EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Wirtschafts- und Sozialwissenschaftliche Fakultät

LEHRSTUHL FÜR MARKETING
Prof. Dr. Dominik Papies

Telefon +49 7071 29-78202
Telefax +49 7071 29-5078
dominik.papies@uni-tuebingen.de
www.uni-tuebingen.de/wiwi/marketing

"Data Science Project" Winter term 2024/2025

Prof. Dr. Stefan Mayer

As part of the M.Sc. „Data Science in Business and Economics"

I. Introduction

All students in the master program M.Sc. Data Science in Business and Economics enrol in an implementation-focussed module "Data Science Project". With 12 ECTS, the project has a substantial weight in the curriculum. This is also reflected in our expectation that students show a high degree of independence in their work in this module.

The basic idea of this module is that students **independently** design and program the entire process of a data science project from start to finish.

This requirement places a focus on **automation**, which means that the procurement of data, the structuring, reading, validation, modification as well as the analysis takes place (as far as possible) without the intervention of the analyst or user of the analysis.

A data science project should include the following components.

1. **Data acquisition**: The project acquires and reads in the data, for example by reading data from web pages ("web scraping"), by using one or several APIs ("application programming interface"), or by reading text documents. It is also possible for students to create a new dataset semi-manually themselves or to use an existing dataset as the basis for the project. The complexity and the effort of the data acquisition is considered in the grading of the project. Ideally, the data contains components that are unstructured, for example text or images.
2. **Data preparation**: The data are read in and made ready for computation and estimation. Among other things, this means that different parts of the data sets are read in and merged with other data sets, unstructured data (e.g., text) are converted into computable formats and correct identifiers for cross-sectional units (e.g., companies, brands, products, countries, persons) and for time points are generated.
3. **Data validation**: The data should be validated, for example by generating descriptive statistics and figures and by identifying implausible observations and outliers. After having defined the criteria for validation, this process should also take place largely without the users' intervention.
4. **Data analysis**: The primary focus of the project is a relevant substantive question, which can, but does not necessarily have to be of an economic nature. For instance, it can relate to a relation between to variables (correlational or causal), or to a prediction

of economic conditions (predictive analytics). The corresponding analysis should be implemented with adequate methods (traditional statistics or machine learning).

5. **Data visualization:** The structure and distribution of the data as well as the relations and results of the data analysis should be presented in an appealing and informative way, according to the state-of-the-art of data visualization.
6. **Presentation of the result**: All outputs of the project (e.g., tables, graphs, results) should be embedded in an interactive environment, for example in the form of a Shiny app. This should allow the user ("reader") to browse and explore the data, its distribution, the variables and the results (e.g., correlations, predictions). Examples of comparable Shiny apps can be found on RStudio's Shiny app website (https://shiny.rstudio.com/gallery/).
7. **Automation**: The entire project with the above-mentioned steps should run automatically. What is to be understood by this? Let us imagine any data set. We now divide this into two parts. The steps (1-6) are now programmed (i.e., the code is created) using one half of the data set. When the second half of the data set is read in, the entire code should be executed, and the results should be displayed without having to make any changes to the code. This is intended to mimic a workflow that is often relevant in practice: a data analysis and visualization structure is created, i.e., programmed and tested. Then, when new data comes in periodically, a user who did not create the code can evaluate and explore the data without having to "touch" the code. The important thing here is that the code is complete and well documented (embedded comments) so that external parties can understand and review the code.
8. **Version control and collaboration:** As soon as the topic, the data sources, the structure and the type of analysis are defined, the code must be organized by state-of-the-art version control. The recommended way is that all team members collaborate via GitHub, where all team members share their and through which version control is implemented. After completion (or earlier, if appropriate, e.g., to obtain support or feedback), the link to the repository is shared with the supervisor.

What is the difference between a master thesis and the Data Science Project?
A master thesis focuses on solving a relevant research problem, which must be embedded in the relevant literature. The analyses are described in the text, and the results are printed in the form of static tables or graphs. The thesis then interprets the results and draws conclusions. There are no requirements for automation and the reader "consumes" the text without interacting with the data themselves.

