# Overview

Primary and secondary education in the United States is a large system. As displayed in Figure 1 below, the total spending on primary and secondary education is comparable to the federal defense budget.

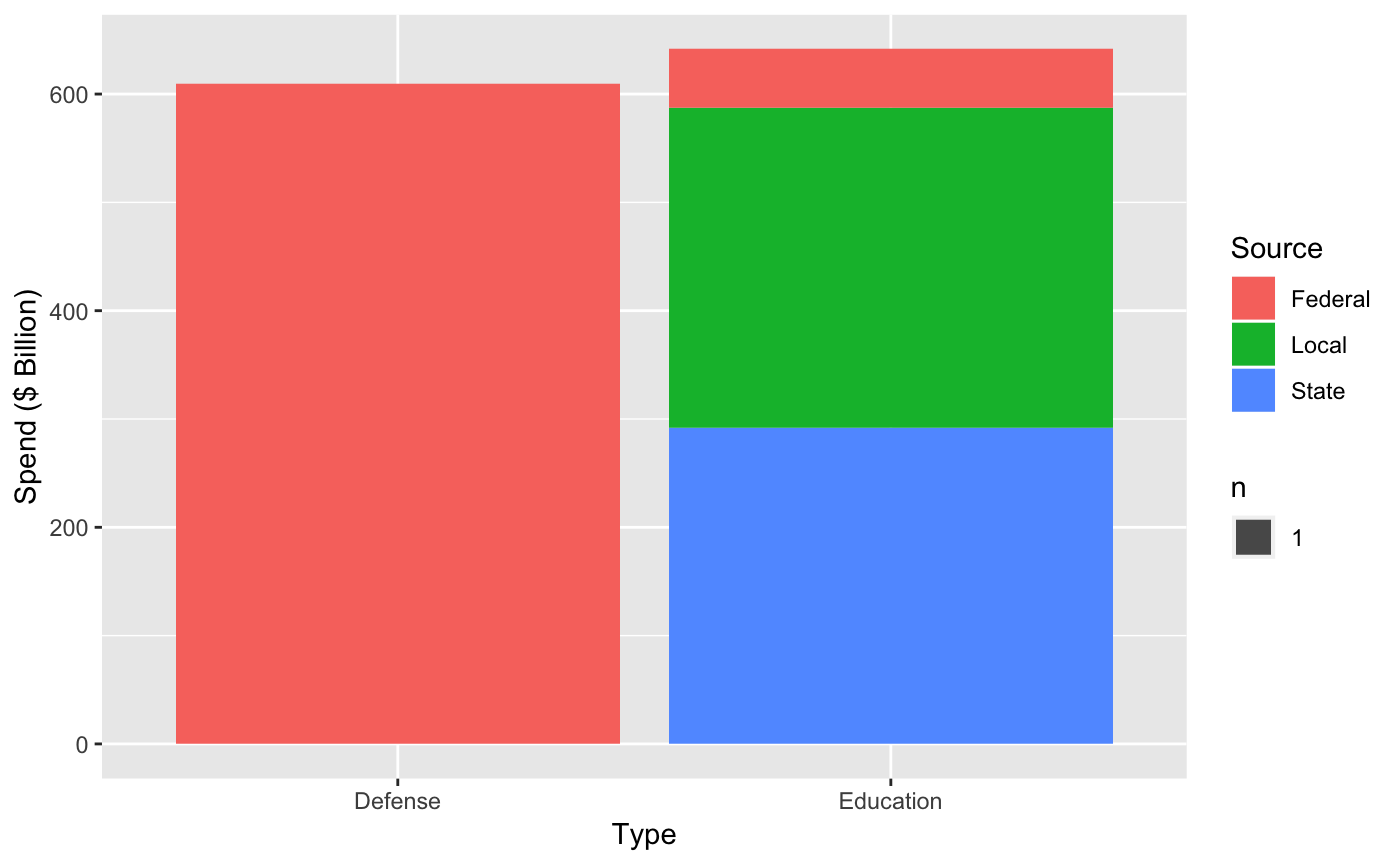


Figure : Total Education Spend and Federal Defense Spend were comparable in FY2014

Given the scale of the operation, incremental improvements have a large potential impact. Of the 18,465 school districts in 2009-2010, only 1,110 were closed by 2014-2015 (six percent), but the total cost of the closing school districts was over $4 billion.

