Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Автоматизированные системы обработки информации и управления»

«Методы машинного обучения»

Отчет по Лабораторной работе №1

Разведочный анализ данных. Исследование и визуализация данных.

Выполнила:
студентка группы ИУ5-22М
Петрова Ирина
Проверил:
доцент, к.т.н. Гапанюк Ю. Е.

Москва, 2020

Текстовое описание набора данных

Используется набор данных, использующий данные химического анализа для установления происхождения вина: https://archive.ics.uci.edu/ml/datasets/Wine (https://archive.ics.uci.edu/ml/datasets/Wine)

Эти данные являются результатами химического анализа вин, выращенных в одном регионе Италии, но полученных из трех различных сортов. В результате анализа было определено 13 компонентов, содержащихся в каждом из трех видов вин.

Датасет содержит следующие колонки:

- Алкоголь
- Яблочная кислота
- Зола
- Щелочность золы
- Магний
- Всего фенолов
- Флаваноиды
- Нефлаваноидные фенолы
- Проантоцианы
- Интенсивность цвета
- Оттенок
- OD280 / OD315 (разбавленность вина)
- Пролин

Импорт библиотек

In [2]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Загрузка данных

In [6]:

```
from sklearn.datasets import *
```

In [7]:

In [8]:

```
type(wine)
```

```
wine = load_wine()
```

Out[8]:

```
sklearn.utils.Bunch
```

In [9]:

```
# Датасет возвращается в виде словаря со следующими ключами
for x in wine:
    print(x)
```

```
data target
target_names
DESCR
feature_names
```

In [10]:

```
wine['target_names']
```

Out[10]:

```
array(['class_0', 'class_1', 'class_2'], dtype='<U7')
```

In [11]:

```
wine['feature_names']
```

Out[11]:

```
['alcohol',
 'malic_acid',
 'ash',
 'alcalinity_of_ash',
 'magnesium',
 'total_phenols',
 'flavanoids',
 'nonflavanoid_phenols',
 'proanthocyanins',
 'color_intensity',
 'hue',
 'od280/od315_of_diluted_wines',
 'proline']
```

In [12]:

```
# Размерность данных
wine['data'].shape
```

Out[12]:

(178, 13) In

[13]:

```
# Размерность целевого признака
wine['target'].shape Out[13]:
```

(178,)

In [14]:

```
data = pd.DataFrame(data= np.c_[wine['data'], wine['target']],
                    columns= wine['feature_names'] + ['target'])
data
```

Out[14]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonf
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	**...**
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	

178 rows × 14 columns

Основные характеристики датасета

In [15]:

```
# Первые 5 строк датасета
data.head()
```

Out[15]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflav
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	

In [17]:

```
# Размер датасета - 178 строк, 14 столбцов
data.shape
```

Out[17]:

```
(178, 14)
```

In [18]:

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 178 In
```

[19]:

```
# Список колонок
data.columns
```

Out[19]:

```
Index(['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
       'total_phenols', 'flavanoids', 'nonflavanoid_phenols',
       'proanthocyanins', 'color_intensity', 'hue',
'od280/od315_of_diluted_wines', 'proline', 'target'],       dtype='object')
```

In [20]:

```
# Список колонок с типами данных
data.dtypes
```

Out[20]:

```
alcohol                         float64
malic_acid                      float64
ash                             float64
alcalinity_of_ash               float64
magnesium                       float64
total_phenols                   float64
flavanoids                      float64
nonflavanoid_phenols            float64
proanthocyanins                 float64
color_intensity                 float64
hue                             float64
od280/od315_of_diluted_wines    float64
proline                         float64
target                          float64
dtype: object
```

In [21]:

```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
 temp_null_count = data[data[col].isnull()].shape[0]    print('{}
 - {}'.format(col, temp_null_count))
```

```
alcohol - 0 malic_acid -
0 ash - 0
alcalinity_of_ash - 0
magnesium - 0
total_phenols - 0
```

