Introduction Single Variables

To evaluate the effectiveness of existing policies. To make informed decisions statistics play an important role in the social world.

example

which school is best for child, which hospital is best for eye operation.-->allow comparisons to be made

every published statistic is the result of a number of different decisions

To display the essential features of one variable at a time:

Univariate analysis is the most fundamental type of statistical data analysis technique.

purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

univariate data include central tendency (mean, mode and median) and dispersion: range , variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation.for example, the height of ten students in a class

The salaries of people in the industry could be a univariate analysis example.

univariate data analysis

A single variable analysis involves examining and summarizing the characteristics of one variable at a time. This type of analysis is often referred to as univariate analysis.

How describing data with univariate data.

- Frequency Distribution Tables
- Bar Charts
- Histograms
- Frequency Polygons
- Pie Charts

Distributions and Variables

population of our country

how many people are graduated

how many people are (Above 18 years)

how many are covid affected

how many are recovered from covid

To answer these kinds of ques_x0002_tions we need to collect information from a large number of people, Conducting large-scale surveys is a time-consuming and

costlydata from survey research

in the social sciences are available free of charge to researchers

Descriptive Statistics

Descriptive Statistics:

Mean: The average of the values in the dataset.

Median: The middle value of the dataset when it is sorted.

Mode: The most frequently occurring value in the dataset.

Range: The difference between the maximum and minimum values.

Standard Deviation and Variance: Measures of the spread or dispersion of the data.

Frequency Distribution:

Creating a table that shows the frequency of each value or range of values in the dataset. This is often presented in the form of a histogram.

Box Plots:

A graphical representation that displays the distribution and central tendency of a dataset. It includes the minimum, first quartile, median, third quartile, and maximum.

Visualization:

Using graphs and charts like histograms, bar charts, pie charts, or line charts to visually represent the distribution and characteristics of the variable.

Outlier Detection:

Identifying and examining values that deviate significantly from the overall pattern of the data.

example : survey of a classroom.

To count the number of boys and girls in the room.

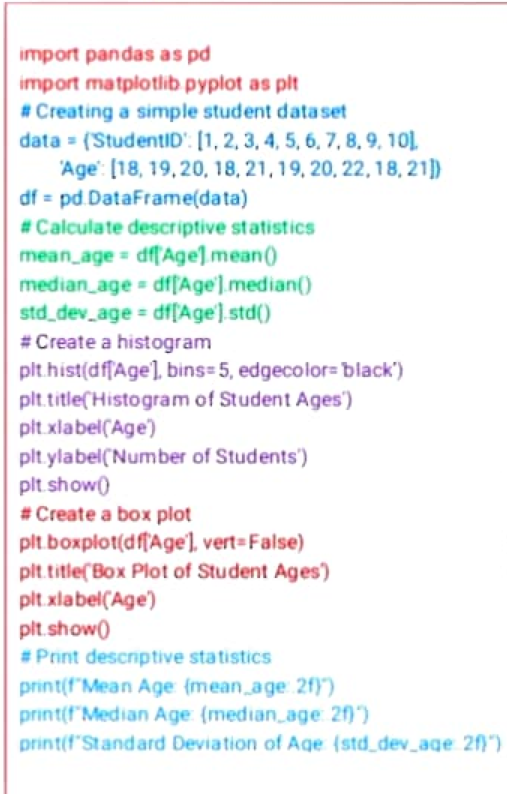
The data here simply talks about the number which is a single variable and the variable quantity.

example : survey of a worldcup2023 match .

"Virat Kohli's score is an example of single variable analysis in data analysis to predict his score in the next match."

