Title: Rough Routes: Patterns in Bus Breakdowns

Team Members: Ben Allen

Description: Pupil transportation 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. 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.

Prior Work: 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. However, as a school bus driver myself, I am familiar with the subtleties of student transport and hope to bring my knowledge with the data set in order to discover insights.

Datasetes: The Bus Breakdown and Delays (New York)
https://catalog.data.gov/dataset/bus-breakdown-and-delays
Downloaded on Ben Allen's home machine,, which is backed-up.

Proposed Work:
Data Cleaning-Find some way to approximate dates. School years are given for each delay, dates are removed. Some attributes are inconsistent. 'schools serviced' for instance, provides a bewildering range of numbers. I would be difficult to even preprocess these values without first having a small idea of what they represent.

Data preprocessing-The time of delay is wildly inconsistent, written in as "10 mins" in some places, a single integer in others, and '1 hour' in other places. This needs some heavy preprocessing to make sure this attribute will be suitable for analysis.

Data integration-because only one data set is currently being employed by this project, integration is not a concern currently.

Tools: Python
Excel
Tableau
Weka
KNIME

Evaluation: By finding strong correlations and patterns in the data not readily obvious.