EDA ASSINGMENT ON BANK DEFAULTER

Business Objective:

The driving factors behind the loan defaulter, i.e the variable which are strong indicator of loan default from the data provided we have to find out the useful informations. From the information company will utilize the knowledge for its portfolio and risk management.

THE ASSINGMENT IS DONE IN VARIOUS STEPS:

1. **Data understanding**
2. **Data Cleaning and Manipulation**
3. **Data analysis**
4. **Recommendations**

DATA UNDERSTANDING:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import files
import io
%matplotlib inline
```

Firstly imported all the library which are going to be used.
After that we are going to upload the two files provided which are

1. Application.csv
2. Previous_application.csv

Then we read the file one by one.

Step 1

Here application_data.csv file and previous_app_data.csv which have shape as (307511,122) and(1670214,37).

DATA CLEANING AND MANIPILATION

Now there are many columns so firstly I removed the columns which have 40% nan values. By which our data set become more easy to be analyse.

By removing the columns by 40% we get application_data 73 column and previous_app_data 26 columns.

The application_data columns:

```
0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   FLAG_OWN_CAR                 307511 non-null  object
 5   FLAG_OWN_REALTY              307511 non-null  object
 6   CNT_CHILDREN                 307511 non-null  int64
 7   AMT_INCOME_TOTAL             307511 non-null  float64
 8   AMT_CREDIT                   307511 non-null  float64
 9   AMT_ANNUITY                  307499 non-null  float64
 10  AMT_GOODS_PRICE              307233 non-null  float64
 11  NAME_TYPE_SUITE              306219 non-null  object
 12  NAME_INCOME_TYPE             307511 non-null  object
 13  NAME_EDUCATION_TYPE          307511 non-null  object
 14  NAME_FAMILY_STATUS           307511 non-null  object
 15  NAME_HOUSING_TYPE            307511 non-null  object
 16  REGION_POPULATION_RELATIVE   307511 non-null  float64
 17  DAYS_BIRTH                   307511 non-null  int64
 18  DAYS_EMPLOYED                307511 non-null  int64
 19  DAYS_REGISTRATION            307511 non-null  float64
 20  DAYS_ID_PUBLISH              307511 non-null  int64
 21  FLAG_MOBIL                   307511 non-null  int64
 22  FLAG_EMP_PHONE               307511 non-null  int64
 23  FLAG_WORK_PHONE              307511 non-null  int64
 24  FLAG_CONT_MOBILE             307511 non-null  int64
 25  FLAG_PHONE                   307511 non-null  int64
 26  FLAG_EMAIL                   307511 non-null  int64
 27  OCCUPATION_TYPE              211120 non-null  object
 28  CNT_FAM_MEMBERS              307509 non-null  float64
 29  REGION_RATING_CLIENT         307511 non-null  int64
 30  REGION_RATING_CLIENT_W_CITY  307511 non-null  int64
 31  WEEKDAY_APPR_PROCESS_START   307511 non-null  object
 32  HOUR_APPR_PROCESS_START      307511 non-null  int64
 33  REG_REGION_NOT_LIVE_REGION   307511 non-null  int64
 34  REG_REGION_NOT_WORK_REGION   307511 non-null  int64
 35  LIVE_REGION_NOT_WORK_REGION  307511 non-null  int64
 36  REG_CITY_NOT_LIVE_CITY       307511 non-null  int64
 37  REG_CITY_NOT_WORK_CITY       307511 non-null  int64
```

