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	107,535
size	variables	36
size	values	3,871,260
size	memory size (MB)	56
duplicated	duplicate observation	0
missing	complete observation	107,535
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	0
data type	integers	25
data type	factors/ordered	0
data type	characters	11
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	107,535 / 107,535 (100%)
job	created	2022-11-28 22:51:11
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
19	21	1

Table 3: Summary of warnings

warnings	status	recommand
accident_index has high(1.00) cardinality, Maybe identifier	cardinality	check
accident_reference has high(1.00) cardinality, Maybe identifier	cardinality	check
accident_year has constant value “2019”	cardinality	remove
accident_severity has a low cardinality. 3 (0%) distinct values	cardinality	judgement
pedestrian_crossing_human_control has a low cardinality. 5 (0%) distinct values	cardinality	judgement
light_conditions has a low cardinality. 5 (0%) distinct values	cardinality	judgement
urban_or_rural_area has a low cardinality. 3 (0%) distinct values	cardinality	judgement
did_police_officer_attend_scene_of_accident has a low cardinality. 3 (0%) distinct values	cardinality	judgement
trunk_road_flag has a low cardinality. 2 (0%) distinct values	cardinality	judgement
carriageway_hazards has 104,094 (96.8%) zeros	zero	check
special_conditions_at_site has 103,392 (96.15%) zeros	zero	check
pedestrian_crossing_human_control has 102,329 (95.16%) zeros	zero	check
pedestrian_crossing_physical_facilities has 81,656 (75.93%) zeros	zero	check
second_road_number has 46,851 (43.57%) zeros	zero	check
junction_detail has 43,947 (40.87%) zeros	zero	check
first_road_number has 43,048 (40.03%) zeros	zero	check
junction_control has 4 (0%) zeros	zero	check
junction_control has 44,305 (41.2%) negatives	negative	check
second_road_class has 44,039 (40.95%) negatives	negative	check
second_road_number has 44,039 (40.95%) negatives	negative	check
road_surface_conditions has 297 (0.28%) negatives	negative	check
special_conditions_at_site has 237 (0.22%) negatives	negative	check

Table 4: Warnings in dataset and variables

	warnings	status	recommand
	warnings	**status**	**recommand**
23	carriageway_hazards has 227 (0.21%) negatives	negative	check
24	pedestrian_crossing_human_control has 157 (0.15%) negatives	negative	check
25	pedestrian_crossing_physical_facilities has 143 (0.13%) negatives	negative	check
26	speed_limit has 79 (0.07%) negatives	negative	check
27	road_type has 29,882 (27.79%) outliers	outlier	judgement
28	pedestrian_crossing_physical_facilities has 25,879 (24.07%) outliers	outlier	judgement
29	weather_conditions has 22,868 (21.27%) outliers	outlier	judgement
30	accident_severity has 21,712 (20.19%) outliers	outlier	judgement
31	number_of_casualties has 21,698 (20.18%) outliers	outlier	judgement
32	speed_limit has 18,602 (17.3%) outliers	outlier	judgement
33	first_road_number has 16,676 (15.51%) outliers	outlier	judgement
34	second_road_number has 16,410 (15.26%) outliers	outlier	judgement
35	junction_detail has 8,958 (8.33%) outliers	outlier	judgement
36	trunk_road_flag has 7,905 (7.35%) outliers	outlier	judgement
37	pedestrian_crossing_human_control has 5,206 (4.84%) outliers	outlier	judgement
38	special_conditions_at_site has 4,143 (3.85%) outliers	outlier	judgement
39	carriageway_hazards has 3,441 (3.2%) outliers	outlier	judgement
40	number_of_vehicles has 2,544 (2.37%) outliers	outlier	judgement
41	road_surface_conditions has 2,496 (2.32%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables (continued)

Variables

variables	types	missing	cardinality	zero	minus	outlier
accident_index	character		identifier			
accident_year	integer		constant			
accident_reference	character		identifier			
location_easting_osgr	character		> high			
location_northing_osgr	character		> high			
longitude	character		> high			
latitude	character		> high			
police_force	integer					
accident_severity	integer		< low			X
number_of_vehicles	integer					X
number_of_casualties	integer					X
date	character					
day_of_week	integer					
time	character					
local_authority_district	integer					
local_authority_ons_district	character					
local_authority_highway	character					
first_road_class	integer					
first_road_number	integer			X		X
road_type	integer					X
speed_limit	integer				X	X
junction_detail	integer			X		X
junction_control	integer			X	X	
second_road_class	integer				X	
second_road_number	integer			X	X	X

Table 5: List of variables diagnosis

variables	types	missing	cardinality	zero	minus	outlier
variables	types	missing	cardinality	zero	minus	outlier
pedestrian_crossing_human_control	integer		< low	X	X	X
pedestrian_crossing_physical_facilities	integer			X	X	X
light_conditions	integer		< low			
weather_conditions	integer					X
road_surface_conditions	integer				X	X
special_conditions_at_site	integer			X	X	X
carriageway_hazards	integer			X	X	X
urban_or_rural_area	integer		< low			
did_police_officer_attend_scene_of_accident	integer		< low			
trunk_road_flag	integer		< low			X
lsoa_of_accident_location	character					

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

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	107,535	100%	identifier	Use as ID
accident_reference	character	107,535	100%	identifier	Use as ID
location_easting_osgr	character	84,383	78.5%	high cardinality	Judgment
location_northing_osgr	character	85,719	79.7%	high cardinality	Judgment
longitude	character	105,115	97.7%	high cardinality	Judgment
latitude	character	103,851	96.6%	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
accident_severity	integer	3	0%	low cardinality	Judgment
pedestrian_crossing_human_control	integer	5	0%	low cardinality	Judgment
light_conditions	integer	5	0%	low cardinality	Judgment
urban_or_rural_area	integer	3	0%	low cardinality	Judgment
did_police_officer_attend_scene_of_accident	integer	3	0%	low cardinality	Judgment
trunk_road_flag	integer	2	0%	low cardinality	Judgment

