Master Capstone Project

Understanding Public Schools System in MA

Yuanyuan Yang

1. Disclaimer

This document represents a dynamic framework for the project, capturing the current status to the best of my knowledge. It outlines the general planning and comprehensive details of the project as it stands now. However, it is important to acknowledge that this project will undergo changes and enhancements as it progresses.

This project uses real data legally scraped from government websites. Unlike conventional projects that are guided by established guidelines or research papers, this initiative stands as an exploration. This project is driven by two-fold motivations: 1) genuine passion for comprehending an unfamiliar domain, seeking to provide useful and practical insights to key stakeholders, 2) desire to apply acquired knowledge to address real-world challenges, and gain experiences of learned technologies.

Hence, it is important to recognize that this document presents my best understanding of the project as current, and the final deliverables could be different from this document.

2. Motivation

The public education (PreK-12 education) stands as the foundational and significant educational phase for children. With the help of Machine Learning and large-scale data, this project is trying to provide us with a comprehensive and detailed portrayal of our public-school districts in the Commonwealth of Massachusetts, which includes traditional public schools and independent public schools: charter schools.

The evidence-centric approach could enhance our understanding and help us to identify key determinants, reveal their interactions and explain how they could influence the final outcomes. It aims to empower our policymakers with actionable insights. Furthermore, it has the potential to catalyze a positive transformation to a more efficient, supportive and impartial educational environment for the future.

3. Potential Usage and Stakeholders

- Policymakers and education institute administrators could have a comprehensive and in-depth understanding of the current situation and enhance the efficiency and fairness of resource allocation through well-informed decisions.
- Educators and caregivers could enhance their understanding of the influencing factors and their respective situations. This meaningful awareness would empower them to make more informed choices based on their unique circumstances.

4. Scope of the Project

This project devotes to analyze the main factors behind the discrepancies of students' performances in public school districts of Massachusetts. Following are the aspects that this project will focus on. Please note that the list could change throughout the project:

- **Quantitative Correlation Between Student Performance and Socioeconomic Status:** We aim to discern the numerical correlation between student academic achievement and their socioeconomic backgrounds.
- **Impact of Increased Funding on Low Socioeconomic Status Students:** We will study whether augmented funding correlates with enhanced performance specifically among students from low socioeconomic backgrounds.
- **Optimizing Resource Utilization:** We will explore methods to optimize resource allocation effectively. This could be distributing funds to increase teacher-student ratios, improve teacher compensation, and more.
- **Gender Ratio's Influence on Student Performance:** We intend to exam the potential impact of gender ratios within schools on overall student performance.
- **Demographic Distribution's Impact on Student Performance:** We will investigate the influence of demographic distribution within schools on the academic performance of students.
- **Charter Schools and Equity Mitigation:** Charter School is established to stimulate development of innovation programs within public education, and to encourage performance-based educational programs. We will delve into whether charter schools, detached from traditional public schools, contribute to alleviating the

inequities among school districts and stimulating improvements within public schools.

5. Pipeline and Techniques

5.1. Pipeline

Exploration – This stage is primarily focused on collecting data from multiple sources and preparing it for further preprocessing.

Data processing and organization – This stage is for cleaning, selecting, and organizing data for the following implementations.

Data Visualization – This stage is for gaining a visual understanding of the dataset in both time and geographical dimensions, identifying patterns, and uncovering hidden trends.

Models Training – This stage is for training different models with processed data, evaluating model performance, as well as understanding the important and influential features.

Results discussions/explanation – This stage is for studying the results of the whole research.

5.2. Techniques

Web scraping, Database Systems (expected), Statistics, Visualization, and Machine learning.

