DATA DIAGNOSIS REPORT

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	153,158
size	variables	18
size	values	2,756,844
size	memory size (MB)	26
duplicated	duplicate observation	0
missing	complete observation	153,158
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	0
data type	integers	16
data type	factors/ordered	0
data type	characters	2
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	data.frame
job	samples	153,158 / 153,158 (100%)
job	created	2022-11-28 12:54:37
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
17	19	1

Table 3: Summary of warnings

warnings	status	recommand
accident_year has constant value "2019"	cardinality	remove
casualty_class has a low cardinality. 3 (0%) distinct values	cardinality	judgement
sex_of_casualty has a low cardinality. 4 (0%) distinct values	cardinality	judgement
casualty_severity has a low cardinality. 3 (0%) distinct values	cardinality	judgement
car_passenger has a low cardinality. 5 (0%) distinct values	cardinality	judgement
pedestrian_road_maintenance_worker has a low cardinality. 4 (0%) distinct values	cardinality	judgement
casualty_home_area_type has a low cardinality. 4 (0%) distinct values	cardinality	judgement
pedestrian_road_maintenance_worker has 151,736 (99.07%) zeros	zero	check
bus_or_coach_passenger has 150,334 (98.16%) zeros	zero	check
pedestrian_location has 131,388 (85.79%) zeros	zero	check
pedestrian_movement has 131,386 (85.78%) zeros	zero	check
car_passenger has 125,535 (81.96%) zeros	zero	check
casualty_type has 21,770 (14.21%) zeros	zero	check
age_of_casualty has 247 (0.16%) zeros	zero	check
casualty_imd_decile has 15,372 (10.04%) negatives	negative	check
casualty_home_area_type has 15,355 (10.03%) negatives	negative	check
age_of_casualty has 3,255 (2.13%) negatives	negative	check
age_band_of_casualty has 3,255 (2.13%) negatives	negative	check
sex_of_casualty has 724 (0.47%) negatives	negative	check
car_passenger has 382 (0.25%) negatives	negative	check
pedestrian_road_maintenance_worker has 74 (0.05%) negatives	negative	check
bus_or_coach_passenger has 61 (0.04%) negatives	negative	check

Table 4: Warnings in dataset and variables

	warnings	status	recommand
	warnings	**status**	**recommand**
23	casualty_type has 5 (0%) negatives	negative	check
24	pedestrian_location has 1 (0%) negatives	negative	check
25	casualty_home_area_type has 42,006 (27.43%) outliers	outlier	judgement
26	casualty_reference has 36,854 (24.06%) outliers	outlier	judgement
27	car_passenger has 27,623 (18.04%) outliers	outlier	judgement
28	casualty_severity has 27,125 (17.71%) outliers	outlier	judgement
29	pedestrian_movement has 21,772 (14.22%) outliers	outlier	judgement
30	pedestrian_location has 21,770 (14.21%) outliers	outlier	judgement
31	age_band_of_casualty has 3,255 (2.13%) outliers	outlier	judgement
32	bus_or_coach_passenger has 2,824 (1.84%) outliers	outlier	judgement
33	pedestrian_road_maintenance_worker has 1,422 (0.93%) outliers	outlier	judgement
34	casualty_type has 1,124 (0.73%) outliers	outlier	judgement
35	vehicle_reference has 991 (0.65%) outliers	outlier	judgement
36	sex_of_casualty has 733 (0.48%) outliers	outlier	judgement
37	age_of_casualty has 178 (0.12%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables (continued)

Variables

variables	types	missing	cardinality	zero	minus	outlier
accident_index	character		> high			
accident_year	integer		constant			
accident_reference	character		> high			
vehicle_reference	integer					X
casualty_reference	integer					X
casualty_class	integer		< low			
sex_of_casualty	integer		< low		X	X
age_of_casualty	integer			X	X	X
age_band_of_casualty	integer				X	X
casualty_severity	integer		< low			X
pedestrian_location	integer			X	X	X
pedestrian_movement	integer			X		X
car_passenger	integer		< low	X	X	X
bus_or_coach_passenger	integer			X	X	X
pedestrian_road_maintenance_worker	integer		< low	X	X	X
casualty_type	integer			X	X	X
casualty_home_area_type	integer		< low		X	X
casualty_imd_decile	integer				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

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

variables	types	unique	unique (%)	status	recommand
accident_index	character	117,536	76.7%	high cardinality	Judgment
accident_reference	character	117,536	76.7%	high cardinality	Judgment

Table 6: 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
accident_year	integer	1	0%	constant	Remove Variable
casualty_class	integer	3	0%	low cardinality	Judgment
sex_of_casualty	integer	4	0%	low cardinality	Judgment
casualty_severity	integer	3	0%	low cardinality	Judgment
car_passenger	integer	5	0%	low cardinality	Judgment
pedestrian_road_maintenance_worker	integer	4	0%	low cardinality	Judgment
casualty_home_area_type	integer	4	0%	low cardinality	Judgment

