ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекция 9

Разведочный анализ данных: EDA

Оспан Ә.Ғ., PhD, доцент-исследователь

КазНУ, кафедра ИИ и Big Data

2026-2027 учебный год

Цель и содержание лекции

Цель и результаты обучения

- Проводить EDA от паспорта до новых гипотез.
- Различать редкость, ошибку и статистический выброс.
- Описывать связи с учетом групп и ограничений источника.

Содержание лекции

1. **Вопрос и паспорт**
2. **Качество и категории**
3. **Распределения**
4. **Связи и срезы**
5. **Выводы и гипотезы**

EDA: вопрос, описание, проверка устойчивости

1

Сформулировать вопрос

Уточнить объект анализа, признаки и единицу наблюдения.

2

Проверить качество

Изучить схему, пропуски, категории и повторения.

3

Описать распределения

Рассчитать центр, разброс, частоты и показать форму.

4

Исследовать связи

Сравнить общую картину с содержательными срезами.

5

Проверить устойчивость

Уточнить наблюдение и сформулировать следующие гипотезы.

- EDA допускает возврат к предыдущим проверкам.
- Изменения правил и причины фиксируются в журнале.

Словарь признаков: тип определяется смыслом

Поле	Семантический тип	Ограничение
Age	Количественный возраст	Поперечный срез, не траектория
Position	Номинальная категория	Нет числового расстояния
Stamina	Рейтинг навыка	Не физическая единица измерения
Name	Имя записи	Не гарантирует уникальный ID
Unnamed: 0	Служебный индекс	Исключается из признаков
Penalties	Рейтинг навыка	Не количество забитых голов

Для Wage/Value неизвестны единицы и период; это отмечаем явно.

Паспорт и первое числовое описание

Пример Python

```
from pathlib import Path
import numpy as np
import pandas as pd

football = pd.read_csv(Path('data') / 'football.csv')
features = ['Age', 'Acceleration', 'SprintSpeed', 'Stamina']
print(football.shape)
print(football.isna().sum().sum())
print(football[features].describe())
```

Результат: Форма: (12897, 42); число пропусков: 0.

Паспорт

- 12897 строк, 42 колонки.
- В комплектном CSV нет пропущенных ячеек.

Описание

- Центр, квартили и диапазон признаков.
- Крайнее значение еще не доказанная ошибка.

Качество: счетчики до содержательных выводов

Проверяем данные

- Пропуски и условные обозначения.
- Точные повторы по выбранным полям.
- Частоты и варианты написания категорий.

Не смешиваем ответы

- Повтор имени не равен повтору сущности.
- Число категорий не равно числу строк.
- Редкая группа не обязательно ошибочна.

Порог частоты - правило отображения, а не закон существования категории.

Категории: прозрачная нормализация

Пример Python

```
category_demo = pd.Series(
    [' Alpha ', 'alpha', 'ALPHA', 'Beta', '', None],
    dtype='string'
)
category_key = category_demo.str.strip().str.casefold()
category_key = category_key.replace({'': pd.NA})
print(category_key.value_counts(dropna=False))
assert len(category_key) == len(category_demo)
```

Что изменилось

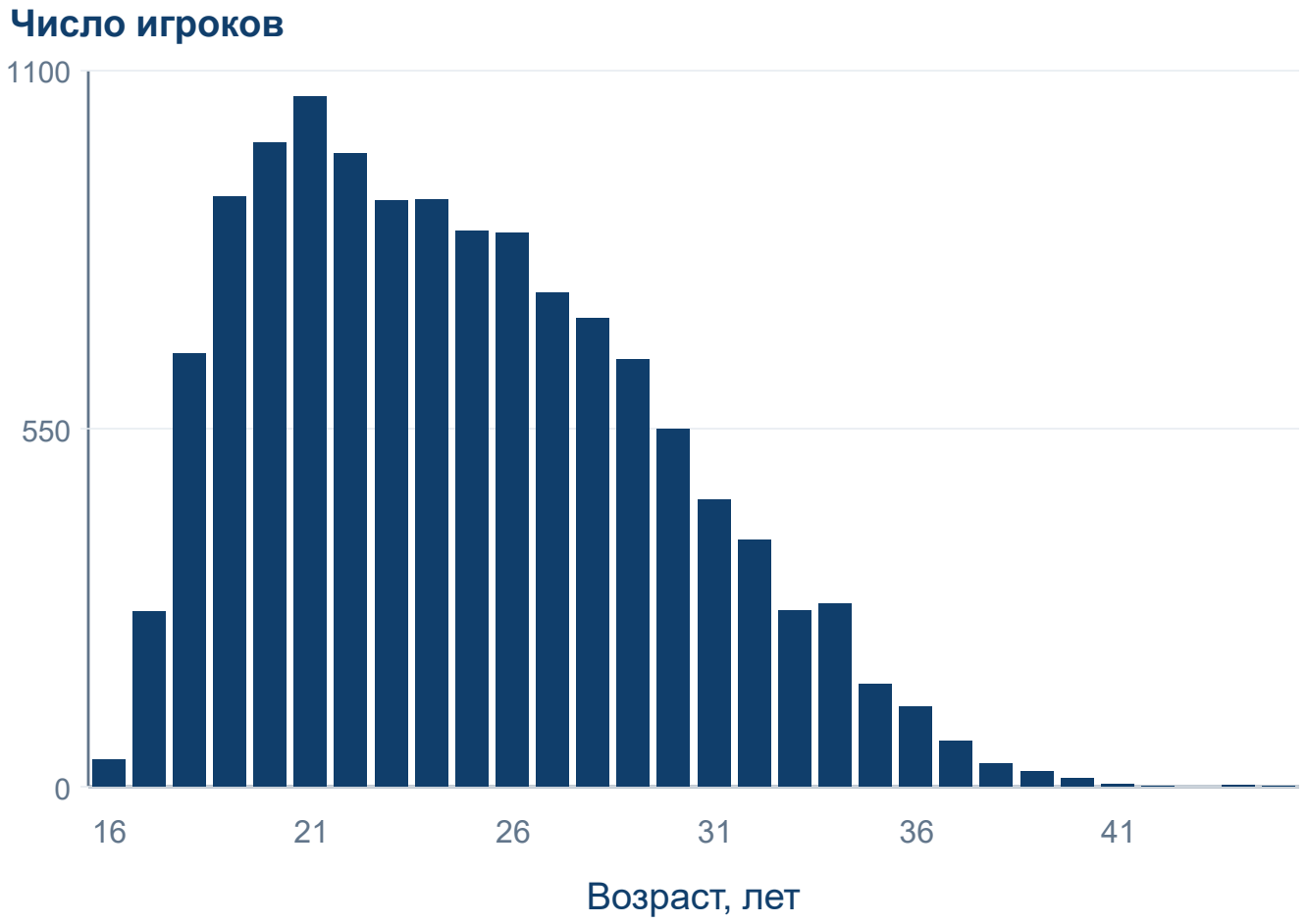
- alpha: 3 записи; beta: 1 запись.
- Пустая строка и None дают два пропуска.

Что сохраняется

- Число строк не изменяется.
- Исходное написание доступно для проверки.

Словарь подтвержденных соответствий безопаснее случайных замен.

Возраст: распределение и состав выборки



Вопросы к распределению

- Где расположен центр и насколько широк разброс?
- Сохраняется ли описание при других bins?
- Как возрастной состав различается между позициями?
- Какие выводы требуют сведений о способе отбора?

Распределение разных игроков не является траекторией одного игрока и требует проверки состава.

IQR: флаг необычности, а не приговор

IQR = Q3 - Q1; границы: Q1 - 1.5*IQR, Q3 + 1.5*IQR

Статистическое правило

- Зависит от квартилей текущего набора.
- Выделяет необычные относительно большинства значения.

Предметное решение

- Требует смысла признака и контекста.
- Редкое реальное значение сохраняют.

Для [1, 2, 3, 4, 100]: Q1=2; Q3=4; границы [-1, 7].

Флаг хранится отдельно; удаление требует дополнительного основания.

Рабочий пример: влияние крайнего значения

Пример Python

```
toy = pd.Series([1., 2., 3., 4., 100.])
q1, q3 = toy.quantile([0.25, 0.75])
iqr = q3 - q1
lower, upper = q1 - 1.5 * iqr, q3 + 1.5 * iqr
flag = ~toy.between(lower, upper)
print(lower, upper)
print(toy.mean(), toy.median())
print(toy.loc[flag].tolist())
```

Результат: -1.0 7.0; 22.0 3.0; [100.0]

Расчет

- Границы: -1 и 7; кандидат: 100.
- Среднее 22, медиана 3.

Интерпретация

- Крайность заметно сдвигает среднее.
- Решение об ошибке из этого не следует.

Возраст и скорость: связь требует контекста



Читаем связь осторожно

- Пары взяты из одного исторического поперечного среза.
- Возрастные группы могут различаться составом позиций.
- Корреляция описывает связь, но не индивидуальное изменение.
- Для проверки объяснения нужны сопоставимые срезы или новые данные.

От описания общей картины переходите к проверке состава групп.

Корреляции и сводка по позициям

Пример Python

```
correlation = football[features].corr(method='pearson')
print(correlation)
top = football['Position'].value_counts().head(4).index
subset = football.loc[football['Position'].isin(top)]
groups = subset.groupby('Position').agg(
    n=('Stamina', 'size'),
    mean_age=('Age', 'mean'),
    median_stamina=('Stamina', 'median')
)
print(groups)
```

Числовая связь

- Пирсон описывает линейное совместное изменение.
- Набор признаков выбирается по смыслу.

Групповой контекст

- Сводка показывает n и состав возрастов.
- Общая связь требует сравнения внутри групп.

Смешение: общий набор и отдельные группы

Общая связь

- Объединяет разные позиции и составы.
- Может отражать различия между группами.
- Не устанавливает причину.

Проверка срезов

- Сравниваем направление и величину связи.
- Сообщаем размеры и правила включения.
- Формулируем альтернативные объяснения.

Срез уточняет наблюдение, но сам по себе не доказывает причинный механизм.

Как оформить проверяемое наблюдение

1

Наблюдение

Указать конкретную величину, группу или форму распределения.

2

Доказательство

Привести таблицу, рисунок, объем и правило расчета.

3

Ограничение

Назвать, чего эти данные и процедура не устанавливают.

4

Следующий вопрос

Предложить срез, новые данные или план отдельной проверки.

- Проверка устойчивости должна быть описана прозрачно.
- Не выбираем только те варианты, которые поддерживают ожидание.

Граница EDA: гипотеза и подтверждение

Разведка

- Замечает структуру и создает гипотезы.
- Допускает широкий поиск и возвраты.
- Показывает ограничения и неопределенность.

Подтверждение

- Требует заранее определенного вопроса.
- Учитывает предпосылки и множественный поиск.
- Может потребовать независимые данные.

Не называйте найденное после поиска различие доказанным эффектом без отдельной проверки предпосылок.

Основные результаты лекции

1. EDA связывает вопрос, качество, распределения и срезы.
2. Тип признака определяется смыслом, а не только dtype.
3. IQR помечает необычность; решение об ошибке требует контекста.
4. Общая связь может отражать неоднородный состав групп.
5. Наблюдение сопровождают доказательством, ограничением и следующим вопросом.

Контрольные вопросы и самостоятельная работа

1. Чем результат EDA отличается от подтвержденной гипотезы?
2. Почему почти уникальный столбец не всегда полезен для модели?
3. Чем предметная ошибка отличается от IQR-кандидата?
4. Почему связь внутри позиции может отличаться от общей?
5. Как проверить устойчивость наблюдения к аналитическому решению?

Самостоятельная работа

Выберите одно наблюдение по football. Предложите два объяснения и для каждого назовите срез или новые данные, которые помогут их различить.

Источники и документация

Документация

1. pandas: описательные расчеты

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

2. pandas: группировка

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

3. NIST: что такое EDA

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Основа лекции

- Лекция 9 EDA.pptx
- Лабораторная работа 9 .docx