand Total	**6518**	**3127**	**9645**	**9132**	**512**	**9644**	**8451**	**1193**	**9644**	**8849**	**795**	**9644**	**38577**

Verification status when mapped with total loan payment leading to the loan status.

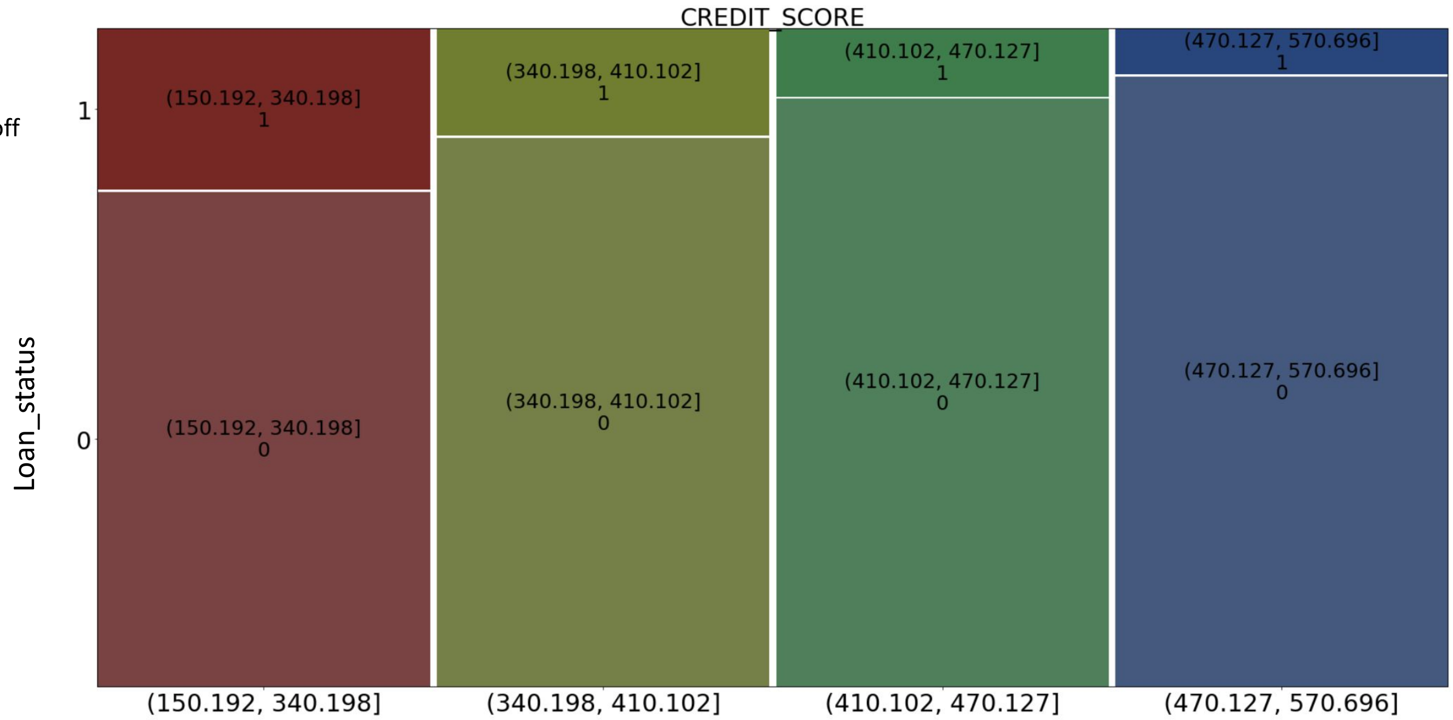
- The unverified small bracket loans are the ones to be most defaulted.
- The most defaults is in the smallest bracket loan, irrespective of the verification status.
- The least default case is the largest bin of the loan amount and not verified.

Derived Analysis:

Using Multiple parameters driving one more parameter

Business-Driven Derived: Credit_score appended.

0 - Paid off
1 - Charged off



Analysis Outcome : As we can see that the derived credit score as high, the default rate comes down.

We can assume anything more than 450 is a good score.

Max credit score can go till 570, the upper limit can be rounded off to 600.

Formula Used :

credit score = (dti +grade)*30%+ (10 -inq last 6mths)*10%) - (total_rec_late_fee *35%)+(purpose*25%)

Summary and other observations:

- Removal of columns without much variance and need of the core fintech knowledge are ignored throughout categories and numeric.
- Few other factors which were not a major driving factor for the default/charged off status, yet were observed:

    - **Home Ownership**: Those who have mortgage are less default, who's home_ownership is on rent or other are likely to default.
    - **States**: NV state defaults the most (less population), FL with standard population defaults next, most loans go to CA, Texas & NY has the least default and PA is the best performing with a small population.
    - **Loan amount**:Through this we can conclude that the loan amount increment is exactly boosting the number of defaults, Also there is not an entire inverse happening at the paid off, it is increasing till the third bracket. Once the loan amount goes passed 75% of the spread,the amount of paid off is declining and default is increasing.
    - **Funded amount** : Loan amount and funded amount are entirely the same, again a correlation that can happen between funded_amount and loan_amount
    - **Installment**: The population is comparable and the increment in installment can be a driving factor to the default status. Although till the 50% of the spread of the installment, the rate of default doesn't make too much difference. After 50% of the spread which is somewhere around 250 is incrementing the default rate.
    - **Annual income**: Higher the income lower the default rate over proportional distribution, the income is a driving factor univariately.
    - **Debt to Income**: The bigger the debt over the income, the more the debt over income increases,it will result in default.
    - Columns like **issue date, title and zip code** had too many discrete information and cannot be binned.

- As per us the major driving factors towards default/charged off : **Grades, Verification, Purpose, Term, Total payment, Interest rate** through the univariate analysis.
- We went through heatmap analysis and scatter plot to figure out correlations. But they do work on numeric to numeric type correlation.
- We tried pivottable to get more clarity on **categorical vs numeric type**.
- We derived two derived metrics:
    - **Type driven metric: Approved date** was extracted from the description and from there we can derive day, month and year.
    - **Business driven metric:** We derived **“Credit score”** which is equivalent to **“FICO score” and “Cibil score”** which can be used by the LC to provide safer loans.

- We used **mosaic plot** mostly as it gives us both info about the population and categorical comparison, we also tried smthn like multivariate analysis, quite useful

Thank You