Table 7: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
accident_index	2019500885809	52	0.0
accident_index	2019220855375	25	0.0
accident_index	2019350900122	20	0.0
accident_index	2019410889448	19	0.0
accident_index	2019440129002	19	0.0
accident_index	2019136AT1088	16	0.0
accident_index	2019051911747	13	0.0
accident_index	20191369T0667	13	0.0
accident_index	2019140838359	13	0.0
accident_index	2019200843528	13	0.0
accident_index	Other levles	152,955	99.9
accident_reference	500885809	52	0.0
accident_reference	220855375	25	0.0
accident_reference	350900122	20	0.0
accident_reference	410889448	19	0.0
accident_reference	440129002	19	0.0
accident_reference	136AT1088	16	0.0
accident_reference	051911747	13	0.0
accident_reference	1369T0667	13	0.0
accident_reference	140838359	13	0.0
accident_reference	200843528	13	0.0
accident_reference	Other levles	152,955	99.9

Table 8: Top 10 levels of categorical variables

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
accident_year	2,019	2,019	2,019.00	2,019	2,019	2,019	0	0	

Table 9: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
vehicle_reference	1	1	1.46	1	2	20	0	0	991
casualty_reference	1	1	1.39	1	1	991	0	0	36,854
casualty_class	1	1	1.49	1	2	3	0	0	0
sex_of_casualty	-1	1	1.39	1	2	9	0	724	733
age_of_casualty	-1	22	36.93	34	50	102	247	3,255	178
age_band_of_casualty	-1	5	6.32	6	8	11	0	3,255	3,255
casualty_severity	1	3	2.81	3	3	3	0	0	27,125
pedestrian_location	-1	0	0.76	0	0	10	131,388	1	21,770
pedestrian_movement	0	0	0.60	0	0	9	131,386	0	21,772
car_passenger	-1	0	0.25	0	0	9	125,535	382	27,623
bus_or_coach_passenger	-1	0	0.06	0	0	9	150,334	61	2,824
pedestrian_road_maintenance_worker	-1	0	0.02	0	0	2	151,736	74	1,422
casualty_type	-1	1	7.14	9	9	98	21,770	5	1,124
casualty_home_area_type	-1	1	1.07	1	1	3	0	15,355	42,006
casualty_imd_decile	-1	2	4.38	4	7	10	0	15,372	0

Zero Values

variables	min	median	max	zero	zero (%)
pedestrian_road_maintenance_worker	-1	0	2	151,736	99.1
bus_or_coach_passenger	-1	0	9	150,334	98.2
pedestrian_location	-1	0	10	131,388	85.8
pedestrian_movement	0	0	9	131,386	85.8
car_passenger	-1	0	9	125,535	82.0
casualty_type	-1	9	98	21,770	14.2
age_of_casualty	-1	34	102	247	0.2

Table 10: List of numerical diagnosis (zero)

Negative Values

variables	min	median	max	minus	minus (%)
casualty_imd_decile	-1	4	10	15,372	10.0
casualty_home_area_type	-1	1	3	15,355	10.0
age_of_casualty	-1	34	102	3,255	2.1
age_band_of_casualty	-1	6	11	3,255	2.1
sex_of_casualty	-1	1	9	724	0.5
car_passenger	-1	0	9	382	0.2
pedestrian_road_maintenance_worker	-1	0	2	74	0.0
bus_or_coach_passenger	-1	0	9	61	0.0
casualty_type	-1	9	98	5	0.0
pedestrian_location	-1	0	10	1	0.0

Table 11: List of numerical diagnosis (minus)

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
casualty_home_area_type	-1	1	3	42,006	27.4
casualty_reference	1	1	991	36,854	24.1
car_passenger	-1	0	9	27,623	18.0
casualty_severity	1	3	3	27,125	17.7
pedestrian_movement	0	0	9	21,772	14.2
pedestrian_location	-1	0	10	21,770	14.2
age_band_of_casualty	-1	6	11	3,255	2.1
bus_or_coach_passenger	-1	0	9	2,824	1.8
pedestrian_road_maintenance_worker	-1	0	2	1,422	0.9
casualty_type	-1	9	98	1,124	0.7
vehicle_reference	1	1	20	991	0.6
sex_of_casualty	-1	1	9	733	0.5
age_of_casualty	-1	34	102	178	0.1

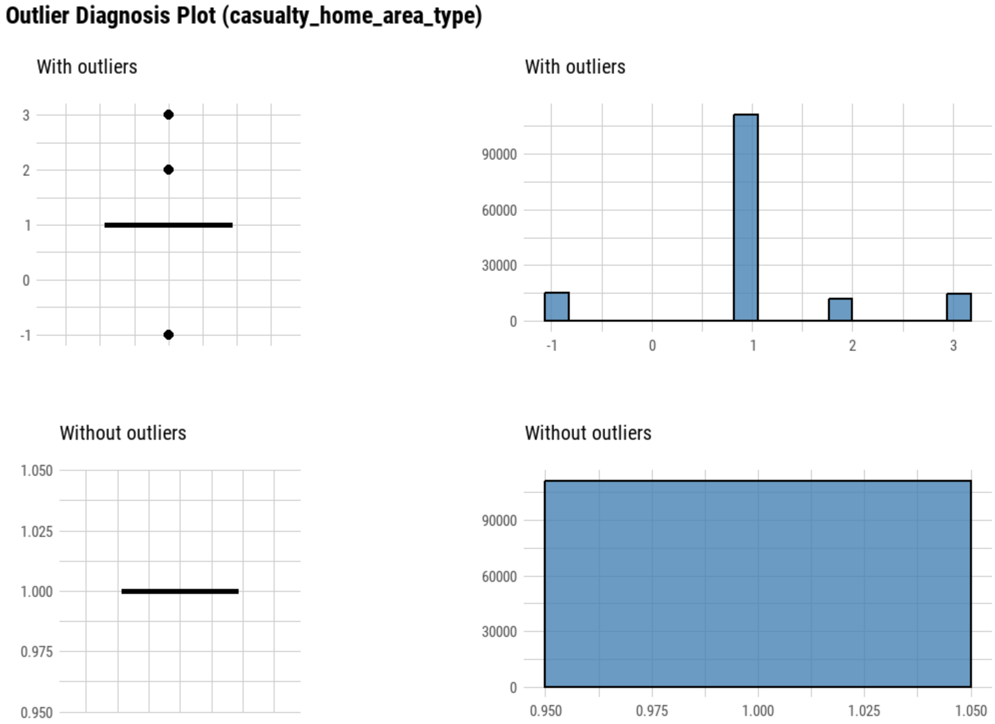
Table 12: Diagnosis of numerical variable outliers

