Lending Club

Problem Statement

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. 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'.

Given a loan data set having loan data for all loans issued through the time period 2007 t0 2011. It is expected to understand the driving factors (or driver variables) which identify direct relation or impact loan default, i.e. the variables which are strong indicators of default. This knowledge can be utilized by organization for its portfolio and risk assessment.

Analysis Approach

The approach followed for the analysis:

- Load data
- Clean and Transform the data for analysis with following steps
  - Check data shape (i.e no of rows and columns)
  - Check columns having all null values and drop them
  - Understand the data dictionary and for the given use drop the columns not required, finally check the data frame shape to get only columns which are relevant for the use case analysis
  - Data types correction
    - Evaluate data type
    - Convert columns which str, object and required to be int
    - Fill 0 or relevant values for empty values in the columns
  - Remove the Current/In-Progress value from target variable “Loan_status”
  - For each of the columns required in analysis remove the outliers

Analysis Approach

The approach followed for the analysis:

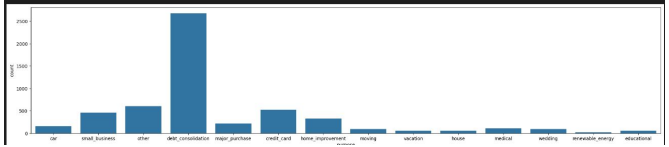
- Once Data is cleaned then start doing analysis
  - UniVariate Analysis
    - Analyze each column in the dataframe for count, freq for - grade, home owner, purpose, annual income, interest rate, loan amount
  - Segmented Analysis
    - Check the key categories
    - For each of these key categories, check the defaulted loan count for various features like - term, grade, home ownership, employee length
    - Create bins (High, MED, LOW) for various categories like - loan amount, interest rate, annual income, DTI and check it’s variation/impact on “loan_status”
  - BiVariate Analysis
    - Have the correlation between the required fields loan amount, interest rate, annual income, installments, DTI etc
    - Using the required bins create the plots between
      - Annual income vs Charged loans
      - Purpose vs Charged loans

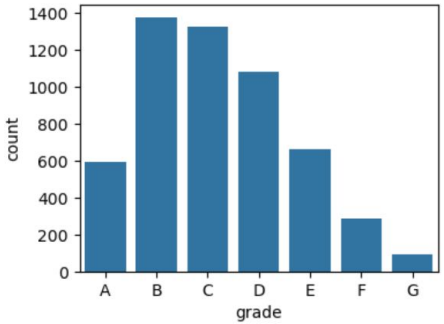
Results Details - Univariate

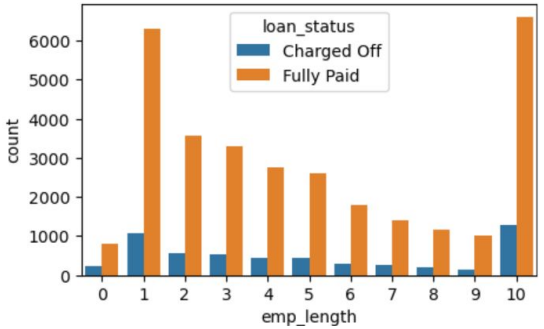
```
# see the count plot for purpose status
fig, ax = plt.subplots(figsize = (25,5))
sns.countplot(x = 'purpose',data = loan_data_df[loan_data_df['loan_status']=='Charged Off'])
```

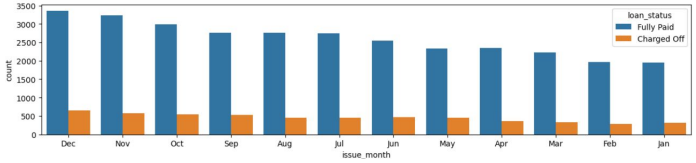
0.2s

```
<Axes: xlabel='purpose', ylabel='count'>
```





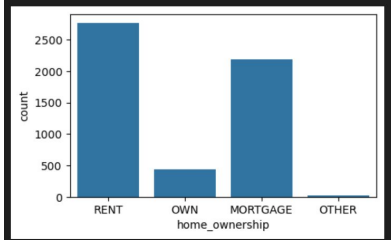





```
plt.figure(figsize=(5,3))
sns.countplot(x = 'home_ownership', data = loan_data_df[loan_data_df['loan_status']=='Charged Off'])
```

