Разработка системы анализа медицинских изображений для эпидемиологического мониторинга COVID-19

Аналитическая система для эпидемиологического мониторинга COVID-19 на основе метаданных рентгеновских снимков, реализованная с использованием распределённых вычислений PySpark. Проект направлен на автоматизацию анализа больших объёмов медицинских данных и выявление эпидемиологических паттернов.

Разработал: Тураносов К.В.

Архитектура проекта и методология обработки данных

Проект реализован в четыре последовательных этапа, обеспечивающих комплексный подход к анализу медицинских данных — от первичной оценки качества до распределённой обработки в PySpark.

01

Анализ качества данных

Проведён детальный анализ распределения пропущенных значений по всем полям датасета. Идентифицированы аномальные значения, включая некорректные возрастные данные, которые систематически обработаны и исключены из дальнейшего анализа для обеспечения достоверности результатов.

02

Предобработка данных

Заполнение пропусков выполнено с применением статистических методов для числовых данных и частотного анализа для категориальных переменных. Унифицированы диагнозы для устранения неоднозначностей (объединение всех вариаций записи COVID-19). Удалены дубликаты для обеспечения точности анализа.

03

SQL-аналитика

Построена базовая статистика по диагнозам и распределению по полу. Применены оконные функции для сегментации данных. Проведена оценка временных трендов по датам исследований и анализ связи проекций снимков с диагнозами.

04

Обработка в PySpark

Разработаны пользовательские функции (UDF) для категоризации возраста и точной сегментации данных. Выполнена унификация диагнозов с применением распределённой фильтрации. Результаты сохранены в формате Parquet для оптимизированного хранения.

Результаты SQL-аналитики

Статистика по диагнозам

Анализ выявил доминирование COVID-19 в выборке с существенным отрывом от других заболеваний:

- **COVID-19:** 438 записей, 321 уникальный пациент
- **Unknown:** 97 записей, 77 уникальных пациентов
- **Pneumonia:** 49 записей, 34 уникальных пациента
- **Tuberculosis:** 17 записей, 11 уникальных пациентов

Возрастной анализ (топ-3 по группам)

COVID-19: 94, 93, 88 лет

Pneumonia: три пациента по 80 лет

Tuberculosis: 78, 70, 58 лет — более широкий возрастной диапазон

Unknown: 90, 90, 78 лет

Гендерное распределение

Выявлено преобладание мужчин во всех диагностических категориях:

- **COVID-19:** 290 мужчин / 148 женщин
- **Pneumonia:** 29 мужчин / 20 женщин
- **Tuberculosis:** 12 мужчин / 5 женщин
- **Unknown:** 57 мужчин / 40 женщин

Временные тренды и проекции снимков

Оценка временных трендов показала резкий всплеск диагностики COVID-19 в январе 2020 года (386 исследований), что соответствует началу пандемии. В 2019 году зафиксировано лишь 4 исследования COVID-19. Следует отметить, что в январь 2020 попали все записи с пропущенными датами, хотя этот месяц действительно был пиковым.