Individual Outliers

variable: casualty_home_area_type

Measures	Values
Outliers count	42,006
Outliers ratio (%)	27.43%
Mean of outliers	1.256392
Mean with outliers	1.07032
Mean without outliers	1

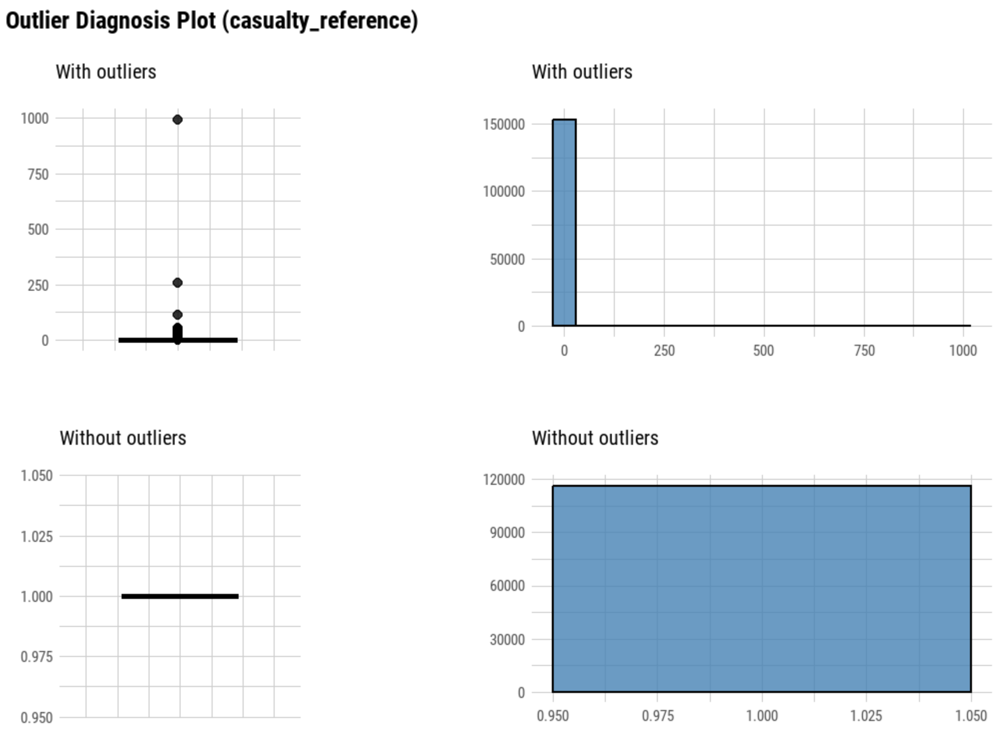
Table 13: casualty_home_area_type



variable: casualty_reference

Measures	Values
Outliers count	36,854
Outliers ratio (%)	24.06%
Mean of outliers	2.637109
Mean with outliers	1.393933
Mean without outliers	1