Table 7: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
accident_index	2019010128300	1	0.0
accident_index	2019010152270	1	0.0
accident_index	2019010155191	1	0.0
accident_index	2019010155192	1	0.0
accident_index	2019010155194	1	0.0
accident_index	2019010155195	1	0.0
accident_index	2019010155196	1	0.0
accident_index	2019010155198	1	0.0
accident_index	2019010155206	1	0.0
accident_index	2019010155207	1	0.0
accident_index	Other levles	107,525	100.0
accident_reference	010128300	1	0.0
accident_reference	010152270	1	0.0
accident_reference	010155191	1	0.0
accident_reference	010155192	1	0.0
accident_reference	010155194	1	0.0
accident_reference	010155195	1	0.0
accident_reference	010155196	1	0.0
accident_reference	010155198	1	0.0
accident_reference	010155206	1	0.0
accident_reference	010155207	1	0.0
accident_reference	Other levles	107,525	100.0
date	04/12/2019	463	0.4
date	29/11/2019	440	0.4
date	20/09/2019	423	0.4

Table 8: Top 10 levels of categorical variables

variables	levels	freq	ratio (%)
date	27/11/2019	415	0.4
date	20/12/2019	413	0.4
date	02/12/2019	407	0.4
date	08/11/2019	405	0.4
date	12/12/2019	401	0.4
date	17/09/2019	400	0.4
date	22/01/2019	397	0.4
date	Other levles	103,371	96.1
latitude	NULL	25	0.0
latitude	51.450232	6	0.0
latitude	51.38271	4	0.0
latitude	51.473943	4	0.0
latitude	51.490723	4	0.0
latitude	51.509646	4	0.0
latitude	51.523918	4	0.0
latitude	51.546307	4	0.0
latitude	51.557147	4	0.0
latitude	51.565088	4	0.0
latitude	Other levles	107,472	99.9
local_authority_highway	E10000016	3,619	3.4
local_authority_highway	E10000030	2,964	2.8
local_authority_highway	E08000025	2,623	2.4
local_authority_highway	E10000012	2,385	2.2
local_authority_highway	E10000014	2,385	2.2
local_authority_highway	E10000017	2,306	2.1
local_authority_highway	E10000032	1,983	1.8
local_authority_highway	E10000015	1,960	1.8
local_authority_highway	E10000019	1,893	1.8
local_authority_highway	E10000020	1,648	1.5

Table 8: Top 10 levels of categorical variables (continued)

variables	levels	freq	ratio (%)
local_authority_highway	Other levles	83,769	77.9
local_authority_ons_district	E08000025	2,623	2.4
local_authority_ons_district	E09000033	1,521	1.4
local_authority_ons_district	E08000035	1,451	1.3
local_authority_ons_district	E09000022	1,191	1.1
local_authority_ons_district	E06000052	1,131	1.1
local_authority_ons_district	E09000030	1,131	1.1
local_authority_ons_district	E09000028	1,096	1.0
local_authority_ons_district	E09000009	983	0.9
local_authority_ons_district	E06000023	967	0.9
local_authority_ons_district	E09000008	955	0.9
local_authority_ons_district	Other levles	94,486	87.9
location_easting_osgr	NULL	25	0.0
location_easting_osgr	533465	8	0.0
location_easting_osgr	526143	7	0.0
location_easting_osgr	527553	7	0.0
location_easting_osgr	530857	7	0.0
location_easting_osgr	531125	7	0.0
location_easting_osgr	531574	7	0.0
location_easting_osgr	532537	7	0.0
location_easting_osgr	533541	7	0.0
location_easting_osgr	533579	7	0.0
location_easting_osgr	Other levles	107,446	99.9
location_northing_osgr	NULL	25	0.0
location_northing_osgr	180819	9	0.0
location_northing_osgr	181092	9	0.0
location_northing_osgr	182780	9	0.0
location_northing_osgr	180979	8	0.0
location_northing_osgr	180992	8	0.0

Table 8: Top 10 levels of categorical variables (continued)

variables	levels	freq	ratio (%)
location_northing_osgr	181066	8	0.0
location_northing_osgr	181087	8	0.0
location_northing_osgr	181131	8	0.0
location_northing_osgr	181323	8	0.0
location_northing_osgr	Other levles	107,435	99.9
longitude	NULL	25	0.0
longitude	-0.080261	4	0.0
longitude	-1.176676	4	0.0
longitude	-1.47504	4	0.0
longitude	-1.524603	4	0.0
longitude	-1.594158	4	0.0
longitude	-1.909639	4	0.0
longitude	-2.715409	4	0.0
longitude	-0.009487	3	0.0
longitude	-0.018565	3	0.0
longitude	Other levles	107,476	99.9
lsoa_of_accident_location	E01032739	194	0.2
lsoa_of_accident_location	E01004736	119	0.1
lsoa_of_accident_location	E01033595	93	0.1
lsoa_of_accident_location	E01002444	72	0.1
lsoa_of_accident_location	E01033708	67	0.1
lsoa_of_accident_location	E01004689	59	0.1
lsoa_of_accident_location	E01003482	58	0.1
lsoa_of_accident_location	E01004733	58	0.1
lsoa_of_accident_location	E01001667	56	0.1
lsoa_of_accident_location	E01004763	56	0.1
lsoa_of_accident_location	Other levles	106,703	99.2
time	17:00	1,011	0.9
time	17:30	935	0.9

Table 8: Top 10 levels of categorical variables (continued)

variables	levels	freq	ratio (%)
time	16:00	911	0.8
time	18:00	893	0.8
time	16:30	890	0.8
time	15:30	872	0.8
time	15:00	805	0.7
time	18:30	770	0.7
time	08:30	769	0.7
time	19:00	733	0.7
time	Other levles	98,946	92.0

Table 8: Top 10 levels of categorical variables (continued)

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									
police_force	1	4	23.34	20	43	55	0	0	

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
accident_severity	1	3	2.79	3	3	3	0	0	21,712
number_of_vehicles	1	1	1.85	2	2	17	0	0	2,544
number_of_casualties	1	1	1.30	1	1	52	0	0	21,698
day_of_week	1	2	4.11	4	6	7	0	0	

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
local_authority_district	1	70	278.46	286	475	647	0	0	

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
first_road_class	1	3	4.20	4	6	6	0	0	

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
first_road_number	0	0	755.67	27	503	6,918	43,048	0	16,676
road_type	1	6	5.21	6	6	9	0	0	29,882
speed_limit	-1	30	36.38	30	40	70	0	79	18,602
junction_detail	0	0	3.62	2	3	99	43,947	0	8,958
junction_control	-1	-1	1.83	2	4	9	4	44,305	

