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): Морозова В.А. 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 импортируем необходимые библиотеки (рисунок 1)



Рисунок 1. Импорт необходимых библиотек

Далее подключаемся к спарку (рисунок 2)



Рисунок 2. Подключение к спарку

Затем идет обзор данных (рисунок 3).

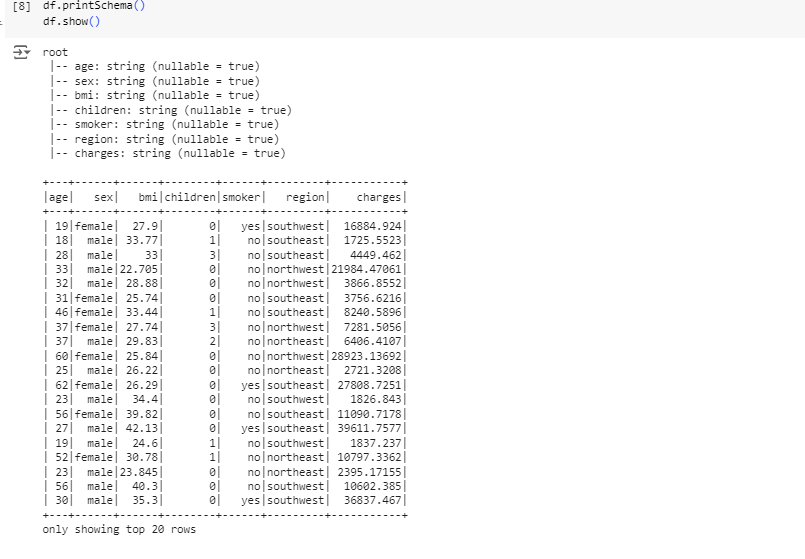


Рисунок 3. Обзор данных

Потом надо узнать типы данных (рисунок 4)

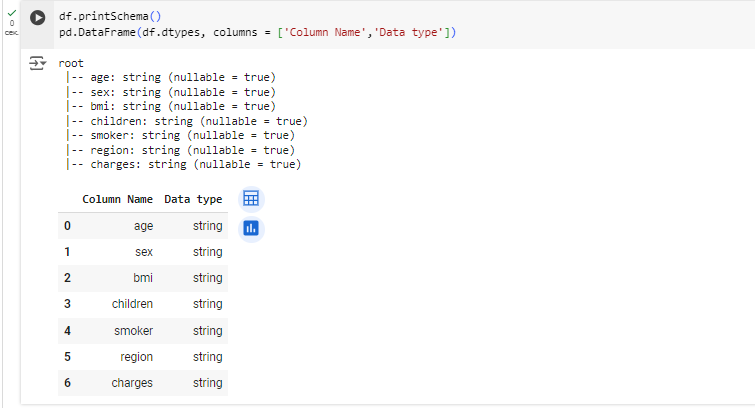


Рисунок 4. Узнавание типов данных

Изменение типов данных и вывод количества строк показано на рисунке 5.

