LENDING CLUB CASE STUDY

Problem Statement:

A **consumer finance company** 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. 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 provided contains 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.

Use Exploratory Data Analysis techniques (Univariate and Bivariate analysis) on the data provided to understand how consumer attributes and loan attributes influence the tendency of the default.

Analysis and Approach:

Data Understanding and Preliminary Analysis

Load the dataset and take a look at the first few rows to understand its structure and contents.
Check for the data types of different columns, missing values, and basic summary statistics.

Missing Value Analysis

Analyze the percentage of missing values in each column.
Depending on the amount of missing data, decide whether to impute, drop, or leave missing values as-is.

- Check the number of null values in the columns
  df.isnull().sum()

- Print the presentage of missing value

  round(df.isnull().sum()/len(df.index) * 100)

- Drop the columns with values missing for more than 90% rows

- Verify the columns that have misssing values desc and mths_since_last_delinq

- The desc column contains the free text. For the EDA related analysis, this column values will not be of any help unless we do NLP related work. Let us drop those columns

Data Reformatting and Analysis

Manually validate the data to see if any column that require transformations

Reformat the int_rate column by removing % to perform mathematical operations

```
df["int_rate"] = df["int_rate"].apply(lambda x: float(x[:-1]))
```

reformat emp_legth

drop the missing values from emp_length

Outlier Analysis

Identify potential outliers in numerical variables using box plots or Z-scores.
Decide whether to treat outliers (remove, transform, or leave as-is) based on domain knowledge and data characteristics.

In the data, the behaviour variable value is generated once after the loan is approved, Hence these column s are not required for our analysis. So drop those columns

Let us also drop the address and URL related variable and this is not needed for analysis

Since purpose and title looks same, we will drop the title

```
#Analysis on the loan_status column and see the value counts

df['loan_status'] = df['loan_status'].astype('category')


df['loan_status'].value_counts()
```

```
Fully Paid     32145
Charged Off     5399
Current         1098
```

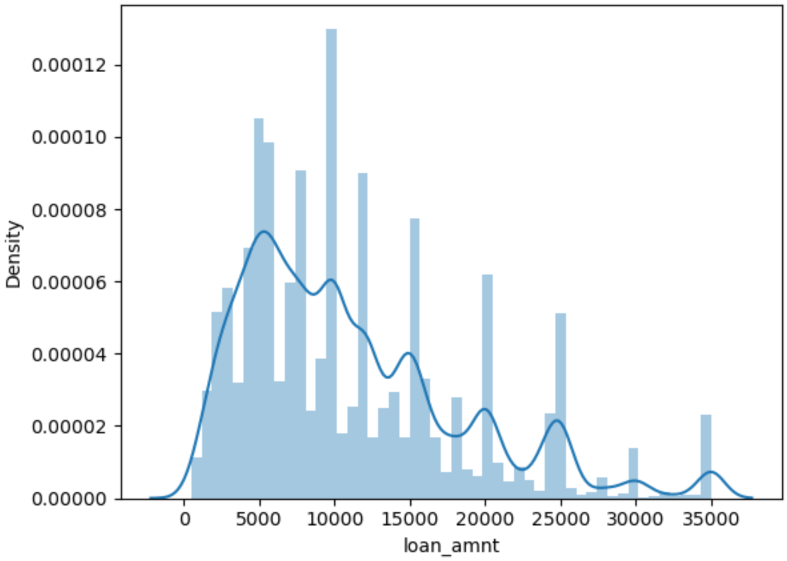
```
#In this case we only have three categories in the loan status.
#Current customers are ongoing paying customers, from whom we cannot find any patterns. So drop those rows

df.drop(df[df['loan_status'] == 'Current'].index, inplace = True)
```

