



REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.12

December 16, 2019

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	29
3.1	Correlation Coefficient	29
3.1.1	Correlation Coefficient by Variable Combination	29
3.1.2	Correlation Plot of Numerical Variables	29
4	Target based Analysis	31
4.1	Grouped Descriptive Statistics	31
4.1.1	Grouped Numerical Variables	31
4.1.2	Grouped Categorical Variables	31
4.2	Grouped Relationship Between Variables	31
4.2.1	Grouped Correlation Coefficient	31
4.2.2	Grouped Correlation Plot of Numerical Variables	31

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 28,534 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
idcode	numeric	0	0.0000000	4711	0.1651013
year	numeric	0	0.0000000	15	0.0005257
birth_yr	numeric	0	0.0000000	14	0.0004906
age	numeric	24	0.0841102	34	0.0011916
race	numeric	0	0.0000000	3	0.0001051
msp	numeric	16	0.0560735	3	0.0001051
nev_mar	numeric	16	0.0560735	3	0.0001051
grade	numeric	2	0.0070092	20	0.0007009
collgrad	numeric	0	0.0000000	2	0.0000701
not_smsa	numeric	8	0.0280367	3	0.0001051
c_city	numeric	8	0.0280367	3	0.0001051
south	numeric	8	0.0280367	3	0.0001051
ind_code	numeric	341	1.1950655	13	0.0004556
occ_code	numeric	121	0.4240555	14	0.0004906
union	numeric	9296	32.5786781	3	0.0001051
wks_ue	numeric	5704	19.9901871	62	0.0021728
ttl_exp	numeric	0	0.0000000	4744	0.1662578
tenure	numeric	433	1.5174879	271	0.0094974
hours	numeric	67	0.2348076	86	0.0030139
wks_work	numeric	703	2.4637275	106	0.0037149
ln_wage	numeric	0	0.0000000	8173	0.2864302

The target variable of the data is 'NULL', and the data type of the variable is NULL(You did not specify a

target variable).

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

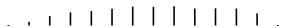
Univariate Analysis


2.1 Descriptive Statistics


21 Variables edaData
28534 Observations

idcode : NLS ID Format:%8.0g 
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
28534 0 4711 1 2601 1717 259.7 518.0 1327.0 2606.0 3881.0 4656.0 4889.0
lowest : 1 2 3 4 5, highest: 5155 5156 5157 5158 5159

year : interview year Format:%8.0g 
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
28534 0 15 0.995 77.96 7.339 69 70 72 78 83 87 88
Value 68 69 70 71 72 73 75 77 78 80 82 83 85 87
Frequency 1375 1232 1686 1851 1693 1981 2141 2171 1964 1847 2085 1987 2085 2164
Proportion 0.048 0.043 0.059 0.065 0.059 0.069 0.075 0.076 0.069 0.065 0.073 0.070 0.073 0.076
Value 88
Frequency 2272
Proportion 0.080

birth_yr : birth year Format:%8.0g 
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
28534 0 14 0.991 48.09 3.455 43 44 46 48 51 52 53
Value 41 42 43 44 45 46 47 48 49 50 51 52 53 54
Frequency 26 574 1522 2095 2311 2707 3040 3017 3095 2718 2765 2722 1935 7
Proportion 0.001 0.020 0.053 0.073 0.081 0.095 0.107 0.106 0.108 0.095 0.097 0.095 0.068 0.000

age : age in current year Format:%8.0g 
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95
28510 24 33 0.998 29.05 7.682 19 21 23 28 34 38 41
lowest : 14 15 16 17 18, highest: 42 43 44 45 46

race Format:%8.0g 
n missing distinct Info Mean Gmd
28534 0 3 0.624 1.303 0.4351
Value 1 2 3
Frequency 20180 8051 303
Proportion 0.707 0.282 0.011

msp : 1 if married, spouse present Format:%8.0g
n missing distinct Info Sum Mean Gmd
28518 16 2 0.718 17194 0.6029 0.4788

nev_mar : 1 if never married Format:%8.0g
n missing distinct Info Sum Mean Gmd
28518 16 2 0.531 6550 0.2297 0.3539

grade : current grade completed Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
28532	2	19	0.874	12.53	2.374	9	10	12	12	14	16	17

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	21	6	4	2	36	41	161	262	671	889	1518	1781	14252	1734
Proportion	0.001	0.000	0.000	0.000	0.001	0.001	0.006	0.009	0.024	0.031	0.053	0.062	0.500	0.061

Value	14	15	16	17	18
Frequency	1751	950	2681	851	921
Proportion	0.061	0.033	0.094	0.030	0.032

collgrad : 1 if college graduate Format:%8.0g

n	missing	distinct	Info	Sum	Mean	Gmd
28534	0	2	0.419	4795	0.168	0.2796

not_smsa : 1 if not SMSA Format:%8.0g

n	missing	distinct	Info	Sum	Mean	Gmd
28526	8	2	0.608	8057	0.2824	0.4054

c_city : 1 if central city Format:%8.0g

n	missing	distinct	Info	Sum	Mean	Gmd
28526	8	2	0.689	10190	0.3572	0.4592

south : 1 if south Format:%8.0g

n	missing	distinct	Info	Sum	Mean	Gmd
28526	8	2	0.725	11683	0.4096	0.4837

ind_code : industry of employment Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
28193	341	12	0.957	7.693	3.355	4	4	5	7	11	11	12

Value	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	241	52	252	5845	1420	4952	2427	849	1712	215	8480	1748
Proportion	0.009	0.002	0.009	0.207	0.050	0.176	0.086	0.030	0.061	0.008	0.301	0.062

occ_code : occupation Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
28413	121	13	0.934	4.778	3.225	1	1	3	3	6	8	13

Value	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	3008	1494	10974	1323	438	4309	571	4300	6	144	194	7	1645
Proportion	0.106	0.053	0.386	0.047	0.015	0.152	0.020	0.151	0.000	0.005	0.007	0.000	0.058

union : 1 if union Format:%8.0g

n	missing	distinct	Info	Sum	Mean	Gmd
19238	9296	2	0.538	4510	0.2344	0.359

wks_ue : weeks unemployed last year Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
22830	5704	61	0.558	2.548	4.537	0	0	0	0	0	8	17

lowest : 0 1 2 3 4, highest: 56 62 73 75 76

ttl_exp : total work experience Format:%9.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50
28534	0	4744	1	6.215	5.147	0.6667	1.0385	2.4615	5.0577
.75	.90	.95							
9.1282	13.2801	15.3269							

lowest : 0.00000000 0.01923077 0.03846154 0.05769231 0.05769231
highest: 26.53846169 26.84615135 27.19230461 27.46153831 28.88461494

tenure : job tenure, in years Format:%9.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50
28101	433	270	1	3.124	3.638	0.08333	0.16667	0.50000	1.66667
.75	.90	.95							
4.16667	8.41667	11.41667							

lowest : 0.00000000 0.08333334 0.16666667 0.25000000 0.33333334
highest: 23.08333397 23.33333397 24.50000000 24.75000000 25.91666603

hours : usual hours worked Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
28467	67	85	0.842	36.56	9.175	15	20	35	40	40	44	48

lowest : 1 2 3 4 5, highest: 99 100 105 112 168

wks_work : weeks worked last year Format:%8.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
27831	703	105	0.996	53.99	32.48	6	14	36	52	72	98	104

lowest : 0 1 2 3 4, highest: 100 101 102 103 104

ln_wage : ln(wage/GNP deflator) Format:%9.0g

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
28534	0	8173	1	1.675	0.5237	0.9928	1.1661	1.3615	1.6405	1.9641	2.2757	2.4562

lowest : 0.000000000 0.004487075 0.004939650 0.008032188 0.017654561
highest: 4.349081993 4.349225998 4.499809742 4.828313828 5.263916016

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

idcode

normality test : Shapiro-Wilk normality test
 statistic : 0.95577, p-value : 1.79068E-36

type	skewness	kurtosis
original	-0.0197	1.8137
log transformation	-2.3132	11.3522
sqrt transformation	-0.6023	2.4766

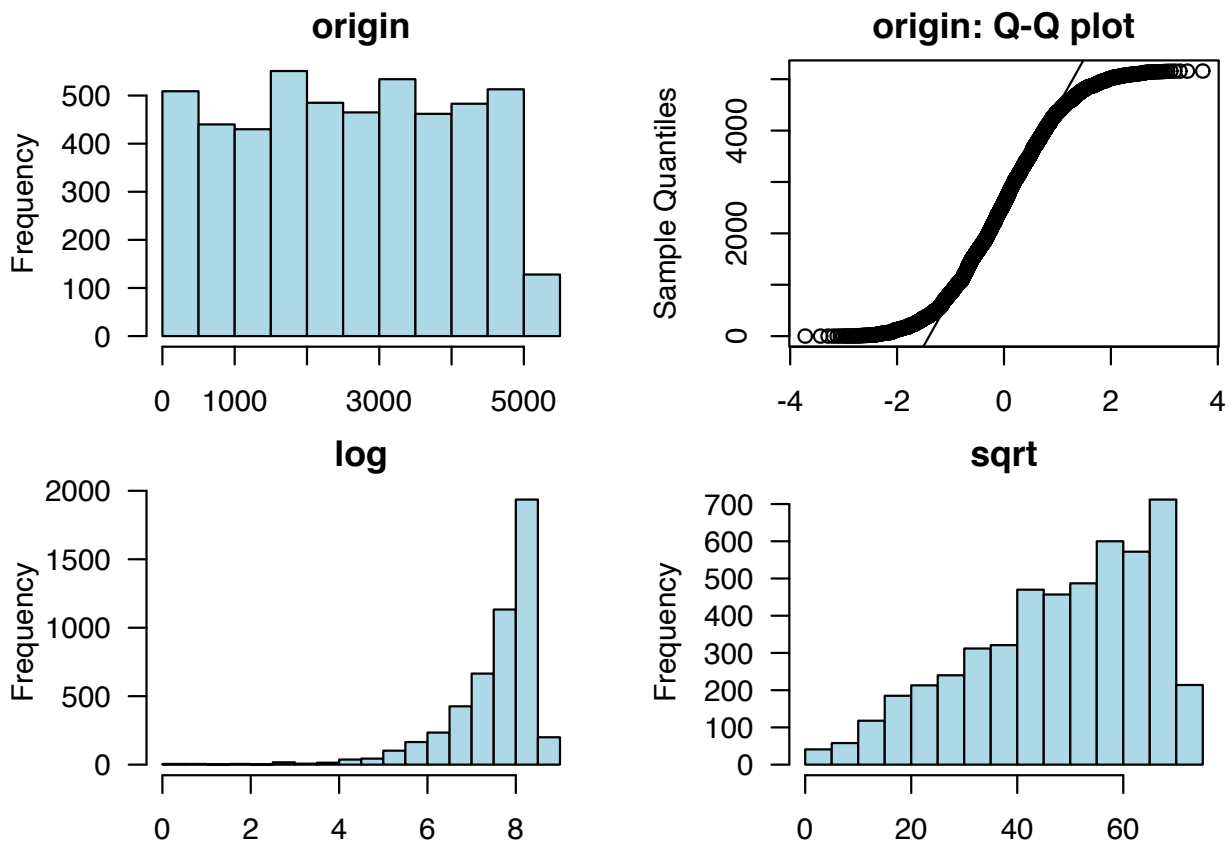


Figure 2.1: idcode

year

normality test : Shapiro-Wilk normality test
 statistic : 0.93183, p-value : 5.56401E-43

type	skewness	kurtosis
original	0.0688	1.6982
log transformation	-0.0160	1.6958
sqrt transformation	0.0264	1.6950

