DATA SCIENCE – UNIVARIATE ANALYSIS

NORMAL DISTRIBUTION

def get_pdf_probability

Functions Code Explanation:-

1. **def get_pdf_probability(dataset, startrange, endrange):**

- This line defines a function named get_pdf_probability. Think of a function like a mini-program within your main program. It takes three inputs:
- dataset: A collection of data (like a list of ages or heights).
- startrange: The beginning of the range you're interested in (e.g., age 10).

- endrange: The end of the range (e.g., age 20).

2. from matplotlib import pyplot

- This line imports a tool called pyplot from the matplotlib library. matplotlib is used for creating graphs and plots.

3. from scipy.stats import norm

- This line imports the norm function from the scipy.stats library. The norm function is used for working with the normal distribution (a common bell-shaped curve in statistics).

4. import seaborn as sns

- This line imports the seaborn library, which is built on top of matplotlib and makes it easier to create nice-looking statistical graphs.

5. ax = sns.distplot(dataset, kde=True, kde_kws={'color': 'blue'}, color='Green')

- This line creates a histogram (a type of graph) of your dataset using seaborn.
- kde=True adds a smooth curve (called a kernel density estimate) to the histogram.
- kde_kws and color control the colors of the curve and histogram.

6. pyplot.axvline(startrange, color='Red')

- This draws a vertical red line on the graph at the startrange value.

7. pyplot.axvline(endrange, color='Red')

- This draws another vertical red line at the endrange value.

8. sample = dataset

- This line creates a variable named sample and assigns it the value of dataset. In this case, the sample is the entire dataset.

9. sample_mean = sample.mean()

- This calculates the average (mean) of the data in the sample and stores it in the variable sample_mean.

10. sample_std = sample.std()

- This calculates the standard deviation of the data in the sample and stores it in the variable sample_std. The standard deviation measures how spread out the data is.

11. print('Mean=%.3f, Standard Deviation=%.3f' % (sample_mean, sample_std))

- This line prints the calculated mean and standard deviation to the console (the text output area).

12. dist = norm(sample_mean, sample_std)

- This creates a normal distribution object (dist) using the calculated mean and standard deviation.

13. (one-liner for loop)
values = [value for value in range(startrange, endrange)]

- This creates a list of numbers from startrange to endrange (not including endrange).

14. probabilities = [dist.pdf(value) for value in values]

- This calculates the probability of each value in the values list occurring in the normal distribution (dist). pdf stands for probability density function.

15. prob = sum(probabilities)

- This calculates the total probability by summing the individual probabilities in the probabilities list.

16. print("The area between range({},{}): {}".format(startrange, endrange, sum(probabilities)))

- This prints the calculated total probability (the area under the curve between the two red lines) to the console.

17. return prob

- This line makes the function return the calculated probability value (prob) when it's called.

Understanding Probability Calculation with Python

What this code does:

1. Visualizes Data: It creates a graph (histogram) of your data and highlights a specific range on it.
2. Calculates Average and Spread: It computes the average (mean) and spread (standard deviation) of your data.
3. Normal Distribution: It uses the normal distribution (bell curve) to model your data.

4. **Probability Calculation: It calculates the probability of data falling within a specified range on the graph.**
5. **Output: It prints the calculated probability and displays it on the graph.**

Key Concepts:

- **Functions: Reusable blocks of code that perform specific tasks.**
- **Data Visualization: Using graphs to understand data better.**
- **Statistics: Calculating mean and standard deviation to summarize data.**
- **Normal Distribution: A common pattern in data, used for probability calculations.**
- **Probability: The chance of an event occurring.**