```
flavanoids - 0
nonflavanoid_phenols - 0
proanthocyanins - 0
color_intensity - 0 hue
- 0
od280/od315_of_diluted_wines - 0
proline - 0 target - 0
```

In [22]:

```
# Основные статистические характеристки набора данных
data.describe()
```

Out[22]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flav
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.2
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.

In [24]:

```
# Определим уникальные значения для целевого признака
data['target'].unique()
```

Out[24]:

```
array([0.,
1., 2.])
```

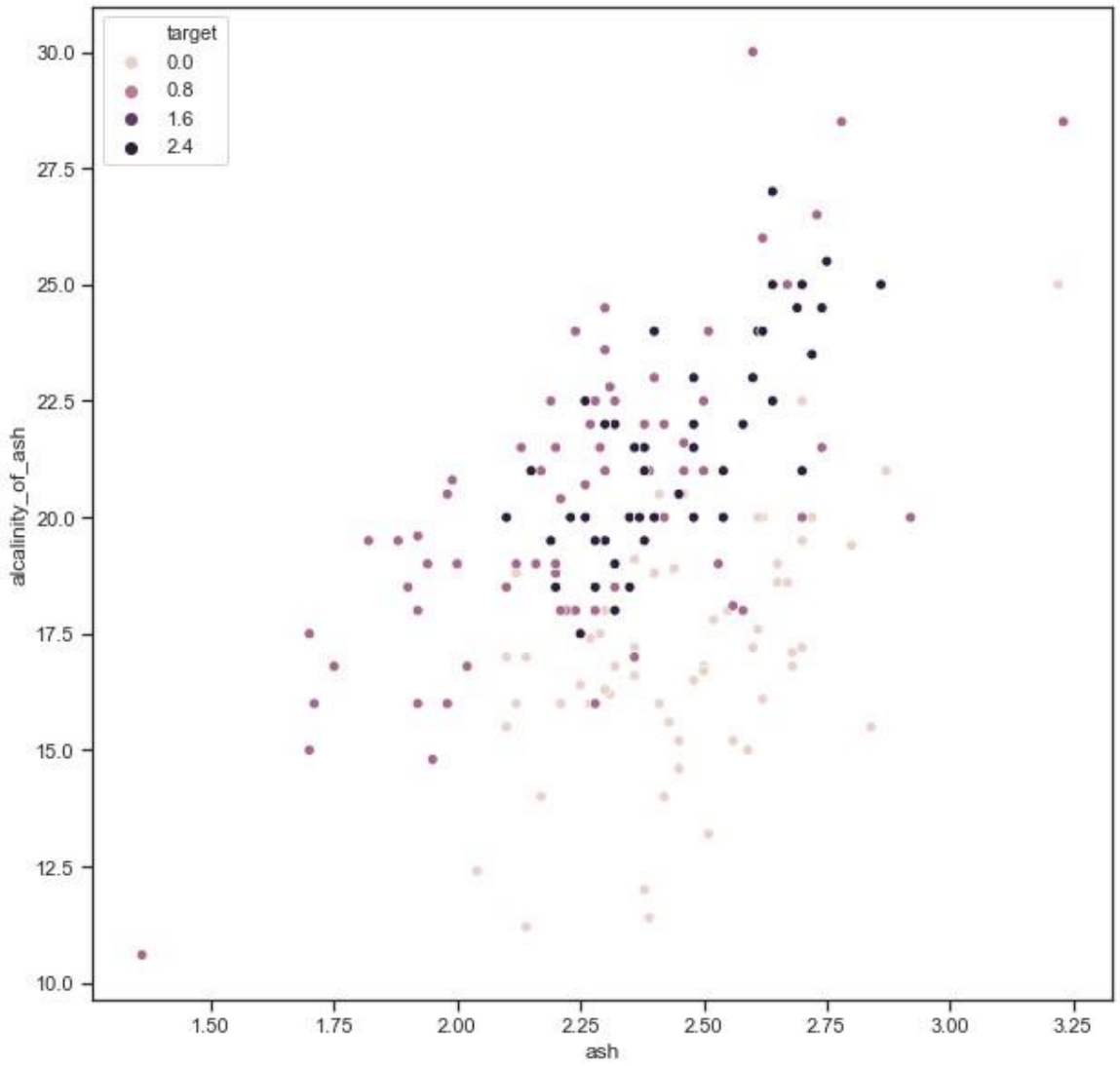
Визуальное исследование датасета

In [25]:

```
fig, ax = plt.subplots(figsize=(10,10))  sns.scatterplot(ax=ax, x='ash',
y='alcalinity_of_ash', data=data, hue='target')
```

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x267a798e6a0>
```



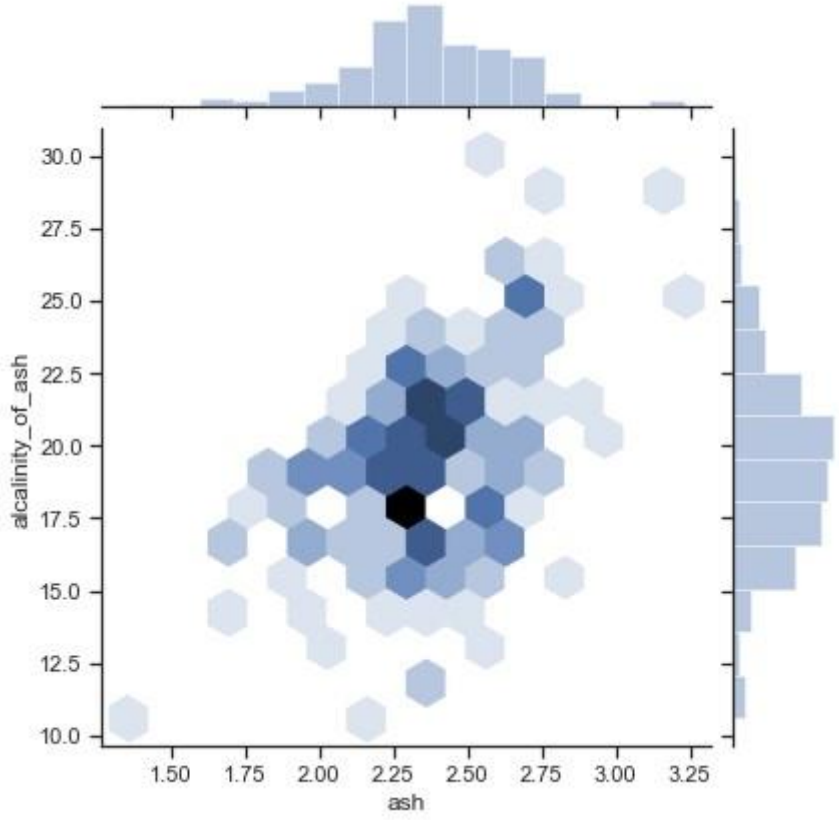

```
sns.jointplot(x='ash', y='alcalinity_of_ash', data=data, kind="hex")
```

