REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.12

December 16, 2019

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 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

edaData
21 Variables 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

---

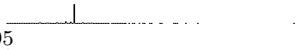
tenure : job tenure, in years Format:%9.0g

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
```

---

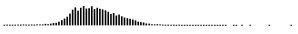
wks_work : weeks worked last year Format:%8.0g

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 : 0.95529, p-value : 1.25423E-36

type	skewness	kurtosis
original	-0.0035	1.8064
log transformation	-2.1379	9.8044
sqrt transformation	-0.5785	2.4471

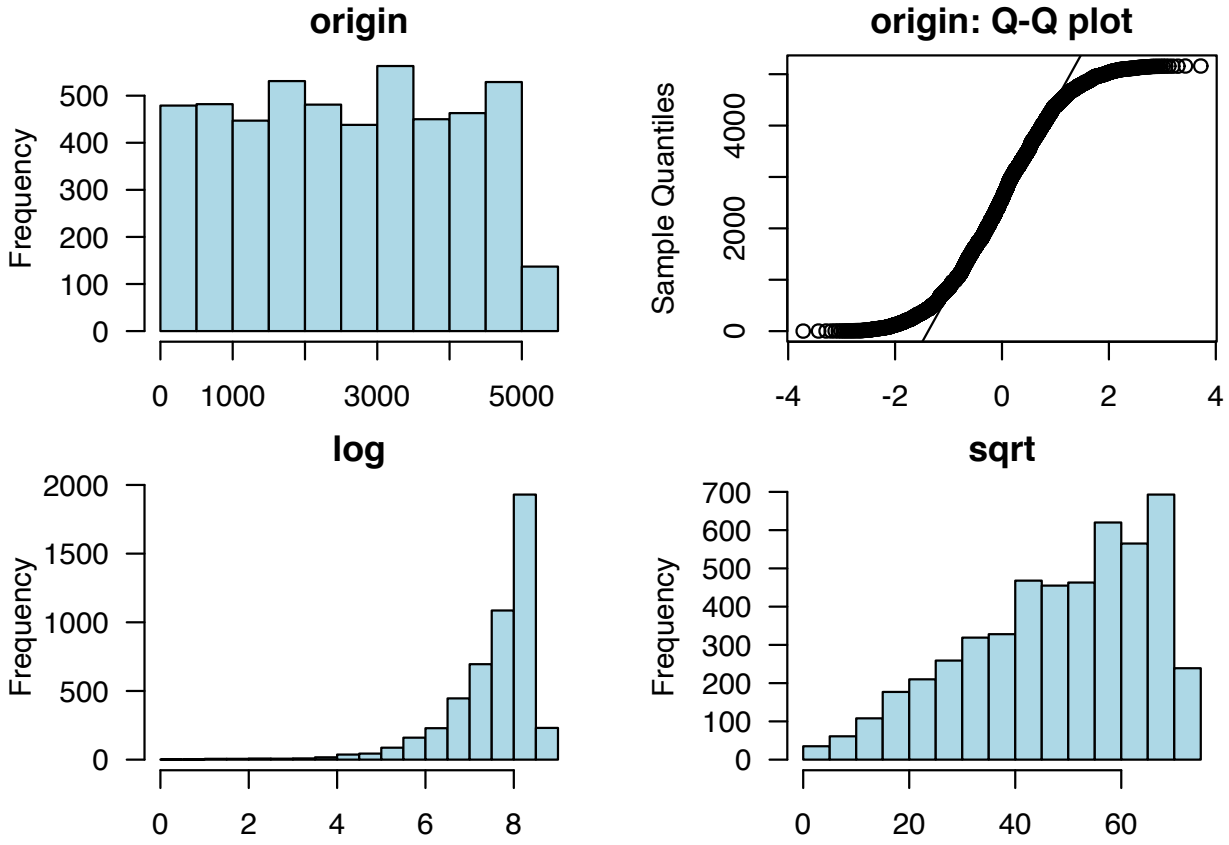


