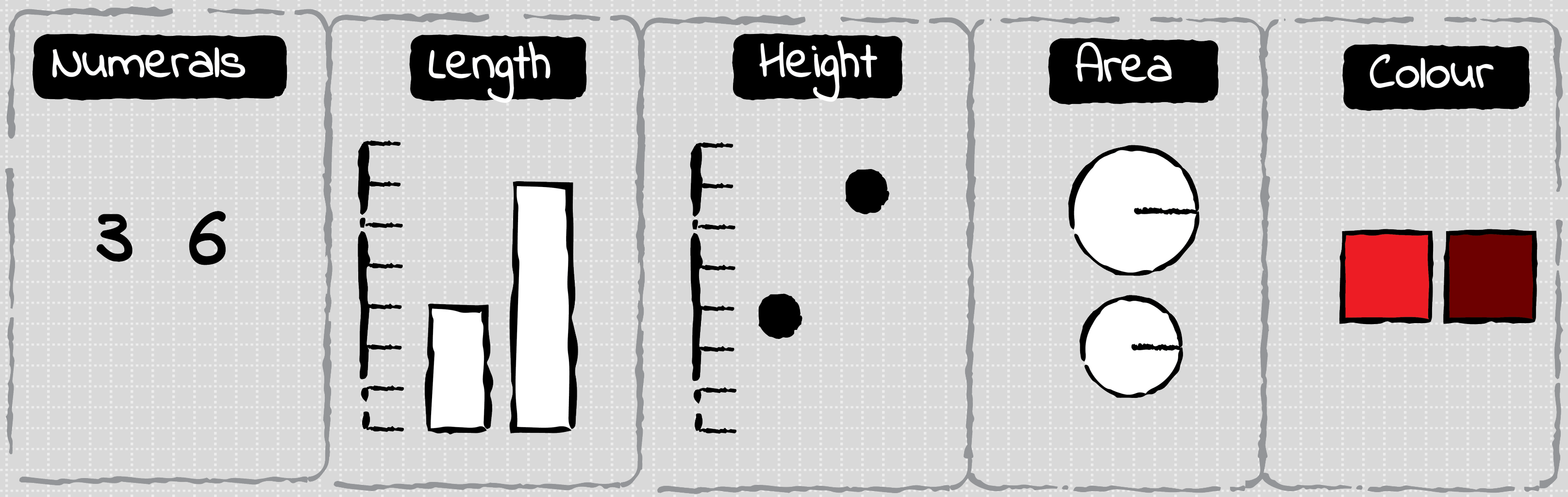
#hackYourData Beautiful

Your Data and Your Visualizations

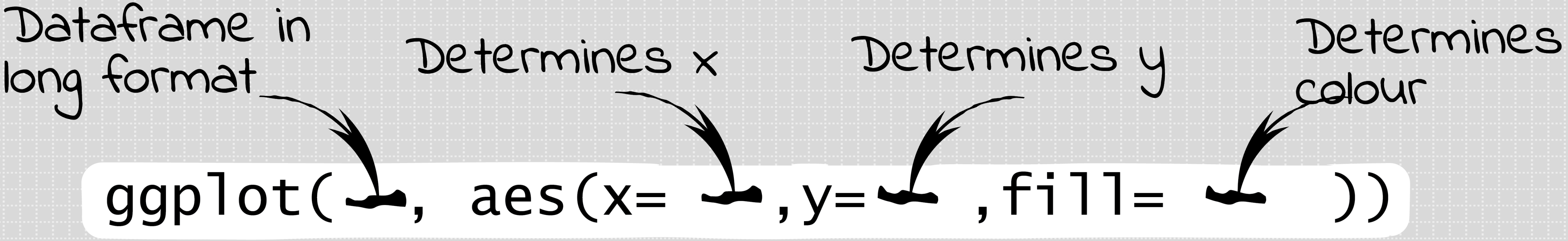
Lovisa Sundin
l.sundin.1@research.gla.ac.uk

Mapping numbers to visual properties



ggplot2 - the Grammar of Graphics

- First, we provide:
  - our data in a long-format data frame
  - which columns should be mapped to visual properties ("**aesthetics**") by default



Dataframe in long format

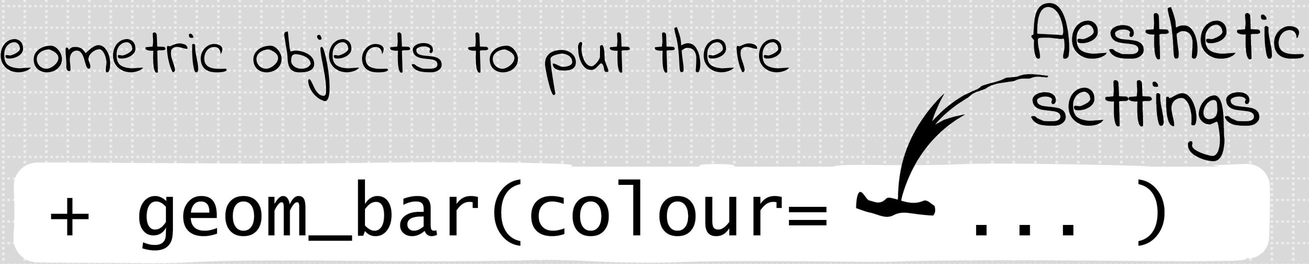
Determines x

Determines y

Determines colour

```
ggplot(  , aes(x=  ,y=  ,fill=   ))
```

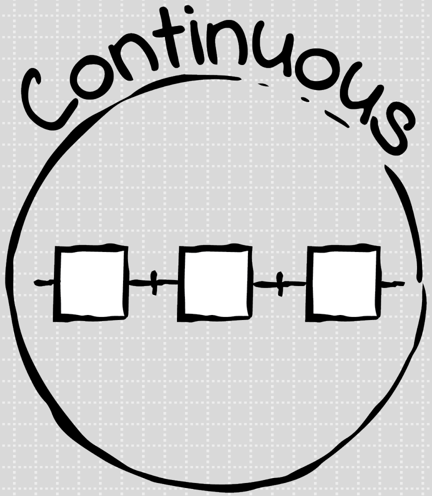
- Second, we tell it *which* geometric objects to put there



Aesthetic settings

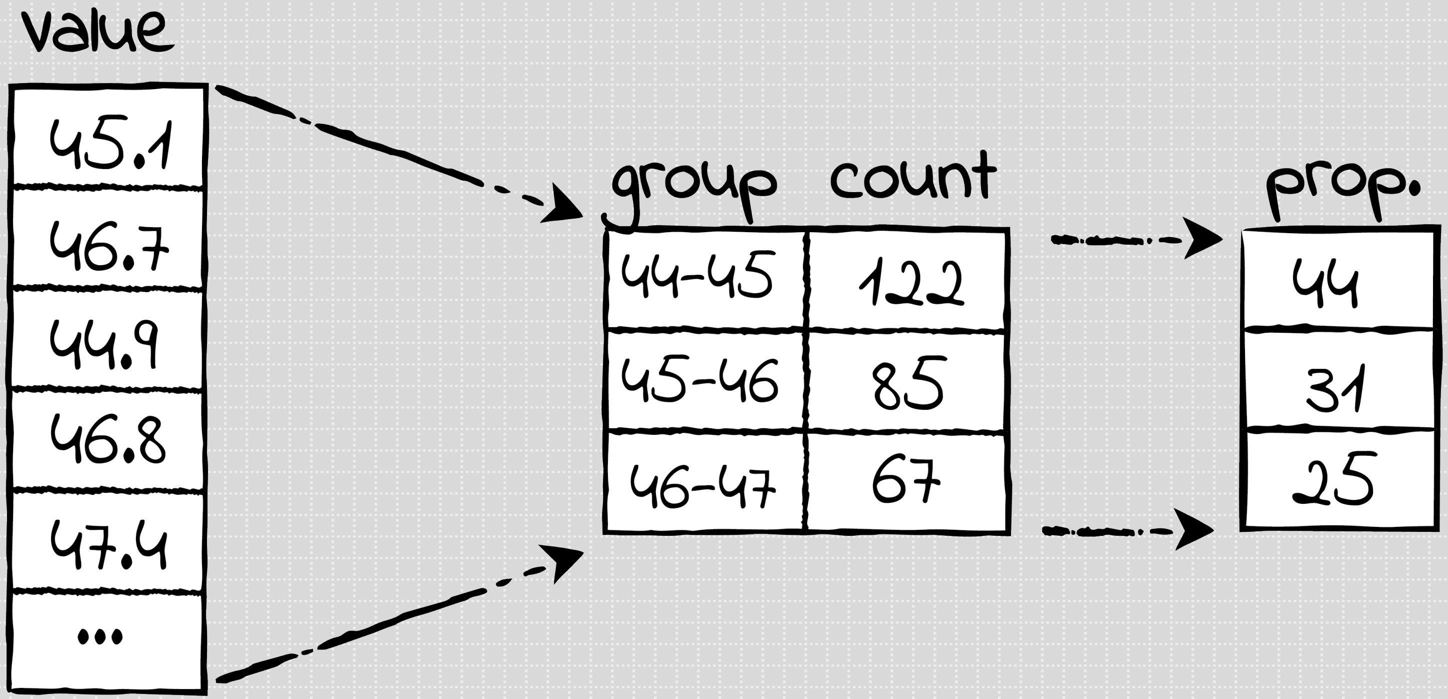
```
+ geom_bar(colour=  ... )
```

which geom to use for which data?



