Lesson Objectives

1. Frame. What changes when there is no y to predict.
2. Apply. Fit PCA on scaled columns with `prcomp` and read the loadings, scores, and proportion of variance.
3. Mean. What a principal component, a loading, and a score are, and how the loadings are chosen.
4. Communicate. Rank and describe states with one or two component scores, and say how far to trust them.

Bottom Line Up Front

1. **No y, no way to check.** Unsupervised learning has no right answer. You judge the result by whether it makes sense.
2. **A principal component is a new column.** A weighted sum of the old columns, with weights picked so the scores spread out as much as possible.
3. **Scale first.** On USArrests, Z_1 is overall crime (62% of the spread) and Z_2 is how urban (another 25%).

Key Definitions

Unsupervised learning: only x's, no response. Find structure: which columns move together, which rows group together.

Principal component Z_m: a new column, a weighted sum of the scaled columns.

Loading ϕ_{jm}: column j's weight in Z_m. In R: `pca$rotation`.

Score z_{im}: one state's number in Z_m. In R: `pca$x`.

Proportion of variance: $\text{Var}(Z_m)$ as a share of the total. Scaled columns each have variance 1, so the total is the number of columns.

Principal Components

Each Z	$Z_1 = \phi_{11}\,\text{Murder} + \phi_{21}\,\text{Assault} + \phi_{31}\,\text{UrbanPop} + \phi_{41}\,\text{Rape}$
Weights	$\phi_{11}, \ldots, \phi_{41}$ chosen so Z_1 is as spread out as possible
Rules	squares sum to 1: $\phi_{11}^2 + \phi_{21}^2 + \phi_{31}^2 + \phi_{41}^2 = 1$
	Z_2: as spread out as possible, uncorrelated with Z_1; and so on
Fit	`pca <- prcomp(arrests |> select(-state), scale. = TRUE)`
	`pca$rotation` loadings, `pca$x` scores, `summary(pca)` proportion of variance
Same answer	`eigen(cor(...))`: eigenvectors are the loadings, eigenvalues are $\text{Var}(Z_m)$

No y: Crime in 50 States

`USArrests` ships with R: 1973 arrests per 100,000 residents for Murder, Assault, and Rape, plus UrbanPop, the percent of residents in cities. One row per state. No response column.

```
arrests |> select(-state) |> cor() |> round(2)

         Murder Assault UrbanPop Rape
Murder     1.00    0.80     0.07 0.56
Assault    0.80    1.00     0.26 0.67
UrbanPop   0.07    0.26     1.00 0.41
Rape       0.56    0.67     0.41 1.00
```

1. In Lesson 26 the leaderboard scored your model against the truth. What can't you do here, and what does ISLR mean by "there is no way to check our work"?

2. The three crimes correlate 0.56 to 0.80. UrbanPop barely moves with Murder (0.07). What does that suggest about how many numbers we need to describe a state?

How Can We Rank the States by Crime?

Each crime crowns a different state: Georgia (Murder), North Carolina (Assault), Nevada (Rape). We want one crime number per state.

Try 1: `added <- arrests |> mutate(crime = Murder + Assault + Rape)`

```
  state          Murder Assault UrbanPop  Rape crime
1 Florida          15.4     335       80  31.9  382.
2 North Carolina   13       337       45  16.1  366.
3 Maryland         11.3     300       67  27.8  339.

cor(added$crime, added$Assault)
[1] 0.9966603
```

3. Why is Try 1 really an Assault ranking?

Try 2: scale each crime, $(x - \bar{x})/s$, then `crime = (Murder + Assault + Rape) / 3`

```
  state      Murder Assault UrbanPop  Rape crime
1 Florida     1.75    1.97        80  1.14  1.62
2 Nevada      1.01   0.975        81  2.64  1.54
3 Alaska     0.508    1.11        48  2.48  1.37
```

4. Try 2 fixes the units. What did we still decide ourselves?

Fit PCA

`pca <- prcomp(arrests |> select(-state), scale. = TRUE)`. R made more crime negative, so we flipped PC1: `pca$rotation[, 1] <- -pca$rotation[, 1]` and the same for `pca$x[, 1]`.