```
 38  LIVE_CITY_NOT_WORK_CITY      307511 non-null  int64
 39  ORGANIZATION_TYPE            307511 non-null  object
 40  EXT_SOURCE_2                 306851 non-null  float64
 41  EXT_SOURCE_3                 246546 non-null  float64
 42  OBS_30_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 43  DEF_30_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 44  OBS_60_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 45  DEF_60_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 46  DAYS_LAST_PHONE_CHANGE       307510 non-null  float64
 47  FLAG_DOCUMENT_2              307511 non-null  int64
 48  FLAG_DOCUMENT_3              307511 non-null  int64
 49  FLAG_DOCUMENT_4              307511 non-null  int64
 50  FLAG_DOCUMENT_5              307511 non-null  int64
 51  FLAG_DOCUMENT_6              307511 non-null  int64
 52  FLAG_DOCUMENT_7              307511 non-null  int64
 53  FLAG_DOCUMENT_8              307511 non-null  int64
 54  FLAG_DOCUMENT_9              307511 non-null  int64
 55  FLAG_DOCUMENT_10             307511 non-null  int64
 56  FLAG_DOCUMENT_11             307511 non-null  int64
 57  FLAG_DOCUMENT_12             307511 non-null  int64
 58  FLAG_DOCUMENT_13             307511 non-null  int64
 59  FLAG_DOCUMENT_14             307511 non-null  int64
 60  FLAG_DOCUMENT_15             307511 non-null  int64
 61  FLAG_DOCUMENT_16             307511 non-null  int64
 62  FLAG_DOCUMENT_17             307511 non-null  int64
 63  FLAG_DOCUMENT_18             307511 non-null  int64
 64  FLAG_DOCUMENT_19             307511 non-null  int64
 65  FLAG_DOCUMENT_20             307511 non-null  int64
 66  FLAG_DOCUMENT_21             307511 non-null  int64
 67  AMT_REQ_CREDIT_BUREAU_HOUR   265992 non-null  float64
 68  AMT_REQ_CREDIT_BUREAU_DAY    265992 non-null  float64
 69  AMT_REQ_CREDIT_BUREAU_WEEK   265992 non-null  float64
 70  AMT_REQ_CREDIT_BUREAU_MON    265992 non-null  float64
 71  AMT_REQ_CREDIT_BUREAU_QRT    265992 non-null  float64
 72  AMT_REQ_CREDIT_BUREAU_YEAR   265992 non-null  float64
```

And the previous_app_data columns are:

```
0   SK_ID_PREV                  1670214 non-null  int64
 1   SK_ID_CURR                  1670214 non-null  int64
 2   NAME_CONTRACT_TYPE          1670214 non-null  object
 3   AMT_ANNUITY                 1297979 non-null  float64
 4   AMT_APPLICATION             1670214 non-null  float64
 5   AMT_CREDIT                  1670213 non-null  float64
 6   AMT_DOWN_PAYMENT            774370 non-null   float64
 7   AMT_GOODS_PRICE             1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START  1670214 non-null  object
 9   HOUR_APPR_PROCESS_START     1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT 1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY      1670214 non-null  int64
 12  RATE_DOWN_PAYMENT           774370 non-null   float64
 13  RATE_INTEREST_PRIMARY       5951 non-null     float64
```

```
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
```

In the data we have some error as there are negative values in some of the columns. The negative value columns in the application_data are as follows:

```
['DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH',
'DAYS_LAST_PHONE_CHANGE']
```

And the negative value column in previous_app_data is ‘DAYS_DECISION’

So we have to remove the negative value and make all the column positive.

Now the data is ready for analysis.

DATA ANLYSIS

We have two data set:

1. Final_app_data
2. Pre_app_data

- First we will analyse final_app_data
- Then we will analyse pre_app_data
- Then we will merge the two data and then analyse the merged data.

So from final_app_data:

- We have analysed different columns and take out mean and median of them.
- Then with the help of bar graph and histogram we see the outliners and the get the facts relate to it.
- Some of the important columns of final_app_data are ATM_CREDIT,AMT_ANNUITY,AMT_GOODS_PRICE,AMT_REQ_CREDIT_BUREAU_YEAR,SK_ID_CURR,TARGET,DAYS_EMPLOYED,NAME_INCOME_TYPE,NAME-FAMILY_STATUS,OCCUPATION_TYPE etc.
-

So now from pre_app_data:

- We have analysed different columns of pre_app_data through mean and median.
- Then with the help of different graphs we have seen outliners.
- Some important columns of pre_app_data are AMT_ANNUITY,AMT_APPLICATION,AMT_CREDIT,AMT_GOODS_PRICE,CNT_PAYMENT,SK_ID_PREV,SK_ID_CURR,NAME_CASH_LOAN_PURPOSE,NAME_PORTFOLIO etc.

After that merged data:

- From the merged data univariate and bivariate analysis is done.
- From this we get the relation between previous loan default condition.
- The important columns of merged data is AMT_ANNUITY,AMT_GOODS_PRICE,SK_ID_CURR,AMT_CREDIT,NAME_CONTRACT_TYPE,WEEKAY_APPR_PROCESS_START,HOUR_APPR_PROCESS_START etc.

RECOMMENDATION:

- From the data imbalance we get that only 8.1% on the previous data have defaulted and rest 91.9% have not defaulted.
- The data was imbalance on higher level.

- With the help of univariate and bivariate analysis we see that there are few important columns which give us clear idea about the data such as NAME_INCOME_TYPE,AMT_ANNUITY,TARGET,SK_ID_CURR etc.
- As there was 121,73 columns (~40%) of the columns have missing values.
- We have analysed the outliers and handled it to get out relative required information.
- The data consists of numeric and categorical columns through which we get information.