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?

Набор данных: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

Столбцы:

- Алкоголь
- Яблочная кислота
- Пепел
- Щелочность золы
- Магний
- Всего фенолов
- Флавоноиды
- Нефлаваноидные фенолы
- Проантоцианы
- Интенсивность цвета
- оттенок
- OD280/OD315 разбавленных вин
- Пролин

Подгружаем необходимые библиотеки и датасет:

```
#Загружаем все бибилиотеки
import numpy as np
import pandas as pd
from sklearn.datasets import *
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

Подключаем DataSet

```
#Преобразование формата в DataFrame - выгрузка датасета про вино
wine = load_wine()
```

```
type(wine)
```

```
sklearn.utils.Bunch
```

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

```
data
target
frame
target_names
DESCR
feature_names
```

```
#Выведем все колонки датасета
wine['feature_names']
```

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

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

Размер набора данных

```
data.shape
```

```
(178, 14)
```

Смотрим на сам датасет

```
data
```

```
     alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  
total_phenols  \
0      14.23        1.71  2.43               15.6      127.0  
2.80
```

```
1      13.20        1.78  2.14              11.2      100.0
2.65
2      13.16        2.36  2.67              18.6      101.0
2.80
3      14.37        1.95  2.50              16.8      113.0
3.85
4      13.24        2.59  2.87              21.0      118.0
2.80
..       ...         ...   ...               ...        ...
...
173    13.71        5.65  2.45              20.5       95.0
1.68
174    13.40        3.91  2.48              23.0      102.0
1.80
175    13.27        4.28  2.26              20.0      120.0
1.59
176    13.17        2.59  2.37              20.0      120.0
1.65
177    14.13        4.10  2.74              24.5       96.0
2.05

     flavanoids  nonflavanoid_phenols  proanthocyanins
color_intensity   hue  \
0          3.06                  0.28             2.29
5.64  1.04
1          2.76                  0.26             1.28
4.38  1.05
2          3.24                  0.30             2.81
5.68  1.03
3          3.49                  0.24             2.18
7.80  0.86
4          2.69                  0.39             1.82
4.32  1.04
..          ...                   ...              ...              ..
.   ...
173        0.61                  0.52             1.06
7.70  0.64
174        0.75                  0.43             1.41
7.30  0.70
175        0.69                  0.43             1.35
10.20  0.59
176        0.68                  0.53             1.46
9.30  0.60
177        0.76                  0.56             1.35
9.20  0.61

     od280/od315_of_diluted_wines  proline  target
0                            3.92   1065.0     0.0
1                            3.40   1050.0     0.0
2                            3.17   1185.0     0.0
```

```
3                            3.45   1480.0     0.0
4                            2.93    735.0     0.0
..                            ...      ...     ...
173                          1.74    740.0     2.0
174                          1.56    750.0     2.0
175                          1.56    835.0     2.0
176                          1.62    840.0     2.0
177                          1.60    560.0     2.0

[178 rows x 14 columns]
```

```
data.head(5)
```

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

   flavanoids  nonflavanoid_phenols  proanthocyanins  color_intensity
hue  \
0        3.06                  0.28             2.29             5.64
1.04
1        2.76                  0.26             1.28             4.38
1.05
2        3.24                  0.30             2.81             5.68
1.03
3        3.49                  0.24             2.18             7.80
0.86
4        2.69                  0.39             1.82             4.32
1.04

   od280/od315_of_diluted_wines  proline  target
0                          3.92   1065.0     0.0
1                          3.40   1050.0     0.0
2                          3.17   1185.0     0.0
3                          3.45   1480.0     0.0
4                          2.93    735.0     0.0
```

типы колонок

```
#Узнаем типы данных каждого столбца
data.dtypes
```

```
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
```

```
#Проверим количество пустых значений
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
```

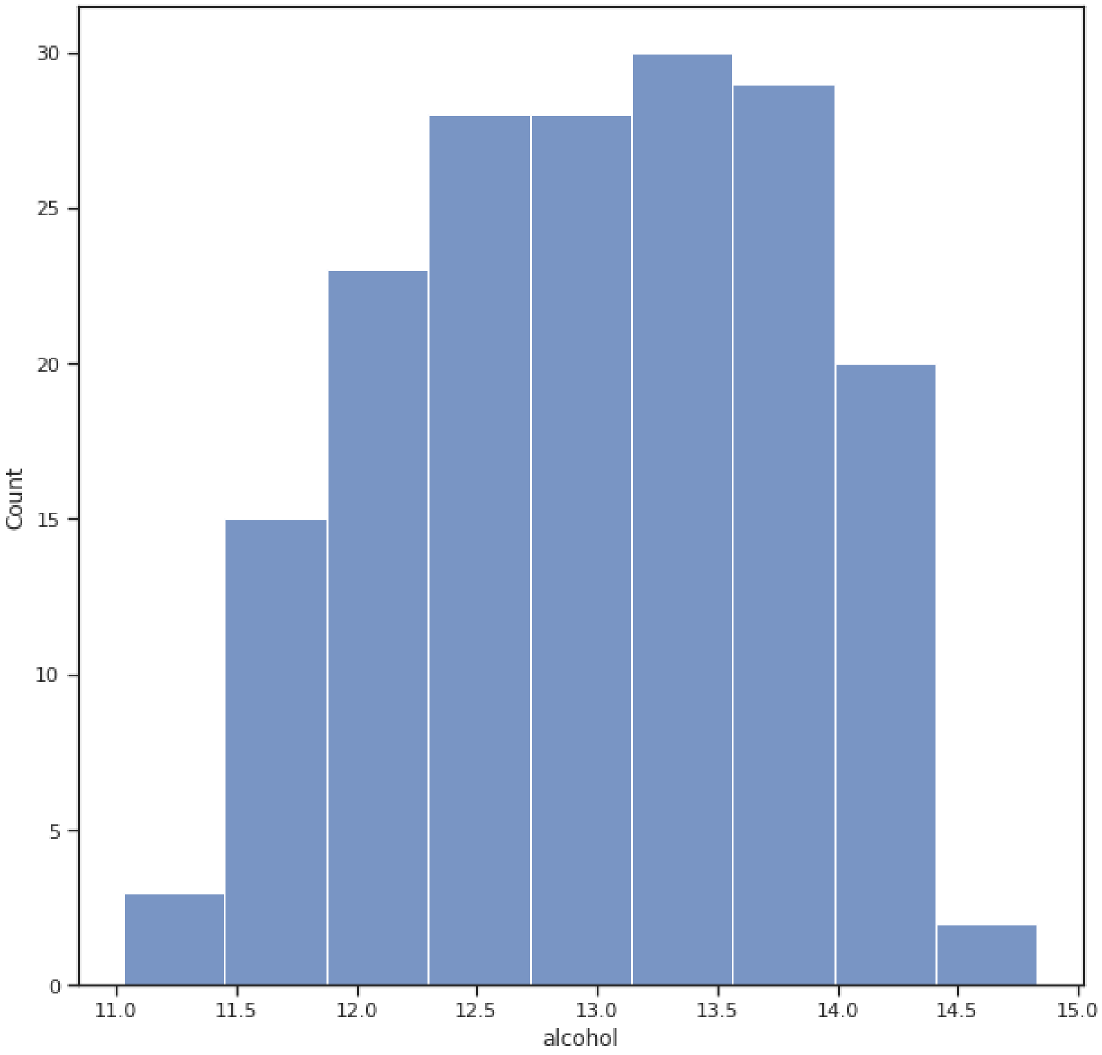
Визуальное исследование датасета

Гистограммы

Гистограмма распределения % алкоголя.

```
fig, ax = plt. subplots (figsize=(10,10))
sns.histplot(data['alcohol'])
```

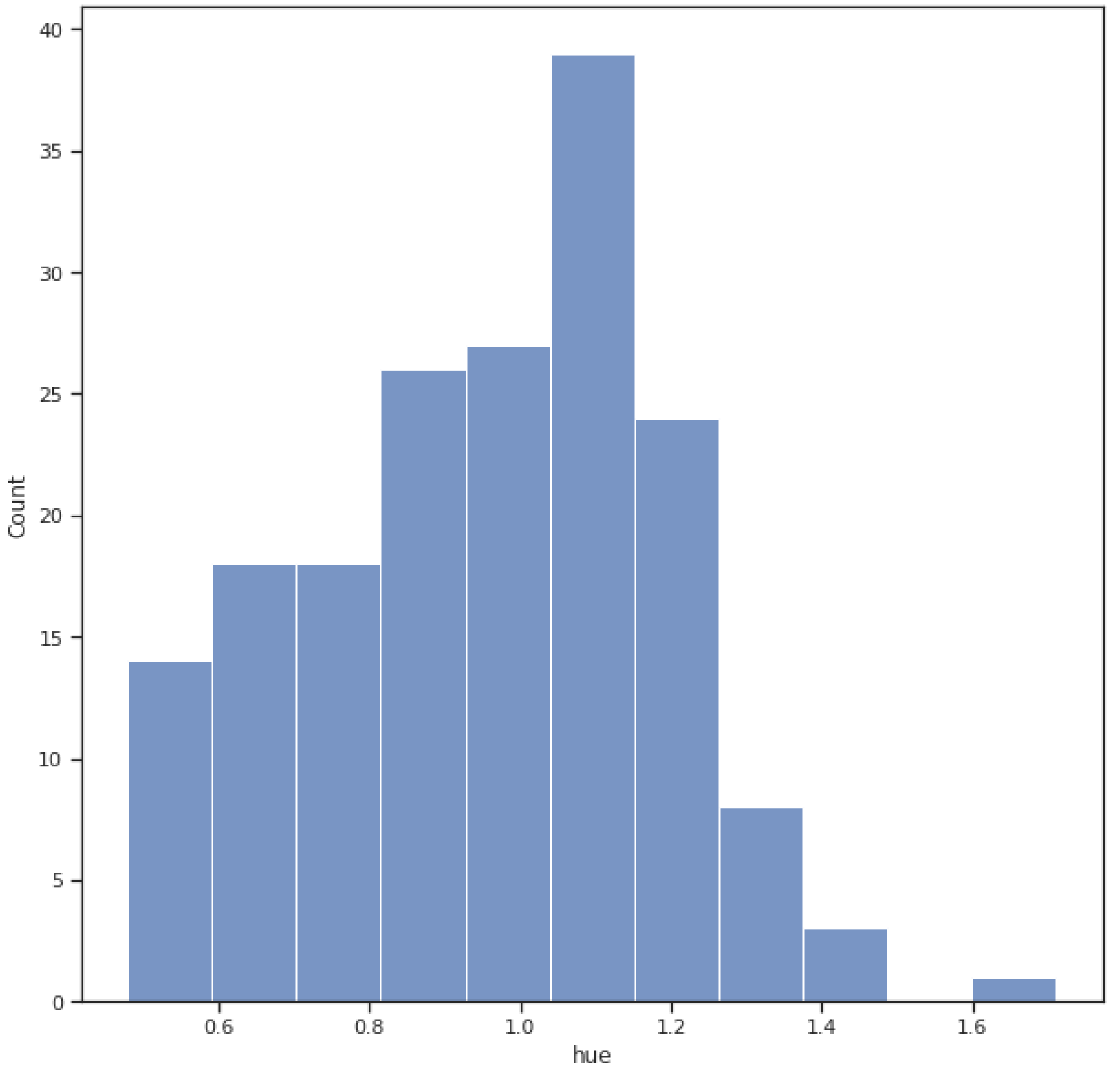
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f28495eded0>
```



Распределение оттенков

