Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

**Лабораторная работа 6.1**

**Обработка данных с использованием Apache Spark**

Выполнил(а): Шведова С.С., группа: АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2024

**Цель и задачи работы:**

1. Познакомиться с понятием «большие данные» и способами их обработки;

2. Познакомиться с инструментом Apache Spark и возможностями, которые он предоставляет для обработки больших данных.

3. Получить навыки выполнения разведочного анализа данных использованием pyspark.

Для начала надо взять датасет. Для этого на Kaggle взялись данные о спортивном зале (https://www.kaggle.com/datasets/mexwell/gym-check-ins-and-user-metadata), причем было две таблицы: users\_data с данными о пользователях и checkin\_checkout\_history\_updated с данными о приходе клиентов. С помощью SQL нужно соединить эти две таблицы для более подробного анализа, это продемонстрировано на рисунке 1.

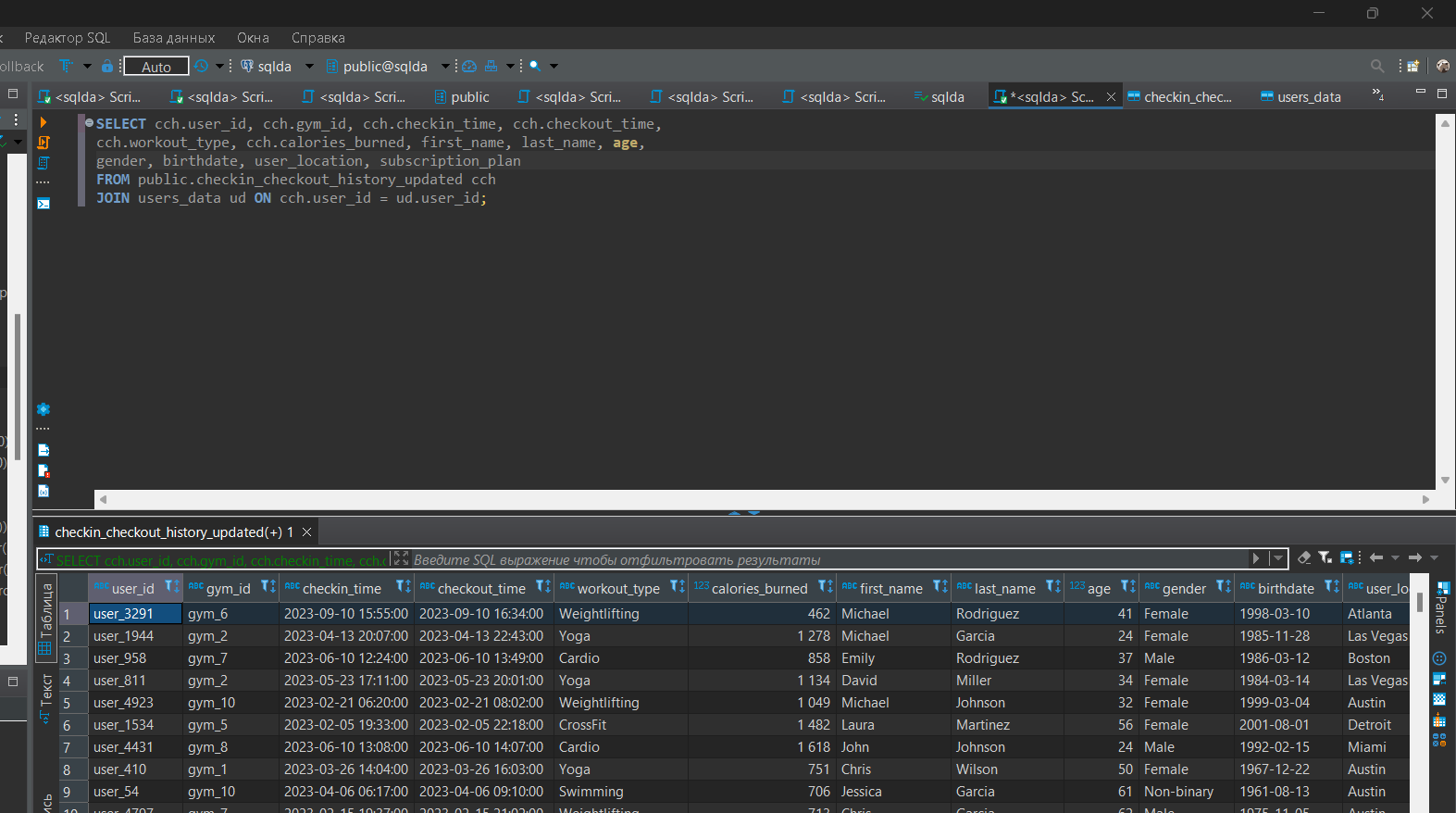


Рисунок 1. Соединение двух таблиц в единую (GYM.CSV)

Затем нужно импортировать необходимые библиотеки, чтобы работать в спарке (рисунок 2).

