**PGPDSE FT Mini-Project -Python**

**Loan Dataset**

**Participant :**

Rohit Dudmal

Shubham Rajput

Parul Bhaiya

Shachi Vaidya

Rupal Sanjay Nikum

**Industry Review**

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

The data given below contains the information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
   * **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
   * **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
   * **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

**Dataset and Domain**

Data Dictionary :

1.**annual\_inc** - The self-reported annual income provided by the borrower during registration.

2.**dti** - A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.

3.**emp\_length** -Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

4.**funded\_amnt** - The total amount committed to that loan at that point in time.

5.**funded\_amnt\_inv** -The total amount committed by investors for that loan at that point in time.

6**.grade** - LC assigned loan grade

7.**id** - A unique LC assigned ID for the loan listing.

8.**installment** - The monthly payment owed by the borrower if the loan originates.

9. **int\_rate** - Interest Rate on the loan

10.**last\_pymnt\_amnt**-Last total payment amount received

11.**last\_pymnt\_d** -Last month payment was received

12.**loan\_amnt** -The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.

13.**loan\_status** - Current status of the loan

14.**member\_id** -A unique LC assigned Id for the borrower member.

15.**purpose** - A category provided by the borrower for the loan request.

16.**term** -The number of payments on the loan. Values are in months and can be either 36 or 60.

17.**total\_acc** -The total number of credit lines currently in the borrower's credit file

18.**total\_pymnt** -Payments received to date for total amount funded

19.**total\_pymnt\_inv** -Payments received to date for portion of total amount funded by investors

20.**total\_rec\_int** -Interest received to date

**Dataset description:**

The ‘loan’ dataset contains all the required data for loans issued during the time period, 2007-2011.

**Business Importance:**

* Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
* Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owned. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.
* If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.
* In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

**Business Questions Discussion:**

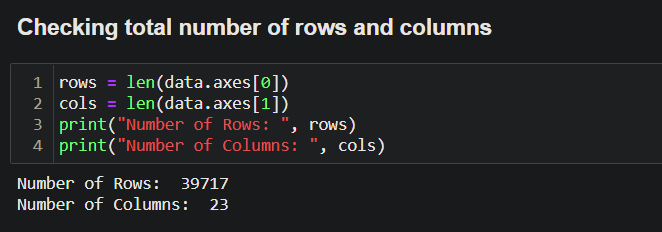
1. **Import the dataset and understand it.**

* To read the dataset we use pandas function pd.read\_csv(“”) and provide the path for location where the file is stored in quotes.

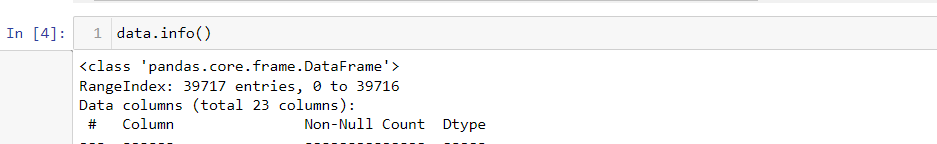
****

1. **List down the number of rows and columns.**

* For listing down rows and columns we use pandas function as pd.shape which return values as (rows\_no,columns\_no)

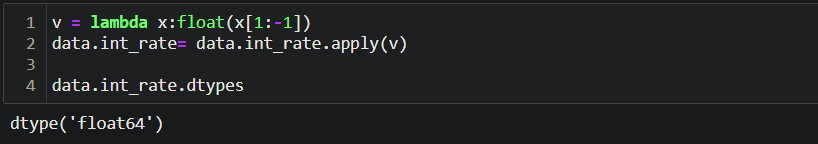


* Another way data.info will also return info of dataset which also gives information of Rows and columns.



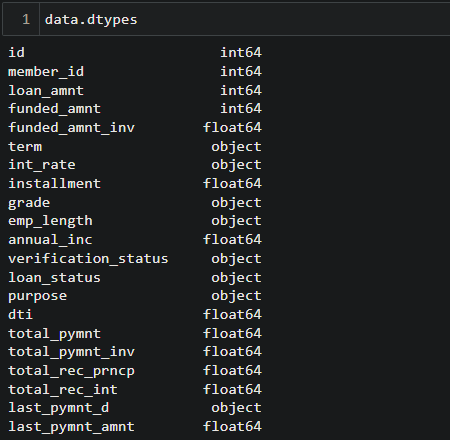
1. **‘Int\_rate’ column is character type. With the help of lambda function** **converting it into float type.**

* First we check datatype of int\_rate column using data.dtypes
* So datatype of int\_rate is ‘ Object’
* As int\_rate column contains values with % sign .First we have to replace that ‘%’ sign with blank\_space.
* Then apply lambda function to change int\_rate to float.



