Alexandra Garthe

BDM Individual Project

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. 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.

a. 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)

Cleaning school_salary_src

```
#Setting school_salary empty text "" values to NULL values
USE Ally;

SELECT *
FROM school_salary_src;

UPDATE school_salary_src
SET mid_career_median = NULL
WHERE mid_career_median = "";

UPDATE school_salary_src
SET mid_career_90 = NULL
WHERE mid_career_90 = "";

SELECT *
FROM school_salary_src;

#removing the special character "$" from cells with special character "$"
UPDATE school_salary_src
SET starting_median=REPLACE(starting_median,"$","");

UPDATE school_salary_src
SET mid_career_median=REPLACE(mid_career_median,"$","");

UPDATE school_salary_src
SET mid_career_90=REPLACE(mid_career_90,"$","");

SELECT *
FROM school_salary_src; #check results

#delete comma
UPDATE school_salary_src
SET starting_median=REPLACE(starting_median,",","");

UPDATE school_salary_src
SET mid_career_median=REPLACE(mid_career_median,",","");

UPDATE school_salary_src
SET mid_career_90=REPLACE(mid_career_90,",","");

SELECT *
```

```
FROM school_salary_src; #check results


Creating school_salary
DROP TABLE school_salary;

CREATE TABLE school_salary (
school                  VARCHAR(70) PRIMARY KEY,
region                  VARCHAR(50),
starting_median         DECIMAL(10,2),
mid_career_median       DECIMAL(10,2),
mid_career_90           DECIMAL(10,2)
);

INSERT INTO school_salary (school, region, starting_median, mid_career_median,
mid_career_90)
SELECT school, region, starting_median, mid_career_median, mid_career_90
FROM school_salary_src;

SELECT * FROM school_salary;

Cleaning degree_salary_src.csv
#check for Null values
SELECT *
FROM degree_salary_src
Where degree = "";

SELECT *
FROM degree_salary_src
Where starting_median_salary = "";

SELECT *
FROM degree_salary_src
Where mid_career_median_salary = "";

SELECT *
FROM degree_salary_src
Where mid_career_90th_percentile_salary = "";

#removing the special character "$" from cells with special character "$"
SELECT *
FROM degree_salary_src;

UPDATE degree_salary_src
SET starting_median_salary=REPLACE(starting_median_salary,"$","");
```

```
UPDATE degree_salary_src
SET mid_career_median_salary=REPLACE(mid_career_median_salary,"$","");

UPDATE degree_salary_src
SET mid_career_90th_percentile_salary=REPLACE(mid_career_90th_percentile_salary,"$","");

#delete comma
SELECT *
FROM degree_salary_src;

UPDATE degree_salary_src
SET starting_median_salary=REPLACE(starting_median_salary,",","");

UPDATE degree_salary_src
SET mid_career_median_salary=REPLACE(mid_career_median_salary,",","");

UPDATE degree_salary_src
SET mid_career_90th_percentile_salary=REPLACE(mid_career_90th_percentile_salary,",","");
```

Create Table degree_salary

```
CREATE TABLE degree_salary (
degree                  VARCHAR(70) PRIMARY KEY,
starting_median         DECIMAL(10,2),
mid_career_median       DECIMAL(10,2),
mid_career_90           DECIMAL(10,2)
);

INSERT INTO degree_salary (degree, starting_median, mid_career_median, mid_career_90)
SELECT degree, starting_median_salary, mid_career_median_salary,
mid_career_90th_percentile_salary
FROM degree_salary_src;

SELECT *
FROM degree_salary;
```

