REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.3.12

March 27, 2020

Contents

1 Introduction 3
1.1 Information of Dataset 3
1.2 Information of Variables 3
1.3 About EDA Report 3

2 Univariate Analysis 5
2.1 Descriptive Statistics 5
2.2 Normality Test of Numerical Variables 7
2.2.1 Statistics and Visualization of (Sample) Data 7

3 Relationship Between Variables 11
3.1 Correlation Coefficient 11
3.1.1 Correlation Coefficient by Variable Combination 11
3.1.2 Correlation Plot of Numerical Variables 11

4 Target based Analysis 13
4.1 Grouped Descriptive Statistics 13
4.1.1 Grouped Numerical Variables 13
4.1.2 Grouped Categorical Variables 18
4.2 Grouped Relationship Between Variables 28
4.2.1 Grouped Correlation Coefficient 28
4.2.2 Grouped Correlation Plot of Numerical Variables 28

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 234 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
manufacturer	factor	0	0	15	0.0641026
model	factor	0	0	38	0.1623932
displ	numeric	0	0	35	0.1495726
year	integer	0	0	2	0.0085470
cyl	integer	0	0	4	0.0170940
trans	factor	0	0	10	0.0427350
drv	factor	0	0	3	0.0128205
cty	integer	0	0	21	0.0897436
hwy	integer	0	0	27	0.1153846
fl	factor	0	0	5	0.0213675
class	factor	0	0	7	0.0299145

The target variable of the data is 'hwy', and the data type of the variable is integer.

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 234 Observations

---

manufacturer

n	missing	distinct
234	0	15

```
lowest : audi       chevrolet  dodge      ford       honda
highest: nissan     pontiac    subaru     toyota     volkswagen
```

audi (18, 0.077), chevrolet (19, 0.081), dodge (37, 0.158), ford (25, 0.107), honda (9, 0.038), hyundai (14, 0.060), jeep (8, 0.034), land rover (4, 0.017), lincoln (3, 0.013), mercury (4, 0.017), nissan (13, 0.056), pontiac (5, 0.021), subaru (14, 0.060), toyota (34, 0.145), volkswagen (27, 0.115)

---

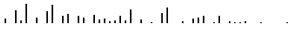
model

n	missing	distinct
234	0	38

```
lowest : 4runner 4wd         a4                  a4 quattro          a6 quattro          altima
highest: ram 1500 pickup 4wd range rover         sonata              tiburon             toyota tacoma 4wd
```

---

displ



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
234	0	35	0.997	3.472	1.471	1.8	2.0	2.4	3.3	4.6	5.4	5.7

```
lowest : 1.6 1.8 1.9 2.0 2.2, highest: 6.0 6.1 6.2 6.5 7.0
```

---

year

n	missing	distinct	Info	Mean	Gmd
234	0	2	0.75	2004	4.519

Value	1999	2008
Frequency	117	117
Proportion	0.5	0.5

---

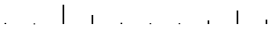
cyl

n	missing	distinct	Info	Mean	Gmd
234	0	4	0.893	5.889	1.761

Value	4	5	6	8
Frequency	81	4	79	70
Proportion	0.346	0.017	0.338	0.299

---

trans



n	missing	distinct
234	0	10

```
lowest : auto(av)   auto(l3)   auto(l4)   auto(l5)   auto(l6)
highest: auto(s4)   auto(s5)   auto(s6)   manual(m5) manual(m6)
```

Value	auto(av)	auto(l3)	auto(l4)	auto(l5)	auto(l6)	auto(s4)	auto(s5)
Frequency	5	2	83	39	6	3	3
Proportion	0.021	0.009	0.355	0.167	0.026	0.013	0.013

Value	auto(s6)	manual(m5)	manual(m6)
Frequency	16	58	19
Proportion	0.068	0.248	0.081

---

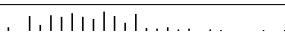
drv

n	missing	distinct
234	0	3

Value	4	f	r
Frequency	103	106	25
Proportion	0.440	0.453	0.107

---

cty

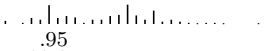


n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
234	0	21	0.993	16.86	4.686	11	11	14	17	19	21	24

```
lowest :  9 11 12 13 14, highest: 26 28 29 33 35
```

---

hwy



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
234	0	27	0.993	23.44	6.668	15.0	16.3	18.0	24.0	27.0	30.0	32.0

```
lowest : 12 14 15 16 17, highest: 35 36 37 41 44
```

---

fl

n	missing	distinct
234	0	5

```
lowest : c d e p r, highest: c d e p r
```

Value	c	d	e	p	r
Frequency	1	5	8	52	168
Proportion	0.004	0.021	0.034	0.222	0.718

---

class

n	missing	distinct
234	0	7

```
lowest : 2seater    compact    midsize    minivan    pickup
highest: midsize    minivan    pickup     subcompact suv
```

Value	2seater	compact	midsize	minivan	pickup	subcompact	suv
Frequency	5	47	41	11	33	35	62
Proportion	0.021	0.201	0.175	0.047	0.141	0.150	0.265

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

displ

normality test : Shapiro-Wilk normality test
statistic : 0.9408, p-value : 3.93641E-08

type	skewness	kurtosis
original	0.4415	2.1074
log transformation	-0.0344	1.8404
sqrt transformation	0.2034	1.8952

