# Suicide Data Analysis Report

## Introduction

This report delves into the intricate patterns of suicide data spanning from 2001 to 2012, specifically focusing on professional profiles, means adopted, and various causes. By analyzing State/UT-wise and age-sex wise details of suicide victims, the study identifies occupations, methods, and root causes associated with higher vulnerability.

## Objective

The primary objective of this project was to investigate and discern any substantial disparities in suicide rates among male and female populations, as well as different age groups.

## Methodology

The analysis involves cleaning the data, filtering relevant categories, and utilizing various statistical methods for exploration. Visualization techniques, including line plots and bar charts, are employed to better understand the trends. Additionally, hypothesis testing and Statistical testing are utilized to investigate differences and relationships within the dataset.

## Exploratory Data Analysis (EDA)

### Dataset Overview

**Original Data Sources (Data.gov.in) :**

1. State/UT-wise distribution of suicides by means adopted during 2001-2012
2. State/UT-wise professional profile of suicide victim during 2001-2012
3. State/UT-wise distribution of suicides by causes during 2001-2012

OR

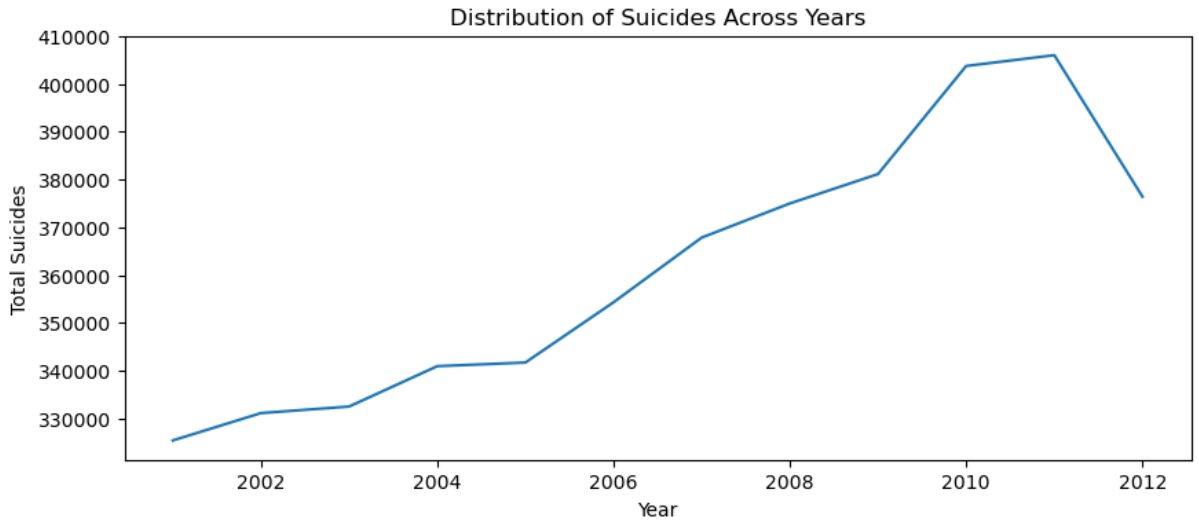
**Kaggle Dataset Download Link:**

- Suicides in India

- The dataset contains 237,519 entries with 7 Variables: State, Year, Type_code, Type, Gender, Age_group, and Total.

