Московский государственный технический

университет им. Н. Э. Баумана

Отчёт по лабораторной работе №1 по курсу «Технологии машинного обучения».

«Разведочный анализ данных. Исследование и визуализация данных».

Выполнил:
Анцифров Н. С.
студент группы ИУ5-61Б

Проверил:
Гапанюк Ю. Е.

Подпись и дата:

Подпись и дата:

Москва, 2022 г.

1. Задание лабораторной работы

- Выбрать набор данных (датасет)
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn
- Для лабораторных работ не рекомендуется выбирать датасеты большого размера
- Создать ноутбук, который содержит следующие разделы: текстовое описание выбранного Вами набора данных, основные характеристики датасета, визуальное исследование датасета, информация о корреляции признаков.
- Сформировать отчет и разместить его в своем репозитории на github.

2. Ячейки Jupyter-ноутбука

1. Текстовое описание датасета

В качестве датасета (набора данных) будем использовать набор данных, содержащий данные для распознавания вин. Данный набор доступен по адресу: https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-recognition-dataset

Набор данных не содержит пропусков в данных.

Набор данных имеет следующие атрибуты:

- Alcohol - алкоголь
- Malic acid - яблочная кислота
- Ash - зола
- Alcalinity of ash - щелочность
- Magnesium - магний
- Total phenols - количество фенолов
- Flavanoids - флавоноиды
- Nonflavanoid phenols - нефлаваноидные фенолы
- Proanthocyanins - проантоцианы
- Color intensity - интенсивность цвета
- Hue - оттенок
- OD280/OD315 of diluted wines - OD280/OD315 разбавленных вин
- Proline - пролин

Импорт библиотек

Импортируем библиотеки с помощью команды import и установим стиль по умолчанию.

In [58]:

```
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 *
import warnings
warnings.filterwarnings('ignore')
```

Загрузка данных

Загрузим набор данных, содеражищй информацию для распознавания вин:

In [21]:

```
wine = load_wine()
```

Проверим загрузку:

In [22]:

```
type(wine)
```

Out[22]:

```
sklearn.utils.Bunch
```

In [23]:

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

Out[23]:

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

In [24]:

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

Out[24]:

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

Преобразуем набор данных в Pandas Dataframe:

In [25]:

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

In [26]:

```
data_wine
```

Out[26]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_i
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	
...	...	...	...	...	...	...	...	...	...	
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_i
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	0.56	1.35	

178 rows × 14 columns

2. Основные характеристики датасета

Выведем первые 5 строк датасета:

In [27]:

```
data_wine.head()
```

Out[27]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_inte
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	

Определим размер датасета:

In [29]:

```
data_wine.shape
```

Out[29]:

```
(178, 14)
```

В датасете 178 строк и 14 столбцов. Определим названия столбцов и их тип:

In [31]:

```
data_wine.columns
```

Out[31]:

```
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 [32]:

```
data_wine.dtypes
```

Out[32]:

```
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 наличие пустых значений:

In [33]:

```
for col in data_wine.columns:
    temp_null_count = data_wine[data_wine[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
```

Видим, что пустых значений в датасете нет.

Основные статистические характеристки набора данных:

In [34]:

```
data_wine.describe()
```

Out[34]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocy
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.0
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.5
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.5
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.4
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.2
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.5
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.9
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.5

Определим уникальные значения для целевого признака (сорт вина):

In [36]:

```
data_wine['target'].unique()
```

Out[36]:

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

Целевой признак содержит только три значения (три сорта).

3. Визуальное исследование датасета

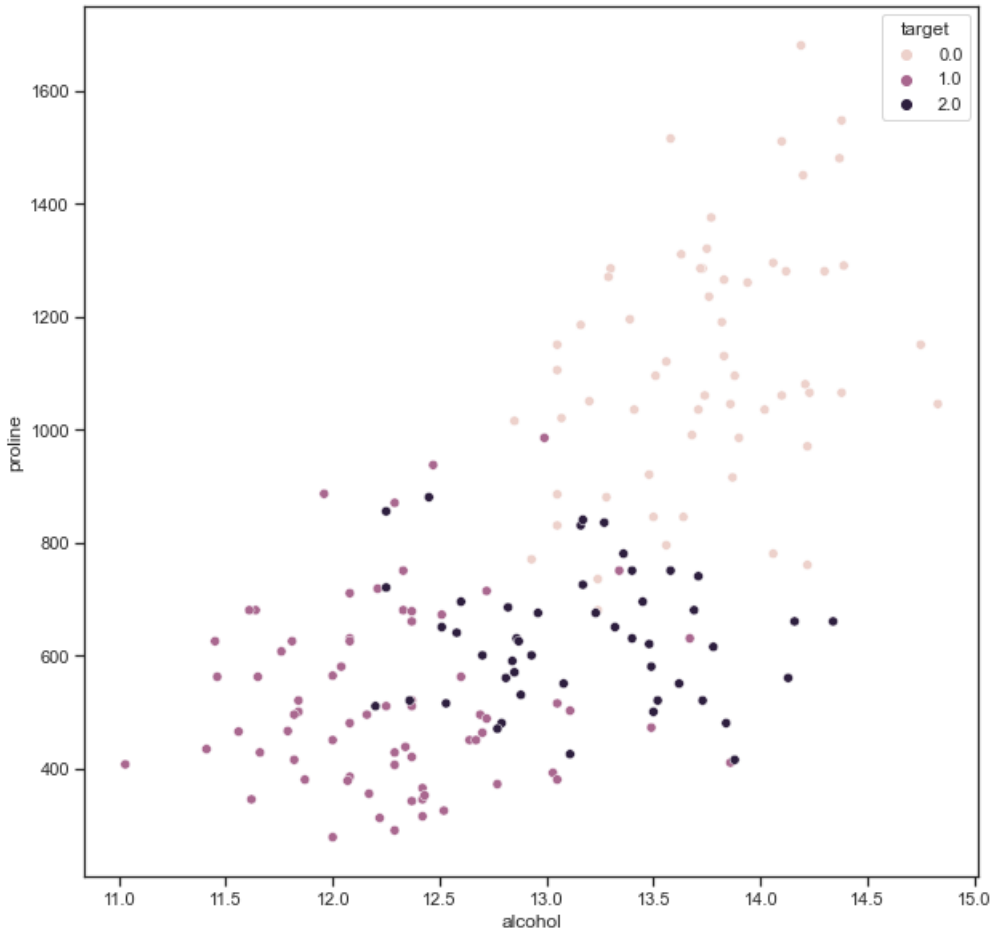
Диаграмма рассеяния - распределение двух столбцов данных и отображение визуальной зависимости между ними:

In [53]:

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

Out[53]:

```
<AxesSubplot:xlabel='alcohol', ylabel='proline'>
```



Из диаграммы можно сделать частичный вывод о том, что чем больше алкоголя в напитке, тем больше в нём пролина. Причём также наблюдается зависимость между 3 сортами напитка (на диаграмме разница по цвету).

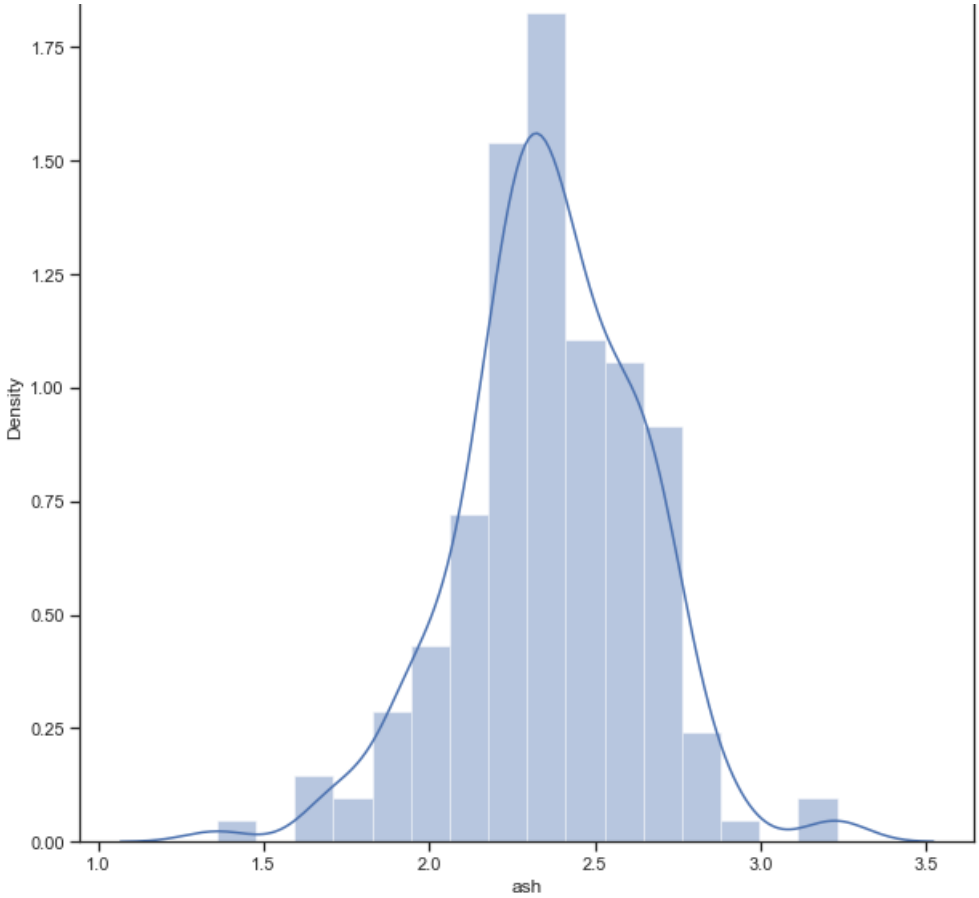
Гистограмма отображает плотность вероятности распределения данных:

