Columns that might be important for analysis with large number of missing values, hard to impute:

* Parent Company: almost all rows have missing values
* CCaaS: technology ecosystem of the client, most are Five9, only one row of NiceCXOne, may not provide much insights into analysis
* Target Account: not in data dictionary, most rows are empty
* Total Agents: column is complete empty
* Close Date: almost all are missing
* # of agents total: almost all are missing
* **ICP Fit Level: ¾ of rows have missing values, claude says it’s likely an indicator for how well clients matches Symtrain’s defined ideal customer profile. Ideal fit/partial fit/long tail**
* Revenue Range: has no values at all
* # of agents contracted: almost all rows have missing values
* Primary Company: no data at all
* **ICP: not in data dictionary. Has a few “Yes” only, 1646 rows only out of 20K**
* Industry Group: no data
* Segmentation: most of rows are missing
* LMS System: almost no data
* Symtrain Use Cases: almost no data
* Associated Company: almost no data
* BPO: almost no data
* SSO application: almost no data
* Symtrain Product: no data
* Contract End Date: almost no data
* SSO Implemented: almost no dat
* Type: not in data dictionary looks like info on type of customer, only 1k out 20K avail.
* BPO Program: almost no data
* WFM: almost no data
* BPO Program Tier: almost no data

Columns with Good availability/potential for imputation:

* Annual Revenue
* Associated Contact: not in data dictionary, imputation to estimate contact for a deal may not be effective but has decent data availability
* Number of form submissions: not in data dictionary, but has good data availability
* Web Technologies: good data availability could be helpful to see link between platform used by client and success rate etc.
* Number of Time contacted: has valid values for little over half of the rows, could be an indicator for success/fail
* Contact with primary company: not in data dictionary but we can assume to be the company alias
* **Record ID: unique identifier for this dataset: links to other datasets for merges**
* Time Zone
* Primary industry: little over half of rows have valid values
* Number of pageviews: not in data dictionary but have good availability, Claude says it’s likely the amount of traffic/engagement the company has on the Symtrain website/platform
* Year Founded: good data availability could help analyze if younger or older companies prefer the service, could do feature engineering to calculate company age
* State/Region: good availability
* Consolidated Industry: good availability
* Number of employees: good availability could proxy for company size
* Primary Sub-Industry: around half of rows have data
* Number of Sessions: good availability, not in dictionary, could be the number of training sessions
* Country/Region: good avail.
* Industry: good avail.
* Create Date: good avail. Date company record created
* Company Name: good avail.
* Last Modified Date: good avail. Could use for feature engineering with Create Date to see case duration