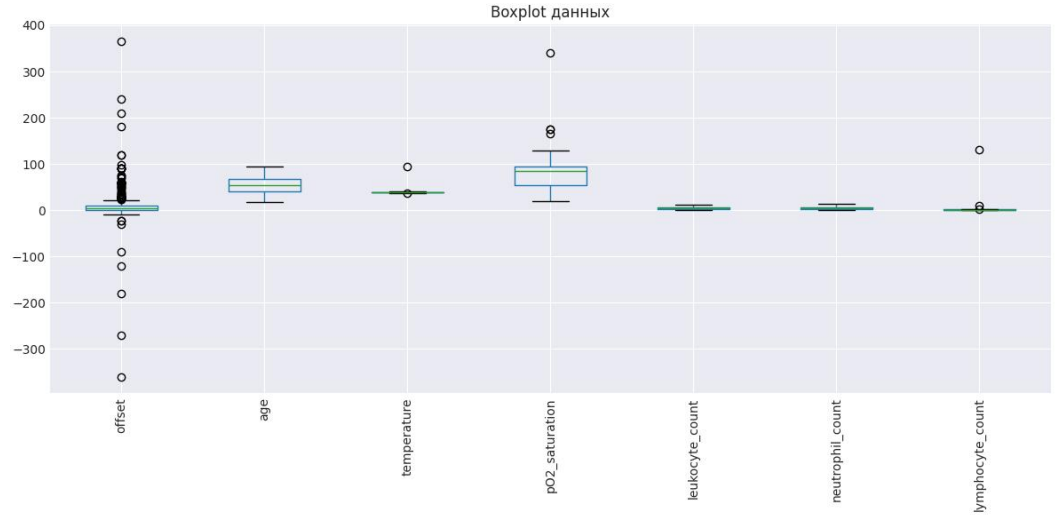
Полный цикл обработки медицинских данных

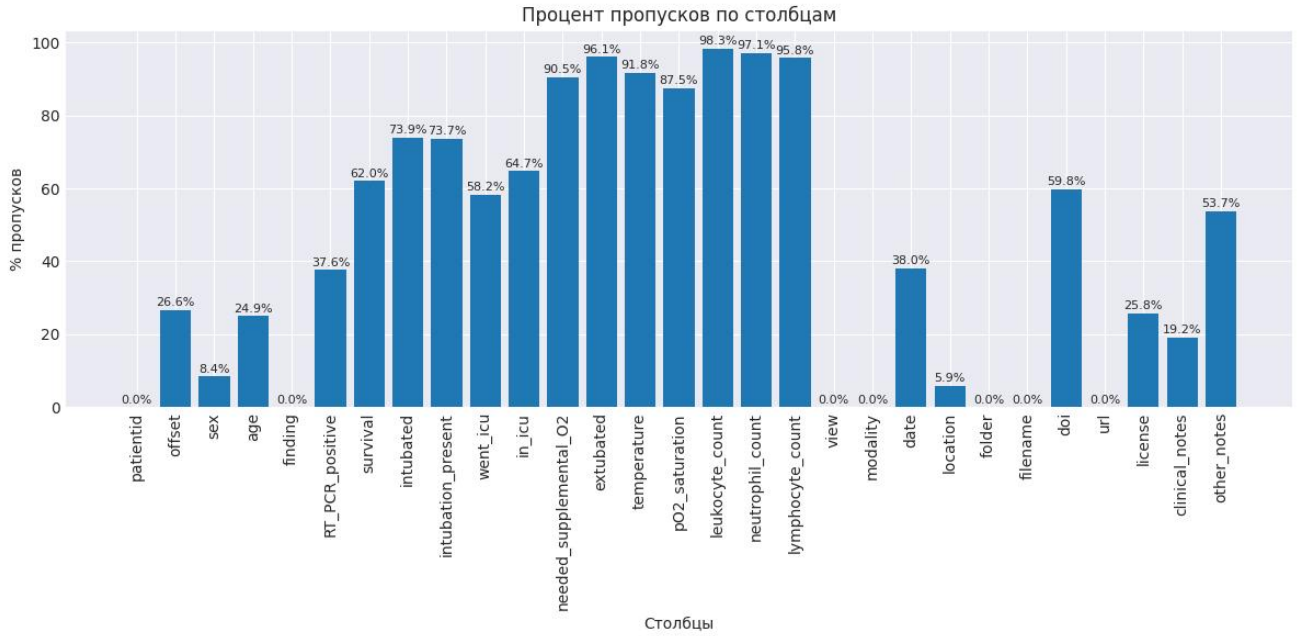
Сбор и предобработка данных

- Загрузка metadata.csv (950 записей, 29 признаков)
- Очистка: удаление дубликатов, обработка пропусков
- Унификация диагнозов и категоризация

SQL-аналитика

- 5 ключевых запросов на Spark SQL
- Анализ распределения по диагнозам, полу, возрасту
- Временные тренды (2004-2020 гг.)





