Data Dictionary: Anonymized Hubspot Deals Dataset

Overview

This dataset contains anonymized deal records from Hubspot CRM. All personally identifiable information (PII) and sensitive business information has been removed or masked. This data describes sales opportunities/deals including their status, value, and progression through the sales pipeline.

Field Descriptions

Field Name	Data Type	Description	Values	Usage Notes
Record ID	Integer	Unique identifier for the deal	Numeric ID	Primary key for identifying deals
Amount	Float	The monetary value of the deal	Numeric value	Used for revenue analysis and deal size segmentation
Amount in company currency	Float	Deal value in company's base currency	Numeric value	Use for consistent monetary analysis across regions
Close Date	Date	Date when the deal was closed (won or lost)	YYYY-MM-DD format	Important for time-based analysis of sales cycles
Create Date	Date	Date when the deal was created	YYYY-MM-DD format	Used for measuring time-in-pipeline
Days to close	Integer	Number of days from creation to closure	Numeric value	Direct measure of sales cycle length
Deal Stage	String	Current stage in the sales pipeline	'Opportunity', 'BANT Deal', 'Deep Dive', 'In Trial', 'Negotiation', 'Contract Sent', 'Closed Won', 'Closed Lost'	Core field for pipeline analysis
Deal Type	String	Classification of deal type	'New', 'Renewal', 'Upsell', 'Cross-sell'	Important for segmenting analysis by deal nature

Field Name	Data Type	Description	Values	Usage Notes
Deal Score	Float	Calculated score indicating deal quality	0-100	Higher numbers indicate better qualified deals
Deal probability	Float	Estimated probability of winning	0.0-1.0	Used for forecasting and win likelihood analysis
Deal source attribution 2	String	Marketing/sales channel that sourced the deal	Various channel names	Useful for channel effectiveness analysis
Forecast amount	Float	Projected revenue from the deal	Numeric value	Used for revenue projections and forecasting
Forecast category	String	Classification for forecasting purposes	'Pipeline', 'Best Case', 'Commit', 'Closed'	Important for pipeline management
Forecast probability	Float	Modified probability for forecasting	0.0-1.0	May differ from standard probability based on sales judgment
Is Closed (numeric)	Integer	Flag indicating if deal is closed	0 or 1	1 means deal is closed (won or lost)
Is closed lost	Boolean	Flag indicating if deal was lost	True/False	Used for win/loss analysis
Is Closed Won	Boolean	Flag indicating if deal was won	True/False	Key field for win rate calculations
Is Deal Closed?	Boolean	Flag for deal closure status	True/False	Similar to "Is Closed (numeric)"
Is Open (numeric)	Integer	Flag indicating if deal is still open	0 or 1	1 means deal is still in progress
Last Activity Date	Date	Date of last recorded activity	YYYY-MM-DD format	Indicator of deal engagement level

Field Name	Data Type	Description	Values	Usage Notes
Last Modified Date	Date	Date record was last updated	YYYY-MM-DD format	Important for data freshness assessment
Pipeline	String	Sales pipeline the deal belongs to	Pipeline name	Some organizations have multiple pipelines
Total contract value	Float	Total value over contract lifetime	Numeric value	Important for multi-year contracts
Weighted amount	Float	Deal amount adjusted by probability	Numeric value	Amount × Probability
Deal Name	String	Anonymized name of the deal	Format: Deal_[hash]	Masked to protect client identity
Deal owner	String	Anonymized sales rep owner	Format: Rep_[hash]	Masked to protect employee identity
Associated Company (Primary)	String	Anonymized company name	Format: Company_[hash]	Masked to protect client identity

Derived/Calculated Fields

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

Field Name	Data Type	Description	Calculation Method
Create_Year	Integer	Year when deal was created	Extracted from Create Date
Create_Month	Integer	Month when deal was created	Extracted from Create Date
Create_Quarter	Integer	Quarter when deal was created	Extracted from Create Date
Close_Year	Integer	Year when deal was closed	Extracted from Close Date
Close_Month	Integer	Month when deal was closed	Extracted from Close Date
YearMonth	String	Year-Month for time-based analysis	Format: YYYY-MM

Field Name	Data Type	Description	Calculation Method
Deal_Size_Category	String	Categorization of deal size	Defined ranges (Small, Medium, Large, Enterprise)

Relationships

This dataset can be linked to other datasets using:

- Associated Company (Primary) → links to companies dataset
- Record ID → links to tickets dataset via CompanyToDeals mapping

Usage Examples

Win Rate Calculation

```
# Calculate overall win rate
win_rate = df['Is Closed Won'].sum() / (df['Is Closed Won'].sum() + df['Is
closed lost'].sum()) * 100

# Win rate by Deal Type
win_rate_by_type = df.groupby('Deal Type').apply(
    lambda x: x['Is Closed Won'].sum() / (x['Is Closed Won'].sum() + x['Is
closed lost'].sum()) * 100
)
```

Pipeline Velocity Analysis

```
# Average days in pipeline by stage
avg_days_by_stage = df.groupby('Deal Stage')['Days to close'].mean()

# Quarterly trend in sales cycle length
quarterly_trend = df.groupby('Create_Quarter')['Days to close'].mean()
```

Revenue Analysis

```
# Total revenue from won deals
total_revenue = df[df['Is Closed Won'] == True]['Amount'].sum()

# Average deal size by Deal Type
avg_deal_size = df.groupby('Deal Type')['Amount'].mean()
```

Data Quality Notes

- Missing values in Amount field (~22% of records) should be handled carefully in analyses
- Days to close may contain outliers that should be addressed for accurate cycle time analysis
- Deal probability values are subjective and may vary by sales representative methodology