



REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.12

November 3, 2020

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	6
2.2.1	Statistics and Visualization of (Sample) Data	6
3	Relationship Between Variables	27
3.1	Correlation Coefficient	27
3.1.1	Correlation Coefficient by Variable Combination	27
3.1.2	Correlation Plot of Numerical Variables	27
4	Target based Analysis	29
4.1	Grouped Descriptive Statistics	29
4.1.1	Grouped Numerical Variables	29
4.1.2	Grouped Categorical Variables	29
4.2	Grouped Relationship Between Variables	29
4.2.1	Grouped Correlation Coefficient	29
4.2.2	Grouped Correlation Plot of Numerical Variables	29

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	haven_labelled	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

```
Error in proxy[i, ..., drop = FALSE]: incorrect number of dimensions
Error in Hmisc::latex(x, file = ""): object 'x' not found
```

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

idcode

normality test : Shapiro-Wilk normality test
 statistic : 0.95471, p-value : 8.13902E-37

type	skewness	kurtosis
original	-0.0245	1.8075
log transformation	-2.1412	9.7793
sqrt transformation	-0.6092	2.4881

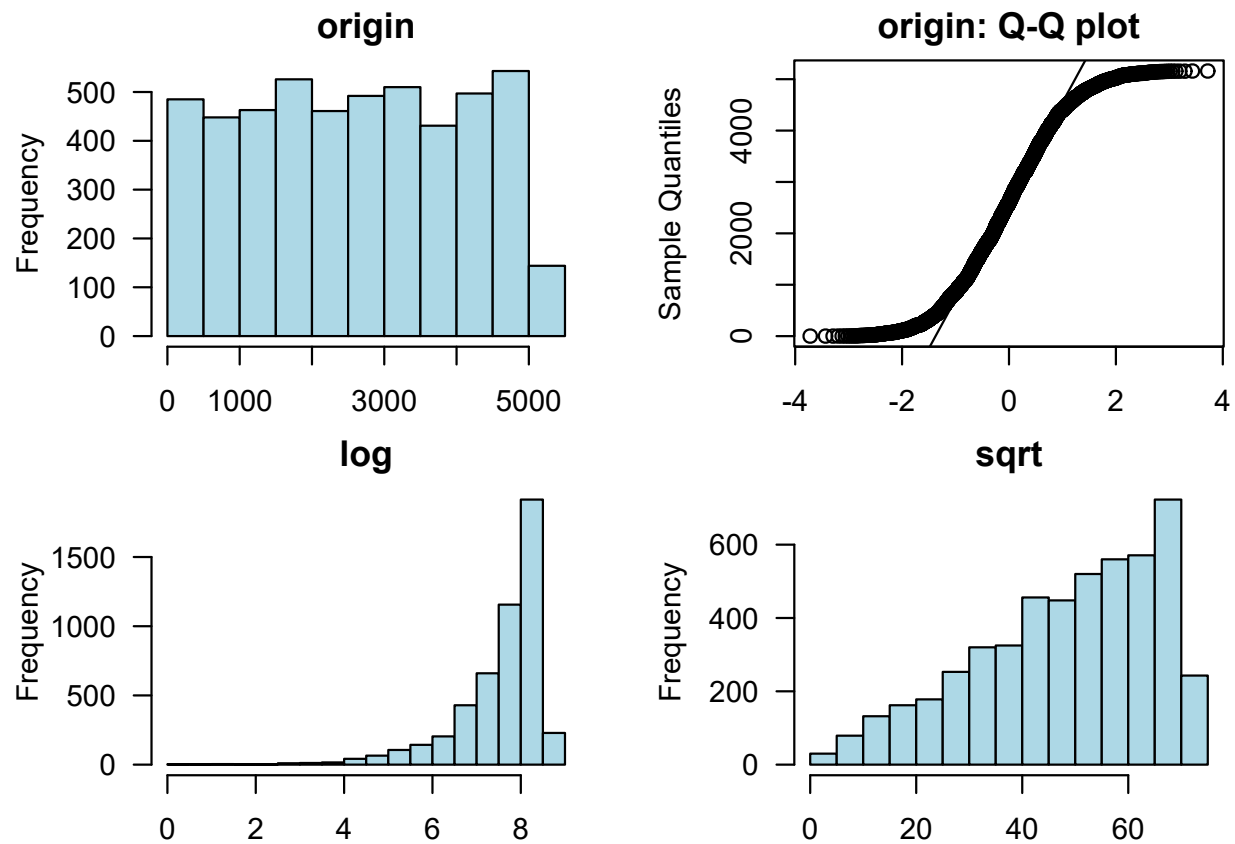


Figure 2.1: idcode

year

normality test : Shapiro-Wilk normality test
 statistic : 0.93137, p-value : 4.35549E-43

type	skewness	kurtosis
original	0.0820	1.6985
log transformation	-0.0030	1.6940
sqrt transformation	0.0395	1.6943

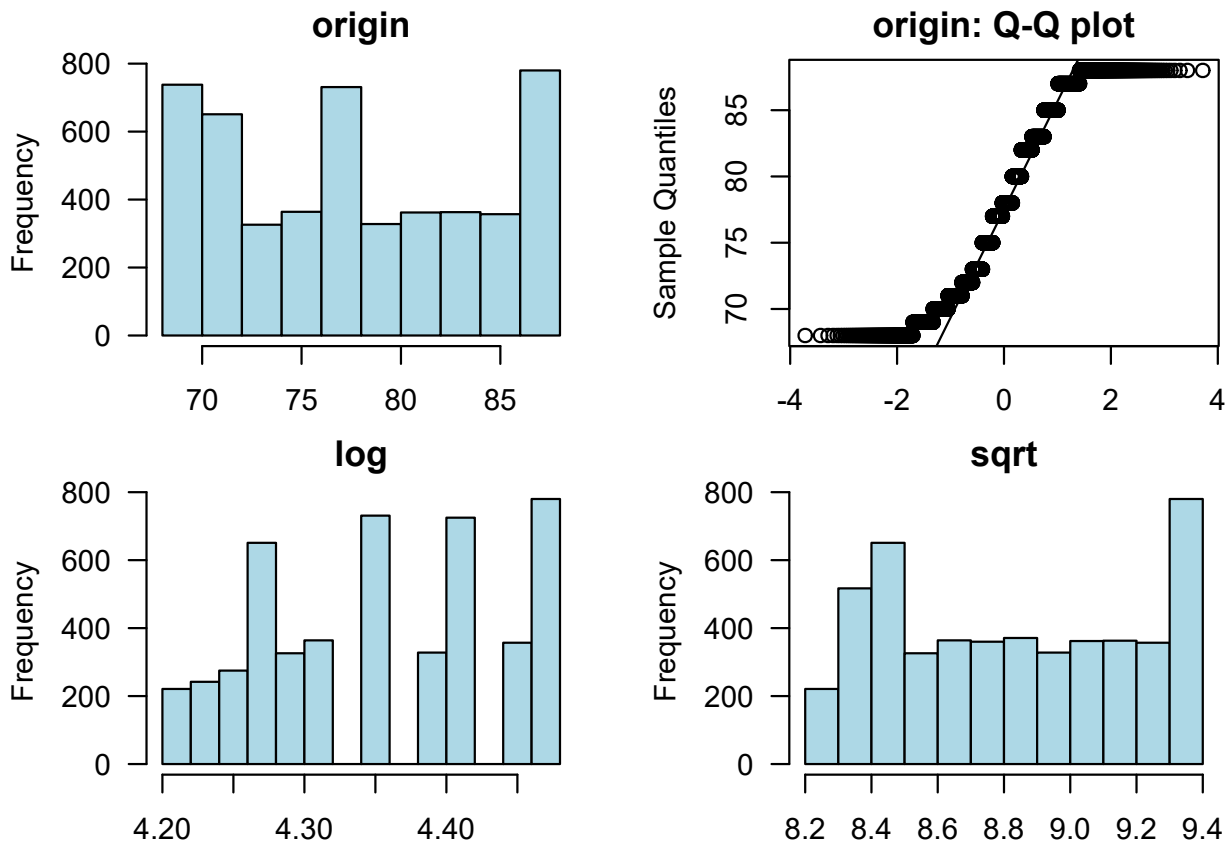


Figure 2.2: year

birth_yr

normality test : Shapiro-Wilk normality test
 statistic : 0.95846, p-value : 1.41833E-35

type	skewness	kurtosis
original	-0.0959	1.9691
log transformation	-0.1889	2.0148
sqrt transformation	-0.1422	1.9893

