# EDA ASSIGNMENT CASE STUDY

Bank Loan Default : Data set contains 3 CSV's

1. Application Data Set : Contains information about the customer's current loan application. This data is about whether the customer has difficulties paying the loan.
2. Previous Application Data set: Contains information about the customer's previous loan application. Whether it has been approved, cancelled, rejected or Unused Offer. Please note that Unused Offer implies that the Client has cancelled the loan application at various stages of the approval process, whereas for Cancelled - the customer has cancelled during the approval process either because he doesn't require the loan or he has received the loan at a high rate of interest due to credit risk.
3. Columns Description: Contains description of the various attributes of the data sets.

Problem Statement 1:

Retail/Consumer Finance Company's wants to identify the drivers/driving factors/driving variables which cause or may cause loan default i.e. variables which are strong indicators of loan default or variables which indicate credit risk. The company wants to use this knowledge to classify their portfolio risk assessment.

Problem Statement 2: What is expected from this assignment?

1. Approach to the Problem Statement in markdown comments.
2. Identify the missing data and use appropriate method to deal with it.
3. Identify if there are any outliers and mention why we think they are an outlier or they are not.
4. Find out if there is any data imbalance. What is the ratio of data imbalance?
5. How will we analyse the data in case of data imbalance?
6. Plot different graphs to explain the data imbalance and its effect on the analysis.
7. Which columns are you likely to delete and why?
8. Understand and identify the variables/columns/attributes which are most likely to influence a customer's repayment abilities. Conduct univariate and multivariate analysis on the various. Identify the top 10 correlations.
9. Identify the relevant attributes from both current application and previous application.
10. Merge the relevant columns ( common Column in both data sets is SK_ID_CURR, you will merge data set on this.

11. The target variable is names as Target in current application data set. You have to ensure that all your correlations contain this target variable. ( Remember the response_flag attribute from bank data set on term deposits).
12. Summarise your findings with the relevant visualization for each finding. 13. Mark your comments at each step, each assumption made, each visualization, each thought process in the python file or a one note or word. (Hyperlink if you can, so that you don't miss anything) This will help in preparing your ppt.

# CREDIT EDA CASE STUDY

## 1. Basic loading, Data overview and routine check

Two data sets are provided for this case study

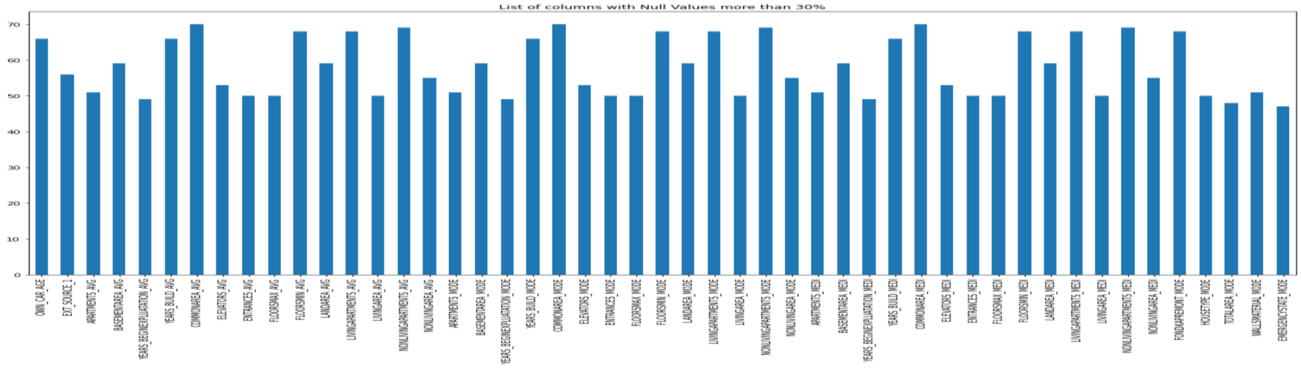
- Application data
- Previous application data

In order to perform quality checks, handling missing values and outliers and analysis, we have considered application data set and perform overall analysis.

## 2. Data Cleaning

**Finding and imputing missing values in current application**.

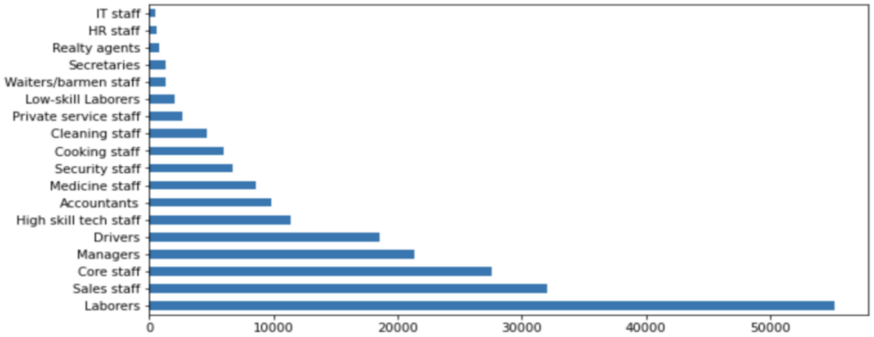
- Visualising columns with null values more than 35%



- Dropping columns with values more than 35% in this case(usually we drop columns that have more than 50% null values but I chose 35% as there these columns with null values more than 35% aren't important for our analysis.)
- Finding columns with missing values less than 35% and imputing it.

**Finding the data types and visualizing columns with missing values for imputing**

Checking the data type and visualizing "OCCUPATION_TYPE" column



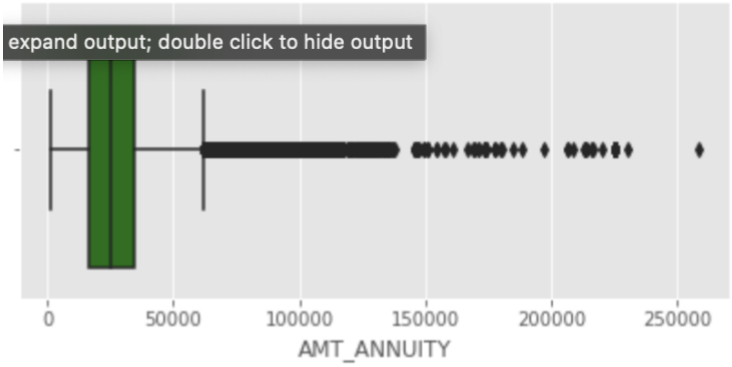
-
- This column is categorical unordered type so we use mode to find the most common value and impute in this column, from  graph also we can visualize that most common value is “Laborers” and mode is also the same so we impute this value.

**Following the same process of imputing mode for categorical columns with missing values as “NAME_TYPE_SUITE” column.**

**Finding missing values for Numerical columns.**

“AMT_ANNUITY”

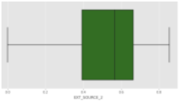
- In numerical columns we check for outliers in that data, if outliers present we use median else we use mean for imputing missing values.
- Visualising this columns using BOXPLOT  for outliers.
-

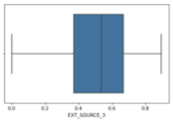


Observation:- We can see from above visualisation that this column 'AMT_ANNUITY' has outliers and  as we go above 99th percentile the values increased drastically So using Mean won't be a good choice, we can use median to impute value that is '24903.0'.

**We follow the same for other columns and imputed median and where we find no outliers we impute mean.**

- AMT_ANNUITY
- AMT_GOODS_PRICE-has outliers so median i.e. 450000.0
- NAME_TYPE_SUITE – is categorical so mode i.e.   ‘unaccompanied
- OCCUPATION_TYPE  - has outliers so median i.e. 2
- CNT_FAM_MEMBERS

- EXT_SOURCE_2  - No outliers so we use mean i.e. 0.5

- EXT_SOURCE_3 -  - No outliers so we use mean i.e. 0.5

- AMT_REQ_CREDIT_BUREAU_HOUR -we use mean as it has outliers

## Fixing error in data types of columns

- By looking through the data we find that columns starting with days have values with negative sign, Finding "DAYS" columns with negative sign and make it positive¶

- We also find that most of the values are in float and we can change that columns that can only contain integer that to integer

## Finding other errors in columns and replacing them with absolute values

- We remove "XNA" from "CODE_GENDER" and replace them with mode of that column which is "F".
- Looping through all the columns to overview errors in data types and fixing wrong values in them.
- There are more than 50k "XNA" in 'ORGANIZATION_TYPE' so we would simply replace with "NaN"

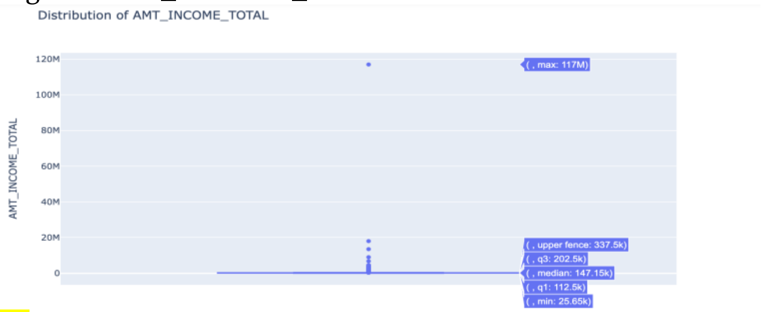
## Binning continuous variables for easy analysis

### Finding continuous variable

- **Binning "AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY","DAYS_BIRTH"**

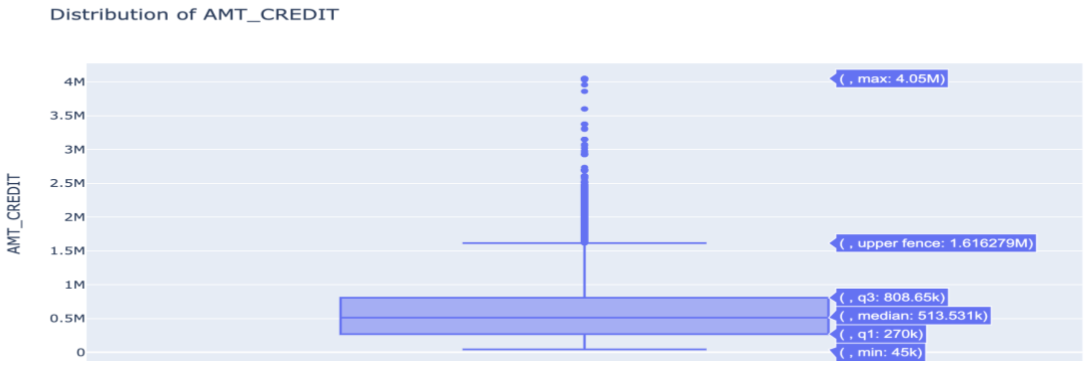
## Finding Outliers- Visualising and checking percentiles data distribution.

- Working on "AMT_INCOME_TOTAL" column to find outliers



**Observation**:- ***In this column we clearly see that there is an outlier in "AMT_INCOME_TOTAL" column, that is 117M***.

- Finding outliers in "AMT_CREDIT" column



Observation:- *We can see that "AMT_CREDIT" also contains outliers values beyond 1.6M-4.05M.*

- Defining a new function and looping multiple columns at once to find outliers.

We can infer from looping and visualisation distributions that multiple columns have outliers, "AMT_ANNUITY"- 255k , "AMT_GOODS_PRICE"- 4.05M,"DAYS_EMPLOYED"-365K, "DAYS_REGISTERATION"-24000 and more.

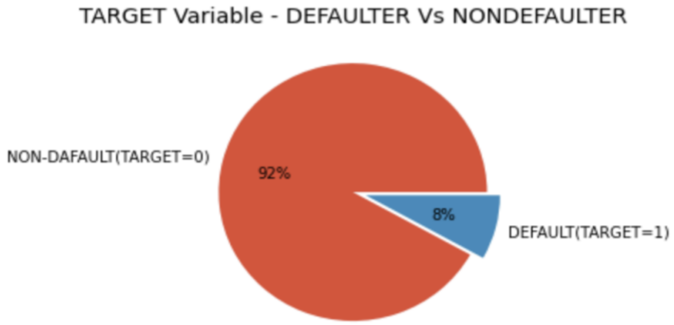
**Observation**:- ***From above visualisation of outliers we find "DAYS_EMPLOYED" is 365k days which equals to almost 1000 years and that's clearly a wrong value, and same inference for "DAYS_REGISTERATION"***

# 3. Analysis

## Checking Data imbalance

Checking imbalance in "TARGET" column

- Finding percentage of different values distribution in "TARGET" column in dataset and Visualising through pie chart.



