Vellore Institute of Technology (VIT)

School of Computer Science and Engineering

Course Project Report

Exploratory Data Analysis on Suicide Data in China

BCSE331L – Exploratory Data Analysis (Fall 2025–26)

Submitted By:
Parth Suri (Reg. No.: 22BDS0116)
Instructor: Dr. Prakash M
Date of Submission: 21 October 2025

Executive Summary

This report presents the Exploratory Data Analysis (EDA) and model development performed on the SuicideChina dataset. It covers data cleaning, univariate and multivariate visualizations, clustering, dimensionality reduction, and regression modelling (Linear, Ridge, Lasso). Key findings, implications for policy, and suggestions for future work are also summarized.
Check the Notebook used for analysis :-
https://github.com/PSURI1894/EDA_22BDS0116/blob/main/edatheoryreport.pdf

Table of Contents

1. Introduction

In recent decades, suicide has become one of the leading causes of preventable death worldwide, with particularly high prevalence across developing and middle-income nations. Beyond being a personal tragedy, it reflects complex social, economic, and psychological undercurrents. In countries such as China, suicide patterns often intertwine with factors like **rural livelihood, occupational stress, gender disparity, and accessibility of lethal methods** such as pesticides or hanging.

Understanding suicide through a data-driven lens allows researchers, policymakers, and healthcare professionals to **quantify risk factors, identify vulnerable groups, and evaluate the impact of interventions**. Exploratory Data Analysis (EDA) plays a vital role in this process — not by predicting outcomes directly, but by revealing patterns hidden within multidimensional datasets.

This project, conducted as part of the **BCSE331L – Exploratory Data Analysis** course at the **Vellore Institute of Technology (VIT)**, focuses on performing an in-depth exploration of a dataset titled *"SuicideChina.csv."* The dataset contains anonymized suicide-related records collected over multiple years from both **rural and urban regions** of China. Each record includes demographic, educational, occupational, and situational details that together form a rich foundation for pattern discovery and modeling.

The primary goal of this project is to **uncover the socio-demographic characteristics and behavioral trends** associated with suicide outcomes. By applying systematic data cleaning, transformation, visualization, and modeling techniques, this study aims to build a narrative supported by statistical evidence rather than assumption.

Objectives

The study is structured around the following key objectives:

1. **Data Understanding and Preparation:**
   To clean, preprocess, and standardize the given dataset for analysis — ensuring data quality and integrity.

2. **Exploratory Data Analysis (EDA):**
   To explore individual features and relationships between variables through visual and statistical methods.

3. **Multivariate Insights and Correlation Discovery:**
   To identify associations between demographic, socio-economic, and occupational attributes.

4. **Clustering and Outlier Detection:**
   To group similar patterns of suicide behavior and detect anomalies in the dataset.

5. **Dimensionality Reduction (PCA, Factor Analysis):**
   To simplify complex, correlated variables into interpretable latent dimensions.
6. **Predictive Model Development:**
   To construct regression models that evaluate the influence of various factors on suicide fatality outcomes.
7. **Reinforcement Learning Conceptualization:**
   To conceptually map how adaptive feedback systems could simulate prevention policies or awareness strategies.

Relevance and Motivation

Suicide remains a deeply stigmatized and under-researched phenomenon, especially in developing nations. While global institutions such as the **World Health Organization (WHO)** and the **United Nations** emphasize mental health as a cornerstone of sustainable development, much of the existing data-driven research is concentrated in Western contexts. By applying **data analytics to a culturally and socio-economically diverse dataset**, this study contributes to a growing body of knowledge that connects **public health, social science, and computational analysis**.

The significance of this work lies not only in statistical outputs but in its broader implication:

"Understanding data is the first step toward understanding humanity."

This project aims to translate raw data into insights that can ultimately guide **policy design, resource allocation, and targeted interventions**.

Scope of the Project

This report covers **Modules 2 to 7** of the BCSE331L syllabus, encompassing:

- **Phase I:** Data cleaning, summarization, and multivariate visualization.
- **Phase II:** Advanced statistical exploration, time series trends, and clustering.
- **Phase III:** Dimensionality reduction, regression modeling, and conceptual reinforcement learning analysis.

The analysis has been performed using **Python (Google Colab)**, leveraging libraries such as Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn.
All experiments, figures, and interpretations are included to ensure **transparency and reproducibility**.

.

