REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

March 27, 2021

Contents

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
tot_credit_debt	numeric	0	0.00	19978	0.999
avg_card_debt	numeric	0	0.00	19607	0.980
credit_age	numeric	0	0.00	410	0.020
credit_good_age	numeric	0	0.00	243	0.012
card_age	numeric	0	0.00	383	0.019
non_mtg_acc_past_due_12_months_num	character	0	0.00	5	0.000
non_mtg_acc_past_due_6_months_num	character	0	0.00	3	0.000
mortgages_past_due_6_months_num	character	0	0.00	2	0.000
credit_past_due_amount	numeric	0	0.00	605	0.030
inq_12_month_num	numeric	0	0.00	11	0.001
card_inq_24_month_num	numeric	0	0.00	19	0.001
card_open_36_month_num	character	0	0.00	3	0.000
auto_open_.36_month_num	character	0	0.00	3	0.000
uti_card	numeric	0	0.00	20000	1.000
uti_50plus_pct	numeric	0	0.00	20000	1.000
uti_max_credit_line	numeric	0	0.00	20000	1.000
uti_card_50plus_pct	numeric	2055	10.27	17946	0.897
ind_acc_XYZ	character	0	0.00	2	0.000
rep_income	numeric	1570	7.85	118	0.006
States	factor	0	0.00	7	0.000
Default_ind	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 : 0.99969, p-value : 0.664656

Table 2.1: skewness and kurtosis : tot_credit_debt

type	skewness	kurtosis
original	-0.0021	2.9688
log transformation	-0.9563	4.9649
sqrt transformation	-0.4233	3.4041

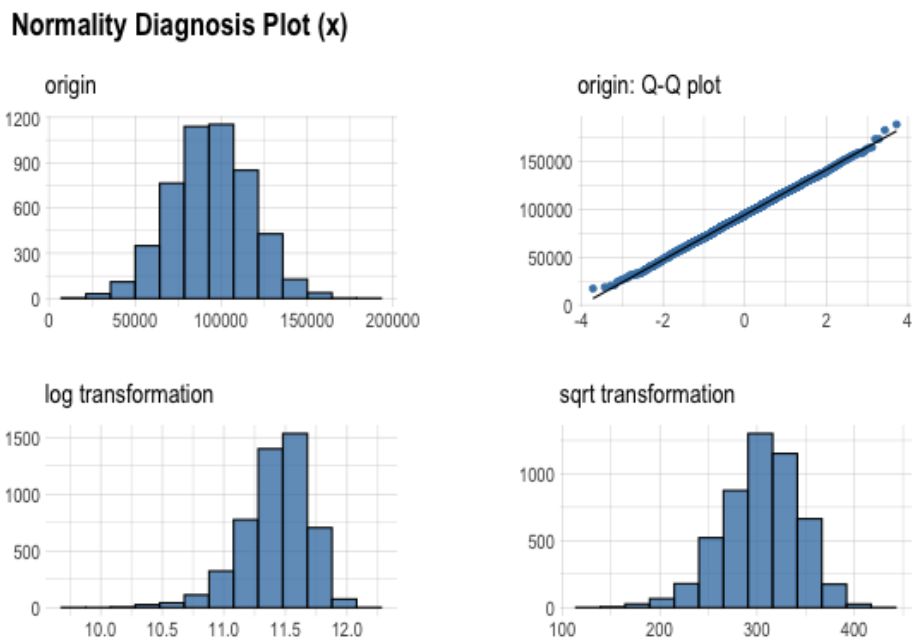

Figure 2.1: tot_credit_debt

avg_card_debt

* normality test : Shapiro-Wilk normality test
- statistic : 0.30203, p-value : 1.58167E-87

Table 2.2: skewness and kurtosis : avg_card_debt

type	skewness	kurtosis
original	8.6894	85.0836
log transformation	2.4857	21.3273
sqrt transformation	6.0134	53.6712

