Workshop - Data visualization using `ggplot2`

Tim Winke, Humboldt University Berlin

The Grammar of Graphics

The Grammar of Graphics

- Visualisation concept created by Wilkinson (1999)
  - to define the basic elements of a statistical graphic

- Adapted for R by Wickham (2009)
  - who created the `ggplot2` package
  - consistent and compact syntax to describe statistical graphics
  - highly modular as it breaks up graphs into semantic components

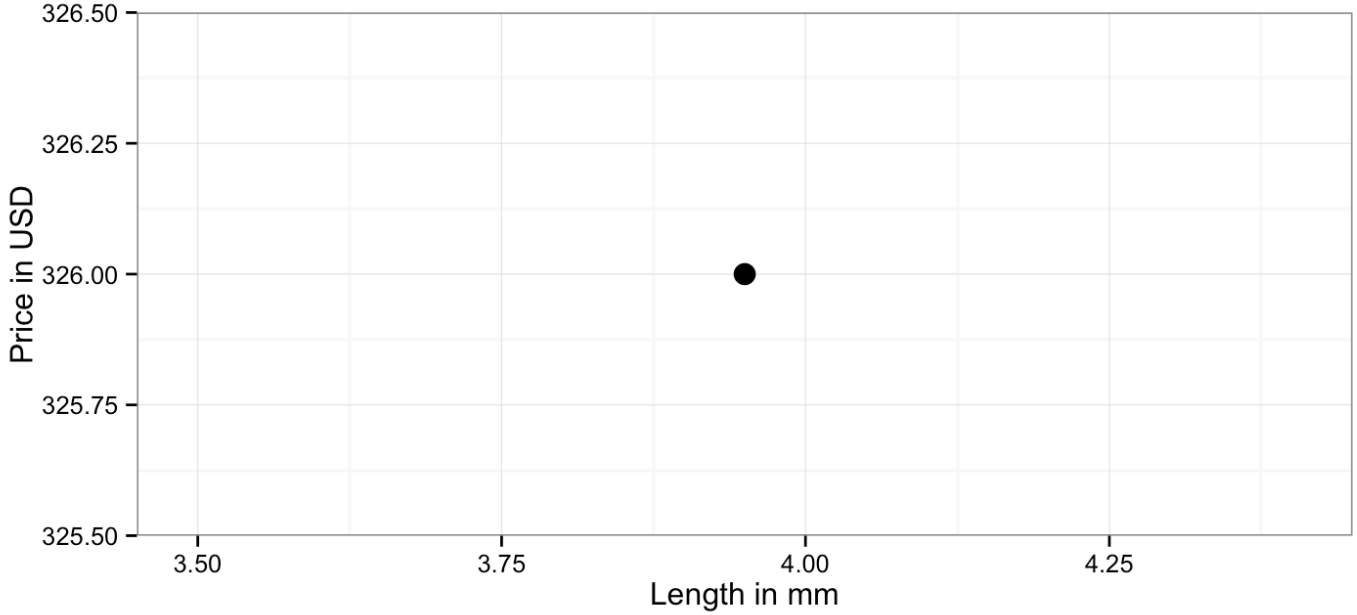
- Is *not* a guide which graph to choose and how to convey information best!

The Grammar of Graphics - Terminology

A statistical graphic is a …

- mapping of **data**
- to **aesthetic attributes** (color, size, xy-position)
- using **geometric objects** (points, lines, bars)
- and using **scaling** (x-scale, y-scale, color-scale, coordinate system)
- with data being **statistically transformed** (summarised, log-transformed)
- and mapped onto a specific **facet** and **coordinate system**

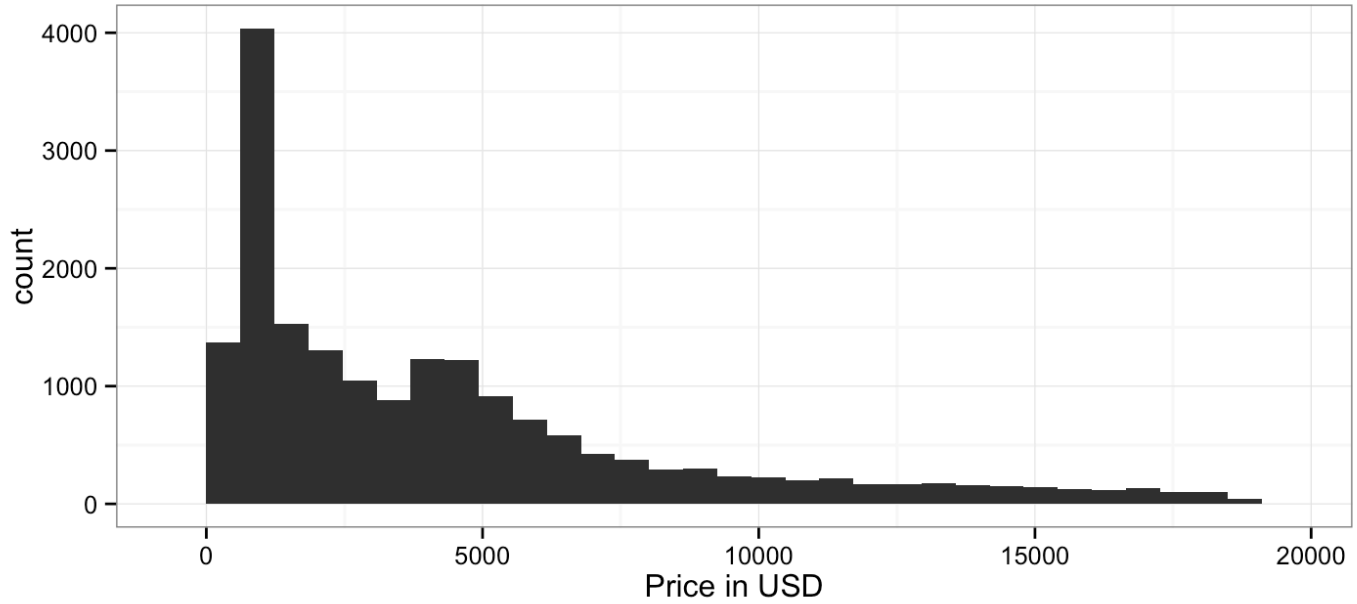
The Grammar of Graphics



The Grammar of Graphics

- Which **data** is used as an input?
- What **geometric objects** are chosen for visualization?
- What variables are **mapped** onto which attributes?
- What type of **scales** are used to map data to aesthetics?
- Are the variables **statistically transformed** before plotting?

The Grammar of Graphics



The Grammar of Graphics

- Which **data** is used as an input?
- What **geometric objects** are chosen for visualization?
- What variables are **mapped** onto which attributes?
- What type of **scales** are used to map data to aesthetics?
- Are the variables **statistically transformed** before plotting?

