1. **Number of applicants for Fall 2025: 9900**
   - I used a COUNT(*) query and I used a WHERE clause so that the results are filtered by term = 'fall 2025` to determine how many submissions were for that specific application cycle. This gives us the total count of Fall 2025 entries from our database.
2. **The percentage of International Students: 57.48%**
   - COUNT(*) FILTER (WHERE ...) **count only rows that match 'international'.**
   - ROUND(..., 2): Rounds the result to 2 decimal places.
   - This query calculates the percentage of applicants labeled as 'international' by dividing their count by the total number of applicants.
3. **The average GPA: 3.70, GRE: 165.26, GRE V: 159.31, GRE AW: 4.24**
   - AVG(...): Aggregates numeric values and ::numeric: Casts the average to numeric so it can be rounded.
   - WHERE: Ensures only use the values in the where clause are included.
   - I used the AVG() function with typecasting (::numeric) to compute the mean of each metric. To avoid distorting the results, we filtered out invalid scores — e.g., GPAs above 4.33 or GRE scores outside their official ranges
4. **The average GPA of American students in Fall 2025: 3.71**
   - Multiple AND clauses filter the data by nationality, term, and valid GPA range.
   - BETWEEN 0 AND 4.33: Ensures GPA is within expected bounds.
   - I queried only american applicants applying for fall 2025 and ensured their GPA was within the 0–4.33 range.
5. **The percent of Acceptances for Fall 2025: 39.71%**
   - ILIKE 'Accepted%': Case-insensitive match on any status that starts with "Accepted".
   - NULLIF(..., 0): Prevents division by zero.
   - This query uses ILIKE 'Accepted%' to match all acceptance-related status messages. It calculates the percent of accepted entries out of all Fall 2025 entries.
6. **The average GPA of Acceptances for Fall 2025: 3.72**
   - WHERE term = 'fall 2025' filters only entries for the Fall 2025 application cycle and AVG(gpa): Computes the average GPA from the filtered rows..
   - AND status ILIKE 'Accepted%' Filters entries where the application status starts with "Accepted"
   - Computed the average GPA among applicants whose status begins with "Accepted" and whose GPAs are within the valid range filtered to fall 2025.
7. **The amount of JHU CS Masters Applicants: 9**
   - ILIKE '% ... %': Finds **substrings** regardless of case.
   - Multiple conditions I narrowed to using WHERE:
     - Specific university and  program '%Johns Hopkins University%Computer Science%'
     - Specific degree ILIKE '%master%';
   - I used a case-insensitive ILIKE filter on the program field to match any entry containing both "Johns Hopkins University" and "Computer Science", and filtered

for masters degree only. This helps identify how many applicants targeted that specific program.

## **Limitations:**

When analyzing the anonymously submitted applicant data from The Grad Café, I can see that it presents inherent challenges in data quality and reliability. Since entries are user-generated and unverified, there is a high risk of inaccurate or inconsistent information whether this is due to typos, misunderstandings, or intentional exaggeration from applicants. For example, before applying validation filters, I observed several unlikely or impossible values such as a GPA of 8.99 and GRE Analytical Writing scores that exceed the possible maximum of 6.0. These data entries clearly violated scoring scales used in academic admissions and significantly threw off our initial analysis. After filtering the dataset to exclude invalid entries, my results became more aligned with what we'd expect.

Some of the analytic results initially seemed surprising for instance, the average GRE score in the dataset was notably higher than the 2023–2024 ETS-reported national average (165.26 vs. ~157 for Quantitative). This discrepancy can be attributed to reporting bias meaning applicants may be more inclined to post strong results, while those with weaker applications may choose not to share. In other cases, users may have misunderstood the GRE field, inputting combined scores for Verbal and Quantitative sections instead of reporting each individually. I noticed this when I observed several entries above 170 for the GRE Quantitative section. This may be one of the biggest issues with using crowd-sourced data; it's often not representative or consistent, especially since the entries are anonymous and unverified, there's no way to ensure accuracy, and people might report things differently or even incorrectly. That's why it's so important to carefully clean and validate the data before relying on it for any meaningful analysis.