CORNELL TECH

CS5785 - Homework 0

Authors
Zen Yui, Will Davis

Supervisor
Serge Belongie

August 29, 2017

Abstract

Setting up the environment in which one codes in is one of the most important steps toward building a program. In this paper, we practice our python skills by gathering and plotting various Iris sample characteristics, while also learning how to set up an effective coding environment.

1 Introduction

The task of homework 0 was to set up and familiarize ourselves with python and our machine learning working environment. As the assignment suggested, we downloaded the Anaconda python distribution and began working in the Jupiter Notebook. In order to "sanity test" our build environment, we used the time-honored tradition of Edgar Anderson's Iris Flower data set.

2 Methodology

We conducted this work in Anaconda Python 3.6 using a Jupyter Notebook[1]. We used the Requests[2] package to obtain the dataset and the standard os[3] package to safely write it to disk. We parsed and explored the dataset using the DataFrame[4] class from the Pandas library - specifically employing the `read_csv` classmethod to parse the tab-delimited dataset and `value_counts()` to find the distribution of labels. Using the `combinations()` function from the itertools[5] package, we generated the distinct combinations of features, which we vizualized as scatterplots using Matplotlib[6]. The full Jupyter Notebook is available in addition to this report in the file `hw0.ipynb`.

3 Observations

How many features/attributes are there per sample?

Anderson measured 4 features/attributes for his Iris samples.

- Sepal length in cm
- Sepal width in cm
- Petal length in cm
- Petal width in cm

How many different species are there?

Anderson found 3 different species in his data set:

- Iris Setosa
- Iris Versicolour
- Iris Virginica

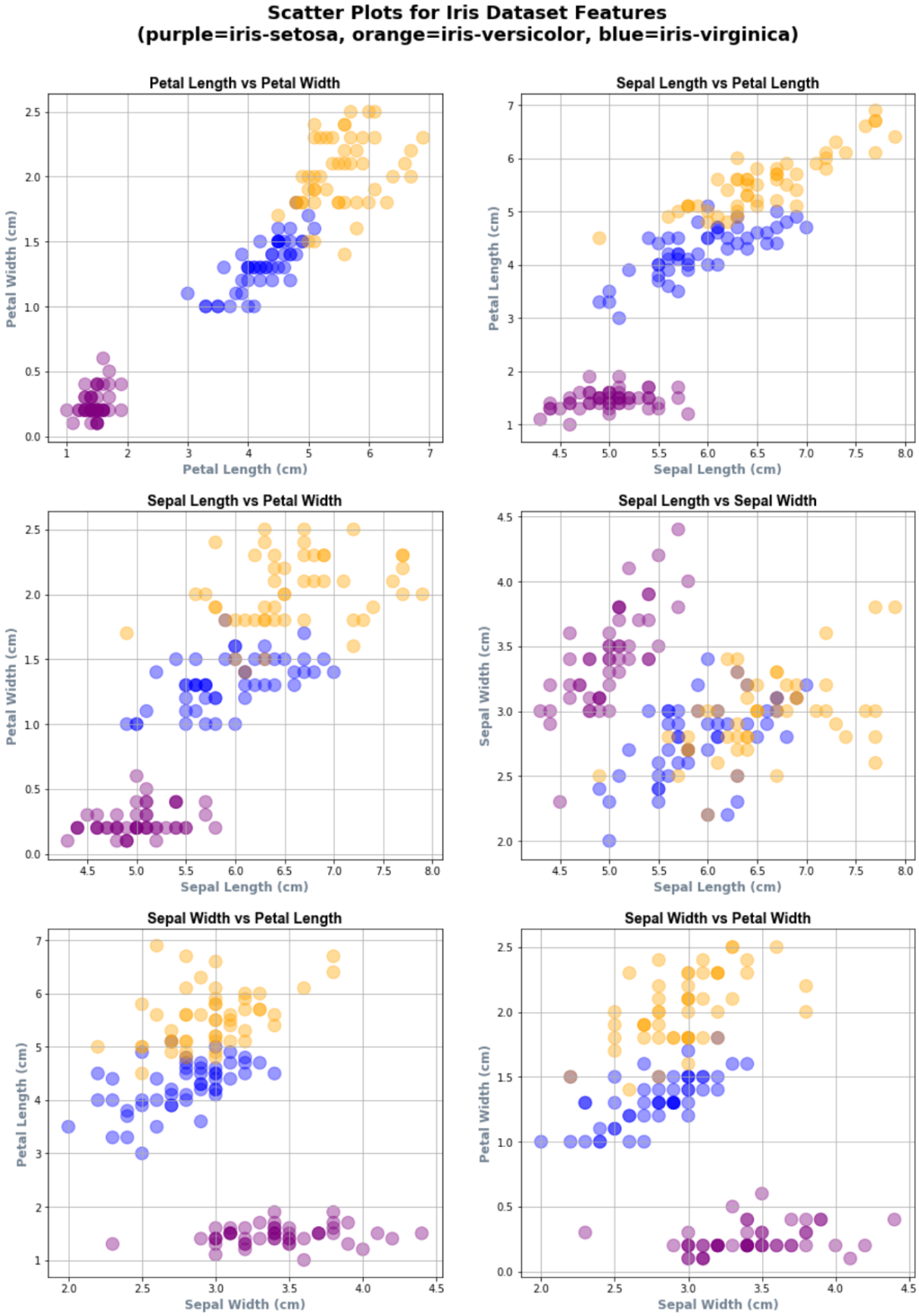
How many samples of each species did Anderson record?

Anderson collected a total of 150 samples, divided equally between each of the species.

- Iris Setosa: 50 samples
- Iris Versicolour: 50 samples
- Iris Virginica: 50 samples

[1] https://www.anaconda.com/what-is-anaconda/
[2] http://docs.python-requests.org/en/master/
[3] https://docs.python.org/3/library/os.html
[4] https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html
[5] https://docs.python.org/3/library/itertools.html#itertools.combinations
[6] https://matplotlib.org/

4 Plots


Scatter Plots for Iris Dataset Features
(purple=iris-setosa, orange=iris-versicolor, blue=iris-virginica)
Petal Length vs Petal Width
Petal Width (cm)
Petal Length (cm)
Sepal Length vs Petal Length
Petal Length (cm)
Sepal Length (cm)
Sepal Length vs Petal Width
Petal Width (cm)
Sepal Length (cm)
Sepal Length vs Sepal Width
Sepal Width (cm)
Sepal Length (cm)
Sepal Width vs Petal Length
Petal Length (cm)
Sepal Width (cm)
Sepal Width vs Petal Width
Petal Width (cm)
Sepal Width (cm)