Using Data Mining to Evaluate Colorado Public Schools Performance

Aaron Holt, Anas Salamah, Hui Soon Kim

Project Overview

- What makes a ‘good’ school?
- What makes a ‘bad’ school?
  - Initial data from a Kaggle competition
  - Combine with census data

- Schools want their students to improve!
- Large amounts of data with little analysis
- End goal to use uncovered trends to help educators improve schools

Proposed Work

Initially process and collect data:

- Colorado school data comes in multiple files with different attribute types
- Census data comes in a different format

Frequent pattern and trend analysis:

- Which attributes correlate to positive academic achievements?
- Negative achievement?

Classifier:

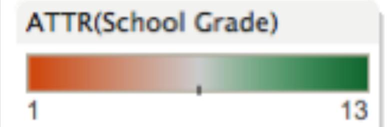
- Colorado’s population is growing and new schools continue to be built
- Predict how well schools will rank based on location and other attributes

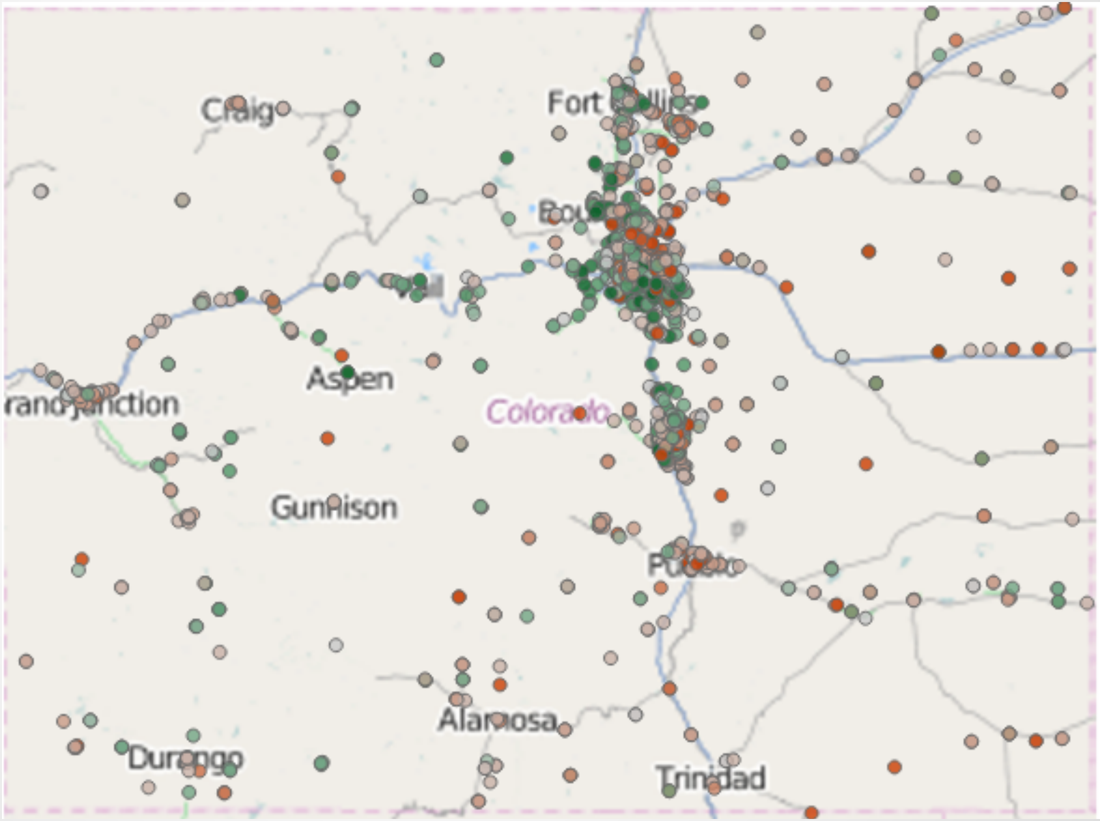
Challenges

Data processing and integration:

- Census data not ‘user friendly’
    - Hundreds of tables, separate tables give location and attribute information
    - Effectively have to integrate 3 tables for one attribute graph
    - Thousands of attributes. Which attributes are important?
    - Tedious, took more time than expected
- After preprocessing census data, we had to integrate with Kaggle data.
    - Deal with missing values etc.

