**BANK LOAN CASE STUDY**

**DESCRIPTION**

A finance company faces multiple challenges: some credit challenges, some customers who don’t have a sufficient credit history take advantage of this and default on their loans. As a data analyst, you are tasked to use data analytics tools and skills to analyse patterns in the data and ensure that capable applicants are not rejected.

The aim of this project is to identify patterns that indicate if a customer will have difficulty paying their instalments. This information can be used to make decisions such as denying the loan, reducing the amount of loan or lending at a higher interest rate to risky applicants.

**TECH STACK**

Python- a programming language.

**DATA OVERVIEW**

Source of data

[**https://drive.google.com/drive/folders/1\_q9XTVBPBG4gIPUPL6SingvJyhj690kb**](https://drive.google.com/drive/folders/1_q9XTVBPBG4gIPUPL6SingvJyhj690kb)

The dataset is divided into 3 csv files:

**previous\_application.csv:** Contains information about previous loan applications.

**application\_data.csv:** Provides details about the current loan applications.

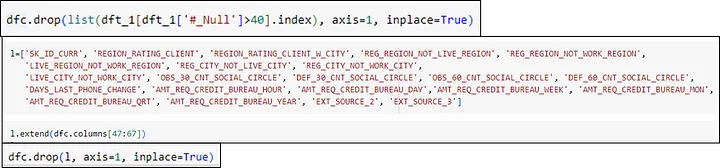
**columns\_description.csv:** Describes the columns present in the other datasets, explaining what each column represents.

**INSIGHTS**

**Identify Missing Data and Deal with it Appropriately:** As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

**Task:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

**Result:** Dropped all the columns where number of null values is greater than 40% and also all the unimportant columns.

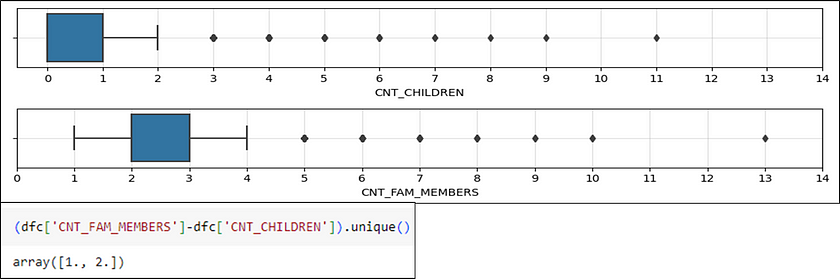
****

* For NULL values in OCCUPATION\_TYPE column, we replaced them with values which have the maximum count of values of NAME\_INCOME\_TYPE and NAME\_EDUCATION\_TYPE of the corresponding null vales. For the null values in OCCUPATION\_TYPE column which are still present after above step, were placed with Laborers, which is the most common occupation type.
* For the null values in the NAME\_TYPE\_SUITE column, the values were replaced with the maximum count of values of NAME\_INCOME\_TYPE and NAME\_FAMILY\_STATUS of the corresponding null values. For the column values of NAME\_INCOME\_TYPE and NAME\_FAMILY\_STATUS, the most common column value of NAME\_TYPE\_SUITE is Unaccompanied So replaced all the null values of NAME\_TYPE\_SUITE with Unaccompanied.
* For all the null values in AMT\_GOODS\_PRICE column, the values were replaced with the corresponding row value of AMT\_CREDIT column.
* Call null values in CNT\_FAM\_MEMBERS column. the values were replaced with the median value of in CNT\_FAM\_MEMBERS.
* There were some error values in GENDER column, the values were replaced with ‘F’ (Female), which is the most common GENDER.
* There were some error values in ORGANISATION\_TYPE column, were replaced them with No Work, as these people were pensioners and unemployed.
* For null values in AMP\_ANNUNITY column, the values were replaced with the median values of AMP\_ANNUNITY for all rows with corresponding values of AMT\_CREDIT and AMT\_INCOME\_TOTAL.