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
0									
second_road_class	-1	-1	2.67	3	6	6	0	44,039	0
second_road_number	-1	-1	225.20	0	0	7,503	46,851	44,039	16,410
pedestrian_crossing_human_control	-1	0	0.31	0	0	9	102,329	157	5,206
pedestrian_crossing_physical_facilities	-1	0	1.17	0	0	9	81,656	143	25,879
light_conditions	1	1	2.05	1	4	7	0	0	0
weather_conditions	1	1	1.66	1	1	9	0	0	22,868
road_surface_conditions	-1	1	1.38	1	2	9	0	297	2,496
special_conditions_at_site	-1	0	0.22	0	0	9	103,392	237	4,143
carriageway_hazards	-1	0	0.16	0	0	9	104,094	227	3,441
urban_or_rural_area	1	1	1.31	1	2	3	0	0	0
did_police_officer_attend_scene_of_accident	1	1	1.41	1	2	3	0	0	0
trunk_road_flag	1	2	1.93	2	2	2	0	0	7,905

Zero Values

variables	min	median	max	zero	zero (%)
carriageway_hazards	-1	0	9	104,094	96.8
special_conditions_at_site	-1	0	9	103,392	96.1
pedestrian_crossing_human_control	-1	0	9	102,329	95.2
pedestrian_crossing_physical_facilities	-1	0	9	81,656	75.9
second_road_number	-1	0	7,503	46,851	43.6
junction_detail	0	2	99	43,947	40.9
first_road_number	0	27	6,918	43,048	40.0
junction_control	-1	2	9	4	0.0

Table 10: List of numerical diagnosis (zero)

Negative Values

variables	min	median	max	minus	minus (%)
junction_control	-1	2	9	44,305	41.2
second_road_class	-1	3	6	44,039	41.0
second_road_number	-1	0	7,503	44,039	41.0
road_surface_conditions	-1	1	9	297	0.3
special_conditions_at_site	-1	0	9	237	0.2
carriageway_hazards	-1	0	9	227	0.2
pedestrian_crossing_human_control	-1	0	9	157	0.1
pedestrian_crossing_physical_facilities	-1	0	9	143	0.1
speed_limit	-1	30	70	79	0.1

Table 11: List of numerical diagnosis (minus)

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
road_type	1	6	9	29,882	27.8
pedestrian_crossing_physical_facilities	-1	0	9	25,879	24.1
weather_conditions	1	1	9	22,868	21.3
accident_severity	1	3	3	21,712	20.2
number_of_casualties	1	1	52	21,698	20.2
speed_limit	-1	30	70	18,602	17.3
first_road_number	0	27	6,918	16,676	15.5
second_road_number	-1	0	7,503	16,410	15.3
junction_detail	0	2	99	8,958	8.3
trunk_road_flag	1	2	2	7,905	7.4
pedestrian_crossing_human_control	-1	0	9	5,206	4.8
special_conditions_at_site	-1	0	9	4,143	3.9
carriageway_hazards	-1	0	9	3,441	3.2
number_of_vehicles	1	2	17	2,544	2.4
road_surface_conditions	-1	1	9	2,496	2.3

Table 12: Diagnosis of numerical variable outliers

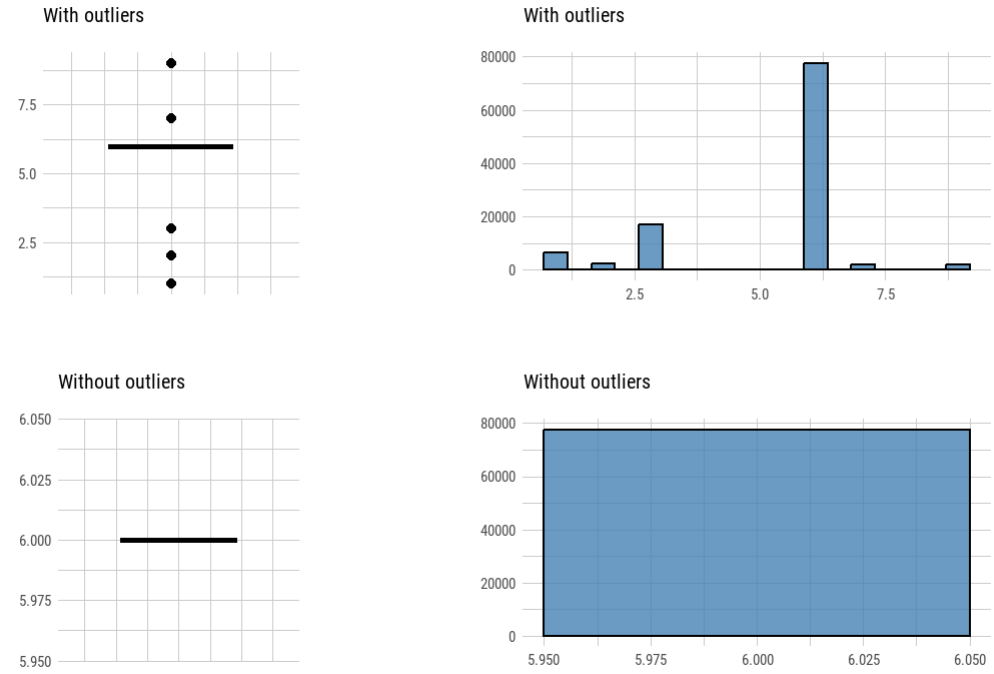
Individual Outliers

variable: road_type

Measures	Values
Outliers count	29,882
Outliers ratio (%)	27.79%
Mean of outliers	3.157787
Mean with outliers	5.210201
Mean without outliers	6

Table 13: road_type

Outlier Diagnosis Plot (road_type)

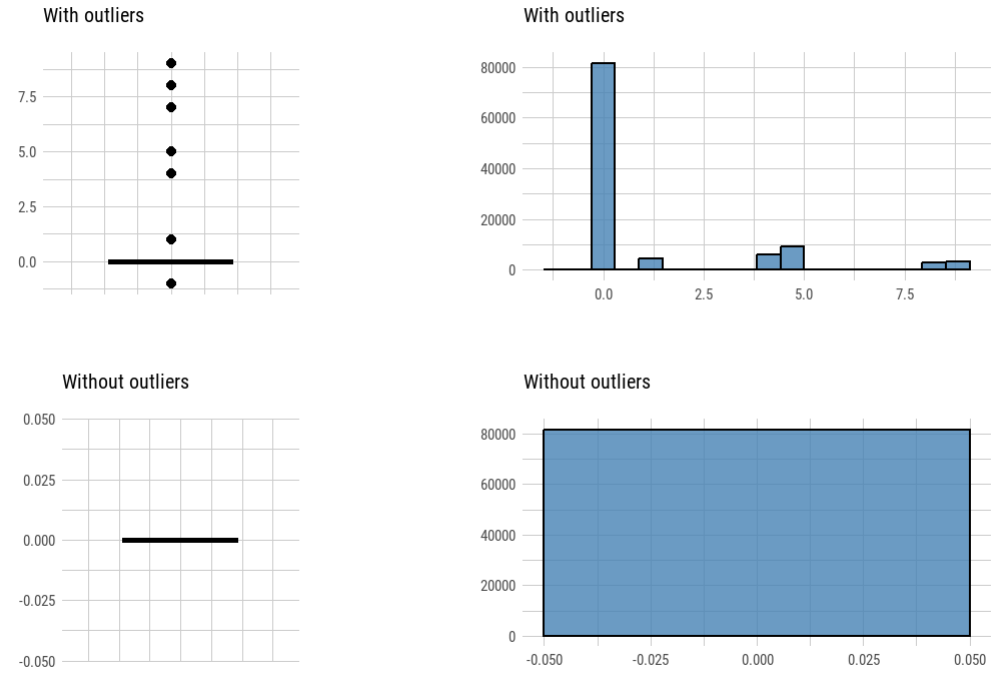


variable: pedestrian_crossing_physical_facilities

Measures	Values
Outliers count	25,879
Outliers ratio (%)	24.07%
Mean of outliers	4.842498
Mean with outliers	1.165379
Mean without outliers	0

Table 13: pedestrian_crossing_physical_facilities

Outlier Diagnosis Plot (pedestrian_crossing_physical_facilities)

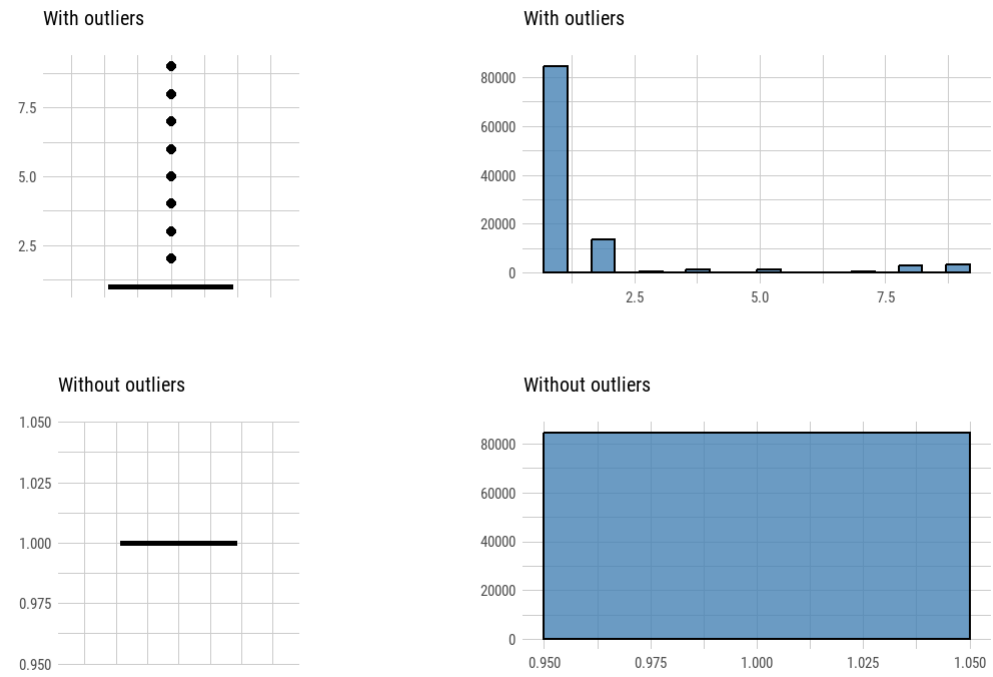


variable: weather_conditions

Measures	Values
Outliers count	22,868
Outliers ratio (%)	21.27%
Mean of outliers	4.10788
Mean with outliers	1.66091
Mean without outliers	1

Table 13: weather_conditions

Outlier Diagnosis Plot (weather_conditions)

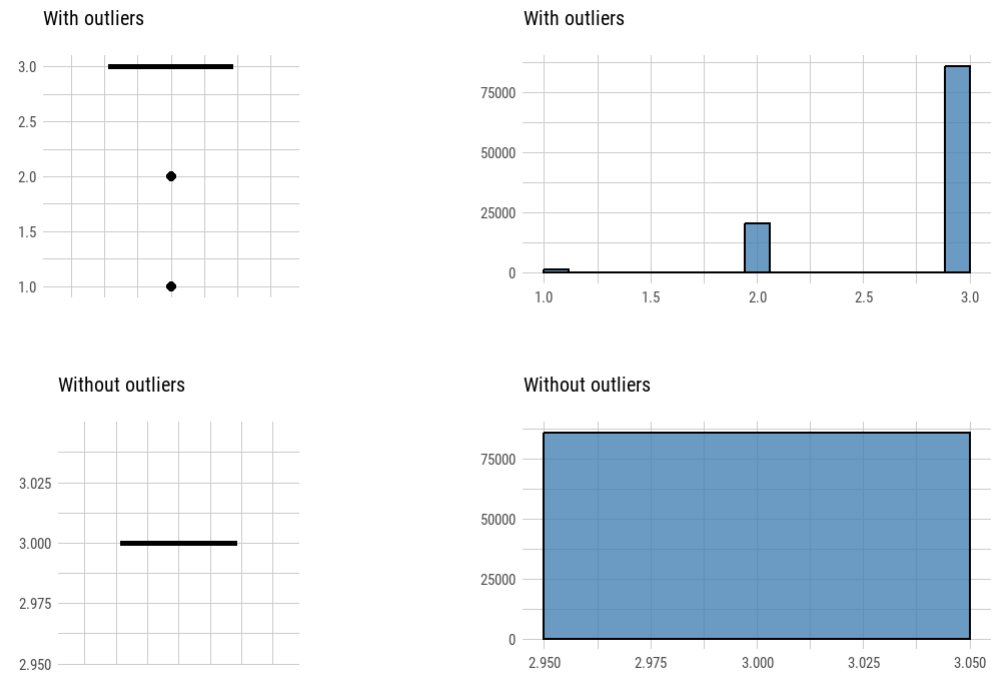


variable: accident_severity

Measures	Values
Outliers count	21,712
Outliers ratio (%)	20.19%
Mean of outliers	1.935381
Mean with outliers	2.785047
Mean without outliers	3

