Đánh giá điểm tín dụng khách hàng - Credit Scoring

Dự đoán tình trạng vay vốn là vỡ nợ hay không vỡ nợ dựa trên bộ dữ liệu **HMEQ dataset** (kaggle dataset).

1. Exploratory Data Analysis - Phân tích khám phá dữ liệu

1.1. Kích thước dữ liệu

Bộ dữ liệu bao gồm 5960 mẫu dữ liệu (bản ghi) và 13 trường dữ liệu

1.2. Ý nghĩa các trường dữ liệu

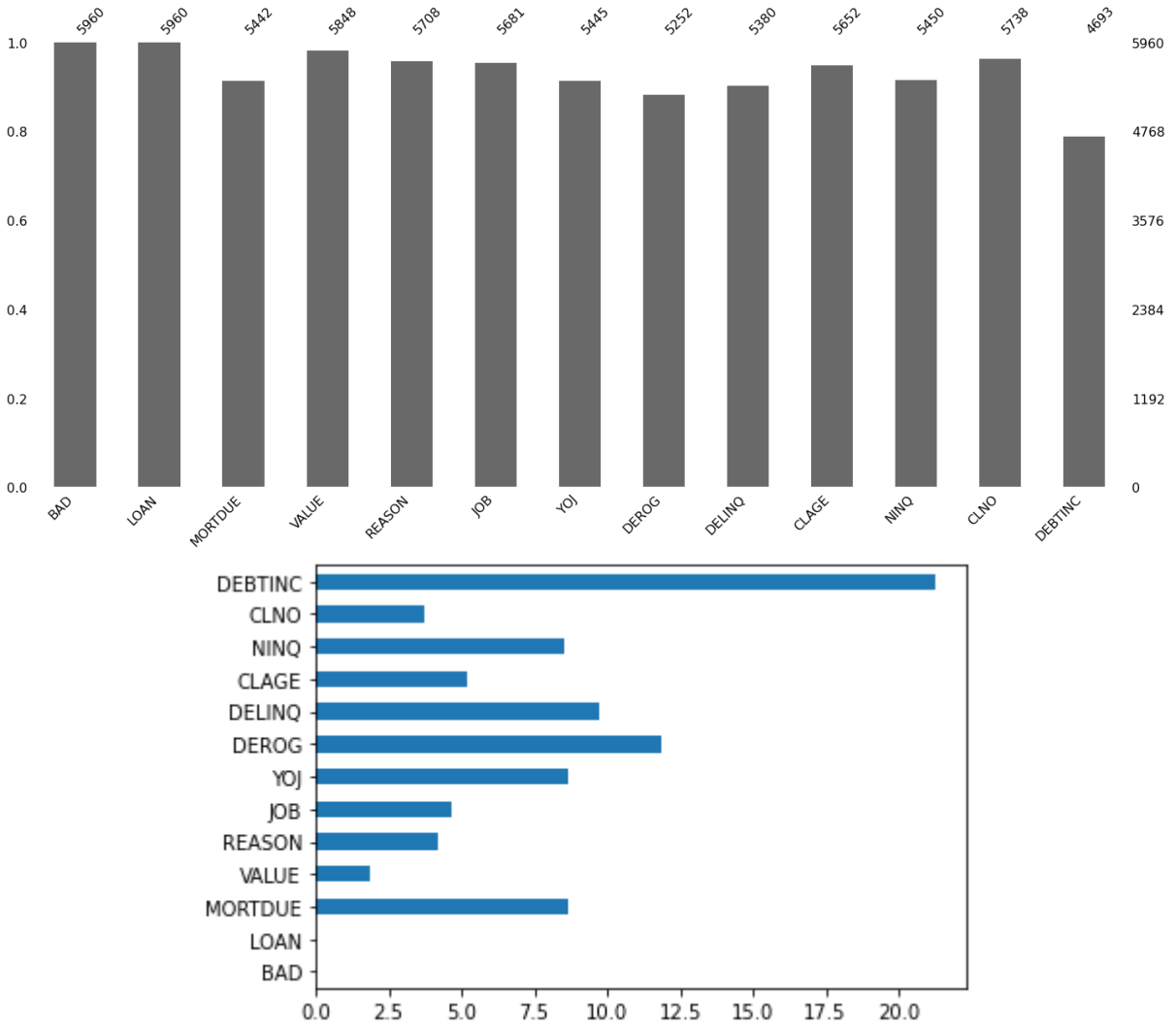
Bảng: mô tả các trường dữ liệu

Column Name	Description	Data Type
loan	khoản vay	numerical - int
mortdue	số tiền đến hạn trên khoản thế chấp hiện có	numerical - int
value	tổng tài sản hiện có	numerical - int
reason	lý do vay vốn	categorical
job	nghề nghiệp	categorical
yoj	số năm làm việc	numerical - int
derog	số lượng vi phạm tiêu cực tín dụng	numerical - int
delinq	số hạn mức quy định quá hạn	numerical - int
clage	số tuổi hạn mức tín dụng cũ nhất tính theo tháng	numerical - double
ninq	số lượng yêu cầu tín dụng gần đây	numerical -int
clno	số hạn mức tín dụng hiện có	numerical- int
debtinc	tỉ lệ nợ trên thu nhập	nujmerical - double (%)
bad (target variable)	rủi ro tín dụng (vỡ nợ, không vỡ nợ)	categorical (0/1)

