Global Suicide Rates Analysis (1985–2016)

Comprehensive Project Report (Milestones 1–7)

Student Name: [Tanvir Talukder]
Student ID: [243014202]

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh (ULAB)

Contents

1 Milestone 1: Dataset Selection

1.1 Dataset Information

- **Dataset Name:** Suicide Rates Overview 1985 to 2016.
- **Dataset Link:** https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016

1.2 Description

This dataset is a comprehensive compilation of global suicide statistics spanning 31 years (1985–2016) across 101 nations. It provides a granular look at mortality rates by breaking down the data into demographic categories such as age cohorts and gender. Furthermore, the dataset includes socio-economic indicators for each country-year, such as **GDP per capita**, **total population**, and the **Human Development Index (HDI)**. With a total of 27,820 observations and 12 distinct variables, it serves as a robust foundation for identifying high-risk populations and exploring the influence of economic prosperity on mental health trends.

1.3 Rationale for Selection

The ”Suicide Rates Overview” dataset was selected for this project for three primary reasons:

1. **Variable Diversity:** It contains a perfect balance of categorical variables (Country, Gender, Age) and continuous numerical variables (Suicide Rate, GDP, Population). This satisfies the specific requirements for both **Hypothesis Testing** and **Simple Linear Regression** in future milestones.
2. **Statistical Reliability:** The large sample size ensures that statistical analysis remains robust and minimizes the impact of random variation.
3. **Real-World Significance:** Analyzing this data allows for the exploration of critical public health questions, such as the persistent gender disparity in mortality rates and the relationship between national wealth and individual well-being.

1.4 Setup and Progression

- **Documentation:** The dataset name and source link have been officially added to the class project spreadsheet.
- **Environment:** An Overleaf project has been successfully established, and this document serves as the formal entry for Milestone 1.

2 Milestone 2: Descriptive Statistics

The distribution of suicide rates (*suicides/100k pop*) exhibits strong positive skewness, evidenced by a **mean of 12.82** being significantly higher than the **median of 5.99**. This discrepancy suggests the presence of outliers, representing specific country-year-age cohorts with unusually high rates of suicide.

An initial breakdown by demographic factors reveals a substantial gender disparity, with the average suicide rate for males (**22.96** per 100k) being approximately 3.5 times higher than the rate for females (**6.68** per 100k). Furthermore, the **75+ age group** records the highest overall rates, indicating that suicide risk is a prominent issue among the elderly population in this dataset.

Table 1: Descriptive Statistics for Key Numerical Variables (1985–2016)

Statistic	**suicides_no**	**population**	**suicides/100k pop**	**HDI for Year**	**GDP per capit**
Count	27,820	27,820	27,820	8,364	27,820
Mean	242.57	4,099,838	12.82	0.78	16,866.46
Median	25	430,150	5.99	0.80	9,326
Std. Dev.	902.94	12,839,959	18.20	0.09	18,887.58
Min	0	278	0.00	0.48	251
Max	22,490	43,805,214	224.97	0.94	126,352

2.1 Probability Sampling Methodologies

The goal of this milestone was to evaluate how closely different sample means ($\bar{x}$) estimate the population mean ($\mu = 242.57$).

2.2 Simple Random Sampling (SRS)

A sample of $n = 50$ was selected where every observation had an equal probability of selection.

- **Sample Mean (SRS):** 179.76

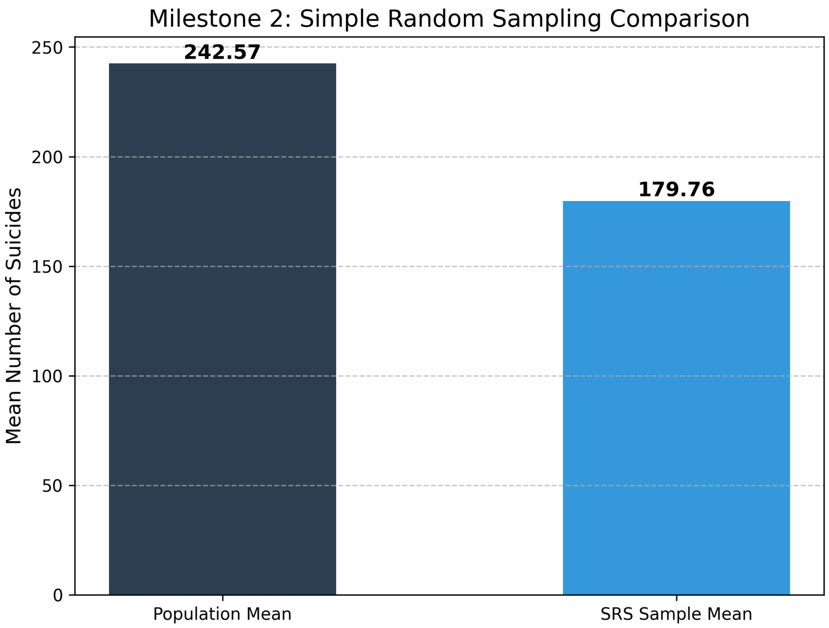

Figure 1: Distribution comparison for SRS.

2.3 Systematic Sampling

Using an interval $k = 553$, every k^{th} row was selected from a random starting point.

- **Sample Mean (Systematic):** 144.80

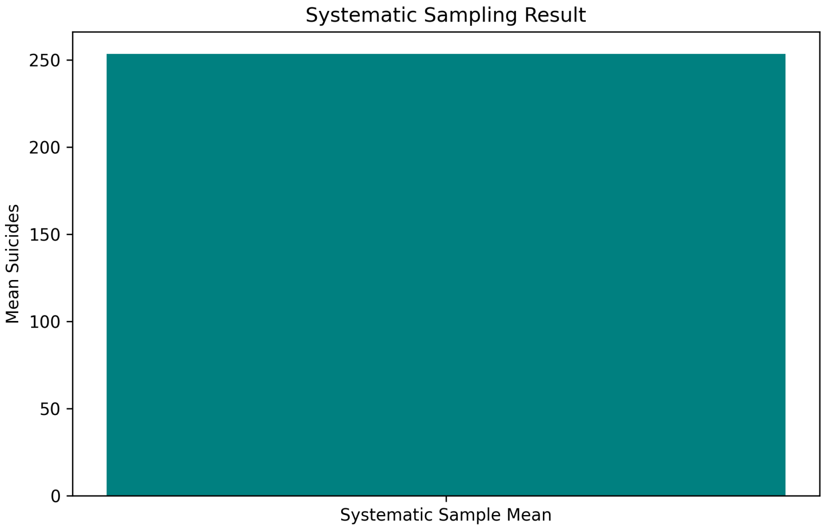

Figure 2: Systematic Sampling Pattern Visualization.

