



REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.6

March 11, 2019

Contents

1	Introduction	3
1.1	Information of Dataset	3
1.2	Information of Variables	3
1.3	About EDA Report	3
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	9
3.1	Correlation Coefficient	9
3.1.1	Correlation Coefficient by Variable Combination	9
3.1.2	Correlation Plot of Numerical Variables	9
4	Target based Analysis	11
4.1	Grouped Descriptive Statistics	11
4.1.1	Grouped Numerical Variables	11
4.1.2	Grouped Categorical Variables	15
4.2	Grouped Relationship Between Variables	16
4.2.1	Grouped Correlation Coefficient	16
4.2.2	Grouped Correlation Plot of Numerical Variables	16

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 6,537 observations and 4 variables.

1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
age	numeric	0	0	65	0.0099434
race	factor	0	0	6	0.0009179
priors_count	numeric	0	0	33	0.0050482
is_recid	factor	0	0	2	0.0003060

The target variable of the data is 'is_recid', and the data type of the variable is factor.

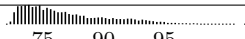
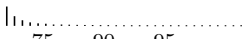
1.3 About EDA Report

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

4 Variables				edaData 6537 Observations											
age															
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
	6537	0	65	0.999	34.68	13.07	21	22	25	31	42	53	58		
lowest : 18 19 20 21 22, highest: 78 79 80 83 96															
race															
	n	missing	distinct												
	6537	0	6												
Value	African-American			Asian			Caucasian			Hispanic					
Frequency	3257			33			2256			535					
Proportion	0.498			0.005			0.345			0.082					
Value	Native American			Other											
Frequency	15			441											
Proportion	0.002			0.067											
priors_count															
	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
	6537	0	33	0.939	2.842	3.891	0	0	0	1	4	8	12		
lowest : 0 1 2 3 4, highest: 28 29 30 33 36															
is_recid															
	n	missing	distinct												
	6537	0	2												
Value	0		1												
Frequency	4317		2220												
Proportion	0.66		0.34												

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

age

normality test : Shapiro-Wilk normality test
 statistic : 0.91363, p-value : 8.91379E-47

type	skewness	kurtosis
original	0.9236	3.1213
log transformation	0.3893	2.2032
sqrt transformation	0.6500	2.5524

