데이터 진단 보고서

DIAMONDS

보고서 개요

이 보고서는 diamonds의 데이터 품질 진단을 위해 작성되었습니다. 탐색적 데이터 분석(EDA, 기술통계)를 수행하기 전, 개별 변수들의 유효성을 판단하기 위해 작성되었습니다.

Contents

Overview

Data Structures

division	metrics	value
size	observations	1,500
size	variables	10
size	values	15,000
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	1,500
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	7
data type	integers	0
data type	factors/ordered	3
data type	characters	0
data type	Dates	0
data type	POSIXcts	0
data type	others	0

Table 1: Data structures and types

Job Informations

division	metrics	value
dataset	dataset	.
dataset	dataset type	tbl_df
job	samples	1,500 / 1,500 (100%)
job	created	2021-10-06 22:12:51
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
1	5	0

Table 3: Summary of warnings

warnings	status	recommand
z has 1 (0.07%) zeros	zero	check
price has 92 (6.13%) outliers	outlier	judgement
depth has 72 (4.8%) outliers	outlier	judgement
carat has 54 (3.6%) outliers	outlier	judgement
table has 14 (0.93%) outliers	outlier	judgement
z has 1 (0.07%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables

Variables

variables	types	missing	cardinality	zero	minus	outlier
carat	numeric					X
cut	ordered					
color	ordered					
clarity	ordered					
depth	numeric					X
table	numeric					X
price	numeric					X
x	numeric					
y	numeric					
z	numeric			X		X

Table 5: List of variables diagnosis

Missing Values

List of Missing Values

No variables including missing values

Visualization

No variables including missing values

Unique Values

Categorical Vaiables

No variable with a high proportion greater than 0.5

Numerical Vaiables

No variable with unique data proportion less than 5

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
clarity	SI1	363	24.2
clarity	VS2	343	22.9
clarity	SI2	259	17.3
clarity	VS1	209	13.9
clarity	VVS2	149	9.9
clarity	VVS1	107	7.1
clarity	IF	41	2.7
clarity	I1	29	1.9
color	G	345	23.0
color	E	290	19.3
color	F	237	15.8
color	H	235	15.7
color	D	189	12.6
color	I	136	9.1
color	J	68	4.5
cut	Ideal	590	39.3
cut	Premium	382	25.5
cut	Very Good	338	22.5
cut	Good	135	9.0
cut	Fair	55	3.7

Table 6: Top 10 levels of categorical variables

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
carat	0.20	0.40	0.79	0.70	1.03	2.80	0	0	54
depth	53.40	61.00	61.79	61.85	62.60	70.20	0	0	72
table	51.00	56.00	57.46	57.00	59.00	66.00	0	0	14
price	365.00	960.00	3,780.12	2,415.50	5,068.00	18,791.00	0	0	92
x	3.81	4.74	5.71	5.68	6.50	8.90	0	0	0
y	3.78	4.75	5.71	5.69	6.50	8.85	0	0	0
z	0.00	2.93	3.53	3.53	4.02	5.53	1	0	1

Table 7: General list of numerical diagnosis

Zero Values

variables	min	median	max	zero	zero (%)
z	0	3.53	5.53	1	0.1

Table 8: List of numerical diagnosis (zero)

Negative Values

No numeric variable with negative value

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
price	365.0	2,415.50	18,791.00	92	6.1
depth	53.4	61.85	70.20	72	4.8
carat	0.2	0.70	2.80	54	3.6
table	51.0	57.00	66.00	14	0.9
z	0.0	3.53	5.53	1	0.1

