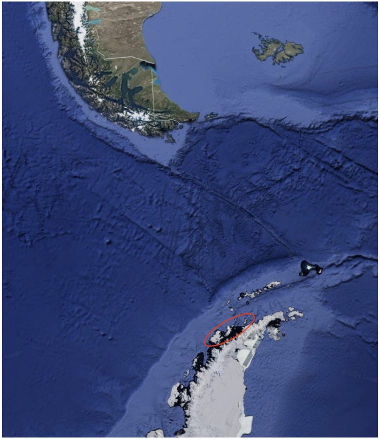
Multivariate Data and the Grammar of Graphics

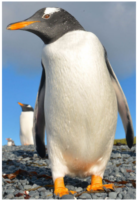
Dennis Sun
Stanford University
DATASCI 112

January 14, 2026

Palmer Penguins

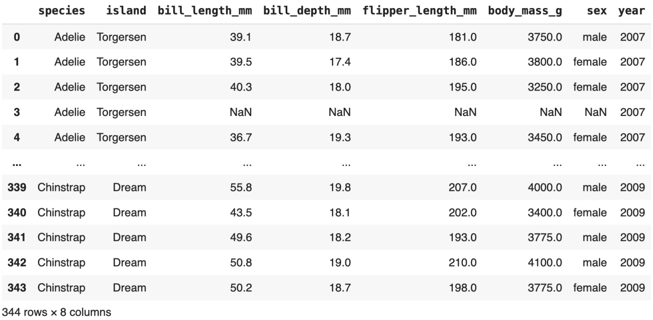
Today's Data: Penguins in the Palmer Archipelago, Antarctica.





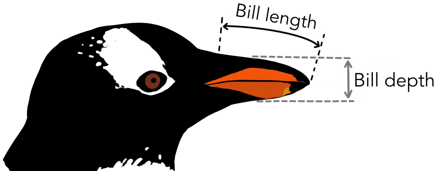
Palmer Penguins

```
import pandas as pd
df = pd.read_csv("https://datasci112.stanford.edu/data/penguins.csv")
df
```



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN	2007
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007
...	...	...	...	...	...	...	...	...
339	Chinstrap	Dream	55.8	19.8	207.0	4000.0	male	2009
340	Chinstrap	Dream	43.5	18.1	202.0	3400.0	female	2009
341	Chinstrap	Dream	49.6	18.2	193.0	3775.0	male	2009
342	Chinstrap	Dream	50.8	19.0	210.0	4100.0	male	2009
343	Chinstrap	Dream	50.2	18.7	198.0	3775.0	female	2009

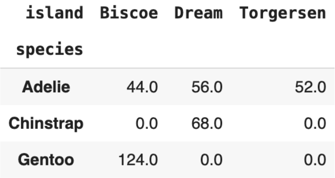
344 rows × 8 columns



Review

1 relationships between two categorical variables

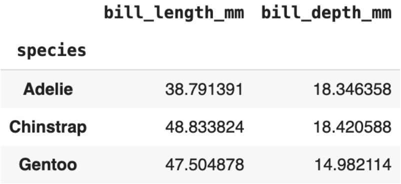
```
df[["species", "island"]].value_counts().unstack().fillna(0)
```



island	Biscoe	Dream	Torgersen
species			
Adelie	44.0	56.0	52.0
Chinstrap	0.0	68.0	0.0
Gentoo	124.0	0.0	0.0

2 relationships between categorical and quantitative variables

```
df.groupby("species")[["bill_length_mm", "bill_depth_mm"]].mean()
```



	bill_length_mm	bill_depth_mm
species		
Adelie	38.791391	18.346358
Chinstrap	48.833824	18.420588
Gentoo	47.504878	14.982114

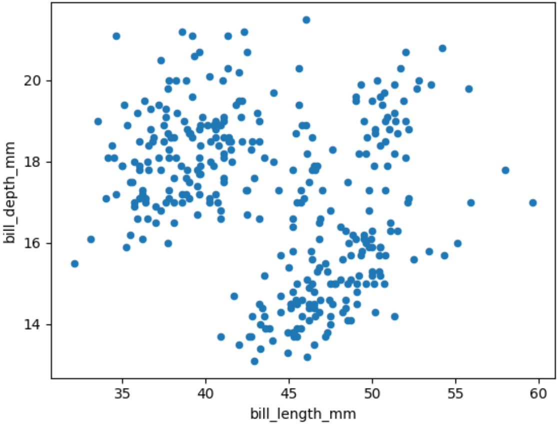
1. Relationships between Quantitative Variables

2. Multivariate Data

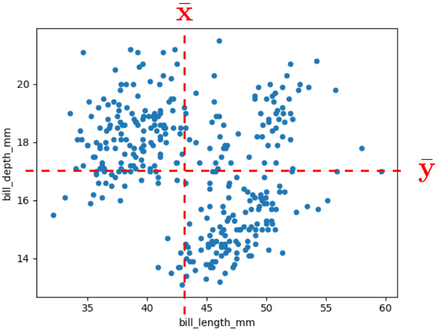
Visualizing the Relationship

The relationship between two quantitative variables can be visualized using a **scatterplot**.

```
df.plot.scatter(x="bill_length_mm", y="bill_depth_mm")
```



