DATA DIAGNOSIS REPORT
SHIP

Report Overview

This report was created for an overview quality diagnosis of . data. It was created for **the purpose of judging the validity of variables** before conducting EDA.

Contents

Overview

Data Structures

division	metrics	value
size	observations	2,154
size	variables	29
size	values	62,466
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	2,119
missing	missing observation	35
missing	missing variables	10
missing	missing values	85

division	metrics	value
data type	numerics	10
data type	integers	17
data type	factors/ordered	0
data type	characters	1
data type	Dates	1
data type	POSIXcts	0
data type	others	0

Table 1: Data structures and types

Job Informations

division	metrics	value
dataset	dataset	.
dataset	dataset type	data.frame
job	samples	2154 / 2,154 (100%)
job	created	2021-10-15 09:56:08
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
4	42	0

Table 3: Summary of warnings

warnings	status	recommand
ldl has 28 (1.3%) missing values	missing	judgement
hdl has 16 (0.7%) missing values	missing	judgement
cholesterol has 15 (0.7%) missing values	missing	judgement
sbp2 has 6 (0.3%) missing values	missing	judgement
dbp2 has 6 (0.3%) missing values	missing	judgement
weight has 4 (0.2%) missing values	missing	judgement
height has 3 (0.1%) missing values	missing	judgement
waist has 3 (0.1%) missing values	missing	judgement
sbp1 has 2 (0.1%) missing values	missing	judgement
dbp1 has 2 (0.1%) missing values	missing	judgement
id has high(1.00) cardinality, Maybe identifier	cardinality	check
sex has a low cardinality. 2 (0.1%) distinct values	cardinality	judgement
dev_length has a low cardinality. 4 (0.2%) distinct values	cardinality	judgement
dev_weight has a low cardinality. 4 (0.2%) distinct values	cardinality	judgement
smoking has a low cardinality. 5 (0.2%) distinct values	cardinality	judgement
myocard has a low cardinality. 5 (0.2%) distinct values	cardinality	judgement
stroke has a low cardinality. 5 (0.2%) distinct values	cardinality	judgement
diab_known has a low cardinality. 3 (0.1%) distinct values	cardinality	judgement
contraception has a low cardinality. 4 (0.2%) distinct values	cardinality	judgement
income has a low cardinality. 5 (0.2%) distinct values	cardinality	judgement
diab_known has 1,946 (90.34%) zeros	zero	check
school has 779 (36.17%) zeros	zero	check

Table 4: Warnings in dataset and variables

	warnings	status	recommand
23	smoking has 733 (34.03%) zeros	zero	check
24	diab_known has 208 (9.66%) outliers	outlier	judgement
25	diab_age has 180 (8.36%) outliers	outlier	judgement
26	myocard has 139 (6.45%) outliers	outlier	judgement
27	stroke has 126 (5.85%) outliers	outlier	judgement
28	income has 116 (5.39%) outliers	outlier	judgement
29	school has 113 (5.25%) outliers	outlier	judgement
30	smoking has 68 (3.16%) outliers	outlier	judgement
31	family has 67 (3.11%) outliers	outlier	judgement
32	hdl has 33 (1.53%) outliers	outlier	judgement
33	sbp2 has 24 (1.11%) outliers	outlier	judgement
34	sbp1 has 23 (1.07%) outliers	outlier	judgement
35	cholesterol has 23 (1.07%) outliers	outlier	judgement
36	ldl has 21 (0.97%) outliers	outlier	judgement
37	dbp1 has 17 (0.79%) outliers	outlier	judgement
38	dbp2 has 17 (0.79%) outliers	outlier	judgement
39	weight has 17 (0.79%) outliers	outlier	judgement
40	obs_bp has 5 (0.23%) outliers	outlier	judgement
41	dev_bp has 3 (0.14%) outliers	outlier	judgement
42	obs_soma has 2 (0.09%) outliers	outlier	judgement
43	obs_int has 2 (0.09%) outliers	outlier	judgement
44	dev_length has 2 (0.09%) outliers	outlier	judgement
45	dev_weight has 2 (0.09%) outliers	outlier	judgement
46	height has 1 (0.05%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables (continued)

Variables

variables	types	missing	cardinality	zero	minus	outlier
id	integer		identifier			
exdate	Date					
age	integer					
sex	integer		< low			
obs_bp	integer					X
obs_soma	integer					X
obs_int	integer					X
dev_bp	integer					X
dev_length	integer		< low			X
dev_weight	integer		< low			X
sbp1	integer	X				X
sbp2	integer	X				X
dbp1	integer	X				X
dbp2	integer	X				X
height	integer	X				X
weight	numeric	X				X
waist	character	X	> high			
cholesterol	numeric	X				X
hdl	numeric	X				X
ldl	numeric	X				X
school	numeric			X		X
family	numeric					X
smoking	numeric		< low	X		X
myocard	numeric		< low			X
stroke	numeric		< low			X

Table 5: List of variables diagnosis

variables	types	missing	cardinality	zero	minus	outlier
diab_known	integer		< low	X		X
diab_age	integer					X
contraception	numeric		< low			
income	integer		< low			X

Table 5: List of variables diagnosis (continued)

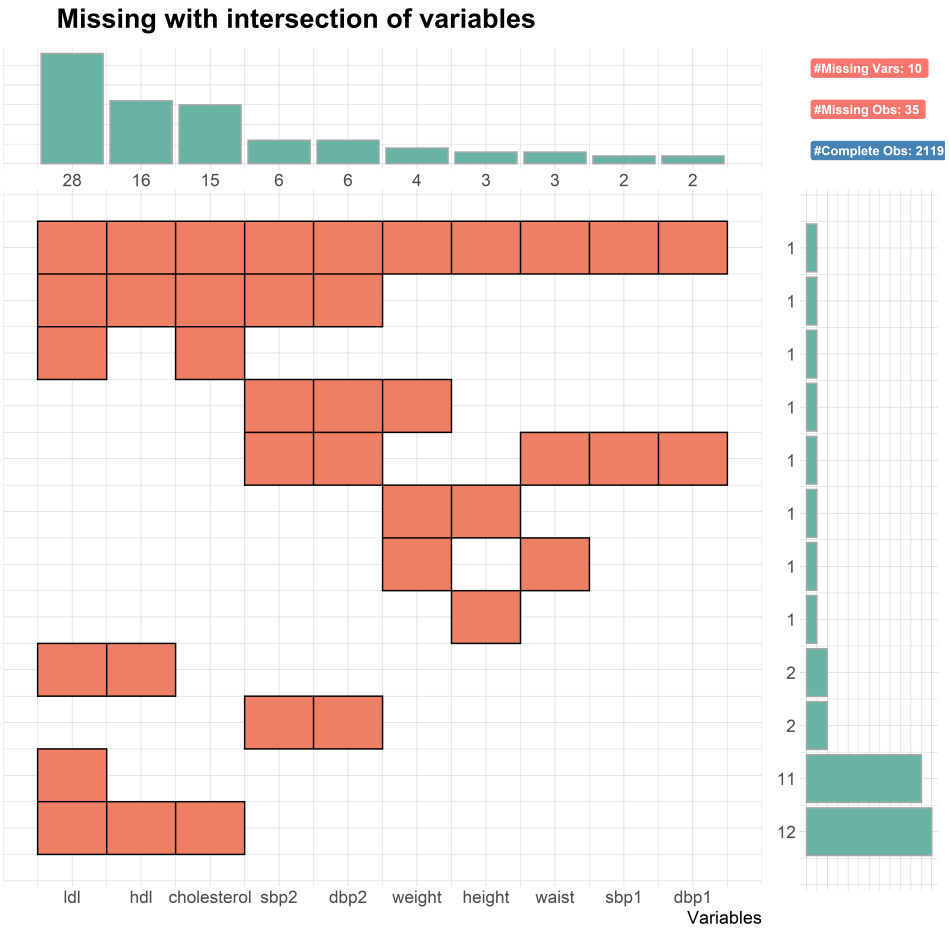
Missing Values

List of Missing Values

variables	missing_count	missing (%)	status	recommand
ldl	28	1.3%	Good	Delete or Imputation
hdl	16	0.7%	Good	Delete or Imputation
cholesterol	15	0.7%	Good	Delete or Imputation
sbp2	6	0.3%	Good	Delete or Imputation
dbp2	6	0.3%	Good	Delete or Imputation
weight	4	0.2%	Good	Delete or Imputation
height	3	0.1%	Good	Delete or Imputation
waist	3	0.1%	Good	Delete or Imputation
sbp1	2	0.1%	Good	Delete or Imputation
dbp1	2	0.1%	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
waist	character	2,106	97.8%	high cardinality	Judgment