```
fig, ax = plt. subplots (figsize=(10,10))
sns.histplot(data['hue'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f284d229150>
```

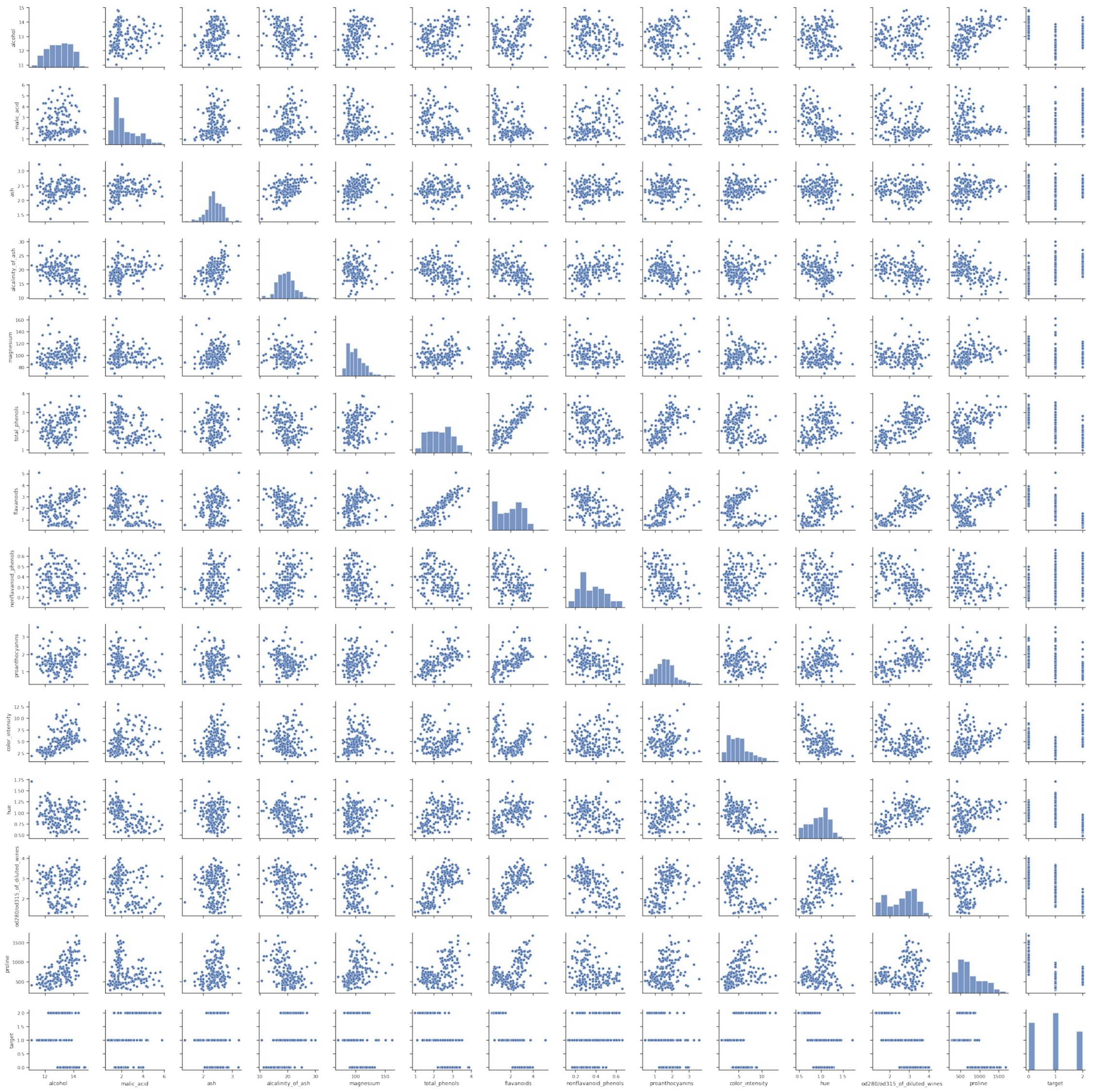


тут виден пропущенный оттенок, а также гистограмма не соотвествует закону нормального распределения.

Парные диаграммы

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

```
<seaborn.axisgrid.PairGrid at 0x7f2848dd67d0>
```



Парные диаграммы позволяют построить большинство диаграмм. На них присутствуют также бессмысленные сравнения данных.

