데이터 진단 보고서
HEALTH_SCREEN

보고서 개요

이 보고서는 health_screen의 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	1,000
size	variables	34
size	values	34,000
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	83
missing	missing observation	917
missing	missing variables	16
missing	missing values	4,247

division	metrics	value
data type	numerics	31
data type	integers	0
data type	factors/ordered	0
data type	characters	3
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	1,000 / 1,000 (100%)
job	created	2021-10-06 22:45:39
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
4	44	5

Table 3: Summary of warnings

warnings	status	recommand
LDL 콜레스테롤 has 668 (66.8%) missing values	missing	judgement
총 콜레스테롤 has 660 (66%) missing values	missing	judgement
트리글리세라이드 has 660 (66%) missing values	missing	judgement
HDL 콜레스테롤 has 660 (66%) missing values	missing	judgement
치아우식증유무 has 582 (58.2%) missing values	missing	judgement
치석 has 582 (58.2%) missing values	missing	judgement
음주여부 has 349 (34.9%) missing values	missing	judgement
요단백 has 14 (1.4%) missing values	missing	judgement
수축기 혈압 has 9 (0.9%) missing values	missing	judgement
이완기 혈압 has 9 (0.9%) missing values	missing	judgement
식전혈당(공복혈당) has 9 (0.9%) missing values	missing	judgement
혈색소 has 9 (0.9%) missing values	missing	judgement
혈청크레아티닌 has 9 (0.9%) missing values	missing	judgement
(혈청지오티)AST has 9 (0.9%) missing values	missing	judgement
(혈청지오티)ALT has 9 (0.9%) missing values	missing	judgement
감마 지티피 has 9 (0.9%) missing values	missing	judgement
가입자 일련번호 has high(1.00) cardinality, Maybe identifier	cardinality	check
기준년도 has constant value “2019”	cardinality	remove
결손치 유무 has constant value “미시행”	cardinality	remove
치아마모증유무 has constant value “미시행”	cardinality	remove
제3대구치(사랑니) 이상 has constant value “미시행”	cardinality	remove
데이터 공개일자 has constant value “20191231”	cardinality	remove

Table 4: Warnings in dataset and variables

	warnings	status	recommand
23	성별코드 has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
24	청력(좌) has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
25	청력(우) has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
26	흡연상태 has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
27	음주여부 has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
28	구강검진 수검여부 has a low cardinality. 2 (0.2%) distinct values	cardinality	judgement
29	치아우식증유무 has a low cardinality. 3 (0.3%) distinct values	cardinality	judgement
30	치석 has a low cardinality. 4 (0.4%) distinct values	cardinality	judgement
31	구강검진 수검여부 has 582 (58.2%) zeros	zero	check
32	치아우식증유무 has 319 (31.9%) zeros	zero	check
33	치석 has 176 (17.6%) zeros	zero	check
34	치아우식증유무 has 99 (9.9%) outliers	outlier	judgement
35	감마 지티피 has 72 (7.2%) outliers	outlier	judgement
36	식전혈당(공복혈당) has 68 (6.8%) outliers	outlier	judgement
37	(혈청지오티)ALT has 58 (5.8%) outliers	outlier	judgement
38	요단백 has 53 (5.3%) outliers	outlier	judgement
39	(혈청지오티)AST has 52 (5.2%) outliers	outlier	judgement
40	청력(우) has 41 (4.1%) outliers	outlier	judgement
41	청력(좌) has 39 (3.9%) outliers	outlier	judgement
42	이완기 혈압 has 37 (3.7%) outliers	outlier	judgement
43	체중(5Kg 단위) has 25 (2.5%) outliers	outlier	judgement
44	트리글리세라이드 has 18 (1.8%) outliers	outlier	judgement
45	혈청크레아티닌 has 12 (1.2%) outliers	outlier	judgement
46	시력(좌) has 11 (1.1%) outliers	outlier	judgement
47	수축기 혈압 has 11 (1.1%) outliers	outlier	judgement
48	혈색소 has 10 (1%) outliers	outlier	judgement
49	시력(우) has 8 (0.8%) outliers	outlier	judgement
50	허리둘레 has 4 (0.4%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables (continued)

	warnings	status	recommand
51	HDL 콜레스테롤 has 4 (0.4%) outliers	outlier	judgement
52	총 콜레스테롤 has 3 (0.3%) outliers	outlier	judgement
53	LDL 콜레스테롤 has 3 (0.3%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables (continued)

Variables

variables	types	missing	cardinality	zero	minus	outlier
기준년도	numeric		constant			
가입자 일련번호	numeric		identifier			
시도코드	numeric					
성별코드	numeric		< low			
연령대 코드(5세단위)	numeric					
신장(5Cm단위)	numeric					
체중(5Kg 단위)	numeric					X
허리둘레	numeric					X
시력(좌)	numeric					X
시력(우)	numeric					X
청력(좌)	numeric		< low			X
청력(우)	numeric		< low			X
수축기 혈압	numeric	X				X
이완기 혈압	numeric	X				X
식전혈당(공복혈당)	numeric	X				X
총 콜레스테롤	numeric	X				X
트리글리세라이드	numeric	X				X
HDL 콜레스테롤	numeric	X				X
LDL 콜레스테롤	numeric	X				X
혈색소	numeric	X				X
요단백	numeric	X				X
혈청크레아티닌	numeric	X				X
(혈청지오티)AST	numeric	X				X
(혈청지오티)ALT	numeric	X				X
감마 지티피	numeric	X				X

Table 5: List of variables diagnosis

variables	types	missing	cardinality	zero	minus	outlier
흡연상태	numeric		< low			
음주여부	numeric	X	< low			
구강검진 수검여부	numeric		< low	X		
치아우식증유무	numeric	X	< low	X		X
결손치 유무	character		constant			
치아마모증유무	character		constant			
제3대구치(사랑니) 이상	character		constant			
치석	numeric	X	< low	X		
데이터 공개일자	numeric		constant			

Table 5: List of variables diagnosis (continued)