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
sex	integer	2	0.1%	low cardinality	Judgment
dev_length	integer	4	0.2%	low cardinality	Judgment
dev_weight	integer	4	0.2%	low cardinality	Judgment
smoking	numeric	5	0.2%	low cardinality	Judgment
myocard	numeric	5	0.2%	low cardinality	Judgment
stroke	numeric	5	0.2%	low cardinality	Judgment
diab_known	integer	3	0.1%	low cardinality	Judgment
contraception	numeric	4	0.2%	low cardinality	Judgment
income	integer	5	0.2%	low cardinality	Judgment

Table 8: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
exdate	1999-03-26	9	0.4
exdate	1998-04-17	7	0.3
exdate	1998-05-19	7	0.3
exdate	1998-05-28	7	0.3
exdate	1998-06-08	7	0.3
exdate	1998-06-25	7	0.3
exdate	1998-09-23	7	0.3
exdate	1998-09-30	7	0.3
exdate	1998-10-10	7	0.3
exdate	1998-11-04	7	0.3
exdate	Other levles	2,082	96.7
waist	101.099	2	0.1
waist	101.585	2	0.1
waist	102.832	2	0.1
waist	103.387	2	0.1
waist	103.844	2	0.1
waist	104.654	2	0.1
waist	104.992	2	0.1
waist	105.206	2	0.1
waist	108.037	2	0.1
waist	Other levles	2,133	99.0
waist	Missing	3	0.1

Table 9: Top 10 levels of categorical variables

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
id	3,301.00	4,356.00	5,431.06	5,428.50	6,490.75	7,607.00	0	0	0
age	19.00	36.00	49.86	50.00	63.00	82.00	0	0	0
sex	1.00	1.00	1.54	2.00	2.00	2.00	0	0	0
obs_bp	1.00	4.00	144.69	4.00	7.00	99,902.00	0	0	5
obs_soma	1.00	4.00	98.40	5.00	7.00	99,902.00	0	0	2
obs_int	1.00	2.00	53.90	3.00	11.00	99,900.00	0	0	2
dev_bp	7.00	10.00	153.65	15.00	18.00	99,902.00	0	0	3
dev_length	3.00	3.00	99.19	3.00	11.00	99,902.00	0	0	2
dev_weight	1.00	1.00	97.97	1.00	11.00	99,902.00	0	0	2
sbp1	83.00	123.00	139.01	137.00	153.00	253.00	0	0	23
sbp2	83.00	120.00	136.40	134.00	150.00	258.00	0	0	24
dbp1	53.00	76.00	84.52	84.00	92.00	198.00	0	0	17
dbp2	55.00	75.00	83.52	83.00	91.00	151.00	0	0	17
height	144.00	161.00	168.22	168.00	175.00	198.00	0	0	1
weight	42.60	66.16	77.63	77.04	87.35	144.44	0	0	17
cholesterol	2.67	4.92	5.76	5.68	6.49	12.12	0	0	23
hdl	0.42	1.14	1.45	1.39	1.70	7.20	0	0	33
ldl	0.70	2.75	3.58	3.52	4.24	9.24	0	0	21
school	0.00	0.00	5,242.27	2.00	2.00	99,914.00	779	0	113
family	1.00	1.00	3,109.63	1.00	3.00	99,914.00	0	0	67
smoking	0.00	0.00	3,155.08	1.00	2.00	99,914.00	733	0	68
myocard	1.00	2.00	3,434.33	2.00	2.00	99,914.00	0	0	139
stroke	1.00	2.00	3,248.82	2.00	2.00	99,914.00	0	0	126
diab_known	0.00	0.00	324.75	0.00	0.00	99,900.00	1,946	0	208
diab_age	12.00	99,801.00	91,790.05	99,801.00	99,801.00	99,900.00	0	0	180

Table 10: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
contraception	1	1	58,750.81	99,801	99,801	99,900	0	0	0
income	1	1	5,381.81	2	3	99,901	0	0	116

Table 10: General list of numerical diagnosis (continued)

Zero Values

variables	min	median	max	zero	zero (%)
diab_known	0	0	99,900	1,946	90.3
school	0	2	99,914	779	36.2
smoking	0	1	99,914	733	34.0

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 (%)
diab_known	0.00	0.00	99,900.00	208	9.7
diab_age	12.00	99,801.00	99,900.00	180	8.4
myocard	1.00	2.00	99,914.00	139	6.5
stroke	1.00	2.00	99,914.00	126	5.8
income	1.00	2.00	99,901.00	116	5.4
school	0.00	2.00	99,914.00	113	5.2
smoking	0.00	1.00	99,914.00	68	3.2
family	1.00	1.00	99,914.00	67	3.1
hdl	0.42	1.39	7.20	33	1.5
sbp2	83.00	134.00	258.00	24	1.1
sbp1	83.00	137.00	253.00	23	1.1
cholesterol	2.67	5.68	12.12	23	1.1
ldl	0.70	3.52	9.24	21	1.0
dbp1	53.00	84.00	198.00	17	0.8
dbp2	55.00	83.00	151.00	17	0.8
weight	42.60	77.04	144.44	17	0.8
obs_bp	1.00	4.00	99,902.00	5	0.2
dev_bp	7.00	15.00	99,902.00	3	0.1
obs_soma	1.00	5.00	99,902.00	2	0.1
obs_int	1.00	3.00	99,900.00	2	0.1
dev_length	3.00	3.00	99,902.00	2	0.1
dev_weight	1.00	1.00	99,902.00	2	0.1
height	144.00	168.00	198.00	1	0.0

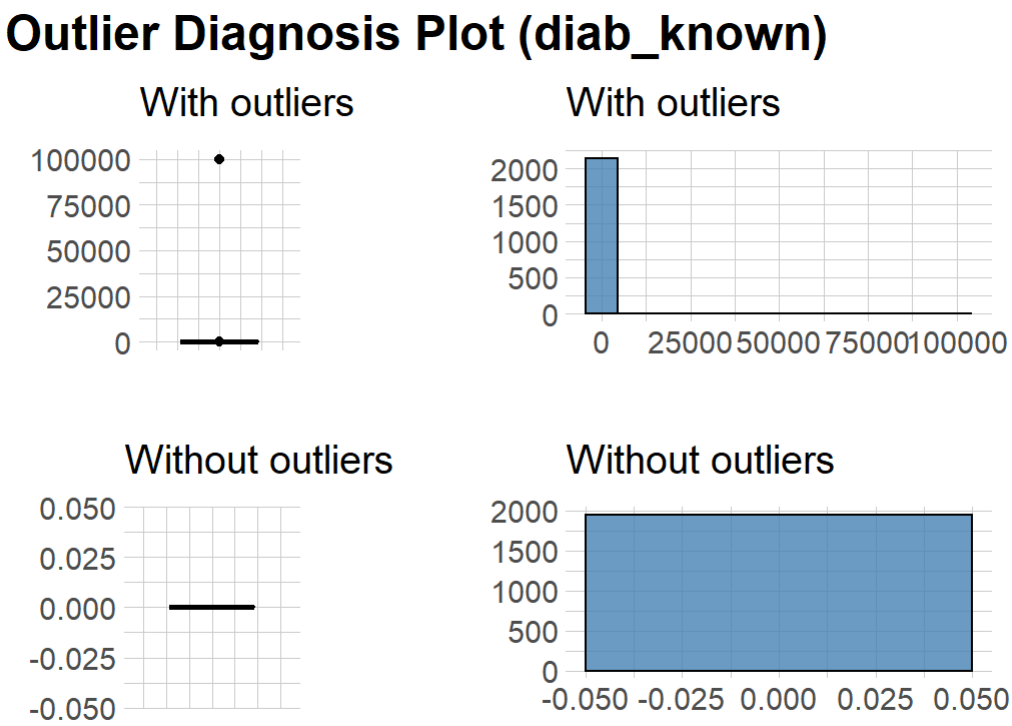
Table 12: Diagnosis of numerical variable outliers

Individual Outliers

variable: diab_known

Measures	Values
Outliers count	208
Outliers ratio (%)	9.66%
Mean of outliers	3362.986
Mean with outliers	324.7451
Mean without outliers	0