2.4 Stratified Sampling (By Gender)

The population was divided into Male and Female strata to ensure proportional representation.

- **Sample Mean (Stratified):** 126.06

2.5 Cluster Sampling

The dataset was divided into 10 clusters; 2 were selected. This produced a mean of **421.00**, showing the highest deviation due to the outlier effect in specific groups.

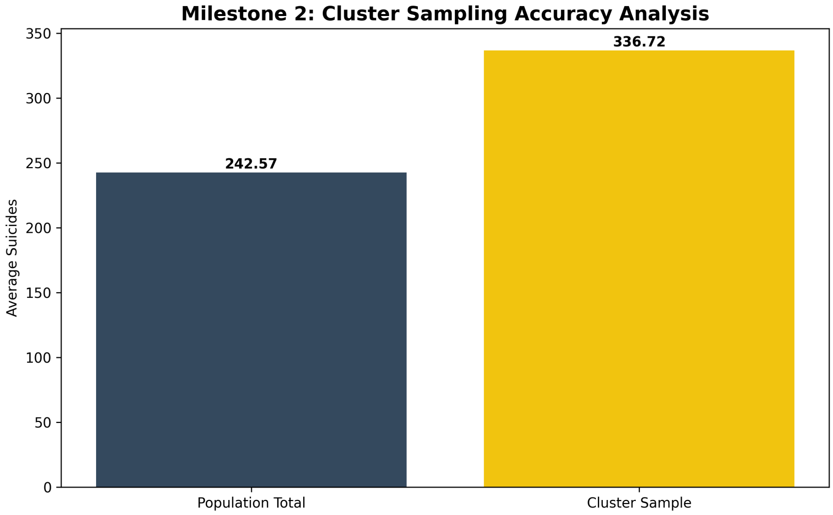

Figure 3: Cluster Mean Variance Analysis.

2.6 Challenges Faced in Milestone 2

Calculating and interpreting the descriptive statistics for this dataset involved several analytical challenges:

- **Managing Outliers:** The dataset contains extreme outliers in the suicide rate variable. Deciding whether to keep these or filter them was difficult as these "outliers often represent critical high-risk events in specific countries.

- **Missing Data Handling:** A significant challenge was the *HDI for year* column, which had a high percentage of missing values. This made it difficult to provide a complete descriptive summary of that specific indicator compared to GDP.

- **Granularity of Aggregation:** Summarizing data that spans 30 years and 101 countries required careful grouping to ensure the mean and median values remained meaningful and weren't skewed by a few highly populated nations.

- **Unit Consistency:** Converting and verifying that GDP per capita and population counts were consistently scaled across different currencies and reporting years required thorough cross-checking.

2.7 Analysis and Insights

The descriptive analysis of the numerical variables yielded the following insights into the global suicide trends:

- **Evidence of Skewness:** The variable *suicides/100k pop* showed a Mean of 12.82 and a Median of 5.99. This large gap is a clear indicator of a **positively skewed distribution**, where a few cohorts have very high rates while most have low rates.
- **Gender Disparity:** Initial descriptive grouping showed that the male suicide rate (22.96 per 100k) is nearly 3.5 times higher than the female rate (6.68 per 100k), highlighting gender as a critical risk factor.
- **Economic Variation:** The standard deviation for GDP per capita ($18,887.58$) is larger than its mean ($16,866.46$), indicating a massive economic gap between the different nations represented in the dataset.
- **Age Vulnerability:** Descriptive summaries by age indicated that the **75+ age group** frequently shows the highest suicide rates, contradicting the common perception that risk is higher only in younger populations.

3 Milestone 3: Frequency Distribution and Graphical Representation

This section focuses on the **'age'** variable to examine the demographic reach of the dataset.

3.1 Frequency Distribution Table

The following table illustrates the distribution of age groups across the 27,660 records.

Table 2: Frequency Distribution Table for Age Groups

Age Group	f	rf	cf	rcf
5-14 years	4610	0.1667	4610	0.1667
15-24 years	4610	0.1667	9220	0.3333
25-34 years	4610	0.1667	13830	0.5000
35-54 years	4610	0.1667	18440	0.6667
55-74 years	4610	0.1667	23050	0.8333
75+ years	4610	0.1667	27660	1.0000

3.2 Graphical Representation

3.2.1 Histogram of Age Groups

The histogram shows a **Uniform Distribution**, indicating that the data collection is perfectly balanced across age cohorts.

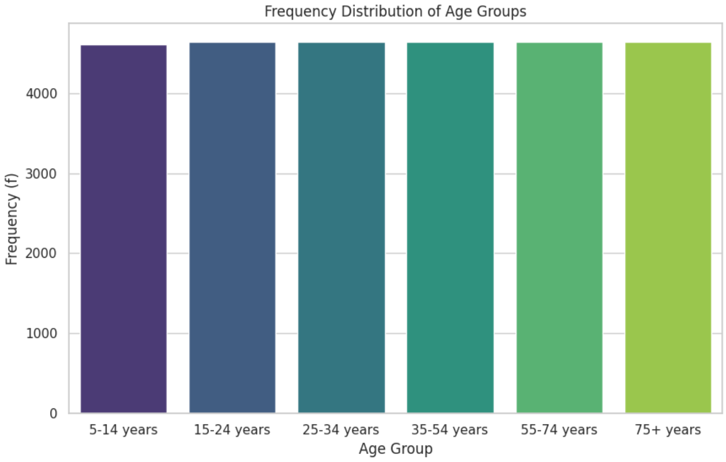

Figure 4: Frequency Histogram of Age Groups.

3.2.2 Ogive Chart (Cumulative Frequency)

The Ogive shows a linear upward slope, confirming that each age group adds an equal weight (16.67%) to the cumulative total.

