



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 20,000 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
<code>tot_credit_debt</code>	numeric	0	0.00	19978	0.999
<code>avg_card_debt</code>	numeric	0	0.00	19607	0.980
<code>credit_age</code>	numeric	0	0.00	410	0.020
<code>credit_good_age</code>	numeric	0	0.00	243	0.012
<code>card_age</code>	numeric	0	0.00	383	0.019
<code>non_mtg_acc_past_due_12_months_num</code>	character	0	0.00	5	0.000
<code>non_mtg_acc_past_due_6_months_num</code>	character	0	0.00	3	0.000
<code>mortgages_past_due_6_months_num</code>	character	0	0.00	2	0.000
<code>credit_past_due_amount</code>	numeric	0	0.00	605	0.030
<code>inq_12_month_num</code>	numeric	0	0.00	11	0.001
<code>card_inq_24_month_num</code>	numeric	0	0.00	19	0.001
<code>card_open_36_month_num</code>	character	0	0.00	3	0.000
<code>auto_open_36_month_num</code>	character	0	0.00	3	0.000
<code>uti_card</code>	numeric	0	0.00	20000	1.000
<code>uti_50plus_pct</code>	numeric	0	0.00	20000	1.000
<code>uti_max_credit_line</code>	numeric	0	0.00	20000	1.000
<code>uti_card_50plus_pct</code>	numeric	2055	10.27	17946	0.897
<code>ind_acc_XYZ</code>	character	0	0.00	2	0.000
<code>rep_income</code>	numeric	1570	7.85	118	0.006
<code>States</code>	factor	0	0.00	7	0.000
<code>Default_ind</code>	character	0	0.00	2	0.000

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

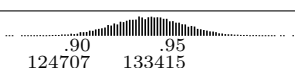
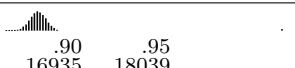



1.3 About EDA Report

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 20000 Observations													
tot_credit_debt													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
20000	0	19978	1	94564	26555	55824	64443	78744	94671	110329	124707	133415	
lowest : 2367.43 3664.49 4662.60 6898.50 11363.34													
highest: 175998.38 179084.56 182094.91 182858.99 188890.96													
avg_card_debt													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
20000	0	19607	1	14088	4913	8454	9555	11322	13244	15196	16935	18039	
lowest : 2363.12 2521.21 2814.66 3074.70 3148.68, highest: 19945.05 19955.42 19959.03 19960.61 99999.00													
Value	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	13000	
Frequency	1	6	8	51	123	258	591	909	1452	2017	2511	2803	
Proportion	0.000	0.000	0.000	0.003	0.006	0.013	0.030	0.045	0.073	0.101	0.126	0.140	
Value	14000	15000	16000	17000	18000	19000	20000	100000					
Frequency	2638	2261	1801	1158	695	396	109	212					
Proportion	0.132	0.113	0.090	0.058	0.035	0.020	0.005	0.011					
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	
20000	0	410	1	296.7	69.64	195	217	255	297	339	375	398	
lowest : 54 78 79 80 82, highest: 521 527 537 539 545													
credit_good_age													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
20000	0	243	1	149.8	38.34	94	106	127	150	172	193	205	
lowest : 21 26 27 28 31, highest: 279 280 281 283 296													
card_age													
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
20000	0	383	1	268	67.04	171	191	227	268	308	344	365	
lowest : 41 56 62 71 75, highest: 481 484 494 516 520													
non_mtg_acc_past_due_12_months_num													
n	missing	distinct											
20000	0	5											
lowest : 0 1 2 3 4, highest: 0 1 2 3 4													
Value	0	1	2	3	4								
Frequency	18502	918	446	119	15								
Proportion	0.925	0.046	0.022	0.006	0.001								

non_mtg_acc_past_due_6_months_num

n	missing	distinct
20000	0	3

Value	0	1	2
Frequency	19481	490	29
Proportion	0.974	0.024	0.001

mortgages_past_due_6_months_num

n	missing	distinct
20000	0	2

Value	0	1
Frequency	19396	604
Proportion	0.97	0.03

credit_past_due_amount

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	605	0.088	329.3	643.9	0	0	0	0	0	0	0

lowest : 0.00 316.39 434.70 602.68 695.96, highest: 27229.53 27726.89 28644.74 29392.72 32662.98

inq_12_month_num

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	11	0.948	1.763	1.893	0	0	0	1	3	4	5

lowest : 0 1 2 3 4, highest: 6 7 8 9 10

Value	0	1	2	3	4	5	6	7	8	9	10
Frequency	6696	3541	3475	2871	1824	968	423	153	37	11	1
Proportion	0.335	0.177	0.174	0.144	0.091	0.048	0.021	0.008	0.002	0.001	0.000

card_inq_24_month_num

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	19	0.984	3.41	3.237	0	0	1	3	5	8	9

lowest : 0 1 2 3 4, highest: 14 15 16 17 18

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	3936	2452	2654	2401	2093	1809	1503	1092	824	521	341	189	93	58
Proportion	0.197	0.123	0.133	0.120	0.105	0.090	0.075	0.055	0.041	0.026	0.017	0.009	0.005	0.003