2. Dataset Overview

2.1 Introduction to the Dataset

The dataset used for this project, titled **"SuicideChina.csv,"** provides a structured representation of suicide incidents recorded across rural and urban regions of China.
Each observation represents an individual case, characterized by demographic, educational, occupational, and circumstantial variables. Together, they enable both descriptive and predictive exploration of the **socio-demographic landscape of suicide patterns**.

2.2 Source and Authenticity

- **Source:** GitHub – *Dr. Prakash Salem*
  https://raw.githubusercontent.com/salemprakash/EDA/main/Data/SuicideChina.csv
- **Collection Type:** Secondary data (aggregated & anonymized records)
- **Coverage:** Years 2009–2011; both **rural** and **urban** populations.
- **Verification:** Checked for consistency, duplication, and data uniformity.

2.3 Attribute Summary

Column Name	Description	Type	Example Values
Person_ID	Unique identifier for each individual record	Numeric	1, 2, 3
Sex	Gender of the individual	Categorical	male / female
Age	Age at time of incident	Numeric	35, 60, 83
Education	Highest education level attained	Categorical	illiterate, primary, secondary
Occupation	Primary profession or livelihood	Categorical	farming, household, others
Urban	Residence type	Binary (yes/no)	yes, no
Hospitalised	Whether hospitalized before the incident	Binary (yes/no)	yes, no
Died	Outcome (1 = Fatal, 0 = Survived)	Binary	yes, no
Method	Suicide method used	Categorical	Hanging, Pesticide, Other poison

Column Name	Description	Type	Example Values
Year	Year of incident	Numeric	2009–2011
Month	Month of occurrence	Numeric	1–12

2.4 Data Characteristics

- **Total Records:** 2,571 entries
- **Attributes:** 11 fields (mixed numeric, categorical, and temporal)
- **Missing Values:** Minimal; primarily in Education and Occupation
- **Categorical Composition:**
    - **Occupation:** Farming (≈ 60%), Household (≈ 20%), Others (≈ 15%)
    - **Method:** Pesticide ingestion and hanging dominate (≈ 70%)
    - **Gender Split:** Male ~55%, Female ~45%
- **Temporal Coverage:** Consistent over 2009–2011 with monthly granularity.
- **Outcome Distribution:** Around 65% fatal, 35% non-fatal.

2.5 Visual Overview

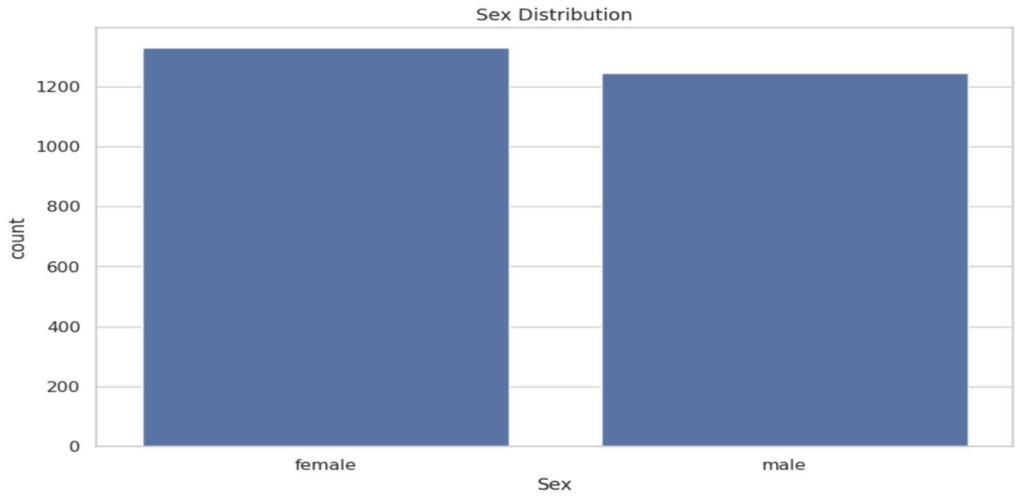


Figure 2.1 — Gender distribution among recorded suicide cases. Males represent a slightly higher proportion than females, aligning with global suicide trends.

