DATA 11900
Ellise Kurczewski
Daniel Babnigg
Enrico Madani
Jonny Brinkley

Impact of Socio-economic Features on PISA Scores

Our project centers on PISA scores, which stands for Programme for International Student Assessment. This metric is used to evaluate the strength of a country's education system by measuring the performance of 15-year-old students in science, math, and reading. We were motivated by articles that linked educational attainment to increases in a country's GDP[1] and an individual's social and physical health,[2] which demonstrate the impact that education can have on a country's well being. This raised some questions for us: what features of a country's economy and culture are most impactful on PISA scores, and therefore, what features should a country "invest" in (either monetarily or otherwise) in order to improve its education system? Throughout our analysis, we also wondered which features might be more impactful to PISA scores, as the variables we used in our analyses were weak predictors.

Our dataset came from Kaggle[5] and contained 634 rows and 20 columns. 39 countries were represented, and data was taken from 2003-2018 over three-year periods. Each year and country contained average scores aggregated by sex, but we did not plan to analyze based on sex, so we dropped non-mixed-gender rows. Some rows contained missing values, which we imputed using the mean for each country on the grounds that observations for a variable within a country in the 15-year period tended to be very similar. Additionally, five countries had missing entries for all years for some variables, so we dropped those countries from the dataset. Our data initially showed a strong left-skew, which is visible in figure 1a. We attempted transformations of the response variable using square root, log, and other functions, which were unsuccessful. We performed a Box-Cox transformation, which succeeded at normalizing the data, but came at the cost of interpretability later in our analysis. For this reason, we instead opted to drop "outlier" observations, which were observations from five countries with significantly lower PISA scores than other countries in the dataset. After doing so, the distribution of the target variable was approximately normal, as shown in figure 1b. Lastly, we transformed the categorical variables using one-hot encoding, including time and country (name), and we standardized all of the numerical variables.

We created a correlation matrix and found that no variables demonstrated a significant correlation (absolute value ≥ 0.5) with PISA score. However, several variables, including gini_index, alcohol_consumption_per_capita, unemployment and suicide_mortality_rate, exhibited moderate correlation scores, hovering around 0.4 in absolute value. We decided to further examine the relationship of these few moderately correlated variables to PISA score by examining their scatter plots in figures 2a-2d. Both gini_index (a measure of wealth inequality) and unemployment features show negative correlation which is intuitively expected whereas alcohol_consumption_per_capita and suicide_mortality_rate interestingly show positive correlation with the PISA score. Understanding these nuanced relationships guided our selection of features for regression models. We created seven feature combinations for our model selection and analysis, including at least one of the more influential variables in each combination. We grouped the features generally by category (economic inequality, mortality, government expenditure, population, wealth, 2 miscellaneous).

We used multilinear, ridge, and LASSO regression methods on each of the feature combinations in order to give us the best chance of finding a highly accurate model in predicting numerical PISA score. First, we created a test-train split on the dataset. Then, we ran Ridge and LASSO cross validations in order to find the best alphas (tuning parameters) for each feature combination using the training dataset. We then created a manual 5-fold analysis with each of the 21 models in order to find which model had the lowest loss (mean squared error) on average across the folds of the training dataset. We printed the average MSE for each model, and found that the model with the lowest MSE varied depending on the random test-train split. To verify the best model, we implemented KFold and ran several trials with shuffled folds, recording the best model and the R^2 for each trial. The ridge model with predicting variables gini_index, mortality_rate_infant, intentional_homicides,

suicide_mortality_rate, alcohol_consumption_per_capita, and government_health_expenditure_pct_gdp performed the best most often.

R^2 is a measure of the fraction of variance in the response variable that can be explained by variance in the independent variables. R^2 values for the best models of the first 9 trials ranged from about 0.20 to 0.24, showing that even our best model performed quite poorly at predicting PISA scores. The weak prediction power of our model tells us that none of the variables that we analyzed were good predictors of PISA score. We believed that there may be other variables that vary by country, which are not included in the dataset, that are stronger predictors. To explore this possibility, we graphed residuals for our best model in figure 3a, as well as our best model with one-hot encoded country variables within the features in figure 3b. In comparing these figures, it is clear that residuals in figure 3b are much smaller than in figure 3a, which means that country variables significantly improve the prediction power of the model. Therefore, there must be variables linked to the country's status that can be used to predict scores.

