DSC 324/424: Assignment #1
Due: Sunday, July 26th, 2020 by 11:59PM (by midnight)
Total: 50 points

Problem 1(5 points – Due **Monday, July 20, 2020** 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) Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

$$Z = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 9 \end{bmatrix}, Y = \begin{bmatrix} 6 \\ 0 \\ 8 \\ 0 \end{bmatrix}, M = \begin{bmatrix} 1 & 11 & 0 \\ 11 & 09 & 35 \\ 0 & 3 & 42 \end{bmatrix}, N = \begin{bmatrix} -50 & -50 & 0 \\ 0 & 50 & 5 \\ 50 & 5 & 50 \end{bmatrix}, v = \begin{bmatrix} 45 \\ 6 \\ 9 \end{bmatrix}, w = \begin{bmatrix} 2 \\ 6 \\ 0 \end{bmatrix},$$

a. $v \cdot w$ (dot product)
b. $-3 * w$
c. $M * v$
d. $M + N$
e. $M - N$
f. $Z^T Z$
g. $(Z^T Z)^{-1}$
h. $Z^T Y$
i. $\beta = (Z^T Z)^{-1} Z^T Y$
j. $\det(Z^T Z)$

Problem 3 (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 to three paragraphs**. **Cite the paper in APA format.**

Choose one of the following regressions:

1. Ridge Regression
2. Lasso Regression
3. Elastic Net Regression
4. Poisson Regression
5. Negative Binomial Regression
6. Cox Regression
7. Robust Regression
8. Jackknife Regression
9. Time Series Regression
10. Polynomial Regression
11. Bayesian Linear Regression

Problem 4 (10 points-Data Ethics or Data Integrity): Using Google Scholar, locate **a journal article**, which discusses data ethics or data integrity in terms of big data in your field of interest. **Write a summary** of the journal article and how it utilizes data ethics or data integrity in **two to three paragraphs**. **Cite the paper in APA format.**

Problem 5: (15 pts – regression analysis, visualization, and interpretation): The insurance_dataset.csv dataset contains 1338 observations (rows) and 7 features (columns). The insurance data contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

We are interested in which independent variables are significant for **predicting the insurance expenses** by the other predictor.

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 must 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.