COVID-19 & Educational Outcomes

Team #: 125

1. **Berend Dumas;** GTID# 903737105: Lead Data Science & AI at a consultancy company based in Amsterdam. BS/MS in business and informatics. Currently working on a simulation case for a public transporter in the EU. Skilled in Python, React.js, and R. Full stack data scientist
2. **Harry "Gill" Potter** GTID# 903743854: 20+ years of working experience in technology development, with 15 of those years in analytics working with healthcare payers, providers, pharmaceuticals, and digital health applications.
3. **Kristina Linn** GTID# 903841121: Product Line Manager in the semiconductor industry; BS in Applied Math with Data Science from Cal Poly. I have worked on multiple analytics projects at work as well as in school, including modeling traffic flow, modeling substance use in a small community, modeling part usage, predicting housing costs, creating GPS
4. **Megha Joshi** GTID# 903759623: Former Senior Data Scientist at Credit Suisse working in the Regulatory Anomaly Detection group under the Investment Banking Division. Received a bachelor's degree from Carnegie Mellon University, majoring in Statistics and Machine Learning. Outside of Data Science enjoys art, hiking, going to the beach, and history.
5. **Sang Park (Team Lead)** GTID# 903849582: Data Analyst in the semiconductor company; BA in Education at Hanyang Univ in S.Korea, AS in Computer Science at Fullerton College. I've worked on some data analytic projects, such as estimating supply/demand for tennis courts in a city and researching success factors for new films. Deeply immersed in tennis and camping. Also, I have a keen interest in History, Architecture, and 19th-century philosophy.

Background

Primary Research Question: Quantify the impact on (measures of) educational outcomes controlling for socioeconomic factors, historical performance, and government restrictions. This analysis is **NOT** analyzing governmental policy as informing morbidity or mortality outcomes; we focus on educational outcomes.

Supporting Research Questions:

1. Did COVID-19 have a disparate impact based on student grade level?
2. How granular can we get on the impact of education? Can we see a difference in math vs. reading? Can we see within math/reading-specific topics?
3. Can we understand how time impacts any effect we detect on educational outcomes?

Business Justification: This analysis has several applications. The most obvious is to help governments and non-profits understand how COVID-19 impacted different communities. Our analysis will examine the impact of various governmental choices, controlling for socioeconomic factors. This would help governments and NGOs decide how best to support a community regarding educational outcomes during future pandemics.

From a commercial perspective, this can also be used to identify market opportunities for new product development or applications of existing products. Armed with this analysis, product development could design new features or products that support future classrooms with targeted interventions, keeping communities on stable footing in education. Additionally, it can help sales teams convince organizations to adopt new products to better prepare for future pandemics.

Data Sources We are relying on two primary data sources. The dependent variables come from a US federal government standardized testing program, reported publicly at the state level. Our predictors/features come from a Google data set assembled from many disparate data sources.

The "Nation's Report Card" is a set of standardized tests given to public and private school students in every state. The US federal government provides state level standardized scores for various subjects each year for 4th, 8th, and 12th grades. For this analysis we are using the reading and math tests for grades 4 and 8. Because of COVID, the test data is not available for 2020 and 2021. Therefore we will be using data for 2019 (pre-COVID) and 2022 (post-COVID).

The Google Data runs from 1/1/2020, slightly before the pandemic, to 9/15/2022. It is an extensive data set that combines data sets from many public data sources across the globe and combines it with internal data from Google on search. For this analysis we are focusing on the US data, which has a line of data for each day, for each locality across all 720+ data elements. Those elements are pulled from a variety of sources, cleaned and aggregated into a single file. We start with the file and winnowed down the data based on our needs. The predictors are grouped into the following types:

Group	Description
Demographics	Various, as of 2022, population statistics
Economy	Various, as of 2022, economic indicators
Epidemiology	COVID-19 cases, deaths, recoveries and tests
Emergency Declarations	Government emergency declarations and mitigation policies
Geography	Geographical information about the region
Health	Health indicators for the region
Hospitalizations	Information related to patients of COVID-19 and hospitals
Mobility	Various metrics related to the movement of people
Search Trends	Trends in symptom search volumes due to COVID-19
Vaccination Access	Metrics quantifying access to COVID-19 vaccination sites
Vaccination search	Trends in Google searches for COVID-19 vaccination information
Vaccinations	Trends in persons vaccinated and population vaccination rate regarding various Covid-19 vaccines
Government response	Government interventions and their relative stringency
Weather	Dated meteorological information for each region
World Bank	Latest record for each indicator from WorldBank for all reporting countries
epidemiology by age or sex	Epidemiology and hospitalizations data stratified by age or sex