	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
0	1	1100	25860.0	39025.0	HomeImp	Other	10.5	0.0	0.0	94.366667	1.0	9.0	NaN
1	1	1300	70053.0	68400.0	HomeImp	Other	7.0	0.0	2.0	121.833333	0.0	14.0	NaN
2	1	1500	13500.0	16700.0	HomeImp	Other	4.0	0.0	0.0	149.466667	1.0	10.0	NaN
3	1	1500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	0	1700	97800.0	112000.0	HomeImp	Office	3.0	0.0	0.0	93.333333	0.0	14.0	NaN

Hình: một số mẫu dữ liệu

1.3 Kiểm tra trường dữ liệu bị khuyết và dữ liệu trùng lặp



Hình: phần trăm số lượng giá trị bị khuyết trên mỗi trường.

Ngoài hai trường dữ liệu là *bad* và *loan* có đầy đủ dữ liệu thì 11 trường dữ liệu còn lại bị thiếu. Các trường có số lượng bị khuyết dưới 30%. có thể thay thế (impute) bằng các giá trị trung vị, trung bình hoặc hạng mục xuất hiện nhiều nhất.

Dữ liệu không có bản ghi bị trùng lặp.

1.4. Phân phối của từng trường dữ liệu

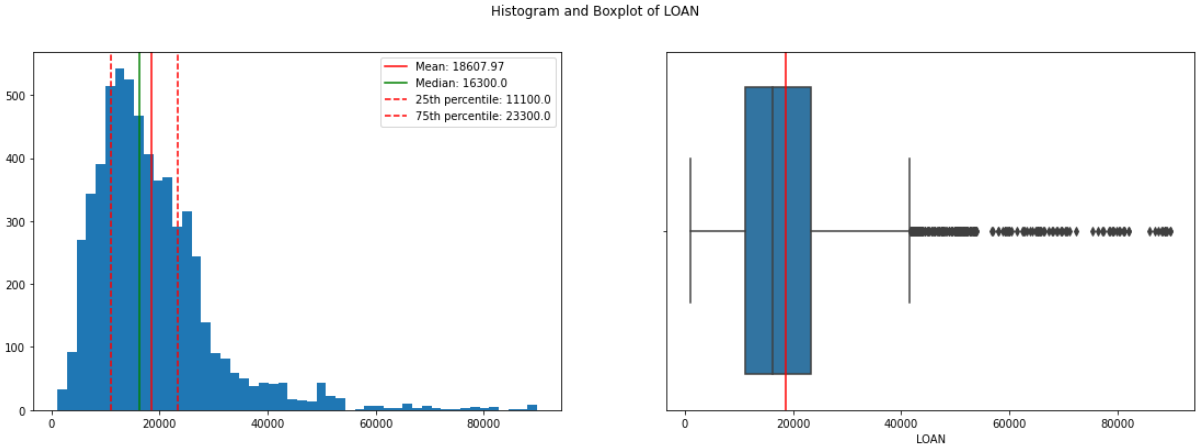
Các trường dạng số bao gồm *'LOAN'*, *'DEBTINC'*, *'DELINQ'*, *'MORTDUE'*, *'YOJ'*, *'CLNO'*, *'DEROG'*, *'CLAGE'*, *'NINQ'*, *'VALUE'*

Các trường dạng hạng mục bao gồm *'REASON'*, *'JOB'*, *'BAD'*

Trong đó trường *BAD* là cột nhãn dự đoán.

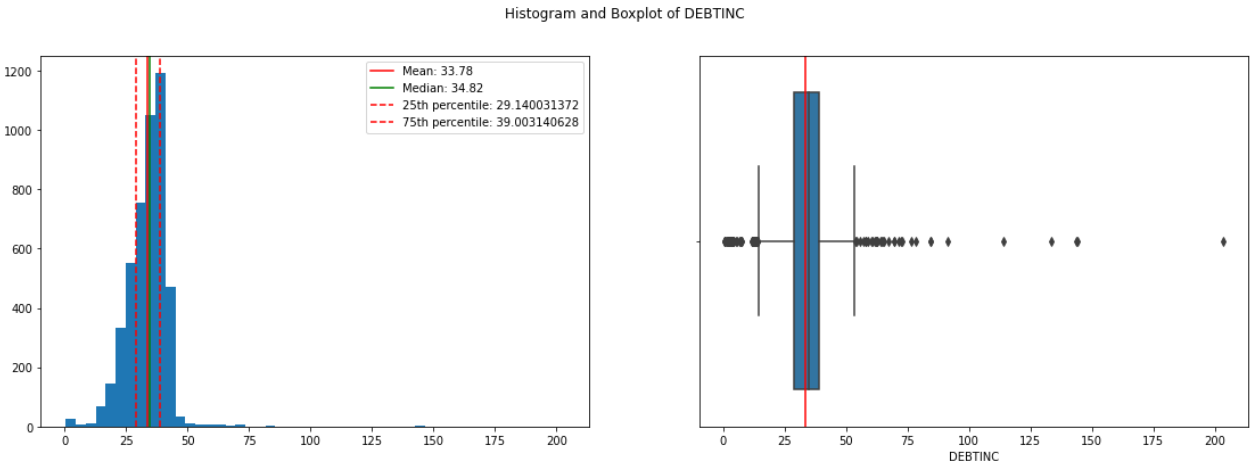
Các trường dạng số:

