Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

DATA SCIENCE

NDBI048

Datasets Overview

https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

PRACTICALS

- Aim: hands-on experimenting with methods discussed in the lectures
  - Real-world data set
    - Chosen from a given list / own sources approved by the instructors
- Result: analytic report
  - description of the solved problem
  - input data description
  - planned process steps, methods, techniques (a short description)
  - data quality check and resolution about data errors or missing data
  - data exploration
  - (possibly simple) predictive or classification model (if not stated differently) for a proper target
  - model performance evaluation
  - summary, conclusion

DATASETS LIST I

Each dataset comes from some real-world branch.

- **King James Bible** – the Bible in an ancient English translation (17th century)
- **Births in US –** daily number of newborns from 1968 to 1988 in USA states
- **Brazil accidents** – data on traffic accidents in Brazil from 2007 to 2023
- **Chess ratings** – official monthly lists of chess players (Dec 2017 to Dec 2022)
- **Card transactions** – transactions at petrol stations of a small Czech bank clients
- **Earthquakes** – recorded events from 1990 to 2023
- **Elections** – results of Czech parliamentary elections in 2013 and 2017 by district

DATASETS LIST II

- **Jokes** – evaluation of English jokes
- **Population in Czech municipalities** – inhabitants in time series 1971–2020
- **Population in world countries** – sampled demographic data 2000–2021
- **Tennis matches** – statistics of top tennis matches in 2015–2023
- **Vaccination** – status and death date for citizens in the Czech Republic 2020–2022
- **Football World Cup** – international football matches in Qatar and long before

MORE ON DATASETS – SEE…

- assignment instructions (Git repository)
- README files inside zips
- if still unclear, just ask us