European Proceedings of
Educational Sciences
EpES

www.europeanproceedings.com

e-ISSN: 2672-815X

DOI: 10.15405/epes.22043.4

ECQEMC 2021
The Fourth Annual International Symposium "Education and City: Quality Education for Modern Cities"

REGIONAL AND URBAN DATA SCIENCE PROJECTS FOR CITIZEN AND YOUTH ENGAGEMENT

Andrey A. Deryabin (a)*, Pavel P. Glukhov (b)
*Corresponding author

(a) Federal Institute for Educational Development, Russian Presidential Academy of National Economy and Public Administration, 9, Chernyakhovskogo st., bld. 1, Moscow, Russia; Department of Sociology and Mass Communications, Novosibirsk State Technical University, 20 Marx Ave., Novosibirsk, Russia, deryabin-aa@ranepa.ru
(b) Federal Institute for Educational Development, Russian Presidential Academy of National Economy and Public Administration, 9 Chernyakhovskogo st., bld. 1, Moscow, Russia; Moscow City University (MCU), 4 Vtoroy Selskohoziajstvenny proezd, Moscow, 129226, Russian Federation. gluhovpav.pav@gmail.com

Abstract

This article considers the psychological and pedagogical implications of data analysis practices for students' decisions about their personal and professional future and the life of local communities. Using the example of a 10-day "Data Campus" Data Analysis Bootcamp, it explores how students without specific mathematical training can learn to perform data research in the CRISP-DM cycle. The method of research was an ascertaining experiment involving 600 students at the age between 14 and 18 years. The research showed an emotional and activity-based engagement of students in the projects with regional and urban data relevant to their interests, real-life situations, or self-determination tasks the participants face. While reproductive activities related to technical programming skills are relatively easy for students to master, productive solutions to research tasks can be somewhat difficult due to a lack of experience in analyzing intersubject relationships and representing the object under research in a multidimensional feature space. In terms of implementing educational programs that teach data analysis methods in social sciences and humanities in a project approach, students should be provided with appropriate methodological assistance in the form of project mentoring, giving them the opportunity to conceptualize the project before they start working with data.

Keywords: Citizen engagement, data literacy, data science, education, Urban data

https://doi.org/10.15405/epes.22043.4
Corresponding Author: Andrey A. Deryabin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-815X

1. Introduction

1.1. Data as a tool for innovation

Regional and urban data refers to data related to infrastructural, economic, environmental, social, and other aspects of regional and urban life (Law & Legewie, 2021). Typically, such data is generated by the activities of people living in the city. These data relate to various topics such as traffic environment, safety, energy consumption, air and water pollution as well as the social and human practices of urban communities: their interests, social connection structures, patterns of behavior, etc. Urban data can be collected through a variety of sensors, smart meters, satellite images, security cameras or cell phones, or derived from surveys of residents.

A data-literate individual is able to understand external data as well as his/her own, and is sensitive to issues of privacy and ethical use of data. In addition to the data literacy skills possessed by the data consumer, the data literate creator also possesses a number of specialized competencies. These include knowing how to find and use open data, how to generate and use data such as data from sensors, and how to integrate data into various products such as websites and mobile apps (Celino et al., 2012).

The use and generation of urban data by citizens is particularly important for urban innovation. Citizens can drive change from the bottom up, they better understand local problems and offer solutions that are more responsive to their needs (Schelings & Elsen, 2019; Tanenbaum et al., 2013). The transition from top-down to bottom-up data-driven innovation shifts citizens from being passive consumers of technology and providers of data to active participants who consciously generate, provide, and research data both to identify problems that can be solved with data (Fotopoulou, 2020) and to innovate urban development (Wolff et al., 2015).