0.1s

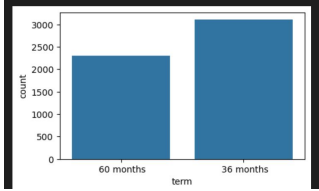
```
<Axes: xlabel='home_ownership', ylabel='count'>
```




```
# see the count plot for loan term
plt.figure(figsize=(5,3))
sns.countplot(x = 'term', data = loan_data_df[loan_data_df['loan_status']=='Charged Off'])
```

0.1s

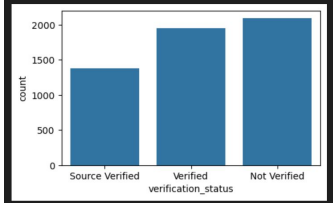
```
<Axes: xlabel='term', ylabel='count'>
```




```
# see the count plot for verification status
plt.figure(figsize=(5,3))
sns.countplot(x = 'verification_status', data = loan_data_df[loan_data_df['loan_status']=='Charged Off'])
```

0.1s

```
<Axes: xlabel='verification_status', ylabel='count'>
```



Results Details - Univariate

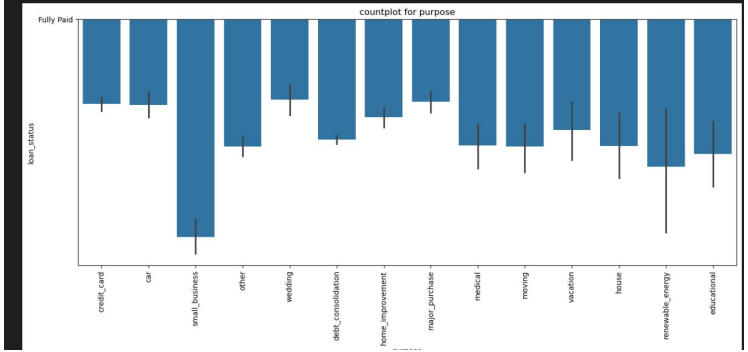
After seeing the various data counts and univariat analysis we can say following from the data set that there is slight more probability of defaulting loans for

- People having employment in either in 0-1 yr or 10 yr or above
- People getting loan are in GRADE B or C
- When loan term is 3 yrs
- When loans are NOT VERIFIED
- People having RENTED/MORTGAED accomodation
- Loan issued during year end Nov-Dec
- People using loans for clear debt collections

Results Details - Segmented Univariate

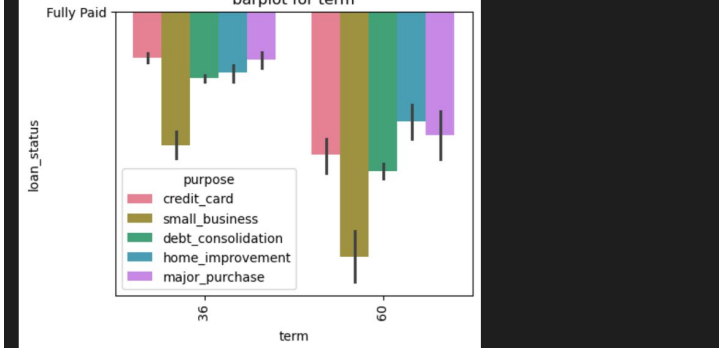
```
loan_data_df_data = loan_data_df
sns.barplot(x='purpose', y='loan_status', data=loan_data_df_data)
plt.show()
```

✓ 0.7s



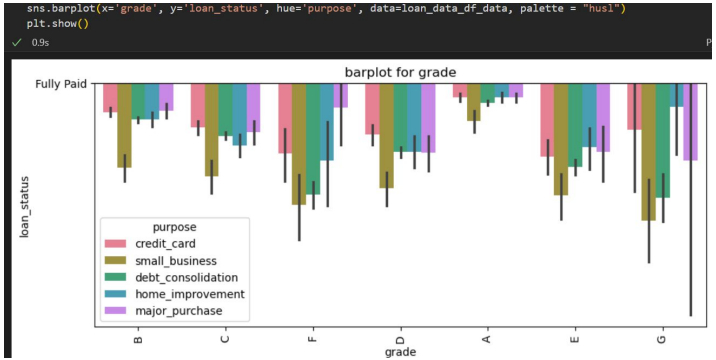

```
sns.barplot(x='term', y='loan_status', hue='purpose', data=loan_data_df_data, palette = "husl")
plt.show()
```

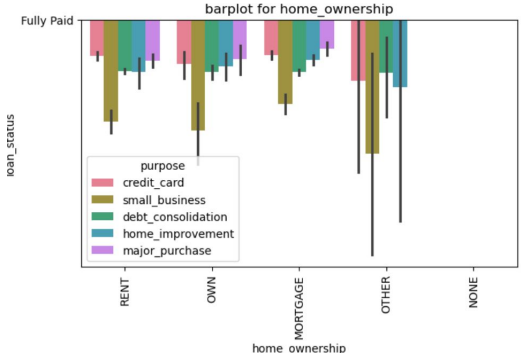
✓ 0.5s

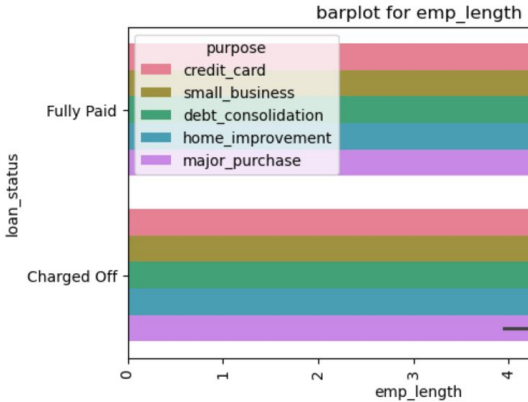



```
sns.barplot(x='grade', y='loan_status', hue='purpose', data=loan_data_df_data, palette = "husl")
plt.show()
```

