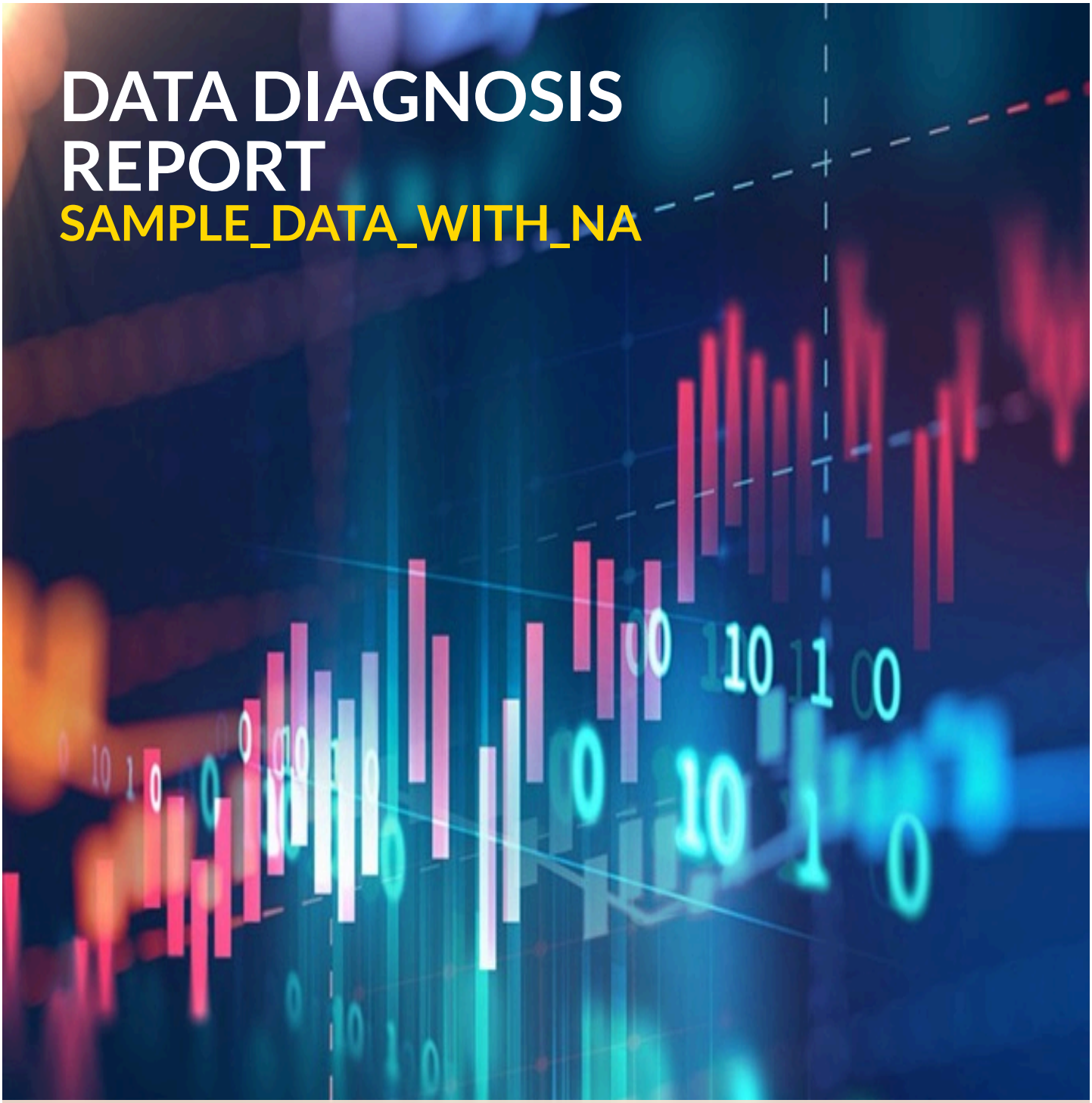


DATA DIAGNOSIS REPORT

SAMPLE_DATA_WITH_NA

Report Overview

This report was created for an overview quality diagnosis of ***sample_data_with_NA*** 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	1,000
size	variables	14
size	values	14,000
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	525
missing	missing observation	475
missing	missing variables	12
missing	missing values	600

division	metrics	value
data type	numerics	3
data type	integers	2
data type	factors/ordered	0
data type	characters	9
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	sample_data_with_NA
dataset	dataset type	data.frame
job	samples	1,000 / 1,000 (100%)
job	created	2024-04-04 11:49:09.04188
job	created by	dlookr

