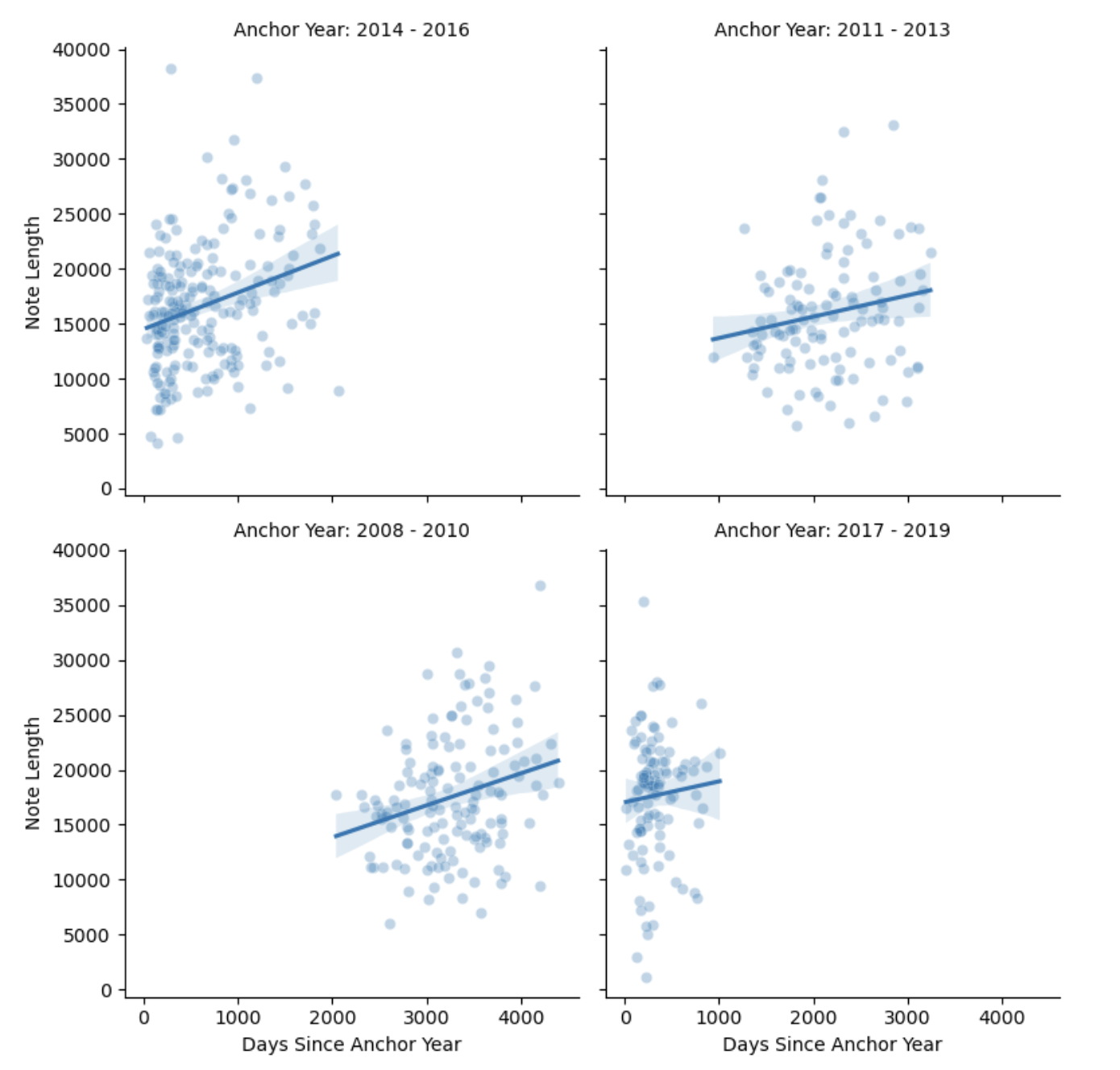
3.27 Meeting

Date since anchor year: for each patient with multiple notes, difference between chart time and anchor year

Notice: all notes were after 2014-2019



**Correlation between days since anchor year and note length**

1. Over all year groups:

Total Patients with multiple notes: 227

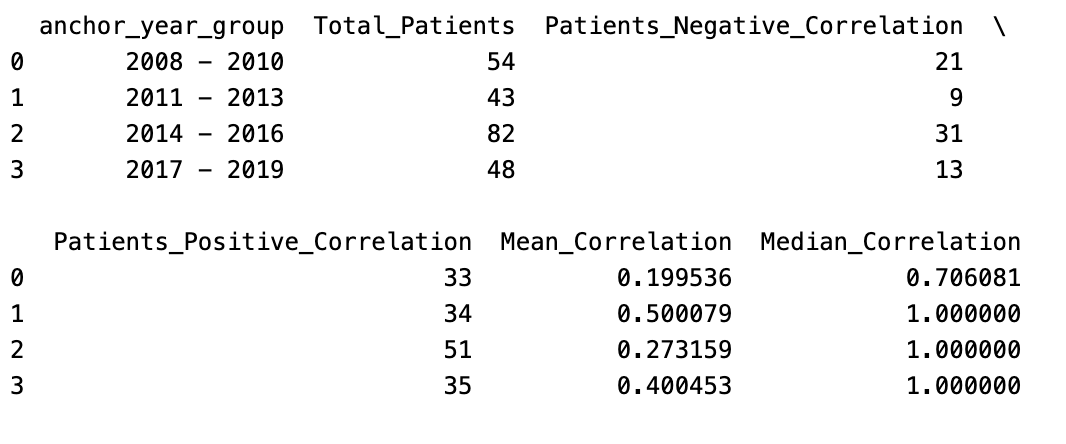
Patients with Negative Correlation: 74

Patients with Positive Correlation: 153

Mean Correlation: 0.3255

Median Correlation: 0.9692

1. Summarize for each year group:



Note length increase is not a interesting thing.

\*May indicate that later visits are more severe.

\*Potential probelem: copy forward across notes. Remove it when possible.

To Do:

1. Remove overlapping paragraph (length above a threshold of ~50 words). Count the number of unique words.
2. (Encode sentence to embeddings (Spacy). Count unique tokens. )
3. Check if we can use a LLM API for MIMIC data.
4. If not, try to put local version of LLMs on Discovery eg. llama, deekseek. Ask Discovery team if other groups have done that.
5. Share batch files and commands useful for Discovery.