1. Importing All The Necessary Modules

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import itertools

( %matplotlib inline )

**Dataset 1 :-**

“application\_data(1).csv”

renamed to application\_data.csv

* **Reading and Understanding the data**.

app\_data = pd.read\_csv("application\_data.csv")

pre\_data = pd.read\_csv("previous\_application.csv")

* **Understanding the Dataset**

First we checked for the null values and missing values in the given dataset.

After that we count the null values and NAN values for each column in our data set.

Later we calculated the percentage of null values in each column.

We arranged the null values in desc order and got an observation of 70% as the highest percentage null value present in app\_data data frame.

1. Data Cleaning and Manipulation

* Finding Null Values
* Dealing with Null Values

After Sorting Null Values in Descending Order According to their Percentage we found out 50 columns containing Null Values And

Highest percentage of Null Values in a Column is 70 %.

We Have shown Graphical Representation of Null Values on Basis of counts using Bar Plots.

* Dropping the Null Values

Here we are dropping Columns Containing Null Values more than 30%.

( first we make list of columns named “ labels “ containing Null Values more than 30%)

No of columns dropped By = 49 Columns

considering above conditions

though we have 50 columns containing null values more than 30%

We manually avoid to drop a column named “ occupation\_type ”

Containing 31% of Null Values As we considered that column

to be crucial and important for the Operation.

We didn’t impute OCCUPATION\_TYPE because it may contain some useful information, so imputing it with mean or median doesn’t make any sense.

We’ll impute ‘OCCUOATION\_TYPE” later by analyzing it.

Here we are using ‘Describe’ command to find all the columns which have higher difference between Max and 75 percentile and also which should

be considered approximately an ‘outlier’ and to be kept out of Consideration as its max value is so high.

to confirm the ‘outliers’ we are using Box plot graph.

**Following outliers conclusions are made on the basis of Box plot.**

It can be seen that in current application data

These columns have some numbers of outliers :-

* AMT\_ANNUITY
* AMT\_CREDIT
* AMT\_GOODS\_PRICE
* CNT\_CHILDREN
* AMT\_INCOME\_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high income when compared to the others.
* DAYS\_BIRTH has no outliers which means the data available is reliable.
* DAYS\_EMPLOYED has outlier values around 350000(days) which

is around 958 years which is impossible and hence this has to be an incorrect entry.

**While doing Cleaning and inspection of the dataset we have encountered Problems as object data is a very crucial type of data to perform operation .**

**So here we are converting object data type to categorical.**

* **The columns which are converted into categorical data type are :-**

1. 'NAME\_CONTRACT\_TYPE' 7. 'NAME\_EDUCATION\_TYPE'

2. 'CODE\_GENDER' 8. 'NAME\_FAMILY\_STATUS'

3. 'FLAG\_OWN\_CAR' 9. 'NAME\_HOUSING\_TYPE'

4. 'FLAG\_OWN\_REALTY' 10. 'OCCUPATION\_TYPE'

5. 'NAME\_TYPE\_SUITE' 11. 'WEEKDAY\_APPR\_PROCESS\_START'

6. 'NAME\_INCOME\_TYPE' 12. 'ORGANIZATION\_TYPE'

* **After imputing and cleaning of the data we have 73 columns which are considered to be clean and can be used for the further operations.**
* **So these columns are being used for further Data Analysis.**

**Dataset (2)**

Here we are dealing with continuous variables "AMT\_ANNUITY", "AMT\_GOODS\_PRICE"

To impute the null values in continuous variables, we plotted the distribution of the columns and used

* median if the distribution is skewed
* mode if the distribution pattern is preserved.
* **Observation :-**

There is a single peak at the left side of the distribution and it indicates the presence of outliers and hence imputing with mean would not be the right approach so therefore we have approached for imputing with median.

* The original distribution is close to the distribution of data imputed with mode.

so in this case, we will impute mode for missing values

( Imputing CNT\_PAYMENT with 0 as the NAME\_CONTRACT\_STATUS for these

indicate that most of these loans were not started )

* There are several peaks along the distribution therefore we imputed using the mode.

