ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекция 8

Визуализация с pandas и Matplotlib

Оспан Ә.Ғ., PhD, доцент-исследователь

КазНУ, кафедра ИИ и Big Data

2026-2027 учебный год

Цель и содержание лекции

Цель и результаты обучения

- Выбирать вид графика по вопросу и типам признаков.
- Проверять агрегацию до графического оформления.
- Создавать читаемые рисунки с подписями и ограничениями.

Содержание лекции

1. **Выбор графика**
2. **Figure и Axes**
3. **Распределения**
4. **Связи и группы**
5. **Композиция и экспорт**

Выбор графика по аналитическому вопросу

Вопрос	Подготовка	График
Сколько записей в категории?	value_counts	Столбцы
Как распределено число?	Числовая колонка	Гистограмма
Чем различаются группы?	Значения по группам	Boxplot
Как связаны два числа?	Пары одних и тех же строк	Scatter
Что меняется во времени?	Разбор и сортировка дат	Линия
Как сравнить несколько срезов?	Одинаковые правила	Сетка панелей

Figure и Axes: управляемый рисунок

Figure

- Весь рисунок и его общий размер.
- Содержит одну или несколько областей.
- fig.savefig сохраняет конкретный результат.

Axes

- Одна система координат с данными.
- ax.set задает подписи и заголовок.
- ax.legend объясняет обозначения групп.

fig, ax = plt.subplots() явно связывает рисунок и область.

Частоты позиций: сначала таблица

Пример Python

```
from pathlib import Path
import pandas as pd
import matplotlib.pyplot as plt

football = pd.read_csv(Path('data') / 'football.csv')
counts = football['Position'].value_counts().head(8)
print(counts)
fig, ax = plt.subplots(figsize=(8, 4), layout='constrained')
counts.plot.bar(ax=ax, color='steelblue')
ax.set(title='Восемь самых частых позиций',
       xlabel='Позиция', ylabel='Число записей')
plt.show()
```

Подготовка

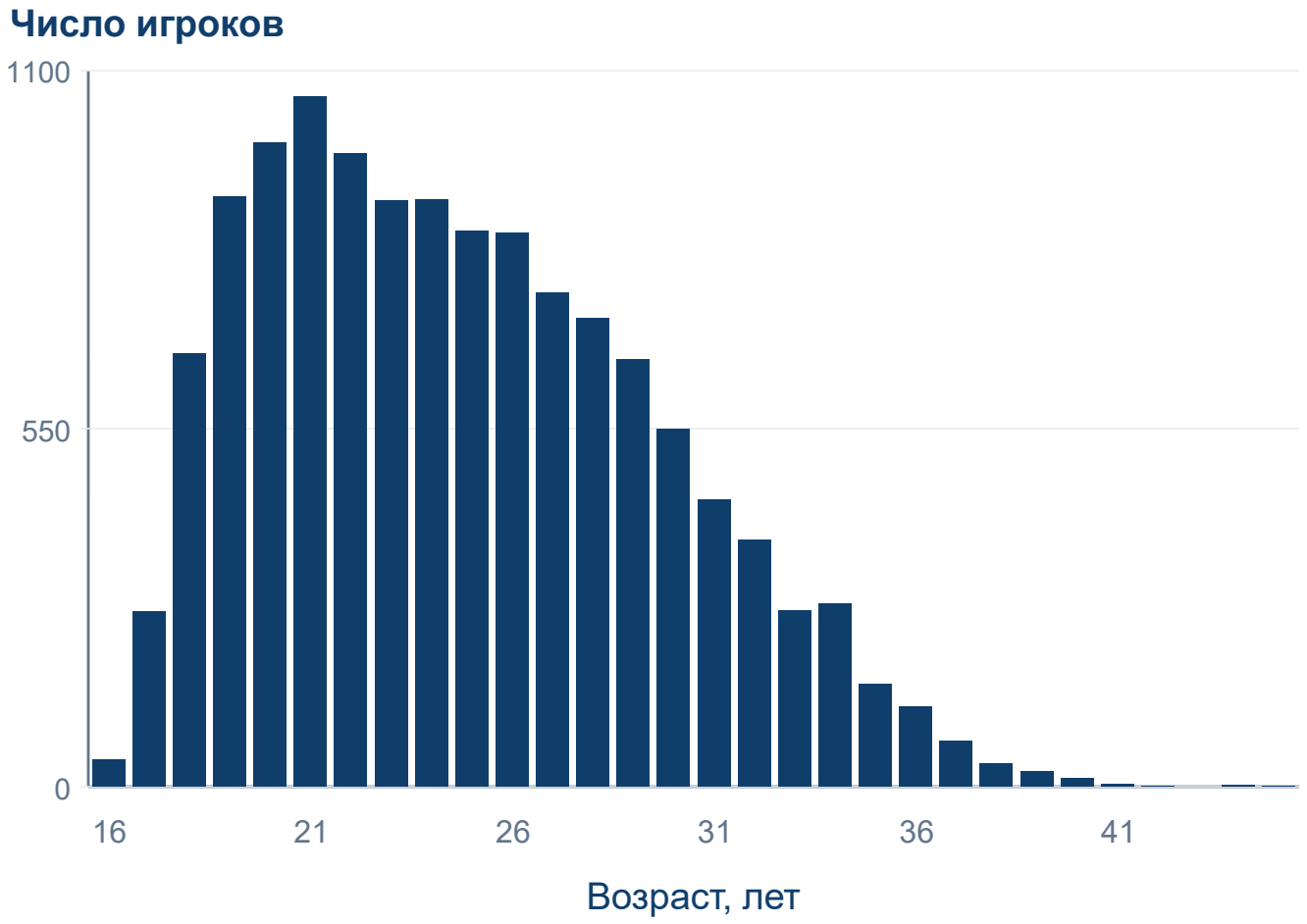
- Частоты рассчитаны до построения.
- Показаны восемь категорий, а не весь состав.

