Exploratory Data Analysis Report
FIFA Players Dataset Analysis

Дата отчета: 2025 год

1. Введение
Данный отчёт представляет собой разведочный анализ данных о футболистах FIFA из датасета fifa_players.csv.

Основные цели анализа:
- Изучение структуры данных и качества данных
- Очистка и предобработка данных
- Анализ распределений признаков
- Выявление аномалий и пропущенных значений
- Подготовка данных для дальнейшего моделирования

2. Описание данных

2.1 Размерность данных
- 17,954 записей (футболистов)
- 51 признак в исходном наборе данных

2.2 Основные группы признаков

Демографические данные:
- name, full_name - имя и полное имя игрока
- age, height_cm, weight_kgs - возраст, рост, вес
- nationality, birth_date - национальность и дата рождения

Игровые характеристики:
- positions - позиции на поле
- overall_rating, potential - текущий и потенциальный рейтинг
- preferred_foot - предпочтительная нога
- international_reputation - международная репутация (1-5)
- weak_foot, skill_moves - слабая нога и навыки финтов (1-5)
- body_type - тип телосложения

Финансовые показатели:

- value_euro - рыночная стоимость в евро
- wage_euro - зарплата в евро
- release_clause_euro - сумма отступных

Технические и физические навыки:
- crossing, finishing, dribbling - навесы, завершение атак, дриблинг
- ball_control - контроль мяча
- acceleration, sprint_speed - ускорение и скорость
- stamina, strength - выносливость и сила
- И другие специализированные навыки

3. Предобработка данных

3.1 Удаление избыточных признаков
Были удалены следующие признаки:
- name - дублирует информацию из full_name
- birth_date - сильно коррелирует с возрастом (age)

3.2 Анализ пропущенных значений
Колонки с критическим количеством пропусков (>95%):
- national_jersey_number: 17,097 пропусков (95.2%)
- national_team_position: 17,097 пропусков (95.2%)
- national_rating: 17,097 пропусков (95.2%)
- national_team: 17,097 пропусков (95.2%)

Колонки с умеренным количеством пропусков:
- release_clause_euro: 1,837 пропусков (10.2%)

Колонки с незначительными пропусками:
- value_euro: 255 пропусков (1.4%)
- wage_euro: 246 пропусков (1.4%)

3.3 Очистка данных
Принятые решения:
1. Полное удаление колонок с >95% пропусков
2. Сохранение колонок с меньшим процентом пропусков

4. Статистический анализ

4.1 Описательная статистика числовых признаков
Возраст игроков (age):
- Средний возраст: 25.6 лет
- Стандартное отклонение: 4.7 лет
- Диапазон: от 17 до 46 лет
- Медиана: 25 лет

Физические параметры:
- Средний рост: 174.9 см
- Средний вес: 75.3 кг

Игровые характеристики:
- Средний общий рейтинг: 66.2
- Диапазон рейтинга: 47-94
- Средний потенциал: 71.4

Финансовые показатели:
- Средняя стоимость: 2.48 млн €
- Медианная стоимость: 0.70 млн €
- Максимальная стоимость: 110.5 млн €
- Средняя зарплата: 9,902 € в неделю

4.2 Распределение ключевых показателей
- Рейтинг: Нормальное распределение с легким смещением вправо
- Возраст: Распределение близко к нормальному
- Стоимость: Сильно правостороннее распределение

5. Анализ категориальных переменных
5.1 Национальность (nationality)
- 160 уникальных национальностей
- Наиболее представленные: England, Germany, Spain, France, Argentina

5.2 Позиции (positions)
- 890 уникальных комбинаций позиций
- Наиболее частые: CB, ST, CM

5.3 Предпочтительная нога (preferred_foot)
- Правша: 13,781 игроков (76.8%)
- Левша: 4,173 игрока (23.2%)

5.4 Тип телосложения (body_type)
- 10 уникальных типов
- Наиболее частые: Normal, Lean, Stocky

6. Качество данных и целостность
6.1 Полнота данных
- Большинство ключевых признаков заполнены
- Технические навыки не содержат пропусков

6.2 Непротиворечивость
- Проверка на дубликаты: 0 дублированных записей

6.3 Проблемные области
- Высокий процент пропусков в данных о национальных сборных

7. Выводы и рекомендации
7.1 Основные выводы
- Данные в хорошем состоянии для большинства признаков
- Минимальное количество пропусков
- Отсутствие дубликатов

7.2 Рекомендации для дальнейшего анализа
- Импутация пропусков
- Feature Engineering
- Обработка выбросов
- Корреляционный анализ
- Кластеризация
- Моделирование
- Визуализация

8. Заключение
Данные готовы для применения современных методов анализа, включая машинное обучение и глубокий статистический анализ.