virat koli scores in worldcup 2023- distribution -- number-Scores

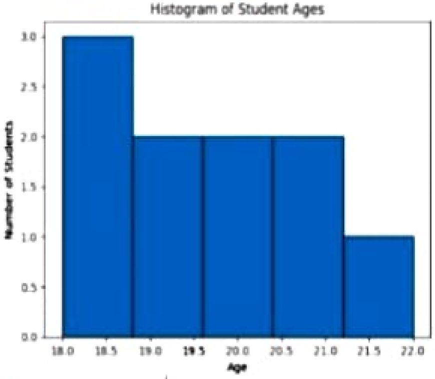
single variable

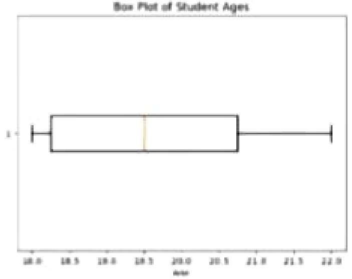


```
import pandas as pd
import matplotlib.pyplot as plt
# Creating a simple student dataset
data = {'StudentID': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
      'Age': [18, 19, 20, 18, 21, 19, 20, 22, 18, 21]}
df = pd.DataFrame(data)
# Calculate descriptive statistics
mean_age = df['Age'].mean()
median_age = df['Age'].median()
std_dev_age = df['Age'].std()
# Create a histogram
plt.hist(df['Age'], bins=5, edgecolor='black')
plt.title('Histogram of Student Ages')
plt.xlabel('Age')
plt.ylabel('Number of Students')
plt.show()
# Create a box plot
plt.boxplot(df['Age'], vert=False)
plt.title('Box Plot of Student Ages')
plt.xlabel('Age')
plt.show()
# Print descriptive statistics
print(f"Mean Age: {mean_age:.2f}")
print(f"Median Age: {median_age:.2f}")
print(f"Standard Deviation of Age: {std_dev_age:.2f}")
```

histogram-- to show the distribution of student ages
box plot--To visualize the central tendency
print the calculated descriptive statistics.

numerical representations (mean, median, standard deviation)to understand the distribution of student ages
visual representations--(histogram, box plot) .





Mean Age: 19.60
Median Age: 19.50
Standard Deviation of Age: 1.43

Task

Name	Age	Gender	City	Prediction(YES/NO)
Hardly	34	M	Simla	
John	18	M	Trichy	
Shanmugam	22	M	Palani	
Subramani	16	M	Ooty	
Farvin Begam	28	F	Chennai	
Assar	75	M	Chennai	
July	29	F	Kashmir	
Subasri	62	F	Simla	
Deepa Lakshmi	21	F	OOty	

"Predict their willingness to have ice cream."

Numerical Summaries of Level and Spread

Numerical summaries of data that describe its central tendency and variability are often referred to as measures of central tendency and measures of dispersion. Here are some common numerical summaries for these two aspects.

numerical summaries provide different perspectives on the characteristics of a dataset, including its central tendency (level) and spread

Measures of Central Tendency (Level):

Mean,Median,Mode,Percentile,Quartiles (Five-Number Summary),Proportion

Measures of Dispersion (Spread)

Range,Variance,Standard Deviation,Interquartile Range (IQR),
Coefficient of Variation,Correlation

Measures of Central Tendency (Level)

