Exploratory Data

Kylie A. Bem

Northeastern Uni
Khoury College of Comput

Northeastern U

Learning go

- What is EDA?
- Variation and covariatio
- "Interesting" visualizatio

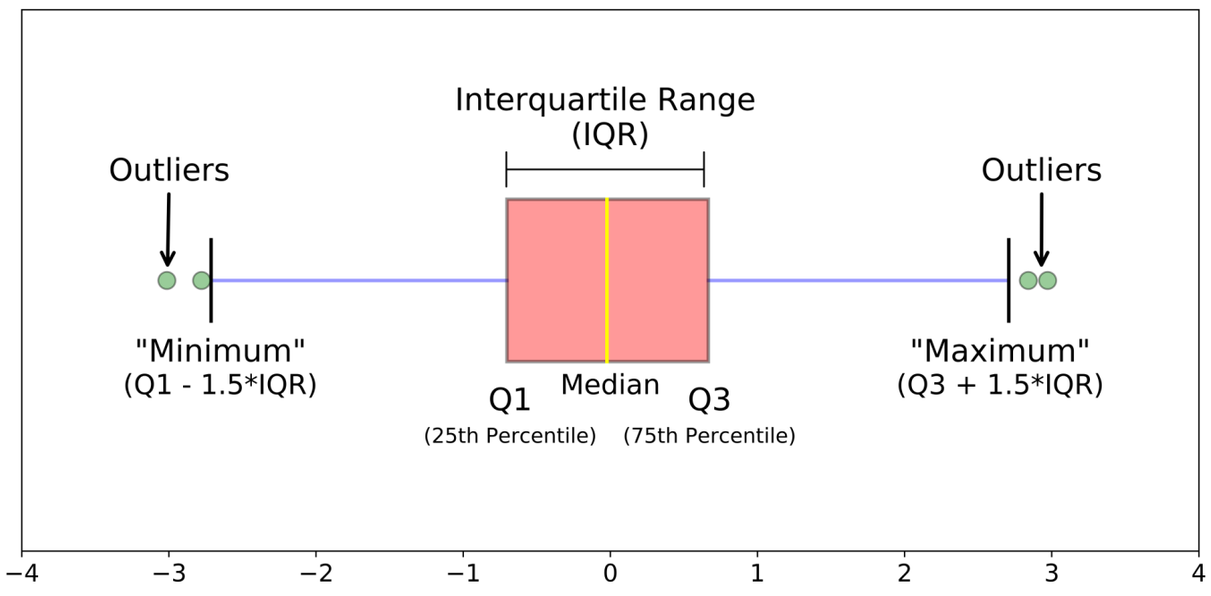
EDA:

EXPLORATORY DA

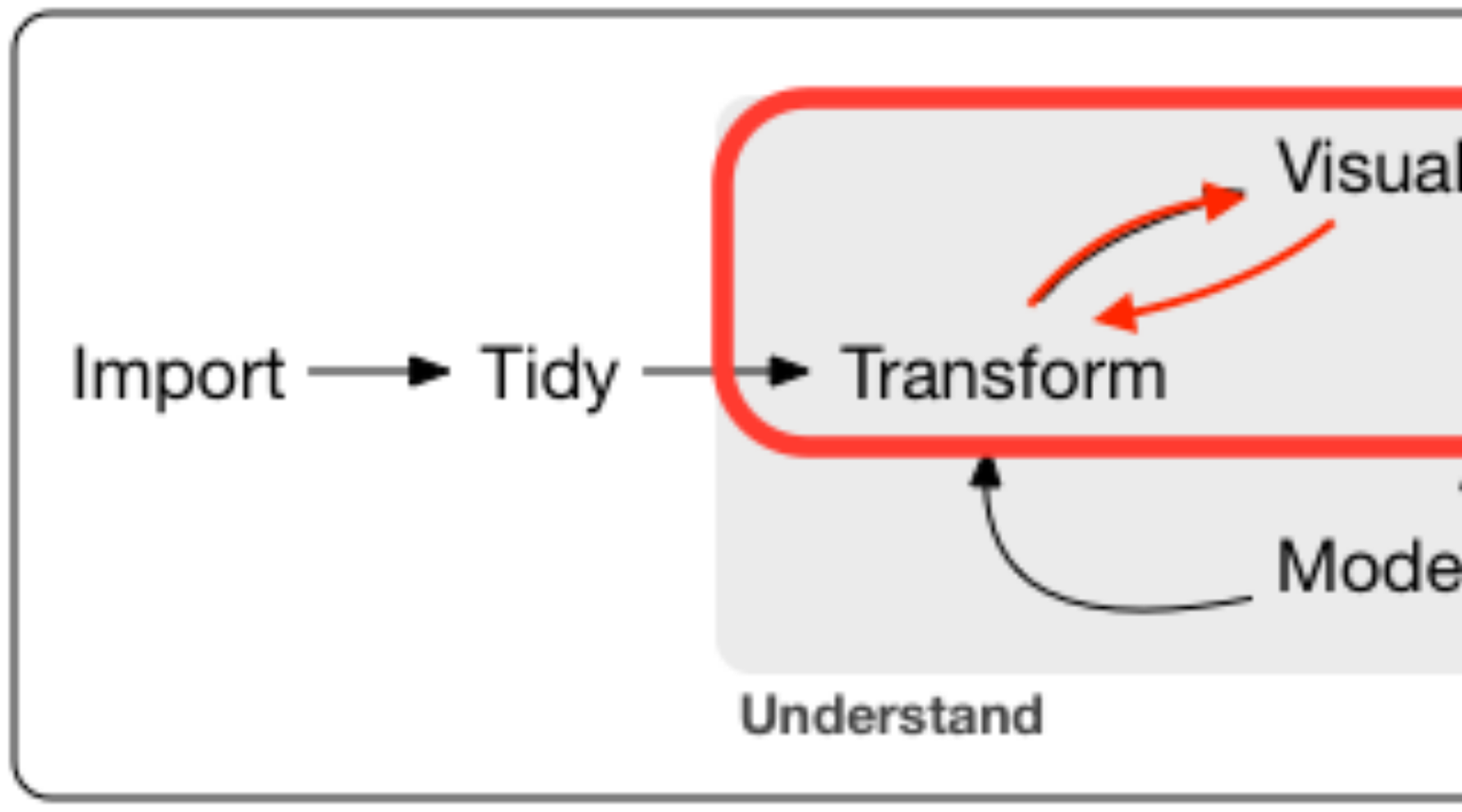
Exploratory data

EDA is an iterative, cyclical process of inform
and relationships between th

- Generate questions abo
- Search for answers by t
  visualizing your data
- Use this to refine your
  generate new questions



Process of E



R for Data Science, Wickham an

Example data: Di

Prices and measurements of

```
diamonds
```

```
## #   tibble: 53,940 x 10
##     carat cut       color clarity depth t
##     <dbl> <ord>     <ord> <ord>   <dbl> <
##  1 0.23  Ideal     E     SI2      61.5
##  2 0.21  Premium   E     SI1      59.8
##  3 0.23  Good      E     VS1      56.9
##  4 0.290 Premium   I     VS2      62.4
##  5 0.31  Good      J     SI2      63.3
##  6 0.24  Very Good J     VVS2     62.8
##  7 0.24  Very Good I     VVS1     62.3
##  8 0.26  Very Good H     SI1      61.9
##  9 0.22  Fair      E     VS2      65.1
## 10 0.23  Very Good H     VS1      59.4
## # ... with 53,930 more rows
```

Available in `ggplot2` p

Question

- What type of variation occurs v
- What type of covariation occur

Useful definit

- A **variable** is a quantity, qu that is measured
- A **value** is the state of a v measure it
- An **observation** is a set o made under similar condit

Types of varia

- Categorical — qualitative
  - Nominal — unordered categories
  - Ordinal — ordered or ranked cat
- Numeric — quantitative a
  - Continuous — infinite possible va
  - Discrete — countable possible va
- Time — dates and times

