REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.12

November 12, 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[, ..., 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 : 0.95505, p-value : 1.04775E-36

type	skewness	kurtosis
original	-0.0221	1.8114
log transformation	-2.1511	9.8380
sqrt transformation	-0.6090	2.4903

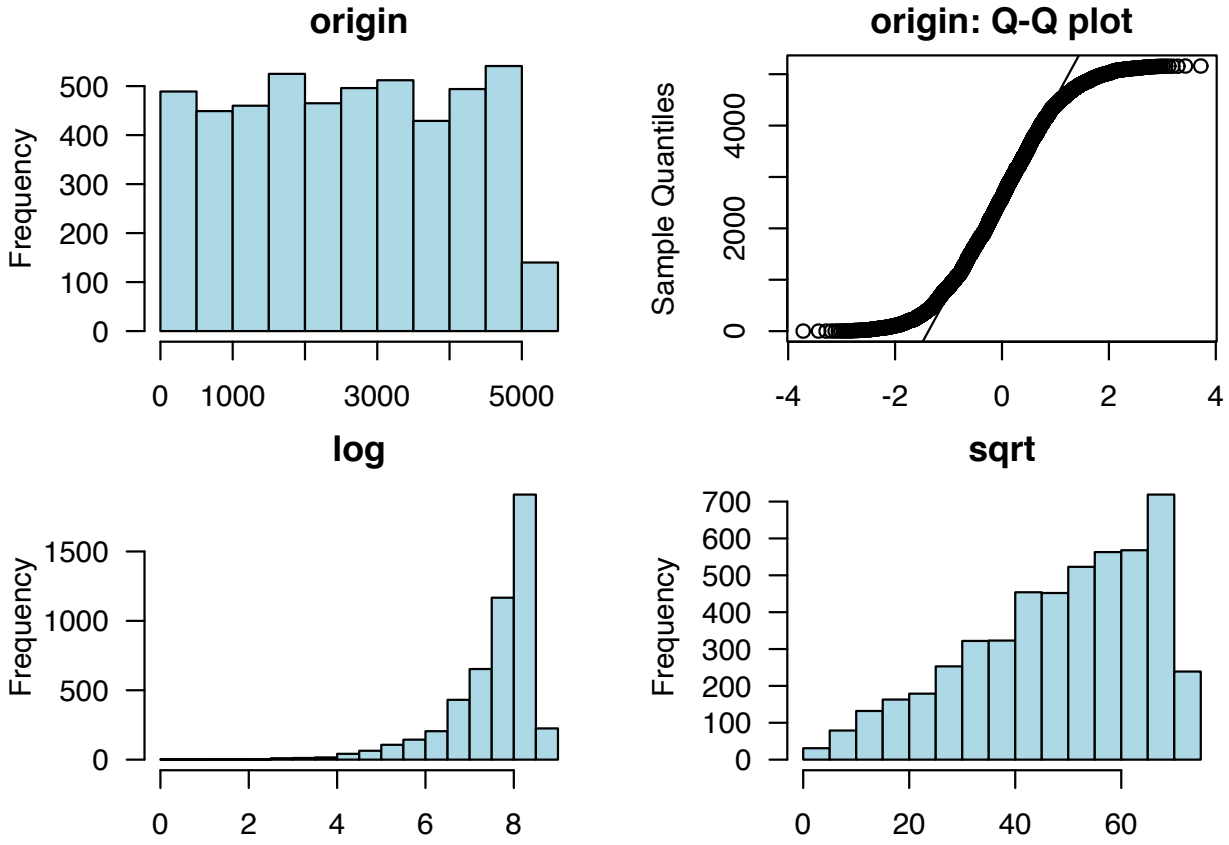

Figure 2.1: idcode

year

normality test : Shapiro-Wilk normality test
statistic : 0.93115, p-value : 3.88156E-43

type	skewness	kurtosis
original	0.0882	1.6983
log transformation	0.0033	1.6930
sqrt transformation	0.0458	1.6937

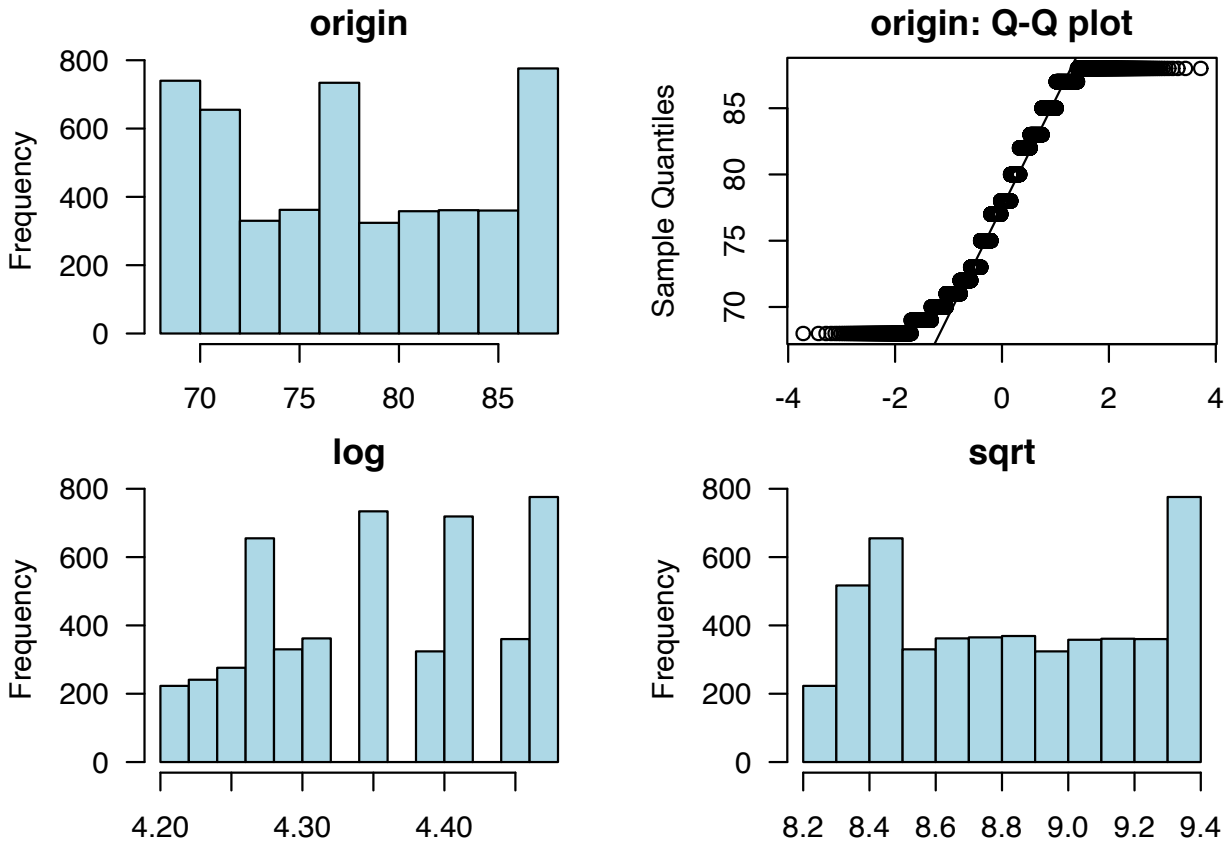

Figure 2.2: year

birth_yr

normality test : Shapiro-Wilk normality test
statistic : 0.95878, p-value : 1.824E-35

type	skewness	kurtosis
original	-0.0990	1.9730
log transformation	-0.1924	2.0198
sqrt transformation	-0.1455	1.9937