- *loan* - khoản vay



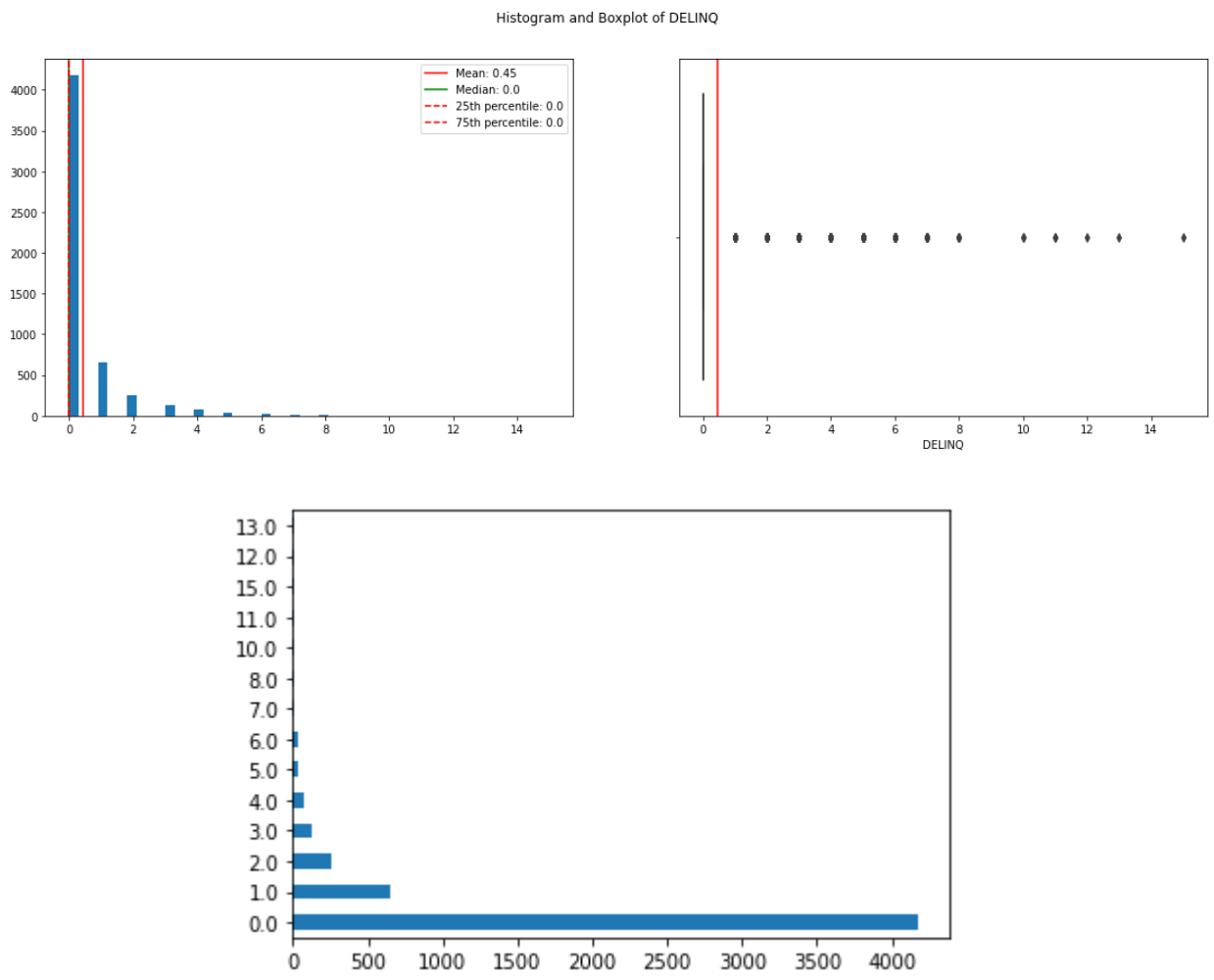
Nhận xét: Từ histogram cho thấy dữ liệu phân bố hầu hết trong khoảng từ 11100 đến 23300. Đồ thị hơi lệch phải (right skewed). phân phối khá gần phân phối chuẩn với skewness gần bằng 0, trung bình gần trung vị.