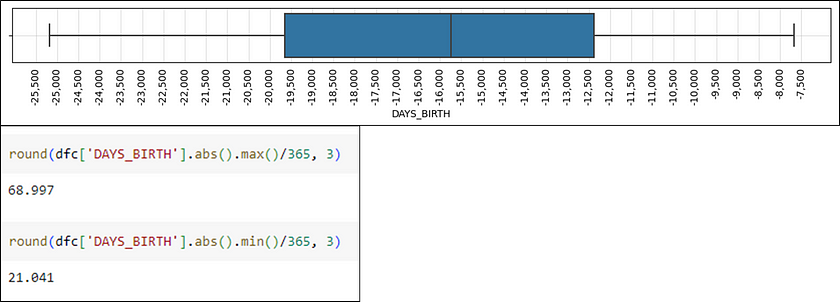
**Identify Outliers in the Dataset:** Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

**Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

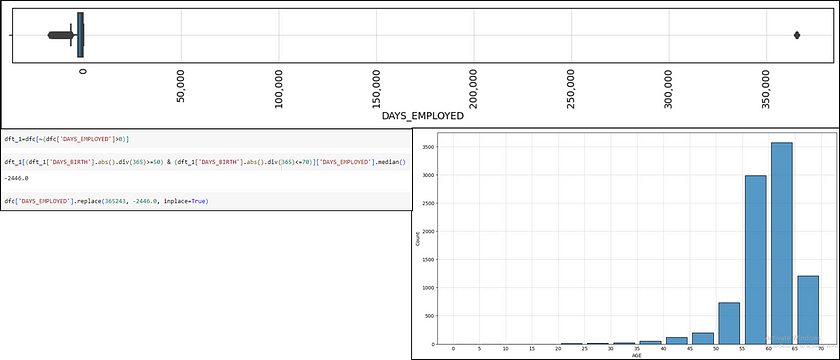
**Result:** For outliers in CNT\_CHILDREN and CNT\_FAM\_MEMBERS column, the number of parents for all row values were checked. Found no issues.



For outliers in DAYS\_BIRTH column, the values for maximum and minimum age were checked. No issues found.



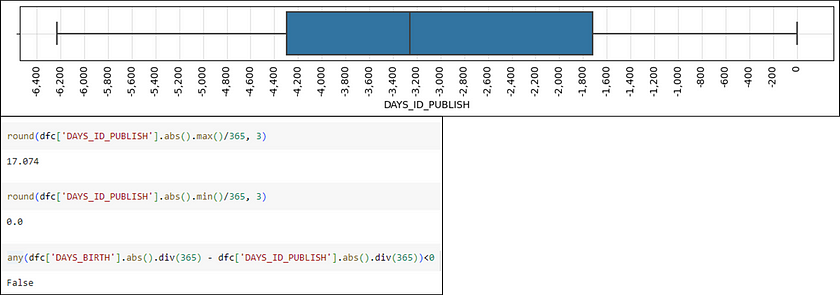
For outliers in DAYS\_EMPLOYED column, a box plot is added, a positive value as an outlier was found. The distribution of age for all those which have positive value of days employed were checked.



The outliers in DAYS\_REGISTRATION column was checked using a box plot considering the values less than -18000 as outliers. Calculated age in years, registration in years and difference between age and registration for all those rows. No issues were found.



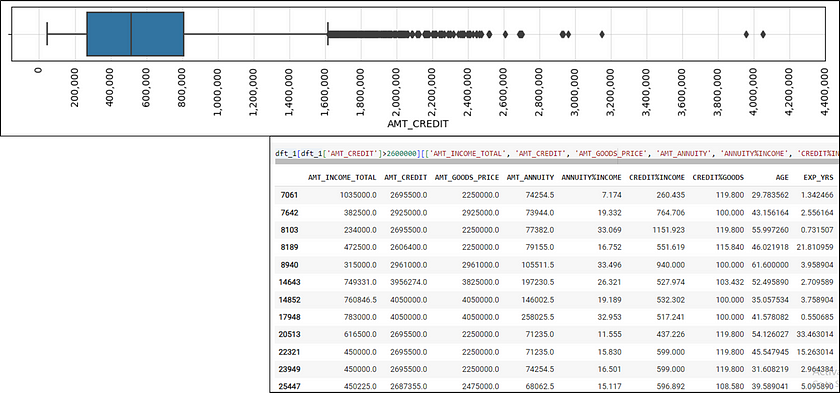
For outliers in DAYS\_ID\_PUBLISH column, the maximum and minimum age and difference between age and DAYS\_ID\_PUBLISH in years were checked. No issues found.