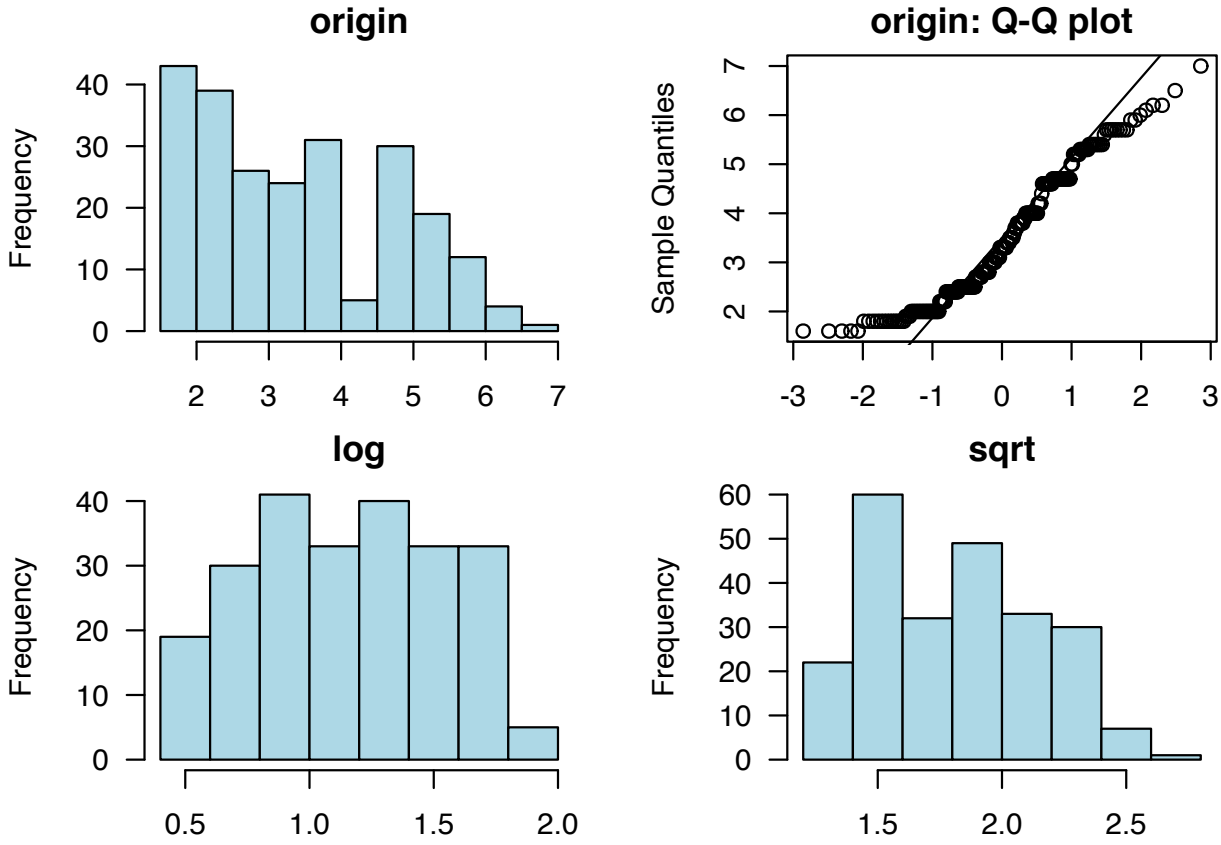


Figure 2.1: displ

year

normality test : Shapiro-Wilk normality test
statistic : 0.63646, p-value : 4.90845E-22

type	skewness	kurtosis
original	0.0000	1.0000
log transformation	0.0000	1.0000
sqrt transformation	0.0000	1.0000

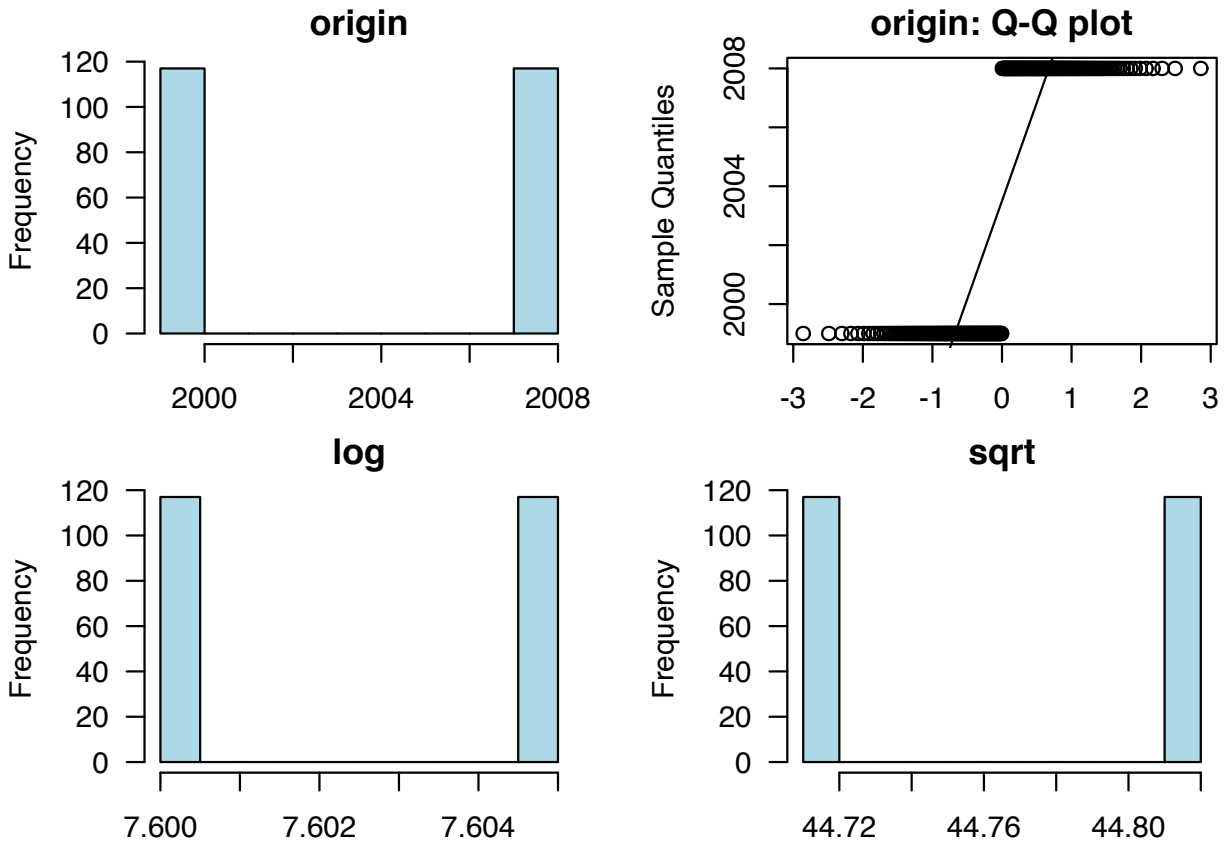


Figure 2.2: year

cyl

normality test : Shapiro-Wilk normality test
statistic : 0.8001, p-value : 1.28609E-16

type	skewness	kurtosis
original	0.1131	1.5491
log transformation	-0.1024	1.4973
sqrt transformation	0.0035	1.5220

