EDA CASE STUDY

- BANK LOAN DATASET

Submitted by

Vasu Teotia & Vignesh M

Understanding of the background

We have been provided Loan.csv dataset from a consumer finance company. The primary attributes captured in data are

- Loans given by the institution
- Loan status which captures type of client (defaulter / fully paid / current) based on behaviour of payment
- Various other variable which are collected at time of loan sanction

Understanding of the problem

Given the above context , and based on the data collected, our team feels ultimate task is to be able t devise a model which can, based on data collected at the time of application, help us in determining if the applicant can be good candidate od defaulter in future

Setting up the problem statement

Given the understanding . We will try to figure out the variable which are the best predictors for defaulting behaviour from the given dataset.

Goals of Data Analysis

The main objective is to be able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Perform an analysis to understand the driving factors (or driver variables) behind loan default, i.e.the variables that are strong indicators of default.

The company can utilize this knowledge for its portfolio and risk assessment.

Importing libraries

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading Loan.csv to variable lcs

```
lcs = pd.read_csv(r"C:\LendingCaseStudy\loan.csv")
lcs.shape
```

Data Cleaning

After load we have 111 columns in lcs data frame variable
Remove Null values from variable lcs

```
lcs.dropna(axis = 1, how = 'all', inplace = True)
lcs.shape
```

After cleaning we have 57 columns in lcs

Drop additional columns we don't need these as these are mostly nulls.

```
lcs.drop(["id", "member_id", "url", "title", "emp_title", "zip_code",
"last_credit_pull_d", "addr_state","out_prncp_inv","total_pymnt_inv",
"funded_amnt", "delinq_2yrs", "revol_bal", "out_prncp", "total_pymnt",
"total_rec_prncp", "total_rec_int", "total_rec_late_fee", "recoveries",
"collection_recovery_fee", "last_pymnt_d", "last_pymnt_amnt",
"chargeoff_within_12_mths"], axis = 1, inplace = True)
```

```
lcs.shape
```

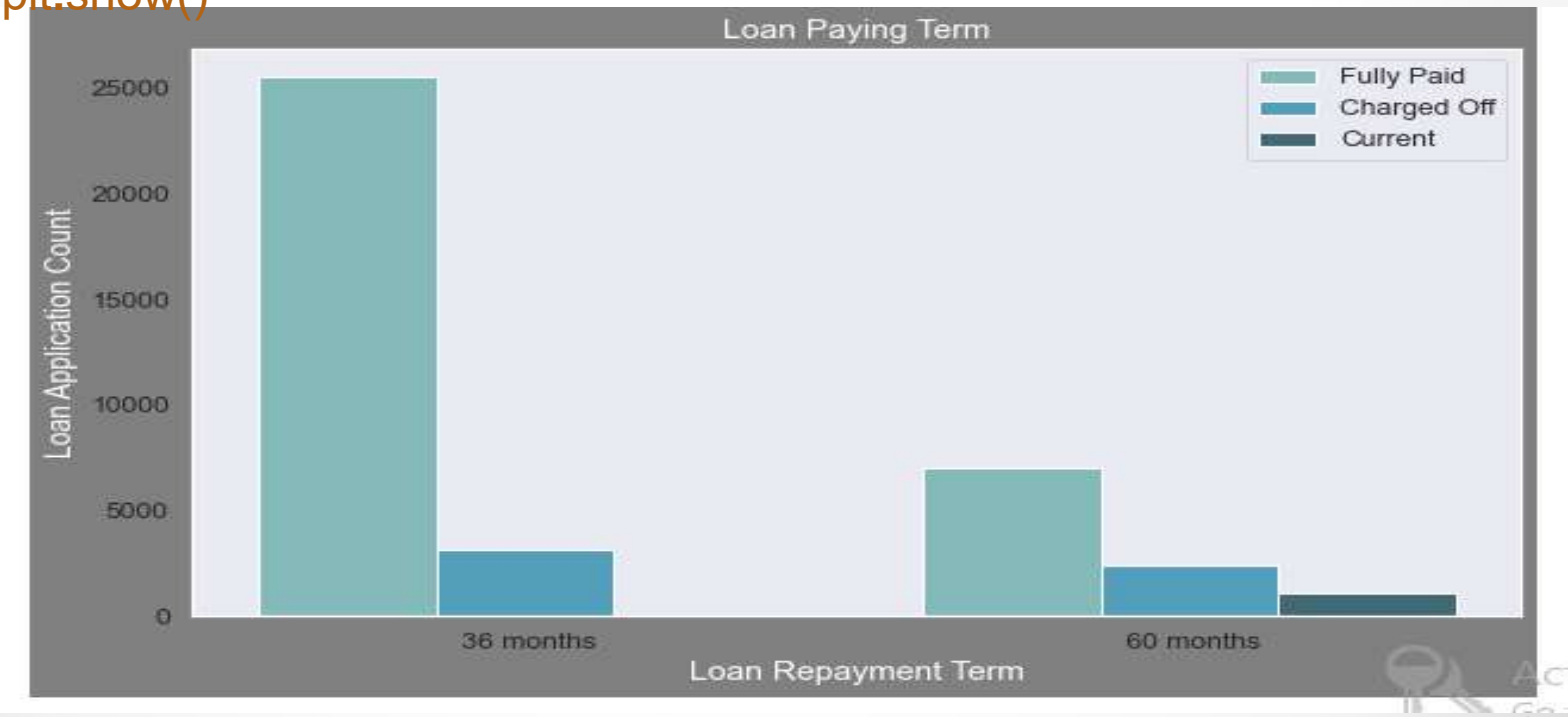
After cleaning we have 23 columns for Data Analysis

```
lcs.columns
```

```
Index(['loan_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade',
 'sub_grade', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status',
 'issue_d', 'loan_status', 'desc', 'purpose', 'dti', 'earliest_cr_line', 'inq_last_6mths',
'mths_since_last_delinq', 'open_acc', 'pub_rec', 'revol_util', 'total_acc'],
dtype='object')
```

