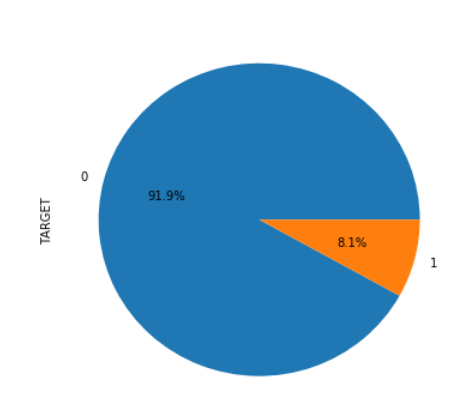
**CREDIT EDA CASE STUDY:**

Analysing application dataset:

* Reading Data: data is read and stored in pandas dataframe
* Data Cleaning:
  + Identifying quantitative and categorical variables
  + Converting datatype of variables where ever necessary

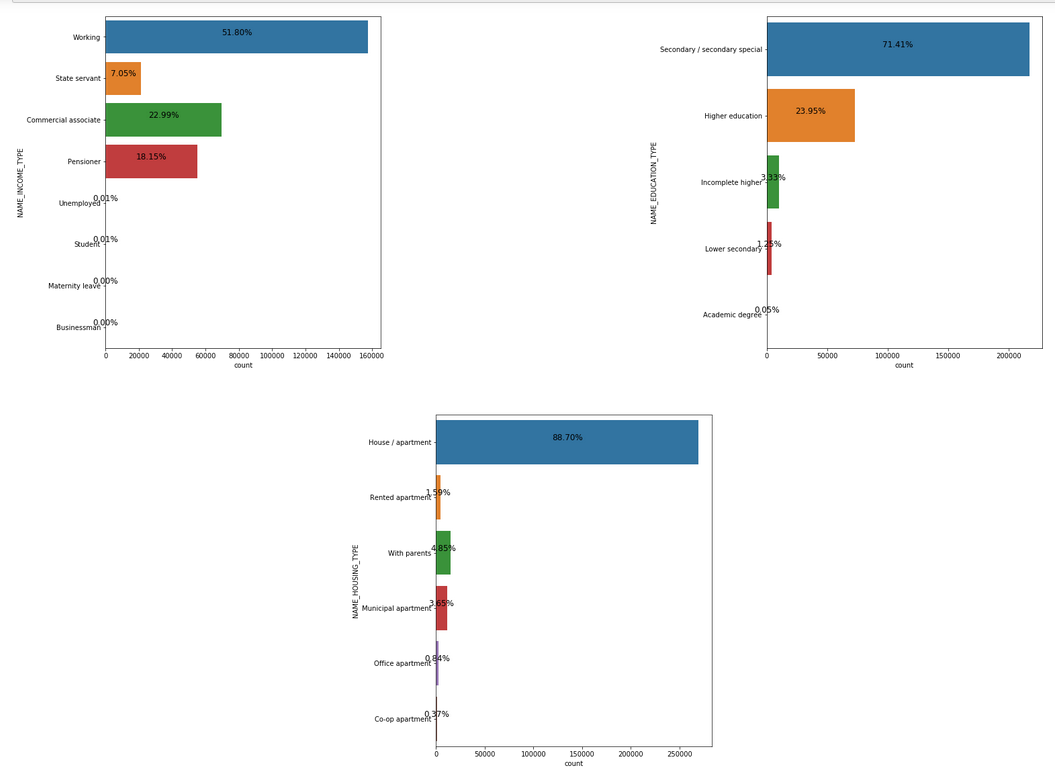
E.g.: SK\_ID\_CURR is id of loan was integer it has to be converted to string

