ENSITE
Environmental Socio-Scientific Issues
in Initial Teacher Education

Module 4

ANALYSING (BIG) DATA

Worksheets

Co-funded by the
Erasmus+ Programme
of the European Union

ENSITE
Environmental Socio-Scientific Issues in Initial Teacher Education

This worksheet is based on the work within the project Environmental Socio-Scientific Issues in Initial Teacher Education (ENSITE). Coordination: Prof. Dr. Katja Maaß, UNIVERSITY OF EDUCATION FREIBURG, Germany. Partners: UNIVERSITEIT UTRECHT, Netherlands; ETHNIKO KAI KAPODISTRIAKO PANEPISTIMIO ATHINON, Greece; UNIVERSITÄT KLAGENFURT, Austria; UNIVERZITA KARLOVA, Czech Republic; UNIVERSITA TA MALTA, Malta; HACETTEPE UNIVERSITY, Turkey; NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET NTNU, Norway; UNIVERSITY OF NICOSIA, Cyprus; INSTITUTE OF MATHEMATICS AND INFORMATICS AT THE BULGARIAN ACADEMY OF SCIENCE, Bulgaria; UNIVERZITA KONSTANTINA FILOZOFA V NITRE, Slovakia.

The project Environmental Socio-Scientific Issues in Initial Teacher Education (ENSITE) has received co -funding by the Erasmus+ programme of the European Union (grant no. 2019-1-DE01-KA203-005046). Neither the European Union/European Commission nor the project's national funding agency DAAD are responsible for the content or liable for any losses or damage resulting of the use of these resources.

© ENSITE project (grant no. 2019-1-DE01-KA203-005046) 2019-2022, lead contributions by International Centre for STEM Education (ICSE) at the University of Education Freiburg, Germany. CC BY-NC-SA 4.0 license granted.

ICSE
International Centre for STEM Education

Pädagogische Hochschule Freiburg
Université des Sciences de l'Education · University of Education

Co-funded by the Erasmus+ Programme of the European Union

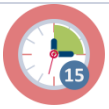
Content Index

ANALYSING (BIG) DATA 1
Worksheets 1
Worksheet 1.1a: your opinion on global warming 1
Worksheet 1.1b: Data in sources 2
Worksheet 1.2 3
Exploring data and visualizations on global temperature change 3
Worksheet 1.3: National temperature change 5
Worksheet 2.2A: Big Data and Algorithms- Smart City 6
Worksheet 2.2B: Big Data and Algorithms: Sampling Bias, Data gaps 8
Worksheet 2.2C: Big Data and Algorithms: feedback loops 10
Worksheet 2.2D: Big Data and Algorithms: Information bias 11
Optional materials for 3.1: see Addendum 11
Worksheet 3.2: Part A: Your ecological footprint 12
Worksheet 3.2 Part B: Comparing countries 12
Worksheet 3.3A: Analysing a large data set (open version) 12
Activity 3.3B: Analysing a data set 13
Worksheet 4.1: Exploring and reviewing a lesson 16
Global warming 17
Graph I 18
Graph II 18
Bar graph I 19
Bar graph II 19
The ecological footprint 21
Addendum 3
two lessons to refresh statistical techniques and skills 3
Before you start the tasks: 4
Task 1. Fill in a questionnaire 4
Task 2. Levels of measurement 4
Task 3. Visualizing data per variable 5
Task 4. Visual overview of the data 5
Task 5. Measures of central tendency 6
Task 6. Exploring the relations between variables 8
Task 7. Central tendency is not enough! 9
Task 8. Information loss in visualizations 9

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 1.1a: your opinion on global warming.

Think-Pair-share



Duration: 15 mins

Think: individually fill in the poll online or below

1. In your opinion: is global warming 'real'?

   Yes/No*

   Explain: …….

2. In your opinion: what is causing global warming?

3. On what sources do you base your opinions? Name at least 3.

Pair

Compare and discuss your results. Summarize the outcomes in a short statement, to share in the whole group.

Share

In whole group present and discuss statements and compare sources.

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 1.1b: Data in sources

15-30 **mins**

Explore ate least one of the sources you came up with in question 3 of the poll or one of the sources presented to you by your educator. The goal is to find out if and how data are used in this source. You can use the guiding questions below:

Reference to your source: ……..

- *Where do the data on which this source is based 'come from'? By whom and how are they collected?*
- *Who uses these data?*
- *How have they been analyzed (selected/filtered/combined) and represented to get their form in this source?*
- *What is the message/story of the datarepresentation/the source?*
- *Could a different representation based on the same data have been made? Would this change the story or message?*
- *What (other) data do you want or need to 'complement' this source?*

Be prepared to give a 1minute pitch on what you found out about data in your source.

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 1.2

Exploring data and visualizations on global temperature change

Duration: 45 minutes

Version A exploring data (online)

On the following websites you can explore data and visualisations on global temperature change:
https://climate.nasa.gov/vital-signs/global-temperature/

https://ec.europa.eu/eurostat/databrowser/view/sdg_13_30/default/line?lang=en

http://www.cru.uea.ac.uk/

Tasks (you may divide these in your small group):

- Study the information on the NASA website: play around with the graph, look at *the downloaded data*. Write a short review in which you describe and discuss: the information on the website, characteristics of the dataset, its source and the way data is represented. Include a comparison of the line graph and the time series on the map (what are strong points of each of those, which one do you prefer, for what audience and why?).

- Do the same for the Eurostat site. Include your reflections on the table compared to the graph.

- Compare the graphs on these sites with the graph on http://www.cru.uea.ac.uk/
  Note differences and similarities as well as strong and weak points of each.

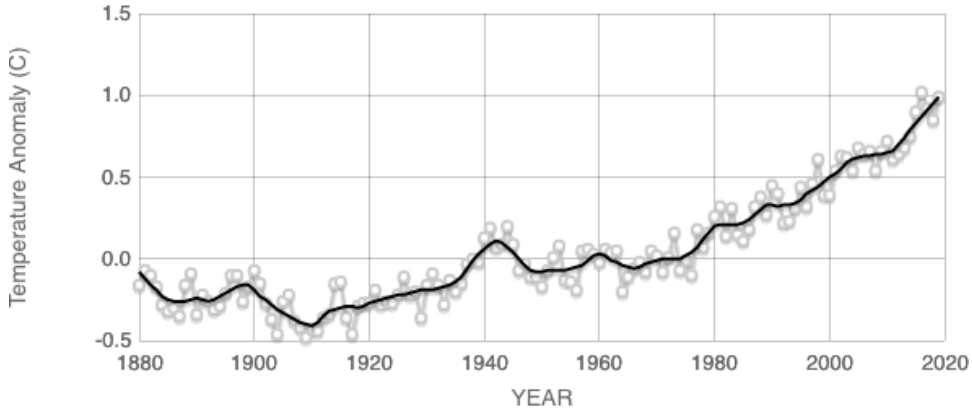
As a small group, make a short presentation on global temperature, based on the representations on one or more of these sites.

Version B (if websites are not available) – representations on the worksheet

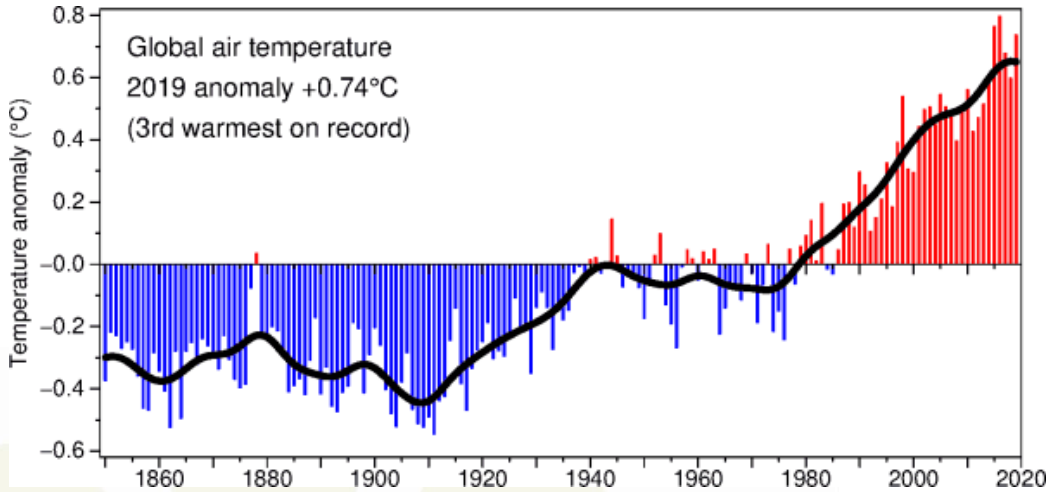
On the next page you see three different graphs representing similar data on global temperature change. Study the graphs and write a review in which you compare characteristics and strong and weak points of these graphs.

Prepare a brief presentation on these graphs for the whole group and include at least one task, problem or question.

Co-funded by the Erasmus+ Programme of the European Union



Source: climate.nasa.gov



Source: http://www.cru.uea.ac.uk/

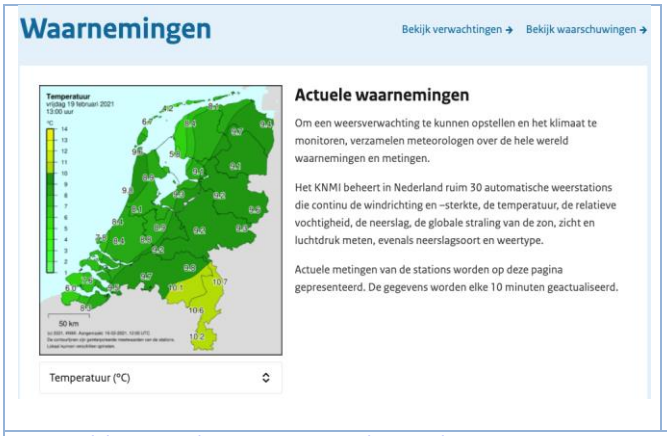
Co-funded by the Erasmus+ Programme of the European Union

Worksheet 1.3: National temperature change

Based on Dutch sources – to be used as a template

45 minutes

The *Dutch meteorological institute (KNMI)* controls 30 automated weather stations spread over the Netherlands, that constantly collect temperature and other weather data.



https://knmi.nl/nederland-nu/weer/waarnemingen

Go this website and explore the data on the map, the table and in the graph. Next answer the questions below.

