NAU NORTHERN ARIZONA UNIVERSITY

Institutional Research & Analysis Applications Programmer – Candidate Interview Exercise

You've been asked to prepare data for the Vice Provost regarding our full-time student cohort populations. The Vice Provost is specifically interested in an analysis that compares our non-retained students and our retained students by Fall Cohort type (freshman, lower-division transfer, upper-division transfer). Consider highlighting both similarities and differences between the various populations.

Please use the fictitious dataset and the data dictionary we provided to you to complete the following:

1. What questions, if any, do you have before beginning work on this project?
   The definitions for lower-division and upper-division transfers (found in Data Dictionary). Do we have a division for full time online vs in person. Which are we trying to learn more about online or in person. Do we have a section where students can put why they are leaving. If not, can we get that information?

2. How do you go about tackling this project? What are your first steps?
   Read through the dictionary. Looking for specific population breakdowns. Download xlxs file as a csv file. Use python and sqlite3 to create a database for info. Within terminal sqlite3 nau.db, .mode csv, .import naudata.csv nau. Perform .schema on the database to see each column title and data type.

3. What tool(s) would you use to extract data from your dataset?
   Python and sqlite3 in an anaconda environment.

4. What programming statements would you write to extract data for this project?

   SELECT DISTINCT COUNT(STUDENT_ID) FROM naudata; Gets the total number of distinct students which is 30832.
   To see the different options logged for each column.
   SELECT DISTINCT FALL_COHORT FROM naudata; Only option 2019.
   SELECT DISTINCT COHORT_TYPE FROM naudata; FTF, N/A, LD, UD
   SELECT DISTINCT GENDER FROM naudata; F, M
   SELECT DISTINCT N__IPEDS_ETHNICITY FROM naudata; HISPA, WHITE, ASIAN, INTL, NSPEC, TWOMORE, BLACK, AMIND, PACIF
   SELECT DISTINCT ENROLLMENT_STATUS FROM naudata; F, P

SELECT DISTINCT EXCLUSION_ELIGIBLE FROM naudata; N, Y
SELECT DISTINCT ENROLLED_FALL_2020 FROM naudata; Y, N/A

Check to see if any students are Exclusion Eligible? SELECT COUNT(STUDENT_ID) FROM naudata WHERE EXCLUSION_ELIGIBLE='Y'; There are 5 who can be excluded.
Check the cohort type for all exclusion eligible students. SELECT COHORT_TYPE FROM naudata WHERE EXCLUSION_ELIGIBLE='Y'; Answer for all 5 was N/A so these will not factor into our analysis of FTF, LD and UD.

Then run percent.py
https://github.com/mxw035/nau/blob/03b2046d04204d88134e4b2d1fe1432dad9dd092/percent.py

5. After extracting the data, what steps do you take to validate your results?
   Run percent.py can be checked by using sql commands within the terminal.

   Every step of the way checking that when the numbers and percentages for subcategories are determined they all add back up to total per category and grand total.

6. After extracting the data and analyzing it, what are the top 3 findings you would highlight for the Vice Provost?

   In general 60% of students where retianed while 40% where lost from Fall 2019-2020.
   NAU is most likely to retain students who identify as white, female, full time and are part of the FTF cohort.
   NAU is most likely to lose students who identify as non-white, male, part time, and are in the LD or UD cohort.

   More data should be collected to identify the cause of withdrawing or not returning the next year. As well as a comparison per ethnicity and gender within ethnicities to see if there are more resources that can be put into place for students.