dAn-oNo: Learning Environment for Data Journalists Teaching Data Analytics Principles

Christina Stoiber, Štefan Emrich, Sonja Radkohl, Eva Goldgruber, and Wolfgang Aigner

Abstract—Several journalistic abilities are required to derive narratives from data, including the skill to discover and construct compelling stories (data storytelling), employ data-driven techniques to research and analyze information (data literacy), utilize visualization methods effectively (visualization literacy), and approach data with a combination of creativity and critical thinking. Despite their expertise in journalism, journalists often encounter challenges in comprehending and utilizing novel visual representations or understanding data analysis methods. The main objective of the dAn-oNo learning environment is to guide journalists through the data analytics process by removing coding hurdles and allowing them to experiment with and understand the code. The learning environment utilizes a Jupyter notebook with Markdown sections. It incorporates a step-by-step approach, covering various stages of the data analytics workflow, such as data importing, inspection, statistical analysis, and in-depth analysis. The learning environment also includes an automated profiler that translates warnings and information into human-understandable insights. The design and implementation of the dAn-oNo learning environment were informed by a literature review, user research, interviews with Austrian data journalists, a phase of exploring different technical possibilities for the learning environment, and rapid prototyping. We also report on the results of a preliminary study with students. The prototype is accessible here: https://github.com/stemrich/SEVA_DA-Onboarding-Tool

Index Terms—data analytics, jupyter notebook, learning, data-journalism

2023 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVis) | 979-8-3503-3030-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/EduVis60792.2023.00013

1 Introduction

Data journalists face a challenge when producing data-centric stories. They are required to possess a certain level of visualization and data literacy [8,36,40] to analyze and interpret data during their research [2]. Visualization literacy is extracting information from data visualizations and gaining insights effectively, efficiently, and safely [36].

The most common steps in data journalism are: Finding stories, working with data, visualizing and preparing the data story, and collaborating. These do not have to take place in a strict linear sequence but are usually found repeatedly in practice, connected via iterative feedback loops [9,34,44,50,53]. To create data stories, several journalistic skills are necessary, such as finding and creating stories (data storytelling), data-driven strategies for researching and analyzing data (data literacy), methods for visualization (visualization literacy), and generally a creative but critical approach to data [11,16,22–25,54].

Although journalists are highly skilled in their field, they usually have difficulties interpreting and working with new visual representations or understanding data analysis methods [10,45,50]. In addition to visualization literacy, other skills required for cognitive activities should also be considered [12]. Information literacy and data literacy are important to interrogate and critique data collection methods. This can help assess the credibility of sources. They are also used to validate conclusions drawn from visual data analysis by consulting other sources of information. Data literacy can be defined as “a compound competence consisting of some level of competence in statistics, data visualization, and more generic competencies in problem-solving using different data.” [40, p.166]. For journalists, data literacy plays an important role in all stages: from finding the data, evaluating data quality, and interpreting data to present the story in a way that takes into account the needs of readers [32]: “Data literacy is needed at every step of the process: to know what data to collect, understand biases in a data set, perform sensitivity analyses and visualize the results properly” [29, p.217].

The data analytics process requires a solid base in the areas of 1) domain knowledge, 2) mathematical/statistical skills, 3) technical skills, and to a certain degree, 4) data visualization tools [46]. These skills are necessary for individuals to understand the data, analyze it effectively or, in the worst case, come up with incorrect or irrelevant conclusions, which are then communicated to readers of a data story, for example. Hence the major challenge of data analytics onboarding is supporting future data journalists to overcome all of the above obstacles at once — a task that is hard, if not impossible, to automate. Data analytics onboarding methods can be utilized to address this challenge and support data journalists learning how to use data analytics methods.

Therefore, we developed an interactive **Data-Analytics-Onboarding-Notebook (dAn-oNo)** — a learning environment for journalists to understand data analytics methods. The notebook is based on a Jupyter notebook with mark-down sections explaining the data analytics pipeline. The main goal was to guide the user step-by-step through the data analytics process of a data set by (1) removing the hurdle of coding, (2) enabling the user to understand the code and experiment with it, (3) helping novice users to understand the challenges and pitfalls of data analytics, (4) allow intermediate users to increase their skill level, and (5) keep the approach flexible and accessible.

Over the last three years, we collaborated closely with Austrian data journalists, thereby iteratively developing a better understanding of their problems using data analytics methods. Methodologically, we relied on interviews for user research and rapid prototyping to build our dAn-oNo learning environment.

In summary, the main contributions of this work are (1) a **problem characterization and abstraction** (Sec. 3); (2) the **design and implementation** of the **Data-Analytics-Onboarding-Notebook** using a Jupyter notebook for data journalists (Sec. 4); (3) the results of a **preliminary study** with students to gain feedback (Sec. 5) and the discussion of results (Sec. 6).

- *Christina Stoiber and Wolfgang Aigner are with St. Pölten University of Applied Sciences, Austria. E-mail: firstname.lastname@fhstp.ac.at.*
- *Sonja Radkohl and Eva Goldgruber are with FH Joanneum - University of Applied Sciences, Graz, Austria. E-mail: firstname.lastname@fh-joanneum.at*
- *Štefan Emrich is drahtwarenhandlung, Landsiedl Popper OG & datengeschichten e.U., Veliki Borištof, Austria E-mail: stefan.emrich@gmx.net.*

979-8-3503-3030-4/23/$31.00 ©2023 IEEE
DOI 10.1109/EduVis60792.2023.00013

2 Related Work

2.1 Challenges in Teaching Data Analytics

We conducted a literature review to characterize problems and main challenges in data journalism that are relevant essentials or intertwined with teaching data visualizations and data analysis. In the next part, we summarize the main challenges we could identify in teaching data analytics to data journalists and students in journalism.

Teaching data journalism presents several challenges that need to be addressed to enhance student's learning experiences and improve the practice of data journalism. First, many students are reluctant to engage in mathematics and statistical analysis due to lacking training. This poses a significant challenge for data journalism education as journalists are traditionally trained in writing stories [43] rather than statistics or coding. Furthermore, teachers often have to teach themselves data journalism skills due to limited time and opportunities for training. This places an additional burden on educators and highlights the need for more comprehensive training programs to support the development of data journalism skills [14, 24, 25, 52]. Both students and teachers encounter challenges in understanding the authenticity of data and dealing with ethical and legal considerations related to data protection. Data can be subjective and manipulated, and journalists need to assess the data sources critically and address potential biases [11, 22]. Moreover, students and teachers come from diverse backgrounds, with differing prior knowledge in journalism, statistics, or data science. Bridging these knowledge gaps and fostering interdisciplinary collaboration is essential in data journalism education [22, 54]. Additionally, recognizing the time constraints journalists and data journalists face is important [4]. Allocating sufficient time for data analysis and visualization within the daily editorial routine [28, 38], along with access to relevant software and tools [49], can support data journalism projects.