Observation *:- From above Visualisation of "TARGET" column we infer that we have data imbalance and most of the values are leaned towards "NON-DAFAULTERS" and most of people who took loan did not default 92% of them, this can took make our analysis biased so we would divide our data-frame in in two parts, those who defaulted and those who didn't.*

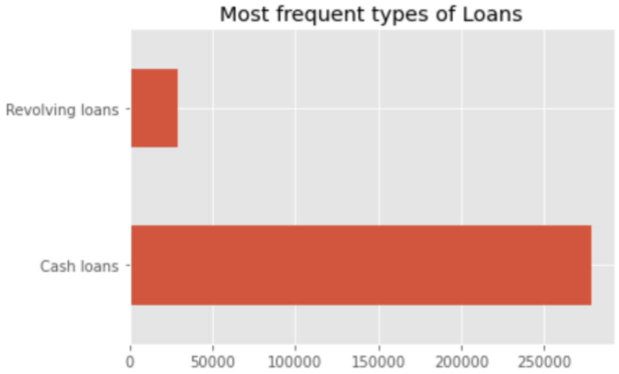
### Dividing data-frame in two Different dataset based on "TARGET" column

# 3. Analysis

## Univariate Analysis

**Univariate analysis of categorical unordered variable of both dataset with details of non-defaulters and defaulters.**

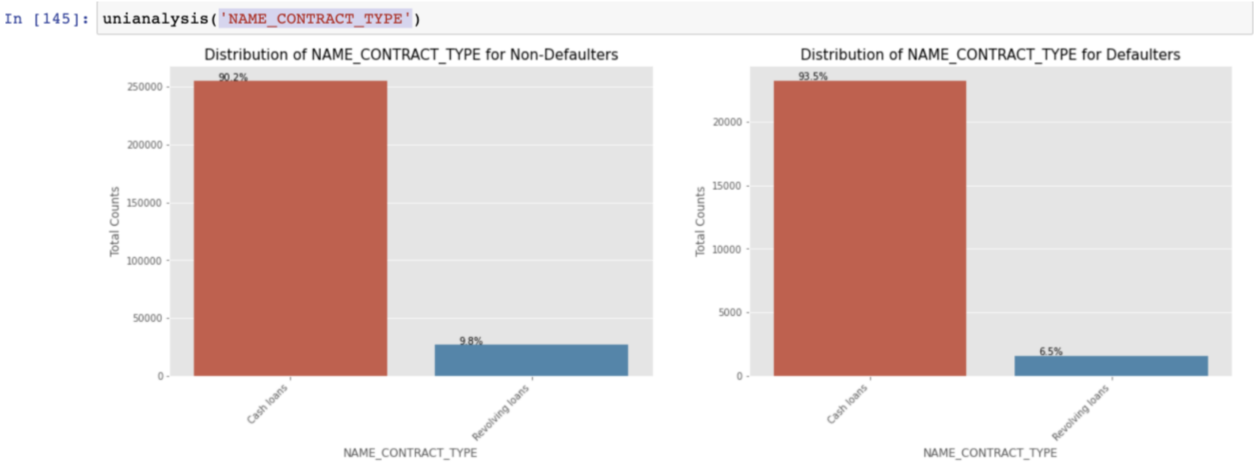
- Analysis of 'NAME_CONTRACT_TYPE' column



**Observation:-Most of the loans are of type "cash laons" as seen from the graph**

Defining a function for univariate analysis of two dataset(defaulters & non-defaulters)

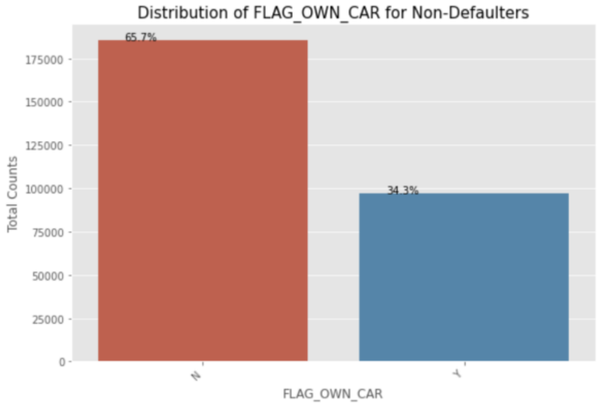
- Analysis of 'NAME_CONTRACT_TYPE'-Type of Loan for defaulters and non-defaulters

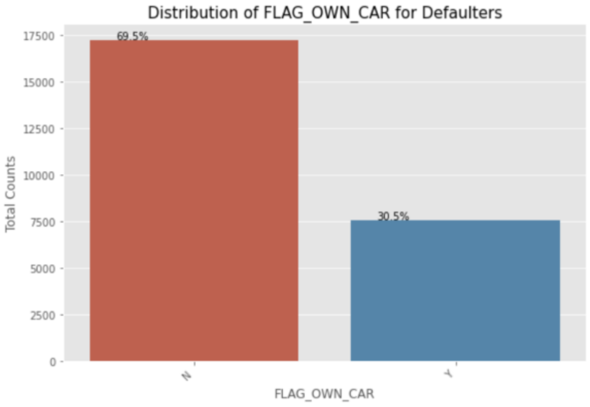


```
In [145]: unianalysis('NAME_CONTRACT_TYPE')
```

**Observation:- Most of the loans are "cash loans" for defaulters and non-defaulters, Defaulters took 3% more cash laons than non-defaulters.**

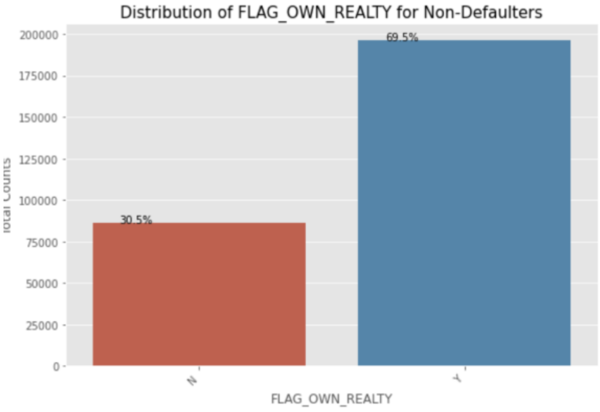
- Analysis of those who owns cars "FLAG_OWN_CAR"

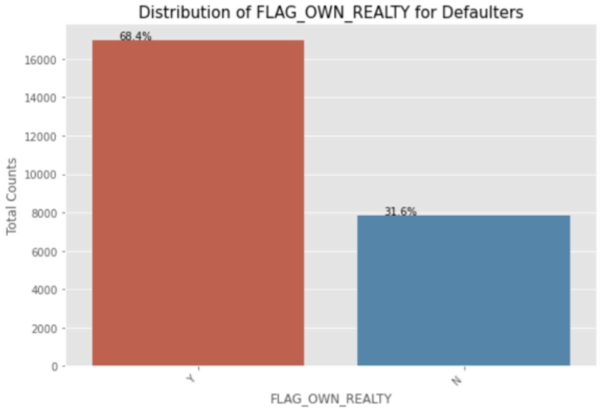




Observation:- Those who default will have slightly almost 4% more chances of not owning a car than those who don’t default . Those who own car are less likely to default

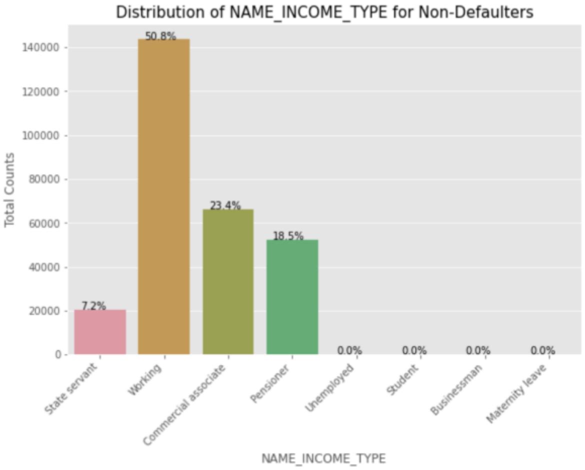
- Analysis of those who owns homes/flat "FLAG_OWN_REALTY"

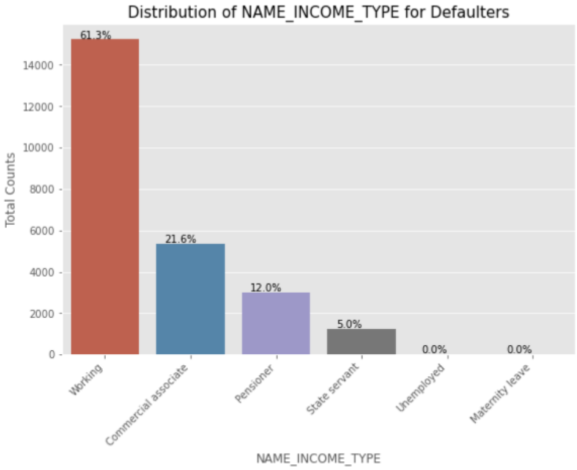




**Observation:- Most of the applicant owns a home.**

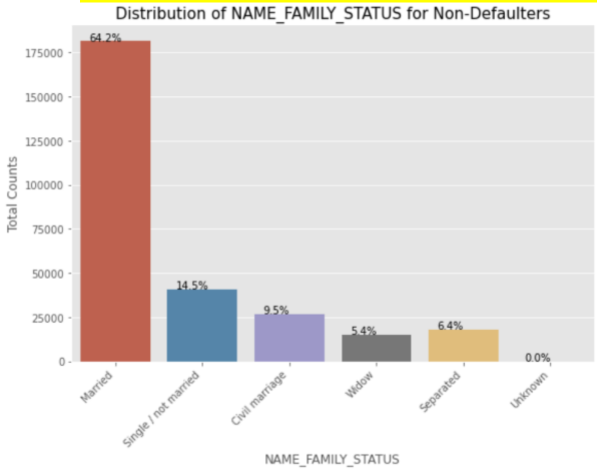
- Analysis "INCOME_TYPE"

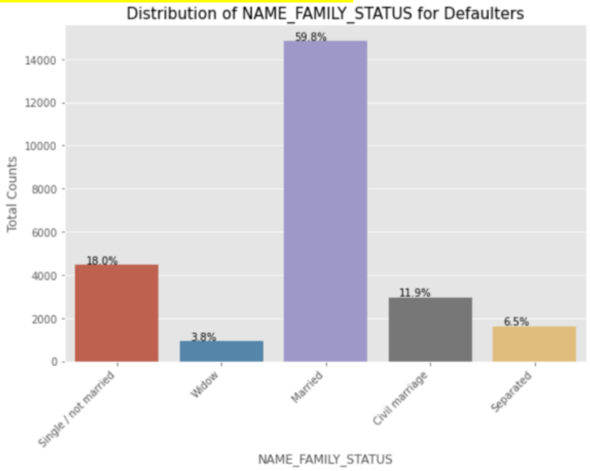




**Observation:- We can see that most of the loans is distributed among working class and there is increase of 11% in their loan payment difficulties, so we have to watch out working applicant as they are also the most. Also the most defaulters are working income type people.**
**-we can observe that pensioner chances of defaulting is also less 4.5% than non-defaulting.**
**-we can observe that State servant defaulters less, means they are less likely to default.**

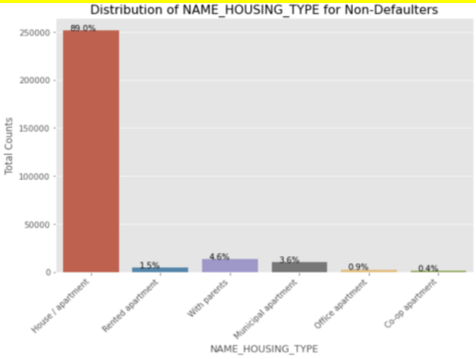
- **Analysis of family status of applicant -"NAME_FAMILY_STATUS**