6. Data Source and Description

6.1. Data Source

6.1.1. Government Repositories

- Massachusetts Department of Elementary and Secondary Education (DESE). This website includes all the data of Elementary and Secondary Education (PreK to 12) in Massachusetts: district level, school level, teacher level and student level.
- The Common Core of Data (CCD). This website is administered by the National Center for Education Statistics (NCES), and it is a universe collection of public

elementary and secondary education data. This website will provide Economic and Social (income, employment, poverty) data for public school district.

6.1.2. Online Platform

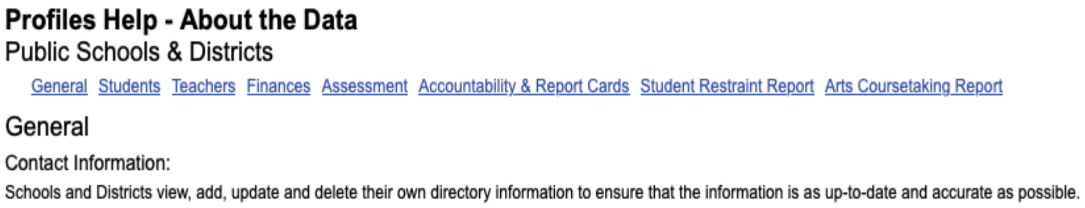
- Kaggle: Massachusetts Public Schools Data. This dataset is only for 2017, through studying it with tools: pandas, matplotlib, and seaborn, I became familiar with each variable and figured out which variables I want to cover in my dataset.

6.1.3. Research Papers

- In the process of the project, keep exploring other relevant resources such as academic research papers and reports, which can provide valuable insights, methodologies, and data for analysis.

6.2. Data Description

For DESE data, the variable definitions can be found on the ***About the Data*** section. There are eight tabs, which are: ***General***, ***Students***, ***Teachers***, ***Finances***, ***Assessment***, ***Accountability & Report Cards***, ***Student Restraint Report,*** and ***Arts Course taking Report***. See Figure 1 below:



Profiles Help - About the Data

Public Schools & Districts

General Students Teachers Finances Assessment Accountability & Report Cards Student Restraint Report Arts Coursetaking Report

General

Contact Information:

Schools and Districts view, add, update and delete their own directory information to ensure that the information is as up-to-date and accurate as possible.

Figure 1. Screenshot of Dataset Tabs

Relevant variables have been selected and classified into three categories (the list would change throughout the project).

- Situational Factors, these are the factors that school authorities have no control:
  - Basic information (10 features) *School Name, School Code, School Type, Town, State, Zip, Grade, District Name, District Code, District Type*.
  - Enrollment counts by *Grade* (PreK to 12), *Race* (African American, Asian, Hispanic, White, Native American, Native Hawaiian, Multi-Race), and *Gender* (Male, Female).

    - Student’s groups by *First Language Not English, English Language Learner, Students with Disabilities, High Needs, and Economically Disadvantaged*.
- Controllable Factors, these are factors that the school authorities have some control over.
    - Educator performance evaluation, as percentage of the four categories, i.e., Exemplary, Proficient, Needs Improvement, and Unsatisfactory.
    - Teacher Data: total number of Teachers, Student/Teacher Ratio, percentage of experienced teachers.
    - Teacher Salaries: Average Salary at district level.
    - Class Data: average class size at school level.
    - Expenditures: total expenditures and expenditures per Pupil at district level.
- Outcomes/evaluation metrics, which are used to evaluate student performance. They include Accountability, Advanced Course Performance, Advanced Placement performance, Dropout, Graduation Rates, Graduation Attending College, SAT Performance, Next Generation MCAS achievement.

With the project continuing, we will narrow down the variables to identify effective controllable variables and evaluation variables to create learning models for predictive purposes.

