REPORT SERIES WITH DLOOKR

Data Quality Diagnosis Report

Author:
dlookr package

Version:
0.3.12

March 27, 2020

Contents

1 Diagnose Data 3
1.1 Overview of Diagnosis 3
1.1.1 List of all variables quality 3
1.1.2 Diagnosis of missing data 3
1.1.3 Diagnosis of unique data(Text and Category) 3
1.1.4 Diagnosis of unique data(Numerical) 3
1.2 Detailed data diagnosis 4
1.2.1 Diagnosis of categorical variables 4
1.2.2 Diagnosis of numerical variables 5
1.2.3 List of numerical diagnosis (zero) 7
1.2.4 List of numerical diagnosis (minus) 7

2 Diagnose Outliers 9
2.1 Overview of Diagnosis 9
2.1.1 Diagnosis of numerical variable outliers 9
2.2 Detailed outliers diagnosis 10

Chapter 1

Diagnose Data

1.1 Overview of Diagnosis

1.1.1 List of all variables quality

Table 1.1: Data quality overview table

variables	type	missing (n)	missing (%)	unique (n)	unique (n/N)
manufacturer	factor	0	0	15	0.064
model	factor	0	0	38	0.162
displ	numeric	0	0	35	0.150
year	integer	0	0	2	0.009
cyl	integer	0	0	4	0.017
trans	factor	0	0	10	0.043
drv	factor	0	0	3	0.013
cty	integer	0	0	21	0.090
hwy	integer	0	0	27	0.115
fl	factor	0	0	5	0.021
class	factor	0	0	7	0.030

1.1.2 Diagnosis of missing data

No variables including missing values

1.1.3 Diagnosis of unique data(Text and Category)

No variable with a high proportion greater than 0.5

1.1.4 Diagnosis of unique data(Numerical)

Table 1.2: Variables where the proportion of unique data is less than 0.1

variables	type	missing (n)	missing (%)	unique (n)	unique (n/N)
cty	integer	0	0	21	0.090
cyl	integer	0	0	4	0.017
year	integer	0	0	2	0.009

1.2 Detailed data diagnosis

1.2.1 Diagnosis of categorical variables

Table 1.3: Categorical variable level top 10

variables	levels	N	freq	ratio(%)	rank
manufacturer	dodge	234	37	15.812	1
manufacturer	toyota	234	34	14.530	2
manufacturer	volkswagen	234	27	11.538	3
manufacturer	ford	234	25	10.684	4
manufacturer	chevrolet	234	19	8.120	5
manufacturer	audi	234	18	7.692	6
manufacturer	hyundai	234	14	5.983	7
manufacturer	subaru	234	14	5.983	8
manufacturer	nissan	234	13	5.556	9
manufacturer	honda	234	9	3.846	10
model	caravan 2wd	234	11	4.701	1
model	ram 1500 pickup 4wd	234	10	4.274	2
model	civic	234	9	3.846	3
model	dakota pickup 4wd	234	9	3.846	4
model	jetta	234	9	3.846	5
model	mustang	234	9	3.846	6
model	a4 quattro	234	8	3.419	7
model	grand cherokee 4wd	234	8	3.419	8
model	impreza awd	234	8	3.419	9
model	a4	234	7	2.991	10
model	camry	234	7	2.991	11
model	camry solara	234	7	2.991	12
model	durango 4wd	234	7	2.991	13
model	f150 pickup 4wd	234	7	2.991	14
model	passat	234	7	2.991	15
model	sonata	234	7	2.991	16
model	tiburon	234	7	2.991	17
model	toyota tacoma 4wd	234	7	2.991	18
trans	auto(l4)	234	83	35.470	1
trans	manual(m5)	234	58	24.786	2
trans	auto(l5)	234	39	16.667	3
trans	manual(m6)	234	19	8.120	4
trans	auto(s6)	234	16	6.838	5
trans	auto(l6)	234	6	2.564	6
trans	auto(av)	234	5	2.137	7
trans	auto(s4)	234	3	1.282	8
trans	auto(s5)	234	3	1.282	9
trans	auto(l3)	234	2	0.855	10
drv	f	234	106	45.299	1
drv	4	234	103	44.017	2
drv	r	234	25	10.684	3
fl	r	234	168	71.795	1
fl	p	234	52	22.222	2
fl	e	234	8	3.419	3

