REPORT SERIES WITH DLOOKR

Data Quality Diagnosis Report

Author:
dlookr package

Version:
0.3.12

August 25, 2020

Contents

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)
Sales	numeric	0	0.00	336	0.840
CompPrice	numeric	0	0.00	73	0.182
Income	numeric	20	5.00	96	0.240
Advertising	numeric	0	0.00	28	0.070
Population	numeric	0	0.00	275	0.688
Price	numeric	0	0.00	101	0.252
ShelveLoc	factor	0	0.00	3	0.007
Age	numeric	0	0.00	56	0.140
Education	numeric	0	0.00	9	0.022
Urban	factor	5	1.25	3	0.007
US	factor	0	0.00	2	0.005

1.1.2 Diagnosis of missing data

Table 1.2: Variables that include missing values

variables	type	missing (n)	missing (%)	unique (n)	unique (n/N)
Income	numeric	20	5.00	96	0.240
Urban	factor	5	1.25	3	0.007

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

variables	type	missing (n)	missing (%)	unique (n)	unique (n/N)
Advertising	numeric	0	0	28	0.070
Education	numeric	0	0	9	0.022

1.2   Detailed data diagnosis

1.2.1   Diagnosis of categorical variables

Table 1.4: Categorical variable level top 10

variables	levels	N	freq	ratio(%)	rank
ShelveLoc	Medium	400	219	54.75	1
ShelveLoc	Bad	400	96	24.00	2
ShelveLoc	Good	400	85	21.25	3
Urban	Yes	400	278	69.50	1
Urban	No	400	117	29.25	2
Urban	NA	400	5	1.25	3
US	Yes	400	258	64.50	1
US	No	400	142	35.50	2

1.2.2   Diagnosis of numerical variables

Table 1.5: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Sales	0	5.39	7.496	7.49	9.32	16.27	1	0	2
CompPrice	77	115.00	124.975	125.00	135.00	175.00	0	0	2
Income	21	42.00	68.329	69.00	90.00	120.00	0	0	0
Advertising	0	0.00	6.635	5.00	12.00	29.00	144	0	0
Population	10	139.00	264.840	272.00	398.50	509.00	0	0	0
Price	24	100.00	115.795	117.00	131.00	191.00	0	0	5
Age	25	39.75	53.322	54.50	66.00	80.00	0	0	0
Education	10	12.00	13.900	14.00	16.00	18.00	0	0	0

1.2.3   List of numerical diagnosis (zero)

Table 1.6: List of numerical diagnosis (zero)

variables	min	median	max	zero	zero ratio(%)
Advertising	0	5.00	29.00	144	36.00
Sales	0	7.49	16.27	1	0.25

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(%)
Price	24	117.00	191.00	5	1.25
Sales	0	7.49	16.27	2	0.50
CompPrice	77	125.00	175.00	2	0.50