```
#Производим коррелляционный анализ
data.corr()
```

```
                              alcohol  malic_acid       ash  \
alcohol                      1.000000    0.094397  0.211545   
malic_acid                   0.094397    1.000000  0.164045   
ash                          0.211545    0.164045  1.000000   
alcalinity_of_ash           -0.310235    0.288500  0.443367   
magnesium                    0.270798   -0.054575  0.286587   
total_phenols                0.289101   -0.335167  0.128980   
flavanoids                   0.236815   -0.411007  0.115077   
nonflavanoid_phenols        -0.155929    0.292977  0.186230   
proanthocyanins              0.136698   -0.220746  0.009652   
color_intensity              0.546364    0.248985  0.258887   
```

```
hue                          -0.071747   -0.561296 -0.074667
od280/od315_of_diluted_wines  0.072343   -0.368710  0.003911
proline                       0.643720   -0.192011  0.223626
target                       -0.328222    0.437776 -0.049643

                              alcalinity_of_ash  magnesium
total_phenols  \
alcohol                                -0.310235   0.270798
0.289101
malic_acid                              0.288500  -0.054575      -
0.335167
ash                                     0.443367   0.286587
0.128980
alcalinity_of_ash                       1.000000  -0.083333      -
0.321113
magnesium                              -0.083333   1.000000
0.214401
total_phenols                          -0.321113   0.214401
1.000000
flavanoids                             -0.351370   0.195784
0.864564
nonflavanoid_phenols                    0.361922  -0.256294      -
0.449935
proanthocyanins                        -0.197327   0.236441
0.612413
color_intensity                         0.018732   0.199950      -
0.055136
hue                                    -0.273955   0.055398
0.433681
od280/od315_of_diluted_wines           -0.276769   0.066004
0.699949
proline                                -0.440597   0.393351
0.498115
target                                  0.517859  -0.209179      -
0.719163

                              flavanoids  nonflavanoid_phenols  \
alcohol                         0.236815             -0.155929
malic_acid                     -0.411007              0.292977
ash                             0.115077              0.186230
alcalinity_of_ash              -0.351370              0.361922
magnesium                       0.195784             -0.256294
total_phenols                   0.864564             -0.449935
flavanoids                      1.000000             -0.537900
nonflavanoid_phenols           -0.537900              1.000000
proanthocyanins                 0.652692             -0.365845
color_intensity                -0.172379              0.139057
hue                             0.543479             -0.262640
od280/od315_of_diluted_wines    0.787194             -0.503270
proline                         0.494193             -0.311385
```

```
target                         -0.847498              0.489109

                              proanthocyanins  color_intensity
hue  \
alcohol                              0.136698         0.546364 -
0.071747
malic_acid                          -0.220746         0.248985 -
0.561296
ash                                  0.009652         0.258887 -
0.074667
alcalinity_of_ash                   -0.197327         0.018732 -
0.273955
magnesium                            0.236441         0.199950
0.055398
total_phenols                        0.612413        -0.055136
0.433681
flavanoids                           0.652692        -0.172379
0.543479
nonflavanoid_phenols                -0.365845         0.139057 -
0.262640
proanthocyanins                      1.000000        -0.025250
0.295544
color_intensity                     -0.025250         1.000000 -
0.521813
hue                                  0.295544        -0.521813
1.000000
od280/od315_of_diluted_wines         0.519067        -0.428815
0.565468
proline                              0.330417         0.316100
0.236183
target                              -0.499130         0.265668 -
0.617369

                              od280/od315_of_diluted_wines   proline
target
alcohol                                           0.072343  0.643720 -
0.328222
malic_acid                                       -0.368710 -0.192011
0.437776
ash                                               0.003911  0.223626 -
0.049643
alcalinity_of_ash                                -0.276769 -0.440597
0.517859
magnesium                                         0.066004  0.393351 -
0.209179
total_phenols                                     0.699949  0.498115 -
0.719163
flavanoids                                        0.787194  0.494193 -
0.847498
nonflavanoid_phenols                             -0.503270 -0.311385
```

```
0.489109
proanthocyanins                                    0.519067  0.330417 -
0.499130
color_intensity                                   -0.428815  0.316100
0.265668
hue                                                0.565468  0.236183 -
0.617369
od280/od315_of_diluted_wines                       1.000000  0.312761 -
0.788230
proline                                            0.312761  1.000000 -
0.633717
target                                            -0.788230 -0.633717
1.000000
```

