

INTERNATIONAL STUDENT APPLICATIONS

DATASET DOCUMENTATION & EXPLORATORY DATA ANALYSIS

DEPAUL UNIVERSITY
FEBRUARY 2026

DATA VISUALIZATION TRAINEE
[TEAM 10]

7,543 Records | 80 Variables

DATASET OVERVIEW

1. Dataset Dimensions

The DePaul International Student Applications dataset comprises 7,543 records across 80 variables, representing comprehensive application and outreach data for prospective international students applying to DePaul University.

2. Dataset Themes and Content Areas

The dataset integrates multiple operational domains within the international student recruitment and admissions pipeline:

Core Thematic Areas:

- **Outreach Tracking** (15 variables): Contact attempts, caller information, outcomes, remarks, and follow-up tracking.
- **Application Processing** (10 variables): Application receipt, processing timestamps, counselor assignments, checklist items, decision statuses.
- **Campaign Management** (9 variables): Campaign IDs, workflow types, templates, attempt counts, escalation flags.
- **Address and Location** (7 variables): Multi-line street addresses, city, region, postal codes.
- **Academic Program Selection** (6 variables): Intended programs, intake terms, degree types, college choices.
- **System Metadata** (6 variables): Record creation and modification timestamps across merged data sources.
- **Contact Information** (3 variables): Primary/secondary phone numbers and email addresses.

- **Personal Information** (3 variables): Student names and date of birth.
- **Academic Background** (3 variables): Previous institution, major, and degree type.
- **Post-Admission Activities** (3 variables): University email creation, housing contracts, RIT email status.
- **Target Institution** (3 variables): University fields across multiple data sources.
- **Admission Decision** (3 variables): Admission dates, decision statuses, admission flags.
- **Identification** (2 variables): Primary and duplicate reference IDs.
- **Demographics** (2 variables): Country of origin and citizenship status.
- **Immigration Compliance** (2 variables): SEVIS ID and I-901 fee payment status.
- **Program-Specific Fields** (2 variables): SLU program start comments and city.
- **Location** (1 variable): City and branch designation.

3. Application Source Distribution

The dataset reveals a dual-channel recruitment approach:

- **Primary Source:** "STUDY GROUP" - Partnership-driven recruitment channel.
- **Secondary Source:** "Non-Study Group" - Direct applications.
- All records target DePaul University exclusively (single institution dataset).
- Applications span multiple intake periods including Fall 2024 and Fall 2025 terms.

4. Geographic Distribution

The dataset reflects substantial international student diversity:

- 95 unique countries represented across the applicant pool.
- 11 missing country values.

5. Address Data Completeness

- City: 7,510 records- 33 missing, requires validation for inappropriate values.
- Region: 6,629 records- 914 missing.
- Postal: 7,317 records-226 missing, requires format validation.
- Street_1: 7,528 records-15 missing, requires validation.
- Street_2: 2,904 records- 4,639 missing (optional field), check for comma usage.

DATA QUALITY ASSESSMENT

1. Data Completeness Analysis

A significant data quality concern is the prevalence of missing information across the dataset:

- 38 variables contain 100% missing data, rendering them non-functional for analysis
- 7 variables have complete data with no missing values (Reference_ID, Last_Name, University, Received_At, Created_At, Modified_At)
- 35 variables exhibit partial missingness.

1.1 Missing Value Analysis

Missing values are a major data quality consideration.

- 10 variables contain 100% missing values (7,543 missing entries each). Examples include Is_Admitted, Is_Global_Grad, SEVIS_ID, Housing_Contract, I_901_Status, and several email- or system-related fields. These columns currently provide no usable information.
- Many variables show partial missingness, especially administrative, outcome-related, and optional tracking fields.
- Only 7 variables are fully complete (0% missing values), including Reference_ID, Last_Name, University, Created_At, Modified_At, Received_At, and Application_Source.
- Some near-complete columns still contain a small number of missing values. For example: Intake has 3 missing values, Given_Name has 2 missing values, Email_ID has 1 missing value.

Overall, missing values are concentrated in non-core fields, but the high number of sparsely populated columns will require careful feature selection.

2. Data Source Integration Issues

The dataset exhibits evidence of multiple data source integration, particularly visible in System Fields category with duplicated timestamp variables (Created_At, Created_At-2, Created_At-3; Modified_At, Modified_At-2, Modified_At-3), suggesting merging from at least three separate operational systems.