The Grammar of Graphics

The Grammar of Graphics

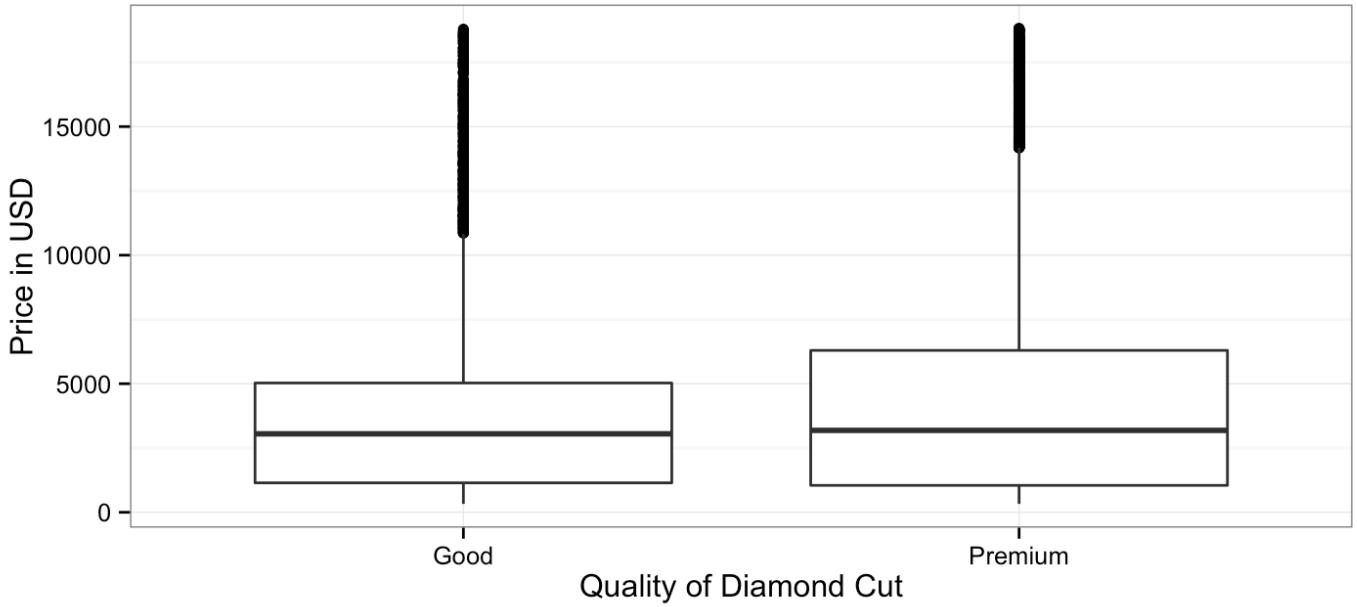
- Which **data** is used as an input?
- What **geometric objects** are chosen for visualization?
- What variables are **mapped** onto which attributes?
- What type of **scales** are used to map data to aesthetics?
- Are the variables **statistically transformed** before plotting?

The Grammar of Graphics

The Grammar of Graphics

- Which **data** is used as an input?
- What **geometric objects** are chosen for visualization?
- What variables are **mapped** onto which attributes?
- What type of **scales** are used to map data to aesthetics?
- Are the variables **statistically transformed** before plotting?
- Is any form of **facetting** applied?

The Grammar of Graphics



Graphics with `ggplot2`

Data preparation

```
library("ggplot2")
packageDescription("ggplot2")
```

```
## Package: ggplot2
## Type: Package
## Title: An Implementation of the Grammar of Graphics
## Version: 1.0.1
## Authors@R: c( person("Hadley", "Wickham", role = c("aut", "cre"),
##        email = "h.wickham@gmail.com"), person("Winston", "Chang",
##        role = "aut", email = "winston@stdout.org") )
## Description: An implementation of the grammar of graphics in R. It
##        combines the advantages of both base and lattice graphics:
##        conditioning and shared axes are handled automatically, and
##        you can still build up a plot step by step from multiple
##        data sources. It also implements a sophisticated
##        multidimensional conditioning system and a consistent
##        interface to map data to aesthetic attributes. See
##        http://ggplot2.org for more information, documentation and
##        examples.
```

Data preparation

```
data("diamonds")
# Prices of 50,000 round cut diamonds

# A dataset containing the prices and other attributes of almost 54,000 diamonds.
# The variables are price in USD, carat, cut quality,...
# help(diamonds)
head(diamonds)
```

```
##   carat       cut color clarity depth table price    x    y    z
## 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 4.20 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
```

Basics: Initiate ggplot object

```
qplot(data, ...) #close to plot() fct with compressed functionality and lots of defaults
ggplot(data, mapping = aes(), ...) #the main plotting function
```

- `data`: the data set employed
- `mapping`: list of asthethic assignments
  - `aes(x, y, color, size, fill, shape)`

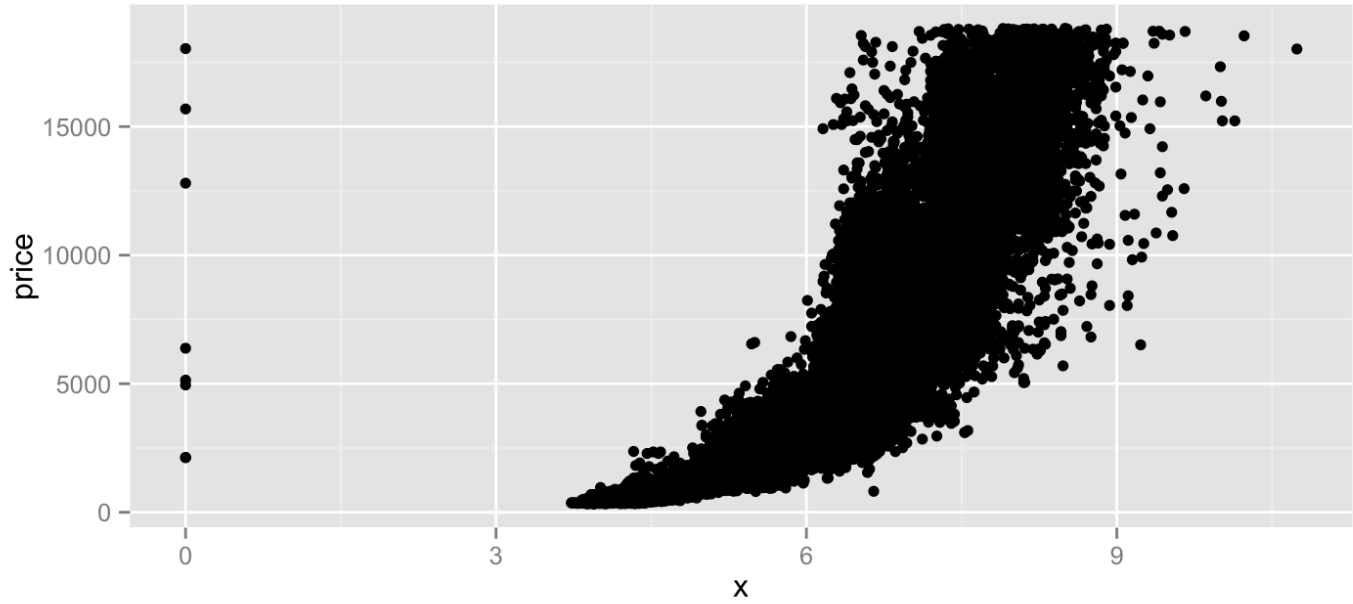
Basics: Initiate ggplot object

```
ggplot(data = diamonds, mapping = aes(x=x, y=price))
# Warning: No layers in plot
```

