R による探索的データ分析入門 - 可視化　演習

発電基盤開発課　高津一誠

2018 年 *10* 月 *12* 日

1　演習

前回勉強した R による可視化を使って、データを分析してみましょう。
R に組み込まれている、diamonds データセットを使ってデータ分析をしてください。diamonds はダイヤモンドの品質と価格を格納したデータで、前回の iris と同じように tidyverse パッケージをロードすると使えるようになります。

```
library(tidyverse)


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
```

変数の説明は以下のとおりです。

表1: diamond データセット

データ列 (変数)	説明
carat	重さ (カラット)
cut	カット等級
color	色

データ列 (変数)	説明
clarity	透明度
x	長さ (mm)
y	幅 (mm)
z	深さ (mm)
depth	深さ比 (z/(x+y)/2)
table	上面幅/最大幅

1.1 演習 1

ダイヤモンドのカラット数と価格には、どのような関係があるでしょうか？グラフを描いて調べてください。

1.2 演習 2

カラット数と価格の関係は、カット等級によって変化するでしょうか？カット等級ごとのサブグラフを描いて調べてください。

2　演習の回答

2.1　演習 1

2 つの変数の相関を視覚的に確認するには散布図を使うのが一般的です。早速描いてみましょう。

```
diamonds %>% ggplot(aes(x = price, y = carat)) +
  geom_point()
```

