Практическая работа № 9.
Визуализация данных на Python.

Задание 9. Визуализация данных файла tips.csv.

9.1. Метод **plot**. Настройка параметров метода. Применение метода ко всему датафрейму, к отдельному показателю (гистограмма распределения признака), к категориальным (нечисловым) переменным. Отображение двух показателей на графике.

9.2. Библиотека **Matplotlib**. Модуль **pyplot** . Методы **axes**() и **hist**().

9.3. Форматирование графика: заголовок диаграммы, подписи осей, легенда.

9.4. Создание линейных графиков. Функция matplotlib.pyplot.plot().

9.5. Графическая библиотека **Seaborn**. Гистограммы распределения признаков. Метод **distplot**(). Метод **countplot**(). Метод **boxplots**(). Метод **heatmap**().

9.6. Напишите отчет.

Примечание.

Файл ***tips.csv*** содержит информацию о ресторанах:

total_bill - общая сумма, уплаченная за заказ;

tip - размер чаевых;

sex - полклиента;

smoker - является ли клиент курильщиком;

day - день недели;

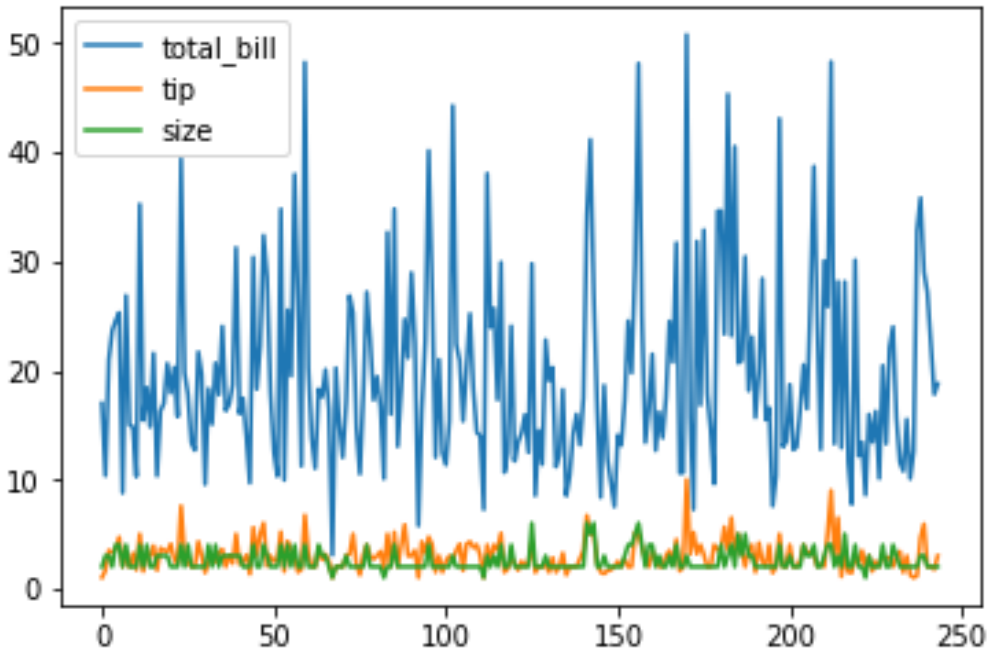
time - время (обед или ужин);

size - количество посетителей, обедавших за столом.

Пример9 .1

Обзор датасета

```
import  pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv('tips.csv')
df.head()
df.plot()
```

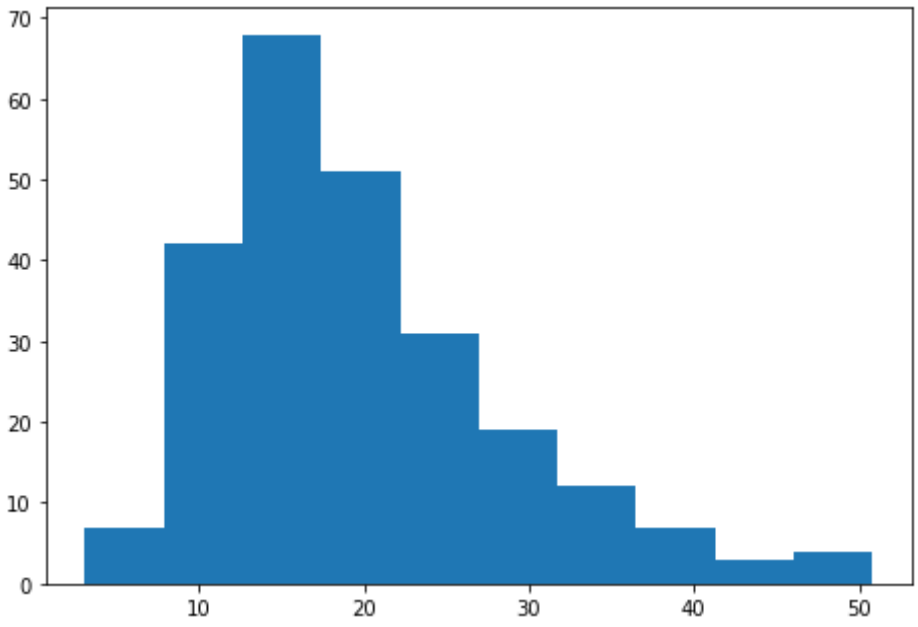


Гистограмма распределения признака

```
fig = plt.figure()
axes = fig.add_axes([0,0,1,1])
axes.hist(df['total_bill'])
```

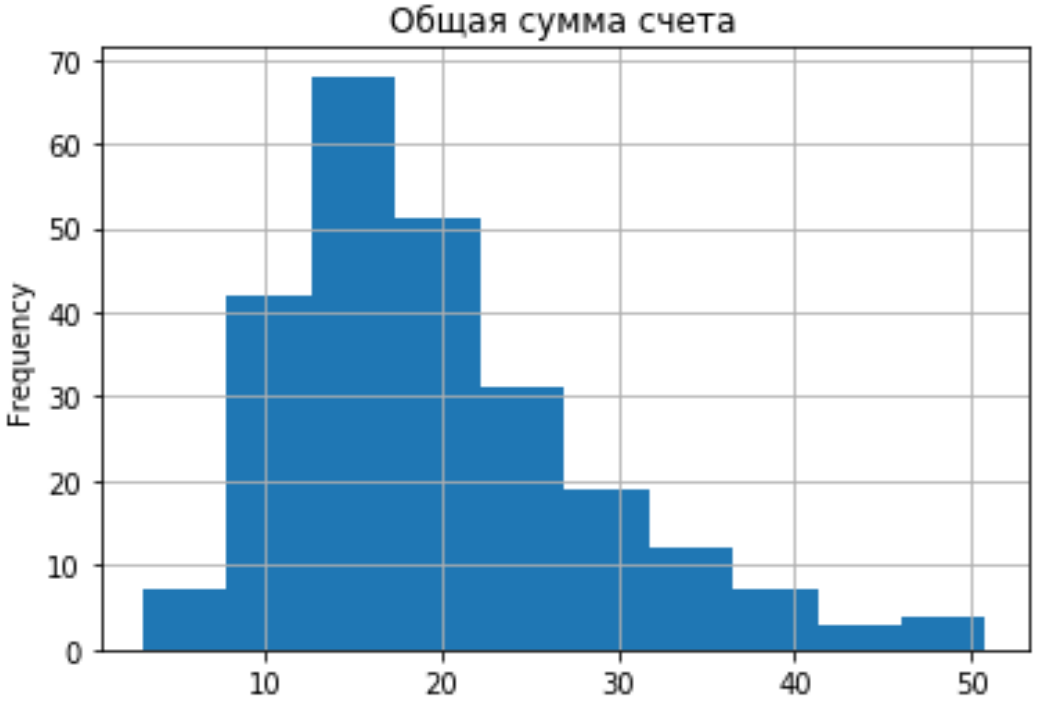
Вывод

```
(array([ 7., 42., 68., 51., 31., 19., 12.,  7.,  3.,  4.]),
 array([ 3.07 ,  7.844, 12.618, 17.392, 22.166, 26.94 , 31.714, 36.488,
        41.262, 46.036, 50.81 ]),
```



Наложим сетку, добавим заголовок на гистограмму.

```
df['total_bill'].plot(kind='hist',grid=True,
                                  title='Общая сумма счета')
```



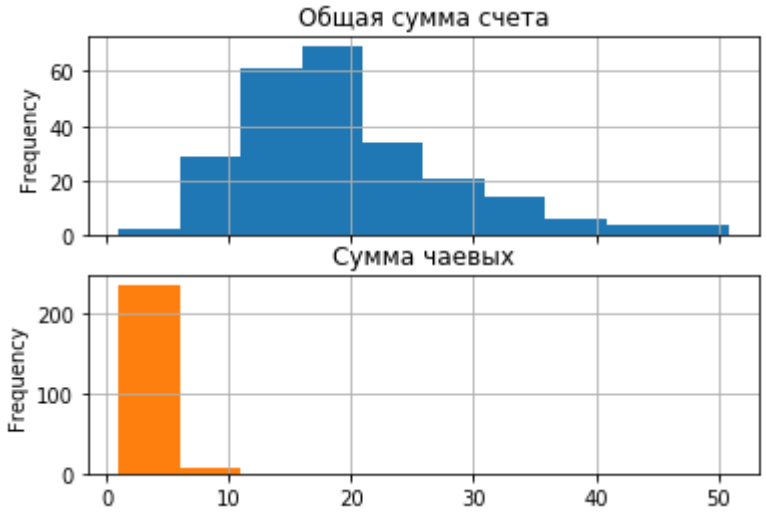
С категориальными (нечисловыми) переменными

```
df['day'].value_counts().plot(kind='bar',
     grid=True,
     color='darkblue',
     title='Количество посетителей по дням')
```

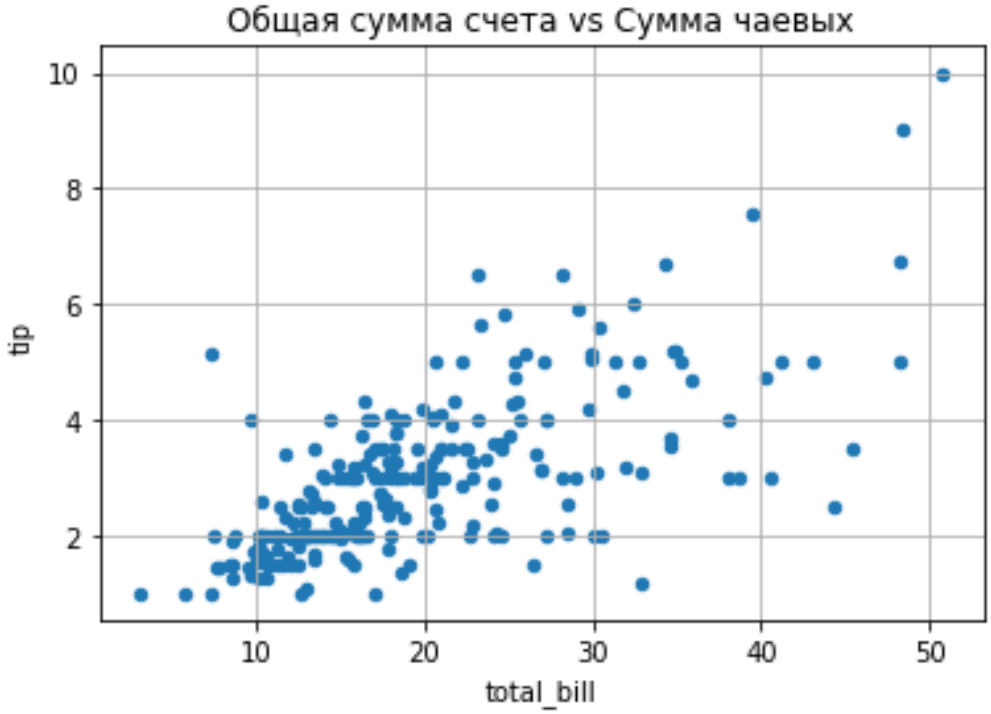
Пример 9.2

Отображение двух показателей на графике

```
df[['total_bill','tip']].plot(kind='hist',
        grid=True,
        subplots=True,
        title=['Общая сумма счета','Сумма чаевых'],
        legend=False)
```




```
df.plot(x='total_bill',y='tip',kind='scatter',grid=True,title='Общая сумма счета vs Сумма чаевых')
```

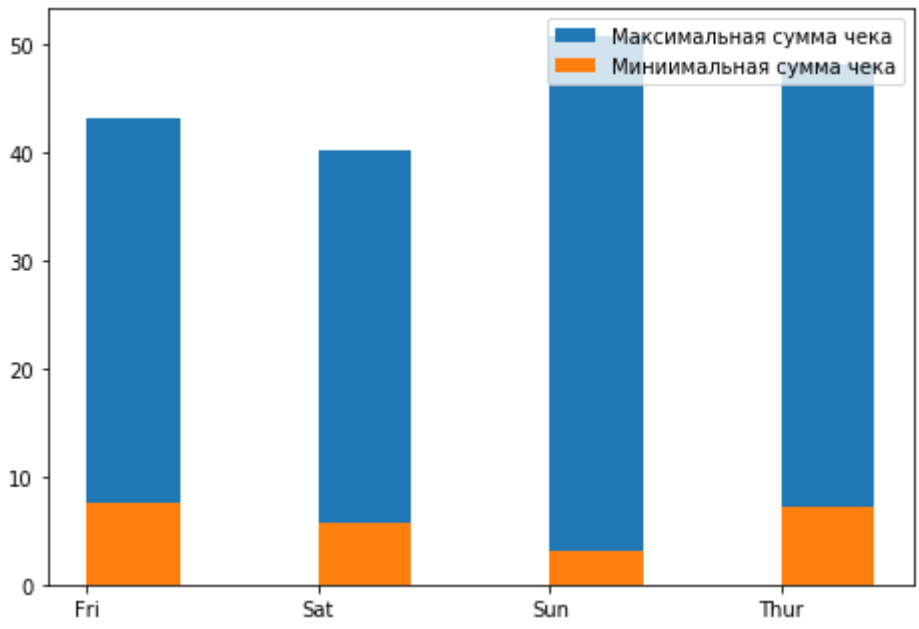


Пример 9.3

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv('tips.csv')
bills_per_day = df['day'].value_counts()
bills_per_day['day']=bills_per_day.index
bmax = df.groupby('day').aggregate(max)
bmin = df.groupby('day').aggregate(min)
bills_per_day['max'] = bmax['total_bill']
bills_per_day['min'] = bmin['total_bill']

fig = plt.figure()
axes = fig.add_axes([0,0,1,1])
axes.bar(x= bills_per_day['day'],
     height=bills_per_day['max'],
     width = 0.4,
     align = 'edge',
     label='Максимальная сумма чека')
axes.bar(x= bills_per_day['day'],
     height=bills_per_day['min'],
     width = 0.4,
     align = 'edge',
     label='Миниимальная сумма чека')
axes.legend(loc=1)
```

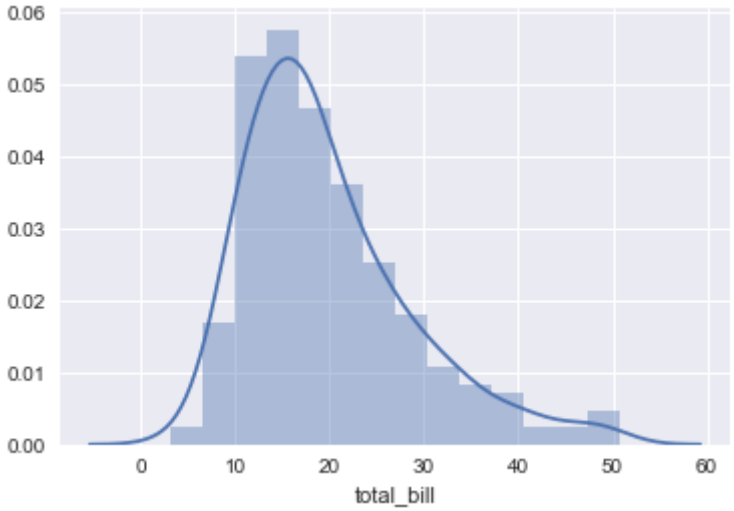
Оформление графика



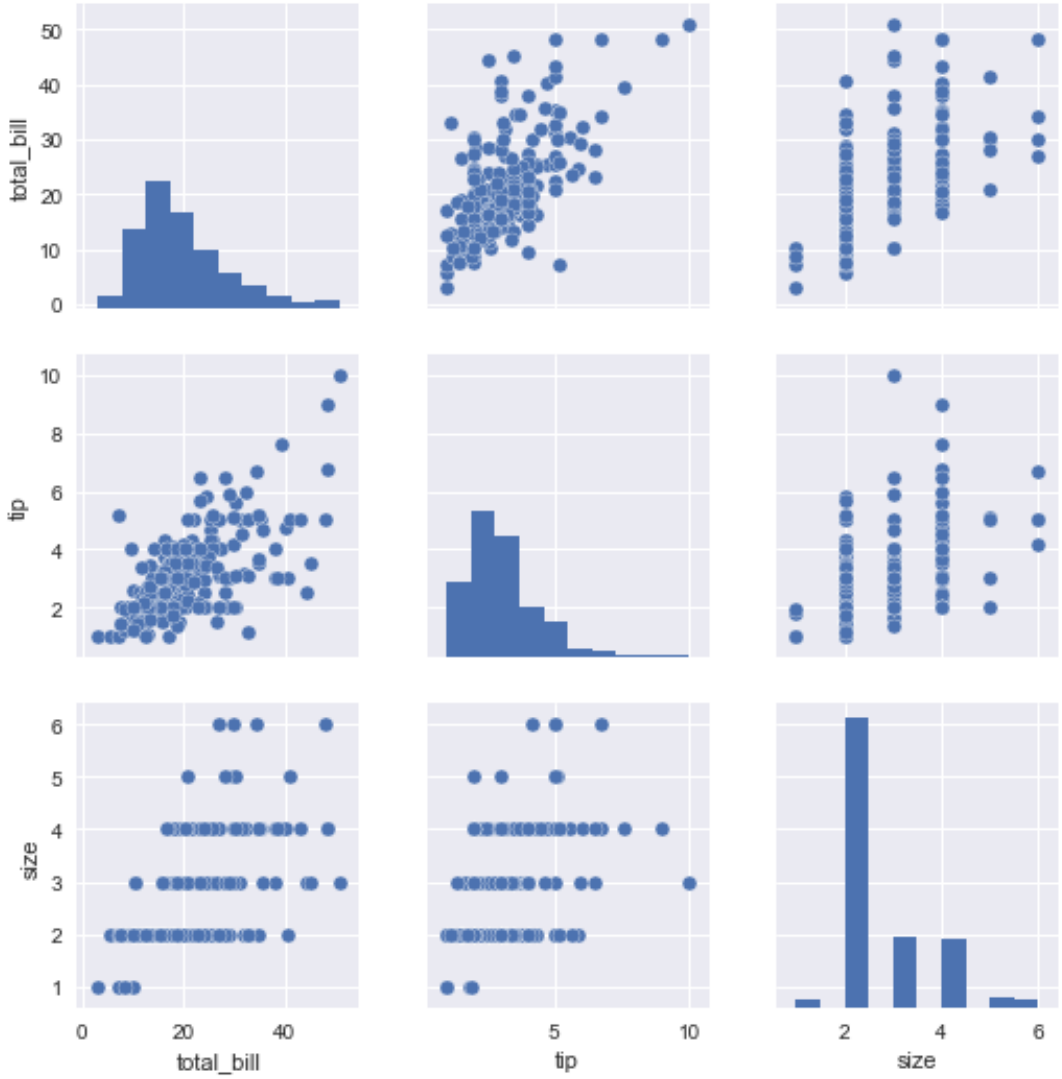
Пример 9.4

Гистограммы распределения признаков

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
df = pd.read_csv('tips.csv')
sns.set()
sns.distplot(df['total_bill'])
```

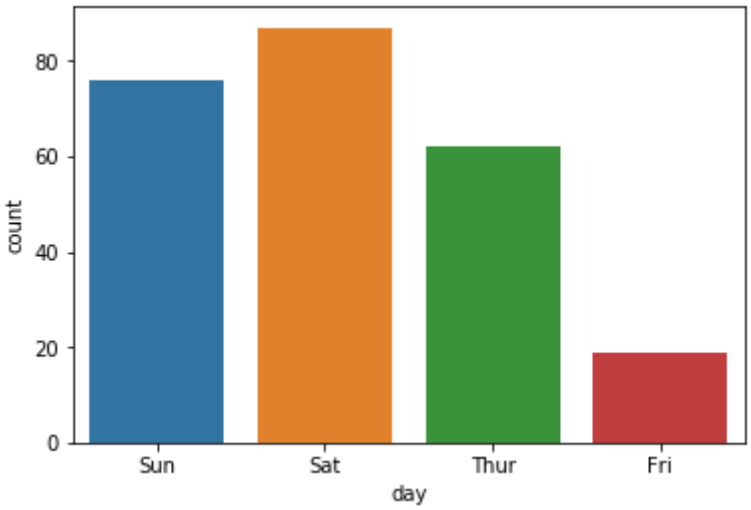



```
sns.pairplot(df)
```



Пример 9.5
Гистограммы распределения категориального признака

```
sns.countplot(x='day',data=df)
```




```
sns.boxplot(x='day',y='tip',data=df)
```

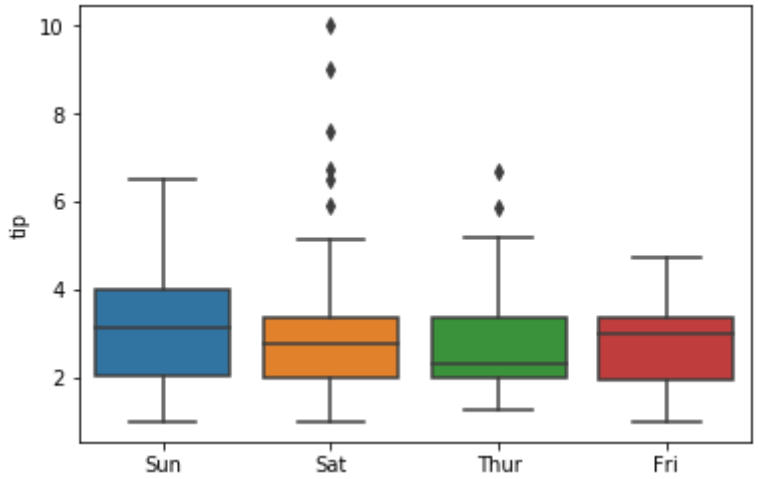
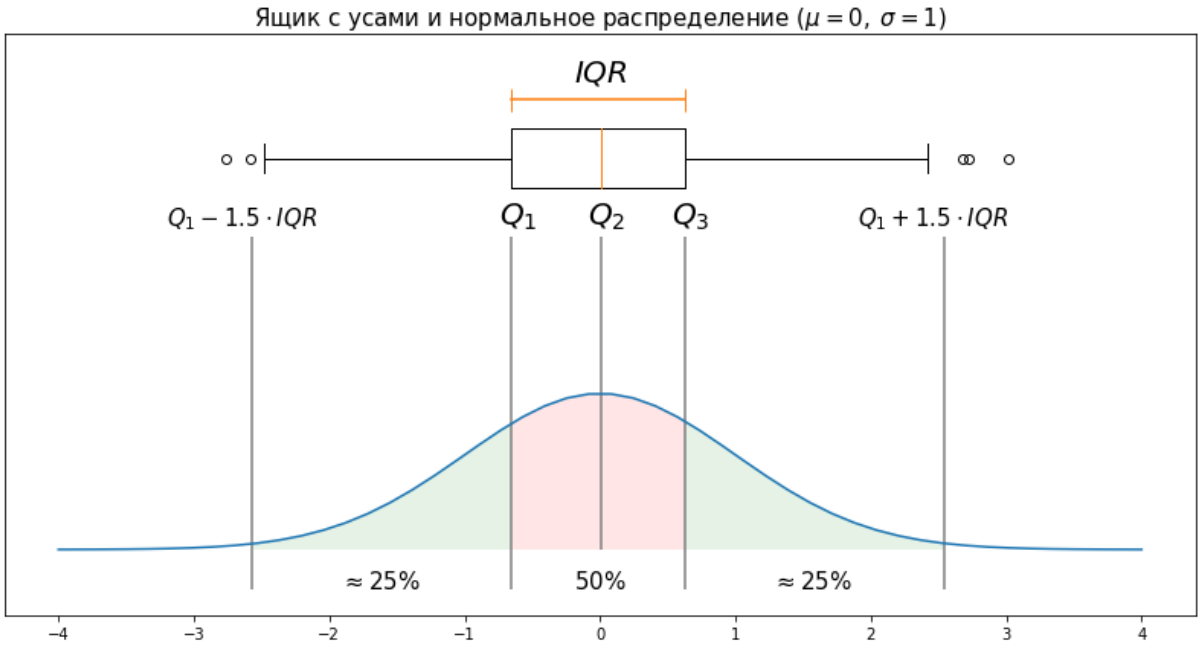


График boxplot и его описание



Границы ящика (прямоугольника) располагаются между нижним и верхним квартилем, или, как еще говорят, они располагаются между первым и третьим квартилем. В общем, они располагаются между Q_1 и Q_3. Q_1 - это, по сути, 0.25 квантиль, т.е. это такое значение в данных, ниже которого располагаются 25% всех остальных значений. Q_3 - это 0.75 квантиль, как вы поняли, ниже этого значения находится 75% всех остальных значений массива данных. Еще внутри ящика располагается полоска, которая соответствует медиане. Медианой, в свою очередь, так же является второй квартиль Q_2 , он же 0.5 квантиль - значение, ниже которого расположена ровно половина всех остальных значений массива данных. Длина ящика IQR равна межквартильному интервалу, т.е. $Q_3 - Q_1$. Также IQR определяет длину "усов" ящика, которая, обычно, не превышает $1.5 \cdot IQR$. Все значения, выходящие за границу усов, считаются выбросами, или, как еще говорят - аномальными значениями, которые обозначаются отдельными точками.

Пример 9.6
Графики табличного типа - корреляционный анализ.

```
correlation = df.corr()
sns.heatmap(correlation,annot=True,cmap='coolwarm')
```