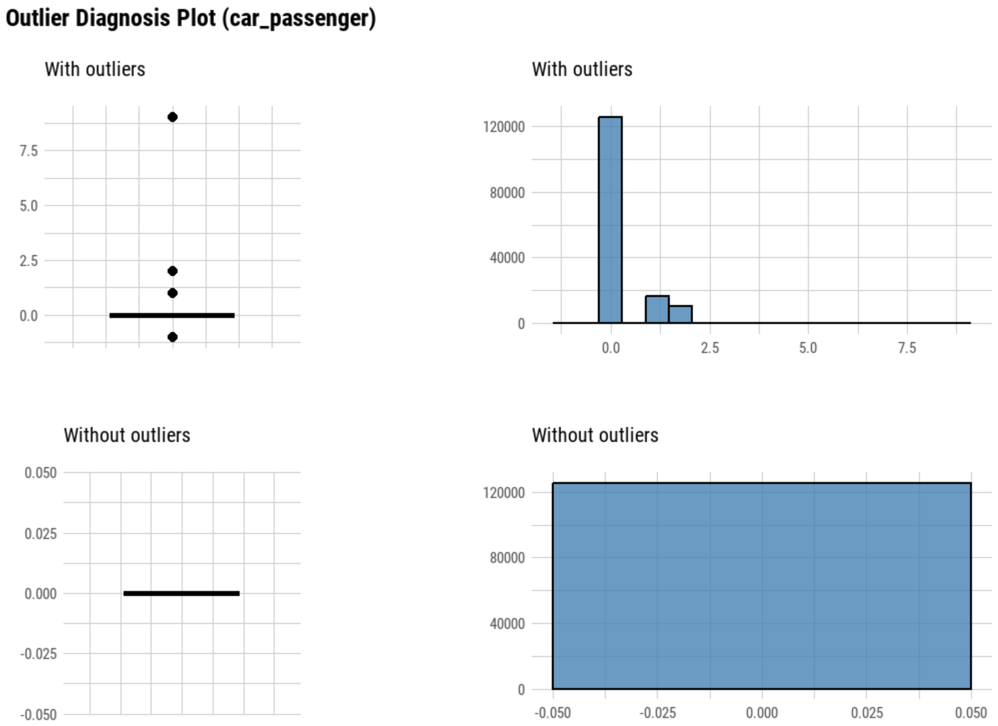
Table 13: casualty_reference



variable: car_passenger

Measures	Values
Outliers count	27,623
Outliers ratio (%)	18.04%
Mean of outliers	1.401875
Mean with outliers	0.2528369
Mean without outliers	0

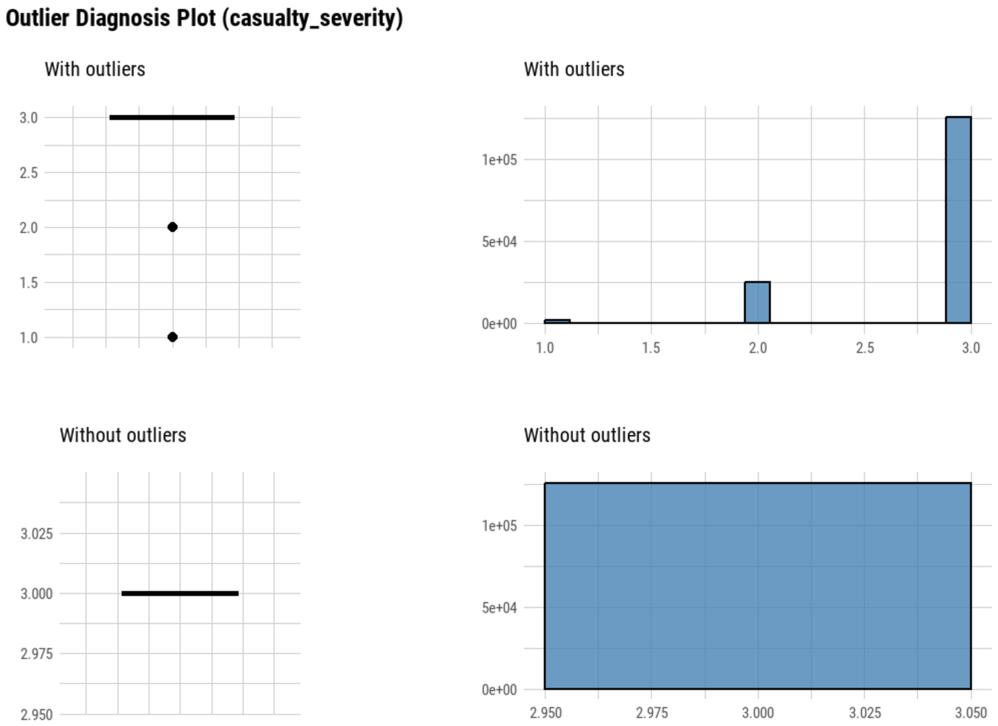
Table 13: car_passenger



variable: casualty_severity

Measures	Values
Outliers count	27,125
Outliers ratio (%)	17.71%
Mean of outliers	1.93541
Mean with outliers	2.811456
Mean without outliers	3

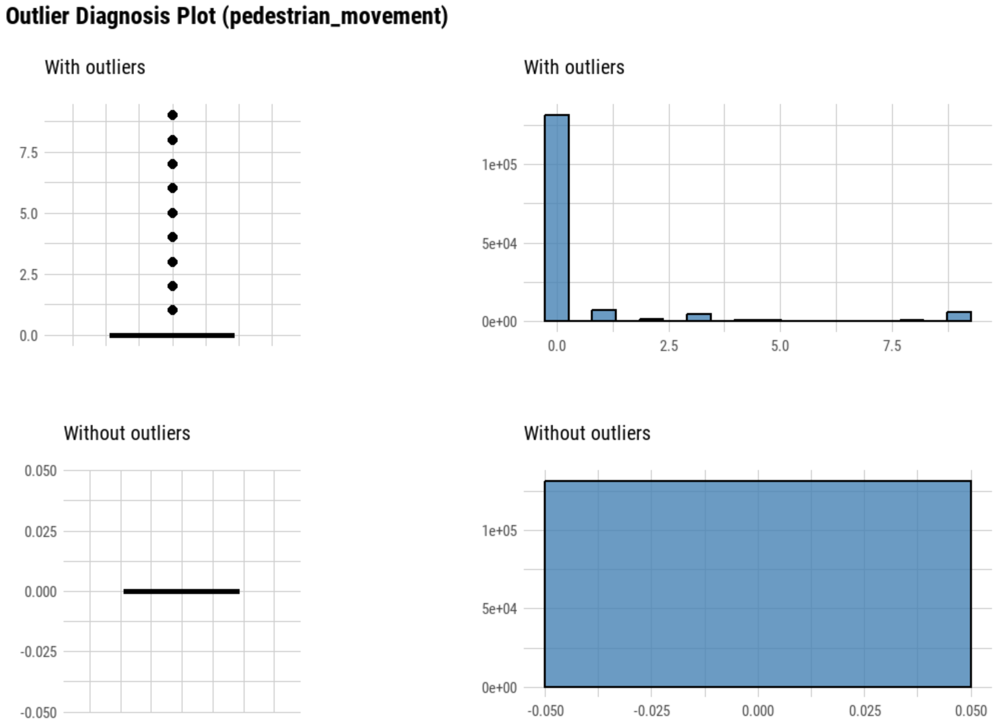
Table 13: casualty_severity



variable: pedestrian_movement

Measures	Values
Outliers count	21,772
Outliers ratio (%)	14.22%
Mean of outliers	4.220283
Mean with outliers	0.5999295
Mean without outliers	0

