upGrad

#LifeKoKaroLift

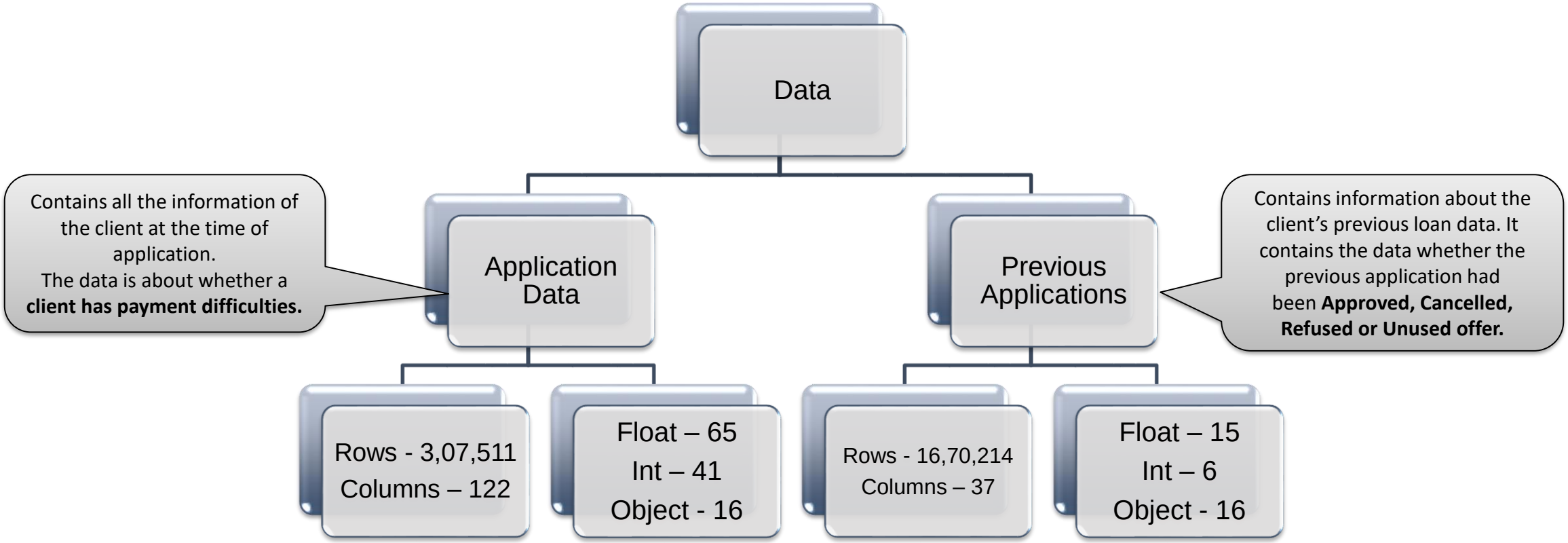
Credit EDA Case Study

Submitted By – Alisha Sharma and Abhishek Chopra

Credit EDA Case Study

Data Description

upGrad



- 43 Columns are selected from Application Data to perform the analysis
- Columns with more than 16% missing values are removed
- Some missing values in other columns are imputed with relevant values
- Data types of some columns are changed for this analysis