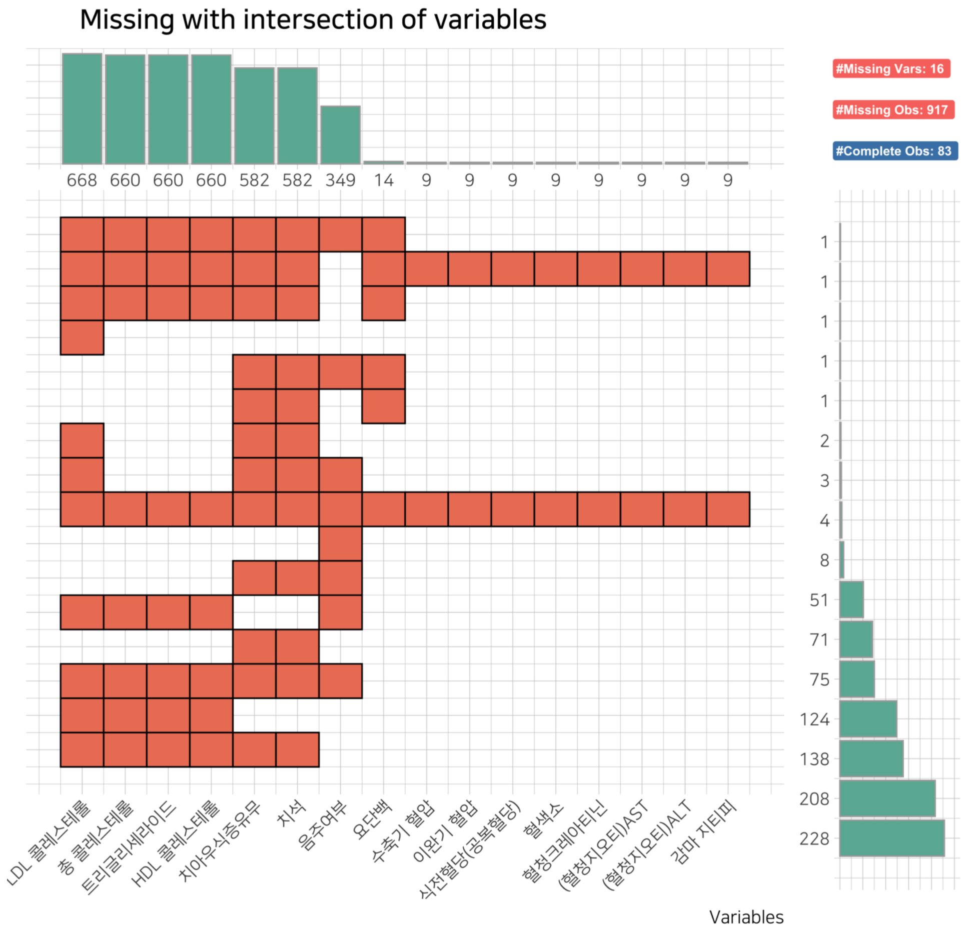
Missing Values

List of Missing Values

variables	missing_count	missing (%)	status	recommand
LDL 콜레스테롤	668	66.8%	Remove	Remove Variable
총 콜레스테롤	660	66%	Remove	Remove Variable
트리글리세라이드	660	66%	Remove	Remove Variable
HDL 콜레스테롤	660	66%	Remove	Remove Variable
치아우식증유무	582	58.2%	Remove	Remove Variable
치석	582	58.2%	Remove	Remove Variable
음주여부	349	34.9%	Bad	Model based Imputation
요단백	14	1.4%	Good	Delete or Imputation
수축기 혈압	9	0.9%	Good	Delete or Imputation
이완기 혈압	9	0.9%	Good	Delete or Imputation
식전혈당(공복혈당)	9	0.9%	Good	Delete or Imputation
혈색소	9	0.9%	Good	Delete or Imputation
혈청크레아티닌	9	0.9%	Good	Delete or Imputation
(혈청지오티)AST	9	0.9%	Good	Delete or Imputation
(혈청지오티)ALT	9	0.9%	Good	Delete or Imputation
감마 지티피	9	0.9%	Good	Delete or Imputation

Table 6: List of variables including missing values

Visualization



Unique Values

Categorical Vaiables

Variables where the proportion of unique data is more than 0.5 or unique is 1.

variables	types	unique	unique (%)	status	recommand
결손치 유무	character	1	0.1%	constant	Remove Variable
치아마모증유무	character	1	0.1%	constant	Remove Variable
제3대구치(사랑니) 이상	character	1	0.1%	constant	Remove Variable

Table 7: Detail warning categorical cardinality

Numerical Vaiables

Variables where the unique cases is less than 5 or unique is 1.

variables	types	unique	unique (%)	status	recommand
기준년도	numeric	1	0.1%	constant	Remove Variable
성별코드	numeric	2	0.2%	low cardinality	Judgment
청력(좌)	numeric	2	0.2%	low cardinality	Judgment
청력(우)	numeric	2	0.2%	low cardinality	Judgment
흡연상태	numeric	2	0.2%	low cardinality	Judgment
음주여부	numeric	2	0.2%	low cardinality	Judgment
구강검진 수검여부	numeric	2	0.2%	low cardinality	Judgment
치아우식증유무	numeric	3	0.3%	low cardinality	Judgment
치석	numeric	4	0.4%	low cardinality	Judgment
데이터 공개일자	numeric	1	0.1%	constant	Remove Variable

Table 8: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
결손치 유무	미시행	1,000	100
제3대구치(사랑니) 이상	미시행	1,000	100
치아마모증유무	미시행	1,000	100

Table 9: Top 10 levels of categorical variables

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
기준년도	2,019.0	2,019.00	2,019.00	2,019.0	2,019.00	2,019.0	0	0	0
가입자 일련번호	1.0	250.75	500.50	500.5	750.25	1,000.0	0	0	0
시도코드	11.0	27.00	33.50	41.0	43.00	50.0	0	0	0
성별코드	1.0	1.00	1.50	1.0	2.00	2.0	0	0	0
연령대 코드(5세단위)	5.0	8.00	10.40	10.0	13.00	18.0	0	0	0
신장(5Cm단위)	135.0	155.00	162.16	160.0	170.00	185.0	0	0	0
체중(5Kg 단위)	40.0	55.00	63.13	60.0	70.00	125.0	0	0	25
허리둘레	56.2	73.45	81.04	81.0	88.00	124.0	0	0	4
시력(좌)	0.1	0.70	0.94	0.9	1.20	2.0	0	0	11
시력(우)	0.1	0.70	0.96	1.0	1.20	9.9	0	0	8
청력(좌)	1.0	1.00	1.04	1.0	1.00	2.0	0	0	39
청력(우)	1.0	1.00	1.04	1.0	1.00	2.0	0	0	41
수축기 혈압	80.0	111.00	122.90	120.0	132.00	210.0	0	0	11
이완기 혈압	46.0	70.00	76.12	76.0	81.00	130.0	0	0	37
식전혈당(공복혈당)	63.0	89.00	100.85	97.0	105.00	322.0	0	0	68
총 콜레스테롤	100.0	166.75	191.57	190.0	218.00	333.0	0	0	3
트리글리세라이드	19.0	73.00	129.25	108.5	160.25	515.0	0	0	18
HDL 콜레스테롤	30.0	46.00	56.42	54.0	65.00	116.0	0	0	4
LDL 콜레스테롤	17.0	84.75	109.41	107.0	134.25	218.0	0	0	3
혈색소	7.3	13.05	14.20	14.3	15.40	19.5	0	0	10
요단백	1.0	1.00	1.10	1.0	1.00	6.0	0	0	53
혈청크레아티닌	0.2	0.70	0.86	0.8	1.00	7.0	0	0	12
(혈청지오티)AST	10.0	19.00	25.22	23.0	28.00	105.0	0	0	52
(혈청지오티)ALT	5.0	15.00	25.29	20.0	30.00	206.0	0	0	58
감마 지티피	6.0	15.00	36.27	23.0	40.00	746.0	0	0	72

Table 10: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
흡연상태	1	1	1.35	1	2	2	0	0	0
음주여부	1	1	1.00	1	1	1	0	0	0
구강검진 수검여부	0	0	0.42	0	1	1	582	0	0
치아우식 증유무	0	0	0.24	0	0	1	319	0	99
치석	0	0	0.64	1	1	2	176	0	0
데이터 공 개일자	20,191,231	20,191,231	20,191,231.00	20,191,231	20,191,231	20,191,231	0	0	0

Table 10: General list of numerical diagnosis (continued)

Zero Values

variables	min	median	max	zero	zero (%)
구강검진 수검여부	0	0	1	582	58.2
치아우식증유무	0	0	1	319	31.9
치석	0	1	2	176	17.6

Table 11: List of numerical diagnosis (zero)

Negative Values

No numeric variable with negative value

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
치아우식증유무	0.0	0.0	1.0	99	9.9
감마 지티피	6.0	23.0	746.0	72	7.2
식전혈당(공복혈당)	63.0	97.0	322.0	68	6.8
(혈청지오티)ALT	5.0	20.0	206.0	58	5.8
요단백	1.0	1.0	6.0	53	5.3
(혈청지오티)AST	10.0	23.0	105.0	52	5.2
청력(우)	1.0	1.0	2.0	41	4.1
청력(좌)	1.0	1.0	2.0	39	3.9
이완기 혈압	46.0	76.0	130.0	37	3.7
체중(5Kg 단위)	40.0	60.0	125.0	25	2.5
트리글리세라이드	19.0	108.5	515.0	18	1.8
혈청크레아티닌	0.2	0.8	7.0	12	1.2
시력(좌)	0.1	0.9	2.0	11	1.1
수축기 혈압	80.0	120.0	210.0	11	1.1
혈색소	7.3	14.3	19.5	10	1.0
시력(우)	0.1	1.0	9.9	8	0.8
허리둘레	56.2	81.0	124.0	4	0.4
HDL 콜레스테롤	30.0	54.0	116.0	4	0.4
총 콜레스테롤	100.0	190.0	333.0	3	0.3
LDL 콜레스테롤	17.0	107.0	218.0	3	0.3

