REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

June 29, 2021

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	0.00	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	numeric	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

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

```
lowest : 68 69 70 71 72, highest: 82 83 85 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

```
lowest : 41 42 43 44 45, highest: 50 51 52 53 54

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 : 1=white, 2=black, 3=other 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 yet 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

```
lowest :  0  1  2  3  4, highest: 14 15 16 17 18

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

```
lowest :  1  2  3  4  5, highest:  8  9 10 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

```
lowest :  1  2  3  4  5, highest:  9 10 11 12 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 : 0.95335, p-value : 3.03016E-37

Table 2.1: skewness and kurtosis : idcode

type	skewness	kurtosis
original	-0.0186	1.7933
log transformation	-2.1552	9.8890
sqrt transformation	-0.5878	2.4208

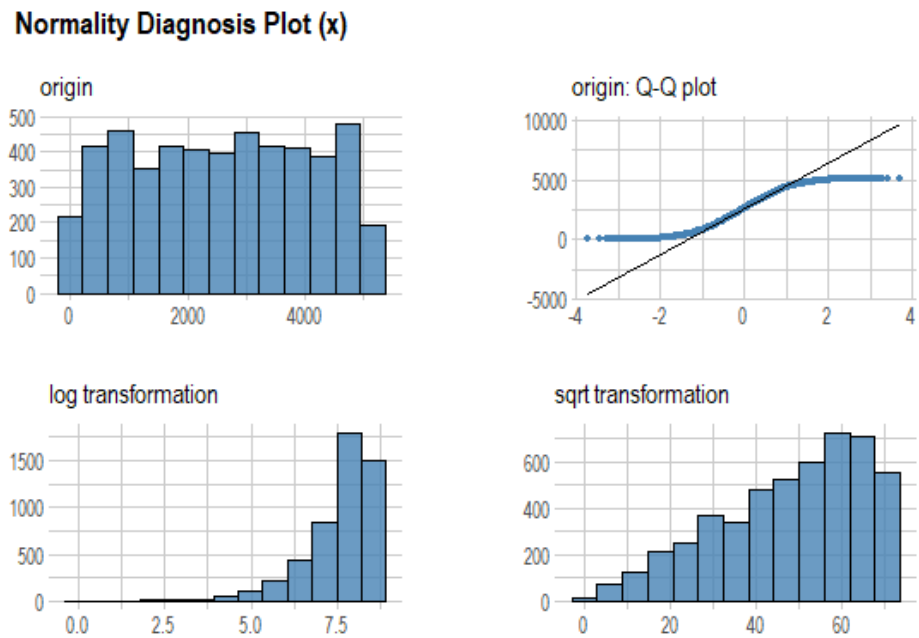

Figure 2.1: idcode

year

* normality test : Shapiro-Wilk normality test
- statistic : 0.9327, p-value : 8.84116E-43

Table 2.2: skewness and kurtosis : year

type	skewness	kurtosis
original	0.0539	1.7063
log transformation	-0.0326	1.7061
sqrt transformation	0.0107	1.7042

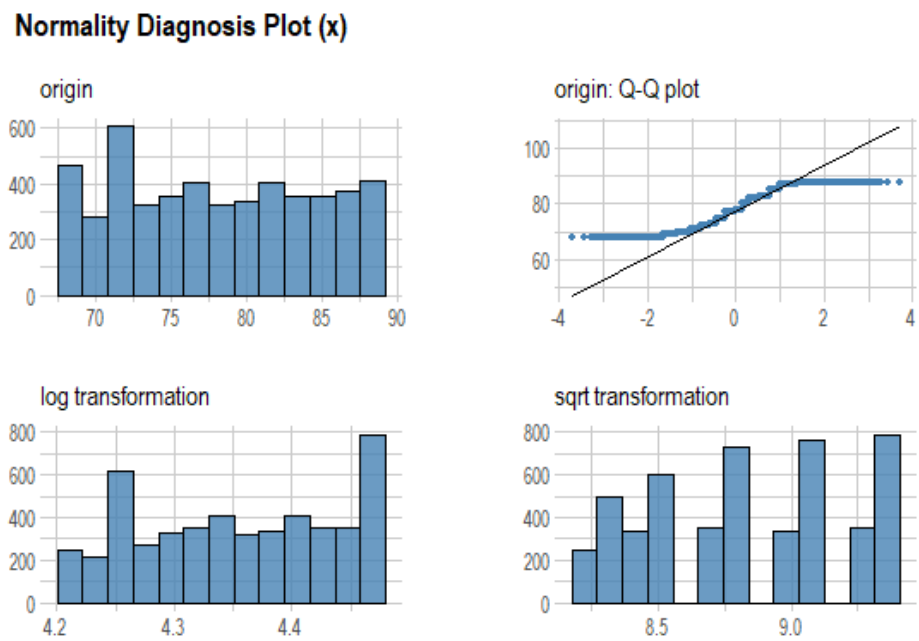

Figure 2.2: year

birth_yr

* normality test : Shapiro-Wilk normality test
- statistic : 0.95927, p-value : 2.68197E-35

Table 2.3: skewness and kurtosis : birth_yr

type	skewness	kurtosis
original	-0.1404	2.0045
log transformation	-0.2351	2.0659
sqrt transformation	-0.1874	2.0323

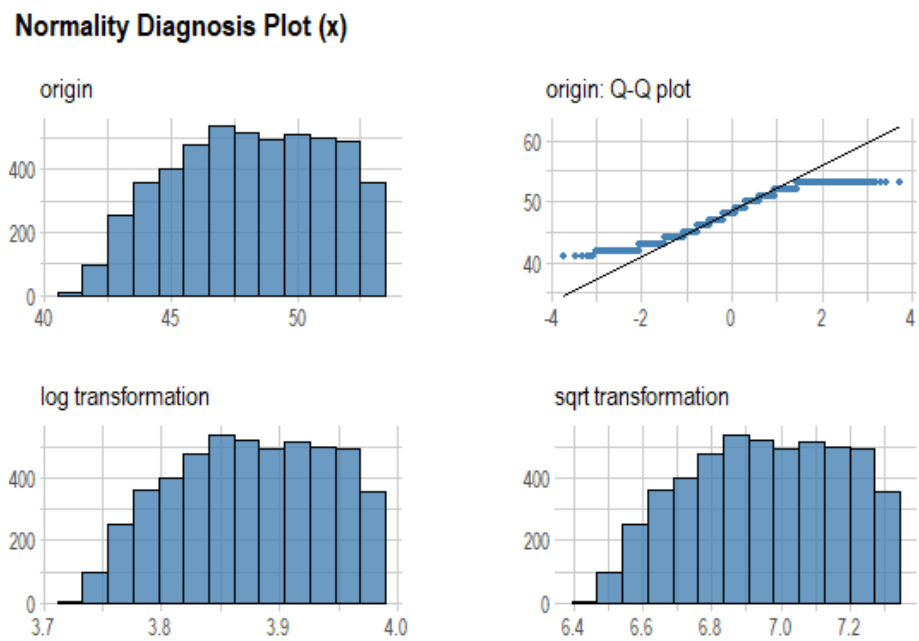

Figure 2.3: birth_yr

age

* normality test : Shapiro-Wilk normality test
- statistic : 0.96982, p-value : 3.47635E-31