For the basic information: School Name, School Code, School Type, Town, State, Zip, Grade, District Name, District Code, District Type, the file and program of generating it are in this link: *https://github.com/yyyang719/MasterCapstone/tree/main/generate%20heading*

The dataset generated as one example year (2018) is in this link: *https://github.com/yyyang719/Master-Capstone/tree/main/DATASETS*

7. Data Collecting and Preparing

7.1. Data Collecting (Web Scraping)

The raw data are stored in government webpages, and it is impossible to download all of them manually. Therefore, web scraping technique is required to collect them efficiently. Scraping a web page involves fetching pages and extracting information from them. Fetching, or downloading pages, is the first step of web scraping, and once fetched, extraction can take place. The content of a page usually needs to be parsed, searched and reformatted to make data available for next step. The government websites are well constructed, and therefore for this project, we can use the most common python libraries, such as *requests*, *beautiful soups*, and *regular expression* to do most of the scraping works.

The example of a web scraping program is uploaded in the following link: *https://github.com/yyyang719/Master-Capstone/tree/main/web%20scraping*

7.2. Data Preparing

The raw data scrapped from website cannot be directly used for visualization and model building, it requires extensive work to clean and prepare the data first. There are few steps that are critical for data preparation.

1. *Data cleaning*, especially handling missing values, duplicate values, and outliers. The first two are very common in scraped data, while the outliers could be important for models, especially K-Means.
2. *Data transformation*, such as *feature encoding* for categorical data, using one-hot and label encoding, *feature normalization/standardization* to bring data to common scale, and *feature engineering* to combine and transform features to be more relevant to the problem.
3. *Feature selection*, to keep features that might be useful and remove features that could causing model overcomplex and overfitting.
4. *Pipeline process,* to combine all steps into pipeline so that we can have a consistent and reproducible data processing procedure.

One more issue that is common in real world problem is the imbalanced dataset. I will conduct some statistical analysis to check the skewness of the data and apply certain techniques to handle these issues, if exist.

8. Visualization

Data visualization is the practice of designing and creating easy-to-communicate and easy-to-understand graphic or visual representations of a large amount of complex data with the help of static, dynamic or interactive visual items such as charts, graphs, and maps. These presentations of information elaborate complex data relationships and data-driven insights in a way that is easier to understand. Data visualization can help make sense of raw data and explore, identify, and understand the patterns behind the data, it can also automate the process of finding predictive information in large databases, and can promptly identify concealed patterns.

In this project, data visualization covers from raw data mining to result visualization, and it is planned to study the educational datasets from following four perspectives:

- Temporal visualization: By visualizing data along the years from 2007 to 2022, we are trying to reveal some trends over time at different resolutions, such as state, or school district levels.
- Spatial visualizing: By displaying data across areas, we are trying to uncover some patterns that may not be obvious when examining one single district and gain a deeper understanding of the educational landscape.
- In-depth visualizing for selected areas of interest: Besides studying public school districts for the entire state, we will also select certain areas and conduct deep dive investigations that are manageable and can yield local insights.
- Interactive visualizations**:** Providing some interactive plots or maps to switch focus to display as we change the criteria.

Following are the existing examples studied from one year's data (2017) with tools: Pandas, Matplotlib, and Seaborn. This sample data will be a small portion of our whole dataset (2007 to 2022), and studying it helps me understand the data types and any potential challenges or complexities I may encounter in the future.

Figure 2 below shows a part of our districts and the number of public schools (PreK to 12) in each of them.

