Близнева А.Е., РК№1, ИУ5Ц-81Б, вариант №26

Задание: 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?

Датасет: "https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine"

Импорт библиотек, загрузка данных

Ввод [1]:
```
import sys
sys.path
import pandas as pd
import numpy as np
import seaborn as sns
np.seterr(divide='ignore', invalid='ignore')
import matplotlib.pyplot as plt
from sklearn.datasets import load_wine
%matplotlib inline
```

Ввод [2]:
```
wine = load_wine()
df = pd.DataFrame(wine.data, columns=wine.feature_names)
df['TARGET'] = wine.target
```

Общее описание датасета

Ввод [3]:
```
# Первые пять строк датасета
df.head()
```

Out[3]:

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	

Ввод [4]:
```
# Описание датасета
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   alcohol                       178 non-null    float64
 1   malic_acid                    178 non-null    float64
 2   ash                           178 non-null    float64
 3   alcalinity_of_ash             178 non-null    float64
 4   magnesium                     178 non-null    float64
 5   total_phenols                 178 non-null    float64
 6   flavanoids                    178 non-null    float64
 7   nonflavanoid_phenols          178 non-null    float64
 8   proanthocyanins               178 non-null    float64
 9   color_intensity               178 non-null    float64
 10  hue                           178 non-null    float64
 11  od280/od315_of_diluted_wines  178 non-null    float64
 12  proline                       178 non-null    float64
 13  TARGET                        178 non-null    int32
dtypes: float64(13), int32(1)
memory usage: 18.9 KB
```

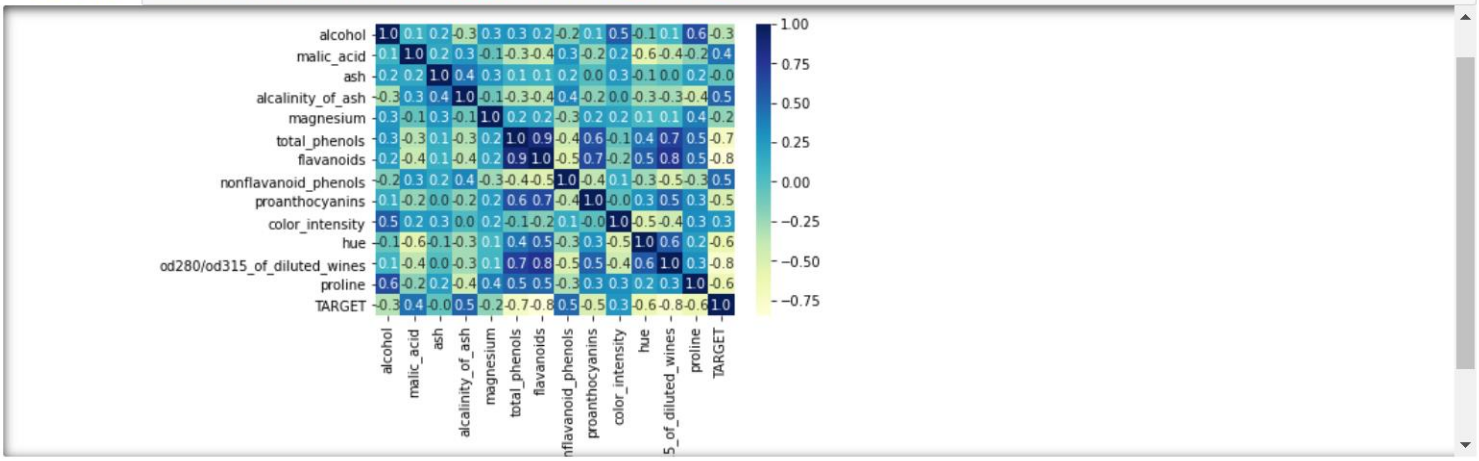
Ввод [5]:
```
# Проверим количество пустых значений
for col_empty in df.columns:
    empty_count = df[df[col_empty].isnull()].shape[0]
    print('{} - {}'.format(col_empty, empty_count))
```

```
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
```

Пустых значений не обнаружено.

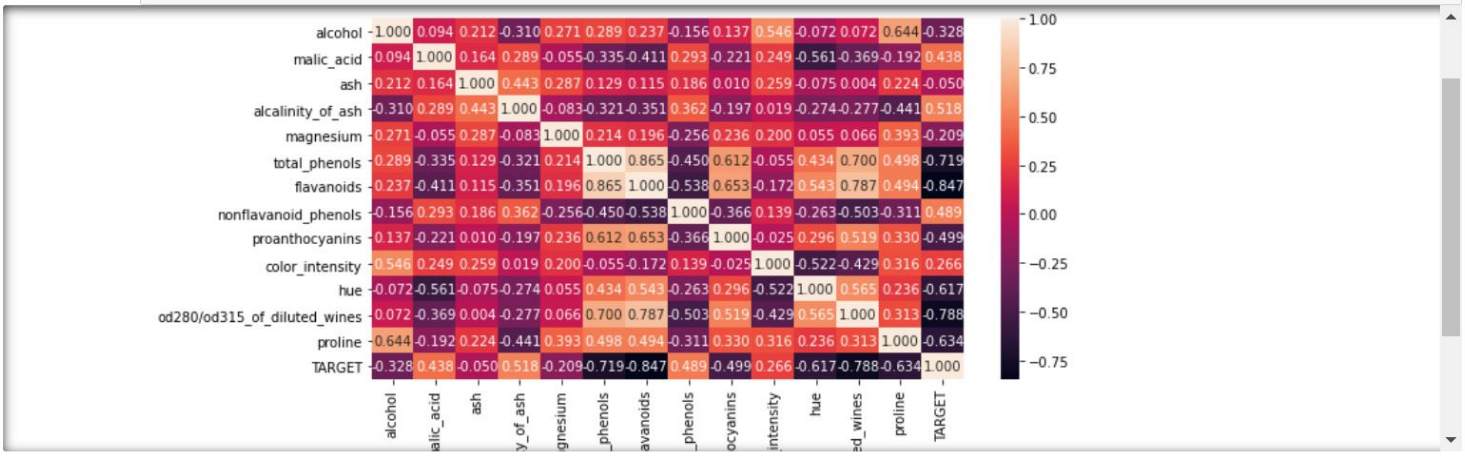
Корреляция признаков

Ввод [6]:
```
sns.heatmap(df.corr(), cmap='YlGnBu', annot=True, fmt='.1f')
```



Наиболее сильную корреляцию имеют признаки total_phenols и flavanoids. Это связано с тем, что флавониды относятся к классу полифенолов.

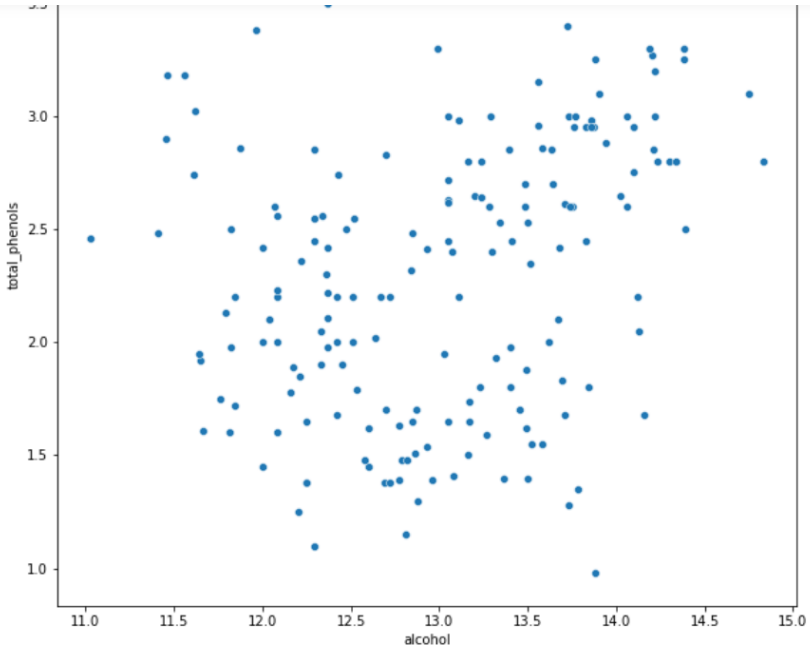
Ввод [7]:
```
fig, ax = plt.subplots(1, 1, sharex='col', sharey='row', figsize=(10,5))
fig.suptitle('Корреляционная матрица')
sns.heatmap(df.corr(), ax=ax, annot=True, fmt='.3f')
```



С целевым признаком TARGET сильнее всего коррелируют признаки "flavanoids", "od280/od315_of_diluted_wines", "total_phenols", "hue", "proline". Соответственно, их стоит учитывать для более информативного построения модели машинного обучения.

Диаграмма рассеивания

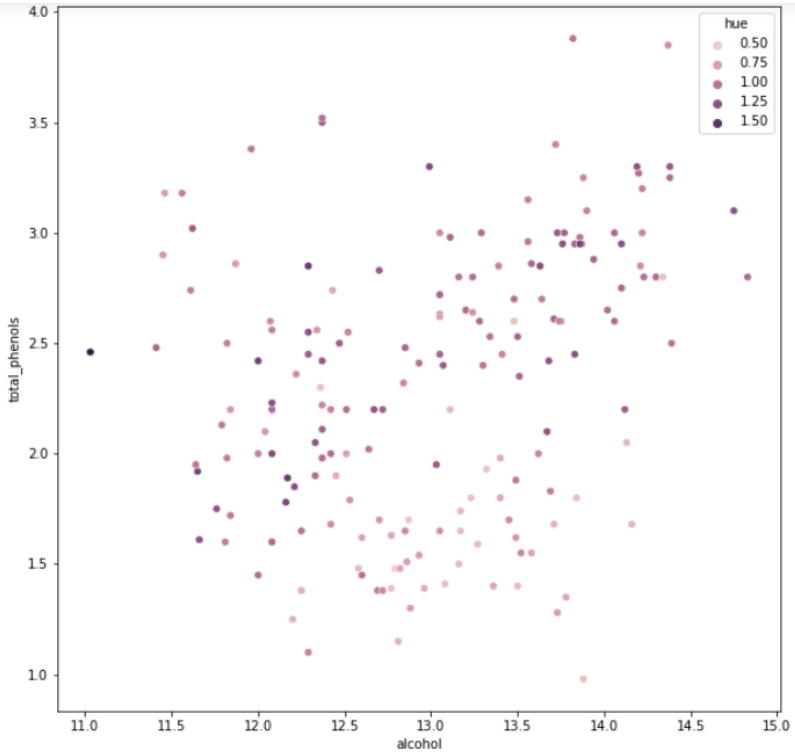
Ввод [8]:
```
#Диаграмма рассеивания для признаков total_phenols и alcohol
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='alcohol', y='total_phenols', data=df)
```



Данная диаграмма показывает количество фенолов в каждом проценте вина.

Ввод [9]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='alcohol', y='total_phenols', data=df, hue='hue')
```

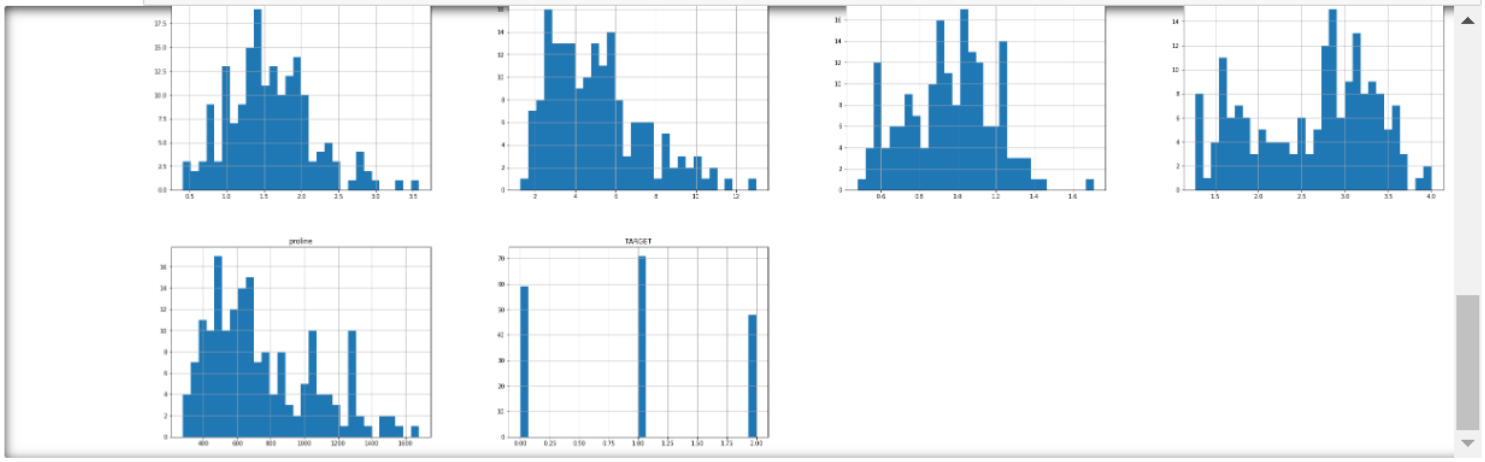
Out[9]: <AxesSubplot:xlabel='alcohol', ylabel='total_phenols'>



Такая же диаграмма показывает количество фенолов в каждом проценте вина, но еще добавили "hue", т.е. в каждой точке можем рассмотреть оттенок конкретного вина.

Гистограмма

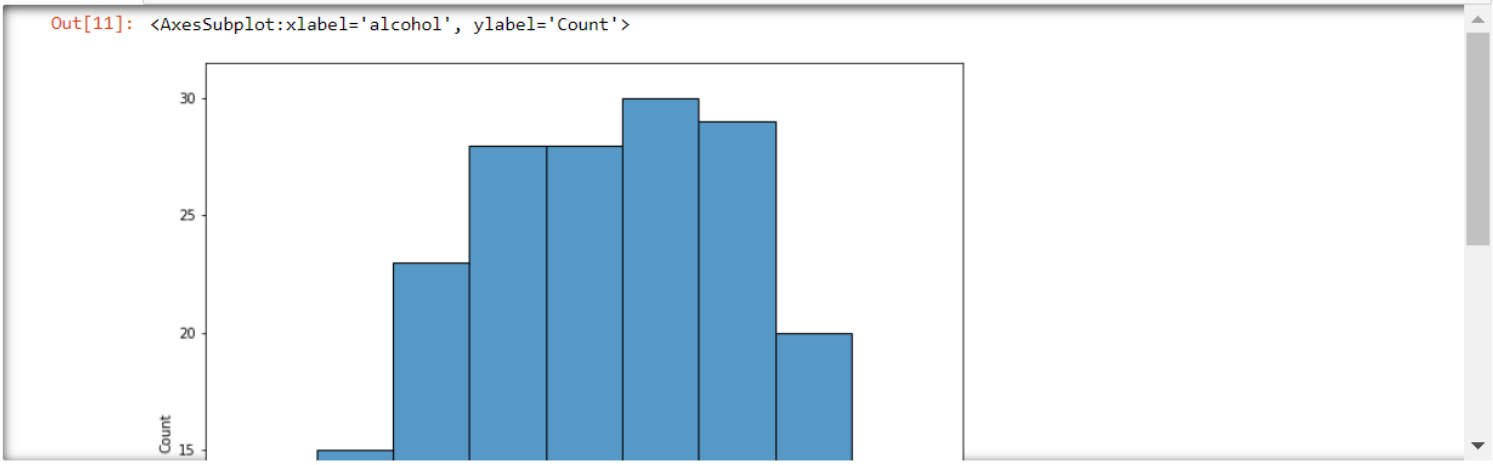
Ввод [10]:
```
# Гистограммы для всех признаков
df.hist(bins=30, figsize = (40,30))
```



Ввод [11]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.histplot(df['alcohol'])
```

Ввод [11]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.histplot(df['alcohol'])
```

Out[11]:
```
<AxesSubplot:xlabel='alcohol', ylabel='Count'>
```

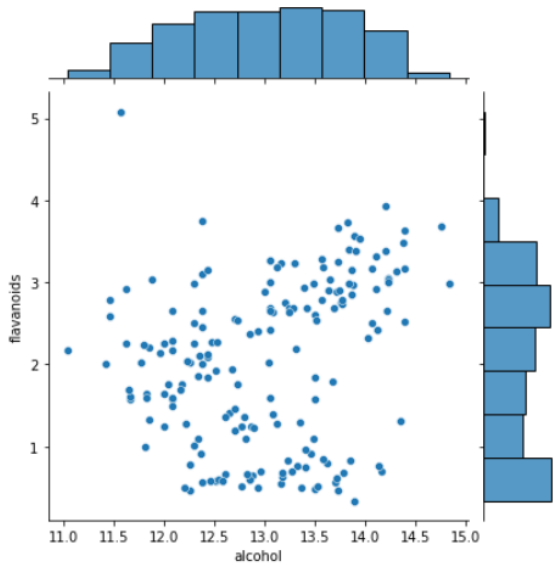


Данная гистограмма показывает наибольшее количество процента алкоголя в вине.

Jointplot

Ввод [12]:
```
sns.jointplot(x='alcohol', y='flavanoids', data=df)
```

Out[12]:
```
<seaborn.axisgrid.JointGrid at 0x246851dd730>
```

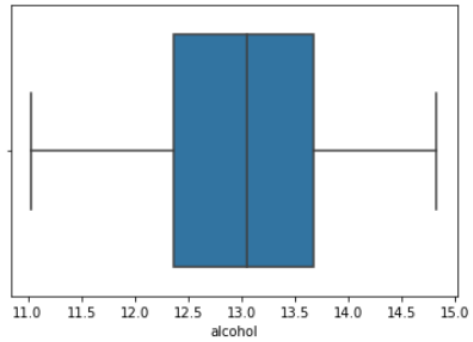


Комбинация гистограмм и диаграмм рассеивания.

"Ящик с усами"

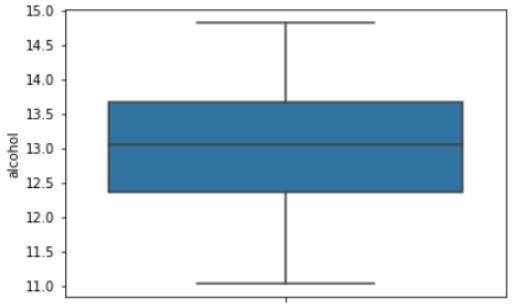
Ввод [13]:
```
# по оси абсцисс.
sns.boxplot(x=df['alcohol'])
```

Out[13]:
```
<AxesSubplot:xlabel='alcohol'>
```



Ввод [14]:
```
# По оси ординат
sns.boxplot(y=df['alcohol'])
```

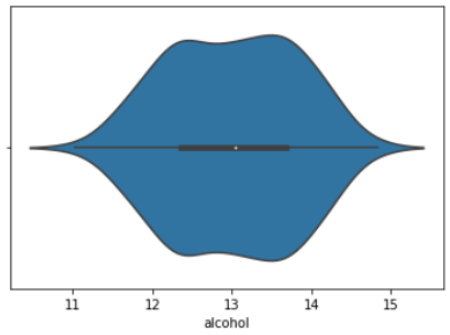
Out[14]:
```
<AxesSubplot:ylabel='alcohol'>
```



Скрипичная диаграмма

Ввод [15]:
```
sns.violinplot(x=df['alcohol'])
```

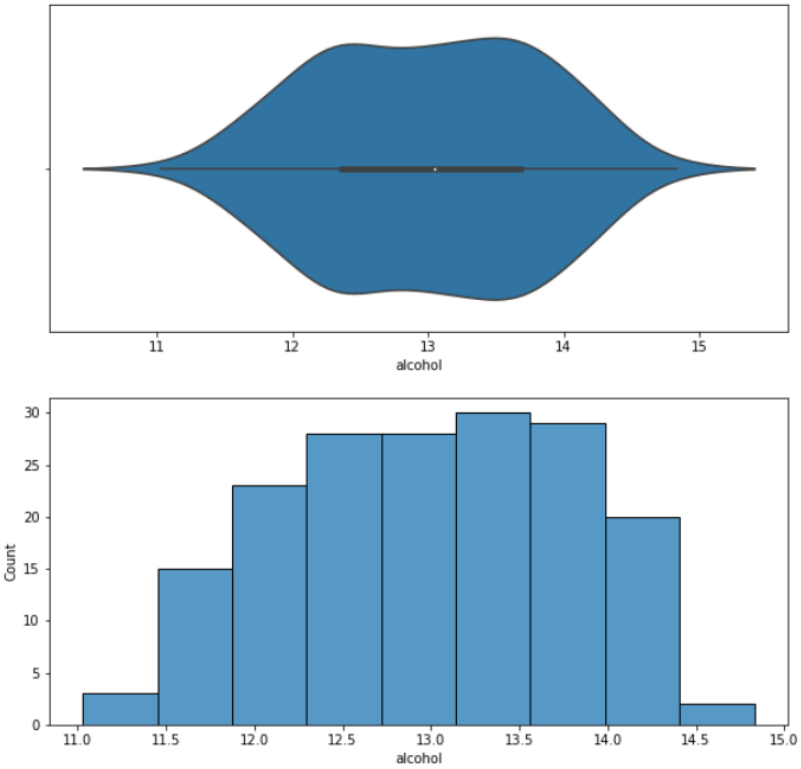
Out[15]:
```
<AxesSubplot:xlabel='alcohol'>
```



Скрипичная диаграмма показывает распределение плотности по краям диаграммы.

Ввод [16]:
```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=df['alcohol'])
sns.histplot(df['alcohol'])
```

Out[16]: <AxesSubplot:xlabel='alcohol', ylabel='Count'>



Из приведенных графиков видно, что скрипичная диаграмма действительно показывает распределение плотности.