**Cost of Living, Economic Development,**

**& Supply Chain**

Rotman Datathon 2025 - **Team Number: 190**

Katarina Jocelyn Chandra

Keisha Valenna Kwek

Siya Gupta

Yana Jakhwal

1. **Introduction**

**1.1 Overview**

Supply chain resilience has become an increasingly important focus in recent years, gaining attention from both the public and private sectors. It plays a critical role in business operations, influencing operational costs and the pricing of goods and services (World Economic Forum, 2022), which can, in turn, impact the cost of living. For policymakers, investing in strategies to strengthen their countries’ roles in the global supply chain is often seen as a way to promote economic growth and foster international trade (OECD, 2021).

With globalization, supply chain demands have increased significantly, and while it brings lowered costs and drives innovation, it also presents increased vulnerabilities against disruptions, which have the ability to disrupt lives and businesses, disruptions which will be explored in this report. Studies have shown that the cost of living and economic development of a country has direct impacts on supply chain costs and economic performance in countries. These disruptions affect labour, raw material costs, customs policies, and infrastructure availability (Deloitte, 2023; McKinsey & Company, 2023).

Throughout this report, thorough insights into the intersectionality of supply chain stability indicators, cost of living, and economic indicators will be provided.

**1.2 Problem Statement**

1. Identify Key Indicators – Determine economic development and cost of living indicators most impactful to supply chain stability.
2. Determine Causal Relationships – Model the relationship dynamics between the cost of living and the supply chain.
3. Develop Predictive Models – Create a predictive model to take into account key indicators.
4. Create Recommendations – Recommend action steps for policymakers and businesses to develop more resilient supply chains.
5. **Data Plan– Methodology**

**2.1 Variables and Datasets Used**

Majority of the exploration was done using three data sets in order to capture economic, political, as well as any environmental factors that possibly affect supply chain stability;

1. Rotman Datathon 2025 Raw Dataset – Contains a number of economic indicators, supply chain indicators, and cost of living indices.
2. World Risk Index – Risk for 193 to be impacted by climate change-driven humanitarian crisis on a scale of 0 to 100. The rationale for introducing this dataset was to take into consideration climate change and natural disasters as one of the prominent causes of supply chain disruption. The Global Supply Chain Report (2023) has cited an increase in disruptions caused by natural disasters and climate change, hence the relevance of this dataset.
3. Political Stability Index – Political stability of 193 countries on a scale of -2.5 to 2.5, with higher values indicating higher stability (World Bank, 2023). The rationale for introducing this dataset is the rising prominence of geopolitical and political instability issues during 2024. Hence, taking this index into account can help provide a layer of dimension to the supply chain resilience model.

**Target Variable**:

The Logistics Performance Index was used as the main supply chain indicator as the dependent variable. It is an index developed by the World Bank that evaluates the performance of logistics companies across 6 indicators (World Bank, 2023). The main rationale behind choosing this specific variable was data availability and relevance. While direct metrics such as the supply chain stability index by KPMG exist, public access is limited, hence the dataset was not obtainable (KMPG, 2024).

LPI was chosen as studies have shown that it has a correlational relationship with supply chain costs, where higher LPIs have indicated reduced costs (World Bank, 2023; Banomyong, Varadejsatitwong, Julagsigorn, 2023). Hence, LPI is deemed to be able to represent supply chain costs, in addition to implying stability.

**Feature Variable:**

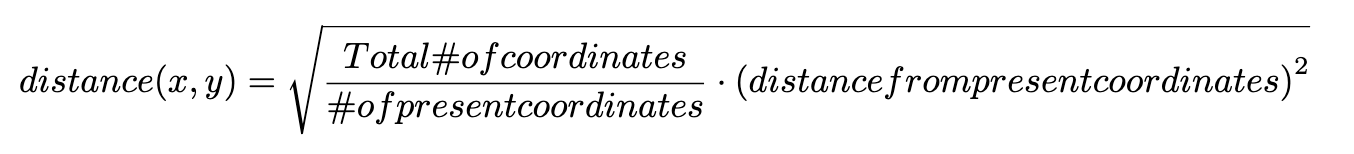
No selection was made for feature variables at the beginning, as it would be part of the further methodology.

