The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Consultancy Style Dissertations in Statistics and Data Science: Why and How

Serveh Sharifi Far, Vanda Inácio, Daniel Paulin, Miguel de Carvalho, Nicole Augustin, Mike Allerhand & Gail Robertson

To cite this article: Serveh Sharifi Far, Vanda Inácio, Daniel Paulin, Miguel de Carvalho, Nicole Augustin, Mike Allerhand & Gail Robertson (2023): Consultancy Style Dissertations in Statistics and Data Science: Why and How, The American Statistician, DOI: 10.1080/00031305.2022.2163689

To link to this article: https://doi.org/10.1080/00031305.2022.2163689

Accepted author version posted online: 11 Jan 2023.

Consultancy Style Dissertations in Statistics and Data Science: Why and How

Serveh Sharifi Far, Vanda Inácio, Daniel Paulin, Miguel de Carvalho, Nicole Augustin, Mike Allerhand, Gail Robertson
School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK
*Corresponding author: Serveh Sharifi Far, serveh.sharifi@ed.ac.uk

Abstract

In this article, we chronicle the development of the consultancy style dissertations of the MSc program in Statistics with Data Science at the University of Edinburgh. These dissertations are based on real-world data problems, in joint supervision with industrial and academic partners, and aim to get all students in the cohort together to develop consultancy skills and best practices as well to promote their statistical leadership. Aligning with recently published research on statistical education suggesting the need for a greater focus on statistical consultancy skills, we summarize our experience in organizing and supervising such consultancy style dissertations, describe the logistics of implementing them, and review the students' and supervisors' feedback about these dissertations.

Keywords: Case study, Consultancy skills, Curriculum design, Statistical leadership, Workforce preparation.

1 Introduction

The fast-evolving demand for Statistics and Data Science education has led to a call for substantial revisions on the Statistics curriculum (Hicks and Irizarry, 2018) as well as for statistical leadership (Gibson, 2018). In particular, there is a growing understanding that experience in statistical consultancy is a key part of the undergraduate and postgraduate curriculum. In a 2007 read paper at *Journal of the Royal Statistical Society, Ser*. A, Taplin (2007) claimed that:

“Recently published papers on statistical education advocate that greater concentration should be put on statistical consulting skills.”

In line with these recommendations, in this paper, we introduce the “consultancy style dissertation” that has been designed for the MSc program in Statistics with Data Science at the University of Edinburgh. In the UK, master’s degrees are primarily one-year “taught” degrees, in which a third of the program credits is dedicated to a research project or a “dissertation” (often called a “thesis“ in the US). In our Statistics with Data Science MSc program, students study a total of 180 credits over a calendar year; 120 credits worth of taught modules and a 60-credit dissertation which is considered as one course.

A “traditional dissertation” is a written report based on a research study done by a master’s student during a specified time, under the supervision of an academic faculty member; the main emphasis of such a traditional master’s dissertation is on critical research, and the project’s research question can be supervisor-led or student-led (Katikireddi and Reilly, 2016). Although this traditional one-to-one dissertation style works well in general, contemplating the nature and size of our MSc program in Statistics with Data Science, we decided to explore a different style for its dissertation course.

The consultancy style dissertation model presented in this paper differs significantly from the classical paradigm, and consists of two independent research projects which students choose based on their interest from a list of projects. Students spend five weeks of full-time work on each of the two projects over the summer period. These projects are based on real-world data problems in joint supervision with industrial and academic experts. Such a dissertation aims to provide a consistent experience for all students in a cohort to develop consultancy skills, with all the subtleties it involves in the problem formulation, the analysis, and its communication. This dissertation reinforces the prominent features of data science, including critical thinking, knowledge exchange, transferable skills, working with real data on a problem-driven approach, group discussions, and the interdisciplinary nature of the field. The consultancy nature of this dissertation style is also beneficial in terms of providing students with a clear description of expectations in future jobs. The

supervision in this model is group-based and discussion focused so it optimizes the number of required consultancy style projects and also the number of faculty members for supervision purposes, which is quite helpful for large cohorts of students.

The dissertation model overviewed in this paper strongly encourages the development of core competencies that align with the recommendations, principles, and guidelines of the American Statistical Association Undergraduate Guidelines Workgroup (2014) and Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report ASA Revision Committee (2016) which we believe apply equally to postgraduate programs. In particular, one recommendation is that students should be exposed to analysing non-textbook data and be able to communicate, both written and verbally, complex statistical methods in basic terms to a non-technical audience. GAISE also recommends "integrating real data with a context and purpose" and "using assessments to improve and evaluate student learning". Another related reference is an NSF report by He et al. (2019) that makes recommendations in six areas, including training the next generation of statisticians and data scientists with modern skills and placing practice at the centre of statistics. In addition, the Royal Statistical Society (2017) accreditation program values similar standards for master's courses in Statistics alongside competence in planning, developing, and presenting an extended project in a dissertation.