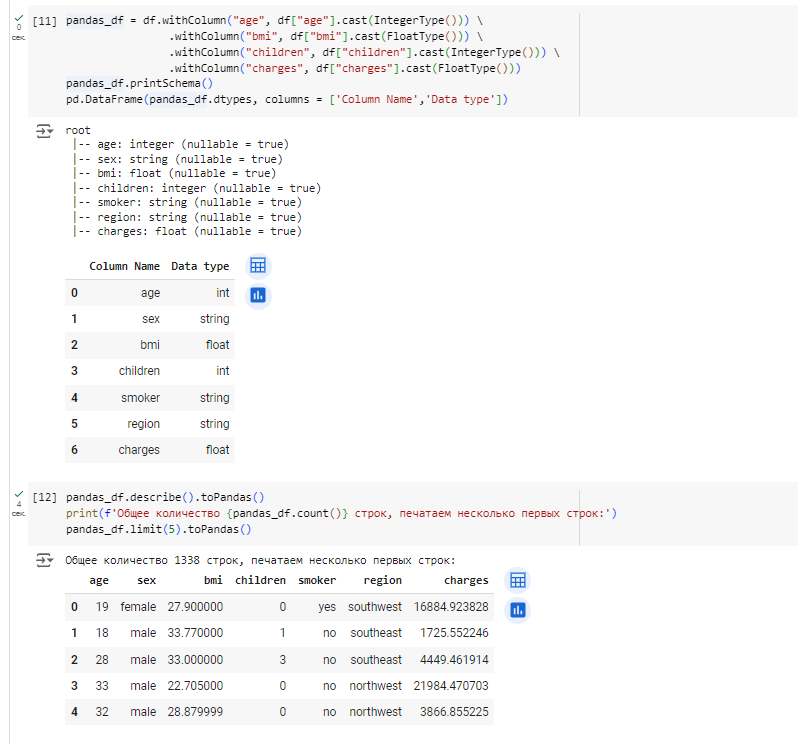


Рисунок 5. Изменение типов данных

Проверка наличия нулевых значений продемонстрировано на рисунке 6. Как можно заметить, их нет



Рисунок 6. Проверка наличия нулевых значений

Расчет статистических показателей (квартили, дисперсия, средняя и медиана) на рисунке 7.



Рисунок 7. Расчет статистических показателей

Параметры по полу, региону и отношению к курению относятся к категориальным данным, что недупустимо для модели, их нужно преобразовать в 0 и 1 для регрессионного анализа и получения точных результатов