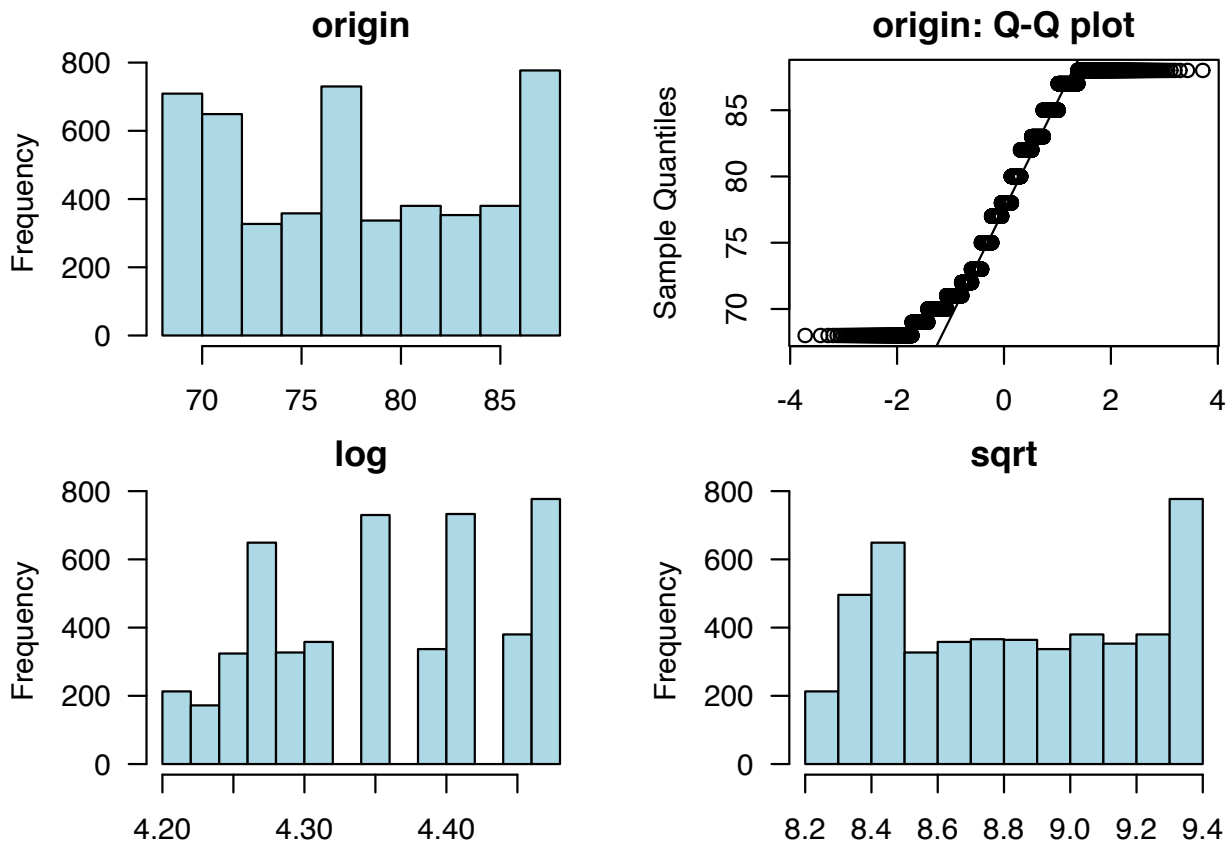


Figure 2.2: year

birth_yr

normality test : Shapiro-Wilk normality test
 statistic : 0.96165, p-value : 1.88666E-34

type	skewness	kurtosis
original	-0.1206	2.0355
log transformation	-0.2185	2.0924
sqrt transformation	-0.1693	2.0610

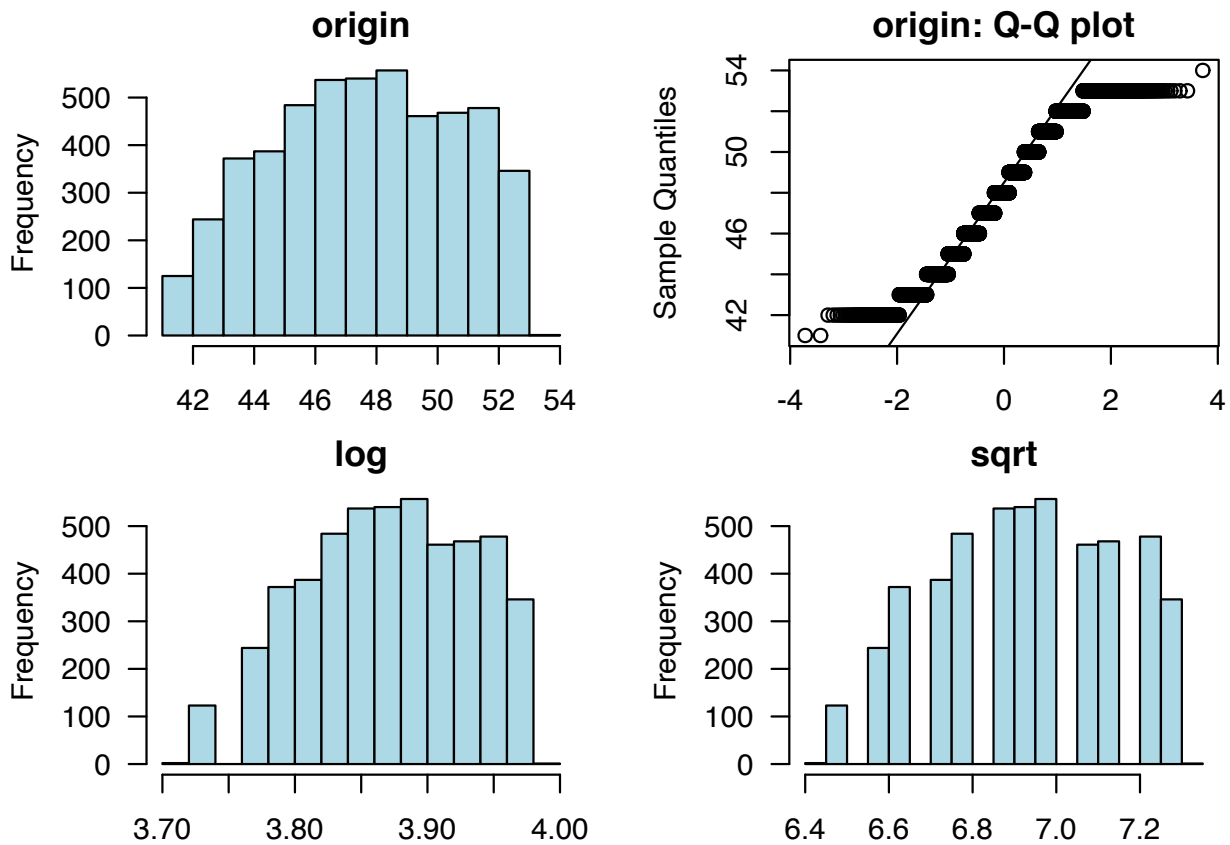


Figure 2.3: birth_yr

age

normality test : Shapiro-Wilk normality test
 statistic : 0.97039, p-value : 6.19944E-31

type	skewness	kurtosis
original	0.2225	2.0776
log transformation	-0.1337	2.0386
sqrt transformation	0.0454	2.0154

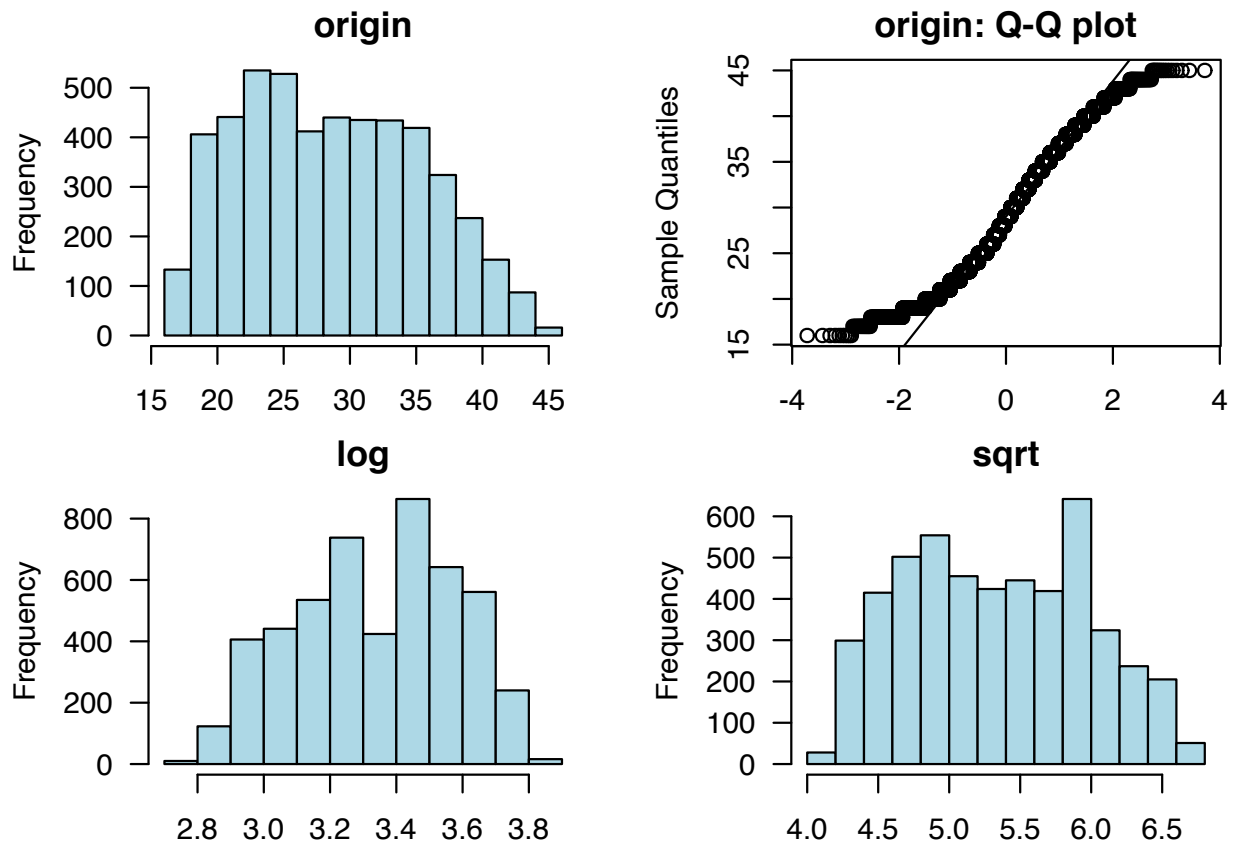


Figure 2.4: age

msp

normality test : Shapiro-Wilk normality test
 statistic : 0.61745, p-value : 1.8183E-74

type	skewness	kurtosis
original	-0.4683	1.2193
log transformation		
sqrt transformation	-0.4683	1.2193

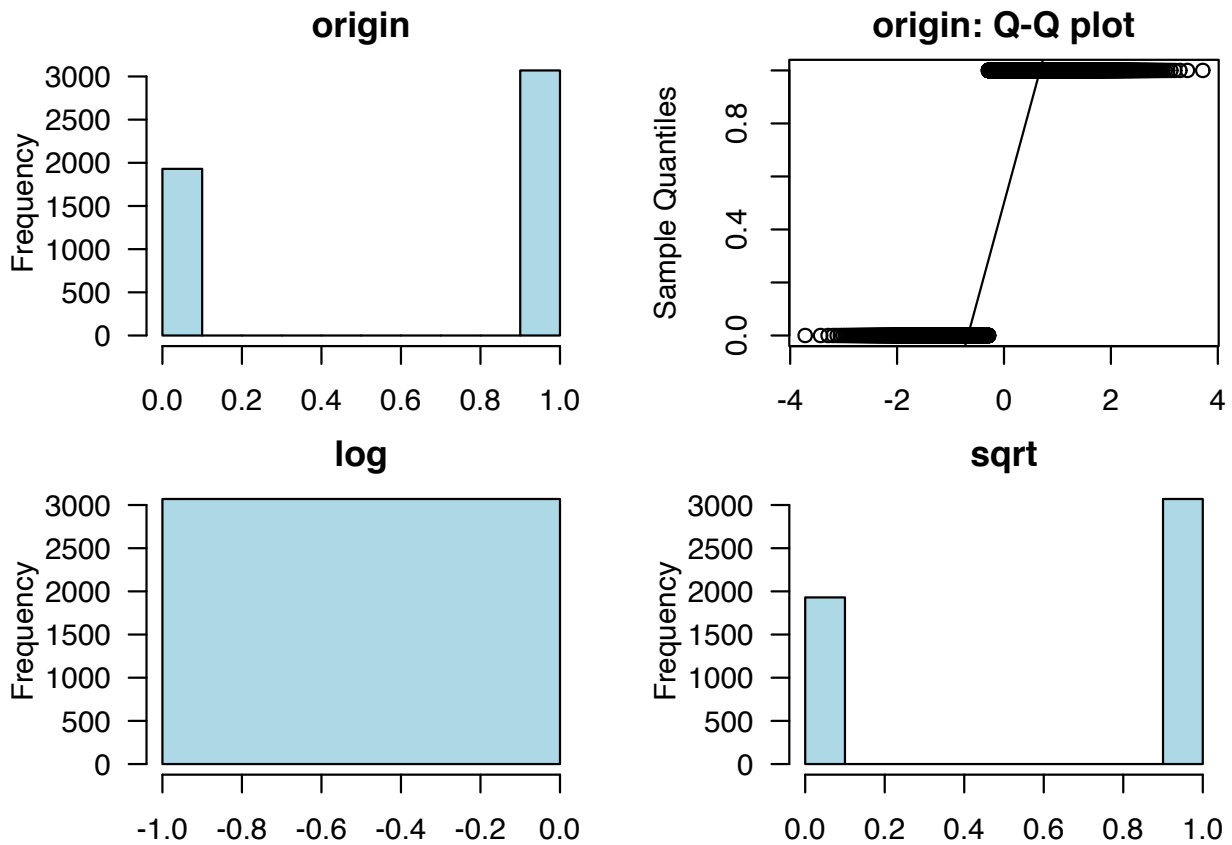


Figure 2.5: msp

nev_mar

normality test : Shapiro-Wilk normality test
 statistic : 0.5197, p-value : 2.81181E-79

type	skewness	kurtosis
original	1.2896	2.6630
log transformation		
sqrt transformation	1.2896	2.6630

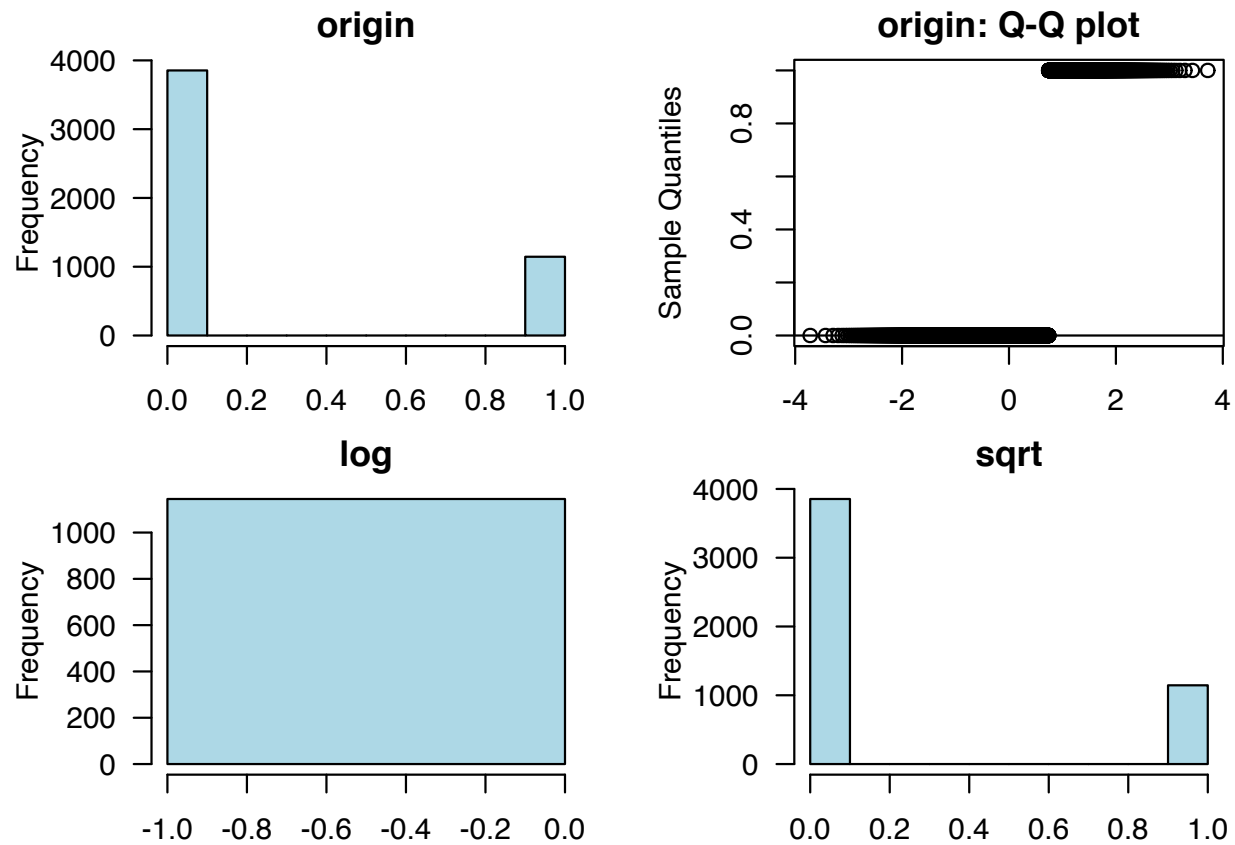


Figure 2.6: nev_mar

grade

normality test : Shapiro-Wilk normality test
 statistic : 0.87934, p-value : 2.01112E-52

type	skewness	kurtosis
original	0.1832	4.4419
log transformation		
sqrt transformation	-0.9099	12.0912

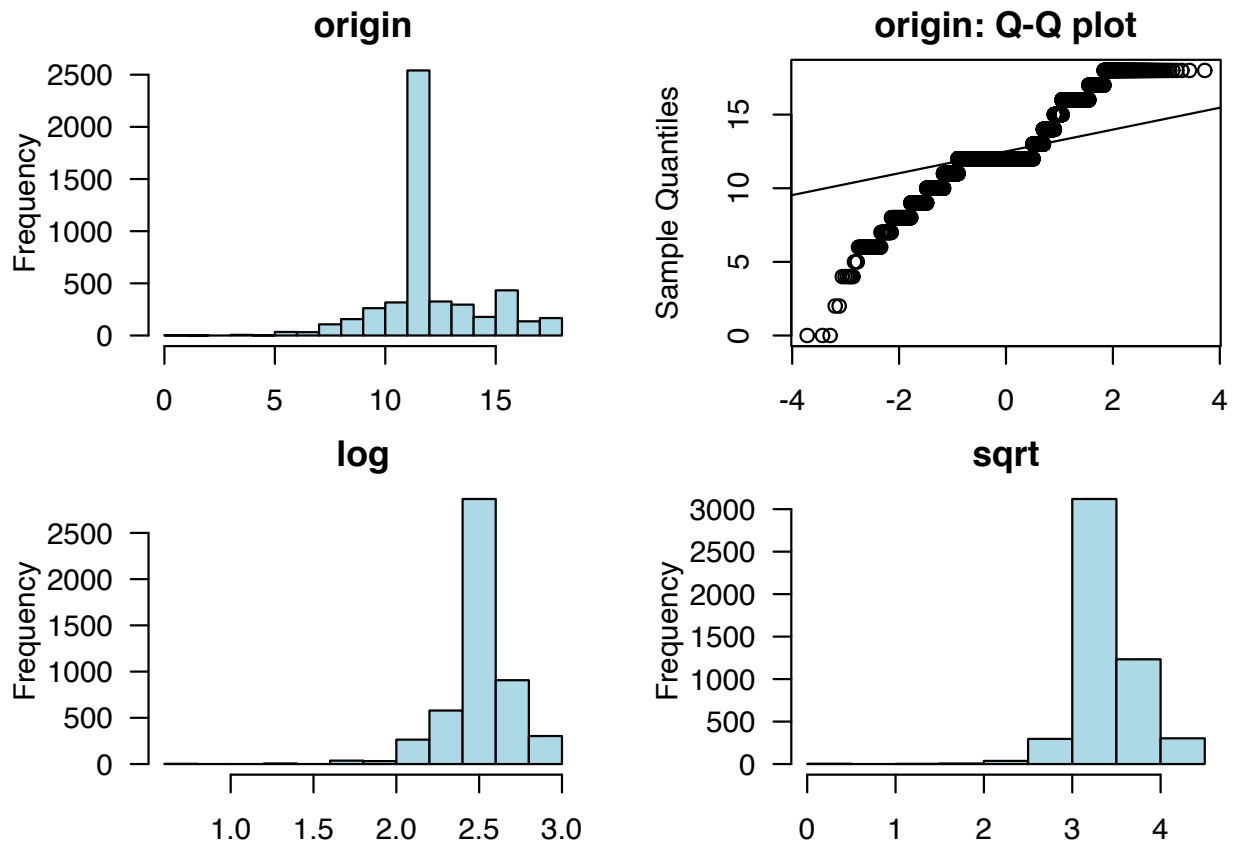


Figure 2.7: grade

collgrad

normality test : Shapiro-Wilk normality test
 statistic : 0.44481, p-value : 1.98196E-82

type	skewness	kurtosis
original	1.8228	4.3225
log transformation		
sqrt transformation	1.8228	4.3225

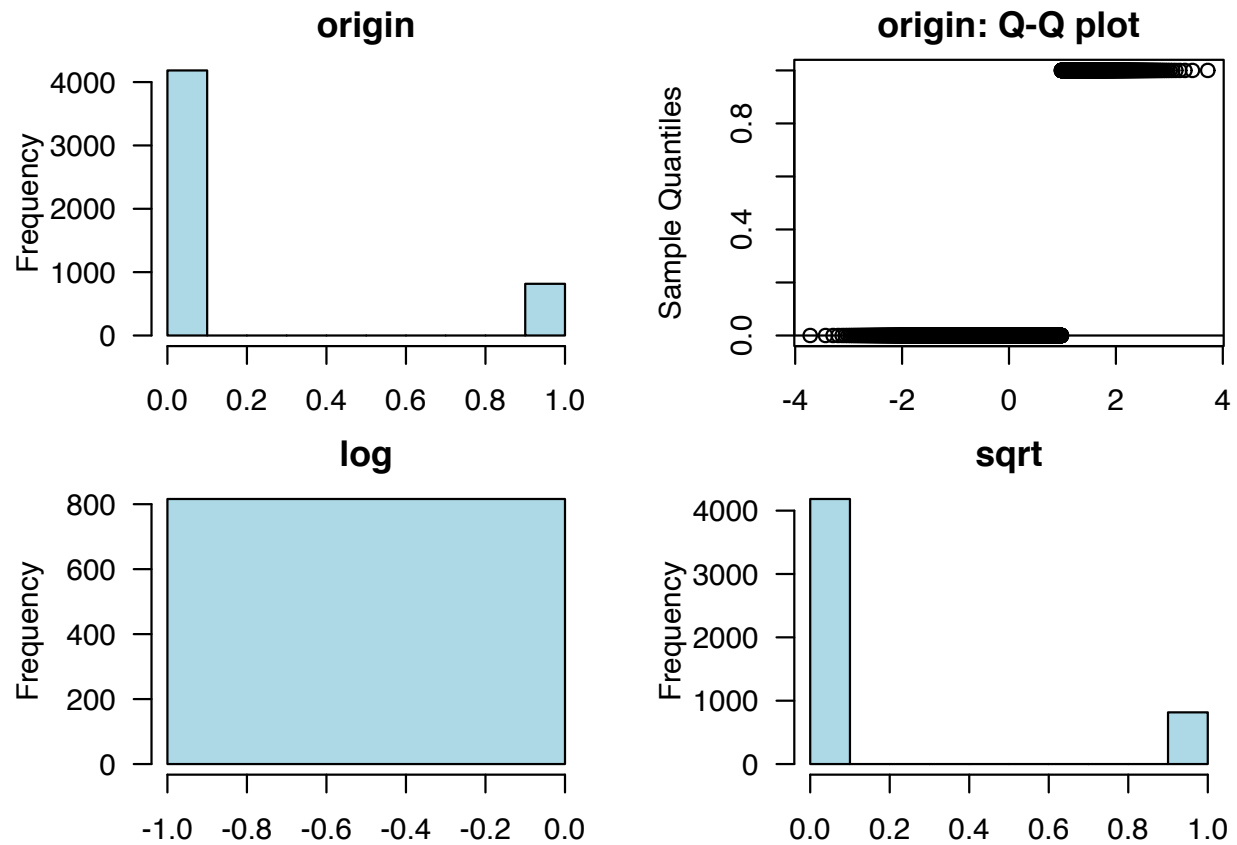


Figure 2.8: collgrad

not_smsa

normality test : Shapiro-Wilk normality test
 statistic : 0.5623, p-value : 2.73849E-77

type	skewness	kurtosis
original	0.9782	1.9569
log transformation		
sqrt transformation	0.9782	1.9569

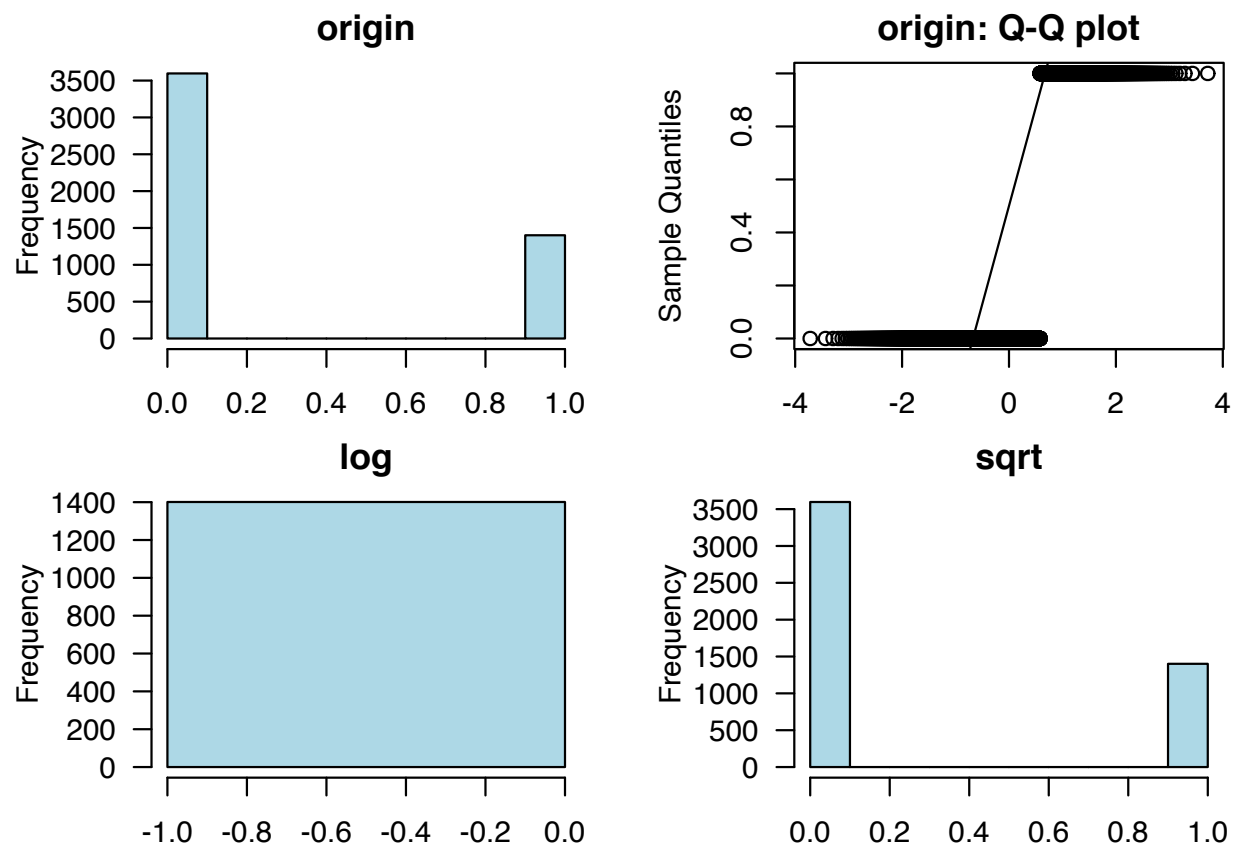


Figure 2.9: not_smsa

c_city

normality test : Shapiro-Wilk normality test
 statistic : 0.60303, p-value : 3.13554E-75

type	skewness	kurtosis
original	0.6292	1.3960
log transformation		
sqrt transformation	0.6292	1.3960