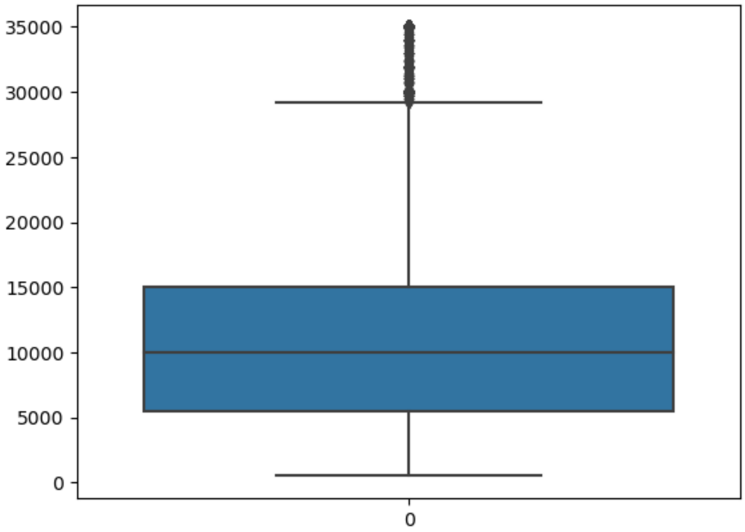
Data Visualization

Create additional visualizations like pair plots, heatmaps, or correlation matrices to identify complex relationships and patterns.