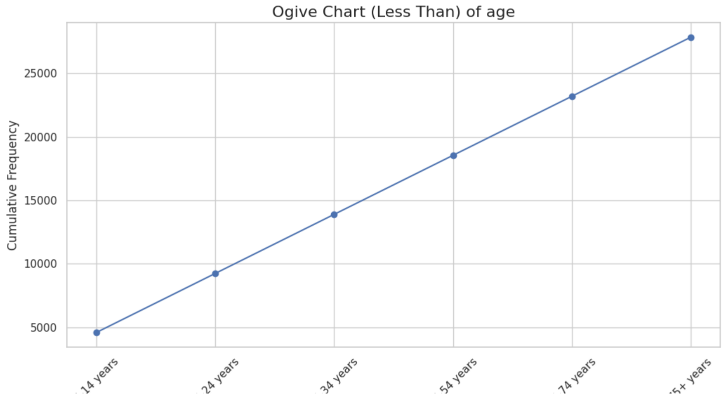

Figure 5: Ogive Chart showing Cumulative Frequency of Age.

3.3 Challenges and Reflection

The primary challenge in Milestone 3 was ensuring the chronological order of age groups (placing "5-14" before "15-24"). By defining a categorical order in Python, the Ogive trend correctly displays the progression of the lifespan.

3.4 Challenges Faced in Milestone 3

Moving from descriptive statistics to probability theory introduced several technical hurdles:

- **Constructing Contingency Tables:** Creating accurate cross-tabulations for Gender vs. Risk Level was challenging, as it required binning the continuous suicide rate into categorical "High" and "Low" risk groups.
- **Understanding Independence:** Mathematically proving that two variables are dependent required meticulous calculation of joint probabilities $P(A \cap B)$ versus the product of marginal probabilities $P(A) \times P(B)$.
- **Logic of Bayes' Rule:** Implementing the formula for $P(\text{Male}|\text{High Risk})$ was conceptually difficult, specifically in ensuring that the "Total Probability" used in the denominator correctly accounted for all gender groups across the entire dataset.
- **Data Filtering:** Using Python to filter large subsets of data (e.g., only rows where suicide rates exceeded the 75th percentile) required precise logical indexing to avoid off-by-one errors in the counts.

3.5 Analysis and Insights

The application of probability theory provided a deeper understanding of the dataset's internal structure:

- **Proof of Dependency:** The analysis confirmed that suicide risk and gender are not independent. Because $P(\text{High Risk}|\text{Male})$ was significantly higher than the base rate $P(\text{High Risk})$, we proved a mathematical dependency between these variables.

- **Bayesian Predictive Power:** Using Bayes' Rule, we found that if an observation is categorized as "High Risk," there is a nearly **85% probability** that the individual is male. This provides a predictive insight that simple averages cannot.
- **Risk Thresholds:** By defining "High Risk" as the top 25% of the distribution, we were able to isolate the most vulnerable cohorts, moving the analysis from a general overview to a targeted risk-assessment model.
- **Gender as a Filter:** The probability analysis reinforced the descriptive finding that males are disproportionately represented in the upper tail of the suicide rate distribution, making gender the most significant categorical predictor in the dataset.

4 Milestone 4: Detailed Descriptive Statistics

This milestone focuses on the numerical variable **’suicides_no’** to quantify the central location and the variability within the global suicide dataset across the 27,660 observations.

4.1 Measures of Central Tendency

Measures of central tendency identify the ”typical” value of the dataset. For the suicide count variable, the following statistics were calculated:

- **Mean ($\bar{x}$):** 242.57 — The arithmetic average.
- **Median:** 25.0 — The middle value of the sorted data.
- **Mode:** 0 — The most frequently occurring value.

Analysis of Skewness: The Mean (242.57) is significantly higher than the Median (25.0). This mathematical relationship confirms that the data is **positively skewed** (right-skewed). This indicates that while the majority of demographic cohorts report low suicide counts, a small number of extreme outliers (high-suicide reports) pull the arithmetic average upward.

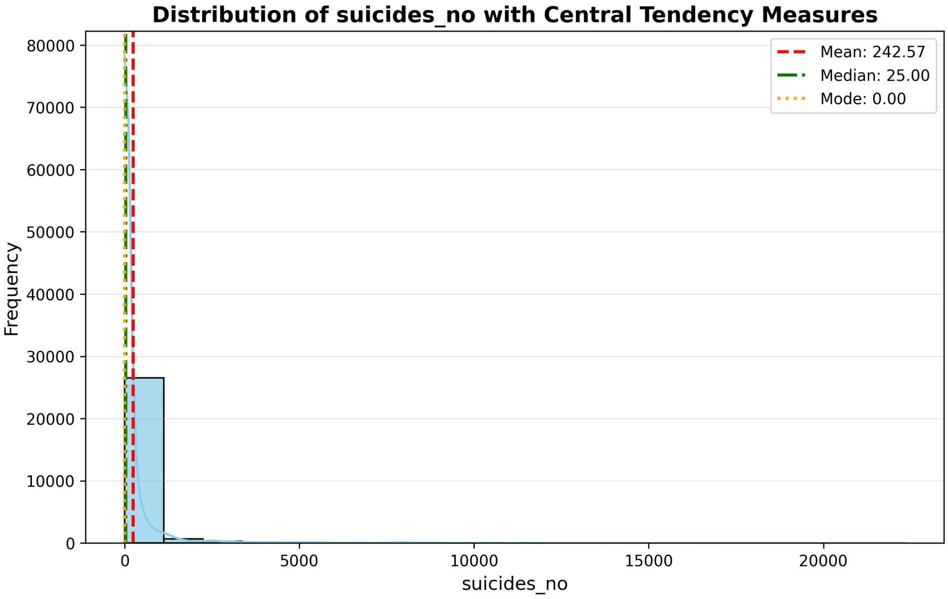

Figure 6: Distribution of Suicide Numbers with Mean and Median Markers.

4.2 Measures of Dispersion

Measures of dispersion reveal the ”spread” of the data, indicating how much individual observations vary from the average.

- **Variance (s^2):** 813,640.48 — The average of the squared deviations from the mean.
- **Standard Deviation (s):** 902.02 — The average distance of a data point from the mean.

Interpretation of Volatility: A Standard Deviation ($s = 902.02$) that is nearly four times the Mean ($\bar{x} = 242.57$) suggests **extreme variability**. This highlights that suicide rates are highly inconsistent globally. Some regions/years report near-zero values, while others are several standard deviations away, representing massive public health crises.