Table 2: Job informations

Warnings

checks	judgements	removes
0	12	0

Table 3: Summary of warnings

warnings	status	recommend
age has 50 (5%) missing values	missing	judgement
workclass has 50 (5%) missing values	missing	judgement
education has 50 (5%) missing values	missing	judgement
marital_status has 50 (5%) missing values	missing	judgement
occupation has 50 (5%) missing values	missing	judgement
relationship has 50 (5%) missing values	missing	judgement
race has 50 (5%) missing values	missing	judgement
sex has 50 (5%) missing values	missing	judgement
capital_gain has 50 (5%) missing values	missing	judgement
capital_loss has 50 (5%) missing values	missing	judgement
hours_per_week has 50 (5%) missing values	missing	judgement
native_country has 50 (5%) missing values	missing	judgement

Table 4: Warnings in dataset and variables

Variables

variables	types	missing	cardinality	zero	minus	outlier
age	numeric	X				
workclass	character	X				
education	character	X				
education_num	integer					
marital_status	character	X				
occupation	character	X				
relationship	character	X				
race	character	X				
sex	character	X				
capital_gain	numeric	X				
capital_loss	numeric	X				
hours_per_week	integer	X				
native_country	character	X				
income	character					

Table 5: List of variables diagnosis

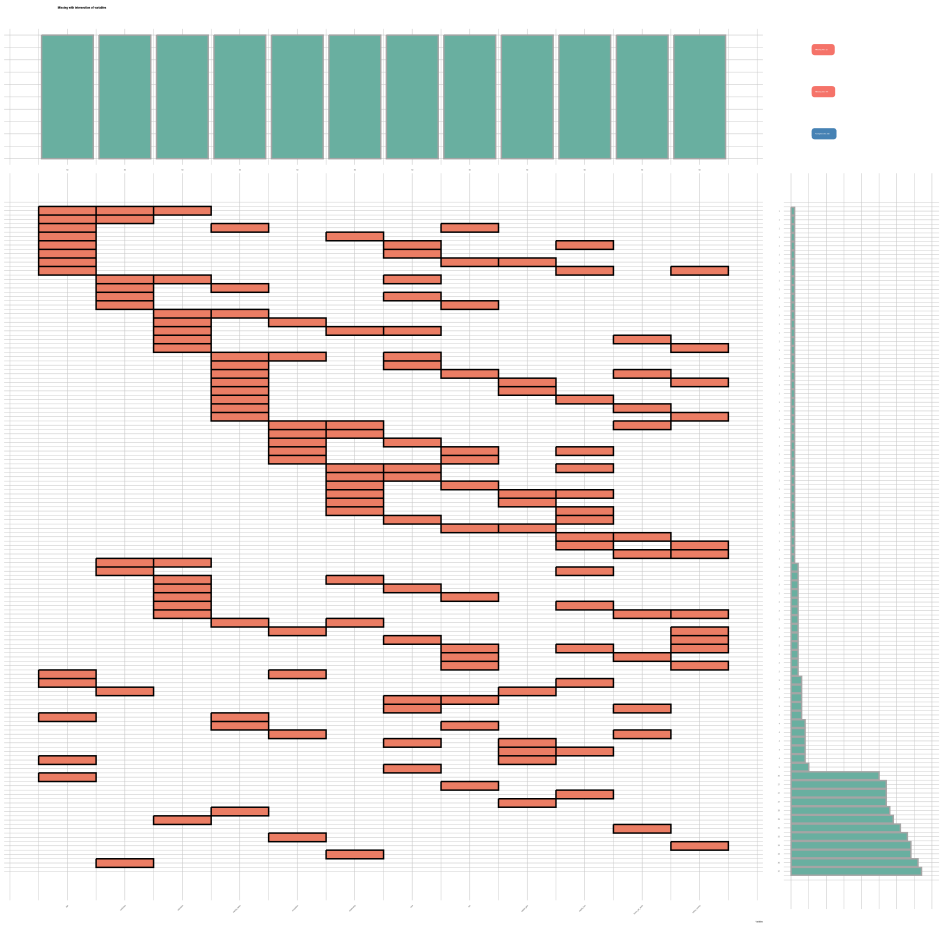
Missing Values

List of Missing Values

variables	missing_count	missing (%)	status	recommend
age	50	5%	Good	Delete or Imputation
workclass	50	5%	Good	Delete or Imputation
education	50	5%	Good	Delete or Imputation
marital_status	50	5%	Good	Delete or Imputation
occupation	50	5%	Good	Delete or Imputation
relationship	50	5%	Good	Delete or Imputation
race	50	5%	Good	Delete or Imputation
sex	50	5%	Good	Delete or Imputation
capital_gain	50	5%	Good	Delete or Imputation
capital_loss	50	5%	Good	Delete or Imputation
hours_per_week	50	5%	Good	Delete or Imputation
native_country	50	5%	Good	Delete or Imputation