We decided to perform further exploratory analysis in order to visualize trends in the data across countries. We created association graphs with the two most influential variables (in order to "control for" the predictive power that exists already within the dataset, while still creating a visual) in figures 4a and 4b. We noticed that the data clustered by country, which encouraged us to run a cluster analysis in order to begin to consider and understand what some stronger influential variables may be.

We used both K-Means clustering and hierarchical clustering, both employing the best model predictive variables along with the PISA score to create the clusters, as we were interested in observing which countries were most similar both in rating and in features. We first made an elbow plot; however, there was not a very clear elbow point. The optimal point seemed to switch between 3 and 4 when we ran the plot multiple times, and because we had no conceptual motivation for a specific number of clusters, we decided to choose K=3 for the K-Means clustering and K=4 for the Hierarchical clustering (this also prevented one cluster from being overly large) . The resulting K-Means cluster is shown in figure 5, the dendrogram for our hierarchical clustering is shown in figure 6, and the results of both cluster analyses are shown geographically in figure 7a and 7b.

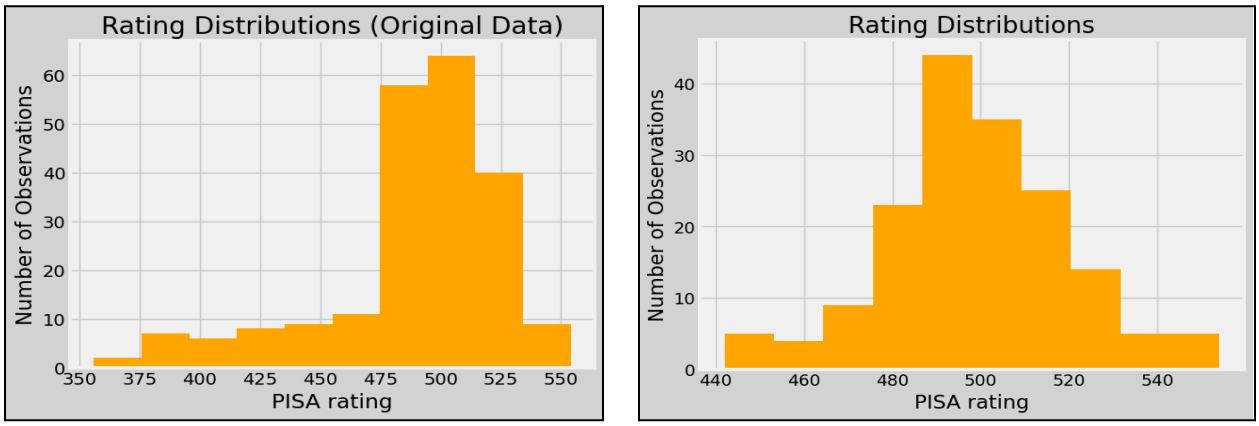
As shown in figures 7a and 7b, there is significant overlap between the clusters formed by the two clustering analyses, and there are similarities between the countries within the clusters. For the hierarchical cluster analysis, the largest cluster included many countries in Western Europe, as well as Nordic countries and Canada and Australia, while another cluster included Central and Northern European countries along with South Korea. Another cluster contained Israel, Greece, and Italy (2003-2006), and the last contained the United States and Latvia (2003-2006). In the K-Means model, a similar breakdown can be seen, as one cluster contains mostly Western European and Nordic countries, one contains largely Central and Northern European countries, and one contains mostly Southern European countries, along with the United States.

The fact that the countries generally clustered by region points to external factors that could be shared by countries that were clustered together, which could affect the mean PISA scores we observe in the dataframe. For example, we can look at internet access, which we considered when explaining the difference in scores between Greece and Spain despite a similar Gini Index. Internet access in Europe varies significantly by region, with a significantly higher proportion of people in Western European and Nordic countries having internet access compared to those in Central and Southern Europe.[3] Countries that are geographically close to each other can also share cultural similarities and views towards education.

Based on the results of our analysis, an effective future step would be to find data for some of the possible predictors that were not included in our dataset but we suspect could affect PISA score. These could be imputed and used to create new models that better predict scores. The PISA website[4] offers data collected on individuals when they take the PISA test, which could be used as cultural variables in addition to country-level statistics such as internet access. These include parents' highest education level, home internet access, time spent outside of

school on homework, highest degree expectation, extracurriculars offered by school, among many others. Furthermore, it would be interesting to look at the effects of these variables on individuals' PISA scores within different countries or globally. While we were unable to answer the questions we originally intended to, our analysis has clearly identified next steps towards this goal.

Figures and Tables



Figures 1a and 1b (left to right). Response variable distributions before and after dropping "outlier" observations.

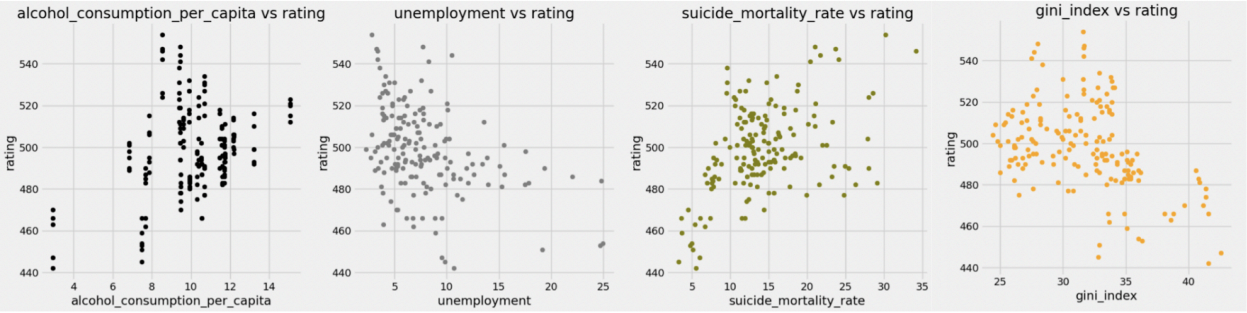
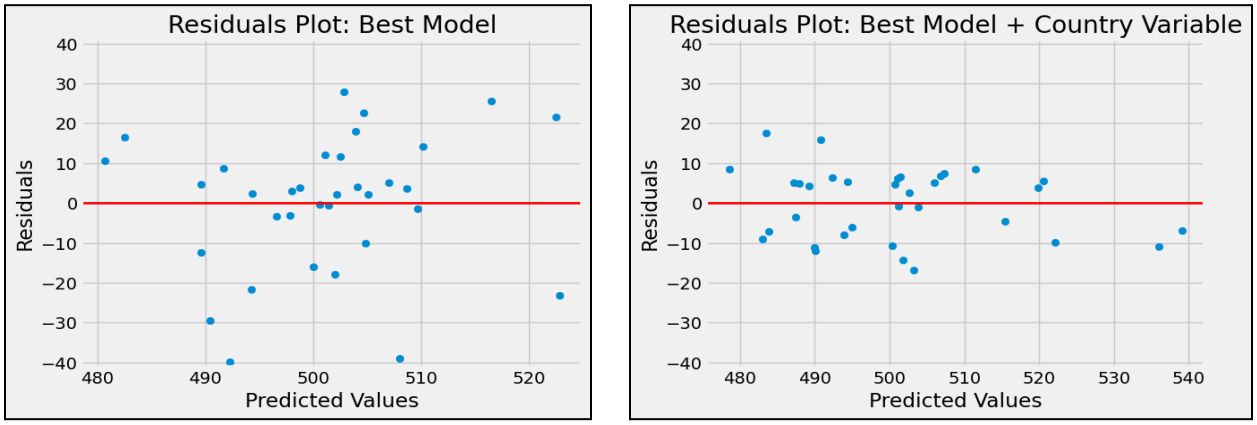
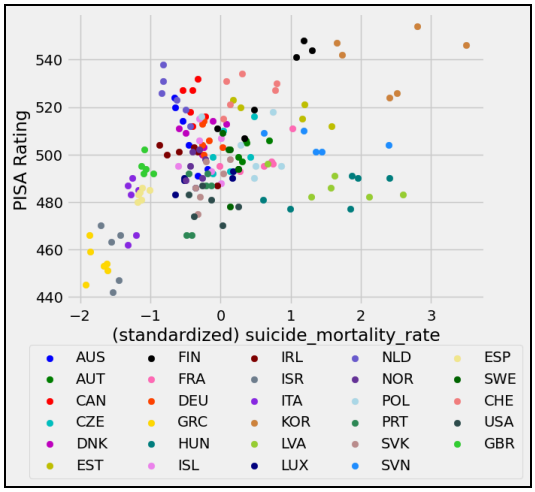
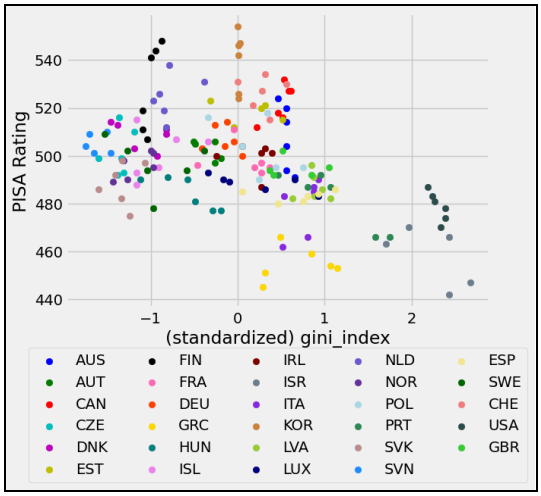


Figure 2a, 2b, 2c, 2d (left to right). Scatter plots of several variables vs PISA score.



Figures 3a and 3b (left to right). Residual plot of the most accurate regression model and residual plot of the most accurate regression model with one-hot encoded country name variables included in its features.





Figures 4a and 4b (left to right). Association graphs of top two influential variables on PISA rating, color-coded by country of observation.

country	cluster_label	country	cluster_label	country	cluster_label	country	cluster_label	country	cluster_label	country	cluster_label	country	cluster_label	country	cluster_label
EST-2006	0	POL-2003	0	AUS-2003	1	CZE-2015	1	DEU-2006	1	LUX-2009	1	SVK-2009	1	GRC-2003	2	ESP-2009	2
EST-2009	0	POL-2006	0	AUS-2006	1	CZE-2018	1	DEU-2009	1	LUX-2012	1	SVK-2012	1	GRC-2006	2	ESP-2012	2
EST-2012	0	POL-2009	0	CAN-2003	1	DNK-2003	1	DEU-2012	1	LUX-2015	1	SVK-2015	1	GRC-2009	2	ESP-2015	2
EST-2015	0	POL-2012	0	CAN-2006	1	DNK-2006	1	DEU-2015	1	LUX-2018	1	SVK-2018	1	GRC-2012	2	ESP-2018	2
HUN-2003	0	POL-2015	0	CZE-2003	1	DNK-2009	1	DEU-2018	1	NLD-2003	1	SVN-2006	1	GRC-2015	2	USA-2003	2
HUN-2006	0	SVK-2003	0	AUS-2009	1	DNK-2012	1	EST-2018	1	NLD-2006	1	SVN-2009	1	GRC-2018	2	USA-2006	2
HUN-2009	0	SVK-2006	0	AUS-2012	1	DNK-2015	1	HUN-2018	1	NLD-2009	1	SVN-2012	1	ISR-2006	2	USA-2009	2
HUN-2012	0	CHE-2003	0	AUS-2015	1	DNK-2018	1	ISL-2003	1	NLD-2012	1	SVN-2015	1	ISR-2009	2	USA-2012	2
HUN-2015	0	CHE-2006	0	AUS-2018	1	FIN-2003	1	ISL-2006	1	NLD-2015	1	SVN-2018	1	ISR-2012	2	USA-2015	2
KOR-2003	0	CHE-2009	0	AUT-2003	1	FIN-2006	1	ISL-2009	1	NLD-2018	1	SWE-2003	1	ISR-2015	2	USA-2018	2
KOR-2006	0	CHE-2012	0	AUT-2006	1	FIN-2009	1	ISL-2012	1	NOR-2003	1	SWE-2006	1	ISR-2018	2		
KOR-2009	0	CHE-2015	0	AUT-2012	1	FIN-2012	1	ISL-2015	1	NOR-2006	1	SWE-2009	1	ITA-2003	2		
KOR-2012	0	CHE-2018	0	AUT-2015	1	FIN-2015	1	ISL-2018	1	NOR-2009	1	SWE-2012	1	ITA-2006	2		
KOR-2015	0			AUT-2018	1	FIN-2018	1	IRL-2003	1	NOR-2012	1	SWE-2015	1	ITA-2009	2		
KOR-2018	0			CAN-2009	1	FRA-2003	1	IRL-2006	1	NOR-2015	1	SWE-2018	1	ITA-2012	2		
LVA-2003	0			CAN-2012	1	FRA-2006	1	IRL-2009	1	NOR-2018	1	GBR-2006	1	ITA-2015	2		
LVA-2006	0			CAN-2015	1	FRA-2009	1	IRL-2012	1	POL-2018	1	GBR-2009	1	ITA-2018	2		
LVA-2009	0			CAN-2018	1	FRA-2012	1	IRL-2015	1	PRT-2009	1	GBR-2012	1	PRT-2003	2		
LVA-2012	0			CZE-2006	1	FRA-2015	1	IRL-2018	1	PRT-2012	1	GBR-2015	1	PRT-2006	2		
LVA-2015	0			CZE-2009	1	FRA-2018	1	LUX-2003	1	PRT-2015	1	GBR-2018	1	ESP-2003	2		
LVA-2018	0			CZE-2012	1	DEU-2003	1	LUX-2006	1	PRT-2018	1			ESP-2006	2		

