

Исследовательский проект

Для НИСа «Анализ данных в Python»

Выполнил: Шикалов Андрей Артемович, БПИ233

Выбор датасета

- В качестве данных для исследования выбран датасет, содержащий информацию о людях, посещающих спортзал.
- Источник: [kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset](https://www.kaggle.com/valakhorasani/gym-members-exercise-dataset)
- Всего наблюдений: 973. Пустых полей: 0
- Переменные:
 - Пол
 - Возраст
 - Рост
 - Вес
 - Частота тренировок
 - Тип тренировок (силовая, кардио, йога, кроссфит)
 - Средняя длительность тренировок
 - Частота сердцебиений во время тренировки (средняя и максимальная)
 - Количество сжигаемых калорий
 - Индекс массы тела
 - Процент подкожного жира

Задачи исследования

- Влияние пола и возраста на количество сжигаемых калорий, индекс массы тела, процент подкожного жира
- Влияние типа тренировок, их длительности и частоты на те же показатели
- Модель определения количества сжигаемых калорий на основе информации о тренировках
- Определение основных тенденций
- Модель определение уровня опытности спортсмена
- Выявление ключевых групп людей, посещающих спортзал

Сжигание калорий

- Построена сводная таблица среднего количества сжигаемых калорий в зависимости от пола, типа тренировок и их частоты
- Мужчины, как правило, сжигают немного больше калорий при прочих равных показателях
- Частота тренировок влияет намного больше, чем их тип

Gender	Female				Male			
Workout_Frequency	2	3	4	5	2	3	4	5
Workout_Type								
Cardio	653.468750	776.244444	947.055556	1172.923077	728.612903	865.361702	1069.975610	1323.900000
HIIT	777.500000	783.400000	951.083333	1199.750000	734.250000	884.585366	1077.903226	1345.166667
Strength	758.136364	773.666667	950.552632	1158.666667	819.521739	832.215686	1015.617021	1412.357143
Yoga	532.133333	828.783784	936.225000	1273.785714	764.750000	835.294118	1036.675676	1269.538462

Процент подкожного жира

- Всё то же самое можно сказать и про процент подкожного жира
- Больше всего влияет частота тренировок
- У женщин показатель в среднем немного больше

Gender	Female				Male			
Workout_Frequency	2	3	4	5	2	3	4	5
Workout_Type								
Cardio	30.150000	30.300000	26.683333	17.800000	24.680645	25.225532	21.521951	12.300000
HIIT	30.272222	29.906667	25.955556	15.825000	24.629167	25.346341	19.325806	12.488889
Strength	30.113636	30.843137	26.844737	17.166667	25.630435	25.305882	22.472340	12.185714
Yoga	31.053333	29.767568	25.632500	17.950000	25.684375	24.562745	20.864865	12.400000

Отличия у мужчин и женщин

- Выдвинута гипотеза: существуют значимые отличия между показателями мужчин и женщин
- Произведен тест хи-квадрат на взаимосвязь между количеством сжигаемых калорий и полом
- $p = 0.313$
- $X = 636.2$
- Существенной взаимосвязи не выявлено

Центральные тенденции

- Средний индекс массы тела – 24.9, медианный – 24.1
- Среднее количество сжигаемых калорий – 905, медианное – 893
- Среднее количество тренировок в неделю – 3
- Средняя частота сердцебиения за тренировку – 143
- Средний процент подкожного жира – 24.9, медианный – 26.2
- Средняя продолжительность тренировки (в часах) – 1.25, медианная – 1.26

Выводы:

- Интенсивность и продолжительность тренировок предположительно распределены симметрично
- Результаты тренировок (процент жира и сжигание калорий) предположительно асимметричны