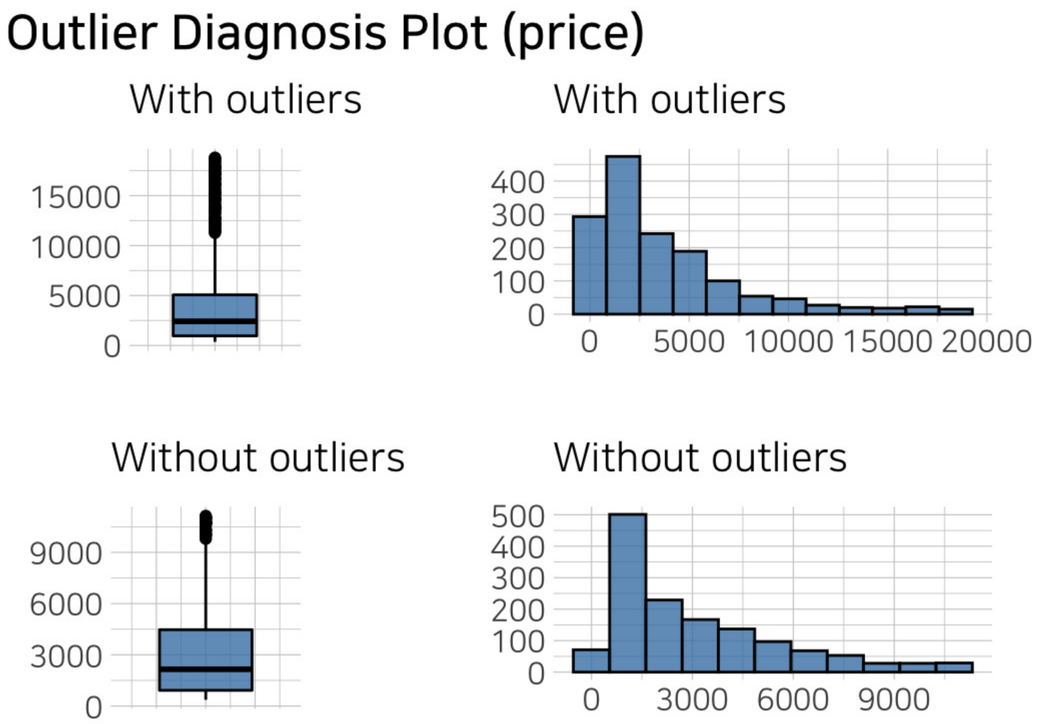
Table 9: Diagnosis of numerical variable outliers

Individual Outliers

variable: price

Measures	Values
Outliers count	92
Outliers ratio (%)	6.13%
Mean of outliers	15037.88
Mean with outliers	3780.121
Mean without outliers	3044.529

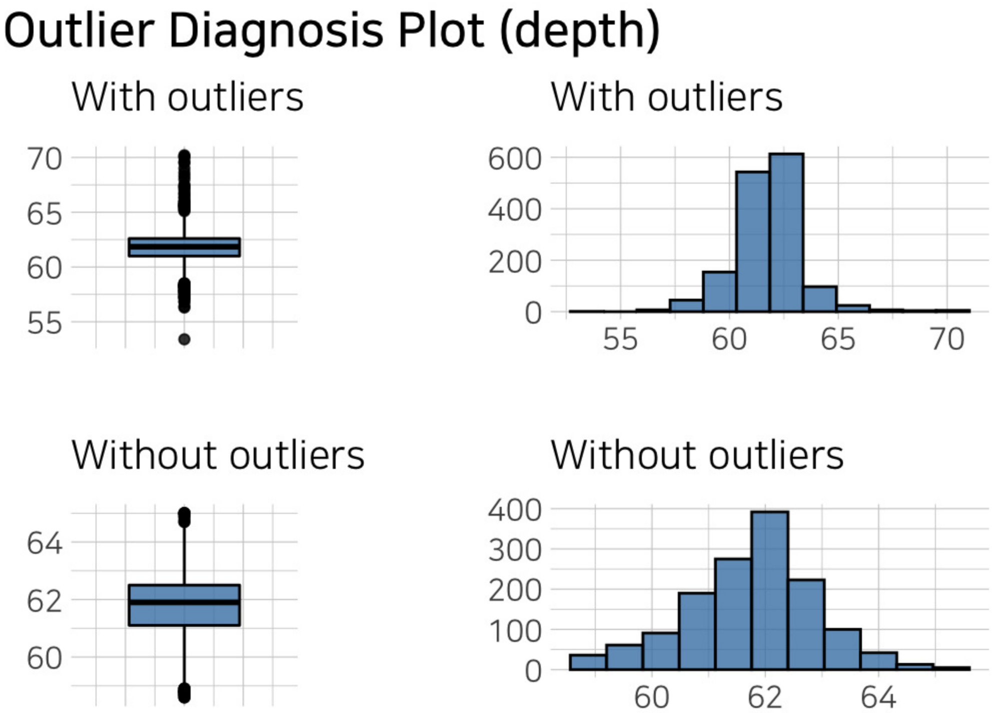
Table 10: price



variable: depth

Measures	Values
Outliers count	72
Outliers ratio (%)	4.8%
Mean of outliers	62.15278
Mean with outliers	61.78753
Mean without outliers	61.76912

