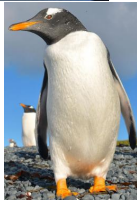


Lecture 5

Multivariate Data and the Grammar of Graphics

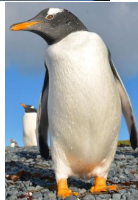
Palmer Penguins

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



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
```

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

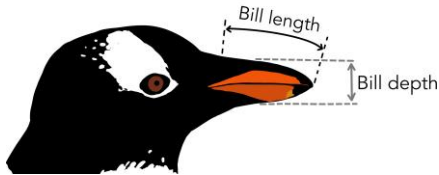
344 rows x 8 columns

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 x 8 columns



Review

- 1 relationships between two categorical variables

Review

- 1 relationships between two categorical variables

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


Review

① relationships between two categorical variables

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

species	island	
Gentoo	Biscoe	124
Chinstrap	Dream	68
Adelie	Dream	56
	Torgersen	52
	Biscoe	44

dtype: int64

Review

① relationships between two categorical variables

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

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

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	

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

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()
```

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

Visualizing the Relationship

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

Visualizing the Relationship

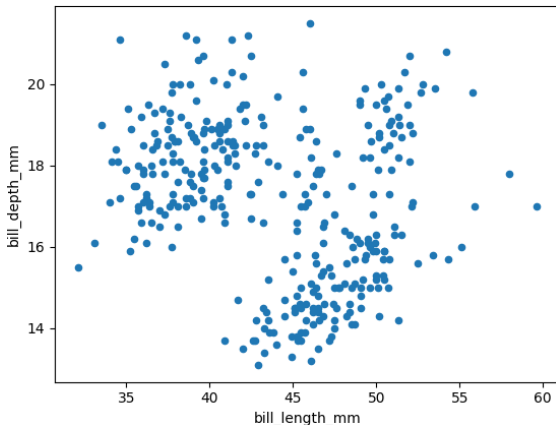
The relationship between two quantitative variables can be visualized using a **scatterplot**.

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

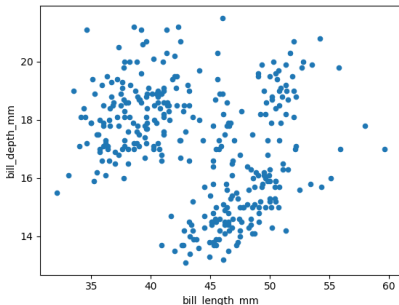
Visualizing the Relationship

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

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



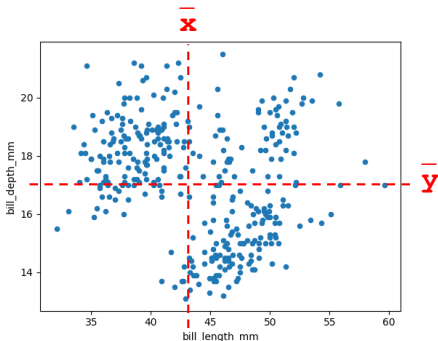
Summarizing the Relationship



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

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

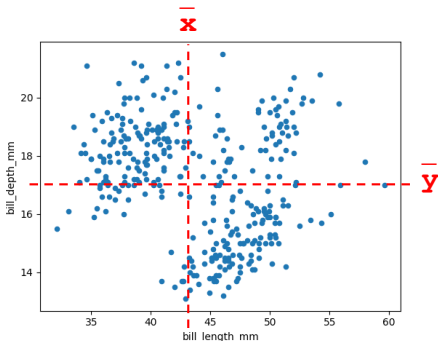
Summarizing the Relationship



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

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

Summarizing the Relationship



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

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

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

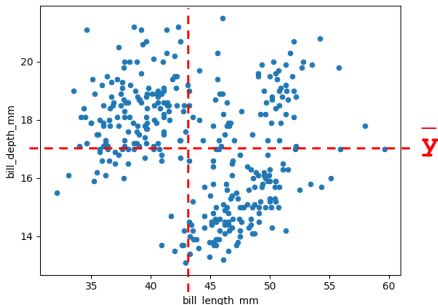
```
((x - x.mean()) / x.std()) * ((y - y.mean()) / y.std()).sum() / (n - 1)
```

`x.isnull()` → gives True for missing (NaN) values and False otherwise.

`~` → is the logical NOT operator, which inverts the Boolean values.

Therefore, `~x.isnull()` is True for valid (non-NaN) elements and False for missing ones.

Summarizing the Relationship



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

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

```
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.

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.

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 .

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.

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 ,

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()
```

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

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**.

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



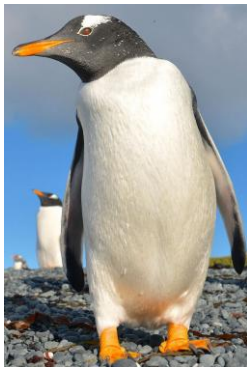
Chinstrap

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?”