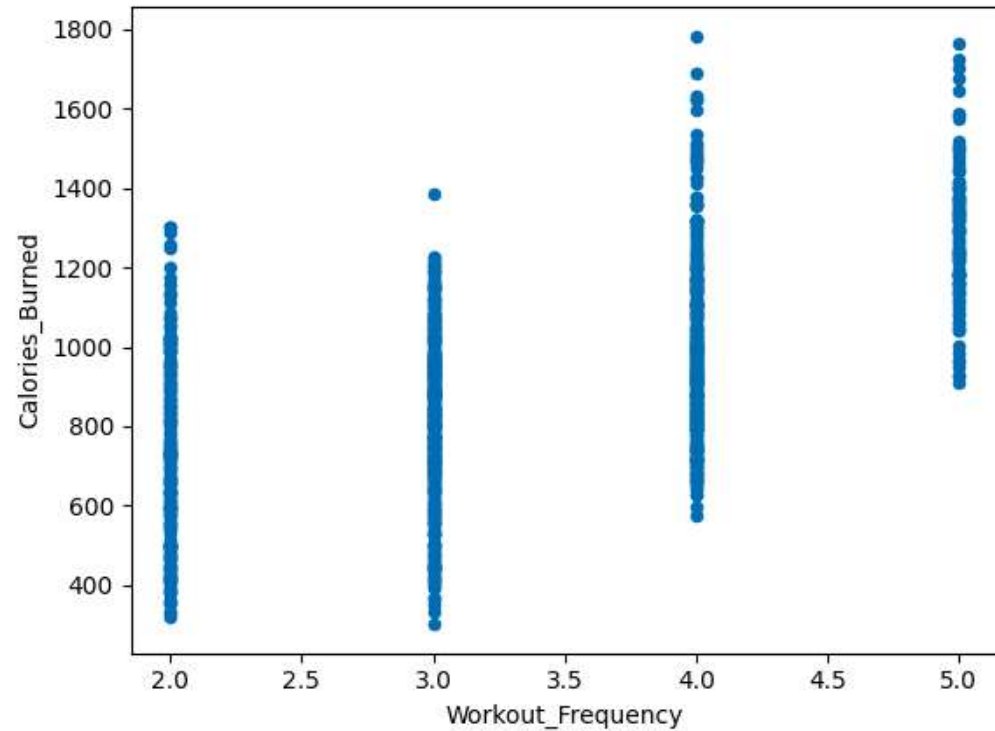
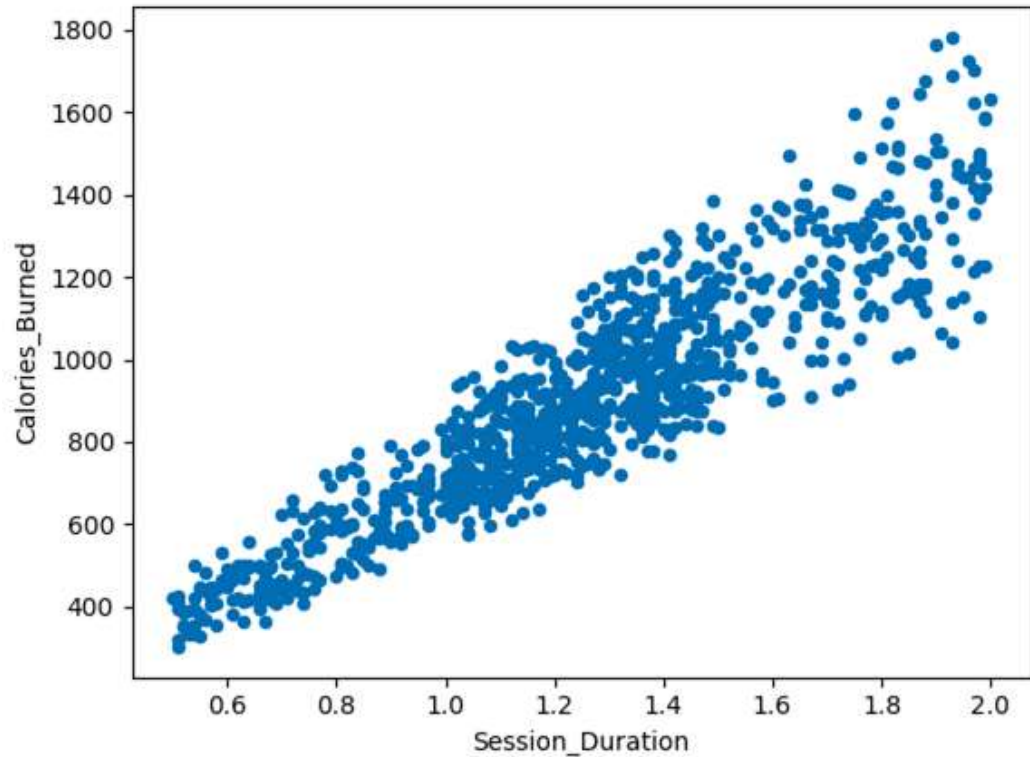
Вариация

