



REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

March 27, 2021

Contents

1	Introduction	3
1.1	Information of Dataset	3
1.2	Information of Variables	3
1.3	About EDA Report	4
2	Univariate Analysis	5
2.1	Descriptive Statistics	5
2.2	Normality Test of Numerical Variables	8
2.2.1	Statistics and Visualization of (Sample) Data	8
3	Relationship Between Variables	21
3.1	Correlation Coefficient	21
3.1.1	Correlation Coefficient by Variable Combination	21
3.1.2	Correlation Plot of Numerical Variables	21
4	Target based Analysis	23
4.1	Grouped Descriptive Statistics	23
4.1.1	Grouped Numerical Variables	23
4.1.2	Grouped Categorical Variables	23
4.2	Grouped Relationship Between Variables	23
4.2.1	Grouped Correlation Coefficient	23
4.2.2	Grouped Correlation Plot of Numerical Variables	23

Chapter 1

Introduction

The EDA Report provides exploratory data analysis information on objects that inherit `data.frame` and `data.frame`.

1.1 Information of Dataset

The dataset that generated the EDA Report is an ‘`data.frame`’ object. It consists of 3,000 observations and 21 variables.

1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
tot_credit_debt	numeric	0	0.00	3000	1.000
avg_card_debt	numeric	0	0.00	2967	0.989
credit_age	numeric	0	0.00	323	0.108
credit_good_age	numeric	0	0.00	181	0.060
card_age	numeric	0	0.00	320	0.107
non_mtg_acc_past_due_12_months_num	character	0	0.00	5	0.002
non_mtg_acc_past_due_6_months_num	character	0	0.00	3	0.001
mortgages_past_due_6_months_num	character	0	0.00	2	0.001
credit_past_due_amount	numeric	0	0.00	105	0.035
inq_12_month_num	numeric	0	0.00	9	0.003
card_inq_24_month_num	numeric	0	0.00	14	0.005
card_open_36_month_num	character	0	0.00	3	0.001
auto_open_36_month_num	character	0	0.00	3	0.001
uti_card	numeric	0	0.00	3000	1.000
uti_50plus_pct	numeric	0	0.00	3000	1.000
uti_max_credit_line	numeric	0	0.00	3000	1.000
uti_card_50plus_pct	numeric	297	9.90	2704	0.901
ind_acc_XYZ	character	0	0.00	2	0.001
rep_income	numeric	253	8.43	96	0.032
States	factor	0	0.00	7	0.002
Default_ind	character	0	0.00	2	0.001

The target variable of the data is ‘`Default_ind`’, and the data type of the variable is character.






1.3 About EDA Report

EDA reports provide information and visualization results that support the EDA process. In particular, it provides a variety of information to understand the relationship between the target variable and the rest of the variables of interest.

Chapter 2

Univariate Analysis

2.1 Descriptive Statistics

edaData													
21 Variables 3000 Observations													
tot_credit_debt													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
3000	0	3000	1	89349	23426	55017	63531	75621	88875	103652	116147	122875	
lowest : 13137.29 21346.34 27222.45 28251.19 28684.23													
highest: 155996.77 158140.85 159389.01 159894.98 170237.01													
avg_card_debt													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
3000	0	2967	1	13531	4663	8193	9208	10884	12756	14602	16180	17173	
lowest : 2910.57 3575.93 4402.31 4687.63 4837.03, highest: 18842.72 18847.68 18923.86 18970.23 99999.00													
Value	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	13000	14000	
Frequency	1	2	7	19	51	108	172	261	362	412	438	383	
Proportion	0.000	0.001	0.002	0.006	0.017	0.036	0.057	0.087	0.121	0.137	0.146	0.128	
Value	15000	16000	17000	18000	19000	100000							
Frequency	330	200	138	68	18	30							
Proportion	0.110	0.067	0.046	0.023	0.006	0.010							
For the frequency table, variable is rounded to the nearest 1000													
credit_age													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
3000	0	323	1	283.3	70.89	182.0	202.9	241.0	282.0	325.0	365.0	388.0	
lowest : 81 99 102 104 111, highest: 488 489 492 503 507													
credit_good_age													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
3000	0	181	1	147.5	34.67	97	109	127	147	169	187	197	
lowest : 24 38 46 53 57, highest: 240 241 242 250 253													
card_age													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
3000	0	320	1	253.2	70.86	152	172	210	251	296	334	359	
lowest : 76 78 82 84 88, highest: 449 453 455 472 479													
non_mtg_acc_past_due_12_months_num													
n	missing	distinct											
3000	0	5											
lowest : 0 1 2 3 4, highest: 0 1 2 3 4													
Value	0	1	2	3	4								
Frequency	2750	152	60	37	1								
Proportion	0.917	0.051	0.020	0.012	0.000								