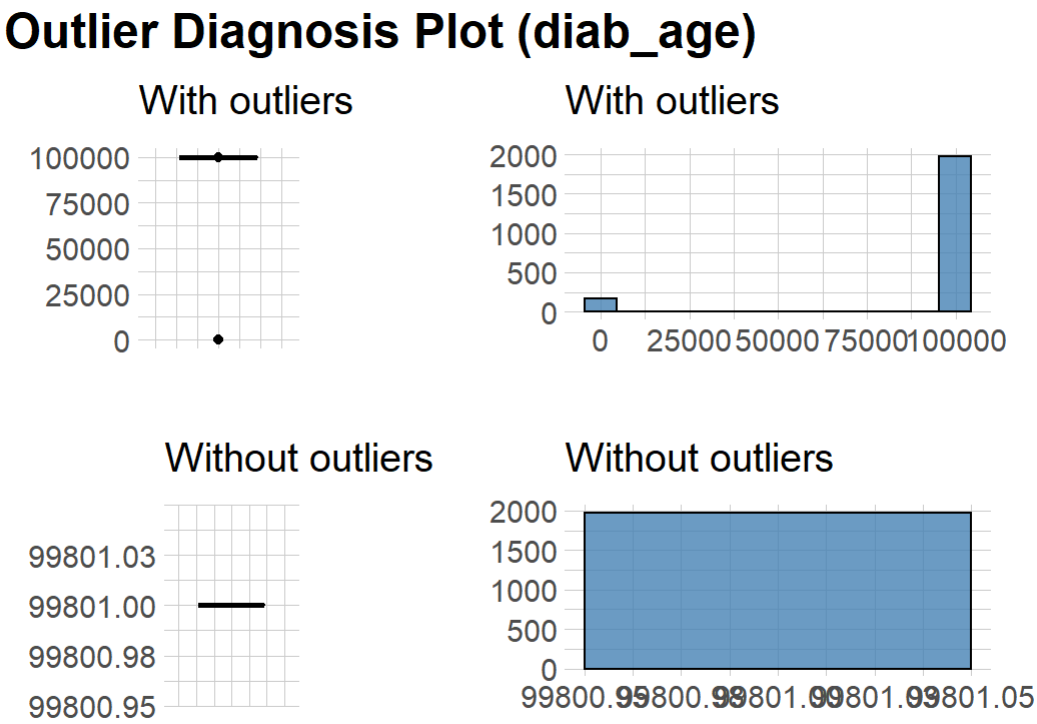
Table 13: diab_known



variable: diab_age

Measures	Values
Outliers count	180
Outliers ratio (%)	8.36%
Mean of outliers	3936.594
Mean with outliers	91790.05
Mean without outliers	99801

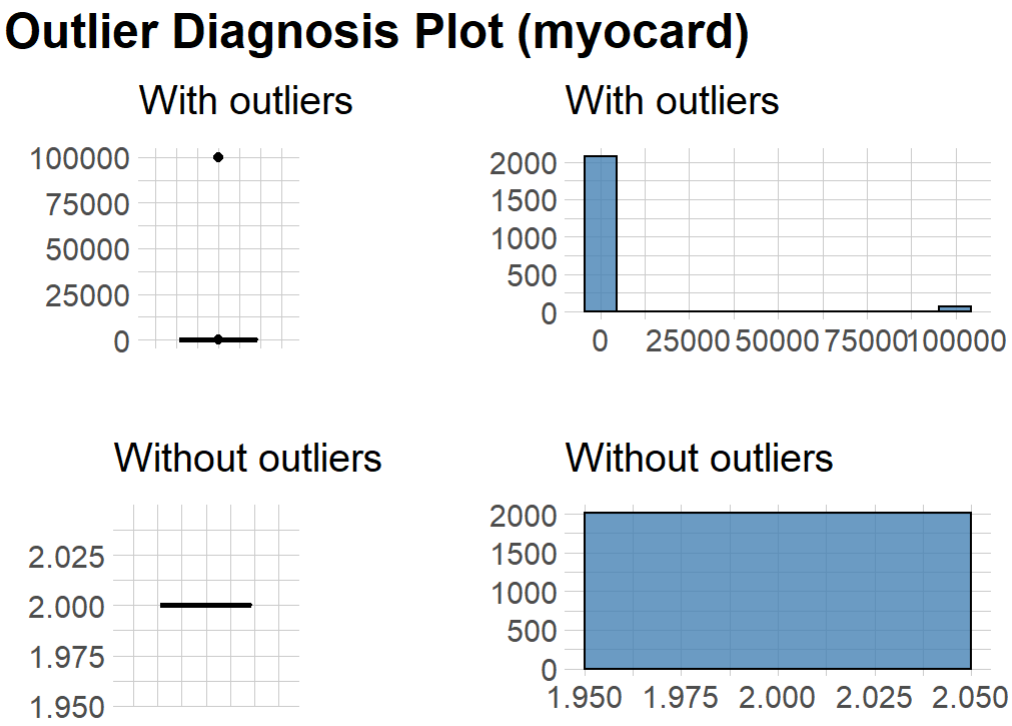
Table 13: diab_age



variable: myocard

Measures	Values
Outliers count	139
Outliers ratio (%)	6.45%
Mean of outliers	53190.73
Mean with outliers	3434.328
Mean without outliers	2

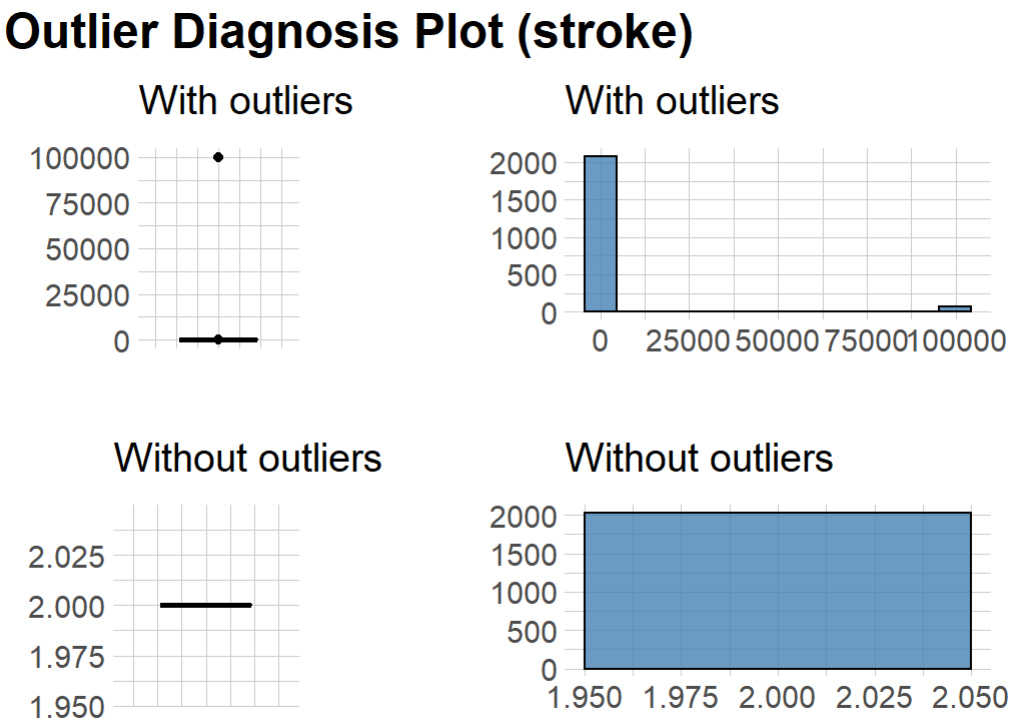
Table 13: myocard



variable: stroke

Measures	Values
Outliers count	126
Outliers ratio (%)	5.85%
Mean of outliers	55507.13
Mean with outliers	3248.818
Mean without outliers	2

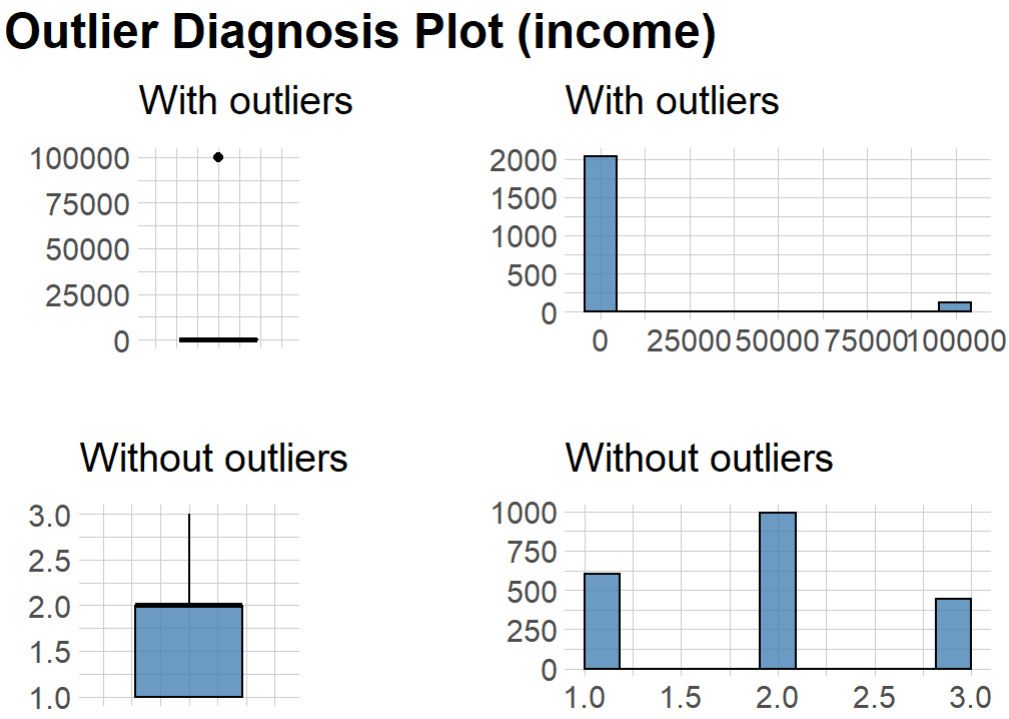
Table 13: stroke



variable: income

Measures	Values
Outliers count	116
Outliers ratio (%)	5.39%
Mean of outliers	99900.86
Mean with outliers	5381.81
Mean without outliers	1.922473