Promoting data literacy among journalists and students can help overcome the aversion to mathematics and statistical analysis. By providing accessible and engaging resources, educators and students can build confidence and competence in working with data.

2.2 Data Analytics Online Courses & Platforms

Online learning platforms like Coursera offer courses specifically focused on data analytics, such as "Introduction to Data Analytics" [3]. This course provides an overview of data analytics, including key steps in the data analytics process, differentiating between various data roles, describing different data structures and sources, and explaining the data analysis process. It is a beginner-level course that requires basic computer literacy and high school-level math. The course is self-paced and takes approximately 10 hours to complete. Additionally, edX offers online courses on data analysis, providing insights into analyzing large data sets and deriving meaningful information [15]. These courses cover various stages of data analysis, including data set establishment, data preparation, modeling, key findings identification, and report creation. Data analysis encompasses data mining, descriptive and predictive analysis, statistical analysis, business analytics, and big data analytics. The online learning platform Linkedin-Learning also provides different courses in data analytics [37]. Those online courses are inappropriate for our target group having limited time and resources to learn and teach themselves in this regard. Besides online courses, several online sources are available such as “The Beginner's Guide to Statistical Analysis — 5 Steps and Examples” by Scribbr [47]. The selection of high-quality material to learn data analytics methods is challenging due to the number of courses and resources.

2.3 Data Analytics Tools

Oracle and Cloudera have proposed a seven-step “value-chain” approach for extracting value using data analytics: Objective identification, business levers identification, data collecting, data cleaning, data modeling, data science team creating, optimize and repeat [1]. All those steps require a certain level of technical skill in programming or using statistical, data analytics, or business intelligence software (R [17], Python [42], SPSS [48], Tableau [51], and the like) [26]. Here “AI”-support proves to be of great help in speeding up the coding process, but this is only possible if the person programming has sufficient experience to formulate good assignments and can bugfix and tweak the automatically generated code.

Data analytics novices face challenges in mathematics and statistics (e.g., lack of familiarity with the underlying concepts or difficulty in selecting and applying appropriate techniques to different data types). They may also need help interpreting statistical results and drawing meaningful conclusions from data, mainly when dealing with complex or large data sets. Overcoming these challenges requires a solid foundation in mathematics and statistics and hands-on experience working with real-world data, which can only be acquired through exercise and practice. Therefore, we designed and implemented a learning environment to support journalists and journalists-to-be in educating themselves in data analytics while working on data stories.

In the next section, we report the problem characterization and abstraction results based on interviews and a survey for data journalists and trainers. Furthermore, we also present design considerations for designing and developing a learning environment for data analytics.

3 Problem Characterization and Abstraction

We conducted ten interviews and a survey (15 answers) with data journalists and data journalism trainers. This survey aimed to determine how familiar data journalists are with (visual) data analysis and their needs, tasks, and goals regarding onboarding methods.

3.1 Method & Participants

We conducted ten semi-structured expert interviews with data journalists (trainers) in the German-speaking area of Europe (Germany, Austria, Switzerland) between fall 2020 and spring 2021 (gender: m = 6, f = 4) to get an understanding of their approach to data journalism, learn from them how they run courses, and what they focus on as well as tools, data, and methods usually used. We decided to conduct our interviews with data journalism trainers, as both fields – onboarding and training – mainly focus on learning processes. The interview guidelines included the following parts and were designed as open-ended questions: a) Understanding of data journalism, b) training, and focus, c) tools, data, and methods, d) challenges and unsolved problems, e) support for learners, and f) data and visualization literacy in the context of covid19. The interview material was analyzed using an approach of content structuring content analysis according to Kuckartz [31] by two coders. Such an approach allows for identifying underlying content patterns in large-scale qualitative data. Also, we developed a survey for the same target group based on our findings. We received 15 complete answers from fall 2021 till spring 2022 (gender: m = 14, no answer = 1). The online survey was spread through emails and also forwarded by data journalists in contact. The extensive questionnaire included the following parts: a) Familiarity with visualizations and analysis of data, b) Visualizations: familiarity and use, c) Data analysis: familiarity and use, d) Visual analysis of data, e) Onboarding: learning how to read and use visualizations as well as f) General information. The survey included questions with a 5-point Likert scale and open-ended questions. Due to the low response rate, the data could only be analyzed descriptively by frequency tables and diagrams.

Most participants were trained in data journalism and worked as (data) journalists. The backgrounds of our interview and survey participants are very diverse. A majority of survey and interview respondents (14 out of 25) has formal training in journalism. Frequently, they worked in different departments such as online or science, which are already close to data journalism topics, and then taught themselves data work, statistics, and partly visualization work. One firth has a mere computer science or statistics background and learned journalism through a second education or in practice. Also worth mentioning are careers that have already combined several fields during their studies, such as in a social science doctorate in which one participant researched journalism with Python. Younger participants, in particular, already have data journalism skills from their studies or have even studied data journalism.

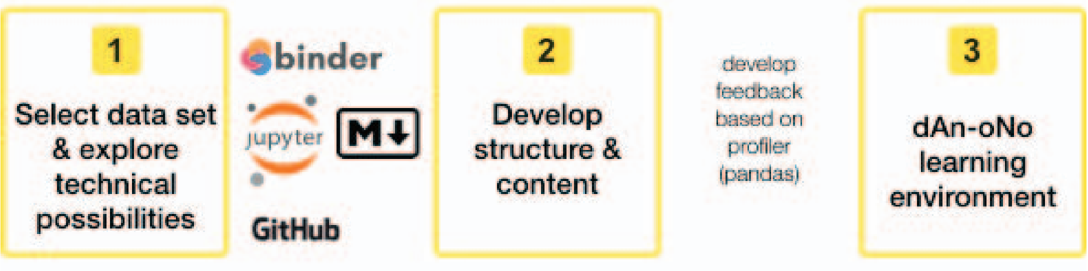

1
Select data set & explore technical possibilities
binder
jupyter
M↓
GitHub
2
Develop structure & content
develop feedback based on profiler (pandas)
3
dAn-oNo learning environment

Fig. 1: Design Process of the dAn-oNo learning environment: (1) **selection of an easy-to-understand data set** & **exploration** and validation of **technical possibilities** for the implementation to overcome the hurdle of coding, support understanding of pitfalls and challenges of data analytics, easy access, and step-wise instructions and feedback. We meet the needs of the journalists by utilizing Jupyter [27] and Markdown [39], binder [6], and GitHub [20], the learning environment. (2) The second step was the **development of the structure and content**, including the following steps: importing data to the notebook, inspecting the data set, statistical analysis, and in-depth analysis. (3) The final step was **implementing** the **dAn-oNo learning environment**, including developing the core component of translating the warnings and information delivered by the automated profiler into human-understandable information.