**2.2 Imputation – Handling Missing Data**

The dataset obtained after merging the fields of interest from the three datasets contains 1413 missing values and 921,167 non-missing values. Three options were considered for handling the missing data: MICE, Interpolation, and KNN. MICE and Interpolation were rejected as those models were unable to handle a large number of missing values, and thus, KNN was the most suitable method for imputation.

As the datasets were obtained from a variety of sources, there are inconsistencies in the presence of data in certain data points. To set a strong foundation for further analysis, we completed the missing data points by imputation through **K-Nearest Neighbour (KNN).** KNN Imputation finds the nearest neighbours of a missing value (using NaN Euclidean distance) and uses their feature values to impute the missing value by taking the mean of these nearest neighbors.

The NaN Euclidean distances is calculated from the following equation:



**2.3 Exploratory Data Analysis**

We plan to perform three types of data analysis given our highly dimensional data. Principal Component Analysis to understand how much variance the features have with one another, and if most of the features are relevant to keep or if we can drop the subsidiary features and only keep the main (i.e. “overall score”) indexes. Next, we will perform a correlation analysis to check for any multicollinearity that exists within the dataset to prevent a bad fitting of our regression model. Finally, we will perform an HDBScan to visualize the highly dimensional data to see if there are any patterns within the observations.

**2.4 Time-series Forecasting– Predictive Modelling**

We will use time-series forecasting to show future predictions of the logistics performance index under different conditions of the economy through the factors we have previously determined.

## **3. Data Analysis**

**3.1 Exploratory Data Analysis**

With 283 variables and 3260 observations, the goal of the EDA was to narrow down the variables of interest to reduce the dimensionality of the data. The data exploration was completed with four dimensions:

1. Correlation Analysis

After running the correlation matrix, we saw many features to be correlated with other features, given that they are subcategories of a “main” feature. This shows that there is multicollinearity in our variables, which will potentially cause incorrect weights as it undermines statistical significance of some variables if not dealt with properly.

The correlation between two variables, with the Logistics Performance Index (LPI) with all of the columns from the source datasets will be generated to allow us to filter variables with little to no correlation. The correlation matrix generated can be found in the Excel Sheet. We used this matrix to select 15 variables that appeared to be the most relevant.

***Table 1.1: Features with Strong Correlations to Overall LPI***

| **Variable** | **Correlation with LPI** |
| --- | --- |
| Human capital index (HCI) (scale 0-1) | 0.7453798529 |
| Current health expenditure per capita, PPP (current international $) | 0.7327386453 |
| Adjusted net national income per capita (current US$) | 0.7292653572 |
| GNI per capita, PPP (current international $) | 0.6441617524 |
| Machinery and transport equipment (% of value added in manufacturing) | 0.6227502313 |
| GDP per capita, PPP (current international $) | 0.6163236977 |
| Individuals using the Internet (% of population) | 0.6131258785 |
| Exports of goods and services (current US$) | 0.5442900112 |
| Imports of goods and services (current US$) | 0.5173294331 |
| Electric power consumption (kWh per capita) | 0.5046114511 |
| GDP per capita (current US$) | 0.4872885212 |
| Power outages in firms in a typical month (number) | -0.3043393908 |
| Cost of business start-up procedures (% of GNI per capita) | -0.36999884 |
| Mortality rate, infant (per 1,000 live births) | -0.5683634254 |
| Poverty headcount ratio at $6.85 a day (2017 PPP) (% of population) | -0.5868382303 |