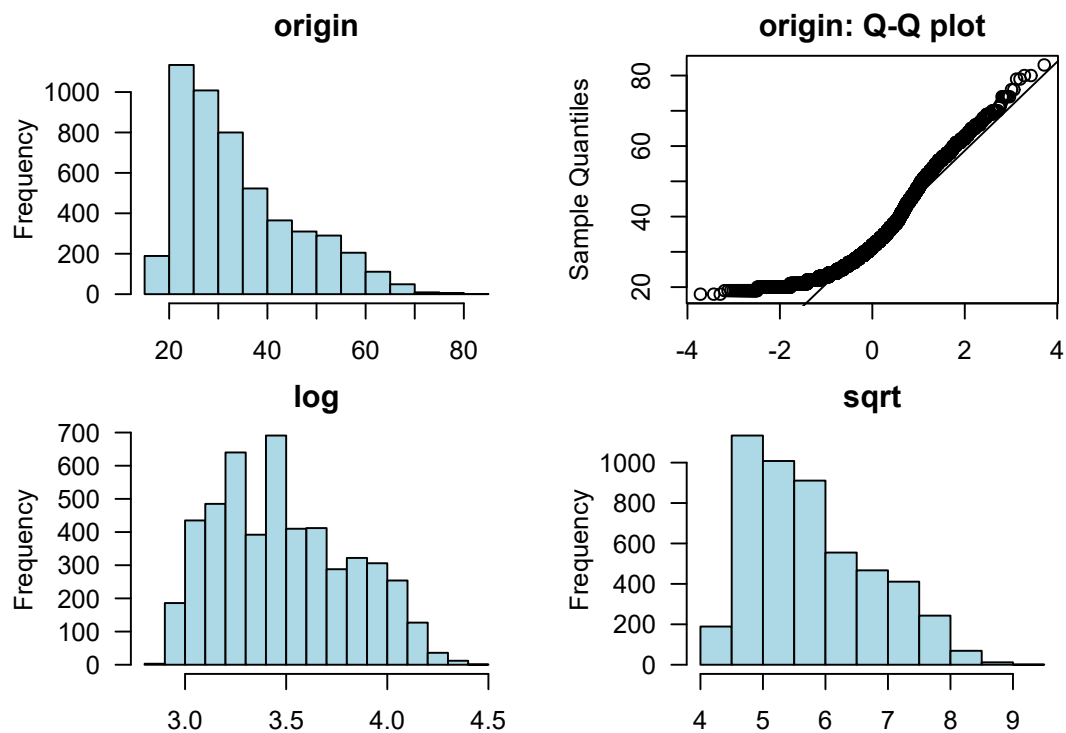


Figure 2.1: age

priors_count

normality test : Shapiro-Wilk normality test
 statistic : 0.68394, p-value : 1.51828E-70

type	skewness	kurtosis
original	2.4362	9.7818
log transformation		
sqrt transformation	0.8127	3.0658

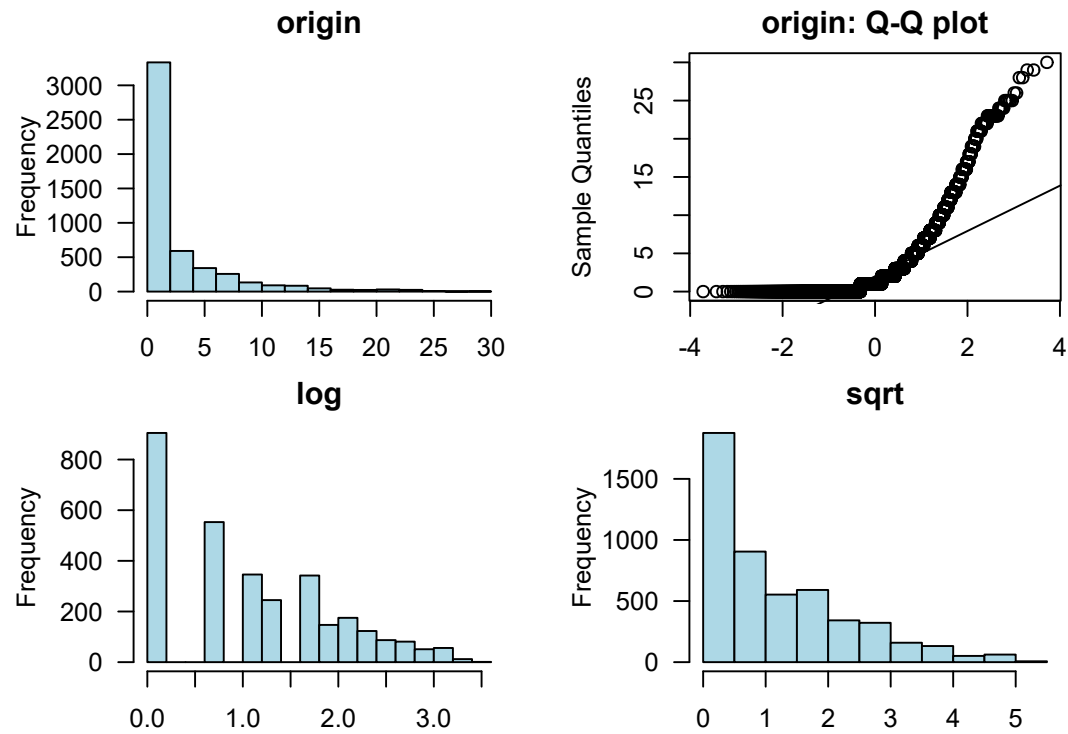


Figure 2.2: priors_count

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

age

Table 4.1: age

	1	0
n	2,220.00	4,317.00
NA	0.00	0.00
mean	32.08	36.02
sd	10.59	12.28
se(mean)	0.22	0.19
IQR	13.00	18.00
skewness	1.21	0.77
kurtosis	1.20	-0.18
0%	18.00	18.00
1%	20.00	20.00
5%	21.00	21.00
10%	21.00	22.00
20%	23.00	25.00
25%	24.00	26.00
30%	25.00	27.00
40%	27.00	30.00
50%	29.00	33.00
60%	31.00	37.00
70%	35.00	42.00
75%	37.00	44.00
80%	40.00	47.00
90%	49.00	54.00
95%	54.00	59.00
99%	62.00	68.00
100%	96.00	83.00

