

# Введение в анализ данных



Домашние задания:

выполнение и оформление



Задача аналитика не только в том,  
чтобы исследовать что-то,  
но и в том, чтобы это что-то  
правильно преподнести.



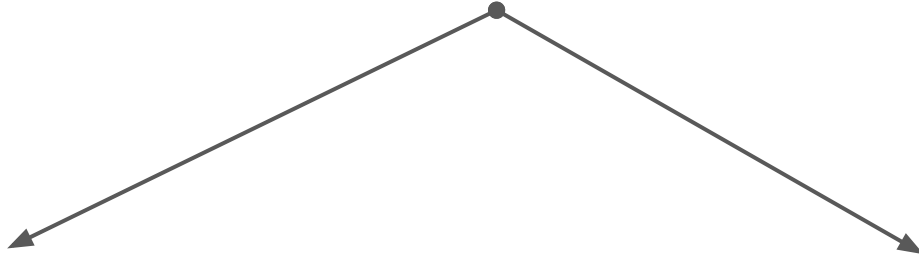


# Интерпретация условия задания



### Дано:

база данных о различных организациях, в которой есть поле categories – строка с описанием категорий, к которым может относиться организация.



### Задание по программированию:

В колонке “Описание” нужно задетектировать строки с подстрокой “Restaurant”, которая отделена “;”.  
Гарантируется корректность строк.

### Задание по анализу данных:

Посчитайте что-то для ресторанов, которые найдите по описанию объектов в данных.

Как они могут быть записаны?

“Restaurant; ...”

“Restaurant ...”

“Best restaurant ...”

“Restarant ...”

“Rest.; ...”

“Rest. ...”

“rest.; ...”



## Задача:

Какая группа имеет наибольший средний рейтинг?

### Решение по алгоритмам:

Смотрю:

Группа	Рейтинг
1	7.858904
2	7.814486

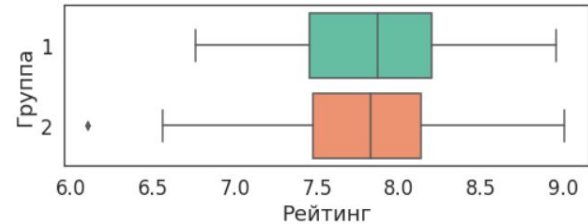
Пишу:

Группа 1 с рейтингом 7.858904

Зачем?

### Решение по анализу данных:

Посмотрим на данные шире.



Кажется, в среднем у них одинаковый рейтинг.



# Главное правило



# Главное правило

В анализе данных важно не написание кода,  
а исследование и результаты.

Код – это удобный инструмент.  
Как арифметика в математике.

*P.S. Если не согласны, можем для вас организовать задания в Excel*





# Оформление решений



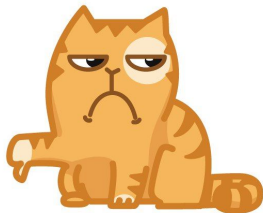


# Оформление решений

## Задание:

- сгенерировать случайные матрицу A и вектор b;
- перемножить;
- посчитать среднее полученных чисел;
- построить гистограмму значений.

Как можно оформить решение в Jupyter Notebook?



## Вариант 1

In [ ]:

Никогда так не делайте

```
import numpy as np
import matplotlib.pyplot as plt
n=50
m=30
A=np.random.uniform(size=(n,m))
b=np.random.uniform(size=(1,n))
c=b@A
print(f'Среднее {c.mean()}')
plt.figure(figsize=(10,4))
plt.hist(c[0],bins=15,color='green')
plt.show()
```

Для такого стиля используйте, например, PyCharm, но не Jupyter Notebook.

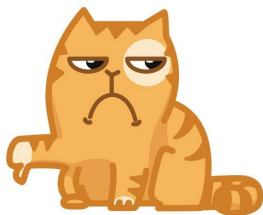


# Оформление решений

## Задание:

- сгенерировать случайные матрицу A и вектор b;
- перемножить;
- посчитать среднее полученных чисел;
- построить гистограмму значений.

Как можно оформить решение в Jupyter Notebook?



## Вариант 1

In [ ]:

Никогда так не делайте

```
import numpy as np
import matplotlib.pyplot as plt
n = 50
m = 30
A = np.random.uniform(size=(n, m))
b = np.random.uniform(size=(1, n))
c = b@A
print(f'Среднее {c.mean()}')
plt.figure(figsize=(10, 4))
plt.hist(c[0], bins=15, color='green')
plt.show()
```

Хотя бы расставим пробелы согласно PEP8

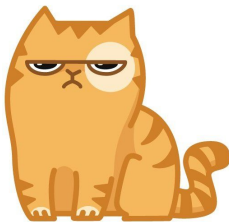


# Оформление решений

## Задание:

- сгенерировать случайные матрицу A и вектор b;
- перемножить;
- посчитать среднее полученных чисел;
- построить гистограмму значений.