- Показатели асимметрии:
 - Индекс массы тела: 0.76
 - Процент жира: -0.63
- Как и ожидалось, в распределении процента жира имеется левосторонняя асимметрия, а в распределении индекса массы тела – правосторонняя. То есть людей с лучшими показателями меньше, чем людей с худшими.
- Показатели вариации:
 - Количество сожженных калорий: 74333
- Наблюдается большой разброс в количестве сжигаемых калорий. То есть разница между профессионалами и новичками велика по сравнению со средним показателем.

Корреляции

- Выявлены существенные взаимосвязи между:
 - Длительностью тренировки и сжиганием калорий
 - Опытом и сжиганием калорий
 - Частотой тренировок и сжиганием калорий
- Отвергнуты взаимосвязи между:
 - Средней частотой сердцебиения и сжиганием калорий
 - Потреблением воды и сжиганием калорий
 - Потреблением воды и процентом подкожного жира

Модель предсказания сожженных калорий



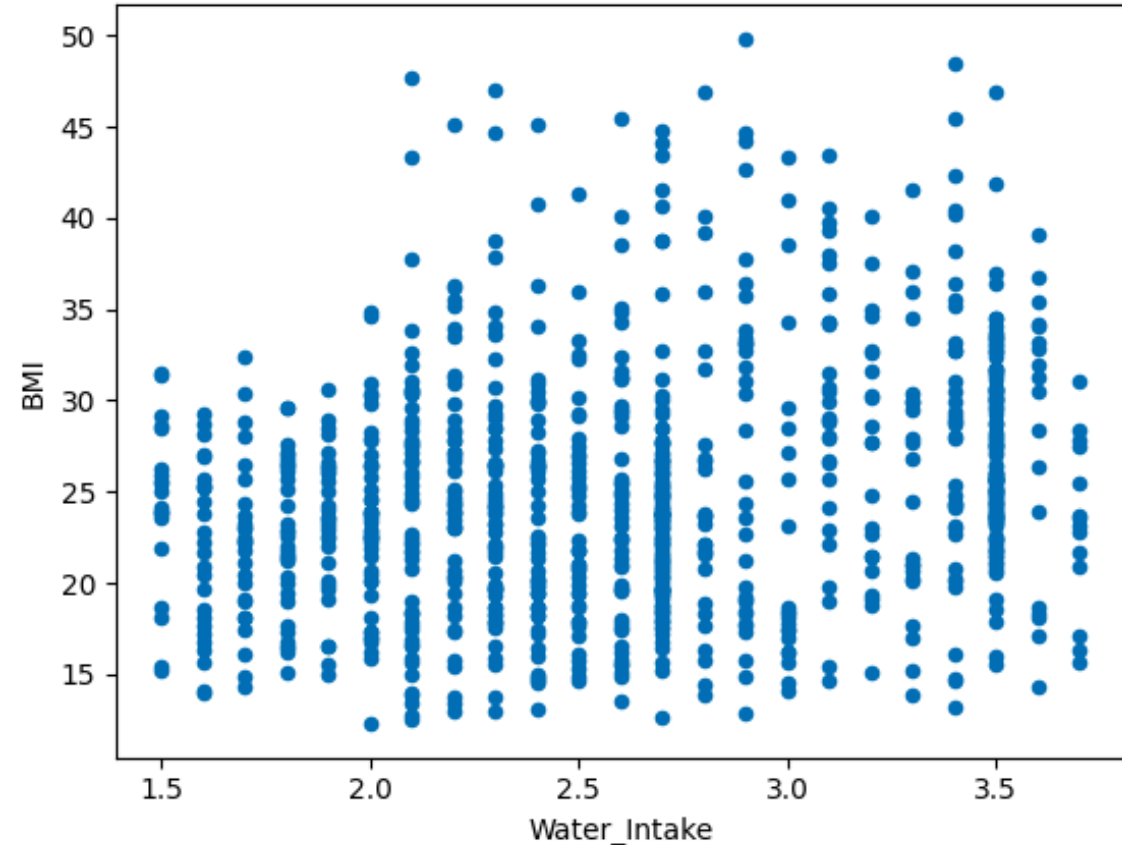
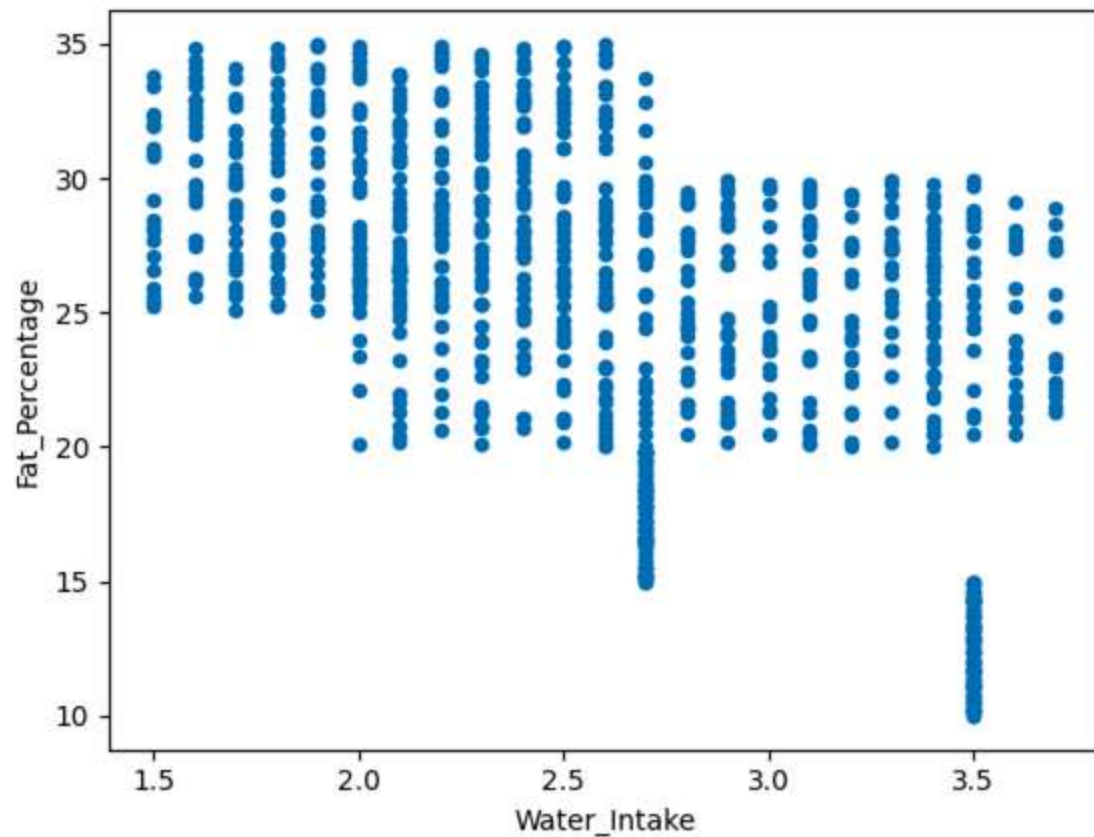
Модель предсказания сожженных калорий

- Использованы предикторы:
 - Пол
 - Потребление воды
 - Тип тренировки (силовая/несиловая)
 - Длительность тренировки
 - Средняя частота сердцебиений на тренировке
- Построена значимая модель, предсказывающая 95.6% изменений
- Тип тренировки и потребление воды как предикторы оказались не значимы

