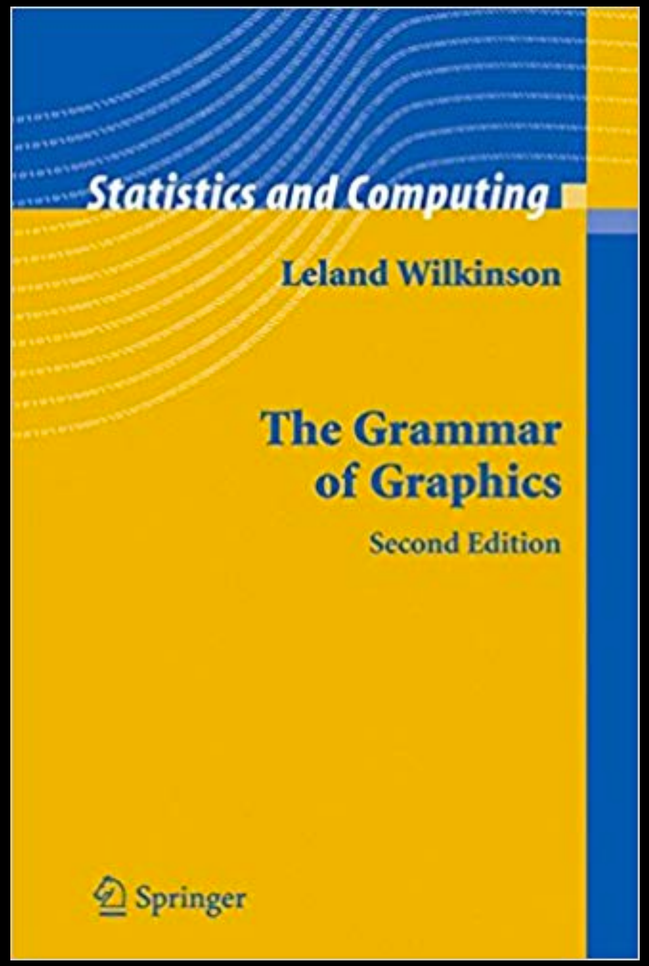
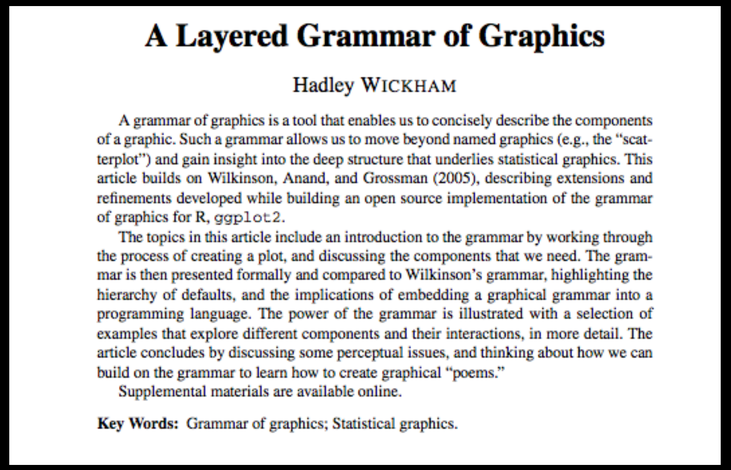
The Grammar of Graphics

Darya Vanichkina

… a personal commentary/interpretation of …





A Layered Grammar of Graphics

Hadley WICKHAM

A grammar of graphics is a tool that enables us to concisely describe the components of a graphic. Such a grammar allows us to move beyond named graphics (e.g., the "scatterplot") and gain insight into the deep structure that underlies statistical graphics. This article builds on Wilkinson, Anand, and Grossman (2005), describing extensions and refinements developed while building an open source implementation of the grammar of graphics for R, `ggplot2`.

The topics in this article include an introduction to the grammar by working through the process of creating a plot, and discussing the components that we need. The grammar is then presented formally and compared to Wilkinson's grammar, highlighting the hierarchy of defaults, and the implications of embedding a graphical grammar into a programming language. The power of the grammar is illustrated with a selection of examples that explore different components and their interactions, in more detail. The article concludes by discussing some perceptual issues, and thinking about how we can build on the grammar to learn how to create graphical "poems."

Supplemental materials are available online.

Key Words: Grammar of graphics; Statistical graphics.

http://vita.had.co.nz/papers/layered-grammar.html

Grammar

“the fundamental principles or rules of an art or science”

(OED Online 1989)

Simple case

x	*y*	Shape
2	4	a
1	1	a
4	15	b
9	80	b

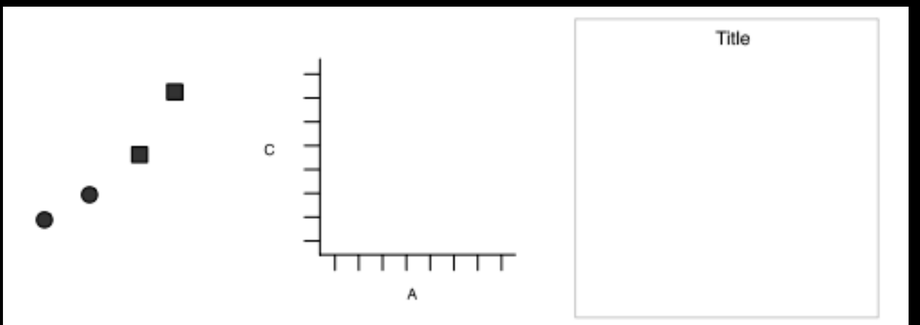


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

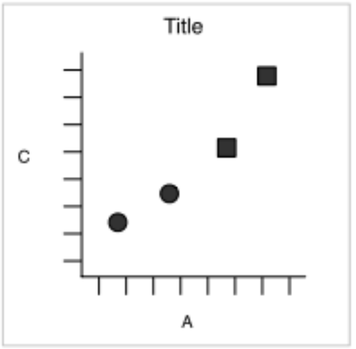


Figure 2. The final graphic, produced by combining the pieces in Figure 1.

Simple case

x	y	Shape
2	4	a
1	1	a
4	15	b
9	80	b

- a default dataset and set of mappings from variables to aesthetics,
- one or more layers, with each layer having one geometric object, one statistical transformation, one position adjustment, and optionally, one dataset and set of aesthetic mappings,
- one scale for each aesthetic mapping used,
- a coordinate system,
- the facet specification.

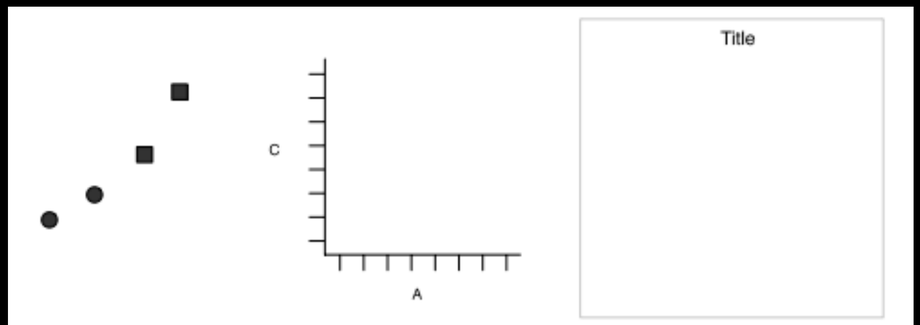


Figure 1. Graphics objects produced by (from left to right): geometric objects, scales and coordinate system, plot annotations.

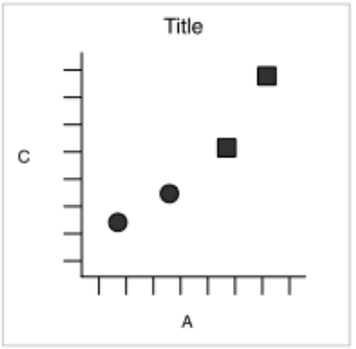


Figure 2. The final graphic, produced by combining the pieces in Figure 1.

Components of the grammar

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

Adapted from DataCamp

Exploring the diamonds data

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

```
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 ...
```

Relationship between table (width of top of diamond relative to widest point) and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

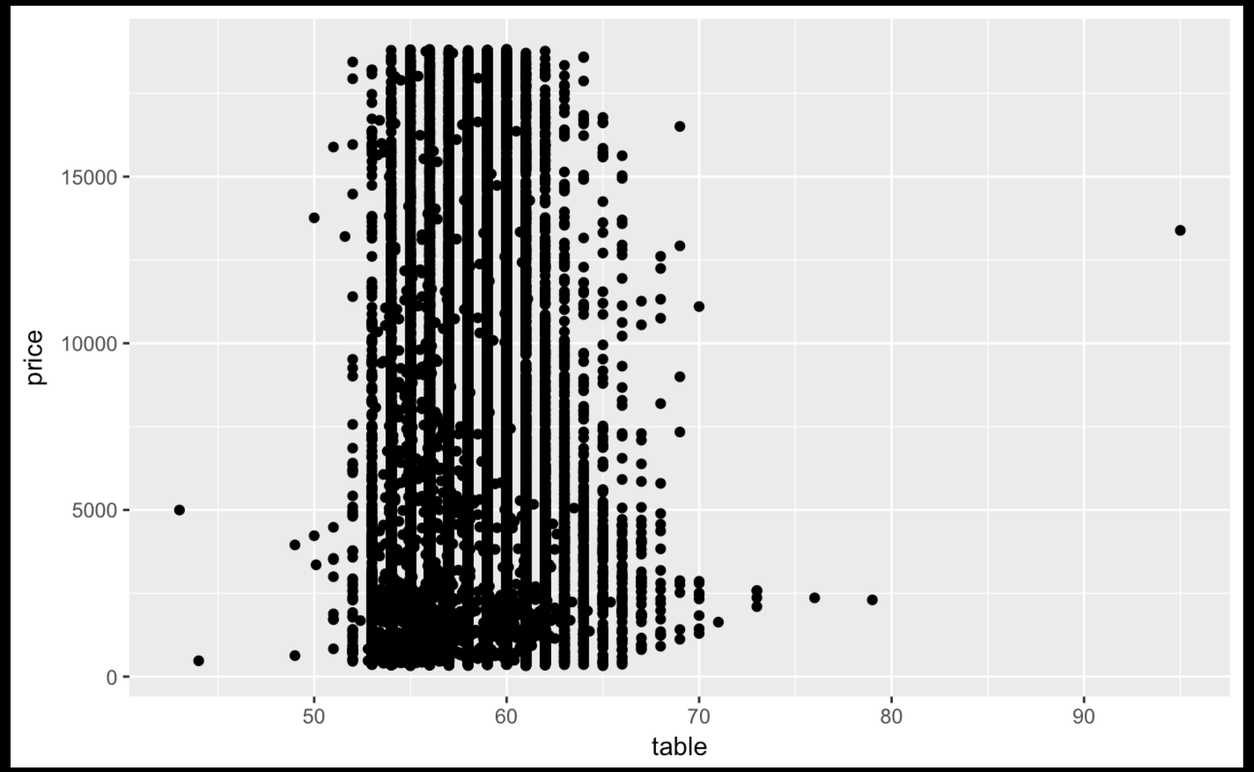
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Relationship between carat (weight) and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

