As an open ended data set was given, narrowing to various factors influencing the default rate of a loan seemed perfect for risk assessment

Importing Libraries

The important libraries that were imported were

- Numpy
- Pandas
- Matplot
- Seaborns

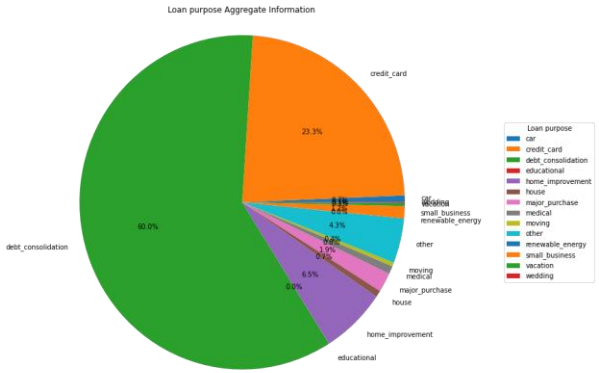
The data was loaded into a variable named **loan**

The data had many empty values that was evident from the isnull() function

Pie Charts

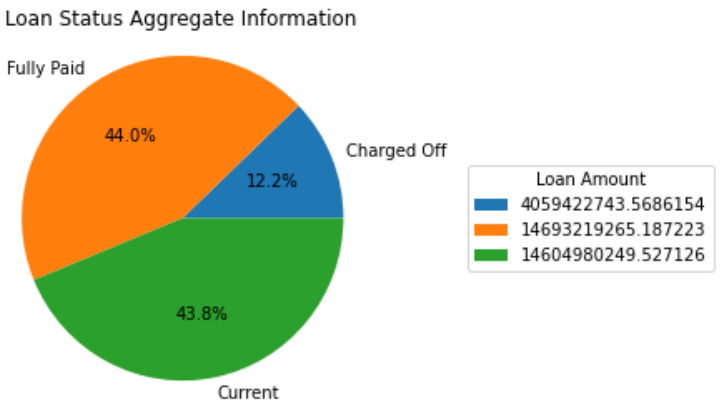
There are two pie charts in the notebook depicting the following

- **Different Types of loan status-** 44% of the loans were fully paid and 43.8% were the ongoing loans
- **Differnet purpose for loans-** This metric has 14 division of which debt consolidation and credit card are the largest



Data Cleaning

- Removing all the columns that are null to reduce the data
- Removing all Demographic and Customer Behavioural data(given in the notebook(
- Removed null value columns from *emp_length* and *revol_util*
- Removing % sign and converting *int_rate* and *revol_util* into numeric value
- Remove the current loan(doesn't give any insight of defaulting but can be used as a test set)
- Change the Charged off to 1 and fully paid to 0(for default analysis
- Using boxplot to find the percentiles for *annual_inc, open_acc, total_acc andpub_rec* and removing the outliers i.e., 99.5-100 percentile
- Binning of continuous data like *loan_amnt, int_rate, annual_inc, installment* and *dti* for effective analysis



Visualizing data

- Checked for the percentage of defaulters using the countplot which depicted 21.81% default percentage
- **Univariate Analysis-** Used ratio bars and barplots to find relationship (direct or indirect) of various continuous and categorial features to the default rate
  - **Categorical Features-** *term, grade, sub_grade, purpose*
  - **Continuous features-** *inq_last_6mths, benificial, total_acc, loan_amnt_range, int_rate_range, annual_inc_range, dti_range, installment*
- **Bivariate Analysis-** Used **pairplot** to find out that *term, grade, purpose, pub_rec, revol_util, funded_amnt_inv, int_rate, annual_inc,* installment are the important features
- **Multivariate Analysis-** Using correlation heatmap to find correlated features

Final Finding

The main driving features for the data set are ***term, grade, purpose, revol_util, int_rate, installment, annual_inc, funded_amnt_inv*** when it comes to defaulting