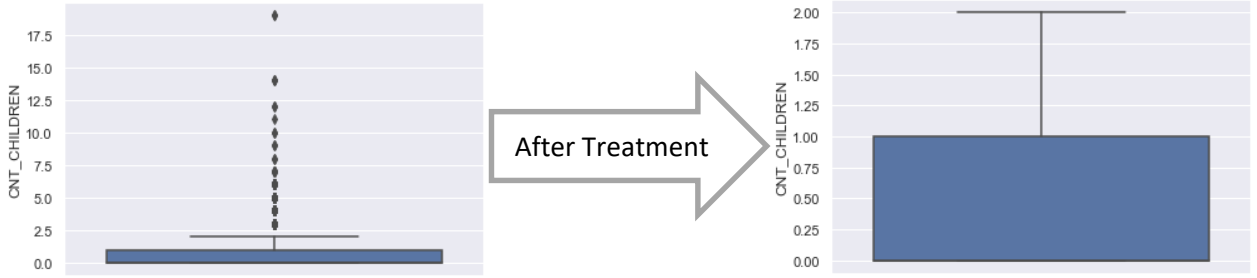
Handling Outliers

upGrad

Outliers checked for **CNT_CHILDREN, AMT_INCOME_TOTAL, CNT_FAM_MEMBERS, DAYS_BIRTH, AMT_CREDIT**

