BUSINESS INTELLIGENCE REPORT

Project #7: Fraud Detection

Instructor: Dr. Nguyen Binh Minh

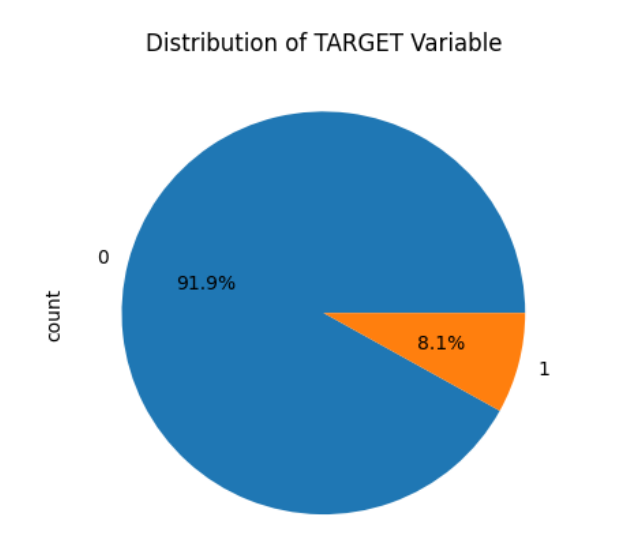
Student: Le Hoang Long - 20232099M

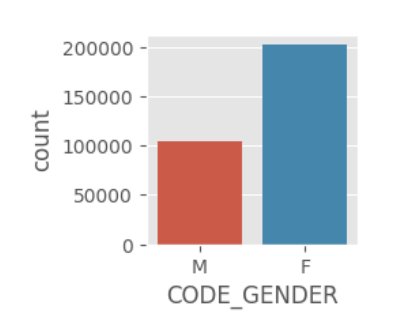
# 

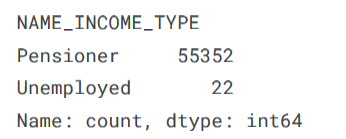
# **Abstract**

* 1. Fraud is a widespread problem in many industries, like banking, sales, and insurance.
  2. The most common form of fraudulent activity is credit card fraud, but there are others, such as identity theft or cyber-attack.
  3. This problem is especially challenging because fraudsters' strategies are constantly adapting and becoming more sophisticated.
  4. This means that there is no one-size-fits-all solution for detecting fraud.

1. **Dataset**
   1. Credit Card Fraud Detection
   2. Explore all possibilities while sanctioning a Loan to any customer
   3. <https://www.kaggle.com/datasets/mishra5001/credit-card>

