**Conclusions & Insights**

This study applied a Retrieval-Augmented Generation (RAG) framework to legal documents, combining exploratory data analysis, retrieval, and evaluation. The dataset exhibited wide variation in document length, necessitating chunking strategies, and showed strong clustering in statutes and contracts while case law was more diverse. Word frequency analysis highlighted common legal terms such as *agreement, section, clause,* and *party*, reflecting the domain-specific nature of the corpus.

The pipeline demonstrated that preprocessing improves data quality, while TF-IDF retrieval works well for keyword-based queries but struggles with semantically phrased questions. Extractive generation produced relevant context snippets, though results lacked fluency and abstraction. Evaluation using ROUGE indicated moderate performance (F1 ≈ 0.3–0.5), with high accuracy for explicit references (e.g., “Section 1”) but weaker results for generic queries.

Overall, the RAG process proved effective in grounding answers in source documents and provided reliable outputs for structured legal queries.