This dissertation model is aligned with a variety of other recommendations and activities. For example, Aerts et al. (2021) advocate that students in Statistics and Data Science have "active experience" in key issues of "ethics, privacy, and data protection" in a consultancy project course or within the MSc thesis. In the same vein, Leman et al. (2015) argue for the need of an emphasis on what they called the qualitative-quantitative-qualitative layers of a problem, which main steps involve developing a qualitative understanding of the problem in question and its context, performing the quantitative analysis, and lastly communicating the results to a nontechnical audience or decision makers. Vance and Smith (2019) adopted and adapted this approach to assist statisticians and data scientists to improve their collaboration skills and their interdisciplinary impact. Our consultancy style dissertations have also borrowed inspiration from the literature on statistical

consulting (e.g., Tweedie, 1998; Pfannkuch and Wild, 2000; Chatfield, 2002; Cabrera and McDougall, 2002; Kauermann and Weihs, 2007; Unwin, 2007) as well as from consultancy activities such as the American Statistical Association Conference on Statistical Practice and the American Statistical Association DataFest. The first one gathers statistical practitioners and consultants, and data scientists to improve their ability to aid customers and organizations in solving real-world problems. The second one is a student event in which teams of undergraduates work during a specified time to answer real-world questions based on a large, rich, and complex dataset.

The setup of the consultancy style dissertation considers good supervision quality and good relationship between the students and supervisors, which is an influential factor in the progress and success of the undertaken research. In this regard, we follow good supervision practices, for example, by creating an enhanced collaborative environment in group supervision and minimizing the privatised nature of the "traditional" one-to-one supervision (Nzimande, 2011). We recognize that successful supervision is a dynamic relationship which requires the active engagement of both students and supervisors, and has a convergent nature of expectations on both partners (Aitken et al., 2020).

Although there are already many programs offering a module in consultancy or embedding it in an existing course (e.g., Smucker and Bailer, 2015; Martonosi and Williams, 2016; Greenhouse and Seltman, 2018; Davidson et al., 2019), to our knowledge MSc dissertations based on a consultancy style are not yet common and our goal in this article is to offer an account of our experience with such a model.

The rest of this manuscript unfolds as follows. Section 2 describes the main ingredients of our consultancy style dissertations, including their operational and logistic aspects, projects sources, examples of past projects, and allocation of students to projects. The students' feedback and supervisors' experience are presented and discussed in Section 3. Concluding remarks are offered in Section 4.

2 Consultancy Style Dissertation

2.1 The Statistics with Data Science MSc Program

This Statistics with Data Science MSc program ran for the first time in the academic year of 2016/17. This one-year program provides a strong foundation in Statistics with additional breadth in the mathematical and computer sciences, with an emphasis on interpersonal and communication skills which are key for either a future career in industry or for future studies and research. The program has experienced substantial growth, starting with about 20 students in 2016/17, having 40 students in the following academic year, and since 2018/19 there have been about 100 students enrolled in each academic year.

The program consists of:

(i) *Compulsory courses (60 credits)*: Six mandatory 10 credits courses to be taken over semesters one and two that provide core training in both classical and Bayesian statistics as well as in statistical computing. These courses help in further standardizing students' background.

(ii) *Optional courses (60 credits)*: Optional 10–20 credits taught courses to be taken over semesters one and two that cover the areas of optimization, operational research, and machine learning, among others.

(iii) *Dissertation (60 credits)*: This is the consultancy style dissertation course to be taken over the summer and it includes two independent research projects selected by students from a range of options.

From the compulsory courses we highlight *Statistical Research Skills*, whose aim is to prepare the students for their consultancy style dissertation by providing an experience of research-related techniques and skills, aspects of the statistical practice, and issues of data ethics. One of this course's assessment components involves orally presenting a poster about a peer-reviewed article published in a statistical journal. As we will see in Section 2.7, presenting a poster about the students' own work is also a component of the consultancy style dissertations.

2.2 Aims and Motivations for a Consultancy Style Dissertation

We created the consultancy style research dissertation prompted by the recognized need of postgraduate students in Statistics, and more broadly in Data Science, to be not only well versed in the methodological aspects of the discipline, but also in computational aspects, data wrangling, and real data analysis, in addition to being able to communicate to potential nontechnical audiences. As mentioned in the Introduction, this format aligns with guidelines and principles suggested by educational research. It also fulfils our belief that training in consulting skills and statistical practice should be part of any modern program in Statistics and Data Science.

The two projects in this dissertation are based on real data, and the particular problem-driven questions raised from them aim to encourage students to engage in critical thinking. The project supervisors guide students in the process by leading them to participate in discussions and exchanging knowledge, and by replying to individual questions when these arise. The interdisciplinary nature of the problems often requires students to conduct their own research in understanding the general and the particular context and later in finding the best ways to communicate their results to a potential non-technical audience.

A practical advantage of this dissertation model is that it scales well with the number of students. For instance, this scheme has allowed us, in the last academic years, to supervise about 100 students with 16 staff members making the dissertation supervision sustainable and scalable, even for a large number of students. For a comparison purpose, assume the traditional one to one supervision setting with the same number of staff members. In that case, each staff member would have to supervise around six to seven students on individual projects for the whole dissertation's duration (or, critically, more staff members would have to be involved to supervise all 100 students). This would make even slight increases in the number of students problematic in terms of having enough supervisors.