1. *How many weather indicators are used? Which data are being collected?*

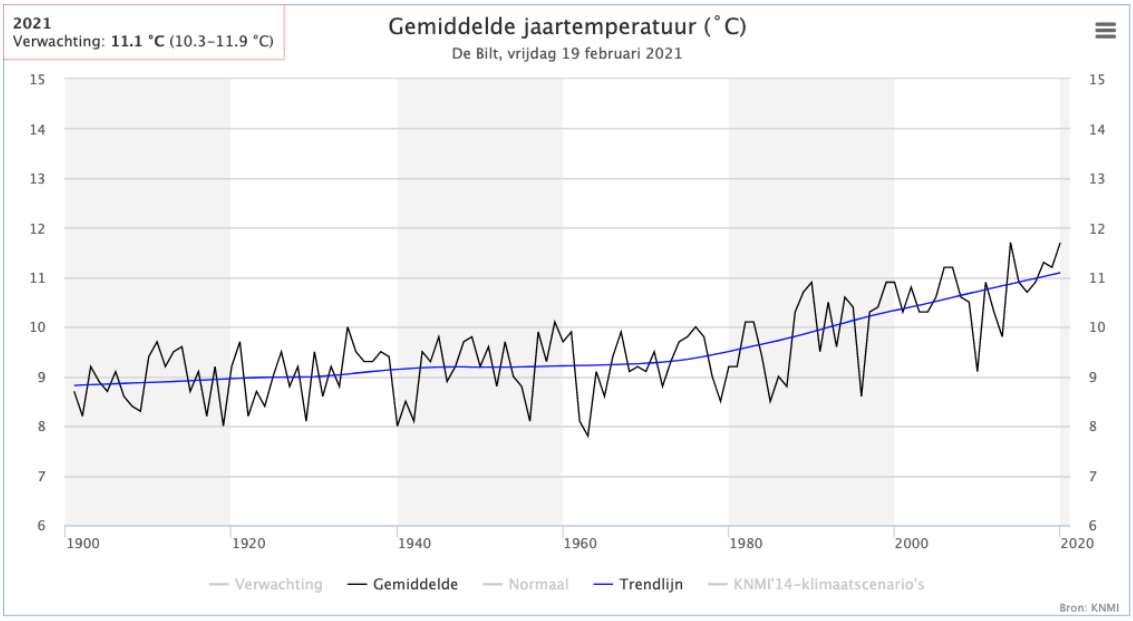
2. *How many datapoints are collected per hour? Per day? Per year?*

3. *What operations on the data in the table are required to make the temperature graph displayed on this site?*

4. *Suppose the temperature data collected (30 stations every 10 minutes) are stored and you are asked to use these data to calculate the mean yearly temperature in the Netherlands: explain your approach to do this.*

Besides the actual weather measurements, the KNMI also has a Climate Dashboard. https://www.knmi.nl/over-het-knmi/nieuws/klimaatdashboard

Co-funded by the Erasmus+ Programme of the European Union



Explore the graph on the website and answer the questions below.

5. *How does the trend compare to the datapoints?*

6. *Compare the Dutch graph with the worldwide graphs from activity 1.2.*
   - What are similarities and differences: between these graphs and data?
   - How does temperature change in the Netherlands compare to the global temperature change?
   Extra 1: Try to make a table and a graph based on the Dutch 'anomalies' (HINT: you will have to calculate these first).

The Dutch mean yearly temperatures are also represented in a 'warming stripes graphic' like the 'global' one shown to you in activity 1.2.
https://www.knmi.nl/over-het-knmi/nieuws/klimaatstreepjescode-warming-stripes

7. *Compare this stripes-graph with the line graphs: what story does each representation tell? For which audience would you use each of these representations? Justify your answer.*

Final task:
Based on your findings (in 1.2 and 1.3) make a presentation in which you summarize how your national temperature change relates to the global temperature change. Include at least one 'self-made' visual representation such as a graph/graphic/diagram to support your findings. Note: ask you educator about the required format.

Worksheet 2.2A: Big Data and Algorithms- Smart City

3 | 20 | Duration: 20-30mins

Co-funded by the Erasmus+ Programme of the European Union

Some definitions (you may look for others):

"A **smart city** is an urban area that uses different types of electronic methods and sensors to collect data. Insights gained from that data are used to manage assets, resources and services efficiently; in return, that data is used to improve the operations across the city. This includes data collected from citizens, devices, buildings and assets that is then processed and analyzed to monitor and manage traffic and transportation systems, power stations, utilities, water supply networks, waste, crime detection, information systems, schools, libraries, hospitals, and other community services." Source: Wikipedia

"In general, a **smart city** is a **city** that uses technology to provide services and solve **city** problems. A **smart city** does things like improve transportation and accessibility, improve social services, promote sustainability, and give its citizens a voice."
Source: https://blog.bismart.com/en/what-is-a-smart-city

"Big data offer the potential for cities to obtain valuable insights collected through various sources and sensors in the real-world environment." (Source Hashem e.a., 2016).

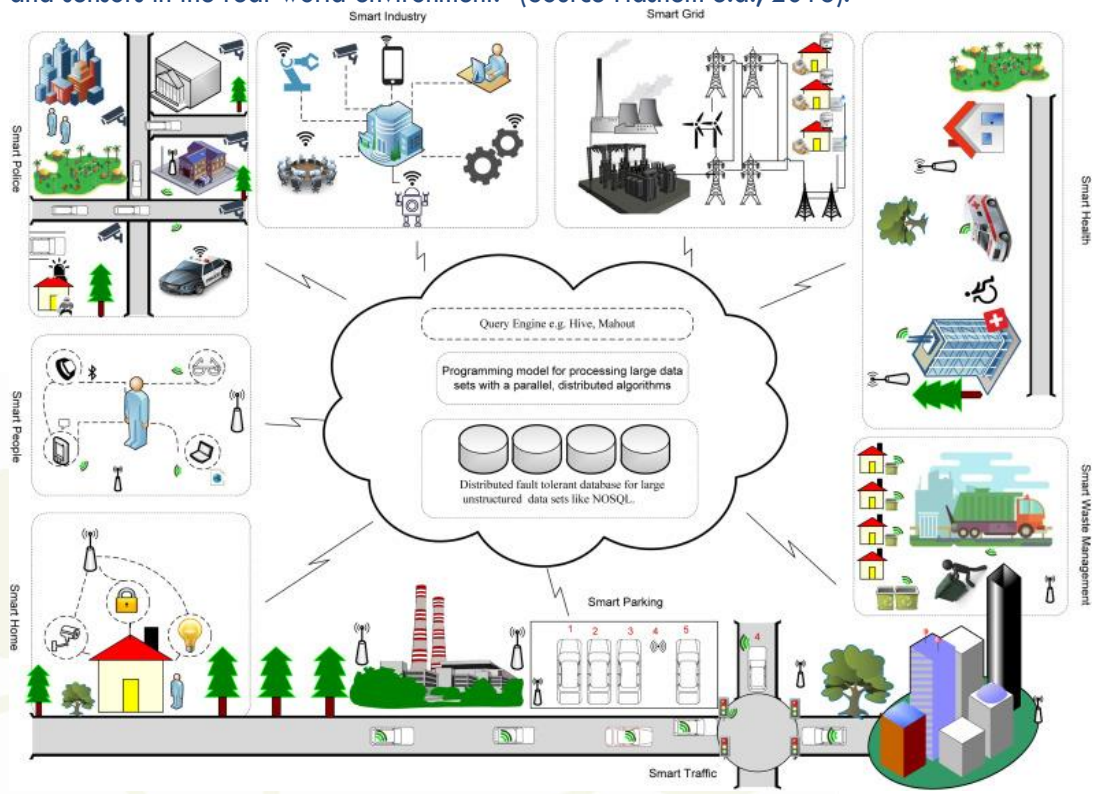


Figure: Landscape of the smart city (Hashem e.a., 2016, p. 749)

Questions for discussing the topic:

- What are characteristics of a smart city? Do you know or do you experience elements of a smart city in your environment?
  - What data are collected and how? For what purpose?
  - Which data are (maybe) connected?
  - What 'algorithms/patterns' maybe used to make decisions?
  - Would you like to live in a smart city? Why (not)?

- How do big data and algorithms impact policy measures or decision making in the context of smart cities? What ethical issues need to be discussed?

Co-funded by the Erasmus+ Programme of the European Union

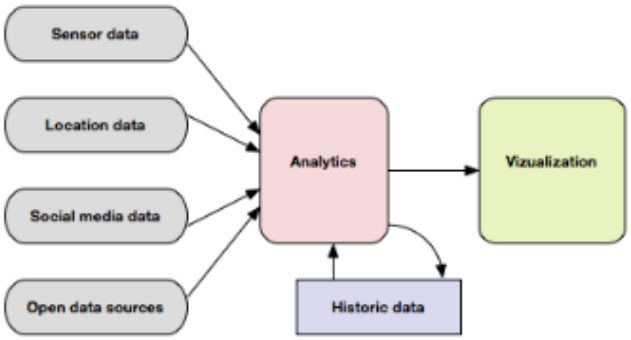


Figure 2. From data to decisions.

Source: Berntzen et al., 2018

Further reading:
https://openarchive.usn.no/usn-xmlui/handle/11250/2682133

https://medium.com/predict/the-smart-city-dilemma-privacy-vs-convenience-9efb2a45c26

Worksheet 2.2B: Big Data and Algorithms: Sampling Bias, Data gaps

Co-funded by the Erasmus+ Programme of the European Union

Duration: 20-30mins

Caroline Criado Perez, author of the book "Invisible Women: Exposing Data Bias in a World Designed for Men", is concerned with gender data gaps.