	Master thesis	**Data Science Project**
Elaborate research problem	High weight	Low weight
Literature work	High weight	Low weight
Automation	Low weight	Very high weight
Reader/user interacts with data	Low weight	Very high weight
Conclusion	Mostly author	Rather reader/ user

Evaluation/ Grading

Students present their results twice (in their groups): In an interim presentation, they present the plan and first steps. In the final presentation, they show the completed project, the central structure of the code, and how users can interact with the result.

The project is evaluated according to the following criteria:

1. Complexity of the overall project
2. Quality of the code
3. Complexity of the investigated problem, quality of implementation of the substantive problem, logical consistency of the problem and implementation

For assessment, students will produce the following components:

1. A presentation that motivates the central question, derives its relevance, and that refers to relevant literature as appropriate. The presentation also contains a discussion in which supervisors and other students ask questions. **The allocated time should be split roughly 50/50 between presentation of the project and discussion**.
2. As part of the final presentation, students present the finished "product", e.g., a website (e.g., a Shiny app) or a comparable way to interact with the data as a user.
3. Students provide the code on GitHub (see above). The code must be comprehensibly commented, and it must be clear from the code, the naming of the files, or any accompanying documentation, which files must be executed and in what order. Paths must be set so that no manual change of path names is necessary when the code is executed on other computers. In addition, students provide the data used in the project, usually through a random (small) sample on GitHub, plus the full data set as a FileTransfer (e.g., Dropbox).
4. Students make the interface (e.g. Shiny app) available to other students and teachers, e.g., by hosting it in the BWCloud.
5. Students create a short video (approx. 3 minutes) that clearly presents the project and the result to a wider circle of users. The video can be a screencast where only the voice of the presenter is heard while the final product is shown, or it can be a video presentation in which the presenters are also visible. The target group is a non-technical audience, e.g., in a business context it would be a manager, not a fellow data scientist.

Evaluation is done at the individual level, not at the group level, i.e., the group members' contributions to the project and the presentations will be considered in the evaluation.

All components (link to code, video, link to interface) are provided to the supervisor on the day prior to the final presentation.

II. Schedule

Friday, October 18, 2024 13:00 s.t. – 14:00 s.t.	Kick-off, assignment of topics, supervisors, and teams	E01 (Mohlstr. 36)
Wednesday, November 6, 2024 9:00 s.t. – 13:00 s.t.	Workshop “Presentation Skills”	room tba
Wednesday, November 27, 2024 14:00 s.t. – 18:00 s.t.	Interim presentation	E07 (Mohlstr. 36)
Friday, January 31, 2025 9:00 s.t. – 13:00 s.t.	Final presentation	E04 (Mohlstr. 36)

Topics and project ideas to be discussed at the kick-off meeting

In addition to our topic ideas that we will discuss in during the kick-off, **it is possible and desirable that students contribute their own topic ideas**. These *can* be inspired by current topics and topics from previous years. Students are invited to present their topic ideas including data sources and analysis methods to the module supervisor (Stefan Mayer) during kick-off. Then it will also be discussed whether additional subject-related supervision is necessary for the topic.

Topics that have been worked on in past years:

1. Does ChatGPT behave like a human?
2. Gender gap in academic publishing:
3. Earnings conference calls & firm valuation
4. Information Retrieval System for Annual Financial Reports:
5. Identifying Informative Content in User-Generated Reviews
6. Automatic Identification of Bicycle-Friendly Streets from Street View Images:
7. Corona database
8. Real estate prices
9. Gender gap in academic publishing
10. Sentiment and stock markets
11. ESG-Dashboard
12. Soccer prediction
13. Understanding and predicting (international) user preferences on Spotify
14. Fashion recommendations
15. NBA manager dashboard
16. Food price prediction
17. Detecting Out-of-Context Images in Online Misinformation
18. Heat Island – Identifying high temperature areas in cities during heatwaves and countermeasures using satellite images and segmentation learning
19. Detecting highway exits suitable for the placement of large-scale solar panels
20. AirBnb Listing Optimiser

...