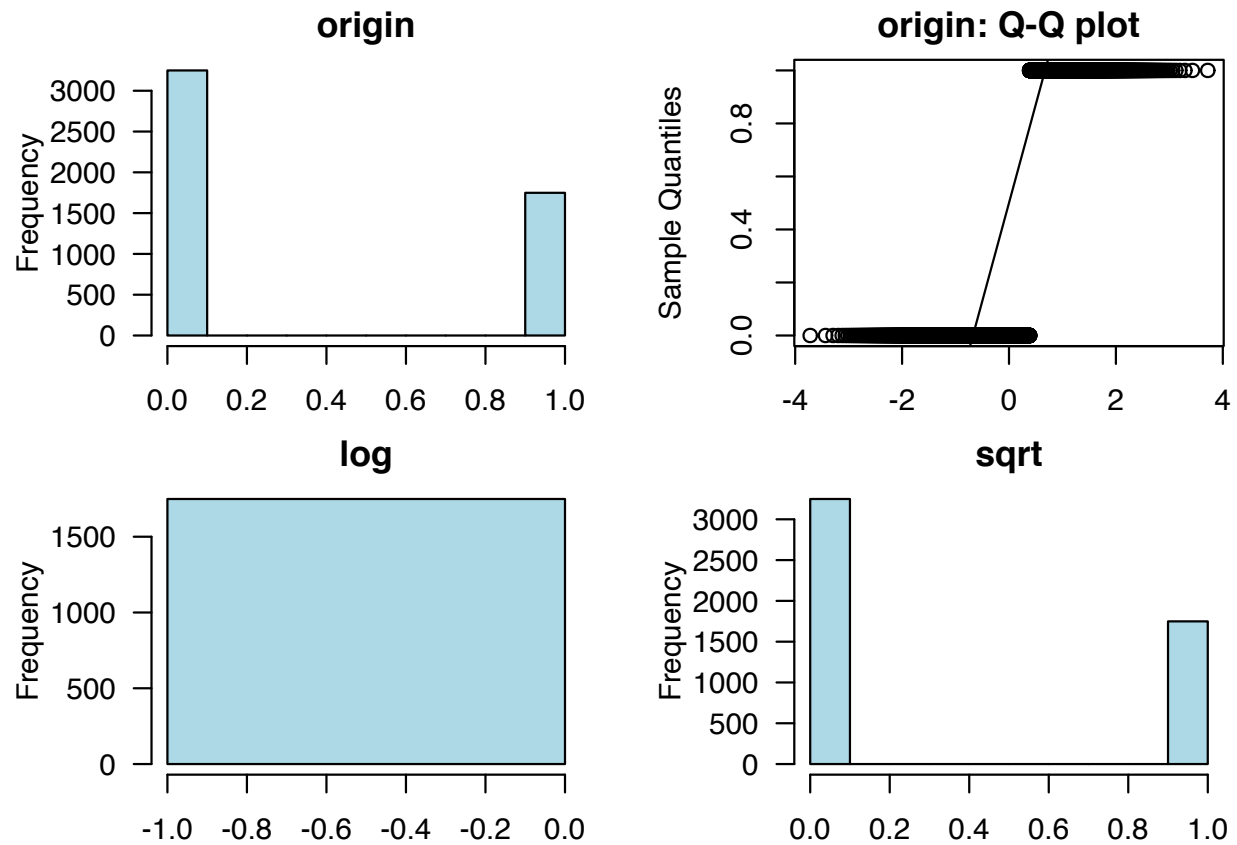


Figure 2.10: c_city

south

normality test : Shapiro-Wilk normality test
 statistic : 0.62199, p-value : 3.32655E-74

type	skewness	kurtosis
original	0.4072	1.1658
log transformation		
sqrt transformation	0.4072	1.1658

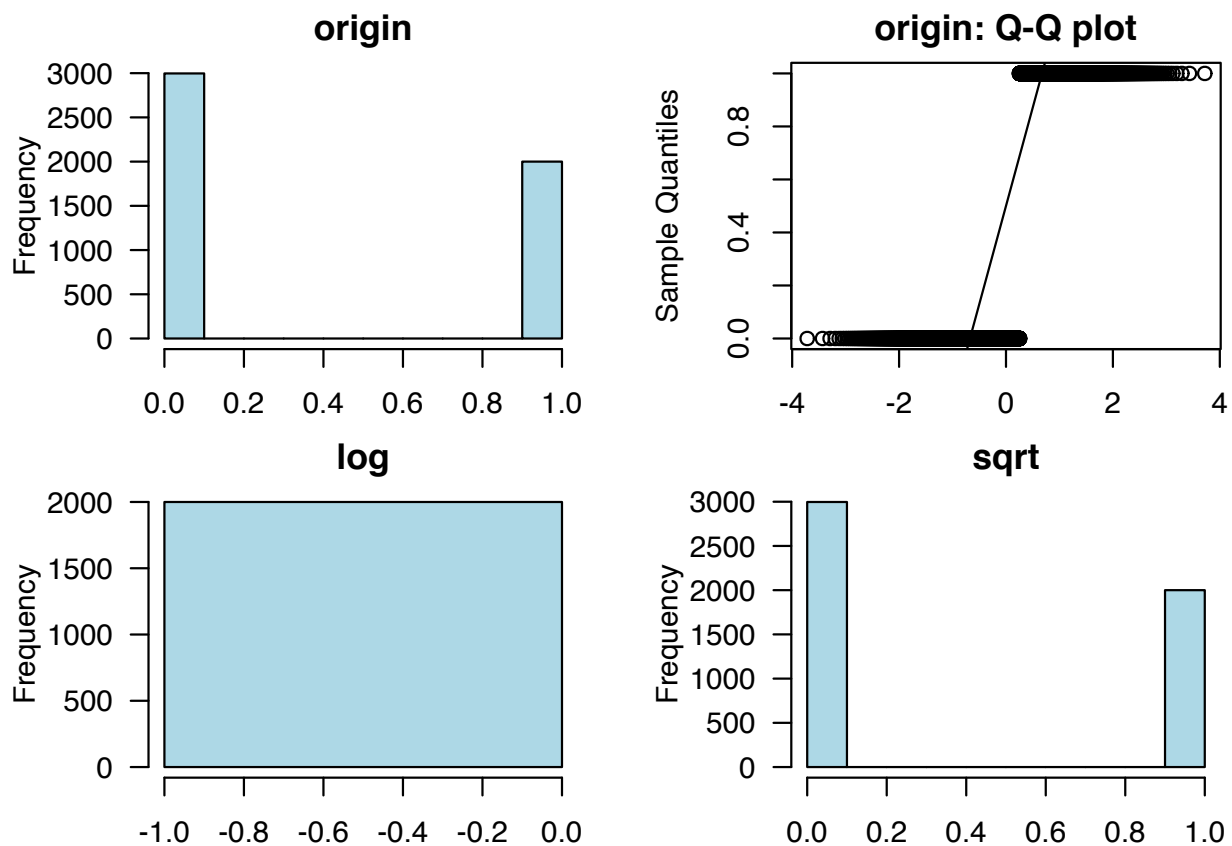


Figure 2.11: south

ind_code

normality test : Shapiro-Wilk normality test
 statistic : 0.86895, p-value : 1.12466E-53

type	skewness	kurtosis
original	-0.0091	1.5282
log transformation	-0.7807	4.0123
sqrt transformation	-0.2565	1.9775

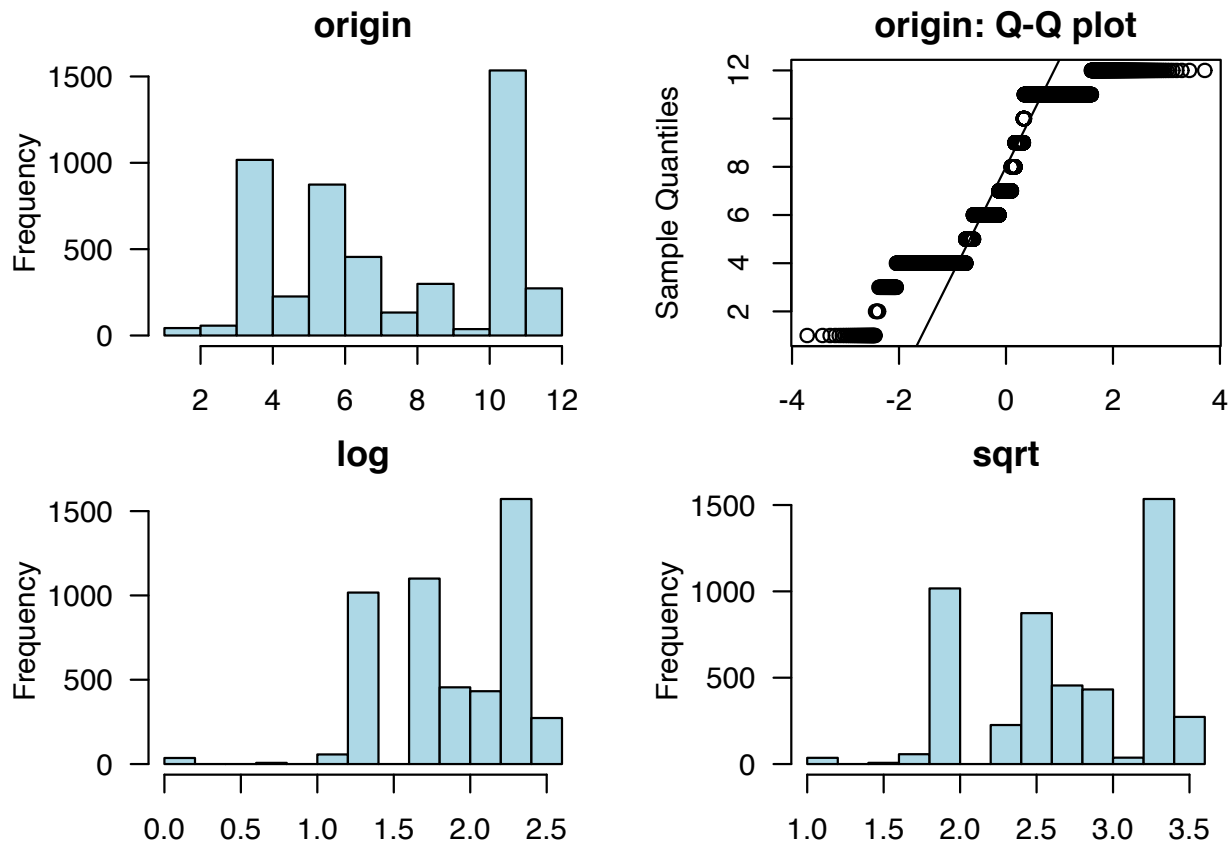


Figure 2.12: ind_code

occ_code

normality test : Shapiro-Wilk normality test
 statistic : 0.85431, p-value : 1.17692E-55

type	skewness	kurtosis
original	1.0725	3.6598
log transformation	-0.3061	2.6630
sqrt transformation	0.4364	2.6148

