Diamond prices

```
library(tidyverse)
library(pander) # for prettier tables
library(scales) # for making prettier axes in plots

theme_set(theme_bw())
```

One exercise is marked with a *. It may be difficult (but also relevant).

Here, we focus on the `diamonds` dataset:

```
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 0.290 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
##  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
```

Read `?diamonds`.

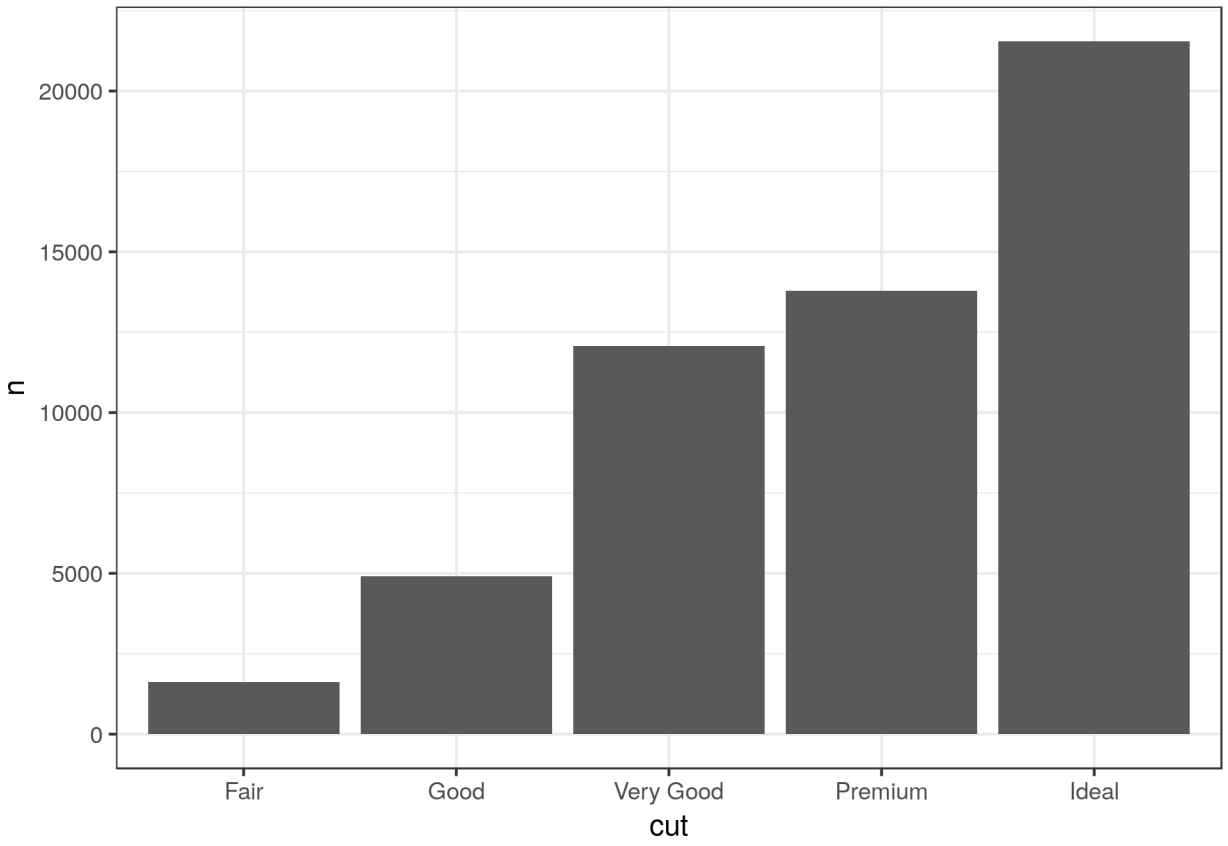
1 Exercises in data wrangling

1.1 Exercise

Construct a summary table of the number of diamonds of each `cut` like this: (Hint: the `pander()` function takes a `big.mark = ','` argument and also a `justify = c('left', 'right')` argument.)

