Findings from EDA (Dataset: anonymized\_hubspot\_companies.csv)

1. Data Cleaning

The raw dataset consisted of 46 features and 19,851 company records. However, many features contained excessive missing values or zero entries, with some exceeding 80%. Columns such as Total Agents, Revenue Range, and Primary Company were removed during the cleaning phase to reduce noise and enhance data reliability. For the remaining columns, missing values were handled through imputation: categorical variables were filled using the mode, and numerical variables were filled using the median. Outlier treatment was then applied to all numerical features using the IQR method. As a result, the statistical distributions of key variables became more balanced and robust.

2. Feature Engineering and Time Dynamics

We created a new variable called duration, which represents the number of days between each company's creation date and its most recent update. This feature was engineered by converting both the Create Date and Last Modified Datecolumns into datetime format and calculating the time difference in days. The duration variable adds a valuable temporal dimension to the dataset and can be used to assess customer retention, lifecycle length, and engagement behavior in future analyses. The analysis of time-based trends revealed several noteworthy patterns. The Create Date field showed that most records were created after mid-2021. In contrast, the Last Modified Date was heavily concentrated on a few specific dates.

3. Analysis of Numerical Features

Distributions of key numerical variables demonstrated noticeable skewness. For instance, both Annual Revenue and Number of Employees were highly right-skewed, indicating that while most companies are small in size and revenue, a few large companies disproportionately inflate the range. The Number of Times Contacted variable showed multiple spikes around values like 6 and 15. The Year Founded variable suggested that most companies were established post-1980, with a smaller concentration of older firms. The new duration feature exhibited a relatively symmetrical distribution, centered around 400–450 days. Together, these results reveal important business characteristics related to company scale, lifecycle stage, and customer engagement patterns.

4. Analysis of Categorical Variables

Categorical feature analysis revealed strong imbalances across many fields. The ICP Fit Level variable showed that over 80% of companies were labeled as “Tier 2: Partial Fit”, suggesting a weak alignment with the company’s ideal customer profile. The Primary Industry and Primary Sub-Industry features were dominated by Manufacturing and Banking, respectively, indicating a sector concentration in the current customer base. Furthermore, the Country/Region variable revealed that around 78% of companies are located in the United States, showing that the dataset is primarily U.S.-centric.

5. Correlation

The correlation matrix provided a useful overview of the linear relationships between numerical variables. A moderately strong correlation (r = 0.50) was observed between Annual Revenue and Number of Employees, supporting the intuitive understanding that larger companies tend to generate more revenue. The duration variable also showed a moderate correlation (r = 0.36) with Number of Times Contacted, suggesting that records with longer lifespans are typically contacted more frequently—possibly due to ongoing follow-up or engagement strategies. No very strong correlations (r > 0.80) were identified, indicating that most variables are relatively independent of one another.

6. Conclusion

This EDA revealed several key characteristics of this dataset: the sample is primarily composed of small to mid-sized companies based in the United States, with most firms having relatively low revenue and employee counts, and a strong geographic concentration in the U.S. Customer engagement behavior varies significantly — some companies are contacted frequently, while others show minimal interaction. Categorical variables exhibit severe imbalance, particularly in fields such as ICP Fit Level, industry, and country. Additionally, moderate linear correlations exist between several key features.