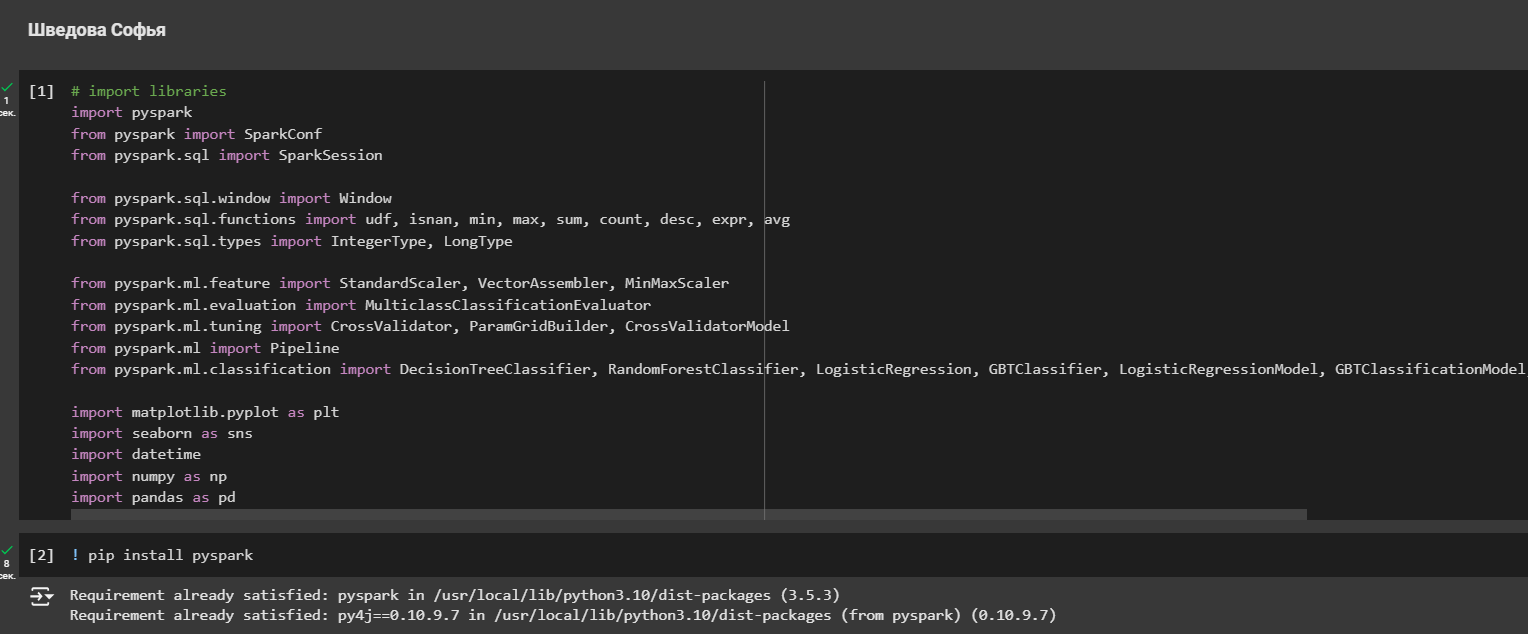


Рисунок 2. Импорт необходимых библиотек

Затем нужно импортировать модули, связанные с PySpark, инициализировать сеанс и загрузить файл (рисунок 3).

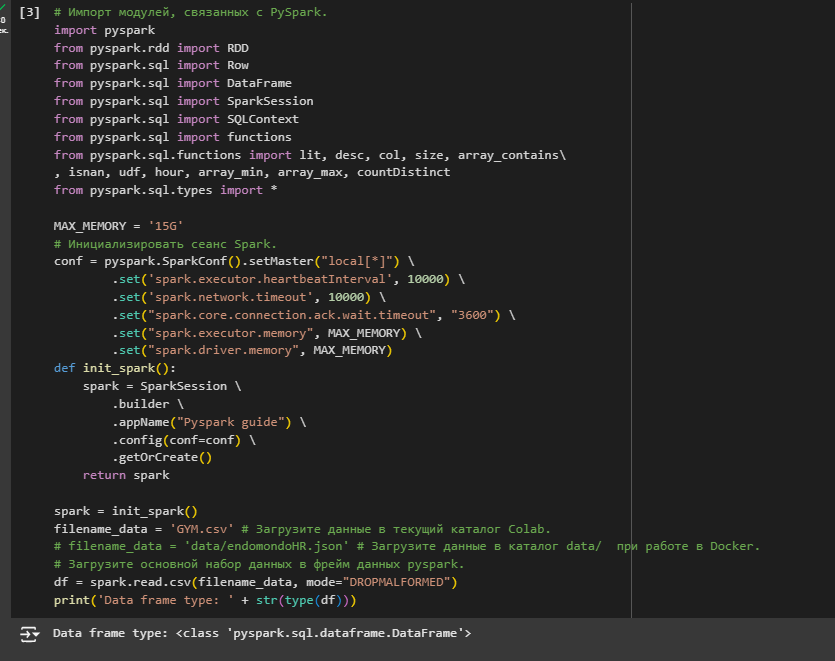


Рисунок 3. Загрузка файла в текущий каталог Colab

Далее нужно сделать обзор набора данных. Для начала нужно переименовать столбцы, так как названия сбились (рисунок 4)