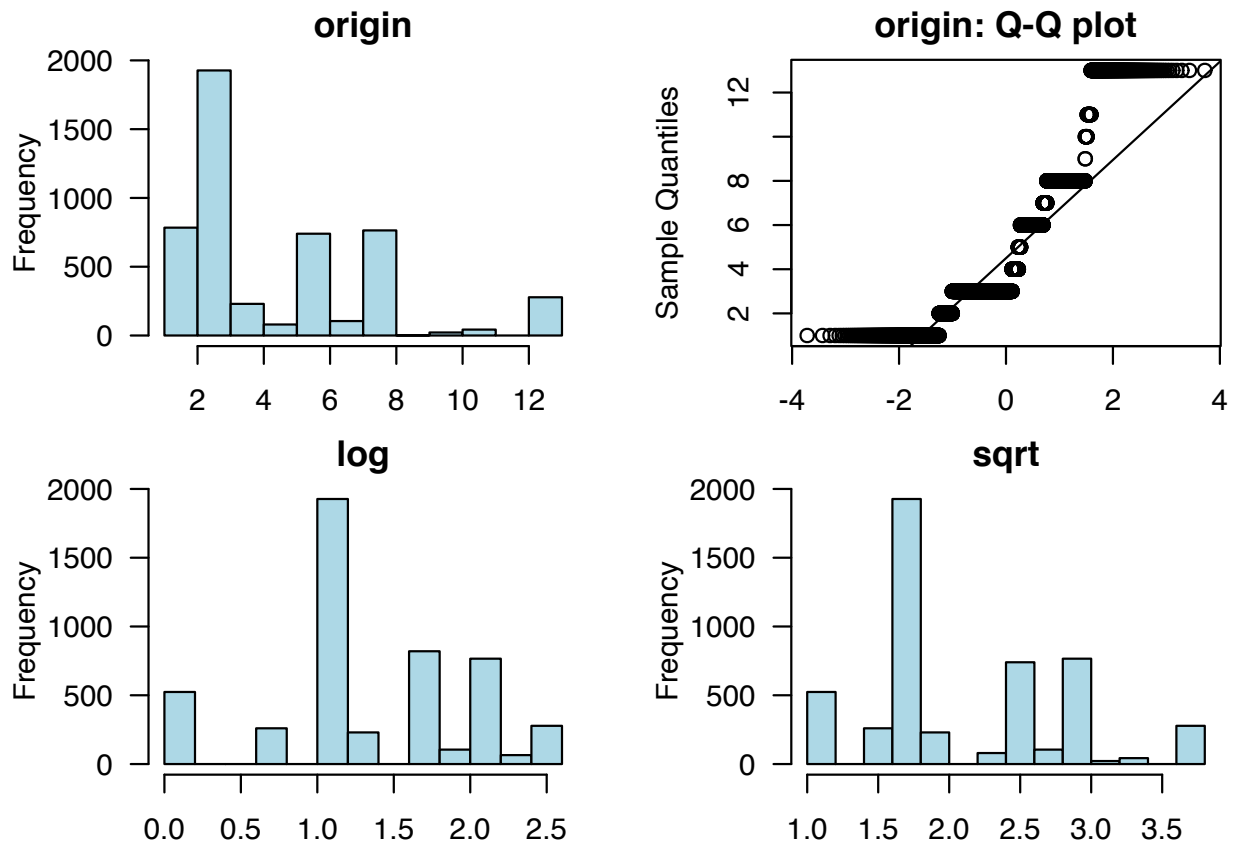


Figure 2.13: occ_code

union

normality test : Shapiro-Wilk normality test
 statistic : 0.52296, p-value : 6.61572E-70

type	skewness	kurtosis
original	1.2664	2.6038
log transformation		
sqrt transformation	1.2664	2.6038

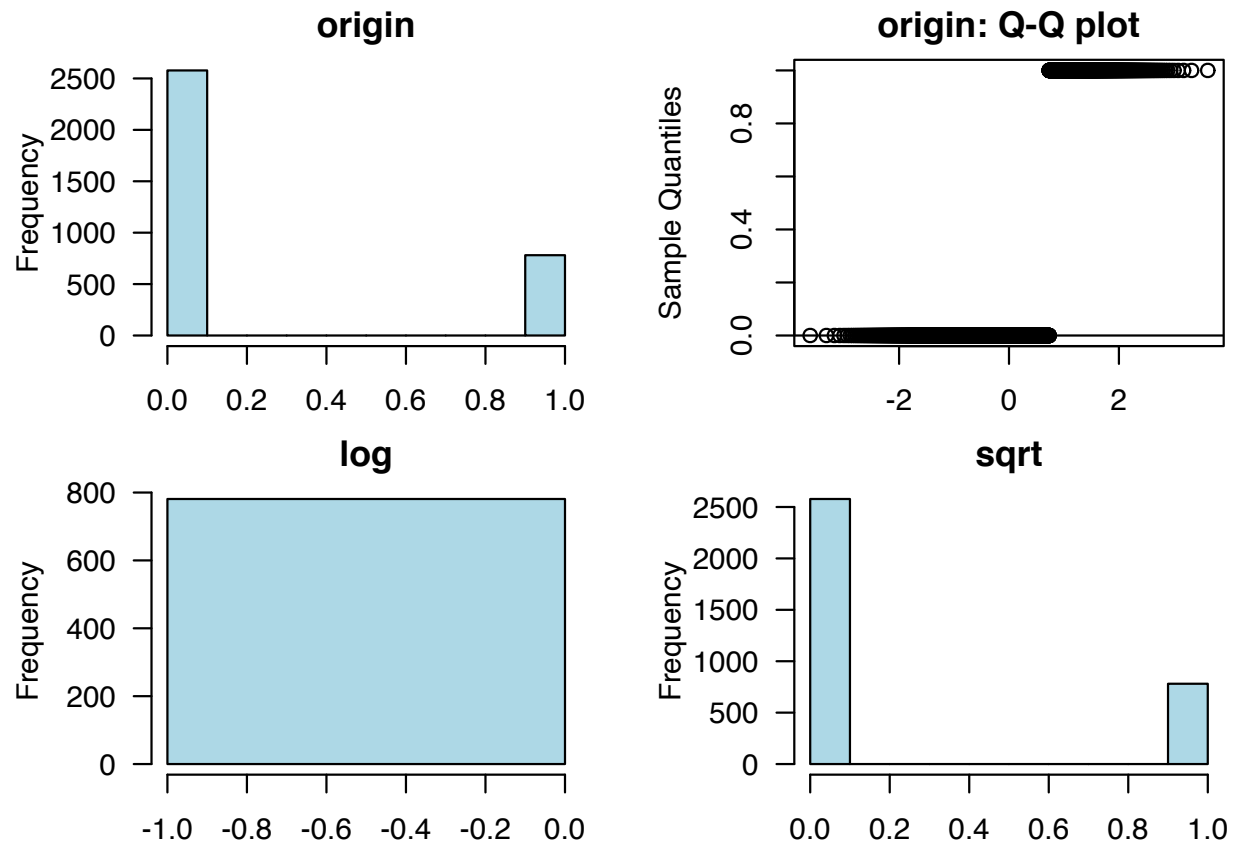


Figure 2.14: union

wks_ue

normality test : Shapiro-Wilk normality test
 statistic : 0.42001, p-value : 6.49128E-78

type	skewness	kurtosis
original	3.8091	19.3596
log transformation		
sqrt transformation	2.2365	7.3800

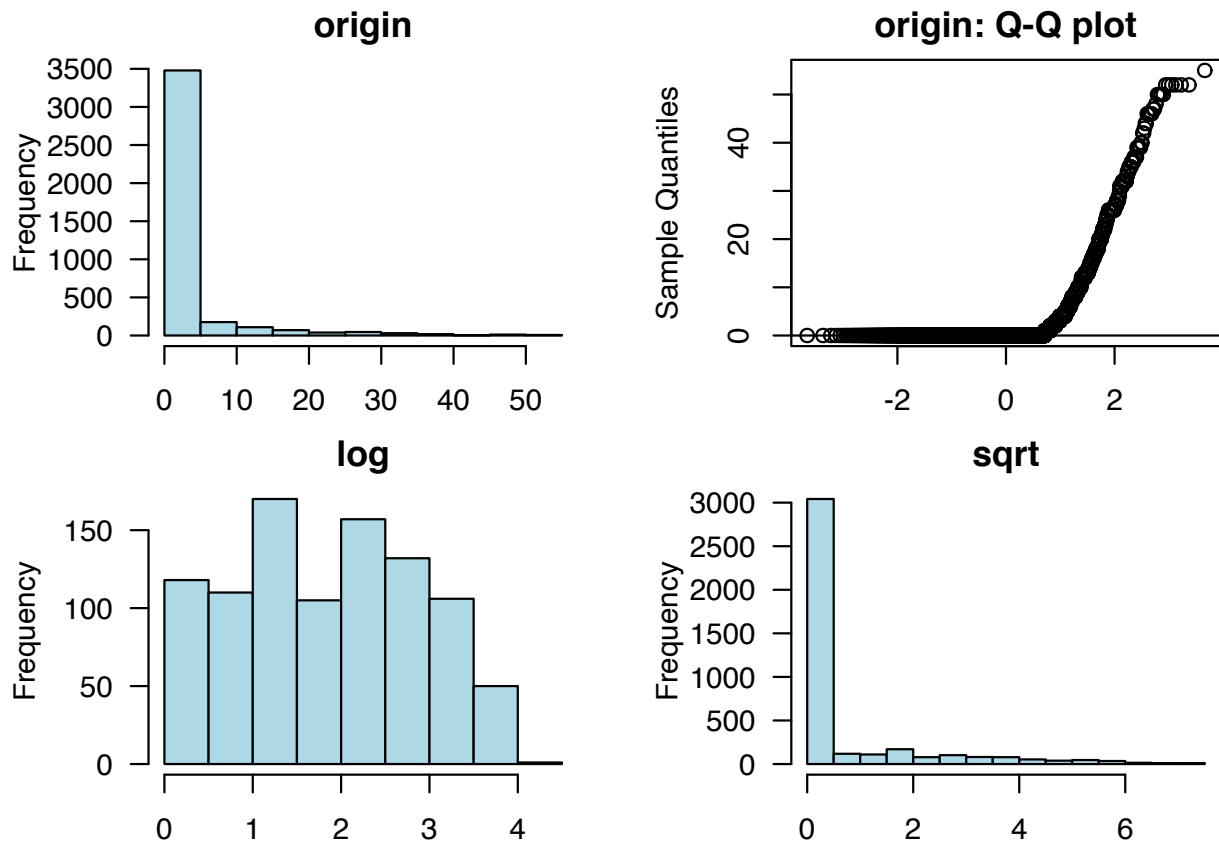


Figure 2.15: wks_ue

ttl_exp

normality test : Shapiro-Wilk normality test
 statistic : 0.92709, p-value : 4.83001E-44

type	skewness	kurtosis
original	0.8390	3.0262
log transformation	-0.9583	4.0385
sqrt transformation	0.1344	2.2322

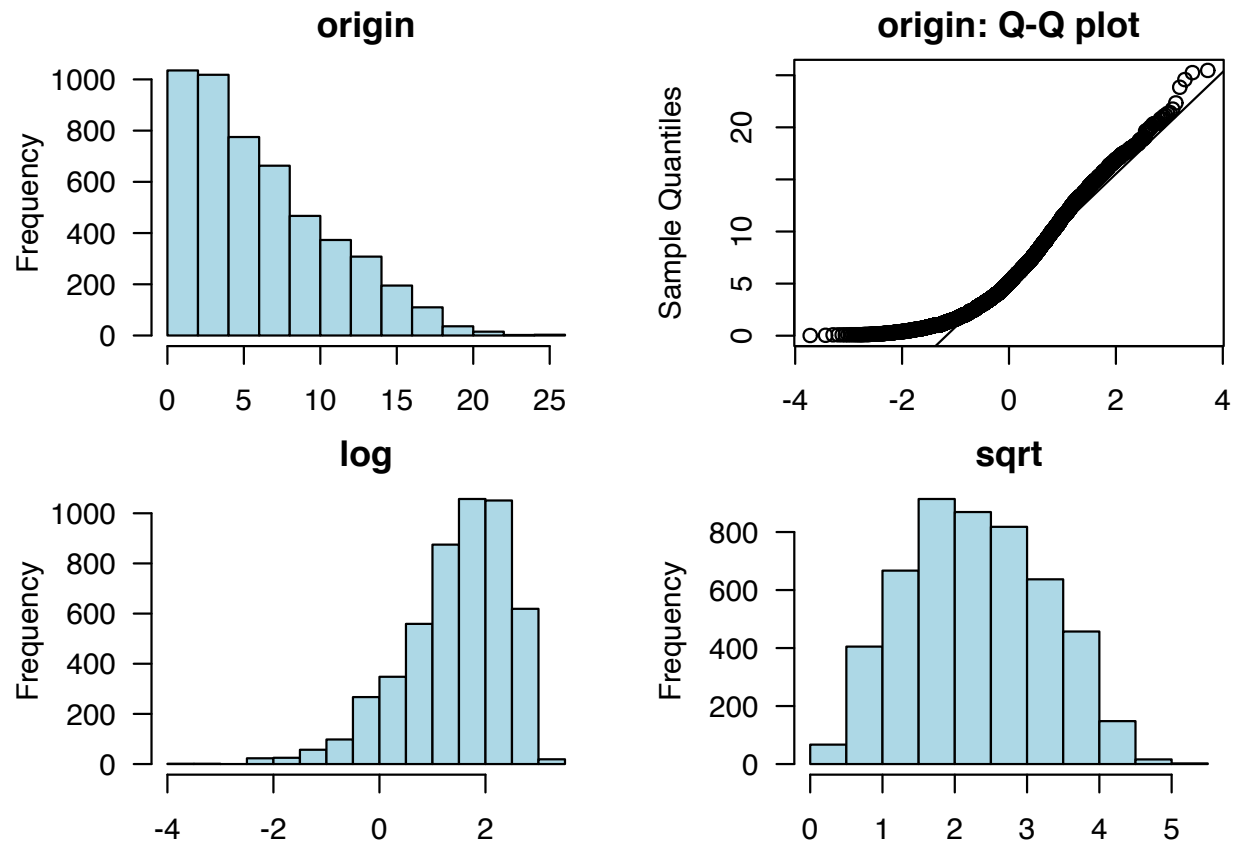


Figure 2.16: ttl_exp

tenure

normality test : Shapiro-Wilk normality test
 statistic : 0.77474, p-value : 1.54979E-63

type	skewness	kurtosis
original	1.8492	6.3967
log transformation		
sqrt transformation	0.7408	2.9651

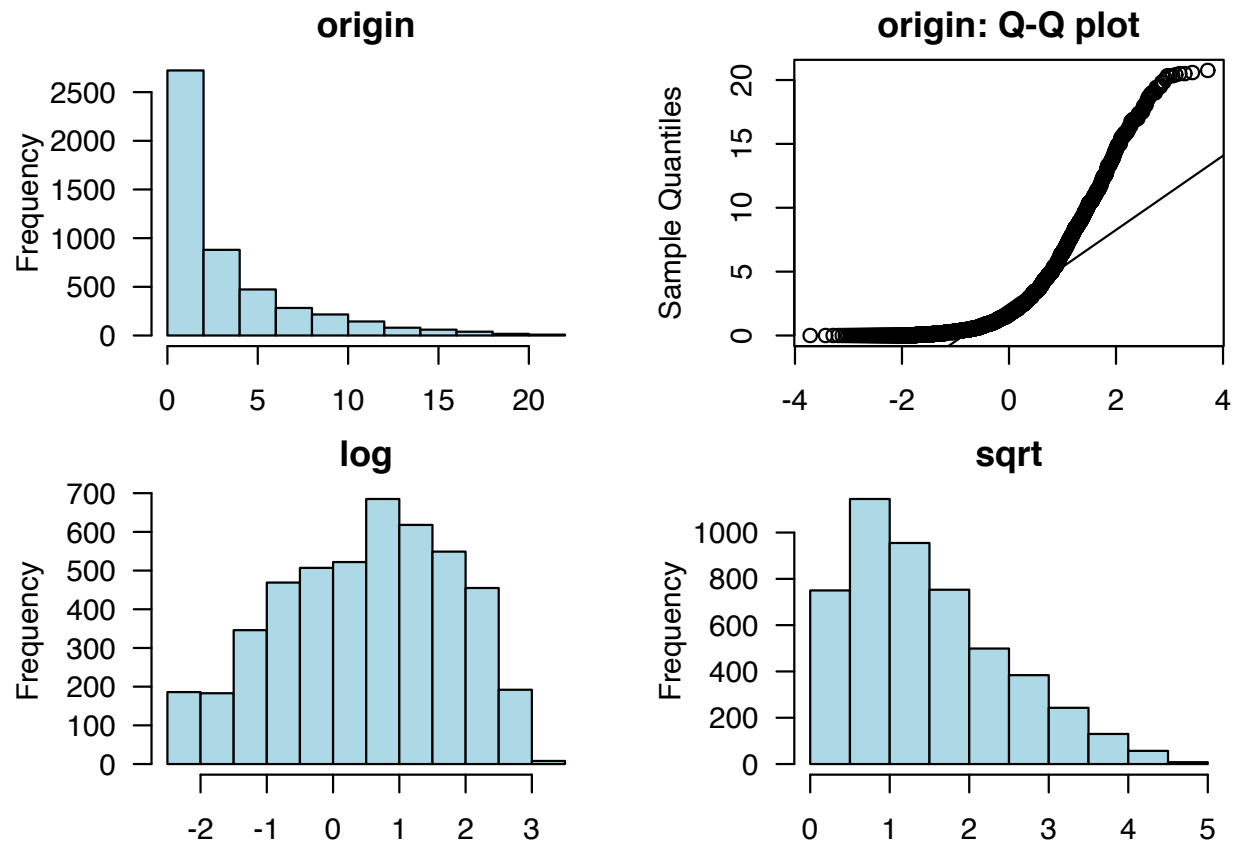


Figure 2.17: tenure

hours

normality test : Shapiro-Wilk normality test
 statistic : 0.76294, p-value : 8.26327E-65

type	skewness	kurtosis
original	-0.5803	12.2148
log transformation	-3.2102	16.6244
sqrt transformation	-1.9003	8.6759

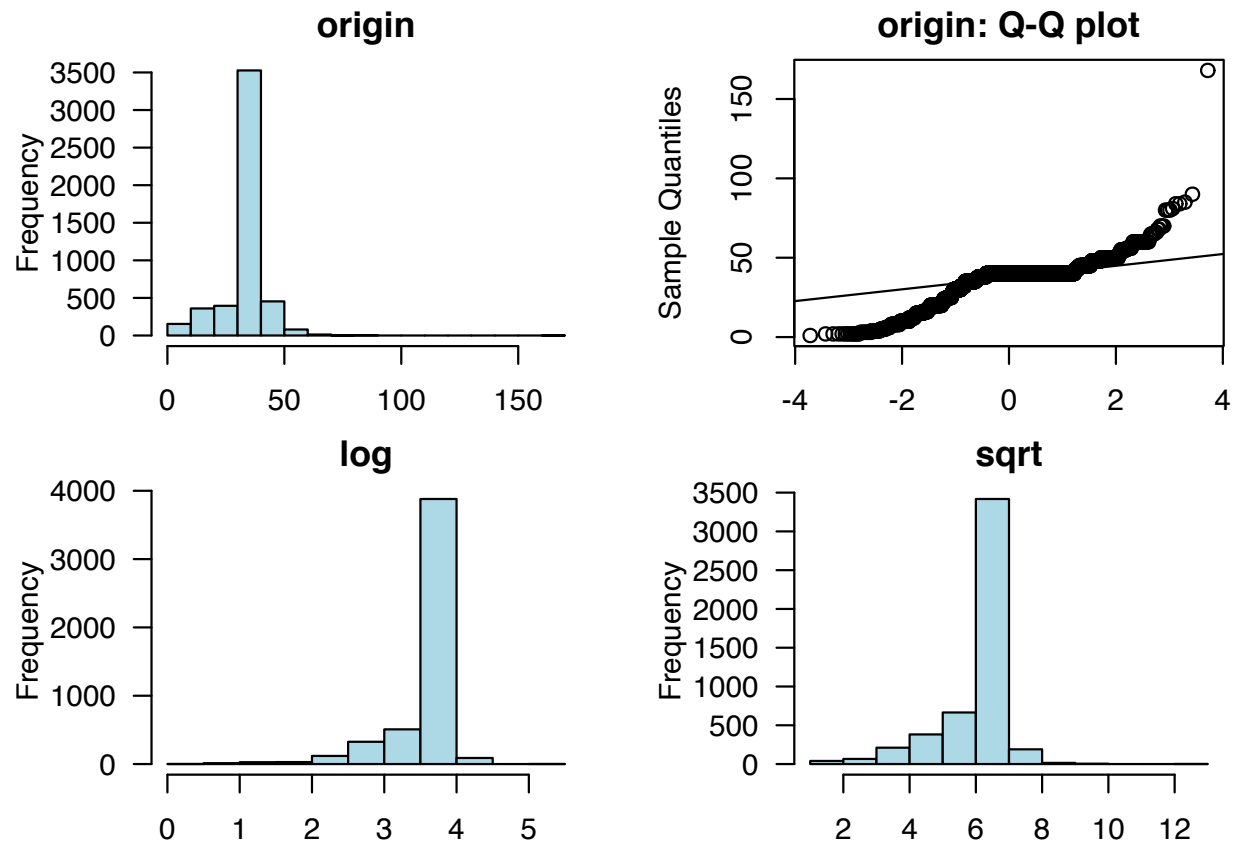


Figure 2.18: hours

wks_work

normality test : Shapiro-Wilk normality test
 statistic : 0.93709, p-value : 2.52751E-41

type	skewness	kurtosis
original	0.1956	2.3285
log transformation		
sqrt transformation	-0.7896	3.5910

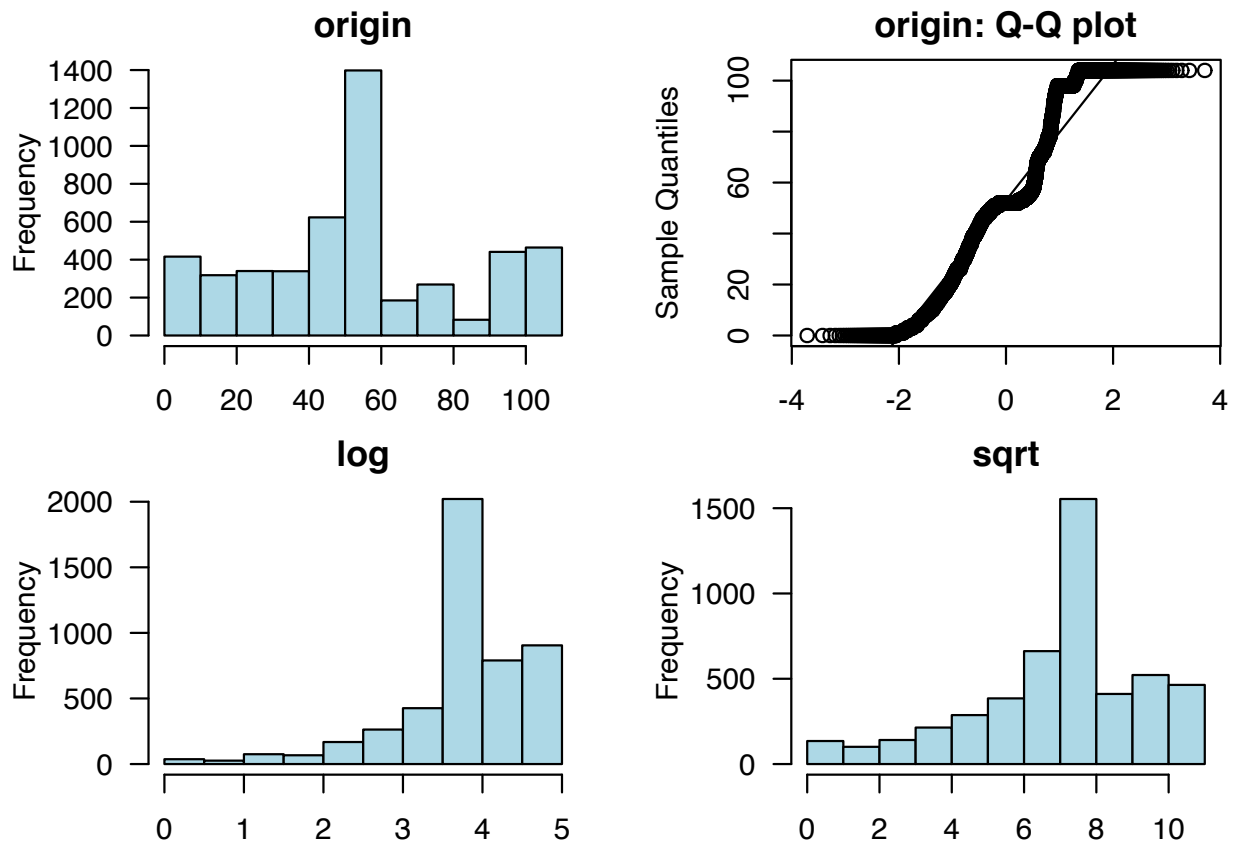


Figure 2.19: wks_work

ln_wage

normality test : Shapiro-Wilk normality test
 statistic : 0.98225, p-value : 1.45277E-24

type	skewness	kurtosis
original	0.3349	4.6155
log transformation	-3.5202	35.0785
sqrt transformation	-0.6646	6.3659

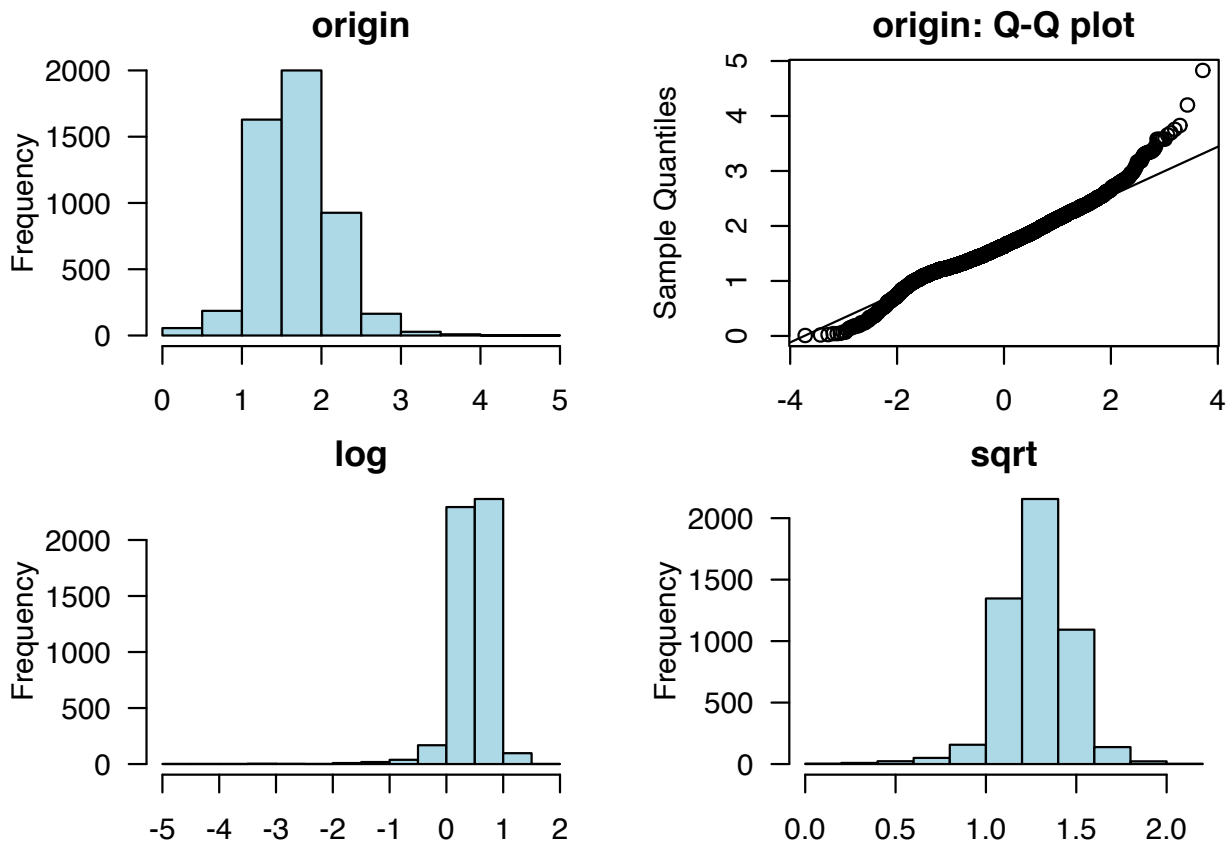


Figure 2.20: ln_wage

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
age	year	0.895
ttl_exp	year	0.777
collgrad	grade	0.757
ttl_exp	age	0.756
tenure	ttl_exp	0.674
nev_mar	msp	-0.673
wks_work	ttl_exp	0.630
wks_work	year	0.565
wks_work	age	0.525

3.1.2 Correlation Plot of Numerical Variables

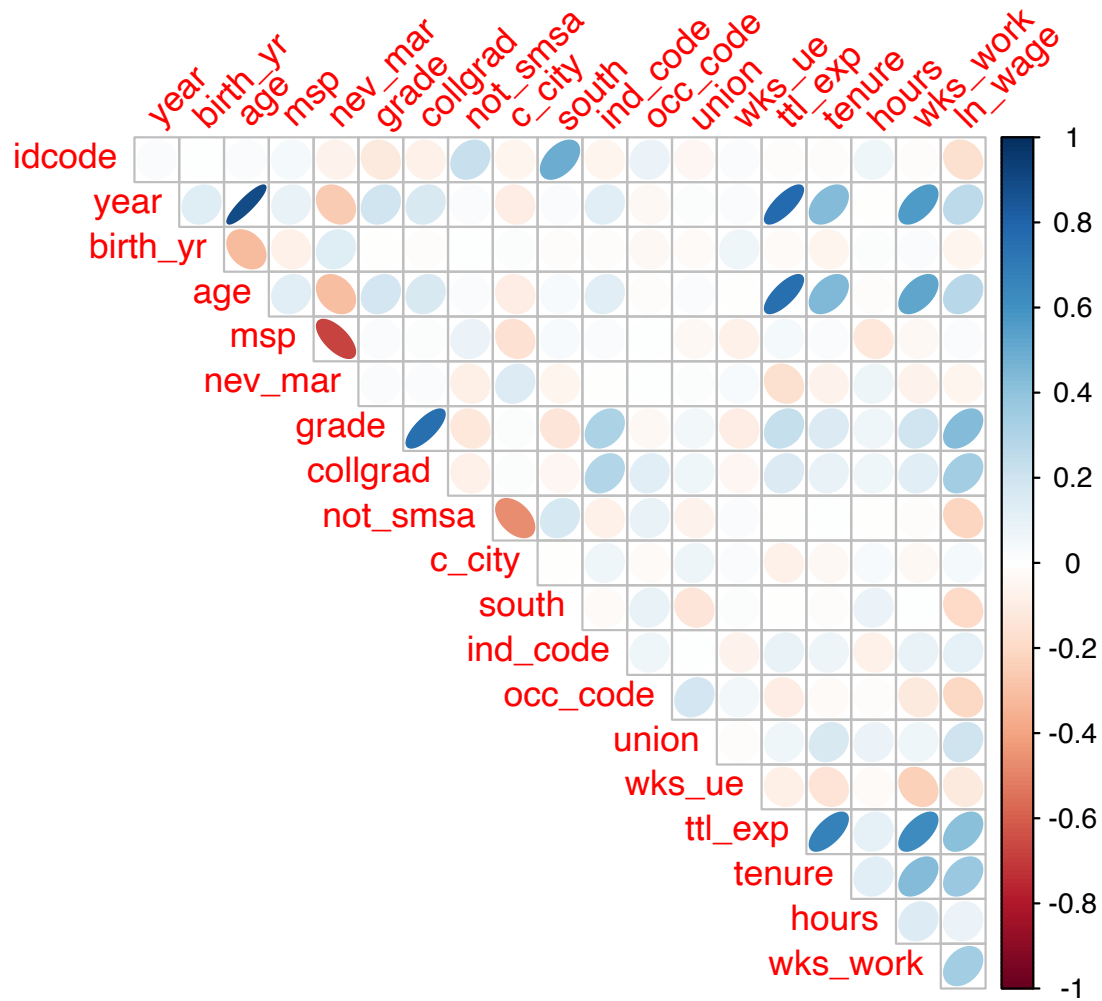


Figure 3.1: The correlation coefficient of numerical variables

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

There is no target variable.

4.1.2 Grouped Categorical Variables

There is no target variable.

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

There is no target variable.

4.2.2 Grouped Correlation Plot of Numerical Variables

There is no target variable.