`pca$rotation`, the loadings

```
               PC1        PC2        PC3         PC4
Murder   0.5358995 -0.4181809  0.3412327  0.64922780
Assault  0.5831836 -0.1879856  0.2681484 -0.74340748
UrbanPop 0.2781909  0.8728062  0.3780158  0.13387773
Rape     0.5434321  0.1673186 -0.8177779  0.08902432
```

```
summary(pca)
```

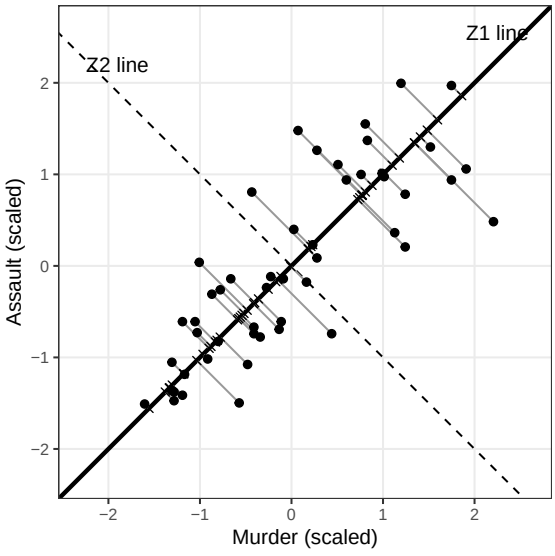
```
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

5. Write Z_1 with its loadings. Which columns drive it, and what would you name it?

6. Name Z_2 from its loadings.

7. We flipped PC1's signs. Why is that allowed?

8. Z_1's proportion of variance is 0.6201. Where do 2.48 and 4 come from, and why is the share 62%?



PCA on just two columns. Dots: states. Gray: each state dropped onto the solid line at a right angle. ×: where it lands.

9. What is a state's Z_1 score in this picture, and why does PCA pick this line over every other line through the middle?

Back to the Ranking

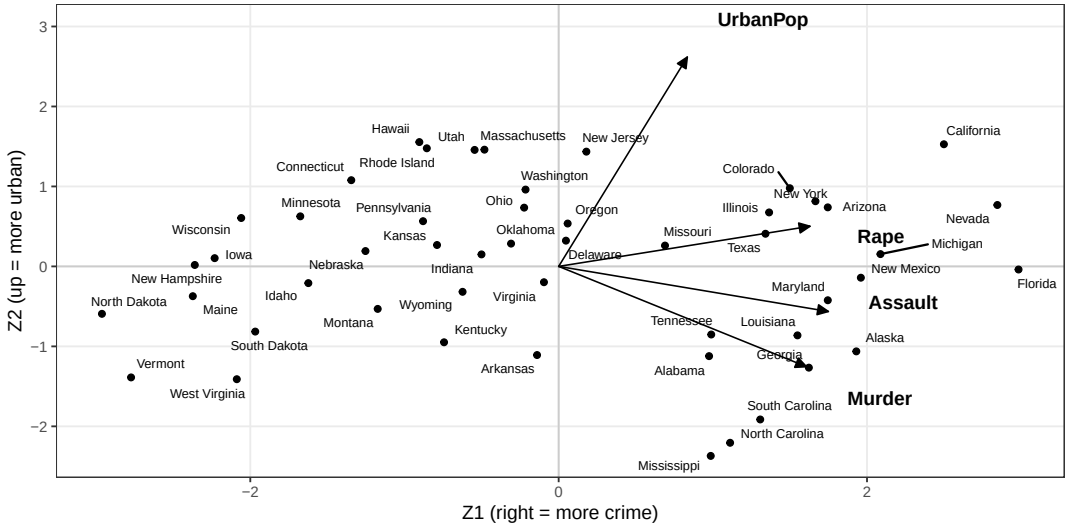
```
scores |> mutate(equal_crimes = averaged$crime) |> slice_max(PC1, n = 5)

  state        PC1 equal_crimes
1 Florida     2.98         1.62
2 Nevada      2.85         1.54
3 California  2.50         1.20
4 Michigan    2.09         1.16
5 New Mexico  1.96         1.12

cor(scores$PC1, averaged$crime)
[1] 0.985298
```

10. PCA nearly matches equal weights on the crimes. So what did PCA buy us?

Every state on Z_1 and Z_2. Arrows: each column's loadings, stretched 3 times.



11. California and Mississippi both score high on Z_1. Use Z_2 to tell them apart.

12. A governor wants a crime ranking from your Z_1. In two sentences, what is it, and how far should they trust it?

Key Takeaway

With no y there is nothing to check against. PCA builds new columns as weighted sums of the scaled columns, with weights picked so the scores spread out as much as possible, each new column uncorrelated with the ones before. On USArrests, Z_1 is overall crime and Z_2 is how urban: two numbers keep 87% of four.