WHERE IS MY DATA FROM? (FONT)

**Acknowledgements** This dataset named “adult” is found in the UCI machine learning repository  
<http://www.cs.toronto.edu/~delve/data/adult/desc.html>

The detailed description on the dataset can be found in the original UCI documentation  
<http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html>

<https://www.kaggle.com/datasets/wenruliu/adult-income-dataset>

<https://www.kaggle.com/datasets/uciml/adult-census-income/data>

PREPROCESSING

BASIC DESCRIPTIVE POST PROCESSING IN THOSE VARIABLES WITH CHANGES

METADATA (DATA DE DATA)

* VARIABLE NAME
* SHORT VARIABLE
* MEANING
* TYPE
* MIN
* MAX
* UNIT
* MODALITY
* SHORT MODALITY

**DATA SELECTION**

Hi ha dues columnes amb la mateixa informació (education i educational.num). Per no tenir-la repetida, mantenim només la numèrica. La resta ho deixem igual.

**Traducció nombres - educació (educational.num vs education)**

1 Preschool

2 1st-4th

3 5th-6th

4 7th-8th

5 9th

6 10th

7 11th

8 12th

9 HS-grad

10 Some-college

11 Assoc-voc

12 Assoc-acdm

13 Bachelors

14 Masters

15 Prof-school

16 Doctorate

**Glossari:**

**Assoc-voc:** Associate’s degree (vocational) a two-year college degree that focuses on practical job skills.

**Assoc-acdm:** Associate’s degree (academical) a two-year college degree, but it's more focused on general academic studies. It's often a stepping stone towards a four-year bachelor's degree.

**Prof-school:** Advanced education for specific professions, like being a doctor, lawyer, or dentist

——————————————————————————————————————

**OUTLIERS**

Utilitzeu els boxplot per explicar els outliers.

Les següents categories tenen outliers que decidim no tractar degut a que son valors extrems de la població, preservant la distribució natural de les dades i sense perdre informació que pot ser rellevant.

**Age**

* Typically, ages range between 18 and 65 in working datasets.
* Very **young (<18)** or very **old (>70-80)** people may be considered outliers.
* If there are **ages above 90**, they may not represent the general population.**Capital Gain & Capital Loss**
* These values usually have a **strong skew** because most people **report zero** capital gains/losses.
* **Very high values** (e.g., extremely rich individuals) might be statistical outliers.

**Capital Gain & Capital Loss**

* These values usually have a **strong skew** because most people **report zero** capital gains/losses.
* **Very high values** (e.g., extremely rich individuals) might be statistical outliers.

**Hours per Week**

* Most people work between **35-50 hours per week**.
* Values below **10 hours** (part-time) or above **80-100 hours** (overwork) might be outliers.

——————————————————————————————————————

**MISSING VALUES**

Per tractar els valors que falten hem decidit emplenar-los amb el valor més freqüent (són dades que falten de manera random, la seva falta no aporta informació i no es causa d’un valor no aplicable (com embarassos d’un home)).

columnes amb missing values (En comptes de NA son “?”): workclass, occupation, native\_country

——————————————————————————————————————

**VARIABLE RENAMING**

Afegeixo els canvis, compareu amb els originals, feel free de canviar-li els noms, aviseu-me i torno a fer els gràfics. He ajuntat algunes categories perquè tenen molts pocs valors (mireu les taules de freqüència (table)).

**Columnes:**

colnames(dd) <- c("age", "workclass", "edu\_num", "marital", "occupation",

"relationship", "race", "sex", "cap\_gain", "cap\_loss",

"hours\_week", "native\_country", "income")

**Modalities:**

Es codi però s'entén, dd$nom\_variable i dintre dels parentesis els canvis de nom. **Hi ha categories juntes per el seu nombre ínfim de registres.**

dd$workclass <- recode(dd$workclass,

"Private" = "Priv",

"Self-emp-not-inc" = "SelfN",

"Self-emp-inc" = "SelfI",

"Federal-gov" = "Fed",

"Local-gov" = "Loc",

"State-gov" = "State",

"Without-pay" = "NoPay",

"Never-worked" = "NoPay")

El selfI vol dir incorporated (com una empresa) i selfn(no inc) vol dir com guanyar diners sense una empresa tipo vendre coses al jardí, llimonada, freelance… (tu gestiones els teus impostos i la teva feina).