non_mtg_acc_past_due_6_months_num

	n	missing	distinct
	3000	0	3

Value	0	1	2
Frequency	2905	93	2
Proportion	0.968	0.031	0.001

mortgages_past_due_6_months_num

	n	missing	distinct
	3000	0	2

Value	0	1
Frequency	2896	104
Proportion	0.965	0.035

credit_past_due_amount

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	3000	0	105	0.1	336.2	655.9	0	0	0	0	0	0	0

lowest : 0.00 103.02 382.75 1399.32 1433.06, highest: 18067.77 18964.98 19219.85 19665.77 20095.43

inq_12_month_num

	n	missing	distinct	Info	Mean	Gmd
	3000	0	9	0.75	0.8317	1.239

lowest : 0 1 2 3 4, highest: 4 5 6 7 8

Value	0	1	2	3	4	5	6	7	8
Frequency	1879	439	300	197	99	56	23	5	2
Proportion	0.626	0.146	0.100	0.066	0.033	0.019	0.008	0.002	0.001

card_inq_24_month_num

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	3000	0	14	0.825	1.346	1.962	0	0	0	0	2	4	6

lowest : 0 1 2 3 4, highest: 9 10 11 12 13

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	1664	439	275	209	130	92	54	62	30	21	14	6	2	2
Proportion	0.555	0.146	0.092	0.070	0.043	0.031	0.018	0.021	0.010	0.007	0.005	0.002	0.001	0.001

card_open_36_month_num

	n	missing	distinct
	3000	0	3

Value	0	1	2
Frequency	2470	509	21
Proportion	0.823	0.170	0.007

auto_open_36_month_num

	n	missing	distinct
	3000	0	3

Value	0	1	2
Frequency	2301	698	1
Proportion	0.767	0.233	0.000

uti_card

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	3000	0	3000	1	0.4921	0.1169	0.3222	0.3603	0.4214	0.4913	0.5610	0.6269	0.6658

lowest : 0.1216411 0.1610491 0.1623538 0.1643655 0.1706756
highest: 0.7926201 0.7959899 0.8038095 0.8148712 0.8488642

uti_50plus_pct

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
	3000	0	3000	1	0.4852	0.1293	0.2993	0.3398	0.4080	0.4800	0.5631	0.6346	0.6751

lowest : 0.06615862 0.08950864 0.13019945 0.15545320 0.15870158
highest: 0.82108601 0.82707938 0.83645644 0.86584674 0.91651642

uti_max_credit_line

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
3000	0	3000	1	0.4588	0.1177	0.2894	0.3253	0.3911	0.4567	0.5306	0.5922	0.6312

lowest : 0.06682402 0.08292235 0.08896100 0.09177259 0.09666043
highest: 0.76533172 0.76876205 0.77164806 0.78304478 0.82486532

uti_card_50plus_pct

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2703	297	2703	1	0.4586	0.1219	0.2817	0.3204	0.3864	0.4589	0.5315	0.5991	0.6339

lowest : 0.05972517 0.07077929 0.07554617 0.11759495 0.13948649
highest: 0.75788080 0.76987987 0.77560435 0.77599721 0.78222749

ind_acc_XYZ

n	missing	distinct
3000	0	2

Value	0	1
Frequency	2229	771
Proportion	0.743	0.257

rep_income

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2747	253	95	1	82895	17743	56300	63000	72000	83000	93500	103000	108000

lowest : 27000 33000 35000 36000 37000, highest: 123000 125000 128000 137000 147000

States

n	missing	distinct
3000	0	7

lowest : AL FL GA LA MS, highest: GA LA MS NC SC

Value	AL	FL	GA	LA	MS	NC	SC
Frequency	451	433	400	413	461	429	413
Proportion	0.150	0.144	0.133	0.138	0.154	0.143	0.138

Default_ind

n	missing	distinct
3000	0	2

Value	0	1
Frequency	2778	222
Proportion	0.926	0.074

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

`tot_credit_debt`

* normality test : Shapiro-Wilk normality test

- statistic : 0.99949, p-value : 0.642511

Table 2.1: skewness and kurtosis : `tot_credit_debt`

type	skewness	kurtosis
original	0.0004	3.0950
log transformation	-0.9784	5.4617
sqrt transformation	-0.4240	3.5855

Normality Diagnosis Plot (x)