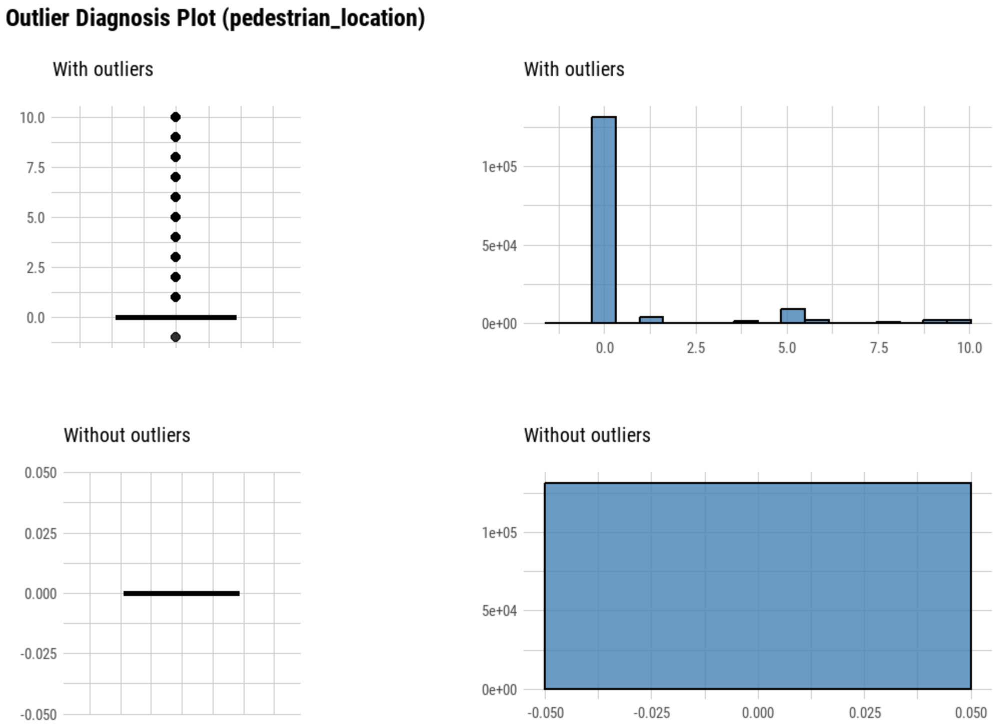
Table 13: pedestrian_movement



variable: pedestrian_location

Measures	Values
Outliers count	21,770
Outliers ratio (%)	14.21%
Mean of outliers	5.351814
Mean with outliers	0.7607112
Mean without outliers	0

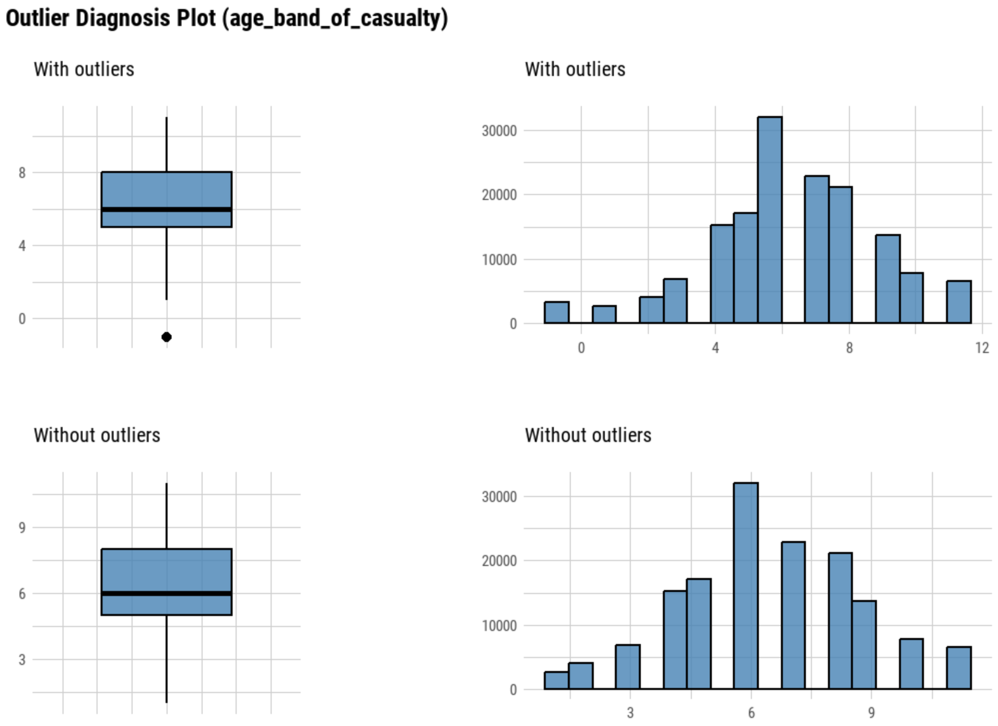
Table 13: pedestrian_location



variable: age_band_of_casualty

Measures	Values
Outliers count	3,255
Outliers ratio (%)	2.13%
Mean of outliers	-1
Mean with outliers	6.322347
Mean without outliers	6.481345

