ATLAS

▼ Summary

Aims to identify and document where data science is being taught across the state covering universities, community colleges, and high schools. After collecting the data, you would apply data science techniques to analyze it, looking for trends, patterns, and gaps. This could include creating visualizations, comparing across regions or institution types, and generating insights that could guide educators, students, and policymakers.

▼ Links

https://www.researchgate.net/

https://www.asee.org/?UserId=d82cd1a9-77c4-4aeb-8407-e77c38a4a309&sso=df7d1bc2-6393-47e6-d09e-08ddeb8a8051

https://guides.library.uwa.edu.au/IEEE/images_tables_figures

▼ Dataset 1

▼ K-12

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. Downloaded initially comes as xls file, but actually contains html code (NCES Data folder → kansas_schools_raw.html)
2. save as html (switch extension from .xls to .html)
3. use python script (convert_html_to_csv.py) → convert html to csv (NCES Data folder → kansas_schools_raw.csv)
4. remove non-table data and edit column names with underscores (edited in VS Code) (NCES Data folder → kansas_schools_clean.csv)
5. upload file into Google BigQuery for data analysis and merging
6. Regex implementation to count number of schools based on these categories: elementary, middle, high, virtual, other
   - elementary: includes grade school, primary school
   - middle: includes junior high
   - high: includes senior high
   - virtual: includes academy and other variants
   - other: not applicable towards first 4 categories
   - for any school names that combine from 2 categories, the school is placed in the upper category (e.g., jr/sr high → high, elem/middle → middle, middle/high → high)
   - regex classification is heavily catered to collected datasets. It is less robust where it might miscategorize unforeseen naming schemes.

▼ K-12 School Category REGEX (SQL)

```
SELECT
  CASE
    -- Virtual (unchanged)
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(virtual|online|academy|ecademy|eacademy|elea|elearn|e-lea)\b')
      THEN 'Virtual'
```

```
    -- Explicit jr + sr (or variants) → High (e.g., "Jr/Sr", "Junior ... Senior")
    WHEN REGEXP_CONTAINS(
        LOWER(school_name),
        r'(\b(jr\.?|junior)\b.*\b(sr\.?|senior)\b)|(\b(sr\.?|senior)\b.*\b(jr\.?|junior)\b)|(\bjr\.?\s*\/\s*sr\.?\b)|(\bjunior\s*
\/\s*senior\b)'
        )
      THEN 'High School'

    -- "High" appears 2+ times (e.g., "Junior High/High School") → High
    WHEN ARRAY_LENGTH(REGEXP_EXTRACT_ALL(LOWER(school_name), r'\bhigh\b')) >= 2
      THEN 'High School'

    -- If name contains BOTH elementary AND junior (but NOT senior) → Middle School
    -- This is the key rule to ensure "Elem/Jr. High" → Middle School.
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(elem|elementary|primary|grade)\b')
      AND REGEXP_CONTAINS(LOWER(school_name), r'\b(junior|jr\.?)\b')
      AND NOT REGEXP_CONTAINS(LOWER(school_name), r'\b(sr\.?|senior)\b')
    THEN 'Middle School'

    -- If "middle" and "high" indicators both appear, treat as High
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\bmiddle\b')
        AND REGEXP_CONTAINS(LOWER(school_name), r'\b(high|hs)\b|senior\s+high|sr\.?\s+high')
    THEN 'High School'

    -- If elementary + high (and NOT the elem+jr special-case above) → High
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(elem|elementary|primary|grade)\b')
        AND REGEXP_CONTAINS(LOWER(school_name), r'\b(high|hs)\b|senior\s+high|sr\.?\s+high')
    THEN 'High School'

    -- "Junior High" or "Jr High" (adjacent) → Middle School
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(junior|jr\.?)\s+high\b')
    THEN 'Middle School'

    -- Plain "middle"
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\bmiddle\b')
    THEN 'Middle School'

    -- Plain "high" (as whole word), or explicit "senior high"
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(high|hs)\b')
        OR REGEXP_CONTAINS(LOWER(school_name), r'senior\s+high|sr\.?\s+high')
    THEN 'High School'

    -- Elementary keywords
    WHEN REGEXP_CONTAINS(LOWER(school_name), r'\b(elem|elementary|primary|grade)\b')
    THEN 'Elementary School'

    ELSE 'Other'
  END AS school_type,
  COUNT(*) AS Num_of_Schools
FROM
  `st-project-454917.kansas_data.nces_kansas_schools`
  GROUP BY school_type
```

```
ORDER BY
  school_type
```

▼ Colleges

1. https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx
2. Download (i.e., HD2023) (Kansas_and_Universities.csv) → Filter results for KS and copy KS rows into another file (Kansas_Colleges_Input.csv)

▼ Census Population

(Note: Used IPUMS NHGIS instead of data.census.gov because having troubles downloading. Some sources lead me to this alternative)

1. https://data2.nhgis.org/main
   - (how to cite: https://www.nhgis.org/citation-and-use-nhgis-data)
2. Filters
   a. Topics → Total Population
   b. Datasets → American Community Survey → 2019_2023_ACS5a 5-Year Data [2019-2023, Block Groups & Larger Areas]
3. Select Geographic Levels → State, County (by State), Place (by State), 5-Digit ZIP Code Tabulation Area
4. Select Geographic Extents → only select Kansas
5. Download
   - nhgis0001_ds267_20235_county.csv, nhgis0001_ds267_20235_place.csv, nhgis0001_ds267_20235_zcta.csv
   - Zip Code file is a national file → USPS Kansas zip code range 66001 - 67954 → used this in SQL query
   - Place file includes cities and CDPs → filter out 'CDP' because they are not legally incorporated municipalities like cities

▼ SQL

```
WITH City_Population AS (
  SELECT
    RTRIM(REPLACE(PLACE, ' city', '')) AS City,
    ASN1E001 AS City_Population
  FROM `st-project-454917.kansas_data.nhgis_kansas_place_pop`
),

-- Schools with their location and population data
Schools_With_Population AS (
  SELECT
    s.School_Name,
    s.District,
    CAST(NULL AS STRING) AS College_Name,
    CAST(NULL AS INT64) AS College_Type,
    s.County_Name,
    copop.ASN1E001 AS County_Population,
    s.City,
    cipop.City_Population,
    CAST(s.ZIP AS STRING) AS ZIP,
    zipop.ASN1E001 AS Zip_Population,
    s.State
  FROM `st-project-454917.kansas_data.nces_kansas_schools` s
```

```
  LEFT JOIN `st-project-454917.kansas_data.nhgis_kansas_county_pop` copop
    ON s.County_Name = copop.COUNTY
  LEFT JOIN City_Population cipop
    ON s.City = cipop.City
  LEFT JOIN `st-project-454917.kansas_data.nhgis_kansas_zip_pop` zipop
    ON CAST(s.ZIP AS STRING) = RIGHT(zipop.NAME_E, 5)
),

-- Colleges with their location and population data
Colleges_With_Population AS (
  SELECT
    CAST(NULL AS STRING) AS School_Name,
    CAST(NULL AS STRING) AS District,
    c.INSTNM AS College_Name,
    c.ICLEVEL AS College_Type,
    c.COUNTYNM AS County_Name,
    copop.ASN1E001 AS County_Population,
    c.CITY AS City,
    cipop.City_Population,
    CAST(NULL AS STRING) AS ZIP,
    CAST(NULL AS INT64) AS Zip_Population,
    'KS' AS State
  FROM `st-project-454917.kansas_data.nces_kansas_colleges` c
  LEFT JOIN `st-project-454917.kansas_data.nhgis_kansas_county_pop` copop
    ON c.COUNTYNM = copop.COUNTY
  LEFT JOIN City_Population cipop
    ON c.CITY = cipop.City
)

-- Combine schools and colleges, replace NULLs with 'DNE'
SELECT
    COALESCE(School_Name, 'DNE') AS School_Name,
    COALESCE(District, 'DNE') AS District,
    COALESCE(College_Name, 'DNE') AS College_Name,
    -- CAST numeric columns to STRING before COALESCE
    COALESCE(College_Type, -1) AS College_Type,
    COALESCE(County_Name, 'DNE') AS County_Name,
    COALESCE(County_Population, -1) AS County_Population,
    COALESCE(City, 'DNE') AS City,
    COALESCE(City_Population, -1) AS City_Population,
    COALESCE(ZIP, 'DNE') AS ZIP,
    COALESCE(Zip_Population, -1) AS Zip_Population,
    COALESCE(State, 'DNE') AS State
FROM (
    SELECT * FROM Schools_With_Population
    UNION ALL
    SELECT * FROM Colleges_With_Population
)
ORDER BY
    County_Name,
    City,
    School_Name,
    College_Name;
```

▼ Dataset 1 - Feature Analysis

▼ 1. School Name

```
SELECT COUNT(DISTINCT School_Name) AS Total_Unique_Schools
FROM st-project-454917.kansas_data.dataset1
WHERE School_Name <> 'DNE'
```

▼ 2. District

```
SELECT District, COUNT(DISTINCT School_Name) AS School_Count
FROM st-project-454917.kansas_data.dataset1
WHERE District <> 'DNE' AND School_Name <> 'DNE'
GROUP BY District ORDER BY School_Count DESC;
```

▼ 3. College Name

```
SELECT COUNT(DISTINCT College_Name) AS Num_of_Colleges
FROM st-project-454917.kansas_data.dataset1
WHERE College_Name <> 'DNE'
```

▼ 4. College Type

```
SELECT
  College_Type,
  COUNT(DISTINCT College_Name) AS Count
FROM `st-project-454917.kansas_data.dataset1`
WHERE College_Name <> 'DNE'
GROUP BY College_Type;
```

▼ 5. County Name

```
SELECT
  County_Name,
  COUNT(
    DISTINCT CASE
      -- Count School_Name only if it is not NULL and not the DNE placeholder
      WHEN School_Name IS NOT NULL AND School_Name <> 'DNE' THEN School_Name
      ELSE NULL
    END
  ) +
  COUNT(
    DISTINCT CASE
      -- Count College_Name only if it is not NULL and not the DNE placeholder
      WHEN College_Name IS NOT NULL AND College_Name <> 'DNE' THEN College_Name
      ELSE NULL
    END
  ) AS Num_of_Institutions
FROM
  `st-project-454917.kansas_data.dataset1`
GROUP BY
  County_Name
```

```
ORDER BY
  Num_of_Institutions DESC;
```

▼ 6. County Population

```
SELECT AVG(County_Population) AS Average_County_Population
FROM
(
 SELECT DISTINCT County_Name, County_Population
 FROM st-project-454917.kansas_data.dataset1
 WHERE County_Population IS NOT NULL AND County_Population <> -1
)
```

▼ 7. City

```
SELECT
  City,
  COUNT(
    DISTINCT CASE
      -- Count School_Name only if it is not NULL and not the DNE placeholder
      WHEN School_Name IS NOT NULL AND School_Name <> 'DNE' THEN School_Name
      ELSE NULL
    END
  ) +
  COUNT(
    DISTINCT CASE
      -- Count College_Name only if it is not NULL and not the DNE placeholder
      WHEN College_Name IS NOT NULL AND College_Name <> 'DNE' THEN College_Name
      ELSE NULL
    END
  ) AS Num_of_Institutions
FROM
  `st-project-454917.kansas_data.dataset1`
GROUP BY
  City
ORDER BY
  Num_of_Institutions DESC;
```

