A brief report on Exploring and analysing different relations in the iris dataset: By TIASA MALITYA, Id – 24005709, https://github.com/Tiasa24/ASSIGHNMENT5709ADS1

Introduction

Ronald Fisher, a British statisticians and biologists first introduced a data set of 150 samples of Iris flower in 1936 which is known as the classic Iris Flower Dataset. He considered three different species of Iris flower which are Iris setosa, Iris virginica, Iris versicolor and set up a multivariate data of sepal length, sepal width, petal length and petal width that consists of 50 samples of each species. The report explores relation between the variables of iris dataset, analyses their distribution and explains with visualisation how they correlate with each other statistically. Furthermore, Iris dataset has wide range applications in the field of machine learning.

Explanation of each Visualisation

The dataset which has been used for analysing is “iris flower dataset.csv”. The code starts with importing every necessary library package which are NumPy, matplotlib, pandas, seaborn, SciPy. Stats.

Fig 1: Distribution of Sepal length

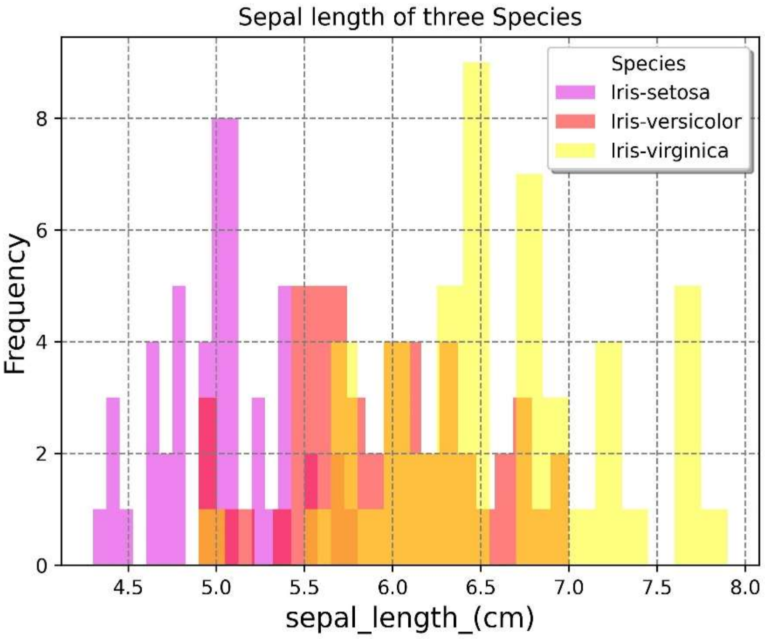
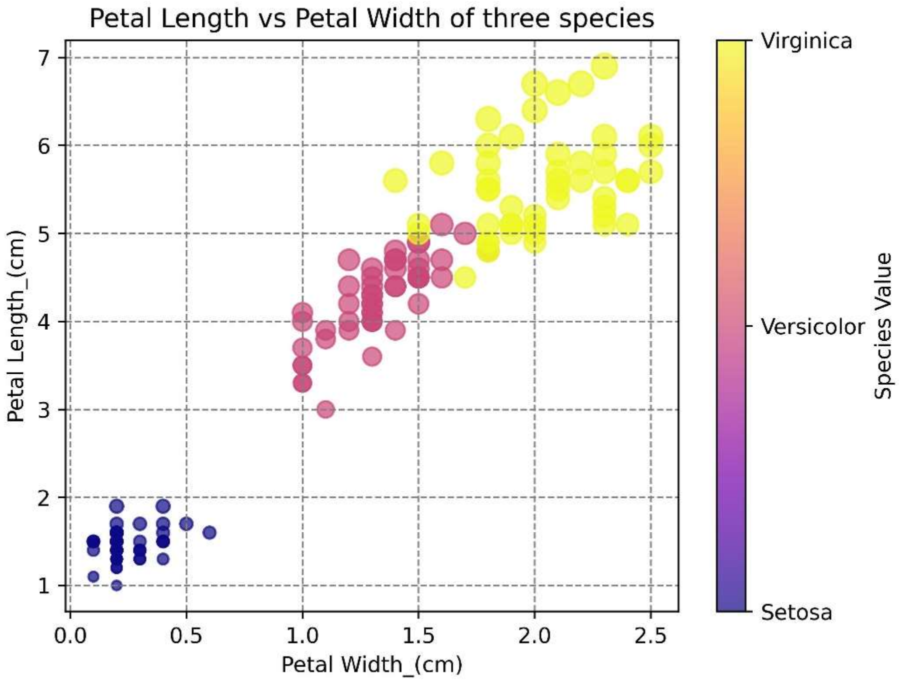


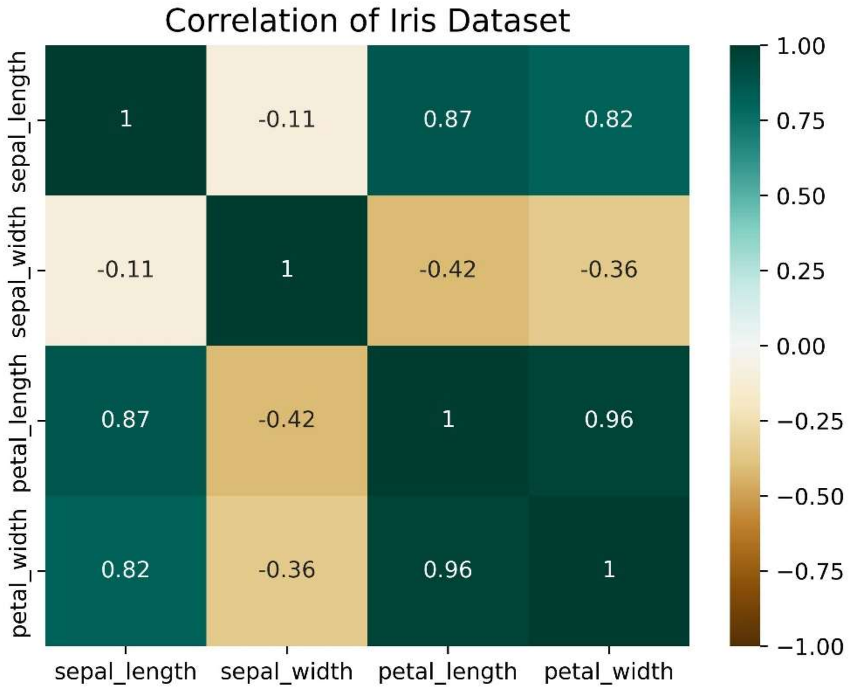
Fig 1 analyses how the sepal length varies with species. From the plot, it is shown that species Setosa has lower sepal lengths as compared to others. The Histogram for Setosa is very mild right skew (0.116) near to zero with kurtosis of -0.346 displays a normal distribution that is cluster around 4.5 to 5.5 with lighter tail. Where versicolor is positively skew (0.102) almost symmetrical and is flatter (as kurtosis is -0.599). From the histogram of virginica, it is clear that generally this species has longer sepal length. The distribution is more like right skewed (0.116) as setosa and kurtosis is nearly zero -0.088 resembling a normal distribution more closely. So, the histogram suggests that sepal lengths for Setosa and Versicolor are more spread out, while Virginica is closer to a normal shape.

Fig 2: Relation Between Petal length and Width



For second plot, it explains the relation between petal length along x axis and petal width along y axis of three species by the use of scatter plot. Each point on the plot represents an individual flower specimen, providing understanding into how these two features correlate across species. The size of each point explains the petal length. As petal width increases, so does petal length. So, it shows a positive linear relation between the length and width of the species petal. Points of setosa is separated from other two species which explain its unique feature. There is a slight overlap between versicolor and virginica that suggest the similarity between them. But it also can be differentiated by their petal measurement.

Fig 3: Correlation Analysis of Four Features of Iris Dataset



The heatmap is generated by the correlation function of iris dataset which shows a visual presentation of how the various features of iris dataset are related to each other. Each square box of heatmap contains the value of correlation coefficients between pairs of features in the corresponding columns and rows. The correlation coefficient which ranges from -1 to 1, gives the idea of linearity among the measurements. If the coefficient is less than 0, it indicates negative relation where as greater than 0 shows a positive relation and coefficient value zero means there are no relation between the features. From the heatmap of Iris dataset, it is clear that petal length and petal width have a strong correlation as the coefficient value is **0.96** which depicts as petal length increases, petal width also increases linearly. In addition, with, sepal length, petal width and sepal length, petal length pairs also have positive correlation. Surprisingly, the pair of Sepal length and sepal width is nearly uncorrelated as its coefficient value is close to 0 (-0.11). Among the remaining pairs of features, there is a negative correlation in the pair of sepal width, petal length and sepal width, petal width. This is how Heatmap gives a deep understanding of visualisation of the relationship among the measurements of Iris dataset.

Basic Statistics of Main Variables of Dataset

The average of sepal lengths, widths, petal lengths, and widths are 5.84, 3.05, 3.76, and 1.20 cm, respectively. Medians are slightly lower at 5.8, 3.0, 4.35, and 1.3 cm. Sepal features have lower variability (standard deviations of 0.83 and 0.43) compared to petal length and width (1.76 and 0.76), indicating that petal lengths and widths vary more among these species.

Conclusion The report analysed the different measurements of Iris flower and visualized the dataset through Histogram, Scatterplot and Heatmap which helped to get an insight of the relationships and distribution of different features of iris flower. Overall, it can be concluded that petal measurements are best for classifying the species. The sample of 150 iris flowers of three different species takes out the clear observation of the unique nature and relations of each species through this report of visualisation.