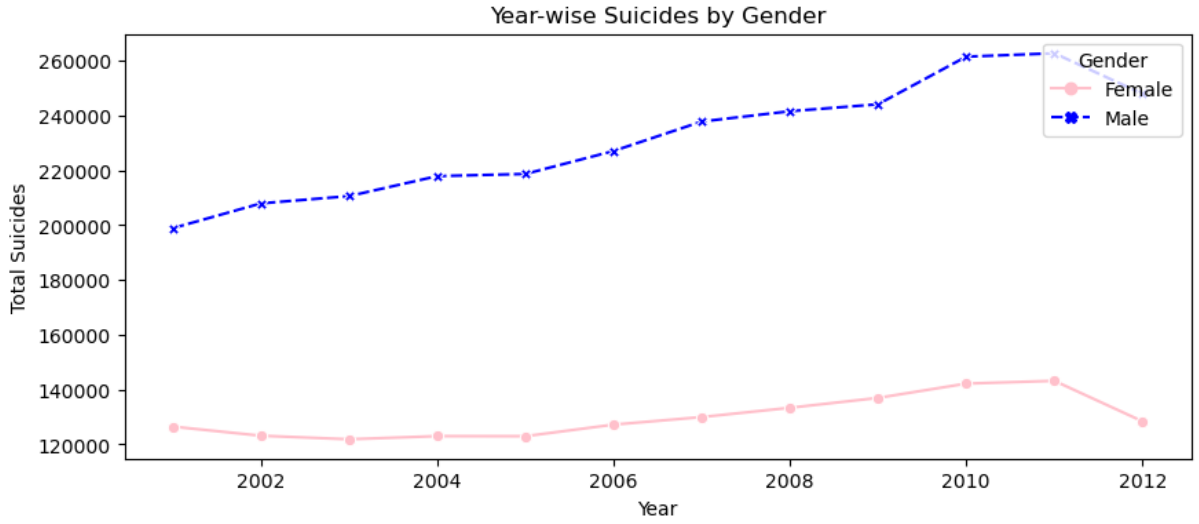
| State | Year | Type_code | Type | Gender | Age_group | Total |
|---|---|---|---|---|---|---|
| A & N Islands | 2001 | Causes | Illness (Aids/STD) | Female | 0-14 | 0 |
| A & N Islands | 2001 | Causes | Bankruptcy | Female | 0-14 | 0 |
| A & N Islands | 2001 | Causes | Cancellation | Female | 0-14 | 0 |
| A & N Islands | 2001 | Causes | Physical Abuse | Female | 0-14 | 0 |
| A & N Islands | 2001 | Causes | Dowry Dispute | Female | 0-14 | 0 |

### Total Suicide Rates in India (2001-2012)



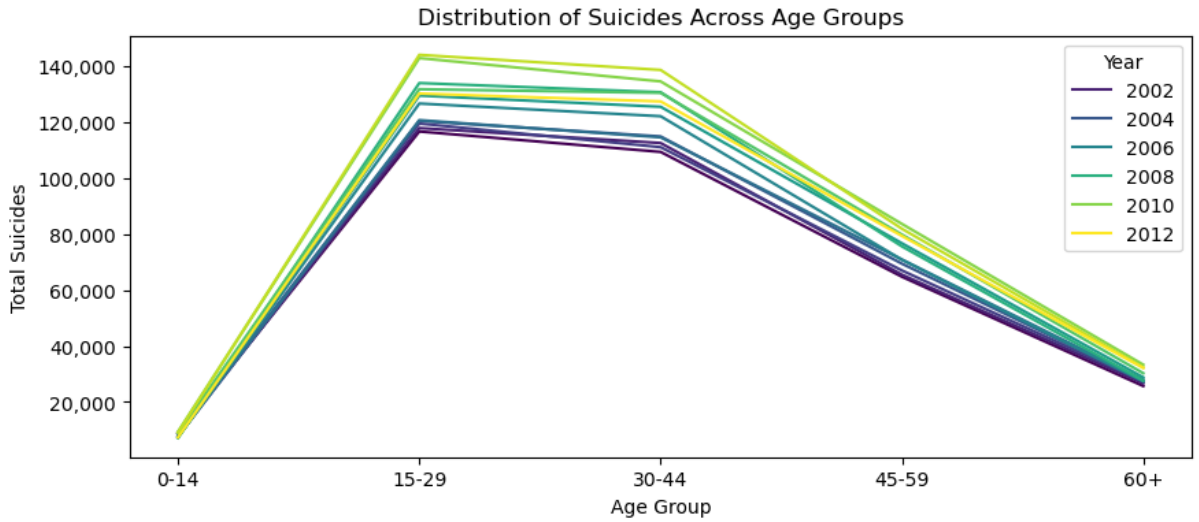

Total suicide rates increased from 2001 to 2010, reaching a peak of 405,989, before declining in 2012.

