# Semestrální práce

## Pokročilé dolování z databází

### Úvodní zpráva 1

Daniel Balarin

Lukáš Vala

# Datasety

Dataset 1 – Adult

* Převzato z: [UCI Machine Learning Repository](https://archive-beta.ics.uci.edu/ml/datasets/adult)
* Počet řádků: 48842
* Počet sloupců: 14

Tento dataset obsahuje data z roku 1994. Obsahuje informace o lidech z celého světa a k tomu navíc, zda daná osoba vydělává více či méně než 50000 dolarů za rok. Můžeme si tedy pokládat otázky, jak ovlivňuje plat národnost, rasa, pohlaví a podobně.

Atributy:

* age: Věk
  + Hodnoty: Celá čísla
* workclass: Pracovní třída
  + Hodnoty: "Private", "Self-emp-not-inc", "Self-emp-inc", "Federal-gov", "Local-gov", "State-gov", "Without-pay", "Never-worked"
* fnlwgt: Počet lidí reprezentovaný záznamem
  + Hodnoty: Celá čísla
* education: Nejvyšší úroveň dosaženého vzdělání.
  + Hodnoty: "Bachelors", "Some-college", "11th", "HS-grad", "Prof-school", "Assoc-acdm", "Assoc-voc", "9th", "7th-8th", "12th", "Masters", "1st-4th", "10th", "Doctorate", "5th-6th", "Preschool"
* education-num: Nejvyšší úroveň dosaženého vzdělání vyjádřená číslem.
  + Hodnoty: Celá čísla
* marital-status: Rodinný stav
  + Hodnoty: "Married-civ-spouse", "Divorced", "Never-married", "Separated", "Widowed", "Married-spouse-absent", "Married-AF-spouse"
* occupation: Oblast zaměstnání
  + Hodnoty: "Tech-support", "Craft-repair", "Other-service", "Sales", "Exec-managerial", "Prof-specialty", "Handlers-cleaners", "Machine-op-inspct", "Adm-clerical", "Farming-fishing", "Transport-moving", "Priv-house-serv", "Protective-serv", "Armed-Forces"
* relationship: Vztah
  + Hodnoty: "Wife", "Own-child", "Husband", "Not-in-family", "Other-relative", "Unmarried"
* race: Rasa
  + Hodnoty: "White", "Asian-Pac-Islander", "Amer-Indian-Eskimo", "Other", "Black"
* sex: Pohlaví
  + Hodnoty: "Female", "Male"
* capital-gain
  + Hodnoty: Celá čísla
* capital-loss
  + Hodnoty: Celá čísla
* hours-per-week: Počet odpracovaných hodin za týden
  + Hodnoty: Celá čísla
* native-country: Země původu
  + Hodnoty: "United-States", "Cambodia", "England", "Puerto-Rico", "Canada", "Germany", "Outlying-US(Guam-USVI-etc)", "India", "Japan", "Greece", "South", "China", "Cuba", "Iran", "Honduras", "Philippines", "Italy", "Poland", "Jamaica", "Vietnam", "Mexico", "Portugal", "Ireland", "France", "Dominican-Republic", "Laos", "Ecuador", "Taiwan", "Haiti", "Columbia", "Hungary", "Guatemala", "Nicaragua", "Scotland", "Thailand", "Yugoslavia", "El-Salvador", "Trinadad&Tobago", "Peru", "Hong", "Holand-Netherlands"
* Income label: Roční příjem v amerických dolarech
  + Hodnoty: „>50k”, “<=50k”

Dataset 2 – GDP per capita – Hrubý domácí produkt na osobu

* Převzato z: <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=1994&most_recent_year_desc=true&start=1993>
* Počet řádků: 266
* Počet sloupců: 2

Tento dataset obsahuje data z roku 1994 (stejně jako první dataset). Obsahuje informace o HDP podle země na osobu. Sloupce datasetu je stát a hrubý domácí produkt na osobu

