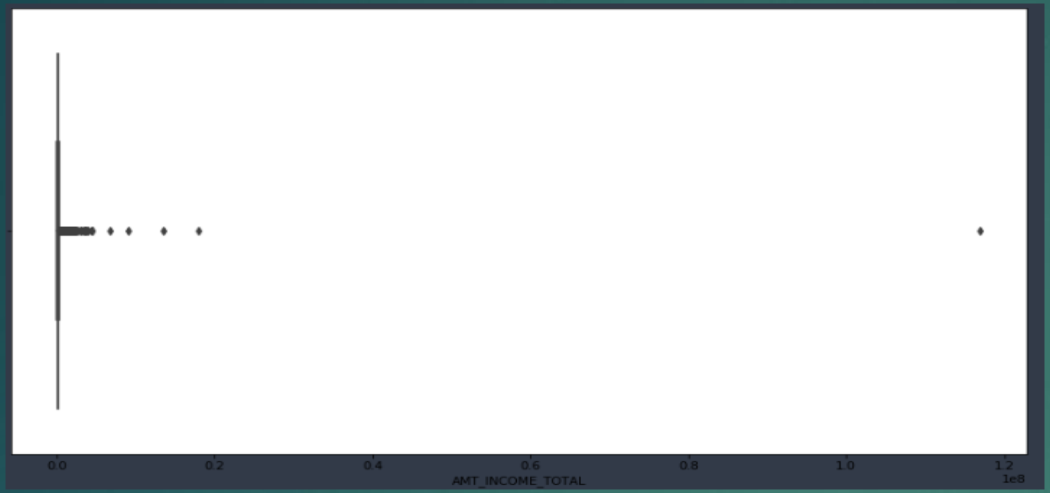
Credit EDA Case Study

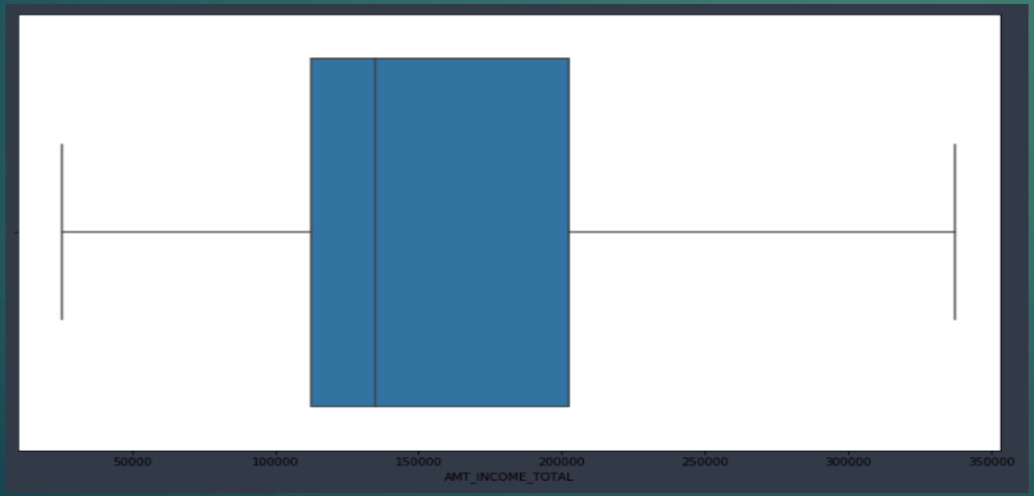
Key Observations

- After loading the source file in a dataframe, selecting the required columns and then performing clean-up activity on them, the first thing that needed to be checked was the presence of outliers.
- For this activity only 3 major columns were considered AMT_INCOME_TOTAL, AMT_CREDIT and AMT_ANNUITY.
- Outliers were present for all the 3 columns mentioned above.

- AMT_INCOMT_TOTAL
  - Before removing the ouliers



  - After removing the outliers



- Detailed steps on how to remove the outliers are mentioned in the jupyter notebook
- In similar manner, outliers were detected in AMT_CREDIT and AMT_ANNUITY column.
- Post detecting the outliers, we moved on to perform the univariate analysis for categorical variables based on target value 0 and1.
- From the plot and results displayed for NAME_CONTRACT_TYPE column it is observed that there are only 2 distinct types of loans which are sanctioned, one is Cash Loans another one is revolving loans. Furthermore, the number of Cash Loans are tremendously high compared to Revolving Loans.
- Looking at subplot 2, it is observed that the number of females who have defaulted the loans are higher than number of males who have defaulted the loan.
- From the plot drawn NAME_INCOME_TYPE column, it is observed that maximum number of people that have defaulted the loan are salaried.

- Similarly the univariate analysis was carried out for other categorical columns and their observations are mentioned with appropriate comments in jupyter notebook
- After completing the univariate analysis for categorical values, we have performed the univariate analysis for continuous columns.
- Distplot was used to perform analysis of continuous variables.
- For all the continuous variables the dist plot was plotted and the appropriate comments are mentiond in jupyter notebook
- Post that, we found correlation between numerical columns.
- From heatmap and python dataframe, we observed that below mentioned columns are highly correlated.

VAR1	VAR2	CORR
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
AMT_ANNUITY	AMT_CREDIT	0.752195

- After findind, the correlation between numerical columns, we moved on to perform bivariate analysis.
- After performing the bivariate analysis on target value equal to 0, we observed that amt_income_total column is highly correlated with amt_credit, amt_annuity and amt_goods_price. (Detailed comments are mentioned in jupyter notebook)
- Similarly, For account holders that have defaulted, amt_income_total column has very less correlation with amt_credit, amt_annuity and amt_goods_price columns. (Detailed comments are mentioned in jupyter notebook)
- Post completing the bivariate analysis, we imported the previous application file, selected only that columns that were required and performed the clean-up activity on them.
- Post completing the clean-up activity we performed the univariate and bivariate analysis on the previous application data. (Detailed comments are mentioned in jupyter notebook)