oki, so here is the output:

```
    n_counts  n_genes
```

min 502.00 99.00
5% 758.00 389.00
25% 2008.00 800.00
median 5141.00 1556.50
75% 11325.00 2408.75
95% 31603.75 4509.25
max 95035.00 7121.00

thing i am thibnking about, whz is it n_genes?i am lookin gat UMI-counts/cell, right? this does not really make any sense to me, can you please explain it to me before we proceed further? adn only then provide me the code for next steps. also, i checked mitochodrial genes, there are no, so it si good:
mitochondrial genes found: 0

Great question—this is a **critical concept** to understand before filtering. Let me clarify the relationship between **n_counts** and **n_genes**:

What Is UMI Count vs. Gene Count?

UMI Count (n_counts):

- **UMI = Unique Molecular Identifier**
- Each UMI represents **one RNA molecule** that was captured and sequenced.
- **n_counts = total number of RNA molecules detected in a cell.**
- Example: If a cell has 5,141 UMIs, it means ~5,141 individual RNA molecules were captured from that cell.
- **Proxy for:** How much mRNA was captured from the cell (cell size, transcriptional activity, capture efficiency).

Gene Count (n_genes):

- **n_genes = number of** *distinct genes* **that have at least 1 count in that cell.**
- Example: If n_genes = 1,556, it means that 1,556 different genes were detected (expressed) in that cell.
- **Proxy for:** Cell complexity, diversity of genes expressed, type of cell (some cell types express more/fewer genes).

Why Both Matter Together:

Imagine two cells:

- **Cell A:** 5,000 UMIs, 500 genes detected → **low complexity** (few genes expressed many times each—possibly a specialized cell or low-quality).
- **Cell B:** 5,000 UMIs, 2,000 genes detected → **high complexity** (many genes expressed once or twice—typical of a healthy, diverse cell).

Both have same UMI count, but very different biology!

How to Interpret Your Stats:

Metric	Your Data	What It Means
Median n_counts	**5,141 UMIs**	Typical cell captured ~5,000 RNA molecules
Median n_genes	**1,556 genes**	Typical cell expresses ~1,500 distinct genes
UMI range	502–95,035	Wide spread: some very sparse, some very complex
Gene range	99–7,121	Also wide, but less extreme than UMI

Why This Pattern Makes Sense:

Your **left spike in UMI histogram** (cells with <2,000 UMIs) but **more centered gene histogram** tells us:

- Many cells captured **few RNA molecules** (likely empty droplets or damaged cells).
- But **when they do express genes, they express many different ones** (good sign for real cells).
- **Cutoff strategy:** Remove cells with very low UMI counts (the spike), keep cells with reasonable UMI + gene diversity.

Recommended Filtering Thresholds (Based on Your Stats):

Looking at your quantiles:

Threshold	Reasoning
Min n_counts ≥ 1,000–1,500	Remove the left spike (empty droplets). Your 25% is 2,008; cutting at ~1,000 removes obvious debris.
Min n_genes ≥ 300–500	Remove cells with extreme low complexity. Your 5% is 389; cutting here removes outliers.
Max n_counts ≤ 50,000	Optional: remove potential doublets (two cells captured as one). Your 95% is 31,603; cutting at 50,000 is conservative.
Max n_genes ≤ 6,000	Optional: remove extreme complexity. Your 95% is 4,509; cutting at 6,000 is very permissive.

Before I Give Filtering Code, Decide:

Which thresholds feel right to you?

1. **Conservative:** Keep most cells (min_counts=1,000, min_genes=300).
2. **Moderate:** Balance quality vs. cell retention (min_counts=2,000, min_genes=500).
3. **Strict:** High-quality only (min_counts=3,000, min_genes=800).

For a **normal sample from your breast cancer cohort**, I'd suggest **moderate** as default. But let me know your preference, and I'll write the filtering code with full documentation.

What do you think?