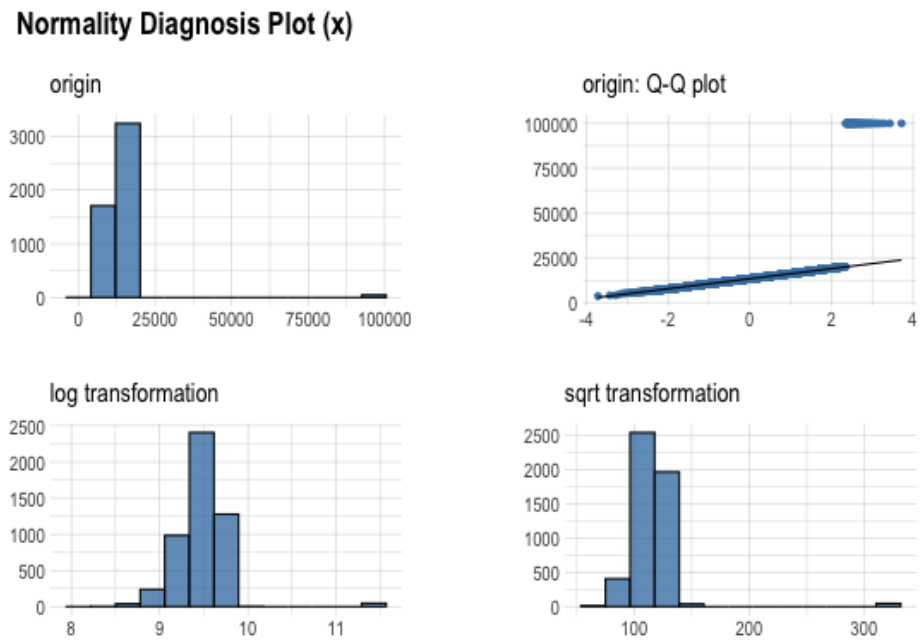

Figure 2.2: avg_card_debt

credit_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.99938, p-value : 0.0864133

Table 2.3: skewness and kurtosis : credit_age

type	skewness	kurtosis
original	0.0760	2.9650
log transformation	-0.6702	4.3595
sqrt transformation	-0.2594	3.2277

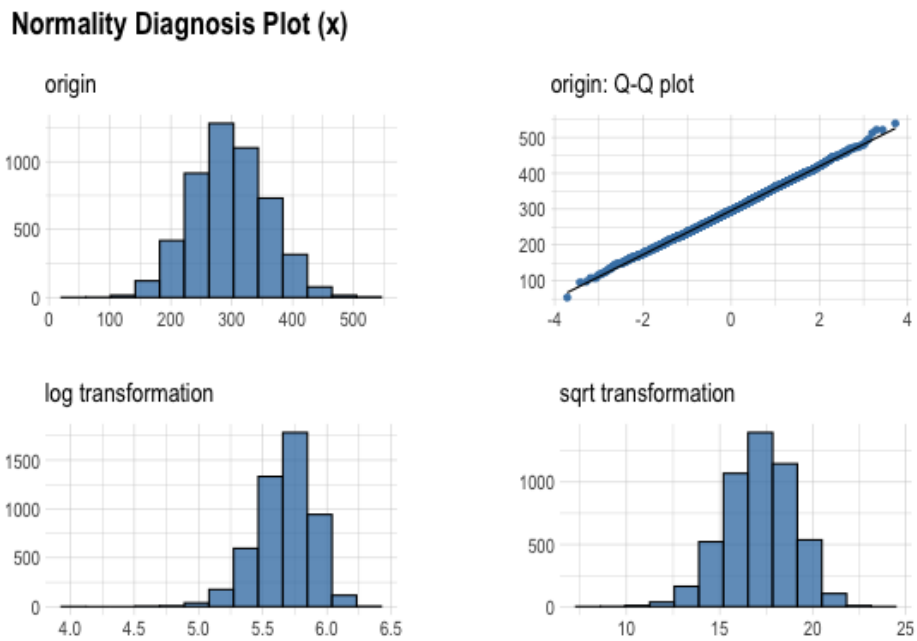

Figure 2.3: credit_age

credit_good_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.99921, p-value : 0.0235765

Table 2.4: skewness and kurtosis : credit_good_age

type	skewness	kurtosis
original	0.0018	3.1086
log transformation	-0.9587	5.5050
sqrt transformation	-0.4117	3.5774

