EDA for Application Data & Previous Application Data:

Imports Libraries:

Import all the necessary Libraries and Load the CSV file as DataFrame.

Data Inspection:

Then inspect the Data thoroughly like its Shape, columns, dtypes, info(), describe() etc.

Check and Handle Missing values:

Check all the missing values and handle them. We can replace any missing value by using fillna().

Syntax- *dataframe*.fillna(value, method, axis, inplace, limit, downcast) where only value is required and all others are optional.

Cleaning Data:

Then Clean the Data by dropping columns which have more than or equals to 40% of missing values

Imputation:

Find 5 columns where the % of missing values <13 to 15(Depends on the Data). Fill the missing values by data imputation techniques.

we can observe that 4 values defined as XNA as we can count this as Female as there are many more numbers of female than male.

we can observe that there are many missing values in "Occupation Type" which we can impute from "Organization Type".

we can take mean or median for any numeric column which have NaN values. Median is mostly preferable.

we can impute/replace No or 0 if we find any missing value in FLAG_MOBIL,FLAG_EMAIL,FLAG_PHONE etc columns

Checking Data Types and conversion:

Check all the data types for the relevant columns and convert it accordingly.

Example: If there is any negative data in Age or any annual income amount column we can change it to positive one.

We can change any Date time Year column to datetime data type.

Checking Outliers:

We can check for any outliers for any numeric columns. Here we can find for total income that a person have huge income than others.

Identifying Categorical and Continuous column and Convert:

We should identify all the continuous and categorical columns and change it to accordingly.

Check the Imbalance of Data:

We must check the imbalance of data and find the ration for this.

Univariate and Bivariate Analysis:

Performing various univariate and Bivariate analysis on both Continuous and Categorical columns we conclude that some points:

1. Females are having more credit amount than male.
2. Repaying of loans is risky whose credit amount is higher than total income amount.
3. Here, we found that number of non-defaulters is more than defaulters.
4. We can conclude that Labourers are more likely to be failed to repay the loan where IT staffs are repaying more loans than others.
5. Females are having more difficulties to pay the loan than males.
6. We can say that customer without payment difficulties having AMT_ANNUITY in between 20000 to 30000.
7. People without payment difficulties take more credit for the annuity.
8. Females are more taking loans and have more amount of credit range.
9. We can conclude that range of the customers without payment is more as compare to the customers with payment.
10. There are more numbers of Approved loan than others.