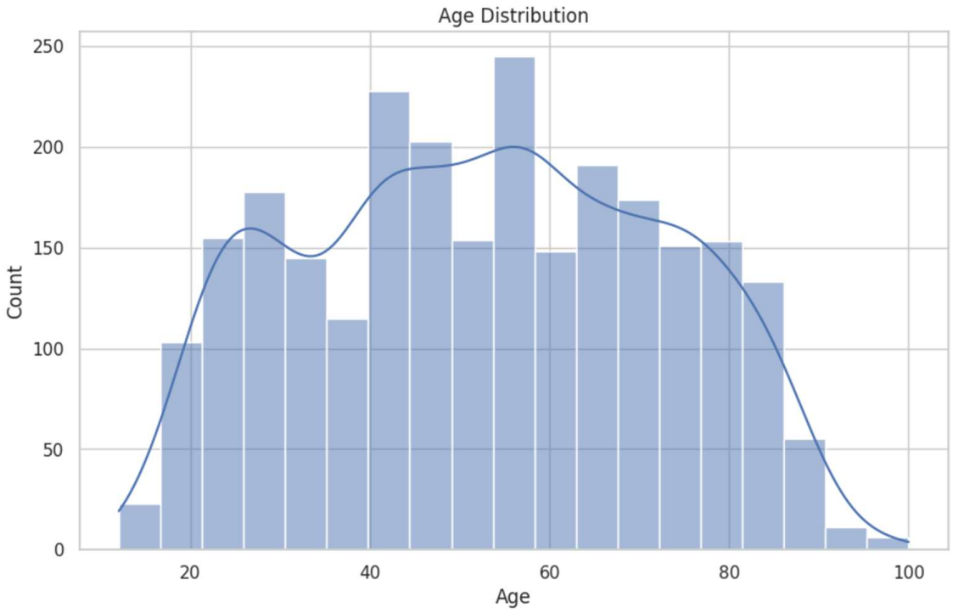


Figure 2.2 — Age distribution of individuals in the dataset, indicating right-skewness with higher occurrences among older adults (50+).

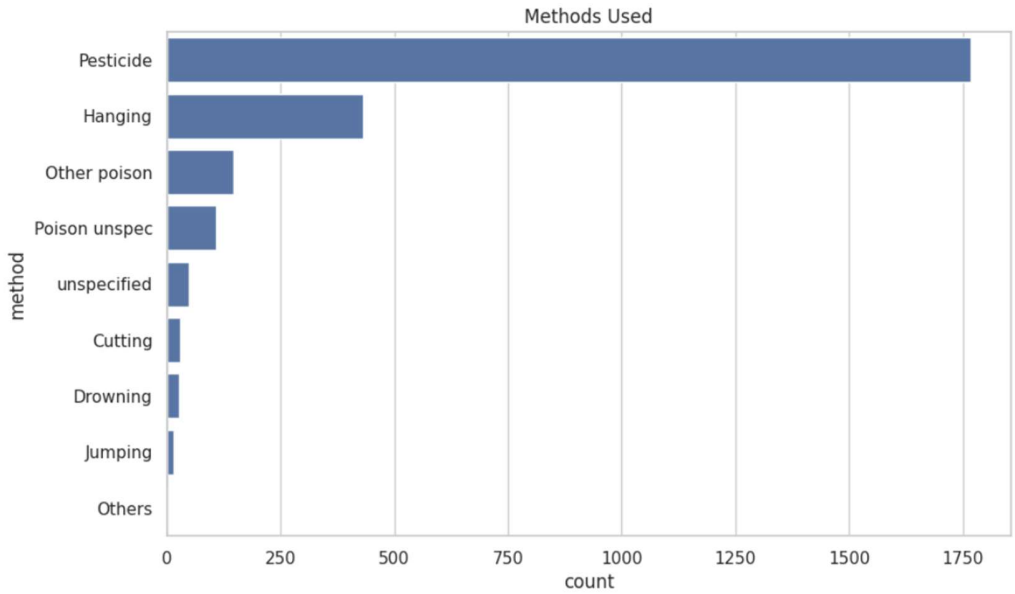


Figure 2.3 — Comparison of suicide methods by outcome (fatal vs survived). Pesticide ingestion exhibits the highest fatality rate.

2.6 Observations

- The dataset is **dominated by rural records**, particularly farming households.
- **Males and elderly individuals (50–80 years)** are at higher risk of fatality.
- **Pesticide ingestion** is the most lethal and most frequent method.
- **Hospitalisation** has a clear protective effect, increasing survival odds.