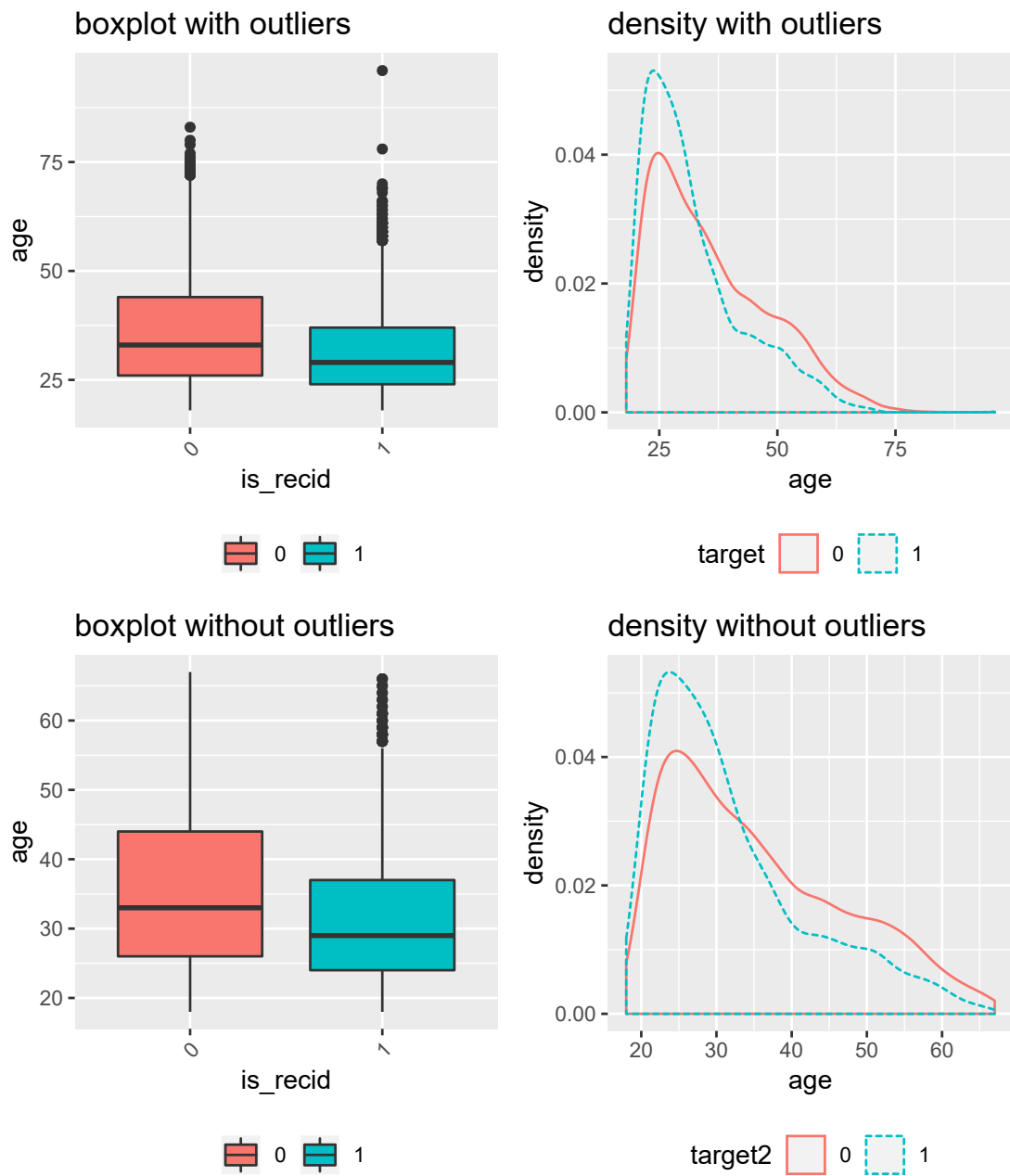
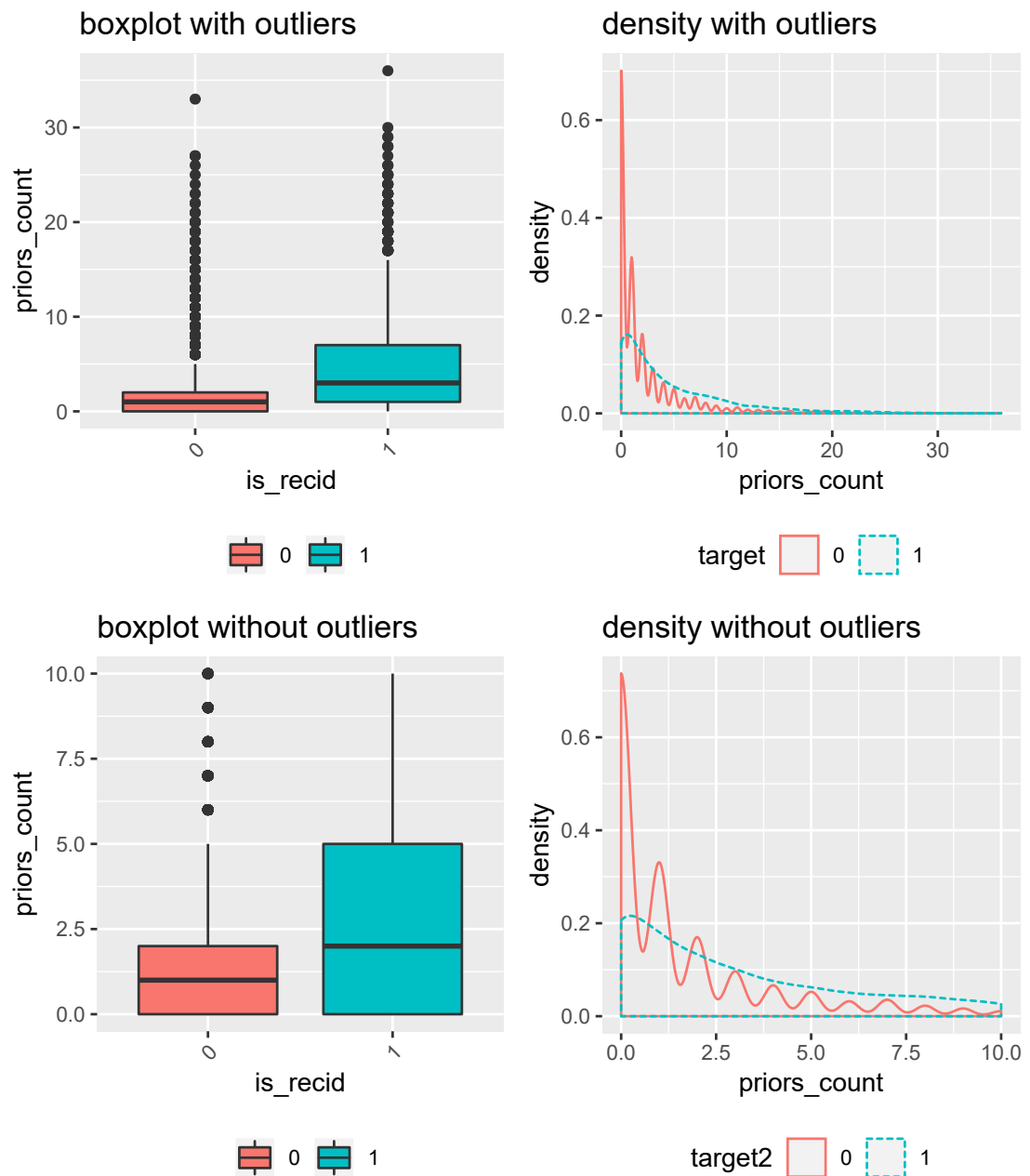