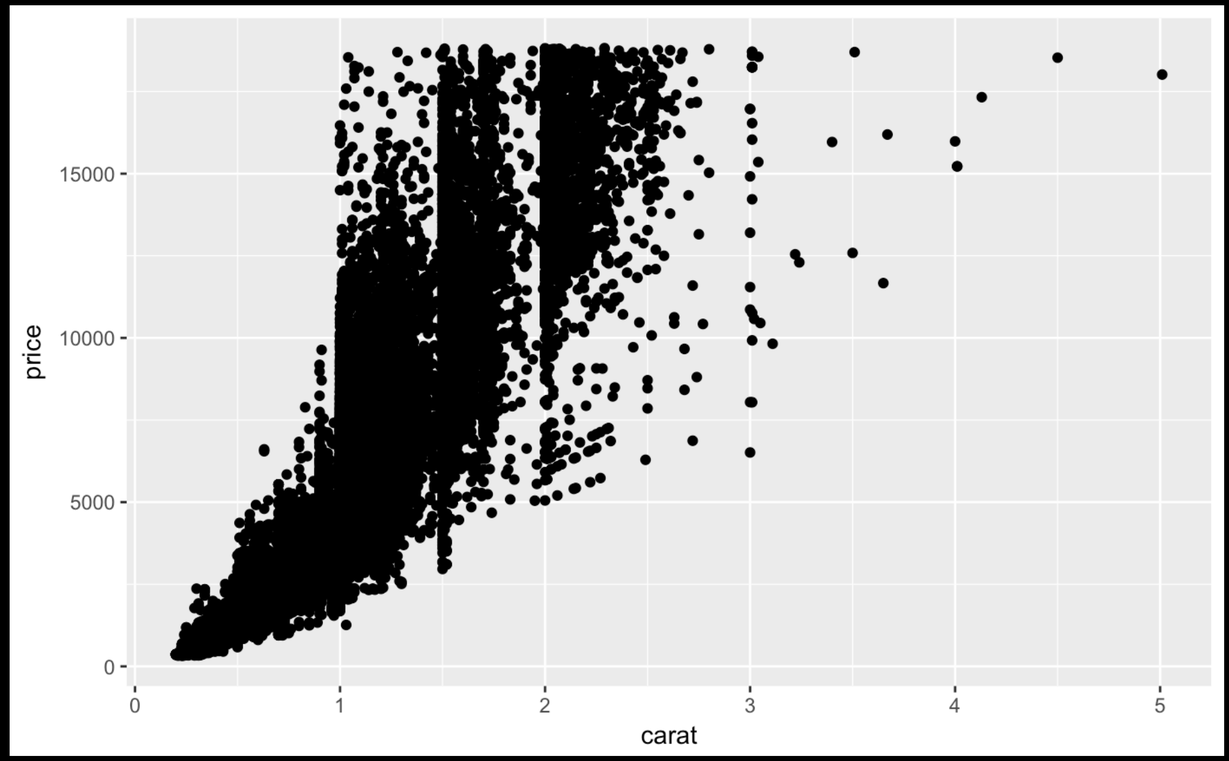
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Relationship between carat (weight) and price? [How many diamonds?]

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

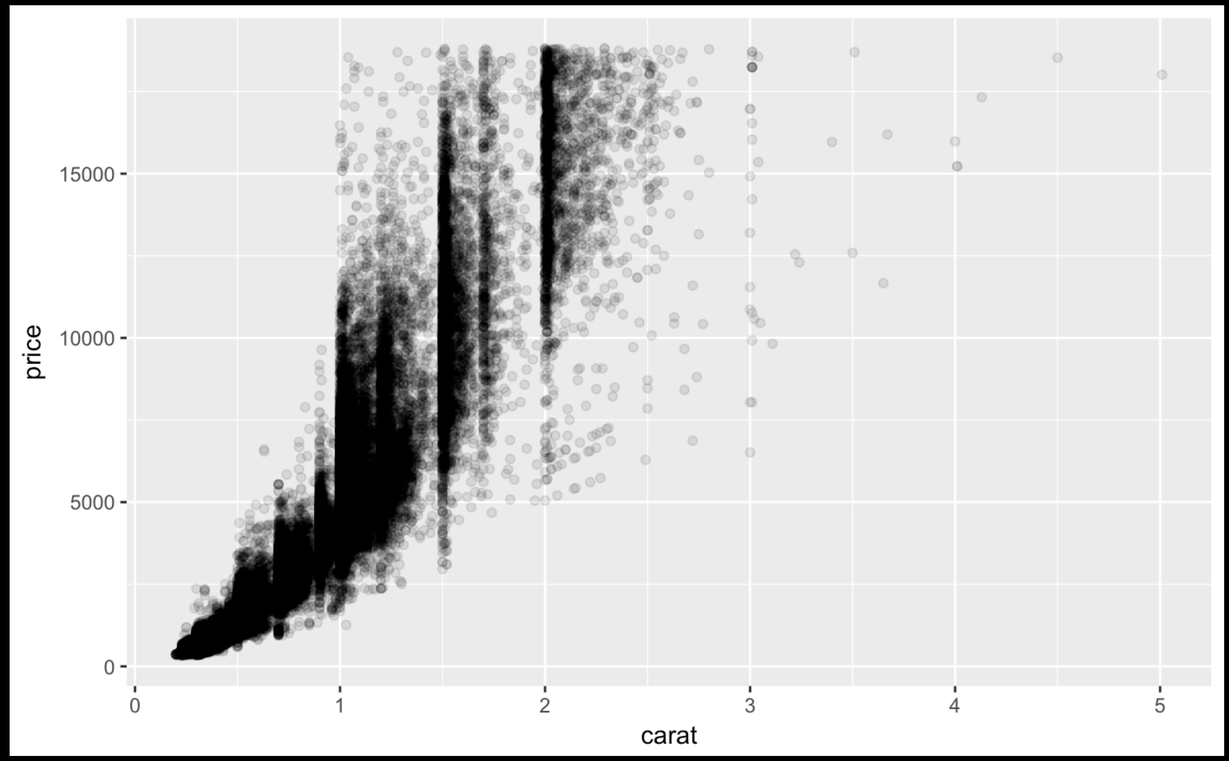
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Relationship between carat (weight) and price? [Lots of small diamonds!]

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

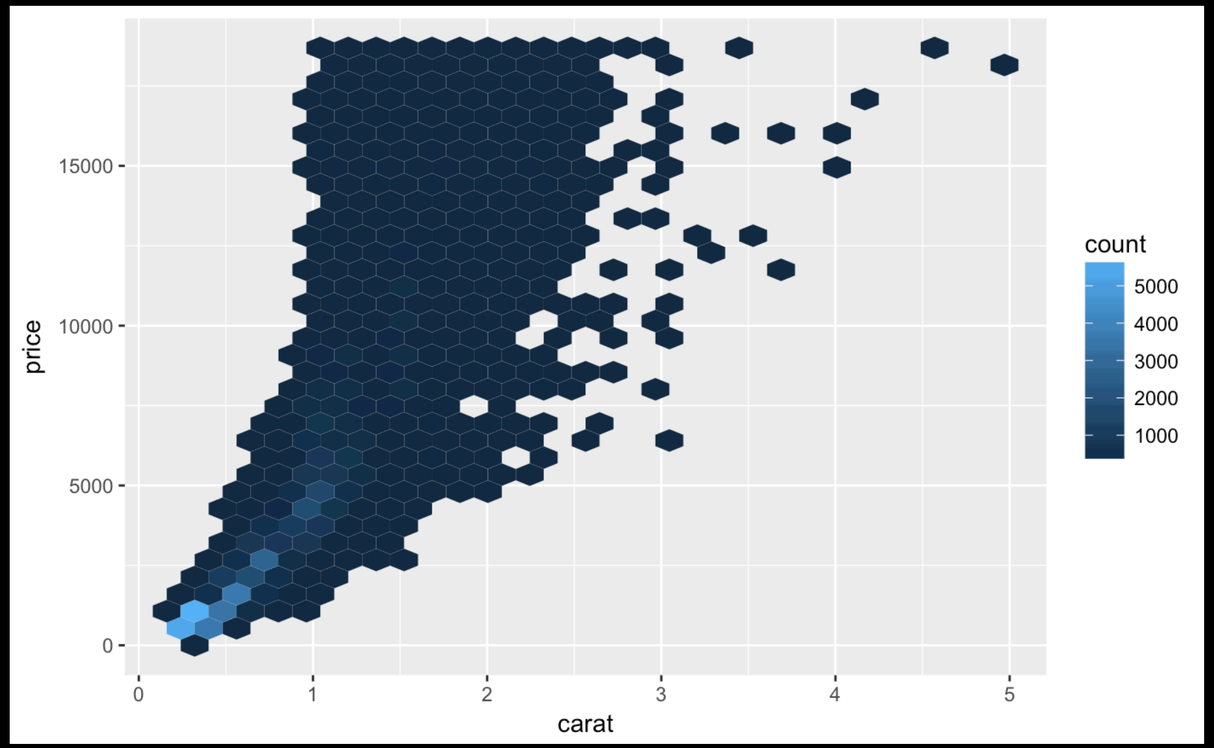
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Components of the grammar

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

Adapted from DataCamp

Relationship between clarity and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

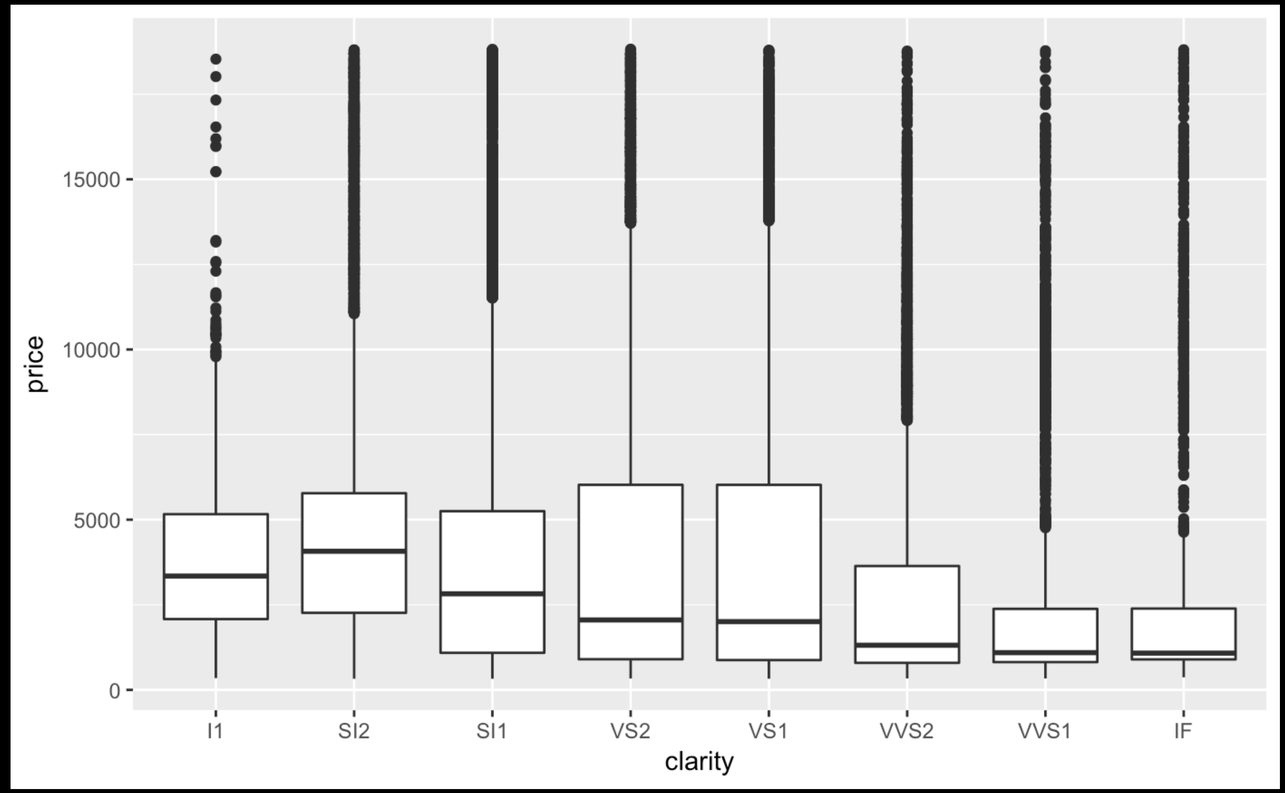
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

```
diamonds %>% ggplot(aes(x = clarity, y = price)) +
 geom_boxplot()
```



Worse ——————————————————> Better

How many diamonds with each clarity?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

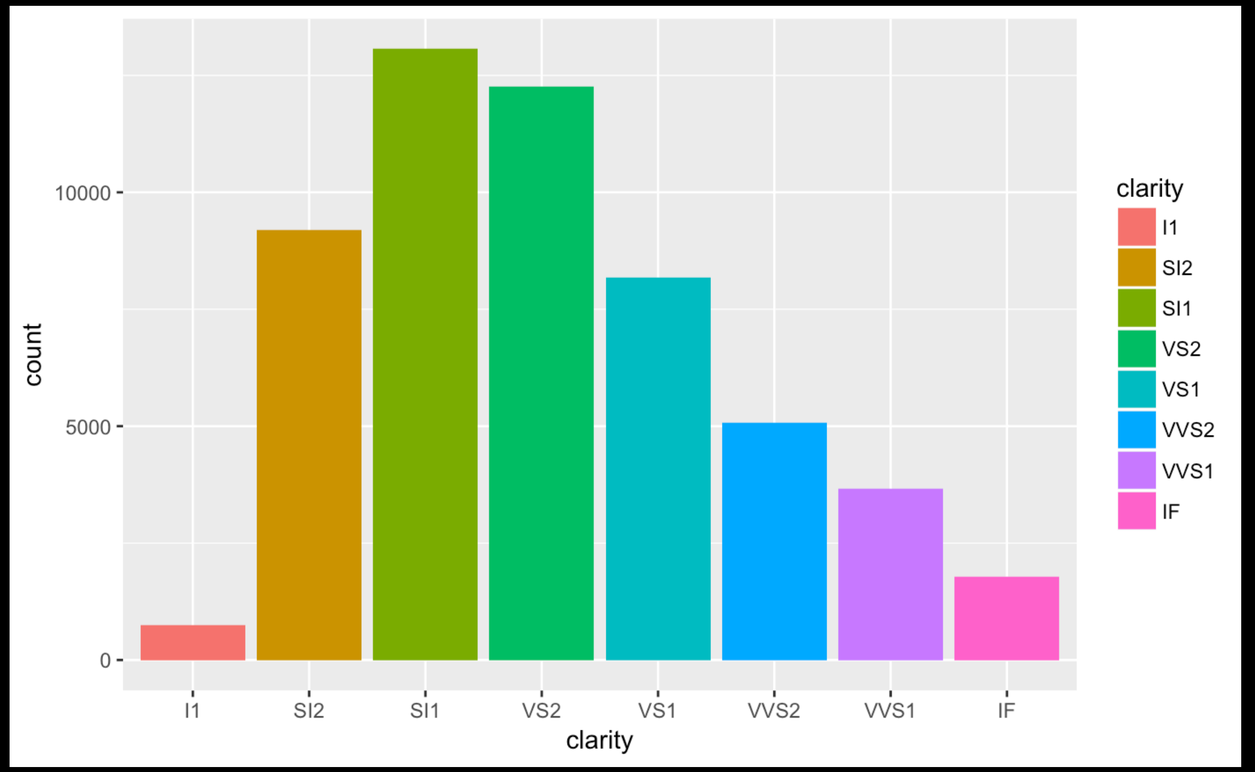
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

```
diamonds %>% ggplot(aes(x = clarity, fill = clarity)) +
 geom_bar()
```



Worse ——————————————————> Better

Relationship between cut and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

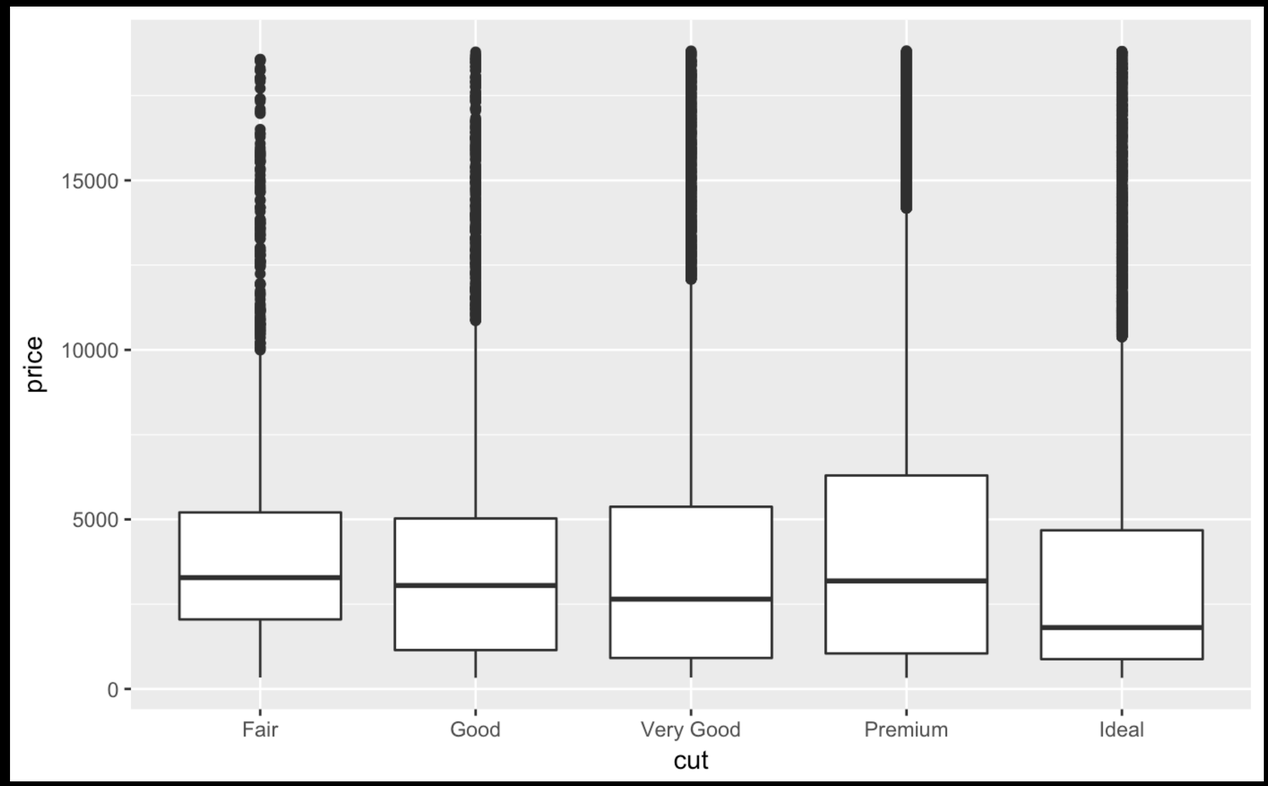
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Worse ——————————————————> Better

Relationship between cut and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

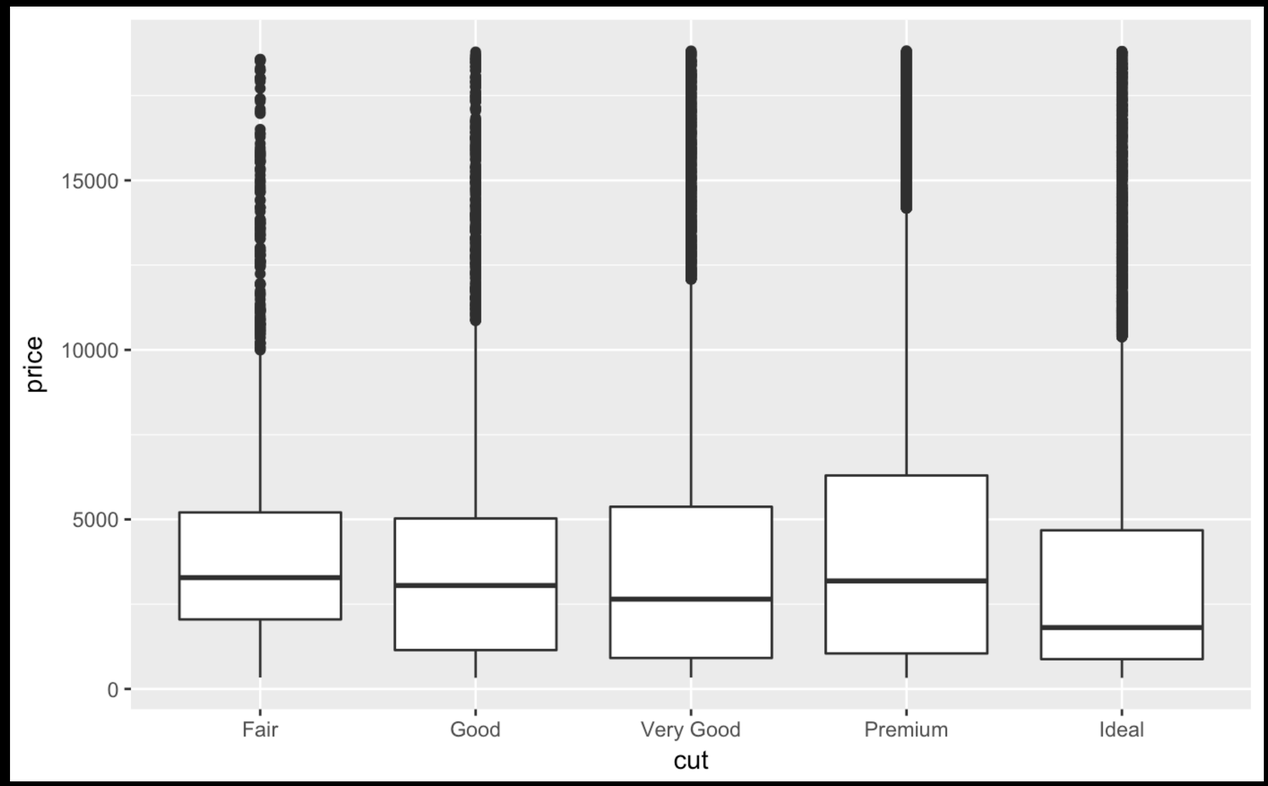
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

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



Worse ——————————————————> Better

Relationship between cut and price?

Prices of 50,000 round cut diamonds

Source: `R/data.R`

A dataset containing the prices and other attributes of almost 54,000 diamonds. The variables are as follows:

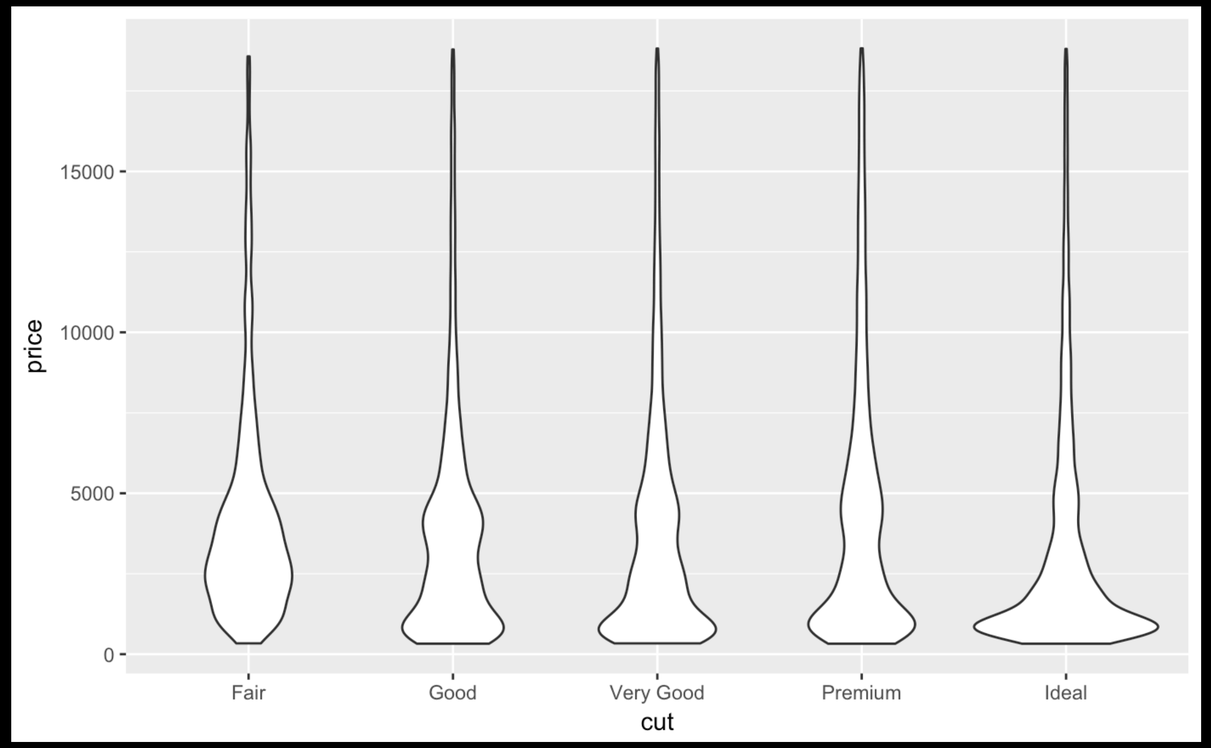
```
diamonds
```

Format

A data frame with 53940 rows and 10 variables:

price	price in US dollars (\$326--\$18,823)
carat	weight of the diamond (0.2--5.01)
cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x	length in mm (0--10.74)
y	width in mm (0--58.9)
z	depth in mm (0--31.8)
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
table	width of top of diamond relative to widest point (43--95)

```
diamonds %>% ggplot(aes(x = cut, y = price)) +
 geom_violin()
```



Worse ——————————————————> Better

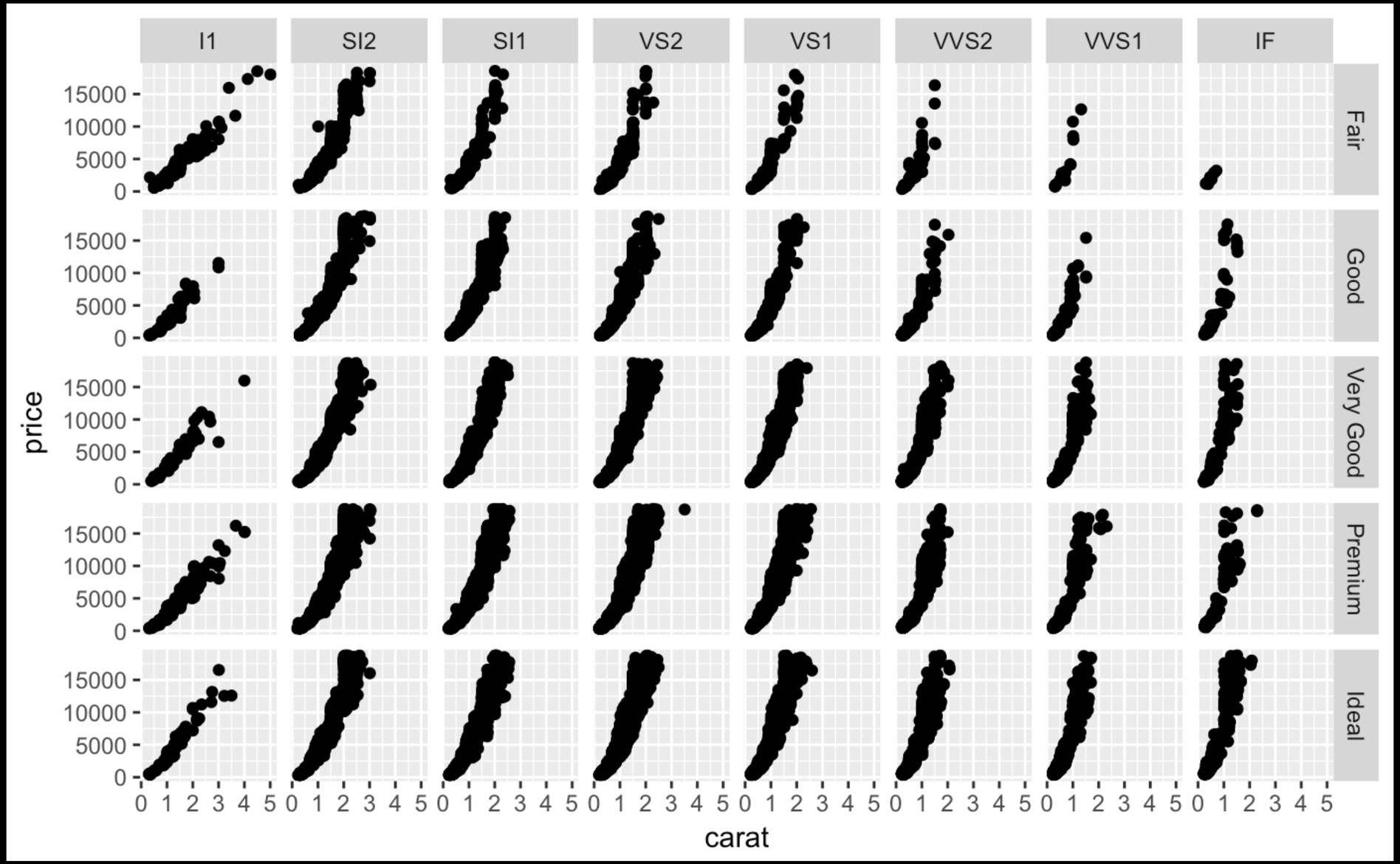
Components of the grammar

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

Adapted from DataCamp

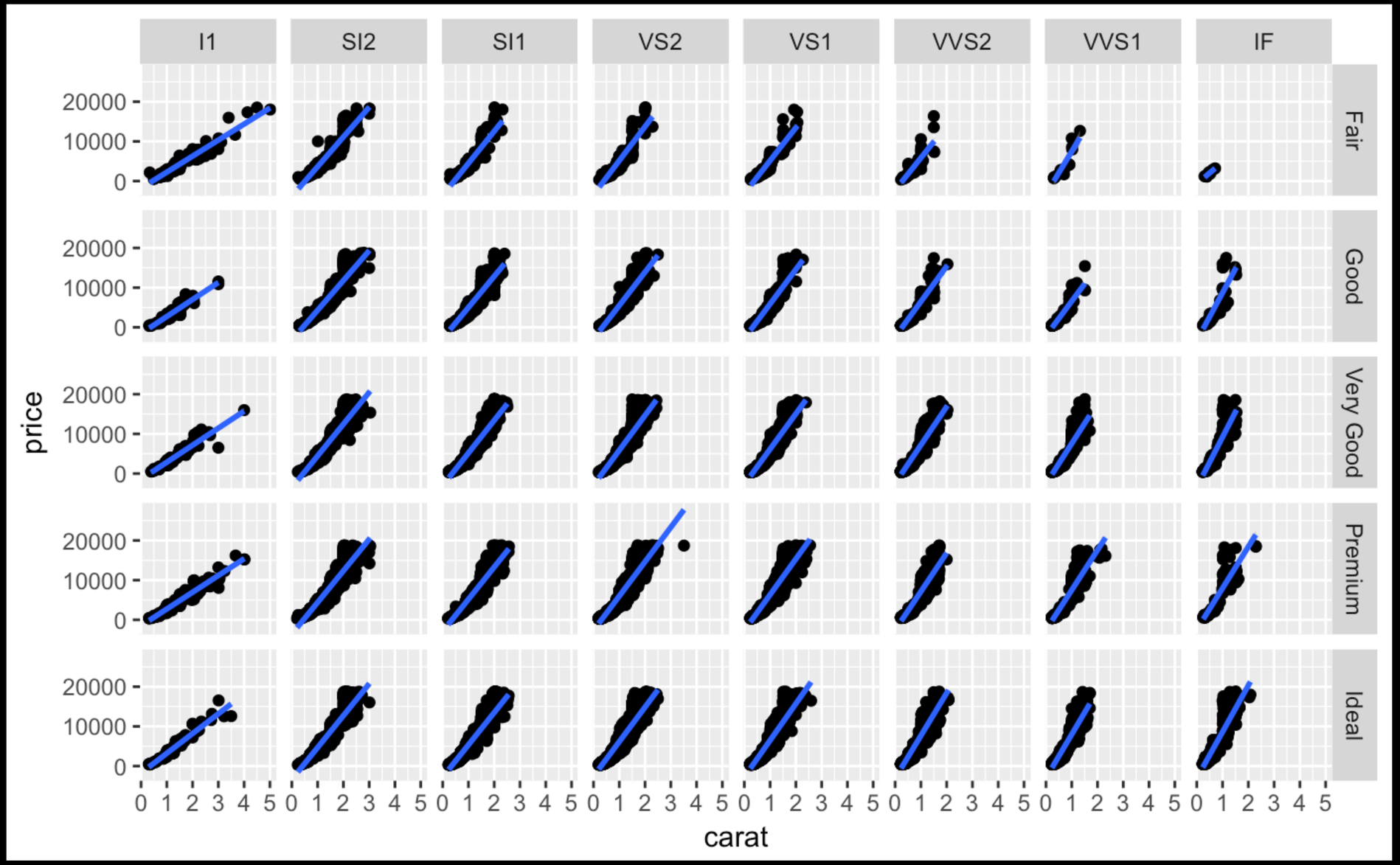
Relationship between price and carat grouped by cut and clarity?

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



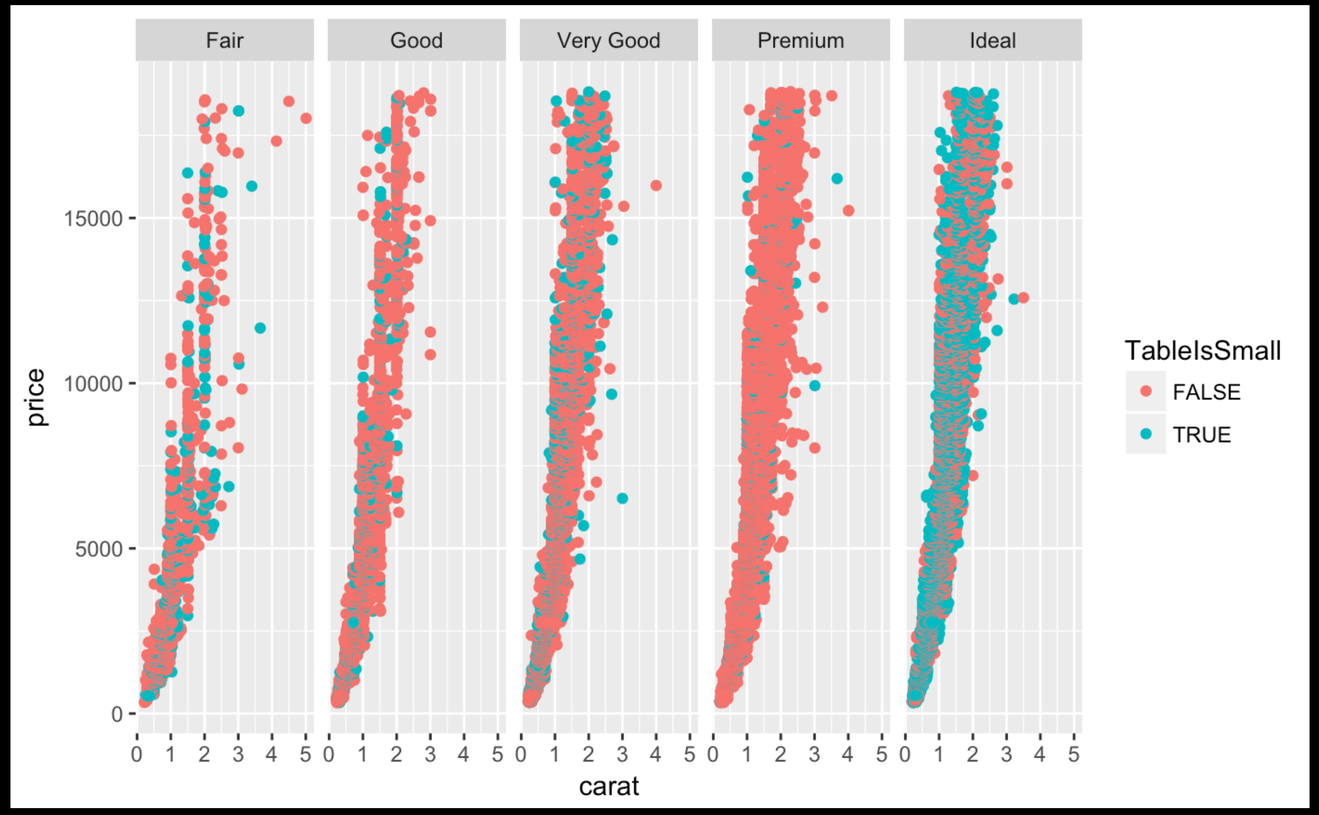
Relationship between price and carat grouped by cut and clarity (with trend)?

```
diamonds %>%  ggplot(aes(x = carat, y = price)) + geom_point() + facet_grid(cut~clarity) + geom_smooth(method =
"lm")
```



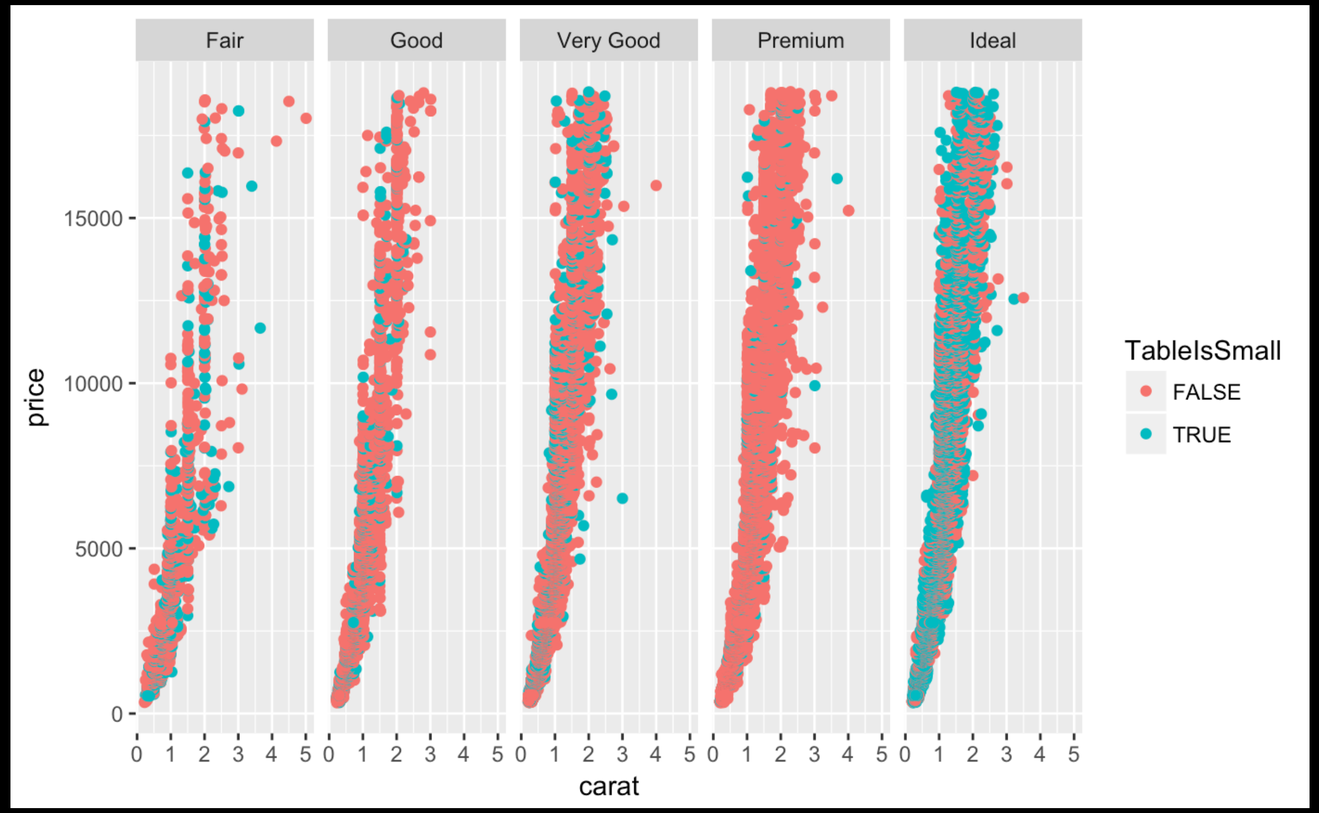
Relationship between table parameter (width of top of diamond relative to widest point, converted on the fly into a factor) and price, faceted by cut?

```
diamonds %>%
  mutate(TableIsSmall = factor(table < median(table)))%>%
  ggplot(aes(x = carat, y = price, colour = TableIsSmall)) + geom_point() + facet_grid(.~cut)
```



Relationship between table parameter (width of top of diamond relative to widest point, converted on the fly into a factor) and price, faceted by cut?

```
diamonds %>%
  mutate(TableIsSmall = factor(table < median(table)))%>%
  ggplot(aes(x = carat, y = price, colour = TableIsSmall)) + geom_point() + facet_grid(.~cut)
```



Most diamonds with a small table (TableIsBig == FALSE) have an “Ideal” cut…

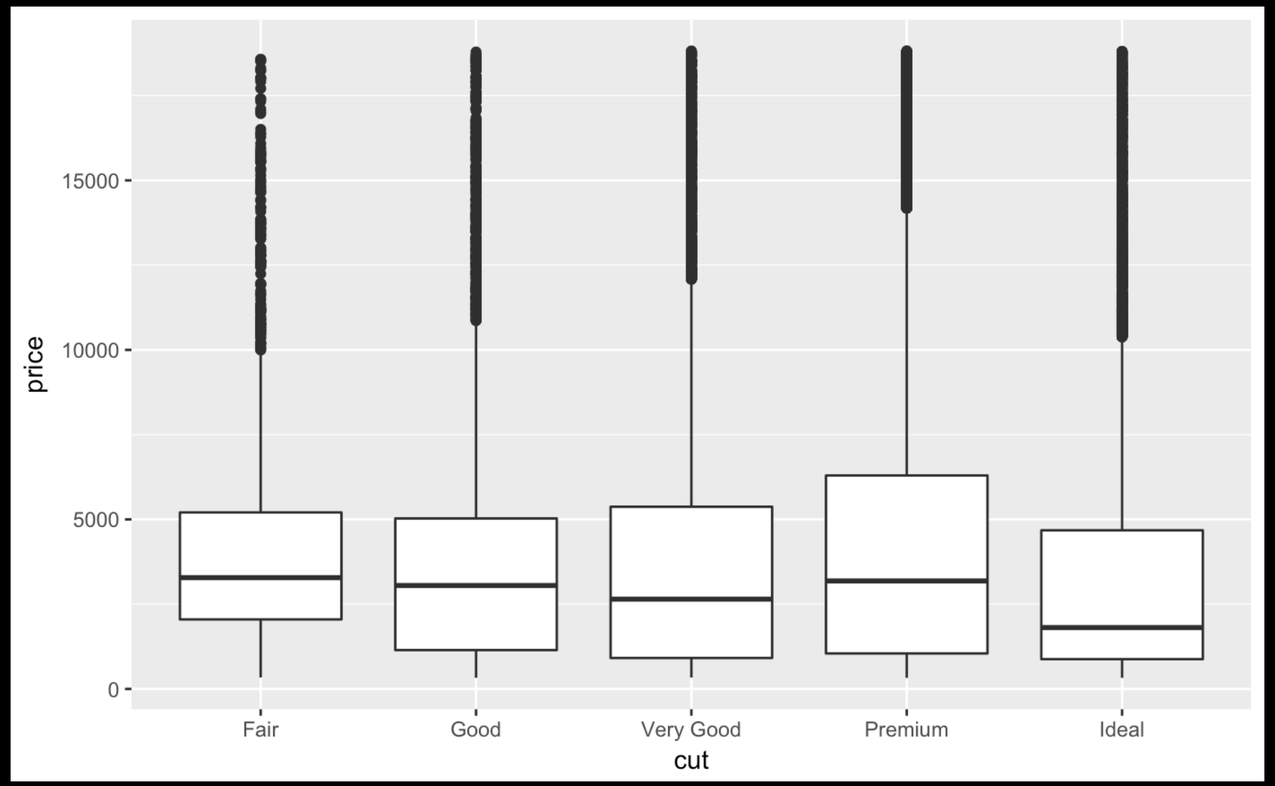
Components of the grammar

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

Adapted from DataCamp

Relationship between cut and price?

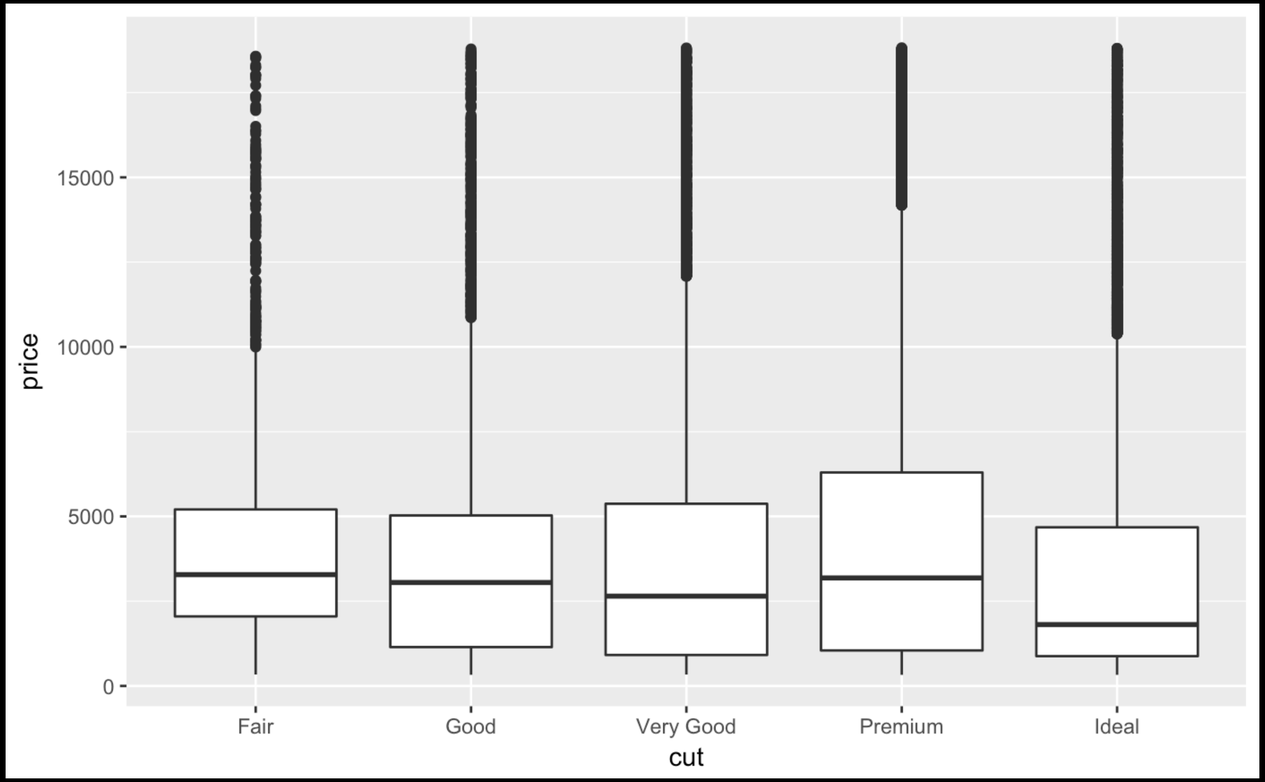
```
diamonds %>% ggplot(aes(x = cut, y = price)) +
 geom_boxplot()
```



Worse ——————————————————> Better

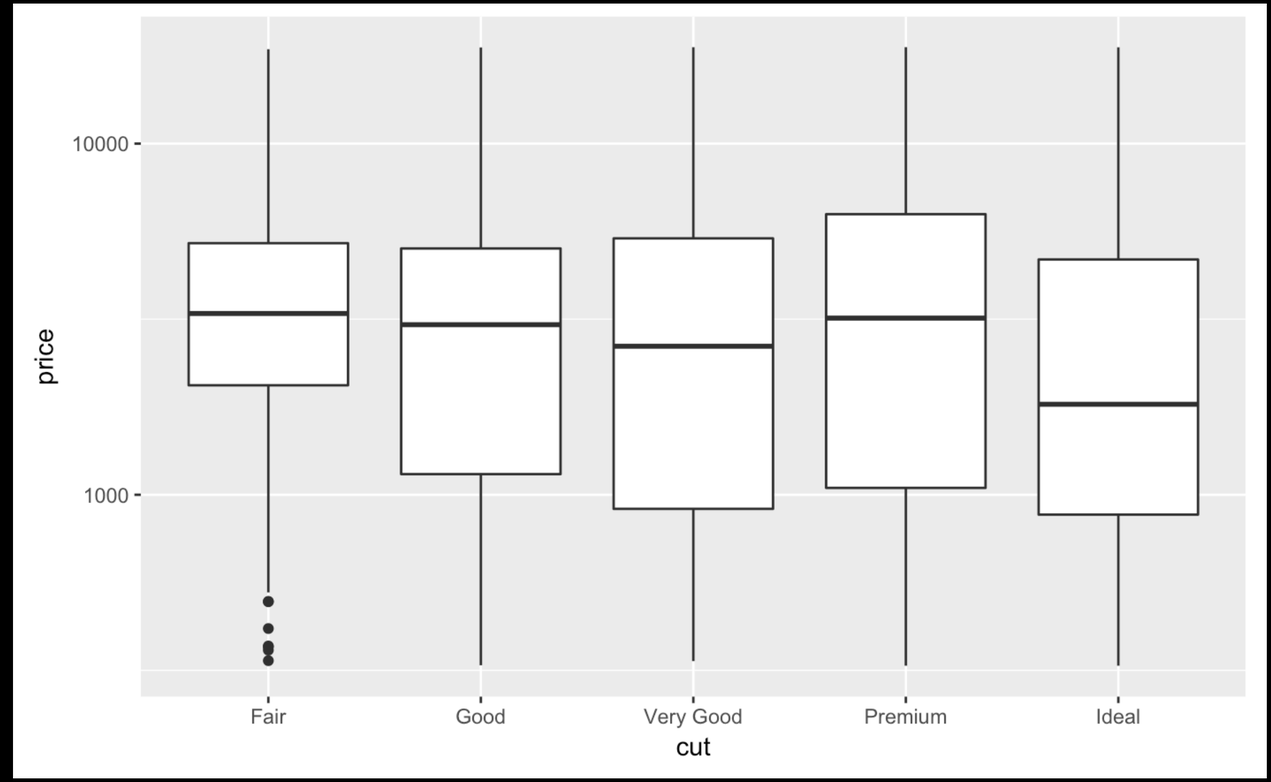
Relationship between cut and price? [log-scale]

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



Worse ——————————————————> Better

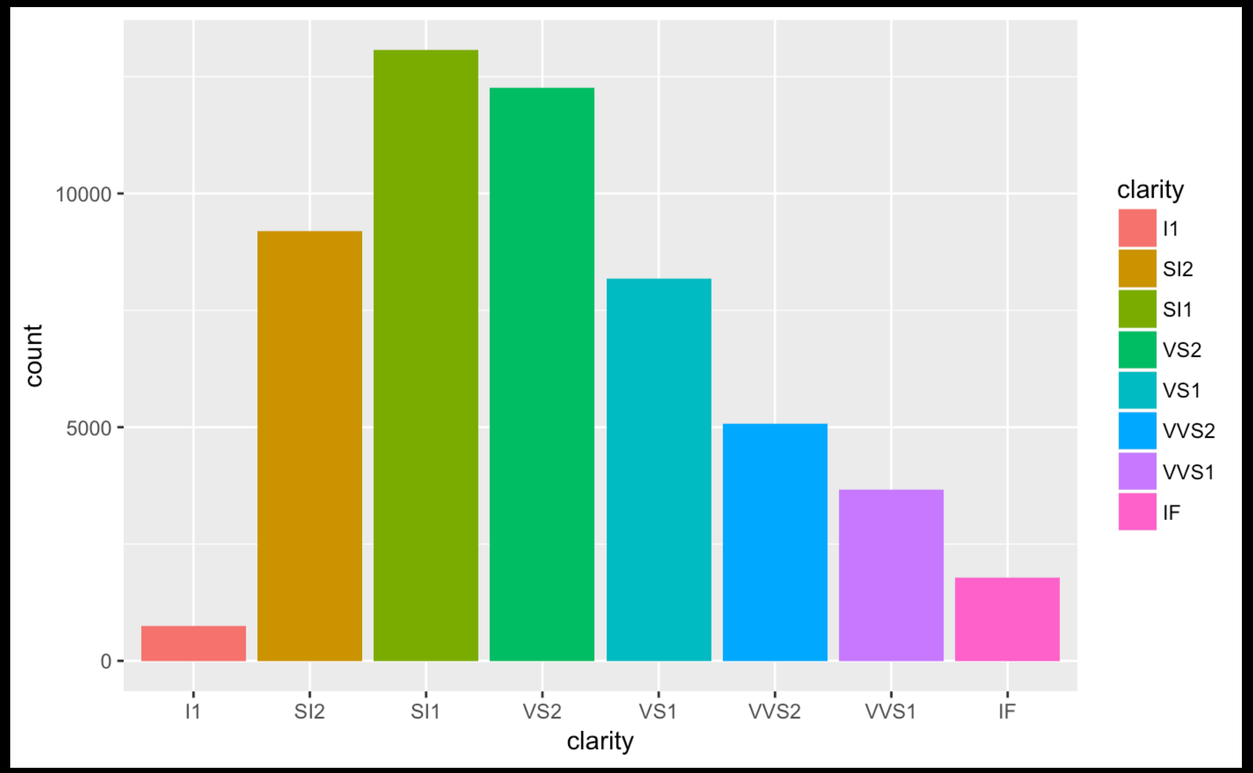
```
diamonds %>% ggplot(aes(x = cut, y = price)) +
 geom_boxplot()+ scale_y_log10()
```



Worse ——————————————————> Better

How many diamonds with each clarity?

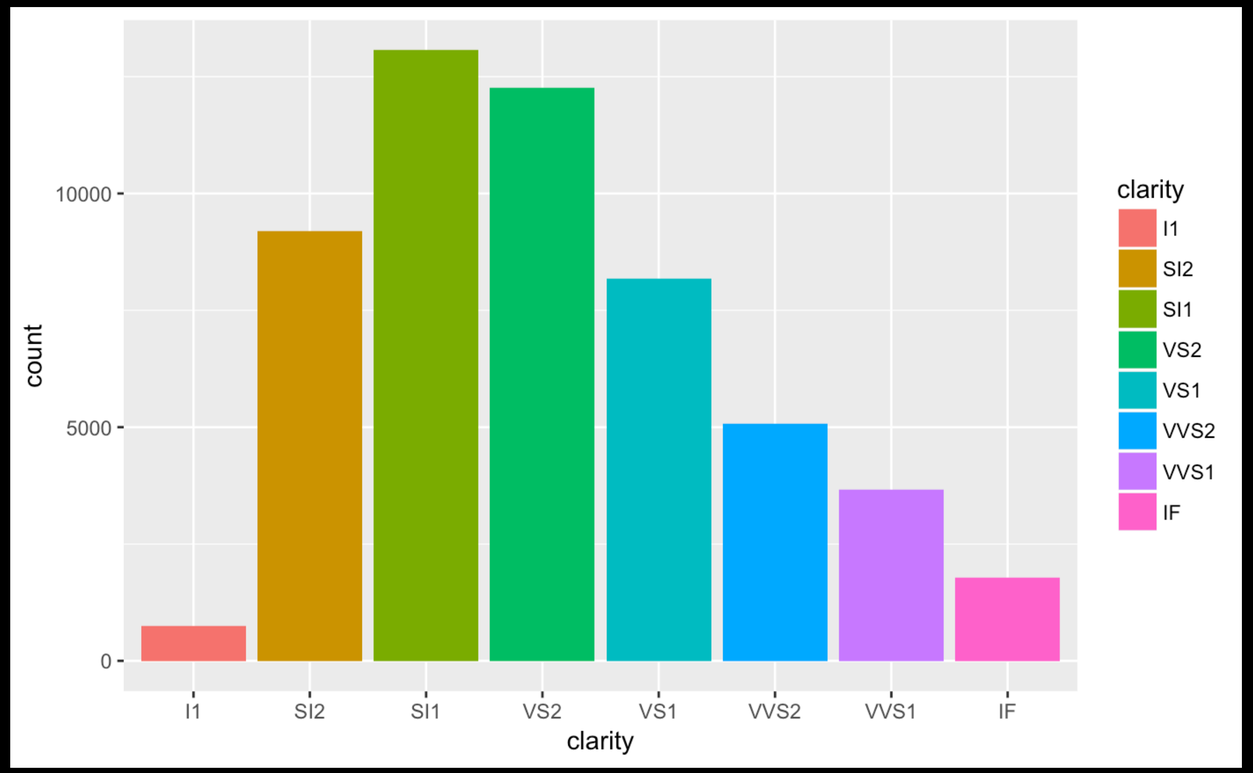
```
diamonds %>% ggplot(aes(x = clarity, fill = clarity))+geom_bar()
```



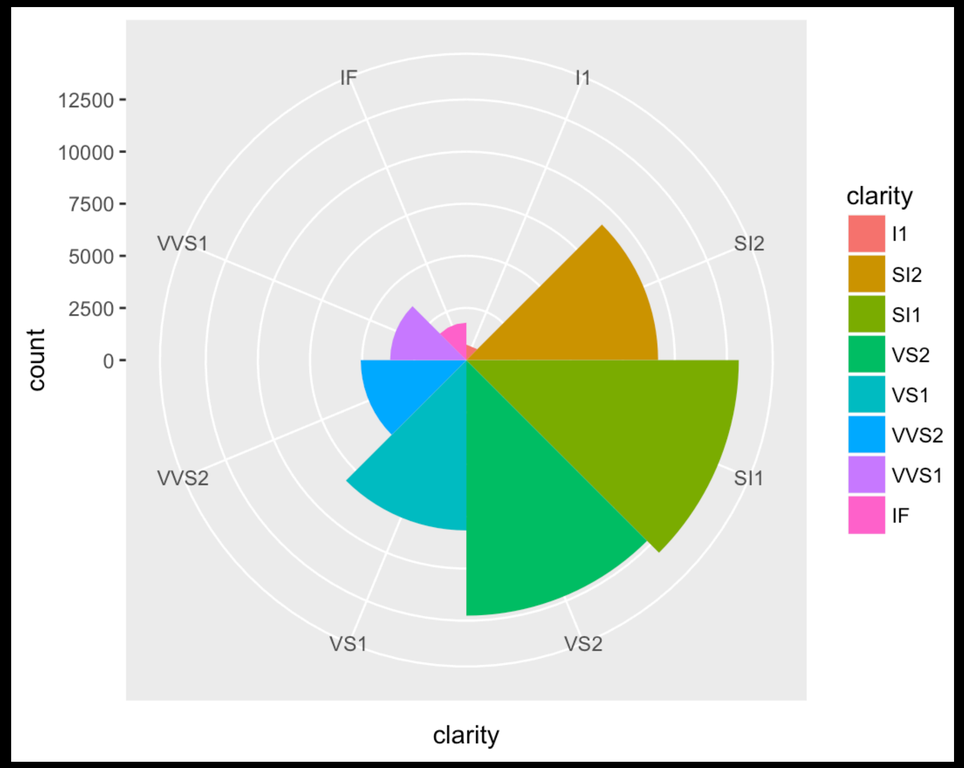
Worse ——————————————————> Better

How many diamonds with each clarity?

```
diamonds %>% ggplot(aes(x = clarity, fill = clarity))+geom_bar()
```




```
diamonds %>% ggplot(aes(x = clarity, fill = clarity))
+geom_bar(width=1) + coord_polar()
```



Worse ——————————————————> Better

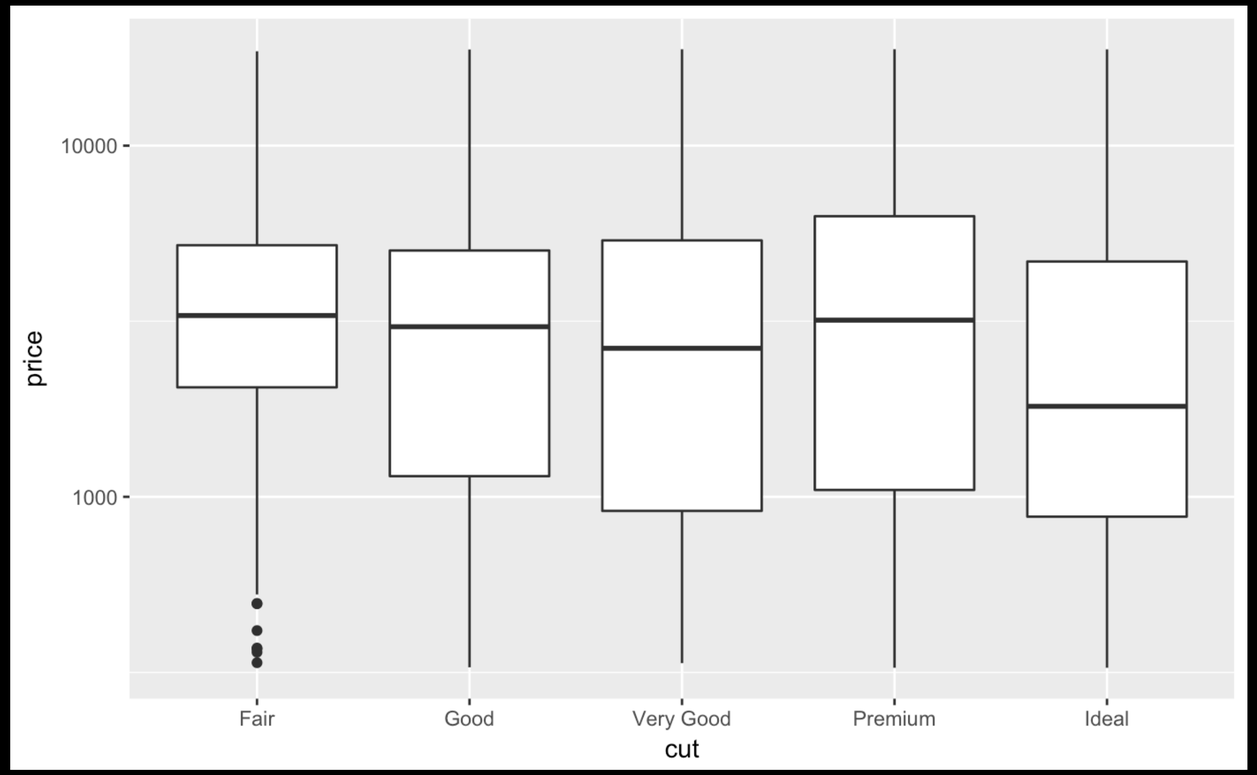
Components of the grammar

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

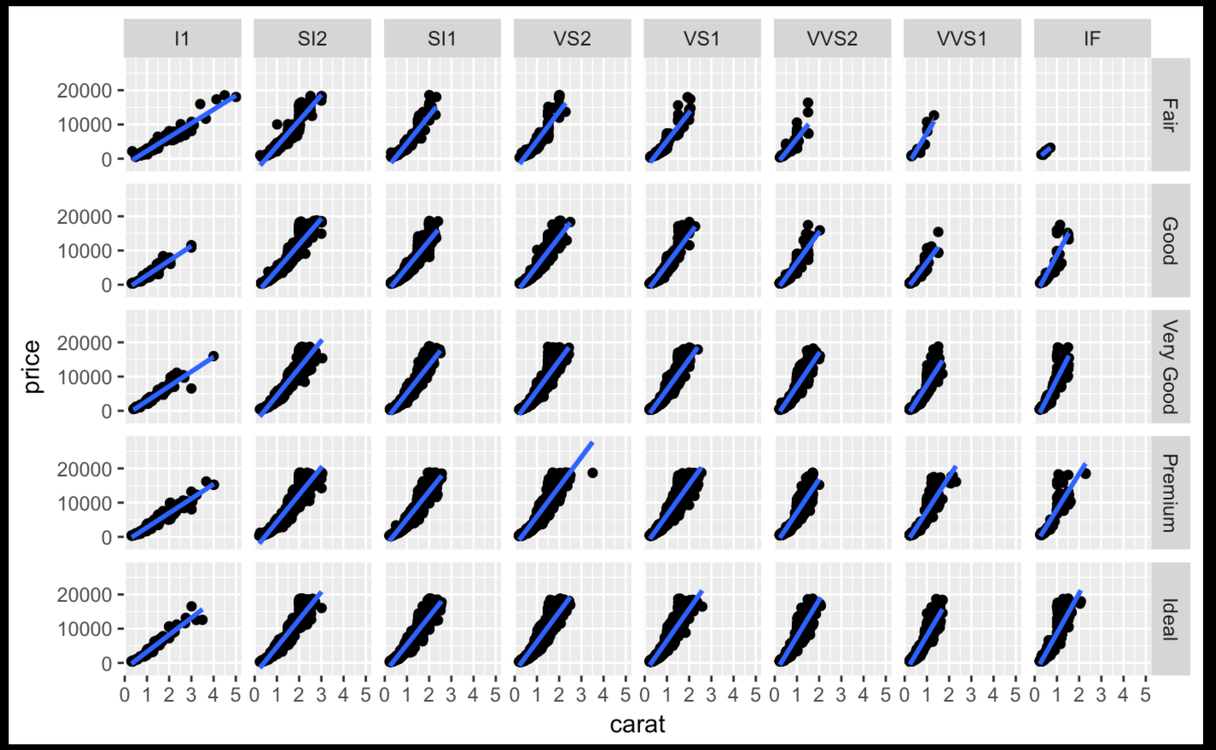
Adapted from DataCamp

Finishing up with the diamonds data

```
diamonds %>% ggplot(aes(x = cut, y = price))
+ geom_boxplot()+ scale_y_log10()
```

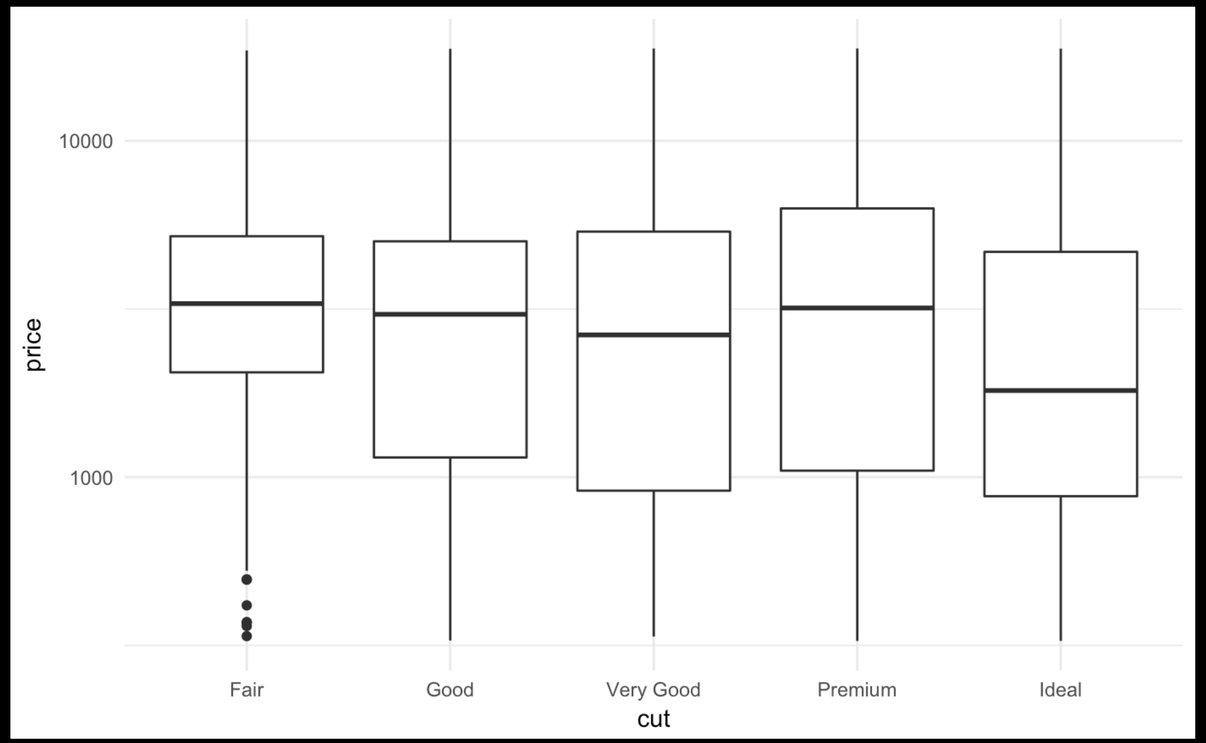



```
diamonds %>%  ggplot(aes(x = carat, y = price)) +
geom_point() + facet_grid(cut~clarity,
scales = "free_x") +  geom_smooth(method = "lm")
```

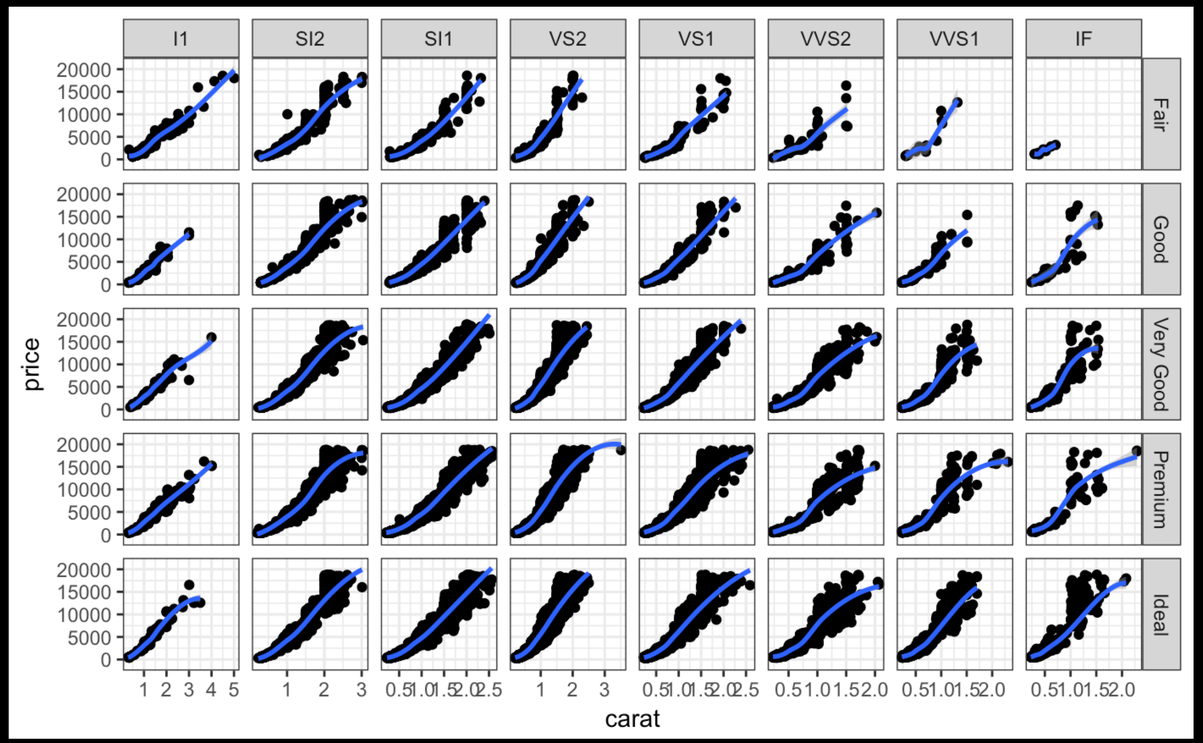


Finishing up with the diamonds data

```
diamonds %>% ggplot(aes(x = cut, y = price))
+ geom_boxplot()+ scale_y_log10() + theme_minimal()
```

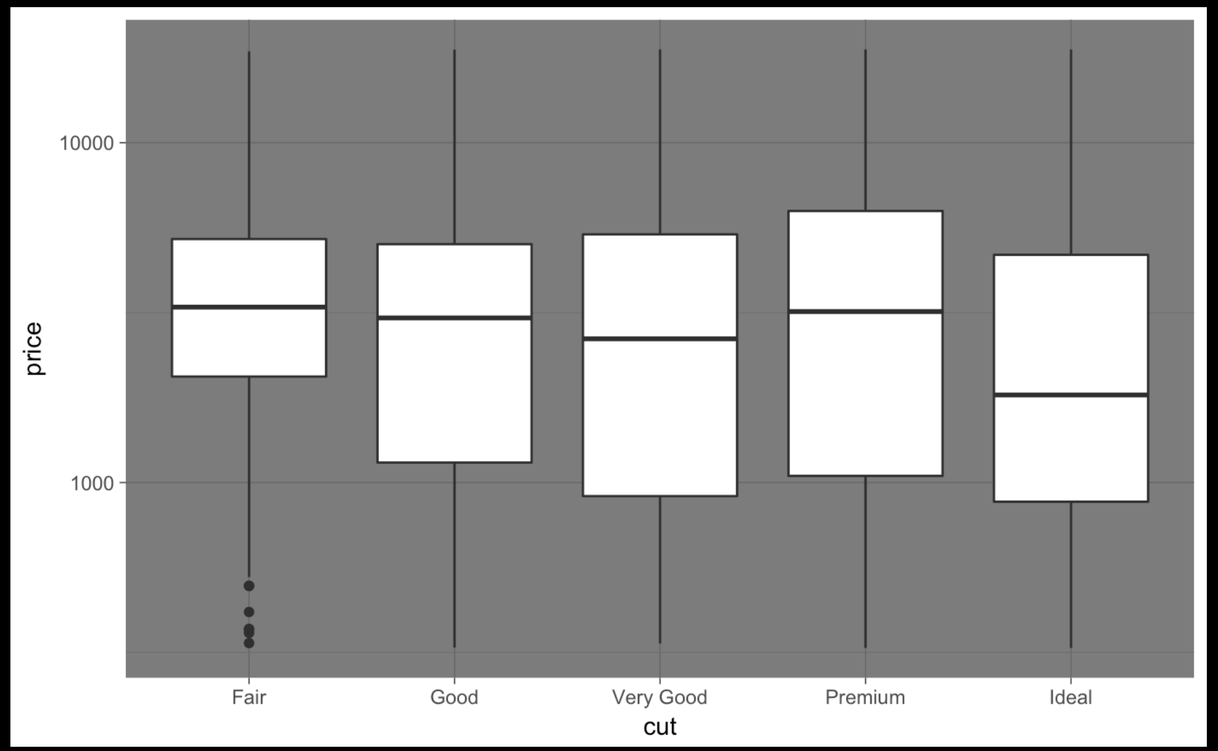



```
diamonds %>%  ggplot(aes(x = carat, y = price)) +
geom_point() + facet_grid(cut~clarity,
scales = "free_x") +  geom_smooth(method = "lm") + theme_bw()
```

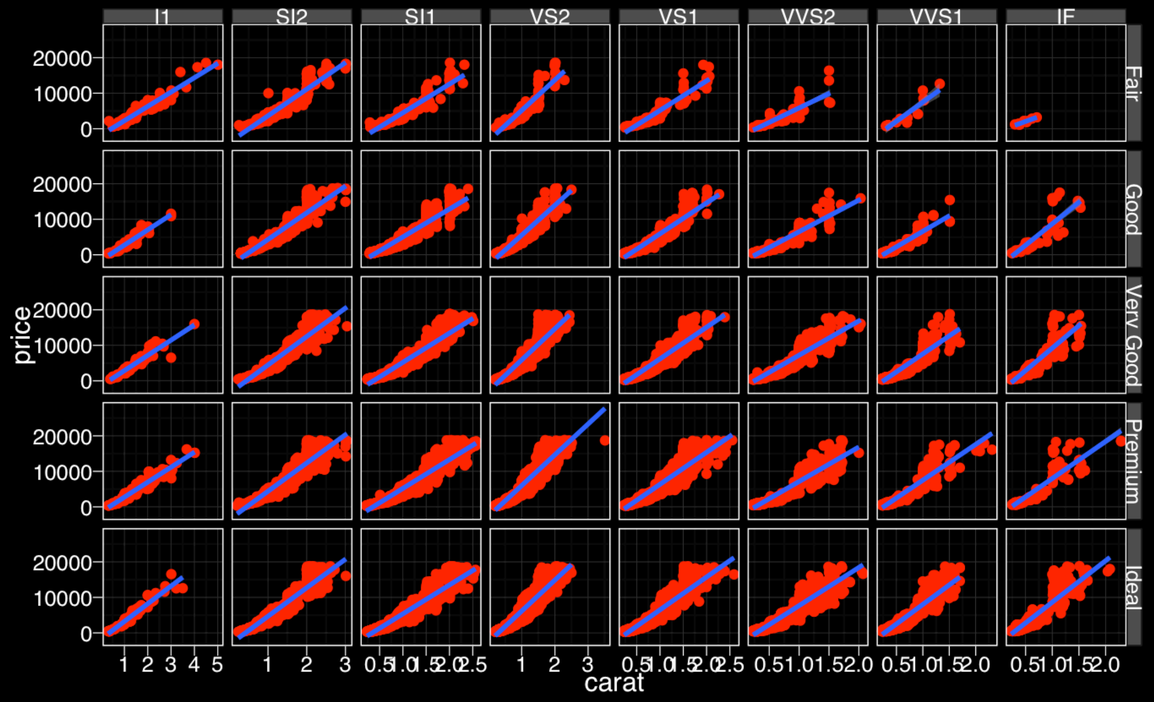


Just changing the theme

```
diamonds %>% ggplot(aes(x = cut, y = price))
+ geom_boxplot()+ scale_y_log10() + theme_dark()
```



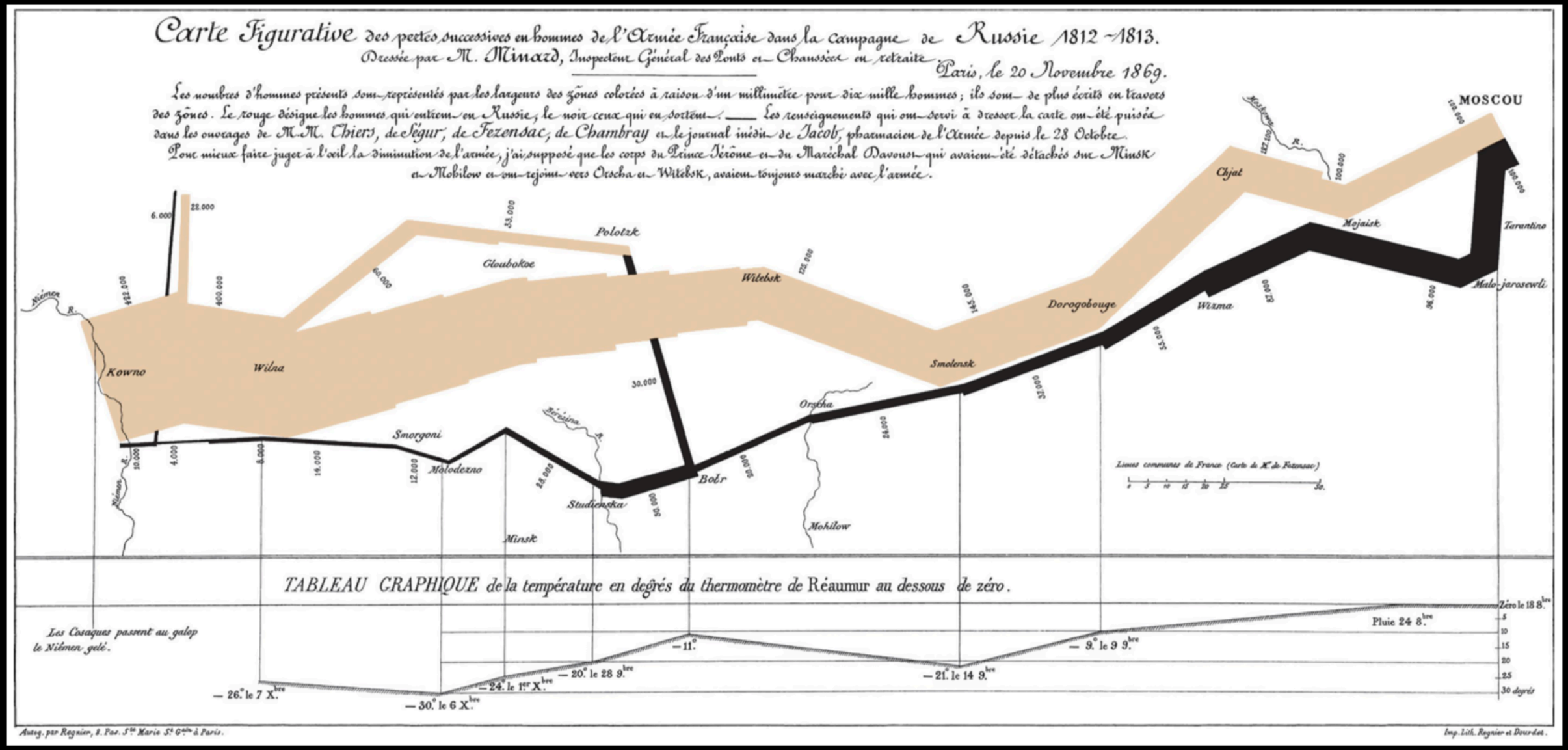

```
diamonds %>%  ggplot(aes(x = carat, y = price)) +
geom_point(col = "red") + facet_grid(cut~clarity,
scales = "free_x") +  geom_smooth(method = "lm") +
theme_black()
```



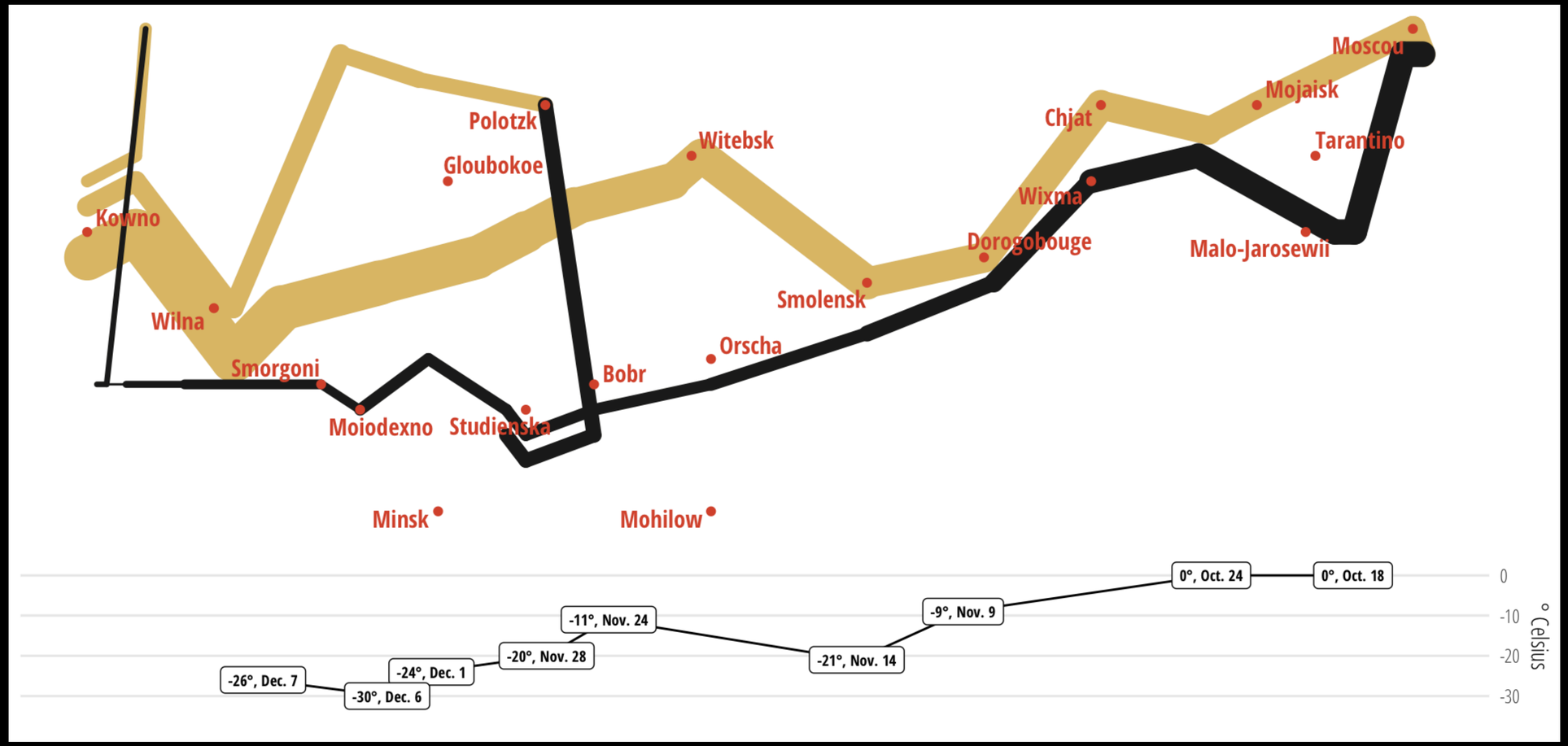
The Grammar of Graphics

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

Limitations



Limitations



https://www.andrewheiss.com/blog/2017/08/10/exploring-minards-1812-plot-with-ggplot2/

The Grammar of Graphics

Data	Variables of interest				
Aesthetics	x-axis y-axis	colour fill	size labels	alpha shape	line width line type
Geometries	point	line	histogram	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	fixed	polar	limits	
Themes	Not data, but important for overall impact				

https://github.com/dvanic/18ResBazPresentation