Table 6: List of variables including missing values

Visualization



Unique Values

Categorical Vaiables

No variable with a high proportion greater than 0.5

Numerical Vaiables

No variable with unique data proportion less than 5

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
education	Masters	74	7.4
education	1st-4th	69	6.9
education	Assoc-voc	69	6.9
education	9th	68	6.8
education	5th-6th	64	6.4
education	Doctorate	63	6.3
education	11th	61	6.1
education	Bachelors	60	6.0
education	Some-college	60	6.0
education	HS-grad	59	5.9
education	Other levles	303	30.3
education	Missing	50	5.0
income	<=50K	737	73.7
income	>50K	263	26.3
marital_status	Never-married	148	14.8
marital_status	Married-spouse-absent	141	14.1
marital_status	Widowed	141	14.1
marital_status	Married-civ-spouse	139	13.9
marital_status	Divorced	136	13.6
marital_status	Separated	132	13.2
marital_status	Married-AF-spouse	113	11.3
marital_status	Missing	50	5.0
native_country	Mexico	37	3.7
native_country	Yugoslavia	33	3.3
native_country	Outlying-US(Guam-USVI-etc)	32	3.2

Table 7: Top 10 levels of categorical variables

	variables	levels	freq	ratio (%)
26	native_country	Greece	30	3.0
27	native_country	Puerto-Rico	30	3.0
28	native_country	Trinadad&Tobago	30	3.0
29	native_country	England	29	2.9
30	native_country	India	29	2.9
31	native_country	Canada	28	2.8
32	native_country	Other levles	672	67.2
33	native_country	Missing	50	5.0
34	occupation	Handlers-cleaners	83	8.3
35	occupation	Machine-op-inspct	77	7.7
36	occupation	Armed-Forces	76	7.6
37	occupation	Sales	75	7.5
38	occupation	Prof-specialty	70	7.0
39	occupation	Protective-serv	69	6.9
40	occupation	Exec-managerial	68	6.8
41	occupation	Transport-moving	67	6.7
42	occupation	Craft-repair	63	6.3
43	occupation	Farming-fishing	63	6.3
44	occupation	Other levles	239	23.9
45	occupation	Missing	50	5.0
46	race	White	205	20.5
47	race	Amer-Indian-Eskimo	199	19.9
48	race	Other	199	19.9
49	race	Black	177	17.7
50	race	Asian-Pac-Islander	170	17.0
51	race	Missing	50	5.0
52	relationship	Husband	175	17.5
53	relationship	Wife	160	16.0
54	relationship	Not-in-family	156	15.6

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

	variables	levels	freq	ratio (%)
55	relationship	Unmarried	155	15.5
56	relationship	Own-child	153	15.3
57	relationship	Other-relative	151	15.1
58	relationship	Missing	50	5.0
59	sex	Female	481	48.1
60	sex	Male	469	46.9
61	sex	Missing	50	5.0
62	workclass	Federal-gov	135	13.5
63	workclass	Local-gov	129	12.9
64	workclass	Without-pay	124	12.4
65	workclass	Self-emp-not-inc	123	12.3
66	workclass	Private	121	12.1
67	workclass	State-gov	109	10.9
68	workclass	Never-worked	107	10.7
69	workclass	Self-emp-inc	102	10.2
70	workclass	Missing	50	5.0

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

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
age	18	36.00	53.81	53.0	72.00	90	0	0	0
education_num	1	5.00	8.59	9.0	13.00	16	0	0	0
capital_gain	85	26,820.00	50,800.46	49,397.5	75,315.50	99,980	0	0	0
capital_loss	55	25,106.75	50,171.11	50,678.0	75,626.25	99,897	0	0	0
hours_per_week	1	26.00	50.25	50.0	76.00	99	0	0	0