cut	n
Fair	1,610
Good	4,906
Very Good	12,082
Premium	13,791
Ideal	21,551

Construct the corresponding barplot:



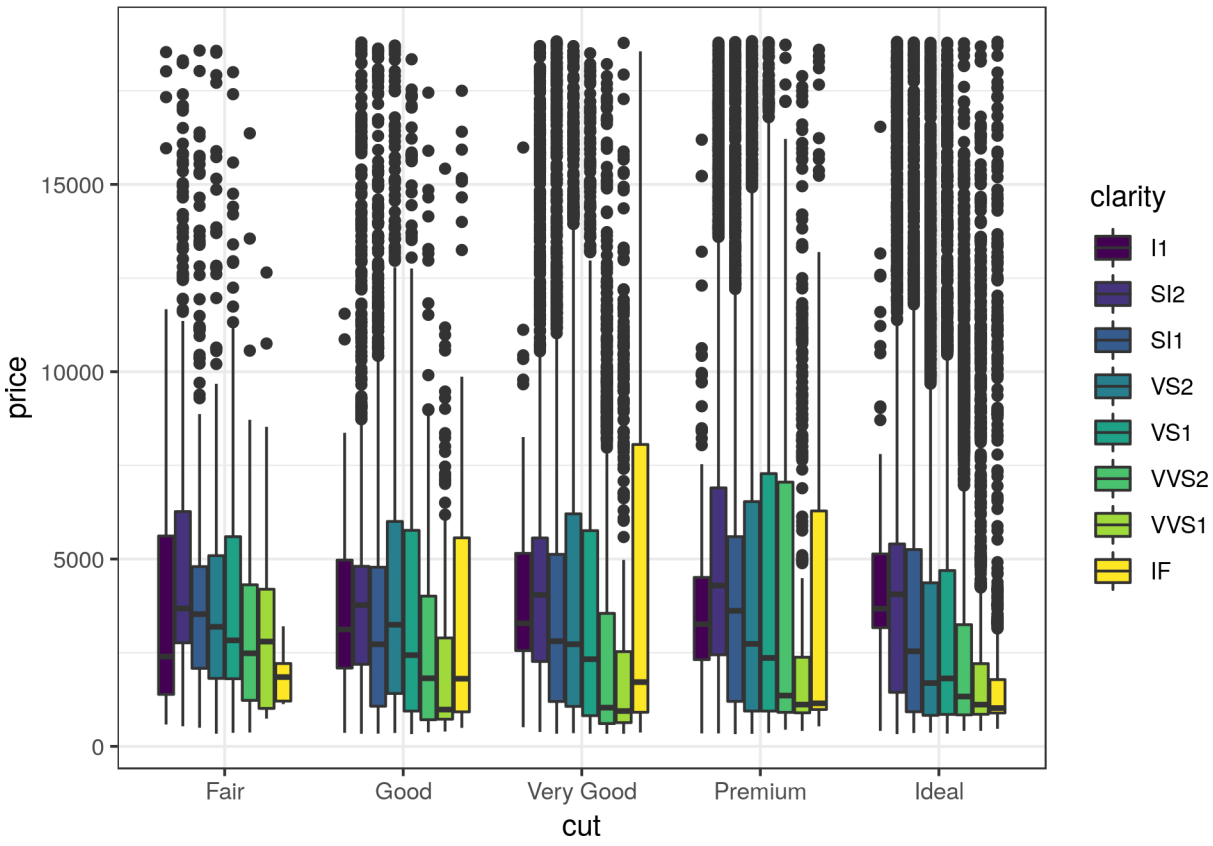
1.2 Exercise

Construct a summary table of the number of diamonds of each `clarity`, but only for diamonds with `cut = "Ideal"` ordered descendingly by `n` like this:

clarity	n
VS2	5,071
SI1	4,282
VS1	3,589
VVS2	2,606
SI2	2,598
VVS1	2,047
IF	1,212
I1	146

1.3 Exercise

Construct a boxplot of the prices of the diamonds for each `clarity` and `cut`, e.g. like this:



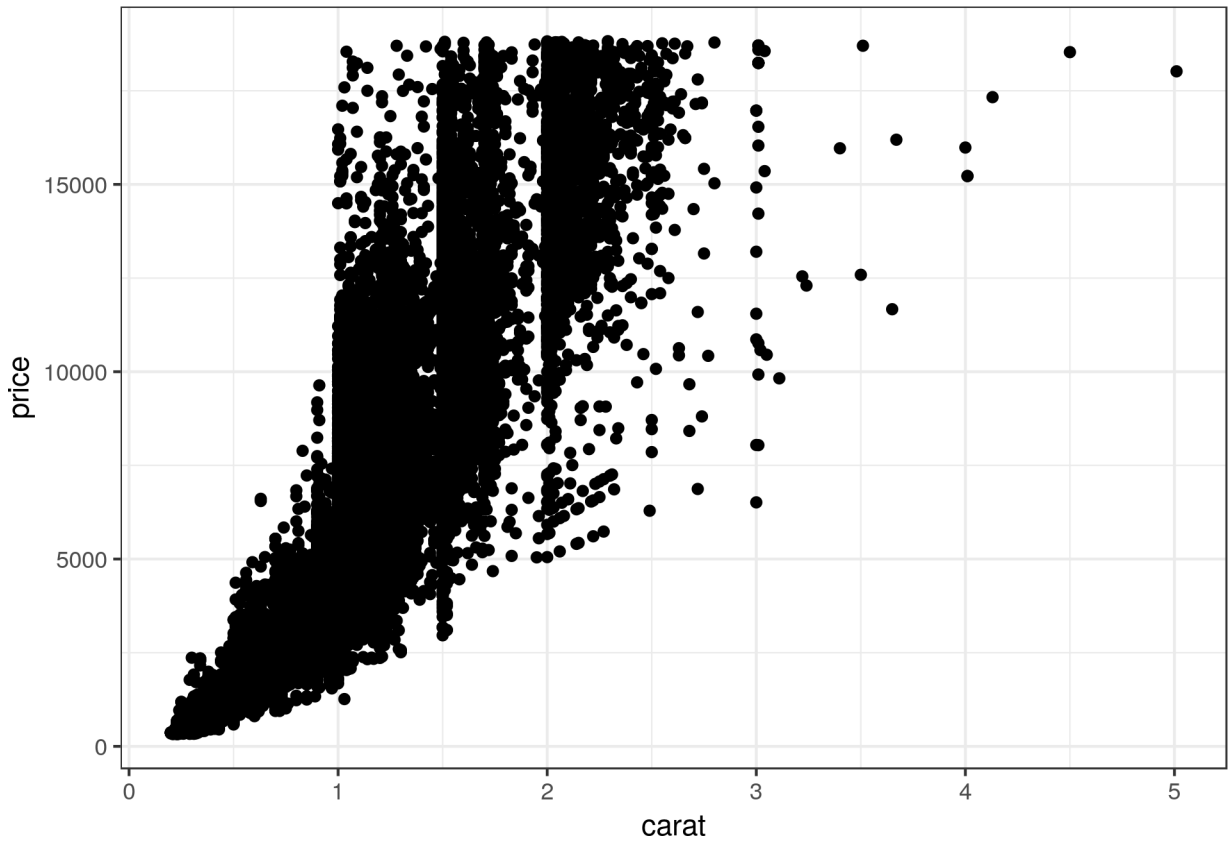
1.4 Exercise

Assuming that the diamonds are rectangular cuboids ("boxes"), then what is the average volume for each `cut`? Order the table descendingly according to mean volume like:

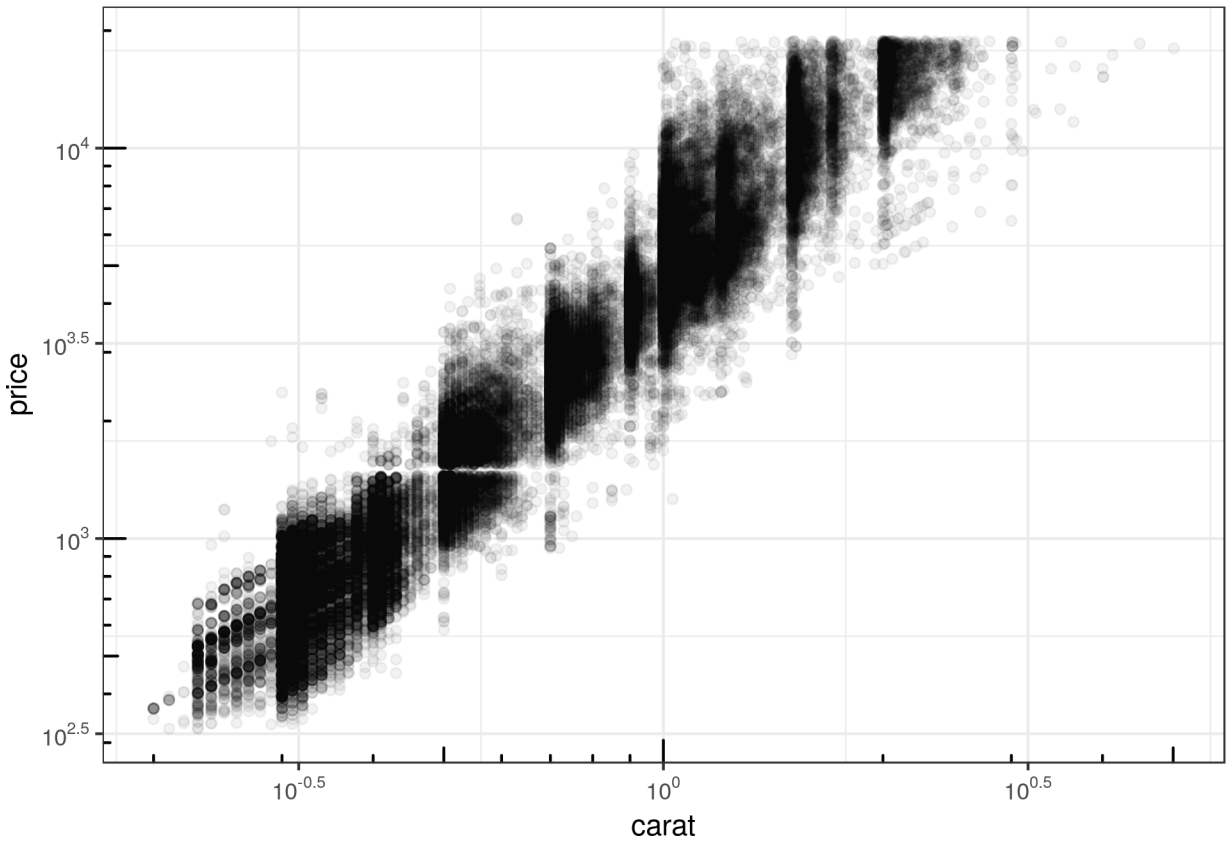
cut	mean_volume
Fair	165
Premium	145.1
Good	136.3
Very Good	131
Ideal	115.4

1.5 Exercise

Do a scatter plot of carat vs price:



Do a log-log plot of carat vs price (scatter plot with both axes on a log scale: http://ggplot2.tidyverse.org/reference/annotation_logticks.html) whre the points are transparent (e.g. with `alpha` of `0.05`):