Figure 2.1: idcode

year

normality test : Shapiro-Wilk normality test
statistic : 0.93286, p-value : 9.6295E-43

type	skewness	kurtosis
original	0.0990	1.7130
log transformation	0.0131	1.7079
sqrt transformation	0.0561	1.7085

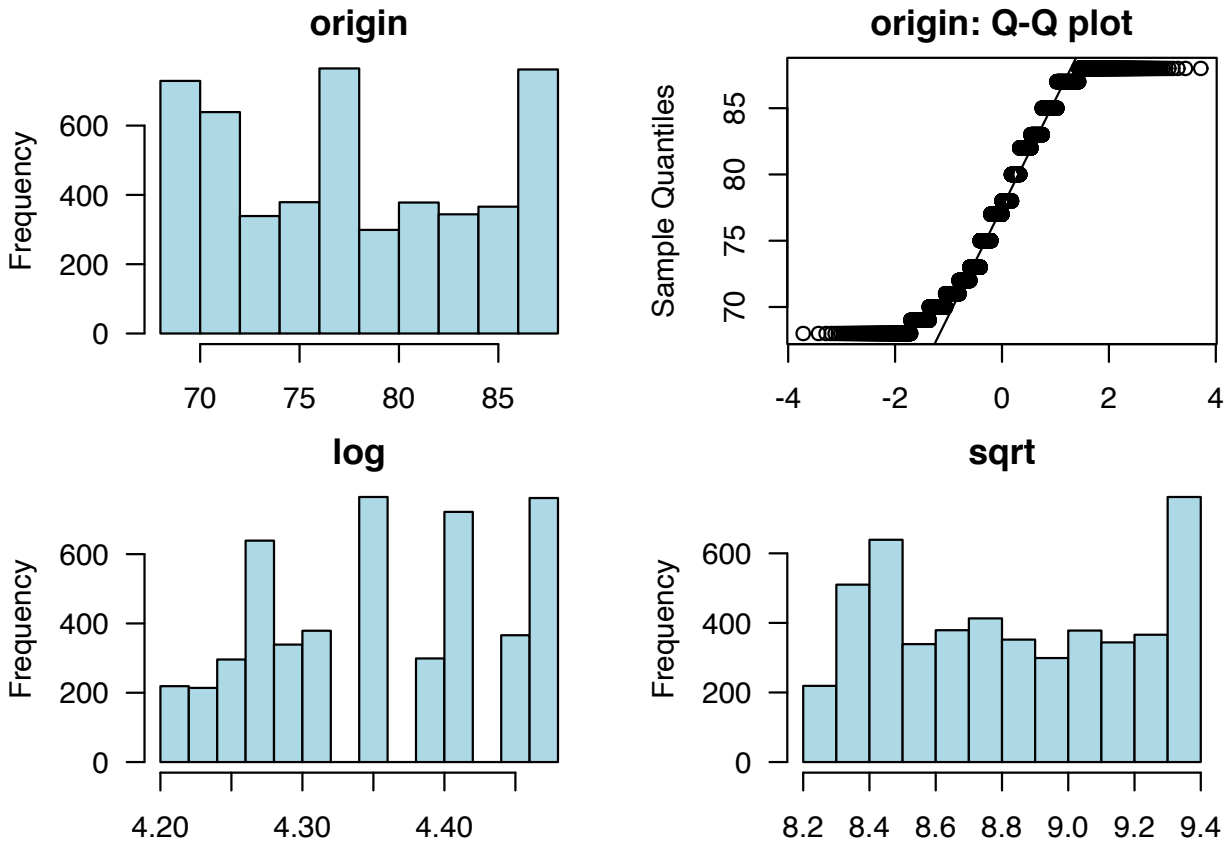


Figure 2.2: year

birth_yr

normality test : Shapiro-Wilk normality test statistic : 0.95854, p-value : 1.50897E-35

type	skewness	kurtosis
original	-0.1462	1.9994
log transformation	-0.2406	2.0574
sqrt transformation	-0.1932	2.0257

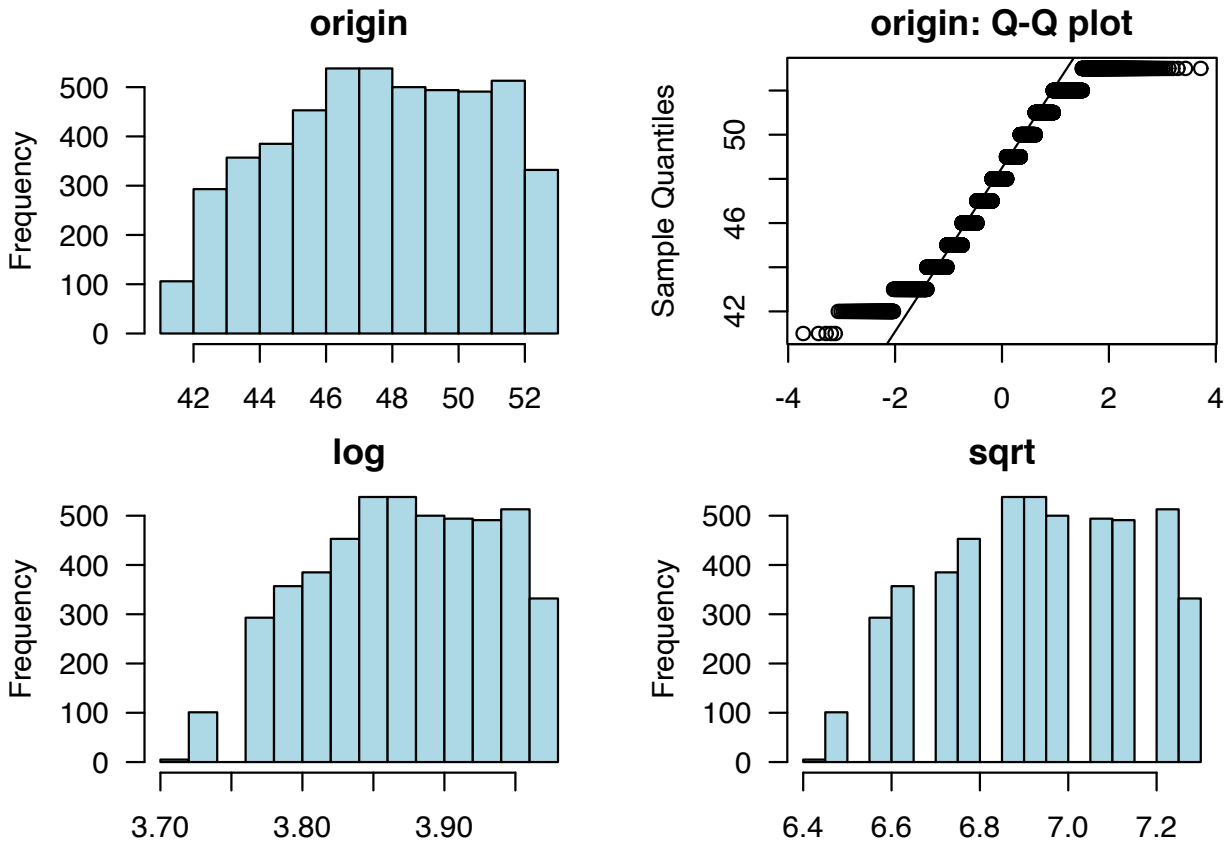


Figure 2.3: birth_yr

age

normality test : Shapiro-Wilk normality test
statistic : 0.96855, p-value : 1.01959E-31

type	skewness	kurtosis
original	0.2349	2.0570
log transformation	-0.1160	2.0211
sqrt transformation	0.0610	1.9956