Table 13: accident_severity

Outlier Diagnosis Plot (accident_severity)

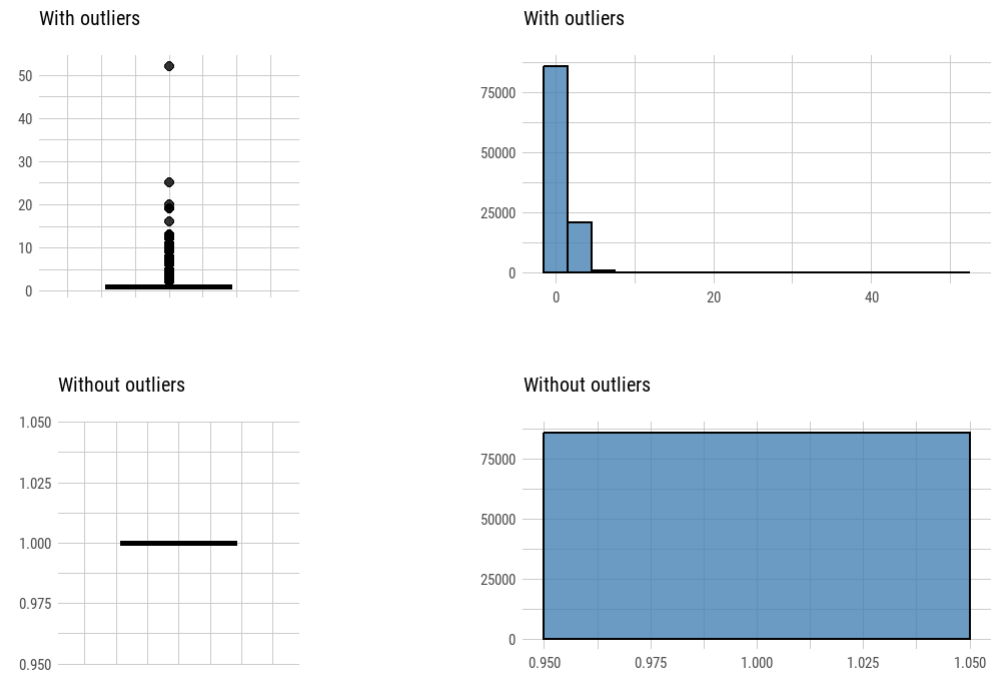


variable: number_of_casualties

Measures	Values
Outliers count	21,698
Outliers ratio (%)	20.18%
Mean of outliers	2.486036
Mean with outliers	1.299847
Mean without outliers	1

Table 13: number_of_casualties

Outlier Diagnosis Plot (number_of_casualties)

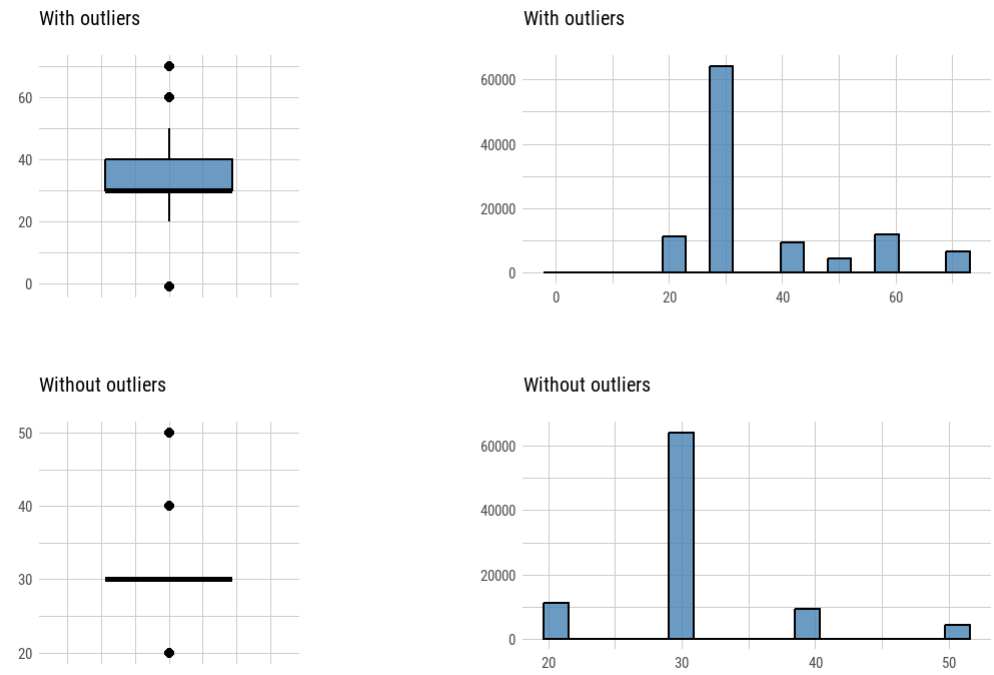


variable: speed_limit

Measures	Values
Outliers count	18,602
Outliers ratio (%)	17.3%
Mean of outliers	63.23358
Mean with outliers	36.37905
Mean without outliers	30.76192

Table 13: speed_limit

Outlier Diagnosis Plot (speed_limit)

