Лабораторная работа №1. Основы анализа данных на языке Python.

Используемый набор данных: Iris (https://archive.ics.uci.edu/ml/datasets/Iris)

In [1]:

```
# Импорт необходимых библиотек
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px

# графики встроены в ноутбук
%matplotlib inline

pd.options.display.max_columns = None
```

Загрузим набор данных средствами библиотеки *Pandas* и выведем загруженные данные на экран. В последней колонке (*Class*) содержатся категориальные данные, поэтому выполним преобразование типа.

In [2]:

```
headers = ["Sepal length", "Sepal width", "Petal length", "Petal width", "Class"]
data = pd.read_csv("dataset/iris.data", names=headers, index_col=None)
data["Class"] = data["Class"].astype("category")
# 40 случайных строк из набора
data.sample(40)
```

Out[2]:

	Sepal length	Sepal width	Petal length	Petal width	Class
57	4.9	2.4	3.3	1.0	Iris-versicolor
132	6.4	2.8	5.6	2.2	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
88	5.6	3.0	4.1	1.3	Iris-versicolor
17	5.1	3.5	1.4	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
144	6.7	3.3	5.7	2.5	Iris-virginica
103	6.3	2.9	5.6	1.8	Iris-virginica
7	5.0	3.4	1.5	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
107	7.3	2.9	6.3	1.8	Iris-virginica
16	5.4	3.9	1.3	0.4	Iris-setosa
37	4.9	3.1	1.5	0.1	Iris-setosa
120	6.9	3.2	5.7	2.3	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
26	5.0	3.4	1.6	0.4	Iris-setosa
106	4.9	2.5	4.5	1.7	Iris-virginica
49	5.0	3.3	1.4	0.2	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
133	6.3	2.8	5.1	1.5	Iris-virginica
41	4.5	2.3	1.3	0.3	Iris-setosa
81	5.5	2.4	3.7	1.0	Iris-versicolor
122	7.7	2.8	6.7	2.0	Iris-virginica
86	6.7	3.1	4.7	1.5	Iris-versicolor
8	4.4	2.9	1.4	0.2	Iris-setosa
64	5.6	2.9	3.6	1.3	Iris-versicolor
99	5.7	2.8	4.1	1.3	Iris-versicolor
137	6.4	3.1	5.5	1.8	Iris-virginica
22	4.6	3.6	1.0	0.2	Iris-setosa
134	6.1	2.6	5.6	1.4	Iris-virginica
75	6.6	3.0	4.4	1.4	Iris-versicolor
117	7.7	3.8	6.7	2.2	Iris-virginica
83	6.0	2.7	5.1	1.6	Iris-versicolor

	Sepal length	Sepal width	Petal length	Petal width	Class
39	5.1	3.4	1.5	0.2	Iris-setosa
25	5.0	3.0	1.6	0.2	Iris-setosa
105	7.6	3.0	6.6	2.1	Iris-virginica

Типы значений набора данных:

In [3]:

```
data.dtypes
```

Out[3]:

```
Sepal length     float64
Sepal width      float64
Petal length     float64
Petal width      float64
Class           category
dtype: object
```

Проверим набор на наличие пустых ячеек.

In [4]:

```
data.isna().sum()
```

Out[4]:

```
Sepal length    0
Sepal width     0
Petal length    0
Petal width     0
Class           0
dtype: int64
```

Пустых ячеек нет.

Визуально оценим статистические данные по числовым признакам.

In [5]:

```
data.describe()
```

Out[5]:

	Sepal length	Sepal width	Petal length	Petal width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

И по категориальным.

In [6]:

```
data.describe(include="category")
```

Out[6]:

	Class
count	150
unique	3
top	Iris-virginica
freq	50

Построим таблицу корреляции между признаками и диаграммы рассеяния для визуализации корреляции между признаками.

In [7]:

```
data.corr()
```

Out[7]:

	Sepal length	Sepal width	Petal length	Petal width
Sepal length	1.000000	-0.109369	0.871754	0.817954
Sepal width	-0.109369	1.000000	-0.420516	-0.356544
Petal length	0.871754	-0.420516	1.000000	0.962757
Petal width	0.817954	-0.356544	0.962757	1.000000

In [8]:

```
plots = px.scatter_matrix(data, dimensions=headers, color="Class")
plots.show()
```