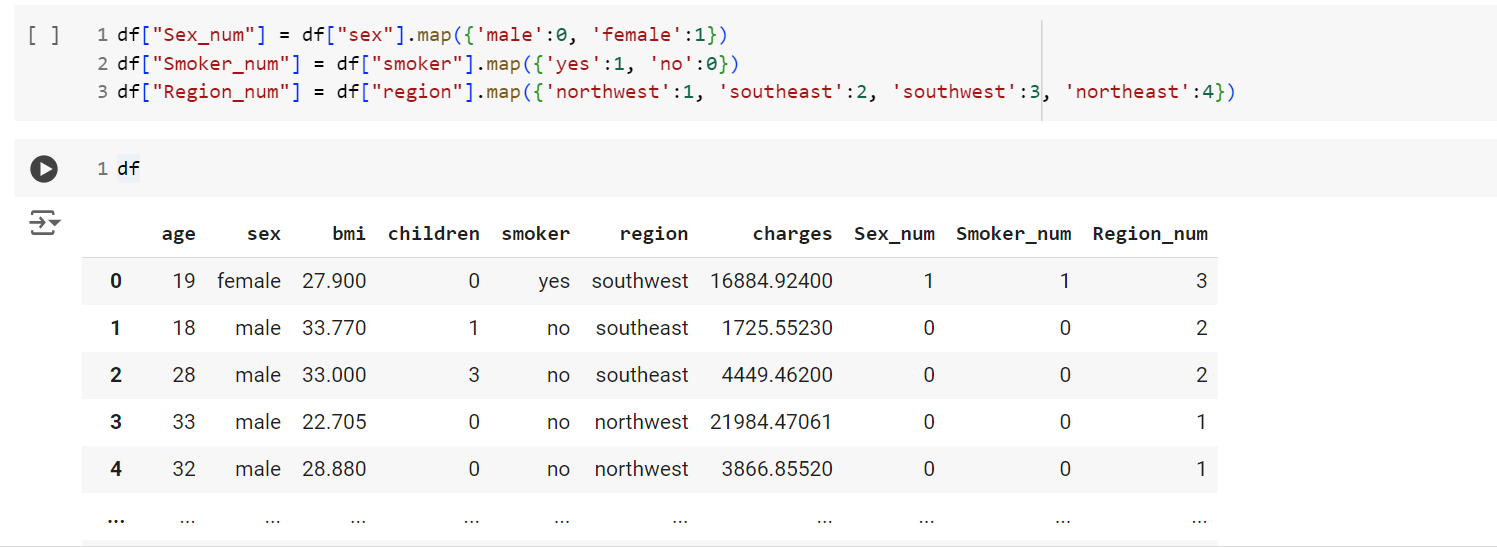


Рисунок 8. Преобразование данных для анализа

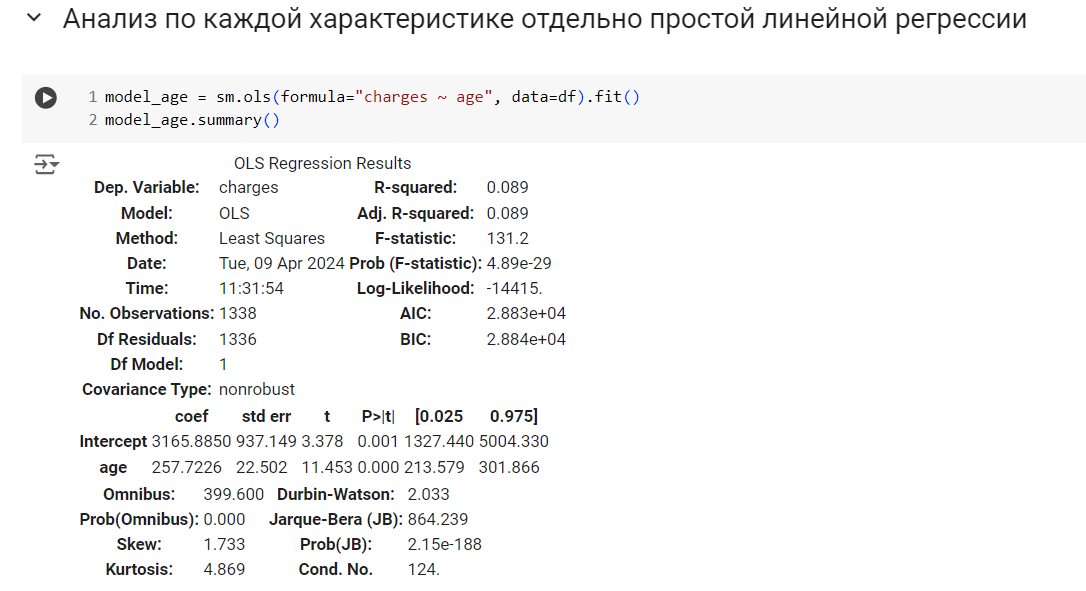


Рисунок 9. Анализ зависимости затрат от возраста

Результат означает, что каждый дополнительный год увеличения возраста увеличивает расходы на 257.7 денежных единиц

R2=8,9% это значит, что модель объясняет 8,9% изменчивости расходов. Это очень маленькое значение, указывающее на нехорошее соответствие модели данным.