3.2 Results

We report on the results of the interviews and the survey results. We asked data journalists how familiar they are with visualizations and data analysis. Specific methods were mentioned in data analysis, but the role of data analysis in doing data journalism was also addressed. One interviewee thinks that *"data journalism is a bit out of the statistics hype (...) We now do much more of what we are sure of, where we also have the competencies for it, so actually relatively simple procedures"*. Another one confirmed that *"good data journalism is not about insanely complex statistical analyses but correlations."*

Regarding the data analytics methods' familiarity and use, one interviewee names the simplest analysis methods as *"counting, median, averages, maximum values, accumulation of points"* but also points out that this could be "completely different for each data set and question." The interviews mentioned methods like filtering, sorting, grouping, aggregating, summing, comparing characteristics, percentages, basic arithmetic, simple math, formulas, references, merging tables/ data sets, S-reference, and pivot tables. For the analysis, they use techniques such as descriptive analysis, median, averages, maximum values, minimum values, categories, frequencies, counting, patterns, outliers (plots), position measures, dispersion, histograms, distribution, X vs. Y, percentiles, deviation, rank comparison (Spearman), correlation/relationship (scatter plot, accumulation points), spatial analysis/statistics (maps), etc.

The answers regarding data analysis indicate that the respondents are most familiar with extracting information from data, followed by extracting relations and distributions in data. In comparison, they are unfamiliar with finding new relations and developing or testing hypotheses. Concerning the formats, the respondents are most familiar with describing the data and least familiar with artificial intelligence/machine learning. Formats like comparisons and formats like trends are still quite familiar to most of the respondents. Data analysis methods frequently applied or known by every participant are mean value and median. The results indicate that most of the methods are known, but often they do not use or do not need ones.

Regarding onboarding, participants reported the following: For data journalists themselves, texts, examples, a learning environment, a help center, or tooltips are possible ways. For data journalists themselves, the preferred way is to interact with learning material/concepts actively. Looking at where the onboarding should take place, the answers show that external onboarding or a learning environment is most imaginable for data journalists. Concerning the point of time (when) help is needed, data journalists prefer help during or before they work with data.

Based on the survey results and the literature review of challenges in teaching data analytics, we identified design considerations (DC) for designing and developing a learning environment for data journalists introducing them to data analysis methods.

- **DC1: Support inexperienced users to understand data analytic methods besides mean and median value.** Inexperienced journalists are – at best – familiar with only fundamental concepts of data analysis, such as the difference between mean value and median. Therefore, the learning environment needs to be designed to guide them step-wise through the process of data analytics pipelines, such as important data, visually and descriptively analyzing data, and pointing out challenges and common mistakes.
- **DC2: Provide a learning environment.** The interviews showed that data journalists prefer an external onboarding that can be accessed before or during data analysis. Therefore, we provide journalists with a learning environment. The learning environments aim to sensitize the users to the pitfalls of data analytics and train them to avoid them right before they analyze their own data.

As pointed out, profound data analysis requires high proficiency in mathematics and statistics, coding, and data and domain knowledge. This set of skills cannot be emulated and generalized without the high risk of undetected mistakes due to the users' lack of experience and know-how. On the other hand, it is impossible to "quickly onboard" a person to a subject that requires years of training and experience. To mitigate this risk and enable inexperienced users, we formulated three design considerations (see above) for developing an approach that guides the user step-by-step through the data analytics process of a data set. As it is impossible to generalize this process for any given data set, we added the following two design considerations:

- **DC3: Use easily comprehensible data set.** An easily comprehensible [18,30,35] example data set is to be used for the analysis, which does not require any specialist knowledge but is understood by a broad audience.
- **DC4: Load users own data set.** The learning environment should be designed (technically and content-wise) so that a data set can replace the default data set the user wishes to analyze.

These design considerations emphasize the importance of customization, context, practical examples, and user-centric design in developing effective data analytics onboarding methods for data journalists. The presented dAn-oNo learning environment covers all design considerations. In the next section, we describe the design and implementation of our dAn-oNo learning environment.

4 dAn-oNo: Data-Analytics-Onboarding-Notebook

Based on our design considerations, we followed a rapid prototype method to implement our learning environment dAn-oNo. The design process consists of three steps which we illustrated in Figure 1: (1) select data set & explore technical possibilities, (2) develop structure and content, (3) implement the learning environment using Jupyter [27] and Markdown [39], binder [6] and Github [20] for hosting, and the development of the core component of translating the warnings and information delivered by the automated profiler [41] into human-understandable information.

Table 1: Overview of tasks the user can perform while using the dAn-oNo learning environment

ID	Tasks	Description
T1	Import data set	Import own data set or use the sample data set integrated into the learning environment.
T2	Prepare it for statistical analysis	Inspect the data set, and normalize columns for comparability.
T3	Automated profiling & interpretation	Human-understandable interpretation of automated profiling of data set allows novices to generate better insights.
T4	In-depth analysis	in-depth visual data analysis using pair plots (scatterplots for each data pair). This introduces the user to correlation, causality, and mavericks.

The central didactical concept is active learning [7], which enables learners to actively engage in learning and teaching activities, assume ownership of their learning, and forge connections between ideas through the processes of analysis, synthesis, and evaluation [21]. We know from the literature review and the interview/survey with data journalists that there is an aversion to mathematics and statistical analysis and that most do not have coding skills. Therefore, we used Markdown [39] to reduce the entry barrier for the users. In table 1, we summarize the tasks the user can perform while using the dAn-oNo prototype.

