Data Dictionary: Anonymized Hubspot Companies Dataset

Overview

This dataset contains anonymized company records from Hubspot CRM. All personally identifiable information (PII) and sensitive business information has been removed or masked. This data describes client and prospect companies, their attributes, and technology stacks.

Field Descriptions

Field Name	Data Type	Description	Values	Usage Notes
Record ID	Integer	Unique identifier for the company	Numeric ID	Primary key for identifying companies
Company name	String	Anonymized company name	Format: Company_[hash]	Masked to protect customer identity
Create Date	Date	Date when company record was created	YYYY-MM-DD format	Used for measuring customer lifecycle
Last Modified Date	Date	Date record was last updated	YYYY-MM-DD format	Important for data freshness assessment
Close Date	Date	Date of first deal closure	YYYY-MM-DD format	Marks transition to customer status
Industry	String	Company's primary industry sector	Various industry names	Essential for segmentation analysis
Industry group	String	Broader grouping of industries	Various group names	Higher-level categorization than Industry
Primary Industry	String	Alternative industry classification	Various industry names	May provide additional detail
Number of Employees	Integer	Company size by employee count	Numeric value	Critical for size-based segmentation
Annual Revenue	Float	Reported annual revenue	Numeric value	Important financial indicator
Country/Region	String	Company's primary country	Country name	Geographic analysis

Field Name	Data Type	Description	Values	Usage Notes
State/Region	String	Company's state or region	State/province/region name	Regional analysis in larger countries
BPO	String	Business Process Outsourcer status	'Yes', 'No'	Identifies BPO companies
BPO Program	String	BPO program participation status	'Yes', 'No', program name	Details on BPO relationship
BPO Program Tier	String	Level within BPO program	Tier name/level	Program stratification indicator
# of Agents Contracted	Float	Number of contracted agents	Numeric value	Scope of agent deployment
# of Agents Total	Float	Total agents at the company	Numeric value	Total addressable agents
CCaaS	String	Contact Center as a Service platform	Platform name	Indicates technology ecosystem
WFM	String	Workforce Management system	System name	Important integration point
LMS System	String	Learning Management System used	System name	Another integration point
Web Technologies	String	Web technology stack used	Semicolon-separated list	Tech stack intelligence
SSO Application	String	Single Sign-On application used	Application name	Security/integration detail
SSO Implemented?	String	SSO implementation status	'Yes', 'No'	Indicates auth integration status
SymTrain Product	String	Product version being used	Product name/version	Product adoption indicator
SymTrain Use Cases	Float	Number of use cases deployed	Numeric value	Depth of product usage
Contract End Date	Date	Date when current contract ends	YYYY-MM-DD format	Important for renewal planning
Parent Company	String	Anonymized parent company name	Format: Company_[hash]	Corporate hierarchy information
Associated Company	String	Anonymized related company	Format: Company_[hash]	Additional relationship information

Field Name	Data Type	Description	Values	Usage Notes
Time Zone	String	Company's primary timezone	Timezone name	Useful for service scheduling

Derived/Calculated Fields

These fields are typically added during analysis and not present in the raw data:

Field Name	Data Type	Description	Calculation Method
Industry_Standardized	String	Normalized industry name	Mapped from Industry with standardized names
Company_Size_Category	String	Size category based on employees	Defined ranges (Very Small, Small, Medium, Large, Enterprise)
Revenue_Category	String	Revenue tier category	Defined ranges (<$1M, $1M-$10M, $10M-$50M, etc.)
Region	String	Broader geographic region	Mapped from Country/Region
Uses_[Technology]	Integer	Flag for specific technology usage	Binary indicator (0/1) from Web Technologies
Technology_Count	Integer	Count of detected technologies	Count of technologies in Web Technologies
Is_BPO	Integer	Flag for BPO companies	Binary indicator (0/1) from BPO field
Create_Year	Integer	Year of record creation	Extracted from Create Date
Create_Month	Integer	Month of record creation	Extracted from Create Date
Create_Quarter	Integer	Quarter of record creation	Extracted from Create Date
Create_YearMonth	String	Year-Month for time-based analysis	Format: YYYY-MM

Relationships

This dataset can be linked to other datasets using:

- Record ID → links to deals dataset via CompanyToDeals mapping
- Record ID → links to tickets dataset via CompanyToTickets mapping
- Company name → may match Associated Company in other datasets

Usage Examples

Industry Distribution Analysis

```
# Count of companies by standardized industry
industry_counts = df['Industry_Standardized'].value_counts()

# Revenue by industry
industry_revenue = df.groupby('Industry_Standardized')['Annual
Revenue'].agg(['mean', 'median', 'sum'])
```

Size Distribution Analysis

```
# Count of companies by size category
size_distribution = df['Company_Size_Category'].value_counts()

# CrossTab of industry and size
industry_size_crosstab = pd.crosstab(df['Industry_Standardized'],
df['Company_Size_Category'])
```

Technology Adoption Analysis

```
# Calculate adoption rates for each technology
tech_columns = [col for col in df.columns if col.startswith('Uses_')]
tech_adoption = {col.replace('Uses_', ''): df[col].mean() * 100 for col in
tech_columns}

# Compare technology adoption by company size
tech_by_size = df.groupby('Company_Size_Category')[tech_columns].mean() *
100
```

Geographic Analysis

```
# Distribution of companies by region
region_distribution = df['Region'].value_counts()

# Average company size by region
size_by_region = df.groupby('Region')['Number of Employees'].mean()
```

Data Quality Notes

- Some records may have incomplete industry or size information
- Web Technologies field parsing requires careful handling of delimiter format

- BPO flags and related fields may be inconsistently populated
- Country/Region standardization may be necessary for global analysis
- Parent/child company relationships may be incomplete