REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

March 2, 2021

Contents

Chapter 1

Introduction

The EDA Report provides exploratory data analysis information on objects that inherit data.frame and data.frame.

1.1 Information of Dataset

The dataset that generated the EDA Report is an ‘data.frame’ object. It consists of 400 observations and 11 variables.

1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
Sales	numeric	0	0.00	336	0.840
CompPrice	numeric	0	0.00	73	0.182
Income	numeric	20	5.00	99	0.248
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

The target variable of the data is ’Sales’, and the data type of the variable is numeric.

1.3 About EDA Report

EDA reports provide information and visualization results that support the EDA process. In particular, it provides a variety of information to understand the relationship between the target variable and the rest of the variables of interest.

Chapter 2

Univariate Analysis

2.1 Descriptive Statistics

edaData
11 Variables   400 Observations

Sales

```
      n   missing  distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    400         0       336        1    7.496    3.192    3.149    4.119    5.390    7.490    9.320   11.300   12.442

lowest :  0.00  0.16  0.37  0.53  0.91, highest: 13.91 14.37 14.90 15.63 16.27
```

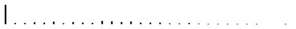
CompPrice

```
      n   missing  distinct     Info     Mean      Gmd    .05    .10    .25    .50    .75    .90    .95
    400         0        73    0.999      125     17.3     98    106    115    125    135    145    150

lowest :  77  85  86  88  89, highest: 157 159 161 162 175
```

Income

```
      n   missing  distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    380        20        98        1    68.12    32.47    25.95    30.00    42.00    68.50    90.00   107.00   115.00

lowest :  21  22  23  24  25, highest: 116 117 118 119 120
```

Advertising



```
      n   missing  distinct     Info     Mean      Gmd    .05    .10    .25    .50    .75    .90    .95
    400         0        28    0.952    6.635    7.337      0      0      0      5     12     16     19

lowest :  0  1  2  3  4, highest: 23 24 25 26 29
```

Population

```
      n   missing  distinct     Info     Mean      Gmd    .05    .10     .25     .50     .75     .90     .95
    400         0       275        1    264.8    170.3   29.0   58.9   139.0   272.0   398.5   467.0   493.1

lowest :  10  12  13  14  16, highest: 503 504 507 508 509
```

Price

```
      n   missing  distinct     Info     Mean      Gmd    .05    .10    .25    .50    .75    .90    .95
    400         0       101        1    115.8    26.52     77     87    100    117    131    146    155

lowest :  24  49  53  54  55, highest: 166 171 173 185 191
```

ShelveLoc

```
      n   missing  distinct
    400         0         3

Value         Bad   Good Medium
Frequency      96     85    219
Proportion  0.240  0.212  0.547
```

Age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
400	0	56	1	53.32	18.71	27.00	30.00	39.75	54.50	66.00	76.00	79.00

```
lowest : 25 26 27 28 29, highest: 76 77 78 79 80
```

Education

n	missing	distinct	Info	Mean	Gmd
400	0	9	0.987	13.9	3.009

```
lowest : 10 11 12 13 14, highest: 14 15 16 17 18

Value         10    11    12    13    14    15    16    17    18
Frequency     48    48    49    43    40    36    47    49    40
Proportion 0.120 0.120 0.122 0.108 0.100 0.090 0.117 0.122 0.100
```

Urban

n	missing	distinct
395	5	2

```
Value         No   Yes
Frequency    116   279
Proportion 0.294 0.706
```

US

n	missing	distinct
400	0	2

```
Value         No   Yes
Frequency    142   258
Proportion 0.355 0.645
```

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

CompPrice

* normality test : Shapiro-Wilk normality test
- statistic : 0.99843, p-value : 0.977151

Table 2.1: skewness and kurtosis : CompPrice

type	skewness	kurtosis
original	-0.0426	3.0262
log transformation	-0.4347	3.3671
sqrt transformation	-0.2347	3.1280

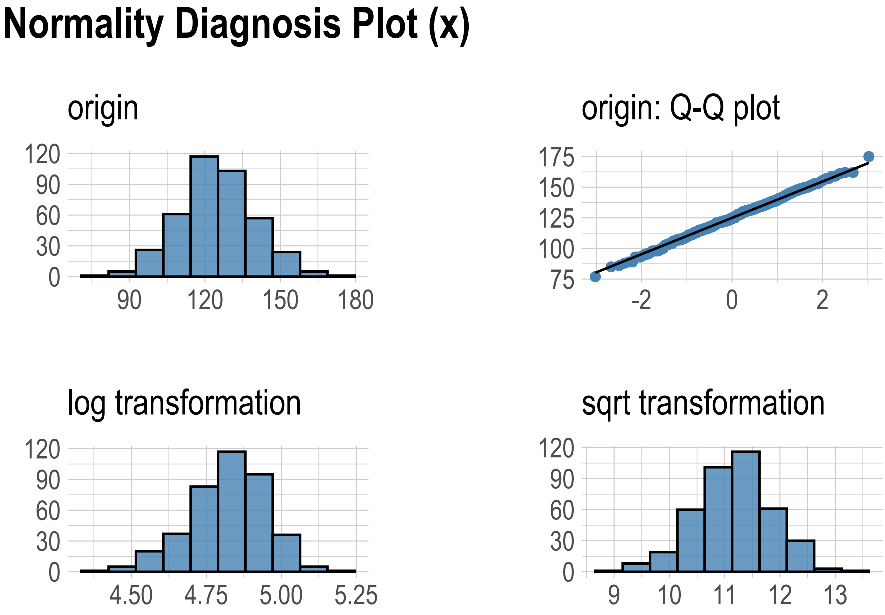


Figure 2.1: CompPrice

Income

* normality test : Shapiro-Wilk normality test
- statistic : 0.95995, p-value : 1.14495E-08

Table 2.2: skewness and kurtosis : Income

type	skewness	kurtosis
original	0.0797	1.9065
log transformation	-0.5412	2.2170
sqrt transformation	-0.2222	1.9480

