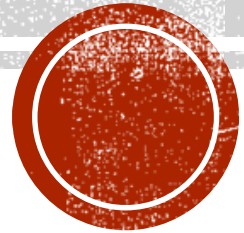
Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

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: two reports
  - Report I. What we found out about the data
    - At the end of part I of lectures (middle of the semester)
  - Report II. What we found in the data
    - At the end of part II of lectures (end of semester)

DATASETS LIST

- **NHL Players** – individual statistics of the NHL players in seasons 2004 to 2018
- **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
- **Jokes** – evaluation of English jokes
- **Tennis matches** – statistics of top tennis matches in 2015–2023
- **King James Bible** – the Bible in an ancient English translation (17th century)
- **Brazil accidents** – data on traffic accidents in Brazil from 2007 to 2023
- **Elections** – results of Czech parliamentary elections in 2013 and 2017 by district
- **Population in world countries** – sampled demographic data 2000–2021
- **Population in Czech municipalities** – inhabitants in time series 1971–2020
- **Song lyrics** – list of popular songs and music compositions with lyrics
- Each dataset comes with one or more real-world problems.

DATASETS – NHL PLAYERS

- What does the players ranking depend on?
- How many times will the player S score next season? Will it be more than the last season?
- Will the player T be traded next season?
- Will the player Z have average ice-time over 16 minutes next season?

DATASETS – CHESS RATINGS

- Will the player N have higher rating next month or next year?
- How many games will player M play next year?
- How did COVID-19 affect playing chess?
- Is rating deflating or inflating – generally or by countries?
- What rating in rapid or blitz games will have player P if we know his/her standard rating?
- What is a typical rating of grandmaster (or other title) and what is the rating distribution?

DATASETS – CARD TRANSACTIONS

- How many transactions will be at station S next day? How many next week in night hours only?
- What will be the total amount spent at station T next day?
- What is a typical distribution of transaction counts during a week?
- What are favourite station brands? Does it change in time?
- What will be the total amount spent by client U next week/month?

DATASETS – JOKES

- What features of jokes have a relationship to the evaluation?
- Are there readers with similar taste (reader A likes/dislikes similar jokes as reader B, joke X is liked by similar readers as joke Y)?
- What evaluation will have a new joke J? What evaluation variance will it have?
- Will reader R like the joke K when reader Q gave it P points?

DATASETS – TENNIS MATCHES

- Is surface type important for winning probability of player A?
- What is the probability of winning a match with respect to players‘ ranking?
- How long will a match of players M a N be?
- Will player S record more than 10 aces in a match on a hard surface?
- Are there changes in average match metrics between, say, 2016 and 2022?

DATASETS – SONG LYRICS

- What words are typical for pop songs compared to other genres?
- How many unique words (and possibly not „stopwords“) are typical for a country song?
- Which interpret sings a song with given lyrics?
- To what genre does belong a song with givem lyrics?

DATASETS – BRAZIL ACCIDENTS

- How many accidents will happen tomorrow?
- How many injured and dead people will be in accidents next week in place/region R?
- Was the driver drunk, given all other information about the accident?
- Can we detect new dangerous places for accidents?

DATASETS – ELECTIONS

- What vote share will get party X in district/municipality Y in 2017, given vote shares by regions?
- What vote share will get party X in district Y in 2017, given vote shares of adjacent districts?
- Given the region/county and size of municipality, what share of unemployed people can we expect?
- Will the turnout at elections in district D be over/under overall average?
- Will the change of vote share for party P from 2013 to 2017 be over/under overall average change?

DATASETS – COUNTRIES POPULATION

- How many people will live in the country A next year? Will it be more or less than this year?
- Will median age in the country B increase next year?
- Given the fertility of adjacent countries, what fertility is in the country C?
- Which countries within a group or in the world are similar to the country D?

DATASETS – MUNICIPALITIES POPULATION

- How many people will live in the town A next year? Will it be more or less than this year?
- Will the migration to the municipality M be higher than from it?
- How many babies will be born next year? Will it be more or less than this year?
- Given the population change in adjacent municipalities with similar size, will the population in municipality S be higher?

DATASETS – KING JAMES BIBLE

- Does the verse (by its text) belong to the Old Testament, or to the New Testament?
- What pairs of books are mostly similar?

DATASETS – THREE MORE...

- for special use during the whole course
- **Titanic** – well known dataset of passengers and their survival
- **Home Credit** – loans (un)paid and people features
- **Stopwords** – for filtering out useless English words