МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа №1
по курсу «Методы машинного обучения»

Тема: «Разведочный анализ данных. Исследование и визуализация данных»

ИСПОЛНИТЕЛЬ: _Бабин В.Е.________
ФИО

группа ИУ5-22 ____________________
подпись

"__"_________2020 г.

ПРЕПОДАВАТЕЛЬ: ____________________
ФИО

____
подпись

"__"_________2020 г.

Москва - 2020

[2]:
```python
from sklearn.datasets import
load_wine import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Dataset about wines
data = load_wine()

# To the table
columns = data['feature_names']
data = data['data']
table_basic = dict()
for j in range(len(columns)):
    table_basic[columns[j]] = [data[i][j] for i in range(len(data))]

dataset = pd.DataFrame(table_basic, index=range(len(data)))

# Got the
dataset dataset
```

[2]:
```
     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
0                            3.92  1065.0
1                            3.40  1050.0
2                            3.17  1185.0
3                            3.45  1480.0
4                            2.93   735.0
..                            …      …
173                          1.74   740.0
174                          1.56   750.0
175                          1.56   835.0
176                          1.62   840.0
177                          1.60   560.0

[178 rows x 13 columns]
```

[3]:
```python
# Check null values at dataset
for col in dataset.columns:
    temp_null_count = dataset[dataset[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
```

[4]:
```python
# Key Statistical Characteristics of a Dataset
dataset.describe()
```

[4]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium \
count	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573
std	0.811827	1.117146	0.274344	3.339564	14.282484
min	11.030000	0.740000	1.360000	10.600000	70.000000
25%	12.362500	1.602500	2.210000	17.200000	88.000000
50%	13.050000	1.865000	2.360000	19.500000	98.000000
75%	13.677500	3.082500	2.557500	21.500000	107.000000
max	14.830000	5.800000	3.230000	30.000000	162.000000

	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins \
count	178.000000	178.000000	178.000000	178.000000
mean	2.295112	2.029270	0.361854	1.590899
std	0.625851	0.998859	0.124453	0.572359
min	0.980000	0.340000	0.130000	0.410000
25%	1.742500	1.205000	0.270000	1.250000
50%	2.355000	2.135000	0.340000	1.555000
75%	2.800000	2.875000	0.437500	1.950000
max	3.880000	5.080000	0.660000	3.580000

	color_intensity	hue	od280/od315_of_diluted_wines	proline
count	178.000000	178.000000	178.000000	178.000000
mean	5.058090	0.957449	2.611685	746.893258
std	2.318286	0.228572	0.709990	314.907474
min	1.280000	0.480000	1.270000	278.000000
25%	3.220000	0.782500	1.937500	500.500000
50%	4.690000	0.965000	2.780000	673.500000
75%	6.200000	1.120000	3.170000	985.000000
max	13.000000	1.710000	4.000000	1680.000000

[5]:

```
# Correlation data
dataset.corr()
```

[5]:

	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

```
                              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

                              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

                              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

                              od280/od315_of_diluted_winesproline
```

```
alcohol                                           0.0723430.643720
malic_acid                                       -0.368710 -0.192011
ash                                               0.0039110.223626
alcalinity_of_ash                                -0.276769 -0.440597
magnesium                                         0.0660040.393351
total_phenols                                     0.6999490.498115
flavanoids                                        0.7871940.494193
nonflavanoid_phenols                             -0.503270 -0.311385
proanthocyanins                                   0.5190670.330417
color_intensity                                  -0.4288150.316100
hue                                               0.5654680.236183
od280/od315_of_diluted_wines                      1.0000000.312761
proline                                           0.3127611.000000
```

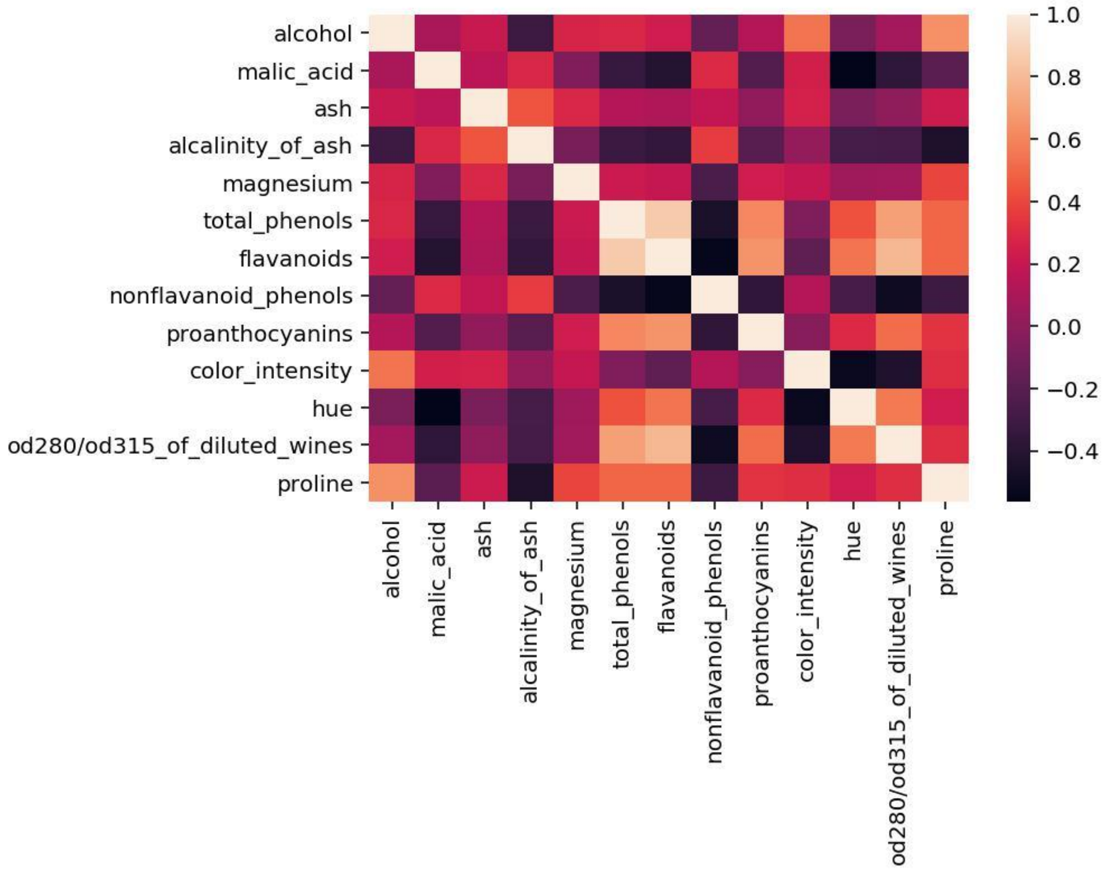
[6]:

```
# Heat correlation diagram
sns.heatmap(dataset.corr())
```

[6]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa3a2422710>
```

[6]:

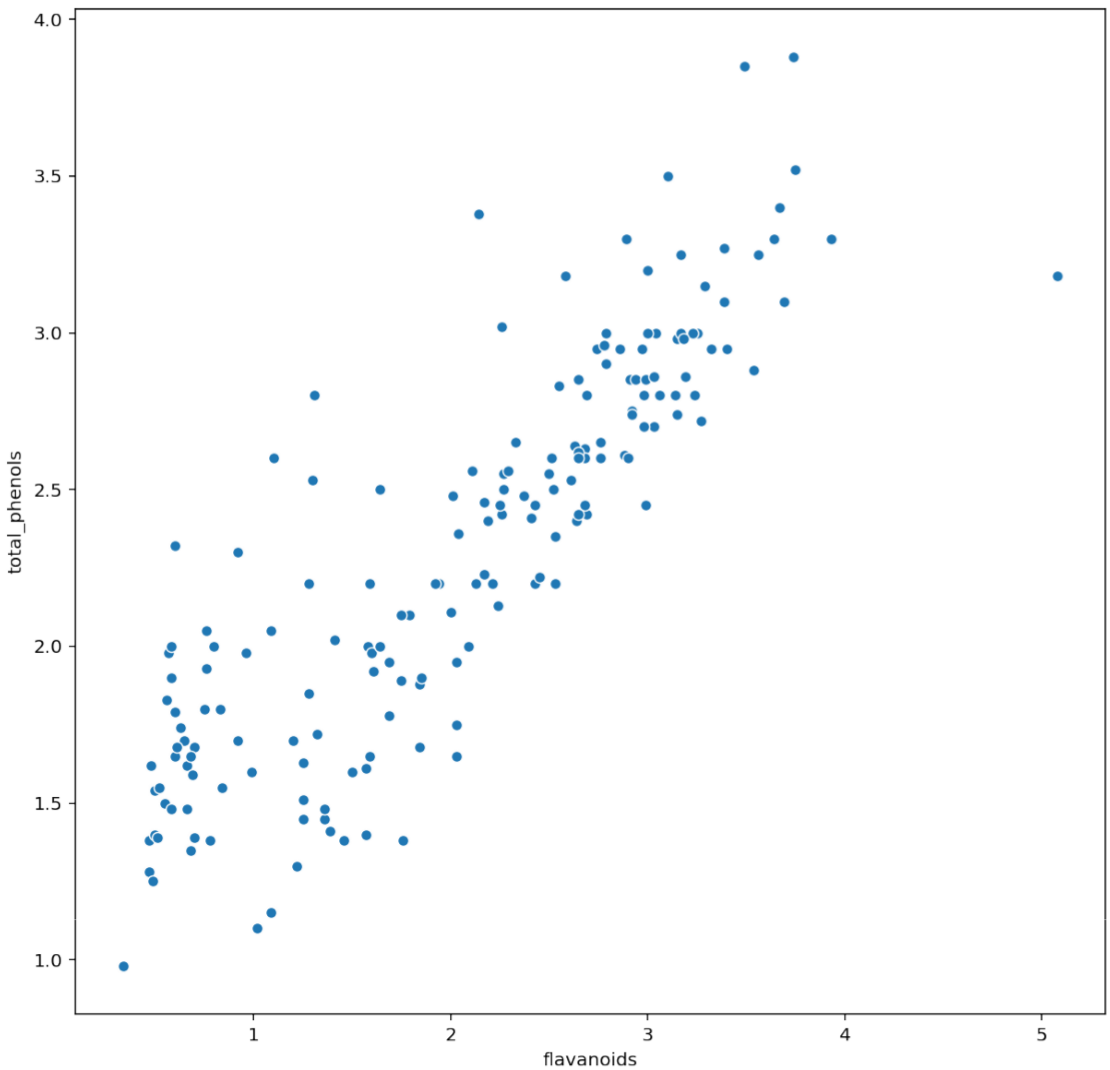


[7]:
```python
 # From the correlation matrix and the heat diagram is clear,
# that more interconnected is attrs flavanoids and total_phenols

