Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему
«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-21М
Андреев К.А.

Москва — 2020 г.

Цель лабораторной работы

Изучить различные методы визуализации данных.

Задание

Требуется выполнить следующие действия:

- Выбрать набор данных (датасет).
- Создать ноутбук, который содержит следующие разделы:

> 1. Текстовое описание выбранного набора данных.
> 2. Основные характеристики датасета.
> 3. Визуальное исследование датасета.
> 4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на GitHub.

Текстовое описание набора данных

Wine recognition dataset

Data Set Characteristics:

Number of Instances	178 (50 in each of three classes)
Number of Attributes	13 numeric, predictive attributes and the class
Attribute Information	Alcohol, , Malic acid, , Ash, , Alcalinity of ash, , Magnesium, , Total phenols, , Flavanoids, , Nonflavanoid phenols, , Proanthocyanins, , Color intensity, , Hue, , OD280/OD315 of diluted wines, , Proline
Missing Attribute Values	None
Creator	R.A. Fisher
Date	July, 1988

This is a copy of UCI ML Wine recognition datasets. https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

The data is the results of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in the three types of wine.

In [0]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
from sklearn.datasets import *
```

In [0]:
```
def make_dataframe(ds_function):
    ds = ds_function()
    df = pd.DataFrame(data= np.c_[ds['data'], ds['target']],
                      columns= list(ds['feature_names']) + ['target'])
    return df
```

In [0]:
```
data = make_dataframe(load_wine)
```

In [4]:
```
data.head()
```

Out[4]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid
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 [5]:
```
print(data.shape)
print('Всего строк: {}'.format(data.shape[0]))
```

```
(178, 14)
Всего строк: 178
```

In [6]:
```
data.columns
```

Out[6]:
```
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 [7]:
```
data.dtypes
```

Out[7]:
```
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
```

проверка пустых значений

In [8]:
```
for col in data.columns:
    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 [9]:
```
data.describe()
```

Out[9]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.00000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.02927
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.99885
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.34000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.20500
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.13500
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.87500
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.08000

In [10]:
```
data['target'].unique()
```

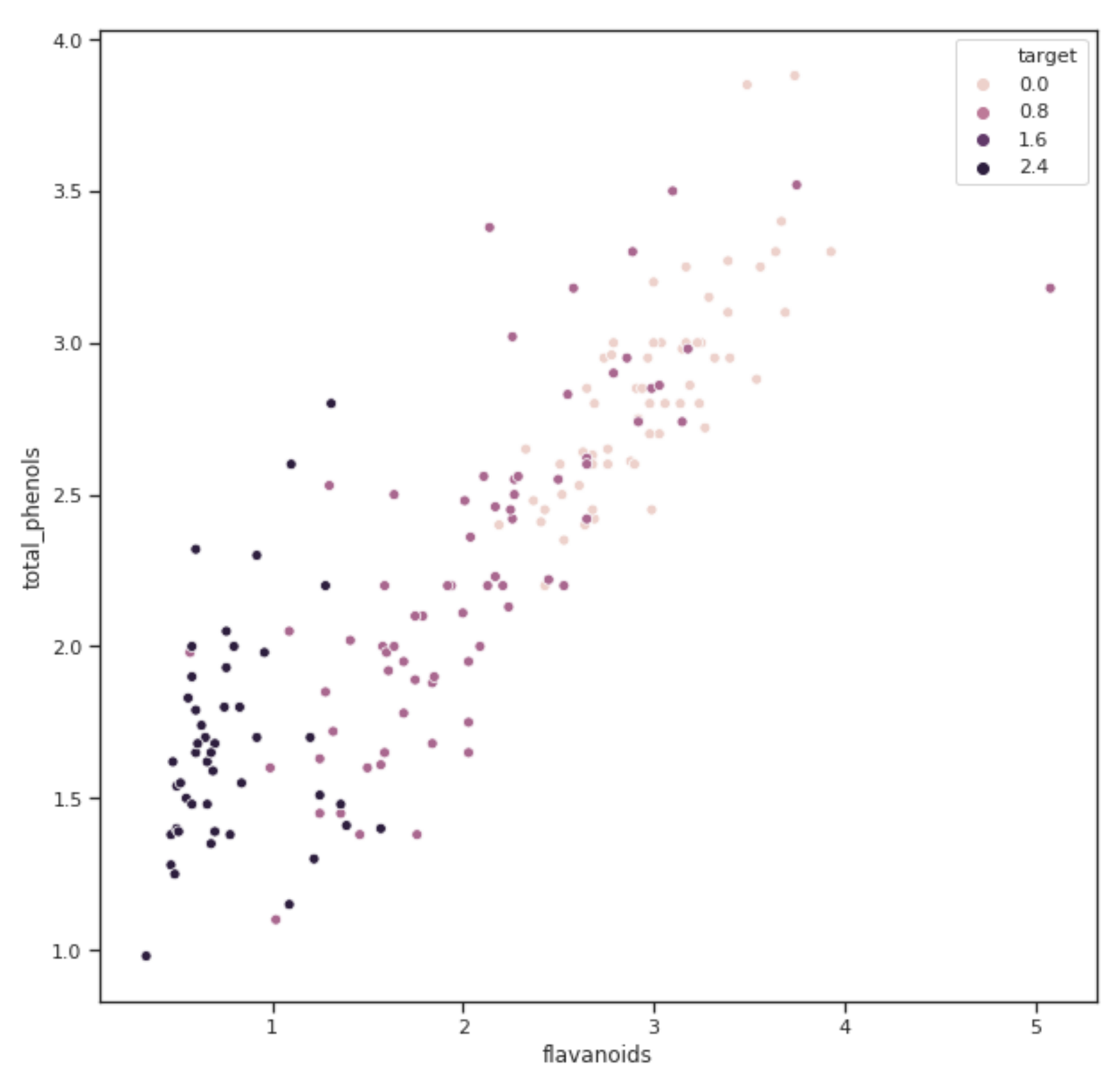
Out[10]:
```
array([0., 1., 2.])
```

Визуальное иследование датасета

Диаграмма рассеяния

In [11]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='flavanoids', y='total_phenols', data=data, hue
='target')
```