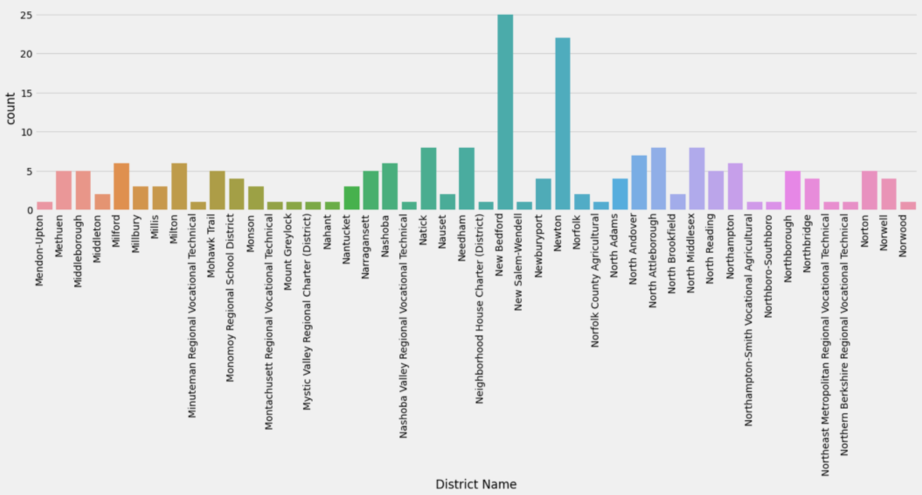


Figure 2. Public school districts with school# in them

Figure 3 gives us a sense that the distribution of % high needs has almost the same distribution of % economically disadvantaged.

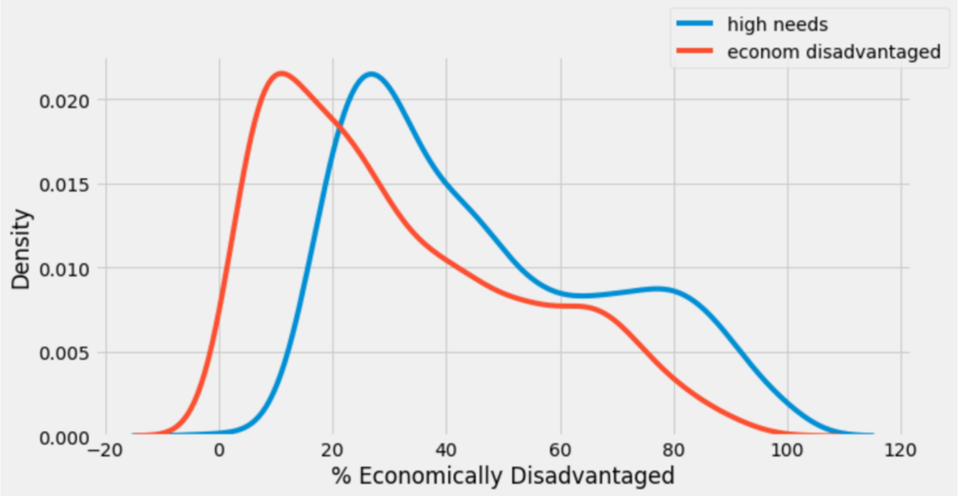


Figure 3. Distributions of % high needs and % economically disadvantaged

Figure 4 presents a relationship between assistance level (level with higher number means needing more assistance) and % economically disadvantaged.