Variation

Variation is the tendency of values of v
measurement to meas

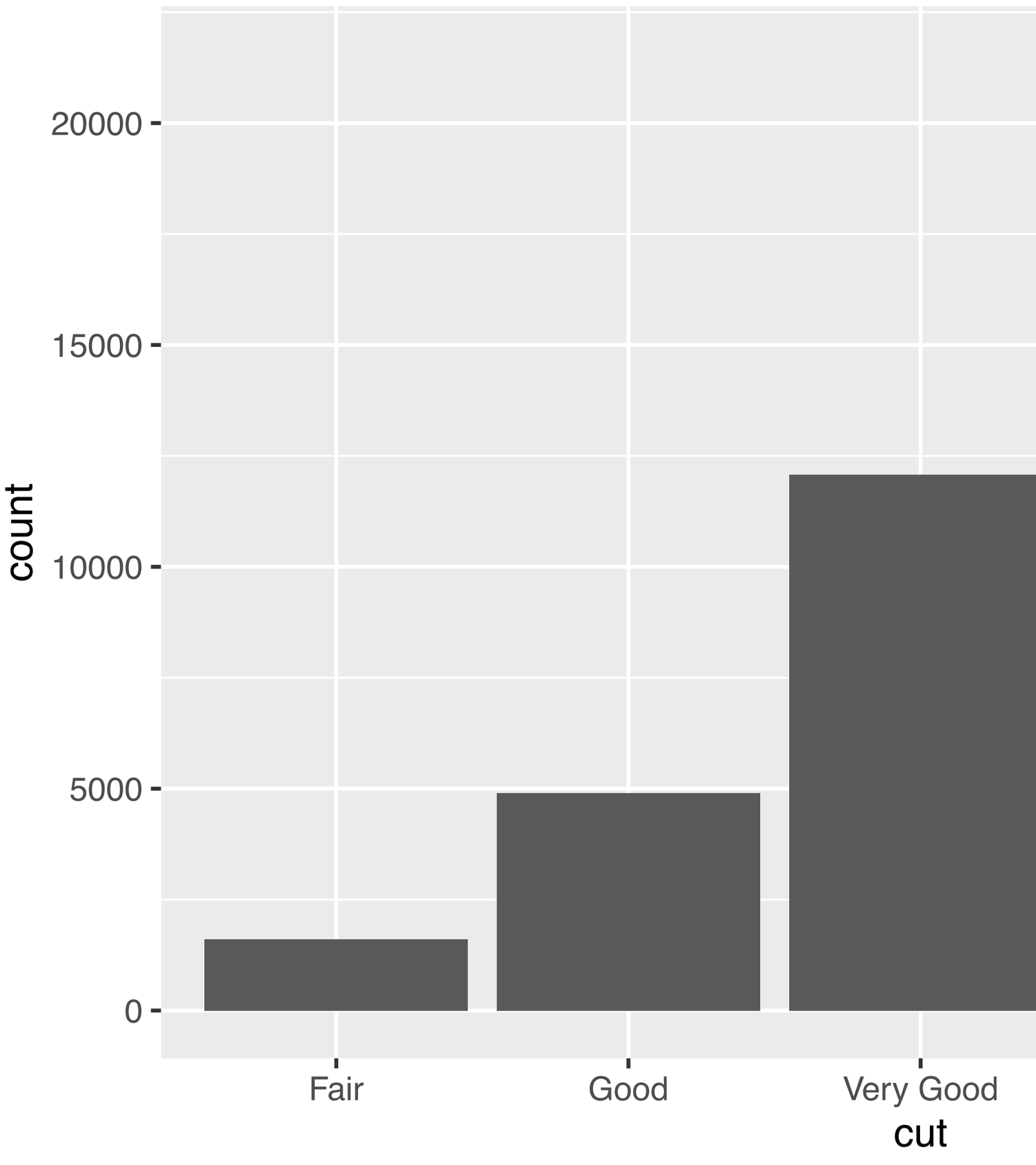
- Natural variation
- Measurement erro
- Time and location
- Between subjects
- Between condition

