Beginner's Guide to Correlation, Probability, and Plots in Python

Prepared for New Learners

Contents

1 Introduction

Data analysis requires both **mathematical understanding** and **visual exploration**. This guide explains:

- Correlation Coefficient
- Probability Basics
- Univariate, Bivariate, and Multivariate Plots

For each plot, we cover:

- Definition
- Purpose and Usage (Why we use it)
- Python Code Example
- Small Exercise

2 Correlation Coefficient

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}}$$

Purpose & Usage

- Measure strength and direction of linear relationship between two variables.
- Used in finance (stock price correlations), healthcare (symptom vs. disease), education (study hours vs. exam scores).

Python Example

```
import numpy as np
import pandas as pd

data = {’StudyHours’: [2,4,6,8,10],
        ’ExamScore’: [50,55,65,70,90]}
df = pd.DataFrame(data)

print(df.corr())
```

3 Univariate Plots

3.1 Histogram

Purpose & Usage:

- Shows frequency distribution of one variable.

- Useful for marks distribution, age distribution, sales per day.

```
plt.hist(data, bins=20, color='skyblue', edgecolor='black')
```

3.2 Boxplot

Purpose & Usage:

- Summarizes data with median, quartiles, and outliers.
- Useful in salary comparison, medical data (blood pressure range).

```
sns.boxplot(x=data, color="lightgreen")
```

3.3 KDE Plot

Purpose & Usage:

- Smooth probability density curve of variable.
- Used when we want continuous estimation of distribution (heights, weights).

```
sns.kdeplot(data, fill=True, color="red", bw_adjust=0.5)
```

4 Bivariate Plots

4.1 Scatter Plot

Purpose & Usage:

- Shows relationship between two variables.
- Useful for: advertising budget vs sales, hours studied vs marks.

```
sns.scatterplot(x="StudyHours", y="ExamScore", data=df)
```

4.2 Joint Plot

Purpose & Usage:

- Combines scatter + marginal histograms/KDEs.
- Good for deeper exploration: income vs expenditure, height vs weight.

```
sns.jointplot(x="StudyHours", y="ExamScore", data=df, kind="kde")
```

4.3 Heatmap

Purpose & Usage:

- Shows correlation matrix in colored grid.
- Widely used in machine learning feature selection.

```
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
```

5 Multivariate Plots

5.1 Pairplot

Purpose & Usage:

- Pairwise scatter plots and histograms for all variables.
- Ideal for exploring datasets with 4–10 features (like Iris dataset).

```
sns.pairplot(df)
```

5.2 3D Scatter Plot

Purpose & Usage:

- Visualizes relationship among 3 variables.
- Useful in physics (speed, distance, time), finance (price, volume, volatility).

```
ax.scatter(df["StudyHours"], df["ExamScore"], np.random.randn(len
    (df)))
```

6 Probability Basics

Purpose & Usage

- Describes uncertainty of events.
- Used in gambling, weather forecasting, quality control, AI decision making.

$$P(E) = \frac{\text{Favorable Outcomes}}{\text{Total Outcomes}}$$

```
import random
outcomes = ["H","T"]
heads = sum([1 for _ in range(10000) if random.choice(outcomes)==
    "H"])
print("P(Head):", heads/10000)
```

7 Quick Reference Table

Plot	**Purpose**	**Python Code**
Histogram	Frequency of values	plt.hist(data,bins=20)
Boxplot	Spread + outliers	sns.boxplot(x=data)
KDE	Smooth density	sns.kdeplot(data)
Scatter	Relationship (x,y)	sns.scatterplot(x,y)
Jointplot	Relationship + marginals	sns.jointplot(x,y,kind="kde")
Heatmap	Correlation matrix	sns.heatmap(df.corr())
Pairplot	Multi-variable exploration	sns.pairplot(df)
3D Scatter	Three-variable relation	ax.scatter(x,y,z)

8 Conclusion

- Histograms, KDEs, Boxplots → explore single variable.
- Scatter, Jointplot, Heatmap → compare two variables.
- Pairplot, 3D Scatter → study multiple variables.
- Probability and correlation provide mathematical backbone.
- Purpose & Usage helps select the right visualization.