To prepare the **companies** dataset for segmentation and clustering, key features were engineered to enhance data quality and interpretability. Companies were grouped by size and revenue, industry names and regions were standardized, and binary flags were created to indicate the use of specific technologies. A technology count variable was added to reflect digital maturity, along with a BPO indicator. Temporal features such as creation year and quarter were also extracted to support time-based analysis.

Feature engineering was also performed on the **tickets** dataset to support implementation tracking and performance analysis. Date fields such as creation and closure dates were converted for time-based calculations, including total implementation duration and days to first symptom presentation. A helper function was used to convert time-to-close values into hours for easier comparison. Ticket statuses were simplified into broader categories (e.g., "Ongoing", "Completed") to support clearer analysis. Training-related fields were consolidated into both count and percentage metrics to assess client onboarding progress. Lastly, the ticket creation date was used to extract year, month, and year-month values for temporal trend analysis.

The **deals** dataset was cleaned and enriched to enable time-series and value-based analysis. Column names were stripped of whitespace, and key date fields for deal creation and closure were converted to datetime format. From these, additional features were derived to capture the year, month, quarter, and a combined year-month string for trend grouping. To contextualize deal values, a custom categorization was applied based on deal size, labeling them as "Small," "Medium," "Large," or "Enterprise" depending on the total amount.

To ensure data quality and reduce noise, a cleaning process was applied to the **companies**, **tickets**, and **deals** datasets. Columns with more than 95% missing values were dropped to eliminate sparsity. Outliers were then identified using the interquartile range (IQR) method across all numeric fields, flagging records with values significantly outside the expected range. This helped surface anomalies that may require further review or transformation before modeling.

Define the relevant columns to keep for further analysis on these three datasets ( remove all the unnecessary columns).

Further refinement was performed on the **companies** dataset to ensure consistency and completeness for segmentation analysis. Records missing critical fields—industry, revenue, or employee count—were removed. The Country/Region field was standardized for case and spacing, and the Web Technologies field was cleaned and converted into structured lists. Additionally, fields related to parent or associated companies were dropped if present but largely empty, reducing unnecessary complexity in the dataset.

Additional processing was applied to the **tickets** dataset to improve feature usability. Time-related fields were converted from string format to total seconds, enabling precise duration analysis. Binary flags were created to indicate whether each of the five training modules had been completed. An Is\_Closed flag was added to mark completed tickets, and another flag captured whether a ticket was linked to an associated deal—supporting downstream analysis of implementation progress and deal connectivity.

To refine the **deals** dataset for analysis, records missing deal amounts were removed to ensure accuracy in value-based assessments. Outliers in the Days to close field were filtered using the IQR method to reduce skew from extreme cases. Additionally, deal probabilities were grouped into categorical bins—Low, Medium, High, and Unknown—providing a clearer view of deal confidence levels for segmentation and modeling.

To evaluate the linkage between support tickets and deals, deal identifiers were standardized as strings in both datasets. A match was then performed between the Associated Deal field in **tickets\_cleaned** and the Deal Name field in **deals\_cleaned**. This allowed identification of overlapping records and enabled further analysis of how many tickets were tied to active deals, revealing the proportion of implementation activity directly related to sales outcomes.

To create a unified view of support and sales activities, ticket-to-deal relationships were established using a provided JSON mapping file. Ticket IDs and deal names were cast to strings for consistent matching, and a new column was created in the **tickets\_cleaned** dataset to reflect the mapped deal names. An inner join was then performed between **tickets\_cleaned** and **deals\_cleaned** using the mapped deal name as the key. This merge enabled deeper analysis of deal-linked tickets, and a summary report was generated to confirm the number of matched records and unique links between tickets and deals.

Following the merge of tickets and deals, a quality assessment was conducted to evaluate its effectiveness. The report quantified how many tickets had mapped deals, how many were successfully merged, and how many unique ticket and deal identifiers were preserved. It also identified any duplicated ticket entries—indicating one-to-many relationships—and counted how many deals were linked to multiple tickets. When available, the number of merged deals marked as "Closed Won" was also summarized, offering early insight into sales success associated with support activities.

To build a unified dataset for segmentation and value analysis, the merged ticket-deal data was joined with the companies\_cleaned dataset using the associated company name as the linking key. After ensuring consistent formatting, a curated set of relevant features was selected, spanning company attributes, technology usage, deal details, ticket outcomes, and training progress. Rows with missing values were removed to ensure data completeness, and categorical variables were explicitly typed to support future modeling and clustering tasks.