Data-Oriented Programming Paradigms 2024W: 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 lecture:

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

Throughout the 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?
- 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
    - 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 may select 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 Review Meeting. Each question may be worked on by a maximum of two groups — there is a poll on TUWEL allowing each group to select a question.

1. How has the severity of floods in Europe developed over time? Is there a regional effect? Do flood defences have an effect on flooding? How well can floods and their severity be predicted?

2. How has the severity of forest fires in Europe developed 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?

3. 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?

4. 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? Can the development of the amount of recycling be predicted?

5. 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?

6. What impact do hotels or Airbnb apartments have on cities in Europe? Is the position or the popularity of hotels/Airbnb apartments related to Points of Interest, public transport or other features of the city? How well can good locations for a new hotel/Airbnb be predicted?

7. How can the innovation level of a country in Europe be measured? What characteristics of a country predict the innovation level? What characteristics of a country predict an increase or decrease in the innovation level?

8. With which means of transport do people move around in cities (modal split) in Europe? How has this changed over time? How has this changed in various countries? Are there specific characteristics of countries/cities that can be shown to correlate with modal split and its evolution?

9. How does the rail infrastructure and road infrastructure of European countries compare? To what extent are goods and people transported by road and rail? How has this developed over time? Do changes in the road or rail infrastructure have an effect on the goods or people being transported?

10. 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?

11. 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.

12. How is the level of carbon emission in Europe evolving? Which countries are emitting and absorbing the most carbon? Are there differences when considering all carbon ever emitted by a country? Are there characteristics of countries that correlate with the changes in the carbon emissions?

13. How is the population throughout Europe changing over time? How is it changing based on characteristics of the people (e.g. age, education level)? How well can movements of people between countries/regions be followed? How well can common migration routes in Europe be identified?

14. How does tourism compare across the European countries and how has it developed over time? Where do European citizens go on vacation? 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?

15. How do university rankings change over time? Which characteristics of universities contribute most to good rankings, or to large changes in the ranking position? How do these characteristics correlate with characteristics of cities or countries in which the university is located? Are there predictors for increases or decreases in the rankings?

16. How do levels of education attained compare across countries in Europe? How have these levels changed over time? How has the internationalisation of education (e.g. Erasmus exchanges) developed over time? How strong are the correlations between education levels obtained and economic development?

17. How well can the level of corruption of a country in Europe be quantified? What differences are there in actual corruption and perceived corruption? Are there different forms of corruption prevalent in different countries in Europe? What characteristics of a country predict the level of corruption? What characteristics of a country predict an increase or decrease in the level of corruption?

18. How is the adoption of renewable energy in Europe evolving? How is the adoption of different types of renewable energy evolving? Are there characteristics of countries that correlate with their level of adoption of renewable energy?

19. 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?

20. 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?

21. Are reviews for some categories of product on Amazon overall more positive than for other categories? Are reviews more subjective for some classes of products than for others? Which aspects of different classes of products are the most important in the reviews? Can one predict the star rating from the review text?

22. How well can the level of healthcare in countries and cities in Europe be quantified? What differences are there across countries in Europe? What is the correlation with life expectancy?

23. How do vaccination rates against various diseases vary across Europe? How has this rate changed over time? Is the occurrence of certain diseases higher in areas in which vaccination is lower? Are there country characteristics that predict vaccination levels, or trends in vaccination levels?

24. How has the occurrence of various causes of death developed across Europe? How do causes of death vary by region in Europe? Are there clusters of regions with similar distributions? Are there correlations with characteristics of the regions?

25. How does illegal drug use vary across countries in Europe? How do laws on illegal drug use vary across countries in Europe? Does the level of illegal drug use affect the health in a country? Is the level of illegal drug use related to other crimes?

26. 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?

27. 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?

28. 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?

29. 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?

30. Which countries in Europe have the highest/lowest crime rates? Are there typical characteristics of countries with high/low crime rates? Are there countries that have different types of crimes that are dominant? Are there country characteristics that predict crime rates, trends in crime rates, or types of crimes?

31. How have the number of natural disasters in Europe changed over the last 100 years? How have the number of deaths per year from natural disasters in Europe changed over this time? How does this vary by country? How does this vary by type of natural disaster? Are there trends visible that could be due to climate change?

32. How many species are currently listed as endangered in Europe? How has this changed over time? Which geographical, natural or country characteristics predict higher numbers of endangered species? Can these characteristics also predict trends in the number of endangered species?

33. How do city quality of life rankings in Europe change over time (e.g. Mercer, Global Liveability Ranking)? How do these rankings correlate with each other? How do they correlate with statistics about the countries in which the cities are found? How do they correlate with the cost of living? What are the determining characteristics for liveability of a city?

34. What makes a good board game? Can the success of a board game be predicted based on its characteristics? What else affects the success of a board game?

35. What makes a good video game? Can the success of a video game be predicted based on its characteristics? What else affects the success of a video game?

36. 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 University Rankings – `https://www.kaggle.com/mylesoneill/world-university-rankings`
- 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/`
- Amazon Review Data – `https://nijianmo.github.io/amazon/index.html`
- 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**. Please add yourself to a group in TUWEL by the 19th of November. Groups of fewer than 4 people are not permitted.

6 Evaluation

The final mark for this exercise will be based on a hand-in uploaded to TUWEL and 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 35% 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? 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 36 hours per person is foreseen for this exercise, which is around half the time foreseen for the course (75 hours). This means that everyone should work for around a standard working week on this exercise, so four weeks effort for a group of four. The evaluation will be based on the expectation of a manager assigning such a task to a group of four junior data scientists for a week. Note that this expectation is not met by submitting an overly long Jupyter notebook — you need to demonstrate that:

- 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 Review Meeting, Submission, and Final Presentation

Review Meeting: The review meetings will take place on the 3rd and 6th of December. Each group should reserve a 15 minute slot in TUWEL. The aim of this session 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 review meeting is the perfect time to get answers. The deadline for the PDF upload and the timeslot reservation in TUWEL is the 2nd of December at 09:00.

Submission: The deadline for uploading the zip file to TUWEL is the 28th of January at 08:00 CET. By the same deadline, you should upload the presentation slides on TUWEL (remember to reserve a timeslot for the final presentation by the 24th of January).

Final Presentation: On the 28th and 30th of January, 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 group should reserve a 15 minute slot in TUWEL by the 24th of January. Everybody is free to attend the presentations. Please attend for about 1 hour before and 1 hour after your presentation slot, so that there is an audience to ask questions on the presentations.

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 DOPP course page in TUWEL):

19.11.2024, 23:55 — All group members must be registered for their Exercise 2 group in TUWEL

26.11.2024, 23:55 — Select the question from Section 3 of this document that your group will work on in the poll in TUWEL (possible from 20.11 at 13:00)

2.12.2024, 09:00 — Upload the 1 page work plan to TUWEL **AND** book a timeslot for the review meeting in TUWEL (possible from 27.11 at 9:00)

3.12.2024 & 6.12.2024 — Review meetings — in presence – there will be two sessions in parallel — note the location given with the timeslot that you reserve in TUWEL

7.1.2025 — Voluntary consultation session on Exercise 2, 12:00-14:00 in EI11

24.1.2025, 23:55 — Deadline for reserving a time slot for the final presentation in TUWEL (possible from 7.1 at 8:00)

28.1.2025, 08:00 — Deadline for uploading the final hand-in (zip file and presentation slides) to TUWEL (possible from 7.1 at 8:00)

28.1.2025 & 30.1.2025 — Presentations from all groups — in presence – there will be two sessions in parallel, so note the location given with the timeslot that you reserve in TUWEL

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`)