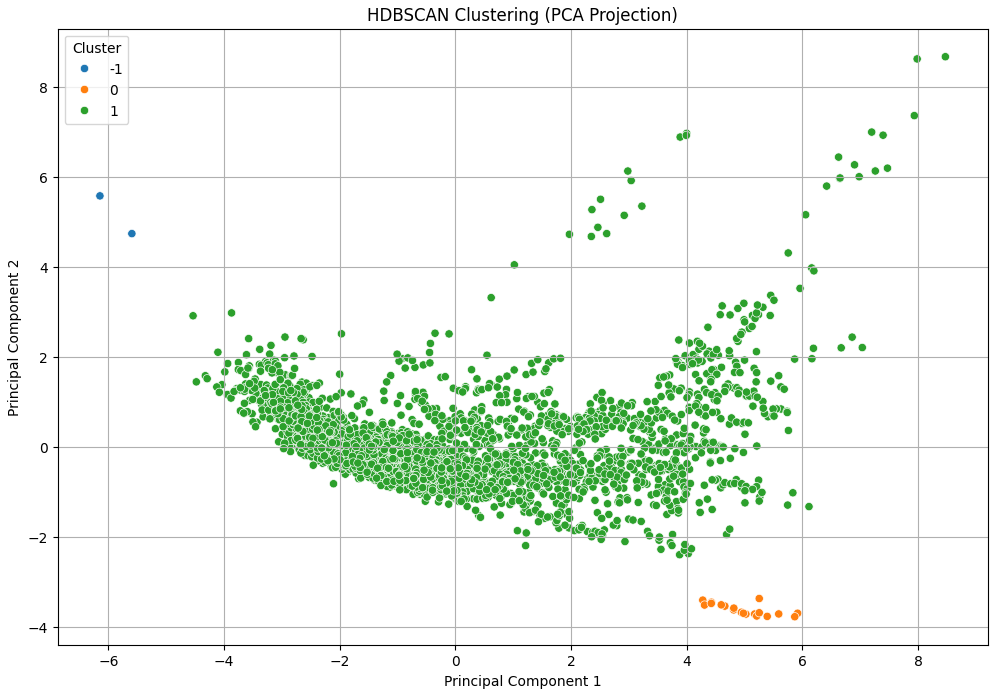
1. Principal Components Analysis (PCA)

PCA was used to help us gauge how much information was captured by the variables. We had to generate 51 principal components to capture 85% cumulative variance. We then found the loadings of these components, and cross-verified them with data from the correlation analysis to arrive at the 15 chosen variables.

PCA will be used to determine which indicators have the most impact by reducing the dimensionality of the data and checking which variables capture the most information.

1. HDBScan

We further performed an HDBScan to judge how different countries perform to identify some regional patterns. We performed PCA on the results and plotted the two components against each other to get a helpful visual of countries and their behaviour.