```
#Корелляционный анализ методом Спирмана
data.corr(method='spearman')
```

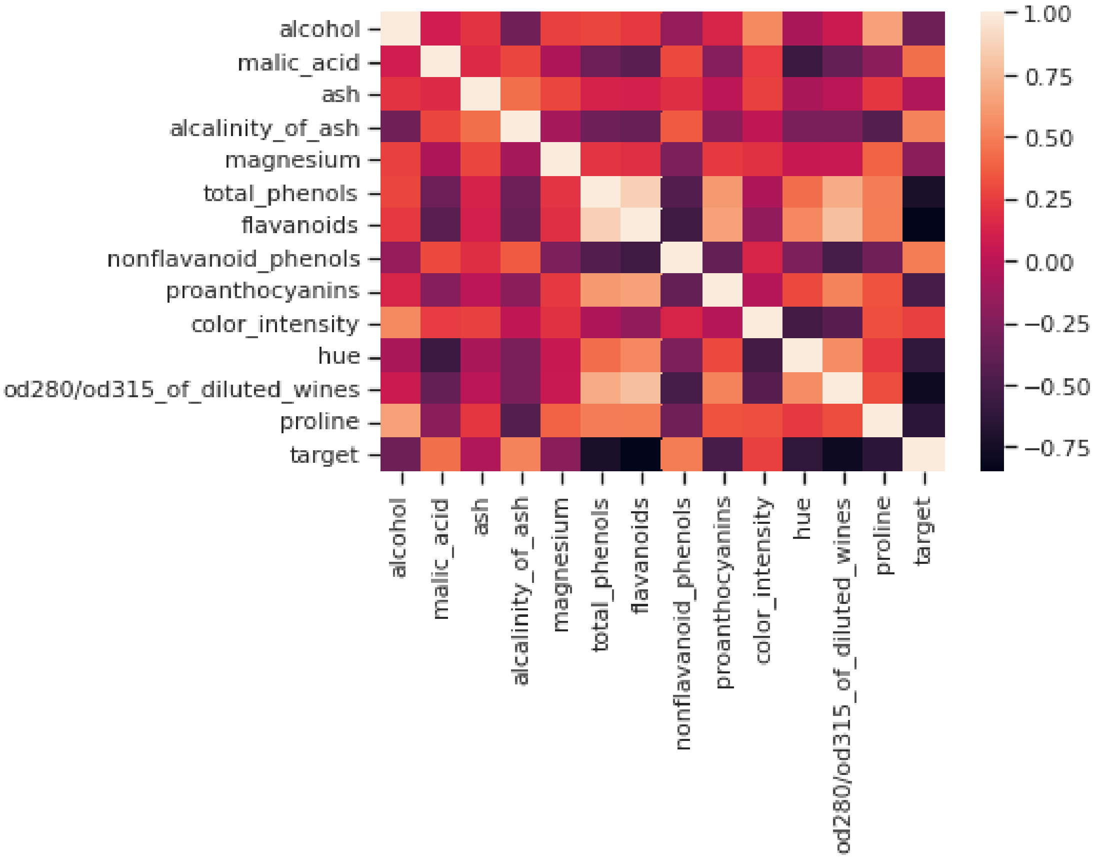
```
                               alcohol  malic_acid       ash  \
alcohol                       1.000000    0.140430  0.243722
malic_acid                    0.140430    1.000000  0.230674
ash                           0.243722    0.230674  1.000000
alcalinity_of_ash            -0.306598    0.304069  0.366374
magnesium                     0.365503    0.080188  0.361488
total_phenols                 0.310920   -0.280225  0.132193
flavanoids                    0.294740   -0.325202  0.078796
nonflavanoid_phenols         -0.162207    0.255236  0.145583
proanthocyanins               0.192734   -0.244825  0.024384
color_intensity               0.635425    0.290307  0.283047
hue                          -0.024203   -0.560265 -0.050183
od280/od315_of_diluted_wines  0.103050   -0.255185 -0.007500
proline                       0.633580   -0.057466  0.253163
target                       -0.354167    0.346913 -0.053988

                              alcalinity_of_ash  magnesium
total_phenols  \
alcohol                               -0.306598   0.365503
0.310920
malic_acid                             0.304069   0.080188      -
0.280225
ash                                    0.366374   0.361488
0.132193
alcalinity_of_ash                      1.000000  -0.169558      -
0.376657
magnesium                             -0.169558   1.000000
0.246417
total_phenols                         -0.376657   0.246417
1.000000
flavanoids                            -0.443770   0.233167
0.879404
nonflavanoid_phenols                   0.389390  -0.236786      -
```

```
0.448013
proanthocyanins                       -0.253695   0.173647
0.666689
color_intensity                       -0.073776   0.357029
0.011162
hue                                   -0.352507   0.036095
0.439457
od280/od315_of_diluted_wines          -0.325890   0.056963
0.687207
proline                               -0.456090   0.507575
0.419470
target                                 0.569792  -0.250498      -
0.726544

                              flavanoids  nonflavanoid_phenols  \
alcohol                         0.294740             -0.162207
malic_acid                     -0.325202              0.255236
ash                             0.078796              0.145583
alcalinity_of_ash              -0.443770              0.389390
magnesium                       0.233167             -0.236786
total_phenols                   0.879404             -0.448013
flavanoids                      1.000000             -0.543897
nonflavanoid_phenols           -0.543897              1.000000
proanthocyanins                 0.730322             -0.384629
color_intensity                -0.042910              0.059639
hue                             0.535430             -0.267813
od280/od315_of_diluted_wines    0.741533             -0.494950
proline                         0.429904             -0.270112
target                         -0.854908              0.474205

                              proanthocyanins  color_intensity
hue  \
alcohol                              0.192734         0.635425 -
0.024203
malic_acid                          -0.244825         0.290307 -
0.560265
ash                                  0.024384         0.283047 -
0.050183
alcalinity_of_ash                   -0.253695        -0.073776 -
0.352507
magnesium                            0.173647         0.357029
0.036095
total_phenols                        0.666689         0.011162
0.439457
flavanoids                           0.730322        -0.042910
0.535430
nonflavanoid_phenols                -0.384629         0.059639 -
0.267813
proanthocyanins                      1.000000        -0.030947
0.342795
```

