МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ   
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ОРЛОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИМЕНИ И.С. ТУРГЕНЕВА»

Кафедра информационных систем

**ЛАБОРАТОРНАЯ РАБОТА №1**

по дисциплине «Основы анализа данных на языке Python»

на тему: «Факторный анализ»

Студент: Первых Е.Е. Шифр 195153

Васильев М.Е. Шифр 195137

Бологов Д.Е. Шифр 195136

Институт приборостроения, автоматизации и информационных технологий

Направление подготовки 09.04.03 Прикладная информатика

Направленность (профиль) Технологии сбора и анализа больших данных

Группа 91-ПИ-м

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Демидов А.В.

Оценка: «\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_» Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Орел 2020

## 1 Цель занятия

Реализовать базовые операции над набором данных на языке программирования Python с применением библиотек NumPy, Pandas и Matplotlib и проанализировать значения рассчитанных показателей.

## 2 Решение

Подключим все библиотеки, которые нам могут пригодиться (рисунок 1).

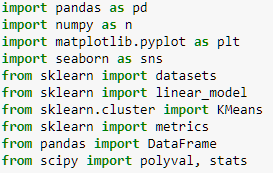


Рисунок 1 – Подключаемые библиотеки

Загрузим данные «Ирисы Фишера», собранные Эдгаром Андерсоном, для удобства манипулирования ими, сделаем из них датафрейм и выведем их на экран (рисунок 2).

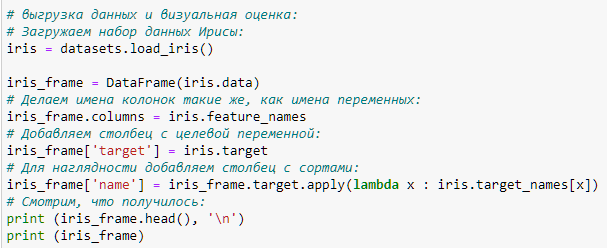


Рисунок 2 – Код для загрузки и вывода на экран собранных данных

Полученный датафрейм достаточно наглядный (рисунок 3).

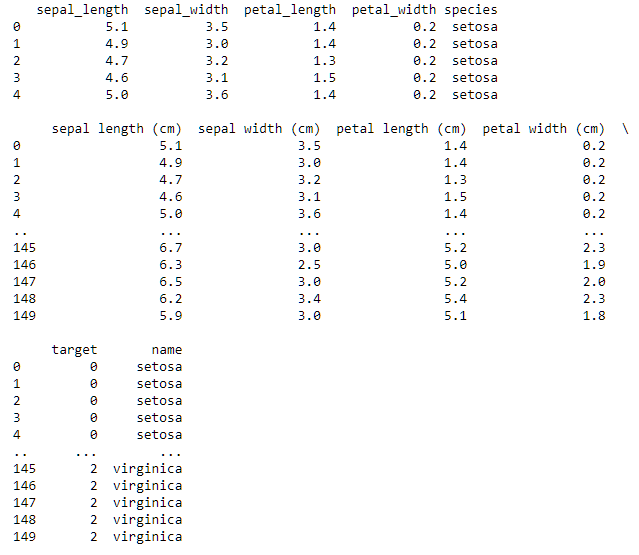


Рисунок 3 – Данные, выведенные на экран

Для большей наглядности, требуется отобразить статистику на экране, при помощи этой команды (Рисунок 4).



Рисунок 4 – Команда для отображения статистики

Как результат – на экране появилось много графиков с зависимостями между признаками (рисунок 5).