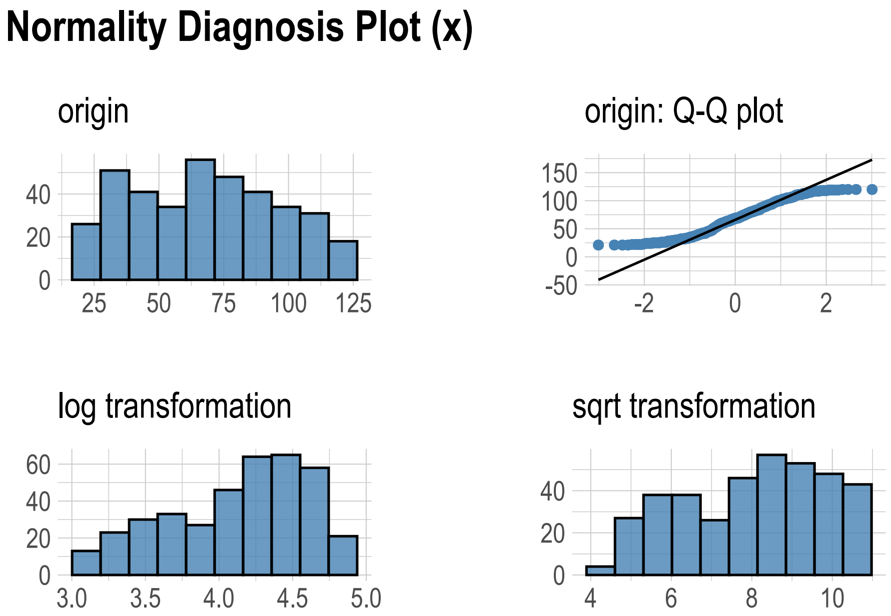


Figure 2.2: Income

Advertising

* normality test : Shapiro-Wilk normality test
- statistic : 0.87354, p-value : 1.49183E-17

Table 2.3: skewness and kurtosis : Advertising

type	skewness	kurtosis
original	0.6372	2.4467
log+1 transformation	-0.1978	1.3423
sqrt transformation	-0.0565	1.4653

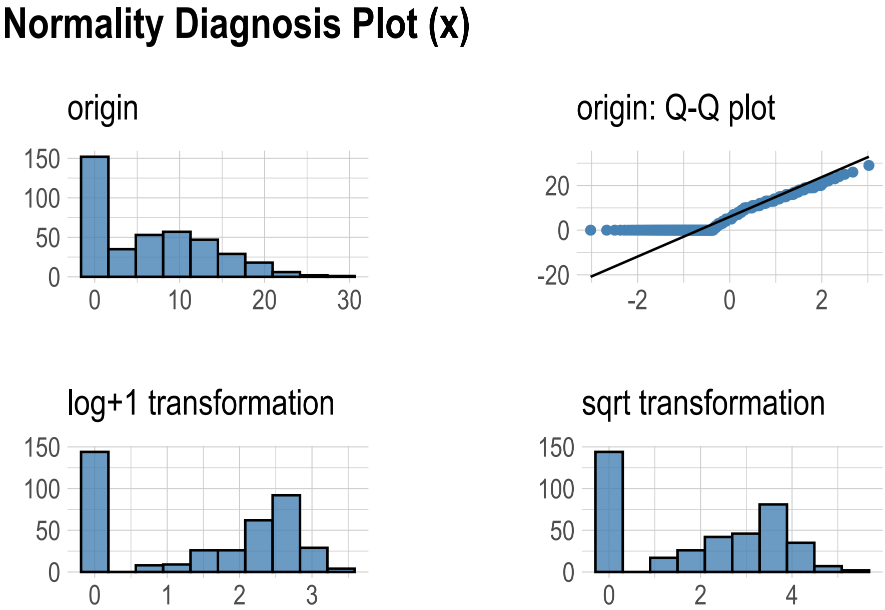


Figure 2.3: Advertising

Population

* normality test : Shapiro-Wilk normality test
- statistic : 0.95201, p-value : 4.08085E-10

Table 2.4: skewness and kurtosis : Population

type	skewness	kurtosis
original	-0.0510	1.7977
log transformation	-1.2945	4.1336
sqrt transformation	-0.5427	2.2584

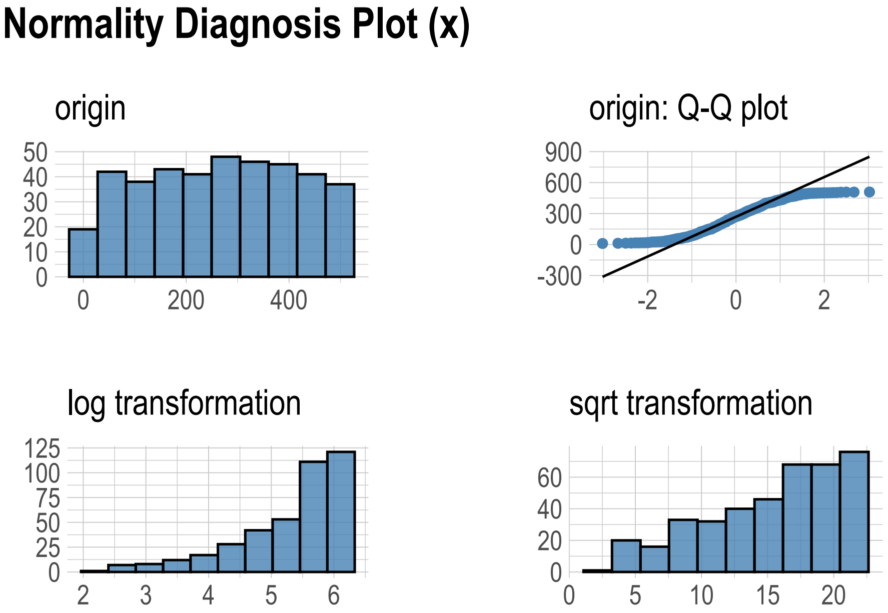


Figure 2.4: Population

Price

* normality test : Shapiro-Wilk normality test
- statistic : 0.99592, p-value : 0.390213

