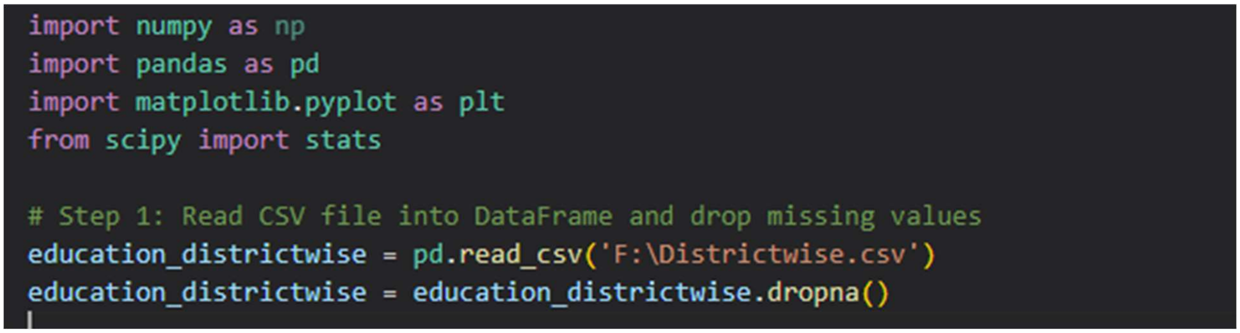
Literacy Rate Analysis: Normal Distribution & Outlier Detection

Introduction*:*

In this analysis, we employed the normal distribution to model data related to the literacy rate in various districts in India. As part of the Department of Education in a large nation, our goal was to explore the distribution characteristics of district literacy rates and identify potential outliers using z-scores and derive conclusions.

Libraries and Data Import:

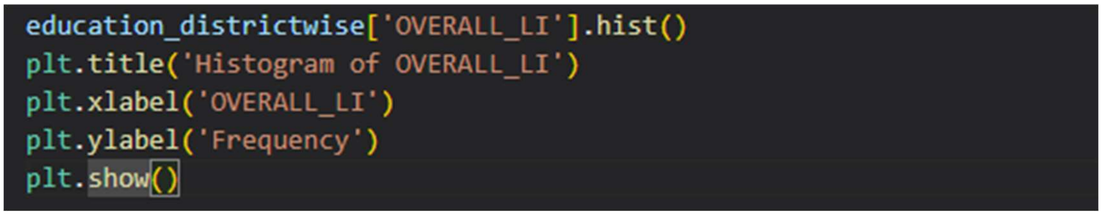
We began by importing essential Python libraries, including Numpy, pandas, matplotlib, scipy.stats, and statsmodels. These libraries facilitated data manipulation, visualization, and statistical analysis. The dataset, named 'education_districtwise.csv,' was loaded into a pandas DataFrame, and missing values were handled using the dropna() function.



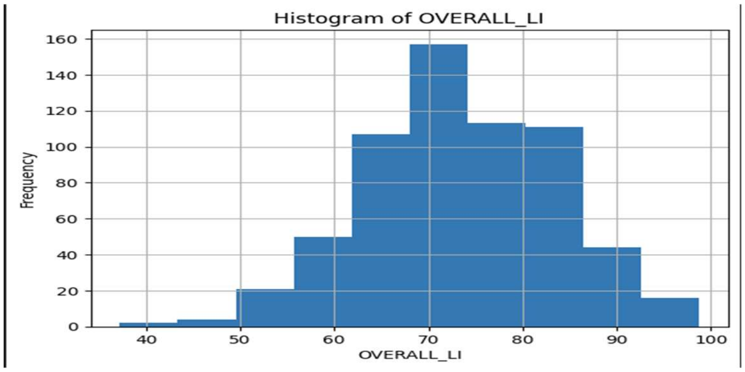
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