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.”

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.”

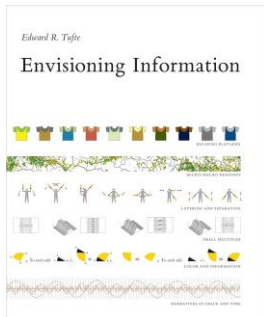
– Edward Tufte, *Envisioning Information*

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

- **What It Shows:**

Napoleon set out with **600,000 soldiers**, by the time he reached **Moscow**, **only about 100,000 remained**, and when he finally returned to **France**, **barely 10,000 soldiers survived**.

- **Importance in Data Science:**

It is an example of **multidimensional data visualization**.

Different variables such as **time, location, temperature, and quantity** (number of soldiers) are meaningfully combined in a single graphic.

It is a classic example of the concept of “**storytelling with data**” in information visualization design.

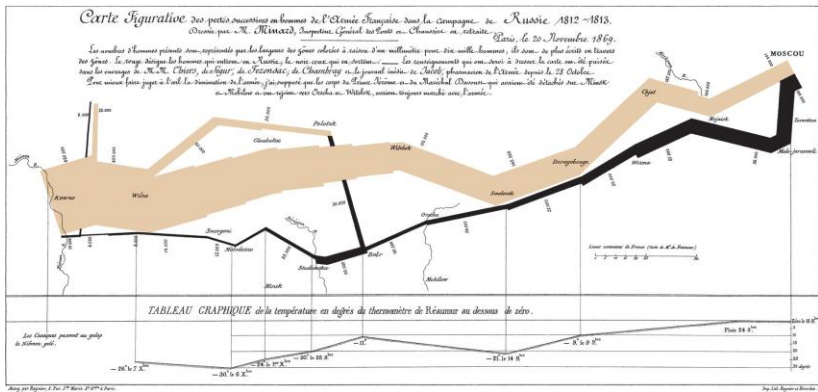
Edward Tufte calls Charles Minard's map of Napoleon's Russian campaign the “best statistical graphic ever drawn” because:

1. **Multidimensional Data in One Graphic** – It shows army size, location, direction, and time simultaneously.
2. **Rich Meaning, Simple Design** – Data-dense and free of unnecessary decoration.
3. **Storytelling** – Conveys the army's tragic losses and journey, not just numbers.
4. **Combines Time and Space** – Allows viewers to understand both geography and chronology at a glance.

Map of Napoleon's Russia Campaign



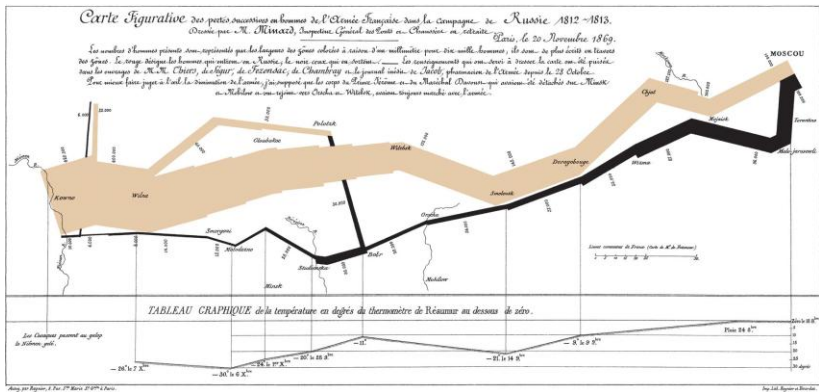
The French civil engineer Charles Joseph Minard (1781–1870) made the following visualization of Napoleon’s Russia campaign of 1812.



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.



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

Aesthetic Mappings

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

Aesthetic (visual mapping)

The mapping of variables in the data to visual properties of the plot. In other words, it defines which variable controls visual elements such as **color, size, position, or shape**.

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.

Aesthetics for quantitative variables:

Size: The larger the value, the larger the point or symbol.

Intensity / Lightness: The color becomes darker or lighter as the value changes.

Position (x, y): The variable's value determines where it is placed on the plot.

Aesthetics for categorical variables:

Hue: Different categories are shown with different colors.

Shape: Each category can be represented by a different marker shape (circle, triangle, etc.).

Line type: Categories can also be distinguished by line styles (solid, dashed, dotted).