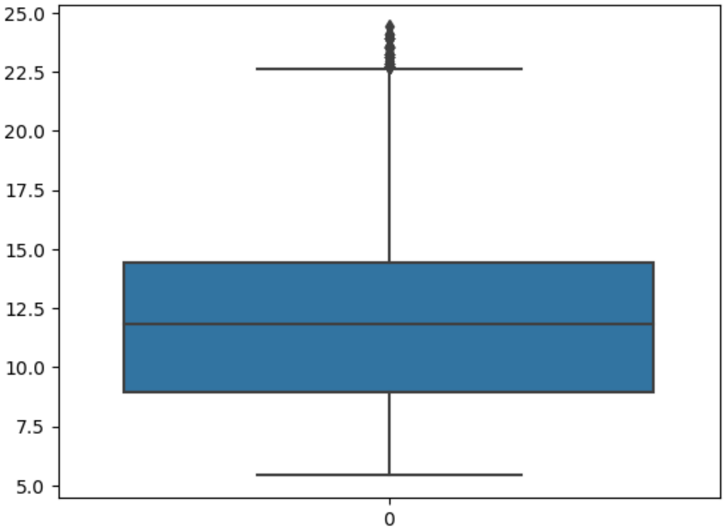
Univariate Analysis

Do Box plot on Loan amount field





Histogram on int_rate and dist plot



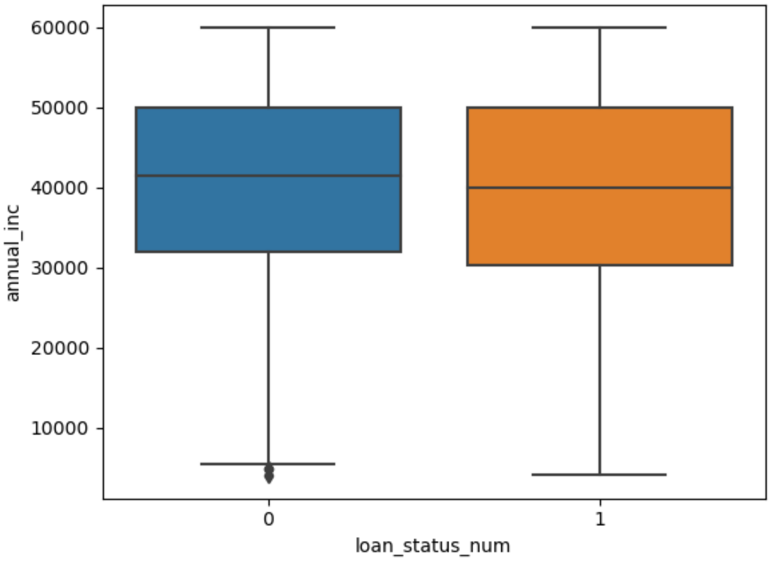
Univariate categorical analysis

Take Grade column.value_counts

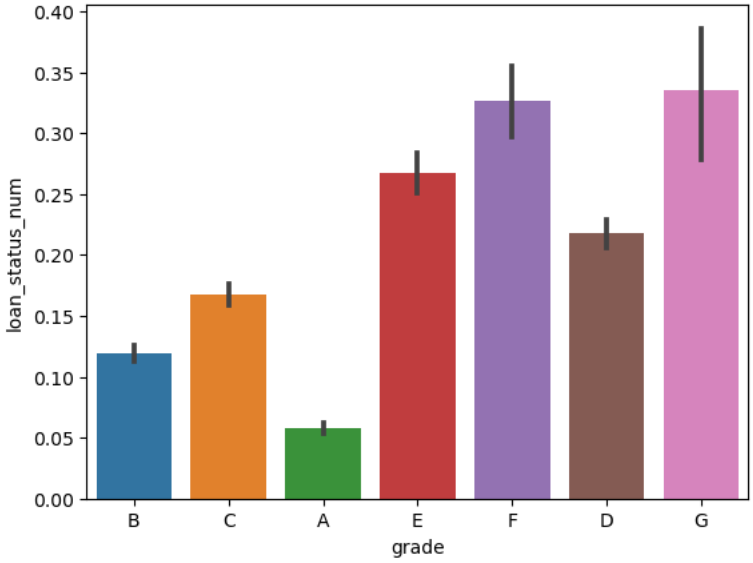
Sns.count_plot(grade)

Bivariate

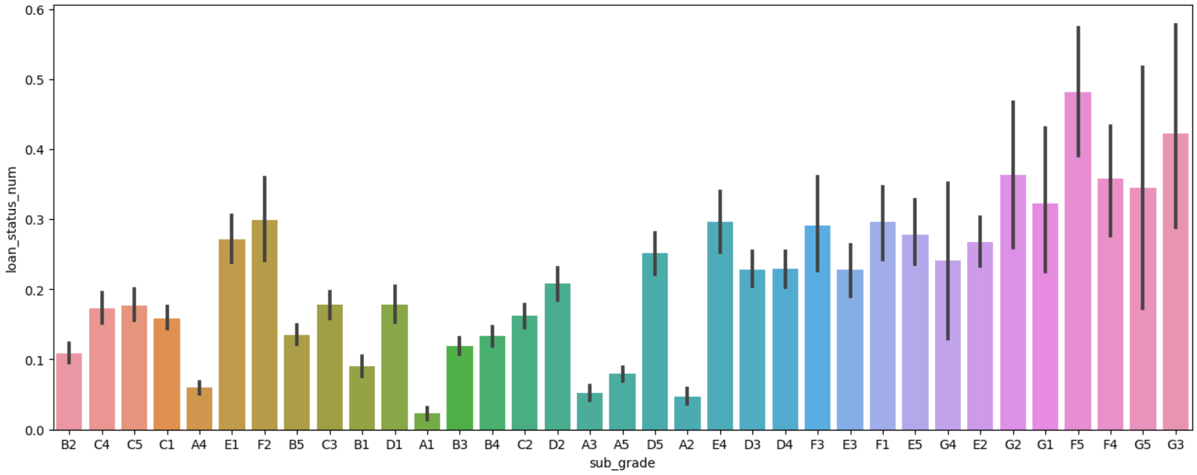
Analysis relation between loan_status and salary
Do a sns.box plot, x = loan_status, y=salary(Use log scale)



Also do box plot on loan status and int_rate. (Shows int rate are high, customers are defaulting)



Also analyse against grade and loan_status



The above graph show the chart against loan status and the sub grade.

Conclusion:

The Loan status varies with different factors. After performing EDA analysis, it is evidently clear that higer interest rate having more defaulters. Similary the Income and the Grade also have huge impact