DATA DIAGNOSIS REPORT

CLEAN_DATA

Report Overview

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

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	clean_data
dataset	dataset type	data.frame
job	samples	1,000 / 1,000 (100%)
job	created	2024-04-04 11:51:13.441176
job	created by	dlookr

Table 2: Job informations

Warnings

No warnings

No warnings in dataset and variables

Variables

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

Table 3: 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

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	79	7.9
education	1st-4th	74	7.4
education	9th	73	7.3
education	Assoc-voc	70	7.0
education	11th	68	6.8
education	5th-6th	68	6.8
education	Doctorate	66	6.6
education	Some-college	64	6.4
education	Bachelors	61	6.1
education	HS-grad	61	6.1
education	Other levles	316	31.6
income	<=50K	737	73.7
income	>50K	263	26.3
marital_status	Never-married	155	15.5
marital_status	Widowed	152	15.2
marital_status	Married-civ-spouse	149	14.9
marital_status	Married-spouse-absent	149	14.9
marital_status	Divorced	140	14.0
marital_status	Separated	139	13.9
marital_status	Married-AF-spouse	116	11.6
native_country	Mexico	39	3.9
native_country	Yugoslavia	36	3.6
native_country	Greece	33	3.3
native_country	Outlying-US(Guam-USVI-etc)	32	3.2
native_country	Puerto-Rico	32	3.2

Table 4: Top 10 levels of categorical variables

	variables	levels	freq	ratio (%)
26	native_country	Trinadad&Tobago	32	3.2
27	native_country	India	31	3.1
28	native_country	Cuba	30	3.0
29	native_country	England	30	3.0
30	native_country	Canada	29	2.9
31	native_country	Other levles	676	67.6
32	occupation	Handlers-cleaners	90	9.0
33	occupation	Machine-op-inspct	82	8.2
34	occupation	Armed-Forces	80	8.0
35	occupation	Sales	79	7.9
36	occupation	Prof-specialty	76	7.6
37	occupation	Exec-managerial	72	7.2
38	occupation	Protective-serv	72	7.2
39	occupation	Transport-moving	72	7.2
40	occupation	Craft-repair	66	6.6
41	occupation	Farming-fishing	65	6.5
42	occupation	Other levles	246	24.6
43	race	White	213	21.3
44	race	Amer-Indian-Eskimo	210	21.0
45	race	Other	207	20.7
46	race	Black	191	19.1
47	race	Asian-Pac-Islander	179	17.9
48	relationship	Husband	181	18.1
49	relationship	Wife	173	17.3
50	relationship	Unmarried	166	16.6
51	relationship	Not-in-family	163	16.3
52	relationship	Own-child	160	16.0
53	relationship	Other-relative	157	15.7
54	sex	Female	507	50.7

Table 4: Top 10 levels of categorical variables (continued)

	variables	levels	freq	ratio (%)
55	sex	Male	493	49.3
56	workclass	Local-gov	143	14.3
57	workclass	Federal-gov	141	14.1
58	workclass	Self-emp-not-inc	130	13.0
59	workclass	Private	127	12.7
60	workclass	Without-pay	126	12.6
61	workclass	State-gov	115	11.5
62	workclass	Never-worked	111	11.1
63	workclass	Self-emp-inc	107	10.7

Table 4: Top 10 levels of categorical variables (continued)

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
age	18	35.00	53.47	53	72.00	90	0	0	0
education_num	1	5.00	8.59	9	13.00	16	0	0	0
capital_gain	85	26,662.50	50,748.19	49,541	75,636.25	99,980	0	0	0
capital_loss	55	25,252.25	50,140.91	50,544	75,620.75	99,897	0	0	0
hours_per_week	1	26.00	50.19	49	76.25	99	0	0	0

