

**Московский государственный технический  
университет им. Н. Э. Баумана**  
**Факультет «Информатика и системы управления»**

Кафедра «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по лабораторной работе №1

Разведочный анализ данных. Исследование и визуализация данных

Группа: РТ5-61

Студент: Савушкин Д.А.

Преподаватель: Гапанюк Ю.Е.

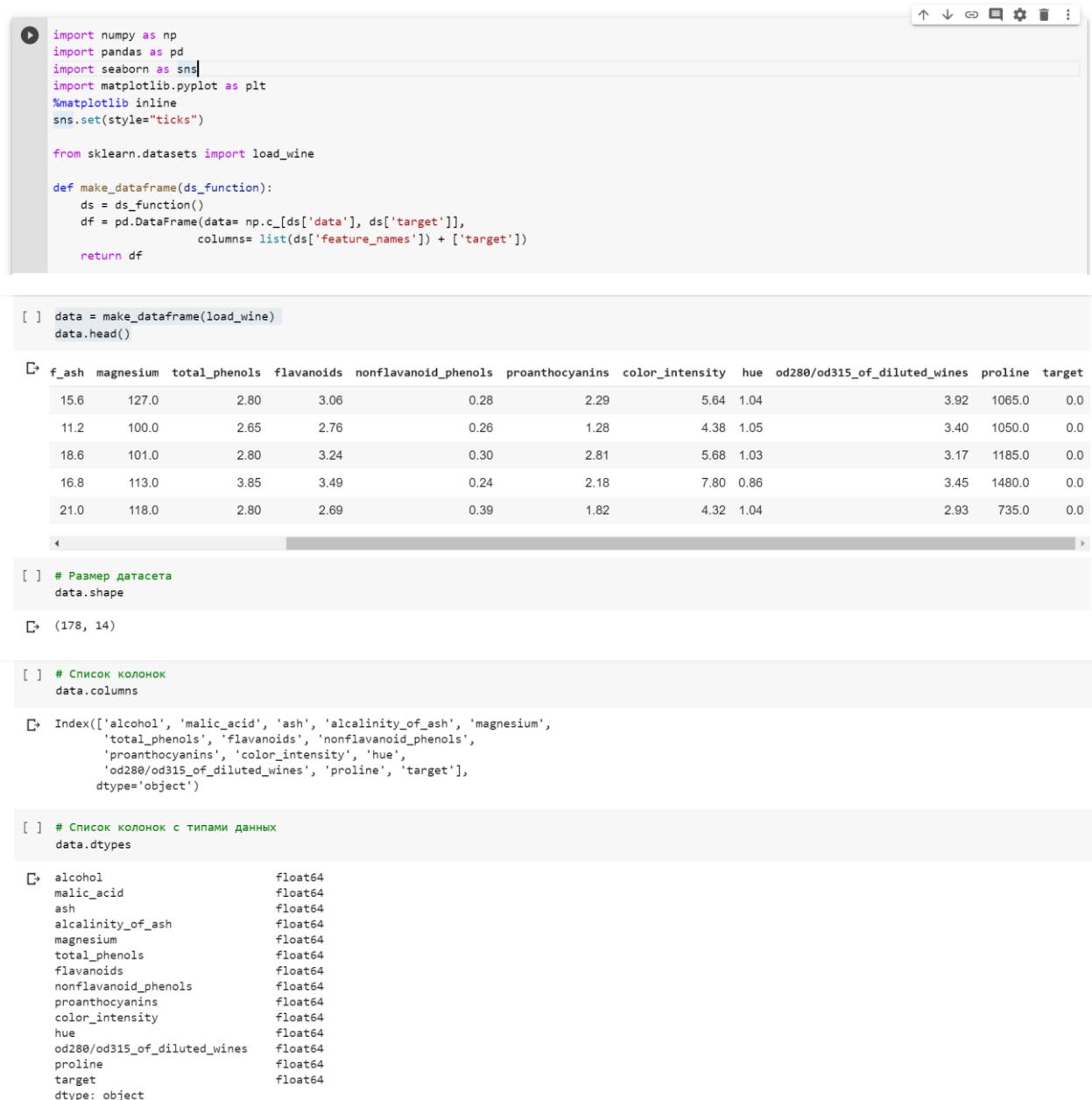
Москва, 2020 г.

