Система анализа медицинских изображений для эпидемиологического мониторинга COVID-19

Архитектура, статистика, выводы

Алексей Сущих

Предобработка данных

При проведении анализа были выявлены пропуски

```
1. Анализ пропущенных значений:
patientid: 0 пропусков (0.0%)
age: 237 пропусков (24.9%)
sex: 80 пропусков (8.4%)
finding: 0 пропусков (0.0%)
view: 0 пропусков (0.0%)
date: 289 пропусков (30.4%)
```

Пропуски в возрасте и поле были заполнены средним

```
2. Заполнение пропущенных значений:
Медианный возраст для заполнения пропусков: 54
Наиболее частый пол: M
```

Формат даты был унифицирован: те даты, которые были просто пропущены в исходном датасете - оставлены как «NULL» для полноты данных

```
4. Обработка даты исследований:
Всего записей: 950
Успешно распарсено дат: 661 (69.6%)
Не удалось распарсить: 289 (30.4%)

Анализ наиболее частых форматов дат:
Для записей с успешным парсингом:
+----------------+-----------+
|date            |date_parsed|
+----------------+-----------+
|January 22, 2020|2020-01-22 |
|January 25, 2020|2020-01-25 |
|January 27, 2020|2020-01-27 |
|January 28, 2020|2020-01-28 |
|January 25, 2020|2020-01-25 |
|January 30, 2020|2020-01-30 |
|2017            |2017-01-01 |
|January 6, 2020 |2020-01-06 |
|January 10, 2020|2020-01-10 |
|2004            |2004-01-01 |
+----------------+-----------+
```

```
Для записей с неудачным парсингом:
+----+-----------+
|date|date_parsed|
+----+-----------+
|NULL|NULL       |
|NULL|NULL       |
|NULL|NULL       |
|NULL|NULL       |
|NULL|NULL       |
+----+-----------+
```

Так же были стандартизированы проекции снимков

```
5. Стандартизация проекций снимков:
Всего записей: 950
Известные проекции: 950 (100.0%)
Неизвестные проекции: 0 (0.0%)

Распределение по стандартизированным проекциям снимков:
+----------------+-----+-----------------+
|view_standardized|count|percentage       |
+----------------+-----+-----------------+
|AP              |438  |46.10526315789474|
|PA              |344  |36.21052631578947|
|Other           |168  |17.68421052631579|
+----------------+-----+-----------------+
```

… и удалены дубликаты

```
6. Удаление дубликатов:
Записей до удаления дубликатов: 950
Записей после удаления дубликатов: 641
Удалено дубликатов: 309
```

SQL-аналитика

1. Базовая статистика по диагн

diagnosis	count	percentage
COVID-19	412	64.27
Pneumonia	164	25.59
Unknown	27	4.21
Normal	20	3.12
Other	18	2.81

2. Распределение по полу

sex	diagnosis	count
F	COVID-19	124
F	Pneumonia	59
F	Normal	9
F	Unknown	6
F	Other	5
M	COVID-19	288
M	Pneumonia	105
M	Unknown	21
M	Other	13
M	Normal	11

3. Оконная функция (топ-3 по возрас

diagnosis	age	patientid	sex	rank
COVID-19	94	326b	M	1
COVID-19	93	324b	F	2
COVID-19	88	200	M	3
Normal	78	325	F	1
Normal	78	315	F	2
Normal	75	313b	M	3
Other	78	421	M	1
Other	70	453	M	2
Other	58	456	M	3
Pneumonia	90	460	M	1
Pneumonia	90	460	M	2
Pneumonia	80	91	F	3
Unknown	54	384	M	1
Unknown	54	387	M	2
Unknown	54	388	M	3

4. Анализ временных трендов по датам исследова

month	studies_count	diagnosis
2003-03	3	Pneumonia
2004-01	5	Pneumonia
2007-01	1	Pneumonia
2009-09	3	Pneumonia
2010-01	3	Pneumonia
2010-05	2	Pneumonia
2010-10	1	Pneumonia
2011-01	3	Pneumonia
2013-01	5	Pneumonia
2014-01	6	Pneumonia
2015-01	11	Pneumonia
2015-05	1	Pneumonia
2016-01	14	Pneumonia
2017-01	3	Pneumonia
2017-06	1	Pneumonia
2018-01	3	Pneumonia
2019-01	2	Normal
2019-02	1	Pneumonia
2019-05	1	Pneumonia
2019-11	1	Pneumonia

