A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data

Daniel Frischemeier | Rolf Biehler | Susanne Podworny | Lea Budde

1Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany
2Institute of Mathematics, Paderborn University, Paderborn, Germany
3Institute of Computer Science, Paderborn University, Paderborn, Germany

Correspondence
Daniel Frischemeier, Faculty of Mathematics and Computer Science, University of Münster, Fliednerstraße 21, Münster D-48149, Germany. Email: dfrische@uni-muenster.de

Abstract
In this paper, we will describe an introduction to Data Science for secondary school students. We will report on the design and implementation of an introductory unit on “Data and data detectives with CODAP” in which secondary school students used the online tool CODAP to explore real and meaningful survey data on leisure time activities and media use (so-called JIM-PB data) in a statistical project setting as a starting point for data science. The JIM-PB data set served as a valuable data set that offered meaningful and exciting opportunities for data exploration for secondary school students, and CODAP proved to be a valuable tool for the first explorations of this data.

KEYWORDS
Teaching, CODAP, data science education, percentages, statistical questions, survey data

INTRODUCTION

A competent understanding and handling of data have become indispensable today because many decision-making contexts in the world, including business, politics, and society are supported by pieces of evidence based on data [7] [12]. To support pupils in their development as citizens, it is crucial that reasoning about data is stimulated in school as early as possible [4]. Data have meaning for society, and citizens’ ability to understand data are vital for a vibrant democracy. For data and statistics about society (such as statistics about poverty, health, economy, migration), the ProCivicStat Project (see: http://iase-web.org/islp/pcs/) defined the term civic statistics, as statistics, which concerns us as citizens, and which are essential since they have a considerable influence in our life because of the evidence-informed decisions made in business, politics, or economy.

There are several definitions of data science. Data science has to be regarded as a discipline situated at the crossroad of statistics, computer science, and subject matter disciplines, where data are used. When considering the implementation of data science education at the school level, two fundamental issue arise: first there is the question of which kind of data (such as survey data, observational data, data from sensors) to focus on and second there is the question about which kind of tools (educational software like TinkerPlots or CODAP, spreadsheet software like Microsoft Excel, or professional programming tools like R or Python) to choose for analyzing data.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. Teaching Statistics published by John Wiley & Sons Ltd on behalf of Teaching Statistics Trust.
An overview of different kinds of tools for analyzing data from an educational perspective can be found in Biehler et al. [3]. TinkerPlots, Fathom, and CODAP were developed as tools for data analysis that offer learners an easy entry into data analysis by requiring no programming skills but of course, are, therefore, limited in their data exploration potential. On the other hand, there are professional tools like R and Python, which offer a vast landscape of different exploration methods but also require learning the programming language alongside the development of statistical thinking, which is demanding for learners at the school level and in introductory college courses. We choose to start with investigating survey data with CODAP as a first step to data science. Our main intention for the choice of the data was that the data used should be of reasonable size, contain multivariate data, and be of interest to high school students.

2 | A COURSE ON DATA SCIENCE FOR SECONDARY SCHOOL

In this paper, we will report on the work of our project ProDaBi (Project Data Science and Big Data at school level). The aim of this interdisciplinary project between statistics and computer science education, which was initiated by the Deutsche Telekom Stiftung, is to investigate in which way and with what topics data science can be implemented in the school curriculum at the secondary level. For this purpose, we have developed a year-long course on data science for upper secondary school (grade 12) in which we can test out several ideas. We have taught the course twice, in the school year 2018/2019 and school year 2019/2020. In this paper, we will report on the course in the school year 2019/2020.