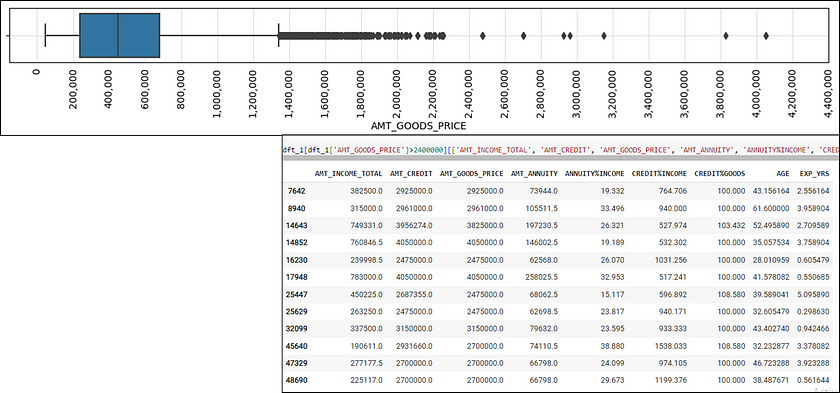
For outliers in AMT\_INCOME\_TOTAL column, a box plot was plotted considering values greater than 5000000 as outlier.



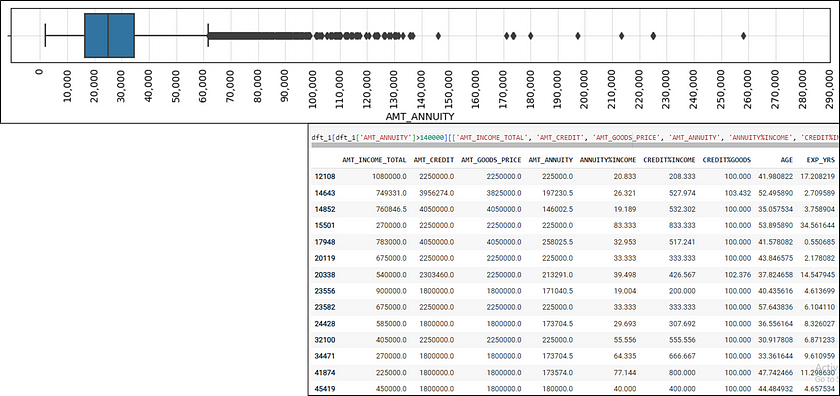
For outliers in AMT CREDIT column, the value is greater than 26000000 as outliers were checked for any possible issues for rows with amt credit greater than 2600000. No issues were found.



For outliers in AMT\_GOODS\_PRICE column. The value is greater than 2400000 as outliers were checked. For any possible issues for rows with AMT\_GOODS\_PRICE. greater than 2400000. No issues were found.



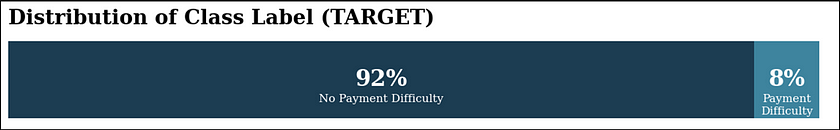
For outliers in AMT\_ANNUITY column, the value is greater than 140000 as outliers were checked for any possible issues for rows with AMT\_ANNUITY greater than 140000. No issues were found.



**Analyse Data Imbalance:** Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

**Task:**Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

**Result:**



The Dataset is highly imbalanced, skewed more towards Class Level 0. From above bar chart it is visible that around 92% of applicants didn’t have any difficulty in paying loans in instalments and around 8% class level one of applicants had difficulty in paying loans in instalments.

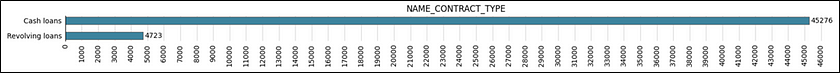
This data imbalance may give wrong predictions during modelling. So it needs to be taken care of by upscaling data of Class Label 1, or downscaling data of Class Label 0.

**Perform Univariate, Segmented Univariate, and Bivariate Analysis:**To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

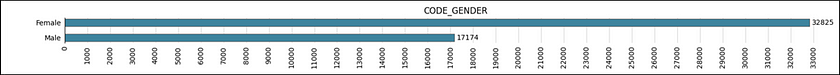
**Task:** Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

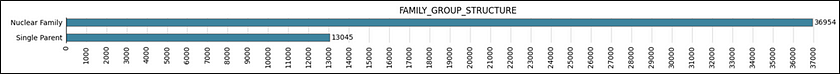
**Result:**

**UNIVARIATE ANALYSIS:**



Most of the loan applications are for CASH Loans and very less or for Resolving Loans.

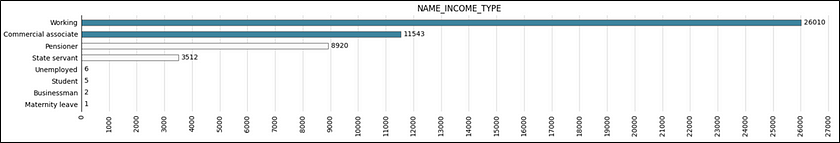
 Most of the loan applicants are females and there are less males.