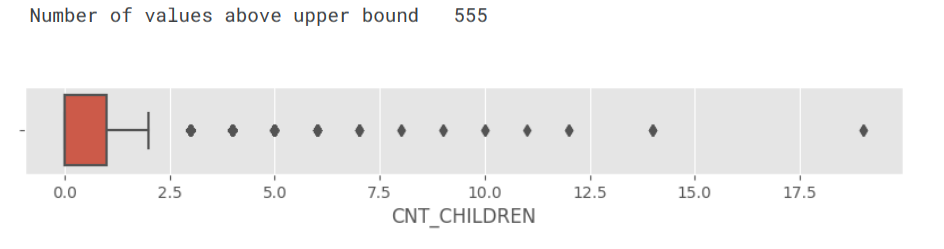


1. **Cleaning Data**
   1. Data shape
      1. 307511 rows
      2. 122 columns
   2. There are 67 columns which have null values
   3. We note that “Occupation Type” plays an important role in predicting whether a person will fail to pay the loan “default”
   4. Removing columns that have more than 32% missing value is a better approach.
   5. List of removed columns:  
      Index(['OWN\_CAR\_AGE', 'EXT\_SOURCE\_1', 'APARTMENTS\_AVG', 'BASEMENTAREA\_AVG',  
       'YEARS\_BEGINEXPLUATATION\_AVG', 'YEARS\_BUILD\_AVG', 'COMMONAREA\_AVG',  
       'ELEVATORS\_AVG', 'ENTRANCES\_AVG', 'FLOORSMAX\_AVG', 'FLOORSMIN\_AVG',  
       'LANDAREA\_AVG', 'LIVINGAPARTMENTS\_AVG', 'LIVINGAREA\_AVG',  
       'NONLIVINGAPARTMENTS\_AVG', 'NONLIVINGAREA\_AVG', 'APARTMENTS\_MODE',  
       'BASEMENTAREA\_MODE', 'YEARS\_BEGINEXPLUATATION\_MODE', 'YEARS\_BUILD\_MODE',  
       'COMMONAREA\_MODE', 'ELEVATORS\_MODE', 'ENTRANCES\_MODE', 'FLOORSMAX\_MODE',  
       'FLOORSMIN\_MODE', 'LANDAREA\_MODE', 'LIVINGAPARTMENTS\_MODE',  
       'LIVINGAREA\_MODE', 'NONLIVINGAPARTMENTS\_MODE', 'NONLIVINGAREA\_MODE',  
       'APARTMENTS\_MEDI', 'BASEMENTAREA\_MEDI',  
      'YEARS\_BEGINEXPLUATATION\_MEDI',  
       'YEARS\_BUILD\_MEDI', 'COMMONAREA\_MEDI', 'ELEVATORS\_MEDI',  
       'ENTRANCES\_MEDI', 'FLOORSMAX\_MEDI',],
2. **Handling Errors in Data**
   1. We note that some columns have negative values of day such as
      1. DAYS\_BIRTH, DAYS\_EMPLOYED
      2. DAYS\_REGISTRATION, DAYS\_ID\_PUBLISH
      3. DAYS\_LAST\_PHONE\_CHANGE
   2. These columns should be converted to year and changed names to
      1. YEARS\_BIRTH, YEARS\_EMPLOYED
      2. YEARS\_REGISTRATION, YEARS\_ID\_PUBLISH
      3. YEARS\_LAST\_PHONE\_CHANGE
   3. We note that 4 rows in CODE\_GENDER have the value “XNA” and note that there are 105059 male records and 202448 female records in the data set
   4. We represent “XNA” to “F”
   5. We note that there are 55374 “XNA” rows out of 307511 rows of column “ORGANIZATION\_TYPE”.
   6. We cannot remove a large amount of data
   7. We also observe that the values of “NAME\_INCOME\_TYPE” columns of these “XNA” rows are “Pensioner” and “Unemployed”
   8. The number of “Pensioner” rows
   9. is much larger than the number
   10. of “Unemployed” rows
   11. We replace the “XNA“ values of column “ORGANIZATION\_TYPE” with “Pensioner”

