Lending Club Case Study

UPGRAD

Table of Contents

- Problem Statement
- Solution at a high Level
- Data Analysis
- Conclusion

Problem Statement

- To develop an understanding of the provided Dataset and provide the company with the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Solution at a high Level

- The following categories help in understanding the work done on the given data
  - Data Understanding
  - Data Cleaning and manipulation
  - Data Analysis

Data Understanding

- A data set and data dictionary were provided to understand the data.
- The dataset contained data of Lending club a finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company
- It contained the complete loan data for all loans issued through the time period 2007 t0 2011.
- borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

Data Cleaning and manipulation

- <u>Cleanup</u>
- The initial data provided had 39717 Rows and 11 Columns.
- There was about 51% NULL values in the sheet
- 53 Columns which had all values NULL were dropped initially
- There were 4 columns which had very less unique values hence they were dropped too
- Unwanted columns that were created after a loan application is approved doesn't make sense in analysis
  - *del_cols = ["out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv", "total_rec_prncp","total_rec_int","total_rec_late_fee","recoveries","collection_recovery_fee","last_pymnt_d","last_pymnt_amnt","last_credit_pull_d"]*
- % symbols were stripped int_rate and revol_util columns
- Two new columns month and year were derived from issue date

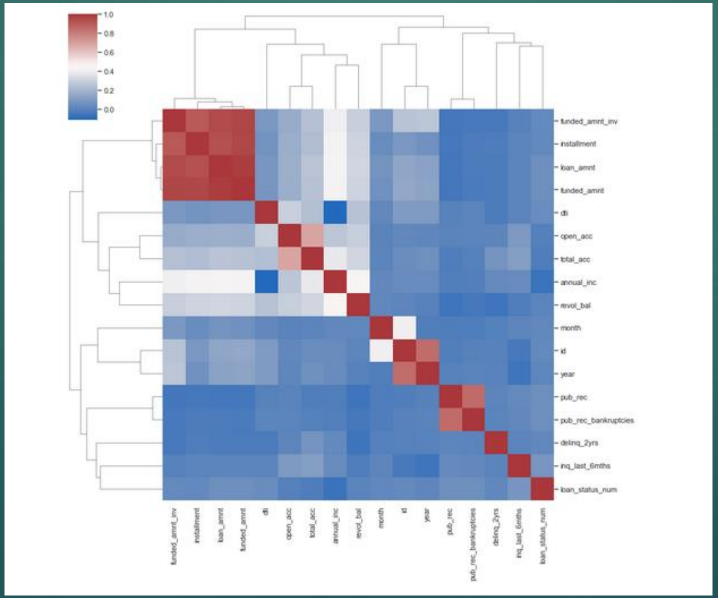
- <u>Outliers</u>
- Removed Outliers quantile .99 from Annual Income

- And hence the data was ready for analysis with 39319 rows and 30 columns

Data Analysis

Univariate Analysis

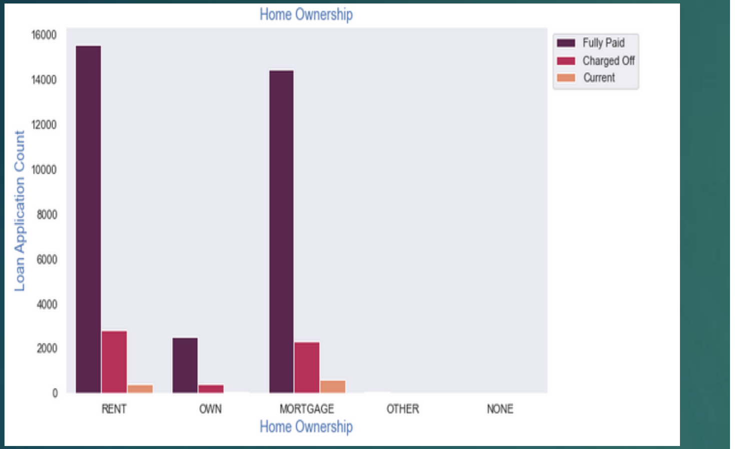
- A correlation chart was arrived which gave us a clear Idea on which of the columns were correlated
- Since we know darker the value higher the correlation, we can clearly see loan_amnt, funded_amnt, funded_amnt_inv and installment have huge correlation to each other. Also, the public records related fields pub_rec & pub_rec_bankrupcies and number of accounts related fields open_acc & total_acc are correlated.



