DATA DIAGNOSIS REPORT
NEW_DATA

Report Overview

This report was created for an overview quality diagnosis of ***new_data*** 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	187
size	variables	36
size	values	6,732
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	187
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	10
data type	integers	0
data type	factors/ordered	26
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	new_data
dataset	dataset type	data.frame
job	samples	187 / 187 (100%)
job	created	2024-03-13 01:08:31.193674
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
2	6	0

Table 3: Summary of warnings

warnings	status	recommend
Donorage has high(1.00) cardinality, Maybe identifier	cardinality	check
survival_status has a low cardinality. 2 (1.1%) distinct values	cardinality	judgement
survival_status has 102 (54.55%) zeros	zero	check
PLTrecovery has 23 (12.3%) outliers	outlier	judgement
CD3dCD34 has 16 (8.56%) outliers	outlier	judgement
CD34kgx10d6 has 12 (6.42%) outliers	outlier	judgement
CD3dkgx10d8 has 4 (2.14%) outliers	outlier	judgement
ANCrecovery has 2 (1.07%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables

Variables

variables	types	missing	cardinality	zero	minus	outlier
Recipientgender	factor					
Stemcellsource	factor					
Donorage	numeric		identifier			
Donorage35	factor					
IIIV	factor					
Gendermatch	factor					
DonorABO	factor					
RecipientABO	factor					
RecipientRh	factor					
ABOmatch	factor					
CMVstatus	factor					
DonorCMV	factor					
RecipientCMV	factor					
Disease	factor					
Riskgroup	factor					
Txpostrelapse	factor					
Diseasegroup	factor					
HLAmatch	factor					
HLAmismatch	factor					
Antigen	factor					
Alel	factor					
HLAgrI	factor					
Recipientage	numeric					
Recipientage10	factor					
Recipientageint	factor					

Table 5: List of variables diagnosis

variables	types	missing	cardinality	zero	minus	outlier
variables	types	missing	cardinality	zero	minus	outlier
Relapse	factor					
aGvHDIIIIV	factor					
extcGvHD	factor					
CD34kgx10d6	numeric					X
CD3dCD34	numeric					X
CD3dkgx10d8	numeric					X
Rbodymass	numeric					
ANCrecovery	numeric					X
PLTrecovery	numeric					X
survival_time	numeric					
survival_status	numeric		< low	X		

Table 5: List of variables diagnosis (continued)

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

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

variables	types	unique	unique (%)	status	recommend
survival_status	numeric	2	1.1%	low cardinality	Judgment

Table 6: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
ABOmatch	1	135	72.2
ABOmatch	0	52	27.8
Alel	-1	94	50.3
Alel	0	54	28.9
Alel	1	32	17.1
Alel	2	6	3.2
Alel	3	1	0.5
Antigen	-1	94	50.3
Antigen	1	65	34.8
Antigen	0	21	11.2
Antigen	2	7	3.7
CMVstatus	2	63	33.7
CMVstatus	0	52	27.8
CMVstatus	3	42	22.5
CMVstatus	1	30	16.0
Disease	ALL	68	36.4
Disease	chronic	45	24.1
Disease	AML	33	17.6
Disease	nonmalignant	32	17.1
Disease	lymphoma	9	4.8
Diseasegroup	1	155	82.9
Diseasegroup	0	32	17.1
DonorABO	0	73	39.0
DonorABO	1	71	38.0
DonorABO	-1	28	15.0

Table 7: Top 10 levels of categorical variables

	variables	levels	freq	ratio (%)
26	DonorABO	2	15	8.0
27	DonorCMV	0	115	61.5
28	DonorCMV	1	72	38.5
29	Donorage35	0	104	55.6
30	Donorage35	1	83	44.4
31	Gendermatch	0	155	82.9
32	Gendermatch	1	32	17.1
33	HLAgrI	0	94	50.3
34	HLAgrI	1	42	22.5
35	HLAgrI	4	19	10.2
36	HLAgrI	2	14	7.5
37	HLAgrI	3	9	4.8
38	HLAgrI	7	5	2.7
39	HLAgrI	5	4	2.1
40	HLAmatch	0	94	50.3
41	HLAmatch	1	65	34.8
42	HLAmatch	2	23	12.3
43	HLAmatch	3	5	2.7
44	HLAmismatch	0	159	85.0
45	HLAmismatch	1	28	15.0
46	IIIV	1	112	59.9
47	IIIV	0	75	40.1
48	RecipientABO	1	76	40.6
49	RecipientABO	-1	50	26.7
50	RecipientABO	0	48	25.7
51	RecipientABO	2	13	7.0
52	RecipientCMV	1	105	56.1
53	RecipientCMV	0	82	43.9
54	RecipientRh	1	160	85.6

