

Applicant Data Summary

Philip Turk - June 8, 2025

This document summarizes key findings extracted from the PostgreSQL applicants table using Python and SQL. Each question was answered using a direct SQL query, and a brief rationale is provided for transparency.

1. How many applicants do I have in my database table who have applied for Fall 2025?

Answer: 9899

Query:

```
SELECT COUNT(*) FROM applicants WHERE term = 'Fall 2025';
```

Why: I filtered on the "term" column to isolate all Fall 2025 entries.

2. What percentage of applicants are from International students (not American or Other)?

Answer: 57.45%

Query:

```
SELECT ROUND(100.0 * COUNT(*) / (SELECT COUNT(*) FROM applicants), 2)
FROM applicants
WHERE us_or_international NOT IN ('American', 'Other');
```

Why: I excluded known domestic categories and calculated the proportion. As I later came to discover, due to the absence of 'Other', I understand that I could have done this by filtering on 'International'.

3. What is the average GPA, GRE, GRE V, GRE AW of applicants who provide these metrics?

Answer: GPA: 3.72, GRE: 271.4, GRE_V: 158.7, GRE_AW: 4.19

Query:

```
SELECT ROUND(AVG(gpa)::numeric, 2) FROM applicants WHERE gpa IS NOT NULL;
SELECT ROUND(AVG(gre)::numeric, 1) FROM applicants WHERE gre IS NOT NULL AND gre
<= 340;
SELECT ROUND(AVG(gre_v)::numeric, 1) FROM applicants WHERE gre_v IS NOT NULL AND
gre_v <= 170;
```

```
SELECT ROUND(AVG(gre_aw)::numeric, 2) FROM applicants WHERE gre_aw IS NOT NULL  
AND gre_aw <= 6.0;
```

Why: I analyzed each metric separately and applied value limits where appropriate to exclude invalid entries.

4. What is the average GPA of American students in Fall 2025?

Answer: 3.71

Query:

```
SELECT ROUND(AVG(gpa)::numeric, 2)  
FROM applicants  
WHERE term = 'Fall 2025' AND us_or_international = 'American' AND gpa IS NOT NULL AND  
gpa <= 4.00;
```

Why: Focused on GPA for U.S. students in the target term, excluding unreasonably high values.

5. What percent of applicants for Fall 2025 are Acceptances?

Answer: 39.77%

Query:

```
SELECT ROUND(100.0 * COUNT(*) / (  
    SELECT COUNT(*) FROM applicants WHERE term = 'Fall 2025'), 2)  
FROM applicants  
WHERE term = 'Fall 2025' AND status ILIKE 'Accepted%';
```

Why: Counted accepted outcomes and compared to total Fall 2025 pool.

6. What is the average GPA of applicants who applied for Fall 2025 and were accepted?

Answer: 3.72

Query:

```
SELECT ROUND(AVG(gpa)::numeric, 2)  
FROM applicants  
WHERE term = 'Fall 2025' AND status ILIKE 'Accepted%' AND gpa IS NOT NULL;
```

Why: I focused on the GPA of accepted applicants in the given term.

7. How many entries are from applicants who applied to JHU for a Masters degree in Computer Science?

Answer: 9

Query:

```
SELECT COUNT(*) FROM applicants
WHERE program ILIKE '%Johns Hopkins University%Computer Science%' AND degree =
'Masters';
```

Why: I filtered for applicants who applied to Johns Hopkins University in Computer Science at the Masters level.

Limitations

Disclaimer: these thoughts are my own and not those from ChatGPT.

It is important to recognize that these applicants represent a type of convenience sample called a *voluntary response sample*. Because the latter is not drawn using a probability sampling design, any conclusions (under the statistical liturgy) apply only to the voluntary respondents and not some larger “population”. Conclusions drawn from a voluntary response sample are typically biased with respect to any attempt at inference about the population (e.g., estimation). For example, voluntary respondents might consist of more motivated, better students inclined to fill out a form on The GradCafe, which would explain the higher average GRE quantitative reasoning score of 165 you cite. Also, students filling out a public-facing form observed by their peers might be inclined to inflate metrics such as GPA, a type of bias known as *measurement bias*.

There are several other areas that represent limitations. First, and as is typically the case when working with real-world data, there were anomalies noted in the data at first glance. For example, the following were **not** included in computing statistics: one student with a GRE exceeding the maximum of 340; 24 students with a GRE AW score exceeding the maximum of 6.0; and 12 American students with a GPA above 4.00. Second, Questions #3 and #6 mix disparate American and international GPA scales (n.b., these scales could be standardized, but I did not do so here). Third, the distribution of GPAs is likely quite left-skewed due to grade inflation. Accordingly, the median should also be reported as a more robust measure of central tendency. Lastly, due to the federal government’s crackdown on U.S. universities and placing tight restrictions on international students, the results we see in Fall 2025 will naturally reflect this intervention and should be interpreted cautiously.