▼ 8. City Population

```
SELECT
  AVG(City_Population) AS Average_City_Population
FROM
(
  SELECT DISTINCT
    City,
    City_Population
  FROM
    `st-project-454917.kansas_data.dataset1`
  WHERE
    City_Population IS NOT NULL
```

```
    AND City_Population <> -1
)
```

▼ 9. ZIP

```
SELECT
  ZIP,
  COUNT(
    DISTINCT CASE
      -- 1. Exclude 'DNE' from School_Name count
      WHEN School_Name IS NOT NULL AND School_Name <> 'DNE' THEN School_Name
      ELSE NULL
    END
  ) +
  COUNT(
    DISTINCT CASE
      -- 2. Exclude 'DNE' from College_Name count
      WHEN College_Name IS NOT NULL AND College_Name <> 'DNE' THEN College_Name
      ELSE NULL
    END
  ) AS Num_of_Institutions
FROM
  `st-project-454917.kansas_data.dataset1`
WHERE
  ZIP IS NOT NULL
  AND ZIP <> 'DNE'  -- 3. Exclude the 'DNE' placeholder from the ZIP column
GROUP BY
  ZIP
ORDER BY
  Num_of_Institutions DESC;
```

▼ 10. ZIP Population

```
SELECT
  AVG(ZIP_Population) AS Average_ZIP_Population
FROM
(
  SELECT DISTINCT
    ZIP,
    ZIP_Population
  FROM
    `st-project-454917.kansas_data.dataset1`
  WHERE
    ZIP_Population IS NOT NULL
    AND ZIP_Population <> -1
)
```

▼ 11. State

Nothing. Every row says ‘KS’

▼ Dataset 2

```
WITH School_Classification AS (
  -- Classify schools by type using REGEX logic from Dataset 1
  SELECT
    County_Name,
    City,
    ZIP,
    County_Population,
    City_Population,
    Zip_Population,
    School_Name,
    CASE
      -- Virtual (highest priority)
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(virtual|online|academy|ecademy|eacademy|elea|elearn|
e-lea)\b')
      THEN 'Virtual'

      -- Explicit jr + sr combinations → High
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'(\b(jr\.?|junior)\b.*\b(sr\.?|senior)\b)|(\b(sr\.?|senior)\b.*\b(j
r\.?|junior)\b)|(\bjr\.?\s*\/\s*sr\.?\b)|(\bjunior\s*\/\s*senior\b)')
      THEN 'High School'

      -- "High" appears 2+ times → High
      WHEN ARRAY_LENGTH(REGEXP_EXTRACT_ALL(LOWER(School_Name), r'\bhigh\b')) >= 2
      THEN 'High School'

      -- Elementary + Junior (but NOT senior) → Middle
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(elem|elementary|primary|grade)\b')
        AND REGEXP_CONTAINS(LOWER(School_Name), r'\b(junior|jr\.?)\b')
        AND NOT REGEXP_CONTAINS(LOWER(School_Name), r'\b(sr\.?|senior)\b')
      THEN 'Middle School'

      -- Middle + High → High
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\bmiddle\b')
        AND REGEXP_CONTAINS(LOWER(School_Name), r'\b(high|hs)\b|senior\s+high|sr\.?\s+high')
      THEN 'High School'

      -- Elementary + High → High
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(elem|elementary|primary|grade)\b')
        AND REGEXP_CONTAINS(LOWER(School_Name), r'\b(high|hs)\b|senior\s+high|sr\.?\s+high')
      THEN 'High School'

      -- Junior High (adjacent) → Middle
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(junior|jr\.?)\s+high\b')
      THEN 'Middle School'

      -- Plain "middle" → Middle
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\bmiddle\b')
      THEN 'Middle School'

      -- Plain "high" or "senior high" → High
      WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(high|hs)\b')
        OR REGEXP_CONTAINS(LOWER(School_Name), r'senior\s+high|sr\.?\s+high')
      THEN 'High School'
```

```
    -- Elementary keywords → Elementary
    WHEN REGEXP_CONTAINS(LOWER(School_Name), r'\b(elem|elementary|primary|grade)\b')
    THEN 'Elementary School'

    ELSE 'Other'
  END AS School_Type
 FROM `st-project-454917.kansas_data.dataset1`
 WHERE School_Name IS NOT NULL  -- Only school rows from Dataset 1
),

-- Get unique (County, City, ZIP) combinations with populations
Unique_Locations AS (
 SELECT DISTINCT
  County_Name,
  City,
  ZIP,
  County_Population,
  City_Population,
  Zip_Population
 FROM `st-project-454917.kansas_data.dataset1`
 WHERE County_Name IS NOT NULL
  AND ZIP IS NOT NULL  -- Only locations that have ZIPs (excludes college-only rows)
),

-- Count schools by type for each (County, City, ZIP)
School_Counts_By_Location AS (
 SELECT
  County_Name,
  City,
  ZIP,
  COUNT(DISTINCT CASE WHEN School_Type = 'Elementary School' THEN School_Name END) AS Elementary_Sch
ool_Count,
  COUNT(DISTINCT CASE WHEN School_Type = 'Middle School' THEN School_Name END) AS Middle_School_Cou
nt,
  COUNT(DISTINCT CASE WHEN School_Type = 'High School' THEN School_Name END) AS High_School_Count,
  COUNT(DISTINCT CASE WHEN School_Type = 'Virtual' THEN School_Name END) AS Virtual_School_Count,
  COUNT(DISTINCT CASE WHEN School_Type = 'Other' THEN School_Name END) AS Other_School_Count
 FROM School_Classification
 GROUP BY County_Name, City, ZIP
),

-- Count colleges by type for each (County, City)
-- Note: Colleges don't have ZIPs, so we count at city level
College_Counts_By_City AS (
 SELECT
  County_Name,
  City,
  COUNT(DISTINCT CASE WHEN College_Type = 1 THEN College_Name END) AS Four_Year_Or_Above_Colleges,
  COUNT(DISTINCT CASE WHEN College_Type = 2 THEN College_Name END) AS Two_Year_Colleges,
  COUNT(DISTINCT CASE WHEN College_Type = 3 THEN College_Name END) AS Less_Than_Two_Year_Colleges
 FROM `st-project-454917.kansas_data.dataset1`
 WHERE College_Name IS NOT NULL  -- Only college rows from Dataset 1
 GROUP BY County_Name, City
```

```
)

-- Final output: One row per (County, City, ZIP) combination
SELECT
  ul.County_Name,
  ul.City,
  ul.ZIP,
  ul.County_Population,
  ul.City_Population,
  ul.Zip_Population,

  -- K-12 school counts (specific to this ZIP)
  COALESCE(scbl.Elementary_School_Count, 0) AS Elementary_Schools,
  COALESCE(scbl.Middle_School_Count, 0) AS Middle_Schools,
  COALESCE(scbl.High_School_Count, 0) AS High_Schools,
  COALESCE(scbl.Virtual_School_Count, 0) AS Virtual_Schools,
  COALESCE(scbl.Other_School_Count, 0) AS Other_Schools,

  -- College counts (same for all ZIPs in the same city)
  COALESCE(ccbc.Four_Year_Or_Above_Colleges, 0) AS Four_Year_Or_Above_Colleges,
  COALESCE(ccbc.Two_Year_Colleges, 0) AS Two_Year_Colleges,
  COALESCE(ccbc.Less_Than_Two_Year_Colleges, 0) AS Less_Than_Two_Year_Colleges

FROM Unique_Locations ul
LEFT JOIN School_Counts_By_Location scbl
  ON ul.County_Name = scbl.County_Name
  AND ul.City = scbl.City
  AND ul.ZIP = scbl.ZIP
LEFT JOIN College_Counts_By_City ccbc
  ON ul.County_Name = ccbc.County_Name
  AND ul.City = ccbc.City

ORDER BY ul.County_Name, ul.City, ul.ZIP;
```

▼ Dataset 3

▼ Collection

1. I collected course info manually (going through individual courses) and using Claude (AI assistant) (dataset2_manual.csv and dataset2_claude.csv).
2. Another round of manual check with ChatGPT's help and combines results from Claude (dataset2.csv)
    - DNE value means 'Does Not Exist' (could not find info)
    - When self-collecting data science/AI courses (dataset 2), I looked for courses that had keywords like AI, Artificial Intelligence, Machine Learning, Data Science, Deep Learning, or Reinforcement Learning in the course name or description

▼ Other Notes

- Professional courses pertain to courses in Continuing Education department or external career training.
-

▼ Dataset 3 - Feature Analysis

▼ 1. School Name

```
-- Enhanced version with percentage
SELECT
  school_name,
  COUNT(*) AS course_count,
  ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER(), 2) AS percentage
FROM `st-project-454917.kansas_data.dataset3`
WHERE school_name <> "school_name"
GROUP BY school_name
ORDER BY course_count DESC;
```

▼ 2. Degree Name

```
-- course count by degree
SELECT
  CASE
    WHEN LOWER(degree_name) LIKE '%associate%' THEN 'Associate'
    WHEN LOWER(degree_name) LIKE '%bachelor%' THEN 'Bachelor'
    WHEN LOWER(degree_name) LIKE '%master%' THEN 'Master'
    ELSE 'Other' -- Certificate, Minor, or Other
  END AS degree_type,
  COUNT(*) AS course_count
FROM
  `st-project-454917.kansas_data.dataset3`
WHERE
  degree_name IS NOT NULL
  AND degree_name != 'DNE'
  AND degree_name != 'degree_name'
GROUP BY
  degree_type
ORDER BY
  course_count DESC;
```

▼ 3. Dept Name

```
SELECT
    dept_name,
    COUNT(*) AS course_count
FROM
    `st-project-454917.kansas_data.dataset3`
WHERE
    dept_name <> 'DNE' AND dept_name <> 'dept_name'
GROUP BY
    dept_name
ORDER BY
    course_count DESC;
```

▼ 4. Course Code

```
-- 4. Course codes
SELECT
```

```
  COUNT(DISTINCT course_code) AS unique_course_codes,
  MIN(course_code) AS min_code,
  MAX(course_code) AS max_code
FROM st-project-454917.kansas_data.dataset3
WHERE course_code IS NOT NULL AND course_code != 'DNE' AND course_code != 'course_code'
```

▼ 5. Course Name

```
-- course frequency
SELECT
  course_name,
  COUNT(*) AS frequency
FROM st-project-454917.kansas_data.dataset3
WHERE course_name IS NOT NULL AND course_name != 'DNE' AND course_name != 'course_name'
GROUP BY course_name
ORDER BY frequency DESC;
```

▼ 6. Description

```
-- 6. Descriptions (length insights)
SELECT
  MIN(LENGTH(description)) AS min_desc_length,
  MAX(LENGTH(description)) AS max_desc_length,
  AVG(LENGTH(description)) AS avg_desc_length
FROM st-project-454917.kansas_data.dataset3
WHERE description IS NOT NULL AND description != 'DNE' AND description != 'description'
```