**Цель лабораторной работы:** изучение различных методов визуализации данных.

**Описание:** построение основных графиков, входящих в этап разведочного анализа данных.

**Текст программы и экранные формы с примерами выполнения программы:**

Набор данных вина.Классический и простой мультиклассовый классификационный набор данных. Данные представляют собой результаты химического анализа вин, выращенных в одном регионе Италии Существует тринадцать различных измерений, проведенных для разных компонентов, найденных в трех типах вина.



```
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 load_wine

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
```

```
[ ] data = make_dataframe(load_wine)
data.head()
```

f_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target	
15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04		3.92	1065.0	0.0
11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05		3.40	1050.0	0.0
18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03		3.17	1185.0	0.0
16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86		3.45	1480.0	0.0
21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04		2.93	735.0	0.0

```
[ ] # Размер датасета
data.shape
```

```
[ ] (178, 14)
```

```
[ ] # Список колонок
data.columns
```

```
[ ] 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')
```

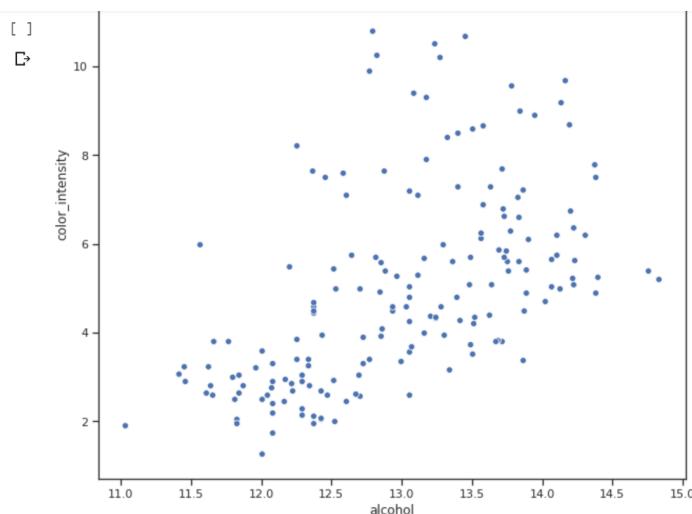
```
[ ] # Список колонок с типами данных
data.dtypes
```

```
[ ] 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
```

```
[ ] # Основные статистические характеристики набора данных  
data.describe()
```

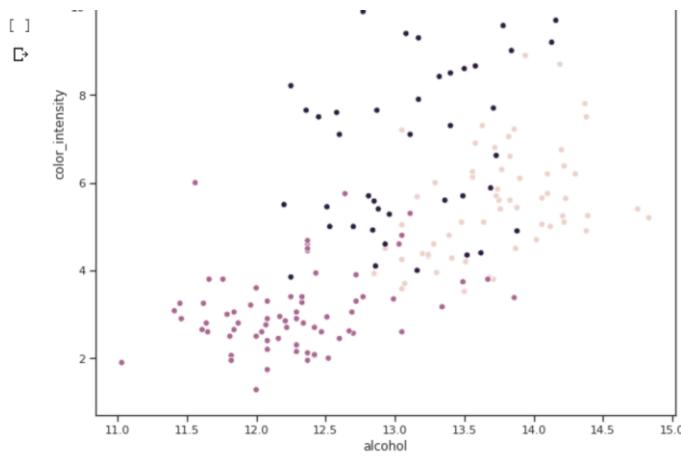
	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intens
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.0581
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.00

```
[ ] #Диаграмма рассеяния  
fig, ax = plt.subplots(figsize=(10,10))  
sns.scatterplot(ax=ax, x='alcohol', y='color_intensity', data=data)
```

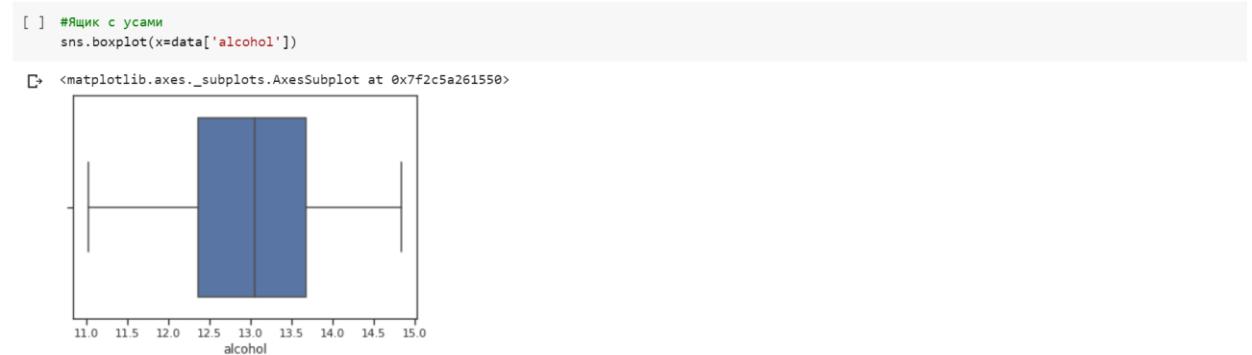
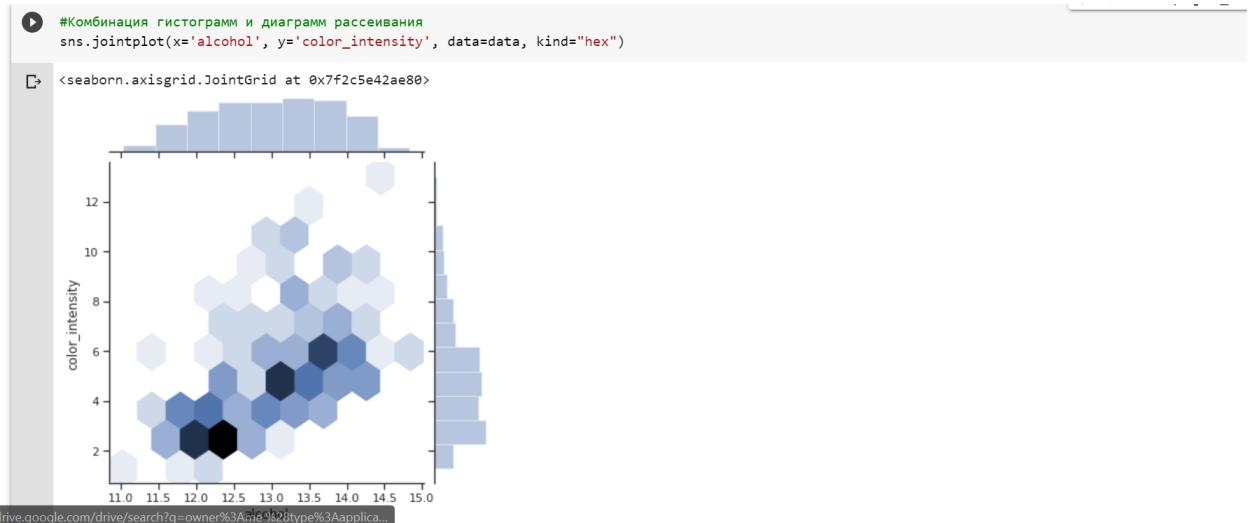
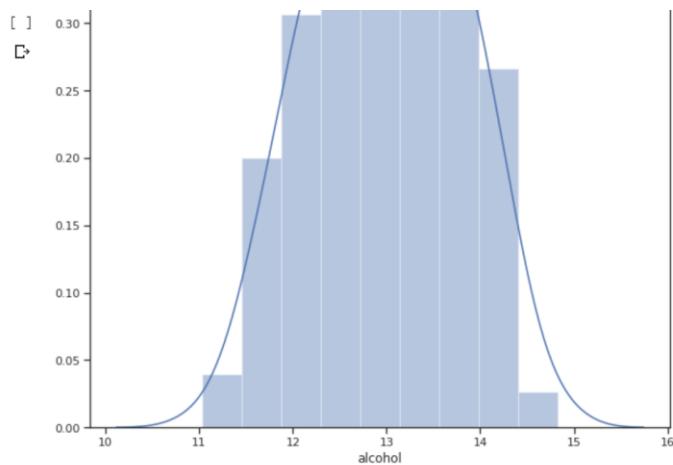


