

ai and ml

Mini Project



Rabia Soomro

133-22-0025

Instructor: Dr. Mumtaz Kaloi

**Student Performance Analysis Report**

**1. Introduction**

This report presents a detailed analysis of student exam scores based on a large dataset. The objective of this study is to evaluate the statistical properties of student performance, visualize the data distribution, and determine whether the scores follow a normal distribution. This analysis helps in understanding student performance trends, identifying common scores, and assessing the overall variability of scores.

**2. Dataset Overview**

* **Dataset Name:** Student Performance Large Dataset
* **Column Analyzed:** Exam\_Score (%)
* **Preprocessing Steps:**
  + Missing values were removed using .dropna() to ensure clean data.
  + No outlier removal was performed, as we aimed to analyze the natural variation in student scores.
  + The dataset consists of a large number of records, providing a reliable representation of exam performance trends.

**3. Code**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

df = pd.read\_csv(r"C:\Users\PMLS\Videos\archive\student\_performance\_large\_dataset.csv")

column\_name = "Exam\_Score (%)"

data = df[column\_name].dropna() # REMOVING THE MISSING PARTS

# CALCULATIONS

mean\_value = data.mean()

mode\_value = data.mode()[0]

std\_dev = data.std()

# RESULTS

print("Mean:", mean\_value)

print("Mode:", mode\_value)

print("Standard Deviation:", std\_dev)

# PLOTTING HISTOGRAM

plt.hist(data, bins=10, color='blue', edgecolor='black', alpha=0.7)

plt.xlabel(column\_name)

plt.ylabel("Frequency")

plt.title("Histogram of Exam Scores")

plt.show()

# CHECKING NORMAL DISTRIBUTION

x = np.linspace(min(data), max(data), 100)

y = (1 / (std\_dev \* np.sqrt(2 \* np.pi))) \* np.exp(-0.5 \* ((x - mean\_value) / std\_dev) \*\* 2)

plt.hist(data, bins=20, density=True, color='lightblue', edgecolor='black')

plt.plot(x, y, color='red', label="Normal Curve")

plt.xlabel(column\_name)

plt.ylabel("Density")

plt.title("Exam Scores Distribution")

plt.legend()

plt.show()

# Interpretation

print("\nAnalysis and Interpretation")

print(f"The mean exam score is {mean\_value:.2f}%, which represents the average student performance.")

print(f"The mode exam score is {mode\_value}%, showing the most frequently achieved score.")

print(f"The standard deviation is {std\_dev:.2f}, indicating how much variation exists in scores.")

print("the histogram follows a bell-shaped curve which means the scores follow a normal distribution.")

**4. Output**

**A black text with numbers

AI-generated content may be incorrect.**

A graph of a test results

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

**5. Statistical Analysis**

To understand the key characteristics of the dataset, we computed the following statistical measures:

|  |  |
| --- | --- |
| Metric | Value |
| Mean (Average Score) | 70.18% |
| Mode (Most Frequent Score) | 55% |
| Standard Deviation (Spread of Scores) | 17.68% |

**5.1 Interpretation of Statistical Results**

* **Mean:** The mean represents the average exam score, providing an overall measure of student performance.
* **Mode:** The mode is the most frequently occurring exam score in the dataset, indicating common performance levels.
* **Standard Deviation:** This value measures the dispersion of scores from the mean. A higher standard deviation indicates greater variability among student performance levels.

**6. Data Visualization and Normal Distribution Check**

**6.1 Histogram of Exam Scores**

* The histogram represents the frequency distribution of student scores.
* Data is divided into **10 bins** to provide a clear visual representation.
* The distribution appears **somewhat uniform**, indicating a diverse range of scores rather than a strict concentration around a central value.

**6.2 Normal Distribution Analysis**

* A probability density histogram is plotted to examine the spread of exam scores.
* A **normal distribution curve** is overlaid to assess whether the data follows a normal pattern.
* The probability density function (PDF) of a normal distribution is given by the formula:

A black and white math equation

AI-generated content may be incorrect.

where:

* + μ is the mean of the dataset
  + σ sigma is the standard deviation
  + x represents individual data points
* The histogram does not form a perfect bell-shaped curve, indicating a possible deviation from normality.

**7. Findings and Interpretation**

* The **mean and mode** are relatively close, suggesting a **symmetrical distribution**, but not perfectly normal.
* The histogram shows a **slightly uniform spread** rather than a distinct peak, indicating variability in student performance levels.
* If the dataset followed a perfect normal distribution, we would expect a **clear bell curve**, with most scores concentrated around the mean and fewer extreme values.
* The deviation from normality suggests that **external factors such as grading policies, exam difficulty, or student preparation levels** may influence score distribution.

**8. Conclusion**

The statistical and graphical analysis indicates that student exam scores are not normally distributed. The histogram suggests a fairly uniform spread rather than a distinct bell-shaped curve. The overlaid normal curve does not fit the data well, further confirming deviation from normality. Additional tests, such as skewness and kurtosis analysis, would be needed to quantify this deviation. This suggests that student performance varies significantly rather than clustering around a single average score.

**9. Recommendations**

* **Additional statistical measures** (such as median, skewness, and kurtosis) should be calculated to further confirm the distribution pattern.
* **Further segmentation** of the data (e.g., analyzing scores based on student demographics or subject-wise performance) may provide deeper insights.
* **More bins in histograms** can be used to improve visualization and detect finer details in the distribution pattern.
* **Machine learning models** could be applied to predict student performance based on past trends and external influencing factors.