Bank Loan Case Study using EDA

By Gaurav Chamate

PROBLEM STATEMENT

- **The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history.**
- **Because of that, some consumers use it to their advantage by becoming a defaulter.**

Two types of risks are associated with the bank's decision:

1. **If the applicant is likely to repay the loan, then not approving the loan results in a loss of business for the company.**

2. **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**

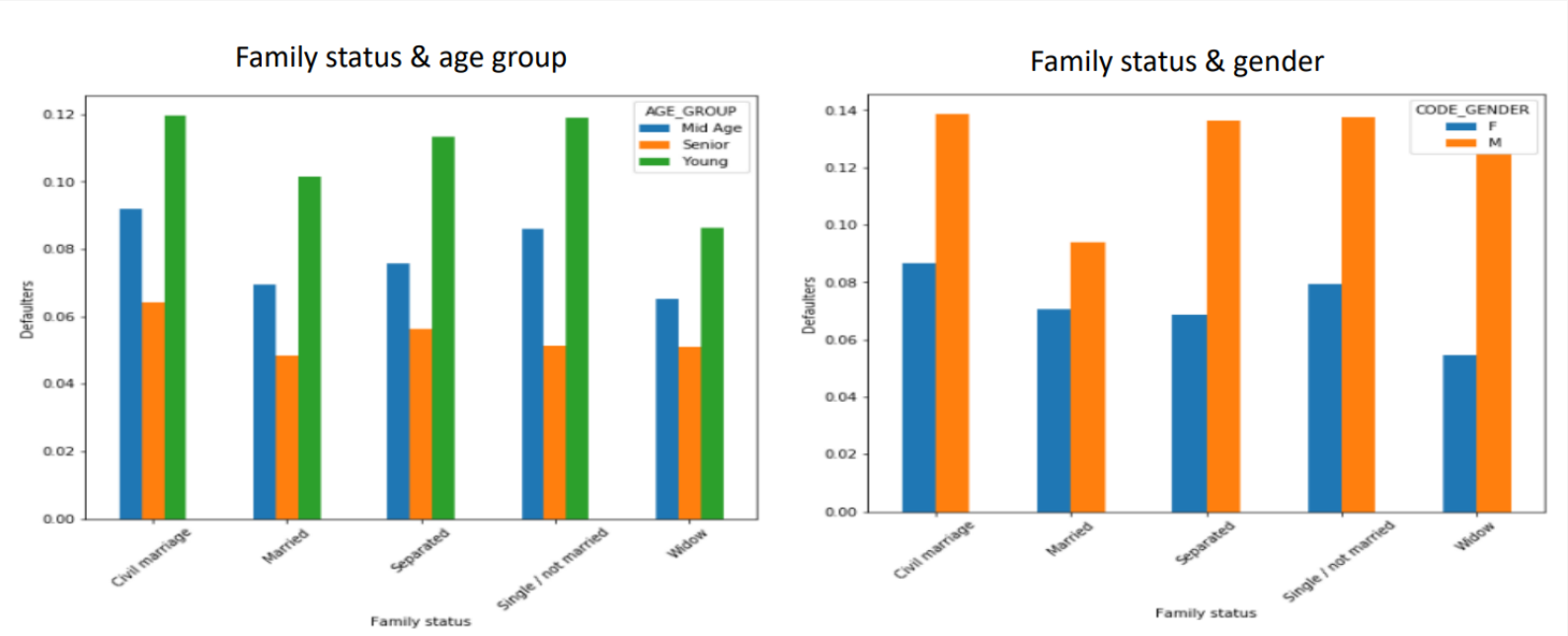
ANALYSIS APPROACH

Use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

1. Data Cleaning

- **Removing all columns with a null value**
- **Removing column with maximum no of missing value**
- **Replacing XAP/XNA with NA**
- **Binning of continuous variable**

Current Applicant Status:



Observations:-

1.The previous applications for portfolio Cards and POS are mostly defaulted .

