

# Введение в машинное обучение

Н.В. Артамонов

8 апреля 2025 г.

## Содержание

|                            |          |
|----------------------------|----------|
| <b>1 Введение в Python</b> | <b>1</b> |
| 1.1 Pandas . . . . .       | 1        |
| 1.2 Визуализация . . . . . | 4        |
| <b>2 Preprocessing</b>     | <b>7</b> |
| <b>3 Кластеризация</b>     | <b>8</b> |

## 1 Введение в Python

### 1.1 Pandas

#1. Загрузите датасет `countries`. Вычислите описательные статистики для каждой переменной.

#2. Загрузите датасет `sleep75`.

1. вычислите размер датасета (число наблюдений & число переменных)
2. Заполните следующую таблицу со значениями переменных

| index | sleep | totwrk | age | male |
|-------|-------|--------|-----|------|
| 0     |       |        |     |      |
| 5     |       |        |     |      |
| 100   |       |        |     |      |
| 700   |       |        |     |      |

3. Вычислите корреляционную матрицу для следующих переменных:  
sleep, totwrk, age
4. Заполните следующую таблицу

| Desc.Stat    | sleep | totwrk | age | hrwage |
|--------------|-------|--------|-----|--------|
| max          |       |        |     |        |
| min          |       |        |     |        |
| mean         |       |        |     |        |
| median       |       |        |     |        |
| st.dev       |       |        |     |        |
| var          |       |        |     |        |
| 1st quartile |       |        |     |        |
| 3rd quartile |       |        |     |        |

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
  - (a) sleep>3000
  - (b) totwrk<2000
  - (c) age>40
  - (d) age<30
6. Сколько наблюдений с условием totwrk=0? Кто эти люди?
7. Есть ли в датасете пропущенные наблюдения? Сколько их?

**#3.** Загрузите датасет **Electricity**.

1. вычислите размер датасета (число наблюдений & число переменных)
2. заполните следующую таблицу со значениями переменных

| index | cost | q | pl | pk | pf |
|-------|------|---|----|----|----|
| 1     |      |   |    |    |    |
| 15    |      |   |    |    |    |
| 48    |      |   |    |    |    |
| 87    |      |   |    |    |    |

3. Вычислите корреляционную матрицу для следующих переменных:  
cost, q, pl, pk, pf
4. Заполните следующую таблицу

| Desc.Stat    | cost | q | pl | pk | pf |
|--------------|------|---|----|----|----|
| max          |      |   |    |    |    |
| min          |      |   |    |    |    |
| mean         |      |   |    |    |    |
| median       |      |   |    |    |    |
| st.dev       |      |   |    |    |    |
| var          |      |   |    |    |    |
| 1st quartile |      |   |    |    |    |
| 3rd quartile |      |   |    |    |    |

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
  - (a)  $\text{cost} > 40$
  - (b)  $q < 5000$
  - (c)  $q > 4000$
  - (d)  $20 < \text{cost} < 50$
6. Есть ли в датасете пропущенные наблюдения? Сколько их?

#4. Загрузите датасет **wage2**.

1. вычислите размер датасета (число наблюдений & число переменных)
2. заполните следующую таблицу со значениями переменных

| index | wage | hours | IQ | educ | exper | age |
|-------|------|-------|----|------|-------|-----|
| 1     |      |       |    |      |       |     |
| 25    |      |       |    |      |       |     |
| 179   |      |       |    |      |       |     |
| 800   |      |       |    |      |       |     |

3. Вычислите корреляционную матрицу для следующих переменных: wage, hours, IQ, educ, exper
4. Заполните следующую таблицу

| Desc.Stat    | wage | hours | IQ | educ | exper | wage |
|--------------|------|-------|----|------|-------|------|
| max          |      |       |    |      |       |      |
| min          |      |       |    |      |       |      |
| mean         |      |       |    |      |       |      |
| median       |      |       |    |      |       |      |
| st.dev       |      |       |    |      |       |      |
| var          |      |       |    |      |       |      |
| 1st quartile |      |       |    |      |       |      |
| 3rd quartile |      |       |    |      |       |      |

Замечание: 1st/3rd квантили – 25%/75% квантили соответственно.

5. Сколько наблюдения соответствуют следующим условиям
  - (a) wage>1000
  - (b) age<40
  - (c) exper>10
  - (d) 100<IQ<130

6. Есть ли в датасете пропущенные наблюдения? Сколько их?

#5. Загрузите датасет **Labour**. Создайте новый датасет, содержащий log-переменные из исходного датасета.

#6. Загрузите датасет **Electricity**. Создайте новый датасет, содержащий log-переменные из исходного датасета.

## 1.2 Визуализация

#7. Загрузите датасет **sleep75**.

1. нарисуйте гистограммы для переменных sleep, totwrk, age, hrwage, educ
2. нарисуйте гистограмму с накоплением для sleep относительно male

3. нарисуйте гистограмму с накоплением для totwrk относительно south
4. нарисуйте гистограмму с накоплением для totwrk относительно smsa
5. нарисуйте диаграмму рассеяния sleep vs totwrk
6. нарисуйте диаграмму рассеяния sleep vs totwrk с группировкой по male
7. нарисуйте диаграмму рассеяния sleep vs age
8. нарисуйте диаграмму рассеяния sleep vs age с группировкой по south
9. нарисуйте диаграмму рассеяния sleep vs edu
10. нарисуйте диаграмму рассеяния sleep vs edu с группировкой по smsa
11. визуализируйте корреляционную матрицу для следующих переменных: sleep, totwrk, age

**#8. Загрузите датасет Labour.**

1. нарисуйте гистограммы для каждой переменной
2. нарисуйте гистограммы для log-переменных output, capital, labour, wage
3. нарисуйте диаграммы рассеяния output vs других переменных
4. нарисуйте диаграммы рассеяния  $\log(\text{output})$  vs log других переменных
5. визуализируйте корреляционную матрицу для всех переменных
6. визуализируйте корреляционную матрицу для log-переменных

**#9. Загрузите датасет Electricity.**

1. нарисуйте гистограммы для переменных cost, q, pf, pk, pl
2. нарисуйте гистограммы для log-переменных cost, q, pf, pk, pl

3. нарисуйте диаграммы рассеяния `cost` vs других переменных
4. нарисуйте диаграммы рассеяния  $\log(\text{cost})$  vs  $\log$  других переменных
5. визуализируйте корреляционную матрицу для всех переменных
6. визуализируйте корреляционную матрицу для  $\log$ -переменных

#10. Загрузите датасет `diamonds`.

1. нарисуйте гистограммы для переменных `price`, `carat`
2. нарисуйте гистограммы для  $\log$ -переменных `price`, `carat`
3. нарисуйте гистограмму с накоплением для `price` относительно `cut`
4. нарисуйте гистограмму с накоплением для `carat` относительно `clarity`
5. нарисуйте гистограмму с накоплением для  $\log(\text{price})$  относительно `color`
6. нарисуйте гистограмму с накоплением для  $\log(\text{carat})$  относительно `color`
7. нарисуйте диаграмму рассеяния `price` vs `carat`
8. нарисуйте диаграмму рассеяния  $\log\text{-price}$  vs  $\log\text{-carat}$
9. нарисуйте диаграмму рассеяния  $\log\text{-price}$  vs  $\log\text{-carat}$  с группировкой по `cut`
10. нарисуйте диаграмму рассеяния  $\log\text{-price}$  vs  $\log\text{-carat}$  с группировкой по `color`
11. нарисуйте диаграмму рассеяния  $\log\text{-price}$  vs  $\log\text{-carat}$  с группировкой по `clarity`

#11. Загрузите датасет `Diamond`.

1. нарисуйте гистограммы для переменных `price`, `carat`
2. нарисуйте гистограммы для  $\log$ -переменных `price`, `carat`

3. нарисуйте гистограмму для price с группировкой относительно переменной certification
4. нарисуйте гистограмму для carat с накоплением относительно clarity
5. нарисуйте гистограмму для log(price) с накоплением относительно colour
6. нарисуйте гистограмму для log(carat) с накоплением относительно colour
7. нарисуйте диаграмму рассеяния price vs carat
8. нарисуйте диаграмму рассеяния log-price vs log-carat
9. нарисуйте диаграмму рассеяния log-price vs log-carat с группировкой по certification
10. нарисуйте диаграмму рассеяния log-price vs log-carat с группировкой по colour
11. нарисуйте диаграмму рассеяния log-price vs log-carat с группировкой по clarity

#12. Загрузите датасет `countries`.

1. Постройте гистограммы для всех переменных
2. Постройте диаграмму рассеяния Население vs ВВП д/н
3. Постройте диаграмму рассеяния ИРЧП vs ВВП д/н
4. Постройте диаграмму рассеяния Безработица vs ВВП д/н

## 2 Preprocessing

**Замечание:** рассмотрите следующие преобразования переменных:

- *квантильное (для гауссового распределения)*
- *Box-Cox*
- *Yeo-Johnson*

#1. Загрузите датасет `Labour`

1. Нарисуйте гистограммы для каждой переменной в уровнях и после стандартных преобразований
2. Нарисуйте диаграммы рассеяния в уровнях и после стандартных преобразований

#2. Загрузите датасет `diamonds`. Для переменных `price`, `carat`, `x`, `y`, `z`

1. Нарисуйте гистограммы для каждой переменной в уровнях и после стандартных преобразований
2. Нарисуйте диаграммы рассеяния в уровнях и после стандартных преобразований

### 3 Кластеризация

**Важно** обязательно проводим предварительную обработку данных:

- удаление пропущенных значений
- нормировка
- преобразование категориальных признаков

#1. Для набора данных `countries` проведите разбиение на кластеры следующими методами:

| Число кластеров | Метод         |
|-----------------|---------------|
| 3               | k-средних     |
| 4               | k-средних     |
| 5               | k-средних     |
| 3               | иерархическая |
| 4               | иерархическая |
| 5               | иерархическая |

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета



#2. Для набора данных `countries` найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin

#3. Из набора данных `sleep75` возьмите переменные `sleep`, `totwrk`, `age`, `educ` и проведите разбиение на кластеры следующими методами:

| Число кластеров | Метод         |
|-----------------|---------------|
| 3               | k-средних     |
| 4               | k-средних     |
| 5               | k-средних     |
| 3               | иерархическая |
| 4               | иерархическая |
| 5               | иерархическая |

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета

#4. Из набора данных `sleep75` возьмите переменные `sleep`, `totwrk`, `age`, `educ` и найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin

#5. Для набора данных `Labour` проведите разбиение на кластеры следующими методами:

| Число кластеров | Метод         |
|-----------------|---------------|
| 3               | k-средних     |
| 4               | k-средних     |
| 5               | k-средних     |
| 3               | иерархическая |
| 4               | иерархическая |
| 5               | иерархическая |

Визуализируйте разбиение на кластеры на диаграмме рассеяния в переменных датасета

#6. Для набора данных **Labour** найдите «оптимальное» число кластеров для метода

1. k-средних
2. иерархической кластеризации

относительно метрик: Silhouette, Calinski-Harabasz, Davies-Bouldin