Continuous

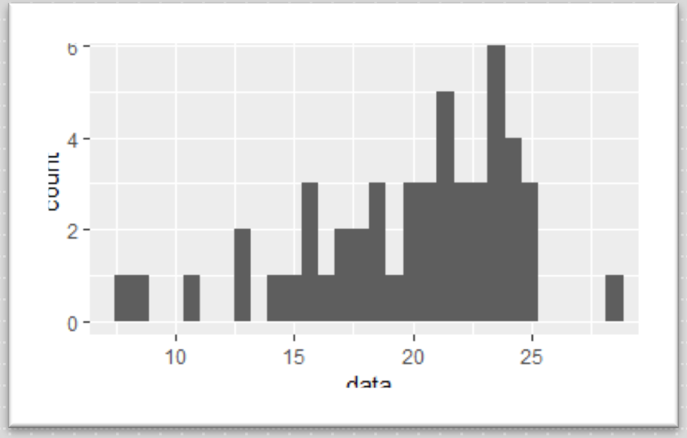
Value vs. Group vs. Count vs. Proportion



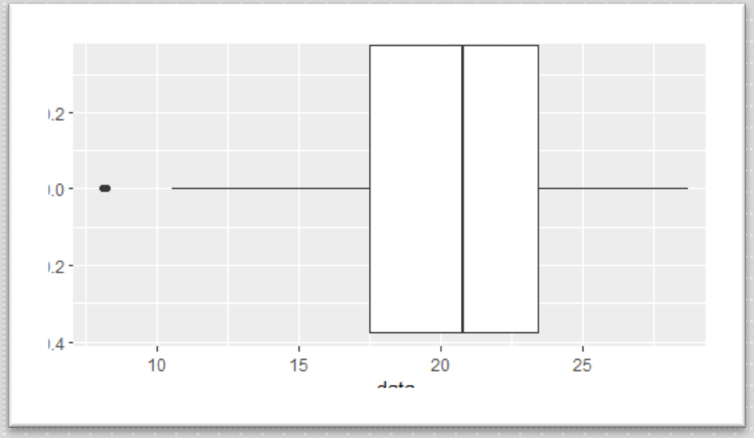
value
45.1
46.7
44.9
46.8
47.4
...

group	count
44-45	122
45-46	85
46-47	67

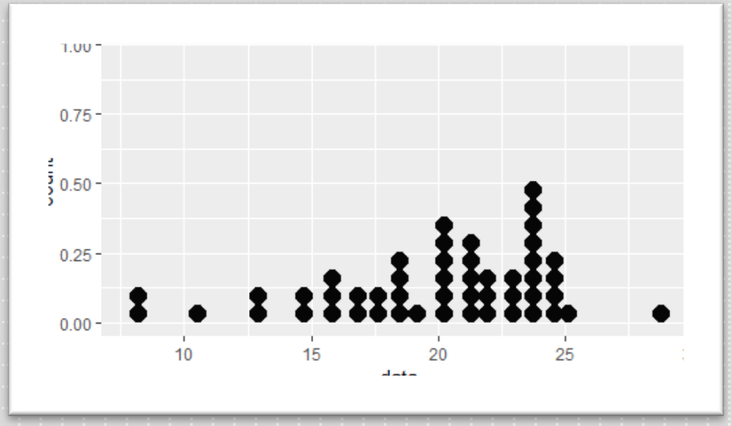
prop.
44
31
25



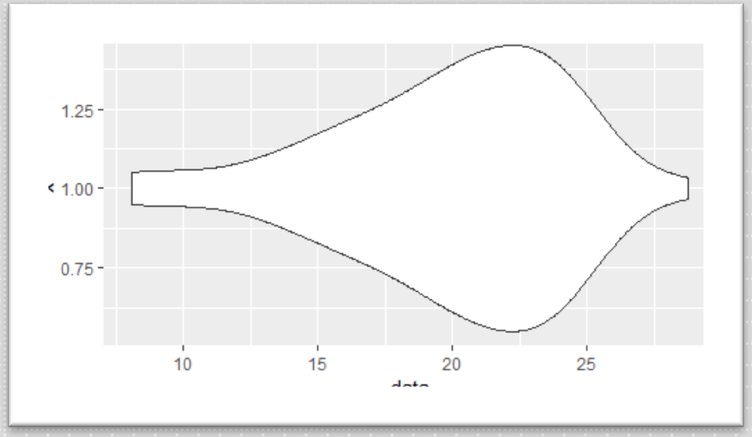
```
ggplot(data, aes(x=data)) +
geom_histogram()
```



```
ggplot(data, aes(y=data,
group=1)) + geom_boxplot() +
coord_flip()
```

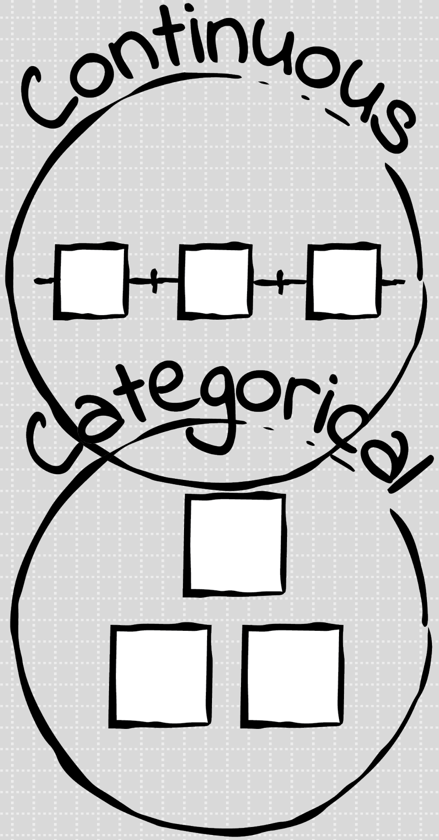


```
ggplot(data, aes(x=data)) +
geom_dotplot()
```



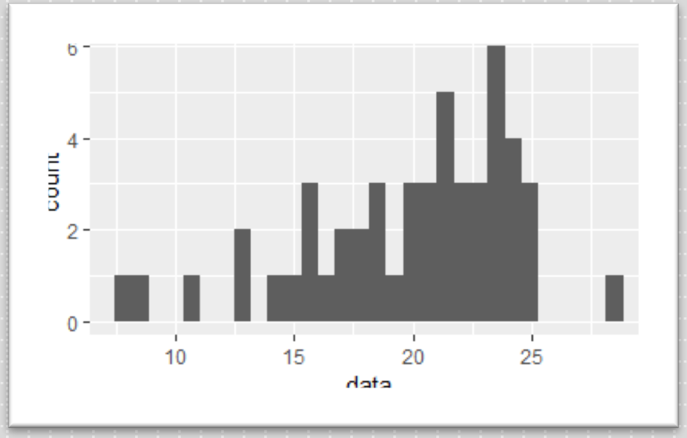
```
ggplot(data, aes(y=data, x=1)) +
geom_violin() + coord_flip()
```

Continuous x Categorical

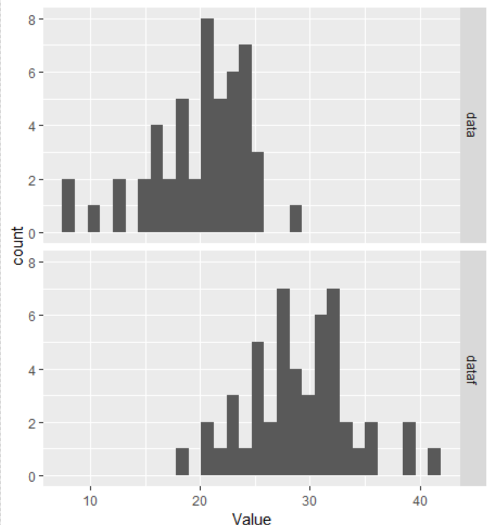


Group	Value
"A"	45.1
"B"	46.7
"B"	44.9
"A"	46.8
"A"	47.4
...	...