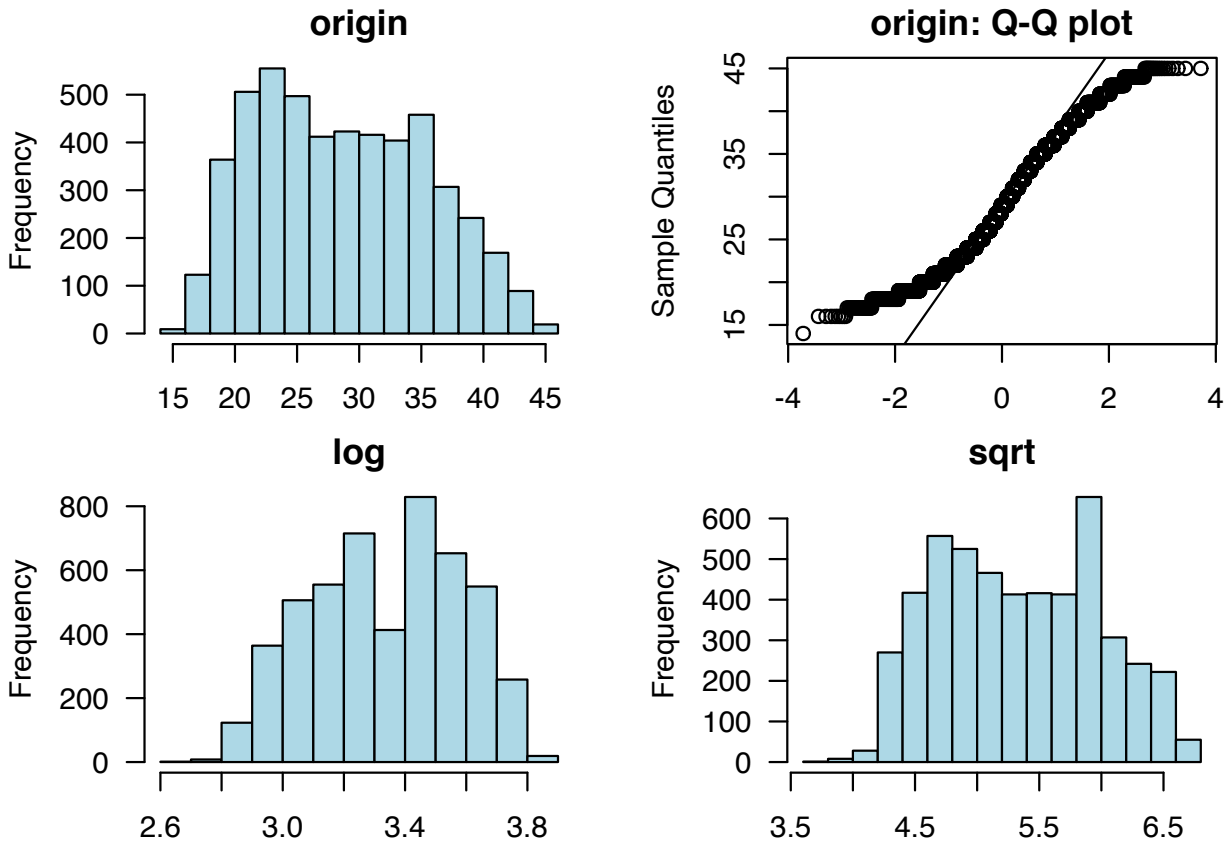


Figure 2.4: age

msp

normality test : Shapiro-Wilk normality test
statistic : 0.62169, p-value : 3.20324E-74

type	skewness	kurtosis
original	-0.4115	1.1693
log transformation		
sqrt transformation	-0.4115	1.1693

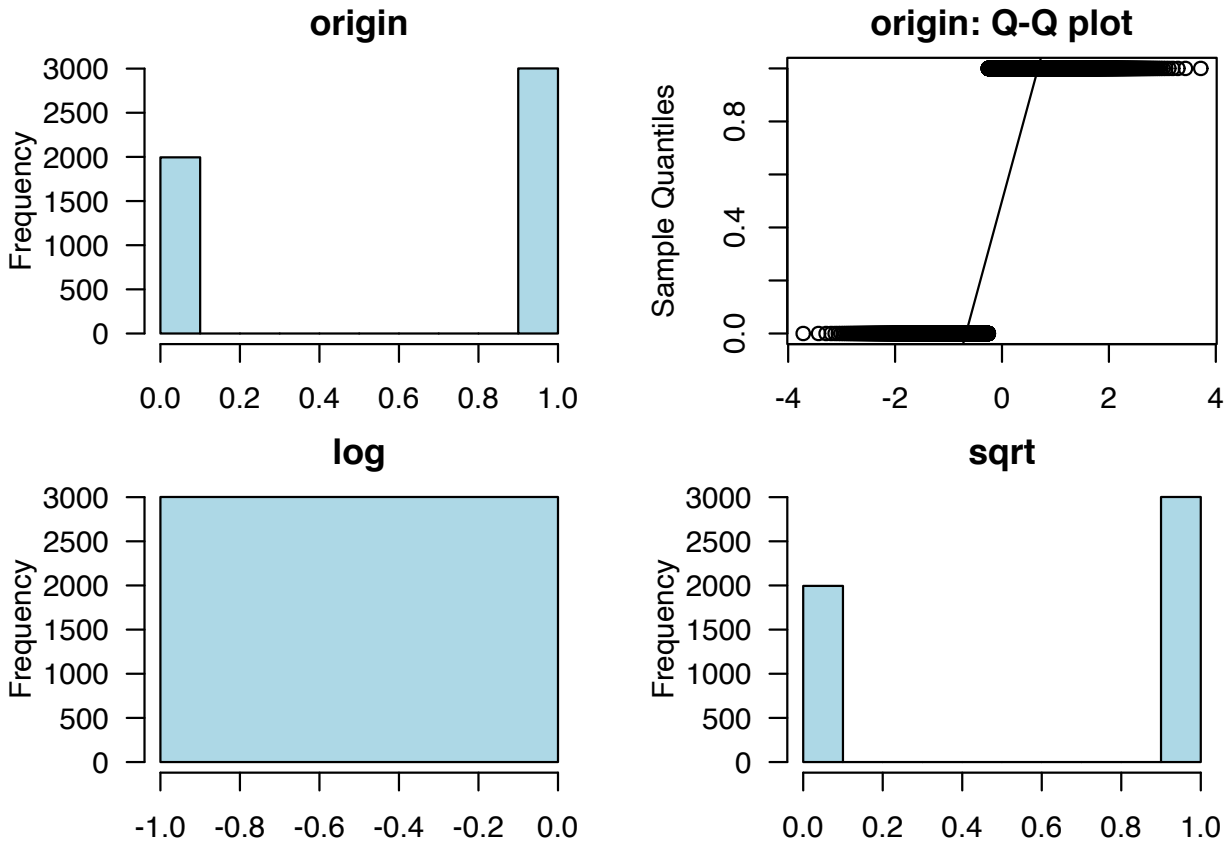


Figure 2.5: msp

nev_mar

normality test : Shapiro-Wilk normality test
statistic : 0.52276, p-value : 3.90135E-79

type	skewness	kurtosis
original	1.2679	2.6075
log transformation		
sqrt transformation	1.2679	2.6075

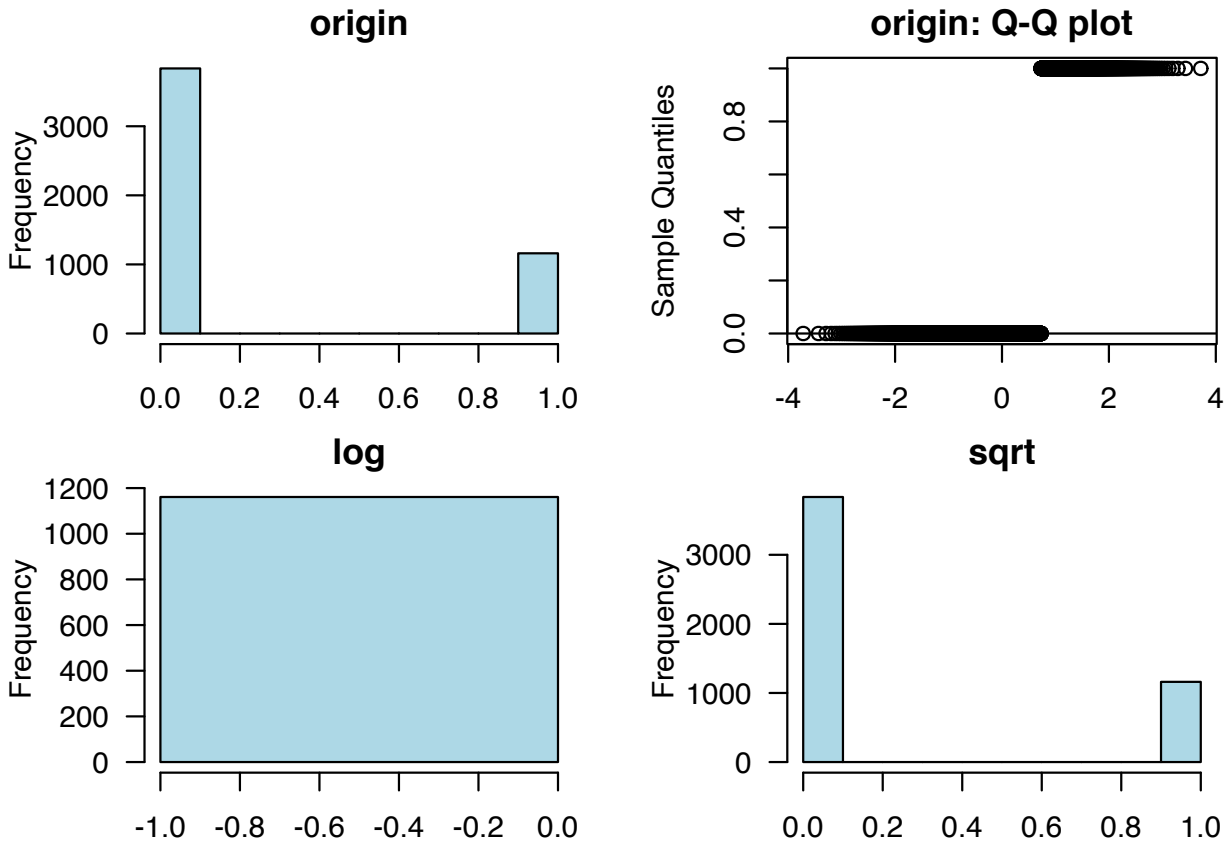


Figure 2.6: nev_mar

grade

normality test : Shapiro-Wilk normality test
statistic : 0.88616, p-value : 2.04355E-51

type	skewness	kurtosis
original	0.0432	4.5261
log transformation		