4.1 Data set

To satisfy our design considerations, we needed to find a suitable (data) (DC3) base for our data analytics (DA) onboarding tool (DC4), around which we could design and implement a generic DA process. The interviews and the survey emphasized the need for a data set that is not too complex and does not require expert knowledge. Similarly, when introducing new visualizations, it is recommended to use an easy and understandable data set that can be assumed to be well-known by the target audience [18, 30, 35]. In this particular case, we needed to find a data set that is “understandable” without expert knowledge (e.g., general education is sufficient to understand how the data is to be interpreted) and large enough so that users cannot process it manually but not too big so that it would require highly sophisticated approaches. Furthermore, the data set should be “real”, hence spotting deficiencies frequently encountered when dealing with data.

During our project, we collaborated closely with the University of Applied Sciences FH JOANNEUM Graz (FHJ), which offers a study program, “Journalism and Public Relations”. To develop our DA onboarding tool, we teamed up with a DDJ course at FHJ, where students were to research their data for the course’s assignments. Here we encountered a (close to) perfect data set for our issue. This data set stems from multiple sources, contains information on all Austrian (2,093) municipalities, and sports a lot of the common errors that data sets tend to have, mainly when composed by beginners: ideal for pointing out the pitfalls of data analytics and also introducing and explain ways how to avoid them.

4.2 Explore Technical Possibilities

Next, we describe the data analytics onboarding tool itself. We chose an approach that utilizes code (which is, to some extent, necessary for almost any serious data analysis) of a very broadly used programming language — Python [42] — while making it still easily readable, hence comprehensible for people new to coding and thus reducing the inhibition threshold using the tool. For the implementation, we chose Jupyter [27], which uses Python code, and Markdown [39], a textual documentation that is easy to understand. With this approach, we could allow users to un-comment code or modify it slightly to obtain the results needed. Only minimal modifications were necessary to get through the basic data analytics processes, but it did lower the entry barrier to interact with code. Motivated users could follow their curiosity, experiment with more commands, and gather more insights.

Aiming at beginners and novices to data analytics (i.e., using Windows, without Python or Jupyter installed on their system), it was obvious that the accessibility obstacle needed to be tackled by offering a hosted version of dAn-oNo. For this, we tested and evaluated several options. Our first choice was Jetbrains [13], which we used for our proof-of-concept design. While accessible without the need for payment, these options still had shortcomings. The most important ones: it required user registration (potentially scaring off users), and it did require the code to be stored directly within the system — hence not freely accessible. Thus, we settled for a hybrid approach in which we keep the code at *GitHub* (the complete code of dAn-oNo is available at `https://github.com/stemrich/SEVA_DA-Onboarding-Tool`) and thus separated from the interactive environment. For the interactive environment, we chose *binder* (https://mybinder.org). The reasons for this are manifold. First, storing the code at *GitHub* [20], a major standard for sharing publicly accessible code, makes it widely accessible. Subsequently, many coders, developers, and scientists can build upon our code without changing their standard approaches and/or interfaces. Second, the code can be inspected and downloaded separately from the “interactive environment” by anybody who wants to host the Jupyter Notebook on their premises, which allows them to, for example, use more powerful hardware. Third, by using a binder as the environment where the code is being executed, we have an accessible platform that does not require any registration/login, not to mention a subscription or fee. Fourth, there is a smooth integration of *binder* [6] and *GitHub* [20]. Finally, while the *binder* platform is far from being weak for “average user projects” (offering a minimum of 1GB of RAM and a maximum of 2GB; kernel-session shut-down happens after 10 minutes of inactivity), it is only providing modest computing power. For development, it is often convenient not to have these restrictions (i.e., more computing power; no kernel shut down). Our approach enabled us to develop locally without restrictions and push the code to GitHub with *binder*, automatically accessing the most up-to-date version of the code and providing it (almost instantly) to our users.