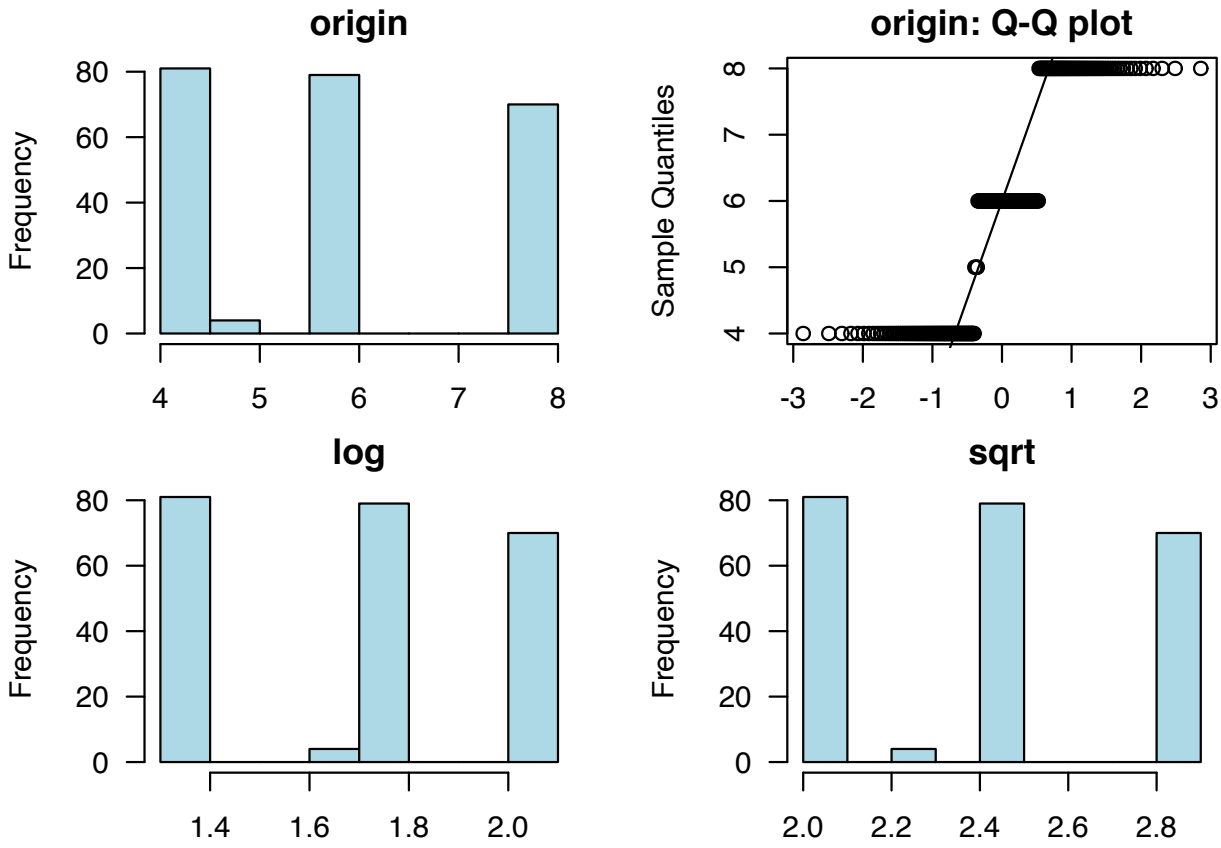


Figure 2.3: cyl

cty

normality test : Shapiro-Wilk normality test
statistic : 0.95679, p-value : 1.7442E-06

type	skewness	kurtosis
original	0.7914	4.4687
log transformation	-0.0247	2.9427
sqrt transformation	0.3572	3.3912

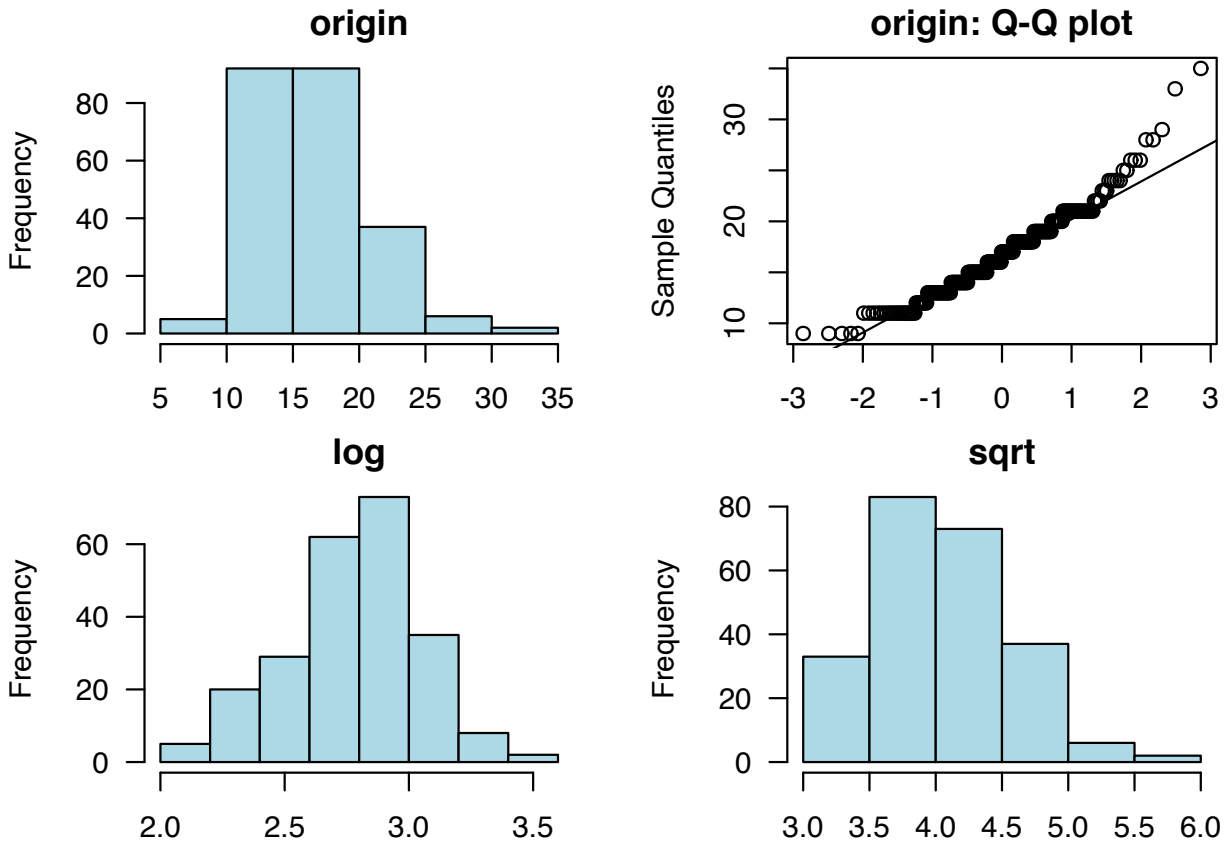


Figure 2.4: cty

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
hwy	cty	0.956
cyl	displ	0.930
cty	cyl	-0.806
cty	displ	-0.799
hwy	displ	-0.766
hwy	cyl	-0.762