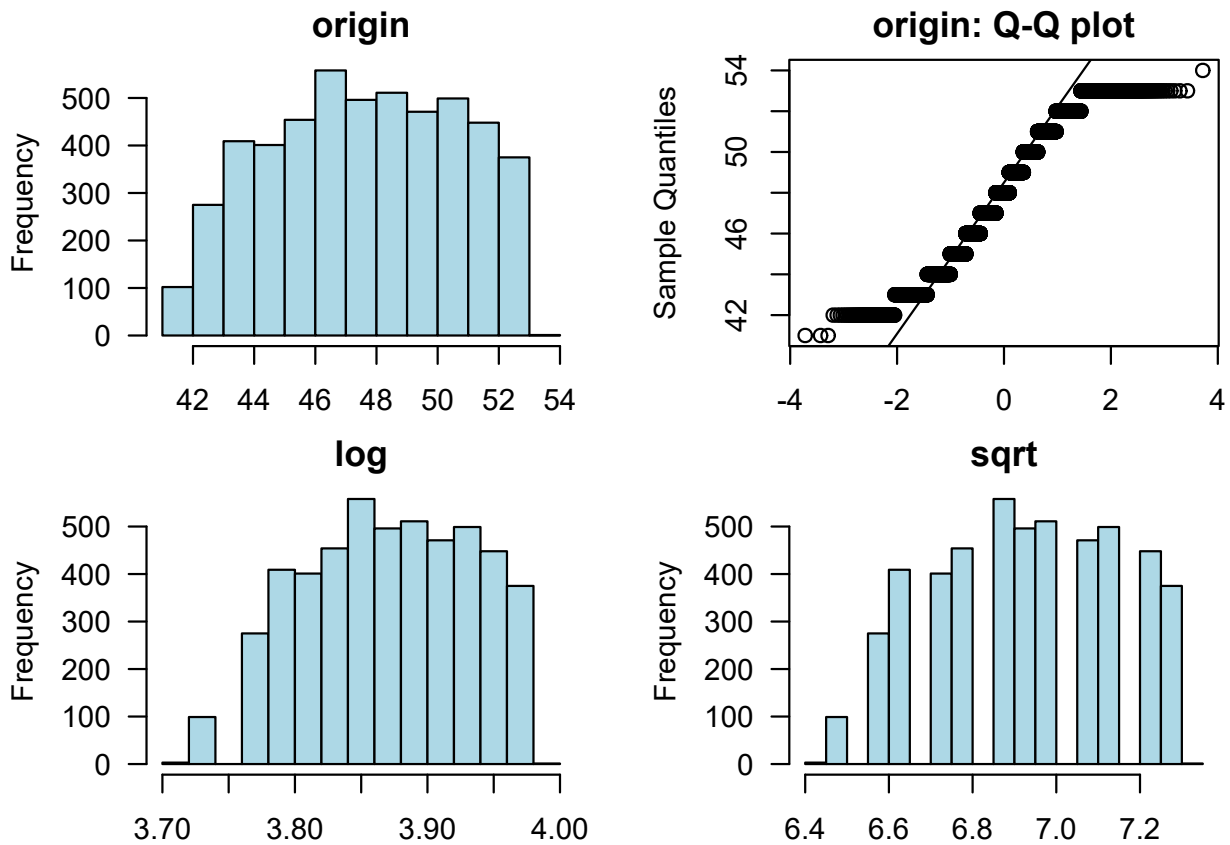


Figure 2.3: birth_yr

age

normality test : Shapiro-Wilk normality test
 statistic : 0.96659, p-value : 1.51567E-32

type	skewness	kurtosis
original	0.2548	2.0586
log transformation	-0.0870	1.9937
sqrt transformation	0.0849	1.9855

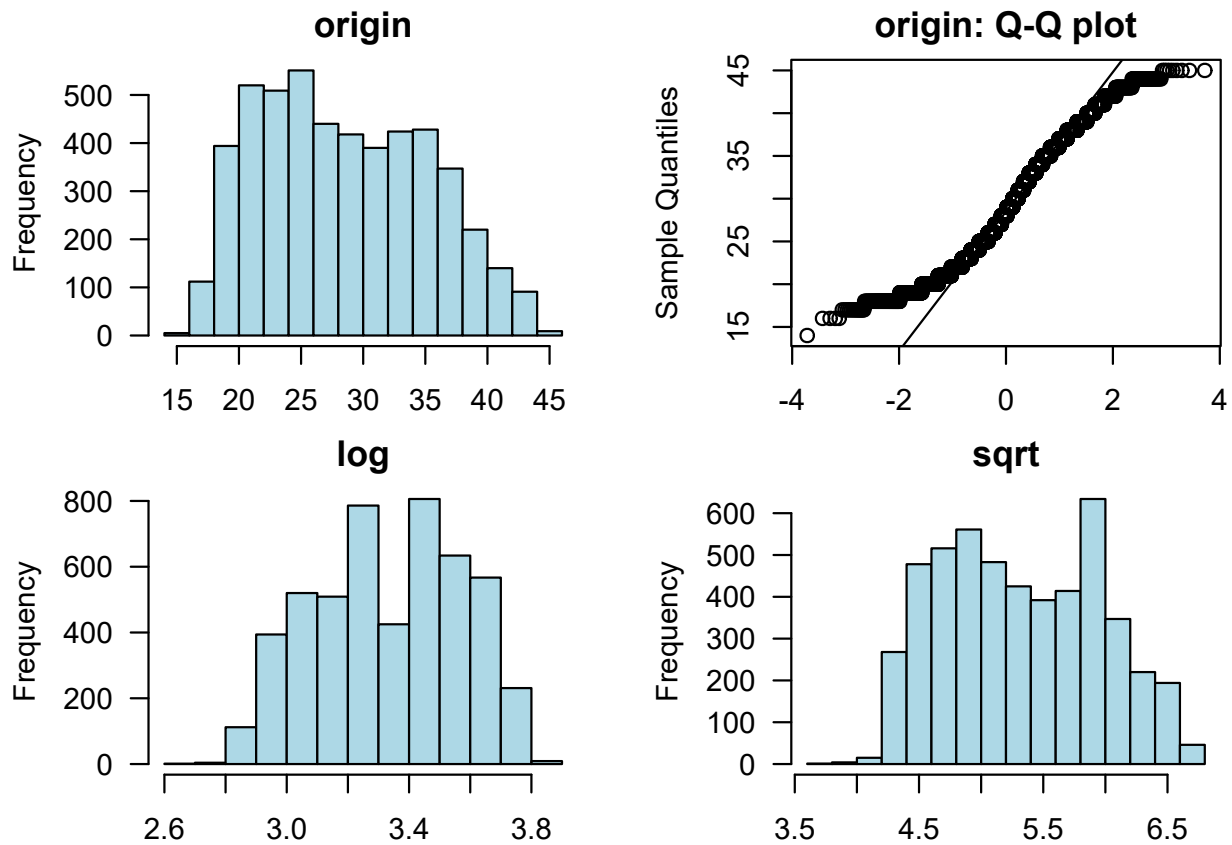


Figure 2.4: age

msp

normality test : Shapiro-Wilk normality test
 statistic : 0.62599, p-value : 5.64809E-74

type	skewness	kurtosis
original	-0.3458	1.1196
log transformation		
sqrt transformation	-0.3458	1.1196

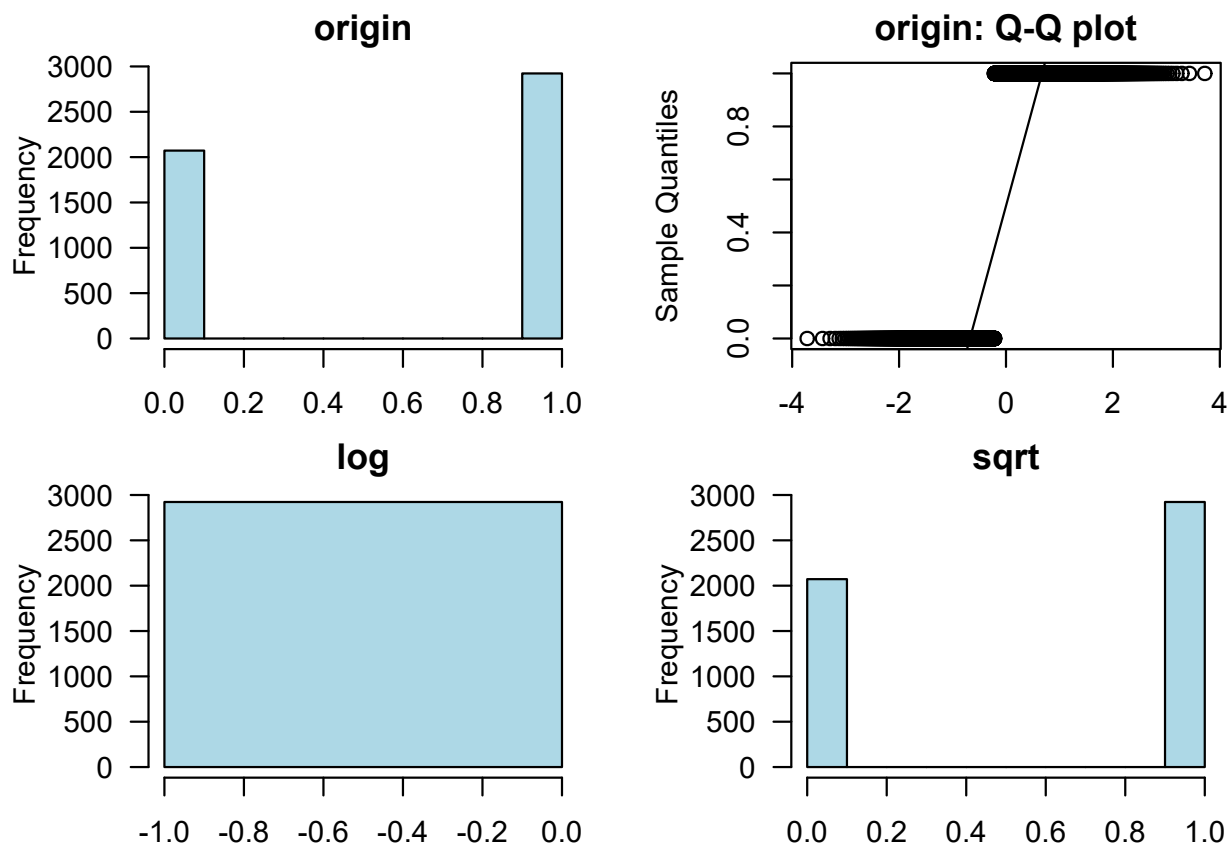


Figure 2.5: msp

nev_mar

normality test : Shapiro-Wilk normality test
 statistic : 0.51937, p-value : 2.74586E-79

type	skewness	kurtosis
original	1.2920	2.6692
log transformation		
sqrt transformation	1.2920	2.6692

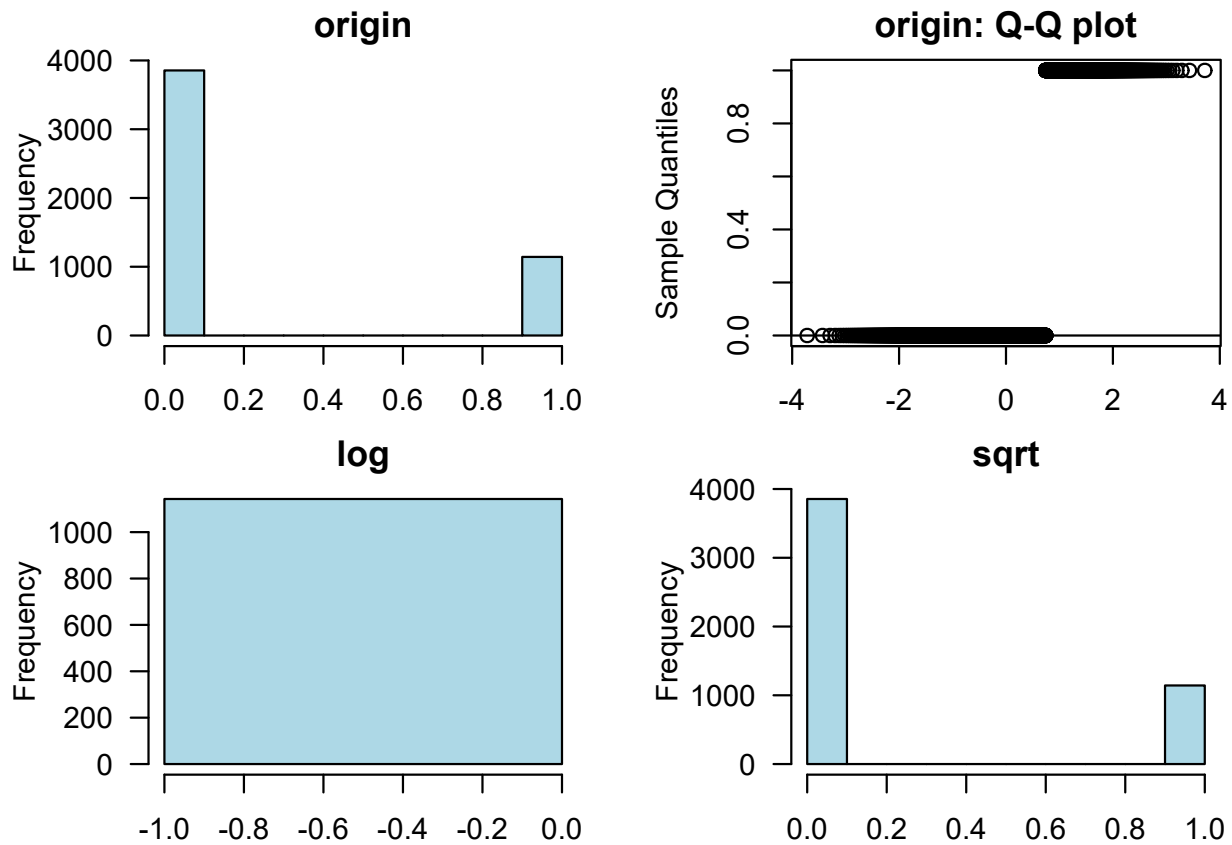


Figure 2.6: nev_mar

grade

normality test : Shapiro-Wilk normality test
 statistic : 0.88882, p-value : 5.25145E-51

type	skewness	kurtosis
original	0.0831	4.4863
log transformation		
sqrt transformation	-1.2718	15.0690

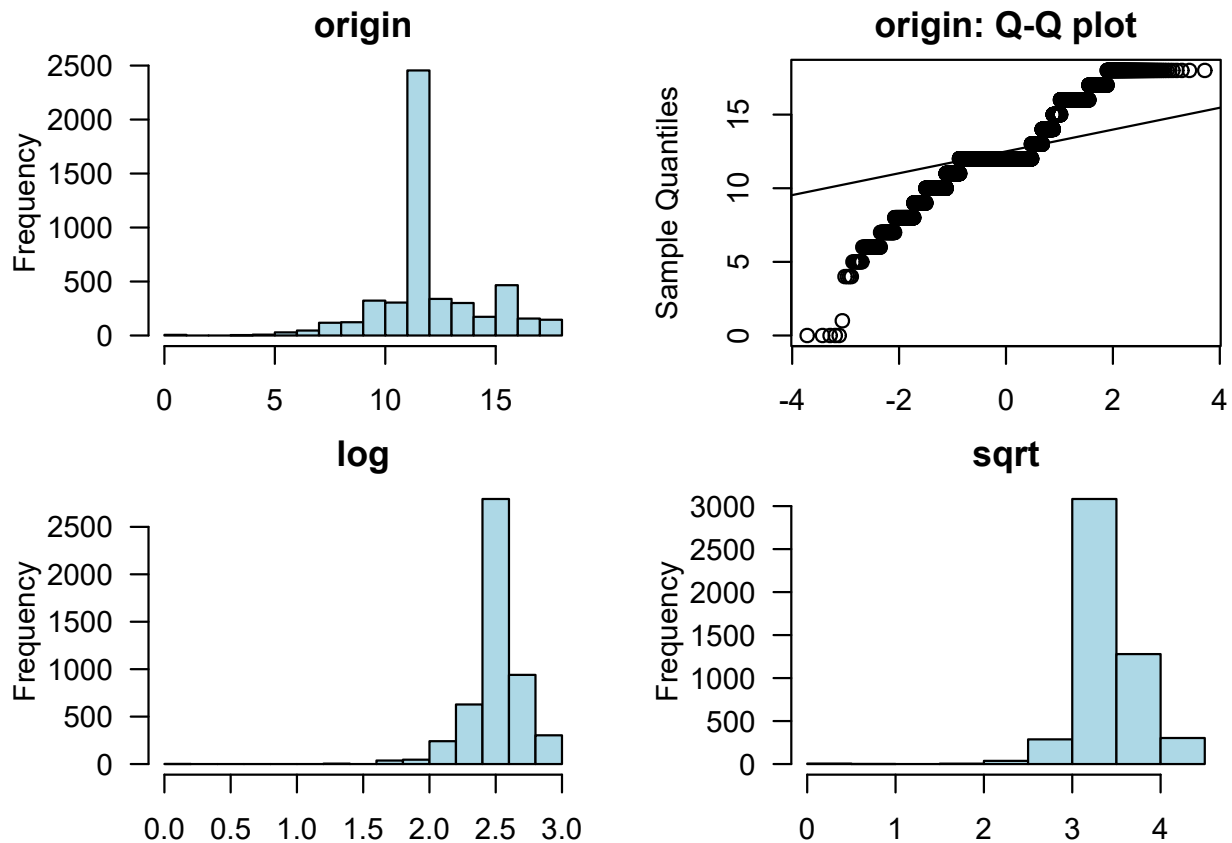


Figure 2.7: grade

collgrad

normality test : Shapiro-Wilk normality test
 statistic : 0.45407, p-value : 4.63132E-82

type	skewness	kurtosis
original	1.7551	4.0806
log transformation		
sqrt transformation	1.7551	4.0806

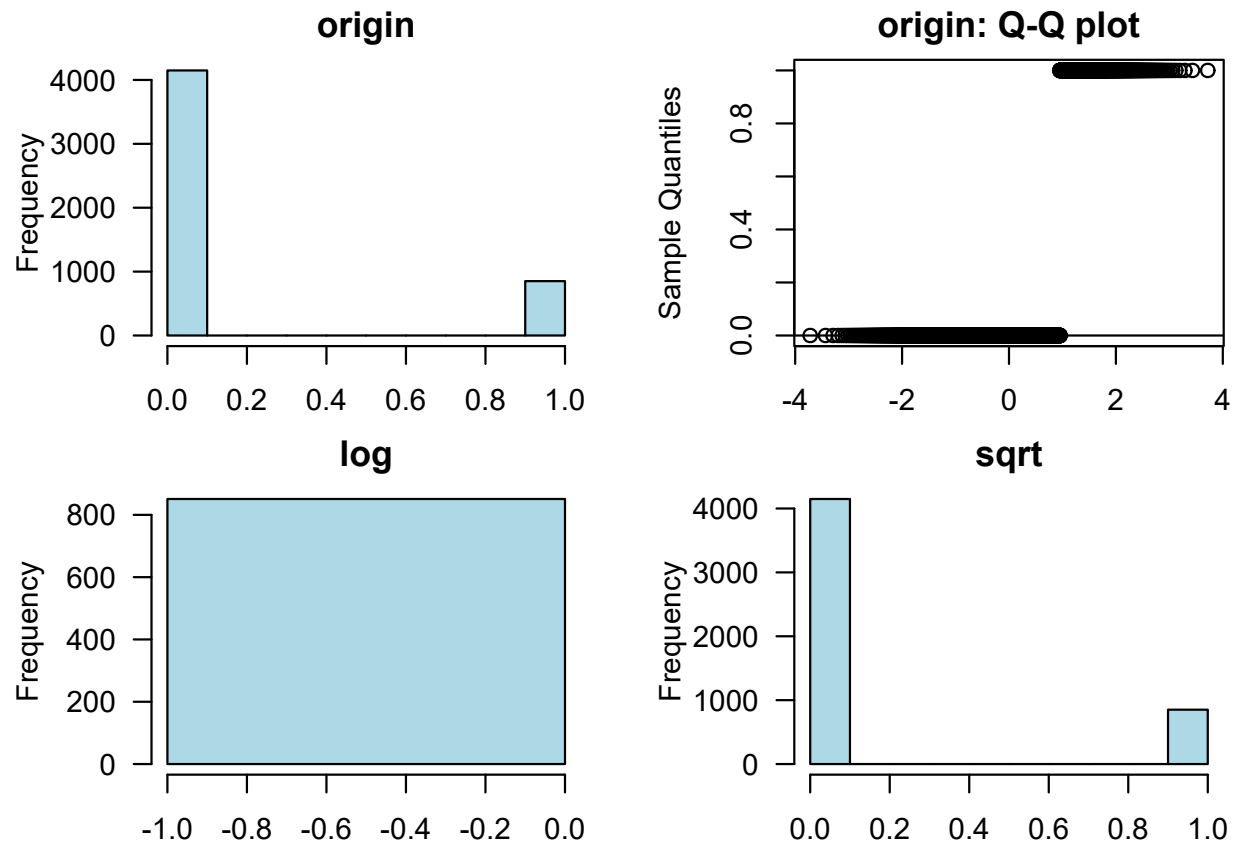


Figure 2.8: collgrad

not_smsa

normality test : Shapiro-Wilk normality test
 statistic : 0.56419, p-value : 3.42199E-77

type	skewness	kurtosis
original	0.9636	1.9286
log transformation		
sqrt transformation	0.9636	1.9286

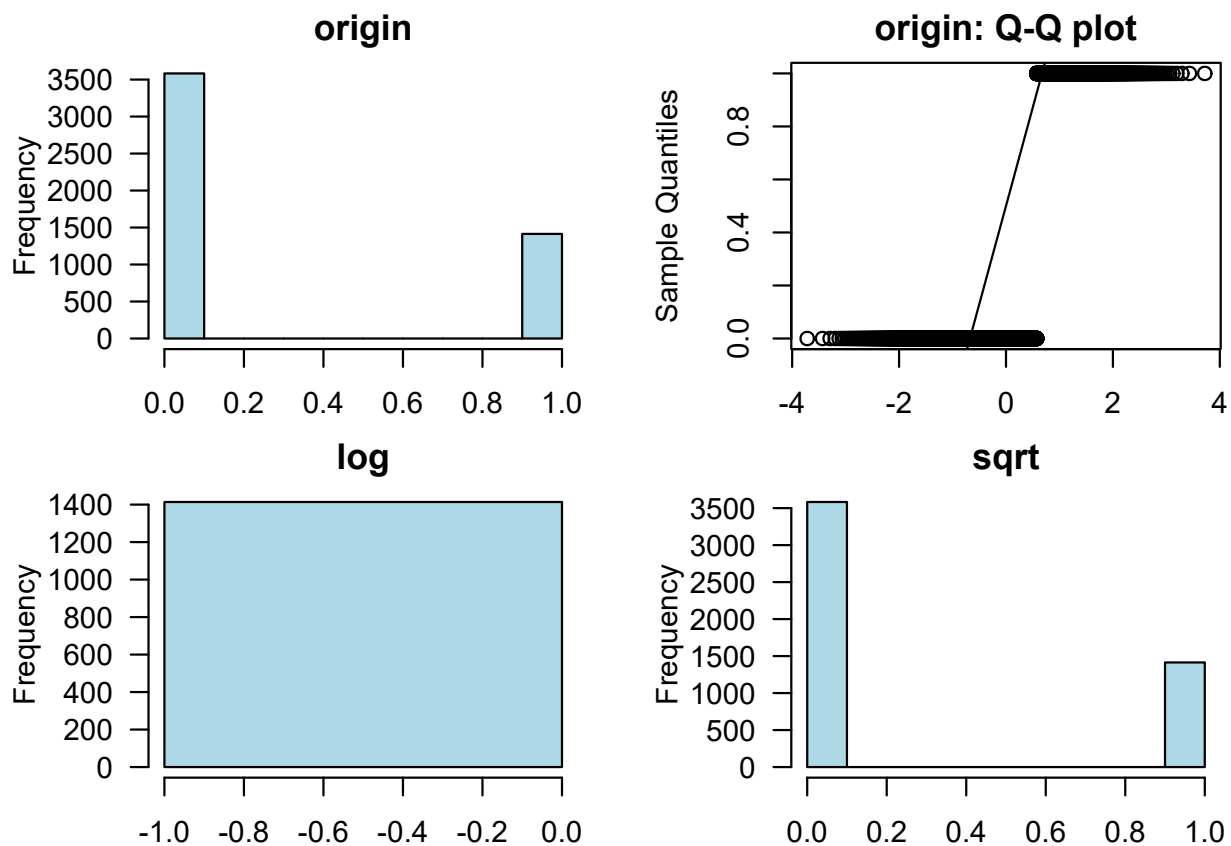


Figure 2.9: not_smsa

c_city

normality test : Shapiro-Wilk normality test
 statistic : 0.60379, p-value : 3.36566E-75

type	skewness	kurtosis
original	0.6216	1.3864
log transformation		
sqrt transformation	0.6216	1.3864

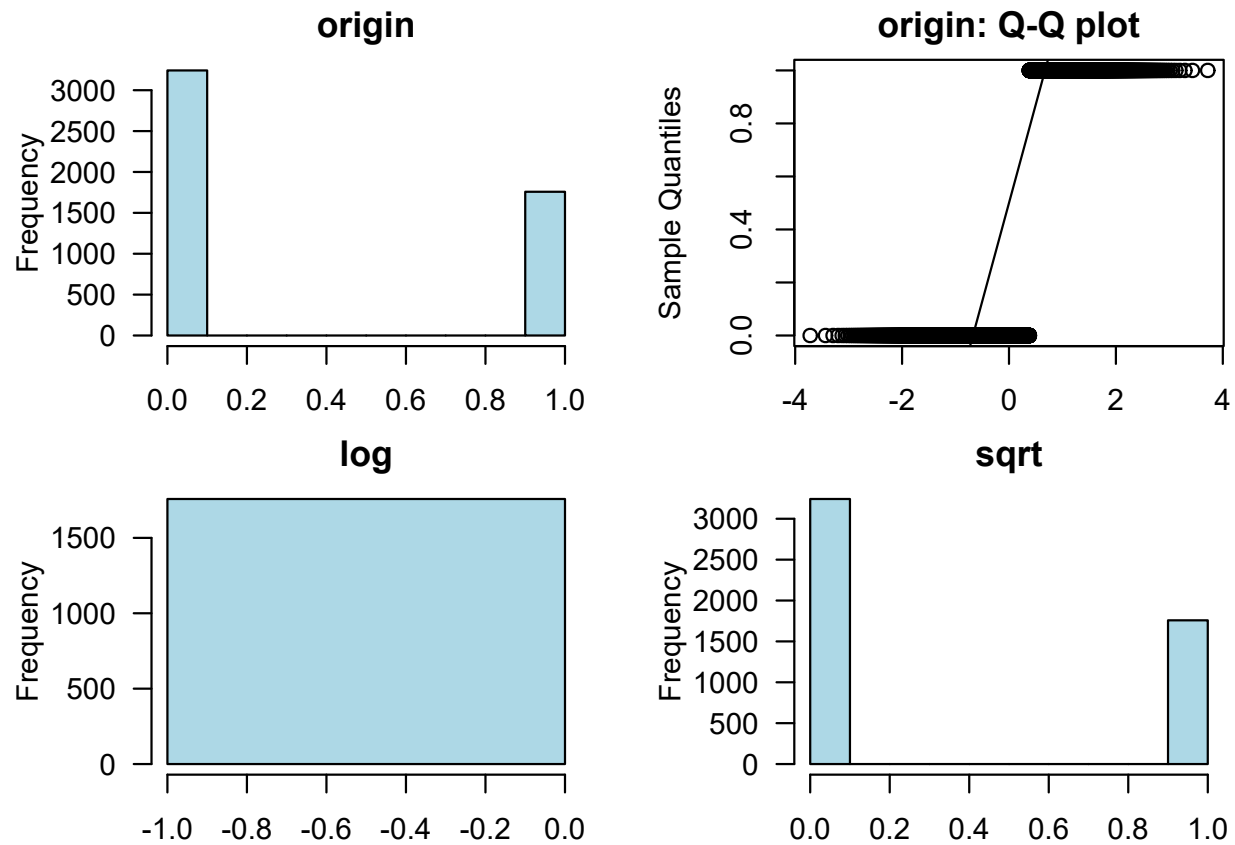


Figure 2.10: c_city

south

normality test : Shapiro-Wilk normality test
 statistic : 0.62577, p-value : 5.32357E-74

type	skewness	kurtosis
original	0.3493	1.1220
log transformation		
sqrt transformation	0.3493	1.1220