```
color_intensity                     -0.030947         1.000000 -
0.418522
hue                                  0.342795        -0.418522
1.000000
od280/od315_of_diluted_wines         0.554031        -0.317516
0.485454
proline                              0.308249         0.457096
0.207740
target                              -0.570648         0.131170 -
0.616570

                              od280/od315_of_diluted_wines   proline
target
alcohol                                           0.103050  0.633580 -
0.354167
malic_acid                                       -0.255185 -0.057466
0.346913
ash                                              -0.007500  0.253163 -
0.053988
alcalinity_of_ash                                -0.325890 -0.456090
0.569792
magnesium                                         0.056963  0.507575 -
0.250498
total_phenols                                     0.687207  0.419470 -
0.726544
flavanoids                                        0.741533  0.429904 -
0.854908
nonflavanoid_phenols                             -0.494950 -0.270112
0.474205
proanthocyanins                                   0.554031  0.308249 -
0.570648
color_intensity                                  -0.317516  0.457096
0.131170
hue                                               0.485454  0.207740 -
0.616570
od280/od315_of_diluted_wines                      1.000000  0.253266 -
0.743787
proline                                           0.253266  1.000000 -
0.576383
target                                           -0.743787 -0.576383
1.000000
```

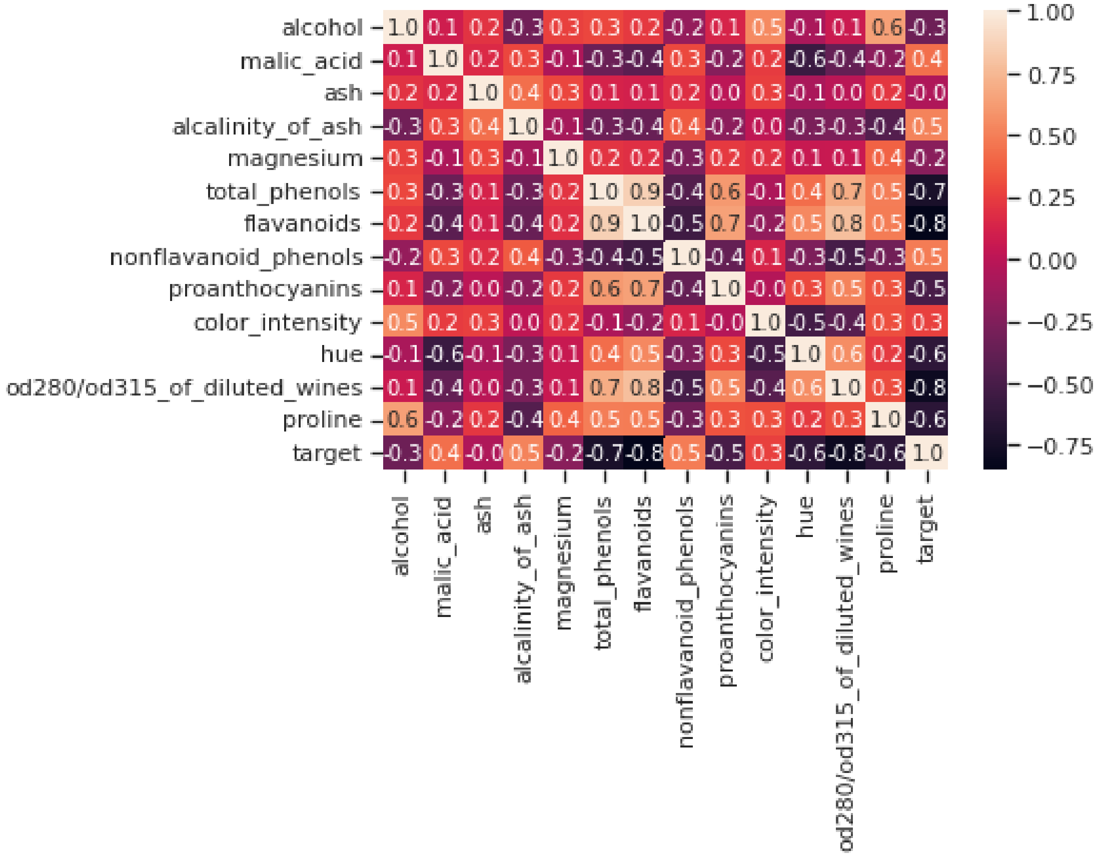
```
#Используем тепловые карты для того, чтобы показать стеень корелляции
различными цветами
sns.heatmap(data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f28524cad50>
```



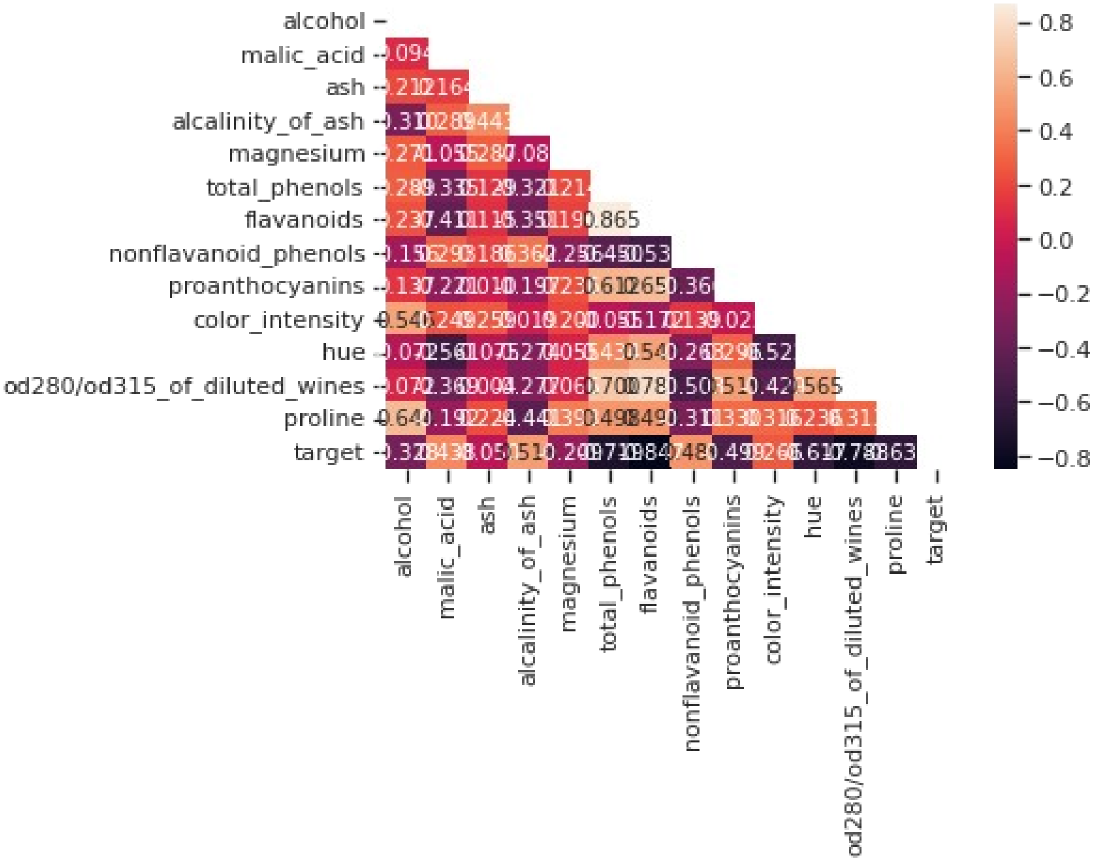
```
sns.heatmap(data.corr(), annot=True, fmt='.1f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f284f688590>
```

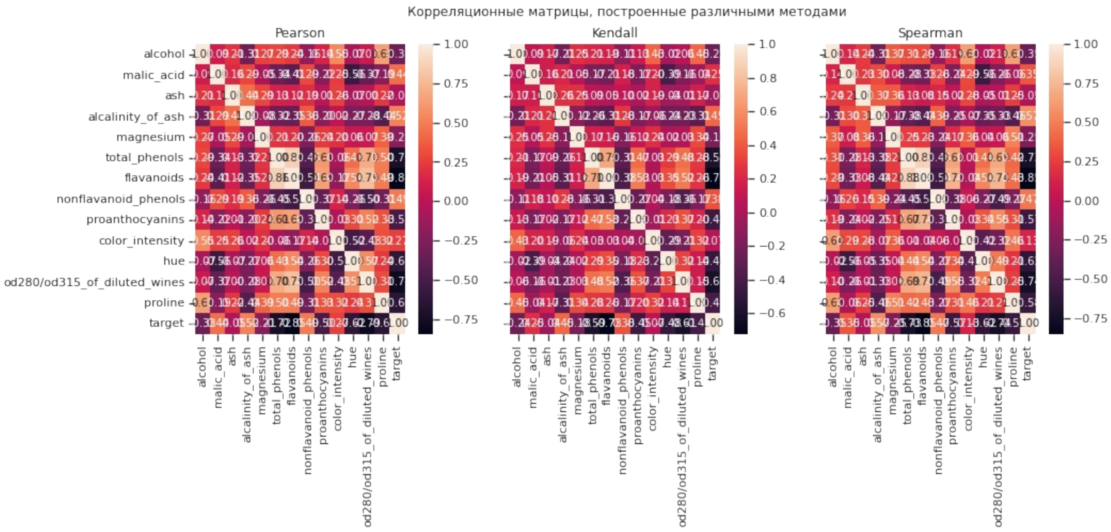