3. Data Quality Flags and Validation Requirements

3.1 Field Naming Issues

Spelling Errors:

- Recieved_At (should be "Received_At") - No missing values.
- Counsler (should be "Counselor") - Column is completely empty.

3.2 Text Field Quality Checks Required

The following fields require validation to identify and remediate inappropriate placeholder values:

- **Given_Name:** Missing values present; check for inappropriate placeholder values (?, -, .)
- **Last_Name:** Check for inappropriate placeholder values (?, -, .)
- **Country:** Missing values present; check for spelling inconsistencies (e.g., "United states" vs "United States").
- **College:** Check formatting consistency (should include country, university name, and degree type).

3.3 Address Field Validation Requirements

- **Street_1:** 15 records missing; validate for inappropriate values.
- **Street_2:** 4,639 records missing; check for comma usage and formatting issues.
- **City:** 33 records missing; validate for inappropriate values such as embedded numbers.
- **Postal:** 226 records missing; requires format validation and consistency checks.

EXPLORATORY DATA ANALYSIS (EDA)

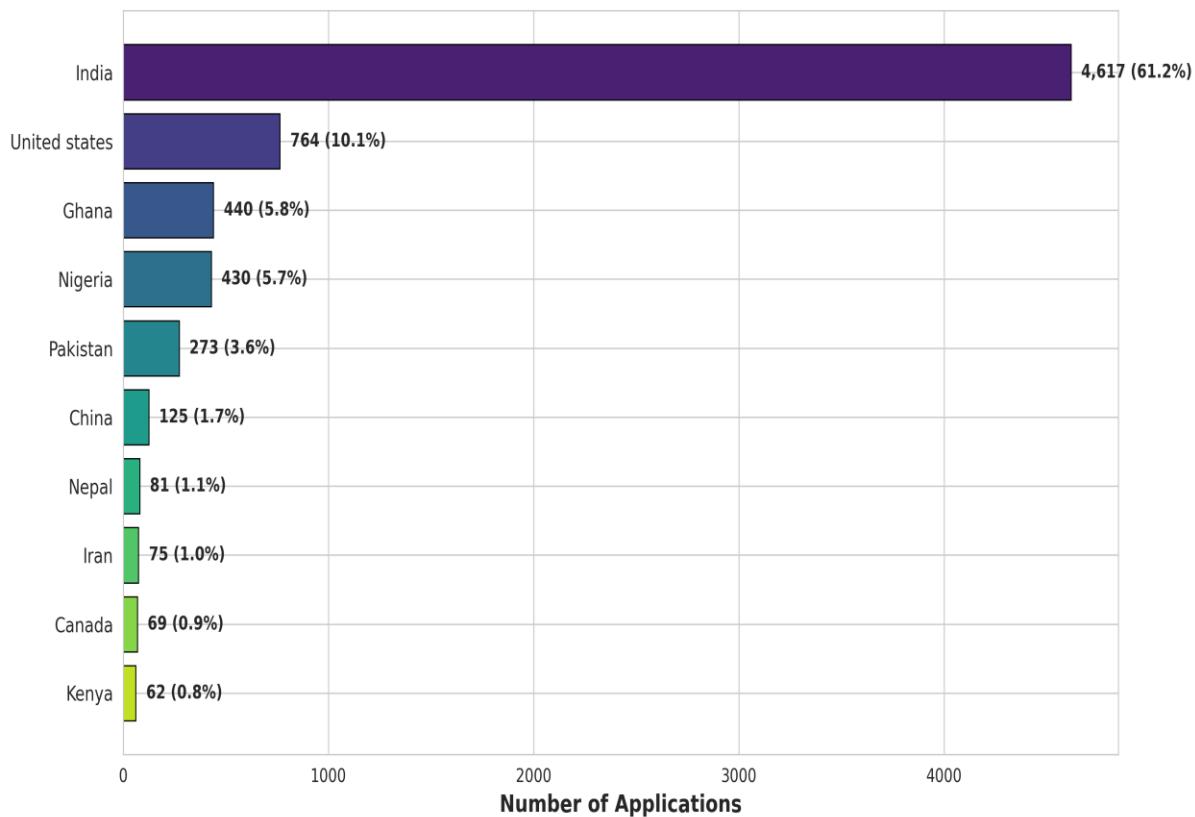
Notable patterns

- **Top 10 Countries by Application Volume**

The geographic distribution shows clear patterns in where applications originate.

India is the primary source country with 4,617 applications, representing 61.2% of the total dataset. This is followed by the United States (764 applications), Ghana (440 applications), Nigeria (430 applications), and Pakistan (273 applications).