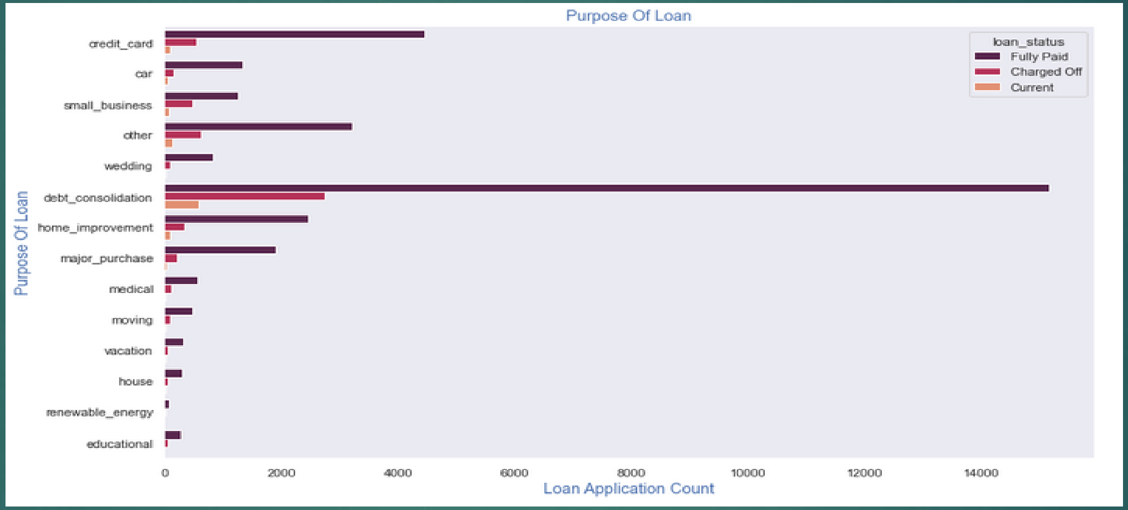
Data Analysis

Unordered Categorical Variables

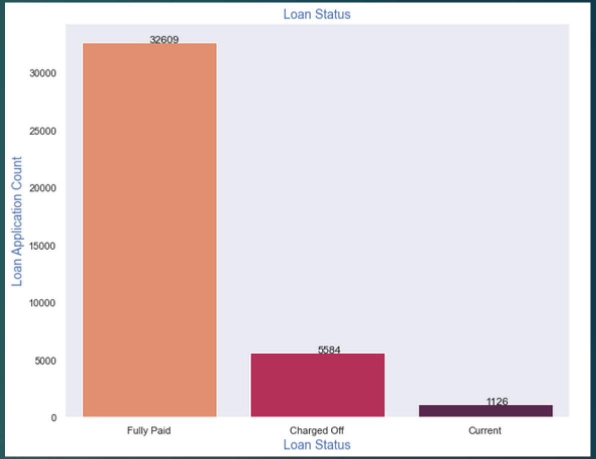
Home Ownership



Purpose Of Loan



Loan Status



- Observations :
- Home Ownership plot shows that most of them living in rented home or have mortgaged their home.
- Applicant numbers are high from these categories so charged off is high too.

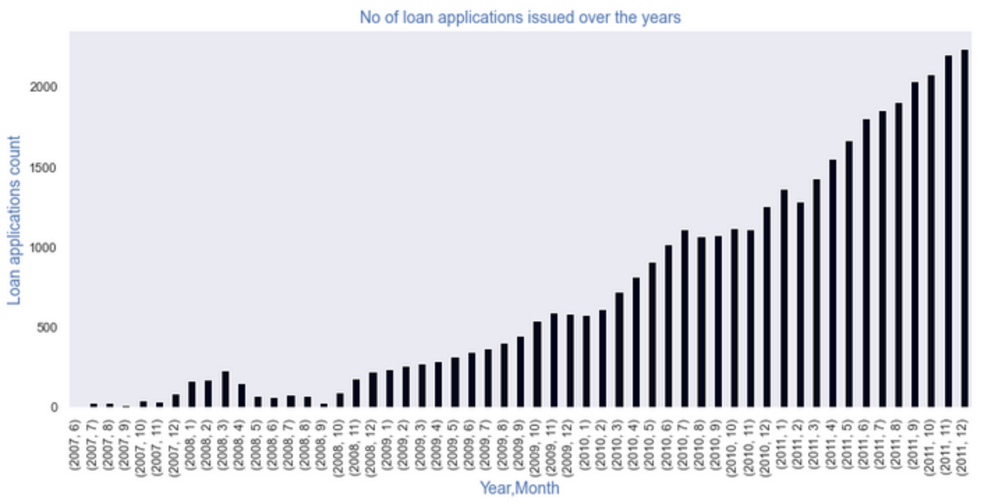
- Purpose Of Loan plot shows that most of the loans were taken for the purpose of debt consolidation & paying credit card bill.
- Number of charged off count also high too for these loans.