* + Dealing with missing values : Removed columns having 50% of null values
  + There are some rows in CODE\_GENDER, NAME\_FAMILY\_STATUS columns where missing values are represented by XNA, Unknown therefore treating them as null values by replacing them by nan
  + Detecting outliers : There were some outliers in columns to be analysed so only 99 percentile of data is considered for those columns('CNT\_CHILDREN','AMT\_INCOME\_TOTAL')
* Considering these Columns for analysis
  + Numerical Variables
    1. CNT\_CHILDREN
    2. AMT\_INCOME\_TOTAL
    3. AMT\_CREDIT
    4. AMT\_ANNUITY
    5. AMT\_GOODS\_PRICE
  + Categorical Variables
    1. CODE\_GENDER
    2. FLAG\_OWN\_CAR
    3. FLAG\_OWN\_REALTY
    4. NAME\_INCOME\_TYPE
    5. NAME\_EDUCATION\_TYPE
    6. NAME\_FAMILY\_STATUS
    7. NAME\_HOUSING\_TYPE
    8. TARGET
* Univariate Analysis of Categorical variables:
  + Analysing Target variable:

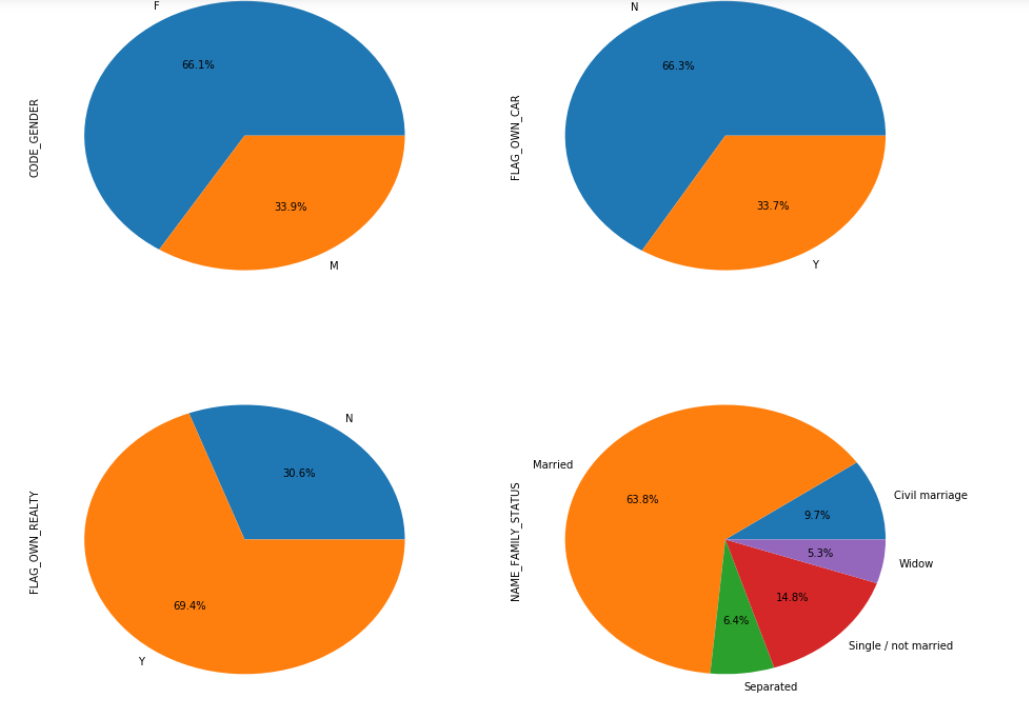


As we can see from above graph there is imbalance of data 91.9% applicants doesn't have any difficulty paying loan while only 8.1% had difficulty

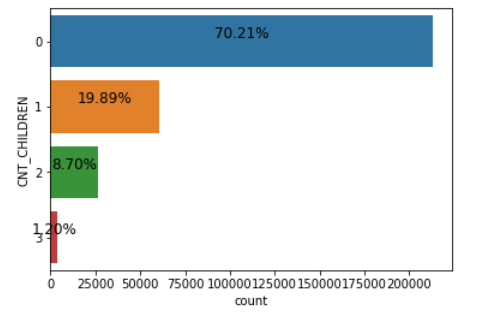
* + Analysing other categorical columns:



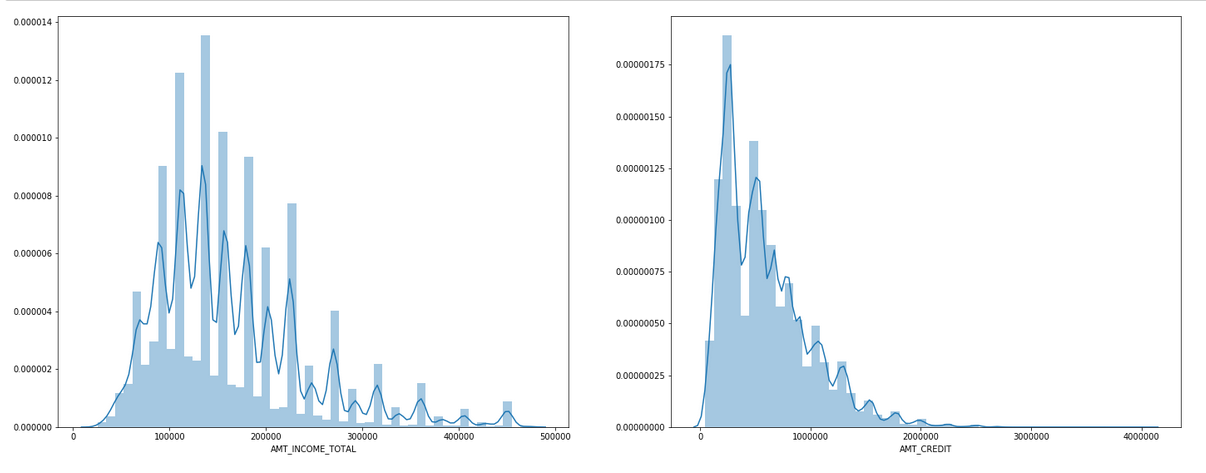
**Inferences from above plots**

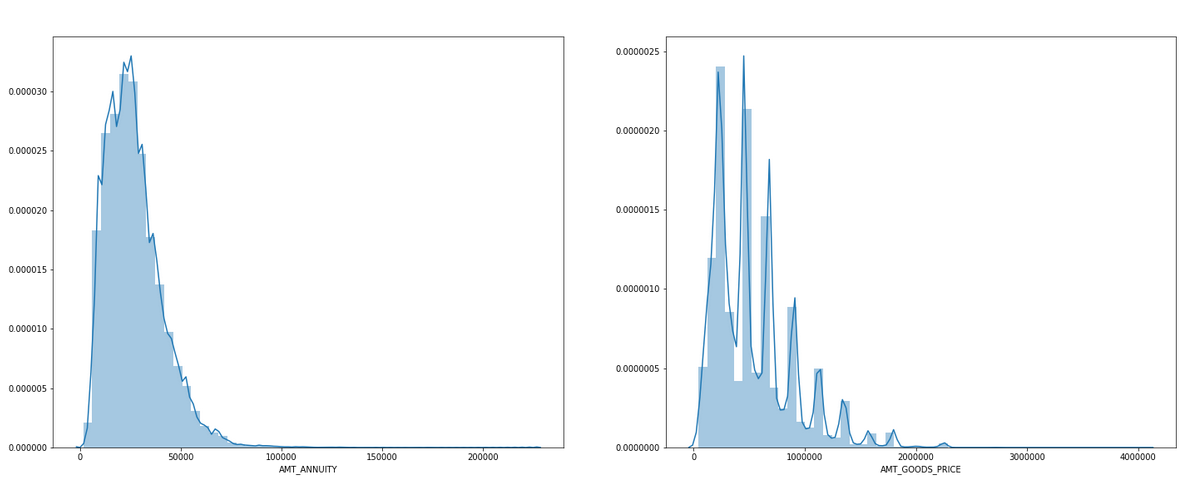
* + 1. NAME\_INCOME\_TYPE :Many applicants are working class and least are the Businessman
    2. NAME\_EDUCATION\_TYPE :Many applicants have Secondary/Secondary Special Education and very few applicants have Academic degree
    3. NAME\_HOUSING\_TYPE :About 88.70% of applicants have House/Apartments and very few applicants 0.37% stay in coop apartments
  + 