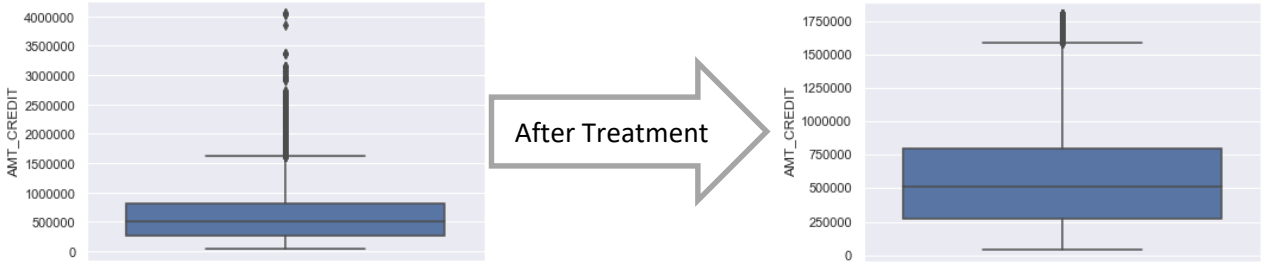
Total Rows after treatment – 2,96,545

CNT_CHILDREN

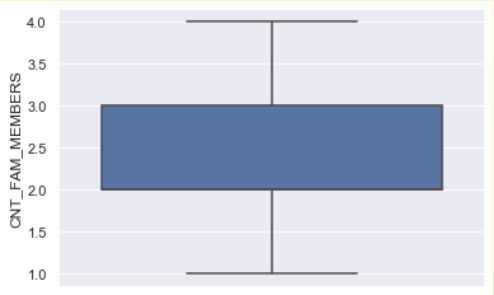


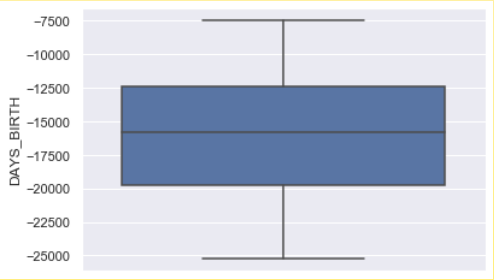
AMT_INCOME_TOTAL

AMT_CREDIT



No treatment required for **CNT_FAM_MEMBERS and DAYS_BIRTH**



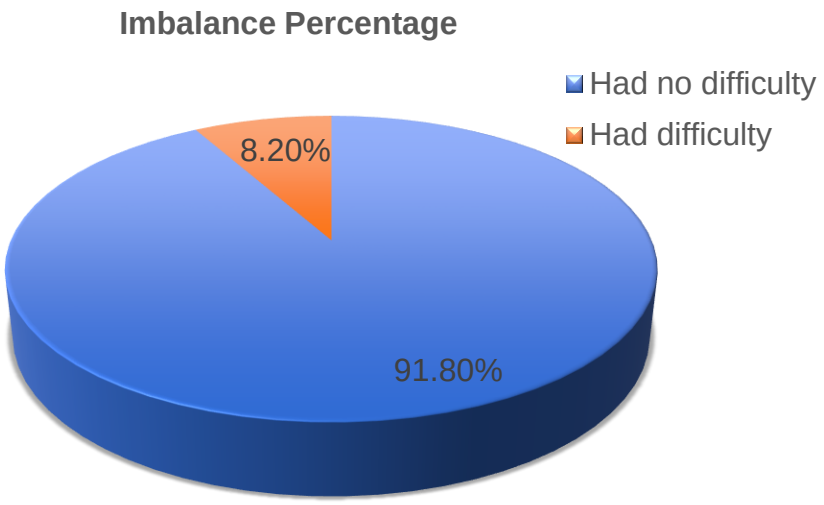


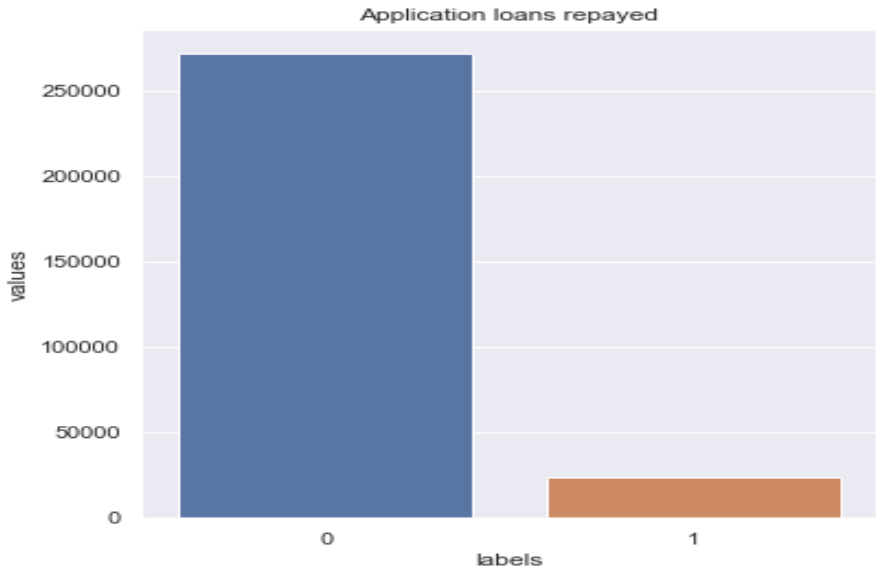
Analysis on Application Data (1/2)

upGrad

Analysis of Target Variable and dividing the data into 2 parts-

1. **0 – Customer did not have any payment difficulties**
2. **1 – Customer had payment difficulty in making payments**



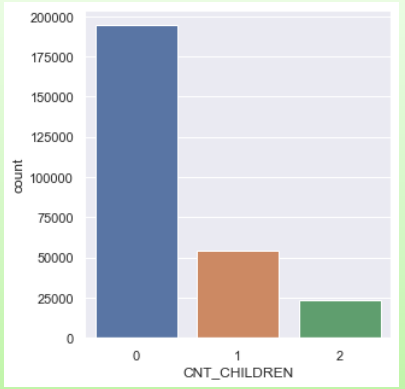


- **91.8% (272460/296545)** of the customers had payment difficulties and **8.2% (24085/296545)** of the customers do not have payment difficulties

Analysis on Application Data (2/2)

