

Лабораторная работа №1  
по дисциплине  
«Методы машинного обучения»  
на тему  
«Разведочный анализ данных. Исследование и  
визуализация данных»

Выполнил:  
студент группы ИУ5-64Б  
Береговая Д. В.

---

# 1. Цель лабораторной работы

Изучить различные методы визуализации данных # Задание Требуется выполнить следующие действия: - Выбрать набор данных для анализа. - Создать ноутбук, который содержит следующие разделы. 1. Текстовое описание выбранного 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.

## 2. Ход выполнения работы

### 2.1. 1) Текстовое описание набора данных

#### 2.1.1. В качестве набора данных используется набор по исследованию качества белых вин

Датасет состоит из одного файла: - wine.csv

Файл содержит следующие колонки: 1. fixed acidity — фиксированная кислотность 2. volatile acidity — летучая кислотность 3. citric acid — лимонная кислота 4. residual sugar — остаточный сахар 5. chlorides — хлориды 6. free sulfur dioxide — свободный диоксид серы 7. total sulfur dioxide — общая двуокись серы 8. density — плотность 9. pH — потенциал водорода 10. sulphates — сульфаты 11. alcohol — алкоголь 12. quality — качество алкоголя (выходной параметр)

#### 2.1.2. Импортируем библиотеки

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

#### 2.1.3. Загрузим данные датасета

```
[4]: data = pd.read_csv('wine.csv', sep=";")
```

#### 2.1.4. Проверим корректность загрузки

```
[5]: data.head()
```

```
[5]:   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides \
0           7.0           0.27           0.36           20.7       0.045
1           6.3           0.30           0.34            1.6       0.049
2           8.1           0.28           0.40            6.9       0.050
3           7.2           0.23           0.32            8.5       0.058
4           7.2           0.23           0.32            8.5       0.058

   free sulfur dioxide  total sulfur dioxide  density  pH  sulphates \
0                45.0                170.0   1.0010  3.00         0.45
1                14.0                132.0   0.9940  3.30         0.49
```

2	30.0	97.0	0.9951	3.26	0.44
3	47.0	186.0	0.9956	3.19	0.40
4	47.0	186.0	0.9956	3.19	0.40

	alcohol	quality
0	8.8	6
1	9.5	6
2	10.1	6
3	9.9	6
4	9.9	6

### 2.1.5. Размер датасета

```
[6]: data.shape
```

```
[6]: (4898, 12)
```

### 2.1.6. Список атрибутов

```
[7]: data.dtypes
```

```
[7]: fixed acidity      float64
     volatile acidity  float64
     citric acid       float64
     residual sugar    float64
     chlorides         float64
     free sulfur dioxide float64
     total sulfur dioxide float64
     density           float64
     pH               float64
     sulphates        float64
     alcohol           float64
     quality           int64
     dtype: object
```

### 2.1.7. Проверка датасета на наличие пустых значений

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

```
fixed acidity - 0
volatile acidity - 0
citric acid - 0
residual sugar - 0
chlorides - 0
free sulfur dioxide - 0
total sulfur dioxide - 0
```

```
density - 0
pH - 0
sulphates - 0
alcohol - 0
quality - 0
```

### 2.1.8. Основные статистические характеристики набора данных

```
[9]: data.describe()
```

```
[9]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar \
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415
std	0.843868	0.100795	0.121020	5.072058
min	3.800000	0.080000	0.000000	0.600000
25%	6.300000	0.210000	0.270000	1.700000
50%	6.800000	0.260000	0.320000	5.200000
75%	7.300000	0.320000	0.390000	9.900000
max	14.200000	1.100000	1.660000	65.800000

	chlorides	free sulfur dioxide	total sulfur dioxide	density \
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	0.045772	35.308085	138.360657	0.994027
std	0.021848	17.007137	42.498065	0.002991
min	0.009000	2.000000	9.000000	0.987110
25%	0.036000	23.000000	108.000000	0.991723
50%	0.043000	34.000000	134.000000	0.993740
75%	0.050000	46.000000	167.000000	0.996100
max	0.346000	289.000000	440.000000	1.038980

	pH	sulphates	alcohol	quality
count	4898.000000	4898.000000	4898.000000	4898.000000
mean	3.188267	0.489847	10.514267	5.877909
std	0.151001	0.114126	1.230621	0.885639
min	2.720000	0.220000	8.000000	3.000000
25%	3.090000	0.410000	9.500000	5.000000
50%	3.180000	0.470000	10.400000	6.000000
75%	3.280000	0.550000	11.400000	6.000000
max	3.820000	1.080000	14.200000	9.000000

### 2.1.9. Уникальные значения для целевого признака

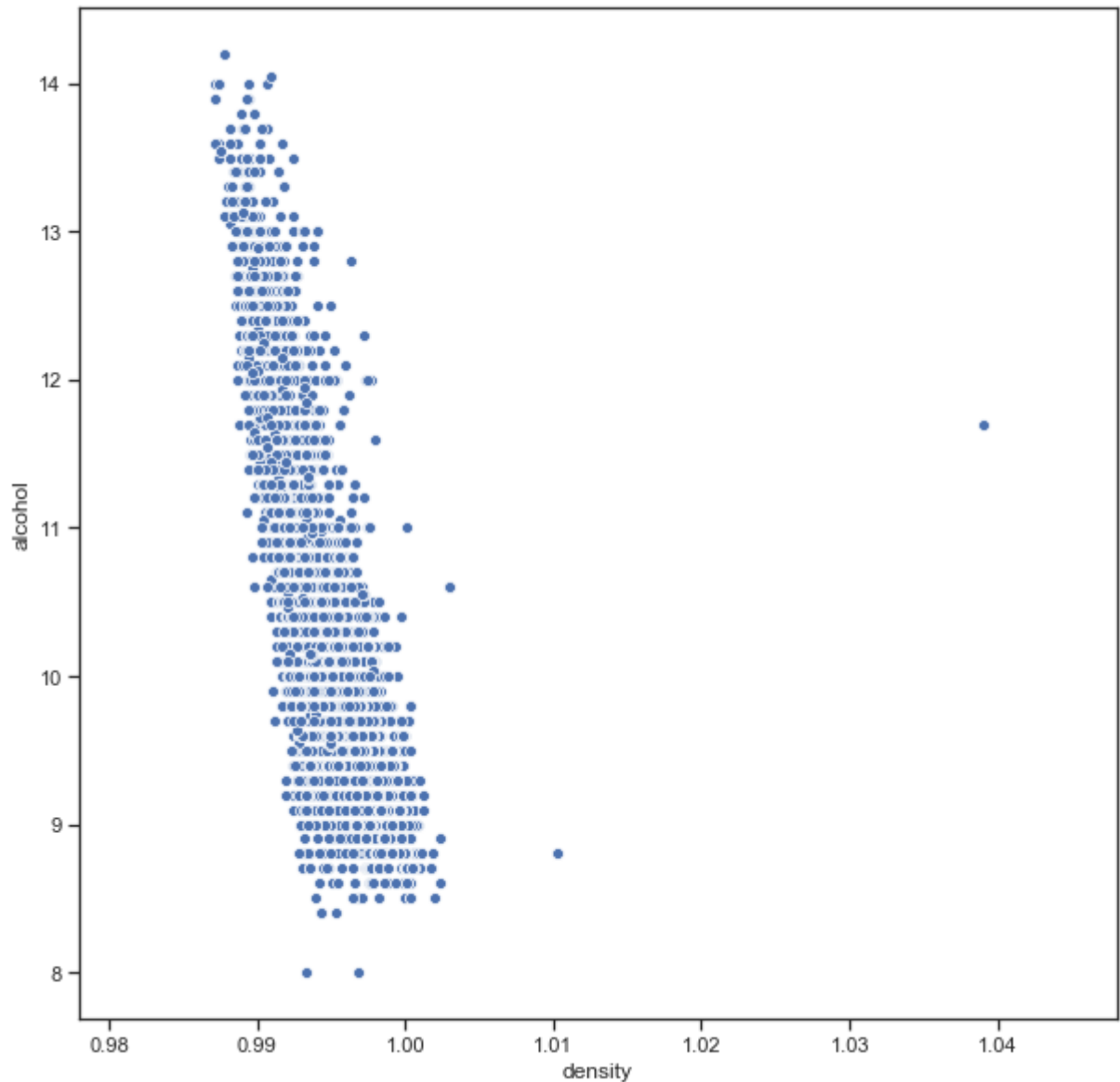
```
[10]: data['quality'].unique()
```