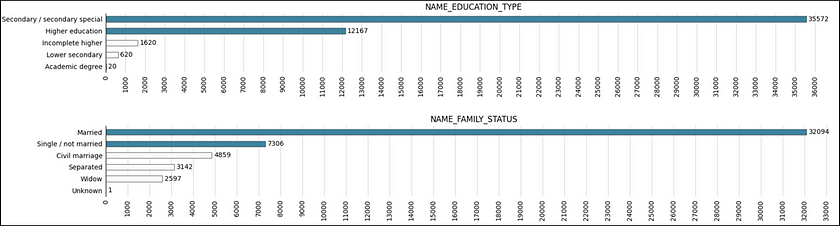
Most of the loan applicants have nuclear family, very less number of applicants are single parents.



Most of the applicants were not accompanied by anyone else followed by applicants who were accompanied by family members.

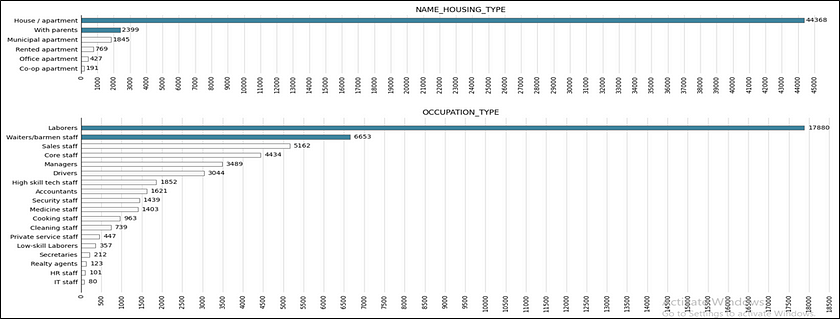


Most of the applicants had working income type followed by commercial associate income type.



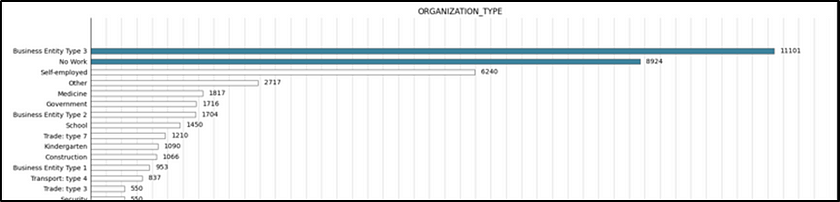
Most of the applicants had education. Up to secondary, followed by higher education

Most of the applicants were normally married, followed by single or not married.

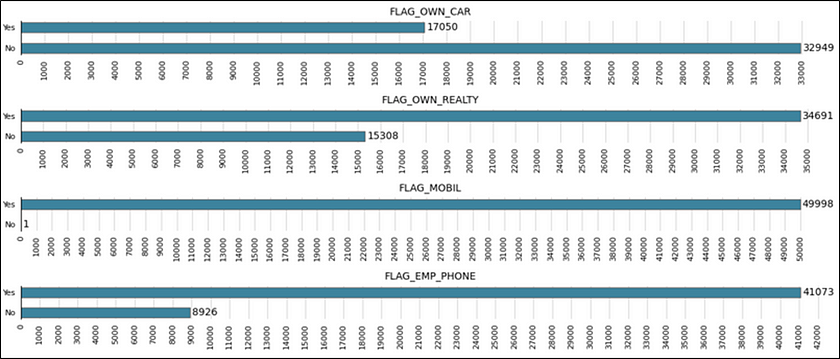


Most of the applicants had own house followed by applicants who were living with parents.

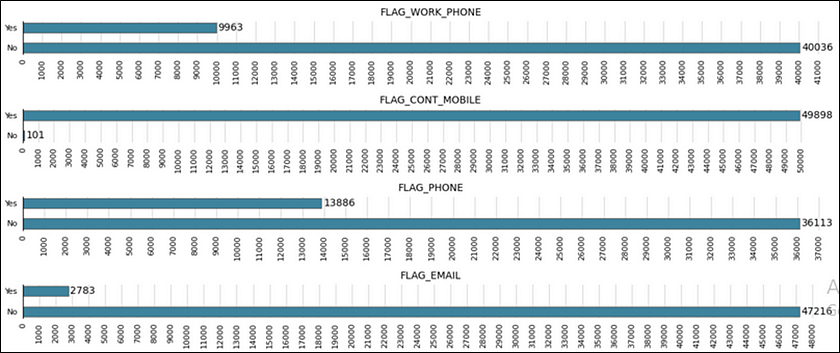
Most of the applicants were labourers, followed by waiters and barmen staff.



