Experiment Design and Execution 2025W: Exercise 2[*]

1 Introduction

This is a rather open ended exercise with the aim to get you some practice working through the steps of the *Data Science Process* covered in the lectures:

- Ask interesting questions
- Get the data
- Explore the data
- Model the data
- Communicate and visualise the results

Throughout this text, various deadlines are referred to. All deadlines and what is due by these deadlines are summarised in the final section of this document (if there are any inconsistencies in the deadlines given in different parts of the document, then the deadline given in Section 8 is the correct one).

2 Task

The task is to take one of the questions listed in Section 3 as a starting point. Then work through the steps of the Data Science process (including steps back as required) to answer the questions. Some of the first cycles through the Data Science Process could also lead to a refinement or modification of the questions. You may use whichever datasets are required to answer the questions (some potentially useful datasets are listed in Section 4). During the data science process steps, you may have to do some of the following:

- Understand what is in the data – are the data measurements or estimates? How accurate are these measurements or estimates? Are there biases in the data (e.g. in the data gathering process)? If you use estimates to make new estimates, how accurate are the new estimates? Note that this list of questions is not exhaustive.
- Clean the data
- Check for missing data points – decide what to do about them
- Check for outliers – decide what to do about them
- Check for inconsistencies – decide what to do about them
- Calculate descriptive statistics
- Transform the data (e.g. changing units of measurements)
- Check if the necessary data is there to answer the questions. If not, then you could:
    - Combine columns in some way to generate the necessary data

[*] Also applicable to students doing the 3 ECTS DOPP version of the course.

  - Find the necessary data in another dataset
  - Change the questions asked (in this case you have the freedom to do this, but this may not be the case if someone else is asking the questions)
  - ...

- Visualise the data
- Calculate correlations
- Check predictions
- ...

The results should be communicated in both a presentation and a report. Be aware of and document potential biases in the data and analysis. Make sure that the answers to the questions are clear and well supported by the data.

As examples of basic analyses, take a look at:

- `https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic`
- `https://www.kaggle.com/mrisdal/happiness-and-open-data`
- `https://www.kaggle.com/somesnm/new-york-parties-eda`.

The Data Stories on `https://data.europa.eu/en/publications/datastories` are useful.

Also take a look at: `https://ourworldindata.org/`

3 Questions

Each group selects one of the following questions as a starting point for your investigations, or come up with your own questions (in the latter case, check with the exercise coordinator for approval). Note that these questions are generally very broad, so a first step could be to reformulate your selected question in a way that is answerable given the time and people available — this will be discussed at the Feedback Meeting. Each question may be worked on by a maximum of two groups — there is a link on TUWEL allowing each group to select a question.

1. How is e-commerce being adopted in Europe over time? Are there certain areas of the economy that have moved to a larger extent to e-commerce? What effect does e-commerce have on traditional "brick-and-mortar" shops? How are digital payments in countries of Europe being adopted over time, compared to cash?
2. How has the severity of floods in Europe evolved over time? Is there a regional effect? Do flood defences have an effect on flooding? How well can floods and their severity be predicted?
3. How has the severity of forest fires in Europe evolved over time? Is there a regional effect? Do measures for preventing forest fires have an effect? How well can forest fires and their severity be predicted?
4. How has the number of commuters (a person who travels some distance to work on a regular basis) in Europe developed over time? How have the modes of transport changed over time? What impact did the COVID lockdown have on commuting? How do major transport infrastructure projects have an effect on commuting?
5. How has the amount of recycling of waste developed in Europe over time? How does recycling compare across countries? How does it compare for specific types of waste? Are there characteristics of countries that could lead to increased recycling? How well can the development of the amount of recycling be predicted?

6. What structural characteristics (natural and artificial) differentiate cities in Europe? Are there useful clusters of city types based on these characteristics? Are the characteristics related to the climate of the city? How well can the climate resilience of cities be predicted?

7. How has the cycling infrastructure developed in Europe over the previous decades? How has transport by bicycle developed in Europe over the previous decades? Are there correlations between the availability of cycling infrastructure and use of bicycles in European cities or countries? What differences are there between rural and urban cycling infrastructure? How do European countries compare?

8. How do rail travel times compare to air travel times between cities in Europe? Are there routes on which high-speed rail leads to shorter journey times than air travel? How can estimates of travel time to and from airports be included? Which is the most well-connected city in Europe in terms of minimising travel times to other cities? Visualisation of isochrones would be useful in answering these questions.

9. How do rail travel times compare to road travel times between cities in Europe? On how many routes does travel by high-speed rail lead to shorter journey times than travel by car? Which is the most well-connected city in Europe in terms of minimising travel times to other cities? Visualisation of isochrones would be useful in answering these questions.

10. How does tourism compare across the European countries and how has it developed over time? What effect did the COVID pandemic have on tourism in Europe? What can be found out about the ecological impact of tourism? How well can tourism at a regional or city level be quantified and compared?

11. How do citizens of various European countries spend their leisure time? Where do European citizens go on vacation? How have these activities changed over time? How much of their income do European citizens spend on leisure activities?

