# Information Visualization

# CHECKPOINT II: Data cleaning and processing

G22

**1. Initial Dataset**

For heath indicators the following datasets will be used: Life expectancy at birth, life expectancy at 65, suicide rates, daily smokers, alcohol consumption, overweight or obese population and deaths from cancer.

And for potential health influencers: Average wages, employment rate, hours worked, adult education level, social spending, gross domestic product (GDP), life satisfactionand air exposure.

All the data used in this project will be static and is available at: https://github.com/jtaca/What-makes-us-healthy-/tree/master/VI%20datasets. The time span of the visualization is from 2007 to 2017.

* + - 1. (from “Life\_expectancy\_at\_birth.csv”) and “Life\_expectancy\_at\_65.csv”

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; LIFEEXP; TOT; YR; A; 1972; 72;

(from “Suicide\_rates.csv”)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; SUICIDE; TOT; 100000PER; A; 1983; 12.1;

(from “Daily\_smokers\_population.csv”) (PC\_POP15 is % of population aged 15+)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; SUICIDE; TOT; 100000PER; A; 1968; 15.6; B

(from “Cancer\_deaths.csv”)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; DEATHCANCER; TOT; 100000PER; A; 1968; 230.6; B

(from “Alcohol\_consumption.csv”) (LT\_CAP15 is Liters/capita aged 15+)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; ALCOHOL; TOT; LT\_CAP15; A; 1966; 10.3;

(from “Overweight\_or\_obese\_population.csv”)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUT; OVEROBESE; SELFREPORTED; PC\_POP15; A; 2014; 46.7;

(from “Average\_wage.csv”) (USD is dollars)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUT; AVWAGE; TOT; USD; A; 2008; 49509.35566;

(from “Employment\_rate.csv”) (THND\_PER is thousands of persons and PC\_WKGROUP is % of working age population)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUT; EMP; MEN; THND\_PER; A; 1974; 1847; B

AUS; EMP; WOMEN; PC\_WKGPOP; A; 1994; 56.85906;

(from “Hours\_worked.csv”) (HR\_WKD is Hours/worker)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; HRWKD; TOT; HR\_WKD; A; 1991; 1773.5;

(from “Adult\_education\_level.csv”) (BUPPSRY is Below upper secondary, TRY is Tertiary and UPPSRY is Upper secondary)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; EDUADULT; BUPPSRY; PC\_25\_64; A; 2017; 19.012659

(from “Air\_pollution\_exposure.csv”) (THND\_TONNE is thousands of tonnes)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

FIN; POLLUTIONEXP; EXPOS2PM25; MICGRCUBM; A; 1990; 7.41045

(from “Social\_spending.csv”) (PC\_GDP is % of GDP)

**Location; Subject; Measure; Time; Value**

AUT; SOCEXP; PRIV; PC\_GDP; A; 1984; 0.624

(from “Gross domestic product (GDP)” (USD\_CAP is US dollars/capita)

**LOCATION; INDICATOR; SUBJECT; MEASURE; FREQUENCY; TIME; Value; Flag Codes**

AUS; GDP; TOT; USD\_CAP; A; 1976; 7377.136633;

(from “happiness-cantril-ladder.csv (Self-Reported happiness)

**Entity,Code,Year,World Happiness Report 2016 (Cantril Ladder (0=worst; 10=best))**

Afghanistan,AFG,2008,3.723589897)

**2. Selected/Derived Data**

**Selected Attributes:** Location, Time, Education, Pollution, Happiness, GDP, Social Spending, Employment rate, Hours worked, Average Wage, Alcohol, Smokers, Cancer, Suicide, Obese, Life expectancy at birth, Life expectancy at age 65.

**Derived Attributes:** Average of life expectancy at 65 – this dataset had the average life expectancy for men and women, so in order to be unisex we did the average of these 2 values.

**3. Data abstraction**

1. The dataset is composed by 8 CSV files
2. **Dataset Type:** Multidimensional table
3. **Availability:** Static
4. **Attributes and semantics (per file):**