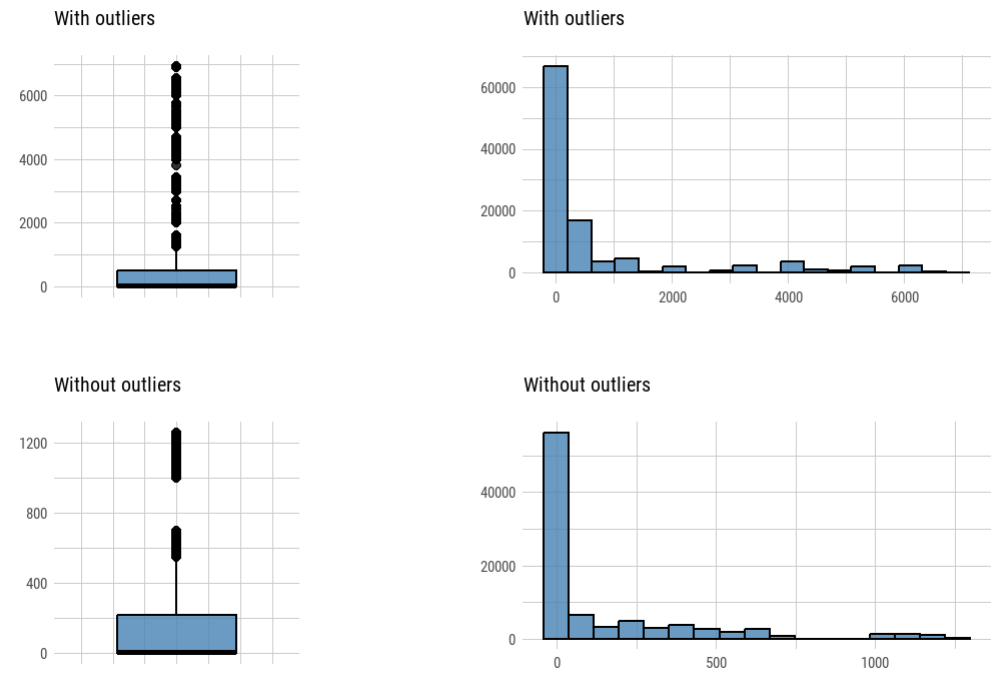


variable: first_road_number

Measures	Values
Outliers count	16,676
Outliers ratio (%)	15.51%
Mean of outliers	4023.146
Mean with outliers	755.6657
Mean without outliers	155.9619

Table 13: first_road_number

Outlier Diagnosis Plot (first_road_number)

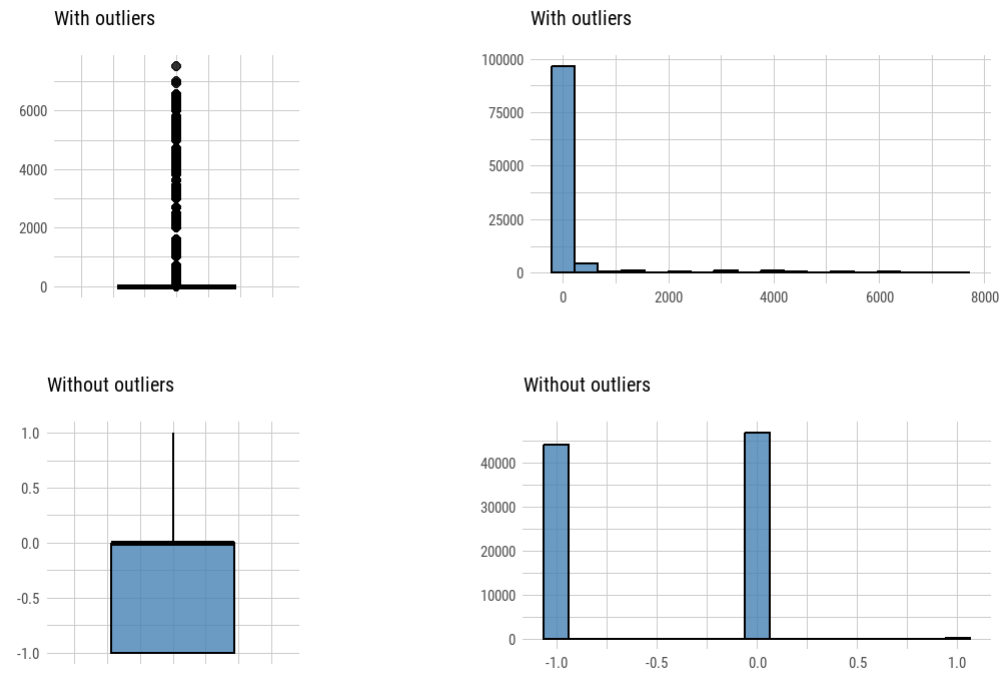


variable: second_road_number

Measures	Values
Outliers count	16,410
Outliers ratio (%)	15.26%
Mean of outliers	1478.393
Mean with outliers	225.1977
Mean without outliers	-0.4807023

Table 13: second_road_number

Outlier Diagnosis Plot (second_road_number)

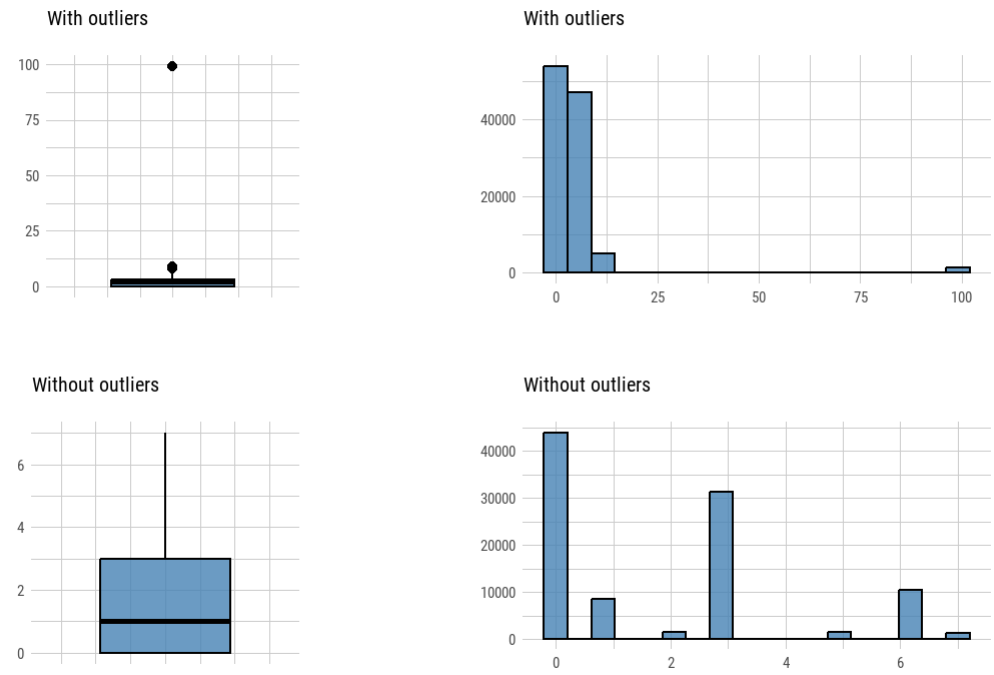


variable: junction_detail

Measures	Values
Outliers count	8,958
Outliers ratio (%)	8.33%
Mean of outliers	22.74793
Mean with outliers	3.617864
Mean without outliers	1.879455

Table 13: junction_detail

Outlier Diagnosis Plot (junction_detail)

