MODULE 3

EXPLANATION OF SQL CODE FOR Q10 AND Q11

QUESTION 10: What are the top 5 most popular programs applied to in gradcafe.com?

Code:

```
SELECT program, COUNT(*) AS count
FROM applicants
WHERE program IS NOT NULL AND program <> ''
GROUP BY program
ORDER BY count DESC
LIMIT 5;
```

I will explain what each line of this code does below (i.e. why I utilized it):

SELECT program, COUNT(*) AS count – SELECT returns the values of program name as output; COUNT (*) counts how many rows are in each group with the * designating to count every row. AS count renames the output to count and makes it readable. This line answers how many applicants applied to each program.

FROM applicants – instructs to read from the table “applicants”

WHERE program IS NOT NULL AND program <> '' – this line filters out junk and/or missing data. “Program is not null” removes rows where program is missing entirely and ‘program <> “’ removes rows where program is an empty string

GROUP BY program – this line groups all rows that share the same program together; this is required because SQL cannot return program alongside COUNT (*) unless program is grouped

ORDER BY count DESC – this line sorts the output in “count” in descending order (largest number of programs applied to are at the top)

LIMIT 5; - this returns only the first 5 results after sorting

QUESTION 11: What are the top 5 universities applied to for Physics PhD?

Code:

```
SELECT university, COUNT(*) AS count
FROM applicants
WHERE program = 'Physics PhD'
  AND university IS NOT NULL AND university <> ''
GROUP BY university
ORDER BY count DESC
LIMIT 5;
```

I will explain what each line of this code does below (i.e. why I utilized it):

SELECT university, COUNT(*) AS count – this selects the university field (the school the applicant applied to) and COUNT (*) counts how many rows are in each group with the * designating to count every row. AS count renames the output to count and makes it readable.

FROM applicants – instructs to read from the table “applicants”

WHERE program = 'Physics PhD' – this isolates only the Physics PhD programs applied to in the data (i.e. looks at a subset of all the programs)

AND university IS NOT NULL AND university <> '' – similar to Q10, filters out missing universities and prevents null from appearing in the data

GROUP BY university – this groups all Physics PhD rows by university allowing COUNT (*) per university

ORDER BY count DESC – this sorts the data in descending order

LIMIT 5; - this shows only the top 5 results

Question 10 groups the full dataset by program and counts frequency to find the top 5 most common programs applied to.

Question 11 filters the dataset to Physics PhD only and then groups by university; it then counts the frequency to find the top 5 target universities for Physics PhD applicants