DATACAMP LIVE CODE-ALONG

Machine Learning for Healthcare

Patient Segmentation with AI-Assisted Python Coding

Maria Eugenia Inzaugarat

Senior Data Scientist · Cofounder at Insight Delta · DataCamp Instructor

Wednesday, April 8, 2026

Agenda

1. Intro: Clustering & health data basics
2. Code-along: Build a patient segmentation model with AI-assisted coding
3. Questions

What Is Clustering?

The Idea

- Unsupervised ML — no labels needed
- Group similar patients together
- Objects in the same cluster are more alike than those in other clusters
- Discovers hidden structure in data

Why Distance Metrics Matter

- Euclidean distance works for numeric data only
- One-hot encoding inflates dimensionality for categoricals
- Gower distance handles mixed data natively
- Wrong metric → meaningless clusters

Our Dataset: Patient Segmentation

2,000 Records · 16 Features

- Demographics: Age, Gender, State, City
- Clinical: BMI, Number of Chronic Conditions, Primary Condition
- Utilization: Annual Visits, Days since last visit
- Financial: Average billing amount, Insurance type

Key Challenge: Mixed Data

- Numeric features: Age, BMI, billing, visits
- Categorical features: Gender, Insurance type, Primary condition
- K-Means + one-hot encoding won't work properly
- We'll use Gower distance + hierarchical clustering

Health Data: What to Check First

Before modeling, always inspect your healthcare dataset for these key considerations:

- ✔ Handle missing values to avoid biased or misleading patterns
- ✔ Ensure features are on comparable scales so no variable dominates the model
- ✔ Account for mixed data types to prevent incorrect distance calculations
- ✔ Remove irrelevant information to reduce noise and improve clustering quality
- ✔ Understand distributions to detect skewness, imbalance, and outliers
- ✔ Critically evaluate AI suggestions to avoid incorrect modeling approaches

What We'll Build Today

We'll segment 2,000 patients using Gower distance and hierarchical clustering — with AI-assisted Python coding throughout.

1. Load & explore: shape, dtypes, missing values, value counts, statistical summary
2. Clean: fill missing values, drop non-informative columns
3. EDA: histograms, count plots, and correlation heatmap
4. First attempt: ask AI for clustering — check if it handles mixed data well
5. Alternative approach: Gower distance + hierarchical clustering with weighted linkage
6. Silhouette scores, dendrogram, cluster profiles & patient segment narratives

Code-Along Steps

A Create a DataCamp account if you haven't already

B Open this link: bit.ly/healthcare-ml

C Code along!

AI-Assisted Coding: Prompt Tips

✕ VAGUE PROMPT

"Analyze the data"

✓ SPECIFIC PROMPT

"Using seaborn, create a boxplot of 'charges' grouped by 'smoker' status. Title it 'Insurance Charges by Smoking Status' and use the 'Set2' palette."

Tip: Include column names, library preferences, and desired output. Always review AI-generated code before running it.

Let’s Code!

Follow along in your DataLab notebook