```
[10]: array([6, 5, 7, 8, 4, 3, 9])
```

## 2.2. 2) Визуальное исследование датасета

```
[11]: fig, ax = plt.subplots(figsize=(10,10))  
      sns.scatterplot(ax=ax, x='density', y='alcohol', data=data)
```

```
[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1a24f23750>
```

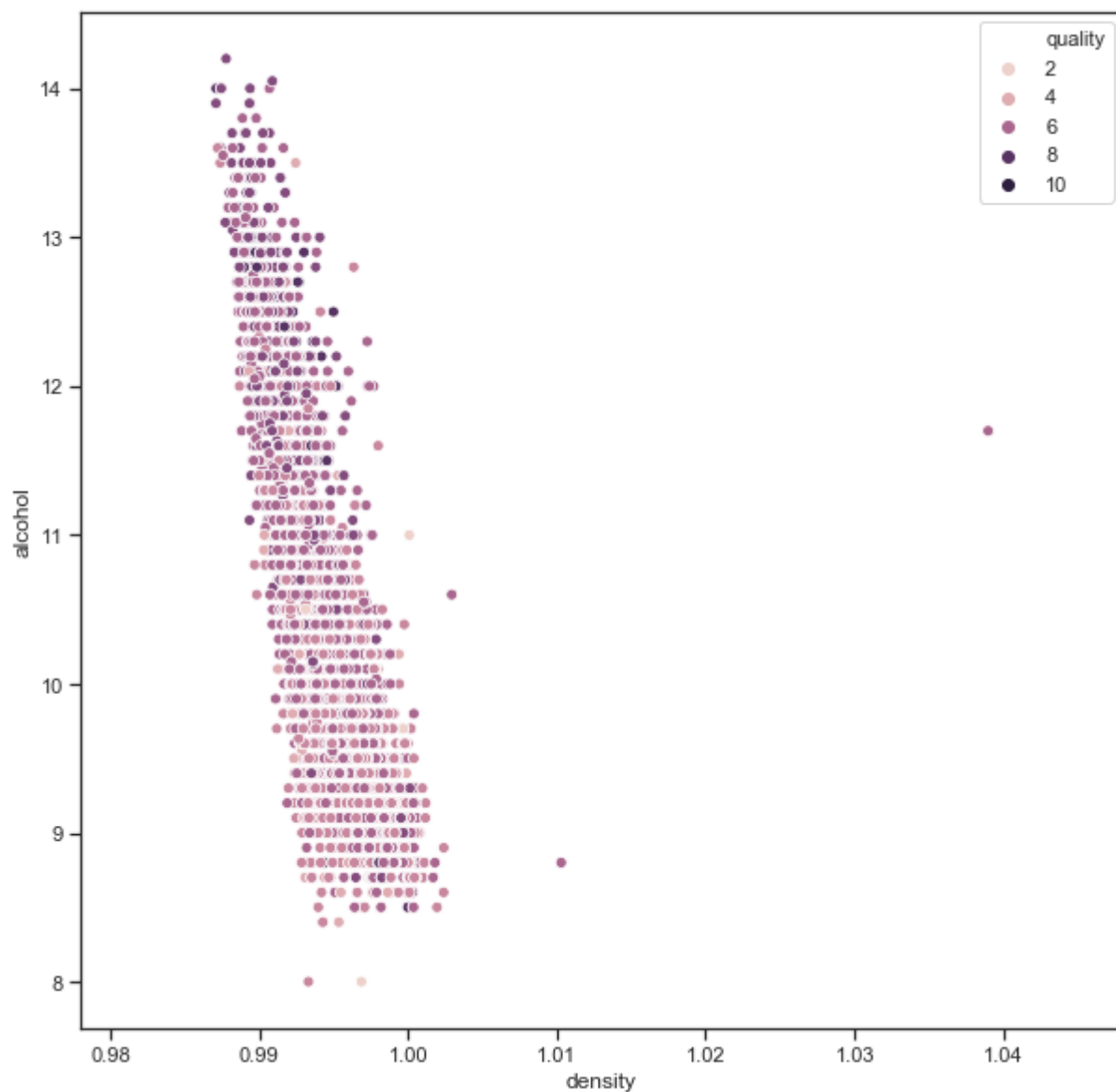


Можно заметить, что между атрибутами density и alcohol присутствует что-то похожее на линейную зависимость

### 2.2.1. Введем в зависимость целевой признак

