

What is the complete guide to the diamonds dataset in R?

Authored by
stats writer

June 30, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the complete guide to the diamonds dataset in R?*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=161782>

The complete guide to the diamonds dataset in R is a comprehensive resource that provides detailed information and instructions on how to work with the diamonds dataset in the R programming language. This dataset contains information on various characteristics of over 53,000 diamonds, such as carat weight, cut, color, and clarity. The guide covers topics such as data importing, cleaning, manipulation, and visualization, as well as statistical analysis and machine learning techniques that can be applied to the dataset. It also includes sample code and step-by-step instructions to help users effectively utilize the dataset for their own projects and analyses. Overall, the complete guide aims to provide a thorough understanding of the diamonds dataset and how it can be utilized in R for data analysis.

A Complete Guide to the diamonds Dataset in R

The diamonds dataset is a dataset that comes built-in with the package in R.

It contains measurements on 10 different variables (like price, color, clarity, etc.) for 53,940 different diamonds.

This tutorial explains how to explore, summarize, and visualize the diamonds dataset in R.

Load the diamonds Dataset

Since the diamonds dataset is a built-in dataset in ggplot2, we must first install (if we haven't already) and load the ggplot2 package:

```
#install ggplot2 if not already installed  
install.packages('ggplot2')
```

```
#load ggplot2  
library(ggplot2)
```

Once we've loaded ggplot2, we can use the data() function to load the diamonds dataset:

```
data(diamonds)
```

We can take a look at the first six rows of the dataset by using the head() function:

```
#view first six rows of diamonds dataset  
head(diamonds)
```

```
carat cut color clarity depth table price x y z  
1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
4 0.290 Premium I VS2 62.4 58 334 4.2 4.23 2.63  
5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48
```

Summarize the diamonds Dataset

We can use the summary() function to quickly

summarize each variable in the dataset:

#summarize diamonds dataset

summary(diamonds)

carat cut color clarity depth

Min. :0.2000 Fair : 1610 D: 6775 SI1 :13065 Min. :43.00

**1st Qu.:0.4000 Good : 4906 E: 9797 VS2 :12258 1st
Qu.:61.00**

**Median :0.7000 Very Good:12082 F: 9542 SI2 : 9194
Median :61.80**

**Mean :0.7979 Premium :13791 G:11292 VS1 : 8171 Mean
:61.75**

**3rd Qu.:1.0400 Ideal :21551 H: 8304 VVS2 : 5066 3rd
Qu.:62.50**

Max. :5.0100 I: 5422 VVS1 : 3655 Max. :79.00

J: 2808 (Other): 2531

table price x y z

**Min. :43.00 Min. : 326 Min. : 0.000 Min. : 0.000 Min. :
0.000**

**1st Qu.:56.00 1st Qu.: 950 1st Qu.: 4.710 1st Qu.: 4.720
1st Qu.: 2.910**

**Median :57.00 Median : 2401 Median : 5.700 Median :
5.710 Median : 3.530**

**Mean :57.46 Mean : 3933 Mean : 5.731 Mean : 5.735
Mean : 3.539
3rd Qu.:59.00 3rd Qu.: 5324 3rd Qu.: 6.540 3rd Qu.: 6.540
3rd Qu.: 4.040
Max. :95.00 Max. :18823 Max. :10.740 Max. :58.900 Max.
:31.800**

For each of the numeric variables we can see the following information:

Min: The minimum value.1st Qu: The value of the first quartile (25th percentile).Median: The median value.Mean: The mean value.3rd Qu: The value of the third quartile (75th percentile).Max: The maximum value.

For the categorical variables in the dataset (cut, color, and clarity) we see a frequency count of each value.

For example, for the cut variable:

Fair: This value occurs 1,610 times.Good: This value occurs 4,906 times.Very Good: This value occurs 12,082 times.Premium: This value occurs 13,791 times.Ideal: This value occurs 21,551 times.

We can use the `dim()` function to get the dimensions of the dataset in terms of number of rows and number of columns:

```
#display rows and columns  
dim(diamonds)
```

```
53940 10
```

We can see that the dataset has 53,940 rows and 10 columns.

We can also use the `names()` function to display the column names of the data frame:

```
#display column names  
names(diamonds)
```