The National Center of Education Statistics (NCES) manages the Common Core of Data, which is a publicly available database of information about primary and secondary schools used by the US Department of Education. The data dates back to 1990, and it contains detailed information about school districts’ financials and demographics. In this project, the Common Core of Data will be used for a data-driven approach to identify at-risk school districts.

# Introduction

The purpose of this project is to predict which school districts will close within five years. The final deliverable is a machine learning model that can be used by government (local, state, and federal) and non-profit organizations to better identify at-risk school districts and implement interventions tailored towards those school districts’ needs.

# Project Description

This project involves three main tasks:

1. Acquire and preprocess the Common Core of Data.
2. Perform exploratory data analysis to identify factors correlated with whether school districts close.
3. Create and tune machine learning models that predict whether school districts close.

# Work Completed

The tasks below have already been completed.

## Acquire and preprocess the Common Core of Data.

Only the 2009-2010 and 2014-2015 Common Core of Data were used for this project. Data for years after 2015 are still being completed by the NCES. Given that labels represent information five years into the future, the 2009-2010 data are the most recent collection of labelled data available.

The dataset is stored on NCES’s website as delimited files. Files were organized by year and type of data they contained (i.e. fiscal data 2009-2010). Along with each file, there was documentation that provided the character delimiter, column specifications, and number of records. Each file was downloaded onto a local folder (data) and renamed with a naming scheme (filetype\_fiscalYYYY.txt.zip).

After each file was downloaded, they were processed by a Python script. The main processing steps were:

1. Encode missing and irrelevant values.
2. Aggregate school demographic data into school district demographic data
3. Create label representing whether school district is still operational in five years.

## Perform exploratory data analysis to identify factors correlated with whether school districts close.

Exploratory data analysis was completed using R programming. The main steps involved were:

1. Remove columns with many missing or non-relevant values.
2. Modeled distribution of remaining columns.
3. Tested columns for correlation with whether school districts close within five years.

The numerical features that are correlated with whether school districts will close within five years are: total\_students, total\_schools, teachers\_total, total\_revenue, total\_federal\_revenue, total\_state\_revenue, total\_local\_revenue, total\_expenditure, total\_salaries, and minority\_students. The categorical features that are correlated with whether school districts will close within five years are: lowest\_grade, highest\_grade, metro\_micro, and charter\_status.

# Work Not Yet Completed

The task below has not yet been completed.

## Create and tune machine learning models that predict whether school districts close.

Because the data is imbalanced (only six percent of 2009-2010 school districts closed in 2014-2015), accuracy is not an adequate evaluation metric. Recall and precision will be used for model selection, and recall will be prioritized over precision based on the assumption that false positive errors are more costly than false negative errors.

Currently, the only model that has been completed is a k-means clustering model. Even without using labels, clustering identified school districts more likely to close within five years. In addition, the clusters at higher risk were also significantly smaller than other clusters. Recall was 0.398 and precision was 0.411. These observations suggest that supervised learning methods will yield greater results.



Figure : Even without using labels, clustering identified school districts more likely to close.

The following iterative process will be used to select models from supervised learning methods.

1. Split data into train-development-test sets.
2. Build model using sklearn pipelines.
3. Evaluate model (calculate recall and precision on development set).

Through this general process, the following methods will be tested.

* Logistic Regression
* Gradient Boosting
* Automated Machine Learning (Tree-based with TPOT library)

This is not an exhaustive list, and additional models will be tested during the process based on results from previously tested model.