```
[12]: fig, ax = plt.subplots(figsize=(10,10))  
      sns.scatterplot(ax=ax, x='density', y='alcohol', data=data, hue='quality')
```

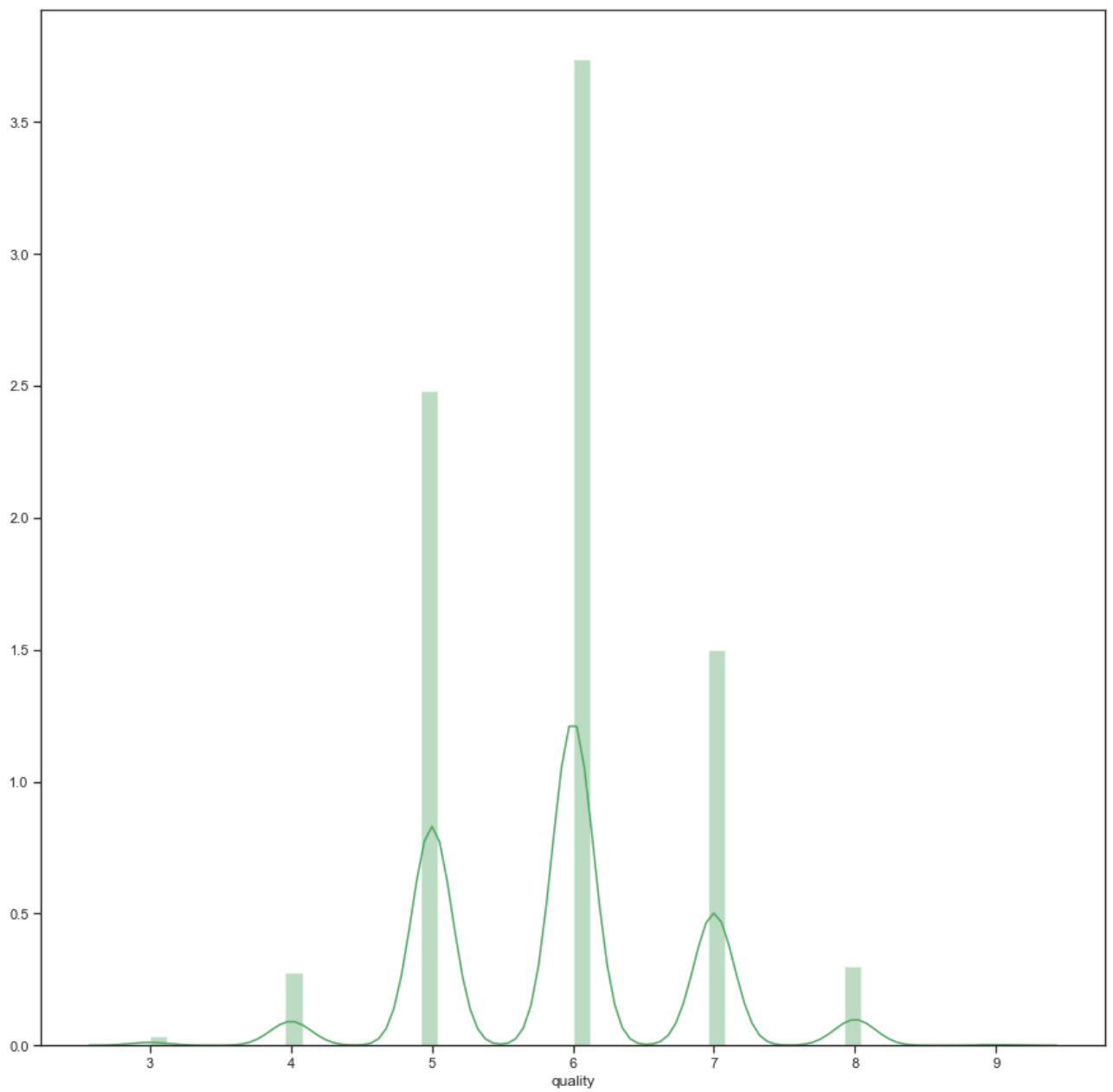
```
[12]: <matplotlib.axes._subplots.AxesSubplot at 0x1a25d25090>
```



### 2.2.2. Распределение качества алкоголя

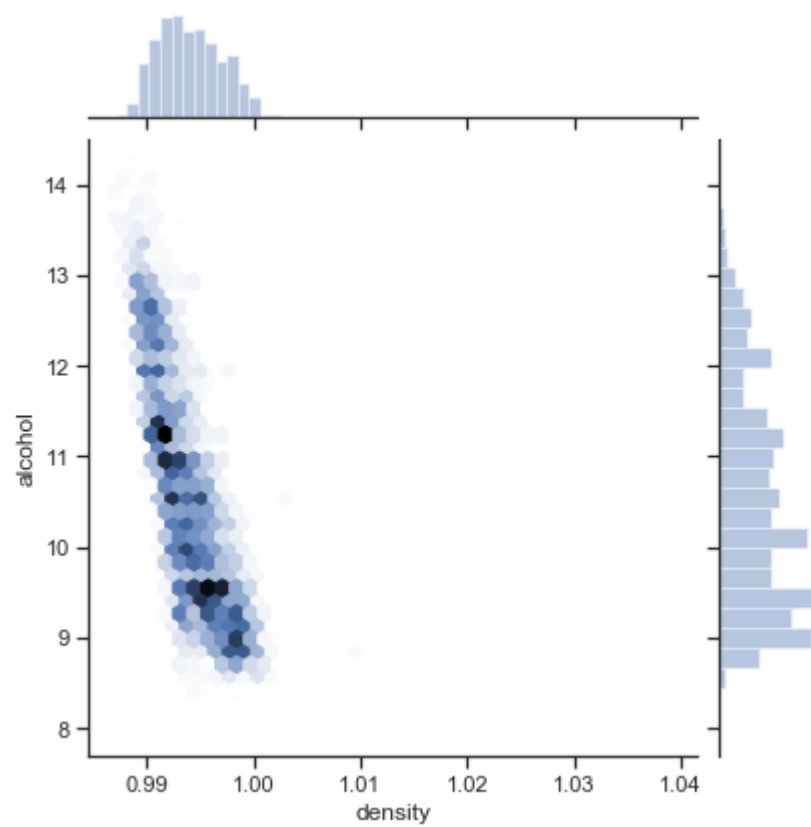
```
[14]: fig, ax = plt.subplots(figsize=(15,15))
      sns.distplot(data['quality'], color="g")
```