**Figure 1: Top 10 Countries of Origin
(India as the Dominant Market Source)**

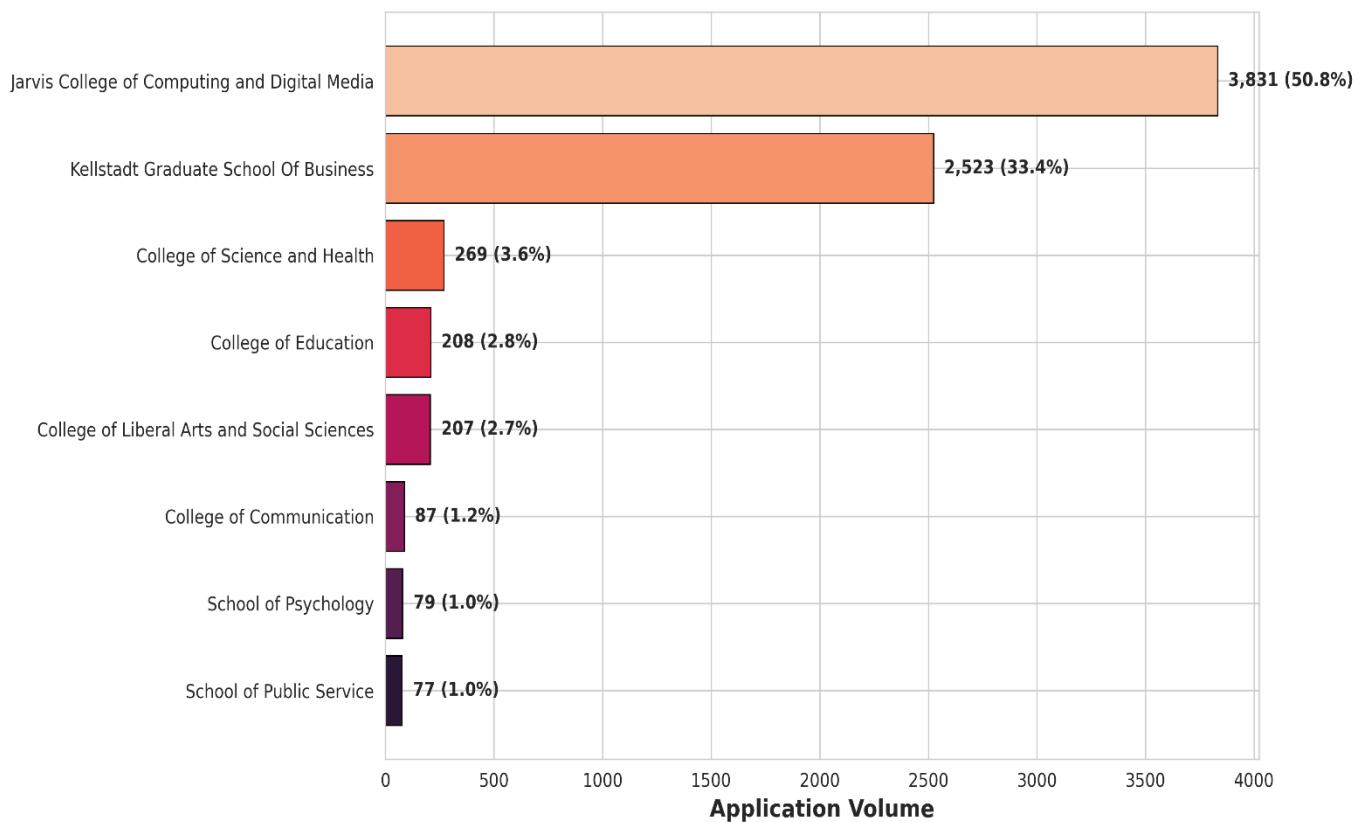


- **Top College Choices by Application Volume**

The distribution of first-choice college selections reveals clear preferences among international applicants regarding academic programs at DePaul.

Jarvis College of Computing and Digital Media dominates international student interest with 3,831 applications (50.8% of total), followed by Kellstadt Graduate School of Business (2,523 applications, 33.5%) and the College of Science and Health (269 applications, 3.6%).

