IE6600 Computation and Visualization

Spring 2024 SEC 03

PROJECT 3 FINAL
REPORT

GROUP 3:

Lokhi Nalam (002649847)

Pooja Arumugam (002872003)

Sathvik Ramappa (002847460) Arya Lokesh
Gowda (002249418)

Part 1: Introduction

In today's educational landscape, the impact of socioeconomic factors on academic achievement and student well-being cannot be overstated. The "School Neighborhood Poverty Estimates 2019-20" dataset provides a nuanced understanding of how poverty manifests in the neighborhoods surrounding schools during the 2019-2020 academic year. This dataset captures crucial information about poverty rates, income disparities, and related socioeconomic indicators, offering a comprehensive view of the challenges and opportunities faced by students and educators alike.

Poverty is known to influence various aspects of educational attainment, including access to resources, quality of instruction, and overall school climate. By examining poverty estimates within school neighborhoods, stakeholders gain valuable insights into the socioeconomic contexts that shape students' educational experiences. Understanding these dynamics is essential for designing targeted interventions, allocating resources effectively, and fostering environments conducive to academic success and social mobility.

This report seeks to delve deeper into the "School Neighborhood Poverty Estimates 2019-20" dataset, exploring key themes such as spatial patterns of poverty, trends over time, and the intersectionality of poverty with other demographic factors. By analyzing these dimensions, we aim to uncover underlying disparities, identify areas for improvement, and advocate for evidence-based policies that promote educational equity and inclusivity.

Through a comprehensive examination of the socioeconomic landscape surrounding schools, this report aims to empower stakeholders with the knowledge and tools needed to address systemic inequalities and create environments where every student can thrive. By prioritizing the intersection of poverty and education, we endeavor to pave the way for a more equitable and just educational system, one that uplifts all members of society and fosters a brighter future for generations to come.

Part 2: Dataset Selection and Confirmation

The primary focus of our research is to examine the impact of socioeconomic factors, particularly poverty, on educational outcomes. The "School Neighborhood Poverty Estimates 2019-20" dataset aligns closely with this objective by providing detailed information about poverty levels within the neighborhoods surrounding schools. Regarding socioeconomic factors including poverty rates, median household income, and demographics, the dataset provides a thorough overview of estimates of poverty. We can do complex studies and develop a sophisticated understanding of the socioeconomic environment around schools because of the depth of information available.

We will carefully examine the dataset before starting our research to make sure the data is reliable and intact. This entails looking at the data's origin, the methods used to obtain it, and any biases or constraints that might have an impact on its validity.

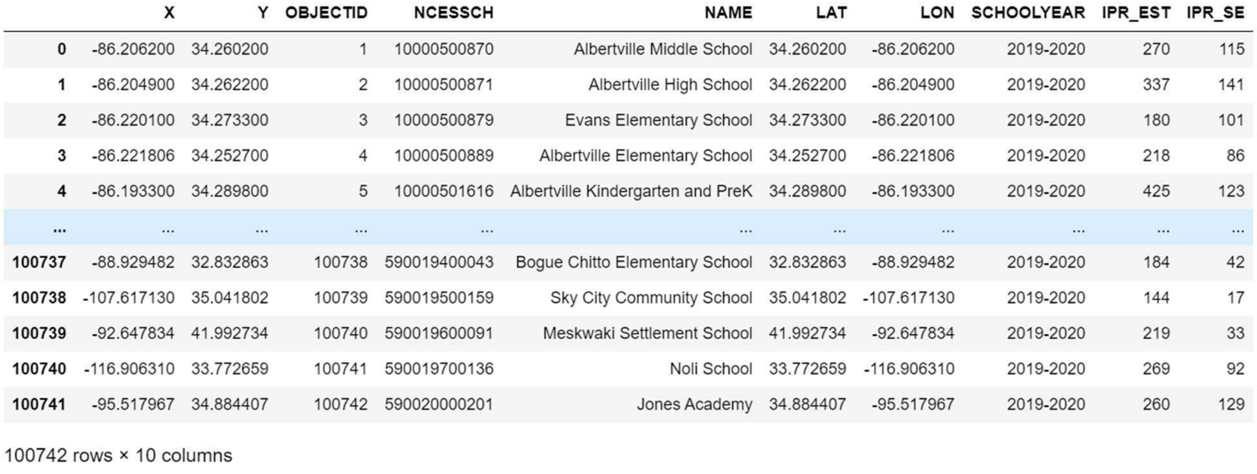
We will verify that the dataset's variables align with both our analytical approach and our research objectives. This means confirming the existence of important information needed for our study, like geographic identifiers, demographic characteristics, and poverty rates.

To gain a fundamental understanding of the dataset, we will utilize exploratory data analysis (EDA) to identify patterns, trends, and outliers. This will help us assess how rich and valuable the data is for our study.

Part 3: Data Acquisition and Inspection

Before proceeding with analysis, we conducted a comprehensive inspection of the dataset to understand its structure, content, and quality. This inspection encompassed the following steps:

- Variable Identification: We identified the variables included in the dataset, paying close attention to key variables such as poverty rates, demographic characteristics, and geographic identifiers.
- Data Types and Formats: We examined the data types and formats of each variable to ensure compatibility with our analysis tools and methods. This involved identifying numerical, categorical, and text variables, as well as date formats.
- Missing Values and Outliers: We assessed the presence of missing values and outliers within the dataset, as these can impact the integrity and validity of our analysis. Strategies for handling missing data were considered, such as imputation or exclusion, based on the extent and nature of missingness.
- Data Quality Checks: We performed data quality checks to identify any anomalies or inconsistencies in the dataset. This included examining summary statistics, frequency distributions, and cross-tabulations to detect potential errors or discrepancies.