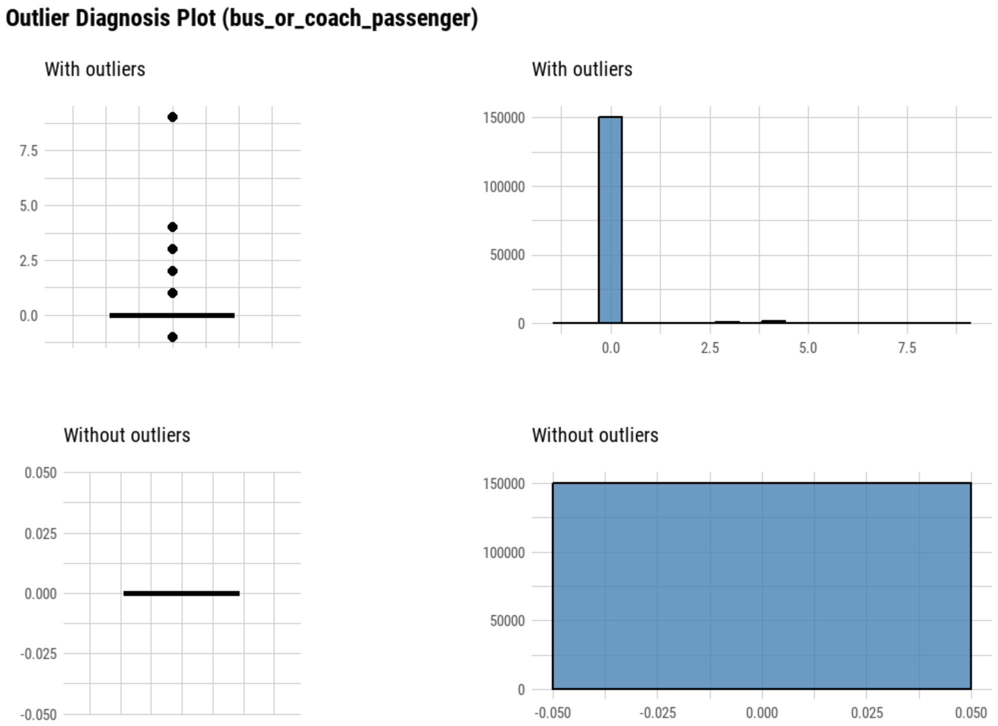
Table 13: age_band_of_casualty



variable: bus_or_coach_passenger

Measures	Values
Outliers count	2,824
Outliers ratio (%)	1.84%
Mean of outliers	3.479816
Mean with outliers	0.0641625
Mean without outliers	0

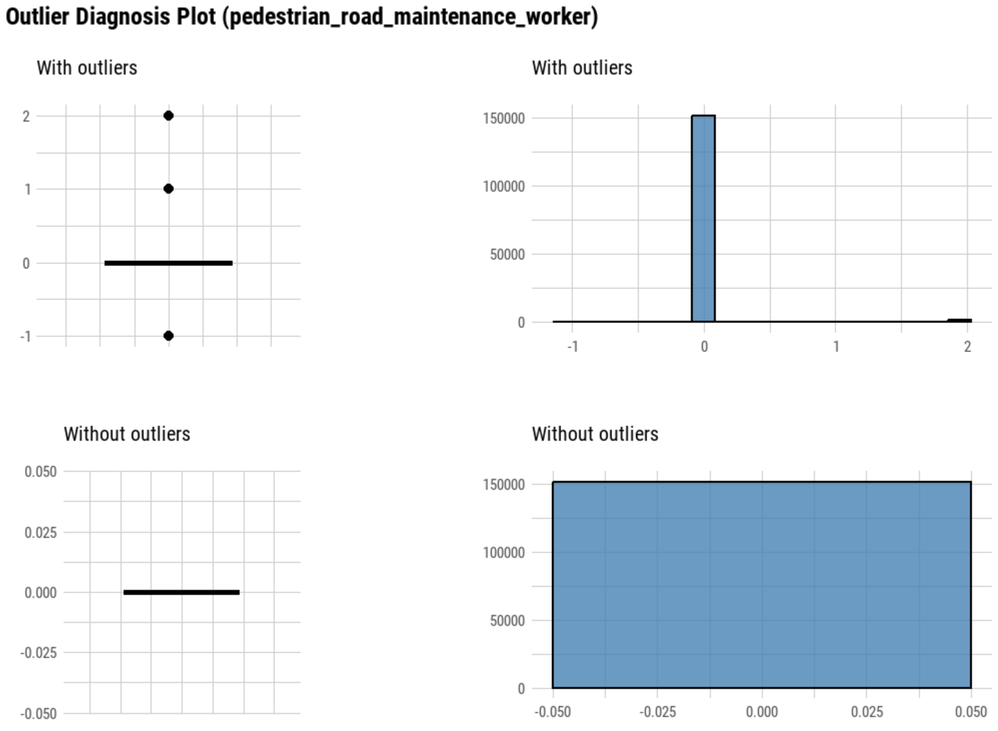
Table 13: bus_or_coach_passenger



variable: pedestrian_road_maintenance_worker

Measures	Values
Outliers count	1,422
Outliers ratio (%)	0.93%
Mean of outliers	1.786217
Mean with outliers	0.01658418
Mean without outliers	0