Модели предсказания индекса массы тела и процента подкожного жира

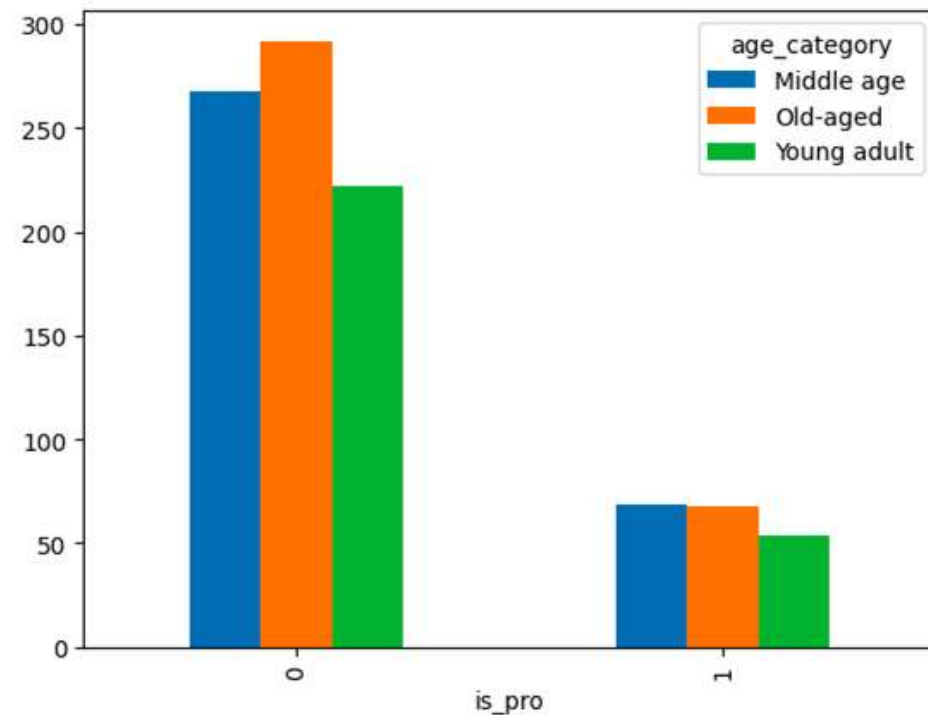
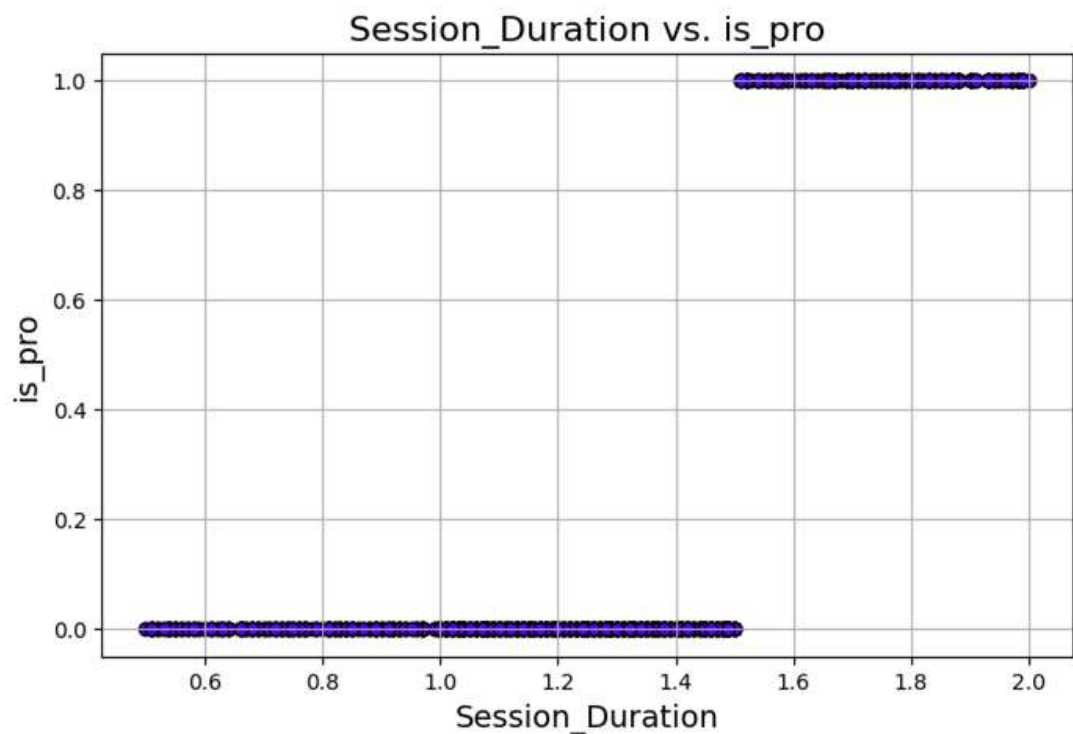
Предиктор с самым большим коэффициентом корреляции - потребление

