Japan Suicide Statistics: Forecasting, Economic Impact, and Policy Simulation

警察庁
National Police Agency

Timeline: Daily ▾ December 6-11, 2025

Reason: Youth suicides reached 526 cases in 2024, the highest on record. Being based in Tokyo and working with youth programs, I saw the data gap firsthand and built a system to address it.

Author: Yulia Chekhovska

Japan Suicide Statistics: Forecasting, Economic Impact, and Policy Simulation

A concise project narrative for portfolio presentation

1. Situation

Japan's annual suicide statistics, published by the National Police Agency, provide a rich but fragmented dataset. Values are spread across multiple PDFs, structured with inconsistent formatting, and use Japanese era dating (昭和, 平成, 令和). Public health stakeholders often lack automated tools for trend monitoring, forecasting, or estimating the economic burden of suicides by demographic group. My task was to build a reproducible, evidence-driven analytical system that transforms raw police reports into actionable insights for prevention policy.

2. Task

I set out to design a pipeline that would:

1. Scrape and consolidate historical suicide statistics.
2. Clean, normalize, and structure age-, gender-, and reason-specific data.
3. Produce interpretable visualizations and comparative analyses.
4. Quantify economic losses using lifetime productivity and tax revenue models.
5. Run policy simulations (e.g., 15 percent reduction scenarios) to estimate potential savings.

6. Explore time-series forecasting feasibility for demographic subgroups.

3. Action

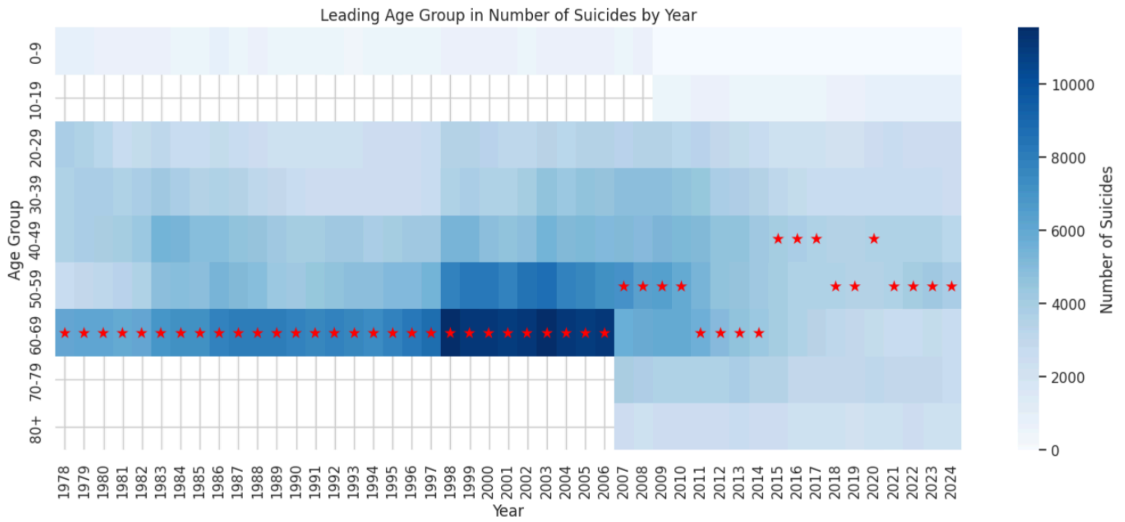
Data Engineering and Preparation

- Automated ingestion of NPA annual tables and conversion of Japanese eras to Gregorian years.
  Constructed three master datasets: age_cleaned, gender_cleaned, and reason_cleaned.
- Standardized demographic categories, cleaned missing values, and harmonized “unknown” and “total” entries.
- Designed a tidy dataset enabling cross-tab analysis by year, age, gender, and cause.

Exploratory and Statistical Analysis

- Produced demographic trend dashboards highlighting shifts in suicide burden across cohorts.

- Built comparative tables of suicides, loss per case, and total annual economic burden by age group.



Economic Impact Model

- Developed a loss-per-case estimation model incorporating:
  - expected remaining working years
  - age-specific productivity curves
  - discounted lifetime tax contributions
- Aggregated results into annual economic losses.

