# Lending Club Case Study

By Salman Ahmad

# Problem Statement

We have a consumer 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

Our goal here is to help this finance company to make good decisions to minimise the risk. This can be done by understanding and analyzing the historical data that this financial institution is willing to provide.

# Objective

The finance company has provided historical data which provides various inputs related to loans they have offered including few behavioural attributes of loan applicants. Data can looked at here and data dictionary which explains what each of the columns mean can be found here
The data provided consists of loan attributes like loan amount , funded amount , interest rate , issued data etc. The data also provides certain glimpse into customers demographic and financial situation using attributes like address , debt to income ratio, revol balance , publicly recorded bankruptcies , details related to delinquency in past 2 years etc. For each of loan application , loan status is also provided which denotes

1. Loan was fully paid
2. Current status i.e payments are still on going
3. Charged off i.e the loan applicant has default and is a loss to the bank.

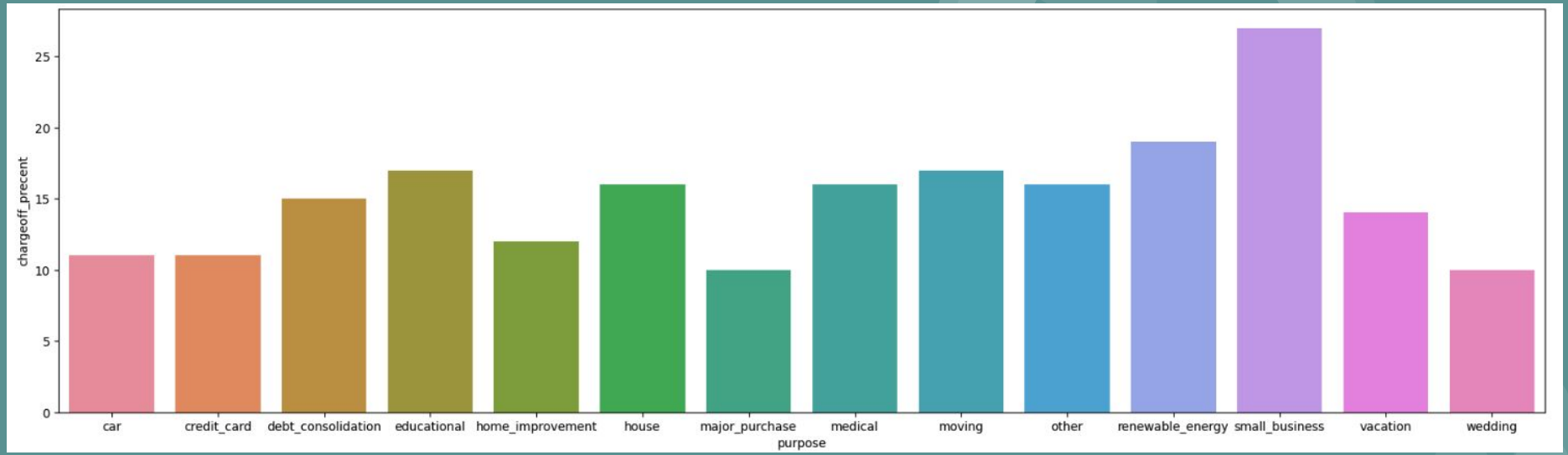
The prime objective of this assignment is the identify the patterns using Exploratory Data Analysis and help the bank to make profitable decisions.

# Approach

There are various steps in the approach followed

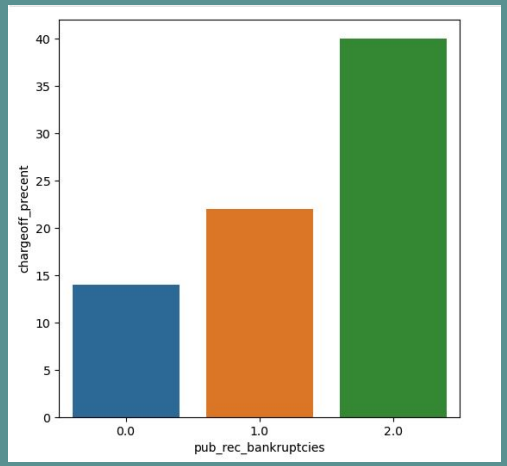
1. Data Loading and Exploration
2. Data Handling and Cleaning
    a. Identifying and dropping redundant columns
    b. Handling incorrect data type
    c. Handling columns with missing values
    d. Imputing data in case of missing values
    e. Data conversion for better analysis
    f. Outliers removal
    g. Dropping columns which are highly correlated
3. Exploratory Data Analysis
    a. Univariate Analysis
        i. Unordered Categorical Variables
        ii. Ordered Categorical Variables
        iii. Quantitative Variables
    b. Bivariate Analysis