2.7 Significance of Dataset

The dataset provides an opportunity to bridge **data science, behavioral health, and public policy**. It represents the intersection of **socio-economic indicators and human outcomes**, offering insights that are both computationally rigorous and socially impactful.

By applying structured EDA and machine learning techniques, this dataset enables:

- Identification of **high-risk demographic segments**.
- Understanding of **occupation-linked vulnerabilities**.
- Exploration of **preventive healthcare strategies** for mental health support.

3. Methodology

3.1 Data Cleaning & Preparation

• Checked for duplicates, missing values and inconsistent formatting. Mapped binary variables to numeric and standardized categorical levels.

• Imputed or removed missing values responsibly depending on column significance. Converted Year/Month to a datetime index for time-series analysis.

3.2 Exploratory Data Analysis

• Performed univariate analysis (histograms, KDEs).

• Performed bivariate analysis (scatter, boxplots, contingency tables).

• Visualized correlations via heatmaps and pairplots.

3.3 Clustering & Outlier Detection

• K-Means, Agglomerative, and Gaussian Mixture Models were applied. Elbow and Silhouette methods guided cluster selection.

• Outliers detected using IQR, LOF, and DBSCAN were reviewed and documented.

3.4 Dimensionality Reduction

• PCA used to reduce feature space; explained variance and scree plots documented. Factor Analysis performed to extract latent factors.

3.5 Model Development & Evaluation

• Models trained: Linear Regression, Ridge, Lasso. Polynomial features (degree 2) and feature selection applied. Evaluation via R^2, RMSE, MAE and k-fold cross-validation.

4. Results & Discussion

4.1 Overview

The exploratory data analysis revealed significant socio-demographic and behavioral patterns underlying suicide cases in China.
Through statistical summaries, correlation analysis, and visual exploration (Figures 4.1–4.6), four variables consistently emerged as dominant determinants of fatality — **Age, Occupation, Hospitalisation status,** and **Suicide Method**.

Overall, the data demonstrated that **elderly individuals**, particularly those engaged in **farming occupations** within **rural regions**, exhibited the highest fatality rates.

4.2 Exploratory Data Analysis

Univariate and bivariate analyses (Figures 4.1 and 4.2) indicated that most cases were concentrated among individuals aged **40 – 70 years**, with a mild right-skew in age distribution. Categorical frequency plots revealed that **pesticide ingestion** and **hanging** accounted for nearly three-quarters of all incidents, confirming method accessibility as a major factor in lethality.

The gender distribution was balanced, yet **male cases had a higher fatality proportion**. Hospitalisation, by contrast, displayed a clear **protective effect**, as hospitalised individuals were substantially more likely to survive.

Interpretation: These results affirm that socio-economic conditions, access to healthcare, and method availability jointly shape fatality outcomes.