Table 5: 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

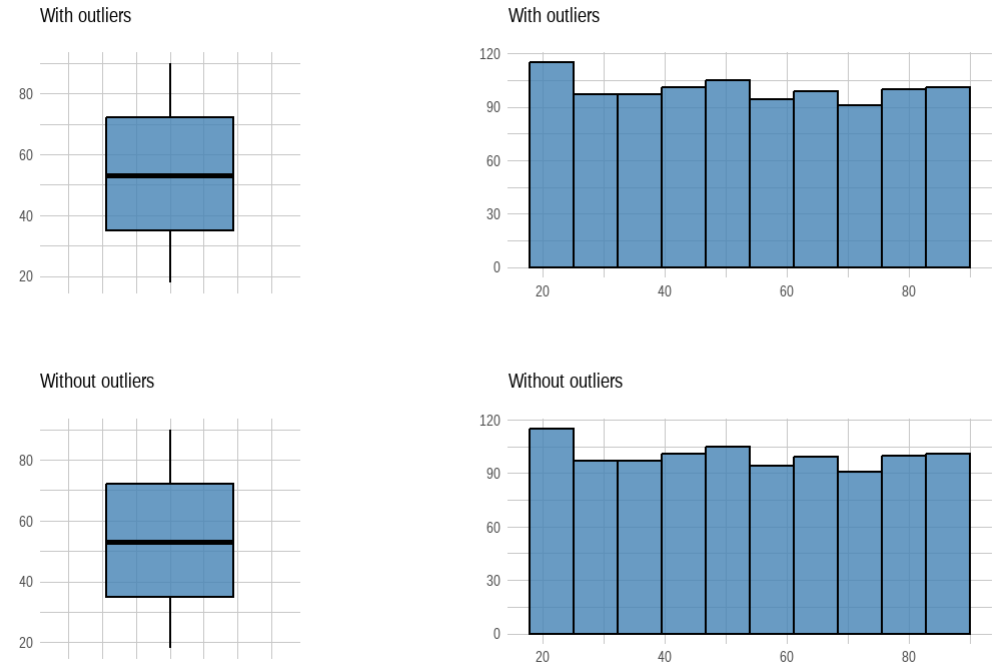
Individual Outliers

variable: age

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	53.47
Mean without outliers	53.47

Table 6: age

Outlier Diagnosis Plot (age)

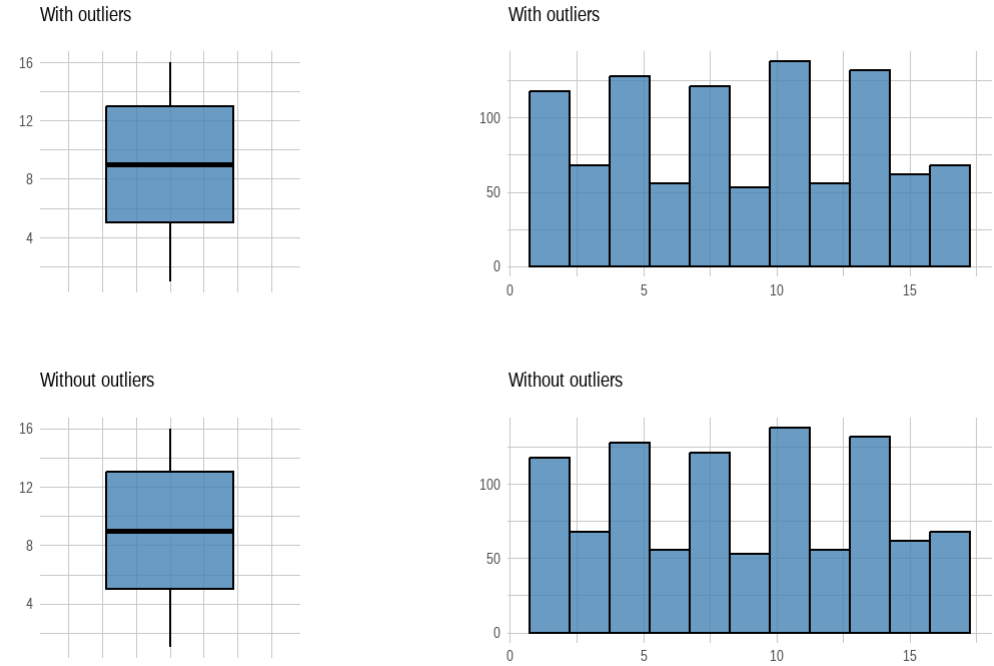


variable: education_num

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

Table 6: education_num

Outlier Diagnosis Plot (education_num)

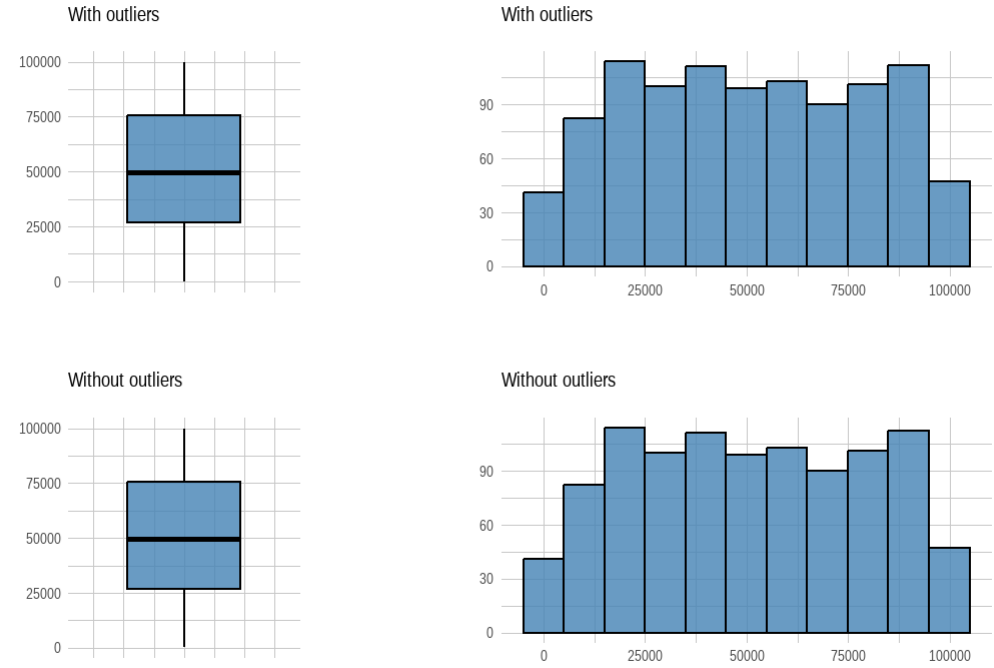


variable: capital_gain

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	50748.19
Mean without outliers	50748.19

Table 6: capital_gain

Outlier Diagnosis Plot (capital_gain)

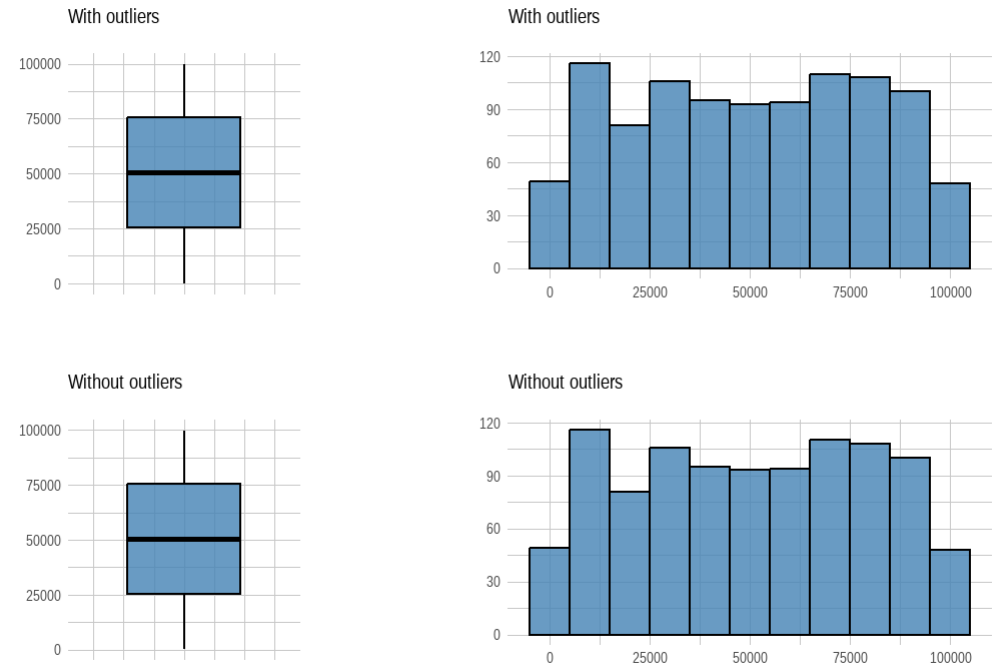


variable: capital_loss

Measures	Values
Outliers count	0
Outliers ratio (%)	0%
Mean of outliers	NaN
Mean with outliers	50140.91
Mean without outliers	50140.91

Table 6: 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	50.188
Mean without outliers	50.188

Table 6: hours_per_week

Outlier Diagnosis Plot (hours_per_week)