12. What is the dependence in Europe on fossil fuels (oil, coal, and natural gas)? How has this developed over time? What are the main uses of fossil fuels in Europe? What effect does the change in price of fossil fuels have on their use? What are the main sources of fossil fuels used in Europe and how have these changed over time? How well are countries in Europe moving toward the goal of reducing fossil fuel use?

13. How have political parties represented in parliament evolved over time in countries in Europe? Are there clusters of countries showing similar trends? How well can similar transitions between parliament configurations in European countries be identified? How well can future trends be predicted?

14. How has fishing developed in Europe and its oceans/seas over time? Do nature conservation laws have an effect on the amount of fishing done? Are there changes in the amount of various species of fish caught over time and are causes for these changes detectable? Are fish imports related to the amount of local fishing in coastal countries?

15. How has agriculture developed in Europe over time? Are there differences in the types of agriculture practised? How has the amount of land used for agriculture changed? How are green house gas emissions form agriculture developing over time and what are further trends for these emissions? Are climate change effects on agriculture detectable?

16. What differences are there in the food consumption across European countries? Are there differences in food standards between European countries? What are the imports and exports of food between European countries? Which food and how much is imported from outside Europe? Which food and how much does Europe export? How much food waste is produced and what happens to it?

17. How is access to the internet developing across Europe? What are the differences between countries and between rural and urban areas? How are internet skills developing in Europe? Do internet skills vary by age? Is the teaching of internet/digital skills at schools increasing? Are there correlations between internet access and country characteristics?

18. Choose a sport that has extensive amounts of data published. Formulate and answer questions on this sport.

4 Datasets

The following datasets could be useful for your analysis (this list is far from complete, so you have to do some searching, too):

- The Official Portal for European Data – `https://data.europa.eu/en`
- United Nations Word Population Prospects – `https://population.un.org/wpp/`
- unicef datasets – `https://data.unicef.org/resources/resource-type/datasets/`
- UN Statistics – `https://unstats.un.org/UNSDWebsite/`
- Gridded Population of the World – `http://sedac.ciesin.columbia.edu/data/collection/gpw-v3`
- Open Government Data Austria – `https://www.data.gv.at/en/`
- Eurostat – `https://ec.europa.eu/eurostat`
- OECD Stats – `https://stats.oecd.org`
- World Bank World Development Indicators – `https://data.worldbank.org`
- International Monetary Fund Data – `https://www.imf.org/en/Data`
- World Health Organisation Statistics – `http://www.who.int/healthinfo/statistics/en/`
- Institute for Health Metrics and Evaluation (IHME) – `http://www.healthdata.org`
- Transparency International Corruption Perception Index – `https://www.transparency.org/research/cpi/overview`
- UN Office on Drugs and Crime Data – `https://dataunodc.un.org/`
- World Values Survey – `http://www.worldvaluessurvey.org/wvs.jsp`
- Taxi trips in New York – `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`
- NASA EarthData – `https://www.earthdata.nasa.gov/`
- DataHub collections – `https://datahub.io/collections`
- Awesome Public Datasets – `https://github.com/awesomedata/awesome-public-datasets`
- Inside Airbnb – `http://insideairbnb.com/`
- Copernicus satellite data – `https://www.copernicus.eu/en/access-data`

Google has a Dataset Search tool: `https://datasetsearch.research.google.com/`

5 Groups

The work should be done in **groups of four**. You will be randomly assigned to a group soon after Exercise 2 is handed out.

6 Evaluation

The final mark for this exercise will be based on a hand-in uploaded to TUWEL, as well as a presentation. The final mark is calculated from the following components:

- 20% for the presentation
- 20% for the management summary document (PDF)
- 60% for the solution presented in the Jupyter notebook

A minimum mark of 40% for each of these components is necessary to pass the exercise.

The *hand-in* should consist of a *zip file* containing the following files:

- All of the results of each group should be documented in a single **Jupyter notebook** (i.e. one notebook submitted per group). Make sure that all cells have been calculated in the submitted version. Document and explain all important steps in the submitted Jupyter notebook, including: Which dataset(s) did you choose? Why? How did you clean/transform the data? Why? How did you solve the problem of missing values? Why? What questions did you ask of the data? Why were these good questions? What were the answers to these questions? How did you obtain them? Do the answers make sense? Were there any difficulties in analysing the data? What were the key insights obtained? What are potential biases in the data and analysis? Which Data Science tools and techniques were learned during this exercise? Be sure to document in the Jupyter notebook how was the work divided up between the members of the group.
- This notebook should be accompanied by a **2-page PDF document** that presents a summary of the main insights into the data obtained — this is a management summary, so should be written in a way that is easy to understand by managers. It should also justify why the insights obtained make sense — include diagrams. Do not try and summarise everything that you did in the two pages, focus on the insights. Only the first two pages of the submitted document will be read — do not add a cover page.
- Data needed by the Jupyter notebook should either be accessed directly at its source in the code, or included in the zip file (in the sub-directories expected in the Python code). If some of the data is too large to include in the zip file and cannot be accessed directly within the code, then include a file named `install_data.txt` that includes full instructions on where to download the data and in which sub-directories to install it so that the Python code in the Jupyter notebook can execute.

