BDA 2021/22

Datasets for projects

more details at this link:
https://bit.ly/3zuF1Kw

List of datasets

- India House Price Prediction
- Mushroom Classification
- No-Show Appointments
- The Movies Dataset
- Rental Bike
- Chicago Taxi
- Police Killings
- Ethereum Fraud

Module 3: laboratory for interactive project development

- Create teams of “data analysts”
- Choose a dataset among those proposed

1. *October*: 1st Mid Term (Data Understanding and Project Formulation)
2. *November*: 2nd Mid Term (Model implementation and evaluation)
3. *December*: 3rd Mid Term (Model interpretation and explanation)
4. *January*: Exam (Final Project results)

Building Teams...

- Register your team by September 29th (3 or 4 members)

  https://forms.gle/UWLUp11QNVCPBgSZ6

- Write me an email if you cannot find any team

  luca.pappalardo@isti.cnr.it
  giuliano.cornacchia@phd.unipi.it

India House Price Prediction

11 variables
(#rooms, square ft., address, under construction, latitude, longitude, etc.)

1 target
(house price)

- **room for feature engineering**
- **≈ 30,000 records**

kaggle

https://www.kaggle.com/anmolkumar/house-price-prediction-challenge

Mushroom Classification

22 variables
(cap-shape, cap-colour, ring number, habitat, etc.)

☠

1 target:
safe to eat (e) or **deadly** poison (p)

kaggle

- **balanced dataset (52% - 48%)**
- **≈ 8,000 records**

https://www.kaggle.com/uciml/mushroom-classification

No-Show Appointments

13 variables
(gender, age, SMS received, booking date, appointment date, place, etc.)

