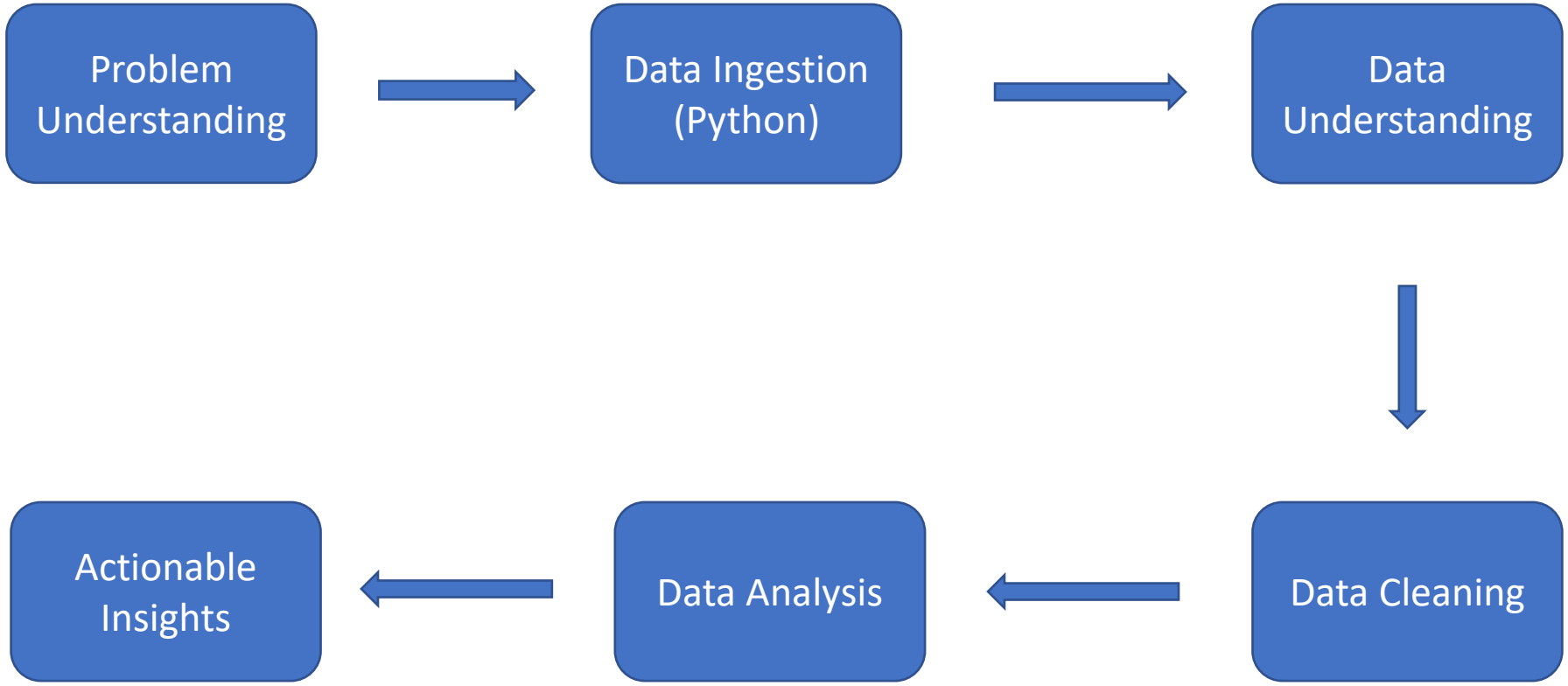
Lending case Study

By:

- Tushar Tyagi

- Pawan

Project Workflow



Problem Statement

Understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. These indicator can be utilized for its portfolio and risk assessment of new loan applicants.

Data Ingestion

- There were two important files:
  - Loan.csv – This file has all the tabular data used for data analysis
  - Dictionary.xlsx – It has one line definition of all the columns in loan.csv file

Data Understanding

- **Data size:** 39K rows and 111 columns
- Found quality issue with the below analysis:
  - **Missing values:** There were 68 columns having null values. There are few columns having 100% blanks like bc_open_to_buy, acc_open_past_24mths, bc_util, etc.
  - **Outlier analysis:** IQR analysis is used to identify outliers and based on the same we found 22 columns having multiple outlier values
  - **Variance analysis:** This technique is used to determine if we have features which doesn't add values or having one unique values in all rows. Few of zero variance columns were pymnt_plan, initial_list_status, application_type, etc.

Data Cleaning and Manuplation

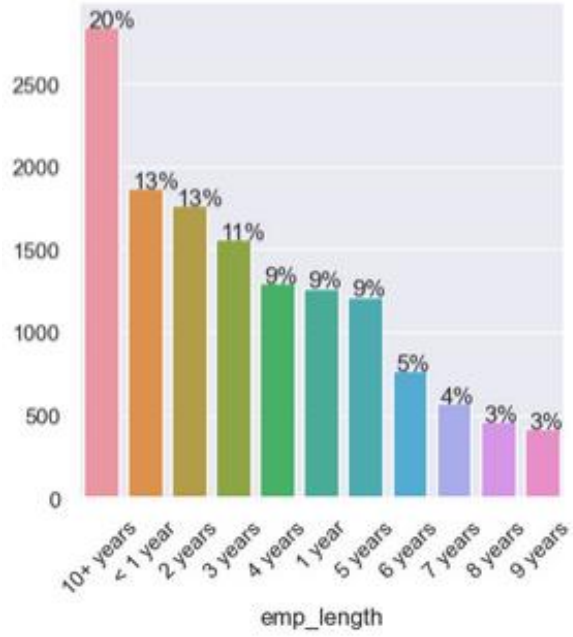
- **Treating blank/null values:**
  - High blank values (>90%): Deleted columns having more then 90% blanks. Like bc_open_to_buy, acc_open_past_24mths, bc_util
  - Low blanks:
    - Categorical columns: Replaced blanks with mode (most frequent value) of the column
    - Numerical columns: Replaced with mean/average value
  - Benefit: Dropped 56 columns having 90% blank
- **Treating outliers:**
  - IQR: Used IQR statistical technique to remove outliers
  - Benefit: Reduced 35.43 % of outlier values using IQR
- **In-scope:**
  - Dropped data regarding “Current” or on-going loan applicants
- **Correcting data types:** There were few columns having incorrect types. Like Interest rates were given as categorical but they were more insightful when converted to float.

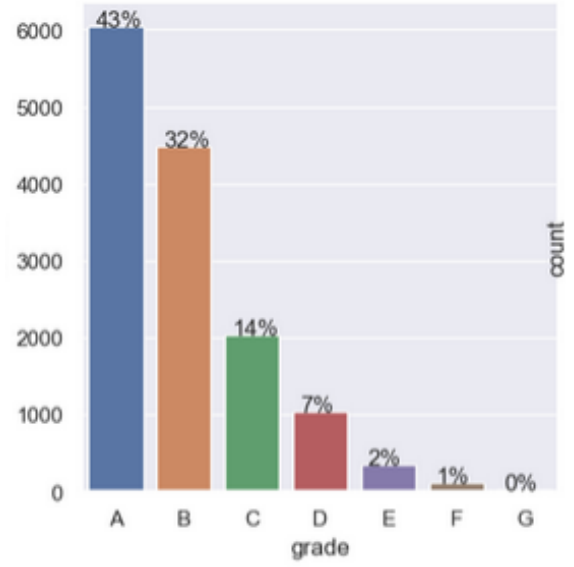
Data Analysis (Univariate analysis - Highlights)

