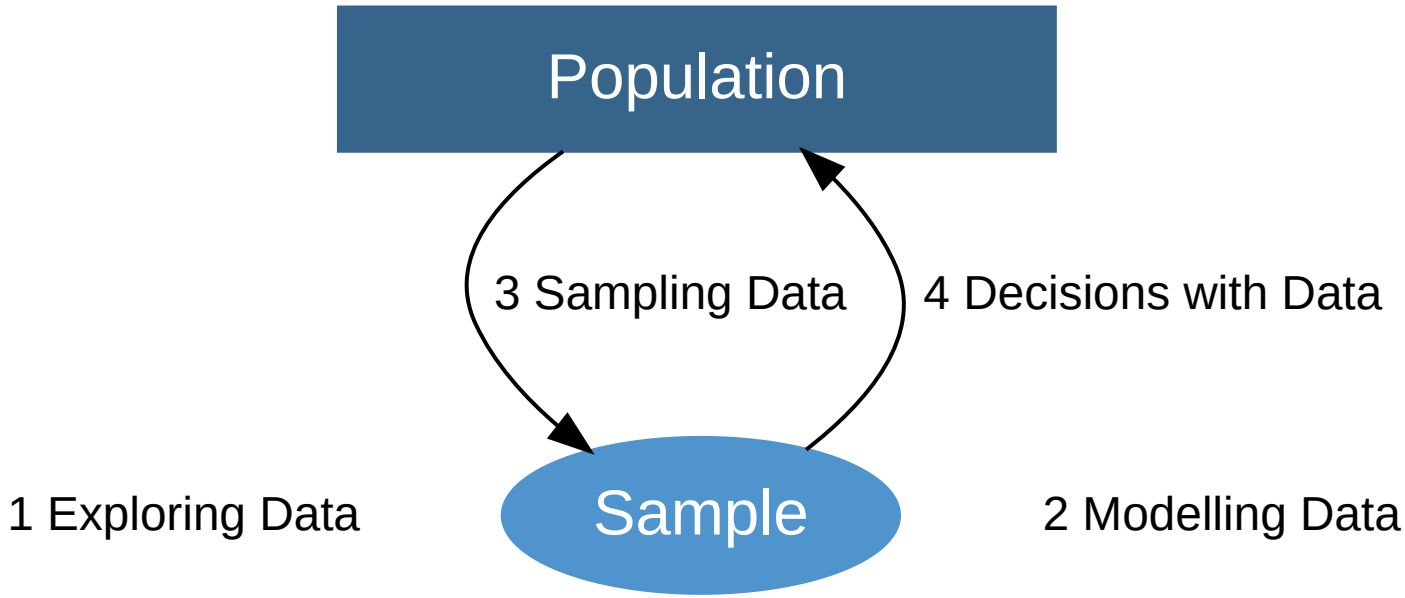
Data Visualisation

Exploring Data | Data & Graphical Summaries

Unit Overview



Module1 Exploring Data

Design of Experiments

Where did the data come from & can we make reliable conclusions?

Data & Graphical Summaries

What type of data do we have & how can we visualise it?

Numerical Summaries

What are the main features of the data?

Data Visualisation

Data Story

What is the price point for a diamond?

The Sydney Morning Herald

WORLD EUROPE ART

Rarest white diamond ever to be auctioned

9 February 2018 — 3:17pm

London: A flawless diamond, the size of a large strawberry, is expected to fetch a world record price when it goes on sale at Sotheby's in London this month.

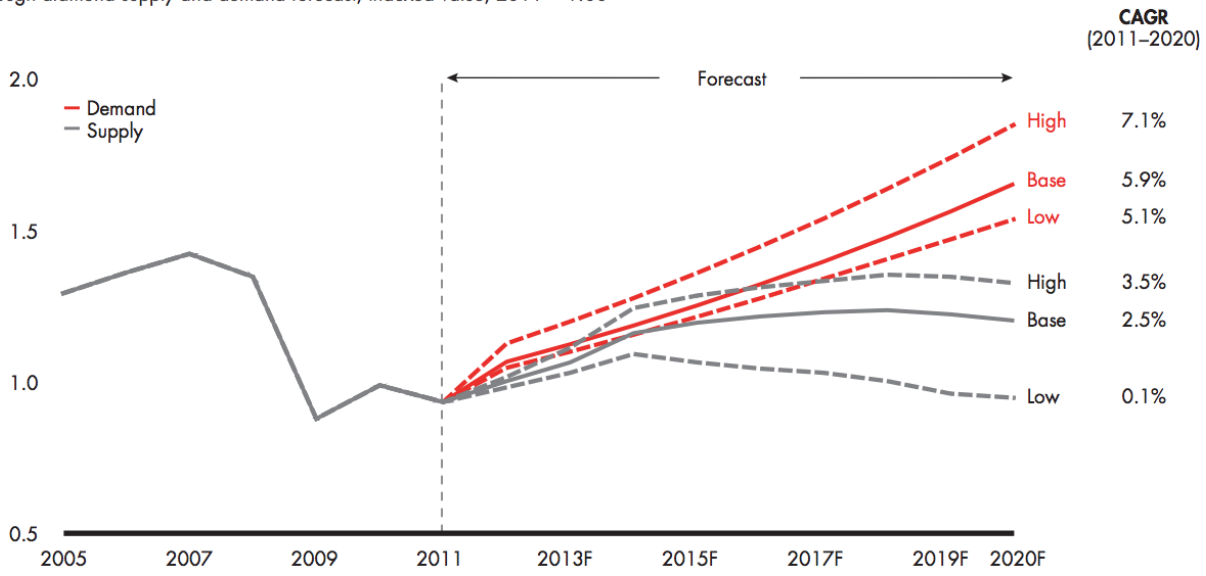
Weighing just over 102 carats, the round, brilliant white stone is smaller than a 163-carat oval diamond sold in Hong Kong in 2013, which holds the record price per carat.

But Sotheby's, which also handled that Hong Kong sale, expects the smaller stone's rarity and high quality will see it attract an even higher price.

SMH

Figure 7.19: We project that global demand for rough diamonds will exceed supply through 2020

Global rough diamond supply and demand forecast, indexed value, 2011 = 1.00



Note: Historical values are presented in 2011 dollars
Sources: IDEX, Tacy Ltd. and Chaim Even-Zohar; Kimberley Process Statistics; company plans; expert interviews; publication analysis

Bain

Diamonds

- Diamonds are the hardest known natural material and one of the world's major natural resources.
  - Australia has the largest reserves estimated at around 210 million carats.
- Diamonds are used for jewellery (30%, sales of US79 Billion in 2015) and industrial applications (70%).
  - The 59.6 carat Pink Diamond is currently the most expensive gemstone selling for $71.2 million in 2017.

