

# Архитектура решения и стек:



1. Исходные данные: metadata.csv из COVID-19 Chest X-Ray Dataset + папка images со снимками.
2. Среда: Jupyter Notebook, PySpark 3.5.1 в режиме local[\*].
3. Библиотеки: pandas, matplotlib, seaborn, pyarrow.
4. Этапы обработки: чтение CSV → очистка и заполнение пропусков → создание признаков (finding\_unified, age\_group, date\_parsed).
5. SQL-аналитика: 5 запросов по временной таблице covid\_data (статистика по диагнозам, полу, возрасту и времени).
6. Результат: отфильтрованный набор только по X-ray, сохранённый в Parquet, и набор визуализаций для отчёта.

# Ключевые статистики по пациентам:

5. Временной анализ показывает резкий рост числа снимков с COVID-19 в конце 2019 — начале 2020 года, до этого почти все случаи — обычные пневмонии.

year_month	diagnosis	cnt
2004-01	Pneumonia (non-COVID)	11
2007-01	Pneumonia (non-COVID)	1
2009-09	Pneumonia (non-COVID)	3
2010-01	Pneumonia (non-COVID)	3
2010-05	Pneumonia (non-COVID)	2
2010-10	Pneumonia (non-COVID)	1
2011-01	Pneumonia (non-COVID)	5
2013-01	Pneumonia (non-COVID)	7
2014-01	Pneumonia (non-COVID)	11
2015-01	Pneumonia (non-COVID)	24
2015-05	Pneumonia (non-COVID)	1
2016-01	Pneumonia (non-COVID)	20
2017-01	Pneumonia (non-COVID)	3
2017-06	Pneumonia (non-COVID)	1
2018-01	Pneumonia (non-COVID)	6
2019-01	Normal	2
2019-02	Pneumonia (non-COVID)	1
2019-05	Pneumonia (non-COVID)	1
2019-12	COVID-19	4
2020-01	COVID-19	393
2020-01	Normal	7
2020-01	Unknown	83
2020-02	COVID-19	18
2020-03	COVID-19	28
2020-03	Normal	1
2020-04	COVID-19	1

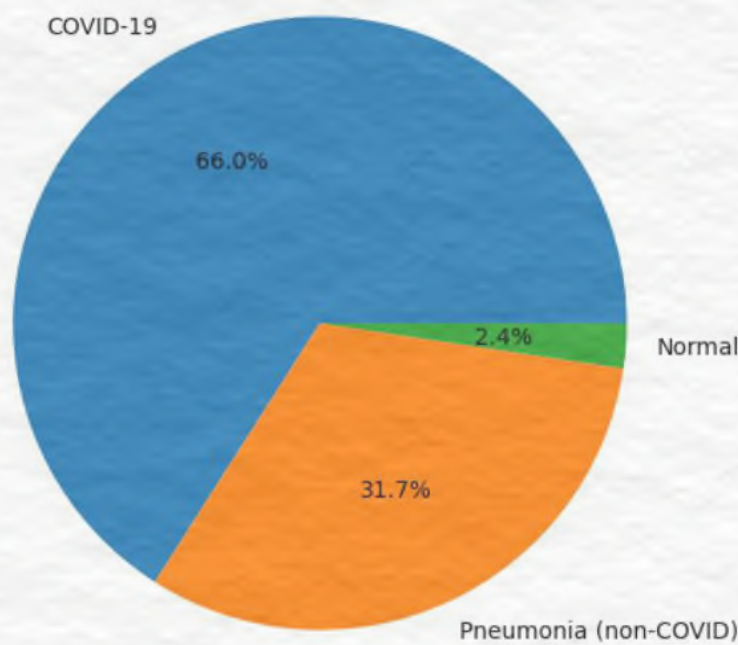
sex	diagnosis	cnt
F	COVID-19	175
F	Pneumonia (non-COVID)	98
F	Unknown	23
F	Normal	10
F	Other	5
M	COVID-19	346
M	Pneumonia (non-COVID)	129
M	Unknown	60
M	Other	13
M	Normal	11
Unknown	COVID-19	63
Unknown	Pneumonia (non-COVID)	15
Unknown	Normal	1
Unknown	Unknown	1

diagnosis	cnt	pct	avg_age
COVID-19	584	61.47	55.8
Pneumonia (non-COVID)	242	25.47	49.3
Unknown	84	8.84	54.0
Normal	22	2.32	52.5
Other	18	1.89	43.1

- 1. В исходном наборе доминирует диагноз COVID-19: ~61 % записей (584 случая).
- 2. Пневмония (non-COVID) около 25 % (242 случая), нормальные снимки ~2-3 %.
- 3. Средний возраст пациентов с COVID-19 около 56 лет, с пневмонией около 49 лет.
- 4. Больше всего случаев COVID-19 у мужчин: 346 против 175 у женщин; есть небольшой хвост записей с неизвестным полом

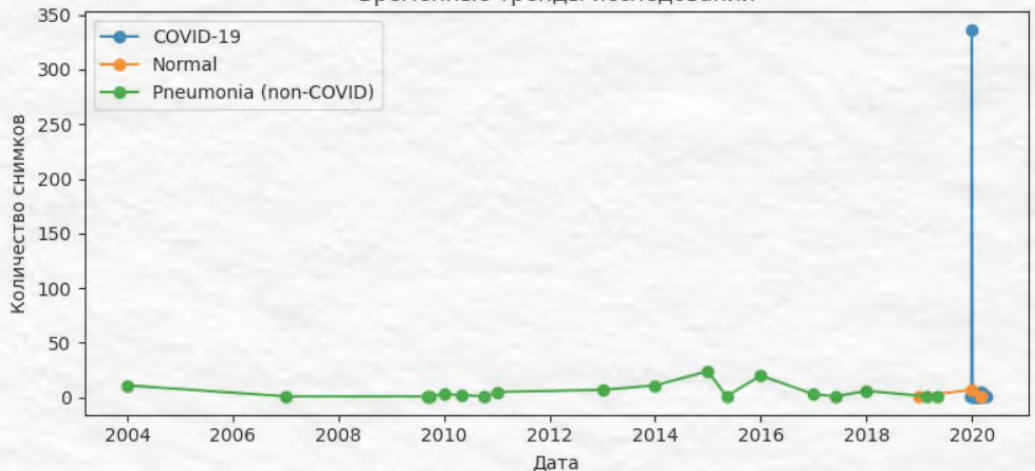
# Визуализация:

Распределение диагнозов



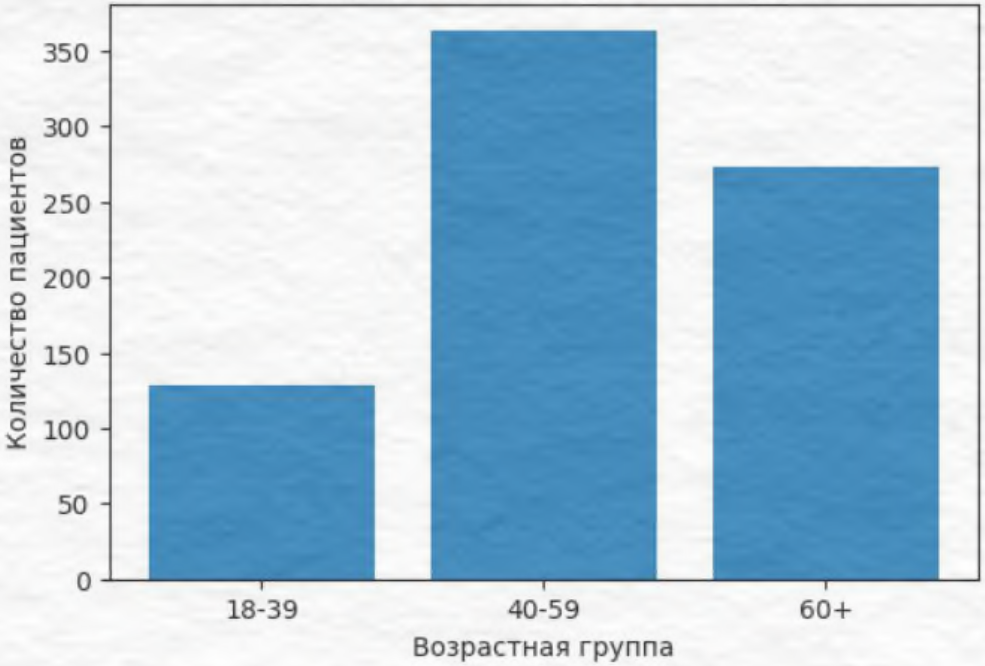
По круговой диаграмме видно, что в очищенной выборке  $\approx 66\%$  снимков относятся к COVID-19,  $\approx 32\%$  — к пневмонии поп-COVID, нормальных случаев всего несколько процентов.

Временные тренды исследований



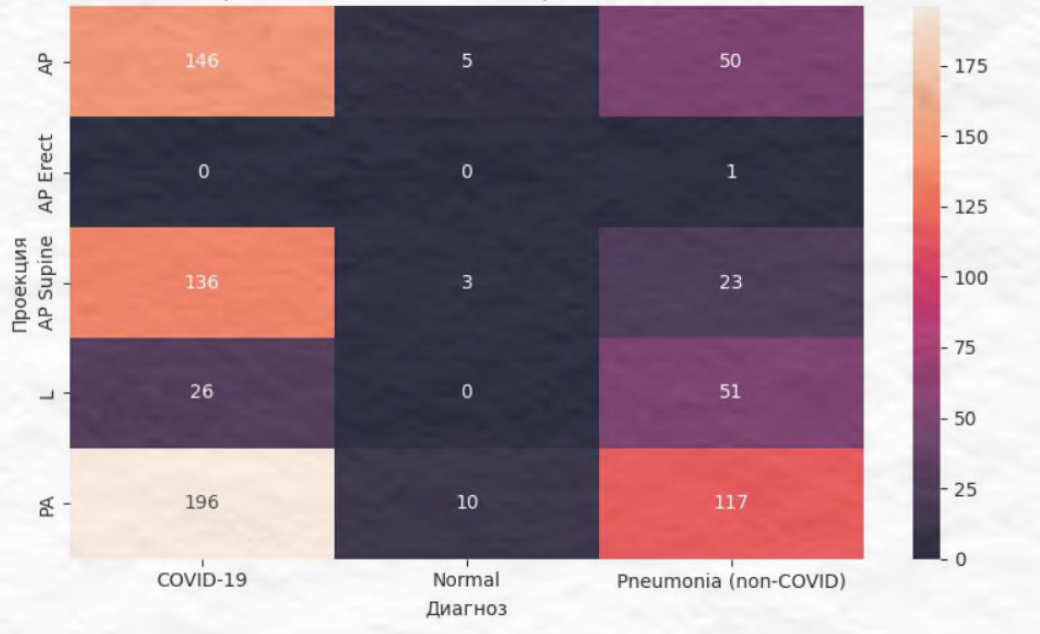
Временной график показывает, что до 2019 года в данных практически только пневмонии, а с 2020 года появляется мощный пик по COVID-19 — отражение начала пандемии.

Распределение по возрастным группам



Самая многочисленная возрастная группа — 40–59 лет, затем 60+; молодые пациенты (18–39) встречаются заметно реже.

Распределение диагнозов по проекциям снимков



Heatmap по проекциям показывает, что основная нагрузка приходится на стандартные проекции PA и AP, а в проекции AP Supine также много тяжёлых случаев COVID-19 и пневмонии.