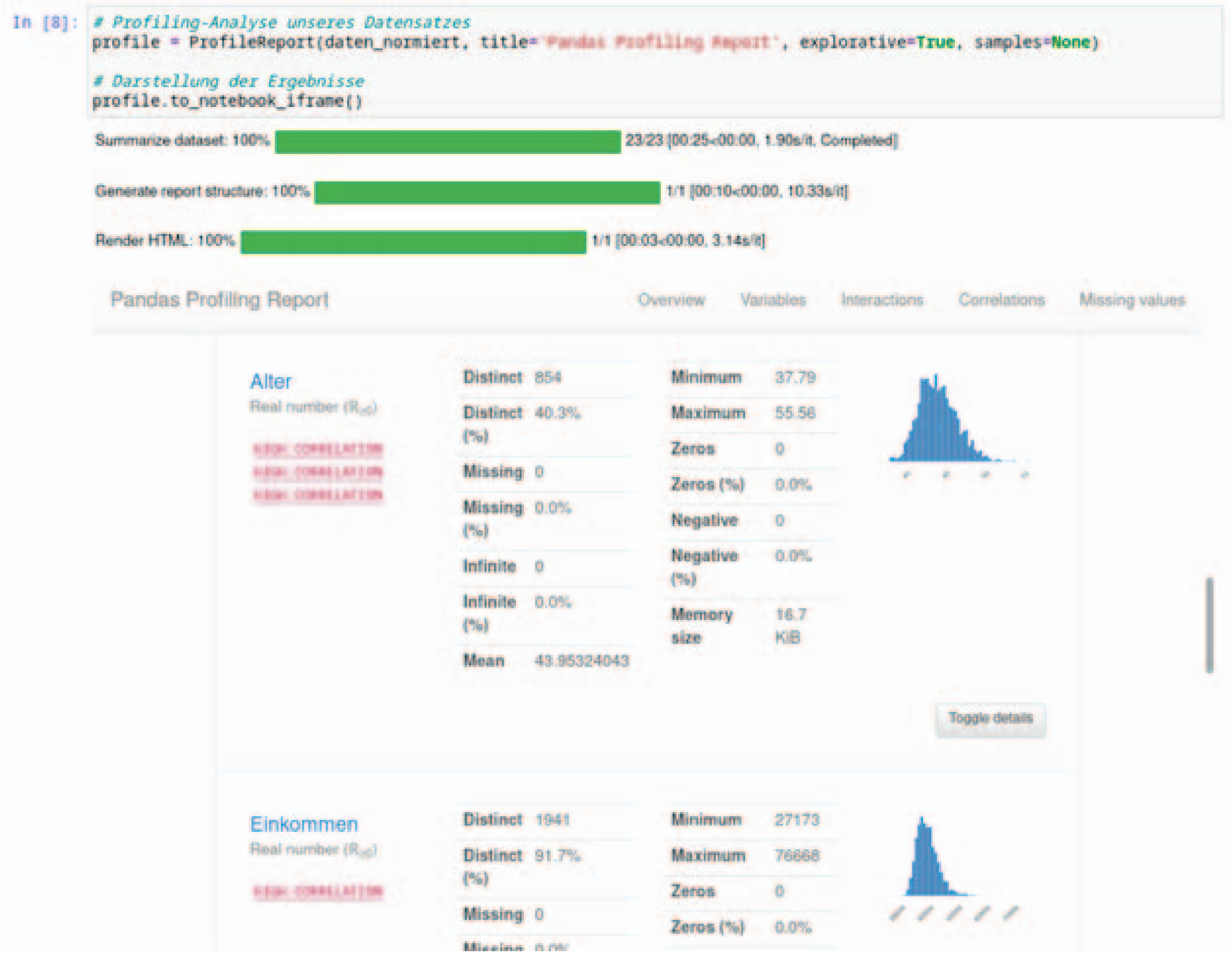
4.3 Develop Structure & Content

For the interactive Data-Analytics-Onboarding-Notebook **dAn-oNo**, we developed a very general data analytics pipeline around our sample data set, which was commented on and explained using the mark-down sections of the Jupyter Notebook. This pipeline started with the import of Python libraries and the data set (also available on GitHub) itself. Next, the data set was inspected visually (printing several lines of the data set to see what data is presented in the columns). Then the data set is prepared for the actual (statistical) analysis. For this, it is necessary to normalize several columns to compare them (e.g., derive the fraction of employees from the absolute number of employees for each municipality).

A major element of this analytics process was to analyze the data set using an automated profiler *ProfileReport* from the library *pandas_profiling*. The challenge was that DDJ beginners and intermediates with little statistical competence were quickly overwhelmed. Firstly by the amount of information and secondly, by the interpretation of fairly technical and statistical feedback, as visible in Figure 2, which shows a screenshot of this profiler report. To tackle this obstacle, we developed the core of dAn-oNo: a human-understandable interpreter for the automated profiler.

This module takes the report of the profiler (the output is a JSON file) and parses through the structure of it collecting all relevant information contained. The function checks whether the JSON data contains information about categorical and numerical values by examining the keys in the ’table’[’types’] section of the JSON data. It then extracts warning messages from the JSON data. It organizes them into different data structures: a set of unique warning messages, a dictionary where warnings are keys and associated columns are values, and another dictionary where columns are keys and associated warnings are values.

This information is processed to generate a textual explanation and interpret the report in HTML format. This allows it to include hyperlinks to additional resources (where users can read up on details and guidelines on dealing with the data set analysis’s warnings and errors),



```
In [8]: # Profiling-Analyse unseres Datensatzes
        profile = ProfileReport(daten_normiert, title='Pandas Profiling Report', explorative=True, samples=None)

        # Darstellung der Ergebnisse
        profile.to_notebook_iframe()
```

```
Summarize dataset: 100%  23/23 [00:25<00:00, 1.90s/it, Completed]

Generate report structure: 100%  1/1 [00:10<00:00, 10.33s/it]

Render HTML: 100%  1/1 [00:03<00:00, 3.14s/it]
```

Alter

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	854	Minimum	37.79
Distinct (%)	40.3%	Maximum	55.56
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	43.95324043	Memory size	16.7 KiB

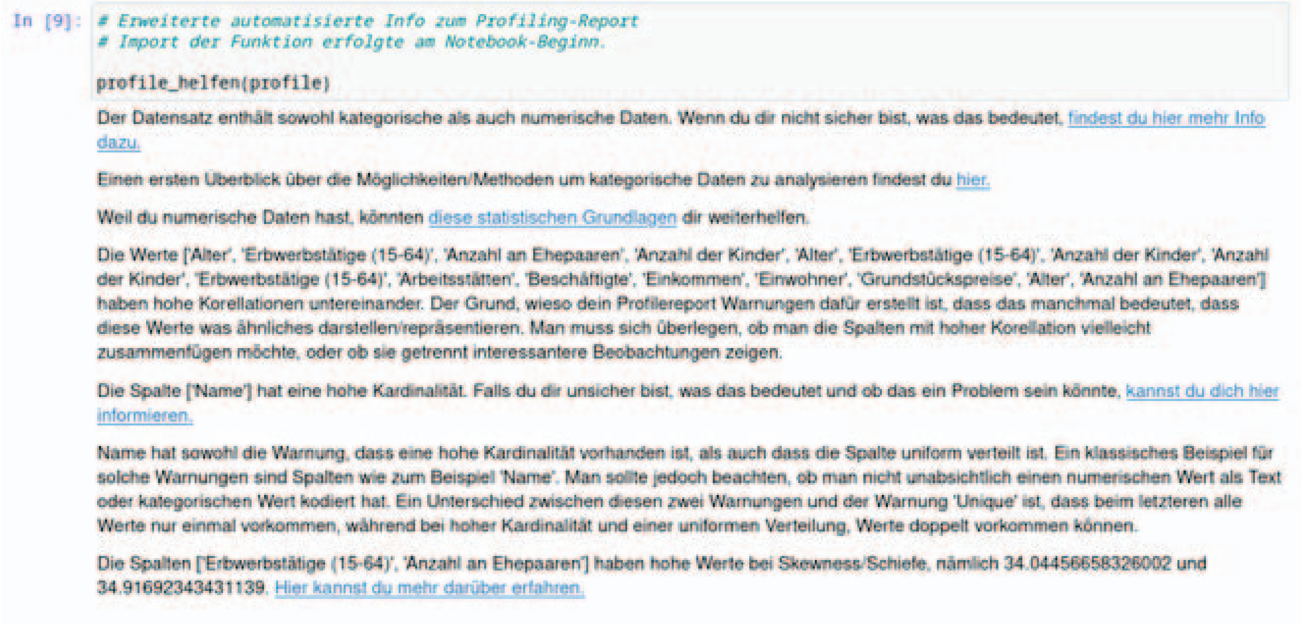
Einkommen

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	1941	Minimum	27173
Distinct (%)	91.7%	Maximum	76668
Missing	0	Zeros	0
Missing	0.0%	Zeros (%)	0.0%

Fig. 2: Automated data set analysis using the Python library pandas_profiling delivering extensive insight, often overwhelming for DA novices



```
In [9]: # Erweiterte automatisierte Info zum Profiling-Report
        # Import der Funktion erfolgte am Notebook-Beginn.

        profile_helfen(profile)
```

Der Datensatz enthält sowohl kategorische als auch numerische Daten. Wenn du dir nicht sicher bist, was das bedeutet, findest du hier mehr Info dazu.