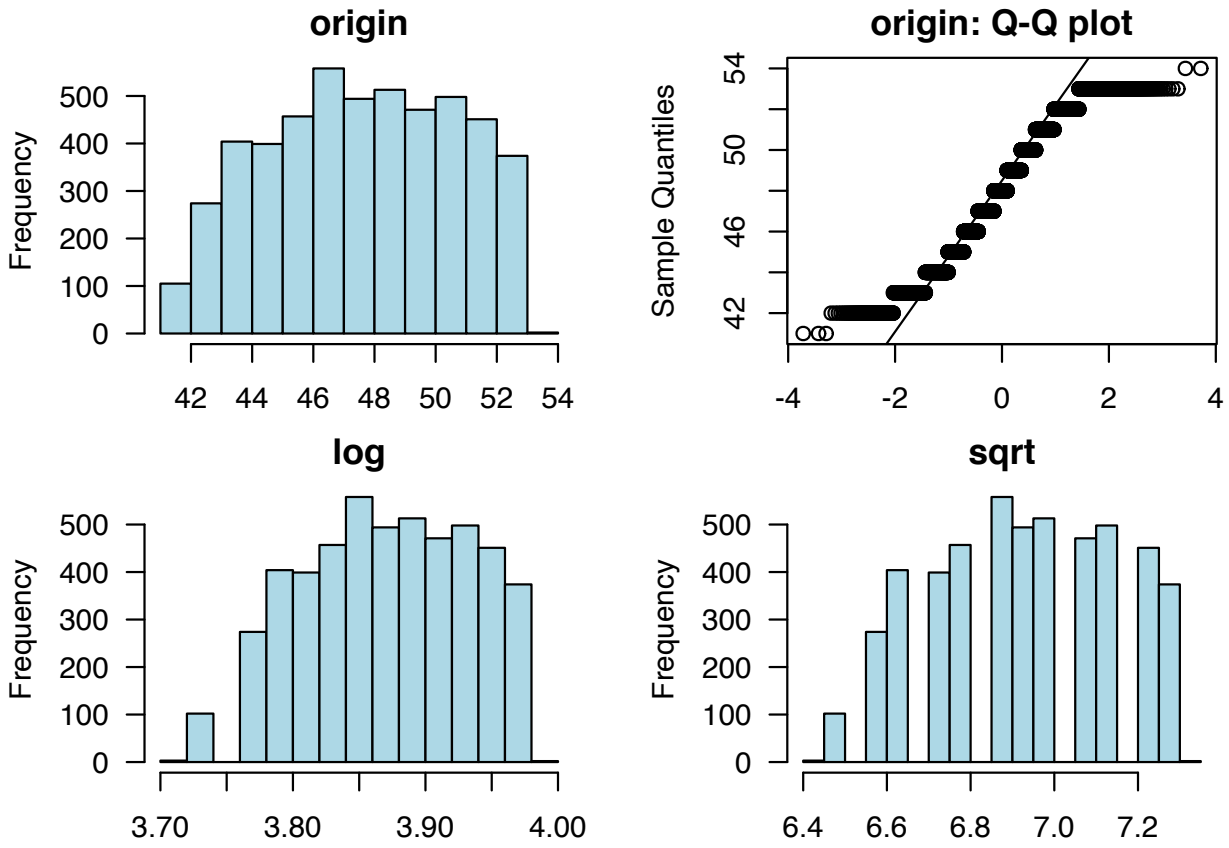

Figure 2.3: birth_yr

age

normality test : Shapiro-Wilk normality test
statistic : 0.96668, p-value : 1.64852E-32

type	skewness	kurtosis
original	0.2567	2.0620
log transformation	-0.0857	1.9930
sqrt transformation	0.0863	1.9868

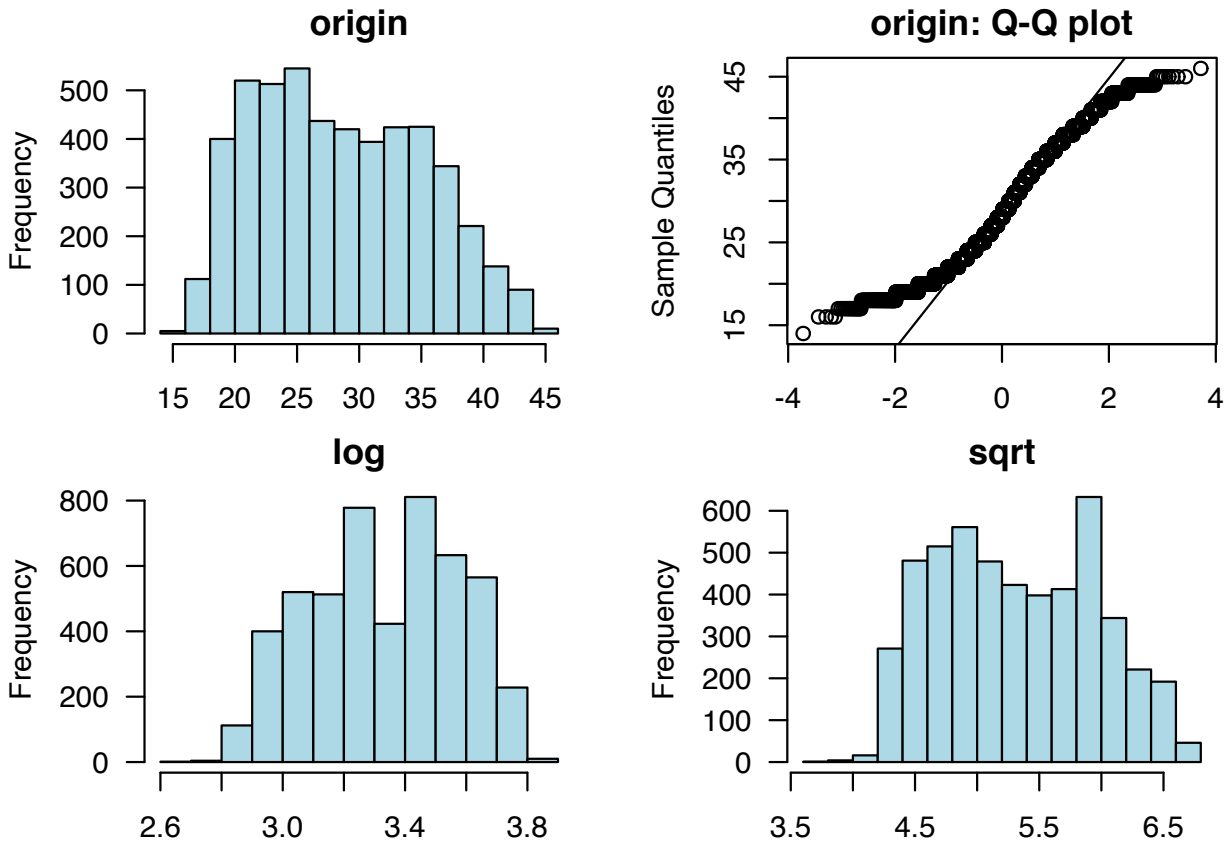

Figure 2.4: age

msp

normality test : Shapiro-Wilk normality test
statistic : 0.62594, p-value : 5.6118E-74

type	skewness	kurtosis
original	-0.3466	1.1202
log transformation		
sqrt transformation	-0.3466	1.1202

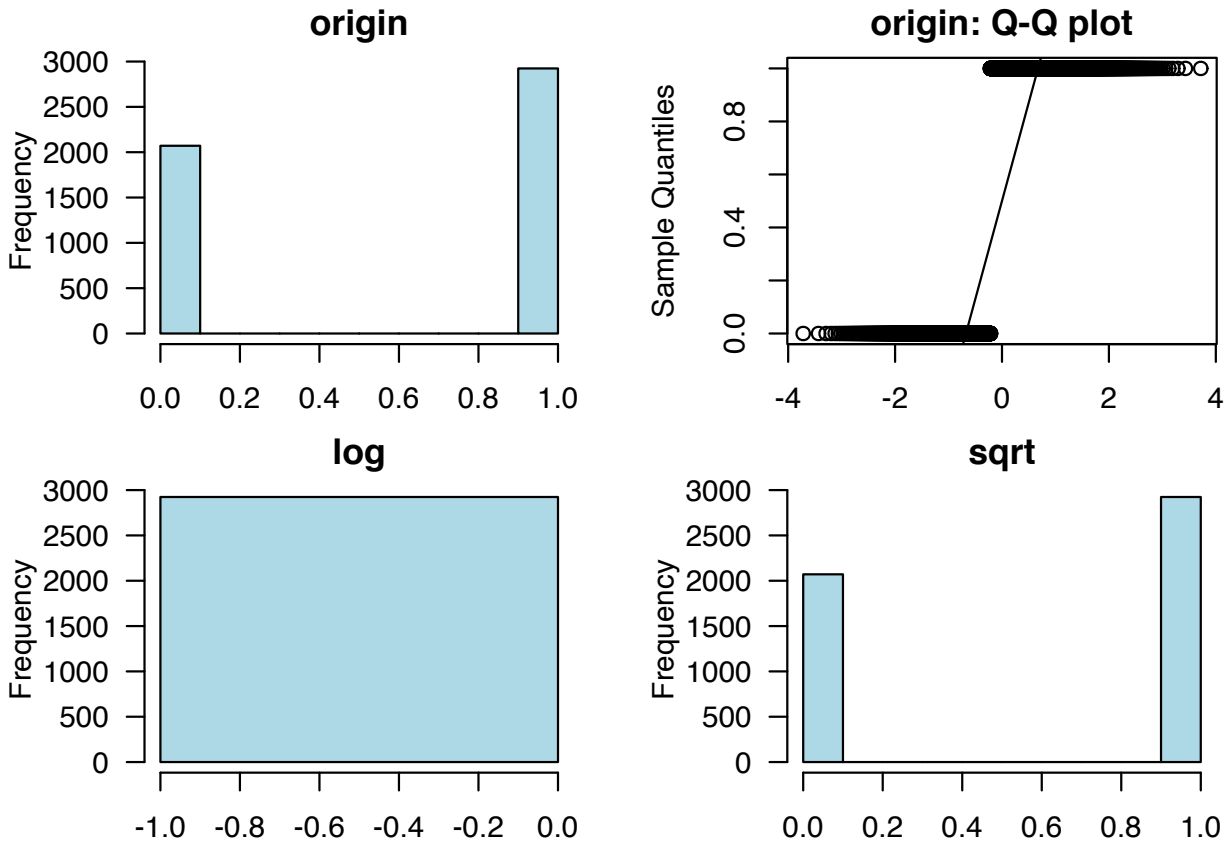

Figure 2.5: msp

nev_mar

normality test : Shapiro-Wilk normality test
statistic : 0.5186, p-value : 2.53812E-79

type	skewness	kurtosis
original	1.2974	2.6832
log transformation		
sqrt transformation	1.2974	2.6832

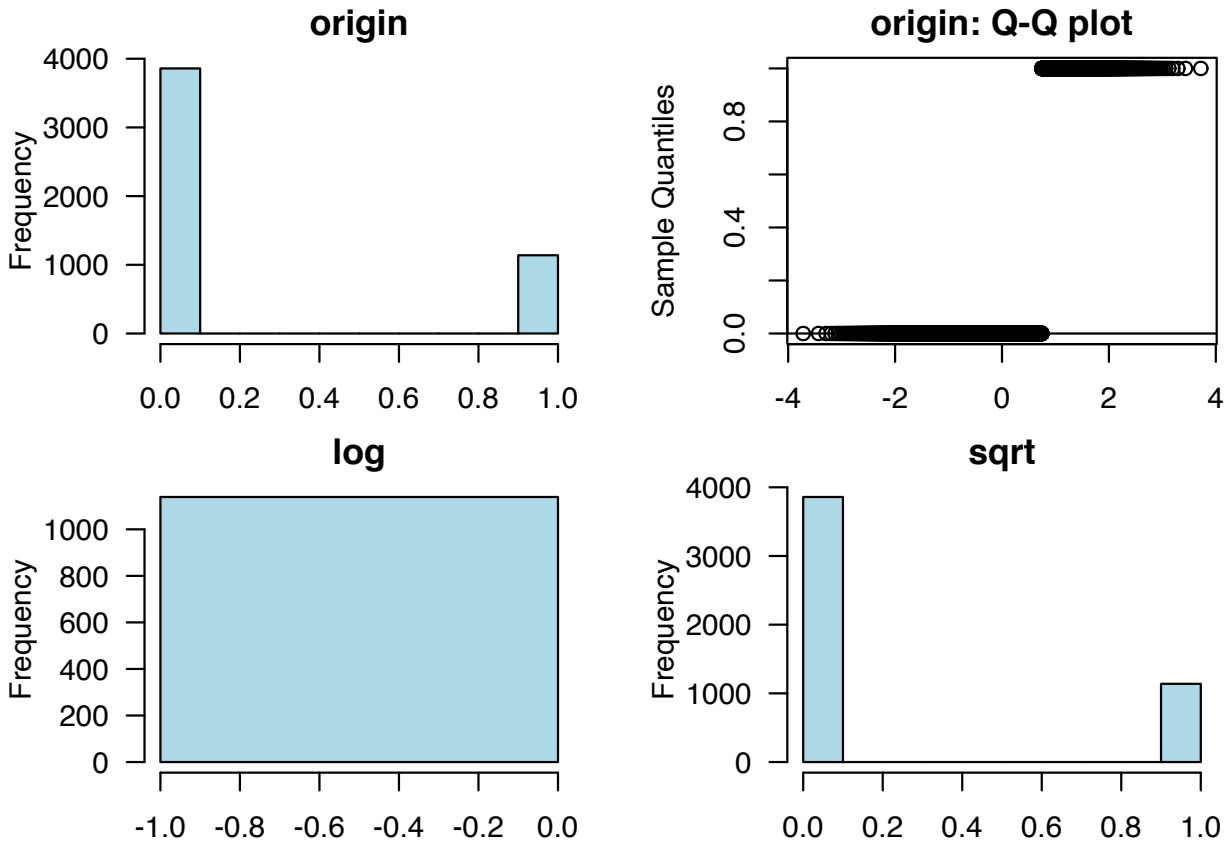

Figure 2.6: nev_mar

grade

normality test : Shapiro-Wilk normality test
statistic : 0.88853, p-value : 4.73598E-51

type	skewness	kurtosis
original	0.0846	4.4767
log transformation		
sqrt transformation	-1.2685	15.0438

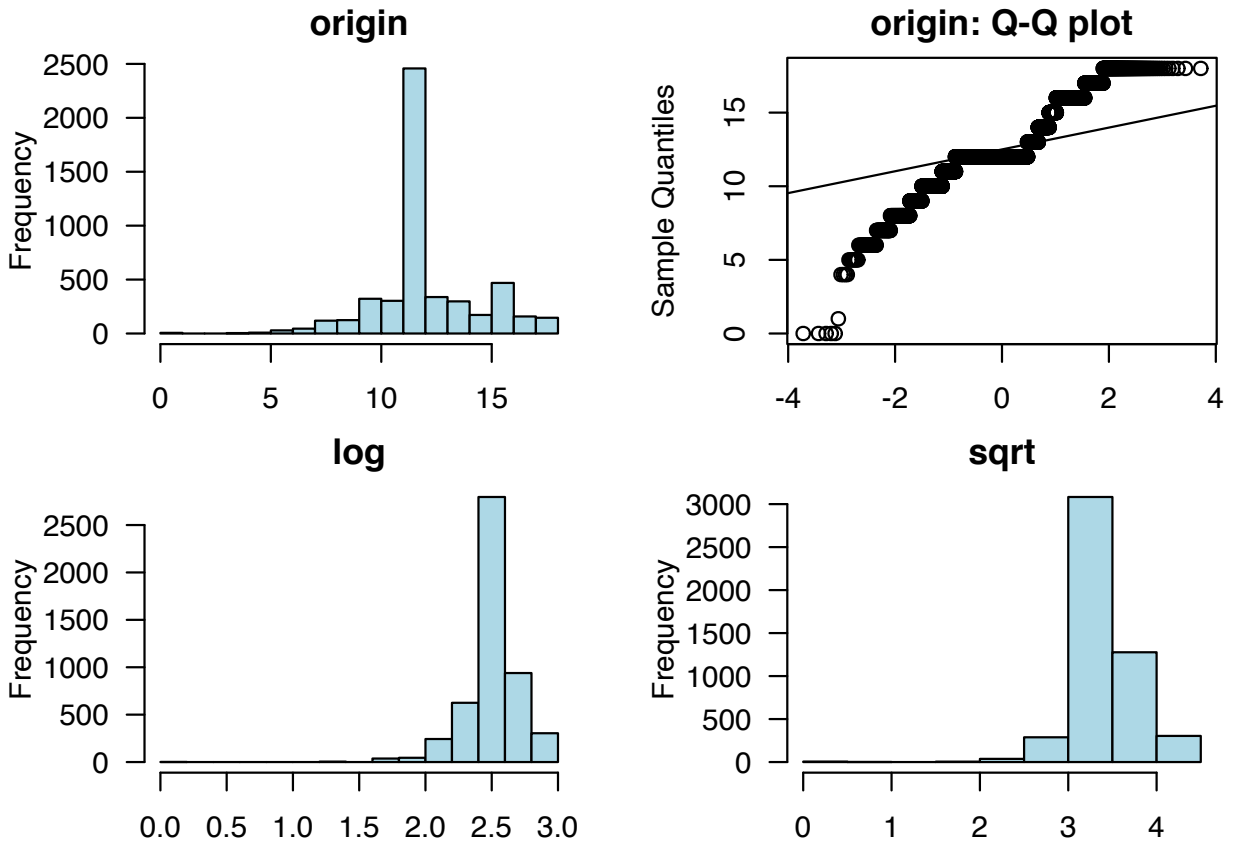

Figure 2.7: grade

collgrad

normality test : Shapiro-Wilk normality test
statistic : 0.45511, p-value : 5.09668E-82

type	skewness	kurtosis
original	1.7476	4.0542
log transformation		
sqrt transformation	1.7476	4.0542

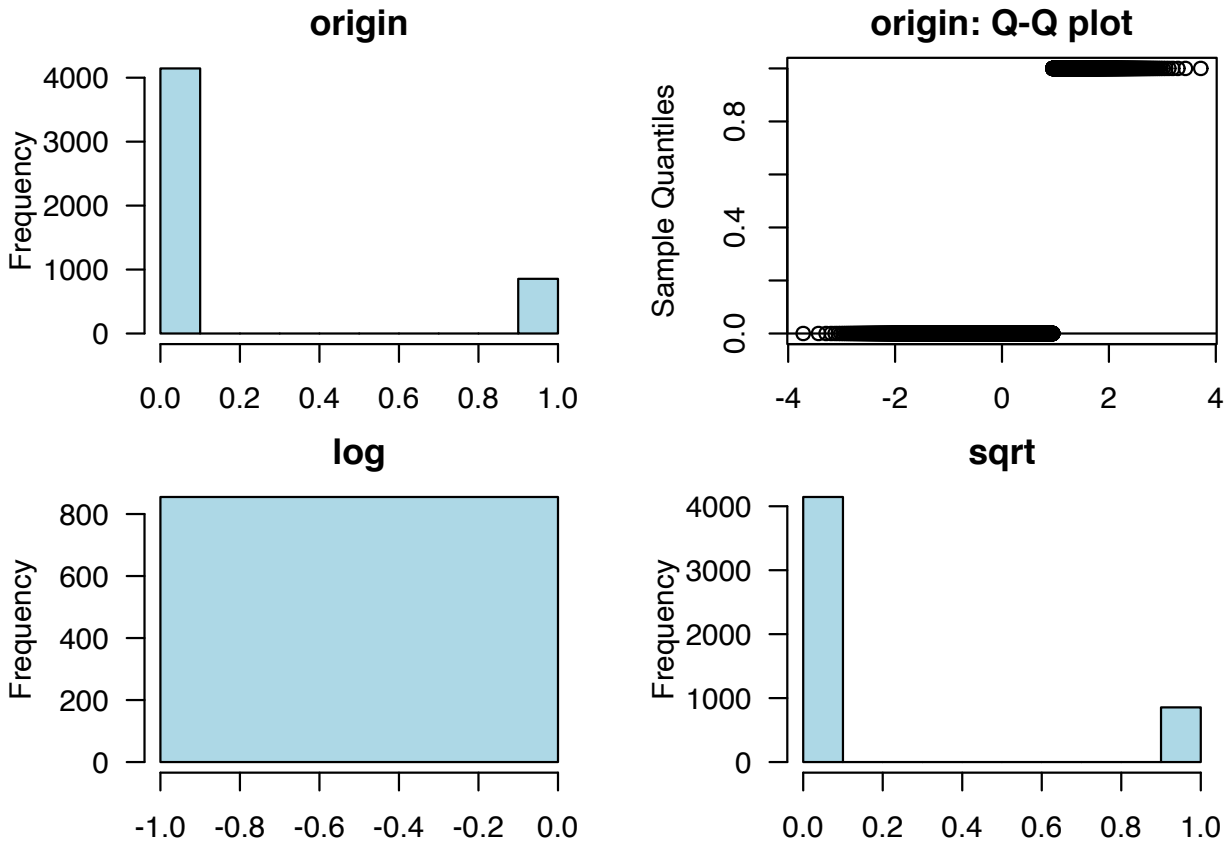

Figure 2.8: collgrad

not_smsa

normality test : Shapiro-Wilk normality test
statistic : 0.56462, p-value : 3.58861E-77

type	skewness	kurtosis
original	0.9604	1.9223
log transformation		
sqrt transformation	0.9604	1.9223