- `ggplot()` itself …
  - is not a plotting layer but initializes a ggplot object
  - declares the input data and some common aesthetics
- Add layers by using the **+** operator

Basics: Geometric objects

```
ggplot(data = diamonds, mapping = aes(x=x, y=price)) + geom_point()
```



Basics: Geometric objects

```
geom_point(mapping = NULL, data = NULL, stat, ...)
```

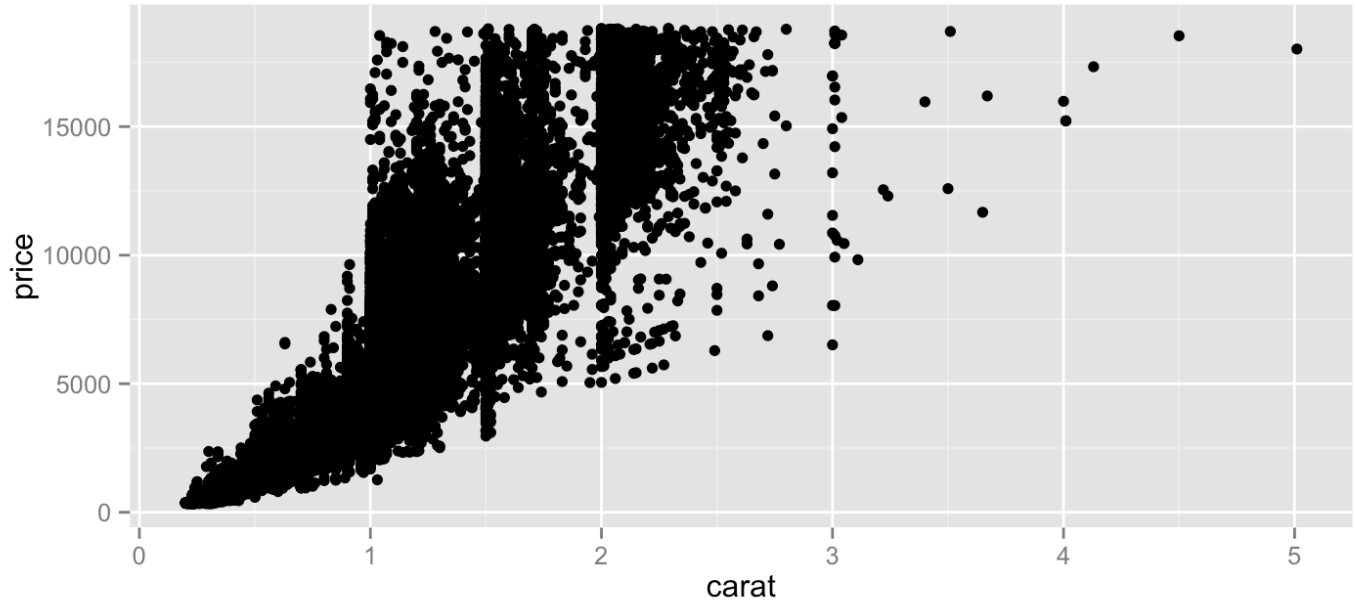
- `mapping`: list of asthethic assignments `aes()` for `geom` object
- `stat`: statistical transformation required for `geom` object
- `NULL`: inhibit values from `ggplot()`
- `...` other arguments,
    - often aesthetics you want to set unconditionally, e.g. color="red"

Geometric objects - Exercise

- Excercise:
  1. Load `library(ggplot2)`
  2. Load `data(diamonds)` from the `ggplot2` package
  3. Create a scatterplot of `price` and `carat`
     - Use `ggplot(data = ..., mapping = aes(x=..., y=...))` to initiate an ggplot object
     - Use `+ geom_point()` to create a scatterplot layer

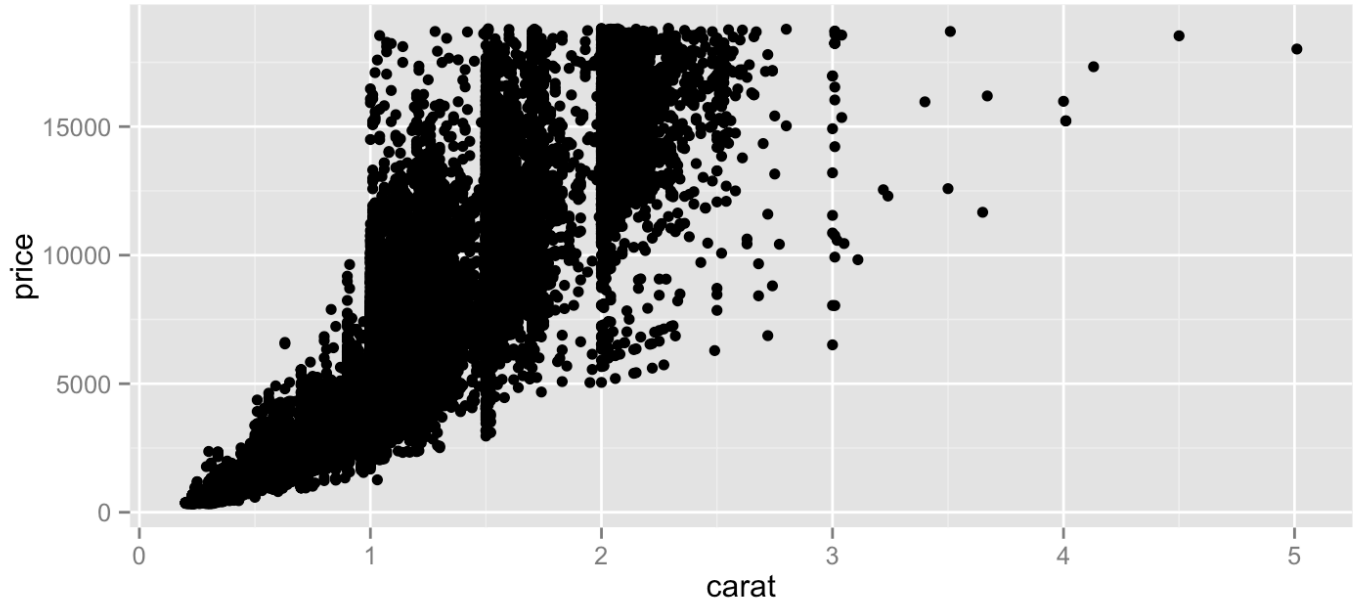
Geometric objects - Exercise

```
ggplot(data = diamonds, mapping = aes(x=carat, y=price)) + geom_point()
```



Geometric objects - Exercise

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



Basics: Geometric objects

Examples for basic `geom_` functions

```
geom_point(mapping = NULL, data = NULL,
stat = "identity", position = "identity", ...)

geom_line(mapping = NULL, data = NULL,
stat = "identity", position = "identity", ...)

geom_boxplot(mapping = NULL, data = NULL,
stat = "boxplot", position = "dodge", outlier.color = "black",
                      outlier.shape = 16, outlier.size = 2, ...)
```

Basics: Geometric objects

- Add, combine and edit layers like a toolbox
- Extensive list of all `ggplot2` objects can be found at
  - docs.ggplot2.org
  - including many examples at the end of each topic

Basics: Mapping aesthetics

- Besides mapping onto x- and y-position
    - variables can be assigned to `geom` aesthetics

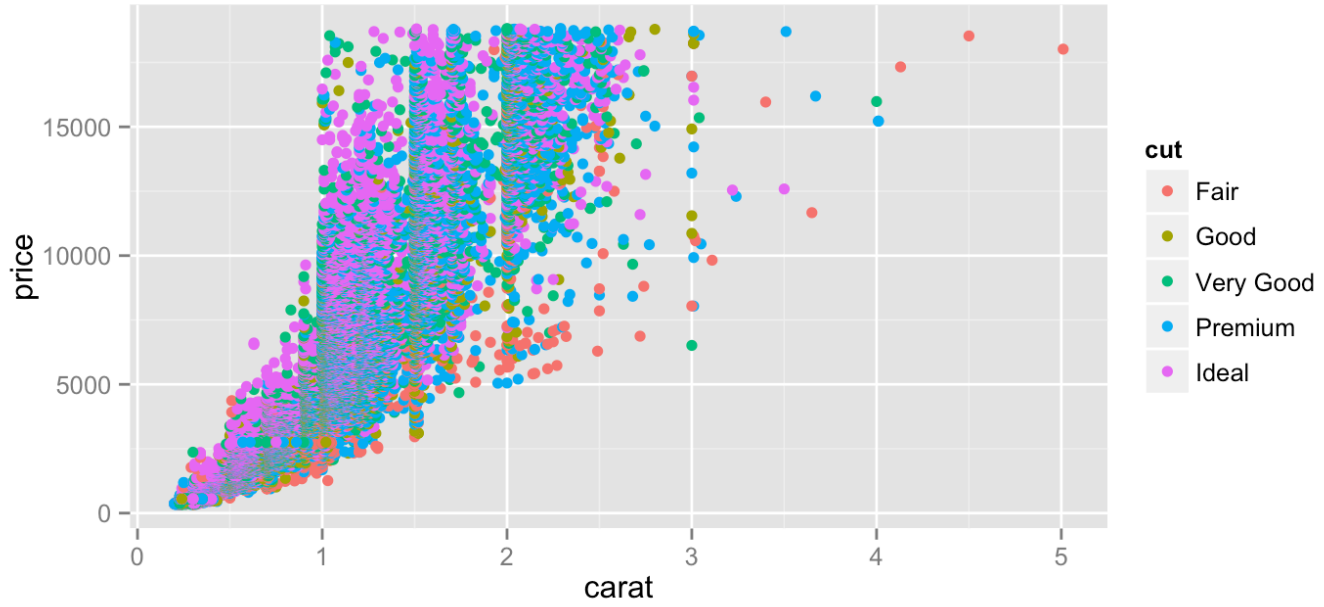
Examples:

```
geom_point(aes(x=carat, y=price, size = carat ))#: point size varies with `carat`
geom_point(aes(x=carat, y=price, color = carat))#: color varies with `carat`
geom_point(aes(x=carat, y=price, fill = carat)) #: fill color varies with `carat`
geom_point(aes(x=carat, y=price, linetype = carat))#: linetype varies with `carat`
```

Basics: Mapping aesthetics

Setting mappings for `geom` extends or replaces `ggplot()` mappings

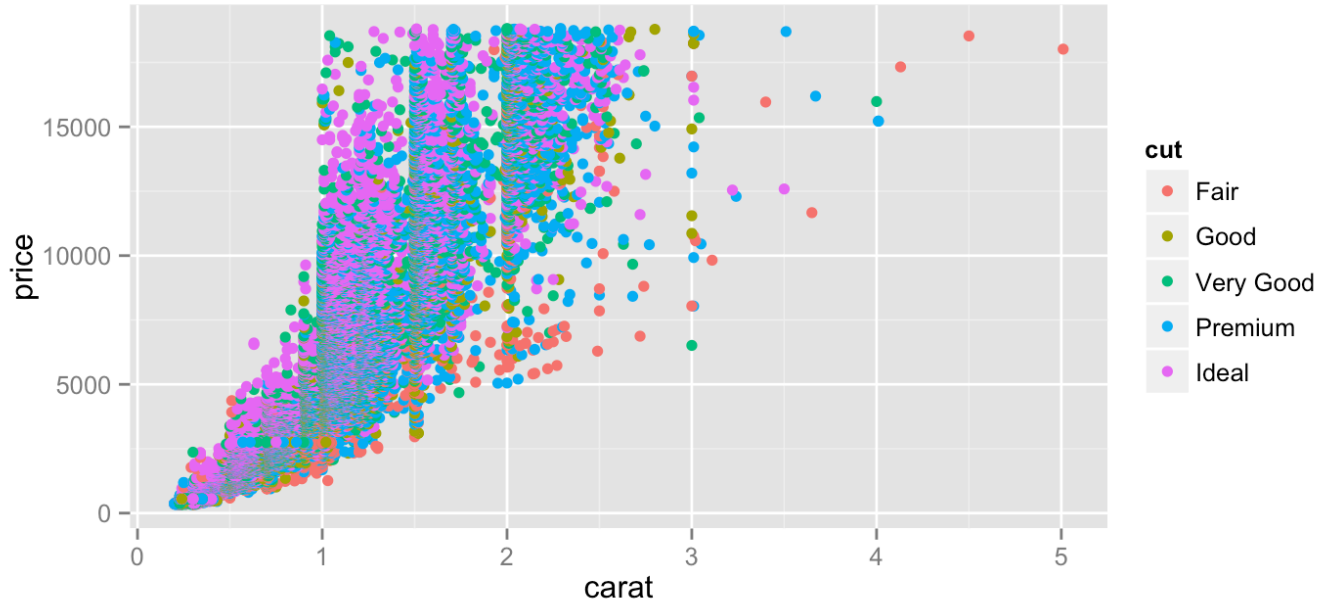
```
ggplot(diamonds, aes(x=carat, y=price)) + geom_point(aes(color = cut))
```



Basics: Mapping aesthetics

But you can also state universal mappings within `ggplot()` objects

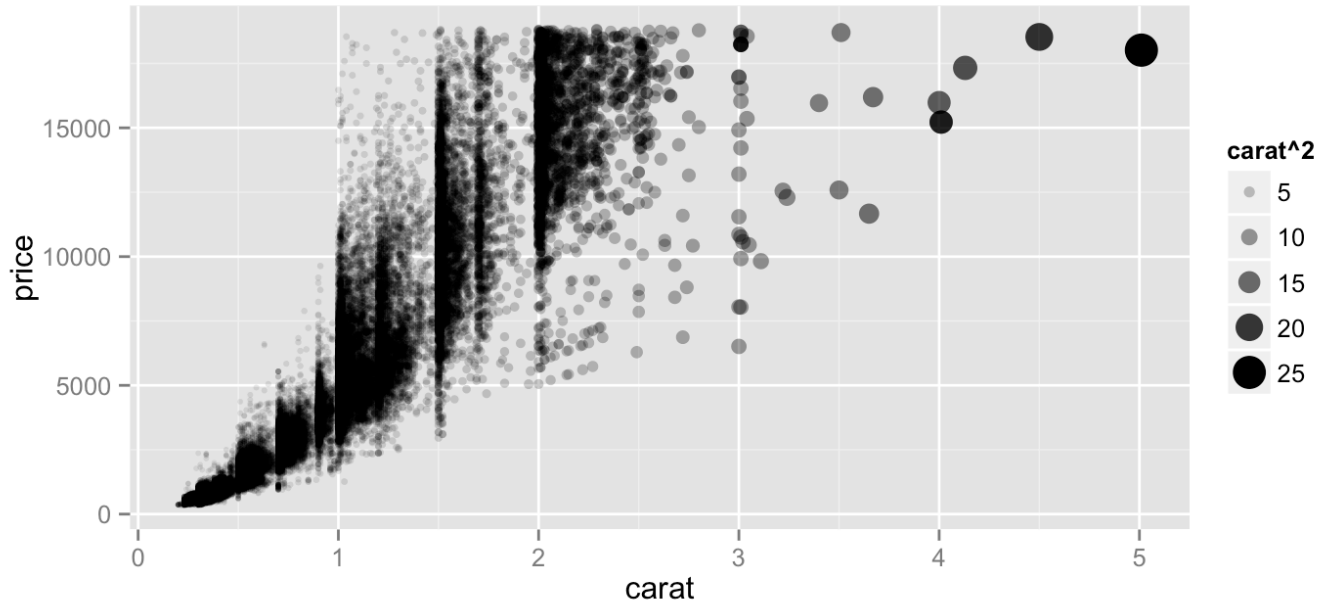
```
ggplot(diamonds, aes(x=carat, y=price, color = cut)) + geom_point()
```



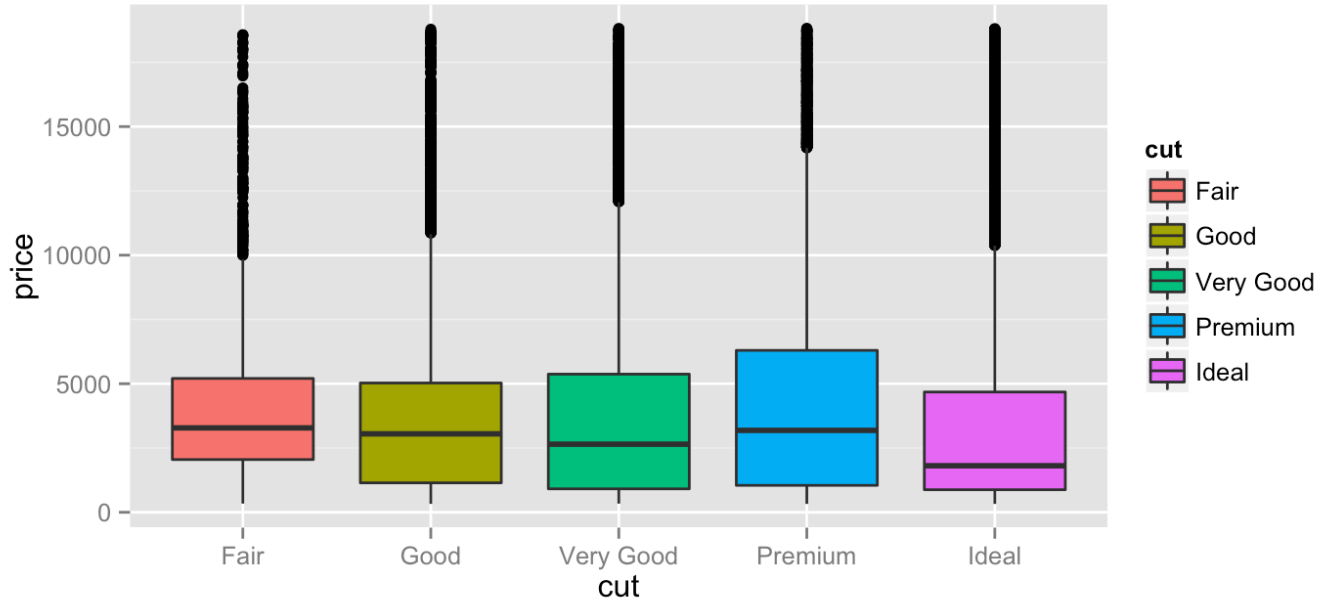
Basics: Mapping aesthetics

Including some additional manipulations of variables

```
ggplot(diamonds, aes(x=carat, y=price, size=carat^2, alpha=carat^2)) + geom_point()
```