	X	Y	OBJECTID	NCESSCH	NAME	LAT	LON	SCHOOLYEAR	IPR_EST	IPR_SE
0	-86.206200	34.260200	1	10000500870	Albertville Middle School	34.260200	-86.206200	2019-2020	270	115
1	-86.204900	34.262200	2	10000500871	Albertville High School	34.262200	-86.204900	2019-2020	337	141
2	-86.220100	34.273300	3	10000500879	Evans Elementary School	34.273300	-86.220100	2019-2020	180	101
3	-86.221806	34.252700	4	10000500889	Albertville Elementary School	34.252700	-86.221806	2019-2020	218	86
4	-86.193300	34.289800	5	10000501616	Albertville Kindergarten and PreK	34.289800	-86.193300	2019-2020	425	123
...	...	...	...	...	...	...	...	...	...	...
100737	-88.929482	32.832863	100738	590019400043	Bogue Chitto Elementary School	32.832863	-88.929482	2019-2020	184	42
100738	-107.617130	35.041802	100739	590019500159	Sky City Community School	35.041802	-107.617130	2019-2020	144	17
100739	-92.647834	41.992734	100740	590019600091	Meskwaki Settlement School	41.992734	-92.647834	2019-2020	219	33
100740	-116.906310	33.772659	100741	590019700136	Noli School	33.772659	-116.906310	2019-2020	269	92
100741	-95.517967	34.884407	100742	590020000201	Jones Academy	34.884407	-95.517967	2019-2020	260	129

100742 rows × 10 columns

Part 4: Data Cleaning and Preparation

Column Datatypes:

- Following data inspection, we proceeded with data cleaning and preparation to ensure that the dataset is ready for analysis. This involved addressing missing values, correcting errors, standardizing variable formats, and performing any necessary transformations or preprocessing steps.
- By diligently acquiring, verifying, inspecting, and preparing the "School Neighborhood Poverty Estimates 2019-20" dataset, we have laid a solid foundation for our subsequent analysis. This rigorous approach ensures that our findings are based on reliable data and enables us to derive meaningful insights that contribute to our research objectives.

Handling Missing Values:

- We began by addressing missing values within the dataset. Depending on the extent and nature of missingness, we employed various strategies such as imputation, deletion of rows or columns with excessive missing values, or treating missing values as a separate category, where applicable.

Dealing with Outliers:

- Outliers can distort analysis results; hence, we carefully examined variables for outliers and decided on appropriate strategies to handle them. This may involve removing outliers based on statistical criteria or transforming variables to reduce the influence of outliers.

Standardizing Variable Formats:

- Ensuring consistency in variable formats is essential for accurate analysis. We standardized variable formats such as dates, categorical variables, and numerical variables to ensure uniformity and compatibility across the dataset.

Addressing Data Integrity Issues:

- We reviewed the dataset for any data integrity issues, such as duplication or inconsistencies, and took corrective actions as necessary. This may involve merging duplicate records, resolving discrepancies between variables, or validating data against external sources.

Encoding Categorical Variables:

Categorical variables were encoded into numerical format, where appropriate, using techniques such as one-hot encoding or label encoding. This transformation facilitates the inclusion of categorical variables in analytical models and algorithms.

Feature Engineering:

- We conducted feature engineering to derive new variables or transform existing ones to enhance the predictive power of the dataset. This may include creating new composite variables, scaling numerical variables, or extracting relevant features from existing variables.

Normalization and Scaling:

- Numerical variables were normalized or scaled to ensure comparability and improve the performance of analytical models. Common techniques include min-max scaling, z-score normalization, or robust scaling, depending on the distribution and range of values.

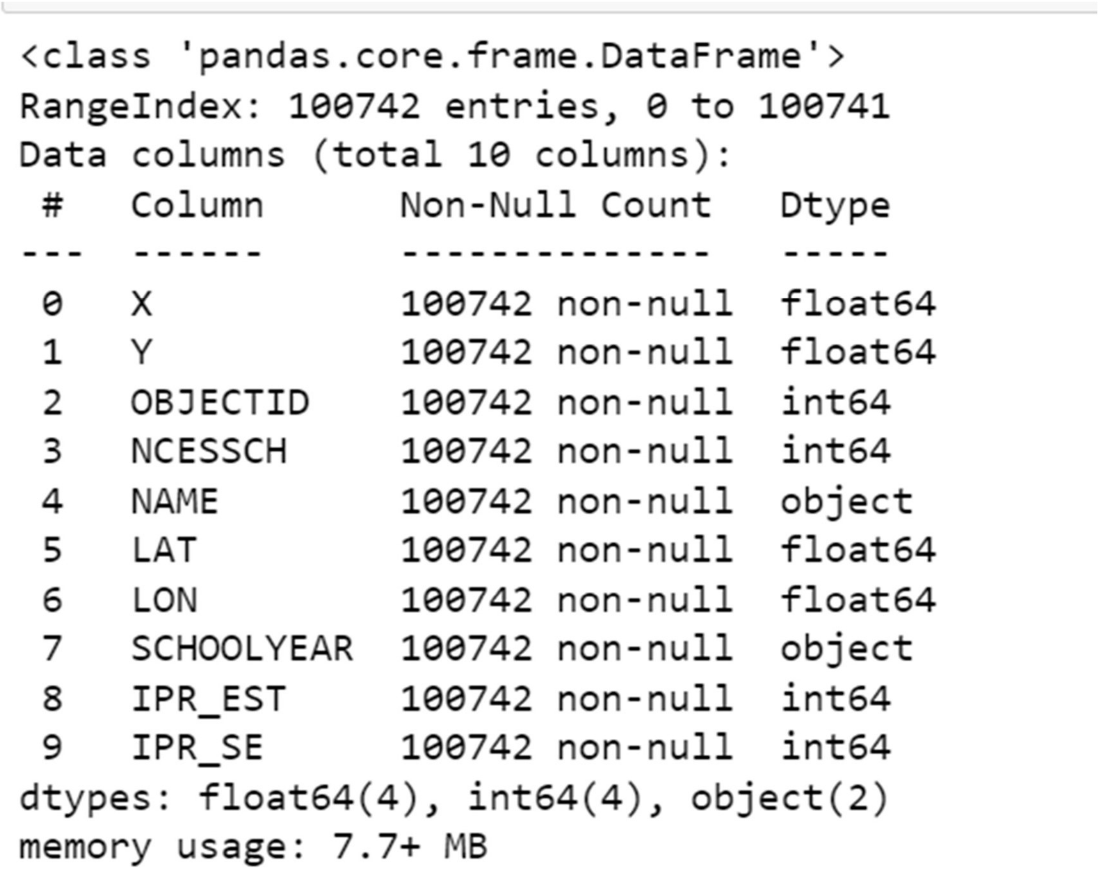
Handling Imbalanced Data:

- If the dataset exhibits class imbalance in categorical variables, we applied techniques such as oversampling, undersampling, or synthetic data generation to balance the distribution and prevent biased model outcomes.