Summary:

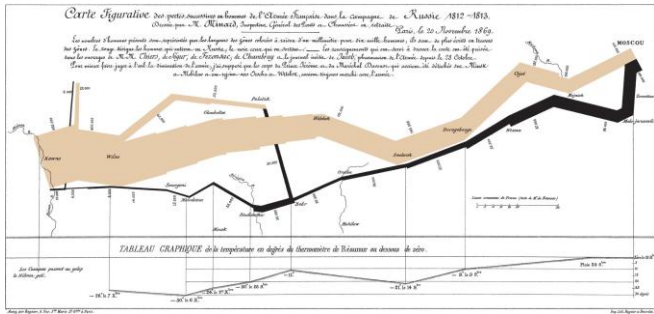
Quantitative → size, intensity, position

Categorical → hue, shape, line type

Aesthetic Mappings

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

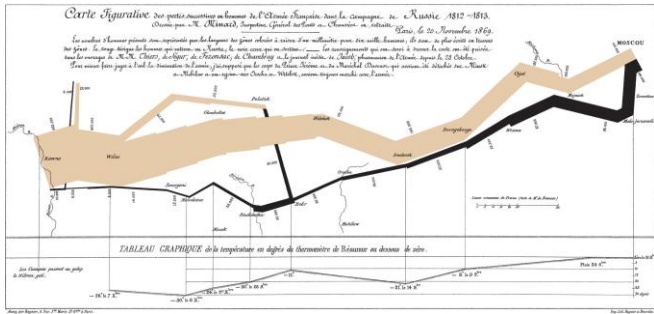
By mapping other dimensions in the data to other dimensions in the graphic.



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.

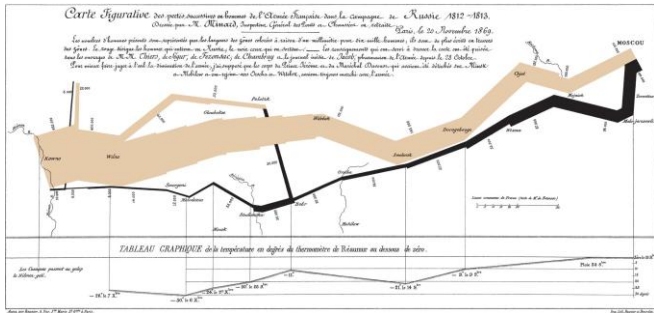


- $x \leftarrow$ longitude
- $y \leftarrow$ latitude

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.

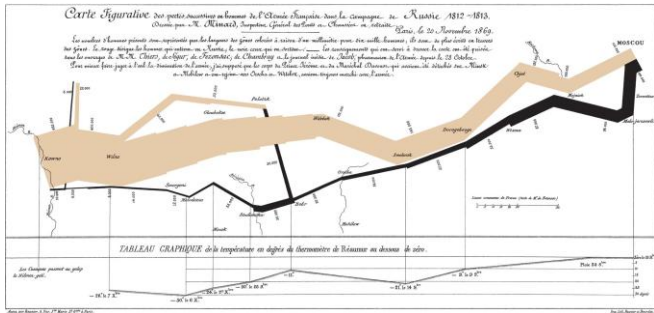


- `x` ← longitude
- `y` ← latitude
- `width` ← size of army

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.

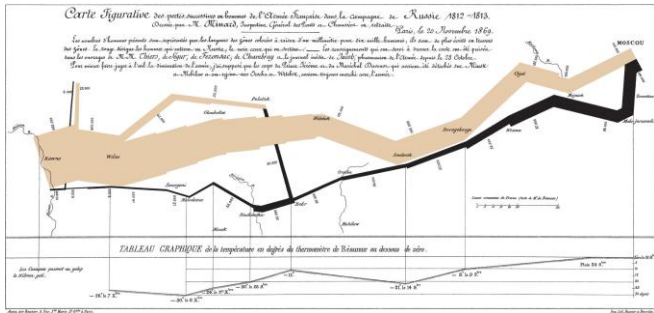


- x ←– longitude
- y ←– latitude
- width ←– size of army
- color ←– direction of army
- y (line graph) ←– temperature

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.



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

Aesthetics



Size



Hue



Intensity

Which aesthetics are associated with quantitative variables?

Which are associated with categorical variables?

Quantitative (ölçülebilir ve sıralanabilir) → size, intensity, position

Categorical → hue, shape, line type (grupları veya sınıfları temsil eder ayrı ve niteliksel görsel farklarla (renk tonu, şekil, çizgi tipi) gösterilir)

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



facet, veriyi alt gruplara ayırıp her grup için küçük bir grafik çizme yöntemidir.

Source: Gelman

Grammar of Graphics

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

grafik oluşturmak belirli kurallar ve bileşenler

Grammar of Graphics

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

- aesthetic mappings

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)

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.

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.

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.



Je voudrais un line plot, where...

- $x \leftarrow$ longitude enlem/boylam
- $y \leftarrow$ latitude
- width \leftarrow size of army
- color \leftarrow 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!

