**Clustering - Financial Documents**

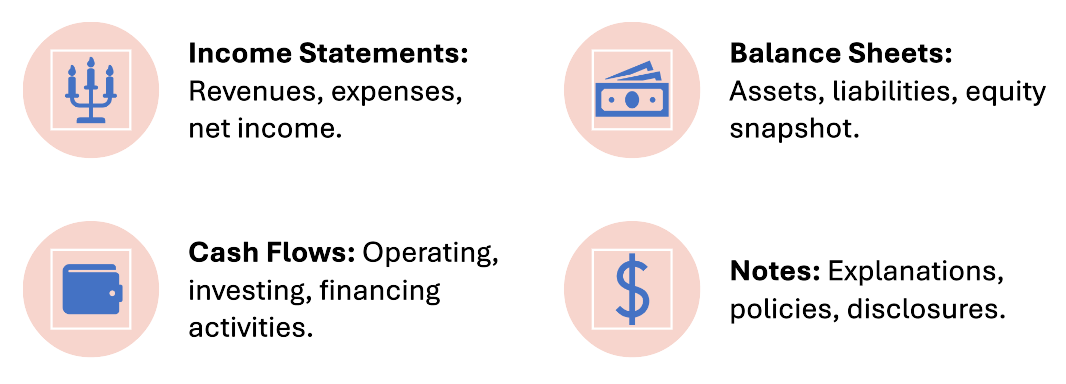
A clipboard with a pen and glasses on top of money

Description automatically generated

Public companies release financial reports to communicate their financial performance, position, and activities to stakeholders. These reports are rich in data, often structured across multiple tables, each serving a distinct purpose.

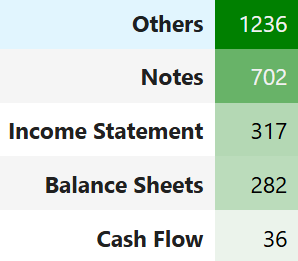
From an analytics perspective, clustering these tables into predefined categories helps streamline data extraction, comparison, and analysis, enabling more effective insights.

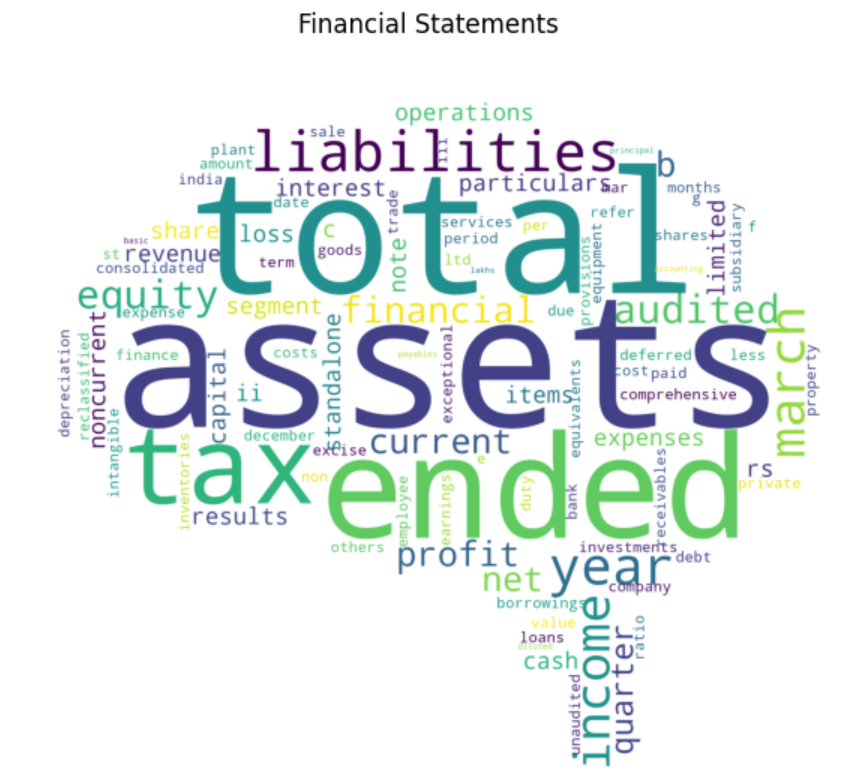
The primary categories include:



By clustering financial tables into these standardized categories, analytics professionals can efficiently identify patterns, detect anomalies, and generate actionable insights.