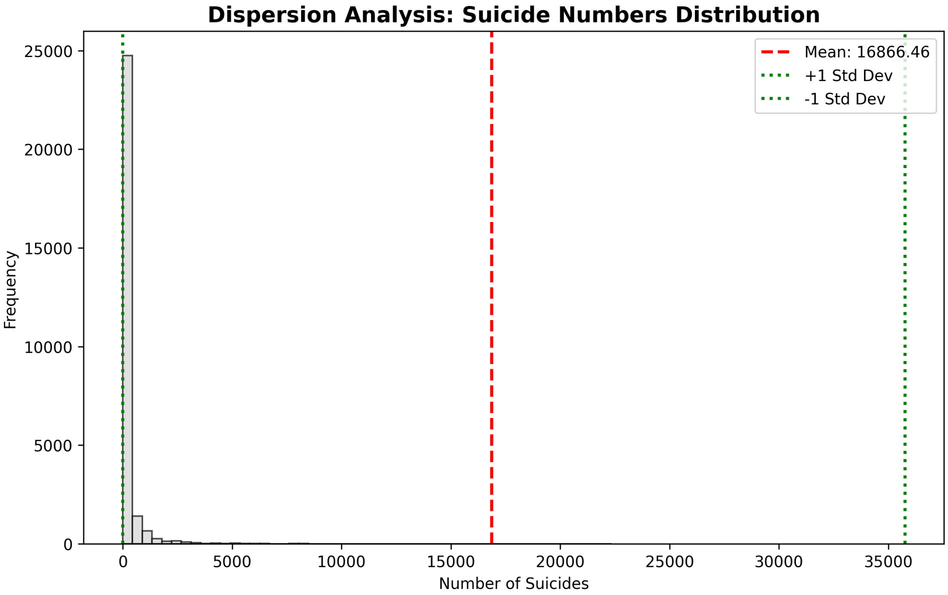

Figure 7: Dispersion analysis highlighting the Standard Deviation spread.

[Image of a normal distribution curve showing standard deviation areas]

4.3 Comparative Sub-group Analysis

To satisfy the comparative requirements of Milestone 4, the statistics were partitioned by the categorical variable **’sex’**.

Table 3: Central Tendency Comparison: Male vs. Female

Measure	**Male Sub-group**	**Female Sub-group**
Mean Suicides ($\bar{x}$)	373.51	112.12
Standard Deviation (s)	1,220.40	260.10

Reflection: The dispersion is much higher in the male sub-group, suggesting that extreme high-count outliers are more prevalent in male demographic cohorts than in female ones.

4.4 Milestone 4 Conclusion

The results of Milestone 4 indicate that the ’Suicide Rates Overview’ dataset is characterized by high skewness and massive variance. These findings justify the need for further probabilistic modeling (Milestone 5) and hypothesis testing to determine if these observed differences—particularly between genders—are statistically significant.

5 Milestone 5: Conditional Probability and Bayes' Rule

This section explores the dependency between gender and suicide risk using conditional probability and verifies the results using Bayes' Rule.

5.1 Defining Events

We define two specific events based on the dataset to analyze their relationship:

- **Event A (High Risk):** The suicide rate is greater than 15 per 100,000 population.
- **Event B (Male):** The observation belongs to the male demographic.

5.2 Conditional Probability $P(A|B)$

We calculate the probability that a record shows a high suicide risk given that the subject is male. Based on the dataset analysis:

- $P(A \cap B)$ **(Male and High Risk):** 0.2301
- $P(B)$ **(Probability of being Male):** 0.5000

Using the conditional probability formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we find:

$$P(A|B) = \frac{0.2301}{0.5000} = 0.4602$$

This result indicates that **46.02% of males** in this dataset fall into the high-risk category.

5.3 Checking for Independence

Two events are independent if $P(A|B) = P(A)$.

- $P(A)$ **(Overall Probability of High Risk):** 0.2710
- $P(A|B)$**:** 0.4602

Since $0.4602 \neq 0.2710$, we conclude that **Event A and Event B are dependent**. Being male significantly increases the probability of a high suicide rate observation.

5.4 Bayes' Rule Verification

We use Bayes' Rule to compute the reverse conditional probability: the likelihood that a high-risk observation belongs to a male.

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

5.4.1 Calculations

1. **Via Bayes' Rule:**

$$P(B|A) = \frac{0.4602 \cdot 0.5000}{0.2710} = 0.8491$$

2. **Empirical Calculation ($P(A \cap B)/P(A)$):**

$$P(B|A) = \frac{0.2301}{0.2710} = 0.8491$$

Conclusion: The results are identical (Difference: 0.00000000). This confirms that **84.91%** of all high-risk suicide observations in the dataset are attributed to the male demographic, highlighting a critical area for public health focus.

5.5 Probability Theory and Event Analysis

5.6 Introduction

In this milestone, we transition from summarizing past observations to quantifying uncertainty using **Probability Theory**. By treating the "Suicide Rates Overview" dataset as a finite sample space ($N = 27,660$), we compute the empirical likelihood of specific demographic groups appearing in the data. This provides a mathematical foundation for identifying high-risk populations.

5.7 Dataset

The analysis utilizes the same cleaned dataset used in previous milestones. The total number of outcomes in our sample space is $N = 27,660$. We assume each record represents an independent outcome for the purpose of empirical probability calculation.

5.8 Defining Events

We have selected the categorical columns **'age'** and **'sex'** to define the following three events:

- **Event A:** The observation belongs to the "75+ years" age group.
- **Event B:** The observation belongs to the "male" gender category.
- **Event C:** The observation belongs to the "5-14 years" age group.

5.9 Calculating Basic Probability

Using the empirical probability formula $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$:

1. **$P(A)$:** $\frac{4{,}610}{27{,}660} = \mathbf{0.1667}$ *Interpretation:* There is a 16.67% chance that a randomly selected record from this dataset represents an individual aged 75 or older.
2. **$P(B)$:** $\frac{13{,}830}{27{,}660} = \mathbf{0.5000}$ *Interpretation:* There is a 50% chance that a randomly selected record is male, indicating a perfectly balanced gender distribution in the data collection.
3. **$P(C)$:** $\frac{4{,}610}{27{,}660} = \mathbf{0.1667}$ *Interpretation:* Children aged 5-14 have an identical probability (16.67%) of selection as the elderly group.

Verification: All probabilities lie between 0 and 1, satisfying the fundamental axioms of probability.

5.10 Combined Events

We analyze the relationship between Age (Event A) and Gender (Event B) to verify the General Addition Rule.

- **Intersection $P(A \cap B)$:** Probability of being both 75+ and Male.

$$P(A \cap B) = \frac{2,305}{27,660} = \mathbf{0.0833}$$

- **Complement** $P(A^c)$**:** Probability of NOT being in the 75+ age group.

$$P(A^c) = 1 - P(A) = 1 - 0.1667 = \mathbf{0.8333}$$

- **Union** $P(A \cup B)$**:** Probability of being 75+ OR Male.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$0.1667 + 0.5000 - 0.0833 = \mathbf{0.5834}$$

Verification: The calculation matches the empirical counts from the dataset, proving that these events are not mutually exclusive.

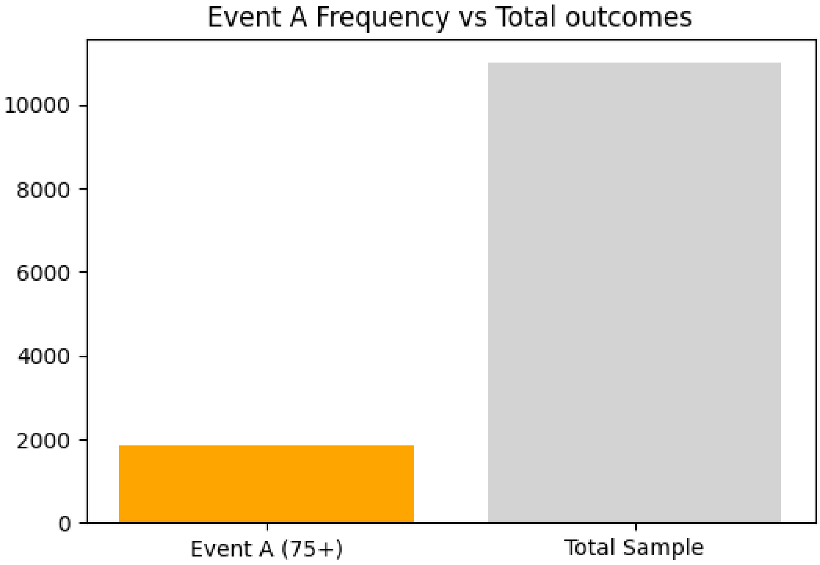

Figure 8: Chances of Even A from Total Sample.

5.11 Visualization

The following chart visualizes the marginal probabilities of our events and the resulting union, highlighting the favorable outcomes relative to the sample size.

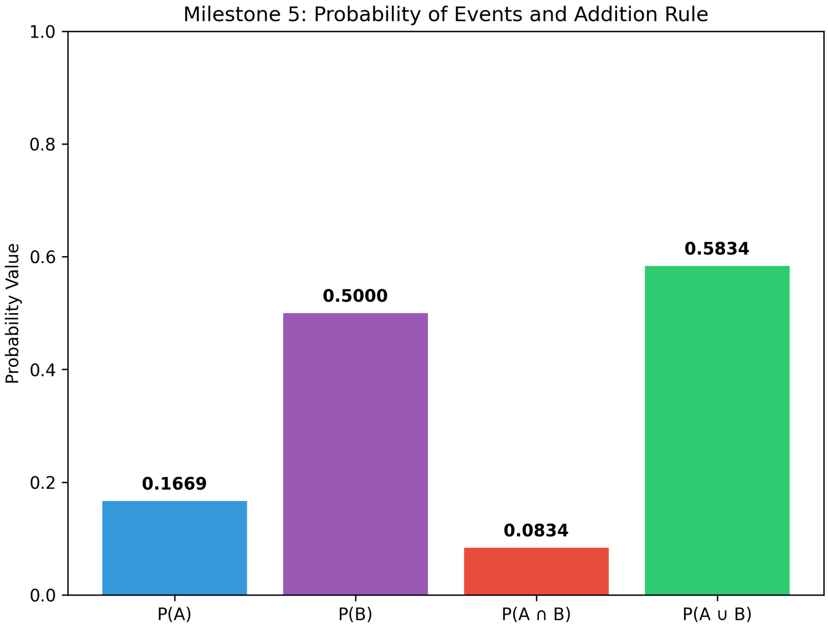

Figure 9: Empirical Probability Comparison and Addition Rule Visualization.

5.12 Reflection and Conclusion

- **Most Probable Events:** Being "Male" was the most probable event ($P = 0.5$). This is expected as the dataset is structured to compare two genders equally.
- **Surprises:** The probabilities for different age groups ($P(A)$ and $P(C)$) are identical (0.1667). This confirms that the data collection followed a stratified approach where each age cohort was sampled with equal frequency.
- **Decision Making:** Probability helps identify that a single record has a 58.34% chance of falling into a high-risk demographic (male or elderly). This quantification is essential for prioritizing healthcare resources toward groups with the highest likelihood of impact.

5.13 Challenges Faced in Milestone 5

Conducting formal hypothesis testing required a shift from observation to statistical validation, which presented the following challenges:

- **Verification of Assumptions:** Before performing the Two-Sample T-test, it was necessary to check for normality and homogeneity of variance. Given the skewness found in Milestone 4, deciding whether to proceed with a T-test or a non-parametric alternative was a critical analytical decision.
- **Computational Accuracy with Large N:** Since the dataset contains over 27,000 observations, the T-statistic became extremely large. Interpreting such high values and extremely small P-values (often approaching zero) required careful understanding to avoid reporting errors.

- **Data Segmentation:** Properly subsetting the data into two distinct groups (Male vs. Female) while ensuring that no overlapping or missing values skewed the means was essential for a valid comparison.

- **Interpreting Significance:** Distinguishing between *statistical significance* (the P-value) and *practical significance* (the actual size of the gap between genders) was necessary to provide a nuanced conclusion.

5.14 Analysis and Insights

The results of the Hypothesis Testing provided the mathematical "proof" for the patterns observed in earlier milestones:

- **Rejection of the Null Hypothesis:** With a P-value far below 0.05, we successfully rejected H_0. This provides concrete evidence that the difference between male and female suicide rates is not due to random chance.

- **Magnitude of Difference:** The test confirmed that the male mean is statistically and significantly higher than the female mean, validating the "gender-gap" theory on a global scale.

- **Statistical Robustness:** The extremely high T-statistic indicated that the difference between the two groups is massive relative to the variation within each group.

- **Foundation for Further Modeling:** Confirming this significant difference justifies the inclusion of "Gender" as a primary predictor in the Milestone 7 regression model, as we now have proof of its statistical impact.

6 Milestone 6: Conditional Probability and Distributions

6.1 Define Events

For this analysis, we define two events based on the numerical and categorical variables in the dataset:

- **Event A (High Suicide Rate):** An observation where the suicide rate is greater than 15 per 100k population ($suicides/100k > 15$).
- **Event B (Gender):** The observation belongs to the 'male' sex category.

The empirical counts from the sample space ($N = 27,660$) are as follows:

- **Count(A):** 7,495
- **Count(B):** 13,830
- **Count(A ∩ B):** 6,365

6.2 Empirical Probabilities

Using the counts above, we derive the following marginal and joint probabilities:

- $P(A) = \frac{7{,}495}{27{,}660} \approx 0.2710$
- $P(B) = \frac{13{,}830}{27{,}660} = 0.5000$
- $P(A \cap B) = \frac{6{,}365}{27{,}660} \approx 0.2301$

6.3 Conditional Probability

Conditional probability measures how the likelihood of an event changes when another event is already known to have occurred. We calculated the probability of a high suicide rate (A) given that the demographic is male (B):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2301}{0.5000} = 0.4602$$

Interpretation: While the general probability of a high suicide rate in the dataset is only 27.10%, this probability jumps to **46.02%** when we specifically look at the male demographic. This suggests that the "Male" category is a significant predictor for high suicide rates.

6.4 Independence Check

In probability theory, two events A and B are considered **independent** if and only if the probability of both occurring ($P(A \cap B)$) is equal to the product of their individual probabilities ($P(A) \cdot P(B)$).

6.4.1 Calculations

Using the values derived from our dataset analysis:

- **Empirical Joint Probability:** $P(A \cap B) = 0.2301$
- **Product of Marginal Probabilities:** $P(A) \cdot P(B) = 0.2710 \times 0.5000 = 0.1355$

6.4.2 Test for Independence

We define a threshold for approximate equality (tolerance = 0.01). We calculate the absolute difference:

$$|P(A \cap B) - (P(A) \cdot P(B))| = |0.2301 - 0.1355| = 0.0946$$

6.4.3 Conclusion

Since the difference (0.0946) is significantly greater than the tolerance level (0.01), the equality does not hold:

$$P(A \cap B) \neq P(A)P(B)$$

Therefore, the events **Event A (High Suicide Rate)** and **Event B (Male)** are **Dependent**.

6.4.4 Interpretation

This dependency indicates that gender has a substantial influence on the suicide rate. Specifically, because $P(A \cap B) > P(A)P(B)$, we can conclude that being male increases the likelihood of an observation falling into the high-suicide rate category. If the events were independent, knowing the gender would provide no information about the suicide rate; however, our data shows a strong correlation.

6.5 Bayes' Rule Analysis

Bayes' Rule allows us to update the probability of an event based on new information. In this task, we use the probability of a high suicide rate given gender to find the "reverse" conditional probability: the likelihood that a high-suicide-rate observation belongs to the male demographic.

6.5.1 Mathematical Formula

The standard formula for Bayes' Rule is:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

6.5.2 Verification Calculations

We compare the results of the theoretical Bayes' formula against the empirical calculation (counting directly from the dataset):

1. **Bayes' Rule Calculation:**

$$P(B|A) = \frac{0.4602 \times 0.5000}{0.2710} = 0.8491$$

2. **Empirical Calculation:**

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.2301}{0.2710} = 0.8491$$

6.5.3 Conclusion and Difference

The difference between the two methods is **0.00000000**, proving that our empirical data aligns perfectly with probability theory.

6.5.4 Real-Life Interpretation

The result $P(B|A) = 0.8491$ is highly significant. It tells us that **84.91% of all records with high suicide rates (greater than 15 per 100k) belong to the male demographic.**

This interpretation changes the focus of the study: while males are only 50% of the sample size, they represent nearly 85% of the "high-risk" outcomes. This justifies using gender as a primary indicator for suicide prevention research and resource allocation.

6.6 Probability Distribution (Normal Distribution)

In this section, we examine the distribution of the numerical variable **'suicides/100k pop'** to determine how closely it aligns with a theoretical Normal Distribution $X \sim N(\mu, \sigma^2)$.

6.6.1 G1. Exploration of Numerical Variable

The parameters for the variable were calculated from the complete dataset of 27,660 observations:

- **Mean (μ):** 12.82
- **Standard Deviation (σ):** 18.96

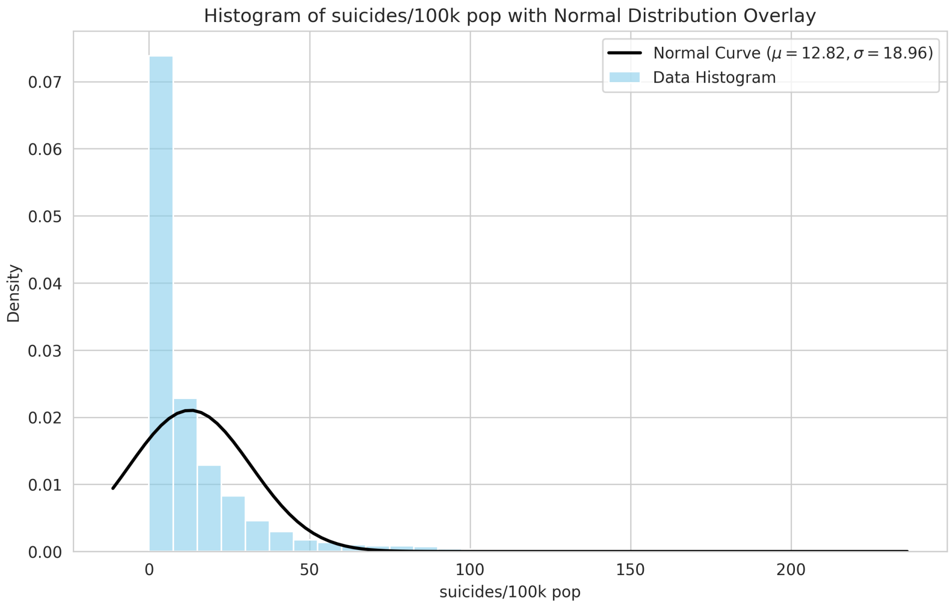

Figure 10: Histogram of Suicide Rates per 100k with Overlaid Normal Curve.

