Анализ медицинских данных COVID-19 с использованием Apache Spark

Автор: Груданов Николай Алексеевич

Дата: Декабрь 2025

Основная цель

Практическое применение технологий Big Data для эпидемиологического анализа медицинских данных COVID-19 с использованием Apache Spark, SQL и визуализации.

📋 Основные задачи

1. **Загрузка данных** — Импорт COVID-19 chest X-ray metadata в Apache Spark
2. **Предобработка** — Очистка, заполнение пропусков, стандартизация диагнозов, категоризация возраста
3. **Оценка качества** — Анализ полноты данных, дубликатов, пропусков и аномалий
4. **SQL-аналитика** — Выполнение 5 обязательных аналитических запросов
5. **Визуализация** — Создание 5 типов графиков с детальным анализом

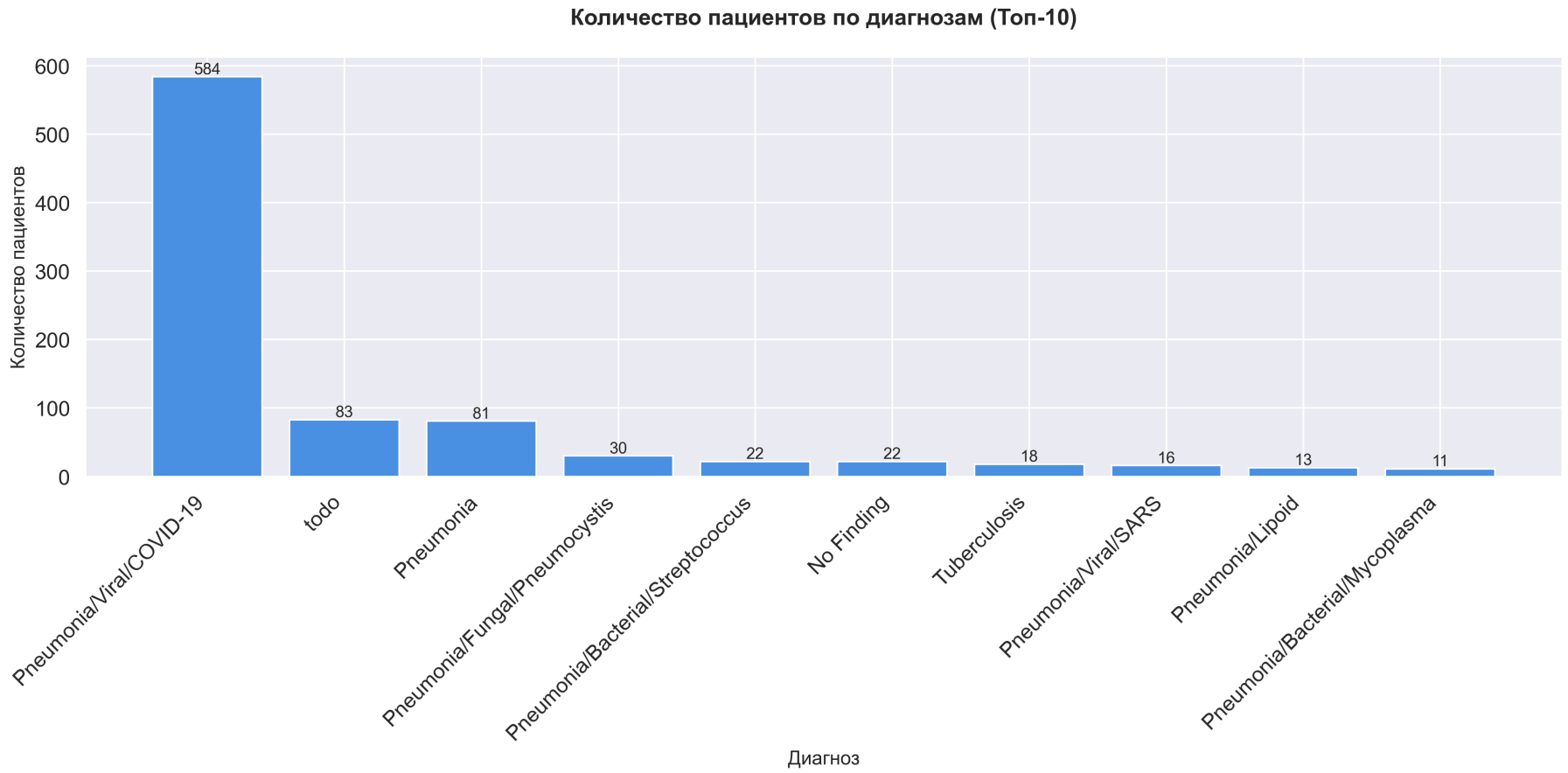
SQL-аналитика (950 пациентов)