Table 7: Top 10 levels of categorical variables (continued)

	variables	levels	freq	ratio (%)
55	RecipientRh	0	27	14.4
56	Recipientage10	0	99	52.9
57	Recipientage10	1	88	47.1
58	Recipientageint	2	89	47.6
59	Recipientageint	1	51	27.3
60	Recipientageint	0	47	25.1
61	Recipientgender	1	112	59.9
62	Recipientgender	0	75	40.1
63	Relapse	0	159	85.0
64	Relapse	1	28	15.0
65	Riskgroup	0	118	63.1
66	Riskgroup	1	69	36.9
67	Stemcellsource	1	145	77.5
68	Stemcellsource	0	42	22.5
69	Txpostrelapse	0	164	87.7
70	Txpostrelapse	1	23	12.3
71	aGvHDIIIIV	1	147	78.6
72	aGvHDIIIIV	0	40	21.4
73	extcGvHD	1	153	81.8
74	extcGvHD	0	34	18.2

Table 7: Top 10 levels of categorical variables (continued)

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Donorage	18.65	27.04	33.47	33.55	40.12	55.55	0	0	0
Recipientage	0.60	5.05	9.93	9.60	14.05	20.20	0	0	0
CD34kgx10d6	0.79	5.35	11.89	9.72	15.41	57.78	0	0	12
CD3dCD34	0.20	1.79	5.41	2.73	6.01	99.56	0	0	16
CD3dkgx10d8	0.04	1.67	4.70	4.26	6.57	20.02	0	0	4
Rbodymass	6.00	19.00	35.78	33.00	50.70	97.80	0	0	0
ANCrecovery	9.00	13.00	15.32	15.00	17.00	26.00	0	0	2
PLTrecovery	9.00	15.50	31.23	21.00	29.50	285.00	0	0	23
survival_time	6.00	168.50	938.74	676.00	1,604.00	3,364.00	0	0	0
survival_status	0.00	0.00	0.45	0.00	1.00	1.00	102	0	0

Table 8: General list of numerical diagnosis

Zero Values

variables	min	median	max	zero	zero (%)
survival_status	0	0	1	102	54.5

Table 9: List of numerical diagnosis (zero)

Negative Values

No numeric variable with negative value

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
PLTrecovery	9.00	21.00	285.00	23	12.3
CD3dCD34	0.20	2.73	99.56	16	8.6
CD34kgx10d6	0.79	9.72	57.78	12	6.4
CD3dkgx10d8	0.04	4.26	20.02	4	2.1
ANCrecovery	9.00	15.00	26.00	2	1.1

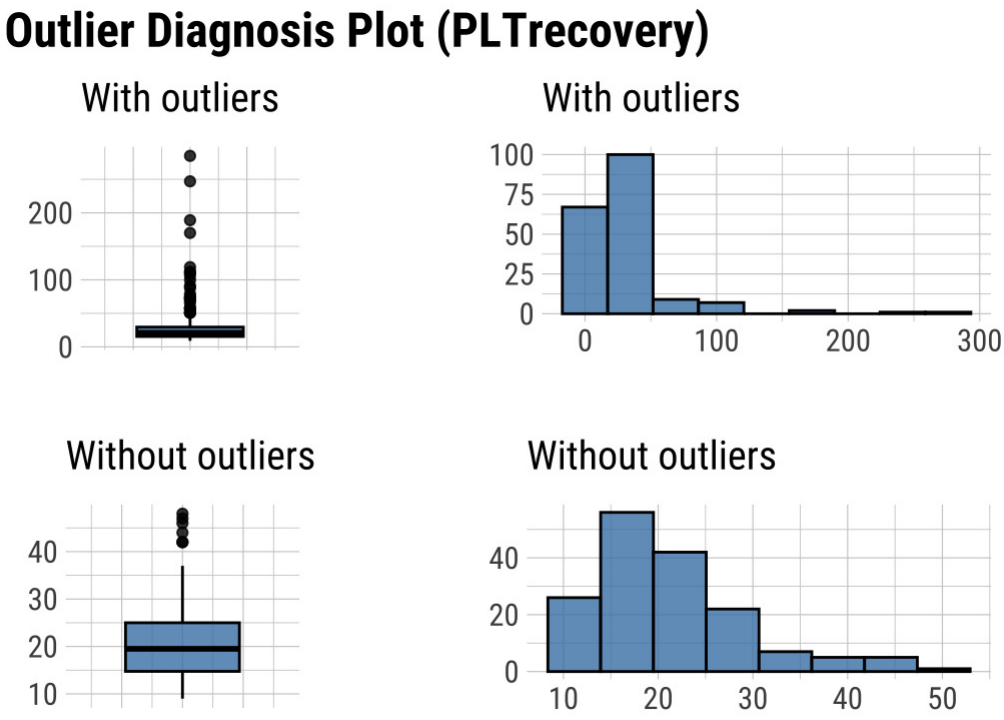
Table 10: Diagnosis of numerical variable outliers