Histogram

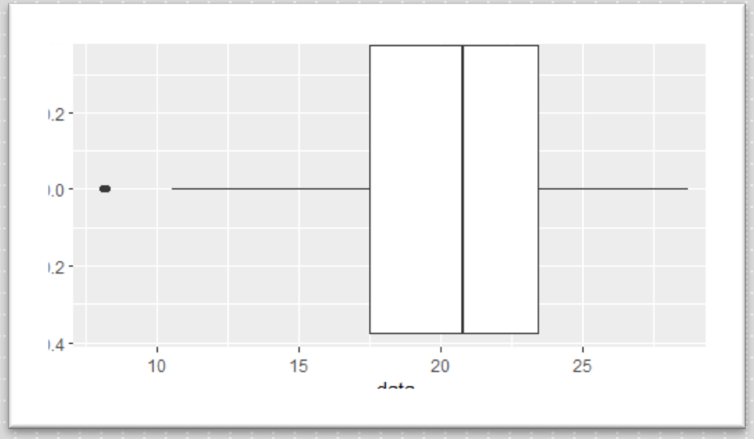


```
ggplot(df, aes(x=data)) +
geom_histogram()
```

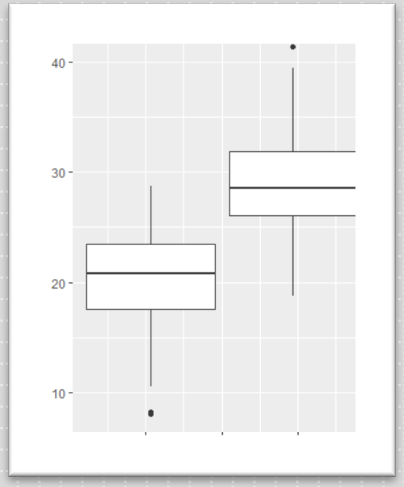


```
ggplot(data, aes(x=Value))
+ geom_histogram()+
facet_grid(rows =
vars(Group))
```

Boxplot

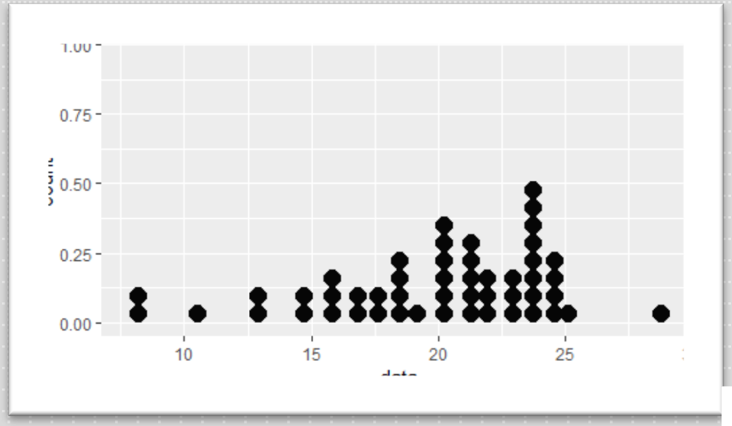


```
ggplot(df, aes(y=data, group=1))
+ geom_boxplot() + coord_flip()
```

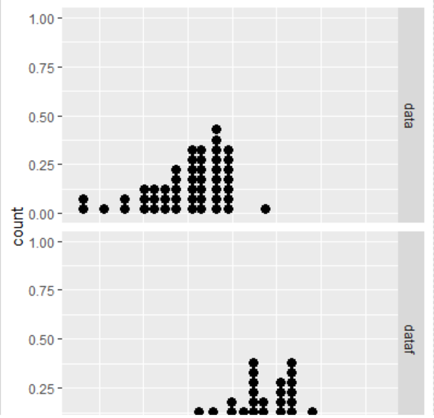


```
ggplot(data, aes(y=Value,
x=Group)) +
geom_boxplot()
```

Dotplot

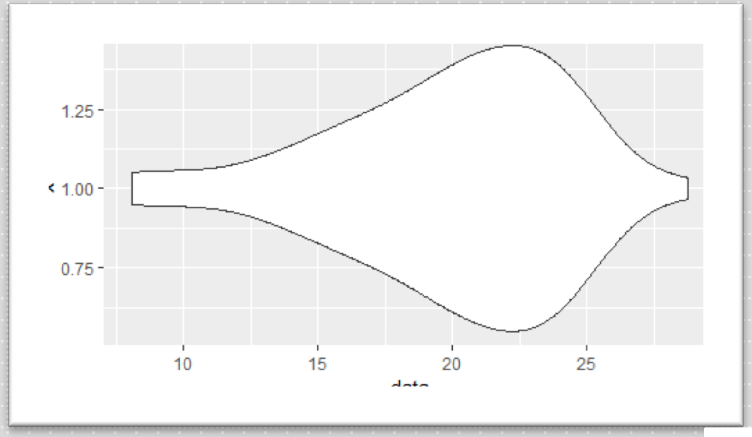


```
ggplot(df, aes(x=data)) +
geom_dotplot()
```

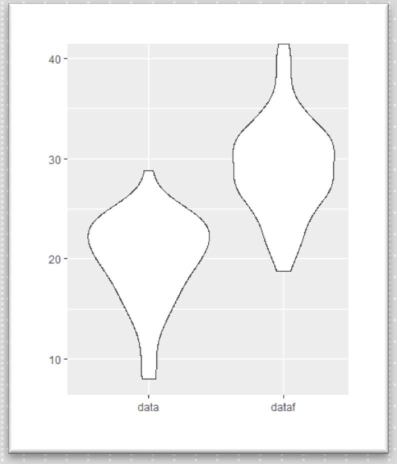


```
ggplot(data, aes(x=Value))
+ geom_dotplot()+
facet_grid(rows =
vars(Group))
```

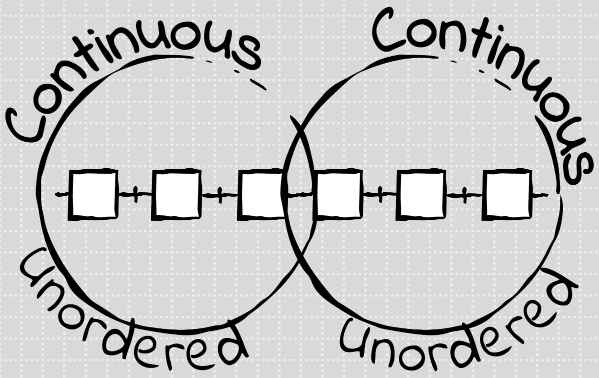
Violinplot



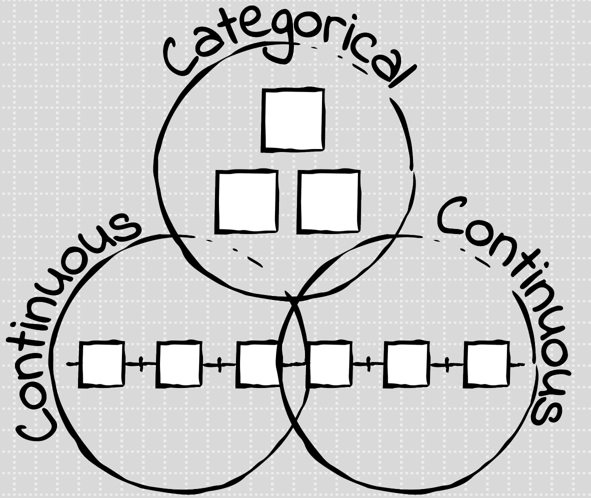
```
ggplot(df, aes(y=data, x=1)) +
geom_violin() + coord_flip()
```



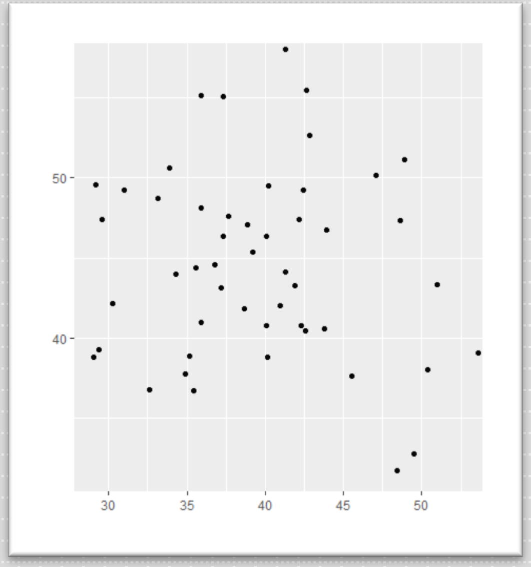
```
ggplot(data, aes(y=Value,
x=Group)) +
geom_violin()
```



