CSC 424, Homework 1
Dr. John McDonald

Deliverables: Turn in your answers in a single PDF file. Each answer should be submitted as a full analysis with the R commands used to create the analysis. Copy any R commands, short output relevant to your answer and graphs into your document. Explain your answer thoroughly but concisely.

Section 1, To be turned in: Make sure that software output is clearly indicated, and your explained answers are clear and easy to find. Note: **R is the official course software,** I will not be providing support for any other software. There are discussion forums provided for other packages, but **they are provided for student discussion only.**

1. (20 points, **can be completed in any software package**) **Looking beyond the surface.** While we will be spending a great deal of time learning more sophisticated model building techniques, often the immediate "parameter of interest" that seems to appear in a dataset is not the most interesting/impactful one, nor are the original variables necessarily in a form that is best for modeling the data. This problem asks you to look below the surface to find a story in the data that is more interesting than the obvious one.

   The Olympics data set concerns the performance of various countries in the 2012 London Summer Olympics. For each country, the data contains separate medal counts, number of athletes by gender, national population figures, and national gross domestic product (GDP). The obvious surface message in the data is that larger countries/teams with higher GDP generally win more medals. It is your job to distill an interesting story or insight in this data, but it should be something other than the obvious positive relationship between raw or aggregate medal counts and population/GDP.

   It will take some investigation to find a suitable message, and you should look at **several relationships, and consider transforming/creating some new variables based on the original variables** before settling on one. There are several opportunities for interesting analyses in this data and you do not need to investigate all of them. Think about whether there is an important trend or lesson that you would like the public to understand? Below are some things to consider. You do not have to investigate all of these. They are provided to help you think about the data.

   a. Do any surprises emerge? Often, the most interesting results are surprising ones because they tell us things we didn't expect.
   b. Are there any transformations or ways of combining variables that can reveal more subtle patterns than simply overall population/GDP? What about per-capita or per-participant measures? Is there any relationship between participation of certain demographics and the country's performance?
   c. Imagine you are an Olympic coach for a small country, what does performance mean if your country has limited resources? The GDP/team size vs. medal count relationships merely say, "grow your economy and increase your team size (larger budget) to win more medals." Is there a way of marginally improving the performance of the athletes you have (how would you measure that)? Even if the resulting model does not have a high R^2, it can still be practically significant.
   d. Are there ways to evaluate a country's "performance" beyond medal counts? Are there any relationships that have nothing do to with medal performance that are interesting or impactful?
   e. Sometimes the most interesting results are not in the nature of the model but in the **nature of the outliers.** These outliers can suggest directions for future study.

   One note: be very careful about multicollinearity (correlation among predictors) in this dataset. In other words, if you are using two predictors that are highly correlated, remember what it can do to the slopes in a regression. **You will not receive full credit if your model(s) contain a significant multicollinearity!**

   You may try different multiple-regressions and plots and can compare these results to automatic variable selection methods, but your writeup should only include your best, most interesting analysis. Be thorough but concise in your write-up and be sure to include the graph(s) and analyses you are using to see the

relationships and **clearly indicate** the intended message of your analysis. There are many possible relationships to consider but you will be graded on the clarity and the thoroughness of your graphs and written analysis. You should be able to fill at least a page.

2. (25 pts, **to be completed in R**) The Housing dataset housing.csv contains a modified version of a dataset of housing values in the suburbs of Boston from the UCI machine learning repository http://archive.ics.uci.edu/ml/datasets/Housing. One parameter has been dropped from the original dataset due to its very slight contribution to the parameter of interest and its biased and mathematically flawed nature.[1]

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. LSTAT: % lower status of the population
13. MEDV: Median value of owner-occupied homes in $1000's

a. (5 points) Fit an initial linear regression model of MEDV based on all the other variables and report R^2, Adjusted R^2, the utility of the model (F-Test), the estimated coefficients, their standard errors, and statistical significance. Interpret your results. Treat the RAD ordinal variable as numeric.

b. (5 points) Plot the dataset in a scatterplot matrix and also the correlation with a corrplot. Interpret the result. Are there variables whose correlation with MEDV are weak? Are their variables whose relationship to MEDV are non-linear, or for which a log transform should be applied (look for a lot of samples on the axis with relatively few at high values)? Look for at least two transformations to apply that can increase the R^2 value of the regression. Transform the variables, rerun the regression, and compare the results to the initial regression.

c. (5 points) Perform a feature selection on the transformed data by using the stepwise selection method of the regression analysis. Which variables are dropped in the stepwise selection model and how is the adjusted R^2 affected? Evaluate the result in comparison to the full model.

d. (5 points) Perform an all-subsets analysis with "regsubsets" (set the "nvmax" parameter high enough that the search will include the regression with all the variables). Write out the model as an equation, plot and interpret the results (using the adjusted R^2 value on the vertical axis). What variables are dropped in the "best" model and how does it compare to the stepwise model? Leave the parameter "nbest" at its default of 1 to reduce the complexity of the graph.

e. (5 points) Suppose you were trying to find parsimonious model (i.e. as few features as possible) to make the result easier to explain and use practically. Investigate the graph of the regsubsets result and determine if there is a model that reduces the number of variables significantly without significantly reducing adjusted R^2 (more than a percent or two). Explain your choice, and discuss which variables are included in the model? Compute that model with lm and interpret and compare the model practically with the stepwise model in terms of the effect of each variable on median house price.

[1] See racist data destruction?. a Boston housing dataset controversy | by Michael Carlisle | Medium

3. (20 points) Perform, by hand, the following calculations from linear algebra. For the following matrices and vectors. Submit a **clear and easy to read scan or photo.** Particularly if you are using a cell phone, make sure that your page is well framed and is square with the camera. If your text is clipped off, blurred or taken at an angle that makes it difficult to read, it will not be graded.

$$Z=\begin{bmatrix}1 & 4\\ 1 & 3\\ 1 & 2\\ 1 & -5\end{bmatrix}, Y=\begin{bmatrix}2\\ 1\\ -1\\ 3\end{bmatrix}, M=\begin{bmatrix}20 & 5 & 0\\ 5 & 25 & -10\\ 0 & 10 & 5\end{bmatrix}, N=\begin{bmatrix}-20 & 0 & 10\\ 5 & 10 & 15\\ 5 & 20 & -5\end{bmatrix}, v=\begin{bmatrix}-1\\ 1\\ 3\end{bmatrix}, w=\begin{bmatrix}2\\ -1\\ 1\end{bmatrix}$$

   a. $v \cdot w$ (dot product)
   b. $-3 * w$
   c. $M * v$
   d. $M + N$
   e. $M - N$
   f. Z^T
   g. $Z^T Z$ (Make sure you get the right dimensions on this matrix)

4. (10 points) In R, write a script to compute each of the parts in problem 4 to check your answers. Submit both the R code and the output. Make sure you correct any discrepancies in the answers, as they indicate that you have an issue either in your manual calculation or in your R code.

5. (Due separately online, see the final project milestones) Complete the first milestone for the final project by posting an interesting dataset or continuing someone else's discussion.

Section 2: Practice Problems (not for turn-in, but if you have any questions, make sure to ask)

1. If $M = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}$ and $M = \begin{bmatrix} -4 & 3 \\ 2 & 8 \end{bmatrix}$, what is the matrix $M - N$?

   a. $\begin{bmatrix} -1 & -1 \\ -3 & -7 \end{bmatrix}$
   b. $\begin{bmatrix} -1 & 5 \\ 1 & 8 \end{bmatrix}$
   c. $\begin{bmatrix} 7 & -1 \\ -3 & -7 \end{bmatrix}$
   d. $\begin{bmatrix} 7 & 5 \\ 3 & 9 \end{bmatrix}$
   e. None of these

2. What is the product of the following two matrices? $\begin{bmatrix} 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}$

   a. The two matrices cannot be multiplied because their sizes don't match
   b. $\begin{bmatrix} 7 \\ -4 \end{bmatrix}$
   c. $\begin{bmatrix} 6 & 1 & 0 \\ -3 & 0 & -1 \end{bmatrix}$
   d. $\begin{bmatrix} 6 & -3 \\ 2 & 0 \\ -2 & -1 \end{bmatrix}$

3. If $v = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$ and $w = \begin{bmatrix} 2 \\ -1 \\ -4 \end{bmatrix}$ what is $v \cdot w$?

   a. $\begin{bmatrix} 2 \\ 3 \\ -8 \end{bmatrix}$
   b. 13
   c. -9
   d. -3
   e. 3

Answers to practice problems: 1) c, 2) b, 3) d