«Архитектура решения»

- Источник данных:

  metadata.csv из репозитория ieee8023/covid-chestxray-dataset.

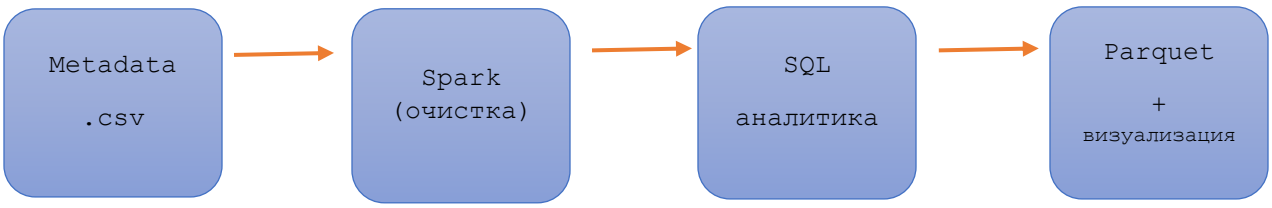
- Ingest: загрузка CSV в **Spark DataFrame**.

- Предобработка:
  - очистка пропусков (age, sex, date),
  - нормализация finding через UDF,
  - удаление дубликатов.

- Аналитика: **Spark SQL** (5 запросов + оконная функция).

- Обработка: фильтрация по критериям (например, COVID-19, возраст ≥ 18), сохранение в **Parquet**.

- Визуализация: **Pandas** графики в matplotlib / seaborn.



«Ключевые статистики»

- **Всего записей в метаданных: 950 , Пропусков в age (по исходному DataFrame): 237, после удаления дублей всего: 667 .**

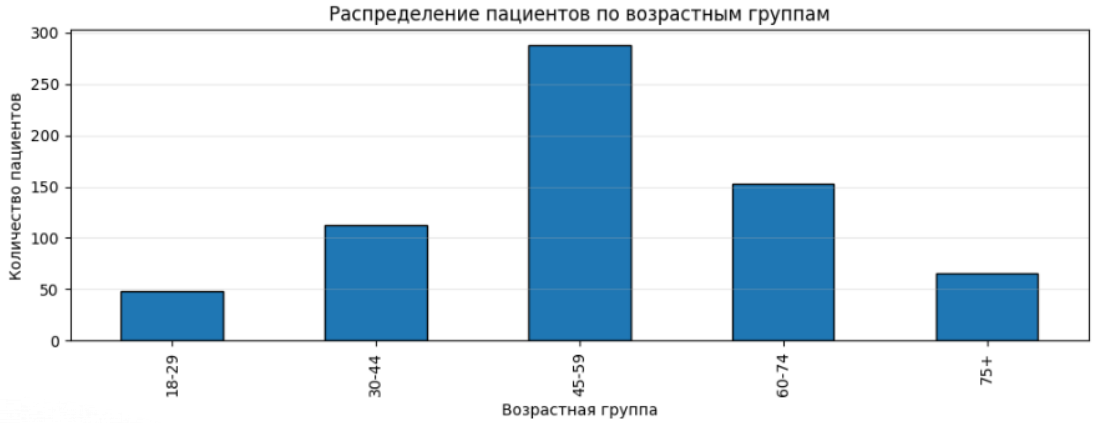
- **Доля COVID-19: ~ 65.5%, Pneumonia: 24.7%, остальные диагнозы — в сумме ~10%. (см 2.2)**

- **Возрастные группы: больше всего пациентов в сегментах 45–59 и 60–74 лет. (см 2.1)**

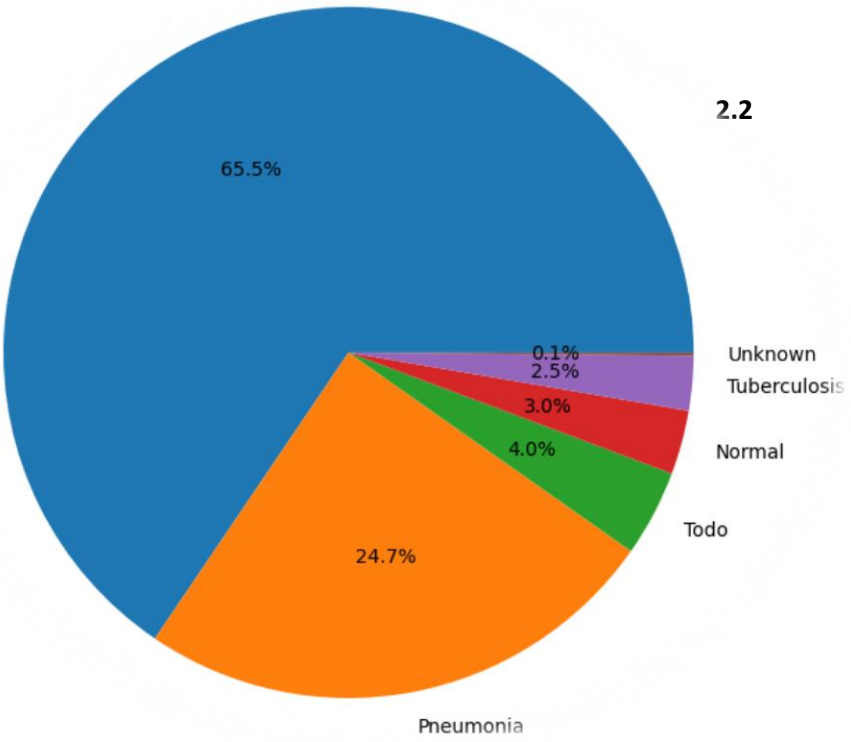
- **Существенные пропуски в age и sex, датам пришлось делать нормализацию и заполнение модой.**

- **Количество случаев COVID заметно выше у мужчин, чем у женщин** `62.02% VS 67.10%`

2.1



2.2



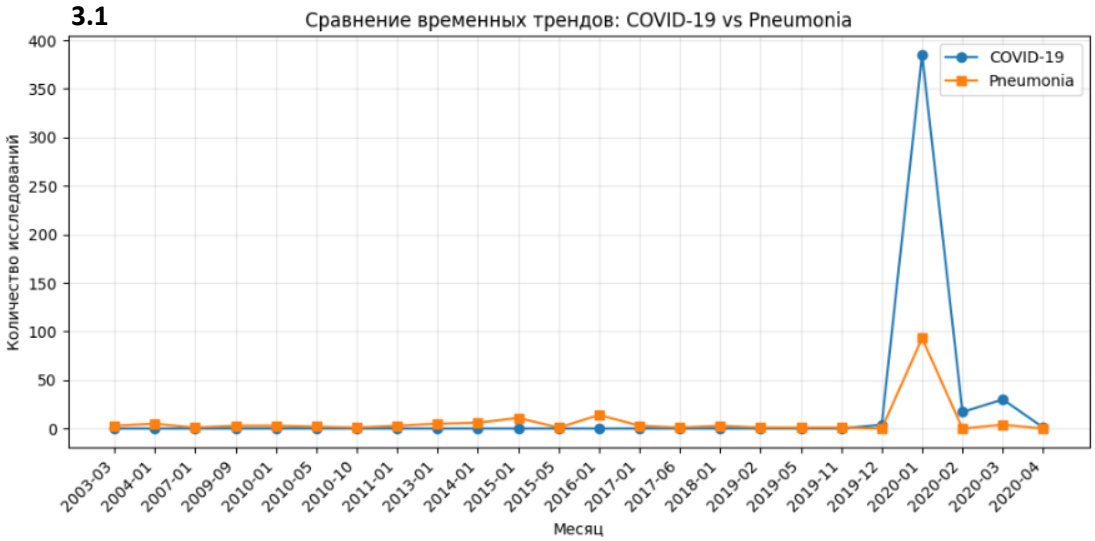
«Визуализации и выводы»

Пик количества исследований приходится на 2020 год, что совпадает с началом пандемии COVID. (см 3.1)

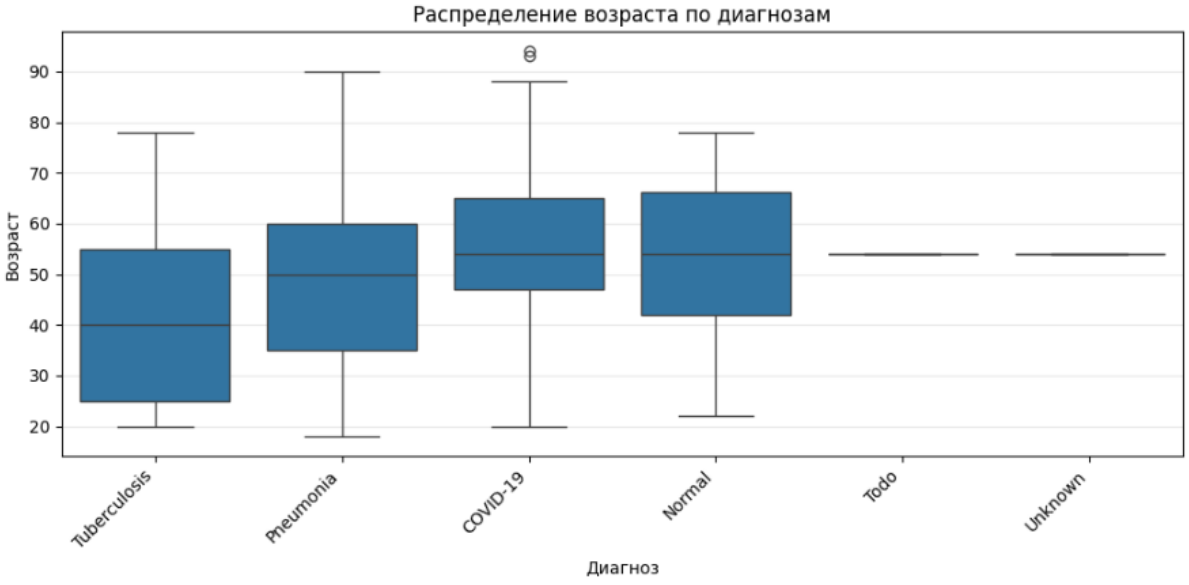
COVID и Pneumonia показывают схожие временные тренды с резким всплеском в начале 2020 года но как мы видим ниже, всплеск ковида гораздо сильнее (см 3.1)

Для Pneumonia и COVID медианный возраст находится примерно в районе 50 лет, с довольно широким разбросом и наличием пожилых пациентов. У Tuberculosis возрастная группа чуть моложе (см 3.2)

3.1

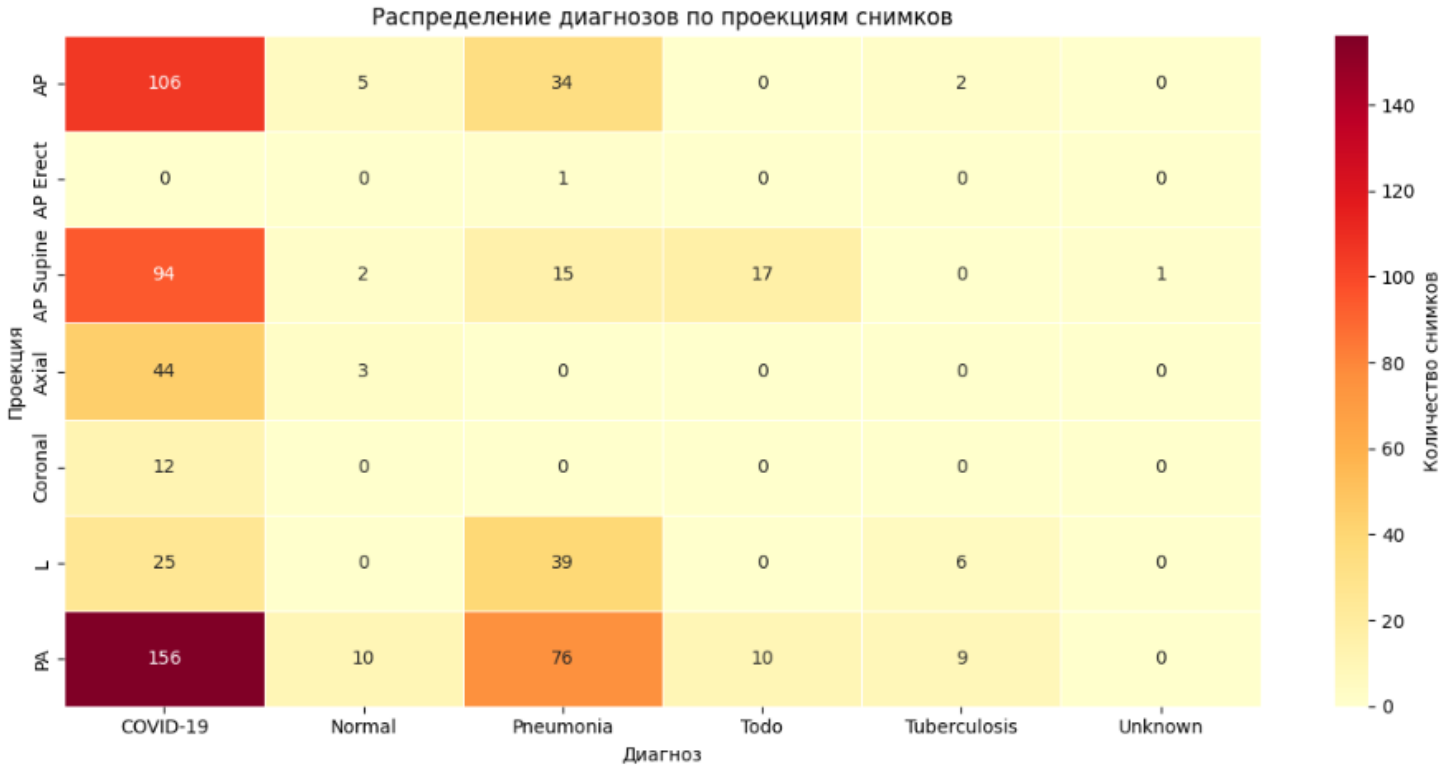


3.2



Основная часть снимков сделана в проекциях PA и AP, именно там концентрируются случаи COVID (см 4.1)

4.1



видно, что для COVID-19 чаще всего используются проекции PA и AP Supine, что логично для тяжёлых пациентов