▼ 7. Level

```
-- 7. Course levels
SELECT
  level,
  COUNT(*) AS course_count,
  ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER(), 2) AS percentage
FROM st-project-454917.kansas_data.dataset3
WHERE level IS NOT NULL AND level NOT IN ('DNE', 'level')
GROUP BY level;
```

▼ 8. Modality

```
-- 8. Modalities (online/in-person/both)
SELECT
  modality,
  COUNT(*) AS course_count,
  ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER(), 2) AS percentage

FROM st-project-454917.kansas_data.dataset3
WHERE modality IS NOT NULL AND modality NOT IN ('DNE', 'modality')
GROUP BY modality
ORDER BY course_count DESC
```

▼ 9. Course URL

```
-- 9. Course URLs (coverage check)
SELECT
  COUNT(DISTINCT course_url) AS unique_urls,
  COUNTIF(course_url IS NULL OR course_url = 'DNE') AS missing_urls
FROM st-project-454917.kansas_data.dataset3
WHERE course_url != 'course_url'
```

▼ Dataset 4

```
-- CTE: Calculate metrics for each school
WITH school_metrics AS (
  SELECT
    school_name,
    COUNT(CASE WHEN course_name IS NOT NULL AND UPPER(TRIM(course_name)) <> 'DNE' THEN 1 END) AS tot
al_courses_school,
    COUNT(DISTINCT CASE WHEN degree_name IS NOT NULL AND UPPER(TRIM(degree_name)) <> 'DNE' THEN de
gree_name END) AS total_programs_school,
    SUM(CASE WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE' AND course_name IS NOT NULL AND UPPER(TRIM
(course_name)) <> 'DNE' THEN 1 ELSE 0 END) AS undergrad_courses_school,
    SUM(CASE WHEN UPPER(TRIM(level)) = 'GRADUATE' AND course_name IS NOT NULL AND UPPER(TRIM(course
_name)) <> 'DNE' THEN 1 ELSE 0 END) AS graduate_courses_school,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE' AND degree_name IS NOT NULL AND
UPPER(TRIM(degree_name)) <> 'DNE' THEN degree_name END) AS undergrad_programs_school,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(level)) = 'GRADUATE' AND degree_name IS NOT NULL AND UPPER
(TRIM(degree_name)) <> 'DNE' THEN degree_name END) AS graduate_programs_school
  FROM
    `st-project-454917.kansas_data.dataset3`
  WHERE
    school_name IS NOT NULL
  GROUP BY
    school_name
)

-- Final query: Join metrics with NCES college location data
SELECT
  sm.school_name,
  kc.COUNTYNM,
  kc.CITY,
  kc.LATITUDE AS latitude,
  kc.LONGITUD AS longitude,
  sm.undergrad_courses_school,
  sm.graduate_courses_school,
  sm.undergrad_programs_school,
  sm.graduate_programs_school,
  (1.0 * sm.undergrad_programs_school) +
  (2.0 * sm.graduate_programs_school) +
  (0.1 * sm.undergrad_courses_school) +
  (0.2 * sm.graduate_courses_school) AS impact_score
FROM
  school_metrics sm
INNER JOIN
  `st-project-454917.kansas_data.nces_kansas_colleges` kc
ON
```

```
  TRIM(LOWER(sm.school_name)) = TRIM(LOWER(kc.INSTNM))
ORDER BY
  impact_score DESC;
```

▼ Dataset 5

▼ FCC - Fixed Broadband

https://broadbandmap.fcc.gov/data-download/nationwide-data

1. Select "Fixed Broadband Summary by Geography Type"
2. Filter by Kansas
3. Download csv

- Data science courses (especially online) require:
  - Stable, high-speed connections for video lectures
  - Large file downloads (datasets, software)
  - Consistent connectivity for real-time collaboration
  - Fixed broadband provides this; mobile is inconsistent and data-capped
- **Household access metric**: Fixed broadband measures home internet access, which is where students take online courses

▼ NHGIS - Computers and Internet

https://data2.nhgis.org/main

1. SELECT DATA → TOPICS → Housing → Computers and Internet
2. DATASETS → 2019_2023_ACS5a
3. Table B28002 - "Presence and Types of Internet Subscriptions in Household" AND Table B28003 - "Presence of a Computer and Type of Internet Subscription"
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ SQL

```
WITH Online_Course_Program_Counts AS (
  SELECT
    school_name,
    SUM(CASE WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
      AND UPPER(TRIM(level)) = 'UNDERGRADUATE'
      AND course_name IS NOT NULL AND UPPER(TRIM(course_name)) <> 'DNE'
      THEN 1 ELSE 0 END) AS online_ugrad_courses,
    SUM(CASE WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
      AND UPPER(TRIM(level)) = 'GRADUATE'
      AND course_name IS NOT NULL AND UPPER(TRIM(course_name)) <> 'DNE'
      THEN 1 ELSE 0 END) AS online_grad_courses,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
      AND UPPER(TRIM(level)) = 'UNDERGRADUATE'
      AND degree_name IS NOT NULL AND UPPER(TRIM(degree_name)) <> 'DNE'
      THEN degree_name END) AS online_ugrad_programs,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
      AND UPPER(TRIM(level)) = 'GRADUATE'
      AND degree_name IS NOT NULL AND UPPER(TRIM(degree_name)) <> 'DNE'
```

```
    THEN degree_name END) AS online_grad_programs
  FROM `st-project-454917.kansas_data.dataset3`
  WHERE school_name IS NOT NULL
  GROUP BY school_name
),

Online_Impact AS (
  SELECT
    school_name,
    (2.0 * online_grad_programs) + (1.0 * online_ugrad_programs) +
    (0.2 * online_grad_courses) + (0.1 * online_ugrad_courses) AS Online_Impact_Score
  FROM Online_Course_Program_Counts
),

County_Online_Density AS (
  SELECT
    kc.COUNTYNM,
    SUM(oi.Online_Impact_Score) AS Total_Online_Impact_Score
  FROM Online_Impact oi
  INNER JOIN `st-project-454917.kansas_data.nces_kansas_colleges` kc
    ON TRIM(LOWER(oi.school_name)) = TRIM(LOWER(kc.INSTNM))
  GROUP BY kc.COUNTYNM
),

-- Get unique counties with population from Dataset 2
All_Kansas_Counties AS (
  SELECT DISTINCT
    County_Name,
    County_Population
  FROM `st-project-454917.kansas_data.dataset2`
  WHERE County_Name IS NOT NULL
),

Digital_Divide_Metrics AS (
  SELECT
    COUNTY AS COUNTY_NAME,
    CAST(COUNTYA AS STRING) AS FIPS_CODE,
    ASWSE001 AS Total_Households,
    ASWSE002 AS Households_With_Internet,
    ASWSE013 AS No_Internet_Households,
    SAFE_DIVIDE(CAST(ASWSE002 AS NUMERIC), ASWSE001) AS Internet_Adoption_Rate,
    SAFE_DIVIDE(CAST(ASWSE013 AS NUMERIC), ASWSE001) AS Pct_No_Internet
  FROM `st-project-454917.kansas_data.nhgis_computer_and_internet`
  WHERE COUNTYA IS NOT NULL
),

FCC_Broadband_Metrics AS (
  SELECT
    t2.geography_desc AS City_Name,
    s.County_Name,
    t2.speed_25_3 AS Broadband_Coverage
  FROM `st-project-454917.kansas_data.fcc_fixed_broadband_summary` t2
  INNER JOIN `st-project-454917.kansas_data.nces_kansas_schools` s
    ON TRIM(t2.geography_desc) = TRIM(s.City)
```

```
  WHERE t2.biz_res = 'R'  -- Residential only
),

County_Broadband_Metrics AS (
  SELECT
    County_Name,
    AVG(Broadband_Coverage) AS Avg_Broadband_Coverage,
    COUNT(DISTINCT City_Name) AS Num_Cities_Measured,
    MAX(Broadband_Coverage) AS Max_Broadband_Coverage
  FROM FCC_Broadband_Metrics
  GROUP BY County_Name
),

Broadband_Access_Index AS (
  SELECT
    ddm.COUNTY_NAME,
    ddm.Total_Households,
    ddm.Internet_Adoption_Rate,
    ddm.Pct_No_Internet,
    fbm.Avg_Broadband_Coverage,
    fbm.Max_Broadband_Coverage,
    fbm.Num_Cities_Measured,
    (
      COALESCE(ddm.Internet_Adoption_Rate, 0) * 0.5 +
      COALESCE(fbm.Avg_Broadband_Coverage, 0) * 0.5
    ) AS Broadband_Access_Index
  FROM Digital_Divide_Metrics ddm
  LEFT JOIN County_Broadband_Metrics fbm
    ON TRIM(REPLACE(ddm.COUNTY_NAME, ' County', '')) = TRIM(REPLACE(fbm.County_Name, ' County', ''))
)

-- FINAL OUTPUT: One row per county (105 rows total)
-- Replace NULLs with -1 for numeric columns to indicate missing data
SELECT
  akc.County_Name,
  akc.County_Population,
  COALESCE(cod.Total_Online_Impact_Score, 0) AS Total_Online_Impact_Score,
  COALESCE(bai.Broadband_Access_Index, -1) AS Broadband_Access_Index,
  COALESCE(ROUND(bai.Internet_Adoption_Rate * 100, 1), -1) AS Internet_Adoption_Pct,
  -- Use -1 for missing FCC data (keeps column numeric)
  COALESCE(
    CASE
      WHEN bai.Num_Cities_Measured IS NULL OR bai.Num_Cities_Measured = 0
      THEN NULL
      ELSE ROUND(bai.Avg_Broadband_Coverage * 100, 1)
    END,
    -1
  ) AS Avg_Broadband_Coverage_Pct,
  COALESCE(ROUND(bai.Pct_No_Internet * 100, 1), -1) AS Pct_No_Internet,
  COALESCE(bai.Total_Households, -1) AS Total_Households,
  COALESCE(cod.Total_Online_Impact_Score, 0) * COALESCE(bai.Broadband_Access_Index, 0) AS Effective_Acc
ess_Score

FROM All_Kansas_Counties akc
```

```
LEFT JOIN County_Online_Density cod
  ON TRIM(akc.County_Name) = TRIM(cod.COUNTYNM)
LEFT JOIN Broadband_Access_Index bai
  ON TRIM(REPLACE(akc.County_Name, ' County', '')) = TRIM(REPLACE(bai.COUNTY_NAME, ' County', ''))

ORDER BY akc.County_Name;
```

- 25 Mbps (download) / 3 Mbps (upload) because it is widely accepted minimum benchmark for broadband → downloading large datasets, videoconferencing (upload; e.g. Zoom, Webex), cloud access (e.g. Google Colab, remote desktop)

▼ Database 5 - Feature Analysis

▼ 1. County Name

```
SELECT
  COUNT(DISTINCT County_Name) as Total_Counties,
  MIN(County_Name) as First_Alphabetically,
  MAX(County_Name) as Last_Alphabetically
FROM `st-project-454917.kansas_data.dataset5`;
```

