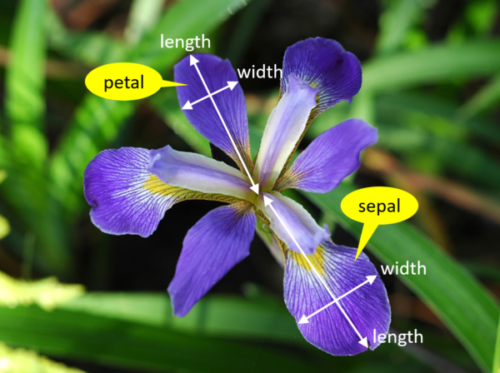
Statistical analysis of the “Iris” dataset

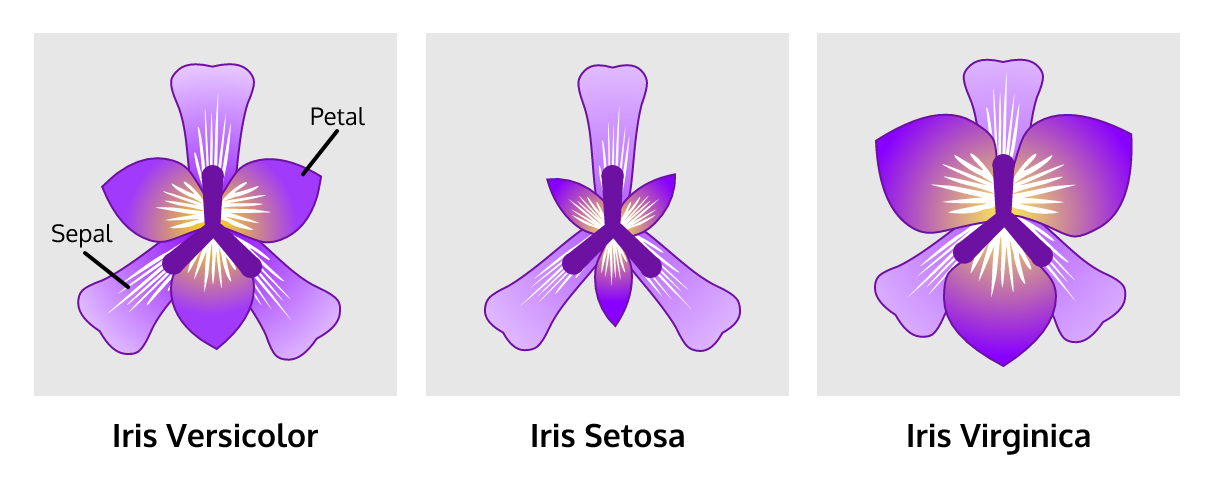
Dorota Bjöörn, AI22

# Introduction

The “Iris” dataset used in the current report was originally collected by Ronald Fisher in 1936. By measuring width and length of sepals and petals, respectively (*Figure 1)*, Fisher created a mathematical model distinguishing between three types of iris flowers: vercicolor, setosa and virginica (*Figure 2*). His findings were published the same year: Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936). [*https://en.wikipedia.org/wiki/Iris\_flower\_data\_set*]



*Figure 1. A photograph of an iris flower indicating the four measurements gathered in the iris dataset: petal length, petal width, sepal length and sepal width.[https://www.integratedots.com/determine-number-of-iris-species-with-k-means/]*



*Figure 2. Images of the three iris species described in the iris dataset: versicolor, setosa and virginica.*

*[https://www.codecademy.com/courses/machine-learning/lessons/machine-learning-clustering/exercises/iris-dataset]*

In the current report data is analysed with SciPy and Statsmodels with focus on descriptive analysis, confidence intervals, hypothesis testing, correlation and linear regression.

# Methods

Dataset was used as provided [https://www.kaggle.com/datasets/arshid/iris-flower-dataset].

Analysis of the data was performed with NumpPy, Pandas, SciPy, Statsmodels. Graphs were generated with Seaborn and Matplotlib.

# Results and discussion

In the following a simple analysis of the dataset is reported. Firstly, the dataset is described and adjusted for further analysis. Secondly the following questions are answered using basic statistical analysis in Python:

Q1 Construct a 95% confidence interval for means for sepal width for each species. Is there a significant difference between sample means for sepal width for iris species?

Q2 Is sepal length for virginica different from sepal length for versicolor?

Q3 Visualize data for pair-wise dependencies.

Q4 Describe dependency between petal length and petal width for setosa and compare with versicolor

All measurements are in cm. Only results are presented and discussed. The underlying code can be found in Appendix.

## Dataset

Dataset consisted of 5 series: sepal length, sepal width, petal length, petal width and class. There were 150 entries; 50 for each of the iris species: setosa, versicolour and virginica. There were no missing values or Nan-values. Numerical values in dataset were already provided as float, while ‘class’ series classifying each entry into a species was an object.

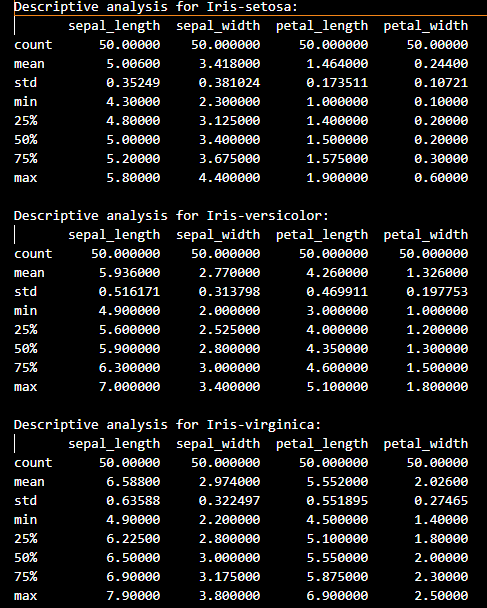
## Descriptive analysis

Descriptive analysis of the dataset summarizing each series showing for example mean, standard deviation (std) and median (50%) is shown in *Table 1*. Values given are for all species pooled together.



*Table 1. Descriptive analyses of the dataset with all species pooled together.*

Of more interest is descriptive analysis of each iris species separately as shown in Tables 2. Results show lower measures of spread as expected.



*Table 2. Descriptive analyses of each species separately.*

Q1. Construct a 95% confidence interval for means for sepal width for each species

To construct confidence intervals, the t-distribution is used since sigma of the population is not known. The following confidence intervals were obtained at 95% level:

* Iris-setosa: 3.31-3.53
* Iris-versicolor: 2.68-2.86
* Iris-virginica: 2.88-3.07

None of the confidence intervals is overlapping thus mean values for all three species are significantly different.

## Q2 Is sepal length for Iris-virginica different from sepal length for Iris-versicolor?

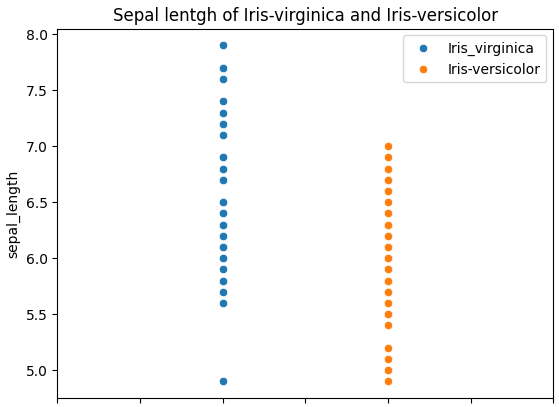
Variances for the two sample groups are not equal since F statistic (= 1.518) is larger than F critical for 49 and 49 degrees of freedom (= 0.622). H0 can thus be discarded.

Following hypothesis is to be tested on the two means:

H0: µ\_virginica = µ\_versicolor

Ha: µ\_virginica ≠ µ\_versicolor

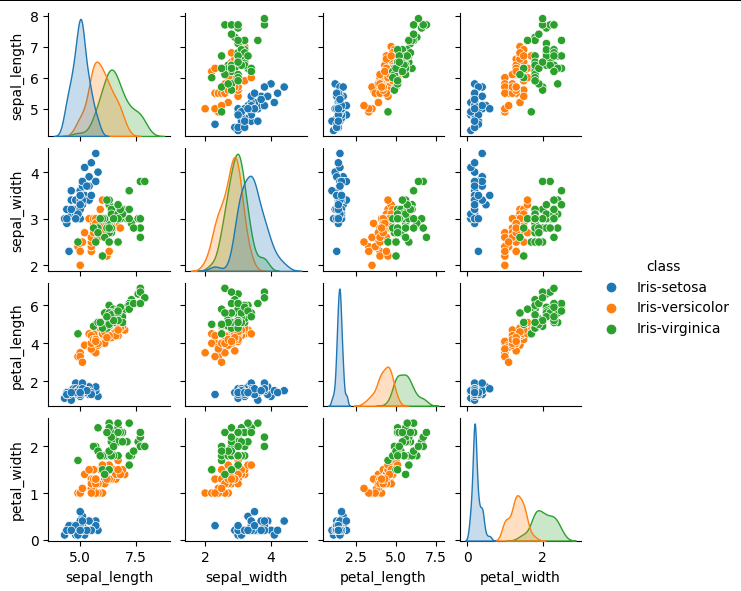
p-value for the t-test is 1.866e-07, which is much smaller than alpha (0.05). Thus H0 hypothesis can be discarded concluding that the sepal lengths for virginica and versicolor are different. A plot of sepal length data points is shown in Figure 3, where visually sepal length of virginica is larger/different than versicolor, supporting the hypothesis test result.



*Figure 3. Sepal length of Iris-virginica and Iris-vesicolor.*

## Q3 Visualize data for pair-wise dependencies.

Visualization of all series pairwise (Figure 4) with species indicated shows that iris setosa (blue) can be distinguished from the other species as the blue circles are clearly isolated from green and orange for all combinations.

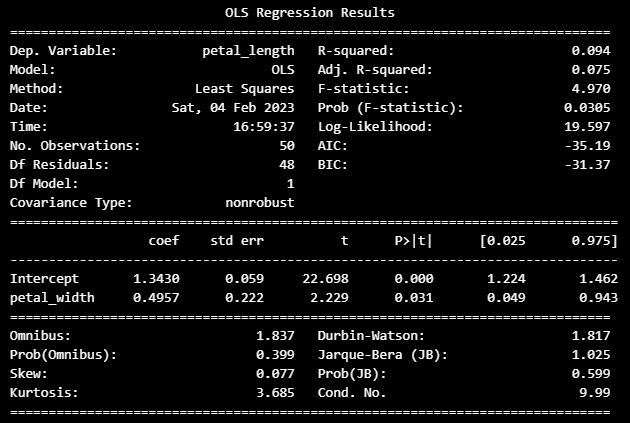


*Figure 4. Pairwise dependance plots for the iris dataset.*

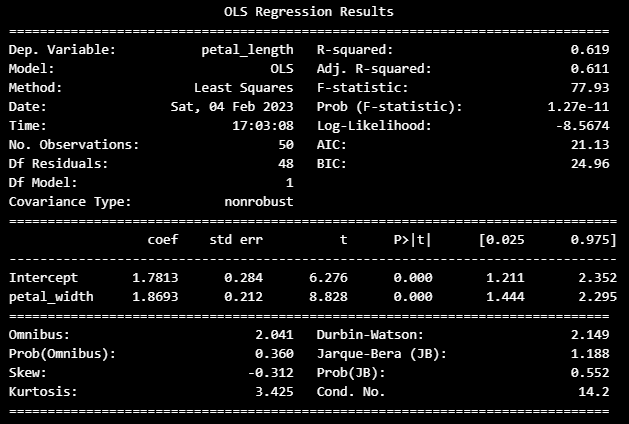
## Q4: Describe dependency between petal length and petal width for setosa and compare with versicolor.

Ordinary least squares (OLS) regression of petal length and width are shown in Table 3 and Table 4. Correlation is very week for setosa with R2 = 0.094 and stronger for versicolor with R2 = 0.619.

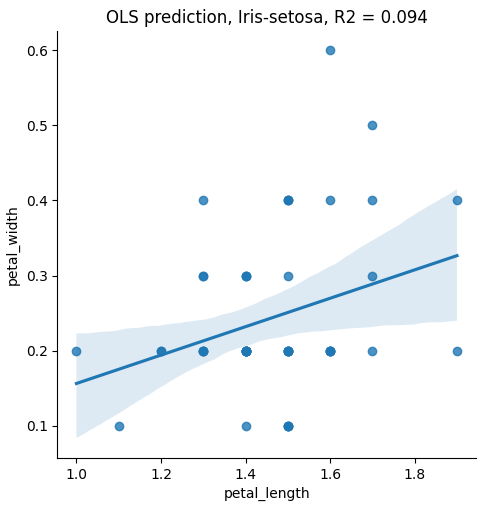
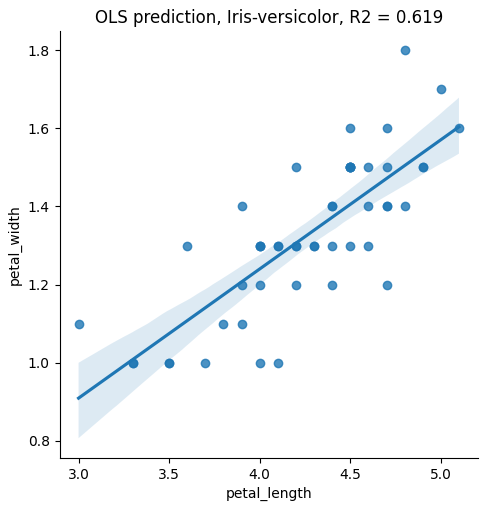
Same trends can be seen in graphs in Figure 5, where data is much more scattered with broad confidence intervals for setosa compared to versicolor. Both linear regression models approach 0 for petal width when petal length approaches 0, which is expected.



*Table 3 Results from OLS regression of petal length and width for setosa.*



*Table 4. Results from OLS regression of petal length and width for versicolor.*

*Figure 5. OLS predictions for correlation of petal width and length for setosa (left) and versicolor (right)*

# Conclusions

Scipy and Statsmodels have together with Pandas and Numpy been a powerful and simple tool for data analysis of Iris dataset.

# Appendix