```
[14]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2635d6d0>
```



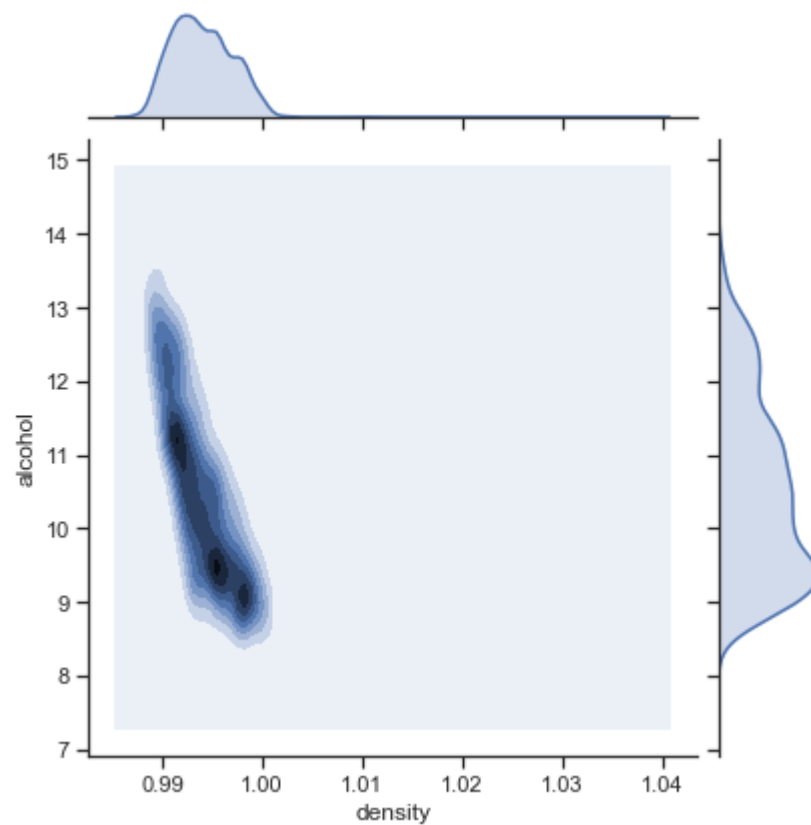
```
[15]: sns.jointplot(x='density', y='alcohol', data=data, kind="hex")
```

```
[15]: <seaborn.axisgrid.JointGrid at 0x1a26aee650>
```



```
[16]: sns.jointplot(x='density', y='alcohol', data=data, kind="kde")
```

```
[16]: <seaborn.axisgrid.JointGrid at 0x1a25e87890>
```

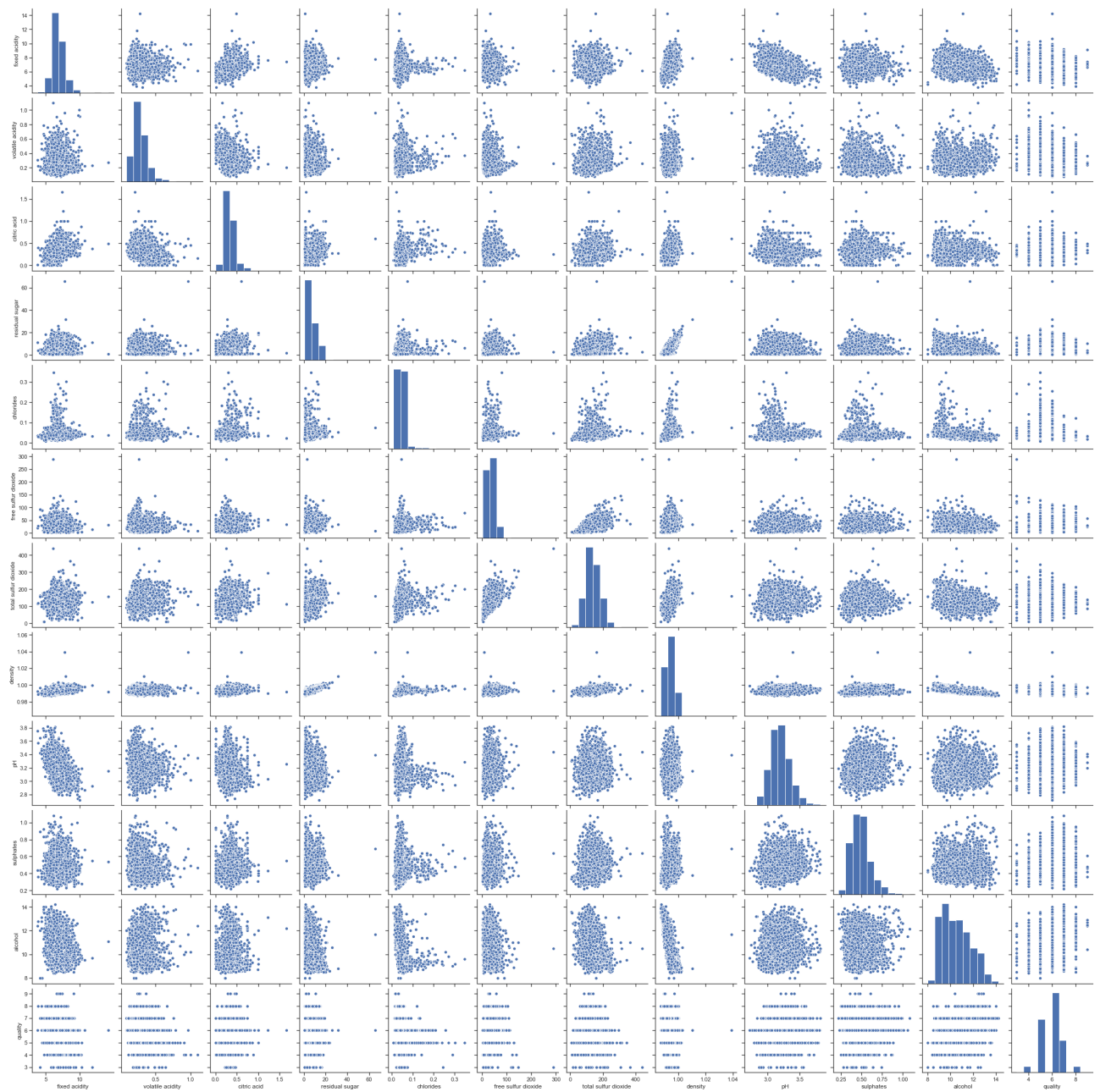




### 2.2.3. Парные диаграммы