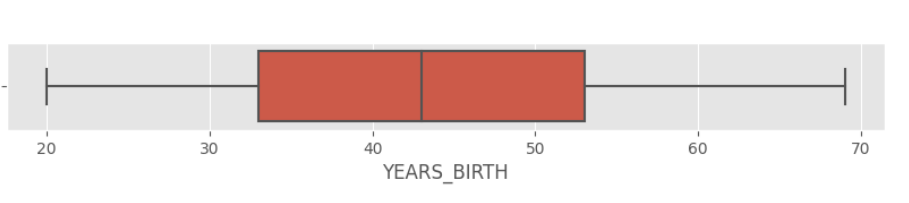


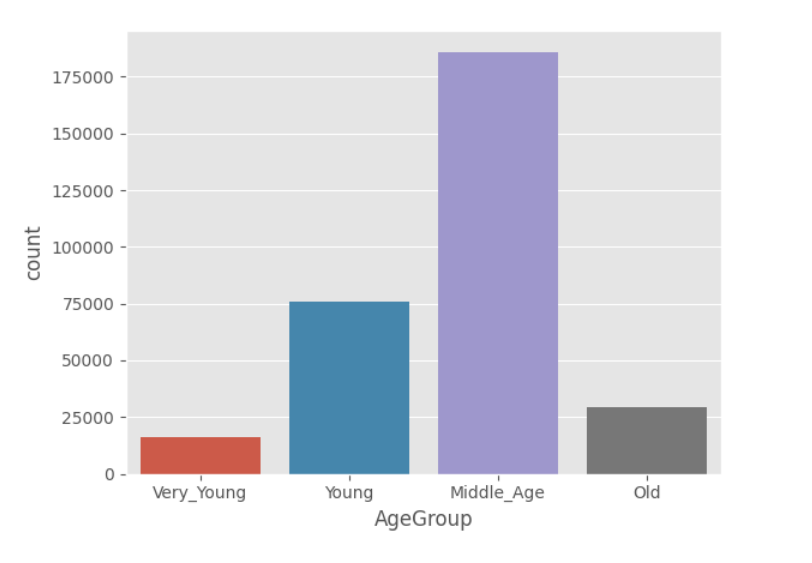
* 1. There are many null values in the dataset, we fill them with 0 if the data type of the column is int64 or float64
  2. We replace the null value of column “NAME\_TYPE\_SUITE” with “Unaccompanied” for this column tells us about the person who was accompanying the client when he was applying for the loan
  3. We replace the null value of column “OCCUPATION\_TYPE” with “Missing” for this column tells us about the kind of occupation that the client has

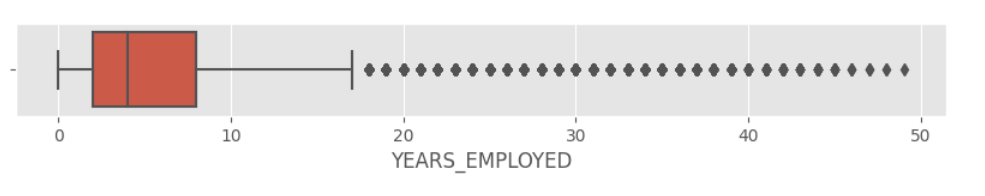
1. **Identifying outliers**
   1. CNT\_CHILDREN
      1. Some people have more than 3 children
      2. 99% of people have less than 4 children



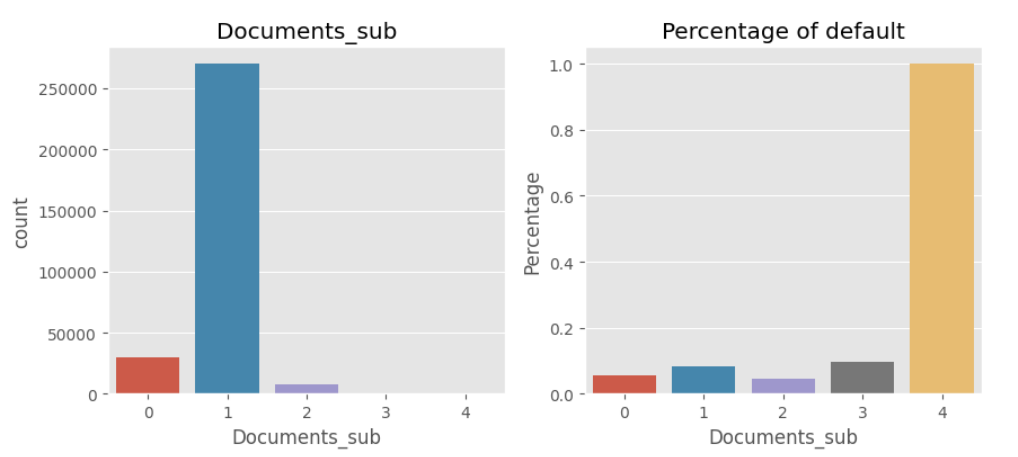
* 1. YEARS\_BIRTH
     1. The value of age spreads from 20 to 69



