Workshop Guidelines

Business Intelligence

MBIN Staff – Georg Krempl, Armel Lefebvre Academic year 2018-2019 INFOMBIN – Business Intelligence

Contents

Business Intelligence: Workshop Guidelines

In this document, you'll find information about the workshop of Business Intelligence (INFOMBIN), Academic Year 2018-2019. The information found in this document is for educational purposes only.

This document might be updated to reflect the latest changes during the course. We will always refer to a version of this document when referring to the BI Workshop guidelines. See Table 1 for the updates.

Table 1 Workshop guidelines for students and staff

Version	Date	Author
2	17/04/2019	Armel
3	29/04/2019	Armel
4	30/04/2019	Armel

Introduction

The capacity of Europe to innovate and lead the battle for significant scientific discoveries and technological applications depends on the ability to understand the global environment. Major challenges are putting some pressure on European (and national) funding programs, quality and impact of publications in academia and, the potential to develop technology in the era of artificial intelligence and Big Data. Recently, the news was not good for the EU[1]. At the European level, it is no secret that the EU is lagging in terms of research and innovation to compete with countries such as the United States or China. But what is exactly going on, and what should be done better? How do we compare to international competitors, whether there are established or emergent ones?

To answer these questions, there is a strong need for a better, coherent, strategy for which the performance has to be evaluated. The evaluation should take advantage of the numerous statistics, indicators and (open) data sets. Also, policymakers and universities should be able to understand what is going on and understand how to get more insights with your dashboards. So, you will have the difficult task to combine a deep understanding of Business Intelligence, Data warehousing, and state-of-the-art dashboards.

Background

The Georg Institute for Intergalactic Research and Innovation Analytics (with business intelligence) better known (or not) under the name GIIRIABI, has an open call for selecting business intelligence solutions providing new insights into the problematic of Research Performance and Diffusion in Europe. The GIIRIABI was mandated by several stakeholders (mainly European universities and policymakers) to compile data to help them identify improvements areas in scientific research quality, funding, impact, and many others.

There is a small caveat though. The GIIRIABI accepted to design and implement several BI solutions to impress its stakeholders without having read the small prints stating that the overview must be

[1] https://ec.europa.eu/growth/content/european-innovation-scoreboard-2018-europe-must-deepen-its-innovation-edge_en

international (i.e., worldwide), as the big players in the research and innovation worldwide are evolving too. Besides, the final deliverable is expected quite early, at the end of June 2019! It looks a bit overwhelming for the GIIRIABI to solve it all by itself…

By registering for the course Business Intelligence 2018-2019 at Utrecht University you were automatically enrolled as BI consultants for the GIIRIABI. So, it is your job now to figure out how to design BI solutions helping universities and policymakers identify improvement areas to boost research, innovation, and diffusion of scientific knowledge in Europe.

Luckily, you quickly notice that there is already a vast amount of work that has been done by others. But maybe some data is still quite disconnected. Some KPIs could be refined or specified for particular issues in a given county. So, it looks like you do not have to generate new data but smartly combine existing (open) data to provide useful insights!

This document guides you through the project by giving pointers to data sets, additional information about the content of the report, a list of deliverables and an agenda. Each week we will provide you with some more details about the final project.

In short, what are we waiting for?

Pre-requisites

For the workshop, you will have to develop a software artifact in R and use a relational database (i.e., Postgres) to store the data. It is assumed that you know how programming languages work and how to code. If you do not now R well, you can understand the tutorials and documentation you can find on the web, as you would do with other similar scripting languages. There is only very basic and limited technical assistance offered during this course. In case you feel that your programming skills are lagging, check with your team how you can contribute more at a strategic level (identification of relevant data sources, KPIs, etc.) which is a crucial role for the success of this project.

We will not judge your project on your level of coding skills (as long as the script is documented and structured). For instance, we will not lower your grade if we believe that some faster or more optimized functions of R could have been used instead of a (slower) loop. It is not a programming course. Nevertheless, we expect that you possess the required skills to work on a data science project at the start of the workshop.

Project Outline

You, together with your team members, will design and implement a Research and Innovation Performance solution which consists of an interactive dashboard and a written report that the GIIRIABI will transfer to relevant policymakers at the institutional or national level. Keep the stakeholders in your mind; they want to quickly grasp the current environment but also decide what action they should take and when.

Team Building

A team consists of four members, a focus and a selection of three to four countries. In your team, we suggest that responsibilities are given to each member according to their interests and skills. Every team member is expected to be able to explain all deliverables. What KPIs are chosen and how, how does the dashboard work, etc. If a team member fails to demonstrate the required knowledge about the project, it will lead to an insufficient grade for team contribution (see Appendix C).