## Gender Distribution in Suicides in India (2001-2012)




In the above graph, we can conclude that males consistently had higher suicide rates than females throughout the period. **Both genders displayed a similar pattern of fluctuations over the years.**

## Suicides across Age Groups Distribution




In the graph from 2001 to 2012, suicide rate trends emerge across age groups in India. The "0-14" age group maintains low figures, **while "15-29" and "30-44" consistently report the highest numbers,** indicating vulnerability in young adults. Interestingly, a decline is observed in the "60+" age group, suggesting a shift in vulnerability towards older demographics. This nuanced pattern reflects age-related trends in suicides over the specified period.

## Yearly Distribution of Suicide Types:

**2001 - Top Suicide Types Among Different Causes:**

| Cause | Count | Percentage |
|---|---|---|
| By Hanging | 29,757 | 9.14% |
| Family Problems | 24,162 | 7.42% |
| House Wife | 21,659 | 6.66% |

**Hanging, family problems, and housewife-related** stress were predominant factors in suicides among different professions.

**2012 - Top Suicide Types Among Different Causes:**

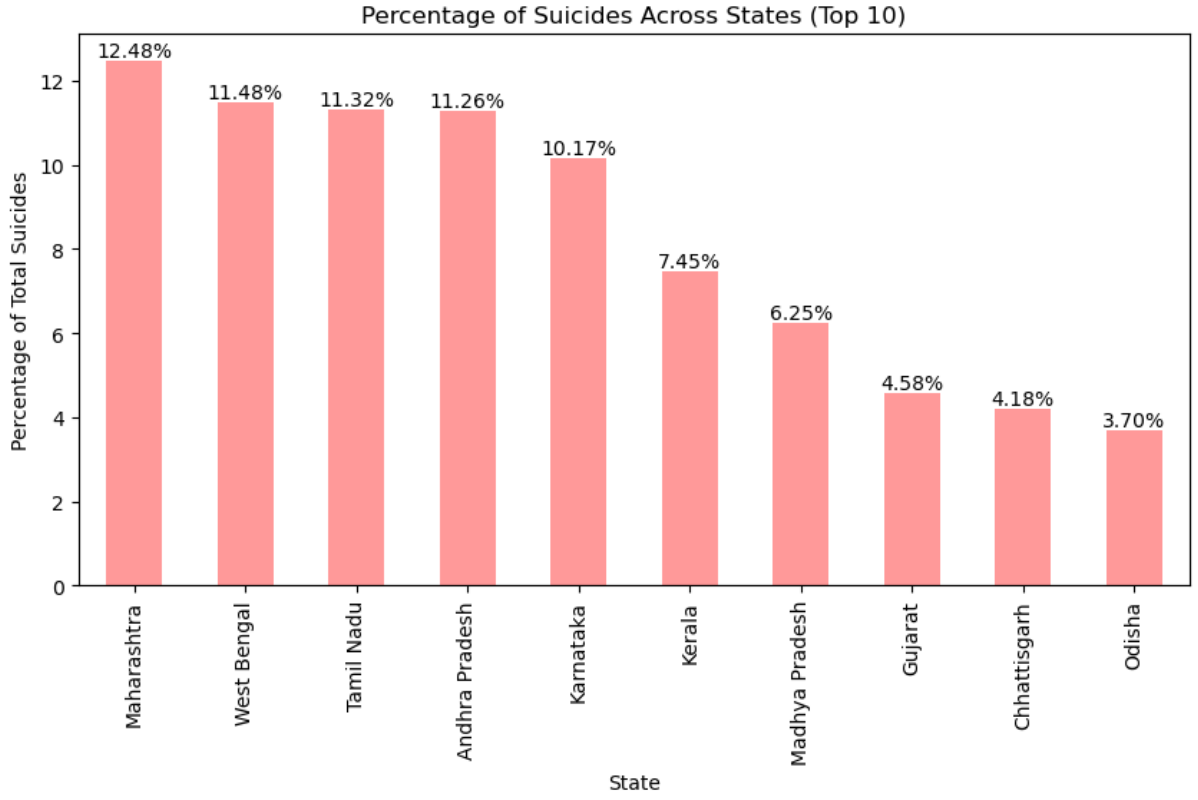
| Cause | Count | Percentage |
|---|---|---|
| By Hanging | 50,062 | 13.30% |
| Family Problems | 30,792 | 8.18% |
| House Wife | 21,904 | 5.82% |

The **prevalence of hanging increased**, while family problems remained significant, and there was a decrease in housewife-related suicides.