▼ 2. County Population

```
-- Top 10 and Bottom 10 by population
(SELECT County_Name, County_Population, 'Top 10' as Category
 FROM `st-project-454917.kansas_data.dataset5`
 ORDER BY County_Population DESC
 LIMIT 10)
UNION ALL
(SELECT County_Name, County_Population, 'Bottom 10' as Category
 FROM `st-project-454917.kansas_data.dataset5`
 ORDER BY County_Population ASC
 LIMIT 10)
ORDER BY County_Population DESC;
```

▼ 3. Total Online Impact Score

```
-- How many people have NO local programs?
SELECT
  SUM(CASE WHEN Total_Online_Impact_Score = 0 THEN County_Population ELSE 0 END) as Pop_Without_Progr
ams,
  SUM(CASE WHEN Total_Online_Impact_Score > 0 THEN County_Population ELSE 0 END) as Pop_With_Program
s,
  ROUND(100.0 * SUM(CASE WHEN Total_Online_Impact_Score = 0 THEN County_Population ELSE 0 END) /
        SUM(County_Population), 1) as Pct_Without_Programs
FROM `st-project-454917.kansas_data.dataset5`;
```

▼ 4. Broadband Access Index

```
-- Summary stats for broadband
SELECT
  ROUND(MIN(Broadband_Access_Index), 3) as Min_Index,
  ROUND(AVG(Broadband_Access_Index), 3) as Avg_Index,
  ROUND(APPROX_QUANTILES(Broadband_Access_Index, 100)[OFFSET(50)], 3) as Median_Index,
```

```
  ROUND(MAX(Broadband_Access_Index), 3) as Max_Index,
  ROUND(STDDEV(Broadband_Access_Index), 3) as Std_Dev,
FROM `st-project-454917.kansas_data.dataset5`;
```

▼ 5. Internet Adoption Percentage

```
SELECT
  CASE
    WHEN Internet_Adoption_Pct >= 90 THEN '90%+'
    WHEN Internet_Adoption_Pct >= 80 THEN '80-89%'
    WHEN Internet_Adoption_Pct >= 70 THEN '70-79%'
    WHEN Internet_Adoption_Pct >= 60 THEN '60-69%'
    ELSE 'Below 60%'
  END AS Adoption_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Internet_Adoption_Pct), 1) AS Avg_Adoption,
  ROUND(MIN(Internet_Adoption_Pct), 1) AS Min_Adoption,
  ROUND(MAX(Internet_Adoption_Pct), 1) AS Max_Adoption
FROM `st-project-454917.kansas_data.dataset5`
GROUP BY Adoption_Range
ORDER BY Min_Adoption DESC;
```

▼ 6. Average Broadband Coverage Percentage

```
SELECT
  CASE
    WHEN Avg_Broadband_Coverage_Pct >= 80 THEN '80%+'
    WHEN Avg_Broadband_Coverage_Pct >= 70 THEN '70-79%'
    WHEN Avg_Broadband_Coverage_Pct >= 60 THEN '60-69%'
    WHEN Avg_Broadband_Coverage_Pct IS NULL THEN 'No Data'
    ELSE 'Below 60%'
  END AS Coverage_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Avg_Broadband_Coverage_Pct), 1) AS Avg_Coverage,
  ROUND(MIN(Avg_Broadband_Coverage_Pct), 1) AS Min_Coverage,
  ROUND(MAX(Avg_Broadband_Coverage_Pct), 1) AS Max_Coverage
FROM `st-project-454917.kansas_data.dataset5`
GROUP BY Coverage_Range
ORDER BY Min_Coverage DESC;
```

▼ 7. No Internet Percentage

```
SELECT
  CASE
    WHEN Pct_No_Internet >= 20 THEN '20%+ (Critical)'
    WHEN Pct_No_Internet >= 15 THEN '15-19% (High)'
    WHEN Pct_No_Internet >= 10 THEN '10-14% (Moderate)'
    WHEN Pct_No_Internet >= 5 THEN '5-9% (Low)'
    ELSE 'Below 5% (Minimal)'
  END AS No_Internet_Range,
  COUNT(*) AS County_Count,
FROM `st-project-454917.kansas_data.dataset5`
GROUP BY No_Internet_Range
```

```
ORDER BY
  CASE
    WHEN No_Internet_Range = '20%+ (Critical)' THEN 1
    WHEN No_Internet_Range = '15-19% (High)' THEN 2
    WHEN No_Internet_Range = '10-14% (Moderate)' THEN 3
    WHEN No_Internet_Range = '5-9% (Low)' THEN 4
    ELSE 5
  END;
```

▼ 8. Total Households

```
SELECT
  CASE
    WHEN Total_Households >= 100000 THEN '100K+ (Urban)'
    WHEN Total_Households >= 50000 THEN '50K-99K (Suburban)'
    WHEN Total_Households >= 10000 THEN '10K-49K (Small City)'
    WHEN Total_Households >= 5000 THEN '5K-9K (Town)'
    WHEN Total_Households >= 1000 THEN '1K-4K (Small Town)'
    ELSE 'Below 1K (Rural)'
  END AS Household_Range,
  COUNT(*) AS County_Count,
  SUM(Total_Households) AS Total_Households_Sum,
FROM `st-project-454917.kansas_data.dataset5`
GROUP BY Household_Range
ORDER BY MIN(Total_Households) DESC;
```

▼ 9. Effective Access Score

```
SELECT
  CASE
    WHEN Effective_Access_Score > 3 THEN 'High Impact (>3)'
    WHEN Effective_Access_Score > 1.5 THEN 'Moderate Impact (1.5-3)'
    WHEN Effective_Access_Score > 0.5 THEN 'Low Impact (0.5-1.5)'
    WHEN Effective_Access_Score > 0 THEN 'Minimal Impact (>0)'
    ELSE 'No Local Programs (0)'
  END AS Impact_Range,
  COUNT(*) AS County_Count,
FROM `st-project-454917.kansas_data.dataset5`
GROUP BY Impact_Range
ORDER BY MIN(Effective_Access_Score) DESC;
```

▼ Dataset 6

▼ NHGIS - Median Household Income

1. SELECT DATA → TOPICS → POPULATION → Household and Family Income
2. DATASETS → 2019_2023_ACS5a
3. Table B19013 - “Median Household Income in the Past 12 Months (in 2023 Inflation-Adjusted Dollars)”
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Employment Status

1. SELECT DATA → TOPICS → POPULATION → Labor Force and Employment Status

2. DATASETS → 2019_2023_ACS5a
3. Table B23025 - "Employment Status for the Population 16 Years and Over"
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Educational Attainment

1. SELECT DATA → TOPICS → POPULATION → Education Attainment
2. DATASETS → 2019_2023_ACS5a
3. Table B15003 - "Educational Attainment for the Population 25 Years and Over"
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Poverty

1. SELECT DATA → TOPICS → POPULATION → Poverty (Income Relative to Poverty Level)
2. DATASETS → 2019_2023_ACS5a
3. Table B17021 - Poverty Status of Individuals in the Past 12 Months by Living Arrangement
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Young Adult Education Levels

1. SELECT DATA → TOPICS → POPULATION → Educational Attainment
2. DATASETS → 2019_2023_ACS5a
3. Table B15002 - Sex by Educational Attainment for the Population 25 Years and Over
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Occupation and Industry for Employees

1. SELECT DATA → TOPICS → POPULATION → Occupation and Industry
2. DATASETS → 2019_2023_ACS5a
3. Table C24010 - Sex by Occupation for the Civilian Employed Population 16 Years and Over
4. Table C24030 - Sex by Industry for the Civilian Employed Population 16 Years and Over
5. GEOGRAPHIC LEVELS → COUNTY
6. GEOGRAPHIC EXTENTS → Kansas

▼ NHGIS - Computer Access by Income Level

1. SELECT DATA → TOPICS → HOUSING → Computers and Internet
2. DATASETS → 2019_2023_ACS5a
3. Table B28004 - Household Income in the Last 12 Months by Presence and Type of Internet Subscription
4. GEOGRAPHIC LEVELS → COUNTY
5. GEOGRAPHIC EXTENTS → Kansas

