Visualize Pairwise Correlations in PalmerPenguins Dataset via Seaborn Package

Dachuan Shi

Abstract

Correlation analysis among variables is a fundamental problem in statistical learning, and Seaborn is a Python package that is dedicated to data visualization. This manuscript conducts experiments on the PalmerPenguins dataset, and demonstrates correlation analysis diagrams created by Seaborn package.

1. Introduction

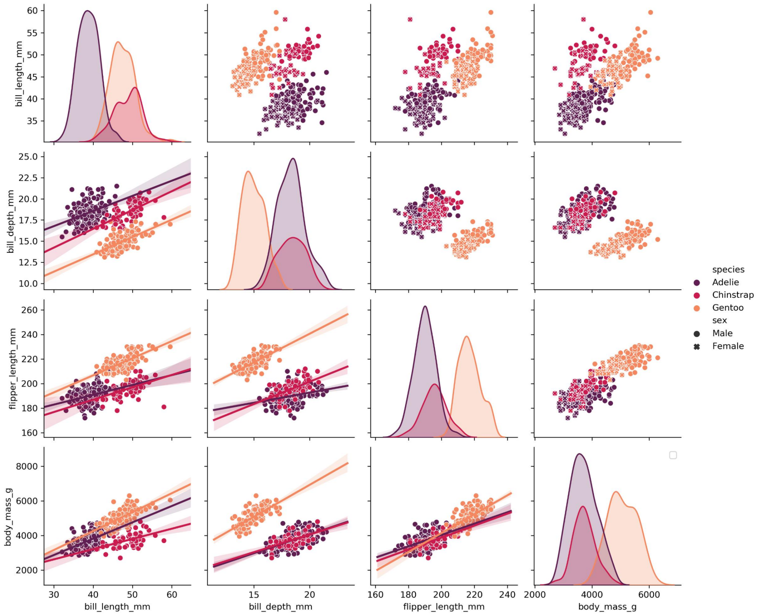
Palmerpenguins is a widely used statistical learning dataset. Penguins is a subset of the Palmerpenguins dataset, and contains data for 344 penguins. Each item in the dataset consists of 6 attributes, that is species, body mass, bill length, bill depth, flipper length, and sex. Species has three categories, that is “Adelie”, “Chinstrap”, and “Getoo”. Sex has two categories, that is “Male” and “Female”. And all other attributes are of numeric type. Seaborn is an open-source data visualization based on Python. With the help of Seaborn, pairwise correlations among attributes of the dataset can be visualized. Furthurmore, linear regression analysis can also be conducted upon each pair of attributes.

2. Visualization Results

Visualization results are illustrated in the bottom figure. PS. The original figure with higer resolution can be found in the attachment. As shown in the legend, data points with different colors represent different species. On the diagonal of the figure is the marginal distribution of each attribute, while off the diagonal of the figure is the correlation scatter diagram between two different attributes. For example, the subfigure in the first row and the first column demonstrates the distribution of different species with respect to the bill length attribute. And the subfigure in the third row and the second column demonstrates the correlation between flipper length attribute and the bill depth attribute. Besides, for subfigure in the lower triangular area, linear regression results with confidence intervals are added on them. And for subfigure in the upper triangular area, data points are further distinguish according to the sex attribute.

3. Conclusion

From the visualization diagram, we can summarize some conclusions. For example, the subfigure in the first row and the first column demonstrates that in general, bill of “Adelie” is the shortest, “Chinstrap” is the second, and ”Gentoo” is the longest. The subfigures in the lower triangular area demonstrate that in general, there is a positive correlation between any pair of four attributes. The subfigures in the upper triangular area demonstrate that in general, male penguins have higher four attributes than female penguins.


bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
bill_length_mm
bill_depth_mm
flipper_length_mm
body_mass_g
species
● Adelie
● Chinstrap
● Gentoo
sex
● Male
✖ Female