Categorical: *diam*

What is the distribution of cut quality

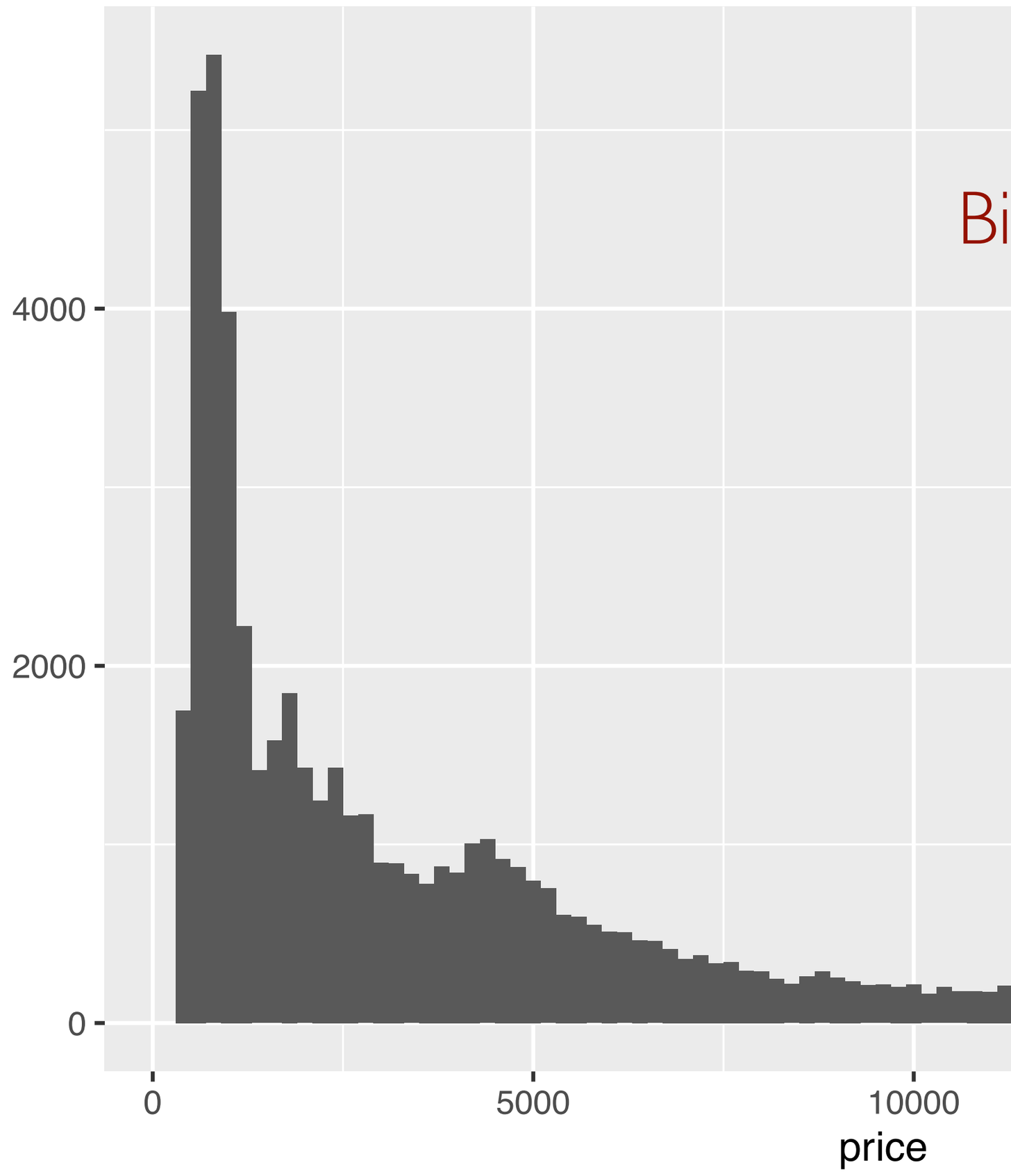
Categorical: *diam*

"Ideal" is the most common qua



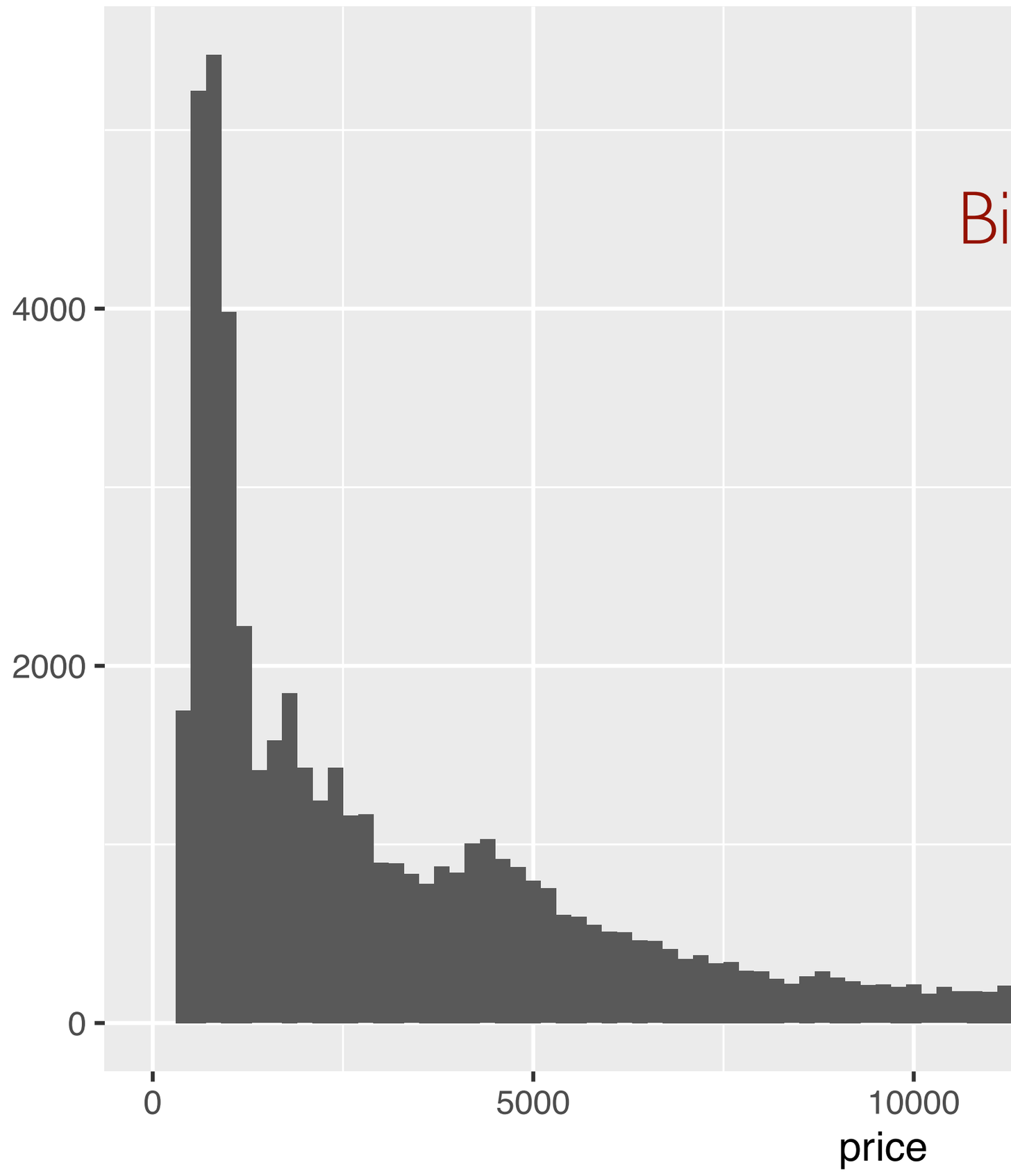
Continuous: *diam*

What is the distribution of price a



Continuous: *diam*

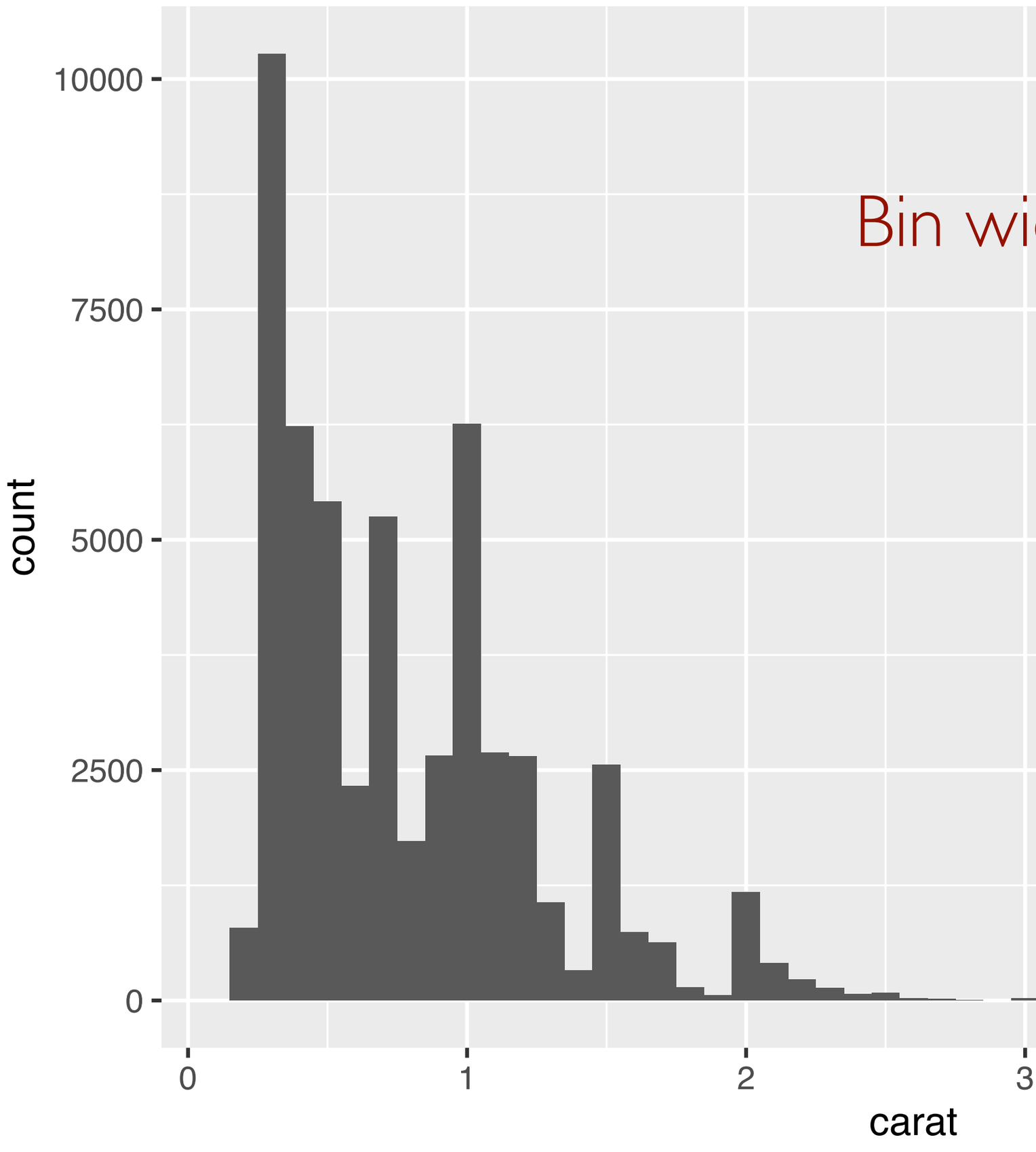
Price is heavily right-skewed with a long tai