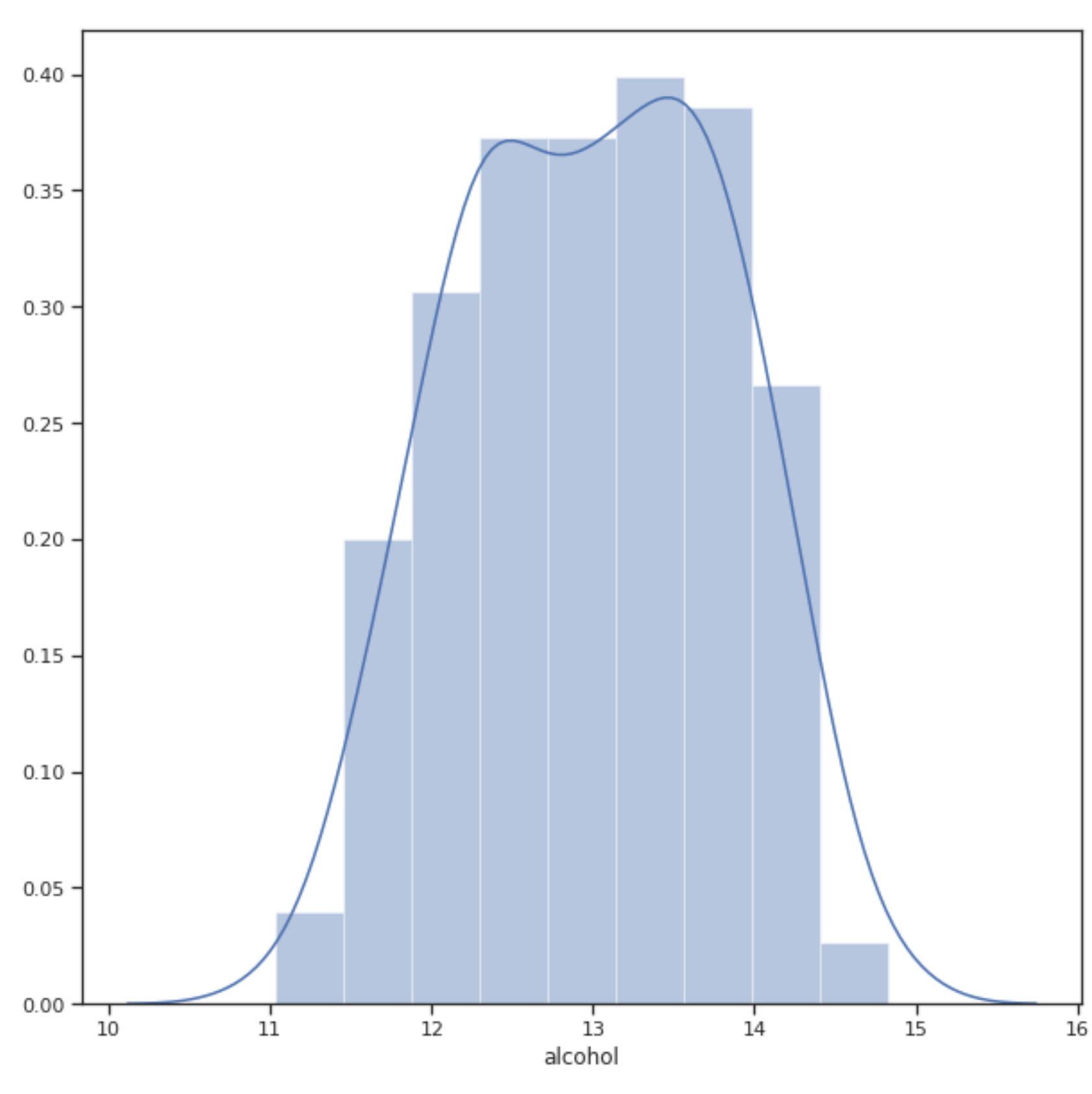
Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f10e9006a58>
```



Гистограмма

In [12]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['alcohol'])
```