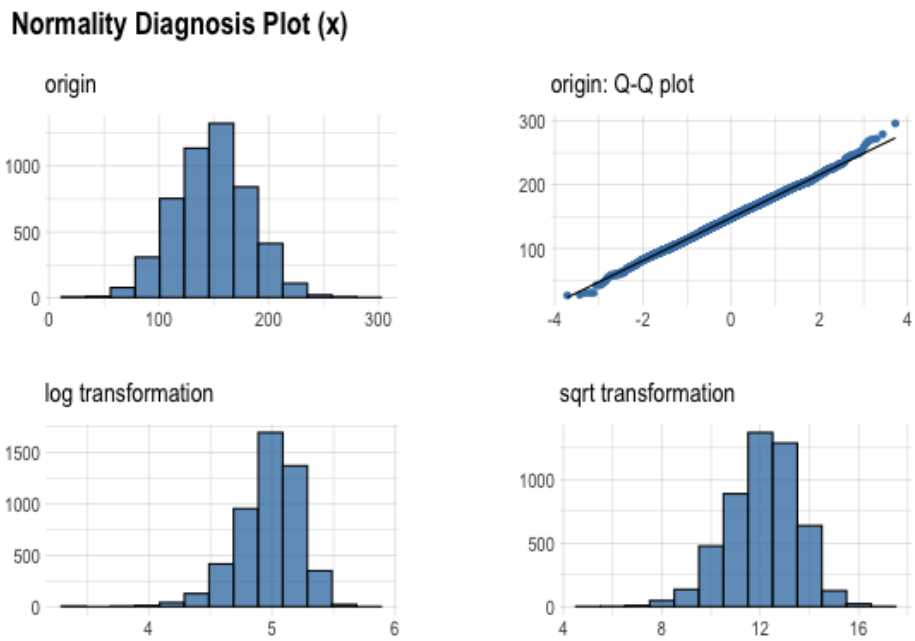

Figure 2.4: credit_good_age

card_age

* normality test : Shapiro-Wilk normality test
- statistic : 0.9996, p-value : 0.408364

Table 2.5: skewness and kurtosis : card_age

type	skewness	kurtosis
original	0.0206	3.0307
log transformation	-0.8108	4.5176
sqrt transformation	-0.3574	3.3755

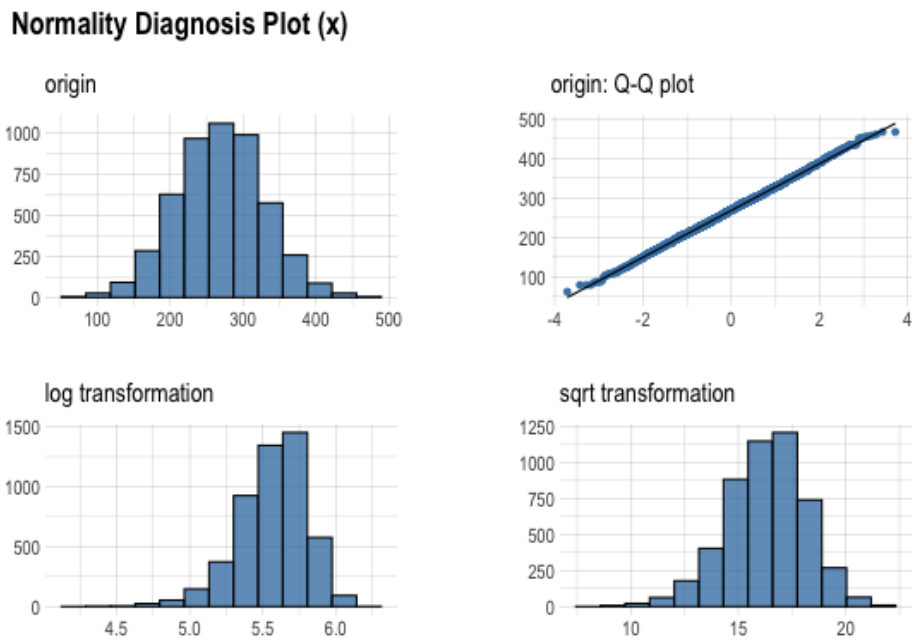

Figure 2.5: card_age

credit_past_due_amount

* normality test : Shapiro-Wilk normality test
- statistic : 0.13549, p-value : 1.95586E-92