Einen ersten Überblick über die Möglichkeiten/Methoden um kategorische Daten zu analysieren findest du hier.

Weil du numerische Daten hast, könnten diese statistischen Grundlagen dir weiterhelfen.

Die Werte ['Alter', 'Erbwerbstätige (15-64)', 'Anzahl an Ehepaaren', 'Anzahl der Kinder', 'Alter', 'Erbwerbstätige (15-64)', 'Anzahl der Kinder', 'Anzahl der Kinder', 'Erbwerbstätige (15-64)', 'Arbeitsstätten', 'Beschäftigte', 'Einkommen', 'Einwohner', 'Grundstückspreise', 'Alter', 'Anzahl an Ehepaaren'] haben hohe Korellationen untereinander. Der Grund, wieso dein Profilereport Warnungen dafür erstellt ist, dass das manchmal bedeutet, dass diese Werte was ähnliches darstellen/repräsentieren. Man muss sich überlegen, ob man die Spalten mit hoher Korellation vielleicht zusammenfügen möchte, oder ob sie getrennt interessantere Beobachtungen zeigen.

Die Spalte ['Name'] hat eine hohe Kardinalität. Falls du dir unsicher bist, was das bedeutet und ob das ein Problem sein könnte, kannst du dich hier informieren.

Name hat sowohl die Warnung, dass eine hohe Kardinalität vorhanden ist, als auch dass die Spalte uniform verteilt ist. Ein klassisches Beispiel für solche Warnungen sind Spalten wie zum Beispiel 'Name'. Man sollte jedoch beachten, ob man nicht unabsichtlich einen numerischen Wert als Text oder kategorischen Wert kodiert hat. Ein Unterschied zwischen diesen zwei Warnungen und der Warnung 'Unique' ist, dass beim letzteren alle Werte nur einmal vorkommen, während bei hoher Kardinalität und einer uniformen Verteilung, Werte doppelt vorkommen können.

Die Spalten ['Erbwerbstätige (15-64)', 'Anzahl an Ehepaaren'] haben hohe Werte bei Skewness/Schiefe, nämlich 34.04456658326002 und 34.91692343431139. Hier kannst du mehr darüber erfahren.

Fig. 3: Example of translating the warnings and information delivered by the automated profiler into human-understandable information

with the following aspects being covered:

- Categorical and numerical data introduction and links
- Analysis of categorical data
- Analysis of numerical data
- High correlation warnings and explanations
- High cardinality warnings and explanations
- High cardinality and uniform distribution combination
- Missing value warnings and information
- Skewed data warnings and skewness values

This HTML content is then displayed in the notebook, making it easier for people without training in data analytics to understand the more complex aspects of it (see Fig. 3). Subsequently, this core function empowers the notebook user to extract meaningful insights from the data processed and guides them toward better choices.

Following this, in-depth visual data analysis is conducted using pair plots (scatterplots for each data pair). This introduces the user to correlation, causality, and mavericks. A focus is then set on finding explanations for this and stressing the importance of understanding the rules that govern the behavior of the data contained in the data set – which will often require consultation of a domain expert. The DA process identifies and explains the data set's shortcomings, such as using unique IDs as keys when merging data sets, not the municipalities' names (string).

Based on this prototype, we conducted a preliminary study to gain feedback on the notebook with students. The following section presents the study's results, discussion, and future work.

5 Preliminary Study

We conducted a preliminary study to assess the quality of our interactive, code-based DA onboarding prototype with students at St. Pölten University of Applied Sciences.

5.1 Participants

Six students participated in our evaluation (Gender: f=5, m=1). They attended a data journalism class in their master's study program, thus having a similar proficiency level to our leading target group: ongoing data journalists. The participants had no coding skills and used Microsoft Excel or SPSS for statistical data analysis. They were familiar with interpreting data visualizations due to their work experience and classes.

5.2 Study Design & Procedure

We chose a "guided logbook" approach, similar to a diary study [33, p.138], in which we prepared a logbook that they filled out parallel to using the dAn-oNo learning environment. This logbook served two primary purposes. First, the results of all tests were standardized and thus comparable, but at the same time, they gave the users the possibility for individual feedback. We used Microsoft PowerPoint slides to provide the tasks and information to the participants. We prepared empty slides where students could give feedback using any media (text, screenshots/images, video, ...) and answer our questions. Second, we used the logbook to include information, instructions and hints about the learning environment using textual instructions and screenshots, e.g., information on how to find errors in the data, read the scatterplots, and provide lessons learned. We chose this approach to support the participants in finding their way through the evaluation. Furthermore, due to the context and course structure participants had to work independently through the prototype.

In general, the logbook was structured as follows: (a) introduction to the learning environment and the evaluation, (b) presentation of the aim, content, and time to work through it, (c) overview of the tasks, including the data import and descriptive analysis (reading the data) and visual and in-depth analysis of the data (reading between the data). The participants worked in groups, following the logbook. After finishing the tasks and instructions, they submitted their logbooks via our internal Moodle system. We analyzed the students' feedback contained in their slides.

5.3 Results

We analyzed and summarized the feedback, which is presented in this section. Overall, the onboarding tool was (still) too demanding for the student's level of expertise. They reported being overwhelmed by written instructions only. They wished for videos or screenshots to support the instructional design and guide them through the steps. They also struggled with the word "kernel" in explaining the Jupyter Notebook. Additionally, they commented that the pandas profiling reporting could be improved, as the timing of information was not appropriate. They positively commented on the data set used in the learning environment. One group reported spending three hours working through the learning environment due to problems with the kernel. The results of this preliminary study revealed some challenges and issues that need to be improved. In the following, we discuss the results and limitations and provide future directions.

6 Discussion, Limitations, and Future Work

The participants had no programming and coding skills, so following the steps in the learning environment was overwhelming. Providing an introduction to Markdown [39] and the Jupyter [27] environment by using a video tutorial, for example, could be an option to help overcome the coding entry hurdle. Second, additional support in the form of screenshots to aid the instructional design and guide them through the steps could be incorporated into the dAn-oNo learning environment. These improvements may increase the user experience and support them in learning. Third, the instructions of the pandas profiler [41] core component should be improved, mainly regarding the timing of the information provided, as participants commented.

