**Data Familiarization**

**1. Dataset Overview**

* **Shape**: 187 rows × 5 columns
* **Columns**:
  1. opportunity\_id – String, looks like a unique ID (e.g., Opportunity#000000000G127E8VYE08TXBT6X).
  2. opportunity\_name – String, descriptive name of the opportunity.
  3. category – String (categorical: Event, Competition, Internship, Career, Course).
  4. opportunity\_code – String, alphanumeric short code (e.g., E501873).
  5. tracking\_questions – String (JSON-like format in some rows), **69 missing values**.

**2. Data Types Check**

| **Column** | **Detected Type** | **Expected Type** | **Notes** |
| --- | --- | --- | --- |
| opportunity\_id | object | string | Looks correct, unique ID. |
| opportunity\_name | object | string | Correct, descriptive text. |
| category | object | string | Correct, categorical. |
| opportunity\_code | object | string | Correct, short identifier. |
| tracking\_questions | object | string/JSON | Mostly text in JSON format, but many missing. |

✅ Data types appear correct — only tracking\_questions might need parsing if used.

**3. Missing Values**

| **Column** | **Missing Count** | **Missing %** |
| --- | --- | --- |
| opportunity\_id | 0 | 0% |
| opportunity\_name | 0 | 0% |
| category | 0 | 0% |
| opportunity\_code | 0 | 0% |
| tracking\_questions | 69 | 36.9% |

⚠ **tracking\_questions** is missing in over **⅓ of records**.

**4. Duplicate Check**

* **Duplicate rows (full match)**: **0** — good.

**5. Categorical Column Insights**

**category** (5 unique values):

| **Value** | **Count** | **%** |
| --- | --- | --- |
| Event | 78 | 41.7% |
| Internship | 48 | 25.7% |
| Competition | 39 | 20.9% |
| Career | 15 | 8.0% |
| Course | 7 | 3.7% |

**opportunity\_code** appears unique per row (no repeats found).

**6. Numeric Columns**

* None detected — no numerical statistical summary.

**7. Potential Data Quality Issues**

* Missing values in tracking\_questions (36.9%).
* tracking\_questions is stored as text containing JSON — should be parsed if analysis requires its details.
* No timestamp/date column — limits time-series analysis.
* No metadata about data source or refresh frequency.