Value	14	15	16	17	18
Frequency	18	5	6	3	2
Proportion	0.001	0.000	0.000	0.000	0.000

card_open_36_month_num

n	missing	distinct
20000	0	3

Value	0	1	2
Frequency	16865	3009	126
Proportion	0.843	0.150	0.006

auto_open_36_month_num

n	missing	distinct
20000	0	3

Value	0	1	2
Frequency	17191	2798	11
Proportion	0.860	0.140	0.001

uti_card

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	20000	1	0.5032	0.1233	0.3238	0.3628	0.4296	0.5028	0.5774	0.6443	0.6816

lowest : 0.06512047 0.06563675 0.07869497 0.10148322 0.11754010
highest: 0.89357072 0.90489927 0.92232634 0.92532315 0.96928868

uti_50plus_pct

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	20000	1	0.511	0.128	0.3254	0.3653	0.4352	0.5099	0.5884	0.6566	0.6975

lowest : 0.03374933 0.07398763 0.08376058 0.11596965 0.12081086
highest: 0.89448028 0.89499581 0.90084806 0.90509788 0.98896404

uti_max_credit_line

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
20000	0	20000	1	0.5076	0.1226	0.3290	0.3680	0.4335	0.5072	0.5814	0.6467	0.6874

lowest : 0.005173925 0.091742468 0.098516713 0.115342939 0.117451965
highest: 0.894630428 0.903665489 0.912962710 0.971640159 1.000000000

uti_card_50plus_pct

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50
17945	2055	17945	1	0.4896	0.1348	0.2923	0.3380	0.4098	0.4901
.75	.90	.95							
0.5690	0.6431	0.6855							

lowest : 0.000000000 0.005784274 0.032522037 0.065678794 0.068748893
highest: 0.918661007 0.929283466 0.931222261 0.949958864 0.970775774

ind_acc_XYZ

n	missing	distinct
20000	0	2

Value	0	1
Frequency	14829	5171
Proportion	0.741	0.259

rep_income

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
18430	1570	117	1	75500	18465	49000	55000	64000	75000	86000	97000	102000

lowest : 12000 18000 19000 20000 22000, highest: 130000 131000 132000 134000 150000

States

n	missing	distinct
20000	0	7

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

Value	AL	FL	GA	LA	MS	NC	SC
Frequency	2893	2857	2857	2849	2827	2898	2819
Proportion	0.145	0.143	0.143	0.142	0.141	0.145	0.141

Default_ind

n	missing	distinct
20000	0	2

Value	0	1
Frequency	18414	1586
Proportion	0.921	0.079

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 : 0.99983, p-value : 0.975065

Table 2.1: skewness and kurtosis : `tot_credit_debt`

type	skewness	kurtosis
original	-0.0023	2.9534
log transformation	-1.0477	6.1720
sqrt transformation	-0.4303	3.4925

Normality Diagnosis Plot (x)

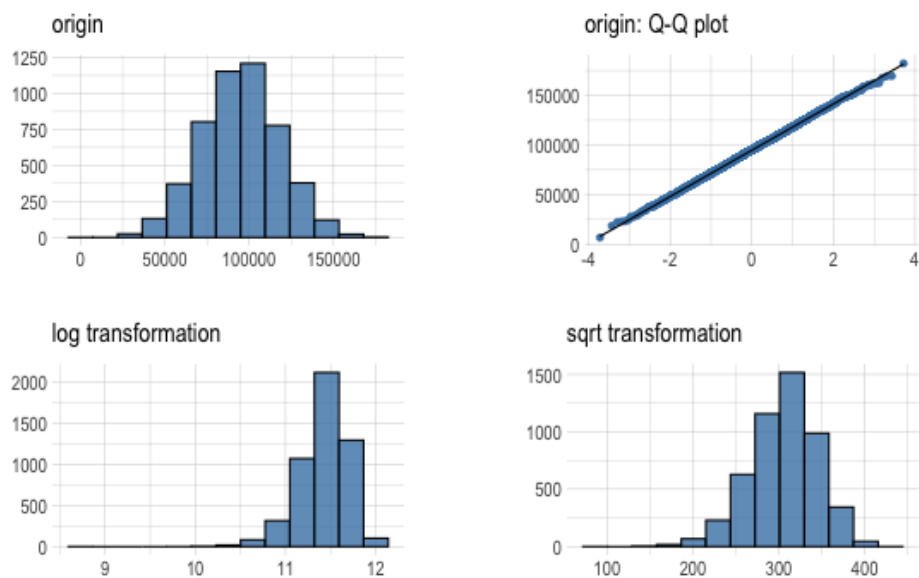


Figure 2.1: `tot_credit_debt`

avg_card_debt

* normality test : Shapiro-Wilk normality test
 - statistic : 0.29711, p-value : 1.09698E-87

Table 2.2: skewness and kurtosis : avg_card_debt

type	skewness	kurtosis
original	8.1202	73.3227
log transformation	2.6827	21.2375
sqrt transformation	5.8910	49.0992

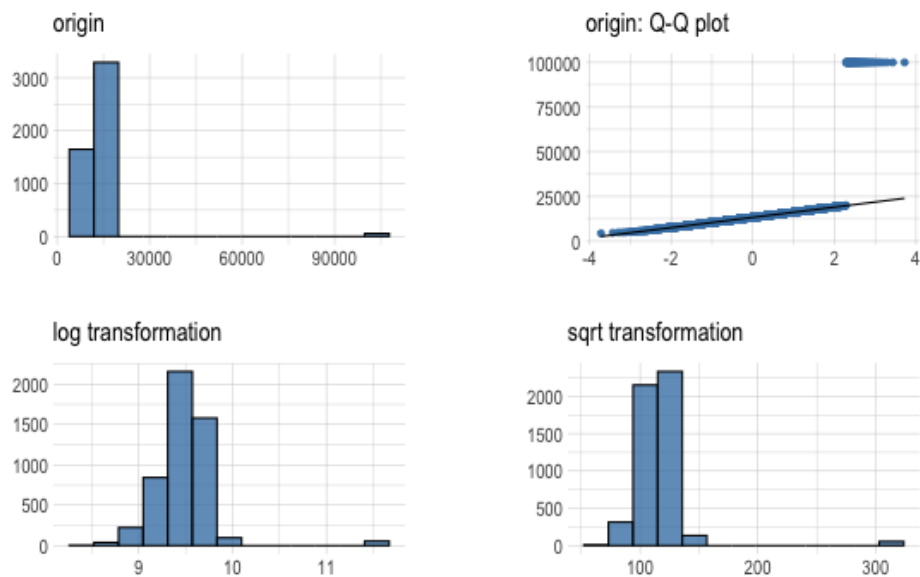
Normality Diagnosis Plot (x)

Figure 2.2: avg_card_debt

credit_age

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99955, p-value : 0.295429

Table 2.3: skewness and kurtosis : credit_age

type	skewness	kurtosis
original	0.0270	2.9560
log transformation	-0.7069	4.1623
sqrt transformation	-0.3102	3.2305

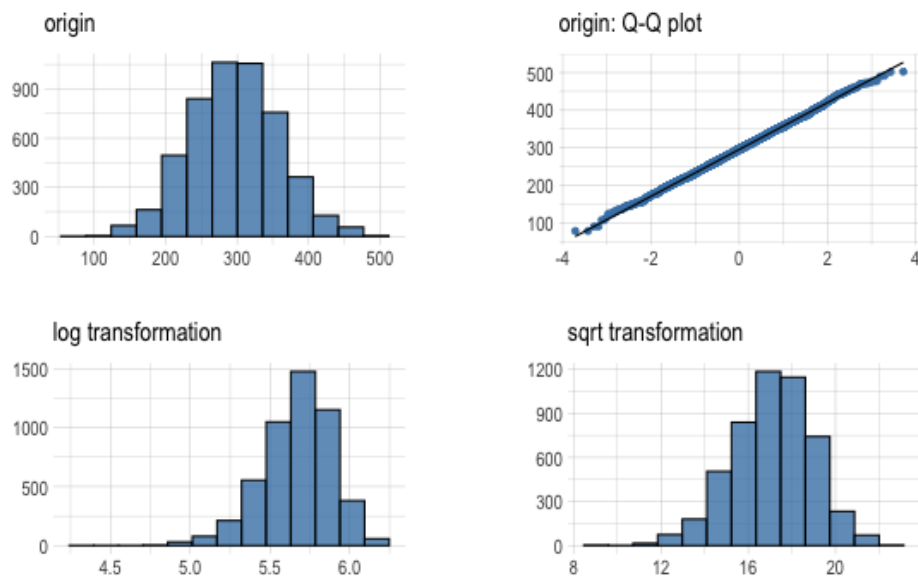
Normality Diagnosis Plot (x)

Figure 2.3: credit_age

credit_good_age

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99934, p-value : 0.0642439

Table 2.4: skewness and kurtosis : credit_good_age

type	skewness	kurtosis
original	0.0124	2.9773
log transformation	-0.8046	4.3875
sqrt transformation	-0.3597	3.2713