Как можно оформить решение в Jupyter Notebook?



## Вариант 1

In [ ]:

Лучше так не делайте

```
import numpy as np
import matplotlib.pyplot as plt

n = 50
m = 30
A = np.random.uniform(size=(n, m))
b = np.random.uniform(size=(1, n))

c = b@A
print(f'Среднее {c.mean()}')

plt.figure(figsize=(10, 4))
plt.hist(c[0], bins=15, color='green')
plt.show()
```

А еще лучше – логическое разделение пустыми строками



# Оформление решений

## Задание:

- сгенерировать случайные матрицу A и вектор b;
- перемножить;
- посчитать среднее полученных чисел;
- построить гистограмму значений.

Как можно оформить решение в Jupyter Notebook?



## Вариант 2

Никогда так  
не делайте

In [ ]:

```
import numpy as np
import matplotlib.pyplot as plt
```

In [ ]:

```
n = 50
m = 30
```

In [ ]:

```
A = np.random.uniform(size=(n, m))
```

In [ ]:

```
b = np.random.uniform(size=(1, n))
```

In [ ]:

```
c = b@A
```

In [ ]:

```
c.mean()
```

In [ ]:

```
plt.figure(figsize=(10, 4))
plt.hist(c[0], bins=15, color='green')
plt.show()
```

Стиль “вермишель ячеек”



# Оформление решений

## Задание:

- сгенерировать случайные матрицу A и вектор b;
- перемножить;
- посчитать среднее полученных чисел;
- построить гистограмму значений.

Как можно оформить решение в Jupyter Notebook?

## Вариант 3

Хорошее решение



In [ ]:

```
import numpy as np
import matplotlib.pyplot as plt
```

Сгенерация данных

In [ ]:

```
n, m = 50, 30
m = 30
A = np.random.uniform(size=(n, m))
b = np.random.uniform(size=(1, n))
```

Посчитаем матричное произведение и усредним результат

In [ ]:

```
c = b@A
c.mean()
```

Построим гистограмму

In [ ]:

```
plt.figure(figsize=(10, 4))
plt.hist(c[0], bins=15, color='green')
plt.show()
```



# Оформление решений

1. Jupyter Notebook это не просто файл с кодом, это в том числе отчет о проведенном исследовании.
2. Читатель должен без труда разобраться в вашем ноутбуке.
3. Комментарии по логике решения должны быть в markdown-ячейках, а не в ячейках с кодом. В них же необходимо пояснять в общих чертах, что за код идет далее.
4. В коде должны быть поясняющие комментарии *непосредственно по коду*, обязательно перед крупными логическими блоками кода.



# Оформление решений

Проверяющие будут снижать баллы  
за плохо оформленные ноутбуки.

В частности, если проверяющий  
за разумное время не поймет логику решения  
из-за недостаточно хорошего оформления,  
он может его не оценивать вообще.



# Оформление графиков





# Важное правило

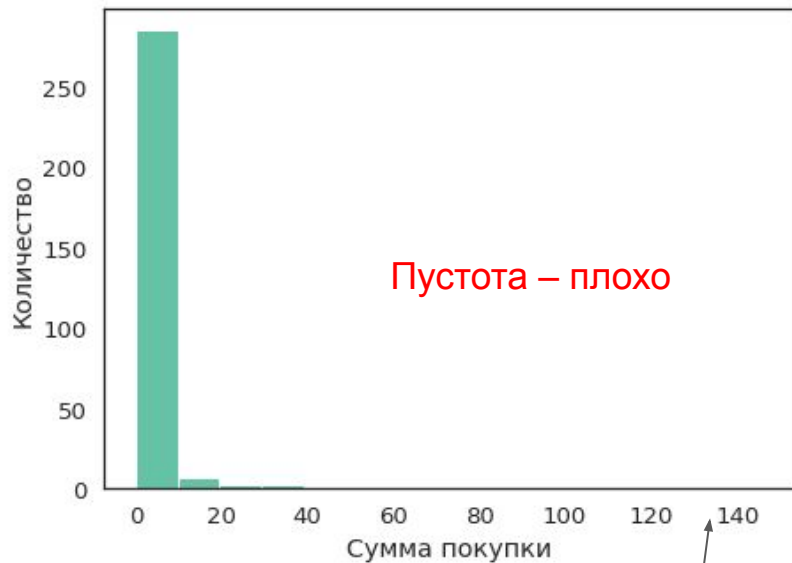
*Необходимые условия* качественного графика:

- на нем все четко видно;
- все объекты сбалансированы;
- если график вырезать из контекста,  
то из него понятно, что на нем изображено.