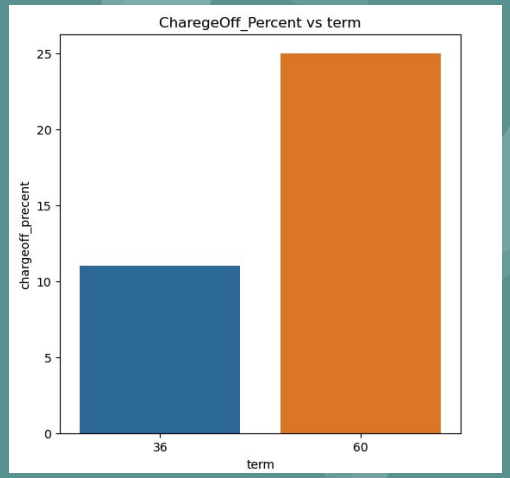
# Exploratory Data Analysis




Small Business charge off ratio is on higher end compared to other purposes.
Banks should be vary when providing loans to small businesses

# Exploratory Data Analysis

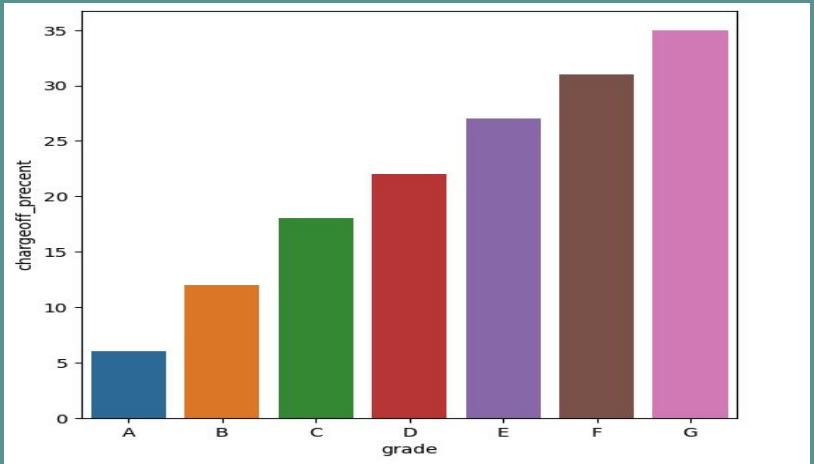


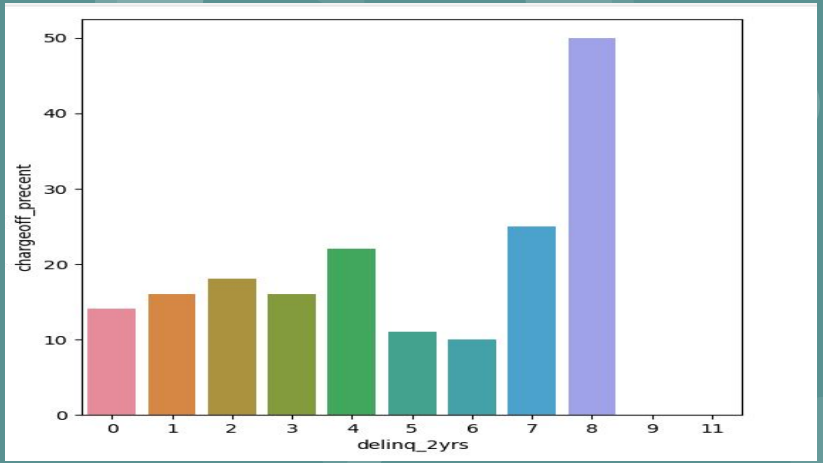

Charge off percent increasing with number of public bankruptcy record.
Banks should not provide loans to applicants with history of bankruptcy