Table 2.4: skewness and kurtosis : age

type	skewness	kurtosis
original	0.2608	2.1138
log transformation	-0.0952	2.0445
sqrt transformation	0.0837	2.0349

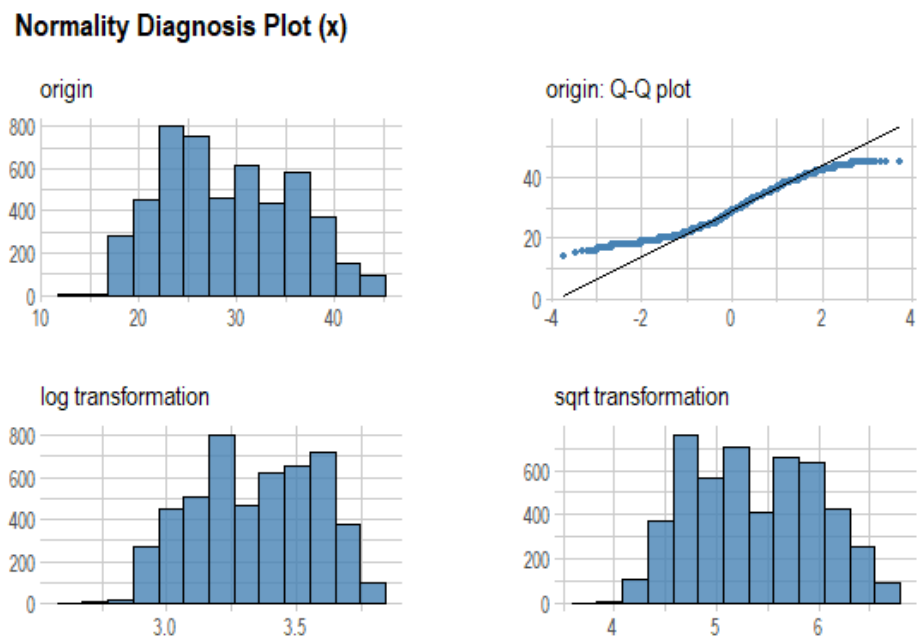

Figure 2.4: age

race

* normality test : Shapiro-Wilk normality test
- statistic : 0.6002, p-value : 2.18026E-75

Table 2.5: skewness and kurtosis : race

type	skewness	kurtosis
original	1.1974	3.2526
log transformation	0.9997	2.2225
sqrt transformation	1.0730	2.5839

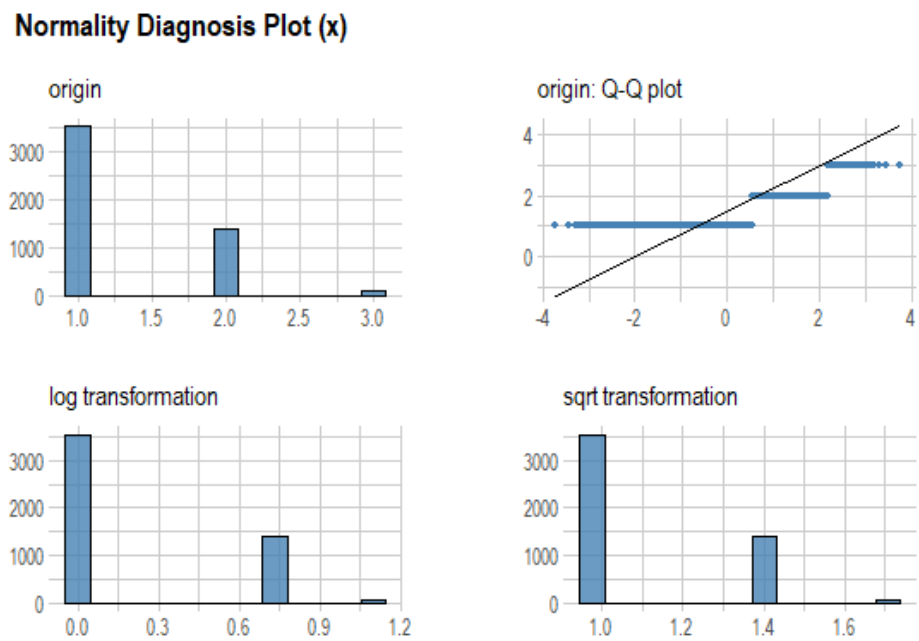

Figure 2.5: race

msp

* normality test : Shapiro-Wilk normality test
- statistic : 0.62095, p-value : 2.82566E-74

Table 2.6: skewness and kurtosis : msp

type	skewness	kurtosis
original	-0.4219	1.178
log+1 transformation	-0.4219	1.178
sqrt transformation	-0.4219	1.178

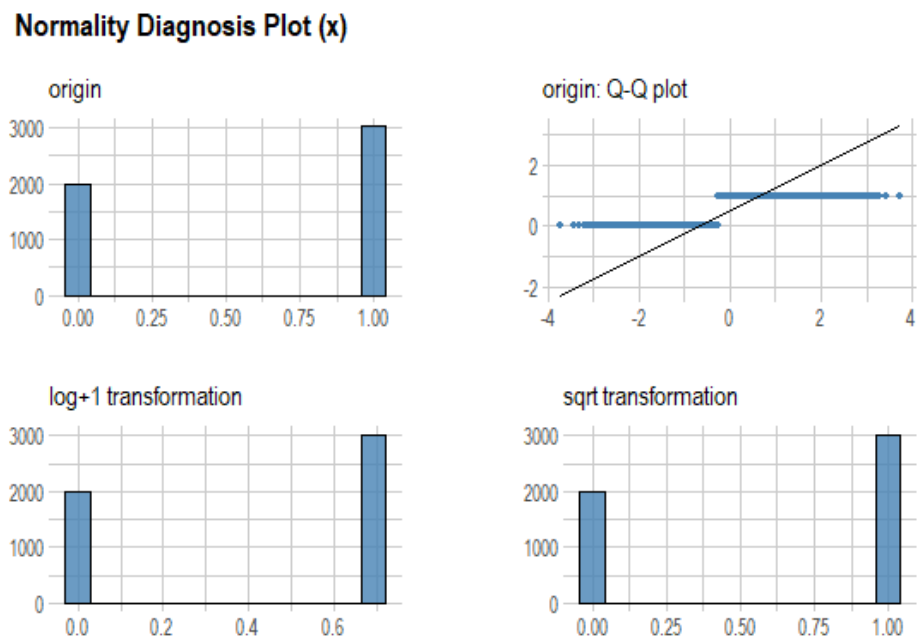

Figure 2.6: msp

nev_mar

* normality test : Shapiro-Wilk normality test
- statistic : 0.52324, p-value : 4.00837E-79

Table 2.7: skewness and kurtosis : nev_mar

type	skewness	kurtosis
original	1.2645	2.599
log+1 transformation	1.2645	2.599
sqrt transformation	1.2645	2.599

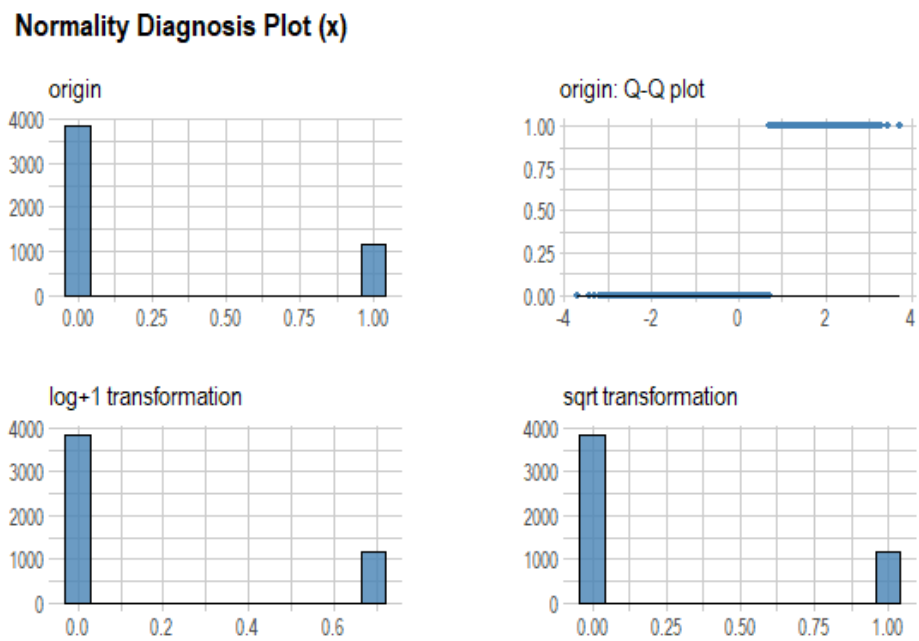

Figure 2.7: nev_mar

grade

* normality test : Shapiro-Wilk normality test
- statistic : 0.88887, p-value : 5.2873E-51

Table 2.8: skewness and kurtosis : grade

type	skewness	kurtosis
original	0.0608	4.4686
log+1 transformation	-2.9747	35.8701
sqrt transformation	-1.2752	14.5555

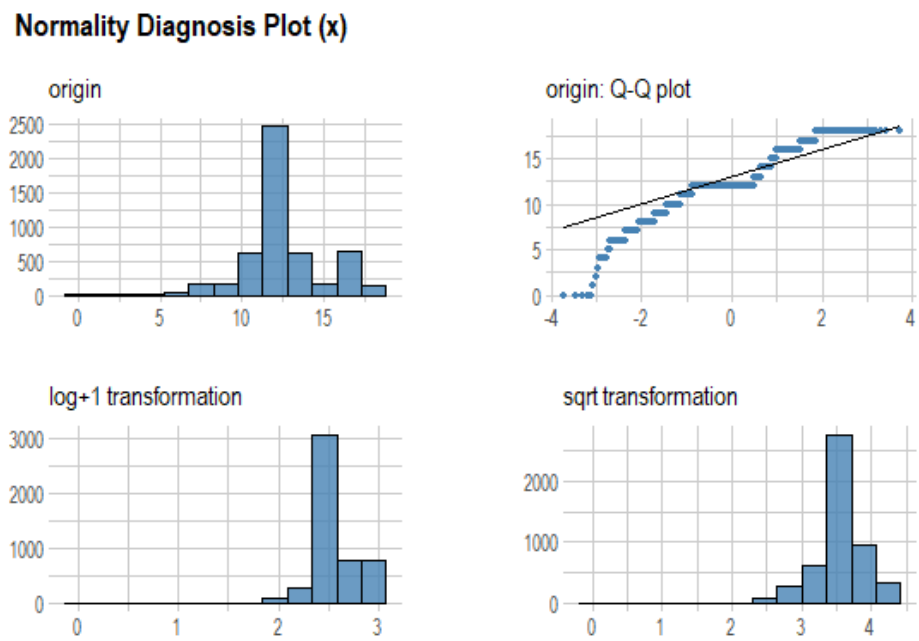

Figure 2.8: grade

collgrad

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

Table 2.9: skewness and kurtosis : collgrad

type	skewness	kurtosis
original	1.7551	4.0806
log+1 transformation	1.7551	4.0806
sqrt transformation	1.7551	4.0806

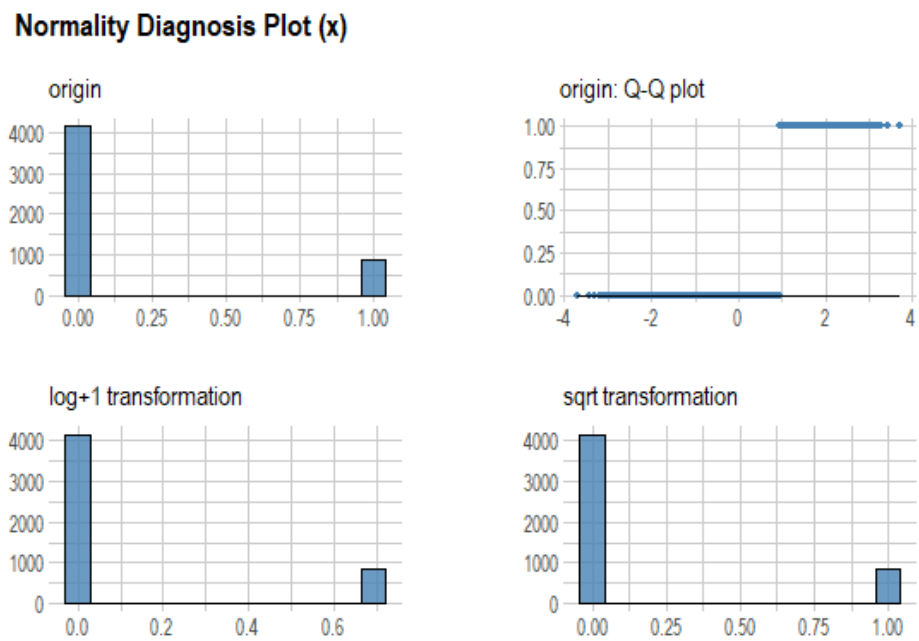

Figure 2.9: collgrad

not_smsa

* normality test : Shapiro-Wilk normality test
- statistic : 0.56701, p-value : 4.54703E-77

Table 2.10: skewness and kurtosis : not_smsa

type	skewness	kurtosis
original	0.9417	1.8868
log+1 transformation	0.9417	1.8868
sqrt transformation	0.9417	1.8868

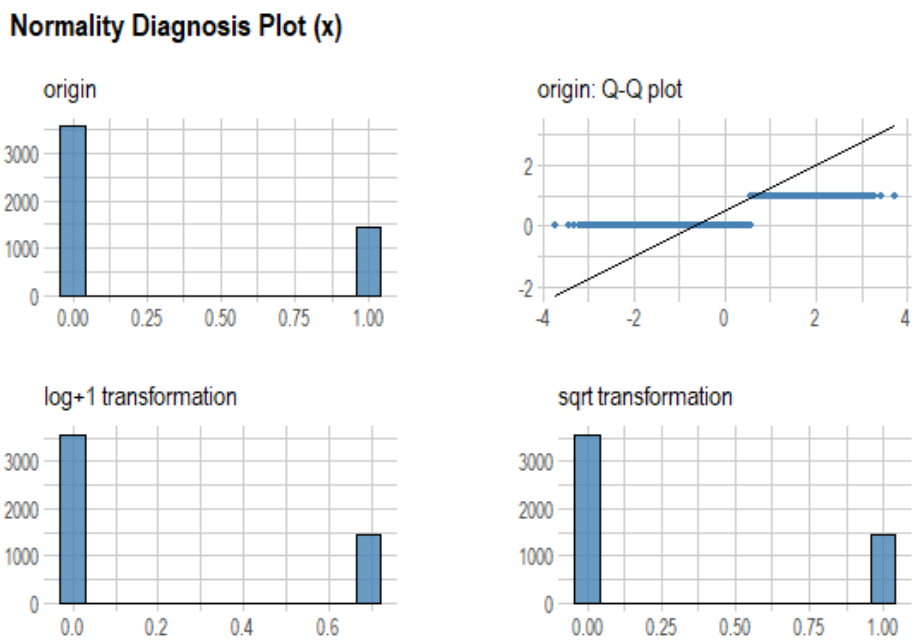

Figure 2.10: not_smsa

c_city

* normality test : Shapiro-Wilk normality test
- statistic : 0.60908, p-value : 6.42866E-75

Table 2.11: skewness and kurtosis : c_city

type	skewness	kurtosis
original	0.5662	1.3206
log+1 transformation	0.5662	1.3206
sqrt transformation	0.5662	1.3206

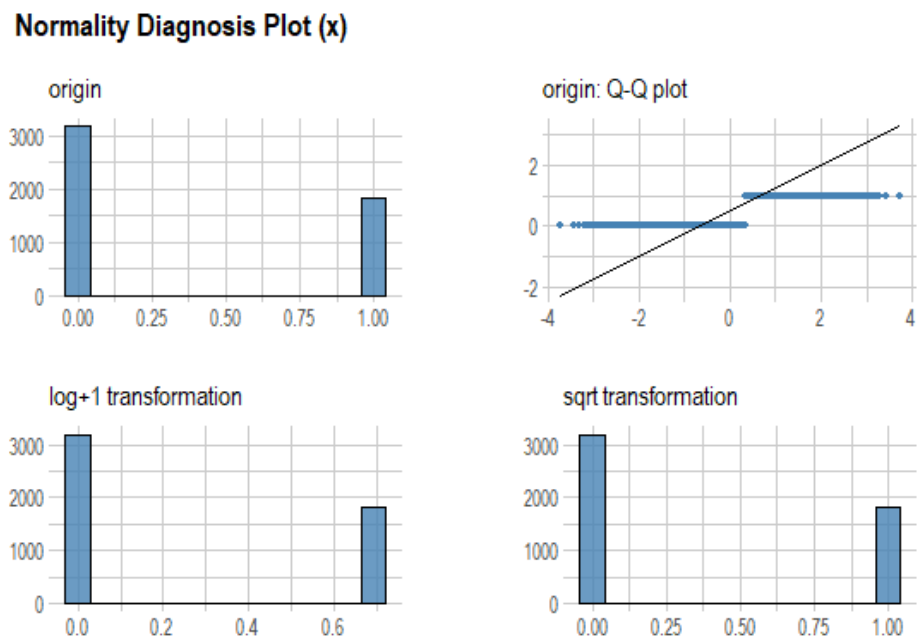

Figure 2.11: c_city

south

* normality test : Shapiro-Wilk normality test
- statistic : 0.62697, p-value : 6.07222E-74

Table 2.12: skewness and kurtosis : south

type	skewness	kurtosis
original	0.3292	1.1084
log+1 transformation	0.3292	1.1084
sqrt transformation	0.3292	1.1084

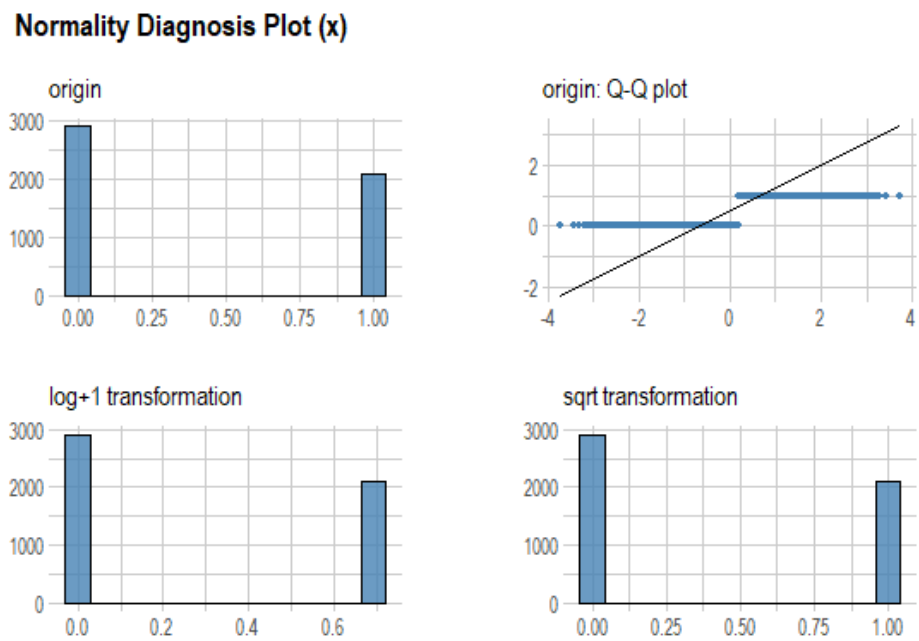

Figure 2.12: south

ind_code

* normality test : Shapiro-Wilk normality test
- statistic : 0.87345, p-value : 2.96237E-53

Table 2.13: skewness and kurtosis : ind_code

type	skewness	kurtosis
original	-0.0322	1.5612
