**Business Data Management Individual Project**

**Fall 2021**

The database in this project is about colleges and universities, hereinafter called schools. The following information about schools is provided: school conferences, starting and mid-career salaries for schools, and starting and mid-career salaries for degrees. The following are the relations you need to create:

schools(school, conference)

degree\_salary(degree, starting\_median, mid\_career\_median, mid\_career\_90)

school\_salary(school, region, starting\_median, mid\_career\_median, mid\_career\_90)

All source files are in **individual\_project.zip**.

The main source file for schools is **schools\_src.csv**. The main source file for degree\_salary is **degree\_salary\_src.csv**. The main source file for school\_salary is **school\_salary\_src.csv**.

Each school is in one of the following conferences: Patriot, Pac-12, SEC, Big 12, ACC, Big Ten, and Independent. Each school is in one of the following regions: Northeastern, Southern, Western, Midwestern, and California. Each school has salary information. Each degree in the relation degree\_salary has salary information. The relation school\_salary is the authoritative source for school names.

The salaries in each of the salary relations are the starting median starting salary for students, the mid-career median salary for graduates, and the 90th percentile (highest 10%) salaries for graduates at mid-career.

Some of the attributes in the relations are unknown and, should therefore, be set to NULL. If the attribute is NULL in a source file, the attribute will be an empty string ''. You will need to do arithmetic on the salaries, so be sure that ultimately, you load them as numbers.

Whenever you start a project, you must clean and integrate the data and check it for consistency. This project is no different. You may find that you need to convert empty strings, '', to null values, change school names in schools to match school\_salary, or change the format of the salaries so they can be treated as numbers. A question you will need to answer is: “How will you find the rows that you need to update?”

The data in these relations come from various sources. Schools is derived from a dataset fivethirtyeight.com used for an article called “Our Guide To The Exuberant Nonsense Of College Fight Songs.” The article is available at <https://projects.fivethirtyeight.com/college-fight-song-lyrics/>. Data was downloaded from <https://fivethirtyeight.datasettes.com/fivethirtyeight/fight-songs%2Ffight-songs> on 11/2/2020. The schools for the Patriot League were compiled by Professor Ordille using Internet Search. Professor Ordille has normalized most, though perhaps not all, the school names in the schools relation to match the names in the school\_salary relation.

Salary information about degrees and schools is derived from a dataset from the Wall Street Journal used for an article called “Where it Pays to Attend College.” Data was downloaded from [https://www.kaggle.com/wsj/college-salaries on 11/2/2020](https://www.kaggle.com/wsj/college-salaries%20on%2011/2/2020). In order to guarantee a match between school names in schools and school names in school salary, Professor Ordille estimated the salaries at some schools using Internet search. Be aware that these values do not come from the validated sources used by the Wall Street Journal and may be inaccurate:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| school | region | starting\_median | mid\_career\_median | mid\_career\_90 |
| Loyola University Maryland | **Southern** | **59900.00** | **111200.00** |  |
| Texas Tech University | **Southern** | **47000.00** |  |  |
| The United States Military Academy (Army) | **Northeastern** | **74000.00** | **120000.00** |  |
| United States Naval Academiy (Navy) | **Southern** | **77100.00** | **131000.00** |  |
| University of Louisville | **Southern** | **38000.00** |  |  |
| University of Miami | **Southern** | **58760.00** |  |  |
| University of Pittsburgh (Pitt) | **Northeastern** | **45700.00** | **74000.00** | **150000.00** |
| Wake Forest University | **Southern** | **63800.00** |  |  |
|  |  |  |  |  |

Answer the following questions by performing SQL operations on the database. Your SQL answer must work for any data or changes to data in this database. Your data answer must be in a single table you generated using SQL, not several tables that need to be combined manually to get the answer to the question. You cannot use data values in a query unless they are specified in the question. If you need another value, generate it with a query and use that query or its resulting table in the query that ultimately produces the answer. Unless specified otherwise, remove duplicates from your answers as appropriate. Show the answer (that is a table of data that answers the question) and the SQL used to generate the answer.

* 1. What is the SQL for creating the relations and loading the files into the relations. Be sure to include any key, foreign key, or check constraints that you see when creating the relations. Include any transformations you do on the data to enable the data to be loaded/used and made consistent. (20 points)
  2. What is all the information in the school\_salary relation about tech schools in descending order by starting median salary? (10 points)
  3. What is the degree and salary information for the degree with the highest 90th percentile mid-career salary? (10 points)
  4. What is all the salary information for the schools in the Big Ten in decreasing order of mid-career 90th percentile salary? (10 points)
  5. List the school and salary information for these NJ schools: Fairleigh Dickinson University, Princeton University, Rider University, Rutgers University, Seton Hall University, Stevens Institute of Technology in ascending order by school. Use FORMAT and CONCAT to create a string for the salary that has a starting $ and a comma after the thousands place, for example: $49,200.00 . The salary columns in the result should be named starting\_median, mid\_career\_median, and mid\_career\_90. (10 points)
  6. List the degree and starting median salary in descending order by median salary for degrees about information, marketing, accounting, finance, or business. Use FORMAT and CONCAT to create a string for the salary that has a starting $ and a comma after the thousands place, for example: $49,200.00 . The result columns should be named degree and starting\_median. (10 points)
  7. What schools in the Big Ten have a higher median starting salary than the median starting salary of Management Information Systems (MIS), and what are their median starting salaries? Format the starting salaries with a starting $ and a comma after the thousands place in a result column called starting\_median. List the schools in the answer in descending order by median starting salary. (10 points)
  8. What are the schools, conferences, regions and starting median salaries for schools that do not have a median mid-career salary listed for the 90th percentile? Format the starting salaries with a starting $ and a comma after the thousands place, and call the column starting\_median in the result. Also include a column called both\_mid\_career\_unknown which should be set to True if both the mid-career median and the mid-career 90th percentile are set to null and False otherwise. Sort the result in ascending order by conference and then school. (10 points)
  9. What is the name, median starting salary, median mid-career salary, and percentage increase from median starting to median mid-career salary for the school(s) with the highest percentage increase? Calculate the percentage increase as ((mid\_career\_median – starting\_median)/starting\_median)\*100. Round the percentage increase to the nearest integer and add a % symbol to the end. The column with the percentage increase should be called percent\_incr. (10 points).