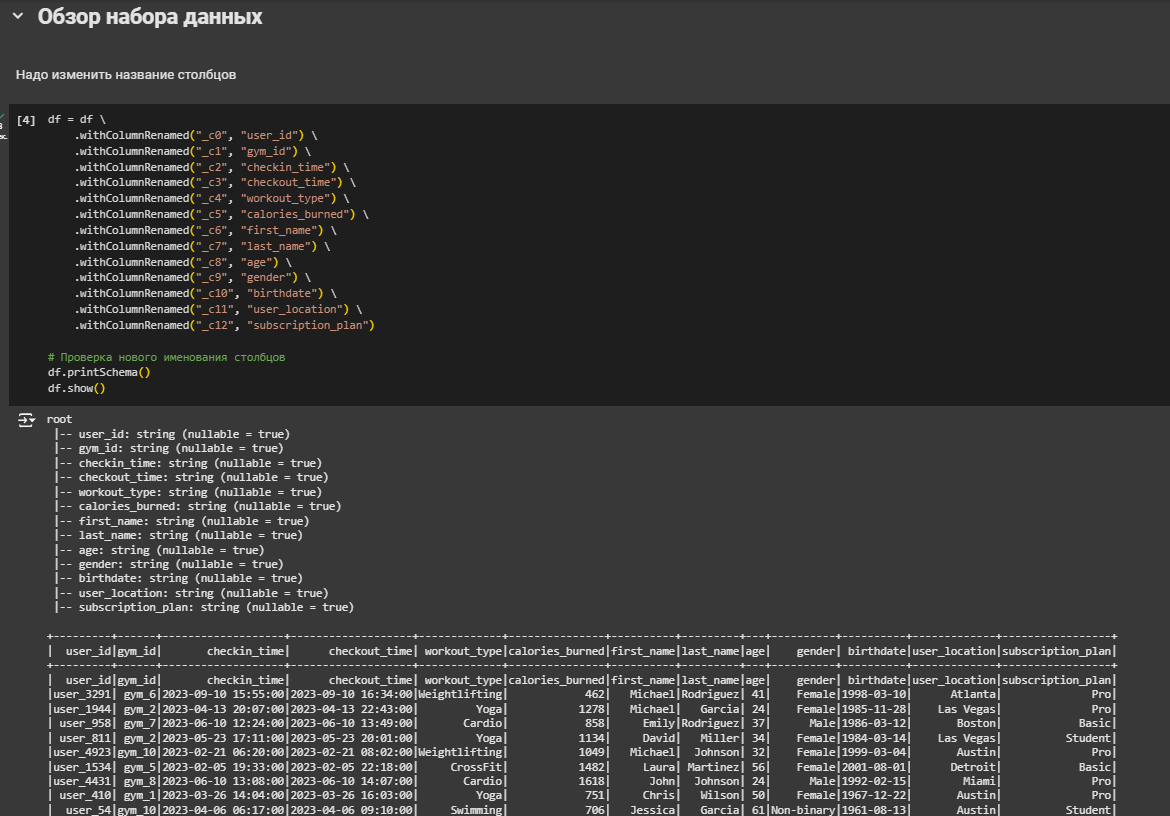


Рисунок 4. Переименование столбцов

Затем нужно удалить первую строку, которая должна была быть именами столбцов, но стала первой строкой (рисунок 5).

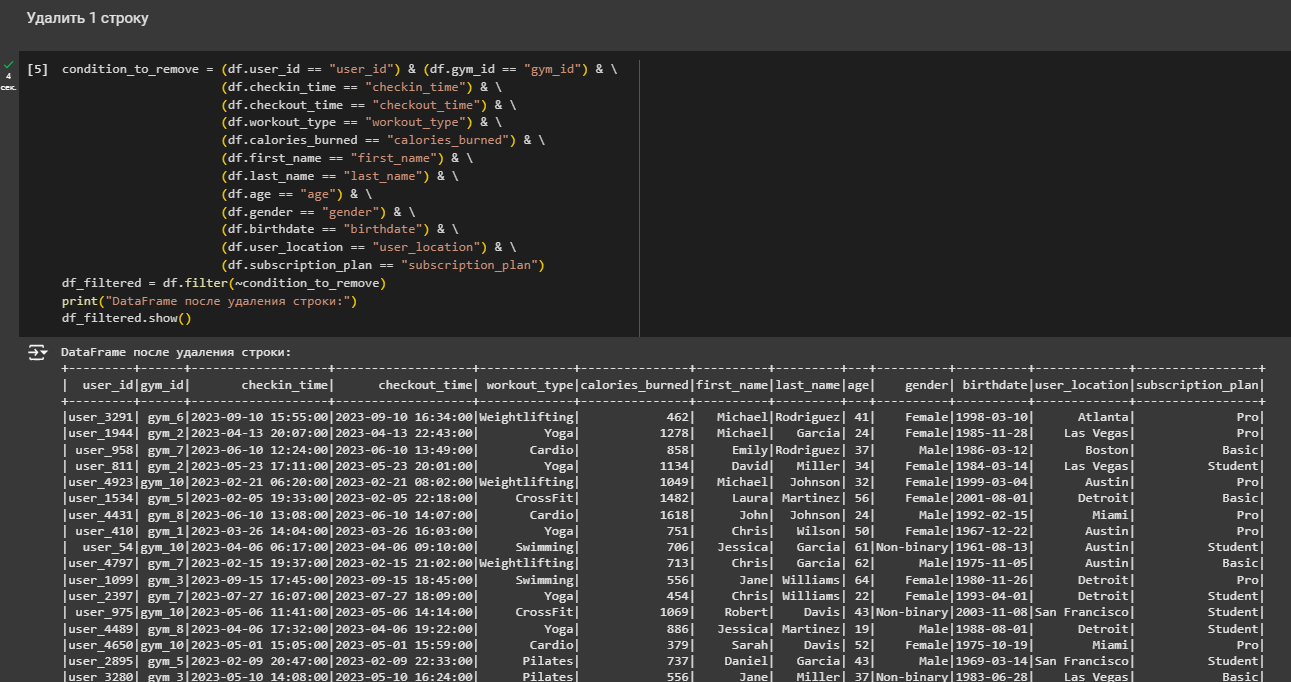


Рисунок 5. Удаление первой строки

Затем нужно посмотреть типы данных. Как можно увидеть на рисунке 6, все типы данных – это строка, поэтому нужно будет их изменить.

