Question 1. Number of Fall 2025 applicants: 20024

Question 1.A) What are the number of entries for each term?

Fall 2022: 11

Fall 2023: 4785

Fall 2024: 17585

Fall 2025: 20024

Fall 2026: 6990

Spring 2022: 10

Spring 2023: 72

Spring 2024: 162

Spring 2025: 173

Spring 2026: 148

I filtered the table to count only rows where the term field equals 'Fall 2025', since each row represents a single application. An exact match was appropriate because the term values in the dataset use consistent formatting.

Questions 2. Percentage of entries from international students: 44.36%

Question 2 A) What are the total entries by citizenship category?

American: 22055

International: 22160

None: 5745

This query counts all rows where the status field is 'Accepted', directly answering how many applicants received an acceptance. I used an exact match because the status field is standardized and contains only a small set of known values.

Question 3. Average GPA, GRE, GRE V, and GRE AW (excluding missing values):

Average GPA: 3.81

Average GRE Total: 204.89

Average GRE Verbal: 160.42

Average GRE AW: 8.51

I filtered for rows where status = 'Rejected' to count all applicants who were denied admission. The field is clean and categorical, so an exact match provides an accurate count.

Question 4: Average GPA of American students in Fall 2025): 3.77

This query groups rows by university and counts how many applications each institution received. Grouping is necessary because each row represents one application, and the question asks for totals per university.

Question 5. Percent of Fall 2025 entries that are Acceptances: 36.00%

I grouped the dataset by the program field to count how many applications were submitted to each program. This structure directly reflects the one-row-per-application format of the table

Question 6. Average GPA of Accepted applicants in Fall 2025: 3.75

This query filters for rows where degree = 'PhD' to count all doctoral-level applications. The degree field is standardized, so an exact match reliably isolates PhD entries.

Question 7. Number of JHU Master's in Computer Science applicants: 6

filtered for rows where the LLM-generated program field equals 'Computer Science', since this field provides the cleanest representation of program names. Using the LLM-normalized field avoids the noise and inconsistencies present in the raw scraped text.

Question 8. Accepted 2025 PhD CS applicants to Georgetown University, Massachusetts Institute of Technology, Stanford Unviersity, or Carnegie Mellon University (using LLM generated values): 11

This query applies all required filters—term, status, degree, program, and university—to count only applicants who meet every condition in the question. I used the LLM-generated fields because they provide a more standardized representation of universities and programs than the raw scraped data.

Question 9 result: Accepted 2025 PhD CS applicants to Georgetown University, Massachusetts Institute of Technology, Stanford Unviersity, or Carnegie Mellon University (using downloaded fields values): 25

=== INTERPRETATION FOR QUESTION 9 ===

Using the LLM-generated fields (Question 8), I found 11 accepted 2025 PhD Computer Science applicants to the four target universities. Using the raw scraped fields with substring matching (Question 9), the count increased to 25.

This difference occurs because the raw program field is significantly noisier and contains inconsistent formatting, abbreviations, and concatenated text. When using ILIKE with broad substrings (e.g., “%Computer%”, “%MIT%”),

many additional entries are matched that would not be considered true Computer Science PhD applications. The LLM-generated fields are more standardized, so the filtering is more precise and produces a lower, more accurate count.