▼ SQL

```
WITH All_Course_Program_Counts AS (
  -- 1. Calculate weighted ALL-MODALITY course and program counts per college from raw data
  SELECT
    school_name,
    SUM(CASE WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE' AND course_name IS NOT NULL AND UPPER(TR
IM(course_name)) <> 'DNE' THEN 1 ELSE 0 END) AS ugrad_courses,
    SUM(CASE WHEN UPPER(TRIM(level)) = 'GRADUATE' AND course_name IS NOT NULL AND UPPER(TRIM(cou
rse_name)) <> 'DNE' THEN 1 ELSE 0 END) AS grad_courses,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE' AND degree_name IS NOT NULL A
ND UPPER(TRIM(degree_name)) <> 'DNE' THEN degree_name END) AS ugrad_programs,
    COUNT(DISTINCT CASE WHEN UPPER(TRIM(level)) = 'GRADUATE' AND degree_name IS NOT NULL AND UPP
ER(TRIM(degree_name)) <> 'DNE' THEN degree_name END) AS grad_programs
  FROM `st-project-454917.kansas_data.dataset3`
  WHERE school_name IS NOT NULL
  GROUP BY school_name
),
Total_Impact AS (
  -- 2. Calculate the total weighted Impact Score per college
  SELECT
    school_name,
    ( (2.0 * grad_programs) + (1.0 * ugrad_programs) + (0.2 * grad_courses) + (0.1 * ugrad_courses) ) AS Total_Pro
gram_Impact_Score
  FROM All_Course_Program_Counts
),
County_Program_Density AS (
  -- 3. Aggregate the Total Impact Score to the County level
  SELECT
    kc.COUNTYNM,
    SUM(ti.Total_Program_Impact_Score) AS Total_Program_Impact_Score
  FROM Total_Impact ti
  INNER JOIN `st-project-454917.kansas_data.nces_kansas_colleges` kc
    ON TRIM(LOWER(ti.school_name)) = TRIM(LOWER(kc.INSTNM))
  GROUP BY kc.COUNTYNM
),
Economic_Demographic_Metrics AS (
  -- 4. Extract CONDENSED economic/demographic metrics
  SELECT
    t1.COUNTY AS COUNTY_NAME,
    t1.ASQPE001 AS Median_Household_Income,

    -- (COMBINED - replaces 4 separate degree populations)
    -- Combines Bachelor's, Master's, Professional, Doctorate into single rate
    SAFE_DIVIDE(
      CAST(t2.ASP3E022 AS NUMERIC) +  -- Bachelor's
      CAST(t2.ASP3E023 AS NUMERIC) +  -- Master's
      CAST(t2.ASP3E024 AS NUMERIC) +  -- Professional
      CAST(t2.ASP3E025 AS NUMERIC),   -- Doctorate
      CAST(t4.ASN1E001 AS NUMERIC)    -- Total Population
    ) AS Advanced_Degree_Rate,

    SAFE_DIVIDE(t3.ASSRE005, t3.ASSRE003) AS Unemployment_Rate,

    t4.ASN1E001 AS Total_Population,
```

```
  SAFE_DIVIDE(
    CAST(t5.ASQME002 AS NUMERIC),
    CAST(t5.ASQME001 AS NUMERIC)
  ) AS Poverty_Rate,

  -- Single metric for ages 25-34 with Bachelor's or higher
  SAFE_DIVIDE(
    CAST(t6.ASP2E015 + t6.ASP2E016 + t6.ASP2E017 + t6.ASP2E018 + t6.ASP2E032 + t6.ASP2E033 + t6.ASP2E
034 + t6.ASP2E035 AS NUMERIC),
    CAST(t6.ASP2E001 AS NUMERIC)
  ) AS Young_Adult_Bachelors_Plus_Rate,

  -- Combines Computer/Math, Engineering, Science occupations
    SAFE_DIVIDE(
    CAST(t7.ASSUE008 + t7.ASSUE044 AS NUMERIC) +
    CAST(t7.ASSUE009 + t7.ASSUE045 AS NUMERIC) +
    CAST(t7.ASSUE010 + t7.ASSUE046 AS NUMERIC),
    CAST(t7.ASSUE001 AS NUMERIC)
  ) AS STEM_Employment_Rate,

  -- Combines Professional/Scientific/Technical + Information industries
  SAFE_DIVIDE(
    CAST(t8.ASS5E018 + t8.ASS5E045 AS NUMERIC) +
    CAST(t8.ASS5E013 + t8.ASS5E040 AS NUMERIC),
    CAST(t8.ASS5E001 AS NUMERIC)
  ) AS Professional_Services_Rate,

  -- Percentage of low-income households (<$35K) with broadband
  SAFE_DIVIDE(
    CAST(t9.ASWUE002 + t9.ASWUE006 + t9.ASWUE010 AS NUMERIC),
    CAST(t9.ASWUE004 + t9.ASWUE008 + t9.ASWUE012 AS NUMERIC)
  ) AS Low_Income_Digital_Access_Rate

  FROM `st-project-454917.kansas_data.nhgis_household_family_income` t1
 INNER JOIN `st-project-454917.kansas_data.nhgis_education_attainment` t2 ON t1.GISJOIN = t2.GISJOIN
 INNER JOIN `st-project-454917.kansas_data.nhgis_employment_status` t3 ON t1.GISJOIN = t3.GISJOIN
 INNER JOIN `st-project-454917.kansas_data.nhgis_county` t4 ON t1.COUNTY = t4.COUNTY
 INNER JOIN `st-project-454917.kansas_data.nhgis_poverty` t5 ON t1.GISJOIN = t5.GISJOIN
 INNER JOIN `st-project-454917.kansas_data.nhgis_young_adult_education_levels` t6 ON t1.GISJOIN = t6.GISJOI
N
 INNER JOIN `st-project-454917.kansas_data.nhgis_occupation_for_employees` t7 ON t1.GISJOIN = t7.GISJOIN

INNER JOIN `st-project-454917.kansas_data.nhgis_industry_for_employees` t8 ON t1.GISJOIN = t8.GISJOIN

INNER JOIN `st-project-454917.kansas_data.nhgis_computer_access_by_income_level` t9  ON t1.GISJOIN = t9.GI
SJOIN
)

-- FINAL SELECT: CONDENSED TO 10 CORE FEATURES (plus target)
SELECT
 edm.COUNTY_NAME AS COUNTYNM,

 -- TARGET VARIABLE (will be used to derive Has_Programs in Dataset 7)
```

```
  COALESCE(cpd.Total_Program_Impact_Score, 0.0) AS Total_Program_Impact_Score,

  -- 9 CONDENSED ECONOMIC/DEMOGRAPHIC FEATURES:

  -- Economic Indicators (3 features)
  edm.Median_Household_Income,            -- Wealth measure
  edm.Poverty_Rate,                       -- Economic distress (NEW)
  edm.Unemployment_Rate,                  -- Labor market health

  -- Education Indicators (2 features)
  edm.Advanced_Degree_Rate,               -- Overall education level (COMBINED)
  edm.Young_Adult_Bachelors_Plus_Rate,    -- Student pipeline (NEW)

  -- Employment Type (2 features)
  edm.STEM_Employment_Rate,               -- Tech job demand (NEW)
  edm.Professional_Services_Rate,         -- White-collar economy (NEW)

  -- Digital Access (1 feature)
  edm.Low_Income_Digital_Access_Rate,     -- Digital divide for vulnerable populations (NEW)

  -- Demographics (1 feature)
  edm.Total_Population                    -- Size/urbanicity proxy

FROM Economic_Demographic_Metrics edm
LEFT JOIN County_Program_Density cpd
  ON TRIM(REPLACE(cpd.COUNTYNM, ' County', '')) = TRIM(REPLACE(edm.COUNTY_NAME, ' County', ''))

ORDER BY Total_Program_Impact_Score DESC;
```

▼ Dataset 6 - Feature Analysis

▼ 1. County

```
SELECT
  COUNT(DISTINCT County) AS Total_Counties,
  MIN(County) AS First_Alphabetically,
  MAX(County) AS Last_Alphabetically
FROM `st-project-454917.kansas_data.dataset6`;
```

▼ 2. Total Program Impact Score

```
SELECT
  CASE
    WHEN Total_Program_Impact_Score = 0 THEN 'No Programs (0)'
    WHEN Total_Program_Impact_Score > 0 AND Total_Program_Impact_Score <= 2 THEN 'Very Low (0-2)'
    WHEN Total_Program_Impact_Score > 2 AND Total_Program_Impact_Score <= 5 THEN 'Low (2-5)'
    WHEN Total_Program_Impact_Score > 5 AND Total_Program_Impact_Score <= 8 THEN 'Moderate (5-8)'
    WHEN Total_Program_Impact_Score > 8 THEN 'High (8+)'
  END AS Impact_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Total_Program_Impact_Score), 2) AS Avg_Score,
FROM `st-project-454917.kansas_data.dataset6`
```

```
GROUP BY Impact_Range
ORDER BY MIN(Total_Program_Impact_Score);
```

▼ 3. Median Household Income

```
SELECT
  CASE
    WHEN Median_Household_Income >= 80000 THEN '$80K+ (High)'
    WHEN Median_Household_Income >= 65000 THEN '$65K-$79K (Above Average)'
    WHEN Median_Household_Income >= 55000 THEN '$55K-$64K (Average)'
    WHEN Median_Household_Income >= 50000 THEN '$50K-$54K (Below Average)'
    ELSE 'Below $50K (Low)'
  END AS Income_Bracket,
  COUNT(*) AS County_Count,
  ROUND(AVG(Median_Household_Income), 0) AS Avg_Income,
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Income_Bracket
ORDER BY MIN(Median_Household_Income) DESC;
```

▼ 4. Poverty Rate

```
SELECT
  CASE
    WHEN Poverty_Rate >= 0.20 THEN '20%+ (Critical)'
    WHEN Poverty_Rate >= 0.15 THEN '15-19% (High)'
    WHEN Poverty_Rate >= 0.10 THEN '10-14% (Moderate)'
    WHEN Poverty_Rate >= 0.05 THEN '5-9% (Low)'
    ELSE 'Below 5% (Minimal)'
  END AS Poverty_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Poverty_Rate) * 100, 1) AS Avg_Poverty_Pct,
  ROUND(MIN(Poverty_Rate) * 100, 1) AS Min_Poverty_Pct,
  ROUND(MAX(Poverty_Rate) * 100, 1) AS Max_Poverty_Pct
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Poverty_Range
ORDER BY MIN(Poverty_Rate) DESC;
```

▼ 5. Unemployment Rate

```
SELECT
  CASE
    WHEN Unemployment_Rate >= 0.06 THEN '6%+ (Very High)'
    WHEN Unemployment_Rate >= 0.045 THEN '4.5-5.9% (High)'
    WHEN Unemployment_Rate >= 0.03 THEN '3-4.4% (Moderate)'
    WHEN Unemployment_Rate >= 0.02 THEN '2-2.9% (Low)'
    ELSE 'Below 2% (Very Low)'
  END AS Unemployment_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Unemployment_Rate) * 100, 2) AS Avg_Unemployment_Pct,
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Unemployment_Range
ORDER BY MIN(Unemployment_Rate) DESC;
```

▼ 6. Advanced Degree Rate

```
SELECT
  CASE
    WHEN Advanced_Degree_Rate >= 0.30 THEN '30%+ (Very High)'
    WHEN Advanced_Degree_Rate >= 0.20 THEN '20-29% (High)'
    WHEN Advanced_Degree_Rate >= 0.15 THEN '15-19% (Moderate)'
    WHEN Advanced_Degree_Rate >= 0.10 THEN '10-14% (Low)'
    ELSE 'Below 10% (Very Low)'
  END AS Education_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Advanced_Degree_Rate) * 100, 1) AS Avg_Advanced_Degree_Pct,
  ROUND(MIN(Advanced_Degree_Rate) * 100, 1) AS Min_Advanced_Degree_Pct,
  ROUND(MAX(Advanced_Degree_Rate) * 100, 1) AS Max_Advanced_Degree_Pct
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Education_Range
ORDER BY MIN(Advanced_Degree_Rate) DESC;
```

▼ 7. Young Adult Bachelors Plus Rate

```
SELECT
  CASE
    WHEN Young_Adult_Bachelors_Plus_Rate >= 0.40 THEN '40%+ (Very High)'
    WHEN Young_Adult_Bachelors_Plus_Rate >= 0.30 THEN '30-39% (High)'
    WHEN Young_Adult_Bachelors_Plus_Rate >= 0.25 THEN '25-29% (Moderate)'
    WHEN Young_Adult_Bachelors_Plus_Rate >= 0.20 THEN '20-24% (Low)'
    ELSE 'Below 20% (Very Low)'
  END AS Young_Adult_Education_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Young_Adult_Bachelors_Plus_Rate) * 100, 1) AS Avg_Young_Adult_Pct,
  ROUND(MIN(Young_Adult_Bachelors_Plus_Rate) * 100, 1) AS Min_Young_Adult_Pct,
  ROUND(MAX(Young_Adult_Bachelors_Plus_Rate) * 100, 1) AS Max_Young_Adult_Pct
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Young_Adult_Education_Range
ORDER BY MIN(Young_Adult_Bachelors_Plus_Rate) DESC;
```

▼ 8. STEM Employment Rate

```
SELECT
  CASE
    WHEN STEM_Employment_Rate >= 0.05 THEN '5%+ (Very High)'
    WHEN STEM_Employment_Rate >= 0.03 THEN '3-4.9% (High)'
    WHEN STEM_Employment_Rate >= 0.02 THEN '2-2.9% (Moderate)'
    WHEN STEM_Employment_Rate >= 0.01 THEN '1-1.9% (Low)'
    ELSE 'Below 1% (Very Low)'
  END AS STEM_Employment_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(STEM_Employment_Rate) * 100, 2) AS Avg_STEM_Pct,
  ROUND(MIN(STEM_Employment_Rate) * 100, 2) AS Min_STEM_Pct,
  ROUND(MAX(STEM_Employment_Rate) * 100, 2) AS Max_STEM_Pct
FROM `st-project-454917.kansas_data.dataset6`
```

```
GROUP BY STEM_Employment_Range
ORDER BY MIN(STEM_Employment_Rate) DESC;
```