- *debtinc* - tỷ lệ nợ trên thu nhập



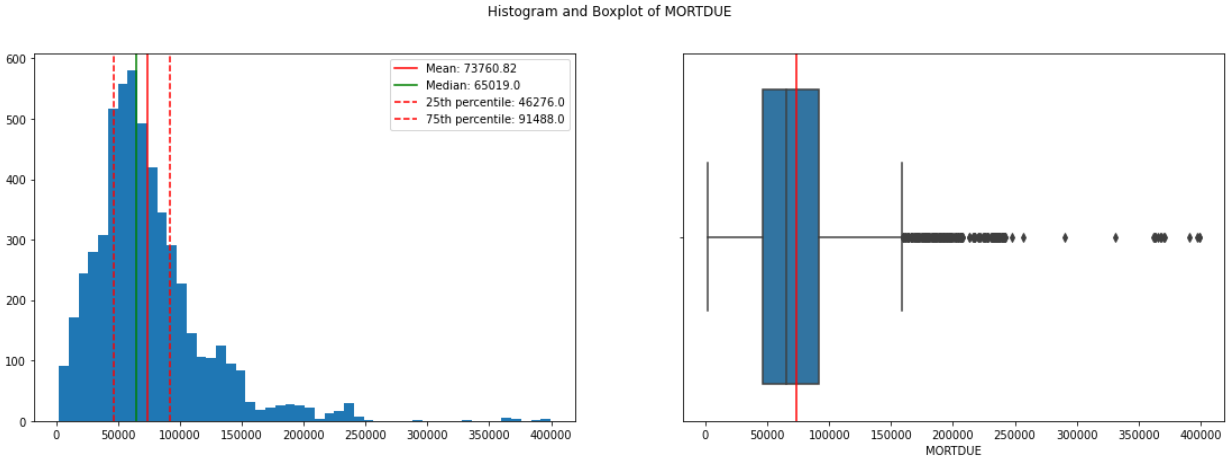
Nhận xét: Dữ liệu hầu hết phân bố trong khoảng 29 đến 39%, trong đó có một vài mẫu dữ liệu là tỷ lệ nợ trên thu nhập lớn hơn 100%, đặc biệt có trường hợp lớn hơn 200%.

- *delinq* - số hạn mức quy định quá hạn



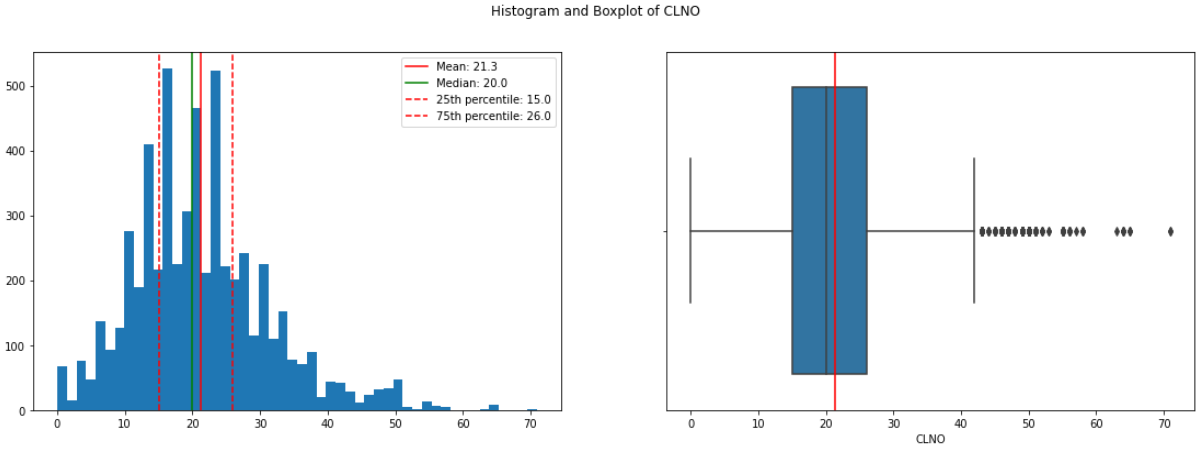
Nhận xét: dữ liệu phân bố rời rạc tại các giá trị cụ thể từ 1 đến 15. với hầu hết các trường hợp có giá trị 0 chiếm tới 70%. có thể coi số hạn mức quy định quá hạn là dạng dữ liệu hạng mục, với hai giá trị 0 và 1 là các giá trị còn lại.