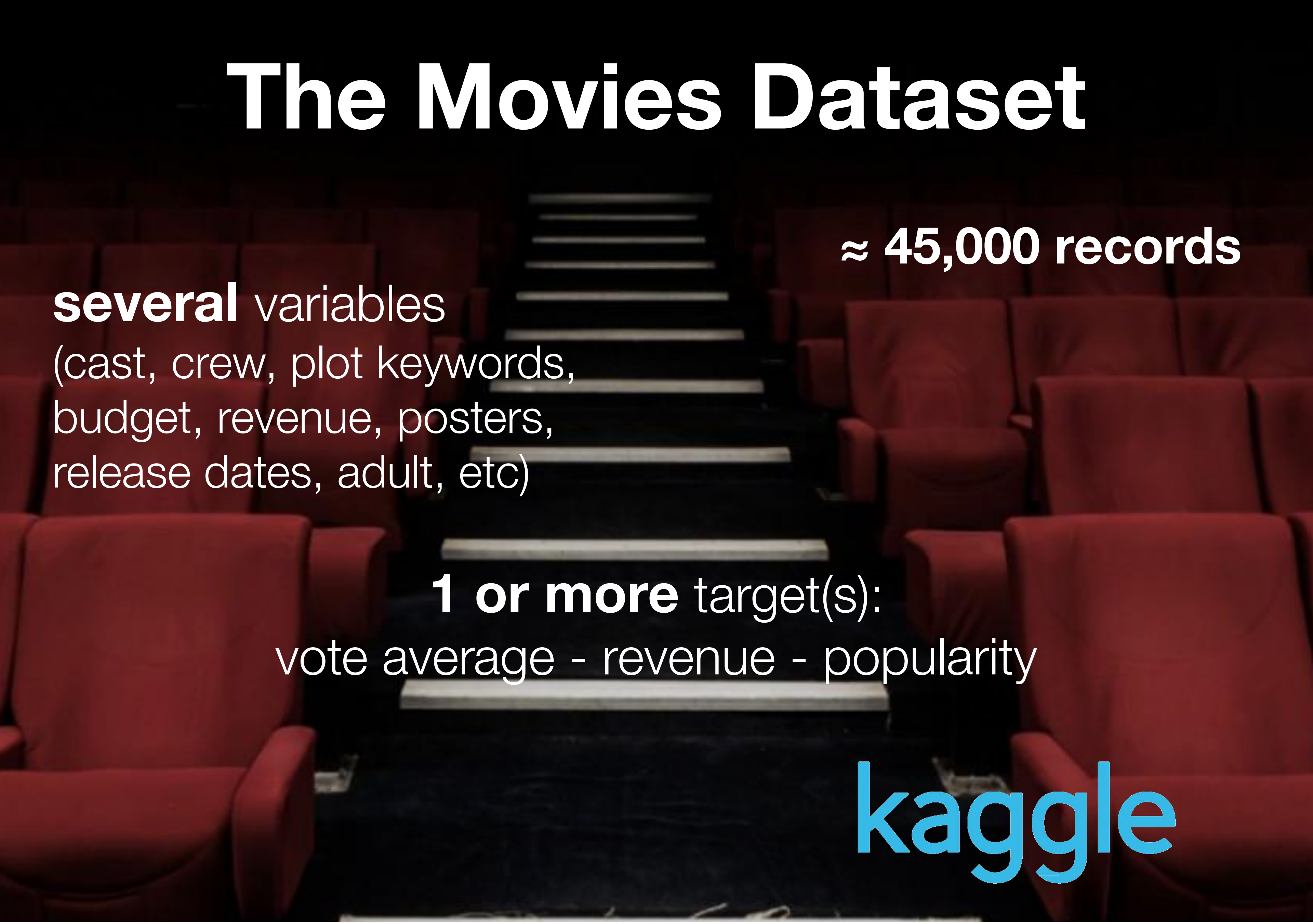
1 target:
show or **no-show**

≈ 110,000 records

kaggle

https://www.kaggle.com/joniarroba/noshowappointments

The Movies Dataset



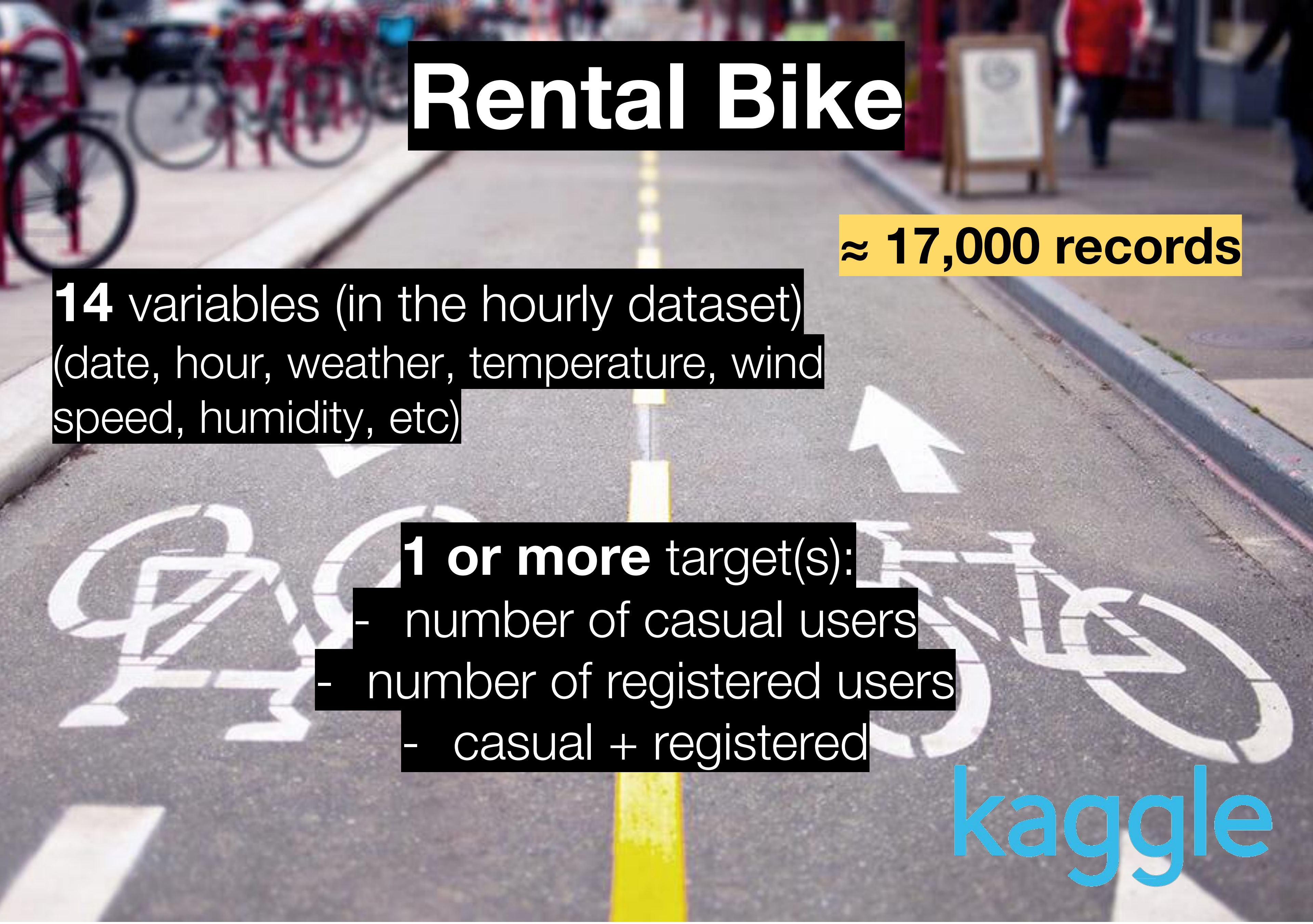
≈ **45,000 records**

several variables
(cast, crew, plot keywords,
budget, revenue, posters,
release dates, adult, etc)

1 or more target(s):
vote average - revenue - popularity

kaggle

https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv



Rental Bike

≈ 17,000 records

14 variables (in the hourly dataset) (date, hour, weather, temperature, wind speed, humidity, etc)

1 or more target(s):

- number of casual users
- number of registered users
- casual + registered

kaggle

https://www.kaggle.com/shelvigarg/rental-bike-demands-on-hourly-and-daily-basis?select=hour.csv

Chicago Taxi

several variables
(taxi ID, start/end time, start/end positions, miles, etc)