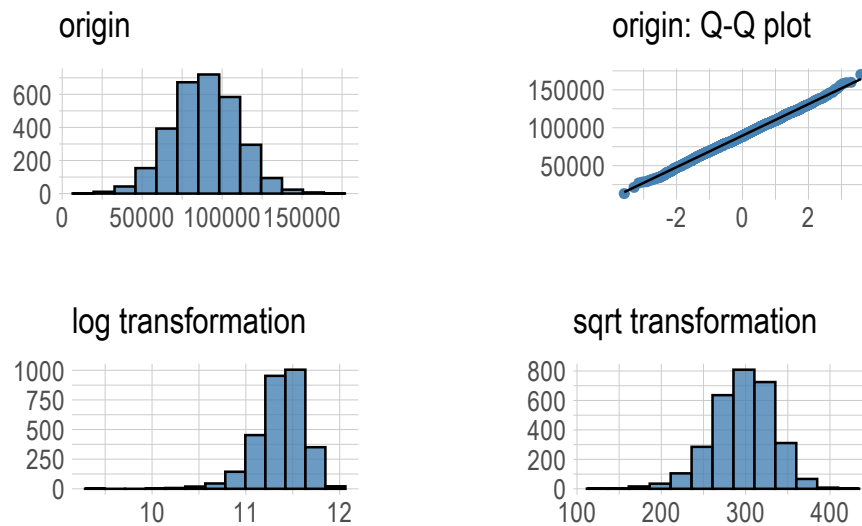


Figure 2.1: `tot_credit_debt`

avg_card_debt

* normality test : Shapiro-Wilk normality test
 - statistic : 0.28576, p-value : 1.7697E-75

Table 2.2: skewness and kurtosis : avg_card_debt

type	skewness	kurtosis
original	8.6152	82.3612
log transformation	2.6728	22.5243
sqrt transformation	6.1734	54.2633

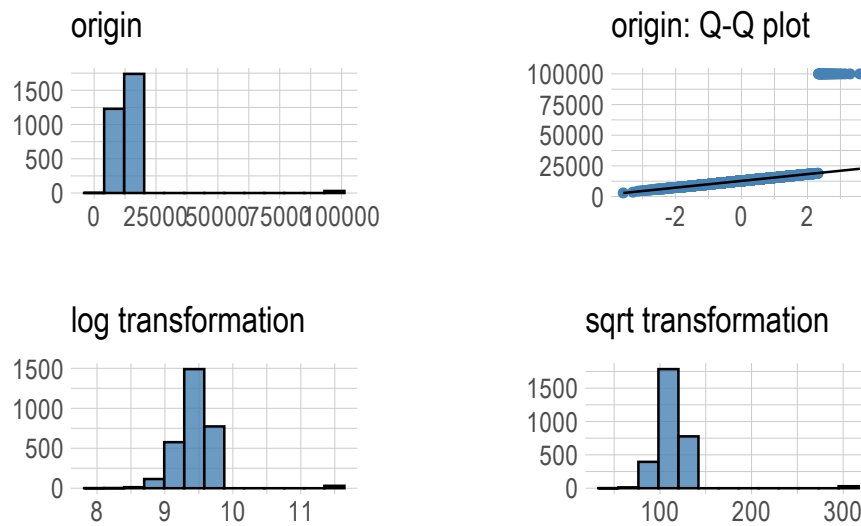
Normality Diagnosis Plot (x)

Figure 2.2: avg_card_debt

credit_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.99927, p-value : 0.282164

Table 2.3: skewness and kurtosis : credit_age

type	skewness	kurtosis
original	0.0843	2.9728
log transformation	-0.6657	3.9110
sqrt transformation	-0.2665	3.1514

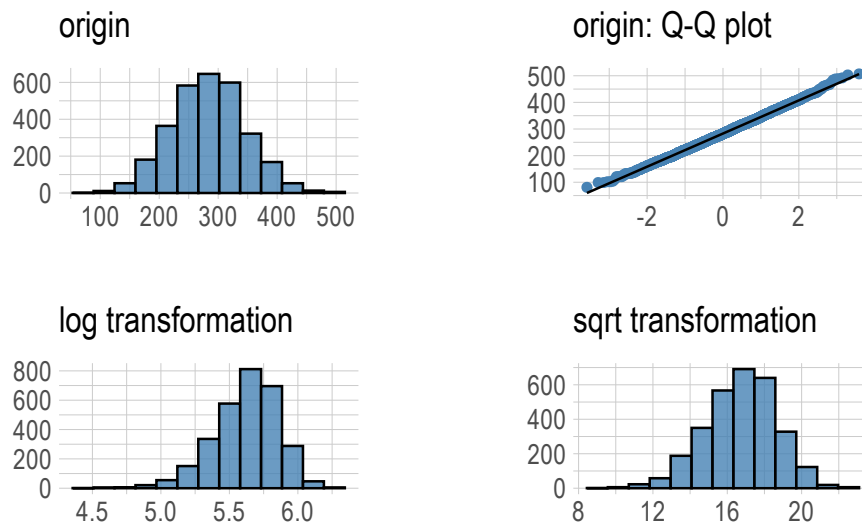
Normality Diagnosis Plot (x)