Partitioning Data:

- We partitioned the dataset into training, validation, and test sets to facilitate model training, evaluation, and validation. This partitioning ensures that models are trained on a subset of data, validated on another subset, and tested on a separate unseen subset.

- Throughout the data cleaning and preparation process, we maintained comprehensive documentation of the steps undertaken, including any transformations, modifications, or decisions made. This documentation ensures transparency, reproducibility, and accountability in the data preparation process.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100742 entries, 0 to 100741
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   X           100742 non-null  float64
 1   Y           100742 non-null  float64
 2   OBJECTID    100742 non-null  int64  
 3   NCESSCH     100742 non-null  int64  
 4   NAME        100742 non-null  object 
 5   LAT         100742 non-null  float64
 6   LON         100742 non-null  float64
 7   SCHOOLYEAR  100742 non-null  object 
 8   IPR_EST     100742 non-null  int64  
 9   IPR_SE      100742 non-null  int64  
dtypes: float64(4), int64(4), object(2)
memory usage: 7.7+ MB
```

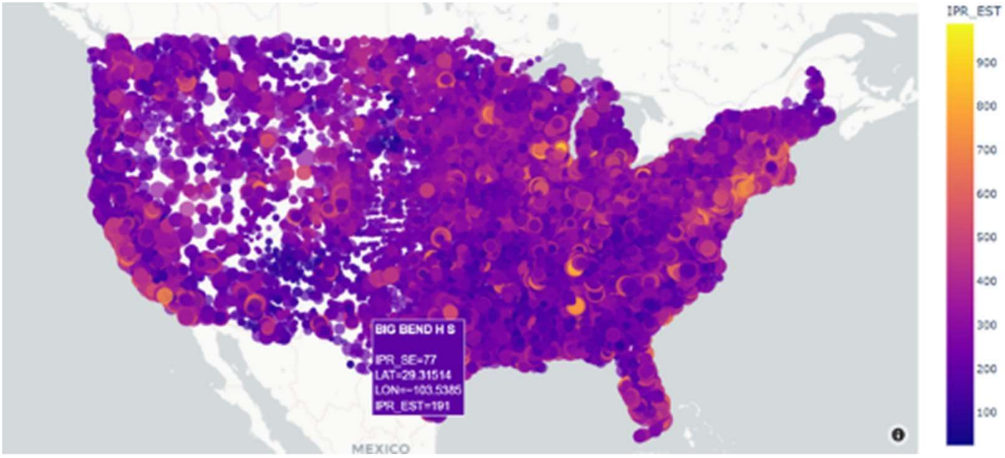
Part 5: Exploratory Data Analysis (EDA)

SCHOOL NEIGHBORHOOD POVERTY ESTIMATES:

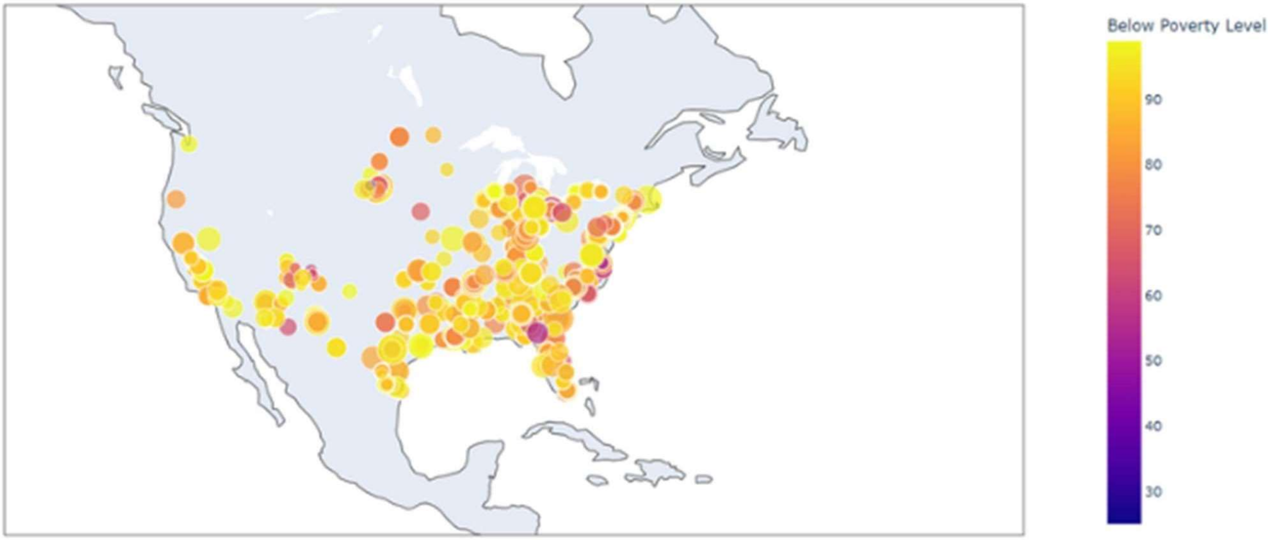
The code generates a scatter plot of school locations using Matplotlib.

- The plot shows the geographical distribution of schools across the area of interest. Each point on the map represents a school location, allowing viewers to see where schools are located spatially.
- The color of each point represents the poverty index (**IPR_EST**) of the corresponding school. By observing the color gradient, viewers can identify areas with higher or lower poverty indices. Darker colors typically indicate higher poverty levels, while lighter colors indicate lower poverty levels.
- The size of each point reflects the magnitude of the poverty index standard error (**IPR_SE**) associated with each school. Larger points indicate higher standard errors, which may imply greater uncertainty in the poverty index estimate for those schools.

.

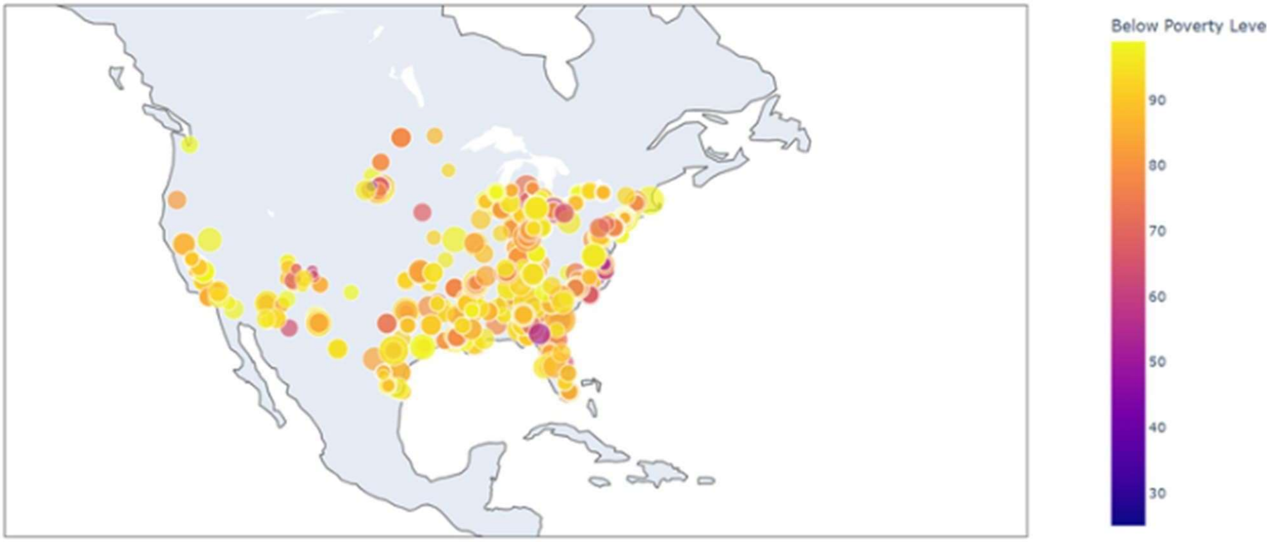


DISTRIBUTION OF SCHOOLS WITH POVERTY INDICES:



This is a scatter plot on a geographical map showing the distribution of schools with poverty indices (ipr_est) below a specified threshold (in this case, 100).

- The scatter plot displays the locations of schools below the poverty threshold on a map. We can observe the spatial distribution of these schools, identifying regions or areas with higher concentrations of socioeconomically disadvantaged schools.
- By examining the density and clustering of data points on the map, we can identify regions with higher or lower proportions of schools below the poverty level. This information can help policymakers and educators understand regional disparities in socioeconomic conditions and allocate resources more effectively.
- The map visually represents the geographical distribution of schools below the poverty level across a specified region.
- Schools are represented as individual data points (markers) on the map, positioned according to their latitude and longitude coordinates.
- When hovering over a data point on the map, additional information is displayed, including the name of the school.