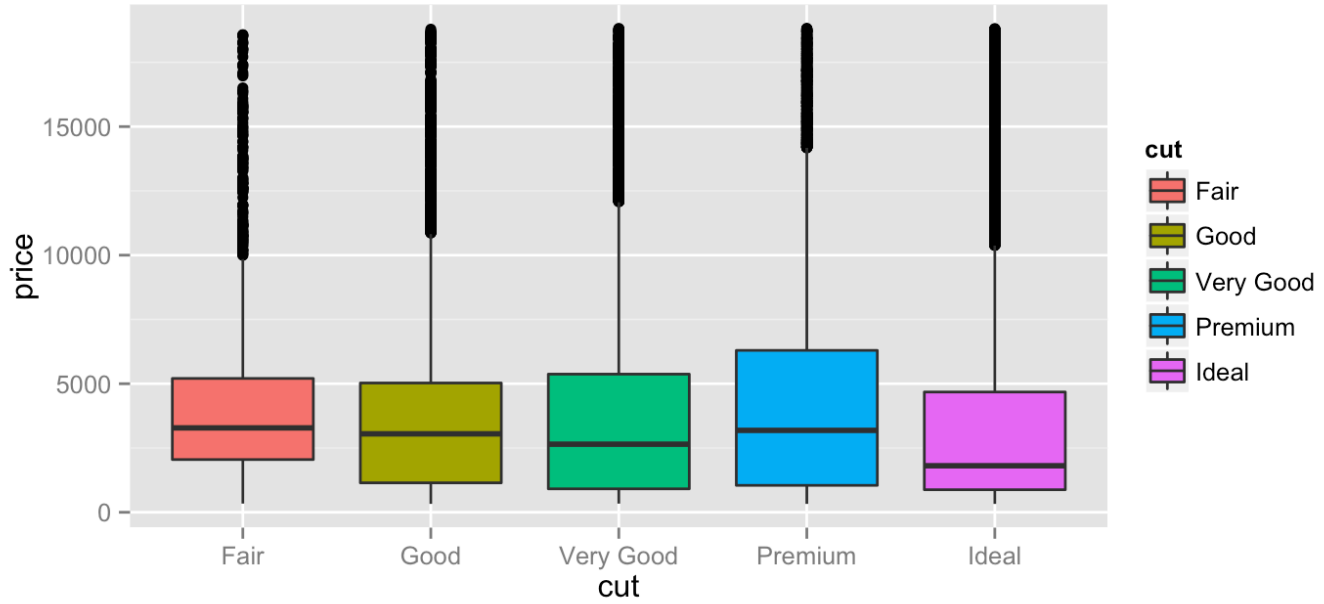


Mapping aesthetics - Exercise

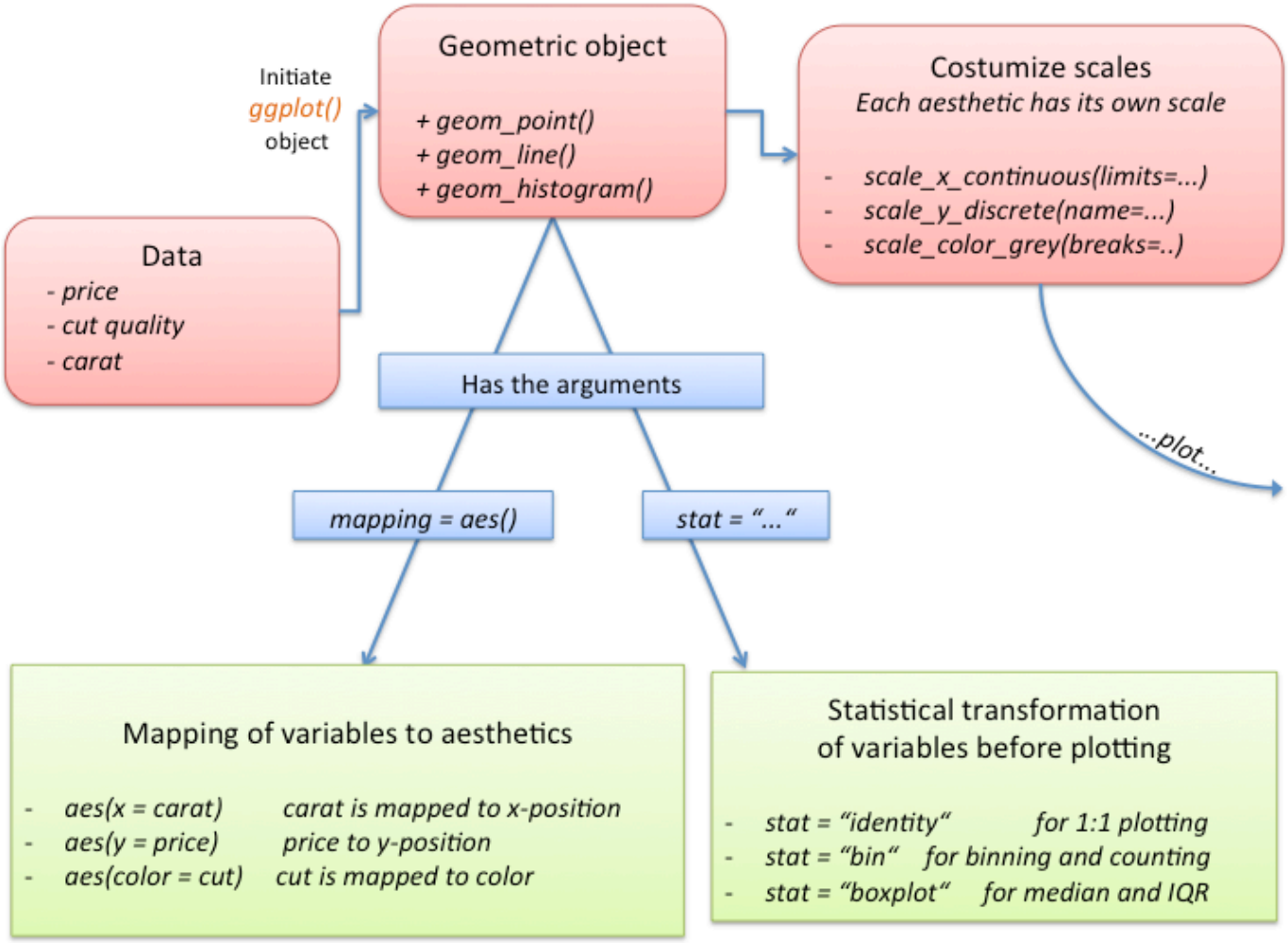


Mapping aesthetics - Exercise

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

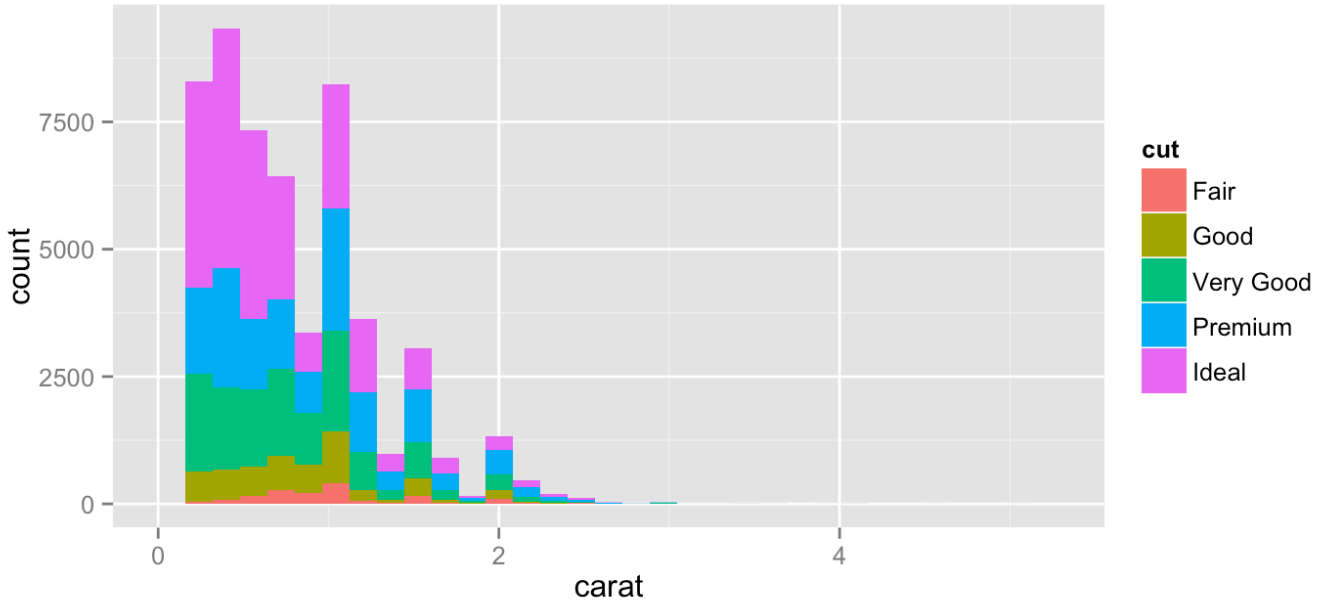


Basics - Summary



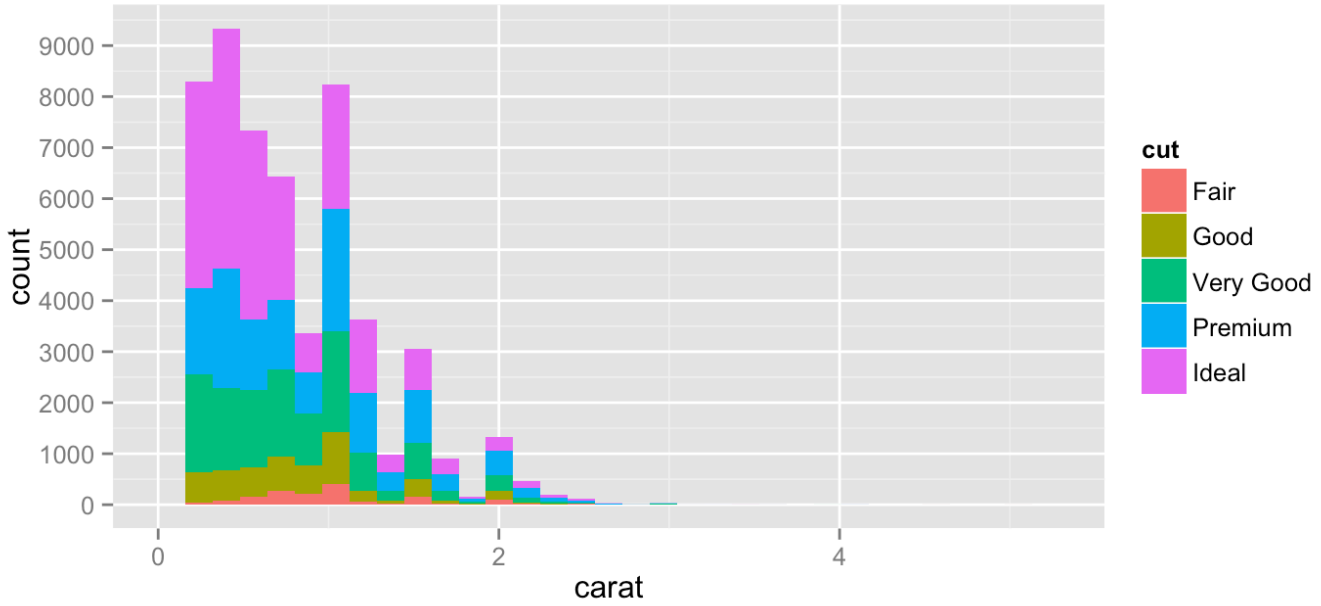
Introduction to scales

```
ggplot(diamonds, aes(x=carat, fill=cut)) +  geom_bar()
```



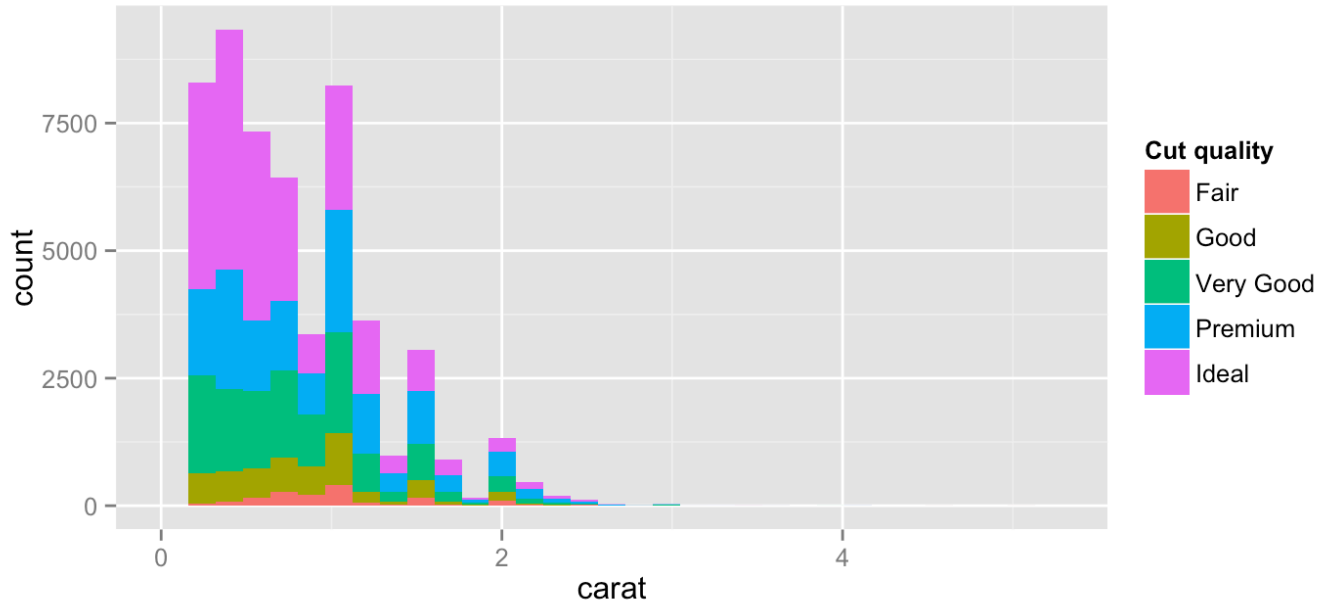
Introduction to scales

```
ggplot(diamonds, aes(x=carat, fill=cut)) +  geom_bar() +
  scale_y_continuous(breaks = seq(0,9000,1000))
```