Figure 2.3: credit_age

credit_good_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.99939, p-value : 0.449667

Table 2.4: skewness and kurtosis : credit_good_age

type	skewness	kurtosis
original	0.0165	3.0639
log transformation	-0.8645	5.4994
sqrt transformation	-0.3586	3.5511

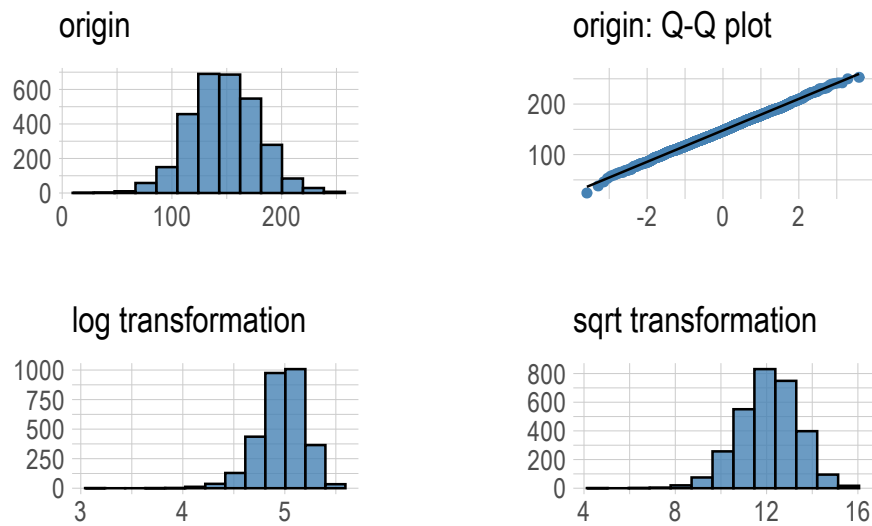
Normality Diagnosis Plot (x)

Figure 2.4: credit_good_age

card_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.99824, p-value : 0.00219145

Table 2.5: skewness and kurtosis : card_age

type	skewness	kurtosis
original	0.1305	2.8417
log transformation	-0.6348	3.6251
sqrt transformation	-0.2294	2.9578

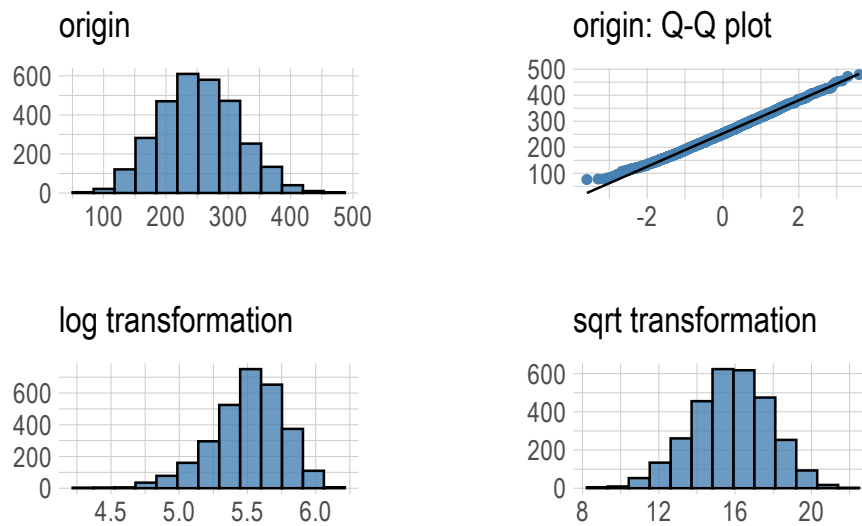
Normality Diagnosis Plot (x)

Figure 2.5: card_age

credit_past_due_amount

* normality test : Shapiro-Wilk normality test
- statistic : 0.16234, p-value : 8.39912E-79

Table 2.6: skewness and kurtosis : credit_past_due_amount

type	skewness	kurtosis
original	6.5191	47.4920
log+1 transformation	5.1402	27.5643
sqrt transformation	5.6053	33.8176

