# Credit EDA Case study

Abhishek Kumar Singh
Karthik Balan

# Business Objective

- **This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.**

- **Two types of risks are associated with the bank's decision:**
    1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
    2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

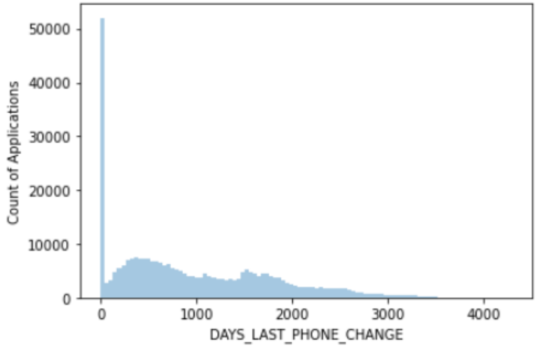
# Step 1- Correcting Data Types

- Analysed variables with only a few unique values but loaded as ‘numeric’ in the data frame
- Maximum Unique values threshold set to 10 for number of categories
- Found 53 such variables that have a very few categories and loaded as continuous, they can be corrected from ‘numeric’ to ‘str’ i.e. continuous to categorical
- Converted variables to their absolute values
  - 'DAYS_BIRTH'
  - 'DAYS_EMPLOYED'
  - 'DAYS_REGISTRATION'
  - 'DAYS_ID_PUBLISH'
  - 'DAYS_LAST_PHONE_CHANGE’

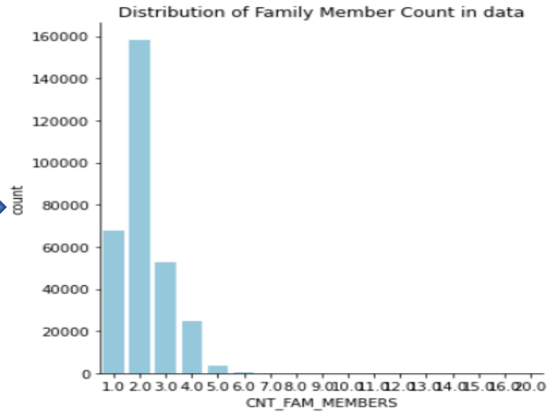
  All of these variables represent no. of days  in history when the event last happened. So these can be co nverted to positive numbers for analysis

# Strategy to handle Missing Values

- We took 5 variable with <13% missing values for showing the strategy
- ***DAYS_LAST_PHONE_CHANGE***
  - No. of observations with missing value = 1
  - Strategy to Impute – Either remove the observation since there is only one such observation.**(Preferred method)**
    If the row is belonging to minority TARGET class
    we may not want to remove the row but replace with
    value 0 (being the most observed value in data)



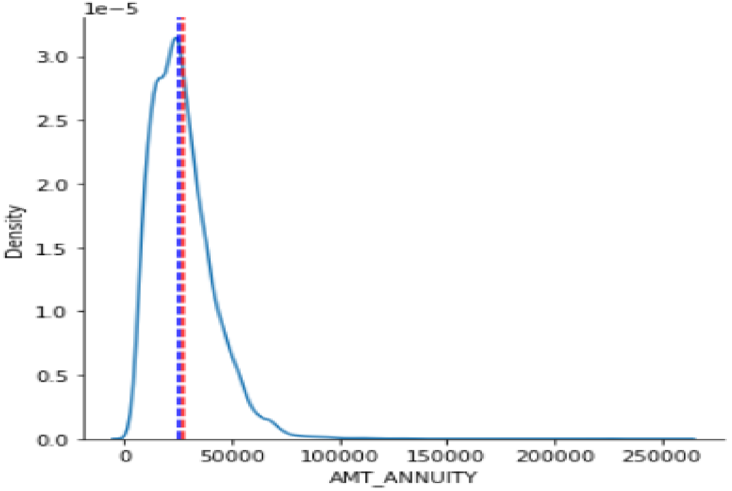
- ***CNT_FAM_MEMBERS***
  - No. of observations with missing value = 2
  - Strategy to Impute – Either remove the observation since there are only 2 observations. **(Preferred method)**
    Or replace with value 2 (being the most observed value in data)