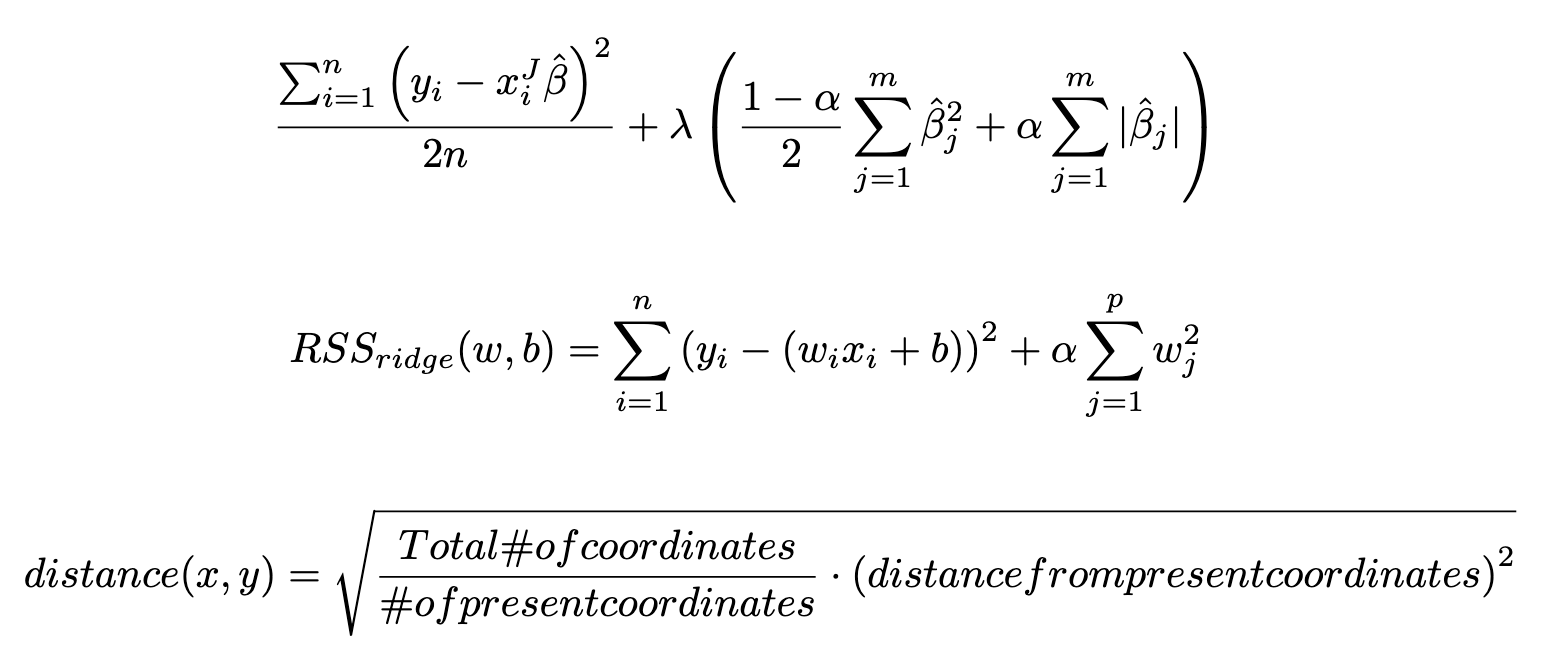


It is interesting to note that cluster 0, which contains the countries of Iceland and the Faroe Islands, has a particularly high LPI index score. On further research, we find that these countries, in addition to convenient geographic locations, have very strong infrastructure and international relations, which helps businesses thrive. This helps us frame our recommendations.

**3.2 Model Selection**

Given that the feature and target variables of the dataset are continuous, a regression model is used. With 280 features, regularization must be performed to estimate the parameters to prevent the model from overfitting. L1 Regularization (also known as LASSO) works well with datasets containing “useless” parameters, therefore reducing its coefficients to zero (i.e. removing parameters entirely). However, given that our data has many features correlated with one another, this tends to cause LASSO to pick the most important feature in the model that is “useful” and eliminate its correlating variables, causing a potential loss of information. L2 Regularization (also known as Ridge) works well with datasets containing mostly “useful” features, as it only performs shrinking of less relevant features but does not remove any of them. However, ridge regression tends to shrink all parameters of the correlated variables together.

The goal is to minimize the incorrect interpretation of feature usefulness for our model, thus, we performed a multivariate version of Elastic Net regression to ensure that features were selected properly. Elastic Net, a combination of LASSO and ridge regression, groups and shrinks parameters associated with the correlated variables and leaves them in the equation or removes them all at once by setting their weights to zero – making it a great choice when dealing with data that has correlation between parameters as we are performing both shrinkage and feature selection. Elastic Net is made out of the sum of squared residuals, a LASSO component, as well as a Ridge component, giving the following equation:

****

The MultiTask Elastic Net was run on a 5-fold cross validation to find the best l1 ratio, demonstrating sparsity and weight shrinkage between 0 (purely LASSO) to 1 (purely Ridge).

