Grouping and Correlation Analysis in Pandas

Using Pandas for Data Analysis, Correlation, and Visualization

Introduction to Grouping and Correlation Analysis

Learning Objectives:

- Use groupby() for summary statistics.

- Apply advanced indexing techniques.

- Perform correlation analysis.

- Visualize correlations with heatmaps.

Using groupby() for Summary Statistics

1. What is groupby()?

- The groupby() function in Pandas is used to split the data into groups based on some criteria, then apply an aggregation function

2. Advanced Indexing in Pandas

Advanced indexing techniques help in selecting subsets of data efficiently.

- **Using .loc[] for Label-based Indexing**
- Using .iloc[] for Position-based Indexing
- Multi-indexing with groupby()

3. Correlation Analysis in Pandas

- Correlation measures the strength of relationships between numerical variables.
- **1.0**: Perfect positive correlation.
- **-1.0**: Perfect negative correlation.
- **0**: No correlation

Using groupby() for Summary Statistics

What is groupby()?

- Splits data into groups based on criteria.

- Apply aggregation functions like sum(), mean(), count().

- Useful for summarizing data by categories.

Example:

df.groupby('Category').agg({'Sales': 'mean', 'Profit': 'sum'})

Advanced Indexing in Pandas

Selecting Subsets of Data:

- .loc[]: Label-based indexing.

- .iloc[]: Position-based indexing.

- Multi-indexing with groupby() for hierarchical data.

Example:

```
df.loc[df['Sales'] > 150]
```

```
df.iloc[0:3]
```

Correlation Analysis in Pandas

Understanding Correlation:

- Measures strength of relationships between numerical variables.

- Correlation ranges from -1.0 to 1.0.

Computing Correlation Matrix:

```
correlation_matrix = df.corr()
```

Identifying Strong Correlations:

```
strong_corr = correlation_matrix[abs(correlation_matrix) > 0.8]
```

Visualizing Correlation with Heatmaps

Creating Heatmaps using Seaborn:

- Heatmaps provide a visual representation of correlation matrices.

Example Code:

```
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
```

- Helps identify strong and weak correlations at a glance.

Function Definitions

Definitions of all functions used in the code:
1. pd.read_csv(): Read CSV file into a DataFrame.
2. df.head(): Display first 5 rows of the DataFrame.
3. df.isnull(): Check for missing values.
4. df.info(): Display DataFrame structure.
5. df.groupby(): Group data by categories.
6. agg(): Apply aggregation functions.
7. reset_index(): Convert index labels back to columns.
8. corr(): Compute correlation matrix.
9. sns.heatmap(): Visualize matrix as a heatmap.
10. plt.figure(): Create new plotting figure.
11. plt.title(): Set plot title.
12. plt.show(): Display the plot.

Summary

What we learned:

- Using groupby() for summary statistics.

- Advanced indexing techniques in Pandas.

- Performing correlation analysis.

- Visualizing correlations with heatmaps.

Would you like to proceed with hands-on exercises?