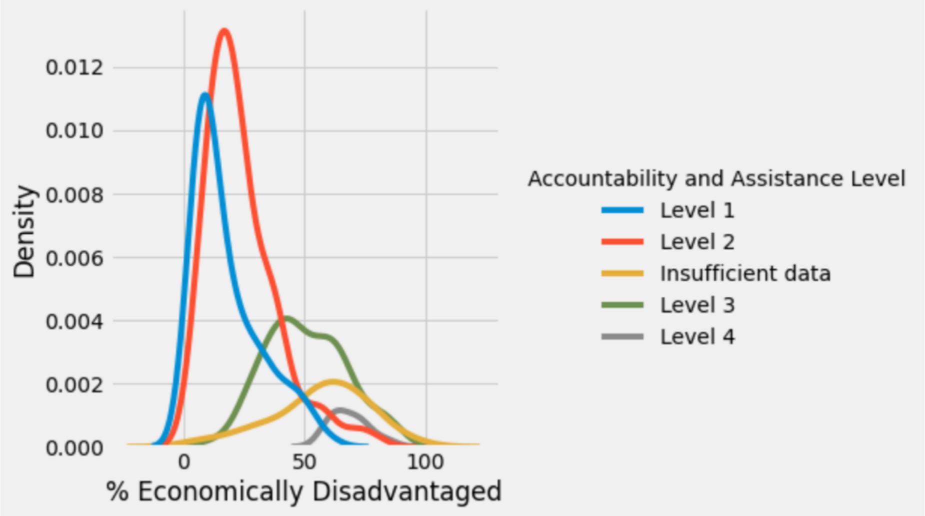


Figure 4. Assistance level vs % economically disadvantaged

Figure 5 provides an interesting phenomenon that with % Female increasing there is a growth in % 4 years college attending, however, there is an opposite relation between % Male and % college attending. It seems that male-dominated schools may result in lower academic performances than gender well-distributed or female-dominated schools.

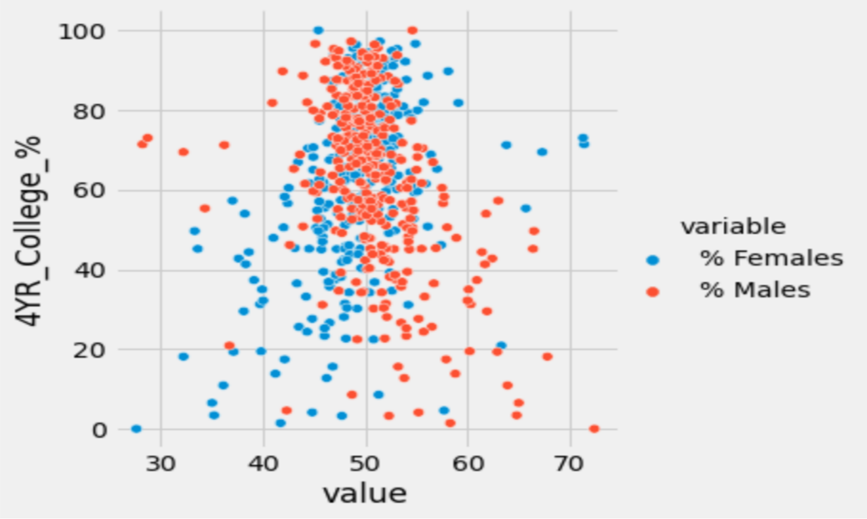


Figure 5. % Gender vs % 4 years college attending

Figure 6 may indicate that compared to traditional public schools, the academic performance for students in Charter schools could be affected less by their economic situation.

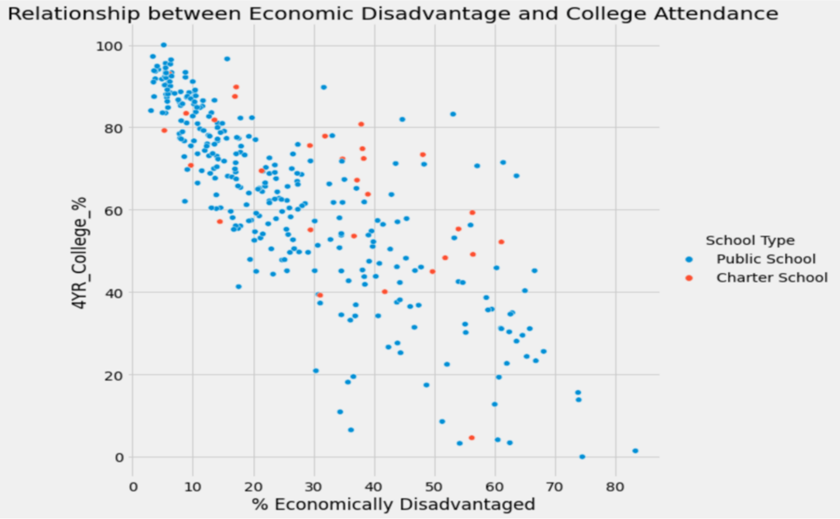


Figure 6. % economically disadvantaged vs % college attend for Traditional public school and Charter school

