Pivot Tables and Cross Tabulation in EDA

Pivot tables and cross-tabulation are essential tools in EDA to summarize and analyze data, especially for categorical variables.

1. Pivot Tables

A pivot table is used to summarize data by grouping it based on one or more categorical variables and applying aggregation functions like sum, mean, count, etc.

Example Dataset

Consider a dataset with the following structure:

Category	Subcategory	Sales	Quantity
A	X	100	10
A	Y	200	15
B	X	150	12
B	Y	300	20

Code Example in Python

```
import pandas as pd

# Sample data
data = {
   'Category': ['A', 'A', 'B', 'B'],
   'Subcategory': ['X', 'Y', 'X', 'Y'],
   'Sales': [100, 200, 150, 300],
   'Quantity': [10, 15, 12, 20]
}

df = pd.DataFrame(data)

# Creating a pivot table
pivot_table = pd.pivot_table(df, values='Sales', index='Category', columns='Subcategory',
aggfunc='sum', fill_value=0)

print(pivot_table)
```

Output

```
Subcategory      X    Y
Category
 A             100  200
 B             150  300
```

- **pd.pivot_table()**: Creates a pivot table from the dataset.
  - **values='Sales'**: Specifies the column to aggregate (e.g., Sales).
  - **index='Category'**: Groups data by the Category column.
  - **columns='Subcategory'**: Breaks the data into columns based on Subcategory.
  - **aggfunc='sum'**: Aggregates the Sales column using the sum function.
  - **fill_value=0**: Replaces missing values with 0.

This pivot table shows the sum of sales for each combination of Category and Subcategory.

2. **Cross Tabulation**

Cross-tabulation is a special case of a pivot table where it summarizes data by counting occurrences or calculating other statistics.

Example Dataset

We use the same dataset as above.

Code Example in Python

```
# Cross-tabulation
cross_tab = pd.crosstab(df['Category'], df['Subcategory'])

print(cross_tab)
```

Output

```
Subcategory  X  Y
Category
 A           1  1
 B           1  1
```

- **pd.crosstab()**: Generates a cross-tabulation table.
  - **df['Category']**: Rows are grouped by Category.
  - **df['Subcategory']**: Columns are grouped by Subcategory.

- The output shows the count of occurrences of each combination of Category and Subcategory.

Advanced Cross-Tabulation with Aggregation

You can also calculate statistics like sum or mean in cross-tabulation.

```
# Cross-tabulation with aggregation
cross_tab_agg = pd.crosstab(df['Category'], df['Subcategory'], values=df['Sales'], aggfunc='sum', margins=True)

print(cross_tab_agg)
```

Output

Subcategory	X	Y	All
Category			
A	100	200	300
B	150	300	450
All	250	500	750

- **values=df['Sales']**: Specifies the column to aggregate.
- **aggfunc='sum'**: Aggregates using the sum function.

- **margins=True**: Adds row and column totals.
- **Pivot Tables**: Provide flexibility to summarize data using multiple aggregation functions.
- **Cross Tabulation**: Ideal for analyzing relationships between categorical variables.