1.2. Psychological and pedagogical implications of data analysis

We consider urban data and, in particular, geodata as a tool for placing a student's Self and Identity, and relevant topics, problems, and prospects on a map of the city, and, in a broader perspective, on a map of his/her "possible future", including projecting possible trajectories and conditions for his/her social and geographical mobility. The answer to the question of the prospects and capitalization of an individual's personal opportunities is impossible without an analysis of the trends in the development or degradation of certain territories and areas of activity, knowledge of which can be provided by data research.

The mastering and instrumental use of this knowledge is possible when the student sees his/her own intentions, propensities, and prospects for implementing personal opportunities expressed in relevant data and projected onto a map of the city, region, country, or world, and their individual layers - economic, educational, and many others, all of which describe his/her present, future, or probable habitat.

"Data Campus" is an ongoing educational program that since 2019 organizes student project activities in data analysis, machine learning, and modeling of socioeconomic, sociocultural, and engineering objects. A notable part of Data Campus student projects is urban, regional and country data

https://doi.org/10.15405/epes.22043.4
Corresponding Author: Andrey A. Deryabin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-815X

analysis. In a sense, such data projects can be considered as a platform for students to model their personal, educational and career futures. Its main objectives are:

- to ensure that participants learn the basics of data analysis and machine learning and have a professional tryout in this field;
- to provide participants with the basics of modern spatial-analytical thinking as a tool for designing their own future and the future of the territory under study;
- to develop rational, data-driven attitudes among participants toward global processes that affect their settlement area and the entire country;
- to actualize self-determination among participants in relation to various professional and social groups, to modern forms of social, educational and professional mobility.

2. Problem Statement

Although the didactic merits of applying data to learning within not only natural science, but also humanities and social science subjects is obvious to many professionals (Gibson & Mourad, 2018; Glukhov et al., 2021; Locke, 2017), one often encounters objections concerning the fact that students who are not familiar with the relevant parts of mathematics (in particular, with mathematical statistics) are hardly capable of any meaningful practical application of data science methods, especially for researching such complex sociocultural constructs as "city" or "region."

We believe, however, that outside the field of professional education in Data Science, the focus of pedagogical efforts may not be on the technical aspects of working with data, but on the professional trial and development of students' understanding of how data analysis can be useful in solving specific problems of cities and communities. In this case, it would be more correct to talk about students' achievement not of craft perfection in programming for data analysis and professional status, but of data literacy as one of the modern digital competencies and of the data-literate individual as a rational and informed citizen (D'Ignazio, 2017; Fotopoulou, 2020; Pangrazio & Selwyn, 2019).

In terms of the educational challenge of developing an analytical approach for students to explore such complex objects and relate them to their own self-determination, a lack of proficiency in mathematical methods or programming is not an insurmountable obstacle. This part of the data research cycle can be done at the reproductive level. Of much greater interest was students' productivity in formulating a research topic, hypothesis, and selecting data to test it, activities consistent with the initial phases of the standard data research process: understanding goals - initial data investigation - data preparation (Chapman et al., 2000).

Reproductive activities here refer to the activities performed by students using a previously mastered algorithm, and reproductive competencies refer to those that make it possible to understand the conditions for applying that algorithm and the ability to apply it. Reproductive competencies are assessed with the help of tasks that have a solution given in advance. A productive activity refers to an activity in a new situation, where a previously learnt algorithm is not applicable. A productive action in cultural-historical psychology is a two-fold event of overcoming one's a habitual way of acting and presenting the emerging result of action (a product) to others. As such, productive action is an act of one's development (Elkonin, 2019). In the course of productive action, the ability of a person to act in a situation of uncertainty acquires special significance, and the question of whether the subject has or does not have certain productive competencies is actualized. Assessment of productive competences is carried out in the

https://doi.org/10.15405/epes.22043.4
Corresponding Author: Andrey A. Deryabin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-815X

course of solving problematic situations, where tasks are yet to be defined by the actors, which also requires the application of knowledge and certain ways of activity (Glukhov, 2016).