Individual Outliers

variable: PLTrecovery

Measures	Values
Outliers count	23
Outliers ratio (%)	12.3%
Mean of outliers	103.4348
Mean with outliers	31.22995
Mean without outliers	21.10366

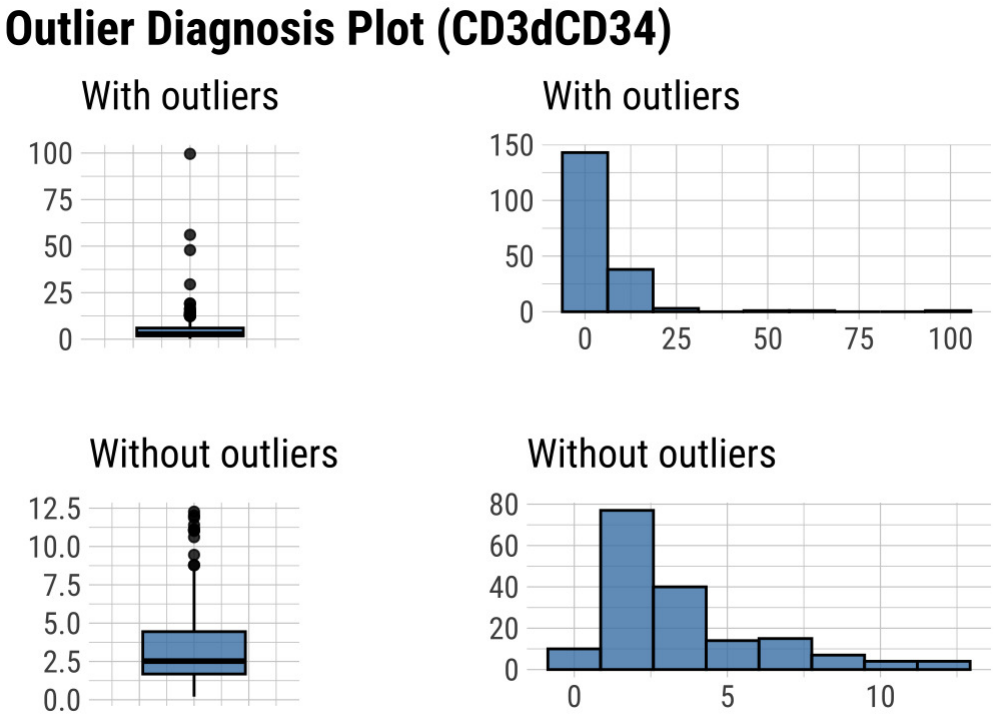
Table 11: PLTrecovery



variable: CD3dCD34

Measures	Values
Outliers count	16
Outliers ratio (%)	8.56%
Mean of outliers	25.63439
Mean with outliers	5.409611
Mean without outliers	3.517234