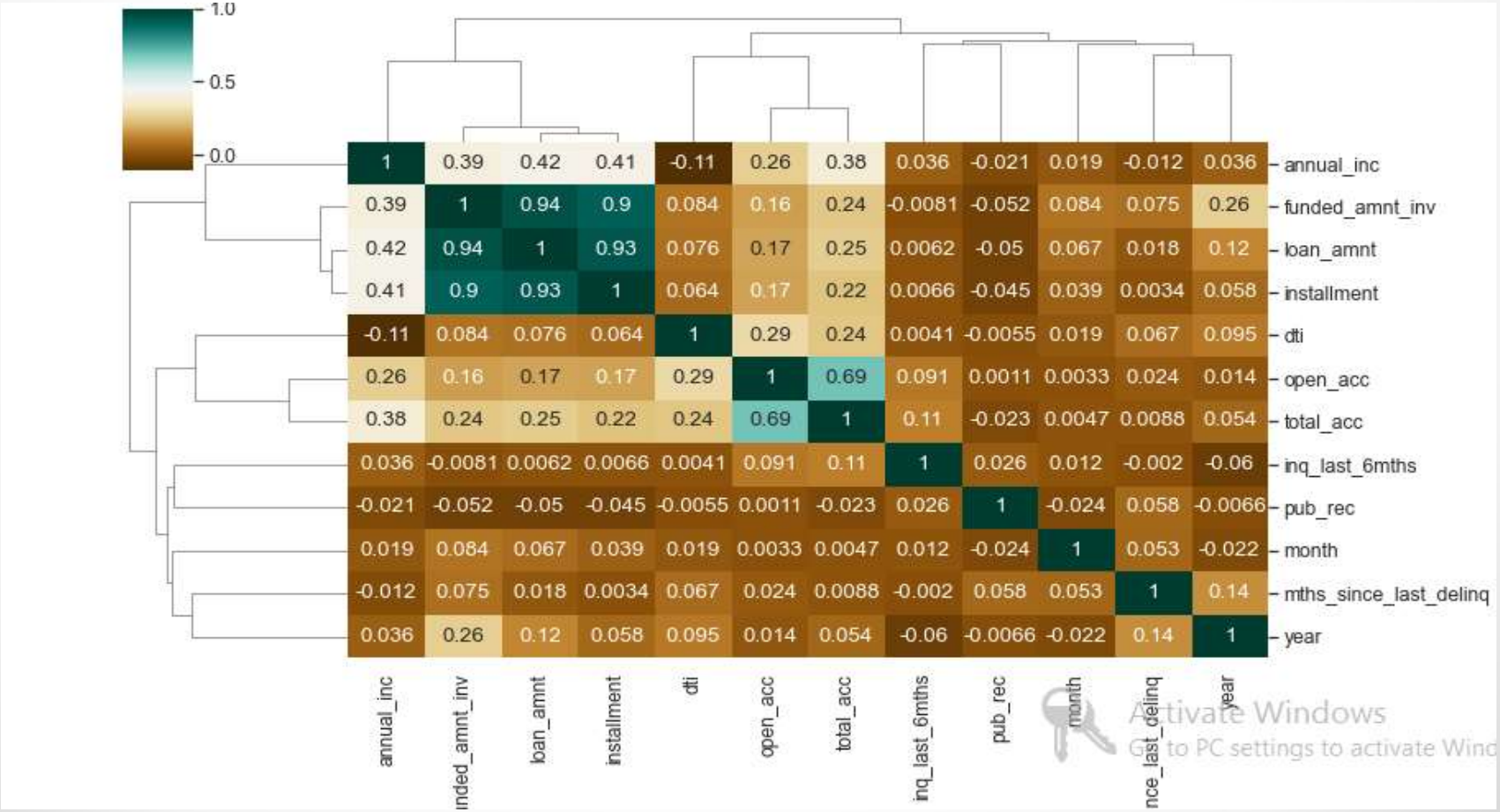
Univariate Analysis

```
plt.figure(figsize=(10,6),facecolor='grey')
ax=sns.countplot(x="term",data=lcs,hue='loan_status',palette='GnBu_d'
)
ax.set_title('Loan Paying Term',fontsize=14,color='w')
ax.set_xlabel('Loan Repayment Term',fontsize=14,color = 'w')
ax.set_ylabel('Loan Application Count',fontsize=14,color = 'w')
ax.legend(bbox_to_anchor=(1, 1))
plt.show()
```



Bivariate Analysis

```
loan_correlation = lcs.corr()
sns.set(font_scale=1.1)
sns.clustermap(loan_correlation,annot=True,figsize=(12, 8),cmap="BrBG")
plt.show()
```



Observations

The above analysis with respect to the charged off loans. There is a more probability of defaulting when :

- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%