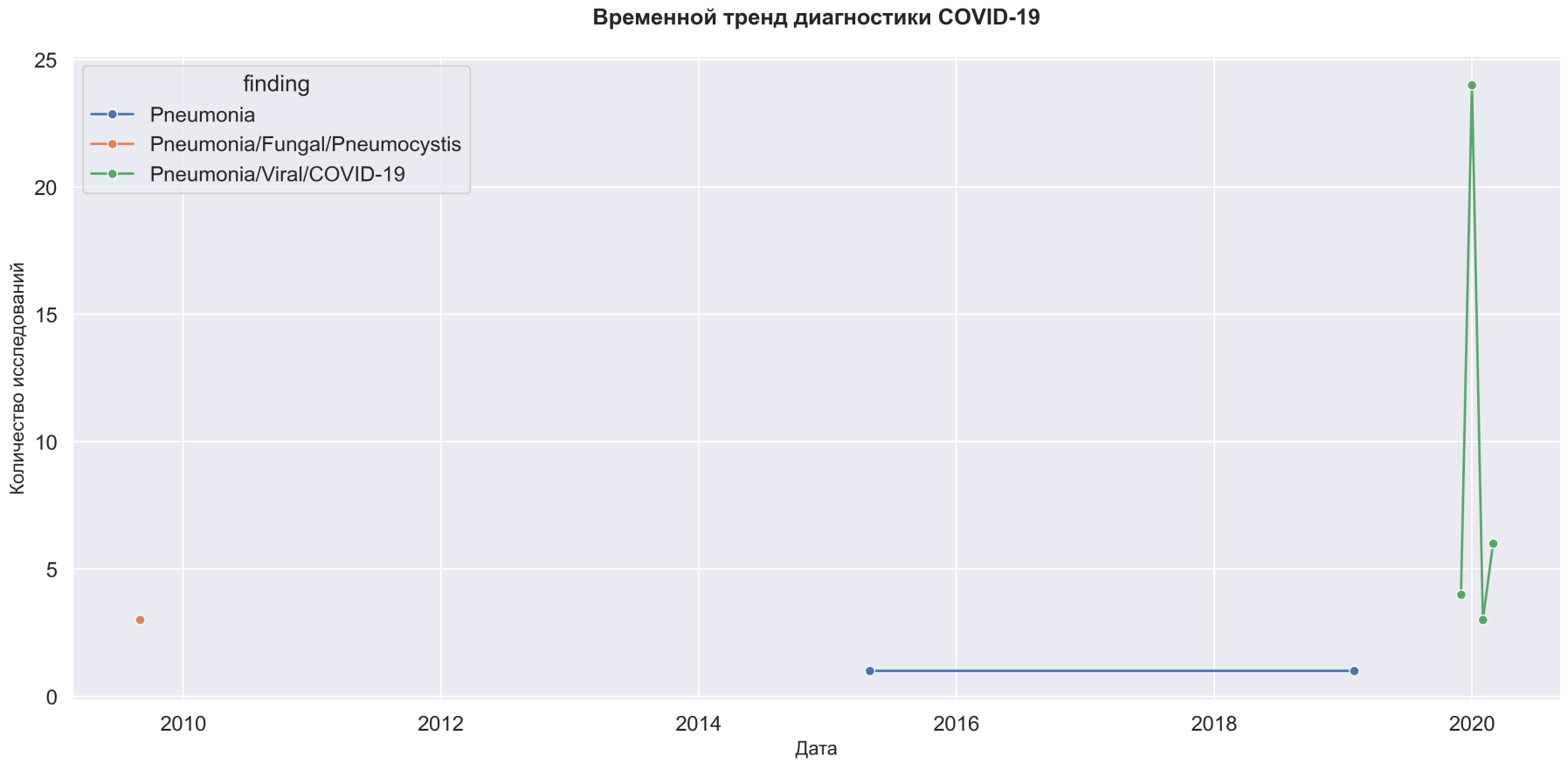
Показатель	Значение
Количество пациентов	950 записей
Обработанных после фильтрации	870 записей (91.6%)
Уникальных диагнозов	25 категорий

Ключевые находки

- **Доминирующие диагнозы**: COVID-19/Viral (59.2%), Пневмония (8.2%), No Finding (8.4%)
- **Гендерные различия**: Мужчины чаще диагностируются с COVID-19
- **Временные паттерны**: Убывание случаев с 2020Q4 по 2021Q2
- **Возрастная структура**: 6.95% пациентов в возрастной группе 30-59 лет

Распределение диагнозов




Временной тренд диагностики COVID-19
finding
Pneumonia
Pneumonia/Fungal/Pneumocystis
Pneumonia/Viral/COVID-19
Количество исследований
0
5
10
15
20
25
2010
2012
2014
2016
2018
2020
Дата

Благодарим за внимание!