Lending Club Case Study (M.Sc. ML/AI)

- Dhananjay Joshi
- Shajiv Kalangath

Executive Summary

- **Objective - Lending Club provided historical data with several customer attributes and loan attributes. Objective is to apply EDA principles to find features influencing loan defaults , given the past data for 37K+ consumers.**

- **Key Conclusions :**
    - LC Grades E, F, G have higher defaulting risk. In general, as grade increases, higher defaulting risks from approved loan requests.
    - Zip code ranges 300xx-400xx , 900xx+ and 800xx-900xx have higher defaulting tendency. (may be certain income groups ?)
    - NE state stands out in percentage terms, however, in absolute terms, CA has higher defaulting rate.
    - By absolute terms and percentage terms - 'small business' and 'debt consolidation' have higher default.
    - Whoever have 2006/2007 as initial credit line year , have higher defaulting rate.
    - %wise higher defaults in 2007, however absolute number wise 2011 was special year as well when default numbers increased.
    - longer term loans (36 months) have higher risk than shorter term loans, specially it also links to employment length. So, employment length 10+ year, higher interest rate and long-term loans have higher risk of defaulting - kind of deadly combination.
    - out of approved loan amount, around 60% of principle could be recovered i.e., remaining average 40% is risk that organisation carries when someone defaults. However, such risk for small loan amounts up to 15K is much higher and Bank may lose from 50-100% of approved amt.
    - Gross recoveries post charge off are higher up to 25K loan amount and for 35K as well. recoveries increased linearly with loan amount up to 25K, which means Bank is not losing money and able to recover from risk position. Likely Bank's recovery infrastructure is strong but maybe it is an additional cost.
    - Loan defaulting increases after 8 years of employment length marginally.
    - For longer term loans given for educational purposes to people with more than 8 years employment do have small risk of defaulting.

Approach

Four Step approach was taken to complete EDA and draw conclusions from the data :

- Data Cleansing
- Descriptive Statistics Review
- Data Visualization
- Inferences

I. Data Cleansing

- **Drop redundant features - ** Any features with just 1 unique value were dropped, descriptions which are more apt for NLP were dropped

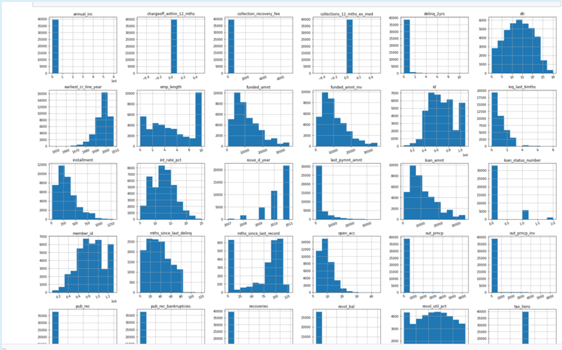
- **Conversion of DateTime Features –** At least five features related to time were in the format Mon-YY, they were converted to Pandas Datetime format

- **Handling null values –** Features with majority having null values were dropped.

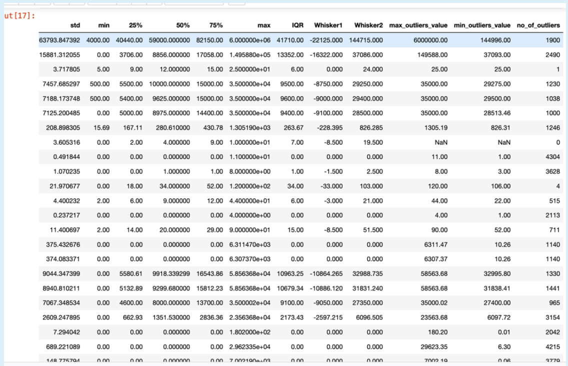
- **Feature Derivation –** From existing features new features were derived (e.g., Interest rate percentage from int_rate, Employment length converted into numerical column, Zip-codes were converted as numbers , year columns were derived for few features.