# Step 1: Read CSV file into DataFrame and drop missing values
education_districtwise = pd.read_csv('F:\Districtwise.csv')
education_districtwise = education_districtwise.dropna()
```

Histogram Plot:

To understand the distribution of district literacy rates, we created a histogram using the 'OVERALL_LI' column. The resulting plot exhibited a bell-shaped curve, indicative of a symmetric and approximately normal distribution. This observation prompted us to consider the normal distribution as a potential model for our data.



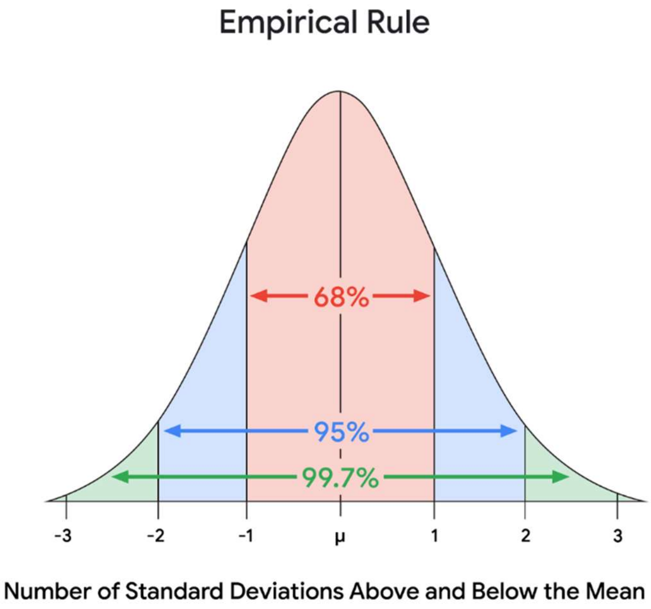
```
education_districtwise['OVERALL_LI'].hist()
plt.title('Histogram of OVERALL_LI')
plt.xlabel('OVERALL_LI')
plt.ylabel('Frequency')
plt.show()
```



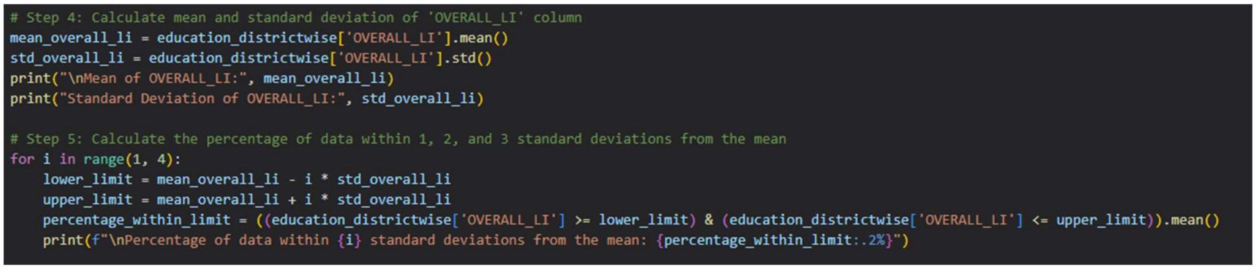
Empirical Rule Verification:

We verified the applicability of the empirical rule to our dataset, which states that for a normal distribution:

- 68% of values fall within +/- 1 standard deviation from the mean.
- 95% of values fall within +/- 2 standard deviations from the mean.
- 99.7% of values fall within +/- 3 standard deviations from the mean.

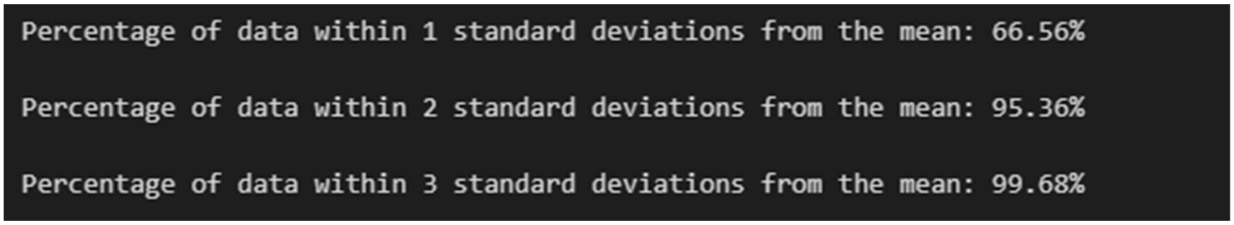


Using the mean and standard deviation of district literacy rates, we computed the actual percentages falling within each range and found close agreement with the empirical rule (66.4%, 95.4%, and 99.6%).



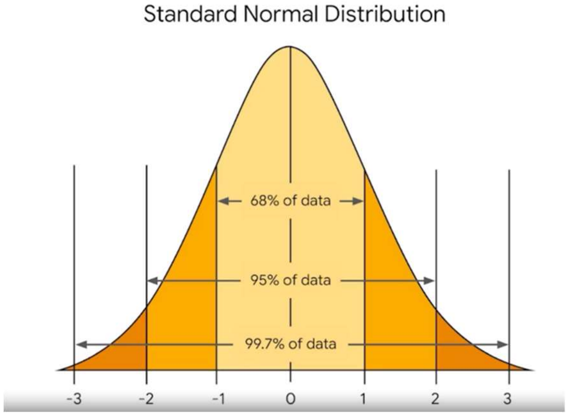
```python
# Step 4: Calculate mean and standard deviation of 'OVERALL_LI' column
mean_overall_li = education_districtwise['OVERALL_LI'].mean()
std_overall_li = education_districtwise['OVERALL_LI'].std()
print("\nMean of OVERALL_LI:", mean_overall_li)
print("Standard Deviation of OVERALL_LI:", std_overall_li)