Table 2.5: skewness and kurtosis : Price

type	skewness	kurtosis
original	-0.1248	3.4313
log transformation	-1.3589	8.6448
sqrt transformation	-0.6083	4.5887

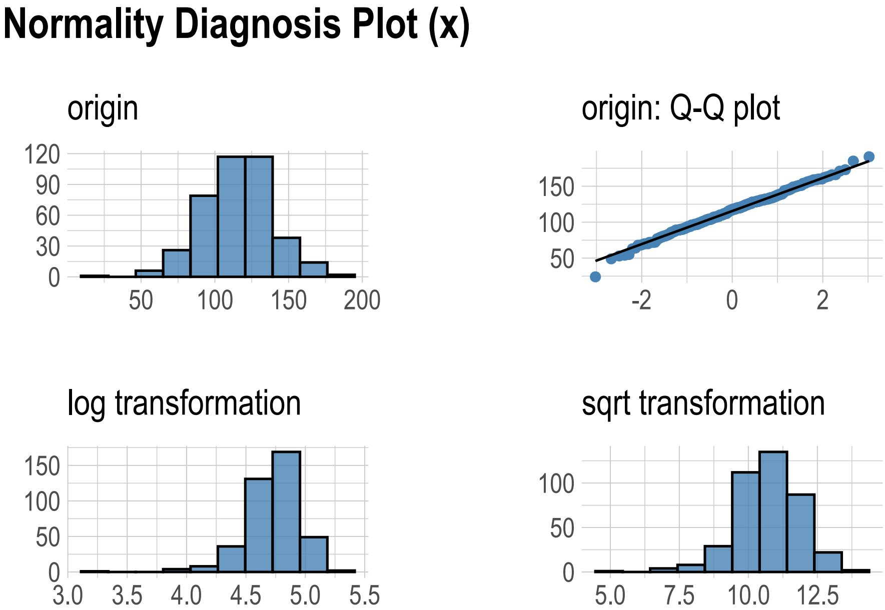


Figure 2.5: Price

Age

* normality test : Shapiro-Wilk normality test
- statistic : 0.95672, p-value : 1.86455E-09

Table 2.6: skewness and kurtosis : Age

type	skewness	kurtosis
original	-0.0769	1.8648
log transformation	-0.5112	2.1718
sqrt transformation	-0.2890	1.9631

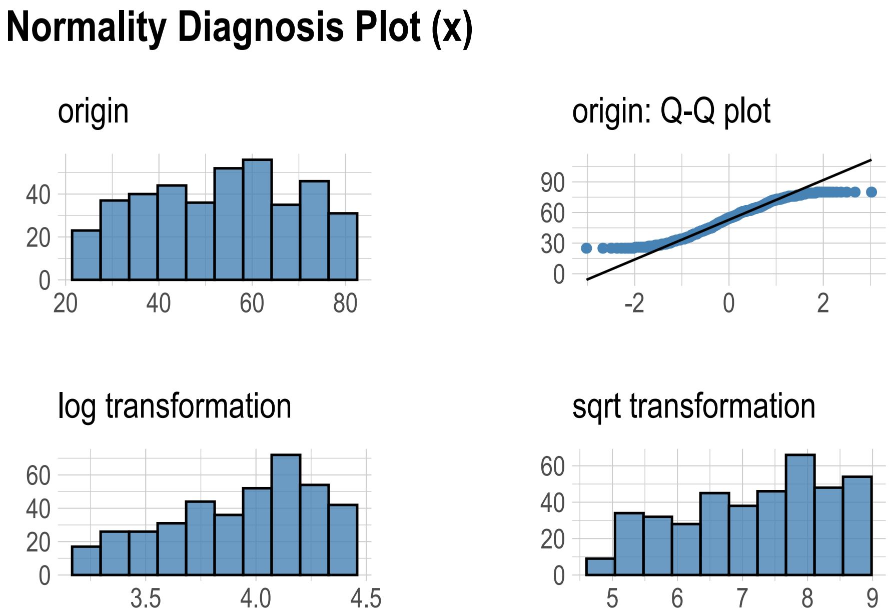


Figure 2.6: Age

Education

* normality test : Shapiro-Wilk normality test
- statistic : 0.9242, p-value : 2.42693E-13

Table 2.7: skewness and kurtosis : Education

type	skewness	kurtosis
original	0.0438	1.7029
log transformation	-0.1599	1.7434
sqrt transformation	-0.0572	1.7118

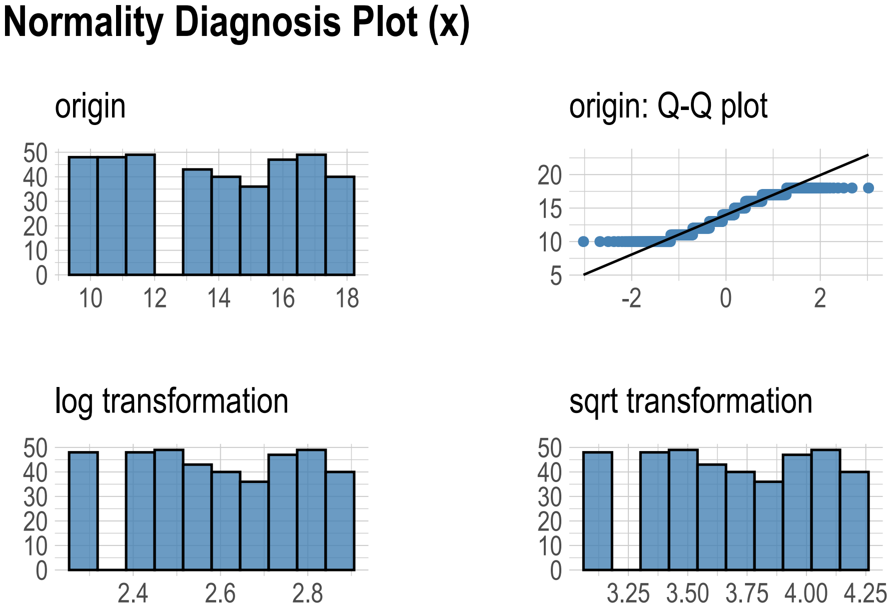


Figure 2.7: Education

Chapter 3

Relationship Between Variables

3.1 Correlation Coefficient

3.1.1 Correlation Coefficient by Variable Combination

Table 3.1: The correlation coefficients (0.5 or more)