The students on the course have no specific knowledge of statistics and data analysis. Therefore, we postulate a learning trajectory consisting of three modules: (1) basics of data analysis and statistical thinking; (2) algorithmic thinking and machine learning; and (3) application in a comprehensive project. Module 1 “Data and data detectives” is supposed to introduce data exploration and enhance statistical thinking with digital tools first with CODAP (unit 1) and then with Python-based Jupyter notebooks (unit 2). In the “Machine Learning” module 2 [5], the students are introduced to data-driven machine learning, mainly decision trees (unit 1) and artificial neural networks (unit 2). In a third module, our students are confronted with a larger data science project, which is carried out in cooperation with partners from industry and administration and in which the students are supposed to use the knowledge on statistical thinking and machine learning methods they have learned in modules 1 and 2. The overarching aim of the course is to introduce secondary students into different facets like statistical thinking with digital tools (in module 1) and computational thinking (in module 2) of data science and to enable them to work collaboratively on real data science projects using statistical and computational methods. Therefore, the first two modules aim to prepare the students in statistical thinking and machine learning methods so that they can apply the knowledge gained to an authentic data science project in module 3. Concerning our students’ introduction into statistical thinking, we followed the idea of introducing our students into the five phases “problem”, “plan”, “data”, “analysis”, and “conclusions” of the Problem, Plan, Data, Analysis, Conclusions (PPDAC) cycle [14]. To realize a smooth introduction into statistical thinking with digital tools, we decided to use survey data first and afterward, we use experimental data (noise and dust data from sensors) within unit 2 in module 1.

The aim was to introduce our students to the main components of statistical thinking.

In this paper, we will concentrate on the very first unit of the first module “Data and data detectives with CODAP” and on the introduction of our students into data science under the frame of the PPDAC cycle. We will share our design ideas and our experiences from the implementation of the course with upper secondary school students who have no previous knowledge of data science or working with multivariate data.

3 | DESIGN IDEAS OF THE UNIT “DATA AND DATA DETECTIVES WITH CODAP”

The idea of the introductory unit “Data and data detectives with CODAP” was that our students are to be introduced to reasoning about data in the frame of a data project. We decided to start with the PPDAC cycle for framing the introductory data project. The JIM study (https://www.mpfs.de/startseite/), a representative survey on media and leisure time activities of 12-19-year-old students in Germany, offers data on the use of classical media (books, journals), on the use of digital media, on the use of social media, on the use of messengers, on the use of YouTube, and the use of games (on a computer, tablet, or console). In total, the JIM data set includes more than 80 categorical questions (eg, “How often do you use Whatsapp?” with possible answers like “daily”, “several times a week”, “once a week”, “two times a month”, “once a month”, “less often”, “never”) on the leisure time activities and the media use of German students. A report with aggregate and summary data from this study is published every year, but the raw data are not available for re-analysis. This is why we used a
modified version of the JIM study questionnaire for collecting our data in the schools, the students of our course belonged to. We wanted our students to explore this kind of microdata on their own and in addition to statistical and computational aspects; we wanted to introduce our students to media education.

All in all, 215 students of a total of about 1800 students of the two schools in Paderborn took part in our online survey (we call it JIM-PB data in the following). We regard the JIM-PB survey data as a suitable example of real and meaningful data for secondary students in Paderborn despite the low response rate and sample bias (just 23% of the respondents were female). Of course, a data set with 215 cases like this JIM-PB data set might not be considered “Big Data” as such. However, the JIM-PB data set contains responses on over 80 different variables. Therefore, it serves as a complex and exciting data set to develop the data handling skills of our students. For our students that was their first encounter with a multivariate data set in their school time. Moreover, even creating frequency distributions and analyzing the statistical relationship between two variables is not part of the regular mathematics curriculum. Further, the PPDAC cycle is not known to the students. We did not want to teach these aspects abstractly, but with a sense-making context, where the students can explore meaningful questions of their own to get the first experiences of data analysis project work.