Figure 5. Results of K-Means cluster analysis with 3 clusters.

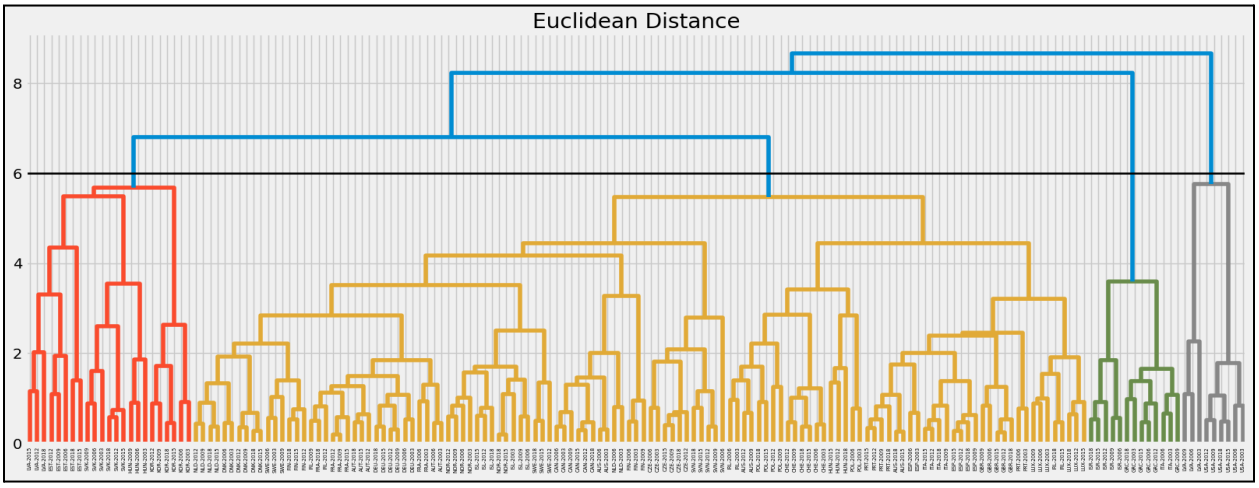
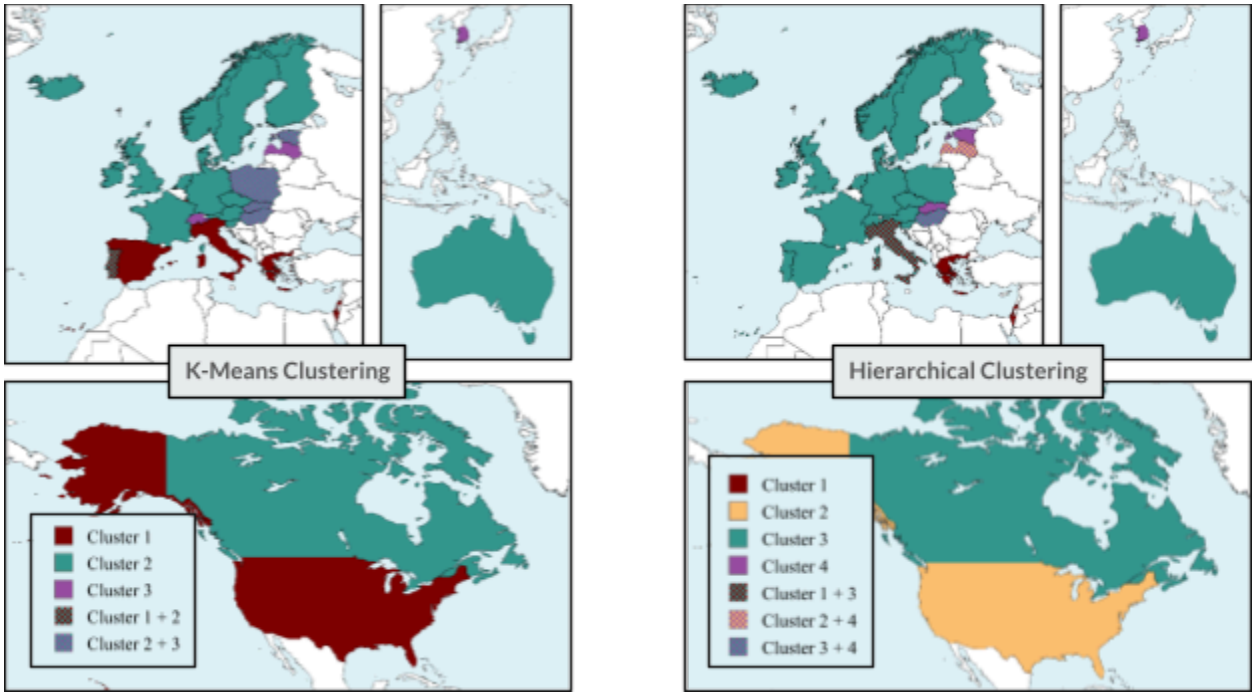


Figure 6. Results of hierarchical cluster analysis with cutoff at 4 clusters.



Figures 7a and 7b (left to right). Map representation of KMeans and hierarchical clustering analyses, respectively.

Citations

1. Holtz-Eakin, D., & Lee, T. (2019, September 17). *The economic benefits of educational attainment*. American Action Forum. https://www.americanactionforum.org/project/economic-benefits-educational-attainment/
2. *Why education matters to health: Exploring the causes*. Virginia Commonwealth University Center on Society and Health. (2015, February 13). https://societyhealth.vcu.edu/work/the-projects/why-education-matters-to-health-exploring-the-causes.html#gsc.tab=0
3. *Digital society statistics at regional level*. eurostat. (2023, October 5). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_society_statistics_at_regional_level&oldid=610768
4. *Programme for International Student Assessment*. OECD. (2023, December 5). https://www.oecd.org/pisa/
5. Tomaz, Walasse. (2024, January). *PISA Results 2000-2022 (Economics and Education)*. Retrieved February, 2024 from https://www.kaggle.com/datasets/walassetomaz/pisa-results-2000-2022-economics-and-education/data

Labor Distribution

Ellise: LassoCV and RidgeCV code, organizing/leading meetings, model creation, K-means clustering code, association graphs (coded by country), presentation, report writing
Enrico: data cleaning, Ridge code, exploratory data analysis, multiple linear regression, presentation, report writing.
Daniel: code cleaning, transformations, visual data exploration, iterative KFold, clustering map synthesis, presentation, report writing
Jonny: LASSO code, report writing, presentation