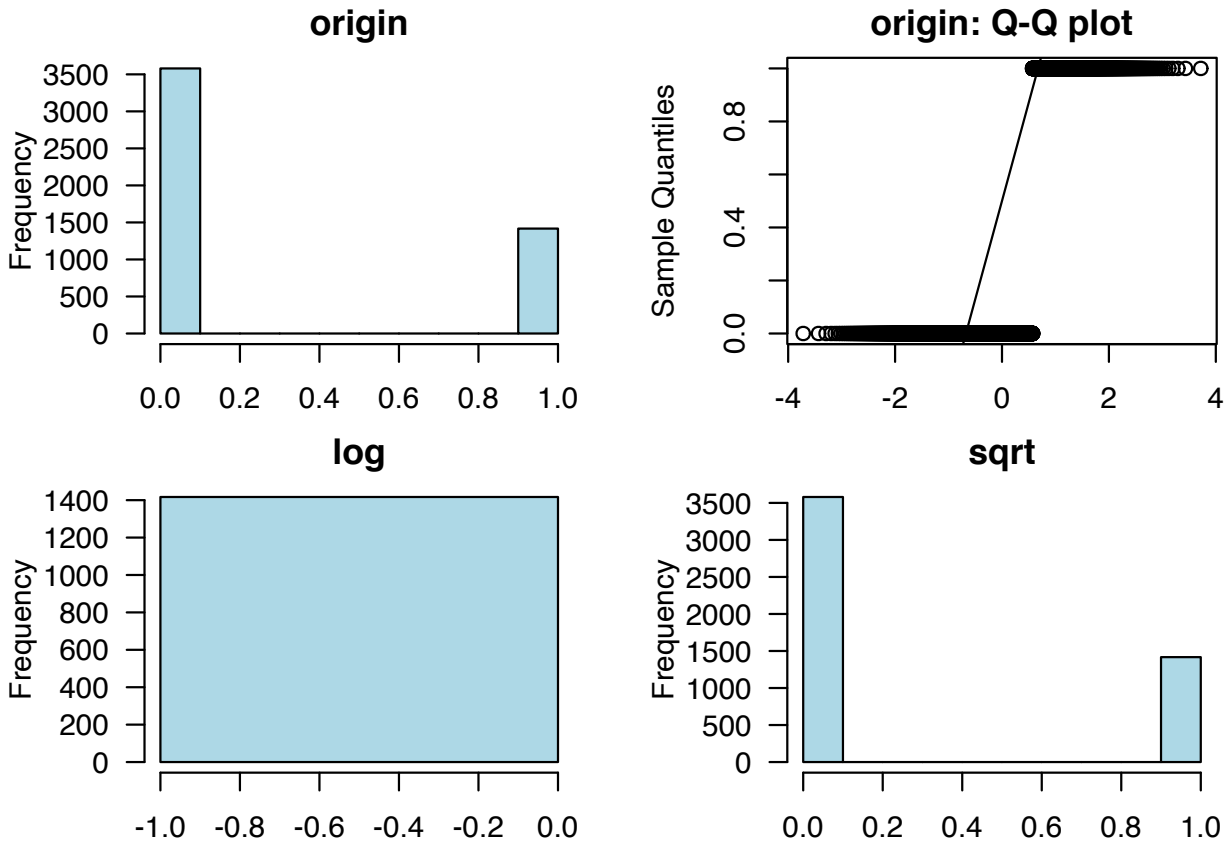

Figure 2.9: not_smsa

c_city

normality test : Shapiro-Wilk normality test
statistic : 0.60324, p-value : 3.14902E-75

type	skewness	kurtosis
original	0.6271	1.3933
log transformation		
sqrt transformation	0.6271	1.3933

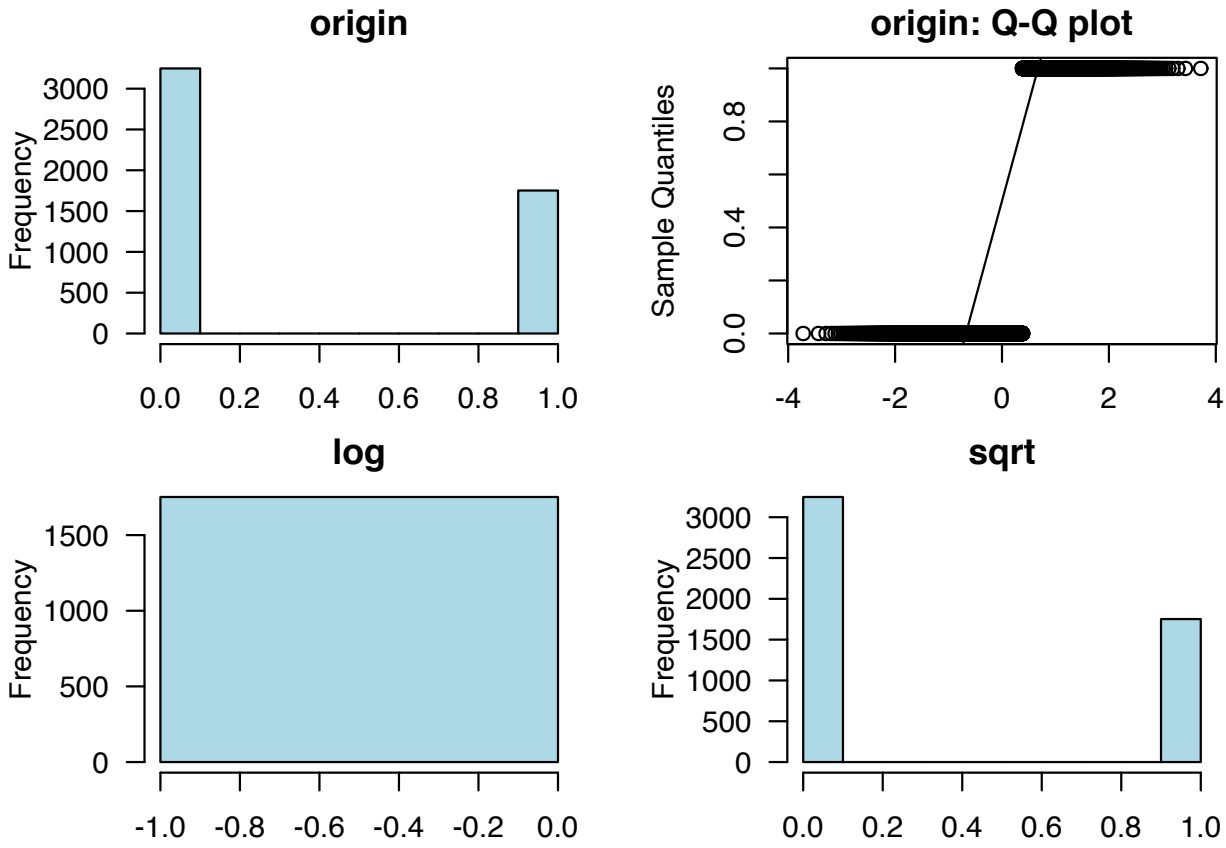

Figure 2.10: c_city

south

normality test : Shapiro-Wilk normality test
statistic : 0.62567, p-value : 5.25465E-74

type	skewness	kurtosis
original	0.3510	1.1232
log transformation		
sqrt transformation	0.3510	1.1232