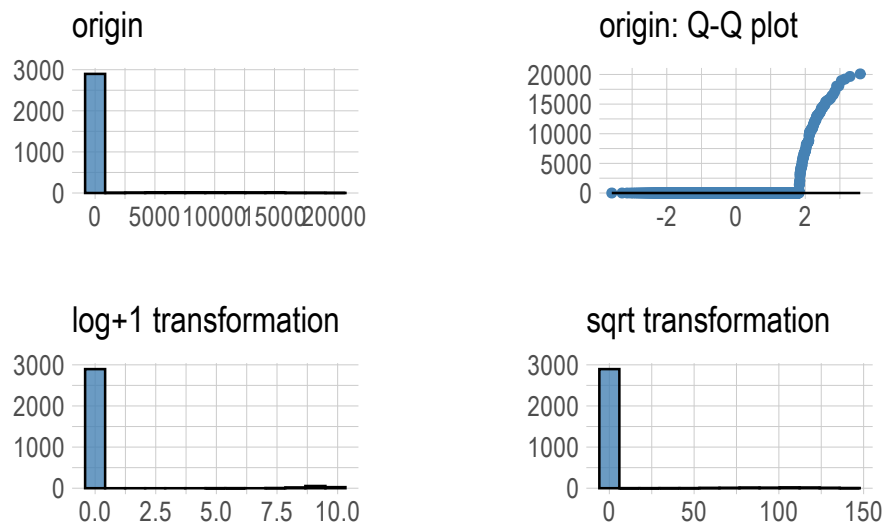
Normality Diagnosis Plot (x)

Figure 2.6: credit_past_due_amount

inq_12_month_num

* normality test : Shapiro-Wilk normality test
 - statistic : 0.67234, p-value : 2.76349E-60

Table 2.7: skewness and kurtosis : inq_12_month_num

type	skewness	kurtosis
original	1.8284	6.0630
log+1 transformation	1.0352	2.6781
sqrt transformation	0.9258	2.4022

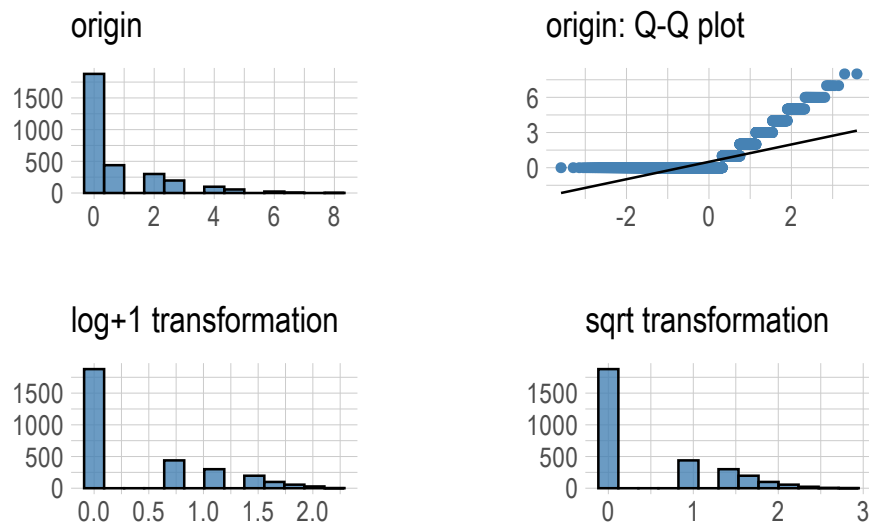
Normality Diagnosis Plot (x)

Figure 2.7: inq_12_month_num

card_inq_24_month_num

* normality test : Shapiro-Wilk normality test
 - statistic : 0.68693, p-value : 1.88563E-59

Table 2.8: skewness and kurtosis : card_inq_24_month_num

type	skewness	kurtosis
original	2.0051	7.0588
log+1 transformation	0.8977	2.5033
sqrt transformation	0.8526	2.5037

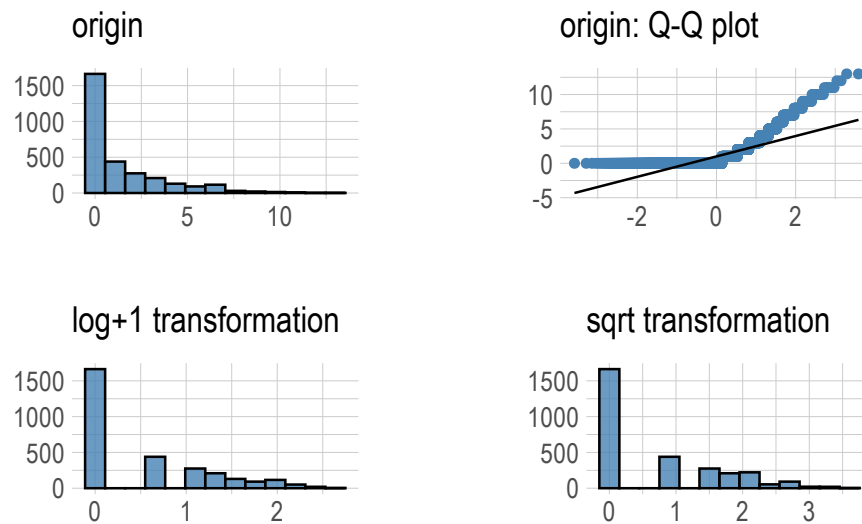
Normality Diagnosis Plot (x)

