Είμαι τριτοετής φοιτήτρια και παρακολουθώ ένα εργαστήριο ανάλυσης δεδομένων. Ο καθηγητής μας είπε πως θα πρέπει να βρούμε ένα dataset για να επεξεργαστούμε κατά τη διάρκεια του εξαμήνου, ξεκινώντας από καθάρισμα, περιγραφικά στατιστικά, διερευνητική ανάλυση, οπτικοποίηση, και εκπαίδευση ενός απλού γραμμικού μοντέλου (παλινδρόμηση ή ταξινόμηση). Είπε πως το dataset θα πρέπει να είναι 500-100000 γραμμές (οκ και παραπάνω) και πάνω από 8-15 στήλες, να έχει ποικιλία τύπων μεταβλητών, να είναι tabular και να μην είναι τετριμμένο (πχ. Iris, Titanic, California Housing). Δε ξέρω τι πεδίο θα με ενδιέφερε και από πού να ξεκινήσω. Ιδέες για datasets και πεδία εφαρμογής; Απάντησε συνοπτικά και με links.

Αφού θες **tabular, μεσαίο–μεγάλο μέγεθος και όχι τετριμμένο**, ξεκίνα από αποθετήρια που έχουν *καθαρά αλλά ρεαλιστικά* δεδομένα:

📦 Γενικές πλατφόρμες datasets (πολλές επιλογές)

- **Kaggle Datasets**
  👉 https://www.kaggle.com/datasets ↗
  (φίλτρα για size, domain, tabular)
- **UCI Machine Learning Repository**
  👉 https://archive.ics.uci.edu ↗
  (κλασικά αλλά όχι “toy”)
- **OpenML**
  👉 https://www.openml.org/search?type=data ↗

📊 Καλές θεματικές για project εξαμήνου (με έτοιμα μεγάλα datasets)

🏠 Οικονομία / Αγορές

- House sales, Airbnb listings, used cars
  - 👉 https://www.kaggle.com/datasets?search=real+estate ↗
  - 👉 https://www.kaggle.com/datasets?search=airbnb ↗

🩺 Υγεία

- Hospital outcomes, insurance costs, patient readmissions
  - 👉 https://www.kaggle.com/datasets?search=healthcare ↗
  - 👉 https://archive.ics.uci.edu/ml/datasets.php?task=classification ↗

🌍 Κοινωνικά / Ανάπτυξη

- **World Bank Open Data**
  - 👉 https://data.worldbank.org ↗

🚦 Μεταφορές / Πόλεις

- Traffic accidents, bike sharing, public transport usage
  - 👉 https://www.kaggle.com/datasets?search=traffic ↗
  - 👉 https://www.kaggle.com/datasets?search=bike+sharing ↗

💳 Επιχειρηματικά / Πελατειακά

- Churn, sales, customer behavior
  - 👉 https://www.kaggle.com/datasets?search=customer+churn ↗
  - 👉 https://www.kaggle.com/datasets?search=sales ↗

✅ Παραδείγματα που ταιριάζουν τέλεια στις απαιτήσεις σου

- Airbnb listings (δεκάδες χιλιάδες γραμμές, mixed variables)
- Credit risk / loan default datasets
- Traffic accidents by city/year
- Retail sales with time + categories
- Health insurance costs + demographics

Αν θες, πες:

- αν προτιμάς κάτι **κοινωνικό, οικονομικό, υγείας ή tech**
- αν θα δουλέψεις σε **Python ή R**

και σου προτείνω **2–3 συγκεκριμένα datasets έτοιμα για κατέβασμα** που ταιριάζουν ακριβώς στο εργαστήριο.

θα με ενδιέφερε κάτι από τον τομέα της υγείας

Ωραία — ο τομέας της **υγείας** είναι ιδανικός για καθάρισμα δεδομένων, EDA και απλά γραμμικά μοντέλα (πολλές κατηγορικές + συνεχείς μεταβλητές).

Παρακάτω είναι **συγκεκριμένα, μη-τετριμμένα tabular datasets** με σωστό μέγεθος:

🏥 Κλινικά & ασθενών (πολύ καλά για ταξινόμηση)

Από Kaggle

- 🫀 Heart disease + lifestyle factors
  https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction ↗
- 🏥 Hospital readmissions
  https://www.kaggle.com/datasets/dansbecker/hospital-readmissions ↗
- 🩺 Diabetes clinical data (50k+ γραμμές)
  https://www.kaggle.com/datasets/mathchi/diabetes-data-set ↗

💊 Κόστη υγείας & δημογραφικά (τέλεια για παλινδρόμηση)

- Insurance charges + age, BMI, smoking κτλ

https://www.kaggle.com/datasets/mirichoi0218/insurance ↗

- Medical expenditure panels
  https://www.kaggle.com/datasets/cdc/medical-expenditure-panel-survey-meps ↗

🧠 Πιο "επιστημονικά" datasets

Από UCI Machine Learning Repository

- Chronic kidney disease
  https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease ↗
- Breast cancer (πολλές μεταβλητές, όχι toy)
  https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) ↗

✅ Γιατί είναι καλά για το εργαστήριο σου

- ✔ Missing values → καθάρισμα
- ✔ Categorical + numerical features
- ✔ Regression ή classification
- ✔ Ρεαλιστικά patterns
- ✔ 500 – 100.000+ γραμμές

📌 Αν θες "σίγουρη επιτυχία" για project:

Top επιλογές:

- Insurance costs (παλινδρόμηση)
- Hospital readmission (ταξινόμηση)
- Diabetes dataset (πλήρες ML pipeline)

Αν θέλεις, μπορώ να σου προτείνω **ένα συγκεκριμένο dataset** και να σου πω:

- τι ερωτήματα να απαντήσεις
- ποιο μοντέλο ταιριάζει
- τι γραφήματα αξίζουν

(πχ *"πρόβλεψη κόστους υγείας με κοινωνικούς παράγοντες"* ή *"ποιοι ασθενείς θα επανεισαχθούν"*).