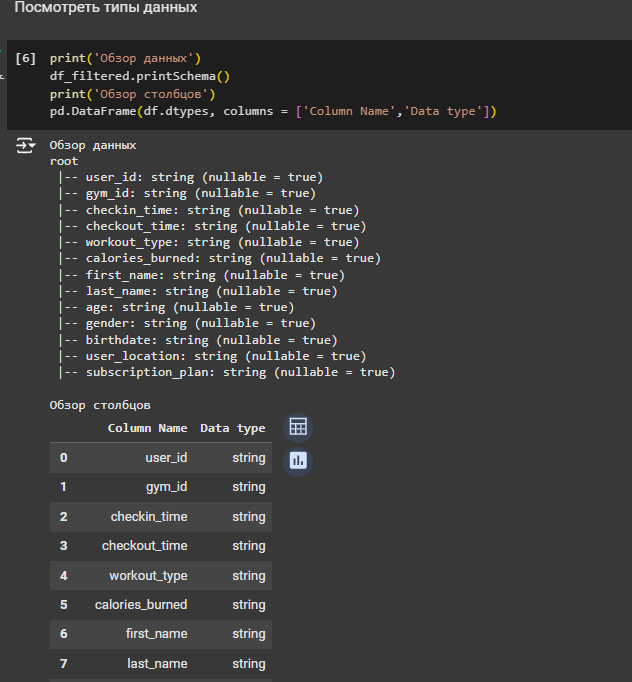


Рисунок 6. Типы данных

Теперь надо изменить типы данных, а именно возраст и калории, которые человек сжег – на целое число, а день рождения, дата прибытие и дата ухода – на время (рисунок 7). Как можно заметить, типы данных изменились.

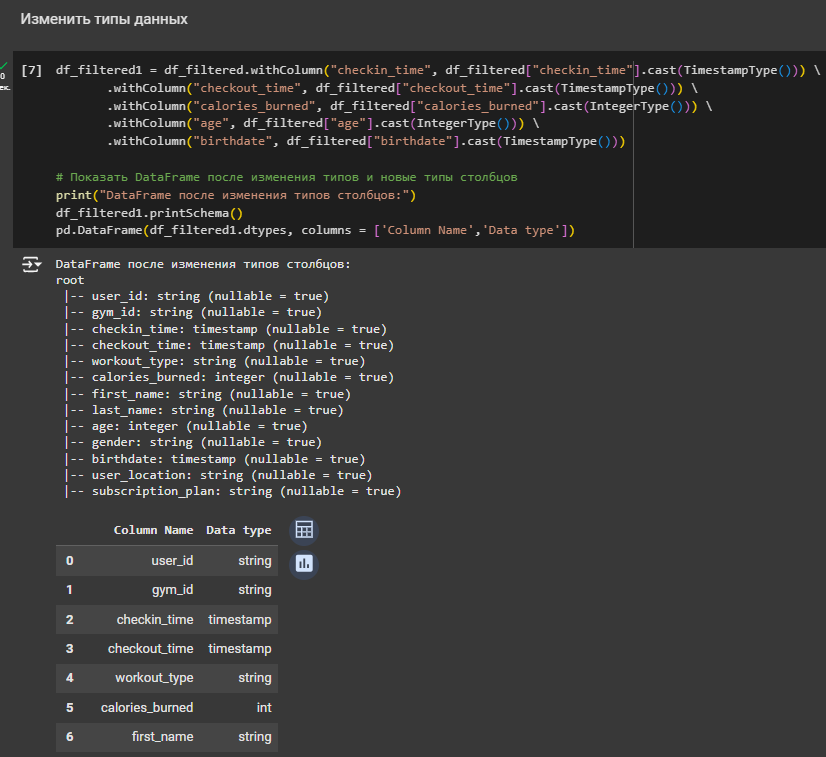


Рисунок 7. Изменение типов данных

Далее нужно выяснить, сколько в датафрейме строк. Их всего оказалось 300 000 (рисунок 8).

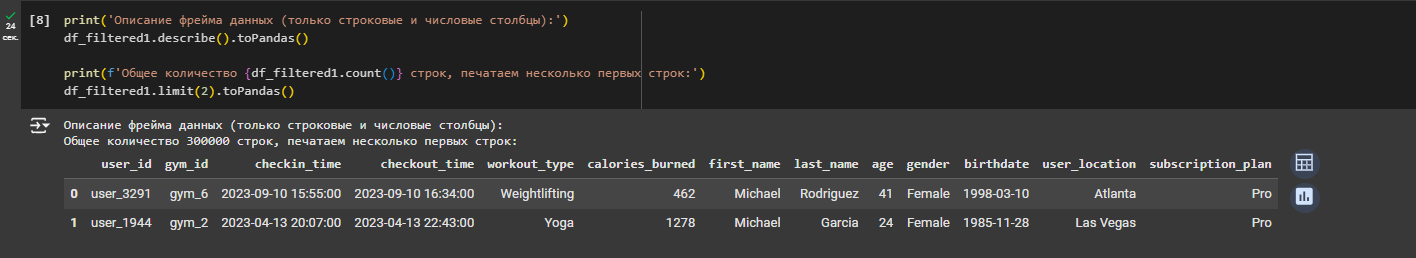


Рисунок 8. Общее количество строк

Затем нужно обнаружить пропущенные значения и аномальных нулей. Для этого для каждого типа данных – целое, числовое и массив надо установить поиск пропущенных значений, а именно 0, NaN, None, Null. Как можно заметить по рисунку 9, их не оказалось.

