Class 5: Data Visualization with Seaborn and Plotly

Overview

Data visualization is essential for understanding patterns in data. This class note covers statistical plotting using Seaborn and Plotly, focusing on relational, distributional, and categorical plots to derive insights, and converting Seaborn plots to Plotly for interactive visualizations.

Objectives

- Create various statistical plots using Seaborn.
- Understand insights derived from statistical plots.
- Convert Seaborn plots to equivalent Plotly visualizations.

Key Concepts

1. **Relational Plots**: Visualize relationships between variables.
    - *Scatter Plot*: Shows correlation between two variables (e.g., total bill vs. tip).
    - *Line Plot*: Displays trends over a continuous variable (e.g., time vs. signal).
2. **Distributional Plots**: Analyze variable distributions.
    - *Univariate*: Distribution of one variable (e.g., histogram of flipper length).
    - *Bivariate*: Distribution of two variables (e.g., bill length vs. depth).
    - *Multivariate*: Pair plots for multiple variables.
3. **Categorical Plots**: Explore relationships involving categorical data.
    - *Strip/Swarm*: Individual data points by category.
    - *Box/Violin*: Distribution summaries by category.
    - *Count*: Frequency of categories.
4. **Plotly**: Interactive visualizations, equivalent to Seaborn plots.

Plot Types and Interpretations

Plot Type	Sub-Type	Interpretation
Relational	Scatter	Shows correlation or relationship strength between two continuous variables.
	Line	Displays trends or changes over a continuous variable (e.g., time).
Distributional	Histogram	Reveals the frequency distribution of a single variable (univariate).
	KDE	Shows the probability density of a variable, smoothing the distribution.
	Bivariate	Visualizes joint distribution of two variables (e.g., scatter with density).
	Pair Plot	Displays pairwise relationships and distributions for multiple variables.
Categorical	Strip/Swarm	Shows individual data points for categories, highlighting spread.
	Box	Summarizes distribution (quartiles, outliers) per category.
	Violin	Combines density and quartiles for categorical data distributions.
	Count	Displays frequency of categorical variable occurrences.
Other	Pie	Shows proportional distribution of categories (e.g., percentage per group).

Table 1: Types of statistical plots and their data interpretations.

Example Code

```
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt

% Setting Seaborn style
sns.set(style="darkgrid", palette="flare", font="Arial", font_scale=1)

% Loading datasets
tips = sns.load_dataset("tips")
penguins = sns.load_dataset("penguins")
fmri = sns.load_dataset("fmri")

% Relational scatter plot (Seaborn)
sns.relplot(
    data=tips, x="total_bill", y="tip", kind="scatter", hue="sex", col="day",
    height=4
).set(xlabel="Total Bill", ylabel="Tip", title="Total Bill vs Tip")

% Equivalent Plotly scatter plot
px.scatter(
    data_frame=tips, x="total_bill", y="tip", color="sex"
).update_layout(xaxis_title="Total Bill", yaxis_title="Tip", title="Total Bill vs
    Tip")

% Distributional histogram (Seaborn)
sns.displot(
    data=penguins, x="flipper_length_mm", hue="species", bins=20, col="species"
).set(xlabel="Flipper length (mm)", ylabel="Count")

% Categorical box plot (Seaborn)
```

```
sns.catplot(
    data=tips, x="day", y="total_bill", kind="box", hue="smoker"
).set(xlabel="Day", ylabel="Total Bill")

% Pair plot for multivariate analysis
sns.pairplot(penguins)

% Pie chart with Matplotlib
x = tips["day"].value_counts()
plt.pie(x, labels=x.index, autopct="%1.1f%%")
plt.title("Distribution of Tips by Day")
```

Key Takeaways

Seaborn and Plotly enable powerful data visualization. Seaborn provides static, statistical plots to uncover data patterns, while Plotly offers interactive equivalents, enhancing data exploration for machine learning and data science applications.