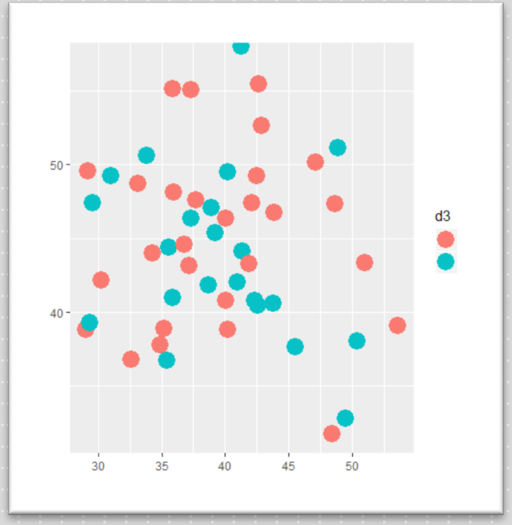
x	y



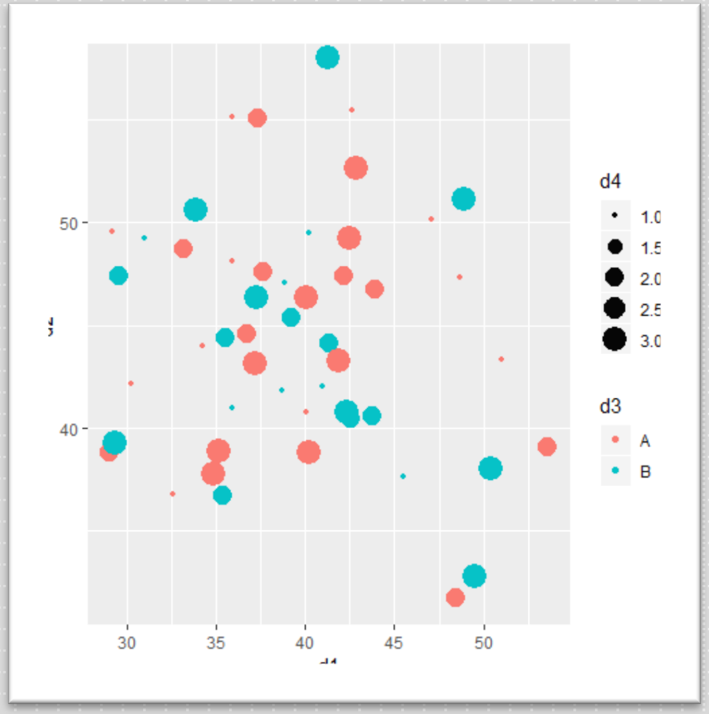
x	y	Group
		"A"
		"B"
		"B"
		"A"
		"A"
		•••



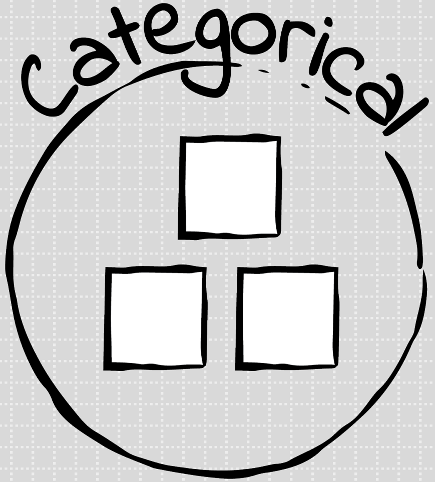
ggplot(df, aes(x=d1, y=d2))
+ geom_point()



ggplot(df, aes(x=d1, y=d2,
color=Group)) + geom_point()



ggplot(df, aes(x=d1, y=d2,
color=Group, size=d4)) +
geom_point()



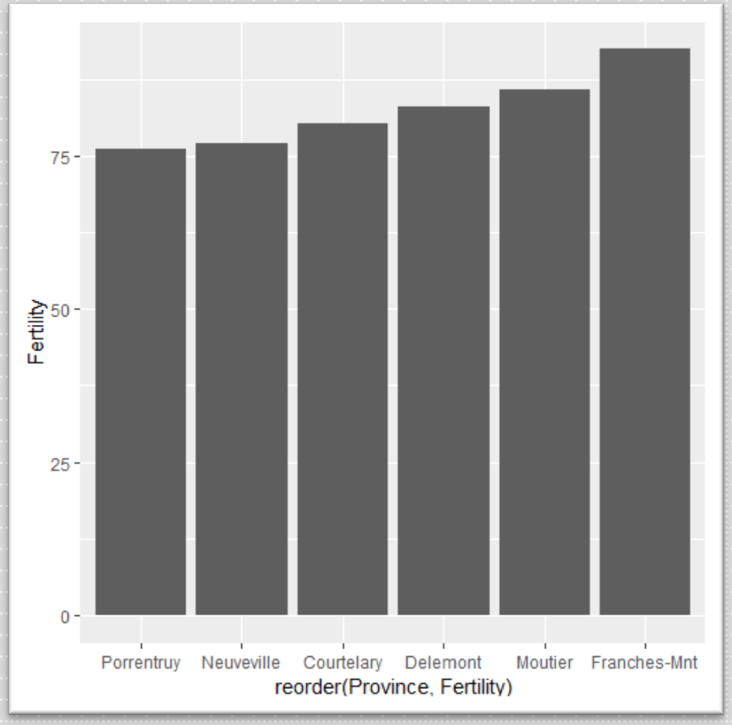
Value vs. Group vs. Count

value
'A'
'A'
'B'
'C'
'B'
...

