EDA CASE STUDY

Group Members:

1. Varsha Venkapally
2. Ashish Kumar Korukonda
3. Shravani Kothur
4. Chayan Naskar

AGENDA

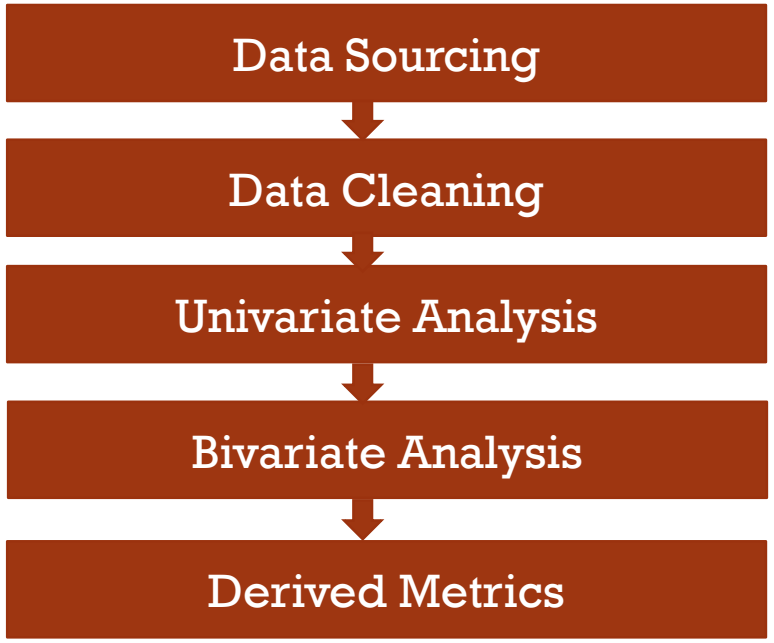
We were given Consumer Finance company dataset which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

As an data analyst , we are supposed to find out which applicant is not likely to repay the loan, which may lead to financial loss of the company and our aim here is to infer various factors as to which applicant is likely default the loan. Our inferences helps the company to predict the defaults before hand to avoid future loses.

EXPLORATORY DATA ANALYSIS (EDA)

We would be looking into the dataset provided and use EDA approach to get insights from the data and accordingly find out the factors as to which applicant is likely to default and provide various recommendations to approve applications for a loan request.

Below are the steps we can follow to get insights from the data:



DATA SOURCING

- As data is the key : **the better the data, the more insights you can get out of it.**
- So the dataset given is of a consumer finance company which contains data about various types of loans given to the customers and various metrics related to the customer credit background.
- This dataset will help us derive decisions on avoid approving loan applications which may lead to defaulters.
- Firstly we will be reading the data into our R console using below command read.csv()

```
loan <- read.csv("loan.csv", header=T, na.strings=c("","NA"))
```

  - We have read the csv formatted file into a data frame in R
  - We have used "na.strings" which will replace blank values with NA's in the whole data set.

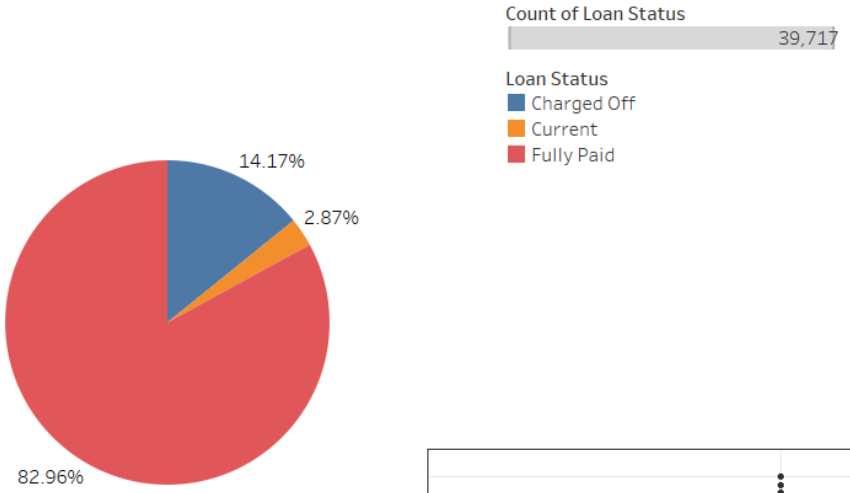
DATA CLEANING

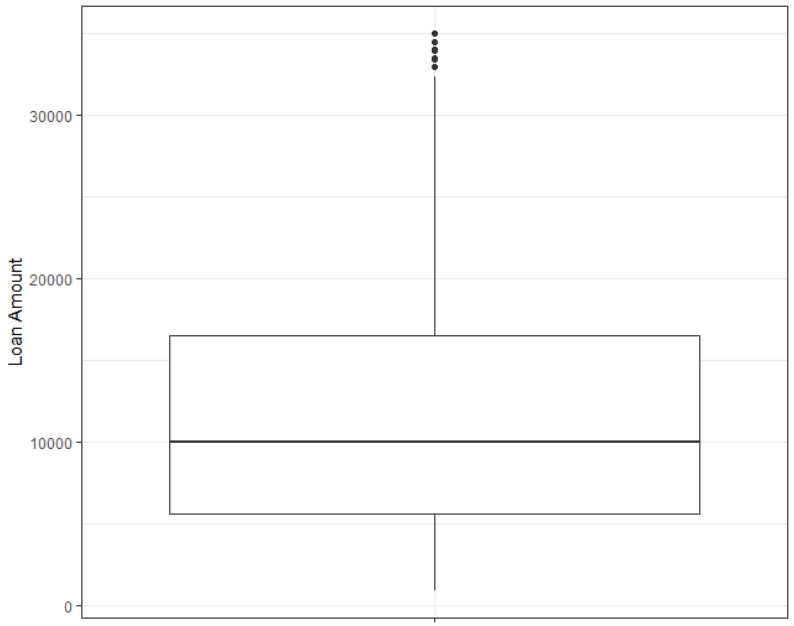
Dataset contains various metrics like loan amount, interest rates, grades etc, our next step would be to do data cleaning in order to perform our analysis on the dataset.

Below are the steps we followed to clean the data:

- Check for duplicates on id and member id fields, as these are the columns which uniquely identify the loan applications.
- Fixing missing values/Rows and columns
  - Identifying & deleting columns which have only NA values.
  - Replace the blank value with NA's
  - Identifying & deleting the columns which has just 1 factor values.
  - Removing unnecessary columns which are not required for our analysis.
- Standardise values
  - Correct wrong structure, convert incorrect data types.
  - Standardise precision (ex: 2.10347 to 2.10).
  - Remove extra characters(ex: in term columns remove "years" text).
  - Rectify In-correct data types.

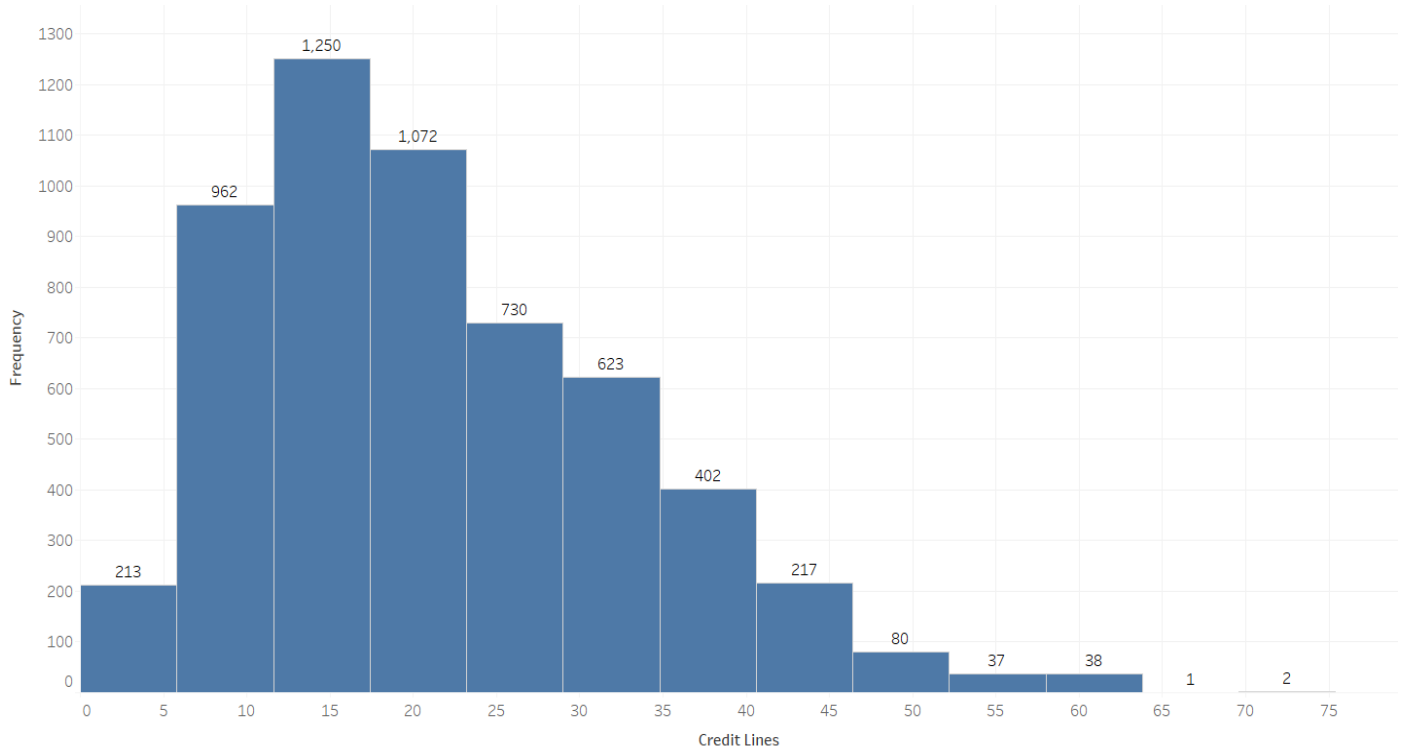
DATA ANALYSIS : DATA DISTRIBUTION





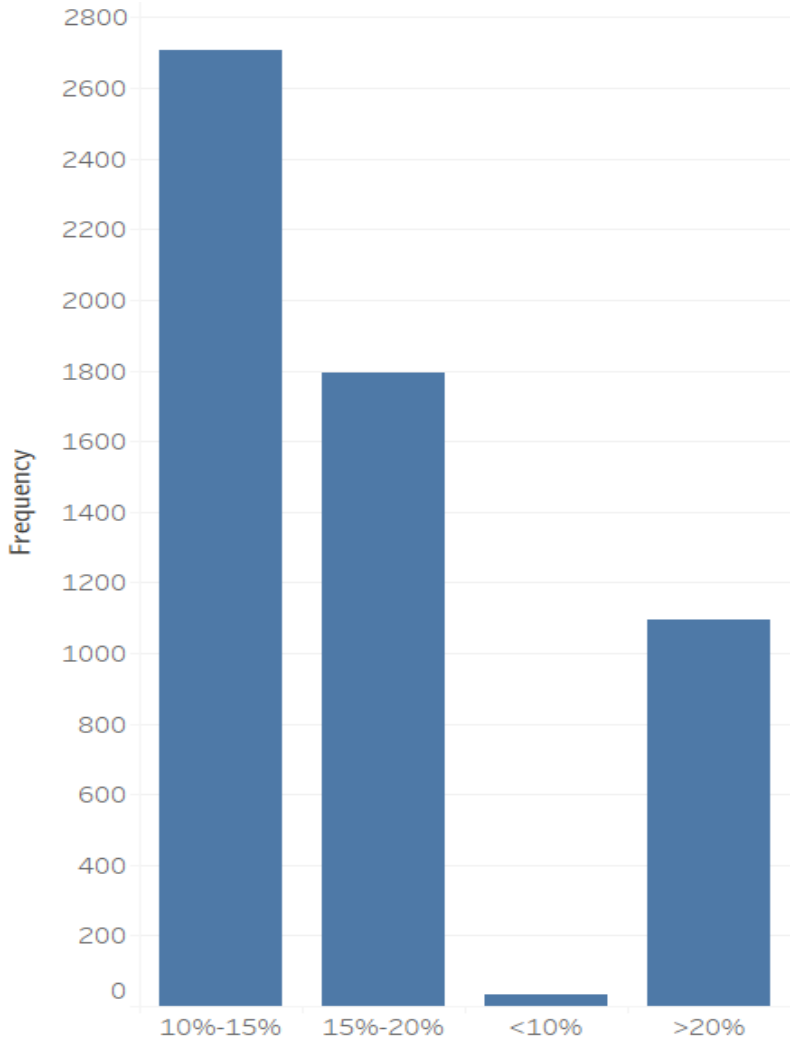
- After data cleaning, in order to get a feel of the data set we see the data distribution and grab insights from the graphs plotted.
- Plotted a pie chart which shows the loan status distribution and we see 82.96% of fully paid customers,14.17% of Charged Off customers and 2.87% Current loan status. Since our business problem is to find the strong reasons for defaults of the loan therefore we have analyse on 14.17% of the data set.
- A box plot is created to show the distribution of loan amount taken by the customers. We can see that the mean loan amount taken is $10000.

DATA ANALYSIS : UNIVARIATE ANALYSIS - I



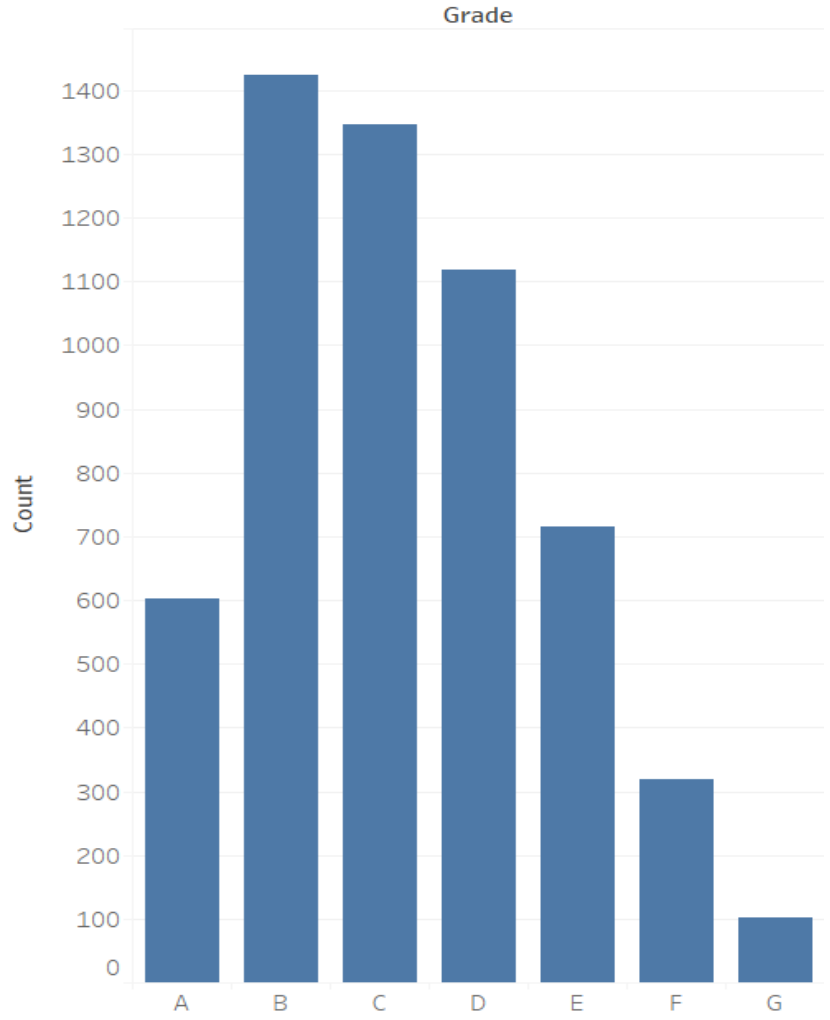
- As part of Univariate Analysis we have created a histogram to understand the distribution of **Credit lines** of all the customers.
- As Credit lines are the available limit for the customer to borrow money therefore less credit lines indicate that the customer has less limit to borrow the money.
- The data in the plot is filtered out to infer only the Charged off loan status – to understand the likeliness of defaulters with respect to no. of credit lines.
- As you can see in the plot – people having less credit lines (ranging from 5-25) are likely to default when compared to the higher credit lines.

DATA ANALYSIS : UNIVARIATE ANALYSIS - II



- In Continuation of Univariate Analysis, we have created a bar graph which shows the frequency of **interest rates** for all the defaulters provided in the dataset.
- X-axis in the plot tells us the range of interest rates given on a loan. The ranges are divided into segments for better understanding.
- The data in the plot is filtered out to infer only the Charged off loan status – to understand the likeliness of defaulters with respect to interest rates.
- We can understand from the plot that customers who have taken loan with interest rates ranging from 10-15% are highly defaulting the loan and they are putting the company as risk.

DATA ANALYSIS : SEGMENTED UNIVARIATE ANALYSIS



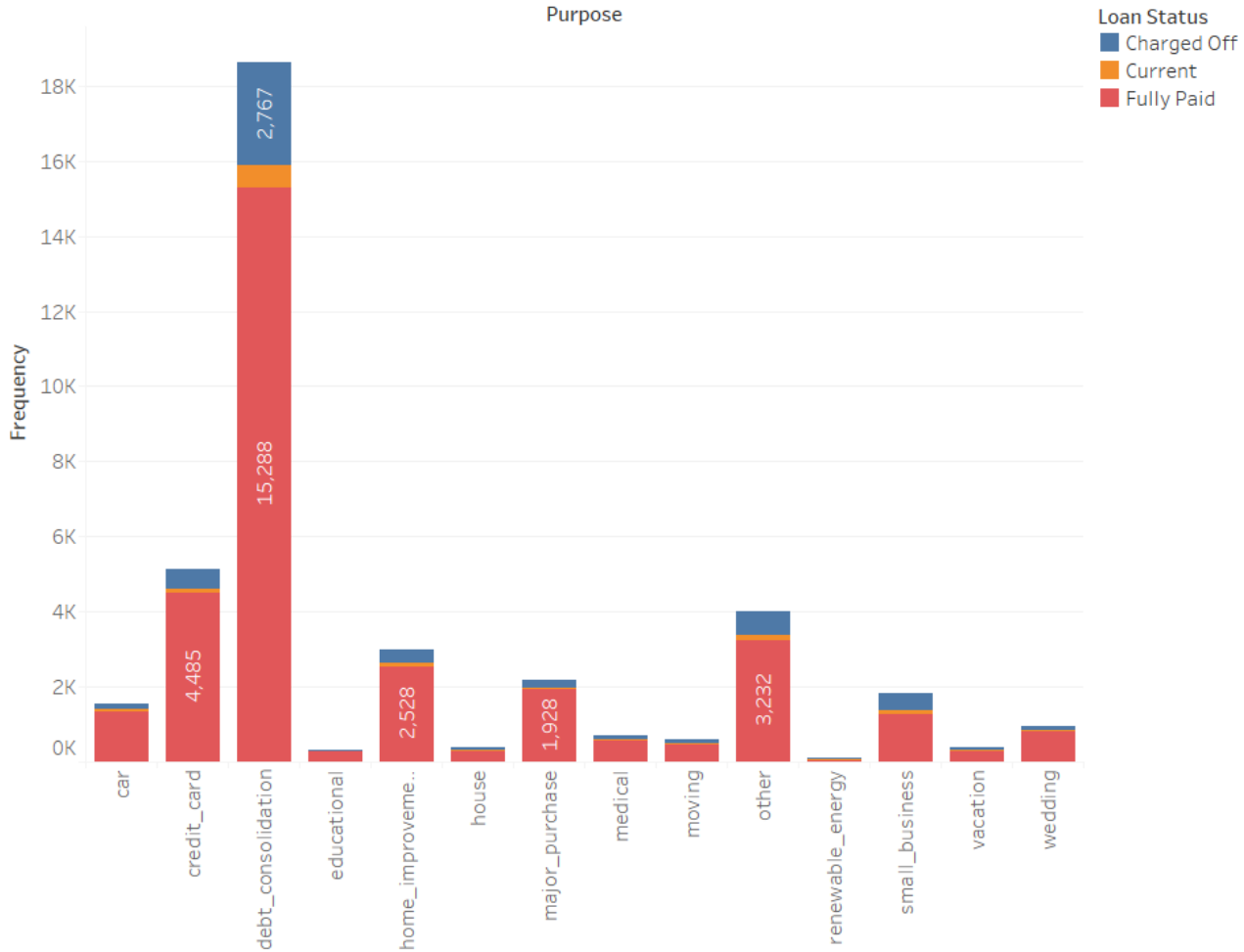
- Segmented Univariate analysis is done on a single variable by segmenting it into one category i.e. we are inferring data only for Charged off loan status.
- The data in the plot is filtered out to infer only the Charged off loan status – to understand the likeliness of defaulters with respect to **grades**.
- From the plot we can understand that customers having B & C grades (comparatively lower risk) tend to default higher than the customers with other grades.

DATA ANALYSIS : OTHER KEY INFERENCES

In similar fashion Univariate/Segmented Univariate analysis can be done on below variables to infer the default rates of the customers.

- Segmented Univariate Analysis is done on Customer's years of employment experience and thereby inferring that customers who are experienced more than 10 years are likely to default than other categories.
- Segmented Univariate Analysis is done on Customer's Home ownership and we can understand that customers who live on rent or pay mortgage have the highest default rate when compared to Owners.
- Analysis performed on States can give the state where the default rate is higher – that turned out to be California. Also, Florida & New York are marginally higher than the other states.

DATA ANALYSIS : BIVARIATE ANALYSIS



- Bivariate Analysis is done by comparing the variables - Purpose of taking the loan and the loan status.
- On the X-axis we have the Purpose and the Y-axis we have frequency distribution.
- We can see that the category "debt consolidation" is higher in all the categories of loan status.
- We can also infer that the customers who lake loan for the purpose of "debt consolidation" have the higher default rate when compared to other categories.

TOP 5 DRIVER VARIABLES TO INDICATE THE HIGH DEFAULT RATE:

No	Driver Variable	Inference Drawn
1.	Total_acc(Credit lines)	Customers having less credit lines (ranging from 5-25) are likely to default when compared to the higher credit lines.
2.	Interest Rates	Customers who have taken loan with interest rates ranging from 10-15% are likely to default the loan when compared to others.
3.	Grades	Customers having B & C grades (comparatively lower risk) tend to default higher than the customers with other grades.
4.	Purpose	Customers who take loan for the purpose of "debt consolidation" have the higher default rate when compared to other categories.
5.	Home ownership	Customers who live on rent or pay mortgage have the highest default rate when compared to Owners.

RECOMMENDATIONS

- Based on the inferences below are the recommendations provided :
  - As inferred, there is a higher default rate of the customers who's purpose of taking loan is "debt consolidation" , therefore if the purpose of the loan is debt consolidation, the approver much perform rigorous background checks on the debts of the customer.
  - If the credit lines of the customer is comparatively lower then the loan approver should check for his previous debt's status and the transactions performed to make sure that the loan given to him is not likely to be defaulted.
  - We found that most of the defaulters fall under interest rate of 10-15% & the Grades and sub grades of the defaulters are B & C which computes the same interest rate, therefore considering the factors such as credit scores, grades, etc will help to predict if the customer is likely to default or not.
  - Other interesting fact we found is that – customer's who are paying rent or are on mortgage have a high tendency to default the loan. Therefore factors such as annual income and credit lines should be monitored to predict the likeliness of loan default.
  - Another key factor to identify the defaulters is DTI (Debt to Income ratio) – If the DTI is higher for the customer then he/she is likely to default the loan. Depending on DTI one can decide if the customer is to be granted a loan or grant a load with high interest rate to avoid any losses.