II. Descriptive Statistics Review

- **Histogram Review –** Studied distribution of individual numeric features



- **Distribution and Outlier Review –** Outliers were reviewed with the help of Box Plot and IQR (Inter Quantile Range) calculations

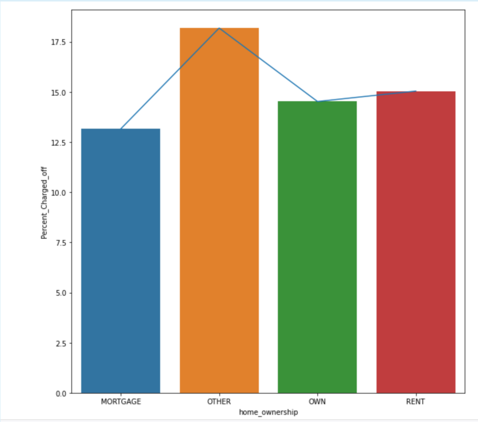


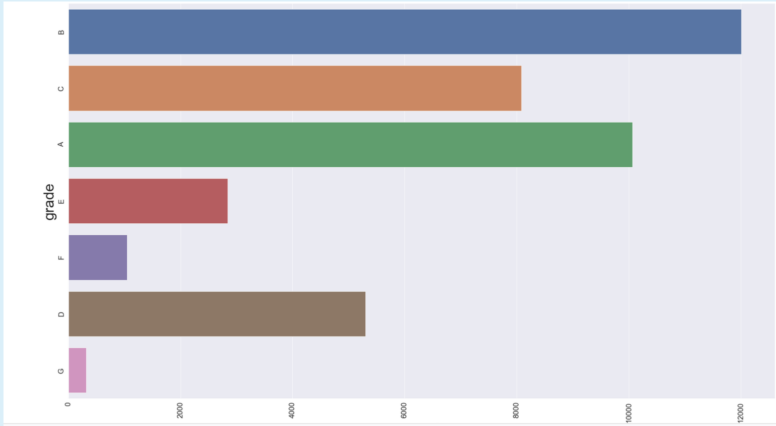
ut[17]:

std	min	25%	50%	75%	max	IQR	Whisker1	Whisker2	max_outliers_value	min_outliers_value	no_of_outliers
63793.847392	4000.00	40440.00	59000.000000	82150.00	6.000000e+06	41710.00	-22125.000	144715.000	6000000.00	144996.00	1900
15881.312055	0.00	3706.00	8856.000000	17058.00	1.495880e+05	13352.00	-16322.000	37086.000	149588.00	37093.00	2490
3.717805	5.00	9.00	12.000000	15.00	2.500000e+01	6.00	0.000	24.000	25.00	25.00	1
7457.685297	500.00	5500.00	10000.000000	15000.00	3.500000e+04	9500.00	-8750.000	29250.000	35000.00	29275.00	1230
7188.173748	500.00	5400.00	9625.000000	15000.00	3.500000e+04	9600.00	-9000.000	29400.000	35000.00	29500.00	1038
7125.200485	0.00	5000.00	8975.000000	14400.00	3.500000e+04	9400.00	-9100.000	28500.000	35000.00	28513.46	1000
208.898305	15.69	167.11	280.610000	430.78	1.305190e+03	263.67	-228.395	826.285	1305.19	826.31	1246
3.605316	0.00	2.00	4.000000	9.00	1.000000e+01	7.00	-8.500	19.500	NaN	NaN	0
0.491844	0.00	0.00	0.000000	0.00	1.100000e+01	0.00	0.000	0.000	11.00	1.00	4304
1.070235	0.00	0.00	1.000000	1.00	8.000000e+00	1.00	-1.500	2.500	8.00	3.00	3628
21.970677	0.00	18.00	34.000000	52.00	1.200000e+02	34.00	-33.000	103.000	120.00	106.00	4
4.400232	2.00	6.00	9.000000	12.00	4.400000e+01	6.00	-3.000	21.000	44.00	22.00	515
0.237217	0.00	0.00	0.000000	0.00	4.000000e+00	0.00	0.000	0.000	4.00	1.00	2113
11.400697	2.00	14.00	20.000000	29.00	9.000000e+01	15.00	-8.500	51.500	90.00	52.00	711
375.432676	0.00	0.00	0.000000	0.00	6.311470e+03	0.00	0.000	0.000	6311.47	10.26	1140
374.083371	0.00	0.00	0.000000	0.00	6.307370e+03	0.00	0.000	0.000	6307.37	10.26	1140
9044.347399	0.00	5580.61	9918.339299	16543.86	5.856368e+04	10963.25	-10864.265	32988.735	58563.68	32995.80	1330
8940.810211	0.00	5132.89	9299.680000	15812.23	5.856368e+04	10679.34	-10886.120	31831.240	58563.68	31838.41	1441
7067.348534	0.00	4600.00	8000.000000	13700.00	3.500002e+04	9100.00	-9050.000	27350.000	35000.02	27400.00	965
2609.247895	0.00	662.93	1351.530000	2836.36	2.356368e+04	2173.43	-2597.215	6096.505	23563.68	6097.72	3154
7.294042	0.00	0.00	0.000000	0.00	1.802000e+02	0.00	0.000	0.000	180.20	0.01	2042
689.221089	0.00	0.00	0.000000	0.00	2.962335e+04	0.00	0.000	0.000	29623.35	6.30	4215
148.775794	0.00	0.00	0.000000	0.00	7.002190e+03	0.00	0.000	0.000	7002.19	0.06	3779

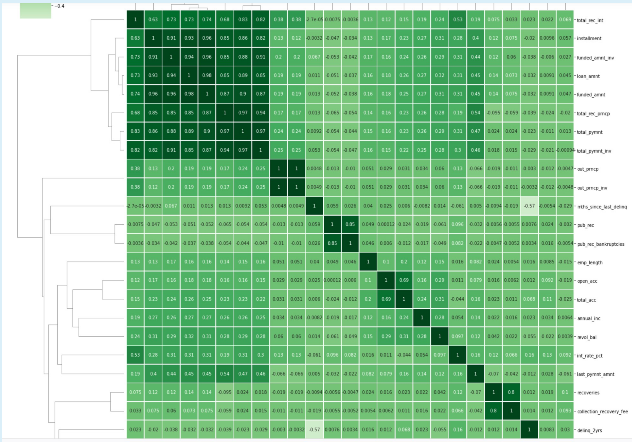
III. Data Visualization

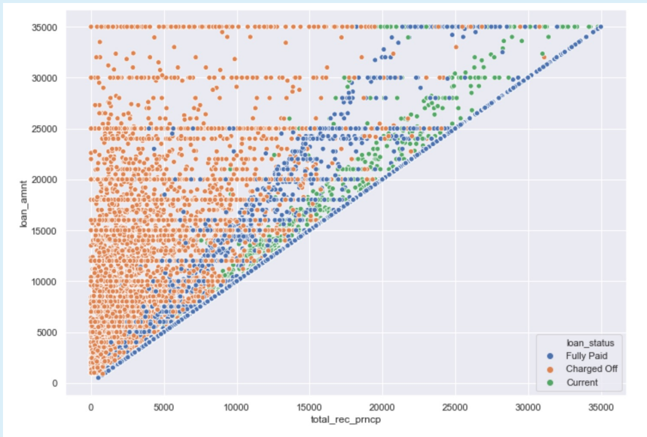
- **Univariate Analysis –** Studied distribution of individual numeric and categorical features





- **Bi-Variate Analysis – using correlation matrix, Pivot Table and graphs, seaborn scatter plots**





IV. Inferences

- **Customer Attributes influencing Loan Default :**

Customer Attributes
Employment Length
Purpose of Loan
First Credit Line year
Zip Code and State Address

- **Loan Attributes influencing Loan Default :**

Customer Attributes
LC Grade
Loan Tenure
Loan Issued Year
Interest Rate
Loan Amount and Recovery