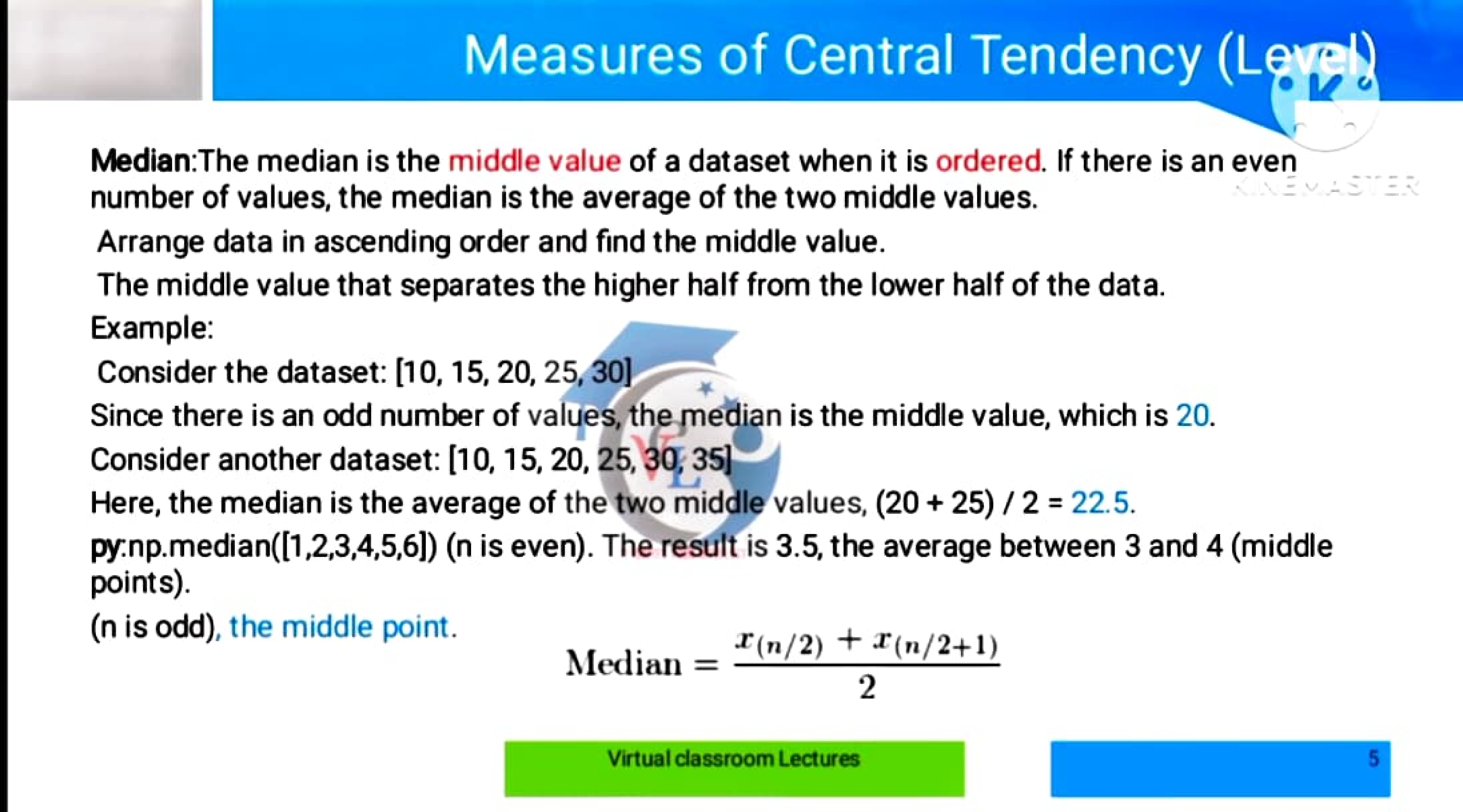
Mean (Arithmetic Average):

The average value of the data.The mean, also known as the arithmetic average, is the sum of all values in a dataset divided by the number of values.

Consider the dataset: [10, 15, 20, 25, 30] Mean=(10+15+20+25+30)/5=100/5=20

py:np.mean([1,2,3,4,5]) The result is 3

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



Measures of Central Tendency (Level)

Median:The median is the middle value of a dataset when it is ordered. If there is an even number of values, the median is the average of the two middle values.

Arrange data in ascending order and find the middle value.

The middle value that separates the higher half from the lower half of the data.

Example:

Consider the dataset: [10, 15, 20, 25, 30]

Since there is an odd number of values, the median is the middle value, which is 20.

Consider another dataset: [10, 15, 20, 25, 30, 35]

Here, the median is the average of the two middle values, (20 + 25) / 2 = 22.5.

py:np.median([1,2,3,4,5,6]) (n is even). The result is 3.5, the average between 3 and 4 (middle points).

(n is odd), the middle point.

$$\text{Median} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Measures of Central Tendency (Level)

Mode:

The most frequently occurring value in the dataset.

Definition: The mode is the value that appears most frequently in a dataset. A dataset may have no mode (if all values are unique), one mode (unimodal), or more than one mode (multimodal).

Example:

Consider the dataset: [5, 10, 10, 15, 20, 20, 25, 30]

Here, the mode is 10 and 20 since these values appear most frequently.

Consider another dataset: [5, 10, 10, 15, 20, 20, 25, 25, 30]

In this case, the dataset is bimodal with modes at 10 and 20.

py:statistics.mode([1,2,2,2,3,3,4,5,6]) The result is 2.

mean-->average value, median --> middle value, mode--> most frequently occurring value

Numerical Summaries of Level

Percentile:

Percentiles provide a measure of level by dividing the data into specific percentage intervals. For example, the median corresponds to the 50th percentile, indicating the middle point of the data.

Example:

Consider the following dataset: [2, 4, 7, 10, 15, 20, 25, 30].

The median (50th percentile) is the middle value, which is 15.

The 25th percentile would be the value below which 25% of the data falls, which in this case is 7.