**Here we are doing Cleaning and inspection on the dataset and**

**After that we are converting object data type to categorical.**

* **The columns which are converted into categorical data type are :-**

1. ‘NAME\_CONTRACT\_TYPE' 9. 'NAME\_CLIENT\_TYPE'

2. 'WEEKDAY\_APPR\_PROCESS\_START ‘ 10. 'NAME\_GOODS\_CATEGORY

3. 'FLAG\_LAST\_APPL\_PER\_CONTRACT ‘ 11. 'NAME\_PORTFOLIO'

4. 'NAME\_CASH\_LOAN\_PURPOSE' 12. 'NAME\_PRODUCT\_TYPE'

5. 'NAME\_CONTRACT\_STATUS' 13. 'CHANNEL\_TYPE'

6. 'NAME\_PAYMENT\_TYPE' 14. 'NAME\_SELLER\_INDUSTRY'

7. 'CODE\_REJECT\_REASON' 15. 'NAME\_YIELD\_GROUP

8. 'NAME\_TYPE\_SUITE' 16. 'PRODUCT\_COMBINATION'

* **Following outliers conclusions are made on the basis of Box plot.**

It can be seen that in previous application data

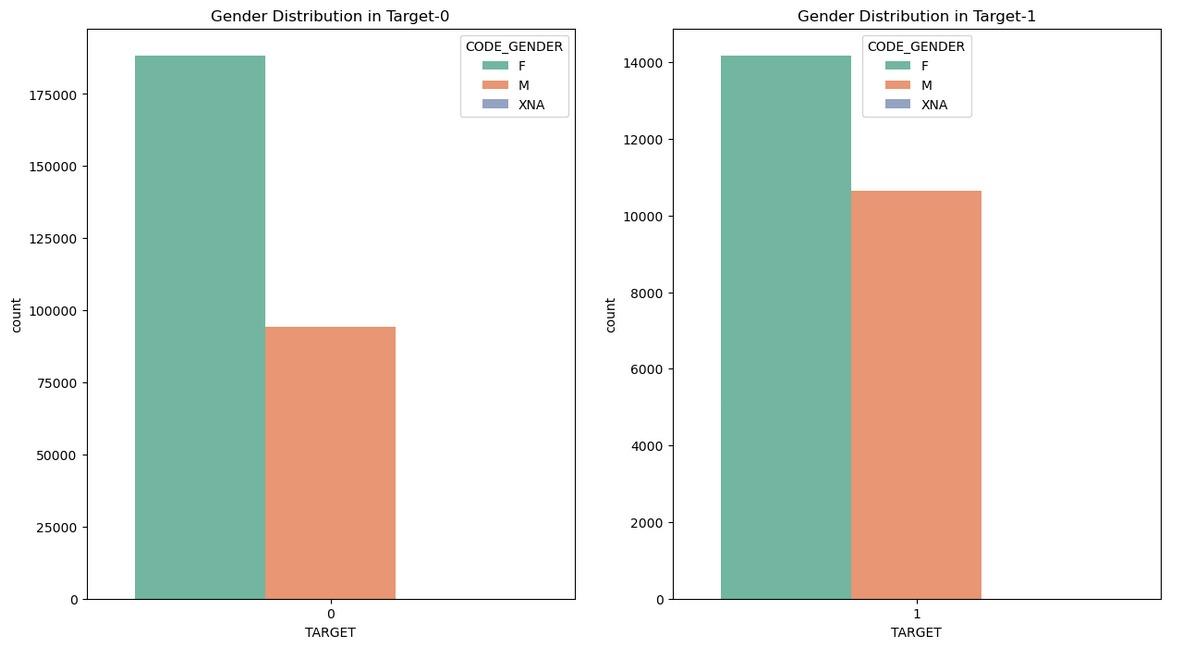
* AMT\_ANNUITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE, SELLERPLACE\_AREA have a huge number of outliers.
* CNT\_PAYMENT has few outlier values.
* DAYS\_DECISION has a small number of outliers indicating that these previous application decisions were taken long back.

**Insights / Business Conclusions :-**

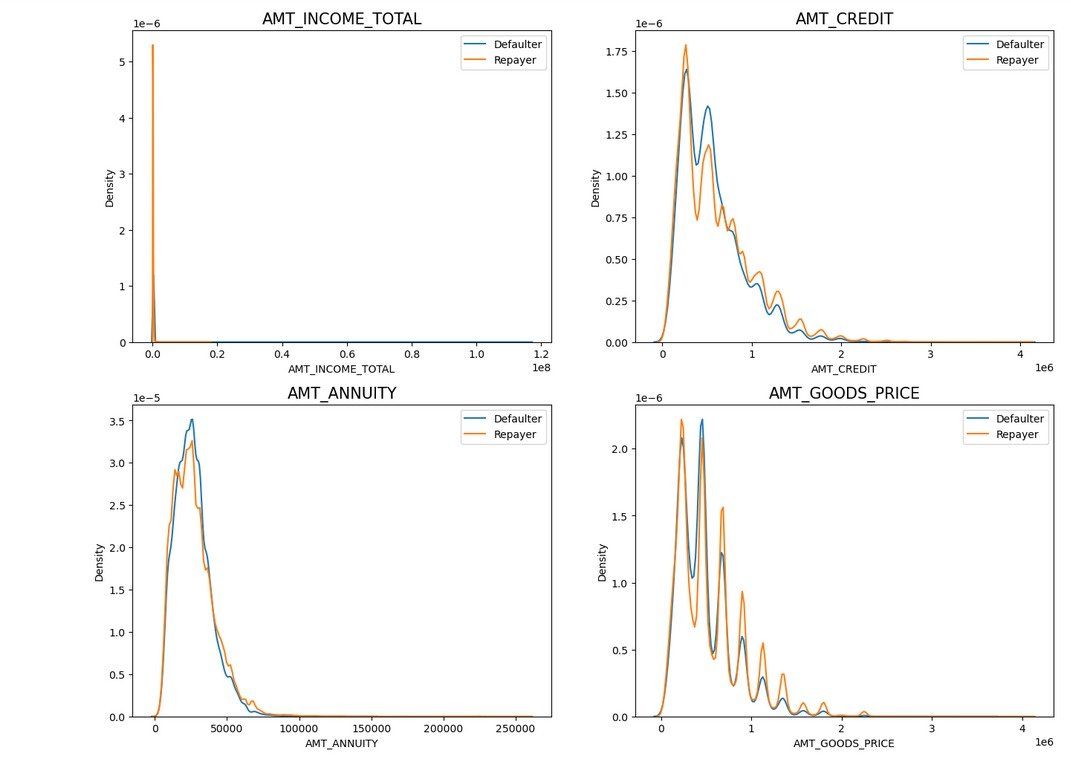
1. **After the Analysis of the categorical data which we have used for operations, It seems like Female clients applied higher than male clients for loan**
2. **Higher number of Female clients are non-defaulters as compared to male clients.**

**Also ,**

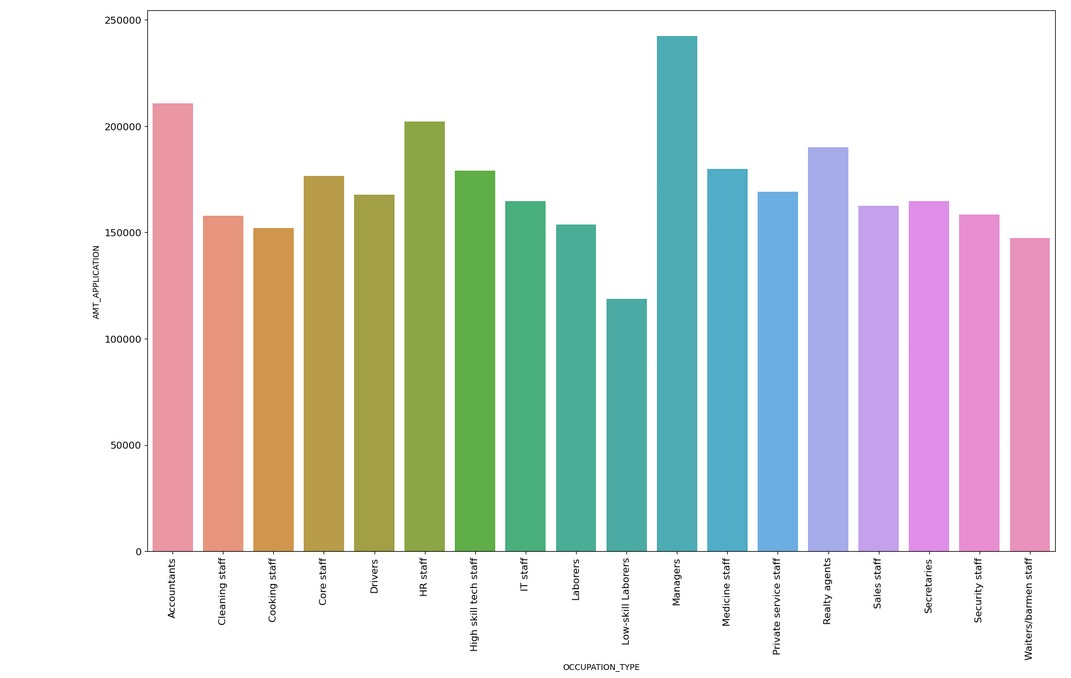
1. **Higher number of Female clients are defaulters as compared to male clients.**

