Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчёт по рубежному контролю №1

По дисциплине:
«Технологии машинного обучения»

Выполнил:

Студент группы ИУ5ц-82Б _____ **Акимкин М.Г.**

(Подпись, дата) (Фамилия И.О.)

Проверил:

_____ **Гапанюк Ю. Е.**

(Подпись, дата) (Фамилия И.О.)

Москва, 2021

Задание

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему?
Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

Для студентов групп ИУ5-62Б, ИУ5Ц-82Б - для произвольной колонки данных построить гистограмму.

Набор данных:

https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris

Акимкин М.Г., РК№1, ИУ5ц-82Б, вариант №26

Задание: для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

Датасет: https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris

Импорт библиотек

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
from sklearn import datasets
from sklearn.datasets import load_iris
from sklearn import linear_model
from sklearn.cluster import KMeans
from sklearn import metrics
from pandas import DataFrame
%pylab inline
```

```
Populating the interactive namespace from numpy and matplotlib
```

In [2]:
```
boston = load_iris()
data = pd.DataFrame(boston.data, columns=boston.feature_names)
data['TARGET'] = boston.target
```

In [3]:
```
# Первые пять строк датасета
data.head()
```

Out[3]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	TARGET
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

In [4]:
```
# Описание датасета
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
sepal length (cm)    150 non-null float64
sepal width (cm)     150 non-null float64
petal length (cm)    150 non-null float64
petal width (cm)     150 non-null float64
TARGET               150 non-null int32
dtypes: float64(4), int32(1)
memory usage: 5.4 KB
```

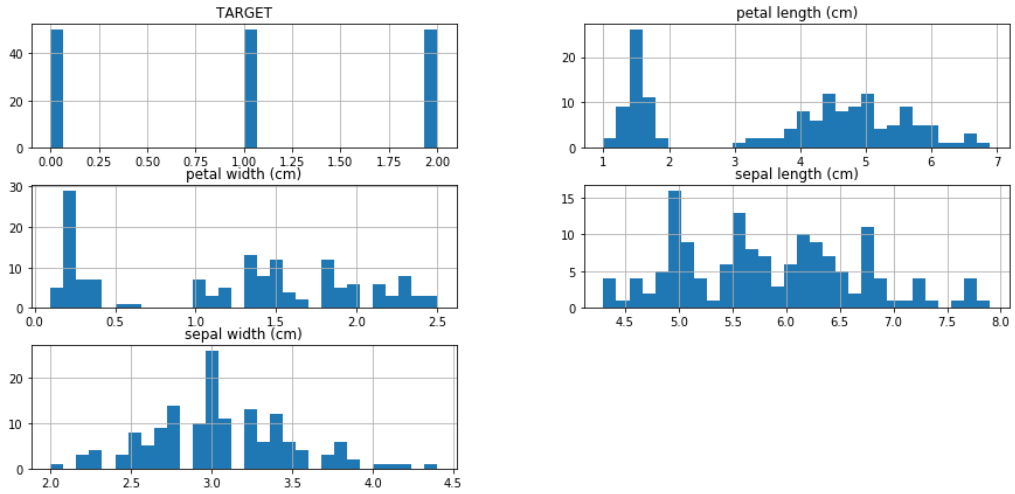
In [5]:
```
# Статистические данные
data.describe()
```

Out[5]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	TARGET
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333	1.000000
std	0.828066	0.435866	1.765298	0.762238	0.819232
min	4.300000	2.000000	1.000000	0.100000	0.000000
25%	5.100000	2.800000	1.600000	0.300000	0.000000
50%	5.800000	3.000000	4.350000	1.300000	1.000000
75%	6.400000	3.300000	5.100000	1.800000	2.000000
max	7.900000	4.400000	6.900000	2.500000	2.000000

In [6]:
```
# Гистограммы для всех признаков
data.hist(bins=30, figsize = (15,7))
```

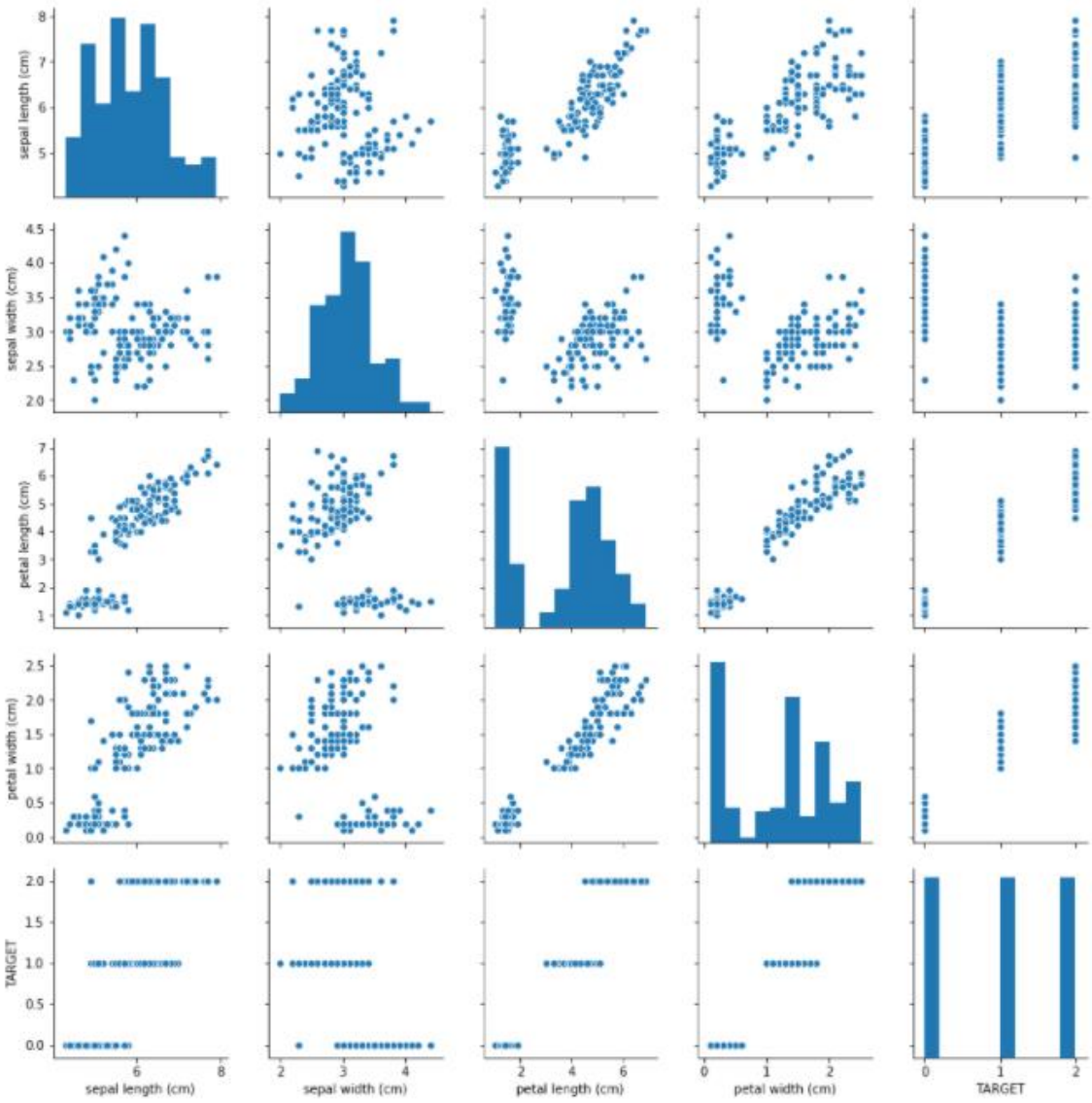
Out[6]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x0000029286ACF8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000029286C5EB00>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000029286C88D68>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000029286CB3FD0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x0000029286CE1278>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000029286D0A4E0>]],
      dtype=object)
```