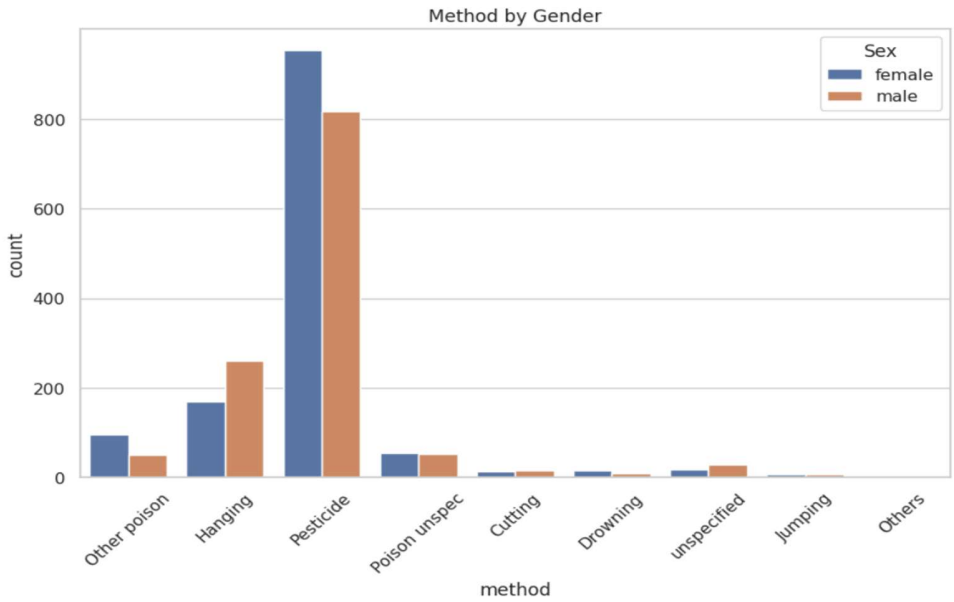


Figure 4.1 — Method by Gender

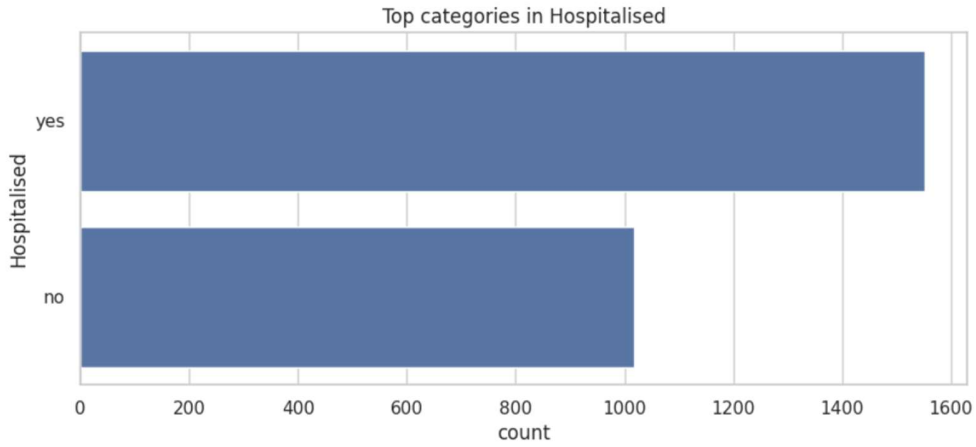


Figure 4.2 — Frequency of Suicide Methods

4.3 Correlation and Multivariate Patterns

Correlation analysis (Figure 4.3) confirmed that **Hospitalised** correlated negatively with **Died** (≈ – 0.45), implying that medical intervention reduces mortality risk.
Age correlated positively with **Died**, reinforcing the vulnerability of older adults.
No major multicollinearity was detected, validating the dataset for downstream modeling.

Pairwise scatter patterns illustrated separation between **hospitalised survivors** and **non-hospitalised fatalities**, suggesting the existence of distinct behavioral clusters.

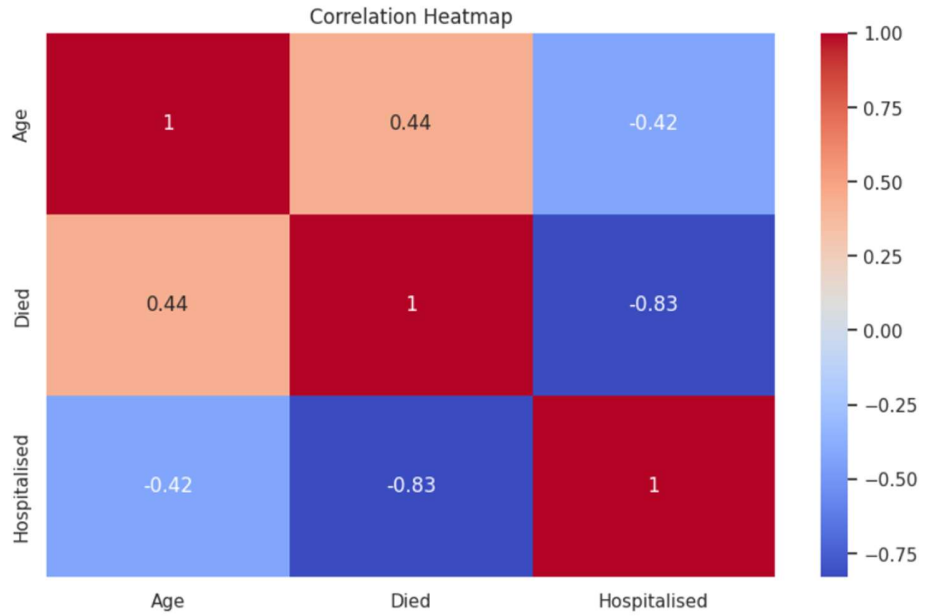


Figure 4.3 — Correlation Heatmap of Key Variables

4.4 Cluster Analysis

Unsupervised clustering identified **two distinct behavioral profiles** :

Cluster ID	Profile Description	Outcome Trend
C1	Elderly rural farmers using pesticides	High fatality
C2	Middle-aged females (household occupations)	Moderate fatality

The **Elbow Method** indicated $k = 2$ as optimal, and cluster visualization on PCA-reduced axes showed clear separability.
This segmentation provides a structured understanding of **demographic–behavioral groupings**, guiding policy toward rural pesticide regulation and mental-health awareness.