3.1.2 Correlation Plot of Numerical Variables

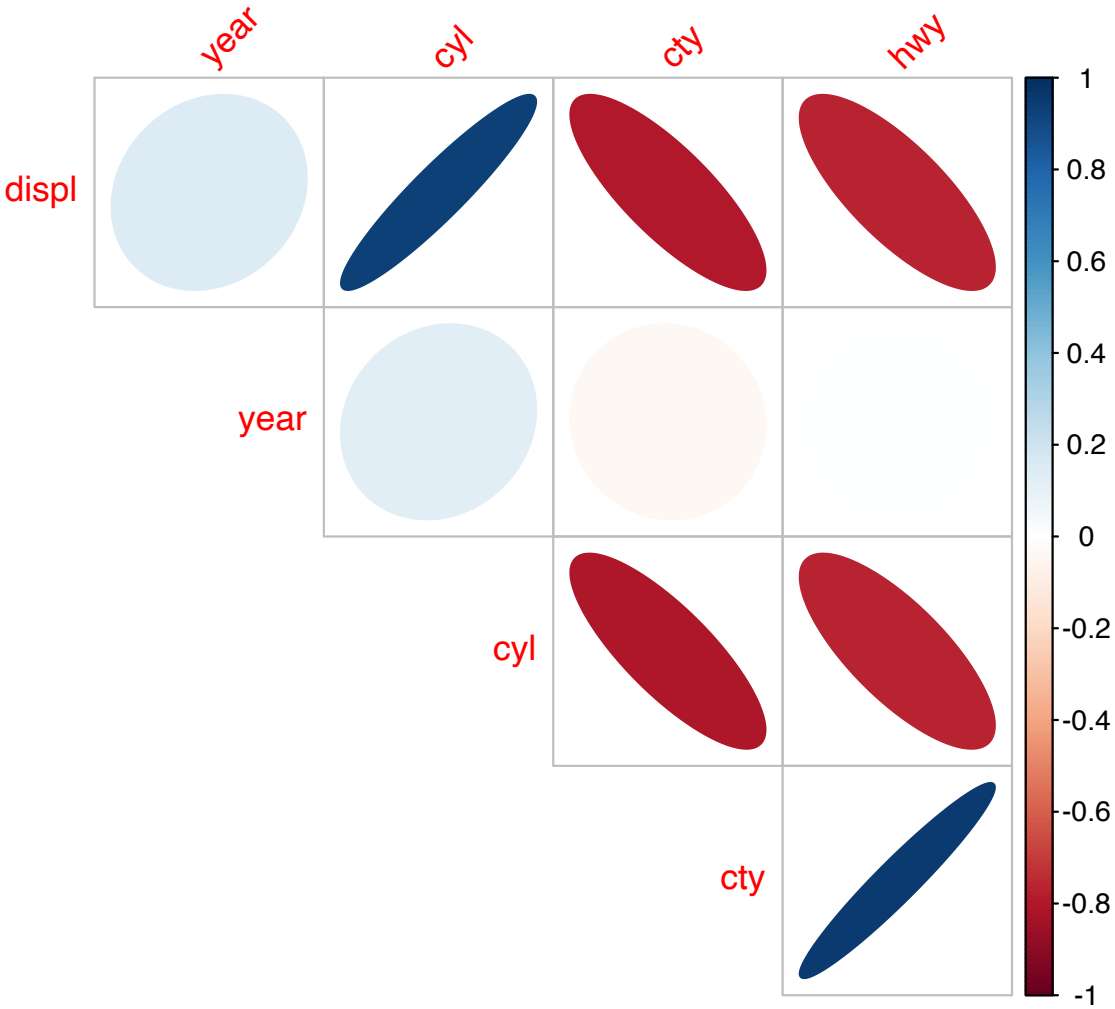


Figure 3.1: The correlation coefficient of numerical variables

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

displ

1. Simple Linear Model Information

Residual standard error: 4 on 232 degrees of freedom
Multiple R-squared: 0.58679, Adjusted R-squared: 0.58501
F-statistic: 329 on 1 and 232 DF, p-value: 0

Table 4.1: Simple Linear Model coefficients : displ

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	35.70	0.72	49.55	0
displ	-3.53	0.19	-18.15	0

2. Visualization - Scatterplots

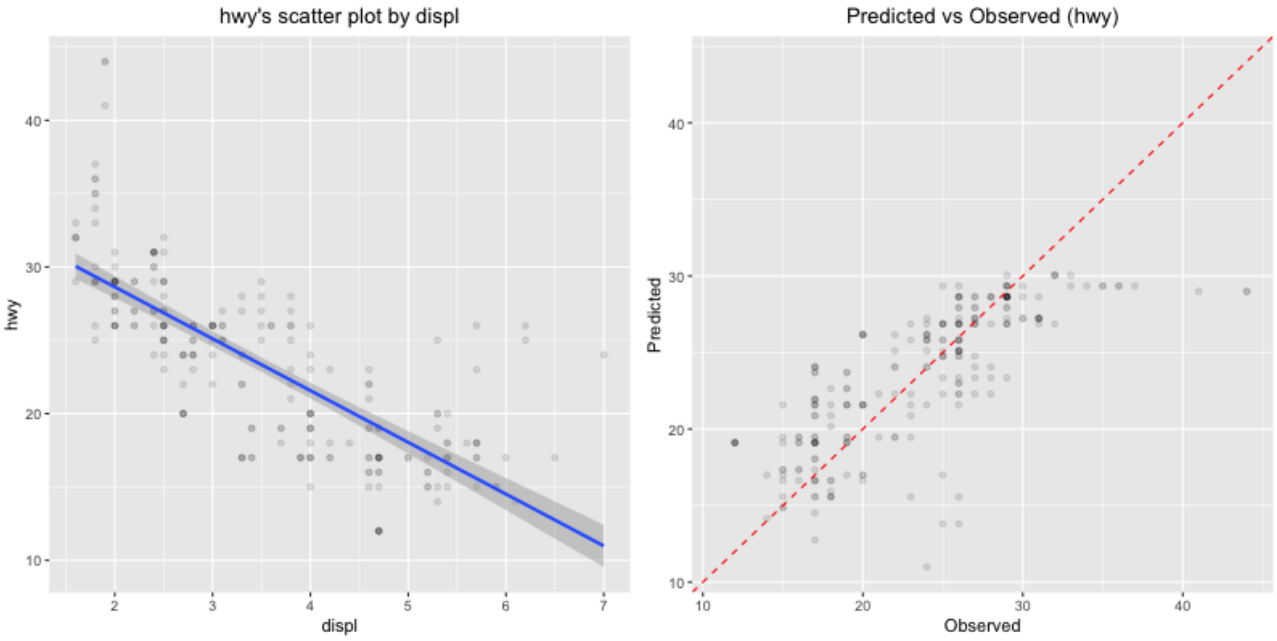


Figure 4.1: displ

year

1. Simple Linear Model Information

Residual standard error: 6 on 232 degrees of freedom
Multiple R-squared: 0, Adjusted R-squared: -0.00431
F-statistic: 0 on 1 and 232 DF, p-value: 0.973811

Table 4.2: Simple Linear Model coefficients : year

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	17.73	173.68	0.10	0.92
year	0.00	0.09	0.03	0.97

