MATHS 7107 Data Taming Tutorial Solutions

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).

```
# Solution
# Create a R Markdown file
```

2. Load the `diamonds` dataset. This is saved in the `tidyverse` package.

```
# Solution
pacman::p_load(tidyverse, inspectdf)
diamonds
```

```
## # A tibble: 53,940 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  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  0.29 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
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
## # ... with 53,930 more rows
```

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

```
# Solution
diamonds %>% inspect_na()
```

```
## # 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`.

```
# Solution
unique(diamonds$cut)
```

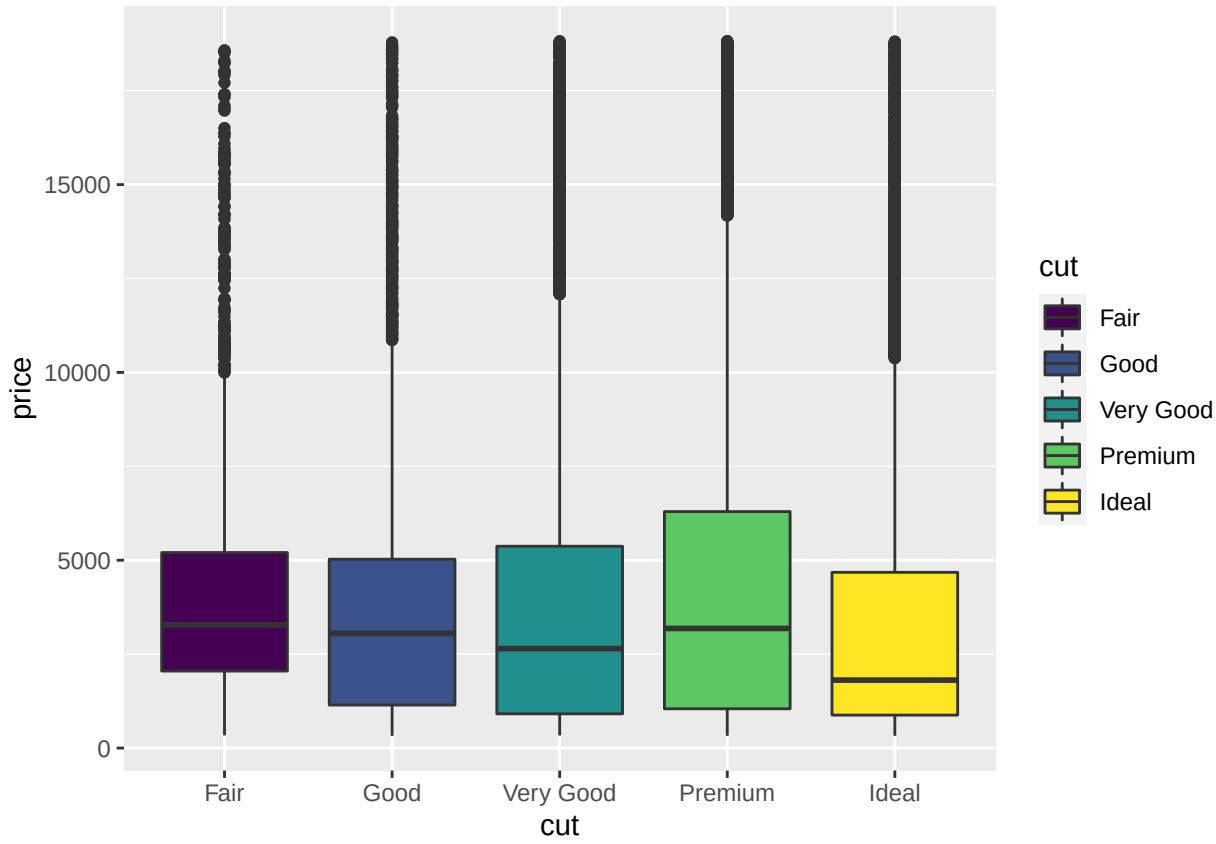
```
## [1] Ideal     Premium   Good      Very Good Fair
## Levels: Fair < Good < Very Good < Premium < Ideal
```