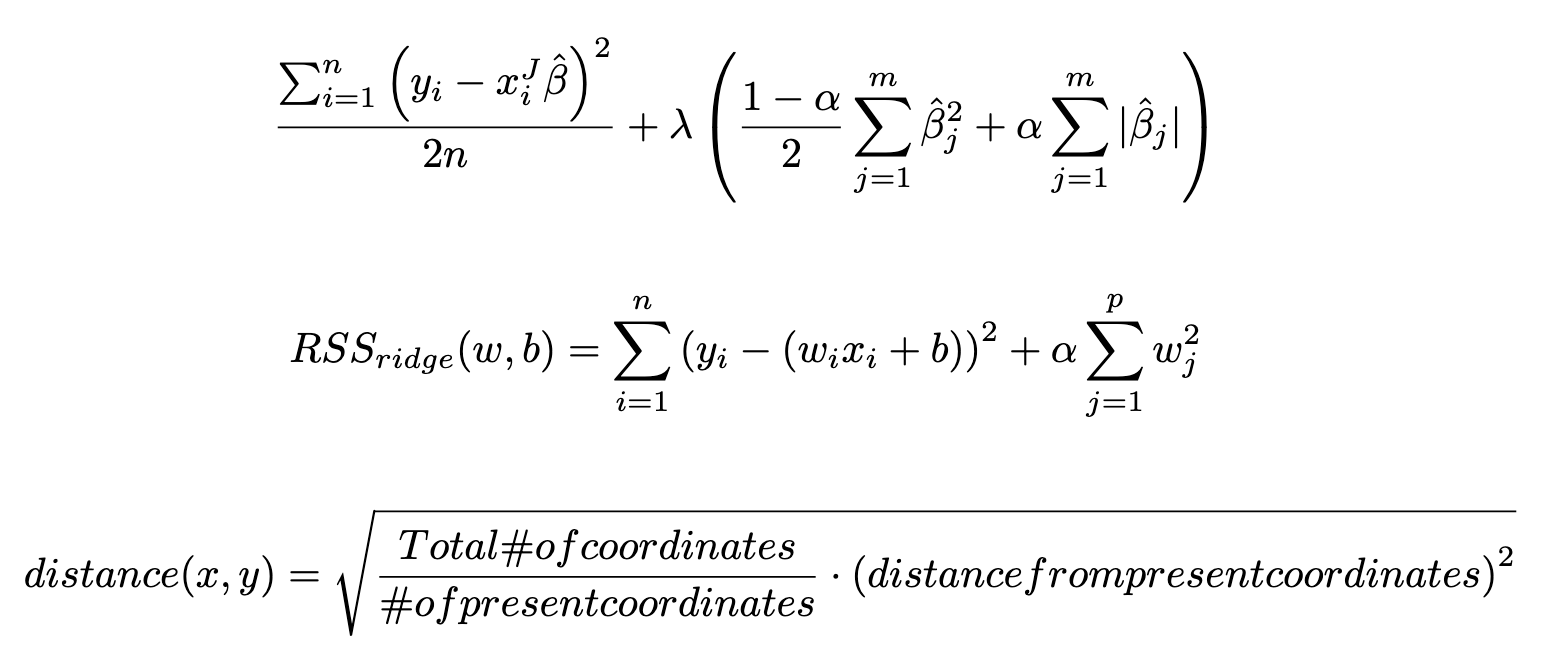
**3.2.3 Model Performance**

| **Metric** | **Result** |
| --- | --- |
| Coefficient of Determination (R2) | 0.82 |
| Root Mean Square Error (RMSE) | 0.43 |
| Mean Absolute Error (MAE) | 0.32 |