COVID-19

PA: 157 снимков
AP: 106 снимков
AP Supine: 94 снимка

Pneumonia

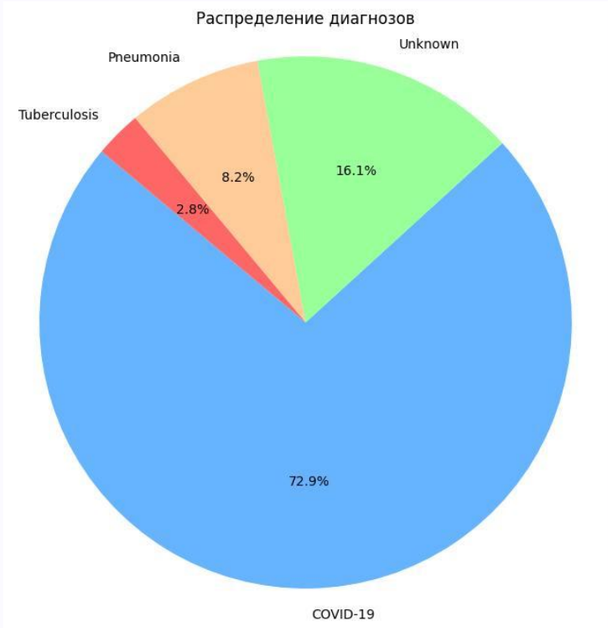
PA: 18 снимков
AP: 18 снимков

Tuberculosis

PA: 18 снимков

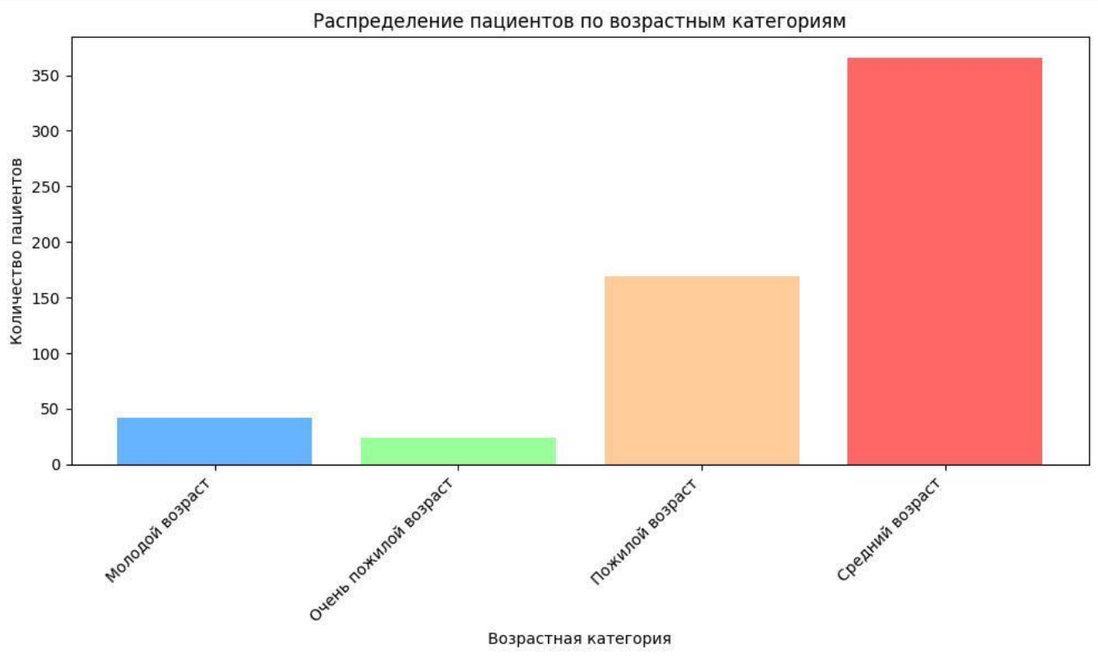
Визуализация данных и ключевые выводы

Для наглядного представления эпидемиологических паттернов и результатов анализа разработан комплекс информативных визуализаций, позволяющих оценить распределение заболеваний, демографические характеристики и временную динамику.



Круговая диаграмма распределения диагнозов

Визуализирует долевое соотношение основных диагнозов в датасете с акцентом на доминирование COVID-19



Столбчатая диаграмма по возрастным группам

Отображает распределение пациентов по возрастным категориям для каждого типа заболевания