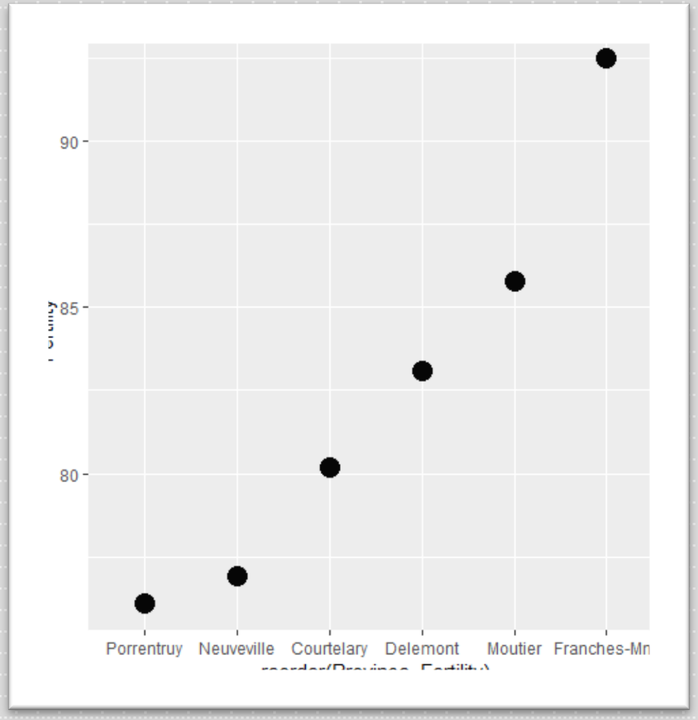
group	count
'A'	45
'B'	85
'C'	67

prop.
22
43
35

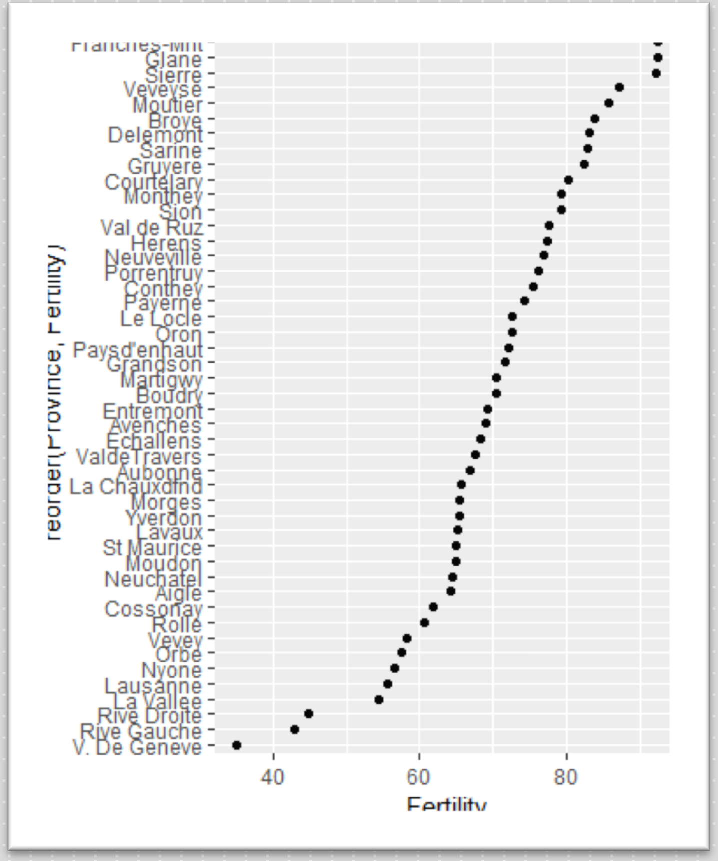
Count as discrete numerical



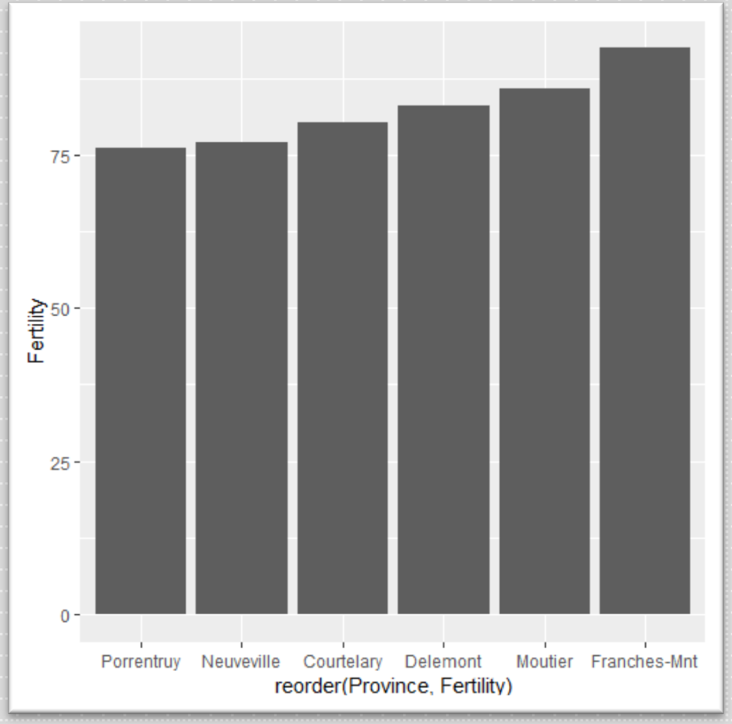
```
ggplot(head(df), aes(y = Fertility, x
= reorder(Province, Fertility))) +
geom_bar(stat="identity")
```



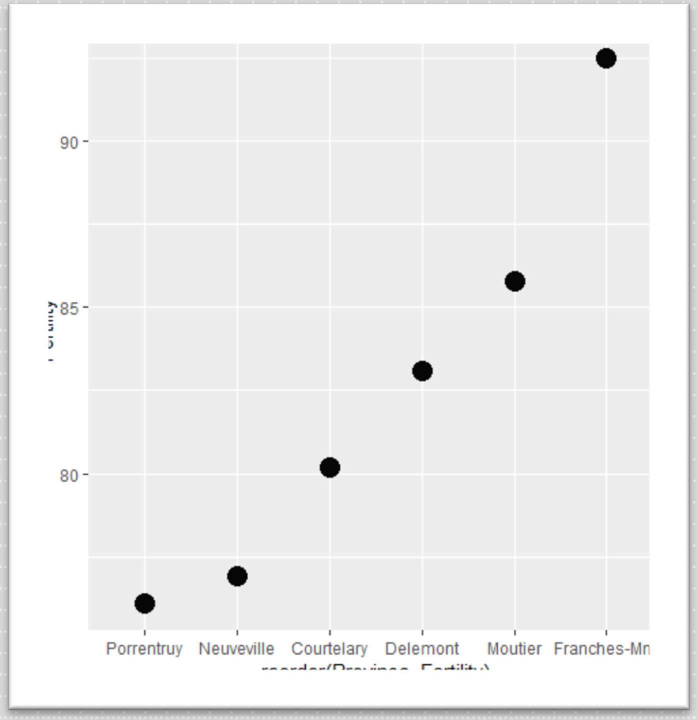
```
ggplot(df, aes(x = Fertility, y =
reorder(Province, Fertility))) +
+     geom_point()
```



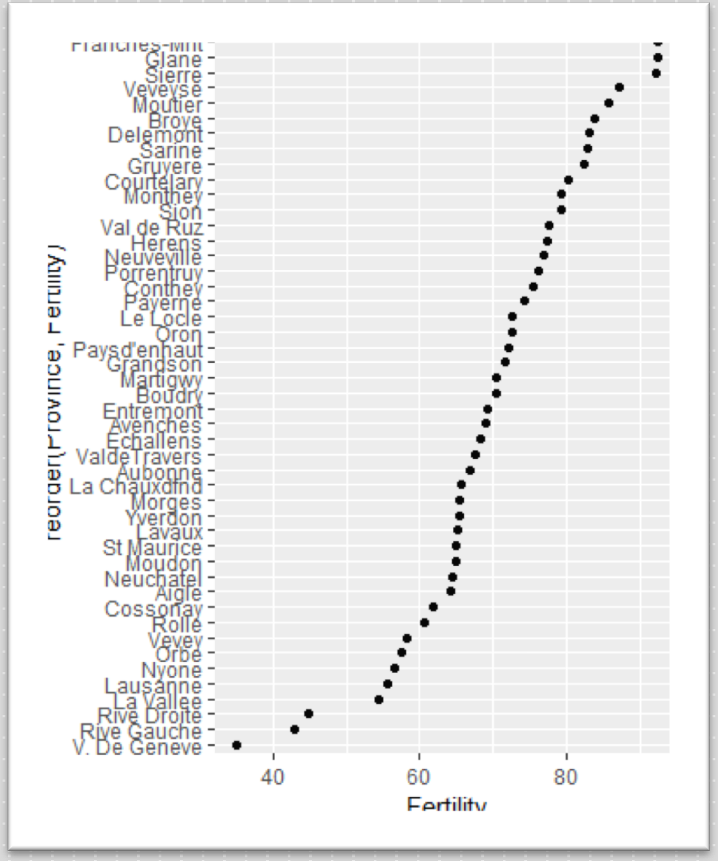
```
ggplot(df, aes(x = Fertility, y =
reorder(Province, Fertility))) +
+     geom_point()
```



```
ggplot(head(df), aes(y = Fertility, x
= reorder(Province, Fertility))) +
geom_bar(stat="identity")
```

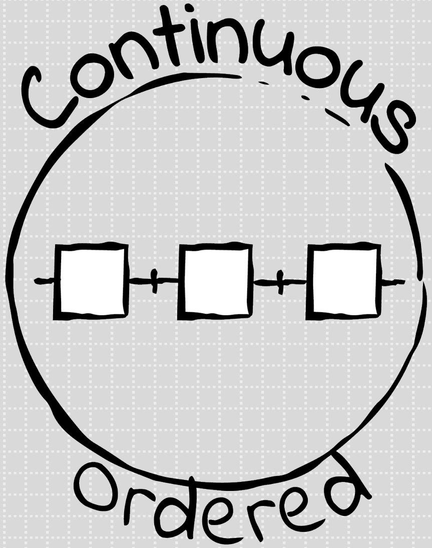


```
ggplot(df, aes(x = Fertility, y =
reorder(Province, Fertility))) +
+     geom_point()
```



```
ggplot(df, aes(x = Fertility, y =
reorder(Province, Fertility))) +
+     geom_point()
```

Unordered vs. ordered data



value
45.1
46.7
44.9
47.4

Unordered

	value
1	45.1
2	46.7
3	44.9
4	46.8
5	47.4
...	...

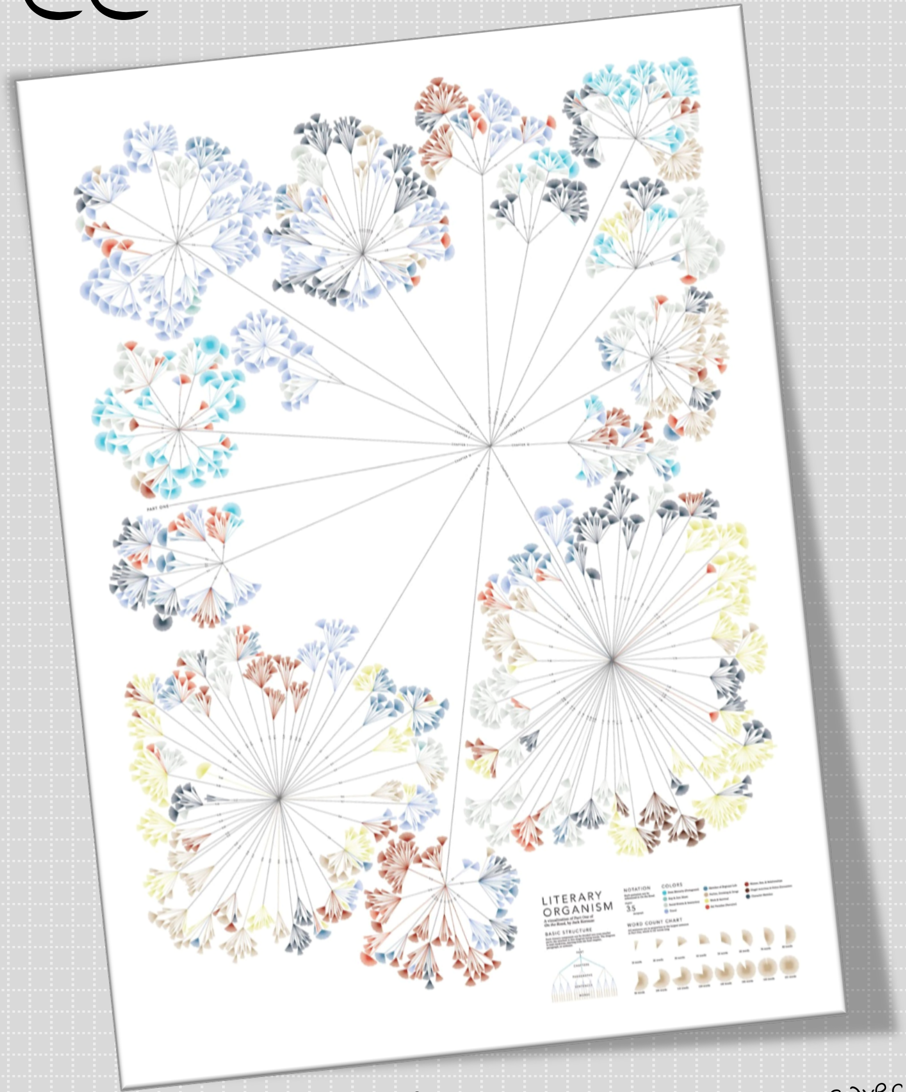
Implicit ordering

index	value
3	45.1
5	46.7
7	44.9
8	46.8
13	47.4
...	...

Explicit ordering

Beyond ggplot2...

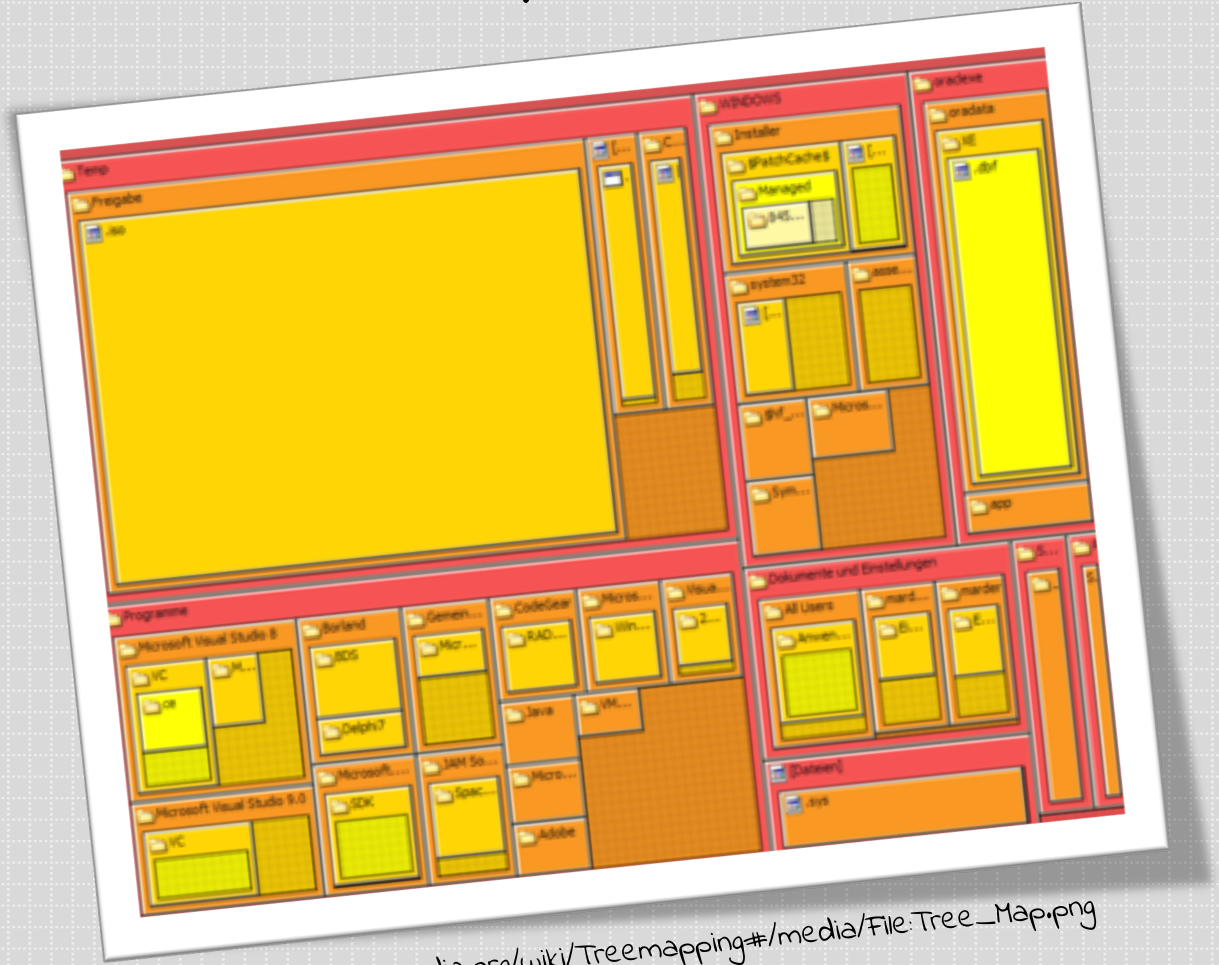
Tree



By Stefanie Posavec
http://www.thefunctionalart.com/2012/06/stefanie-posavec-
infographics-and.html