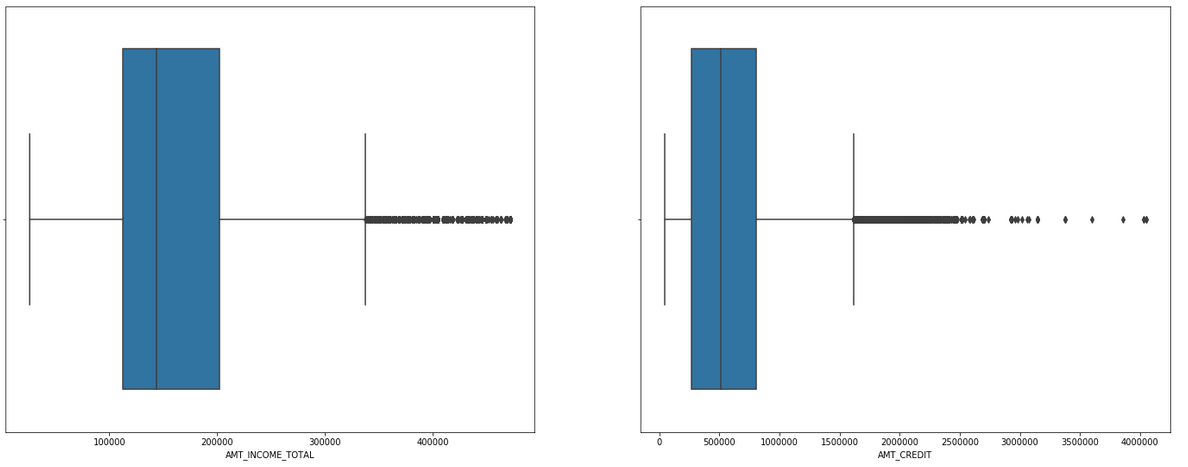
**Inferences from above plots**

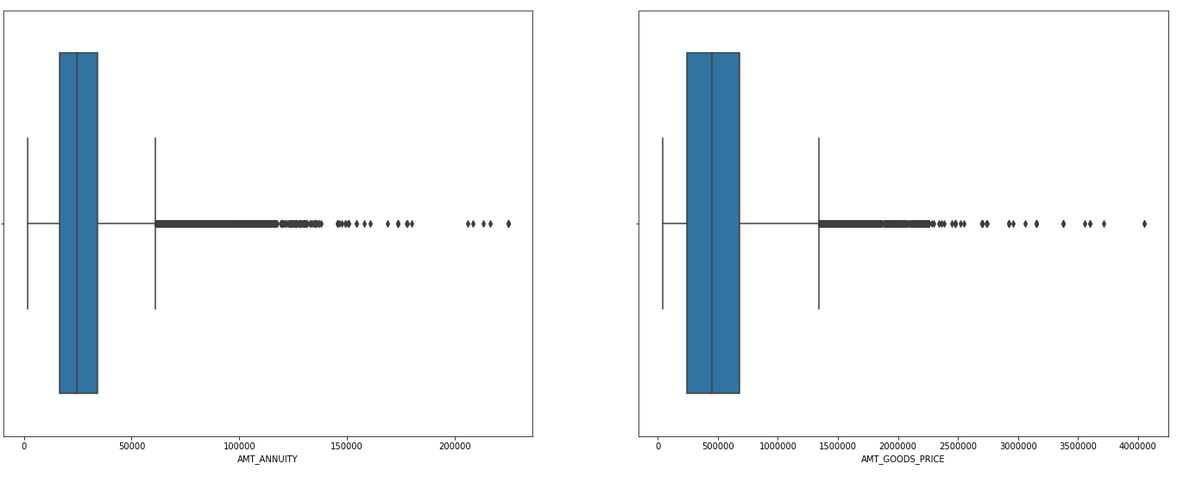
* + 1. CODE\_GENDER :Female applicants are more than male
    2. FLAG\_OWN\_CAR :Majority of applicants don't own car
    3. FLAG\_OWN\_REALTY :Majority of applicants own Realty
    4. NAME\_FAMILY\_STATUS :Majority of applicants are married and 5.3% of applicants are widow
* Univariate Analysis of Categorical variables:
  + 

