Project 1
Databases and visualization with new data

This project is worth 20% of your course grade. It is to be completed individually. You may not collaborate with anyone else, and you may not look for examples on the Internet which use this data. Your submission should take the form of a document which includes everything – your written work, queries, samples of results, and images of your visualizations. You will submit this document to the assignment for this project on the Blackboard page.

Section 1: Setting up the database

You are going to be exploring a new dataset called the Employees Sample Database. It is a decently-sized database created to test MySQL systems. To install the database in your RDS instance, do the following:

1. Log into your EC2 instance.

2. Download the zip file (which has the setup scripts) to your EC2 instance from my website:
   ```
   wget analytics.drake.edu/~manley/employee_db.zip
   ```

3. Unzip the employee_db.zip file with the command
   ```
   unzip employee_db.zip
   ```

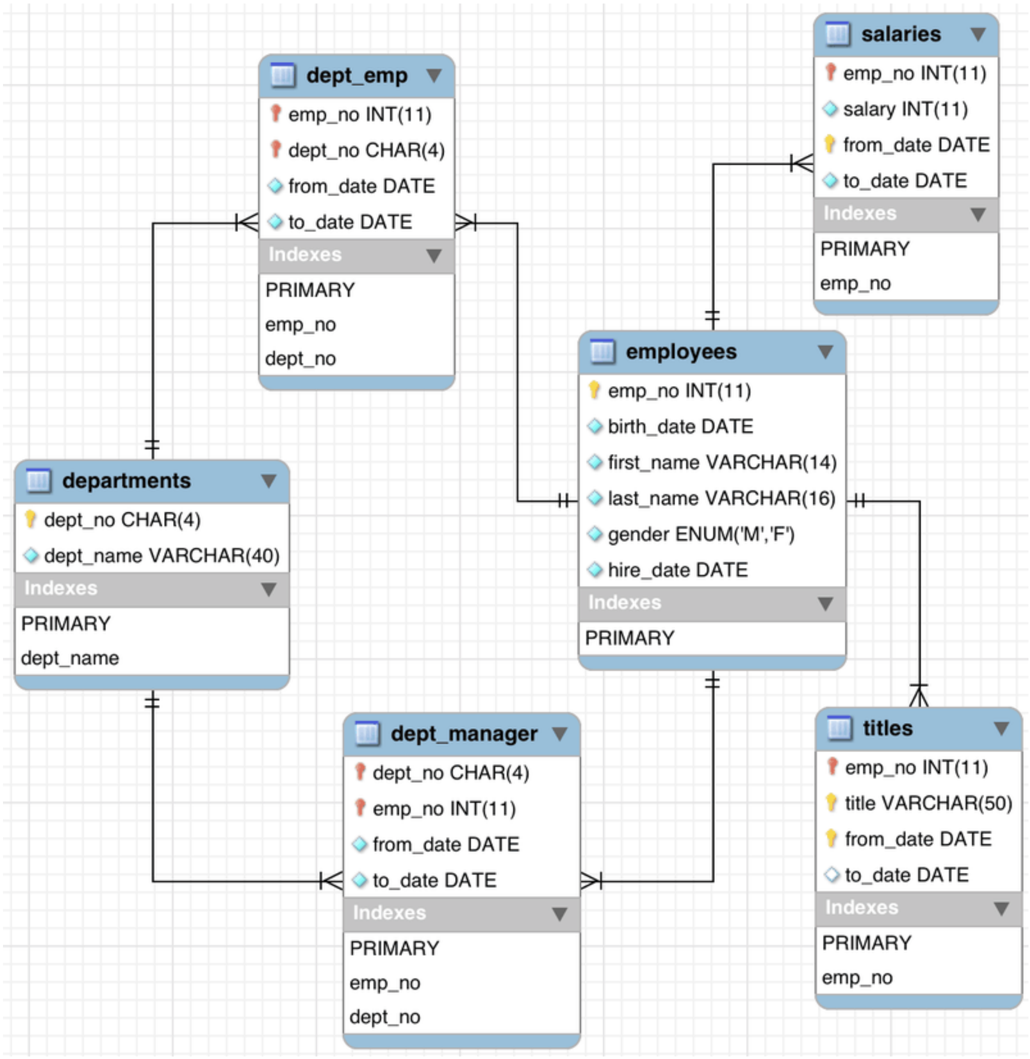
4. This should create a folder in your EC2 instance's home directory called test_db-master. Change to this directory:
   ```
   cd test_db-master/
   ```

5. Use the install script to create the database and insert data into your RDS instance (this is just like Step 2 from Section 4 of Lab 9 – you need to replace the red text below with the address of your RDS instance's endpoint and username respectively):
   ```
   mysql -h mysql-instance1.123456789012.us-east-1.rds.amazonaws.com -P 3306 -u mymasteruser -p < employees.sql
   ```

This will create a database called "employees" on your RDS instance. You can log into your RDS instance and use it like any other database.

Section 2: Getting to know the employees database

The Entity Relationship Diagram for the employees database looks like this:



In this notation ─|< means “many” and ||─ means “one”. For example, this means each key in the employees table (emp_no) may appear many times in the salaries table.

I suggest writing some queries to look at examples of the rows from each table. However, this is a large database, so you will probably want to make use of LIMIT.

Section 3: Questions and queries

For this section, think of some English-language questions that someone might be interested to know about this data and the MySQL queries you’d need to answer them. Each of these queries should involve more than one table. Furthermore, I want to see

1. at least one query that involves three or more tables
2. at least one query that involves a subquery
3. at least one query that involves grouping

In your write-up document, write down at least three English-language questions. For each one, give the query that answers it, and show a sample of what the results look like after you run it.

Section 4: Visualizations

Connect the employees database to Tableau, and create a dashboard with visualizations that provide additional insight about the data. Your dashboard should have the following:

1. at least three different charts/graphs
2. at least one chart/graph that uses multiple tables
3. proper labels, titles, legends, etc. for all charts/graphs
4. good visualization design based on what you've learned in this course

Your write-up document should include an image of your entire dashboard, and you may include additional images which show individual graphs if the detail is not clear from the full dashboard image. Additionally, you should answer the following questions about each visualization in your dashboard:

1. What insight or story does this visualization communicate that isn't clear from looking at a table alone?
2. In what ways does this visualization reflect good design principles? You can talk about chart type, color, etc.

Section 5: Rubric

Your project will be graded based on the following 3 criteria:

1. Demonstrated ability to construct correct queries which answer questions about data using joins, subqueries, and groupings
2. Demonstrated ability to explain questions that are addressed by a query
3. Demonstrated ability to select appropriate visualization types and use them to effectively communicate insights about data

Each of these will be scored from 0 to 4 points, meaning

- 0 points: not attempted
- 1 point: a serious attempt has been made, but the ability has not been demonstrated
- 2 points: the ability has been partially demonstrated, though there are problems that need correction
- 3 points: the ability has been mostly demonstrated, and only minor corrections are needed
- 4 points: full proficiency in the ability has been demonstrated, and no changes are necessary

Note that these 12 points are on a completely different scale than the points on the labs – this score will ultimately be scaled to reflect 20% of your overall grade for the course.