2. Previously refused applications for Cash are also defaulted in higher rate. 3. Low external source

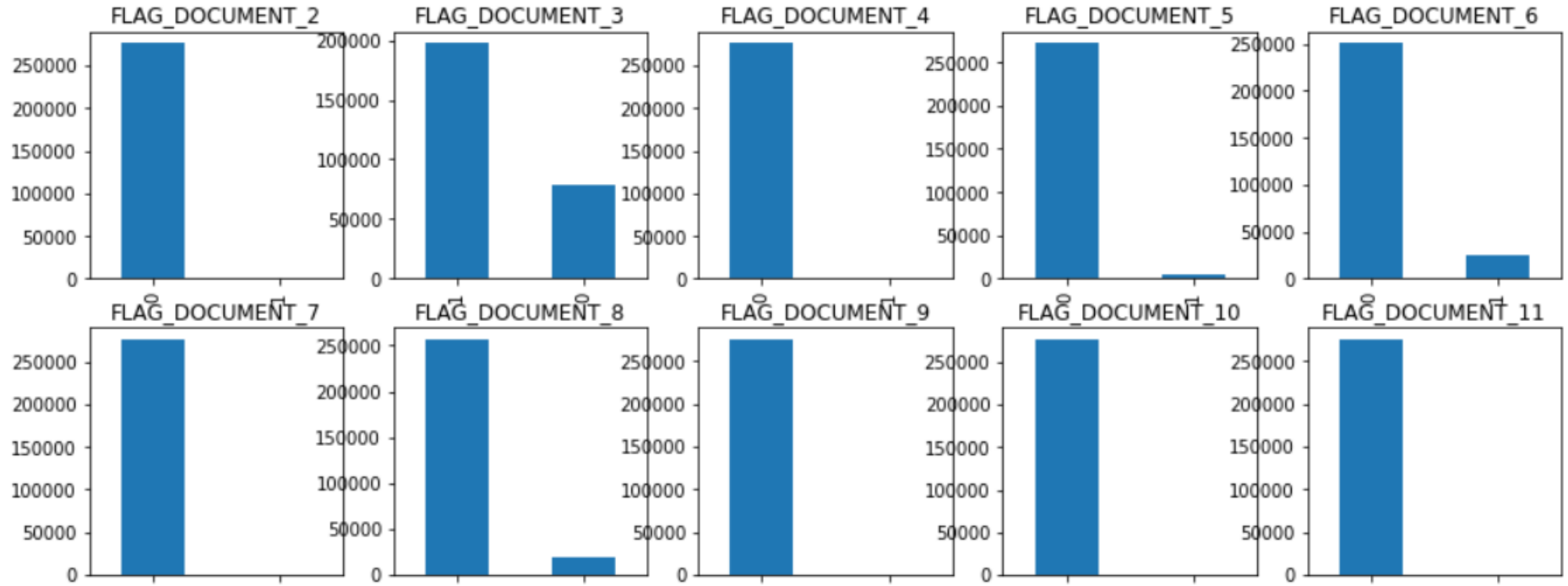
Recommendations:-

1. It is safer to grant loans for any portfolio for previously approved applicants.

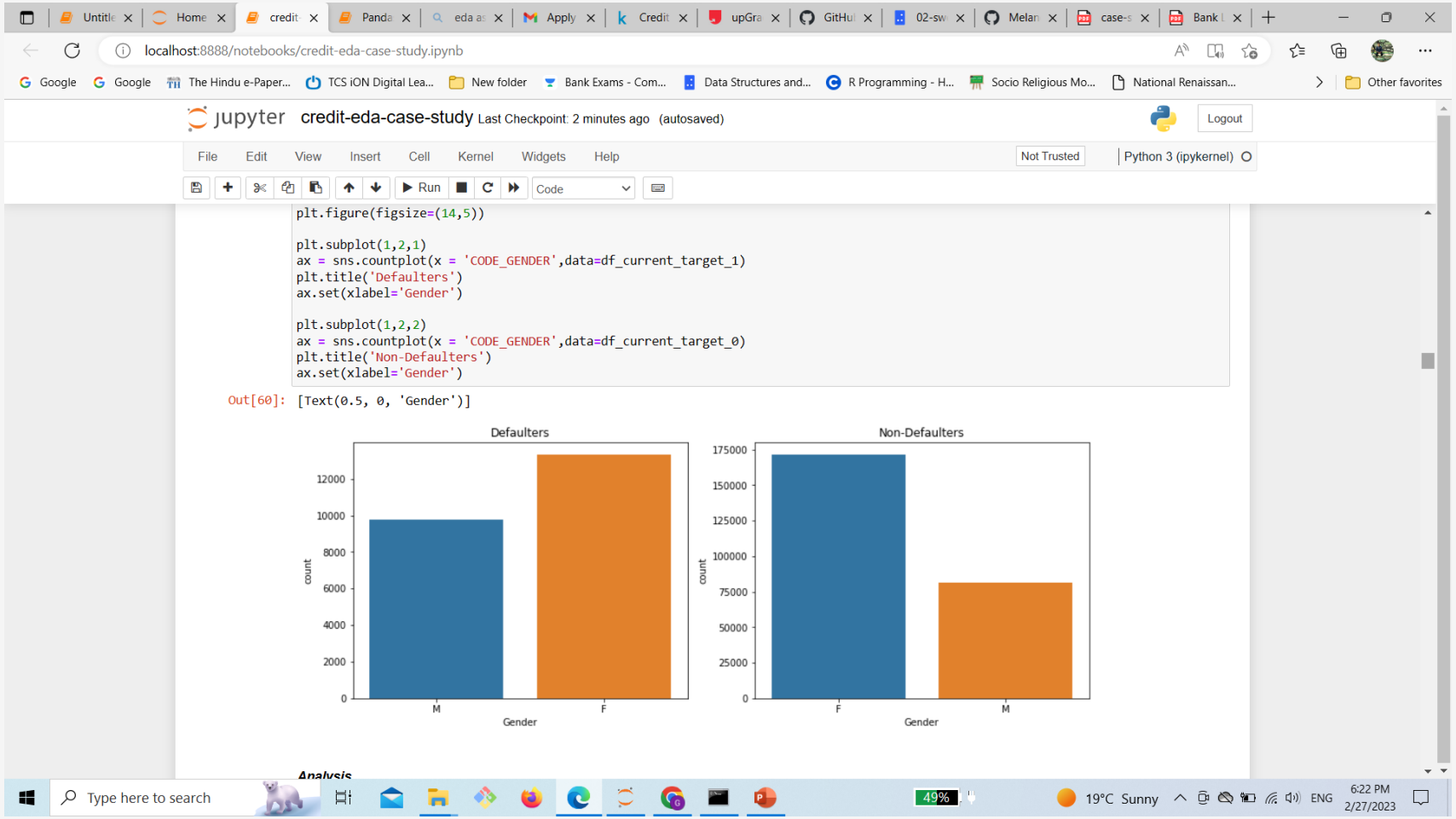
2. It is high risk to grant loans for applicants, who have poor external source score specially whose loan were previously refused, unused or cancel.