****

1. **Most no of loans are given for goods price below 10 lakhs**
2. **Most people pay annuity below 50K for the credit loan**
3. **Credit amount of the loan is mostly less than 10 lakhs**
4. **The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision**

****

1. **Most of the loans are taken by Laborers, followed by Sales staff.**
2. **IT staff are less likely to apply for Loan.**
3. **Category with highest percent of defaultess are Low-skilled Laborers , followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff**

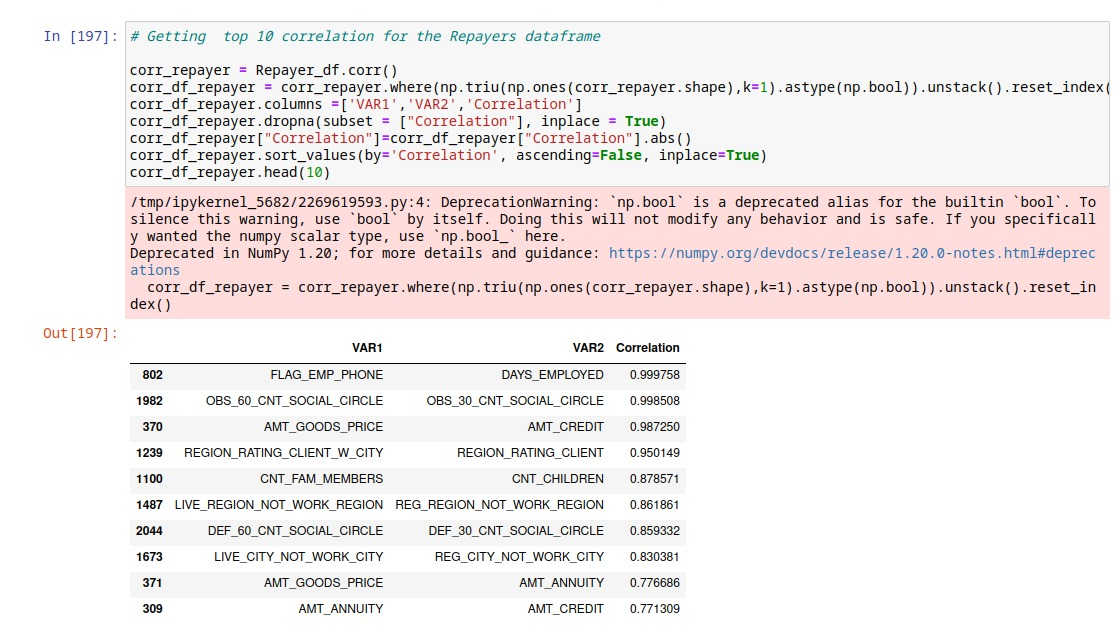
****

* **Inferences of Top 10 Correlating factors amongst repayers**

1. Credit amount is highly correlated with:

* Goods Price Amount
* Loan Annuity
* Total Income

2. We can also see that repayers have a high correlation in the number of days employed.



***THANK YOU***