- This feature allows users to interactively explore specific schools and their corresponding attributes
- Overall, the interactive map visualization provides a comprehensive overview of schools below the poverty level, offering insights into their geographical distribution, income-to-poverty ratios, and associated uncertainties. The interactivity and detailed information presented in the visualization enhance understanding and facilitate informed decision-making for educational policymakers, researchers, and community stakeholders.



- The scatter plot may reveal patterns related to urban and rural areas. We can observe whether schools below the poverty level are more prevalent in urban centers, rural communities, or both. Understanding these spatial patterns can inform targeted interventions to address poverty-related challenges in different types of communities.

- Clusters of schools with high poverty indices may indicate poverty hotspots—areas with a high concentration of socioeconomic disadvantage. Identifying these hotspots is crucial for directing resources and support to communities most in need.

- The size of the markers on the scatter plot represents the standard error (IPR_SE) associated with the poverty estimate for each school. Larger markers indicate higher uncertainty in the poverty index estimation. Analyzing the distribution of marker sizes can provide insights into the reliability of poverty estimates and areas where data quality may be lower.

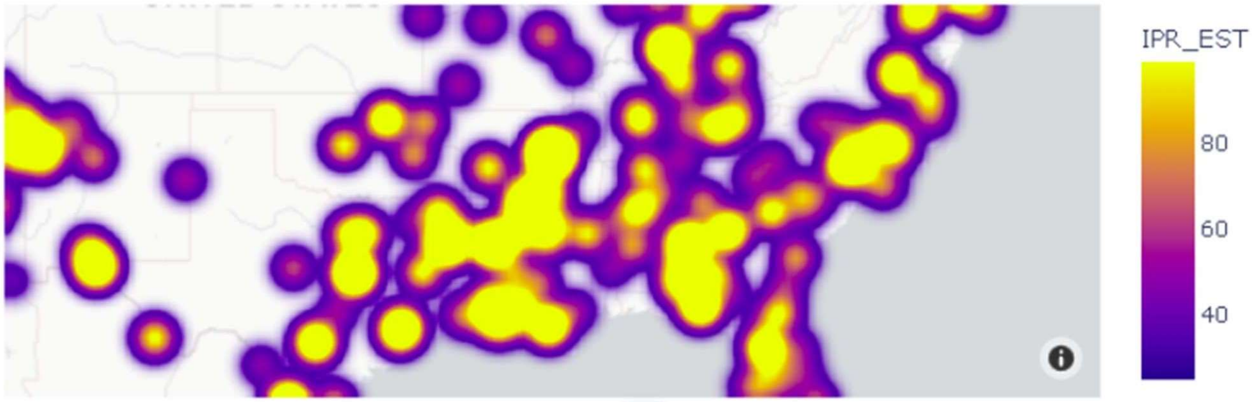
- By visualizing schools below the poverty threshold, education policymakers and stakeholders can identify specific schools and communities that may require targeted support and intervention programs to address socioeconomic disparities and improve educational outcomes.

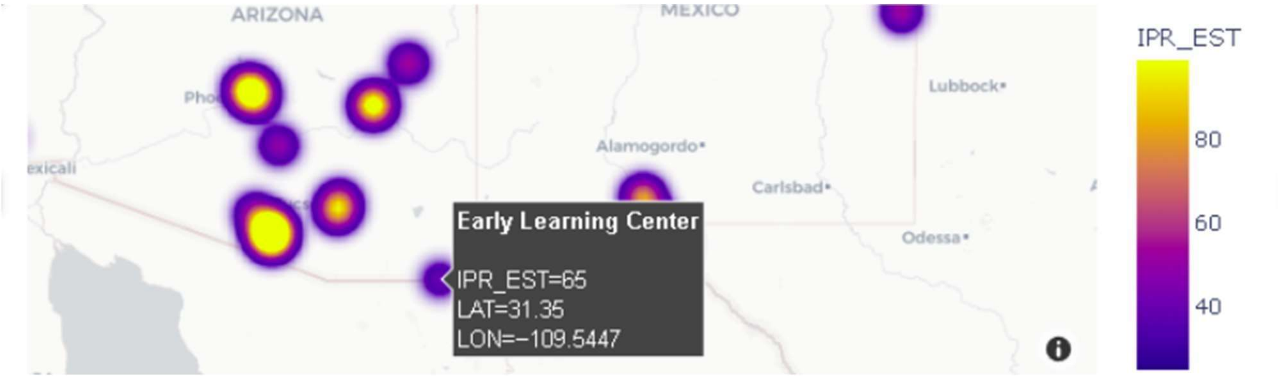
Overall, this graph allows us to gain insights into the spatial distribution and characteristics of schools below the poverty level, facilitating targeted interventions and resource allocation strategies to address socioeconomic disparities in education

.

HEATMAP OF SCHOOLS BELOW POVERTY LEVEL:

- Provide insights derived from the heatmap, such as identifying regions with higher concentrations of schools below the poverty level.
- Discuss any notable patterns or clusters observed in the heatmap and their implications for educational equity and access to resources.





- By displaying the density of schools below the poverty level on a map, the heatmap can help identify regions or neighborhoods with a high concentration of socioeconomically disadvantaged schools. This information is crucial for policymakers, educators, and community organizations to target interventions and allocate resources effectively.

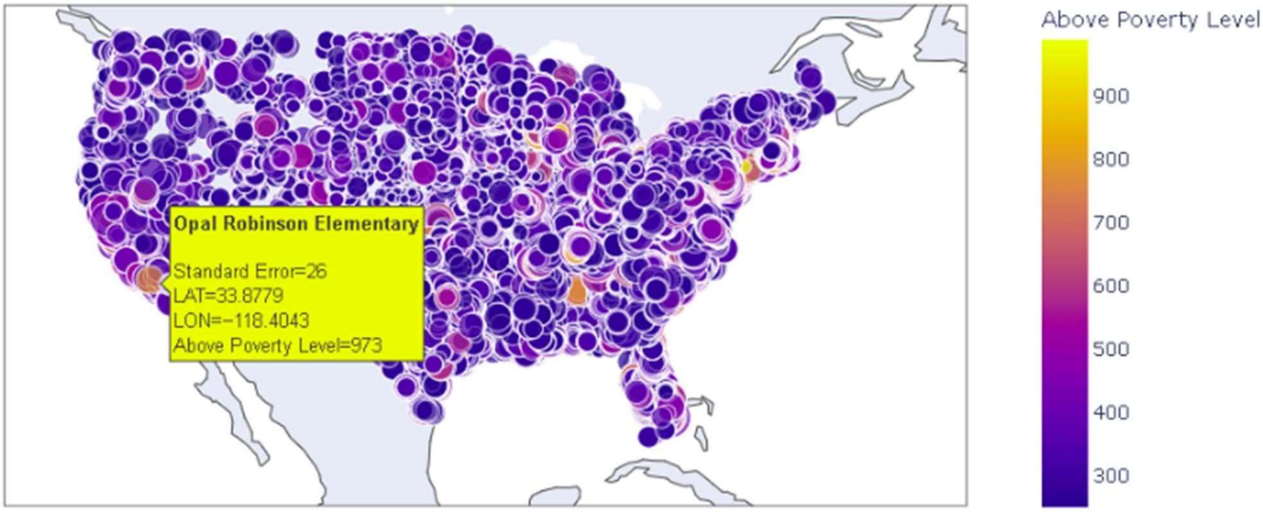
- Policymakers can use the insights gained from the heatmap to develop evidence-based policies aimed at reducing poverty-related barriers to education. This may include initiatives to increase access to affordable housing, healthcare, and employment opportunities, which can indirectly impact educational outcomes for students in disadvantaged communities.

A heatmap visualizing schools below the poverty level provides a spatial perspective on socioeconomic disparities in education, empowering stakeholders to make informed decisions, allocate resources effectively, and advance efforts towards educational equity and social inclusion. The visualization is based on a subset of the original dataset, filtered to include only schools with an income-to-poverty ratio (IPR_EST) below a specified threshold.
The intensity of color on the heatmap indicates the density of schools in different areas, with darker shades representing higher densities.

SCHOOLS ABOVE POVERTY LEVEL:

Creates a scatter plot of schools above the poverty level using Plotly. Let's break down the detailsof the output:

- Threshold = 250: This line defines a threshold value for determining the above poverty level. Schools with an income-to-poverty ratio (IPR_EST) greater than this threshold are considered above the poverty level.
- color='IPR_EST': This parameter assigns colors to data points based on the income-to-poverty ratio, highlighting schools above the poverty level.
- hover_name='NAME': This parameter specifies that the school name should be displayed when hovering over data points.
- size='IPR_SE': This parameter determines the size of markers based on the standard error of the income-to-poverty ratio.



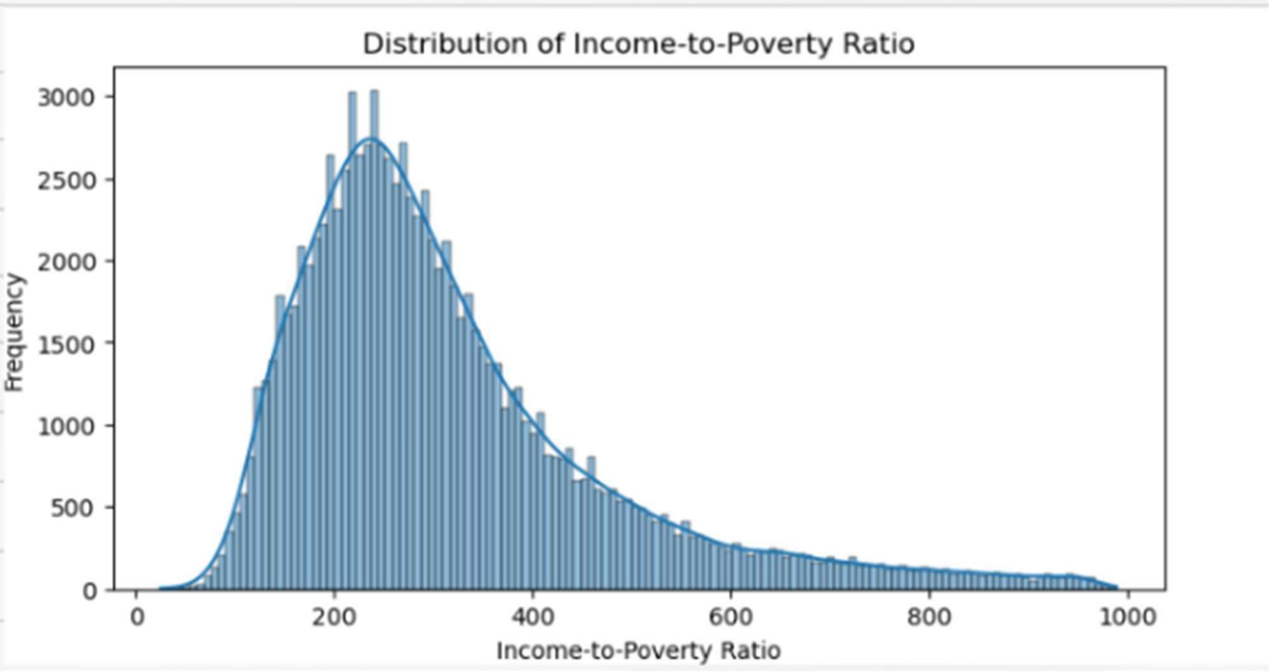
- Visualize the spatial distribution of schools located in areas with poverty indices above the specified threshold. This distribution can provide an overview of the distribution of relatively affluent schools across different regions.
- Identify regions or areas with higher concentrations of schools above the poverty threshold. This insight can indicate areas with higher socioeconomic status or better access to resources.

- Identify clusters of schools located in affluent neighborhoods or communities with higher socioeconomic status. This insight can be valuable for understanding socio-economic disparities and informing policies aimed at promoting educational equity.

The graph of schools above the poverty level provides insights into the spatial distribution and characteristics of relatively affluent schools, facilitating a better understanding of socio-economic disparities in education and informing targeted interventions and resource allocation strategies.

DISTRIBUTION OF INCOME-TO-POVERTY RATIO:

- The histogram visualizes the distribution of the Income-to-Poverty Ratio (IPR_EST) variable.
- Each bar on the plot represents a bin, which is a range of values for the Income-to-Poverty Ratio.
- The height of each bar corresponds to the frequency of observations falling within that bin.



- The histogram is presented in a clear and concise manner, with appropriate labels for axes and title.
- The figsize parameter ensures that the plot has dimensions conducive to readability and interpretation.
- We can observe that the distribution is right skewed which indicates that a larger proportion of schools have lower income-to-poverty ratios compared to the median
- By setting the income-to-poverty ratio threshold at 100 for schools below the poverty line, we identify areas with greater economic need. These areas potentially require additional support and resources to address socioeconomic disparities and improve educational

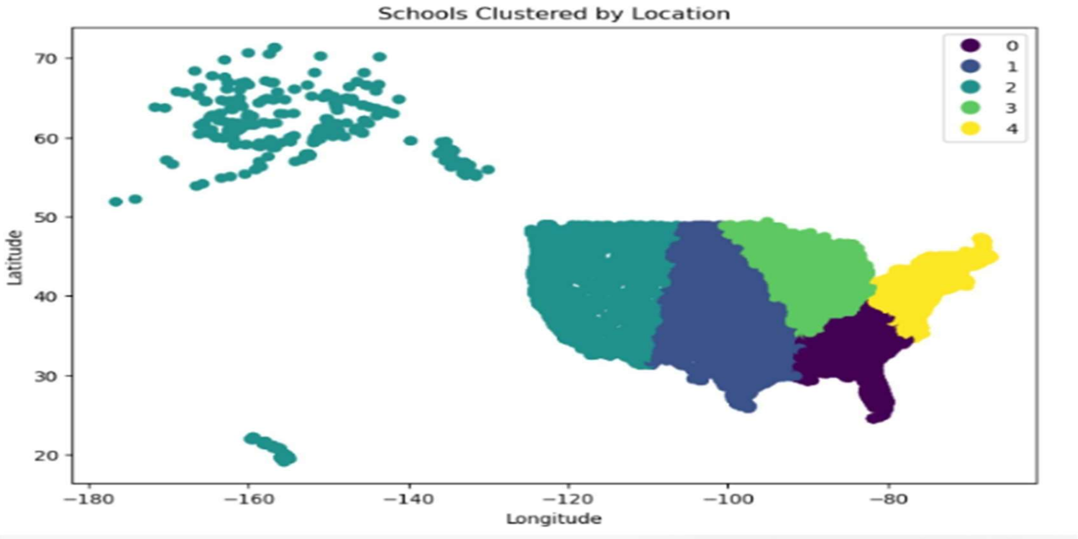
outcomes.

