**CANADA**

**Missing values (%):**

Valuation at IPO 99.785282

Actively Hiring 99.232221

Investment Stage 98.880864

Investor Type 98.100072

Price Currency (in USD) 97.800768

Apptopia - Downloads Last 30 Days 95.204633

Number of Events 92.283167

Apptopia - Number of Apps 84.761533

IPqwery - Patents Granted 76.010150

Last Funding Amount Currency (in USD) 67.649164

G2 Stack - Total Products Active 65.918407

Number of Investors 64.890364

Total Funding Amount Currency (in USD) 63.621576

Number of Articles 61.318238

Estimated Revenue Range 60.992908

SEMrush - Monthly Visits 59.431323

SEMrush - Page Views / Visit 59.431323

SEMrush - Global Traffic Rank 59.431323

SEMrush - Visit Duration 59.431323

SEMrush - Bounce Rate 59.431323

Funding Status 58.286160

Number of Funding Rounds 53.705511

Last Funding Type 53.705511

Headquarters Regions 53.445247

Number of Founders 48.675906

BuiltWith - Active Tech Count 17.021277

Number of Employees 5.224803

Industries 3.240289

Company Type 2.296831

Founded Date 0.208211

CB Rank (Company) 0.026026

Headquarters Location 0.000000

| **Column** | **Type** | **% Missing** | **Strategy** | **Logic** |
| --- | --- | --- | --- | --- |
| Valuation at IPO | Num | 99.78% | Drop column | Too sparse to impute meaningfully |
| Actively Hiring | Num | 99.23% | Infer from other signals like funding, size, growth metrics | Funding activity can indicate active hiring |
| Investment Stage | Cat | 98.88% | Use related info: Funding type, amount, age of startup | Early stage vs Late stage inference |
| Investor Type | Cat | 98.10% | Use same as above | Based on type and round of investments |
| Price Currency (in USD) | Num | 97.80% | Drop or use mode if only 1 country | Irrelevant for single-country (Canada) |
| Apptopia - Downloads Last 30 Days | Num | 95.20% | Infer from Monthly Visits or Apps count | Proxy via engagement |
| Number of Events | Num | 92.28% | Median imputation by industry/region | Tech/startup events often align by sector |
| Apptopia - Number of Apps | Num | 84.76% | Fill 0 for non-tech sectors; median/grouped imputation for others | Only tech orgs have apps |
| IPqwery - Patents Granted | Num | 76.01% | Fill 0 for non-R&D sectors; median/grouped imputation for others | Not all orgs file patents |
| G2 Stack - Total Products Active | Num | 65.91% | Fill 0 or median by industry | Only SaaS/product-based companies |
| Estimated Revenue Range | Cat | 60.99% | Predictive modeling | Related to funding, visitors, years |
| Number of Investors | Num | 64.89% | Grouped median by stage/industry | Funding/rounds correlate |
| Total Funding Amount Currency (in USD) | Num | 63.62% | Predictive modeling | Based on all funding and traffic metrics |
| Funding Status | Cat | 58.28% | Mode imputation by stage/type | Usually aligns with round/status |
| Last Funding Amount Currency (in USD) | Num | 67.65% | Predictive modeling or group median | Same as above |
| Last Funding Type | Cat | 53.70% | Use stage/type/revenue to infer |  |
| Number of Founders | Num | 48.68% | Mode or median by company type | Solo vs team startups |
| Company Type | Cat | 2.30% | Mode | Low missing |
| Number of Employees | Num | 5.22% | Predict from revenue + traffic | Clear relation to size |
| Industries | Cat | 3.24% | Mode | Minor missing |
| Others (~15 columns) | Mostly under 20% | Use median/mode, depending on type | Justified per column context |  |

| **Column Name** | **Method** |
| --- | --- |
| Founded Date / Founded Year | Extracted & filled from text |
| Headquarters Regions | Mode (most frequent value) |
| Company Type | Mode |
| Number of Employees | Mode |
| Estimated Revenue Range | Random Forest Classifier |
| SEMrush - Monthly Visits | LightGBM Regression |
| SEMrush - Visit Duration | LightGBM Regression |
| SEMrush - Page Views / Visit | LightGBM Regression |
| SEMrush - Bounce Rate | LightGBM Regression |
| SEMrush - Global Traffic Rank | LightGBM Regression |
| Funding Status | Random Forest Classifier |
| Last Funding Type | Random Forest Classifier |
| Last Funding Amount Currency (in USD) | LightGBM Regression |
| Total Funding Amount Currency (in USD) | LightGBM Regression |
| Number of Investors | LightGBM Regression |
| Number of Lead Investors | LightGBM Regression |
| Column | Method |
| CB Rank (Company) | Grouped median |
| Founded Date | Grouped median |
| Company Type | Grouped mode |
| Industries | Grouped mode |
| Number of Employees | RandomForestRegressor |
| Estimated Revenue Range | RandomForestClassifier |
|  |  |

| **Column Name** | **Data Type** | **Missing % (Approx)** | **Cleaning / Imputation Logic** |
| --- | --- | --- | --- |
| **Founded Date** | Date | Low | Converted to datetime format, filled missing with median year per industry. |
| **Number of Founders** | Numeric | Low | Filled missing with median per industry group. |
| **Company Type** | Categorical | Low | Filled missing with 'Unknown'. |
| **Number of Employees** | Numeric (range converted) | Low | Converted ranges to numeric midpoints, filled missing with median. |
| **Industries** | Categorical | Low | Filled missing with 'Unknown'. |
| **Headquarters Location** | Categorical | Low | Filled missing with 'Unknown'. |
| **Headquarters Regions** | Categorical | Low | Filled missing with 'Unknown'. |
| **Number of Investors** | Numeric | ~65% | Predicted using RandomForestRegressor with funding, employees, industry features. |
| **Actively Hiring** | Categorical (Y/N) | ~99% | Filled missing with 'Unknown' (no reliable prediction possible). |
| **Number of Funding Rounds** | Numeric | Low | Filled missing with median per industry. |
| **Last Funding Amount** | Numeric | Low | Converted to numeric, filled missing with median per industry. |
| **Funding Status** | Categorical | Low | Filled missing with 'Unknown'. |
| **Last Funding Type** | Categorical | Low | Filled missing with 'Unknown'. |
| **Estimated Revenue Range** | Categorical | Low | Filled missing with 'Unknown'. |
| **IPqwery - Patents Granted** | Numeric | ~76% | Converted to numeric, predicted using RandomForestRegressor with related features. |
| **SEMrush - Monthly Visits** | Numeric | Low | Converted to numeric, filled missing with median per industry. |
| **SEMrush - Visit Duration** | Numeric | Low | Converted time strings to seconds, filled missing with median. |
| **SEMrush - Page Views / Visit** | Numeric | Low | Converted to float, filled missing with median. |
| **SEMrush - Bounce Rate** | Numeric (percentage) | Low | Converted to float %, filled missing with median. |

| **Column Name** | **Data Type** | **Missing % (Approx)** | **Cleaning / Imputation Logic** |
| --- | --- | --- | --- |
| **Number of Events** | Numeric | ~92% | Predicted using RandomForestRegressor with industry, funding, and activity features. |
| **BuiltWith - Active Tech Count** | Numeric | Low | Converted to numeric, filled missing with median per industry. |
| **G2 Stack - Total Products Active** | Numeric | ~65.9% | Converted to numeric, predicted using RandomForestRegressor with related features. |
| **Number of Articles** | Numeric | ~61% | Predicted using RandomForestRegressor with funding, industry, and company rank features. |
| **CB Rank (Company)** | Numeric | Low | Converted to numeric, filled missing with median per industry. |
| **Total Funding Amount** | Numeric | Low | Converted to numeric, filled missing with median per industry. |
| **Valuation at IPO** | Numeric | ~87% | Restricted to IPO companies, converted to numeric, predicted using RandomForestRegressor. |
| **Price** | Numeric | ~97.8% | Converted to numeric, predicted using RandomForestRegressor with funding & industry features. |
| **Number of Exits** | Numeric | ~98% | Predicted using RandomForestRegressor with industry, investors, and rank features. |
| **Industry Groups** | Categorical | Low | Filled missing with 'Unknown'. |
| **Apptopia - Downloads Last 30 Days** | Numeric | Low | Converted to numeric, filled missing with median. |
| **Apptopia - Number of Apps** | Numeric | Low | Converted to numeric, filled missing with median. |