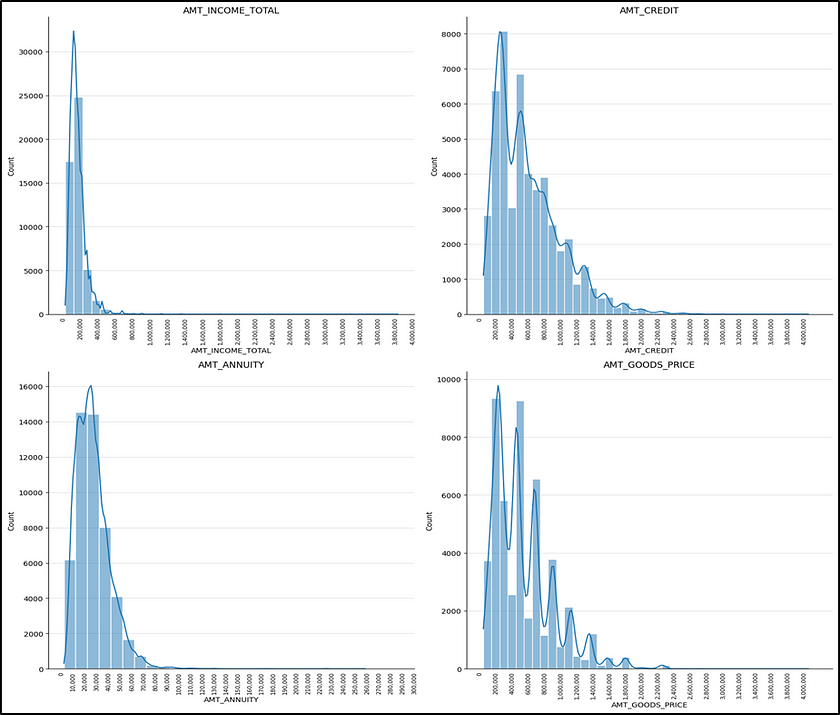
Most of the applicants worked in business entity, type iii, followed by applicants who didn’t work at that time. There were more pensioners and fewer unemployed.



Most of the applicants didn’t own a car, but had own property. Almost everyone had a mobile phone and most applicants had a phone from their employer.



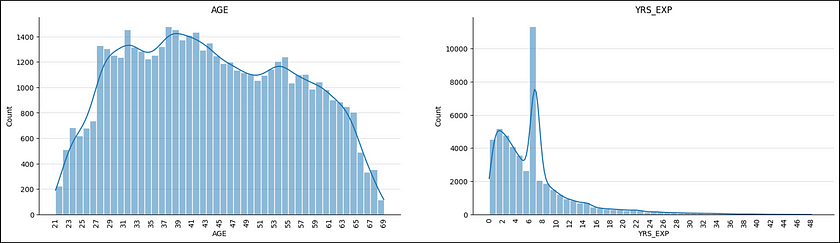
Most of the applicants didn’t have provided home phone, but most of the applicant’s phone provided were reachable. Also, most applicants didn’t had provided their email ids.



Most applicant’s income was less than Unit 400000 but a lot of their credit amount is more than unit 400000 showing that they have applied for loans of amount greater than their income.

Most applicant’s annuity amount is less than Unit 50000 which is close to 10% of income of most applicants.

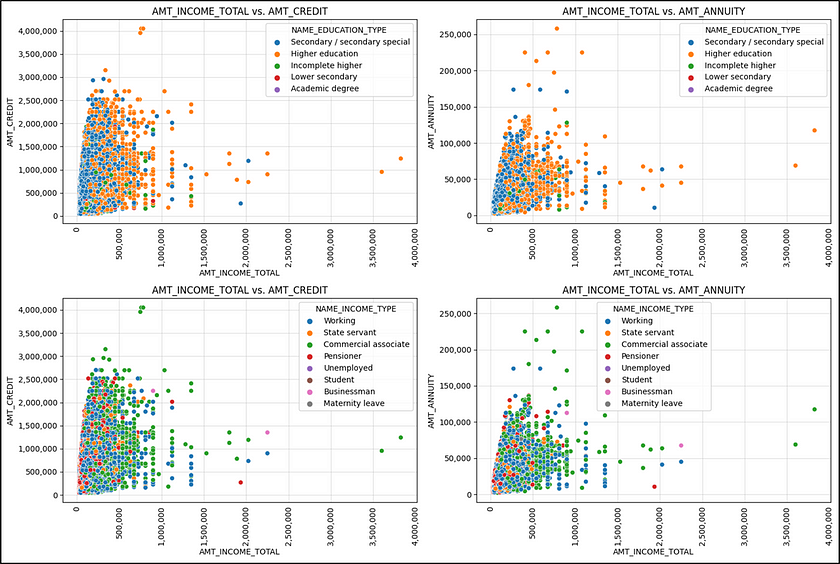
The distribution of AMT\_GOODS\_PRICE closely follows the distribution of AMT\_CREDIT.