sqrt transformation	-1.3694	15.3867

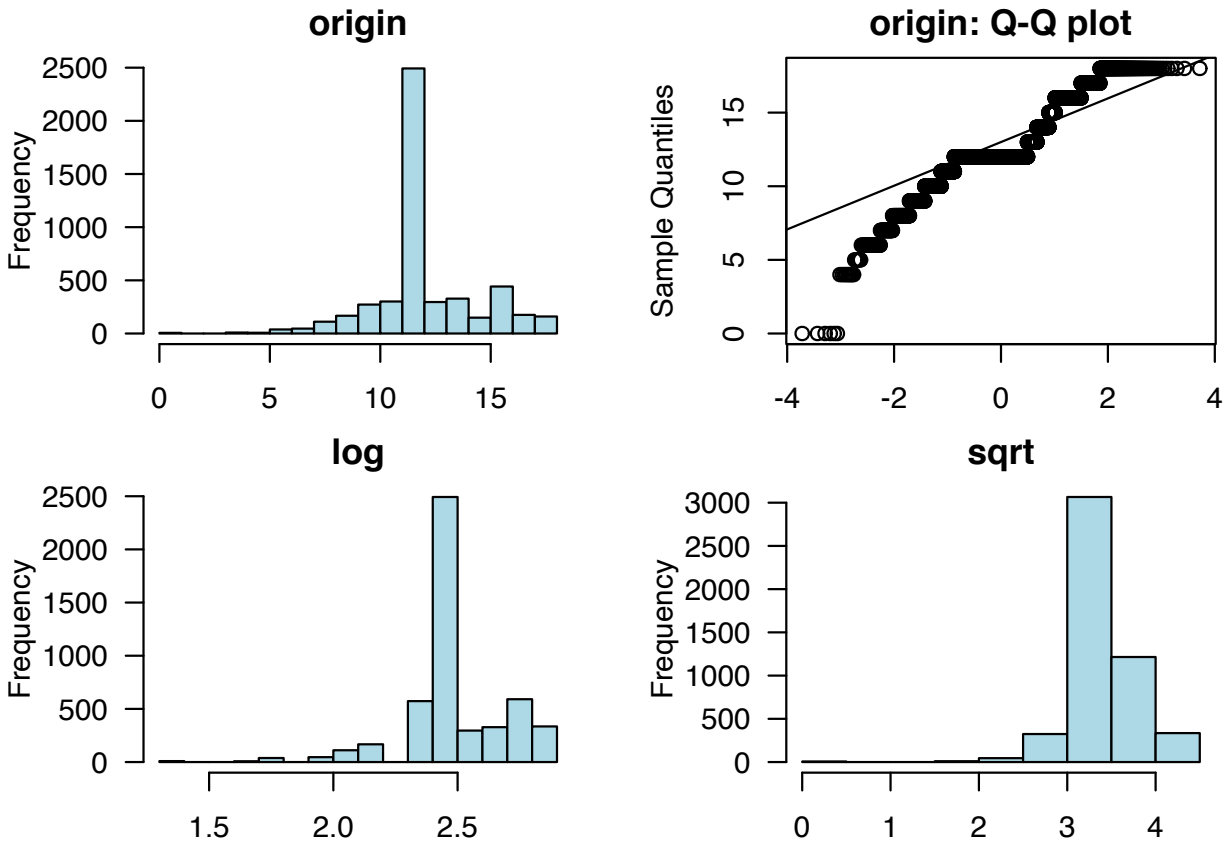


Figure 2.7: grade

collgrad

normality test : Shapiro-Wilk normality test
statistic : 0.44155, p-value : 1.47477E-82

type	skewness	kurtosis
original	1.8468	4.4105
log transformation		
sqrt transformation	1.8468	4.4105

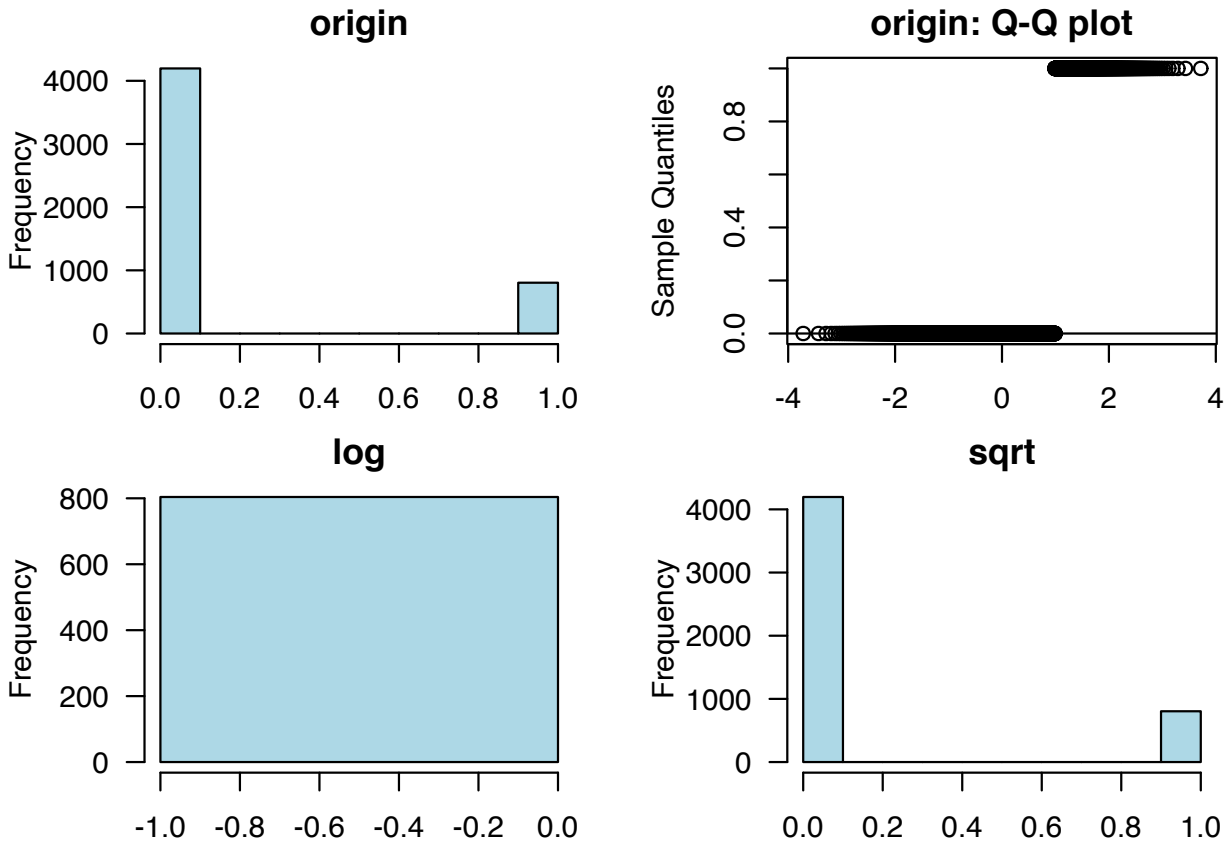


Figure 2.8: collgrad

not_smsa

normality test : Shapiro-Wilk normality test
statistic : 0.56546, p-value : 3.9451E-77

type	skewness	kurtosis
original	0.9538	1.9097
log transformation		
sqrt transformation	0.9538	1.9097

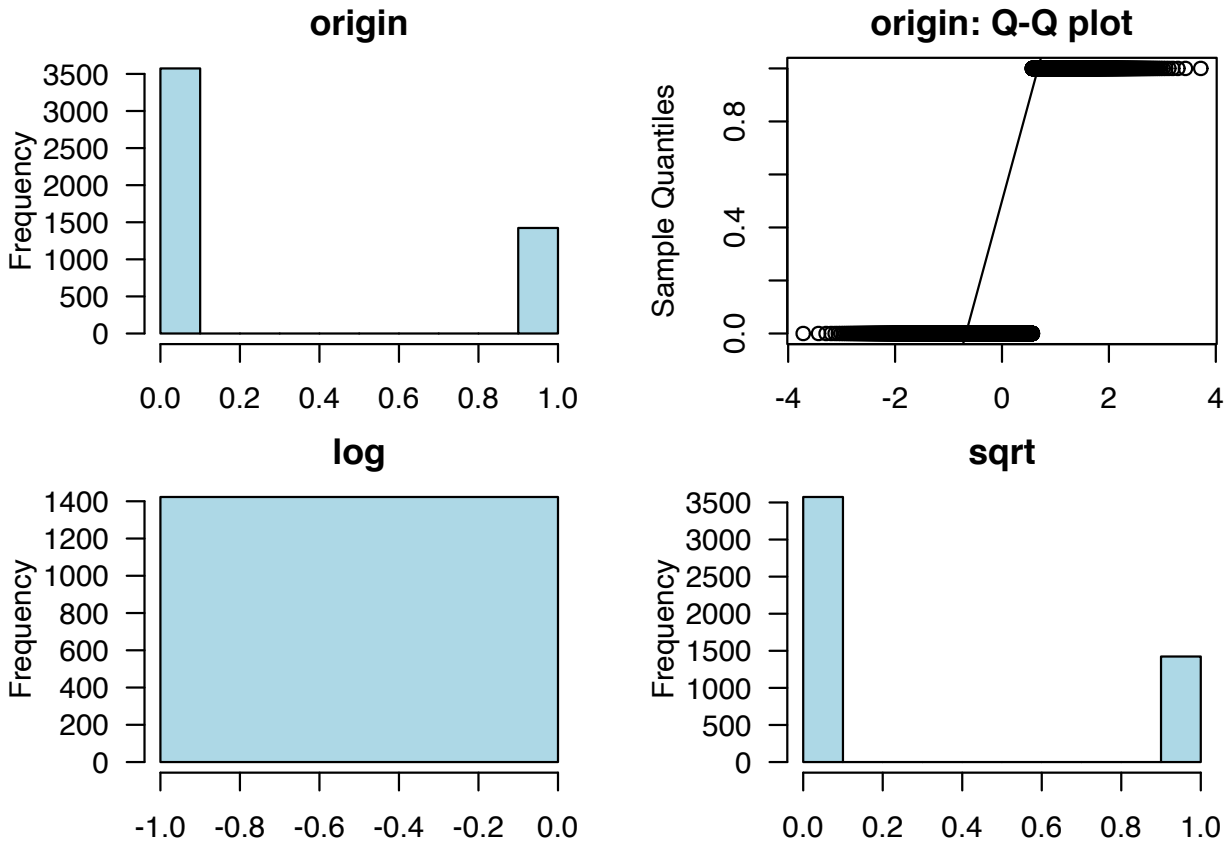


Figure 2.9: not_smsa

c_city

normality test : Shapiro-Wilk normality test
statistic : 0.60394, p-value : 3.50353E-75

type	skewness	kurtosis
original	0.6200	1.3845