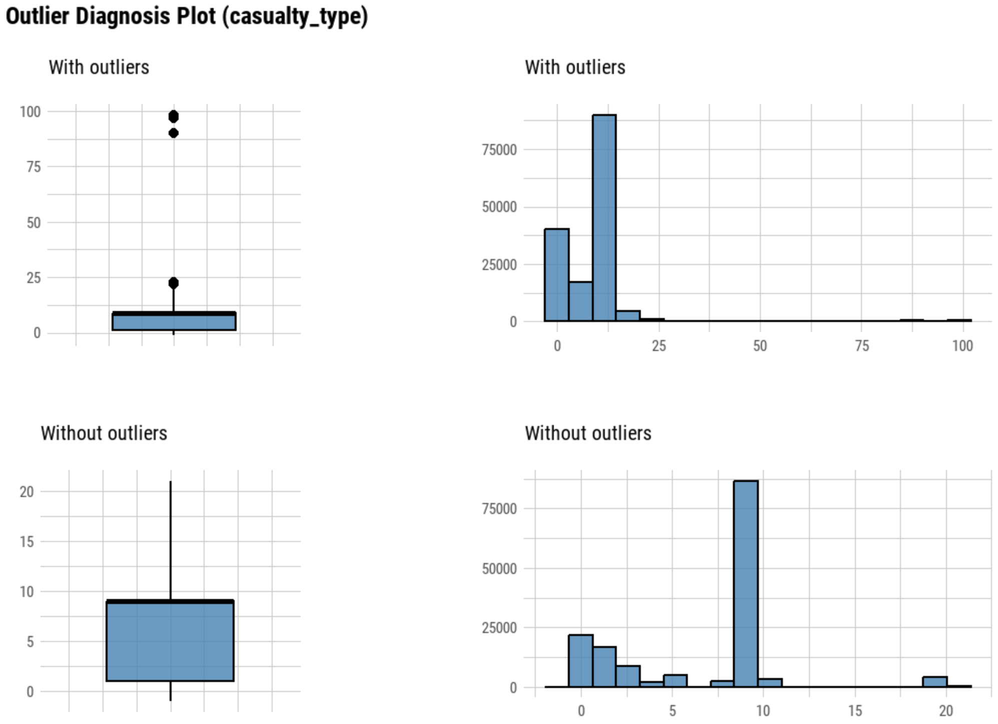
Table 13: pedestrian_road_maintenance_worker



variable: casualty_type

Measures	Values
Outliers count	1,124
Outliers ratio (%)	0.73%
Mean of outliers	76.57384
Mean with outliers	7.141148
Mean without outliers	6.627827

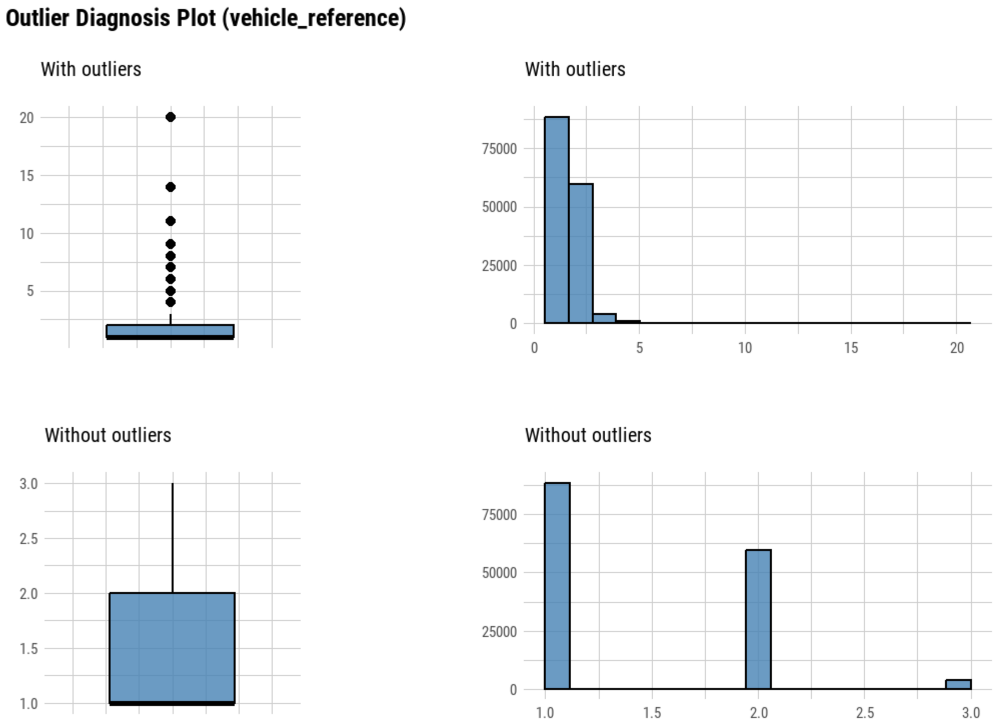
Table 13: casualty_type



variable: vehicle_reference

Measures	Values
Outliers count	991
Outliers ratio (%)	0.65%
Mean of outliers	4.526741
Mean with outliers	1.46484
Mean without outliers	1.444899

