CSC334/424: Assignment #1
Due: Sunday, June 17, 2018 (by midnight)
Total: 50 points

Problem 1(5 points – Due Monday, June 11, 2018 at 5PM) Introduce yourself on D2L by posting to the Class Introductions forum on D2L. Include a bit of information about yourself including some of the following. Note, this

- Name
- Undergraduate Degree
- Major/Degree Program(Concentration)/Time in Program (e.g. 3rd quarter, 2nd yr, graduating this quarter)
- Position at Work, if applicable
- What is your experience with R? Have you used it for any courses? For work?
- What interests you about Advanced Data Analysis?
- Field(s) of Interest and/data
- Hobbies

Problem 2 (10 Points): Post to the final project forum with the following:

- Project Type: Data / Literature Review
- Project Team: Individual / Group Members
- For Literature Review:
    - Topic to be Covered
    - Draft Outline of Techniques to be Covered
- For Data Project:
    - Subject Area or Field of Interest
    - Source of Data
    - Specific dataset(s)
    - description of its scope (# metric variables, #categorical variables, #samples, multiple related tables?)
    - Technology group plans to use for Project

    - In addition, as you are forming your groups, remember the following requirements for datasets and groups:

        a. Your group should have 4-5 people in it.

        b. Your group should have at least one **in-class** student. This helps me check in with each group if I have at least one in-class student in each group.

        c. Your dataset should be a real and rich dataset with at least 15 to 20 variables mixed between categorical and metric. It should have at least (10 * #var, but better yet 20 * #var) samples (we will see that some techniques like PCA require this for significance/stability). You will need a large sample size if you have a large n umber of variables. See me if you have any doubts about your dataset.

Problem 3 (10 points) Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

$$Z = \begin{bmatrix} 1 & 4 \\ 1 & 6 \\ 1 & -3 \\ 1 & 2 \end{bmatrix}, Y = \begin{bmatrix} -1 \\ 1 \\ 7 \\ 8 \end{bmatrix}, M = \begin{bmatrix} 1 & 5 & 0 \\ 30 & 50 & 15 \\ 0 & 20 & 10 \end{bmatrix}, N = \begin{bmatrix} -20 & -5 & 0 \\ 0 & 20 & 5 \\ 20 & 10 & 20 \end{bmatrix}, v = \begin{bmatrix} -2 \\ 2 \\ 6 \end{bmatrix}, w = \begin{bmatrix} 3 \\ -7 \\ 5 \end{bmatrix},$$

a. $v \cdot w$ (dot product)
b. $-3 * w$
c. $M * v$
d. $M + N$
e. $M - N$
f. $Z^T Z$ (Make sure you get the right dimensions on this matrix)
g. $(Z^T Z)^{-1}$ (You do not have to perform Gaussian-Elimination here … i.e. this is a hint on what the dimensions of the matrix should be ☺)
h. $Z^T Y$
i. $\beta = (Z^T Z)^{-1} Z^T Y$
j. $\det(Z^T Z)$

Problem 4 (10 points –other types of regression models): There are other types of regression models outside of linear and logistic regression. **Using Google Scholar**, locate **a journal article**, which utilizes **one** of the **types of regressions** listed below or another regression outside of linear/logistic that interests you. **Write a summary** of the journal article and how it utilizes the regression model in **two paragraphs**. **Cite the paper in APA format.**

Choose one of the following regressions:

1. Ridge Regression
2. Lasso Regression
3. Elastic Net Regression
4. Polynomial Regression
5. Poisson Regression
6. Cox Regression
7. Robust Regression
8. Jackknife Regression
9. Time Series Regression

Problem 5: (15 pts – regression analysis, visualization, and interpretation): The data in the file *housedata.csv* are collected from 1,000 homes being sold in King County, Washington from May 2014 through May 2015. There are 21 columns in the data file but not all are relevant here. The response variable of interest is the Price (price of the house). The predictor variables are bedrooms, bathrooms, sqft_living (the living space area), sqft_lot (the area of the land the house sits on), floors (the number of levels of the house), waterfront (is it near the waterfront-values are higher near the waterfront), view (the number of times the house was viewed), condition (overall condition of the house), grade (provided by the King County auditor), sqft_above (area of the house excluding the basement), sqft_basement (basement area), yr_built (year the house was built), yr_renovated (the year the house was renovated), zipcode (zipcode of the home), sqft_living15 (area of the living room in 2015-could be different from above if the house was renovated, also affecting the lotsize area), sqft_lot15 (the lot size in 2015, different from above if there was a renovation), lat (latitude coordinate of the home), and long (longitude coordinate of the home). We are interested in which predictors are significant variables of determining price values of houses.

a. (5 points) Before running any regressions make sure to check for multicollinearity. How did you check for multicollinearity? If there is multicollinearity, how do you plan to resolve it? Are there any other issues with the dataset we have to consider before running the regressions?

b. Run a multiple regression of price on the variables listed above.
    i. (5 points) Run the model **using an automatic method** (i.e. stepwise, forward, backward). Explain why you chose the method. Comment on the overall significance of the regression fit. Which predictors have coefficients that are significantly different from zero at the .05 level?

    ii. (5 points) Using the variables above, **create a visualization**, which will provide an interesting story or insight within this data and present it to the general public. Is there an important trend or lesson that you would like the public to understand about housing prices within King County?