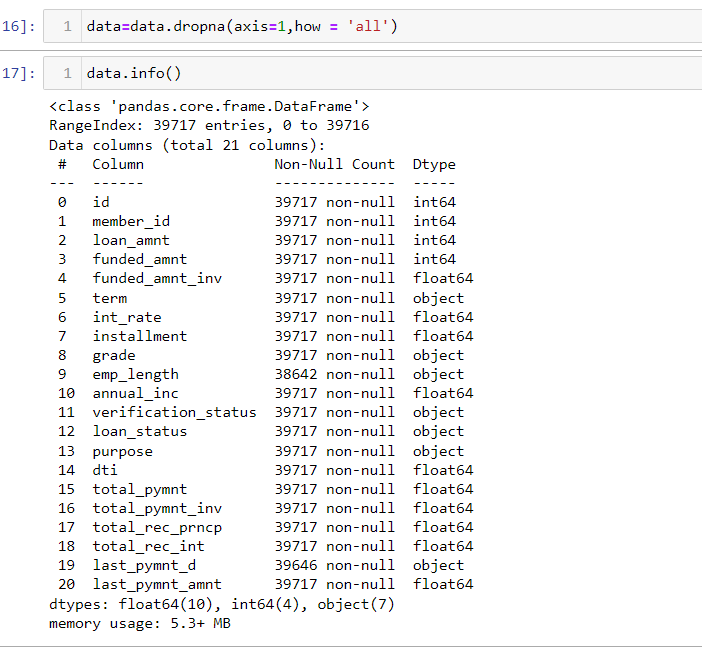
1. **Check the datatype of each column.**

* For checking datatypes of each column we use **data.dtypes**

****

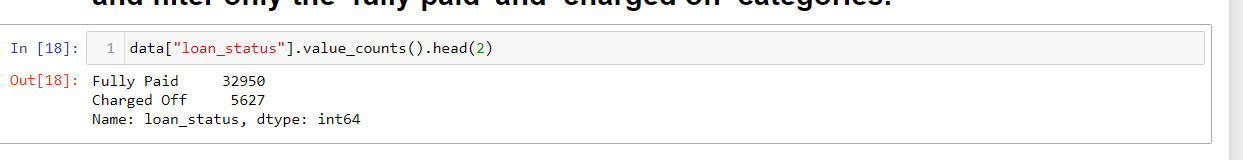
1. **Cleaning the dataset- Remove the columns having complete NaN value in the entire dataset.**

* **First we will check which columns contain all null values .**



1. **Write the code to find the value counts of the ‘loan\_status’ category.** **column and filter only the ‘fully paid’ and ‘charged off’ categories.**

* For reading values counts from particular department we use value\_counts **.**
* It return values of Loan\_status which give Fully\_paid and Charged\_off count.