Results So Far: Locations





Data Preprocessing : Used Data

School Data(17 Separate CSV files)	Census Data(8 Separate CSV files)
Final School Grade Free and Reduced Meal Enrollment School Address School GPS, etc.	Family Income Single Parent Home Race Distribution Education Level

Ex) Attributes for Final School Grade(2011)
School Name, rank_tot, Overall_ACH_Grade, Read_Ach_Grade, Math_Ach_Grade
Write_Ach_Grade, Sci_Ach_Grade, etc.

Ex) Attributes for Signle Family Home
ZIP_CODE, In married-couple_families, Female_householder_no_husband_present,
Male_ householder_no_wife_present, etc

Data Integration and Cleaning

Data Integration Result

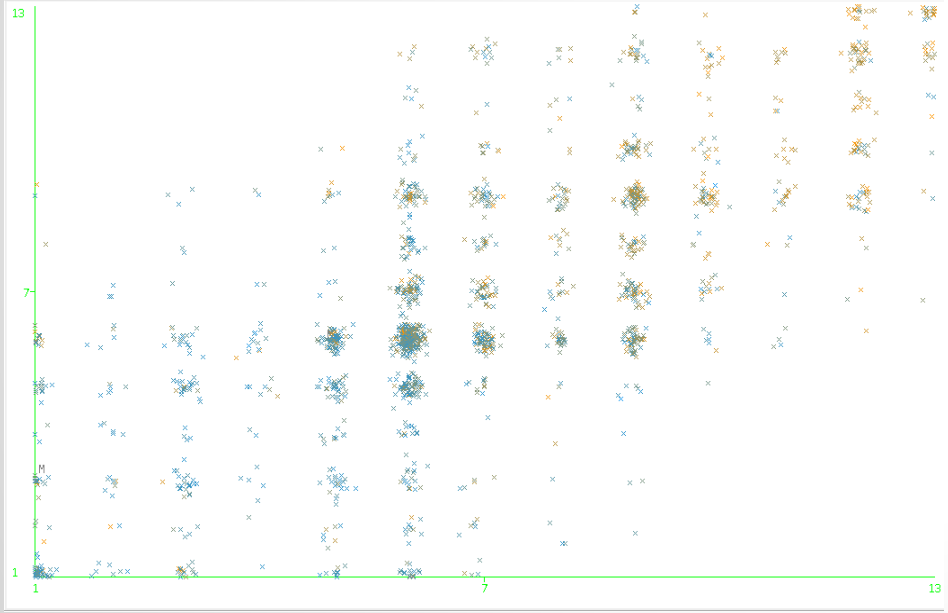
Year	Number of Object	Attributes(Total number : 49)
2011	2,061	School Name, School Grade, Student’s Subject Grade, Race Distribution, Free and Reduced Meal, Annual Income, Single Family Home, Education Level, etc.
2012	1,962	

Data Cleaning Result

Year	Total object number used (After Data Cleaning)
2011	1,814
2012	1,812

Data Reduction and Discretization

[Scatter Plot(X:Math Grade, Y:School Grade)]



[Correlation Coefficient
Between School Grade and Each Student Grade]

Subject	Correlation Coefficient
Reading	0.72
Math	0.71
Writing	0.73
Science	0.69

Frequent Patterns (2012)

Apriori Specifications	
Metric	Lift
Minimum Lift	2.0
Minimum Support	0.075

Attributes	
School Grade Pct American Indian Pct Asian Pct Black Pct White Pct Pacific Islander	Pct free meal Pct single family Pct family income $60k+ Pct Bachelors+

Strongest Patterns:

Good school? Lift = 3.2

School Grade = 9.5+	PCT Hispanic = 0-0.110	PCT Free Meal = 0-0.2

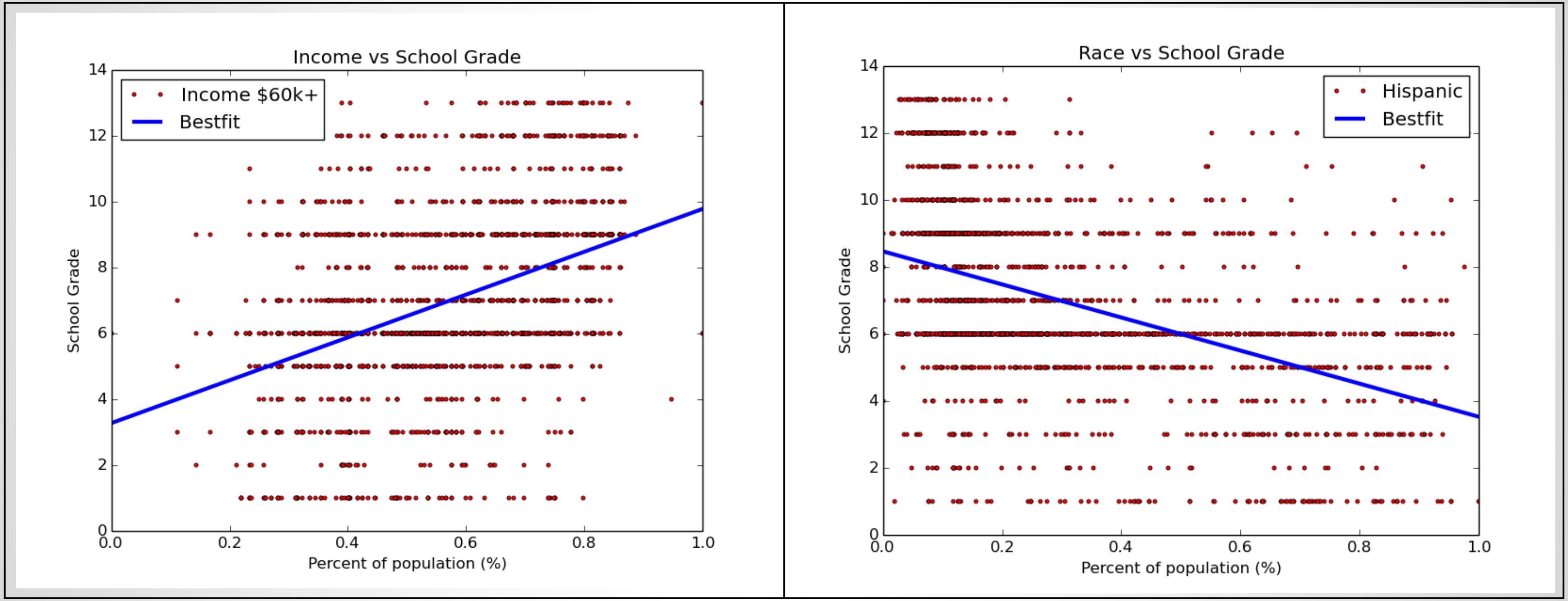
Average school? Lift = 2.77

School Grade = 6.5-9.5	PCT $60,000+ = 0.7+	PCT Free Meal = 0-0.2

Bad school? Lift = 4.13

School Grade = 0-5.5	PCT Hispanic = 0.45+	PCT White = 0-0.38	PCT Free Meal = 0.65+

Results So Far: Verification



Remaining Tasks

Finish mining 2011 data:

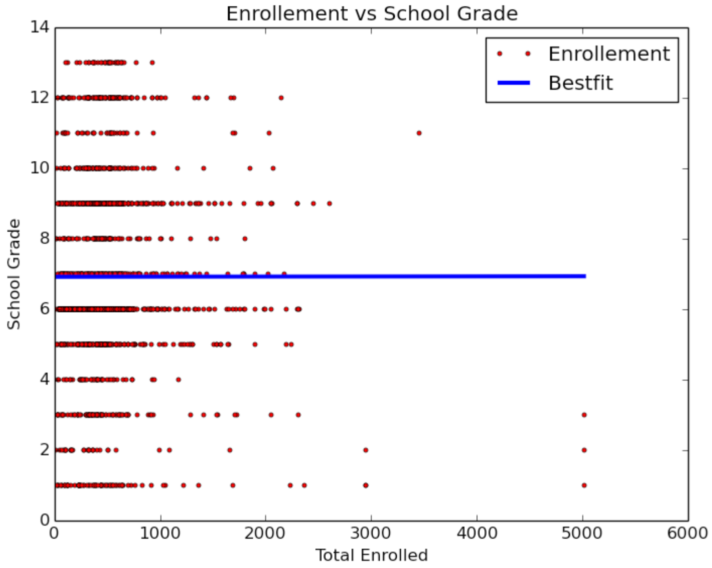
- Already have helper functions from 2012 dataset
- Compare and contrast results to 2012 data