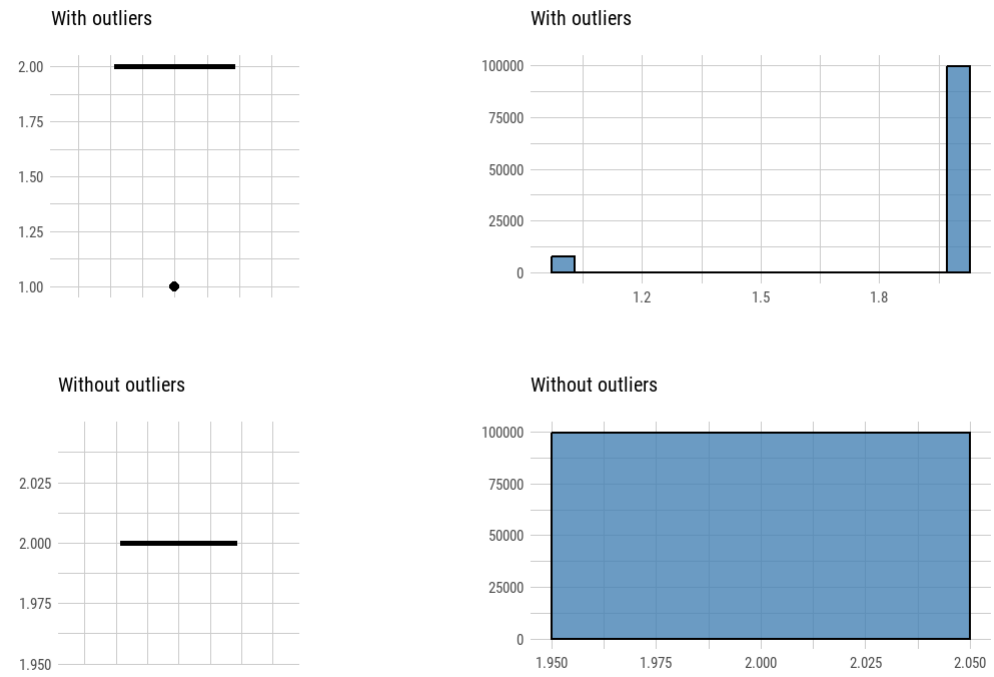


variable: trunk_road_flag

Measures	Values
Outliers count	7,905
Outliers ratio (%)	7.35%
Mean of outliers	1
Mean with outliers	1.926489
Mean without outliers	2

Table 13: trunk_road_flag

Outlier Diagnosis Plot (trunk_road_flag)

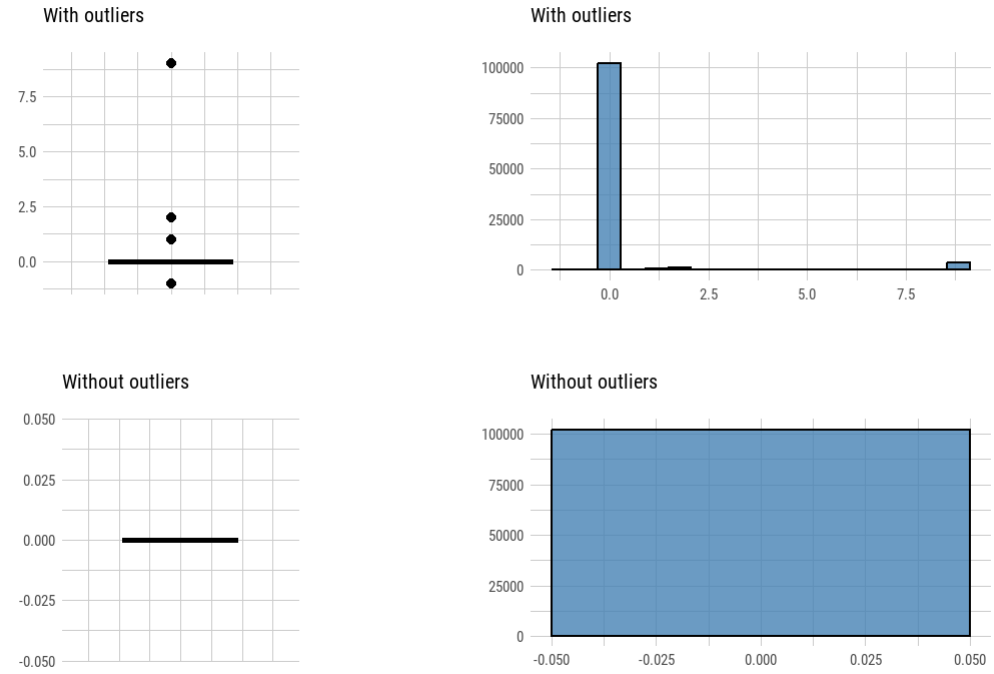


variable: pedestrian_crossing_human_control

Measures	Values
Outliers count	5,206
Outliers ratio (%)	4.84%
Mean of outliers	6.494045
Mean with outliers	0.3143907
Mean without outliers	0

Table 13: pedestrian_crossing_human_control

Outlier Diagnosis Plot (pedestrian_crossing_human_control)

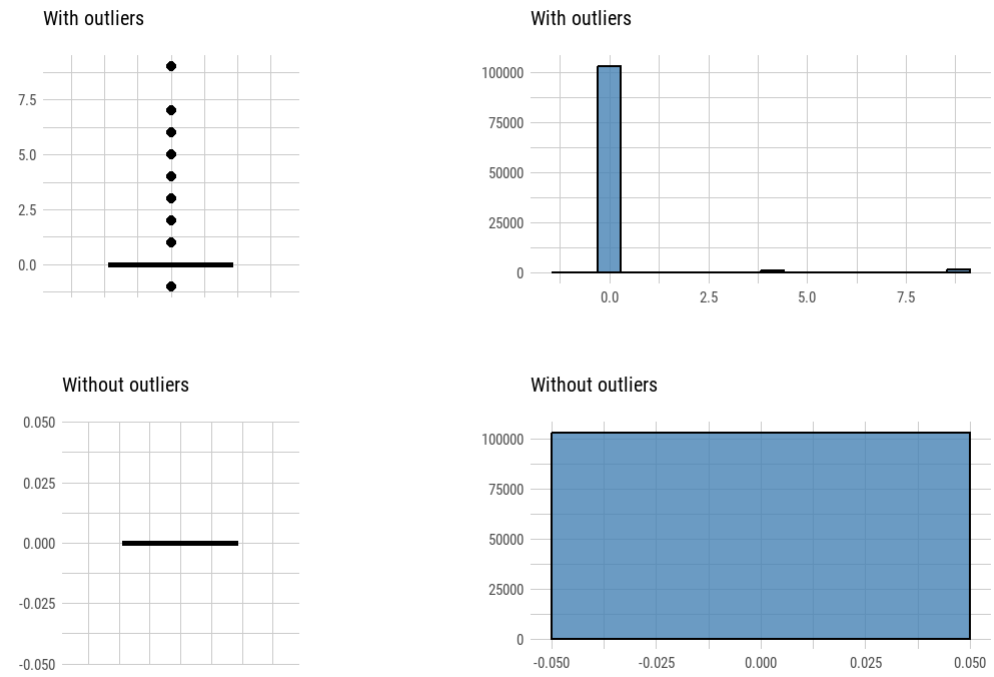


variable: special_conditions_at_site

Measures	Values
Outliers count	4,143
Outliers ratio (%)	3.85%
Mean of outliers	5.69008
Mean with outliers	0.2192216
Mean without outliers	0