Зависимости между алкоголем и интенсивностью цвета не наблюдается.

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

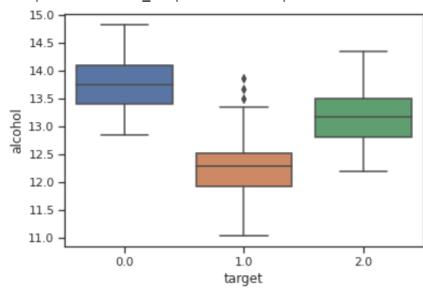


```
[ ] #Гистограмма  
fig, ax = plt.subplots(figsize=(10,10))  
sns.distplot(data['alcohol'])
```



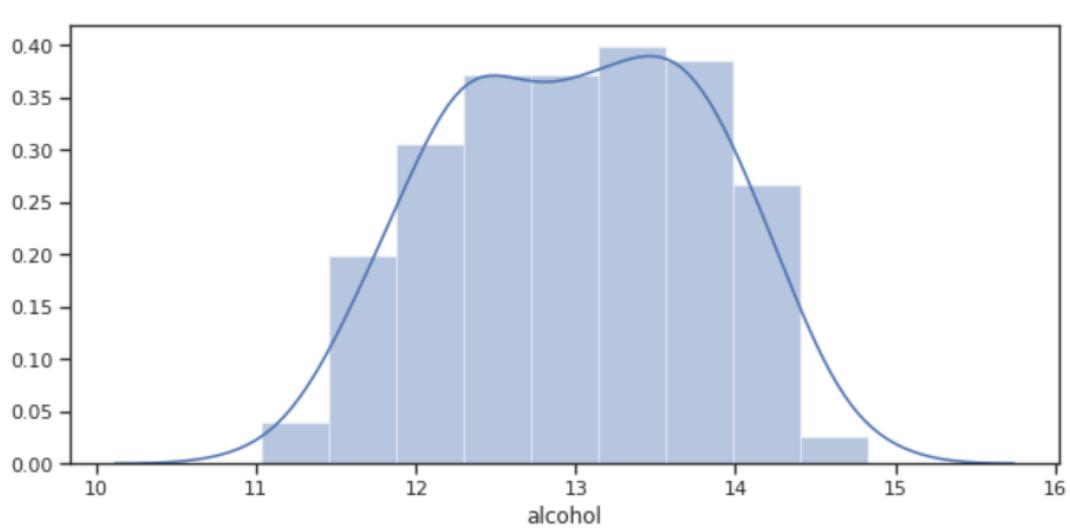
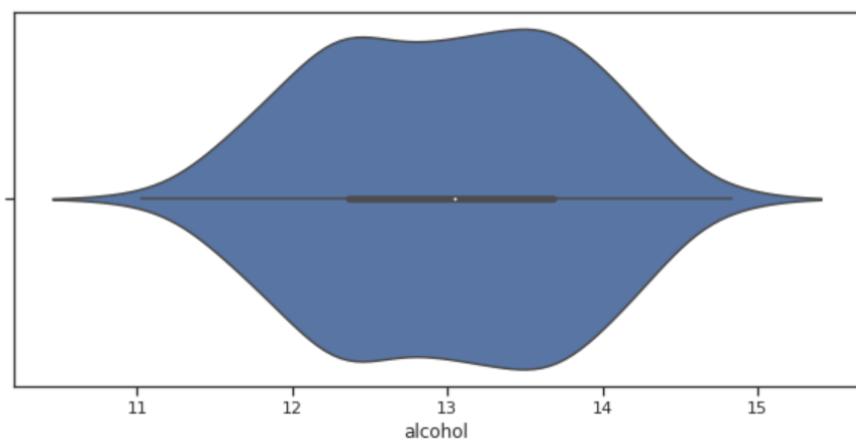
```
[ ] # Распределение параметра alcohol сгруппированные по target.  
sns.boxplot(x='target', y='alcohol', data=data)
```

```
↳ <matplotlib.axes._subplots.AxesSubplot at 0x7f2c5a289f98>
```