# Dispersion chart for these attrs
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='flavanoids', y='total_phenols', data=dataset)
```

[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa39e2f6940>
```

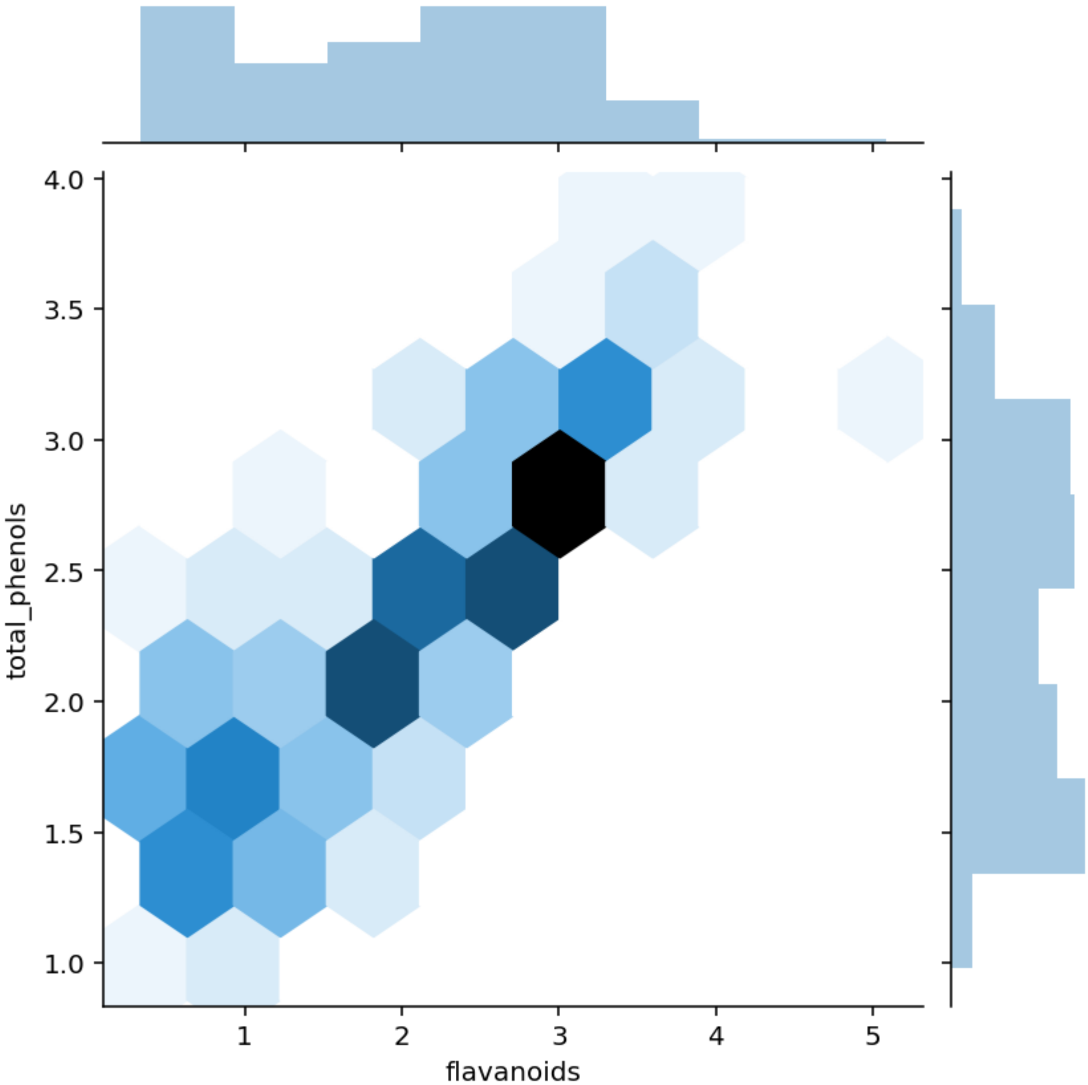
[7]:



[10]:
```python
sns.jointplot(x='flavanoids', y='total_phenols', data=dataset, kind="hex")
```

[10]:
```
<seaborn.axisgrid.JointGrid at 0x7fa38da4e208>
```

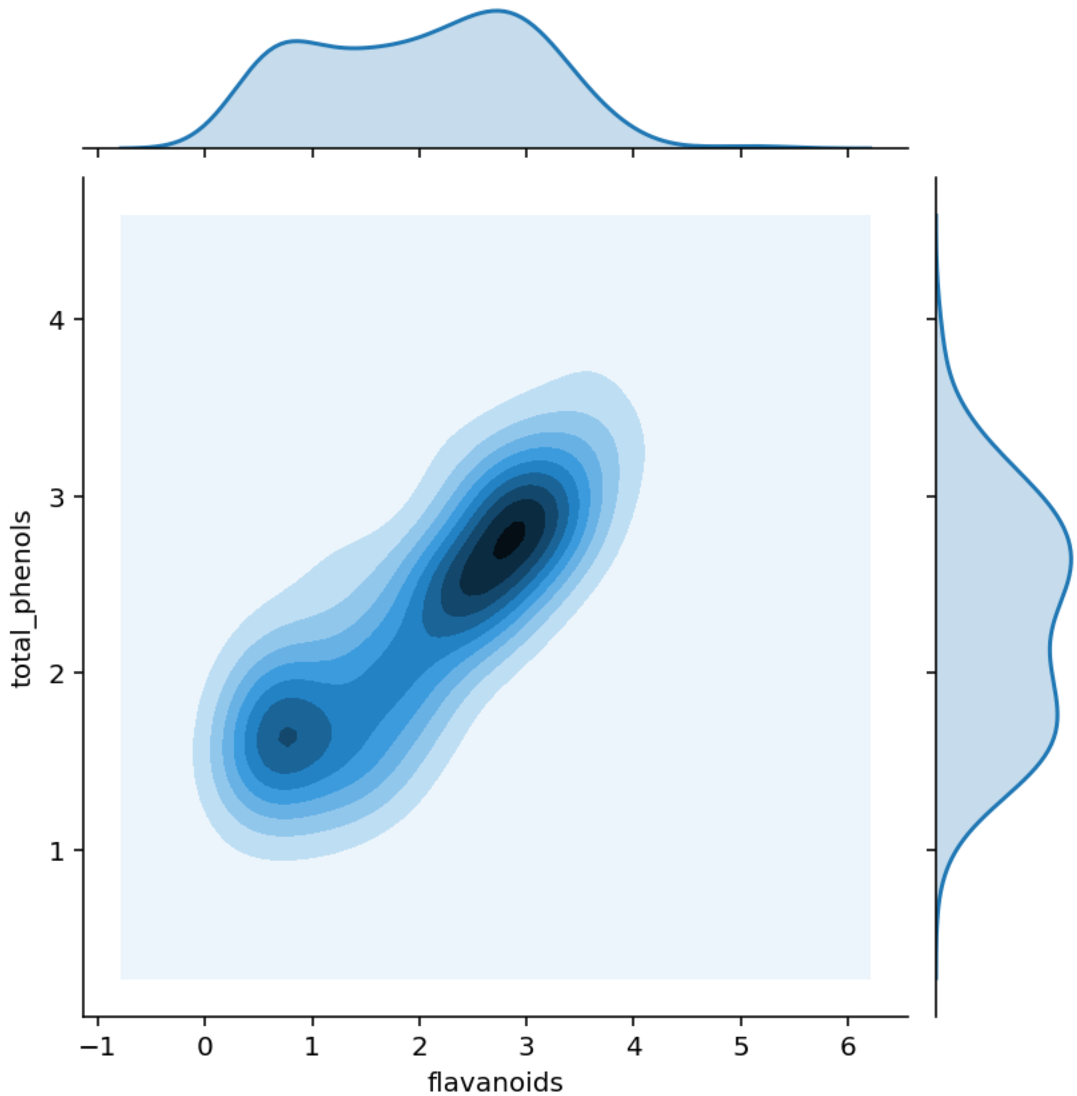
[10]:



[12]:
```
sns.jointplot(x='flavanoids', y='total_phenols', data=dataset, kind="kde")
```