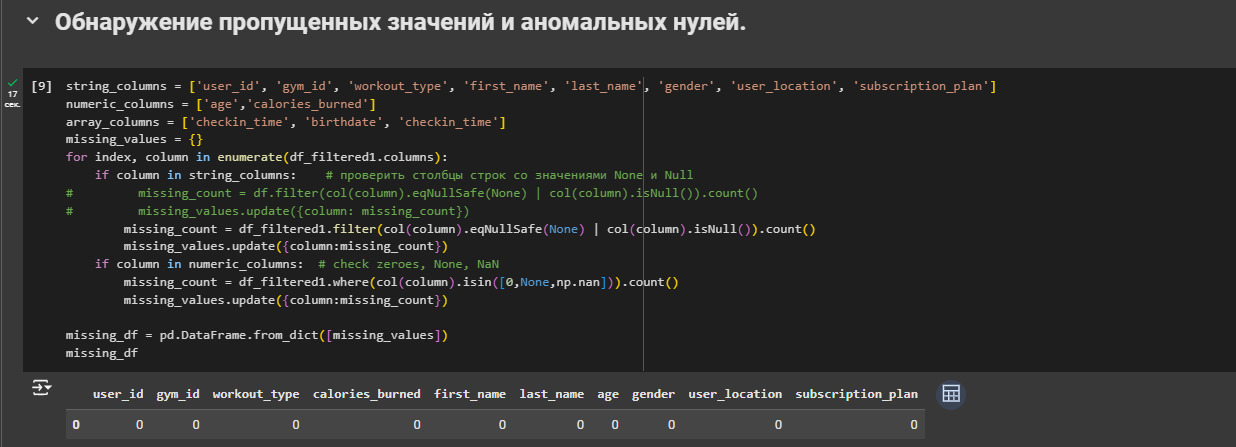


Рисунок 9. Проверка на пропущенные значения

Далее нужно рассчитать статистические показатели. Расчет квартилей показан на рисунке 10, а именно 1 квартиль равен 534 калории, второй 834 калории, а третий – 1179.

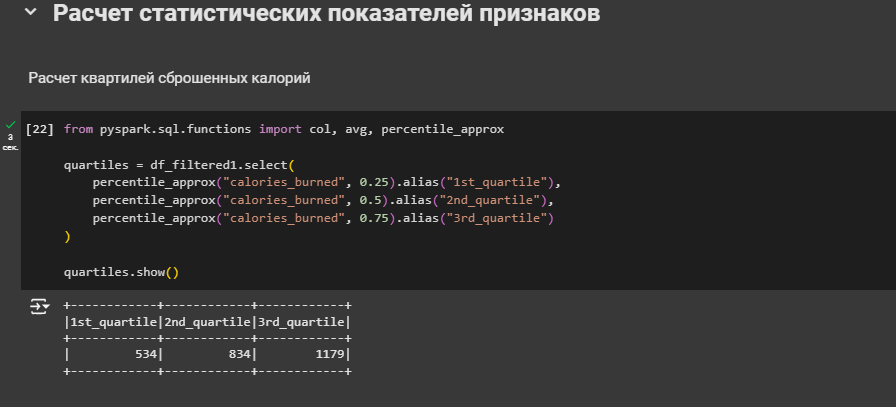


Рисунок 10. Расчет квартилей

Затем нужно вывести аномальные значения, и тем самым аномальными значения будут считаться, где сброшенных калорий больше 2146, а это только одно значение (рисунок 11).

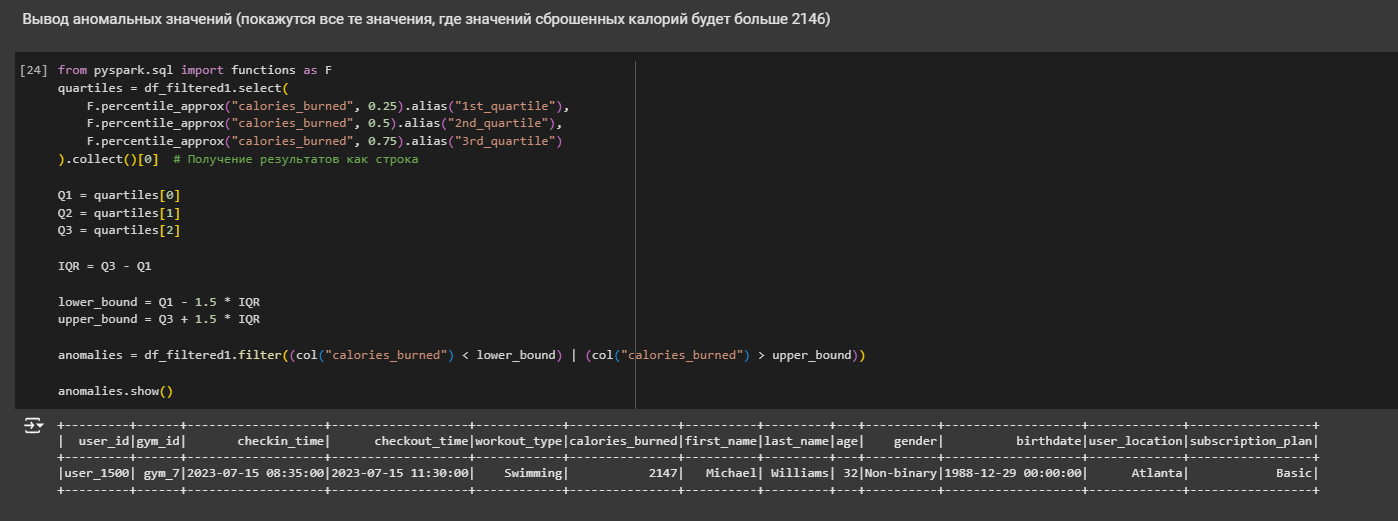


Рисунок 11. Аномальное значение

Затем нужно рассчитать средние, медиану, моду, дисперсию и стандартное отклонение для возраста и сожжённых калорий. Как можно заметить на рисунке 12, средний возраст – 41, медиана тоже 41, мода 64, дисперсия равна 183, а среднее отклонение – 13 лет. Среднее количество сожженных калорий – 886, медиана 834, мода – 490, дисперсия 187225, среднее отклонение 432.