Out[26]:

```
<seaborn.axisgrid.JointGrid at 0x26792189a90>
```

In [26]:

```
```

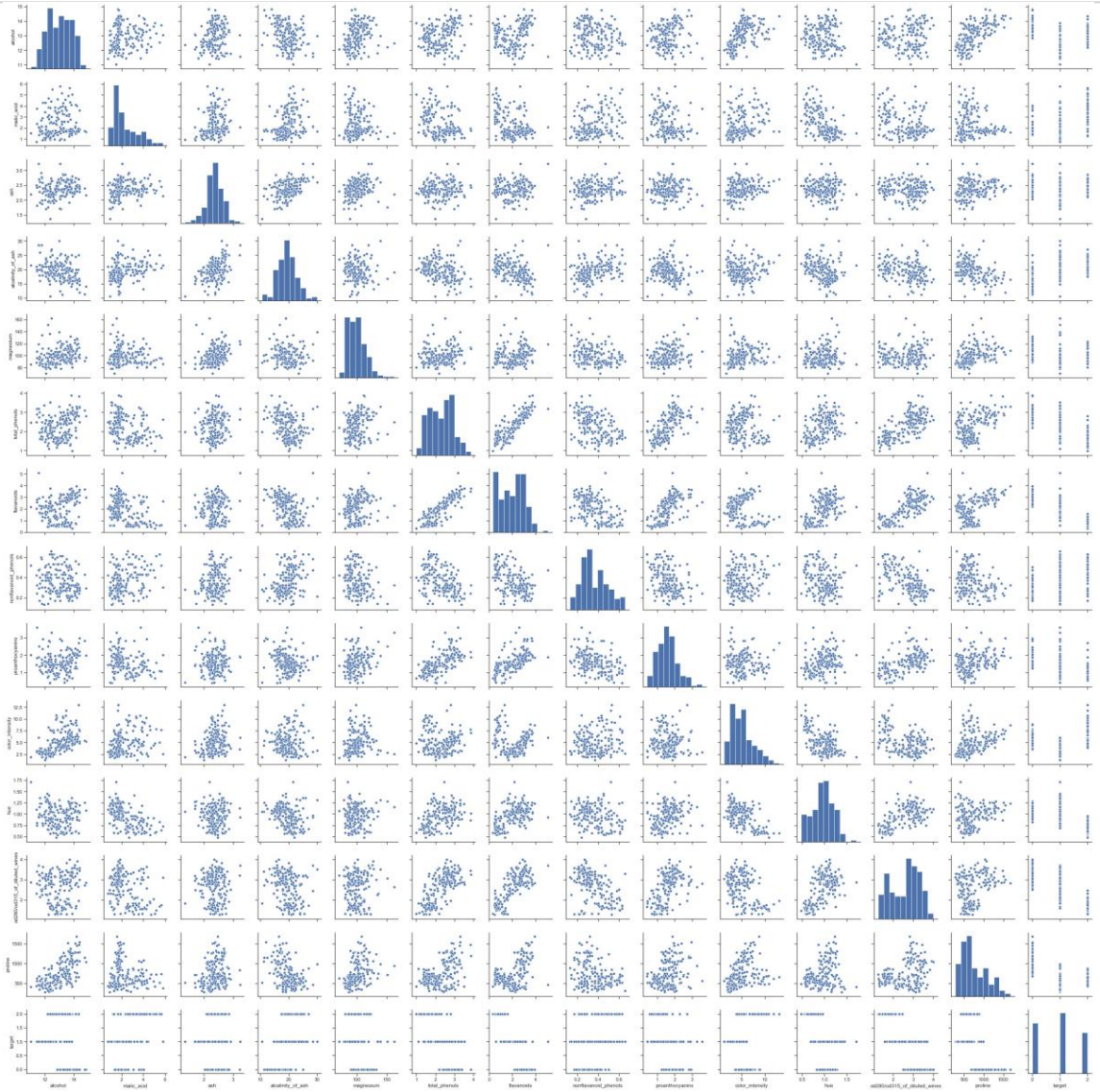



```
sns.pairplot(data)
```

Out[27]:

```
<seaborn.axisgrid.PairGrid at 0x267a9b5f5f8>
```

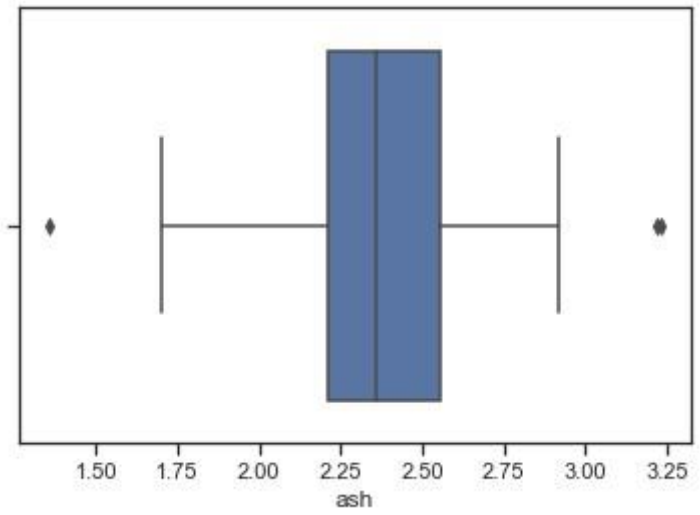
In [27]:



```
sns.boxplot(x=data['ash'])
```