Summarizing the Relationship



The relationship between two quantitative variables $\mathbf{x}$ and $\mathbf{y}$ can be summarized using the **correlation coefficient** r.

$$r = \frac{\sum_{i=1}^{n} \frac{x_i - \bar{\mathbf{x}}}{\text{sd}(\mathbf{x})} \cdot \frac{y_i - \bar{\mathbf{y}}}{\text{sd}(\mathbf{y})}}{n-1}$$

```
x = df["bill_length_mm"]
y = df["bill_depth_mm"]
n = (~x.isnull() & ~y.isnull()).sum()

(((x - x.mean()) / x.std()) * ((y - y.mean()) / y.std())).sum() / (n - 1)
```

```
-0.2350528703555327
```

Correlation Coefficient

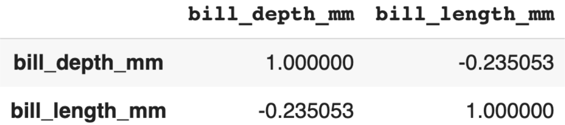
- A positive correlation means that as $\mathbf{x}$ increases, $\mathbf{y}$ tends to increase also.
- A negative correlation means that as $\mathbf{x}$ increases, $\mathbf{y}$ tends to decrease.
- The correlation coefficient r is always between -1 and 1.
- The closer the correlation coefficient is to ± 1, the stronger the relationship.

Since the correlation coefficient between bill length and bill depth is -0.235, bills that are longer tend to be less deep.

Correlation Coefficient

Of course, there's a built-in function for calculating r.

```
df[["bill_length_mm", "bill_depth_mm"]].corr()
```



	bill_depth_mm	bill_length_mm
bill_depth_mm	1.000000	-0.235053
bill_length_mm	-0.235053	1.000000

This is called the **correlation matrix**.

Why are the correlation coefficients on the diagonal equal to 1.0?

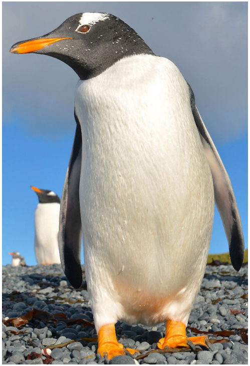
1. Relationships between Quantitative Variables

2. Multivariate Data

Beyond Two Variables

But wait! There were also different penguin species.

Adelie



Gentoo

Chinstrap

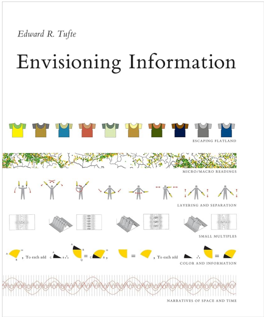
How do we incorporate another variable into our analysis?

Multivariate Data

“The world is complex, dynamic, multidimensional; the paper is static, flat. How are we to represent the rich visual world of experience and measurement on mere flatland?”

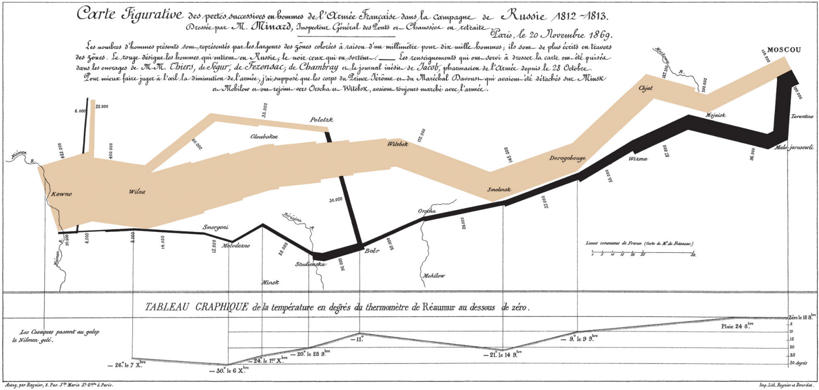
“Escaping this flatland is the essential task of envisioning information—for all the interesting worlds (physical, biological, imaginary, human) that we seek to understand are inevitably and happily multivariate in nature.”

– Edward Tufte, *Envisioning Information*



Map of Napoleon’s Russia Campaign

The French civil engineer Charles Joseph Minard (1781–1870) made the following visualization of Napoleon’s Russia campaign of 1812.

