Column-wise Cleaning & Imputation Notes

Dataset: Canada startups dataset (provided by hackathon organizers).

Goal: Impute missing values across 31 key columns using domain logic, group statistics and ML-based imputations. Preserve all data and document logic.

Key challenges: High missingness (up to ~99%), mixed formats (strings for numbers), and IPO-only fields.

Approach (summary):

- Profile missingness and unique value patterns across columns.
- Standardize formats: remove commas, % signs; parse dates.
- Impute low-missing columns using mode/median/grouped statistics.
- For high-missing but predictable columns, use ML-based imputations (RandomForest/LightGBM).
- Train IPO-only models for IPO-specific columns such as Valuation at IPO.
- Iteratively impute from low->high missing columns so earlier imputations become predictors.

Selected initial missing % highlights:

Valuation at IPO 99.79%, Actively Hiring 99.23%, Price 97.80%, Apptopia Downloads 95.20%

Number of Events 92.28%, Apptopia Apps 84.76%, IPqwery - Patents 76.01%, G2 Stack 65.92%

Number of Investors 64.89%, Total Funding 63.62%, Number of Articles 61.32%, SEMrush features ~59.43%

Column	Missing %	Method	Short explanation
Founded Date	0.21%	Extracted & parsed	Parsed to datetime; missing values filled by grouped median year (industry-based).
Number of Founders	48.68%	Median / RF	Filled with grouped median or predicted with RF where correlated features present.
Company Type	2.30%	Mode	Filled missing with most frequent company type or 'Unknown'.
Number of Employees	5.22%	RandomForestRegressor	Converted ranges to numeric midpoints; predicted missing using company size, revenue, funding.
Industries	3.24%	Mode / mapping	Cleaned tags and filled missing from similar companies by CB Rank or industry groups.
Headquarters Location	0.00%	Standardize / Mode	Standardized city/state strings; filled missing with most frequent location or 'Unknown'.
Headquarters Regions	53.45%	Mode / inference	Mapped from location; filled with most frequent region when missing.
Number of Investors	64.89%	LightGBM/RandomForest	Predicted using funding rounds, last funding amount, industry and CB Rank.
Actively Hiring	99.23%	Heuristic/Mode	Inferred from funding activity and growth signals; otherwise 'Unknown'.
Number of Funding Rounds	53.71%	Median / RF	Filled with median or predicted using funding history and company age.
Last Funding Amount	67.65%	LightGBM/RandomForest	Cleaned numeric strings, converted to float, predicted or filled by grouped median.
Funding Status	58.29%	RandomForestClassifier	Predicted stage/status using funding signals and company features.
Last Funding Type	53.71%	RandomForestClassifier	Predicted funding type from amount, stage and investors.
Estimated Revenue Range	60.99%	RandomForestClassifier	Predicted ordinal revenue bracket using employees, traffic and funding.
IPqwery - Patents Granted	76.01%	RandomForestRegressor	Converted strings to numbers; predicted using industry, CB Rank and size.
SEMrush - Monthly Visits	59.43%	LightGBM/RandomForest	Imputed using web metrics, revenue and employees.

Continued: Column-wise methods & explanations

Column	Missing %	Method	Short explanation
SEMrush - Visit Duration	59.43%	LightGBM/RandomForest	Predicted using related SEMrush features.
SEMrush - Page Views / Visit	59.43%	LightGBM/RandomForest	Imputed using traffic and engagement features.
SEMrush - Bounce Rate	59.43%	LightGBM/RandomForest	Converted percentages and predicted using visit duration and page views.
Number of Events	92.28%	RandomForestRegressor	Predicted using publicity and engagement features (articles, CB rank).
BuiltWith - Active Tech Count	17.02%	Median / RF	Filled with median per industry, predicted for tech-heavy firms.
G2 Stack - Total Products Active	65.92%	RandomForestRegressor	Predicted for product companies using industry, web presence and funding.
Number of Articles	61.32%	RandomForestRegressor	Predicted visibility using funding, CB Rank and events.
CB Rank (Company)	0.03%	Grouped median	Converted to numeric and filled by grouped median per industry.
Total Funding Amount	63.62%	LightGBM/RandomForest	Cleaned and converted to numeric; predicted using funding history and investor counts.
Valuation at IPO	99.79%	IPO-only RandomForest	Model trained only on IPO companies then used to predict IPO valuations.
Price	97.80%	RandomForestRegressor	Cleaned numeric values then predicted using funding and valuation signals.
Number of Exits	Low/Var	RandomForestRegressor	Predicted using investor activity, funding and company age.
Industry Groups	3.24%	Mode / mapping	Mapped from Industries or filled with most frequent industry group.
Apptopia - Downloads Last 30 Days	95.20%	RandomForestRegressor	Converted to numeric and predicted using app and traffic features.
Apptopia - Number of Apps	84.76%	RandomForestRegressor	Predicted using downloads and company category.

Note: ML models used ensemble trees (RandomForest/LightGBM) with OrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=-1) for categorical features. Numeric cleaning removed commas and %-characters where applicable.