REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

June 28, 2021

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	0.00	4711	0.165
year	numeric	0	0.00	15	0.001
birth_yr	numeric	0	0.00	14	0.000
age	numeric	24	0.08	34	0.001
race	haven_labelled	0	0.00	3	0.000
msp	numeric	16	0.06	3	0.000
nev_mar	numeric	16	0.06	3	0.000
grade	numeric	2	0.01	20	0.001
collgrad	numeric	0	0.00	2	0.000
not_smsa	numeric	8	0.03	3	0.000
c_city	numeric	8	0.03	3	0.000
south	numeric	8	0.03	3	0.000
ind_code	numeric	341	1.20	13	0.000
occ_code	numeric	121	0.42	14	0.000
union	numeric	9296	32.58	3	0.000
wks_ue	numeric	5704	19.99	62	0.002
ttl_exp	numeric	0	0.00	4744	0.166
tenure	numeric	433	1.52	271	0.009
hours	numeric	67	0.23	86	0.003
wks_work	numeric	703	2.46	106	0.004
ln_wage	numeric	0	0.00	8173	0.286

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 : 0.95324, p-value : 2.80638E-37

Table 2.1: skewness and kurtosis : idcode

type	skewness	kurtosis
original	-0.0027	1.7873
log transformation	-2.2843	10.9307
sqrt transformation	-0.5748	2.4369

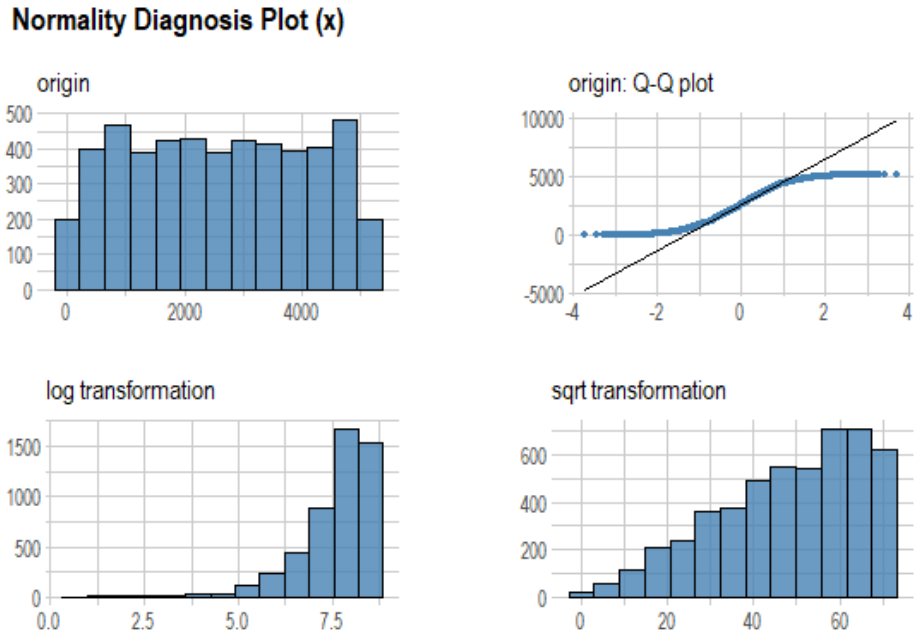

Figure 2.1: idcode

year

* normality test : Shapiro-Wilk normality test
- statistic : 0.9291, p-value : 1.33857E-43

Table 2.2: skewness and kurtosis : year

type	skewness	kurtosis
original	0.0647	1.6765
log transformation	-0.0187	1.6748
sqrt transformation	0.0230	1.6738

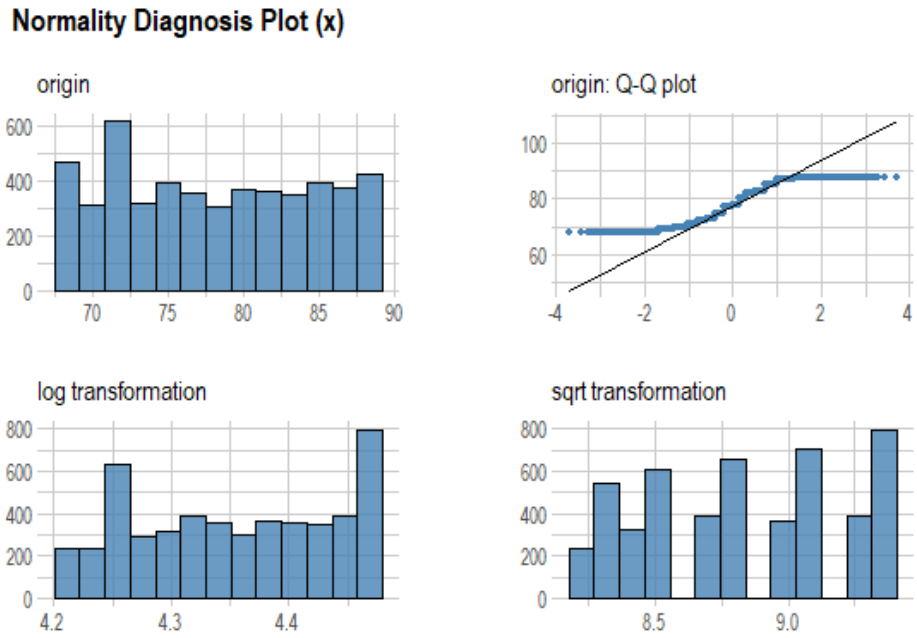

Figure 2.2: year

birth_yr

* normality test : Shapiro-Wilk normality test
- statistic : 0.96135, p-value : 1.4725E-34

Table 2.3: skewness and kurtosis : birth_yr

type	skewness	kurtosis
original	-0.1230	2.0250
log transformation	-0.2190	2.0759
sqrt transformation	-0.1708	2.0477