Forecasting Attempts

- Implemented baseline linear regression, ARIMA, XGBoost, and Prophet to model age- and gender-specific suicide trajectories.
- All time-series approaches underperformed (negative R^2 in several cohorts) due to short time windows and high volatility, demonstrating that reliable forecasting for small demographic segments is not feasible.
- Documented reasons for failure and proposed alternative approaches (e.g., rolling means, hierarchical modeling).

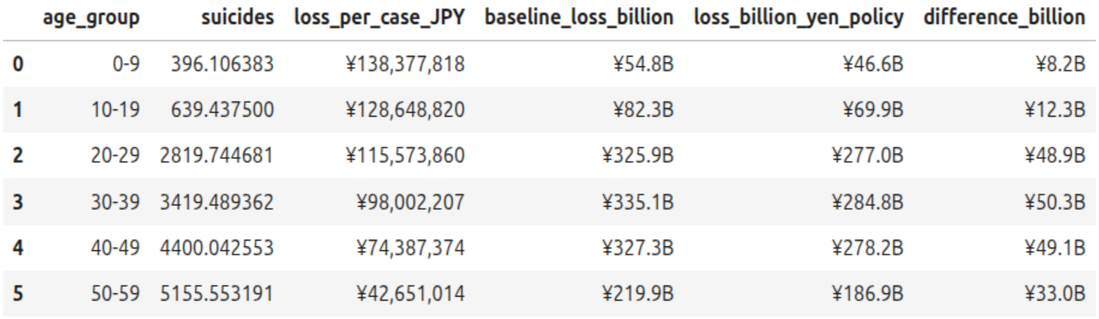
4. Results

Analytical Findings

- Middle-aged groups (30–49) generate the highest economic loss despite lower loss per case, due to high suicide count.

- Younger groups (0–19) show high loss per case but low total impact because suicides are less frequent.

- A 15 percent reduction policy scenario yields very large savings, with the strongest economic effects in the 20–49 cohorts.



	age_group	suicides	loss_per_case_JPY	baseline_loss_billion	loss_billion_yen_policy	difference_billion
0	0-9	396.106383	¥138,377,818	¥54.8B	¥46.6B	¥8.2B
1	10-19	639.437500	¥128,648,820	¥82.3B	¥69.9B	¥12.3B
2	20-29	2819.744681	¥115,573,860	¥325.9B	¥277.0B	¥48.9B
3	30-39	3419.489362	¥98,002,207	¥335.1B	¥284.8B	¥50.3B
4	40-49	4400.042553	¥74,387,374	¥327.3B	¥278.2B	¥49.1B
5	50-59	5155.553191	¥42,651,014	¥219.9B	¥186.9B	¥33.0B

Forecasting Evaluation

- Demonstrated that off-the-shelf forecasting is unsuitable for age-specific suicide data in Japan due to limited data size.
- Provided justification for switching to non-parametric smoothing or future deep learning models once more years accumulate.

This project integrates public health analytics, applied econometrics, and data engineering. It shows domain understanding, technical fluency, and honest scientific judgment about model limitations. The output is a practical decision-support tool that turns government statistics into interpretable risk profiles and cost-saving scenarios relevant to policymakers, NGOs, and researchers.

Data Cleaning and Processing

- Pandas GroupBy and merge operations
- Regular expressions for category normalization
- Custom logic for age group harmonization and “Total/Unknown” handling
- Datetime preprocessing for era → Gregorian conversion
- Quality checks: missing-value auditing, distribution checks, cross-dataset consistency

Visualization and Reporting

- Matplotlib
- Seaborn
- Horizontal and vertical bar charts, heatmaps, demographic trend visualizations
- Jupyter outputs rendered as portfolio-friendly figures

Statistical and Machine Learning Modeling

- Scikit-learn
  - Linear Regression
  - Train-test split
  - Cross-validation attempts

  XGBoost
  - XGBRegressor
- Prophet
  - Additive regression model for seasonality and long-term trend
- Statsmodels
  - ARIMA and SARIMA