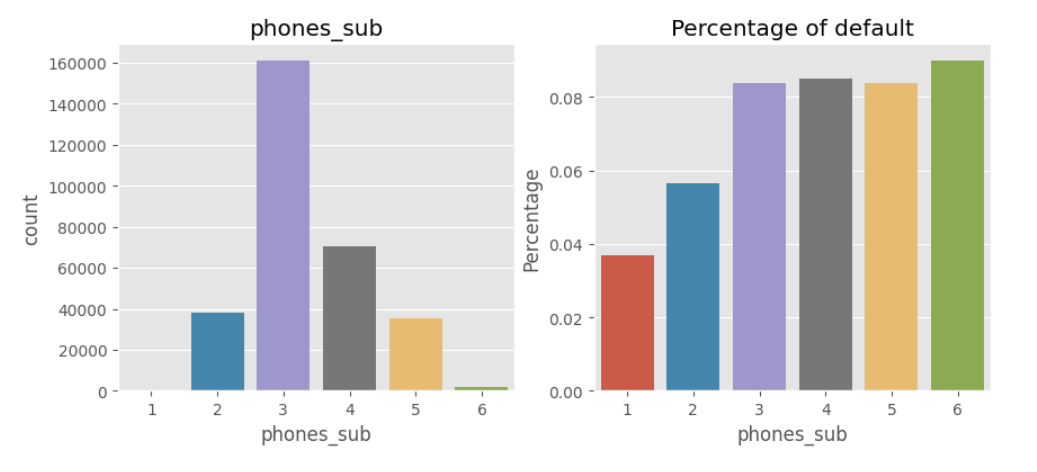
* + 1. We put them into
       1. 4 groups
       2. Very young
       3. Young
       4. Middle age
       5. Old
  1. YEARS\_EMPLOYED
     1. Pensioner and unemployed have default value of 999, they create noises in the dataset,
     2. We set them to NA so that we could ignore them while working on the column
     3. The new range is from 1 year to 50 years



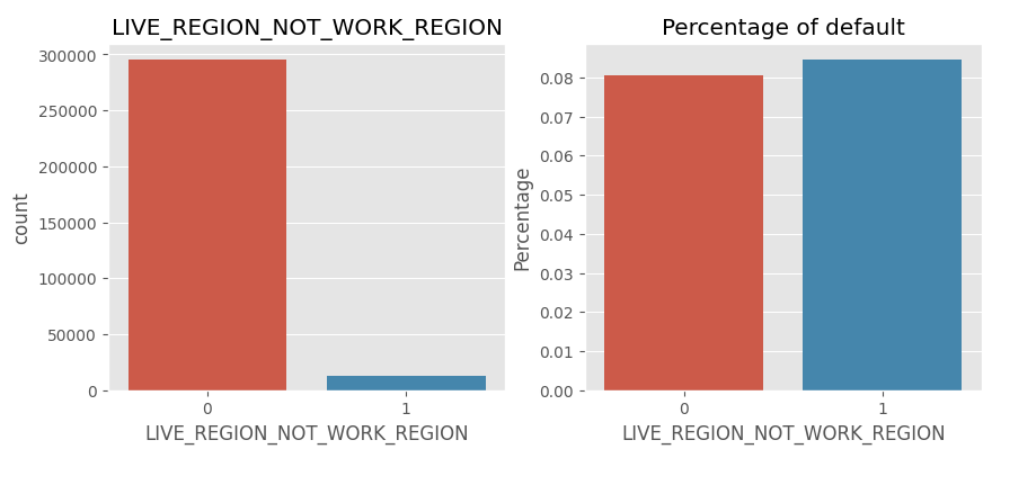
1. **Analysis**
   1. We note that the more documents submitted by a certain client the more risky the loan is



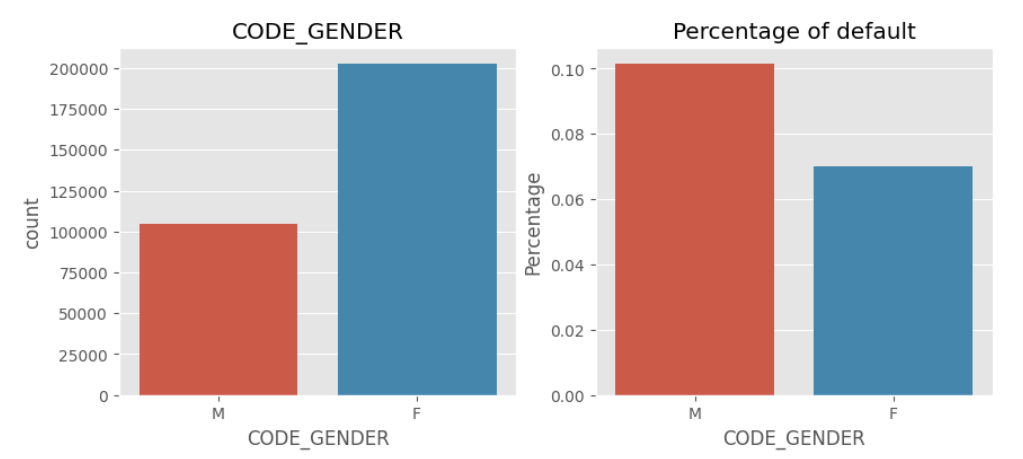
* 1. We note that the more phone numbers provided by a certain client the more risky the loan is



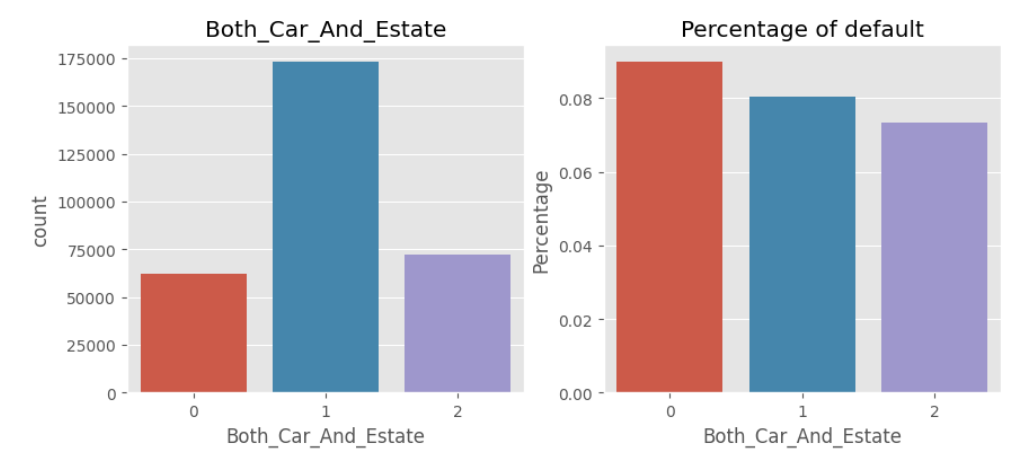
* 1. We note that there are more risky to loan a person who lives far from his workplace. 1 stands for payment difficulties



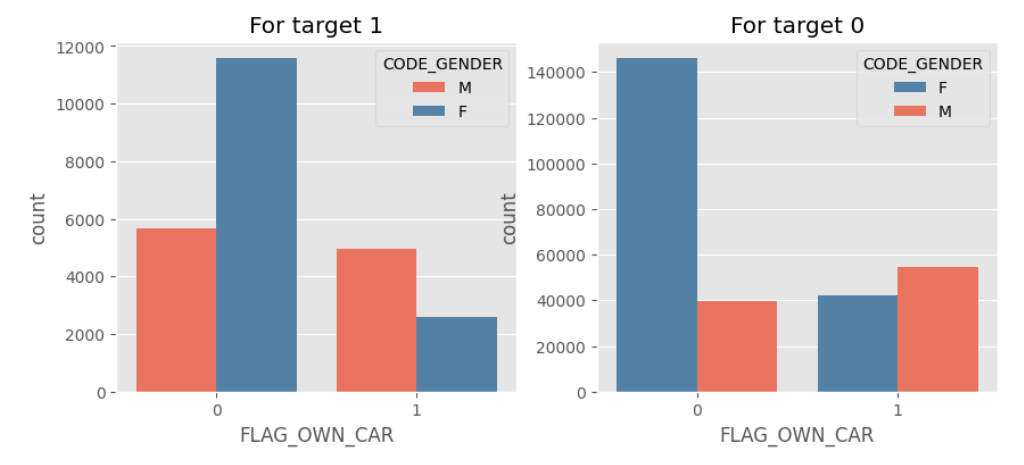
* 1. We note that female clients have more loans than male clients do but males client has a higher percentage of payment difficulties



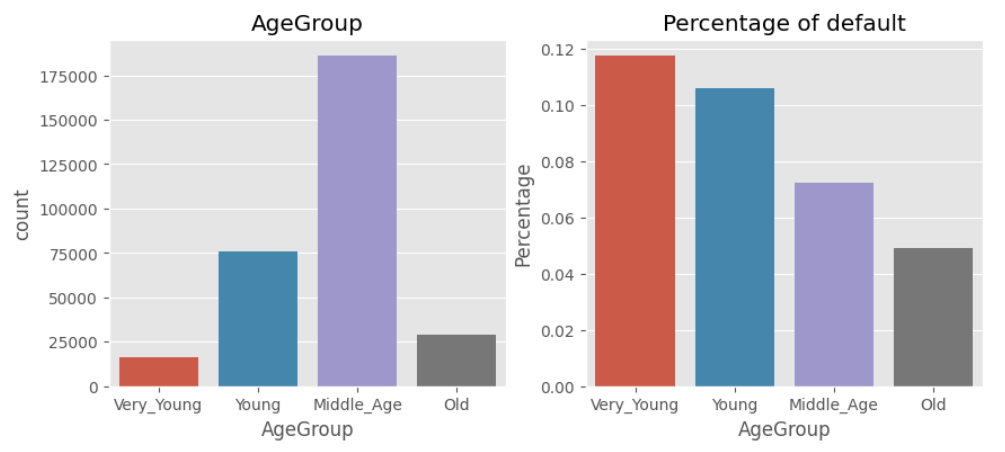
* 1. We note that clients who don’t have a car or any estate have more chance to face payment difficulties



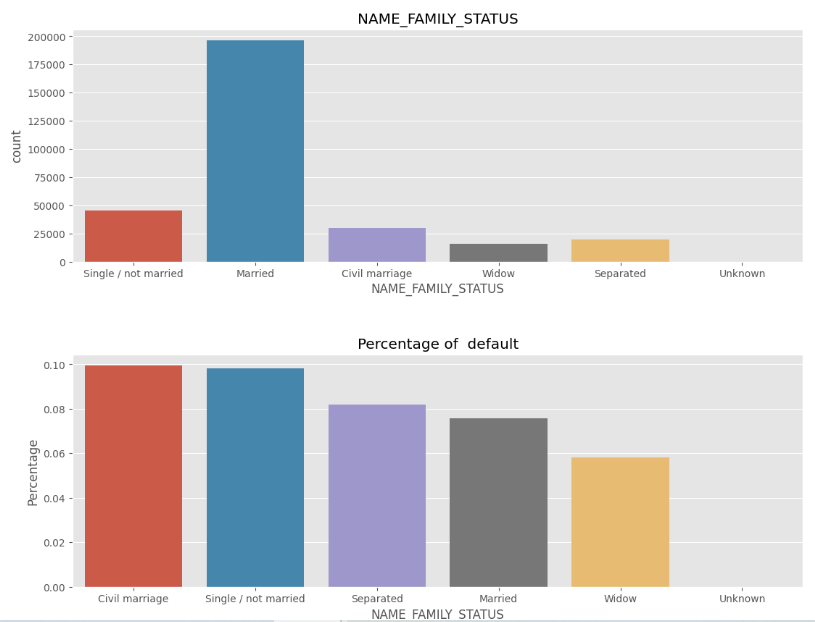
* 1. We note that female clients who don’t have a car have more chance to face payment difficulties than female clients, who have a car, and male clients in both cases