- *mortdue* - số tiền đến hạn trên khoản thế chấp hiện có



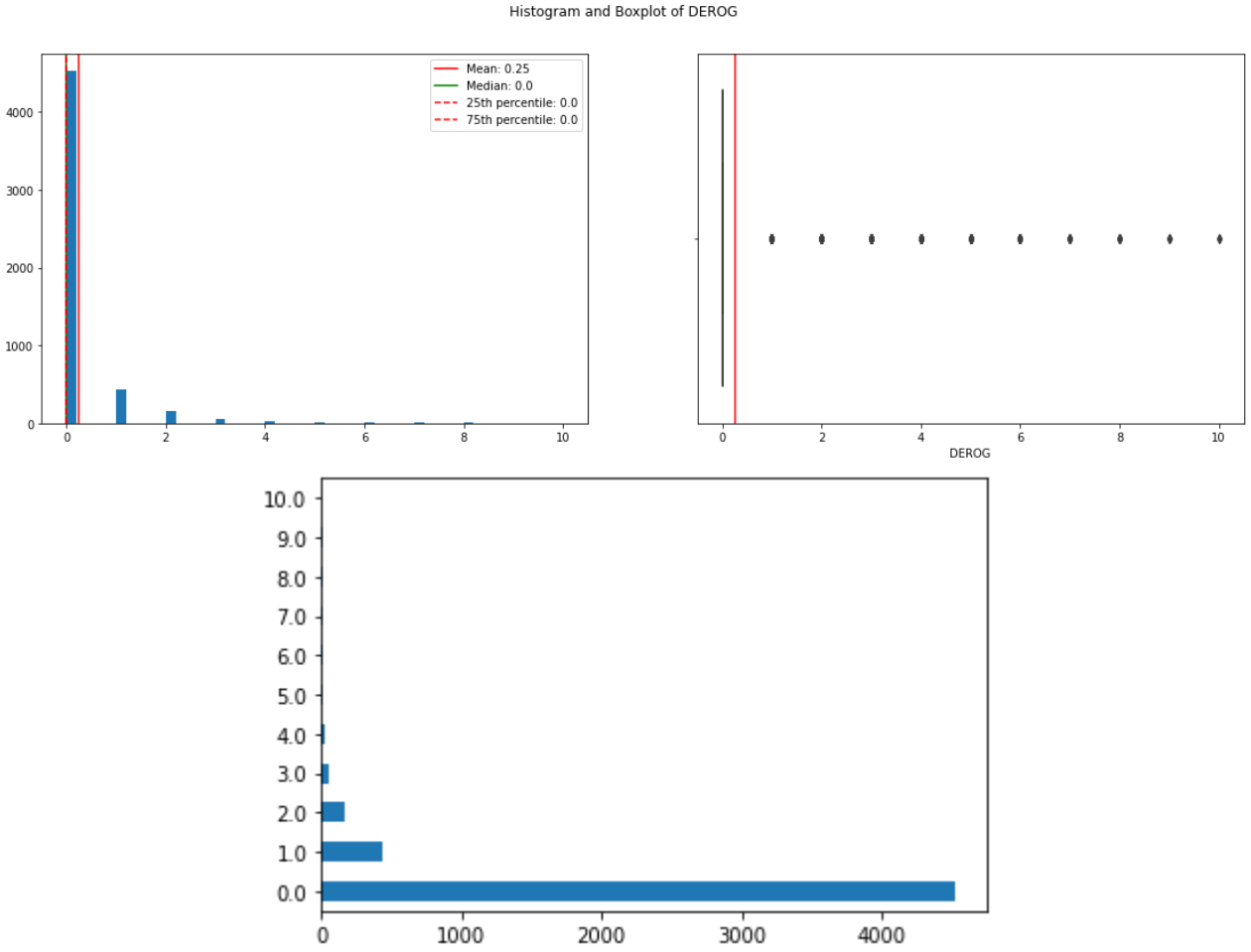
Nhận xét: dữ liệu hầu hết tập trung trong khoảng 46276 và 91488. một vài trường hợp nhiễu có giá trị gần 400000.

- *clno* - số hạn mức tín dụng hiện có



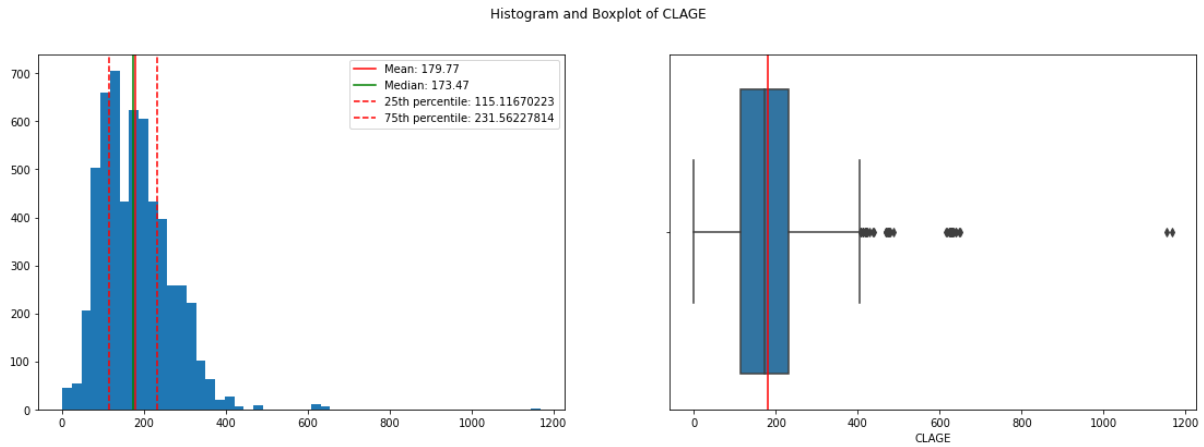
Nhận xét: dữ liệu gần phân phối chuẩn, với skewness gần bằng 0. tập trung hầu hết tại các giá trị 15-26.