log transformation	-0.9377	4.6286
sqrt transformation	-0.3148	2.1313

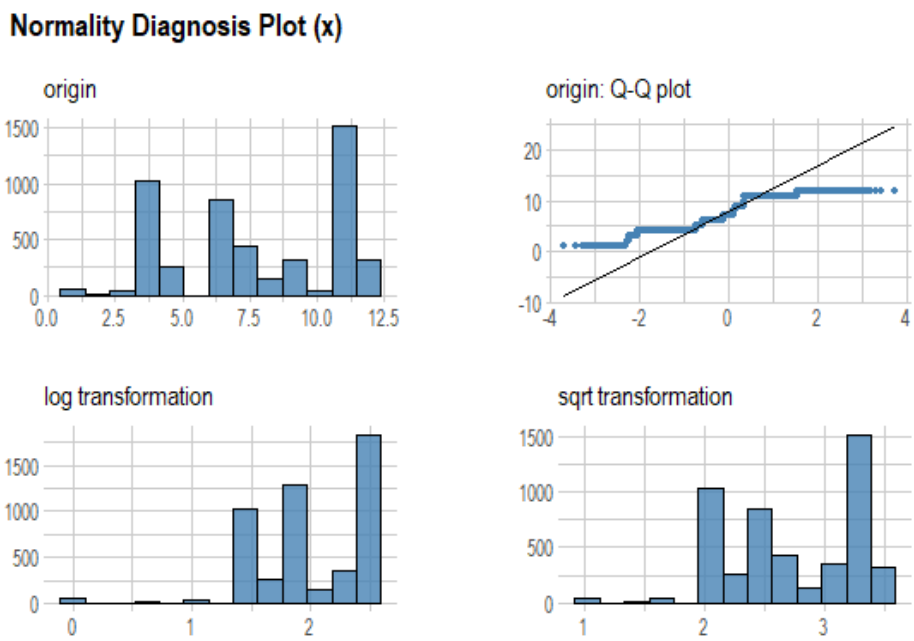

Figure 2.13: ind_code

occ_code

* normality test : Shapiro-Wilk normality test
- statistic : 0.85606, p-value : 1.54761E-55

Table 2.14: skewness and kurtosis : occ_code

type	skewness	kurtosis
original	1.0589	3.6479
log transformation	-0.3237	2.6963
sqrt transformation	0.4218	2.6135

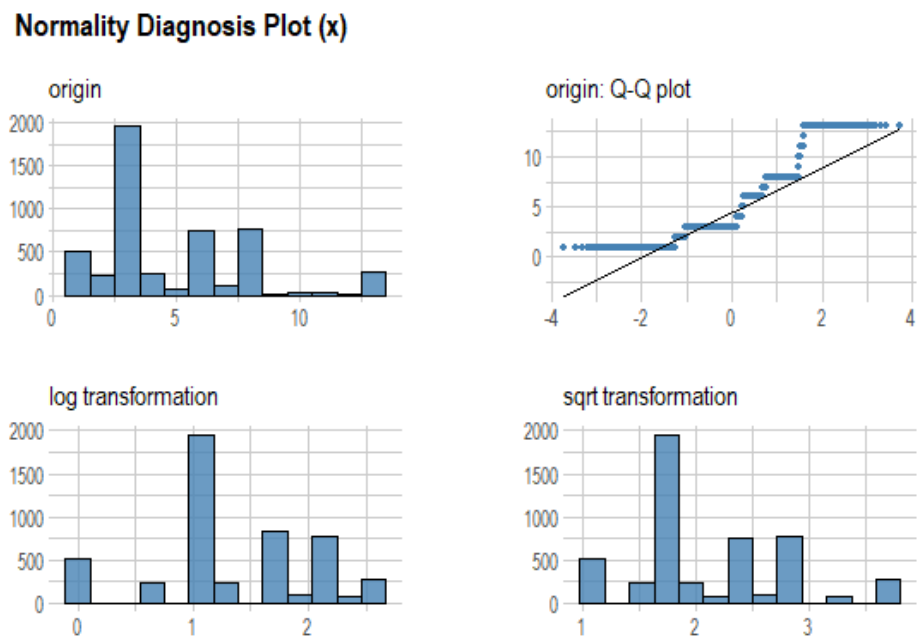

Figure 2.14: occ_code

union

* normality test : Shapiro-Wilk normality test
- statistic : 0.51698, p-value : 2.10091E-79

Table 2.15: skewness and kurtosis : union

type	skewness	kurtosis
original	1.3089	2.7132
log+1 transformation	1.3089	2.7132
sqrt transformation	1.3089	2.7132

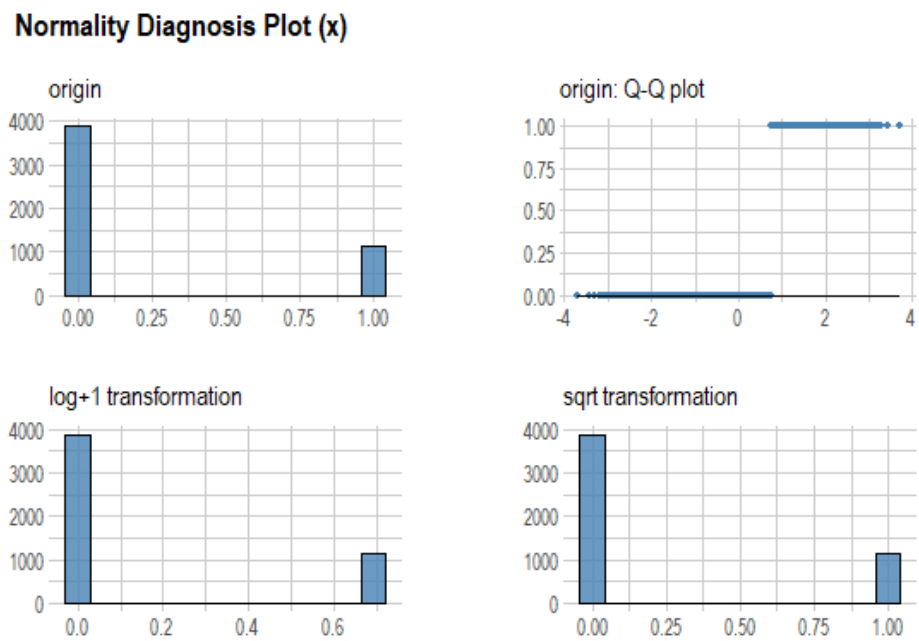

Figure 2.15: union

wks_ue

* normality test : Shapiro-Wilk normality test
- statistic : 0.41664, p-value : 1.61372E-83

Table 2.16: skewness and kurtosis : wks_ue

type	skewness	kurtosis
original	3.9020	20.1987