Limitations: It is worth noting that it seemed as if students did not take the evaluation too seriously. They were more assignment-oriented (get over the teaks and answer the logbook) than goal-oriented (find out what stories were in the data) in working through the notebook and giving feedback. In addition, with six participants, the number of participating students in the preliminary study is relatively low.

The learning environment provides a general data analytics pipeline, including importing data, inspecting data sets, statistical analysis, and in-depth data analysis, as we followed the derived design consideration (DC1, DC2) for the learning environment. More advanced methods, such as machine learning or more complex statistical tests, are not yet incorporated. However, the predominant feedback of "feeling overwhelmed" with the assignments suggests that additional topics would increase the overload.

Besides, the participants of our preliminary study performed the tasks in groups due to the constraints of the course in which we conducted our preliminary study. However, we designed the dAn-oNo prototype to be used individually.

Future Work: We plan to incorporate the students' feedback in a future learning environment to improve the instructions using different media, such as videos, animations, or screenshots. Besides, further research is required to evaluate the learning environment's impact on journalists' data literacy. We plan to conduct a user experience evaluation with data journalists and explore in detail: (1) how easy it is for them to use the Markdown language, (2) how we can improve the instructions, and (3) how they would use the learning environment in their daily journalistic workflow. During the user experience evaluation we plan to let participants work individually with the prototype.

Our literature review (see Section 2.1) showed that the curriculum of journalists does not include coding or data analysis methods. They are trained to write stories, not to code or use statistical methods. Therefore, we suggest including classes for data analytics and coding in the curriculum of journalism education to equip journalism students with these skills and prepare them for a data-driven work environment.

A general focus should also be set on finding ways to support students (in general) and data journalists in overcoming their aversion to mathematics and statistical analysis (see Section 2.1), e.g., gamification [19] or data comics [5].

7 Conclusion

This paper presents the design considerations based on a problem characterization and abstraction phase, exploring different technical possibilities for designing an interactive learning environment for data analytics for journalists. Furthermore, we elaborate on the learning environment's implementation, content, and structure. Based on literature reviews and our interviews and survey, journalists need to gain coding skills and have limited experience with data analytics methods besides median and mean (see design considerations presented in Section 3). Therefore, the focus was on developing a learning environment that allows inexperienced journalists to sensitize themselves to the pitfalls in the data analytics workflow using a real-world data set. Furthermore, we designed the learning environment to replace the default data with their data set to provide a blueprint. The resulting technical approach combines a Jupyter notebook with Markdown, hosted using GitHub and the binder environment (see Section 4). We guide the user step-by-step through a basic data analytics workflow ((1) importing data, (2) inspecting data sets, (3) statistical analysis, and (4) in-depth analysis), letting them uncomment and comment on the Markdown code. The core component of the learning environment is the development of an automated profiler [41] that translates the warnings and information delivered into human-understandable information.

Acknowledgments

This work was funded by the BMK under the ICT of the Future program via the SEVA project (no. 874018), and by the Austrian Science Fund as part of the Vis4Schools project (I 5622-N).

References

[1] S. S. Abdul-Jabbar and A. k. Farhan. Data analytics and techniques: A review. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, 10(2):45–55, 2022. doi: 10.14500/aro.10975 2

[2] A. Abellán. State of Data Journalism Survey 2021: 11 surprising findings. https://datajournalism.com/read/blog/data-journalism-survey-2021, 2022. Accessed: 2022-11-14. 1

[3] R. Ahuja. Coursera – Introdcution to Data Analytics. https://www.coursera.org/learn/introduction-to-data-analytics, 2023. Accessed: 2023-06-13. 2

[4] E. Appelgren and G. Nygren. Data Journalism in Sweden: Introducing new methods and genres of journalism into "old" organizations. *Digital Journalism*, 2(3):394–405, 2014. doi: 10.1080/21670811.2014.884344 2

[5] B. Bach, N. H. Riche, S. Carpendale, and H. Pfister. The emerging genre of data comics. *IEEE CG&A*, 37(3):6–13, 2017. doi: 10.1109/MCG.2017.33 7

[6] https://mybinder.org/, 2023. Accessed: 2023-06-20. 3, 4

[7] C. C. Bonwell and J. A. Eison. Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. Technical report, ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183 ($17, 1991. ISBN: 9781878380081 ISSN: 0884-0040 ERIC Number: ED336049. 4

[8] J. Boy, R. A. Rensink, E. Bertini, and J. D. Fekete. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984 1

[9] P. Bradshaw. The inverted piramyd of data journalism - online journalism blog. https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/, 2011. Accessed: 2023-06-13. 1

[10] P. Bradshaw. Reporting beyond the case numbers: How to brainstorm covid-19 data story ideas. https://datajournalism.com/read/longreads/brainstorm-covid-19-data-story-ideas, 2022. Accessed: 2022-10-25. 1

[11] L. S. Burns and B. J. Matthews. First things first: Teaching data journalism as a core skill. *Asia Pacific Media Educator*, 28(1):91–105, 2018. doi: 10.1177/1326365X18765530 1, 2

[12] F. Chevalier, N. H. Riche, B. Alper, C. Plaisant, J. Boy, and N. Elmqvist. Observations and reflections on visualization literacy in elementary school. *IEEE Computer Graphics and Applications*, 38(3):21–29, 2018. doi: 10.1109/MCG.2018.032421650 1

[13] https://datalore.jetbrains.com, 2023. Accessed: 2023-06-20. 4

[14] K. Davies and T. Cullen. Data journalism classes in australian universities: Educators describe progress to date. *Asia Pacific Media Educator*, 26(2):132–147, 2016. Publisher: SAGE Publications India. doi: 10.1177/1326365X16668969 2

[15] edx. edx - Data Analytics Online Course. https://www.edx.org/learn/data-analysis, 2023. Accessed: 2023-06-13. 2

[16] A. Feigenbaum, T. Einar, D. Weissmann, and O. Demirkol. Visualising data stories together: Reflections on data journalism education from the Bournemouth University Datalabs Project. *Journalism Education*, 2(5):59–74, 2016. 1

[17] T. R. Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2023. Accessed: 2023-06-13. 2

[18] E. G. Freedman and P. Shah. Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, and N. H. Narayanan, eds., *Diagrammatic Representation and Inference*, Lecture Notes in Computer Science, pp. 18–30. Springer, 2002. doi: 10.1007/3-540-46037-3_3 3, 4