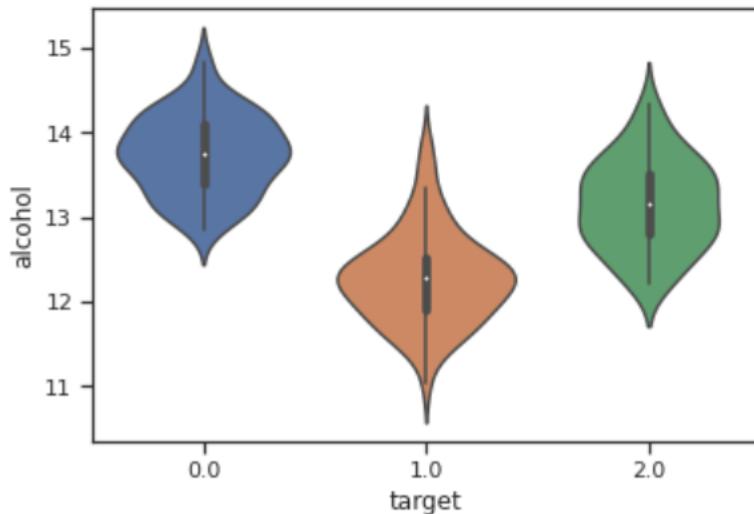
```
▶ fig, ax = plt.subplots(2, 1, figsize=(10,10))  
sns.violinplot(ax=ax[0], x=data['alcohol'])  
sns.distplot(data['alcohol'], ax=ax[1])
```

```
↳ <matplotlib.axes._subplots.AxesSubplot at 0x7f2c58912978>
```



```
[ ] # Распределение параметра alcohol сгруппированные по target.  
sns.violinplot(x='target', y='alcohol', data=data)
```

```
↳ <matplotlib.axes._subplots.AxesSubplot at 0x7f2c5889b898>
```



```
[ ] #Проверка корреляции  
data.corr()
```