Figure 4.1: age

priors_count

Table 4.2: priors_count

	1	0
n	2,220.00	4,317.00
NA	0.00	0.00
mean	4.48	2.00
sd	5.33	3.41
se(mean)	0.11	0.05
IQR	6.00	2.00
skewness	1.81	3.04
kurtosis	3.60	12.11
0%	0.00	0.00
1%	0.00	0.00
5%	0.00	0.00
10%	0.00	0.00
20%	0.00	0.00
25%	1.00	0.00
30%	1.00	0.00
40%	2.00	0.00
50%	3.00	1.00
60%	4.00	1.00
70%	5.00	2.00
75%	7.00	2.00
80%	8.00	3.00
90%	12.00	6.00
95%	16.00	9.00
99%	23.81	17.00
100%	36.00	33.00

Figure 4.2: `priors_count`

4.1.2 Grouped Categorical Variables

race

	0	1	Sum
African-American	1,956	1,301	3,257
Asian	27	6	33
Caucasian	1,606	650	2,256
Hispanic	386	149	535
Native American	10	5	15
Other	332	109	441
Sum	4,317	2,220	6,537

	0	1	Sum
African-American	45.31	58.60	49.82
Asian	0.63	0.27	0.50
Caucasian	37.20	29.28	34.51
Hispanic	8.94	6.71	8.18
Native American	0.23	0.23	0.23
Other	7.69	4.91	6.75
Sum	100.00	100.00	100.00

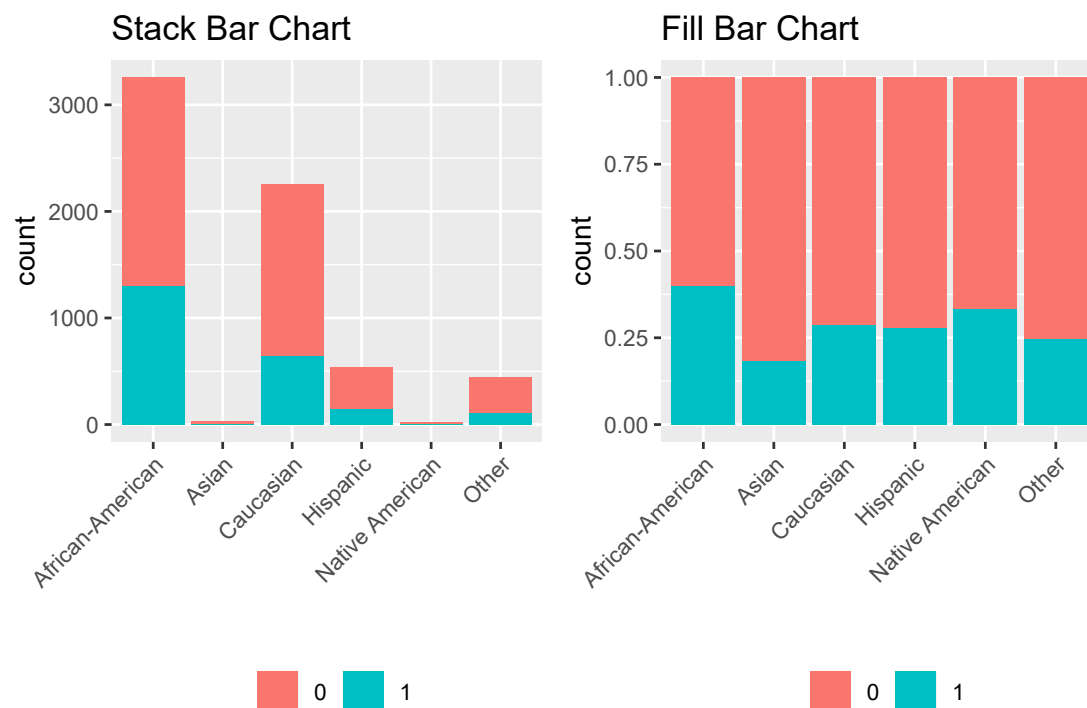


Figure 4.3: race

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

Number of numerical variables is less than 2.

4.2.2 Grouped Correlation Plot of Numerical Variables

Number of numerical variables is less than 2.