log transformation		
sqrt transformation	0.6200	1.3845

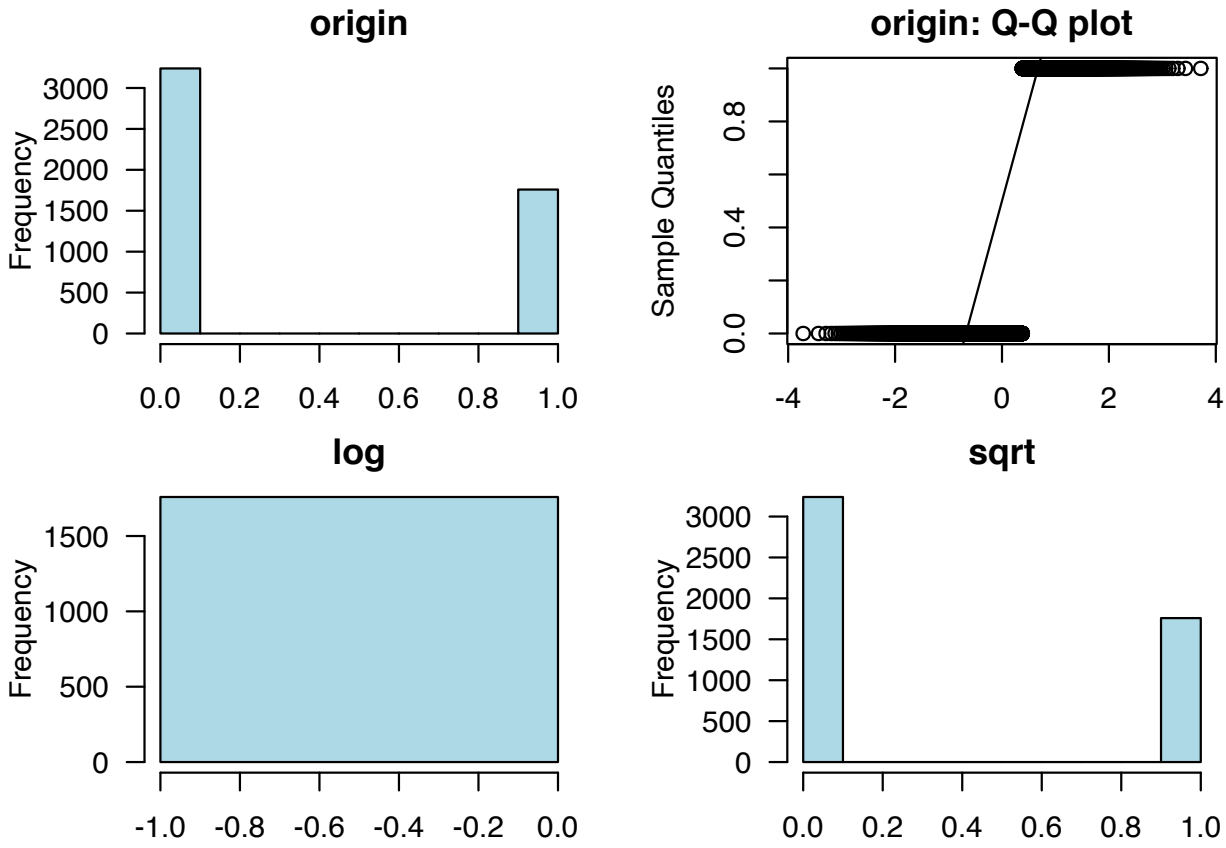


Figure 2.10: c_city

south

normality test : Shapiro-Wilk normality test
statistic : 0.62382, p-value : 4.06184E-74

type	skewness	kurtosis
original	0.3803	1.1446
log transformation		
sqrt transformation	0.3803	1.1446

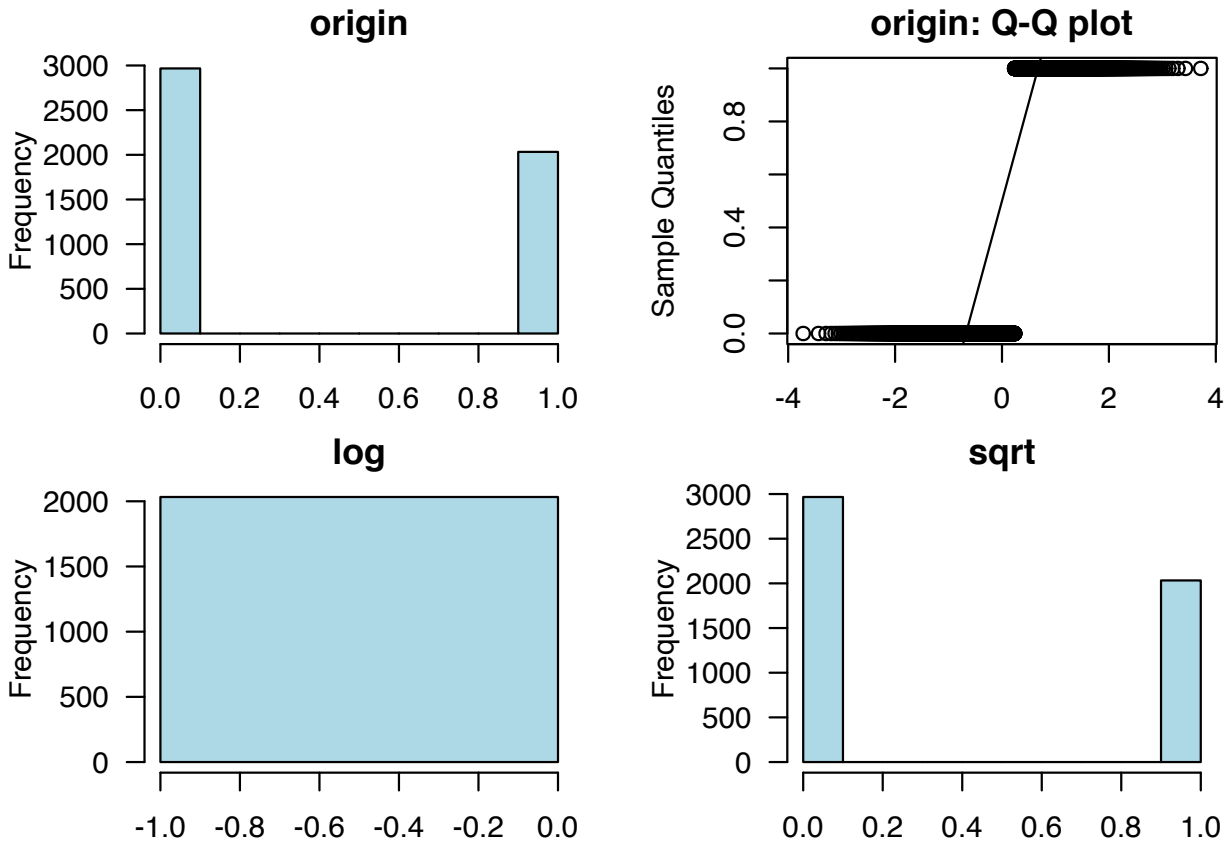


Figure 2.11: south

ind_code

normality test : Shapiro-Wilk normality test
statistic : 0.87002, p-value : 1.72595E-53

type	skewness	kurtosis
original	-0.0063	1.5386
log transformation	-0.8614	4.3499
sqrt transformation	-0.2745	2.0514

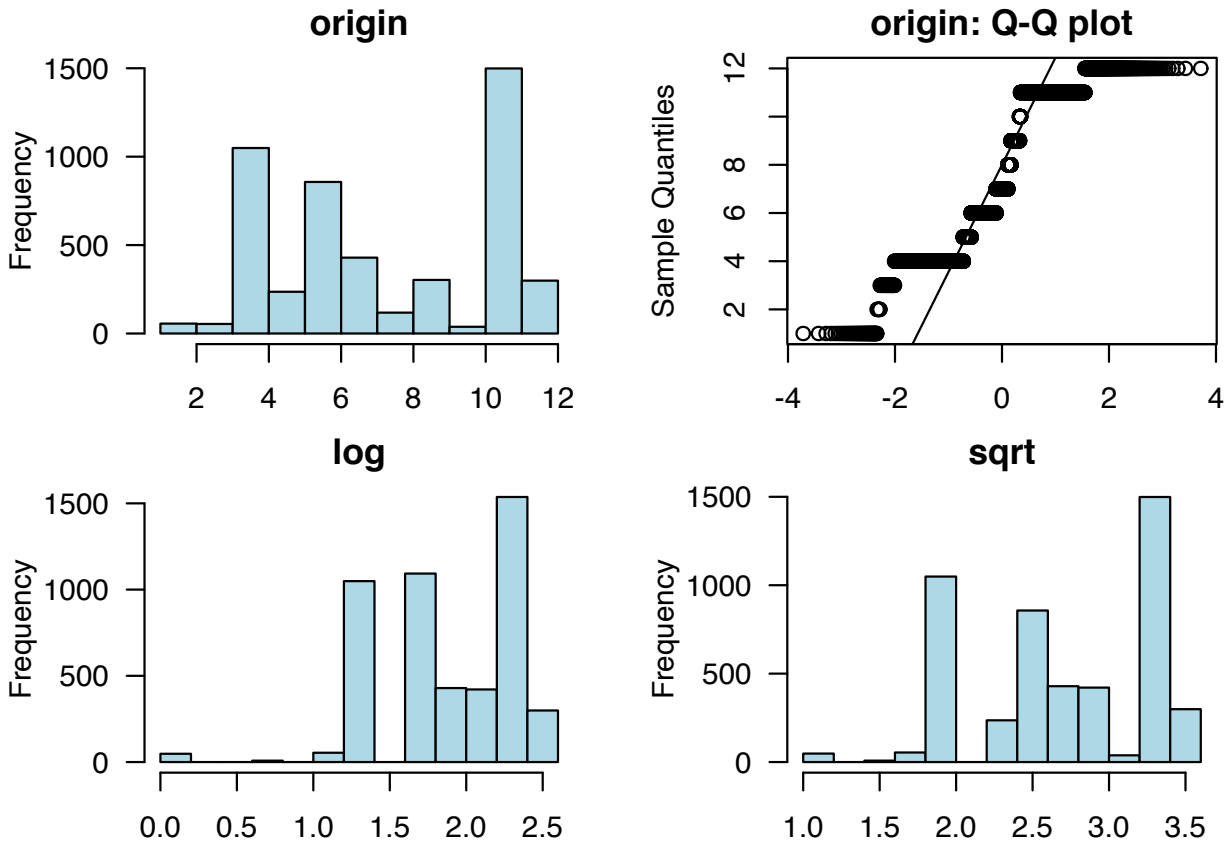


Figure 2.12: ind_code

occ_code

normality test : Shapiro-Wilk normality test
statistic : 0.84785, p-value : 1.86192E-56