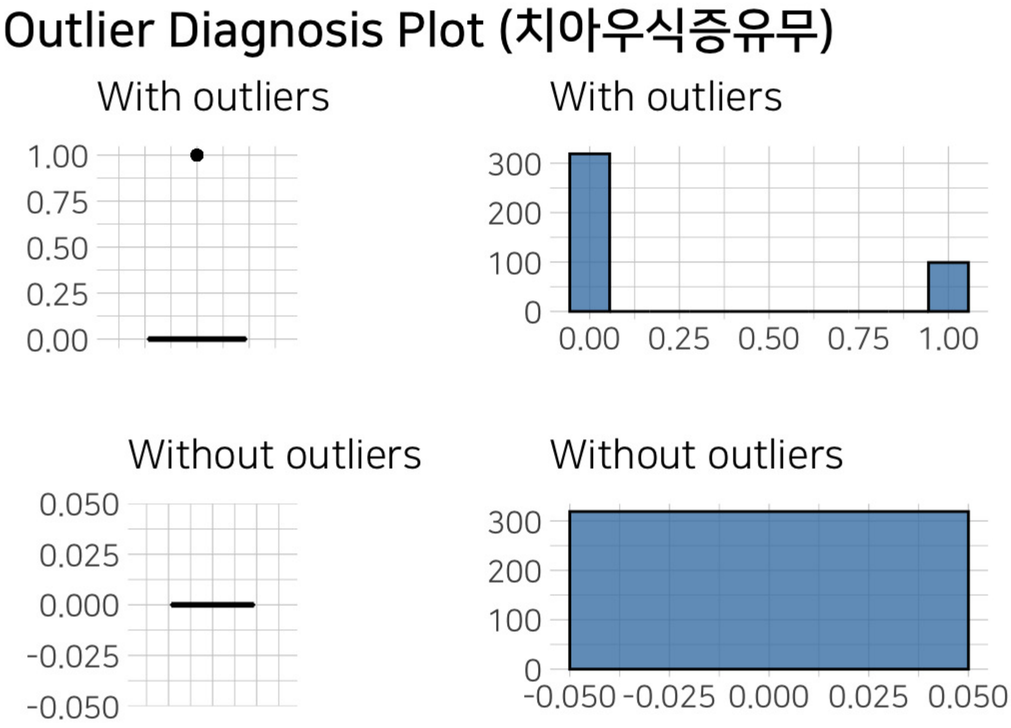
Table 12: Diagnosis of numerical variable outliers

Individual Outliers

variable: 치아우식증유무

Measures	Values
Outliers count	99
Outliers ratio (%)	9.9%
Mean of outliers	1
Mean with outliers	0.2368421
Mean without outliers	0

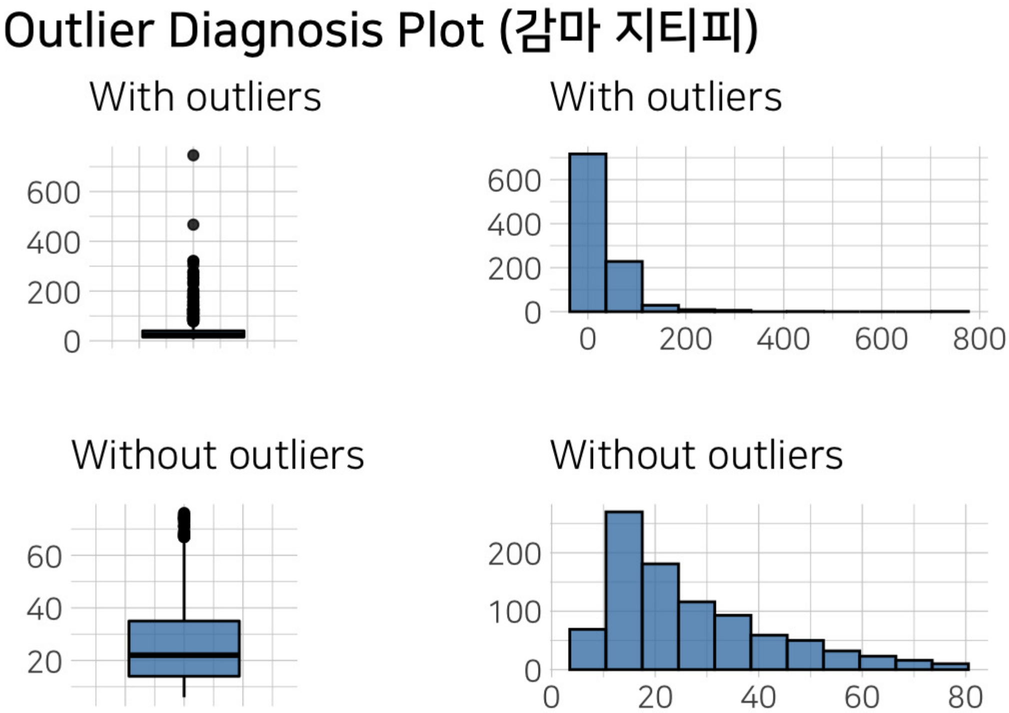
Table 13: 치아우식증유무



variable: 감마 지티피

Measures	Values
Outliers count	72
Outliers ratio (%)	7.2%
Mean of outliers	159.8056
Mean with outliers	36.27245
Mean without outliers	26.59412

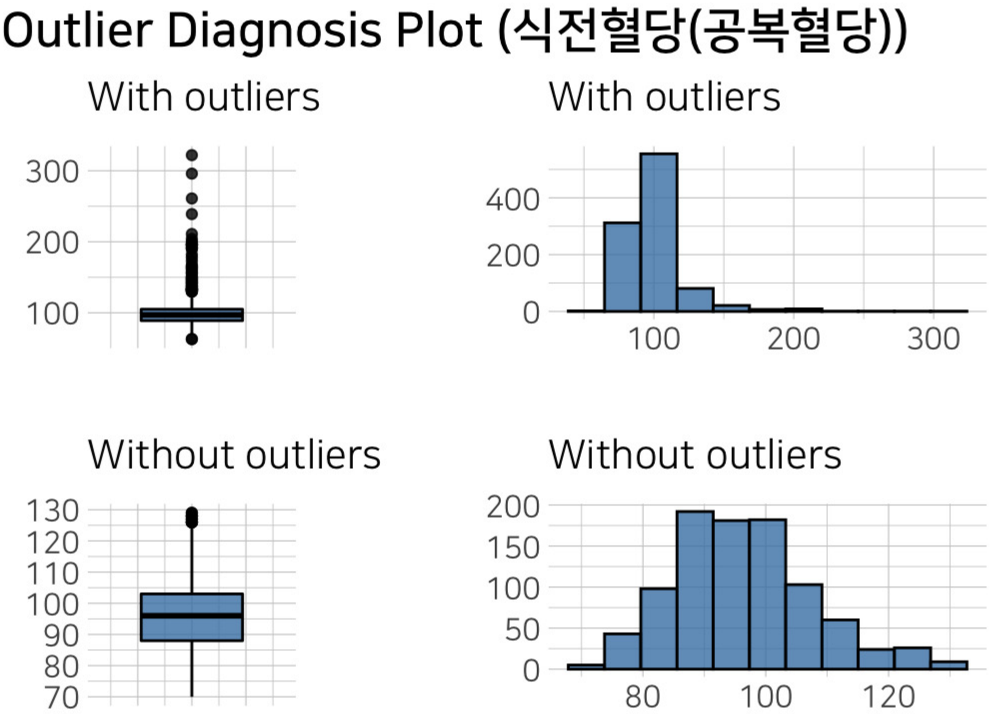
Table 13: 감마 지티피



variable: 식전혈당(공복혈당)

Measures	Values
Outliers count	68
Outliers ratio (%)	6.8%
Mean of outliers	160.8824
Mean with outliers	100.8476
Mean without outliers	96.4247

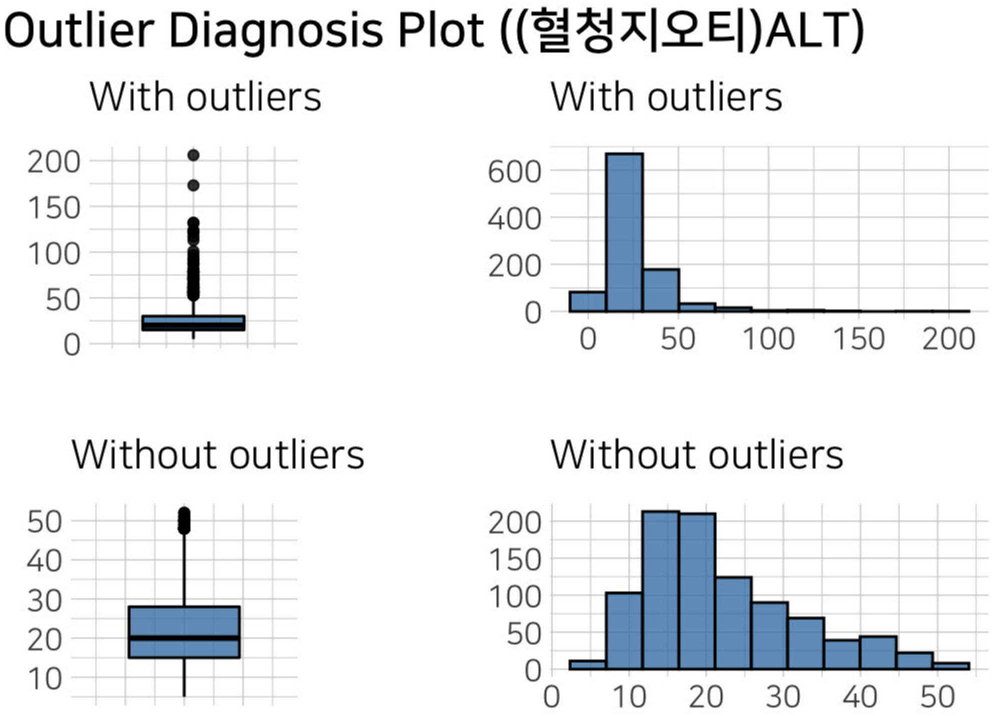
Table 13: 식전혈당(공복혈당)



variable: (혈청지오티)ALT

Measures	Values
Outliers count	58
Outliers ratio (%)	5.8%
Mean of outliers	79
Mean with outliers	25.29162
Mean without outliers	21.95284