parent	child	theme1	leaf
ch1_p1	ch1_p1_s1	"love"	TRUE

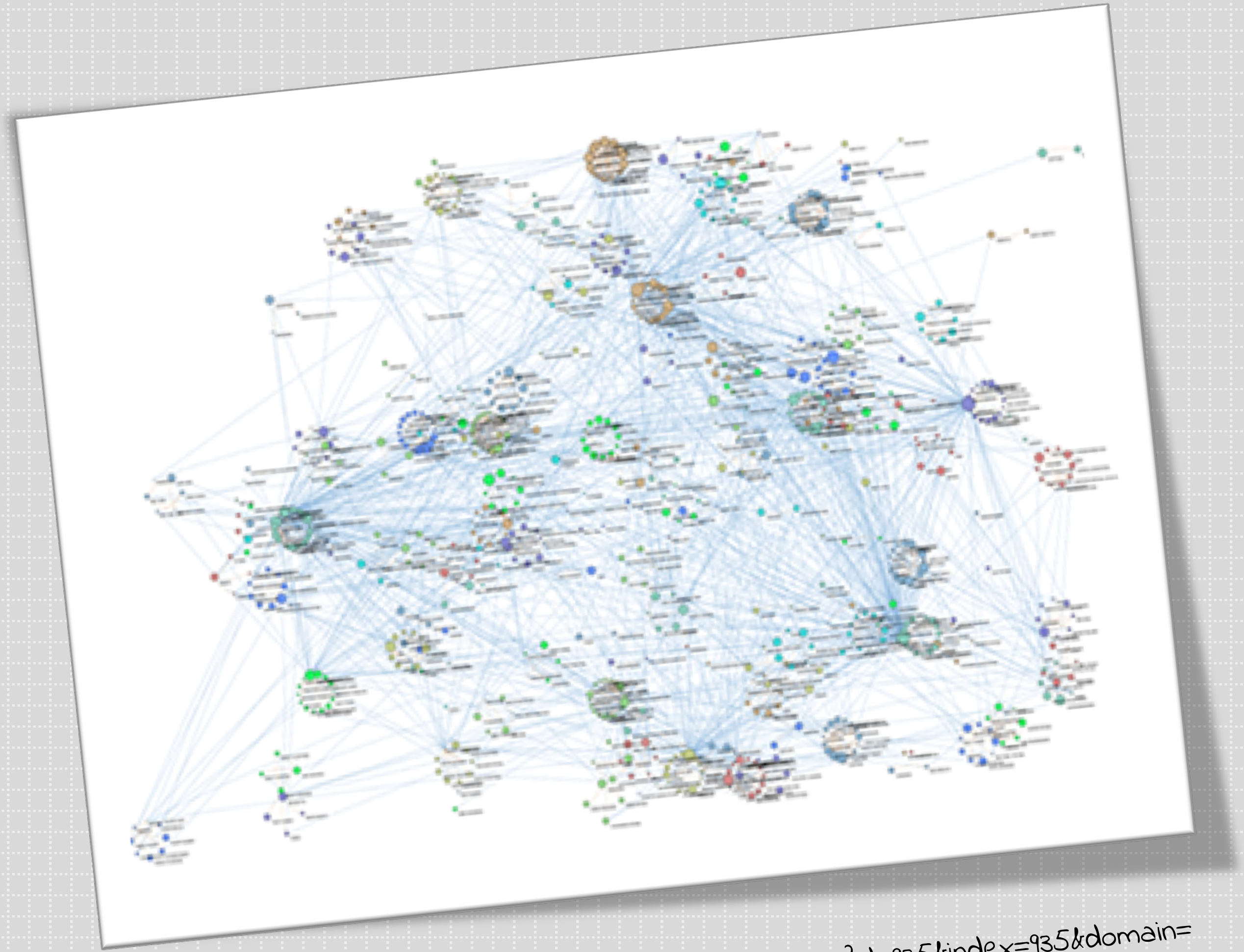
Treemap



https://en.wikipedia.org/wiki/Treemapping#/media/File:Tree_Map.png

parent	node	size
User"	"Docs"	56

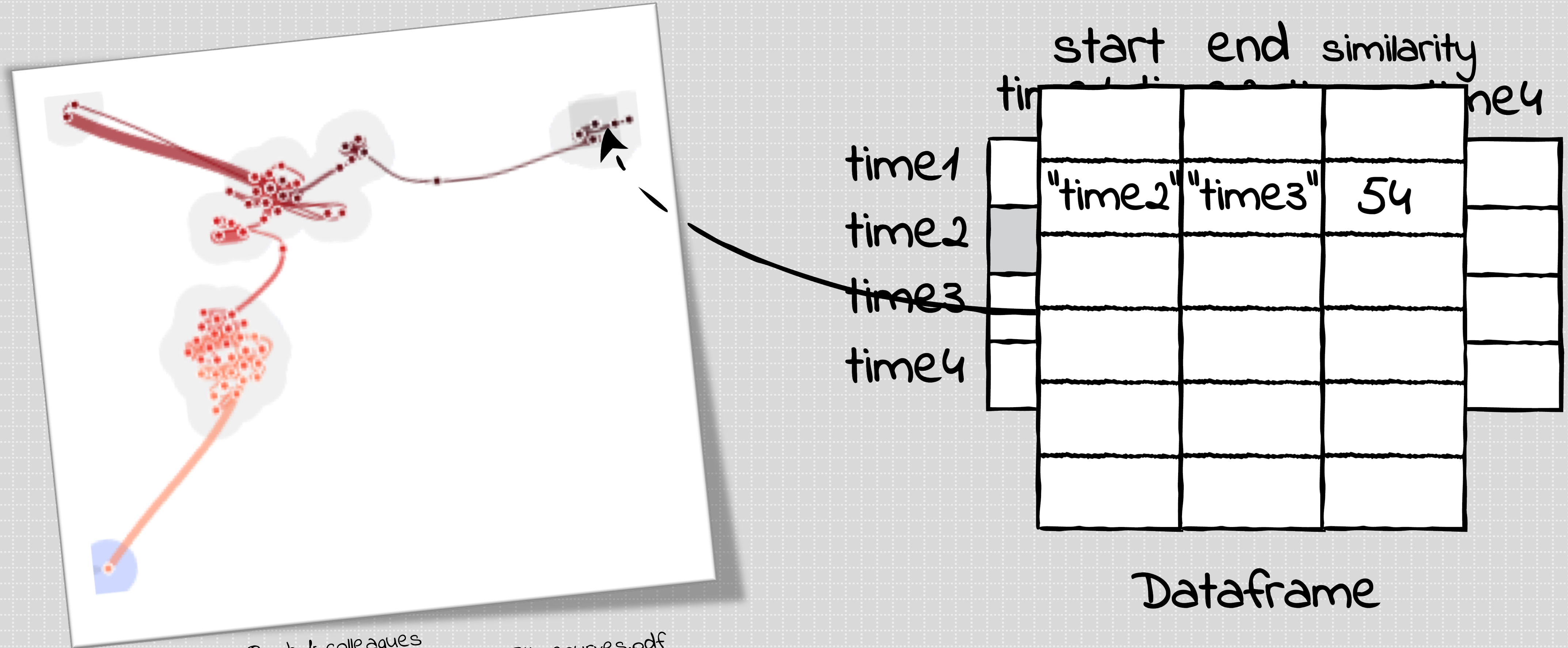
Network



By The Guardian via http://www.visualcomplexity.com/vc/project_details.cfm?id=935&index=935&domain=

node1	node2	size
"LAX"	"JFK"	56

Time Curve

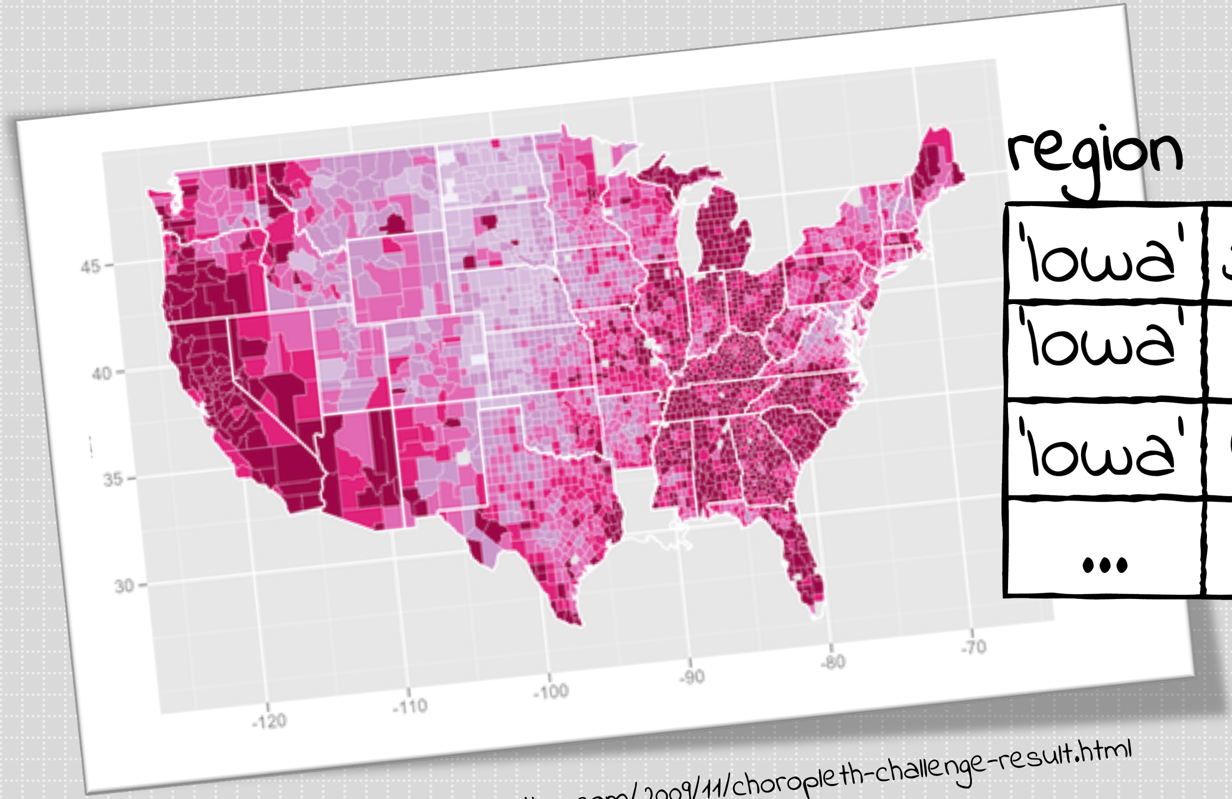


start	end	similarity
"time2"	"time3"	54

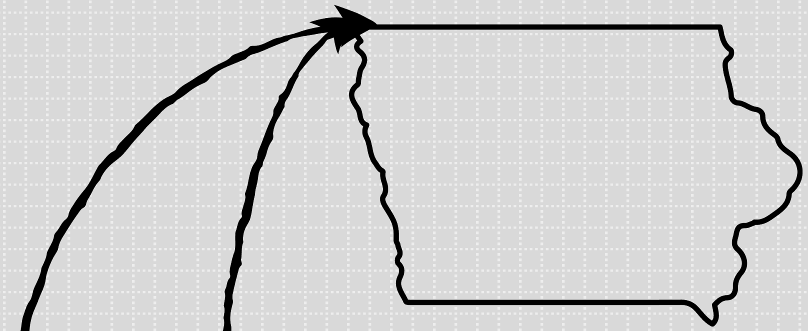
Dataframe

By Benjamin Bach & colleagues
https://aviz.fr/~bbach/timecurves/Bach2015timecurves.pdf