There are various online tools for collaborating on Jupyter notebooks. One free possibility is Kaggle Notebooks: `https://www.kaggle.com/docs/notebooks`

Note that 47 hours per person is foreseen for this exercise, which is around one third of the time foreseen for the course (150 hours).[1] This means that everyone should work for just over a standard working week on this exercise, so four weeks effort for a group of four.[2] The evaluation will be based on the expectation of a manager assigning such a task to a group of four junior data scientists for the given time period. Note that this expectation is not met by submitting an overly long Jupyter notebook — you need to demonstrate that:

[1] For students doing the 3 ECTS DOPP version of the course, this is 22 hours.

[2] For students doing the 3 ECTS DOPP version of the course, this is just over half a week per person, so 2 weeks of effort for a group of four.

- You have approached the analysis in a logical and structured way.
- You have learned some new data science tools and techniques.
- You have gained new insights into the data.

Overly long notebooks with little substance will be penalised.

Use any additional information that you wish — document which information you use in the Jupyter notebook. If you use Large Language Models (LLMs) then add a section in the Jupyter Notebook documenting exactly what LLMs were used for and how they were used. Releasing your Notebook as a public Kaggle Notebook will be well received.

7 Feedback Meeting, Submission, and Final Presentation

Feedback Meeting: The Feedback Meetings will take place on the 1st and 2nd of December. Each group should reserve a 15 minute slot in TUWEL. The aim of this Feedback Meeting is to present and discuss your plan for the exercise and get feedback. You should outline the plan, including:

1. the questions that you plan to answer,
2. the datasets that you plan to use,
3. how you plan to answer the questions,
4. how the work will be divided up between the group members.

This should be a maximum of 1 page PDF. All key information should be on this page in an easy-to-follow way. No presentation slides are permitted. Be sure to prepare questions for the course coordinators — the Feedback Meeting is the perfect time to get answers. The deadline for the PDF upload and the timeslot reservation in TUWEL is the 1st of December at 09:00.

Submission: The deadline for uploading the zip file to TUWEL is the 8th of January 2026 at 23:55 CET. Remember to reserve a session for the final presentation by the 15th of December 2025. The presentation slides for the final presentation should be uploaded on TUWEL by the 26th of January 2026 at 9:00.

Note that for Exercise 3, your peers will have the task of testing your code and writing a review on your submission. Make sure that everything is well-documented and reproducible.

Final Presentation: On the 26th and 27th of January 2026, each group will present the main results of their work in a 15 minute presentation. The format is 10 minutes presentation and 5 minutes of questions — we will be very strict with the timing, and stop the presentation at the 10 minute mark. The presentation should be aimed at data science colleagues, so highlight which questions you answered, which techniques you used, which data you used, and the insights obtained. Use slides for the presentation. Make it clear in the presentation which member of each group did which part of the work. Each presentation will be followed by short presentations of four reviews on the project (Exercise 3). Each presentation session has space for five groups to present and get the reviews on their projects – all members of all groups registered for a presentation session are required to attend the full duration of the session.[3]

Each group should reserve presentation session in TUWEL by the 15th of December.[4]

[3] For students doing the 3 ECTS DOPP version of the course, there will be no reviews, but there will be dedicated presentation sessions for the groups of DOPP students.

[4] TUWEL will ensure that students in DOPP groups and ExDEx groups can only register in DOPP sessions and ExDEx sessions, respectively.

8 List of Deadlines and Meetings

Here is a list of the deadlines and what should be done by each deadline (all TUWEL links are under the *Exercise 2* heading on the ExDEx course page in TUWEL):

24.11.2025, 23:55 — Select the question from Section 3 of this document that your group will work on in the poll in TUWEL (possible from 18.11 at 13:00)

1.12.2025, 09:00 — Upload the 1 page work plan to TUWEL **AND** book a timeslot for the Feedback Meeting in TUWEL (possible from 24.11 at 13:00)

1.12.2025 & 2.12.2025 — Feedback Meetings — in presence – there will be two tracks in parallel — note the location given with the timeslot that you reserve in TUWEL

15.12.2025, 23:55 — Deadline for reserving a presentation session for the final presentation in TUWEL (possible from 3.12 at 8:00)

8.1.2026, 23:55 — Deadline for uploading the final hand-in (zip file) to TUWEL (possible from 3.12 at 8:00)

26.1.2026, 9:00 — Deadline for uploading the presentation slides for the final presentations

26.1.2026 & 27.1.2026 — Presentations from all groups — in presence – there will be two tracks in parallel, so note the location given with the presentation session that you reserve in TUWEL

Note that there will be no further possibilities for the Feedback Meeting or final presentation beyond the timeslots on the dates given above. Also note that in order to pass Exercise 2, attendance is mandatory at one Feedback Meeting and one complete presentation session.

Allan Hanbury's office hours are on Thursdays, 13:00-14:00 (see changes on this TISS page: `https://tiss.tuwien.ac.at/person/48222`)