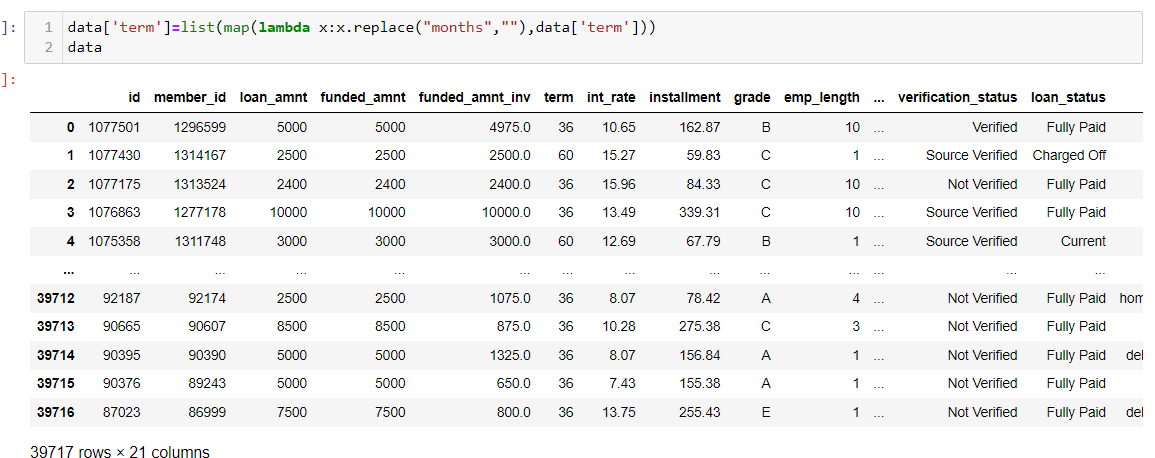


**7. Filter the ‘Emp\_Len’ column to extract the numerical value from the string.**

**Hint - Emp\_len : < 1year, 2 years , 3 years as 1 , 2, 3 so on.**

**8.Using the Lambda function, remove the month from the ‘term’ column such that ‘36 months’, ‘60 months’ appear as 36 and 60 respectively.**

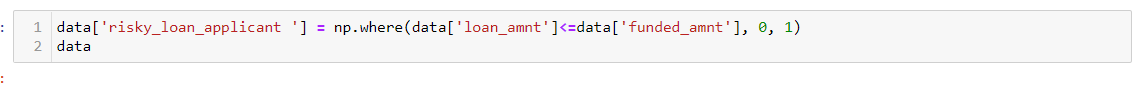
* We going to use replace along with lambda function to separate 36 from months.



**9. Create a new column as risky\_loan\_applicant by comparing loan\_amnt and funded\_amnt with the following criteria -**

**If loan\_amnt is less than equals to funded\_amnt set it as ‘0’ else set it as‘1’.**

* First, here we use numpy library with where condition to check given condition. Then we will give 0 or 1 as per the given criterias.

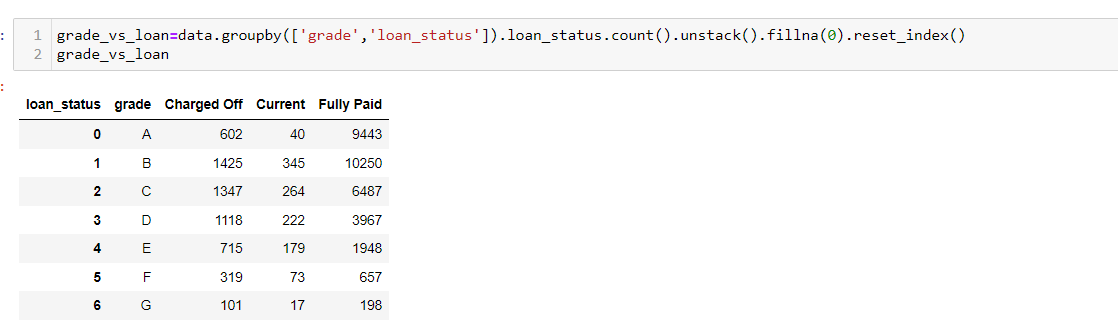


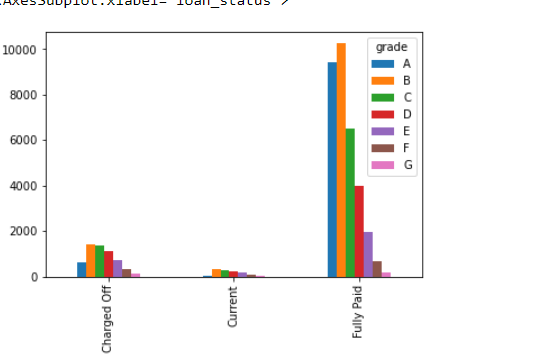
**10. Using the bar plot visualize the loan\_status column against categorical.column grade, term, verification\_status . Write the observation from each graph.**

* Loan\_status vs Grade:

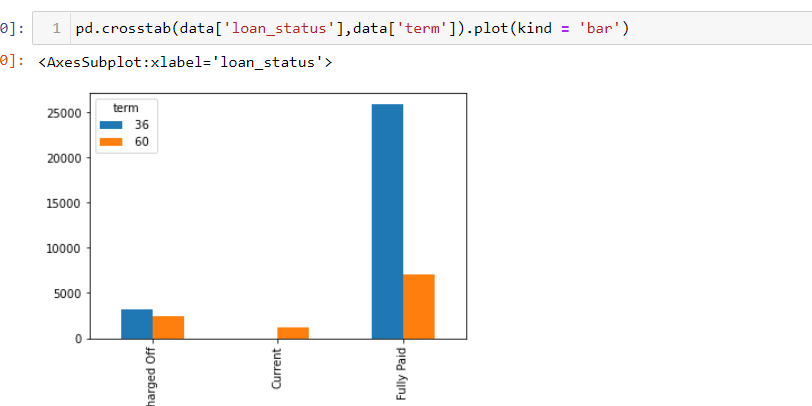
First we group by loan\_status and grades columns to get count of people.

Then will apply bar chart to get categories in which max defaulter are in.

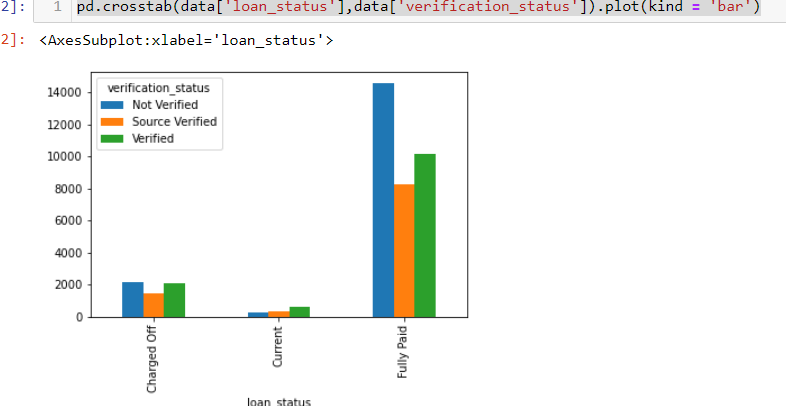




**2] plot term vs Loan\_status**



**3]Verification status vs Loan\_status**



**Observation :**

1.There are comparatively greater number of people with grade B.

2. In the given dataset, we are having more number of people with term 36 months than 60 months. But the probability of loan getting defaulted is more for 60 months than 30 months.

3. In the given data , there are more records for which the status is non verified, and the defaulter rate is more for non-verified status.

**11.Using a user defined function convert the ‘emp\_len’ column into**

**categorical column as follows -**

If emp\_len is less than equals to 1 then recode as ‘fresher’.

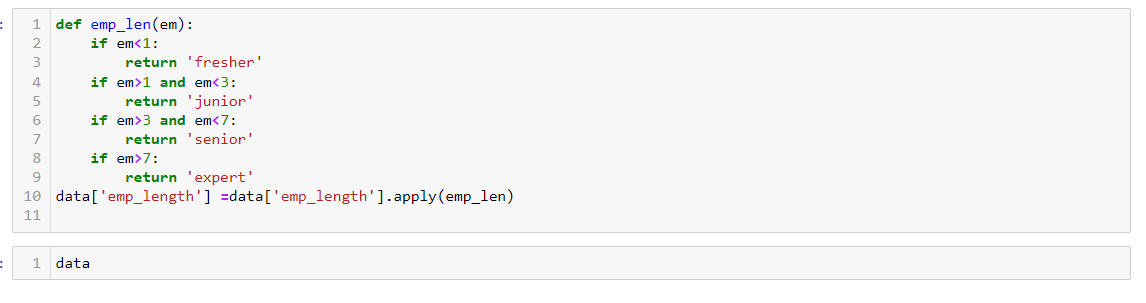
If emp\_len is greater than 1 and less than 3 then recode as ‘junior’.

If emp\_len is greater than 3 and less than 7 then recode as ‘senior’

If emp\_len is greater than 7 then recode as ‘expert’.

* In this explanation, we are using user-defined function
* Before that we have to separate the emp\_len column

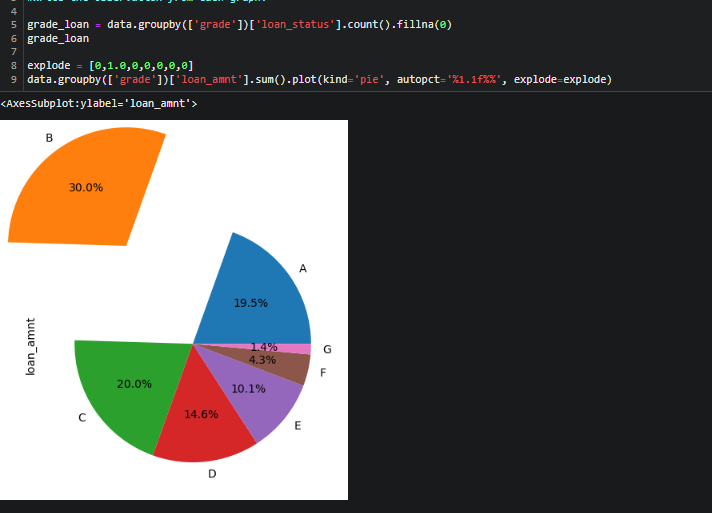
By numeric data using replace function.