1 or more target(s):

- predict the number of trips in a given area at a specific hour of the day
- predict the trip’s fare

https://www.kaggle.com/chicago/chicago-taxi-trips-bq

https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data

Police Killings

kaggle

29 variables
(age, gender, race/ethnicity, city, cause, imputation, etc)

1 or more target(s):

- Predict the number of killings in a given region
- Verify whether a race/ethnicity bias is supported from an objective analysis.

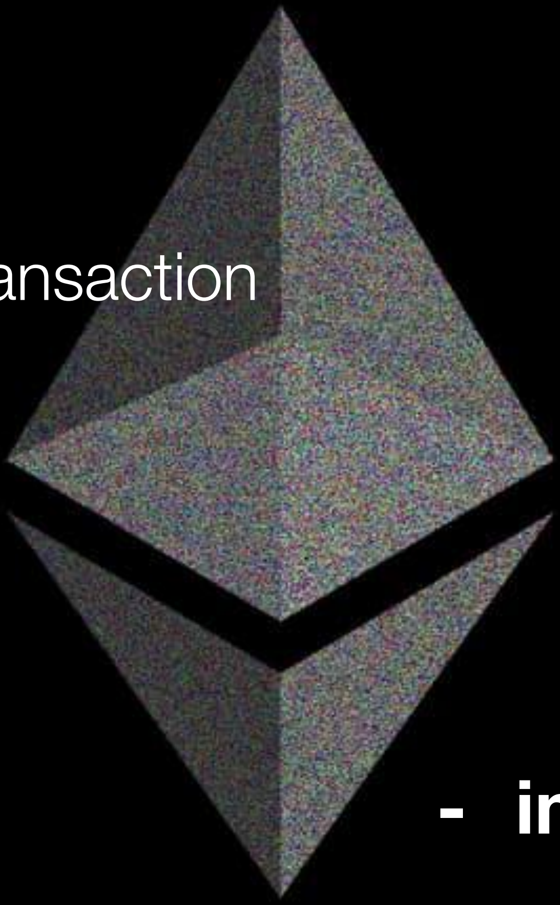
≈ 28,000 records

https://www.kaggle.com/djonafegnem/police-fatalities-in-the-us-from-2000-to-2020

Ethereum Fraud

kaggle

50 variables
which describe the transaction



1 target:
fraud or **no-fraud**

- **imbalanced dataset**

≈ 10,000 records

https://www.kaggle.com/vagifa/ethereum-frauddetection-dataset