Clean data from school_src.csv and Create Table Schools

```
#create table schools
DROP TABLE schools;
CREATE TABLE schools (
school              VARCHAR(70) PRIMARY KEY,
conference          VARCHAR(50),
FOREIGN KEY (school) REFERENCES school_salary(school)
);

#check which schools are not the same in school_salary_src
SELECT school
```

```
FROM schools_src
WHERE school NOT IN (
                SELECT school
      FROM school_salary_src);

# update: Rutgers and notre dame
UPDATE schools_src
SET school = "University of Notre Dame" WHERE school LIKE "%Notre Dame%";

UPDATE schools_src
SET school = "Rutgers University" WHERE school LIKE "%Rutgers%";

#double check
SELECT school
FROM schools_src
WHERE school NOT IN (
                SELECT school
      FROM school_salary_src);

#insert data
INSERT INTO schools (school, conference)
SELECT school, conference
FROM schools_src;

SELECT*
FROM schools;
```

b. What is all the information in the school_salary relation about tech schools in descending order by starting median salary? (10 points)

```
SELECT*
FROM school_salary
WHERE school LIKE "%tech%"
ORDER BY starting_median DESC;
```

school	region	starting_median	mid_career_median	mid_career_90
California Institute of Technology (CIT)	California	75500.00	123000.00	NULL
Massachusetts Institute of Technology (MIT)	Northeastern	72200.00	126000.00	220000.00
Polytechnic University of New York, Brooklyn	Northeastern	62400.00	114000.00	190000.00
Rensselaer Polytechnic Institute (RPI)	Northeastern	61100.00	110000.00	182000.00
Worcester Polytechnic Institute (WPI)	Northeastern	61000.00	114000.00	180000.00
Stevens Institute of Technology	Northeastern	60600.00	105000.00	185000.00
Georgia Institute of Technology	Southern	58300.00	106000.00	183000.00
Illinois Institute of Technology (IIT)	Midwestern	56000.00	97800.00	165000.00
South Dakota School of Mines & Technology	Midwestern	55800.00	93400.00	147000.00
Virginia Polytechnic Institute and State Universit…	Southern	53500.00	95400.00	163000.00
Wentworth Institute of Technology	Northeastern	53000.00	96700.00	153000.00
New Mexico Institute of Mining and Technology…	Western	51000.00	93400.00	NULL
Rochester Institute of Technology (RIT)	Northeastern	48900.00	84600.00	159000.00
Texas Tech University	Southern	47000.00	NULL	NULL
Tennessee Technological University	Southern	46200.00	80000.00	121000.00
Fashion Institute of Technology	Northeastern	42800.00	81000.00	138000.00

c. What is the degree and salary information for the degree with the highest 90th percentile mid-career salary? (10 points)

```
SELECT degree, starting_median, mid_career_median, mid_career_90
FROM degree_salary
WHERE mid_career_90 = (
        SELECT MAX(mid_career_90)
        FROM degree_salary);
```

degree	starting_median	mid_career_median	mid_career_90
Economics	50100.00	98600.00	210000.00

d. What is all the salary information for the schools in the Big Ten in decreasing order of mid-career 90th percentile salary? (10 points)

```
SELECT schools.school, conference, starting_median, mid_career_median, mid_career_90
FROM schools INNER JOIN school_salary ON schools.school = school_salary.school
WHERE conference = "Big Ten"
ORDER BY mid_career_90 DESC;
```

school	conference	starting_median	mid_career_median	mid_career_90
Northwestern University	Big Ten	52700.00	95900.00	205000.00
University of Michigan	Big Ten	52700.00	93000.00	182000.00
Indiana University (IU), Bloomington	Big Ten	46300.00	84000.00	178000.00
University of Illinois at Urbana-Champaign (UIUC)	Big Ten	52900.00	96100.00	177000.00
Rutgers University	Big Ten	50300.00	91800.00	176000.00
University of Wisconsin (UW) - Madison	Big Ten	48900.00	87800.00	170000.00
Michigan State University (MSU)	Big Ten	46300.00	85300.00	170000.00
Purdue University	Big Ten	51400.00	90500.00	168000.00
University of Maryland, College Park	Big Ten	52000.00	95000.00	166000.00
University of Iowa (UI)	Big Ten	44700.00	83900.00	163000.00
Ohio State University (OSU)	Big Ten	44900.00	83700.00	162000.00
Pennsylvania State University (PSU)	Big Ten	49900.00	85700.00	160000.00
University of Nebraska	Big Ten	45700.00	80900.00	156000.00
University of Minnesota	Big Ten	46200.00	84200.00	148000.00