Bi

Continuous: *diam*

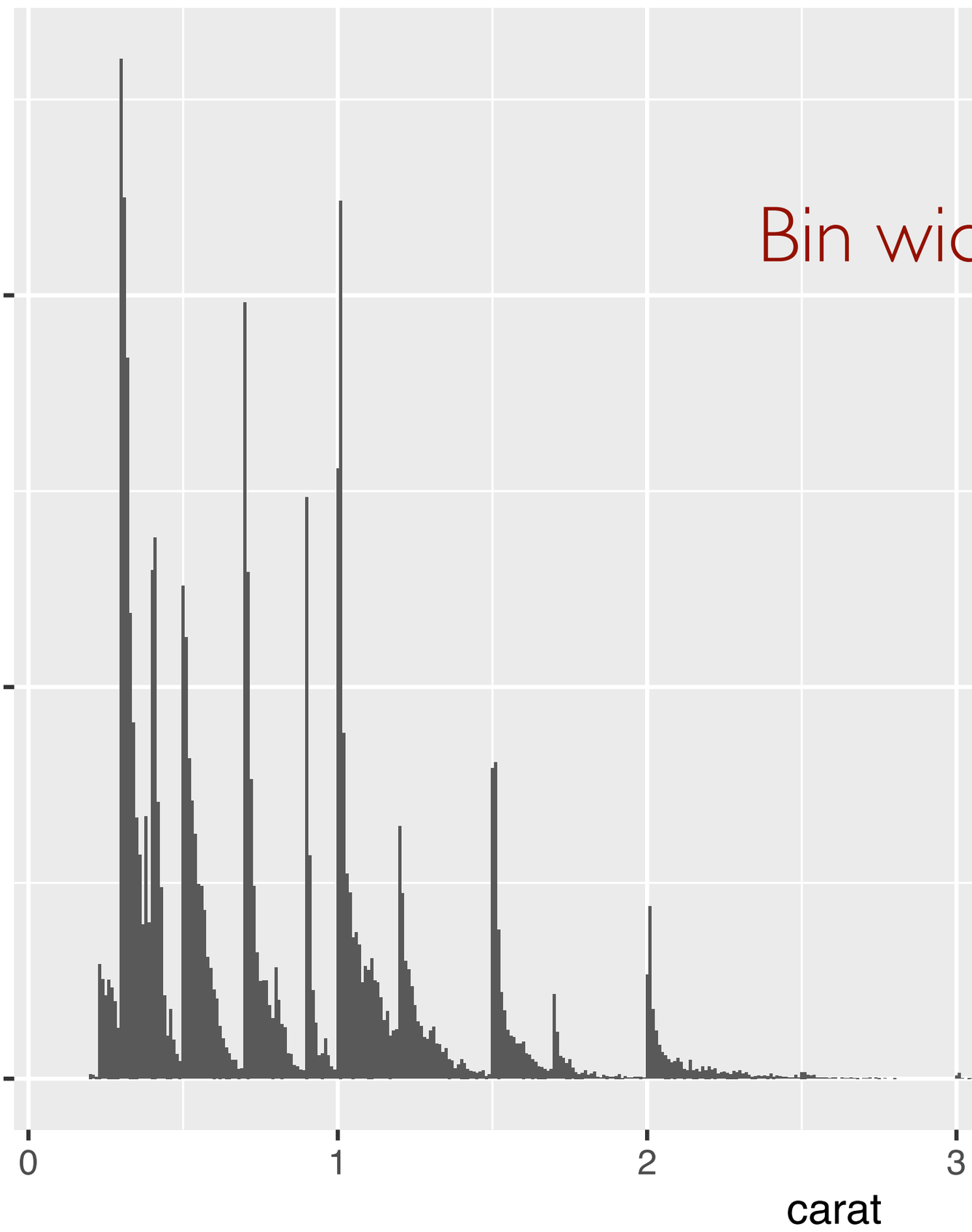
What is the distribution of carat a



Bin wi

Continuous: *diam*

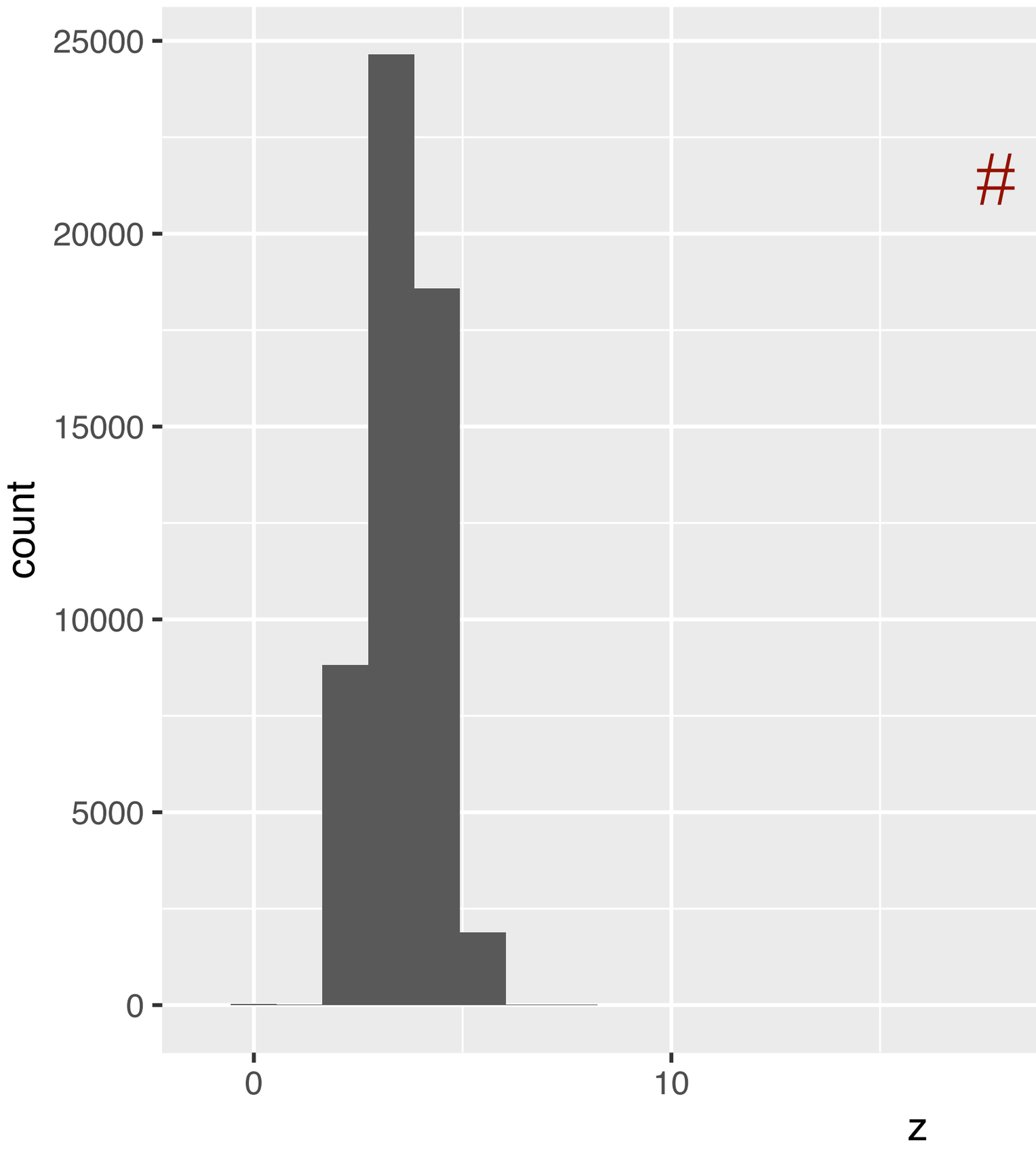
Diamonds are often cut in quar



Try different bin widths
or # of bins

Identify outl

Investigate z-dimension measure



Identify outl

Diamonds with *z* > *30* or *z* <=

#

Covariatio

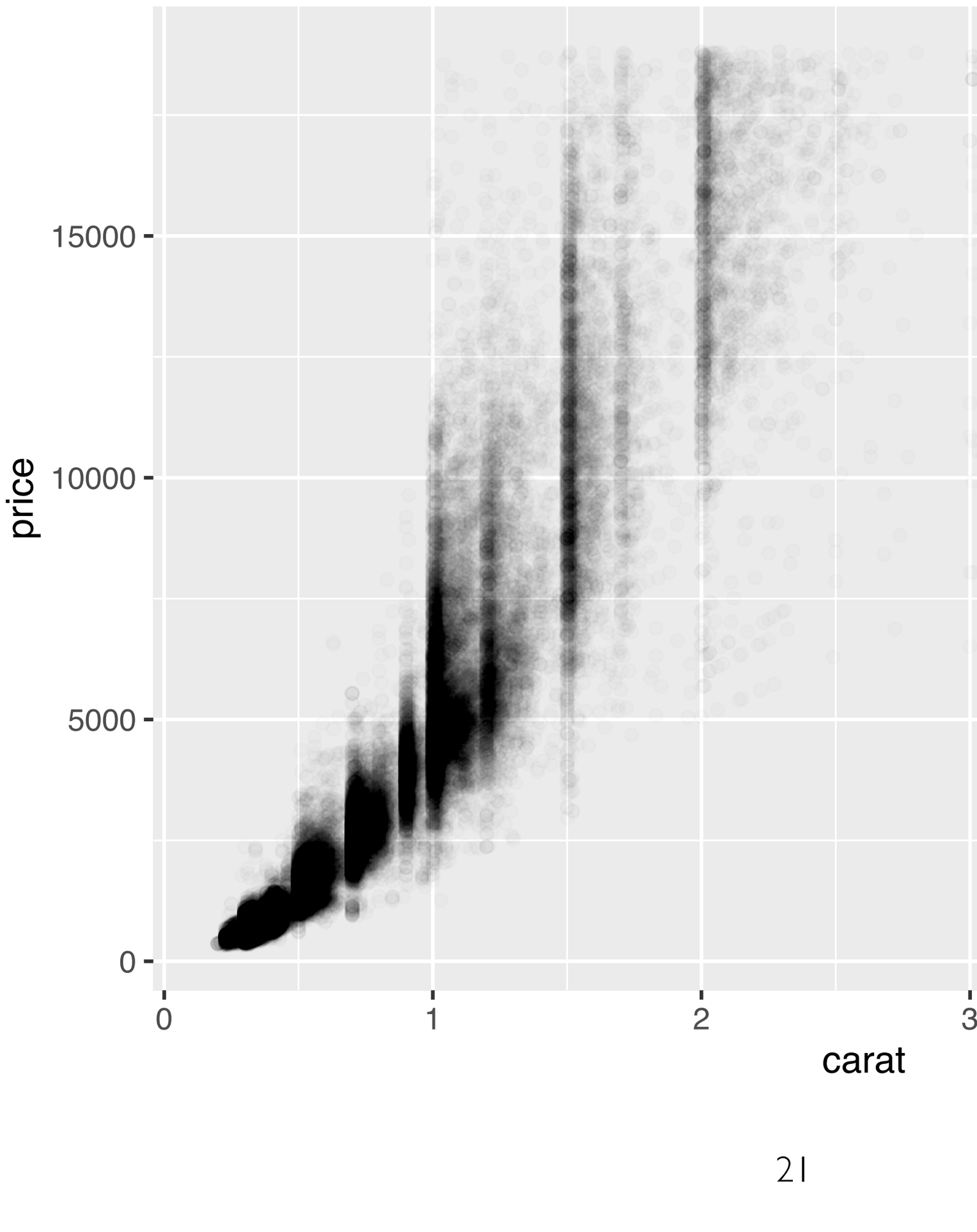
Covariation is the tendency of values
to vary together in a re