ВОДЫ



Модели предсказания индекса массы тела и процента подкожного жира

Бинарная модель определения опыта



Бинарная модель определения опыта

- Зная возраст, пол, длительность тренировок, частоту тренировок, потребление воды и среднюю частоту сердцебиений за тренировку можно с вероятностью 100% сказать, является ли человек профессионалом (уровень опыта = 3)
- Модель прошла кросс-валидацию без ошибок. Обучение на 20% от выборки, тест на 80% от выборки.
- Наиболее значимой переменной в уравнении регрессии является длительность тренировки
- Наименее значимыми переменными оказались средняя частота сердцебиений и потребление воды

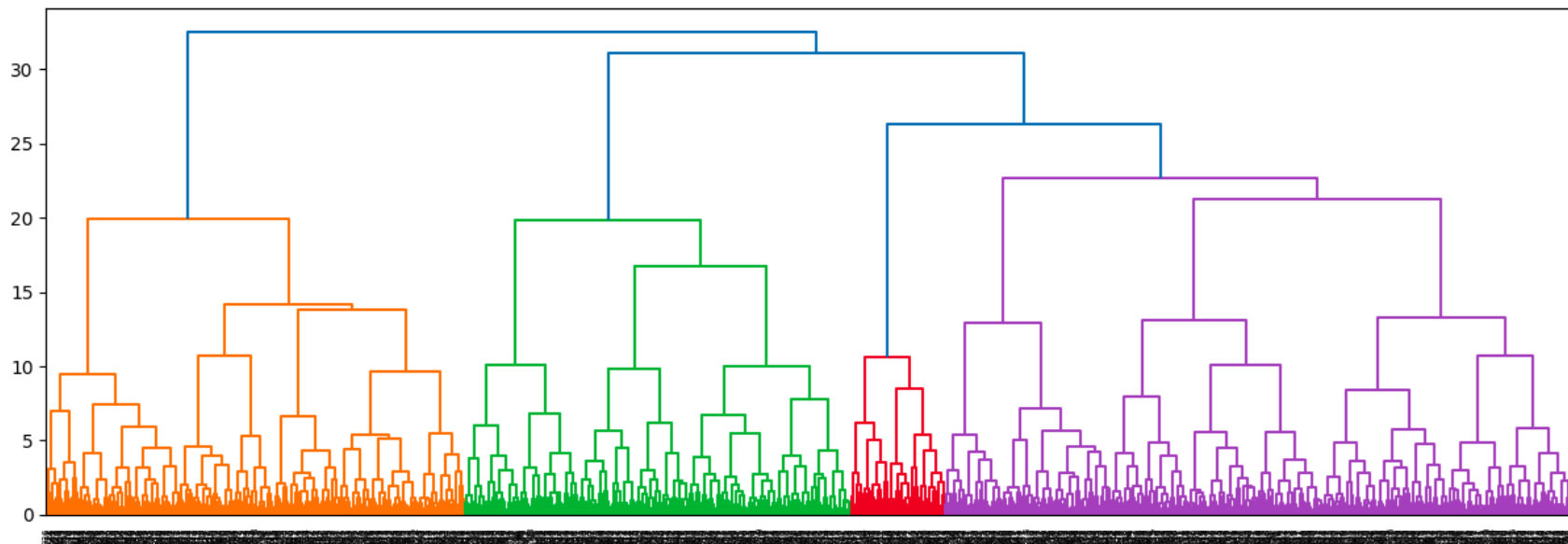
Выводы:

Характерные для профессиональных тренировок черты: их частота и длительность.

Потребление воды и интенсивность работы сердца не связаны с уровнем опыта человека.