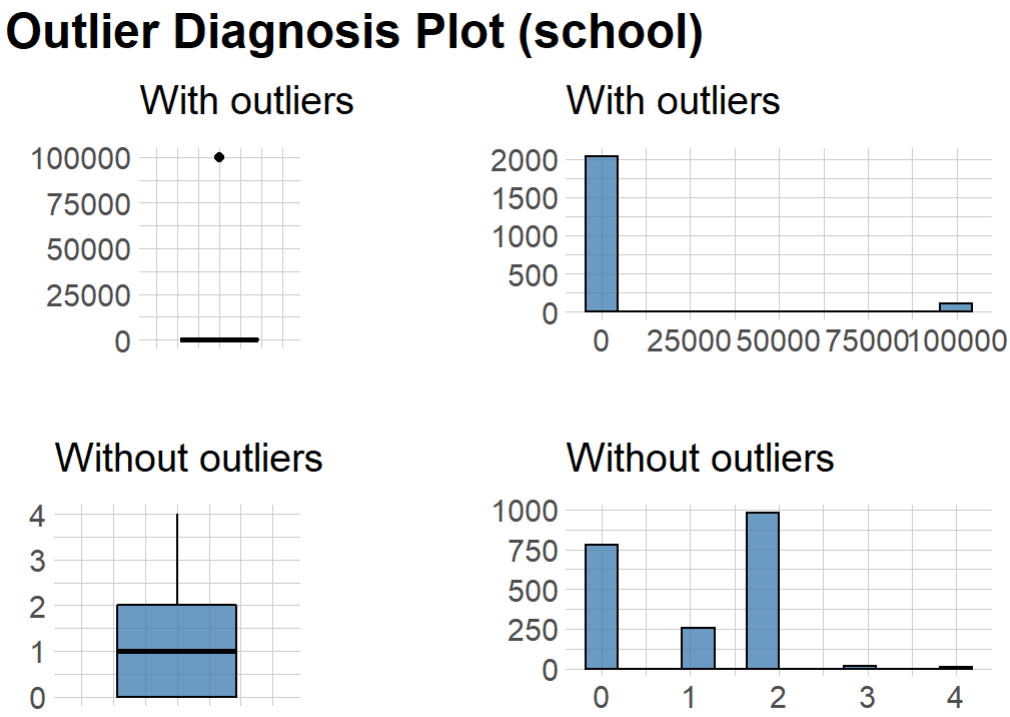
Table 13: income



variable: school

Measures	Values
Outliers count	113
Outliers ratio (%)	5.25%
Mean of outliers	99907.43
Mean with outliers	5242.266
Mean without outliers	1.126899

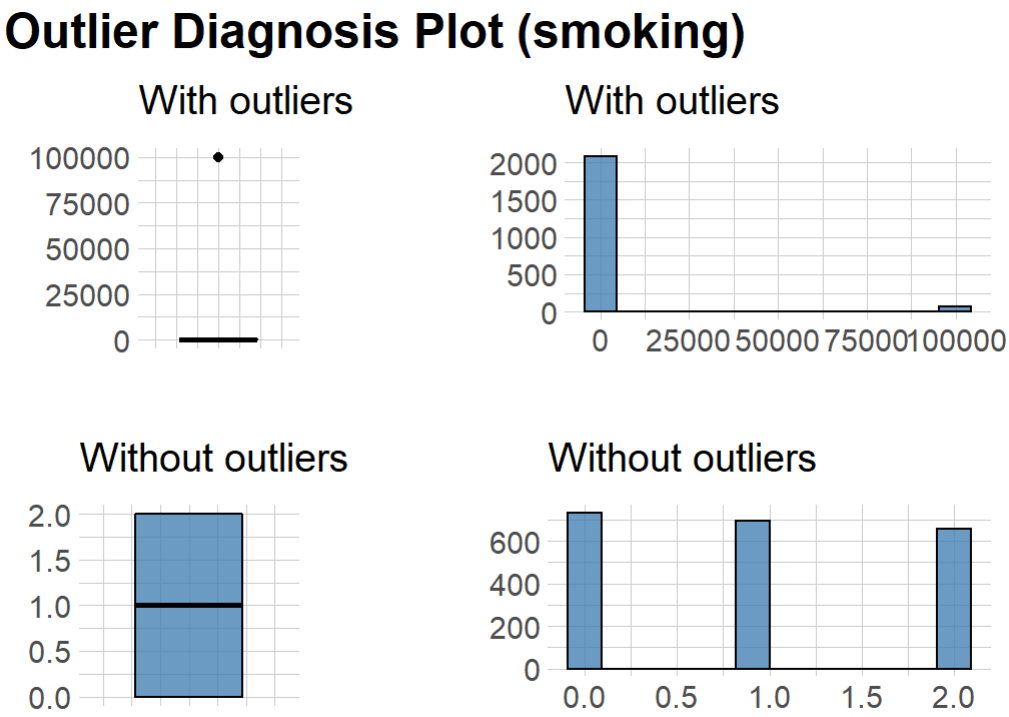
Table 13: school



variable: smoking

Measures	Values
Outliers count	68
Outliers ratio (%)	3.16%
Mean of outliers	99912.35
Mean with outliers	3155.083
Mean without outliers	0.9630872

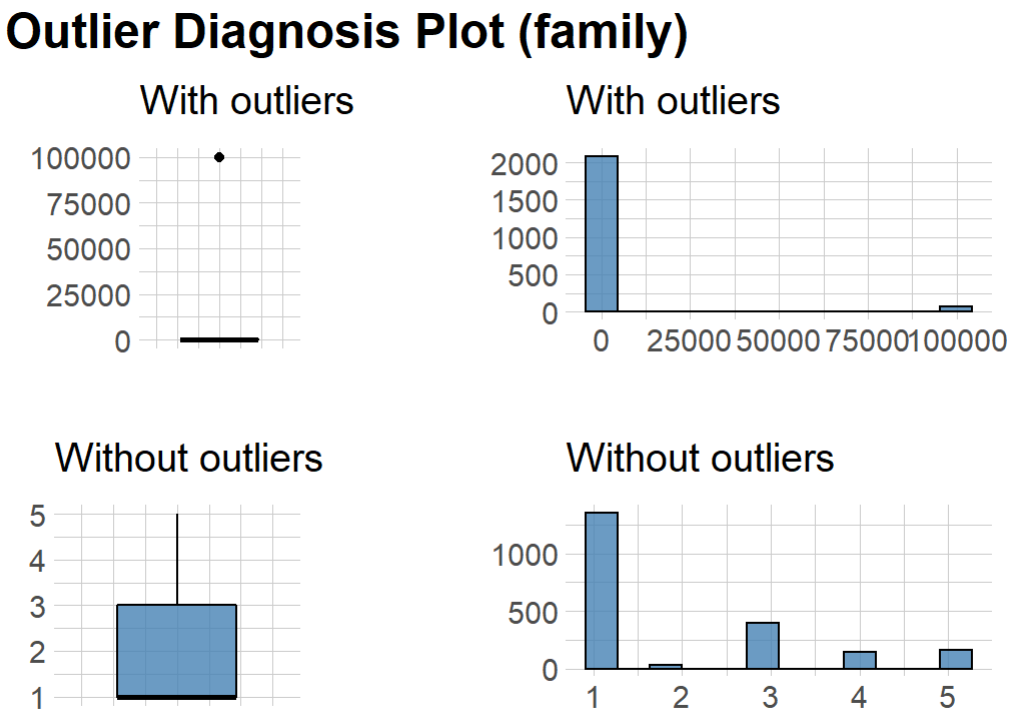
Table 13: smoking



variable: family

Measures	Values
Outliers count	67
Outliers ratio (%)	3.11%
Mean of outliers	99912.54
Mean with outliers	3109.625
Mean without outliers	1.913273