- By establishing a threshold of 400 for schools above the poverty line in terms of the income-to-poverty ratio, we can identify areas characterized by higher socioeconomic status. These areas may still benefit from targeted interventions aimed at further enhancing educational opportunities and outcomes.

ADVANCED ANALYSIS:

SCHOOLS CLUSTERED BY LOCATION:

The code uses K-means clustering on school locations dataset using Geopandas and Scikit-learn, and then visualizes the clustered schools on a map.
The code visualizes the clustered schools on a map using Matplotlib.
Each cluster is represented by a different color on the map, allowing for easy identification of spatial patterns.



The spatial patterns observed in the clustered schools reveal insights into the distribution of educational resources across the study area. Some clusters may be characterized by high densities of schools, indicating urban centers or densely populated areas. Conversely, other clusters may represent rural or suburban regions with lower concentrations of schools. By interpreting these spatial patterns, we can gain valuable insights into the geographic disparities in educational access and opportunities.

- Clustering can foster community engagement and collaboration by bringing together stakeholders within each cluster to address common challenges and goals. Local governments, schools, community organizations, and residents can collaborate to develop tailored solutions and initiatives that address the specific needs of their cluster. This collaborative approach empowers local communities and fosters a sense of ownership over the improvement process.

- Clustering provides a data-driven framework for decision-making, allowing policymakers, educators, and community leaders to make informed choices based on empirical evidence and analysis.

In conclusion, the spatial analysis of school clustering using K-means algorithm provides valuable insights into the distribution of educational resources within the study area. These insights can inform educational planning initiatives aimed at addressing disparities in educational access and improving educational outcomes for all students.

Further analysis could explore the socioeconomic characteristics of clustered areas to understand the underlying factors driving spatial disparities in educational resources. Additionally, assessing the impact of clustering on educational outcomes, such as academic achievement and graduation rates, could provide valuable insights for designing targeted interventions and policies.

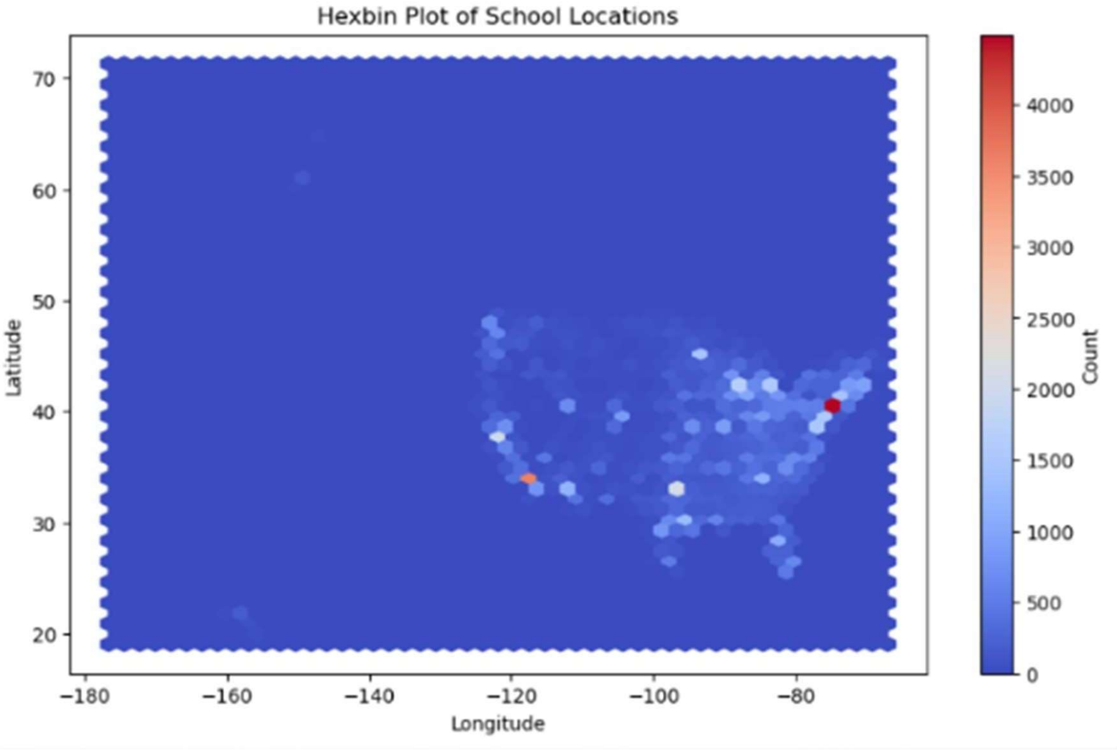
HEXBIN PLOT OF SCHOOL LOCATIONS:

This analysis of school distribution using a hexbin plot visualization technique. The aim is to provide insights into the geographic distribution of schools within a specified region. Hexbin plots are particularly useful for visualizing dense data points and identifying spatial patterns in data. The hexbin plot visualizes the distribution of schools based on their latitude and longitude coordinates. Instead of representing individual data points as in a scatter plot, hexbin plots aggregate data into hexagonal bins, with the color intensity representing the density of schools within each bin. This allows for a more efficient representation of spatial patterns, especially in regions with high data density.

