Data-Oriented Programming Paradigms: Exercise 3

December 4, 2019

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.

2 Task

The task is to take one of the questions listed in Section 3 as a starting point. Then work through all 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 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 exploration and modelling 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? 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. 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`.

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

3 Questions

You 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). 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. Each question may be worked on by a maximum of two groups — there is a poll in Doodle allowing each group to select a question (link from TUWEL).

1. Does the majority of the word's population live in low-income countries? Which characteristics are predictive for the income level of a country? Which characteristics are predictive for change in the income level of a country?
2. In all countries across the world today, how many girls finish primary school? How does this differ for low-, medium- and high-income countries? Is it possible to identify objectively reasons for them not finishing school? How well can predictions be made for the figures 10 years in the future?
3. What percentage of the world population lives in extreme poverty? Which characteristics are predictive for countries with large populations living in extreme poverty? Which characteristics are predictive for populations emerging from extreme poverty?
4. Is terrorism highest in low-, medium-, or high-income countries? Which characteristics of countries are predictive for high/low levels of terrorism? Which characteristics are predictive for how the level of terrorism in a country develops?
5. What is the average human life expectancy in the world? How does this differ between low-, medium, and high-income countries? What characteristics of countries are predictors for differences in life expectancies? What characteristics are predictors for changes in life expectancy in a country over time?
6. How did the number of deaths per year from natural disasters change over the last hundred years? 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?

7. How many of the world's 1-year-old children today have been vaccinated against some disease? How many against more diseases? How has the rate of vaccination for different diseases changed over time? Are there country characteristics that predict vaccination levels, or trends in vaccination levels?

8. How many species are currently listed as endangered? 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?

9. How many people in the world have some access to electricity? How has this changed with time? Are there correlations between electricity supply and changes in levels of living or education?

10. 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?

11. How can the innovation level of a country 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?

12. How can the level of corruption of a country be measured? 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?

13. What is the most accurate overview of immigration between countries that can be obtained? Are there typical characteristics of immigration origin and destination countries? Are there typical characteristics of large flows of immigration? Can sources of immigration be predicted? Can changes in immigration patterns be predicted?

14. What is the most accurate overview of flows of refugees between countries that can be obtained? Are there typical characteristics of refugee origin and destination countries? Are there typical characteristics of large flows of refugees? Can countries that will produce large numbers of refugees be predicted? Can refugee flows be predicted?

15. How can the crime rate of a country be measured? Which countries 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 also country characteristics that predict trends in crime rates or types of crimes?

16. How do city quality of life rankings 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 livability of a city?

17. With which means of transport do people move around in cities (modal split)? 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?

18. Is an optimal level of taxation predictable, given the characteristics of multiple countries? This one is pretty difficult, as it will be necessary to define what should be regarded as optimal.

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):

- United Nations Word Population Prospects – `https://population.un.org/wpp/`
- Gridded Population of the World – `http://sedac.ciesin.columbia.edu/data/collection/gpw-v3`
- United Nations Population Division – `http://www.un.org/en/development/desa/population/`
- EU Open Data Portal – `http://data.europa.eu/euodp/en/home`
- Eurostat – `https://ec.europa.eu/eurostat`
- UN Refugee Agency (UNHCR) Data – `http://www.unhcr.org/data.html`
- OECD Stats – `https://stats.oecd.org`
- World Bank World Development Indicators – `https://data.worldbank.org`
- 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`
- Global Terrorism Database – `https://www.kaggle.com/START-UMD/gtd` and `https://www.start.umd.edu/gtd`
- World University Rankings – `https://www.kaggle.com/mylesoneill/world-university-rankings`
- World Values Survey – `http://www.worldvaluessurvey.org/wvs.jsp`
- Taxi and Uber trips in New York – `https://github.com/toddwschneider/nyc-taxi-data`

5 Groups

The work should be done in **groups of four**. Please add yourself to a group in TUWEL before the 11th of December.

6 Evaluation

The final mark will be based on a hand-in uploaded to TUWEL and a presentation. 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). Document all important steps — 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? 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, not data scientists. It should also justify why the insights obtained make sense — include diagrams.
- All data needed by the Jupyter notebook should be 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, include a file named `install_data.txt` the 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 is a good tutorial on Jupyter notebooks here: `https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook`. Notice also the tips on using Jupyter notebooks in teams toward the end of this tutorial.

Note that 42 hours per person is foreseen for this exercise, which is more than half the time foreseen for the course (75 hours). This means that everyone should work for slightly more than a standard (40 hour) 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 just be 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. Releasing your Notebook as a public Kaggle Kernel will be well received. See here: `https://www.kaggle.com/mrisdal/new-to-kernels-start-here-1`

7 Review Meeting, Submission, and Final Presentation

Review Meeting: The review meetings will take place on the 17th and 18th of December. Each group should reserve a 15 minute slot in TUWEL. The aim of this session is to present your plan for the exercise and get feedback. You should outline the plan, including the questions that you plan to answer, the datasets that you plan to use, how you plan to answer the questions, and 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 – the number of words written is not a criterion. The deadline for the PDF upload and the timeslot reservation in TUWEL is the 16th of December at 23:55.

Submission: The deadline for uploading the zip file to TUWEL is the 27th of January 2020 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 20th of January).

Final Presentation: On the 27th of January 2020 between 9:00 and latest 18:00 (depending on the number of groups), each group will present the main results of their work in a 15 minute presentation in the Hauptgebäude Hörsaal 6. 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. 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 for the presentations.

8 List of Deadlines

Here is a list of the deadlines and what should be done by each deadline (all TUWEL links are under the *Exercise 3* heading):

11.12.2019, 23:55 — All group members must be registered for their Exercise 3 group in TUWEL

11.12.2019, 23:55 — Select the question from Section 3 that your group will work on in the Doodle Poll (link from TUWEL)

16.12.2019, 23:55 — Upload the 1 page work plan to TUWEL **AND** book a timeslot for the review meeting in TUWEL (possible from 10.12 at 8:00)

17.12.2019, 14:00-18:00 & 18.12.2018, 10:00 – 14:00 — Review meetings

20.1.2020, 23:55 — Deadline for reserving a timeslot for the final presentation in TUWEL (possible from 7.1.2020 at 8:00)

27.1.2020, 08:00 — Deadline for uploading the final hand-in (zip file and presentation slides) to TUWEL (possible from 7.1.2020 at 8:00)

27.1.2020, 9:00–18:00 — Presentations from all groups — All presentations will be given from a single computer (that supports Powerpoint and PDF), using the slides uploaded on 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`)