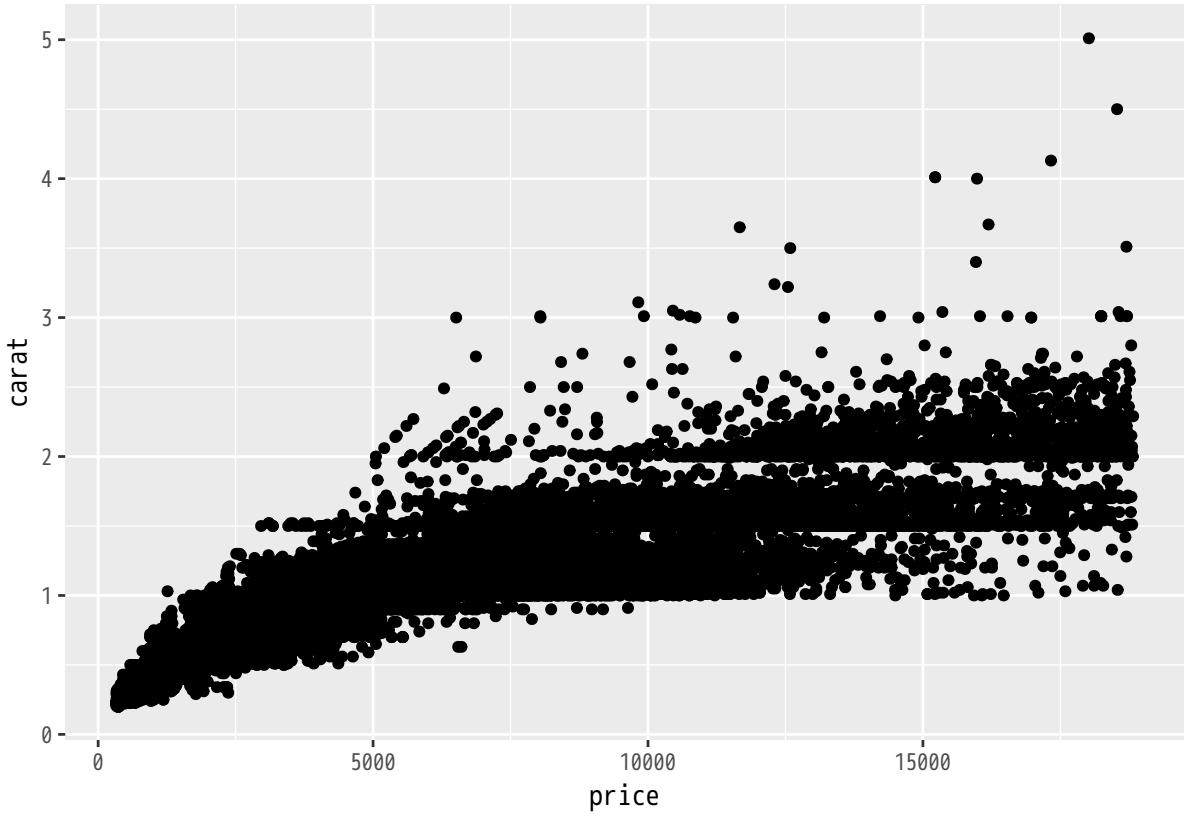

図1　演習 1：散布図

カラット数が大きくなると価格も大きくなる傾向があり、相関はありそうです。ただデータ数が 50,000 以上あるので点が重なってしまい、どこに多くのデータが分布しているのかが分からなくなっています。そこで少し工夫してみましょう。

- 点を半透明にする
  グラフ属性 `alpha` に値を指定して透明度を変化させることができます。ここではデータを割り当てるのではなく固定値にするので、`aes()` の中でなく直接指定します。

```
diamonds %>% ggplot(aes(x = price, y = carat)) + geom_point(alpha = 0.1)
```