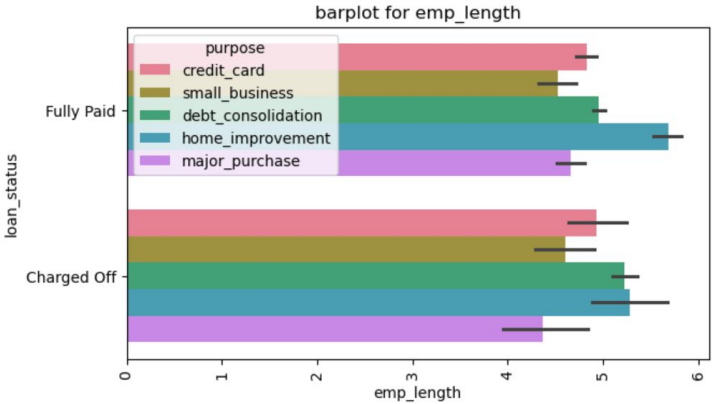
✓ 0.9s

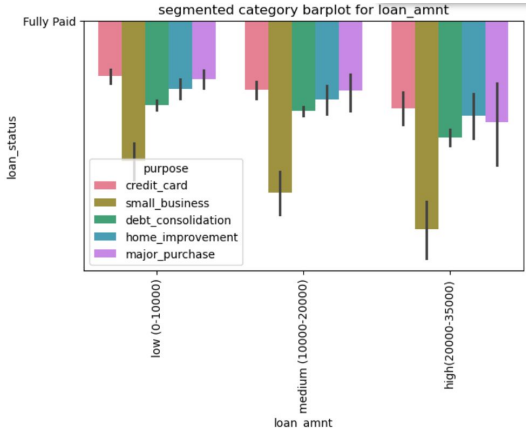


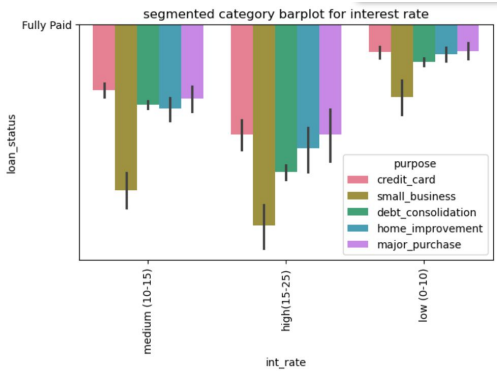


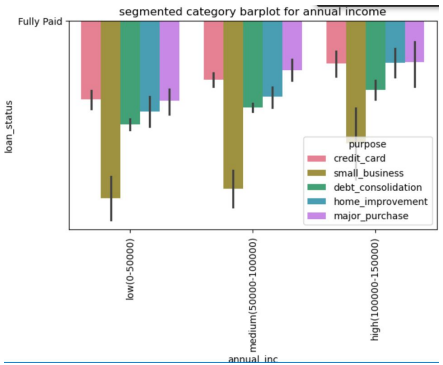


Results Details - Segmented Univariate







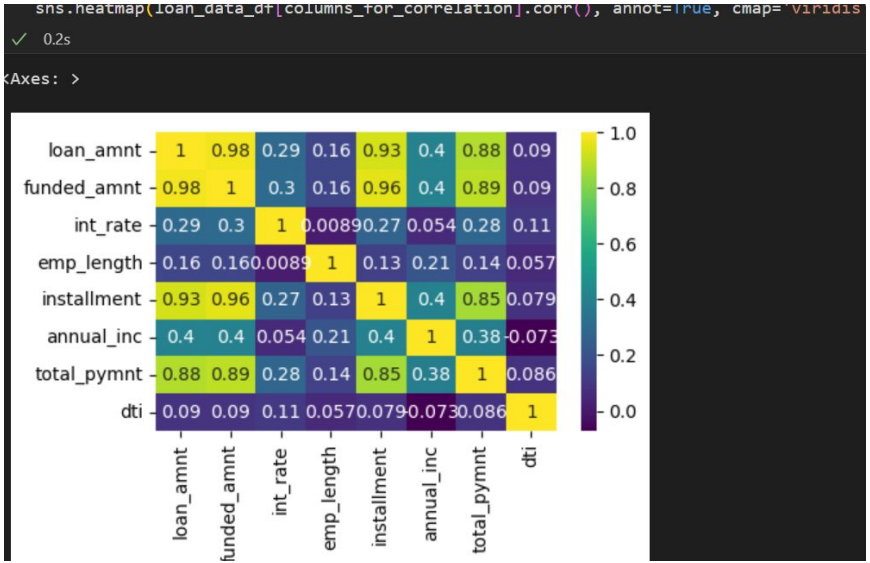


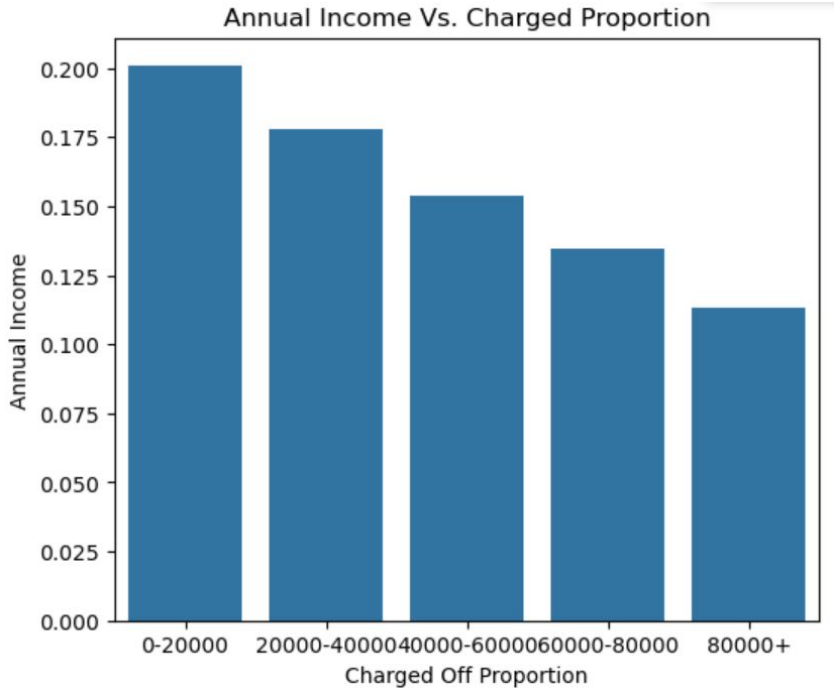
Results Details - Segmented Univariate

Observation:

- Overall "small business" have high tendency for loan defaults followed by "debt_collection" in all various areas with following order
- loan amount -> HIGH > MEDIUM > LOW
- interest rate -> HIGH > MEDIUM > LOW
- Annual income -> LOW > MEDIUM > HIGH
- "DTI" -> HIGH > MEDIUM > LOW

Results - Bivariate





Results Details - Bivariate

Observation:

- Income range 80000+ has less chances of charged off.
- Income range 0-20000 has high chances of charged off.
- Notice that with increase in annual income charged off proportion got decreased. So, they are inversely proportional.
- Small Business applicants have high chances of getting charged off.

Observations:

- Loan Amount is positively correlated to funded amount and installemnets i.e higher the loan amount , higher is funded amount and installments
- Annual income and employment length are positively correlated i.e higher the employment length higher the annual income
- Annual income and Debt to income are negatively correlated i.e higher the annual income lower is Debt to income ratio