Title: Rough Routes: Patterns in Bus Breakdowns

Team Members: Ben Allen

Problem Statement and Motivation:

The motivation to study pupil transportation lies in the fact that it is a major part of the national education system: many students rely on school district provided transportation as their sole means to arrive at school each day. With 480,000 school busses in operation and 26,000,000 riders across the country[1], transportation for students is an enormous endeavor with many moving parts. Delays, breakdowns, and other obstacles that hinder bus routes are a chief motivation when considering how to best operate this critical operation. Pupil transportation affects a large portion of the student population as about 55% of K-12 students use the bus to arrive at and return home from school each day.[2] Furthermore, the author has a special vested interest in the study of student transportation as he is a school bus operator (and furthermore, a relief operator that is assigned to many different routes as needed), based in Nederland, Colorado.

Could there be patterns within a vast record of breakdowns that could be identified to help lessen the occurrence or severity of delays? Are there strong correlations between certain routes (or other factors) that influence tardiness in pupil transportation? This project aims to find out.

More specifically, at the beginning of this project, the author seeks to explore simple correlations such as longest delay times against reason for delay, delays occurring at certain times of time or certain days if the week, reoccurring delays and when and where they occur, etc. Ultimately, this project endeavors to pursue correlations and discover where there are any pertinent factors in school bus delays.

Literature Survey:

Keeping with schedule is a major imperative for any transportation entity. Thus, there are a myriad of different studies in several areas to exist for improving timeliness no matter the transportation mode. Looking at a variety of transportation analysis surveys and studies, even those outside of pupil transportation, will undoubtably lend not only direction to the project by building upon existing work, but also opportunities to adopt and adapt research methodologies to this project, ensuring the project to be as effective as possible.

Below, please find a few studies that either share a dataset with this project (in part or in whole) or otherwise are in some way useful or connected with studies of transportation delays:

School Bus Transportation Services 2022
https://github.com/NewYorkCityCouncil/schoolBusDelays2022/blob/main/README.md
Description: Data and visuals for November 21, 2022 meeting of the Oversight and Investigations committee and the Education Committee of New York City to discuss severity of bus delays, the most common reasons, identity trends by route type, and recommendations for better reporting. Overall, this study shows some main takeaways from the data and will serve as a good springboard in searching for further insights. This study uses the same dataset as this project, though it focuses on specific years rather than the entire

[1] https://www.nysbca.com/fastfacts

[2] https://www.nysbca.com/fastfacts

span of the dataset. There are associated webpages to this meeting, however the github page provides to quickest access to the data and visuals that will be of most use to this project.

School Bus Breakdowns in New York City: https://www.granthaalayahpublication.org/journals/index.php/granthaalayah/article/view/36_IJRG20_B02_3076
Description: Taken from the abstract "In this paper, we examine the dataset representing bus breakdowns and delays in the New York public school system. We analyze several measures involving the companies involved in delays, the season/date of the delays, the causes of the delays and other measures. We have several conclusions and recommendations." While this paper will also serve as a good starting point, this project aims to find deeper insights than the takeaways presented here.

Data Visualizations of Bus Breakdowns in NYC https://www.kellyameredith.com/bus-breakdowns-in-nyc
Description: "This group was tasked with discovering and analyzing a dataset to create visual representations and suggest recommendations for improvement. The data reviewed consisted of records of NYC school bus breakdowns based on such information as location, route, reason, and date and time." This visuals are crafted from the dataset that this project employs. Thus, this project can study the visualizations from this work in order to gain more immediate understanding of the data, as well as give thought to which sort of visualizations that this project will use.

Using Telematics Data to Evaluate Breakdown Risk for NYC School Buses https://cusp.nyu.edu/2023-capstone-projects/using-telematics-data-to-evaluate-breakdown-risk-for-nyc-school-buses/
Description: While not a complete project, the milestones and methodologies outlined here are a good resource for this project to adapt from. The overall goal of Telematics Data is to reduce breakdowns though predictive data strategies.

