Рубежный контроль №1 ¶

Группа: ИУ5Ц-83Б

Студент: Соловьева Александра

Тема: Технологии разведочного анализа и обработки данных.

- Номер варианта: 27
- Номер задачи: 4
- Номер набора данных, указанного в задаче: 3

Дополнительные требования по группам: ИУ5Ц-83Б - для произвольной колонки данных построить график "Ящик с усами (boxplot)".

Задача №4.

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

In [1]:

```
import sys
sys.path
import numpy as np
import pandas as pd
np.seterr(divide='ignore', invalid='ignore')
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [2]:

```
from sklearn.datasets import load_wine
```

In [3]:

```
data = load_wine()
```

Конвертация датасета

In [4]:

```
# Конвертируем загруженный набор данных из sklearn в формат pandas dataframe.
df = pd.DataFrame(data.data, columns=data.feature_names)
type(df)
```

Out[4]:

```
pandas.core.frame.DataFrame
```

Основные характеристики датасета

In [5]:

```
# Первые 5 строк датасета
df.head()
```

Out[5]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonfla
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	

In [6]:

```
# Размер датасета 178 строк, 13 колонок
df.shape
```

Out[6]:

```
(178, 13)
```

In [7]:

```
# Список колонок
df.columns
```

Out[7]:

```
Index(['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
       'total_phenols', 'flavanoids', 'nonflavanoid_phenols',
       'proanthocyanins', 'color_intensity', 'hue',
       'od280/od315_of_diluted_wines', 'proline'],
      dtype='object')
```

In [8]:

```
# Список колонок с типами данных
df.dtypes
```

Out[8]:

```
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
```

In [9]:

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

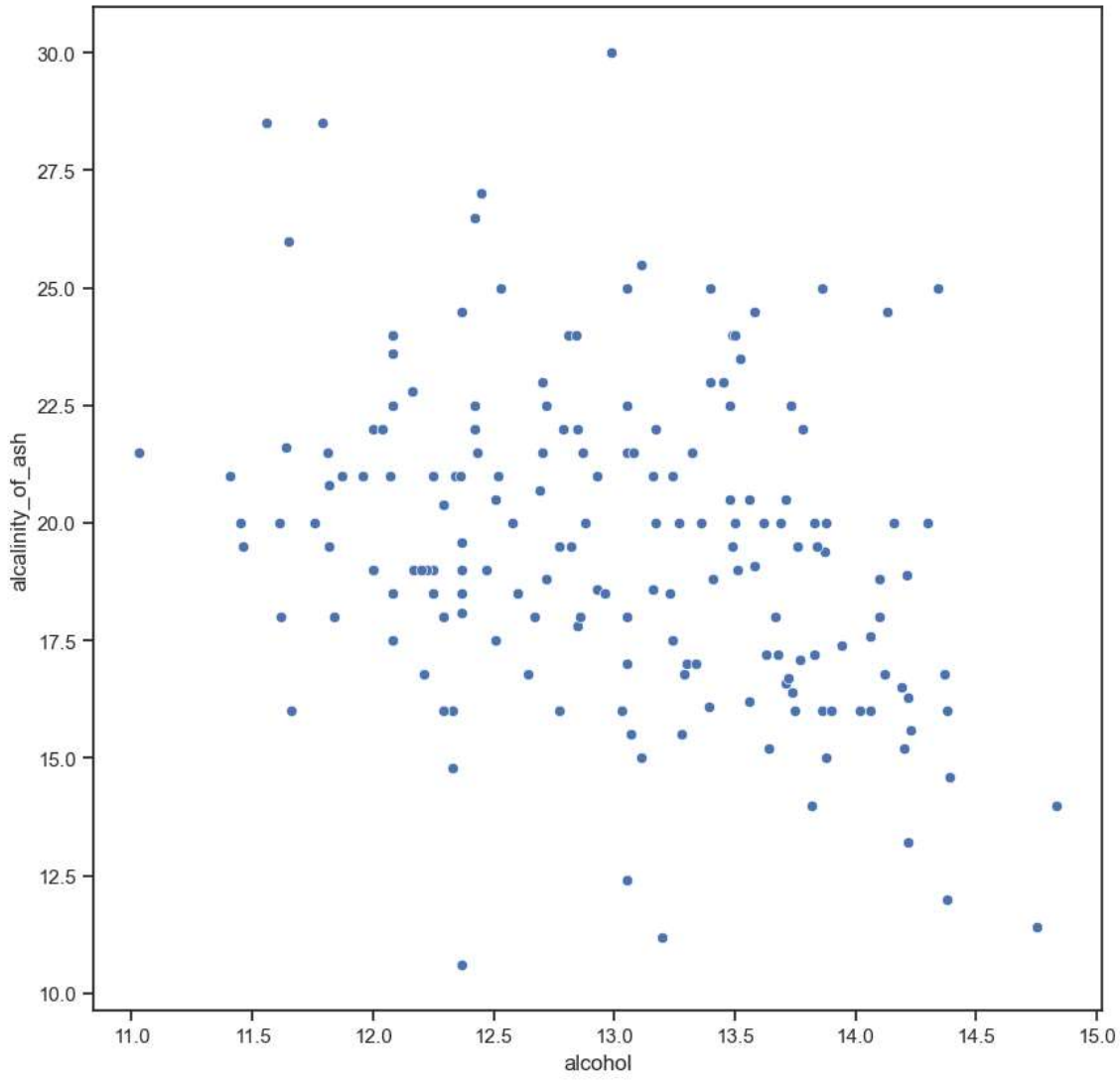
Визуальное исследование датасета

In [10]:

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

Out[10]:

```
<Axes: xlabel='alcohol', ylabel='alcalinity_of_ash'>
```

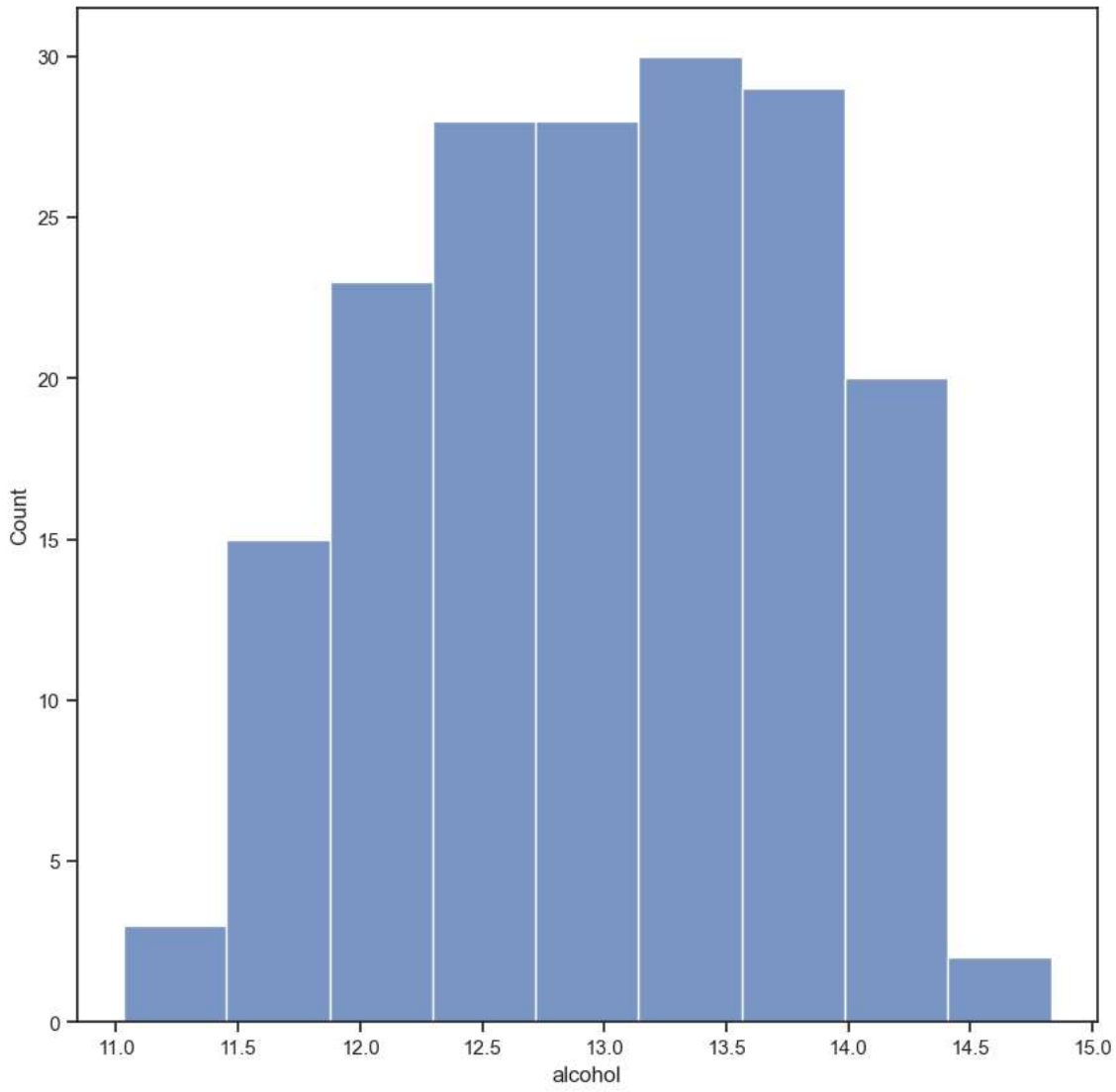


In [11]:

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

Out[11]:

```
<Axes: xlabel='alcohol', ylabel='Count'>
```

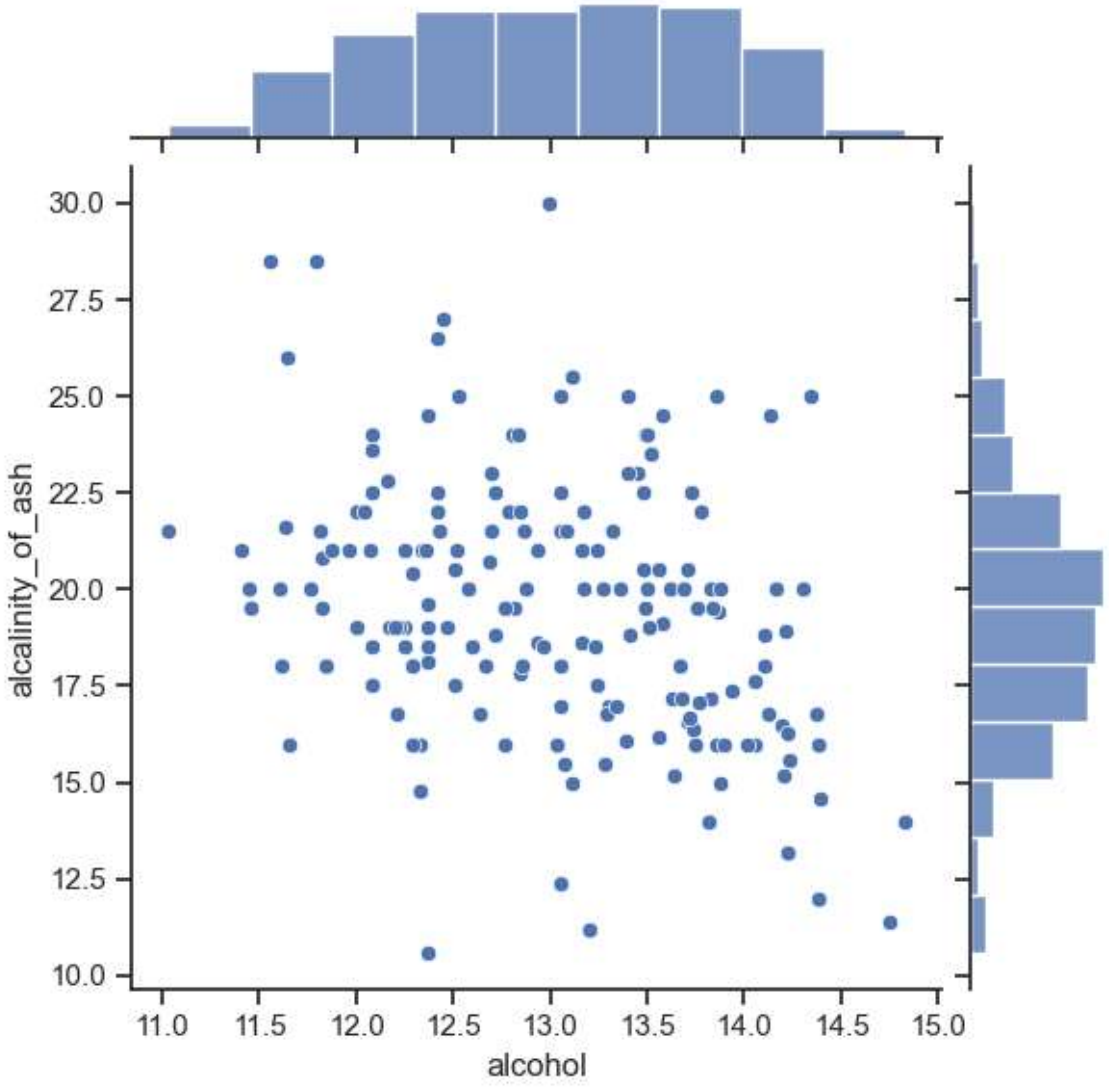


In [12]:

```
# Комбинация гистограмм и диаграмм рассеивания
sns.jointplot(x='alcohol', y='alcalinity_of_ash', data=df)
```

Out[12]:

```
<seaborn.axisgrid.JointGrid at 0x2465ad502e0>
```



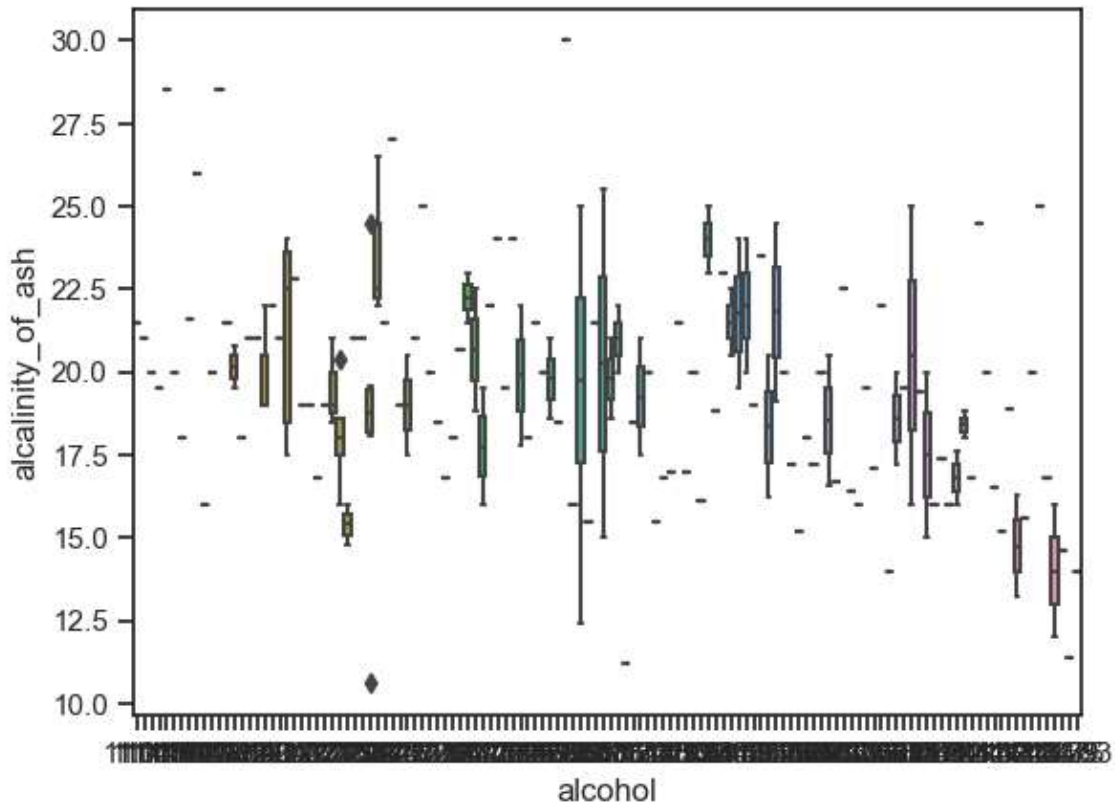
Построить график "Ящик с усами (boxplot)"

In [14]:

```
sns.boxplot( x=df["alcohol"], y=df["alcalinity_of_ash"])
```

Out[14]:

```
<Axes: xlabel='alcohol', ylabel='alcalinity_of_ash'>
```



In [ ]:

```

```