- *Find information on the internet about the notion of a 'gender data gap'. Describe the notion and give at least one example of a gender data gap.*
- *Think of other groups of people that might be underrepresented in data collections and consequently create biases in the analyses and the presentation of the results.*
- *Check presentations of data-based research on their account of the data collection. Does the context of the sample (location, selection, …) cover the reach of the presented results?*

Further reading:

Giest, S., Samuels, A. 'For good measure': data gaps in a big data world. *Policy Sci* **53,** 559–569 (2020). https://doi.org/10.1007/s11077-020-09384-1

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 2.2C: Big Data and Algorithms: feedback loops

Duration: 20-30mins

Many policy measures are nowadays supported by algorithms based on big data collections. In particular this is the case with safety measures (e.g. where and who to search for weapons on airports) or to detect fraud (e.g. who's tax returns to check).

Read this fictional example[1]:

Suppose that a population exists of two equally sized group: the Hippos and the Raves. Statistics show that the Hippos are responsible for 51% of all crimes and the Raves for 49%. Further research shows that 51% of the Hippos are involved in these crimes, while 49% of the Raves show in criminal behaviour. The police want to be more efficient and decides to start to not randomly check 1000 inhabitants for one month, but to check 510 Hippos – of which 51% is criminal resulting in arresting 260 criminals. At the same time the police check 490 Raves – of which 49% is criminal, resulting in 240 arrested criminals. The police is satisfied with the results of the month, and argue that even 52% of the 500 criminals appeared to be a Hippo. Next month they decide to check 520 Hippos and 480 Raves. And what happened? That month even 53% of the arrested criminals appeared to be a Hippo. The method was implemented and after two year 73% of all criminal behaviour was assigned to Hippos.

Tasks:

1. Check the calculations and discuss the example.

2. In 2019 in New York 95% of the black boys of 20 years old were at least once checked because of 'reasonable suspicion', while this happened with 17,5% of the white boys of the same age. Can that be the results of a similar mechanism?

Further reading: **https://en.wikipedia.org/wiki/Algorithmic_bias** .

[1] This example is based on the essay "510 staandehoudingen en de zelfversterkende feedback-loop" by Ionica Smeets, which was published in the Dutch newspaper Volkskrant of June 21, 2019.

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 2.2D: Big Data and Algorithms: Information bias

Duration: 20-30mins

Discuss in your group:
What is **your** primary source for actual news: traditional media? Social media? Other…..?
Compare this to the sources discussed in activity 1.1.

Statement: *The number of young people reading newspapers decreases. This leads to filter bubbles and fake news.*

- Discuss the statement above. Do you agree or not? What can be the mechanism behind this happening here? How do big data and algorithms play a part in this?

Read the Introduction and the part titled 'Dangers' in the following article on Wikipedia: https://en.wikipedia.org/wiki/Filter_bubble

- Do you recognize the dangers described here? Do you want to change your opinion about the statement above?

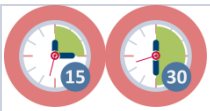
Optional materials for 3.1: see Addendum

Optional 2 lessons of 45-60 mins

The materials in The addendum (at the end of this file) can be used to refresh (basic) statistical skills for analysing and visualising data (lesson 1) and to practice reason about advanced data visualisations (lesson 2)

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 3.2: Part A: Your ecological footprint



15 - 30 mins

Part A:
Individually
Go to the website: https://www.footprintcalculator.org/ to calculate your personal footprint.
Pair
In a pair compare your personal footprints and discuss similarities and differences. What causes these?
Share
Estimate and discuss (based on the results of all pairs) what you expect your national ecological footprint to be in 'number of earths'.

Worksheet 3.2 Part B: Comparing countries

20 min

Go to the website: https://data.footprintnetwork.org and explore the data of the country your group has been assigned. Be sure to do the following:

- *Study the three graphs and their tables for the country: what data are used in each visualisation? How are the graphs connected? (click on "learn more')*
- *Summarize the trends for the country in terms of the ecological footprint, deficit/reserve and biocapacity and prepare to explain these trends briefly in the whole group.*

Optional (preparation for 3.3): further explore the data and visualisations on the website.

- *What data are used to 'determine the footprint' (the deficit and the biocapacity) and how are these used?*
- *Study the data tables: what are characteristics of the data? How have they been measured/collected? how are the data organized in the table?*

Worksheet 3.3A: Analysing a large data set (open version)

30 min

Use the excel data file presented to you by your educator or download it here:
Database in excel format

Co-funded by the Erasmus+ Programme of the European Union

Using excel make a combined line graph comparing the 'ecological footprint per person' and 'the biocapacity per person' over time for two countries (which you may select yourselves). The following questions may guide you in doing so.

Exploration: understanding the data-file

- How many rows? How many columns? How many cells? What types of data?
- Why are some numbers big and others small (even in the same column)?
- What is meant by the headings of columns? What special columns are there? What is in them?
- How is the data organised in the table?
  HINT: use the codebook to further explore and understand

Selecting the required data

- Which countries do you wish to compare overtime? Why these?
- Where can you find the variables you need (in which columns)?
- Filter the data to have just the ones you need to make your graph.
  HINT: paste these data in a new empty Excel file.

Representing the data

- Use the tools in excel to make the combined line graph as indicated in the task (see top of page).
- What is the story for this graph?
- Can you adjust (manipulate) the graph in such a way that the story will be different?

Reflection

Compare your graph to the two graphs of the countries you selected from the website.

- What are advantages of combining the graphs of two countries?
- What are drawbacks?

Activity 3.3B: Analysing a data set

2 | 30 min

Use the excel data file presented to you by your educator to make a combined line graph comparing the 'ecological footprint per person' and 'the biocapacity per person' over time for The Netherlands and Finland. The following steps guide you in doing so.

Step 1: Download the data (if not yet available for you)
Database in excel format derived from data.footprintnetwork.org (1961 - 2017)

Have a close look at the data. We will concentrate on the two variables:

- Ecological footprint per person (EFConsPerCap)
- Biocapacity per person (BiocapPerCap)

Step 2: Two separate files

Co-funded by the Erasmus+ Programme of the European Union

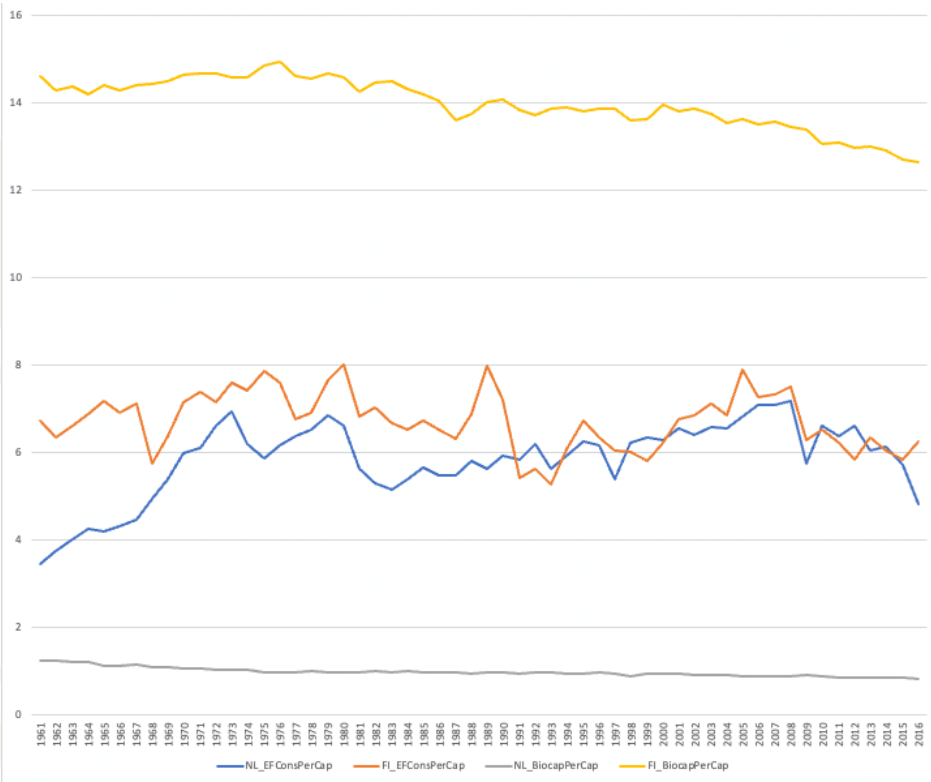
- Make two separate excel files: one for Finland (code 67) and one for the Netherlands (code 150).
    - To do so, take out (copy) all rows with the correct country code (from the excel file with all countries) and save these in the separate file(s). Be careful to also copy the first row (with the column names).
- Sort the data by 'record' (that's the column with the variable names). Now all data from EFConsPerCap and BiocapPerCap can be copied easily.

Step 3: One new file to combine data and make the graph

- Make a new excel file where you will combine the data of the two countries, like this:

	A	B	C	D	E
1	Year	NL_EFConsPerCap	FI_EFConsPe	NL_BiocapPe	FI_BiocapPerCap
2	1961	3,441279604	6,72844908	1,24884015	15
3	1962	3,757064707	6,34715741	1,25412626	14,2721372
4	1963	4,030625503	6,59753656	1,21405536	14,3614576
5	1964	4,263565614	6,8900676	1,22610153	14,1894186
6	1965	4,207671426	7,16218402	1,13798012	14,3847131
7	1966	4,31488108	6,90693547	1,10958268	14,2690707
8	1967	4,472822469	7,11714645	1,14716588	14,4039301
9	1968	4,950798414	5,75129775	1,09716033	14,4111914
10	1969	5,373565251	6,38517923	1,09200401	14,4834109
11	1970	5,989664773	7,14404093	1,05979802	14,6272478
12	1971	6,117044515	7,38592589	1,07292496	14,6694126
13	1972	6,600457811	7,15770908	1,03539347	14,661672

Step 4: Now you can make a line-graph of the four columns, using the tools in excel. Note: you can also 'color' the regions between line-graphs.