Table 2.6: skewness and kurtosis : credit_past_due_amount

type	skewness	kurtosis
original	7.8391	71.6313
log+1 transformation	5.8579	35.4717
sqrt transformation	6.4162	44.3816

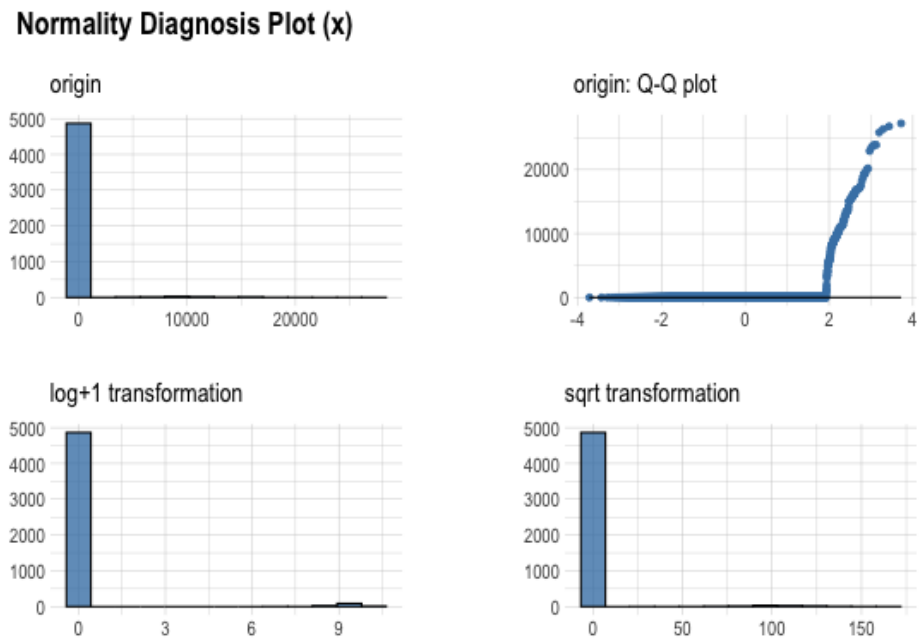

Figure 2.6: credit_past_due_amount

inq_12_month_num

* normality test : Shapiro-Wilk normality test
- statistic : 0.87243, p-value : 2.14514E-53

Table 2.7: skewness and kurtosis : inq_12_month_num

type	skewness	kurtosis
original	0.8487	3.0990
log+1 transformation	0.0139	1.6455
sqrt transformation	-0.0868	1.6673

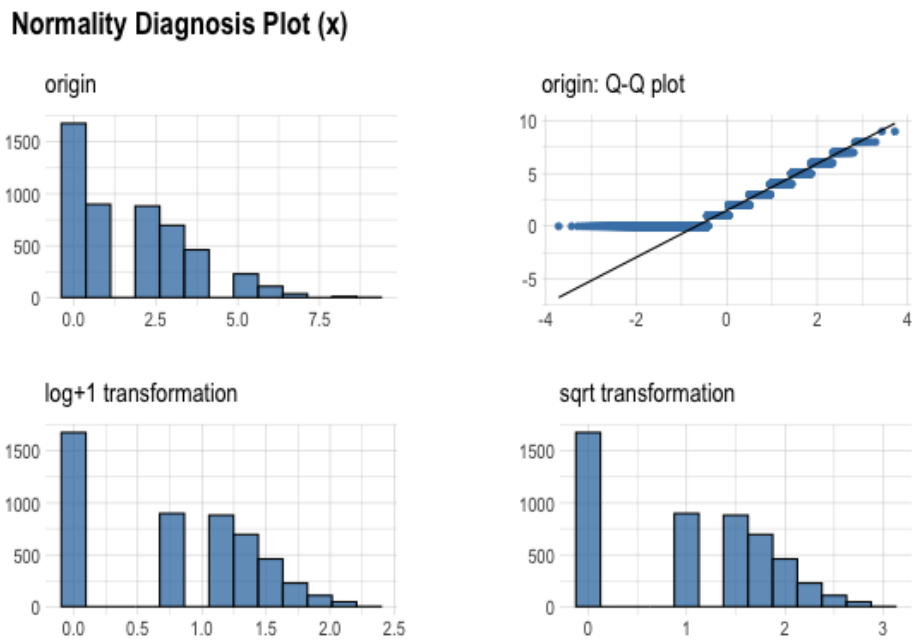

Figure 2.7: inq_12_month_num

card_inq_24_month_num

* normality test : Shapiro-Wilk normality test
- statistic : 0.92051, p-value : 1.99831E-45

