

CMPU 100 · Programming with Data

Data Exploration and Visualization

Class 10



Assignment 3 due on Thursday

Exam 1 will be graded as soon as I can – hopefully by Thursday.

Across sections of CMPU 100, several students still need to take the exam.

We've seen that tables give us a powerful and flexible way to represent many kinds of data.

And there are a staggering number of publicly available datasets that we can load from CSV files!

Data Is Plural

... is a newsletter of useful/curious datasets, published by [Jeremy Singer-Vine](#). There have been [400 editions](#), dating from [October 21, 2015](#) to [August 27, 2025](#). To receive future editions, sign up here:

Subscribe

Recent Editions

[2025.08.27](#) • Congressional campaign policy platforms, Billboard hits, EU antitrust cases, Texas oil and gas, and 100,000+ wines.

[2025.07.16](#) • Landfills, car crash datasets, AI legal flubs, British literary prizes, and subway art.

[2025.05.21](#) • Unemployment insurance, FiveThirtyEight's public data, corporate contracts, positions of power in Italy, and state formation/dissolution.

[2025.05.14](#) • Deportation records, European workforces, California ghost guns, US dams, and what the nose knows.

[2025.04.30](#) • Refugee and asylum policies, tens of millions of flights, US sewer overflow sites, previously unmapped waterways, and canoe marathons.

Read Full Archive

Some Nice Things People Have Said

- “a treasure trove of interesting datasets” — [Julia Silge](#)
- “delivers exactly what it promises, it's delightful” — [Simon Willison](#)
- “required reading for anyone interested in data journalism” — [Julia Angwin](#)
- “consistently fascinating” — [Robin Sloan](#)
- “the only newsletter I open immediately” — [Paul Ford](#)

data-is-plural.com

In this part of the course, we'll go deeper into how to analyze & visualize the contents of tables – and we'll expand from “clean” data to “messy” real-world datasets.

Arctic climate data

Lots of interesting climate datasets are available from the National Snow and Ice Data Center (NSIDC) at the University of Colorado, Boulder:

nsidc.org/data/explore-data



The Arctic region is often defined as the area where the average temperature for the warmest month is below 10°C.

*Arto Vitikka,
Arctic Centre,
University of Lapland*

Data isn't abstract; it's often the result of efforts "in the field".

"Constructed in the 1960s, the USS Pargo was in operation when the United States and Russia agreed to declassify Arctic observations for scientific study."

nsidc.org/data/ewg



USS Pargo in the Arctic, September 1993



→ Notebook: *Setup & Data cleaning*

Categorical

Year	Month	Air temperature (C)	Sea level pressure (mbar)	Precip (mm)
1873	9	0.4	9999.9	15
1873	10	-5.3	9999.9	34
1873	11	-9.4	9999.9	30
1873	12	999.99	9999.9	29
1874	1	-29.6	9999.9	9
1874	2	-19.6	9999.9	22
1874	9	0.1	1010.7	68
1874	10	-5.4	1002.7	24
1874	11	-8	1010.5	15
1874	12	-8.4	1005.1	69

... (1355 rows omitted)

Numerical

Year	Month	Air temperature (C)	Sea level pressure (mbar)	Precip (mm)
1873	9	0.4	9999.9	15
1873	10	-5.3	9999.9	34
1873	11	-9.4	9999.9	30
1873	12	999.99	9999.9	29
1874	1	-29.6	9999.9	9
1874	2	-19.6	9999.9	22
1874	9	0.1	1010.7	68
1874	10	-5.4	1002.7	24
1874	11	-8	1010.5	15
1874	12	-8.4	1005.1	69

... (1355 rows omitted)

