QUESTIONS/QUERIES/ANSWERS:

Q1: 'How many entries do you have in your database who have applied for Fall 2026?'

QUERY:

```
SELECT COUNT(p_id)
    FROM applicants
    WHERE term = 'Fall 2026';
```

ANSWER: 6810

EXPLANATION: Used simple SELECT query to get the number of unique IDs where the term was Fall 2026

Q2: 'What percentage of entries are from international students (not American or Other) (to two decimal places)?'

QUERY:

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

ANSWER: 50.27

EXPLANATION: Uses FILTER to count the number of rows where the criteria is true, which is divided by the number of all rows. That number is multiplied by 100 to represent a percentage and rounded to 2 decimal places.

Q3: 'How many entries do you have in your database who have applied for Fall 2026?'

QUERY:

```
SELECT
        ROUND(AVG(gpa)::numeric, 2),
        ROUND(AVG(gre)::numeric, 2),
        ROUND(AVG(gre_v)::numeric, 2),
        ROUND(AVG(gre_aw)::numeric, 2)
    FROM applicants;
```

ANSWER: GPA: 3.75 GRE: 230.98 GRE V: 161.53 GRE AW: 4.42

EXPLANATION: Found the average of each of the columns in question, rounded to two decimal points and returns result.

Q4: 'What is their average GPA of American students in Fall 2026?'

QUERY:

```
SELECT ROUND(AVG(gpa)::numeric, 2)
    FROM applicants
    WHERE us_or_international = 'American' AND term = 'Fall 2026';
```

ANSWER: 3.80

EXPLANATION: “Filtered” table to the relevant criteria and then averaged the GPA.

Q5: 'What percent of entries for Fall 2026 are Acceptances (to two decimal places)?'

QUERY:

```
SELECT ROUND(
        100.0 * COUNT(*) FILTER (WHERE term = 'Fall 2026' AND status = 'Accepted')
        / COUNT(*),
    2
)
FROM applicants;
```

ANSWER: 4.70

EXPLANATION: Uses FILTER to count the number of rows where the criteria is true, which is divided by the number of all rows. That number is multiplied by 100 to represent a percentage and rounded to 2 decimal places.

Q6: 'What is the average GPA of applicants who applied for Fall 2026 who are Acceptances?'

QUERY:

```
SELECT ROUND(AVG(gpa)::numeric, 2)
FROM applicants
WHERE status = 'Accepted' AND term = 'Fall 2026';
```

ANSWER: 3.76

EXPLANATION: This query finds the AVG of the GPA field and treats it as a numeric, it rounds it to 2 decimal points and “filters” it for only the data applicable to when the criteria is valid.

Q7: 'How many entries are from applicants who applied to JHU for a masters degree in Computer Science?'

QUERY:

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

ANSWER: 6

EXPLANATION: This “filters” down the table by using WHERE for multiple columns. It then reports out the length of the table with those “Filters” applied.

Q8: 'How many entries from 2026 are acceptances from applicants who applied to Georgetown University, MIT, Stanford University, or Carnegie Mellon University for a PhD in Computer Science?'

QUERY:

```
SELECT COUNT(*)
FROM applicants
WHERE status = 'Accepted'
AND EXTRACT(YEAR FROM date_added) = 2026
AND degree = 'PhD'
AND TRIM(SPLIT_PART(program, ',', 2)) IN ('Georgetown University', 'MIT', 'Stanford University', 'Carnegie Mellon University')
AND TRIM(SPLIT_PART(program, ',', 1)) = 'Computer Science';
```

ANSWER: 0

EXPLANATION: This “filters” down the table by using WHERE for multiple columns. It then reports out the length of the table with those “Filters” applied. The answer is 0 because there is not much data for 2026 and this is a very specific case (filters down a lot). No entries matched this criteria.

Q9: 'Do your numbers for question 8 change if you use LLM Generated Fields?'

QUERY:

```
SELECT COUNT(*)
    FROM applicants
    WHERE status = 'Accepted'
    AND EXTRACT(YEAR FROM date_added) = 2026
    AND degree = 'PhD'
    AND llm_generated_program = 'Computer Science'
    AND llm_generated_university IN ('Georgetown University', 'MIT', 'Stanford University',
'Carnegie Mellon University');
```

ANSWER: 0

EXPLANATION: This “filters” down the table by using WHERE for multiple columns. It then reports out the length of the table with those “Filters” applied. The answer is 0 because there is not much data for 2026 and this is a very specific case (filters down a lot). No entries matched this criteria.

Q10: 'How many unique program names and university names are in the data set?'

QUERY:

```
SELECT
      COUNT(DISTINCT TRIM(SPLIT_PART(program, ',', 1))),
      COUNT(DISTINCT TRIM(SPLIT_PART(program, ',', 2)))
    FROM applicants;
```

ANSWER: 3001, 2064

EXPLANATION: I came up with this question because I was curious about the extent of the variance in the dataset when it came to the program and university names. The COUNT function allows me to see the number of unique entries in a column.

Q11: 'How many unique llm-generated program names and university names are in the data set?'

QUERY:

```
SELECT
      COUNT(DISTINCT llm_generated_program),
      COUNT(DISTINCT llm_generated_university)
    FROM applicants;
```

ANSWER: 2933, 1503

EXPLANATION: I came up with this question because I was curious to compare the results to the previous question. When compared, these two results give us a metric of how effective the llm is in standardizing the data. From these results, it is much more effective at standardizing the universities (1503 vs 2064) than the the programs (2933 vs 3001). This could also be due to the university names having more variance.

Write-Up

There are certainly strong limitations with conducting analytics on a dataset of user-provided information. The two strongest limitations are in the form of poor data quality and biases. First off, if the data of interest is collected in the form of text fields where there is no limitation on the data value, it dramatically increases the chances of typos, misspellings, missing data, etc. For example, the module_2 data set from Gradcafe could contain 20 unique strings for the ‘university’ field when they were all referring to the same university. There are also cases of entries where the data is fake or misleading. For example, there was one such case where someone put in a fake university name and results just to leave a comment complaining about the types of comments people were making. This leads to greater variance in the aggregate data, making it much more difficult to derive meaning from the data analysis. As a solution, when collecting data from user input, forms should be designed to limit the variance in the input (ex: use a dropdown list of universities rather than have the user enter the name).

Another limitation with deriving meaning from user-entered information is that it can skew data in favor of a bias that a third party would not have. When reporting their own results, people are more likely to inflate their scores and acceptances. This gets amplified when they are able to remain anonymous. An anonymous user is far more likely to post incorrect or inflated information since there is no way for someone to fact check them. Also, users that perform better are more likely to report their results to others when compared to those who did not do well. This is the most likely reason why the average GRE score was 165 for applicants on GradCafe, but the actual average GRE score for 2023 was 157.