**Metadata for cleaned datasets**

1. cleaned\_theUS\_adult dataset:

* Structure:

27504 rows, 8 columns

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Unit | Type | Meaning |
| Unnamed: 0 | \ | Int64 | Original indices(ignored) |
| age | Years old | int64 | Ages of people in the U.S. |
| education.num | \ | int64 | Educational attainment by number: 1(low)-16(high) |
| occupation | \ | object | \ |
| race | \ | object | \ |
| sex | \ | object | Male / Female |
| hours.per.week | hour | int64 | Working hours per week in the U.S. |
| native country | \ | object | \ |
| income | dollar | object | Income of people in the U.S.: <=50K or >50K |

1. The education and earning dataset (cleaned\_earning.csv):

* Structure:

123611 rows, 7 columns

|  |  |  |  |
| --- | --- | --- | --- |
| Field | Unit | Type | Meaning |
| Unnamed: 0 | \ | Int64 | Original indices(ignored) |
| Country | \ | object | States |
| ISC1\_LEVEL\_CAT | \ | object | Level of education background |
| Gender | \ | object | Gender of data acquisition object |
| income | percentage | int64 | Income of data acquisition object |
| Education level and programe orientation | \ | object | Education level and program orientation |
| Grad\_Rate | percentage | int64 | Graduation rate of education background |

1. cleaned\_HR\_Employee\_Attrition dataset:

* Structure:

1470 rows, 28 columns

A picture containing graphical user interface, text

Description automatically generated

1. The salary dataset (project1\_salary(clean).csv):

* Structure:

26 columns, 224 rows

|  |  |  |  |
| --- | --- | --- | --- |
| Field | unit | type | meaning |
| Name | \ | object |  |
| Type | \ | object | eg. Engineering, State |
| str\_50\_type | dollar | float64 | the median salary at the beginning of the career |
| mid\_50\_type | float64 | Different percentile of the median salary at the middle career  (low-high) by school type |
| mid\_10\_type | float64 |
| mid\_25\_type | float64 |
| mid\_75\_type | float64 |
| mid\_90\_type | float64 |
| Region | \ | object |  |
| str\_50\_region | dollar | float64 | First 50% value of the median salary at the beginning of the career  (low-high) by region |
| mid\_50\_region | float64 | Different percentile of the median salary at the middle career  (low-high) by region |
| mid\_10\_region | float64 |
| mid\_25\_region | float64 |
| mid\_75\_region | float64 |
| mid\_90\_region | float64 |
| Major | \ | object |  |
| str\_50\_major | dollar | float64 | First 50% value of the median salary at the beginning of the career  (low-high) by major |
| mid\_50\_major | float64 | Different percentile of the median salary at the middle career  (low-high) by major |
| mid\_10\_major | float64 |
| mid\_25 | float64 |
| mid\_75\_major | float64 |
| mid\_90\_major | float64 |
| percentage\_str\_to\_mid\_major | % | float64 | growth rate from the beginning career to mid career by major |

## The rights associated with the data

The data owner is University of Sydney DATA1002 labR16A group4 (Kanglu Pei, Xu Jia, Yuxuan Xiao, Songyin Li).

The license of these four datasets is CC0, public domain, with the visibility of the public, where states explicitly that there is no copyright and thereby anyone can copy, modify, distribute and perform the work without asking permission.

These datasets were cleaned and summarised by group4 members from the original datasets as follows:

|  |  |  |
| --- | --- | --- |
| filename | Source /data owner | License & applicable policy |
| adult.csv | University College London | CC0, public domain, with the visibility of the public, where states explicitly that there is no copyright and thereby anyone can copy, modify, distribute and perform the work without asking permission |
| EAG\_EARNINGS.csv | OECD.Stat |
| EAG\_GRAD\_ENTR\_RATES.csv |
| degrees-that-pay-back.csv | The Wall Street Journal |
| salaries-by-college-type.csv |
| salaries-by-region.csv |
| HR-Employee-Attrition | Pavansubhash | OdbL, open dataset and no limitation to be applied |