log+1 transformation	1.8752	5.2900
sqrt transformation	2.2432	7.5069

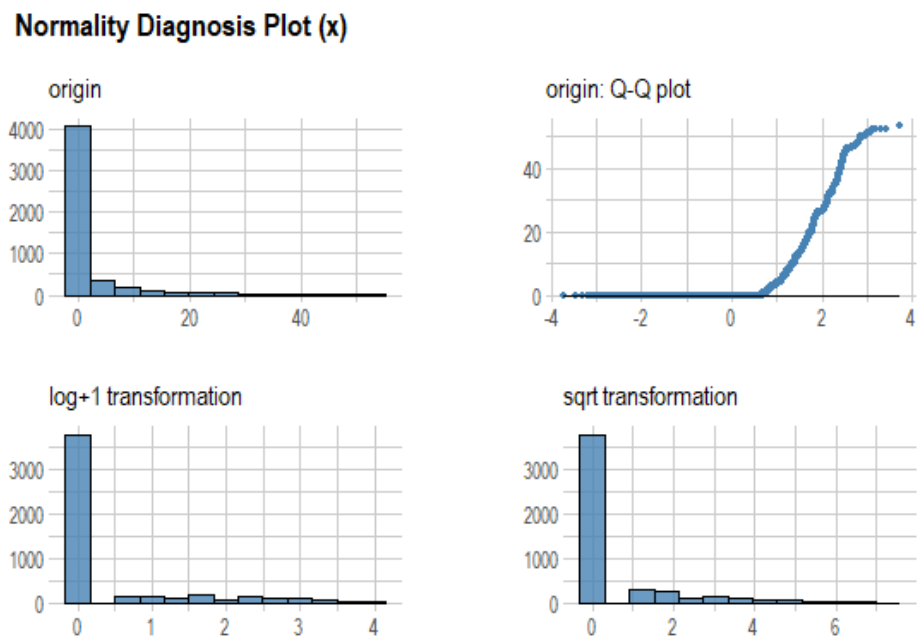

Figure 2.16: wks_ue

ttl_exp

* normality test : Shapiro-Wilk normality test
- statistic : 0.92189, p-value : 3.82888E-45

Table 2.17: skewness and kurtosis : ttl_exp

type	skewness	kurtosis
original	0.8734	3.0762
log+1 transformation	-0.2756	2.2406
sqrt transformation	0.1473	2.2534

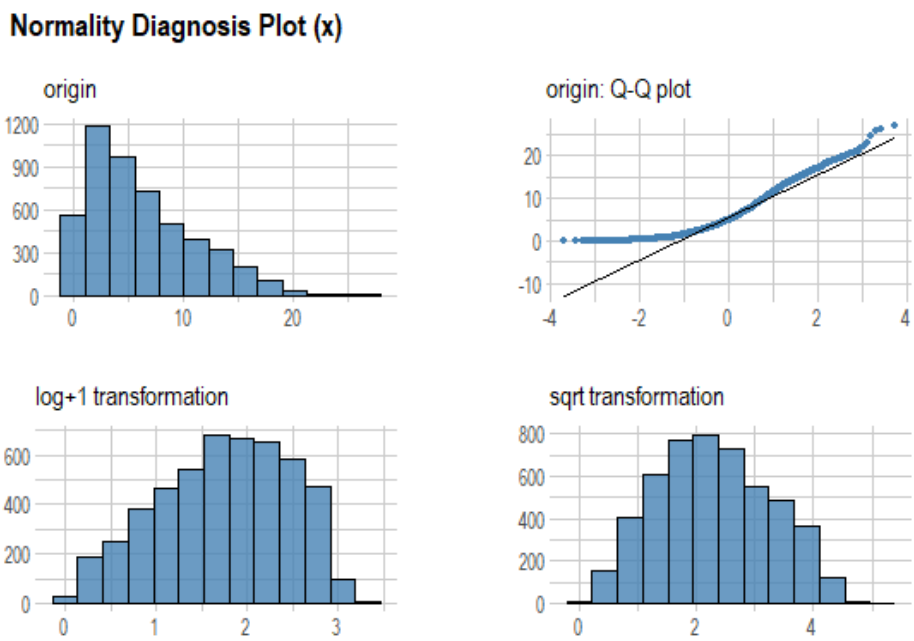

Figure 2.17: ttl_exp

tenure

* normality test : Shapiro-Wilk normality test
- statistic : 0.76061, p-value : 4.94015E-65

Table 2.18: skewness and kurtosis : tenure

type	skewness	kurtosis
original	1.9800	7.0899
log+1 transformation	0.5098	2.3353
sqrt transformation	0.7878	3.1482

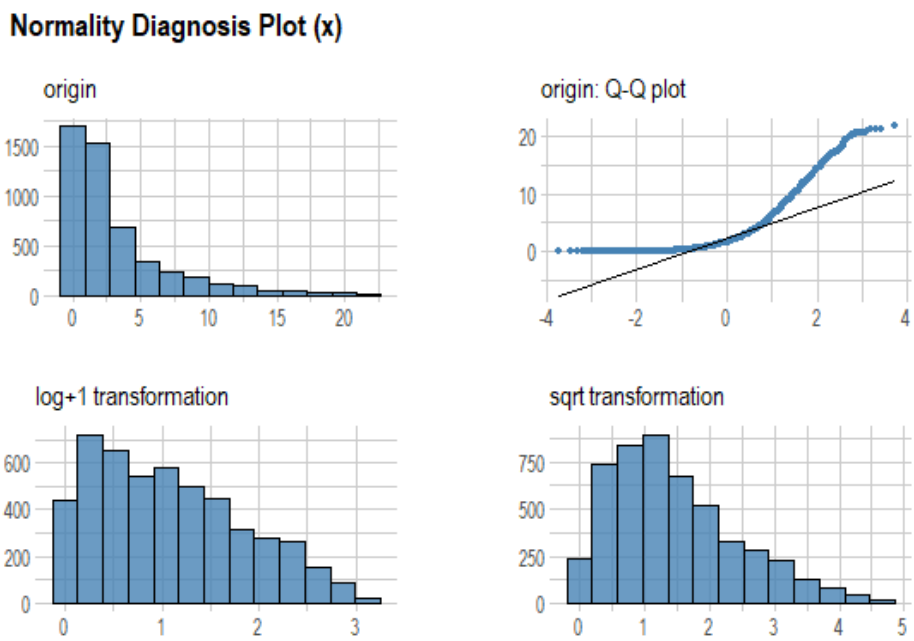

Figure 2.18: tenure

hours

* normality test : Shapiro-Wilk normality test
- statistic : 0.79068, p-value : 1.88059E-62

Table 2.19: skewness and kurtosis : hours

type	skewness	kurtosis
original	-1.1536	5.6712
log transformation	-3.2688	17.3195