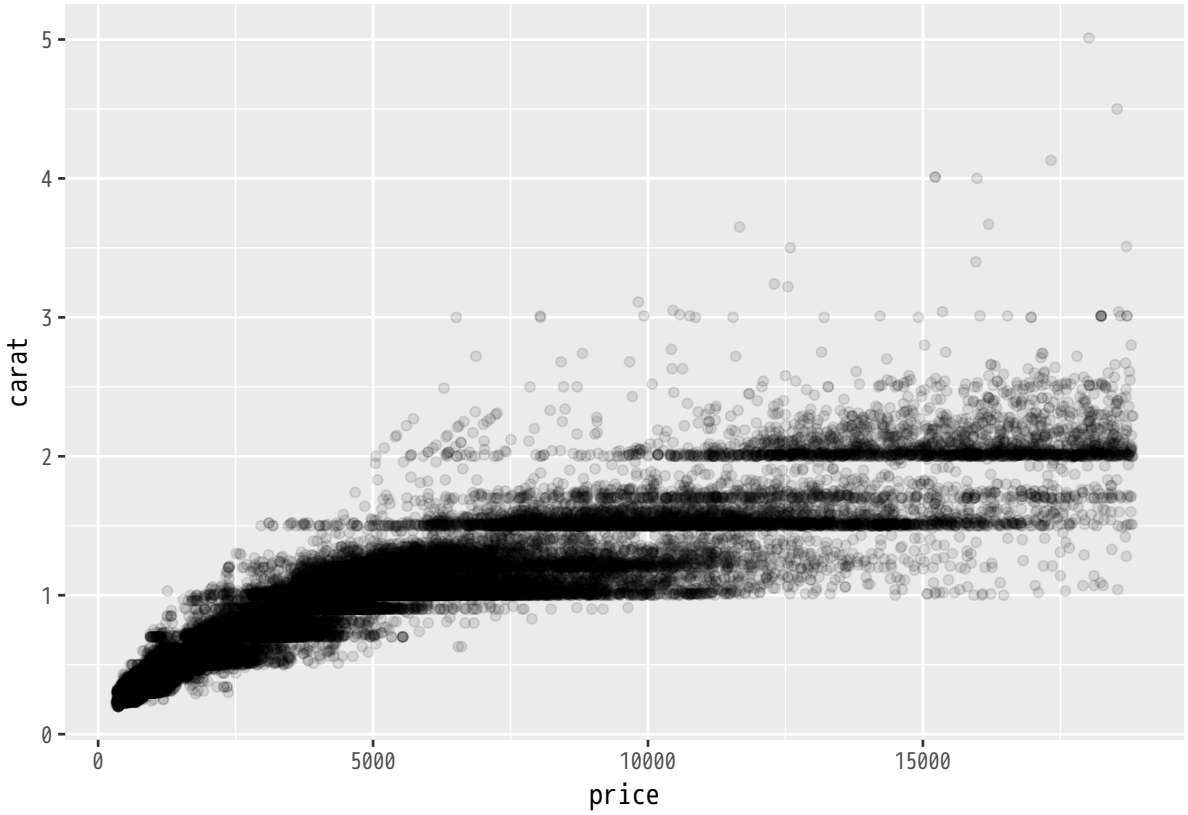

図2　演習 1：オーバープロット対策 1

- 2 次元ヒストグラム
  おおまかな分布を確認したいなら、先ほど勉強した 2 次元ヒストグラムを使うと分かりやすくなります。

```
diamonds %>% ggplot(aes(x = price, y = carat)) + geom_bin2d()
```

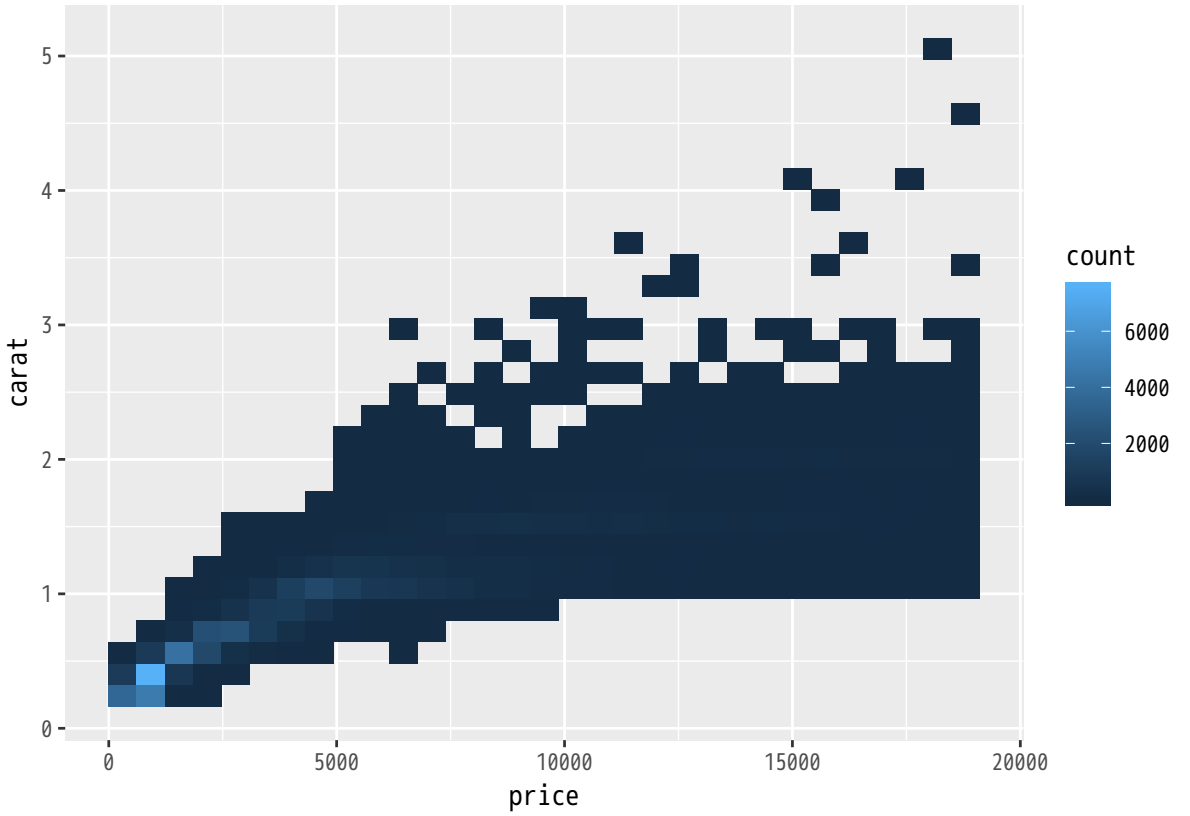

図3　演習 1：オーバープロット対策 2

以降では、2 次元ヒストグラムを使ってプロットすることにします。

では、カラット数と価格の関係をより詳しく見てみましょう。

このデータを直線近似でモデル化しても、関係を明確にできません。その理由は、カラット数と価格が比例関係にないことと、価格が高くなるにつれてカラット数のバラつきが大きくなることです。つまり多項式近似でのモデル化も不適切ということです。

```
diamonds %>% ggplot(aes(x = price, y = carat)) +
  geom_bin2d() +
  geom_smooth(method = "lm", color = "red")
```

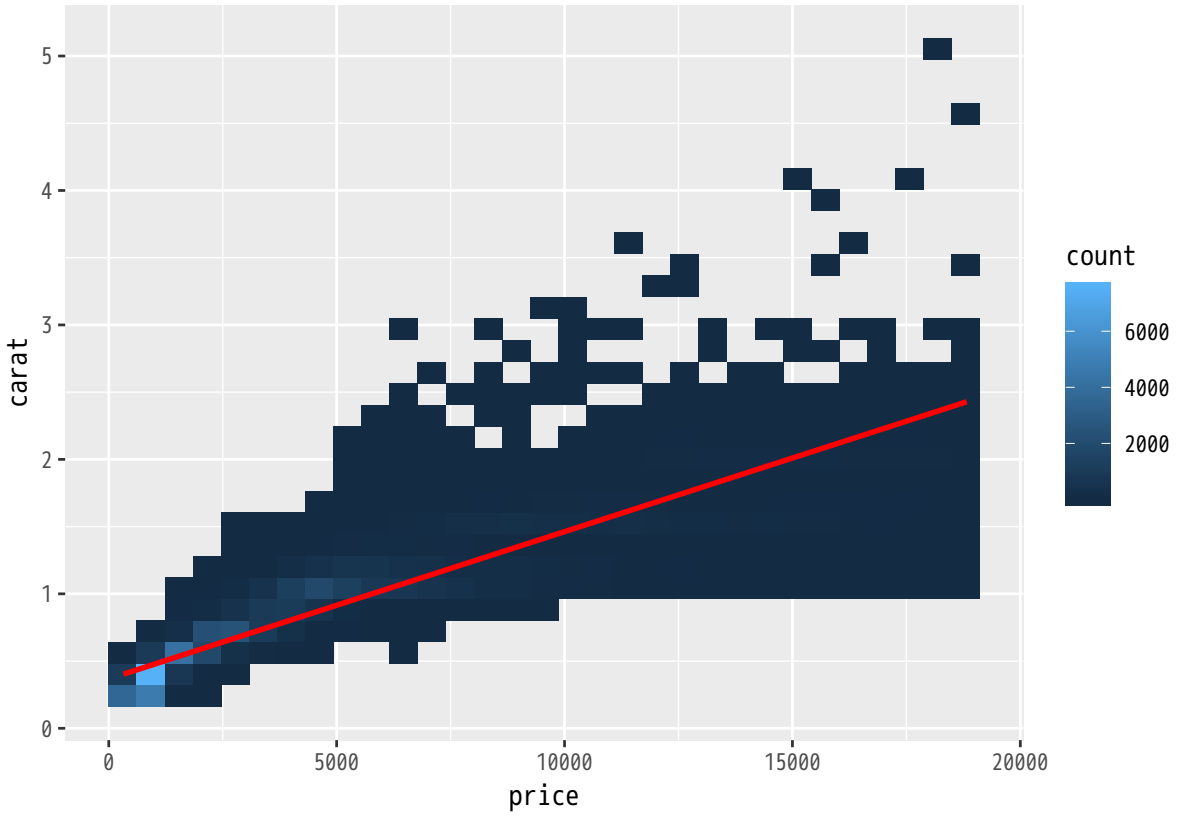

図4　演習 1：直線近似モデル

```
diamonds %>% ggplot(aes(x = price, y = carat)) +
  geom_bin2d() +
  geom_smooth(method = "lm", color = "red", formula = y ~ poly(x, 4, raw = T))
```