Рисунок 12. Расчет статистических показателей

Далее нужно провизуализировать наиболее важные признаки. Для начала нужно посмотреть, как распределяются калории, сжигаемые за тренировку. Как можно увидеть, сжигается в основном от 400 до 1000 калорий (рисунок 13).

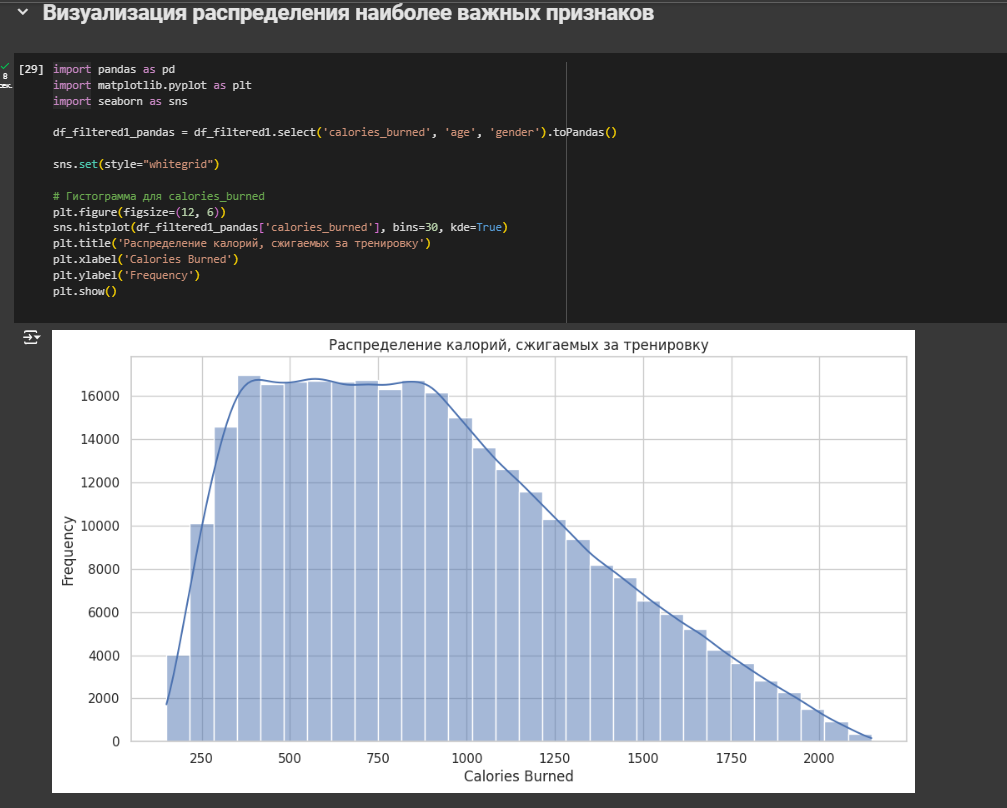


Рисунок 13. Распределение калорий, сжигаемых за тренировку

Затем нужно посмотреть, как распределяется возраст клиентов, которые посещают спортзал, для этого подойдет тип диаграммы боксплот. Диапазон от 30 до 55 лет (рисунок 14).