[19] J. Gäbler, C. Winkler, N. Lengyel, W. Aigner, C. Stoiber, G. Wallner, and S. Kriglstein. Diagram safari: A visualization literacy game for young children. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, CHI PLAY '19 Extended Abstracts, p. 389–396. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3341215.3356283 7

[20] https://github.com/, 2023. Accessed: 2023-06-20. 3, 4

[21] A. Gogus. *Active Learning*, pp. 77–80. Springer US, Boston, MA, 2012. doi: 10.1007/978-1-4419-1428-6_489 4

[22] C. Graham. By the numbers: Data journalism projects as a means of teaching political investigative reporting. *Asia Pacific Media Educator*, 25(2):247–261, 2015. doi: 10.1177/1326365X15604936 1, 2

[23] C. Graham. A diy, project-based approach to teaching data journalism. *Asia Pacific Media Educator*, 28(1):67–77, 2018. doi: 10.1177/1326365X18768308 1

[24] B. R. Heravi. *Teaching Data Journalism*, chap. Data Journalism: Past, Present and Future, p. 8. Abramis Academic Publishing, 2017. 1, 2

[25] J. Hewett. Learning to teach data journalism: Innovation, influence and constraints. *Journalism*, 17(1):119–137, 2016. doi: 10.1177/1464884915612681 1, 2

[26] M. Islam. Data analysis: Types, process, methods, techniques and tools. *Journal of data science*, 6:10, 2020. 2

[27] https://jupyter.org/, 2023. Accessed: 2023-06-20. 3, 4, 6

[28] J. Karlsen and E. Stavelin. Computational journalism in norwegian newsrooms. *Journalism Practice*, 8(1):34–48, 2014. doi: 10.1080/17512786.2013.813190 2

[29] N. Kayser-Bril. Don't ask too much from data literacy. *The Journal of Community Informatics*, 12(3), Aug. 2016. Number: 3. doi: 10.15353/joci.v12i3.3286 1

[30] S. M. Kosslyn. Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3):185–225, 1989. doi: 10.1002/acp.2350030302 3, 4

[31] U. Kuckartz. *Qualitative Text Analysis: A Guide to Methods, Practice & Using Software*. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2014. doi: 10.4135/9781446288719 2

[32] R. Kõuts-Klemm. Data literacy among journalists: A skills-assessment based approach. *Central European Journal of Communication*, 12(3(24)):299–315, Aug. 2019. doi: 10.19195/1899-5101.12.3(24).2 1

[33] J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, Boston, second ed., 2017. 6

[34] B. Lee, N. H. Riche, P. Isenberg, and S. Carpendale. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications*, 35(5):84–90, 2015. doi: 10.1109/MCG.2015.99 1

[35] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-A. Kang, and J. S. Yi. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):499–508, 2016. doi: 10.1109/TVCG.2015.2467195 3, 4

[36] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 1

[37] Linkedin. Linkedin learning. https://www.linkedin.com/

learning/search?keywords=data%20analytics, 2023. Accessed: 2023-06-13. 2

[38] J. Lugo-Ocando and R. F. Brandão. Stabbing news. *Journalism Practice*, 10(6):715–729, 2016. doi: 10.1080/17512786.2015.1058179 2

[39] https://www.markdownguide.org/getting-started/, 2023. Accessed: 2023-06-20. 3, 4, 6

[40] A. Y. Pedersen and F. Caviglia. Data literacy as a compound competence. In T. Antipova and A. Rocha, eds., *Digital Science*, pp. 166–173. Springer International Publishing, Cham, 2019. 1

[41] https://pypi.org/project/pandas-profiling/, 2023. Accessed: 2023-06-20. 4, 6, 7

[42] Python. https://www.python.org/, 2023. Accessed: 2023-06-13. 2, 4

[43] S. Reilly. The need to help journalists with data and information visualization. *IEEE Computer Graphics and Applications*, 37(2):8–10, 2017. doi: 10.1109/MCG.2017.32 2

[44] N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-Driven Storytelling*. A K Peters/CRC Press, Boca Raton, 3 2018. doi: 10.1201/9781315281575 1

[45] J. C. Robert Kosara, Sarah Cohen and M. Wattenberg. Panel: Changing the world with visualization. https://kosara.net/papers/2009/Kosara-InfoVisPanel-2009.pdf, 2009. 1

[46] A. Roy, S. Ray, and R. T. Goswami. Approaches and challenges of big data analytics—study of a beginner. In J. K. Mandal, S. C. Satapathy, M. K. Sanyal, and V. Bhateja, eds., *Proceedings of the First International Conference on Intelligent Computing and Communication*, pp. 237–245. Springer Singapore, Singapore, 2017. 1

[47] Scribbr. The Beginner's Guide to Statistical Analysis | 5 Steps & Examples. https://www.scribbr.com/category/statistics/, 2023. Accessed: 2023-06-13. 2

[48] SPSS Statistics. https://www.ibm.com/de-de/products/spss-statistics, 2023. Accessed: 2023-06-13. 2

[49] F. Stalph. *Datenjournalismus: Eine Dekonstruktion aus feldtheoretischer und techniksoziologischer Perspektive*. PhD thesis, Universität Passau, 2020. 2

[50] C. Stoiber, A. Rind, F. Grassinger, R. Gutounig, E. Goldgruber, M. Sedlmair, S. Emrich, and W. Aigner. netflower: Dynamic network visualization for data journalists. *Comput. Graph. Forum*, 38(3):699–711, 2019. doi: 10.1111/cgf.13721 1

[51] https://www.tableau.com/products/desktop, 2023. Accessed: 2023-01-18. 2

[52] G. Treadwell, T. Ross, A. Lee, and J. K. Lowenstein. A numbers game: Two case studies in teaching data journalism. *Journalism & Mass Communication Educator*, 71(3):297–308, 2016. Publisher: SAGE Publications Sage CA: Los Angeles, CA. 2

[53] T. I. Uskali and H. Kuutti. Models and Streams of Data Journalism. *The Journal of Media Innovations*, 2(1):77–88, Mar. 2015. Number: 1. doi: 10.5617/jmi.v2i1.882 1

[54] L. Zhu and Y. R. Du. Interdisciplinary learning in journalism: A hong kong study of data journalism education. *Asia Pacific Media Educator*, 28(1):16–37, 2018. doi: 10.1177/1326365X18780417 1, 2