Подписи

- Указаны признак и единица счетчика.
- Таблица позволяет проверить высоту столбцов.

Точный числовой источник сопровождает графическое сравнение.

Гистограмма возраста: что показывает форма



Как читать распределение

- Каждый столбец объединяет возрастной интервал.
- Высота показывает число записей, а не причинный механизм.
- Ширина bins влияет на заметность деталей.
- Возрастной состав относится к историческому снимку.

Перед объяснением причин зафиксируйте наблюдаемую форму и проверьте ее по таблице частот.

Гистограмма, столбцы и плотность

Изображение	Ось x	Смысл высоты
Столбцы категорий	Отдельные названия	Частота или агрегат
Гистограмма counts	Числовые интервалы	Число наблюдений
Гистограмма density	Числовые интервалы	Плотность; площадь равна 1
Сравнение групп	Общие границы bins	Зависит от нормировки

Для сравнения формы учитывайте объемы и используйте одинаковые интервалы.

Границы интервалов: проверяем счетчики

Пример Python

```
import numpy as np

ages = np.array([18, 19, 21, 22, 23, 24])
edges = np.array([18, 21, 25])
hist_counts, _ = np.histogram(ages, bins=edges)
print(hist_counts.tolist(), hist_counts.sum())
fig, ax = plt.subplots(layout='constrained')
ax.hist(ages, bins=edges, edgecolor='black')
ax.set(xlabel='Возраст, лет', ylabel='Число наблюдений')
plt.show()
```

Результат: [2, 4] 6

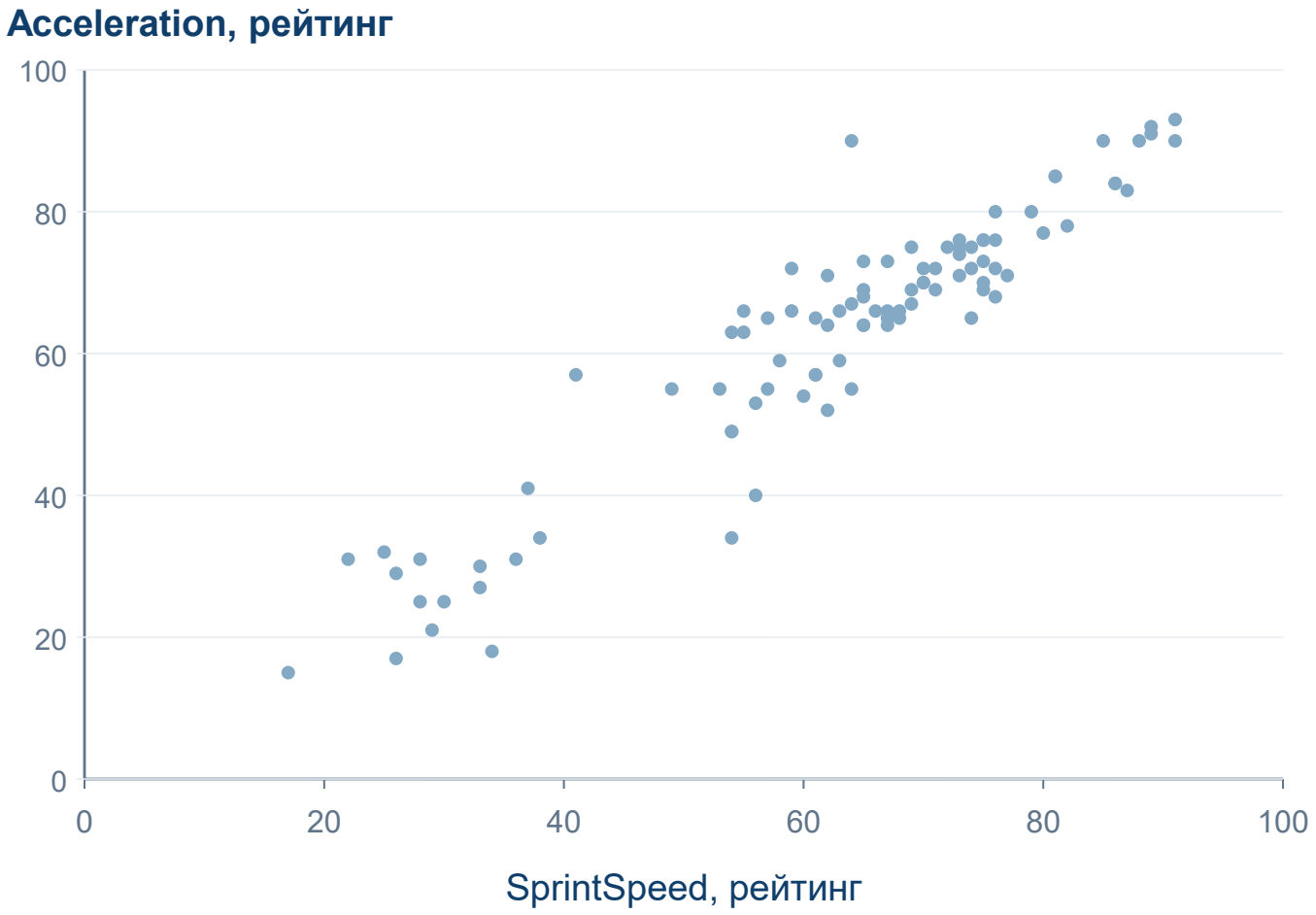
Границы

- [18, 21) содержит два значения.
- [21, 25] содержит четыре значения.

Контроль

- Сумма частот совпадает с объемом.
- Граничное число не попадает в два интервала.

Scatter: ускорение и скорость



Что означает точка

- Координаты взяты из одной и той же записи.
- Прозрачность помогает заметить перекрытие.
- Совместное изменение рейтингов не доказывает причинность.
- Позиция и система оценивания могут объяснять часть связи.

Не соединяйте игроков линией по случайному порядку строк: этот порядок не задает время.

Сравнение групп: показатель и объем

Что сравниваем

- Среднее, медиану или число записей.
- Доли требуют явного знаменателя.
- Одинаковые единицы нужны для общей шкалы.

Что сообщаем рядом

- Размер каждой группы n.
- Показатель разброса и правило отбора.
- Охват полного набора выбранными категориями.

Агрегат без объема группы легко переинтерпретировать.

Минимум и максимум: две серии столбцов

Пример Python

```
summary = football.groupby('Position')['Stamina'].agg(
    ['size', 'min', 'max']
)
shown = summary.loc[summary['size'].nlargest(6).index]
print(shown)
fig, ax = plt.subplots(figsize=(8, 4), layout='constrained')
shown[['min', 'max']].plot.bar(ax=ax)
ax.set(xlabel='Позиция', ylabel='Stamina, рейтинг',
       title='Минимум и максимум: шесть частых позиций')
ax.legend(['Минимум', 'Максимум'])
plt.show()
```

Таблица

- Размер и экстремумы рассчитаны явно.
- Выбраны шесть самых частых позиций.

Ограничение

- Размах не описывает типичный уровень.
- Для центра нужны медиана и квартили.

Boxplot: центр, ящик и необычные точки

Медиана и ящик

- Линия: медиана.
- Края ящика: Q1 и Q3; IQR = Q3 - Q1.

Усы

- До наблюдаемых значений внутри границ.
- Обычные границы: Q1/Q3 плюс 1.5 IQR.

Отдельные точки

- Расположены за усами.
- Не являются автоматически ошибками.

Boxplot дополняет распределение, но не заменяет его полностью.

Сетка панелей и вставная область

Сетка subplots

- 2x3 означает шесть областей.
- axes[0, 1] выбирает верхнюю среднюю панель.
- Общие шкалы облегчают прямое сравнение.

Вставка add_axes

- [left, bottom, width, height].
- Все координаты - доли размера Figure.
- Не закрывает основной рисунок и подписи.

Композиция должна уточнять сравнение, а не украшать пустое место.

Экспорт: проверить рисунок в размере отчета

1

Сохранить объект Figure

Задать разумный размер и dpi; использовать fig.savefig.

2

Открыть результат

Проверить подписи, легенду, обрезку и размер текста.

3

Сверить числа

Сопоставить столбцы и агрегаты с исходной таблицей.

4

Добавить пояснение

Указать вопрос, срез, единицы и границы интерпретации.

- Данные для рисунков сохраняются отдельно от изображений.
- Локальные результаты записываются в work_results/.

Ошибки визуализации и их последствия

Ошибка	Искажение	Исправление
Линия по индексу игроков	Мнимая последовательность	Scatter или распределение
Обрезанный ноль столбцов	Преувеличение различий	Отсчет длины от нуля
Разные bins у групп	Несопоставимая форма	Общие границы
Среднее без n	Скрытый объем данных	Показать размеры групп
Частичный срез без подписи	Иллюзия полного охвата	Указать правило и долю
Нечитаемый экспорт	Недоступный результат	Просмотр в размере отчета

Основные результаты лекции

1. Вопрос определяет агрегацию и тип графика.
2. Figure содержит рисунок; Axes задает конкретную систему координат.
3. Гистограмма требует осмысленных интервалов и нормировки.
4. Сравнение групп сопровождается объемами и разбросом.
5. Экспорт проверяют визуально и по числовой таблице.

Контрольные вопросы и самостоятельная работа

1. Чем гистограмма отличается от столбцов категорий?
2. Почему номер строки футболиста не является временем?
3. Что означает высота при density=True?
4. Почему точка за усами boxplot не доказывает ошибку?
5. Какие сведения нужны рядом со средними групп?

Самостоятельная работа

Постройте три гистограммы Age с разными bins. Назовите устойчивые свойства и детали, зависящие от оформления; проверьте сумму частот.

Источники и документация

Документация

1. pandas: визуализация

https://pandas.pydata.org/docs/user_guide/visualization.html

2. Matplotlib: краткое руководство

https://matplotlib.org/stable/users/explain/quick_start.html

3. Matplotlib: boxplot

https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.boxplot.html

Основа лекции

- Лекция 8 Визуализация данных [Autosaved].pptx
- Лекция 10 Визуализация данных.pptx
- Лабораторная работа 8.docx
- Лабораторная работа 8 с ответами.docx