▼ 9. Professional Services Rate

```
SELECT
  CASE
    WHEN Professional_Services_Rate >= 0.06 THEN '6%+ (Very High)'
    WHEN Professional_Services_Rate >= 0.04 THEN '4-5.9% (High)'
    WHEN Professional_Services_Rate >= 0.03 THEN '3-3.9% (Moderate)'
    WHEN Professional_Services_Rate >= 0.02 THEN '2-2.9% (Low)'
    ELSE 'Below 2% (Very Low)'
  END AS Professional_Services_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Professional_Services_Rate) * 100, 2) AS Avg_Prof_Services_Pct,
  ROUND(MIN(Professional_Services_Rate) * 100, 2) AS Min_Prof_Services_Pct,
  ROUND(MAX(Professional_Services_Rate) * 100, 2) AS Max_Prof_Services_Pct
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Professional_Services_Range
ORDER BY MIN(Professional_Services_Rate) DESC;
```

▼ 10. Low Income Digital Access Rate

```
SELECT
  CASE
    WHEN Low_Income_Digital_Access_Rate >= 1.50 THEN '150%+ (Anomaly)'
    WHEN Low_Income_Digital_Access_Rate >= 1.20 THEN '120-149% (Very High)'
    WHEN Low_Income_Digital_Access_Rate >= 1.00 THEN '100-119% (High)'
    WHEN Low_Income_Digital_Access_Rate >= 0.80 THEN '80-99% (Moderate)'
    ELSE 'Below 80% (Low)'
  END AS Digital_Access_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Low_Income_Digital_Access_Rate) * 100, 1) AS Avg_Digital_Access_Pct,
  ROUND(MIN(Low_Income_Digital_Access_Rate) * 100, 1) AS Min_Digital_Access_Pct,
  ROUND(MAX(Low_Income_Digital_Access_Rate) * 100, 1) AS Max_Digital_Access_Pct
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Digital_Access_Range
ORDER BY MIN(Low_Income_Digital_Access_Rate) DESC;
```

▼ 11. Total Population

```
SELECT
  CASE
    WHEN Total_Population >= 100000 THEN '100K+ (Very Large)'
    WHEN Total_Population >= 50000 THEN '50K-99K (Large)'
    WHEN Total_Population >= 25000 THEN '25K-49K (Medium-Large)'
    WHEN Total_Population >= 10000 THEN '10K-24K (Medium)'
    WHEN Total_Population >= 5000 THEN '5K-9K (Small)'
    ELSE 'Below 5K (Very Small)'
  END AS Population_Range,
  COUNT(*) AS County_Count,
  ROUND(AVG(Total_Population), 0) AS Avg_Population,
  SUM(Total_Population) AS Total_Population_In_Range
```

```
FROM `st-project-454917.kansas_data.dataset6`
GROUP BY Population_Range
ORDER BY MIN(Total_Population) DESC;
```

▼ Dataset 7

- **Note:** Super dataset combining datasets 1-6

▼ TIGER/Line Shapefiles

1. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html
2. 2025
3. FTP Archive
4. COUNTY/
5. tl_2025_us_county.zip
6. to be uploaded into Google BigQuery, dbf converted into csv using anyconv.com (tl_2025_us_county_dbf_to_csv.csv)
7. upload in excel and filter for kansas rows (i.e. STATEFP = 20) (tl_2025_us_county_kansas.xlsx)
8. download excel → copy contents into excel → only preserve these columns (STATEFP, COUNTYFP, GEOID, NAME, NAMELSAD, INTPTLAT, INTPTLON) because BigQuery was getting error loading the rest of the columns (tl_2025_us_county_kansas_v2.csv)
9. Must extract columns INTPTLON and INTPTLAT to get coordinates of every Kansas county centroid

▼ SQL

```
WITH Online_Course_Program_Counts AS (
 -- Calculate online-specific course/program counts from Dataset 3
 SELECT
  school_name,
  SUM(CASE
   WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
    AND UPPER(TRIM(level)) = 'UNDERGRADUATE'
    AND course_name IS NOT NULL
    AND UPPER(TRIM(course_name)) <> 'DNE'
   THEN 1 ELSE 0
  END) AS online_ugrad_courses,
  SUM(CASE
   WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
    AND UPPER(TRIM(level)) = 'GRADUATE'
    AND course_name IS NOT NULL
    AND UPPER(TRIM(course_name)) <> 'DNE'
   THEN 1 ELSE 0
  END) AS online_grad_courses,
  COUNT(DISTINCT CASE
   WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
    AND UPPER(TRIM(level)) = 'UNDERGRADUATE'
    AND degree_name IS NOT NULL
    AND UPPER(TRIM(degree_name)) <> 'DNE'
   THEN degree_name
  END) AS online_ugrad_programs,
  COUNT(DISTINCT CASE
   WHEN UPPER(TRIM(modality)) IN ('ONLINE', 'BOTH')
```

```
      AND UPPER(TRIM(level)) = 'GRADUATE'
      AND degree_name IS NOT NULL
      AND UPPER(TRIM(degree_name)) <> 'DNE'
    THEN degree_name
  END) AS online_grad_programs
 FROM `st-project-454917.kansas_data.dataset3`
 WHERE school_name IS NOT NULL
 GROUP BY school_name
),
All_Course_Program_Counts AS (
 -- Calculate total (all modality) course/program counts from Dataset 3
 SELECT
  school_name,
  SUM(CASE
   WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE'
    AND course_name IS NOT NULL
    AND UPPER(TRIM(course_name)) <> 'DNE'
   THEN 1 ELSE 0
  END) AS total_ugrad_courses,
  SUM(CASE
   WHEN UPPER(TRIM(level)) = 'GRADUATE'
    AND course_name IS NOT NULL
    AND UPPER(TRIM(course_name)) <> 'DNE'
   THEN 1 ELSE 0
  END) AS total_grad_courses,
  COUNT(DISTINCT CASE
   WHEN UPPER(TRIM(level)) = 'UNDERGRADUATE'
    AND degree_name IS NOT NULL
    AND UPPER(TRIM(degree_name)) <> 'DNE'
   THEN degree_name
  END) AS total_ugrad_programs,
  COUNT(DISTINCT CASE
   WHEN UPPER(TRIM(level)) = 'GRADUATE'
    AND degree_name IS NOT NULL
    AND UPPER(TRIM(degree_name)) <> 'DNE'
   THEN degree_name
  END) AS total_grad_programs
 FROM `st-project-454917.kansas_data.dataset3`
 WHERE school_name IS NOT NULL
 GROUP BY school_name
),
College_DS_AI_Indicator AS (
 SELECT
  kc.COUNTYNM,
  kc.INSTNM,
  kc.ICLEVEL,
  CASE
   WHEN acp.school_name IS NOT NULL
    AND ((acp.total_ugrad_programs > 0) OR (acp.total_grad_programs > 0)
        OR (acp.total_ugrad_courses > 0) OR (acp.total_grad_courses > 0))
   THEN 1
   ELSE 0
  END AS Has_DS_AI_Programs
 FROM `st-project-454917.kansas_data.nces_kansas_colleges` kc
```

```
  LEFT JOIN All_Course_Program_Counts acp
    ON TRIM(LOWER(kc.INSTNM)) = TRIM(LOWER(acp.school_name))
),
County_DS_AI_College_Counts AS (
  SELECT
    COUNTYNM,
    SUM(CASE WHEN ICLEVEL = 1 AND Has_DS_AI_Programs = 1 THEN 1 ELSE 0 END) AS Four_Year_Colleges_Wit
h_DS_AI,
    SUM(CASE WHEN ICLEVEL = 2 AND Has_DS_AI_Programs = 1 THEN 1 ELSE 0 END) AS Two_Year_Colleges_Wit
h_DS_AI,
    SUM(CASE WHEN ICLEVEL = 3 AND Has_DS_AI_Programs = 1 THEN 1 ELSE 0 END) AS Less_Than_Two_Year_C
olleges_With_DS_AI
  FROM College_DS_AI_Indicator
  GROUP BY COUNTYNM
),
County_Online_Impact AS (
  -- Aggregate online impact scores to county level
  SELECT
    kc.COUNTYNM,
    SUM(
      (2.0 * ocp.online_grad_programs) +
      (1.0 * ocp.online_ugrad_programs) +
      (0.2 * ocp.online_grad_courses) +
      (0.1 * ocp.online_ugrad_courses)
    ) AS Total_Online_Impact_Score
  FROM Online_Course_Program_Counts ocp
  INNER JOIN `st-project-454917.kansas_data.nces_kansas_colleges` kc
    ON TRIM(LOWER(ocp.school_name)) = TRIM(LOWER(kc.INSTNM))
  GROUP BY kc.COUNTYNM
),
County_Total_Impact AS (
  -- Aggregate total impact scores to county level
  SELECT
    kc.COUNTYNM,
    SUM(
      (2.0 * acp.total_grad_programs) +
      (1.0 * acp.total_ugrad_programs) +
      (0.2 * acp.total_grad_courses) +
      (0.1 * acp.total_ugrad_courses)
    ) AS Total_Program_Impact_Score
  FROM All_Course_Program_Counts acp
  INNER JOIN `st-project-454917.kansas_data.nces_kansas_colleges` kc
    ON TRIM(LOWER(acp.school_name)) = TRIM(LOWER(kc.INSTNM))
  GROUP BY kc.COUNTYNM
),
County_School_Counts AS (
  -- Aggregate K12 school counts from Dataset 2
  SELECT
    County_Name,
    SUM(Elementary_Schools) AS Total_Elementary_Schools,
    SUM(Middle_Schools) AS Total_Middle_Schools,
    SUM(High_Schools) AS Total_High_Schools,
    SUM(Virtual_Schools) AS Total_Virtual_Schools,
    SUM(Other_Schools) AS Total_Other_Schools,
```

```
    SUM(Four_Year_Or_Above_Colleges) AS Total_4_Year_Colleges,
    SUM(Two_Year_Colleges) AS Total_2_Year_Colleges,
    SUM(Less_Than_Two_Year_Colleges) AS Total_Less_Than_2_Year_Colleges
  FROM `st-project-454917.kansas_data.dataset2`
  GROUP BY County_Name
),
Digital_Infrastructure AS (
  -- Extract digital infrastructure metrics from Dataset 5
  SELECT
    County_Name,
    Broadband_Access_Index,
    Internet_Adoption_Pct,
    Avg_Broadband_Coverage_Pct,
    Pct_No_Internet,
    Total_Households,
    Effective_Access_Score
  FROM `st-project-454917.kansas_data.dataset5`
),
Economic_Demographics AS (
  -- Extract economic and demographic data from Dataset 6
  SELECT
    County,
    Total_Program_Impact_Score AS DS6_Total_Impact,
    -- Economic Indicators (3 features)
    Median_Household_Income,
    Poverty_Rate,
    Unemployment_Rate,
    -- Education Indicators (2 features)
    Advanced_Degree_Rate,
    Young_Adult_Bachelors_Plus_Rate,
    -- Employment Type (2 features)
    STEM_Employment_Rate,
    Professional_Services_Rate,
    -- Digital Access (1 feature)
    Low_Income_Digital_Access_Rate,
    -- Demographics (1 feature)
    Total_Population
  FROM `st-project-454917.kansas_data.dataset6`
),
Base_County_Data AS (
  -- Start with unique counties and their core population metrics
  SELECT DISTINCT
    County_Name,
    MAX(County_Population) AS County_Population
  FROM `st-project-454917.kansas_data.dataset2`
  WHERE County_Name IS NOT NULL
  GROUP BY County_Name
)
-- FINAL SELECT: Combine all datasets with DS/AI college counts
SELECT
  bcd.County_Name,
  bcd.County_Population,
  roid.INTPTLAT AS County_Latitude,
  roid.INTPTLON AS County_Longitude,
```

```
  -- Educational Infrastructure (K12)
  COALESCE(csc.Total_Elementary_Schools, 0) AS Elementary_Schools,
  COALESCE(csc.Total_Middle_Schools, 0) AS Middle_Schools,
  COALESCE(csc.Total_High_Schools, 0) AS High_Schools,
  COALESCE(csc.Total_Virtual_Schools, 0) AS Virtual_Schools,
  COALESCE(csc.Total_Other_Schools, 0) AS Other_Schools,

  -- Total College Counts (all colleges)
  COALESCE(csc.Total_4_Year_Colleges, 0) AS Four_Year_Colleges,
  COALESCE(csc.Total_2_Year_Colleges, 0) AS Two_Year_Colleges,
  COALESCE(csc.Total_Less_Than_2_Year_Colleges, 0) AS Less_Than_Two_Year_Colleges,

  -- Colleges with DS/AI Programs by Type
  COALESCE(cdac.Four_Year_Colleges_With_DS_AI, 0) AS Four_Year_Colleges_With_DS_AI,
  COALESCE(cdac.Two_Year_Colleges_With_DS_AI, 0) AS Two_Year_Colleges_With_DS_AI,
  COALESCE(cdac.Less_Than_Two_Year_Colleges_With_DS_AI, 0) AS Less_Than_Two_Year_Colleges_With_DS_AI,

  -- Data Science/AI Program Metrics
  COALESCE(coi.Total_Online_Impact_Score, 0) AS Online_Impact_Score,
  COALESCE(cti.Total_Program_Impact_Score, 0) AS Total_Program_Impact_Score,

  -- Digital Infrastructure (from Dataset 5)
  COALESCE(di.Broadband_Access_Index, 0) AS Broadband_Access_Index,
  COALESCE(di.Internet_Adoption_Pct, 0) AS Internet_Adoption_Pct,
  COALESCE(di.Avg_Broadband_Coverage_Pct, 0) AS Avg_Broadband_Coverage_Pct,
  COALESCE(di.Pct_No_Internet, 0) AS Pct_No_Internet,
  COALESCE(di.Total_Households, 0) AS Total_Households,
  COALESCE(di.Effective_Access_Score, 0) AS Effective_Access_Score,

  -- Economic & Demographic Indicators (from Dataset 6)
  -- Economic Indicators (3 features)
  COALESCE(ed.Median_Household_Income, 0) AS Median_Household_Income,
  COALESCE(ed.Poverty_Rate, 0) AS Poverty_Rate,
  COALESCE(ed.Unemployment_Rate, 0) AS Unemployment_Rate,

  -- Education Indicators (2 features)
  COALESCE(ed.Advanced_Degree_Rate, 0) AS Advanced_Degree_Rate,
  COALESCE(ed.Young_Adult_Bachelors_Plus_Rate, 0) AS Young_Adult_Bachelors_Plus_Rate,

  -- Employment Type (2 features)
  COALESCE(ed.STEM_Employment_Rate, 0) AS STEM_Employment_Rate,
  COALESCE(ed.Professional_Services_Rate, 0) AS Professional_Services_Rate,

  -- Digital Access (1 feature)
  COALESCE(ed.Low_Income_Digital_Access_Rate, 0) AS Low_Income_Digital_Access_Rate,

  -- ML TARGET: Binary indicator of DS/AI curriculum existence
  CASE
    WHEN COALESCE(cti.Total_Program_Impact_Score, 0) > 0
    THEN 1 ELSE 0
  END AS Has_Programs

FROM Base_County_Data bcd
```

```
LEFT JOIN County_School_Counts csc
  ON TRIM(bcd.County_Name) = TRIM(csc.County_Name)
LEFT JOIN County_DS_AI_College_Counts cdac
  ON TRIM(REPLACE(bcd.County_Name, ' County', '')) = TRIM(REPLACE(cdac.COUNTYNM, ' County', ''))
LEFT JOIN County_Online_Impact coi
  ON TRIM(REPLACE(bcd.County_Name, ' County', '')) = TRIM(REPLACE(coi.COUNTYNM, ' County', ''))
LEFT JOIN County_Total_Impact cti
  ON TRIM(REPLACE(bcd.County_Name, ' County', '')) = TRIM(REPLACE(cti.COUNTYNM, ' County', ''))
LEFT JOIN Digital_Infrastructure di
  ON TRIM(bcd.County_Name) = TRIM(di.County_Name)
LEFT JOIN Economic_Demographics ed
  ON TRIM(REPLACE(bcd.County_Name, ' County', '')) = TRIM(REPLACE(ed.County, ' County', ''))
LEFT JOIN `st-project-454917.kansas_data.census_kansas_county_centroid` roid
  ON TRIM(bcd.County_Name) = TRIM(roid.NAMELSAD)

ORDER BY bcd.County_Name;
```