Figure 2.8: card_inq_24_month_num

uti_card

* normality test : Shapiro-Wilk normality test
- statistic : 0.99967, p-value : 0.92782

Table 2.9: skewness and kurtosis : uti_card

type	skewness	kurtosis
original	-0.0121	2.9754
log transformation	-0.7974	4.5305
sqrt transformation	-0.3664	3.3655

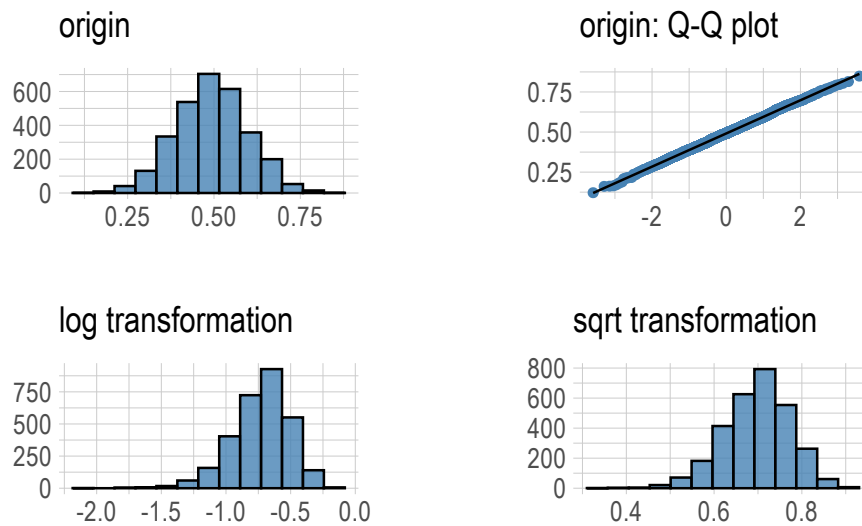
Normality Diagnosis Plot (x)

Figure 2.9: uti_card

uti_50plus_pct

* normality test : Shapiro-Wilk normality test
- statistic : 0.99934, p-value : 0.37859

Table 2.10: skewness and kurtosis : uti_50plus_pct

type	skewness	kurtosis
original	0.0481	2.9088
log transformation	-0.8995	5.4613
sqrt transformation	-0.3484	3.3836

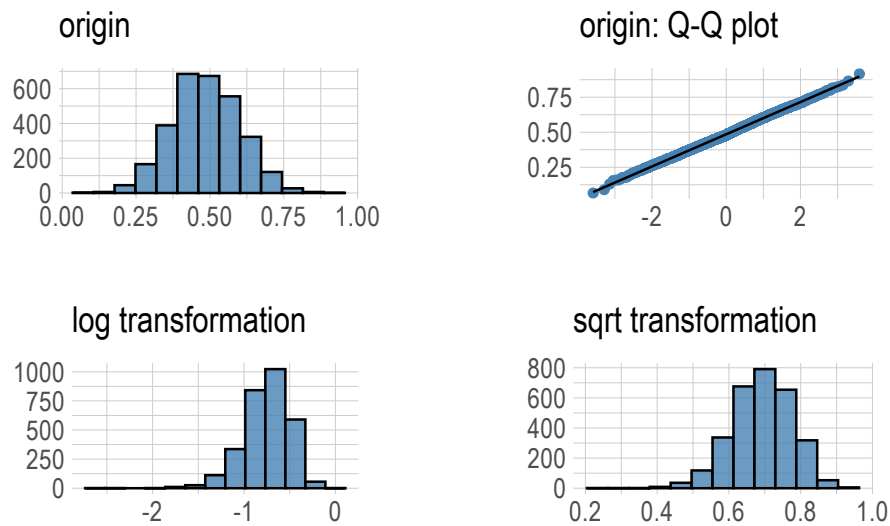
Normality Diagnosis Plot (x)

Figure 2.10: uti_50plus_pct

uti_max_credit_line

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99897, p-value : 0.0715545

Table 2.11: skewness and kurtosis : uti_max_credit_line

type	skewness	kurtosis
original	-0.0385	3.1383
log transformation	-1.2121	7.3422
sqrt transformation	-0.5045	4.0204