Focus

You need to choose one of the three focuses. A focus will scope the key performance indicators you select. Also, it will allow you to deepen some aspects of Research and Innovation Performance.

For each focus, we selected some suggested online background. But we invite you to review the literature using search engines you know, e.g., Scopus, WebOfKnowledge, Google Scholar, ScienceDirect, to complete the background information in your report. You might get inspired by existing metrics published in the literature, and, why not, improve them.

In case you feel you are getting lost in the vast amount of literature available, feel free to discuss your findings with the course staff.

Education

In the "Education" focus, you will explore and present findings related to the capacity of European (and non-EU) countries to educate their citizens in an evolving knowledge economy.

Suggested background:

- http://www.oecd.org/education/
- https://ec.europa.eu/eurostat/statistics-explained/index.php/Education_and_training_in_the_EU_-_facts_and_figures
- https://www.cbs.nl/nl-nl/maatschappij/onderwijs (in Dutch)

Research Quality

In the "Research quality" focus, you will explore university rankings, publications, citations, impact factors, Open Access, and Open Science. Or at least a few of these items as we do not expect your dashboard to cover them all, of course.

Suggested background:

- http://libraryguides.missouri.edu/impact
- https://www.universityrankings.ch/methodology/on_rankings/ranking_methodology

Research and Innovation

In the "Research and Innovation" focus, you will explore other essential prosperity drivers than the quality of research. This is a perfect focus for those of you who feel comfortable with economic, productivity and technology (ICT) indicators.

Suggested background:

- https://ec.europa.eu/info/research-and-innovation/strategy/support-policy-making/support-national-research-and-innovation-policy-making/srip-report_en
- http://www.sgi-network.org/2018/Policy_Performance/Economic_Policies/Research_and_Innovation

Countries

The list of countries shown in Table 2and Table 3 is selected from existing data made available by the European Commission (EC) see here: https://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards_en and here: https://rio.jrc.ec.europa.eu/EN/COUNTRY-ANALYSIS.

EU

Select two or three EU countries. There is no strict criterion for including countries. Think about your interests and curiosity. If your country of origin differs from the Netherlands, you are more than welcome to add your country of origin. In case the country is not in these lists, but you could identify the data you need, contact the course staff, and we might accept that you integrate an unlisted country in your assignment.

Table 2 EU countries

Austria	Belgium	Bulgaria	Croatia
Cyprus	Czech Republic	Denmark	Estonia
Finland	France	Germany	Greece
Hungary	Ireland	Italy	Latvia
Lithuania	Luxembourg	Malta	Netherlands
Poland	Portugal	Romania	Slovakia
Slovenia	Spain	Sweden	United Kingdom

Non-EU

Select a maximum of two international countries, minimum 1. Note, data from foreign countries might be harder to obtain for some scenarios and KPIs. Make sure to collect sufficient information about data availability and quality before you make a final selection of the focus and countries you want.

Table 3 Non-EU countries

Iceland	Israel	North Macedonia		
Norway	Serbia	Switzerland	Turkey	Ukraine
Brazil	China	India	Mexico	United States

Weekly schedule and assignments

Here, we describe what you can expect each week. This section is a more elaborated version of Appendix D (the agenda). Read it carefully as it contains strict requirements for some assignments. Not satisfying the (hard) requirements for the assignments lead to penalties, or worse, failing to pass the workshop (and hence the course). The hard requirements are not added to bother you but to ensure a smooth collection of the documents and a fast grading process. It is a course with around 100 enrolled students where we need to evaluate the quality of written reports and coding assignments. That requires some logistics which also rely on you. Reading and applying the requirements makes it even more comfortable for us (the course staff) to guarantee that you obtain feedback and grades on time.

W17: Register Teams

Deadline: 28-04-19 at 23:59

After an introduction to the project on Thursday 25/04/2019, you will have the opportunity to team up with other students.

You will register your team on Blackboard.

W18: Create Strategy

Deadline: 05-05-2019 at 23:59

In week 18 you will be looking for data based on indicators you find useful. At the same time, you work on getting a clearer view of what kind of sources you can use for your project. Don't look for a very high granularity, as we are using available open data; most have aggregated values already.

Many of the sources used (OECD, European Commission) have embedded exploratory tools. Make use of them to get a quick glimpse of the data — download files (e.g., xlsx and csv) to check the completeness and granularity of the datasets.

You can also refine your list of indicators based on the data you find; this is not a problem. What is important is that you get a coherent list of indicators with data you deem appropriate.

At the end of week 18, you will submit a PDF (1-3) pages with the following information:

- Team Number
- Team Members
- Contact Person
- Focus
- Countries
- (Potential) Data sources
- (Preliminary) Key performance indicators
- Relevance statement: a short section where you explain why your dashboard will make a difference and why the KPIs are relevant for the selected countries and focus. You might add what kind of findings you expect (e.g., trends, patterns). The relevance statement should not exceed 500 words.
- A roadmap is describing the main milestones, tasks, and responsibilities in your team. Carefully consult the agenda.

Next, a short recap of the distinction between a measure, a metric an indicator, and a key indicator.

- Measure: a value
- Metric: a value and a unit (context)
- (Performance) Indicator: a metric that reflects the performance
- Key (performance) indicator: Selected set of indicators

But why some indicators are selected and others not? You will need critical success factors for that. Key performance indicators are indicators that operationalize critical success factors. If it does sound a bit complex, an example will help! Let's take a look at the Council Decision 2013/743/EU (ANNEX II) which states a bunch of CSFs and indicators for assessing H2020, a funding program from the European Commission (EC).

For the CSF (or what the EC calls priority) "Excellent science":

Key performance indicators are:

- Publications in peer-reviewed high impact journals
- Number of researchers who have access to research infrastructures through Union support

For CSF "Industrial leadership":

Key performance indicators are:

- Number of joint public-private publications
- Growth and job creation in participating SMEs

If we take "Publications in peer-reviewed high impact journals," what we see is that it is more than a number (value). There is some context attached to it: peer-reviewed, high impact, journals. It is more than a metric because it is used to assess the performance of the H2020 program. Here it is even a Key performance indicator as it has been selected to evaluate how "excellent" science is. So it's an indicator which is crucial to evaluate the strategy of H2020.

On Thursday 02/05/2019 there is a tutorial about Data warehousing in R.

W19: Implement Dashboard Prototype

Deadline: 15-05-2019 at 23:59

It's time making sure that the technical stuff works. You will upload the implementation of a subset of KPIs and push the code on Gitlab. Please consult Appendix B to make sure you follow the required file hierarchy.

Consult the gallery http://shiny.rstudio.com/gallery/ to get some technical info about how R shiny works (and try it).

In short, these are the basic functionalities each of your dashboards should have:

- Slice
- Dice
- Drill-up/Drill-down
- Pivot
- (Key) Indicator selection
- Charts
- Quick highlights
- Targets

As the tutorial on dashboards will be on Thursday 09/05/2019, we leave a little bit more time to complete this part of the assignment. Please send your code on Gitlab before the project kick-off.

W20: Kick-off Presentation

!! **Deadline: 15-05-2019 at 23:59 !!**

Time to show your plan. You will have a spot for a 5 min presentation about:

1) Your focus and countries
2) Your KPIs
3) A mockup dashboard, which does not have to be implemented in R. The mockup can be a drawing in PowerPoint or even pen and paper (if you are skilled at it). It should show us how the dashboard is organized and what kind of visualizations you are planning to use.

W21: Business as usual

No deadline, work on the assignment

W22: Update Gitlab

Deadline: 02-06-2019 at 23:59

Submit an update of your Dashboard to Gitlab. It should already be quite complete in terms of data sets used and ETL. You still have time to perfect the Dashboard itself (in Shiny).

W23: Peergrade

Deadline: 06-06-2019 at 23:59

Send a draft report to Peergrade. The strategy and ETL must be properly described. Your peers will understand that the dashboard is still work in progress but provide maximum information about what already works so you can benefit from their feedback.

Please make sure you send an anonymized version of your report. Remove all identifying information from your report. Also, make sure the PDF itself does not contain identifying meta-data. See here for more instructions about how to anonymize PDFs:

https://www.siam.org/Publications/Journals/Related/Journal-Policies/Detail/protecting-referee-personal-information

W24: Review reports

Deadline: 13-06-2019 at 12:00

You will have to review two reports. You will receive a request on Tuesday evening, and we expect to receive your review on June 13 at 12:00.

W25: Final push

Deadline: 23-06-2019 at 23:59

Final update of the dashboard. You will present this week, and we want to make sure that everything is working well.

The final presentation is in plenary. You will get 8 min. to present your dashboard. Show the functionalities and the main findings. We will also judge if it looks professional and user-friendly.

The code that will be reviewed for grading your project is this version. In case you modify your code right before the presentation, it will have no consequence on the grading of the Dashboard deliverable.

W26: Submit the final report

Deadline: 30-06-2019 at 23:59

Send your final report as a PDF.

Configuration

For the technical part you will need:

- PostgreSQL version 10.7 + PgAdmin 4 + PsqlODBC drivers

    - Install from here: https://www.enterprisedb.com/downloads/postgres-postgresql-downloads , please install stackbuilder (when prompted during the install). Otherwise, you need to install ODBC drivers separately.
- R >= 3.4.3 or 3.5+ branch with at least shiny, tidyverse and RPostgres packages
- RStudio > 1.1.4X
- Git client

To host a PostgreSQL database for your team, you might be interested in creating an Amazon free tier RDS server:

➔ https://aws.amazon.com/free/?awsf.Free%20Tier%20Categories=productcategories%23database

Demo Project

There is a demo project available; you can clone it from Gitlab to start with the right project structure.

