Suggested Exercises and Guided Projects

DATA SCIENCE ESSENTIALS

Between Sessions

Session 1 to Session 2

- complete the exercises of session 1
- download the datasets from the website
- read Programming Primer (sections 1 – 4)
- install R / RStudio (Posit)
- install the following R packages: dplyr, xts, knitr, tidyverse, ggplot2, pastecs, Hmisc, e1071, psych, quantmod, ggm, kerndwd, MASS, DMwR, ROCR, car, forcats, corrplot

Session 2 to Session 3

- complete the exercises of session 2

Session 3 to Session 4

- complete the exercises of session 3

After Session 4

- complete the exercises of session 4
- attempt the guided projects

Guided Project I

Select a data project of interest (personally or professionally) and provide a planning draft for it, touching on the topics discussed in this course. The following questions can help:

1. What are some questions associated with the project?
2. What is the conceptual model of the underlying situation?
3. What kind of dataset(s) exist that could help you answer these questions?
4. Are there data or analytical limitations?
5. Do you need to collect new data to handle such questions?
6. How is the data stored/accessed? What are the infrastructure requirements?
7. What do deliverables look like?
8. How would successes be quantified/qualified?
9. What are your timelines and availability?
10. What skillsets are required to work on this project?
11. Would you work on this alone or as part of a team?
12. How costly would it be to initiate and complete this project?
13. What does the data analysis pipeline look like?
14. What software and analytical methods will be used?

Guided Project II

Write a paper discussing some of the ethical issues surrounding the use of artificial intelligence, data science, and/or machine learning (M.L.) algorithms.

Establish a list of the 3 most important ethical principles that the use of such algorithms should abide by. Explain why you have selected each of these principles.

Describe (at least) 2 real-life instances of the use of A.I./D.S./M.L. in the public sector, the private sector, or in academia, when the ethical principles you have chosen were violated. Discuss how the failure to abide by your selected ethical principles have caused (or could cause) harm to individuals, organizations, countries, etc.

Suggest how the projects discussed above could have been modified so that their use of A.I./D.S./M.L. algorithms would abide by your selected ethical principles.

Guided Project III

This project uses the Gapminder Tools (there is also an offline version).

1. Take some time to explore the tool. In the online version, the default starting point is a bubble chart of 2020 life expectancy vs. income, per country (with bubble size associated with total population). In the offline version, select the "Bubbles" option.
2. Can you identify the available variable categories and some of the variables? [You may need to dig around a bit.]
3. Why do you think that Gapminder has selected Life Expectancy and Income as the default plotting variables?
4. Replace Life Expectancy by Babies per woman. Observe and discuss the changes from the default plot.
5. Formulate a few questions that could be answered with the default data.
6. Formulate a few questions that could be answered using some of the other variables.
7. At what point in the data science workflow do you think that visualizations of this nature could be useful?
8. Do these visualizations provide a sound understanding of the system under investigation (the geopolitical Earth)?

Guided Project III (cont.)

9. What do you think the data sources are for the underlying dataset? [You may need to dig around the internet to answer this question].
10. Are all variables and measurements equally trustworthy? How could you figure this out?
11. Is the underlying dataset structured or unstructured?
12. Provide a potential data model for the dataset.
13. What are the types of the 4 default variables (Life Expectancy, Income, Population, World Regions)?
14. Play around with the charts for a bit. Can you find pairs of variables that are positively correlated? Negatively correlated? Uncorrelated?
15. Among those variables that are correlated, do any seem to you to exhibit a dependent-independent relationship? How could you identify such pairs?
16. Can you provide an eyeball estimate of the mean, the median, and the range of various numerical variables?
17. Can you provide an eyeball estimate of the mode of the categorical variables?
18. Can you identify epochal moments (special temporal points) in the data where a shift occurs, say?
19. Is the tool and its underlying dataset useable? What factors does your answer depend on?

Guided Project III (cont.)

20. Do you think that there could be problems with the reported values? For instance, select Sweden and the United States from the checkbox menu on the right and follow their path from 1799 to 2018/2020. From what point onwards are the values sensible? What do you think is happening at the start of the series?
21. Follow Eritrea for the same duration. Look up the country's independence date from Ethiopia. What do you think the measurements prior to that date represent?
22. Follow Austria for the same duration. Look up the historical timeline of the country's boundaries (Austria-Hungary, Anschluss, modern borders, etc.). What does that imply for the measurements?
23. Follow Finland for the same duration. What happens in 1809? Does that tell you anything about the way data is coded in the dataset?
24. De-select all countries and let the simulation run from 1799 to 2018/2020. Can you identify instances where a large subset of observations behaves in unexpected manners? If so, do you think that this is due to data cleaning/data processing issues?
25. Continue exploring the dataset. You may change which variables are displayed or work with some of the other visualization methods. Overall, do you think that the dataset is sound? Would you use it to run analyses? What are some of its strengths and weaknesses?

Guided Project IV

Select a dataset from the list below (or any other set of interest to you):

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- polls_us_election_2016.csv
- HR_2016_Census_simple.xlsx

For your dataset(s):

1. Create a "data dictionary" to explain the different fields and variables. Can you find a source for these datasets online?
2. Develop a list of questions you would like answered about the datasets.
3. Investigate individual variables (through simple charts, univariate statistics, etc.).
4. Repeat the process with bivariate investigations (though simple charts, joint distributions, variable interactions, etc.).
5. Do you trust the dataset, or not? Support your answer. If you do not trust the dataset, flag potential invalid entries, anomalous observations, missing values, or outliers. How should these entries be treated?
6. Does any of your analysis suggest that some of the variables should be transformed? Do any of the questions you developed in step 2 support such transformations? If so, transform the data appropriately.