```
# Треугольный вариант матрицы
mask = np.zeros_like(data.corr(), dtype=np.bool)
# чтобы оставить нижнюю часть матрицы
mask[np.triu_indices_from(mask)] = True
# чтобы оставить верхнюю часть матрицы
#mask[np.tril_indices_from(mask)] = True
sns.heatmap(data.corr(), mask=mask, annot=True, fmt='.3f')
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2:
DeprecationWarning: `np.bool` is a deprecated alias for the builtin
`bool`. To silence this warning, use `bool` by itself. Doing this will
not modify any behavior and is safe. If you specifically wanted the
numpy scalar type, use `np.bool_` here.
Deprecated in NumPy 1.20; for more details and guidance:
https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations


<matplotlib.axes._subplots.AxesSubplot at 0x7f284da7e090>
```

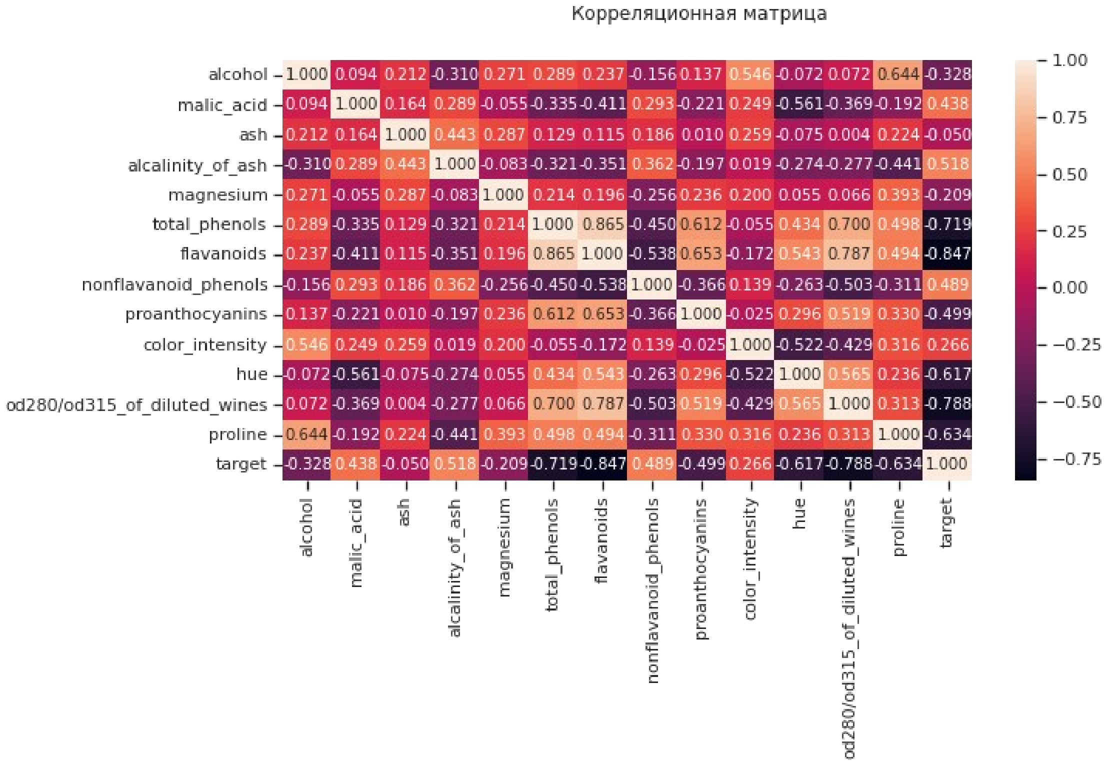


```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row',
figsize=(15,5))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True,
fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True,
fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True,
fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными
методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```



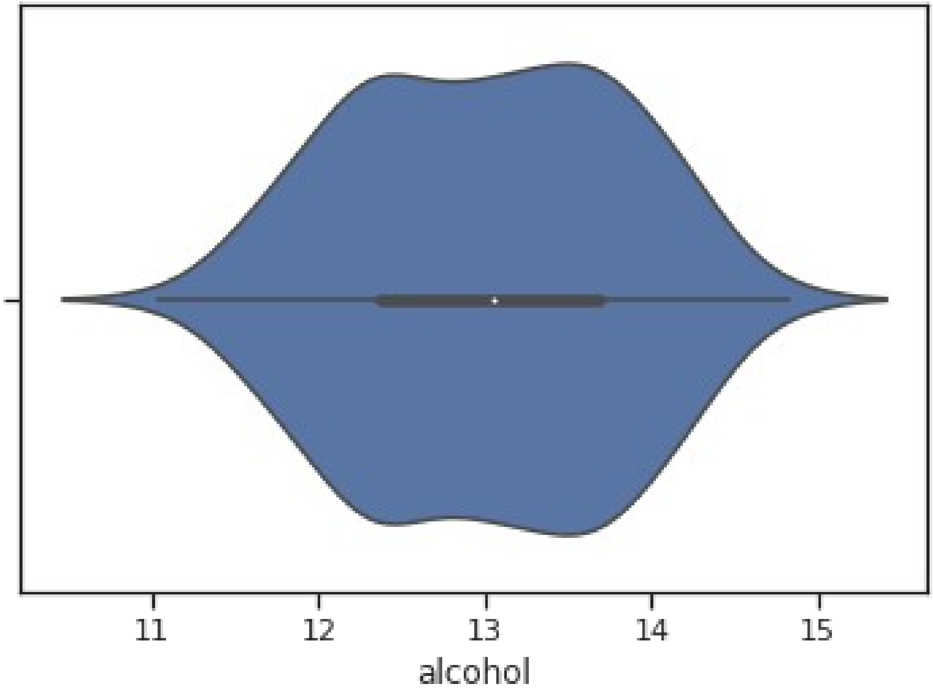
```
fig, ax = plt.subplots(1, 1, sharex='col', sharey='row',
figsize=(10,5))
fig.suptitle('Корреляционная матрица')
sns.heatmap(data.corr(), ax=ax, annot=True, fmt='.3f')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f284d114350>
```



```
#Дополнительное задание для группы ИУ5Ц-84Б - Скрипичная диаграмма
(violin plot).
sns.violinplot(x=data['alcohol'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f284cc7f350>
```


11
12
13
14
15
alcohol