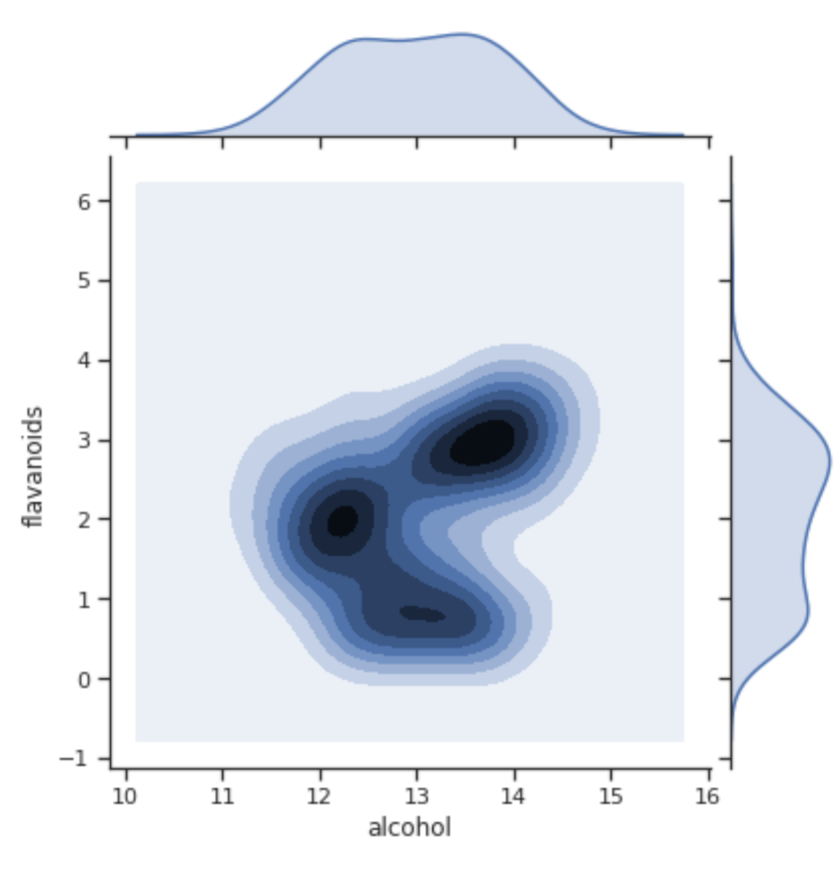
Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f10e62aeb70>
```



Комбинация гистограмм и диаграмм рассеивания

In [13]:
```
sns.jointplot(x='alcohol', y='flavanoids', data=data, kind="kde")
```

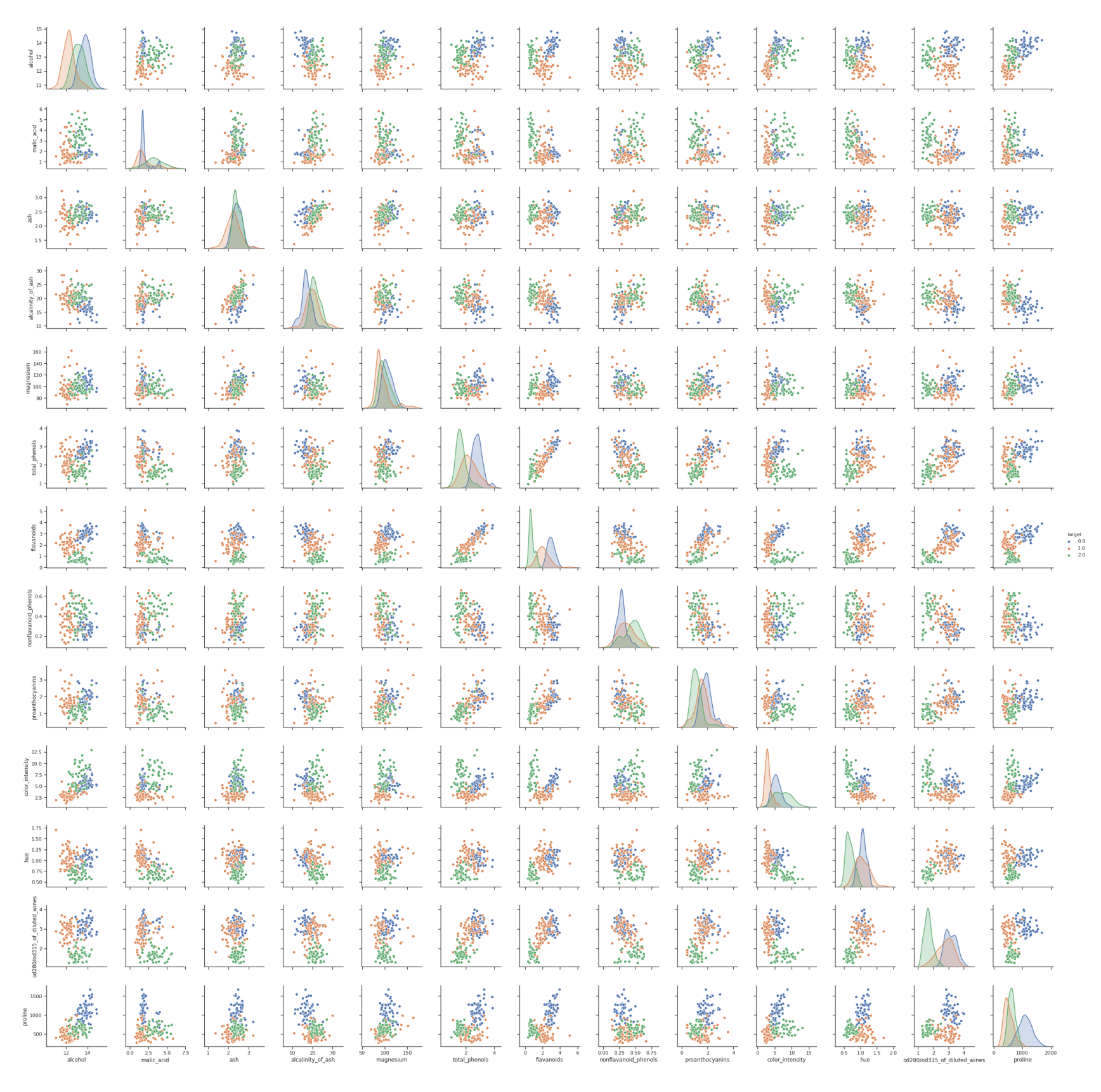
Out[13]:
```
<seaborn.axisgrid.JointGrid at 0x7f10e67214e0>
```



Парные диаграммы

In [14]:
```
sns.pairplot(data, hue='target')