- Loan Status plot shows that close to 14% loans were charged off out of total loan issued.

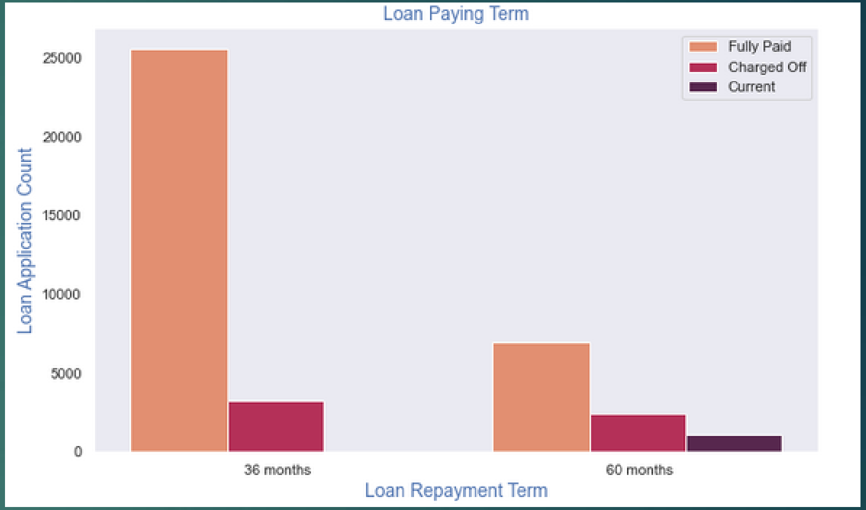
Data Analysis

Ordered Categorical Variables

Plot to check pattern of loan issuing over the years



Loan Paying Term



- Observations :
- Loan Paying term plot shows that those who had taken loan to repay in 60 months had more % of number of applicants getting charged off as compared to applicants who had taken loan for 36 months.

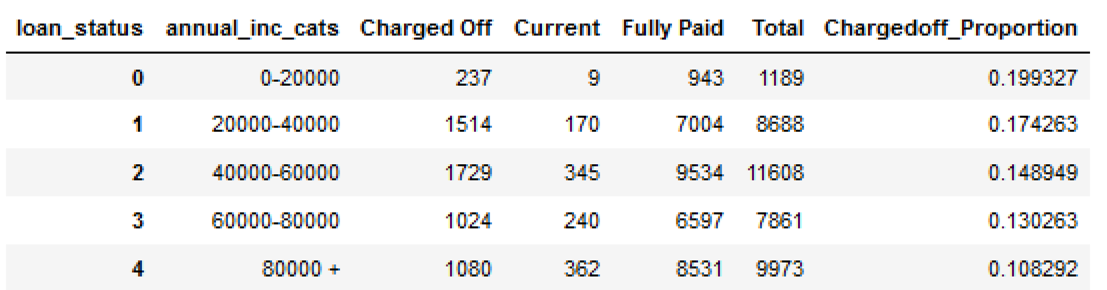
- The count of loan application is increasing every passing year. So increase in number of loan applications are adding more to number of charged off applications.
- Number of loans issued in 2008(May-October) got dipped, may be due to Recession.