Variable1	Variable2	Correlation Coefficient
Price	CompPrice	0.585

3.1.2 Correlation Plot of Numerical Variables

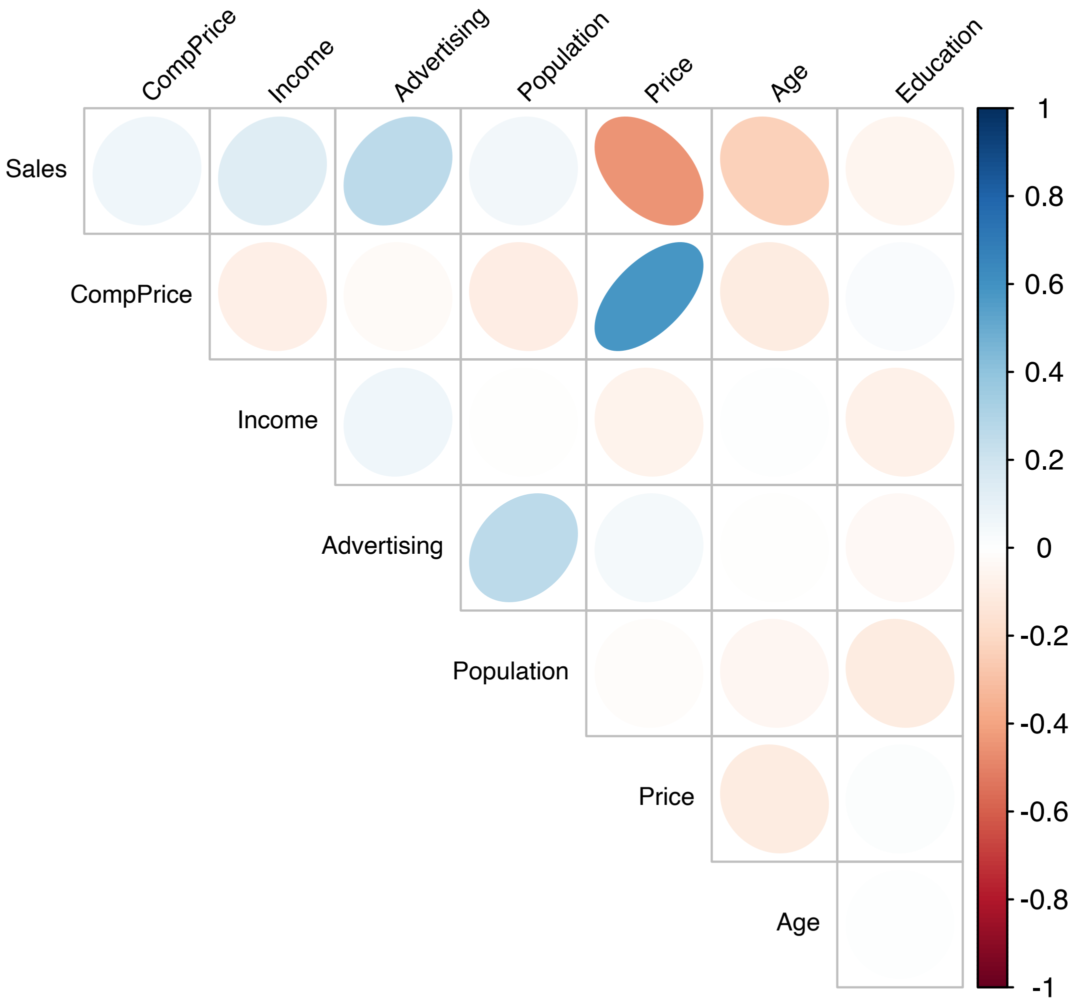


Figure 3.1: The correlation coefficient of numerical variables

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

CompPrice

1. **Simple Linear Model Information**

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.00411, Adjusted R-squared: 0.0016
F-statistic: 2 on 1 and 398 DF, p-value: 0.2009398

Table 4.1: Simple Linear Model coefficients : CompPrice

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	6.02	1.16	5.19	0.0
CompPrice	0.01	0.01	1.28	0.2