2. Visualization - Scatterplots

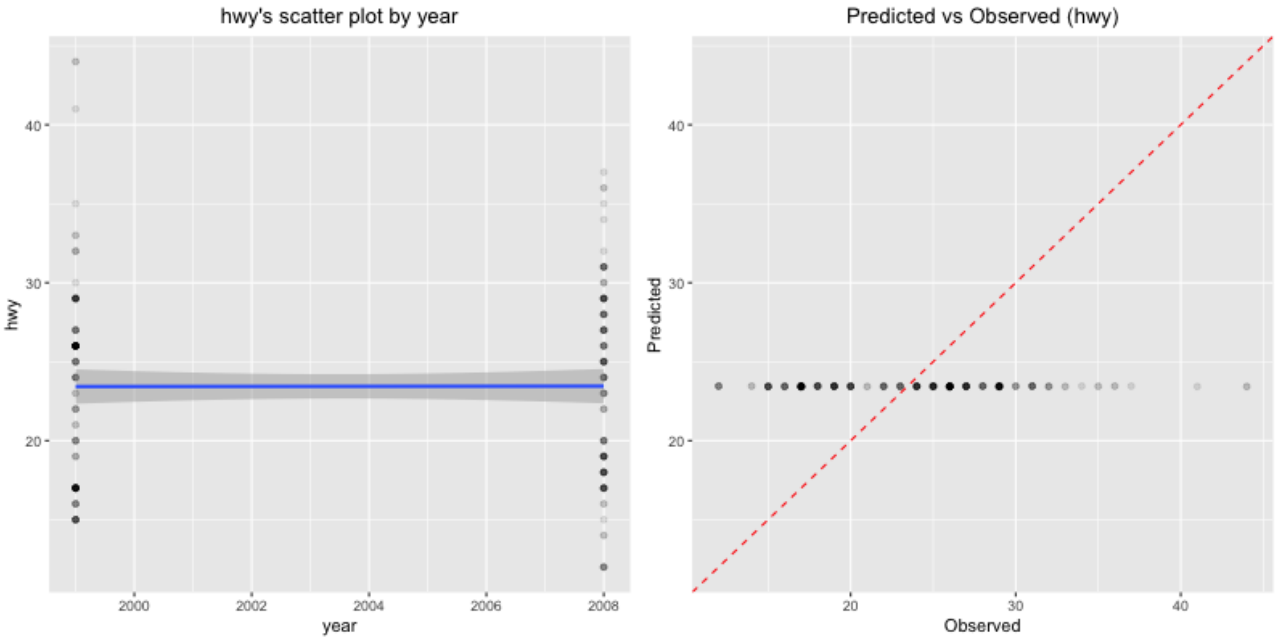


Figure 4.2: year

cyl

1. Simple Linear Model Information

Residual standard error: 4 on 232 degrees of freedom
Multiple R-squared: 0.58051, Adjusted R-squared: 0.5787
F-statistic: 321 on 1 and 232 DF, p-value: 0

Table 4.3: Simple Linear Model coefficients : cyl

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	40.02	0.96	41.72	0
cyl	-2.82	0.16	-17.92	0

2. Visualization - Scatterplots

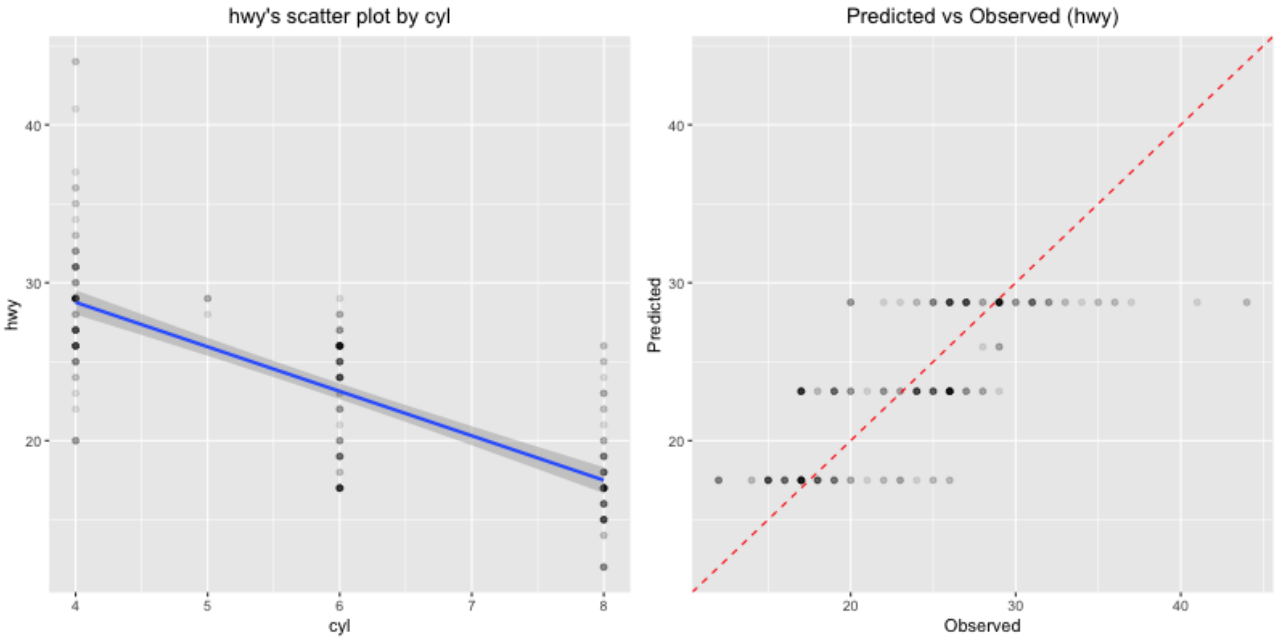


Figure 4.3: cyl

cty

1. Simple Linear Model Information

Residual standard error: 2 on 232 degrees of freedom
Multiple R-squared: 0.91378, Adjusted R-squared: 0.9134
F-statistic: 2459 on 1 and 232 DF, p-value: 0

Table 4.4: Simple Linear Model coefficients : cty

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	0.89	0.47	1.90	0.06
cty	1.34	0.03	49.58	0.00

2. Visualization - Scatterplots

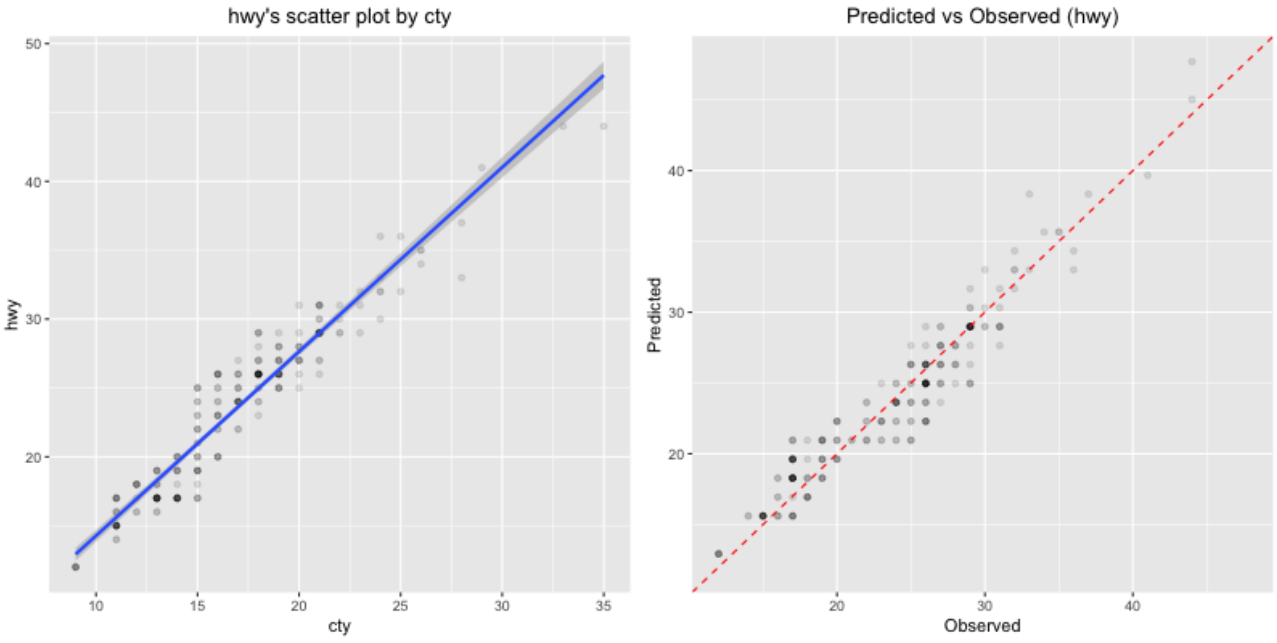


Figure 4.4: cty

4.1.2 Grouped Categorical Variables

manufacturer

1. Analysis of Variance

Table 4.5: Analysis of Variance Table : manufacturer

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
manufacturer	14	4459.86	318.56	18.35	0
Residuals	219	3801.80	17.36	NA	NA

2. Simple Linear Model Information

Residual standard error: 4 on 219 degrees of freedom
Multiple R-squared: 0.53983, Adjusted R-squared: 0.51041
F-statistic: 18 on 14 and 219 DF, p-value: 0.0010548

Table 4.6: Simple Linear Model coefficients : manufacturer

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	26.44	0.98	26.93	0.00
manufacturerchevrolet	-4.55	1.37	-3.32	0.00
manufacturerdodge	-8.50	1.20	-7.10	0.00
manufacturerford	-7.08	1.29	-5.50	0.00
manufacturerhonda	6.11	1.70	3.59	0.00
manufacturerhyundai	0.41	1.48	0.28	0.78
manufacturerjeep	-8.82	1.77	-4.98	0.00
manufacturerland rover	-9.94	2.30	-4.32	0.00
manufacturerlincoln	-9.44	2.60	-3.63	0.00
manufacturermercury	-8.44	2.30	-3.67	0.00
manufacturernissan	-1.83	1.52	-1.21	0.23
manufacturerpontiac	-0.04	2.11	-0.02	0.98
manufacturersubaru	-0.87	1.48	-0.59	0.56
manufacturertoyota	-1.53	1.21	-1.26	0.21
manufacturervolkswagen	2.78	1.27	2.19	0.03

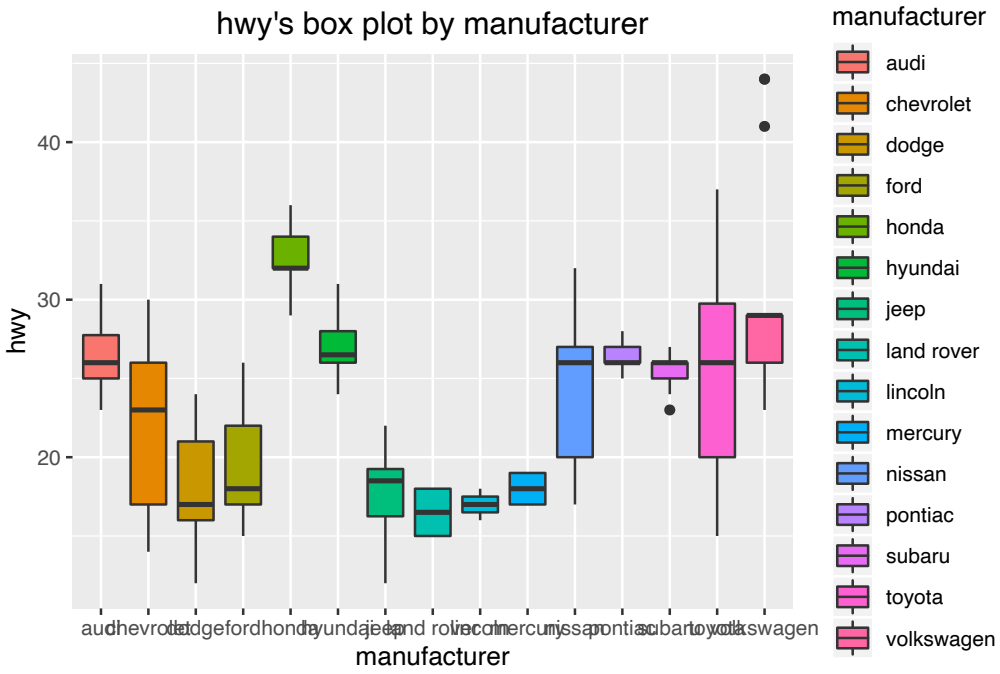


Figure 4.5: manufacturer

model

1. Analysis of Variance

Table 4.7: Analysis of Variance Table : model

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
model	37	7000.91	189.21	29.42	0
Residuals	196	1260.76	6.43	NA	NA

2. Simple Linear Model Information

Residual standard error: 3 on 196 degrees of freedom
Multiple R-squared: 0.8474, Adjusted R-squared: 0.81859
F-statistic: 29 on 37 and 196 DF, p-value: 0

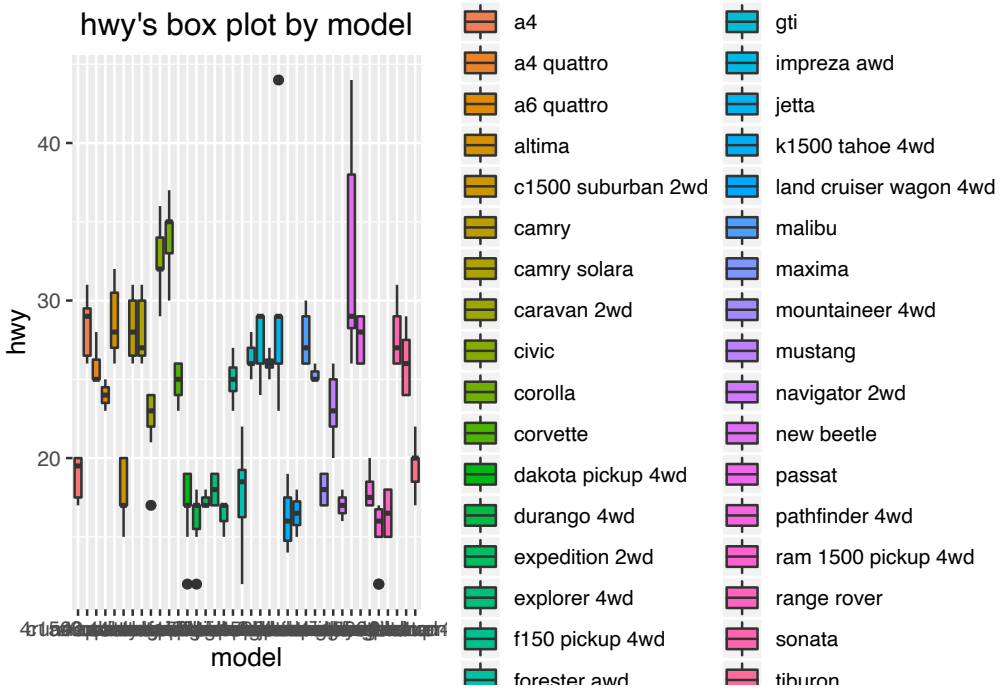


Figure 4.6: model

Table 4.8: Simple Linear Model coefficients : model

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	18.83	1.04	18.19	0.00
modela4	9.45	1.41	6.70	0.00
modela4 quattro	6.92	1.37	5.05	0.00
modela6 quattro	5.17	1.79	2.88	0.00
modelaltima	9.83	1.46	6.72	0.00
modelc1500 suburban 2wd	-1.03	1.54	-0.67	0.50
modelcamry	9.45	1.41	6.70	0.00
modelcamry solara	9.31	1.41	6.60	0.00
modelcaravan 2wd	3.53	1.29	2.74	0.01
modelcivic	13.72	1.34	10.27	0.00
modelcorolla	15.17	1.54	9.88	0.00
modelcorvette	5.97	1.54	3.89	0.00
modeldakota pickup 4wd	-1.83	1.34	-1.37	0.17
modeldurango 4wd	-2.83	1.41	-2.01	0.05
modelexpedition 2wd	-1.50	1.79	-0.84	0.40
modelexplorer 4wd	-0.83	1.46	-0.57	0.57
modelf150 pickup 4wd	-2.40	1.41	-1.70	0.09
modelforester awd	6.17	1.46	4.21	0.00
modelgrand cherokee 4wd	-1.21	1.37	-0.88	0.38
modelgrand prix	7.57	1.54	4.93	0.00
modelgti	8.57	1.54	5.58	0.00
modelimpreza awd	7.17	1.37	5.23	0.00
modeljetta	10.28	1.34	7.69	0.00
modelk1500 tahoe 4wd	-2.58	1.64	-1.58	0.12
modelland cruiser wagon 4wd	-2.33	2.07	-1.13	0.26
modelmalibu	8.77	1.54	5.71	0.00
modelmaxima	6.50	1.79	3.62	0.00
modelmountaineer 4wd	-0.83	1.64	-0.51	0.61
modelmustang	4.39	1.34	3.28	0.00
modelnavigator 2wd	-1.83	1.79	-1.02	0.31
modelnew beetle	14.00	1.46	9.56	0.00
modelpassat	8.74	1.41	6.19	0.00
modelpathfinder 4wd	-0.83	1.64	-0.51	0.61
modelram 1500 pickup 4wd	-3.53	1.31	-2.70	0.01
modelrange rover	-2.33	1.64	-1.43	0.16
modelsonata	8.88	1.41	6.29	0.00
modeltiburon	7.17	1.41	5.08	0.00
modeltoyota tacoma 4wd	0.60	1.41	0.42	0.67

trans

1. Analysis of Variance

Table 4.9: Analysis of Variance Table : trans

	Df	Sum Sq	Mean Sq	F value	Pr(> | *F* |)
trans	9	1219.13	135.46	4.31	0
Residuals	224	7042.53	31.44	NA	NA

2. Simple Linear Model Information

Residual standard error: 6 on 224 degrees of freedom
Multiple R-squared: 0.14756, Adjusted R-squared: 0.11331
F-statistic: 4 on 9 and 224 DF, p-value: 0.8647474

Table 4.10: Simple Linear Model coefficients : trans

	Estimate	Std. Error	t value	Pr(> | *t* |)
(Intercept)	27.80	2.51	11.09	0.00
transauto(l3)	-0.80	4.69	-0.17	0.86
transauto(l4)	-5.84	2.58	-2.26	0.02
transauto(l5)	-7.08	2.66	-2.66	0.01
transauto(l6)	-7.80	3.40	-2.30	0.02
transauto(s4)	-2.13	4.09	-0.52	0.60
transauto(s5)	-2.47	4.09	-0.60	0.55
transauto(s6)	-2.61	2.87	-0.91	0.36
transmanual(m5)	-1.51	2.61	-0.58	0.56
transmanual(m6)	-3.59	2.82	-1.27	0.20

Figure 4.7: trans

drv

1. Analysis of Variance

Table 4.11: Analysis of Variance Table : drv

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
drv	2	4384.53	2192.27	130.62	0
Residuals	231	3877.13	16.78	NA	NA

2. Simple Linear Model Information

Residual standard error: 4 on 231 degrees of freedom
Multiple R-squared: 0.53071, Adjusted R-squared: 0.52665
F-statistic: 131 on 2 and 231 DF, p-value: 0

Table 4.12: Simple Linear Model coefficients : drv

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	19.17	0.40	47.50	0.00
drvf	8.99	0.57	15.85	0.00
drvr	1.83	0.91	2.00	0.05

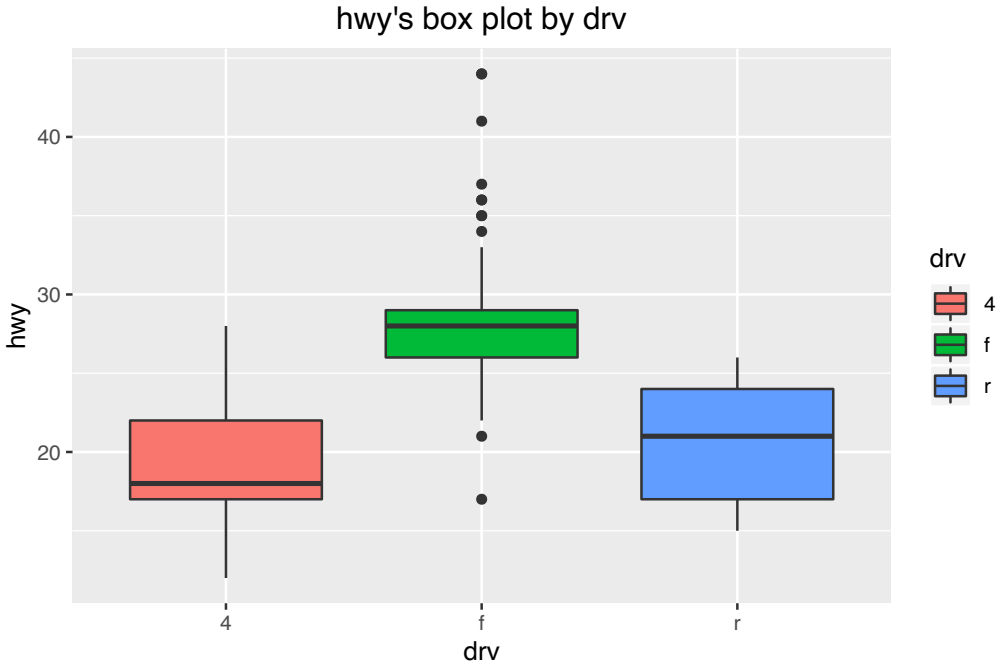


Figure 4.8: drv

fl

1. Analysis of Variance

Table 4.13: Analysis of Variance Table : fl

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
fl	4	1704.74	426.18	14.88	0
Residuals	229	6556.92	28.63	NA	NA

2. Simple Linear Model Information

Residual standard error: 5 on 229 degrees of freedom
Multiple R-squared: 0.20634, Adjusted R-squared: 0.19248
F-statistic: 15 on 4 and 229 DF, p-value: 0.6826009

Table 4.14: Simple Linear Model coefficients : fl

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	36.00	5.35	6.73	0.00
fld	-2.40	5.86	-0.41	0.68
fle	-22.75	5.68	-4.01	0.00
flp	-10.77	5.40	-1.99	0.05
flr	-13.01	5.37	-2.42	0.02

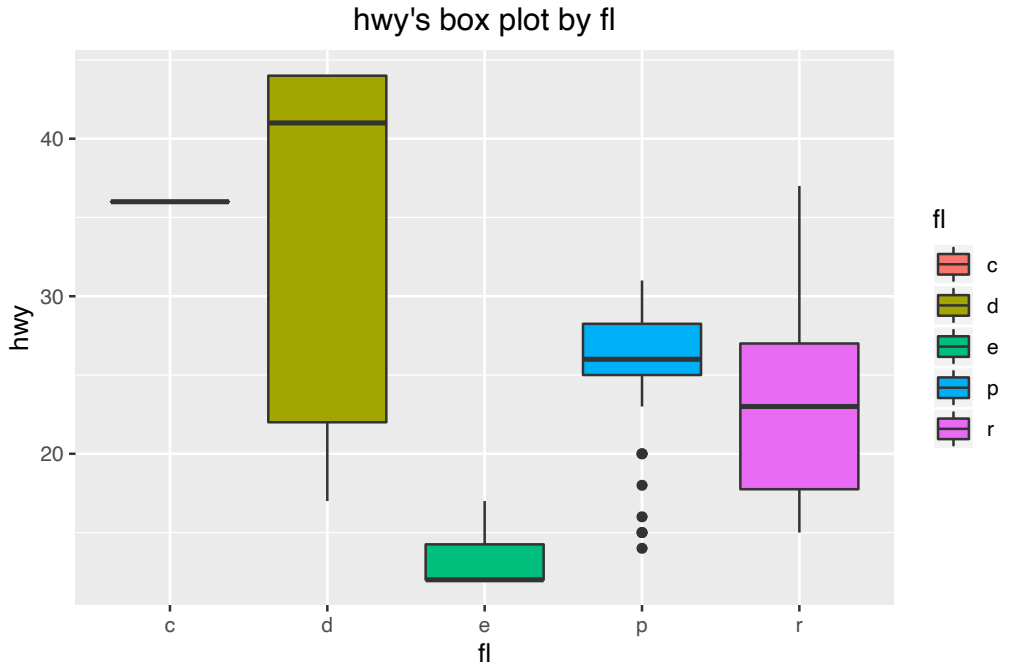


Figure 4.9: fl

class

1. Analysis of Variance

Table 4.15: Analysis of Variance Table : class

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
class	6	5683.23	947.21	83.39	0
Residuals	227	2578.43	11.36	NA	NA

2. Simple Linear Model Information

Residual standard error: 3 on 227 degrees of freedom
Multiple R-squared: 0.6879, Adjusted R-squared: 0.67965
F-statistic: 83 on 6 and 227 DF, p-value: 0.0283629

Table 4.16: Simple Linear Model coefficients : class

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	24.80	1.51	16.45	0.00
classcompact	3.50	1.59	2.21	0.03
classmidsize	2.49	1.60	1.56	0.12
classminivan	-2.44	1.82	-1.34	0.18
classpickup	-7.92	1.62	-4.90	0.00
classsubcompact	3.34	1.61	2.07	0.04
classsuv	-6.67	1.57	-4.26	0.00

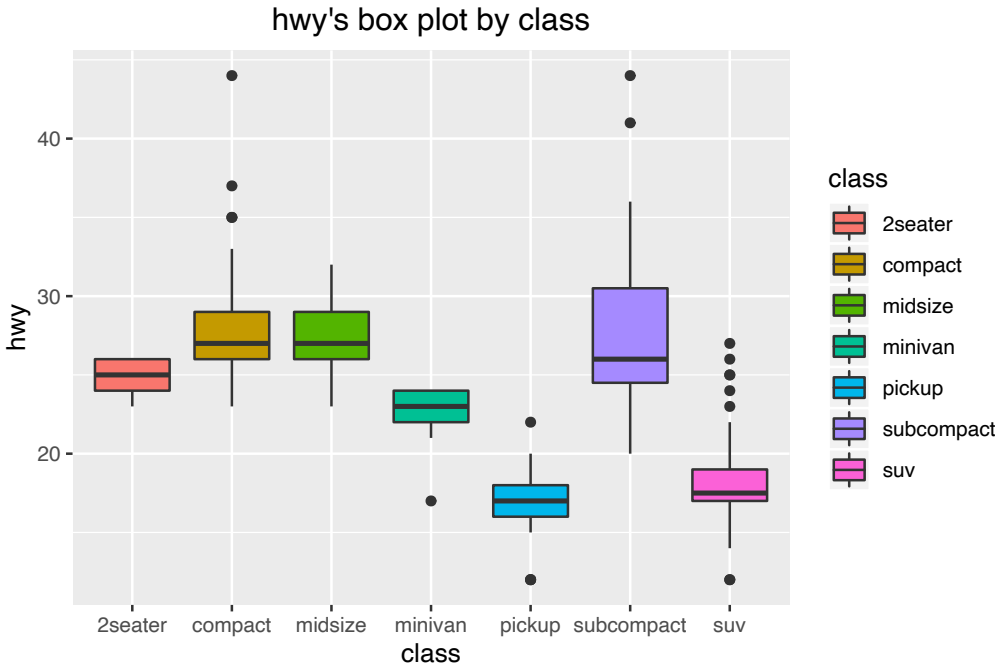


Figure 4.10: class

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.