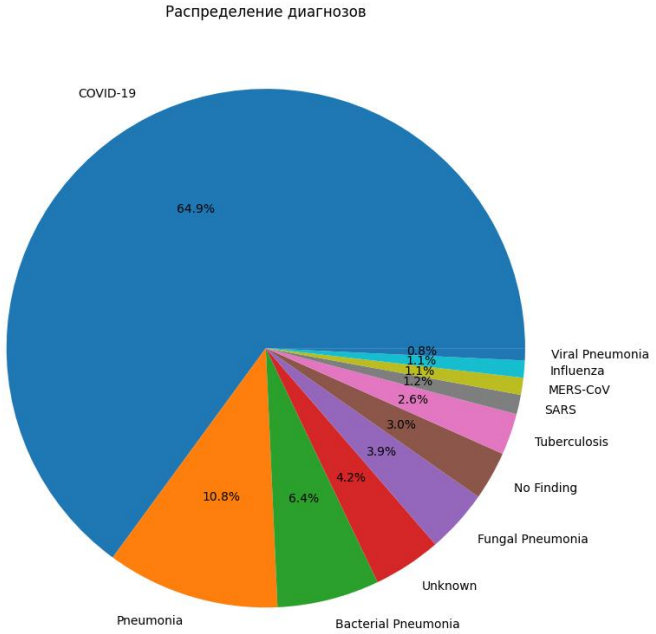
Ключевые статистики

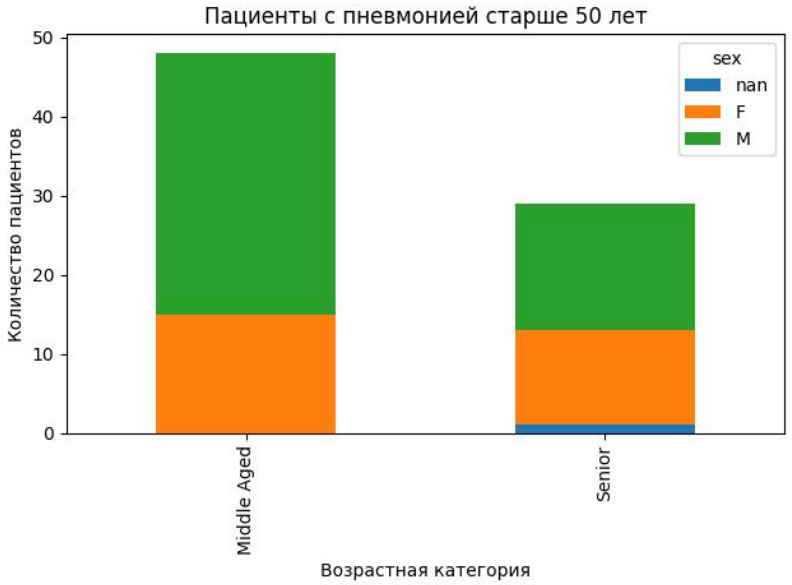
Основные метрики:

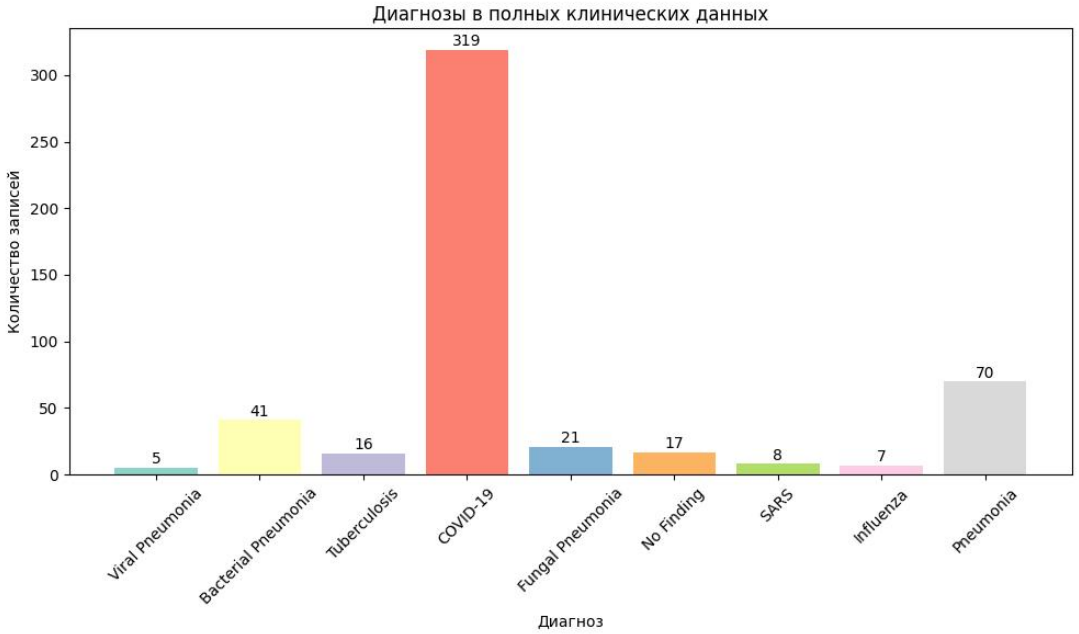
- 594 случаев пневмонии (62.5%)
- 53.5 года средний возраст пациентов
- 347 мужчин vs 186 женщин с пневмонией
- 302 исследования в 2020 году (пик активности)
- 77 пациентов с пневмонией старше 50 лет
- 504 записи с полными клиническими данными