Pricing Diamonds

- As each diamond is unique, buyers need to investigate the price point.
- Diamonds are graded by 4 qualities, know as the “4Cs”:
  - **carat** (weight, where a metric ‘carat’ is 200mg)
  - **cut** (quality of the cut according to proportions, symmetry and polish)
  - **color** (colour-graded from D-colourless to Z-saturated)
  - **clarity** (graded from flawless to inclusions)

How to Choose a Diamond: Four-Minute GIA Diamond Grading Guide by GIA

Diamond Grading

Diamonds dataset

The `diamonds` dataset contains the prices and 9 other attributes of almost 54,000 diamonds. Data Dictionary

```
# install.packages("ggplot2")
library(ggplot2)
# Or: install.packages("tidyverse")
# 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
```

```
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Statistical Thinking

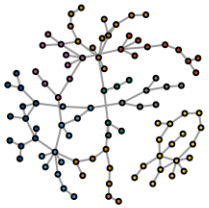
- Pose a research question about diamonds.
- Sketch what type of graphical summary you could use.

🔗 Data Visualisations of the Diamond Data

Data Visualisation

Data Visualisation (Data Viz)

- Data Visualisation is graphical summaries on steriods.
- Examples include infographics, heat maps and networks. See 50 Top Viz in R.



💬 Why is Data Viz such a massive growth industry?

Good Data Visualisations

- Good Data Viz tells an interesting story in a visually appealing way.
- This requires an understanding of both data and design.

For example:

- What question are we trying to answer?
- What variables are we highlighting?
- What does the eye focus on?
- What is the effect of different colours?

Remember: Each plot is a statistic: ie a summary of the data. It is built to be informative and communicate insights.

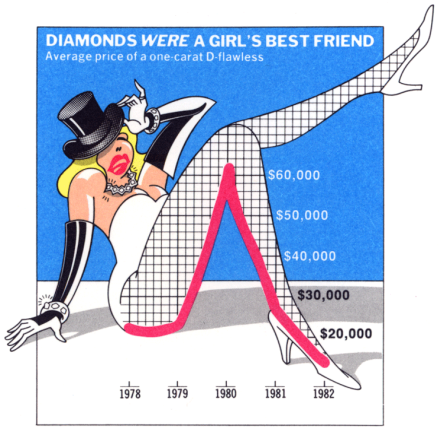
Hans Rosling

New York City's greenhouse gas emissions as one-ton spheres of carbon dioxide gas

Carbon Visuals

Poor and Bad Data Visualisations

- **Poor** Data Viz tells a story in a visually boring way, or a visually distracting way ("chartjunk").



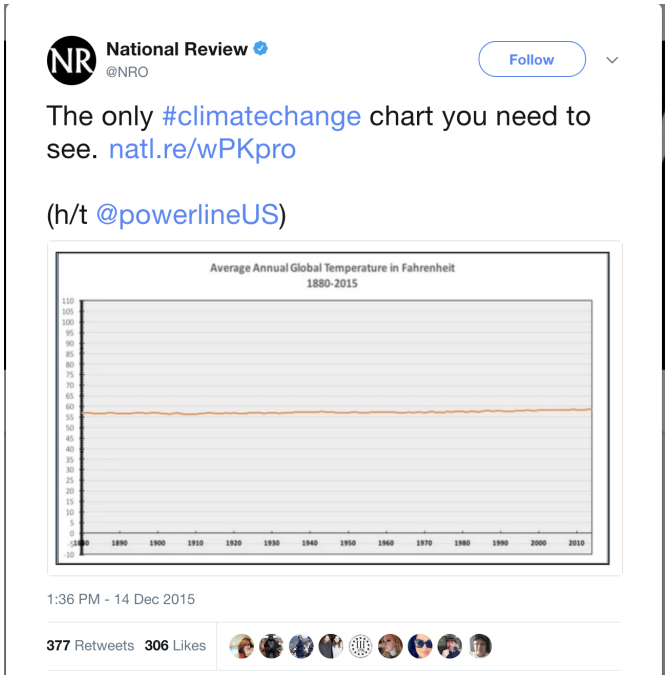
- **Bad** Data Viz tells a misleading story (especially in an appealing way!)

Obama and US economy

💬 What's wrong?

🔗 Obama's 2011 State of the Union Address

Cimate Change

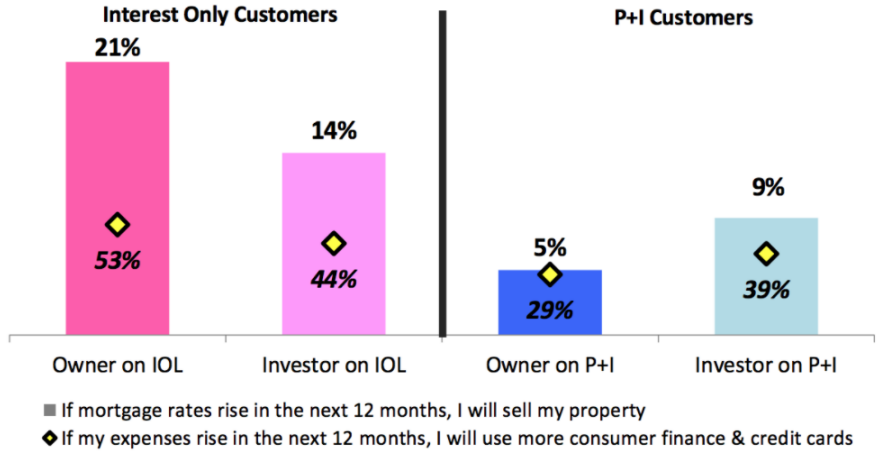


💬 What's misleading?

🐦 Twitter Other examples

Australian mortages

Exhibit 1: Interest-only borrowers are more likely to use credit cards / consumer finance to manage higher costs and more likely to sell the property if rates rise



Source: AlphaWise, Morgan Stanley Research estimates

💬 What's confusing?

🔗 Domain

ggplot2

Introducing ggplot2

- `gglot2` is an intuitive way to produce data visualisations.
  - It's a plotting system based on on "The Grammar of Graphics".
  - That's why it's called `ggplot` (**g**rammar of **g**raphics plot).
  - It's part of the `tidyverse` package.
- ggplot looks more complicated than Base R graphics, but allows much more complex plots with layers.

Building blocks & parameters

- `ggplot` allows you to specify the individual building blocks of your plot, and then combine them to create just about any kind of visualisation you want.
- The **aesthetic** `aes` is “what you can see”.

Eg: position, outside color, inside colour (fill), shape of points, linetype, size.

- The **geometric objects** `geom_xxx` are the actual marks we put on a plot.

Eg: points (`geom_point`), lines (`geom_line`), boxplot (`geom_boxplot`)

- The **facet** is a subset of the data.

🔗 Example

Steps to using ggplot

Step1: Install the package

```
# install.packages("ggplot2")
library(ggplot2)   # Or instead: "tidyverse"
data = diamonds
```

Step2: Check the data is “tidy”

The data needs to be in a data frame, with subjects as rows and variables as columns.