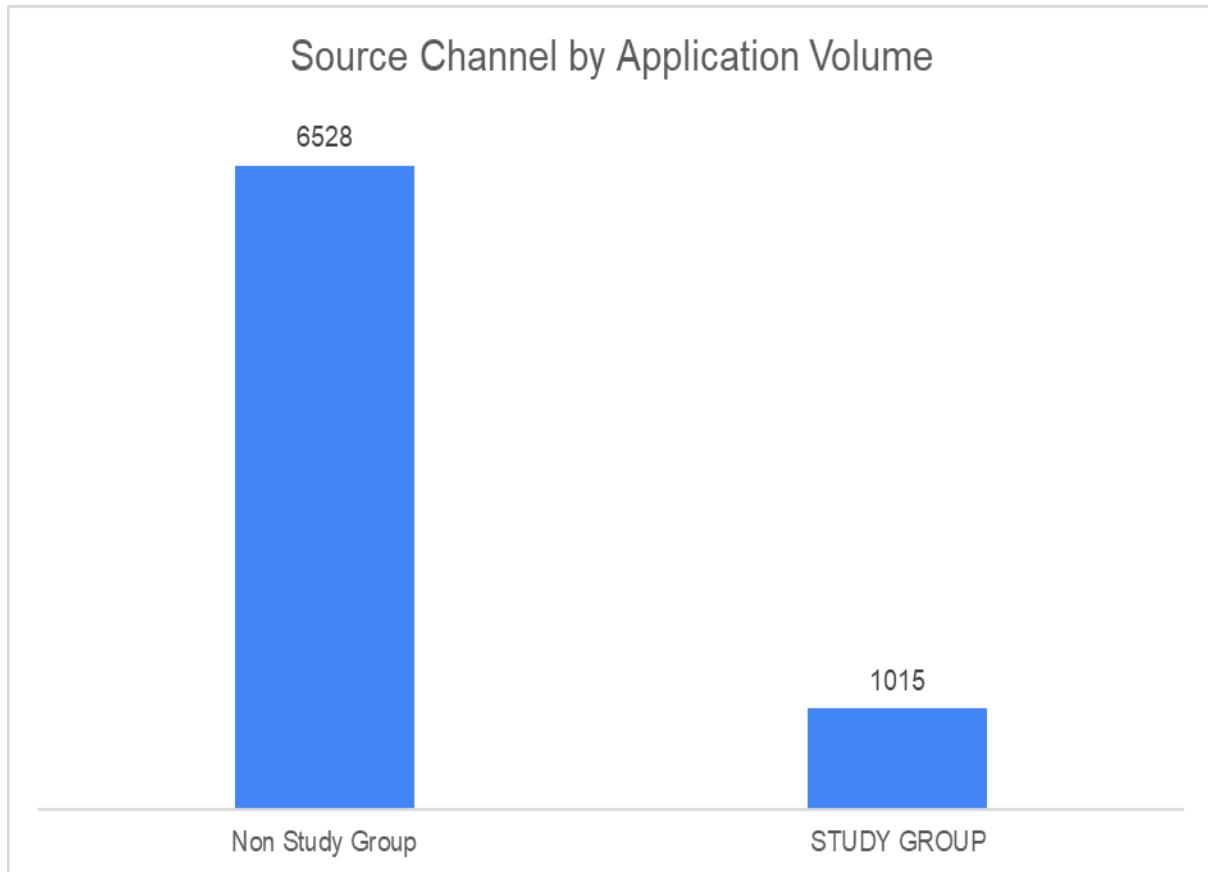
**Figure 2: Application Volume by College Choice
(Jarvis and Kellstadt as Lead Units)**



- **Application Source Channel Distribution**

The distribution of application sources reveals the recruitment channels through which international students discover and apply to DePaul.

Non-Study Group applications dominate the dataset with 6,528 applications (86.5% of total), while Study Group partnership applications account for 1,015 applications (13.5%).



- **The Application Funnel Distribution**

The analysis of Outcome_1 provides critical insight into the current state of the admissions pipeline and identifies operational bottlenecks. Among the 4,349 records with outreach data (57.7% of total):

Top Outreach Outcomes:

Follow up: 1,375 instances

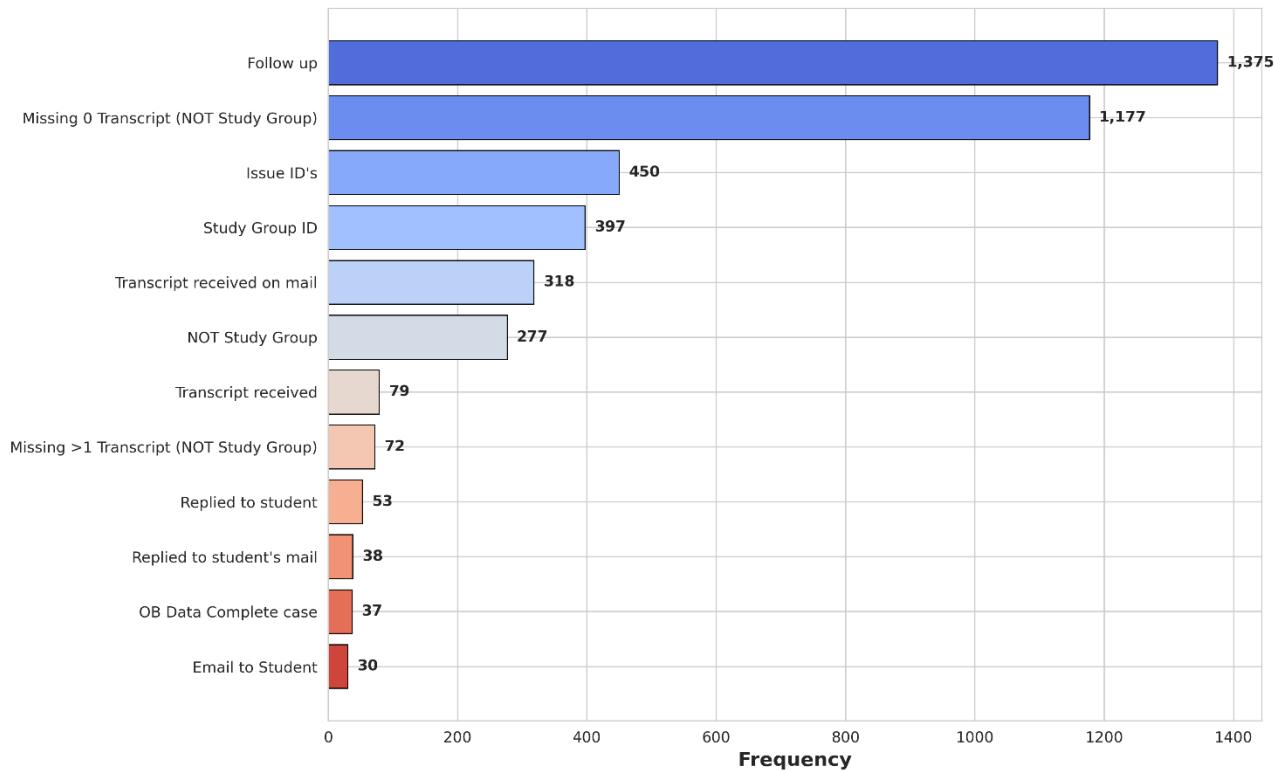
Missing 0 Transcript (NOT Study Group): 1,177 instances