```

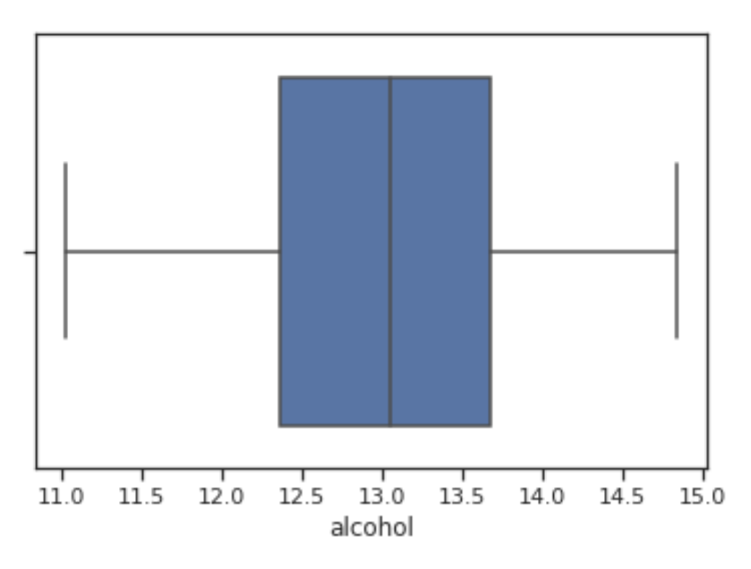
Out[14]:
```
<seaborn.axisgrid.PairGrid at 0x7f10e5eee940>
```



"Ящик с усами"

In [15]:
```
sns.boxplot(x=data['alcohol'])
```

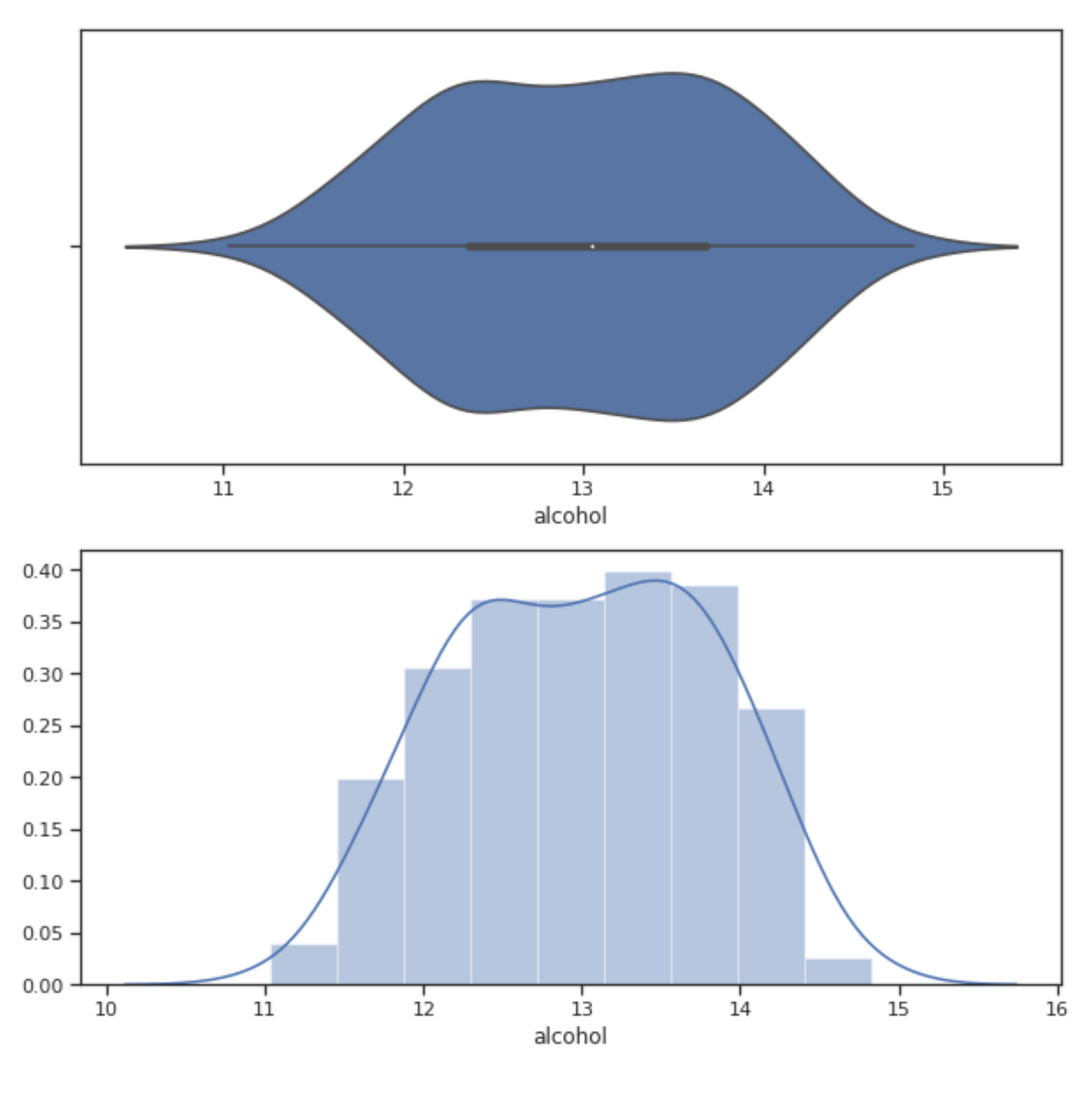
Out[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f10e286e4a8>
```



"Скрипка"

In [16]:
```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['alcohol'])
sns.distplot(data['alcohol'], ax=ax[1])
```

Out[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f10e29bda90>
```



Корреляция признаков

Корреляционная матрица

In [17]:
```
data.corr()
```

Out[17]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total
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	

Heatmap

In [20]:
```
sns.heatmap(data.corr())
```

Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f10df5a4208>
```