```
head(data,2)
```

```
## # A tibble: 2 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
```

Step3: Check classification of variables

```
str(data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Step4: Sketch by hand

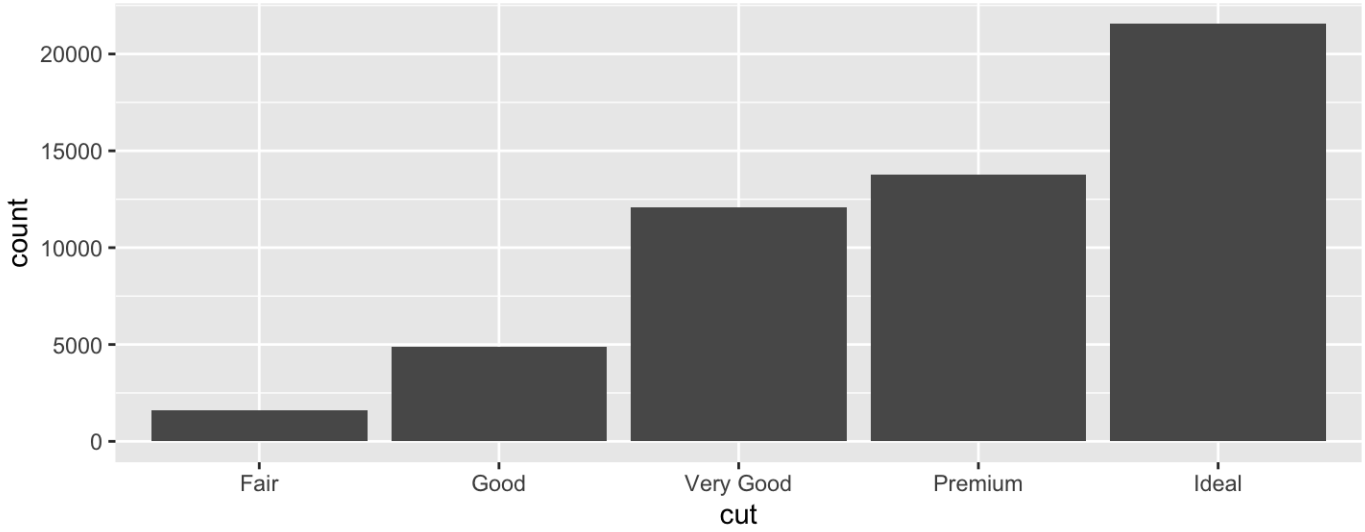
- Sketch by hand what you want to produce, labelling the variables.
- Then write the code.

Step5: Run your ggplot and then customise to improve visual design

Example of ggplot: Barplot

Simple Barplot [1 qual]

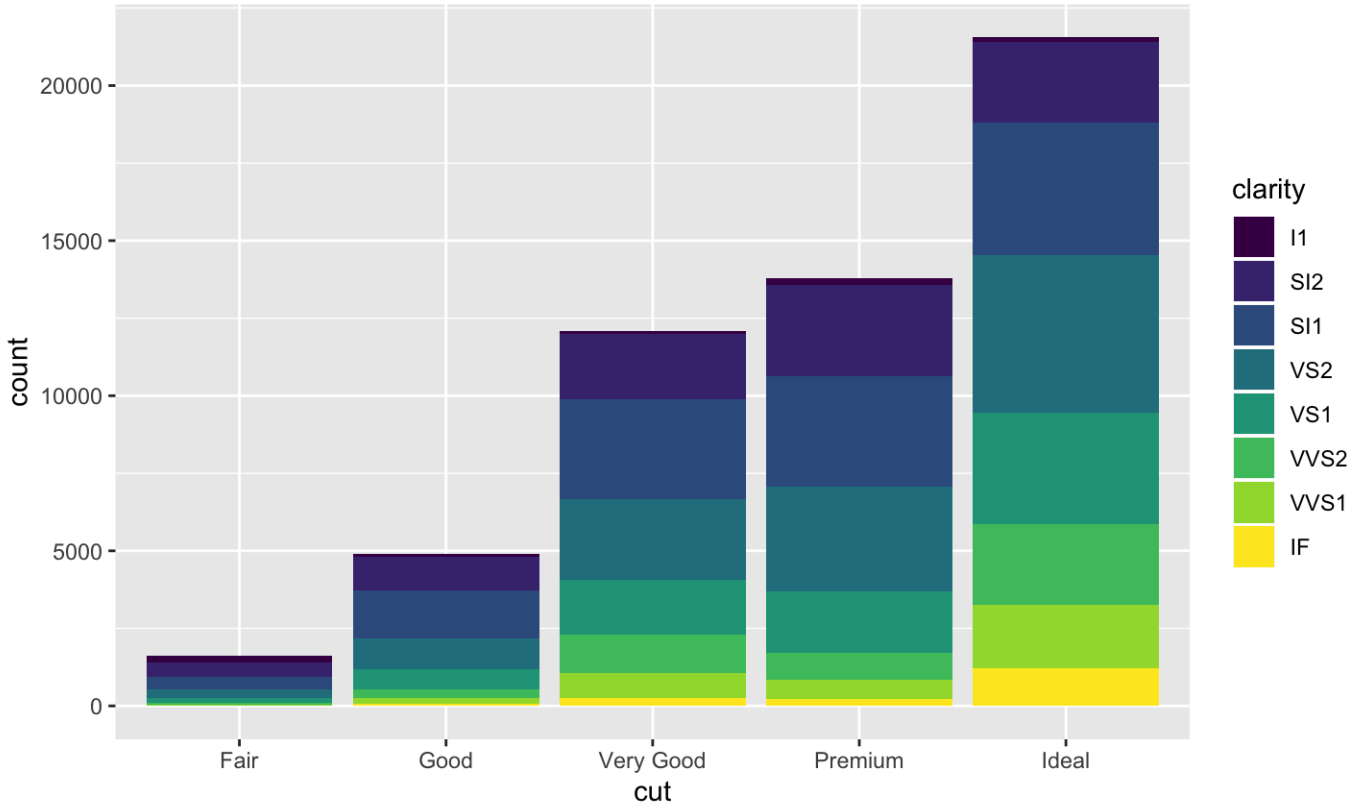
```
p = ggplot(diamonds, aes(x=cut)) # Defines the x axis (1 variable).
p + geom_bar()  # Represents the data by bar chart.
```




```
# Compare in Base R: barplot(table(diamonds$cut))
```

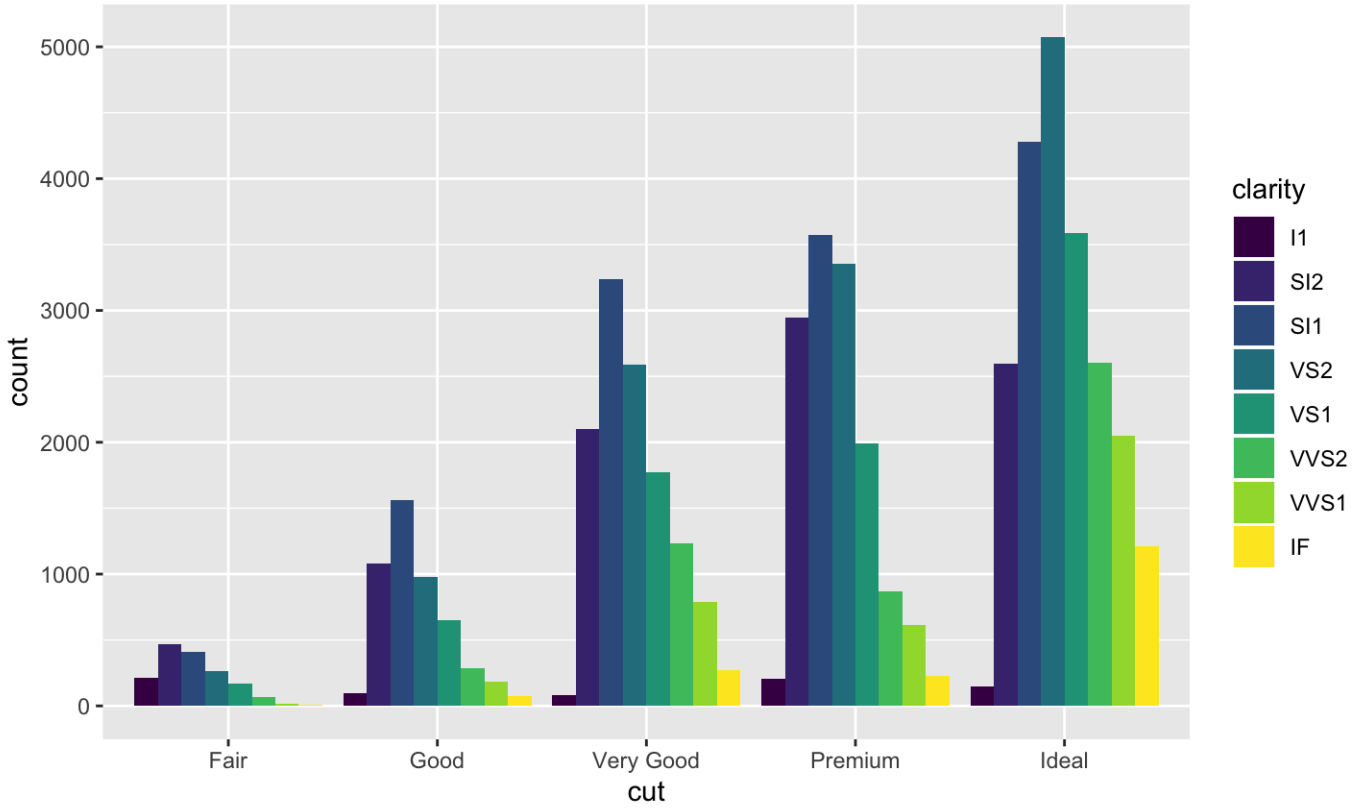
Barplot [1 qual] + aesthetic [1 qual]

```
p + geom_bar(aes(fill=clarity))  # Adds colour by a 2nd variable (clarity)
```



Barplot [1 qual] + aesthetic [1 qual]: Side-by-side

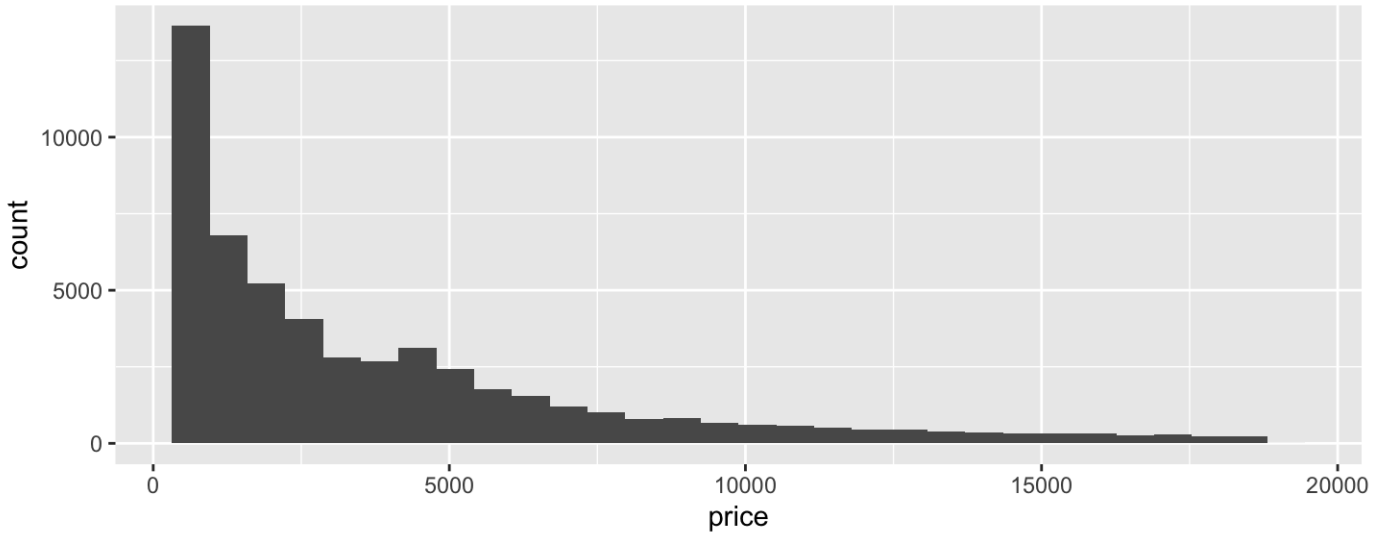
```
p + geom_bar(aes(fill=clarity),position="dodge")
```



More Examples

Histogram [1 quant]

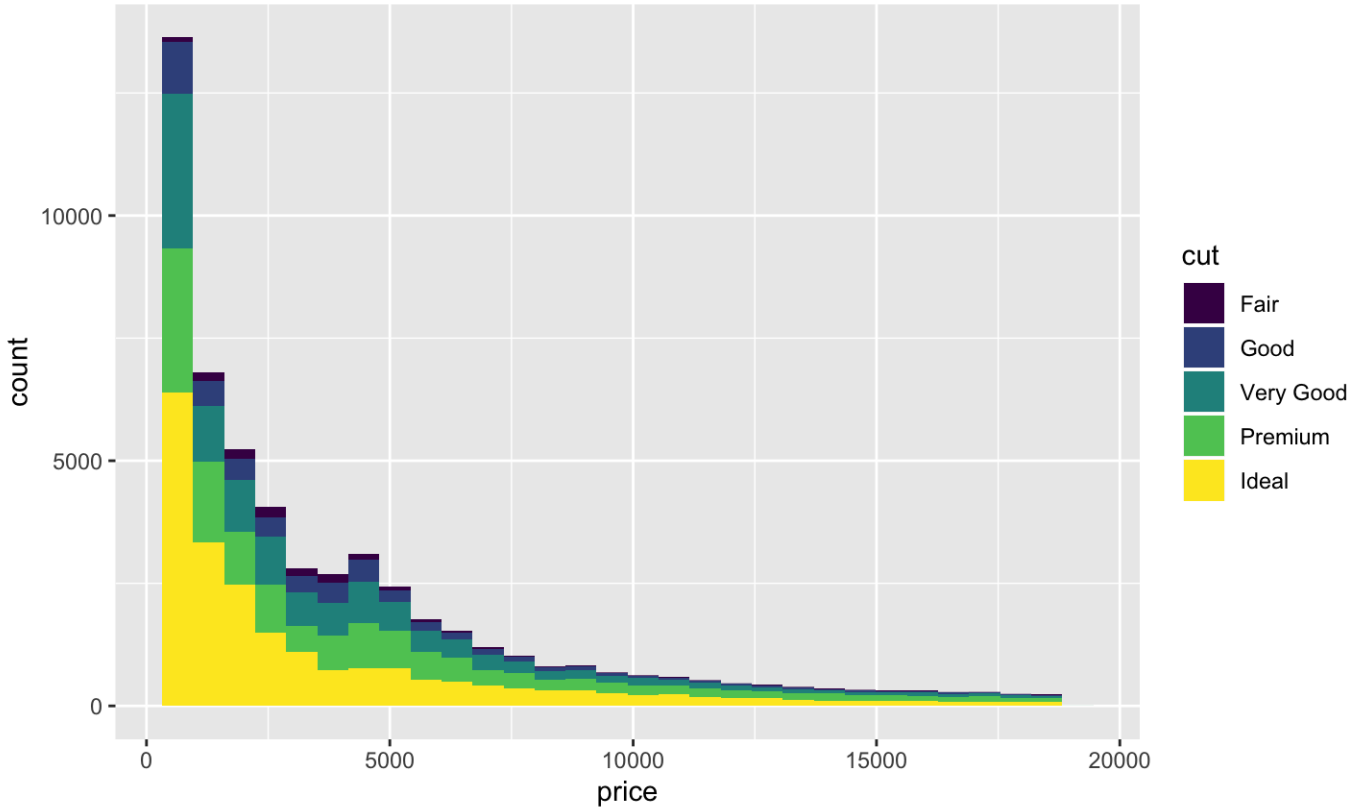
```
p1 = ggplot(diamonds, aes(x=price)) # Defines the x axis (1 variable).
p1 + geom_histogram() # Represents the data by histogram.
```




```
# Compare in Base R: hist(diamonds$price)
```

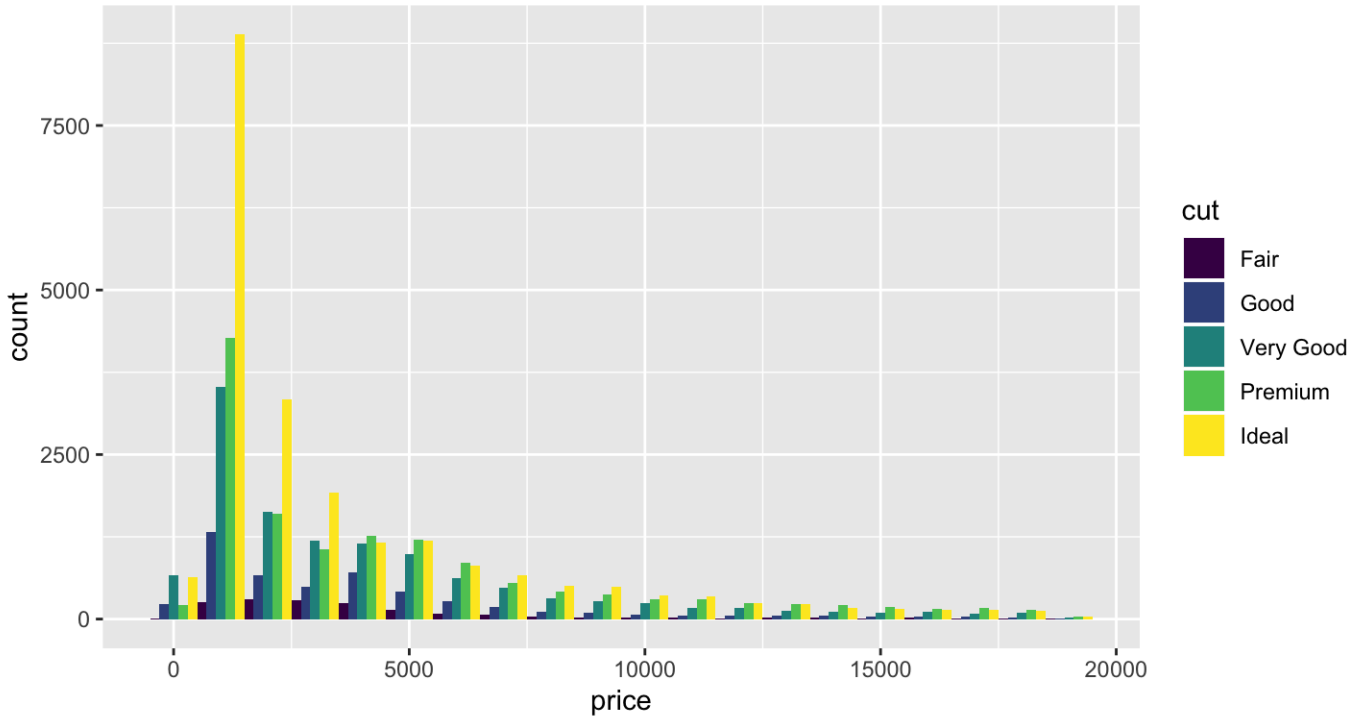
Histogram [1 quant] + aesthetic [1 qual]

```
p1 + geom_histogram(aes(fill=cut)) # Adds colour by a 2nd variable (cut).
```



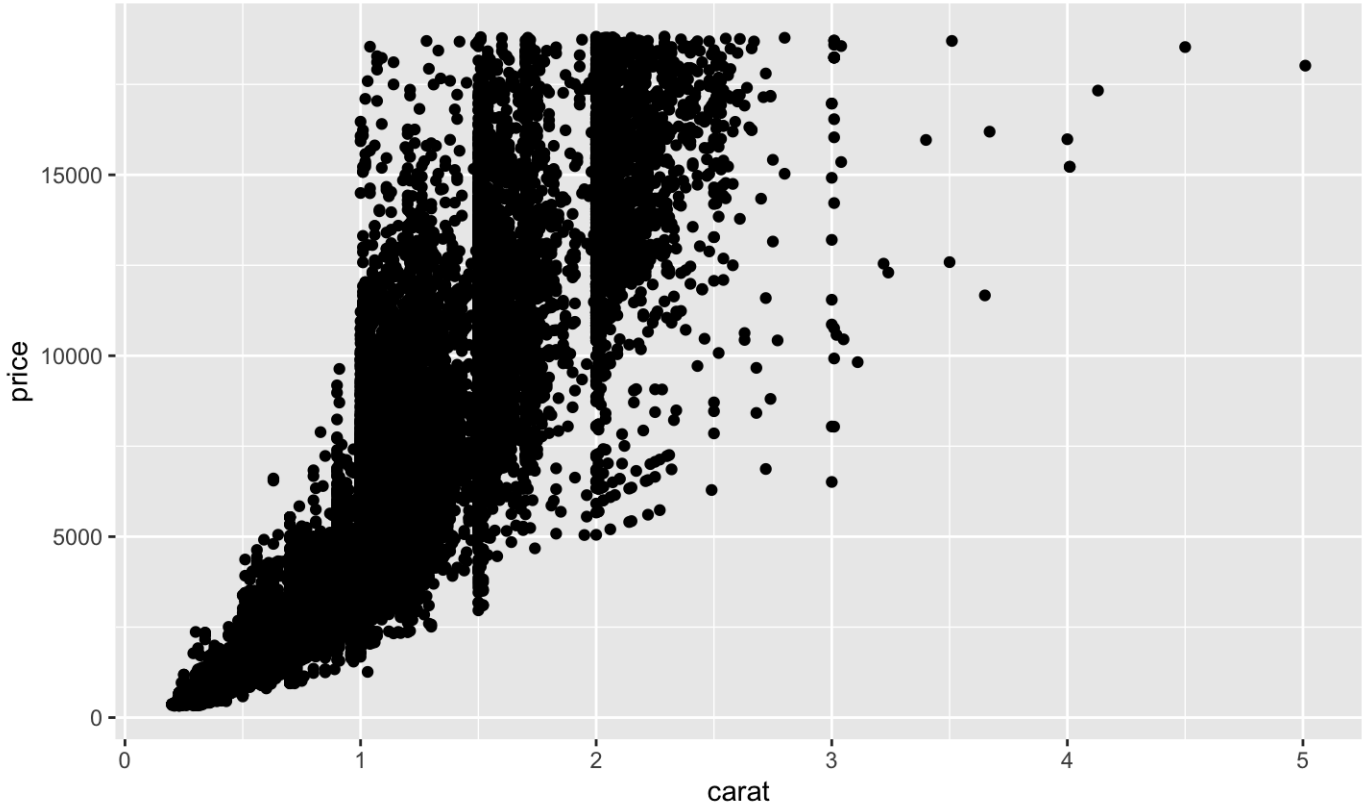
Histogram [1 quant] by aesthetic [1 qual] : side-by-side

```
p1 +  geom_histogram(aes(fill=cut),position = "dodge",binwidth=1000)
```



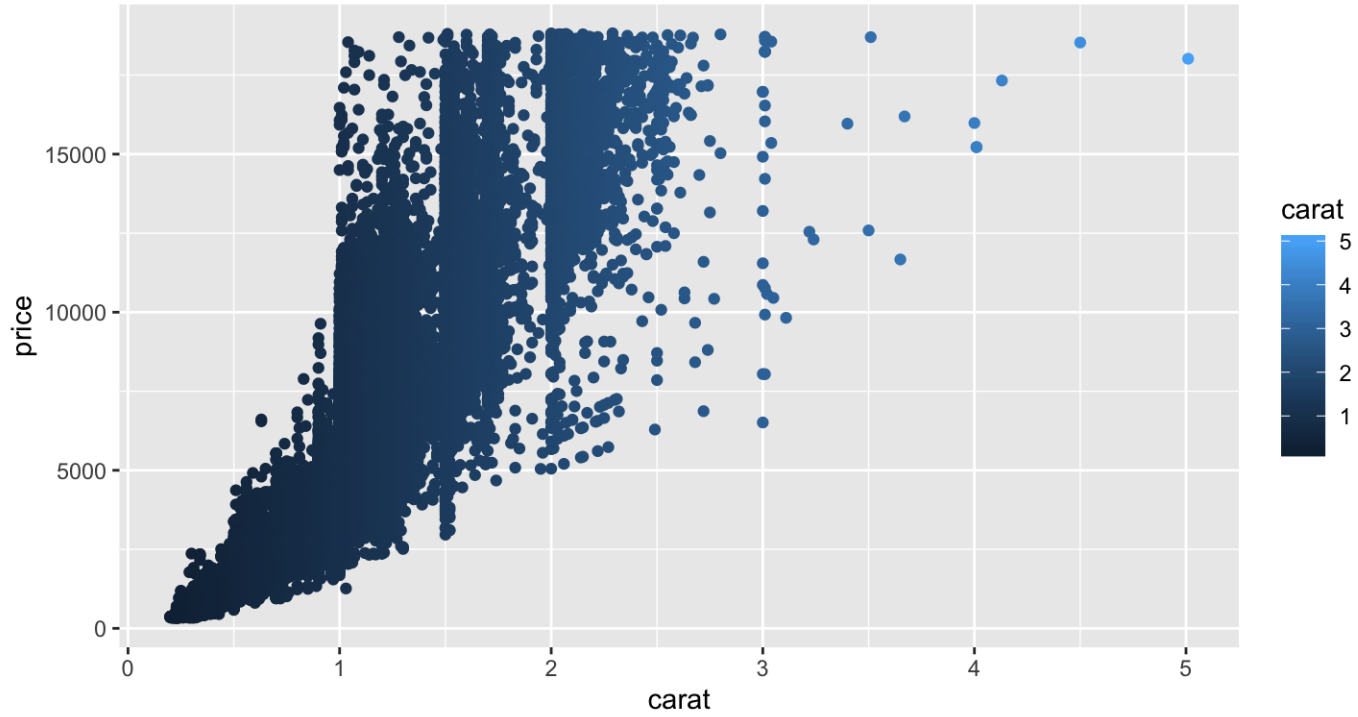
Simple Scatter Plot [2 quant]

```
p2 = ggplot(diamonds, aes(x=carat, y=price)) # Defines the x and y axis (2 variables).
p2 + geom_point()  # Represents the data by points.
```



Scatter Plot [2 quant] + aesthetic [1 quant]