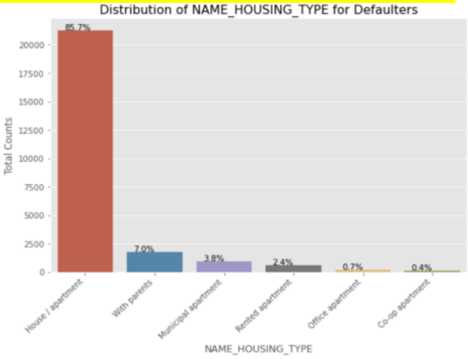




**Observation:- As from graph most of the applicant are married and slight less chances of around 5% less of non-defaulting loan and also they account for most Non- Defaulters**
**-we can observe that there is slight risk of single/not married and civil marriage applicant to default.**
**-widows also tend to be less defaulters**

- **Analysis of Housing situation of applicant "NAME_HOUSING_TYPE"**

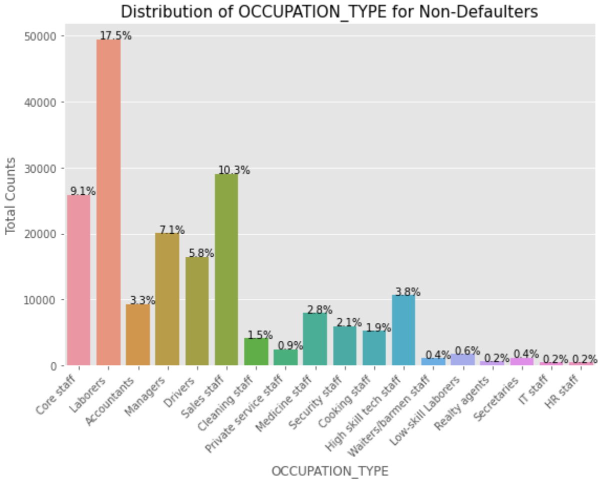


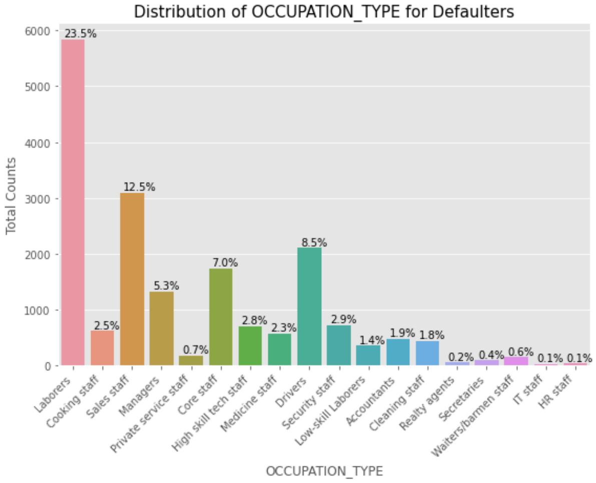


Observation:- we can infer that applicant with House/apartments apply for most loans.

-we can observe that applicant living with parents have slight more chances of defaulting a loan and so as applicant living in Rented apartments.

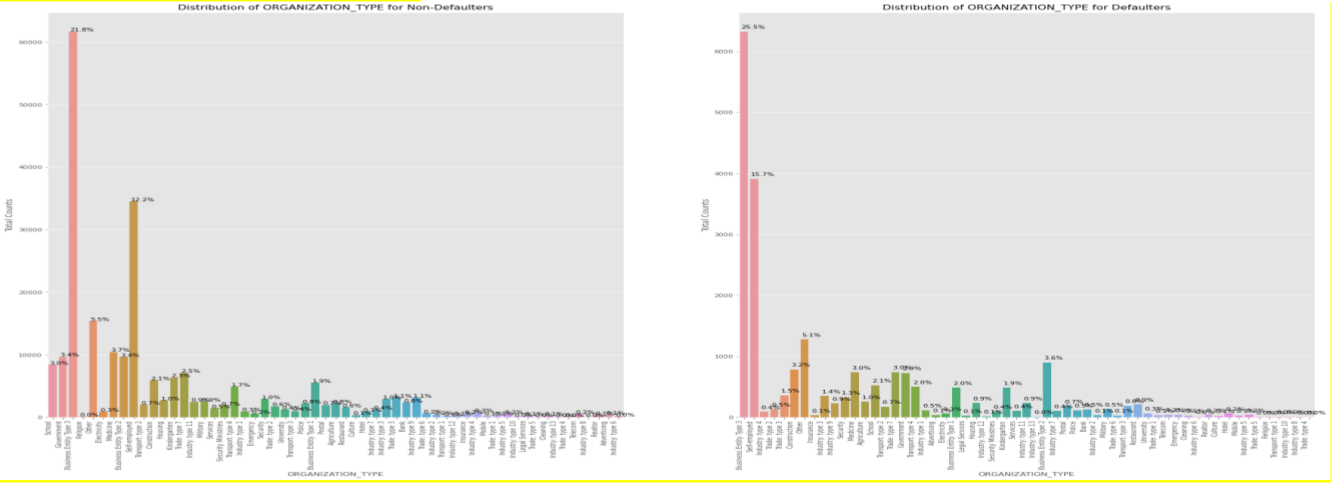
- **Analysis of type of occupation "OCCUPATION_TYPE"**





**Observation:- we can observe that most of the applicants are Laborers and also there is slight risk with them to default payment. While sales staff, drivers are more likely to default.**
**-we can infer that core staff, managers, accountant ,high skill tech staff are less likely to default their payment.**

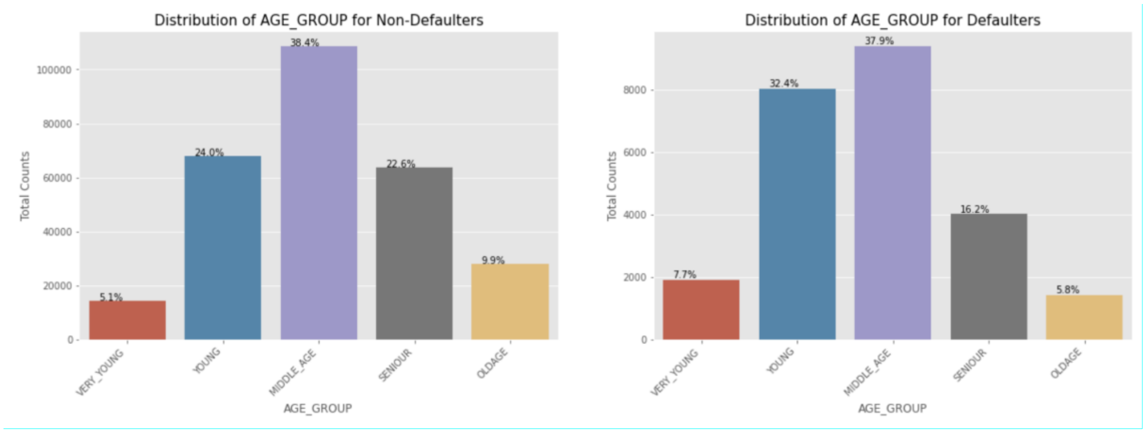
- **Analysis of Organization type**



Observation:- we can find Business entity type3 applied most for loan and also there is slight chances of them defaulting.¶

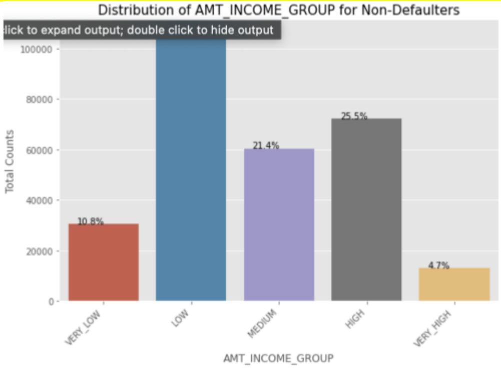
we can also observe that self-employed people have higher chance of defaulting loan.

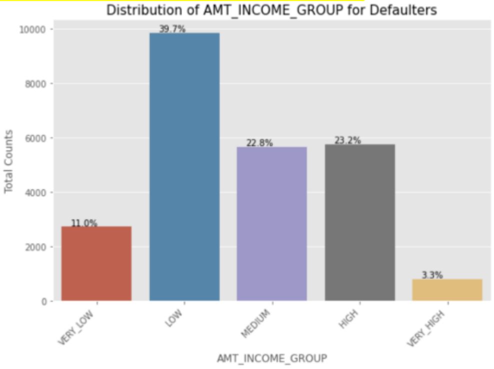
- **Univariate Analysis of Categorical Ordered variable -Analysis of "Age Group"**



**Observation:- we can analyze that most of the applicants comes in Middle age group the chances of defaulting payment or not defaulting are almost same.**
**-we can also observe that there are higher risk of giving loan to Young as their chances of defaulting payment is higher and so does for very Young .**
**-It can be concluded that applicant who are seniour and old has less chances of defaulting payment.**
**-It comes with less risk if loan is approved for seniours or elderly and more risk if approved for young or very young.**

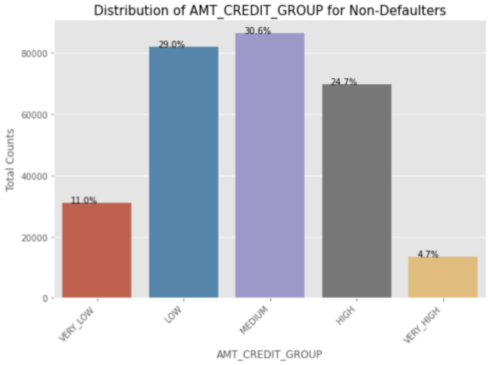
- **Analysis of "AMT_INCOME_GROUP" total income of applicant**

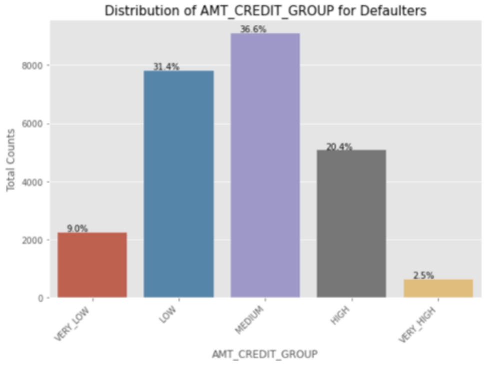




**Observation:- we can observe that most of the applicant for laon have low income and also their risk of defaulting is almost 2% less of non-defaulting.**
**-we can also observe that people with high and very high income have 1-2% slight less chances of defaulting a payment.**

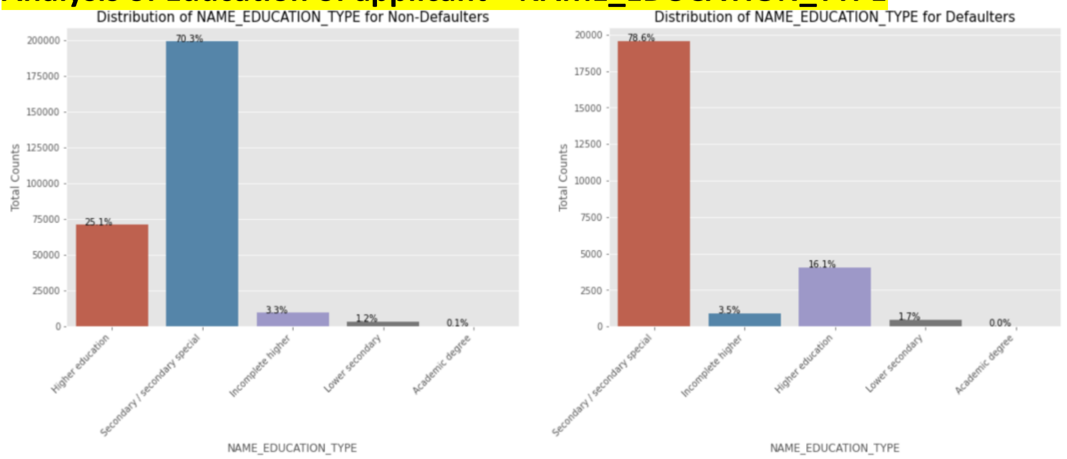
- **Analysis of credit amount of loan "AMT_CREDIT_GROUP"**





**Observation:- we can analyze that applicant who are in medium loan credit group tend to apply for the most loan and their chance of defaulting is highest in all loan credit groups around 36%.**
**-we can also observe that applicant in very low, high and very high loan credit group has less chance of defaulting compared with others.**

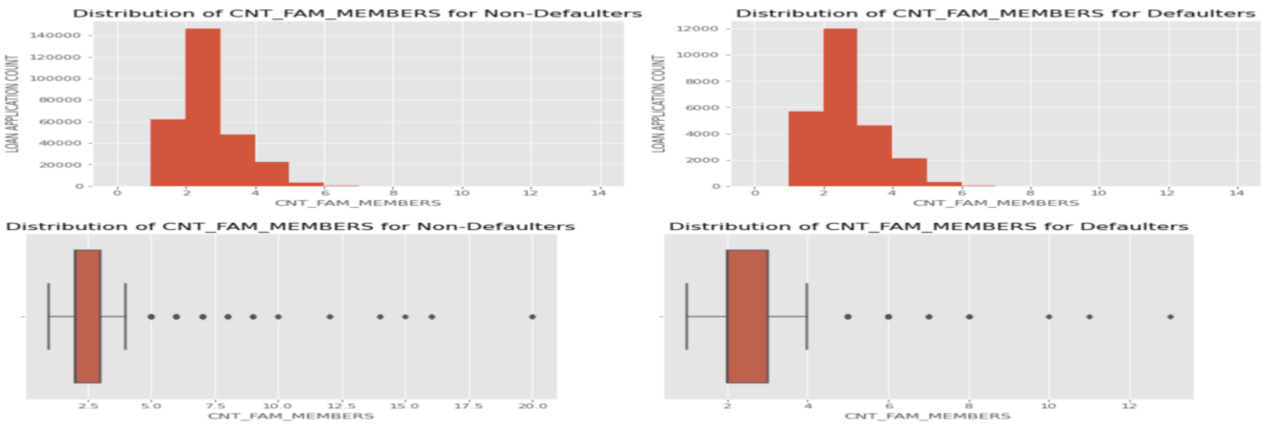
- **Analysis of Education of applicant " NAME_EDUCATION_TYPE**



Observation:- we can observe that Secondary/secondary special education group applies for loan the most and there’s also risk as they default 8% more than non-defaulters of same group.

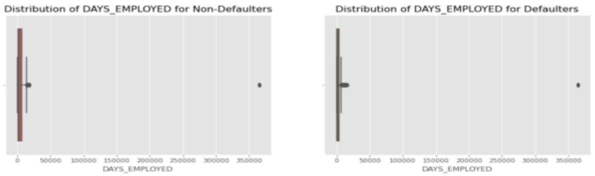
-we can also observe that applicant in Higher Education Group has good less chance of defaulting payment mean least risk in this group.

- **Univariate Analysis of Numerical Variable- Analysis family members of applicants**



**Observation:-Most of the loan applicant have family of three and this column has outliers.**

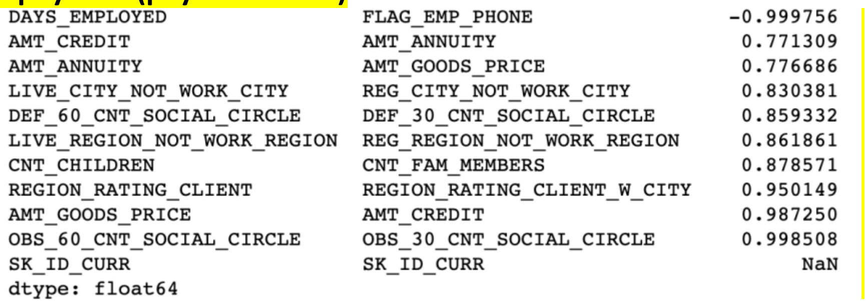
- **Analysis of "DAYS_EMPLOYED"**



**Observation:- There are clearly some outliers and wrong inputs in this column as days cant be close to 350k.**

# Finding top correlation in dataset of defaulters and non-defaulter

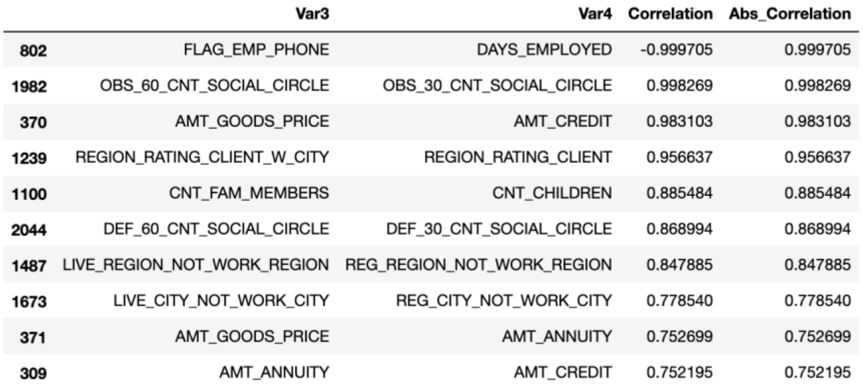
## Finding top correlation between variables of dataset of applicant those who didnt defaulted a payment(payed on time)



```
DAYS_EMPLOYED               FLAG_EMP_PHONE             -0.999756
AMT_CREDIT                  AMT_ANNUITY                 0.771309
AMT_ANNUITY                 AMT_GOODS_PRICE             0.776686
LIVE_CITY_NOT_WORK_CITY     REG_CITY_NOT_WORK_CITY      0.830381
DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE    0.859332
LIVE_REGION_NOT_WORK_REGION REG_REGION_NOT_WORK_REGION  0.861861
CNT_CHILDREN                CNT_FAM_MEMBERS             0.878571
REGION_RATING_CLIENT        REGION_RATING_CLIENT_W_CITY 0.950149
AMT_GOODS_PRICE             AMT_CREDIT                  0.987250
OBS_60_CNT_SOCIAL_CIRCLE    OBS_30_CNT_SOCIAL_CIRCLE    0.998508
SK_ID_CURR                  SK_ID_CURR                       NaN
dtype: float64
```

**Observation:-These are the top 10 correlation variables of applicants those who paid loan on time.**

## Finding top correlation between variables of dataset of applicant those who did defaulted a payment(Not payed on time)



| | Var3 | Var4 | Correlation | Abs_Correlation |
|---|---|---|---|---|
| 802 | FLAG_EMP_PHONE | DAYS_EMPLOYED | -0.999705 | 0.999705 |
| 1982 | OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998269 | 0.998269 |
| 370 | AMT_GOODS_PRICE | AMT_CREDIT | 0.983103 | 0.983103 |
| 1239 | REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.956637 | 0.956637 |
| 1100 | CNT_FAM_MEMBERS | CNT_CHILDREN | 0.885484 | 0.885484 |
| 2044 | DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.868994 | 0.868994 |
| 1487 | LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.847885 | 0.847885 |
| 1673 | LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.778540 | 0.778540 |
| 371 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.752699 | 0.752699 |
| 309 | AMT_ANNUITY | AMT_CREDIT | 0.752195 | 0.752195 |

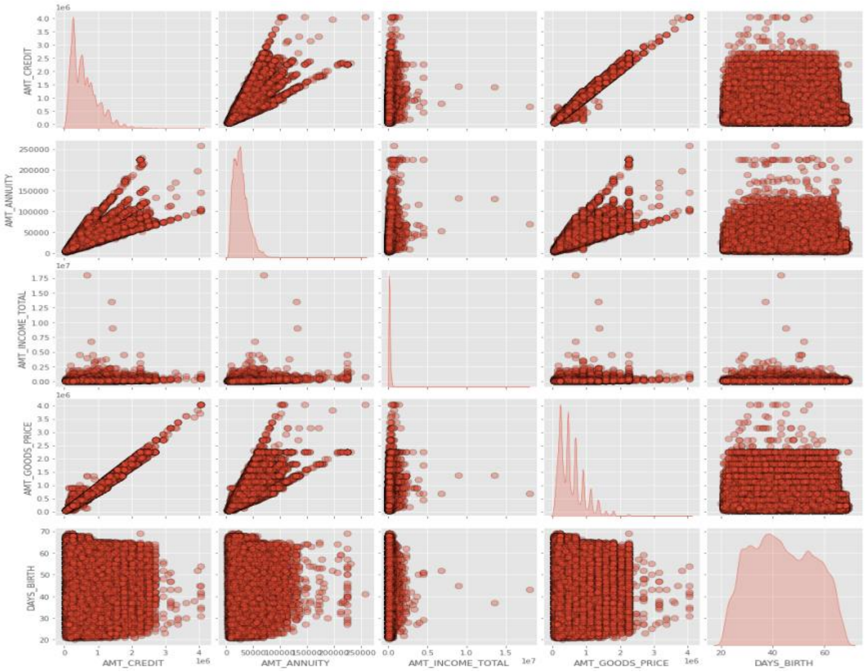
**Observation:- These are top 10 correlation between variables of those who didn't paid loan on time**

# Bivariate Analysis

## Bivariate Analysis of Numerical - Numerical variables

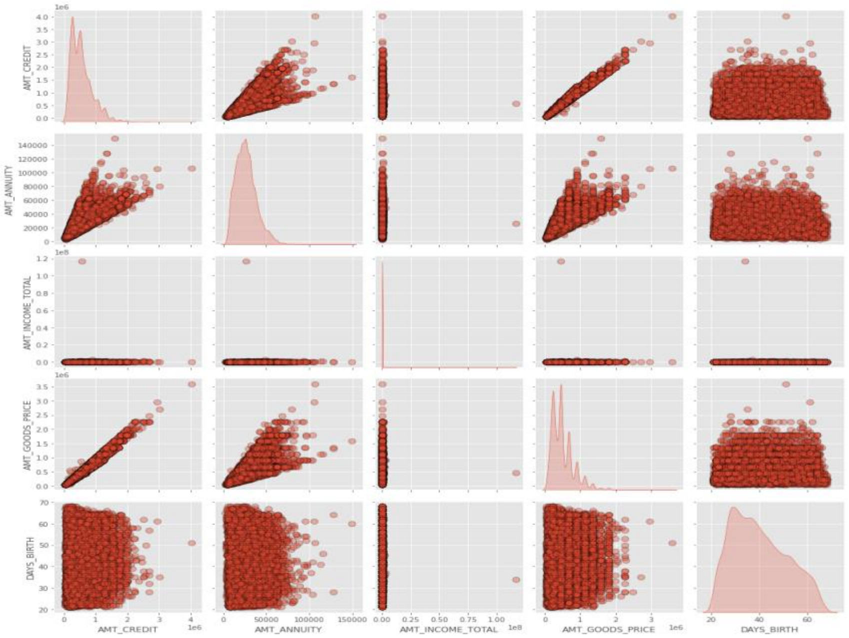
Pair plots and scatter plots of those who didn't defaulted a loan or not late on payment

- **Pair polts between different variables 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE', 'DAYS_BIRTH'.**



**Pair plots of those who did defaulted or late on their payment.**

- **Pair plot between variables 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE', 'DAYS_BIRTH'**



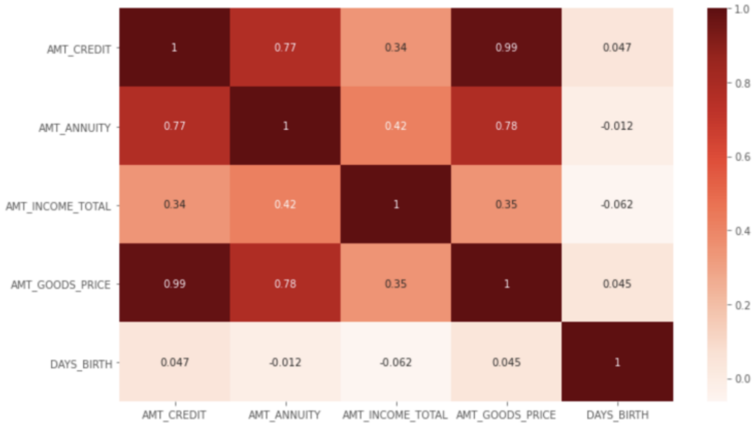
**Observation:-**

**-We observe from these visualisation, that credit amount of loan and goods price for which loan is being taken has good linear relation, one increases with other.**