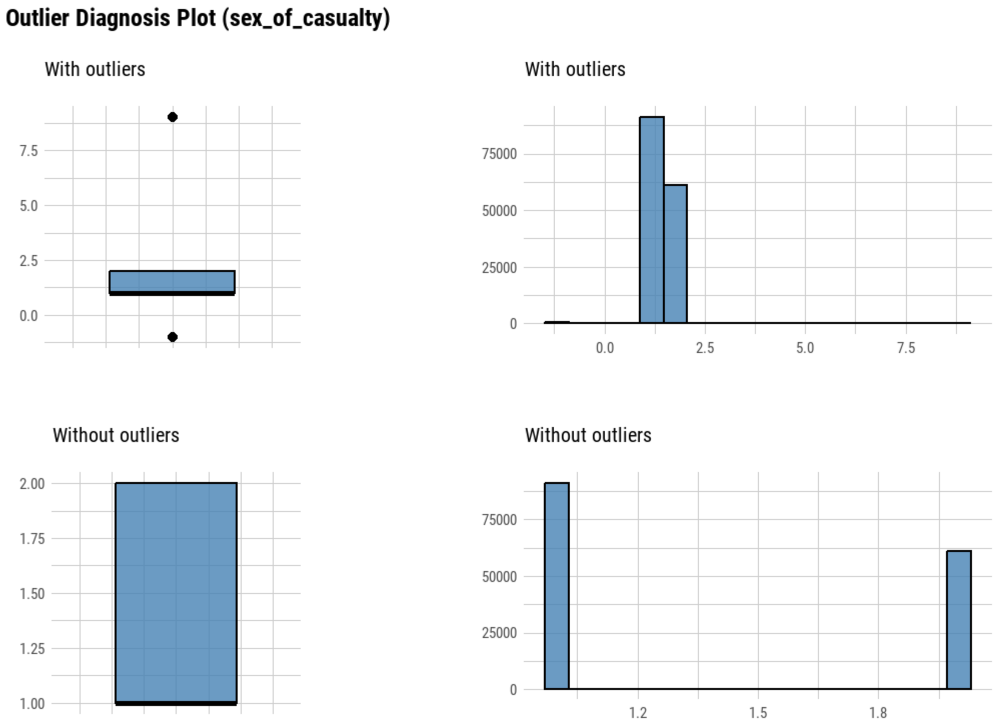
Table 13: vehicle_reference



variable: sex_of_casualty

Measures	Values
Outliers count	733
Outliers ratio (%)	0.48%
Mean of outliers	-0.8772169
Mean with outliers	1.390342
Mean without outliers	1.401247

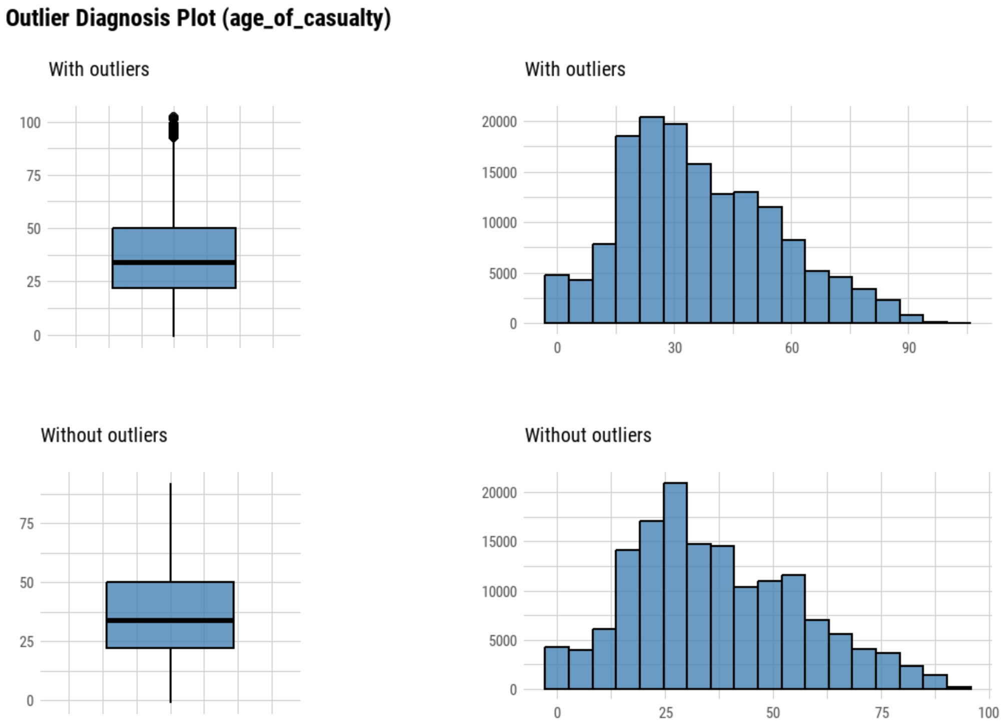
Table 13: sex_of_casualty



variable: age_of_casualty

Measures	Values
Outliers count	178
Outliers ratio (%)	0.12%
Mean of outliers	94.54494
Mean with outliers	36.93261
Mean without outliers	36.86558

Table 13: age_of_casualty


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