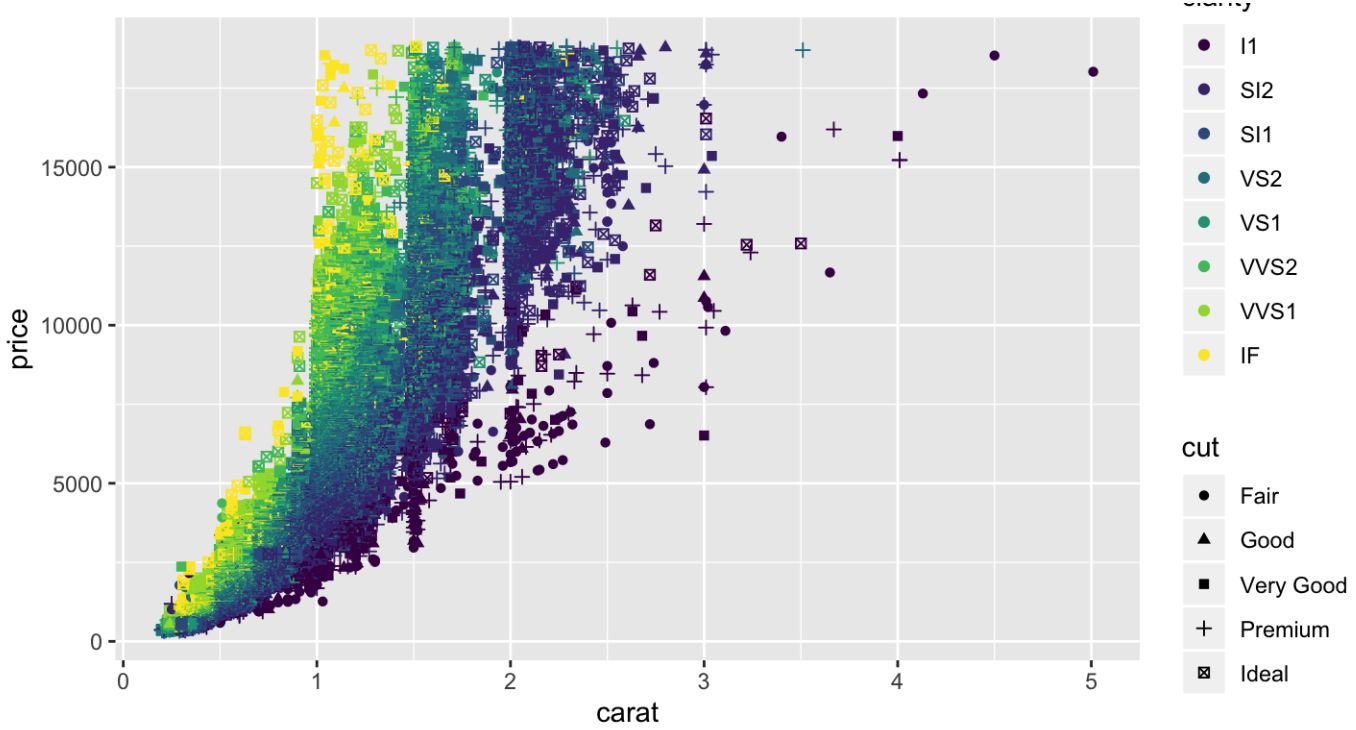
```
p2 + geom_point(aes(colour = carat)) # Adds colour by a 3rd variable (carat).
```



Scatter Plot [2 quant] + aesthetics [2 qual]

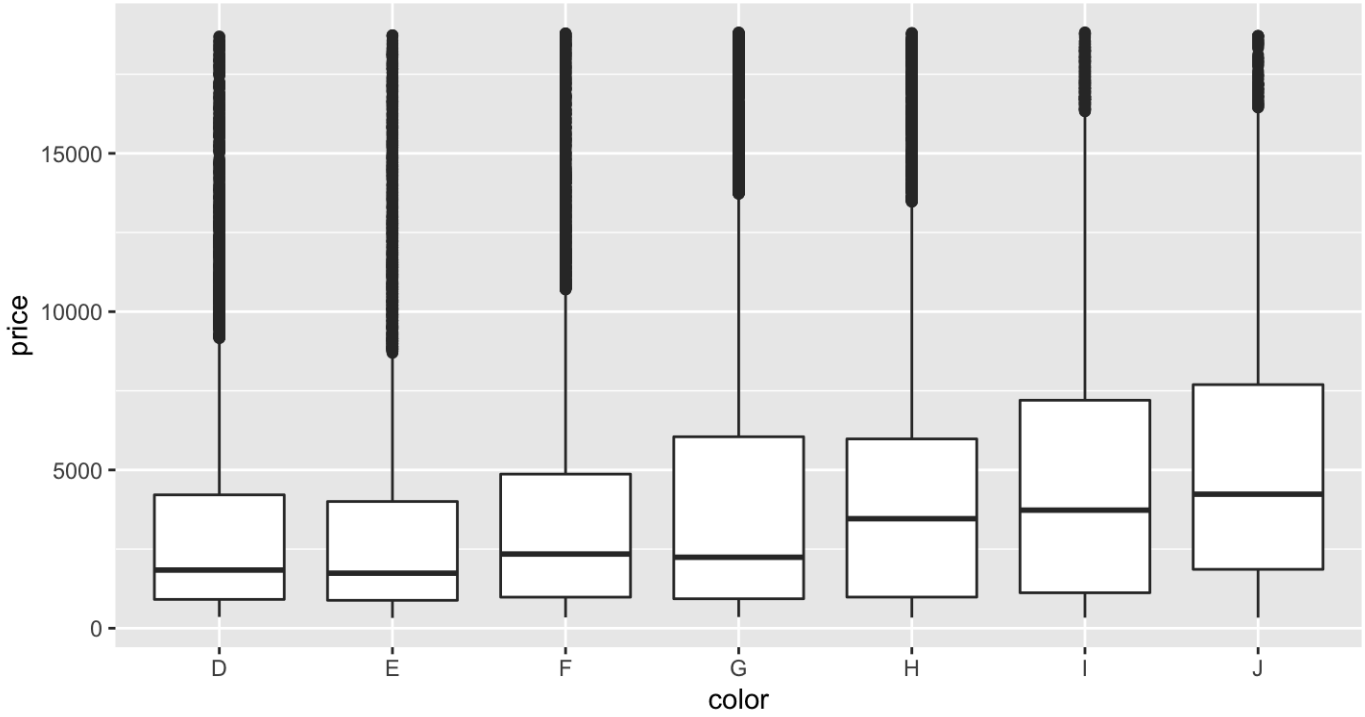
```
p2 + geom_point(aes(color=clarity, shape=cut)) # Adds colour by a 3rd variable (clarity), and shape by a 4th variable (cut).
```

```
## Warning: Using shapes for an ordinal variable is not advised
```



