Lending Club Case Study

Exploratory Data Analysis

Business Overview

- The company specializes in lending loans to urban customers.
- 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, 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.
- There are a total of 111 columns and 39717 records in the loan.csv file.

The aim is to identify patterns that 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.

Data Cleaning

- There are a total of 111 columns and 39717 rows.
- Majority of the columns has only Null values -
  - Hence dropping all the columns whose null values are more than 20000
- We are left with 16 columns and 39717 rows.
- There are no rows in the data with all null records.
- Dropping all the rows whose *loan_status = Current* as they have no effect on our analysis

Output -The final cleaned dataset has a total of 16 columns and 36800 rows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   loan_amnt             39717 non-null  int64
 1   term                  39717 non-null  object
 2   int_rate              39717 non-null  object
 3   installment           39717 non-null  float64
 4   grade                 39717 non-null  object
 5   emp_length            38642 non-null  object
 6   home_ownership        39717 non-null  object
 7   annual_inc            39717 non-null  float64
 8   verification_status   39717 non-null  object
 9   issue_d               39717 non-null  object
 10  loan_status           39717 non-null  object
 11  purpose               39717 non-null  object
 12  dti                   39717 non-null  float64
 13  pub_rec               39717 non-null  int64
 14  revol_util            39667 non-null  object
 15  pub_rec_bankruptcies  39020 non-null  float64
dtypes: float64(4), int64(2), object(10)
memory usage: 4.8+ MB
```

```
final_loan_data.isnull().all(axis=1).sum() # Number of records with all null values
✓ 0.0s
0
```

Data Formatting

- The following columns with inconsistent data formats are corrected for appropriate analysis
  - Employment length
    - Special chars & Strings are removed
    - Converted into float
  - Loan issue date
    - Object to date
  - Term of the loan
    - String object to years
  - Interest rate & Revolving credit utilization rate
    - Removed % and converted to float

```
final_loan_data.info()
```

✓ 0.0s

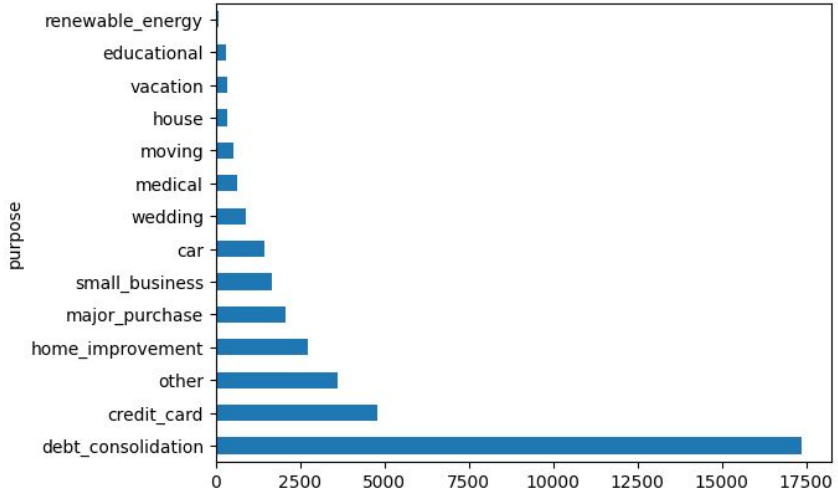
```
<class 'pandas.core.frame.DataFrame'>
Index: 36800 entries, 0 to 39680
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   loan_amnt             36800 non-null  int64
 1   term                  36800 non-null  float64
 2   int_rate              36800 non-null  float64
 3   installment           36800 non-null  float64
 4   grade                 36800 non-null  object
 5   emp_length            36800 non-null  object
 6   home_ownership        36800 non-null  object
 7   annual_inc            36800 non-null  float64
 8   verification_status   36800 non-null  object
 9   issue_d               36800 non-null  datetime64[ns]
 10  loan_status           36800 non-null  object
 11  purpose               36800 non-null  object
 12  dti                   36800 non-null  float64
 13  pub_rec               36800 non-null  int64
 14  revol_util            36800 non-null  float64
 15  pub_rec_bankruptcies  36800 non-null  float64
dtypes: datetime64[ns](1), float64(7), int64(2), object(6)
memory usage: 4.8+ MB
```

Output - Cleaned data without data inconsistencies and ready for further EDA steps

Univariate Analysis

1. Purpose of the loan

Debt Consolidation, Credit card and Home improvement are some of the popular reasons for taking loan

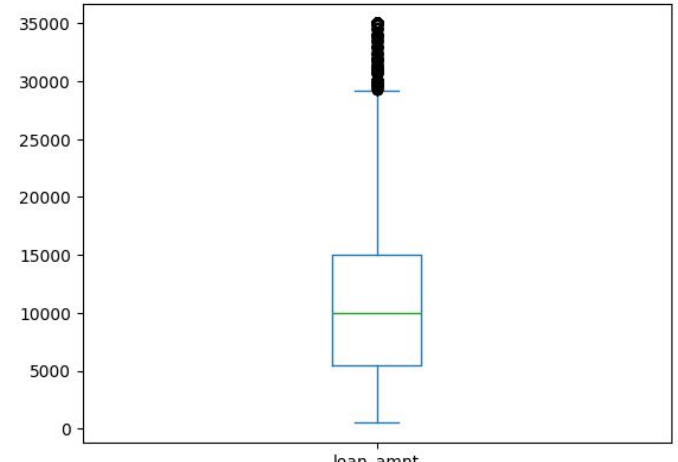


2. Loan Amount

- Median loan amount: 10000.0

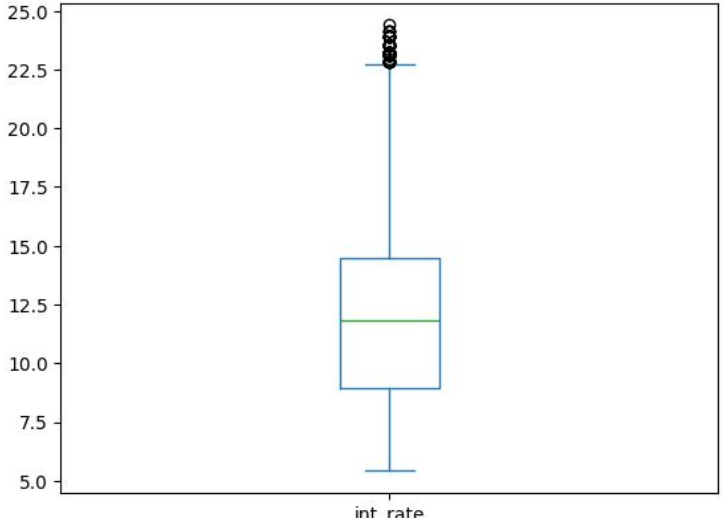
- Largest loan amount: 35000

- Smallest loan amount: 500



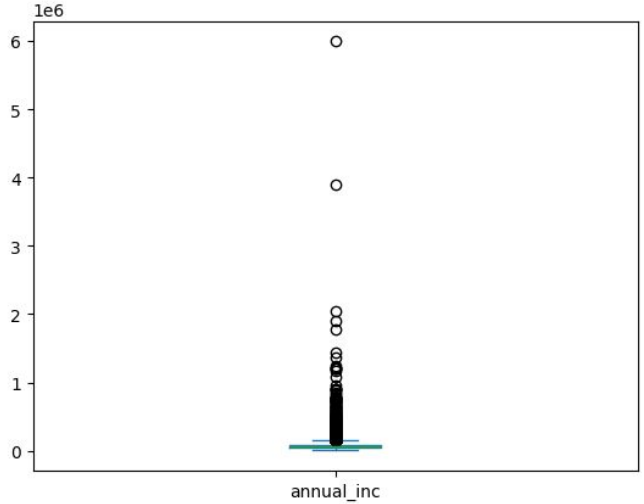
3. Interest Rate

- Median interest rate: 11.83
- Max interest rate: 24.4
- Lowest interest rate: 5.42



4. Annual Income

- Median Annual Income: 10000.0
- ***Largest Annual Income: 6000000.0****
- Lowest Annual Income: 500



Post mean and median comparison the outliers in the columns are dropped to ensure unbiased analysis.

Segmented Univariate Analysis

Create bins for quantitative columns like :1.Debt to income ratio (DTI) | 2.Interest rate | 3.Revolving line utilization rate.

1. Debt to income ratio:

- DTI <=8 : Very Low
- 8 - 12 : Low
- 12 - 14 : Moderate
- 16 - 20 : High

DTI > 20 : Very High

2. Interest Rate:

- Interest Rate < = 9 : Very Low
- 9 - 11 : Low
- 11 - 13 : Moderate
- 13 - 15 : High

Interest Rate > 15 : Very High

3. Revolving line utilization rate

- 0 - 20 : Very Low
- 20 - 40 : Low
- 40 - 50 : Moderate
- 50 - 60 : High
- 60 - 100 : Very High

Segmented Univariate Analysis

loan_status	dti_b	
Charged Off	Very Low	246
	Very High	223
	Moderate	205
	High	204
	Low	171
Fully Paid	Very Low	1245
	Very High	1071
	Moderate	919
	High	860
	Low	793

Name: count, dtype: int64

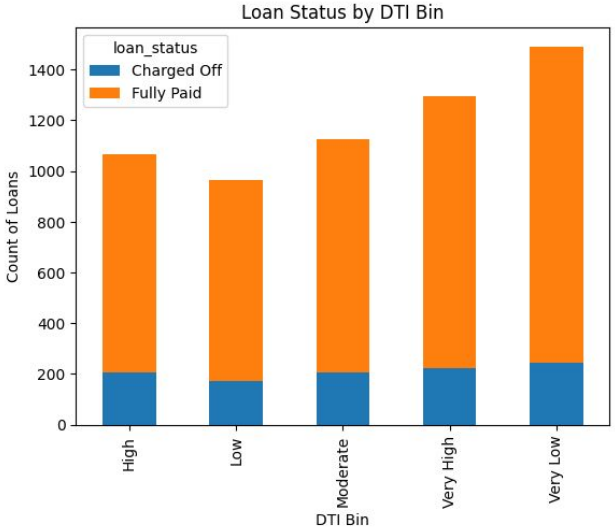
loan_status	int_rate_b	
Charged Off	Very High	293
	Moderate	253
	High	226
	Very Low	144
	Low	133
Fully Paid	Very Low	1362
	Moderate	1009
	High	922
	Low	879
	Very High	716

Name: count, dtype: int64

loan_status	revol_util_b	
Charged Off	Very High	897
	Very Low	116
	Moderate	14
	High	12
	Low	10
Fully Paid	Very High	3971
	Very Low	599
	High	112
	Low	104
	Moderate	102

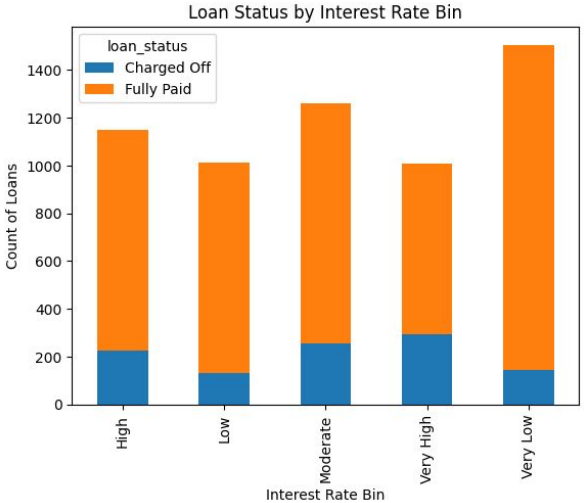
Name: count, dtype: int64

Bi Variate Analysis -



Segmented DTI ratios Vs Loan Status

- **Conclusion** : Loans given to higher DTI ratio > 16 for applicants are more likely to be defaulted

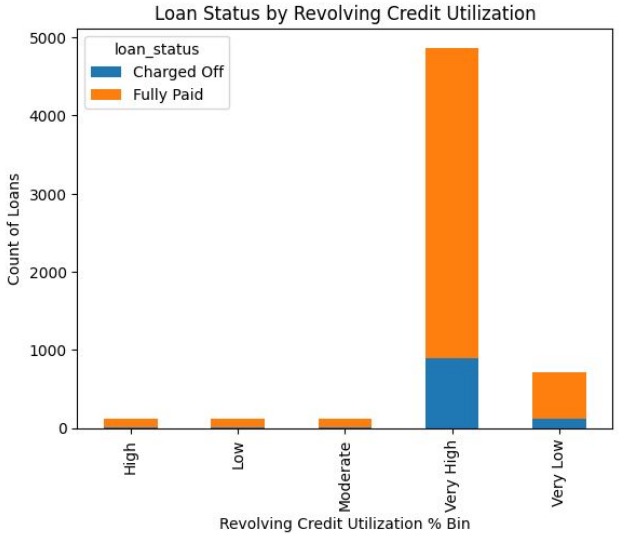


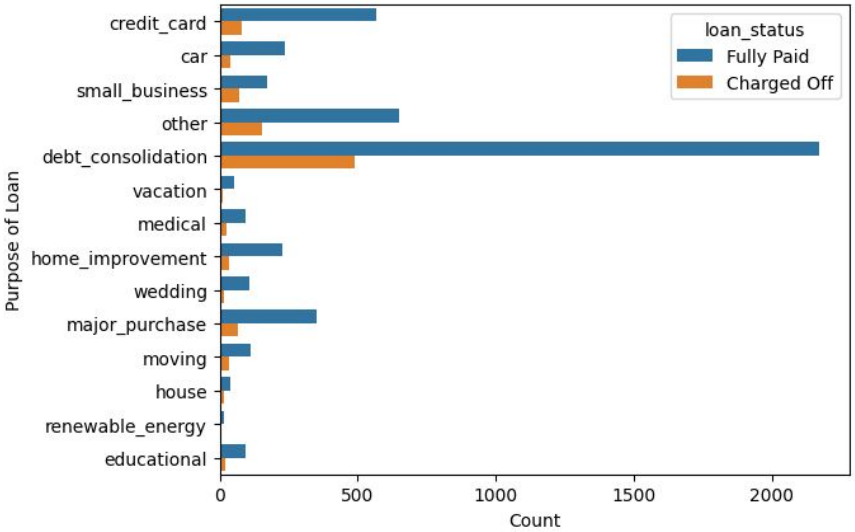
Segmented interest rates Vs Loan Status

- **Conclusion** : Higher interest rates loans > 13% are more likely to get defaulted.

Segmented Credit line utilization Rate Vs Loan Status

- **Conclusion** : Customers who have maxed out their credit utilization (high %) are more likely to default



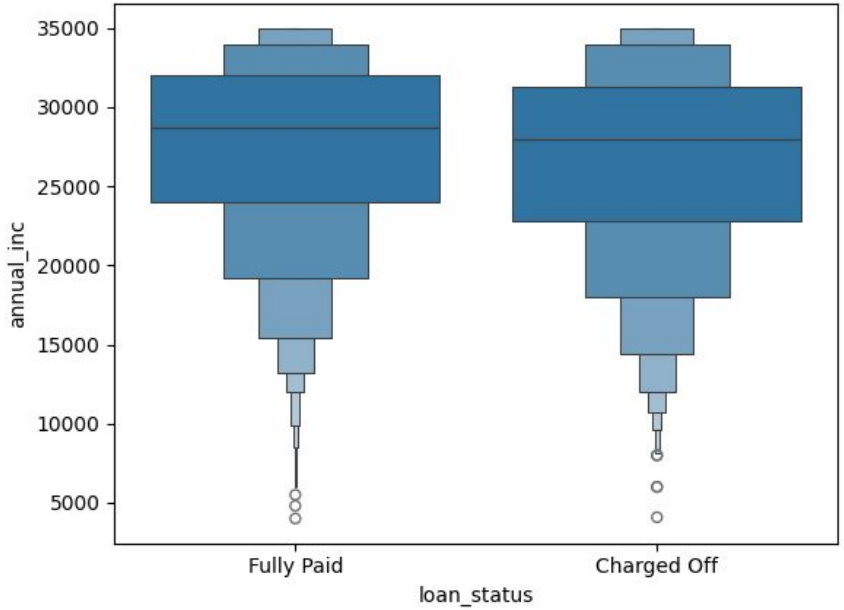


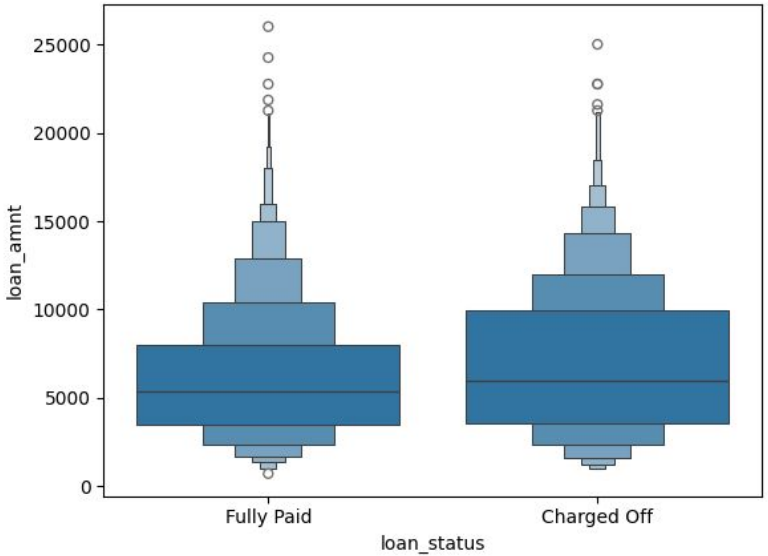
Purpose of the Loan Vs Loan Status

- **Conclusion** : Most of the loan defaults are occurring under debt consolidation category.

Annual income Vs Loan Status

- **Conclusion** : Majority of loan defaults are happening in the case of applicants whose annual income < 20000



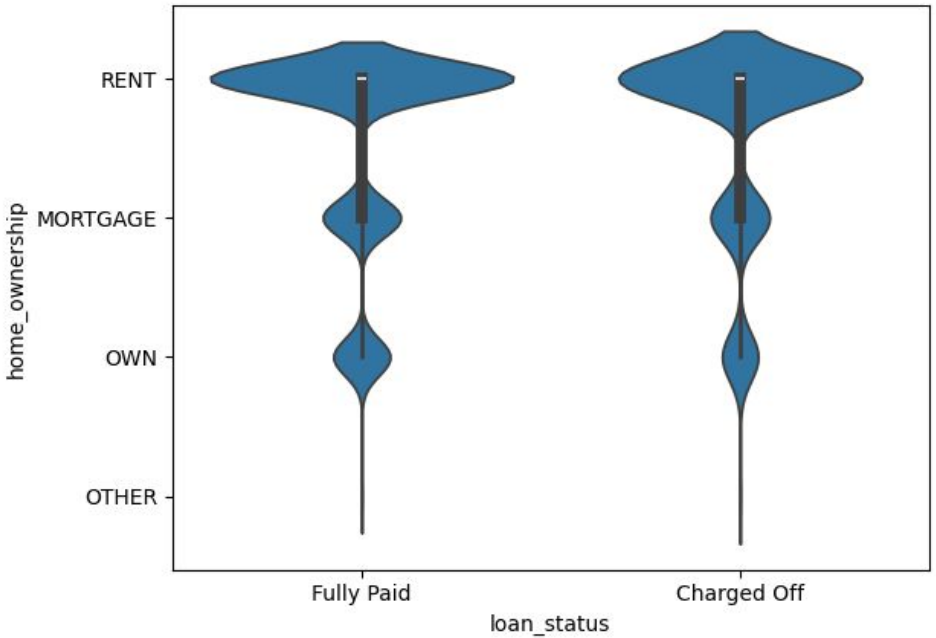


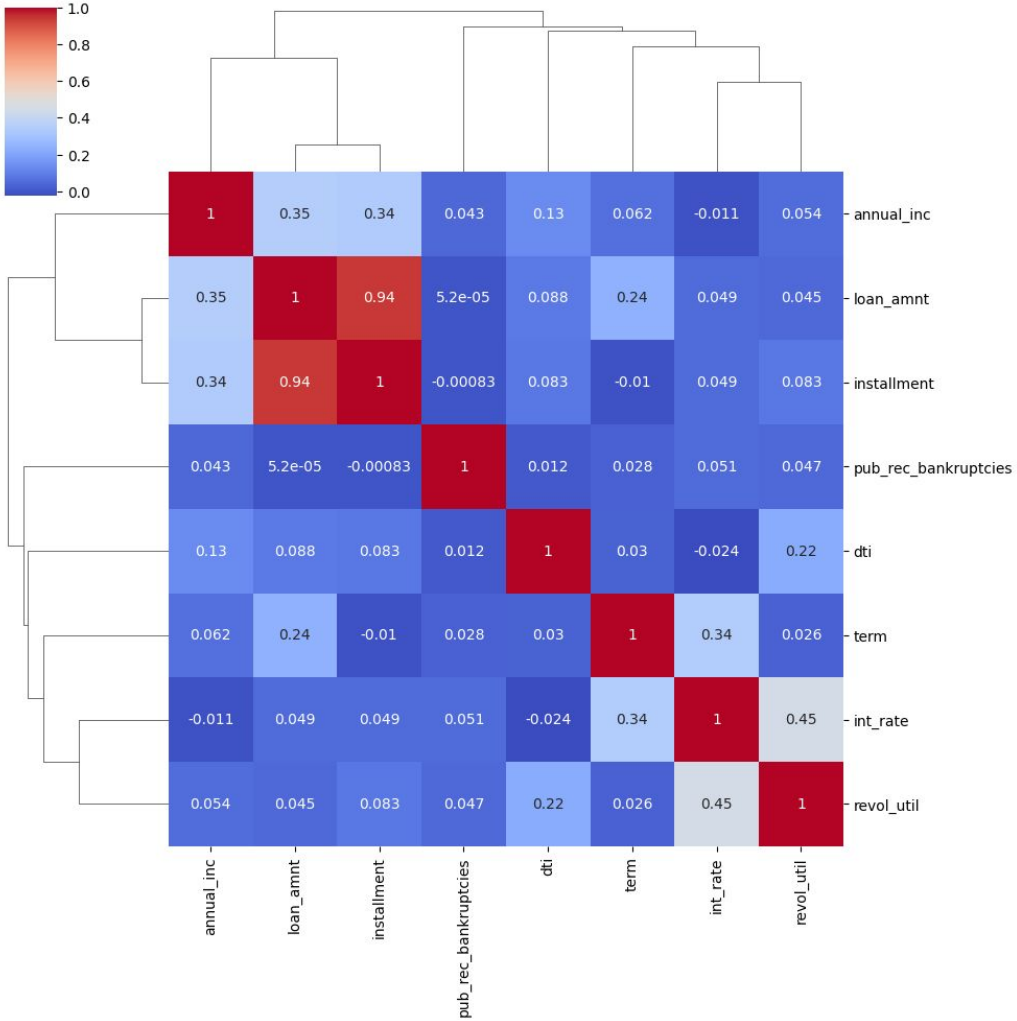
Loan amount taken Vs Loan Status

- **Conclusion** : We can infer that lower value loans has a higher probability of being charged off. Especially true for loan amounts between 2500 to 7500 with the average loan amount being 5000

Home Ownership Vs Loan Status

- **Conclusion** : Most of the loan defaults are by applicant who are living in rented homes




1.0
0.8
0.6
0.4
0.2
0.0
1 0.35 0.34 0.043 0.13 0.062 -0.011 0.054
0.35 1 0.94 5.2e-05 0.088 0.24 0.049 0.045
0.34 0.94 1 -0.00083 0.083 -0.01 0.049 0.083
0.043 5.2e-05 -0.00083 1 0.012 0.028 0.051 0.047
0.13 0.088 0.083 0.012 1 0.03 -0.024 0.22
0.062 0.24 -0.01 0.028 0.03 1 0.34 0.026
-0.011 0.049 0.049 0.051 -0.024 0.34 1 0.45
0.054 0.045 0.083 0.047 0.22 0.026 0.45 1
annual_inc
loan_amnt
installment
pub_rec_bankruptcies
dti
term
int_rate
revol_util

Correlation Analysis

- Find correlation between the most impacted factors from the lessons drawn from previous analysis

All the factors given below has strong correlation with Loan status

- Loan Amount
- Term of the loan
- Interest Rate
- Installment amount
- DTI
- Annual Income

- Conclusions

1. Annual Income, Loan Amount and installment shows a high degree of correlation that confirms our earlier analysis observations.
2. Interest rate granted to the loan is highly r=dependant on the credit utilization percentage of the customer.
3. The loan term/duration is highly influenced by the interest rate.
4. Interest rates are negatively correlated chances of default DTI and Annual Income of the applicant which implies that borrowers with strong indicators of financial health get low interest rates hence less chances of default.