



***«Методы машинного обучения»***

Отчет по

Лабораторной работе №1

Разведочный анализ данных. Исследование и визуализация данных.

Выполнил:

студент

Егоров С. А. ИУ5-22М

Проверил:

доцент, к.т.н.

Гапанюк Ю.Е.

Москва, 2020

## Цель лабораторной работы

Изучить различные методы визуализация данных. Построить основные графики, входящие в этап разведочного анализа данных. Корреляционный анализ данных. Формирование выводов о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

## Реализация задания

```
In [13]: #Подключаем библиотеки для анализа
```

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

```
In [ ]: #Подключаем данные
```

```
data = pd.read_csv('wine.data', sep=",")
```

```
In [15]: # Разменность датасета
```

```
total_count = data.shape
print('Всего строк: {}'.format(total_count[0]))
print('Всего колонок: {}'.format(total_count[1]))
```

```
Всего строк: 178
Всего колонок: 14
```

```
In [16]: # Список колонок с типами данных
```

```
data.dtypes
```

```
Out[16]: class                int64
Alcohol                float64
Malic acid              float64
Ash                    float64
Alcalinity of ash       float64
Magnesium               int64
Total phenols           float64
Flavanoids              float64
Nonflavanoid phenols    float64
Proanthocyanins         float64
Color intensity         float64
Hue                    float64
OD280/OD315 of diluted wines float64
Proline                 int64
dtype: object
```

```
In [12]: #Выведем первые 5 строк
```

```
data.head()
```

```
Out[12]:
```

	class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD315 of diluted wines	Proline
0	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
2	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
3	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
4	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735

```
In [17]: # Проверим наличие пустых значений
```

```
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
class - 0
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
```

```
In [18]: # Основные статистические характеристики набора данных
data.describe()
```

Out[18]:

	class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	1.938202	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	0.957440
std	0.775035	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	0.228510
min	1.000000	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	0.480000
25%	1.000000	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	0.782500
50%	2.000000	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	0.965000
75%	3.000000	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	1.120000
max	3.000000	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	1.710000

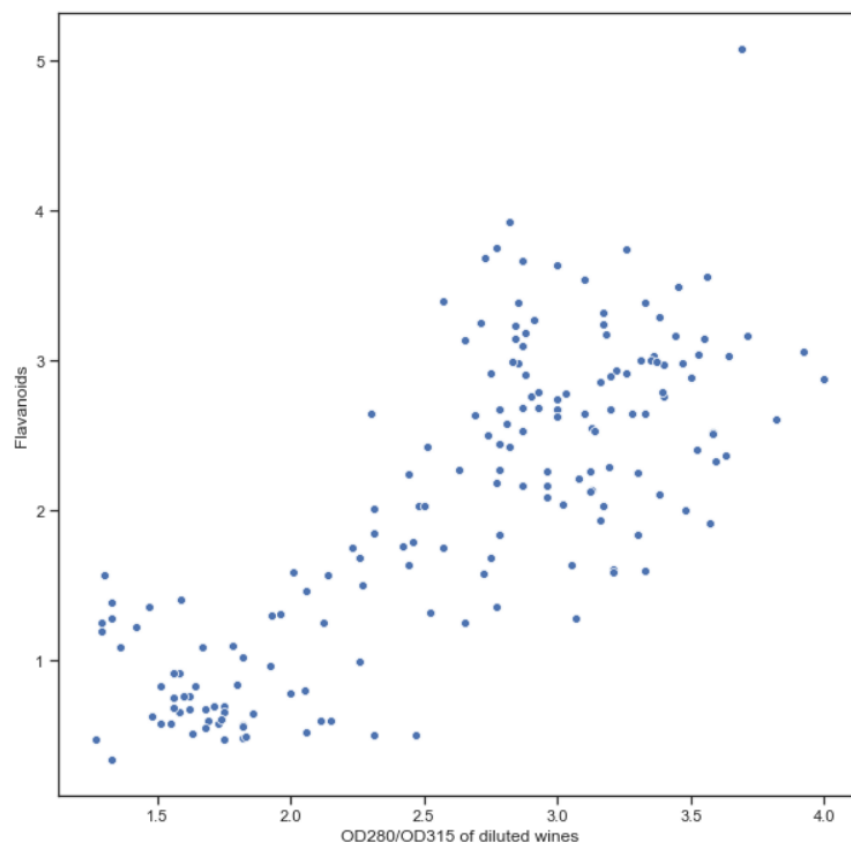
```
In [19]: #Проведем корреляционный анализ
data.corr()
```

Out[19]:

	class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue
class	1.000000	-0.328222	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.499130	0.265668	-0.617369
Alcohol	-0.328222	1.000000	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.136698	0.546364	-0.071747
Malic acid	0.437776	0.094397	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.220746	0.248985	-0.561296
Ash	-0.049643	0.211545	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.009652	0.258887	-0.074667
Alcalinity of ash	0.517859	-0.310235	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.197327	0.018732	-0.273955
Magnesium	-0.209179	0.270798	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.236441	0.199950	0.055398
Total phenols	-0.719163	0.289101	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.612413	-0.055136	0.433681
Flavanoids	-0.847498	0.236815	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.652692	-0.172379	0.543479
Nonflavanoid phenols	0.489109	-0.155929	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.365845	0.139057	-0.262640

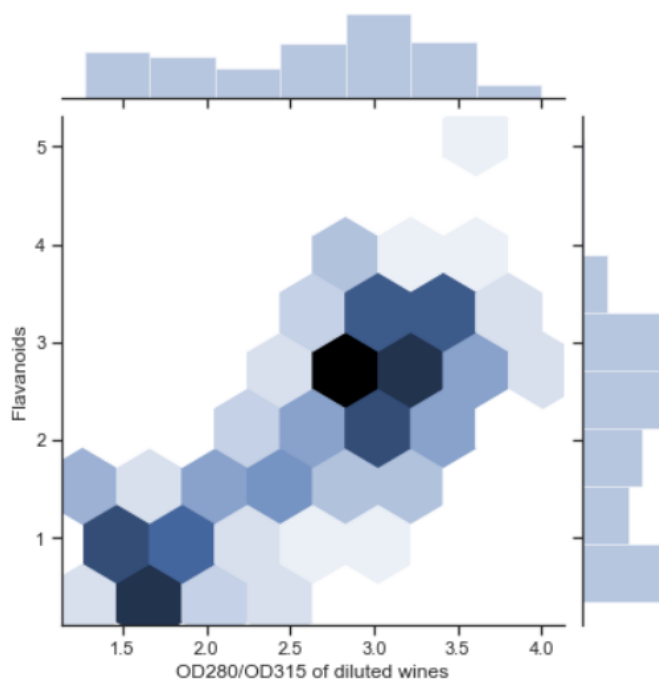
```
In [22]: #Диаграмма рассеивания
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='OD280/OD315 of diluted wines', y='Flavanoids', data=data)
```

Out[22]: <matplotlib.axes.\_subplots.AxesSubplot at 0x25ac9066940>



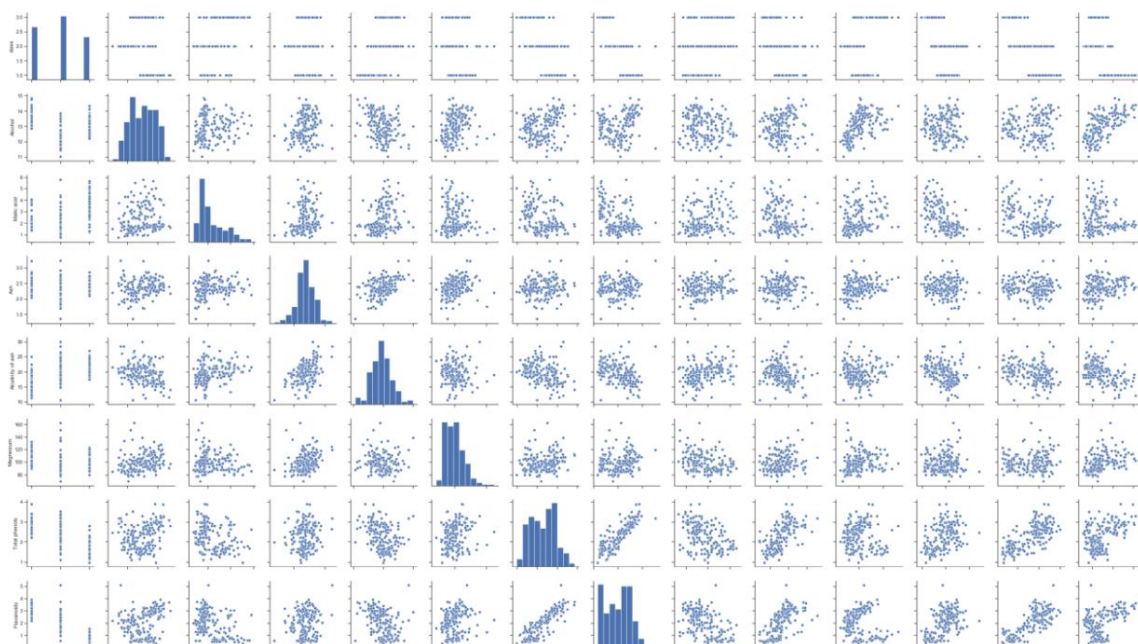
```
In [26]: #Диаграмма рассеивания + гистограмма
sns.jointplot(x='OD280/OD315 of diluted wines', y='Flavanoids', data=data, kind="hex")
```

```
Out[26]: <seaborn.axisgrid.JointGrid at 0x25ac90816d8>
```



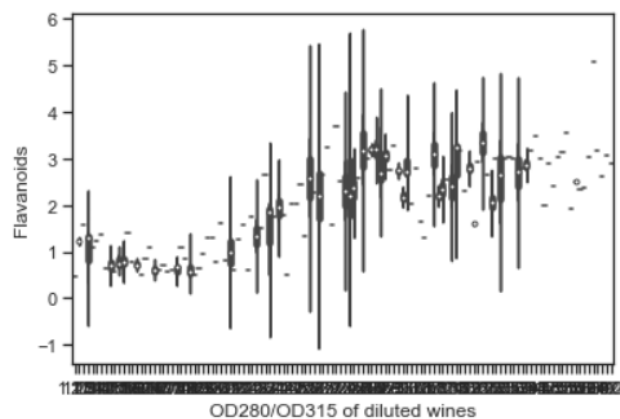
```
In [27]: #Парные диаграммы"
sns.pairplot(data)
```

```
Out[27]: <seaborn.axisgrid.PairGrid at 0x25ac994f7f0>
```



```
In [28]: # Распределение параметра 'Flavanoids' сгруппированные по 'OD280/OD315 of diluted wines'.
sns.violinplot(x='OD280/OD315 of diluted wines', y='Flavanoids', data=data)
```

```
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x25ad129eeb8>
```



```
In [37]: # Треугольный вариант матрицы корреляции
mask = np.zeros_like(data.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
f, ax = plt.subplots(figsize=(15, 10))
sns.heatmap(data.corr(), mask=mask, cmap='YlGnBu', annot=True, fmt='.3f',)
```