Table 2.8: skewness and kurtosis : card_inq_24_month_num

type	skewness	kurtosis
original	0.7914	3.2630
log+1 transformation	-0.3682	2.0017
sqrt transformation	-0.3254	2.1888

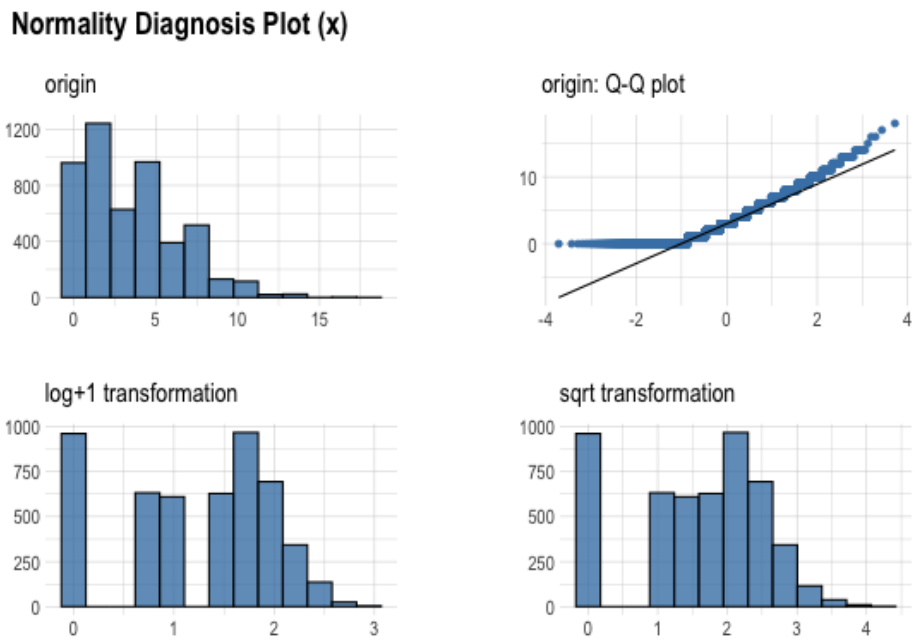

Figure 2.8: card_inq_24_month_num

uti_card

* normality test : Shapiro-Wilk normality test
- statistic : 0.9997, p-value : 0.691747

Table 2.9: skewness and kurtosis : uti_card

type	skewness	kurtosis
original	-0.0310	3.0784
log transformation	-1.0056	6.1384
sqrt transformation	-0.4354	3.6879

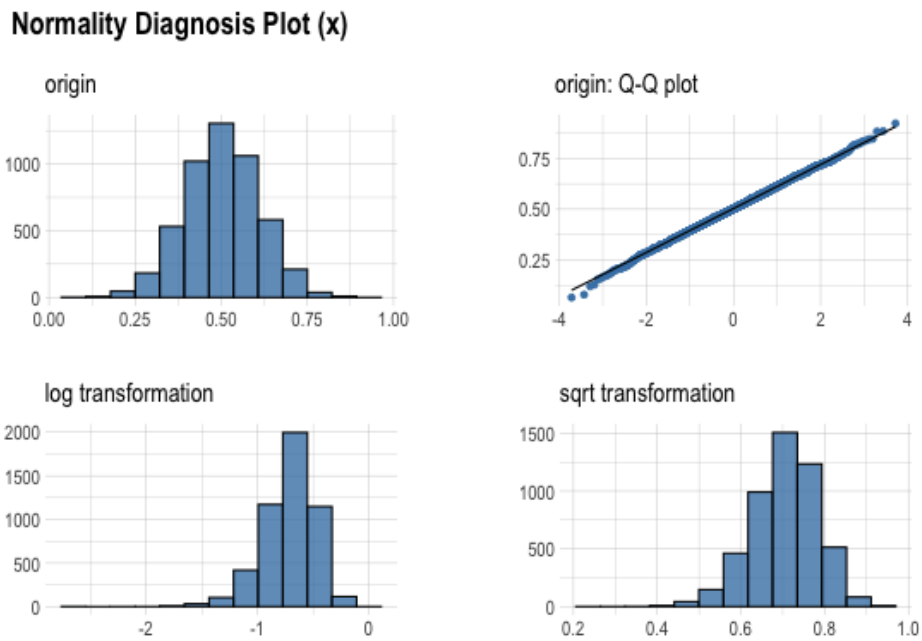

Figure 2.9: uti_card

uti_50plus_pct

* normality test : Shapiro-Wilk normality test
- statistic : 0.99919, p-value : 0.0194484

Table 2.10: skewness and kurtosis : uti_50plus_pct

type	skewness	kurtosis
original	0.0101	3.1087
log transformation	-1.3102	10.5233
sqrt transformation	-0.4412	4.0394

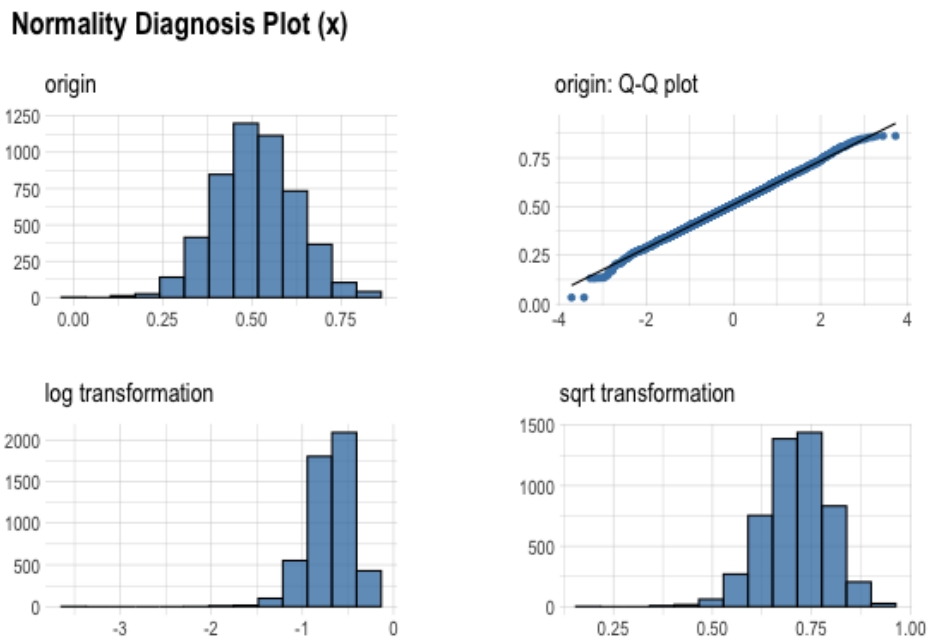

Figure 2.10: uti_50plus_pct

uti_max_credit_line

* normality test : Shapiro-Wilk normality test
- statistic : 0.99974, p-value : 0.831569

Table 2.11: skewness and kurtosis : uti_max_credit_line

type	skewness	kurtosis
original	0.0254	2.9725
log transformation	-0.7417	4.1567
sqrt transformation	-0.3293	3.2562

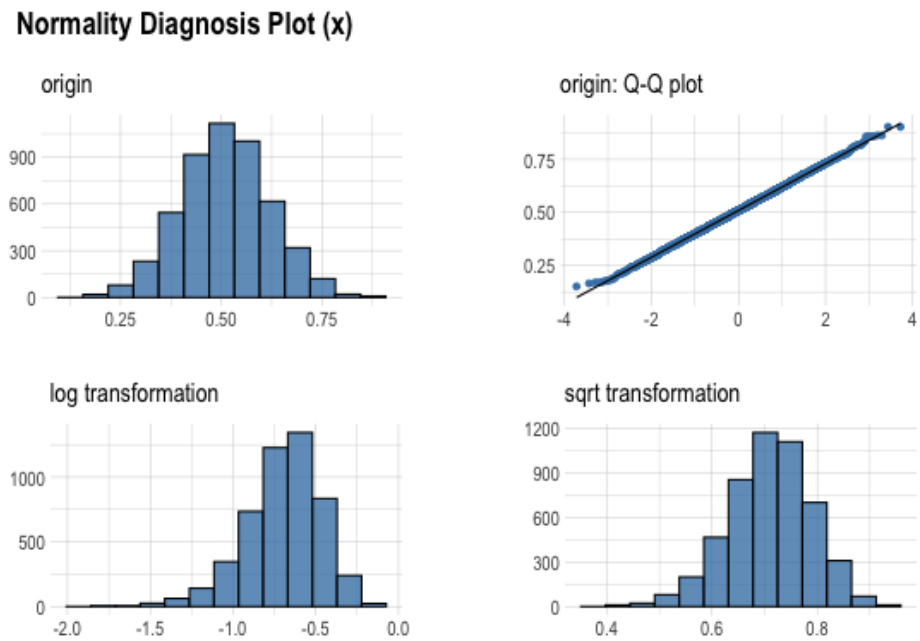

Figure 2.11: uti_max_credit_line

uti_card_50plus_pct

* normality test : Shapiro-Wilk normality test
- statistic : 0.99972, p-value : 0.764923

Table 2.12: skewness and kurtosis : uti_card_50plus_pct

type	skewness	kurtosis
original	-0.0177	3.0390
log transformation	-1.7945	17.9181
sqrt transformation	-0.5141	4.0357

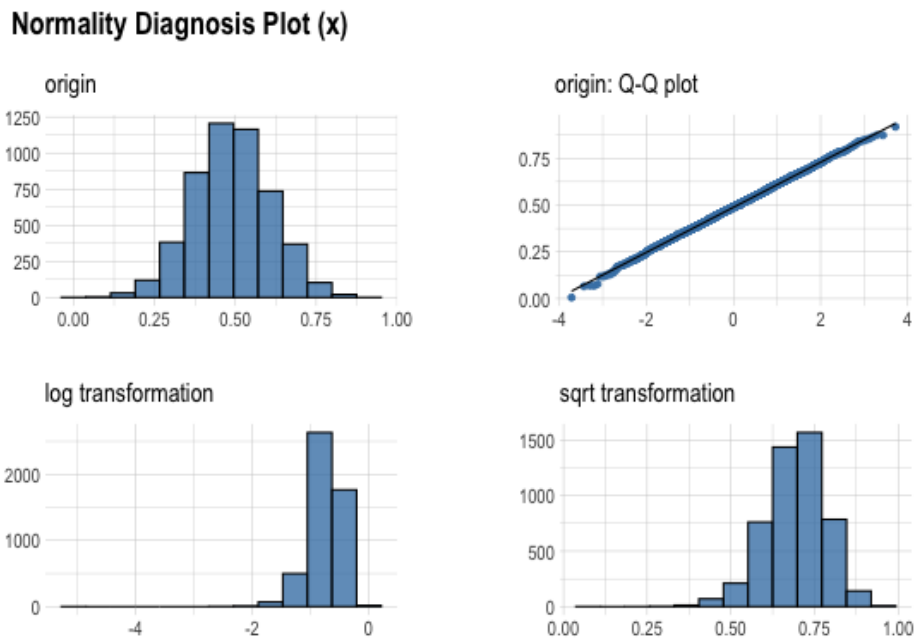

Figure 2.12: uti_card_50plus_pct

rep_income

* normality test : Shapiro-Wilk normality test
- statistic : 0.99926, p-value : 0.0350482

Table 2.13: skewness and kurtosis : rep_income

type	skewness	kurtosis
original	-0.0181	2.8906
log transformation	-0.7722	4.2830
sqrt transformation	-0.3610	3.2528

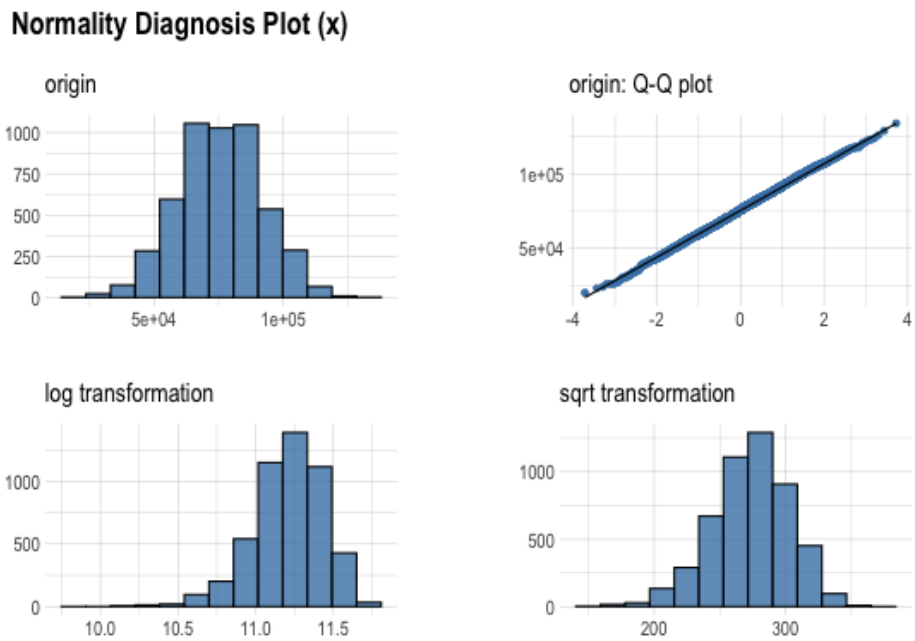

Figure 2.13: rep_income

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

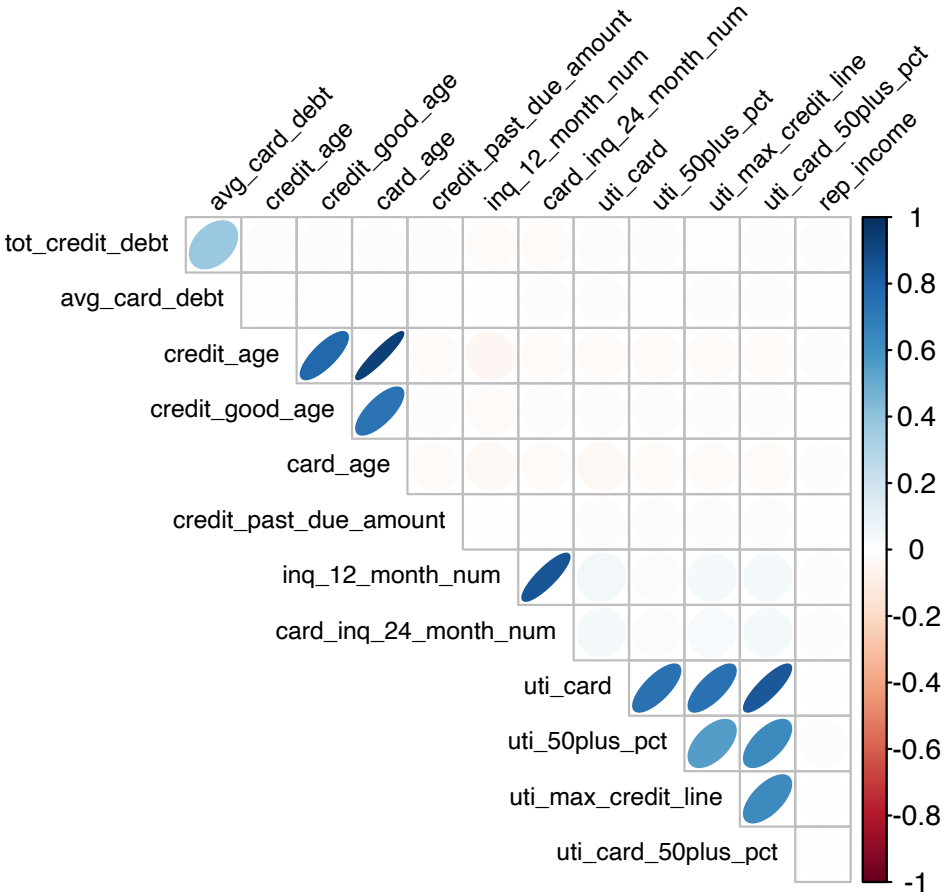

Figure 3.1: The correlation coefficient of numerical variables

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