- What is the story for this graph?
- Can you adjust (manipulate) your graph in such a way that the story will be different?

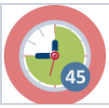
Reflection

Co-funded by the Erasmus+ Programme of the European Union

Compare your graph to the two graphs of The Netherland and Finland from the website.
- What are advantages of combining the graphs of two countries?
- What are drawbacks?

Co-funded by the Erasmus+ Programme of the European Union

Worksheet 4.1: Exploring and reviewing a lesson



45 minutes

For one of the teaching materials in the appendix do the following:

Individually

- Work through the tasks in the materials as if you were a secondary school student (15 min).
  Note: you may use the outcomes of the activities in section 1 and 3 of this module.

In a small group

- Share your results on the lesson tasks and discuss your experiences and opinions about: the level of difficulty – the time it took - your interest in the topic – improvements you would make when teaching this lesson to (your) secondary school pupils (and the reasons why) etc.

- Find out how this topic fits in the curriculum of your teaching-subject (or other STEM subjects). Which curricular topics and goals are addressed in this lesson?

- Think about what is needed for you to teach this lesson.

Be prepared to share your findings in the whole group.

Co-funded by the Erasmus+ Programme of the European Union

APPENDIX 1A – Exemplary teaching materials on global Warming for lower secondary school students – to be used in activity 4.1

Global warming

You may have heard people say that the world temperature is rising. Ice on the north and south pole is melting and summers seem to be warmer. This is called global warming.

Source: https://pixabay.com/nl/photos/ijs-ijsberg-gletsjer-ijsland-water-3544836/

1. ” What do you think: is global warming happening?”
    a. Do you notice signs of global warming in your own environment?
    b. Do you hear of global warming at school, at home in the news or other media? What is the message?
    c. What sources would you need to be sure global warming is happening?

Worldwide organisations like NASA collect data on temperature. They observe **temperature anomalies** in the data. This means that they look for **differences** between the measured temperature and the **average** temperature over a long time.

2. On the next page you see two graphs of ‘temperature anomalies’. Study these graphs carefully. For each graph write if it supports the statement: ‘global warming is happening’. Also explain how the graph does or does not support this statement.

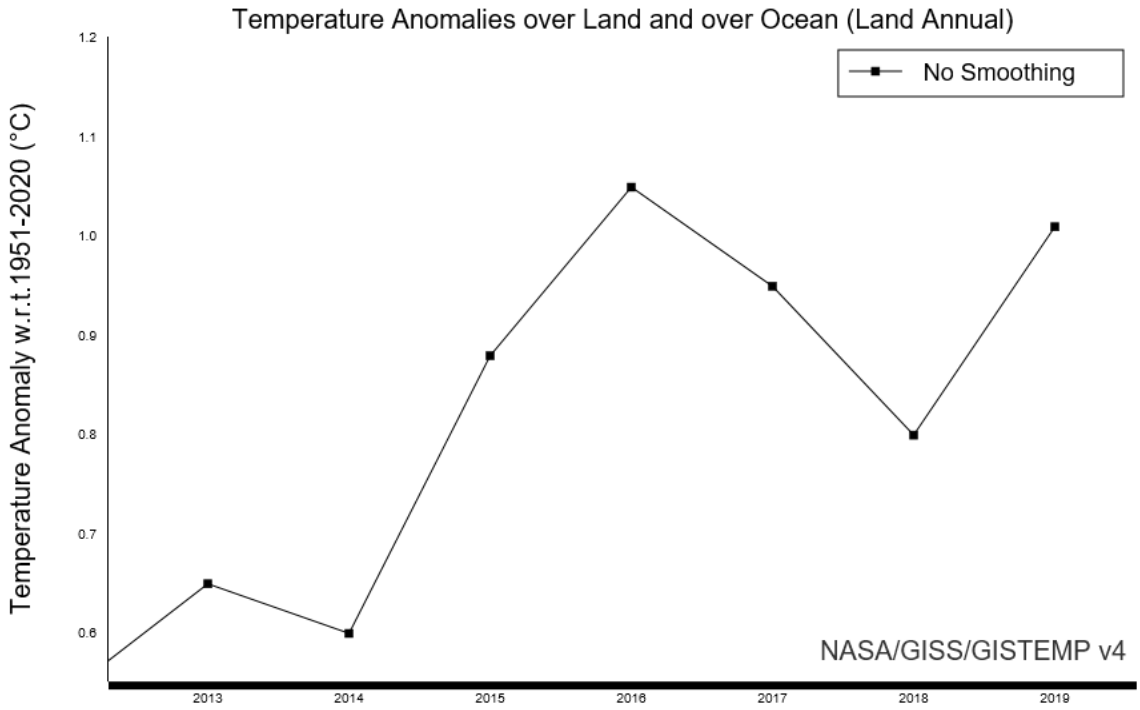
    Graph I supports/does not support this statement because,
    ……

    Graph II supports/does not support this statement because,
    ……

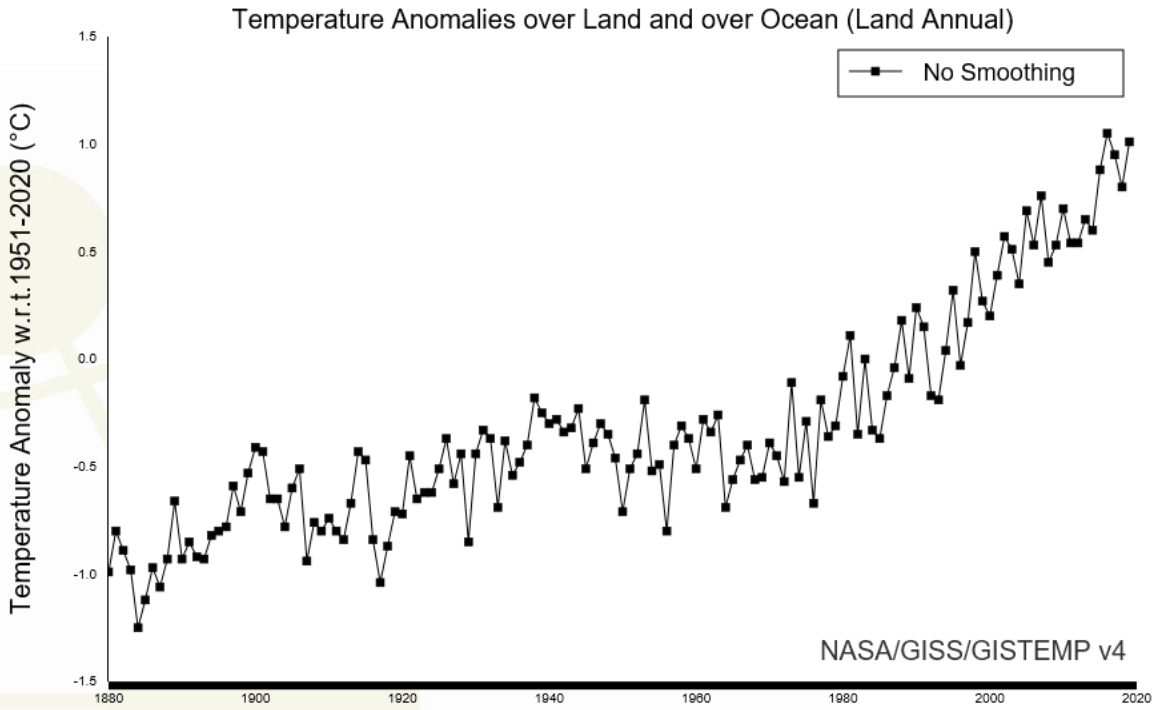
3. Compare your answers in class. What conclusions can you derive from the graphs?

Co-funded by the Erasmus+ Programme of the European Union

Graph I



Graph II



Source:

- GISTEMP Team, 2021: *GISS Surface Temperature Analysis (GISTEMP), version 4*. NASA Goddard Institute for Space Studies. Dataset accessed 2020-06-10 at data.giss.nasa.gov/gistemp/.
- Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, 124, no. 12, 6307-6326, doi:10.1029/2018JD029522.

Co-funded by the Erasmus+ Programme of the European Union

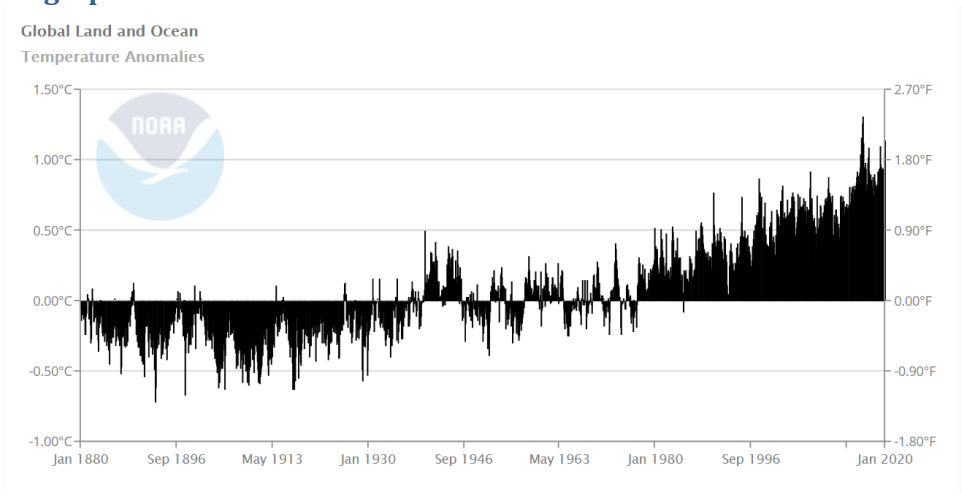
4. Both graphs I and II on the previous page are based on the same data about temperature. Look again very closely at the graphs.

    a. Explain what is the same and what is different.