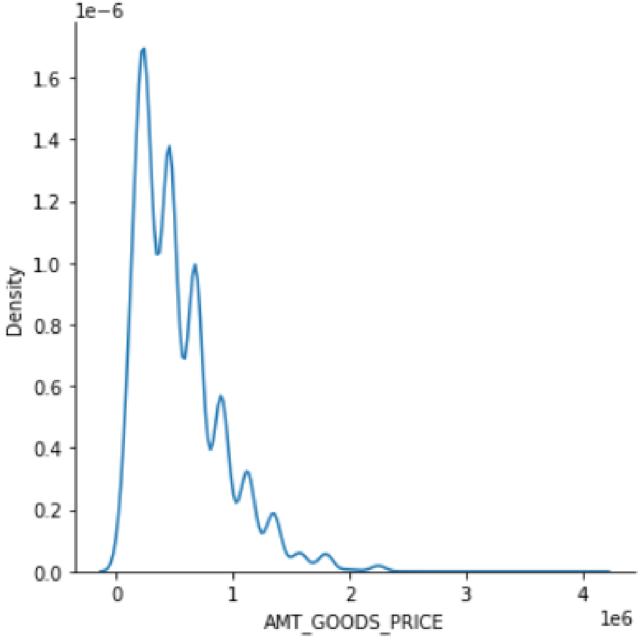
- ***CODE_GENDER***
  - No. of observations with missing value = 4
  - Females = 66%, Males = 34%
  - There are almost double Female applicants compared to Males, therefore the best way to impute CODE_GENDER is as Female (value=F)

# Strategy to handle Missing Values – Continued…

- ***AMT_ANNUITY***
  - No. of observations with missing value = 12
  - Strategy to Impute –
    - Mean and Median are very close with high positive kurtosis
    - High probability of a missing value to fall near median value.
    - We may impute missing values with Median value = 24903

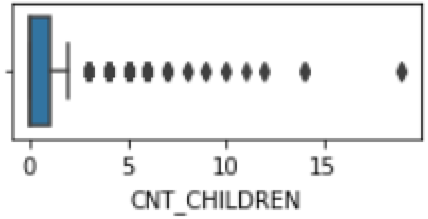


- ***AMT_GOODS_PRICE***
  - No. of observations with missing value = 278
  - Strategy to Impute –
    - Variability in data distribution, data for the variable is widely spread
    - Tried to find additional strategies to Impute
    - Found that variable is highly positively correlated (0.99) to another variable AMT_CREDIT, which has no missing values
    - We may use value 0.9*AMT_CREDIT to best impute the variable
    - By definition this makes sense too, more the amount of goods a person owns, Higher are hos chances of getting bigger loan amounts.

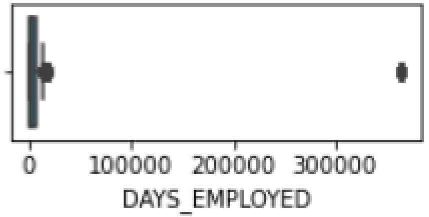


# Strategy to handle Outliers

- ***CNT_CHILDREN***
  - Strategy to handle outliers –
    - Used a box plot to see the distribution
    - Most of the observations fall under 15 CHILDREN
    - No. of observations over value 15 are only 2
    - Simply removing the rows could be the best option

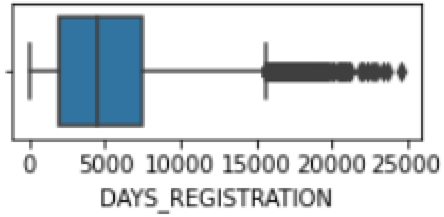


- ***DAYS_EMPLOYED***
  - Strategy to handle outliers –
    - The box plot clearly indicates the outliers in data
    - There are 55374 observations in data with over 20000 days of employments
    - This looks unrealistic in real world so these values can be considered as outliers
    - Interestingly all of them have same value present as 365243, this could be assumed as either a data entry error or a default value used for this variable to represent missing value.
    - Strategy to handle this outlier could be to replace this max value of 365243 with the median value of variable DAYS_EMPLOYED i.e. 2219 days



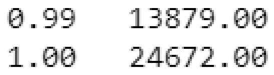
- ***DAYS_REGISTRATION***
  - Strategy to handle outliers –
    - The variable has outliers but they are spread and not like the above 2 variables
    - The data is fairly distributed between 25th, 50th and 75th percentiles