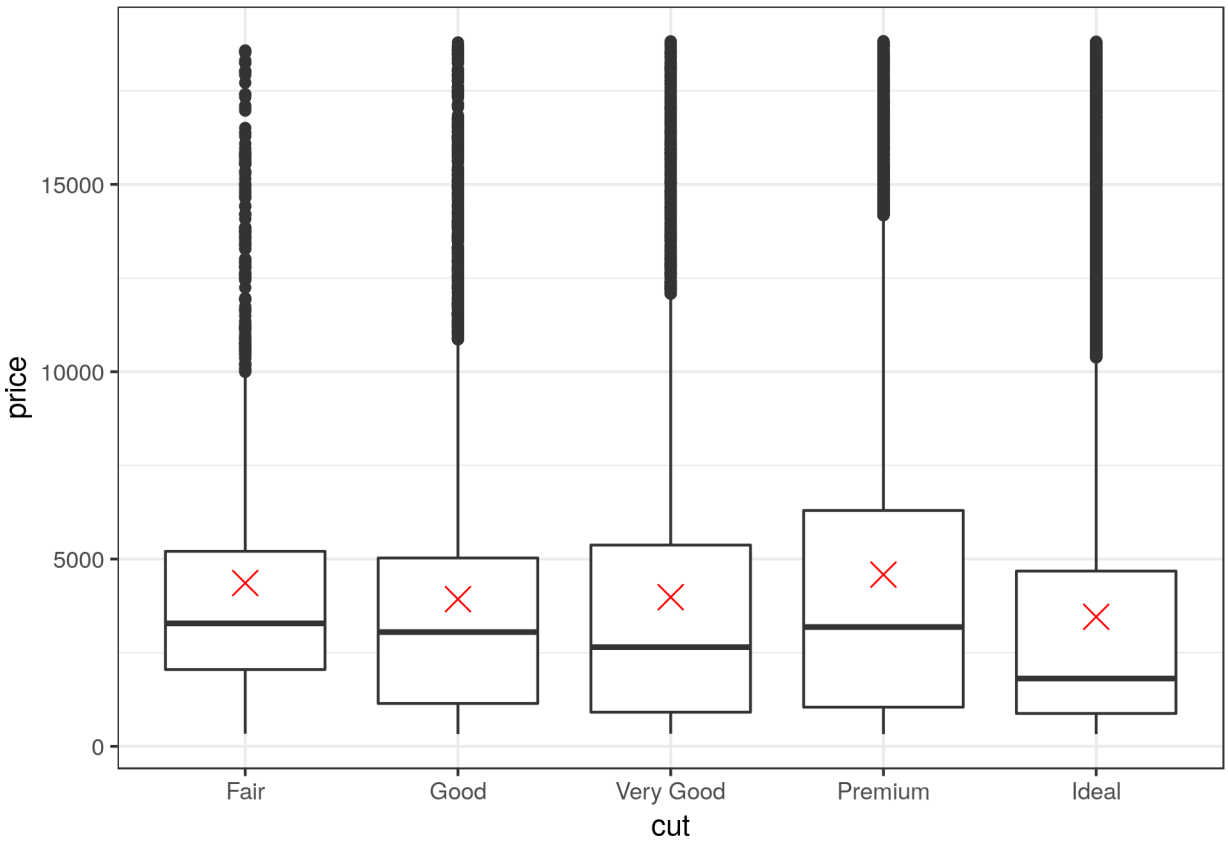
1.6 Exercise

What is the mean `price` per `cut`? You should get something like:

cut	mean_price
Fair	4,359
Good	3,929
Very Good	3,982
Premium	4,584
Ideal	3,458

1.7 Exercise

Construct a boxplot of the prices of the diamonds for each `cut` and include the average prices as points (hint: `geom_point` can take a `data` argument, a new mapping, a `col` and a `pch`), e.g. like this:



1.8 Exercise

For each cut, how many diamonds are at least 5 mm. in length, width or depth?

cut	xyz_gte_5	n
Fair	FALSE	144
Fair	TRUE	1,466
Good	FALSE	1,169
Good	TRUE	3,737
Very Good	FALSE	3,510
Very Good	TRUE	8,572
Premium	FALSE	3,974
Premium	TRUE	9,817
Ideal	FALSE	8,662
Ideal	TRUE	12,889

cut	sum(xyz_gte_5)
Fair	1,466
Good	3,737
Very Good	8,572
Premium	9,817
Ideal	12,889

What is the mean `price` per `cut` for diamonds that is at least 5 mm. in length, width or depth? You should get something like:

cut	mean_price
Fair	4,690
Good	4,946
Very Good	5,329
Premium	6,108
Ideal	5,226