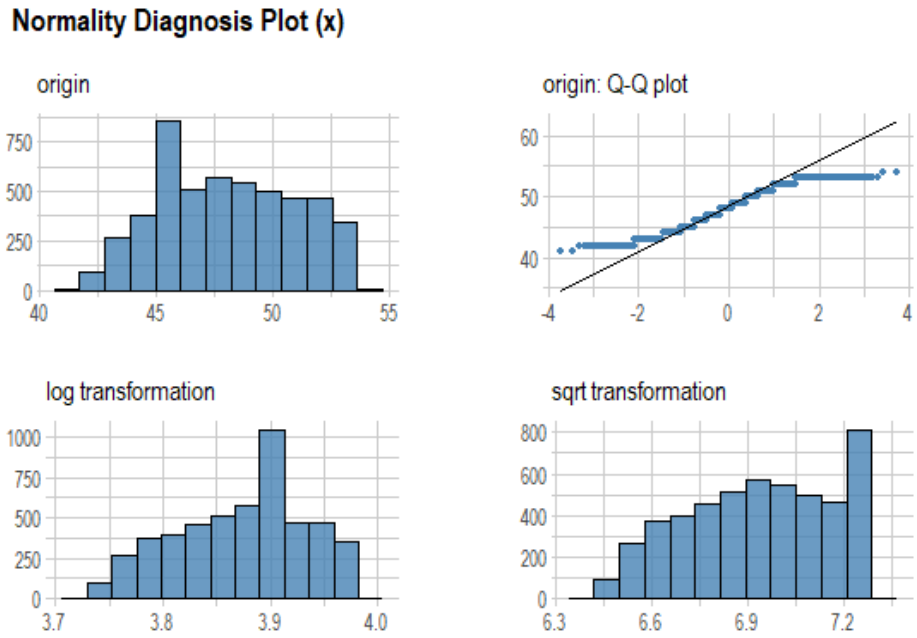

Figure 2.3: birth_yr

age

* normality test : Shapiro-Wilk normality test
- statistic : 0.96648, p-value : 1.34979E-32

Table 2.4: skewness and kurtosis : age

type	skewness	kurtosis
original	0.2904	2.1000
log transformation	-0.0627	2.0184
sqrt transformation	0.1151	2.0171

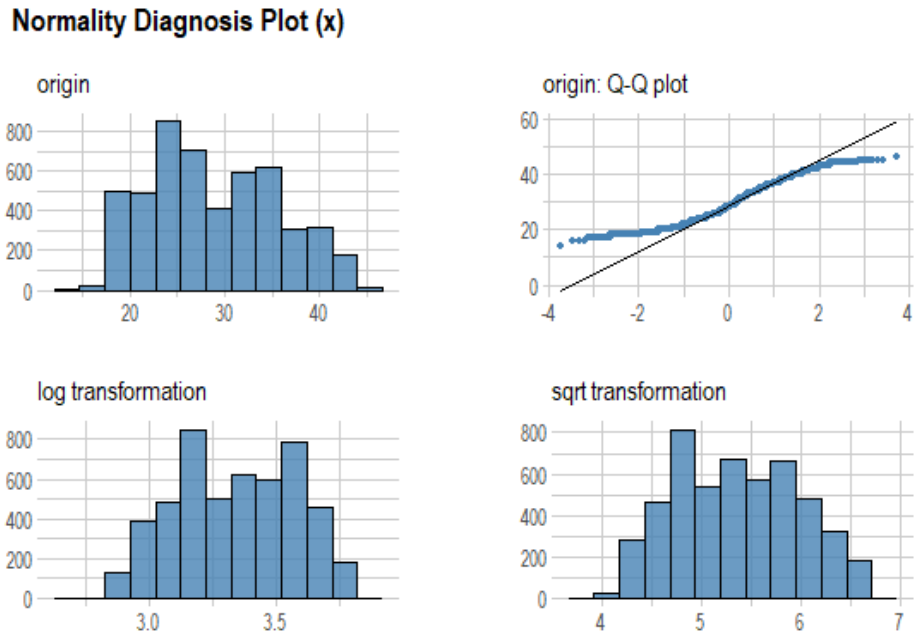

Figure 2.4: age

msp

* normality test : Shapiro-Wilk normality test
- statistic : 0.6242, p-value : 4.2656E-74

Table 2.5: skewness and kurtosis : msp

type	skewness	kurtosis
original	-0.3744	1.1402
log+1 transformation	-0.3744	1.1402
sqrt transformation	-0.3744	1.1402

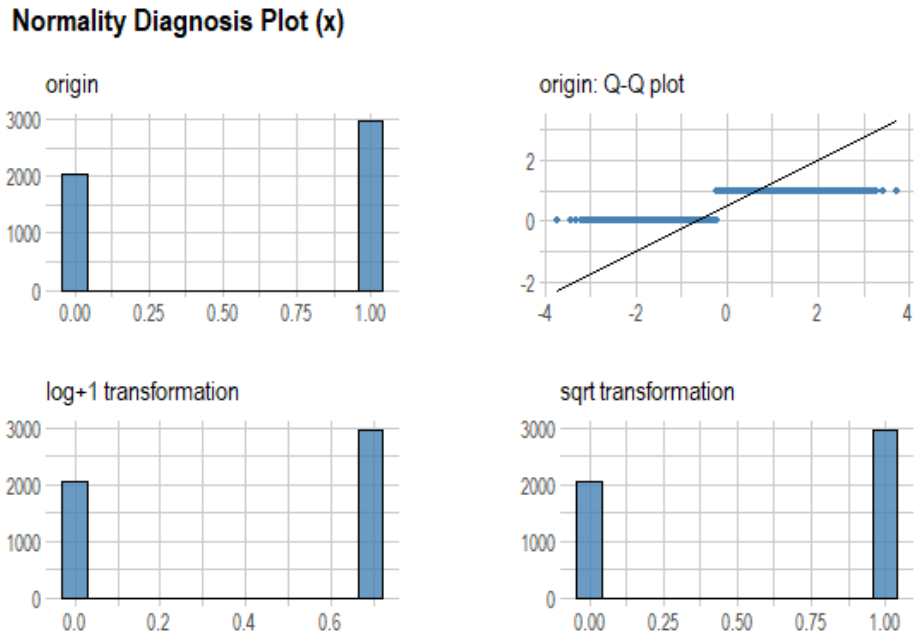

Figure 2.5: msp

nev_mar

* normality test : Shapiro-Wilk normality test
- statistic : 0.51871, p-value : 2.50949E-79

Table 2.6: skewness and kurtosis : nev_mar

type	skewness	kurtosis
original	1.2967	2.6813
log+1 transformation	1.2967	2.6813
sqrt transformation	1.2967	2.6813