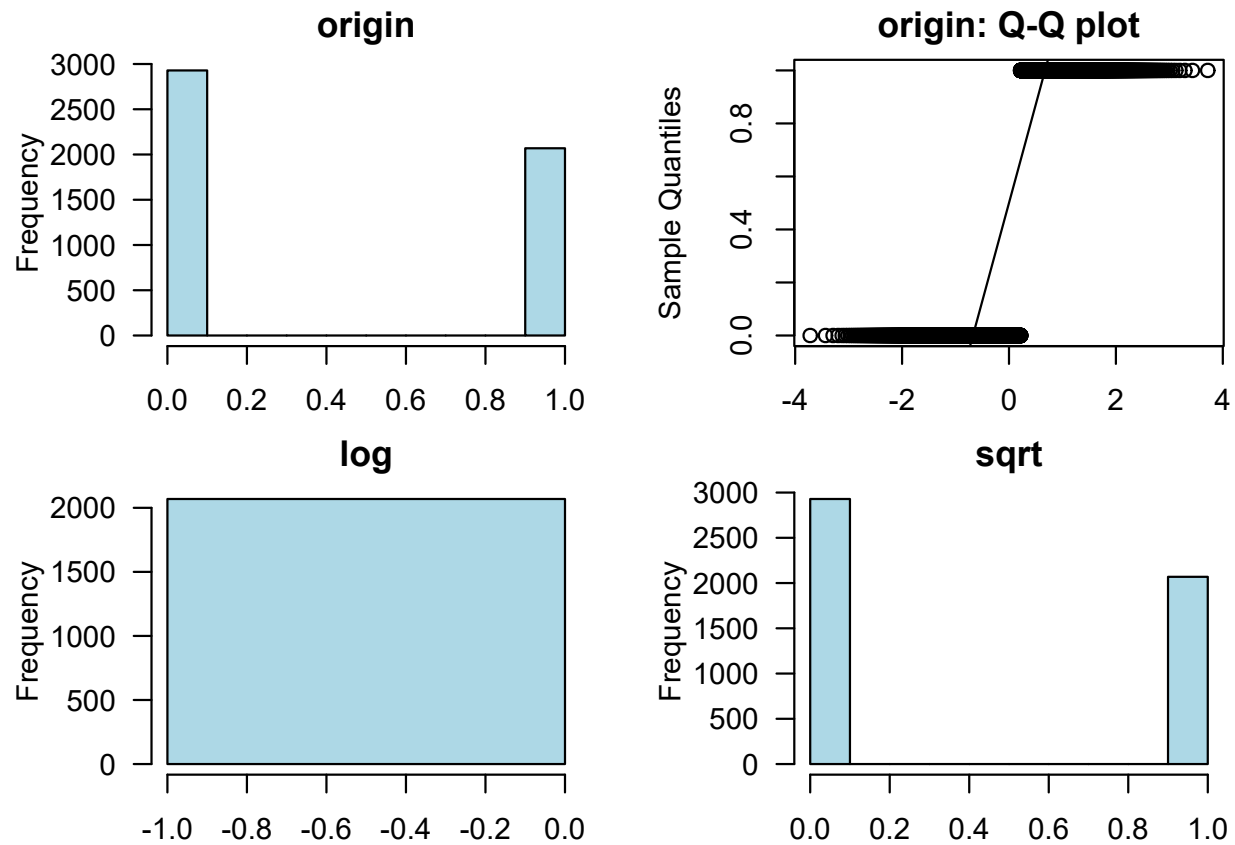


Figure 2.11: south

ind_code

normality test : Shapiro-Wilk normality test
 statistic : 0.8753, p-value : 8.57421E-53

type	skewness	kurtosis
original	-0.0070	1.5784
log transformation	-0.9567	4.7754
sqrt transformation	-0.3030	2.1755

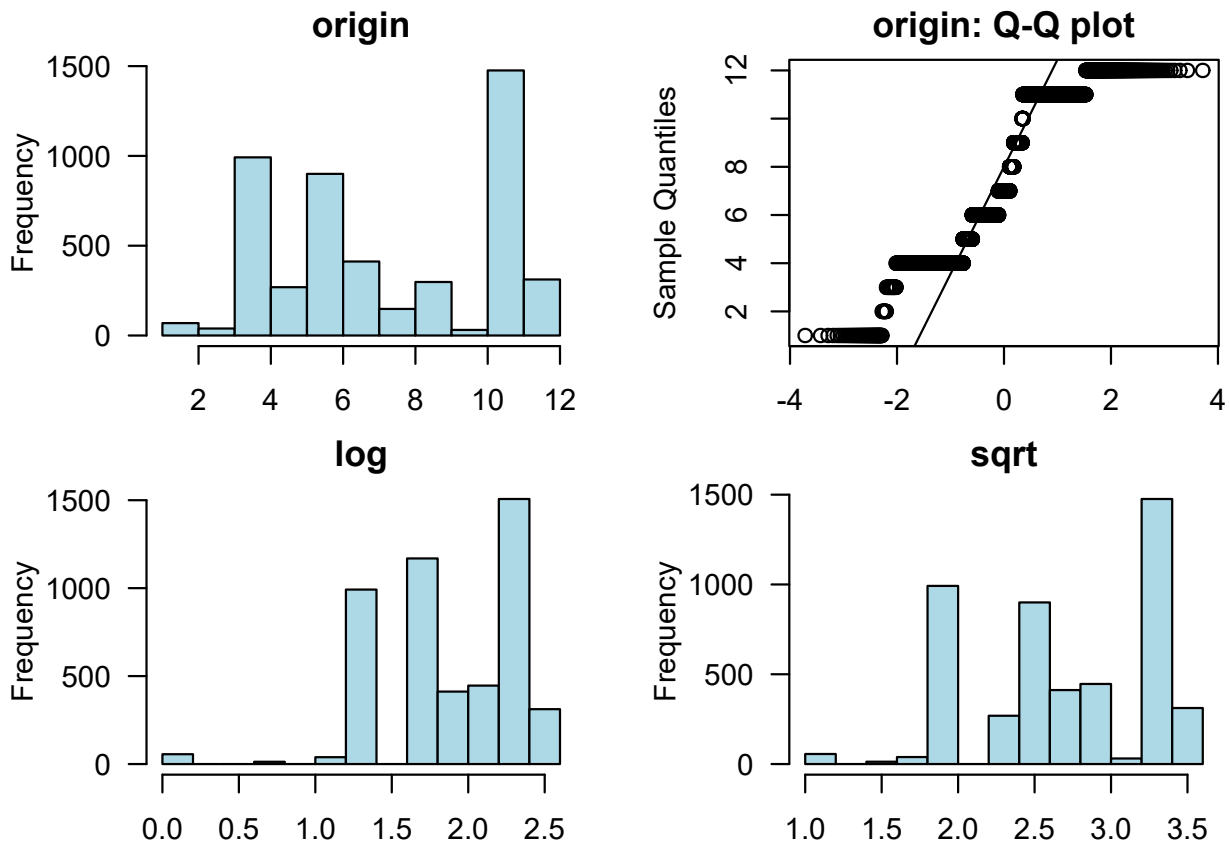


Figure 2.12: ind_code

occ_code

normality test : Shapiro-Wilk normality test
 statistic : 0.85143, p-value : 5.09415E-56

type	skewness	kurtosis
original	1.0923	3.6919
log transformation	-0.2909	2.6626
sqrt transformation	0.4553	2.6357

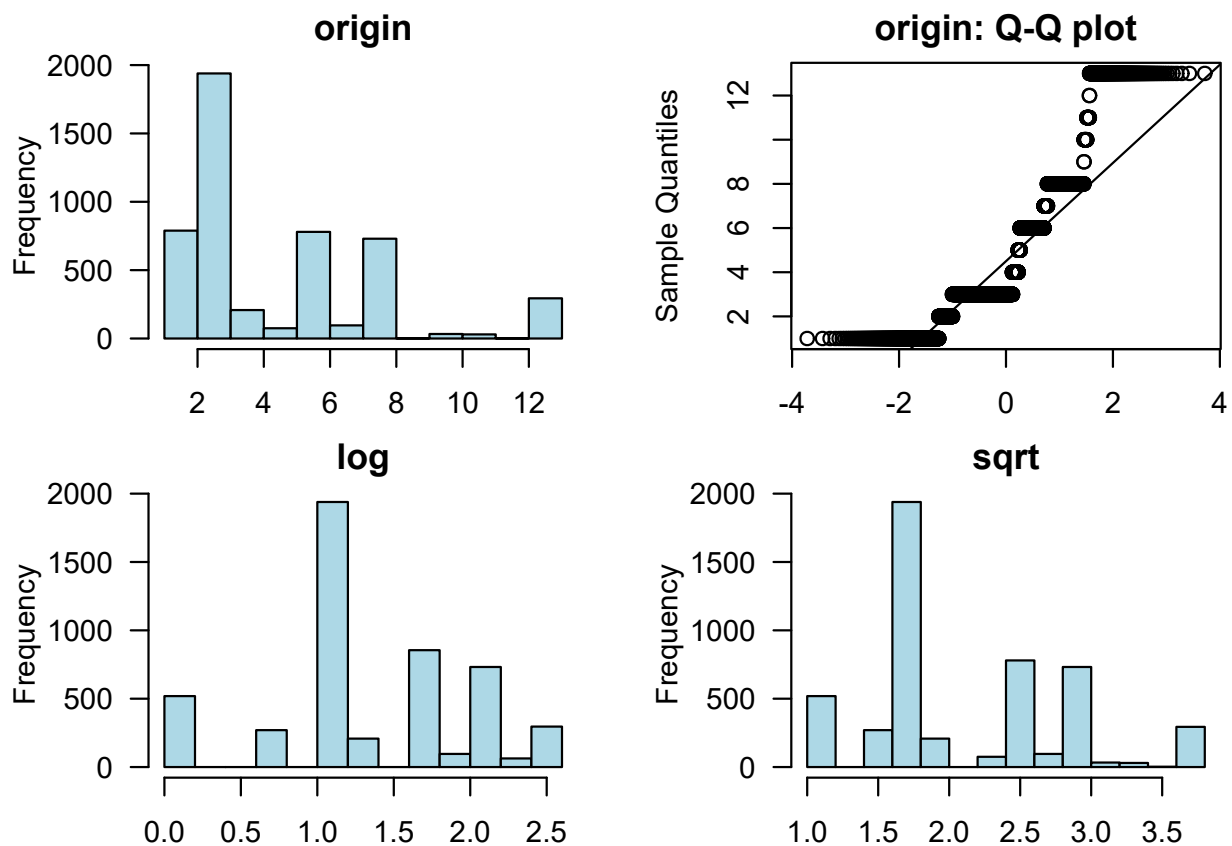


Figure 2.13: occ_code

union

normality test : Shapiro-Wilk normality test
 statistic : 0.51696, p-value : 1.01762E-70

type	skewness	kurtosis
original	1.3090	2.7134
log transformation		
sqrt transformation	1.3090	2.7134

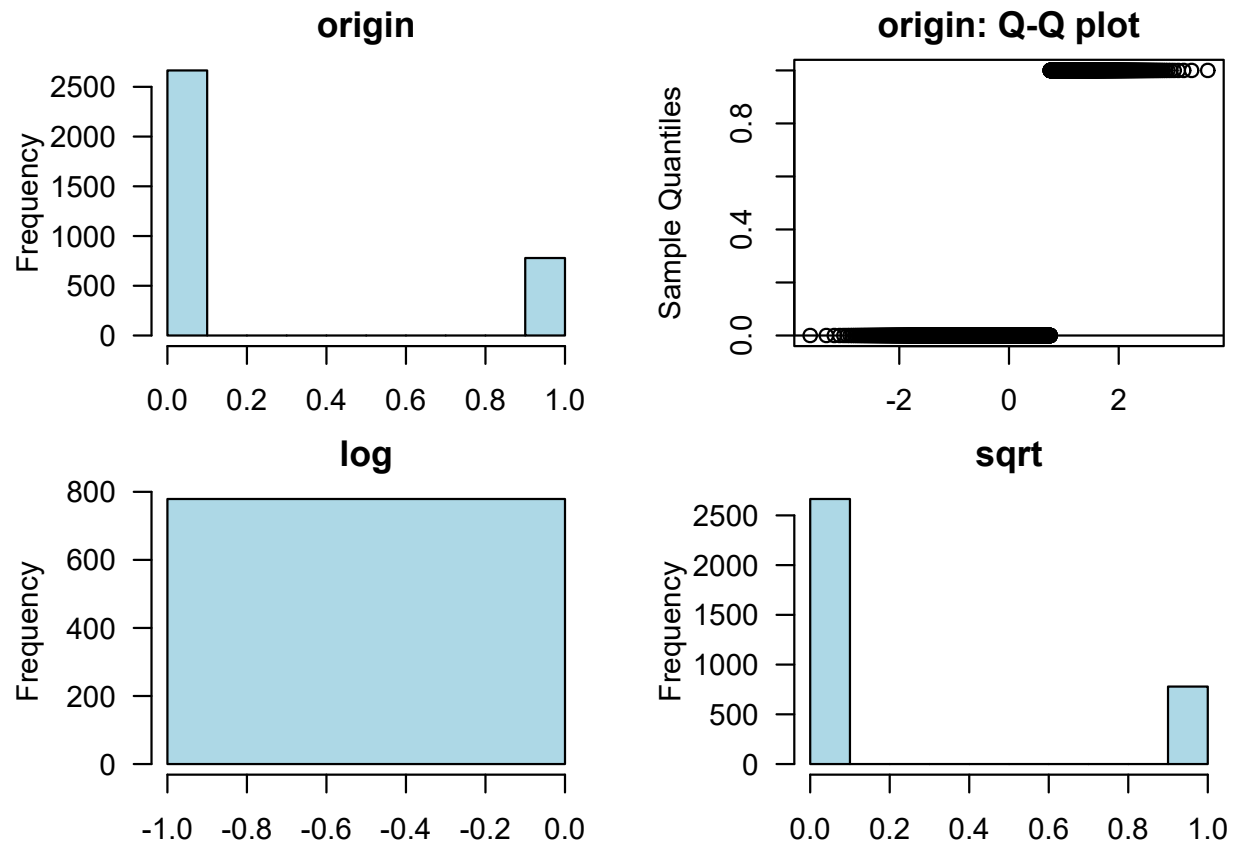


Figure 2.14: union

wks_ue

normality test : Shapiro-Wilk normality test
 statistic : 0.40521, p-value : 5.40433E-78

type	skewness	kurtosis
original	3.9695	20.6532
log transformation		
sqrt transformation	2.3255	7.9054

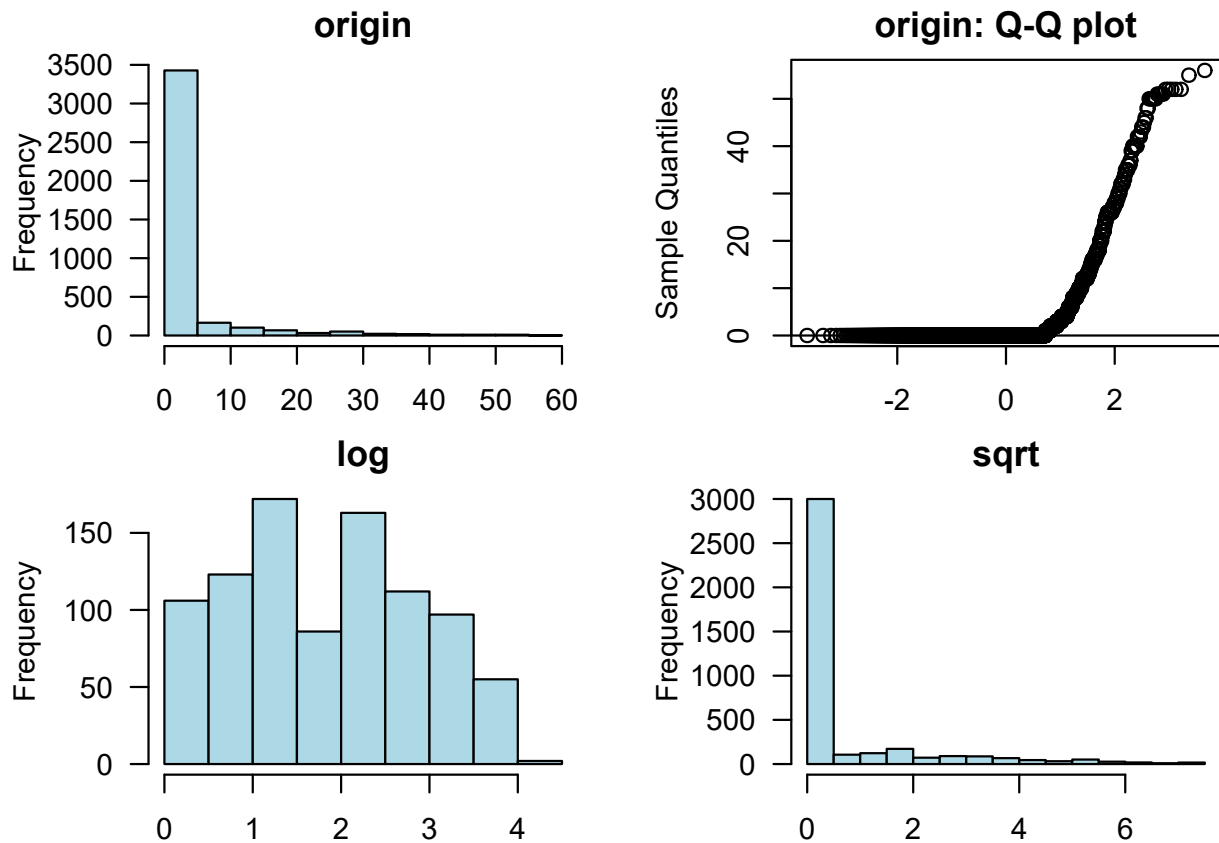


Figure 2.15: wks_ue

ttl_exp

normality test : Shapiro-Wilk normality test
 statistic : 0.92491, p-value : 1.6422E-44

type	skewness	kurtosis
original	0.8811	3.1852
log transformation		
sqrt transformation	0.1344	2.2991

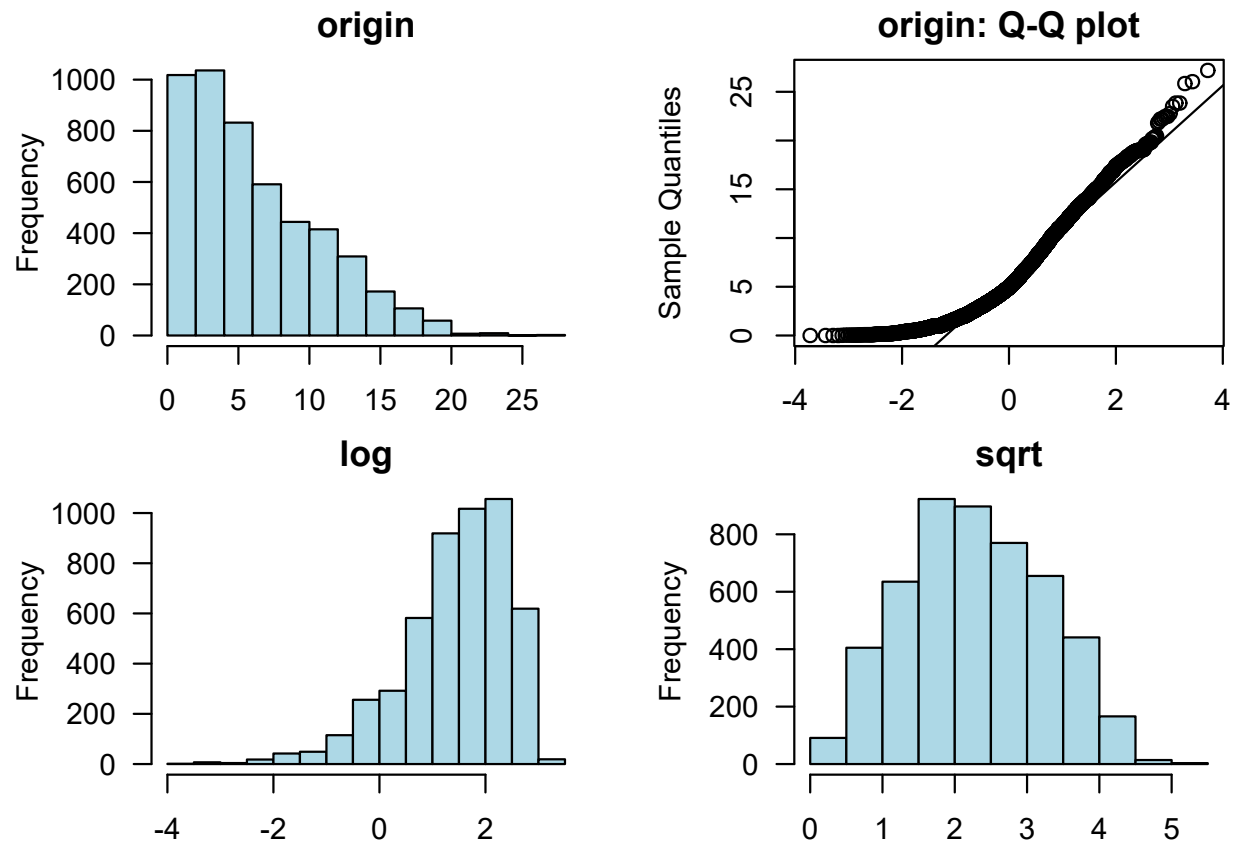


Figure 2.16: ttl_exp

tenure

normality test : Shapiro-Wilk normality test
 statistic : 0.76779, p-value : 3.76788E-64

type	skewness	kurtosis
original	1.9097	6.7001
log transformation		
sqrt transformation	0.7690	3.0527

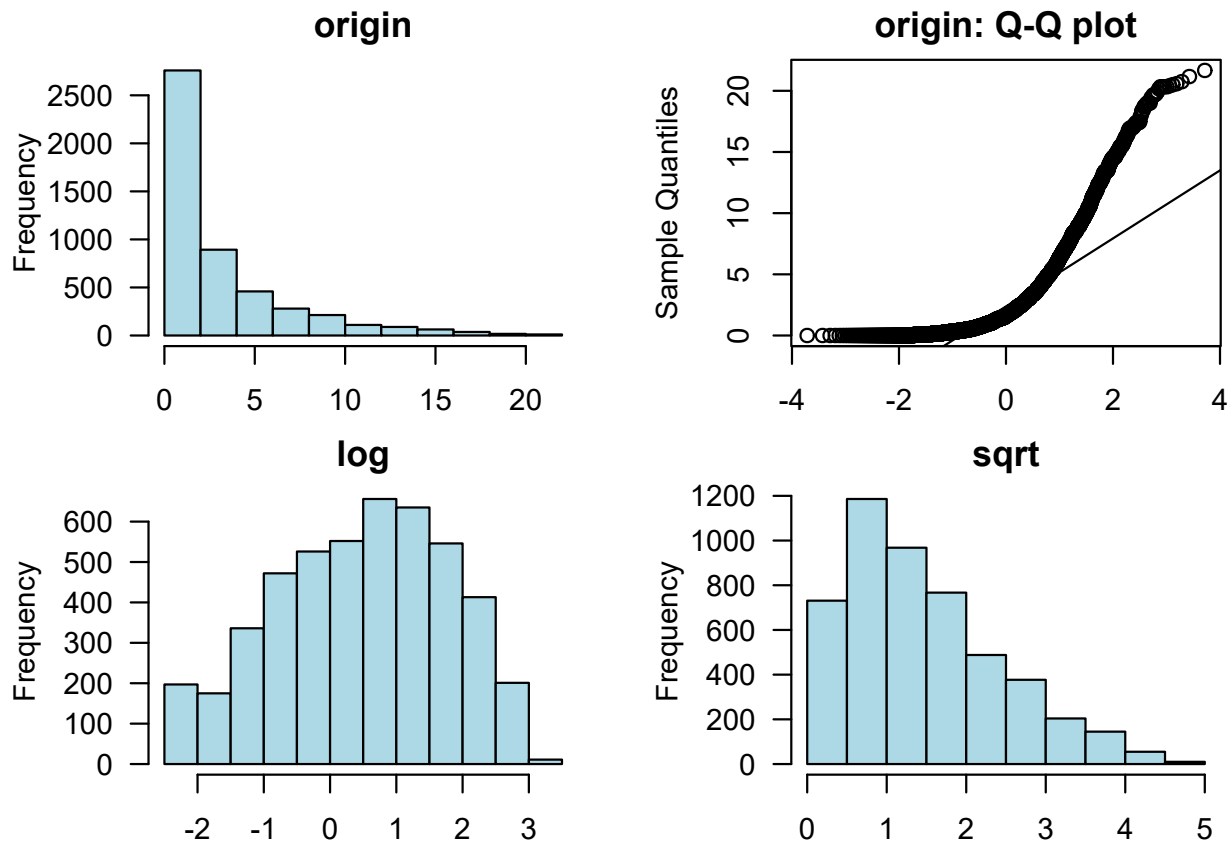


Figure 2.17: tenure

hours

normality test : Shapiro-Wilk normality test
 statistic : 0.77474, p-value : 8.5639E-64

type	skewness	kurtosis
original	-0.3831	12.6962
log transformation	-3.1754	16.9745
sqrt transformation	-1.7906	8.7119

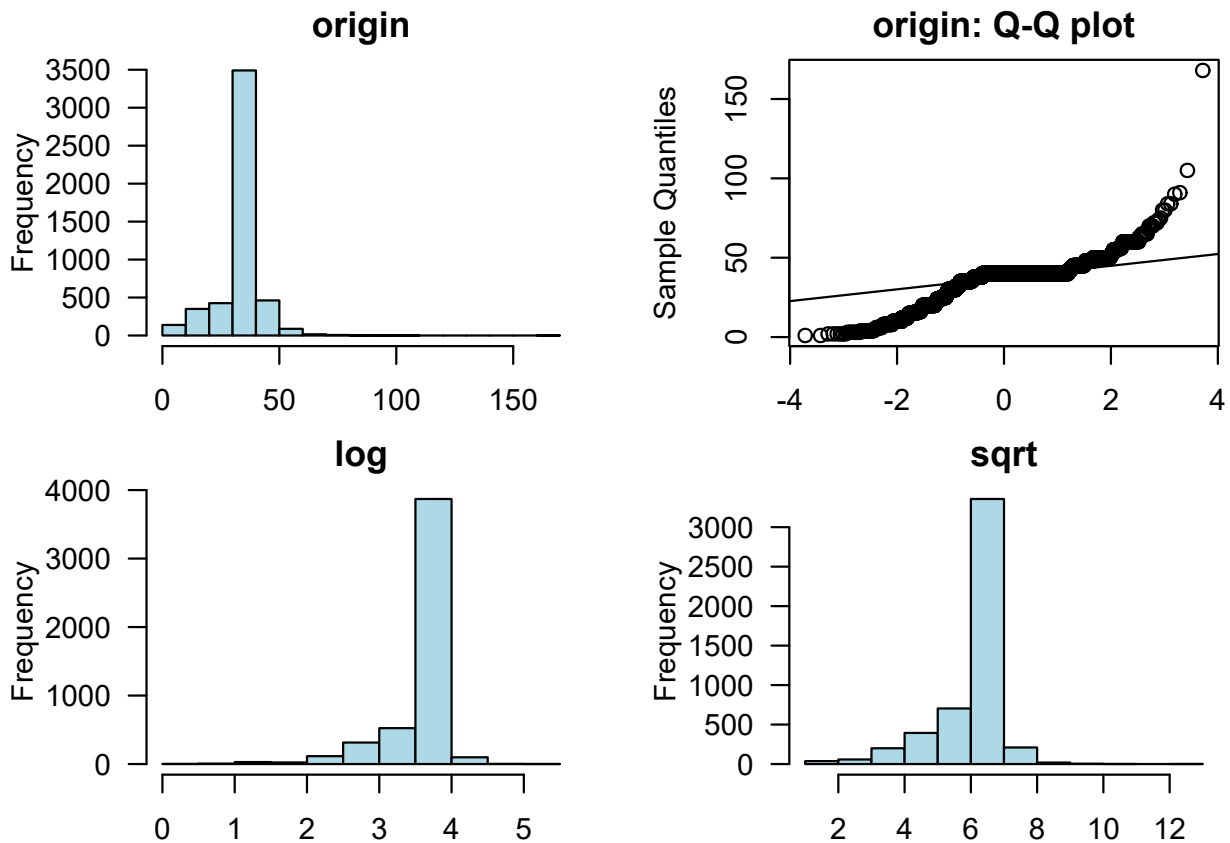


Figure 2.18: hours

wks_work

normality test : Shapiro-Wilk normality test
 statistic : 0.9383, p-value : 5.19531E-41

type	skewness	kurtosis
original	0.1933	2.3432
log transformation		
sqrt transformation	-0.7641	3.5887