Numerical Summaries of Level

Quartiles (Five-Number Summary):

Quartiles, specifically the median (Q2), also provide a measure of level by dividing the data into quarters. Q1 and Q3 represent the 25th and 75th percentiles, respectively.

Five-number summary is composed of:

1.Minimum
2.25th percentile (lower quartile)
3.50th percentile (median)
4.75th percentile (upper quartile)
5.100th percentile (maximum)

Example:

dataset [2, 4, 7, 10, 15, 20, 25, 30],

the five-number summary is: Minimum = 2,

Q1 = 6.75 (average of 4 and 7),

Q2=Median = 15

, Q3 = 23.75 (average of 20 and 25),

Maximum = 30.

Minimum, Lower Quartile (Q1) ,Median (Q2) ,Upper Quartile (Q3), Maximum.

Numerical Summaries of Level

Proportion:

Proportion represents the distribution of data within specific percentage intervals, contributing to the understanding of the data's level within those intervals.

Example:

For the dataset [2, 4, 7, 10, 15, 20, 25, 30], we can say that 50% of the data falls below the median, 25% below the first quartile, and 75% below the third quartile.

$$\hat{p} = \frac{x}{n}$$

Measures of Dispersion (Spread)

Range:

The difference between the maximum and minimum values in the dataset.

$$x_{(n)} - x_{(1)}$$

Example:

For the dataset [2, 4, 7, 10, 15, 20, 25, 30], the range is 30 - 2 = 28.

Python: np.max(n) – np.min(x)

Variance:

A measure of how far each data point in the set is from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Example:

Using the same dataset as above, the variance is 68.25 (calculated by averaging the squared differences from the mean).

Python: np.var(x)

Measures of Dispersion (Spread)

Standard Deviation:

A more interpretable measure of the spread, as it is in the same units as the data.

Example:

Consider the dataset [2, 4, 7, 10, 15, 20, 25, 30]. The mean is 15. The squared differences from the mean are [169, 121, 64, 25, 0, 25, 64, 121]. The variance is the average of these squared differences, which is 68.25. The standard deviation is the square root of the variance, approximately 8.26.

Python: np.std(x)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Measures of Dispersion (Spread)

Interquartile Range (IQR):

Definition: The Interquartile Range (IQR) is a measure of statistical dispersion, it represents the range in which the middle 50% of the data values . It is the difference between the third quartile (Q3) and the first quartile (Q1).

Example:

Consider the dataset: [12, 15, 18, 20, 22, 24, 28, 32, 40]

1.Order the dataset:

12,15,18,20,22,24,28,32,40

2.Calculate Quartiles:

Q1=18 (median of the first half),Q3=28 (median of the second half)

3.Calculate IQR:

IQR=Q3−Q1=28−18=10

The IQR in this example is 10, indicating that the middle 50% of the data falls within the range of 18 to 28.

Measures of Dispersion (Spread)

Coefficient of Variation (CV):

Definition: The Coefficient of Variation (CV) is a relative measure of variability that expresses the standard deviation as a percentage of the mean. It is used to compare the variability of datasets with different units or scales.

Formula:

$CV =(s/x)\times100\%$s is the standard deviation and x is the mean

Example:

Consider two datasets:

Dataset A: [10, 12, 15, 18, 20]

Dataset B: [50, 55, 60, 65, 70]

Measures of Dispersion (Spread)

Calculate Mean and Standard Deviation for Each Dataset:

Dataset A: Mean (x) = (10+12+15+18+20)/5=75 /5=15 Standard Deviation (s) ≈ 3.16

Dataset B: Mean (x) = (50+55+60+65+70/)5=60 Standard Deviation (s) ≈ 7.07

Calculate Coefficient of Variation:

Dataset A: CV=(3.16/15)×100% ≈21.07%

Dataset B: CV=(7.07/60)×100% ≈11.78%

The coefficient of variation allows us to compare the relative variability between the two datasets. In this example, Dataset A has a higher CV (21.07%) compared to Dataset B (11.78%), indicating that Dataset A has a higher relative variability compared to its mean.

