Improving New York City School Bus Operations: Learning from 2015-2022

Jennifer Berryman

Western Governor’s University

# Contents

[Contents 2](#_Toc139399415)

[A. Project Overview 4](#_Toc139399416)

[A1. Research Question or Organizational Need 4](#_Toc139399417)

[A2. Context and Background 4](#_Toc139399418)

[A3. Summary of Published Works 4](#_Toc139399419)

[A4. Summary of Data Analytics Solution 4](#_Toc139399420)

[A5. Benefit to Organization and Decision-Making Process 4](#_Toc139399421)

[B. Data Analytics Plan 5](#_Toc139399422)

[B1. Goals, Objectives, and Deliverables 5](#_Toc139399423)

[B2. Scope of Project 5](#_Toc139399424)

[B3. Methodology 5](#_Toc139399425)

[B4. Timeline and Milestones 5](#_Toc139399426)

[B5. Resources and Cost 5](#_Toc139399427)

[B6. Criteria for Success 5](#_Toc139399428)

[C. Hypothesis 6](#_Toc139399429)

[C1. Hypothesis 6](#_Toc139399430)

[C2. Analytical Method 6](#_Toc139399431)

[C2a. Justification of Analytics Method 6](#_Toc139399432)

[C3. Tools and Environments of Solution 6](#_Toc139399433)

[C4. Methods and Metrics to Evaluate Statistical Significance 6](#_Toc139399434)

[C4a. Justification of Methods and Metrics 6](#_Toc139399435)

[C5. Practical Significance 6](#_Toc139399436)

[C6. Visual Communication 6](#_Toc139399437)

[D. Description of Dataset 7](#_Toc139399438)

[D1. Source of Data 7](#_Toc139399439)

[D2. Appropriateness of Dataset 7](#_Toc139399440)

[D3. Data Collection Methods 7](#_Toc139399441)

[D4. Data Quality 7](#_Toc139399442)

[D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulator Compliance 7](#_Toc139399443)

[D5a. Precautions 7](#_Toc139399444)

[E. Sources 8](#_Toc139399445)

# Project Overview

## A1. Research Question or Organizational Need

How can the New York City school bus operations be improved to support potential breakdowns and delays based on known factors from the 2015-2022 school year?

Many school systems in the United States maintain their own transportation, and must plan how to support changing transportation needs. Due to limited school budgets, schools would benefit from knowing factors that contribute to potential hurdles in bringing students to school safely and effectively. New York City public schools cataloged different aspects of buses that experienced breakdowns or delays between 2015-2022, which can help identify some of these hurdles and help plan for ways to improve transportation.

## A2. Context and Background

School operations are a complex component of managing a school budget to ensure that students are brought to school safely and efficiently. With schools facing budget cuts, it is important that school systems know how to optimize the allocation of resources in their transportation departments.

## A3. Summary of Published Works

Funding for public schools has been decreasing given the rate of rising inflation. It is becoming increasingly more important for schools to manage budgets for physical costs like school buildings, buses, and technology, so that they can optimize the resources they can provide for instruction.

The Center on Budget and Policy Priorities, in 2015, found that at least 31 states provided less funding per student in 2014 than 2008 (prior to the 2008 recession that restricted many state government budgets). Additionally, they found that in at least 18 states, local government funding also fell over the same period. States and local funding represent, on average, about 91% of a school’s revenue, with the other 9% being federal revenue. School reform initiatives, such as improving teacher quality, are endangered by these funding cuts, which further exacerbates schools’ abilities to fund their transportation departments (Leachman et al., 2015).

Due to the restriction of school funding, being able to identify ways to properly allocate resources within the school system would be essential to help optimize the resources the school system already has on hand. It may also provide assistance in acquiring grant funding by justifying the need for additional resources for school transportation by identifying features that could be improved with targeted funding.

The distance between home to schools has increased in recent decades, necessitating a growth in the number of students travelling to school (Easton & Ferrari, 2015). This means that more students are in need of bussing, or require their parents to drive, in order to reach school, and that these busses would need to travel longer distances. Given that funding is decreasing for schools, but the need for transportation is rising, knowing how to efficiently plan for the allocation of resources for transportation is paramount for school systems.

Another consideration that affects United States public school transportation is school choice. Public schools that compete with charter schools have to provide services that their competing charter school may not, such as offering school transportation. According to the Center for Education Policy Analysis at the University of Connecticut, public district per pupil spending was roughly $16,200 vs. $14,866 spent by charter schools in the 2015 school year for students in the state. Public schools must account for student transportation, causing them to have to spend more per student (Cotto, 2017). To optimize the school budget, schools should know how and when to allocate resources to transportation.