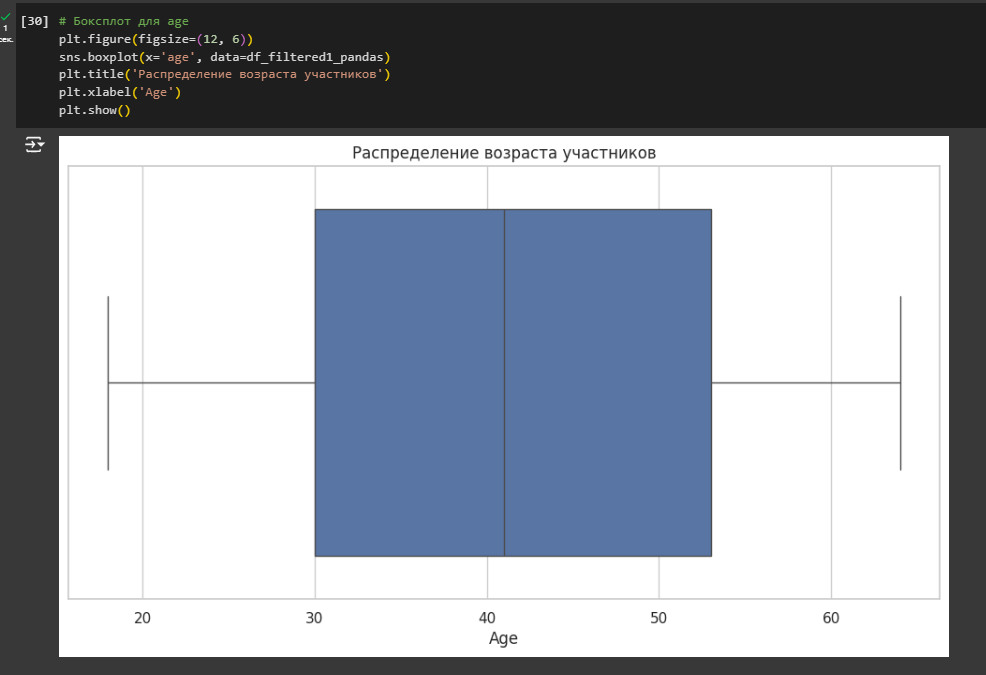


Рисунок 14. Диапазон возраста

Осталось посмотреть, какое распределение идет по полу клиентов, для этого успешно справится столбчатая диаграмма. Как можно увидеть, что мужчин на 9 тысяч больше, чем женщин, но при этом есть небинарные, их 30 тысяч (рисунок 15).

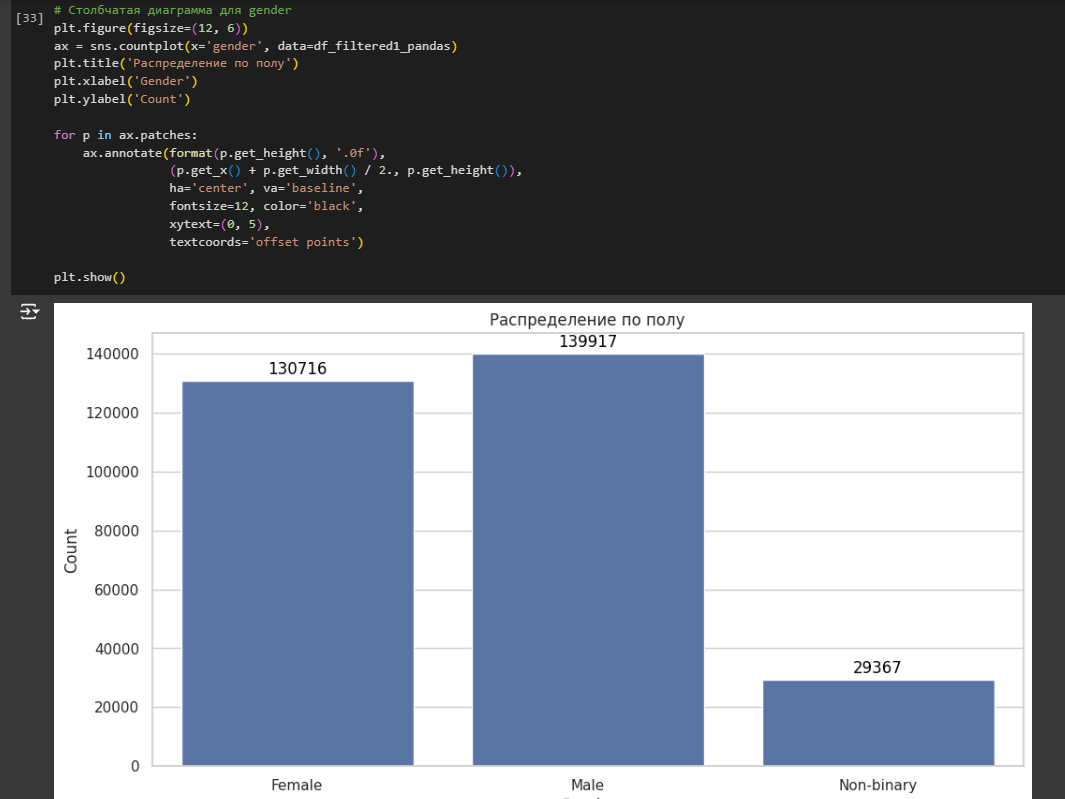


Рисунок 15. Распределение по полу

Корреляционный анализ между признаками возраст и сожженные калории показывают, что между ними связи нет, то есть разное количество калорий человек может сжечь в любом возрасте (рисунок 16).

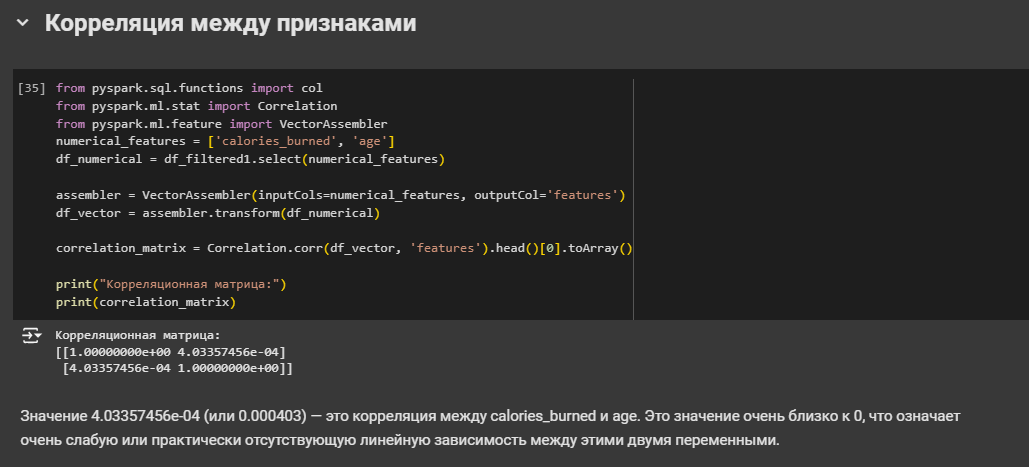


Рисунок 16. Корреляционный анализ

Затем нужно сделать несколько UDF для работы с данными. UDF используются для расширения функций платформы и повторного использования этих функций в нескольких DataFrame. Например, требуется преобразовать каждую первую букву слова в строке имени в заглавную; Встроенные функции PySpark не имеют этой функции, поэтому можно создать UDF и повторно использовать ее по мере необходимости во многих кадрах данных. После создания UDF их можно повторно использовать в нескольких выражениях DataFrame и SQL. Сначала нужно сделать извлечение возраста из даты рождения (рисунок 17).