➔ https://git.science.uu.nl/bi2019/demo

Final remarks

There is some room left in this project for exploring the vast environment of Research and Innovation (R&I). At the same time, you will learn to create a Business Intelligence solution independently of the R&I domain. If you are not sure about the scope of your project, the selection of KPIs or the data sets, do not hesitate to ask the course staff. We are there to keep you on track!

Have fun!

Appendix A

The **Final Report** should be structured as follows:

Section	Pg*	Description	Bonus	Weight
Cover page	1	Cover page with team number	No	Mandatory
Team	1-2	Present team, a short biography, and expertise	No	5
Executive Summary	1	Main findings, action points	No	10
Introduction	1-2	(Worldwide) Context, relevance	No	10
Strategy and Key Performance Indicator	3-5	Document the strategy, vision, and needs of the countries and explain how you operationalized them into KPIs	No	20
Data sources	2-3	Present the data sources used.	No	10
ETL, Dashboard and Data Analytics	10-15	Document the ETL process and present the (main) features of the Dashboard. In this section, you should also describe the multidimensional model. Describe what kind of descriptive and predictive statistical/data mining methods you implemented.	No	30
Limitations and Future Work	2-3	Any project has some limitations, think about the data sources, an incomplete strategy, too much focus, etc. Also, explain what you would do if you had more time and resources to extend the dashboard. Which new kind of findings would you expect?	No	10
Conclusion	1	Relate KPIs to the main findings	No	5
References	1-2	The references you used for the strategy and KPIs	Yes, max. 10	-
Appendix A	1	Team contribution	No	Mandatory
Appendix B	1	Link to the working dashboard	Yes, max. 10	-

*** (A4, size 11, Calibri or equivalent)**

Appendix B

The Dashboard project should contain the following material:

Section	Format	Description	Weight
Shiny dashboard	R	The graphical user interface to interact with the data	30
ETL Markdown	Rmd	The documented ETL pipeline is written in R Markdown	30
Queries	R	A queries.R file containing the queries to extract data from the database	20
Copy of datasets	CSV	A copy of the raw data used as input to make the RMD re-executable	10
Export of Postgres	SQL	Export the Postgres schema and data as a SQL file	10

The final folder structure in the git repository:

- **Dashboard**
  - Dashboard.R
  - Queries.R
- **Presentations**
  - Kick_off_X.pdf
  - Final_X.pdf
- **ETL**
  - ETL.Rmd
- **Data**
  - Raw_data
    - SourceA_data.csv
    - SourceB_data.csv
  - Schema
    - export.sql
- .gitignore
- README.md
- TeamX.RProj

Where X is the number of your team.

Caution, submissions in the Git repository which do not comply with this folder structure will not be graded

Appendix D

Tutorials and Assignment

WEEK	Date	Time	Topic	Type
17	**25-04-19**	**15:15-17:00**	Introduction to the final assignment	Tutorial
18	02-05-19	17:00-18:00	Data warehousing with R and Postgres	Tutorial
19	09-05-19	17:00-18:00	Dashboards	Tutorial
20	16-05-19	15:15-19:00	Project kick-off	Plenary
21	23-05-19	17:00-18:00	Feedback	Meeting
22	30-05-19	-	Ascension Day	Holiday
23	06-06-19	-	-	-
24	13-06-19	17:00-18:00	Questions	Q/A Lecture
25	20-06-19	15:15-19:00	Final project presentation	Plenary
26	27-06-19	-	-	-
27	04-07-19	Appointment	Retake final assignment	Oral examination

Deliverables

WEEK	Date	Time	Deliverable	Where
17	28-04-19	23:59	Teams	Blackboard
18	05-05-19	23:59	Focus, Countries, Data sources, KPIs, Relevance statement	Blackboard
19	15-05-19	23:59	First data set for a subset of KPIs	https://git.science.uu.nl/bi2019
20	**15-05-19**	23:59	Send slides of project kick-off	Blackboard
21	26-05-19	23:59	-	-
22	02-06-19	23:59	Working ETL and Dashboard Scripts	https://git.science.uu.nl/bi2019
23	**06-06-19**	**23:59**	Draft Report	Peergrade
24	**13-06-19**	**12:00**	Peer-review	Peergrade
25	23-06-19	23:59	Final Dashboard	https://git.science.uu.nl/bi2019
26	30-06-19	23:59	Final Report	Blackboard