**All Files:  
 - TIME (continuous, sequential, time-based hierarchic):** the year relative to data. **- LOCATION (nominal, non-sequential, not hierarchic):** Country relative to each data.  **Adult\_education\_level.csv:   
 - Education (ratio, non-sequential, not hierarchic):** number of people with secondary education. **Air\_Pollution.csv:   
 - Pollution (ratio, non-sequential, not hierarchic):** Population exposure to more than 10 micrograms/m3 and are expressed as annual average. **Happiness.csv:   
 - Happiness (ordinal, non-sequential, not hierarchic):** Self-reported Happiness from 0 to 10. **Social\_spending\_and\_GDP.csv:   
 - GDP (ordinal, non-sequential, not hierarchic):** GDP relative to each country. **- Social spending (ordinal, not hierarchic):** Money spent on socialist policies per capita. **Employment\_rate.csv:  
 - Employment rate (ratio, non-sequential, not hierarchic):** Employment rate of active population. **Hours\_worked.csv:   
 - Hours worked (ordinal, not hierarchic):** Average hours worked per year.  
**Average\_wage.csv:   
 - Average Wages (ordinal, not hierarchic):** Average earning per year. **Healthy.csv:  
 - Cancer (ordinal, non-sequential , not hierarchic):** Death by cancer per 100000 people  
  **- Suicide (ordinal, non-sequential, not hierarchic):** Suicide by cancer per 100000 people. **- Alcohol (ratio, non-sequential, not hierarchic):** annual sales of pure alcohol in litres per person aged 15 years and older. **- Smokers (ratio, non-sequential, not hierarchic):** Percentage of population aged 15 years and over who are reporting to smoke every day.  **- Obese (ratio, non-sequential, not hierarchic):** Percentage of inhabitants with excessive weight presenting health risks because of the high proportion of body fat. **- LifeAtBirth (ordinal, non-sequential, not hierarchic):** Life Expectancy at birth **-** Average life expectancy at birth   
 **- LifeAtOld (ordinal, non-sequential , not hierarchic):**Life Expectancy at age 65 Average life expectancy at age 64

1. **All the data is time based and linked to country.**
   * + 1. **4. Dataset processing**
       2. All datasets were filtered to limit the “Time” from 2007 to 2017 and sorted by location and time.
       3. Alcohol consumption, daily smokers, deaths from cancer, suicide rates and life expectancy at birth are filtered by “SUBJECT” with value “TOT” in order to select both men and women.
       4. Obese or overweight population is filtered by “SUBJECT” with value “MEASURED” because we discard the “SELFREPORTED” measure.
       5. Then, the we select the rows we want (LOCATION, TIME and (name of potential health factor)). Then we merge the datasets.
       6. Life expectancy at 65 is divided in 2 datasets: filtered “SUBJECT” by “MEN” and by “WOMEN” (because in this dataset we don’t have the “TOT” value). Then we do the average of the values from “MEN” and “WOMEN”, select the rows again and we merge with other datasets.
       7. Adult Education is also filtered by “SUBJECT”, with value “UPPSRY” which means adults which education is higher than high school.
       8. Air Pollution is also filtered by “MEASURE” with value “PC\_POP” because we want the percentage of population as measure.
       9. GDP and Social Spending are also filtered by “MEASURE” with value “USD\_CAP” because we want the USD/capita to work as a measure of GDP, then the rows are selected (as it was done before) and these 2 files are merged, then the rows are selected again.
       10. Employment Rate needed to be filtered at first by “FREQUENCY” with value “A”, because we want the frequency to be yearly. After, they are filtered by “SUBJECT” with value “TOT”, “MEASURE” with value “PC\_WKGPOP” because we want the percentage of working age population.

**5. Mapping (Data sample / Questions)**

* + - 1. **Question 1:** Does a better wage mean a healthier life or a longer life expectancy?
         1. - Based on: Obesity and overweight population, Stress (smokers and alcohol consumption), Life expectancy, Average Wage
    1. **Question 2:** What is the optimal number of hours to work that lead to a healthier life or more life expectancy?
       1. - Based on: Obesity and overweight population, Stress (smokers and alcohol consumption), Life expectancy, Suicide rates, Hours worked
       2. **Question 3:** What is the relationship between, more people working and being healthier and live more?
       3. - Stress (smokers and alcohol consumption) Life expectancy
       4. **Question 4:** How does adult education influence our health?
       5. - Based on: Education, Life expectancy (at birth and 65), suicide, alcohol, smoking and Obese.
       6. **Question 5:** How does air exposure affect our health?
       7. - Based on: Pollution, Cancer, life expectancy and Life Expectancy at birth and 65
       8. **Question 6:** How does self-reported happiness corelates to Alcohol consumption, Smoking habits and Suicide rates?
* Based on: Happiness, suicide, alcohol, smoking.

**Question 7:** Does a bigger social spending or a higher GDP in general influences people to live more and suicide, smoke and drink less?

* Based on: GDP, Social Spending, Happiness, suicide, alcohol, smoking.