type	skewness	kurtosis
original	1.1157	3.6995
log transformation	-0.2677	2.6410
sqrt transformation	0.4820	2.6493

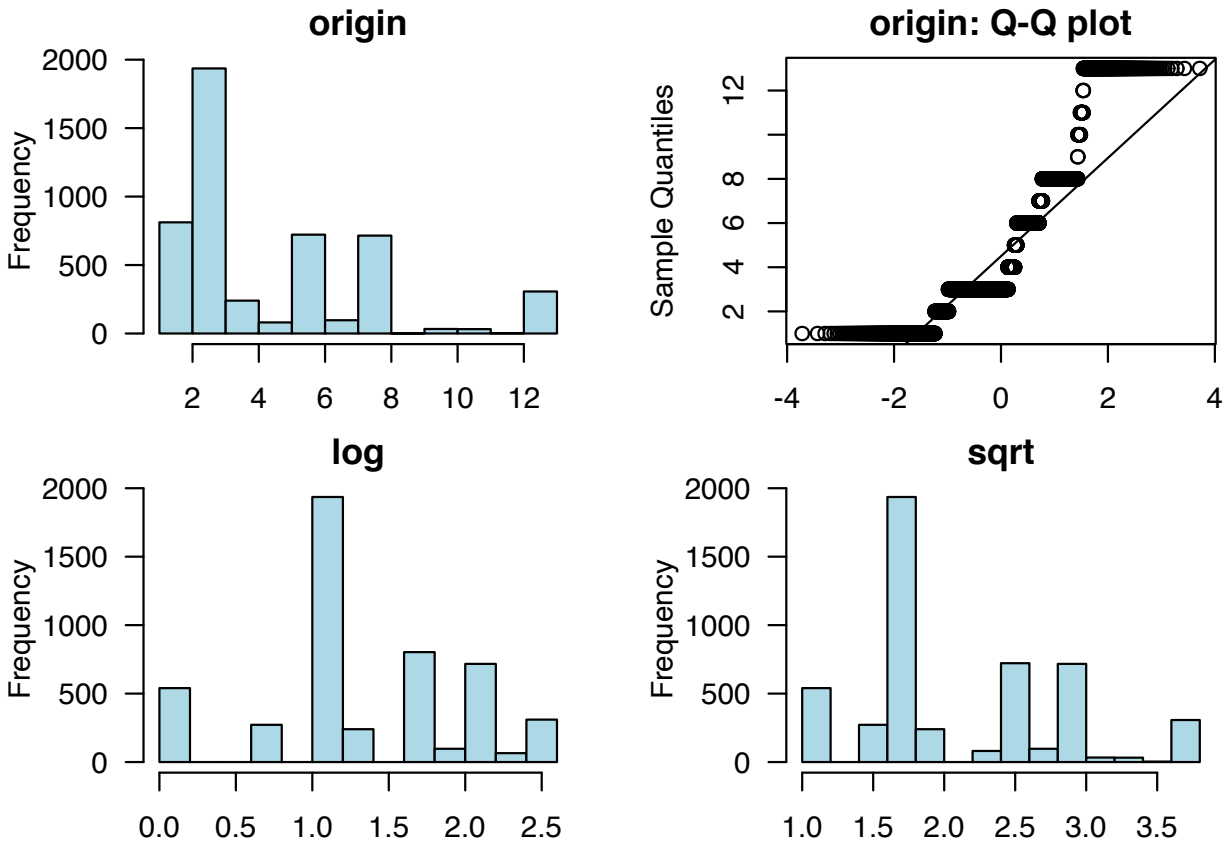


Figure 2.13: occ_code

union

normality test : Shapiro-Wilk normality test
statistic : 0.5318, p-value : 1.18229E-69

type	skewness	kurtosis
original	1.2033	2.4480
log transformation		
sqrt transformation	1.2033	2.4480

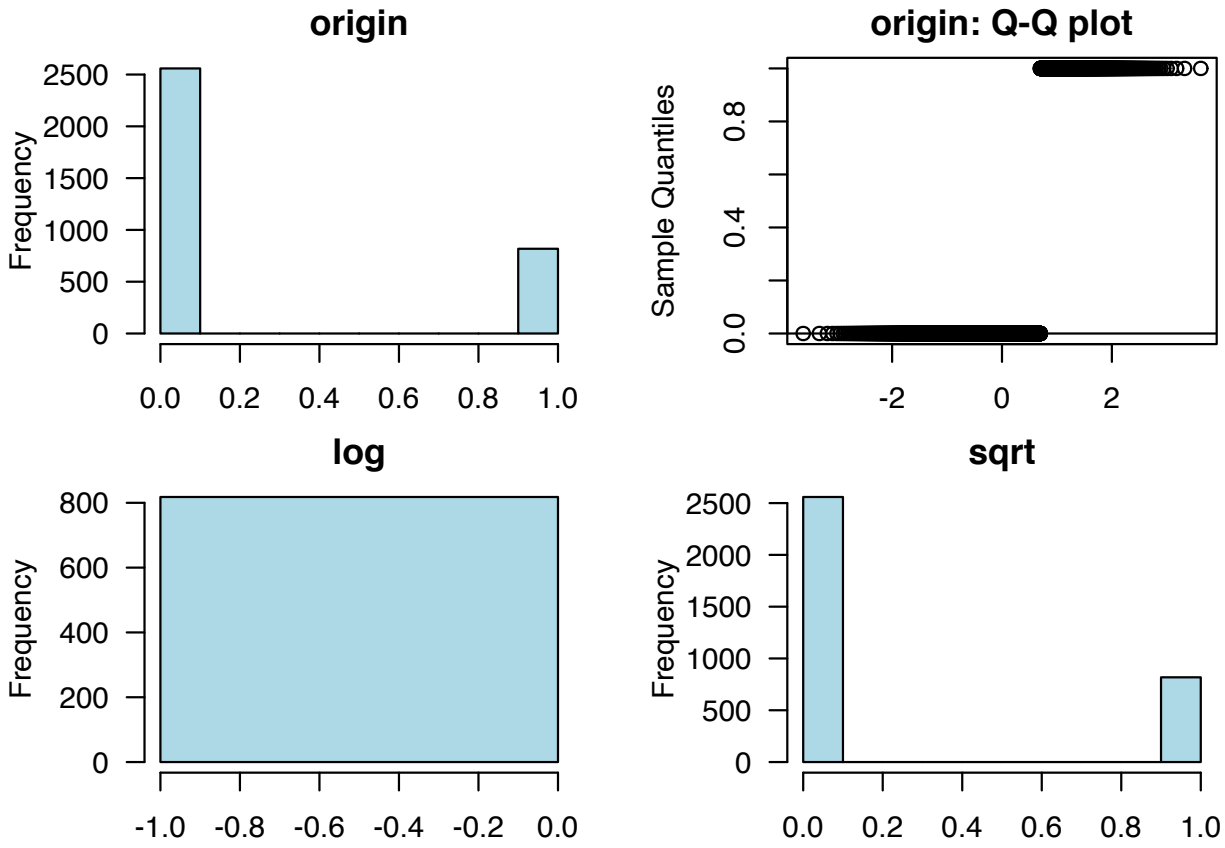


Figure 2.14: union

wks_ue

normality test : Shapiro-Wilk normality test
statistic : 0.3877, p-value : 6.48802E-79

type	skewness	kurtosis
original	4.1148	21.6355
log transformation		
sqrt transformation	2.4406	8.5699

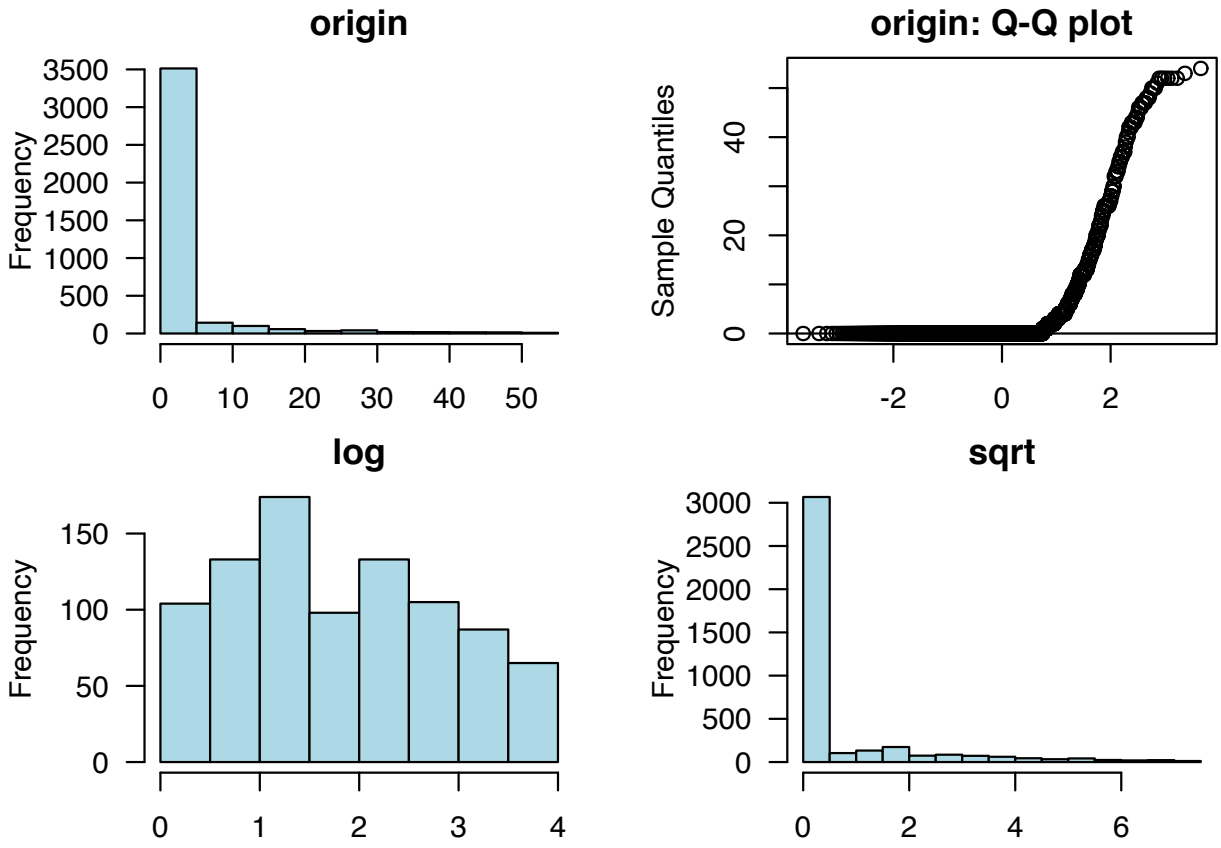


Figure 2.15: wks_ue

ttl_exp

normality test : Shapiro-Wilk normality test