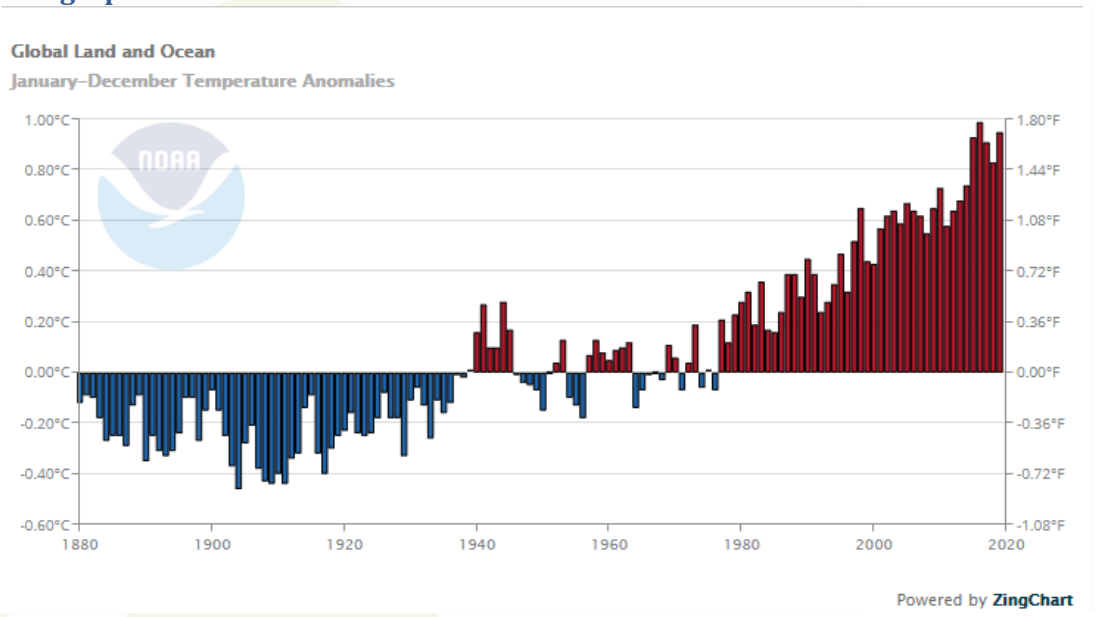
    b. How is it possible that the same data give two different views on the situation?

Below you see two bar graphs on global land and ocean temperature anomalies form a different source. The data were collected by NOAA (National Oceanic and Atmospheric Administration). Both bar graphs are made using the same data set.

Bar graph I



Bar graph II



5. Which of the two bar-graph gives more information? How do you know? Can you give an example? What are the lowest and highest temperature anomalies for each bar graph?

6. Compare the line graphs of NASA and the bar graphs of NOAA

    a. Do these graphs show a similar trend?

Co-funded by the Erasmus+ Programme of the European Union

b. Do these graphs from have similar data? How do you know?

7. Discuss which graph(s) you would use to show that Global warming is happening. How does this graph support your argument.
……
……
……
……
……

8. Do you think it is possible to use the temperature anomalies from the NASA and draw a graph for these data to support the statement that Global Warming is not happening?

   What would such a graphs show?

Co-funded by the Erasmus+ Programme of the European Union

APPENDIX 1 B – Exemplary teaching materials on the ecological footprint for lower secondary school students – to be used in activity 4.1

The ecological footprint

The ecological footprint is a measure to compare how much ecological resources are used by individuals, groups and countries against the earth's capacity for biological regeneration. Humans now worldwide use as much ecological resources as if we lived on 1.6 Earths. This means this resources are quickly diminishing.
There are large differences between countries in the size of their ecological footprint and in how this develops over time.

On the next page you see the graphs of two countries showing the trend of ecological footprint and biocapacity from 1961 to 2016. Using the information on the graphs answer the following questions:

1. What are the trends in each of the countries (Germany and Pakistan) in terms of the ecological footprint? Please explain briefly.

2. What are the trends of these countries in terms of biocapacity? Please explain briefly.

3. What are the similarities and differences between the trends of these countries?

Worldwide there are 4.7 biologically productive acres available per person, and this doesn't include all of the other plants' and animals' needs.

4. Based on this information and the graphs, how fair is the consumption of given countries comparing with the world?

Co-funded by the Erasmus+ Programme of the European Union

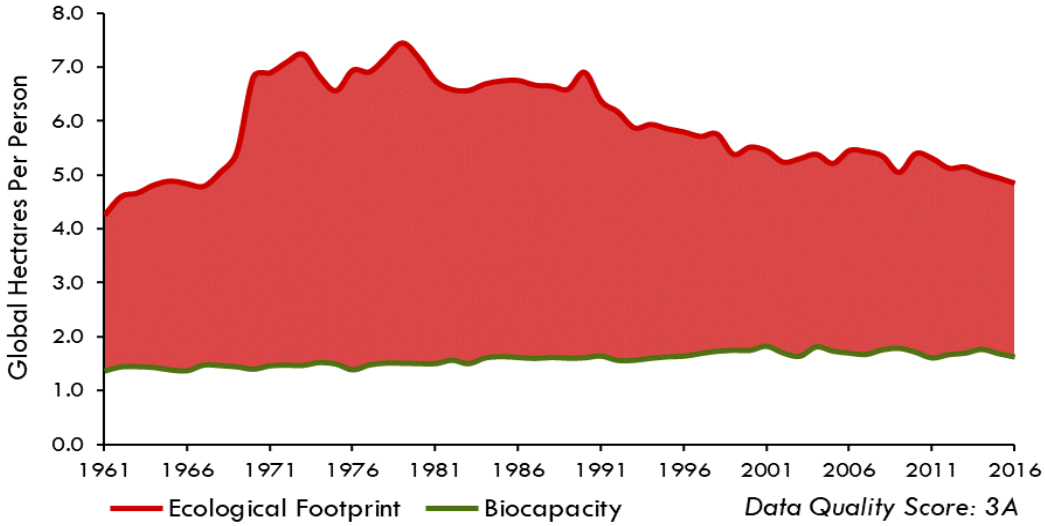


Figure 1: Germany

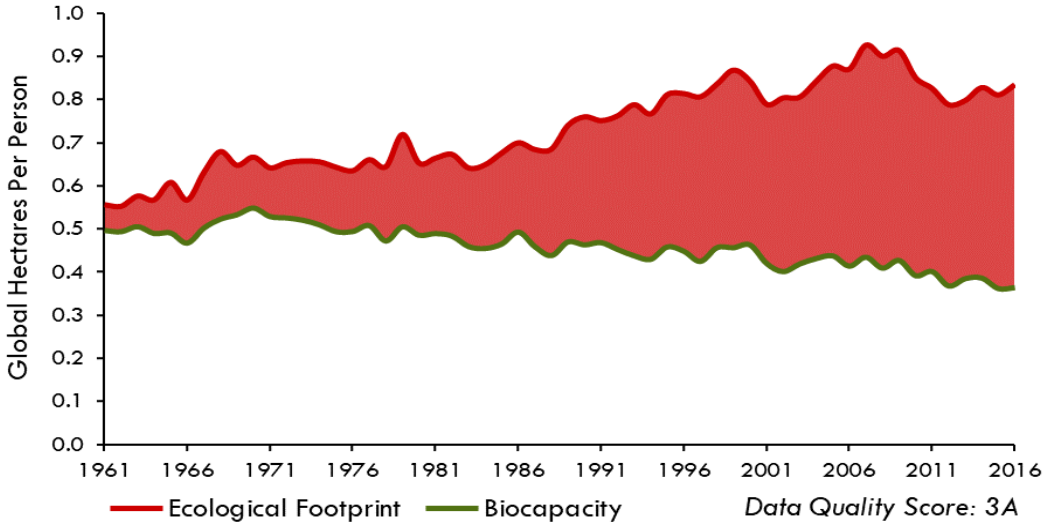


Figure 2: Pakistan

Source: 2019 Global Footprint Network-Open Data

Co-funded by the Erasmus+ Programme of the European Union

Below you see similar graphs as on the previous page for two other countries. Look closely to these graphs and answer the questions below.

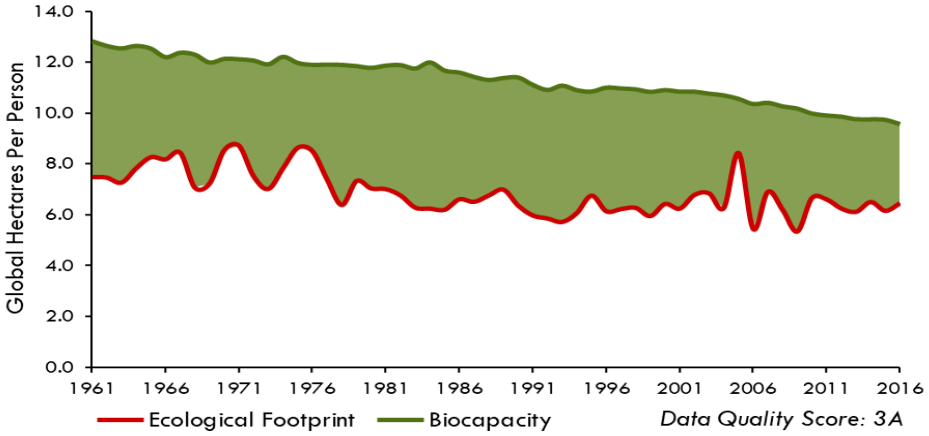


Figure 3: Sweden

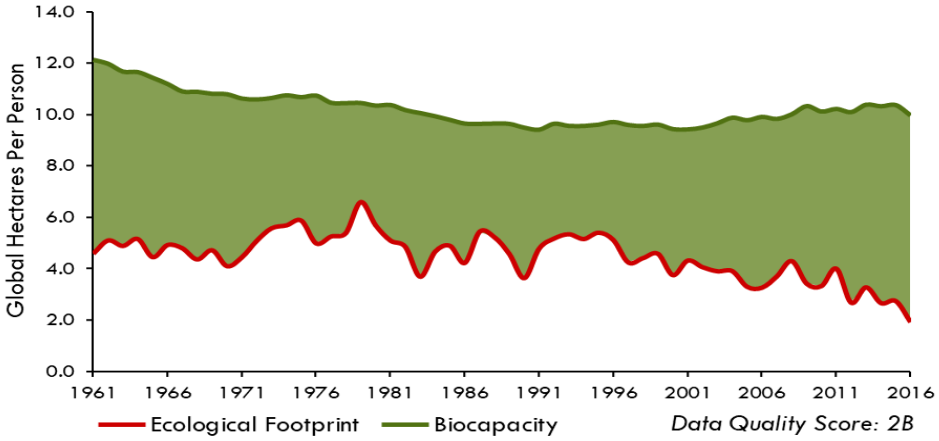


Figure 4: Uruguay

Source: 2019 Global Footprint Network-Open Data

1. What are the trends of each of these countries (Sweden and Uruguay) in terms of the ecological footprint? Please explain briefly.
2. What are the trends of these countries in terms of biocapacity? Please explain briefly.
3. What are the similarities and differences between the trends of these countries?
4. Worldwide there are 4.7 biologically productive acres available per person, and this doesn't include all of the other plants' and animals' needs. Based on this information and the graphs, how fair is the consumption of given countries comparing with the world?
5. What are the main similarities and differences between the first two graphs and these two?