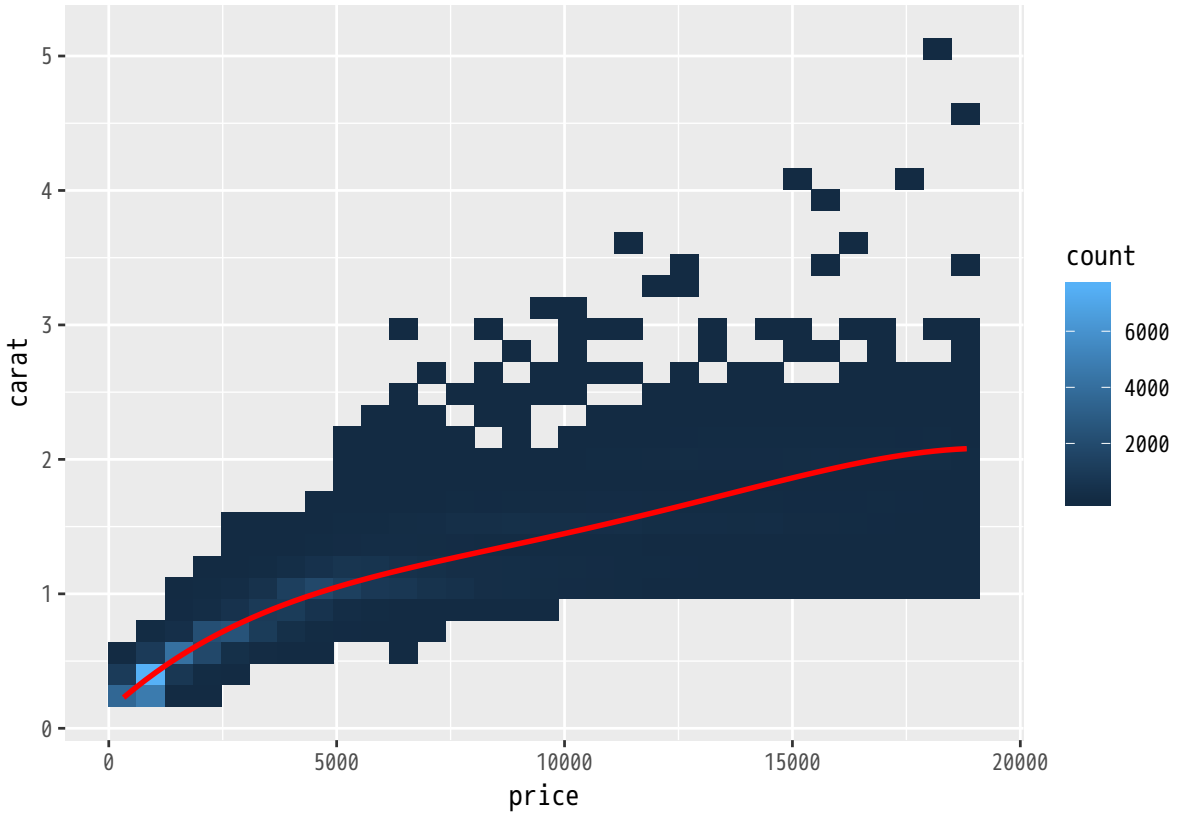

図5　演習 1：多項式近似モデル

では、どのようなモデルならよいのでしょうか？
べき乗関数近似モデルは、変数の値によってもう一方の変数バラつきが大きくなるので試してみましょう。このモデルは対数変換すると直線近似モデルと同じになるので両対数グラフを描いてみます。

R では X や Y のスケールの設定もレイヤーとして追加するので、以下のように記述します。

```
diamonds %>% ggplot(aes(x = price, y = carat)) +
  geom_bin2d() +
  geom_smooth(method = "lm", color = "red") +
  scale_x_log10() + scale_y_log10()
```