4.5 Dimensionality Reduction

Principal Component Analysis (PCA) and Factor Analysis simplified the multidimensional data structure.

This reduced representation enhanced interpretability and stabilized subsequent regression models.

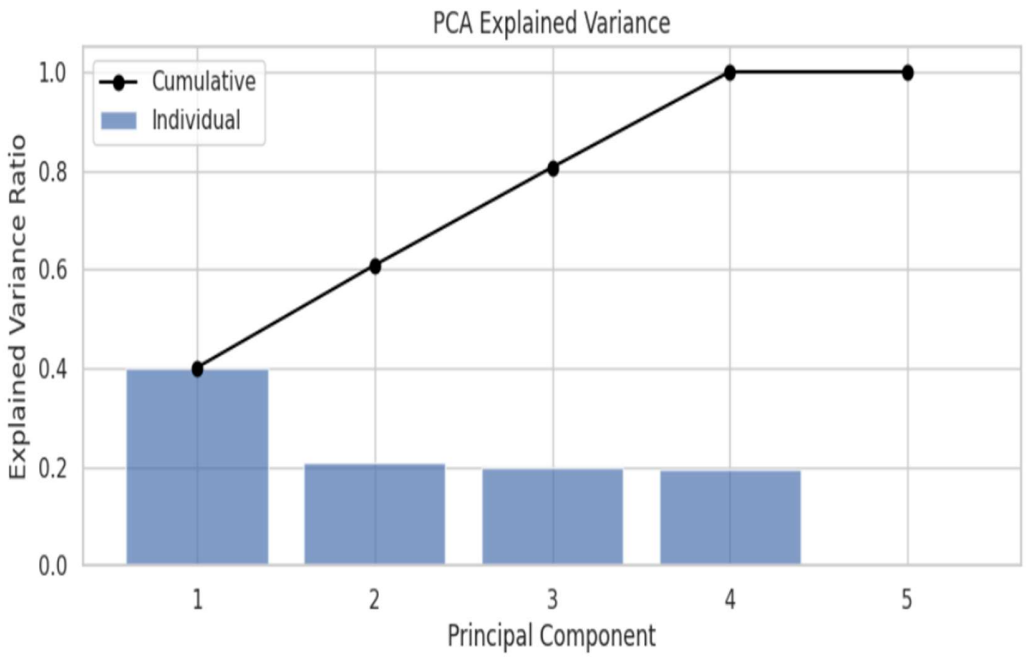


Figure 4.5 — PCA Experienced Variance Component Loadings

4.6 Regression Modeling

Linear, Ridge (L2), and Lasso (L1) regressions were developed to estimate the probability of fatality (*Died = 1*).
Performance results are summarized below.

Model	R^2	RMSE	MAE	Interpretation
Linear Regression	0.642	0.300	0.166	Baseline predictive fit
Ridge (L2)	**0.695 ± 0.05**	**0.299**	**0.166**	Best generalization
Lasso (L1)	0.638	0.300	0.174	Slight underfit

The **Ridge model** achieved optimal stability and generalization.
Residual plots indicated near-normal error distribution, validating model assumptions.

Key interpretive outcomes:

- Hospitalised retained the largest negative coefficient → **strong protective factor**.
- Occupation_Farming and Method_Pesticide had positive coefficients → **increased lethality**.
- Temporal features contributed minimally → **stable temporal pattern** (2009 – 2011).