Out[28]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x267b1402f60>
```

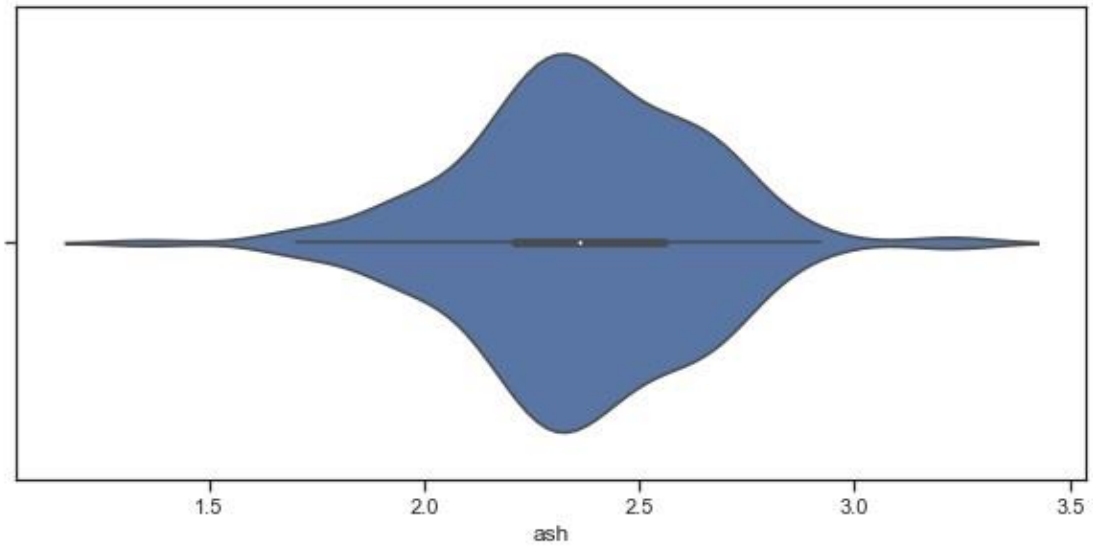


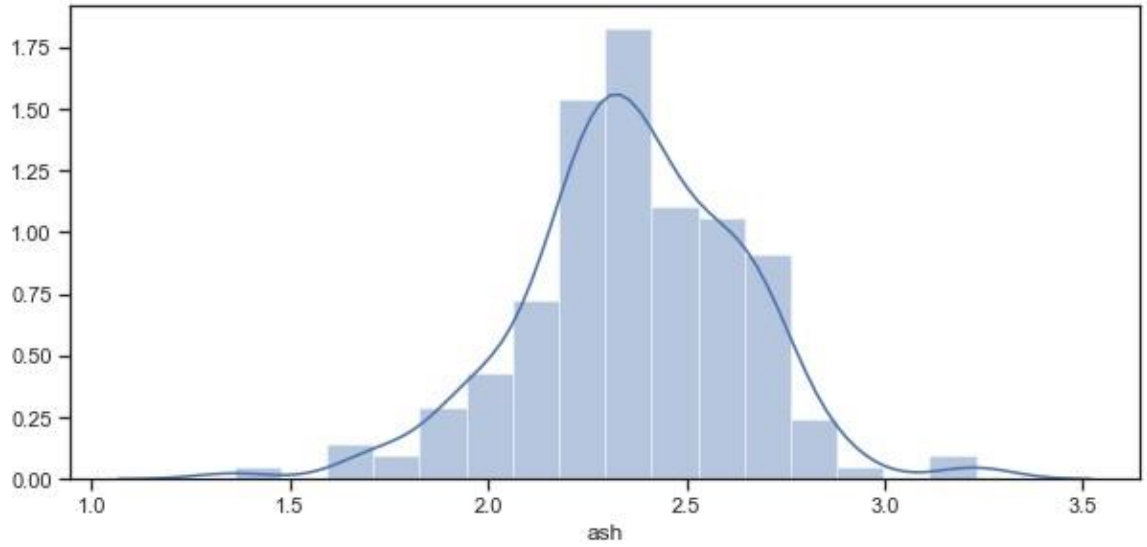
In [29]:

```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['ash'])
sns.distplot(data['ash'], ax=ax[1])
```

Out[29]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x267b146de80>
```


1.5
2.0
2.5
3.0
3.5
ash


1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
1.0
1.5
2.0
2.5
3.0
3.5
ash

Проверка корреляции признаков

In [30]:

```
data.corr()
```

Out[30]:

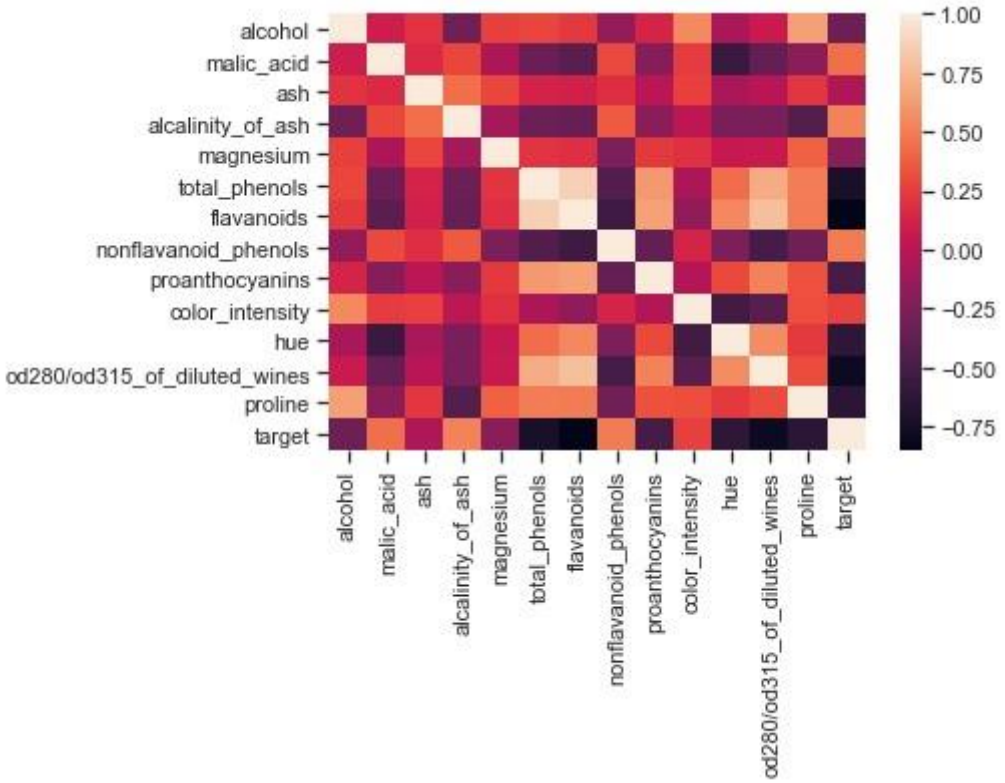
	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575
ash	0.211545	0.164045	1.000000	0.443367	0.286587
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179

In [34]:

```
sns.heatmap(data.corr())
```

Out[34]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x267b434bc50>
```


alcohol
malic_acid
ash
alcalinity_of_ash
magnesium
total_phenols
flavanoids
nonflavanoid_phenols
proanthocyanins
color_intensity
hue
od280/od315_of_diluted_wines
proline
target
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75