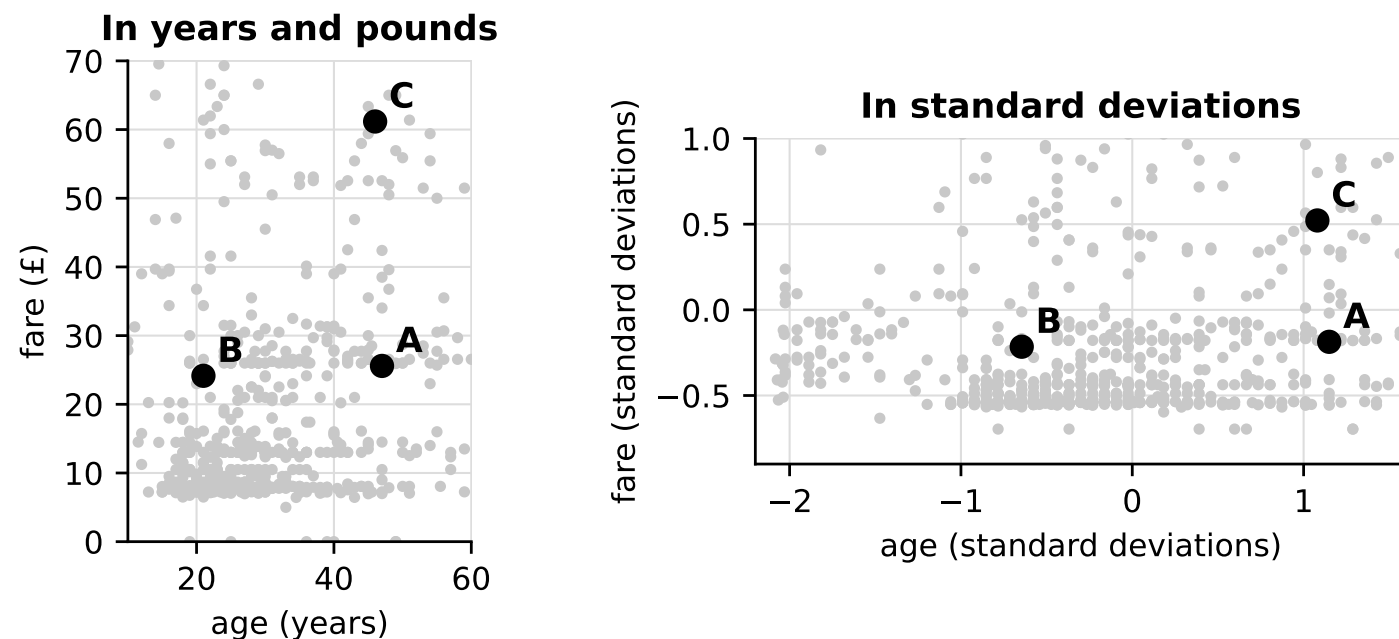


DATA 202 — Scaling, Encoding and Pipelines · Week 9 Monday

Same passengers as today's slides: the 1,043 *Titanic* passengers with age, fare and port recorded. Your professor will tell you when to do each part.

1 · Who is closest?



passenger	class	age	fare
A — Colley, Mr. Edward	1st	47	£25.59
B — Davies, Mr. John	3rd	21	£24.15
C — Chaffee, Mr. Herbert	1st	46	£61.18

1. **In years and pounds** (left plot): draw the lines A–B and A–C. Closer to A: _____
2. **In standard deviations** (age ÷ 14.5, fare ÷ 50.2): A–B differ by _____ in age and _____ in fare; A–C by _____ and _____. Closer to A: _____

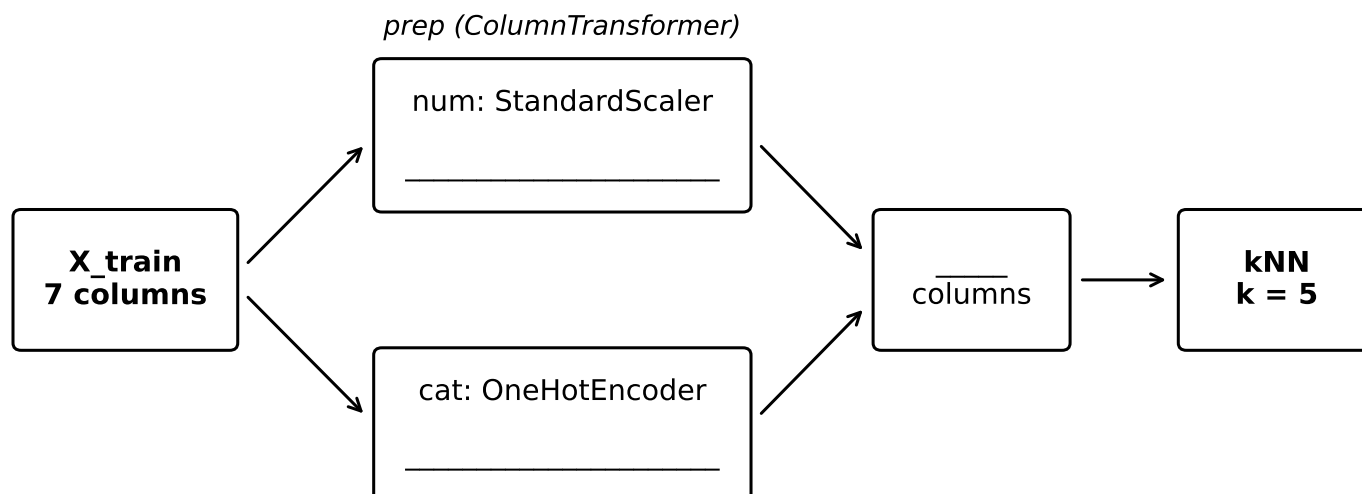
2 · Encode four passengers

The three port columns are `embarked_Cherrybourg`, `embarked_Queenstown` and `embarked_Southampton`.

sex	embarked	sex_male	Cherrybourg	Queenstown	Southampton
male	Southampton	_____	_____	_____	_____
female	Cherrybourg	_____	_____	_____	_____
male	Queenstown	_____	_____	_____	_____
female	Southampton	_____	_____	_____	_____

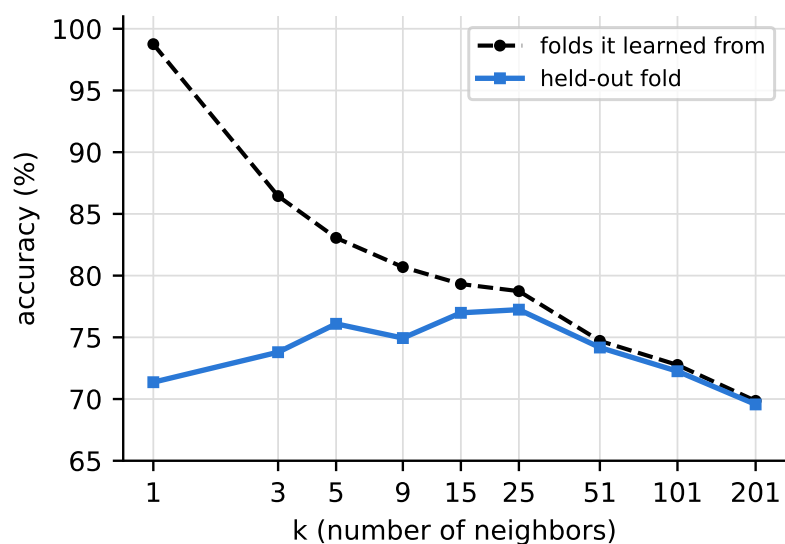
Why is there no `sex_female` column? _____

3 · Follow the pipe



1. In each branch, write the columns that go in.
2. How many columns come out of `prep` and into `kNN`? _____

4 · Choose k



1. Accuracy of the whole pipeline for each `k`: on the folds it learned from, and on the held-out fold.
2. **Circle** the best `k`.
3. **Label** the overfitting side and the underfitting side.
4. `k = 1`: learned-from folds _____ %, held-out fold _____ %.