DATA IMBALANCE

```
# Plotting all the FLAG_DOCUMENT columns to check data imbalance
k=0
plt.figure(figsize=(15,15))
for i in range(2,22) :
    k=k+1
    plt.subplot(5, 5,k)
    col_name = 'FLAG_DOCUMENT_'+str(i)
    df_application_current[col_name].value_counts().plot(kind='bar');
    plt.title(col_name)
```



Gender wise Distribution



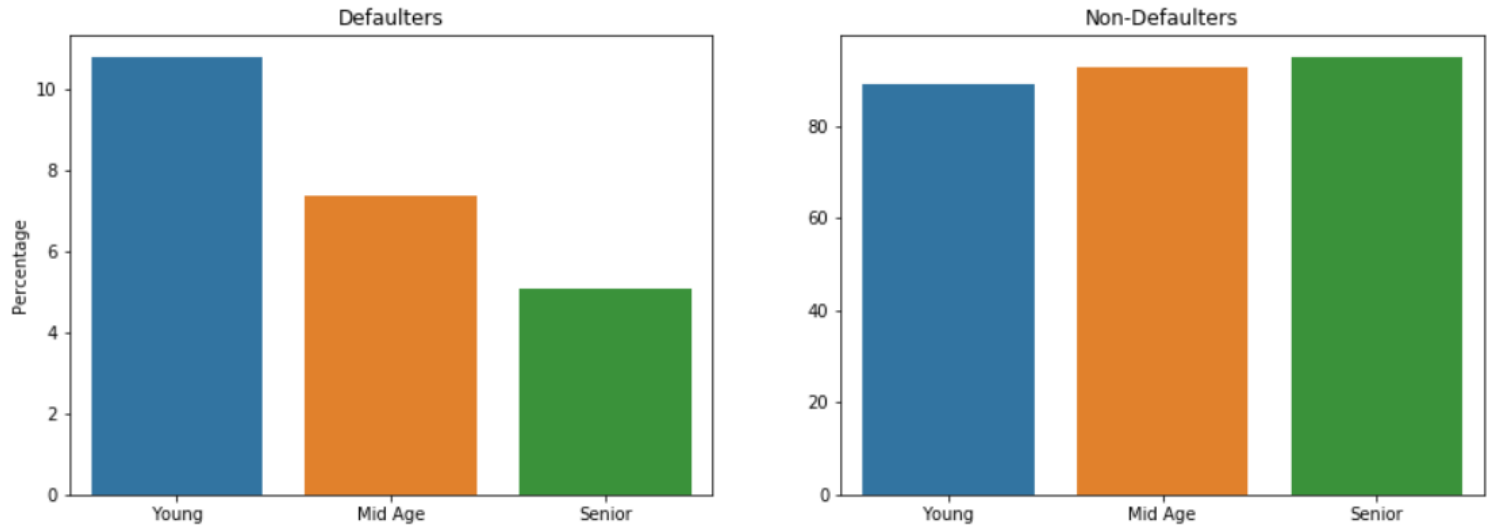
Categorical Analysis

In [72]:
```
# Plotting the above two dataframes
plt.figure(figsize=(15,5))

plt.subplot(1,2,1)
sns.barplot(x='AGE_GROUP',y='Percentage',data=df_agegroup_default)
plt.title('Defaulters')

plt.subplot(1,2,2)
sns.barplot(x='AGE_GROUP',y='Percentage',data=agegroup_nondefault)
plt.title('Non-Defaulters')
```

Out[72]: Text(0.5, 1.0, 'Non-Defaulters')



- **Previously loans refused people are most likely to default.**

CONCLUSION

Highly recommended groups:-

1. Approved clients in their previous applications.
2. Highly educated clients with higher income.
3. Clients with higher external source score.
4. Senior citizens in all categories.
5. Married clients compared to other family status.
6. Females are comparatively favorable than male.

High risk groups:-

1.Previously refused, cancelled or unused offer clients.
2. Low income groups with previously refused status.
3. Unemployed clients.
4. Poor external source scorer.
5. Young clients are comparatively riskier than mid age clients and senior citizens.
6. Lower secondary and secondary educated clients.

Thank You