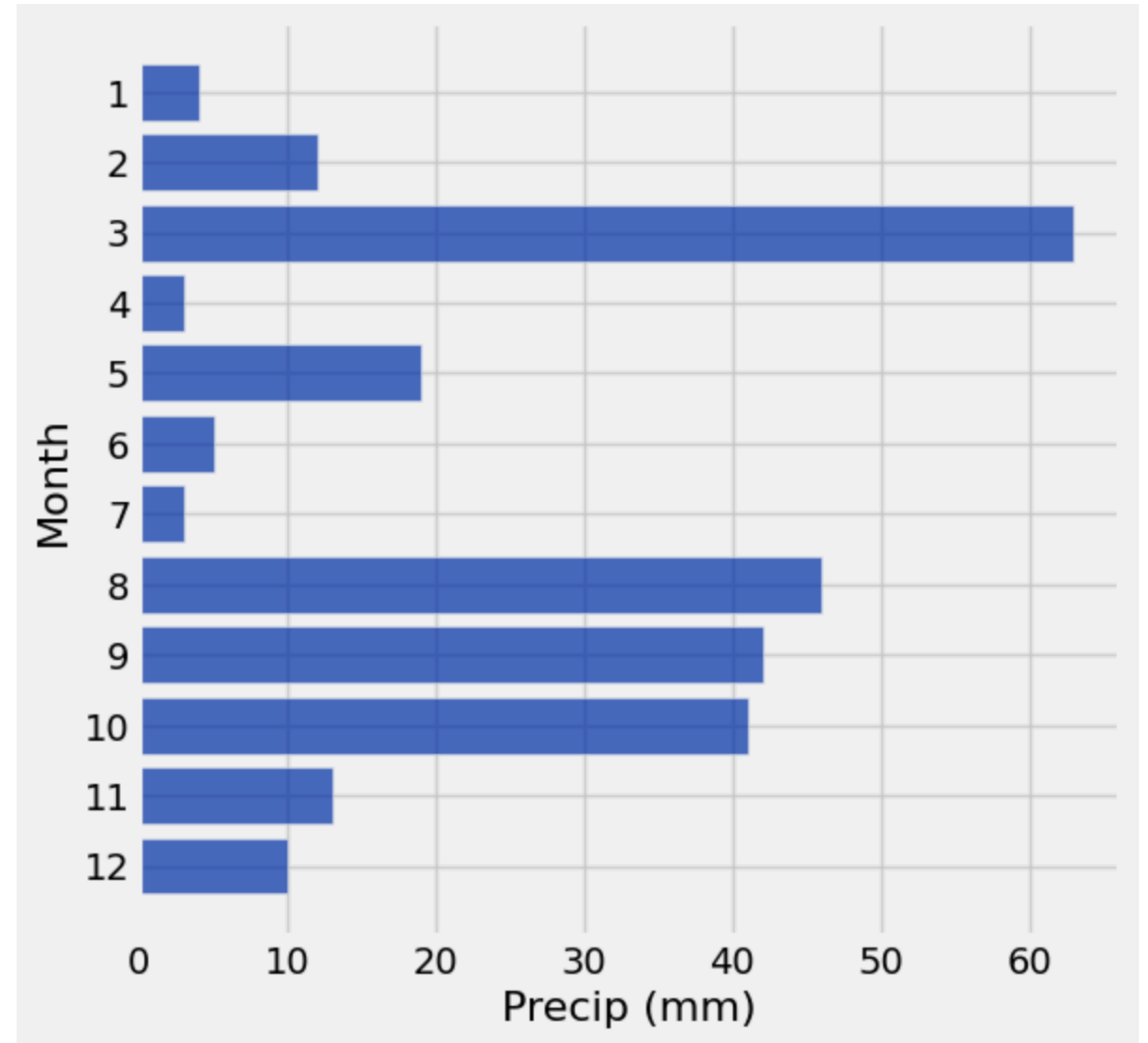
<i>Variable type</i>	<i>Properties</i>	<i>Examples</i>
<i>Numerical</i>	<p>Values are numbers</p> <p>Scale is ordered</p> <p>Mathematical operations (mean, sum) are meaningful</p>	<p>Air temperature (°C)</p> <p>0 . 1, -5 . 4, -8</p>
<i>Categorical</i>	<p>Values are categories</p> <p>May or may not have an ordering</p> <p>Mathematical operations are <i>not</i> meaningful</p>	<p>Months</p> <p>Option 1:</p> <p>"Jan", "Feb", "Mar", ...</p> <p>Option 2:</p> <p>1, 2, 3, ...</p>

<i>Variable type</i>	<i>Properties</i>	<i>Examples</i>
<i>Numerical</i>	Values are numbers Scale is ordered Mathematical operations (mean, sum) are meaningful	Air temperature (°C) 0.1, -5.4, -8
<i>Categorical</i>	Values are categories May or may not have an ordering Mathematical operations are <i>not</i> meaningful	Months Option 1: "Jan", "Feb", "Mar", ... Option 2: 1, 2, 3, ...

What's wrong with taking the sum here?

→ Notebook: *Visualizing temperature & Visualizing precipitation*

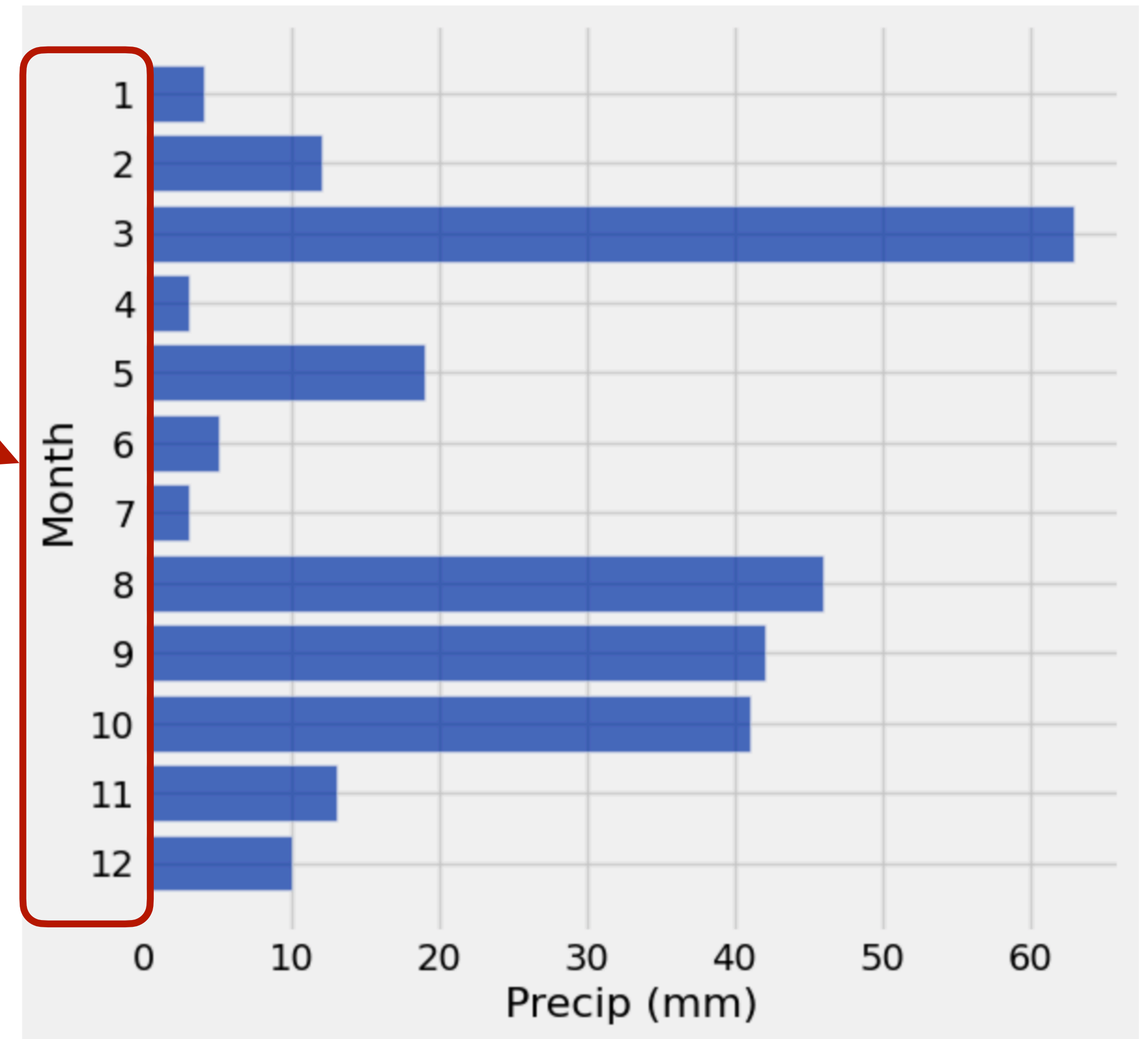
A *bar chart* shows the relationship between a categorical and a numerical variable.



```
t.barh("Month", "Precip (mm)")
```

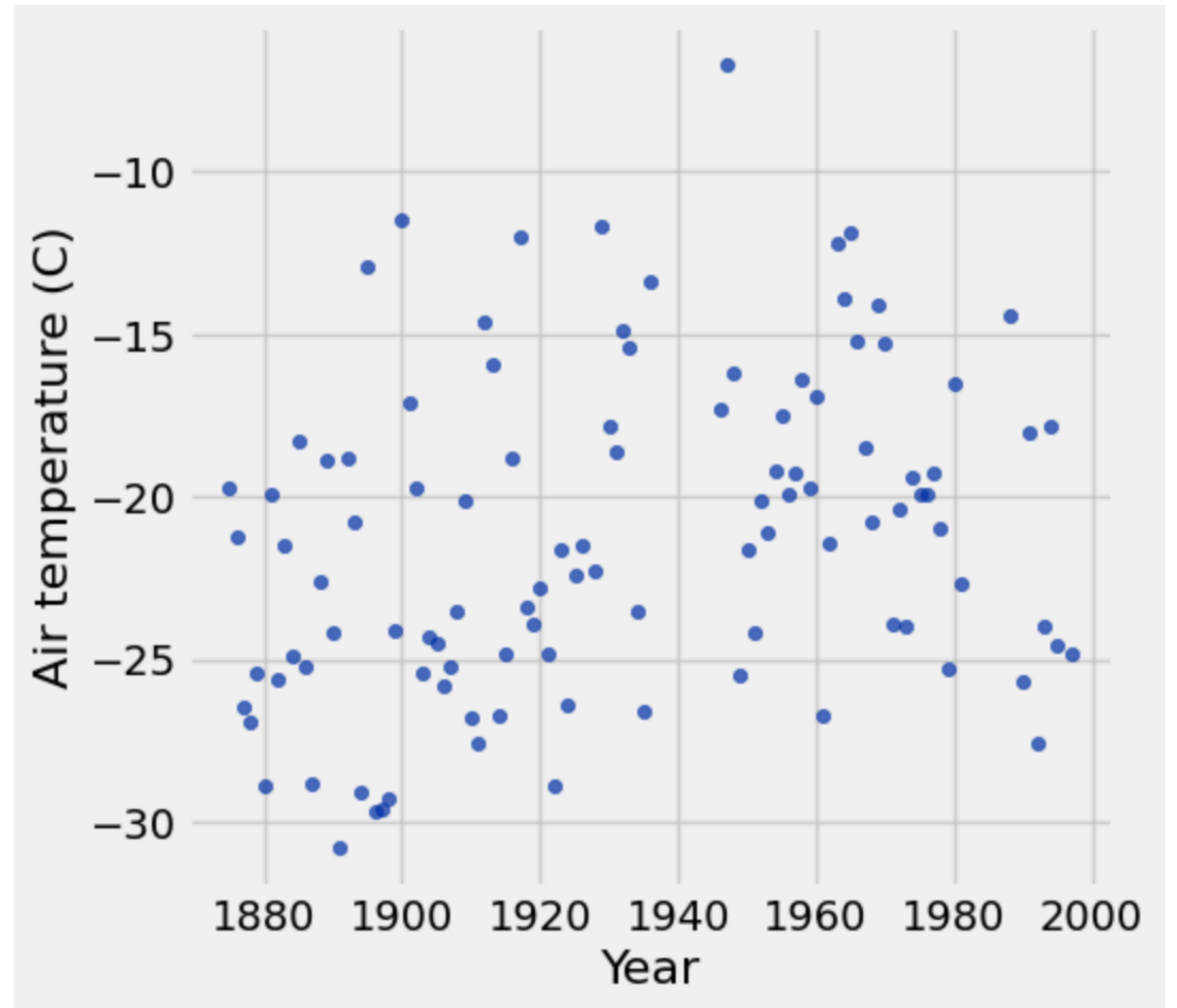
In this class, we'll typically put the categorical variable on the vertical axis.

A *bar chart* shows the relationship between a categorical and a numerical variable.



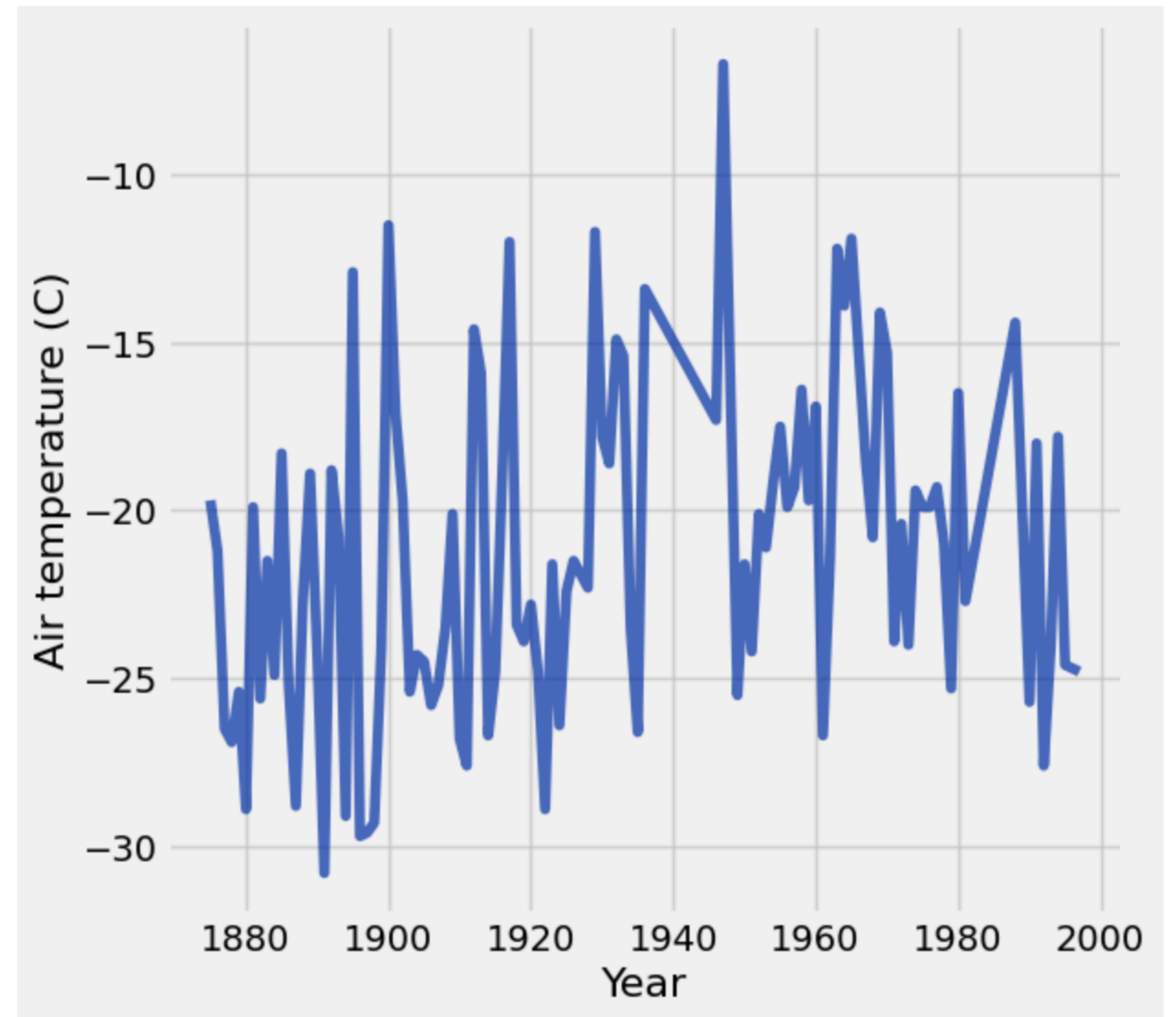
```
t.barh("Month", "Precip (mm)")
```

A *scatter plot* shows the relationship between two numerical variables.



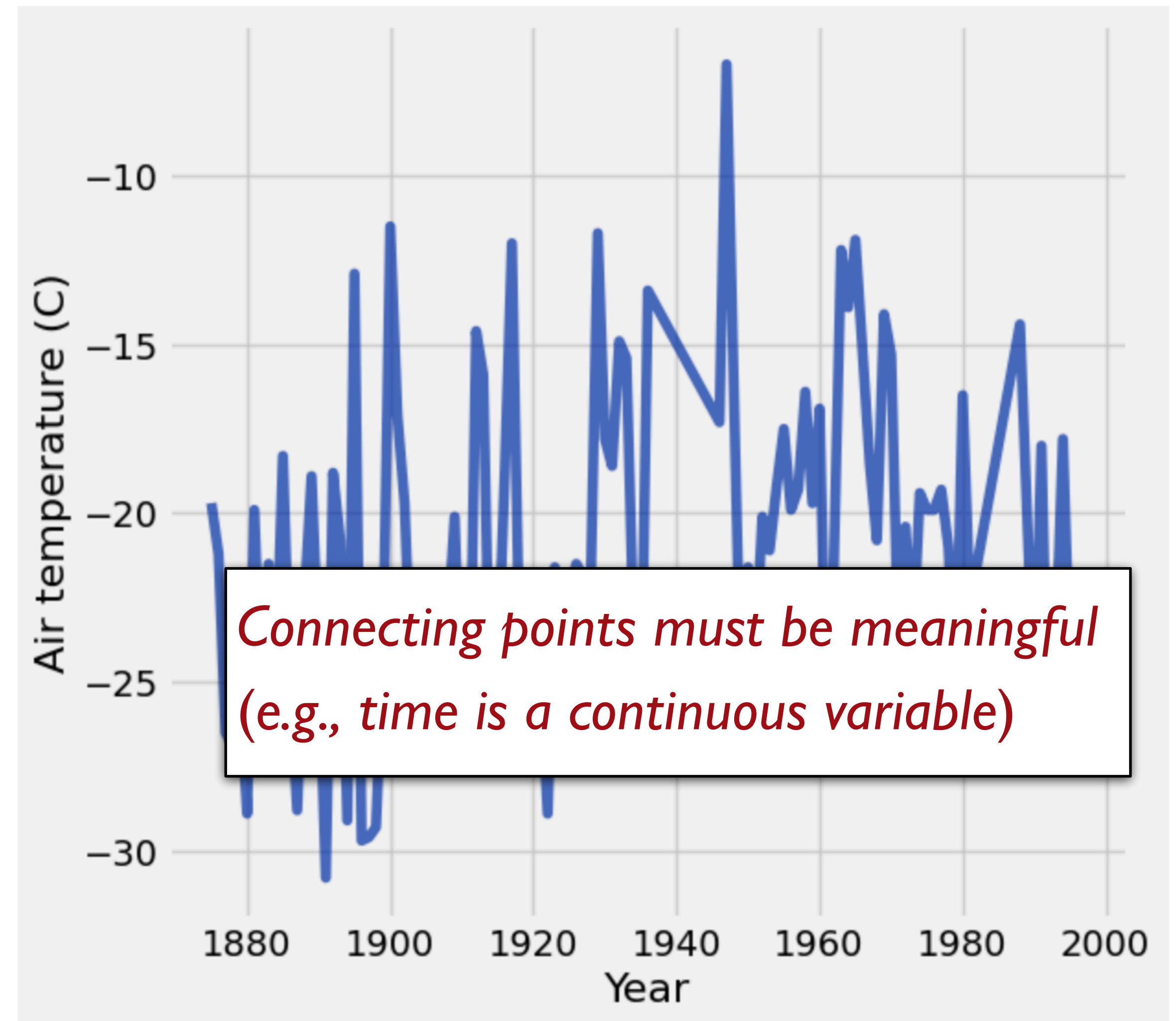
```
t.scatter("Year", "Air temperature (C)")
```

A *line plot* shows a trend – typically chronological – between two numerical variables.



```
t.plot("Year", "Air temperature (C)")
```

A *line plot* shows a trend – typically chronological – between two numerical variables.



```
t.plot("Year", "Air temperature (C)")
```

There are more types of visualizations – some of which we'll see later in the course!