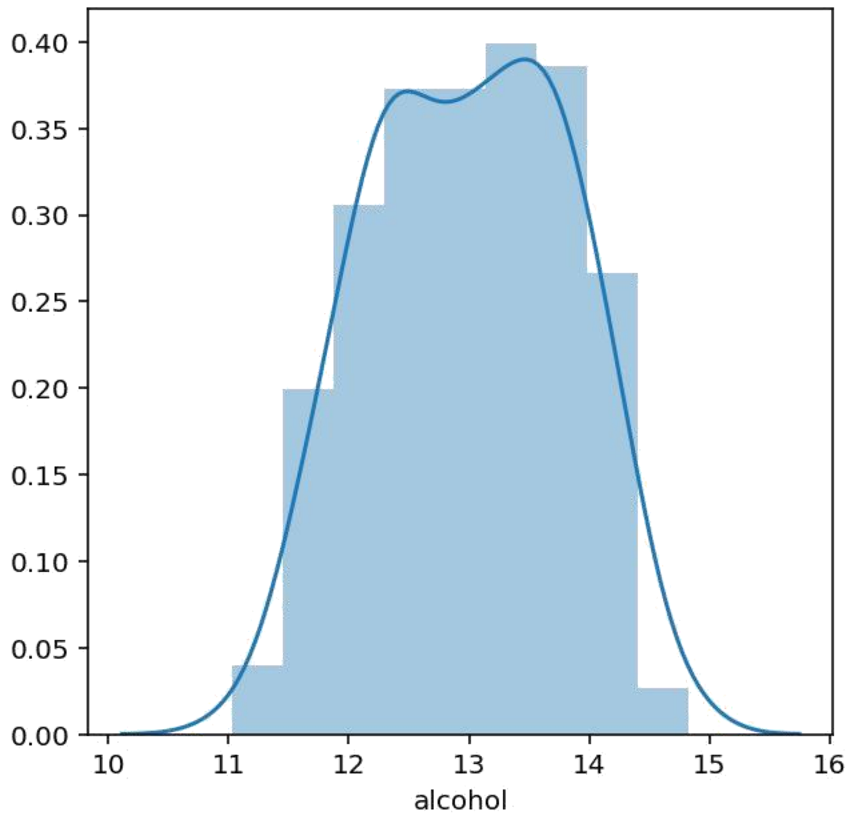
[12]:
```
<seaborn.axisgrid.JointGrid at 0x7fa392c562e8>
```

[12]:

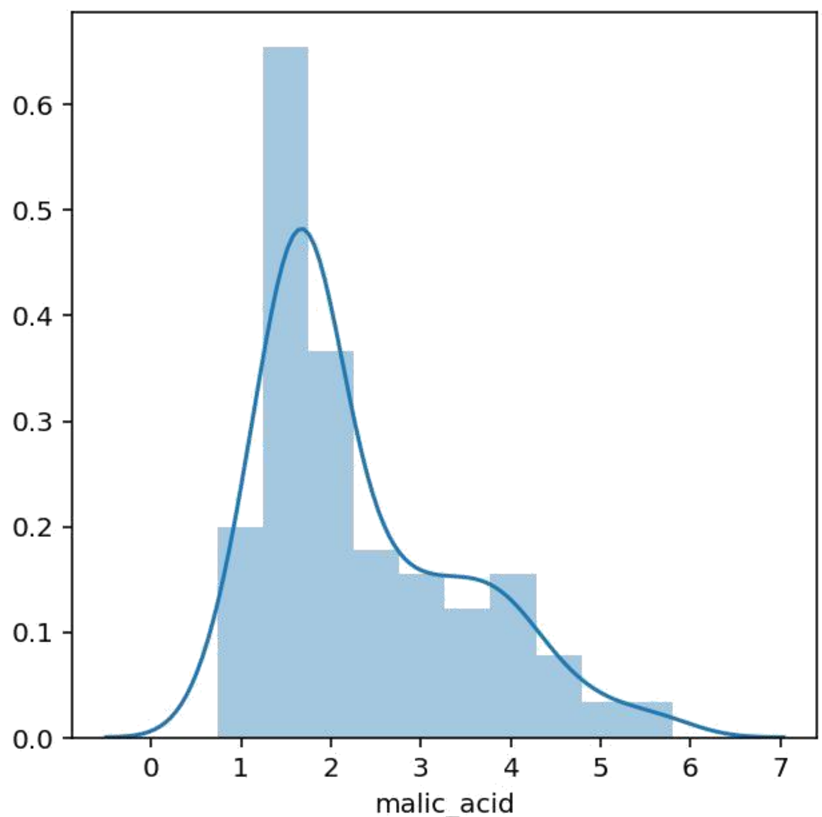


[8]:
```
 # distribtion of features
for column in columns:
    fig, ax = plt.subplots(figsize=(5,5))
    sns.distplot(dataset[column])
```

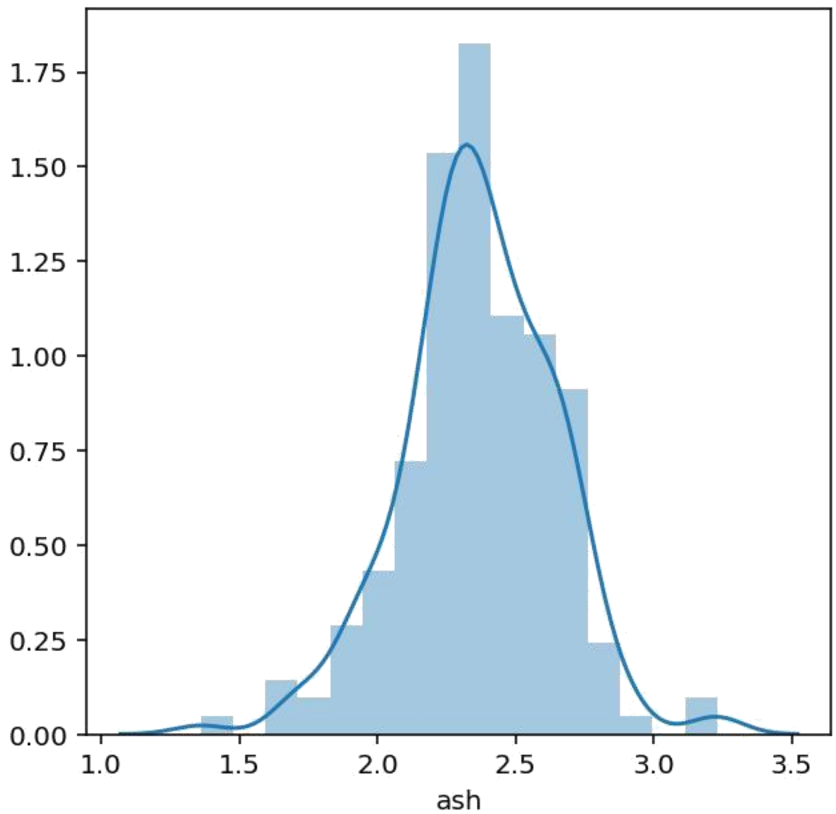
[8]:



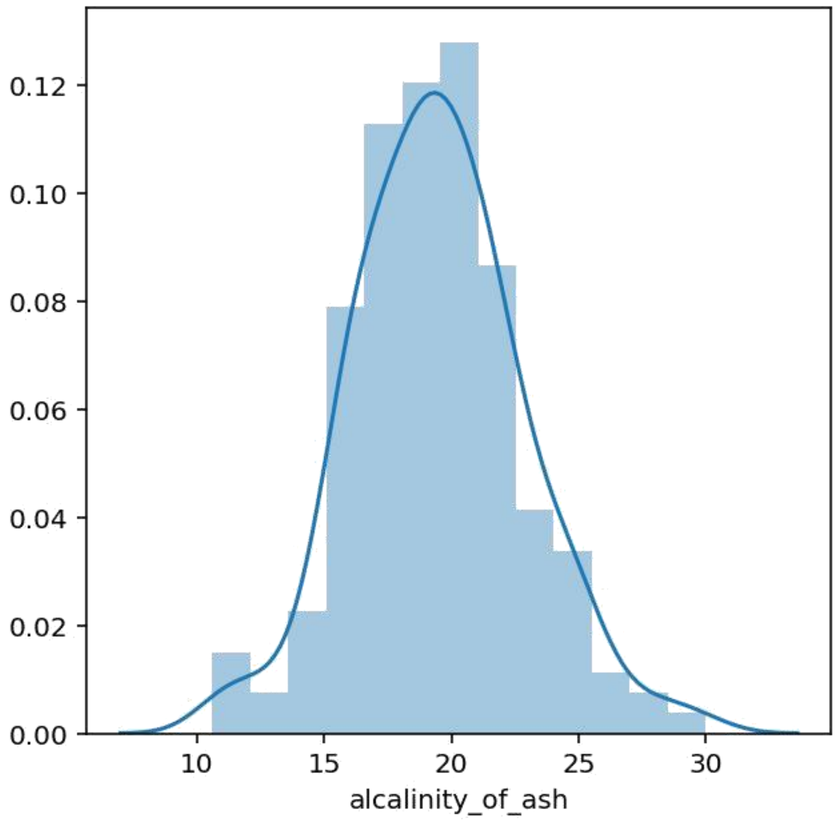
[8]:



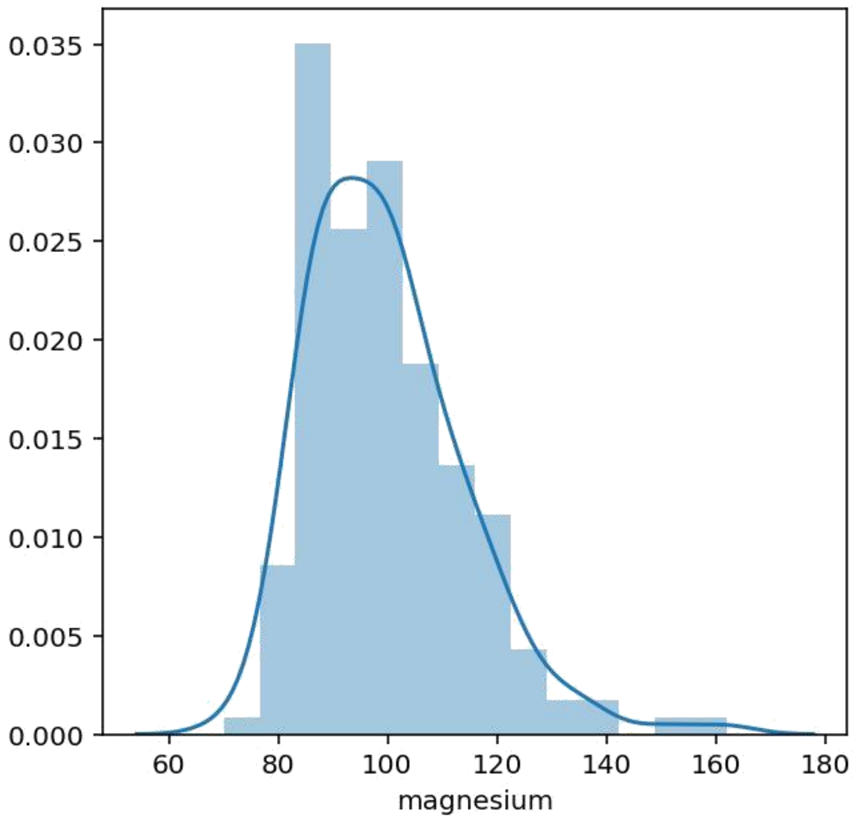
[8]:



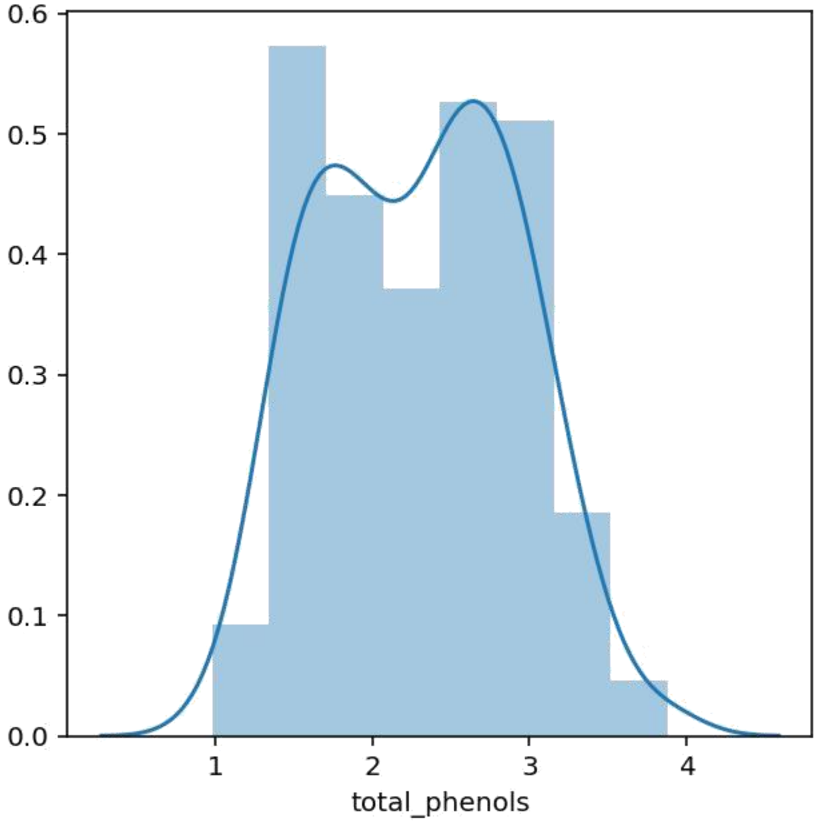
[8]:



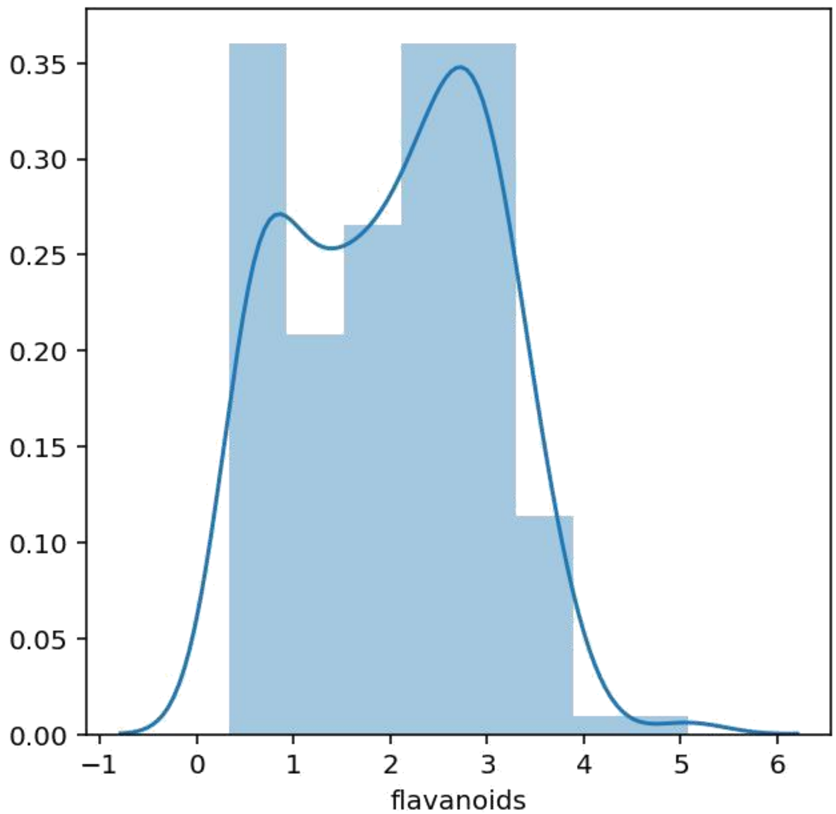
[8]:


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
60
80
100
120
140
160
180
magnesium

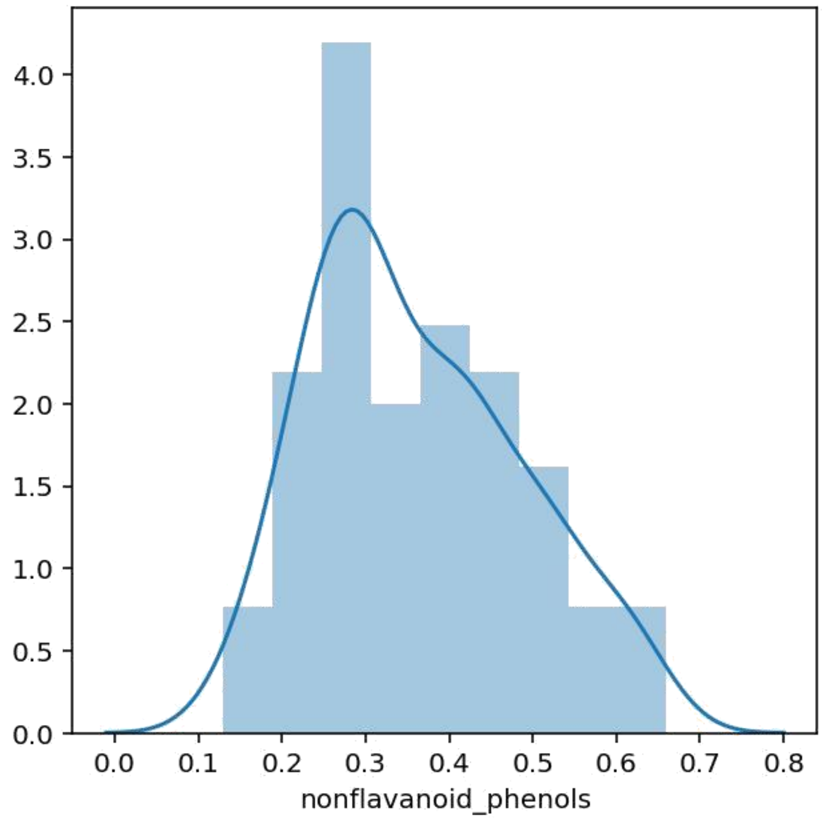
[8]:



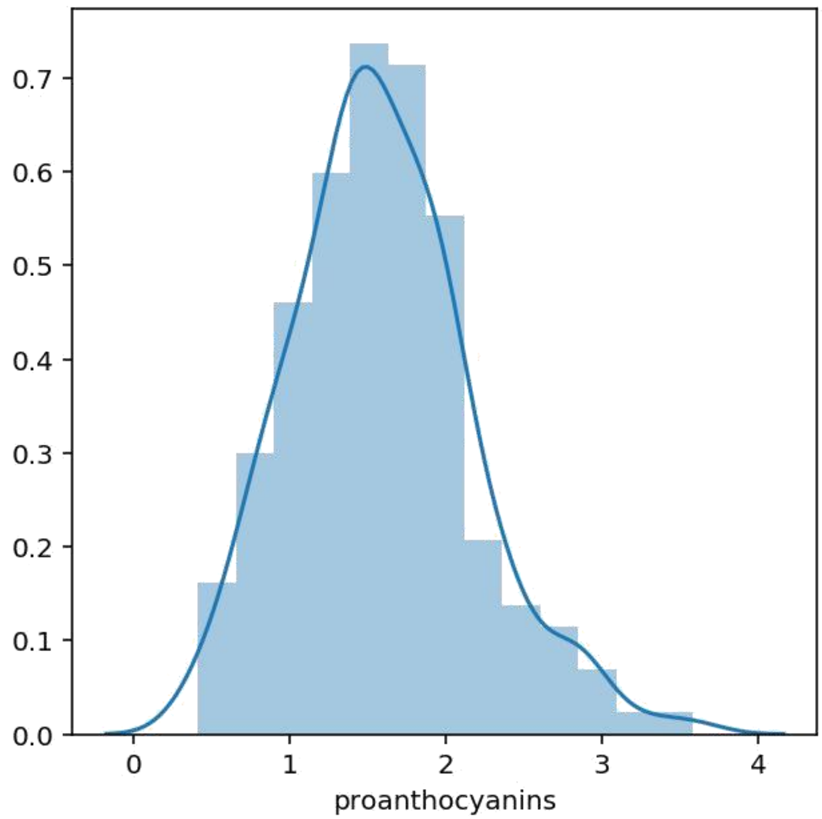
[8]:



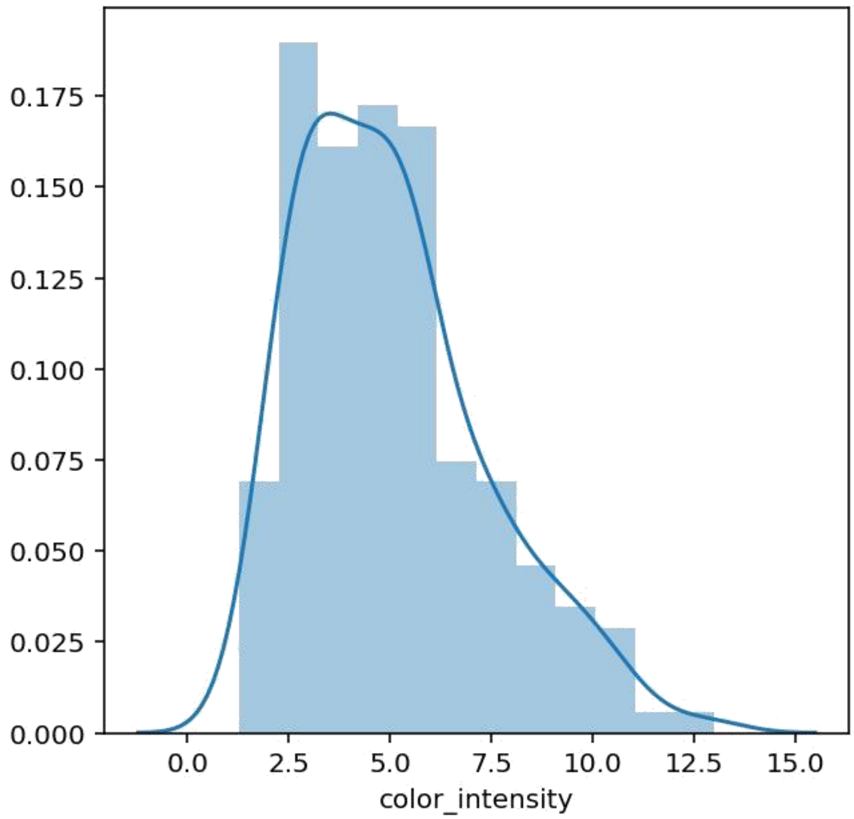
[8]:


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
nonflavanoid_phenols

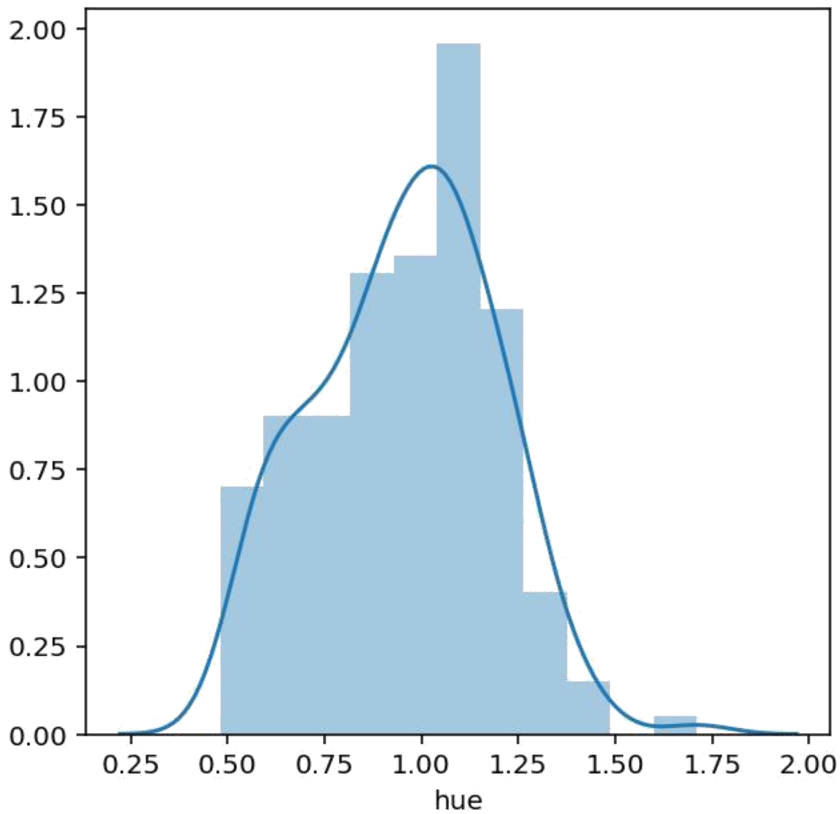
[8]:



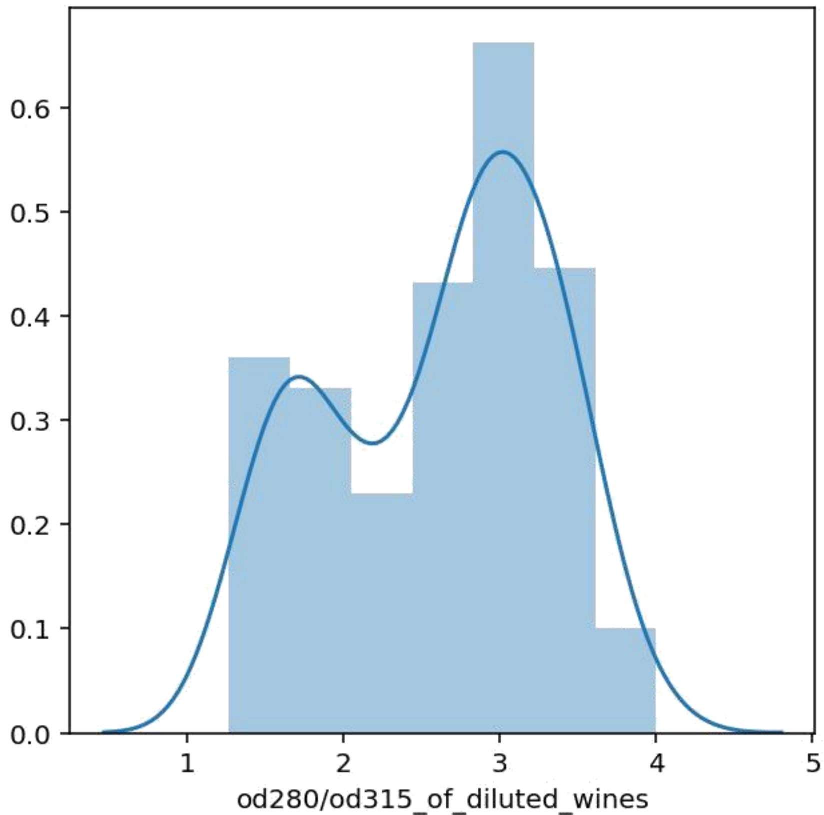
[8]:


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
color_intensity

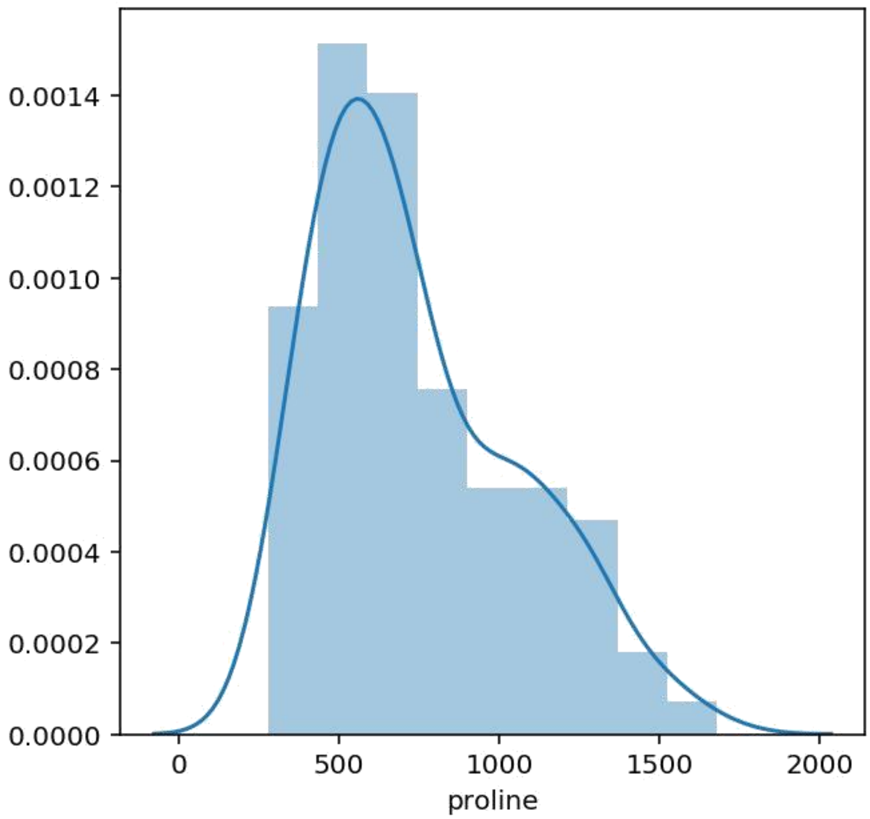
[8]:


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.25 0.50 0.75 1.00 1.25 1.50 1.75 2.00
hue

[8]:



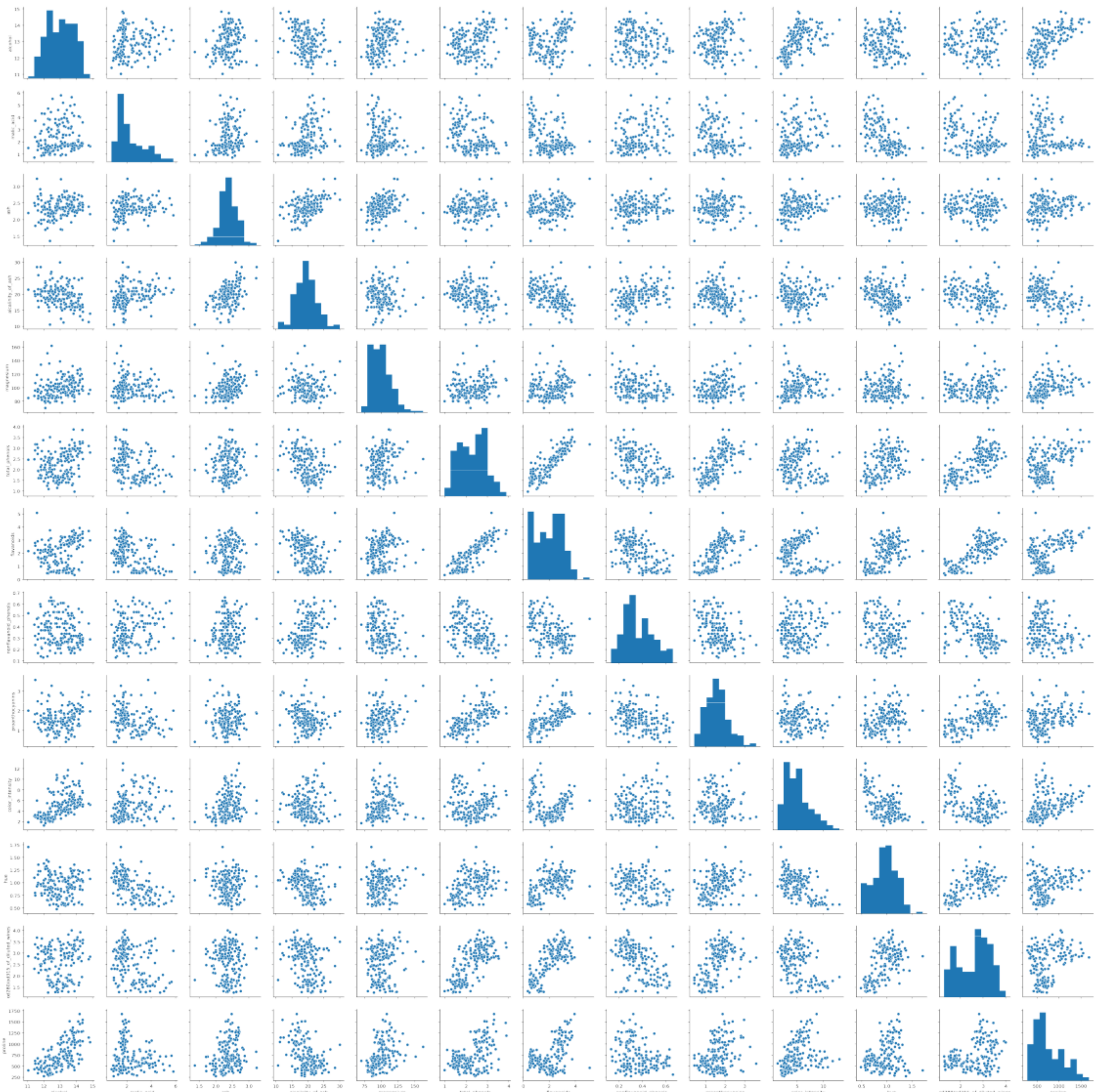
[8]:



[9]:
```
  # Pair diagrams
sns.pairplot(dataset)
```

[9]:
```
<seaborn.axisgrid.PairGrid at 0x7fa39c051cf8>
```

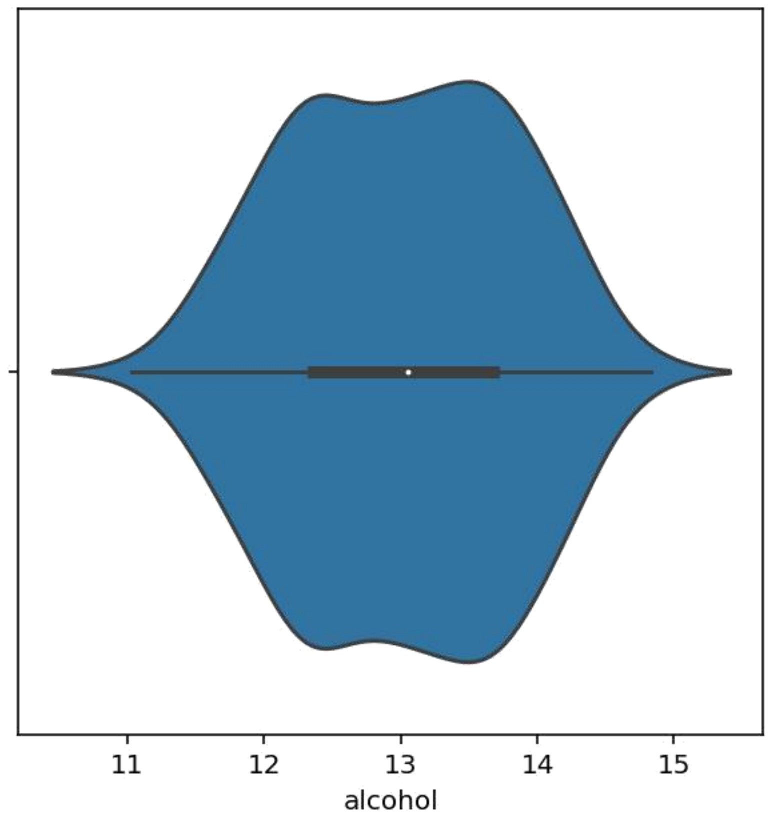
[9]:



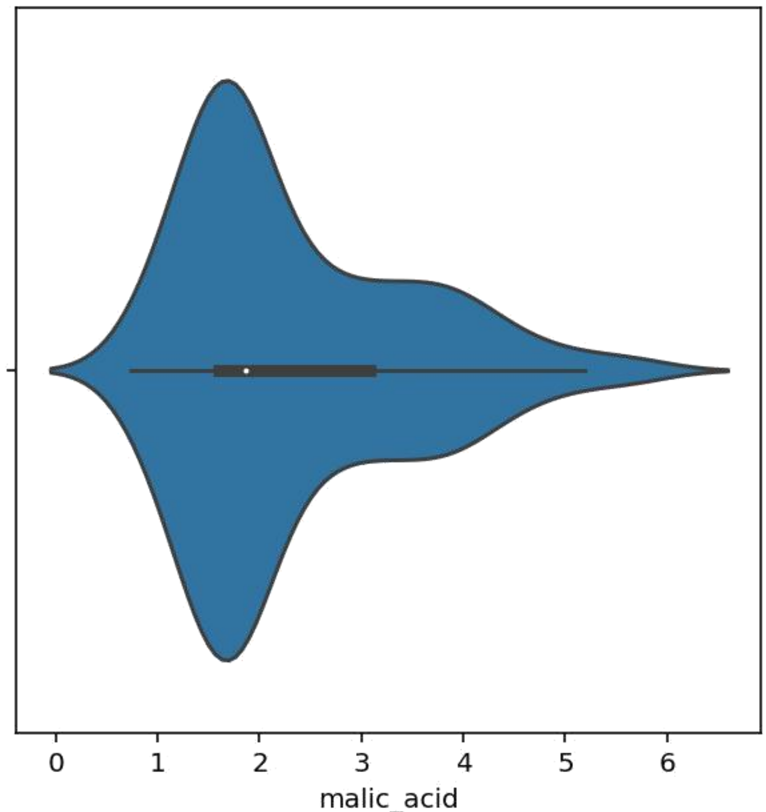
[15]:

```
for column in columns:
    fig, ax = plt.subplots(figsize=(5,5))
    sns.violinplot(x=dataset[column])
```

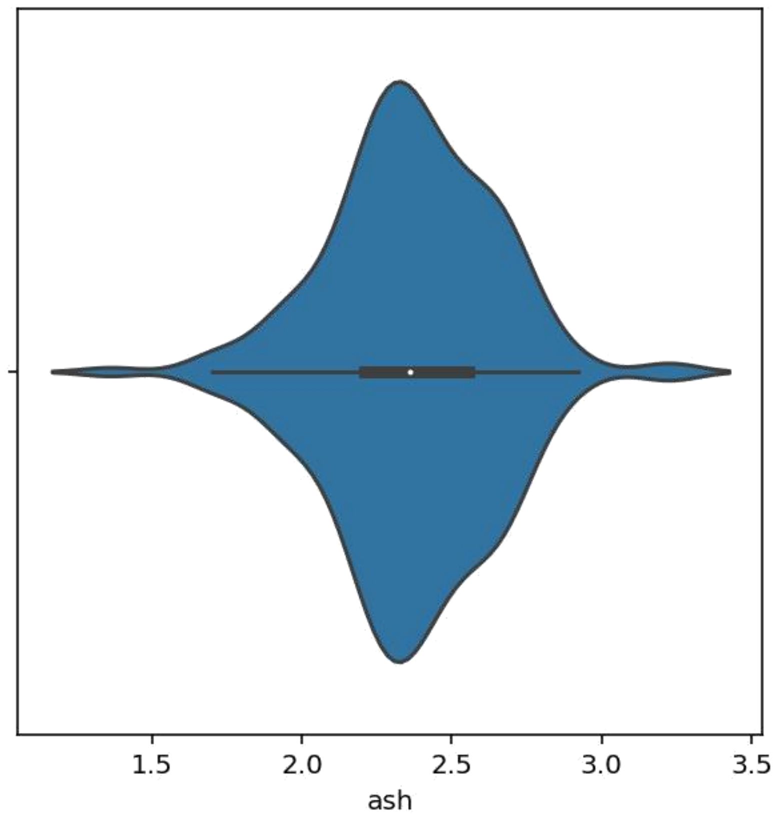
[15]:



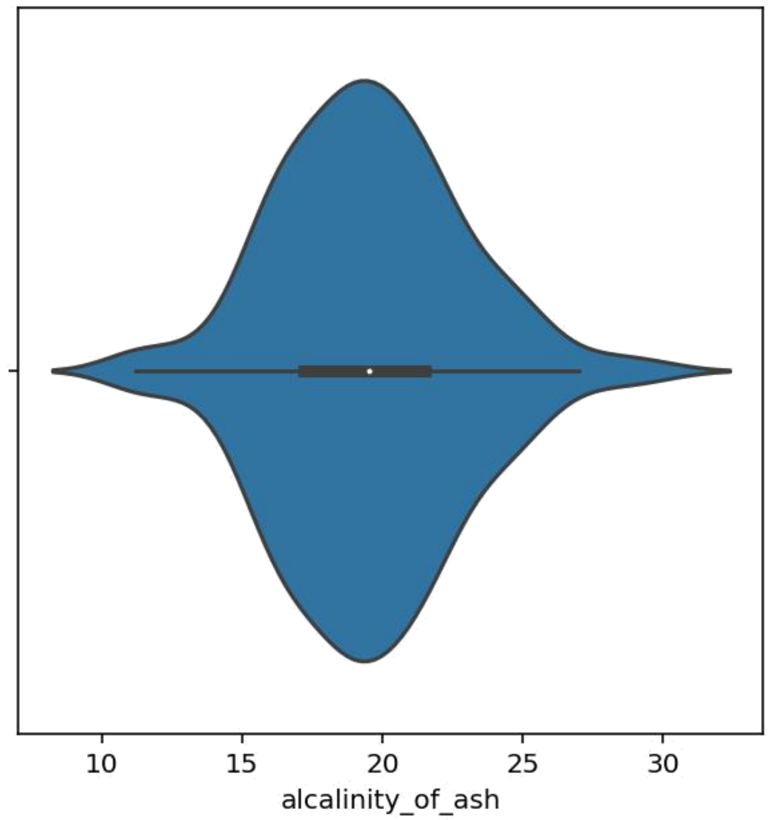
[15]:



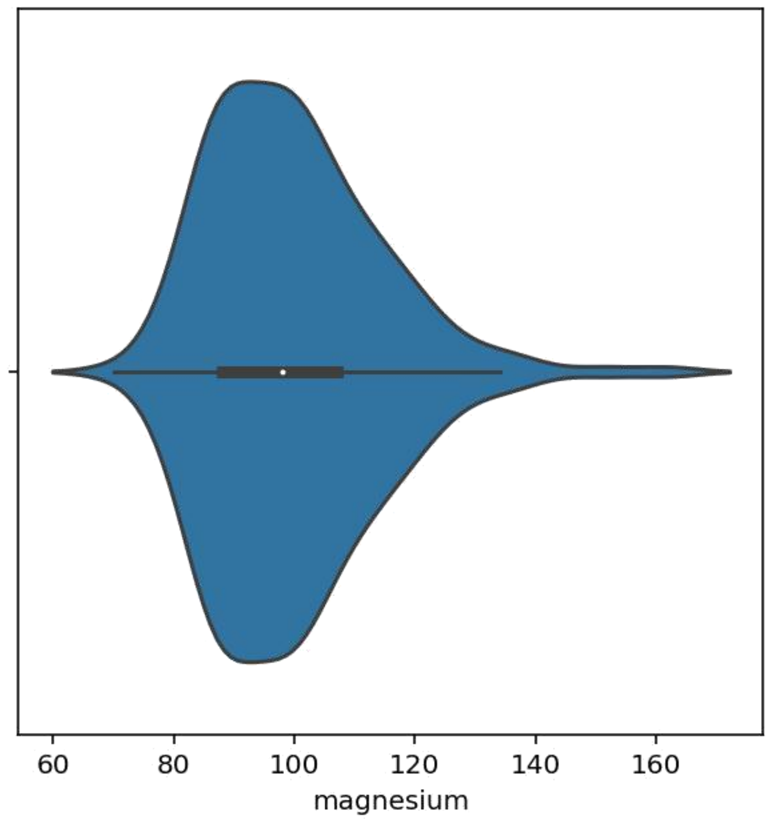
[15]:



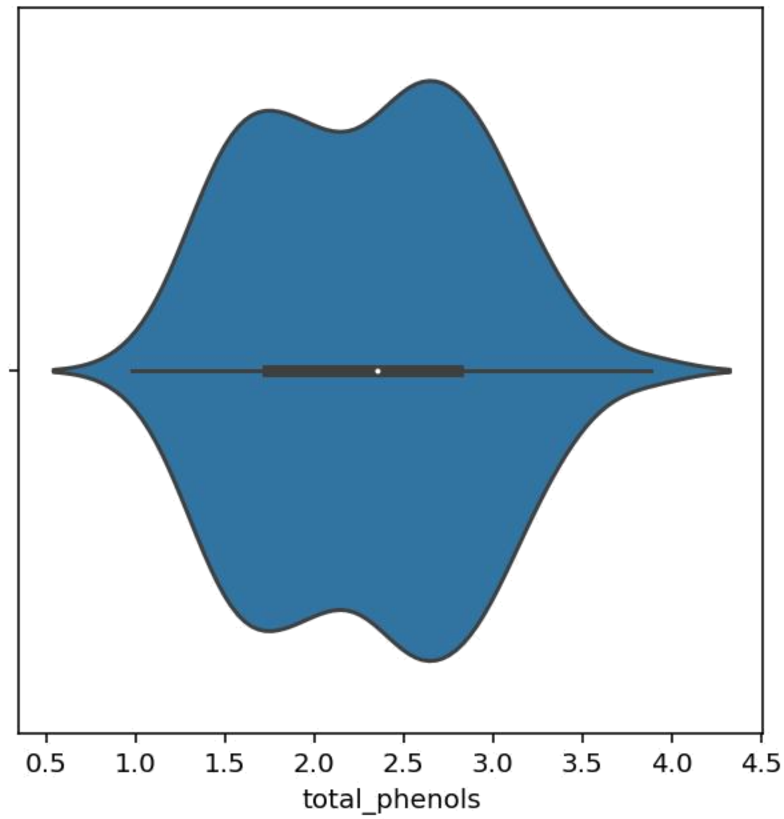
[15]:



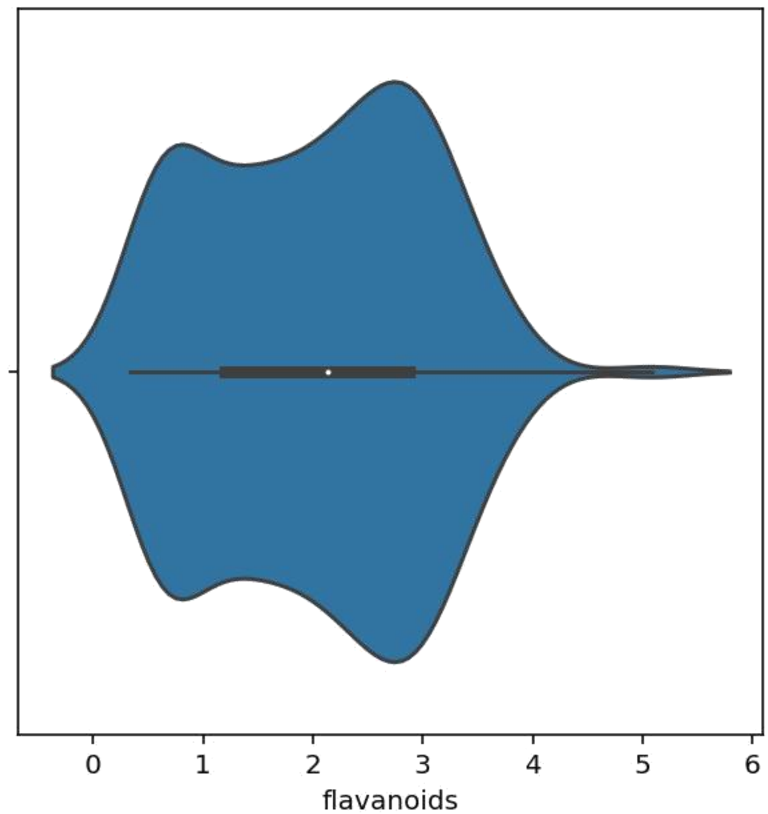
[15]:



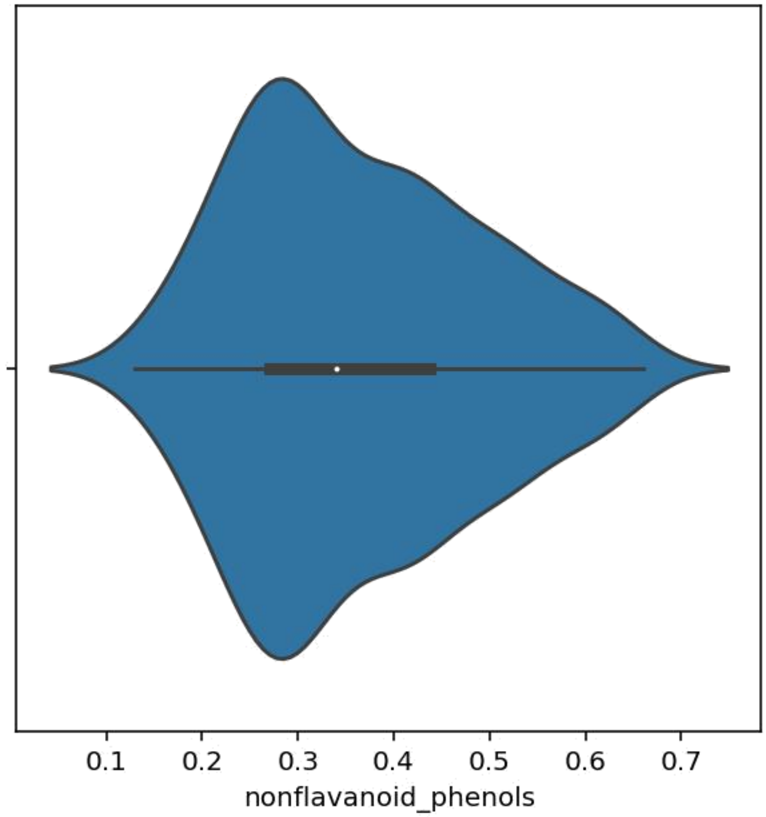
[15]:



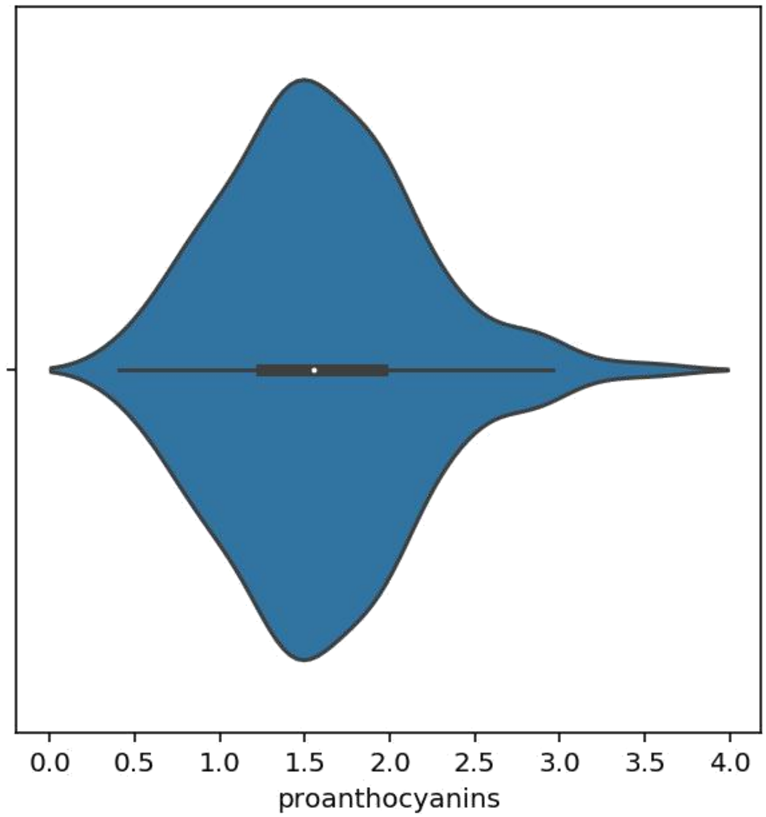
[15]:



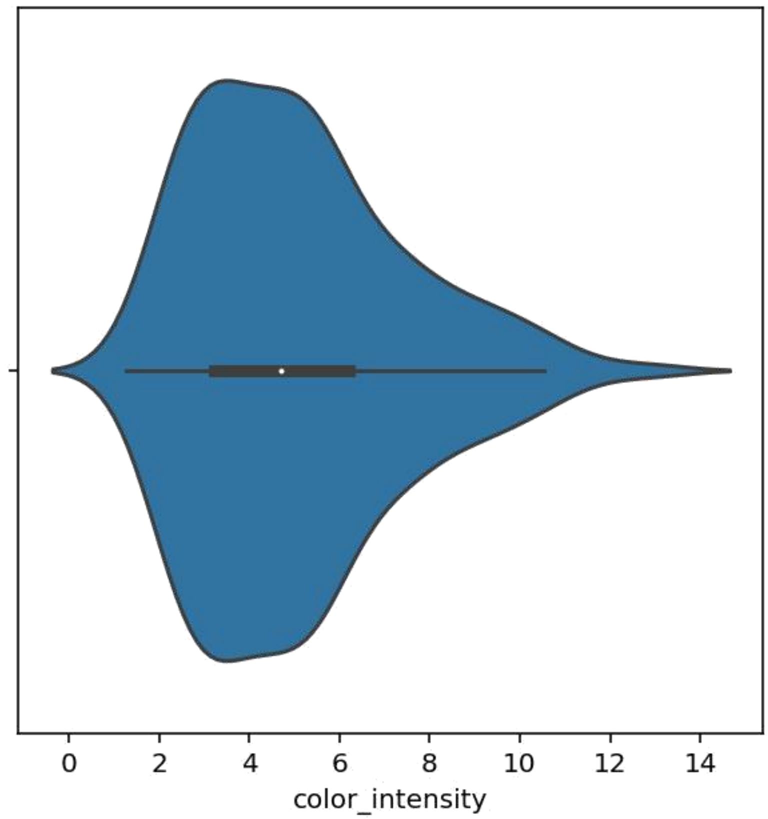
[15]:



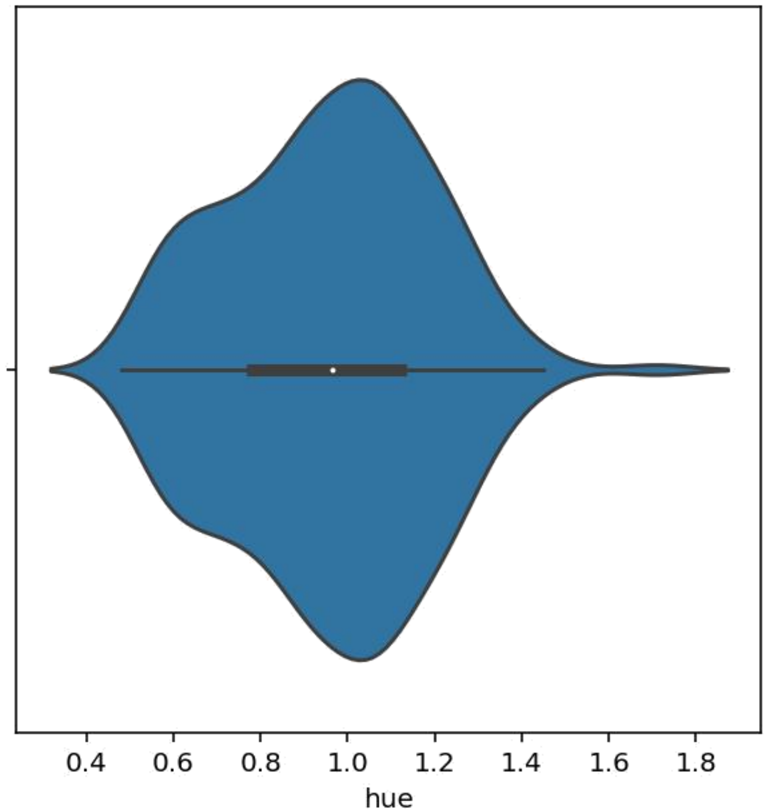
[15]:



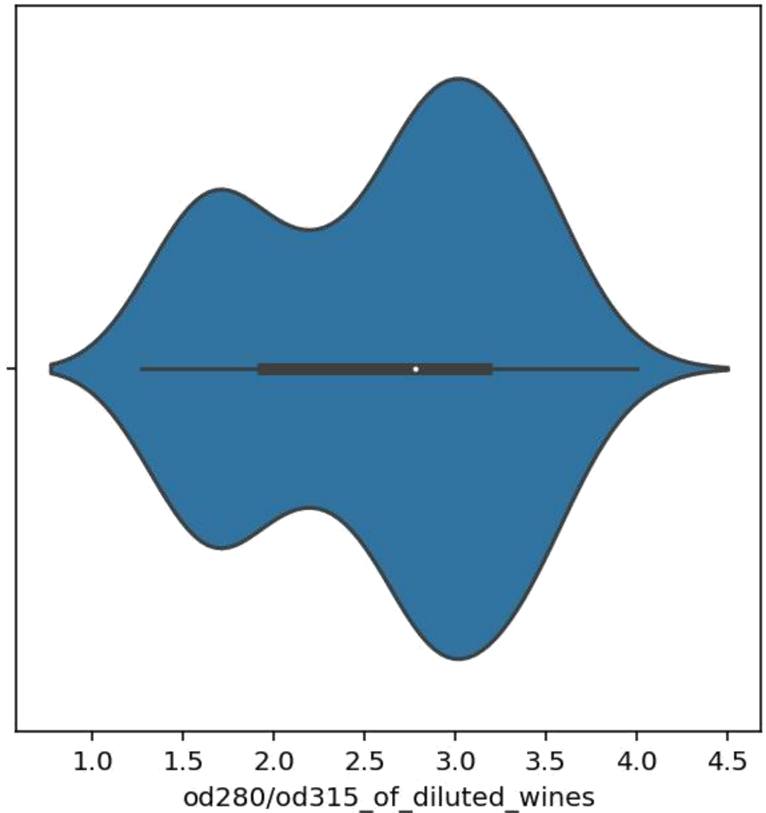
[15]:



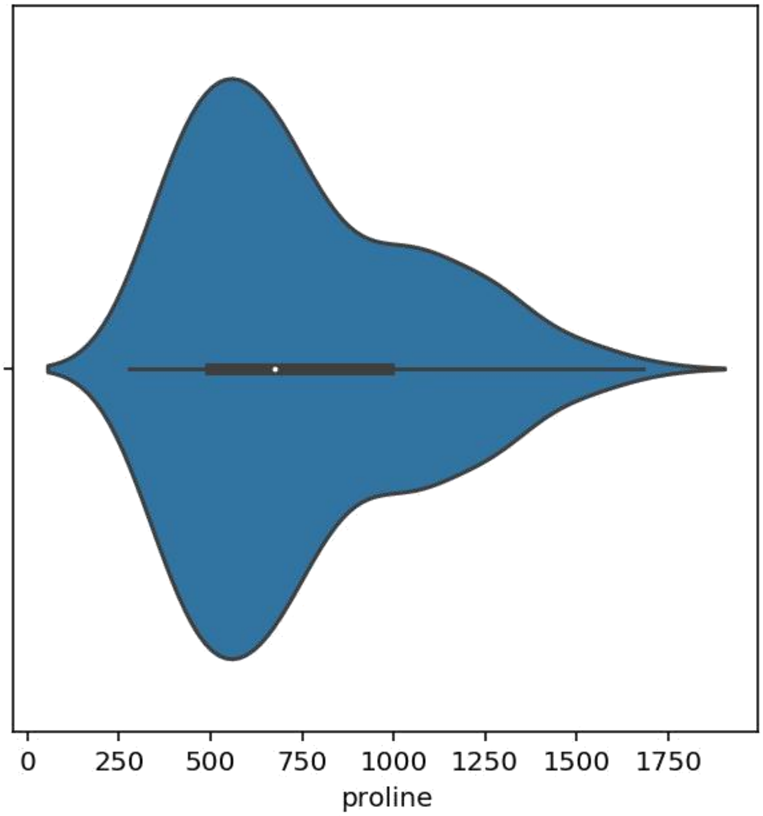
[15]:



[15]:



[15]:




```
[0]:
```