**12.Find the sum of ‘loan\_amnt’ for each grade and display the distribution of ‘loan\_amnt’ using a pie plot**

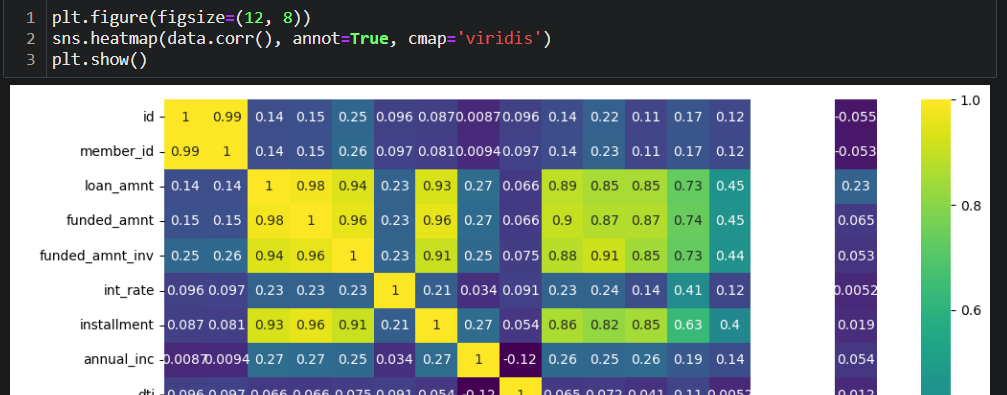
* Using groupby() function – which is an inbuilt function in python pandas, the given set of data is grouped according to the set of columns given.
* Then apply pie plot to display loan\_amount as per each grade.



****

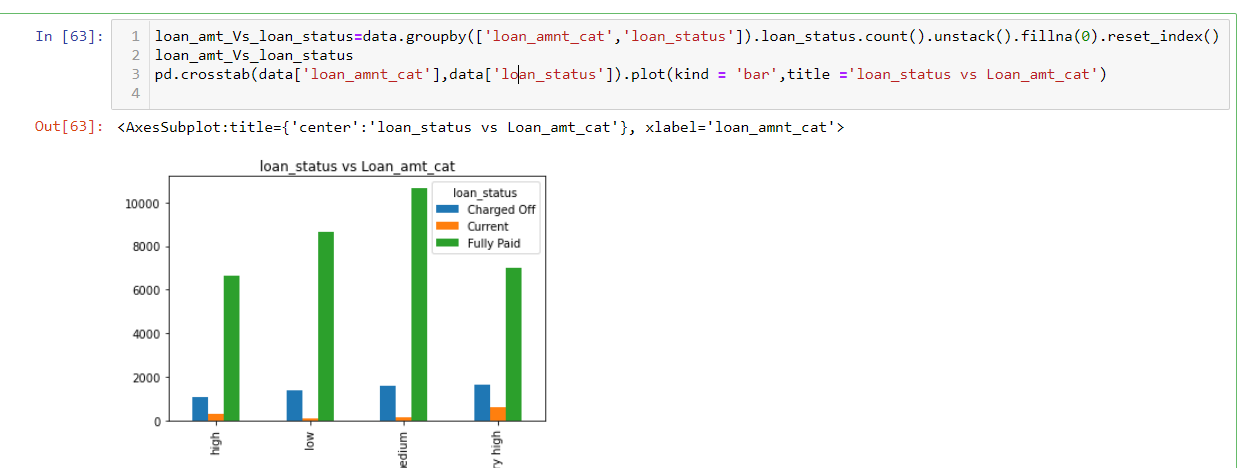
Observation: Loan\_amnt is max for Grade B.