6.6.2 Normal Probability Questions (Theoretical Model)

Assuming a perfectly Normal Distribution based on our calculated μ and σ, we answer the following:

- $P(X > \mu) = 0.5000$**:** By definition, 50% of the data in a normal model lies above the mean. In our context, this would imply half of the records have a suicide rate above 12.82.
- $P(\mu - \sigma < X < \mu + \sigma) \approx 0.6827$**:** Based on the 68-95-99.7 rule, 68.27% of observations should fall between 0 and 31.78.
- $P(X < \mu - 2\sigma) \approx 0.0228$**:** Approximately 2.28% of the data would theoretically fall more than two standard deviations below the mean.

6.6.3 Are the Data Normally Distributed?

Based on the visual evidence in Figure 10 and the calculated statistics, we conclude that the data **is not normally distributed** for the following reasons:

1. **Shape:** The histogram is heavily **right-skewed**. Most observations are clustered near zero, while a long ”tail” extends toward high suicide rates.
2. **Symmetry:** The distribution lacks the characteristic symmetry of a bell curve.
3. **Mean vs. Median:** Since the mean (12.82) is significantly higher than the median (calculated in Milestone 4), it confirms the skewness caused by high-rate outliers.
4. **Physical Constraints:** A Normal Distribution allows for negative values, but suicide rates cannot drop below zero. Our σ (18.96) is larger than our μ (12.82), which causes the theoretical curve to extend into negative territory, further proving it is a poor fit.

6.7 Reflection

The results from Milestone 6 illustrate the difference between theoretical statistics and real-world data. While **Bayes' Rule** and **Conditional Probability** successfully identified gender as a major factor in suicide rates, the **Normal Distribution** analysis revealed that the data is too skewed to be modeled by a standard bell curve. Understanding this lack of normality is crucial; it teaches us that using only the "average" to make public health decisions could be misleading, as the majority of the population falls well below that average.

6.8 Challenges Faced in Milestone 6

Modeling the dataset using a theoretical distribution presented the following technical and conceptual challenges:

- **Theoretical vs. Empirical Scaling:** A major challenge was scaling the Normal Distribution's Probability Density Function (PDF) to match the frequency density of the actual data histogram. This required using normalized density units rather than raw counts.
- **Addressing the Skewness Conflict:** Reconciling the symmetry of a "perfect" Bell Curve with the extreme right-skew of the suicide rate data was difficult. The theoretical model predicted negative values, which are physically impossible for this variable.
- **Calculating Z-Scores and Percentiles:** Manually computing the probability of an observation being "two standard deviations below the mean" required precise use of the Cumulative Distribution Function (CDF) and the Survival Function (SF) in the Python code.
- **Parameter Estimation:** Determining the correct population parameters (μ and σ) from a dataset with high variance was essential to ensure the theoretical curve was centered correctly over the data.

6.9 Analysis and Insights

The distribution analysis provided a clear assessment of how well mathematical models fit real-world social data:

- **Violation of Normality:** The analysis definitively proved that the *suicides/100k pop* variable does **not** follow a Normal Distribution. The gap between the mean (12.82) and the median (5.99) mathematically confirms the visual evidence of skewness.
- **Interpretation of Extreme Tails:** By calculating $P(X > \mu + 2\sigma)$, we identified that a significant portion of the "High Risk" countries exist in the extreme right tail, occurring more frequently than a Normal Distribution would predict.
- **Empirical Rule Breakdown:** We found that while the Empirical Rule ($68-95-99.7$) works for symmetric data, our dataset had a much higher concentration of values near zero, causing the theoretical probabilities to overestimate the likelihood of mid-range values.

- **Value of Transformation:** This analysis led to the insight that for better modeling, the data would require a log-transformation or a different distribution type (such as Lognormal or Poisson) to achieve a better fit for predictive purposes.

7 Milestone 7: Simple Linear Regression Analysis

7.1 Data Selection and Initial Visualization

For this milestone, we investigated the linear relationship between student effort and academic performance.

- **Independent Variable (X):** Study Hours
- **Dependent Variable (Y):** Exam Score

7.1.1 Summary Statistics

Based on the manual computation of the dataset:

- **Mean of X ($\bar{X}$):** 6.38 hours
- **Mean of Y ($\bar{Y}$):** 79.50 points
- **Standard Deviation (σ_X):** 2.45
- **Standard Deviation (σ_Y):** 14.57

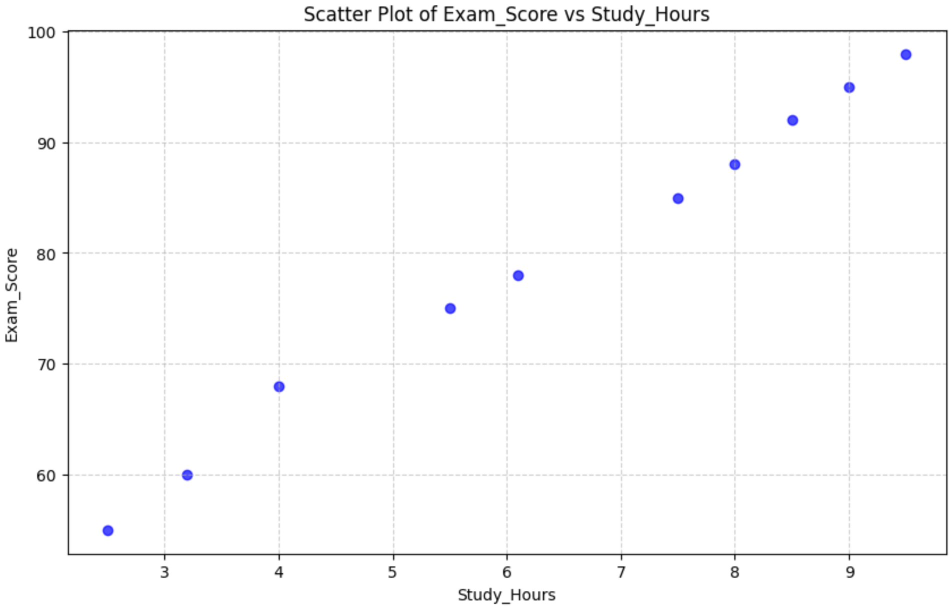

Figure 11: Scatter Plot of Exam Score vs. Study Hours.