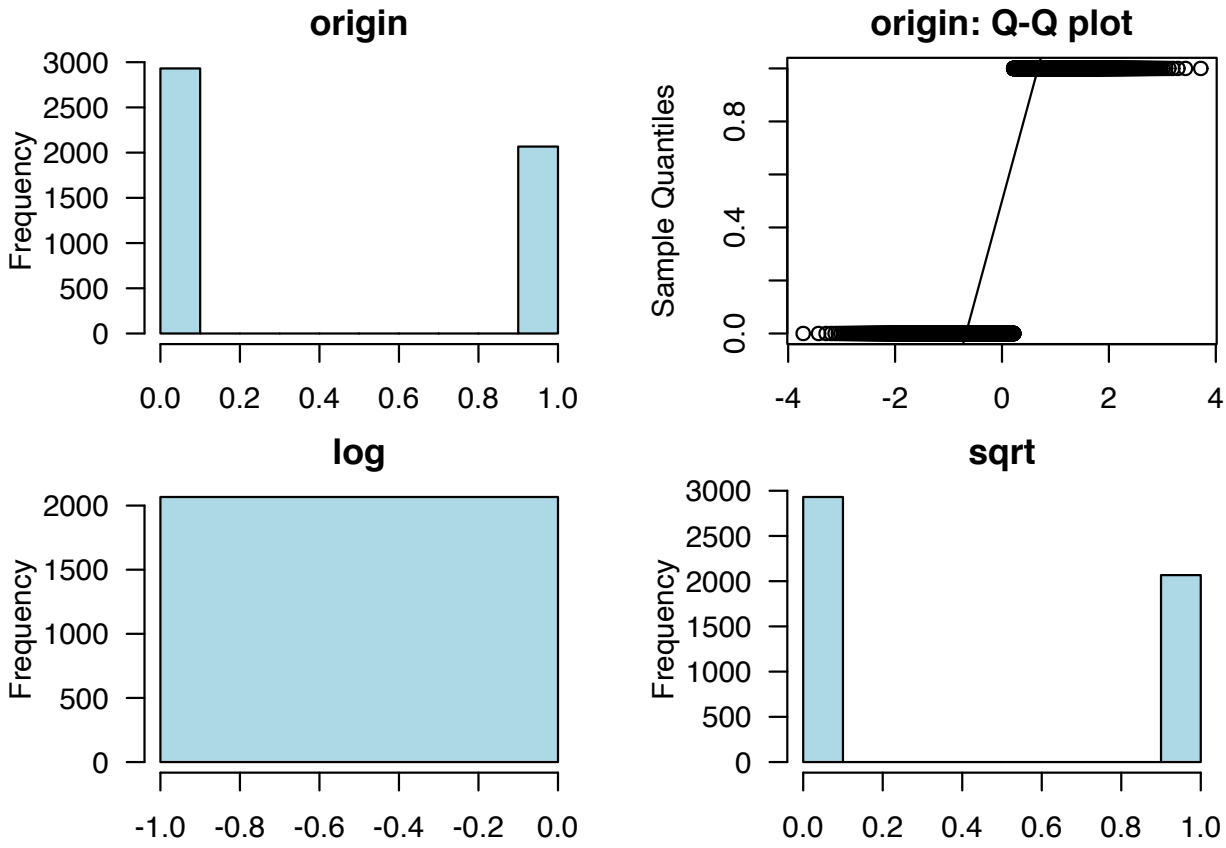

Figure 2.11: south

ind_code

normality test : Shapiro-Wilk normality test
statistic : 0.87558, p-value : 9.29691E-53

type	skewness	kurtosis
original	-0.0082	1.5780
log transformation	-0.9524	4.7572
sqrt transformation	-0.3028	2.1710

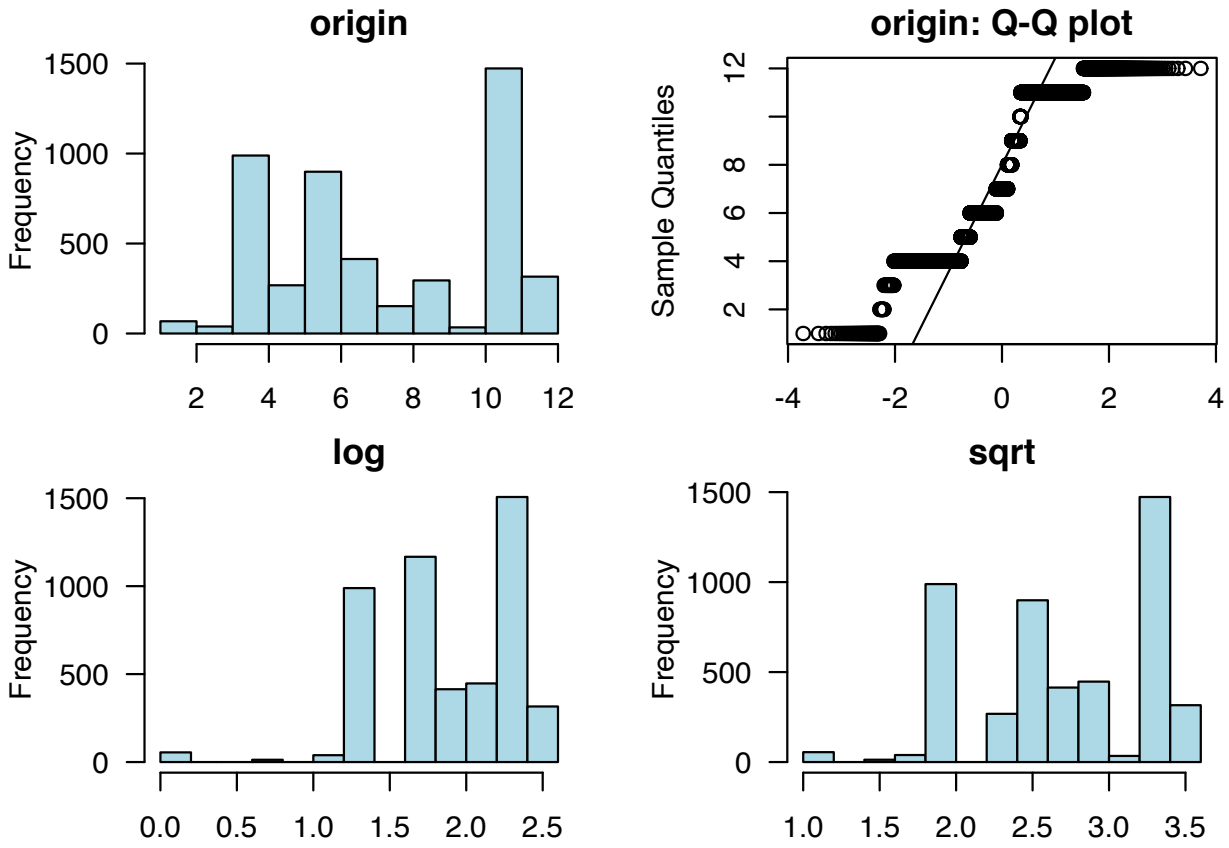

Figure 2.12: ind_code

occ_code

normality test : Shapiro-Wilk normality test
statistic : 0.85115, p-value : 4.71516E-56

type	skewness	kurtosis
original	1.0928	3.6853
log transformation	-0.2886	2.6559
sqrt transformation	0.4569	2.6328

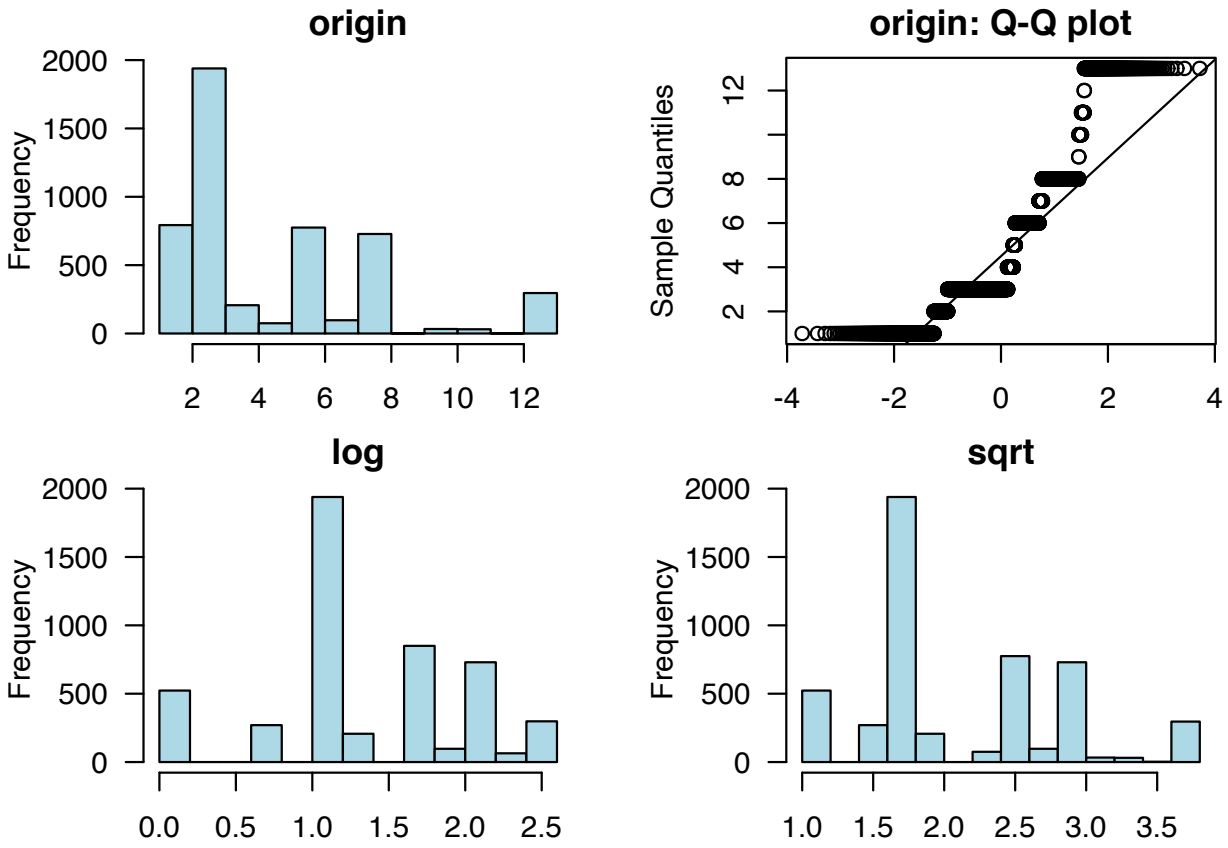

Figure 2.13: occ_code

union

normality test : Shapiro-Wilk normality test
statistic : 0.51514, p-value : 8.30218E-71

type	skewness	kurtosis
original	1.3218	2.7471
log transformation		
sqrt transformation	1.3218	2.7471