- *derog* - số lượng vi phạm tiêu cực tín dụng



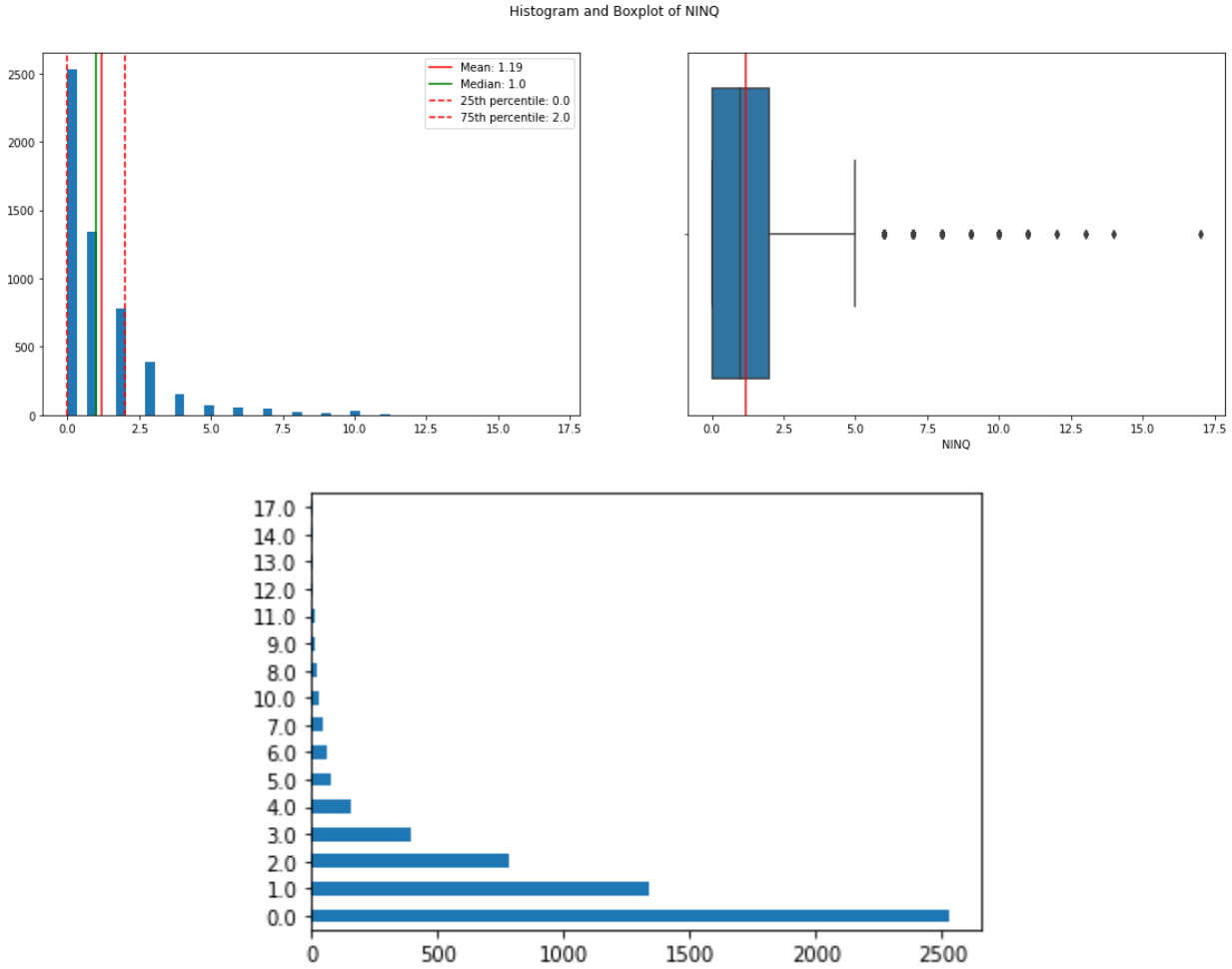
Nhận xét: dữ liệu có các giá trị rời rạc từ giá trị cụ thể 1-10. tương tự như trường *delinq* - số hạn mức quy định quá hạn, có số lượng giá trị 0 hơn 4000 chiếm 75% tổng số giá trị trong cột. có thể biến đổi thành dạng giá trị hạng mục với giá trị 0 và giá trị 1 là các giá trị khác 0.

- *clage* - số tuổi hạn mức tín dụng cũ nhất tính theo tháng



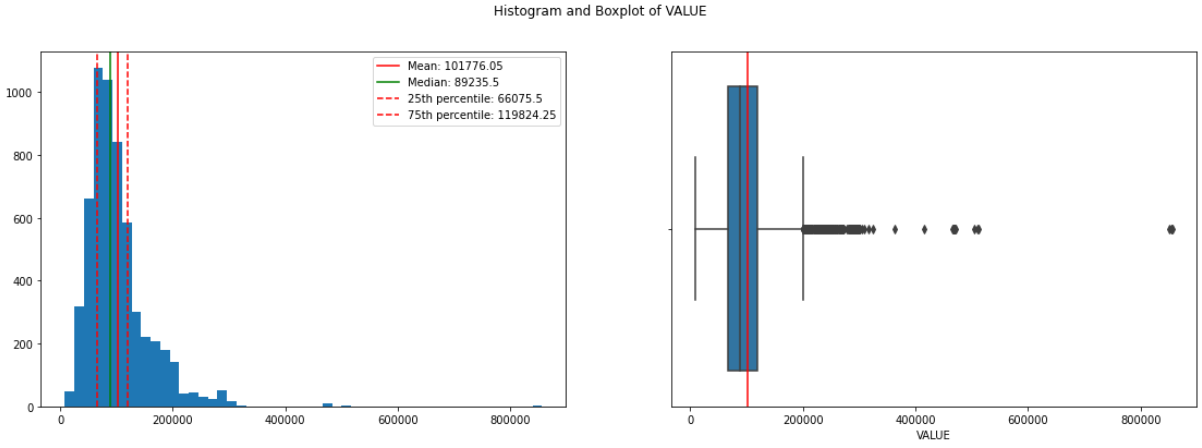
Nhận xét: hầu hết các giá trị tập trung khoảng 115 và 231, số ít có giá trị lớn hơn 600. có thể là nhiễu.