**-we also observe that credit amount has linear relation with loan annuity, and loan annuity has good linear relation with goods price**

## Visualising all correlation with heatmap

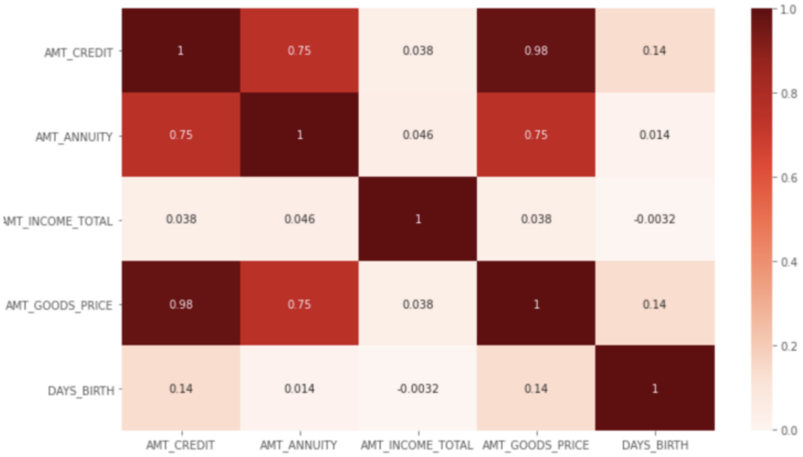
- Correlation heatmap of applicant who didn't default and paid on time



| | AMT_CREDIT | AMT_ANNUITY | AMT_INCOME_TOTAL | AMT_GOODS_PRICE | DAYS_BIRTH |
|---|---|---|---|---|---|
| AMT_CREDIT | 1 | 0.77 | 0.34 | 0.99 | 0.047 |
| AMT_ANNUITY | 0.77 | 1 | 0.42 | 0.78 | -0.012 |
| AMT_INCOME_TOTAL | 0.34 | 0.42 | 1 | 0.35 | -0.062 |
| AMT_GOODS_PRICE | 0.99 | 0.78 | 0.35 | 1 | 0.045 |
| DAYS_BIRTH | 0.047 | -0.012 | -0.062 | 0.045 | 1 |

**Observation:- Applicant who didn't defaulted-**
**We observe that there is some kind of linear relation between goods price and credit amount also we observe that, loan annuity and good price also follow a linear relation and also loan annuity and credit amount also follow a linear relation.**

- **Correlation heatmap of applicant who did default and did not paid on time¶**

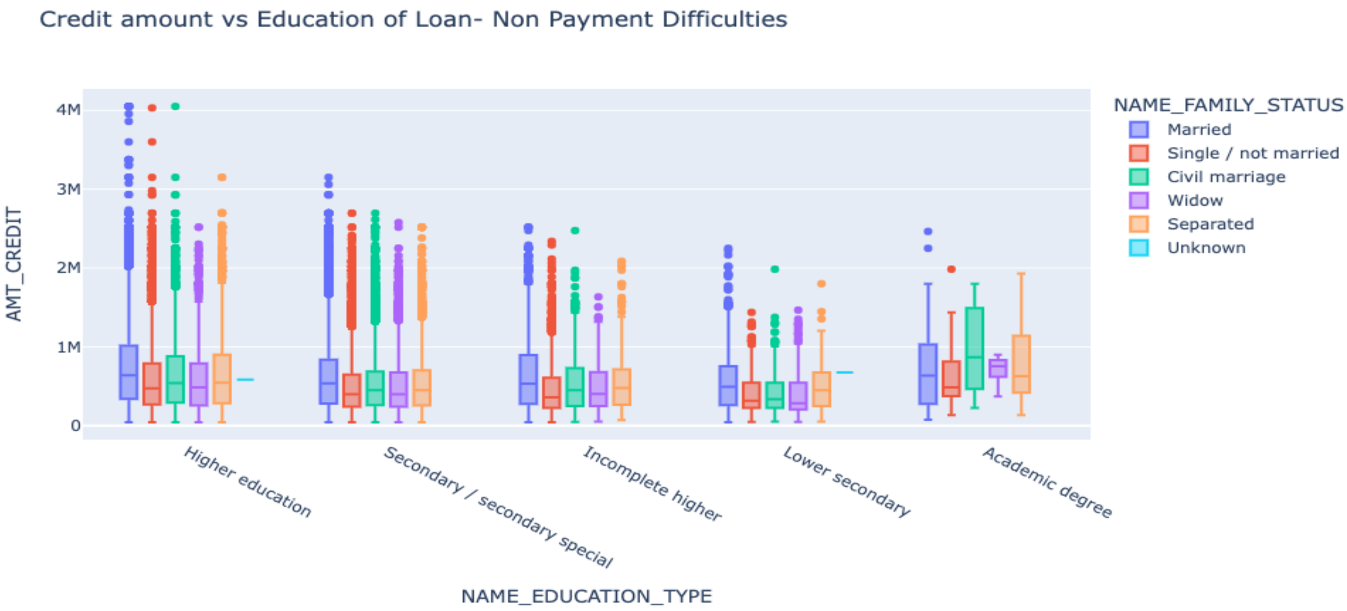


| | AMT_CREDIT | AMT_ANNUITY | AMT_INCOME_TOTAL | AMT_GOODS_PRICE | DAYS_BIRTH |
|---|---|---|---|---|---|
| AMT_CREDIT | 1 | 0.75 | 0.038 | 0.98 | 0.14 |
| AMT_ANNUITY | 0.75 | 1 | 0.046 | 0.75 | 0.014 |
| AMT_INCOME_TOTAL | 0.038 | 0.046 | 1 | 0.038 | -0.0032 |
| AMT_GOODS_PRICE | 0.98 | 0.75 | 0.038 | 1 | 0.14 |
| DAYS_BIRTH | 0.14 | 0.014 | -0.0032 | 0.14 | 1 |

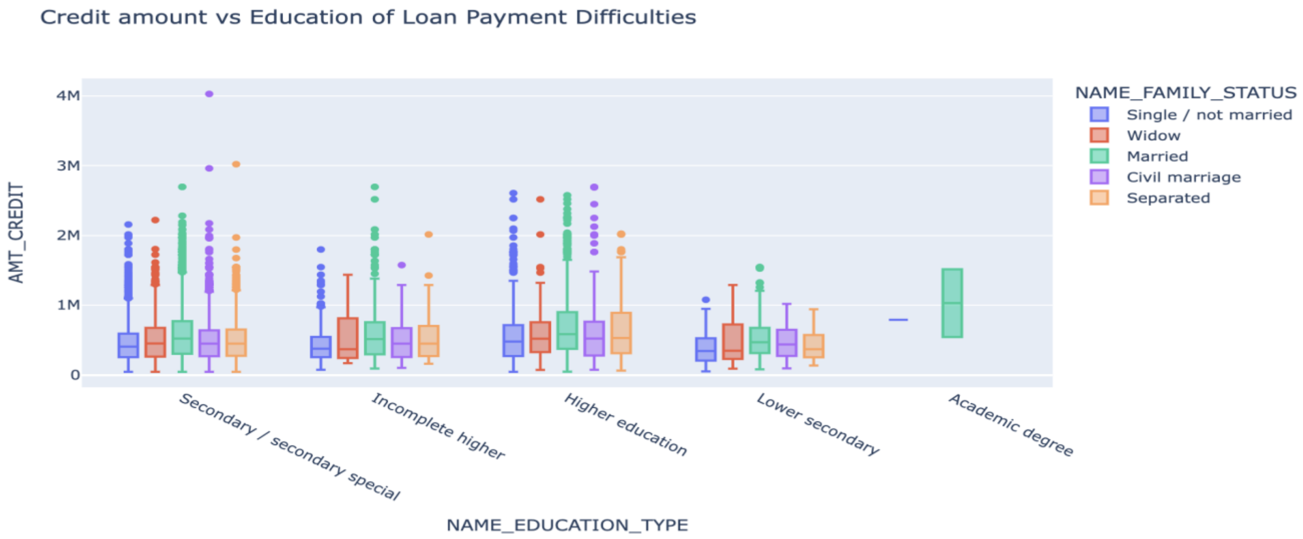
**Observation:- Applicants who defaulted- ('Difference between applicants who defaulted and those who paid on time')**
**-We observer that those who defaulted have weak linear relation between total income and price of goods(0.038) contrary to applicant who didnt defaulted and have good linear relation(0.35)**
**-we also observe that those who defaulted have weak linear relation between total income and laon credit(0.038),contrary to applicants who paid on time and have good linear relation between total income and laon credit(0.34).**
**-we observe that those who defaulted have weak linear relation between total income and laon annuity(0.046),contrary to applicants who paid on time having better relation between total income and laon annuity(0.42)**

# Bivariate Analysis/Multivariate of Numerical and Categorical variable

- Analysis of 'NAME_EDUCATION_TYPE' and "Family status" vs 'AMT_CREDIT' for Loan - Non Payment Difficulties



- 'NAME_EDUCATION_TYPE' and "Family status" vs 'AMT_CREDIT' for Loan -Payment Difficulties

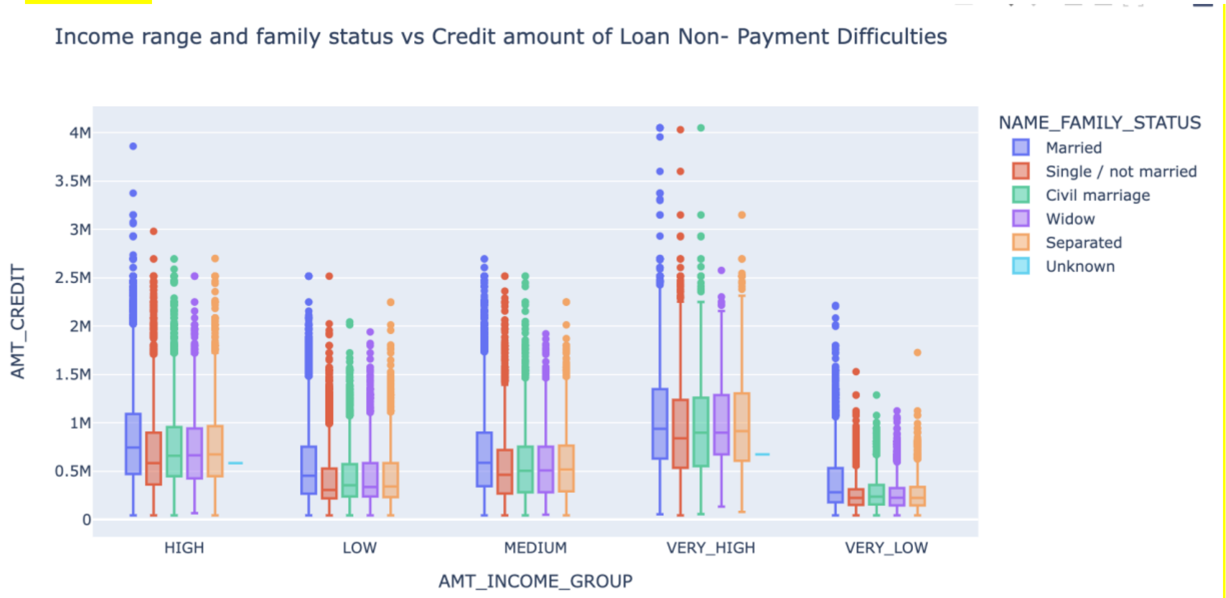


**Observation:-**

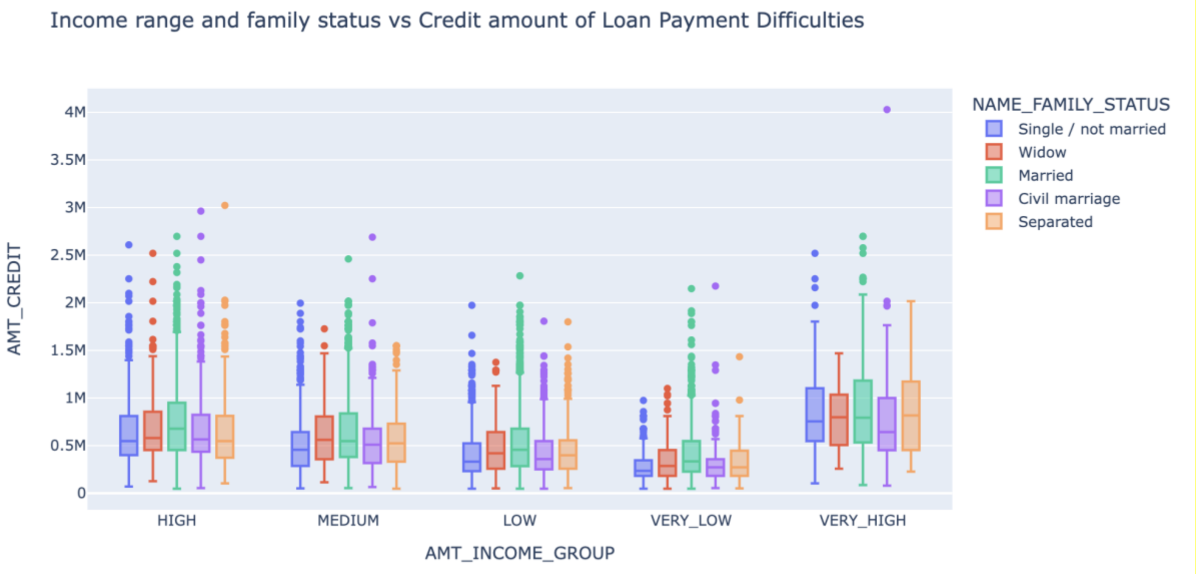
-There doesn't seem to be a lot difference but we do observe people with Academic degree and married have low loan credit and also little high chance of defaulting payment.

-People who are married with Higher education have good credit and are also less defaulting and also single with higher education.

- Analysis of 'AMT_INCOME_GROUP' and family status vs 'AMT_CREDIT' for Loan - Non Payment Difficulties



- AMT_INCOME_GROUP' and family status vs 'AMT_CREDIT' for Loan Payment Difficulties

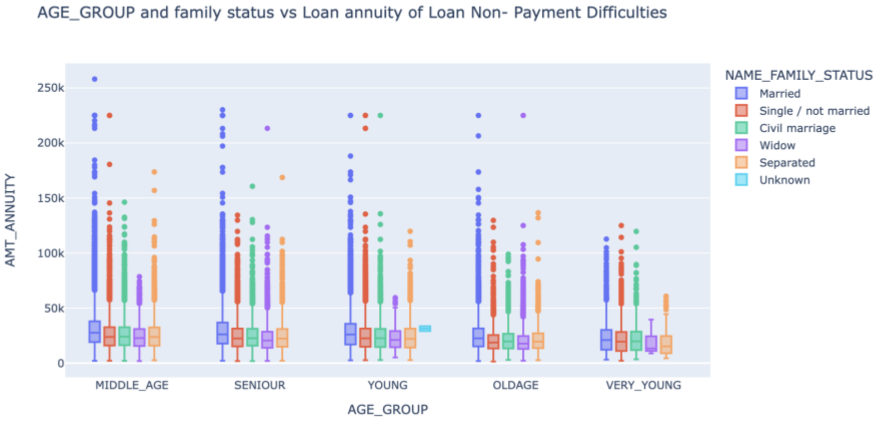


**Observation:-**

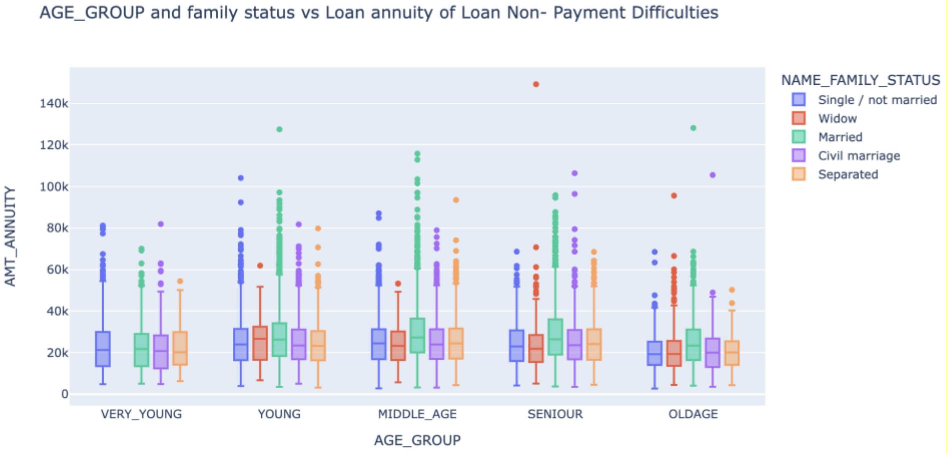
-we observe("Income range and family status vs Credit amount of Loan Non- Payment Difficulties") that applicant who are in very high income range and are married or civil married have good chance of paying loan on time and also applicant who are widow and separated in very high income group have less chance of defaulting. Applicant in high income group who are single also have less chance of defaulting.

-we observe("Income range and family status vs Credit amount of Loan Payment Difficulties") that applicant in high income group and married have slight more chance of defaulting also applicant in high income group and married also have slight more chance of defaulting-This could be the reason because people in high in come group applied for more loan.

- 'AGE_GROUP' and 'family status' vs 'AMT_ANNUITY' for Loan - Non Payment Difficulties



- 'AGE_GROUP' and family status vs 'AMT_ANNUITY' for Loan - Non Payment Difficulties



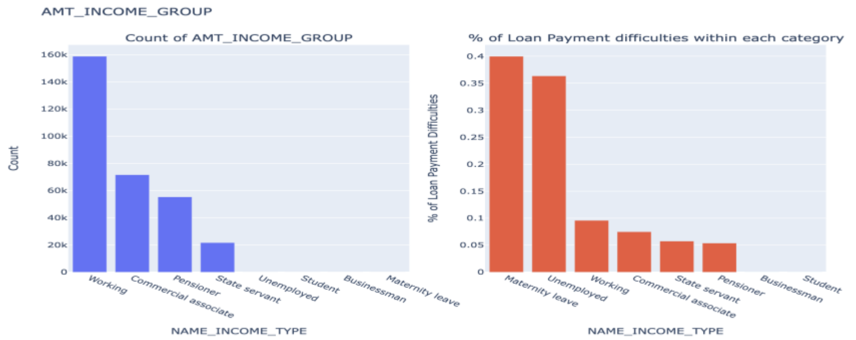
**Observation:-**

-We observe that those who didn't defaulted are applicant who are single and middle aged and have high loan annuity it could be also the reason as they applied for lots of loan.

-we observe that those who defaulted comes in married and middle age group and good loan annuity.

## Bivariate Analysis/Multivariate of Categorical and Categorical variable

- Analysis between 'NAME_INCOME_TYPE' and 'AMT_INCOME_GROUP

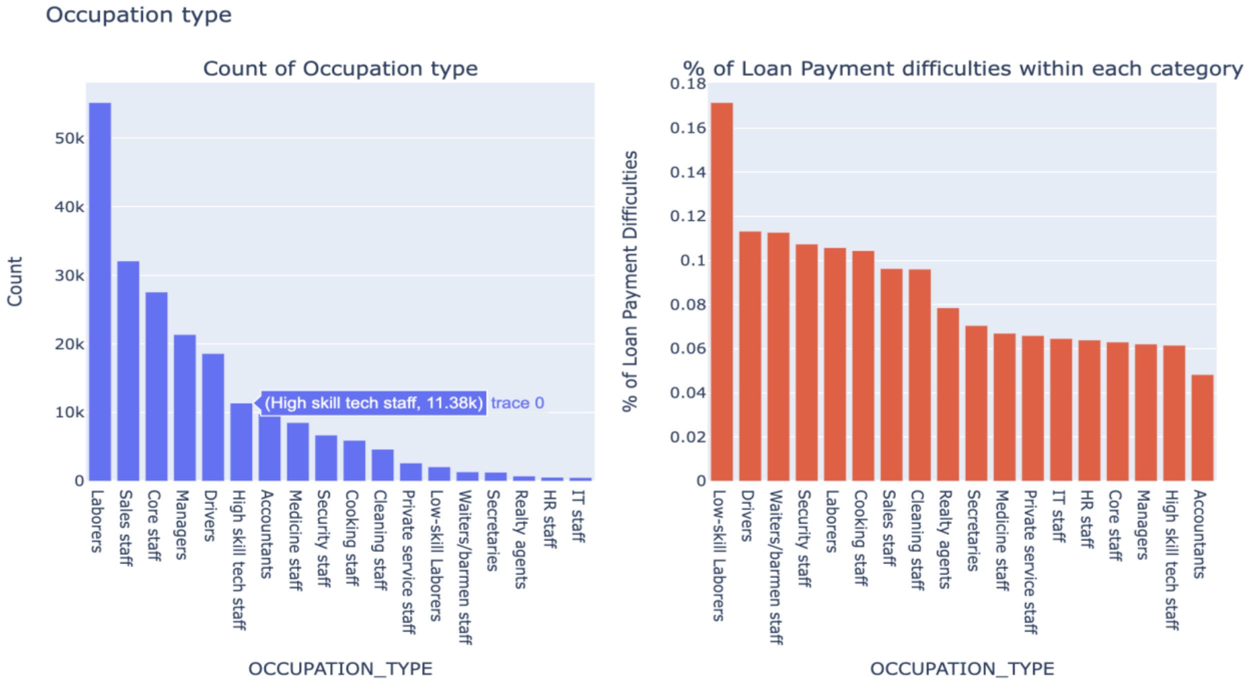


**Observation:-**

-we can observe that applicant who are working and in have high income groups have lower chance of defaulting a payment and applicant who are in in Maternity leave group have highest chance of loan payment difficulty.
-we can also observe that "Unemployed" applicant also have 2nd highest chance of defaulting payment.

- Analysis between 'OCCUPATION_TYPE' and 'Occupation type'

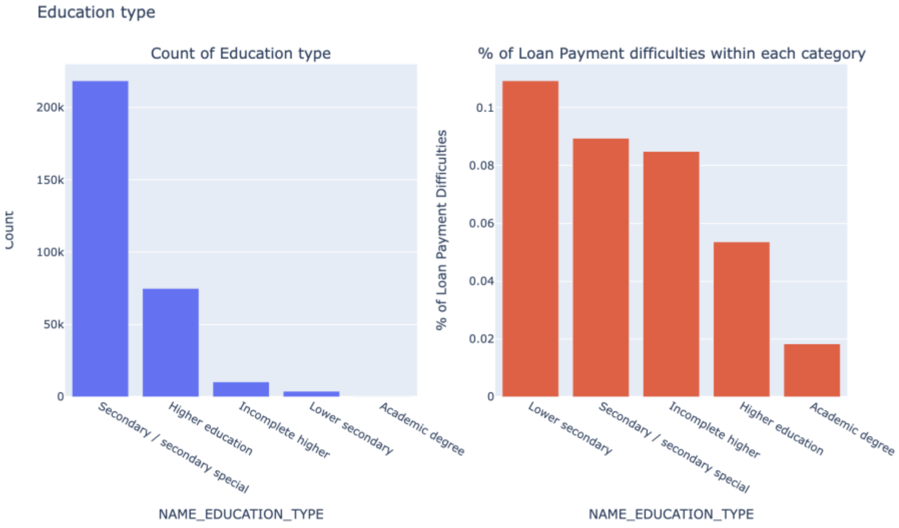
Distribution of Occupation Type and the category with maximum Loan-Payment Difficulties



**Observation:-**

we can say that clients with 'Lower skill Laborers' occupation type have maximum % of Loan-Payment Difficulties.
-Most of the applicants are 'laborers'

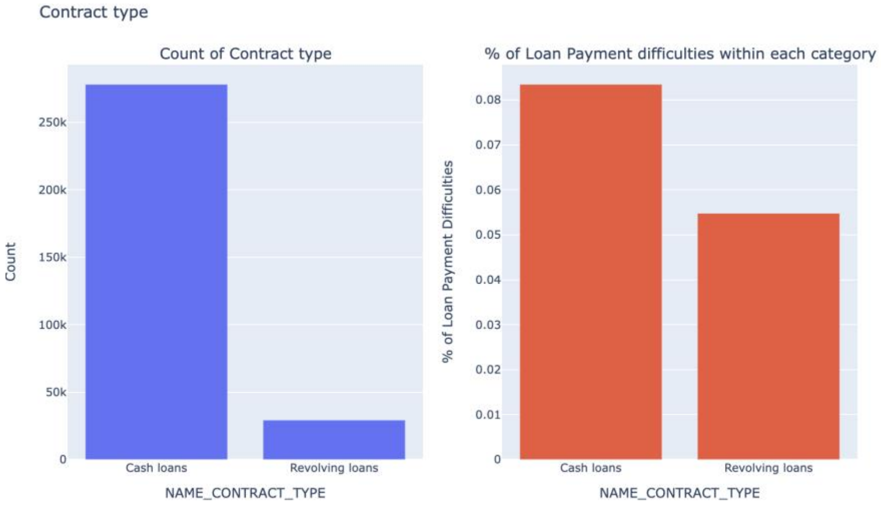
- Analysis between 'Education type' and 'Education type



**Observation:-**
-we can observe that most loans are applied by Secondary/secondary special and they have chances of defaulting payment while Lower secondary have the highest loan payment difficulties.

- Analysis between 'NAME_CONTRACT_TYPE','Contract type'

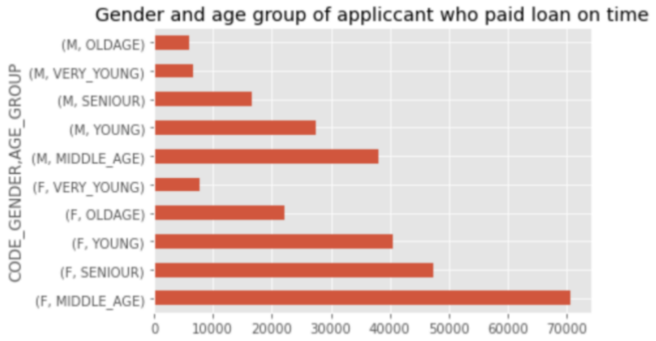
Distribution of Contract Type and the category with maximum Loan-Payment Difficulties



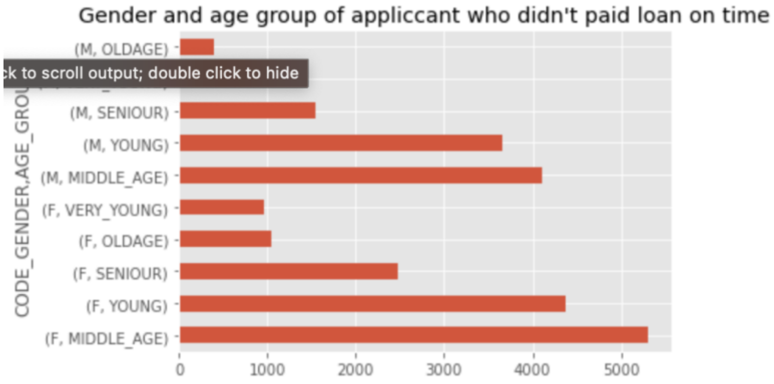
Observation:- Most of applied laon are of cash type hence they have also most defaulters.

- Analysis of age group and gender

-Gender and age group of applicant who paid loan on time



-Gender and age group of applicant who didnt paid loan on time



**Observation:-¶**

-we can observe that female middle age group took the most loan and also paid on time, and also female seniors took loans and paid on time and female seniors have high chance of paying on time and not defaulting.

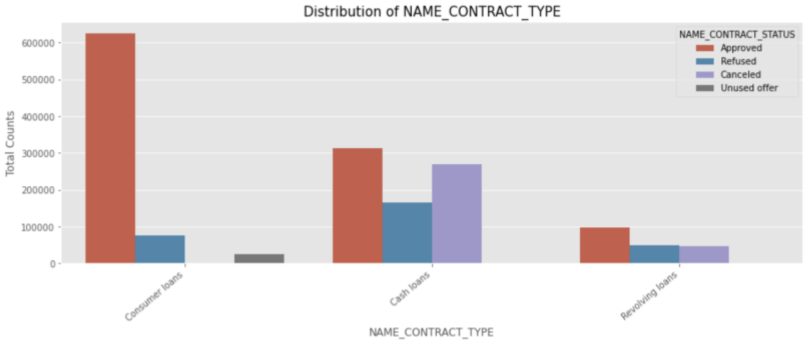
-we can also observe that young females and male middle aged applicant have higher chance of defaulting

## Data Analysis of Previous application

- Routine check of dataset
- Finding and imputing missing values- *Removing all the columns with more than 50% of null values*
- Finding errors-*Changing the negative values in the columns whose name start with DAYS to positive values.*

## Univariate Analysis

- Analysis of NAME_CONTRACT_TYPE

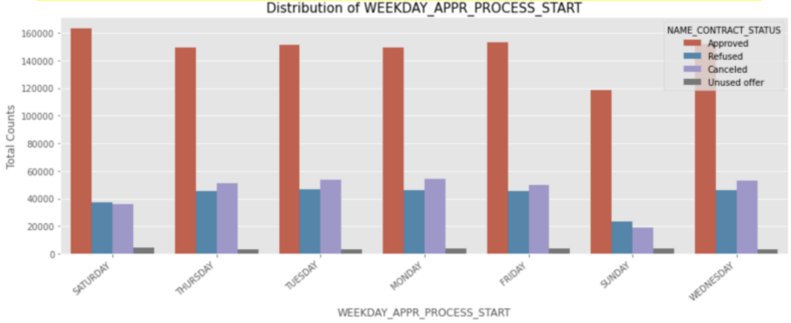


### Observation:-¶

-we can observe that most of application are for consumer loan and cash loan but most of the loans approved are consumer loans.

-we can observe that in cash loans some are approved and some are canceled and less are refused.

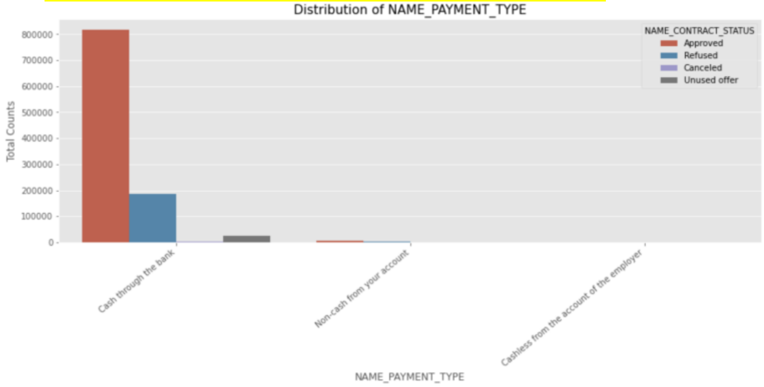
- Analysis of WEEKDAY_APPR_PROCESS_START



### Observation:-

-we can observe that loan approval process started on Sunday have lesser chance to get approved compared with others.

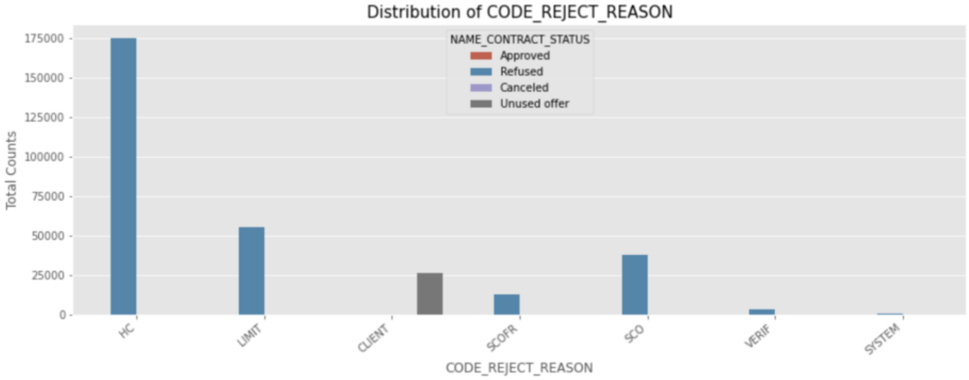
- Analysis of NAME_PAYMENT_TYPE



**Observation:-**

-we can see that payment cash through the bank are done the most and also they have highest chance of getting approved.

- Analysis of CODE_REJECT_REASON(previous application rejection)

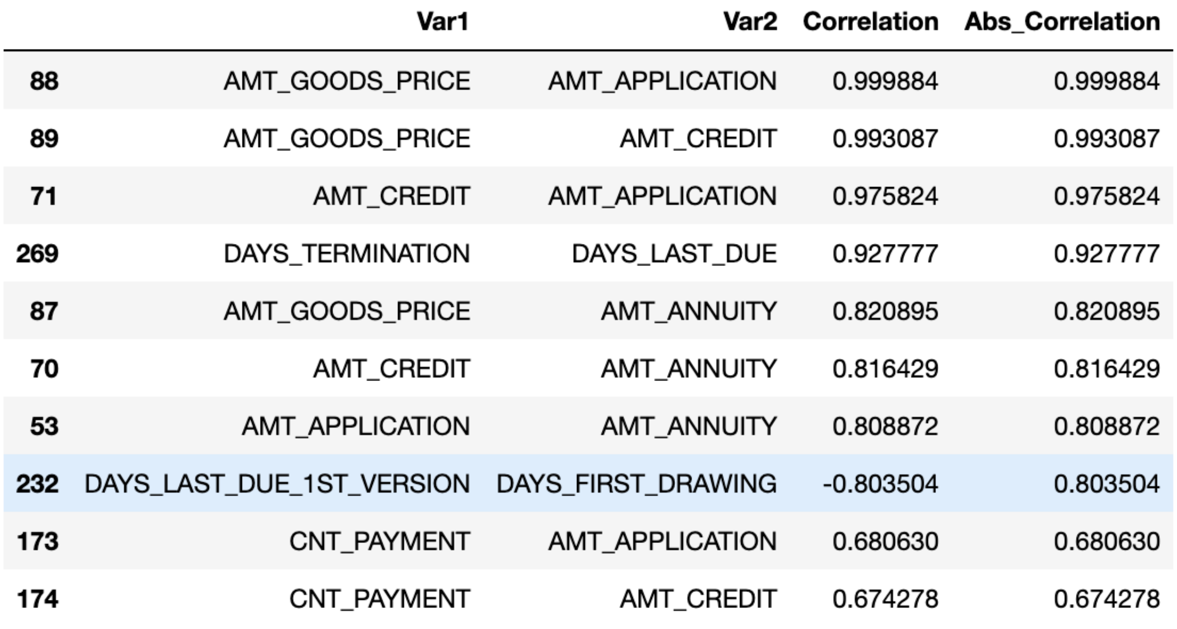


**Observation:-**

-we observe that most of the application are rejected because of HC and at second reason is LIMIT.

# Bivariate/Multivariate analysis

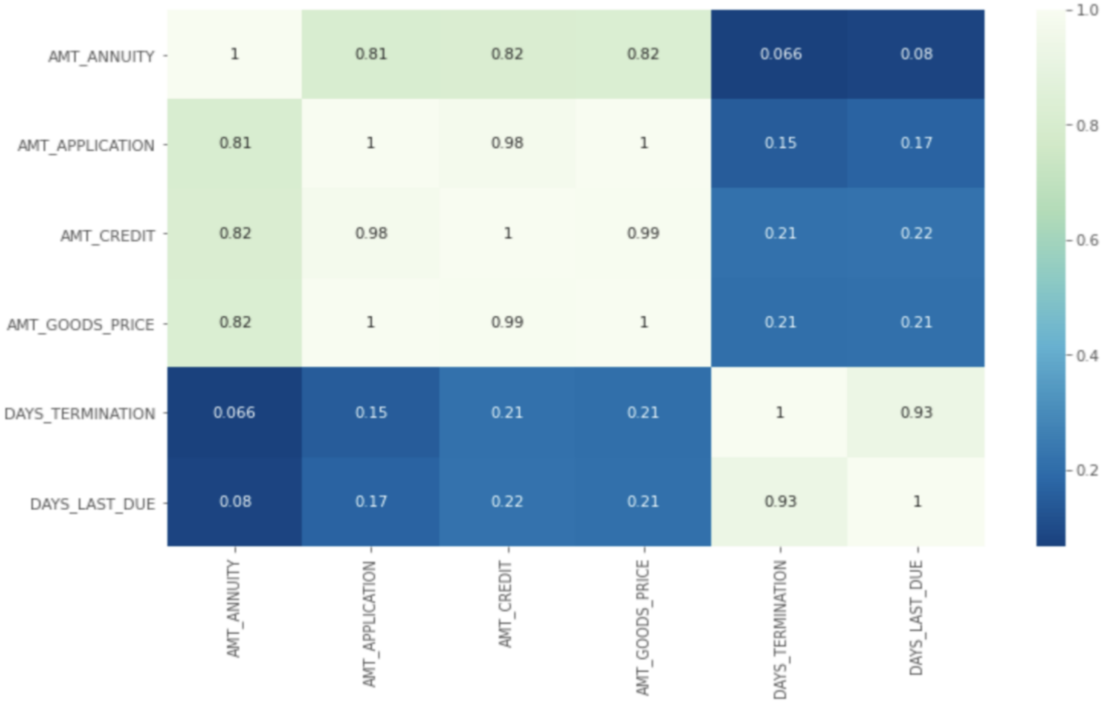
- **Finding top 10 correlation in previous application**



| | Var1 | Var2 | Correlation | Abs_Correlation |
|---|---|---|---|---|
| 88 | AMT_GOODS_PRICE | AMT_APPLICATION | 0.999884 | 0.999884 |
| 89 | AMT_GOODS_PRICE | AMT_CREDIT | 0.993087 | 0.993087 |
| 71 | AMT_CREDIT | AMT_APPLICATION | 0.975824 | 0.975824 |
| 269 | DAYS_TERMINATION | DAYS_LAST_DUE | 0.927777 | 0.927777 |
| 87 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.820895 | 0.820895 |
| 70 | AMT_CREDIT | AMT_ANNUITY | 0.816429 | 0.816429 |
| 53 | AMT_APPLICATION | AMT_ANNUITY | 0.808872 | 0.808872 |
| 232 | DAYS_LAST_DUE_1ST_VERSION | DAYS_FIRST_DRAWING | -0.803504 | 0.803504 |
| 173 | CNT_PAYMENT | AMT_APPLICATION | 0.680630 | 0.680630 |
| 174 | CNT_PAYMENT | AMT_CREDIT | 0.674278 | 0.674278 |

- Plotting pairplots and heatmaps for this correlation

Plotting heatmap for 'AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE','NAME_CONTRACT_STATUS',' DAYS_TERMINATION','DAYS_LAST_DUE



Observation:-

-we observe from this scatter plot that 'AMT_APPLICATION' and AMT_GOODS_PRICE has strong positive linear relationship meaning amount that client has asked for on previous application is highly influenced by amount of goods price on previous application. -we can also observe that AMT_ANNUITY annuity of previous has high influence over

1- AMT_GOODS_PRICE-price of goods on previous application,

2- AMT_CREDIT-Final credit amount on the previous application that was approved by bank and

3- AMT_APPLICATION-how much credit did client ask on previous application

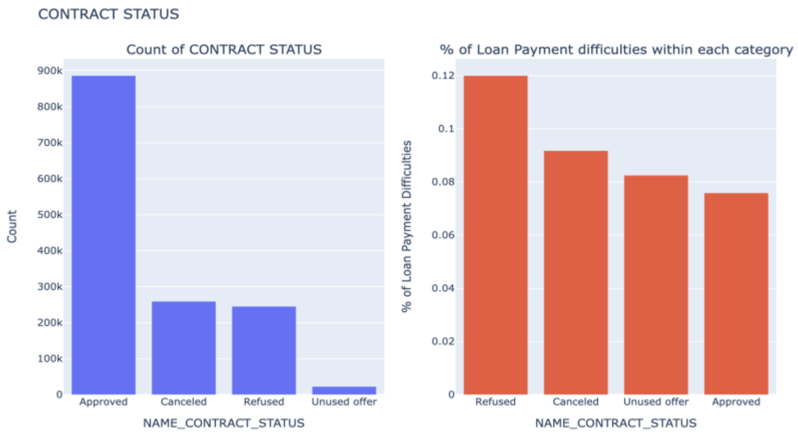
-we also observe that final credit amount on previous application is influenced by goods price and credit amount on previous application.

# Merging the two files and analysing the data

- Merging the two files on 'SK_ID_CURR'  to do some analysis
- Routine check

- Analysis of Contract Status and its category with maximum % of Loan-Payment Difficulties

*Distribution of Contract Status and its category with maximum % of Loan-Payment Difficulties*

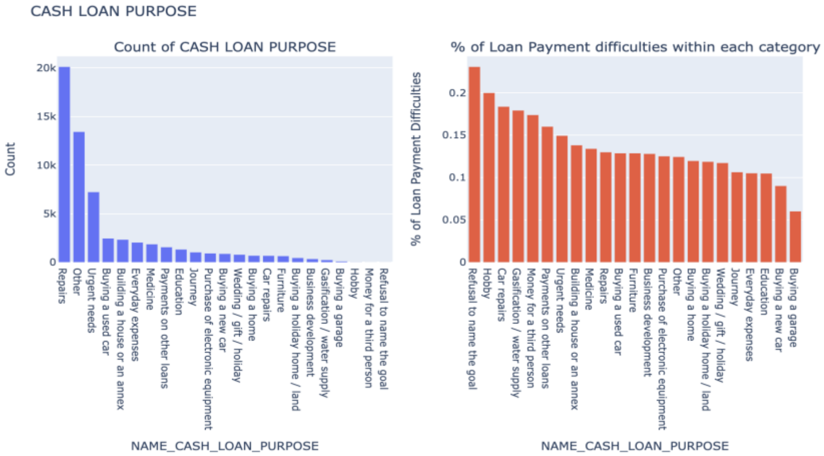


## Observation:-

-we can observe that
-Previously refused contract applicants have higher chances of defaulting.
-Approved contracts applicants have lower risk of defaulting.

- Analysis of 'NAME_CASH_LOAN_PURPOSE', 'CASH LOAN PURPOSE'

*Distribution of Cash Loan Purpose and its category with max % of LoanPaymentsDifficulties*

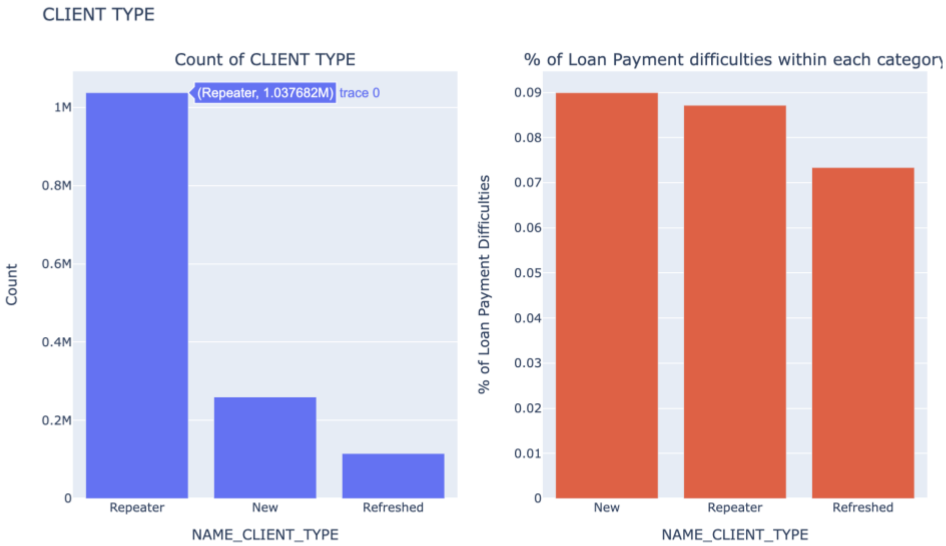


**Observation:-**

-we can see a good pattern that those who refused to mention their purpose of laon have much higher rate and chance of defaulting, we can be careful from those applicants.

- Analysis of 'NAME_CLIENT_TYPE', 'CLIENT TYPE'

  *Distribution of Client Type and its category with maximum % of Loan-Payment Difficulties*



**Observation:-**

From the first graph it can be seen that most of the clients from previous application are 'Repeater'

-'New' clients from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.

-'Refreshed' clients from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

# Recommendations:

## These are my recommendations to consider.

### Current Application

-Most of loan applicants is working class also they are the highest in not defaulting i.e. 50.8% and also defaulting i.e. 61.3% so bank should be careful other factors also with these applicants.
-Passing loans to Pensioners have positive effect as most of them are on time with their payment.
-Married applicants and widows have lower chance of defaulting a payment while there's slight risk of single/not married and civil marriage applicant to default.
-Applicants who live with their parents have more risk associated with them to default and so as applicants who lives in rented apartment.
-Company should also be vigilant with applicants whose occupation sales staff, drivers ,labors as they high slight risk with them to default while t core staff, managers, accountant ,high skill tech staff are less likely to miss their payments.
-There's higher risk with young and teenage applicants as more risk is associated with them to default while applicants who are seniors above 45, have less risk associated with them to default.
-Applicants who have low income have slight risk to default while applicants with high and very income are slightly less defaulters.
- The applicant who are in medium loan credit group tend to apply for the most loan and their chance of defaulting is highest in all loan credit groups around 36%.

- Secondary/secondary special education group applies for loan the most and there's also risk as they default 8% more than non-defaulters of same group while applicant in Higher Education Group has good less chance of defaulting payment mean least risk in this group.

-We observer that those who defaulted have weak linear relation between total income and price of goods(0.038) contrary to applicant who didn't defaulted and have good linear relation(0.35)
-People who are married with Higher education have good credit and are also less defaulting and also single with higher education.

-We observe("Income range and family status vs Credit amount of Loan Non- Payment Difficulties") that applicant who are in very high income range and are married or civil married have good chance of paying loan on time and also applicant who are widow and separated in very high income group have less chance of defaulting. Applicant in high income group who are single also have less chance of defaulting.
-We observe ("Income range and family status vs Credit amount of Loan Payment Difficulties") that applicant in high income group and married have slight more chance of defaulting also applicant in high income group and married also have slight more chance of defaulting-This could be the reason because people in high in come group applied for more loan.
-We can observe that applicant who are working and in high income groups have lower chance of defaulting a payment.

- Applicant who are in Maternity leave group have highest chance of loan payment difficulty so higher risk associated with them.
-Applicants who are "Unemployed" have 2nd highest chance of defaulting payment so company should be careful handling these applicants.
- 'Lower skill Laborers' occupation type applicants have maximum % of Loan-Payment Difficulties.
-Lower secondary education applicants have the highest loan payment difficulties.
-Female seniors have high positive chance of not defaulting payment as well as female middle aged applicants, while young female and middle aged male high risk associated with them to default the payment.

## Previous Application

**-** Most of the loans approved are consumer loans.
- We can observe that loan approval process started on Sunday have lesser chance to get approved compared with others.
we observe from this scatter plot that 'AMT_APPLICATION' and AMT_GOODS_PRICE has strong positive linear relationship meaning amount that client has asked for on previous application is highly influenced by amount of goods price on previous application

## Merged Dataset

-Previously refused contract applicants have higher chances of defaulting.
-Approved contracts applicants have lower risk of defaulting
-we see that car ownership have no effect on approval of laon earlier we observe that owning a car applicant have slight less risk of defaulting.
-Those who refused to mention their purpose of laon have much higher rate and chance of defaulting, we can be careful from those applicants.
-'New' clients from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.
-'Refreshed' clients from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.