As a data analysis tool, we decided to use the online data exploration platform CODAP (see https://codap.concord.org—for a detailed description of CODAP features see for example Haldar et al. [11]). The reasons include: CODAP does not require one to learn a programming language; it is very easily and quickly usable by beginners; and it provides even unexperienced learners with a quick start into data exploration. Therefore, it easily enables young learners to explore the JIM-PB data with their own questions and interests. We preferred CODAP over Fathom and TinkerPlots as CODAP is a free web-based tool, while being aware of its limitations and affordances.

Our main design ideas for the first introduction of our students into data science are ...

- ... explore and analyze a multivariate data set via appropriate statistical investigation questions,
- ... use/apply basic terms of descriptive statistics and statistical concepts,
- ... use and evaluate digital tools like CODAP for their data exploration,
- ... document and present their data analysis in an acceptable form.

We allocated three sessions (135 min each) for the unit “Data and data detectives with CODAP”. As is already made transparent in the learning goals, the phases of the PPDAC cycle were considered the framework of the JIM-PB project (see Figure 1), and we want our students to experience all five phases of the investigative cycle: problem, plan, data, analysis, and conclusions. One of our fundamental design ideas was that our participants are enabled to work on the JIM-PB data project on their own. For this reason, they need a first introduction in reasoning with data. We want our students to concentrate on the statistical content rather than on procedures associated with using a digital tool. We knew that students struggle when posing good statistical questions [1], when exploring relationships between categorical variables [13], when comparing distributions [4], or when synthesizing their findings in a report or a presentation [10]. We, therefore, prepared prompts (see blue boxes in Figure 1) to support our students in their data exploration process in the JIM-PB project.

Corresponding to the five phases of the PPDAC cycle, we see in Figure 1, Table 1 displays the content of the three sessions of the unit.

4 IMPLEMENTATION OF THE UNIT “DATA DETECTIVES WITH CODAP”

The 14 participants were secondary school students (17-18 years old) in grade 12 who had taken an advanced course in computer science and who had very little prior knowledge of descriptive statistics. In session 1, our students were introduced to the project, the basics of descriptive statistics, and the survey. We provided them with four different topic areas of the JIM-PB study for their project work: (a) using information media, (b) using online services, such as messengers, (c) using YouTube, and (d) playing games, playing games on the computer, and playing games on the tablet. The students formed groups of two to work collaboratively, chose one of the topics and worked on the topic for the whole duration of the course.

We will focus on the first two sessions and thus on two aspects in our description of the course in this paper.
First, we discuss how to develop learners competence in generating statistical questions and second on how to explore data such as JIM-PB data using different types of percentages.

As we know from empirical studies like Arnold [1] or Frischemeier and Biehler [8], learners face challenging problems when generating statistical questions. Significant problems are that inexperienced learners pose questions that can be answered by yes/no or by a single value, and therefore, do not allow a decent and in-depth data exploration. Further typical problems are that learners often tend to pose questions that cannot be answered with the given data, or which concentrate on the distributions of one variable but do require an investigation of relationships between variables. So, one central design idea was to design a collaborative setting like think-pair-share in which the students can develop and improve their statistical questions collaboratively and in peer processes [9].

In the first phase of session 2 (think phase), our students were asked to generate an initial statistical question for the exploration of the JIM-PB data set. In this first phase (think phase), our students worked in pairs. They were enabled to communicate with each other about an initial statistical question and to discuss the quality of their initial statistical question. In the second phase, the pair phase, two pairs (pair 1 and pair 2) of students came together. Pair 1 reviewed the statistical question of pair 2 and pair 2 reviewed the statistical question of pair 1. So in concrete terms, pair 1 gave feedback to pair 2, and pair 2 provided feedback for pair 1. After this, both pairs are given some time to revise their statistical questions taking into account the received feedback. Since we expected our students to face difficulties in the feedback process, we provided our students with a kind of checklist (see Table 2), which can be found in Frischemeier and Leavy [9], and which was developed taking into account the work of Arnold [1], Frischemeier and Biehler [8], and Frischemeier and Leavy [9]. The checklist was distributed to our students at the beginning of phase 2 (pair phase).

