Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет

Лабораторная работа № 1

По курсу «Технологии машинного обучения»

«Разведочный анализ данных. Исследование и визуализация данных»

ИСПОЛНИТЕЛЬ:

Харчевников
Александр
Группа ИУ5-64

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

_____

"__"_____2020 г.

Москва 2020

1. Цель работы

Изучение различных методов визуализация данных.

2. Описание задания

- Выбрать набор данных
- Создать ноутбук, который содержит следующие разделы:
    1. Текстовое описание выбранного наборы данных
    2. Основные характеристики датасета
    3. Визуальное исследование датасета
    4. Информация о корелляции признаков
- Сформировать отчет и разместить его на своем репозитории GitHub

3. Текст программы и экранные формы с примерами выполнения

1. Текстовое описание выбранного набора данных

Исследуемый набор данных - https://scikit-learn.org/stable/datasets/index.html#wine-dataset (https://scikit-learn.org/stable/datasets/index.html#wine-dataset).

Данные представляют собой результаты химического анализа вин, выращенных в одном регионе Италии тремя различными культиваторами. Существует тринадцать различных измерений (содержание алкоголя, интенсивность цвета, оттенок и др.), проведенных для разных компонентов, найденных в трех типах вина.

In [39]:

```
from sklearn.datasets import load_wine
import numpy as np
import pandas as pd
```

2. Основные характеристики датасета

In [53]:

```
dataset = load_wine()
df = pd.DataFrame(dataset.data, columns=dataset.feature_names)
df.head()
```

Out[53]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflava
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	

Размер датасета

In [23]:

```
df.shape
```

Out[23]:

```
(178, 13)
```

Список колонок с типами данных

In [26]:

```
df.dtypes
```

Out[26]:

```
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
dtype: object
```

Проверка на наличие пустых значений

In [27]:

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

Основные статистические характеристки набора данных

In [28]:

```
df.describe()
```

Out[28]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flava
count	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	2.0
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.9
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.3
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.2
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.1
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.8
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.0

Уникальные значения для целевого признака (содержание алкоголя)

In [37]:

```
df['alcohol'].unique()
```

Out[37]:

```
array([14.23, 13.2 , 13.16, 14.37, 13.24, 14.2 , 14.39, 14.06, 14.8
3,
       13.86, 14.1 , 14.12, 13.75, 14.75, 14.38, 13.63, 14.3 , 13.8
3,
       14.19, 13.64, 12.93, 13.71, 12.85, 13.5 , 13.05, 13.39, 13.3
,
       13.87, 14.02, 13.73, 13.58, 13.68, 13.76, 13.51, 13.48, 13.2
8,
       13.07, 14.22, 13.56, 13.41, 13.88, 14.21, 13.9 , 13.94, 13.8
2,
       13.77, 13.74, 13.29, 13.72, 12.37, 12.33, 12.64, 13.67, 12.1
7,
       13.11, 13.34, 12.21, 12.29, 13.49, 12.99, 11.96, 11.66, 13.0
3,
       11.84, 12.7 , 12.  , 12.72, 12.08, 12.67, 12.16, 11.65, 11.6
4,
       12.69, 11.62, 12.47, 11.81, 12.6 , 12.34, 11.82, 12.51, 12.4
2,
       12.25, 12.22, 11.61, 11.46, 12.52, 11.76, 11.41, 11.03, 12.7
7,
       11.45, 11.56, 11.87, 12.07, 12.43, 11.79, 12.04, 12.86, 12.8
8,
       12.81, 12.53, 12.84, 13.36, 13.52, 13.62, 12.87, 13.32, 13.0
8,
       12.79, 13.23, 12.58, 13.17, 13.84, 12.45, 14.34, 12.36, 13.6
9,
       12.96, 13.78, 13.45, 12.82, 13.4 , 12.2 , 14.16, 13.27, 14.1
3])
```

In [41]:

```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

3. Визуальное исследование датасета

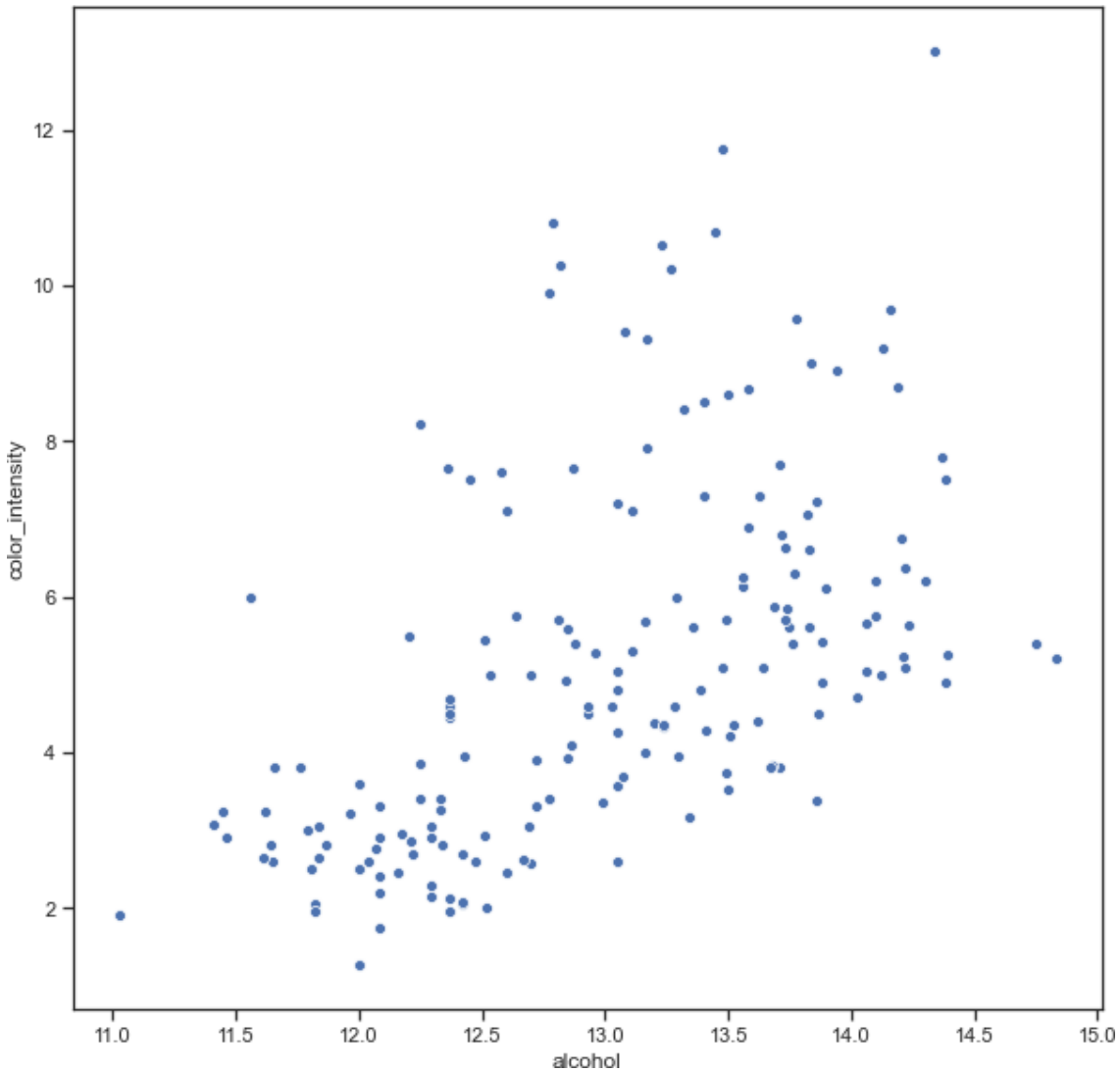
Диаграмма рассеяния

In [52]:

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

Out[52]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x115ccf940>
```



In [ ]:

Можно видеть, что чем выше крепость вина, тем больше интенсивность его цвета.

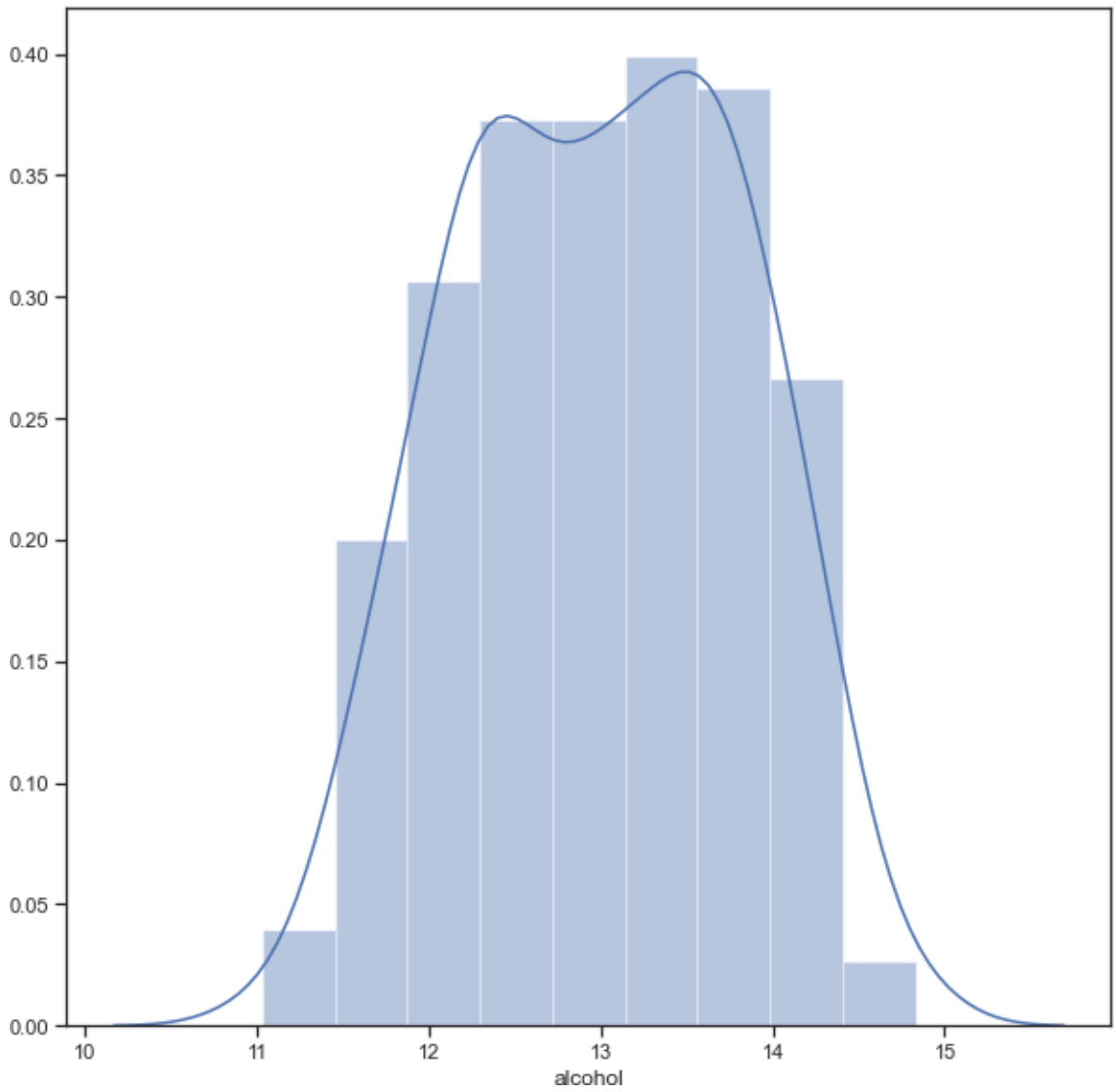
Гистограмма

In [54]:

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

Out[54]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x12210ea58>
```



In [ ]:

Как видим, среднее содержание содержание алкоголя составляет 13%.

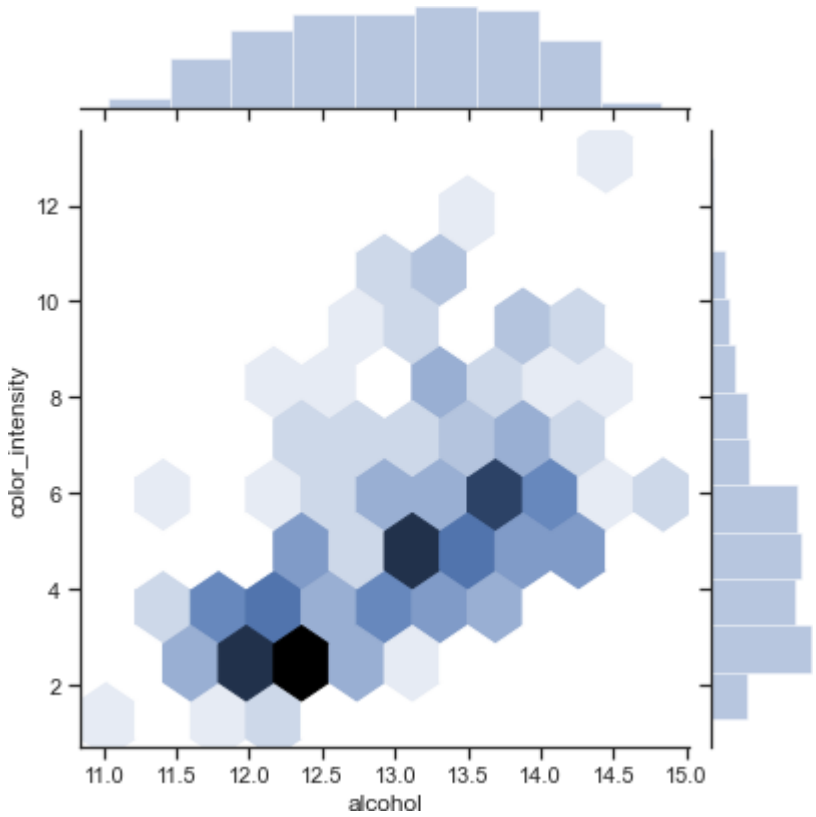
Jointplot - комбинация гистограмм и диаграмм рассеивания

In [58]:

```
sns.jointplot(x='alcohol', y='color_intensity', data=df, kind="hex")
```

Out[58]:

```
<seaborn.axisgrid.JointGrid at 0x115ccae48>
```



Можно сделать вывод, что в основном интенсивность цвета вина от 2 до 6, то есть вин, имеющих неинтенсивный оттенок цвета, больше.

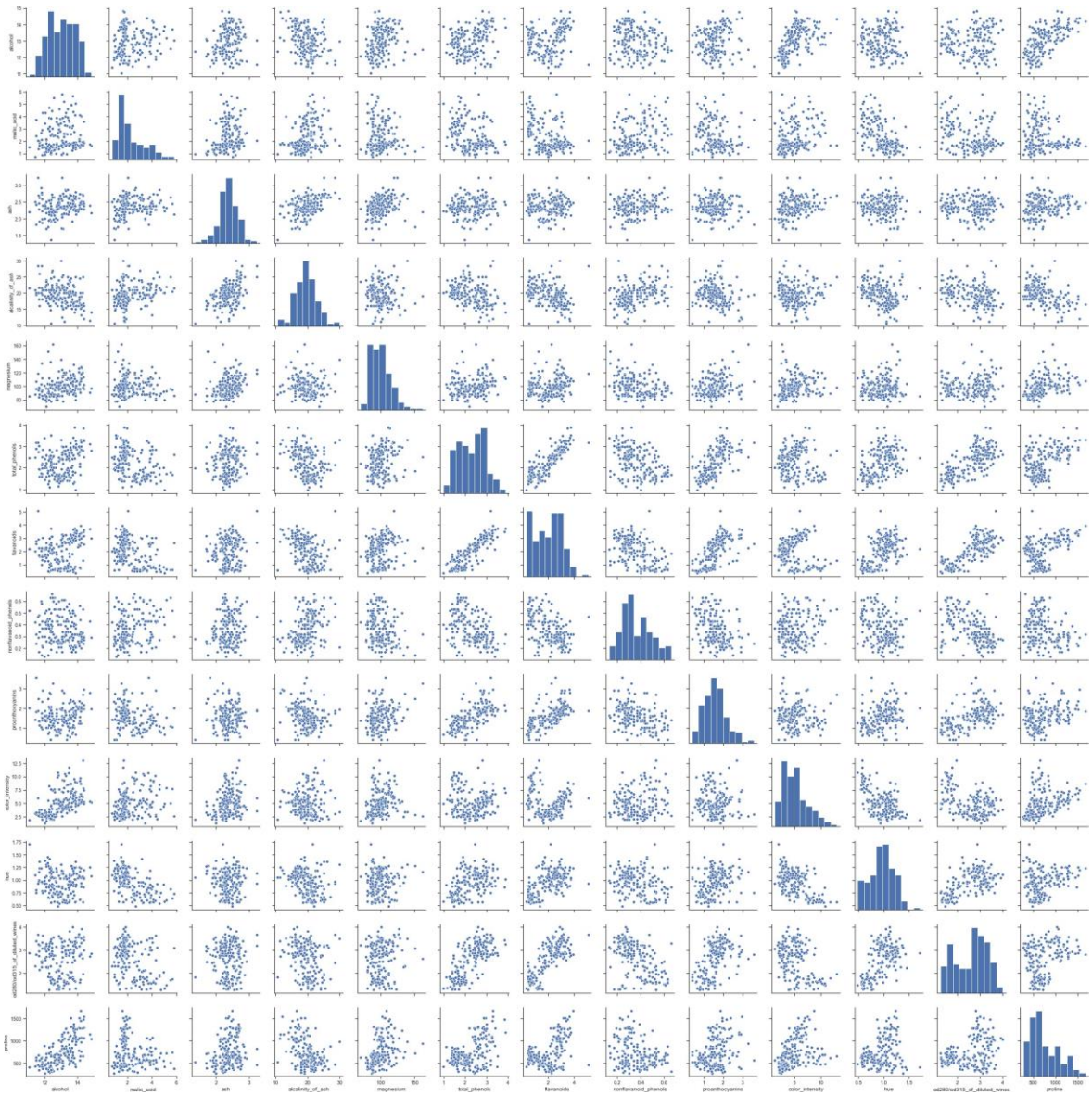
Парные диаграммы

In [60]:

```
sns.pairplot(df)
```

Out[60]:

```
<seaborn.axisgrid.PairGrid at 0x127651da0>
```



Как видим, на некоторых диаграммах наблюдается почти линейная зависимость. Например, в ячейке (7, 6) зависимость flavanoids от total_phenols

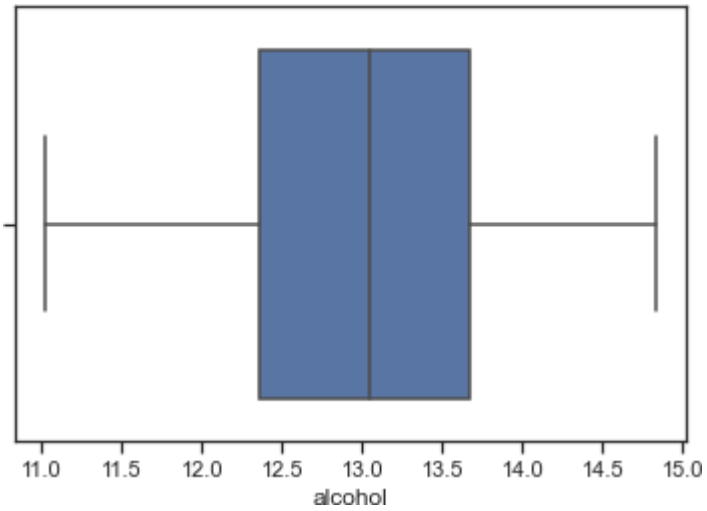
Ящик с усами

In [62]:

```
sns.boxplot(x=df['alcohol'])
```

Out[62]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x12d7d1fd0>
```



Можно сделать вывод, что медиана равна 13, нижний квартиль - 12.3, верхний квартиль - 13.6. Наблюдаемый минимум - 11, наблюдаемый максимум - 14.7.

4. Информация о корреляции признаков

Построим корреляционную матрицу по всему набору данных. Целевой признак - alcohol.

In [63]:

```
df.corr()
```

Out[63]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium
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

Как видим, целевой признак сильнее всего коррелирует с proline (0.64) и color_intensity (0.54).

Heatmap

In [66]:

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

Out[66]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x12f715080>
```