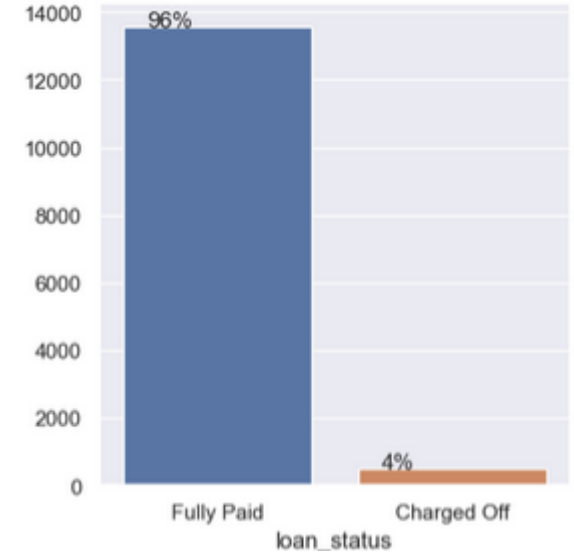
Univariate analysis helps us understand about the data using one variable at a time. We got the following insights:

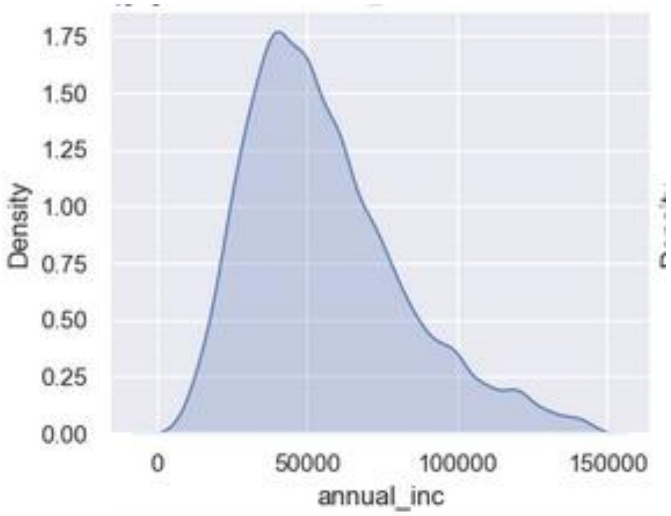
- Loan status: Data has 96% of data of fully paid applicants and 4% of Charged off
- Term: Majority (85%) of applicants belongs to 36 months term
- Home ownership: Rent (54%) and Mortgage (38%) together accounts to 92% of total applicants under home ownership
- Loan grade: A (43%)and B (32%) accounts to 75% of applicants
- Employee tenure: 10+ years applicants are high (20%) and then less then 1 years are 13% of the data
- Purpose: Majority of applicants raised to loans to cover debt (45%) and second highest is to clear credit card amount (15%)
- Interest rate: Most applicants has between 6-8%
- Loan amount: Data is skewed around 30k amount. The highest amount goes to 150K
- Income: Data is skewed around low income group (5k). The highest amount goes to ~27K
- DTI: The majority of applicants are b/w 11-14

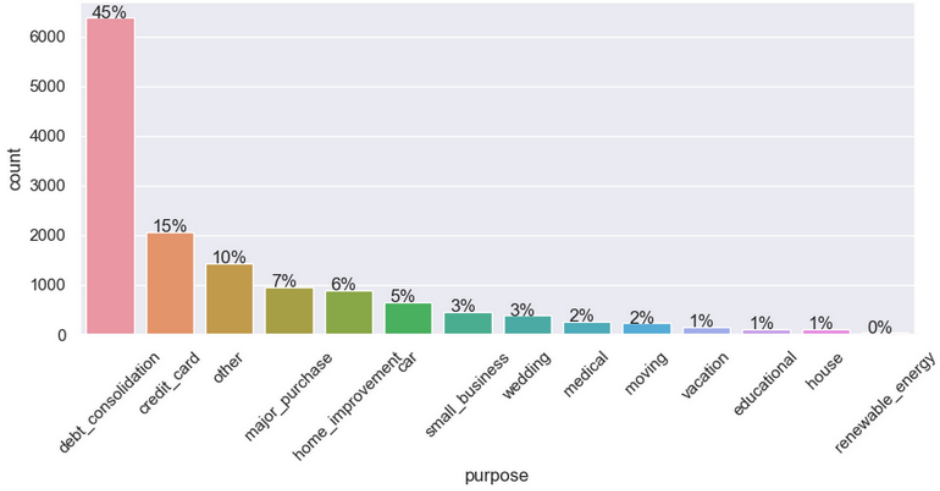
Data Analysis (Univariate analysis - Graphs)

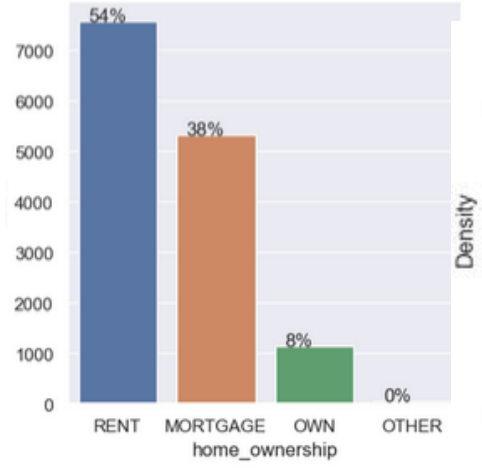


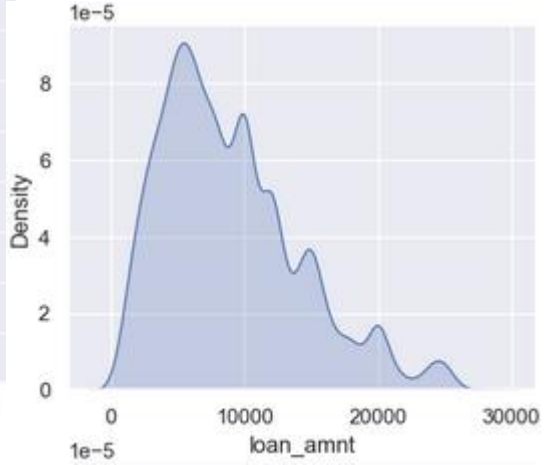


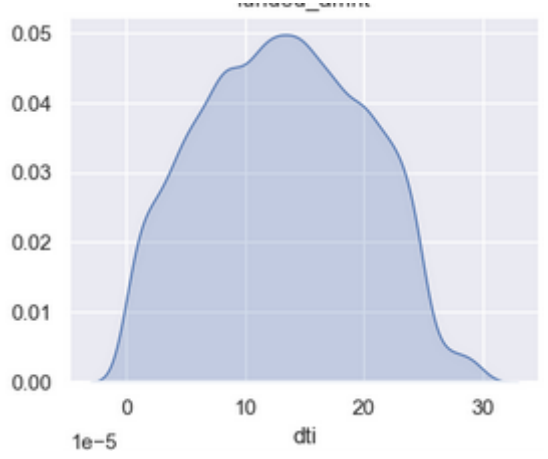












Data Analysis (Bivariate analysis - Highlights)

- Term: High defaulters are in 60 months (62%)
- Grade: B (35%) and C (22%) has high defaulters compare to other grades
- Home ownership: Rent (47%) and Mortgage (44%) has high defaulter
- Purpose: Debt consolidation (51%) and Other (13%) of defaulters
- Loan amount: Defaulters has higher amount of loans. Defaulter mean 10k and non-defaulters has 8.8k
- Annual income: Defaulter has lower income compare to non-defaulters. Mean of non-defaulter is 50K and non-defaulter 56K
- Installments: Defaulters has high installments. Mean of non-defaulter is 267 and non-defaulter 297
- DTI: DTI is high for defaulters. Mean of non-defaulter is 13 and non-defaulter 15

Data Analysis (Bivariate analysis - Graph)