Siametrics Consulting Data Scientist Pre-interview Assignment July 2018

Instructions:

This assignment contains 4 questions and 1 bonus question. Don't worry if you don't have time to do the bonus one. You **will not** be penalized for not completing it.

****Do everything in Jupyter notebook if possible. If not, make sure you send the script and output as well as the answers.*

Some questions may seem confusing and may require you to "do your best" at gauging what they actually mean. Some fields may have inconsistent naming convention. This is common in real situations. Do your best and make your best guess but remember to make a note about why you make certain decisions.

All of these should not take you more than 2 hours.

Please complete the assignment on your own.

You have until **Saturday August 4th** to finish.

Please email all of your answers and code to contact@siametrics.com with subject "Data Scientist Assignment." We will get back to you as soon as we can.

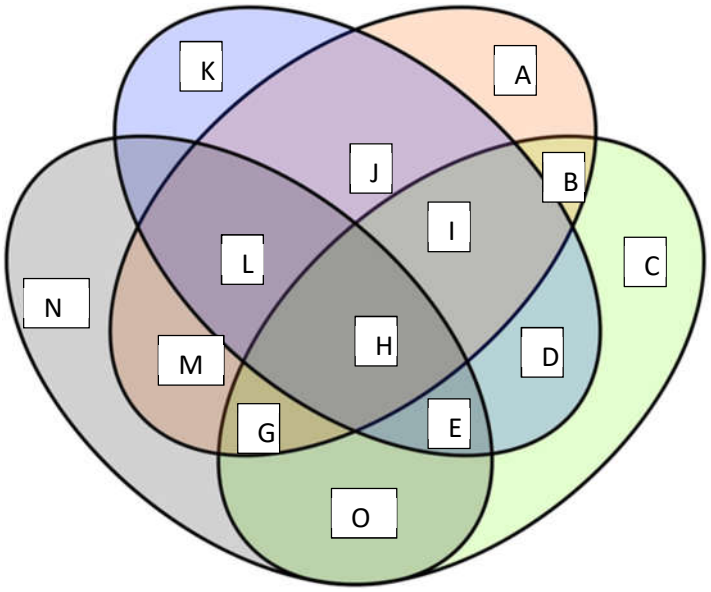
Data Sets:

****These are fake data.*

1. Performance.csv
    a. This contains performance measures (a and b) for each employee. person_id is a unique identifier for each employee.
2. Performance_c.csv
    a. This contains another type of performance measure (c) for each employee.
3. Absence.csv
    a. This contains absences data for each employee
4. Department.csv
    a. This contains information about which department each employee works at. id_s_fr is department id.

Q1: Load the four raw data sets, merge them all together by *person_id*, *id_s_fr*, and *year*. Then, fill every area in the below Venn diagram with the number of unique people.

For example, area A contains the people who **only** appear in the job performance data. Area H contains people who show up in **all** four data sets. And area J contains the people who appear in all data sets but the performance_c data set. You get the point. Save the final data file as *panel.csv*.



Blue (absence.csv):

Orange (performance.csv):

Green (performance_c.csv):

Grey (department.csv):

Q2: Are there any features we should worry about? Check for missingness and other anomalies. Report the top three issues with the data I should worry about.

Q3: Create a new variable for a person's average lifetime performance score (pick any of the three scores we have). Call this new variable, *lperf*. What is the average of *lperf* among female employees?

Q4: Transform the data from the current long format (each row is for each person-year record so there'll be many rows per one person) into the wide format (each row is for each person). Save this as *wide.csv*.

Q5 (BONUS): Go back to the long format file, *panel.csv*. Calculate the rolling average performance score (pick any of the three) for all previous years up until the year for that record. For example, if that row is year =2014, calculate this average using performance score from all years <=2014. If there are missing scores in some year(s) in between, use the latest available rolling average. Call this variable, *rperf*. What's the average *rperf* among male employees in year 2015?