```
count(diamonds, cut)
```

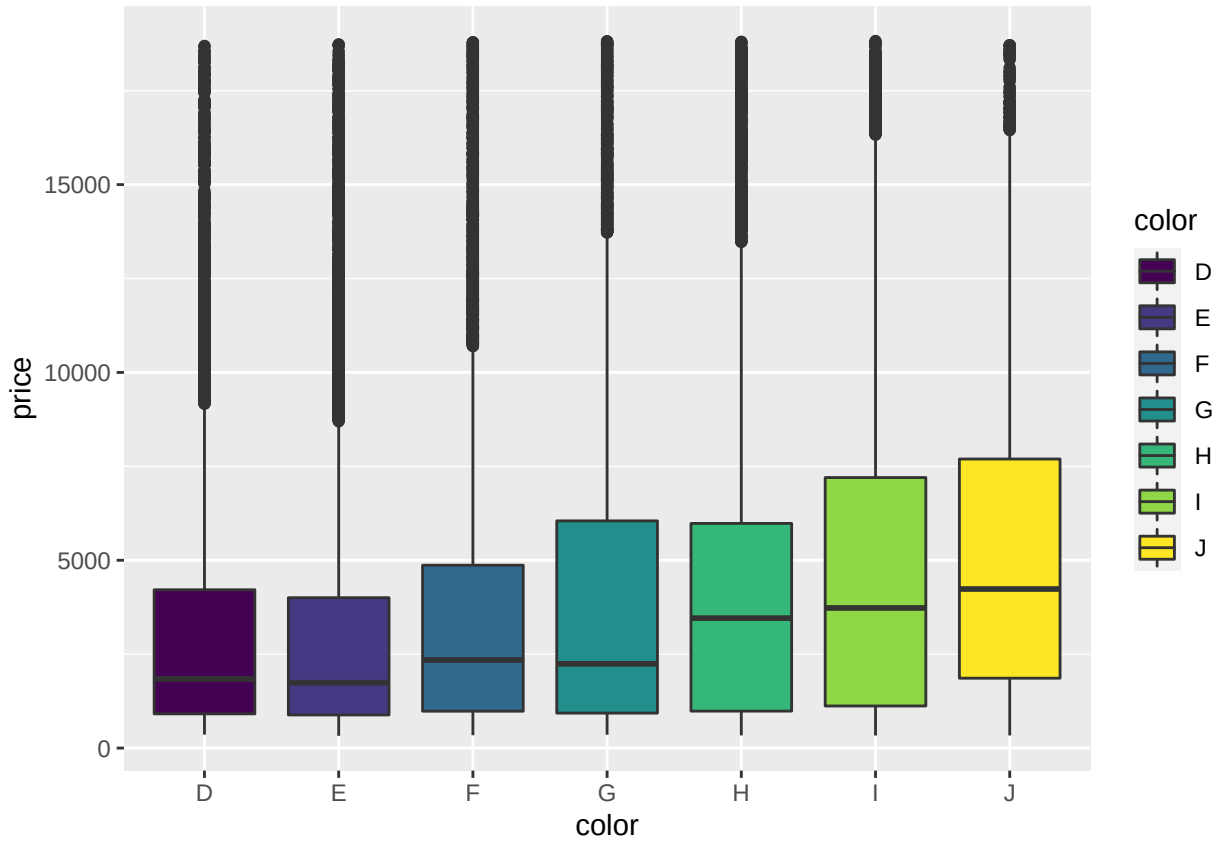
```
## # A tibble: 5 x 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

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`?

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




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




```
# Colour appears to affect price more
```

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

```
 # Solution
diamonds %>%
  filter(cut == "Premium", color == "J") %>%
  summary()
```

```
##      carat               cut        color       clarity        depth      
##  Min.   :0.300   Fair      :  0   D:  0   SI1     :209   Min.   :58.00  
##  1st Qu.:0.810   Good      :  0   E:  0   VS2     :202   1st Qu.:60.70  
##  Median :1.250   Very Good:  0   F:  0   SI2     :161   Median :61.60  
##  Mean   :1.293   Premium   :808   G:  0   VS1     :153   Mean   :61.39  
##  3rd Qu.:1.700   Ideal     :  0   H:  0   VVS2    : 34   3rd Qu.:62.30  
##  Max.   :4.010                   I:  0   VVS1    : 24   Max.   :63.00  
##                                  J:808   (Other): 25                  
##      table           price             x                y         
##  Min.   :54.00   Min.   :  363   Min.   : 4.22   Min.   :4.210  
##  1st Qu.:58.00   1st Qu.: 2203   1st Qu.: 6.04   1st Qu.:5.987  
##  Median :59.00   Median : 5063   Median : 6.92   Median :6.900  
##  Mean   :58.87   Mean   : 6295   Mean   : 6.81   Mean   :6.771  
##  3rd Qu.:60.00   3rd Qu.: 9050   3rd Qu.: 7.62   3rd Qu.:7.580  
##  Max.   :62.00   Max.   :18710   Max.   :10.02   Max.   :9.940  
##                                                                 
##        z         
##  Min.   :2.590  
```

```
##  1st Qu.:3.660
##  Median :4.260
##  Mean   :4.168
##  3rd Qu.:4.673
##  Max.   :6.240
##
```

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!