**Continued on next slide ….**



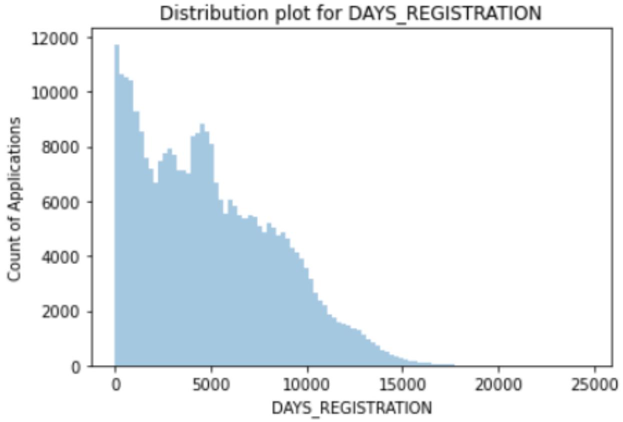
# Strategy to handle Outliers- Continued…

- Looked at 99th and 100th percentiles for the data



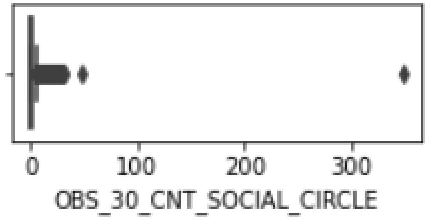
```
0.99    13879.00
1.00    24672.00
```

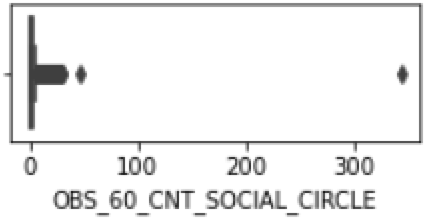
- Looking at the data distribution, we suggest to Cap the data at maximum of value 15000
- So Capping data is the suggested strategy for outlier handing for this variable.



- ***OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE***
  - Strategy to handle outliers –
    - Both the variable are very similar in description
    - They define the number of people in applicant’s social circle who are defaulting payments for over 30 and 60 days
    - It is highly unlikely that a person has over 30 people in social circle defaulting payments
    - Looking at number of rows where value for the variable is >30  we find only 2 rows for both the variables

It is therefore recommended to remove the 2 rows to handle these outliers.





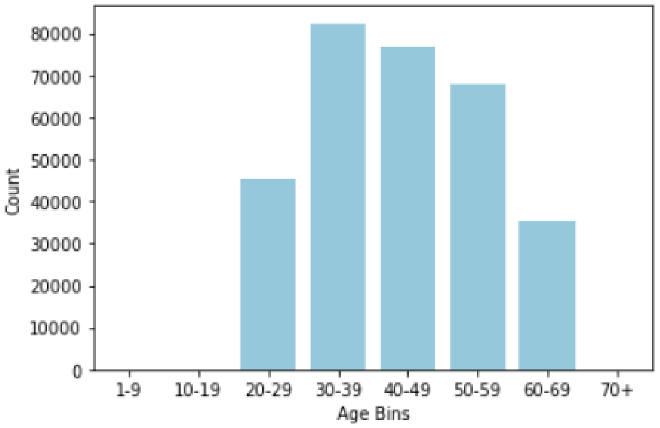
# Binning variables

Identified 2 variables for which binning might be useful

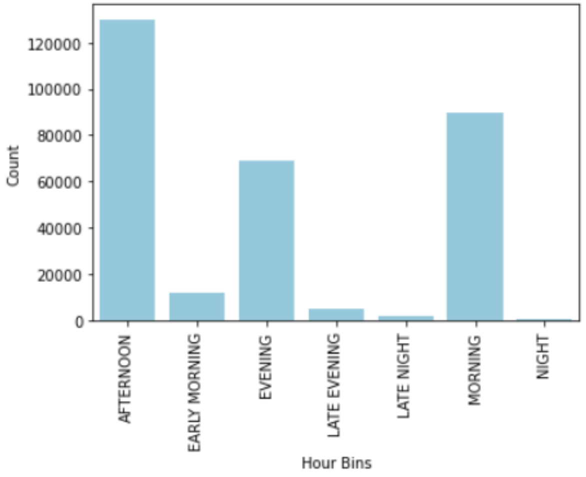
1. DAYS_BIRTH
   - Converted the days to years
   - Binned the years to "1-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70+“

2. HOUR_APPR_PROCESS_START
   - Binned the variable into Early_Morning, Morning, Afternoon, Evening, Late_Evening, Night, Late_Night
   - On a 24 hour clock value of 23:00 Hours and 01:00 Hours are very close but on data it will look very different in magnitude, therefore it is good to bin to more meaningful categories.