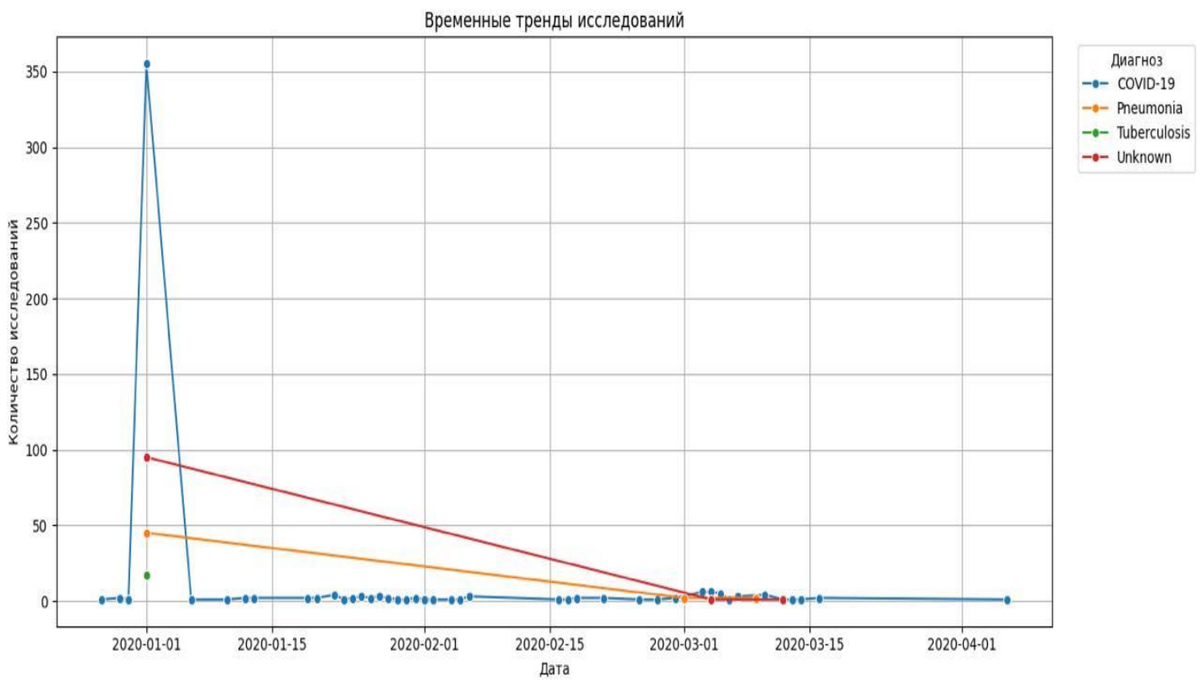
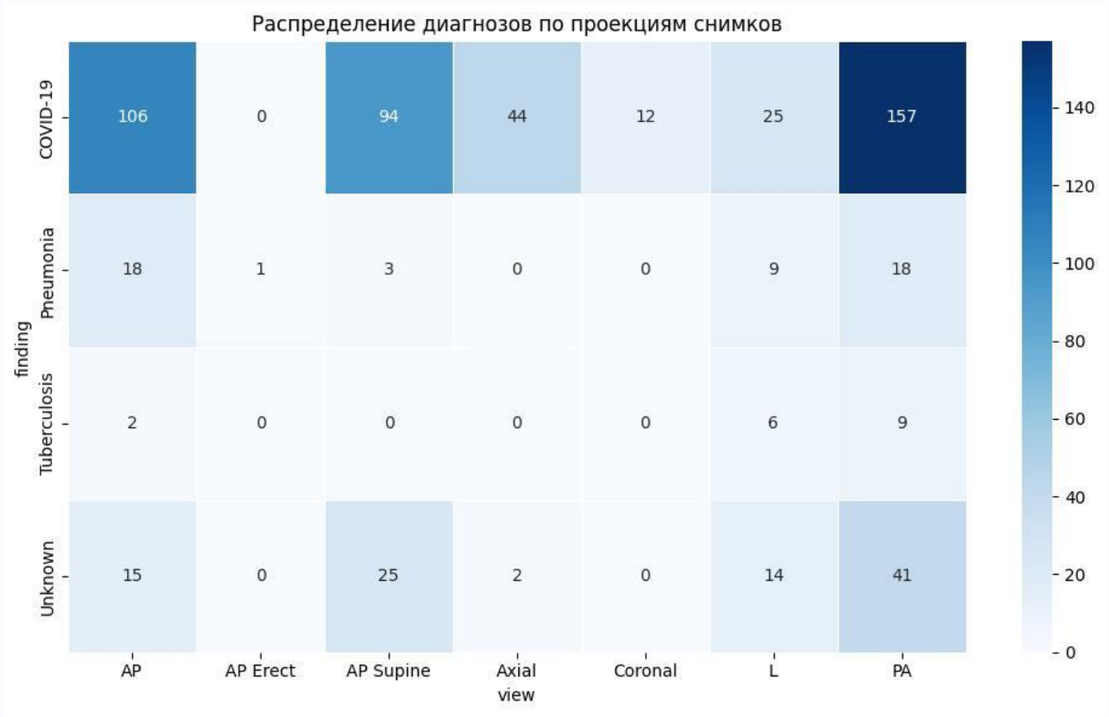


График временных трендов

Демонстрирует динамику количества исследований по месяцам с пиком в начале пандемии



Тепловая карта проекций снимков

Показывает взаимосвязь между типами рентгеновских проекций и диагнозами

Ключевые выводы проекта

Этот проект позволил на практике применить знания по работе с большими данными, используя PySpark для предобработки данных и SQL для анализа. Полученные результаты имеют важное значение для эпидемиологического мониторинга, поскольку помогают выявить ключевые закономерности, которые могут повлиять на дальнейшее принятие решений в области здравоохранения.