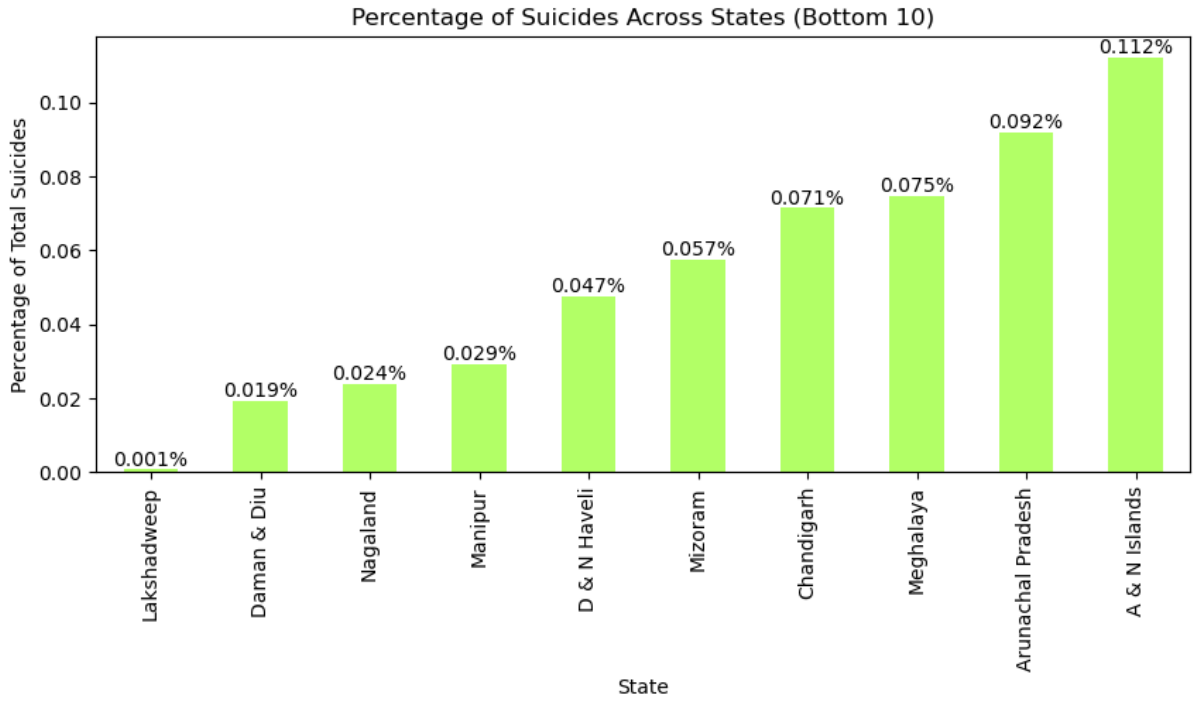
### *Insights*

Comparing occupational trends in 2001 and 2012, both years highlighted hanging as a common method across professions. Family problems consistently played a significant role, emphasizing the enduring impact of personal relationships. **A decline in housewife-related suicides in 2012.**

## State-wise suicide distribution




The dataset encompasses data from 35 states. The states with the ***highest suicide rates are Maharashtra (12%), West Bengal (11%), Tamil Nadu (11%), Andhra Pradesh (11%), and Karnataka (10%).***




The states with the lowest suicide rates, namely ***Lakshadweep, Daman & Diu,, Nagaland, Manipur, and D & N Haveli each report less than 0.1% of the total suicides.***

## *Insights*

The top 5 suicide-contributing states, including Maharashtra, West Bengal, Tamil Nadu, Andhra Pradesh, and Karnataka, collectively constitute a significant portion of national suicides. In contrast, D & N Haveli, Manipur, Nagaland, Daman & Diu, and Lakshadweep exhibit remarkably low rates, each representing less than 0.1% of the total suicides.

# Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences about population parameters based on a sample of data. The process involves formulating a hypothesis, collecting and analyzing data, and drawing conclusions. Here are the key steps in hypothesis testing:

## Test -1: Significant Disparity in Total Suicides between Genders

### Formulate Hypotheses:

- **Null Hypothesis (H0):** There is no significant difference in the suicide rates among male and Female.
- **Alternative Hypothesis (H1):** There is significant difference in the suicide rates among male and Females

### Set Significance Level (α):

- The significance level (often denoted as α) is the probability of rejecting the null hypothesis when it is actually true. Commonly used values are 0.05.

### Collect and Analyze Data:

**Shapiro-Wilk Test - Normality Assumption:**

| Group | P-Value |
|---|---|
| Male | 0.0 |
| Female | 0.0 |

This test assesses whether the data from both the male and female groups follow a normal distribution. The obtained p-values of 0.0 suggest that the **data significantly deviate from normality.**

**Levene's Test - Homogeneity of Variances:**

| Test | P-Value |
|---|---|
| Levene's | 3.954428262660545e-224 |

**Levene's test** evaluates whether the variances of the male and female data are approximately equal. The extremely small p-value (3.954428262660545e-224, close to zero) indicates a significant **violation of the homogeneity of variance assumption.**

These violations suggest caution in applying certain statistical tests, prompting consideration of alternative methods or **non-parametric tests for more reliable analyses.**

### Calculate Test Statistic:

In the **Mann-Whitney U Test**, the U Statistic is a rank-based test statistic, and the P-Value represents the probability of obtaining a U Statistic as extreme as the one calculated.

| Statistic | P-Value |
|---|---|
| U | 6939029794.0 |
| P-Value | 0.0 |

The extremely low P-Value of 0.0 indicates a significant difference in the total number of suicides between males and females. This leads to the rejection of the null hypothesis, suggesting that there is indeed a notable distinction in the total number of suicides between the two genders.

### Decision:

The extremely low p-value (0.0) suggests a significant difference in the total number of suicides between males and females. Therefore, ***we reject the null hypothesis, indicating a substantial distinction in the total number of suicides between the two genders.***

## Test -2:  Significant Disparity in Total Suicides among Different Age Groups

### Formulate Hypotheses:

- **Null Hypothesis (H0):** There is no significant difference in suicide rates among different age groups.
- **Alternative Hypothesis (H1):** There is a significant difference in suicide rates among different age groups.

### Set Significance Level (α):

- The chosen significance level (α) is commonly set at 0.05.

### Collect and Analyze Data:

**Shapiro-Wilk Test - Normality Assumption:**

| Age Group | P-Value |
|---|---|
| 0-14 | 0.0 |
| 15-29 | 0.0 |
| 30-44 | 0.0 |
| 45-59 | 0.0 |
| 60+ | 0.0 |

The obtained p-values of 0.0 suggest significant deviations from normality in all age groups.

**Levene's Test - Homogeneity of Variances:**

| Test | P-Value |
|---|---|
| Levene's | 0.0 |

**Levene's test** reveals a significantly low p-value, indicating a violation of the homogeneity of variance assumption among different age groups.

These deviations caution against using certain statistical tests, suggesting **consideration of non-parametric tests** for robust analyses.

### Calculate Test Statistic:

**Kruskal-Wallis Test - Suicide Rates Among Age Groups:**

| H-Statistic | P-Value |
|---|---|
| 18323.25 | 0.0 |

The extremely low P-Value of 0.0 indicates a significant difference in suicide rates among different age groups. This leads to the rejection of the null hypothesis, suggesting that there is indeed a notable distinction in suicide rates across various age groups.

### Decision:

The extremely low p-value (0.0) indicates a significant difference in suicide rates among different age groups. Therefore, we reject the null hypothesis, concluding that there is a ***substantial distinction in suicide rates across various age groups.***

## Conclusion

The hypothesis test and analysis reveal significant trends in suicide rates in India from 2001 to 2012. The apparent gender disparity, consistently reflecting higher suicide rates among males, emphasizes the urgent need for targeted interventions in men's mental health. Notably, young adults (15-29) and middle-aged (30-44) individuals shoulder the greatest impact of these trends, suggesting potential influences from educational pressures, career anxieties, and overlooked mental health concerns. Regional variations, highlighted by heightened suicide rates in states like Maharashtra and West Bengal, underscore the pivotal role of socio-economic factors in shaping mental health outcomes.