- *ninq* - số lượng yêu cầu tín dụng gần đây



Nhận xét: dữ liệu là các giá trị cụ thể rời rạc, với số lượng tập trung chính vào năm giá trị đầu tiên 0-4. có thể chuyển thành dữ liệu hạng mục với 6 giá trị, 5 giá trị từ 0-4 và 1 giá trị là 5 bao gồm các giá trị còn lại > 4.

- *value* - tổng tài sản hiện có

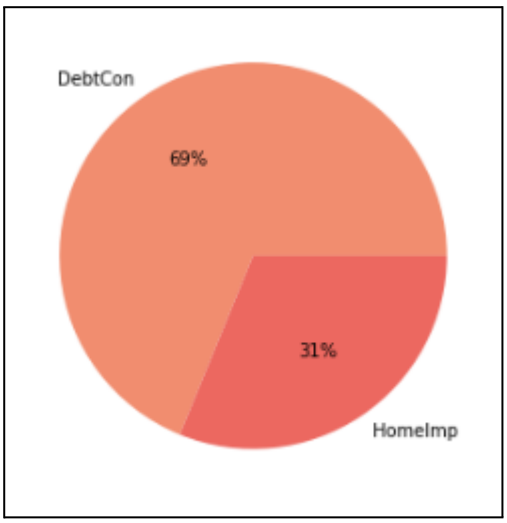


Nhận xét: dữ liệu tập trung trong khoảng 66075 và 119824. số ít có giá trị lớn hơn 400000 có thể cắt về giá trị lớn nhất.

Các trường dạng hạng mục:

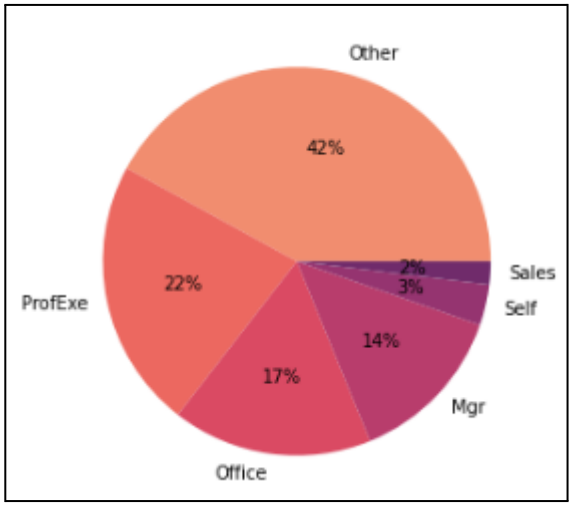
- *reason* - lý do vay vốn

	REASON	Frequencies	% Percentages
0	DebtCon	3928	68.82
1	HomeImp	1780	31.18



- *job* - nghề nghiệp của khách hàng

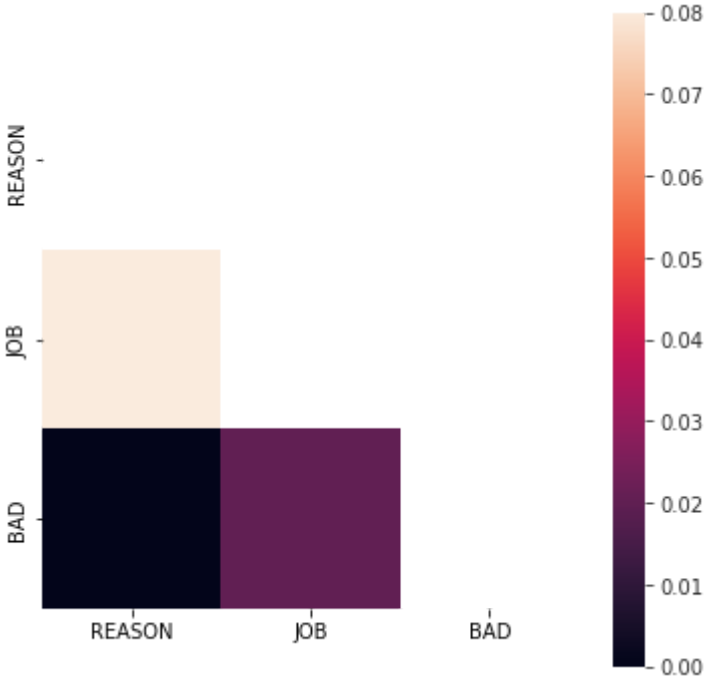
	JOB	Frequencies	% Percentages
0	Other	2388	42.03
1	ProfExe	1276	22.46
2	Office	948	16.69
3	Mgr	767	13.50
4	Self	193	3.40
5	Sales	109	1.92