Co-funded by the Erasmus+ Programme of the European Union

Addendum

For activity 3.1

two lessons to refresh statistical techniques and skills

Co-funded by the Erasmus+ Programme of the European Union

Lesson 1

Before you start the tasks:

Open the link below and make and save a copy of this file (google sheet) for yourself to work in.

https://docs.google.com/spreadsheets/d/12Rwfftm-MOm2S43Bls6xWOXa3Hn-tf3wxX41nDKoF04/edit#gid=0

Tip: Use the tabs to navigate between sheets as is indicated in the tasks.

Task 1. Fill in a questionnaire

Answer the small questionnaire below. Fill your data in the sheet (tab) *FillInData*, or add to the existing data in the sheet *PreFilledData* (based on the teacher's instruction).

Questionnaire

1. Name (Pseudonym)
2. Gender
    - Female (1)
    - Male (2)
    - Other (3)
3. Your height (cm)
4. Your size of shoes (European size)
7. What is the level of you statistics knowledge and skills?
    - I know nothing (1)
    - I know basic concepts (2)
    - I am quite familiar with the main ideas (3)
    - I am an expert user (4)
    - I am happy to assist in teaching (5)
8. What was your favorite subject at school?
    - Physics (1)
    - History (2)
    - Language (3)
    - Mathematics (4)
    - Other (5)

Task 2. Levels of measurement

There are four main levels of measurement (also called types of data):

- Nominal: the entities can be grouped by a label
- Ordinal: it is possible to compare the entities based on the measurement: more/less
- Interval: each entity is assigned with a number per entry and the distance between numbers is meaningful
- Ratio: each entity is assigned with an exact number per entry and the ratio of two numbers is meaningful

Consider watching this video in which the levels are explained
https://www.youtube.com/watch?v=eghn__C7JLQ

For each of the measurements from Task 1, write which type of data it represents:

Co-funded by the Erasmus+ Programme of the European Union

- Name
- Gender
- Height (cm)
- Size of shoes (European size)
- Statistics knowledge
- Favorite subject

Task 3. Visualizing data per variable

3a: Visualize data manually

Look at the visualization of the prefilled data
(Sheet *ManualVisualisation:Height*)

Built the visualization of the data yourself (create new sheets).
Use the following variables:

- Height (for your own group; or add your own data to the ready-made diagram)
- Statistics knowledge (for your own group, or for prefilled data)
- Favorite subject (for your own group, or for prefilled data)

Discuss the differences and similarities between these visualizations:

- what does the vertical axis represent?
- what does the horizontal axis represent?
- what does the order of the bars mean? How is it related to the type and level of the variable?
- what does the width of the vertical bars mean? How is this related to the type and measurement level of the variable?

(3b) Build a histogram automatically

Watch the video on how to build a histogram in Google Sheets:
https://www.youtube.com/watch?v=9LCJ33MnOIA

Apply this to build a histogram of your data of height.
Manipulate the width of the bars by adjusting the width of the "Buckets" (intervals). How does the diagram change? Can you make it the same as your manual visualization?

Built the diagrams automatically for other variables.

Discuss: For which variables is this type of data visualization suitable? For which not? Why? What does the horizontal axis represent in each case? In which cases would you call it a histogram?

Task 4. Visual overview of the data

You may use another program that allows you to create bar charts and circular diagrams directly from your data. In that case, go to tasks (c) and (d).

(a) Summarize the variable Gender manually

Co-funded by the Erasmus+ Programme of the European Union

Open the Sheet *ManualSummary*. Based on the data of your group (or prefilled data), fill in the summary.
Compare with the automatically built data on the sheet *PivotTable:Gender*
What do you notice? Is there a difference? If so, why?

(b) Compute the summaries automatically

Compute a Pivot Table for the variables in your data (or other pre-filled variables).
Read the instruction:
https://support.google.com/fusiontables/answer/2592773?hl=en
Watch the video if needed:
https://youtu.be/Tty0RyD1KLw

Summarize each variable in a new sheet.

(b) Build a bar chart

Read the instruction:
https://support.google.com/docs/answer/9142829?hl=en
Watch the video if needed:
https://www.youtube.com/watch?v=-x_mBMkB9KQ

Build bar charts for each variable from the Pivot Tables.
- What is the difference between a histogram and a bar chart for Height?
- For which variables a bar chart makes sense, but a histogram does not?

(c) Built the pie chart (a circular diagrams)

https://support.google.com/docs/answer/9143036?hl=en
Watch the video if needed:
https://www.youtube.com/watch?v=sVz-5Sm2Y-Q

Built pie charts for each variable from the Pivot Tables.

For which variables a pie chart is better than a bar chart? Why? What is the difference between these two visualizations?
What does a pie chart allow seeing better than a bar chart?
What does a bar chart allow seeing better than a pie chart?

Task 5. Measures of central tendency

A few measures of central tendency are used in most of the analysis: (arithmetic) mean (or average, center of mass), mode, median.

Watch this video:
https://www.youtube.com/watch?v=k3aKKasOmlw
Formulate in your own words, what each of these measures of central tendency means.

Calculate each measure for the variables in your data set
- Gender:
- Height
- Statistics knowledge
- Favorite subject

Co-funded by the Erasmus+ Programme of the European Union

What does the number mean in each case? In which cases these numbers are not meaningful?

Conclude, which central measures are appropriate for which level of measurement:

	Mean (average)	Median	Mode
Nominal			
Ordinal			
Interval			
Ratio			

Co-funded by the Erasmus+ Programme of the European Union

Task 6. Exploring the relations between variables

In this task, you will be answering questions about relations between two variables, using visualization. Mind, those answers stand only for your group of participants and cannot be generalized for the entire population.

(a) Do people of different genders have the same height?

Answer this question relying on your data:
- Look at the Sheet *PivotTable:HeightGender* that expresses Mean of the Height for each gender. Build the same for your data (or include your data to initial data).
- Which visualization would help you to answer this question? Build it.
- What is your conclusion?

(b) Is statistical knowledge related to gender?

- Think, what is different between this question and the previous question. Which parameter will you summarize in the Pivot Tables?
- Built Pivot Tables and suitable visualization to answer the question.

(c) Do people who like different subjects have different statistical knowledge?

- Think, which Pivot Table you will need to build to answer this question?
- Built a bar chart. Which variable is on x-axis? Which on y-axis? Think, what each color means.
- Try to build different bar charts answering this question: position different variables along the x-axis; use cumulative diagram and non-cumulative. Which visualization helps better in interpreting the data?

(d) What is the relation between height and size of shoes? Are they tightly interconnected or not?

- To answer this question use Scatter chart. https://support.google.com/docs/answer/9143294?hl=en

Watch this video if needed:
https://www.youtube.com/watch?v=YC1Is9YJX0k

Adjust minimal and maximal values on the axes. When do you easily see the relation? How can you make it almost invisible?

(e) Pose other questions about relations between variables. Find a way to answer the questions! (The task can be done in pairs, asking a peer to answer the question).

Co-funded by the Erasmus+ Programme of the European Union

Task 7. Central tendency is not enough!

Look at the Sheet *Spread:HeightGender*.
In this sheet, not only the mean value of height for each gender is calculated, but also the standard deviation of the height. This parameter allows seeing how different participants in the group are. In the second PivotTable you can see that variability of the height is similar for two genders.

Come back to the *PreFilledData*. Alter the height of some students from the same gender so that:

(a) *Mean* does not change, but *standard deviation* changes a lot.
(b) *Standard deviation* stays approximately the same, but *mean* changes

(You may use unrealistic values for the height)

Unfortunately, Google Sheets do not provide a straightforward way of working with Standard deviation and Error bars.

Yet, on the Bar Chat in the document you can see Error bars for the initial data. Error Bars on the Bar Chat do not immediately represent changes in the data that you make. You may manually change it by clicking on an error bar, entering a new value for standard deviation at the place for "Constant".

You may watch this video to see how the Bar Chat with error bars was built.
https://www.youtube.com/watch?v=_R8aFZ_I1cw

Task 8. Information loss in visualizations

Look back to the diagrams that you have built in tasks 3, 4, and 7.

Think: in which diagrams information about values of *each case* is preserved? In which diagrams some information is lost? What exactly is lost? How might the loss of information mislead our conclusions?

Co-funded by the Erasmus+ Programme of the European Union

Lesson 2

Tasks

Work on the tasks (1)-(4) using the ***Diagrams*** on the next pages one by one. For the tasks (5) and (6) consider all ***Diagrams*** together

(Task 1) Which story does the diagram tell?
What do you notice first of all? Discuss and describe your first impression. What story does the diagram tell?

(Task 2) Let's get deeper:

a) Which variables are depicted?
b) How are they represented in the diagram?
(Think of: size, color, height, area, etc.)
c) Which type of values is used for each variable?
(Think of: absolute value, relative values, percentage, mean, variance, etc.)
d) What is the level of measurement of each variable?

e) Answer the questions below the diagram.
Which aspects are easy to see? Which aspects are difficult to notice? Which questions require additional data to answer?
Write down what you noticed that was not clear at the beginning.

