REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.6

March 15, 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 1,000 observations and 9 variables.

1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
ID	character	0	0.0	1000	1.000
Race	factor	107	10.7	8	0.008
Age	character	122	12.2	17	0.017
Sex	factor	0	0.0	2	0.002
Height(cm)	numeric	0	0.0	365	0.365
IQ	numeric	102	10.2	58	0.058
Smokes	logical	0	0.0	2	0.002
Income	factor	100	10.0	901	0.901
Died	logical	0	0.0	2	0.002

The target variable of the data is 'Died', and the data type of the variable is logical.

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
9 Variables 1000 Observations

ID

n	missing	distinct
1000	0	1000

```
lowest : 0001 0002 0003 0004 0005, highest: 0996 0997 0998 0999 1000
```

Race

n	missing	distinct
893	107	7

Value	White	Hispanic	Black	Asian	Bi-Racial	Native	Other
Frequency	579	146	114	28	18	7	1
Proportion	0.648	0.163	0.128	0.031	0.020	0.008	0.001

Age

n	missing	distinct
878	122	16

Value	20	21	22	23	24	25	26	27	28	29	30	31	32	33
Frequency	53	65	59	55	51	53	62	64	44	60	42	41	58	59
Proportion	0.060	0.074	0.067	0.063	0.058	0.060	0.071	0.073	0.050	0.068	0.048	0.047	0.066	0.067

Value	34	35
Frequency	48	64
Proportion	0.055	0.073

Sex

n	missing	distinct
1000	0	2

Value	Male	Female
Frequency	521	479
Proportion	0.521	0.479

Height(cm)

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1000	0	365	1	175.1	11.17	159.0	161.8	168.2	175.3	182.0	187.7	190.9

```
lowest : 146.3 146.7 147.4 147.9 150.7, highest: 198.6 199.0 200.4 201.1 207.2
```

IQ

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
898	102	57	0.999	100.2	11.32	84	87	93	100	107	113	117

```
lowest :  68  73  75  76  77, highest: 125 127 128 129 137
```

Smokes

n	missing	distinct
1000	0	2

```
Value      FALSE  TRUE
Frequency    809   191
Proportion 0.809 0.191
```

Income

n	missing	distinct
900	100	900

```
lowest : 592.09    1241.66   1288.4    1551.24   1751.98
highest: 144469.75 157951.53 158298.38 162248.13 167203.34
```

Died

n	missing	distinct
1000	0	2

```
Value      FALSE  TRUE
Frequency    465   535
Proportion 0.465 0.535
```

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

Height(cm)

normality test : Shapiro-Wilk normality test
statistic : 0.9974, p-value : 0.109544

type	skewness	kurtosis
original	-0.0851	2.7017
log transformation	-0.2291	2.7637
sqrt transformation	-0.1569	2.7240