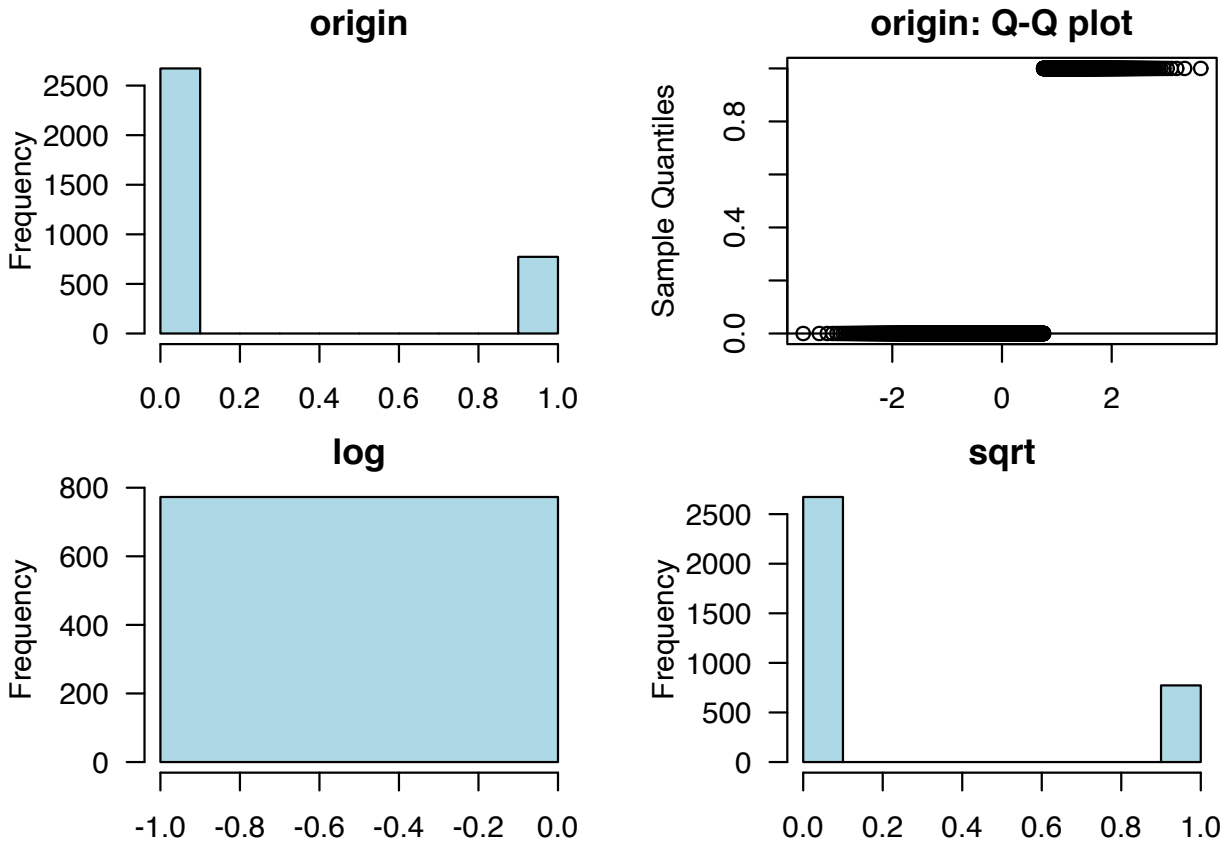

Figure 2.14: union

wks_ue

normality test : Shapiro-Wilk normality test
statistic : 0.40279, p-value : 3.96006E-78

type	skewness	kurtosis
original	3.9977	20.9285
log transformation		
sqrt transformation	2.3417	7.9947

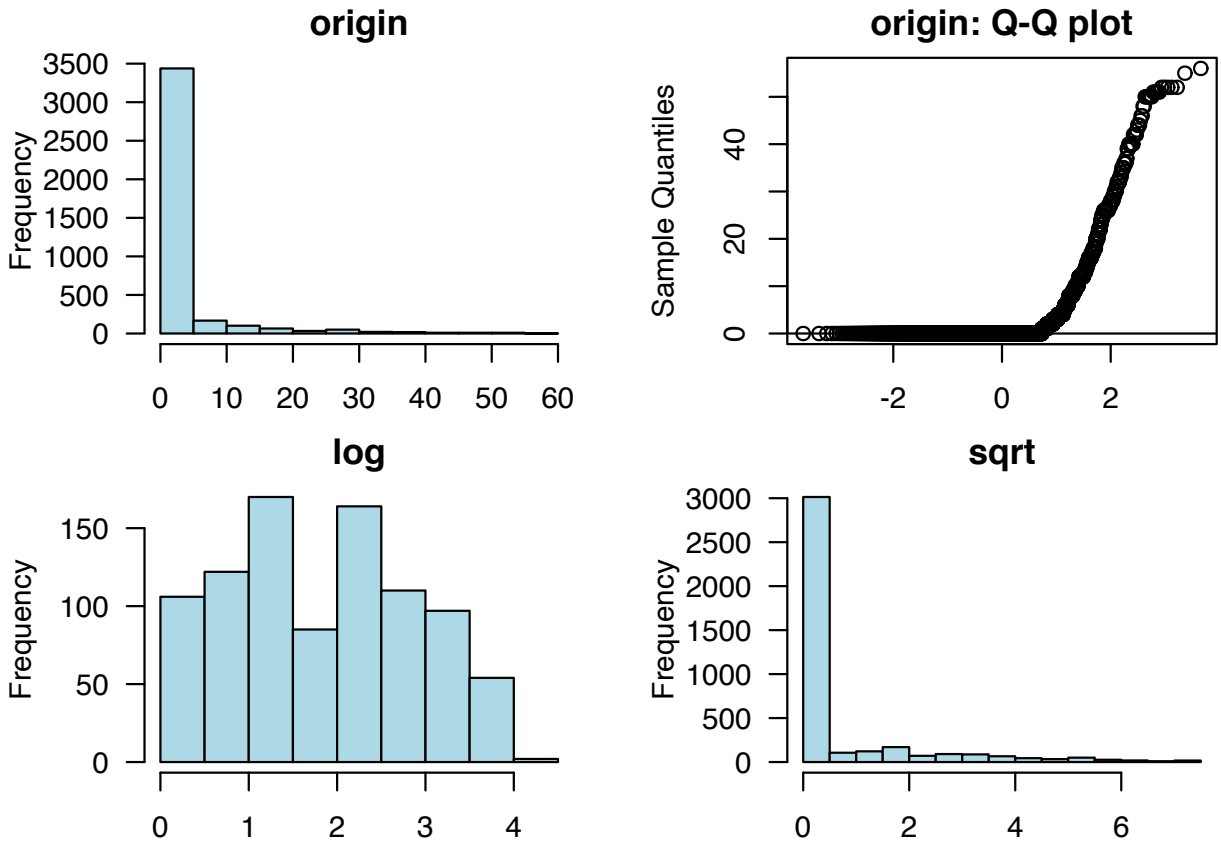

Figure 2.15: wks_ue

ttl_exp

normality test : Shapiro-Wilk normality test
statistic : 0.9249, p-value : 1.63252E-44

type	skewness	kurtosis
original	0.8823	3.1882
log transformation		
sqrt transformation	0.1367	2.3015

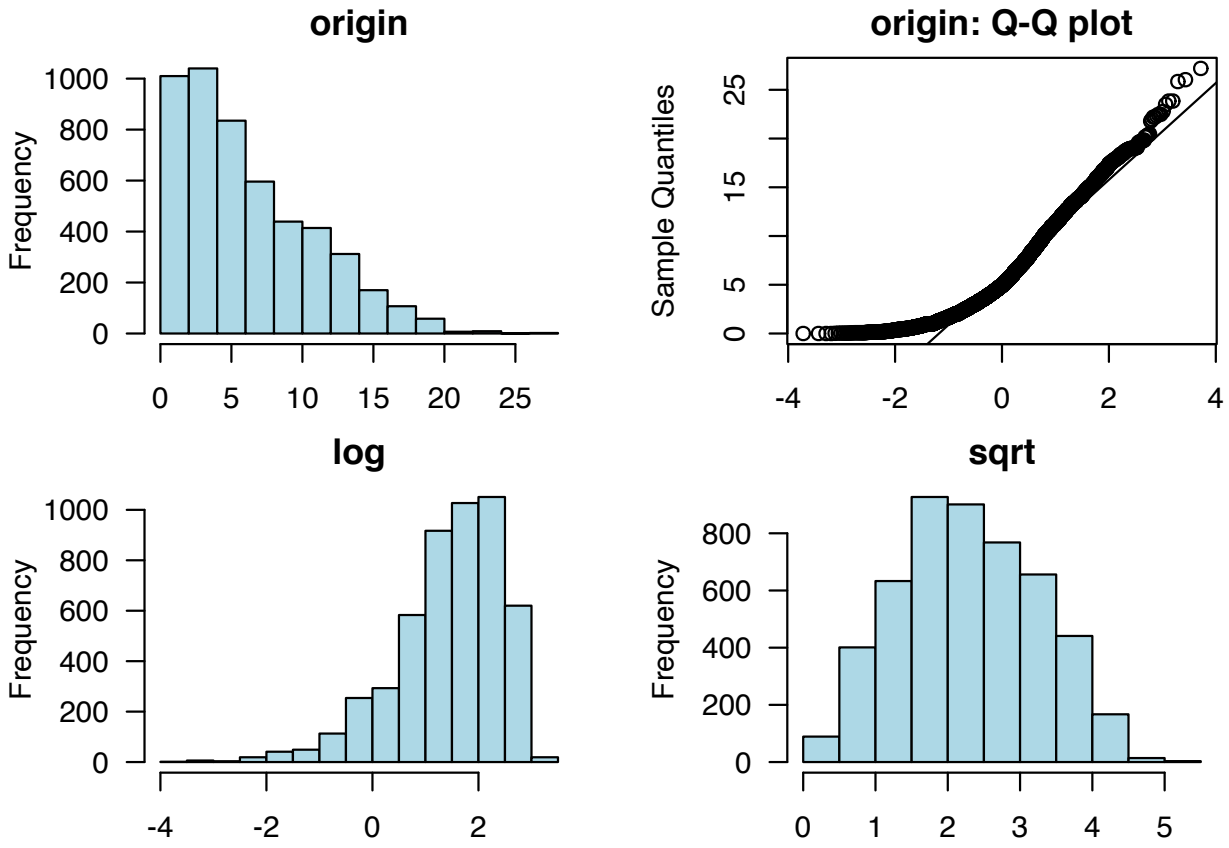

Figure 2.16: ttl_exp

tenure

normality test : Shapiro-Wilk normality test
statistic : 0.76695, p-value : 3.20981E-64

type	skewness	kurtosis
original	1.9162	6.7217
log transformation		
sqrt transformation	0.7734	3.0679