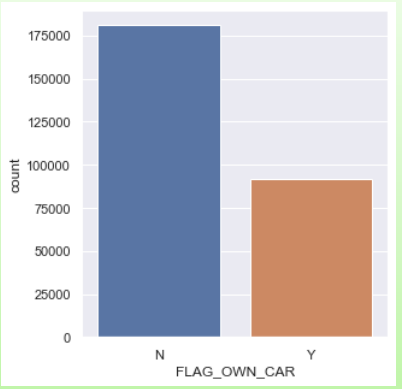
upGrad

0 – Where customers had **no difficulty**

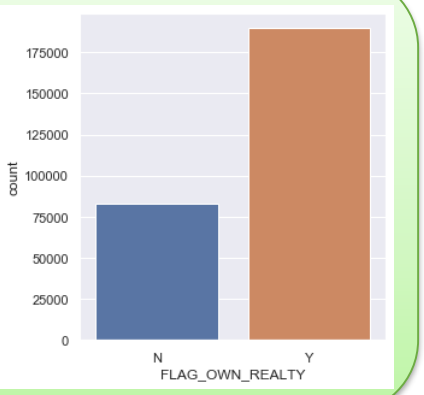


No difference

Number of females is higher in both cases but in ‘0’ the proportion is more while in ‘1’ it is almost equally distributed

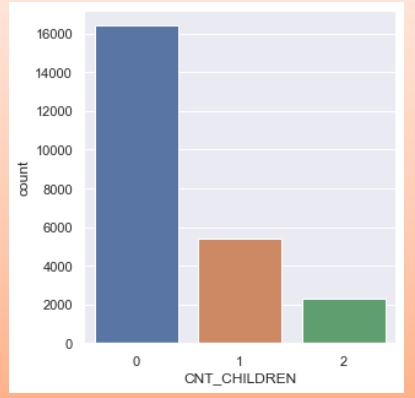


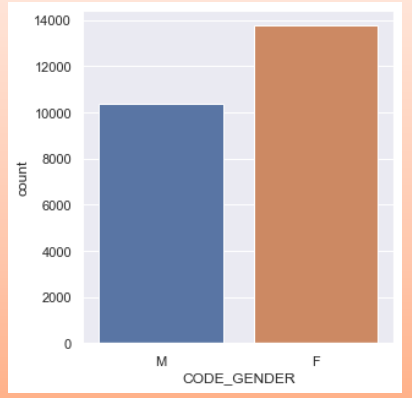
‘0’ have larger proportion of people who own the car