Státy: Aruba, Africa Eastern and Southern, Afghanistan, Africa Western and Central, Angola, Albania, Andorra, Arab World, United Arab Emirates, Argentina, Armenia, American Samoa, Antigua and Barbuda, Australia, Austria, Azerbaijan, Burundi, Belgium, Benin, Burkina Faso, Bangladesh, Bulgaria, Bahrain, Bahamas, The, Bosnia and Herzegovina, Belarus, Belize, Bermuda, Bolivia, Brazil, Barbados, Brunei Darussalam, Bhutan, Botswana, Central African Republic, Canada, Central Europe and the Baltics, Switzerland, Channel Islands, Chile, China, Cote d'Ivoire, Cameroon, Congo, Dem. Rep., Congo, Rep., Colombia, Comoros, Cabo Verde, Costa Rica, Caribbean small states, Cuba, Curacao, Cayman Islands, Cyprus, Czechia, Germany, Djibouti, Dominica, Denmark, Dominican Republic, Algeria, East Asia & Pacific (excluding high income), Early-demographic dividend, East Asia & Pacific, Europe & Central Asia (excluding high income), Europe & Central Asia, Ecuador, Egypt, Arab Rep., Euro area, Eritrea, Spain, Estonia, Ethiopia, European Union, Fragile and conflict affected situations, Finland, Fiji, France, Faroe Islands, Micronesia, Fed. Sts., Gabon, United Kingdom, Georgia, Ghana, Gibraltar, Guinea, Gambia, The, Guinea-Bissau, Equatorial Guinea, Greece, Grenada, Greenland, Guatemala, Guam, Guyana, High income, Hong Kong SAR, China, Honduras, Heavily indebted poor countries (HIPC), Croatia, Haiti, Hungary, IBRD only, IDA & IBRD total, IDA total, IDA blend, Indonesia, IDA only, Isle of Man, India, Not classified, Ireland, Iran, Islamic Rep., Iraq, Iceland, Israel, Italy, Jamaica, Jordan, Japan, Kazakhstan, Kenya, Kyrgyz Republic, Cambodia, Kiribati, St. Kitts and Nevis, Korea, Rep., Kuwait, Latin America & Caribbean (excluding high income), Lao PDR, Lebanon, Liberia, Libya, St. Lucia, Latin America & Caribbean, Least developed countries: UN classification, Low income, Liechtenstein, Sri Lanka, Lower middle income, Low & middle income, Lesotho, Late-demographic dividend, Lithuania, Luxembourg, Latvia, Macao SAR, China, St. Martin (French part), Morocco, Monaco, Moldova, Madagascar, Maldives, Middle East & North Africa, Mexico, Marshall Islands, Middle income, North Macedonia, Mali, Malta, Myanmar, Middle East & North Africa (excluding high income), Montenegro, Mongolia, Northern Mariana Islands, Mozambique, Mauritania, Mauritius, Malawi, Malaysia, North America, Namibia, New Caledonia, Niger, Nigeria, Nicaragua, Netherlands, Norway, Nepal, Nauru, New Zealand, OECD members, Oman, Other small states, Pakistan, Panama, Peru, Philippines, Palau, Papua New Guinea, Poland, Pre-demographic dividend, Puerto Rico, Korea, Dem. People's Rep., Portugal, Paraguay, West Bank and Gaza, Pacific island small states, Post-demographic dividend, French Polynesia, Qatar, Romania, Russian Federation, Rwanda, South Asia, Saudi Arabia, Sudan, Senegal, Singapore, Solomon Islands, Sierra Leone, El Salvador, San Marino, Somalia, Serbia, Sub-Saharan Africa (excluding high income), South Sudan, Sub-Saharan Africa, Small states, Sao Tome and Principe, Suriname, Slovak Republic, Slovenia, Sweden, Eswatini, Sint Maarten (Dutch part), Seychelles, Syrian Arab Republic, Turks and Caicos Islands, Chad, East Asia & Pacific (IDA & IBRD countries), Europe & Central Asia (IDA & IBRD countries), Togo, Thailand, Tajikistan, Turkmenistan, Latin America & the Caribbean (IDA & IBRD countries), Timor-Leste, Middle East & North Africa (IDA & IBRD countries), Tonga, South Asia (IDA & IBRD), Sub-Saharan Africa (IDA & IBRD countries), Trinidad and Tobago, Tunisia, Turkiye, Tuvalu, Tanzania, Uganda, Ukraine, Upper middle income, Uruguay, United States, Uzbekistan, St. Vincent and the Grenadines, Venezuela, RB, British Virgin Islands, Virgin Islands (U.S.), Vietnam, Vanuatu, World, Samoa, Kosovo, Yemen, Rep., South Africa, Zambia, Zimbabwe

## Analytické otázky

Při práci s již zmíněnými daty se budeme zabývat například těmito otázky:

* Kterými atributy je nejvíce ovlivněn roční příjem osob?
* Je roční příjem ovlivněn rasou daného člověka (a jak silně?)
* Je roční příjem ovlivněn pohlavím daného člověka (a jak silně?)
* Jak silně ovlivňuje roční příjem jednotlivých osob HDP na osobu státu, ve kterém tyto osoby žijí?