**Question 13: Examine the correlation among the variables- Loan Amount, Funded Amount and Installment Amount**

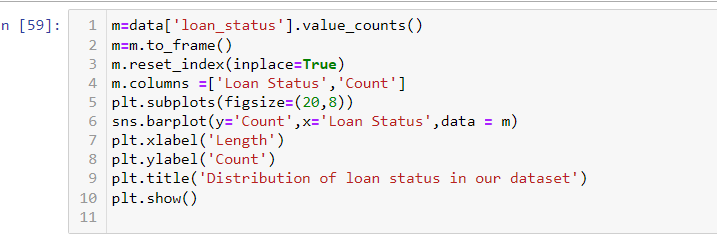
****

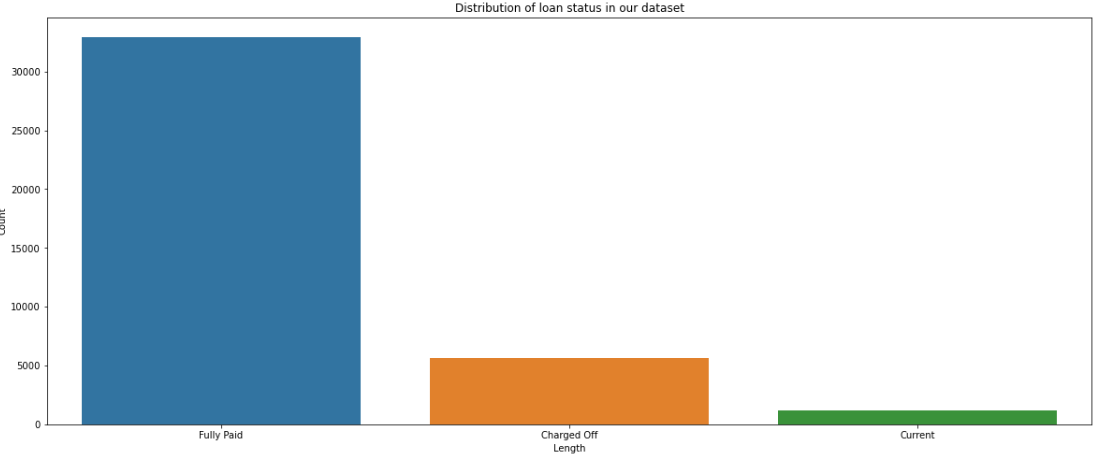
**Question 14: Divide Loan amount into groups based on intervals and plot the graph against loan\_status.**





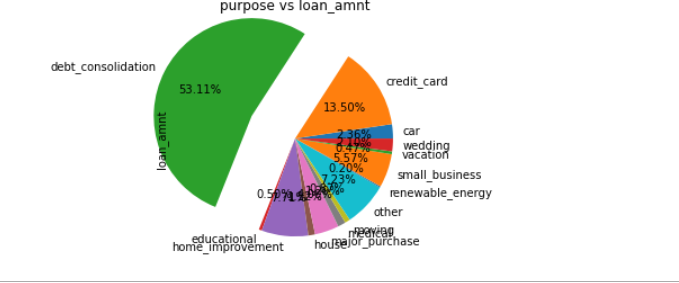
**Question 15: Distribution of loan status values**

****

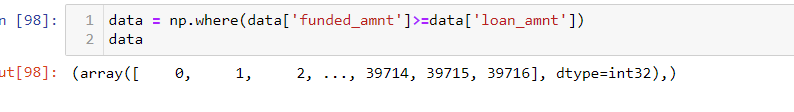
****

# Question 16: Findout Purpose for which max amount is taken as Loan.



****

Question 17: Find the ids of safe loan\_applicant.

****

**Inferences :**

**1) Loan Term:** Average Interest rate for defaulted applications is very high for 36 months and low for 60 months term.  
  
**(2) Grade:** Default Rate is high in high risk loan applicants. It would be important for LC to thoroughly vet high risk loan applications.  
  
**(3) Loan Amount:** Defaulter rate increases as the requested loan amount increases.

(4)**Purpose:** Maximum loan\_amount is taken for debt\_consolidation.

**Conclusion**

Customer Bank Loan data, is a collection of values, wherein the complete data has been provided with reference to the bank data and specifically relates to loan dataset. Identification of risky loan applicants, and cutting down the amount of credit loss is the main aim of our project.

Along with this, comparing data values with other numerical data types to effectively state the most likely features that will help us derive to a set of conclusions and provide inferences according to that is the aim. This way, the bank won’t lose potential customers and also at the same time derive conclusions on the risky loan applicants and take according measures to reduce financial barriers and credit risks for the bank.

Being in the shoes of a Data Scientist, our aim was to provide with effective inferences and conclusions based on previous customer data and provide analysis on predicting conclusions on whether a customer is likely to default a loan or not. Here we used skills such as:

1.Numpy functions

2.Pandas functions

3. Visualization techniques such as : Barplot, pieplot with all their attributes.

4. Various visualization libraries like matplotlib, seaborn and plotly.