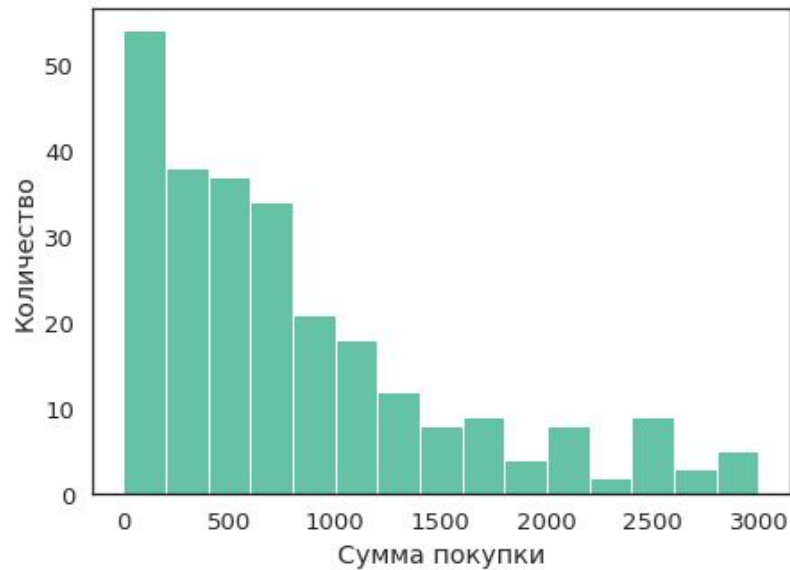
# Гистограммы

Неинформативно



Тут один элемент

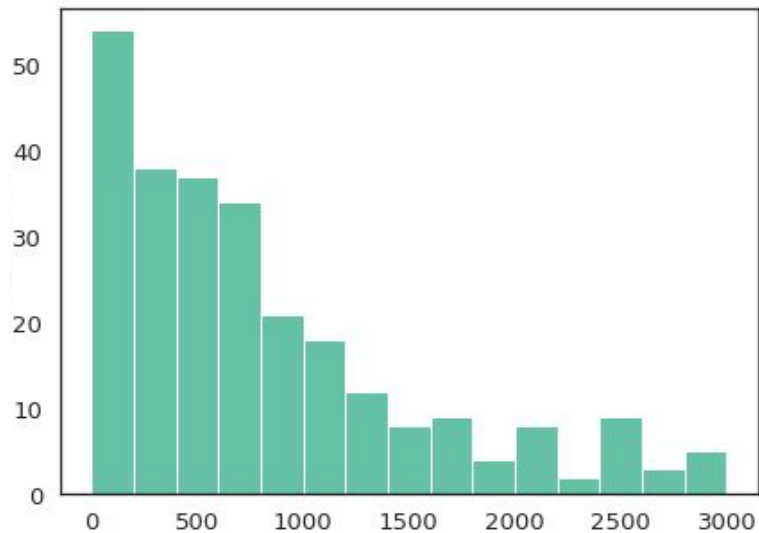
Информативно



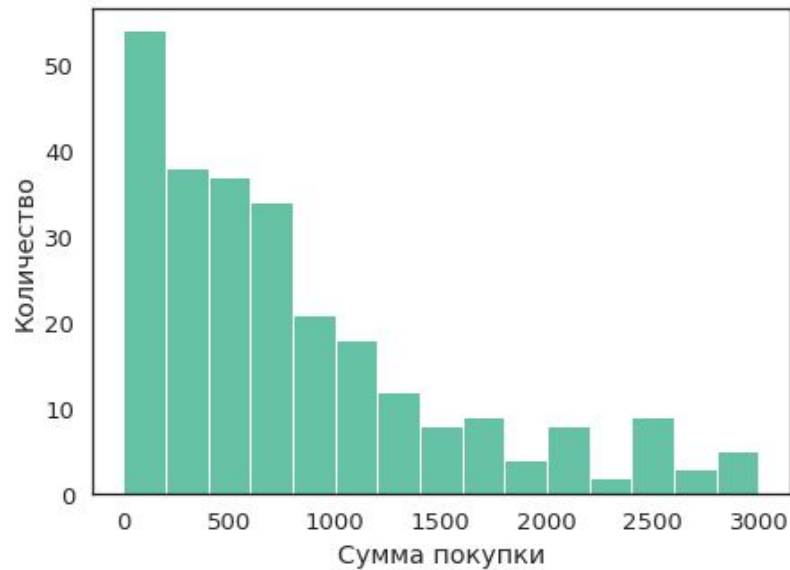


# Подписывайте оси

Плохо



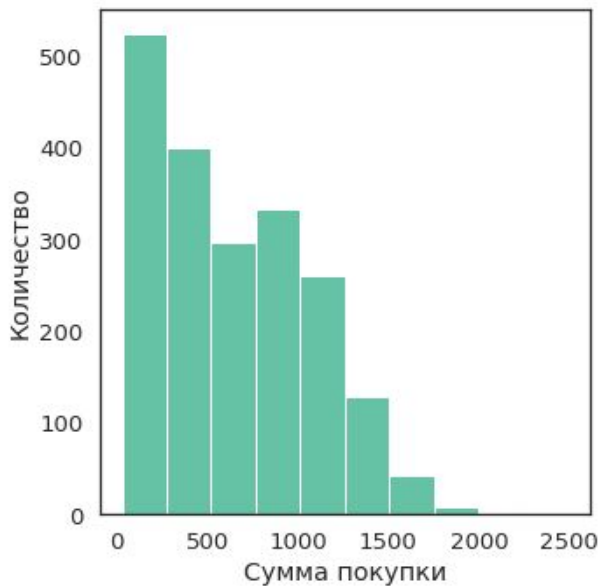
Хорошо



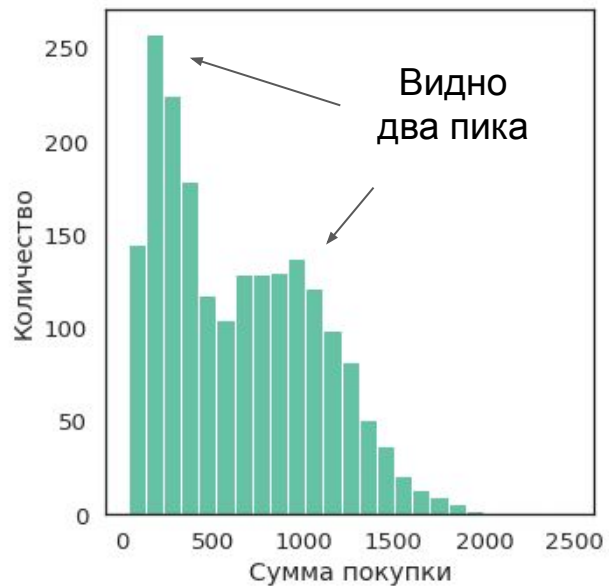


# Подбирайте количество бинов

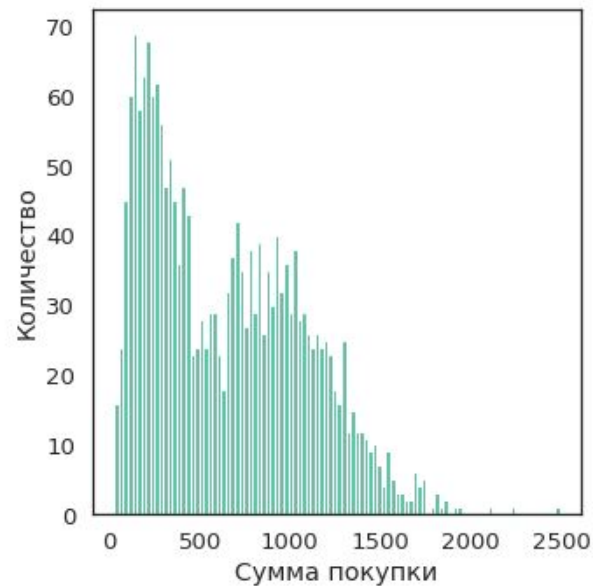
Плохо