Table 8: General list of numerical diagnosis

Zero Values

No numeric variable with zero value

Negative Values

No numeric variable with negative value

Outliers

List of Outliers

No numeric variables including outliers

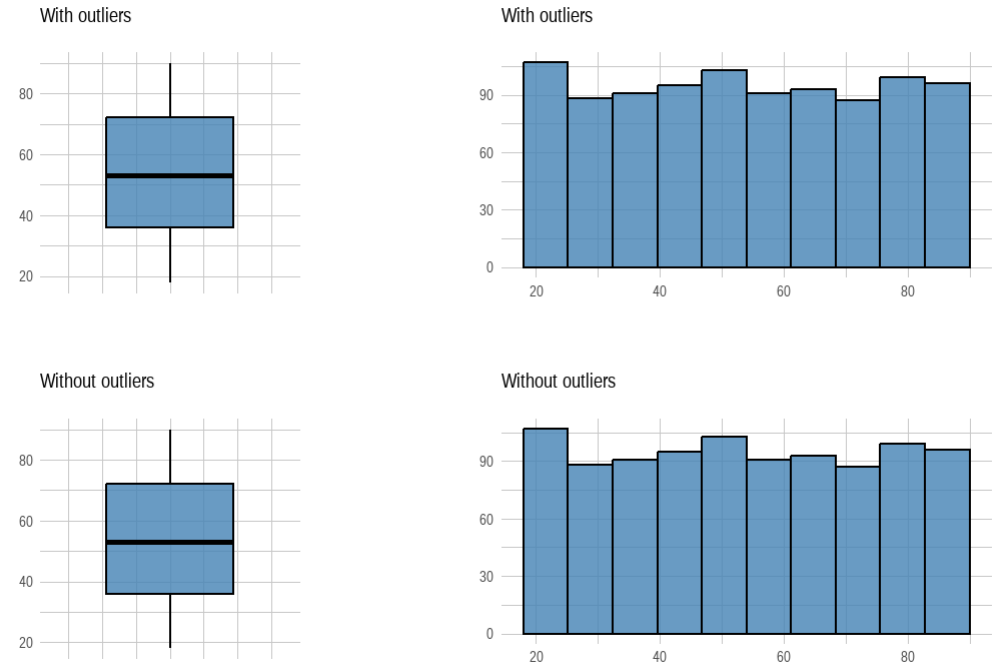
Individual Outliers

variable: age

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	53.81263
Mean without outliers	53.81263

Table 9: age

Outlier Diagnosis Plot (age)

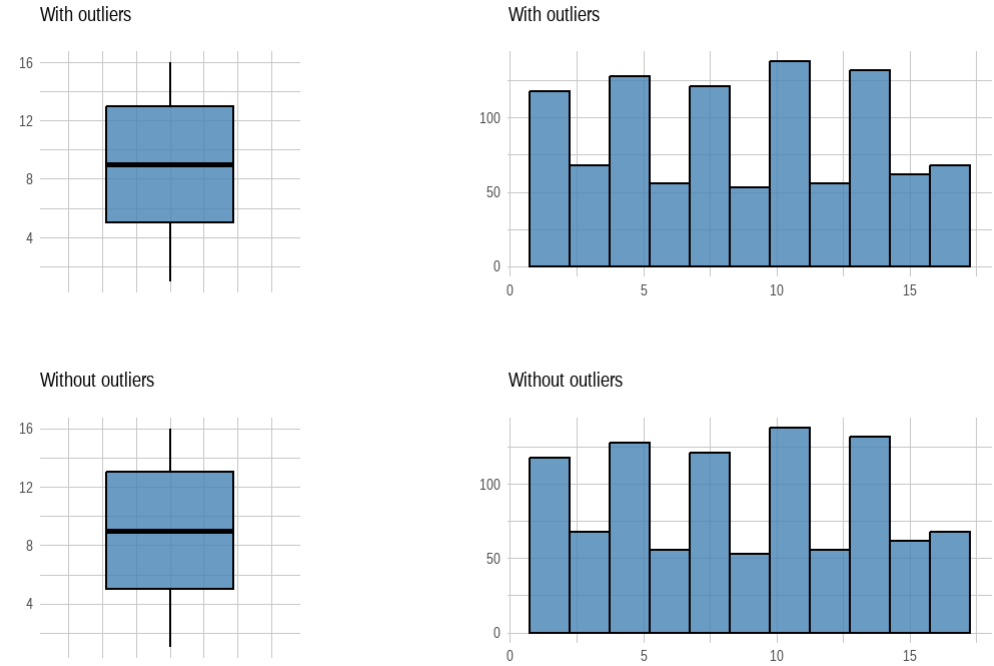


variable: education_num

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	8.587
Mean without outliers	8.587

Table 9: education_num

Outlier Diagnosis Plot (education_num)

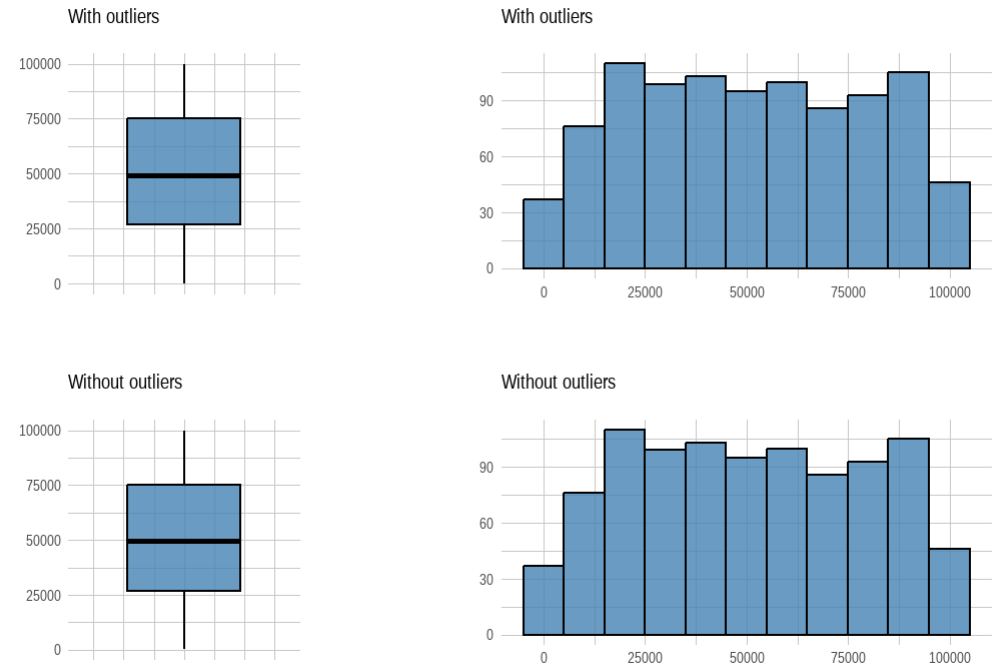


variable: capital_gain

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	50800.46
Mean without outliers	50800.46

Table 9: capital_gain

Outlier Diagnosis Plot (capital_gain)

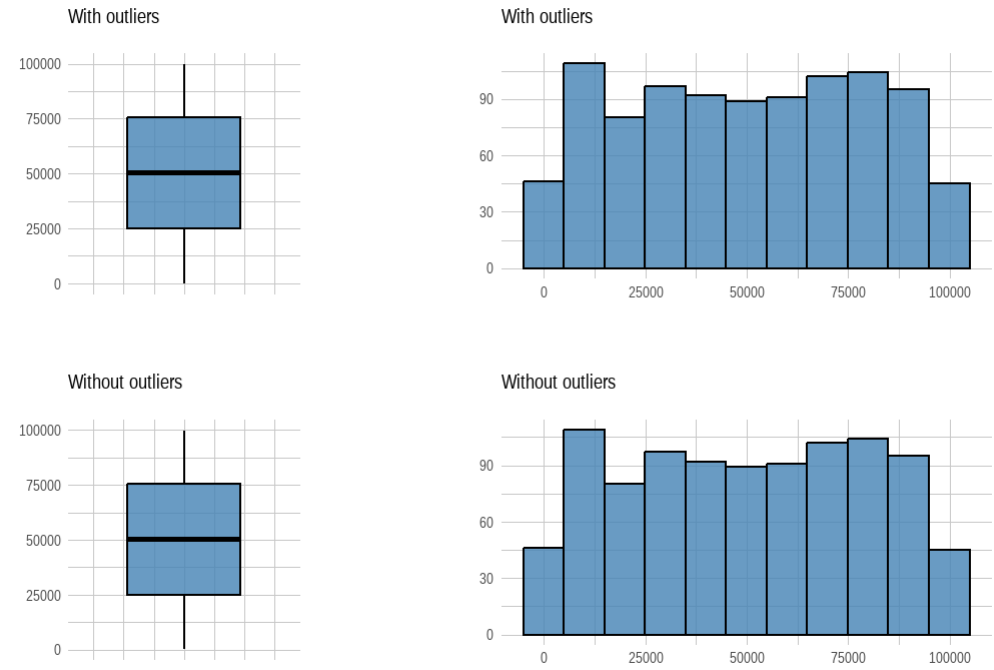


variable: capital_loss

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	50171.11
Mean without outliers	50171.11

Table 9: capital_loss

Outlier Diagnosis Plot (capital_loss)



variable: hours_per_week

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	50.25368
Mean without outliers	50.25368

Table 9: hours_per_week

Outlier Diagnosis Plot (hours_per_week)