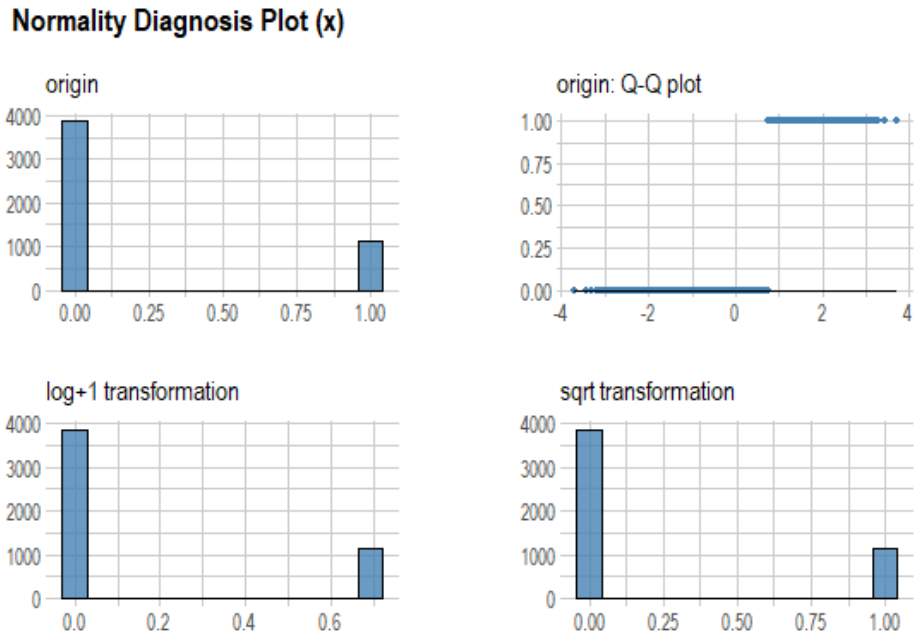

Figure 2.6: nev_mar

grade

* normality test : Shapiro-Wilk normality test
- statistic : 0.88574, p-value : 1.76218E-51

Table 2.7: skewness and kurtosis : grade

type	skewness	kurtosis
original	0.1296	4.1300
log+1 transformation	-1.7158	17.9144
sqrt transformation	-0.6608	8.0986

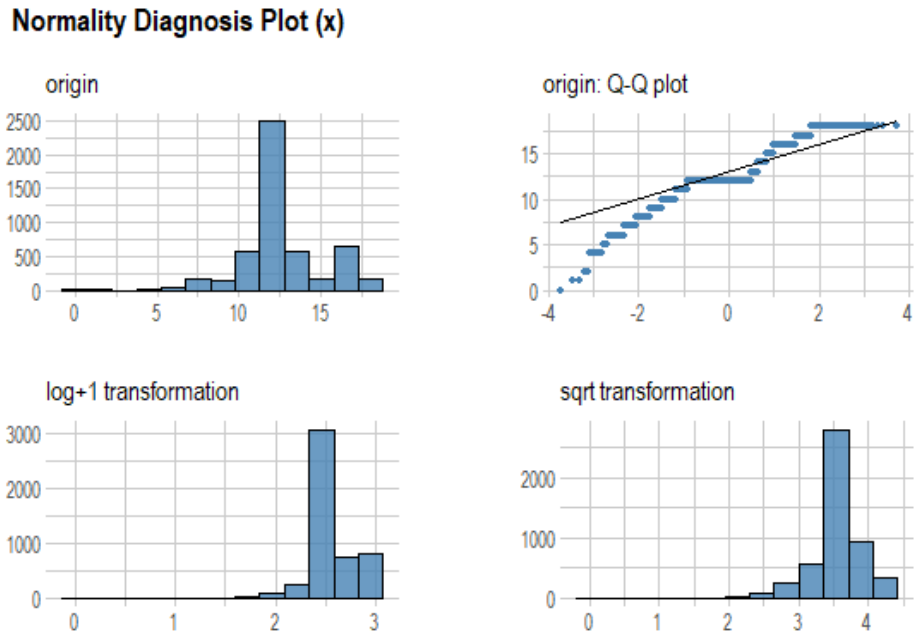

Figure 2.7: grade

collgrad

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

Table 2.8: skewness and kurtosis : collgrad

type	skewness	kurtosis
original	1.7476	4.0542
log+1 transformation	1.7476	4.0542
sqrt transformation	1.7476	4.0542

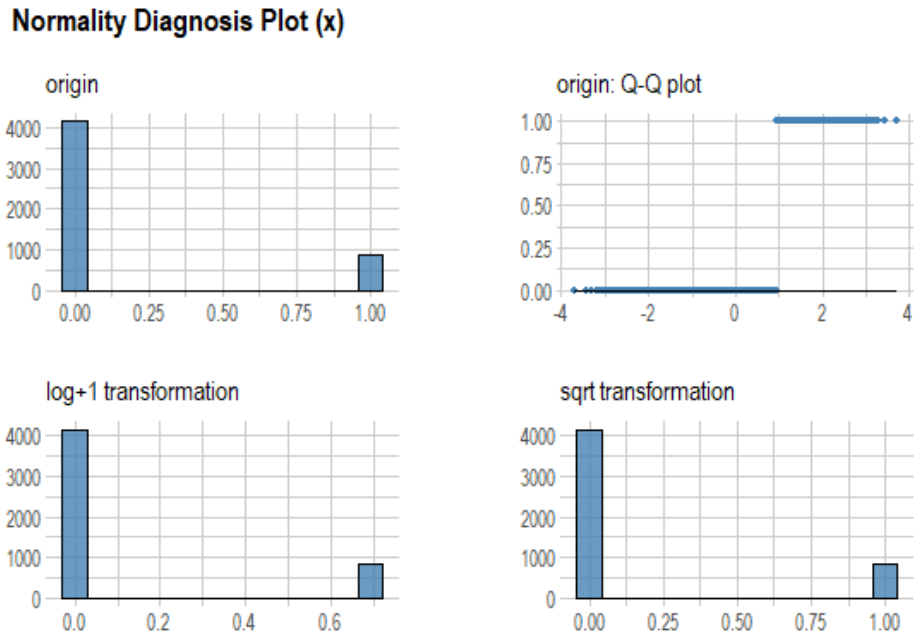

Figure 2.8: collgrad

not_smsa

* normality test : Shapiro-Wilk normality test
- statistic : 0.55798, p-value : 1.65759E-77