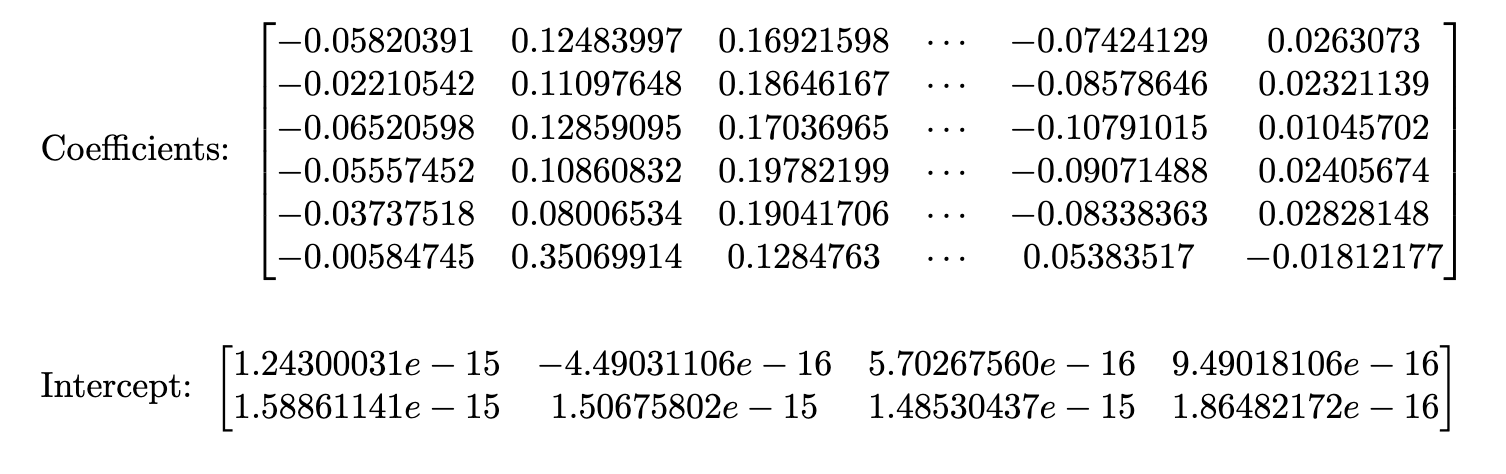
Considering the positive performance of the model, we re-ran the model on all our data as our final model, where we discuss the results of the hyperparameters and values in the model results. This model is uploaded in the submission folder, under the name “Multitask\_Elastic\_Net\_Model.joblib”.

**3.2.4 Model Results**

The chosen l1 ratio after cross-validation was 1.0, meaning that purely using Ridge regression resulted in the best outcome during training the model. Our model resulted in the optimization of the loss function for Ridge Regression:

****

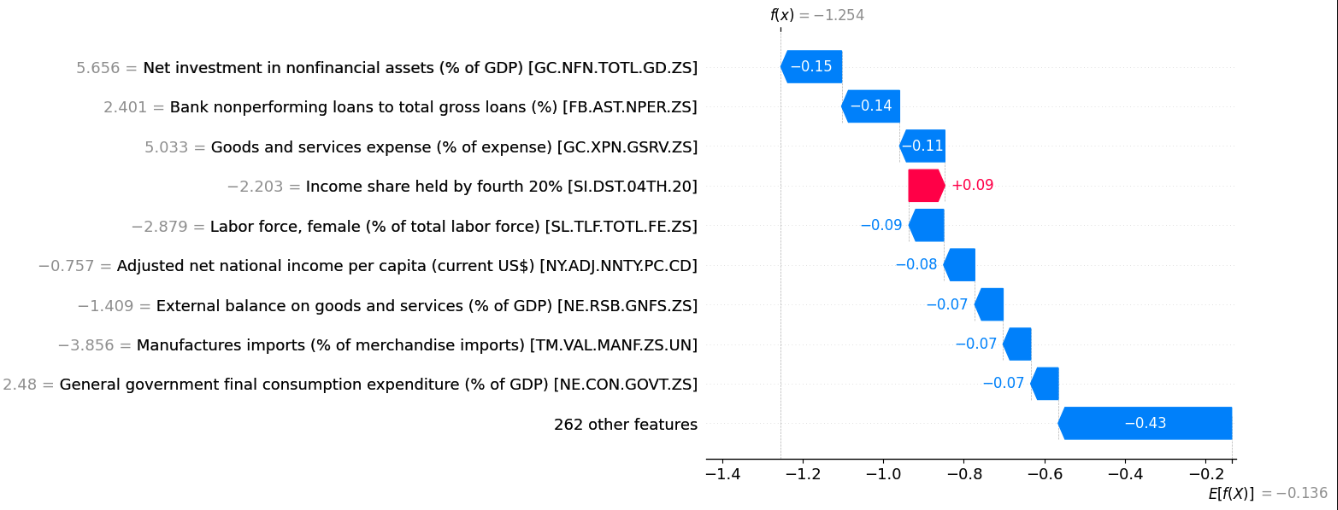
Alongside an alpha value of 0.002, we have the following coefficients and intercepts for our model:

****

Using these coefficients, intercepts, and hyperparameters, we performed SHapley Additive exPlanations (SHAP) to view in-depth what features are most relevant to the model predictions.



Taking a closer look, we can see how the model predicts the first observation. It starts at E[f(X)] = 0, and is moved forwards and backwards by the features seen below as the model makes predictions based on trained information.



Note: Scikit-learn’s elastic net regression uses one regularization strength coefficient alongside a proportional l1 ratio instead of two coefficients for each regularization method.

**4. Implications and Discussion**

The Human Capital Index (HCI), which measures how good countries are at mobilising their human capital through health and education, is the measure that positively correlates to logistical performance the most. Other measures include Current Health Expenditure, Gross National Income per capita, Imports of Goods and Services, and Electric Power Consumption. Conversely, high Poverty Headcount, Mortality Rates, and Power Outages severely impact supply chain performance.

## *Indicators with Positive Correlation:* Human Capital Index, Current Health Expenditure, Gross National Income per capita, Imports of Goods and Services, and Electric Power Consumption

Indicators that show investments in health, infrastructure (including electricity), as well as higher income per capita and higher import costs have positive correlation with LPI. From this, it is implied that countries with better healthcare and education systems, sufficient electronic infrastructure, and generally higher income (and therefore higher living costs) enjoy optimised logistics performance, driving down supply chain costs.

## *Indicators with Negative Correlation:* High Poverty Headcount, Mortality Rates, and Power Outages

Completely converse with the previous segment, it is implied that countries with high poverty, hence high socioeconomic disparity, mortality rates and lower electronic infrastructure development have higher supply chain costs as indicated by lower LPI values.

## **5. Recommendations**

**5.1 For Policymakers**

Based on the data, HCI and health expenditure directly correlate to supply chain performance. Hence, it is imperative that policymakers invest heavily in health, infrastructure, and education at a national level. Investing in public healthcare will also counteract the opposite end of the spectrum, such as mortality rate and power outages. Hence, the following recommendations are made:

1. Investing in public healthcare and education

Investment in human capital allows growth in the workforce and reduces mortality rates, contributing to the supply chain’s labour requirements .By investing in public healthcare, the government ensures equitable healthcare and reduces the cost per individual, which can decrease mortality rates, ensuring that more children grow up to productively contribute to the supply chain system.

1. Investing in Infrastructure

Investing in infrastructure can help reduce power outages, which are known to cause disruptions in just-in-time manufacturing processes, transportation, and inventory management. By ensuring that businesses have access to reliable sources of energy, policymakers essentially give them more leeway to pick manufacturing and inventory storage processes best suited to their products, thus facilitating a healthier supply chain. Furthermore, robust infrastructure allows for route optimization in logistics, enabling lower cost and faster delivery times.

1. Increase Ease of Doing Business by Regulatory Incentives for Businesses

Governments must also increase the ease of doing business, as the data shows that high costs of business start-up procedures negatively affect the supply chain. Several small businesses in an economy make it competitive and healthy. A healthy market is less susceptible to supply chain disruptions than monopolies or oligopolies. Hence, governments must create a supportive legal framework by simplifying regulations, streamlining approval processes, and subsidizing set-up costs for businesses in health, education, and green energy industries. This enables the market to fulfil its own needs: good infrastructure and cheap healthcare.

Thus, by investing in areas that improve the productivity of future citizens and promoting a healthy economy through supporting laws, the government can not just ensure a stable supply chain, but also reduce the cost of living.

**5.2 For Businesses**

In order to reduce costs and optimize its supply chains, businesses can take on a number of proactive strategies, such as:

1. **Optimize Location and Energy Resilience**Businesses should strategically choose locations for manufacturing and distribution based on access to reliable infrastructure, proximity to markets, and energy stability. In regions with unreliable power sources, companies can invest in alternative energy solutions like solar panels or independent power generation to avoid disruptions in manufacturing and logistics.
2. **Adopt Technology and Diversify Supply Chains**Leveraging advanced technologies like AI and automation can help businesses mitigate the risks of labour shortages and infrastructure inefficiencies. Diversifying suppliers and logistics networks across multiple regions reduces reliance on any single area, ensuring stability even in times of geopolitical or natural disruptions.
3. **Invest in Workforce Development and Community Support**By collaborating with local educational institutions and providing training programs, businesses can develop a skilled workforce to meet supply chain demands. Additionally, supporting healthcare and infrastructure projects through corporate social responsibility (CSR) initiatives contribute to long-term regional stability, which benefits both the local community and the supply chain

**Conclusion** In this report, we explored the intricate relationships between cost of living, economic development, and supply chain stability, leveraging diverse datasets and advanced modeling techniques. Through our analysis, we identified key indicators such as the Human Capital Index, health expenditure, Gross National Income per capita, and infrastructure metrics as critical drivers of supply chain resilience. Conversely, poverty, mortality rates, and power outages were found to negatively impact supply chain performance.

By implementing techniques such as KNN imputation, correlation analysis, and Elastic Net regression, we addressed data inconsistencies, reduced dimensionality, and built robust predictive models. The insights derived underscore the importance of investing in health, education, and infrastructure for policymakers and adopting technological advancements, diversification, and workforce development for businesses. These strategies not only enhance supply chain stability but also contribute to long-term economic growth and improved cost-of-living conditions.

Ultimately, fostering resilient supply chains requires a collaborative effort between governments and businesses. Policymakers must prioritize public investments and supportive regulations, while businesses must leverage innovation and strategic planning to mitigate risks. This synergy can lead to reduced costs, optimized logistics performance, and a stable, sustainable global supply chain system.

As global challenges continue to evolve, this report highlights the need for proactive and data-driven approaches to safeguard supply chain resilience in an interconnected world.

**7. Reference**

Behzadi, G., O’Sullivan, M. J., & Olsen, T. L. (2020). On metrics for supply chain resilience. *European Journal of Operational Research*, *287*(1), 145–158.<https://doi.org/10.1016/j.ejor.2020.04.040>

*GES\_ValueChain\_Ex2\_v1\_online*. (n.d.). Retrieved January 17, 2025, from<http://ceros.mckinsey.com/autocx-ex2-v1-online-2-2-2-2-2-2-3-1-2-1-2-4-2-1-2>

*Issue Brief: Supply Chain Resilience | CEA*. (2023, November 30). The White House.<https://www.whitehouse.gov/cea/written-materials/2023/11/30/issue-brief-supply-chain-resilience/>

(PDF) Exploring the relationship between National Logistics Cost and the Logistics Performance Index: An ASEAN perspective. (2024, October 10). *ResearchGate*.<https://www.researchgate.net/publication/384701638_Exploring_the_relationship_between_National_Logistics_Cost_and_the_Logistics_Performance_Index_An_ASEAN_perspective>

*Supply chain resilience in the face of change | McKinsey*. (n.d.). Retrieved January 17, 2025, from<https://www.mckinsey.com/capabilities/operations/our-insights/supply-chains-to-build-resilience-manage-proactively>

*Supply Chain Stability Index*. (n.d.). Retrieved January 17, 2025, from<https://kpmg.com/us/en/articles/2023/supply-chain-stability-index.html?utm_source=chatgpt.com>

1. Deloitte. (2023). *Global supply chain resilience amid disruptions*. Retrieved from https://www2.deloitte.com
2. McKinsey & Company. (2023). *Supply chains: To build resilience, manage proactively*. Retrieved from [https://www.mckinsey.com](https://www.mckinsey.com/capabilities/operations/our-insights/supply-chains-to-build-resilience-manage-proactively)
3. Organisation for Economic Co-operation and Development (OECD). (2021). *The economic impact of supply chain disruptions*. Retrieved from [https://www.oecd.org](https://www.oecd.org/)
4. United Nations Conference on Trade and Development (UNCTAD). (2021). *Global supply chain vulnerability exposed*. Retrieved from<https://unctad.org>
5. World Bank. (2023). *Logistics performance index 2023 report*. Retrieved from<https://www.worldbank.org>
6. World Economic Forum. (2022). *The cost of supply chain resilience*. Retrieved from<https://www.weforum.org>