Binning variable DAYS_BIRTH

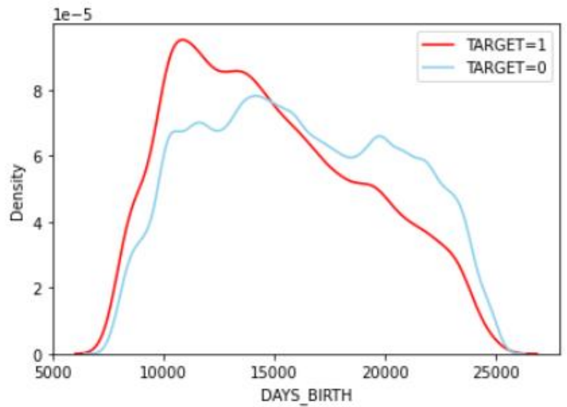


Binning variable HOUR_APPR_PROCESS_START

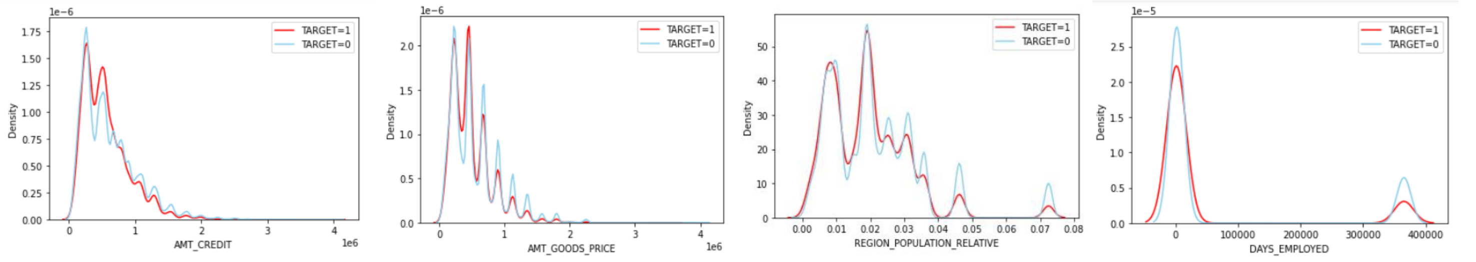


# Analysis

- **Imbalance percentage for TARGET class –**
  - Data is imbalanced with 92% TARGET=0 and 8% TARGET=1
- **Univariate Analysis – We selected only 5 variables to do detailed analysis**
  - **Categorical Variables**
    - 1. CODE_GENDER - Males have less defaults(TARGET=1) compared to Females
    - 2. FLAG_OWN_CAR - People who own a car are more likely to default(TARGET=1)
    - 3. NAME_INCOME_TYPE - Categories like Pensioners are more likely to default(TARGET=1), this variable can be further tested for importance during Categorical-Continuous bivariate analysis
    - 4. FLAG_EMP_PHONE - Applicants who do not provide Employer Phone number are more likely to default(TARGET=1)
    - 5. REG_CITY_NOT_LIVE_CITY - Applicants whose Permanent address does not match Contact address are more likely to make defaults(TARGET=1)
  - **Continuous Variables – We selected 5 values for analysis**
    - DAYS_BIRTH(Age) - Out of the selected variables only this variable seems to be one which varies for TARGET variable classes(0 and 1) , defaulters seems to be more concentrated around an age group