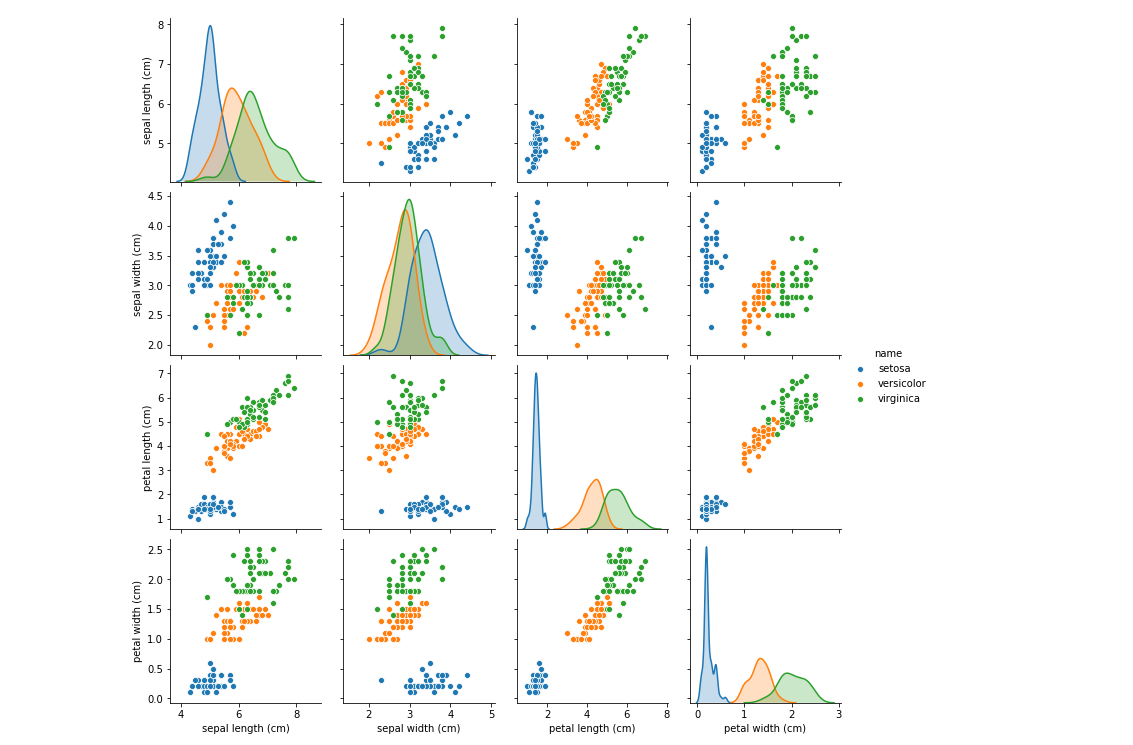


Рисунок 5 – Графики зависимостей между признаками

На данном этапе можно предположить, что «petal width (cm)» и «petal length (cm)» имеют сильную зависимость - точки вытянуты вдоль одной линии. Но предположений в визуальной оценке нам мало, поэтому требуется построить таблицу кореляции, используя данную команду (рисунок 6).



Рисунок 6 – Команда для отображения таблицы кореляции

В результате получаем вот такую наглядную таблицу (рисунок 7).

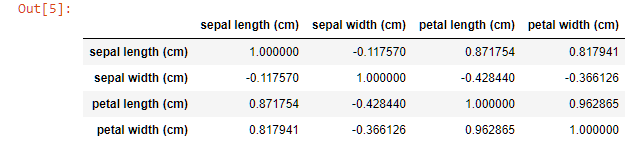


Рисунок 7 – Таблица кореляции

Можно обратить внимание, что между переменными «petal length (cm)» и «petal width (cm)» выявлена очень сильная зависимость 0.96.

Уже на основе данных из таблицы, можно построить линейную регрессию, используя эти команды (рисунок 8).

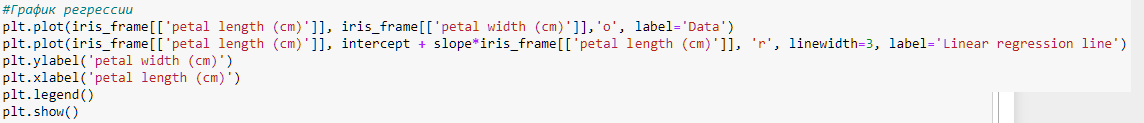


Рисунок 8 – Команды для построения линейной регрессии

После этих команд, на экране мы увидим такой график (рисунок 9).

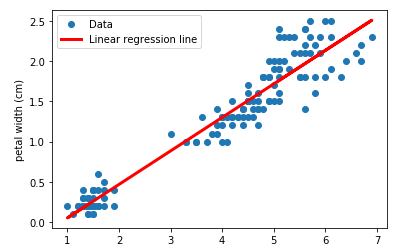


Рисунок 9 – График линейной регрессии

В итоге, теперь, мы сможем с большой точностью определить, какая у листочка pental ширина, зная его длину.

## 3 Вывод

На основе представленной обучающей выборки, могут быть решены задачи, целью которых является нахождение длины или ширины листочка определенного типа, если тип и длина или ширина данного листочка известны заранее.

## 4 Контрольные вопросы

**1. Какие программные библиотеки языка Python используют для решения задач обработки и анализа данных? Для чего предназначена каждая из перечисленных программных библиотек?**

**NumPy** – это библиотека языка Python, добавляющая поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых математических функций для операций с этими массивами.

**Matplotlib** — это библиотека Python для построения качественных двумерных графиков. Matplotlib является гибким, легко конфигурируемым пакетом.

**Pandas** – это библиотека Python, которая является мощным инструментом для анализа данных. Пакет дает возможность строить сводные таблицы, выполнять группировки, предоставляет удобный доступ к табличным данным, а при наличии пакета matplotlib дает возможность рисовать графики на полученных наборах данных.

**Seaborn** – это библиотека визуализации на Python. Он построен на вершине Matplotlib. Позволяет работать с датафреймами. Является дополнением и расширением Matplotlib.

Листинг программы  
  
import pandas as pd

import numpy as n

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import datasets

from sklearn import linear\_model

from sklearn.cluster import KMeans

from sklearn import metrics

from pandas import DataFrame

from scipy import polyval, stats

# выгрузка данных и визуальная оценка:

# Загружаем набор данных Ирисы:

iris = datasets.load\_iris()

iris\_frame = DataFrame(iris.data)

# Делаем имена колонок такие же, как имена переменных:

iris\_frame.columns = iris.feature\_names

# Добавляем столбец с целевой переменной:

iris\_frame['target'] = iris.target

# Добавляем столбец с сортами:

iris\_frame['name'] = iris\_frame.target.apply(lambda x : iris.target\_names[x])

# Результат:

print (iris\_frame.head(), '\n')

print (iris\_frame)

# Графики

sns.pairplot(iris\_frame[['sepal length (cm)','sepal width (cm)','petal length (cm)','petal width (cm)','name']], hue = 'name')

#Таблица кореляции

iris\_frame[['sepal length (cm)','sepal width (cm)','petal length (cm)','petal width (cm)']].corr()

#График регрессии

fit\_output = stats.linregress(iris\_frame[['petal length (cm)','petal width (cm)']])

slope, intercept, r\_value, p\_value, slope\_std\_error = fit\_output

print(slope, intercept, r\_value, p\_value, slope\_std\_error)

plt.plot(iris\_frame[['petal length (cm)']], iris\_frame[['petal width (cm)']],'o', label='Data')

plt.plot(iris\_frame[['petal length (cm)']], intercept + slope\*iris\_frame[['petal length (cm)']], 'r', linewidth=3, label='Linear regression line')

plt.ylabel('petal width (cm)')

plt.xlabel('petal length (cm)')

plt.legend()

plt.show()