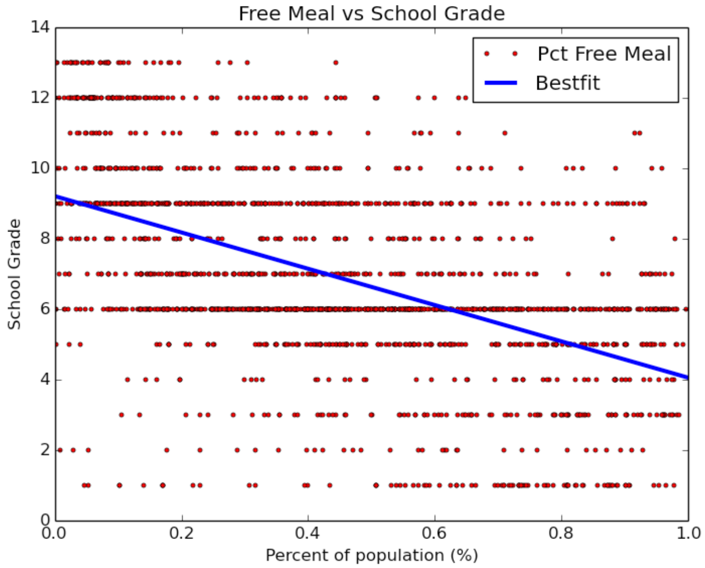
Classifier:

- Split data into test and training sets
- Pick a classifier
- Feature engineering
  - Already have many features after frequent pattern analysis

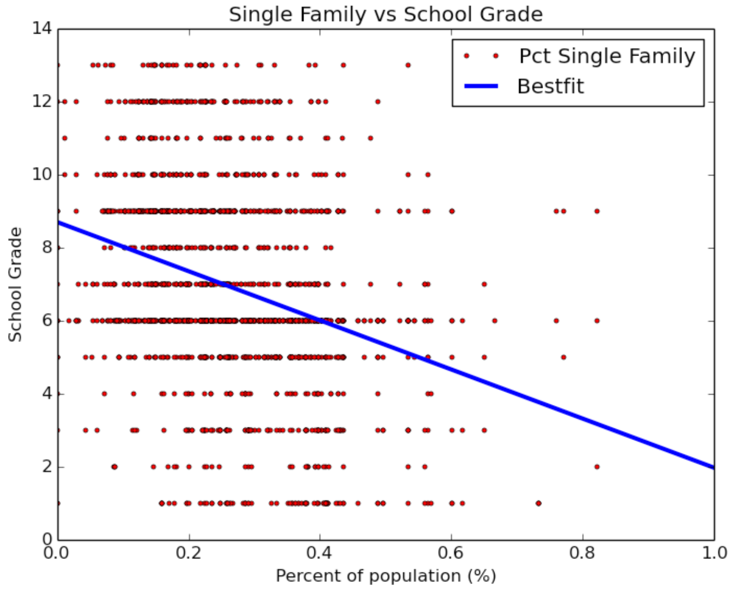
Questions?