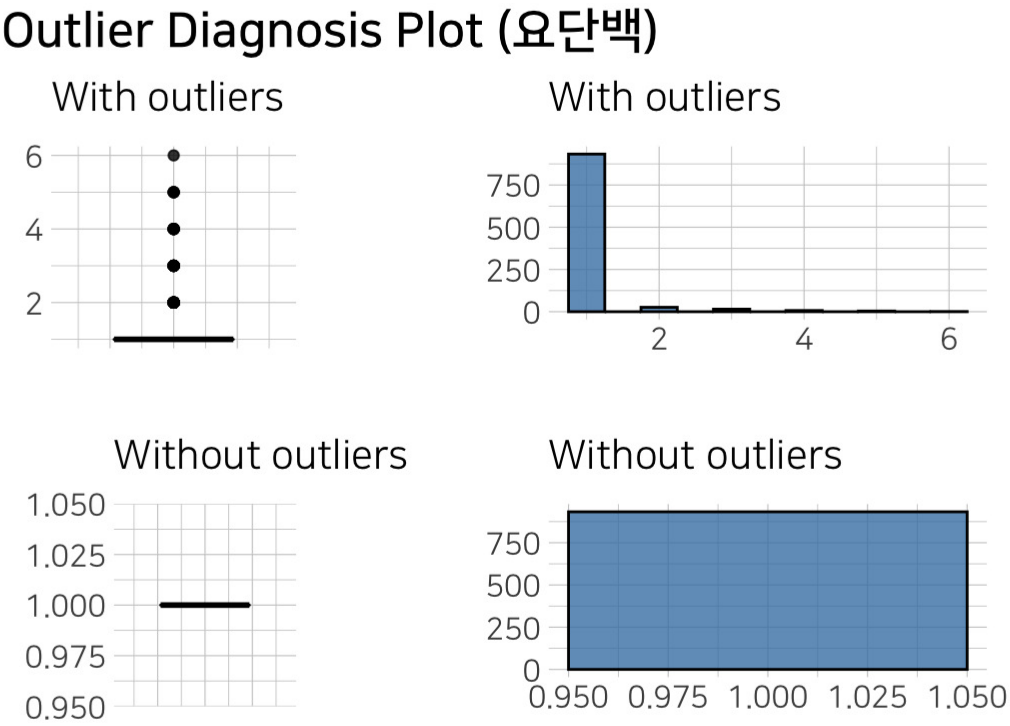
Table 13: (혈청지오티)ALT



variable: 요단백

Measures	Values
Outliers count	53
Outliers ratio (%)	5.3%
Mean of outliers	2.849057
Mean with outliers	1.099391
Mean without outliers	1

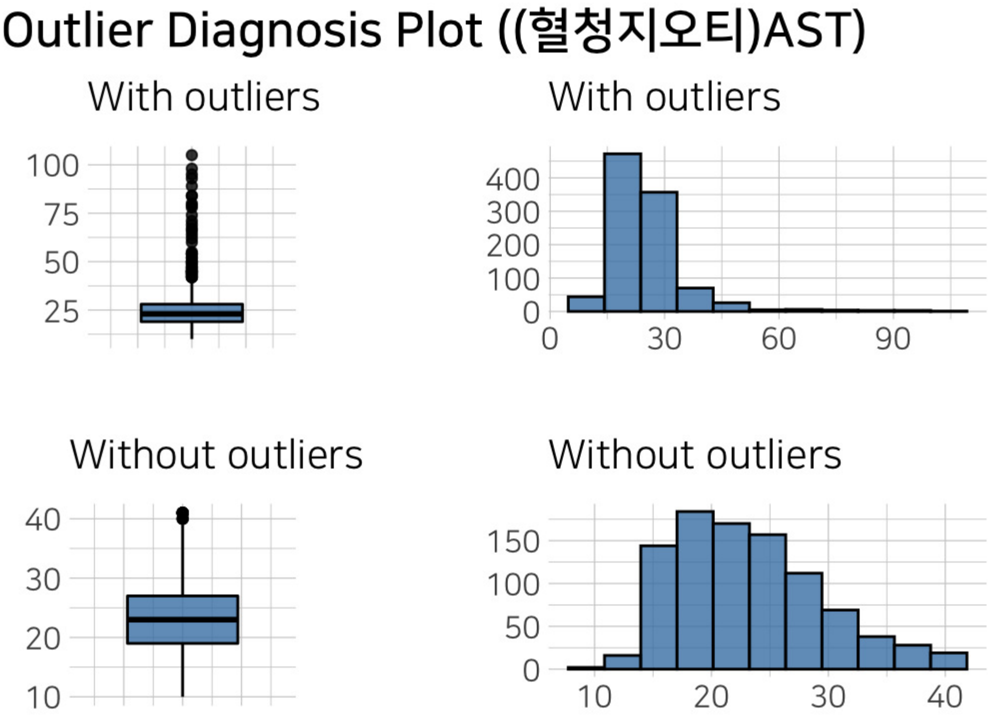
Table 13: 요단백



variable: (혈청지오티)AST

Measures	Values
Outliers count	52
Outliers ratio (%)	5.2%
Mean of outliers	57.90385
Mean with outliers	25.22301
Mean without outliers	23.41321

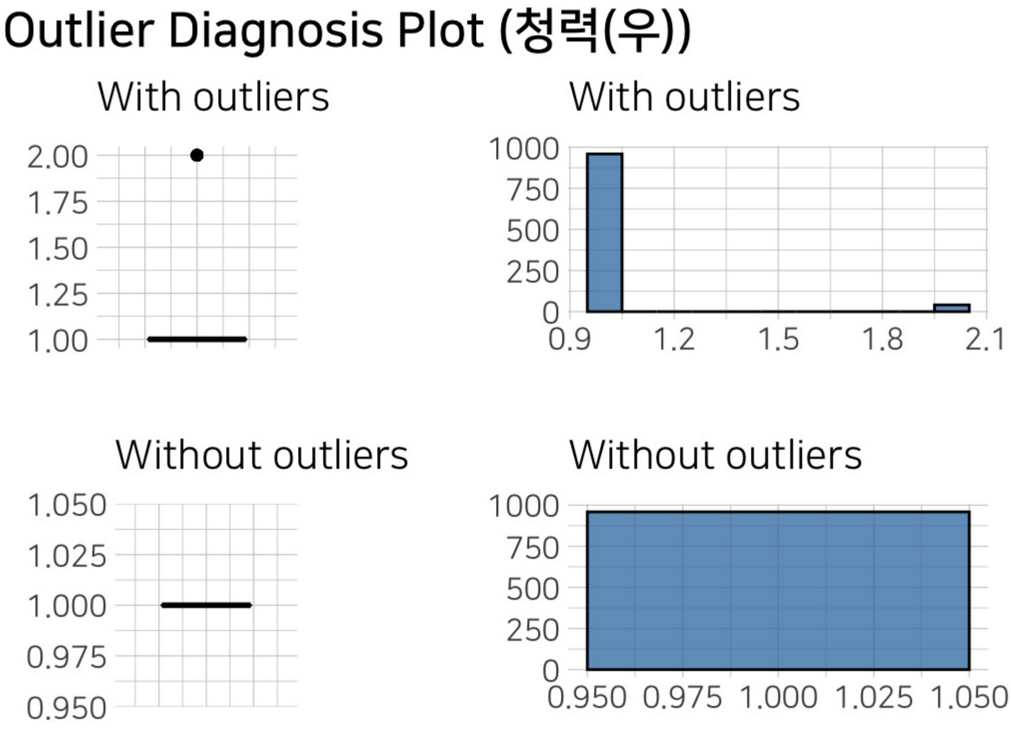
Table 13: (혈청지오티)AST



variable: 청력(우)

