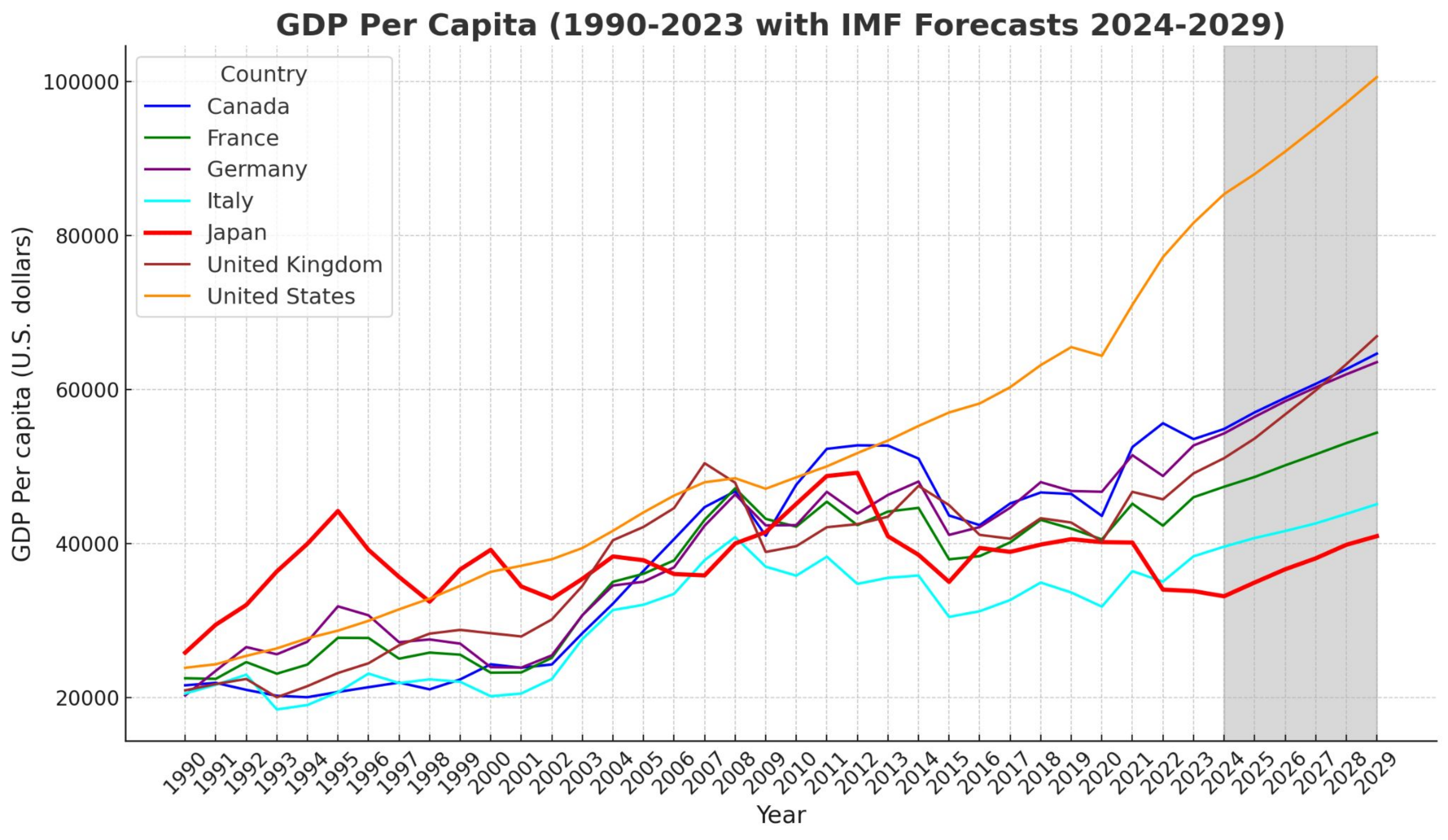
Data Analytics Using Python

Final Project

John Wu

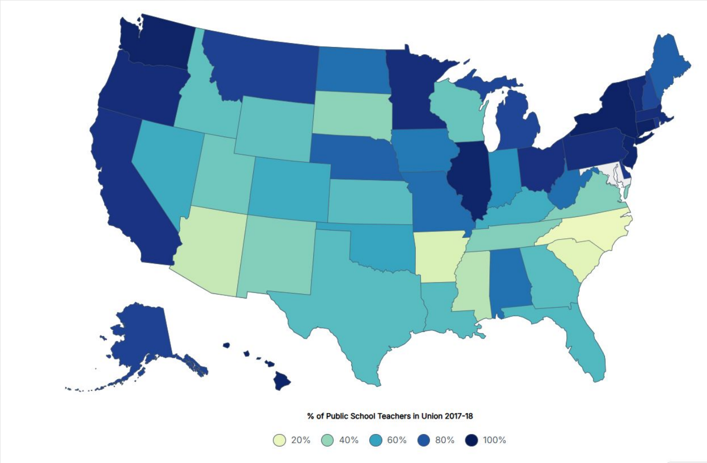
01

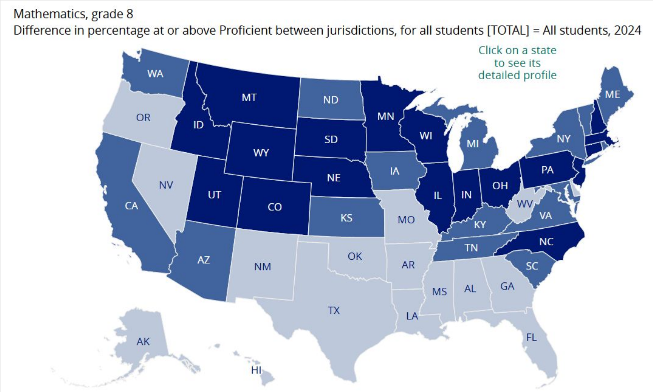
The Problem


GDP Per Capita (1990-2023 with IMF Forecasts 2024-2029)
Country
Canada
France
Germany
Italy
Japan
United Kingdom
United States
GDP Per capita (U.S. dollars)
100000
80000
60000
40000
20000
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
Year

02

Investigation

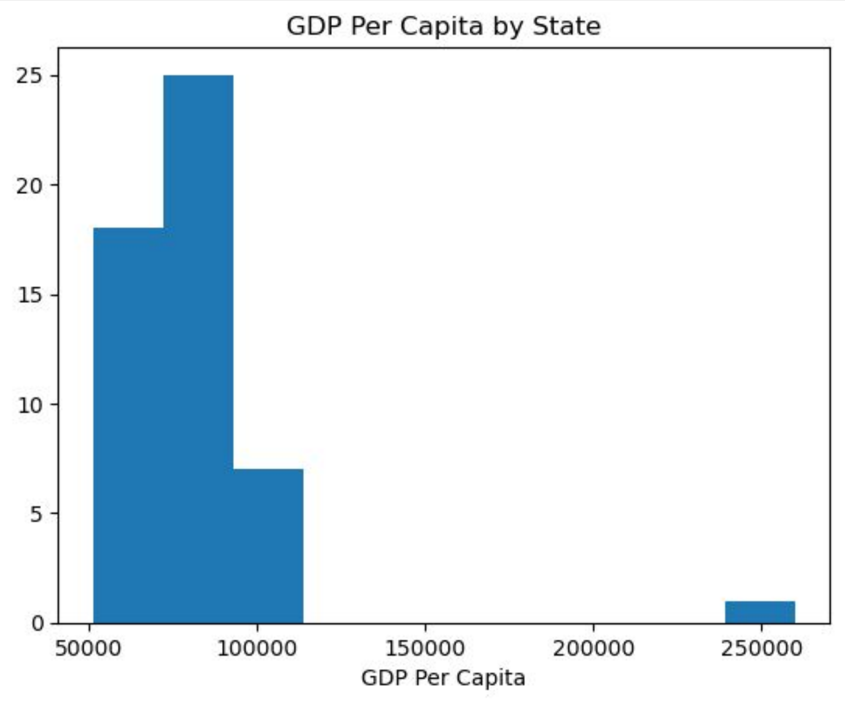


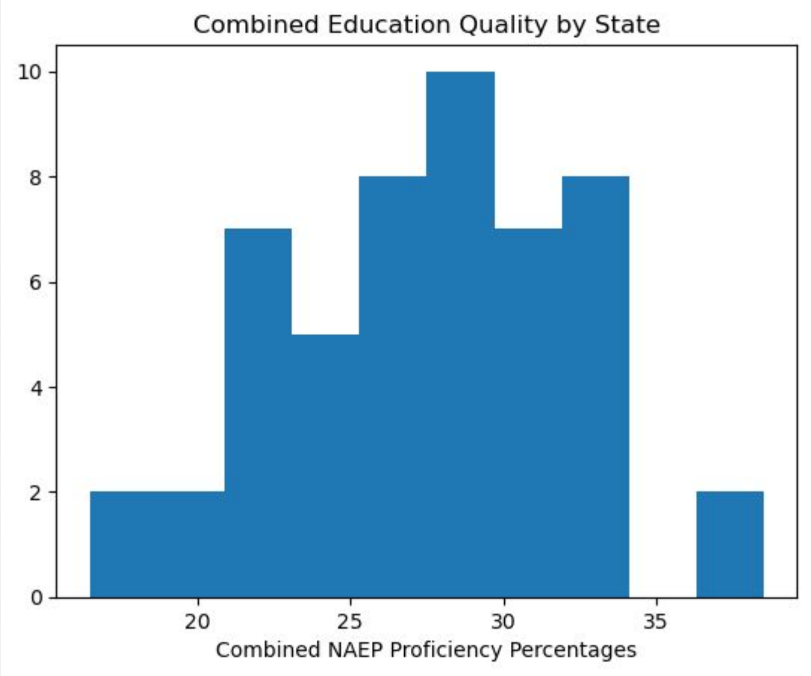

Mathematics, grade 8
Difference in percentage at or above Proficient between jurisdictions, for all students [TOTAL] = All students, 2024

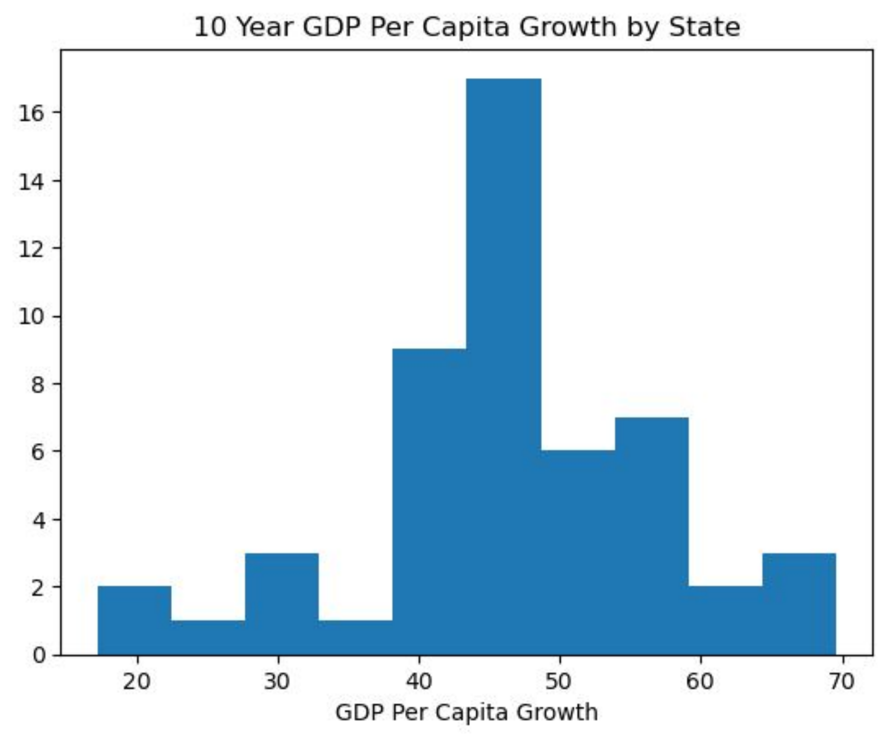
Click on a state
to see its
detailed profile

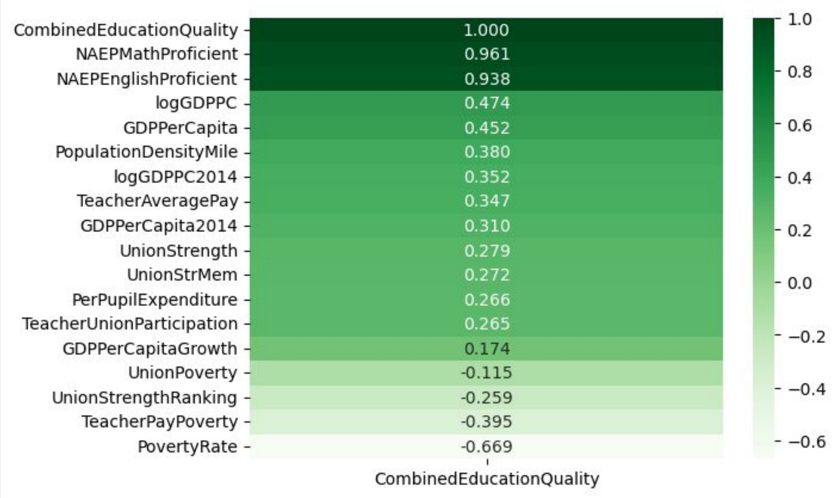
03

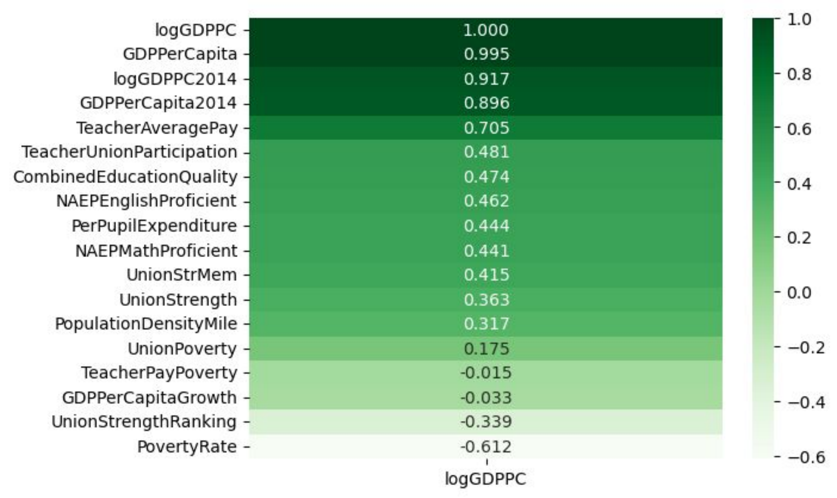
Exploration

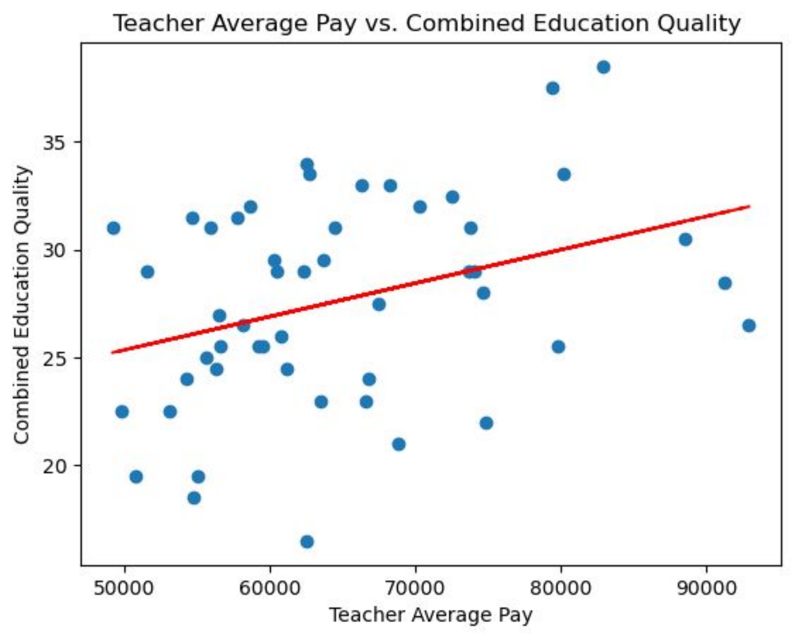

GDP Per Capita by State
25
20
15
10
5
0
50000
100000
150000
200000
250000
GDP Per Capita

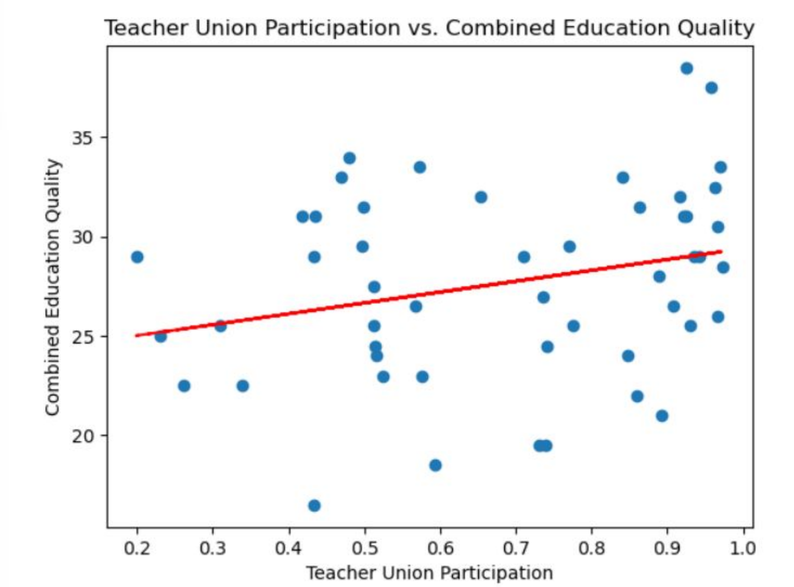

Combined Education Quality by State
10
8
6
4
2
0
20
25
30
35
Combined NAEP Proficiency Percentages

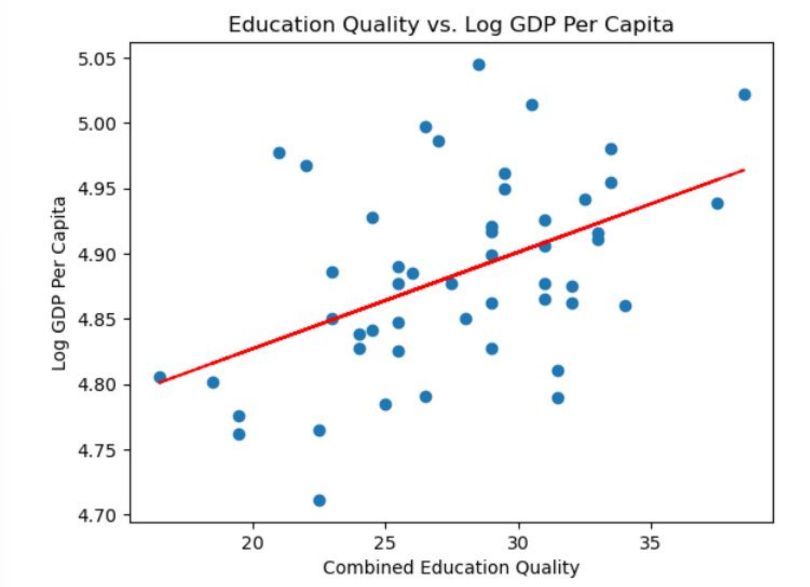

10 Year GDP Per Capita Growth by State
0
2
4
6
8
10
12
14
16
20
30
40
50
60
70
GDP Per Capita Growth

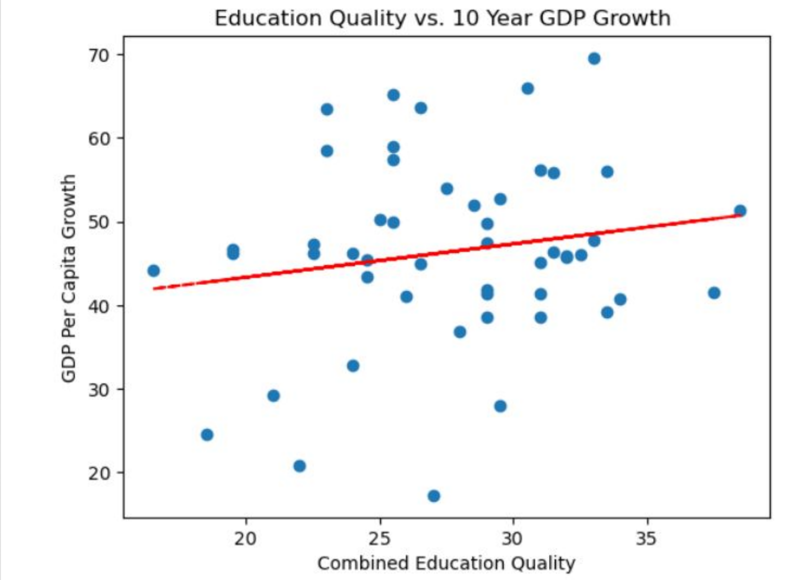

CombinedEducationQuality 1.000
NAEPMathProficient 0.961
NAEPEnglishProficient 0.938
logGDPPC 0.474
GDPPerCapita 0.452
PopulationDensityMile 0.380
logGDPPC2014 0.352
TeacherAveragePay 0.347
GDPPerCapita2014 0.310
UnionStrength 0.279
UnionStrMem 0.272
PerPupilExpenditure 0.266
TeacherUnionParticipation 0.265
GDPPerCapitaGrowth 0.174
UnionPoverty -0.115
UnionStrengthRanking -0.259
TeacherPayPoverty -0.395
PovertyRate -0.669
CombinedEducationQuality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

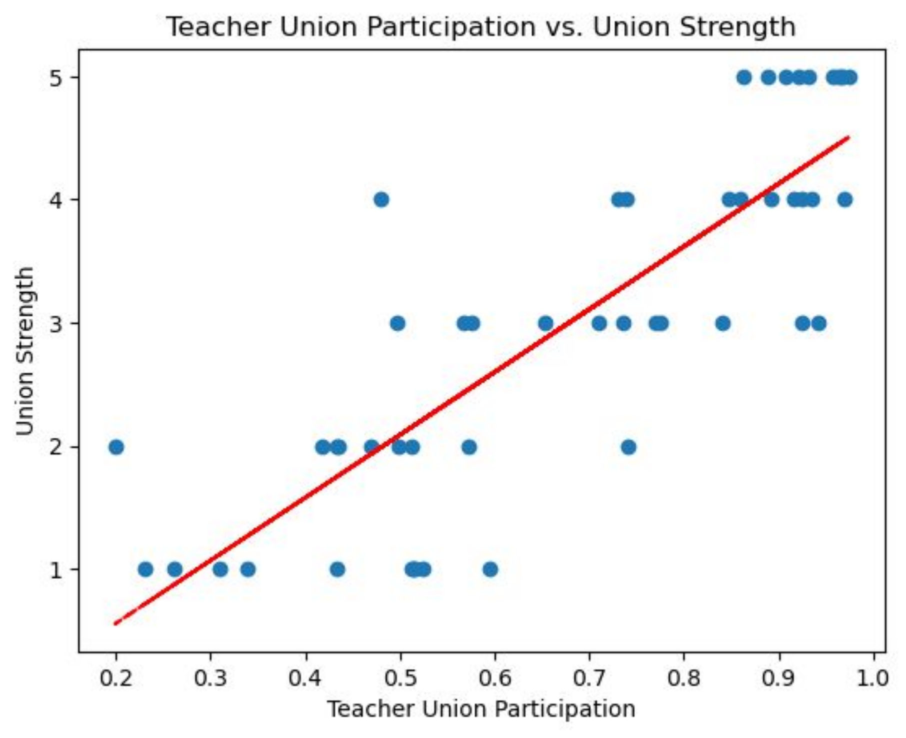

logGDPPC 1.000
GDPPerCapita 0.995
logGDPPC2014 0.917
GDPPerCapita2014 0.896
TeacherAveragePay 0.705
TeacherUnionParticipation 0.481
CombinedEducationQuality 0.474
NAEPEnglishProficient 0.462
PerPupilExpenditure 0.444
NAEPMathProficient 0.441
UnionStrMem 0.415
UnionStrength 0.363
PopulationDensityMile 0.317
UnionPoverty 0.175
TeacherPayPoverty -0.015
GDPPerCapitaGrowth -0.033
UnionStrengthRanking -0.339
PovertyRate -0.612
logGDPPC
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Teacher Average Pay vs. Combined Education Quality
Combined Education Quality
20
25
30
35
50000
60000
70000
80000
90000
Teacher Average Pay


Teacher Union Participation vs. Combined Education Quality
Combined Education Quality
20
25
30
35
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Teacher Union Participation

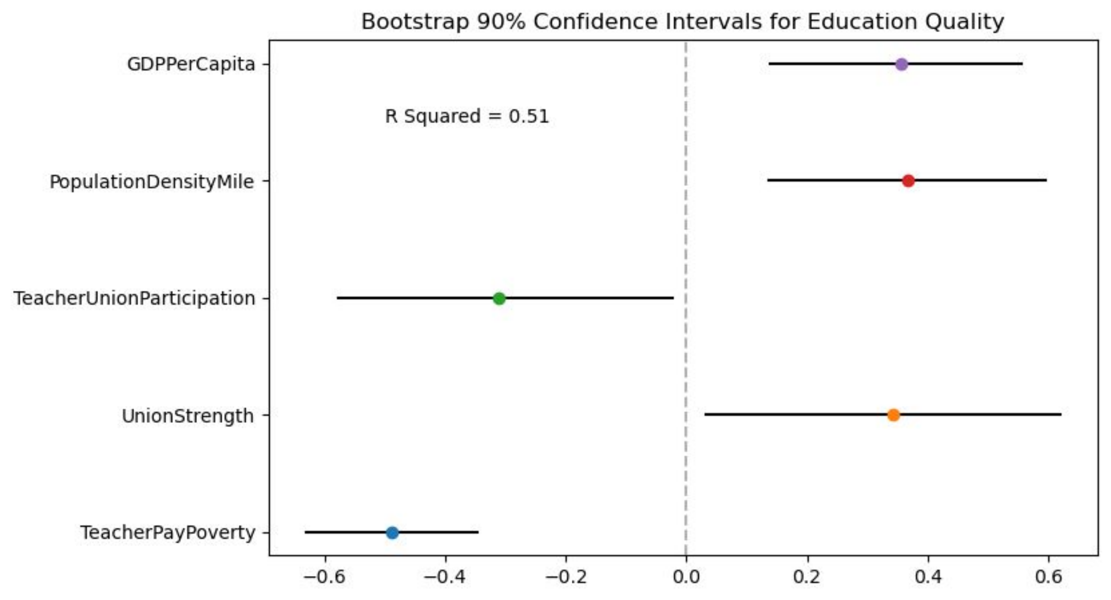

Education Quality vs. Log GDP Per Capita
Log GDP Per Capita
5.05
5.00
4.95
4.90
4.85
4.80
4.75
4.70
20
25
30
35
Combined Education Quality

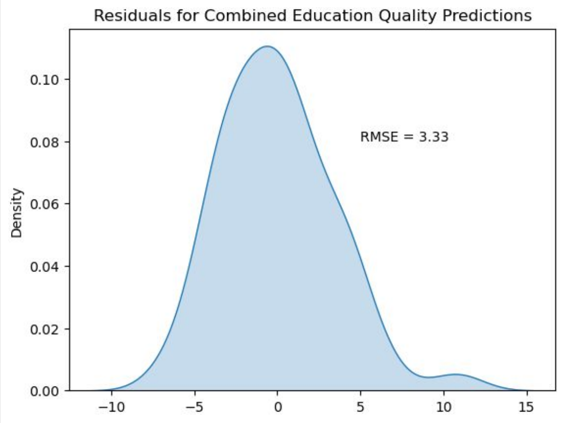

Education Quality vs. 10 Year GDP Growth
GDP Per Capita Growth
70
60
50
40
30
20
20
25
30
35
Combined Education Quality

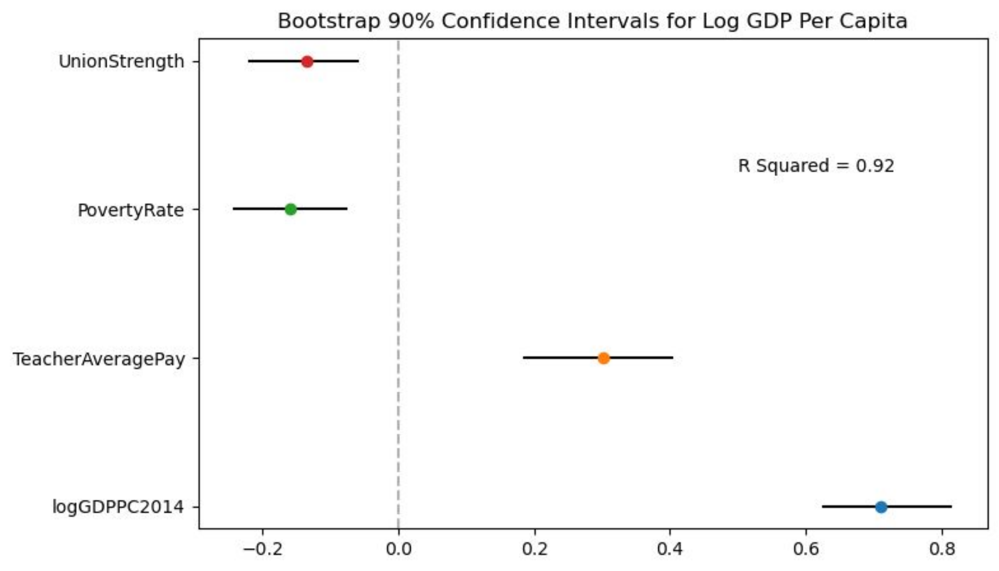

Teacher Union Participation vs. Union Strength
Union Strength
5
4
3
2
1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Teacher Union Participation

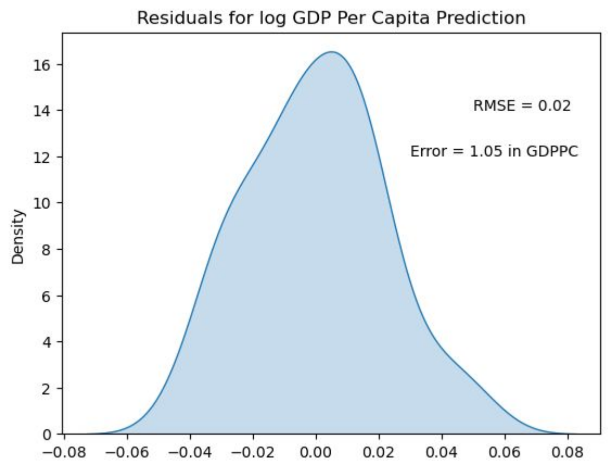
04

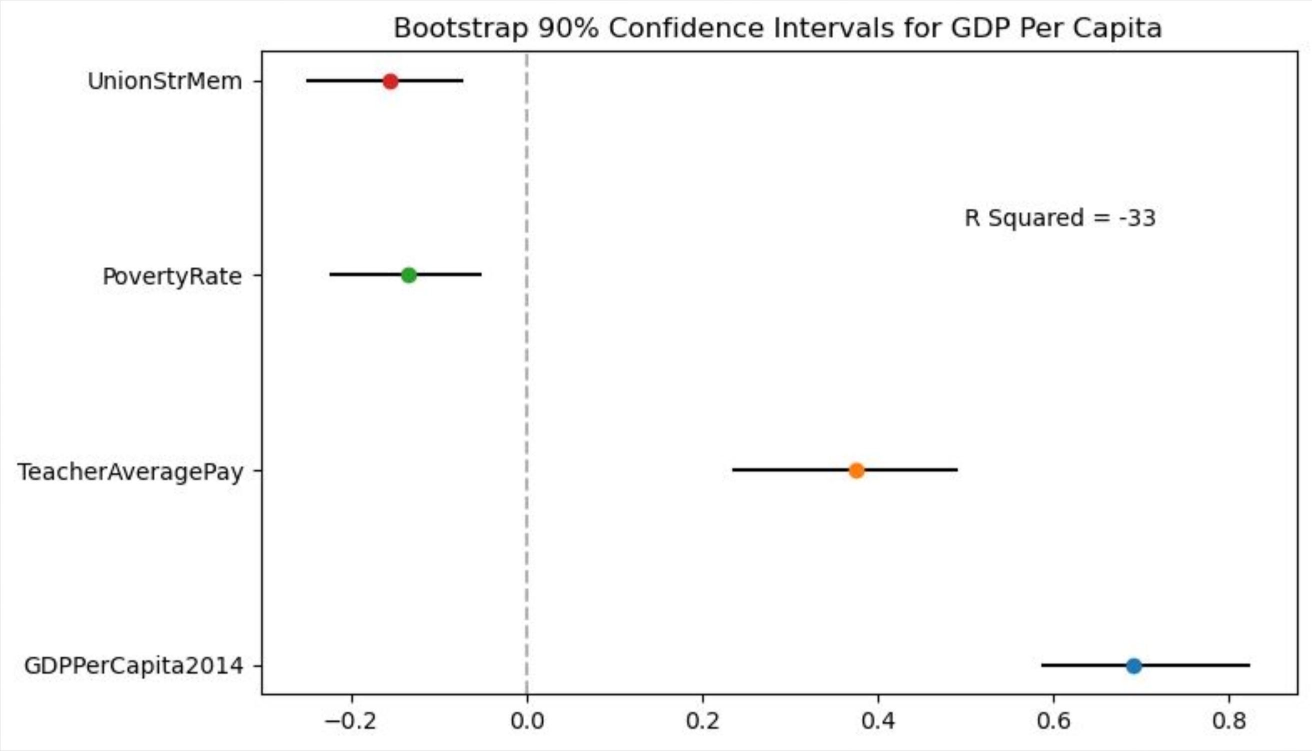
Models


Bootstrap 90% Confidence Intervals for Education Quality
GDPPerCapita
PopulationDensityMile
TeacherUnionParticipation
UnionStrength
TeacherPayPoverty
R Squared = 0.51
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

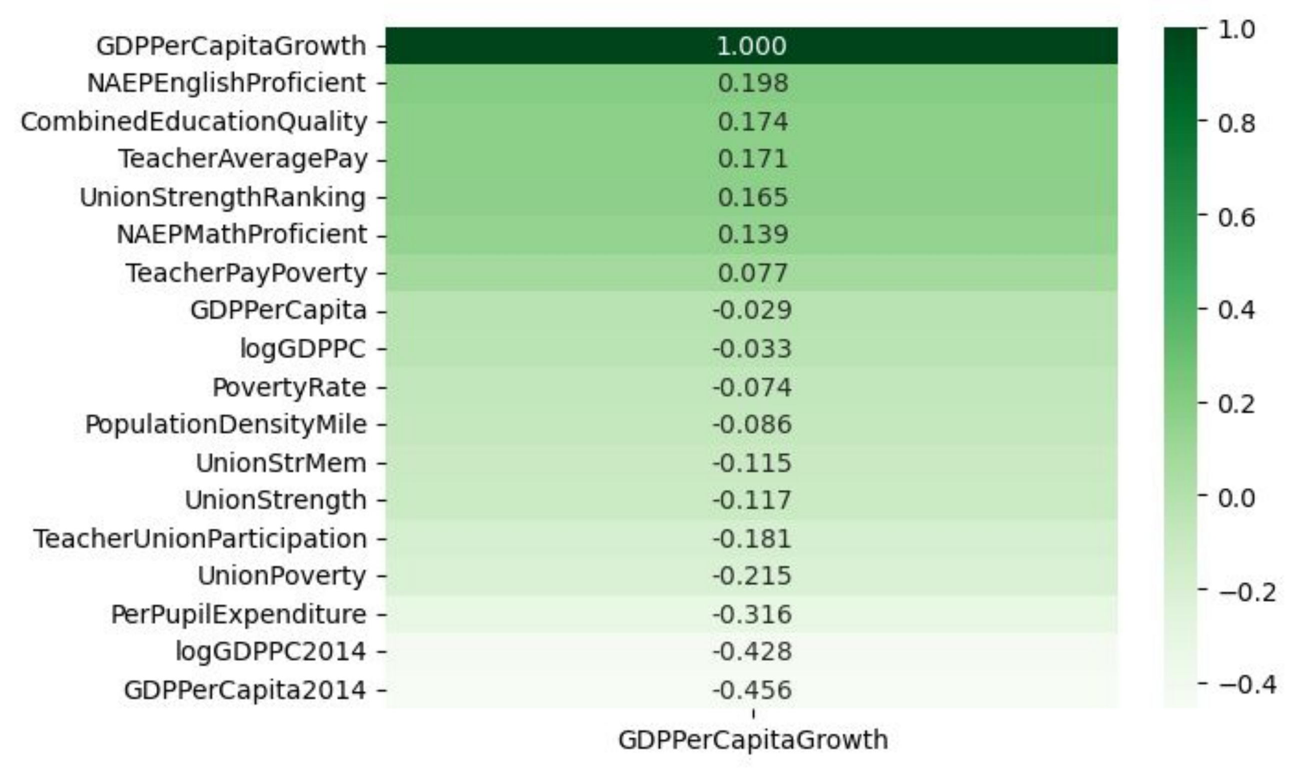

Residuals for Combined Education Quality Predictions
RMSE = 3.33
Density
0.00
0.02
0.04
0.06
0.08
0.10
−10
−5
0
5
10
15

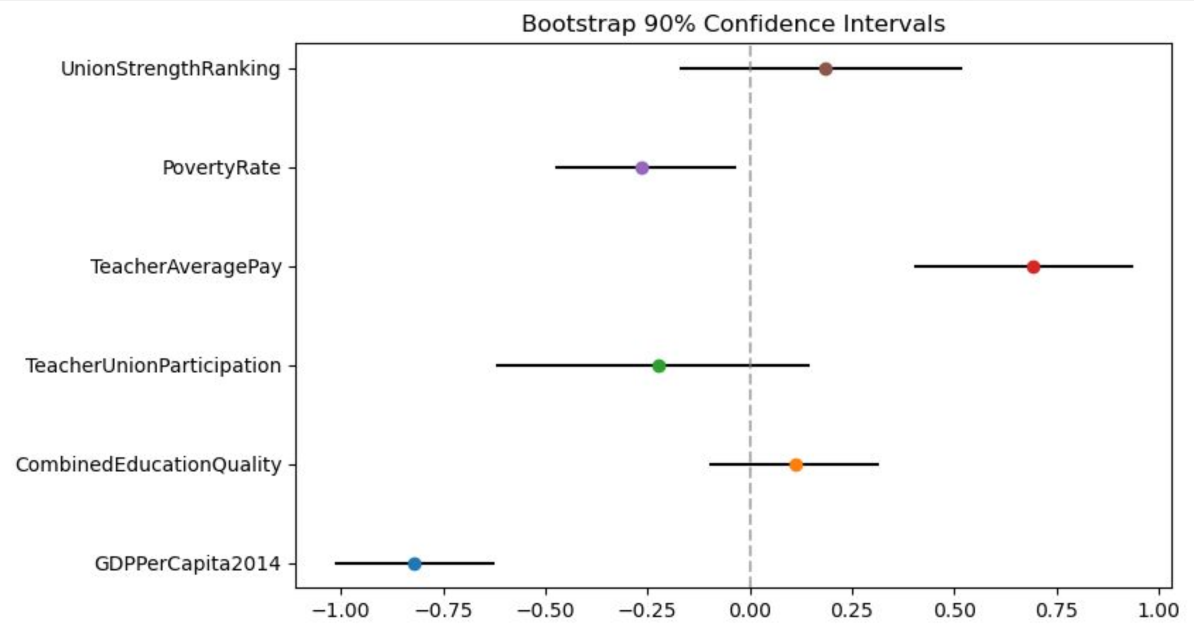

Bootstrap 90% Confidence Intervals for Log GDP Per Capita
UnionStrength
PovertyRate
TeacherAveragePay
logGDPPC2014
R Squared = 0.92
−0.2
0.0
0.2
0.4
0.6
0.8

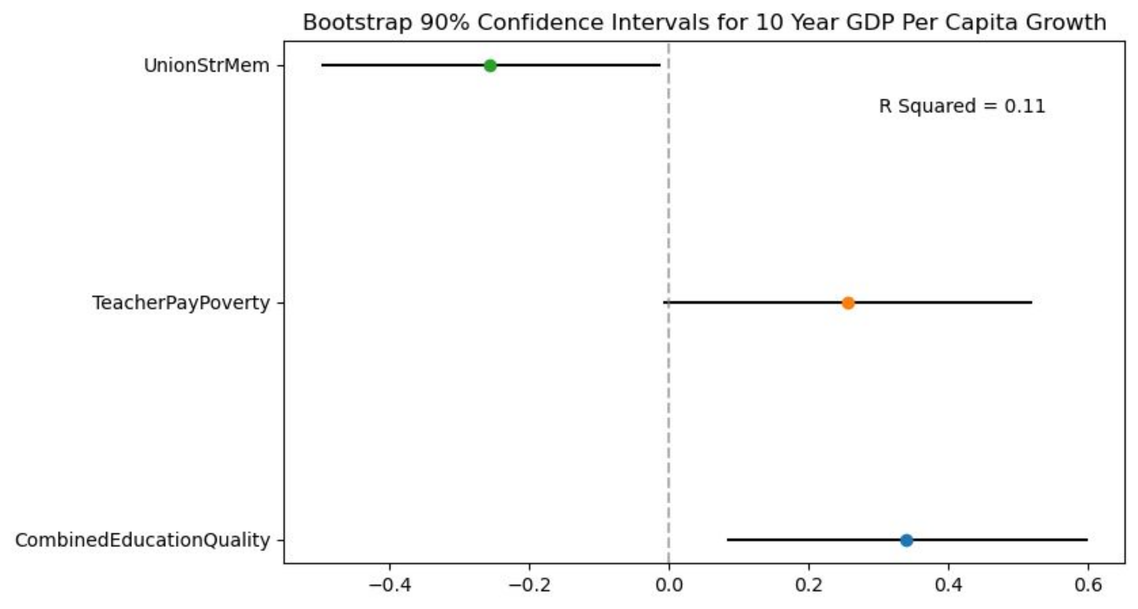

Residuals for log GDP Per Capita Prediction
Density
RMSE = 0.02
Error = 1.05 in GDPPC
0
2
4
6
8
10
12
14
16
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08

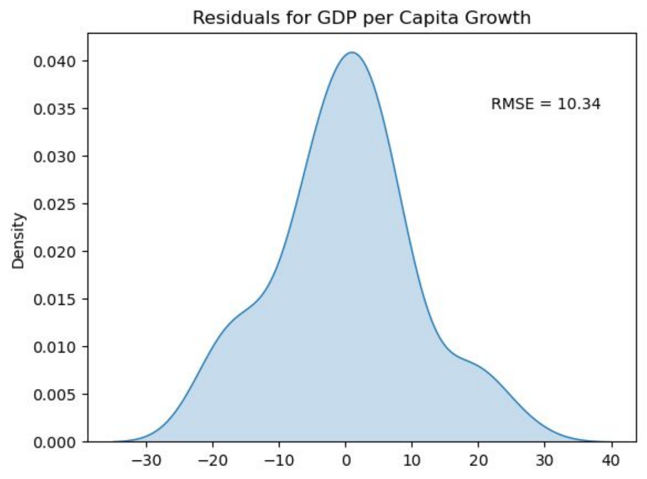


Bootstrap 90% Confidence Intervals for GDP Per Capita


GDPPerCapitaGrowth 1.000
NAEPEnglishProficient 0.198
CombinedEducationQuality 0.174
TeacherAveragePay 0.171
UnionStrengthRanking 0.165
NAEPMathProficient 0.139
TeacherPayPoverty 0.077
GDPPerCapita -0.029
logGDPPC -0.033
PovertyRate -0.074
PopulationDensityMile -0.086
UnionStrMem -0.115
UnionStrength -0.117
TeacherUnionParticipation -0.181
UnionPoverty -0.215
PerPupilExpenditure -0.316
logGDPPC2014 -0.428
GDPPerCapita2014 -0.456
GDPPerCapitaGrowth
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

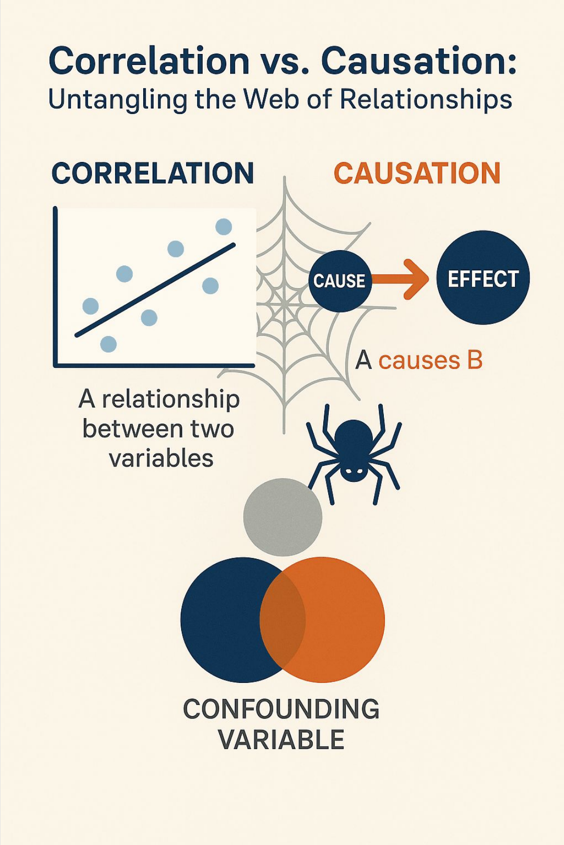

Bootstrap 90% Confidence Intervals
UnionStrengthRanking
PovertyRate
TeacherAveragePay
TeacherUnionParticipation
CombinedEducationQuality
GDPPerCapita2014
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

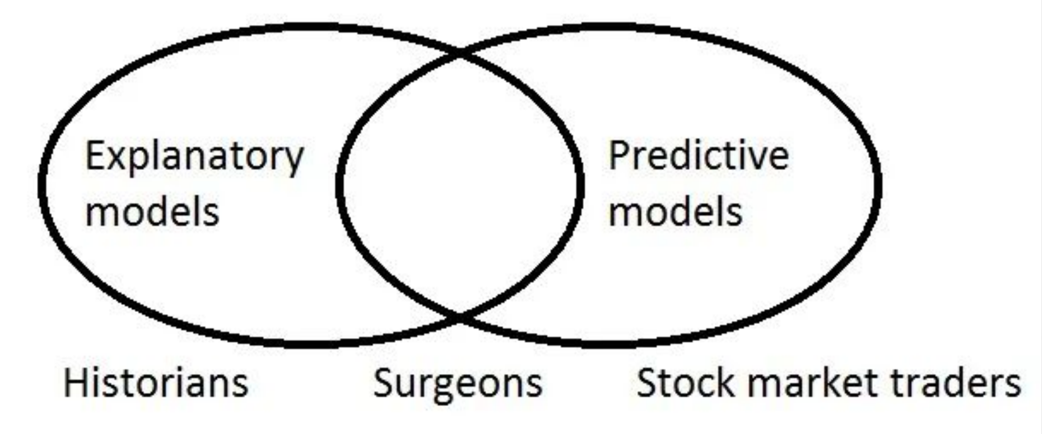

Bootstrap 90% Confidence Intervals for 10 Year GDP Per Capita Growth
UnionStrMem
TeacherPayPoverty
CombinedEducationQuality
R Squared = 0.11
−0.4
−0.2
0.0
0.2
0.4
0.6

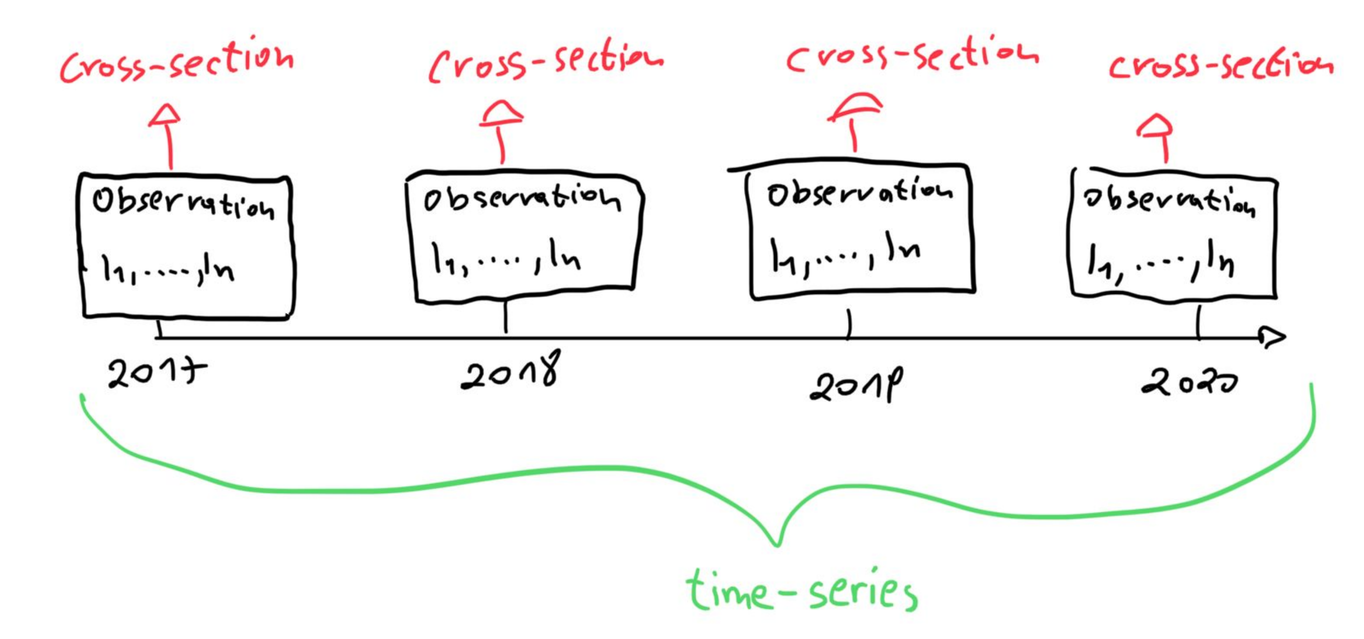

Residuals for GDP per Capita Growth
Density
RMSE = 10.34
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−30
−20
−10
0
10
20
30
40

05

Summary & Future


Correlation vs. Causation:
Untangling the Web of Relationships
CORRELATION
CAUSATION
CAUSE
EFFECT
A causes B
A relationship between two variables
CONFOUNDING VARIABLE


Explanatory models
Predictive models
Historians
Surgeons
Stock market traders


Cross-section
Cross-section
Cross-section
Cross-section
Observation
$l_1, \dots, l_n$
Observation
$l_1, \dots, l_n$
Observation
$l_1, \dots, l_n$
Observation
$l_1, \dots, l_n$
2017
2018
2019
2020
time-series

When considering macroeconomic measures of long-term growth, prioritization of growing capital is widely regarded as one of the most important ways to continue increasing laborer productivity. On the side of firms and businesses, this comes through in terms of investments in physical capital and expansion that improves employment and increases the total output of the firm, then aggregated over many firms to create a largely felt effect on the economy. However, on the other side of the coin, the government's investments and usage of fiscal policy in fostering long-term, sustainable growth generally has two focuses. Assuming rule of law and strongly defined property rights are already defined and not largely changed, one way spending in government expenditures can promote long-term economic growth through spending on physical infrastructure like roads, public transportation, bridges, or sewage, hopefully outweighing the "crowding out" effect of running deficits on government budgets on national saving. The other way, and the method I am investigating further with this project, is with investment into human capital through funding of educational institutions, programs, services, and access. The return on human capital cannot be understated: when we examine how countries have recovered from great devastation to become international leaders in manufacturing and production, like Japan and Germany after World War II or China after the Cultural Revolution, one key trait we find in common is high investment into human capital by the government. From an extensive apprenticeship system to an on-the-job training emphasis, human capital has not only been a key driver in innovation and thus productivity spikes, but also in ensuring overall economic prosperity in the long-term rather than short-term relief.

Reviews, literature, and studies on the topic of the effects of fiscal policy as well as general policies around education and its workers on education quality and its subsequent link on economic growth are indeed extensive; however, there exists a lack of recent data collection and analysis on the differences between states. There exists a unique opportunity with the United States of America, having fifty largely varying territories with differing state laws, natural resources, populations, and demographics, while still having many variables controlled through federal laws and regulations, language, and general culture surrounding work and labor. My project is to empirically determine a link between variables on education spending by the state, policies and culture promoting or discouraging union presence, membership, and collective bargaining strength, as well as their interaction through interaction terms in regression, on GDP per capita long term growth, if there indeed is one. Such variables would include public education spending as a percent of the states' total GDP, per-pupil expenditure, instructional teacher union participation rates, and teacher union strength. I began with a two model approach, testing two questions: first, can we establish a link between expenditure alongside union variables with education quality? Next, if there is a link, I will include education quality and outcomes, as measured by state standardized testing data in the National Assessment of Educational Progress, as an independent variable alongside the original independent variables from the first model to test a link between education quality and economic growth. Due to a lack of extensive time-series data for all variables, though I would like to run panel regression to compare each state to itself as its output, spending, and other educational variables changed, I instead will be comparing states versus each other. Thus, I will attempt to control for some differences in each state that can affect growth, like poverty rate, population density, and initial GDP per capita from 10 years prior in the case of current GDP per capita, by including them in the multiple linear regression. In my research, I also attempted to measure any statistically significant correlative factors between the variables measured and 10-year GDP per capita growth by state. I gathered all of this data from various public sources, including NAEP and

BEA, and compiled them into an Excel spreadsheet before importing into Python and performing exploratory data analysis to look for distributions and outliers, cleaning, and adding interaction terms.

For visualizations, I used heatmaps to determine correlation between dependent variables and independent variables to gain a first good understanding for which ones are relevant to be included within the next machine learning models. For variables with high correlation, scatterplots would be used to gain a better understanding and explore the data's ranges, alongside adding trendlines to better visualize relationships. Additionally, to see how certain variables have changed, I include trend lines of historical data on independent variables. After creating models, I included coefficient plots to more succinctly show variable effects' significance alongside residual diagnostics to determine if model assumptions were met. In all of these, I used pandas, scikitlearn, seaborn, matplotlib, numpy, math, and statsmodels libraries. I additionally used bootstrapping in place of testing and training data splits due to small sample size and the importance of each state, and measured error in the models, built off the mean of bootstrapping coefficients using Root Mean Squared Error.

Here are links to the main datasets I will be pulling from and combining:

BEA Interactive Data Application (GDP)
Union Density Estimates by State, 1964-2021
States with Teacher Unions 2025
Teacher Union Strength
Nation's Report Card
Poverty Rate by State 2025
Real per capita GDP by state U.S. 2024| Statista
COE - Public School Expenditures
Useful Stats: Per Capita Gross State Product, 1998-2018 | SSTI
BEA Interactive Data Application (Population)
Fiscal Effort Index.
Population Density