1.5 Mối tương quan và liên hệ giữa các biến

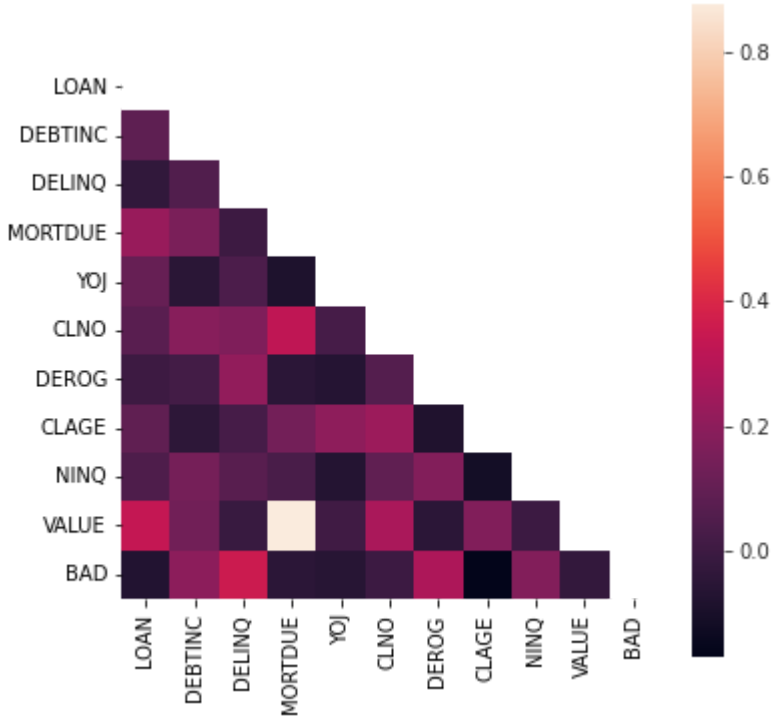
1.5.1 giữa các biến dạng thư mục

Sử dụng phép đo Cramer's V, có giá trị trong khoảng 0-1. Không có mối liên hệ hay tương quan giữa các biến khi giá trị này gần bằng 0, có sự tương quan mạnh khi giá trị này gần bằng 1.



1.5.2 giữa các biến dạng số

Sử dụng phép đo tương quan Pearson để đo mối liên hệ tuyến tính giữa hai biến dạng số. Có giá trị trong khoảng [-1, 1]. Không có mối tương quan khi giá trị gần bằng 0, có mối tương quan dương mạnh khi giá trị gần băng 1, có mối tương quan âm mạnh khi giá trị gần bằng -1.



Nhận xét:

- Các cặp trường dữ liệu MORTDUE-CLNO, CLAGE-CLNO, VALUE-CLNO, DELINQ-DEROG, CLAGE-YOJ,LOAN-VALUE, LOAN-MORTDUE có mối tương quan nhẹ với nhau
- MORTDUE-VALUE có mối tương quan mạnh.
- mối tương quan với biến mục tiêu BAD: BAD-DEROG, DELINQ-BAD có mối tương quan nhẹ.

2. Weight of Evidence - IV (trọng số giới thiệu - chỉ số giá trị thông tin)

Phương pháp xếp hạng các biến có liên quan đến biến nhãn dự đoán, độ quan trọng của biến độc lập với biến mục tiêu dựa vào chỉ số giá trị thông tin IV. Giá trị IV được tính toán thông qua trọng số giới thiệu WOE.
Giá trị IV:

- <= 0.02: Biến không có tác dụng trong việc phân loại hồ sơ Good/Bad
- 0.02 - 0.1: yếu
- 0.1 - 0.3: trung bình
- 0.3 - 0.5: mạnh
- => 0.5: Biến rất mạnh thể hiện mối quan hệ trực tiếp để định nghĩa hồ sơ good/bad.

col name	IV	rank
REASON	0.008618	Useless
MORTDUE	0.051314	Weak
CLNO	0.060437	Weak
YOJ	0.067147	Weak
JOB	0.123731	Medium
VALUE	0.141889	Medium
LOAN	0.160156	Medium
NINQ	0.165559	Medium
CLAGE	0.221710	Medium
DEROG	0.374912	Strong
DELINQ	0.410678	Strong
DEBTINC	1.511125	suspicious

Hình: xếp hạng độ quan trọng của các biến đối với biến mục tiêu BAD

Nhận xét:

- Biến *REASON* không có tác dụng trong việc dự báo khả năng vỡ nợ của khách hàng.
- Các biến *MORTDUE, CLNO, YOJ, JOB, VALUE, LOAN, NINQ, CLAGE* có tác dụng nhẹ.
- Đặc biệt các biến còn lại có tác động mạnh trong việc dự đoán.