**Data:** Financial tables given in html format distributed as below with total files size **2573**

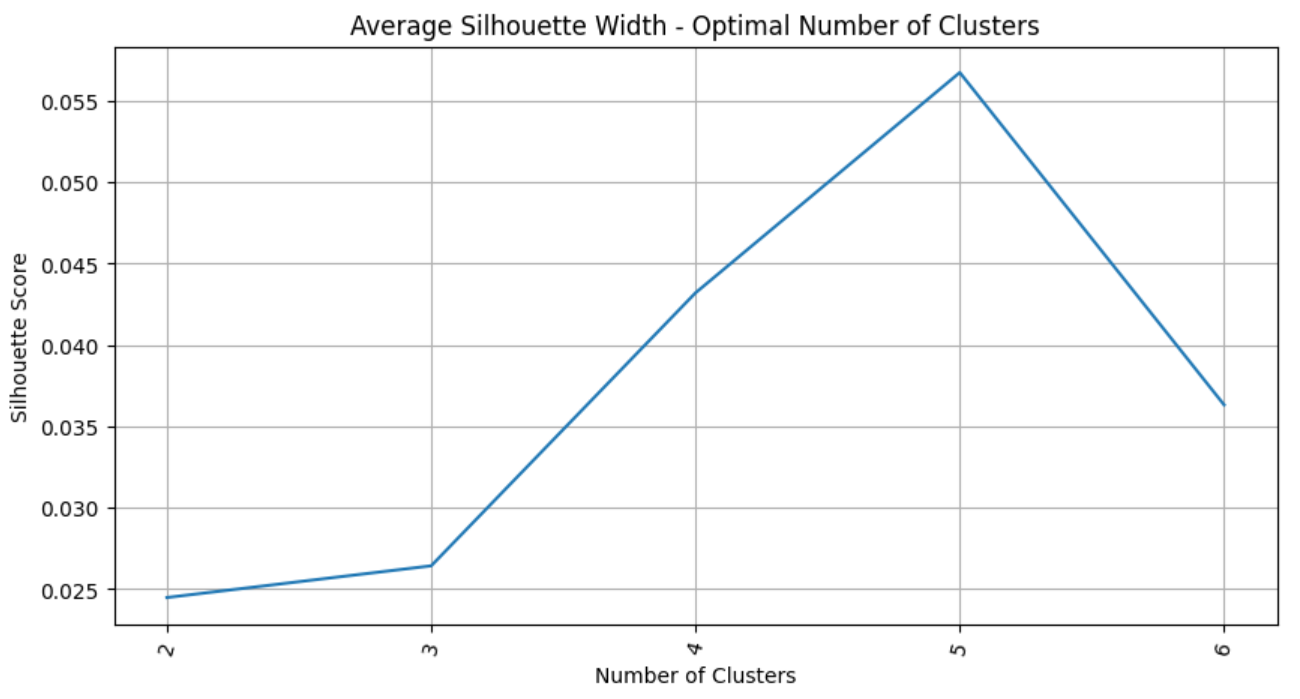
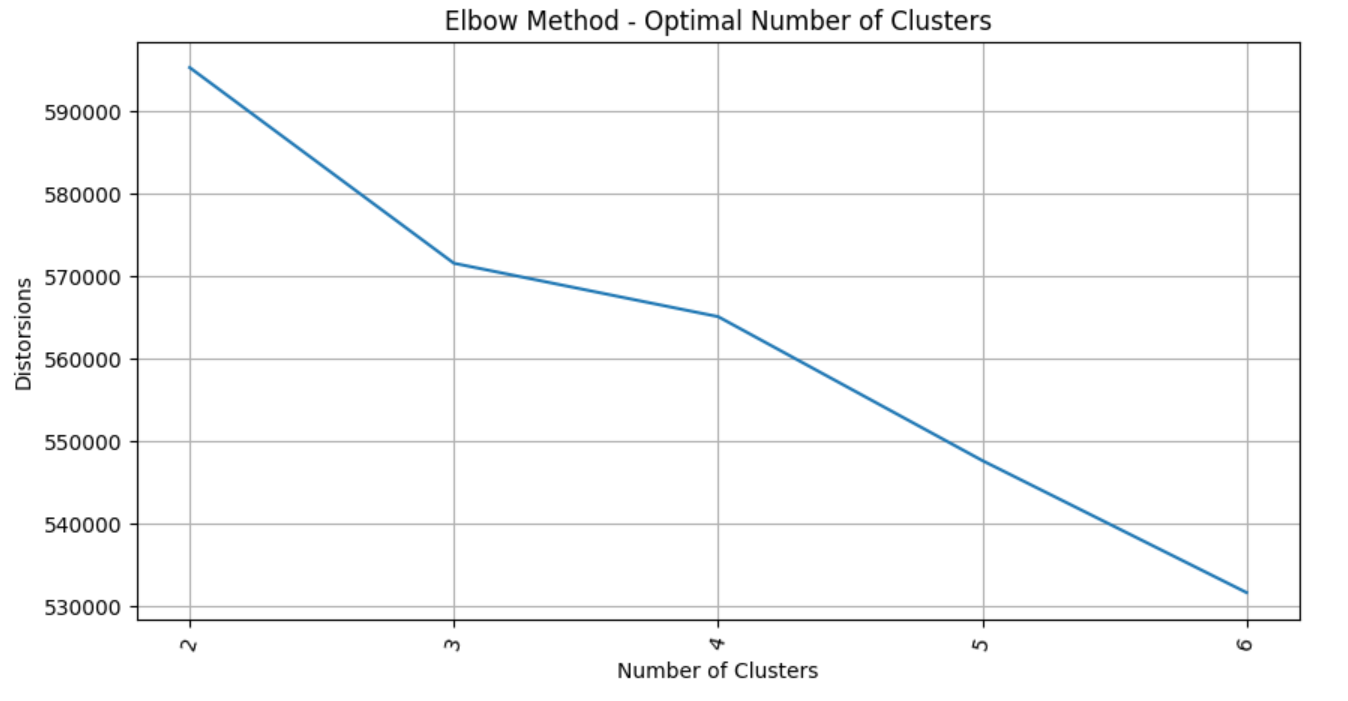




**Financial Table Clustering Workflow**

* **Extract Tables from HTML:** Import financial tables from HTML files and convert them into text strings.
* **Filter Key Metrics:** Retain only relevant financial metric names for analysis.
* **Text Preprocessing:** Clean and standardize text data using NLTK.
* **TF-IDF Matrix Creation:** Build a TF-IDF matrix to represent textual data numerically.
* **Dimensionality Reduction with NMF:** Decompose the sparse TF-IDF matrix using NMF to identify key features from each component.
* **Clustering with K-Means:** Apply the K-Means algorithm (K=5) to classify tables into five predefined categories.
* **Cluster Validation:** Evaluate clustering performance using the Elbow Method and Average Silhouette Width to determine optimal cluster numbers.

For above financial documents it was observed that Average Silhouette Width was optimal with 5 clusters.

The model can further be used for classifying new financial documents into labels returned by K-Means model.