```
[17]: sns.pairplot(data)
```

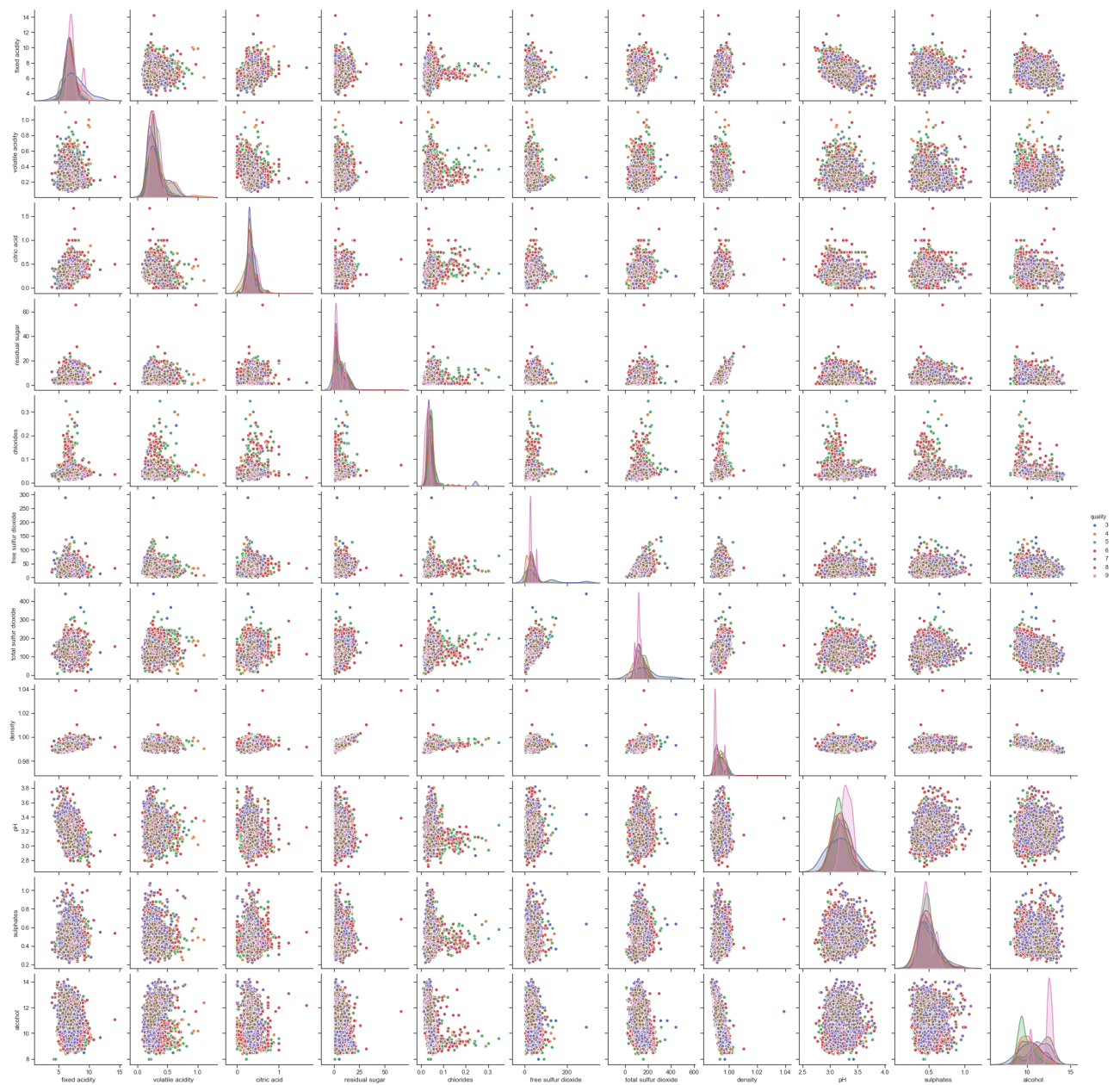
```
[17]: <seaborn.axisgrid.PairGrid at 0x1a2740eb10>
```



### 2.2.4. Группировка по значению целевого признака

```
[18]: sns.pairplot(data, hue="quality")
```

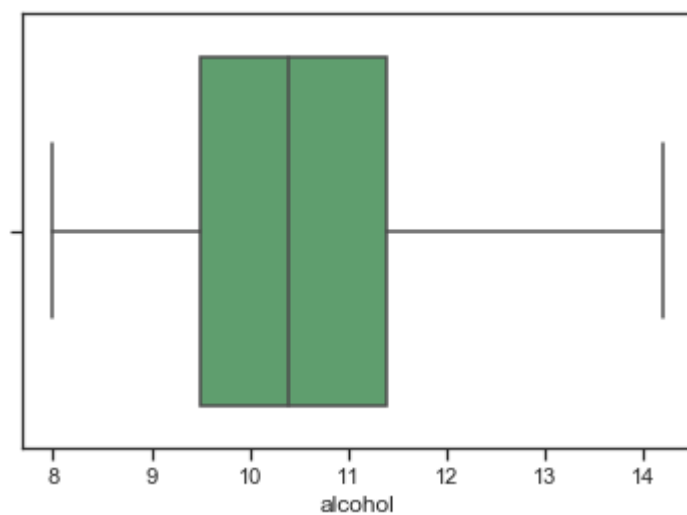
[18]: <seaborn.axisgrid.PairGrid at 0x1a2c014b10>



### 2.2.5. Одномерное распределение вероятности

[22]: `sns.boxplot(x=data['alcohol'], color="g")`

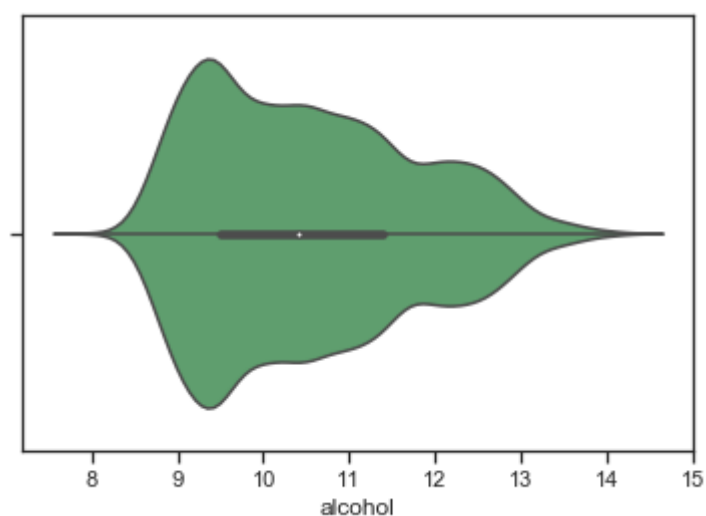
[22]: <matplotlib.axes.\_subplots.AxesSubplot at 0x1a2c674490>



### 2.2.6. Скрипичные диаграммы