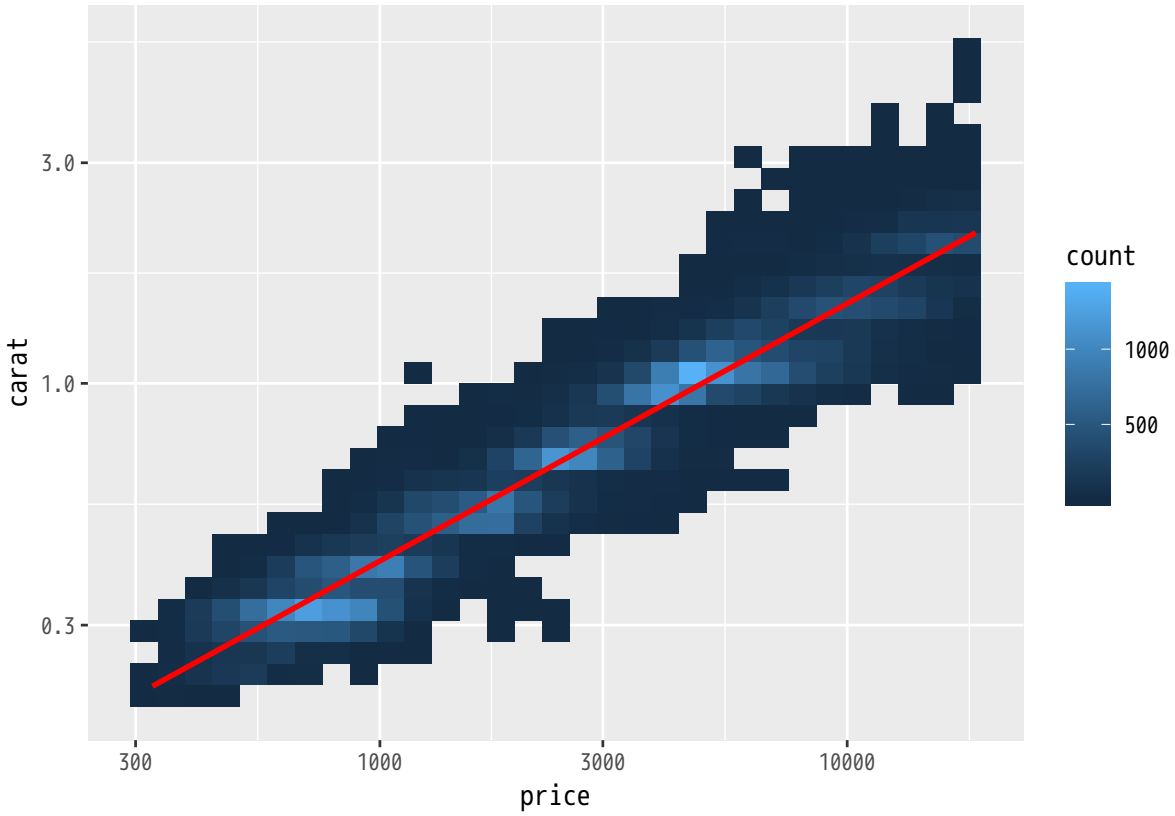

図6　演習 1：直線近似モデル

グラフを見るとうまくモデル化できているようです。
最後にモデル式を以下に示します。誤差項がどのようにモデル化されているかに着目してください。

表2: モデル式

近似モデル	モデル式
直線近似	$y_i = ax_i + b + \epsilon_i$
n 次多項式近似	$y_i = \sum_{j=0}^{n} a_j x_i^j + \epsilon_i$
べき乗関数近似	$y_i = ax_i^b e^{\epsilon_i}$
べき乗関数近似の対数変換モデル	$\log y_i = \log a + b \log x_i + \epsilon_i$