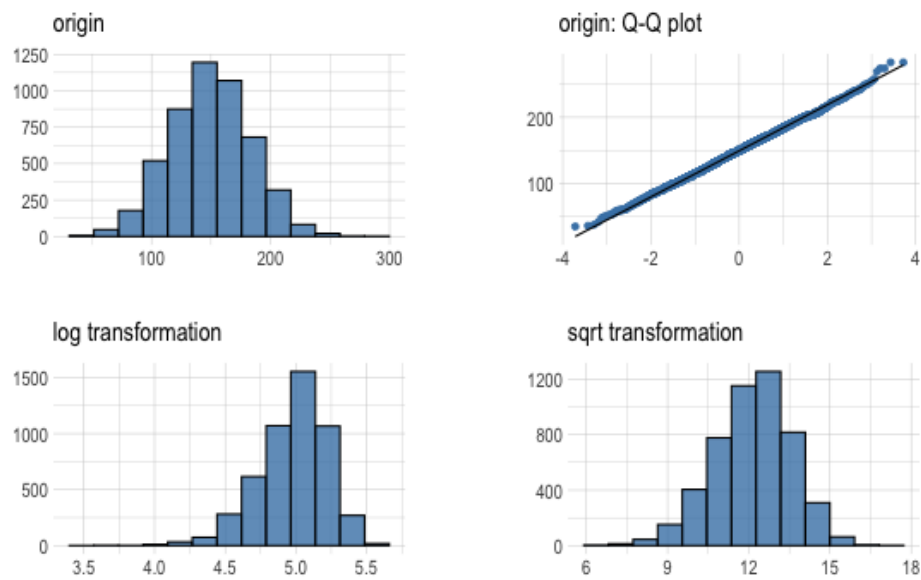
Normality Diagnosis Plot (x)

Figure 2.4: credit_good_age

card_age

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99936, p-value : 0.0721741

Table 2.5: skewness and kurtosis : card_age

type	skewness	kurtosis
original	0.0104	2.9944
log transformation	-0.7644	4.2006
sqrt transformation	-0.3476	3.2398

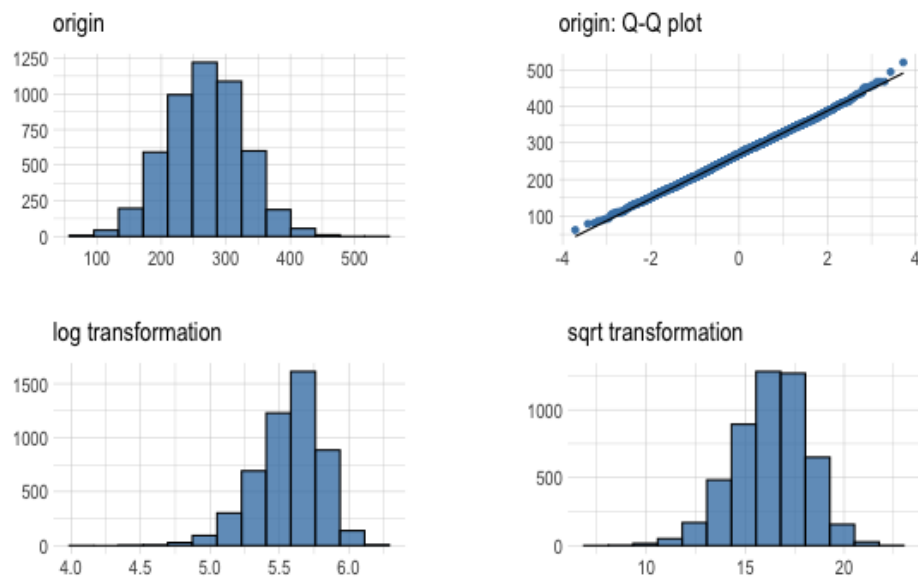
Normality Diagnosis Plot (x)

Figure 2.5: card_age

credit_past_due_amount

* normality test : Shapiro-Wilk normality test
 - statistic : 0.14871, p-value : 4.45892E-92

Table 2.6: skewness and kurtosis : credit_past_due_amount

type	skewness	kurtosis
original	7.3081	63.2718
log+1 transformation	5.5926	32.3566
sqrt transformation	5.9966	38.7595

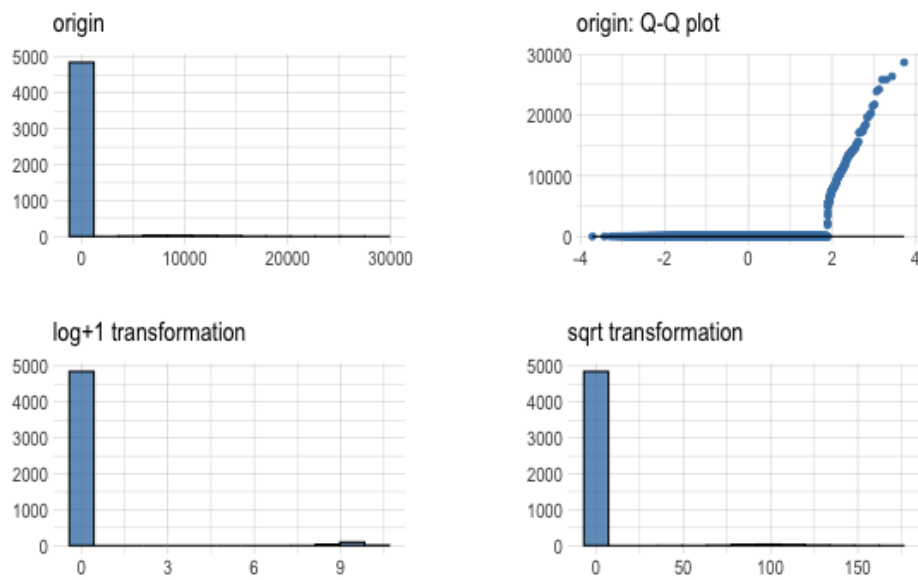
Normality Diagnosis Plot (x)

Figure 2.6: credit_past_due_amount

inq_12_month_num

* normality test : Shapiro-Wilk normality test
 - statistic : 0.87592, p-value : 6.55607E-53

Table 2.7: skewness and kurtosis : inq_12_month_num

type	skewness	kurtosis
original	0.8060	3.0021
log+1 transformation	-0.0184	1.6253
sqrt transformation	-0.1107	1.6512

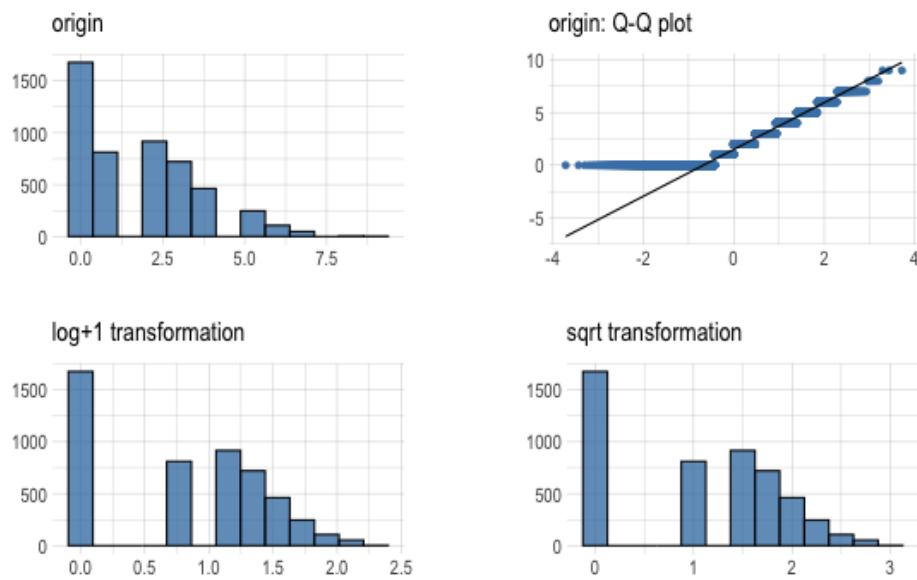
Normality Diagnosis Plot (x)

Figure 2.7: inq_12_month_num

card_inq_24_month_num

* normality test : Shapiro-Wilk normality test
 - statistic : 0.91491, p-value : 1.56517E-46

Table 2.8: skewness and kurtosis : card_inq_24_month_num

type	skewness	kurtosis
original	0.8246	3.2083
log+1 transformation	-0.3243	1.9789
sqrt transformation	-0.2870	2.1592

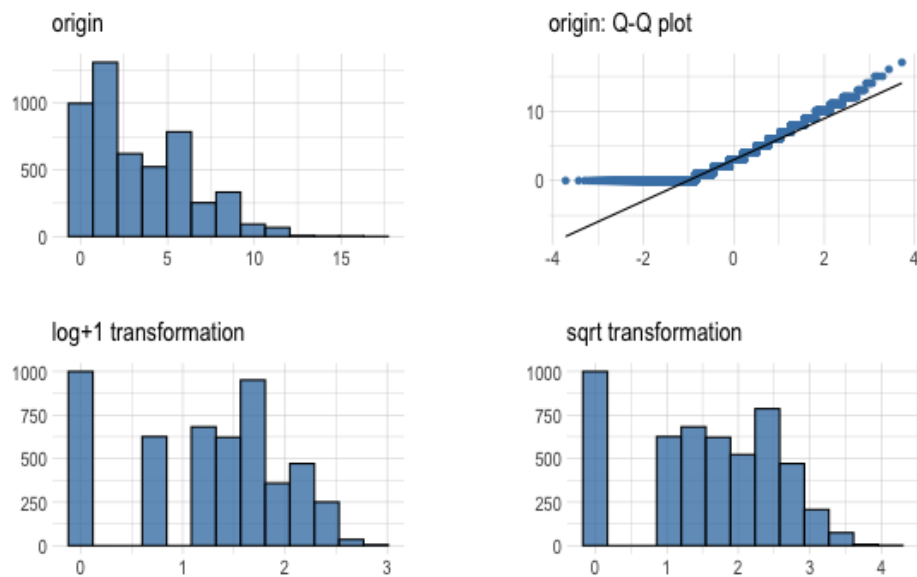
Normality Diagnosis Plot (x)

Figure 2.8: card_inq_24_month_num

uti_card

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99973, p-value : 0.787722

Table 2.9: skewness and kurtosis : uti_card

type	skewness	kurtosis
original	-0.0217	2.9832
log transformation	-0.8130	4.5072
sqrt transformation	-0.3801	3.3563

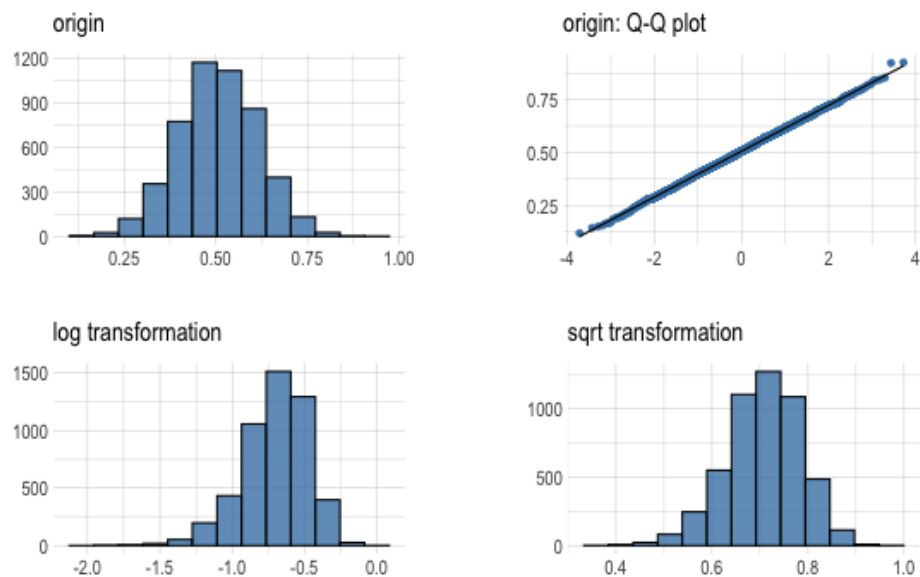
Normality Diagnosis Plot (x)

Figure 2.9: uti_card

uti_50plus_pct

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99939, p-value : 0.0941562

Table 2.10: skewness and kurtosis : uti_50plus_pct

type	skewness	kurtosis
original	0.0307	2.9962
log transformation	-0.7520	4.0590
sqrt transformation	-0.3355	3.2244

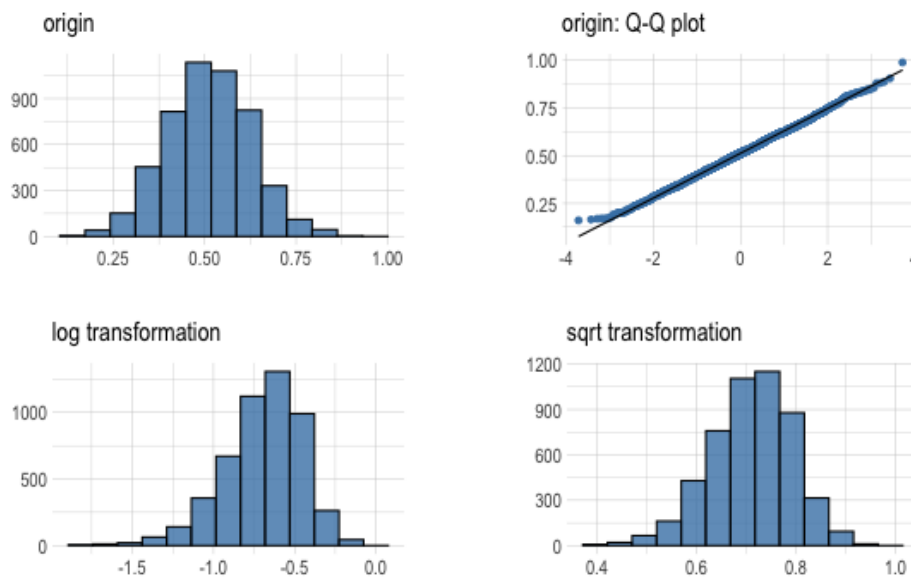
Normality Diagnosis Plot (x)

Figure 2.10: uti_50plus_pct

uti_max_credit_line

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99913, p-value : 0.0120447

Table 2.11: skewness and kurtosis : uti_max_credit_line

type	skewness	kurtosis
original	0.0587	3.1549
log transformation	-0.7350	4.2272
sqrt transformation	-0.3124	3.3502

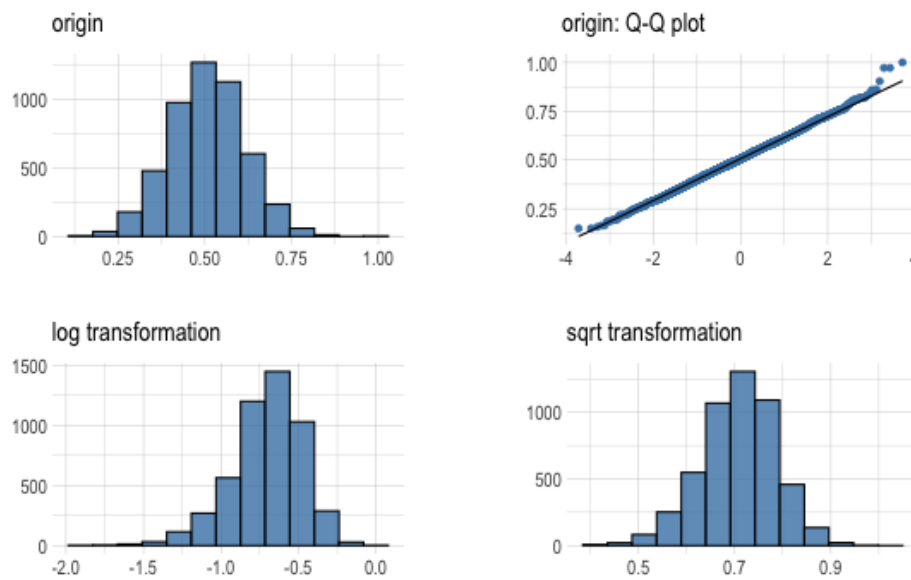
Normality Diagnosis Plot (x)

Figure 2.11: uti_max_credit_line

uti_card_50plus_pct

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99904, p-value : 0.00623877

Table 2.12: skewness and kurtosis : uti_card_50plus_pct

type	skewness	kurtosis
original	-0.0088	3.2331
log transformation	-1.1805	6.4044
sqrt transformation	-0.5044	3.8793

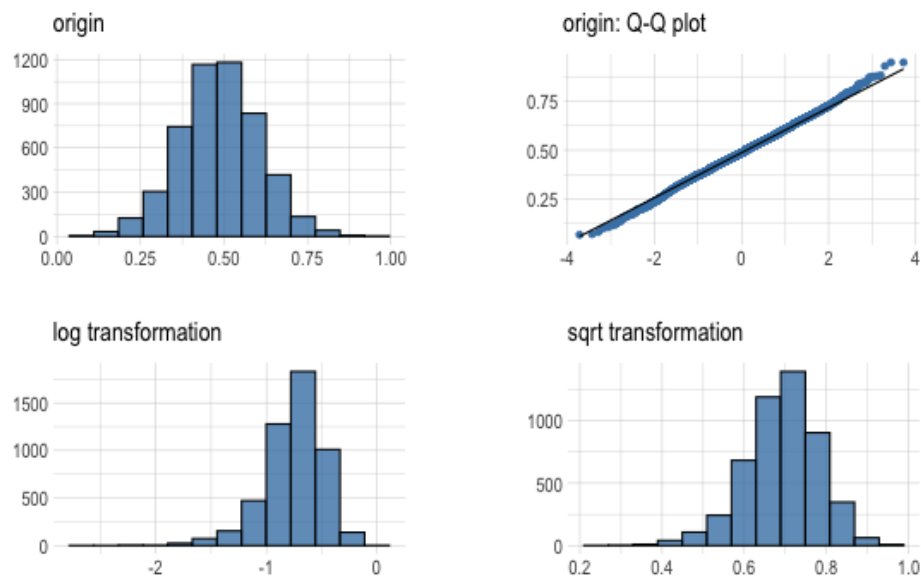
Normality Diagnosis Plot (x)

Figure 2.12: uti_card_50plus_pct

rep_income

* normality test : Shapiro-Wilk normality test
 - statistic : 0.99896, p-value : 0.00333005

Table 2.13: skewness and kurtosis : rep_income

type	skewness	kurtosis
original	-0.0003	2.8535
log transformation	-0.7117	4.0161
sqrt transformation	-0.3277	3.1321