```
[21]: sns.violinplot(x=data['alcohol'], color="g")
```

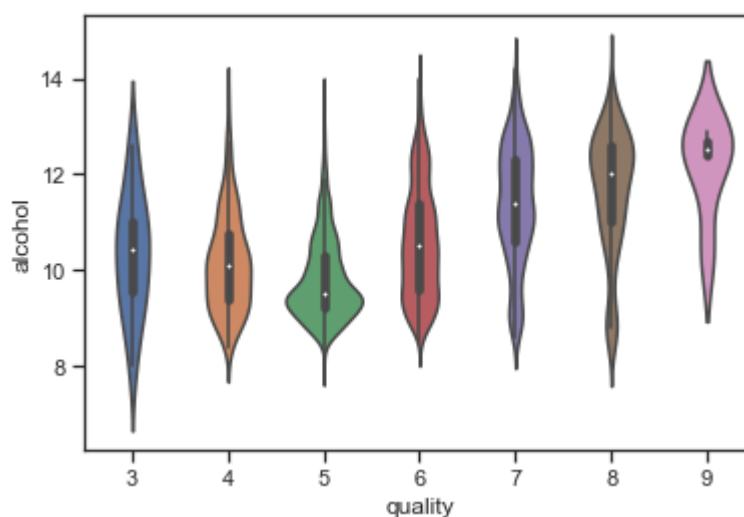
```
[21]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2c650810>
```



### 2.2.7. Сгруппируем по целевому признаку

```
[23]: sns.violinplot(x='quality', y='alcohol', data=data)
```

```
[23]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2c6b0910>
```



### 2.3. 3) Информация о корреляции признаков

[24]: `data.corr()`

```
[24]:
```

	fixed acidity	volatile acidity	citric acid	\
fixed acidity	1.000000	-0.022697	0.289181	
volatile acidity	-0.022697	1.000000	-0.149472	
citric acid	0.289181	-0.149472	1.000000	
residual sugar	0.089021	0.064286	0.094212	
chlorides	0.023086	0.070512	0.114364	
free sulfur dioxide	-0.049396	-0.097012	0.094077	
total sulfur dioxide	0.091070	0.089261	0.121131	
density	0.265331	0.027114	0.149503	
pH	-0.425858	-0.031915	-0.163748	
sulphates	-0.017143	-0.035728	0.062331	
alcohol	-0.120881	0.067718	-0.075729	
quality	-0.113663	-0.194723	-0.009209	

	residual sugar	chlorides	free sulfur dioxide	\
fixed acidity	0.089021	0.023086	-0.049396	
volatile acidity	0.064286	0.070512	-0.097012	
citric acid	0.094212	0.114364	0.094077	
residual sugar	1.000000	0.088685	0.299098	
chlorides	0.088685	1.000000	0.101392	
free sulfur dioxide	0.299098	0.101392	1.000000	
total sulfur dioxide	0.401439	0.198910	0.615501	
density	0.838966	0.257211	0.294210	
pH	-0.194133	-0.090439	-0.000618	
sulphates	-0.026664	0.016763	0.059217	
alcohol	-0.450631	-0.360189	-0.250104	
quality	-0.097577	-0.209934	0.008158	

	total sulfur dioxide	density	pH	sulphates \
fixed acidity	0.091070	0.265331	-0.425858	-0.017143
volatile acidity	0.089261	0.027114	-0.031915	-0.035728
citric acid	0.121131	0.149503	-0.163748	0.062331
residual sugar	0.401439	0.838966	-0.194133	-0.026664
chlorides	0.198910	0.257211	-0.090439	0.016763
free sulfur dioxide	0.615501	0.294210	-0.000618	0.059217
total sulfur dioxide	1.000000	0.529881	0.002321	0.134562
density	0.529881	1.000000	-0.093591	0.074493
pH	0.002321	-0.093591	1.000000	0.155951
sulphates	0.134562	0.074493	0.155951	1.000000
alcohol	-0.448892	-0.780138	0.121432	-0.017433
quality	-0.174737	-0.307123	0.099427	0.053678

	alcohol	quality
fixed acidity	-0.120881	-0.113663
volatile acidity	0.067718	-0.194723
citric acid	-0.075729	-0.009209
residual sugar	-0.450631	-0.097577
chlorides	-0.360189	-0.209934
free sulfur dioxide	-0.250104	0.008158
total sulfur dioxide	-0.448892	-0.174737
density	-0.780138	-0.307123
pH	0.121432	0.099427
sulphates	-0.017433	0.053678
alcohol	1.000000	0.435575
quality	0.435575	1.000000

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

```
[25]:
```

	fixed acidity	volatile acidity	citric acid \
fixed acidity	1.000000	-0.022697	0.289181
volatile acidity	-0.022697	1.000000	-0.149472
citric acid	0.289181	-0.149472	1.000000
residual sugar	0.089021	0.064286	0.094212
chlorides	0.023086	0.070512	0.114364
free sulfur dioxide	-0.049396	-0.097012	0.094077
total sulfur dioxide	0.091070	0.089261	0.121131
density	0.265331	0.027114	0.149503
pH	-0.425858	-0.031915	-0.163748
sulphates	-0.017143	-0.035728	0.062331
alcohol	-0.120881	0.067718	-0.075729
quality	-0.113663	-0.194723	-0.009209

	residual sugar	chlorides	free sulfur dioxide \
fixed acidity	0.089021	0.023086	-0.049396
volatile acidity	0.064286	0.070512	-0.097012
citric acid	0.094212	0.114364	0.094077
residual sugar	1.000000	0.088685	0.299098
chlorides	0.088685	1.000000	0.101392

free sulfur dioxide	0.299098	0.101392	1.000000
total sulfur dioxide	0.401439	0.198910	0.615501
density	0.838966	0.257211	0.294210
pH	-0.194133	-0.090439	-0.000618
sulphates	-0.026664	0.016763	0.059217
alcohol	-0.450631	-0.360189	-0.250104
quality	-0.097577	-0.209934	0.008158

	total sulfur dioxide	density	pH	sulphates	\
fixed acidity	0.091070	0.265331	-0.425858	-0.017143	
volatile acidity	0.089261	0.027114	-0.031915	-0.035728	
citric acid	0.121131	0.149503	-0.163748	0.062331	
residual sugar	0.401439	0.838966	-0.194133	-0.026664	
chlorides	0.198910	0.257211	-0.090439	0.016763	
free sulfur dioxide	0.615501	0.294210	-0.000618	0.059217	
total sulfur dioxide	1.000000	0.529881	0.002321	0.134562	
density	0.529881	1.000000	-0.093591	0.074493	
pH	0.002321	-0.093591	1.000000	0.155951	
sulphates	0.134562	0.074493	0.155951	1.000000	
alcohol	-0.448892	-0.780138	0.121432	-0.017433	
quality	-0.174737	-0.307123	0.099427	0.053678	

	alcohol	quality
fixed acidity	-0.120881	-0.113663
volatile acidity	0.067718	-0.194723
citric acid	-0.075729	-0.009209
residual sugar	-0.450631	-0.097577
chlorides	-0.360189	-0.209934
free sulfur dioxide	-0.250104	0.008158
total sulfur dioxide	-0.448892	-0.174737
density	-0.780138	-0.307123
pH	0.121432	0.099427
sulphates	-0.017433	0.053678
alcohol	1.000000	0.435575
quality	0.435575	1.000000

```
[26]: data.corr(method='kendall')
```

```
[26]:
```

	fixed acidity	volatile acidity	citric acid	\
fixed acidity	1.000000	-0.029565	0.208569	
volatile acidity	-0.029565	1.000000	-0.104012	
citric acid	0.208569	-0.104012	1.000000	
residual sugar	0.074946	0.072757	0.015329	
chlorides	0.065361	-0.003523	0.022292	
free sulfur dioxide	-0.016940	-0.054751	0.060809	
total sulfur dioxide	0.077272	0.081319	0.062188	
density	0.185510	0.006600	0.061542	
pH	-0.294796	-0.030385	-0.101307	
sulphates	-0.008724	-0.011580	0.054489	
alcohol	-0.073241	0.023495	-0.019981	



quality	-0.065474	-0.154787	0.014557
---------	-----------	-----------	----------

	residual sugar	chlorides	free sulfur dioxide	\
fixed acidity	0.074946	0.065361	-0.016940	
volatile acidity	0.072757	-0.003523	-0.054751	
citric acid	0.015329	0.022292	0.060809	
residual sugar	1.000000	0.155274	0.236748	
chlorides	0.155274	1.000000	0.113851	
free sulfur dioxide	0.236748	0.113851	1.000000	
total sulfur dioxide	0.293319	0.257075	0.444696	
density	0.588989	0.349119	0.217295	
pH	-0.125553	-0.037891	-0.005229	
sulphates	-0.002545	0.062555	0.035621	
alcohol	-0.305601	-0.404039	-0.182539	
quality	-0.063087	-0.244856	0.017164	

	total sulfur dioxide	density	pH	sulphates	\
fixed acidity	0.077272	0.185510	-0.294796	-0.008724	
volatile acidity	0.081319	0.006600	-0.030385	-0.011580	
citric acid	0.062188	0.061542	-0.101307	0.054489	
residual sugar	0.293319	0.588989	-0.125553	-0.002545	
chlorides	0.257075	0.349119	-0.037891	0.062555	
free sulfur dioxide	0.444696	0.217295	-0.005229	0.035621	
total sulfur dioxide	1.000000	0.388378	-0.008421	0.108697	
density	0.388378	1.000000	-0.075630	0.064202	
pH	-0.008421	-0.075630	1.000000	0.095823	
sulphates	0.108697	0.064202	0.095823	1.000000	
alcohol	-0.325826	-0.635104	0.102631	-0.026410	
quality	-0.151230	-0.266598	0.084441	0.026403	

	alcohol	quality
fixed acidity	-0.073241	-0.065474
volatile acidity	0.023495	-0.154787
citric acid	-0.019981	0.014557
residual sugar	-0.305601	-0.063087
chlorides	-0.404039	-0.244856
free sulfur dioxide	-0.182539	0.017164
total sulfur dioxide	-0.325826	-0.151230
density	-0.635104	-0.266598
pH	0.102631	0.084441
sulphates	-0.026410	0.026403
alcohol	1.000000	0.346672
quality	0.346672	1.000000

```
[27]: data.corr(method='spearman')
```

```
[27]:
```

	fixed acidity	volatile acidity	citric acid	\
fixed acidity	1.000000	-0.042865	0.297878	
volatile acidity	-0.042865	1.000000	-0.150410	
citric acid	0.297878	-0.150410	1.000000	

residual sugar	0.106725	0.108627	0.024621
chlorides	0.094691	-0.004934	0.032659
free sulfur dioxide	-0.024542	-0.081213	0.088314
total sulfur dioxide	0.112649	0.117614	0.093219
density	0.270031	0.010124	0.091425
pH	-0.418341	-0.045204	-0.146193
sulphates	-0.013238	-0.016902	0.079766
alcohol	-0.106827	0.033967	-0.029170
quality	-0.084485	-0.196562	0.018333

	residual sugar	chlorides	free sulfur dioxide	\
fixed acidity	0.106725	0.094691	-0.024542	
volatile acidity	0.108627	-0.004934	-0.081213	
citric acid	0.024621	0.032659	0.088314	
residual sugar	1.000000	0.227844	0.346107	
chlorides	0.227844	1.000000	0.167046	
free sulfur dioxide	0.346107	0.167046	1.000000	
total sulfur dioxide	0.431252	0.375244	0.618616	
density	0.780365	0.508302	0.327822	
pH	-0.180028	-0.054006	-0.006274	
sulphates	-0.003844	0.093931	0.052252	
alcohol	-0.445257	-0.570806	-0.272569	
quality	-0.082070	-0.314488	0.023713	

	total sulfur dioxide	density	pH	sulphates	\
fixed acidity	0.112649	0.270031	-0.418341	-0.013238	
volatile acidity	0.117614	0.010124	-0.045204	-0.016902	
citric acid	0.093219	0.091425	-0.146193	0.079766	
residual sugar	0.431252	0.780365	-0.180028	-0.003844	
chlorides	0.375244	0.508302	-0.054006	0.093931	
free sulfur dioxide	0.618616	0.327822	-0.006274	0.052252	
total sulfur dioxide	1.000000	0.563824	-0.011829	0.157825	
density	0.563824	1.000000	-0.110061	0.095079	
pH	-0.011829	-0.110061	1.000000	0.140243	
sulphates	0.157825	0.095079	0.140243	1.000000	
alcohol	-0.476619	-0.821855	0.148857	-0.044868	
quality	-0.196680	-0.348351	0.109362	0.033319	

	alcohol	quality
fixed acidity	-0.106827	-0.084485
volatile acidity	0.033967	-0.196562
citric acid	-0.029170	0.018333
residual sugar	-0.445257	-0.082070
chlorides	-0.570806	-0.314488
free sulfur dioxide	-0.272569	0.023713
total sulfur dioxide	-0.476619	-0.196680
density	-0.821855	-0.348351
pH	0.148857	0.109362
sulphates	-0.044868	0.033319
alcohol	1.000000	0.440369

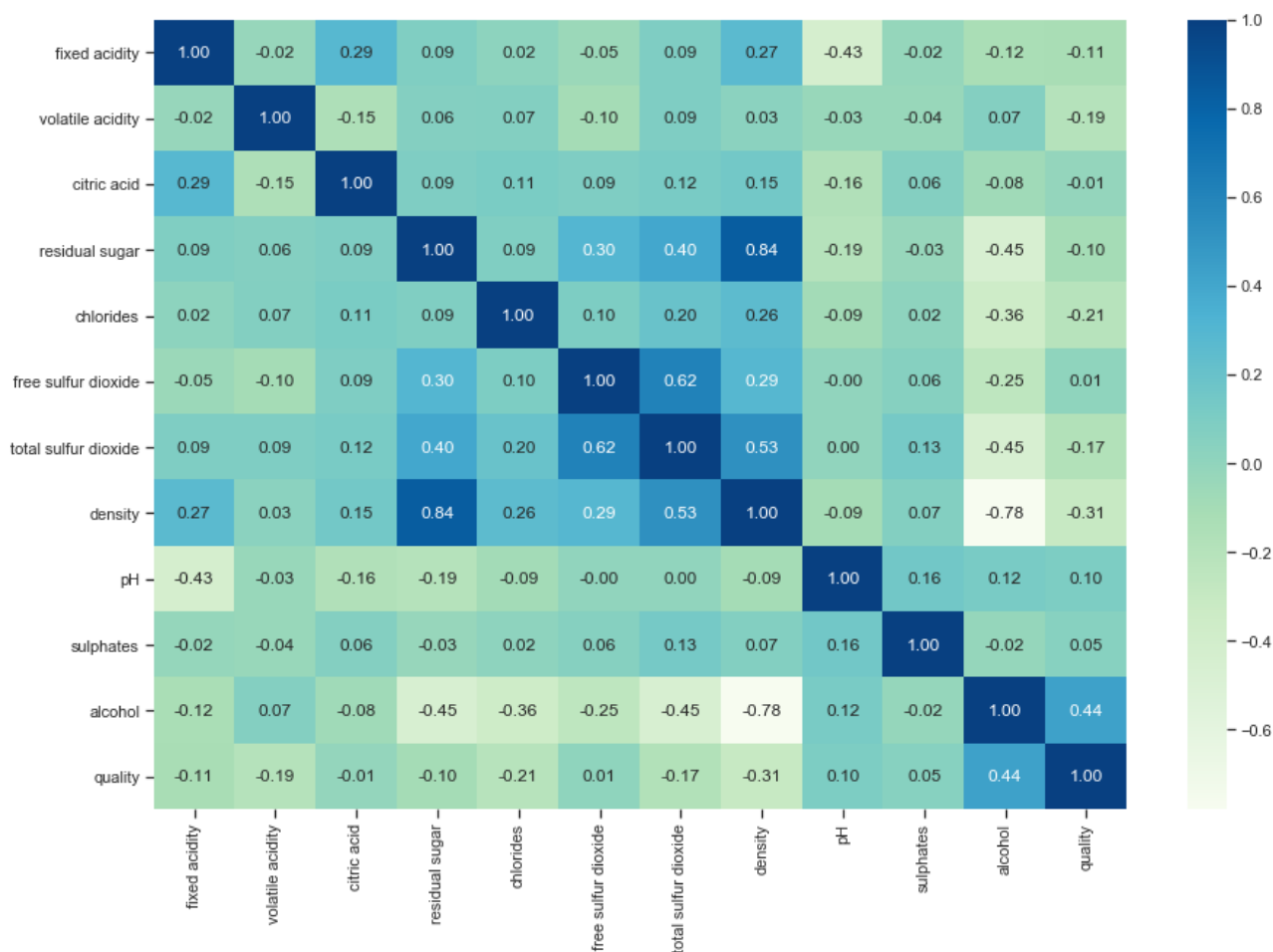


quality 0.440369 1.000000

### 2.3.1. Корреляционная матрица

```
[28]: fig, ax = plt.subplots(figsize=(15,10))  
sns.heatmap(data.corr(), annot=True, fmt='.2f', cmap='GnBu')
```

```
[28]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2df1a590>
```



### 2.4. Выводы о коррелирующих признаках

1. Коэффициенты корреляции в данном наборе достаточно низкие.
2. Если рассуждать чисто логически, то все представленные входные параметры влияют на качество алкоголя, так как они определяют его химический состав.
3. 'alcohol' и 'density' лучше всего коррелируют с целевым признаком, однако они очень сильно коррелируют друг с другом ( $|0.78|$ ).
4. 'free sulfur' и 'total sulfur' довольно неплохо коррелируют друг с другом ( $|0.62|$ ), что логично, так как общий диоксид серы является суммой связанной и свободной серы. У них прослеживается явная зависимость.

[ ]: