Lending Club Case Study

Prepared by :
Sandip Mahato
Subhadip Ghosh

Date: 7 Feb 2023

Business Problem Statement:

- **What:** The Lending Club company want to understand the **driver factors behind loan default**, i.e. the variables which are strong indicators of default.
- **Why:** The company can utilise this knowledge for its portfolio and risk assessment. If risky loan applicants can be identified, then such loans can be reduced thereby cutting down the amount of credit loss.
- **How:** Doing EDA on the complete loan data for all loans issued through the time period 2007 t0 2011

Our Interpretation of the business problem:

- The Target Variable is Loan_Status alone, as this the only attributes that identifies which loans were Defaulted or not.
- We need to analyze the past data to identify what all other variables influnce the target variable strongly.
- We need to indentify the likelyhood of Applicants going to default, prior to the loan is approved.

Our Approach to Analysis:

- Our objective is to predict the chance of defaulting prior to loan disbursement, hence loan’s operational attributes (e.g. total_pymnt, total_rec_late_fee) will be excluded from analysis.
- The records for ongoing loans(i.e. loan_status=’Current’) will be excluded for analysis. Because those loans are already disbursed, active and the outcome(i.e. Default or Not Default) is unknown.
- The LC assigned Grade/ Sub-Grade have significant impact on Default. However, we are not including this in our conclusion, because this not new information for the Lending Club business. Objective is to find other attributes.

Steps for the Data Analysis:

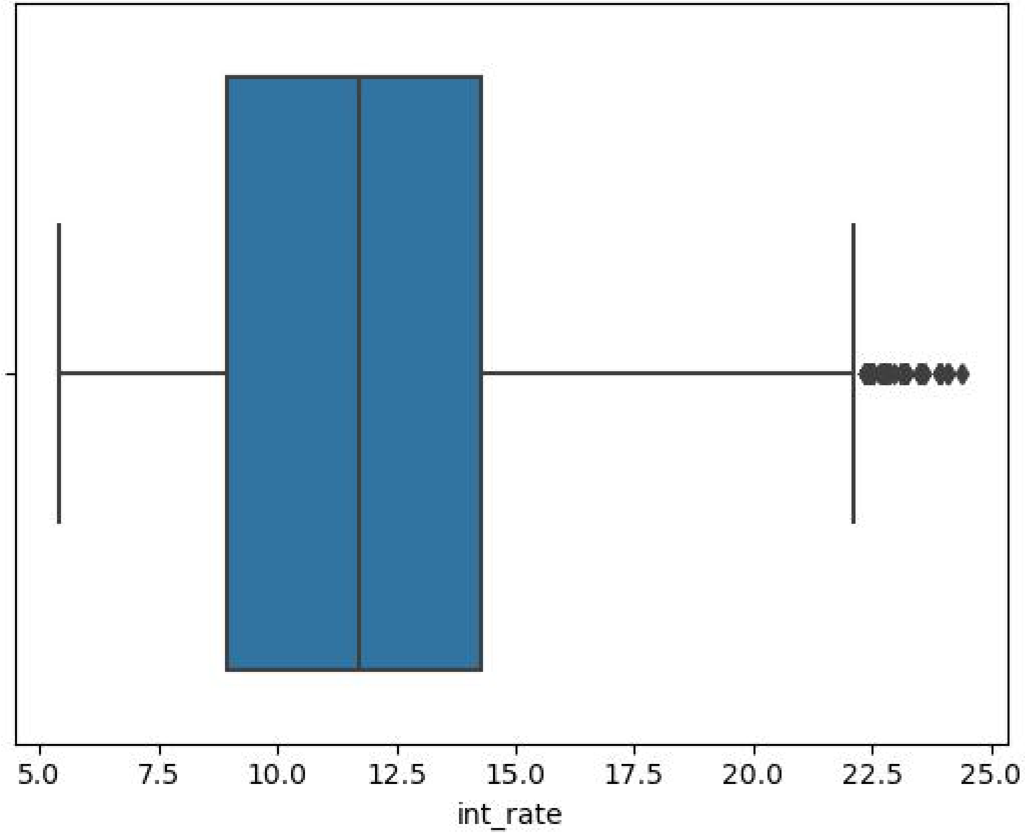
1. Understand the data
    1. Going through data dictionary
    2. glance through the data
2. Clean the data
    1. Drop junk data i.e. duplicate rows and columns with high nulls/NA/0
    2. Remove rows with outliers, columns with constants
    3. Fill missing values, Standardize values
    4. Filter out irrelavant data rows and columns from business perspective
3. Univariate Analysis
    1. Ordered Categotical Univariate Analysis
    2. Unordered Univariate Analysis
    3. Derive Columns as necessary
4. Bivariate Analysis
5. Conclusion

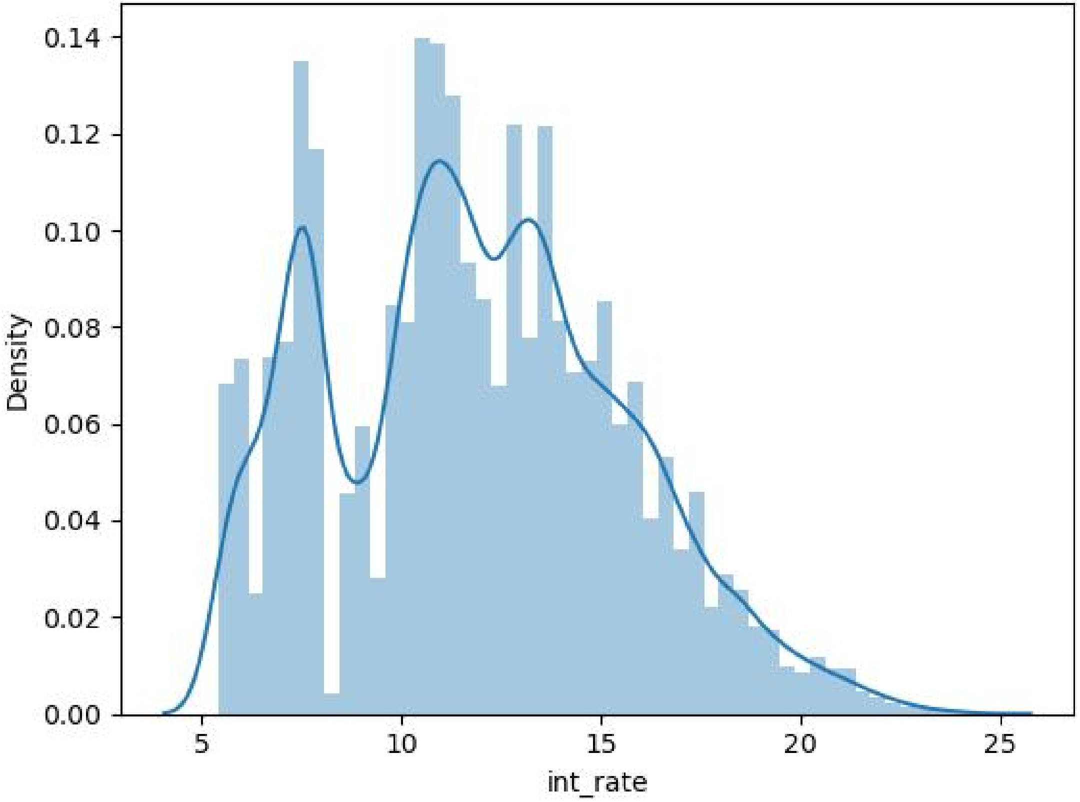
Univariate Analysis - Approach:

- Check distribution and skewness of relevant columns (following slides)
- Standardize the values of columns (e.g. term, emp_length, int_rate, revol_util etc.)
- Remove outliers from columns (e.g. annual_inc)
- dropping NULL columns (e.g. out_prncp, out_prncp_inv)
- Fill missing values with appropriate values

Univariate Analysis - Result:

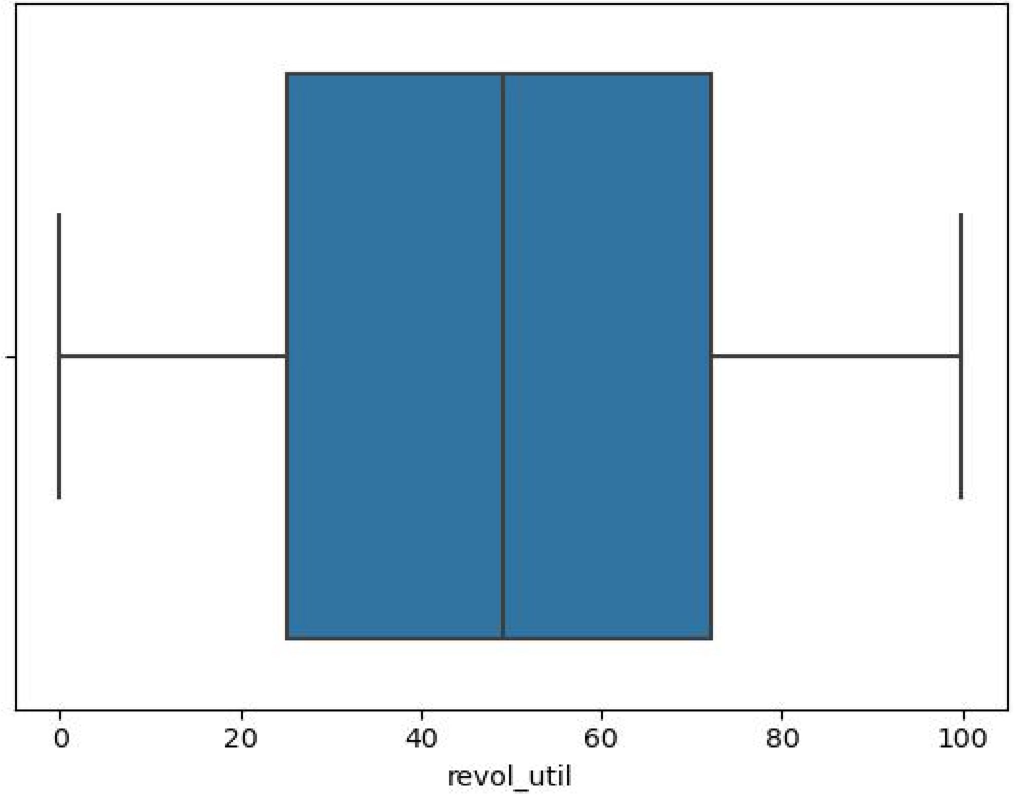
- Interest Rate: The data is skewd to the left, hence median will be better indicator than mean

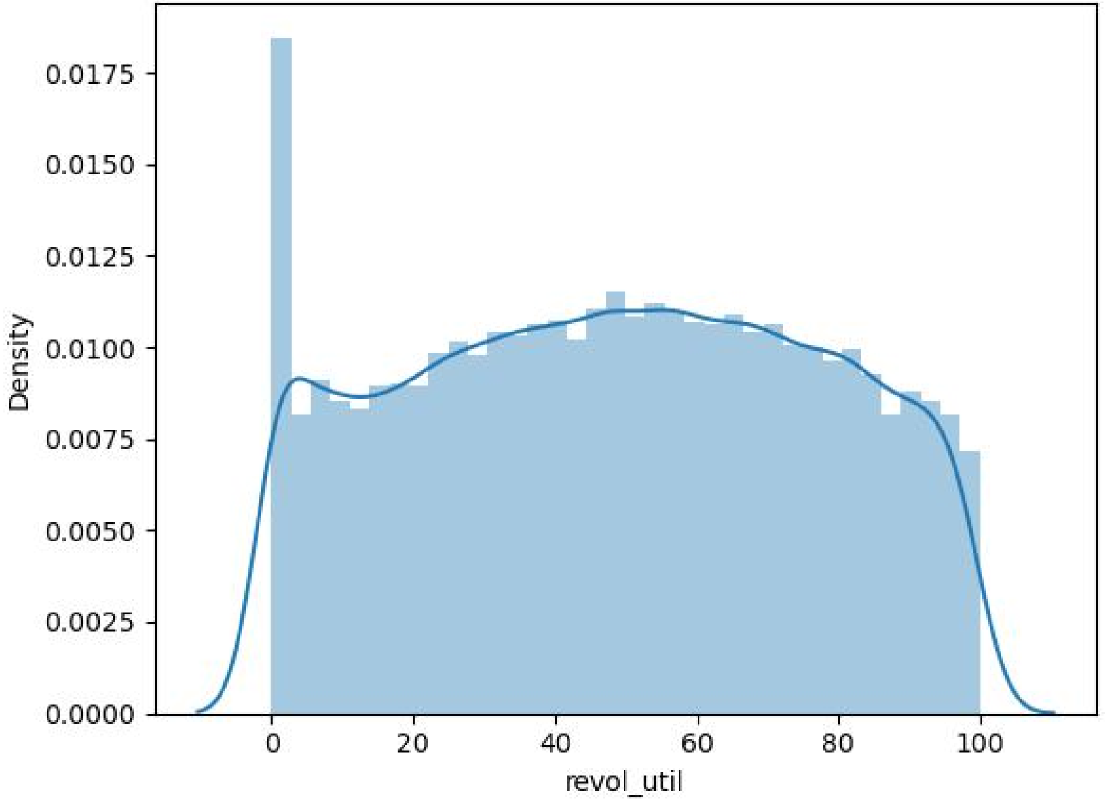




Univariate Analysis (continued:)

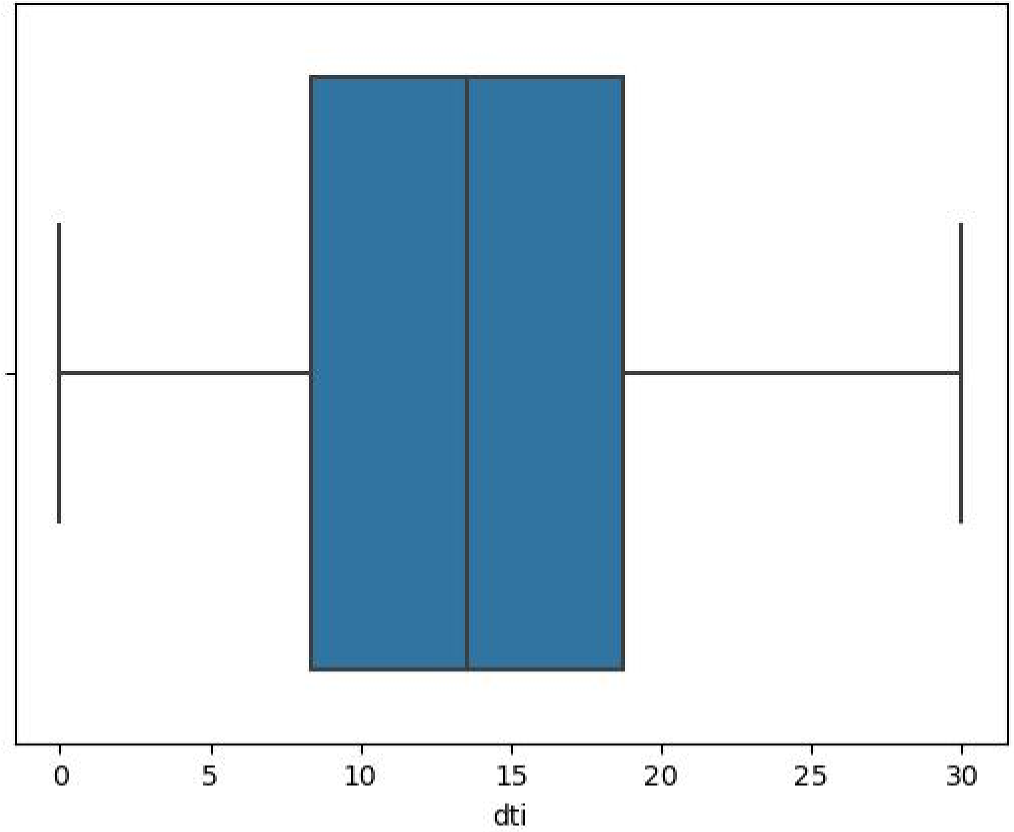
- Revolving Credit Utilization: The data is evenly distributed hence mean can be considered

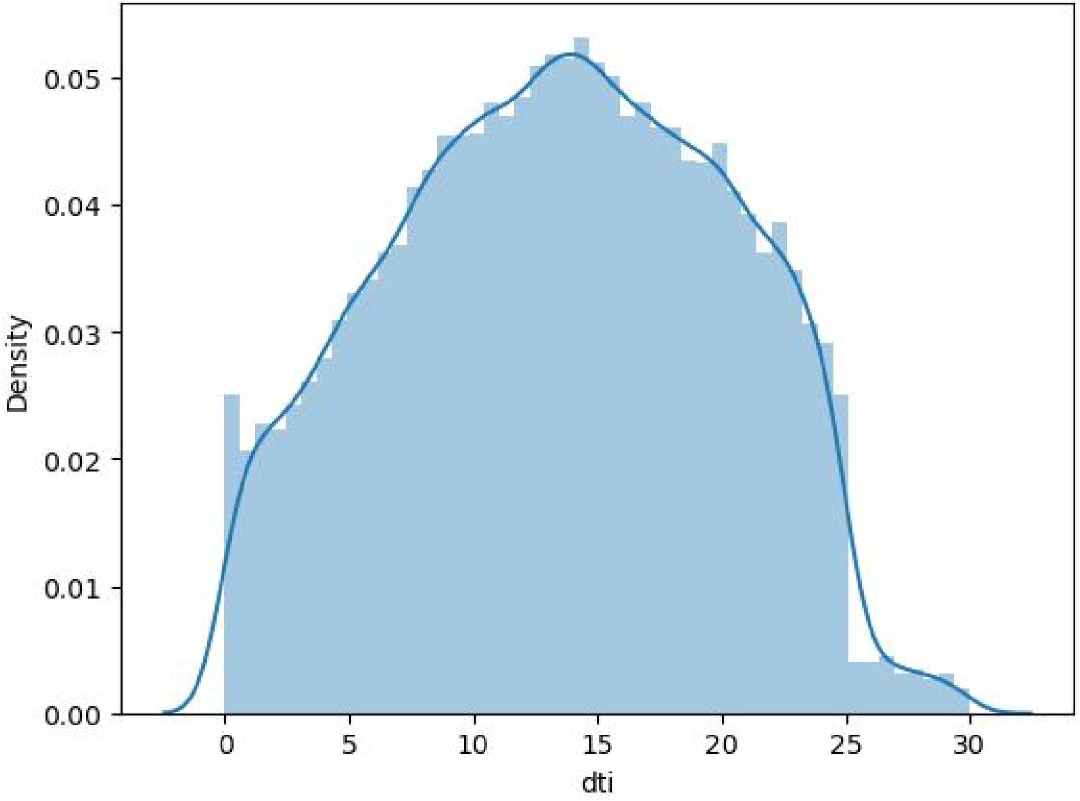




Univariate Analysis (continued)

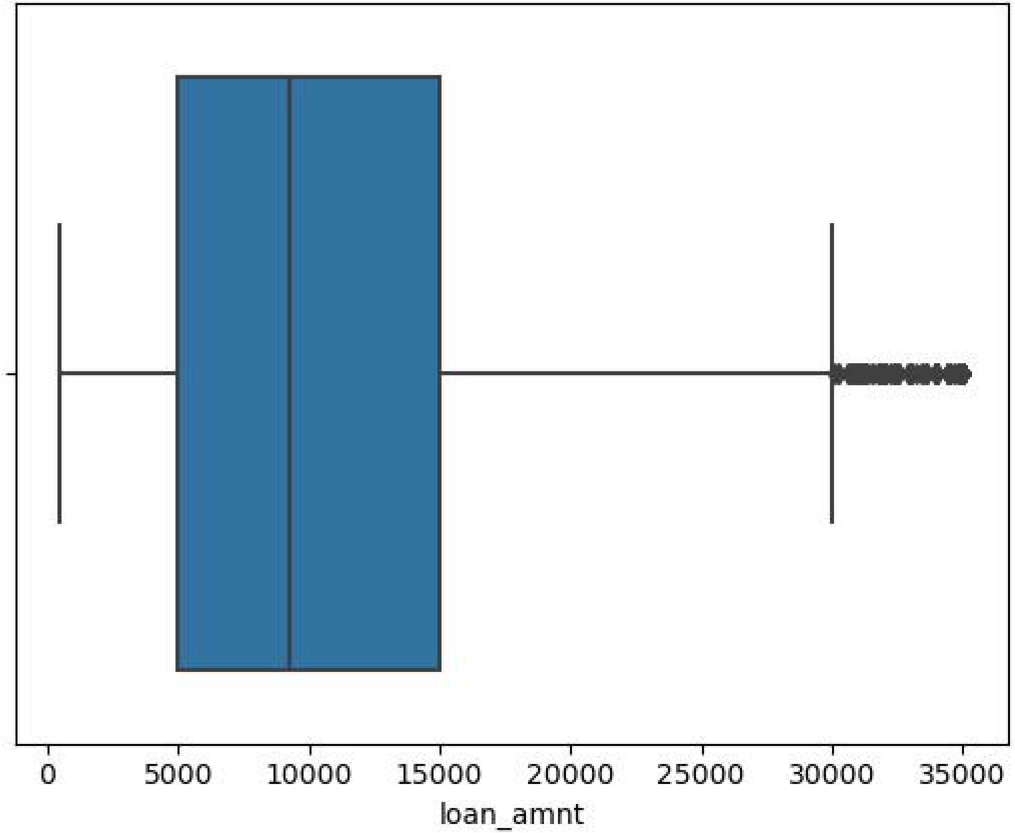
- DTI (Debt to Income Ratio): The data is evenly distributed hence mean can be considered

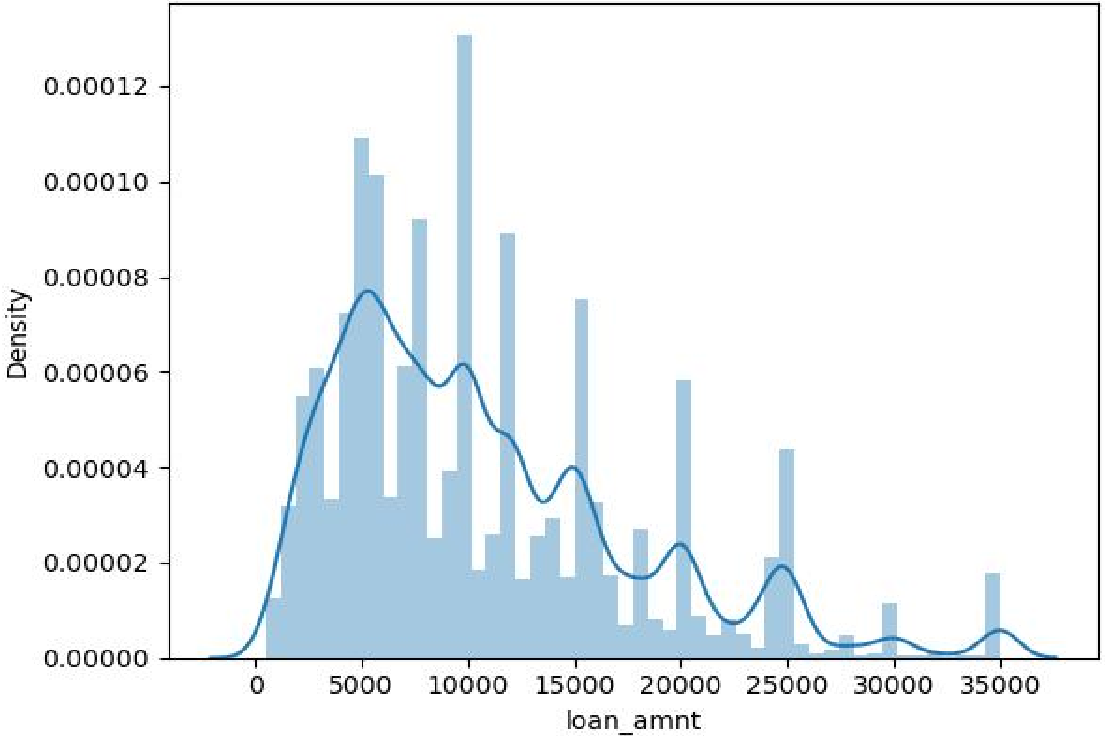




Univariate Analysis (continued)

- Loan Amount: The data skewed to left, so median will be a good representation





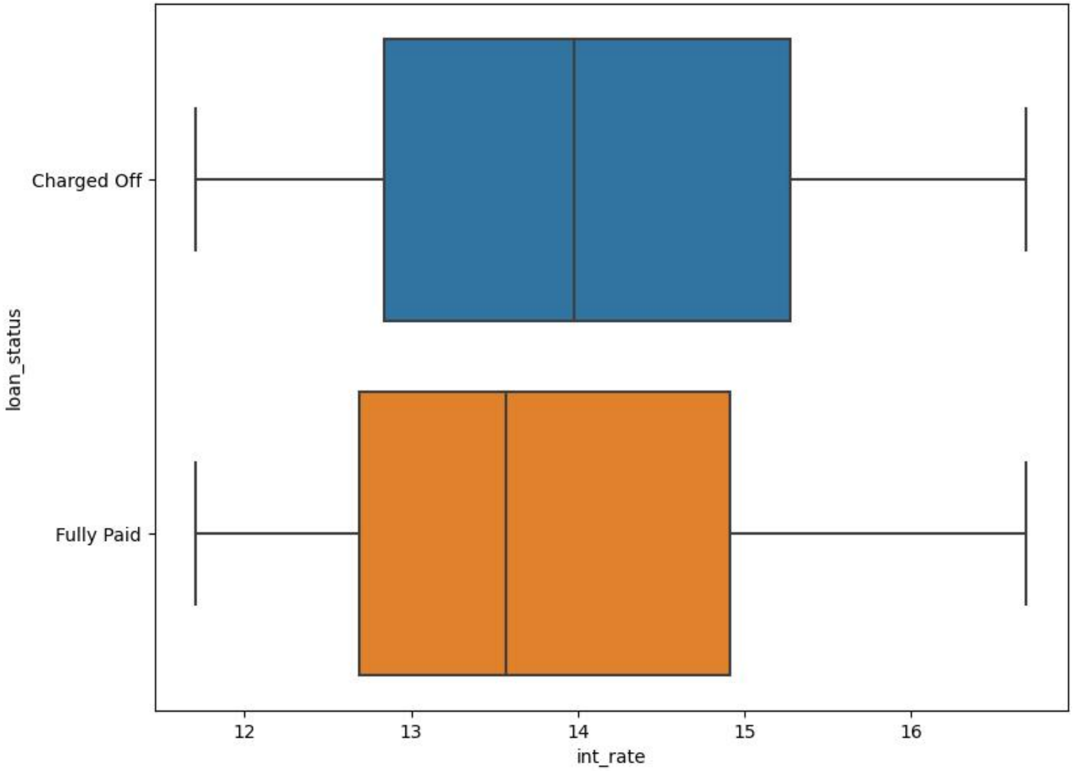
Segmented Univariate Approach:

- Each Continuous variable is anlyzed with respect to loan_status to understand their pattern and impact.
- Each Categorical variable is anlyzed to identify the impact (if any) it has on loan_status
- The variables with higher impact are highlighted in slides, details avaialble in notebook file.

Segmented Univariate Result:

- Impact of Interest Rate on Defaulting:

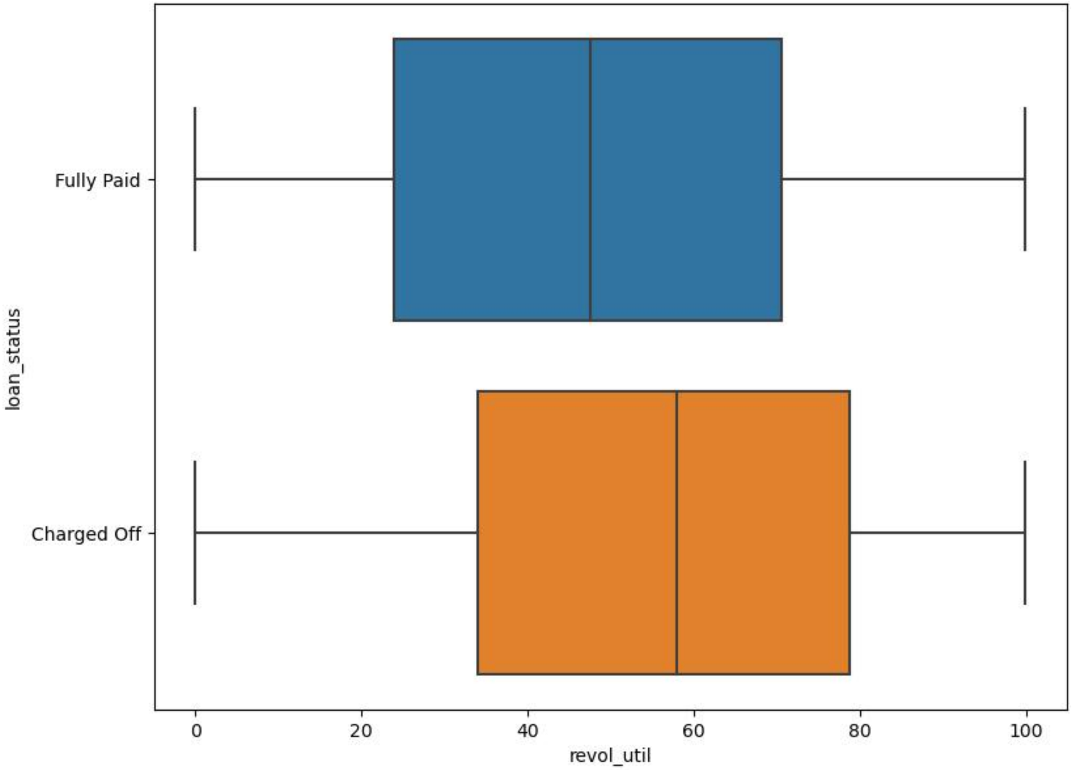
Loans with int_rate > 11.7(median) are 23% more probable to default



Segmented Univariate Result:

- Impact of Revolving Credit Utilization on Defaulting:

Loans with revol_util > 50 are 12% more probable to default



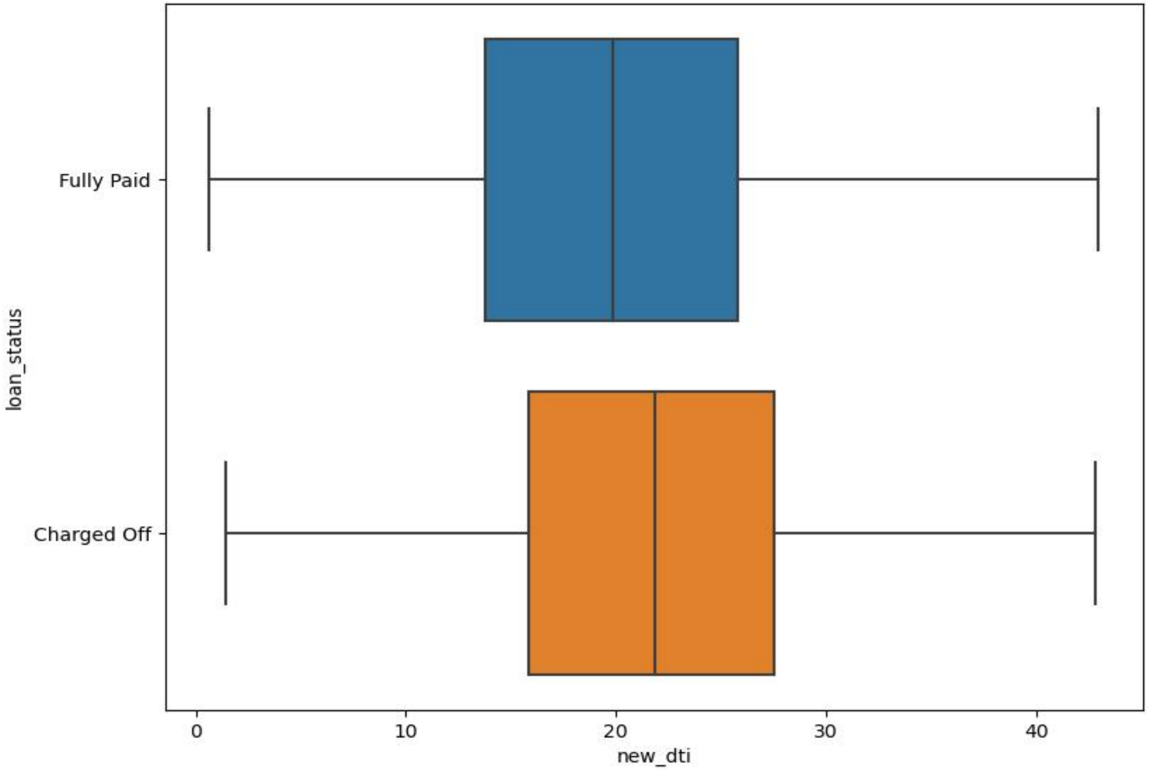
Segmented Univariate Result:

- Impact of new_dti (Derived Column) on Defaulting:

Existing DTI column will not be True indicatior of DTI after teh loan is taken, hence a **derived column** is created for True DTI calculation

new_dti = Current lc Installment to income(monthly) Ratio + dti (before appliction)

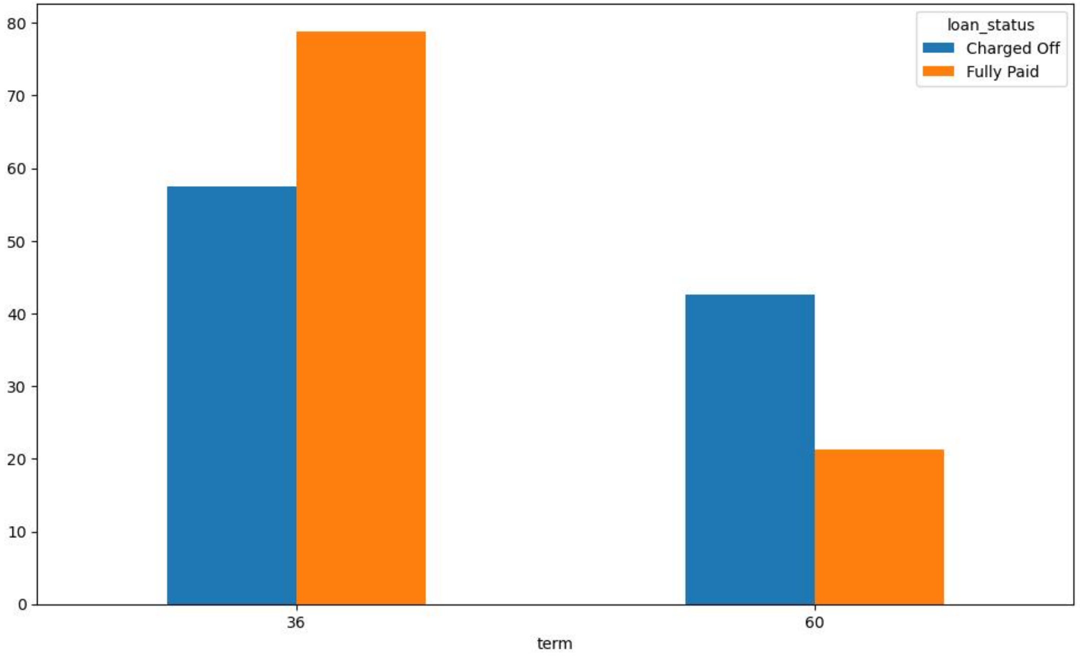
Loans with new_dti > 20 are 9% more probable to default. Whereas DTI was 5%



Segmented Univariate Result (Categorical):

- Impact of Term on Defaulting:

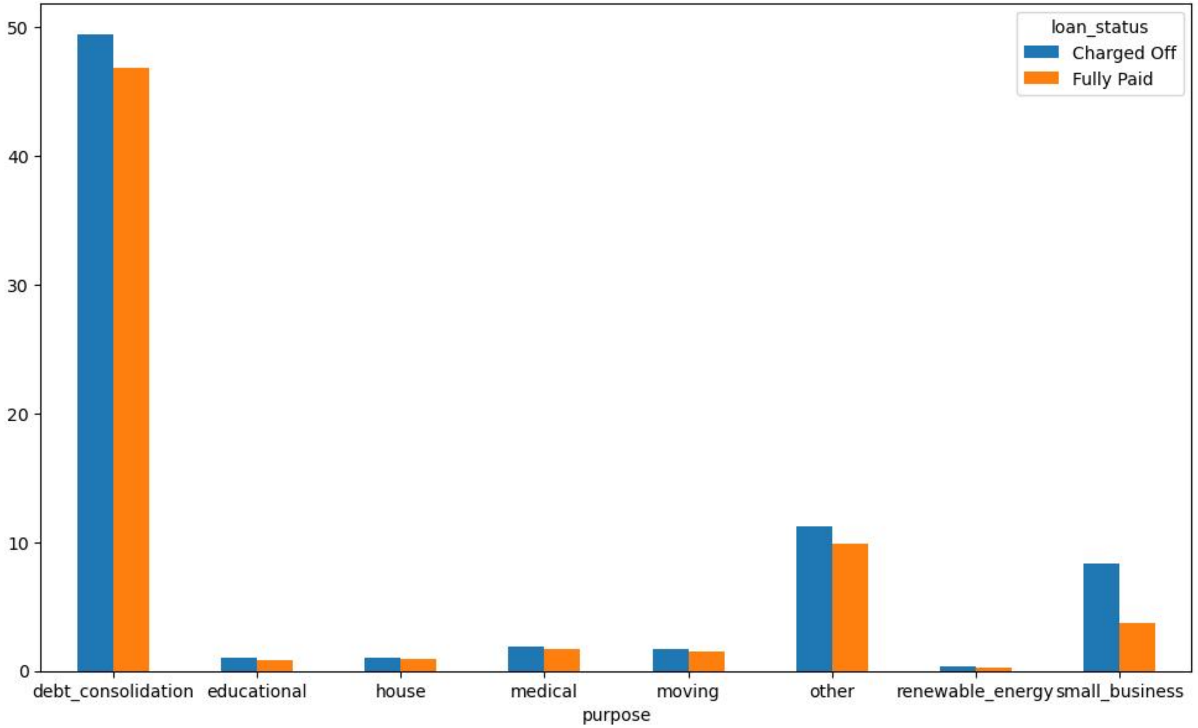
Loans with higer term (60months) are 21% more likely to default



Segmented Univariate Result (Categorical):

- Impact of purpose on Defaulting:

Loans taken for purpose of **'small business'** and **'debt consolidation'** are ~4% more likely to default

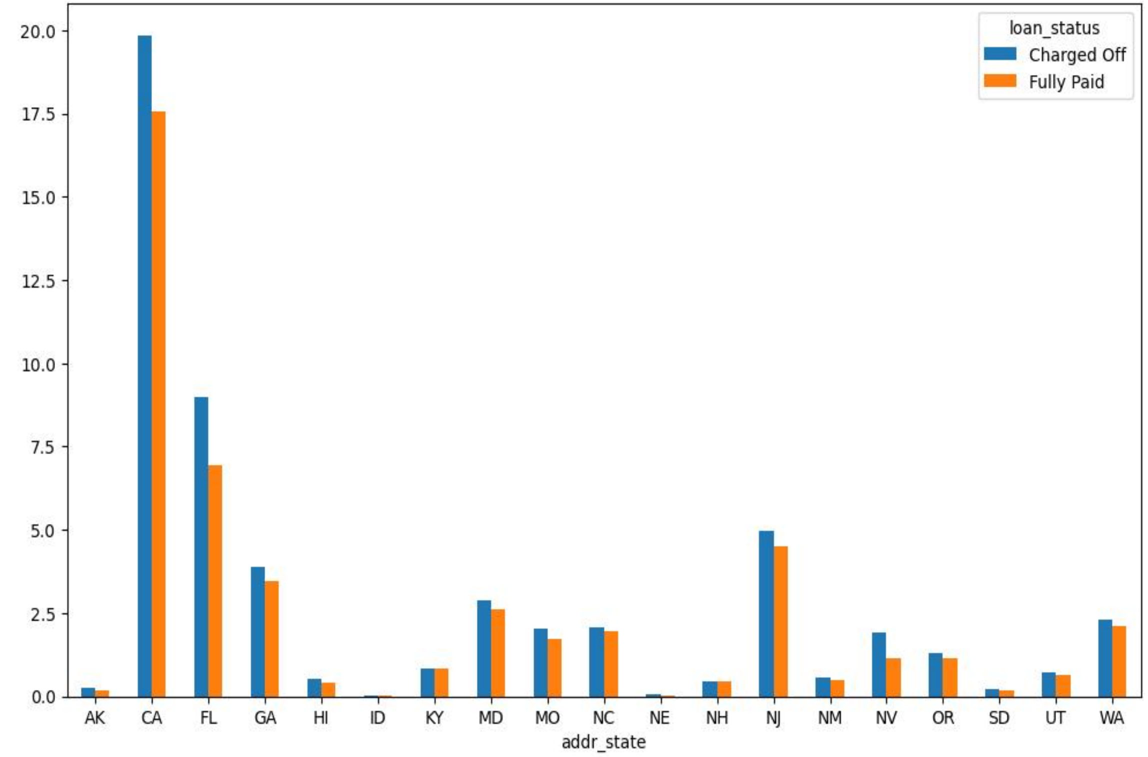


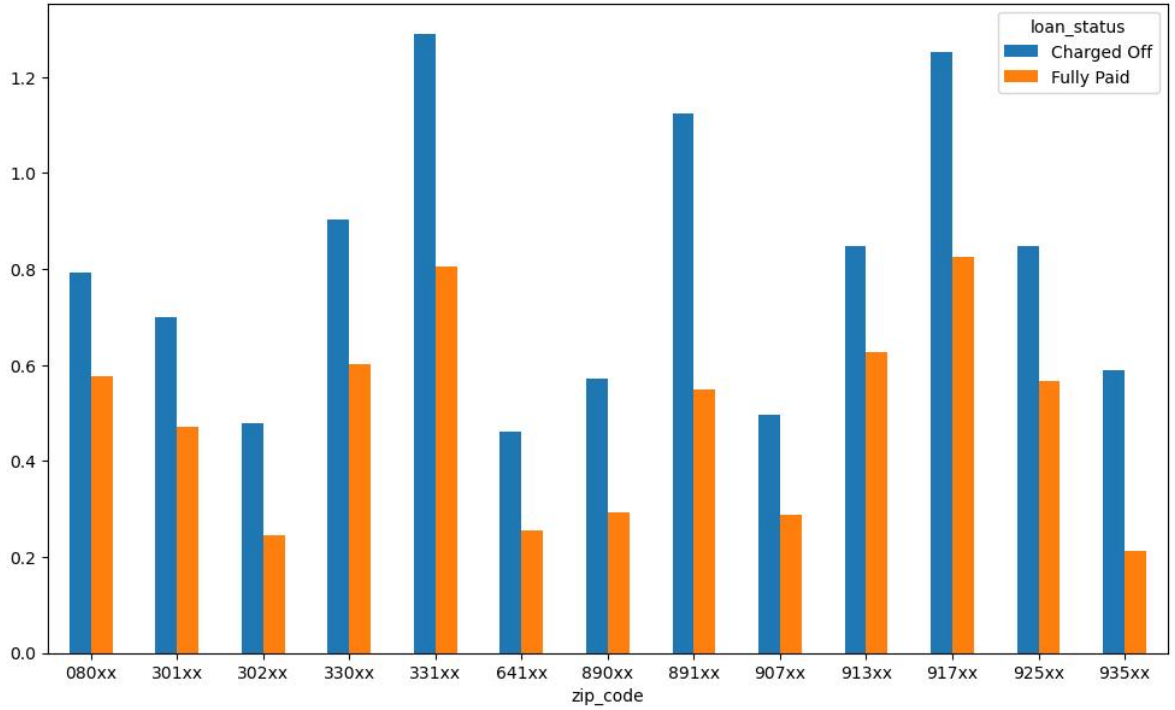
Segmented Univariate Result (Categorical):

