

# Topic 4 DQ 2

Exploring relationships between variables in a dataset is a fundamental step in data analysis, offering insights into patterns, trends, and potential causal connections. Python's rich ecosystem of data visualization libraries, such as Matplotlib, Seaborn, and Plotly, provides powerful tools for this purpose. Here, we'll discuss two creative ideas for visualizing relationships between variables and demonstrate how to implement these ideas using Python.

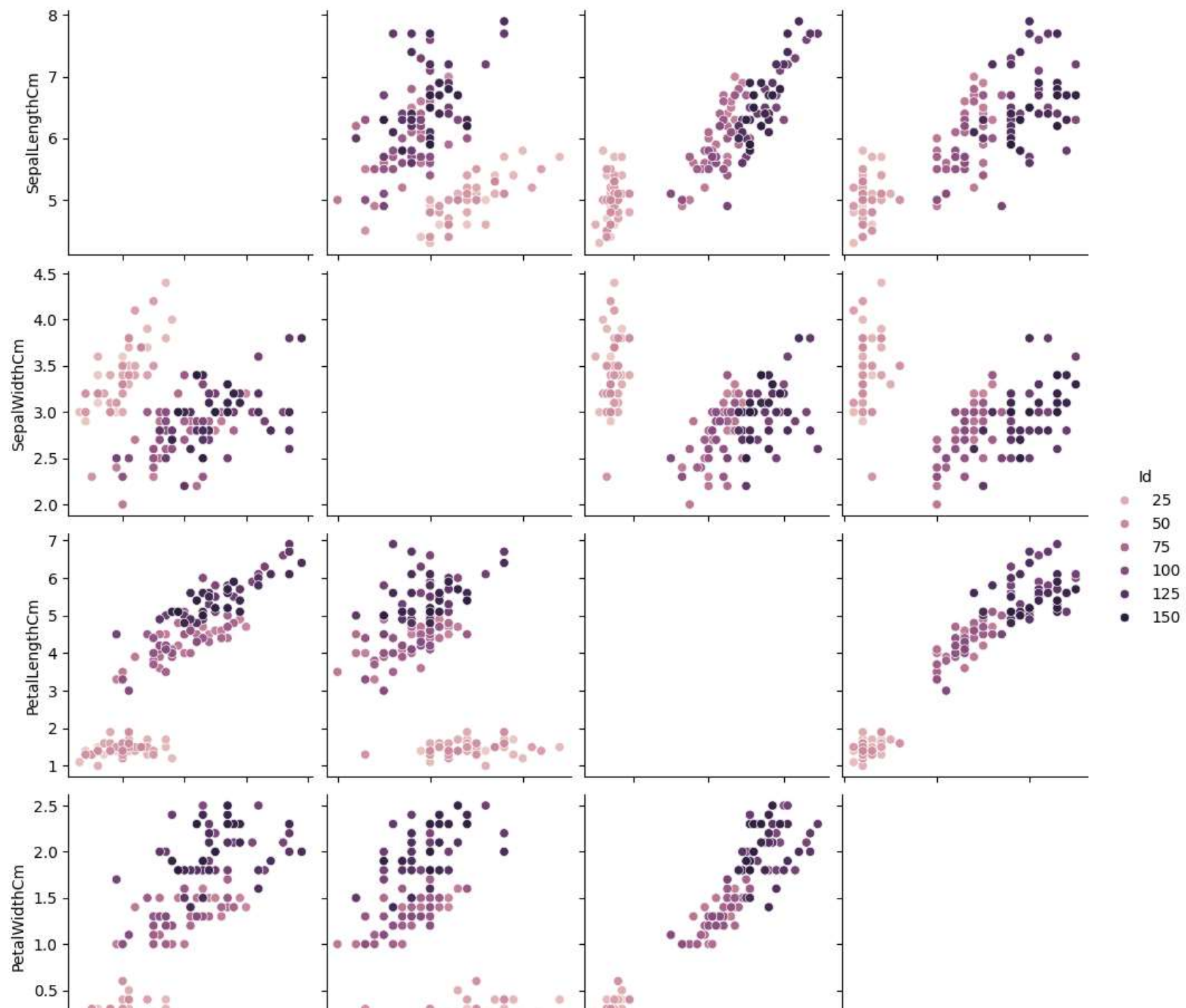
## Pair Plots

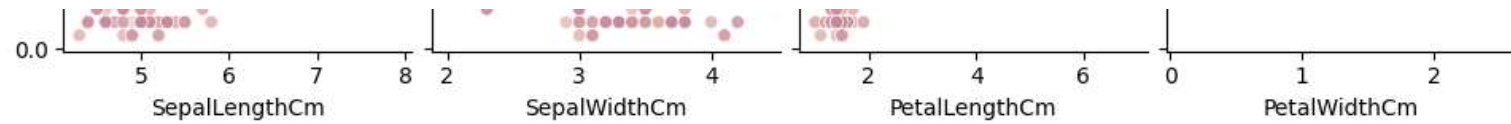
A pair plot is a fantastic way to visualize paired relationships across multiple variables. It provides a comprehensive overview of how each variable relates to the others, allowing for the identification of correlations, clusters, and potential outliers.

A pair plot showing the relationships between four variables. Each cell in the matrix represents a scatter plot for the variable pair, with histograms along the diagonal showing the distribution of each variable.

```
In [ ]: import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv('iris.csv')
sns.pairplot(df, hue='Id')
plt.show()
```





The pair plot of the Iris dataset provides a comprehensive overview of the relationships between features. By examining these plots, you can identify features that correlate, assess feature importance for classifying the species, and spot any outliers or anomalies in the dataset.

## Heatmap of Correlation Matrix

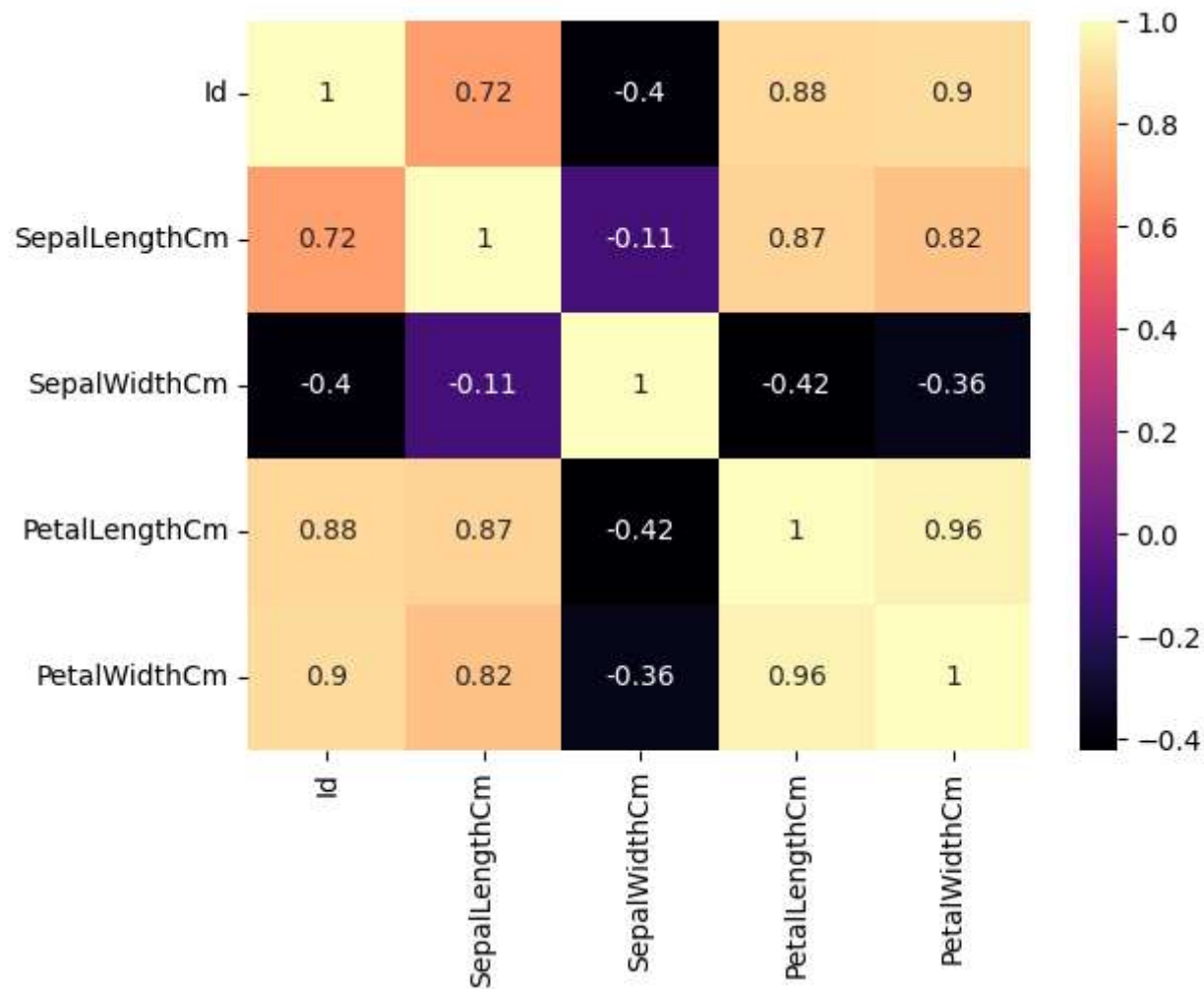
A heatmap of the correlation matrix is another powerful tool for identifying relationships between variables. It visualizes the correlation coefficients between each pair of variables in the dataset, making it easy to spot highly correlated variables at a glance.

```
In [ ]: #assuming df is your DataFrame
df = pd.read_csv('iris.csv')

#select only the numeric columns for correlation calculation
numeric_df = df.select_dtypes(include=['float64', 'int64'])

#calculate the correlation matrix on the numeric columns only
corr = numeric_df.corr()

#generate a heatmap
sns.heatmap(corr, annot=True, cmap='magma')
plt.show()
```



A heatmap displaying the correlation matrix of a dataset. Each cell's color intensity represents the strength and direction of the correlation between variables, with warmer colors indicating positive correlations and cooler colors indicating negative correlations.

## Sources

Cheng, J. (2018). Data Exploration and Visualization with Seaborn Pair Plots. Medium. Retrieved from <https://medium.com/@jaimejcheng/data-exploration-and-visualization-with-seaborn-pair-plots-40e6d3450f6d>

Chugani, V. (2024). Mastering Pair Plots for Visualization and Hypothesis Creation in the Ames Housing Market. Machine Learning Mastery. Retrieved from <https://machinelearningmastery.com/pair-plots/>

Rogel-Salazar, J. (2023). Dazzling Data Designs – Creating Charts. In Statistics and data visualisation with Python (1st ed.). CRC Press.