Measures	Values
Outliers count	41
Outliers ratio (%)	4.1%
Mean of outliers	2
Mean with outliers	1.041
Mean without outliers	1

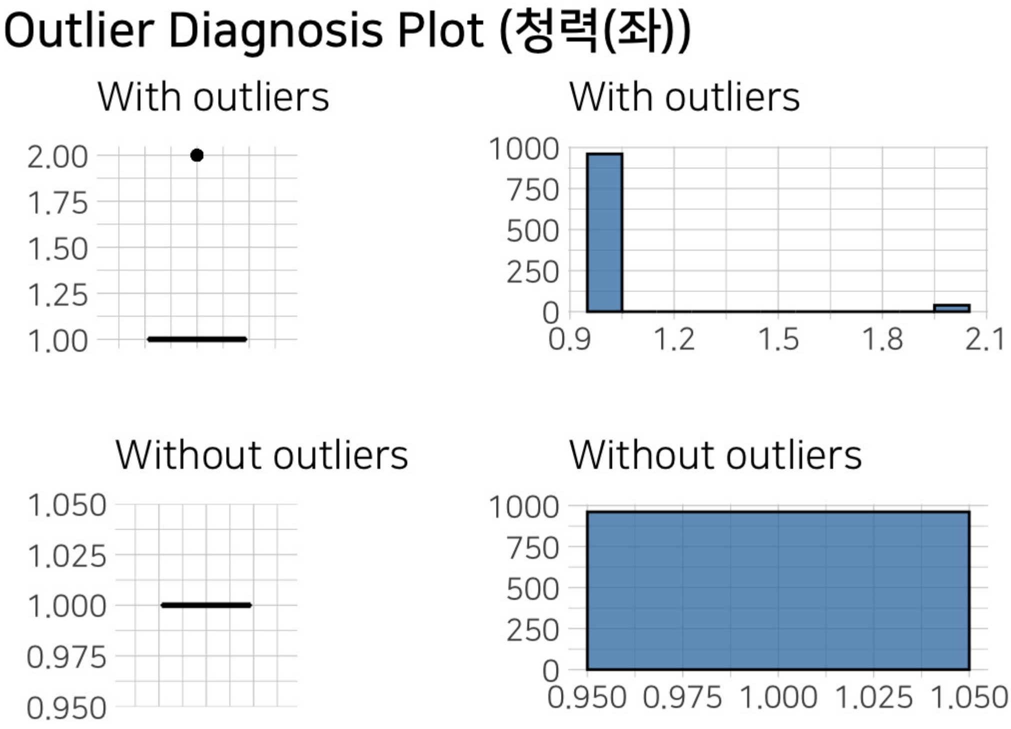
Table 13: 청력(우)



variable: 청력(좌)

Measures	Values
Outliers count	39
Outliers ratio (%)	3.9%
Mean of outliers	2
Mean with outliers	1.039
Mean without outliers	1

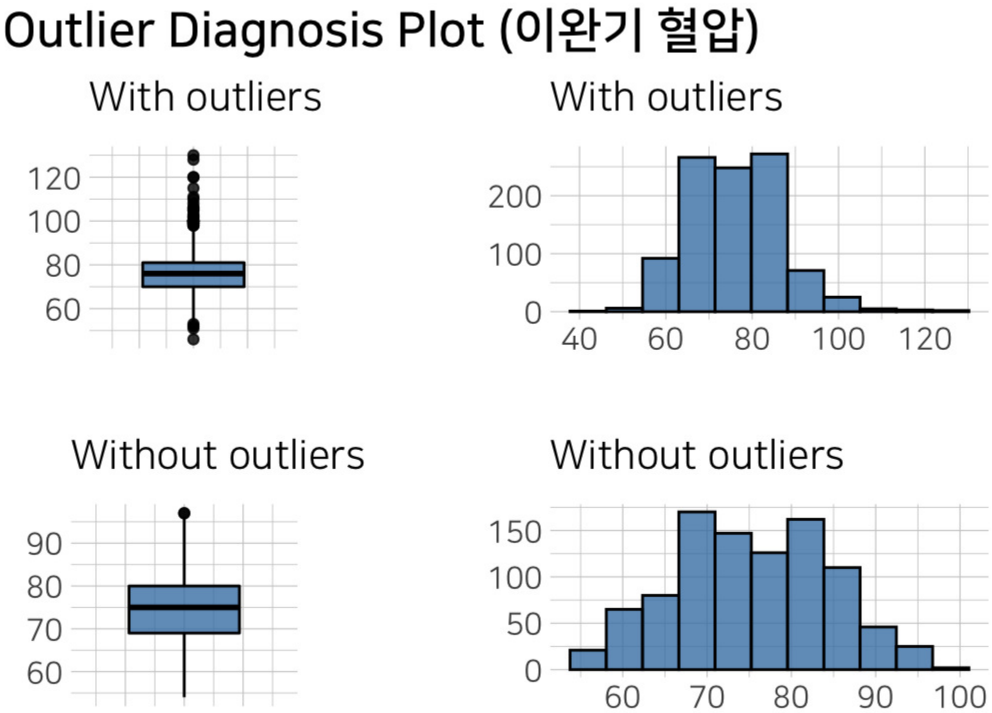
Table 13: 청력(좌)



variable: 이완기 혈압

Measures	Values
Outliers count	37
Outliers ratio (%)	3.7%
Mean of outliers	99
Mean with outliers	76.12008
Mean without outliers	75.2327

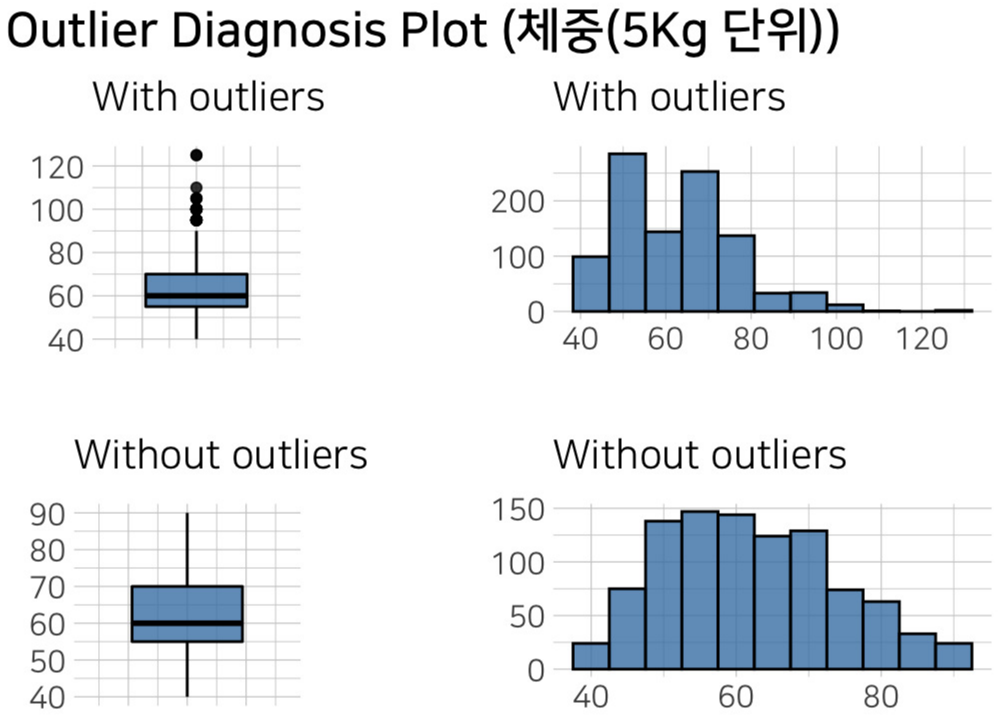
Table 13: 이완기 혈압



variable: 체중(5Kg 단위)

Measures	Values
Outliers count	25
Outliers ratio (%)	2.5%
Mean of outliers	101
Mean with outliers	63.13
Mean without outliers	62.15897