Table 2.9: skewness and kurtosis : not_smsa

type	skewness	kurtosis
original	1.0112	2.0225
log+1 transformation	1.0112	2.0225
sqrt transformation	1.0112	2.0225

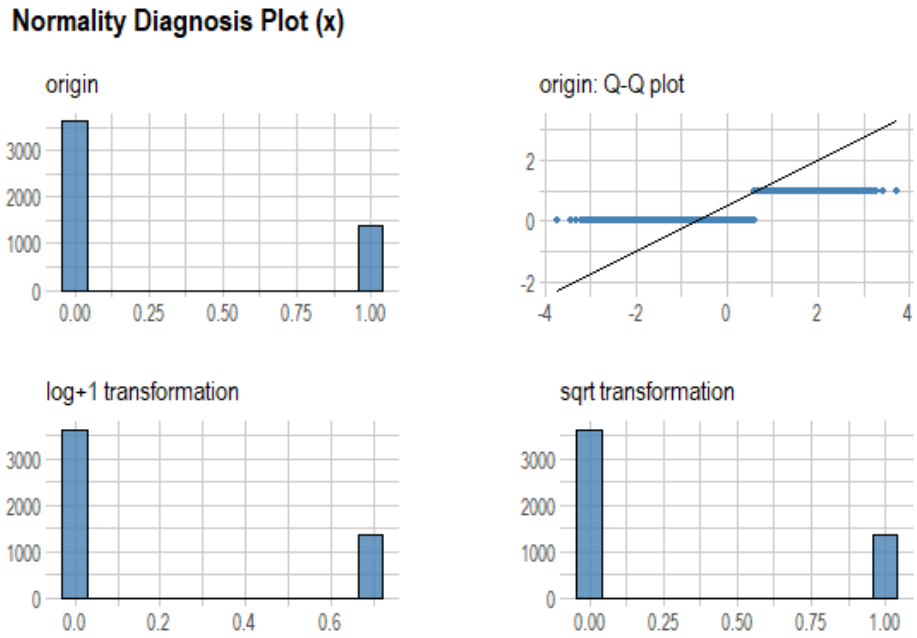

Figure 2.9: not_smsa

c_city

* normality test : Shapiro-Wilk normality test
- statistic : 0.60213, p-value : 2.75242E-75

Table 2.10: skewness and kurtosis : c_city

type	skewness	kurtosis
original	0.6382	1.4073
log+1 transformation	0.6382	1.4073
sqrt transformation	0.6382	1.4073

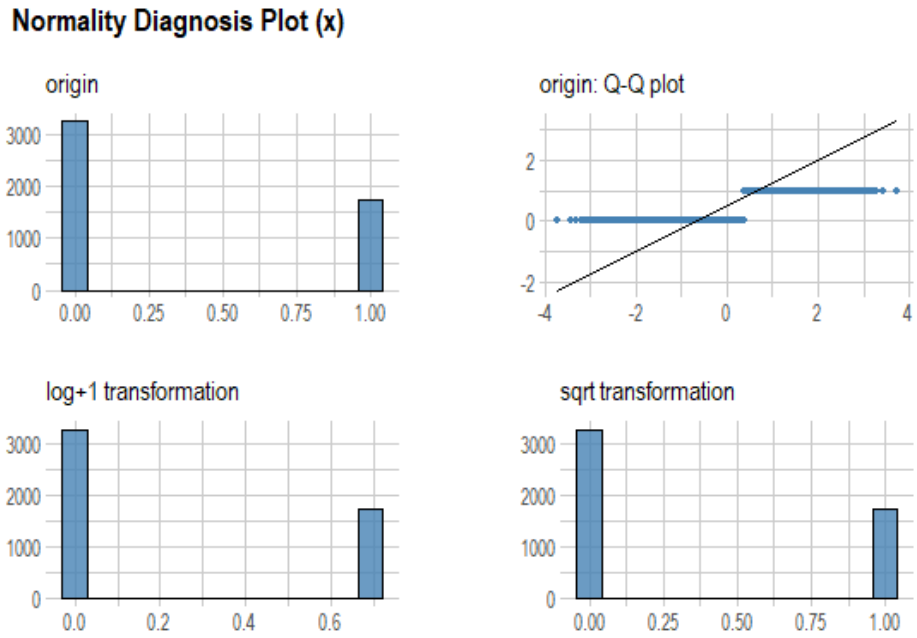

Figure 2.10: c_city

south

* normality test : Shapiro-Wilk normality test
- statistic : 0.62398, p-value : 4.14832E-74

Table 2.11: skewness and kurtosis : south

type	skewness	kurtosis
original	0.3778	1.1427
log+1 transformation	0.3778	1.1427
sqrt transformation	0.3778	1.1427

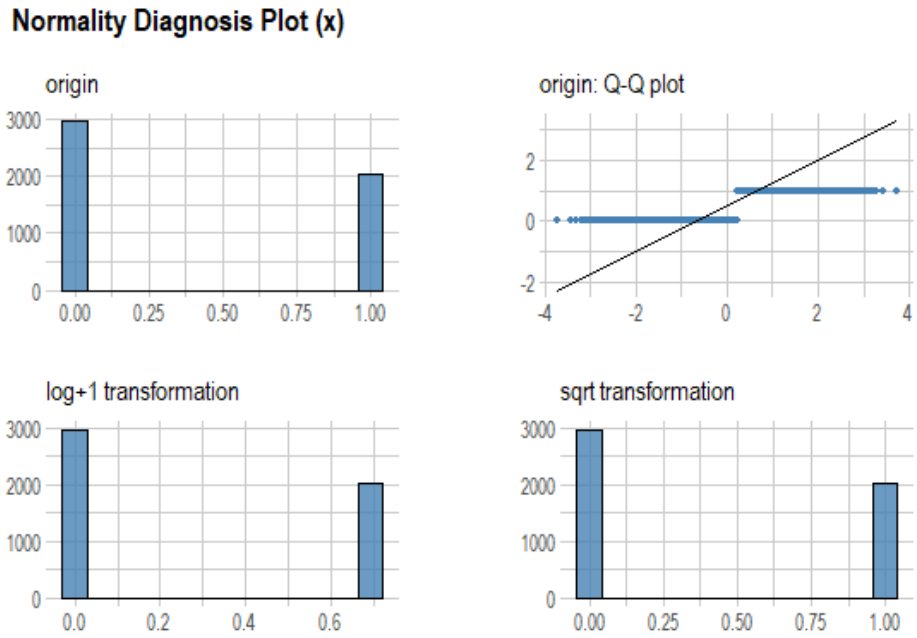

Figure 2.11: south

ind_code

* normality test : Shapiro-Wilk normality test
- statistic : 0.8675, p-value : 4.60024E-54