## A4. Summary of Data Analytics Solution

By analyzing descriptive statistics of school bus data breakdown/delay data from New York City in the 2015-2022 school year, important factors that impact breakdowns and delays can be identified. Machine learning will also be used to predict how/when/where a bus may experience a breakdown or delay. This will empower schools to provide resources optimally to their transportation departments to account for these events.

## A5. Benefit to Organization and Decision-Making Process

The data analytics solution should provide a tool for school systems to use to determine how to allocate resources with a school system’s transportation department. This will be done by identifying what factors contribute to breakdowns/delays, which would allow school systems to provide resources such as additional buses/backup drivers, tools and components such as tires for common issues, or other resources such as personnel. It may also be used to justify the need for additional funding, such as through grants, to acquire the resources necessary to promote safe and effective transportation for students.

# Data Analytics Plan

## B1. Goals, Objectives, and Deliverables

The goal of this project is to identify factors that are most often associated with bus breakdown/delays based on the dataset, and to create a model that can accurately predict the likelihood of a breakdown/delay.

To do this, the following objectives should be met:

* Determine which factors are most often associated with bus breakdowns/delays.
  + Conducting descriptive statistical analysis on the cleaned dataset will determine these factors.
* Create a model that can accurately predict the likelihood of a breakdown/delay.
  + Creating a subset of the data to create a model, then comparing the dataset to the model created will determine the success at creating a good model.
* Communicate the findings of the previous objectives into a report.
  + A final report containing the findings, breakdown of data, and visualizations will be created to communicate findings.

## B2. Scope of Project

The project is aimed at identifying the factors that contributed to bus breakdowns and delays for New York City school buses between 2015-2022, with a goal of creating an model that can predict what may lead to these instances. It will not include suggestions for how school transportation resources should be allocated, but instead is aimed at providing a tool to allow school systems to determine that themselves.

## B3. Methodology

This project will use CRISP-DM as it’s methodology to organize and implement the project using the following phases, as well as to provide structure to the final communication of findings (Hotz, 2023):

1. Business Understanding
   1. Describe business need/objectives using the business understanding provided above.
   2. The goal of data mining will be to predict the likelihood of bus breakdowns/delays for a specific route provided the bus route’s characteristics. These characteristics should be based upon the factors that lead to bus breakdowns/delays.
2. Data Understanding
   1. Collecting the data has already been completed, and documented below. It will be further explained in the final report.
   2. Describe data in detail, providing descriptions for the columns in the data, and what values may mean.
   3. Explore data and provide a breakdown numerically and visually in the final report.
   4. Verify data quality to ensure that the data is aligned with the goal of the research question, and to provide a foundation for a successful model to be built.
3. Data Preparation
   1. Selecting the data has been completed, and a description of why this data is appropriate will be contained in the final report.
   2. Clean data to remove duplicate or unnecessary values, convert data types, address missing values, and remove unwanted outliers.
   3. Format data to prepare it for descriptive statistical analysis and the create of a model to predict the likelihood of breakdowns/delays.
4. Modeling
   1. Select Modeling Techniques
   2. Generate Test Design
   3. Build Model
   4. Assess Model
5. Evaluation
   1. Evaluate results based on the success criteria described in B6. Criteria for Success.
   2. Review process and make corrections as needed.
   3. Determine next steps to see if deployment or reworking needs to be completed next.
6. Deployment
   1. Produce final report to communicate findings.
   2. Review project criteria using rubric provided to ensure that all components are met.

## B4. Timeline and Milestones

|  |  |  |
| --- | --- | --- |
| Business Understanding | 1 Day | July 6th |
| Data Understanding | 2 Days | July 8th |
| Data Preparation | 2 Days | July 10th |
| Modeling | 4 Days | July 14th |
| Evaluation | 6 Days | July 20th |
| Deployment | 5 Days | July 25th |

## B5. Resources and Cost

* Jupyter Notebook – No Cost
* Dataset - No Cost
* Tableau – No Cost
* Microsoft Word – No Cost
* 50 work hours – No Cost

## B6. Criteria for Success

|  |  |  |
| --- | --- | --- |
| Criterion/Metric | Required Component | Success |
| Which factors are most often associated with bus breakdowns/delays. | Descriptive statistical analysis on the cleaned dataset will determine these factors. | Success if clear relationships between factors and bus breakdowns/delays can be identified. |
| Create a model that can accurately predict the likelihood of a breakdown/delay. | A subset of the data to create a model, then comparing the dataset to the model created will determine the success at creating a good model. | The model predicts school bus breakdowns/delays with at least 85% accuracy when applied to the whole dataset, which would present realistic performance and not indicate overfitting. |
| Communicate the findings of the previous objectives into a report. | A final report containing the findings, breakdown of data, and visualizations will be created to communicate findings. | The report meets all criteria set forth in the rubric for task three with a score of competent. |

# Hypothesis

## C1. Hypothesis

If there is a breakdown or delay it is most likely going to be due to physical components on the school bus, such as tires, brakes, or other similar items. While other factors, such as routes, or time of day, may be related to experiencing a breakdown or delay, the physical components themselves will have a stronger relationship.