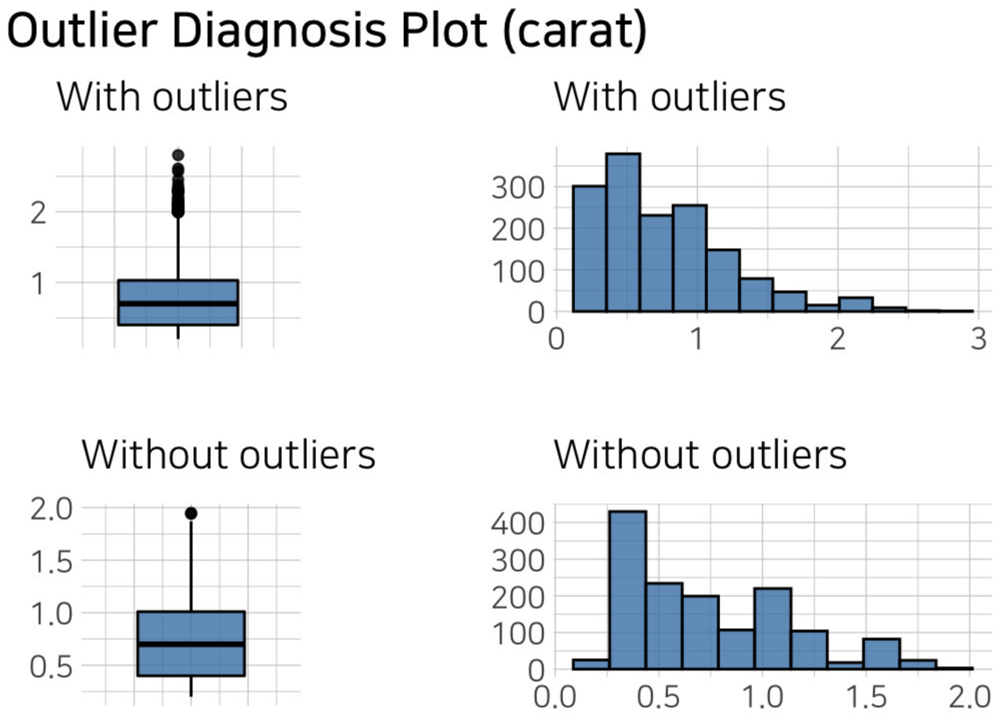
Table 10: depth



variable: carat

Measures	Values
Outliers count	54
Outliers ratio (%)	3.6%
Mean of outliers	2.129815
Mean with outliers	0.7877267
Mean without outliers	0.7376072