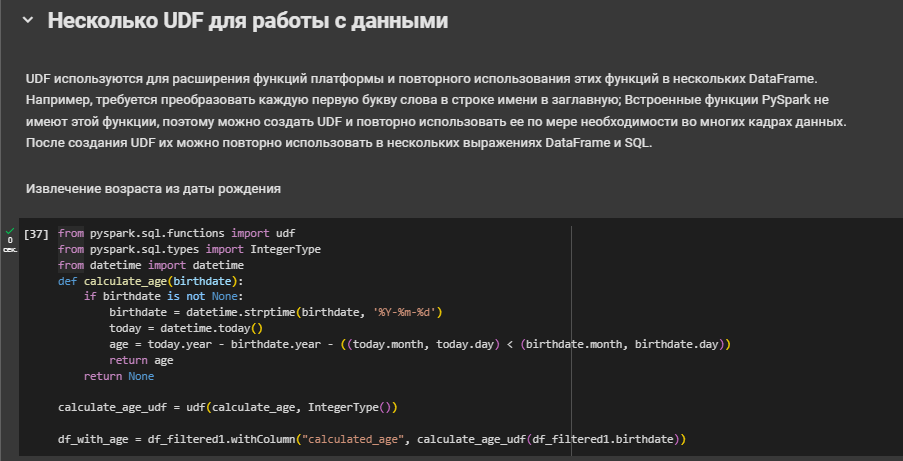


Рисунок 17. Извлечение возраста из даты рождения

Далее преобразование времени в минуты и кодировка пола, а затем дальнейшее объединение (рисунок 18).

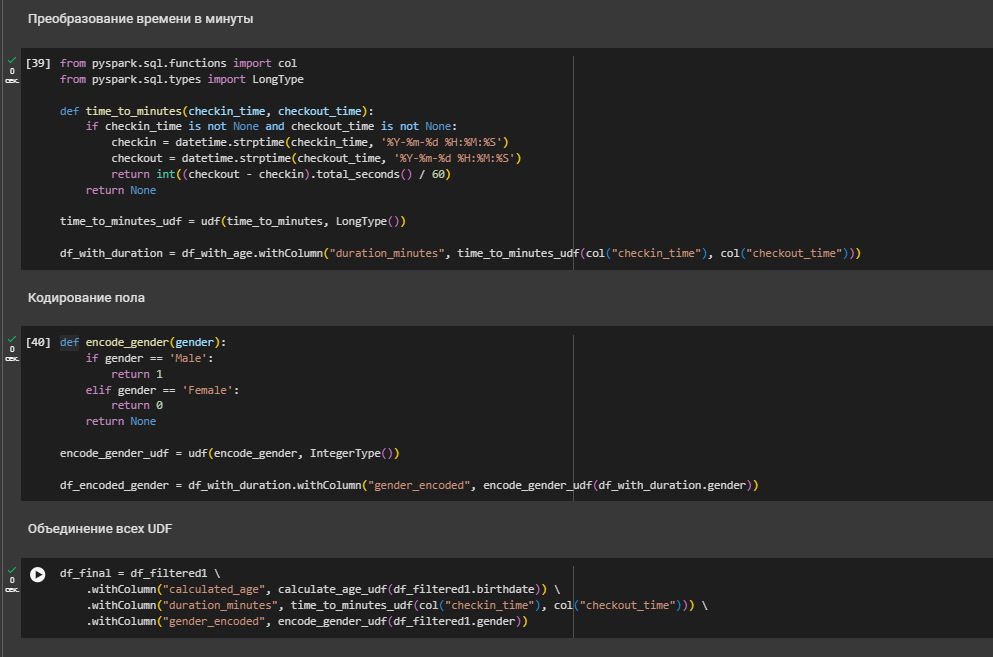


Рисунок 18. Преобразование времени в минуты и кодировка пола

Тем самым возраст может быть критически важным показателем для дальнейшей аналитики, кодировка пола полезна в машинном обучении, в том числе и в корреляционном анализе, а преобразование времени в минуты упростит расчёты.

**Выводы:**

1. При анализе типов признаков в датасете было установлено наличие как количественных, так и категориальных признаков. Различие в типах признаков важно для выбора подходящих методов анализа и визуализации данных.
2. Пропущенных значений в представленном датасете не оказалось.
3. Среди аномальных значений (выбросов) попало только 1 значение.
4. Были рассчитаны основные статистические показатели, такие как средние значения, медианы, квартильные значения, мода, дисперсия и стандартное отклонение. Эти показатели помогают в понимании распределения данных.
5. Визуализация данных проводилась с помощью таких типов диаграмм, как гистограмма, боксплот и столбчатая диаграмма.
6. Анализ корреляций проводился с использованием матрицы корреляции, но зависимостей не было обнаружено.
7. Было реализовано несколько пользовательских функций (UDF), таких, как преобразование времени в минуты, кодировка пола и извлечение возраста из даты рождения.