- Impact of Location on Defaulting:

Loans from specific **States and ZipCode** are more likely to default than others

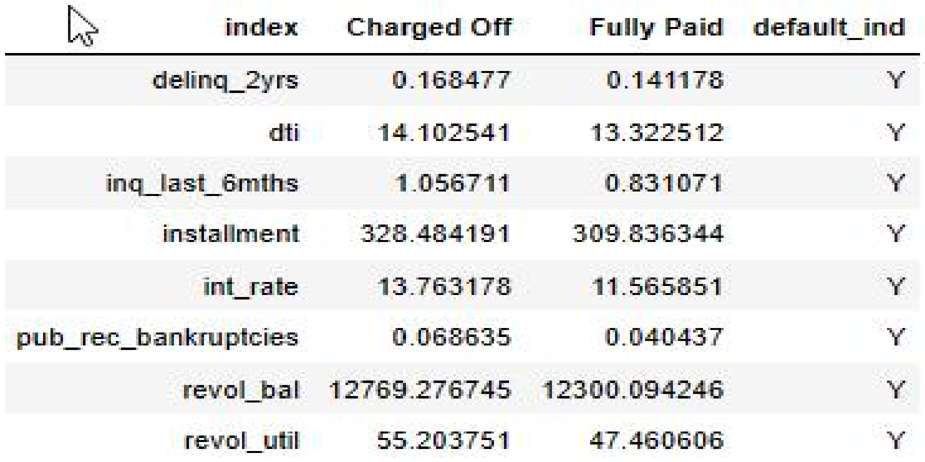
e.g. Loans from CA and FL are more prone to default, ZipCode 331xxx, 891xxx, 917xxx have more default loans in past than others





Bivariate Analysis Result:

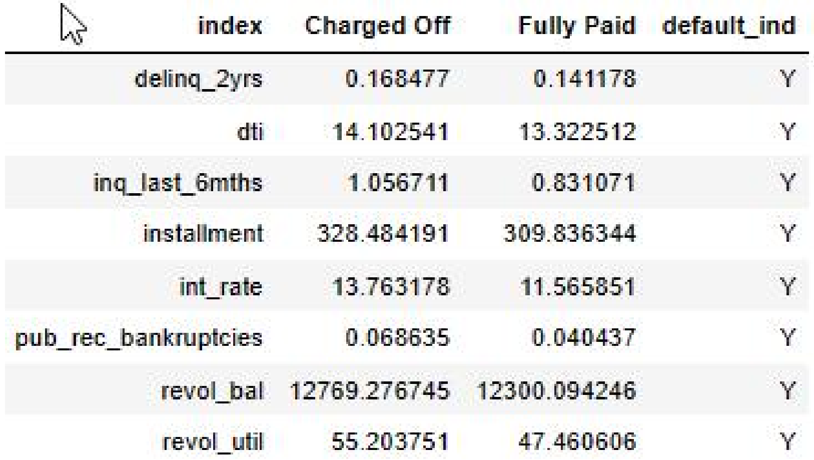
- We have done correlation analysis
- We found each of the following variables has Higher Mean Value for “Charged Off” compared to their corresponding values for mean values of “Fully Paid”. All these variables to be considered as a group and if each of these should have Higher values than the “Fully Paid” column in table below increase chance of of Default.
- Note: These values needs to be refreshed regularly.



index	Charged Off	Fully Paid	default_ind
delinq_2yrs	0.168477	0.141178	Y
dti	14.102541	13.322512	Y
inq_last_6mths	1.056711	0.831071	Y
installment	328.484191	309.836344	Y
int_rate	13.763178	11.565851	Y
pub_rec_bankruptcies	0.068635	0.040437	Y
revol_bal	12769.276745	12300.094246	Y
revol_util	55.203751	47.460606	Y

Bivariate Analysis Result:

- We have done correlation analysis
- We found each of following variables has Higher Mean Value for “Charged Off” compared to corresponding Value in “Fully Paid”. in all together contribute towards Higher chance of Default.



index	Charged Off	Fully Paid	default_ind
delinq_2yrs	0.168477	0.141178	Y
dti	14.102541	13.322512	Y
inq_last_6mths	1.056711	0.831071	Y
installment	328.484191	309.836344	Y
int_rate	13.763178	11.565851	Y
pub_rec_bankruptcies	0.068635	0.040437	Y
revol_bal	12769.276745	12300.094246	Y
revol_util	55.203751	47.460606	Y

Conclusion of Analysis

- Higher Interest Rate and Longer Term together contributes towards default of loan. Lending Club(LC) should try to give lower interest rate with longer term or Higher Interest Rate with Shorter Term.
- LC should be cautious approving new loans with applicants having more than 50% of Revolving Credit Utilization.
- LC should recalculate the DTI column after including the current loan installment and mortgage payment amount.
- LC should consider the Purpose of the Loan, State and ZipCode before approving. Specific location and purpose have higher Default Rate.
- Combination of 8 attributes (i.e. delinq_2yrs, inq_last_6mths, pub_rec_bankruptcies,int_rate,revol_util,revol_bal,installment,dti) indictates towards High Risk of Default.

THANK YOU

Q & A