3. Research Questions

The hypothesis of the research was that students at the age of 15-18 are able to implement the cycle of data research in a short time, with some of its phases at the productive level and only on the basis of existing knowledge and thinking skills ("data understanding", "initial data investigation" and "data preparation"), and the part requiring the possession of special mathematical and software tools - at the reproductive one ("modeling", "evaluation", "implementation").

4. Purpose of the Study

The research objective was to test the "Data Campus" educational program on data analysis and machine learning, involving participants in collaborative project activities.

5. Research Methods

The method of the research was an ascertaining experiment in which 600 students of 8-11 grades of schools from Siberian cities, at the age between 14 and 18, were trained in the "Data-Campus" bootcamp between June and November 2020. The participants underwent 8 days of training for the basics of Data Science and Machine Learning with lectures and master classes, in addition, they carried out a team data project. The students were asked to set their own research objective with approximately the following instructions: "Each team should formulate the topic of its project, the problem it solves, and define its goals and objectives. The project can be either exploratory or applied. Both suggested data sets and any data from the Internet can be used to implement the project."

Teams presented the result in a form that is standard in the data analysis and machine learning industry - as files in Jupyter Notebooks format containing the developed program code with comments, sample datasets, infographics, results and conclusions, and presented the work results publicly. Collaborative programming was carried out in the Google Colab environment. All required content, data sets, as well as didactic and test materials, were provided via cloud drives and Google Classroom.

Due to the fact that the students carried out research projects, the quality of their mastery of the concepts, categories and data that they used in their work was of considerable importance. Accordingly, the evaluation of the educational result had a three-level structure:

1) the conceptual part is concerned, on the one hand, with the understanding of what data he/she deals with, and on the other hand, with the understanding and interpretation of the concepts used in the subject field; at this level, a student is supposed to understand both and be able explain them;

2) the analytical part concerns the description of the chosen object or process of analysis; a student can apply the concepts, analyze the object with their help;

3) the modeling part is associated with the elaboration of scenarios for the behavior of objects and processes in question; a student easily uses this or that concept in connection with other concepts, can build models on this basis, discuss predictions and scenarios within the conceptual framework.

https://doi.org/10.15405/epes.22043.4
Corresponding Author: Andrey A. Deryabin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-815X

A generalized correspondence of the didactic phases of the project to the CRISP-DM data research methodology is given in Table 1.

Table 1. Project work phases and Examples of projects of participants of "Data-Campus" on regional analytics and data-urbanism

Project work phases and evaluation parameters	Phases of the CRISP-DM data research cycle
Conceptual (mastering the concepts of a particular subject area)	data understanding, initial data investigation
Analytical (multidimensional description of a chosen data-driven subject)	initial data investigation, data preparation
Predictive (interpretation of results, identification of analytical models, forecasts and scenarios)	modeling, evaluation, implementation (presentation of results)

Here is an example of one of the projects: "Representation of the region in the federal media". Conceptual phase: students may ask themselves the question "How is their region represented in the news agenda published by news agencies and news websites?" and are introduced to: (1) basic communication models, principles, and metaphors of media functioning - as a "mirror" of reality, a "window," a "filter," a "pointer," a "forum," a "barrier"; (2) the social functions of media (to inform, coordinate, reproduce, entertain, mobilize). The data set can be an array of texts from news agencies or other media.

Analytical phase: students perform standard textual data processing operations (stemming, lemmatization, etc.) and conduct quantitative analysis using frequency methods (counting the frequency of words, phrases, etc.), compare these metrics across geographical regions and time periods (years). This stage is equivalent to Exploratory Data Analysis stage in data research, and it itself can lead to some insights about the objects and processes the data represent.

Modeling phase: in this case, it is possible to apply "topic modeling" technique, which identifies hidden topics from the entire text array – sequences of words that occur together and most frequently in the sample. A necessarily creative objective here is the selection of the optimal number of computer-generated, derived from data yet interpretable topics, their interpretation and naming. The same actions can be performed not only on the regional, but also on the federal data sample for their subsequent comparison. Further, it is possible to analyze the temporal and geographical distribution of the topics identified and their variations in relation to known socio-political, economic, and other events. Finally, it is possible to compare regions by similarity in their news agenda and identify clusters of thematically similar regions, which may be geographically very distant from each other. Visualization of the results is also possible - overlaying the tags of topics on a computer map in the form of layers.

Obviously, the results of such task-activity practices in some respects are different from traditional disciplinary education aimed at reproducing a previously known and didactically “correct” result. In this case, much attention is paid to ensuring that students independently determine the goals and objectives of the project, cultivating their active subject position.

Such organization of project activities facilitates the students' reconstruction of the assumed areas and places of implementation of their interests and intentions. Working with data and machine learning models in this case provides a basis for choosing a preferred space for education and career, a

https://doi.org/10.15405/epes.22043.4
Corresponding Author: Andrey A. Deryabin
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2672-815X

consideration of desirable and rejected places of residence and self-actualization and a critical attitude to them, allowing to make an informed choice.

6. Findings

Examples of student projects related to regional or urban issues are given in Table 2.

Table 2. Examples of "Data Campus" participants' projects on regional analytics and data-urbanism.

Summary	Methods
Analysis of public transport waiting time in Russian cities	Regression
Classification of traffic accidents in Kemerovo region	Clustering
Predicting the quality of life in a region by economic indicators	Regression
Exploring the socio-economic factors affecting cultural consumption in Russian regions	Regression
Investigating the relationships between the Human Development Index and economic indicators of the region	Regression
Classification of architectural styles by photographs	Computer vision

Most of the teams used the datasets that were prepared for them by the organizers. However, a number of teams who formulated topics related to local urban or regional issues found themselves in a situation where they needed to find or create their own datasets. In this regard, the result of one team, which organized an online social media survey on perceived public transportation waiting times in different cities, is interesting, as a result of which this team quickly collected a dataset of 7 attributes and 30 thousand observations. The team made a link between fare and public transportation waiting time and interpreted the relationship between satisfaction with public transportation and access of private carriers to provide services in the city. This example shows how the application of authentic data (Kjelvik & Schultheis, 2019), which participants generate independently, has the greatest educational effect (Wolff et al., 2019) through their "appropriation" of this data. This educational situation allows for productive student activities and keeps students highly motivated.

Students who had a good command of Python programming successfully applied the methods and code examples given to them in Data Science classes to their project, even though their knowledge of statistics and the mathematical foundations of the algorithms behind machine learning models was shallow. This allows us to conclude that reproductive activity in this area was relatively easy for them.

From the experts' perspective, more difficult although manageable in most cases, was the conceptual work on some projects, which required students to define or elaborate concepts, operationalize hypotheses, or tasks in terms of available data, understand inter-subject connections, and represent the object under research in a multidimensional space of attributes.

7. Conclusion

The research showed an emotional and activity-based engagement of students in projects with regional and urban data relevant to the participants' place of living, interests, life situations, or self-determination tasks. In this respect, working with such data has great educational potential.

While reproductive activities related to technical programming skills were relatively easy for students to master without deep knowledge of the mathematical foundations of machine learning algorithms and mathematical statistics, students' productive research activities were somewhat impeded. This was especially noticeable when they performed an analysis of sociocultural objects, which often required from students (i) decomposing research objects into lower-level entities that can be represented in the data, and (ii) actualizing students' knowledge in Social Science, History and Economics. In addition, research activities required students to know the basics of scientific research methodology and the ability to apply them in a new situation.

In terms of implementing educational programs that teach data analysis methods in social sciences and humanities in a project mode, students should be provided with mentoring methodological assistance in the form of pre-project consultations, giving them the opportunity to conceptualize the research project before they start working with data.

Acknowledgments

The article was prepared as a part of the state-assigned research work of the Russian Presidential Academy of National Economy and Public Administration.

References

Celino, I., Contessa, S., Corubolo, M., Dell'Aglio, D., Della Valle, E., Fumeo, S., & Krüger, T. (2012). Linking smart cities datasets with human computation: the case of urbanmatch. In P. Cudré-Mauroux, & J. Heflin (Eds.), *Proceedings of the 11th international conference on The Semantic Web: Vol. Part II. The Semantic Web–ISWC 2012* (pp. 34-49). Springer-Verlag.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS. https://www.the-modeling-agency.com/crisp-dm.pdf

D'Ignazio, C. (2017). Creative data literacy: Bridging the gap between the data-haves and data-have nots. *Information Design Journal, 23*(1), 6-18. https://doi.org/10.1075/idj.23.1.03dig

Elkonin, B. D. (2019). Productive Action. *Cultural-historical psychology, 15*(1), 116-122. https://doi.org/10.17759/chp.2019150112

Fotopoulou, A. (2020). Conceptualising critical data literacies for civil society organisations: agency, care, and social responsibility dilemma. *Information Communication and Society, 24*(2), 1-18. https://doi.org/10.1080/1369118X.2020.1716041

Gibson, P., & Mourad, T. (2018). The growing importance of data literacy in life science education. *American journal of botany, 105*(12). https://doi.org/10.1002/ajb2.1195

Glukhov, P., Deryabin, A., & Popov, A. (2021). Data Literacy as a meta-skill: options for Data Science curriculum implementation. *SHS Web Confernces, Vol. 98, The Third Annual International Symposium "Education and City: Education and Quality of Living in the City".* https://doi.org/10.1051/shsconf/20219805006

Glukhov. P. (2016). Kompetentnostnye ispytaniya kak sovremennaya forma otsenki obrazovatel'nykh dostizhenii. [Competence-based Examinations as a Modern Forms of Educational Achievements Evaluation]. *Philosophy of Education, 4*(67), 99-110. https://doi.org/10.15372/PHE20160410

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: Exploring the potential of using data from scientific research to support student data literacy. *CBE – Life Sciences Education, 18*(2). https://doi.org/10.1187/cbe.18-02-0023

Law, T., & Legewie, J. (2021). Urban Data Science. In R.A. Scott, S.M. Kosslyn and M. Buchmann (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1-12). John Wiley & Sons. https://doi.org/10.1002/9781118900772.etrds0450

Locke, B. T. (2017). Digital Humanities Pedagogy as Essential Liberal Education: A Framework for Curriculum Development. *DHQ: Digital Humanities Quarterly, 11*(3). http://www.digitalhumanities.org/dhq/vol/11/3/000303/000303.html

Pangrazio, L., & Selwyn, N. (2019). "Personal data literacies": A critical literacies approach to enhancing understandings of personal digital data. *New Media and Society, 21*(2), 419-437. https://doi.org/10.1177/1461444818799523

Schelings, C., & Elsen, C. (2019). "Smart" Participation: Confronting Theoretical and Operational Perspectives. *International Journal on Advances in Intelligent Systems, 12*(1-2), 1-13.

Tanenbaum, J. G., Williams, A. M., Desjardins, A., & Tanenbaum, K. (2013). Democratizing technology: pleasure, utility and expressiveness in DIY and maker practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'13: CHI Conference on Human Factors in Computing Systems* (pp. 2603-2612). ACM. https://doi.org/10.1145/2470654.2481360

Wolff, A., Gooch, D., Mir, U., Cavero, J., & Kortuem, G. (2015). Removing barriers for citizen participation to urban innovation. In de Lange M., de Waal M. (Eds.), *The Hackable City*. Springer. https://doi.org/10.1007/978-981-13-2694-3_8

Wolff, A., Wermelinger, M., & Petre, M. (2019). Exploring design principles for data literacy activities to support children's inquiries from complex data. *International Journal of Human Computer Studies, 129, 41-54.* https://doi.org/10.1016/j.ijhcs.2019.03.006