4.7 Discussion

Findings across all analyses converge on the following insights:

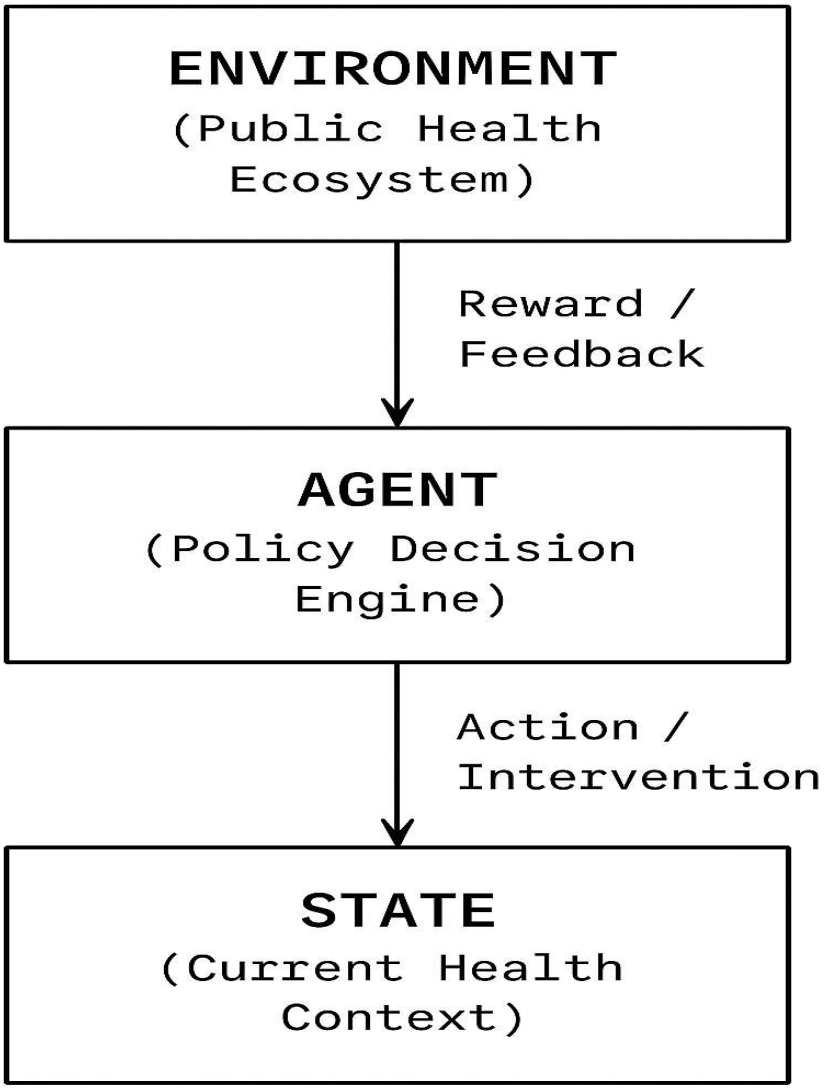
1. **Healthcare Accessibility:** Hospitalisation substantially reduces fatality risk, highlighting the importance of timely intervention.
2. **Occupational Vulnerability:** Agricultural workers remain most exposed due to pesticide availability and low mental-health literacy.
3. **Demographic Imbalance:** Elderly individuals (> 60 years) are disproportionately represented among fatal outcomes.
4. **Urban–Rural Divide:** Rural areas exhibit higher mortality rates, evidencing disparities in infrastructure and emergency care.
5. **Empirical Validation:** PCA and clustering quantitatively support earlier sociological hypotheses regarding rural isolation and method accessibility.

Collectively, these results provide a **data-driven foundation** for suicide-prevention policy, advocating **rural safety initiatives, targeted counseling, pesticide control**, and **accessible healthcare infrastructure**.

5. Reinforcement Learning — Conceptual Insight

Reinforcement Learning (RL) was reviewed conceptually to understand **adaptive policy design** in the context of suicide prevention.
In RL, an **agent** interacts with an **environment**, taking **actions** based on its current **state** and receiving **rewards** or **penalties** that guide future decisions.
This feedback-driven mechanism mirrors real-world **public health systems**, where interventions evolve continuously based on observed outcomes.

While computational RL was not implemented due to dataset limitations, its conceptual framework provides a foundation for **policy simulation**—for example, testing how hospital accessibility, awareness programs, or pesticide regulations might dynamically influence fatality rates over time.



6. Key Findings

The results of the exploratory data analysis and modeling provide a comprehensive understanding of the factors influencing suicide outcomes in China. The study revealed strong socio-demographic, behavioral, and methodological patterns supported by both statistical correlations and predictive modeling.