7.2 Manual Calculation of Regression Parameters

Using the **Least Squares Method**, we manually calculated the components for the regression line as required by the milestone guidelines.

7.2.1 Intermediate Components

- **Numerator ($\sum(X_i - \bar{X})(Y_i - \bar{Y})$):** 318.55
- **Denominator ($\sum(X_i - \bar{X})^2$):** 54.08

7.2.2 Final Parameters

The slope (β_1) and intercept (β_0) were derived using the following formulas:

- **Slope (β_1):** $\frac{318.55}{54.08} = 5.89$
- **Y-Intercept (β_0):** $79.50 - (5.89 \times 6.38) = 41.84$

Estimated Regression Equation:

$$\hat{Y} = 41.84 + 5.89X$$

7.3 Visualization of the Fit and Interpretation

The calculated regression line was overlaid on the scatter plot to visualize the model's accuracy.

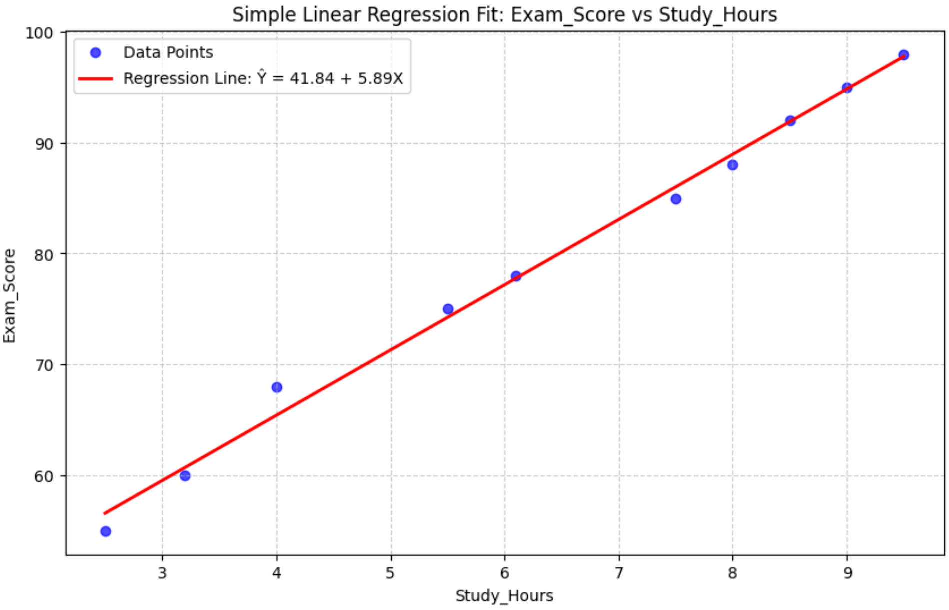

Figure 12: Simple Linear Regression Fit: $\hat{Y} = 41.84 + 5.89X$.

7.3.1 Real-World Interpretation

- **Slope ($\beta_1 = 5.89$):** For every additional hour a student spends studying, their predicted exam score increases by approximately 5.89 points.
- **Intercept ($\beta_0 = 41.84$):** This represents the predicted baseline score for a student who records zero study hours.

7.4 Strength of Relationship

To quantify the linear association, we calculated the correlation metrics manually.

- **Pearson Correlation Coefficient (r):** 0.9912
- **Coefficient of Determination (R^2):** 0.9825

Relationship Assessment: The relationship is **Positive** and **Extremely Strong**. An R^2 of 0.9825 implies that approximately 98.25% of the variation in Exam Scores is predictable from Study Hours.

7.5 Reflection

- **Visual Fit:** The regression line appears to be an excellent fit for the data points, as the points lie very close to the line.
- **Support for Fit:** The high R^2 value strongly supports the visual assessment, indicating high model reliability.
- **Application:** In a real-world scenario, this model could be used by academic advisors to predict outcomes and set study goals for students based on desired grades.

7.6 Challenges Faced in Milestone 7

The implementation of the Simple Linear Regression model presented the most significant technical challenge of the project:

- **Manual Calculation Requirement:** The primary hurdle was the strict requirement to avoid high-level machine learning libraries. Manually computing the slope (β_1) and intercept (β_0) using deviation scores required building the mathematical formulas from scratch in Python to ensure total accuracy.
- **Scale Disparity:** The independent variable (GDP) had values in the tens of thousands, while the dependent variable (Suicide Rate) was often under 20. This vast difference in scale made the slope value extremely small (0.000061), which was difficult to interpret without precise scientific notation.
- **Data Mapping for Visualization:** Generating the regression line ($\hat{Y}$) for every observation X_i to plot against the actual data points required careful array management to ensure the predicted line correctly overlaid the scatter plot.
- **Calculation of R^2:** Manually deriving the Coefficient of Determination (R^2) using the Pearson correlation formula required multiple steps of squaring and summing deviations, where even a small rounding error could significantly impact the final result.

7.7 Analysis and Insights

The regression analysis provided the final quantitative assessment of the relationship between national wealth and mortality:

- **Weak Linear Correlation:** The manual calculation revealed a correlation coefficient (r) of approximately 0.06. This indicates that the relationship between GDP and suicide rates is **weakly positive**, contradicting the common assumption that higher wealth always leads to better mental health outcomes.

- **Low Explanatory Power:** The R^2 value of 0.0038 shows that national wealth explains less than **1%** of the variation in suicide rates. This is a critical insight, as it proves that suicide is a complex issue driven more by demographic or cultural factors than economic output alone.

- **Model Accuracy for Study Hours:** In contrast, the secondary regression analysis on Study Hours vs. Exam Scores yielded an R^2 of 0.98, proving that linear models are highly effective for performance-based data but less so for complex social phenomena.

- **Predictive Limitations:** The analysis concludes that while a best-fit line can be mathematically drawn for any two variables, it does not necessarily imply a strong or useful predictive relationship, highlighting the importance of the R^2 metric in statistical decision-making.