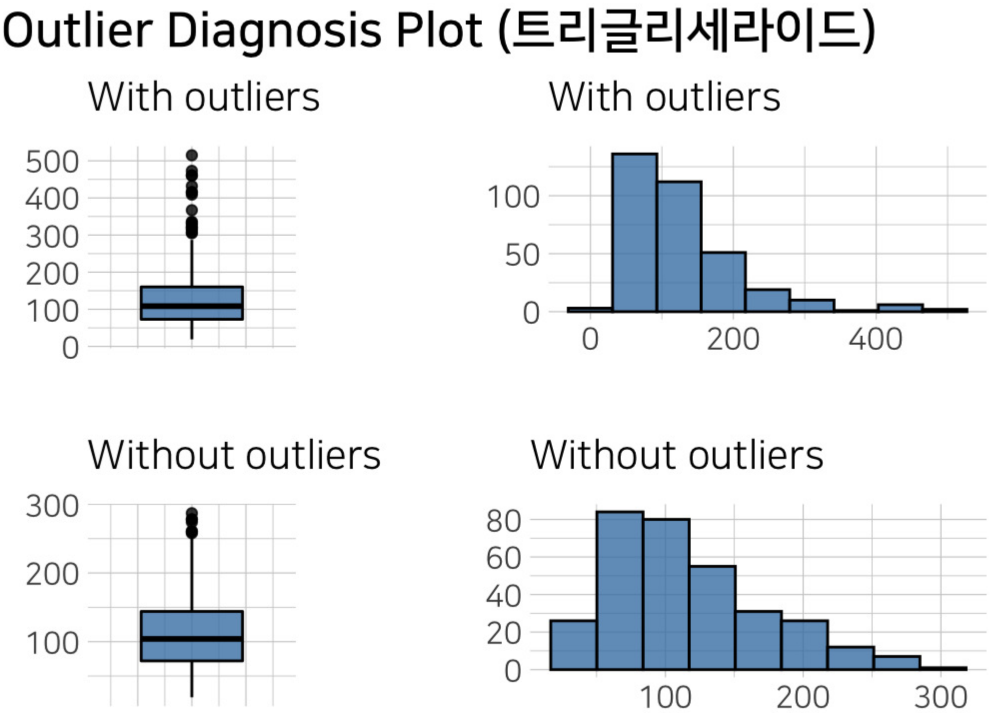
Table 13: 체중(5Kg 단위)



variable: 트리글리세라이드

Measures	Values
Outliers count	18
Outliers ratio (%)	1.8%
Mean of outliers	379.6111
Mean with outliers	129.2471
Mean without outliers	115.2516

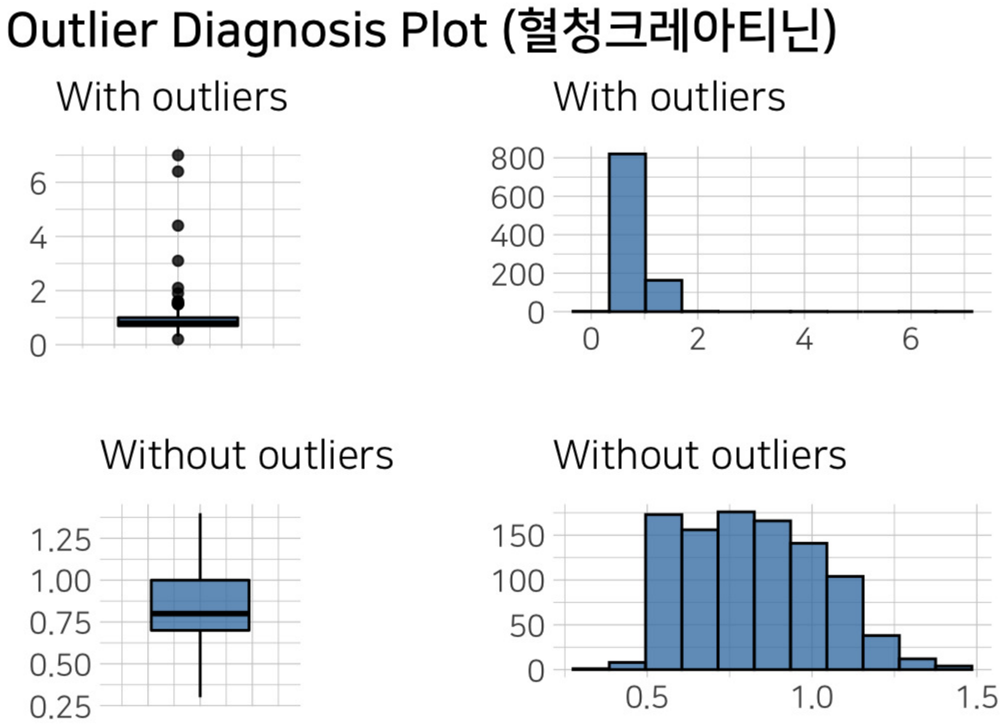
Table 13: 트리글리세라이드



variable: 혈청크레아티닌

Measures	Values
Outliers count	12
Outliers ratio (%)	1.2%
Mean of outliers	2.733333
Mean with outliers	0.864783
Mean without outliers	0.8418795

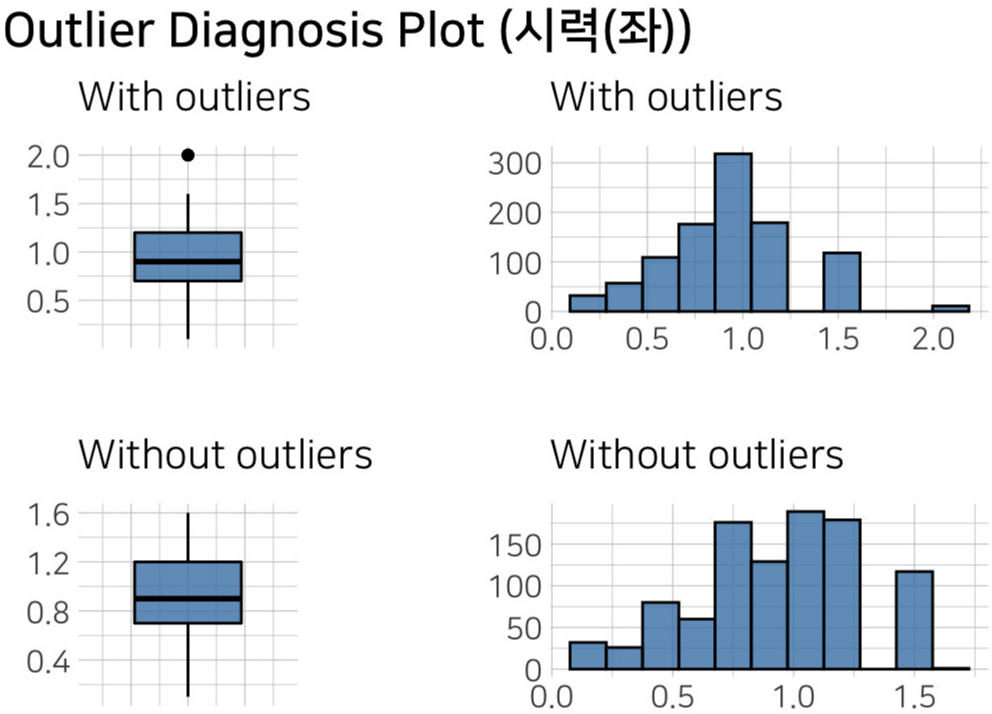
Table 13: 혈청크레아티닌



variable: 시력(좌)

Measures	Values
Outliers count	11
Outliers ratio (%)	1.1%
Mean of outliers	2
Mean with outliers	0.9361
Mean without outliers	0.9242669

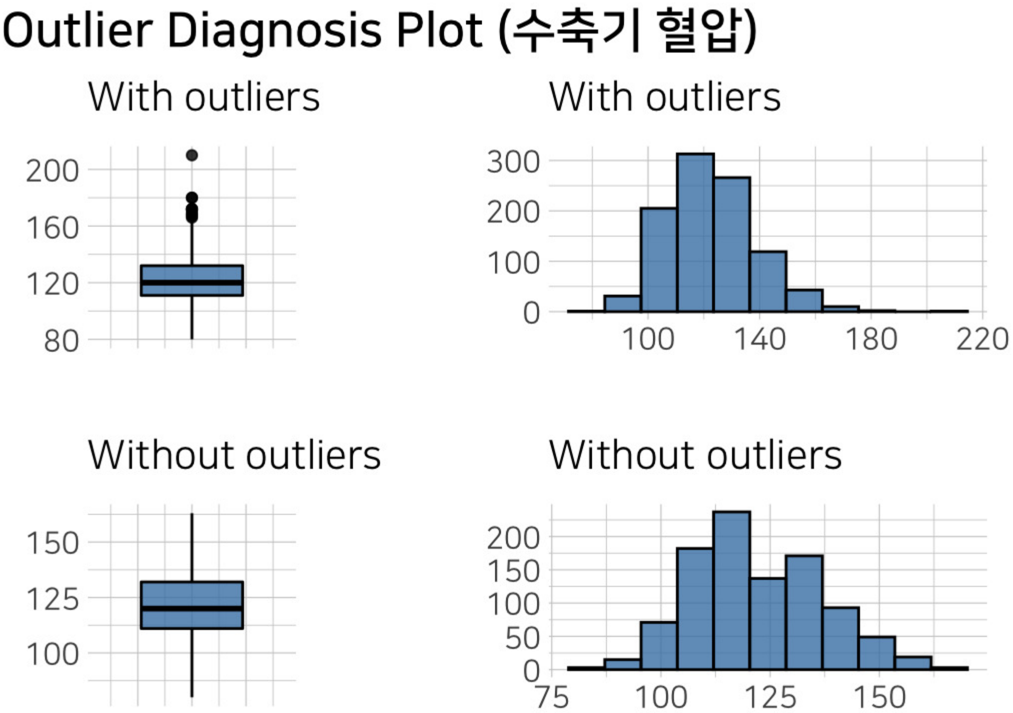
Table 13: 시력(좌)



variable: 수축기 혈압

Measures	Values
Outliers count	11
Outliers ratio (%)	1.1%
Mean of outliers	175.1818
Mean with outliers	122.9001
Mean without outliers	122.3133

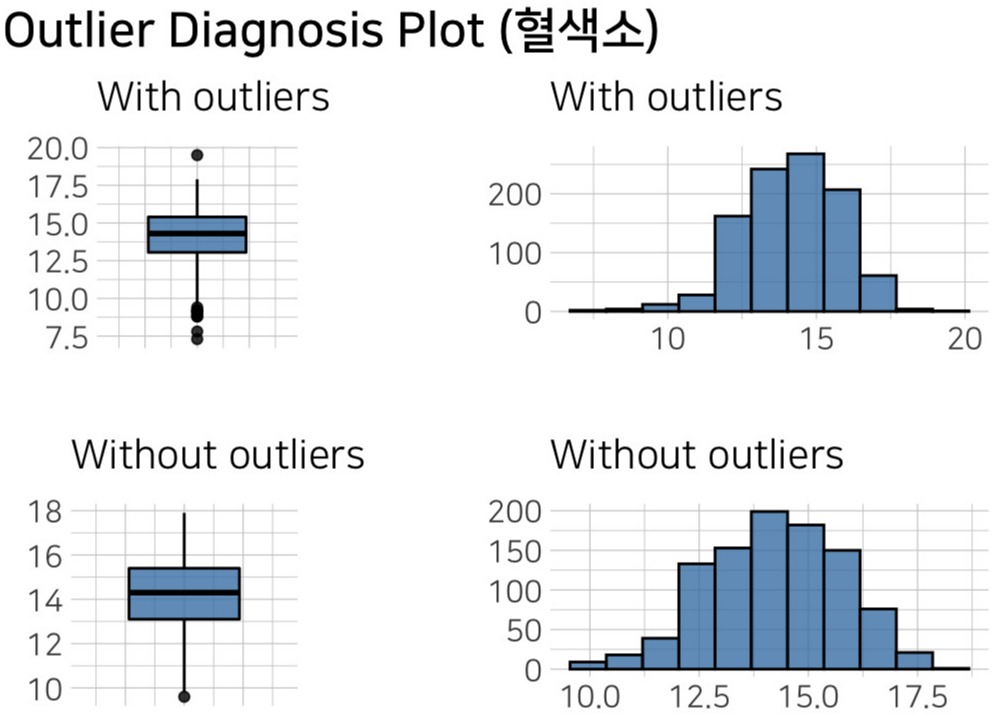
Table 13: 수축기 혈압



variable: 혈색소

Measures	Values
Outliers count	10
Outliers ratio (%)	1%
Mean of outliers	9.81
Mean with outliers	14.19828
Mean without outliers	14.24302

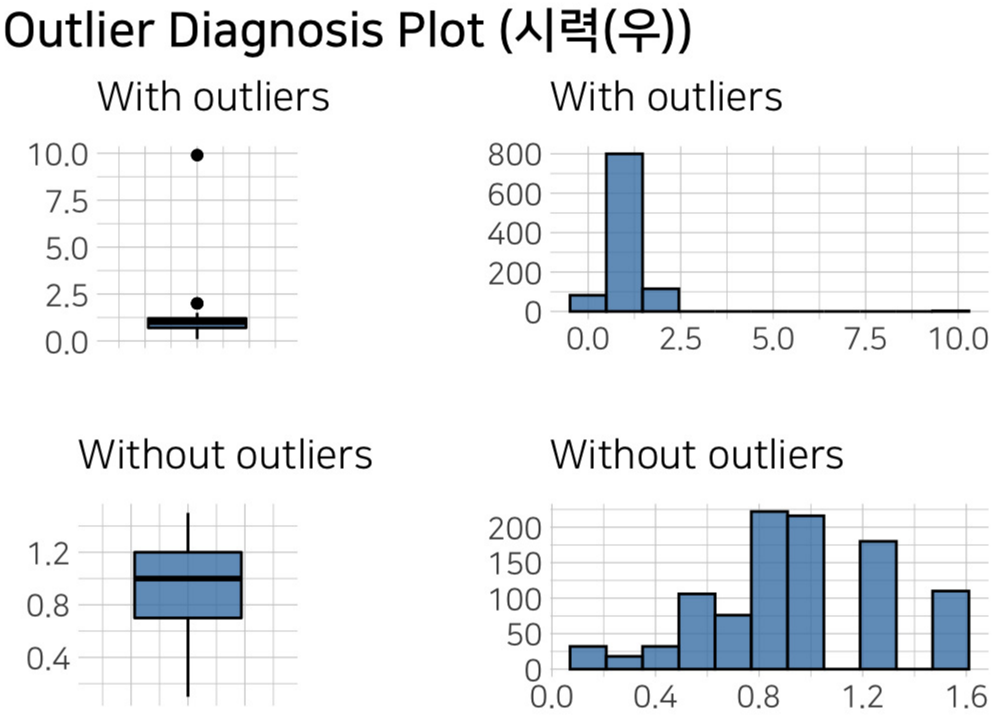
Table 13: 혈색소



variable: 시력(우)

Measures	Values
Outliers count	8
Outliers ratio (%)	0.8%
Mean of outliers	4.9625
Mean with outliers	0.9609
Mean without outliers	0.928629

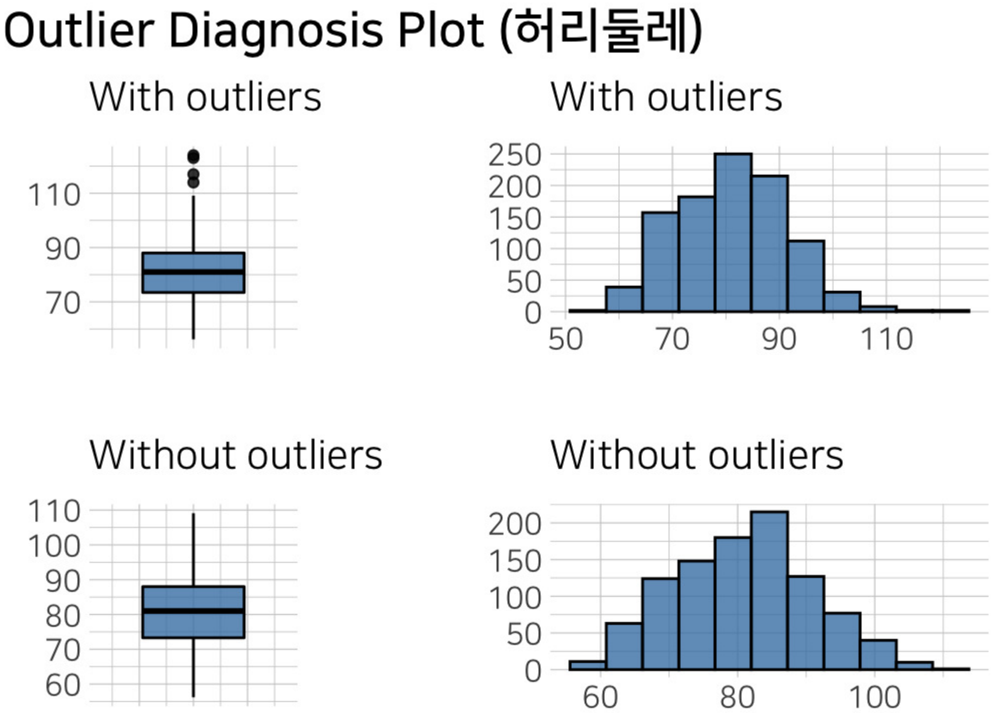
Table 13: 시력(우)



variable: 허리둘레

Measures	Values
Outliers count	4
Outliers ratio (%)	0.4%
Mean of outliers	119.5
Mean with outliers	81.0371
Mean without outliers	80.88263