Measures of Dispersion (Spread)

Correlation:

Correlation measures the strength and direction of a linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Example:

Consider two variables, X and Y. If as X increases, Y consistently increases, there is a positive correlation. If as X increases, Y consistently decreases, there is a negative correlation. If there's no systematic relationship, the correlation is close to 0.

$$r = \frac{\sum \left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1}$$

summary

percentiles and quartiles provide information about the distribution of data within specific percentage intervals.

standard deviation, variance, and range offer measures of spread.

Proportion is more closely related to percentiles and quartiles, representing the distribution within specific intervals.

Correlation the linear relationship between two variables.

These measures provides unique insights into different aspects of the data, and their relevance depends on the specific questions you want to answer about the dataset.

summary

The mean represents the average

The median is the middle value,

The mode is the most frequently occurring value

percentiles, quartiles, and proportion contribute to the understanding of the level of the data within specific intervals

standard deviation, variance, and range provide insights into the spread or variability of the entire dataset.

Correlation, not a direct measure of spread, informs about the relationship between two variables.

The coefficient of variation allows us to compare the relative variability between the two datasets

Together, these numerical summaries help to characterize and interpret different aspects of a dataset's distribution.

Analyse the run scores of Virat Kohli in 2023 matches

ODI World Cup 2023: Virat Kohli's Score In Each Match

85 vs Australia. Chennai.

55 not out vs Afghanistan. Delhi.

16 vs Pakistan. Ahmedabad.

103 not out vs Bangladesh. Pune.

95 vs New Zealand. Dharamsala.

0 vs England. Lucknow.

88 vs Sri Lanka. Mumbai.

101 not out vs South Africa. Kolkata.

Analyse the run scores of Virat Kohli in 2023

Numeric Summaries:

```
        MatchID       Runs
count  10.00000  10.000000
mean    5.50000  71.100000
std     3.02765  39.190277
min     1.00000   0.000000
25%     3.25000  52.000000
50%     5.50000  86.500000
75%     7.75000  99.500000
max    10.00000 117.000000
```

Analyse the run scores of Virat Kohli in 2023 matches

ODI World Cup 2023: Virat Kohli's Score In Each Match

85 vs Australia. Chennai.

55 not out vs Afghanistan. Delhi.

16 vs Pakistan. Ahmedabad.

103 not out vs Bangladesh. Pune.

95 vs New Zealand. Dharamsala.

0 vs England. Lucknow.

88 vs Sri Lanka. Mumbai.

101 not out vs South Africa. Kolkata.

Analyse the run scores of Virat Kohli in 2023

```
Numeric Summaries:
        MatchID        Runs
count  10.00000   10.000000
mean    5.50000   71.100000
std     3.02765   39.190277
min     1.00000    0.000000
25%     3.25000   52.000000
50%     5.50000   86.500000
75%     7.75000   99.500000
max    10.00000  117.000000
```

Task

1.Calculate the Mean, Median, Mode and Range for sample data

Set contains the numbers 2, 2, 3, 5, 5, 7, 8

2.provide numeric summaries for student marks data in Python ,use libraries like NumPy and Pandas.

Student Marks Data:

s.no	S.ID	Sub1	Sub2	Subt3
0	1	90	88	76
1	2	85	91	89
2	3	78	82	91
3	4	92	79	84
4	5	88	95	87

"TRY IT"

Answer 1

Mean: Adding the numbers gives:

2 + 2 + 3 + 5 + 5 + 7 + 8 = 32

There are 7 values, so you divide the total by 7: 32 ÷ 7 = 4.57

So the mean is 4.57

Median The numbers in order: 2 , 2 , 3 , (5) , 5 , 7 , 8

The middle value is marked in brackets, and it is 5.

So the median is 5

Mode The data values: 2 , 2 , 3 , 5 , 5 , 7 , 8

The values which appear most often are 2 and 5. They both ppear more time than any of the other data values.