Variables that do not show any special difference in TARGET classes

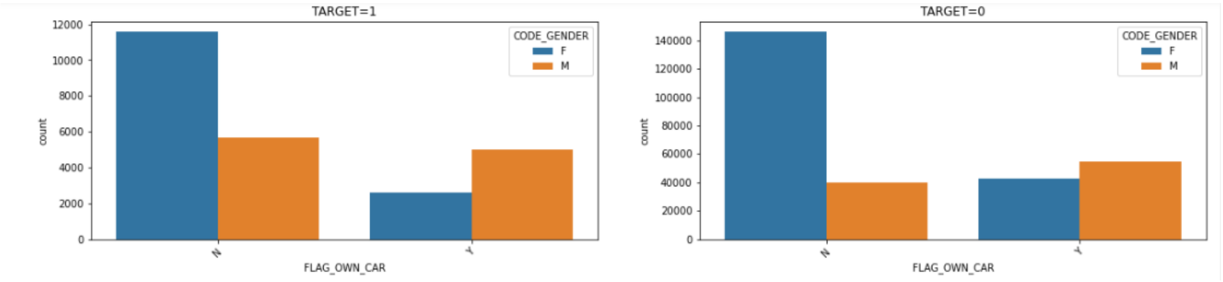


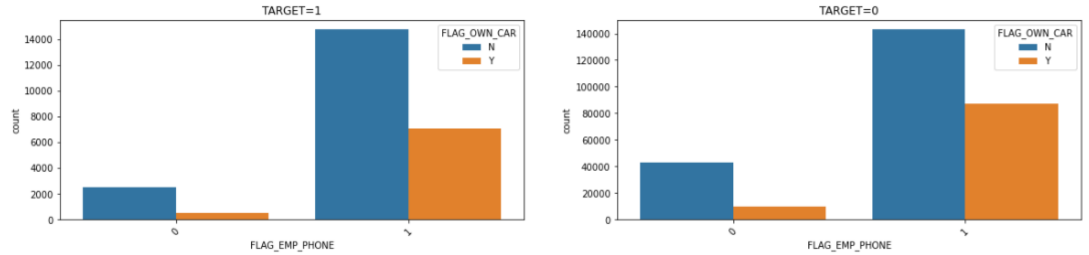
# Analysis – Continued…

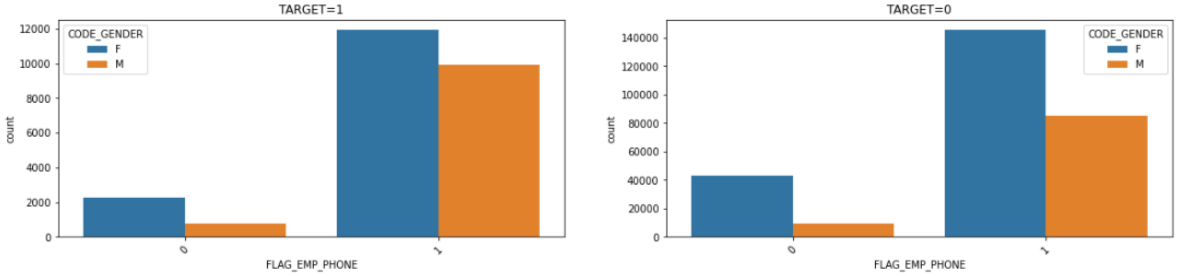
- **Bivariate Analysis –**
    - **Categorical – Categorical Variables**

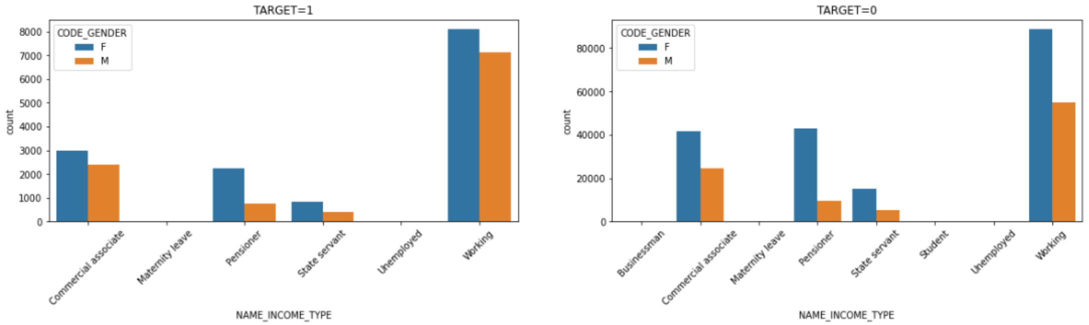
**Out of our selected categorical variables only a few show significant evidence-**

- CODE_GENDER vs. FLAG_OWN_CAR- Males who do not own a car are more likely to default
- CODE_GENDER vs. NAME_INCOME_TYPE- Working professional who own a car are less likely to default
- CODE_GENDER vs. FLAG_EMP_PHONE – Male candidates seems to make more default even after providing employer phones, therefore is a strict employer check is required for male applicants.
- FLAG_OWN_CAR vs. FLAG_EMP_PHONE- Candidates who own a car and have also provided the employer phone number are less likely to default.