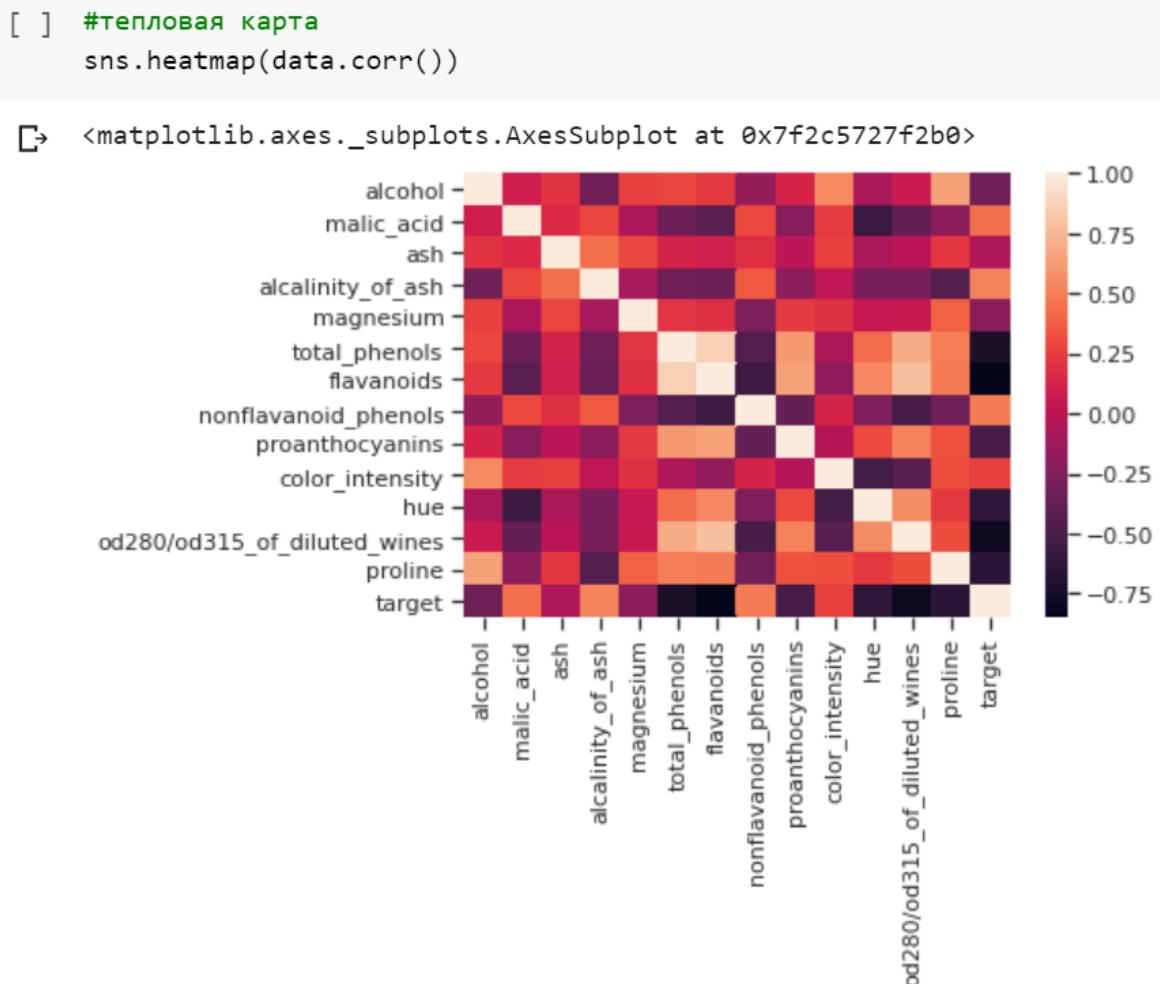
```
↳
```

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthoc
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0..
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.i
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0..
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0..
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.i
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.i
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0..
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1..
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0..
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0..
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0..
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0..
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0..

```
[ ] data.corr(method='pearson')
```

```
↳
```

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocy
alcohol	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.13
malic_acid	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.22
ash	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.0C
alcalinity_of_ash	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.1%
magnesium	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.2%
total_phenols	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.61
flavanoids	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.6%
nonflavanoid_phenols	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.3%
proanthocyanins	0.136698	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1.0C
color_intensity	0.546364	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0.02
hue	-0.071747	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0.2%
od280/od315_of_diluted_wines	0.072343	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0.51
proline	0.643720	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0.3%
target	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.4%



```
[ ] # Треугольный вариант матрицы
mask = np.zeros_like(data.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
sns.heatmap(data.corr(), mask=mask)
```

↪ <matplotlib.axes.\_subplots.AxesSubplot at 0x7f2c5710acc0>