Хорошо



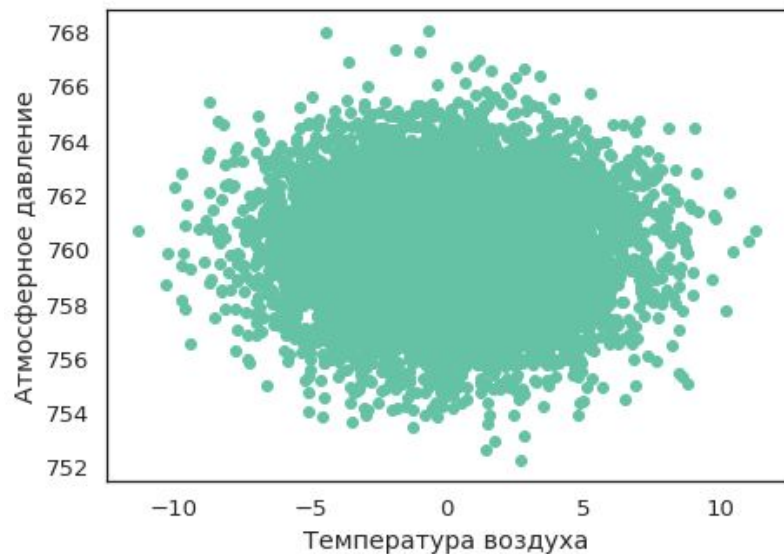
Плохо



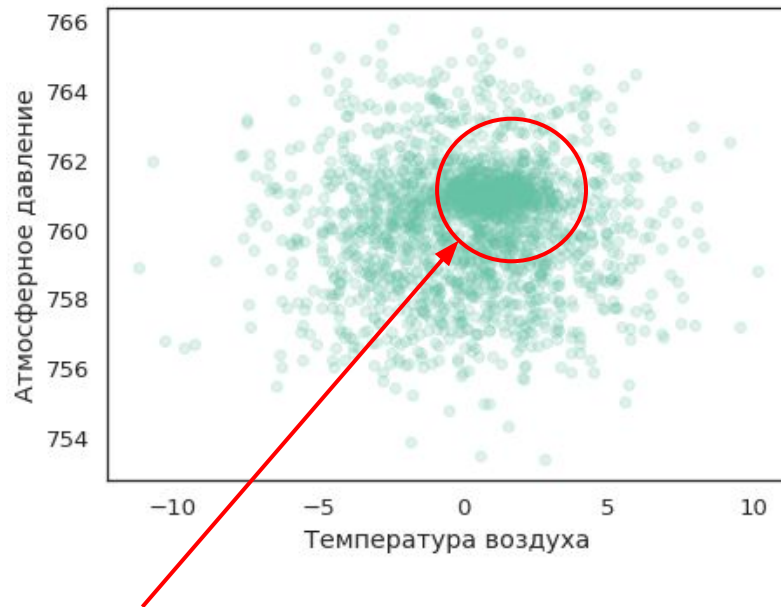


# Устанавливайте прозрачность точек

Плохо



Хорошо



Виден более плотный сгусток точек

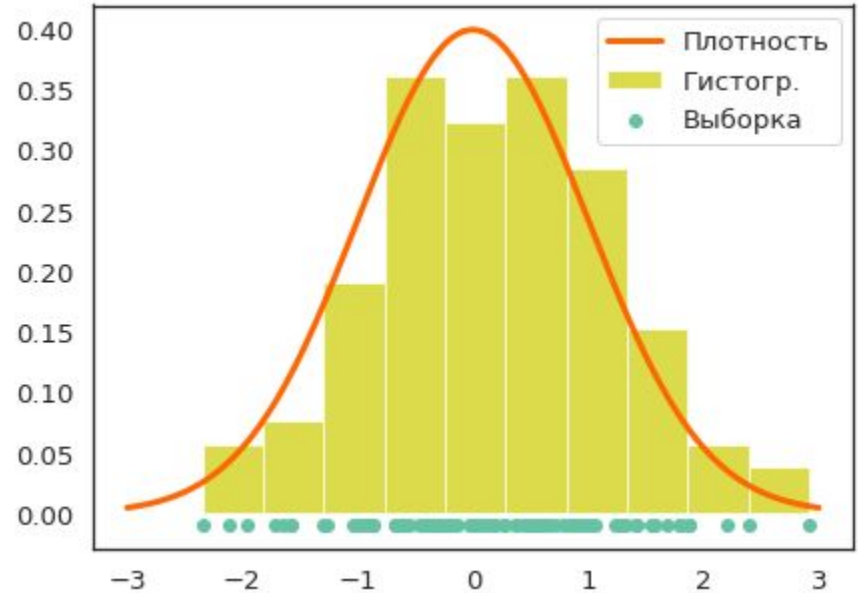


# Легенда – хорошо, когда текст краткий и емкий

Плохо



Хорошо



Не пишите тривиальное



# Оформление выводов



# Оформление выводов

## Плохо

- “я решил задачу”
- Практика подтвердила теорию (без пояснения)
- Пересказ условия задачи
- Вывод-отписка
- Огромное сочинение

## Хорошо

- “Эксперимент показал, что данные ведут себя так-то ...”
- “Клиенты хорошо разделяются на три кластера – ...”
- “Эксперимент подтвердил теорию тем, что ...”
- Желательно явно показывать, из какой части исследования какой вывод следует.





**Успехов в домашнем задании!**