statistic : 0.92752, p-value : 6.00198E-44

type	skewness	kurtosis
original	0.8166	2.9277
log transformation		
sqrt transformation	0.1117	2.2189

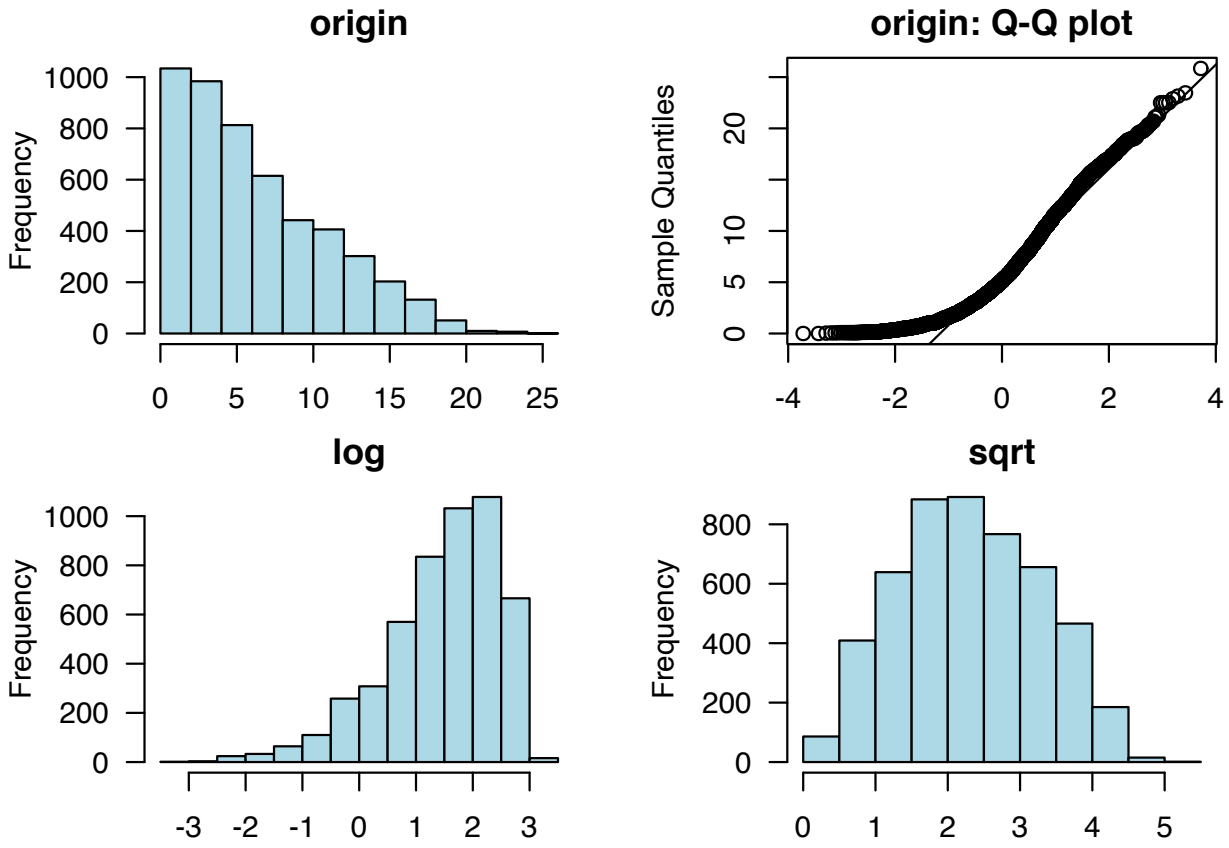


Figure 2.16: ttl_exp

tenure

normality test : Shapiro-Wilk normality test
statistic : 0.76576, p-value : 2.51124E-64

type	skewness	kurtosis
original	1.9082	6.6416
log transformation		
sqrt transformation	0.7659	3.0460

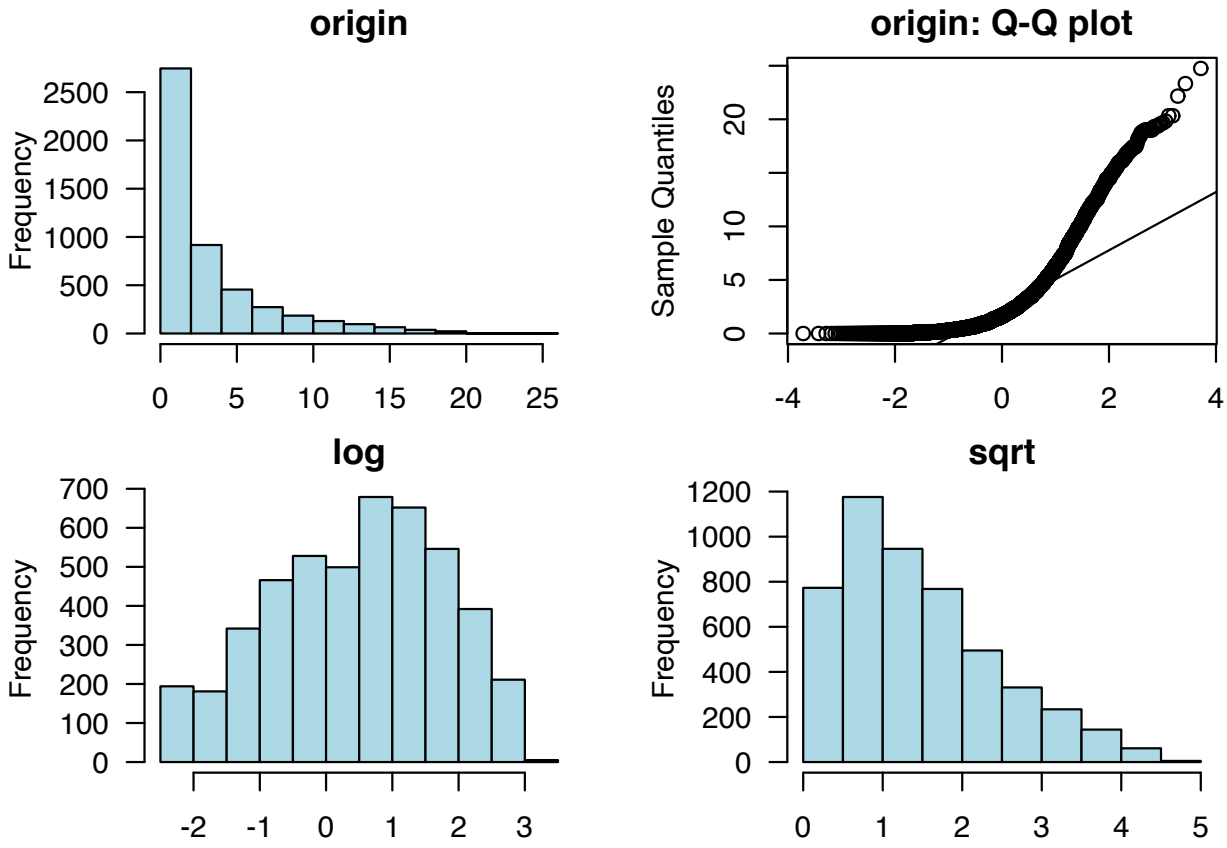


Figure 2.17: tenure

hours

normality test : Shapiro-Wilk normality test
statistic : 0.78796, p-value : 1.22006E-62

type	skewness	kurtosis
original	-0.9924	6.0220
log transformation	-3.0105	14.8091
sqrt transformation	-1.8686	7.4196

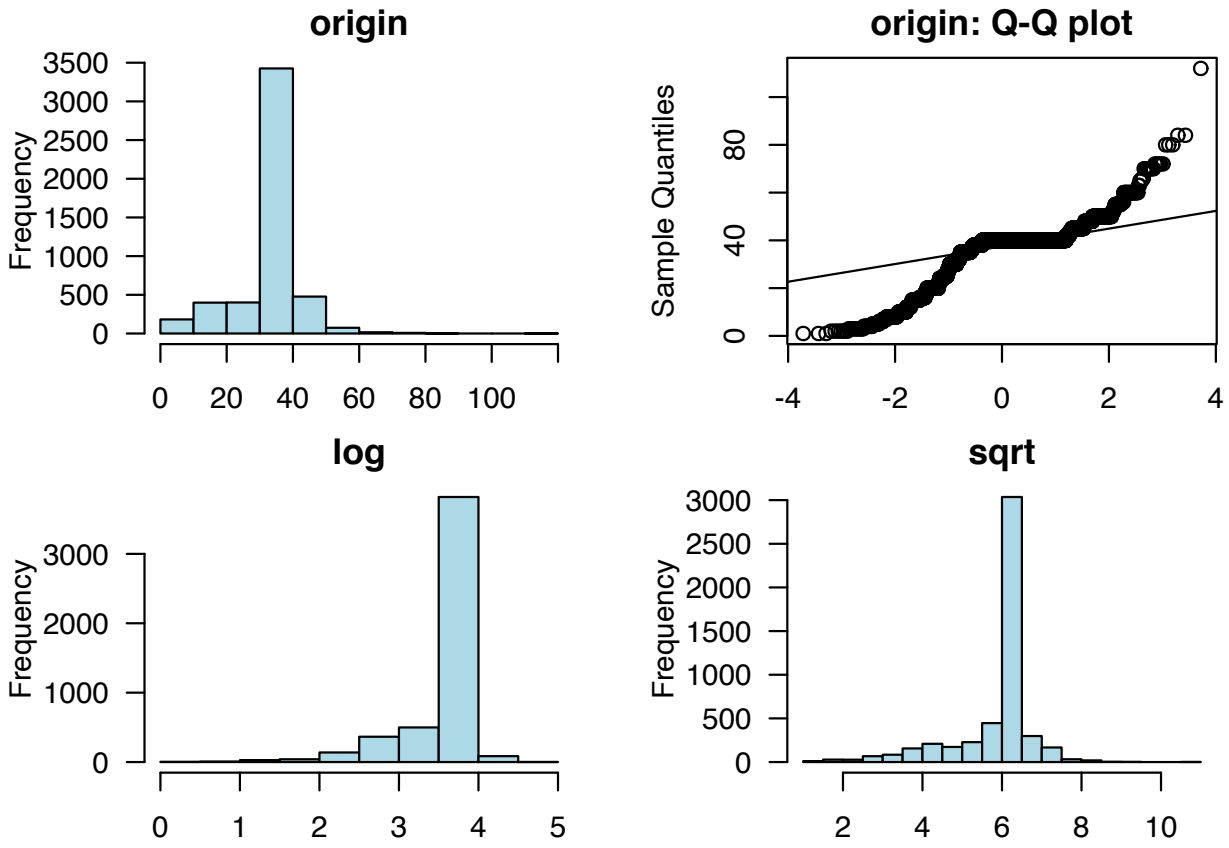


Figure 2.18: hours

