The Kansas Data Science Education Atlas

Srisurya Chandramouli, Supriya Bolla, Dr. Safia Malallah, Dr. Lior Shamir
sschandr@ksu.edu, supriyab@ksu.edu, safia@ksu.edu, lshamir@ksu.edu
Kansas State University, Manhattan

Introduction

Despite the increasing importance of data science and AI in education [4], there is currently no comprehensive, publicly available information on where and how these subjects are taught across Kansas. The availability of such data is crucial for enabling educators, policymakers, and community leaders to make informed decisions when planning new programs or initiatives.

The Kansas Data Science Education Atlas project addresses this critical gap by systematically mapping educational opportunities in data science and AI at the K–12, community college, and university levels. By integrating NCES school data, IPUMS NHGIS population statistics, and manually collected course offerings from university websites, we constructed new datasets and applied data science techniques to analyze geographic and institutional patterns across the state.

Our findings reveal significant disparities, with data science and AI programs concentrated in urban counties like Johnson and Sedgwick, while many rural regions continue to have limited access

Data Collection & Sources

College and K-12 school information came from NCES [1][2], which were used in Dataset 1 and 2. Population information came from IPUMS NHGIS [3], which was used in Dataset 1. All sources are reputable because data is collected by government and university institutions.

Data Processing & Cleaning

For Dataset 1, NCES data was initially downloaded in .xls format, but contained embedded HTML tables requiring additional cleaning. The data was converted to CSV, headers were standardized, and the dataset was loaded into BigQuery for further processing. (Cont.)

Data Processing (cont.)

School types were classified using defined rules based on keywords in the school names (e.g., “elementary”, “middle”, “high”, “virtual”), and colleges were categorized by ICLEVEL (4-year, 2-year, <2-year). If a school fit multiple categories, it was assigned to the higher-level group. Population data from NHGIS were linked at city, county, and ZIP code levels to provide demographic context. For Dataset 2, program and course information was manually collected from university and college websites, then standardized for consistency across institutions.

Data Structure / Features

As seen in Table 1, the datasets capture both the geographic and academic dimensions of Kansas education. Dataset 1 provides features such as school name, district, college name, college type, county, city, ZIP code, and population data at multiple geographic levels. Dataset 2 details academic offerings with features including school name, degree type, department, course code, course name, description, level (undergraduate, graduate, professional), modality (online, in-person, both), and course url. Together, these features enable multi-level analysis of educational opportunities, allowing us to explore how the distribution of schools and population relate to the availability and diversity of academic programs across the state.

Table 1: Dataset Feature Comparison

Feature	Dataset 1 (Schools/Context)	Dataset 2 (Programs/ Courses)
K-12I Identification	School Name	—
Geographic Info	District, County Name, City, Zip	—
Population	ZIP, County, City	—
College Identification	College Name, College Type	School Name
Course Info	—	Course Code, Course Name, Description, Course URL
Delivery Method	—	Modality
Program Info	—	Department, Degree Type

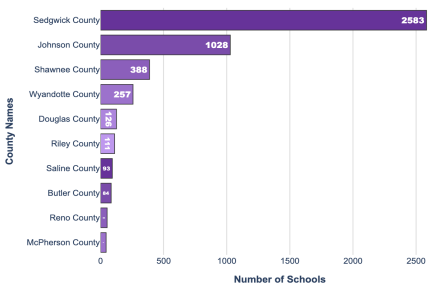
Analysis & Methodology

SQL and BigQuery were used for data joining and querying. Python, VS Code, and regular expressions were used for data cleaning and school type classification. Power BI and Plotly (python pacakge) were used for interactive visualization. Manual review of program and course data ensured accuracy and consistency across institutions. The combined approach enabled detailed geographic and categorical analysis of educational opportunities across Kansas.

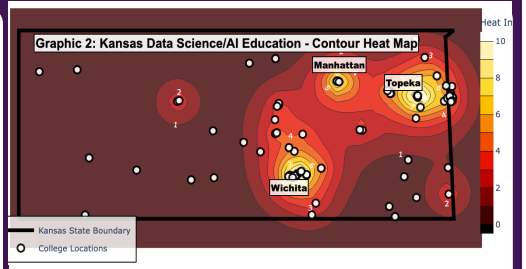
Key Results / Insights

For visualization, aggregated Dataset 1 (school counts) and Dataset 2 (courses, programs, and Impact Score {(1 × Undergrad Programs) + (2 × Graduate Programs) + (0.1 × Undergrad Courses) + (0.2 × Graduate Courses)}) were created. Graphic 1 shows significant geographic imbalance in data science education. Johnson and Sedgwick countieshave the highest concentration of programs, while rural areas lack access.

Graphic 1: Top 10 Counties by Number of Schools



Analysis of Dataset 2 revealed that Data Science and AI programs are highly concentrated in a small number of urban institutions, with most offerings located at large universities in major counties (e.g. K-State, KU, Wichita State). As seen in Graphic 2, most offerings are near the cities of Topeka, Manhattan,



and Wichita. This shows are more rural areas heavily lack Data Science/AI course offerings.

Conclusions & Future Work

This project mapped data science education across Kansas and revealed significant gaps, with programs mainly concentrated in urban areas. There is a clear need to expand opportunities in underserved regions.

Next steps include improving the accuracy of AI/ data science program data, automating data collection with AI tools, and building interactive dashboards and maps to better inform policymakers and educators.

Challenges & Limitations

This project faced several challenges, including mislabeled and messy data files, inconsistent school naming, and scattered course information that was difficult to standardize. Limited features in Power BI also restricted visualization options, and some program details in Dataset 2 were missing, which reduced the depth of analysis.

References

[1] U.S. Department of Education, National Center for Education Statistics, "Search for Public Schools." Common Core of Data (CCD). Accessed: Sep. 30, 2025. [Online]. Available: https://nces.ed.gov/ccd/schoolsearch/school_list.asp?Search=1&State=20&SchoolType=1&SchoolType=2&SchoolType=3&SchoolType=4&SpecificSchlTypes=all&IncGrade=-1&LoGrade=-1&HiGrade=-1&SchoolPageNum=1.
[2] U.S. Department of Education, National Center for Education Statistics, "Complete Data Files." Integrated Postsecondary Education Data System (IPEDS) Data Center. Accessed: Sep. 30, 2025. [Online]. Available: https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?sid=2e642020-4879-43f9-96a1-4f602b6bdf43&rtid=7
[3] IPUMS NHGIS, "Main Page." National Historical Geographic Information System. Accessed: Sep. 30, 2025. [Online]. Available: https://data2.nhgis.org/main.
[4] S. A. Malallah, L. Shamir, W. H. Hsu, J. L. Weese, and S. Alfailakawi, "Data Science (Dataying) for Early Childhood," in *2023 ASEE Annu. Conf. & Expos.*, Baltimore, MD, USA, Jun. 25, 2023, doi: 10.18260/1-2--42867. [Online]. Available: https://peer.asee.org/data-science-dataying-for-early-childhood

This material is based primarily on work funded by the National Science Foundation grant OIA-2148878. Any opinions, findings, conclusions, or recommendations expressed here are those of the author(s) for this project, & do not necessarily reflect the views of NSF.