dd$marital <- recode(dd$marital,

"Never-married" = "NevMarr",

"Married-civ-spouse" = "Married",

"Married-AF-spouse" = "Married",

"Married-spouse-absent" = "Sep",

"Separated" = "Sep",

"Divorced" = "Div",

"Widowed" = "Widow")

dd$occupation <- recode(dd$occupation,

"Exec-managerial" = "ExecMan",

"Prof-specialty" = "Prof",

"Adm-clerical" = "AdminCler",

"Sales" = "Sales",

"Craft-repair" = "CraftRep",

"Transport-moving" = "Trans",

"Handlers-cleaners" = "HandlCl",

"Machine-op-inspct" = "MachOp",

"Tech-support" = "Tech",

"Protective-serv" = "ProtServ",

"Armed-Forces" = "Army",

"Farming-fishing" = "FarmFish",

"Priv-house-serv" = "House",

"Other-service" = "Other")

dd$native\_country <- recode(dd$native\_country,

"United-States" = "USA",

.default = "Other") # Group all other countries as "Other"

***Dels altres països tenim menys de 1000 registres.***

| **VARIABLE NAME** | **SHORT VARIABLE** | **MEANING** | **TYPE** | **MIN** | **MAX** | **UNIT** | **MODALITY** | **SHORT MODALITY** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AGE | age | Number of years a person had at that moment | Integer | 17 | 90 | Years | - | - |
| WORKCLASS | workclass | Sector or type of work.In some rows we will see Private as a workclass, that means that this person is working in the private sector. | Categorical | - | - | - | - Private  Self-emp-not-inc  -Self-emp-inc  -Federal-gov  -Local-gov  -State-gov  -Without-pay -Never-worked | -Priv  -SelfN  -SelfI  -Fed  -Loc  -State  -NoPay  -NoPay |
| FNLWGT | fnlwgt | Final weight assigned to the person, representing the estimated number of people with similar characteristics. | Integer | 12285 | 1490400 | - | - | - |
| EDUCATION | education | The highest level of education achieved | Categorical | - | - | - | **\*\*** | - |
| EDUCATIONAL-NUM | edu\_num | Numerical representation of education level. | Integer | 1 | 16 | - | **\*\*** | - |
| MARITAL-STATUS | marital | Marital status of the individual. | Categorical | - | - | - | -Never-married  -Married-civ-spouse  -Married-AF-spouse  -Separated  -Divorced  -Widowed | -NevMarr  -Married  -Married  -Sep  -Div  -Widow |
| OCCUPATION | occupation | Type of occupation or profession. | Categorical | - | - | - | -Exec-managerial  -Prof-specialty  -Adm-clerical -Craft-repair  -Transport-moving -Handlers-cleaners  -Machine-op-inspct  -Tech-support -Protective-serv  -Armed-Forces  -Farming-fishing  -Priv-house-serv  -Other-service | -ExecMan  -Prof  -AdminCler  -CraftRep,  -Trans  -HandlCl  -MachOp  -Tech  -ProtServ  -Army  -FarmFish  -House  -Other |
| RELATIONSHIP | relationship | Relationship status within a household. | Categorical | - | - | - | -Husband  -Wife  -Own-child  -Not-in-family  -Other-relative  -Unmarried | - |
| RACE | race | Race of the individual. | Categorical | - | - | - | White, Black, Asian-Pac-Islander | - |
| GENDER (SEX) | sex | Sex of the individual. | Categorical | - | - | - | Male, Female | - |
| CAP-GAIN | cap\_gain | Income from capital gains (e.g., asset sales). | Integer | 0 | 99999 | USD | - | - |
| CAP-LOSS | cap\_loss | Losses incurred from asset sales. | Integer | 0 | 4356 | USD | - | - |
| HOURS-WEEK | hours\_week | Number of hours worked per week. | Integer | 1 | 99 | Hours | - | - |
| NATIVE-COUNTRY | native\_country | Country of origin. | Categorical | - | - | - | -United States  -Other | -USA  -Other |
| INCOME  (Target Variable) | income | Income category (>50K or <=50K per year). | Categorical | - | - | USD | >50K,  <=50K | - |