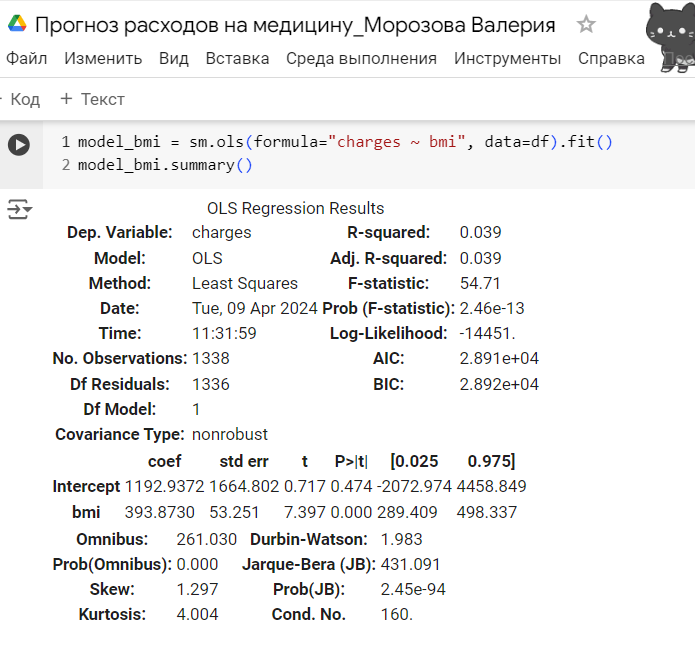


Рисунок 10. Зависимость затрат от ИМТ

Показывает, что увеличение ИМТ на каждую единицу увеличивает расходы на 393.8 денежных единиц.

R2=3,9% это значит,что модель объясняет 3,9% изменчивости расходов. Это очень маленькое значение, указывающее на нехорошее соответствие модели данным.

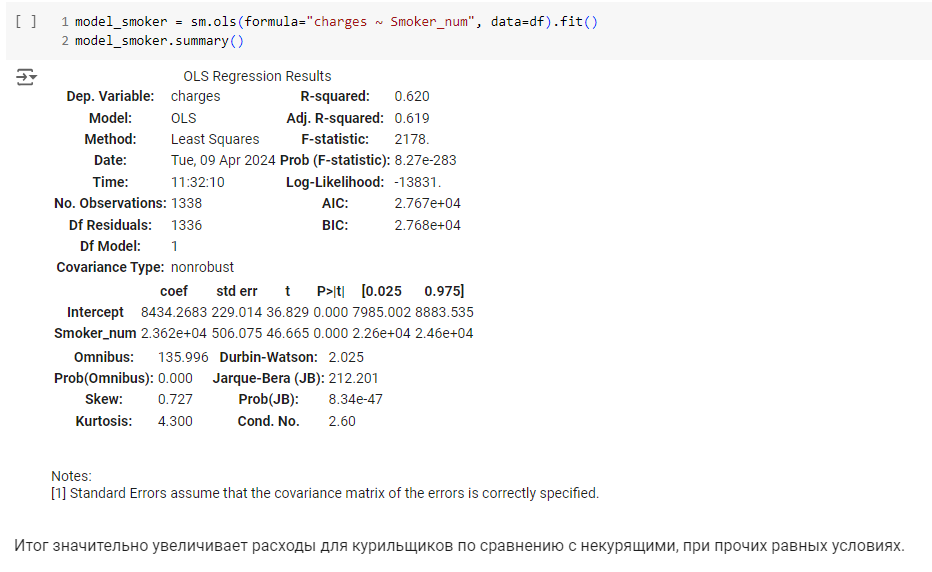


Рисунок 11. Зависимость затрат от статуса курения

Итог значительно увеличивает расходы для курильщиков по сравнению с некурящими, при прочих равных условиях.

