LIPSCOMB, COLLEGE OF COMPUTING AND TECHNOLOGY

MSDS 5013 – Principles of Data Science

Abstract of Hypothetical Problem for Data Scientist
SQL, Python, and Basic Analytics

Due: September 20 at 5:30pm

1 DESCRIPTION

You are new data scientist for a the **Planetary Best Health Care Insurance Company Incorporated (PBHCICI)** which is looking into expanding service to new non-human creatures on newly discovered planets in outer space, named Boran and Radan. As such, your manager has obtained a small preliminary data set for you to perform experiments upon while the actual data set is being reviewed and verified. The data include the number of earth years of life for each alien creature, and several medical diagnostics taken 6-months prior to the alien's death.

GOALS

At a minimum, you will need to perform the following:

- Define the properly planned schema of the Alien Planet Medical Database for SQL,
- Populate the schema using the given data files using the Python SQL interface. This will require some transformation of the data files, and probably creating some temporary files or tables.

(Adapted from the 2016 MS5013 final project by Dr. Tim Wallace.)

- Check that the database enforces the appropriate consistency and validity checks by performing a series of appropriate updates that should be rejected by the database.
- Construct a set of queries for the database to generate the required SQL joins for the arrays needed to compute the analytical features as a function of the de-identified patient ID.

2 Details

You must develop prototype code in Python or iPython notebook to implement the assignments below as part of your data science group effort. As such, the provided data set may be updated within the next few weeks if additional data arrive in time for your analysis, so please make certain you can easily re-run your Python program to regenerate the results of the assignment using the new data set when provided per the below assignments. Placing the Python code in scripts along with database / data in a folder is a nice way to save the methods for future reuse.

3 Project Assignment

Your class assignment is to complete as many levels of the following project plan as follows:

1. Problem Analysis: From inside Python, create schemas needed to import the CSV data into respective database holding tables for each of the two planets, with proper segmentation according to planets; select the primary and/or foreign keys for each table as appropriate. All planets' data may be placed inside a single database instance, with proper multiple table management.

   You must design the database with appropriate choices of:

   a) Attribute constraints (basic data types and additional constraints, such as NOT NULL, CHECK constraints, or DEFAULT values)

   b) Keys. Choose appropriate attributes or sets of attributes to be candidate keys, and decide on the primary keys.

   c) Referential Integrity Constraints. Determine all foreign keys, and decide what should be done if the tuple referred to is deleted or modified.

2. Using Python, write the appropriate SQL query to select data attributes using blood pressure and age attributes for the kth planet. Transfer the results of the query into appropriate Python array variables. This step will include:

   a) Find the best linear regression fit of blood pressure to age relationship,

$$Y_k = m_k X_k + B_k$$

where m_k and B_k are to be determined for the kth planet, P$_k$ and where X_k is the age random variable. Use blood pressure and age as dependent and independent variables respectively. This step will require the proper identified data from an appropriate SQL join.

b) Repeat the above for each planet.

3. Generate plots for the histogram age distribution for each planet, P$_k$ using the http://matplotlib.org/ function calls from Python.
   a) What is the mean (average) life expectance of the creatures on each planet?
   b) What is the probability of a creature living past the mean life expectancy on each planet?

4 Data

Data are provided in three CSV text files: *boran.csv*, *radan.csv*, and *deidentify_list_cross_ref.csv*. You must design the appropriate schemas, create the database and tables, and import the data using Python and sqlite3 DB interface code as was demonstrated in the tutorial in class (URL http://pymotw.com/2/sqlite3/) or using other references, such as Grus book provided.

4.1 Data Conversion

Each of these files could be converted directly into a relation in the database. However, this would not constitute a complete design, without some discussion, planning, and analysis.

4.2 Data Features

In *boran.csv* and *radas.csv*, each record contains: patient ID, blood pressure, exercise, weight, glucose, BMI, and planet ID.
In *deidentify_list_cross_ref.csv*, the features are patient ID (may not be unique) and patient age.

5 Report Format

The report should be formated in AAAI style. You can download the template from the AAAI 18 site

6 Presentation of Project

Each group will be required to present their results and provide discussion of hypothesis and conclusions on the last day of class. The projector/HDTV will be available with HDMI in the Spark classroom. Teams are encouraged to demonstrate running programs of various components or analytics developed using iPython notebooks.

NOTES

1. *During the work on this project, one goal is to try to determine a statement of hypothesis, prediction, and suitable tests to evaluate the appropriate hypothesis.*

2. *Data files have been regenerated without the tabs. This will ensure you do not have to remove the tabs so that Python will not require tab removal prior to numerical processing during linear regression.*

3. *Example SQL queries for a given database are shown in* test-queries.py