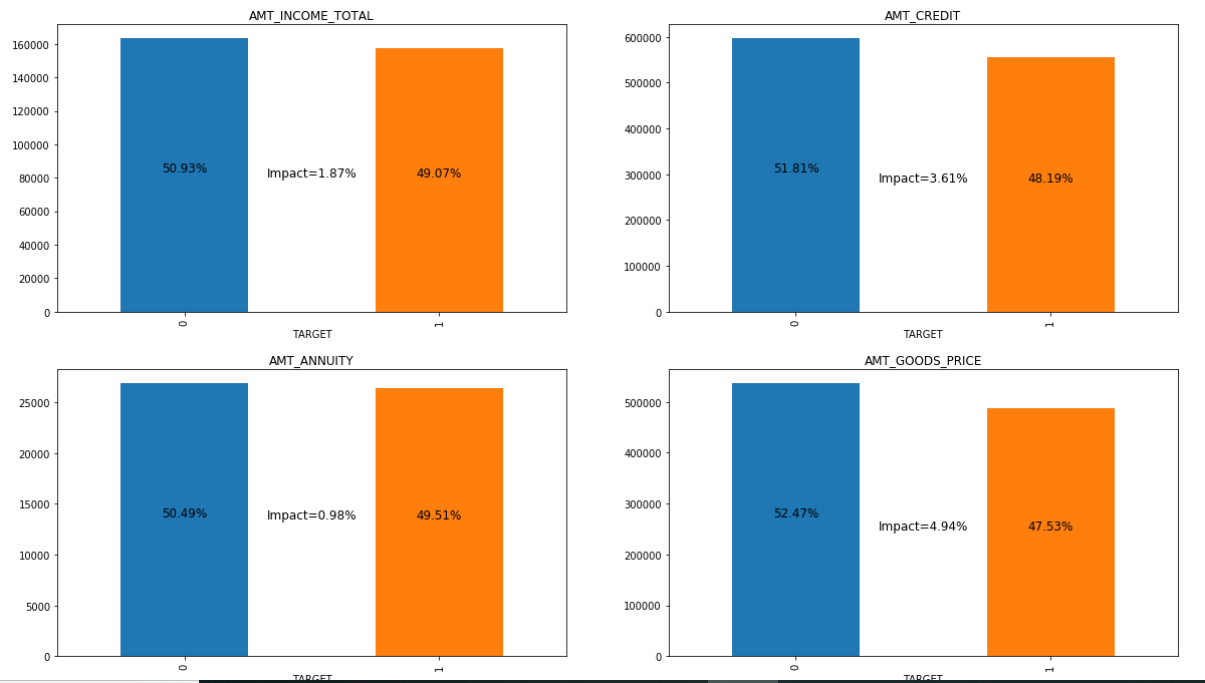
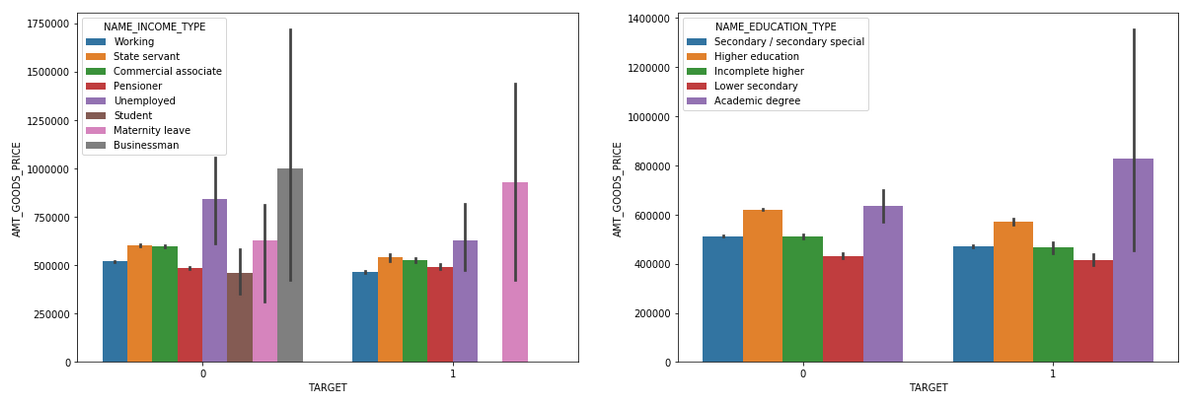
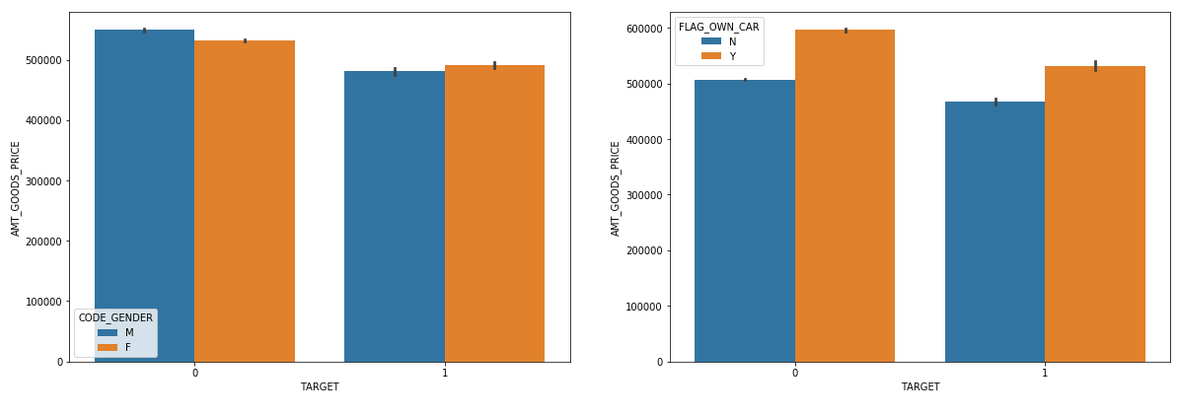
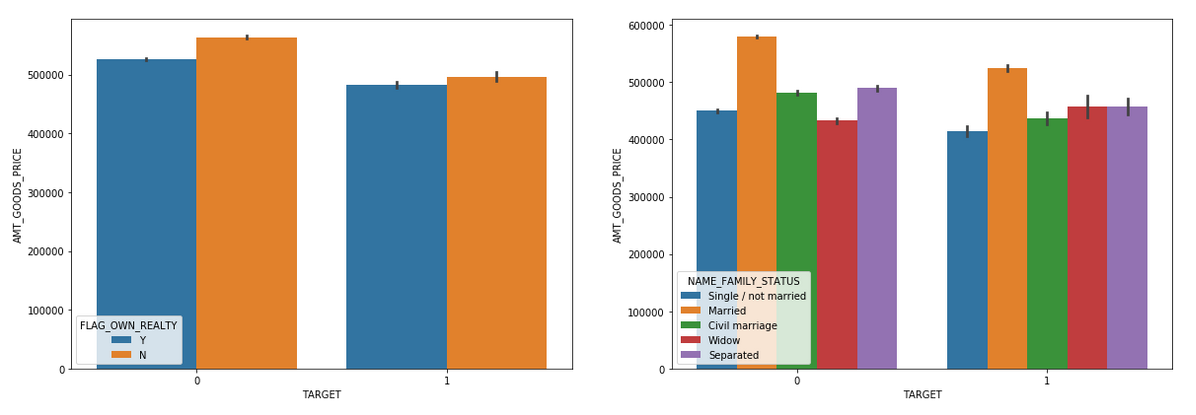
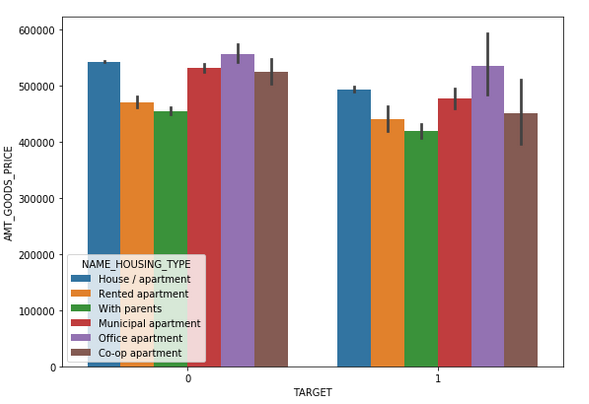
70.21% of the applicants don't have children while only 1.2% of applicants have children