Технические показатели:

- Обработано 5 SQL-запросов
- Создано 2 UDF функции для категоризации
- Сохранено 3 датасета в Parquet формате
- Выявлено 14 выбросов в числовых полях







Визуализации

Heatmap распределения диагнозов по проекциям:

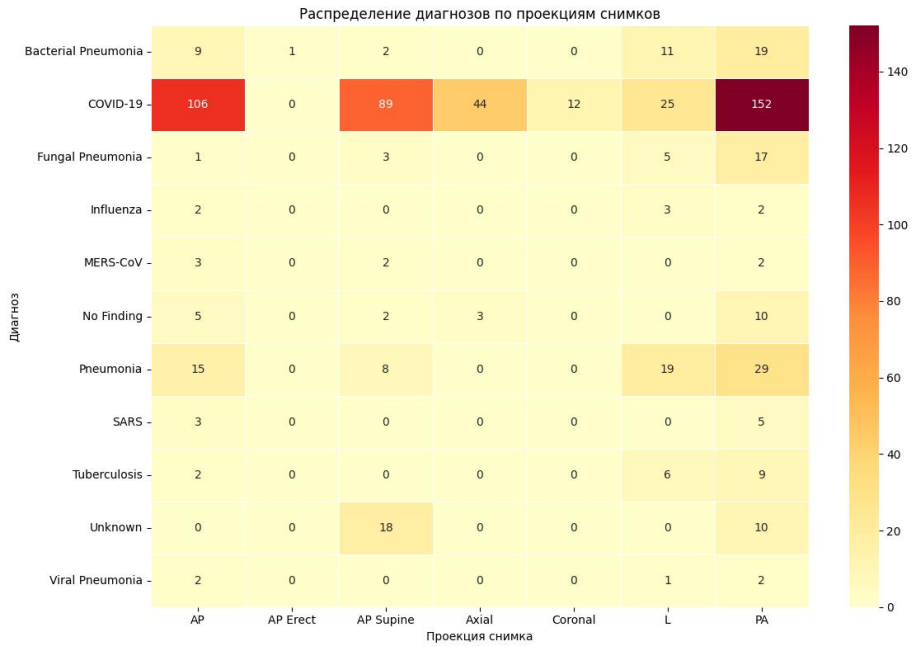
- PA и AP - наиболее частые проекции для пневмонии

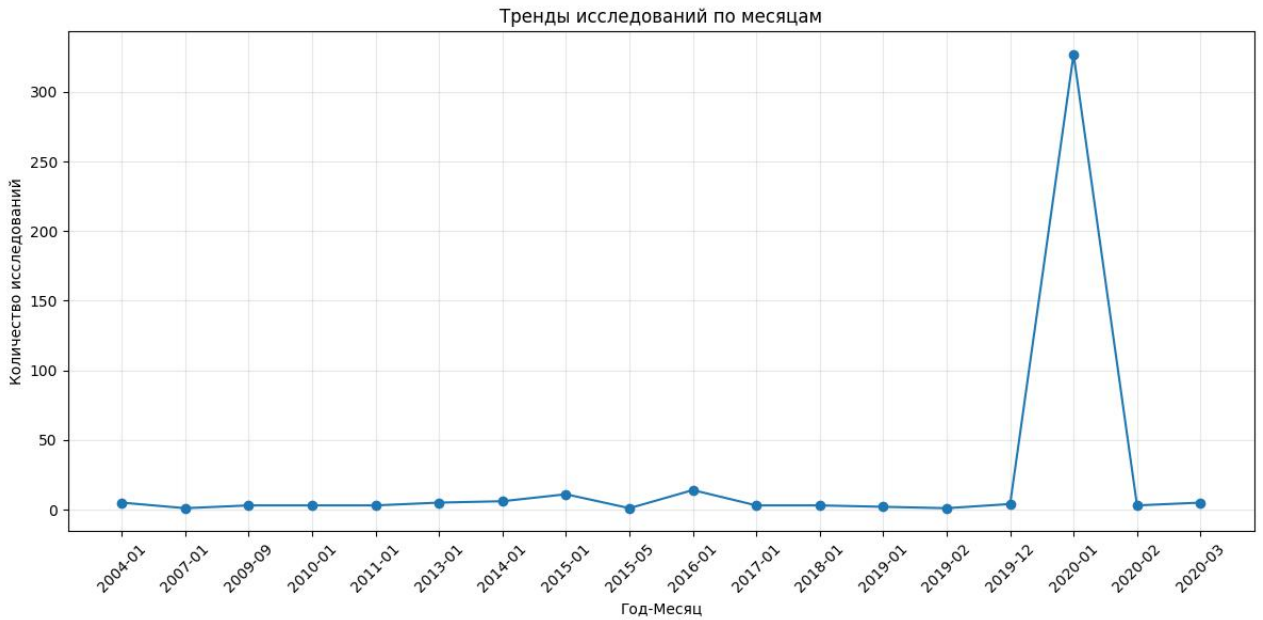
Распределение по возрастным группам:

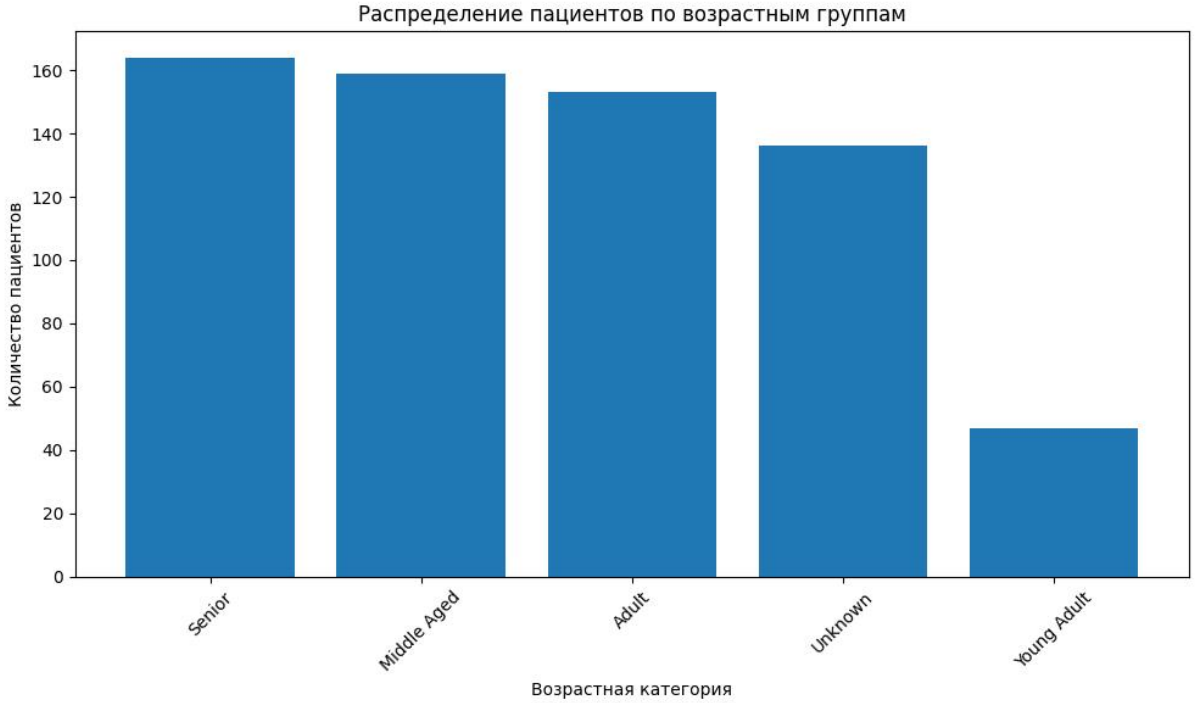
- Преобладание пациентов категории Senior (65+ лет)

Временной тренд исследований:

- Экспоненциальный рост в 2020 году (пик пандемии)







Ключевые выводы:

Качество данных: Высокий процент пропусков в критических полях (survival - 62%, лабораторные - 95-98%)

Демография: Четкое преобладание мужчин (347 vs 186) и пациентов старше 50 лет

Диагностика: 62.5% случаев - пневмония, из них 89.9% с указанием типа (534 из 594)

Временные тренды: 302 исследования в 2020 году подтверждают связь с пандемией