▼ ML Modeling

▼ Links

https://www.youtube.com/watch?v=t5myIGHE2Fg

Link to download WEKA software: https://sourceforge.net/projects/weka/

▼ Process

1. CSV Preparation
   - ensure 32 cols, 105 rows, no nulls, target var (Has_Programs) is last col
2. Load CSV and ensure data integrity
   - preferred by WEKA
   1. Explorer
   2. Open File → Change file type to csv → load dataset 7
   3. change the data type numeric → nominal for Has_Programs
      a. Preprocess → Filter → Choose → unsupervised → attribute → NumericToNominal
      b. Configuration Window → attributeIndices → 32 (feature index of Has_Programs) → OK → Apply
      c. ensure 88 (has programs) and 17 (no programs)
   4. exclude features from the dataset for training
      a. Preprocessing → Filter → Choose → unsupervised → attribute → Remove
      b. Configuration Window → attributeIndices → 1, 13, 14, 15, 16, 17 → OK → Apply
3. Run ML Experiments
   a. Random Forest
      i. Classify → Classifier → Choose → trees → Random Forest
      ii. maxDepth: 5; numIterations: 50; seed: 42 (for reproducibility)
      iii. right-click on configuration string → set -M parameter to 5.0 (min number of instances (data points) in leaf node)
      iv. [Primary Analysis - 10 Fold Cross Validation]
         1. Test options → Cross-validation → Folds → 10

v. [Secondary Analysis - Percentage Splits]
    1. Test options → Percentage split → 70, 60, 50
vi. Start → save results

b. SVM
    i. Classify → Classifier → Choose → functions → SMO
    ii. kernel: PolyKernel
    iii. [Primary Analysis - 10 Fold Cross Validation]
        1. Test options → Cross-validation → Folds → 10
    iv. [Secondary Analysis - Percentage Splits]
        1. Test options → Percentage split → 70, 60, 50
    v. Start → save results

c. Naives Bayes
    i. Classify → Classifier → Choose → bayes → NaiveBayes
    ii. Leave all defaults
    iii. [Primary Analysis - 10 Fold Cross Validation]
        1. Test options → Cross-validation → Folds → 10
    iv. [Secondary Analysis - Percentage Splits]
        1. Test options → Percentage split → 70, 60, 50
    v. Start → save results

4. Record All Metrics (see ML Analysis folder → ML Experiment Metrics.xlsx)
5. Top Features by Rank
    a. Select attributes → Attribute Evaluator → Choose → InfoGainAttributeEval → use defaults
    b. Select attributes → Search Method → Choose → Ranker → configure ranker parameters (numToSelect: -1 (default), threshold: (default))
    c. Select attributes → Attributes Selection Mode → Use full training set
    d. Start → save result
6. Correlation Matrix
    a. Select attributes → Attribute Evaluator → Choose → PrincipalComponents → use defaults
    b. Select attributes → Search Method → Choose → Ranker → configure ranker parameters (numToSelect: -1 (default), threshold: (default))
    c. Select attributes → Attributes Selection Mode → Use full training set
    d. Start → save result
    e. dont include County_Latitude or County_Longitude in the visualization
7. Cross Correlation Matrix
    a. Preprocess → Open File → Change file type to csv → load dataset 7
    b. change the data type numeric → nominal for Has_Programs
        i. Preprocess → Filter → Choose → unsupervised → attribute → NumericToNominal
        ii. Configuration Window → attributeIndices → 32 (feature index of Has_Programs) → OK → Apply
        iii. ensure 88 (has programs) and 17 (no programs)

c. exclude features from the dataset for training
    i. Preprocess → Filter → Choose → unsupervised → attribute → Remove
    ii. Configuration Window → attributeIndices → 1, 3, 4, 16, 17 → OK → Apply
d. run the algorithm
    i. Select attributes → Attribute Evaluator → Choose → PrincipalComponents → use defaults
    ii. Select attributes → Search Method → Choose → Ranker → configure ranker parameters (numToSelect: -1 (default), threshold: (default))
    iii. Select attributes → Attributes Selection Mode → Use full training set
    iv. Start → save result
    v. only include Four_Year_Colleges_With_DS_AI
    Two_Year_Colleges_With_DS_AI
    Less_Than_Two_Year_Colleges_With_DS_AI as the cols in the matrix and all the 23 predictors as the rows

▼ Notes

- Dataset 7: excluding Total_Program_Impact_Score and Online_Impact_Score from training
    - When kept in, Random Forest showed 100% accuracy (90-10, 80-20, 70-30, 60-40) and SVM showed 100% accuracy (90-10, 80-20). This is because of overfitting and these metrics are calculated from the target variable. Meaning, they already know the answer because they're derived from the same source as Has_Programs.
- also excluding Four_Year_Colleges_With_DS_AI, Two_Year_Colleges_With_DS_AI, and Less_Than_Two_Year_Colleges_With_DS_AI for the same reason (deriving from the same source as target variable)
- also excluding County Name because this is an identifier attribute and not a feature. The 1-to-1 mapping of county name to has_programs makes it easy for model to memorize NOT learn
- Removing 90-10 and 80-20 splits because it is causing Random Forest to overfit (even it with Total_Program_Impact_Score and Online_Impact_Score removed)

▼ Sources

▼ MINK Poster Sources

[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

[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] A. Singh - PowerBIPro. "How to Create Power BI TOP N Report (Power BI Top 10 / Top N Filter)." (Oct. 31, 2018). Accessed: Sep. 30, 2025. [Online Video]. Available: https://m.youtube.com/watch?v=4N-KocuYyNU.