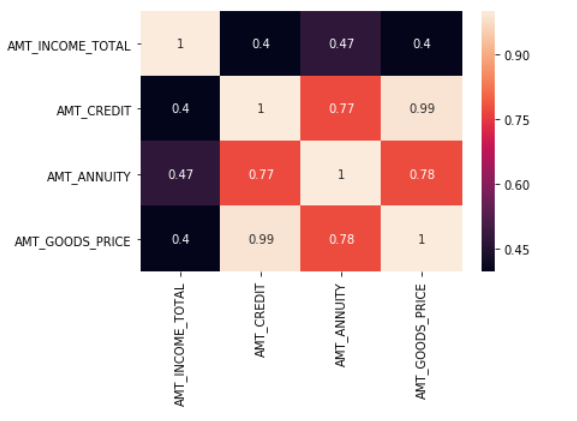
* + 



* + 



* + **Inferences from above plots**
    1. AMT\_INCOME\_TOTAL :Most of the applicants total income lies between 20,000 to 2,00,000
    2. AMT\_CREDIT :Majority of applicants Total credit lies between 10,000 to 10,00,000
    3. AMT\_ANNUITY :Many of the applicants annuity amount lies between 1000 to 50,000
    4. AMT\_GOODS\_PRICE :Majority of applicants Goods price for which they are taking loan lies between 10,000 to 8,00,000
* Segmented univariate analysis on quantitative variables by choosing target as the basis of segmenting:
  + 
  + **Inferences from the above plots(we have done impact analysis i.e. which quantitative column is impacting more to Target variable):**
    1. In AMT\_GOODS\_PRICE there is decrement of 4.94% of mean amount from applicants who doesn't have any difficulty to applicants who has difficulty paying loans, applicants who has difficulty paying loan tend to purchase cheaper goods compared to those who don't have difficulty
    2. In AMT\_ANNUITY there is decrement of 0.98% of mean amount from applicants who doesn't have any difficulty to applicants who has difficulty paying loans.
    3. so the goods price(AMT\_GOODS\_PRICE) is doing more impact to applicants who has difficulty paying loans(TARGET) compared to AMT\_ANNUITY
    4. the impact of quantitative variables on TARGET are arranged in ascending order :
    5. AMT\_ANNUITY,AMT\_INCOME\_TOTAL,AMT\_CREDIT,AMT\_GOODS\_PRICE
* Bivariate Analysis: Since AMT\_GOODS\_PRICE is causing more impact on TARGET so doing bivariate analysis on AMT\_GOODS\_PRICE
  + 
  + 
  + Some interesting Inferences from the above plots
    1. NAME\_EDUCATION\_TYPE: Usually applicants who has difficulty paying loan tend to purchase cheaper products than who don't(we have seen this in segmented univate analysis) but here those who have academic degree have purchased costly products and they are having trouble paying loan.
    2. NAME\_INCOME\_TYPE: Maternity Leave applicants tend to purchase costly products and they are having trouble repaying loan,Businessmans don't have any problem of reaying bills and they tend to purchase costly products
    3. NAME\_FAMILY\_STATUS:Widows tend to spend more on goods and have problem repaying loan.
    4. other than the columns and dimensions mentioned above there is similar pattern between those who don't have trouble to those who have trouble repaying loans
* Finding Correleation between quantitative variables:



If the Goods price are higher, applicants tend to take higher credit amount therefore AMT\_GOODS\_PRICE and AMT\_ANNUITY are highly correleated

**Recommendation:**

* since amount of goods price is creating more impact on Target so it must be thoroughly checked before giving loans
* loans can be given to businessman since they don't have any problem repaying loans
* banks should be careful while giving loans to academic degree holders who are purchasing costly goods because they have problem paying loans
* banks should be careful while giving loans to widows who are purchasing costly goods because they have problem paying loans