Table 13: special_conditions_at_site

Outlier Diagnosis Plot (special_conditions_at_site)

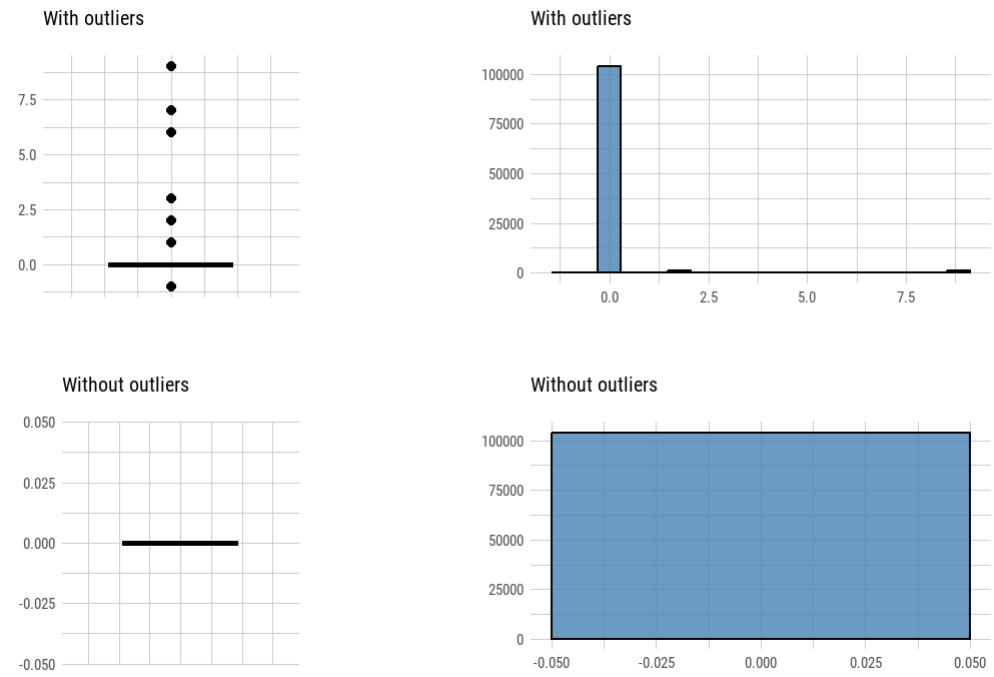


variable: carriageway_hazards

Measures	Values
Outliers count	3,441
Outliers ratio (%)	3.2%
Mean of outliers	5.132229
Mean with outliers	0.1642256
Mean without outliers	0

Table 13: carriageway_hazards

Outlier Diagnosis Plot (carriageway_hazards)

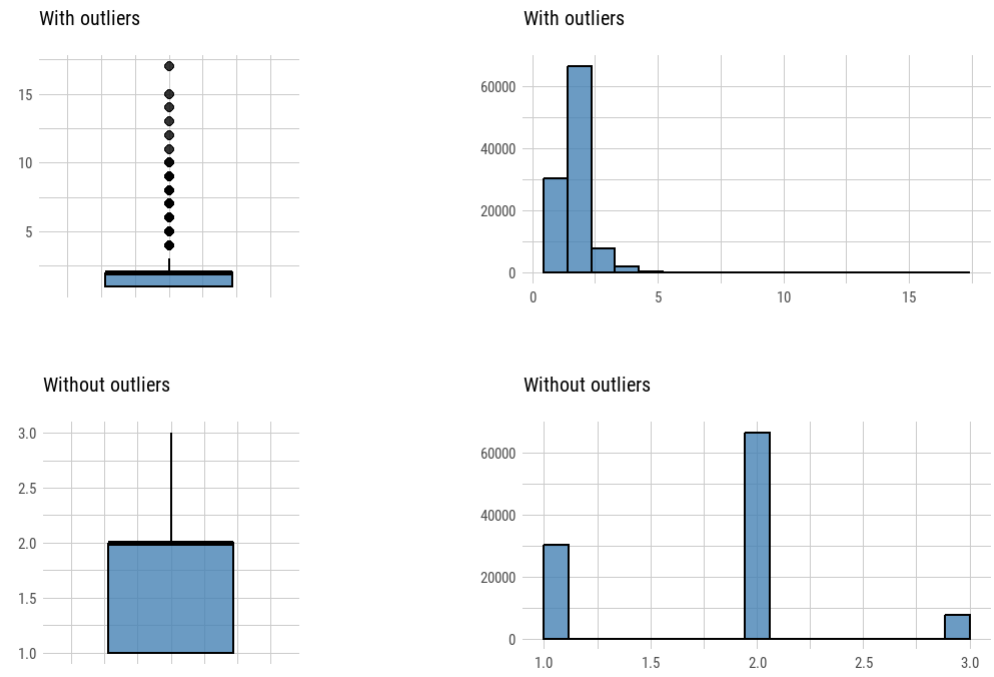


variable: number_of_vehicles

Measures	Values
Outliers count	2,544
Outliers ratio (%)	2.37%
Mean of outliers	4.438679
Mean with outliers	1.847668
Mean without outliers	1.784886

Table 13: number_of_vehicles

Outlier Diagnosis Plot (number_of_vehicles)



variable: road_surface_conditions

Measures	Values
Outliers count	2,496
Outliers ratio (%)	2.32%
Mean of outliers	5.567708
Mean with outliers	1.379393
Mean without outliers	1.279867

Table 13: road_surface_conditions

Outlier Diagnosis Plot (road_surface_conditions)