2.2   Detailed outliers diagnosis

variable : Price

Table 2.2: Outliers information of Price

Measures	Values
Outliers count	5.00
Outliers ratio (%)	1.25
Mean of outliers	100.40
Mean with outliers	115.80
Mean without outliers	115.99
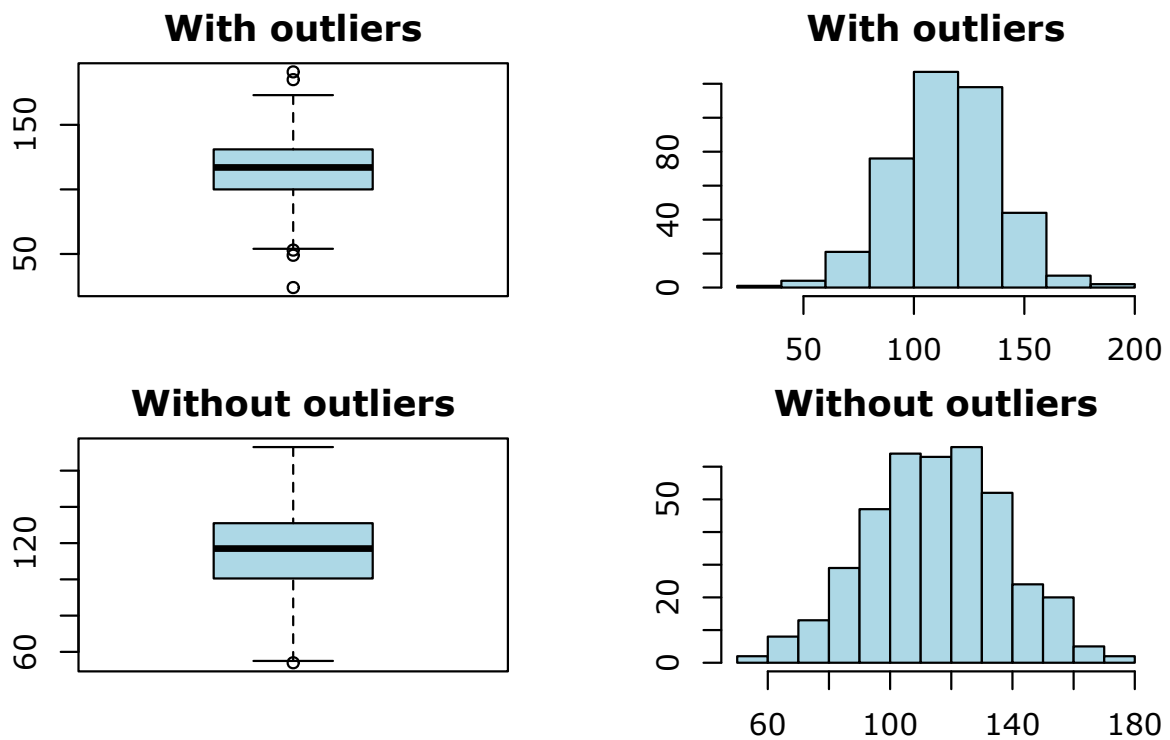
Figure 2.1: Distribution of Price

variable : Sales

Table 2.3: Outliers information of Sales

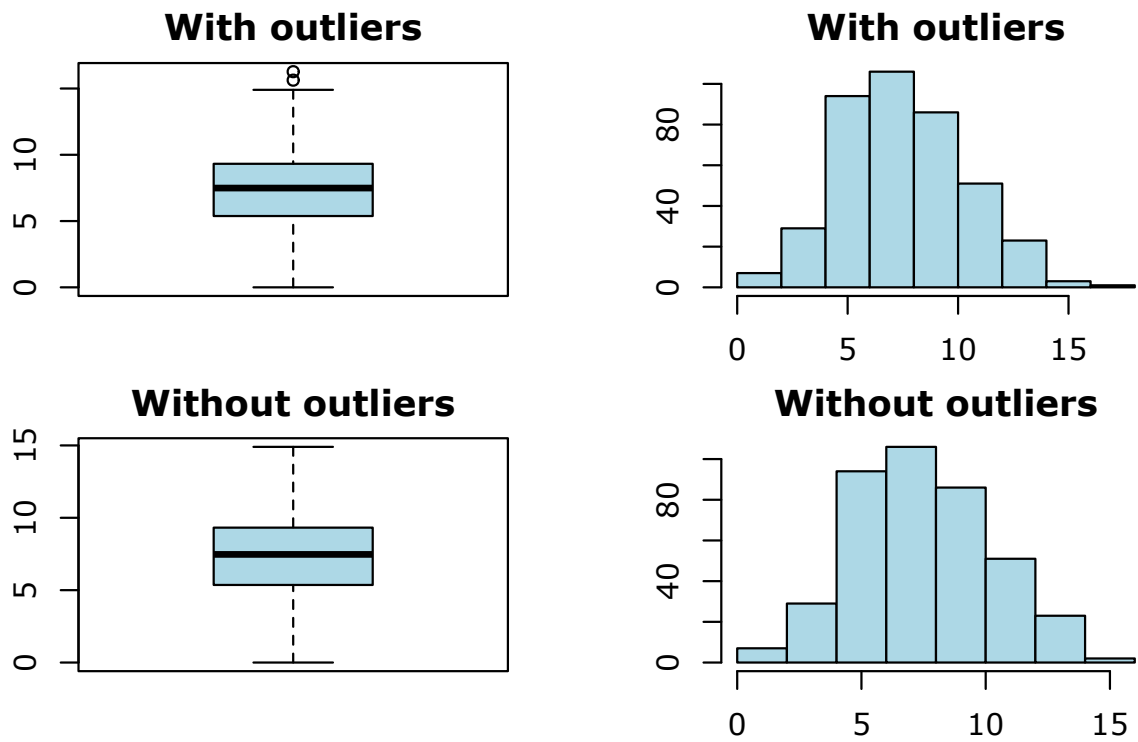| Measures	Values
Outliers count	2.00
Outliers ratio (%)	0.50
Mean of outliers	15.95
Mean with outliers	7.50
Mean without outliers	7.45

Figure 2.2: Distribution of Sales

variable : CompPrice

Table 2.4: Outliers information of CompPrice

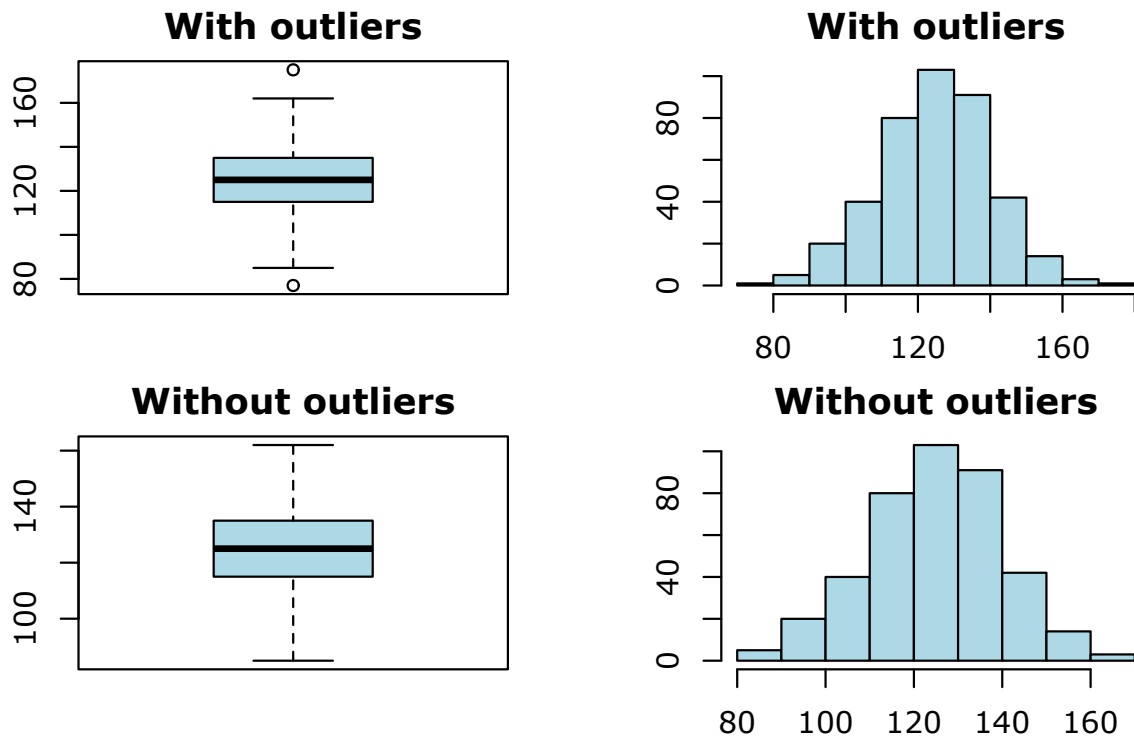| Measures	Values
Outliers count	2.00
Outliers ratio (%)	0.50
Mean of outliers	126.00
Mean with outliers	124.97
Mean without outliers	124.97

Outlier Diagnosis Plot (CompPrice)

Figure 2.3: Distribution of CompPrice