Table 2.12: skewness and kurtosis : ind_code

type	skewness	kurtosis
original	-0.0030	1.5222
log transformation	-0.7754	4.0086
sqrt transformation	-0.2495	1.9683

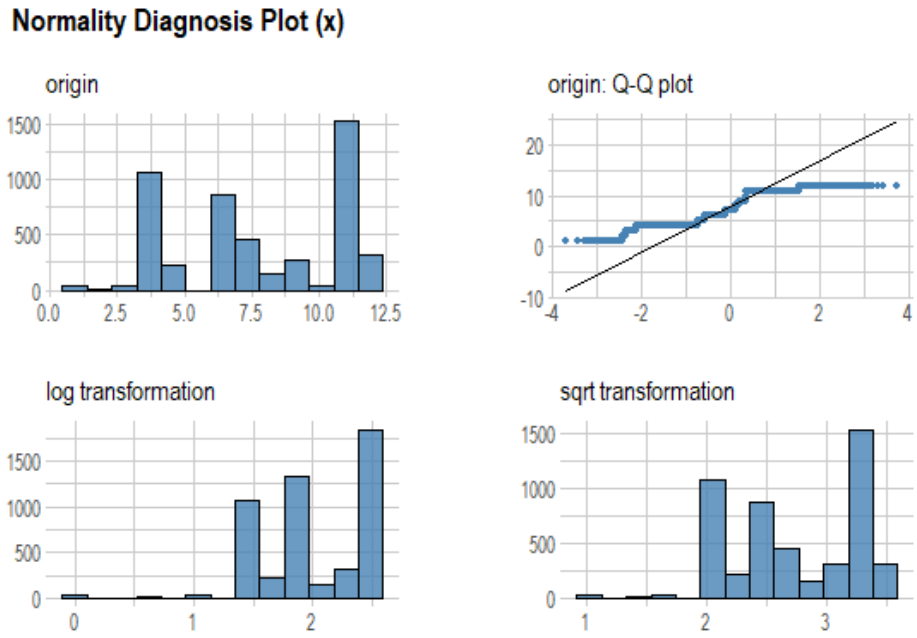

Figure 2.12: ind_code

occ_code

* normality test : Shapiro-Wilk normality test
- statistic : 0.86041, p-value : 5.47663E-55

Table 2.13: skewness and kurtosis : occ_code

type	skewness	kurtosis
original	1.0594	3.7031
log transformation	-0.3386	2.7126
sqrt transformation	0.4093	2.6344

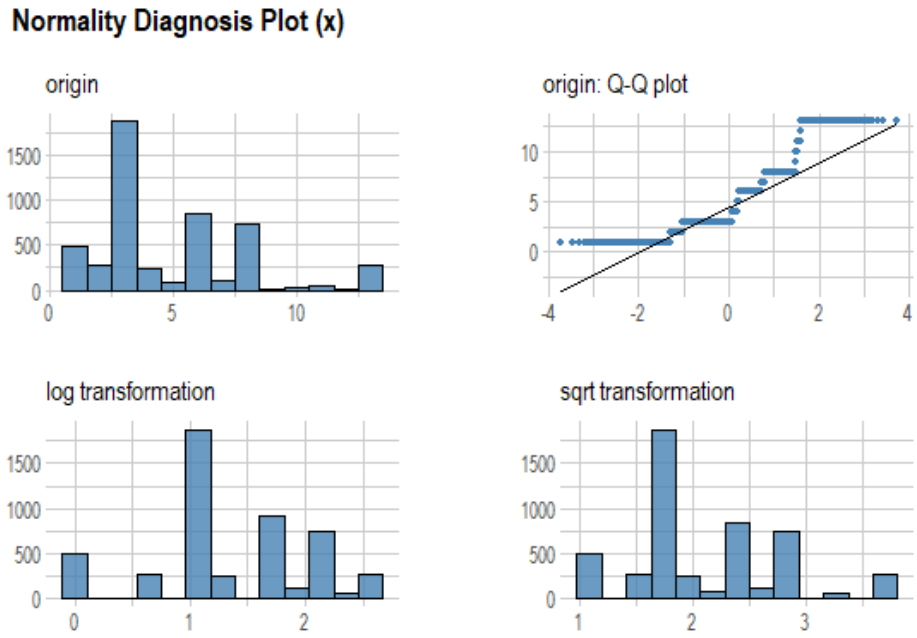

Figure 2.13: occ_code

union

* normality test : Shapiro-Wilk normality test
- statistic : 0.52212, p-value : 3.56822E-79

Table 2.14: skewness and kurtosis : union

type	skewness	kurtosis
original	1.2725	2.6192
log+1 transformation	1.2725	2.6192
sqrt transformation	1.2725	2.6192