Data Analysis

Bivariate Analysis

- The following Derived Columns were created from few of the existing columns
  - loan_amnt_cats from loan_amnt
  - annual_inc_cats from annual_inc
  - dti_cats from dti_cats
- These columns were used to analyse further

Annual income against Chargedoff_Proportion

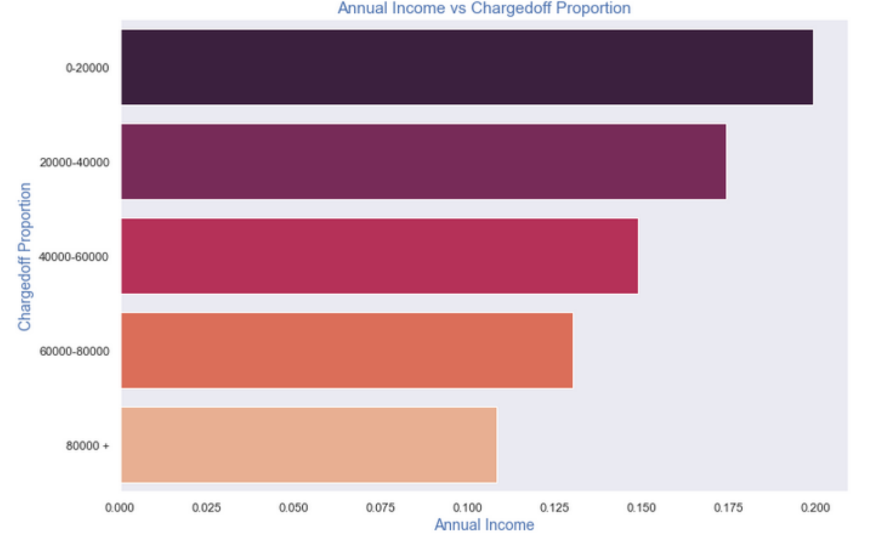


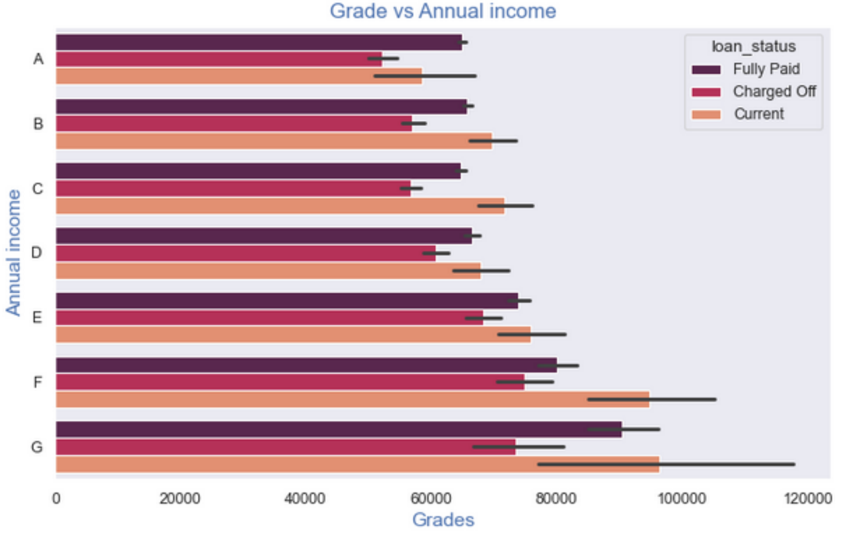
loan_status	annual_inc_cats	Charged Off	Current	Fully Paid	Total	Chargedoff_Proportion
0	0-20000	237	9	943	1189	0.199327
1	20000-40000	1514	170	7004	8688	0.174263
2	40000-60000	1729	345	9534	11608	0.148949
3	60000-80000	1024	240	6597	7861	0.130263
4	80000 +	1080	362	8531	9973	0.108292

Data Analysis

Bivariate Analysis

Bar plots for calculated data





- Observations:

- Income range 80000+ has less chances of charged off.
- Income range 0-20000 has high chances of charged off.
- Notice that with increase in annual income charged off proportion got decreased.

- From this we can conclude that the ones getting 'charged off' have lower annual incomes than the ones who paid fully' for each and every grade (i.e. at same interest range)

Conclusion

We come to an end of the EDA of the loan data set and finding some of the drivers for loan default. Apart from the ones highlighted below, I am sure there will be multiple others too; however, according to me, these are the most impactful ones.

Driving Factors for defaulting a loan

- Purpose of Loan
- Loan Paying term
- Loan Application Count
- Annual Income