CS 457 - Homework Assignment 5: SQL

Due Date: Monday, February 20 at 11:59 pm

Purpose:
Demonstrate exploration of data via creation of statistical tables using RDBMS/SQL; connecting Python with database and perform exploratory data analysis.

Tools:

- PostgreSQL, Oracle, MySQL, etc. (your choice)
    - PostgreSQL: https://www.postgresql.org/download/

Part 1 (70 points):
Deliverables: You can either include screenshots of your pgadmin screen showing query and output table or copy paste your queries and output table for your answers and submit PDF version.

- Create a SQL database and separate tables for both datasets `EmployeeAttrition1.csv` and `EmployeeAttrition2.csv` using a RDBMS (PostgreSQL preferred). You need to submit create table query as well in the final document.
- Load/Import the dataset into the table.
- Query the database table for `EmployeeAttrition1.csv` and interpret the results, displaying:
    1. the count of total number of records in the table
    2. the count of records for each JobRole in descending order of count
    3. the average MonthlyIncome and PercentSalaryHike for each JobRole in ascending order of JobRole
    4. the average JobSatisfaction for each Gender and MaritalStatus
    5. the range (Min and Max) of Age and HourlyRate for each JobRole
    6. Join two tables for `EmployeeAttrition1.csv` and `EmployeeAttrition2.csv` and display 20 records with the following columns
        - EmployeeNumber, Age, Gender, JobRole, OverTime and Attrition

Part 2 (30 points):

- Connect to your EmployeeAttrition database tables in Python and load into pandas dataframe
    - Perform **three** interesting analysis on this data with visualization and tell the story about interesting insights in your analysis
        - Analysis could be anything such as univariate analysis, bivariate analysis, correlation etc.

Deliverables: Submit your pdf file for Part 1 and Jupyter Notebook ipynb file for part 2 with all the code and analysis.