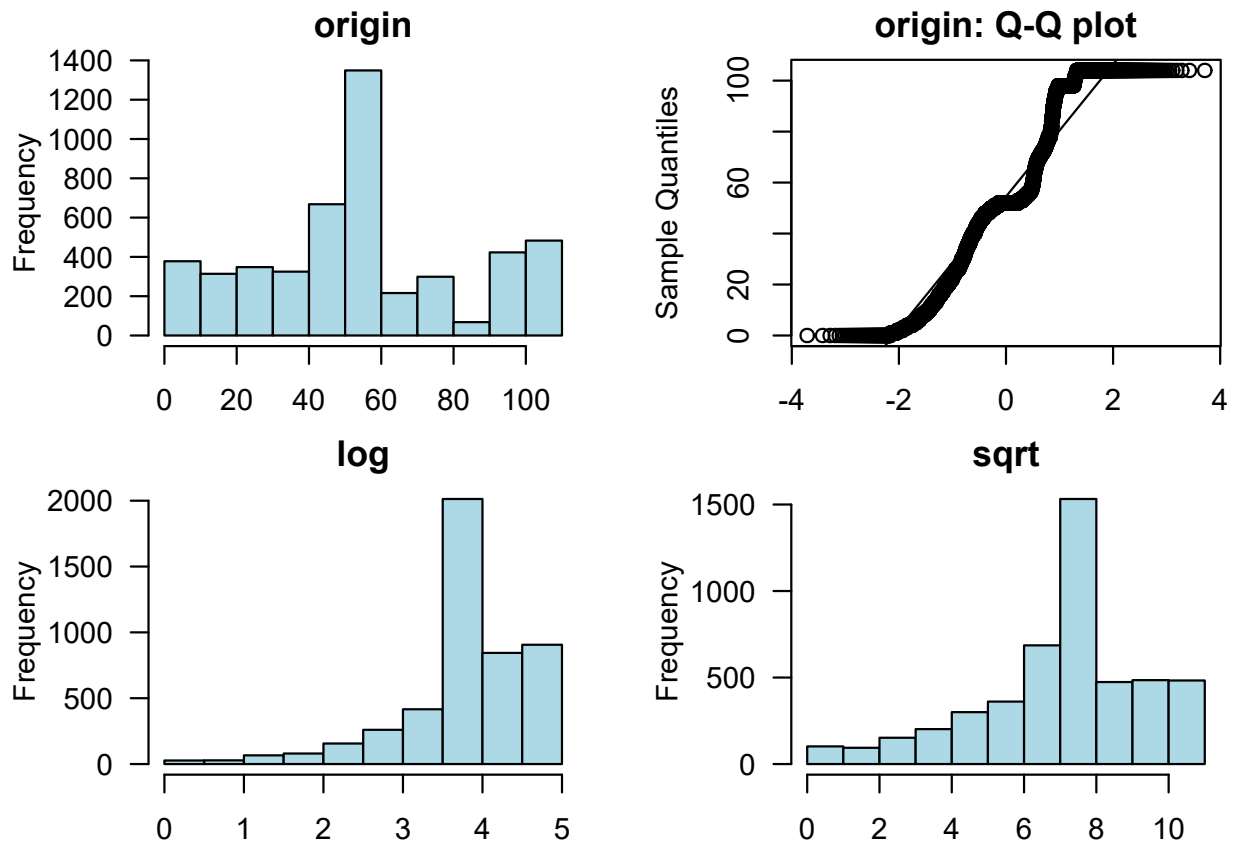


Figure 2.19: wks_work

ln_wage

normality test : Shapiro-Wilk normality test
 statistic : 0.97666, p-value : 7.08836E-28

type	skewness	kurtosis
original	0.3992	4.8522
log transformation	-3.4099	29.2277
sqrt transformation	-0.6972	6.6385

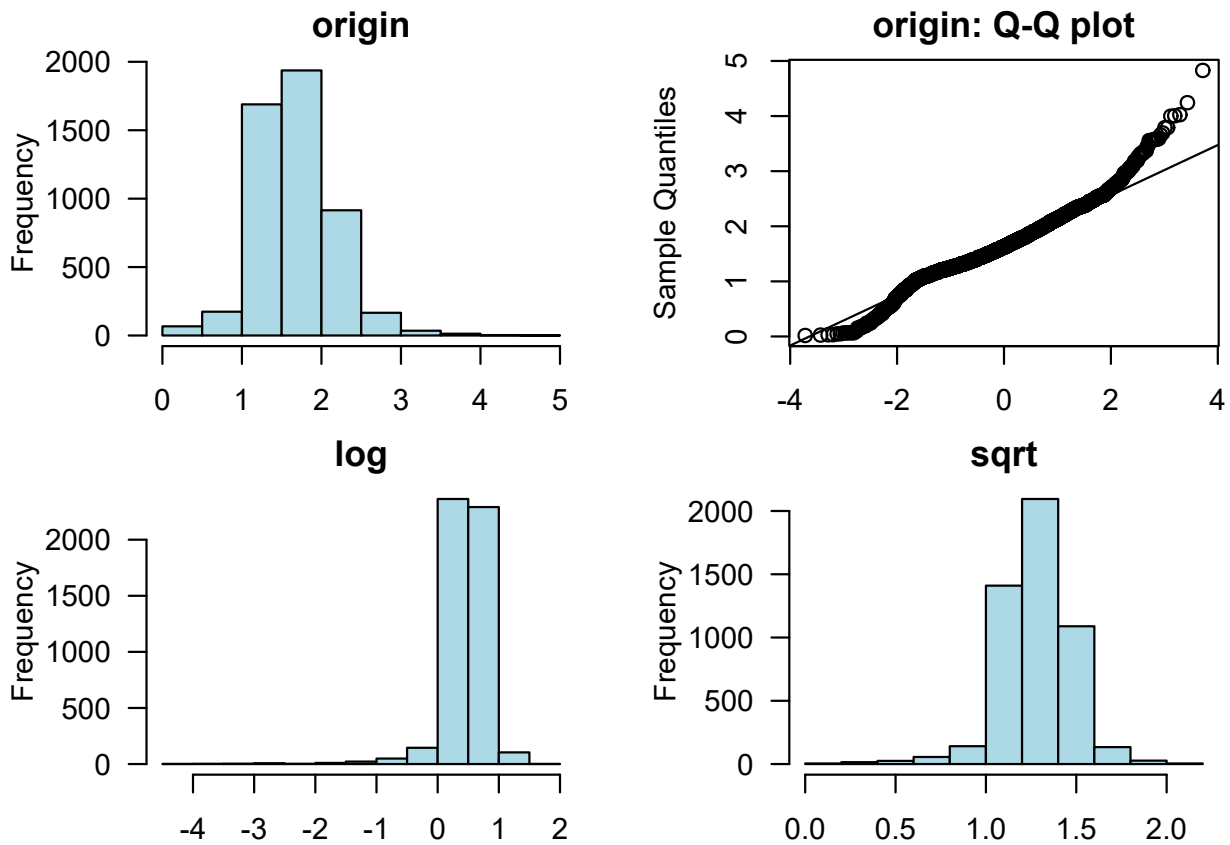


Figure 2.20: ln_wage

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

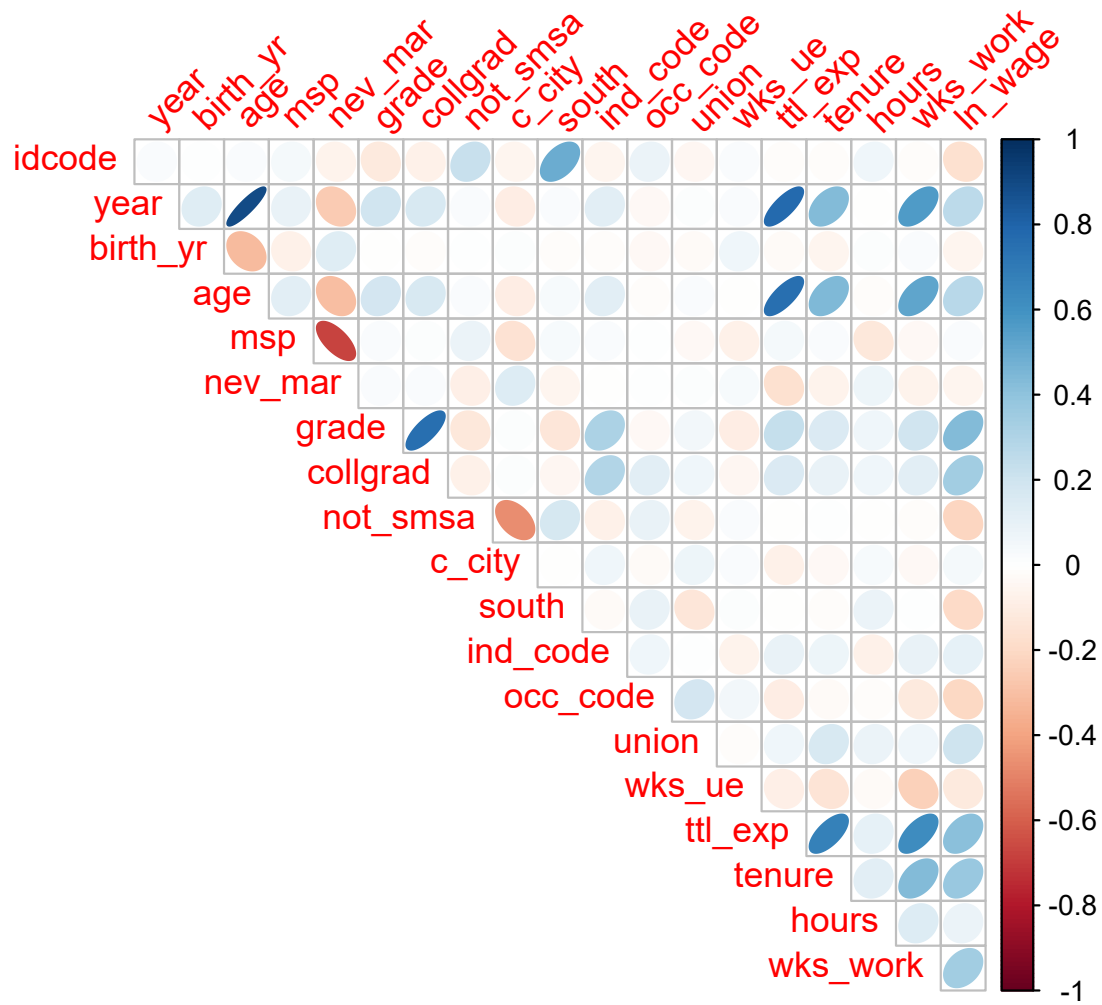


Figure 3.1: The correlation coefficient of numerical variables

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.