- Correlation and de
  - Variables share a statistic
- Causation
  - One variable directly infl
- Confounding
  - A third variable influence

https://xkcd.com/55

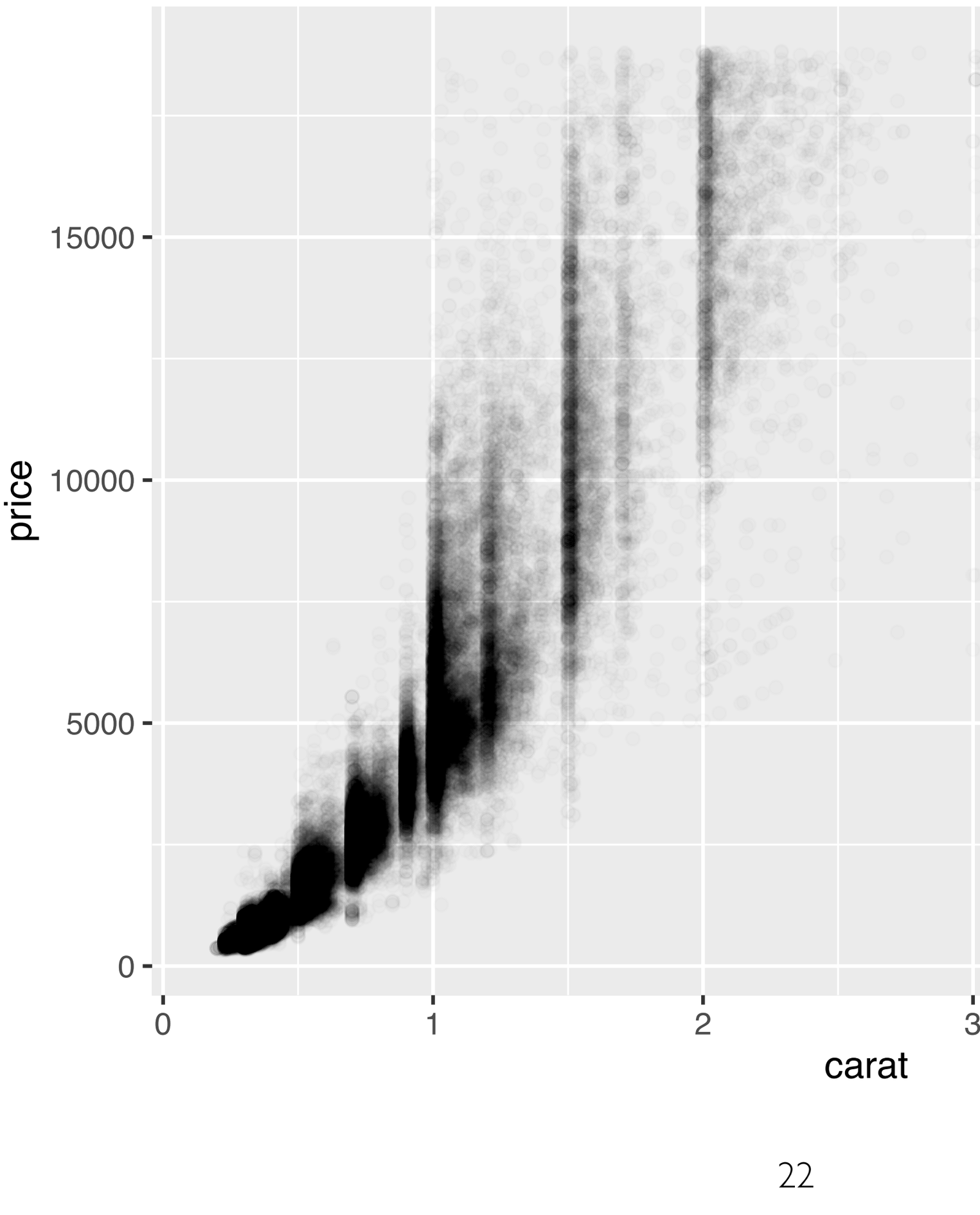
Two continuous: *pr*

What is the relationship betwee



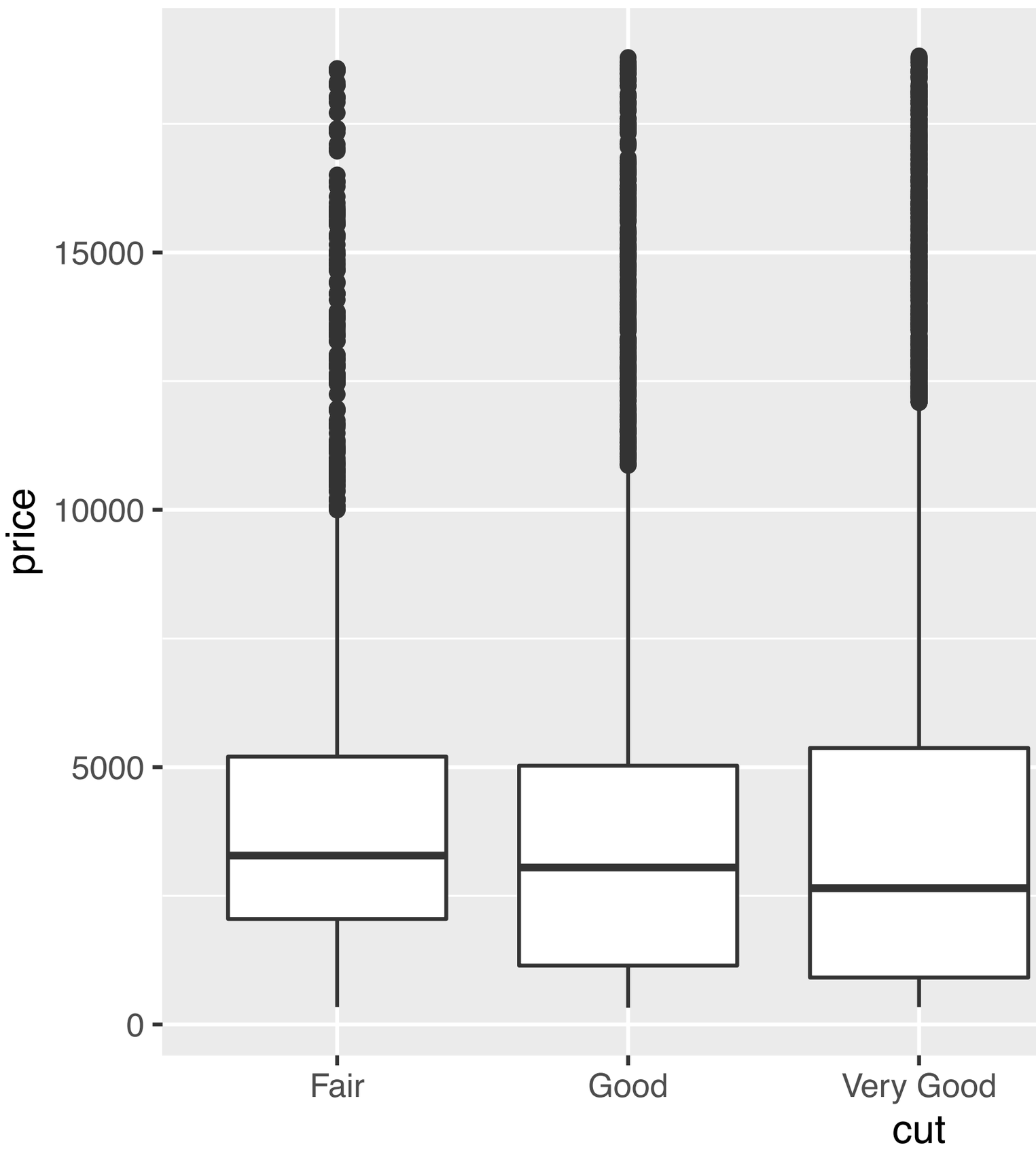
Two continuous: *pr*

Larger diamonds tend to be



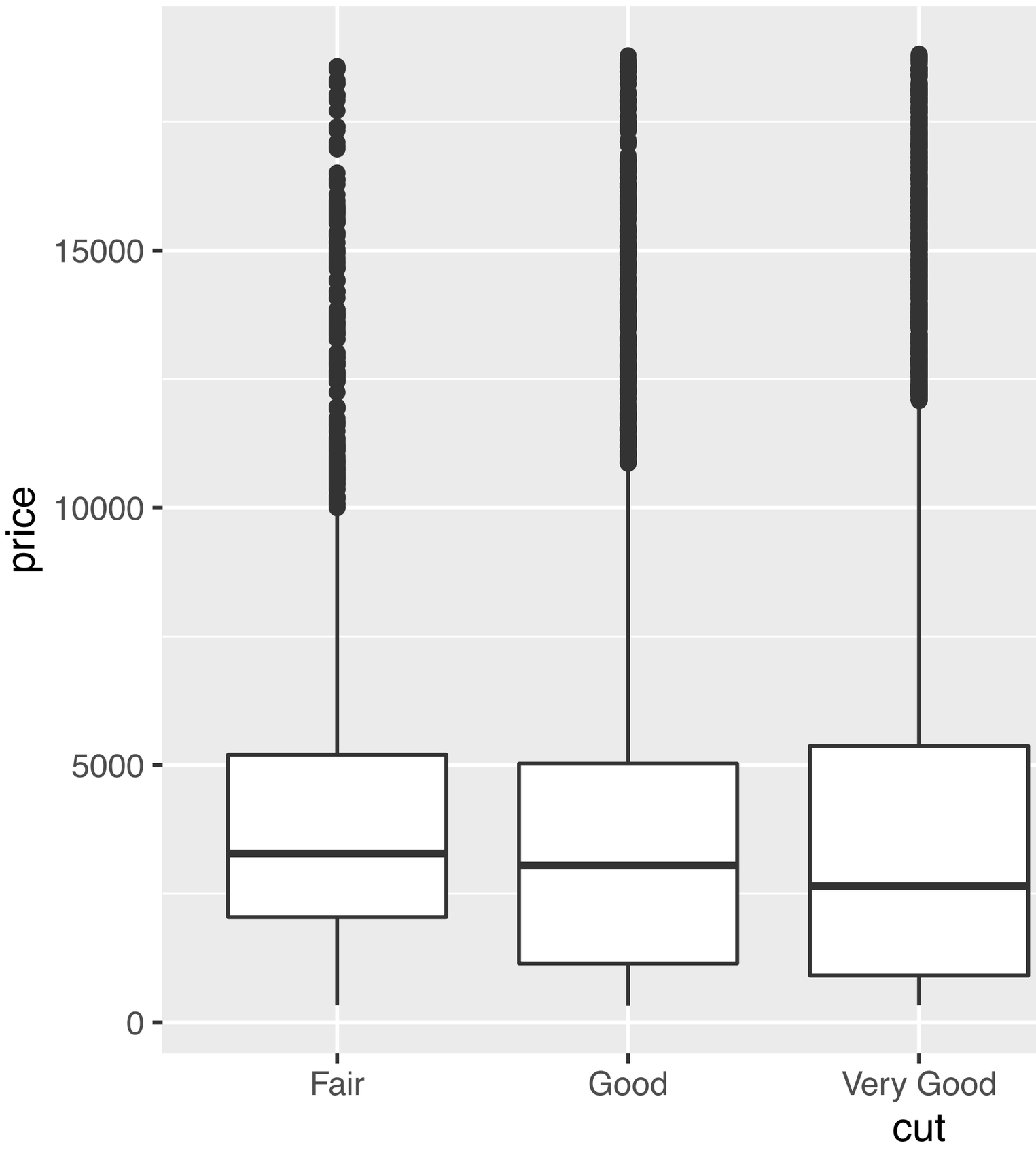
Categorical and continu

What is the relationship betwe



Categorical and continu

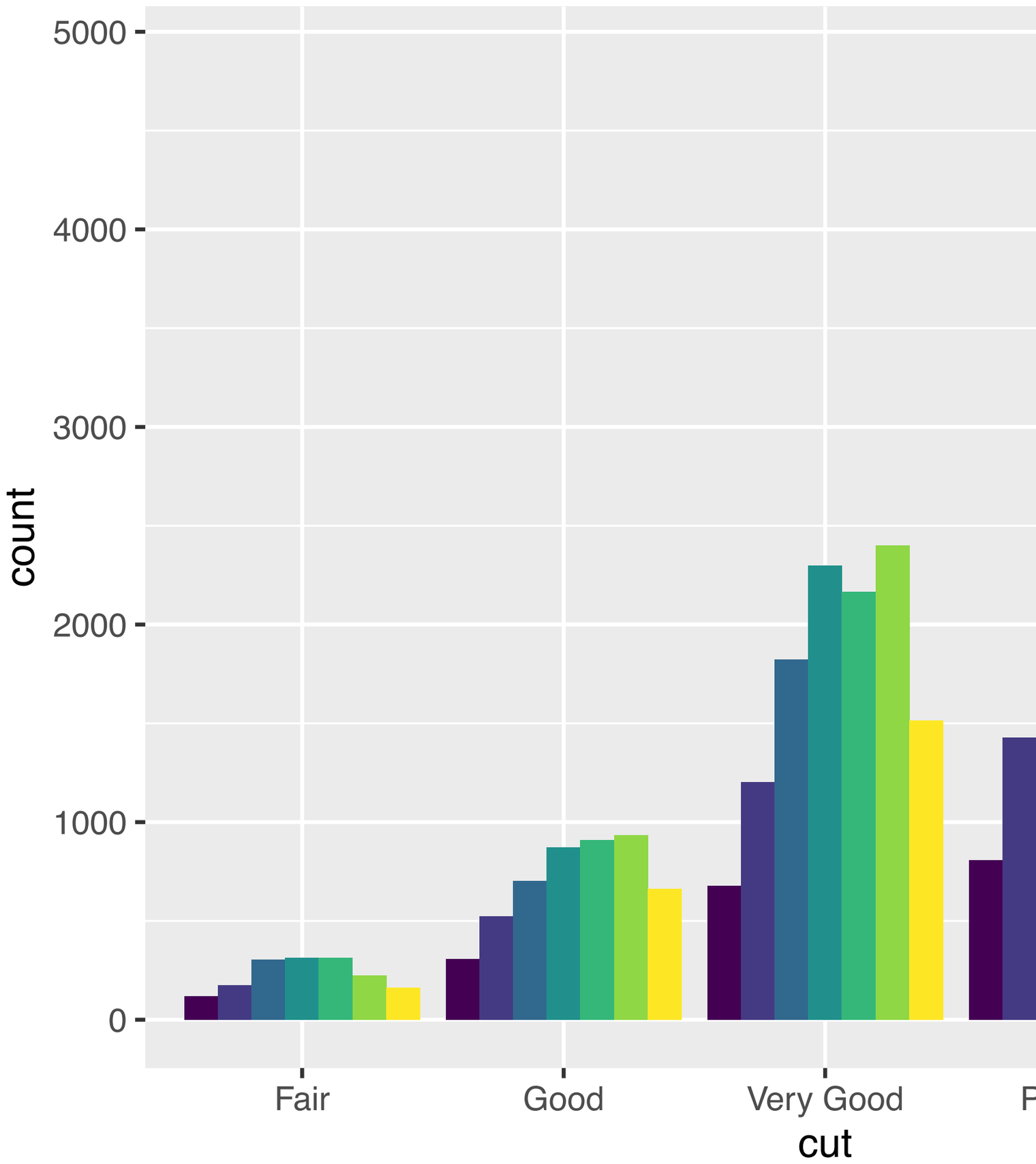
Higher quality cuts tend to be



Weird.