2.2　演習 2 の回答

演習 1 のグラフにサブグラフのレイヤーを追加するのですが、同じことを繰返し書くのは面倒です。特に後で修正が必要になったときに修正し忘れることもあります。そこで、以下のように共通の部分まで作ったグラフに名前を付けて保存することで使いまわすことができます。

```
p <- diamonds %>% ggplot(aes(x = price, y = carat)) +
  geom_bin2d() +
  geom_smooth(method = "lm", color = "red") +
  scale_x_log10() + scale_y_log10()
print(p)
```

```
p + facet_wrap(~ cut)
```

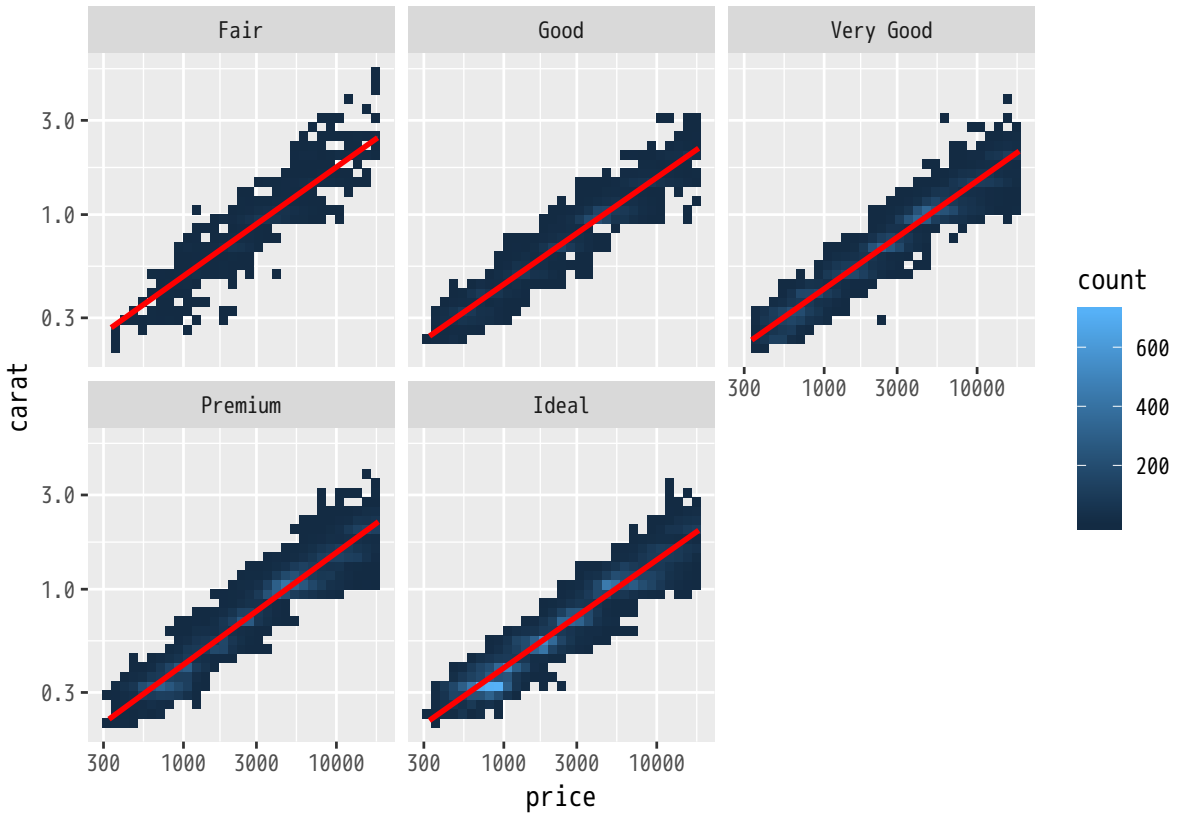

図7　演習 2：カット等級の影響

グラフを見るとカットの影響はほとんどなく、どのサブグラフでも同じような関係になっていることが分かります。