So the modes are 2 and 5

Range The data values: 2 , 2 , 3 , 5 , 5 , 7 , 8 The lowest value is 2 and the highest value is 8. Subtracting the lowest from the highest

gives: 8 - 2 = 6

So the range is 6

Answer2

```
import pandas as pd
import numpy as np
# Sample student marks data
data = {
  'StudentID': [1, 2, 3, 4, 5],
  'Subject1': [90, 85, 78, 92, 88],
  'Subject2': [88, 91, 82, 79, 95],
  'Subject3': [76, 89, 91, 84, 87]
}
# Create a DataFrame
df = pd.DataFrame(data)
# Display the student marks data
print("Student Marks Data:")
print(df)
print("\n")
# Numeric summaries
summary = df.describe()
# Display numeric summaries
print("Numeric Summaries:")
print(summary)
```

```
Student Marks Data:
  StudentID  Subject1  Subject2  Subject3
0        1       90       88       76
1        2       85       91       89
2        3       78       82       91
3        4       92       79       84
4        5       88       95       87
Numeric Summaries:
       StudentID   Subject1   Subject2   Subject3
count   5.000000   5.000000   5.000000   5.000000
mean    3.000000  86.600000  87.000000  85.400000
std     1.581139   5.476163   6.782330   5.615057
min     1.000000  78.000000  79.000000  76.000000
25%     2.000000  85.000000  82.000000  84.000000
50%     3.000000  88.000000  88.000000  87.000000
75%     4.000000  90.000000  91.000000  89.000000
max     5.000000  92.000000  95.000000  91.000000
```

Scaling

Data are produced not given 'data' must be treated with caution.

"scaling" in univariate analysis generally refers to the process of transforming or standardizing a single variable. This can be done for various reasons, such as making variables comparable, improving the interpretability of results, or preparing data for certain statistical methods.

common techniques used for scaling in univariate analysis

Min-Max Scaling:

Formula: X scaled = [X−min(X)] / max(X)−min(X)

This scales the variable to a specific range, usually [0, 1]. It preserves the relative differences between values.

Z-Score Standardization (Standard Scaling):

Formula: Z=[X−mean(X)] / std(X)

This transforms the variable to have a mean of 0 and a standard deviation of 1. It is useful when comparing variables with different units or scales.

Robust Scaling:

Formula: Xscaled = [X-median(X)] / IQR(X)

This is similar to standardization but uses the interquartile range (IQR) instead of the standard deviation. It is less sensitive to outliers.

Log Transformation:

For variables with highly skewed distributions, taking the logarithm can help stabilize variance and make the distribution more symmetric.

Square Root Transformation:

Similar to the log transformation, the square root transformation is useful for stabilizing variance, especially when dealing with count data.

Box-Cox Transformation:

A more generalized power transformation that includes the log and square root transformations as special cases. It requires the data to be positive, and the transformation parameter is chosen to maximize normality.

Always consider the context of your analysis and the requirements of the specific statistical or machine learning methods you plan to use.

Standardizing

Standardizing in univariate analysis refers to the process of transforming a single variable to have a mean of 0 and a standard deviation of 1.

This transformation is also known as z-score standardization or standard scaling.

It is a common technique used to make different variables comparable and is particularly useful when dealing with variables that have different units or scales.

The formula for standardizing a variable x is given by:

Z= [(X−mean(X)] / std(X)

Here:

Z is the standardized value.

X is the original value of the variable.

mean(X) is the mean of the variable X.

std(X) is the standard deviation of the variable .

The resulting Z values have a mean of 0 and a standard deviation of 1.

z-scores

The standardized values (z-scores) provide a measure of how many standard deviations an observation or data point is from the mean.

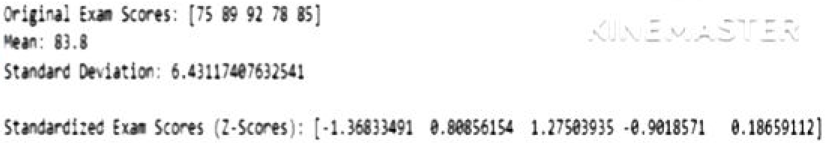
Positive z-scores indicate values above the mean, while negative z-scores indicate values below the mean.

These standardized scores are useful when you want to compare scores on a common scale.

Standardization is a common preprocessing step in data analysis, and it helps in making different variables comparable.