We can observe that AGE column somewhat follows a normal distribution and most applicants are between age 27 and 65 i.e most were in the working age group.

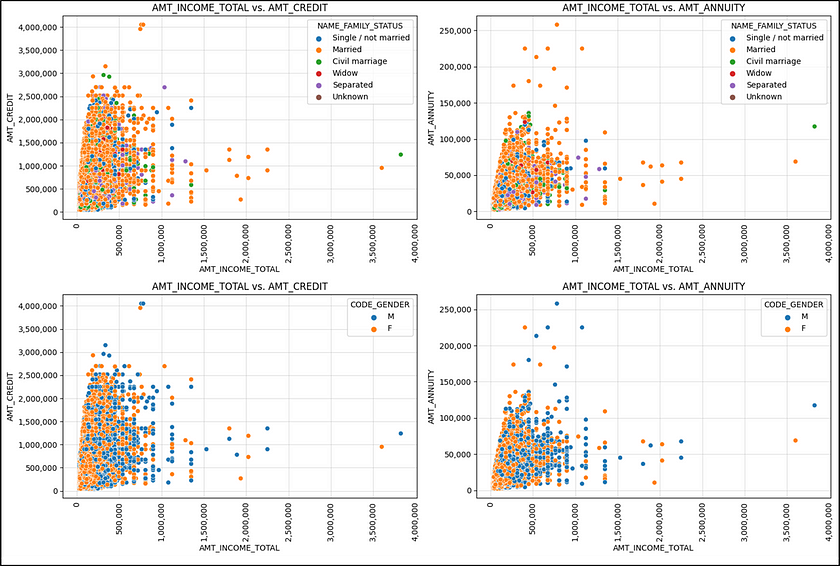
Also, most applicants had less than 8 years of work experience.

**SEGMENTED UNIVARIATE ANALYSIS:**



The first two plot shows relation between AMT\_INCOME\_TOTAL -AMT\_CREDIT and AMT\_INCOME\_TOTAL- AMT\_ANNUITY with segmented NAME\_EDUCATION\_TYPE. Also, blue points are more prominent as most applicants had secondary education.

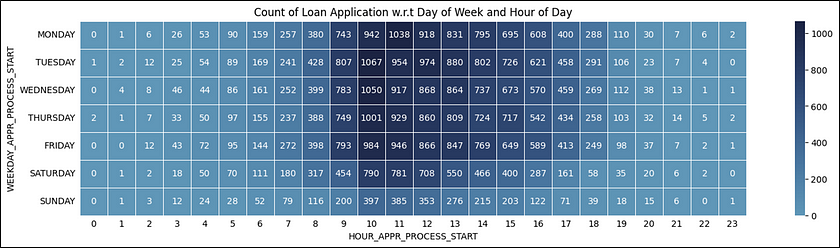
The last two plot shows relation between AMT\_INCOME\_TOTAL- AMT\_CREDIT and AMT\_INCOME\_TOTAL- AMT\_ANNUITY with segmented NAME\_INCOME\_TYPE. It is observable that applicants with higher income and higher loan credit as well as higher income and higher loan annuity has mostly commercial associate type income.



