Data science Assessment-1 (Reunion)

* The dataset given has two files (applicant.csv & loan.csv). One column is common between two given files and target is to predict low credit risk or high predict risk. so, I combined the two data csv files.
* After combining we have 27 attributes (columns) and 1000 applications information (rows).
* This is a imbalanced dataset with target having 700 low credit risk and 300 high credit risk. As dataset is small so I don’t want to make it balanced.

# Task-1

* We have separated numerical columns and categorical columns from the final dataset to check some insights.

**Insights: -**

1.70% of applicants in dataset are Male

2. Around 55% of applicants in dataset are single (un-married)

3. Around 34% of applicants have been employed for at least 1 year

4. Around 65% of the applicants are skilled-employed/official

5.More than 52% of applicants have good loan history (repaid their existing loans)

* Segmentation of customers based on: -

1. Gender
2. Age
3. Income
4. Education level
5. Occupation
6. Marital status
7. Loan history

* Age between 25- 50 are credit worthy and person with good credit accounts to be credit worthy. we consider both
* On numerical data, using displot we can see data was partially skewed.

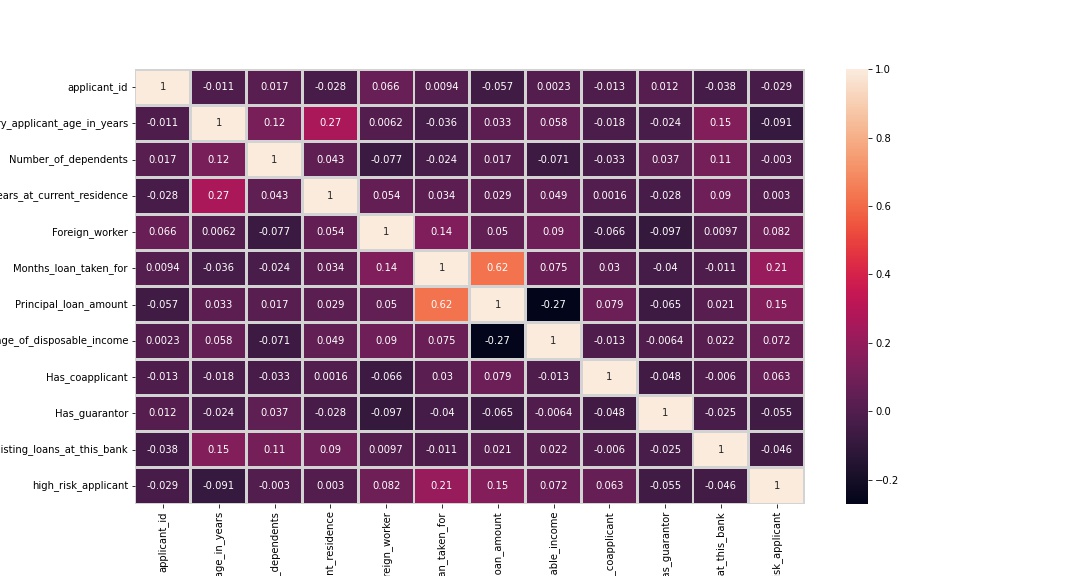
## **Task-2**

1. As the dataset is small so we have included all features for modelling. But after saving the best model I have also tried feature selection (chi2 and RFE technique) to 15 features but the accuracy is not much difference its same to original data.
2. No new feature is derived from the existing features
3. There are no missing values in numerical columns but for categorical columns there are huge number of missing values. Out of 15 categorical features 4 columns have more than 400 missing values remaining 5 columns have a smaller number of missing values.

I have used two methods to fill null values. Random sampling for higher null values and mode sampling for lower null values in categorical features

1. All Categorical features are converted to numeric using Label Encoding
2. Foreign worker and co-applicant features are highly correlated with target feature. Number of dependents and years at current residence are less correlated with target column because this years at current residence is not that much useful for paying back loan. Foreign worker has high chance of paying back the loan.
3. We are not going to drop any feature after the correlation matrix data. Because our dataset is small so we keep all features in model Building
4. I will try Logistic regression model first as it is binary classification problem and Random Forest, Gradient Boosting Algorithm.
5. For random forest confusion matrix, TN= 201, TP= 18, FP= 13, 13 times model predicted 1 but it is 0 (low credit risk). For GBC confusion matrix TN = 192, TP= 25, FP = 22.
6. Tuning number of estimators i.e., no. of decision trees using GridSearchCV. As the no. of trees increases accuracy also increases but after few trees this change in accuracy is insignificant. As the no. of trees increases time complexity and cost increases.
7. Precision metric for random forest after hyper-parameter tuning is best metric that is high credit risk (low chance of paying back loan) is very good up to 59%
8. Random forest has 72% test accuracy score, we will save this model into proper file name using pickle and load it. If we want to create web app with this model data which is saved we can do it by using streamlit or Flask

**Correlation matrix data**



**Conclusion**: - Model is trained with all features 26 independent variables and scaling is not done to our final data before model building because all data values are not much out of range. For choosing best model i preferred precision score rather than accuracy\_score. Hyper-parameter is not done to all models but i should have cross-checked once