e. 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)

```
SELECT school, CONCAT("$",FORMAT(starting_median,2)) AS starting_median,
CONCAT("$",FORMAT(mid_career_median,2)) AS mid_career_median,
CONCAT("$",FORMAT(mid_career_90,2)) AS mid_career_90
FROM school_salary
WHERE school IN ("Fairleigh Dickinson University", "Princeton University", "Rider
University", "Rutgers University", "Seton Hall University", "Stevens Institute of
Technology")
ORDER BY school;
```

school	starting_median	mid_career_median	mid_career_90
Fairleigh Dickinson University	$45,700.00	$78,700.00	$171,000.00
Princeton University	$66,500.00	$131,000.00	$261,000.00
Rider University	$43,600.00	$88,900.00	$150,000.00
Rutgers University	$50,300.00	$91,800.00	$176,000.00
Seton Hall University	$48,900.00	$89,200.00	$195,000.00
Stevens Institute of Technology	$60,600.00	$105,000.00	$185,000.00

f. 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)

```
SELECT degree, CONCAT("$",FORMAT(starting_median,2)) AS starting_median
FROM degree_salary
WHERE degree LIKE "%information%" OR degree LIKE "%marketing%"OR degree LIKE
"%accounting%"OR degree LIKE "%finance%" OR degree LIKE "%business%"
ORDER BY starting_median DESC;
```

degree	starting_median
Management Information Systems (MIS)	$49,200.00
Information Technology (IT)	$49,100.00
Finance	$47,900.00
Accounting	$46,000.00
Business Management	$43,000.00
Marketing	$40,800.00

g. 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)

```
SELECT schools.school, CONCAT("$",FORMAT(starting_median,2)) AS starting_median
FROM schools JOIN school_salary ON schools.school = school_salary.school
WHERE conference = "Big Ten" AND starting_median > (
      SELECT starting_median
      FROM degree_salary
      WHERE degree = "Management Information Systems (MIS)")
ORDER BY starting_median DESC;
```

school	starting_median
University of Illinois at Urbana-Champaign (UIUC)	$52,900.00
Northwestern University	$52,700.00
University of Michigan	$52,700.00
University of Maryland, College Park	$52,000.00
Purdue University	$51,400.00
Rutgers University	$50,300.00
Pennsylvania State University (PSU)	$49,900.00

h. 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)

```
SELECT schools.school, conference, region, CONCAT("$",FORMAT(starting_median,2))
AS starting_median, IF ((mid_career_median IS NULL) AND (mid_career_90 IS
NULL),"TRUE","FALSE" ) AS both_mid_career_unknown
FROM schools JOIN school_salary ON schools.school = school_salary.school
WHERE mid_career_90 IS NULL
ORDER BY conference, school;
```

school	conference	region	starting_median	both_mid_career_unknown
University of Louisville	ACC	Southern	$38,000.00	TRUE
University of Miami	ACC	Southern	$58,760.00	TRUE
Wake Forest University	ACC	Southern	$63,800.00	TRUE
Texas Tech University	Big 12	Southern	$47,000.00	TRUE
College of the Holy Cross	Patriot	Northeastern	$50,200.00	FALSE
Loyola University Maryland	Patriot	Southern	$59,900.00	FALSE
The United States Military Academy (Army)	Patriot	Northeastern	$74,000.00	FALSE
United States Naval Academiy (Navy)	Patriot	Southern	$77,100.00	FALSE

i. 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).

```
SELECT school, starting_median, mid_career_median,
CONCAT((ROUND((mid_career_median - starting_median)/(starting_median)*100)),"%")
AS percent_incr
FROM school_salary
WHERE (ROUND((mid_career_median - starting_median)/(starting_median)*100))  = (
        SELECT MAX((ROUND((mid_career_median -
        starting_median)/(starting_median)*100))) AS percent_incr
        FROM school_salary);
```

school	starting_median	mid_career_median	percent_incr
Dartmouth College	58000.00	134000.00	131%