Then in a final step, the statistical questions of all pairs were discussed with the instructor in a whole class session, and the instructor provided expert feedback on all statistical questions after the pair phase. After this discussion, each pair finalized their statistical questions.

Let us have a look at the development of two exemplary statistical questions in the course. The pair Paul and Sven (both pseudonyms) had the initial question “Do male or female students use Instagram more?” in the thinking phase. This question is a typical statistical question that is at this stage not well-formed because it is answerable with yes or no. In the process of the pair and
share feedback, this question was elaborated into “In which ways do male and female students differ in their Instagram use?” This question shows a higher quality because it enables learners to make more in-depth explorations and to work out differences on several levels between male and female students about their Instagram use.

Another instructive example is provided by the statistical questions of the pair Mirco and Nicole (both pseudonyms). Their initial question was “Which grade shows the highest reading frequency?” This kind of question points to an answer, which consists of a single value (the grade). In the feedback pair and share process, this question was revised to a more open and complex question “In which ways do the students differ in their reading habits across the grades?”, preparing for a more in-depth and more sophisticated data exploration. So all in all, we can say that the quality of the statistical questions of the pairs increased considerably in the think-pair-share process.

As already mentioned above, the JIM-PB data were collected in several schools in Paderborn. In total, 215 students of approx. 1800 students took part in the online survey. Data handling was done by the teachers in advance: variables were defined, a little data cleaning was necessary (eg, correction of German umlauts). We imported the data into CODAP and then gave our students a short introduction to CODAP. We explained to our students how to (a) display the distribution of a categorical or numerical variable in CODAP, (b) how to explore the relationship between two categorical variables using different types of percentages (row, column, cell), and (c) how to calculate absolute and relative

| TABLE 2 | Checklist for improving statistical questions (adapted from Frischemeier and Leavy [9], p. 60) |
|---------|------------------------------------------------------------------------------------------------|
| **Look at the question** | Is it meaningful? Will the question sustain interest and curiosity? Is the intent clear and unambiguous? |
| **Think about the variables of interest** | Is each variable described clearly? |
| **Look at the relationship between the question and the data it will generate** | Can the question be answered with a simple “yes/no” response [avoid these types of questions]? Will the question generate quantitative data (ie, numbers)? Does the question promote group comparison of data? |
| **Look at (or imagine) the data** | Can you answer the question with the given data? Is there sufficient data collected to answer the question? Is there sufficient variability in data collected (is there the potential for a wide range of possible data values)? |

FIGURE 2  Screenshot of comparison visualizations between the variables “eReader” (x-axis) and “Gender” (y-axis) with row percentages in CODAP

FIGURE 3  Screenshot of comparison visualizations between the variables “eReader” (x-axis) and “Gender” (y-axis) with column percentages in CODAP
frequencies and center and spread measures with CODAP. We were aware from studies and research reports such as Batanero et al. [2] and Watson and Callingham [13] that learners face significant problems when exploring relationships between two categorical variables. One crucial question is which kind of percentage to use in which situation? Therefore, we provided our students with supporting help cards for the exploration of the relationship between categorical variables in different situations, which we illustrate in Figures 2-4 (the orange dots in the CODAP displays represent individual respondents). We started with the exploration of the relationship of two categorical variables, which have only two values each: Gender (m/f) and eReader (yes/no). This analysis leads to a 2 × 2 table in CODAP. We used all three kinds of percentages to derive all possible statements from the graph (see Figures 2-4).

Together with the instructor, the students worked out in Figure 2 that 37% of the female students in the sample had an eReader and only 12% of the male students in the sample had an eReader. So there was the conclusion that female students in the JIM-PB sample are more likely to own an eReader than are males. We can quantify this statement by referring to the two percentages. The information (see Figure 3) that 69% of the eReader owners are female and only 31% of the eReader owners are male cannot be used to answer the same question because these numbers can be (and are) distorted by different numbers of males and females in the sample.

In a second example, the instructor asked the students to investigate the question “How do the female and male students differ in their Instagram use?” The CODAP screenshot for this investigation can be seen in Figure 5.

The teacher discussed with the students, which percentages would be appropriate to tackle the question “How do the female and male students differ in their Instagram use?” Concerning Figure 5, the students decided to use row percentages in this case. They determined that 53% of the male students, but 67% of the female students used Instagram daily—this is the first difference between both distributions. The instructor then proposed to consider the distribution more from a global perspective, and therefore, suggested aggregating bins/categories to get a clearer summary. For example, within the class conversation, they aggregated “daily”

---

**Figure 4** Screenshot of comparison visualizations between the variables “eReader” (x-axis) and “Gender” (y-axis) with cell percentages in CODAP [Colour figure can be viewed at wileyonlinelibrary.com]

**Figure 5** Screenshot of comparison visualization between the variables “Using Instagram” and “Gender” with row percentages in CODAP [Colour figure can be viewed at wileyonlinelibrary.com]
and “several times a week” to a “frequent Instagram use” and “less often” and “never” to a “non-frequent Instagram use.” With these redefined broader categories, the instructor together with the students found that 81% of the female students are “frequent Instagram users” compared to 67% of the male students in the JIM-PB sample. In addition to that, the students summarized that 28% of the male students are non-frequent Instagram users, and 17% of the female students are non-frequent Instagram users.

In a more complex challenge (interpretation of a 7×7 table), our students were then asked to describe the pattern usage of Snapchat and Facebook (Figure 6).

The students decided to use this plot to create a two-dimensional display of the data. Together with the instructor, the students distinguished several groups: “Non-frequent Facebook & Non-frequent Snapchat” users (29%) in the lower-left corner, “Non-frequent Facebook, but frequent Snapchat” users (42%) in the lower right corner, and “frequent Facebook and frequent Snapchat” users (14%) in the upper right corner. After this introduction, the instructor showed the students how to compare distributions of numerical variables using boxplots and measures of center and spread. After all these inputs, in session 3, the students were asked to continue their project work and explore their statistical questions on their own. In this working phase, the students also got short inputs about appropriate statistical diagrams and the preparation of PowerPoint presentations. Finally, these PowerPoint presentations were shown and discussed in the classroom as the final stage of the unit “Data and data detectives with CODAP.”

5 | CONCLUSION AND OUTLOOK

From a retrospective view of our project course, we can say that the unit “Data and data detectives with CODAP” can offer a first starting activity into data science education in secondary school. The PPDAC cycle is a helpful framework for conducting a first statistical project at the beginning of the course. However, in the further stages of the project course and the data science curriculum, this cycle will and should be expanded and revised to the CRISP-DM cycle [6], which turned out to be more appropriate, for example, for the conduction of the larger data science projects in module 3. Statistical investigative questions and their quality are a vital concern in the PPDAC cycle, and the Think-Pair-Share setting seemed to support our students when revising and developing their statistical questions. The JIM-PB data set, which with its 215 cases is not to be considered to be big, messy data, served as a valuable data set that offered meaningful and exciting insights for secondary school students and it also provided plenty of different variables to explore. CODAP served as a valuable tool for initial data exploration of the JIM-PB data, with its user-friendly interface. CODAP also facilitated the data analysis process (and the exploration of a large variety of statistical questions) and decreased the cognitive load on the students, who could put their focus instead on data analysis and exploration rather than on the tool use. The introductory unit could well have been part of traditional statistical teaching, which, however, was not part of the students’ curricular experience. Later in the courses, we build on the first unit, but we also contrast later data
science activities to the first unit, for instance, discussing the CRISP-DM as compared to the PPDAC problem-solving cycle. Moreover, the limitations of CODAP motivated our students to progress to a more complex tool. Python and Jupyter notebooks were introduced from a data exploration perspective in unit 2 that was built on the initial unit 1. Specific libraries such as pandas and plotly were recognized as elaborations and improvement of methods learned with CODAP. The JIM-PB data was also used in module 2 for building predictive decision trees [5]. In module 2, students used their understanding of relationships between variables in the data set gained in module 1 as a basis for decision trees (they learned the importance of identifying strong associations between variables, and finding correlations usable for predictions) in the first unit of module 2. The manual building of a decision tree was to be contrasted with automated tree generation. For all this, experiences from unit 1 were necessary.

ACKNOWLEDGEMENTS

The ProDaBi project (www.prodabi.de) was initiated and is funded by the Deutsche Telekom Stiftung. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Daniel Frischemeier https://orcid.org/0000-0002-8957-7660
Rolf Biehler https://orcid.org/0000-0002-9815-1282

REFERENCES

1. P. M. Arnold, Statistical investigative questions—an enquiry into posing and answering investigative questions from existing data, Doctor of Philosophy, Univ of Auckland, 2013, available at https://researchspace.auckland.ac.nz/handle/2292/21305.
2. C. Batanero, A. Estepa, J. D. Godino, and D. R. Green, Intuitive strategies and preconceptions about association in contingency tables, J Res Math Educ 27(2) (1996), 151–169.
3. R. Biehler et al., Technology for enhancing statistical reasoning at the school level, in Third international handbook of mathematics education, M. A. Clements, A. J. Bishop, C. Keitel-Kreidt, J. Kilpatrick, and F. K.-S. Leung, Eds., Springer Science + Business Media, New York, 2013, 643–689.
4. R. Biehler et al., Reasoning about data, in International handbook of research in statistics education, D. Ben-Zvi, K. Makar, and J. Garfield, Eds., Springer International, Cham, 2018, 139–192.
5. R. Biehler et al., Data science education in secondary schools: Teaching and learning decision trees with CODAP and Jupyter Notebooks as an example of integrating machine learning into statistics education, in New Skills in the Changing World of Statistics Education Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE), P. Arnold, Ed., ISI/IASE, Voorborg, The Netherlands, 2020.
6. M. R. Berthold et al., Guide to intelligent data analysis: How to intelligently make sense of real data, Springer Science & Business Media, London, 2010.
7. J. Engel, Statistical literacy for active citizenship: A call for data science education, Stat Educ Res J 16(1) (2017), 44–49.
8. D. Frischemeier and R. Biehler, Stepwise development of statistical literacy and thinking in a statistics course for elementary preservice teachers, in Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education, T. Dooley and G. Gueudet, Eds., DCU Institute of Education and ERME, Dublin, Ireland, 2018, 756–763.
9. D. Frischemeier and A. Leavy, Improving the quality of statistical questions posed for group comparison situations, Teach Stat 42 (2020), 58–65.
10. G. Franzis, An Approach to Report Writing in Statistics Courses, in Paper presented at the IASE/ISI Satellite Conference on Statistics Education, Sydney, New South Wales, Australia, 2005.
11. L. C. Haldar et al., Students making sense of multi-level data, Technol Innovat Stat Educ 11(1) (2018).
12. J. Ridgway, Implications of the data revolution for statistics education, Int Stat Rev 84(3) (2016), 528–549.
13. J. Watson and R. Callingham, Two-way tables: Issues at the heart of statistics and probability for students and teachers, Math Think Learn 16(4) (2014), 254–284.
14. C. J. Wild and M. Pfannkuch, Statistical thinking in empirical enquiry, Int Stat Rev 67(3) (1999), 223–248.

How to cite this article: D. Frischemeier, R. Biehler, S. Podworny, and L. Budde, A first introduction to data science education in secondary schools: Teaching and learning about data exploration with CODAP using survey data, Teaching Statistics 43 (2021), S182–S189. https://doi.org/10.1111/test.12283