Introduction to scales

```
ggplot(diamonds, aes(x=carat, fill=cut)) +  geom_bar() +
  scale_fill_discrete(name="Cut quality")
```



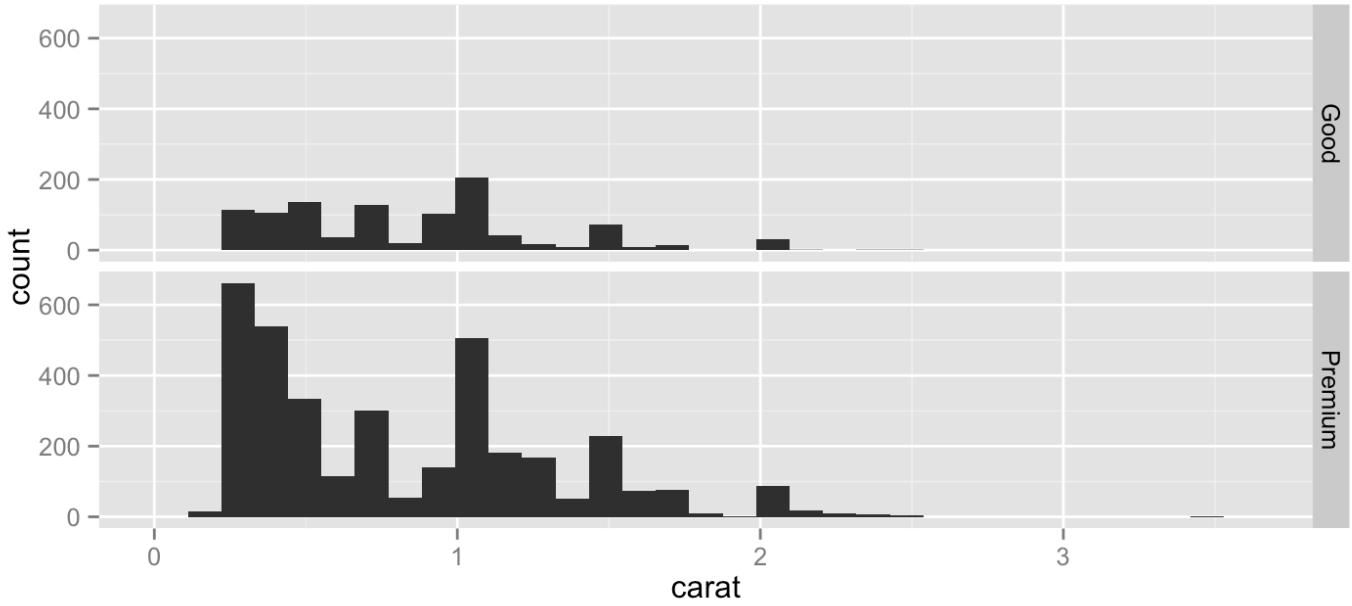
Faceting

Faceting


3000
2000
1000
0
count
Good
Premium
0
5000
10000
15000
20000
Price in USD
Quality of Diamond Cut
Good
Premium

Faceting using `facet_grid`

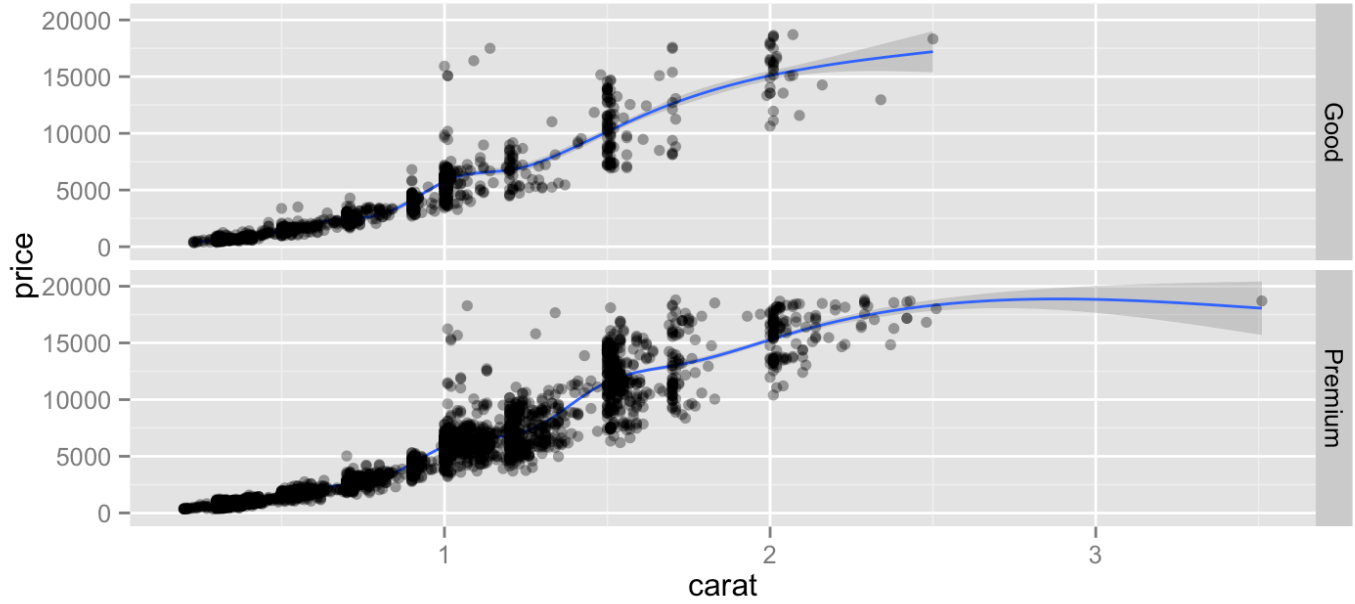
```
ggplot(diamondssub, aes(x=carat)) + geom_histogram() +
  facet_grid(cut ~ .)
```



- single column useful to compare distributions

Faceting using `facet_grid` - Exercise

```
ggplot(diamondssub, aes(x=carat, y= price)) + geom_smooth() +
  facet_grid(cut ~ .) +
  geom_point(alpha=0.4)
```



Final remarks on `ggplot2`

Final remarks on `ggplot2`

- Why to use `ggplot2`
  - Large community as one of the most popular R packages
  - Uses sensible and attractive decisions about
    - dimensions, scales and colors by default
  - Additional packages in-the-same-vain like `ggplot2` for
    - geographical information (`ggmap`),
    - genomic data (`ggbio`),
    - Markov Chain Monte Carlo simulations (`ggmcmc`)
    - and interactive graphics (`ggvis`)

Helpful website and books

- docs.ggplot2.org with many examples at the end of each topic
- cookbook-r.com/Graphs/ provides solutions for frequent problems
- CEB institute handout ggplot2 from the Basel Biometric Section.

Thank you for your attention!

If you have any questions or remarks don't hesitate to contact me!
tim.winke@gmail.com