What is Categorical plot?

Categorical Plots in Seaborn are a family of visualizations designed specifically to show the **relationship between a numerical variable and one or more categorical variables**, or to visualize the distribution of a categorical variable itself. They are essential for comparing groups, understanding variations across categories, and identifying patterns within distinct segments of your data.

Purpose of Categorical Plots

The primary **purpose** of categorical plots is to **visualize and compare statistical properties (like central tendency, spread, or counts) across different discrete groups or categories** within your dataset. This allows you to:

- **Compare Group Means/Medians:** See how the average or typical value of a numerical variable differs between categories.
- **Assess Variability within Groups:** Understand the spread or distribution of data within each category.
- **Identify Outliers per Category:** Spot unusual data points within specific groups.
- **Show Counts:** Visualize the frequency of observations in each category.
- **Explore Multi-Variable Relationships:** Understand how a numerical variable is influenced by one or more categorical factors.
- **Generate Insights for Segmentation:** Gain insights into the characteristics of different customer segments, product categories, or regions.

How Categorical Plots Work and Why They Are Required

Seaborn provides a unified interface for categorical plots through the **figure-level function catplot()**, which acts as a "wrapper" around several axes-level categorical functions. This allows for easy faceting (creating grids of subplots) based on additional categorical variables.

1. **Core Mapping (x, y):**
    - **What it does:** You typically map one categorical column to one axis (e.g., x) and a numerical column to the other axis (e.g., y). If you're

just counting categories, you might only specify one categorical axis.

- **Why it's required:** These are the primary variables whose relationship or distribution you want to visualize across categories.

2. **Plot Type (kind):**
   - **What it does:** This crucial parameter determines the specific type of categorical plot to draw, each offering a different perspective on the data:
     - **kind='strip':** Draws a scatter plot where one variable is categorical. Points are "jittered" to prevent overlap, showing the raw distribution of values within each category.
     - **kind='swarm':** Similar to strip, but points are adjusted (without overlapping) along the categorical axis, giving a better representation of density.
     - **kind='box':** Draws a box plot (boxplot), showing the median, quartiles (25th and 75th percentiles), and potential outliers for the numerical data within each category. Excellent for comparing distributions.
     - **kind='violin':** Draws a violin plot, which is similar to a box plot but also shows the kernel density estimate of the data's distribution within each category, giving a richer view of density.
     - **kind='bar':** Draws a bar plot, where the height of each bar represents the mean (by default) of the numerical variable for each category, with error bars indicating variability.
     - **kind='count':** Draws a bar plot where the height of each bar represents the number of observations (count) in each category. Useful for visualizing the frequency of categorical values.
     - **kind='point':** Draws a point plot, showing the point estimate (e.g., mean) and confidence intervals for a numerical variable

across categories. Useful for comparing changes across ordered categories.

  - **Why it's required:** Allows you to choose the most appropriate visual representation to answer specific questions about your categorical data (e.g., comparing medians, showing raw data points, or visualizing counts).

3. **Faceting/Gridding (col, row):**
   - **What it does:** As a figure-level function, catplot() can create separate subplots (facets) arranged in columns (col) or rows (row) based on additional categorical variables.
   - **Why it's required:** Essential for comparing how the relationship between your primary variables changes across different groups defined by a third or fourth categorical variable. For example, comparing product sales by region, faceted by store type.
4. **Semantic Mappings (hue):**
   - **What it does:** Maps another categorical variable to the **color** of the plot elements within each subplot.
   - **Why it's required:** Allows for visualizing an additional layer of categorical information within each plot, enabling more complex comparisons (e.g., comparing male vs. female sales within each product category).

Conceptual Example:

Imagine you have a DataFrame with customer transaction data, including columns:

- Product_Category (e.g., 'Electronics', 'Apparel', 'Home Goods')
- Purchase_Amount (numerical)
- Customer_Segment (e.g., 'New', 'Loyal', 'VIP')
- Payment_Method (e.g., 'Credit Card', 'Cash', 'Online Wallet')

Using catplot() to visualize purchase amounts by product category and customer segment:

If you wanted to understand the distribution of Purchase_Amount for each Product_Category, and how this varies across Customer_Segment, you could use catplot():

- **x='Product_Category'** (categorical variable on the x-axis)
- **y='Purchase_Amount'** (numerical variable on the y-axis)
- **kind='box'** (to see the distribution and outliers using box plots)
- **col='Customer_Segment'** (to create separate columns of plots for each customer segment)
- **hue='Payment_Method'** (to color the box plots by payment method within each segment)

What you would see:

You would get a grid of box plots. Each column in the grid would represent a different Customer_Segment (e.g., one plot for 'New' customers, one for 'Loyal', etc.). Within each of these plots, you would see box plots of Purchase_Amount for each Product_Category, with each box plot colored according to the Payment_Method. This allows for a rich comparison: "Do 'VIP' customers spend more on 'Electronics' than 'New' customers?" and "Does the typical purchase amount vary by Payment_Method within each Product_Category and Customer_Segment?"

Why are Categorical Plots Required?

Categorical plots are indispensable in data science for:

- **Group Comparisons:** They are the go-to tools for visually comparing numerical distributions or counts across distinct groups.
- **Understanding Variability:** They help in assessing the spread and outliers within each category, which is crucial for robust analysis.
- **Identifying Drivers:** By visualizing how a numerical outcome changes with different categorical factors, you can identify key drivers or segments.
- **Data Quality Checks:** Spotting unexpected distributions or missing categories.

- **Feature Engineering:** Informing decisions about how to encode or transform categorical variables for machine learning models.
- **Business Insights:** Providing clear, actionable insights into customer behavior, product performance, or regional differences.

In summary, categorical plots in Seaborn provide a powerful and intuitive way to visualize relationships between numerical and categorical data, making them fundamental for exploratory data analysis and communicating group-based insights.