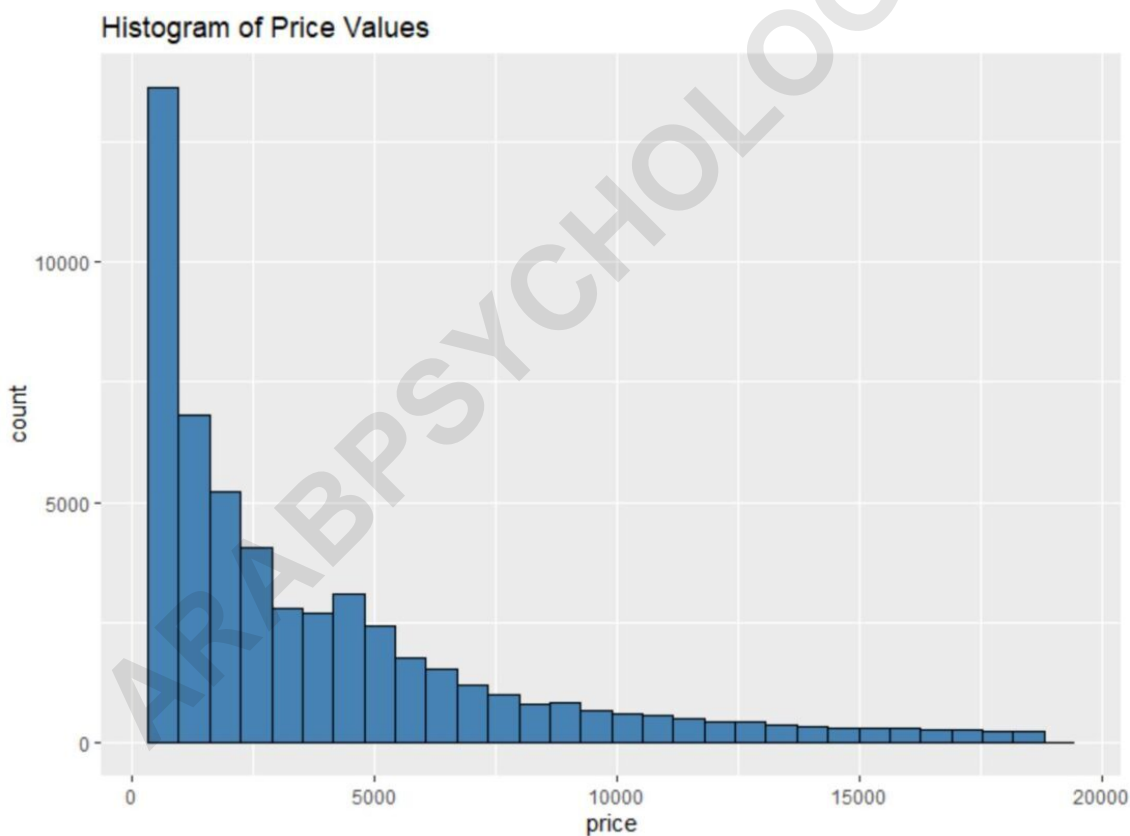
```
"carat" "cut" "color" "clarity" "depth" "table" "price"  
"x"  
"y" "z"
```

Visualize the diamonds Dataset

We can also create some plots to visualize the values in the dataset.

For example, we can use the `geom_histogram()` function to create a histogram of the values for a certain variable:

```
#create histogram of values for price  
ggplot(data=diamonds, aes(x=price)) +  
geom_histogram(fill="steelblue", color="black") +  
ggtitle("Histogram of Price Values")
```



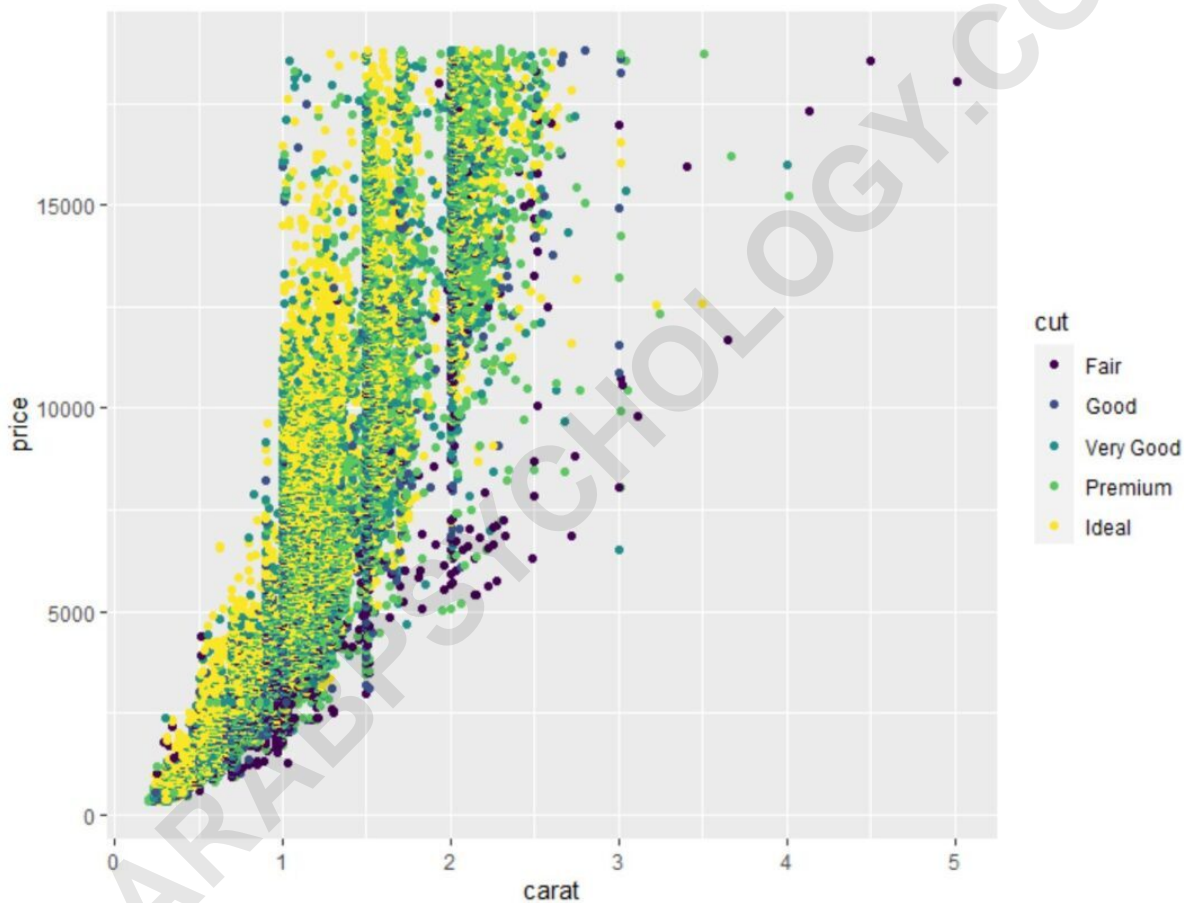
We can also use the `geom_point()` function to create a scatterplot of any pairwise combination of variables:

#create scatterplot of carat vs. price, using cut as color variable

```
ggplot(data=diamonds, aes(x=carat, y=price, color=cut))
```

+

```
geom_point()
```

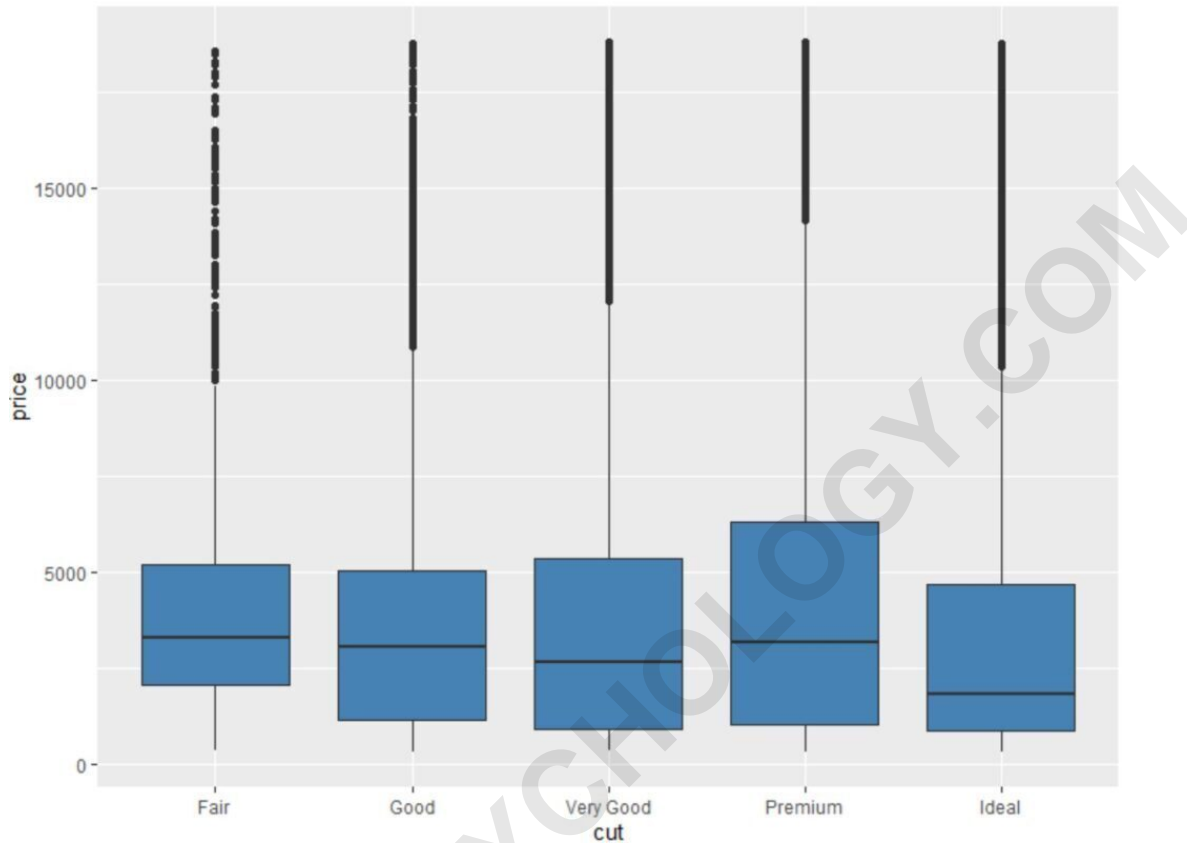


We can also use the geom_boxplot() function to create a boxplot of one variable grouped by another variable:

#create scatterplot of price, grouped by cut

```
ggplot(data=diamonds, aes(x=cut, y=price)) +
```

`geom_boxplot(fill="steelblue")`



By using these functions from `ggplot2`, we can learn a great deal about the variables in the diamonds dataset.

Additional Resources

The following tutorials explain how to explore other datasets in R: