**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет «Радиотехнический»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №1

«Разведочный анализ данных. Исследование и визуализация данных.»

Выполнил: Проверил:

студент группы РТ5-61Б преподаватель каф. ИУ5

Агеев Алексей Гапанюк Ю.Е.

Подпись и дата: Подпись и дата:

Москва, 2023 г

# Описание задания

* Выбрать набор данных (датасет).
* Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из [Scikit-learn.](https://scikit-learn.org/stable/datasets/toy_dataset.html)

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

* Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

* Сформировать отчет и разместить его в своем репозитории на github.

Дополнительно примеры решения задач, содержащие визуализацию, можно посмотреть в репозитории курса mlcourse.ai - <https://github.com/Yorko/mlcourse.ai/wiki/Individual-projects-and-tutorials-(in-Russian)>

# Ход работы

В качестве набора данных используется набор данных химического анализа вин - <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data>

Данные являются результатом химического анализа вин, выращенных в одном и том же регионе Италии тремя разными культиваторами. Существует тринадцать различных измерений различных компонентов, содержащихся в трех типах вина.

## Набор данных содержит следующие параметры: Alcohol - Алкоголь; Acid - Яблочная кислота; Ash - Пепел; Alcalinity of Ash - Щелочность пепла; Magnesium - Магний; Total Phenols - Всего фенолов; Flavanoids - Флавоноиды; Nonflavanoid Phenols - Нефлаваноидные фенолы; Proanthocyanins - Проантоцианы; Colour Intensity - Интенсивность цвета; Hue - Оттенок; OD280/OD315 of diluted wines - OD280/OD315 разбавленных вин; Proline - Пролин.

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

# Набор данных для распознавания вин

wine = load\_wine()  
for x in wine:  
 print(x)

data  
target  
frame  
target\_names  
DESCR  
feature\_names

wine['target\_names']

array(['class\_0', 'class\_1', 'class\_2'], dtype='<U7')

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']

wine['target'].shape

(178,)

data = pd.DataFrame(data= np.c\_[wine['data'], wine['target']],  
 columns= wine['feature\_names'] + ['target'])

data.head()

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

# Размер датасета - 178 строк, 14 колонок  
data.shape

(178, 14)

total\_count = data.shape[0]  
print('Всего строк: {}'.format(total\_count))

Всего строк: 178

# 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

# Проверим наличие пустых значений  
# Цикл по колонкам датасета  
for col in data.columns:  
 # Количество пустых значений - все значения заполнены  
 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

# Основные статистические характеристки набора данных  
data.describe()

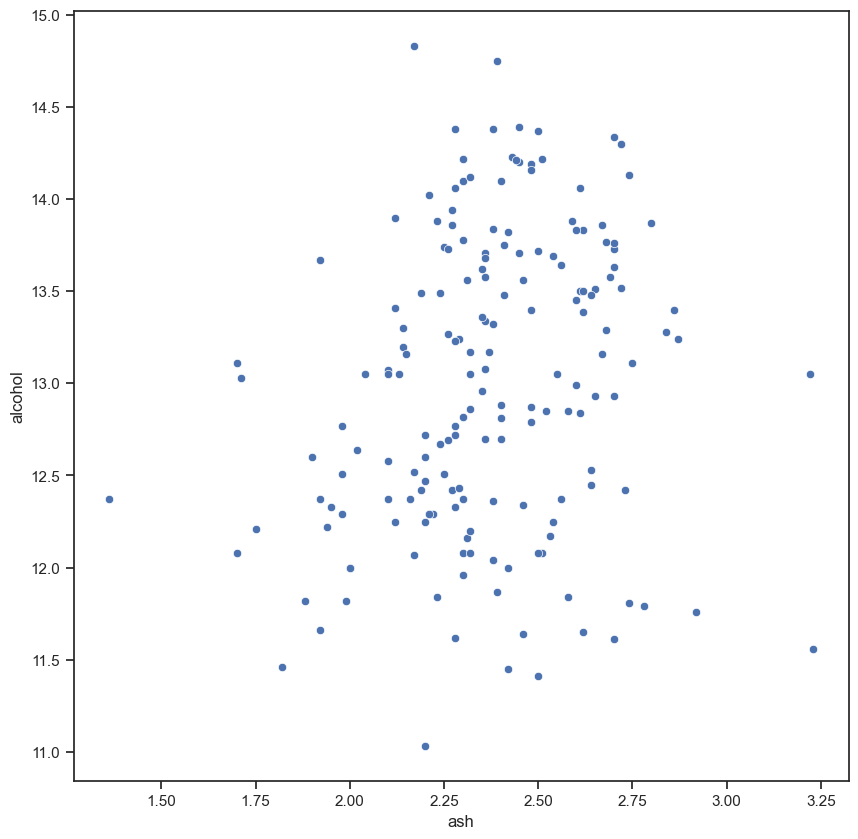
alcohol malic\_acid ash alcalinity\_of\_ash magnesium \  
count 178.000000 178.000000 178.000000 178.000000 178.000000   
mean 13.000618 2.336348 2.366517 19.494944 99.741573   
std 0.811827 1.117146 0.274344 3.339564 14.282484   
min 11.030000 0.740000 1.360000 10.600000 70.000000   
25% 12.362500 1.602500 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   
  
 target   
count 178.000000   
mean 0.938202   
std 0.775035   
min 0.000000   
25% 0.000000   
50% 1.000000   
75% 2.000000   
max 2.000000

# Определим уникальные значения для целевого признака  
data['target'].unique()

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

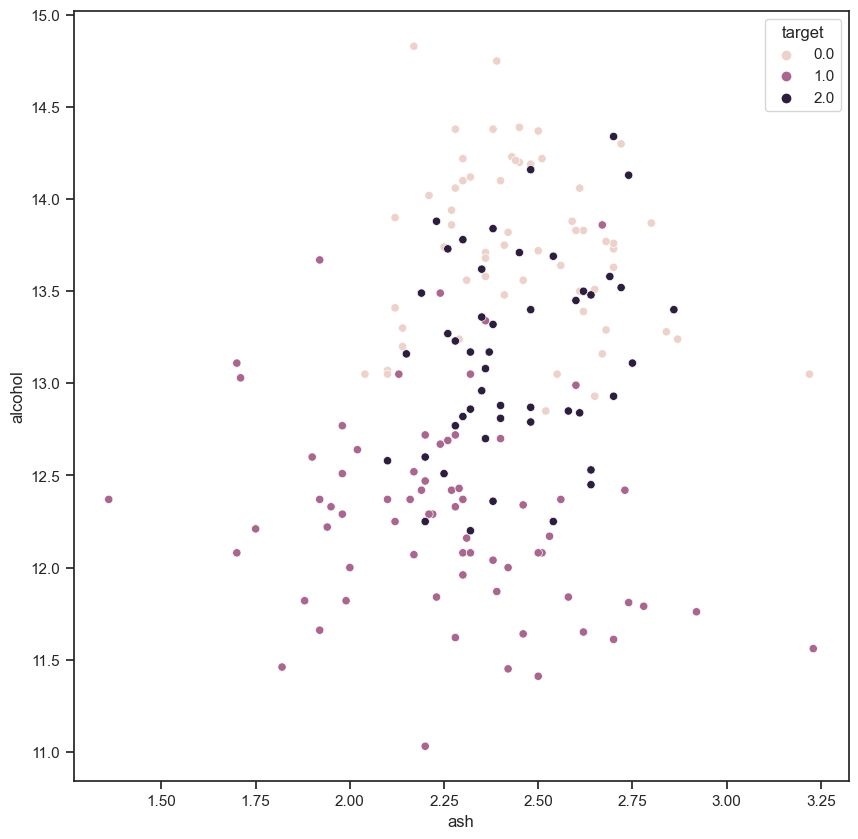
fig, ax = plt.subplots(figsize=(10,10))   
sns.scatterplot(ax=ax, x='ash', y='alcohol', data=data)

<AxesSubplot:xlabel='ash', ylabel='alcohol'>

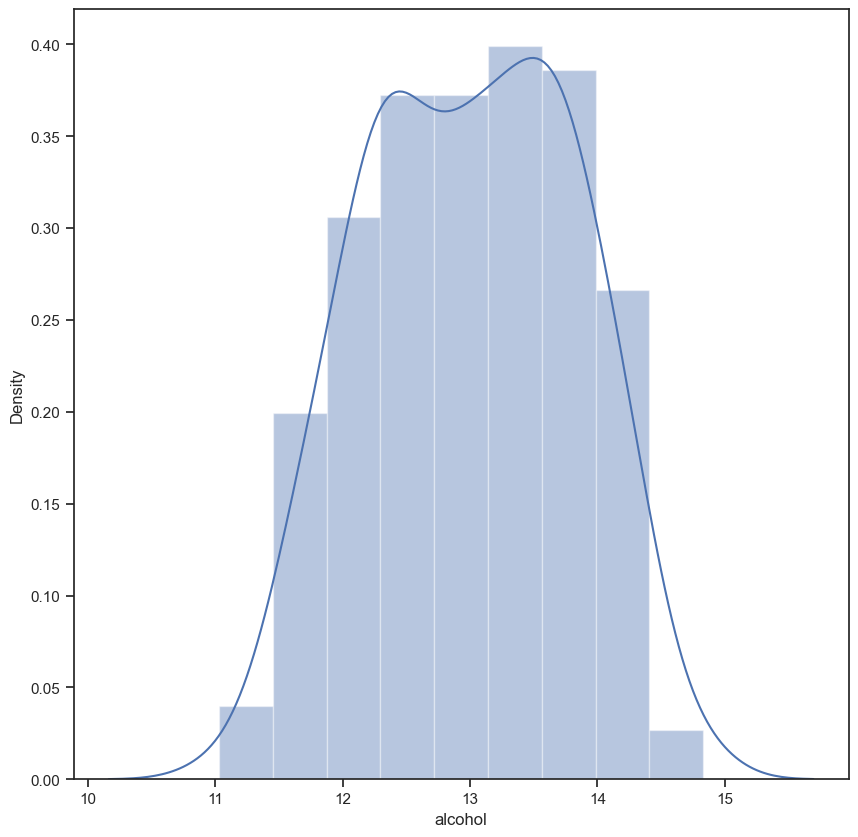


fig, ax = plt.subplots(figsize=(10,10))   
sns.scatterplot(ax=ax, x='ash', y='alcohol', data=data, hue='target')

<AxesSubplot:xlabel='ash', ylabel='alcohol'>



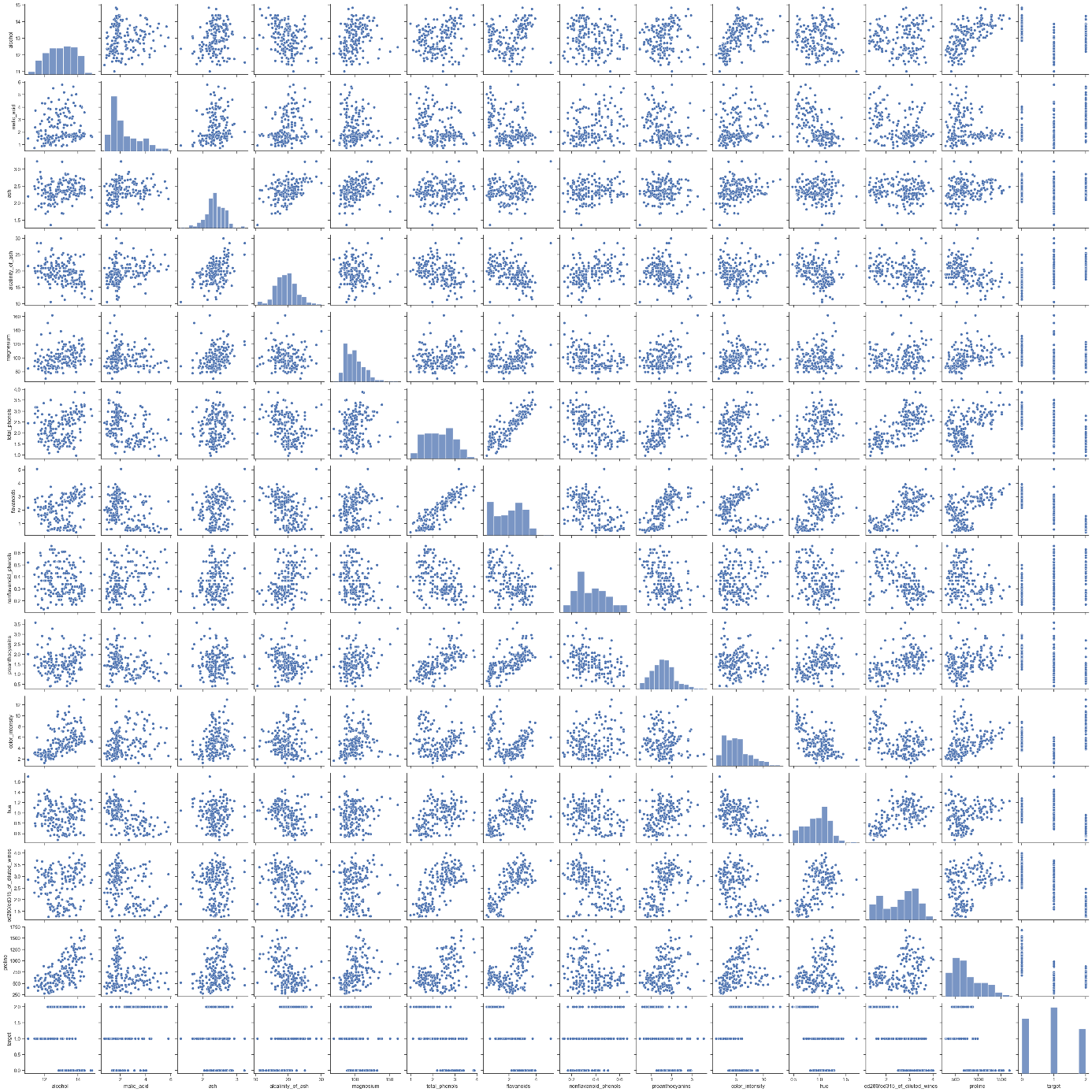
fig, ax = plt.subplots(figsize=(10,10))   
sns.distplot(data['alcohol'])

C:\Users\prite\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).  
 warnings.warn(msg, FutureWarning)

<AxesSubplot:xlabel='alcohol', ylabel='Density'>

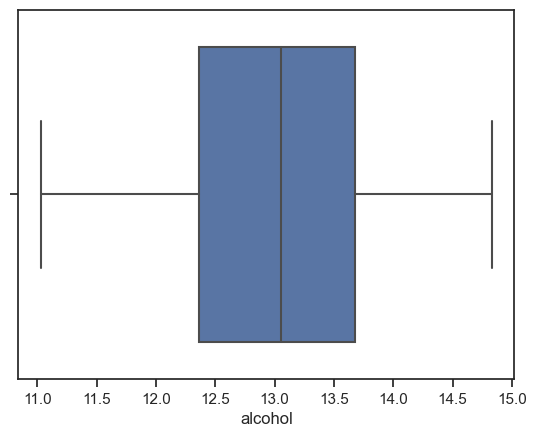
sns.pairplot(data)

<seaborn.axisgrid.PairGrid at 0x20853180a60>



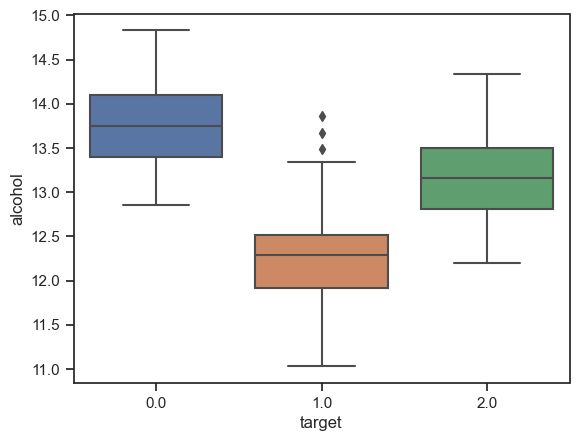
# По горизонтали  
sns.boxplot(x=data['alcohol'])

<AxesSubplot:xlabel='alcohol'>



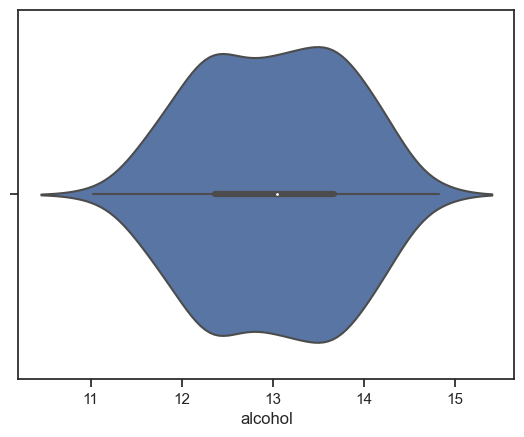
sns.boxplot(x='target', y='alcohol', data=data)

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



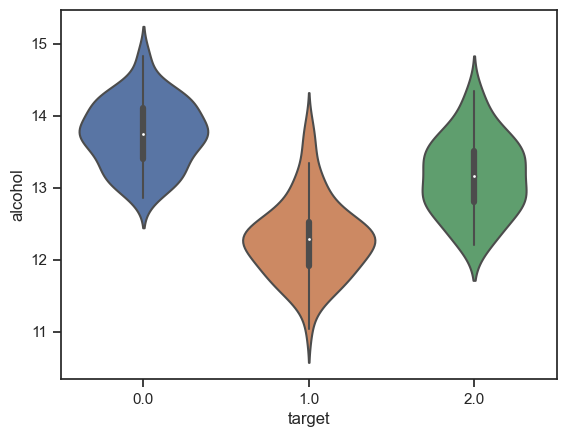
sns.violinplot(x=data['alcohol'])

<AxesSubplot:xlabel='alcohol'>



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

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



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

На основе корреляционной матрицы можно сделать следующие выводы:

* Целевой признак наиболее сильно коррелирует с флавоноидами (-0.85), OD280/OD315 разбавленных вин (-0.79), количеством фенолов (-0.72). Эти признаки обязательно следует оставить в модели.
* Целевой признак отчасти коррелирует с концентрацией пролина (-0.64), оттенком (-0.62), нефлаваноидными фенолами (-0.49), концентрацикй проантоцианы(-0.5), щелочностью золы (0.51), концентрацикй яблочной кислоты (0.44). Эти признаки стоит также оставить в модели.
* Целевой признак слабо коррелирует с концетрациикй алкоголя (-0.32), пепла (0.05), магния (-0.2), интенсивностью цвета (0.27). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат качество модели.

# Вывод значений в ячейках  
fig, ax = plt.subplots(figsize=(10,10))  
sns.heatmap(data.corr(), annot=True, cmap='YlGnBu', fmt='.2f', linewidths=.5, ax=ax)

<AxesSubplot:>