Datasets:

This project employs one dataset at this time: Bus Breakdown and Delays (New York) https://catalog.data.gov/dataset/bus-breakdown-and-delays[i]
The dataset may be found at the link above. Additionally, this data set is downloaded to the author's home machine, as well as preserved in the github repository as a compressed file. Further iterations of this dataset will be available on the github repository as data cleaing and pre-processing is performed. New York City boasts the largest fleet of school busses in the country[3], making this dataset a good one to study even for numbers alone.

Data Description: 21 attributes and 654,531 entries making 13,745,151 data points. Delays in this dataset span from 11/5/2015 to 3/1/2024.

Attribute	Description
School Year	
Busbreakdown_ID	
Run_type	
Bus_No	
Route_Number	
Reason	
Schools_Serviced	
Occurred_On	
Created_On	
Boro	
Bus_Company_Name	
How_Long_Delayed	
Number_Of_Students	
Notified_Schools	

[3] https://www.nysbca.com/fastfacts

Notified_Parents	
Altered_OPT	
Informed_On	
Incident_Number	
Last_Updated	
Breakdown_or_Late	
Age_Group	

Proposed Work:

Data Cleaning:

There are several attributes that need attention in this dataset. For example, the attribute "how_long_delayed" is sometimes an integer, sometimes followed my "min" (to indicate minutes, surely) and sometimes instead given as a range of minutes instead of one value (15-30min, 60-90min, and so on) and in the worst instances the value is completely blank (which leaves the viewer to wonder if there was no time delay at all or this is simply an error of recording.)

For one attribute "Incident_number", no values are listed for any incident. It would be perhaps best to abscond completely with this attribute.

Another data cleaning task is ensuring that "Bus_company_name" is consistent so that each entry of one company matches each other and there are no inconsistencies between names of the same company.

Data preprocessing:

One step of preprocessing would be to make a list of all different possibilities for certain attribute types. For example, knowing all of the different bus companies (and which years they span) as well as all reasons for a delay will be important in considering if there is any data being missed.

One of the preprocessing steps required with this data is to split some of the current attributes into a sub set of attributes. For example, looking at the attribute"run_type", just on the first page, one example entry is "Special Ed AM Run." Here, I am seeing two different facets of the run that are part of one attribute; AM versus PM runs and whether or not pupils from special education or general education are riding. Separating out this attribute (and others like it) will give the data more flexibility which will aloow for greater insights.

On the same note, applying new attributes to the data that can be extrapolated from current data could be very useful. Specifically, I am thinking about which day of the week each of these delays occurred. While this is not an attribute, the date of the delay is an attribute, and perhaps there could be a simple script or the like that assigns a day of the week to each delay. While this might not seem significant at first, the author asserts the certain routes will commonly be affected by weekly traffic patterns.

Data integration:

Because only one data set is currently being employed by this project, integration is not a concern currently. The enormous amount of data points means that the data cleaning and data preprocessing should be plenty to occupy the project. While other data sources could be very interesting and add further complexity, at the moment this one data set gives more than enough material to offer insight and exploration.

Evaluation Methods:

Evaluation methods will be drawn from course material 4502 Data Mining. The relative strength or weakness of correlations, and how to find them, is one of the core tenants of the course. By first processing the data and identifying larger trends, evaluation methods can later be more solidly established.

Tools:

Python – A powerful and effective language for data processing.

Excel – Being able to view the data all at once can be very effective in answering some questions that may be more difficult with SQL or the like.
Tableau – A visualizations tool that enables easy access to multiple kids of visualizations instantly simply by entering in a data file (in this case, a CSV.)
SQL-Entering the data as a database may be a good way to answer certain questions. SQL is a simple and powerful way to look at different attributes quickly and also compare them in a way that is not always possible in a spreadsheet format.

Milestones:
Data Cleaning is the nearest and also the most straightforward task. Next, preprocessing the data in some of the ways described above. Only then can exploration techniques be applied and correlations discovered. Further, continuing research on existing studies on this same data set will certainly be worthwhile.