Assignment 3
Due Nov 25th

SQLite or SAS (with PROC SQL) can used as a database.
This work is to be done individually and submitted through webcourses.

There are two datasets GDP.csv and GEP.csv . The files contain country names and numerical data. They come from the worldbank databank. The GDP is the gross domestic product for each country and the GEP is the Global Economic Prospect for each country over a set of years in the past's prediction and into the future. The headers have been removed but for the GDP.csv they are:
Country code, Country name, gdp

for GEP.csv:
Country Name, CountryCode, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018

Question 1 (20pnts)

a) Load in the data from CSVs into SQLite. Where each CSV goes into a new table. You can handle the column names as you see fit. (4 pnts)

b) Produce a histogram for the GDP values in the dataset after extracting the values from the database. (1 pnt)

c) Use SQL to select from GEP the countries (United States,USA), (Greece,GRC), (China,CHN), (United Kingdom,UK), (Argentina,ARG) and print the mean GEP. For each. The plot the GEP values over the years (5 counties will produce 5 lines). (2 pnts)

d) Select the countries which have an above average GDP using SQL. (1 pnt)

e) Select all the countries whose country name starts with a letter 'G' (1 pnt)

d) Join the 2 tables together using SQL. (3 pnts)

e) Using the data from the joined table as training data you will try to predict the GEP values from he file GEPsupplementRecent.csv which now has a recent addition of GEP information. Using a subset of the rows for training and the remaining rows as testing calculate the MSE or RMSE for your model predictions. Use 2 different models (eg linear regression and random forests) and then compare the MSE or RMSE values (5 pnts)

f) If you remove the countries with below average GDP, how does the RMSE or MSE change the predictive quality of the models (discuss). (3 pnts)