5. Статистика по проекциям снимков и их связи

view_type	diagnosis	count	percentage_in_view
AP	COVID-19	189	71.32
AP	Pneumonia	49	18.49
AP	Unknown	17	6.42
AP	Normal	7	2.64
AP	Other	3	1.13
Other	COVID-19	67	58.26
Other	Pneumonia	39	33.91
Other	Other	6	5.22
Other	Normal	3	2.61
PA	COVID-19	156	59.77
PA	Pneumonia	76	29.12
PA	Unknown	10	3.83
PA	Normal	10	3.83
PA	Other	9	3.45

Обработка в PySpark

1. Реализация пользовательских функций (UDF):
Применение UDF для категоризации возраста:

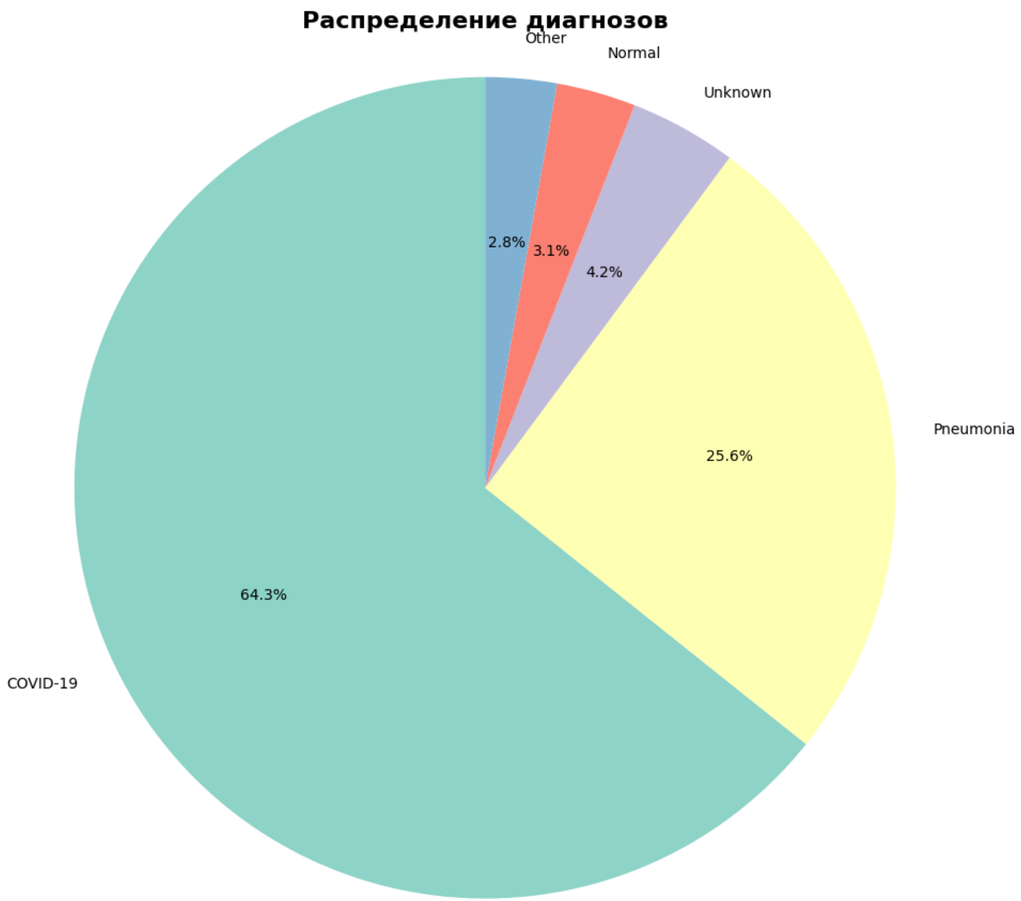
age_processed	age_category
65	51-65
52	51-65
29	21-35
35	21-35
54	51-65
54	51-65
54	51-65
54	51-65
53	51-65
55	51-65

Применение UDF для расширенной унификации диагнозов:

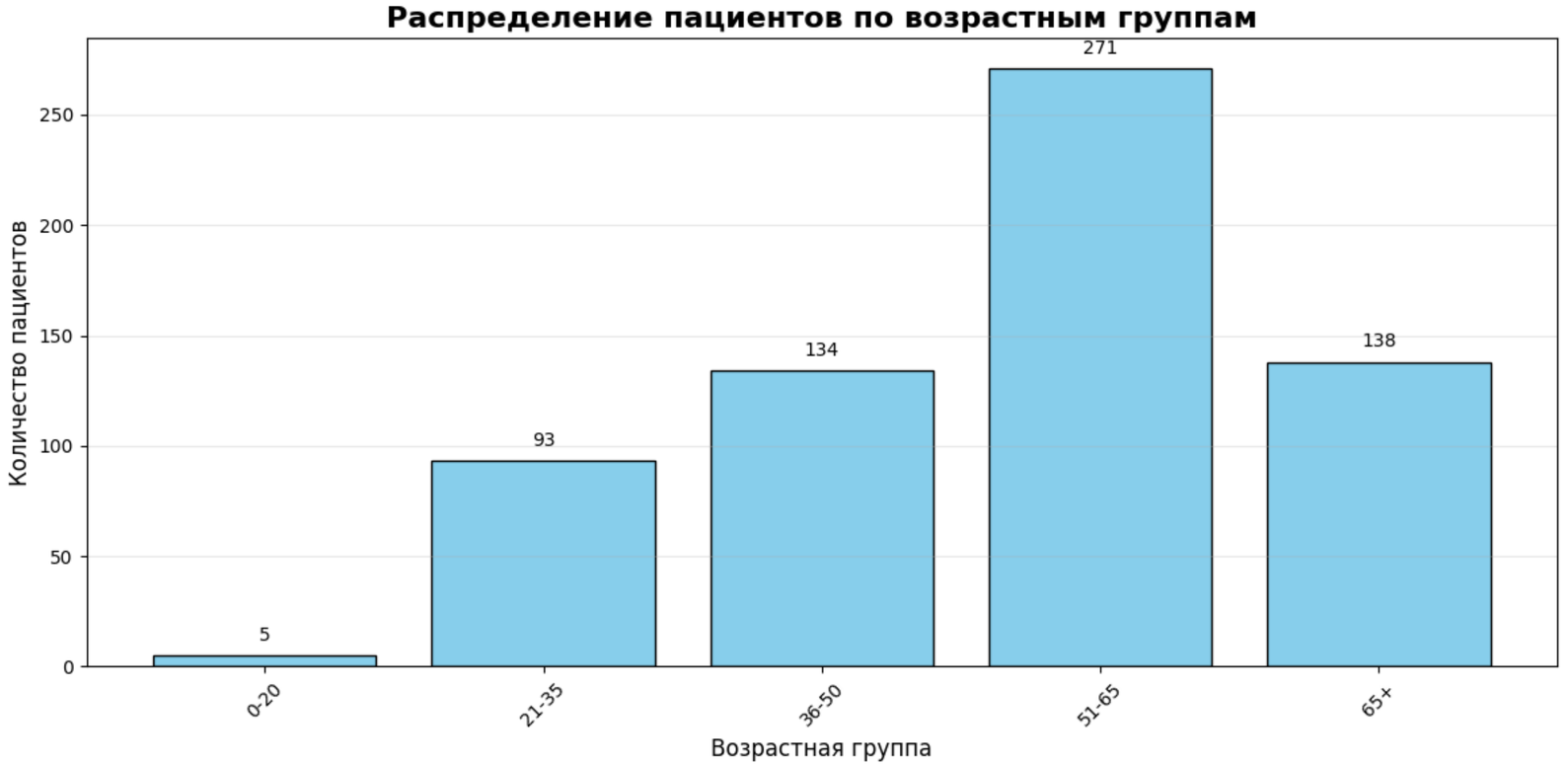
finding	diagnosis_extended
Tuberculosis	Other
Pneumonia/Fungal/...	Pneumonia
Pneumonia/Bacteri...	Bacterial Pneumonia
No Finding	Normal
Pneumonia	Pneumonia
Pneumonia/Viral/C...	COVID-19
Pneumonia	Pneumonia
Pneumonia	Pneumonia
Pneumonia/Lipoid	Pneumonia
Pneumonia/Bacteri...	Bacterial Pneumonia

Распредление диагнозов

Как видно по графику, COVID-19 и Пневмония - самые распространенные заболевания в датасете.

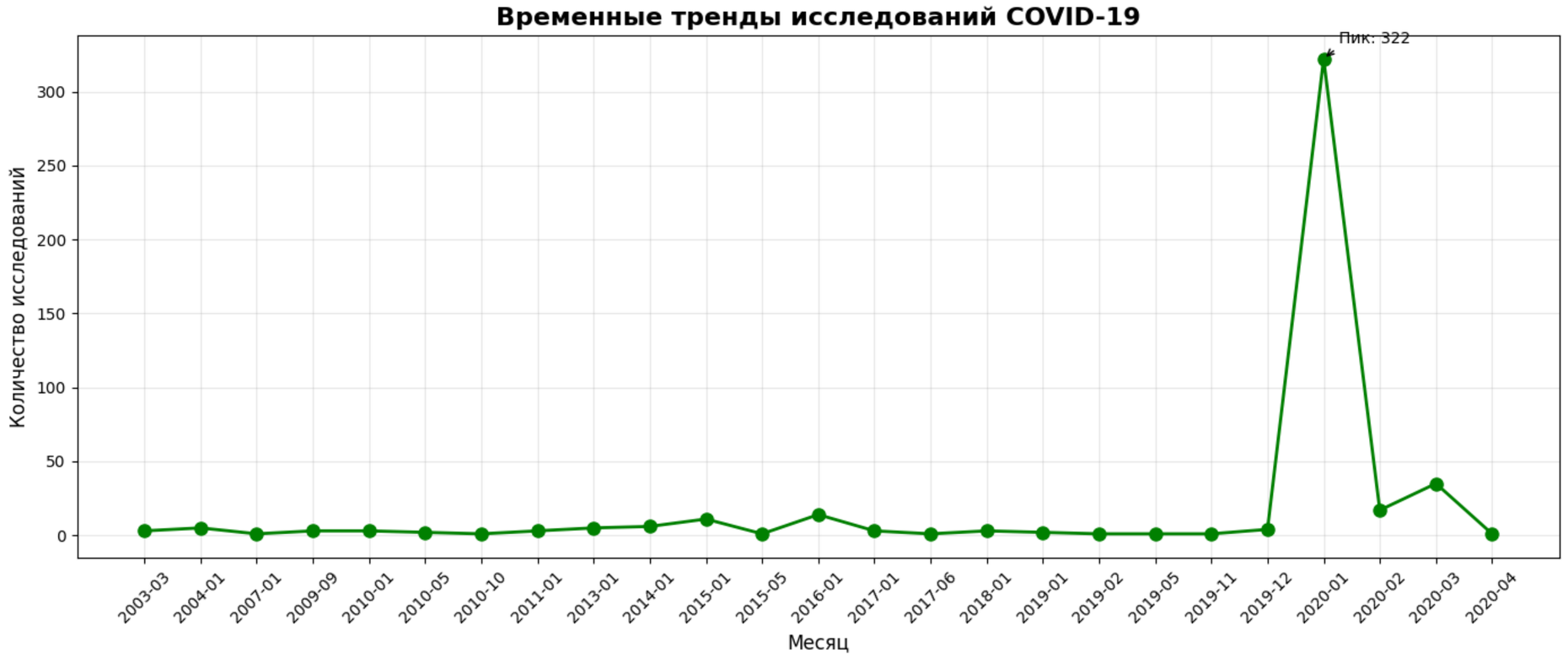


Распределение пациентов по возрастным группам



Наиболее многочисленная группа заболевших в возрасте от 51 до 65.

Тренд исследований



Видим явный скачок исследований в 2020 году.

Заключния и выводы

По итогу анализа, можно сделать несколько заключений:

1. Чаще всего диагностировались COVID-19 и Пневмония.

2. Чаще заболевали люди фозрастом от 51 о 60 лет.

3. Чаще заболевали мужчины.

4. Пик заболеваемости пришелся на 2020 год.

В процессе анализа также были выявлены основные проблемы данных:

- неоднородность формата дат,

- пропущенные значения возраста, дат и пола,

которые были успешно решены.