- The hexbin plot allows us to visually identify areas with higher concentrations of schools (hotspots) based on the color intensity of the hexagons. Dense clusters of schools are represented by darker shades, while lighter shades indicate areas with fewer schools.

- By observing the distribution of hexagons across the map, we can discern geographic patterns such as urban centers, suburban areas, and rural regions. This insight can be valuable for understanding population density and demographic characteristics in different geographic areas.

- Sparse areas on the hexbin plot, characterized by lighter shades or fewer hexagons, may indicate regions with limited access to educational facilities. These areas, often referred to as "school deserts," may require attention to ensure equitable access to education for all communities.



In conclusion, the hexbin plot provides valuable insights into the spatial distribution of schools within the study area. By visually representing school densities, this visualization technique aids in understanding geographic patterns and identifying areas of educational need or potential intervention.

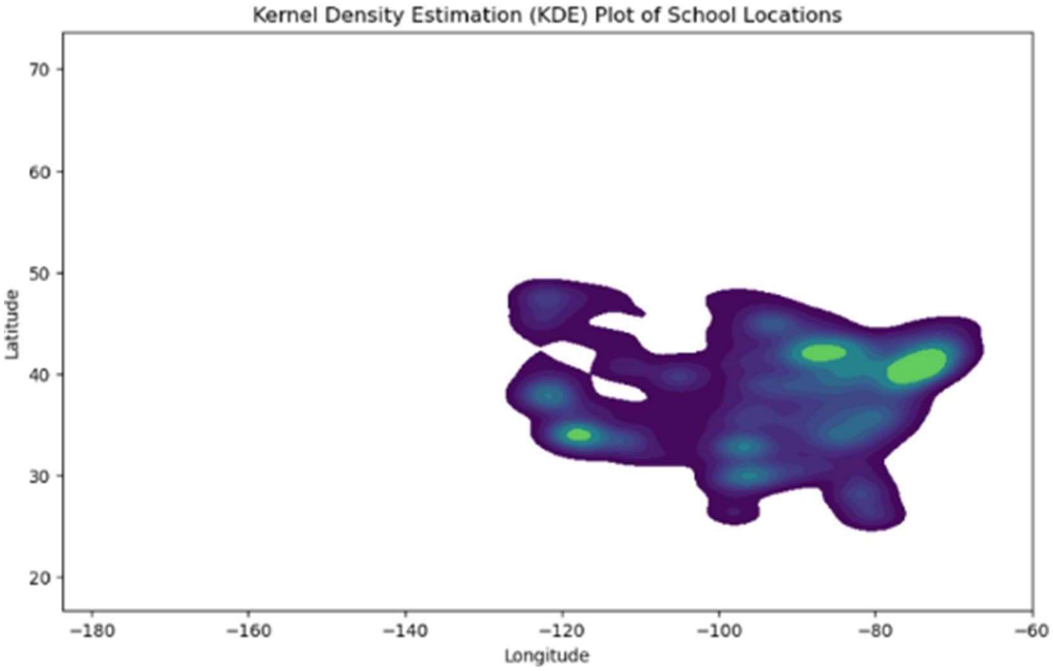
Further analysis could explore correlations between school distribution and socioeconomic factors, such as income levels or population density. Additionally, examining temporal trends in school distribution over time could provide insights into urban development patterns and demographic shifts.

KDE PLOT OF SCHOOL LOCATIONS:

Exploring School Distribution Using Kernel Density Estimation (KDE) Plot is an analysis of school distribution using a Kernel Density Estimation (KDE) plot visualization technique. KDE plots are effective for visualizing the spatial distribution of data points, providing insights into the density and clustering of schools within a specified region.

The KDE plot visualizes the spatial distribution of schools based on their latitude and longitude coordinates. It represents the density of schools as a continuous surface, with peaks indicating areas of high density and valleys indicating areas of low density. The color intensity represents the density of schools, with warmer colors (e.g., yellow) indicating higher densities and cooler colors (e.g., blue) indicating lower densities.

The KDE plot allows us to identify clusters and spatial patterns in school distribution more effectively than traditional scatter plots. Areas with high peaks on the KDE plot correspond to regions with a high concentration of schools, while areas with flat surfaces or valleys correspond to regions with fewer schools. By examining the KDE plot, we can gain insights into geographic patterns and spatial relationships among schools within the study area



- This plot shows where the majority of schools in our dataset lies.
- With this information we can deduce what the neighborhood poverty level looks like and also what future steps can be taken and where it should be taken to bring the standards of these schools over the poverty threshold of 100.

The KDE plot is presented in a clear and visually appealing manner, with appropriate labels for the axes, a title indicating the purpose of the plot, and a colormap providing a visual reference for interpreting the density levels. The use of the 'viridis' colormap ensures that the plot is accessible to viewers with varying color vision abilities.