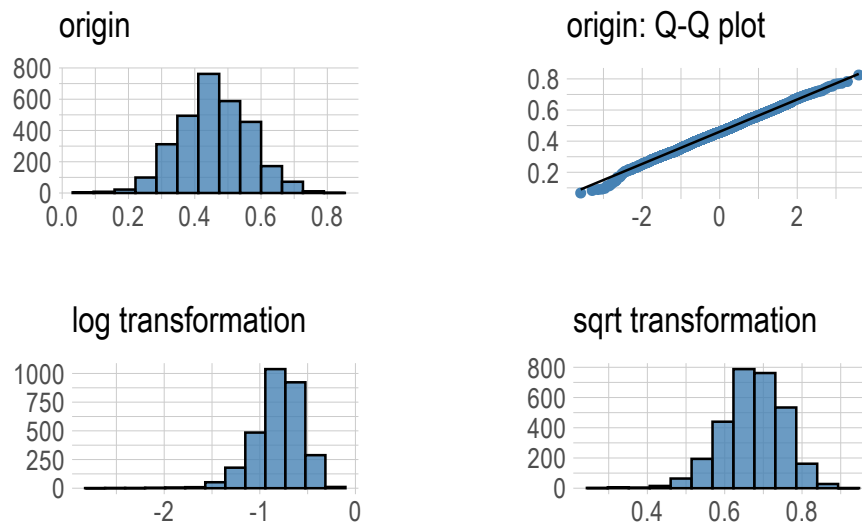
Normality Diagnosis Plot (x)

Figure 2.11: uti_max_credit_line

uti_card_50plus_pct

* normality test : Shapiro-Wilk normality test
- statistic : 0.99927, p-value : 0.375725

Table 2.12: skewness and kurtosis : uti_card_50plus_pct

type	skewness	kurtosis
original	-0.0711	2.9940
log transformation	-1.1792	6.8556
sqrt transformation	-0.5140	3.7823

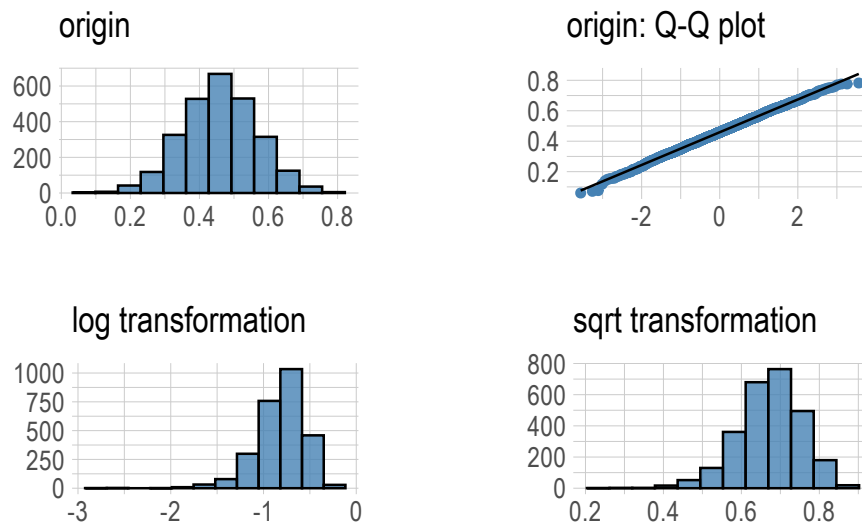
Normality Diagnosis Plot (x)

Figure 2.12: uti_card_50plus_pct

rep_income

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99866, p-value : 0.026873

Table 2.13: skewness and kurtosis : rep_income

type	skewness	kurtosis
original	-0.0790	2.9742
log transformation	-0.7427	4.1178
sqrt transformation	-0.3879	3.2928

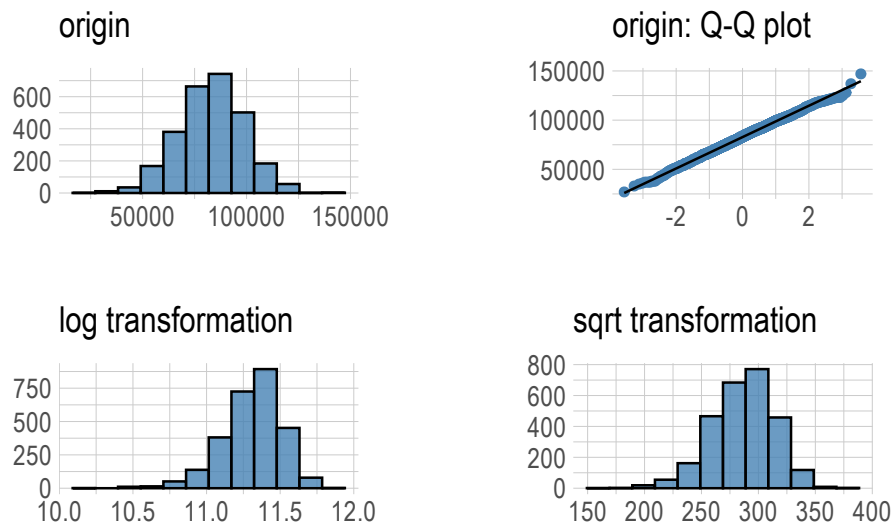
Normality Diagnosis Plot (x)