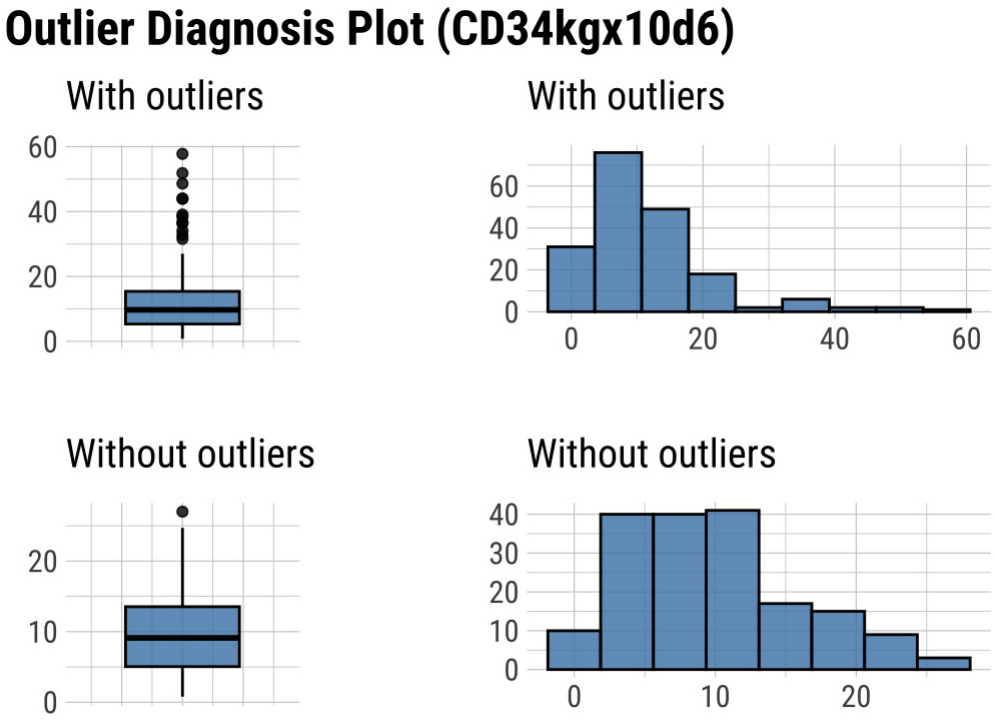
Table 11: CD3dCD34



variable: CD34kgx10d6

Measures	Values
Outliers count	12
Outliers ratio (%)	6.42%
Mean of outliers	41.25
Mean with outliers	11.89178
Mean without outliers	9.878646

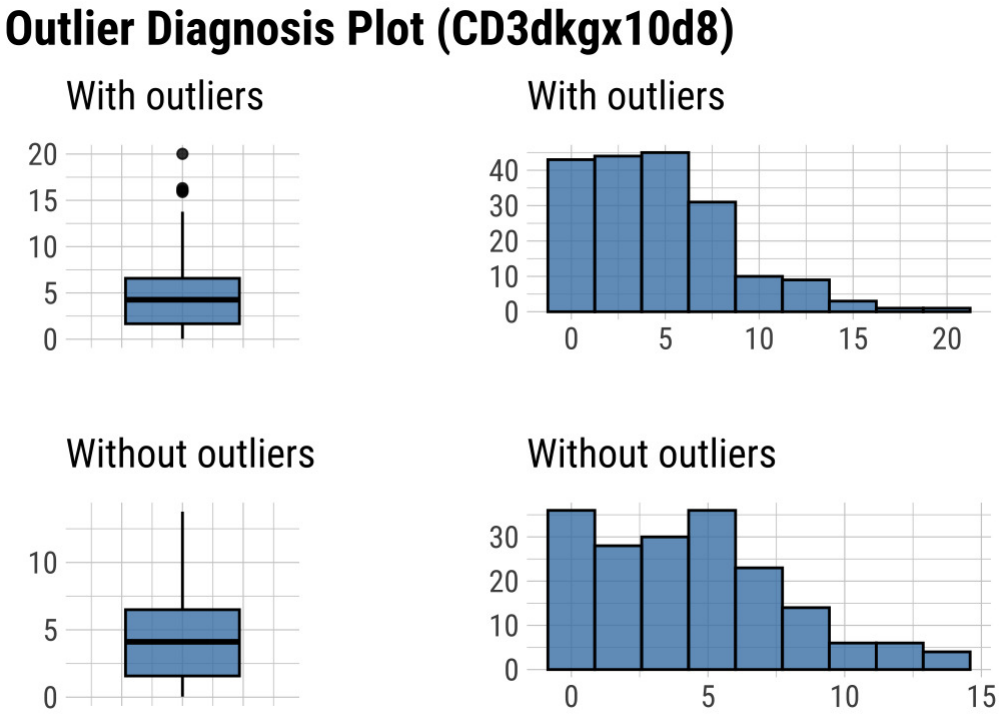
Table 11: CD34kgx10d6



variable: CD3dkgx10d8

Measures	Values
Outliers count	4
Outliers ratio (%)	2.14%
Mean of outliers	17.075
Mean with outliers	4.697219
Mean without outliers	4.426667

Table 11: CD3dkgx10d8



variable: ANCrecovery

Measures	Values
Outliers count	2
Outliers ratio (%)	1.07%
Mean of outliers	25
Mean with outliers	15.31551
Mean without outliers	15.21081

Table 11: ANCrecovery