Issue ID's: 450 instances

Study Group ID: 397 instances

Transcript received by mail: 318 instances

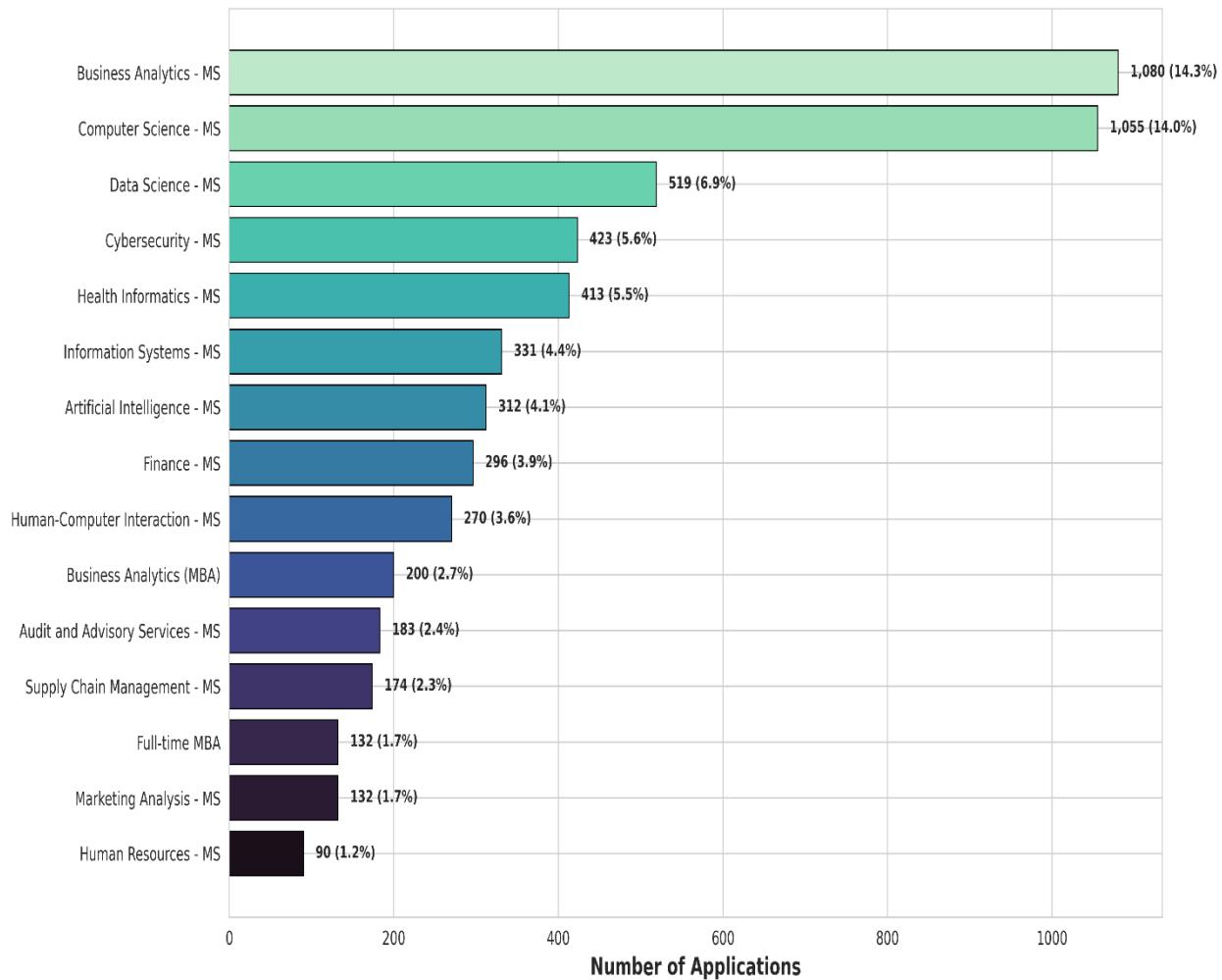
Figure 3: Distribution of Outreach Outcomes (Outcome 1)



- **Program Demand Pattern**

Figure 5 presents the top 15 programs by application volume, providing granular visibility into student academic interests. The visualization confirms technology and business analytics as dominant themes, with Computer Science, Business Analytics, and Data Science programs leading demand.

Figure 5: Top 15 Programs by Application Volume



KEY INSIGHTS OVERVIEW

Based on the exploratory analysis, the following early patterns emerge from the DePaul International Student Applications dataset:

Geographic Concentration:

- India dominates with 61.2% of all applications (4,617 out of 7,543), indicating high market concentration risk.
- Top 5 countries account for 86.4% of applications, showing limited geographic diversification.
- African markets (Ghana, Nigeria) are emerging as notable contributors.

Academic Preferences:

- Computing and Business programs drive international recruitment, with Jarvis College (50.8%) and Kellstadt (33.5%) representing 84.3% of applications.
- Strong STEM and professional graduate program preference aligns with global education trends.
- Significant application volume drop-off after top 2 colleges.

Recruitment Channels:

- Non-Study Group sources account for 86.5% of applications, showing diversified recruitment pathways.
- Study Group partnership provides supplementary (13.5%) rather than primary application volume.
- Balanced channel strategy reduces dependency on single partnership.

Data Quality Concerns

- 47.5% of variables (38 fields) are completely empty, limiting analytical capabilities.
- Only 7.5% of variables have complete data across all records.
- Data collection processes require systematic review and improvement.

CONCLUSION

This documentation provides a comprehensive overview of the DePaul International Student Applications dataset, containing 7,543 application records across 80 variables spanning demographic, academic, outreach, and operational data.

Summary of Findings

The exploratory analysis reveals three primary patterns in the current applicant pool:

1. Geographic concentration in India (61.2% of applications) presents both opportunity and risk, with emerging contributions from African markets.
2. Academic preferences heavily favor Computing and Business programs, with Jarvis College and Kellstadt accounting for 84.3% of first-choice selections.
3. Recruitment channels show balanced diversification, with non-partnership sources contributing 86.5% of application volume.