9. Models

After the data has been prepared, we are going to try several models and compare the performances, more importantly, we will try to explain the correlations and identify most impactful determinants with the help of these trained models. We will try three models and they are described below:

9.1. Linear/Logistic Regression Model

Linear regression is a statistical modeling technique used for analyzing the relationship between a dependent variable and one or more independent variables. It's primarily employed for predicting a continuous numeric outcome based on input features. The fundamental assumption in linear regression is that there's a linear relationship between the independent variables and the dependent variable.

It is represented in Equation 1.

$$y = b + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \tag{1}$$

The goal of linear regression is to determine the optimal values of the intercept b and the coefficients w_i that minimize the difference between the predicted values (calculated using the linear equation) and the actual target values y in the training dataset. This process is often done using the least squares method, which minimizes the sum of squared differences between the predicted and actual values.

Once the model is trained, it can be used to make predictions on new data by plugging in the values of the independent variables into the linear equation. And it is evaluated using various metrics such as the coefficient of determination (R-squared), mean squared error (MSE), and root mean squared error (RMSE), among others. These metrics help assess how well the model fits the data and how accurate its predictions are.

It's important to note that linear regression assumes a linear relationship between the variables, which might not always true in real-world scenarios. However, it serves as a good baseline and starting point for evaluating other models. In addition, the weights w_i can reveal some level of importance of the independent variables x_i, if these x_i are normalized.

9.2. Decision Tree

A decision tree (Figure 7) is a graphical representation of a decision-making process that involves repeatedly splitting a dataset into subsets based on the values of input features. It's a popular for both classification and regression tasks. It is particularly effective for tasks whose relationships between features and outcomes are nonlinear.

A decision tree consists of nodes and edges. The nodes (except leaf nodes) represent features, and the edges represent the decisions that lead to other nodes.

The tree starts with a root node and further split into multiple children's nodes or leaf nodes. Internal nodes contain decision rules, and leaf nodes represent the predicted outcomes.

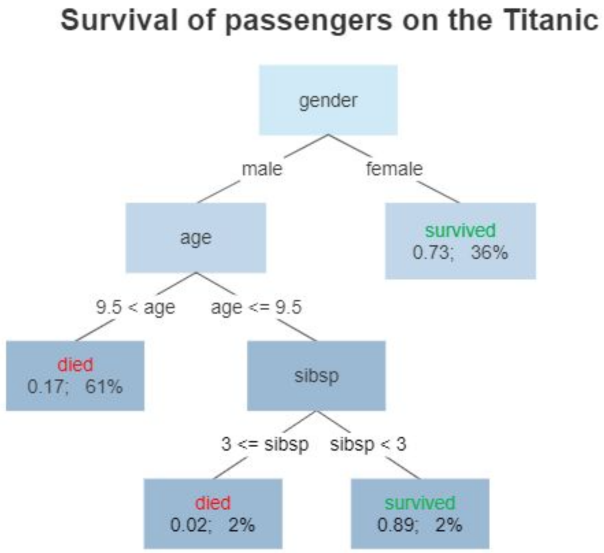


Figure 7. Example of Decision Tree (Source: Wikipedia)

At each internal node, the dataset is split into subsets based on the values of one feature. The goal is to find the feature and value that best separates the data into more homogeneous subsets with respect to the targeted outcomes. The decision of which feature to split on and with what value is determined based on methods such as Gini impurity, entropy, or mean squared error, depending on whether the task is classification or regression.

This process of node splitting is repeated on each derived subset in a recursive manner called recursive partitioning. Some commonly used approaches to stop splitting could be predefined tree depths, minimum number of samples in one node, or when the subsets are pure enough (classification) or variance is below predefined threshold (regression).

