MATHS 7107 Data Taming Tutorial

Chang Dong

2023-02-17

Questions

You work as a data scientist at the multi-million dollar Australian jewellery company sparkles and glitter. Your boss has asked you to do some research on diamonds to better understand which diamonds have a higher price so eventually the Company can increase profits (and hopefully pay you more money!!)

Your boss has specifically told you that your work must be in a report form so it can be forwarded to the sister company shine and shimmer located in the United States of America. Your boss wants them to be able to run your analysis on the data they have collected on diamonds they have sold.

The price of the diamonds has already been converted to US dollars.

Specifically you will need to complete the following for your boss:

1. Create a file - some sort of reproducible report - that can incorporate your explanations, code and output (analysis and plots etc).
2. Load the diamonds dataset. This is saved in the tidyverse package.

```
pacman::p_load(tidyverse,inspectdf)
data("diamonds")
```

3. Check the data to see if there are any entries missing (i.e. are there any NA's?).

```
inspect_na(diamonds)
```

```
## # A tibble: 10 x 3
##    col_name   cnt  pcnt
##    <chr>    <int> <dbl>
##  1 carat        0     0
##  2 cut          0     0
##  3 color        0     0
##  4 clarity      0     0
##  5 depth        0     0
##  6 table        0     0
##  7 price        0     0
##  8 x            0     0
##  9 y            0     0
## 10 z            0     0
```

4. Determine how many types of cut there are. What are they? Show how many diamonds there are of each particular cut.

```
unique(diamonds$cut)
```

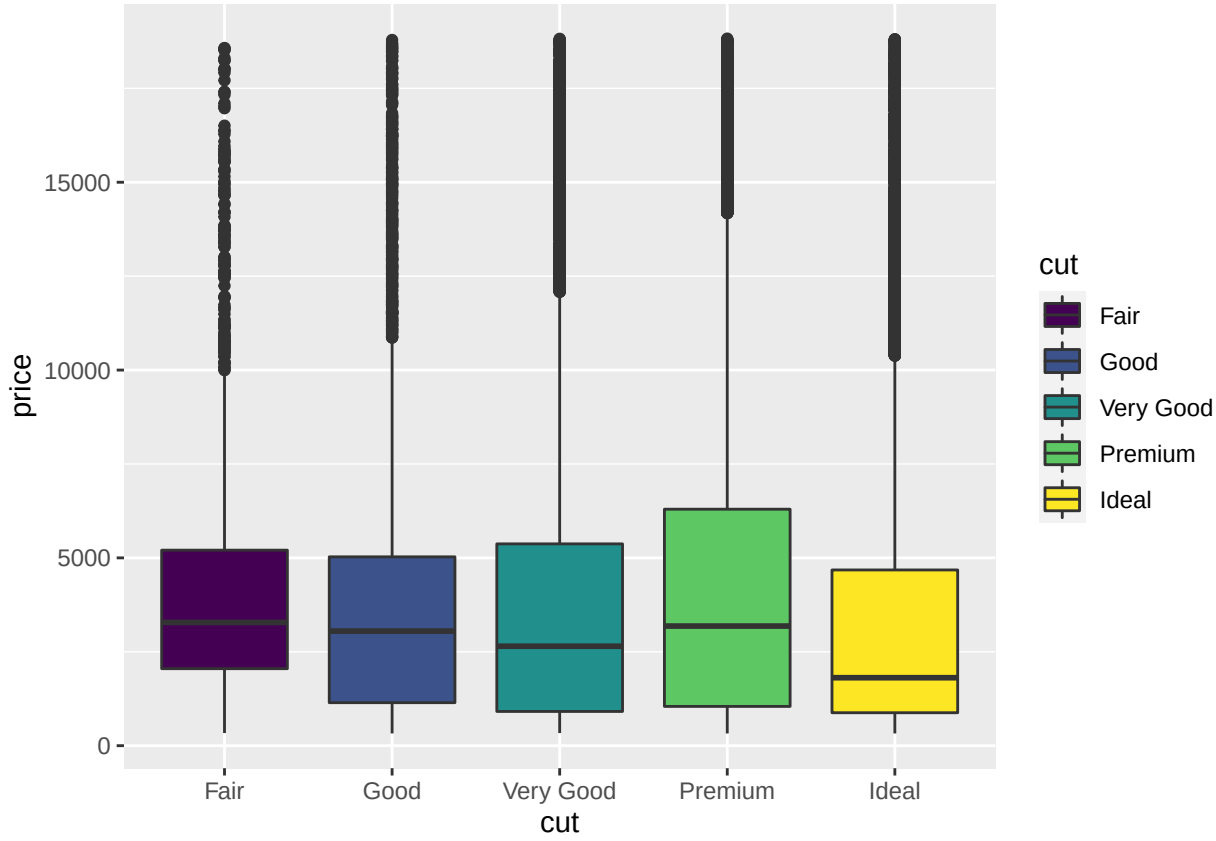
```
## [1] Ideal     Premium   Good      Very Good Fair
## Levels: Fair < Good < Very Good < Premium < Ideal
```

```
diamonds %>% count(cut)
```