Another important point is that although the traditional dissertation format does not preclude collaborating with industry or other academic partners, it does make it very difficult in terms of providing the required number of high-quality consultancy style projects, academic supervisors, and industrial partners for a large cohort of students.

2.3 Dissertation's Format

This master's dissertation consists of two consultancy style research projects in, potentially, different application areas. The dissertation runs in two five weeks sessions over the summer and in each session, students work on one favourite project they have selected from four options and submit individually written reports for it. The projects have a consultant-client style where an industrial or academic partner presents a problem and students act as the consultants. In this section, we explain the main components of the consultancy style dissertation (and elaborate on some of them in the next sections) and provide a timeline of its main constituent parts.

Selection and allocation of projects: In March, students are given one-page descriptions of all available projects, which include a few core references for further reading and a list of the courses which are most related to that project. There are four available projects running in each session, which results in having about 25 students working on each project, this has shown to be an optimal number for us in terms of group size and students' satisfaction with the diversity of choices. However, we must mention that in the early years of running the program with smaller cohorts, we started with two or three choices of projects per session.

Help with academic writing: Before the dissertation season begins, around May, we provide a series of three online asynchronous workshops on "writing skills" to help and guide students in technical writing. The covered topics are: *Marking scheme and audience*, *Structure and line of argument*, and *Academic writing style*. These are further supported by an in-person questions and answers session to address students' enquiries about the format, writing style, and marking of the two projects. Access to a couple of dissertations written in the previous academic years that can serve as examples is also provided.

Supervision: The first of the two projects starts in June. Each project involves two faculty members as the leader and helper supervisors. The leader supervisor is the main person in charge of supervising the project and setting the main direction and further optional routes of the project, and the helper supervisor gives support to the

leader. For each project, there is an initial meeting where the project supervisor and the industrial partner introduce the problem to be investigated, set the main questions to be addressed, and answer any questions students may have. A week prior to the initial meeting, the data, a summary of the problem and relevant background readings are made available to students. After the initial meeting, students start working on one or more particular aspects of the problem. There are two one-hour group supervision meetings every week and students can ask the supervisors or industrial experts questions about the data, the client's goal and background, the methods they want to apply and their technical details.

Submission and feedback: After the first five weeks period, in early July, each student submits a written report for their first project. After one week of break, the second project starts for a period of five weeks, until about mid-August. The individual mark and corresponding feedback on the first project are given within two weeks of submission, thus allowing students to improve any structural bottleneck before the report about the second project is submitted.

Report and presentation: Each student writes two reports based on their own work, one per project. Each report has a limit of 5000 words, describing the statistical techniques used for the analysis, results obtained, and corresponding interpretation in the context of the problem. The dissertation's final mark consists of, 40% for project one's report, 40% for project two's report, and 20% for a poster presentation. The last element aims to assess the communication skills of students via a ten-minute poster presentation in late August after the two written reports have been submitted. Students choose one of the two projects they worked on to present. Each presentation is followed by five minutes of questions from the marker.

Results: At the end of the dissertation period, we usually send a few of the best final reports to the academic or industrial partner to investigate the answers produced by students to their given questions. They are also invited to be present at the poster presentations. We have had at least one case that resulted in a student applying and securing a job, immediately after the end of their MSc, in the external company with which they did one of their dissertation projects.

2.4 Projects Sources and Topics

We provide eight new projects in each academic year. For a project to be considered appropriate for a consultancy style dissertation, it must cater for a wide range of abilities in students. There should be at least one question posed which can be answered in part or completely with the data. In other words, there must be a minimal analysis to attempt to answer the question which can be conducted by all students in the cohort based on the training they have received during the program. Also, there must be several interesting avenues in terms of statistical methods and approaches to data analysis, or the possibility of using extra data which can be explored by the more ambitious students. These questions are given to students by the project supervisor and the industrial/academic partner in the initial session of the project. We also encourage students to explore any other angle of the project that interests them after discussion with the supervisors. However, we make it clear to them that the quality of the work is more important than the quantity and the number of attempted aims and questions, and well written reports that include a comprehensive analysis of just one of the project aims would be well received. Given the time frame of each project, we aim to avoid projects with no specific question from the industrial/academic partner, or a vague one that may not be possible to answer with the given data. We acknowledge, however, the fact that it is important that students learn about investigating what can and cannot be answered with the available data.

Our *sources* of finding these consultancy style projects, can be distinguished into two categories:

(i) *Projects based on academic collaboration with other university departments or external organizations*: These projects are usually proposed by the academic staff members based on their ongoing research works in collaboration with academics from other domain areas than Statistics. Often, they are proposed by our Statistical Consultancy Unit, which is a good source of such projects due to its nature of collaborating with academic and industrial partners.

(ii) *Projects based on industrial problems brought up by external organizations*: This type of projects is provided by the School of Mathematics' Business Development Team, which helps with fostering collaboration with industry partners. This has resulted in collaborations with large and well established companies like Amazon or innovative start-ups like Thrift, or with government based institutions, like Public Health Scotland.