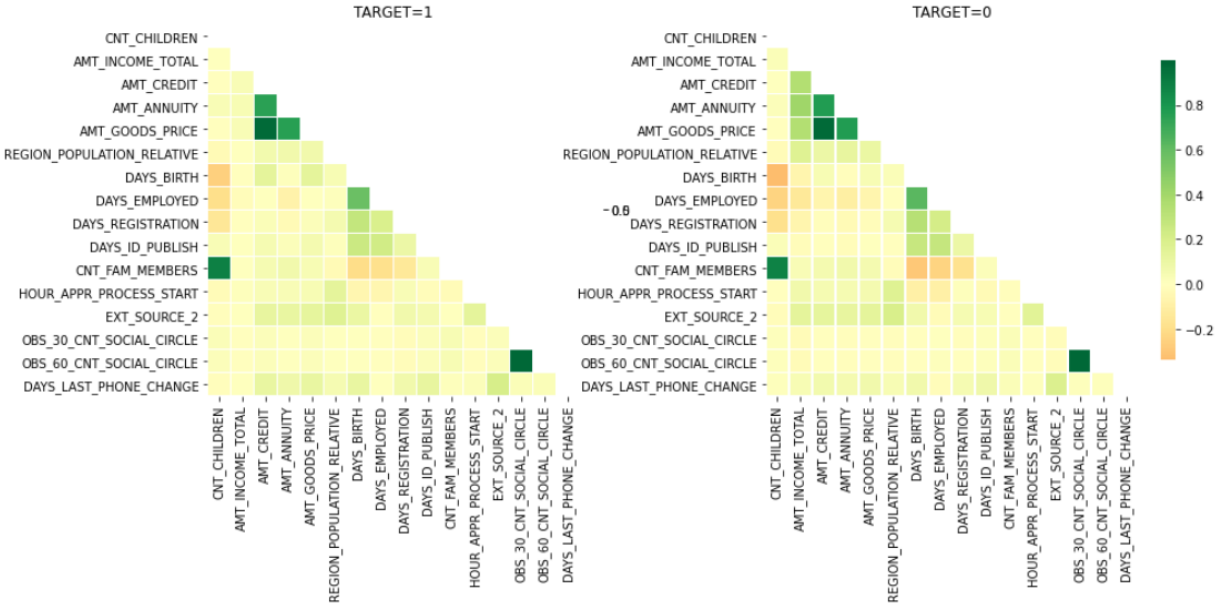






# Analysis – Continued…

- **Bivariate Analysis –**
  - **Continuous – Continuous Variables**
    - Did a correlation check on selected continuous variables for both TARGET=1 and TARGET=0
    - Selected variables with over 0.5 correlation (on absolute value)



We can observe that most of the variables show very similar correlation between TARGET 1 and 0 except variable AMT_INCOME_TOTAL where it seems to be more correlated to variable AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE when TARGET=0.
This means when Credited amount is based on total income and amount of goods an applicant owns then there are less chances of a default.

## Explaining high correlations based on column description

1. AMT_CREDIT and AMT_ANNUITY - This is explainable, most of the high amount loans are paid over a longer period of time
2. AMT_CREDIT and AMT_GOODS_PRICE - Loan amount are mostly decided based on how much a person is worth, so more the amount of goods a person owns, more the credit amount to the loan
3. DAYS_EMPLOYED and DAYS_BIRTH - Age and employment tenure are highly correlated and it makes sense
4. CNT_FAM_MEMBERS and CNT_CHILDREN - More the no. of children more will be the no. of family members, clearly visible in correlation as well

# Analysis – Final Columns

- **Bivariate Analysis –**
    - **Continuous – Categorical Variables**
        - Did analysis for by plotting boxplots for selected continuous variable by each selected categorical variable
        - Female applicants with more than 2 family members are more likely to default compared to applicants with a family of 2 or less
        - Smaller families(<=2) who do not own a car are more likely to default

- **Based on our analysis following variables seems to be important for deciding the TARGET class value in order for the bank to take decision**
    - CODE_GENDER
    - FLAG_OWN_CAR
    - NAME_INCOME_TYPE
    - FLAG_EMP_PHONE
    - REG_CITY_NOT_LIVE_CITY
    - DAYS_BIRTH
    - AMT_CREDIT
    - AMT_ANNUITY
    - AMT_GOODS_PRICE
    - DAYS_EMPLOYED
    - DAYS_BIRTH
    - CNT_FAM_MEMBERS
    - CNT_CHILDREN
    - DAYS_EMPLOYED
    - DAYS_REGISTRATION