Building the data sources

This will be a common theme in our report, the data work in this project was pretty extensive. As you will see, we started with a gigantic data set from Google that was time series in nature and global in scope. We then combined that with state/year specific measures of educational outcomes pre/post pandemic. While the ending data set isn't huge, the amount of pre-processing was significant. As always, the data work ends up being way more time than the modeling

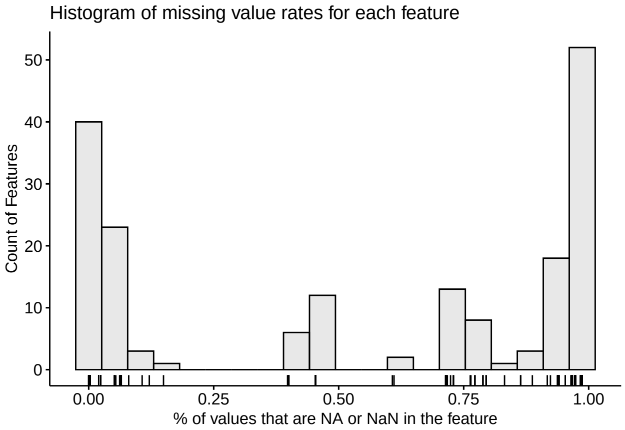
The raw Google data includes sources from all over the world at various levels of detail. For this project we are focusing on the US data. After initial cleaning of the raw data to US only, we have 719 columns and 49,550 rows. There is a mismatch between our dependent variable which is measured annually, in this case 2019 and 2022 and our predictors. As mentioned the candidate predictors are entered every day for every geography. There are obviously a lot of holes in that data that we needed to correct.

Not surprisingly, the Google Data has very good data on search terms. There also seem to be a lot of predictors that are all NA, as the data is really only available at the national level. Therefore we first drop all the predictors that are just NA's.

This process removed 95 columns from the full data set. After evaluating the many search related columns, we couldn't find a business perspective to keep them. There may be insights to be gained about how people are searching the web that could inform marketing channels for future products.

There where 421 columns related to search that were removed (Thanks Google!). There are also many variables that are simply state level features that are repeated every day. These are things like state populations, sub-populations, and other state level measures.

We found 50 in the remaining data set that were state level repetitions. Some of these can be just a single state reporting, for example `new_confirmed_age_0`. Our next sieve came to understanding the missing value percentage. Not all measures in the data set are captured for all states on all dates.



The histogram show us that there is a barbell distribution to the missing value rates. This is largely driven by COVID reporting starting in different states and different attitudes to reporting data by each state. There are roughly three types of features, low or less than 15% missing values, a group clustered around 50% missing values and those above 70% missing values. For this analysis we focus on those features with less than 15% missing values. That leaves us with 67 features to work with. Of those remaining, there are 35 that have interpretable meanings.

For example, it is hard to build products that are responsive to the dew point, but we can gather meaning from state level policy choices or mobility. Choices like stay at home requirements can help decide to focus on more remote learning options. Features around mobility changes can help us understand where to target different promotions. There are 8 population demographic measures, 16 policy measures, 6 mobility and 5 morbidity/mortality measures.

The policy measures, are categorical variables with 4 levels. A zero indicates that there was no policy in place, up to 3, indicating that the policy was strictly enforced. The mobility measures are percentage changes in the visits to different locations (parks, transit stations, residential, etc.). The population, morbidity and mortality measures are M/F %, infections, hospitalizations, etc. All of these measures have some missing data, but relatively low as some states did not report at all some measures or some measures where not captured for the full time period. We ultimately had to drop RI from the data set, as it had very incositant reporting in the Google data set.

The biggest challenge in marrying these data sets has been the change in time measures. To better summarize the predictors, we created means over the entire time period. Since each variable is reported each day, we can create a mean of each variable to use as a predictor. This also helps with the scaling between the variables. Since our dependent variable will be a "% change" we wanted to model with variables on a similar scale.

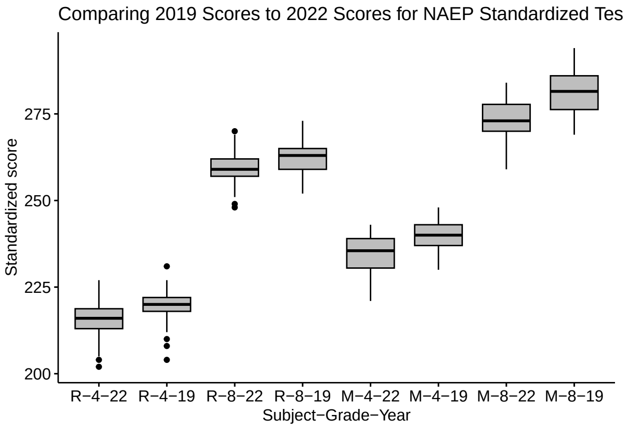
In future analysis, including more details in the predictors like standard deviation, or other transformation could yield more insights.

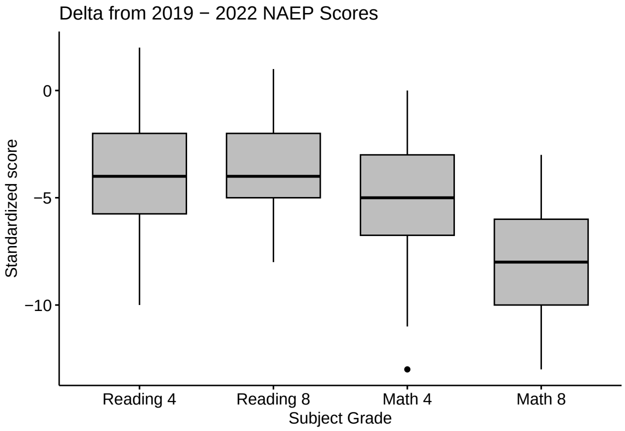
Dependent Analysis

The NAEP data sets represent 50 US states. This is fewer than the number of US "states" captured by

Google, we decided to drop those geographies (think Guam, D.C. or Puerto Rico). We built dependent variables that are subject (Math or Reading), grade (4th or 9th) and year (2019 or 2022) specific by state. From those variables we built difference measures (2022 - 2019) to test in modeling. For each state we had reading and math scores for 4th and 8th graders for 2022 and 2019.

When we look at the distributions of the dependent variables, we see a negative mean difference in all measures between 2019 to 2022.





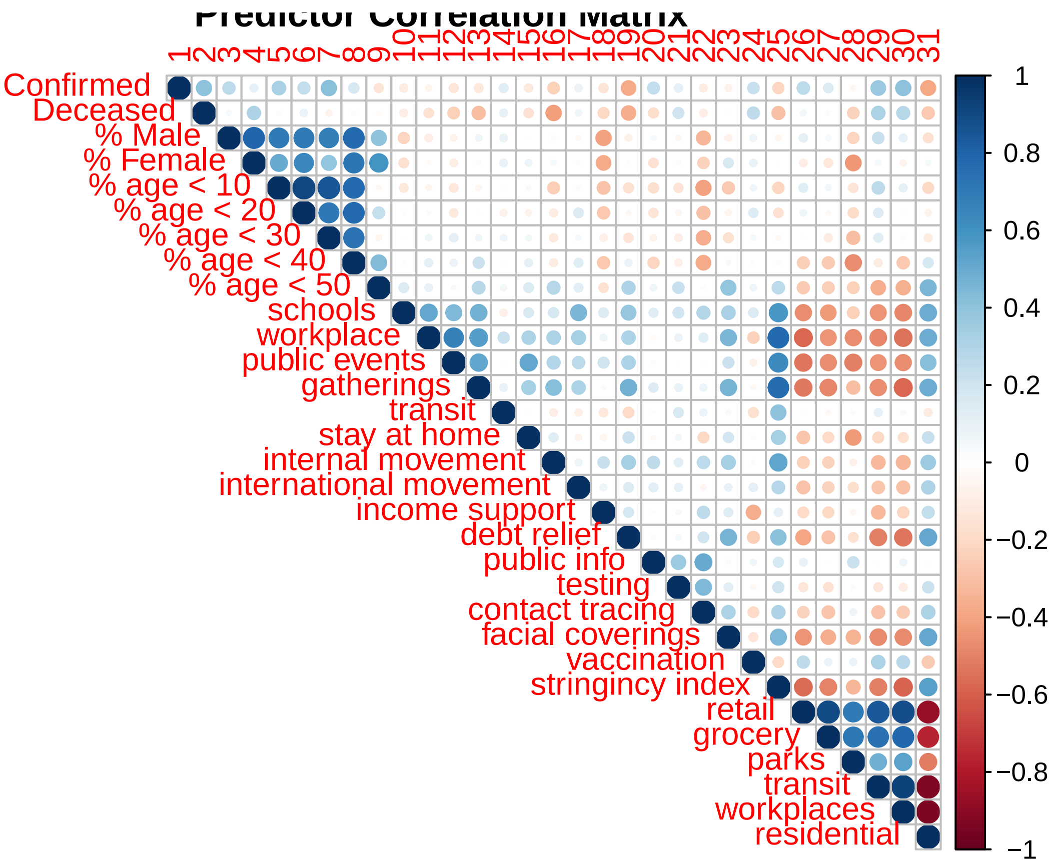
This is expected given the literature search and is in line with our initial hypothesis.

Modeling

Since this is intended to be an interpretable model rather than predictive, we chose LASSO for predictor selection, using Adjusted R^2 to compare **between** different models. We broke the predictors into three groups.

1. Population (All of the demographic, morbidity and mortality data)
2. Policy (All of the policy measures captured by Google, such and masking, transportation, etc.)
3. Mobility (Not surprisingly all of the mobility measures from Google)

Next we looked at correlations within each predictor group.



There is correlation within the above listed groups, so this may impact our LASSO methodology. LASSO will tend to reduce the impact of variables that are potentially explanatory and under represent the impact of each potential predictor. Future work could work to use PCA to help reduce the colinearity. Given our data set and questions, we decided to build 4 models for each dependent variable using different groups of predictors. We earlier grouped the predictors into Population, Policy and Mobility, to that we added “All Predictors” as a forth option. In our modeling process, we decided to use Adjusted R^2 to compare between models, as LASSO would use AIC to best choose predictors within each model. This will help adjust for models with different number of predictors.

We then ran each each model for each outcome and selected those models with the highest Adjusted R^2.

predictors	dependent	adj_rsq
mobility_predictors	perc_diff_math_grade4	0.1142958
all_predictors	perc_diff_math_grade8	0.1606242
all_predictors	perc_diff_reading_grade4	0.1174897
all_predictors	perc_diff_reading_grade8	0.3487635

For the most part all_predictors have the best performance given adjusted R^2 as a measure. The overall Adjusted R^2 is not large. From there we looked at each of the best models, the different predictors and the corresponding βs.

Table 2: Math Grade 4

	Beta
(Intercept)	-0.009873
mean_mobility_residential	-0.001926

Table 3: Math Grade 8

	Beta
(Intercept)	-0.0513995
perc_population_age_00_09	0.0110615
perc_population_age_20_29	0.1890057
mean_facial_coverings	-0.0024086
mean_mobility_grocery_and_pharmacy	0.0000142
mean_mobility_residential	-0.0001554

Table 4: Reading Grade 4

	Beta
(Intercept)	-0.0432062
perc_cumulative_confirmed	0.0653836
mean_stay_at_home_requirements	0.0134811
mean_mobility_parks	-0.0000155

Table 5: Reading Grade 8

	Beta
(Intercept)	-0.0320277
perc_population_male	0.0124696
perc_population_age_20_29	0.2165470
mean_school_closing	0.0082048
mean_public_transport_closing	-0.0032023
mean_contact_tracing	-0.0030960
mean_vaccination_policy	-0.0095232
mean_retail_and_recreation	-0.0001395

These coefficients give us an interesting insight on products to create for reducing drops in test scores during a pandemic or to have on the shelf for other emergencies. In every model, the intercept is negative. Meaning we should expect a drop in scores in either subject or in either grade. What was more interesting is the drop in scores for Math were tied to not being in school. There are negative βs in both 4th and 8th grade math models in "mean mobility residential". This indicates, more time spent at home, holding everything else constant, means lower math. This is somewhat off set in the 8th grade model for math, if the population is generally younger (higher β for percent of population from 0 -> 29).

Reading seems to offer a somewhat different story. Here the βs for mobility out of the house trips like parks, combined with more stringent rules was a negative to reading, but school closings oddly lead to lower loses.

These models show that, while reading is impacted by school closures and more strictures, math is the area where there is opportunity for more distance learning tools. We also see that as students get into the 8th grade, having more people closer to their age in the general population is helpful. This points to a product possibility for infection safe activities that focus on reading. Think of supporting library programs, drag queen story hours or specific products to help teens read at home.

We think more complex work should be done to better understand the offsets in the United States. This could mean more complex models that use the respondent level data or the development