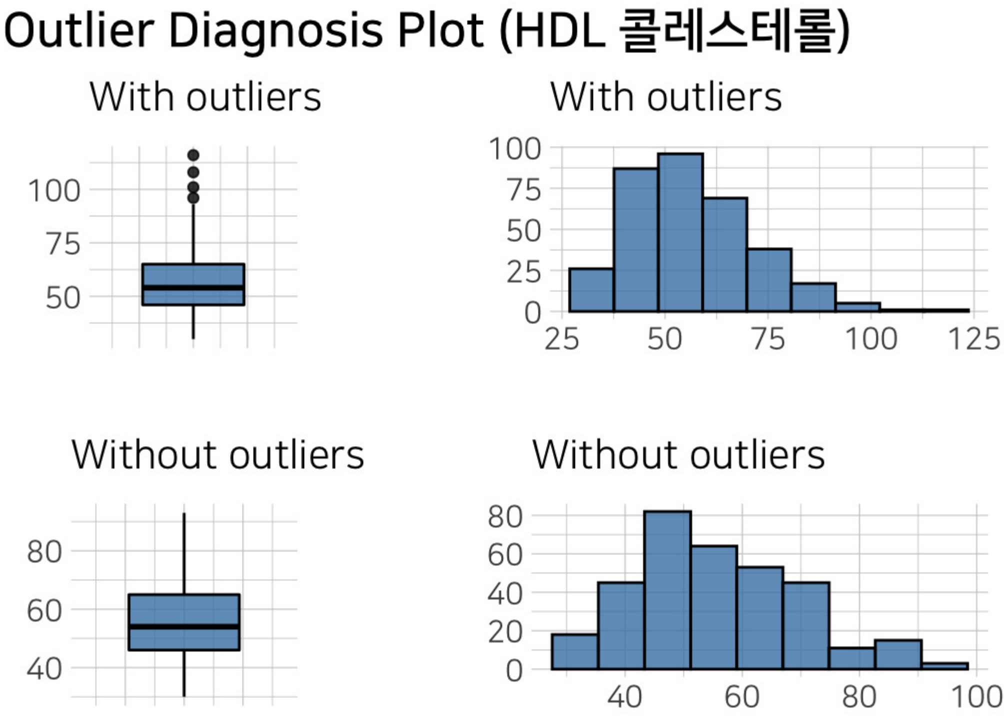
Table 13: 허리둘레



variable: HDL 콜레스테롤

Measures	Values
Outliers count	4
Outliers ratio (%)	0.4%
Mean of outliers	105.25
Mean with outliers	56.42059
Mean without outliers	55.83929

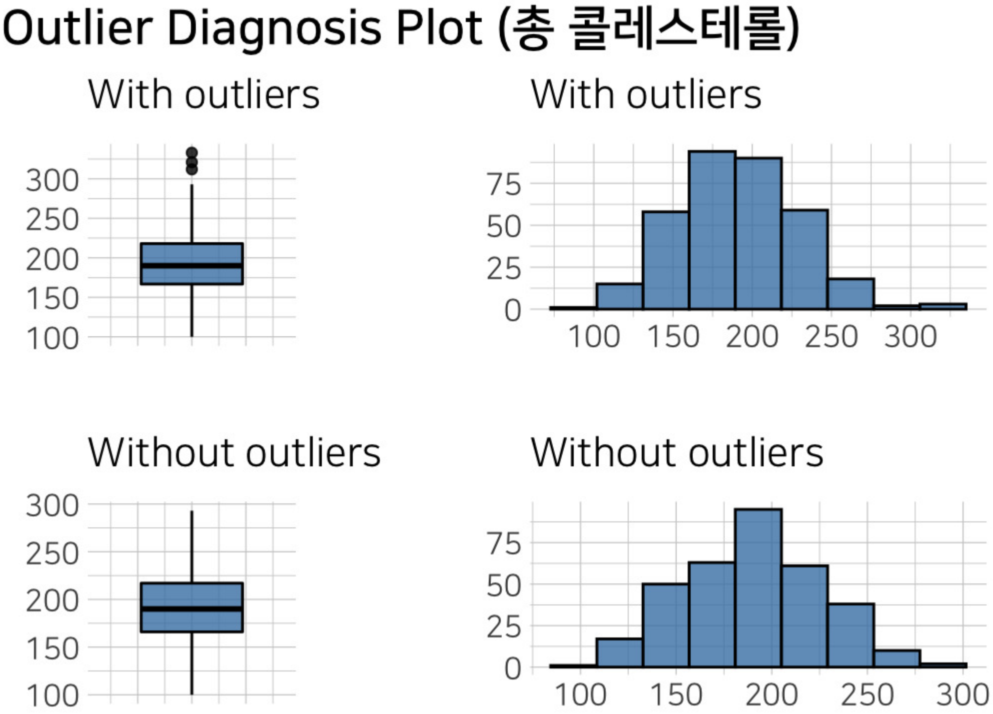
Table 13: HDL 콜레스테롤



variable: 총 콜레스테롤

Measures	Values
Outliers count	3
Outliers ratio (%)	0.3%
Mean of outliers	322
Mean with outliers	191.5706
Mean without outliers	190.4095

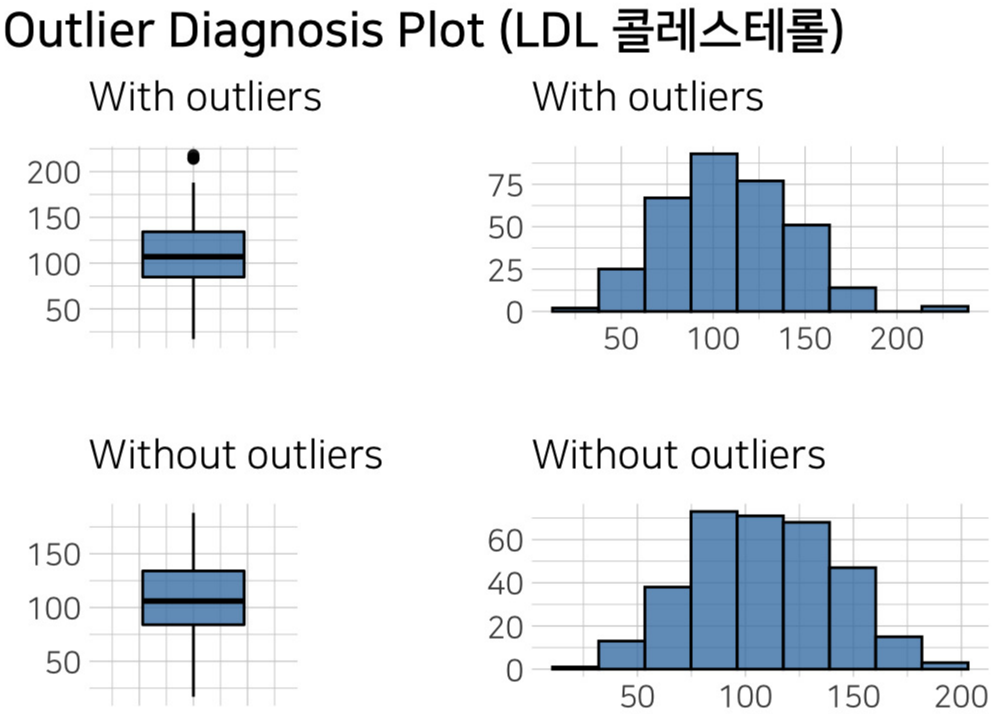
Table 13: 총 콜레스테롤



variable: LDL 콜레스테롤

Measures	Values
Outliers count	3
Outliers ratio (%)	0.3%
Mean of outliers	215.6667
Mean with outliers	109.4127
Mean without outliers	108.4438

Table 13: LDL 콜레스테롤


Outlier Diagnosis Plot (LDL 콜레스테롤)
With outliers
With outliers
Without outliers
Without outliers