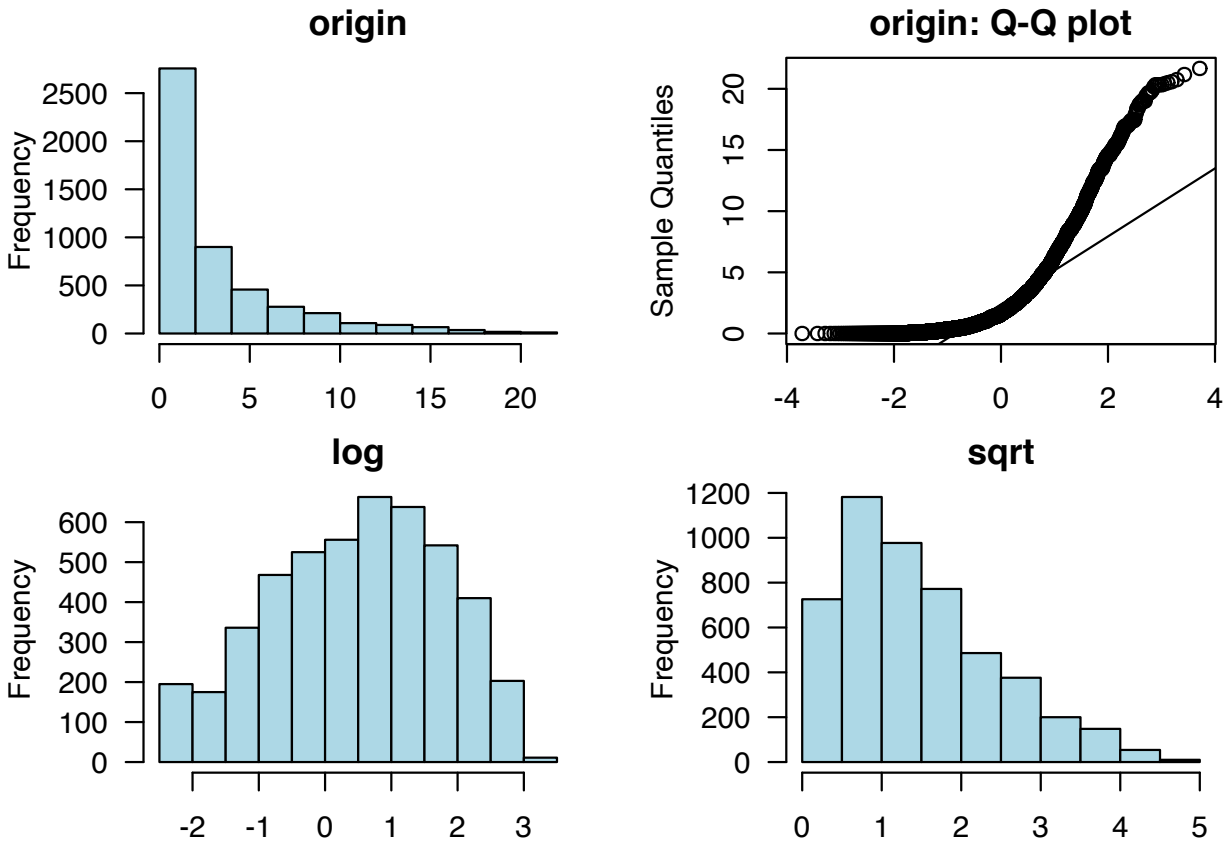

Figure 2.17: tenure

hours

normality test : Shapiro-Wilk normality test
statistic : 0.77469, p-value : 8.49081E-64

type	skewness	kurtosis
original	-0.3885	12.6432
log transformation	-3.1435	16.6391
sqrt transformation	-1.7844	8.6346

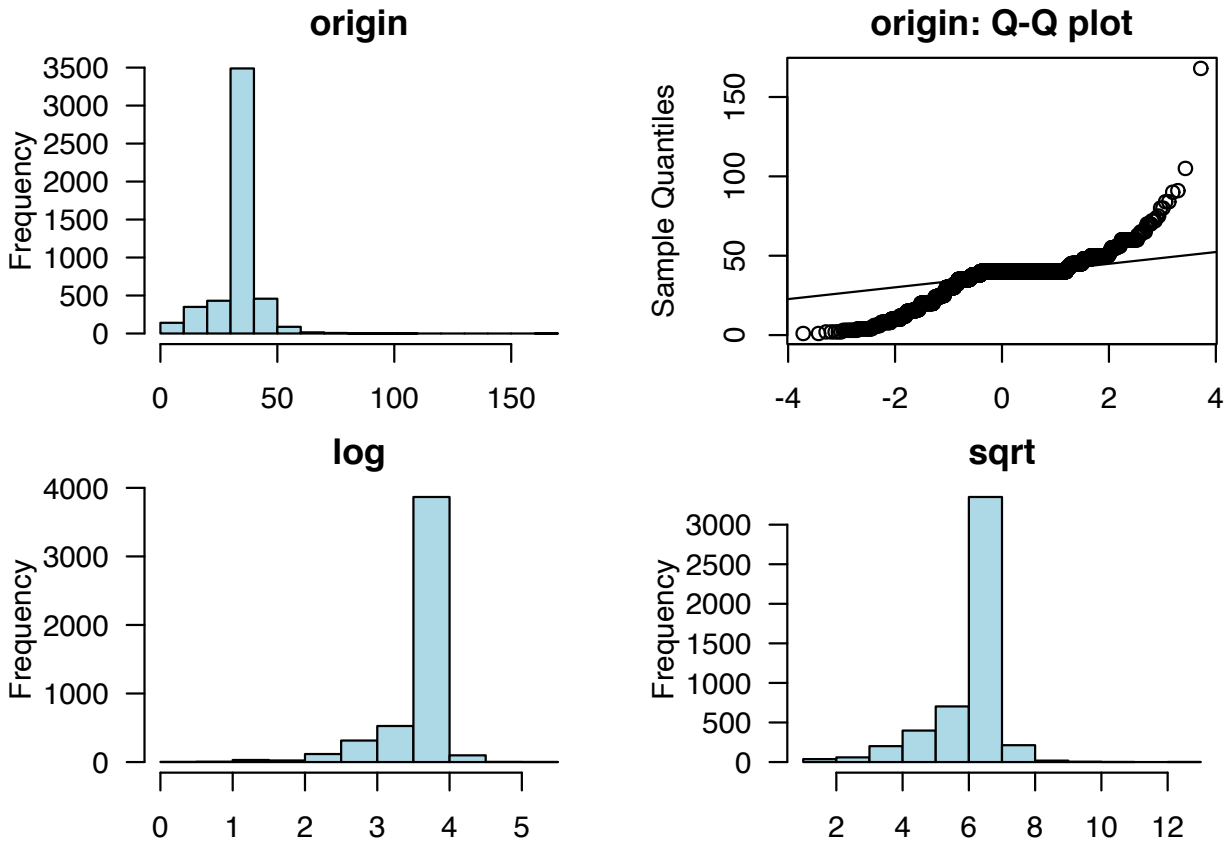

Figure 2.18: hours

wks_work

normality test : Shapiro-Wilk normality test statistic : 0.93875, p-value : 6.77921E-41

type	skewness	kurtosis
original	0.1909	2.3419
log transformation		
sqrt transformation	-0.7683	3.5957