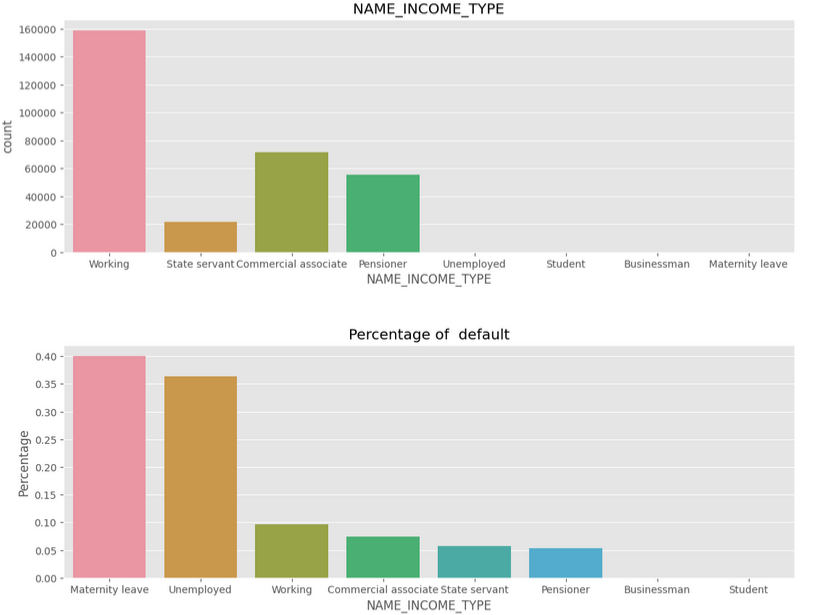
* 1. We note that the middle age group needs more loans than others do. Younger people face more payment difficulties than older people do



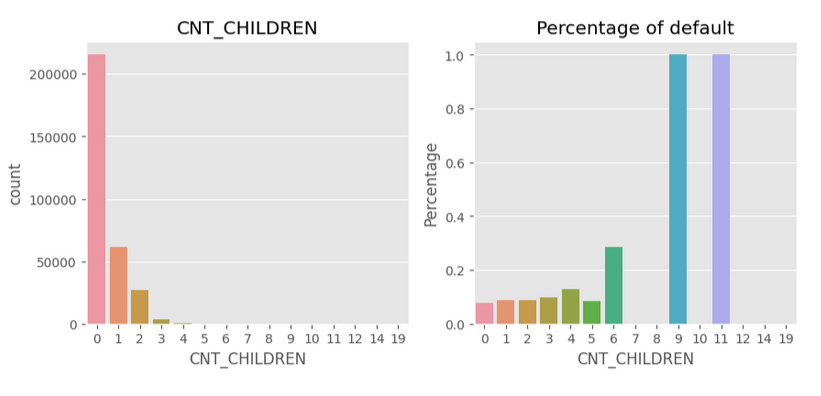
* 1. We note that most clients are married people
  2. Civil marriage and single people tend to have difficulty paying their loan



* 1. We note that most clients belong to working groups but most of them could pay their debt on time
  2. Group of maternity leave and unemployed have a high percentage of payment difficulties



* 1. We note that clients with 9 or 11 children are risky
  2. while clients, having less than 6 children, have more chance to repay the debt on time.



1. **Implementation**
   1. Kaggle notebook: https://www.kaggle.com/code/lehoanglonglong/hust-business-intelligence-29-11-2024
2. **Conclusion**
   1. Clients who submitted more than 3 documents, registered more than 2 phone numbers or using different addresses tend to high percentage of unsuccessful payments
   2. A client who doesn’t have both a car and a house has more probability of unsuccessful payments
   3. Young clients tend to miss their loan payments more than old clients do
   4. Married clients tend to pay their debt in time while civil married and single clients get trouble while repaying their loans.
   5. Most client of the working group pays the debt on time, while client of the non-working groups has payment difficulties
   6. Providing loans to clients who have a high number of children or family members is not a safe decision.
3. **Suggestion**
   1. We note that people don’t want to get into payment difficulties
   2. We should give clients more chance to pay their debt
   3. Young clients will become old clients, and old clients tend to pay the debt in time
   4. Single people will get married, and it is a natural thing, properly married clients tend to take more loans and pay the debt in time, and separated clients get fewer payment difficulties. We guess that separated clients have broken hearts so they don’t want to have any broken account