The proposed projects are then inspected by the dissertation course organiser (who is responsible for finalising the projects and making sure that the course runs smoothly) and the program director, and eight of the most appropriate ones are selected for the next run of the course. We include an example of each project type here with a brief explanation of the problems and the methods students applied.

(i) **Understanding the relationship between antimicrobial resistance and hospital prescription rates**

Problem and data: This project was a collaboration between academics in the statistics group and a clinical researcher working at the University of Edinburgh. Antimicrobial resistance is a global public health crisis requiring widespread surveillance to determine the prevalence of resistant organisms. One of the challenges for the medical and research communities is understanding where and how antimicrobial resistance arises, its implications for human health, and the effectiveness of measures to prevent resistance arising. Over-prescription of antimicrobial drugs is known to drive antimicrobial resistance in local populations (Costelloe et al., 2010). There is evidence that residing in high antimicrobial consuming communities, such as hospitals and countries with unrestricted use, affects faecal carriage of resistant organisms. However, it is not fully established how antimicrobial use within a hospital community affects resistance in the full range of cultured pathogens. This project aimed to quantify the relationship between the amount of antibiotics prescribed per hospital ward and antimicrobial resistance found in clinical specimens taken from a representative sample of patients on these wards. The given data set provided contained information on patient demographics

and antimicrobial prescription rates per ward in a specific hospital, as well as accompanying data on resistance found in various cultural pathogens.

Methods and outcome: Students used generalised linear mixed models to identify variables explaining variation in resistance rate among patients in different wards, accounting for pseudoreplication and confounding within the dataset. Some students additionally explored the more challenging question of examining potential lag effects between drug prescribing in wards and antimicrobial resistance detected in patients. The majority of students chose to drop the missing values from the dataset, but some of them explored the more challenging option of imputing the missing values via different methods and comparing the results with or without imputation. This work assisted the clinician in identifying factors important in influencing hospital infection control procedures prescribed antimicrobials to patients within hospitals and have public health policy implications for the management of antimicrobial usage.

(ii) Neural de-duplication and record linkage

Problem and data: This project was suggested by Amazon, giving students an opportunity to apply their machine learning abilities to work on record linkage or entity resolution as the task of grouping similar entities across one or more data sources. Within Amazon, these methods are used to find relationships between products, drive search and discoverability of products, and improve the shopping experience on the website. A key component in many Record Linkage systems is a matching component that determines whether pairs of records refer to the same entity. Students investigated some related questions, for example, whether we can learn a model from raw text data to outperform hand-crafted features, how different word segmentation techniques affect the performance, or if we can learn unsupervised representations and apply transfer learning for this domain. We used open-source datasets (see Leipzig, 2022, for some examples) for this project and ran two versions of it in two years.

(iii) *Methods and outcome:* Students used natural language processing and deep learning methods to match items in different databases (record linkage) or within the same database (deduplication). These algorithms are quite complex since several preprocessing steps and a significant amount of parameter tuning is required. Even though most of students did not have previous experience with natural language processing and deep learning methods, they managed to learn and apply such methods after getting help at the supervision meetings. Students achieved different levels of prediction performance depending on their undertaken decisions. Some of them managed to obtain outstanding prediction performance on the provided datasets (F1 score of over 99.8%), which makes us optimistic about the potential of deep learning methods for such problems.

2.5 Allocation of Projects to Students

Since the 2018/19 academic year, we have run eight projects per year, meaning that we give students four options in each set to choose from. Our overall student numbers meant that we had between 20–30 students per consultancy style project during the 2018-2021 period. The allocation of students to projects has been a non-trivial question as we need to take into account their interests and also the practical constraints on the group sizes (i.e. they should be similar between different projects). We list the criteria that we took into account during this process:

1. *Student preferences*: We announce the projects to students before the end of March by sharing a brief description of each project created by the industrial expert or the supervisors, and ask them to submit their preferences via an online form. They have to rank the four projects in both sets from 1 (most preferred) to 4 (least preferred).
2. *Constraint on the number of students per project*: We compare how many students make each project their first choice, and decide on what is the number of students that are allocated to that project. We make group sizes for popular projects slightly larger, but ensure that the difference between group sizes is no more than 15%.

3. *Academic record of students*: Since we sometimes need to make a choice between several students vying for a popular project, we decide to make the available academic record at the time of allocation a factor in our considerations (we have used the mean marks from the courses taken in the semester 1 as the quantitative indicator of academic performance).

We have formulated a loss function that takes all of these criteria into account in a balanced way, and found an efficient optimization algorithm that is guaranteed to find the optimal allocation according to this loss function. Details of the algorithm for students' allocation to projects are given in Appendix A.