In [59]:

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data_wine['ash'])
```

Out[59]:

```
<AxesSubplot:xlabel='ash', ylabel='Density'>
```



Видно распределение золы в напитках.

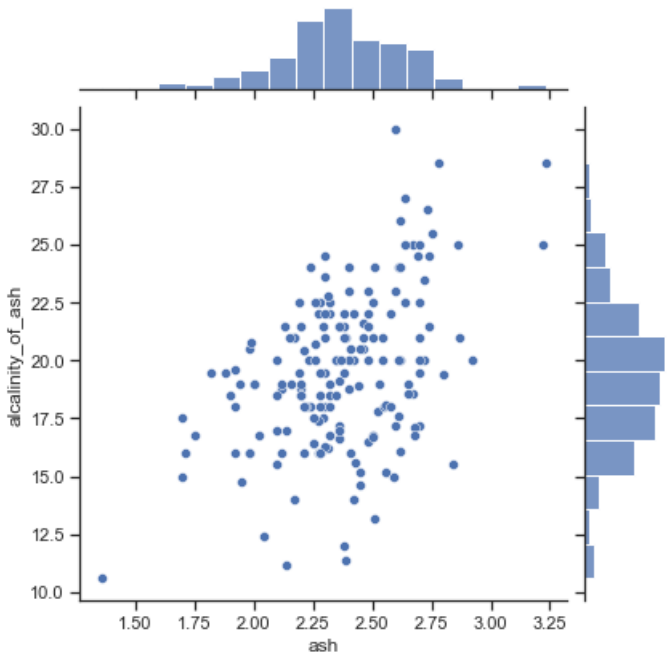
Комбинация гистограмм и диаграмм рассеивания выполняется с помощью jointplot:

In [64]:

```
sns.jointplot(x='ash', y='alcalinity_of_ash', data=data_wine)
```

Out[64]:

```
<seaborn.axisgrid.JointGrid at 0x21b1a481e20>
```



Данные можно представить в виде парных диаграмм - матрицы графиков:

In [67]:

```
sns.pairplot(data_wine, hue="target")
```

Out[67]:

```
<seaborn.axisgrid.PairGrid at 0x21b23d94880>
```

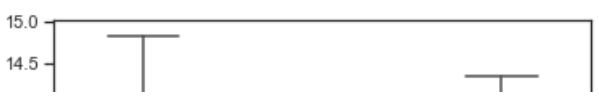
Отображение в виде "Ящика с усами":

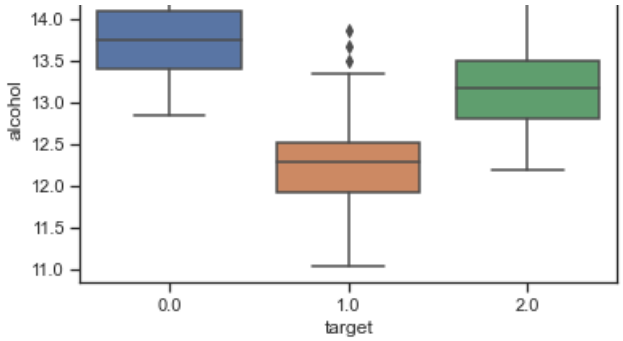
In [71]:

```
sns.boxplot(x='target', y='alcohol', data=data_wine)
```

Out[71]:

```
<AxesSubplot:xlabel='target', ylabel='alcohol'>
```





Он показывает количество алкоголя напитков в зависимости от сортов.

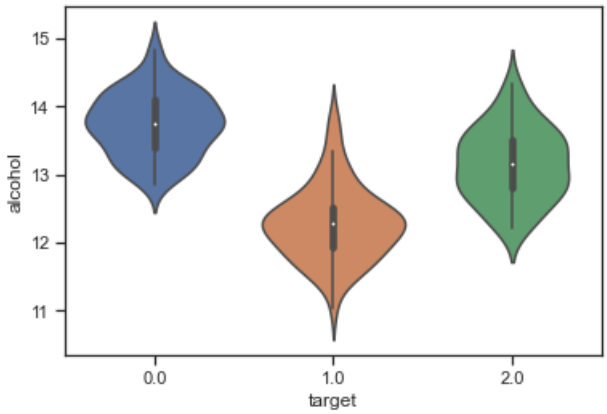
Violin plot дополнительно показывает распределение плотности:

In [76]:

```
sns.violinplot(x='target', y='alcohol', data=data_wine)
```

Out[76]:

```
<AxesSubplot:xlabel='target', ylabel='alcohol'>
```



4. Информация о корреляции признаков

Проверка корреляции помогает найти корреляции с целевом признаком (информативные для машинного обучения), а также выявить линейно независимые нецелевые признаки:

Построим корреляционную матрицу:

In [78]:

```
data_wine.corr()
```

Out[78]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.1
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.2
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.1
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.3

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
magnesium	0.270399	0.054575	0.286587	0.083333	1.000000	0.214401	0.195784	0.2
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.4
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.5
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.0
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.3
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.1
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.2
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.5
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.3
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.4

Можно сделать следующие выводы:

- Целевой признак наиболее сильно коррелирует OD280/OD315 разбавленных вин (-0.78), количеством фенолов (-0.72) и флаваноидами (-0.85) - эти признаки очень важны для модели
- Целевой признак частично коррелирует с нефлаваноидными фенолами (0.49) и проантоцианами (0.5) и щелочностью (0.52) - эти признаки также можно оставить в модели
- Целевой признак слабо коррелирует с алкоголем (-0.33), золой (-0.05), магнием (-0.21) и интенсивностью света (0.27). Такие признаки стоит исключить из модели, так как они ухудшат её качество.

Выше была построена матрица корреляции по Пирсону, но также можно построить матрицы по критерию Кендалла и Спирмена, но разница в значениях будет невелика:

In [80]:

```
data_wine.corr(method='kendall')
```

Out[80]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
alcohol	1.000000	0.093844	0.170154	-0.212978	0.250506	0.209099	0.191087	-0.1
malic_acid	0.093844	1.000000	0.158178	0.210119	0.050869	-0.174929	-0.211918	0.1
ash	0.170154	0.158178	1.000000	0.258352	0.254246	0.089855	0.049474	0.0
alcalinity_of_ash	-0.212978	0.210119	0.258352	1.000000	-0.121005	-0.256669	-0.309865	0.2
magnesium	0.250506	0.050869	0.254246	-0.121005	1.000000	0.172195	0.161603	-0.1
total_phenols	0.209099	-0.174929	0.089855	-0.256669	0.172195	1.000000	0.701999	-0.3
flavanoids	0.191087	-0.211918	0.049474	-0.309865	0.161603	0.701999	1.000000	-0.3
nonflavanoid_phenols	-0.109554	0.175129	0.098937	0.278091	-0.158361	-0.310443	-0.378099	1.0
proanthocyanins	0.133526	-0.168714	0.018240	-0.171404	0.117871	0.466517	0.534615	-0.2
color_intensity	0.434353	0.195607	0.187786	-0.057281	0.241781	0.028264	0.028674	0.0
hue	-0.021717	-0.388707	-0.037234	-0.239210	0.023760	0.289210	0.354372	-0.1

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
od280/od315_of_diluted_wines	0.061513	-0.162909	0.006341	-0.226253	0.034307	0.478267	0.520448	-0.3
proline	0.449387	-0.044660	0.171574	-0.313218	0.343016	0.280203	0.263661	-0.1
target	-0.238984	0.247494	-0.038085	0.449402	-0.184992	-0.590404	-0.725255	0.3

In [81]:

```
data_wine.corr(method='spearman')
```

Out[81]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_ph
alcohol	1.000000	0.140430	0.243722	-0.306598	0.365503	0.310920	0.294740	-0.1
malic_acid	0.140430	1.000000	0.230674	0.304069	0.080188	-0.280225	-0.325202	0.2
ash	0.243722	0.230674	1.000000	0.366374	0.361488	0.132193	0.078796	0.1
alcalinity_of_ash	-0.306598	0.304069	0.366374	1.000000	-0.169558	-0.376657	-0.443770	0.3
magnesium	0.365503	0.080188	0.361488	-0.169558	1.000000	0.246417	0.233167	-0.2
total_phenols	0.310920	-0.280225	0.132193	-0.376657	0.246417	1.000000	0.879404	-0.4
flavanoids	0.294740	-0.325202	0.078796	-0.443770	0.233167	0.879404	1.000000	-0.5
nonflavanoid_phenols	-0.162207	0.255236	0.145583	0.389390	-0.236786	-0.448013	-0.543897	1.0
proanthocyanins	0.192734	-0.244825	0.024384	-0.253695	0.173647	0.666689	0.730322	-0.3
color_intensity	0.635425	0.290307	0.283047	-0.073776	0.357029	0.011162	-0.042910	0.0
hue	-0.024203	-0.560265	-0.050183	-0.352507	0.036095	0.439457	0.535430	-0.2
od280/od315_of_diluted_wines	0.103050	-0.255185	-0.007500	-0.325890	0.056963	0.687207	0.741533	-0.4
proline	0.633580	-0.057466	0.253163	-0.456090	0.507575	0.419470	0.429904	-0.2
target	-0.354167	0.346913	-0.053988	0.569792	-0.250498	-0.726544	-0.854908	0.4

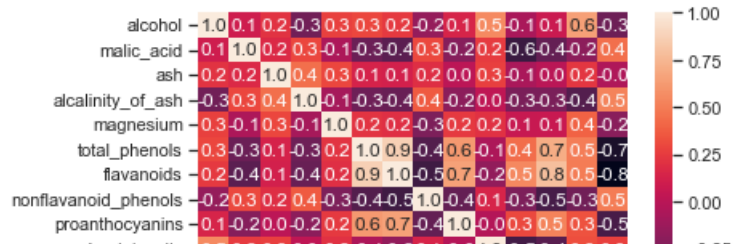
Для визуализации корреляционных матриц используют тепловую карту:

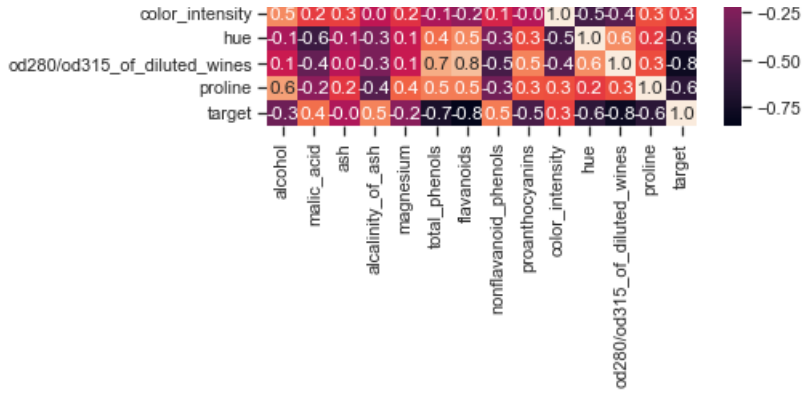
In [92]:

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

Out[92]:

<AxesSubplot:>



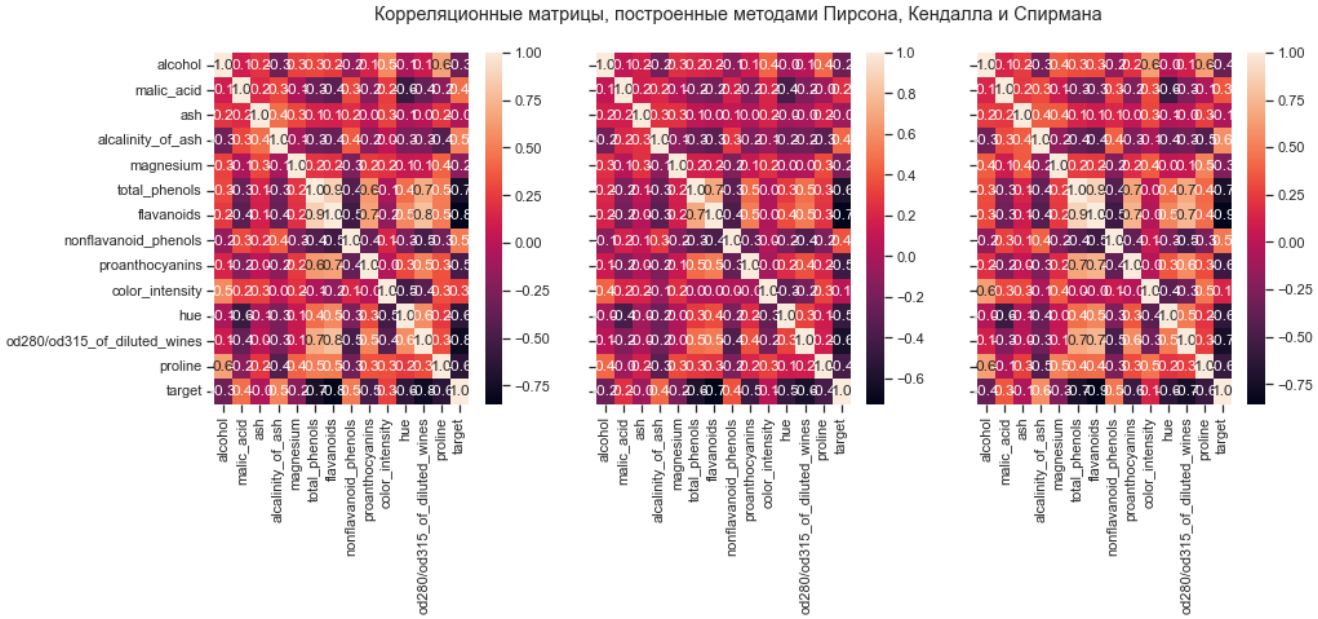


In [90]:

```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))
sns.heatmap(data_wine.corr(method='pearson'), ax=ax[0], annot=True, fmt='.1f')
sns.heatmap(data_wine.corr(method='kendall'), ax=ax[1], annot=True, fmt='.1f')
sns.heatmap(data_wine.corr(method='spearman'), ax=ax[2], annot=True, fmt='.1f')
fig.suptitle('Корреляционные матрицы, построенные методами Пирсона, Кендалла и Спирмана')
```

Out[90]:

```
Text(0.5, 0.98, 'Корреляционные матрицы, построенные методами Пирсона, Кендалла и Спирмана')
```



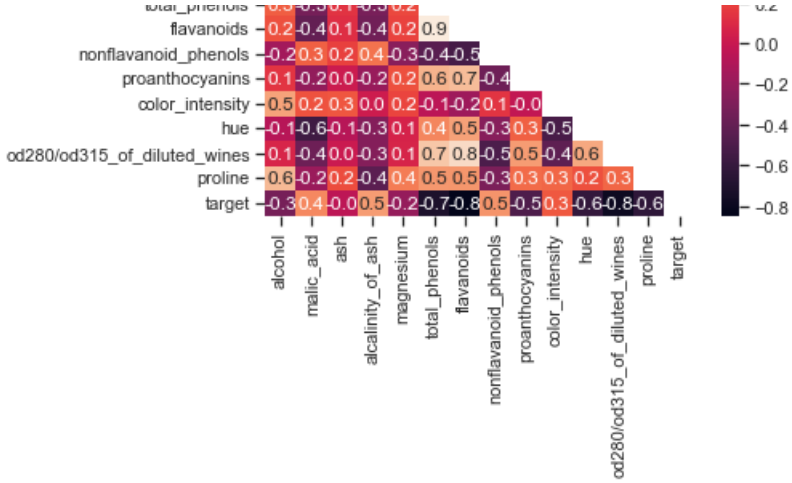
Также можно вывести треугольную матрицу:

In [96]:

```
mask = np.zeros_like(data_wine.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
mask[np.triu_indices_from(mask)] = True
sns.heatmap(data_wine.corr(), mask=mask, annot=True, fmt='.1f')
```

Out[96]:

```
<AxesSubplot:>
```


flavanoids
nonflavanoid_phenols
proanthocyanins
color_intensity
hue
od280/od315_of_diluted_wines
proline
target
alcohol
malic_acid
ash
alcalinity_of_ash
magnesium
total_phenols
flavanoids
nonflavanoid_phenols
proanthocyanins
color_intensity
hue
od280/od315_of_diluted_wines
proline
target
0.0
−0.2
−0.4
−0.6
−0.8