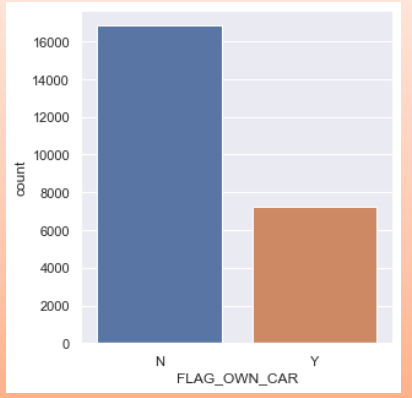


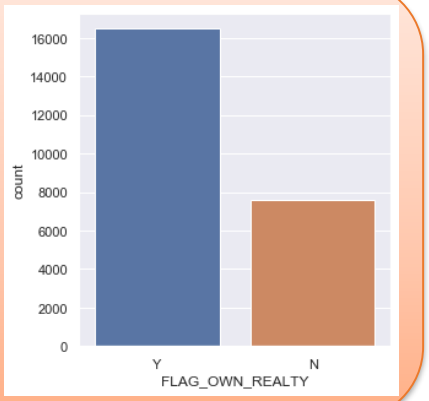
No difference

1 – Where customers had **difficulty**



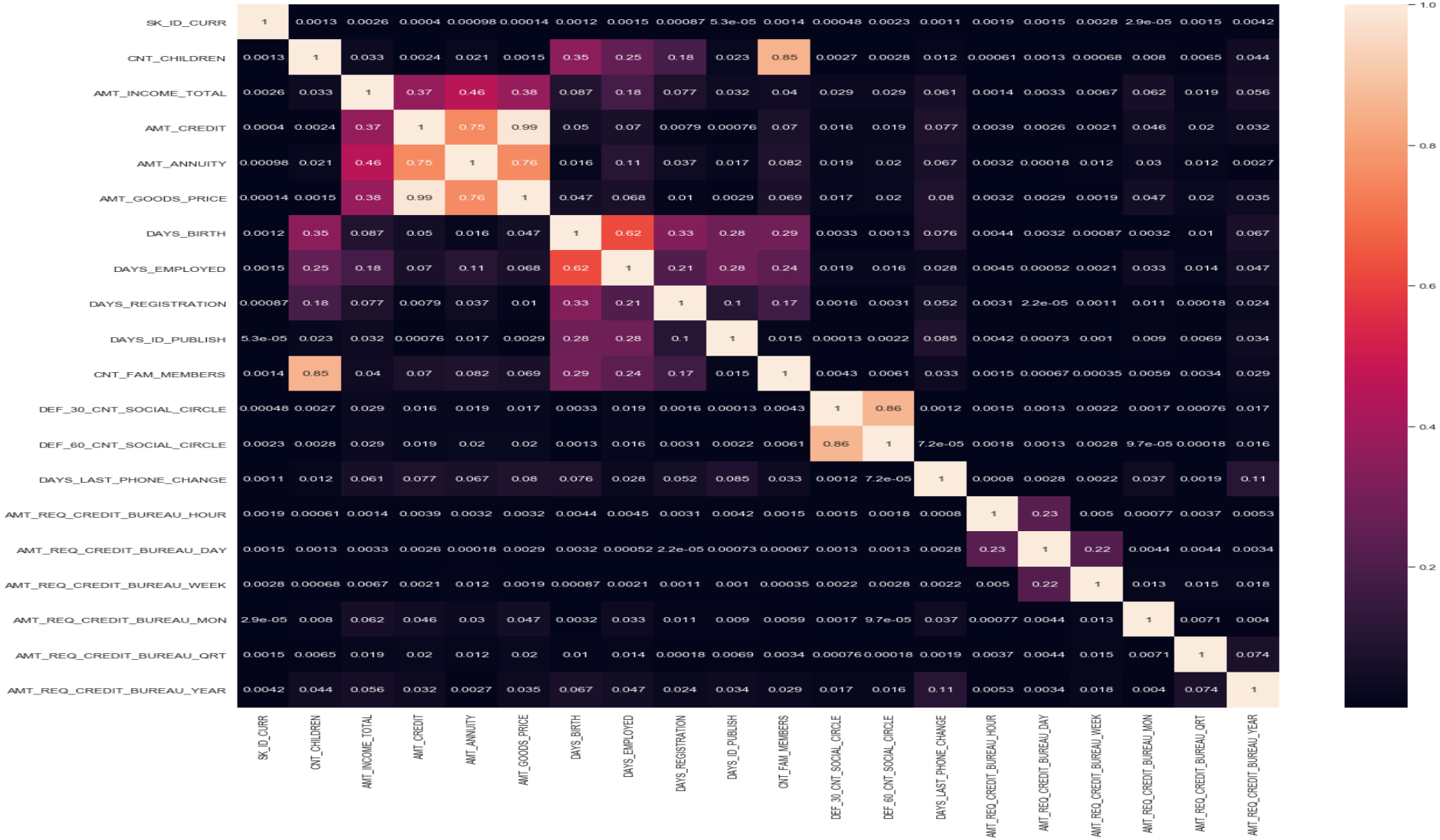






Correlation of ‘0’ Safe Customers

upGrad

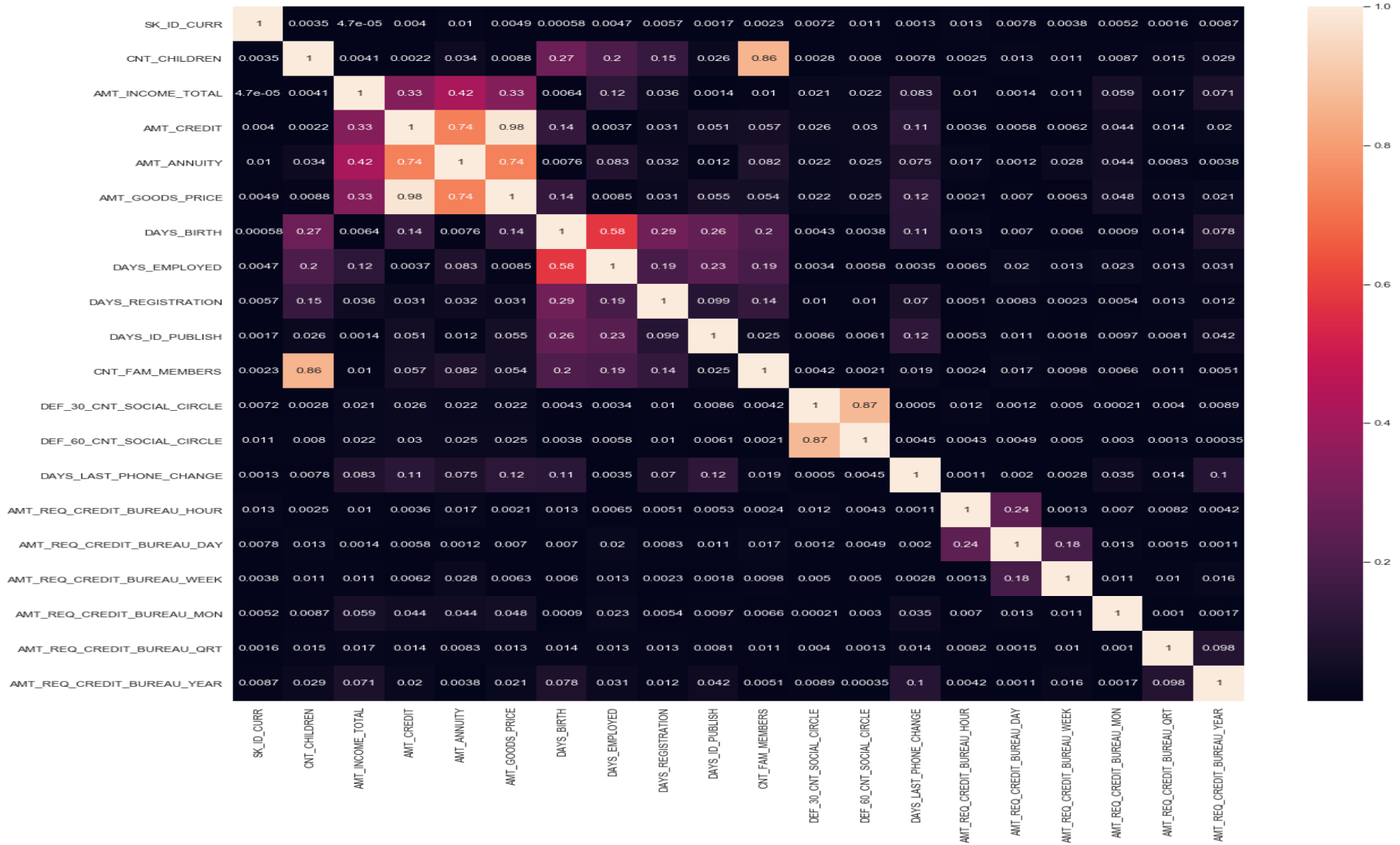


Top 10

1. Amount of goods price and amount of credit
2. DEF_30_CNT_Social_Circle and DEF_60_CNT_Social_Circle
3. Count of family members and count of children
4. Amount of goods price and amount of annuity
5. Amount of annuity and amount of Credit
6. Days Employed and Days Birth
7. Amount of annuity and amount of income total
8. Amount of goods price and amount of income total
9. Amount of credit and amount of income total
10. Days of birth and count of children