(Task 3) Let's get critical:

a) How is a particular story highlighted in the diagram? Think of the colors, the order of grey tones, the order of entities, etc. What pop-ups for you, and what stays less noticeable? How do these visual effects help the authors to tell their story?

b) What are other aspects that might influence how the story is perceived? (Think of the period for which data is represented. Recall, what the levels of measurement for each variable are. How are the groups chosen and why?

c) Which other values for the same variables might tell a different story? (Think of alternating absolute and relative values, think of central tendencies versus variability)

(Task 4) Let's get creative!

a) Ask your own questions and search for the answers based on the diagrams and beyond.
Which story might the diagram not reveal?

b) Look at the diagram 4 and consider the story it tells. Use the data from which the graph (4) has been built. Think of another story and try to tell it with the same data. You may transform the data if your wish.

(Task 5) Combine the stories: a glimpse on complexity

Co-funded by the Erasmus+ Programme of the European Union

Come back to the diagrams (3), (5), and (6). Combine the information from these diagrams. What can you observe? How money is distributed across the world? How does it change in time? Discuss it with your peers.

Discuss the following statement:
"Some countries reduced inequality successfully and thereby reduced poverty. Lower inequality in the future can further reduce poverty. But because the average income in the majority of countries in the world is much lower than $30-poverty-line, strong growth is necessary for global poverty to decline." (by Max Roser, January 11, 2022)

(Task 6) Different visualizations for different stories: ground your choices!

a) Come back to all the diagrams that you have worked with. Think, what are the weak and strong aspects of each visualization? In which situations you would use this type of a diagram? Discuss in groups and write a short summary.

b) What are the aspects of a visualization that you need to care about so that the visualization represents the story trustfully?

c) What are the aspects of a visualization that might help you in highlighting your story?

Co-funded by the Erasmus+ Programme of the European Union

Diagrams

Note: work on the ***Tasks*** presented above, applying them to each diagram.

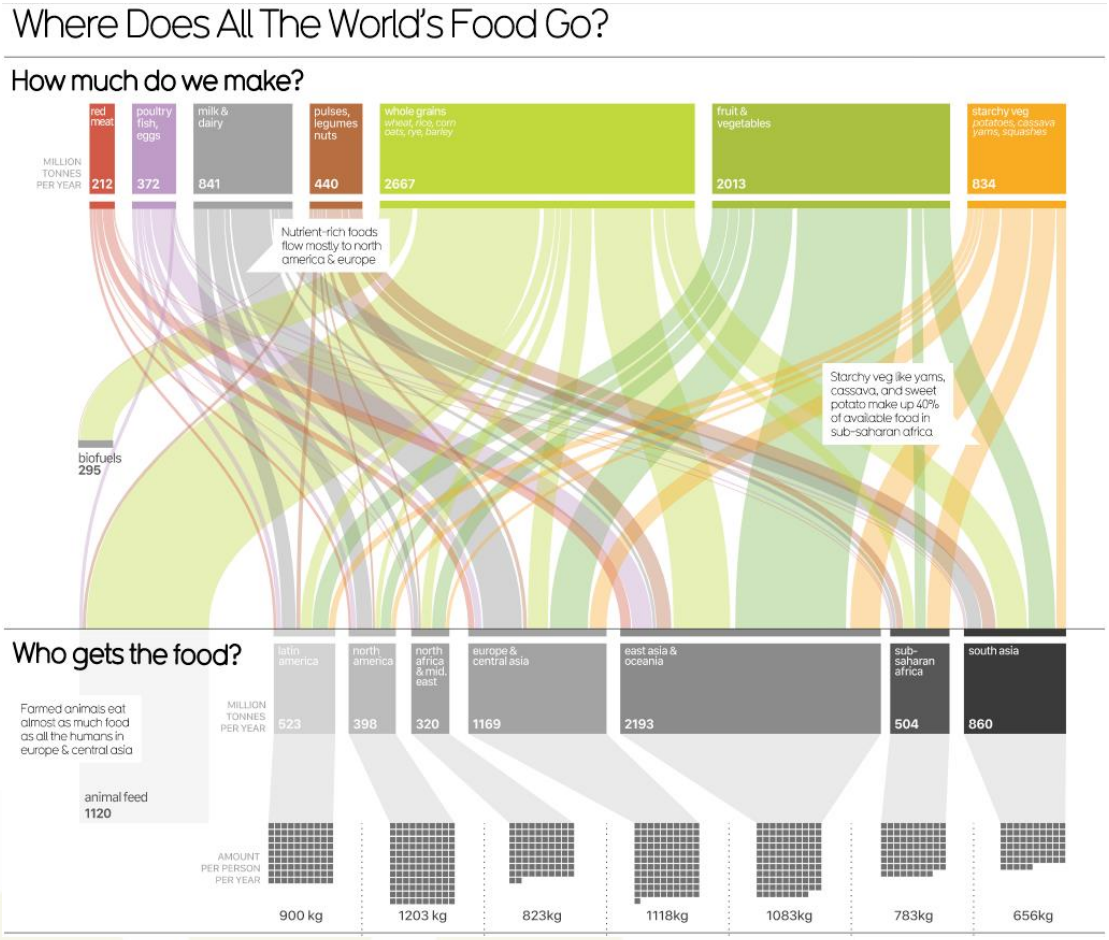
(Diagram 1) Cartogram: interesting way to represent data
https://ourworldindata.org/world-population-cartogram

Questions:

Where do more people live, in China or in Africa? In Japan or in Australia?
Which countries of Europe are the most populated?
In which counties in Europe do people live less dense than in others?
In which countries in the world is the density very low?

Co-funded by the Erasmus+ Programme of the European Union

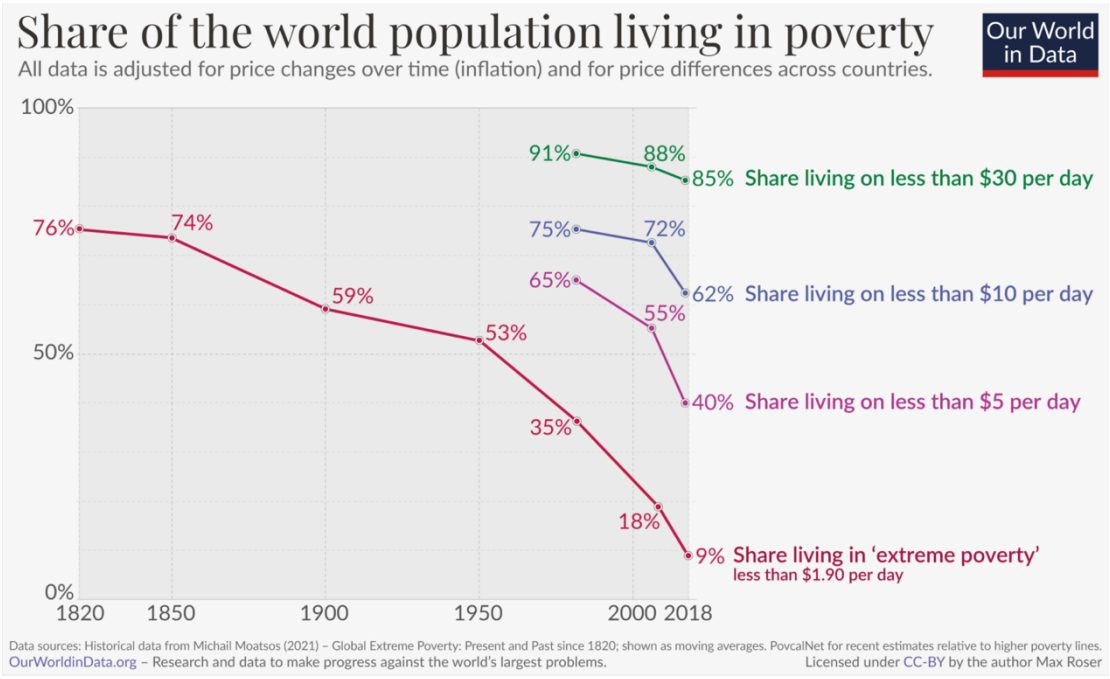
(Diagram 2) Sankey diagram: Who eats what?
https://informationisbeautiful.net/visualizations/global-food-supply-where-does-all-the-worlds-food-go/



Questions:
People of which region eat the biggest part of vegetables and fruits?
People of which region eat the biggest part of red meat?
Which type of food is lacking in South Asia?
In which region are starchy vegetables the most popular in a diet?
In which region are vegetables the most popular in a diet, compared to other regions?
In which region is red meat the most popular in a diet, compared to other regions?