Charge off percentage is more for loans of term 60 months.
Banks should evaluate properly when providing loans for 6o months term

# Exploratory Data Analysis

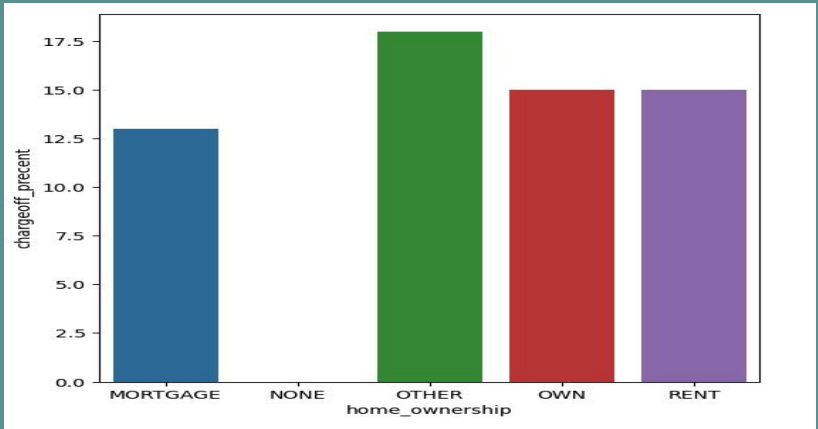



Charge off percent is more for loans of grades D,E,F,G
Banks should be wary when providing loans that fall into these grades




Charge off percentage is increasing with delinquency count.
Banks should avoid providing loans to applicants with delinquencies in their credit file.

# Exploratory Data Analysis



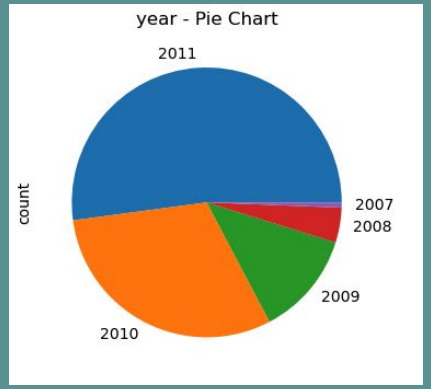

Charge off percent is bit less for applicants with home is mortgaged compared to rent.

# Univariate Analysis Summary

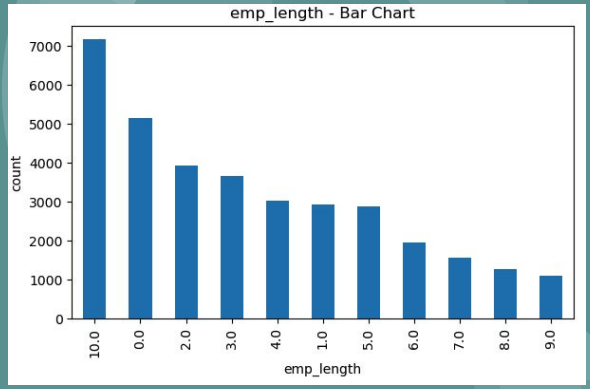
- Most of loans are for the purpose of deb_consolidation followed by credit_card.
- Most of the loan beneficiaries have home rented or mortgaged.
- Most of loan applications are in CA followed by NY.
- Most of the loan applications are with term as 36 months.
- Most of the loan applications fall in B and A grades.
- Most of the loan applicants have emp_length of 10+ years, followed by 0 and 2 years.
- Most of the loan applicants don't have any public record of bankruptcies.
- Most of the loan applications are towards the end of the year in the months of Oct/Nov/Dec. Then lowest number of loan applications are in first quarter of the year Jan/Feb/March.
- Loan applications increasing are over the years.