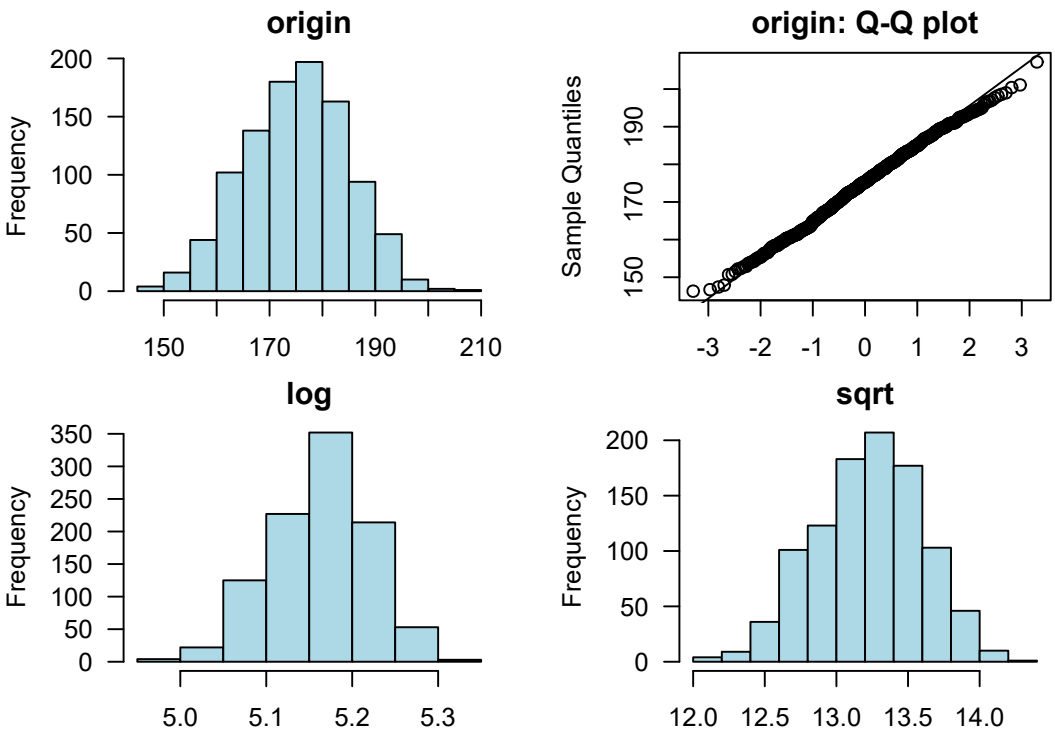

Figure 2.1: Height(cm)

IQ

normality test : Shapiro-Wilk normality test
statistic : 0.99821, p-value : 0.47445

type	skewness	kurtosis
original	0.0753	2.9651
log transformation	-0.2225	3.0516
sqrt transformation	-0.0725	2.9698

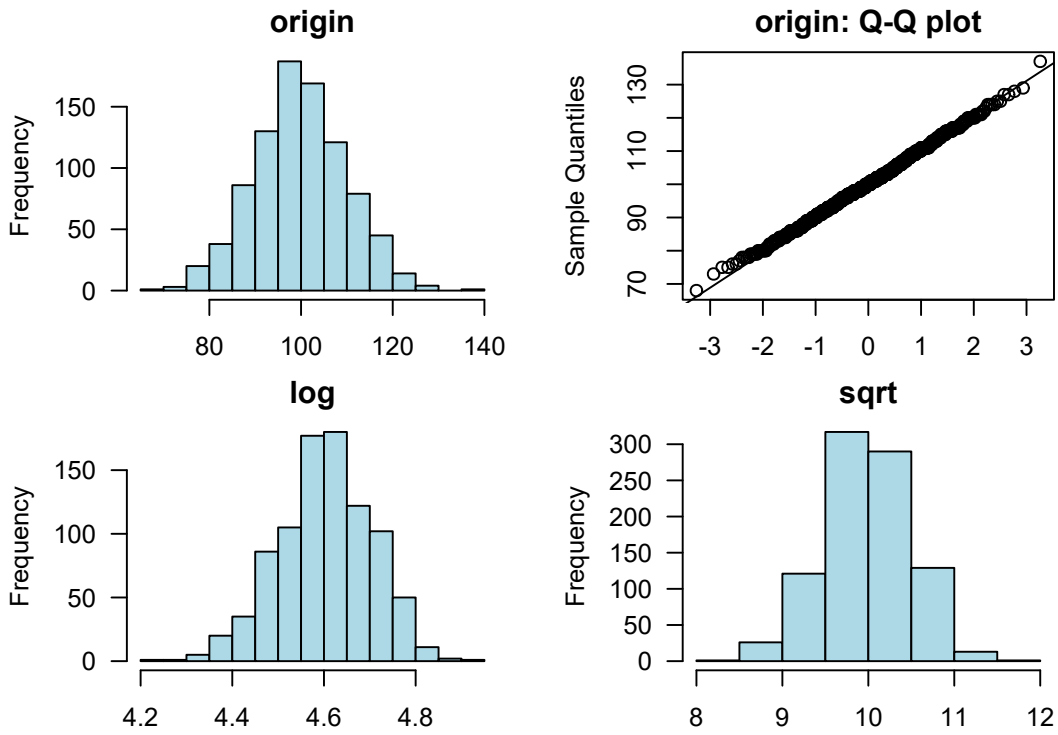

Figure 2.2: IQ

Chapter 3

Relationship Between Variables

3.1 Correlation Coefficient

3.1.1 Correlation Coefficient by Variable Combination

Number of numerical variables is less than 2.

3.1.2 Correlation Plot of Numerical Variables

```
Number of numerical variables is less than 2.
```

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

4.1.2 Grouped Categorical Variables

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

4.2.2 Grouped Correlation Plot of Numerical Variables