## C2. Analytical Method

The data will have descriptive statistics analyzed to determine initial findings from the data in regards to the most common occurrences of values for different factors, including the routes, boroughs, and reason for breakdown/delay. Machine learning, using ensemble methods, will then be applied to create a model that can predict the factors that will affect bus breakdowns/delays.

### C2a. Justification of Analytics Method

I chose these two methods because the goal of the project is to predict whether a route is likely to experience a breakdown or delay, and the most frequent factors combined with a model to predict the likelihood would fulfill that goal.

## C3. Tools and Environments of Solution

The following tools will be used for the analysis and final report:

* Jupyter Notebook for Python analysis and machine learning model
* Tableau for data visualization following the analysis
* Microsoft Word for the final report.

## C4. Methods and Metrics to Evaluate Statistical Significance

A subset of the dataset will be used to create a model that will predict the likelihood of bus breakdown/delays, which will then be tested on the whole dataset. Rather than using all features, a subset of features will be used in this model for efficiency. If the model, when tested, has an 85% or higher accuracy the model will be determined to be suitable.

### C4a. Justification of Methods and Metrics

An accuracy of at least 85% indicates that the model can predict school bus breakdown/delays based on the factors of the school bus without the fear of the model being overfitted.

## C5. Practical Significance

If the findings can predict the likelihood of a breakdown/delay and identify the most common factors associated with these events, it will provide practical benefits by allowing school systems to use the data to inform their decisions around spending in their transportation budgets. School transportation departments will be able to analyze their own fleet of school buses and determine what types of resources can help address any need they may see with their vehicles.

## C6. Visual Communication

Tableau will be used to create visualizations identifying the most important features affecting school bus breakdown/delays based on the analyzed dataset and to share descriptive statistics in the final report. The model created to predict the likelihood of a school bus breakdown/delay will also be graphically displayed to show which factors the model is assessing and its importance in the prediction to help convey the importance of each factor when the model makes its prediction.

# Description of Dataset

## D1. Source of Data

The New York City Bus Breakdown and Delays (2015-2022) dataset, which was created by the New York school system. When a breakdown or delay occurred for their buses they logged the information provided in the dataset. It can be found at <https://www.kaggle.com/datasets/mattop/new-york-city-bus-breakdown-and-delays?resource=download>

## D2. Appropriateness of Dataset

This dataset is the culmination of New York City’s school bus breakdown/delay catalog entries. This will provide information to help determine which factors are most related to school bus breakdown/delays. This information will be applicable to New York City schools, which is why they are the focus of this report. Its findings, however, can be used to help other school systems when evaluating their own fleet of buses and determining how they could allocate resources to support safe and efficient transportation.

## D3. Data Collection Methods

The data was gathered by New York City’s school bus operations entering information regarding break downs and delays for the 2015-2022 school years. I accessed the data through Kaggle.

## D4. Data Quality

There is some data missing from the dataset, but overall, it looks like the data has a robust amount of information. It will require cleaning to address the missing data, create consistency, and to pre-process to fit to a model.

## D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulator Compliance and Precautions

The dataset is in the public dataspace and does not contain any information that would pose a security or privacy risk, so no precautions are currently deemed necessary. The data describes specific instances of a bus breakdown delay, but do not provide information that is not publicly available under the Freedom of Information Act.

# Sources

Cotto, R. (2017). (rep.). *A Guide to Comparing Charter &* *Public School Funding and Spending in Connecticut*. Center for Education Policy Analysis. Retrieved June 16, 2023, from <https://cepa.uconn.edu/wp-content/uploads/sites/3130/2017/03/Cotto-Comparing-Charter-School-and-Public-School-District-Funding-March-2017.pdf>.

Easton, S., & Ferrari, E. (2015). Children’s travel to school—the interaction of individual, neighbourhood and school factors. *Transport Policy*, *44*, 9–18. <https://doi.org/10.1016/j.tranpol.2015.05.023>

Hotz, N. (2023, January 19). *What is CRISP DM?*. Data Science Process Alliance. <https://www.datascience-pm.com/crisp-dm-2/>

Leachman, M., Masterson, K., & Figueroa, E. (2015). (rep.). *A Punishing Decade for School Funding*. Washington, DC: Center on Budget and Policy Priorities.