2. **Visualization - Scatterplots**

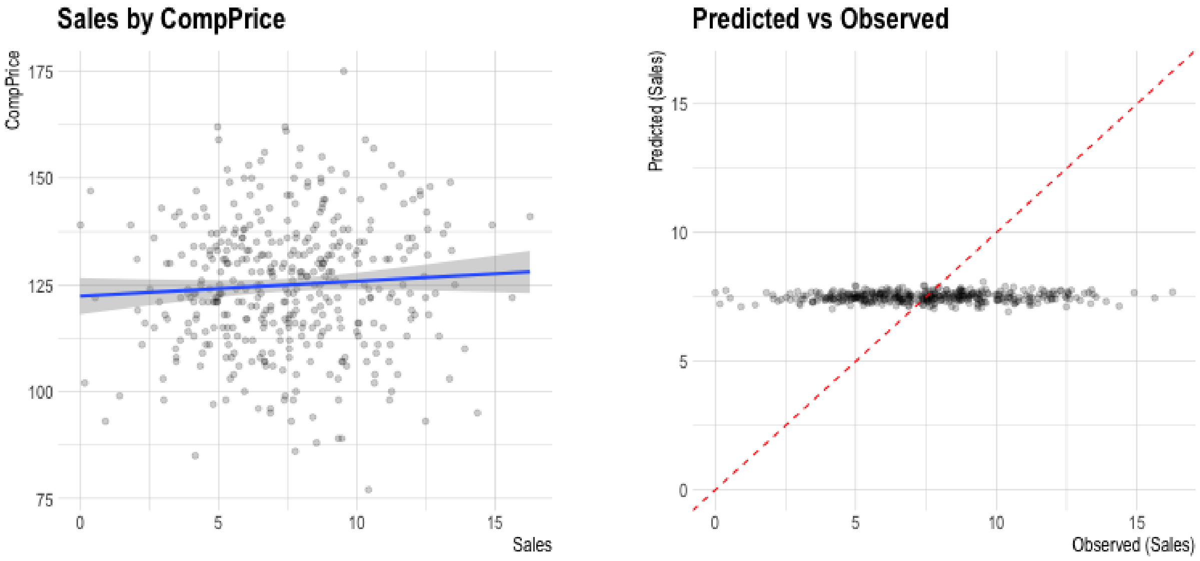


Figure 4.1: CompPrice

Income

1. Simple Linear Model Information

Residual standard error: 3 on 378 degrees of freedom
Multiple R-squared: 0.01817, Adjusted R-squared: 0.01558
F-statistic: 7 on 1 and 378 DF, p-value: 0.0085045

Table 4.2: Simple Linear Model coefficients : Income

	Estimate	Std. Error	t value	$\Pr(>\mid t\mid)$
(Intercept)	6.55	0.38	17.20	0.00
Income	0.01	0.01	2.65	0.01

2. Visualization - Scatterplots

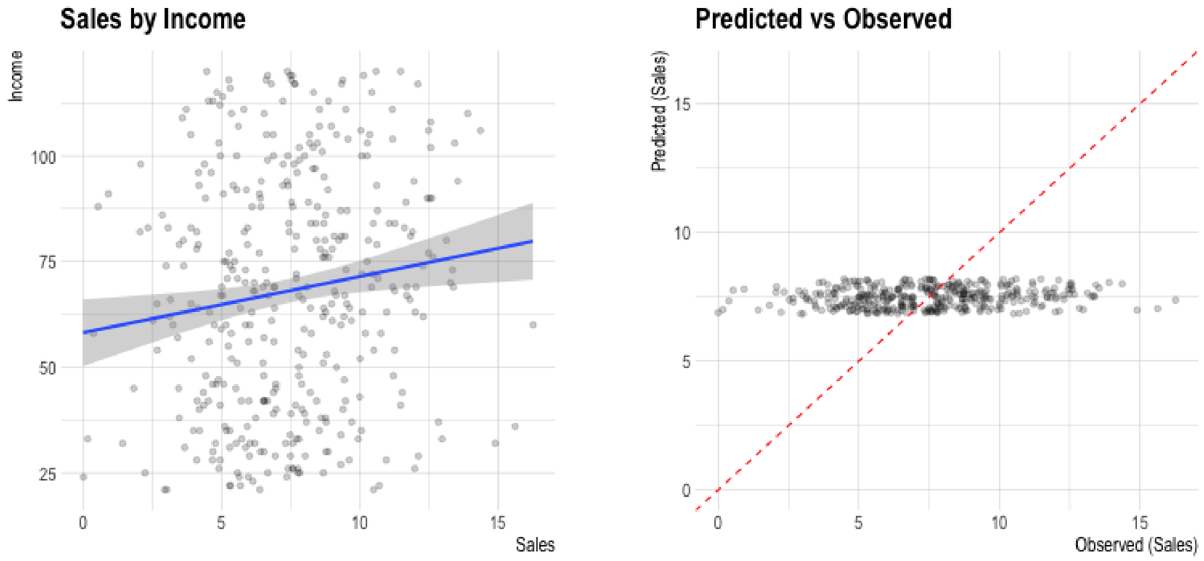


Figure 4.2: Income

Advertising

1. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.07263, Adjusted R-squared: 0.0703
F-statistic: 31 on 1 and 398 DF, p-value: 0

Table 4.3: Simple Linear Model coefficients : Advertising

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	6.74	0.19	35.01	0
Advertising	0.11	0.02	5.58	0

2. Visualization - Scatterplots

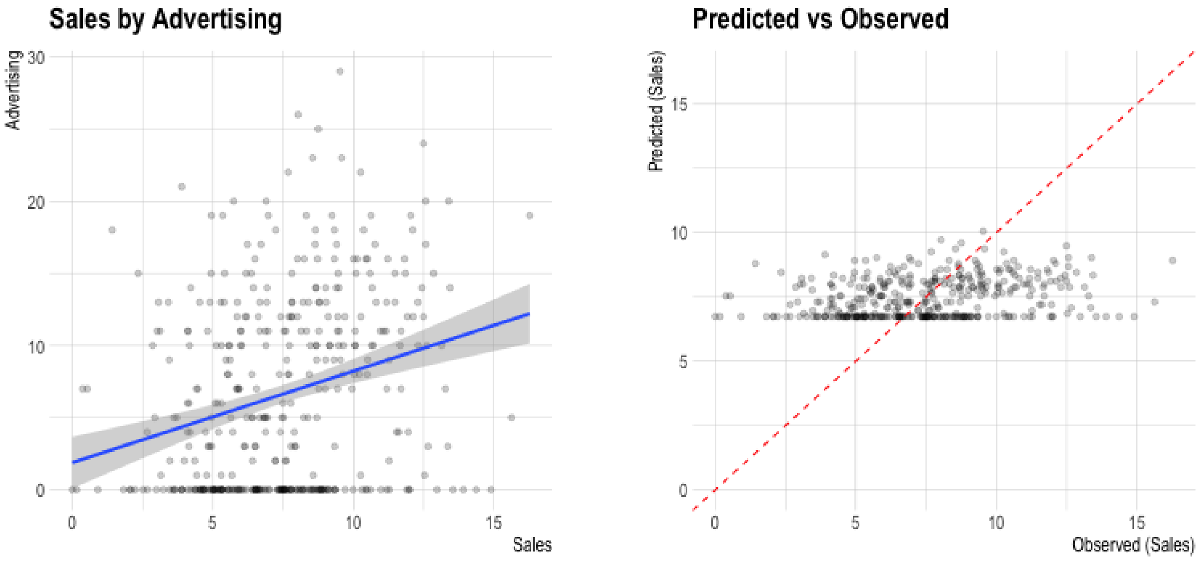


Figure 4.3: Advertising

Population

1. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.00255, Adjusted R-squared: 4e-05
F-statistic: 1 on 1 and 398 DF, p-value: 0.3139816

Table 4.4: Simple Linear Model coefficients : Population

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	7.24	0.29	24.91	0.00
Population	0.00	0.00	1.01	0.31

2. Visualization - Scatterplots

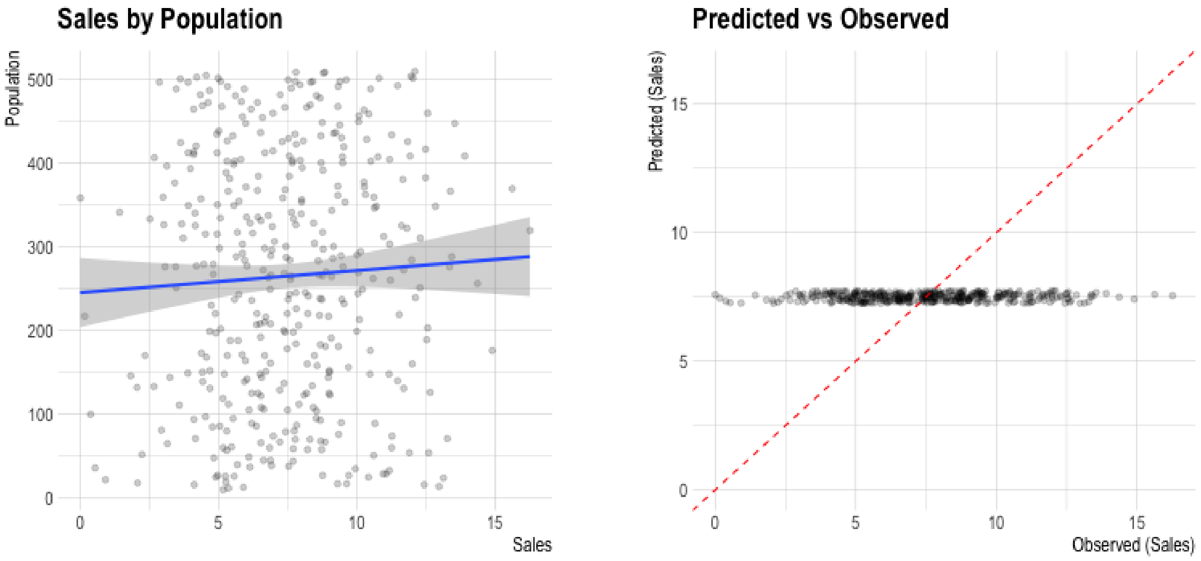


Figure 4.4: Population

Price

1. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.19798, Adjusted R-squared: 0.19597
F-statistic: 98 on 1 and 398 DF, p-value: 0

Table 4.5: Simple Linear Model coefficients : Price

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	13.64	0.63	21.56	0
Price	-0.05	0.01	-9.91	0