**Визуализация**

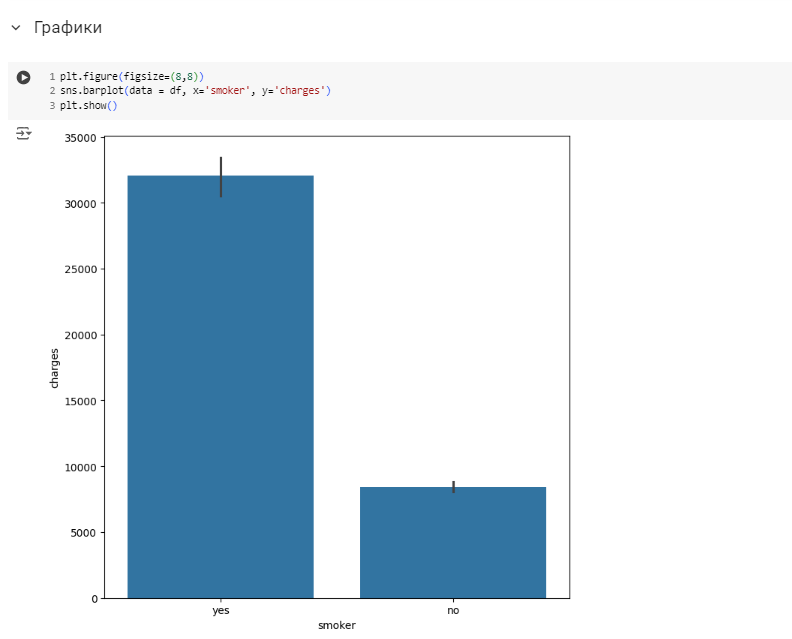


Рисунок 12. Количество курящих и не курящих



Рисунок 13. Демонстрация зависимости количества и статуса к курению и затрат на лечение

Количество людей по возрасту на рисунке 14. Как можно заметить, больше всего людей в возрасте до 20 лет

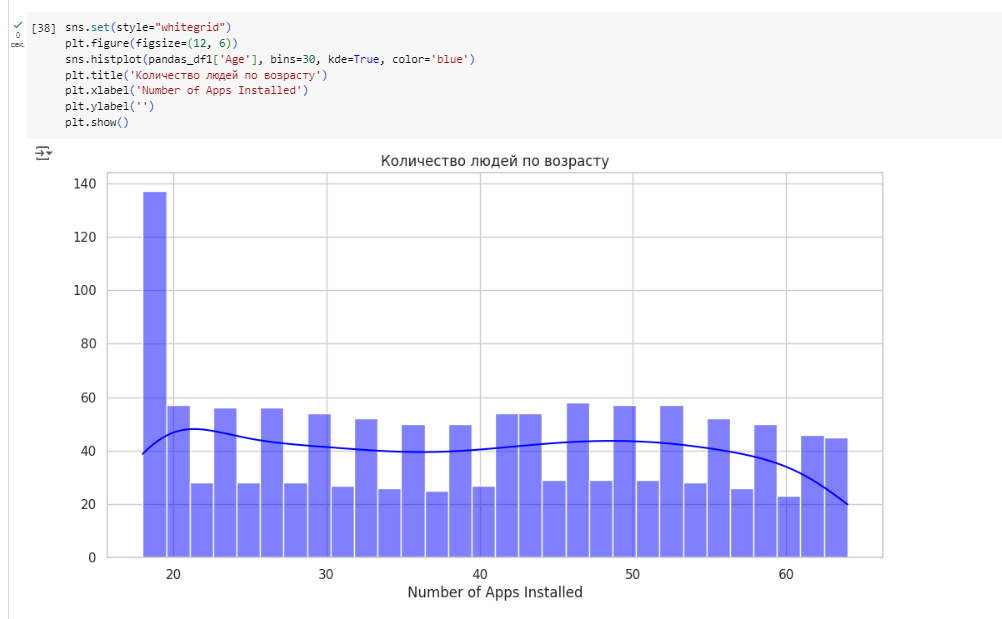


Рисунок 14. Количество людей по возрасту

Затем идет UDF: первый на возвращение отношения charges к age, а второй на классификацию, а второй на классификацию индекса масса тела (рисунок 15).