Table 1.3: Categorical variable level top 10 *(continued)*

variables	levels	N	freq	ratio(%)	rank
fl	d	234	5	2.137	4
fl	c	234	1	0.427	5
class	suv	234	62	26.496	1
class	compact	234	47	20.085	2
class	midsize	234	41	17.521	3
class	subcompact	234	35	14.957	4
class	pickup	234	33	14.103	5
class	minivan	234	11	4.701	6
class	2seater	234	5	2.137	7

1.2.2 Diagnosis of numerical variables

Table 1.4: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
displ	1.6	2.4	3.472	3.3	4.6	7	0	0	0
year	1,999.0	1,999.0	2,003.500	2,003.5	2,008.0	2,008	0	0	0
cyl	4.0	4.0	5.889	6.0	8.0	8	0	0	0
cty	9.0	14.0	16.859	17.0	19.0	35	0	0	5
hwy	12.0	18.0	23.440	24.0	27.0	44	0	0	3

1.2.3 List of numerical diagnosis (zero)

No numeric variable with zero value

1.2.4 List of numerical diagnosis (minus)

No numeric variable with negative value

Chapter 2

Diagnose Outliers

2.1 Overview of Diagnosis

2.1.1 Diagnosis of numerical variable outliers

Table 2.1: Diagnosis of numerical variable outliers

variables	min	median	max	outlier	outlier ratio(%)
cty	9	17	35	5	2.137
hwy	12	24	44	3	1.282

2.2 Detailed outliers diagnosis

variable : cty

Table 2.2: Outliers information of cty

Measures	Values
Outliers count	5.00
Outliers ratio (%)	2.14
Mean of outliers	30.60
Mean with outliers	16.86
Mean without outliers	16.56