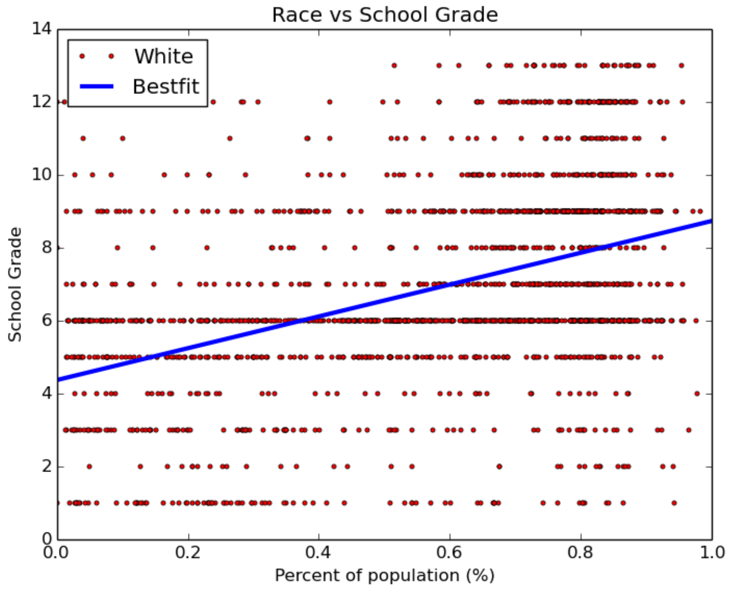
Additional Graphs



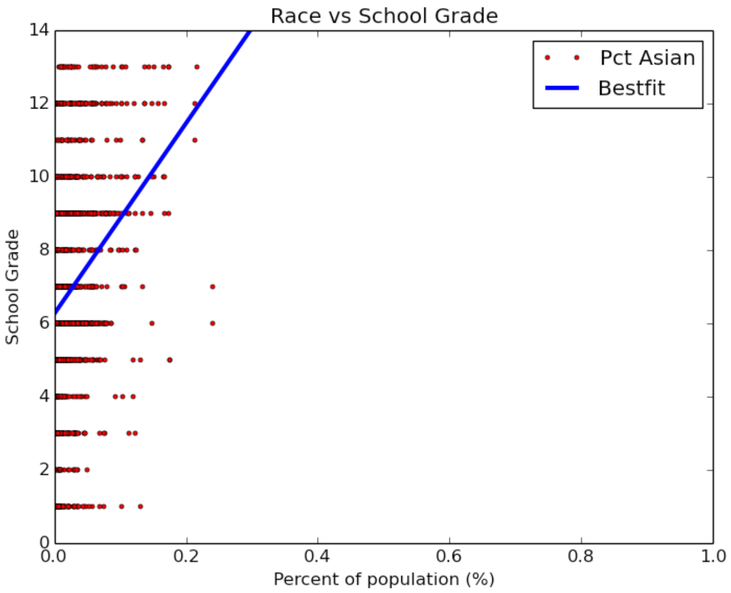


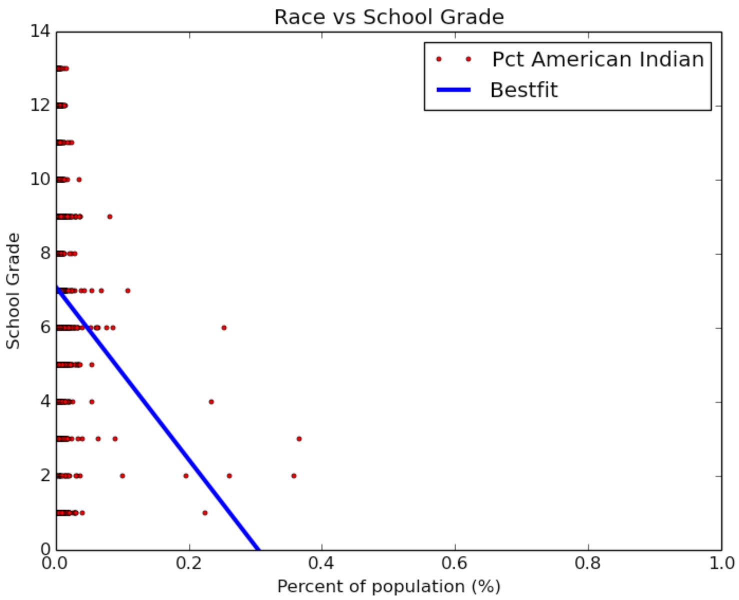
Additional Graphs





Additional Graphs


Race vs School Grade
School Grade
Percent of population (%)
Pct Asian
Bestfit


Race vs School Grade
School Grade
Percent of population (%)
Pct American Indian
Bestfit