sample code

```
import numpy as np
# Example student dataset with exam scores
exam_scores = np.array([75, 89, 92, 78, 85])
# Standardizing the exam scores
mean_scores = np.mean(exam_scores)
std_scores = np.std(exam_scores)
# Z-score standardization
z_scores = (exam_scores - mean_scores) / std_scores
# Display the original scores and standardized scores
print("Original Exam Scores:", exam_scores)
print("Mean:", mean_scores)
print("Standard Deviation:", std_scores)
print("\nStandardized Exam Scores (Z-Scores):", z_scores)
```



```
Original Exam Scores: [75 89 92 78 85]
Mean: 83.8
Standard Deviation: 6.43117407632541

Standardized Exam Scores (Z-Scores): [-1.36833491  0.80856154  1.27503935 -0.9018571   0.18659112]
```

```
t with exam scores
5, 89, 92, 78, 85])
scores
xam_scores)
scores)

mean_scores) / std_scores
es and standardized scores
s:", exam_scores)
```

```
Original Exam Scores: [75 89 92 78 85]
Mean: 83.8
Standard Deviation: 6.43117407632541

Standardized Exam Scores (Z-Scores): [-1.36833491  0.80856154  1.27503935 -0
```

output

```
Original Exam Scores: [75 89 92 78 85]
Mean: 83.8
Standard Deviation: 6.43117407632541

Standardized Exam Scores (Z-Scores): [-1.36833491  0.80856154  1.27503935 -0.9018571   0.18659112]
```

Inequality

The overall happiness rating of a country will go up if income is distributed more equally.The definition of income usually only includes money spent on goods and services that are consumed privately.

But many things of great value to different people are organized at a collective level: health services, education, libraries, parks, museums.

Sources of income are often grouped into three types:

- **Earned income,** from either employment or self-employment;
- **unearned income** which accrues from ownership of investments, property, rent and so on;
- **transfer income**, that is benefits and pensions transferred on the basis of entitlement, not on the basis of work or ownership, mainly by the government but occasionally by individuals.

National Insurance and pension contributions, savings.

shares,deductions ,tax,Expenditure and Food Survey-- gross income

Inequality

Inequality in univariate analysis, it refers to the distribution of values within that variable.

Range: The range is the simplest measure of inequality and is calculated as the difference between the maximum and minimum values in a dataset. A larger range indicates greater variability.

Variance and Standard Deviation: These measures quantify the spread or dispersion of values around the mean.

A higher variance or standard deviation suggests greater inequality.

Interquartile Range (IQR): The IQR is the range of the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). A larger IQR can indicate more variability in the central part of the distribution.

Inequality

Coefficient of Variation (CV): The coefficient of variation is the ratio of the standard deviation to the mean, expressed as a percentage. It is a relative measure of variability, allowing for comparison across variables with different scales.

Skewness: Skewness measures the asymmetry of a distribution.

A positive skewness indicates a longer right tail, while a negative skewness indicates a longer left tail.

Inequality

Gini Coefficient: The Gini coefficient is commonly used to measure income or wealth inequality. It ranges from 0 to 1, where 0 represents perfect equality, and 1 represents perfect inequality. A higher Gini coefficient indicates greater inequality.

A measure that summarizes what is happening across all the distribution is the Gini coefficient.

Lorenz Curve: The Lorenz curve is a graphical representation of income or wealth distribution. It plots the cumulative percentage of total income received by the cumulative percentage of the population, highlighting inequality visually.

Lorenz curves and Gini coefficients have been most widely applied to the study of inequality

Smoothing time series

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time.What sets time series data apart from other data is that the analysis can show how variables change over time.time series data can be used for forecasting—predicting future data based on historical data.Ex Weather data,Stock prices,Rainfall measurements.

Smoothing is usually done to help us better see patterns

Smoothing time series data is a common technique used to remove noise and highlight underlying trends or patterns.

Moving Averages:

Simple Moving Average (SMA): Calculates the average of a specified number of adjacent data points. This helps in reducing short-term fluctuations.

Exponential Moving Average (EMA): Places more weight on recent observations, making it responsive to changes in the underlying trend.

Lowess (Locally Weighted Scatterplot Smoothing):

A non-parametric method that fits a series of local weighted regressions to the data, providing a smoothed curve that captures trends.

Smoothing time series

Kernel Smoothing:

Utilizes a kernel function to smooth the data. The choice of the kernel and its bandwidth influences the level of smoothing.

Exponential Smoothing:

A time series forecasting method that assigns exponentially decreasing weights to past observations. It includes methods like Single Exponential Smoothing, Double Exponential Smoothing (Holt's method), and Triple Exponential Smoothing (Holt-Winters method).

Smoothing time series

Fourier Transform:

Decomposes a time series into its frequency components, allowing the removal of high-frequency noise.

Wavelet Transform:

Decomposes a time series into different frequency bands, making it useful for both short-term and long-term trend extraction.