Once the tree is built, it can be used for prediction for a new instance. We can start at the root node and move down along the tree by following the decision rules at each node. Eventually, we will reach one leaf node that corresponds to a predicted outcome.

Decision trees are easy to understand and visualize, and capable capturing complex nonlinear relationships. However, they could have overfitting issue if the tree grows too deep or complex. To overcome this, techniques such as pruning (combine certain nodes) or ensemble methods like Random Forests (combine multiple simpler trees) are worth exploring.

9.3. K-Mean Clusters

K-Means clustering is a popular unsupervised machine learning algorithm that aims to partition n observations into k cluster that are distinct and non-overlapping, where each cluster is represent by the cluster center or centroid. In other words, it aims to group similar data points together while keeping the clusters as distinct as possible.

The model requires a predefined k as number of clusters and starts with k randomly initialized cluster centroids. Then for each data point, it calculates distance to each of the k centroids and assign the data point to the nearest centroid. After all data points have been assigned, it calculates the mean of the data points of each cluster and update the cluster centroid with that calculated mean. Repeat the assign and update centroid steps until the model converges, such as the changes of cluster centroids below thresholds or targeted iterations have reached (Figure 8).

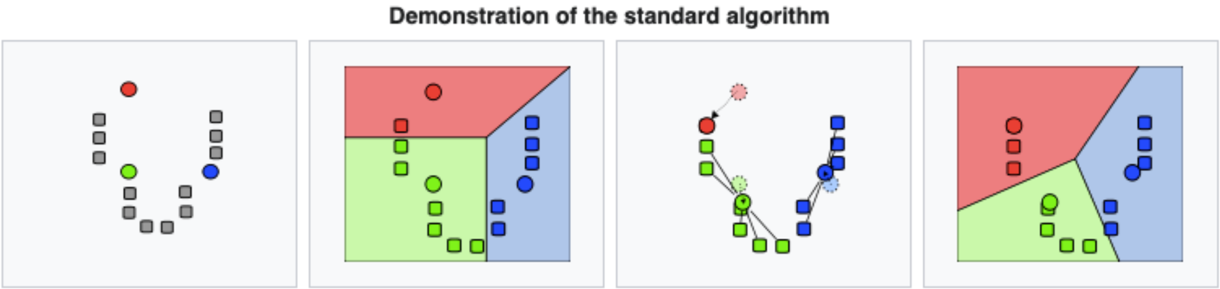


Figure 8. Example of K-Means (Source: Wikipedia)

K-Means is sensitive to initial centroids and each run might lead to slightly different clusters, if the data is truly separatable, or very different clusters, if the data is close by natural. K-Means also requires domain knowledge to determine the best cluster number k, and it is sensitive to outliers, which limits its applications.

Despite these limitations, it is very useful for data segmentation and anomaly detection, which are helpful for the purpose of this project. For example, we can group schools into several clusters based on certain features, exclude these features' impacts and conduct analysis within each cluster to identify other impactful factors.

10. Results Discussion

Once the models have been trained and validated, they will be studied to understand the inner logics of the system. They will also be used to conduct sensitivity tests to identify important features that will impact targeted outcomes.

Data will also be visualized as results, such as plots and maps. These visualizations could reveal some temporal and spatial patterns that are not straightforward by just studying numerical values within tables.

Model limitations will be discussed, as well as possible future directions.

11. Implementation Timeline

- Phase 1, 07/31, *Data Collection*: download data from multiple sources through web-scraping techniques.
- Phase 2, 08/31, *Data Cleaning and Engineering*: clean and prepare data for visualization and model training.
- Phase 3, 09/30, *Data Visualization*: display data from different perspectives through data visualization techniques, to better understand the data and observe some trends and insights.
- Phase 4, 10/20, *Model Training*: train different models with cleaned data, evaluate model performance, as well as understanding the important and influential features.
- Phase 5, 11/15, *Review Results and Discussions*: finalize project with discussions and insights derived from results.