Correlation of ‘1’ Non-Safe Customers

upGrad



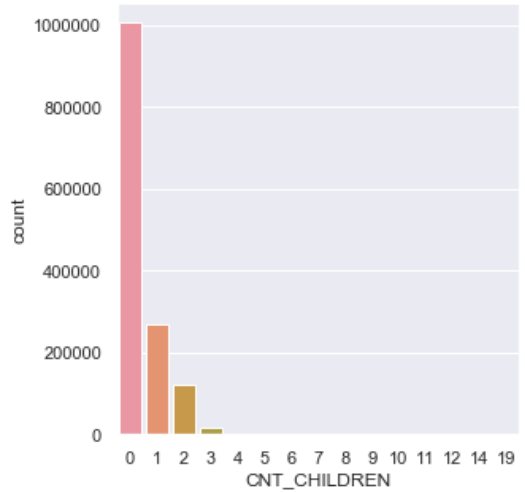
Top 10

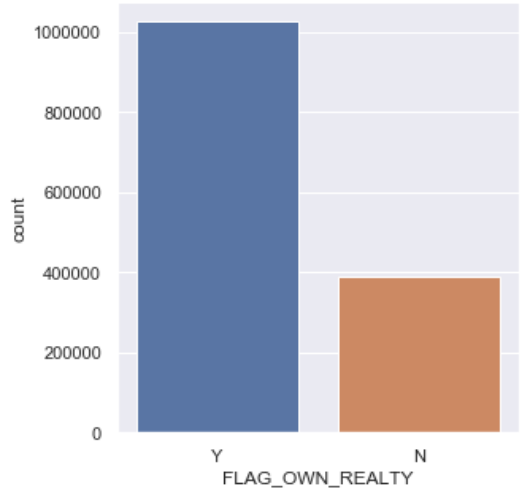
1. Amount of goods price and amount of credit
2. DEF_30_CNT_Social_Circle and DEF_60_CNT_Social_Circle
3. Count of family members and count of children
4. Amount of annuity and amount of Credit
5. Amount of goods price and amount of annuity
6. Days Employed and Days Birth
7. Amount of annuity and amount of income total
8. Amount of goods price and amount of income total
9. Amount of credit and amount of income total
10. **Days of Registration and Days of Birth – This combination was not in top 10 of ‘Safe’ customers**