2.6 Supervision

During each of the five weeks a project lasts, there are weekly drop-in discussion sessions attended by both the project leader and helper supervisors. It is important that the leader supervisor is experienced in the relevant area to be able to lead the group. These sessions are the main place for students to seek advice, get feedback, and regularly check the work they are developing with the supervisors. Students are also encouraged to discuss their work and ideas with their peers during these drop-in sessions. Students and supervisors also use a text-based online discussion forum named Piazza, to keep the discussions rolling during the week. Occasionally, the industrial/academic partner may also attend some of the weekly contact sessions. However, they are only required to attend the initial meeting, so if there are students questions for them after the initial meeting, the supervisors collect those weekly, email the industry partners, and communicate the reply back to students via email or Piazza. Usually students have questions for clients in the first 1-2 weeks, then questions are more about analyses and are answered by the academic supervisors.

While in-person project supervision allowed project leaders and helpers to make sure all students received enough contact time, this was more challenging following the changes that needed to be made to project supervision during the pandemic, when all supervision duties had to be delivered online. After the Covid-19 outbreak during the academic year of 2019/20, we held the supervision of the dissertation projects fully online. The weekly two-hour contact session was broken down to two

one-hour sessions on two days per week to ensure continued support for students throughout the week, and to ensure that students had contact time with supervisors at least once a week, which was particularly important for those who were not able to attend both sessions due to conflicting time zones. The Zoom platform was used to deliver online sessions. A typical session would start with a general introduction by the supervisors in which questions asked by students via Piazza in the previous week were discussed with a focus on those which were thought to be of interest to the whole group. Following the introduction, students were invited to attend breakout rooms to continue discussions in smaller groups. Students were often allocated to breakout rooms randomly, and efforts were made by supervisors to ensure that students of different abilities and confidence levels were evenly mixed among breakout rooms. There were also other ways of managing the breakout rooms based on the supervisors' preference, for example, specific problem-based rooms during the later weeks of the process.

In the following academic year, 2020/21, some of the pandemic restrictions were lifted and allowed us to have a more flexible supervision format. In this round, we held the two weekly sessions in a hybrid format, in which students could join the Zoom call from a distance or from a campus classroom. In this classroom, students were accompanied by one of the two project supervisors and had an opportunity to meet and interact in-person.

The main challenge of online and hybrid modes was the difficulty of engagement between students with supervisors and each other. Although breakout rooms were used for small group discussions, some students and supervisors believed that they could not exactly replicate the collaborative atmosphere that was common in an in-person session.

In the academic year of 2021/22, the two-hour weekly drop-in sessions were held again fully on-campus but with the added advantage, compared to the pre-pandemic sessions, of the industrial/academic experts joining more frequently online via Zoom.

2.7 Marking

A detailed and descriptive marking scheme is made available to students a week before the start of the first project. Both the project leader and helper contribute to the mark of the project. Recognizing that there is no single correct analysis for this type of project, marks on the written component are allocated on a combination of statistical approach, and justification and interpretation of results in context and its presentation.

The marking scheme for each of the two written reports is made of:

(i) 60% for how well the report answers the question: “Can the report be presented to the client?”. In Table 1, we transcribe descriptors of this component.
(ii) 10% for executive summary. This is different in nature from a typical dissertation’s abstract, and it should provide an accurate description of the problem under investigation and summarise the statistical findings in a highly effective manner, as well as, propose solutions to subject matter experts in a compelling way.
(iii) 10% for style and clarity of writing and organization and presentation of the report.
(iv) 10% for providing appropriate background material and references.
(v) 10% for the quality of coding and its reproducibility.

We use an online marking system which enables the marker to select appropriate grades (A–E) for each of the five marking criteria in marking a report. Each grade corresponds to a general and brief description about the quality of the work, so selecting it provides a general feedback for the student. After this, the marker types comments on the specific aspects of that report to produce a written individual feedback.

For the poster presentation, which as mentioned before is assigned 20% of the whole dissertation mark, the main evaluation criteria are: structure, clarity, and engagement with the audience, and each of these criteria is equally worth. Again,

written comments for each criterion are entered into the marking system by the marker.

3 Feedback and Experience

3.1 Students' Feedback

The weekly contact sessions provide an informal opportunity for getting direct feedback from students on how the project evolves. We also rely on a midway written and anonymous feedback from students which would give the supervisors a clear idea of what aspects of the supervision could use improvements. Since the 2020/21 academic year, we decided to collect official student feedback at the end of the projects, similar to what is commonly done for all of our other courses.

In a cohort of 108 students in 2020/21, 16 responses (15% of the class) were submitted for the anonymous questionnaire given to students at the end of the second project. Although this response rate is relatively low, the feedback is consistent with informal comments we have received from students over the past years. Table 2, summarises the responses for six Likert scale questions (originally with five categories but we have reported them in three categories here). Students were asked to reflect on the whole dissertation experience in answering the questions. Questions 1–3 focus on the scientific quality of the two projects and questions 4–5 address the supervision quality. Question 6 asks about their overall satisfaction and about 70% of the students agreed that they were satisfied with the quality of their dissertation.

Three descriptive questions were also given to obtain more detailed information on students' experiences. Students were asked "What did you find most valuable about the dissertation course?" and the answers covered some of our main purposes of doing a consultancy style dissertation. They appreciated learning about working with "real data" in a "hands-on" experience, which gave them an understanding of how statisticians and data scientists work. They mentioned the "active role" of supervisors in the process and their "useful feedback", and also engaging with external data experts as valuable elements.

The next question was about their experience and what improvements, if any, they would make to the dissertation course. Some responses were about the specific ways of managing the hybrid sessions and breakout rooms for discussions. We had chosen to allow the academic supervisors to decide how to run the sessions and set the breakout rooms arrangements, and also encouraged students to communicate their opinions on these directly to the supervisors during the sessions. A comment regarding the planning of the projects was that they would prefer to know the allocation sooner. Also, several students stated that having a break between the two projects would be helpful to reset and prepare for the second one. Both these requests were reasonable and we implemented them in the schedule for the next academic year. Three students stated that they would prefer to work for a longer time on one project in more detail, and one student said they would prefer to have more than two projects. Although we recognize these are contradicting opinions, we believe having “two” projects has worked well in fulfilling the aims of a consultancy style dissertation and in giving students further breadth on domain application areas. There was a request for more “personal” supervision. This is not an unexpected comment, especially from students who may be familiar with the traditional one-to-one MSc supervision. However, we make the objectives of a consultancy style dissertation clear to students early in the program, remind the academic supervisors to aid every student in their group, and encourage students to be proactive in the supervision sessions.

Another question specifically asked for feedback about the academic supervisors. The responses were generally positive and described the academic supervisors’ performance as “helpful”, “supportive” and “exemplary”. The responses to this question about the industry partners were similar, with one potential point of improvement: being given the opportunity to have more time with them. We have observed the interaction between students and industry experts to be interesting and productive for students and we do encourage more of such exchanges. However, there are limitations on the industry partners’ side in terms of the duration and nature of the time they can spend with students. The industry experts also get invited to join the final dissertation poster presentations. We have been running these

presentations online since the beginning of the pandemic and this has increased the participation of industry experts.

3.2 Supervisors' Experience

The dissertation is also a learning experience and most students measurably improve in the second of the two successive projects in terms of familiarity with consultancy work based on real data. For example, students tend to ask more relevant questions about the project aims and data during introductory sessions for the second round of projects than in the first one, and the citations they use in their reports become more relevant and correctly formatted during the second project.

According to our supervision experience, students whose prior experience has mainly been classroom lectures may feel intimidated when it comes to addressing a real-world question, especially if they have had little or no previous experience of analysing data in a consultancy type project, and these students benefit from more personal attention from the project's supervisors. Less confident students are more prone to form groups and mimic each other's work. There are also students who want to apply a particular method or approach they have learned and are enthusiastic about, which may not be the most appropriate one for the problem at hand.

There is anecdotal evidence that students are more comfortable describing a statistical or machine-learning method than using it in practice or interpreting the meaning of its results. Possibly this is due to their prior learning experience. For example, students tend to reproduce textbook material about missing data techniques but they usually struggle to deal with missing values in their analysis. Another common example is when students extract formulae and computer code from online resources, but then have difficulties trying to adapt it to their project.

4 Closing Remarks

With the fast-evolving demand for statisticians and data scientists, the need to prepare students to tackle real-world problems is more important than ever. In this article, we have introduced the "consultancy style dissertation" offered in our MSc in

Statistics with Data Science program. This dissertation, consisting of two independent projects, runs in collaboration with industrial and academic partners from different domain areas and requires students to analyse real (likely very large) datasets and to communicate their findings, written and verbally, to both experts and clients. These are problem-centric projects and any method or statistical technique used should be justified in the context of the problem being solved. The consultancy style dissertation strongly aligns with modern recommendations for the curriculum in statistics. Since a group of students work independently on the same problem or case study, this dissertation's format has the added advantage of scaling well, even for large cohorts.

Overall, we believe that our consultancy style dissertation leads to a reinforcement, in most students, of our initial goals of critical thinking, knowledge exchange, and transferable skills. Nevertheless, not all students finish the MSc program with the same maturity and this is especially true in what concerns their data analysis skills. Yet, evaluating how much students have developed their data analysis skills (on which we include interpretation of the results) is made clear when interacting with them in the weekly discussion sessions and when marking their submitted reports.

Common feedback received from students prior to the starting of this dissertation process is that the projects feel "a little controversial" to them (compared to traditional dissertations), however after experiencing the process, the same students recognized them as very good preparation for future statisticians and data scientists. Indeed, skills such as statistical leadership, and written and verbal communication aimed at nontechnical audiences, promoted under the umbrella of these dissertations, in addition to the technical skills gained during the courses, should be of high value in the job market and also in academia.

Acknowledgements

We thank the Editor, Associate Editor, and two anonymous Reviewers for their comments and insightful remarks. We are grateful to our colleagues of the Statistics research group at the University of Edinburgh, as well as to the external projects experts, for having provided through the years interesting and challenging projects,

and for their contribution to supervising the consultancy style dissertations. We are also grateful to the School of Mathematics' business development team for their help in finding industrial partners.