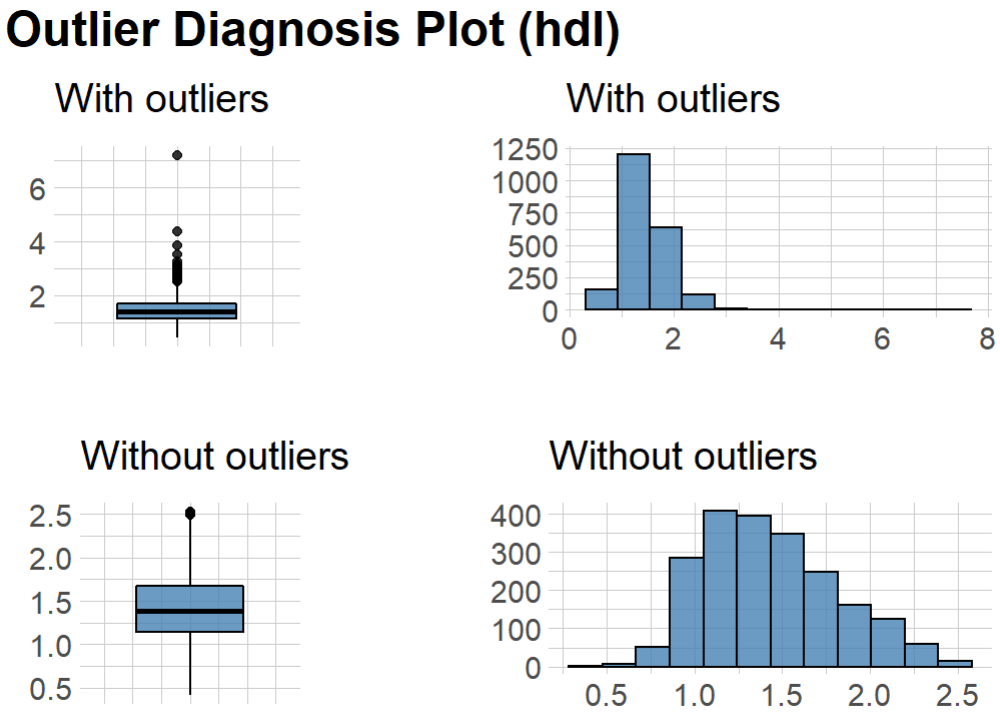
Table 13: family



variable: hdl

Measures	Values
Outliers count	33
Outliers ratio (%)	1.53%
Mean of outliers	3.010848
Mean with outliers	1.452273
Mean without outliers	1.427839

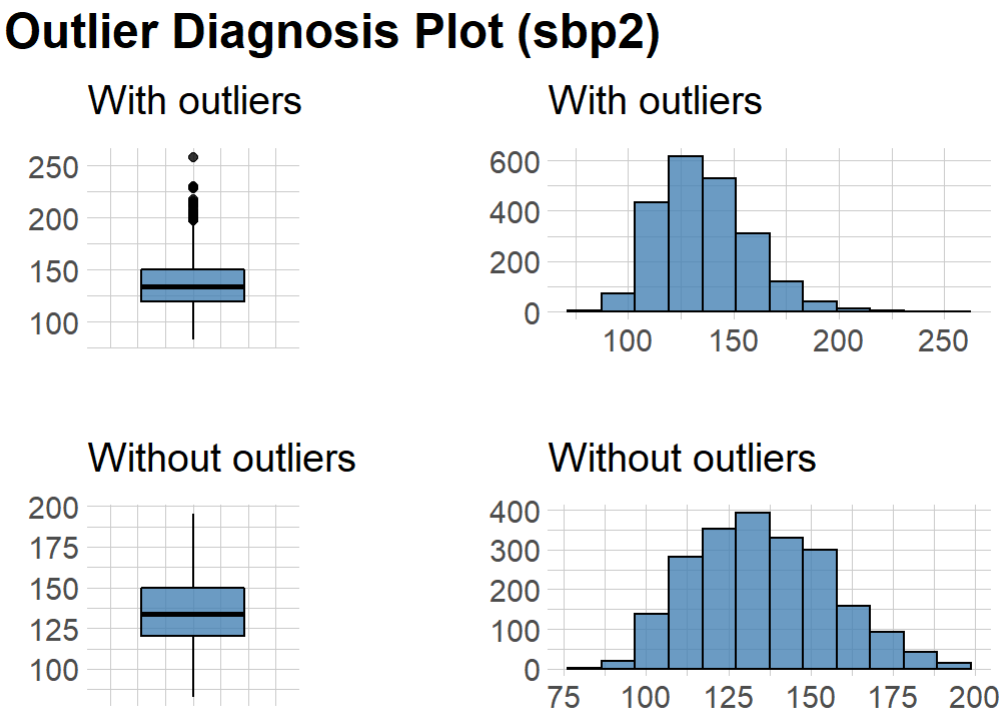
Table 13: hdl



variable: sbp2

Measures	Values
Outliers count	24
Outliers ratio (%)	1.11%
Mean of outliers	209.2917
Mean with outliers	136.3966
Mean without outliers	135.573

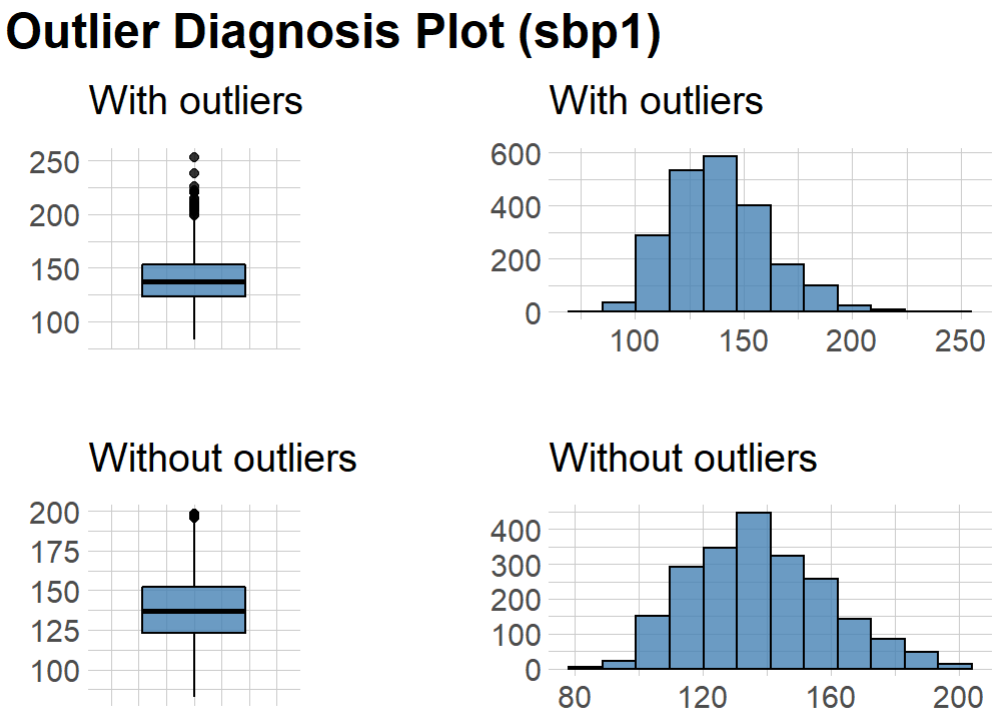
Table 13: sbp2



variable: sbp1

Measures	Values
Outliers count	23
Outliers ratio (%)	1.07%
Mean of outliers	213.9565
Mean with outliers	139.0051
Mean without outliers	138.1954

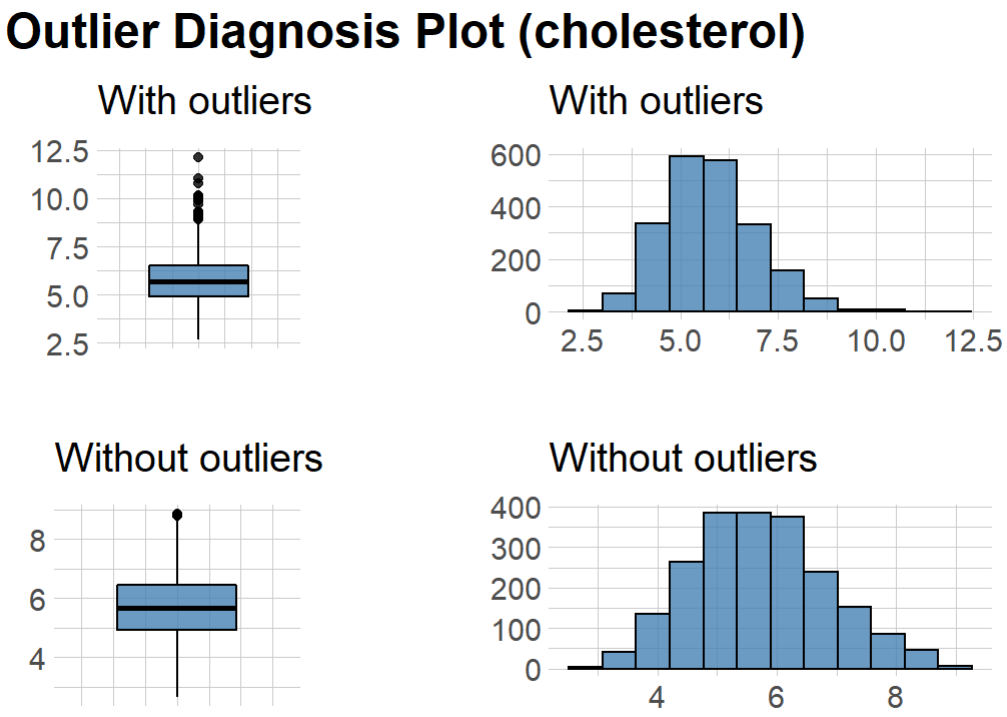
Table 13: sbp1



variable: cholesterol

Measures	Values
Outliers count	23
Outliers ratio (%)	1.07%
Mean of outliers	9.715174
Mean with outliers	5.763767
Mean without outliers	5.720817