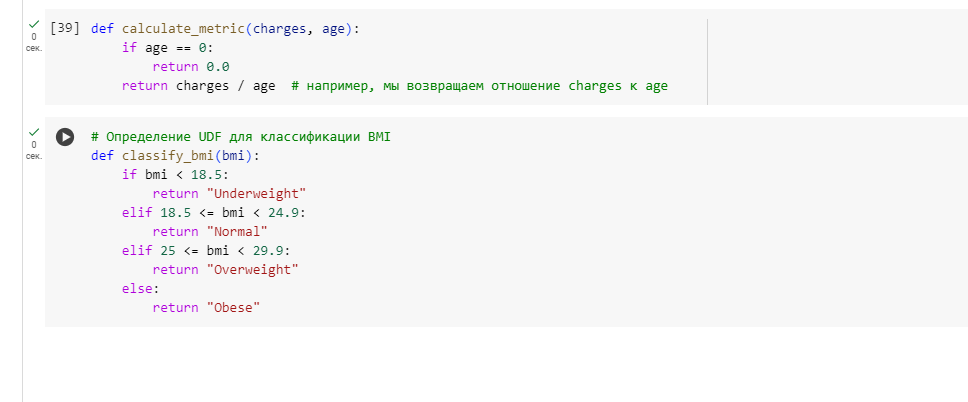


Рисунок 15. UDF

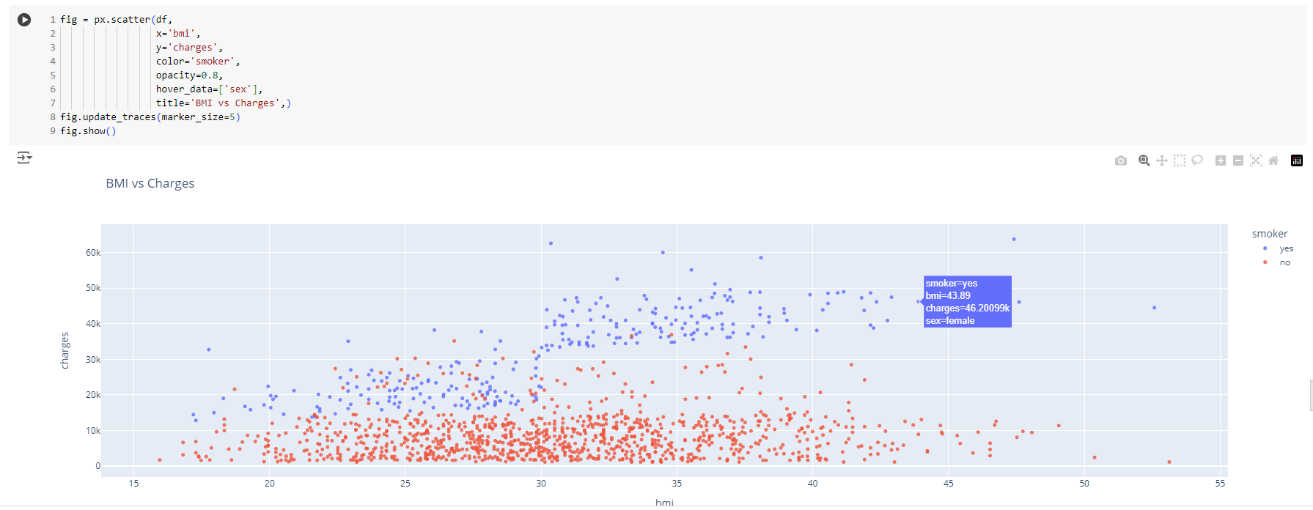


Рисунок 16. Отображение взаимозависимости ИМТ, курения и затрат

**Вывод:**

Чтобы выяснить, есть ли зависимость между демографическими показателями человека и расходами денежных средств, которые он готов потратить на медицинское обслуживание, был проанализирован набор данных, включающий такие метрики, связанные с клиентом, как пол, возраст, индекс массы тела, количество детей, регион и статус в отношении курения.

1. При анализе типов признаков в датасете было установлено наличие как количественных, так и категориальных признаков.
2. Пропущенных значений в представленном датасете не оказалось.
3. Были рассчитаны основные статистические показатели, такие как средние значения, медианы, квартильные значения и дисперсия.
4. Проведен анализ линейной регрессии, выявлена взаимосвязь.
5. Было реализовано несколько пользовательских функций (UDF).
6. Проведена наглядная визуализация.

Apache Spark — это целостная вычислительная система с набором библиотек для параллельной обработки данных на кластерах компьютеров.

Spark производит вычисления в реальном времени и обеспечивает низкую задержку благодаря их резидентному выполнению. В отличие от Hadoop, более ориентирован для обработки данных с использованием оперативной памяти на основе распределения данных.