Appendix

Allocation Algorithm

Here we briefly overview the method for allocating the students to projects that was mentioned in Section 2.5. Our method proceeds by optimizing a loss function that is defined in terms of the students' preferences, the number of students per project, and the academic record of the students. Let n be the number of students (typically 80–120 in our case), m be the number of projects per set (m = 4 in our case). Let S_i be the academic score of student i and let $p_i(j)$ be the preference of student i for project j, for $i = 1, \dots, n$ and $j = 1, \dots, m$. Specifically, $p_i(j)$ is the order of project j amongst their list of preferences (so $p_i(j) = 1$ if j is their most preferred project, and $p_i(j) = m$ if j is their least preferred project). Now, let $[m] = \{1, \dots, m\}$ be the set of projects and let $\mathbf{a} \in [m]^n$ be a possible allocation of students to projects, with a_i denoting the number of the project allocated to student i. Assume the dissertation's organizers decided to allocate $N_1, \dots, N_m$ students to each project. The goal is to find an allocation $\mathbf{a}$ that takes into account all of these factors in a balanced way. We do this by choosing $\mathbf{a}$ as a minimizer of the loss function

$$L(\mathbf{a}) := \sum_{i=1}^{n} p_i(a_i) S_i, \quad (1)$$

subject to the constraints

$$\sum_{i=1}^{n} 1[a_i = k] = N_k, \quad 1 \le k \le m,$$

i.e., there are $N_1, \dots, N_m$ students allocated to each project, respectively. The loss function in (1) sums up the order of the allocated project in their list of preferences for each student, weighted by their academic score.

Now we are going to rewrite this problem into a well-known combinatorial optimization question called the assignment problem (see Burkard et al. (2012)). In this problem, there are n agents and n tasks, and each agent has a certain cost for performing each task (i.e., the costs can be represented in an $n \times n$ matrix). The goal is to allocate exactly one task to each agent in a way that minimizes the total cost. The so-called Hungarian algorithm was the first one to offer a polynomial time solution to this problem. As far as we know, the most efficient method that is currently available is the Jonker–Volgenant algorithm (Jonker and Volgenant, 1987), which finds an optimal solution in $O(m^3)$ time.

To rewrite our problem as an assignment problem, for every $1 \leq j \leq m$ (m is the number of projects), we create N_j tasks for project j. The total number of tasks created is $\sum_{i=1}^{m} N_j = n$, i.e., this is equal to the total number of students. Hence effectively each task corresponds to a place in the corresponding project. For student i and task k corresponding to project $j = P(k)$ (here $P : [n] \to [m]$ denotes a mapping from tasks to projects), we set the cost function of the student performing that project

$$C_{i,k} = S_i p_i(P(k)). \qquad (2)$$

This is the academic score of the student multiplied by their ranking for project $P(k)$ corresponding to task k.

The Assignment problem aims to finds an optimal permutation π on $[n]$ (i.e. $\pi(1), \ldots, \pi(n)$ is a reordering of $1, \ldots, n$) that minimizes

$$A(\boldsymbol{\pi}) = \sum_i C_{i,\pi(i)}.$$

It is easy to see that when $C_{i,k}$ is chosen as (2), finding a permutation π that minimizes $A(\pi)$ also yields a minimizer of (1) as $\mathbf{a} = P(\boldsymbol{\pi})$ (i.e. $a_i = P(\pi(i))$ for every $1 \leq i \leq n$). Hence we are able to use this reformulation to optimize (1). A highly efficient implementation of the Jonker–Volgenant algorithm is available in the `TreeDist` R package (Smith, 2020), which was able to solve the assignment problem for 110 students in less than a second.

We have implemented this approach in R. Our code assigning students to projects is available on the Github repository Paulin (2022).

References

Aerts, M., G. Molenberghs, and O. Thas (2021). Graduate education in statistics and data science: The why, when, where, who, and what. *Annual Review of Statistics and Its Application 8* (1), 25–39.

Aitken, G., K. Smith, T. Fawns, and D. Jones (2020). Participatory alignment: a positive relationship between educators and students during online masters dissertation supervision. *Teaching in Higher Education 0* (0), 1–15.

American Statistical Association Undergraduate Guidelines Workgroup (2014). 2014 curriculum guidelines for undergraduate programs in statistical science. *Alexandria, VA: American Statistical Association.*

Burkard, R., M. Dell'Amico, and S. Martello (2012). *Assignment problems: revised reprint*. SIAM.

Cabrera, J. and A. McDougall (2002). *Statistical Consulting*. New York: Springer.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society: Series D (The Statistician) 51* (1), 1–20.

Costelloe, C., C. Metcalfe, A. Lovering, D. Mant, and A. D. Hay (2010). Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *BMJ 340*.

Davidson, M. A., C. M. Dewey, and A. E. Fleming (2019). Teaching communication in a statistical collaboration course: A feasible, project-based, multimodal curriculum. *The American Statistician 73* (1), 61–69.

Gibson, E. W. (2018). Leadership in statistics: increasing our value and visibility. *The American Statistician 73* (2), 109–116.

Greenhouse, J. B. and H. J. Seltman (2018). On teaching statistical practice: From novice to expert. *The American Statistician 72* (2), 147–154.

Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report ASA Revision Committee (2016). Guidelines for assessment and instruction in statistics education (gaise) reports. http://www.amstat.org/education/gaise.

He, X., C. Madigan, J. Wellner, and B. Yu (2019). Statistics at a crossroads: who is for the challenge? https://www.nsf.gov/mps/dms/documents/Statistics_at_a_Crossroads_Workshop_Report_2019.pdf.

Hicks, S. C. and R. A. Irizarry (2018). A guide to teaching data science. *The American Statistician 72* (4), 382–391.

Jonker, R. and A. Volgenant (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing 38* (4), 325–340.

Katikireddi, S. V. and J. Reilly (2016). Characteristics of good supervision: a multi-perspective qualitative exploration of the Masters in Public Health dissertation. *Journal of Public Health 39* (3), 625–632.

Kauermann, G. and C. Weihs (2007). Statistical consulting. *AStA Advances in Statistical Analysis 91* (4), 343–347.

Leipzig, D. G. (2022). Benchmark datasets for entity resolution. https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution.

Leman, S., L. House, and A. Hoegh (2015). Developing a new interdisciplinary computational analytics undergraduate program: A qualitative-quantitative-qualitative approach. *The American Statistician 69* (4), 397–408.

Martonosi, S. E. and T. D. Williams (2016). A survey of statistical capstone projects. *Journal of Statistics Education 24* (3), 127–135.

Nzimande, M. N. (2011). Exploring students' experiences of producing a masters dissertation. Master's thesis, University of KwaZulu-Natal, Edgewood.

Paulin, D. (2022). Allocation algorithm for SwDS projects. https://github.com/paulindani/projectallocation.

Pfannkuch, M. and C. J. Wild (2000). Statistical thinking an statistical practice: Themes gleaned from professional statisticians. *Statistical Science 15* (2), 132–152.

Royal Statistical Society (2017). Master's (level 7) standards in statistics. https://rss.org.uk/RSS/media/File-library/Membership/Prof\%20Dev/rss-level7-standards.pdf.

Smith, M. R. (2020). Information theoretic generalized robinson-foulds metrics for comparing phylogenetic trees. *Bioinformatics 36* (20), 5007–5013.

Smucker, B. J. and A. J. Bailer (2015). Beyond normal: Preparing undergraduates for the work force in a statistical consulting capstone. *The American Statistician 69* (4), 300–306.

Taplin, R. (2007). Enhancing statistical education by using role-plays of consultations. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170* (2), 267–300.

Tweedie, R. (1998). Consulting: Real problems, real interactions, real outcomes, with a resources appendix by sue taylor. *Statistical Science 13* (1), 1–29.

Unwin, A. (2007). Statistical consulting interactions: a personal view. *AStA Advances in Statistical Analysis 91* (4), 349–359.

Vance, E. A. and H. S. Smith (2019). The asccr frame for learning essential collaboration skills. *Journal of Statistics Education 27* (3), 265–274.

Table 1 Marking descriptors for the question “Can the report be presented to the client?”

Grade	Descriptor
A1: 90 – 100%	The report could be presented to the client or collaborator without revision. The analysis is sound so that conclusions are well supported statistically. Interpretation is mature. The project demonstrates a clear overview of the work, without getting lost in details, and is free of all but the most minor statistical errors, not altering the conclusions. The report clearly links the statistical analysis to the practical problem/application, including the limitations of the analysis.
A2: 80 – 89%	The report could be presented to the client or collaborator with little or no revision. The analysis is sound so that conclusions are well supported statistically. Interpretation is reasonably mature. The project demonstrates a clear overview of the work, without getting lost in details, and is free of all but minor statistical errors. The report clearly links the statistical analysis to the practical problem/application, including the limitations of the analysis.
A3: 70 – 79%	The report could be presented to the client or collaborator with little revision. The analysis is sound so that conclusions are well supported statistically. Interpretation is mostly reasonably mature. The project demonstrates a clear overview of the work, without getting lost in details, and is free of all but minor statistical errors.
B: 60 – 69%	The project could be presented to the client or collaborator after a round of revision, but without having to re-do much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but it is basically sound. A good grasp of the statistics and context, so that interpretation is reasonable.
C: 50 – 59%	Major re-working required before the project could be presented, but containing some sound statistics demonstrating understanding of statistical modelling and its application. Reasonable presentation and organization.

Grade	Descriptor
D: 40 – 49%	Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.
E-H: < 40%	Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or badly organized presentation.

Table 2 Students' responses to six questions on the quality of the dissertations

	Disagree	Neither agree or disagree	Agree
1. The dissertation has been intellectually challenging.	6.25%	6.25%	87.5%
2. The dissertation has developed my skills and abilities.	12.5%	0%	87.5%
3. The dissertation has given me a good idea of what statisticians and data scientists do.	12.5%	0%	87.5%
4. The amount of available supervision was satisfactory.	25%	18.75%	56.25%
5. The type of hybrid (online/on-campus) supervision was satisfactory.	25%	37.5%	37.5%
6. Overall I am satisfied with the quality of the dissertation course.	18.75%	12.5%	68.75%