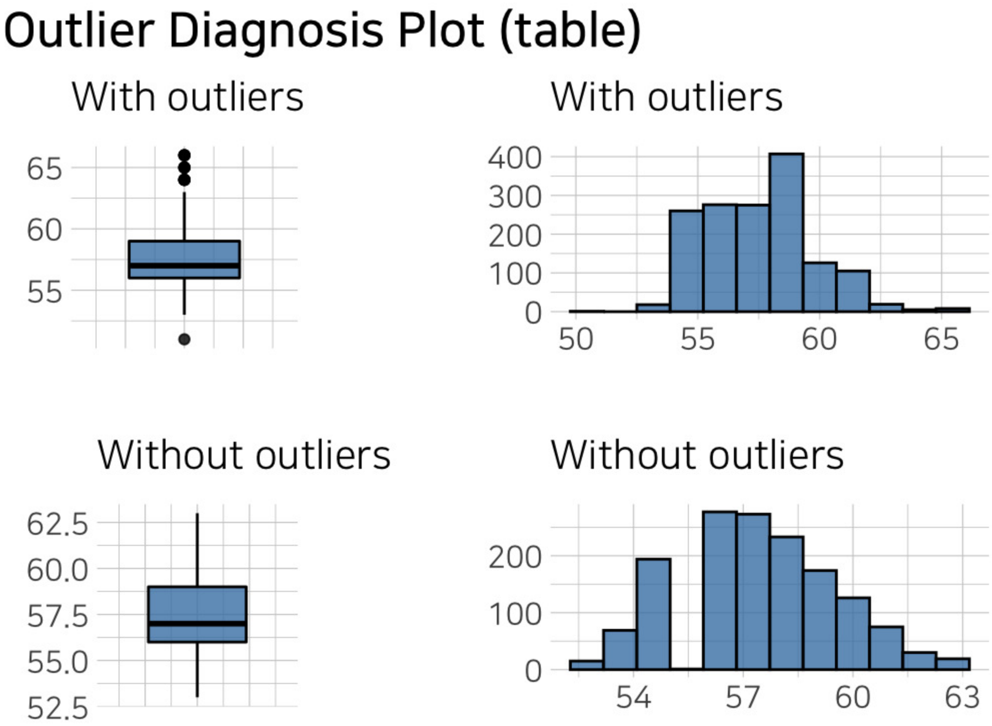
Table 10: carat



variable: table

Measures	Values
Outliers count	14
Outliers ratio (%)	0.93%
Mean of outliers	63.92857
Mean with outliers	57.45647
Mean without outliers	57.39549

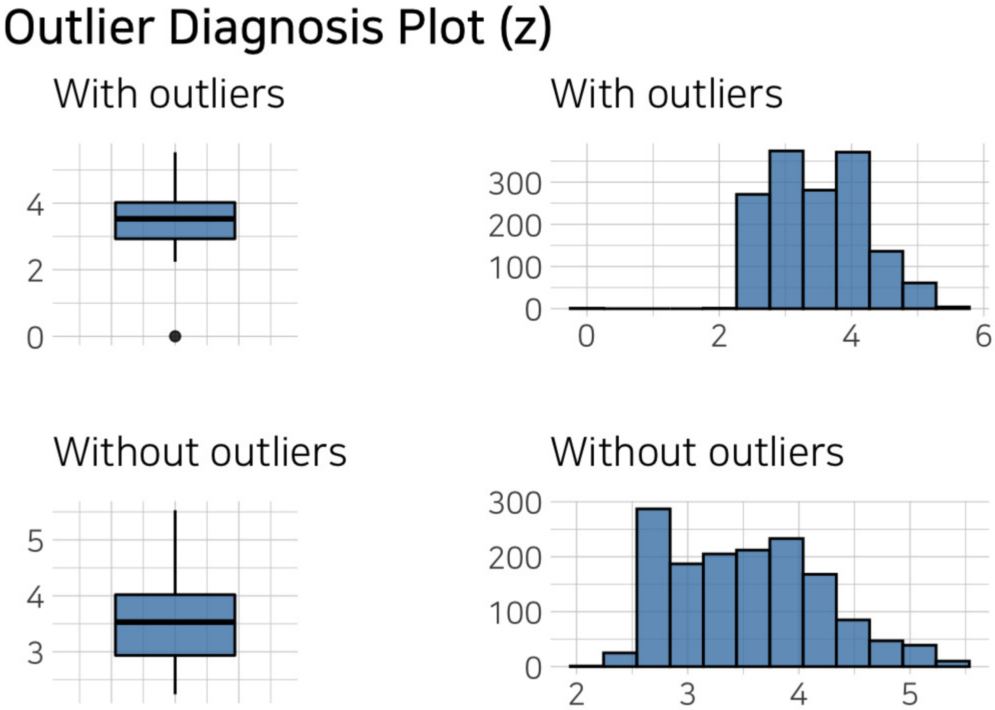
Table 10: table



variable: z

Measures	Values
Outliers count	1
Outliers ratio (%)	0.07%
Mean of outliers	0
Mean with outliers	3.525547
Mean without outliers	3.527899

Table 10: z


Outlier Diagnosis Plot (z)
With outliers
With outliers
Without outliers
Without outliers