Co-funded by the Erasmus+ Programme of the European Union

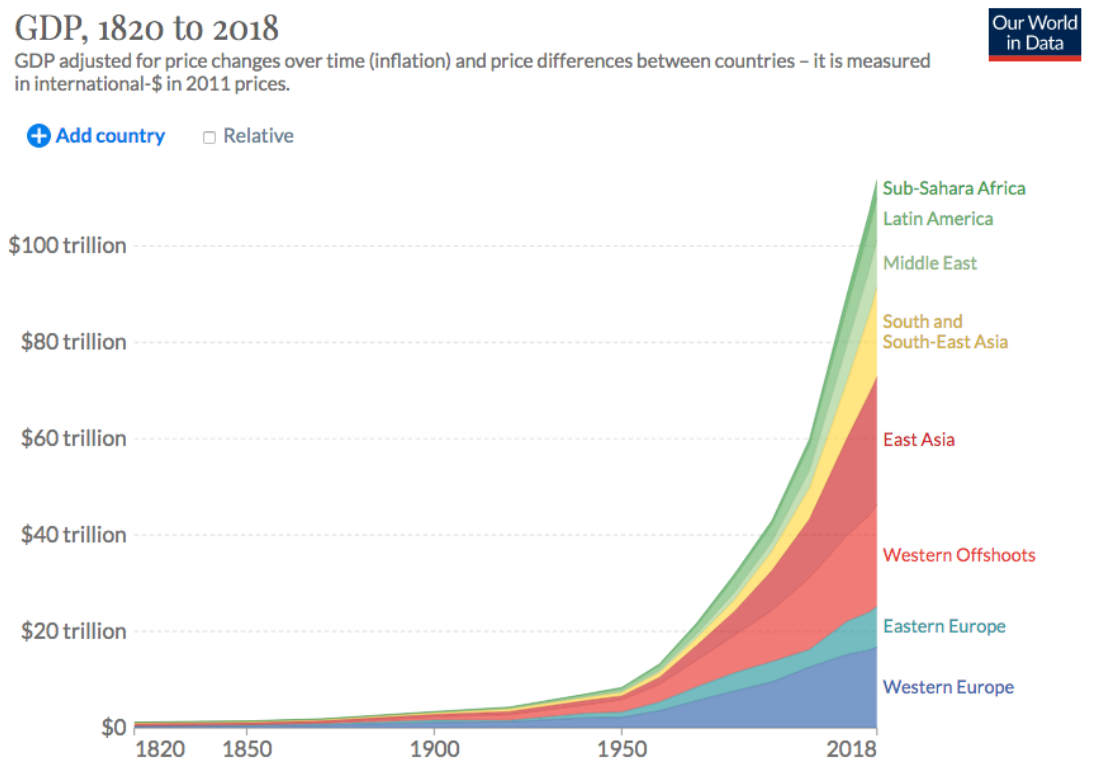
(Diagram 3) Graph in time: history of poverty (highlighted decline, global fetishism not taken into account – think?)
https://ourworldindata.org/history-of-poverty-has-just-begun



Questions:
What was the share of people living in extreme poverty 100 years ago?
How much did the share of people living in extreme poverty fall in the last 100 years? In the last 40 years?
How much did the share of people living in poverty (below 30 dollars per day) fall in the last 40 years?
Which decrease was faster: a decrease in the *share* of people living in poverty in 21th century or a decrease in the share of people living in extreme poverty in 19th century?
Which decrease was faster: a decrease in the *number* of people living in poverty in 21 century or a decrease in the number of people living in extreme poverty in 19 century?
What do you think: will the number of people living in poverty decrease or increase in the 21st century? In extreme poverty? (try to find data that will help in answering this question).

Co-funded by the Erasmus+ Programme of the European Union

(Diagram 4) Graph in time: Economic Growth:
https://ourworldindata.org/grapher/gdp-world-regions-stacked-area?country=Sub-Sahara+Africa~Latin+America~Middle+East~South+and+South-East+Asia~East+Asia~Western+Offshoots~Eastern+Europe~Western+Europe

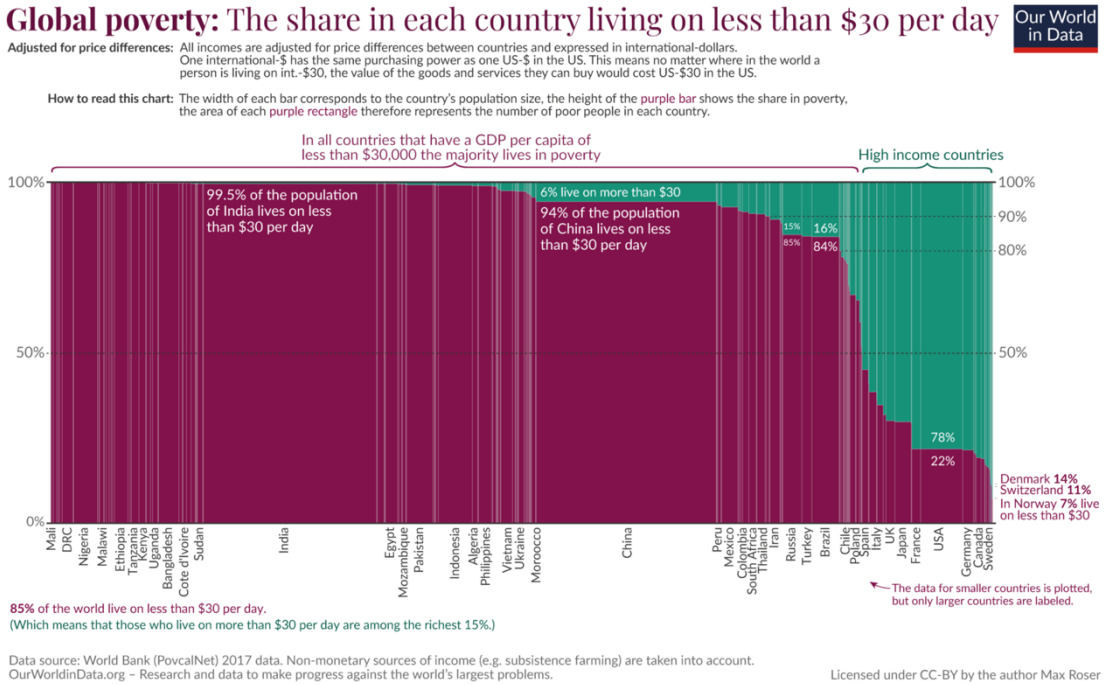


GDP (gross domestic product) is a monetary measure of the market value of all the final goods and services produced in a specific time period by countries.

Questions:
How much did GDP grow since 1950?
How much did GDP grow since 1950 in East Asia?
How much did GDP grow since 1950 in the countries in East Asia?
Does GDP grow quicker in Eastern or Western Europe? In Western Offshoots (USA, Canada, Australia, New Zealand) or in East Asia?
Concentrate on the last 20 years. How does GDP change in Latin America?
What is the share of GDP in Western Europe compared to the GDP of the entire world? Think, what is the relation between GDP per capita (per family) in Western Europe and in East Asia. (Try to find extra data that might help you answer this question).

Co-funded by the Erasmus+ Programme of the European Union

(Diagram 5) Stacked bar graph + bars width: Poverty around the world.
https://ourworldindata.org/history-of-poverty-has-just-begun



Questions:

What is the share of people living on more than 30$ per day in Indonesia? (Further, we will call such people *rich* and others *poor*)

In which country the share of poor people is bigger: in Egypt or in Brazil?

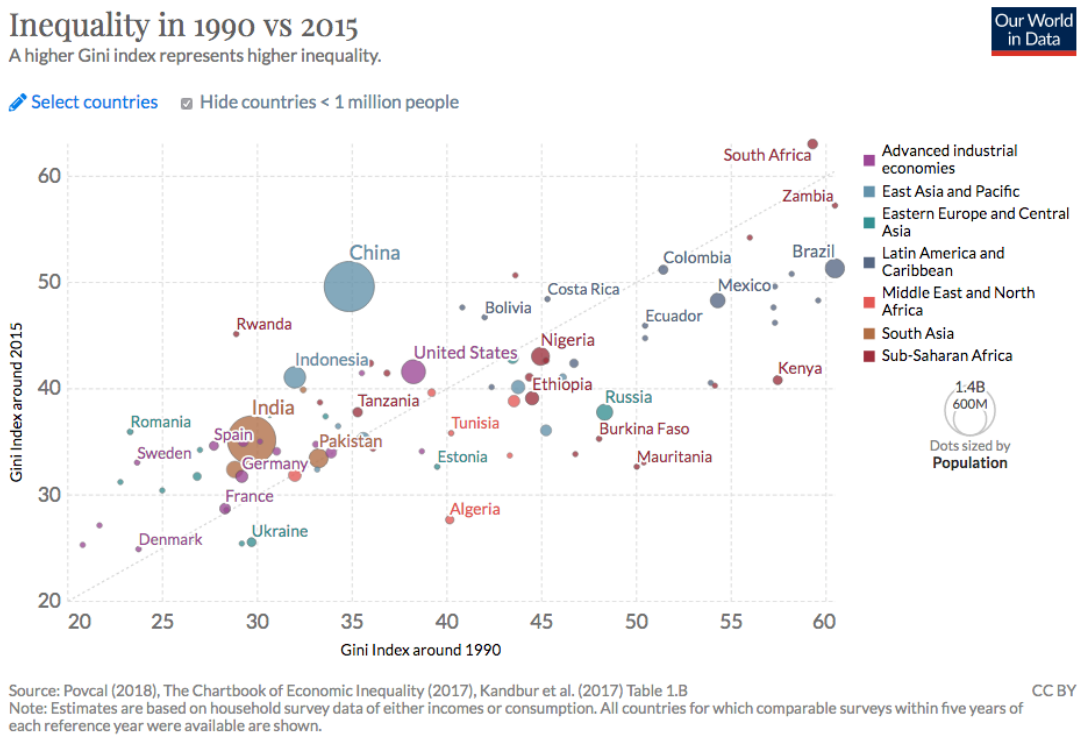
In which country the number of poor people is bigger, in Egypt or in Brazil?

In which country the number of relatively rich people is bigger, in the USA or in Russia? What about the number of poor people in these countries?

Think of this graph together with the previous one. What is the approximate proportion between the money operated by people in the green area of this graph and by purple?

Co-funded by the Erasmus+ Programme of the European Union

(Diagram 6) Scatter plot + size: income equality
https://ourworldindata.org/income-inequality-since-1990



Source: Povcal (2018), The Chartbook of Economic Inequality (2017), Kandbur et al. (2017) Table 1.B CC BY
Note: Estimates are based on household survey data of either incomes or consumption. All countries for which comparable surveys within five years of each reference year were available are shown.

Gini coefficient is a measure of the income variability between people within one country.

Questions:

Which country had higher inequality in 1990, France or Ukraine? In 2015?
What was the Gini coefficient for Brazil in 1990? In 2015? Did it decrease or increase? How can you see if inequality increased or decreased in a country?
In which countries has inequality grown the most? Has diminished?
Where do the poorest people live? Where do the richest people live? Can you see it on this graph?
What can you say about the dynamics of inequality in European countries and in the USA in general? In other regions?
What can you say about the dynamics of inequality in general in the world based on this graph?

Co-funded by the Erasmus+ Programme of the European Union