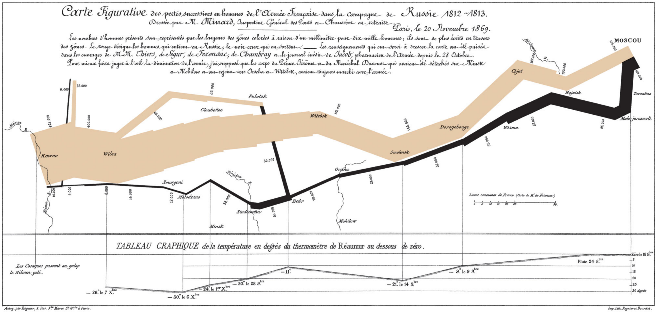

Carte Figurative des pertes successives en hommes de l’Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d’hommes présents sont représentés par les largeurs des zônes colorées à raison d’un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l’Armée depuis le 28 Octobre.
Pour mieux faire juger à l’œil la diminution de l’armée, j’ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l’armée.
MOSCOU
Niémen R.
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wixma
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Moskowa R.
Orscha
Mohilow
Bobr
Studienska
Bérézina R.
Minsk
Molodezno
Smorgoni
422.000
400.000
22.000
6.000
60.000
33.000
30.000
175.000
145.000
100.000
127.100
100.000
96.000
87.000
55.000
37.000
24.000
28.000
50.000
20.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France (Carte de M.r de Fezensac)
0 5 10 15 20 25 30.
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
– 26.° le 7 X.bre
– 30.° le 6 X.bre
– 24.° le 1er X.bre
– 20.° le 28 9.bre
– 11.°
– 21.° le 14 9.bre
– 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

Tufte calls this the “best statistical graphic ever drawn.”

Aesthetic Mappings

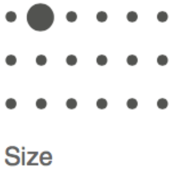
How do we visualize multivariate data on two-dimensional paper or screen?

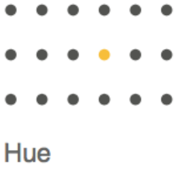
By mapping other dimensions in the data to other dimensions in the graphic.

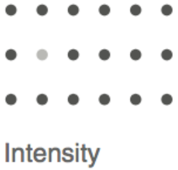

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
6.000
22.000
33.000
Polotzk
Gloubokoe
Witebsk
175.000
Niémen
422.000
400.000
60.000
Kowno
Wilna
Smorgoni
Molodezno
30.000
Bérézina
Studienska
Bobr
Minsk
Orscha
Mohilow
Smolensk
145.000
Dorogobouge
Wizma
Chjat
Mojaisk
MOSCOU
Tarantino
Malo-jarosewli
100.000
10.000
4.000
8.000
14.000
12.000
28.000
50.000
20.000
24.000
37.000
55.000
87.000
96.000
127.100
Moskowa
Lieues communes de France
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24.° le 1er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

- x ⟵ longitude
- y ⟵ latitude
- width ⟵ size of army
- color ⟵ direction of army
- y (line graph) ⟵ temperature
- x / text (line graph) ⟵ date

Aesthetics



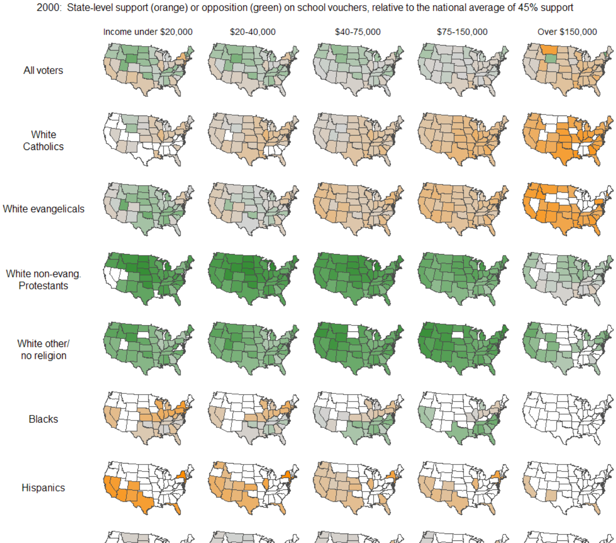




Which aesthetics are associated with quantitative variables?
Which are associated with categorical variables?

Facets

One way to pack more variables without overplotting is to show many small plots. (Tufte calls this “small multiples.”)


2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support
Income under $20,000
$20-40,000
$40-75,000
$75-150,000
Over $150,000
All voters
White Catholics
White evangelicals
White non-evang. Protestants
White other/ no religion
Blacks
Hispanics

Source: Gelman

Grammar of Graphics

The **grammar of graphics** says that every plot can be described by just a few components:

- aesthetic mappings
- geometric objects (e.g., points, lines, bars)
- ...and a few other things.

The ideal library generates a plot from a specification of the aesthetic mappings and the geometric object.

- x ⟵ longitude
- y ⟵ latitude
- width ⟵ size of army
- color ⟵ direction of army

Libraries for the Grammar of Graphics

Libraries that implement the grammar of graphics include `ggplot2` in R and `plotly` in Python.

Let’s try out `plotly` in a Colab!