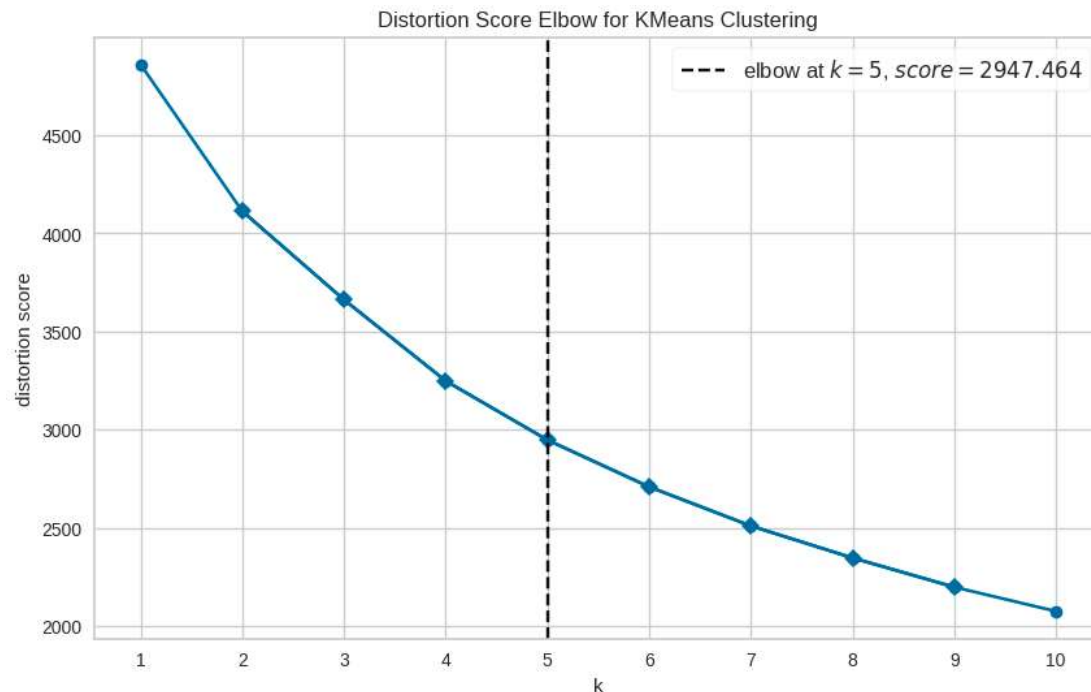
Основные группы посетителей спортзала

- Для кластеризации использовался метод Варда
- Выбраны 5 переменных для кластеризации:
 - Возраст
 - Индекс массы тела
 - Частота тренировок



Определение числа кластеров

- Для определения числа кластеров использовалось два метода
- Индекс Калински-Харабаша:
 - Оптимальное число кластеров - 3 или 4
- Метод локтя
 - Оптимальное число кластеров - 5



Характеристика кластеров

- **Кластер 1: опытные.** Это группа людей среднего возраста. Они тренируются достаточно часто - в среднем 3.74 раз в неделю и имеют нормальный индекс массы тела. Также у этих людей самый низкий пульс покоя (менее 60 ударов в минуту) - известный феномен спортивной брадикардии. Вероятно, в эту группу входят тренеры и завсегдатаи спортзала.
- **Кластер 2: начинающие.** Люди в этой группе одни из самых молодых - средний возраст 30 лет и имеют самый низкий индекс массы тела - 22.92. Среднее количество занятий в неделю также самое низкое - 2.48. Это люди, которые начали заниматься, чтобы набрать мышечную массу.
- **Кластер 3: худеющие.** Люди в этой группе имеют средний индекс массы тела 39.67, что соответствует 1-й степени ожирения. Они занимаются 3 раза в неделю и имеют самый высокий пульс во время тренировки.
- **Кластер 4: спортсмены.** Люди в этой группе самые молодые и тренируются чаще всего - в среднем 4.13 раз в неделю. У них нормальный индекс массы тела и высокая нагрузка на тренировках. Это вторая самая малочисленная группа - в неё вошло 106 человек. Это, скорее всего, люди, выступающие на соревнованиях.
- **Кластер 5: люди в возрасте, поддерживающие форму.** В эту группу вошли люди с самым высоким средним возрастом - 48.5 лет. Они тренируются 3 раза в неделю и не имеют много

Исследовательский проект

Для НИСа «Анализ данных в Python»

Выполнил: Шикалов Андрей Артемович, БПИ233