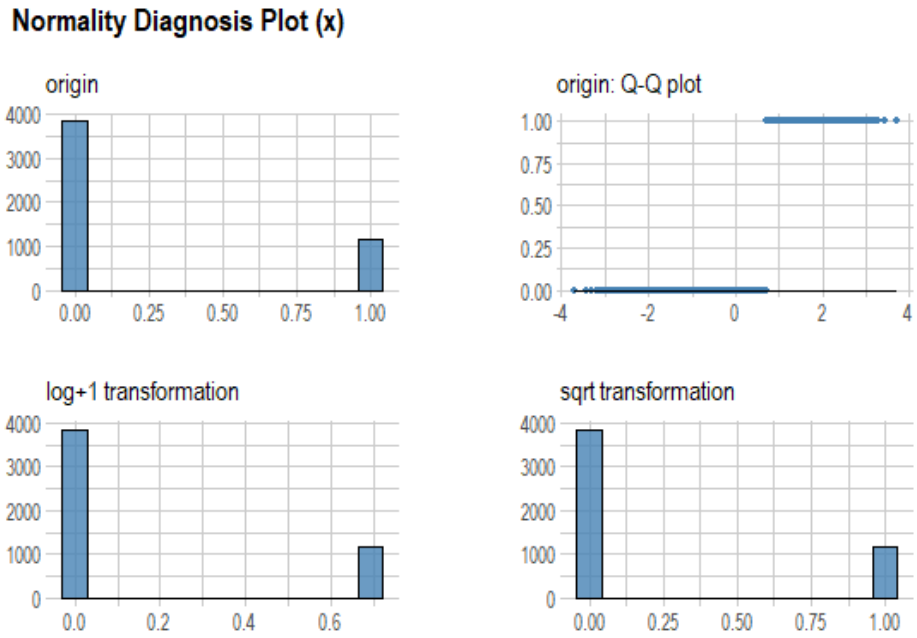

Figure 2.14: union

wks_ue

* normality test : Shapiro-Wilk normality test
- statistic : 0.40393, p-value : 5.3878E-84

Table 2.15: skewness and kurtosis : wks_ue

type	skewness	kurtosis
original	4.0445	21.5400
log+1 transformation	1.9007	5.4332
sqrt transformation	2.3213	7.9828

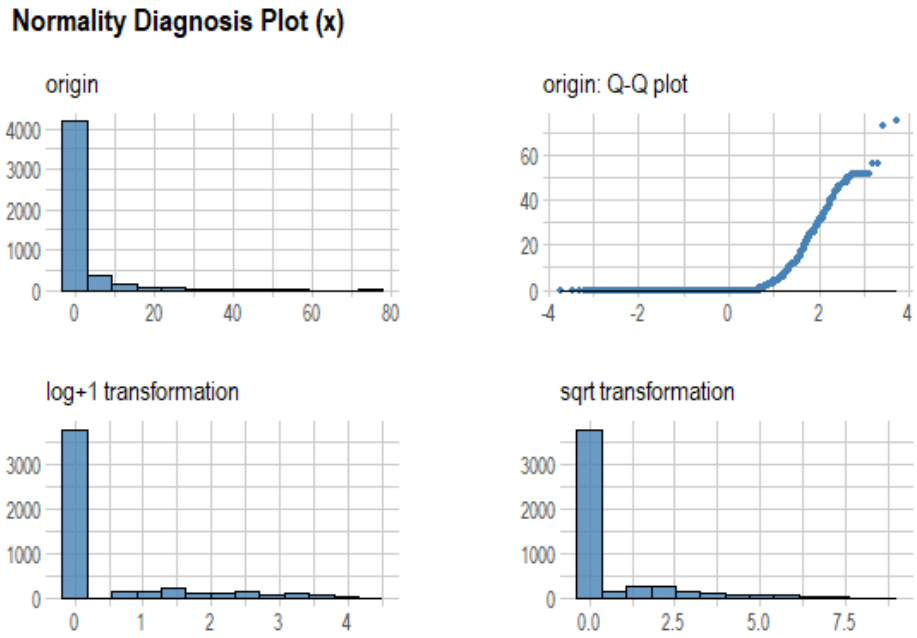

Figure 2.15: wks_ue

ttl_exp

* normality test : Shapiro-Wilk normality test
- statistic : 0.92646, p-value : 3.52492E-44

Table 2.16: skewness and kurtosis : ttl_exp

type	skewness	kurtosis
original	0.8432	2.9967
log+1 transformation	-0.3150	2.2917
sqrt transformation	0.1119	2.2643

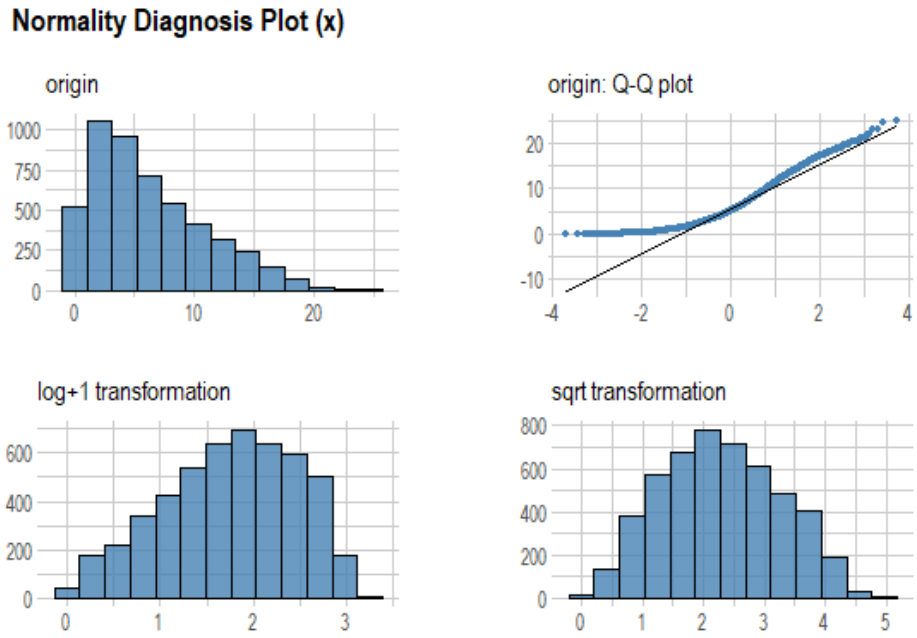

Figure 2.16: ttl_exp

tenure

* normality test : Shapiro-Wilk normality test
- statistic : 0.76993, p-value : 2.90219E-64

Table 2.17: skewness and kurtosis : tenure

type	skewness	kurtosis
original	1.9476	7.0068
log+1 transformation	0.4743	2.3095
sqrt transformation	0.7539	3.1117

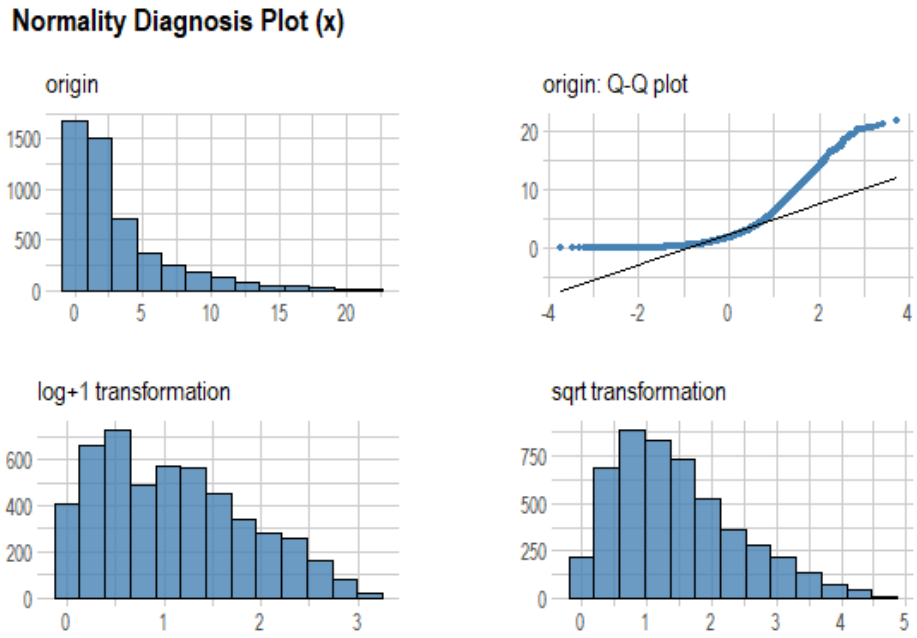

Figure 2.17: tenure

hours

* normality test : Shapiro-Wilk normality test
- statistic : 0.79611, p-value : 5.92868E-62

Table 2.18: skewness and kurtosis : hours

type	skewness	kurtosis
original	-0.8136	6.3747
log transformation	-2.9608	14.8960
sqrt transformation	-1.7858	7.7306

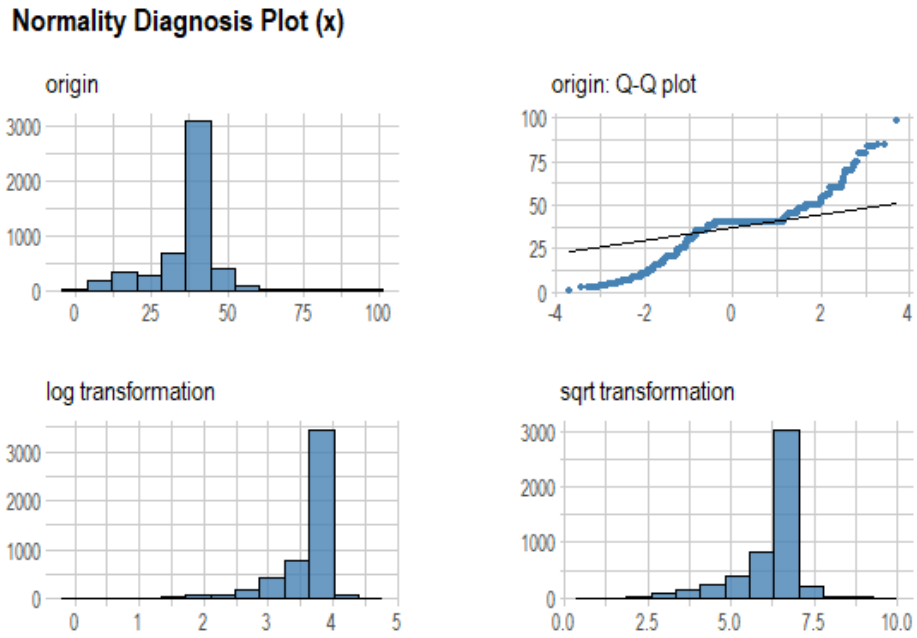

Figure 2.18: hours

wks_work

* normality test : Shapiro-Wilk normality test
- statistic : 0.94005, p-value : 5.4598E-41

Table 2.19: skewness and kurtosis : wks_work

type	skewness	kurtosis
original	0.1840	2.3422
log+1 transformation	-2.0768	8.1977
sqrt transformation	-0.7640	3.5478

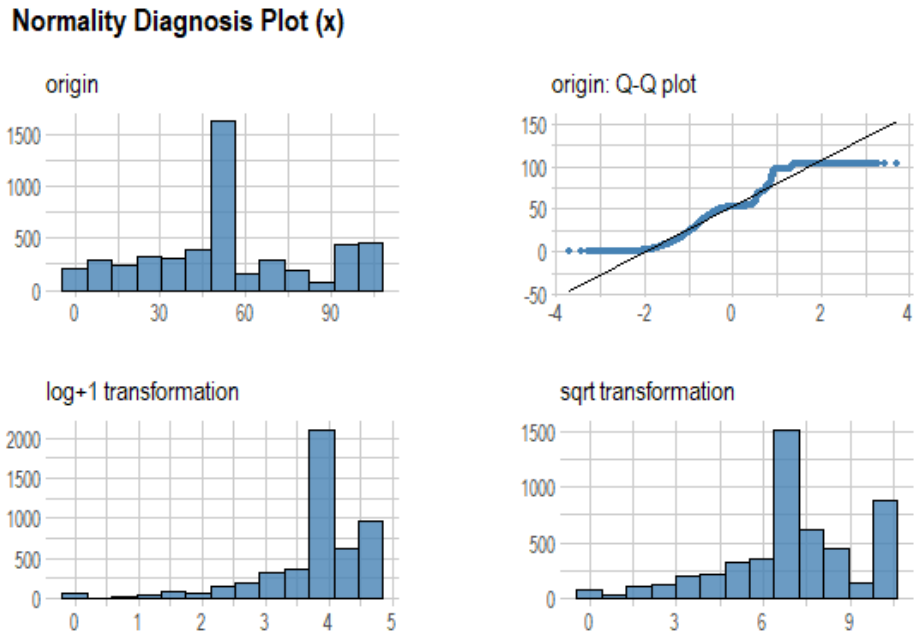

Figure 2.19: wks_work

ln_wage

* normality test : Shapiro-Wilk normality test
- statistic : 0.98043, p-value : 1.02353E-25

Table 2.20: skewness and kurtosis : ln_wage

type	skewness	kurtosis
original	0.3302	4.9095
log+1 transformation	-0.7450	5.8664
sqrt transformation	-0.7766	6.7710

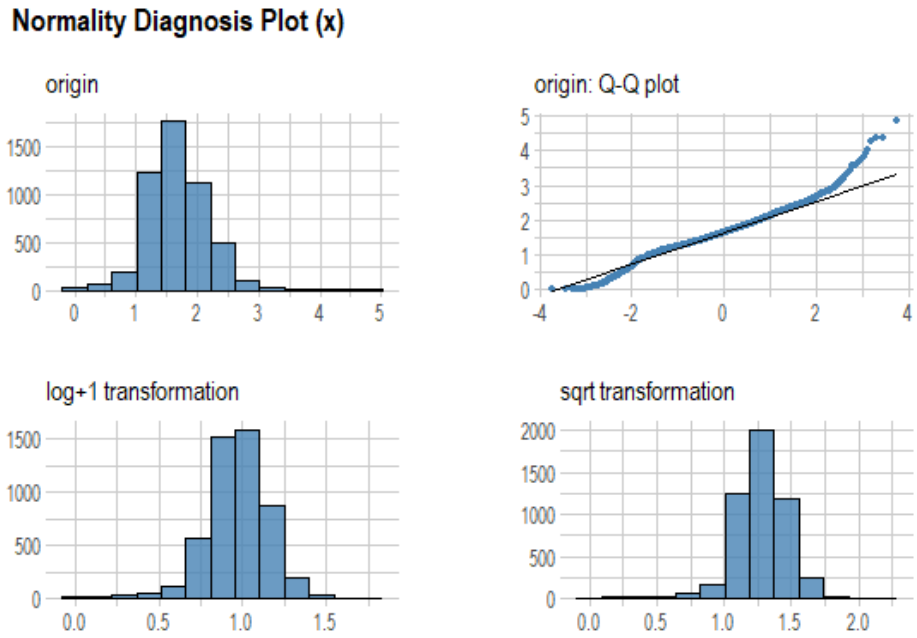

Figure 2.20: ln_wage

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

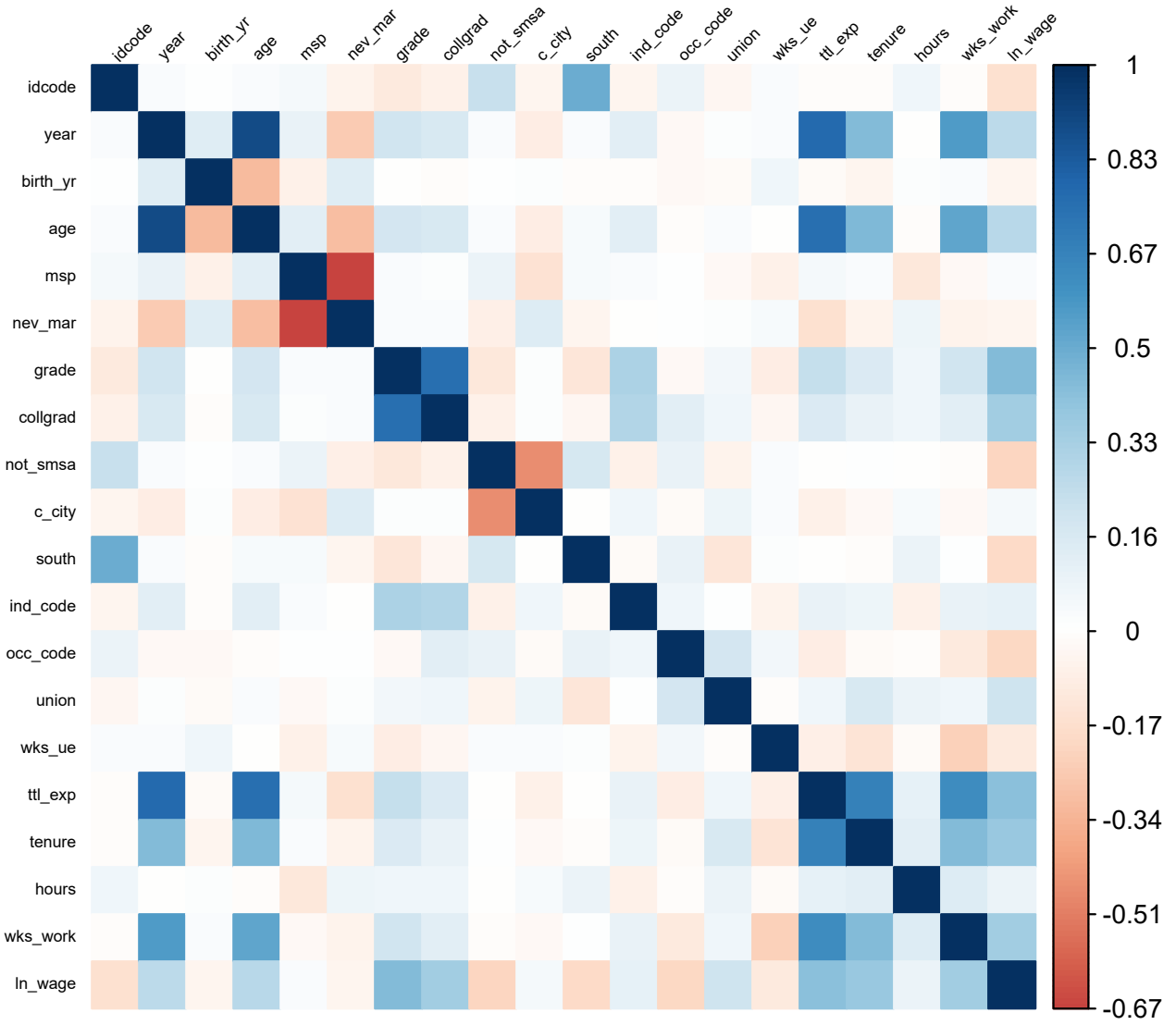

Figure 3.1: The correlation coefficient of numerical variables

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.