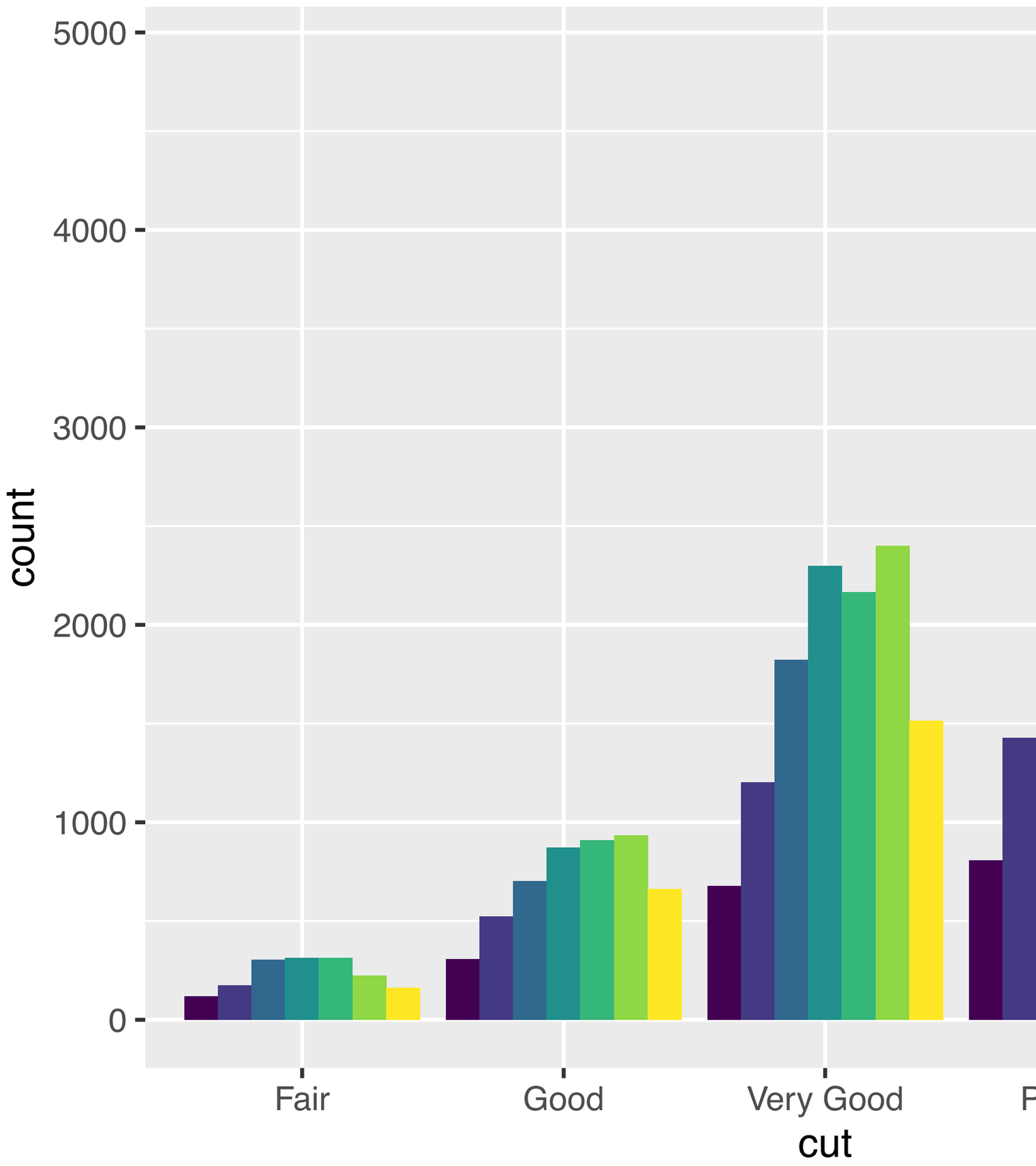
Two categorical: co

What is the relationship betwe



Two categorical: co

Higher quality cuts tend to have "G" col

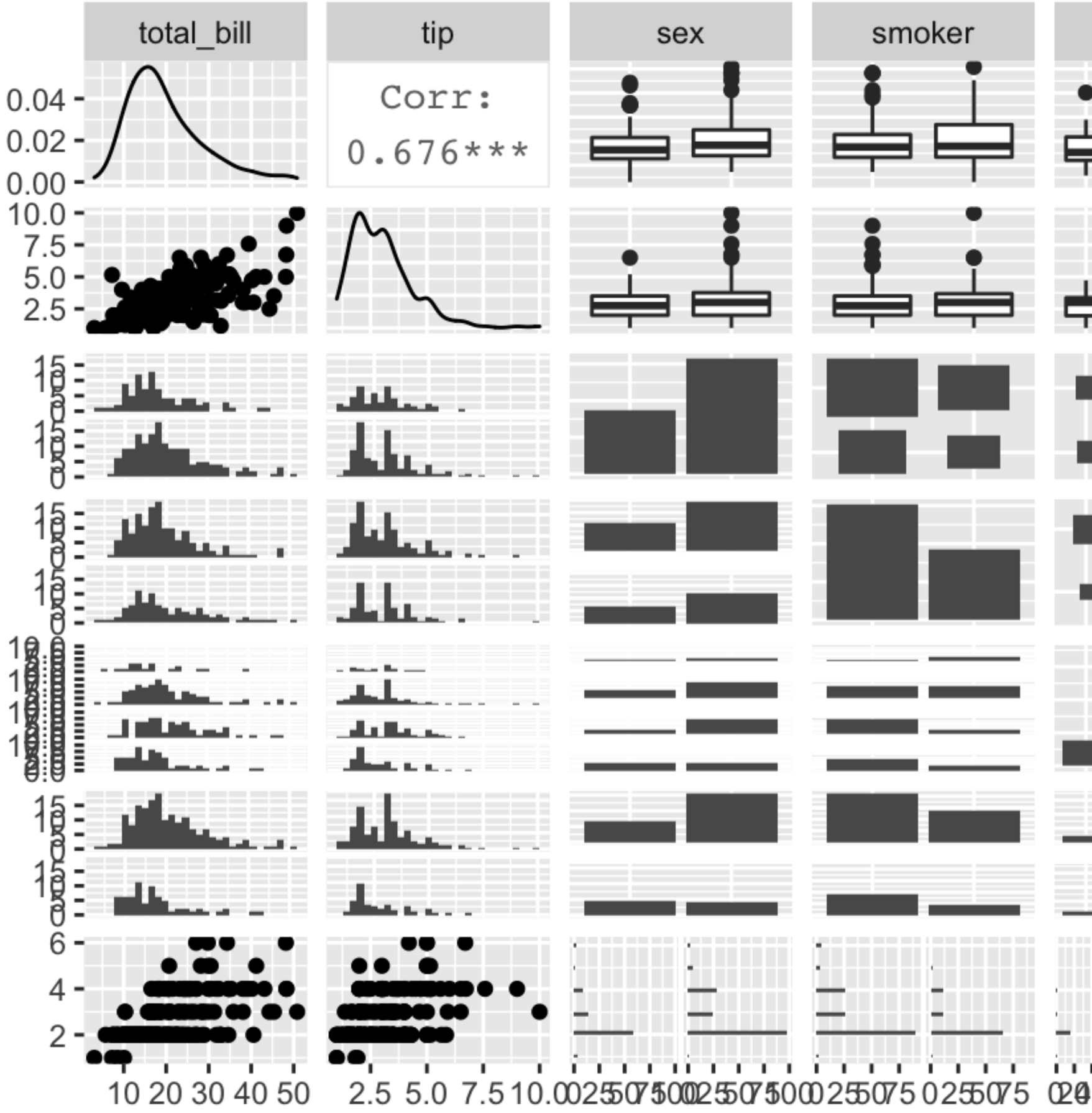


Approach to

- State research question
- Identify target variables answer questions
- Visualize target variable answers the questions
- Tell "the story" of the d