Merging Data Sets

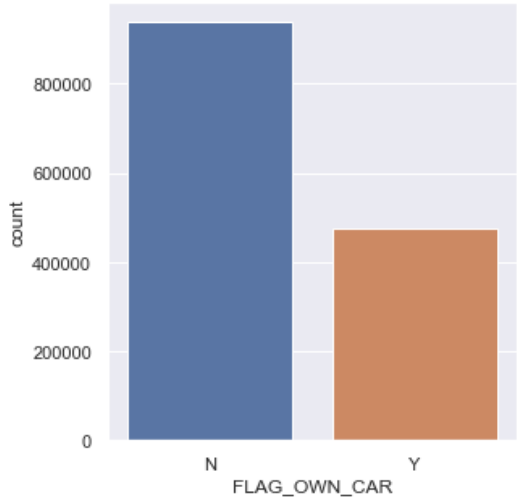
upGrad

Inner Join Used to merge both the datasets

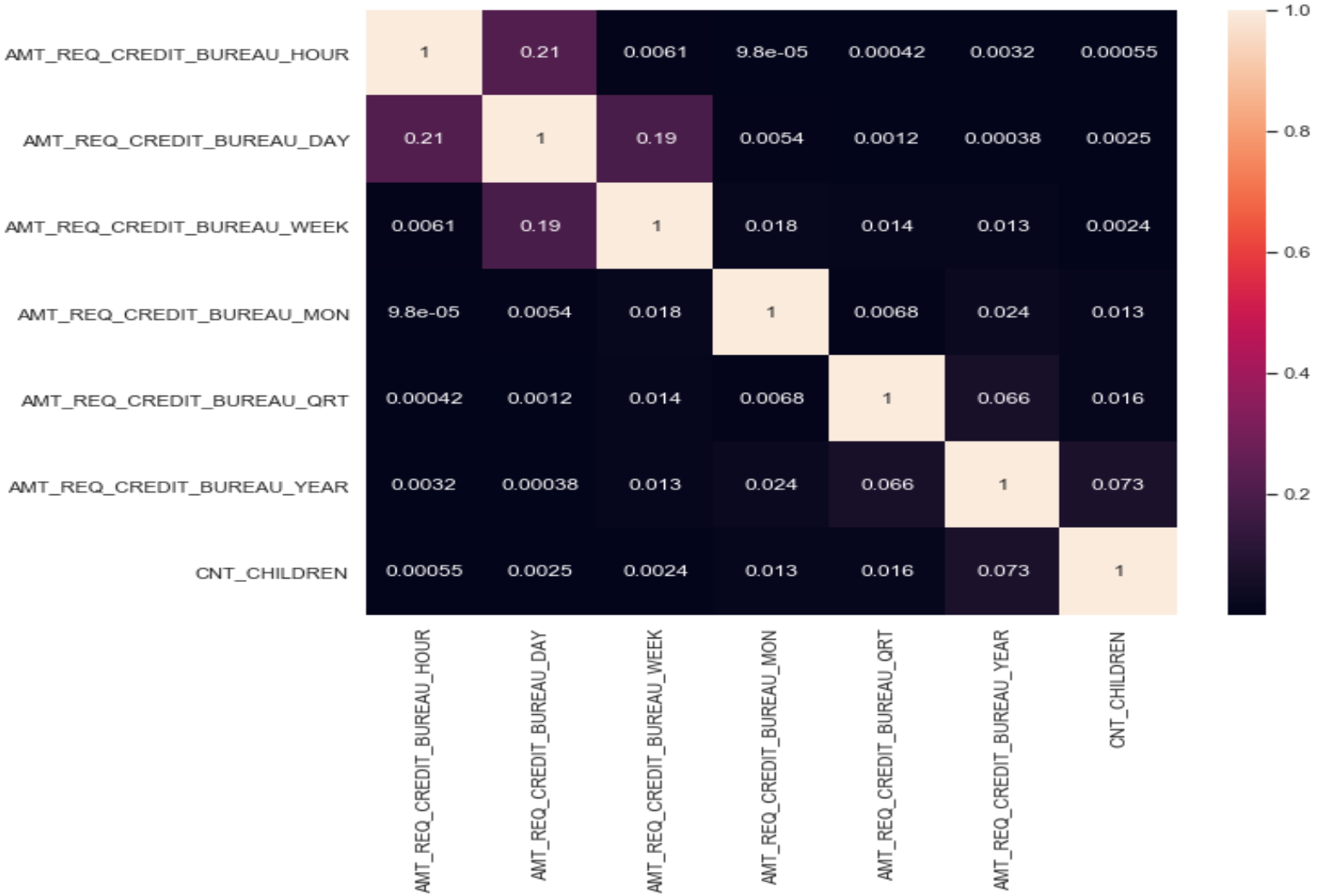




There **is no difference between the categorical variables** of the merged data. They show the same picture as applications data did



Correlation on Merged Dataset



1. AMT_REQ_Credit_Bureau_Day and AMT_REQ_Credit_Bureau_Hour are most correlated
2. AMT_REQ_Credit_Bureau_Week and AMT_REQ_Credit_Bureau_Day are second most correlated

Conclusion

upGrad

- Number of females is higher in both cases but in ‘0’ the proportion is more while in ‘1’ it is almost equally distributed
- ‘0’ have larger proportion of people who own the car
- Correlation matrix shows that variables correlated for both the cases are almost same
- The only exception is Days of Registration and Days of Birth which is present in Non-Safe customers ‘1’
- Merged file also shows the same trend for categorical variables
- No major findings can be drawn from correlation matrix of merged files