2. Visualization - Scatterplots

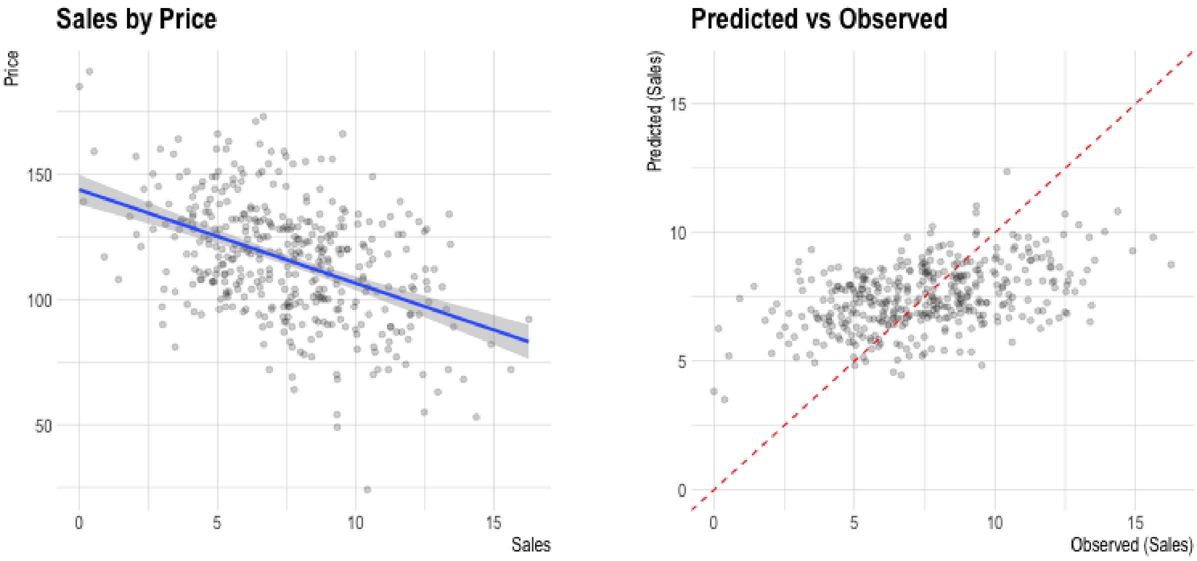


Figure 4.5: Price

Age

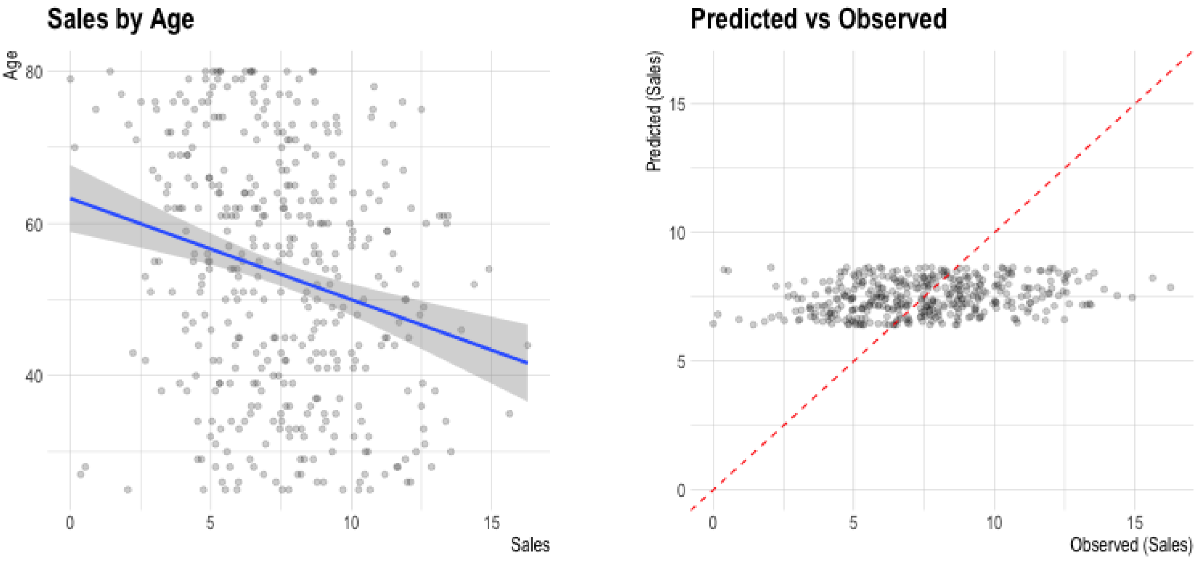
1. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.05374, Adjusted R-squared: 0.05136
F-statistic: 23 on 1 and 398 DF, p-value: 2.8e-06

Table 4.6: Simple Linear Model coefficients : Age

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	9.65	0.47	20.38	0
Age	-0.04	0.01	-4.75	0

2. Visualization - Scatterplots



Sales by Age

Predicted vs Observed

Figure 4.6: Age

Education

1. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.0027, Adjusted R-squared: 0.00019
F-statistic: 1 on 1 and 398 DF, p-value: 0.2999442

Table 4.7: Simple Linear Model coefficients : Education

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	8.27	0.76	10.84	0.0
Education	-0.06	0.05	-1.04	0.3

2. Visualization - Scatterplots

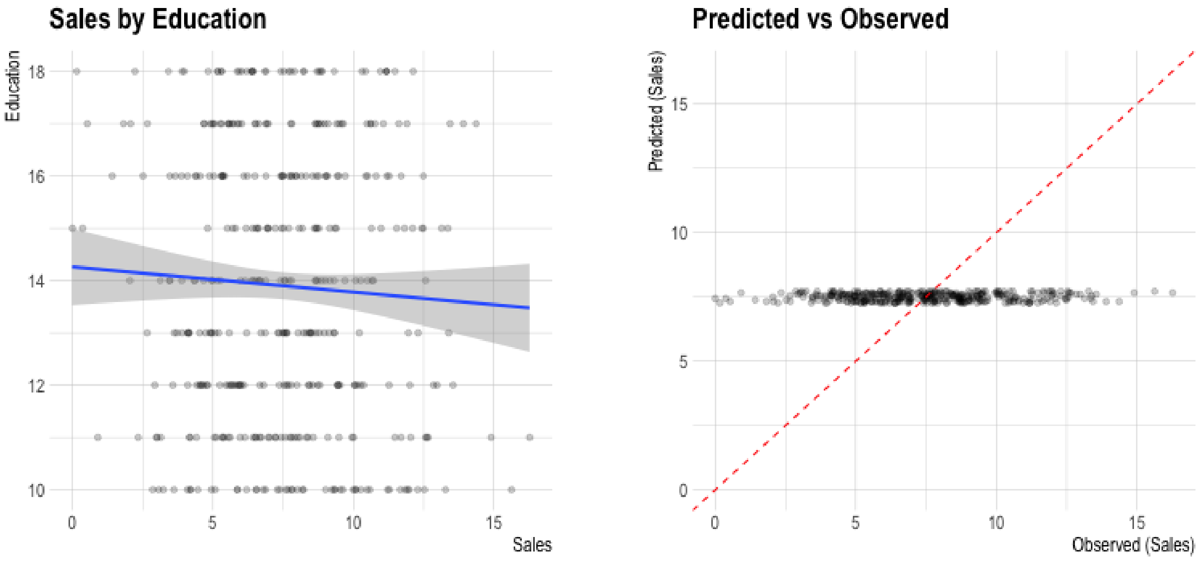


Figure 4.7: Education

4.1.2 Grouped Categorical Variables

ShelveLoc

1. **Analysis of Variance**

Table 4.8: Analysis of Variance Table : ShelveLoc

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
ShelveLoc	2	1009.53	504.77	92.23	0
Residuals	397	2172.74	5.47	NA	NA

2. **Simple Linear Model Information**

Residual standard error: 2 on 397 degrees of freedom
Multiple R-squared: 0.31724, Adjusted R-squared: 0.3138
F-statistic: 92 on 2 and 397 DF, p-value: 0

Table 4.9: Simple Linear Model coefficients : ShelveLoc

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	5.52	0.24	23.13	0
ShelveLocGood	4.69	0.35	13.46	0
ShelveLocMedium	1.78	0.29	6.23	0

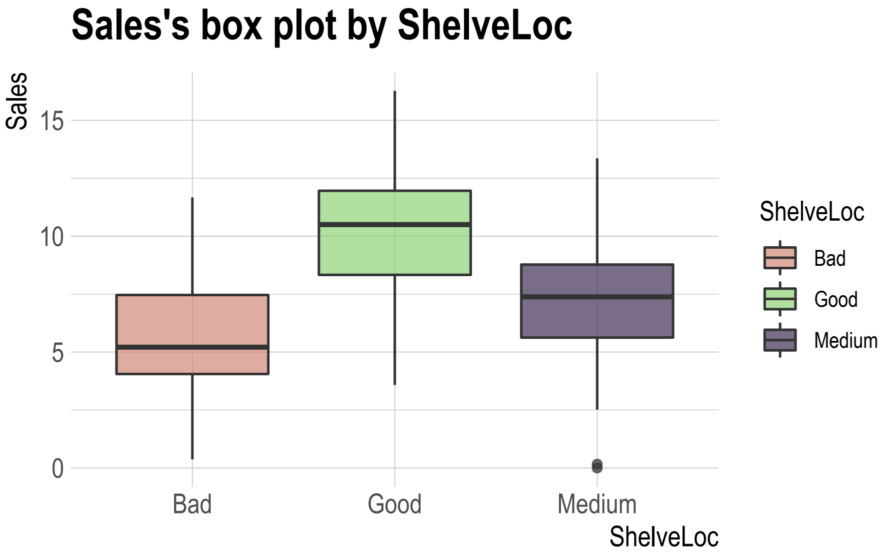


Figure 4.8: ShelveLoc

Urban

1. Analysis of Variance

Table 4.10: Analysis of Variance Table : Urban

	Df	Sum Sq	Mean Sq	F value	$\Pr(>\mid F \mid)$
Urban	1	0.31	0.31	0.04	0.84
Residuals	393	3139.23	7.99	NA	NA

2. Simple Linear Model Information

Residual standard error: 3 on 393 degrees of freedom
Multiple R-squared: 1e-04, Adjusted R-squared: -0.00245
F-statistic: 0 on 1 and 393 DF, p-value: 0.8444621

Table 4.11: Simple Linear Model coefficients : Urban

	Estimate	Std. Error	t value	$\Pr(>\mid t \mid)$
(Intercept)	7.53	0.26	28.71	0.00
UrbanYes	-0.06	0.31	-0.20	0.84

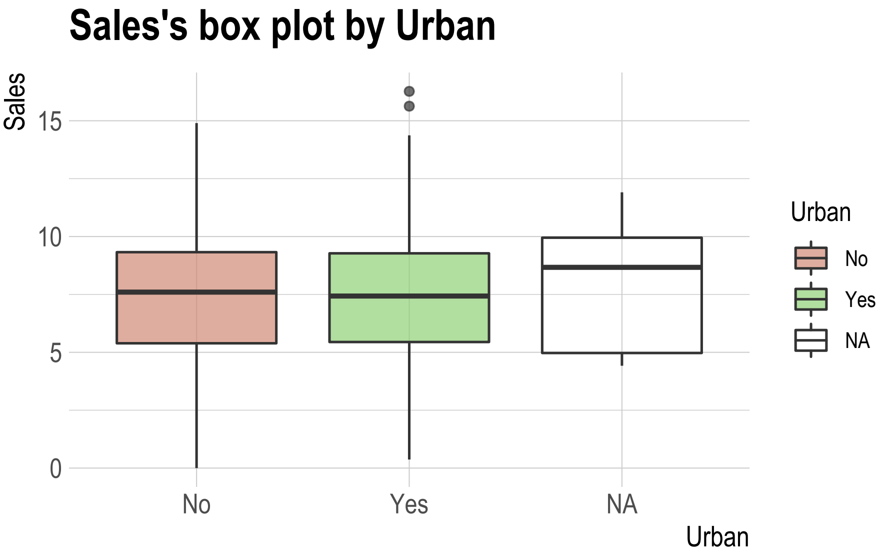


Figure 4.9: Urban

US

1. Analysis of Variance

Table 4.12: Analysis of Variance Table : US

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
US	1	99.80	99.80	12.89	0
Residuals	398	3082.47	7.74	NA	NA

2. Simple Linear Model Information

Residual standard error: 3 on 398 degrees of freedom
Multiple R-squared: 0.03136, Adjusted R-squared: 0.02893
F-statistic: 13 on 1 and 398 DF, p-value: 0.0003723

Table 4.13: Simple Linear Model coefficients : US

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	6.82	0.23	29.22	0
USYes	1.04	0.29	3.59	0

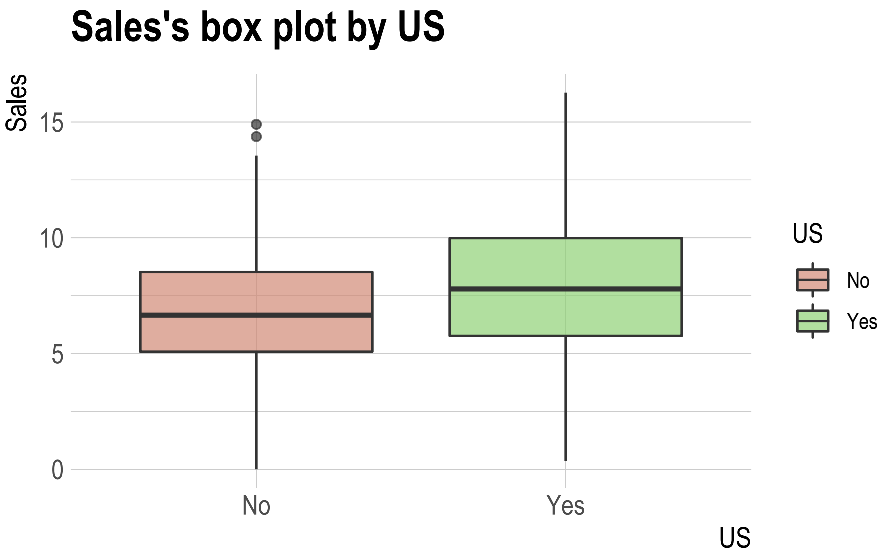


Figure 4.10: US

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

Numerical target variables are not supported.

4.2.2 Grouped Correlation Plot of Numerical Variables

Numerical target variables are not supported.