Data Challenge 2025 Report: U.S. Trade Flows Analysis

The challenge focuses on several key research questions that define the scope of our analysis: **(1)** How do trade volumes vary across countries when accounting for factors such as geography, development status, and population? **(2)** What country-specific factors help explain changes in U.S. exports over time? **(3)** We also made predictive models to predict the monthly import and exports in USD of all countries and commodities that the US deals with.

Data

We worked with several monthly datasets covering 2015–2023:

1. Countries Over Time.csv: a country-level panel with Country, calendar fields (Year, month), and the target trade values for imports and exports
2. Commodities Over Time.csv: a commodity-level panel with Commodity, calendar fields (Year, month), and target import and export values
3. Country Commodity [Year].csv: a series of yearly files (2015-2023) at the country x commodity x year level, which also contains import and exports

The Country Commodity [Year] dataset was used to answer the first two questions of analysis while the Country Over Time and Commodities Over Time were used to make the predictive modeling.

Analysis: Determinants of Trade Volume Differences Across Countries

In the data exploration, we analyzed what drives variation in U.S. exports across trading partners in several factors: population of countries, political rank, and distance of country from U.S. To disentangle short-term within-country effects from structural cross-country differences (panel data constraint), we employed two models:

1. **Fixed-Effects Model:** To capture how a country's evolving time-variant characteristics influence U.S. trade.
   Applied fixed-effects model:
   $$Exports_{it} = \alpha_i + \beta_1 Population_{it} + \beta_2 PoliticalRank_{it} + \varepsilon_{it}$$
2. **Random-Effects Model:** To account for both time-variant and time-invariant such as distance of country from U.S.
   Applied Random-effects model:

$$Exports_{it} = \alpha_i + u_i + \beta_1 Population_{it} + \beta_2 PoliticalRank_{it} + \beta_3 CapitalDistance_{it} + \varepsilon_{it}$$ [1]

Fixed-Effects		Random-Effects	
Variables	P-Value	Variables	P-Value
Population	**3.826e-09*** **	Population	**5.15e-15*** **
Political Rank	**0.006283** **	Political Rank	**1.08e- 1**
		Capital Distance	**4.47e-20*** **

[1] Population: value in thousands, Political Rank: Index score 0-100, Distance: Distance from nation's capital to Washington DC.

In both models, population emerges as a strong and positive predictor of U.S. export. In the fixed-effects model, political stability shows a negative and significant effect that implies that more stable countries may rely on domestic or alternative suppliers, while less-stable markets attract higher U.S. exports. In the random-effects model, distance is strongly negative, confirming that countries farther from the U.S. trade less due to higher costs and weaker integration, while political stability becomes insignificant once population and distance are controlled for which suggests its influence is indirect through broader economic factors.
The Hausman test indicates that fixed effects are the appropriate main specification for inference, since unobserved country traits are correlated with observed regressors. Random effects are still useful descriptively because they allow us to include time-invariant distance and see the classic distance penalty.

Country-level Predictive Modeling
We predict 2024 monthly imports/exports (USD) for each country using 2015-2023 country-month history, producing a full 12-month prediction of 2024 for all countries.

We compared two models that are frequently used in time series analysis:
1. **CNN-LSTM:** Conv1D to learn month-of-year patterns + LSTM for dynamics; L=24, H=12, and early stopping
2. **ETS:** Additive trend + additive seasonality (m=12) , optional damping

We evaluated models on countries with no missing data using four rolling out-of-sample folds (train through December of the prior year, predict the next year for 2020–2023), scored by sMAPE and combined with recency weights {2020: 0.1, 2021: 0.2, 2022: 0.3, 2023: 0.4}. The model with the lower weighted sMAPE was selected. CNN-LSTM consistently outperformed ETS on the countries with no missing data.

For countries with missing data, we route by data sufficiency:

Non NA Months (2015-2023)	2023 Months Completeness	Model Used
≥100/108	12/12	CNN-LSTM
<100	12/12	Seasonal-naive
<100	<12	Theta + light imputation

To guard against over-extrapolation, we compare 2024 vs 2023 totals and flag any extreme changes: >+200% or <-60%. We use 2023 vs 2022 as context to sort the flags:
1. *Consistent extremes* (2022 -> 2023 and 2023 -> 2024 both had extreme changes, both in the same direction): Keep the model prediction unchanged
2. *Direction flip* (2022 -> 2023 and 2023 -> 2024 both had extreme changes, but in the opposite direction): Apply a 40% seasonal-naive blend to move towards last year's monthly pattern reduce sudden reversals
3. *New extreme* (2022 -> 2023 has no extreme changes but 2023 -> 2024 does): Apply a 20% seasonal-naive blend to slight move towards last year's pattern

Commodity-level Predictive Modeling

We extend the pipeline to predicting the import and export of all commodities in 2024 using 2015-2023 monthly data. The core setup of model comparison (CNN-LSTM vs ETS), folds, recency weights, and over-extrapolation guard are the same. Just like the country-level model, the recency-weighted sMAPE shows CNN-LSTM as the better model. What differed was the data routing used for the imperfect commodities;

Case	Criteria	Model / Method
All zero / NA	No data from 2015 - 2023	Set prediction = 0
Late start (2 years)	Values only in 2022 & 2023	ETS on log1p, m=12
Late start (1 year)	Values only in 2023	Seasonal-Naive
Ends early	Values end before 2013-12	Donor bridge (ridge on donor MoM) to 2023-12 then predict using ETS
Gaps in data (low zeros)	zero_eff_frac < 0.50	ETS on log1p, m=12
Gaps in data (high zeros)	zero_eff_frac ≥ 0.50	TSB

References

World Bank. (2025). Population, Total. The World Bank; World Bank Group.

World Bank. (2023). Political Stability and Absence of Violence/Terrorism: Percentile Rank. World Bank Open Data

Mayer, Thierry, and Soledad Zignago. “CEPII - Notes on CEPII’s Distances Measures: The GeoDist Database.”, CEPII Research and Expertise on the World Economy, 2011