[5] r/gis. "Census Data Download Stuck at 'Triggering Download'." Reddit. Mar. 12, 2022. Accessed: Sep. 30, 2025. [Online]. Available: https://www.reddit.com/r/gis/comments/tcg6zm/census_data_download_stuck_at_triggering_download/.

[6] Learn 2 Excel, “Selecting Heat Map Theme for ESRI Map Power BI,” (Oct. 19, 2019). Accessed: Oct. 1, 2025. [Online Video]. Available: https://www.youtube.com/watch?v=CGoA40hZB0c.

▼ Research Paper

https://www.overleaf.com/project → used word instead because overleaf was giving too much space for visuals

▼ Literature Review (Research Paper)

[1] STEM Pathways of Rural and Small-Town Students: Opportunities to Learn, Aspirations, Preparation, and College Enrollment

https://journals.sagepub.com/doi/epub/10.3102/0013189X211027528

- keyword search:
- insights: geographic disparities → rural and small town students less likely to enroll in postsecondary STEM degree programs, have limited access to advanced courses, and have less teaching capacity → shifted away from STEM fields even though they had similar interest levels for STEM as urban students
- quantitative methods: datasets → least-squares models and linear probability models
- visuals: comparison charts, statistical tables

[2] Data Science (Dataying) for Early Childhood

- keyword search:
- insights: Dataying framework (6 steps)
- method: conceptual framework, heavy synthesis of lit review
- visual: synthesis tables, framework graphics

[3] A Review of K-12 Data Science Education in the United States: Trends, Tools, and Gaps

https://peer.asee.org/a-review-of-k-12-data-science-education-in-the-united-states-trends-tools-and-gaps

- keyword search:
- insights: (good for explaining data science) emphasis on integrating data science education onto existing subjects; alternatives are data science camps and after school programs but can't appeal to socioeconomic disadvantaged
- method: lit review synthesis
- visual: tables

[4] Emergency remote teaching across urban and rural contexts: perspectives on educational equity

https://www.researchgate.net/profile/Earl-Aguilera/publication/342539219_Emergency_remote_teaching_across_urban_and_rural_contexts_perspectives_on_edu remote-teaching-across-urban-and-rural-contexts-perspectives-on-educational-equity.pdf?_sg%5B0%5D=started_experiment_milestone&origin=journalDetail&_rtd=e30%3D

- keyword search: "rural urban educational equity STEM access geographic disparities” (Google Scholar)
- insights:
  - rural students - (during COVID-19) expected to provide (internet connection, safe space for studying (homelessness))
  - both - kids not showing up, parents not enforcing homework

- method: dialogue b/w urban and rural teacher
- visual: none (all dialogue)

[5] Transforming Curriculum and Building Capacity in K-12 Data Science Education

https://hdsr.mitpress.mit.edu/pub/p9iemahk/release/1

- keyword search: "K-12 data science education curriculum pipeline undergraduate programs" (Google Scholar)
- insights: 31% of high school and 40% of middle school teachers feel very well prepared to teach statistics and probability → must train teachers (MOOCs, etc) and collab with data science specialists
- method: analysis of curriculum and school policy
- visual: none

[6] Spatial inequality in higher education: a growing urban–rural educational gap?

https://academic.oup.com/esr/article/40/6/1067/7633330

- keyword search: "rural urban educational equity STEM access geographic disparities" (Google Scholar)
- insights: rural woman (more inclined to leave and find alternatives) and urban men more inclined for HE, rural men have higher costs because they are tied to local community through gendered activity and work; overall rural people are less inclined (lack of development and less concentration of talent)
- method: theoretical framework that links inequalities b/w places and social inequalities b/w individuals and households, statistical analysis
- visual: tables, line graphs

[7] Disparities in Technology and Broadband Internet Access across Rurality: Implications for Health and Education

https://journals.lww.com/familyandcommunityhealth/abstract/2021/10000/disparities_in_technology_and_broadband_i

- keyword search: "online education digital divide broadband rural access barriers" (Google Scholar)
- insights: all of results section (KEY PAPER)
- method: survey data, analysis of student internet access data linked to geographic info
- visual: survey prompt, choropleth map, tables

[8] Rural school districts can be creative in solving the internet connectivity gap—but they need support

https://www.brookings.edu/articles/rural-school-districts-can-be-creative-in-solving-the-internet-connectivity-gap-but-they-need-support/

- keyword search: "online education digital divide broadband rural access barriers" (Google Scholar → Brookings)
- insights:
  - rural districts less likely to provide WIFI hotspots or devices
  - rural remedies - buying cellular data for students owning phones, connecting families with free or low internet, hotspots outside school grounds, teachers dropping off USB sticks with pre-recorded lectures
  - families don't want internet → teachers call about progress
  - few states pushing policy
- method: looking at CRPE study
- visual: bar chart

[9] Wealth Inequality Is a Barrier to Education and Social Mobility

https://www.urban.org/sites/default/files/publication/89976/wealth_and_education_2.pdf

- keyword search: "economic factors county wealth STEM education programs correlation" (Google Scholar)
- insights: family wealth is strongly associated with both higher educational attainment and upward educational mobility

- method: analysis from survey data
- visual: horizontal bar chart, tables

[10] STEM education and regional economic development

https://www.bls.gov/opub/mlr/2022/book-review/stem-education.htm

- keyword search: "economic factors county wealth STEM education programs correlation" (Google Scholar)
- insights: more high STEM or soft jobs → higher wages, only focusing on STEM is also not solution (regions could have more high paying soft jobs)
- method: book review (survey response analysis)
- visual: none

[11] Preparing K-12 Students to Meet their Data: Analyzing the Tools and Environments used in Introductory Data Science Contexts

https://dl.acm.org/doi/pdf/10.1145/3594781.3594796

- keyword search: "K-12 data science education equity access curriculum" (ACM Digital Library)
- insights: tools allow exploring exciting and diverse datasets, data visualization and statistical analysis tools seem to be prioritized in the reviewed tool (need more emphasis on data cleaning and choosing good datasets)
- method: systematic analysis of data science tools
- visual: table, graphics of GUIs

[12] https://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf

[13] https://seas.harvard.edu/news/what-data-science-definition-skills-applications-more

[14] https://holisticds.com/en/articulo/ai-economy/

[15] https://theacademicinsights.com/analytics-and-beyond-extracurricular-activities-that-make-learning-fun/

[16] https://sigai.acm.org/main/aboutus/

▼ Section Overview

1. title, authors, abstract (200-300 words, introduce gap in world, mention work, main insight)
2. intro - same as abstract but more details (where/how research, gap)
3. background and related work - give info to non-technical people, italics to separate sub-sections
4. methodology - qualitative (emotions, judgement) and quantitative (max, min) → focusing on quantitative
    a. Design - how you built plan of research
        i. EXPLORE WHAT DESIGN TODO
    b. Data Collection
    c. Sample
    d. Instrument
    e. Keywords, Database, Inclusion criteria
5. Results - what tables you created with what tools
6. Discussion - research question with answers based on tables and findings (something new)
7. Limitations - something beyond your control
    a. public college course info while manual scrape
8. Conclusion - similar to intro

9. Future work
10. Acknowledgment - people who funded project
11. References - IEEE format (use endpoint for citation, tool in Word)

▼ Research Questions

font

- body text - 12
- inside table text - 10
- table and figure label - 11

- 12-14 page limit
- figure titles at bottom; table titles top
- second paragraph is indented
- IEEE format

▼ Other

▼ VPN

https://support.cs.ksu.edu/CISDocs/wiki/Remote_Access#Remote_Windows_Access

https://support.ksu.edu/TDClient/30/Portal/KB/ArticleDet?ID=175

https://gpvpn.ksu.edu/global-protect/login.esp

- Used Global Protect VPN (Palo Alto product) and Windows App to connect to remote windows server to access PowerBI
  - VPN
    1. Portal: gpvpn.ksu.edu
    2. Kansas State eID
    3. Kansas State password
    4. Verification code: PUSH (sends MFA notification to secondary device via Duo Mobile app)
  - Windows App
    1. '+' sign → Add PC
    2. PC name: remote.cs.ksu.edu
    3. host: WIN2\eID
    4. Kansas State password

▼ VPN Backup

- Paralells Desktop was used to host a Windows virtual machine on my Mac to access PowerBI Desktop, which is only available to Windows users. PowerBI Desktop contains features not supported by PowerBI Online (e.g. creating custom fields)

▼ MINK Poster Feedback

- get private sector buy in on the importance of Data Science in commercial and small family generational farms. Why is the gap important to address?
  - It is obvious that where there are more people → more colleges → more course offerings. So what? What is this project trying to explore and what is the call to action?
  - Rural people can still take online classes → might show that rural people not getting access is not necessary true. How to account for this?

- If you can quantify the gap, then show it: 2 heat maps - one with schools and one with data science offerings, or rework it based on the population to offering
    - Can be easily shown
- Use anecdotes of small town farms that have benefitted from data science education
    - This will help the relevancy of the research paper.
    - How to collect anecdotes? Maybe from Kansas Academy of Science?
- connect with Kansas Academy of Science if it exists (coming from a faculty member in Nebraska that is the president of Nebraska's Academy of Science)
    - They might be a source for sponsorship, datasets, or information on why data science education is important for youth
    - Kansas Academy of Science does exist
- Create a call to action that can be acted upon or clearly state the goal- are you just creating another data set?
    - We must define more meaningful research questions for this project to specify the end goals/deliverables.
- Define "Data Science" clearly (comes back to using graphics so it isnt lost in the text, and what courses are classified as a data science offering is clear)
    - “Data Science” definition - Can reference word bubble diagram of course descriptions (made already)
    - Data Science course classification - if keywords like AI, Artificial Intelligence, ML, Machine Learning, Deep Learning, Reinforcement Learning, Data Science
- Are courses in business degrees also considered in the data science bin?
    - Irrespective degree or department, any course is considered if it passes the classification logic
- Why collect or represent elementary and middle schools at all if dataset two is just universities and high schools?
    - We can simplify the datasets and analysis only to colleges and high schools.
    - We should justify either direction:
        - Why were only high schools and colleges considered relevant? OR
        - Why was including elementary and middle schools relevant?
- Homeschool consideration
    - Maybe can find the number of rural areas who employ homeschooling methods. That is the only chance of getting a government/public dataset (maybe). Other than that, we would have to send surveys or intervew families about how they incorporate data science courses in their home-school curriculum.
- how virtual is represented in the location data
    - If school name has “Virtual” or “Academy” in the title.

Presenter Notes

- Seemed much more like a teacher that is motivated to stay in academia to fill the gap in education offerings than a student that is going to use data analytics in industry
- should address that the two data sets overlap and our point of interest is only the overlap (or lack thereof) [goes back to why we have the elementary and middle data at all]

Other

- Grinnell College (Iowa) students and faculty from various Nebraskan colleges were interested in expanding this to the general midwest if we find that gap and have a stronger basis, since they are also highly agricultural states with rural populations where the schools don’t have super specific data science offerings

▼ Misc

IT Help Desk: 785-532-7722

CS IT 785-532-6350

help@cs.ksu.edu

https://artsci.k-state.edu/research/undergraduate/