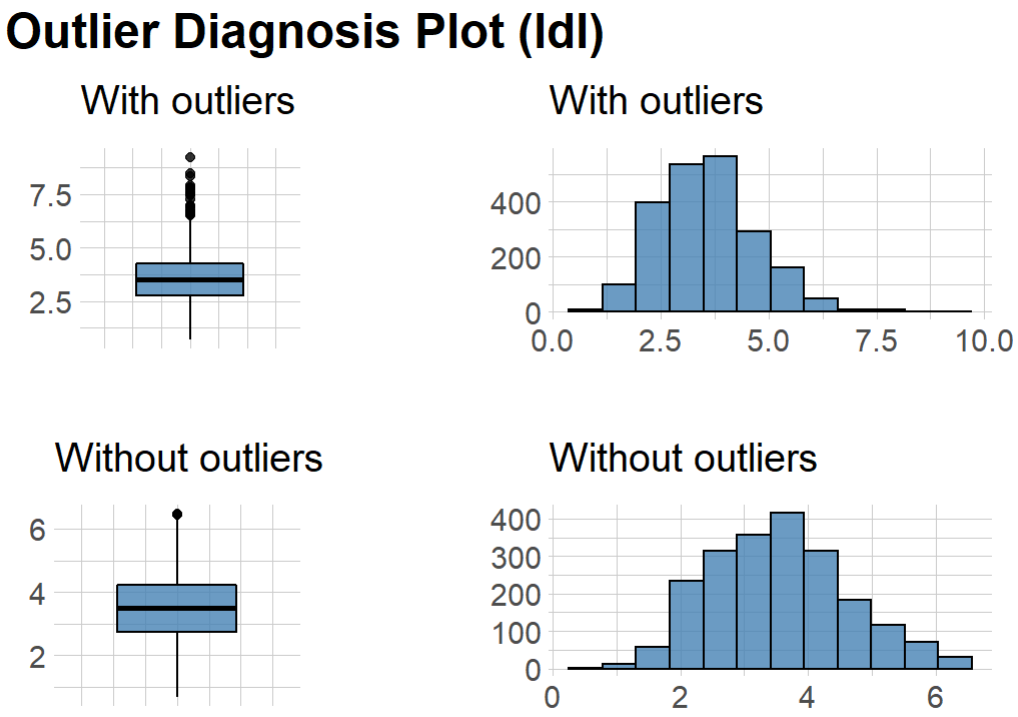
Table 13: cholesterol



variable: ldl

Measures	Values
Outliers count	21
Outliers ratio (%)	0.97%
Mean of outliers	7.395779
Mean with outliers	3.58402
Mean without outliers	3.545993

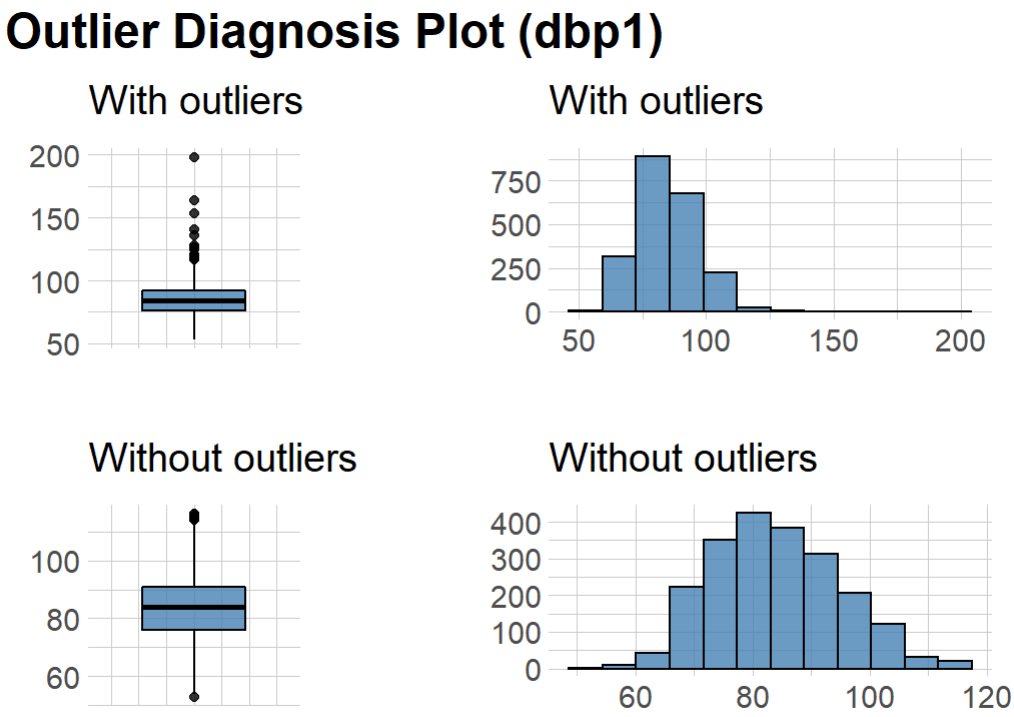
Table 13: ldl



variable: dbp1

Measures	Values
Outliers count	17
Outliers ratio (%)	0.79%
Mean of outliers	132.1765
Mean with outliers	84.52138
Mean without outliers	84.14192

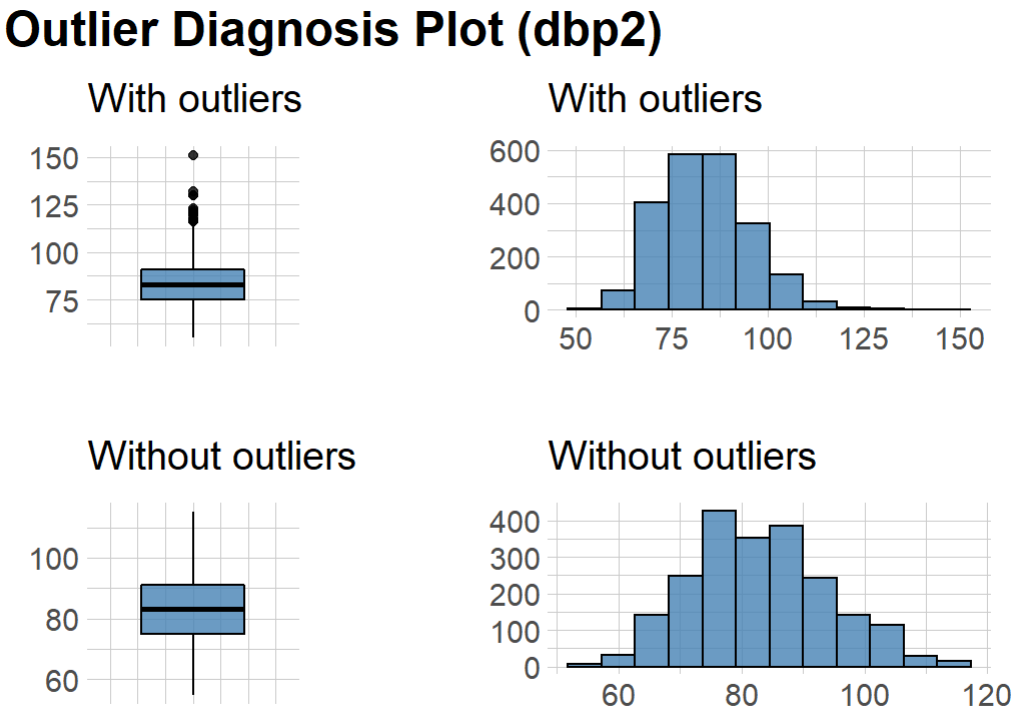
Table 13: dbp1



variable: dbp2

Measures	Values
Outliers count	17
Outliers ratio (%)	0.79%
Mean of outliers	122.7059
Mean with outliers	83.51676
Mean without outliers	83.20413