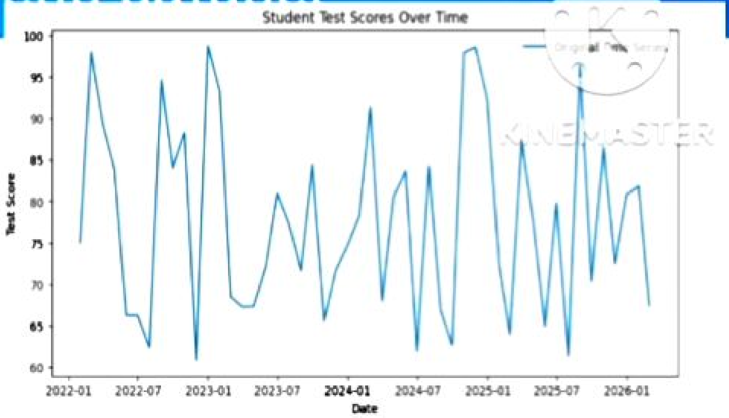
choosing a smoothing method, it's essential to consider the characteristics of the time series, such as the presence of seasonality, trend, and the nature of the noise.

To experiment with different smoothing methods and assess their performance based on the specific goals of your analysis.

Keep in mind that over-smoothing can result in the loss of important information, so a balance must be struck based on the characteristics of your data and the objectives of your analysis.

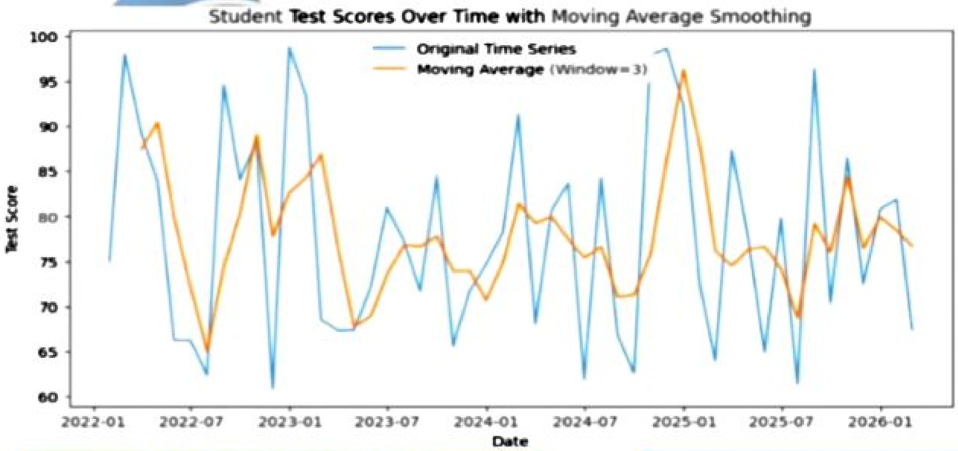
generate a synthetic dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# Generate synthetic student dataset
np.random.seed(42)
# Create a time series with 50 data points
time_series = pd.date_range(start='2022-01-01', periods=50, freq='M')
# Generate random test scores between 60 and 100
test_scores = np.random.uniform(60, 100, 50)
# Create a DataFrame
student_data = pd.DataFrame({'Date': time_series, 'Test_Score': test_scores})
# Plot the original time series
plt.figure(figsize=(10, 5))
plt.plot(student_data['Date'], student_data['Test_Score'], label='Original Time Series')
plt.title('Student Test Scores Over Time')
plt.xlabel('Date')
plt.ylabel('Test Score')
plt.legend()
plt.show()
```



Moving Average Smoothing

```
# Applying a simple moving average with a window size of 3
student_data['MA_Smoothed'] = student_data['Test_Score'].rolling(window=3).mean()
# Plot the original time series and the smoothed version
plt.figure(figsize=(10, 5))
plt.plot(student_data['Date'], student_data['Test_Score'], label='Original Time Series', alpha=0.7)
plt.plot(student_data['Date'], student_data['MA_Smoothed'], label='Moving Average (Window=3)')
plt.title('Student Test Scores Over Time with Moving Average Smoothing')
plt.xlabel('Date')
plt.ylabel('Test Score')
plt.legend()
plt.show()
```



Virtual Classroom Lectures