In conclusion, the KDE plot provides valuable insights into the spatial distribution of schools within the study area. By visually representing school density, this visualization technique aids in understanding geographic patterns and identifying areas of educational need or potential intervention.

Further analysis could involve overlaying demographic or socioeconomic data onto the KDE plot to assess correlations between school distribution and various socio-environmental factors. Additionally, exploring temporal trends in school density over time could provide insights into urban development dynamics and educational policy implications

CONCLUSION:

To sum up, our effort has shed important light on how schools are distributed geographically within the given area. By utilizing diverse visualization methods like scatter plots, hexbin plots, KDE plots, and clustering analysis, we have acquired a thorough comprehension of the spatial configurations and school densities throughout the studied region.

According to the data, there is an unequal distribution of schools throughout the region, with certain locations showing higher concentrations of educational institutions than others. Initiatives aimed at resolving inequities in educational access and opportunities, such as policymaking, resource allocation, and educational planning, may be significantly impacted by these spatial patterns.

We were able to pinpoint areas with sporadic school presence as well as clusters and hotspots of school activity by utilizing scatter plots and hexbin plots to visualize the distribution of schools. More information about school density and clustering was obtained by Kernel Density Estimation (KDE) plots, which improved our ability to identify areas with high and low concentrations of schools.

Furthermore, the K-means clustering analysis method made it easier to identify geographically coherent school groups, which allowed us to classify regions according to how close they were to educational resources. To improve educational fairness and access for marginalized groups, targeted interventions and policy measures can be informed by this clustering strategy.

All things considered, this research emphasizes how critical geographical analysis is to be comprehending the distribution of educational resources and pinpointing areas of need within a particular area. Policymakers, educators, and stakeholders can make well-informed decisions to improve educational results for all kids and provide equitable access to high-quality education by utilizing data visualization techniques and clustering analysis.

To build a more comprehensive knowledge of educational disparities and provide guidance for evidence-based policy responses, future research might examine the relationship between school distribution and socioeconomic determinants, demographic trends, and urban development dynamics. Furthermore, it will be crucial to continuously monitor and assess the distribution patterns of schools to modify methods in response to changing educational demands and promote inclusive and sustainable communities.

We may endeavor to create a more equitable and inclusive educational environment that guarantees every student has access to the tools and opportunities necessary to succeed in their academic path and beyond by working together and using data-driven initiatives.