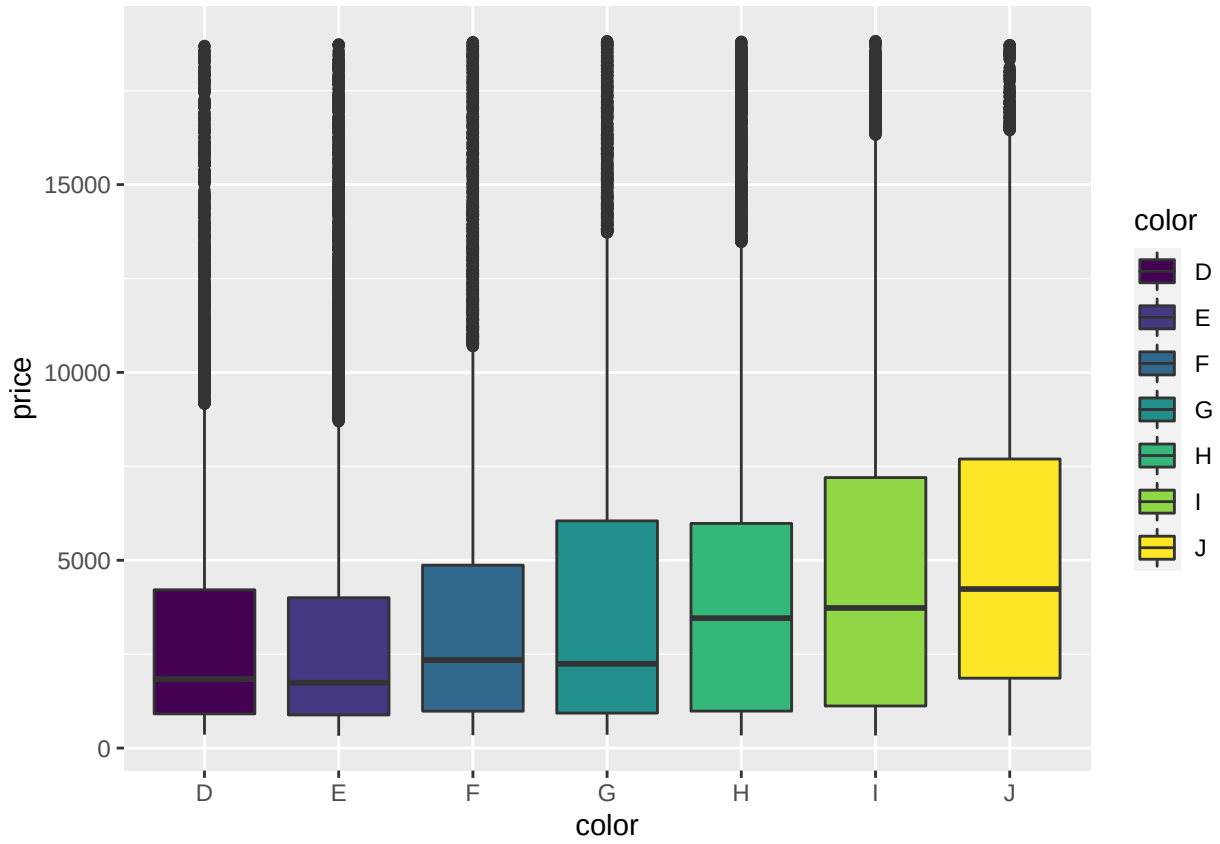
```
## # A tibble: 5 x 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

5. Your boss wants to know whether the price of the diamonds depends more on cut or color Using ggplot, produce two side-by-side boxplots of price, one using cut and one using color. Which variable appears to affect price more, cut or color?

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




```
ggplot(diamonds, aes(x=color, y = price, fill = color)) + geom_boxplot()
```



6. If a customer wants to buy a Premium diamond, with color rating J, how much should they expect to pay on average?

```
diamonds %>%
  filter(cut == "Premium", color == "J") %>%
  pull(price) %>%
  na.omit() %>%
  mean(na.rm = TRUE)
```

```
## [1] 6294.592
```

7. Write a short summary outlining exactly what you did so your boss is prepared when his colleague from America zooms next week. This will mean your research is reproducible to the sister company and your boss won't get cranky when he doesn't know an answer!

This PDF file was knited by Rmarkdown. In this report, I loaded "diamonds" dataset and checked there is no missing value in any column. And this report shows the information of cut feature in this diamonds data set, by visualizing with box plot, we can see more informative details of this data set. Besides, we can see the average price of an specific type of diamond. If you want to change another feature to have a see, just modify the origin Rmarkdown file, and this is a reproducible file, so it is easy for you to make any changes without modify the file format.