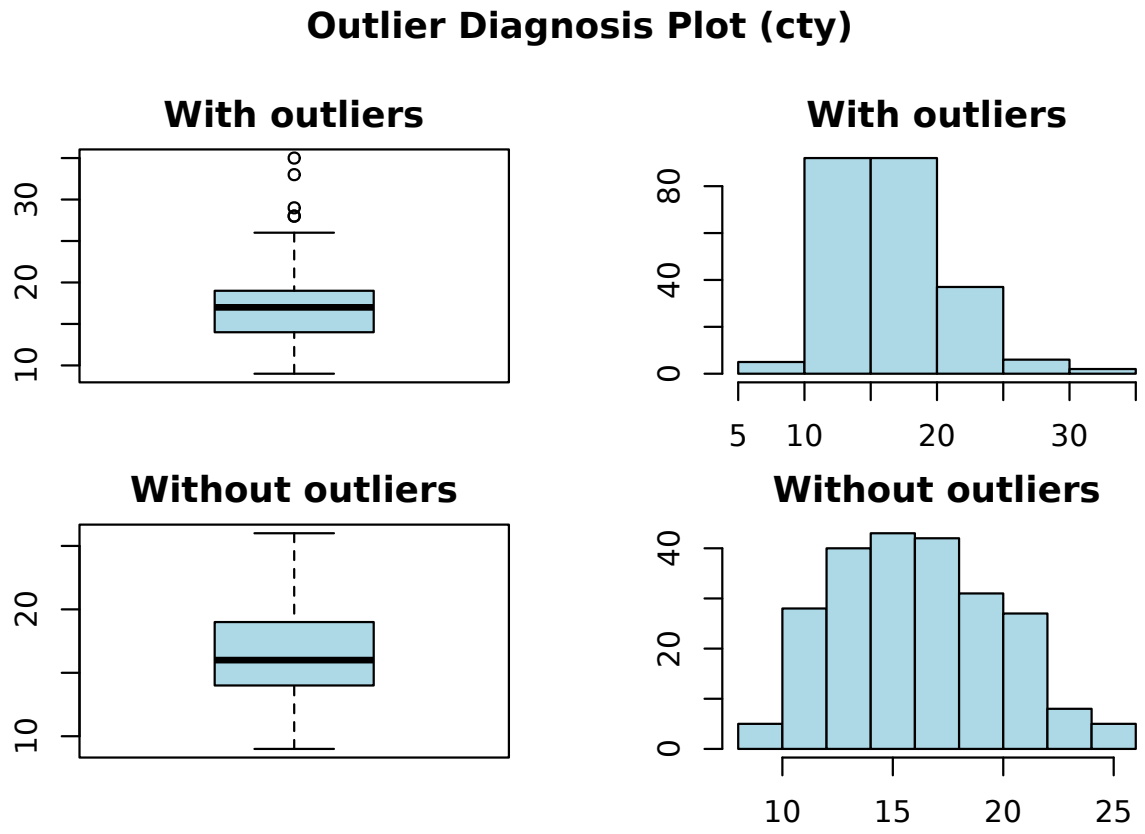

Figure 2.1: Distribution of cty

variable : hwy

Table 2.3: Outliers information of hwy

Measures	Values
Outliers count	3.00
Outliers ratio (%)	1.28
Mean of outliers	43.00
Mean with outliers	23.44
Mean without outliers	23.19

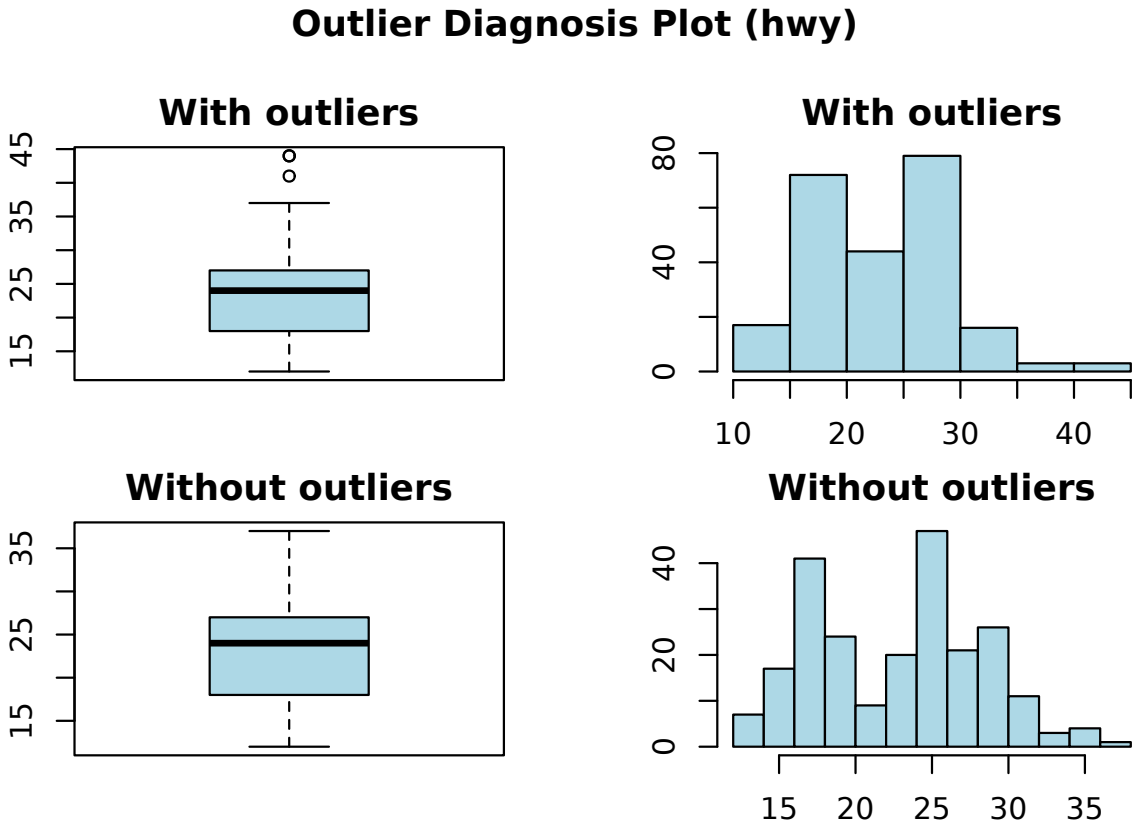

Figure 2.2: Distribution of hwy