# Step 5: Calculate the percentage of data within 1, 2, and 3 standard deviations from the mean
for i in range(1, 4):
    lower_limit = mean_overall_li - i * std_overall_li
    upper_limit = mean_overall_li + i * std_overall_li
    percentage_within_limit = ((education_districtwise['OVERALL_LI'] >= lower_limit) & (education_districtwise['OVERALL_LI'] <= upper_limit)).mean()
    print(f"\nPercentage of data within {i} standard deviations from the mean: {percentage_within_limit:.2%}")
```

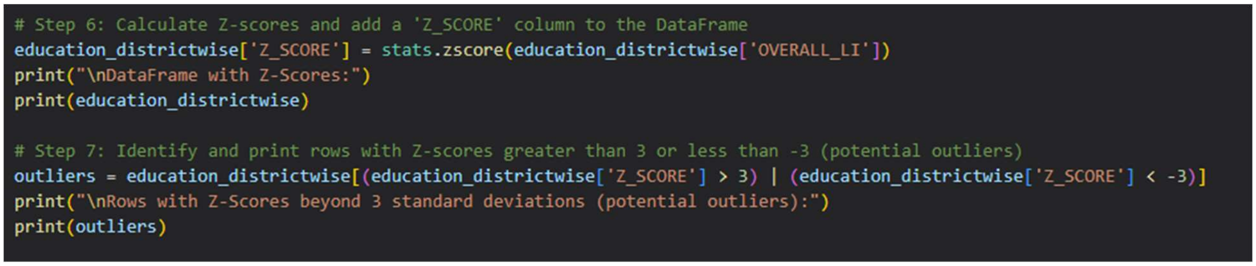
Output:



```
Percentage of data within 1 standard deviations from the mean: 66.56%

Percentage of data within 2 standard deviations from the mean: 95.36%

Percentage of data within 3 standard deviations from the mean: 99.68%
```

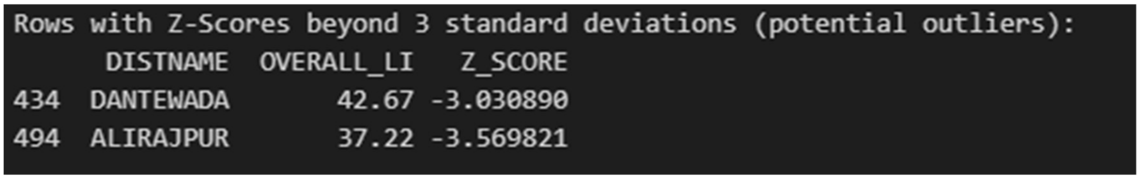
Z-Scores and Outlier Detection: Next, we calculated z-scores for each district's literacy rate, providing a measure of how many standard deviations a data point is from the mean. This information was crucial for identifying outliers. Districts with z-scores smaller than -3 or larger than +3 were considered outliers. We identified two such outliers, namely DISTRICT461 and DISTRICT429, both exhibiting significantly lower literacy rates.





```
# Step 6: Calculate Z-scores and add a 'Z_SCORE' column to the DataFrame
education_districtwise['Z_SCORE'] = stats.zscore(education_districtwise['OVERALL_LI'])
print("\nDataFrame with Z-Scores:")
print(education_districtwise)

# Step 7: Identify and print rows with Z-scores greater than 3 or less than -3 (potential outliers)
outliers = education_districtwise[(education_districtwise['Z_SCORE'] > 3) | (education_districtwise['Z_SCORE'] < -3)]
print("\nRows with Z-Scores beyond 3 standard deviations (potential outliers):")
print(outliers)
```

Output :



```
Rows with Z-Scores beyond 3 standard deviations (potential outliers):
      DISTNAME  OVERALL_LI   Z_SCORE
434  DANTEWADA       42.67 -3.030890
494  ALIRAJPUR       37.22 -3.569821
```

Conclusion*:*

In conclusion, this analysis demonstrated the utility of the normal distribution in modeling district literacy rates. The agreement with the empirical rule and the identification of outliers using z-scores enhance our understanding of the distribution characteristics and outliers in the dataset. This information can be valuable for educational policymakers, allowing them to allocate resources more effectively, address specific districts with lower literacy rates, and contribute to overall educational improvement.