BUINT Group Project: Task Description

Strongly growing availabilities of qualitative data provide new sources for gathering hidden information about the customer behaviour, market situation and the company's current situation. Separately or in combination, business data enables explaining sales rates or market changes and can give insights to key aspects which can be used for valuable estimated trends or forecasts. In addition, combining data of different sources may allow for profiling characters in customer segments or product/service groups. Such information might be of significant relevance for strategic decisions made by the management or stake holders for suitable positioning business in hard competitive environments.

Extracting infromation from large data sets and obtaining new insights for answering strategic questions is one data scientists' tasks. Depending on the given situation, a serious analysis also comprises acquiring additional data for better understanding or explaining the behaviour of a main data set. Reportings summarize the findings and present the interaction of different variables in a complex business environment for defining new strategies and making fact based decisions. Such a scenario of a complex business case shall be assumed and used for a case study where you analyze data sets in a group project. The results will be reported in a scientific paper and presentd to the class and lecturer that represent decision makers comprising the management and/or stake holders.

1. Project overview

The project comprises the combination, analysis and reporting of findings to specific questions on business data from different souces. By using the statistic analysis tool R, the group inpects, cleans and merges data sets from multiple souces (at least two) by passing through the whole Extract-Transform-Load (ETL) process as depicted in Figure 1. The resulting merged data provide a complex data set with several variables that will be used for answering the specific business questions. In contrast to a project documentation, the results of the analysis and the answers to the questions shall be reported in the form of an academic paper and presented to the full class by end of the semester. In the style of a real case data analysis project, there are no guidelines to the task split. Yet, after this project, each team member should be an expert about the investigated data set, the applied ETL process steps (including cleaning) and the conclusions drawn from the resulting merged data set.

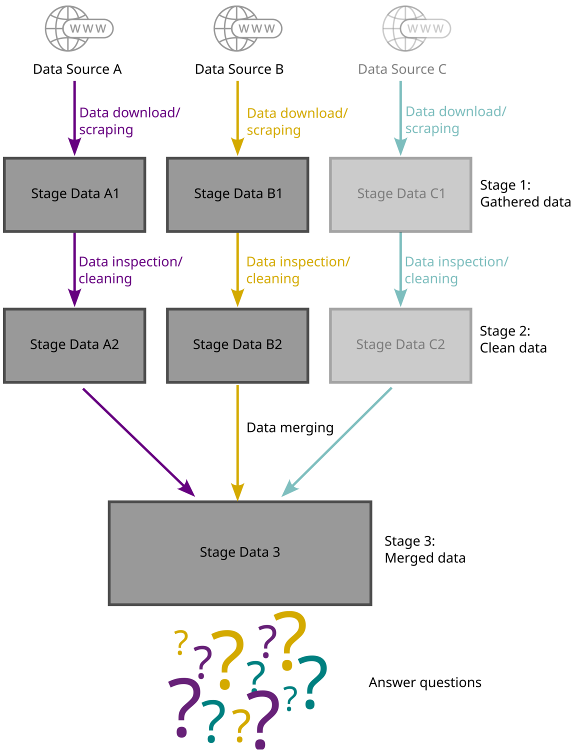

Figure 1: Data flow diagram of BUINT group project

2. Data sources

Data are available in different forms and shapes. It ranges from complex, multi dimensional structures to tabular representations on webpages, product information in webstores or as physical properties of goods for instance. For data scientist, some representations allow for smooth imports into data analysis tools while others have to be recorded, scraped ord extracted first. The rules of this projec do not make any restrictions about the data acquisitions. However, for efficiency reasons to keep the effort on a reasonable level, any data source can be used. Looking for data, it is recommended to start with large data pools such as kaggle.com, data.europa.eu, data.gov. Yet, the final data set for answering the business questions must be merged of at least two different data sets. Thereby, the source of the different data sets does not matter meaning, both can be from the same source (i.e. same website, shop, catalogue, etc.).

Depending on the contribution of the data set, it should comprise 1’000 … 10’000 observations or even more. While the main data set is obliged to have an extensive number of data points, additional data sets have less restrictions and may only comprise several hundreds of observations.

3. Narrative/Questions

Giving the project a serious and reasonable content for analyzing data and focussing on specific properties in chosen data sets, the group has to set up the project with a realistic narrative. In fact, the association of all the data analysis to strict defined problem statements implies the definition of three clear and accurately formulated business questions as they might be formulated from the management or other decision makers. Thus, the questions shall be formulated as close as possible but as open as it allows for a reasonable interpretation of the data.

The questions must be formulated in a way that they can be answered by means of the final merged data set and will rely on the main data set that defines the topic of the business that is tackled in this project. This means, the questions have to be formulated in such a way, that they require adding additional data to the main data resulting in a large merged data set. In addition, each of the three questions shall relate to one of the analysis categories: descriptive, predictive and prescriptive analytics.

4. Project reporting/presentation

Findings from the data analysis shall be reported in the form of an academic paper and presented in the last lecture. In contrast to a project documentation, the paper should focus on the problem statement (business questions), the applied methodology for answering the questions and the findings. In fact, the paper should have a length of maximum 8 pages and present the following structure:

1. Abstract
2. Introduction (with the research (i.e. business) questions at the end)
3. Data source (covering the source, quality and cleaning of data)
4. Data analysis (data organization and applied methodologies for analysis)
5. Findings (statistical results, figures, graphs)
6. Conclusion (answering questions)
7. References

For a uniform appearance, all reports shall be in a double column layout and follow the style guide of the IEEE template available for Word, LaTeX or Overleaf from: https://www.ieee.org/conferences/publishing/templates.html In addition, the report can also directly be written in R markdown using a double column layout by considering IEEE guidelines as good as possible.

Finally, the full work will be presented to the class and lecturer(s) in a short presentation of approximately 12’ followed by a 3’ question and answer sesssion. Thereby, all group members must have the expertise of the whole group work and being able to answer any questions about data sourcing, cleaning, merging, analsis, models and the drawn conclusions for answering the business questions.