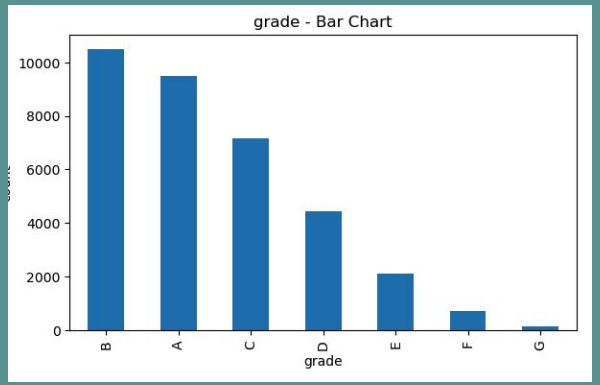
** All these insights are derived from data can be seen here in detail

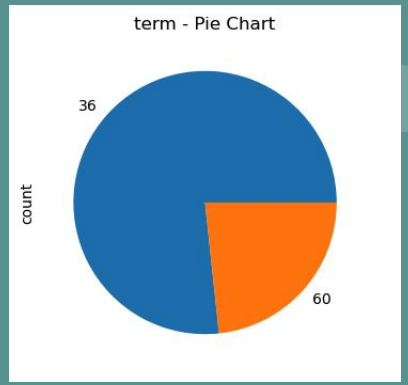
# Univariate Analysis Summary



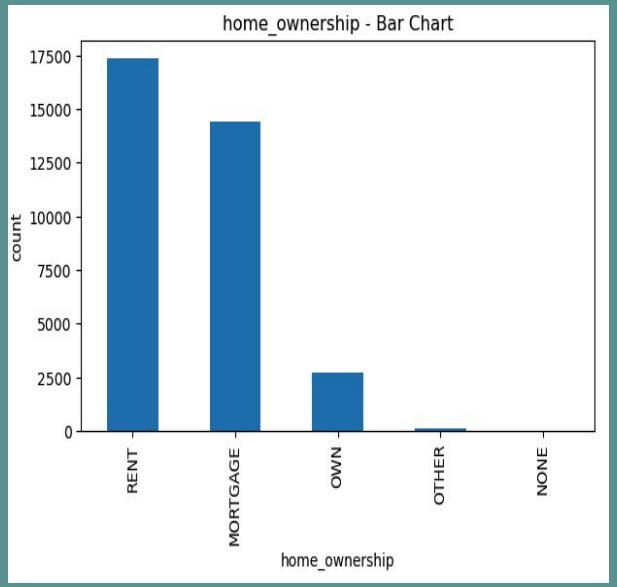


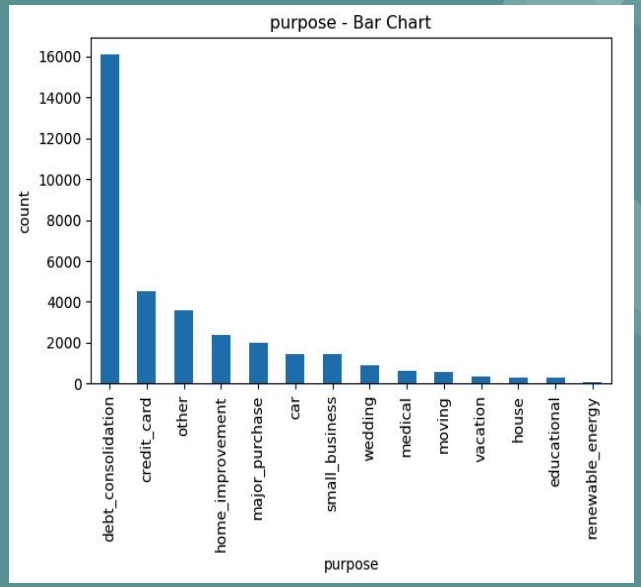


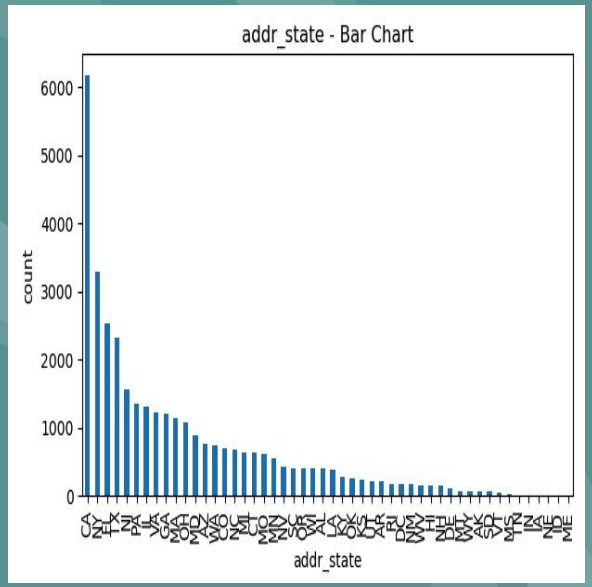



# Univariate Analysis Summary



# Data Handling And Cleaning

- Rows where loan_status is current means that loan payments are still in progress and we can not make any conclusions based on that data. Hence dropping those rows.
- There are few columns where there is no data at all , all the the values are NA. So they can be dropped from further analysis.Here is the list `next_pymnt_d', 'mths_since_last_major_derog', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit`

** All these insights are derived from data can be seen here in detail

# Data Handling And Cleaning Cntd..

- There are few columns which have a different value for each row, such columns don't contribute to the analysis and hence dropped. Here is the list `member_id,url`. Id also falls in this category but we are not dropping it.
- There are columns which are descriptive and textual in nature . They don't contribute to current analysis and hence dropped. One such column we have is `desc`
- There are few columns which has NA values more than 40% . Hence dropping them. Here is the list 'mths_since_last_delinq','mths_since_last_record'
- There are few columns where all the values are 0. Hence dropping them.
- There are few columns where few values are NA and rest all values are 0.Hence dropping them. Here is the list 'collections_12_mths_ex_med','chargeoff_within_12_mths','tax_liens'
- There are few columns where there is only one unique value and rest all are NA. These columns won't contribute to current analysis and hence dropping them. Here is the list pymnt_plan,initial_list_status,policy_code,application_type

** All these insights are derived from data can be seen here in detail

# Data Handling And Cleaning Cntd..

**Handling Incorrect Data Types**

- loan_amnt and funded_amnt are to be changed to float.
- term column has months suffix. The suffix is removed and converted to int.
- int_rate has % suffix. The suffix is removed and converted to float.
- revol_util has % suffix. The suffix is removed and converted to float.
- issue_d is converted to data time and formatted.
- emp_length columns has value like <1 year, 2 years, 3 years etc. This converted to a numeric values <1 year mapped to 1, 2 years mapped to 2 and 3 years mapping to 3. Similar all other values are mapped to a numeric value.

** All these insights are derived from data can be seen here in detail

# Data Handling And Cleaning Cntd..

**Imputing**

- emp_length column has few NA values. It is imputed with 0 years are experience.
- pub_rec_bankruptcies has few NA values . It is imputed with 0 value.

** All these insights are derived from data can be seen here in detail

# Data Handling And Cleaning Cntd..

**Outliers Removal**

Any values which are falling outside of [1st Quantile - 1.5 * IQR, 3rd Quartile + 1.5 * IQR ] are considered as outlier. Following columns underwent outlier removal

1. 1089 are dropped due to outlier in loan_amnt column.
2. After this further analysis revealed that there are 161 rows which has outliers in funded_amnt column and hence dropped.
3. After this further analysis revealed that there are 102 rows which has outliers in funded_amnt_inv column and hence dropped.
4. After this further analysis revealed that there are 62 rows which has outliers in int_rate column and hence dropped.
5. After this further analysis revealed that there are 1028 rows which has outliers in installment column and hence dropped.
6. After this further analysis revealed that there are 1570 rows which has outliers in annual_inc column and hence dropped.

** All these insights are derived from data can be seen here in detail

# Data Handling And Cleaning Cntd..

**Dropping columns based on correlation**

There are few columns which are highly correlated and some of them can be dropped

1. funded_amnt is highly correlated with loan_amnt . Hence dropped.
2. funded_amnt_inv is highly correlated with loan_amnt. Hence dropped.

** All these insights are derived from data can be seen here in detail

# Thank You