sqrt transformation	-1.9935	8.0835

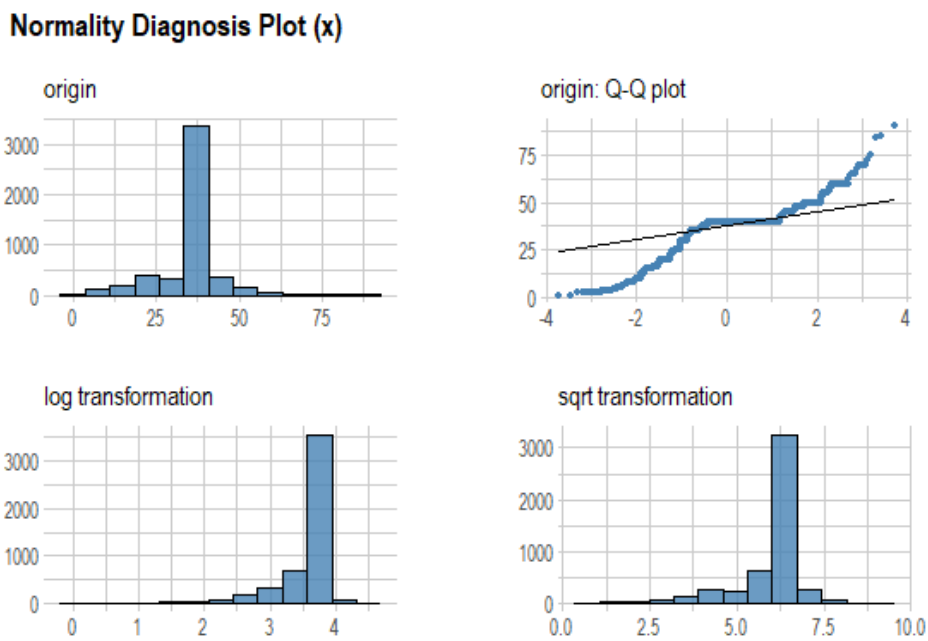

Figure 2.19: hours

wks_work

* normality test : Shapiro-Wilk normality test
- statistic : 0.93704, p-value : 9.60248E-42

Table 2.20: skewness and kurtosis : wks_work

type	skewness	kurtosis
original	0.1972	2.2772
log+1 transformation	-2.0342	7.8397
sqrt transformation	-0.7422	3.4341

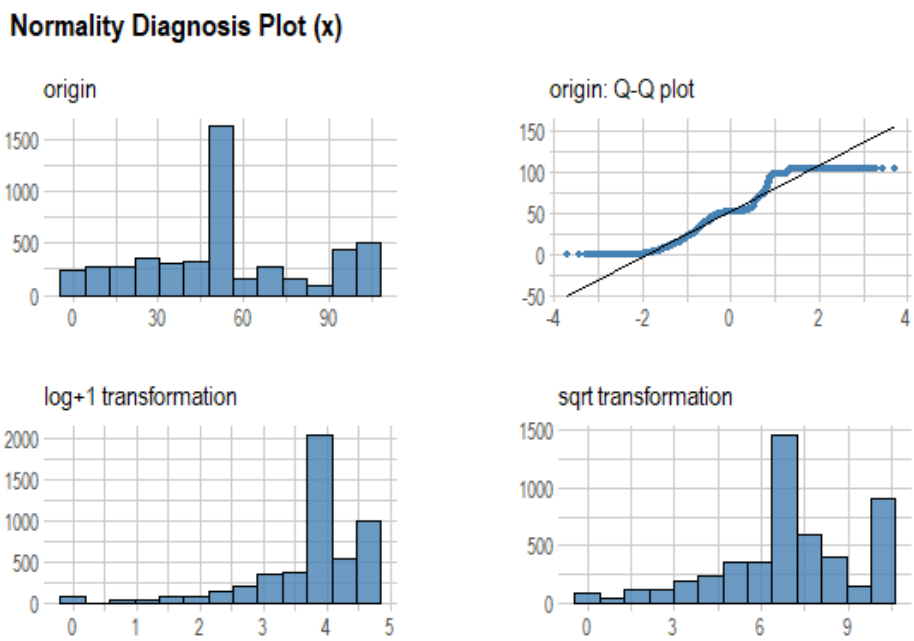

Figure 2.20: wks_work

ln_wage

* normality test : Shapiro-Wilk normality test
- statistic : 0.98245, p-value : 1.96031E-24

Table 2.21: skewness and kurtosis : ln_wage

type	skewness	kurtosis
original	0.2668	4.6209
log transformation	-4.0570	41.3402
sqrt transformation	-0.8189	6.9250

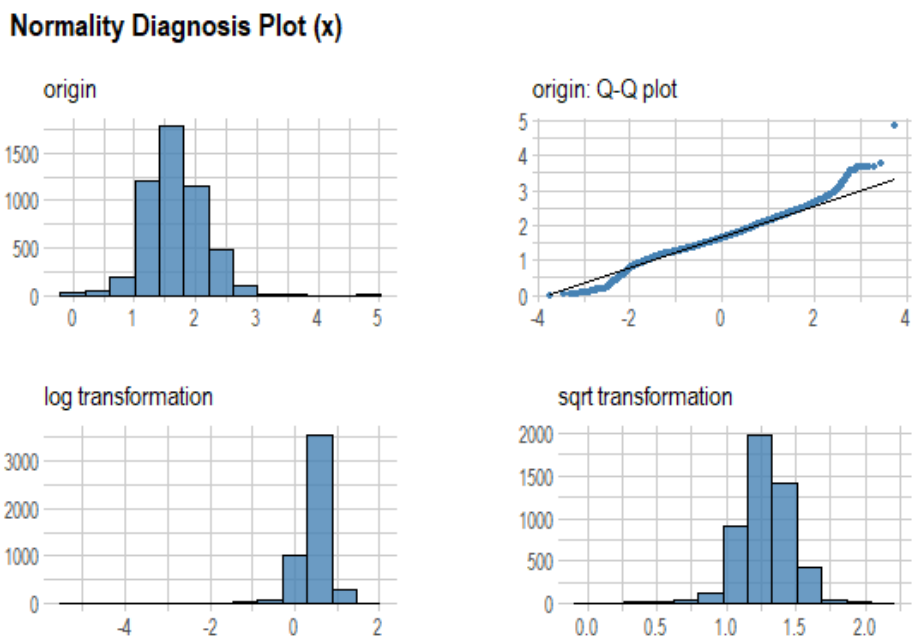

Figure 2.21: ln_wage

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

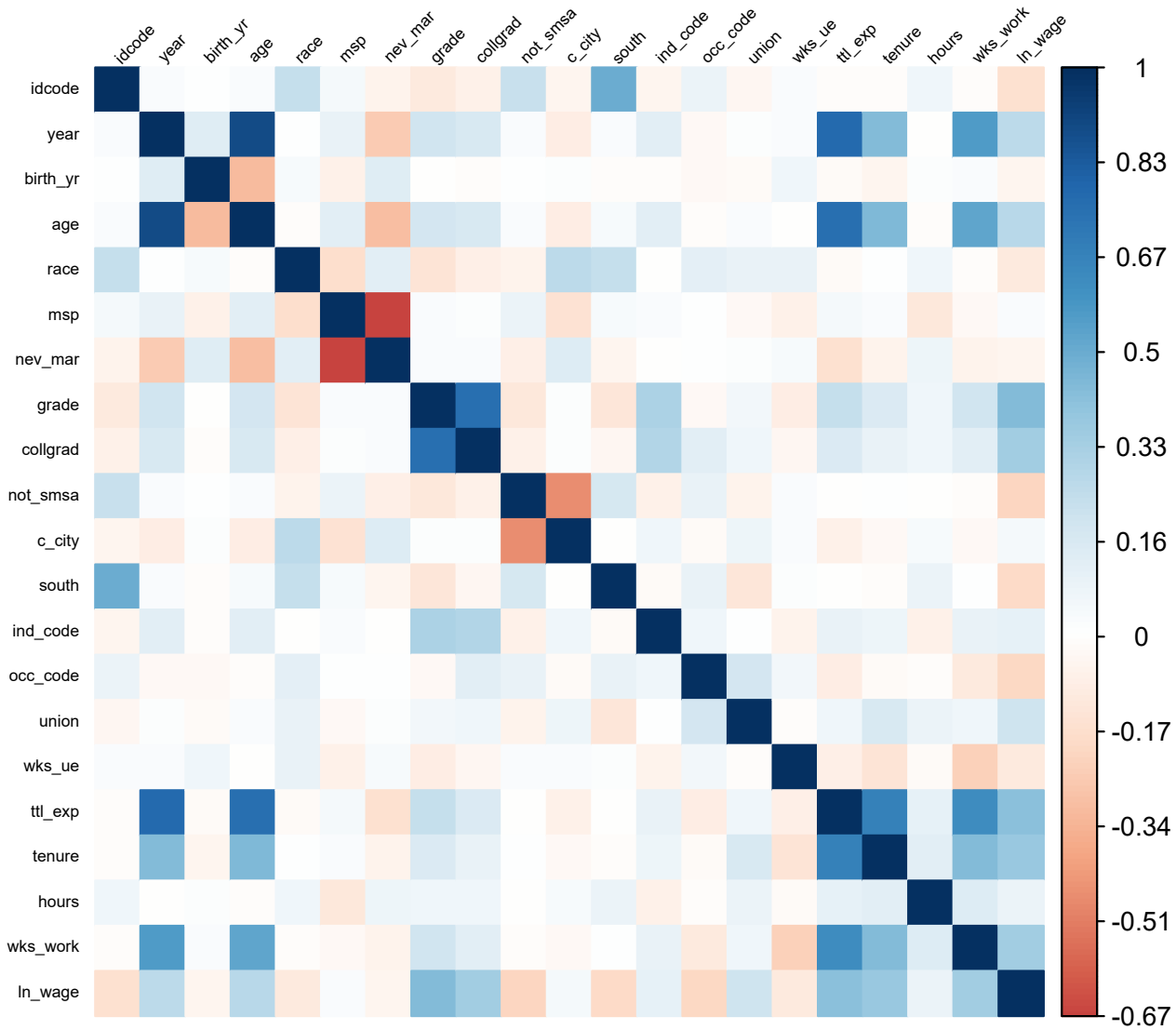

Figure 3.1: The correlation coefficient of numerical variables

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.