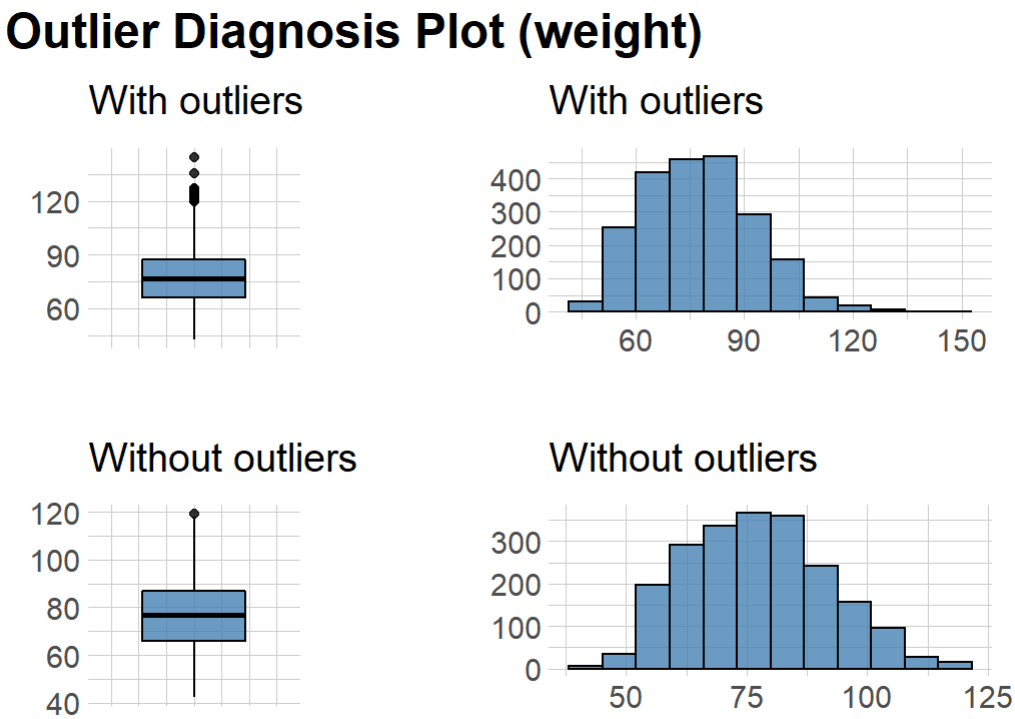
Table 13: dbp2



variable: weight

Measures	Values
Outliers count	17
Outliers ratio (%)	0.79%
Mean of outliers	125.1424
Mean with outliers	77.62963
Mean without outliers	77.25095

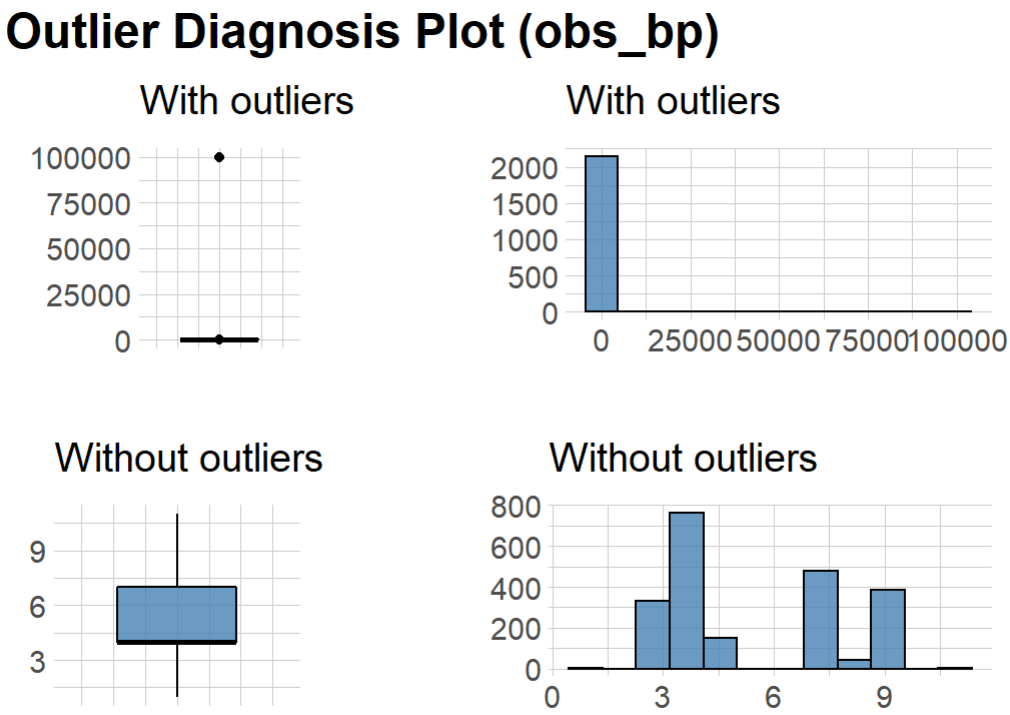
Table 13: weight



variable: obs_bp

Measures	Values
Outliers count	5
Outliers ratio (%)	0.23%
Mean of outliers	59948
Mean with outliers	144.6913
Mean without outliers	5.549093

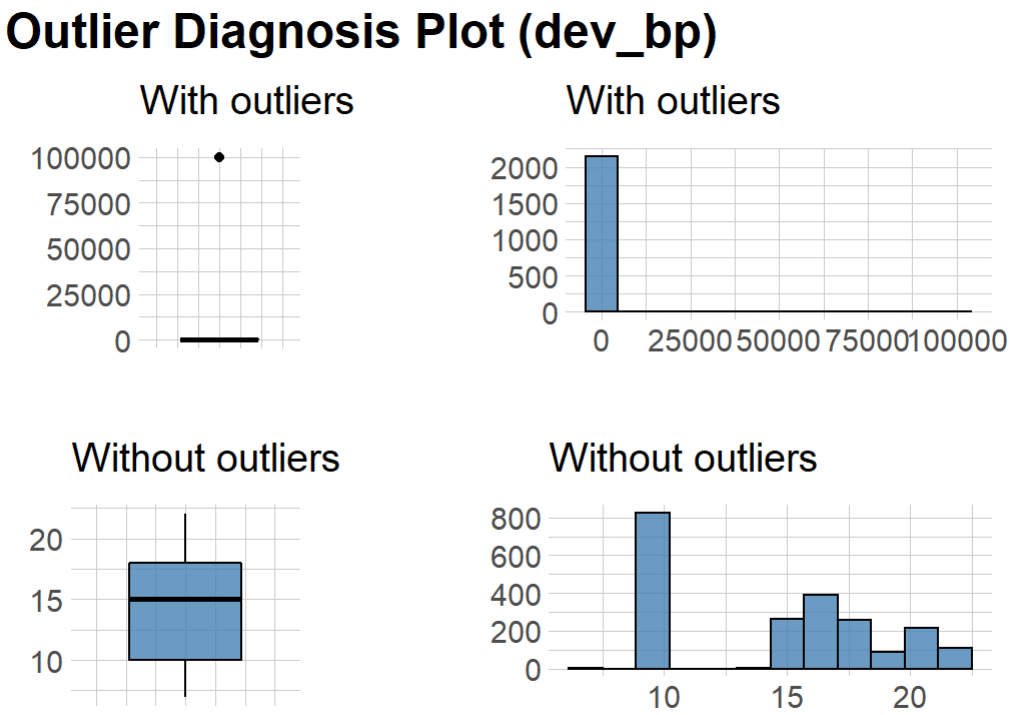
Table 13: obs_bp



variable: dev_bp

Measures	Values
Outliers count	3
Outliers ratio (%)	0.14%
Mean of outliers	99902
Mean with outliers	153.6486
Mean without outliers	14.52952

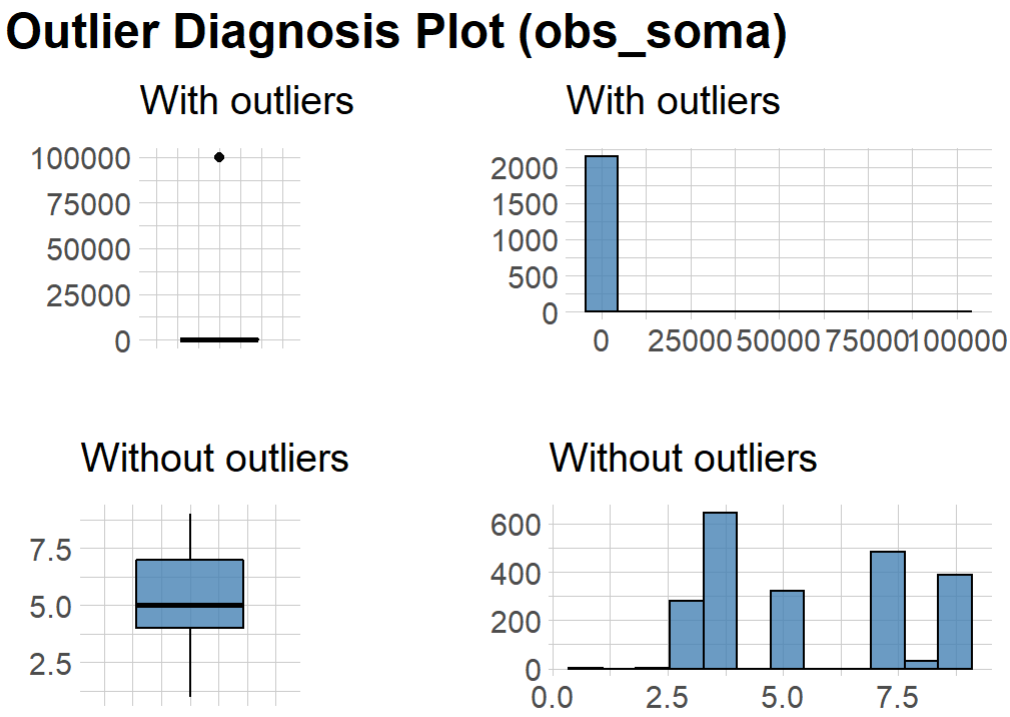
Table 13: dev_bp



variable: obs_soma

Measures	Values
Outliers count	2
Outliers ratio (%)	0.09%
Mean of outliers	99902
Mean with outliers	98.39926
Mean without outliers	5.644981