In [7]:
```
# Диаграммы рассеяние для всех признаков
plt.figure(figsize=(12,6))
sns.pairplot(data)
```

Out[7]: <seaborn.axisgrid.PairGrid at 0x292814d82e8>

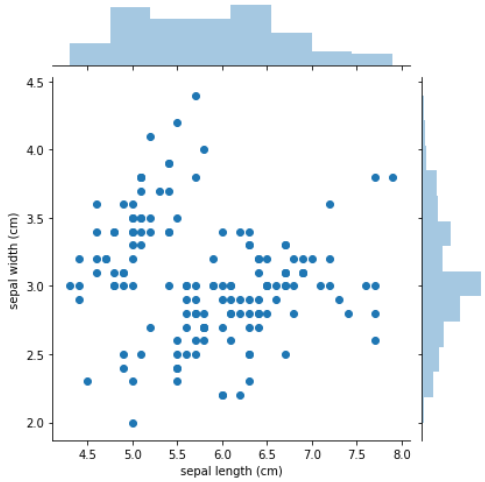
<Figure size 864x432 with 0 Axes>



In [8]:
```
# Увеличенные диаграммы рассеяния для признаков, которые имеют зависимость
sns.jointplot(x = "sepal length (cm)", y = "sepal width (cm)", kind="scatter", data = data)
```

```
C:\Users\MSI GL72 7RD\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for mult
idimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an
array index, `arr[np.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

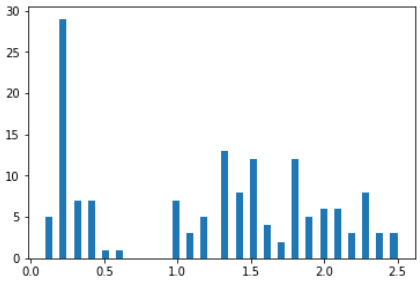
Out[8]:
```
<seaborn.axisgrid.JointGrid at 0x29287e327f0>
```



In [9]:
```
from sklearn.preprocessing import StandardScaler, MinMaxScaler, StandardScaler, Normalizer
```

In [10]:
```
sc1 = MinMaxScaler()
sc1_data = sc1.fit_transform(data[['petal width (cm)']])
```

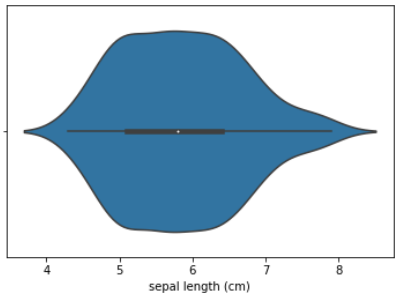
In [11]:
```
plt.hist(data['petal width (cm)'], 50)
plt.show()
```



In [12]:
```
#Скрипичная диаграмма
sns.violinplot(x=data['sepal length (cm)'])
```

```
C:\Users\MSI GL72 7RD\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for mult
idimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an
array index, `arr[np.array(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x29287c49be0>
```



In [13]:
```
# Одномерное распределение вероятности
sns.boxplot(x=data['sepal length (cm)'])
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x29289356160>
```