**School Exercise**

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

school(school\_name, conference)

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

All the attributes in these tables must be typed correctly. In particular, starting\_median, mid\_career\_median, mid\_career\_90 are numeric. The fractional parts of these numbers should be preserved. The sizing of the fields should be as small as possible including VARCHAR types which should have a maximum size just big enough to fit the longest string.

All source files are in **school\_exercise.zip**.

The main source file for schools is **school\_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. The relation school\_salary is the authoritative source for school names.

The salaries 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 school 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. School 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 school relation to match the names in the school\_salary relation.

Since the school\_src dataset was created, changes have occurred to the membership of the athletic conferences, only some of which have been updated in the dataset. After loading the dataset, you will need to use SQL to update the conferences for the following schools as part of creating the school table.

|  |  |  |
| --- | --- | --- |
| **School** | **Old Conference** | **New Conference** |
| Stanford University | Pac-12 | ACC |
| University of California, Berkeley | Pac-12 | ACC |
| Arizona State University (ASU) | Pac-12 | Big 12 |
| University of Arizona | Pac-12 | Big 12 |
| University of Colorado - Boulder (UCB) | Pac-12 | Big 12 |
| University of Utah | Pac-12 | Big 12 |
| University of California at Los Angeles (UCLA) | Pac-12 | Big Ten |
| University of Oregon | Pac-12 | Big Ten |
| University of Washington (UW) | Pac-12 | Big Ten |

**Table of Changes Needed to Athletic Conferences for Schools in the School\_Src Dataset**

Salary information about 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. In order to guarantee a match between school names in school 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 Academy (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** |  |  |
|  |  |  |  |  |

To complete this assignment, you **MUST** create a MySQL script that runs from beginning to end to:

* create the target tables school and school\_salary,
* clean and insert the data into the target tables from the uploaded input files, and
* generate answers to all the questions.

You **MUST** also submit the answers to the the questions in this file via individual Gradescope questions. Also submit your sql file (2 points) through Gradescope. You must submit the requested information for each question in Gradescope. You cannot just submit the script. You can submit your answers to individual questions to Gradescope as you create them, and update your answers/resubmit through the end of the submission period.

It is not enough to simply have the right SQL somewhere in the script file. We will test your answer by running your script on our upload of the input files. Note that you **CANNOT** change the input csv files. If you do so, we will not be able to execute your script successfully from the input files, and you will lose points. You may change the source tables (school\_src and school\_salary\_src) after you upload the source input files into those tables.

You will find the following commands helpful when running your script repeatedly:

DROP TABLE IF EXISTS school;

DROP TABLE IF EXISTS school\_salary;

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. You will need to submit the answer (that is a table of the data that answers the question for 2-9), and the SQL used to generate the answer.

If you have a Windows machine, you can create a picture of the answer table for each query to submit in Gradescope by following these steps:

* Press the Windows logo key ![A black square with blue and orange lines

  AI-generated content may be incorrect.]() and the shift and s keys to activate the snippet tool.
* Click in the upper left-hand corner of the area you wish to capture in the screenshot, and drag a rectangle around the area.
* Open the draw tool and paste the picture.
* Save the picture as a .png of .jpeg file.
* Upload this file to the correct question in Gradescope.

If you have a Mac, you can create a screenshot of the answer table for each query to submit in Gradescope by following these steps:

* Press the Command key A black symbol with a white background

  AI-generated content may be incorrect. and the shift and 4 keys.
* Click in the upper left-hand corner of the area you wish to capture in the screenshot, and drag a rectangle around the area.
* The screenshot will appear on your desktop.
* Upload this file to the correct question in Gradescope.

Note that you must include text for the SQL code, not a picture of the SQL, in the box associated with each question. The queries for Questions 2-9 must be done on the target tables **NOT** the source tables you uploaded from the files.

1. What is the SQL for creating the target relations and inserting the data into the target relations? Be sure to include any key, foreign key, or check constraints that you see when creating the target relations. Include any transformations you do on the data to enable the data to be transferred into the target tables, used and made consistent. (6 points)
2. What is all the salary information about Rutgers University? 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. (2 points)
3. How many schools are in each conference? Name the column with the number of schools: "Number of Schools". List the results in descending order of number of schools. If two conferences have the same number of schools, let the conferences in ascending order. (2 points)
4. What is all the information in the school\_salary relation about tech schools in descending order by starting median salary? (2 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. (2 points)
6. What are the regions for the schools in each conference? List the conference and region ordered by conference and then region. For each conference, list each region once. (2 points)
7. What is all the salary information for the schools in the Big Ten in decreasing order of mid-career median salary? (2 points)
8. What is the school, conference, region and all the salary information for schools in the Northeastern region with the highest mid\_career\_median salary? Be sure to list qualifying schools even if they do not have an athletic conference listed. You cannot use LIMIT to answer this question. LIMIT 1 will not work if there is a tie for the highest salary. (2 points)
9. Excluding the Big Ten, what conference has the most member schools and how many member schools does it have? The answer to this question would be the conference or conferences that have the second most schools of any conferences. You cannot use LIMIT to answer this question. LIMIT 1 will not work if there is a tie for the highest number of members. (2 points)
10. You must create a script that creates the target tables, does any necessary data integration, populates the target tables, and then generates the answers to each of the queries. The script must run from beginning to end successfully when nothing is highlighted and the lightning bolt is pressed. Submitting only the script will cost you points. You must submit the query and the picture for each question to the corresponding Gradescope question. (2 points)