“INTEREST
VISUALIZAT

Too many pl



GGally: https://ggobi.github.io/ggally/a

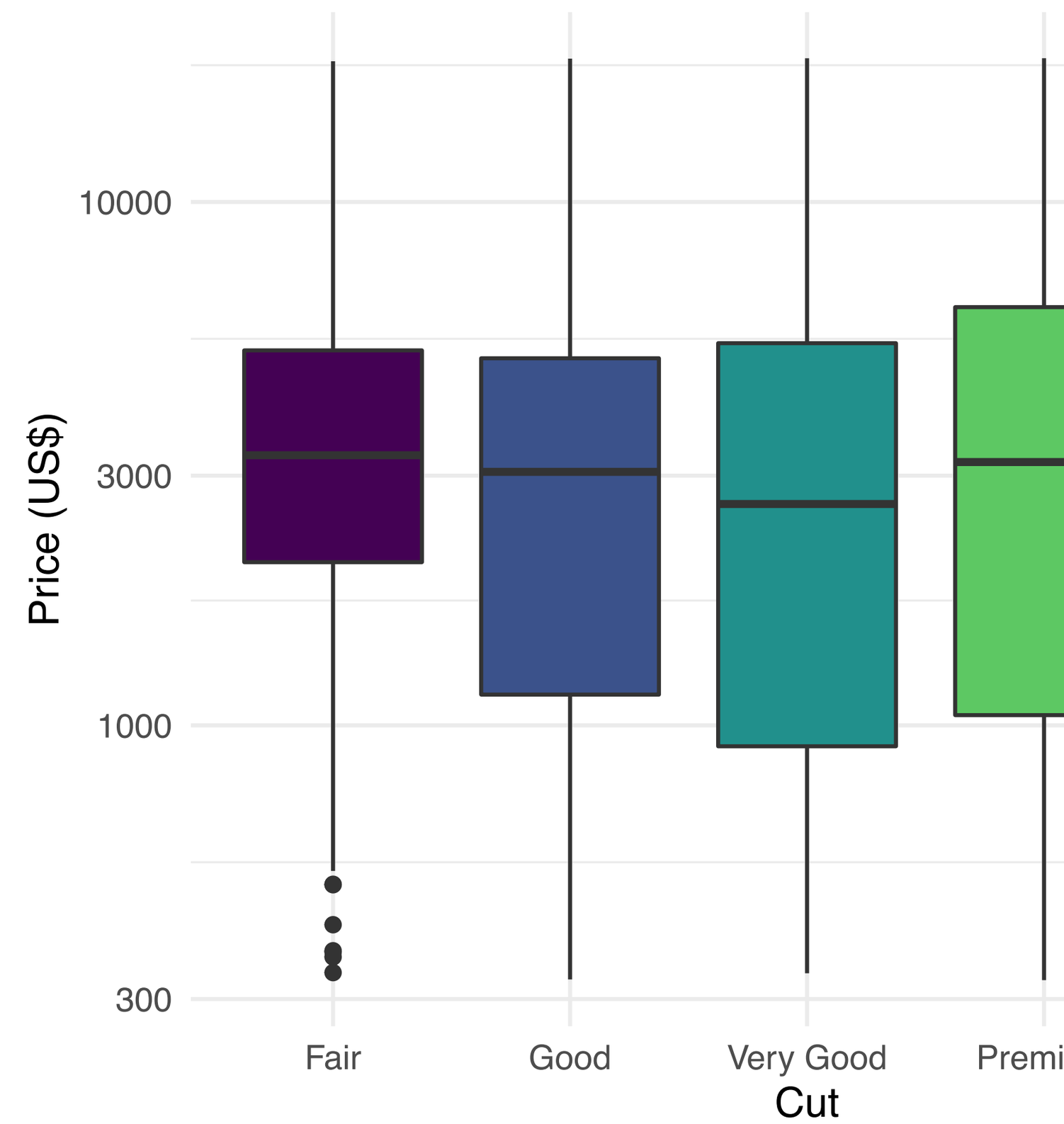
What makes a visualizatio

An “interesting” vis

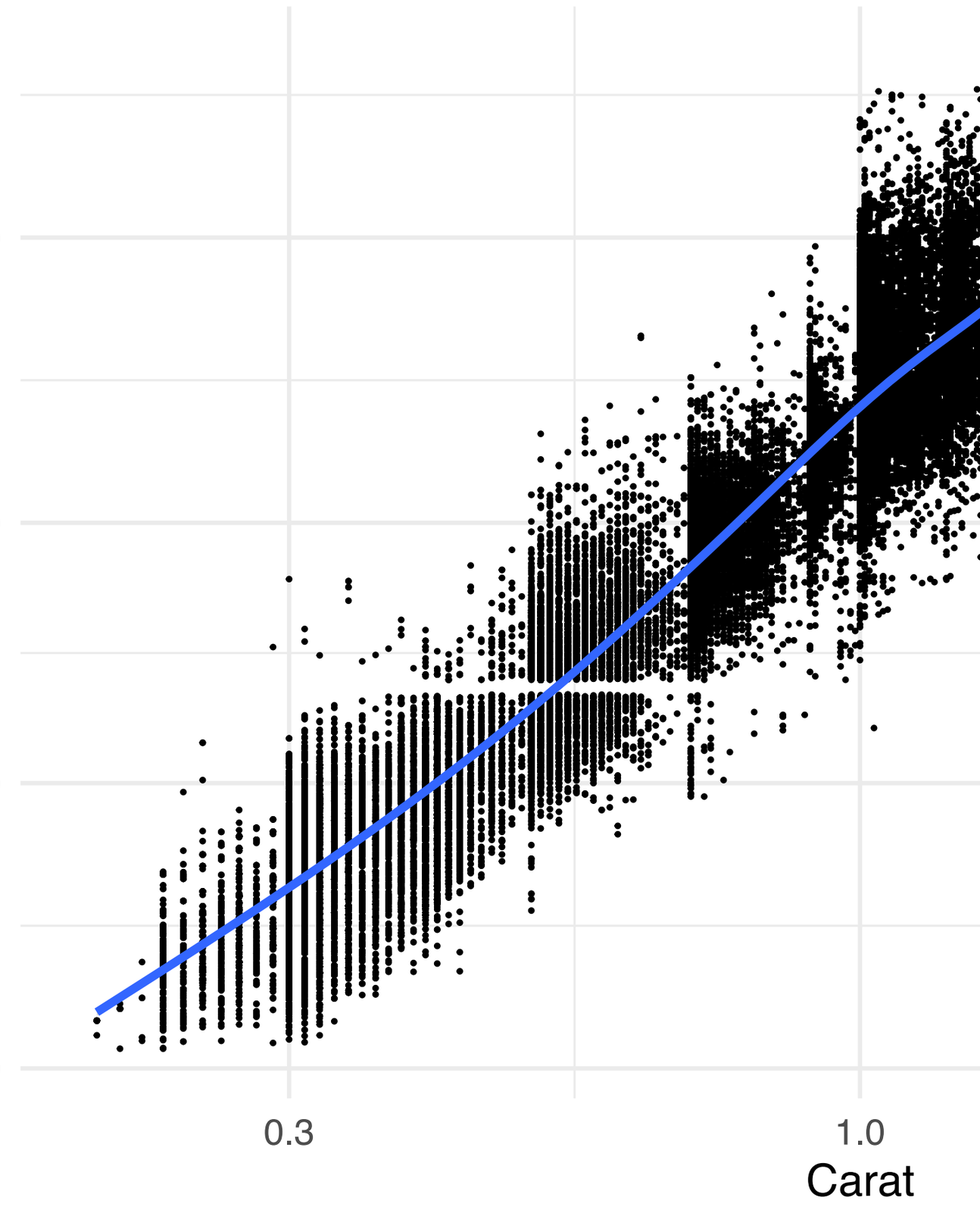
- establishes a relationshi
- explains something
- deviates from expectati

Deviates from exp

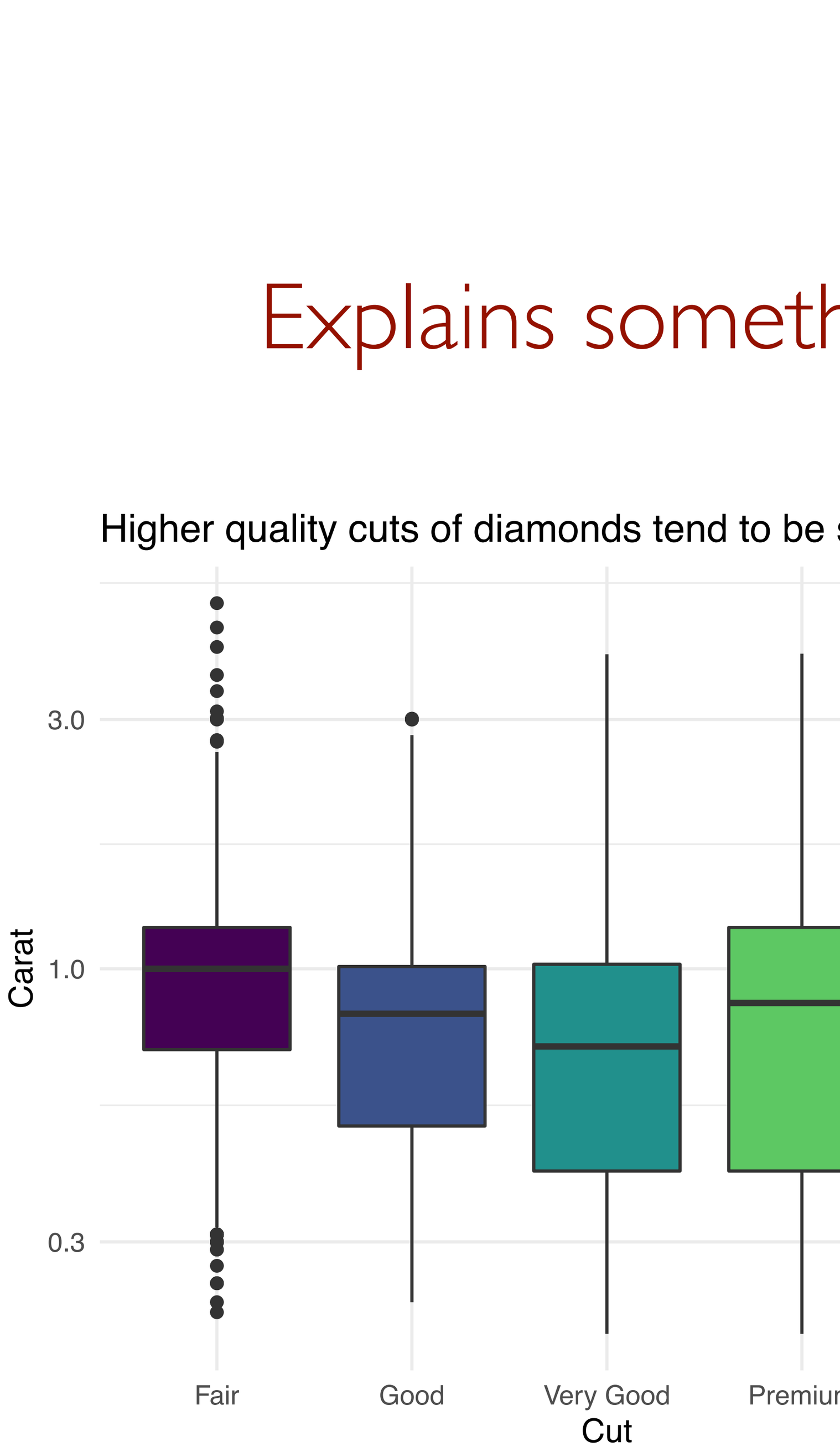
Higher quality cuts of diamonds tend to b



Establishes a rela



Explains someth



Explains someth