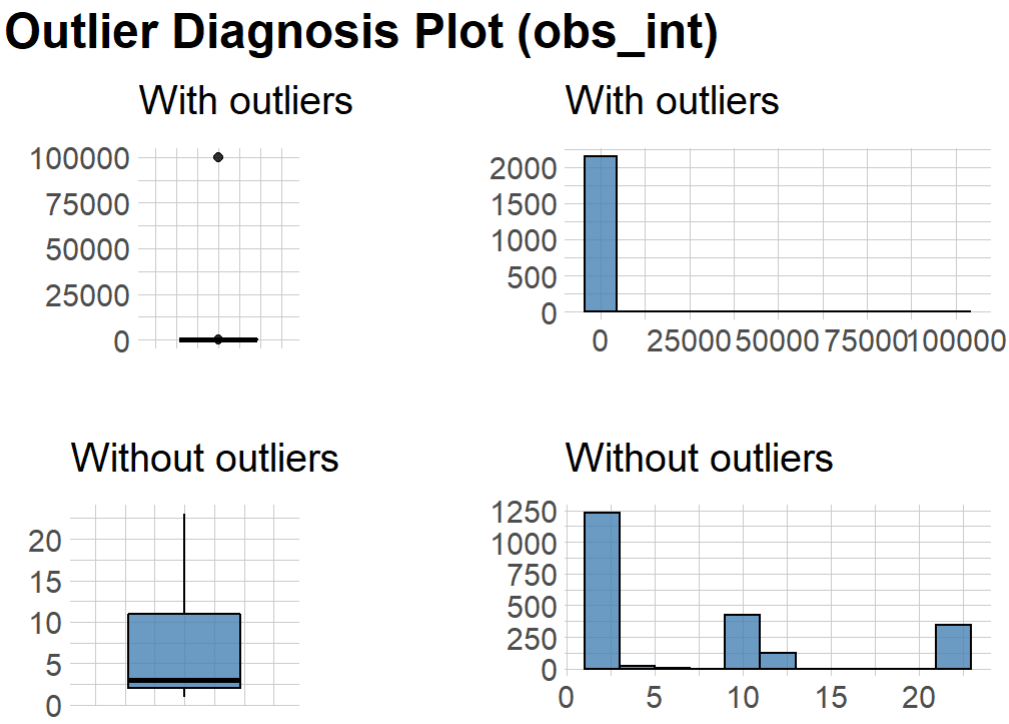
Table 13: obs_soma



variable: obs_int

Measures	Values
Outliers count	2
Outliers ratio (%)	0.09%
Mean of outliers	49962.5
Mean with outliers	53.89554
Mean without outliers	7.512082

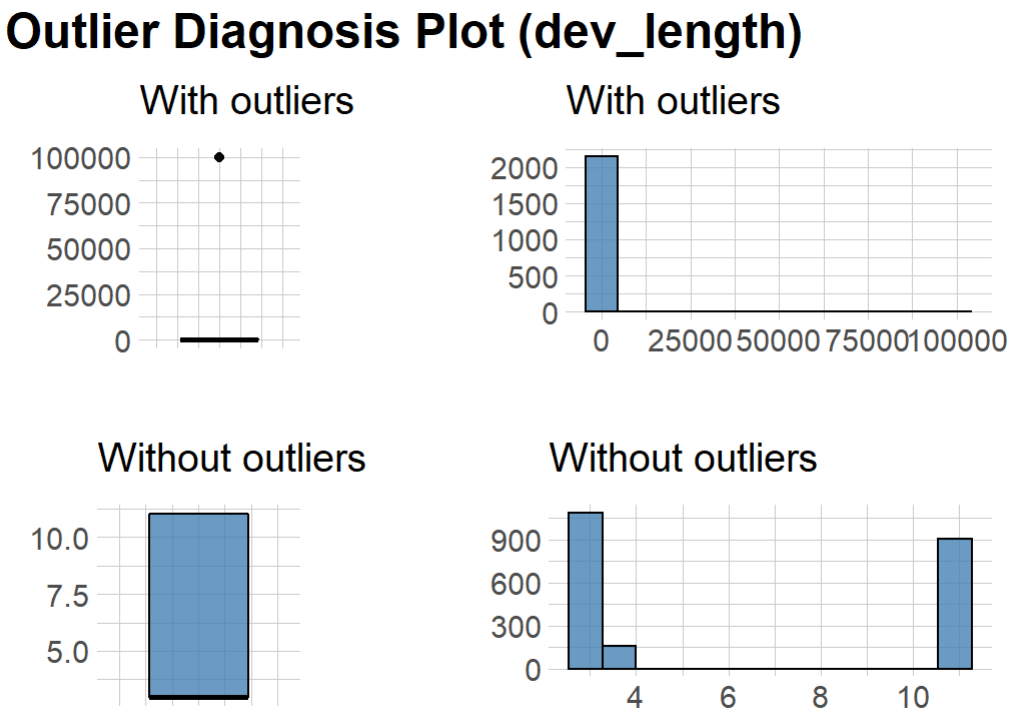
Table 13: obs_int



variable: dev_length

Measures	Values
Outliers count	2
Outliers ratio (%)	0.09%
Mean of outliers	99902
Mean with outliers	99.18849
Mean without outliers	6.434944

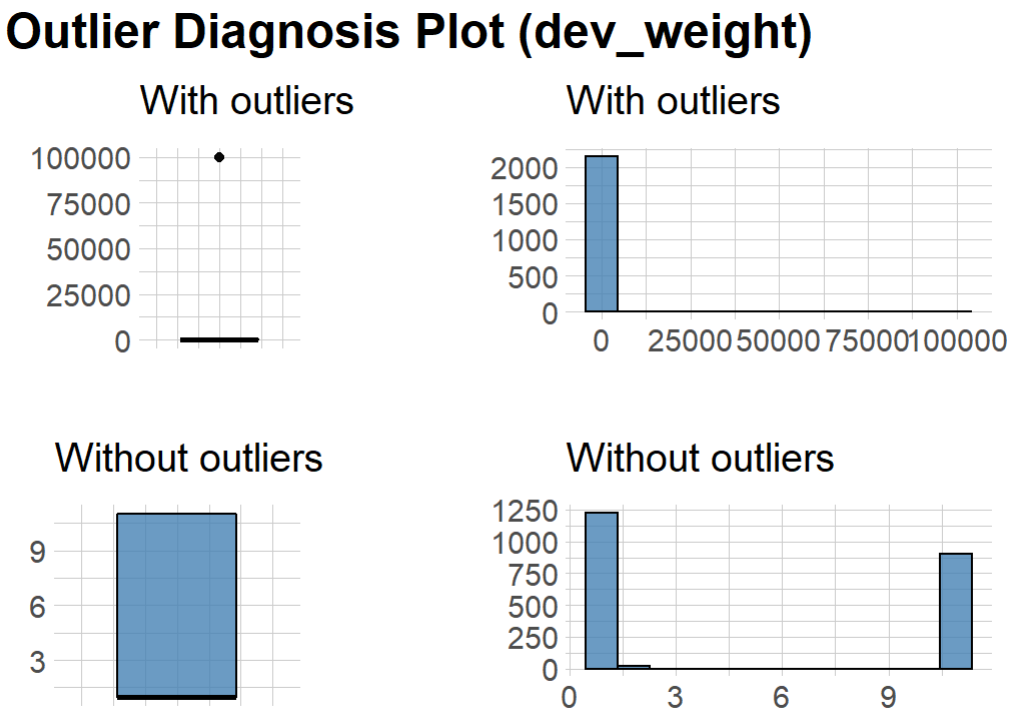
Table 13: dev_length



variable: dev_weight

Measures	Values
Outliers count	2
Outliers ratio (%)	0.09%
Mean of outliers	99902
Mean with outliers	97.96565
Mean without outliers	5.210967

Table 13: dev_weight



variable: height

Measures	Values
Outliers count	1
Outliers ratio (%)	0.05%
Mean of outliers	198
Mean with outliers	168.2194
Mean without outliers	168.2056

Table 13: height