Math, Minds, and Meteorology: A Global Education Insight

Aleksandra Krasnokutskaia, Marton Nagy, Istvan Jaray, Giorgi Machavariani

1 INTRODUCTION

This project focuses on exploring relationships between performance in International Math Olympiads (IMOs) and broader societal, economic, and environmental factors. By integrating IMO results with data on Nobel laureates, macro-level country indicators, and weather data, we aim to uncover patterns that may explain or predict variations in performance. The project combines multiple data sources into a unified dataset using advanced data engineering techniques and tools like KNIME, SQL, and APIs.

2 SOLUTION OVERVIEW

The proposed solution leverages key concepts from data engineering to build a reproducible data product that facilitates analysis. The International Mathematics Olympiad (IMO) dataset serves as the core data source, enriched with Nobel laureate data, macroeconomic indicators such as GDP per capita and education spending, and weather data obtained from relevant APIs. The primary objectives of the analysis include identifying potential relationships between IMO results and Nobel laureates by country, examining correlations between IMO results and macroeconomic factors such as GDP and education indices, and assessing how climatic conditions at competition venues impact participants' performance. The ETL workflow, implemented using KNIME, integrates data cleaning, enrichment, and analysis processes. The workflow utilizes a MySQL database hosted on Azure for data persistence and reproducibility. Additionally, country names across datasets were unified to ensure seamless integration and avoid discrepancies caused by inconsistent nomenclature.

3 DATA SOURCES

The analysis relies on four key datasets:

International Mathematics Olympiad (IMO) Results: The IMO dataset, compiled from GitHub and other scraped sources, serves as the foundation of this project. It captures individual-level competition results and is aggregated to provide country-year-level observations. Key variables include country, participant scores, competition dates, and locations.

Nobel Prize Data: Nobel laureate data was obtained via the Nobel Prize API. This dataset provides individual-level observations, which were aggregated to the country-year level for integration with the IMO results. Relevant variables include laureate names, birth countries, affiliation countries, award years, and prize categories.

Macroeconomic Data: Macroeconomic indicators were sourced from the World Bank API, offering country-level data aggregated annually. Important variables include GDP per capita, public education spending, and secondary school completion rates. Challenges arising from missing data were addressed using prior value imputation to ensure continuity.

Weather Data: Weather data was retrieved from the NOAA API, capturing daily weather observations for competition locations. A geo-proximity algorithm in KNIME was employed to match competition cities with the nearest weather stations, ensuring accurate representation of climatic conditions.

4 TECHNICAL CHOICES

To achieve the project's objectives, several technical decisions were made. The level of observation was carefully tailored to fit the nature of each dataset. For the IMO results and Nobel laureates, individual-level data was utilized to capture granular details, while country-level aggregations were used for macroeconomic and weather indicators to maintain consistency across sources.

Data cleaning was critical to ensuring dataset compatibility. Country names were unified across datasets using GPT-based translations and manual verification. Missing time-series data from the World Bank API was addressed by imputing prior observations, while moving averages and lagged variables were implemented to better understand temporal relationships.

The ETL pipeline was designed in KNIME, emphasizing modularity and reproducibility. The project leveraged APIs for data enrichment, and the MySQL database was hosted on Azure Cloud for scalability and ease of access.

5 DATA MODEL

The integration of datasets followed a star schema: Central Table: IMO results dataset enriched with Nobel laureate, macroeconomic, and weather data. Linked Tables: Nobel laureate data joined by country and year. Macroeconomic data joined by country and year. Weather data joined by city and competition dates.

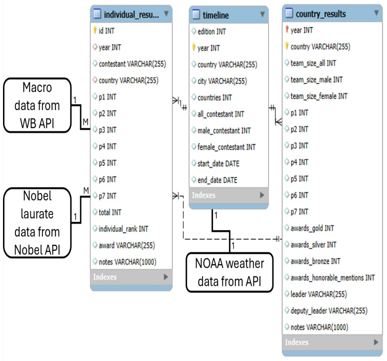

Figure 1: IMO ER Diagram

6 ANALYSIS AND VISUALISATION

Relationship Between IMO Results and Nobel Laureates: Regression analysis was conducted on the log-transformed values of average medals and Nobel Prizes per capita. While the results showed a weak negative correlation, they suggest potential differences in focus between short-term academic achievements and long-term innovation.

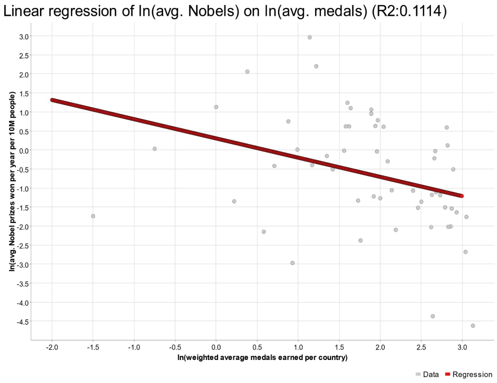

Figure 2: Avg Nob on Avg Medals Regression

Correlation Between IMO Results and Macroeconomic Indicators: Scatterplots revealed positive correlations between weighted medals and macroeconomic factors such as GDP per capita and education spending. This highlights the role of national investments in education and economic stability in fostering competitive performance.

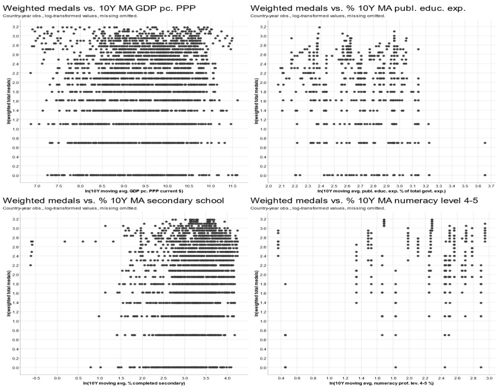

Figure 3: Macroeconomic indicators vs weighted medals scatterplot

Climate's Effect on IMO Performance: Conditional box plots compared IMO performance across different climate categories. The analysis suggests that Nordic countries tend to perform worse in hotter climates, whereas tropical countries exhibit improved performance in warmer conditions.

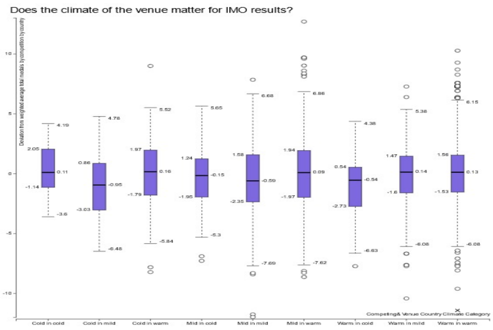

Figure 4: Conditional Box Plot

7 CHALLENGES AND RESOLUTIONS

The project faced several challenges that were addressed using innovative solutions: Country Name Consistency: Historical variations in country names, such as Czechoslovakia, were resolved through GPT-based translations and manual validation. Missing Time-Series Data: Missing observations in the World Bank API dataset were imputed using prior data values to ensure temporal continuity. Weather Data Matching: Geo-proximity algorithms were used to accurately match competition cities with their nearest weather stations, addressing complexities in aligning meteorological data with competition venues.

8 REPRODUCIBILITY

The KNIME workflow and all associated scripts are hosted on GitHub, along with detailed documentation for replication. Key artifacts include the KNIME workflow file, SQL scripts for database setup, and documentation for the APIs and datasets used. The modular design ensures that the project can be reproduced seamlessly.

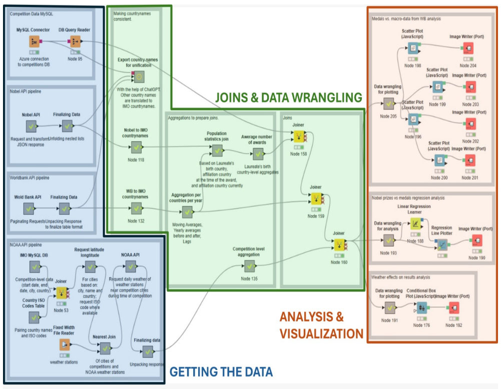
Competition Data MySQL
MySQL Connector
DB Query Reader
Azure connection to competitions DB
Node 95
Nobel API pipeline
Nobel API
Finalizing Data
Request and transform JSON response
Unfolding nested lists
Worldbank API pipeline
Wold Bank API
Finalizing Data
Paginating Requests
Unpacking Response to finalize table format
NOAA API pipeline
IMO MySQL DB
Competition-level data (start date, end date, city, country)
Joiner
Node 53
Country ISO Codes Table
Pairing country names and ISO codes
Fixed Width File Reader
weather stations
Request latitude longtitude
For cities based on: city, name and country; request ISO code where available
NOAA API
Request daily weather of weather stations near competition cities during time of competition
Nearest Join
Of cities of competitions and NOAA weather stations
Finalizing data
Unpacking response
GETTING THE DATA
Making countrynames consistent.
Export country names for unification
With the help of ChatGPT Other country names are translated to IMO countrynames.
Nobel to IMO countrynames
Node 118
WB to IMO countrynames
Node 132
JOINS & DATA WRANGLING
Aggregations to prepare joins.
Population statistics join
Based on Laureate's birth country, affiliation country at the time of the award, and affiliation country currently
Average number of awards
Laureate's birth country-level aggregates
Aggregation per countries per year
Moving Averages, Yearly averages before and after, Lags
Competition level aggregation
Node 135
Joins
Joiner
Node 158
Joiner
Node 159
Joiner
Node 160
Medals vs. macro-data from WB analysis
Data wrangling for plotting
Node 205
Scatter Plot (JavaScript)
Node 198
Image Writer (Port)
Node 204
Scatter Plot (JavaScript)
Node 199
Image Writer (Port)
Node 203
Scatter Plot (JavaScript)
Node 196
Image Writer (Port)
Node 202
Scatter Plot (JavaScript)
Node 200
Image Writer (Port)
Node 201
Nobel prizes vs medals regression analysis
Data wrangling for analysis
Node 193
Linear Regression Learner
Node 188
Regression Line Plotter
Image Writer (Port)
Node 190
Weather effects on results analysis
Data wrangling for plotting
Node 191
Conditional Box Plot (JavaScript)
Node 176
Image Writer (Port)
Node 192
ANALYSIS & VISUALIZATION

Figure 5: Knime Workflow

9 CONCLUSION

This project successfully integrates diverse datasets to uncover connections between IMO performance, Nobel laureates, macroeconomic indicators, and climate conditions. The KNIME pipeline demonstrates advanced ETL capabilities, while the analyses provide actionable insights into factors influencing IMO outcomes. Future work could include additional Olympiads and further enrichment with global datasets to deepen the understanding of the interplay between academic achievement, socioeconomic factors, and environmental conditions.

REFERENCES

[1] KNIME Community. 2024. Geospatial analytics examples - nearest join. Center for Geographic Analysis at Harvard University, (Ed.) [Online; accessed 29-November-2024]. (2024). %5Curl%7Bhttps://hub.knime.com/center%20for%20geographic%20analysis%20at%20harvard%20university/spaces/Geospatial%20Analytics%20Examples/Geospatial%20Analytics%20for%20Beginners/Spatial%20Manipulation/Nearest%20Join~RyxE8-ZuCjTV-dSt/current-state%7D.

[2] National Centers for Environmental Information (NOAA). 2024. Ghcn daily - station data. NOAA, (Ed.) [Online; accessed 29-November-2024]. (2024). %5Curl%7Bhttps://www.ncei.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt%7D.

[3] The Nobel Prize Organization. 2024. Developer zone for nobel prize data. Nobel Prize Organization, (Ed.) [Online; accessed 29-November-2024]. (2024). %5Curl%7Bhttps://www.nobelprize.org/organization/developer-zone-2/%7D.

[4] TidyTuesday. 2024. Tidytuesday dataset overview - 2024-09-24. R for Data Science Community, (Ed.) [Online; accessed 29-November-2024]. (Sept. 2024). %5Curl%7Bhttps://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-09-24/readme.md%7D.