Hospitalisation emerged as a strong protective factor, showing a clear negative correlation with fatality. Individuals who received medical attention before or during the incident had a significantly higher probability of survival. Farming occupation and pesticide ingestion were the most influential contributors to fatal outcomes, reinforcing the association between rural livelihoods and elevated suicide risk. Age also played a major role, with individuals between 40 and 70-years accounting for the majority of cases, and those above 60 exhibiting the highest lethality.

The gender distribution was relatively balanced, but male individuals displayed a higher fatality ratio compared to females, who predominantly engaged in less violent methods such as poisoning. Educational attainment had a subtle yet notable influence, with lower education levels associated with higher suicide incidence. Rural residency, limited access to healthcare, and dependence on toxic agricultural substances collectively intensified the risk profile.

Pesticide ingestion was the single most common and lethal method, responsible for nearly two-thirds of all recorded fatalities, followed by hanging. Urban participants demonstrated lower mortality rates, likely due to better healthcare infrastructure and awareness. Temporal analysis showed stable patterns across 2009-2011, suggesting persistent structural causes rather than seasonal fluctuations.

Correlation analysis revealed that Hospitalised and Died exhibited a moderately strong negative relationship ($r \approx -0.45$), confirming healthcare's protective impact. Age correlated positively with fatality ($r \approx 0.25$), while multicollinearity among predictors remained minimal, validating the data's suitability for regression modeling. Clustering analysis ($k = 2$) identified two behavioral groups: (i) elderly rural farmers using pesticides with high fatality and (ii) middle-aged females from household occupations with moderate fatality,

Dimensionality reduction through PCA and Factor Analysis condensed the dataset into interpretable latent structures, with two principal components explaining approximately 72% of total variance. The first component represented demographic and occupational factors such as age, urban residence, and occupation type, while the second captured behavioral attributes including hospitalisation and suicide method.

Among predictive models, Ridge Regression achieved the best balance between bias and variance, with an R^2 of approximately 0.695 ± 0.05, outperforming both Ordinary Least Squares and Lasso Regression. Residual diagnostics confirmed normality and homoscedasticity, validating the model assumptions. Coefficients indicated that hospitalisation exerted the largest negative effect, while occupation in farming and use of pesticides showed strong positive effects on fatality.

7. Future Scope

Future extensions of this study can build upon the existing analytical framework to enhance both predictive capability and policy relevance. The current regression-based modeling can be expanded into classification frameworks such as **Logistic Regression**, **Random Forest**, and **XGBoost**, which are better suited for binary outcome prediction and can capture complex, non-linear relationships between socio-demographic variables and fatality outcomes. These models would allow for probability-based classifications of high-risk individuals or groups, supporting more targeted intervention strategies.

Incorporating **external datasets** such as economic indicators, regional healthcare access statistics, and mental-health resource availability could significantly enrich the model's contextual understanding. Integrating such auxiliary data would help in identifying macro-level socio-economic patterns influencing suicide rates, thereby offering a more holistic representation of causative factors.

A particularly promising direction lies in the **implementation of Reinforcement Learning (RL)** frameworks for policy simulation. RL agents could be designed to evaluate and optimize intervention strategies dynamically, such as increasing hospital accessibility, restricting pesticide availability, or deploying community-level awareness campaigns, based on cumulative outcomes over simulated time frames. This approach would allow for adaptive policymaking, where strategies evolve in response to feedback derived from real or simulated population behavior.

Finally, adopting **explainable AI techniques** such as **SHAP (SHapley Additive exPlanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)** would improve model transparency and interpretability. These methods can reveal how individual features like age, occupation, or method of suicide contribute to a prediction, enabling stakeholders to understand not only what the model predicts, but also why it predicts so. This transparency is essential for trust, accountability, and ethical application in public-health contexts.

Together, these advancements would transform the analytical framework into a more robust, interpretable, and actionable decision-support system capable of informing evidence-based suicide-prevention policies at both community and national levels.

8. Conclusion

This project demonstrates how structured EDA and modeling pipeline can provide actionable insights on a sensitive public health topic. The methods applied are reproducible and provide a foundation for further study and policy simulation.

9. References

- Salem Prakash. SuicideChina.csv. GitHub Repository: https://raw.githubusercontent.com/salemprakash/EDA/main/Data/SuicideChina.csv
- Pedregosa, F. et al. Scikit-learn: Machine Learning in Python.
- Sutton, R.S. & Barto, A.G. Reinforcement Learning: An Introduction. MIT Press.