Figure 2.13: rep_income

Chapter 3

Relationship Between Variables

3.1 Correlation Coefficient

3.1.1 Correlation Coefficient by Variable Combination

Table 3.1: The correlation coefficients (0.5 or more)

Variable1	Variable2	Correlation Coefficient
card_age	credit_age	0.934
card_inq_24_month_num	inq_12_month_num	0.883
uti_card_50plus_pct	uti_card	0.856
credit_good_age	credit_age	0.792
uti_50plus_pct	uti_card	0.752
card_age	credit_good_age	0.750
uti_max_credit_line	uti_card	0.741
uti_card_50plus_pct	uti_50plus_pct	0.634
uti_card_50plus_pct	uti_max_credit_line	0.627
uti_max_credit_line	uti_50plus_pct	0.561

3.1.2 Correlation Plot of Numerical Variables

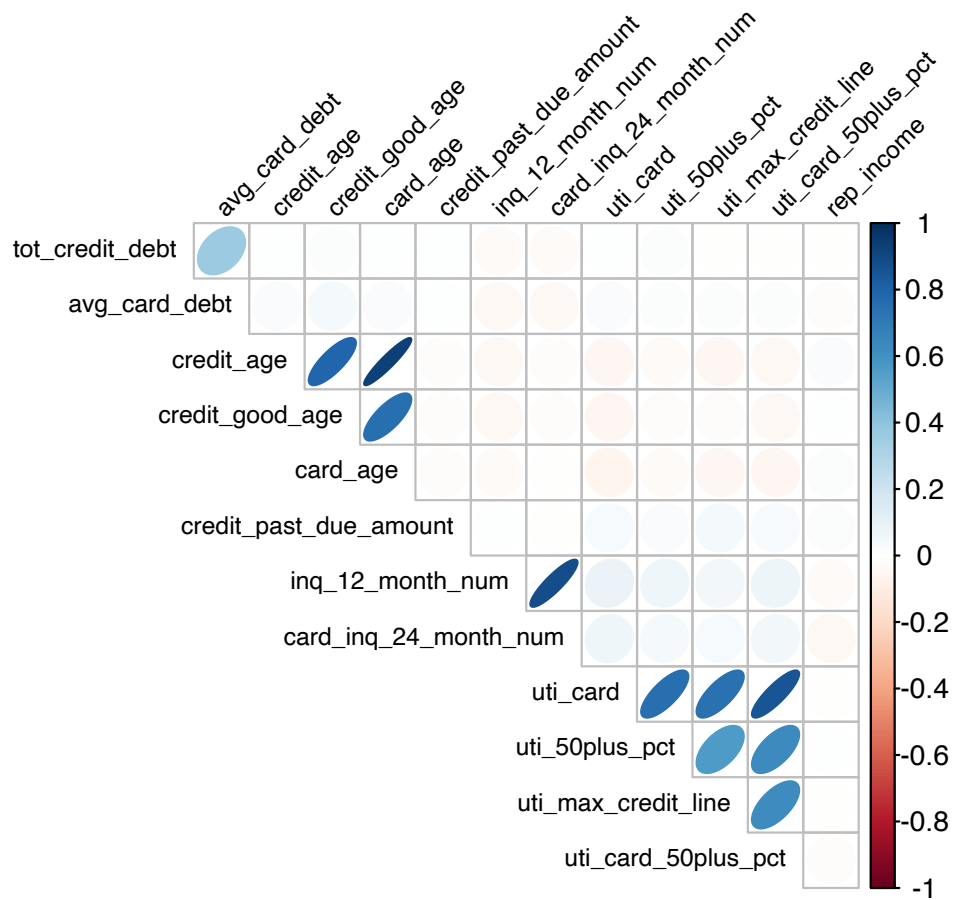


Figure 3.1: The correlation coefficient of numerical variables

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

4.1.2 Grouped Categorical Variables

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

4.2.2 Grouped Correlation Plot of Numerical Variables