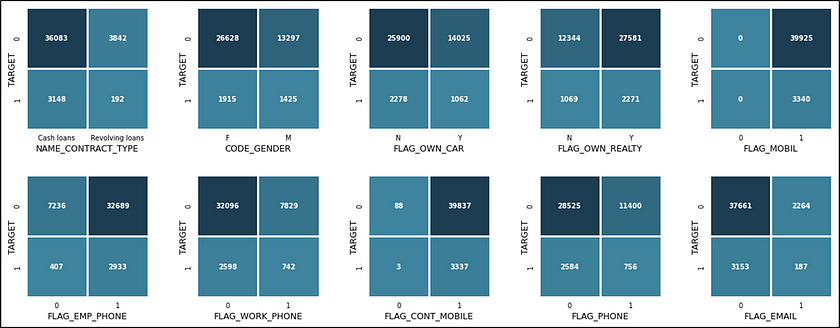
The first two plots shows relationship between AMT\_INCOME\_TOTAL -AMT\_CREDIT and AMT\_INCOME\_TOTAL- AMT\_ANNUITY with segmented NAME\_FAMILY\_STATUS. It is observable that the orange points are more prominent, as most of the applicants were married but it wasn’t sure of any relationship between income amount, credit amount and annuity amount with segmented family status.

The last two plots AMT\_INCOME\_TOTAL -AMT\_CREDIT and AMT\_INCOME\_TOTAL- AMT\_ANNUITY with segmented CODE\_GENDER. It is observable that blue points are more prominent towards right and upper side that shows male have mostly higher income, higher traded amount and higher and medium amount.

**BIVARIATE ANALYSIS:**

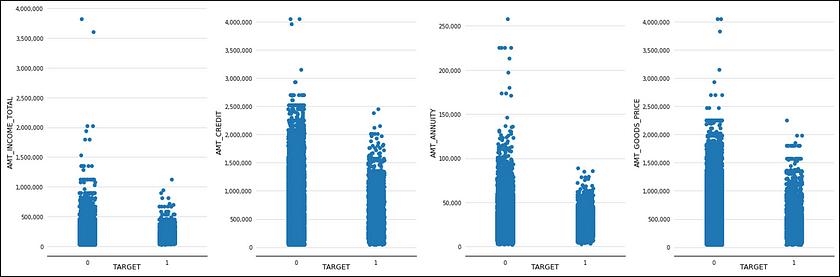


It is observed that most clints applied for loans during weekdays between 9:00 AM to 4:00 PM. But they were also very few clients who applied loans late at night as well.

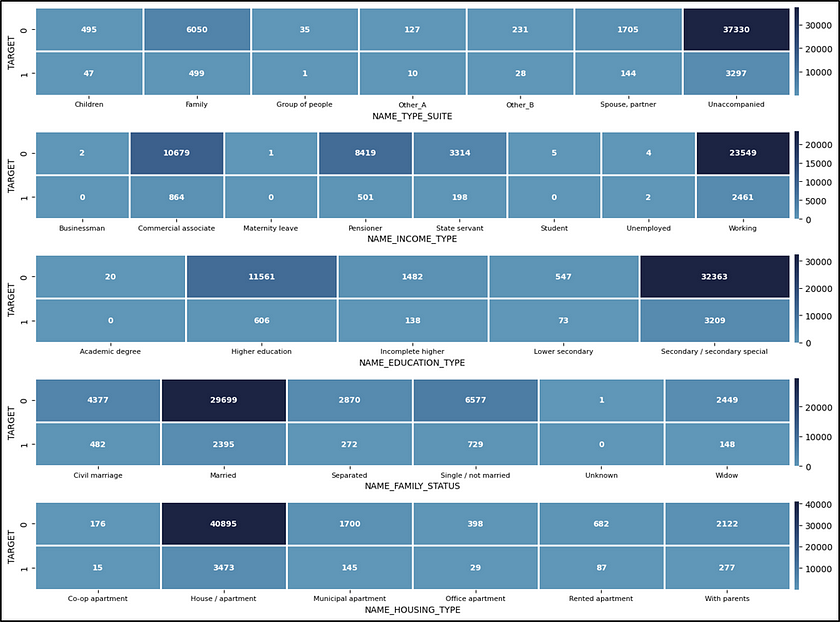


For columns with two unique values, it is observable that for both 0 and 1 of TARGET:

* Most contract types are Cash Loans.
* Most applicants are Females.
* Most applicants have phone number was reachable.
* Most applicants didn’t provide their e-mail id.
* Most applicants don’t possess own car, but have our own property.
* All applicants provided their mobile phone numbers.
* Most applicants provided their work phone number, but didn’t provide their home phone number.