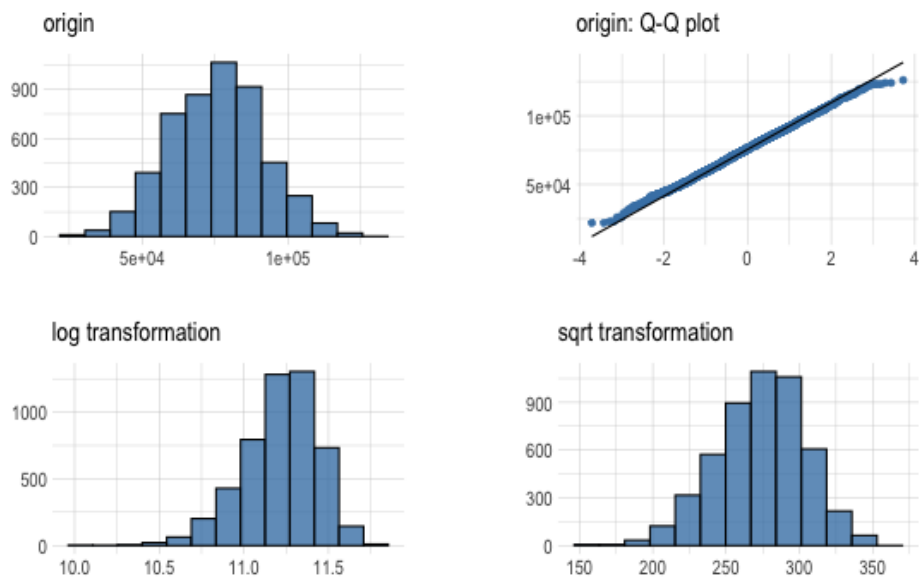
Normality Diagnosis Plot (x)

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	0.937
card_inq_24_month_num	inq_12_month_num	0.859
uti_card_50plus_pct	uti_card	0.847
credit_good_age	credit_age	0.787
uti_50plus_pct	uti_card	0.748
uti_max_credit_line	uti_card	0.746
card_age	credit_good_age	0.736
uti_card_50plus_pct	uti_50plus_pct	0.635
uti_card_50plus_pct	uti_max_credit_line	0.634
uti_max_credit_line	uti_50plus_pct	0.555

3.1.2 Correlation Plot of Numerical Variables

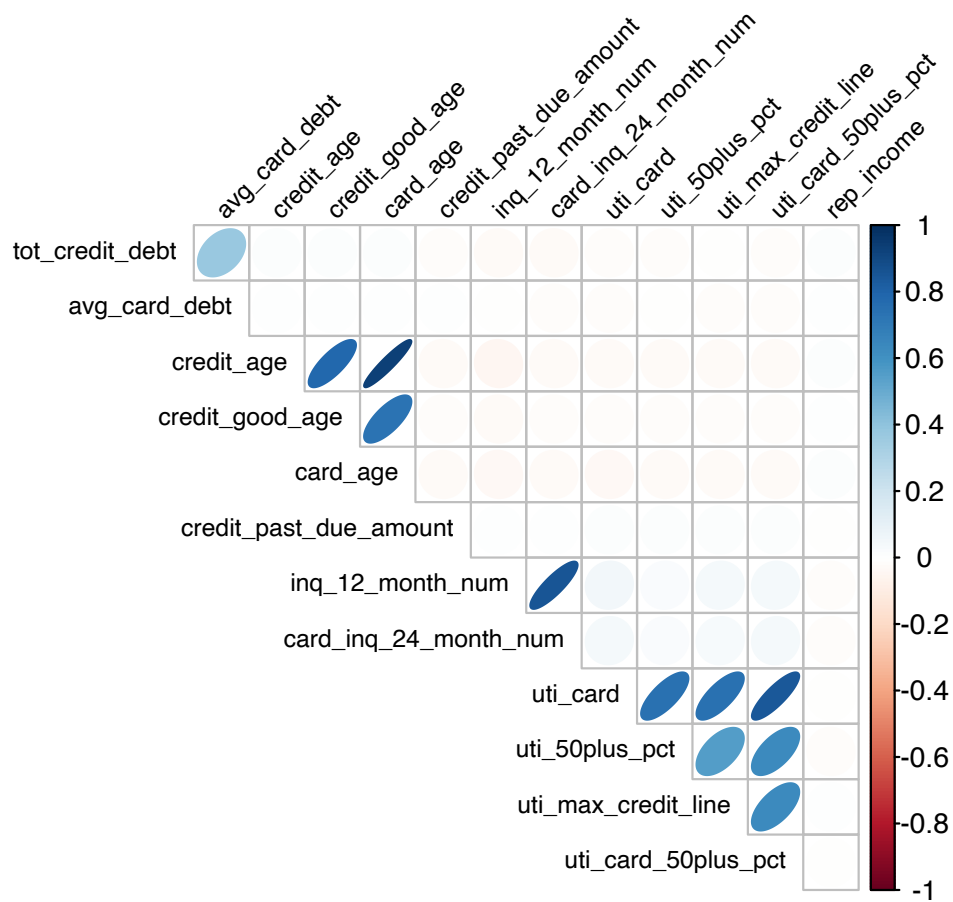


Figure 3.1: The correlation coefficient of numerical variables

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