2d Choropleth



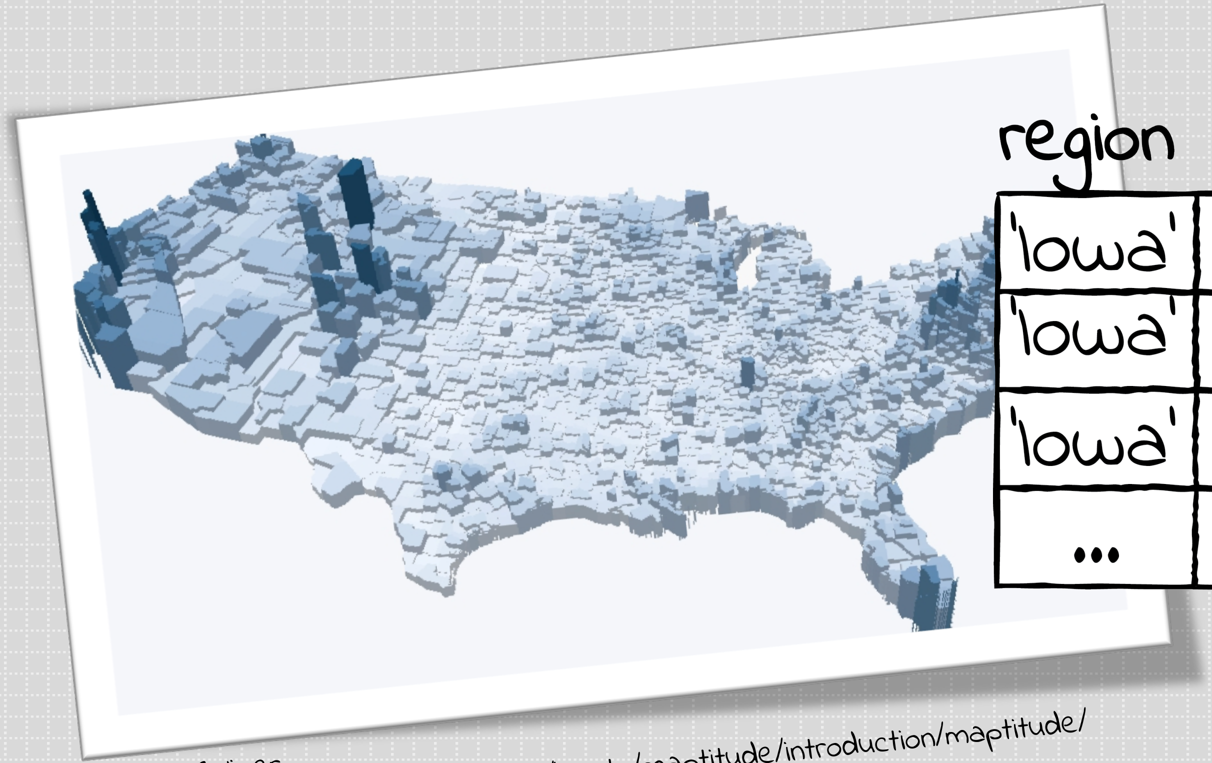
https://blog.revolutionanalytics.com/2009/11/choropleth-challenge-result.html



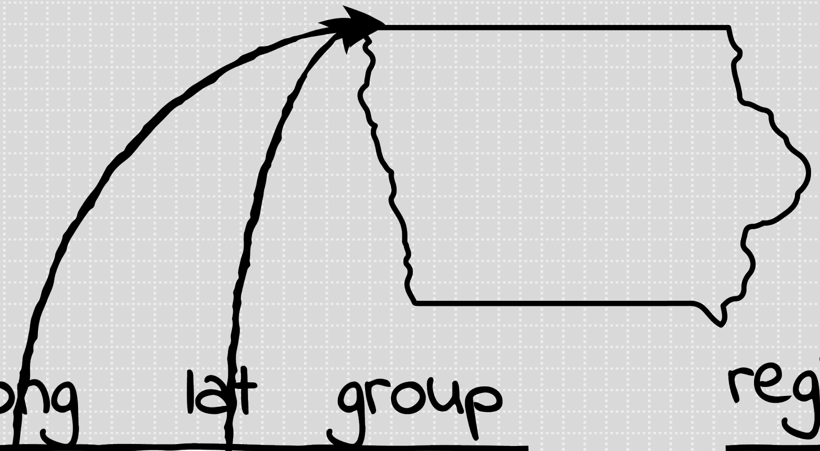
region	long	lat	group
'Iowa'	56.73	23.30	5
'Iowa'	47.93	23.83	5
'Iowa'	45.74	22.63	5
...	...	...	...

region	value
'Iowa'	0.78

3d Choropleth



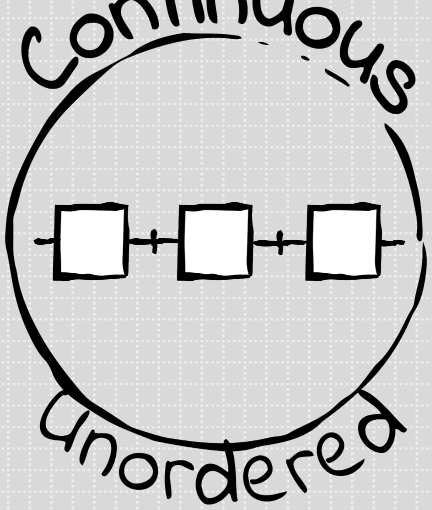
By Caliper
https://www.mapping-tools.com/howto/maptitude/introduction/maptitude/

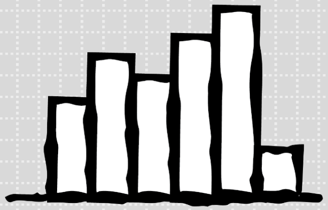


region	long	lat	group
'Iowa'	56.73	23.30	5
'Iowa'	47.93	23.83	5
'Iowa'	45.74	22.63	5
...	...	...	...

region	value
'Iowa'	0.78

which geom for which data?

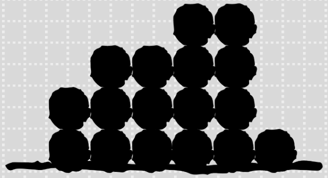




Histogram

`x=value`

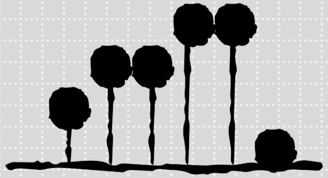
`geom_bar()`



Wilkinson dot plot

`x=group, y=count`

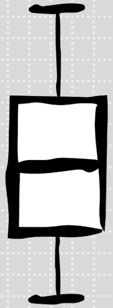
`geom_dotplot()`



Cleveland dot plot

`x=group, y=count`

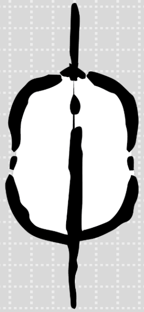
`geom_point()`



Box plot

`x=1, y=value`

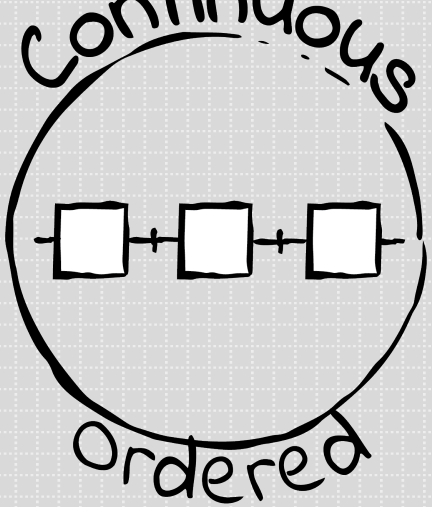
`geom_boxplot()`

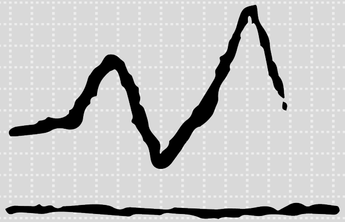


Violin plot

`x=1, y=value`

`geom_violin()`

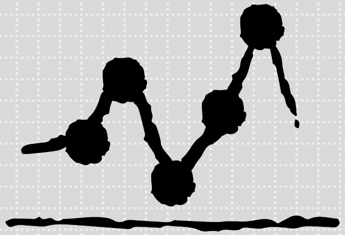




Line graph

```
x=index, y=value
```

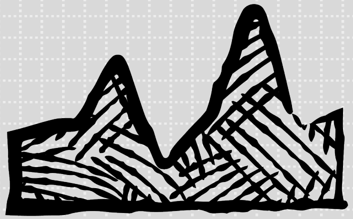
```
geom_line()
```



... with dots

```
x=index, y=value
```

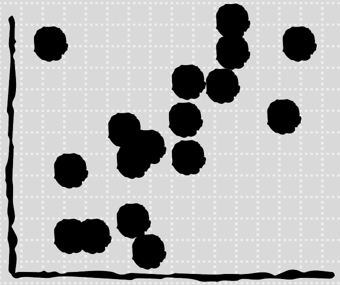
```
geom_line() +

geom_point()
```



Shaded area

```
x=index, y=value
```

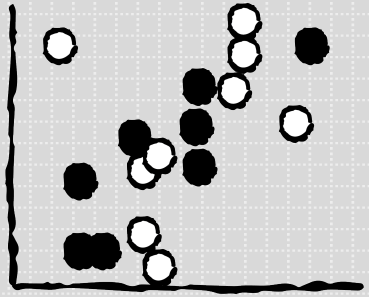
```
geom_area()
```



Scatter plot

```
x=x, y=y
```

```
geom_point()
```



Scatter plot

```
x=x, y=y,
colour=group
```

```
geom_point()
```

- Is your data ordered? (time)
- What are your variables? Are they categorical or continuous?