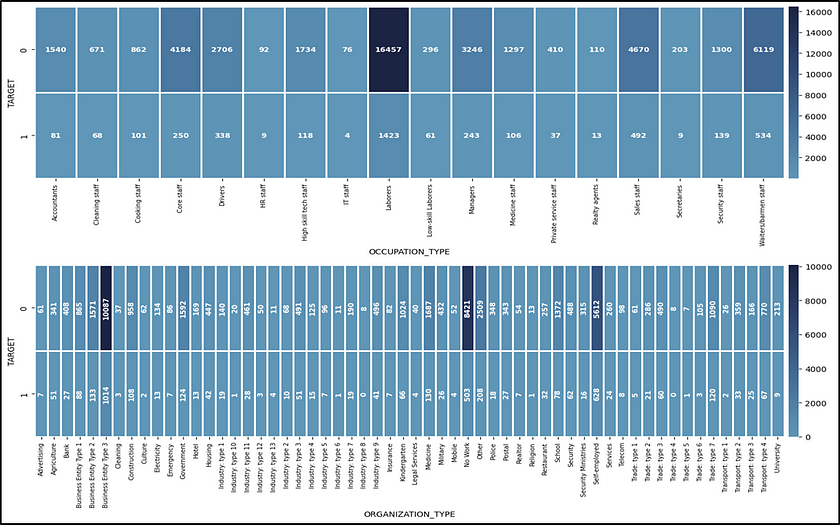


We can observe that for all the plot that is AMT\_INCOME\_TOTAL,AMT\_CREDIT AMT\_ANNUITY, AMT\_GOODS-PRICE, TARGET value of 0 has higher data point than TARGET value of 1, that is applicants with comparatively higher income, credit amount and annuity amount did not have any difficulty in paying their annuity amount in time as compared to applicants who face difficulty in paying their annuity amount in time.



For categorial columns. It is observable that for both 0 and 1 of TARGET:

* Most applicants were unaccompanied, followed by family.
* Most applicants income type were working, followed by commercial associate.
* Most applicants education was up to secondary followed by higher education.
* Most applicants were married followed by single.
* Most applicants were living their own house apartment followed by applicants with living with parents.



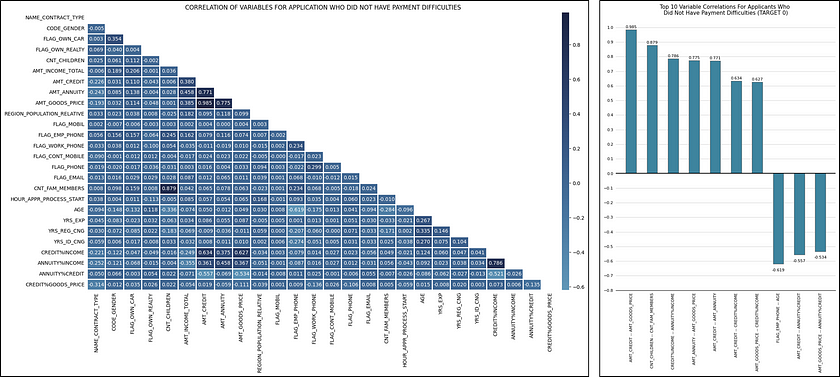
For categorical columns it is observable that:

* Most applicants were laborers followed by waiters and barmen staff for TARGET 0.
* Most applicants were laborers followed by waiters and barmen staff for TARGET 1.
* Most applicants were working in Business Entity Type 3 followed by applicants who had no work (pensioners and unemployed) for TARGET 0, whereas most applicants were working in Business Entity Type 3 followed by applicants who were self-employed for TARGET 1.

**Identify Top Correlations for Different Scenarios:**Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

**Task:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

**Result:**



Considering 0.5 as threshold for higher correlation, it is observable that:

* AMT\_CREDIT and AMT\_GOODS\_PRICE are highly and positively correlated, as the credit amount request is for the goods whose prices in AMT\_GOODS\_PRICE column.
* CNT\_FAM\_MEMBERS and CNT\_CHILDREN are highly and positively correlated as the applicants were either single parents or had a nuclear family.
* CREDIT%INCOME and UNITY%INCOME, AMT\_ANNUITY and AMT\_GOODS\_PRICE, AMT\_CREDIT and AMT\_ANNUITY are highly and positively correlated and have almost same correlation values.
* FLAG\_EMP\_PHONE and AGE are highly and negatively correlated possibly because older generation people are less used to phones.
* AMT\_CREDIT and CREDIT%INCOME, AMT\_GOODS\_PRICE and CREDIT%INCOME are highly and positively correlated as CREDIT%INCOME is a derived feature which is directly proportional to AMT\_CREDIT and AMT\_CREDIT and AMT\_GOODS\_PRICE are highly and positively correlated as in point 1.

**CONCLUSION**

Through this project, I was able to understand the importance of data analytics and bank loan analysis at it provides valuable insights which help in making data driven decisions.