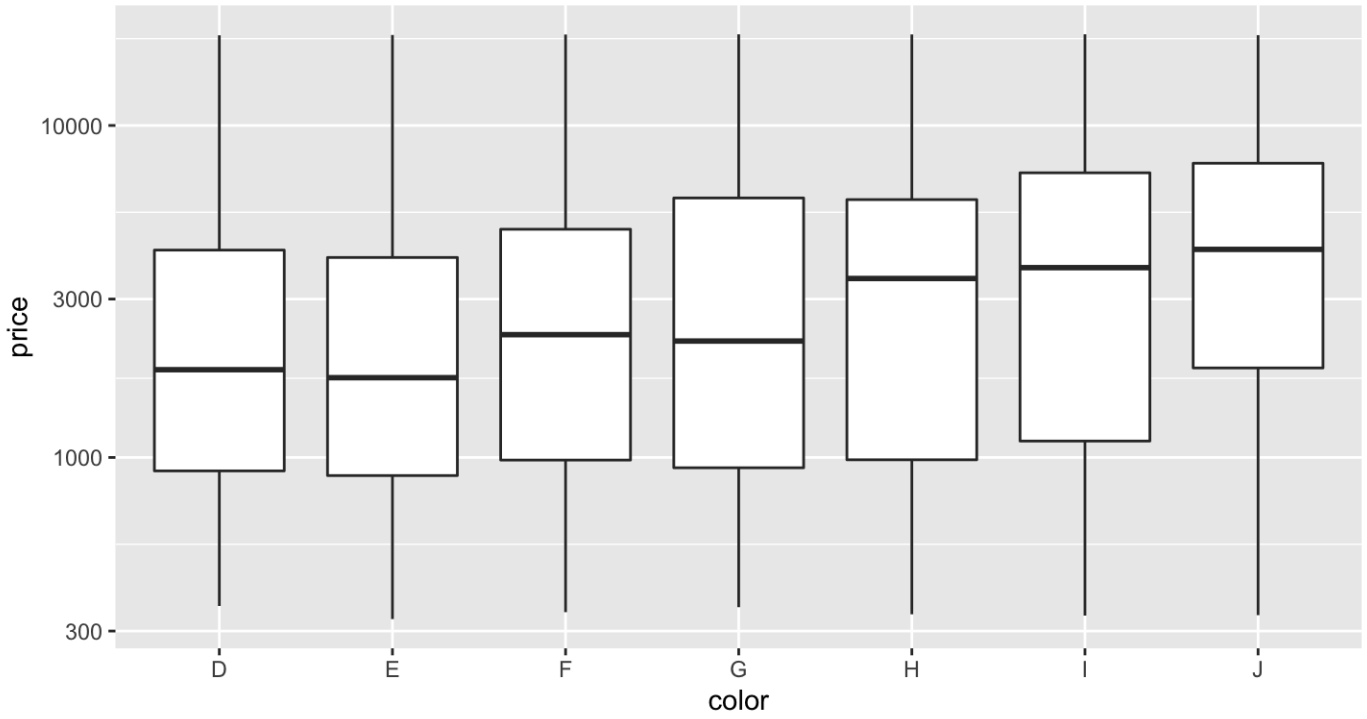
Box plot [1 quant, divided by 1 qual]

```
p3 = ggplot(diamonds, aes(x=color, y=price)) # Defines the x and y axis (2 variables).
p3 + geom_boxplot() # Represents the data by box plot.
```



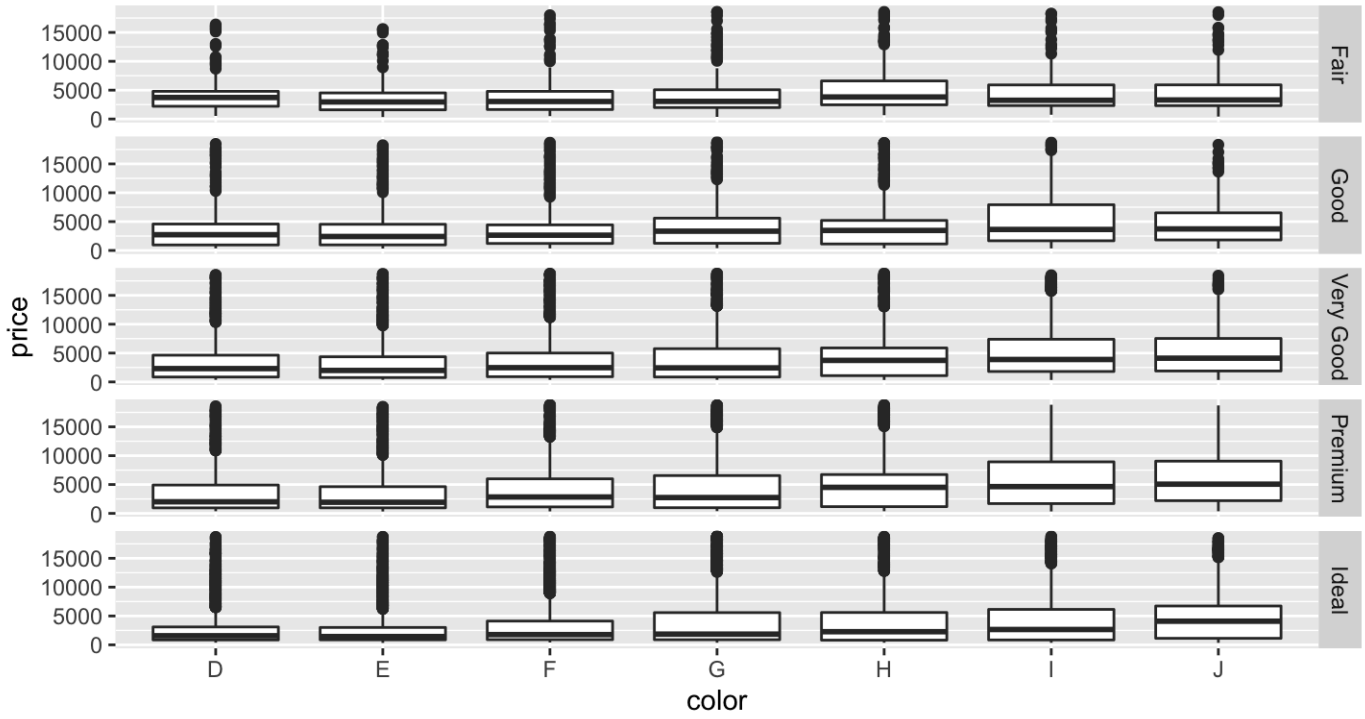
Box plot [1 quant, divided by 1 qual] with scaled y axis

```
p3 + geom_boxplot() + scale_y_log10() # Changes scale of y axis
```



Box plot [1 quant, divided by 1 qual], with facet [1 qual]

```
p3 + geom_boxplot() + facet_grid(cut~.)
```



Have a go

Statistical Thinking

Write down the base R and ggplot code for:

- a bar plot of clarity filled by colour.
- a scatterplot of price vs length, with depth indicated by colour.

Summary

Homework

Run all the code in this lecture yourself.

Key Words

Data visualisation, aesthetics, facets

Further Thinking

- ggplot2 Cheat Sheet
- Hadley Wickham
- Liz Sander
- Selva Prabhakaran