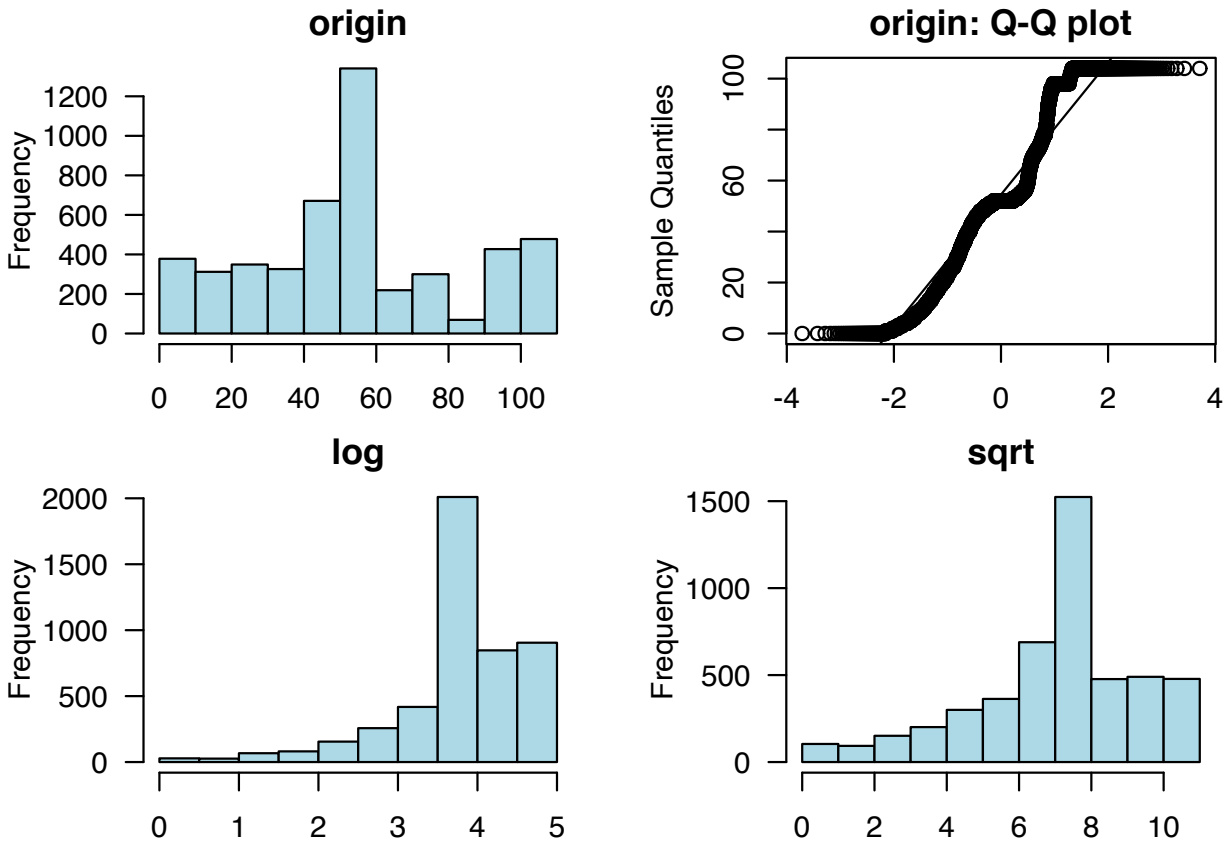

Figure 2.19: wks_work

ln_wage

normality test : Shapiro-Wilk normality test
statistic : 0.97636, p-value : 4.92624E-28

type	skewness	kurtosis
original	0.4047	4.8749
log transformation	-3.4241	29.5334
sqrt transformation	-0.6933	6.6628

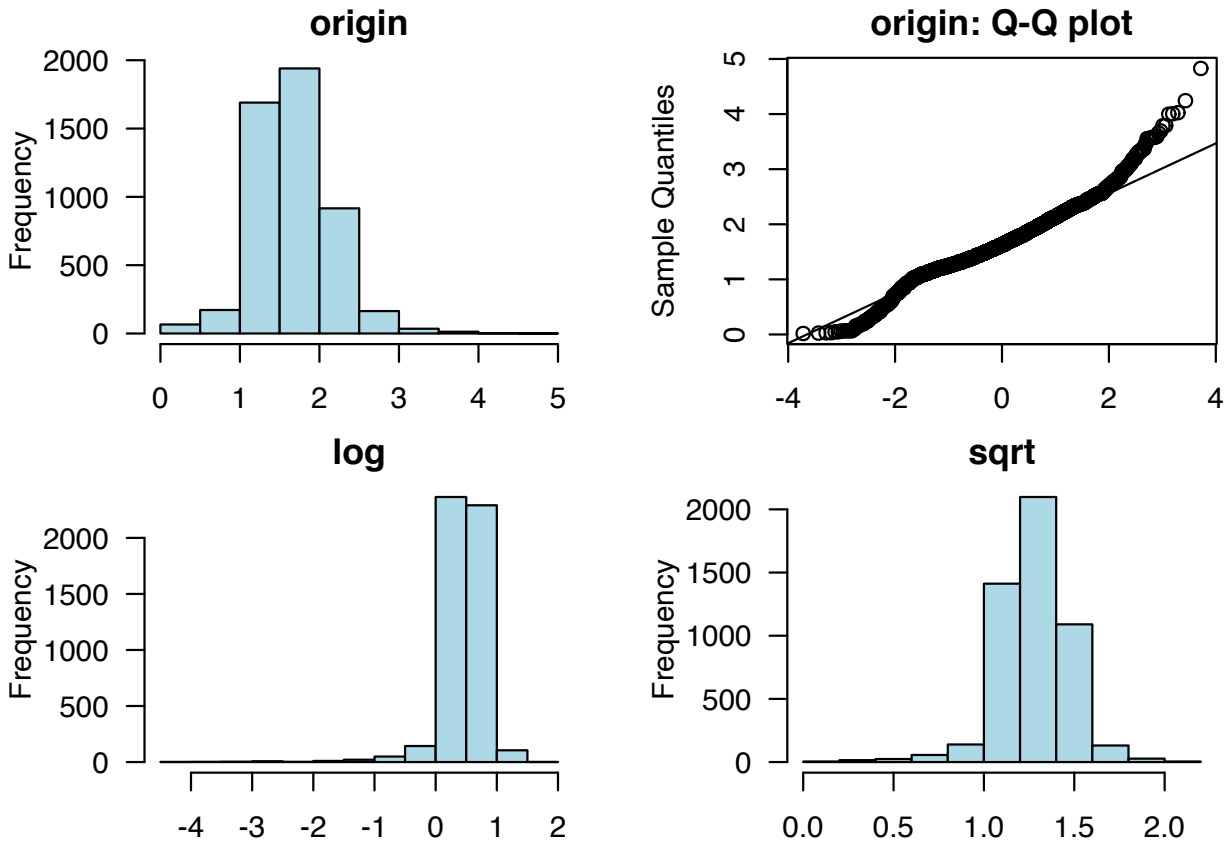

Figure 2.20: ln_wage

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

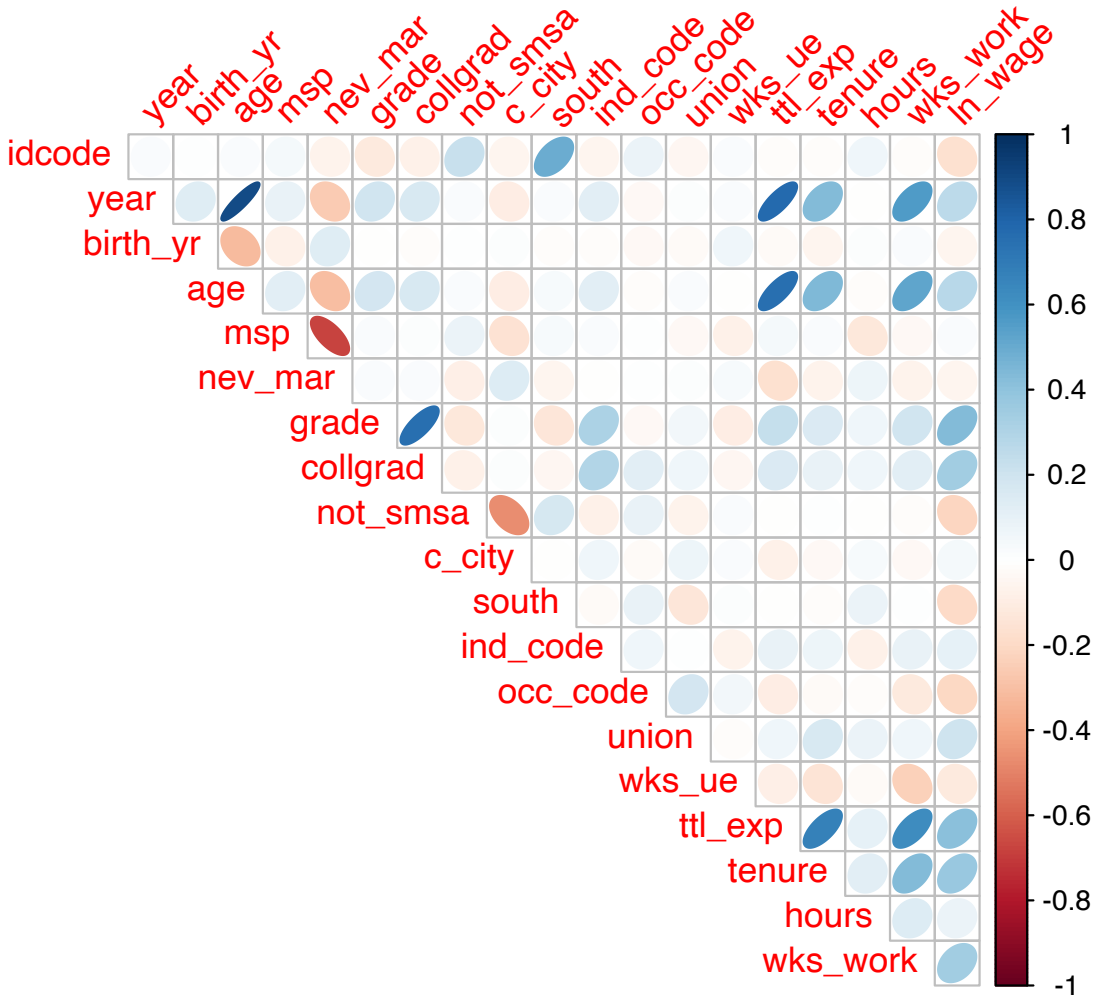

Figure 3.1: The correlation coefficient of numerical variables

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.