wks_work

normality test : Shapiro-Wilk normality test
statistic : 0.93799, p-value : 3.63666E-41

type	skewness	kurtosis
original	0.1615	2.3096
log transformation		
sqrt transformation	-0.8032	3.5871

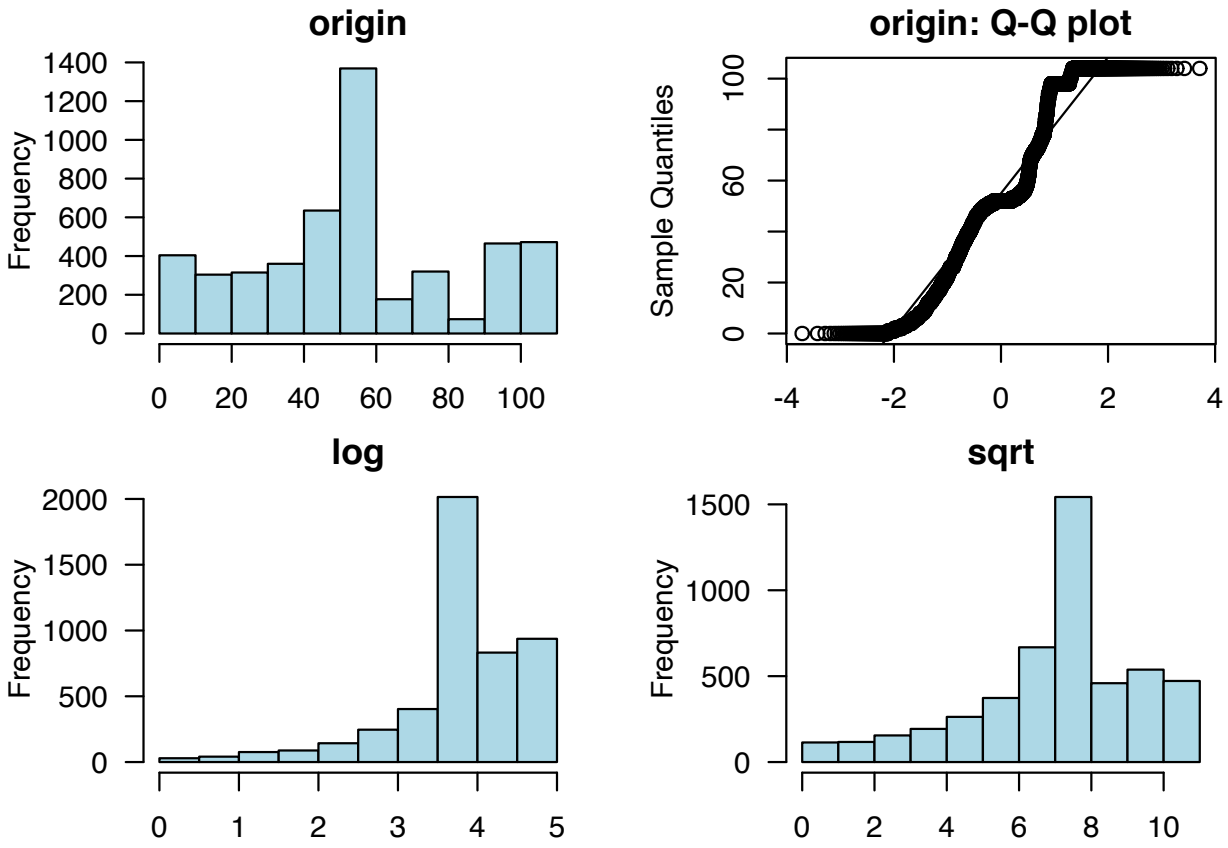


Figure 2.19: wks_work

ln_wage

normality test : Shapiro-Wilk normality test
statistic : 0.97995, p-value : 5.20474E-26

type	skewness	kurtosis
original	0.4143	4.5261
log transformation	-3.4170	36.5940
sqrt transformation	-0.5430	5.9949

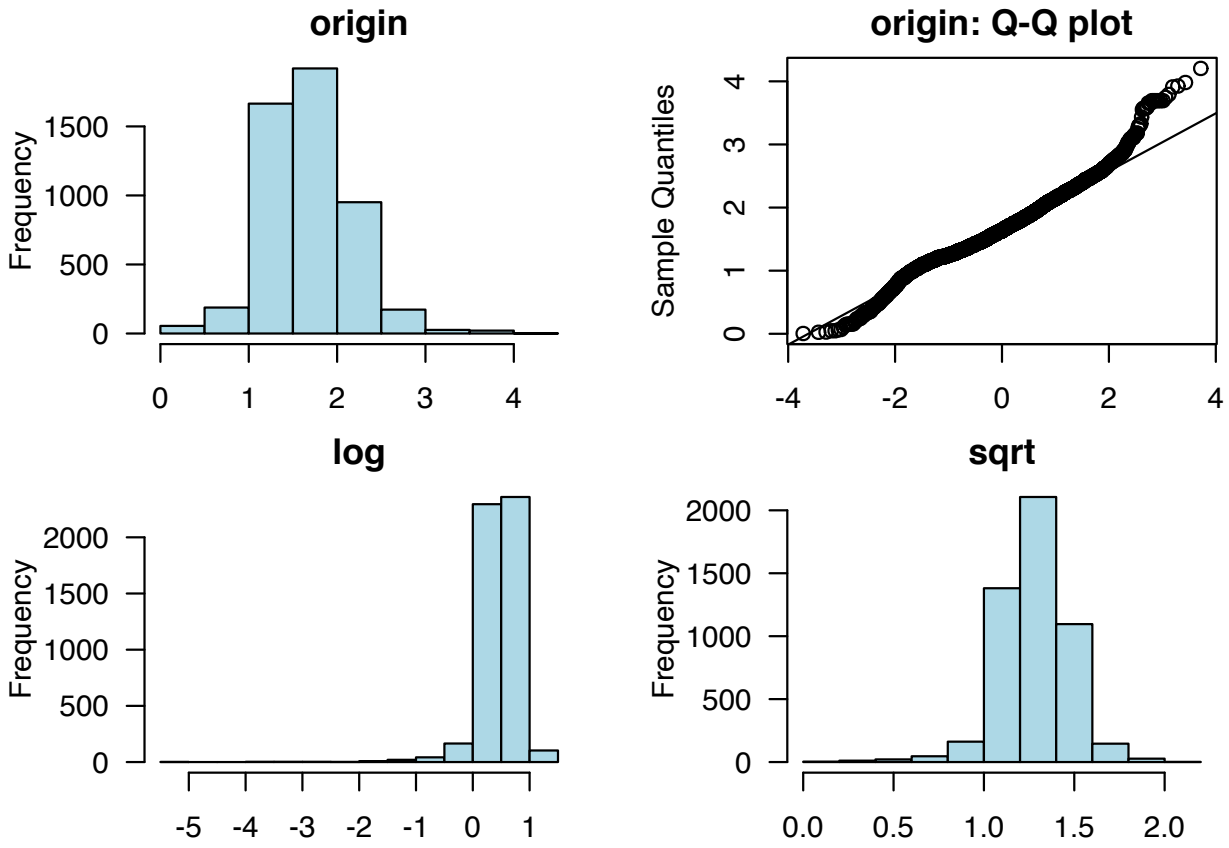


Figure 2.20: ln_wage

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

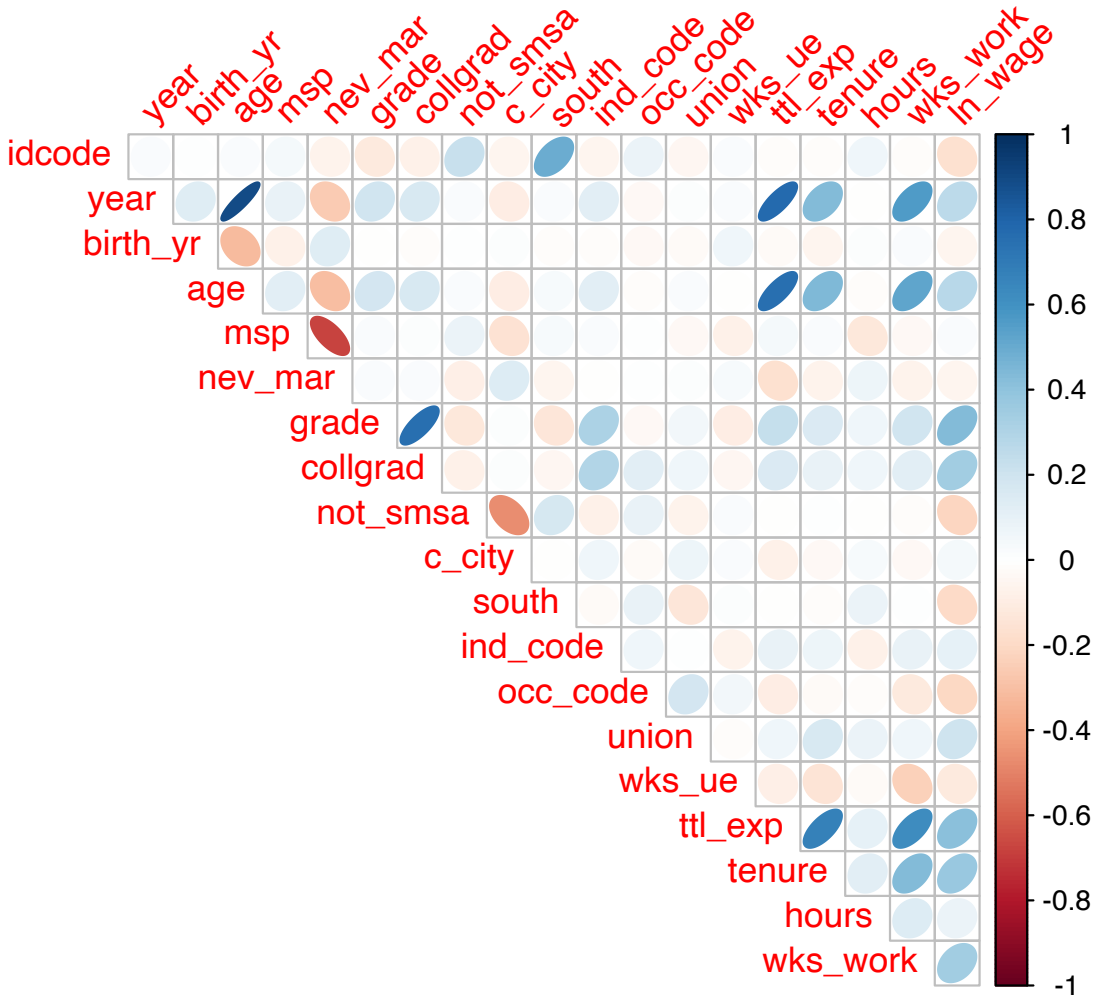


Figure 3.1: The correlation coefficient of numerical variables

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.