Instructions: students need to code these questions and submit before 4.00pm

**M.Tech-I Year-I Sem FDS Mid Lab Exam-November 2020**

**Set-1**

**Roll no. (1,3,5,8,11,14,19,23)**

1. Develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median.

(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01?

(c) Split the data into a training set and a test set.

(d) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

(e) Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

Q2. (a) Write a function, Power(), that prints out the result of raising 2 to the 3rd power. In other words, your function should compute 23 and print out the results.

(b) Create a new function, Power3(), that actually returns the result x^a as an object, rather than simply printing it to the screen. That is, if you store the value x^a in an object called

result within your function.

(c) Using the Power3() function, create a plot of f(x) = x2. The x-axis should display a range of integers from 1 to 10, and the y-axis should display x2. Label the axes appropriately, and use an appropriate title for the figure. Consider displaying either the x-axis, the y-axis, or both on the log-scale. You can do this by using log="x", log="y", or log="xy" as arguments to the plot() function.

**M.Tech-I Year-I Sem FDS Mid Lab Exam-November 2020**

**Set-2**

**Roll no. (2,4,6,9,12,15,20,24)**

1. Develop a model in order to predict whether a given suburb has a crime rate above or below the median based on the Boston data set.

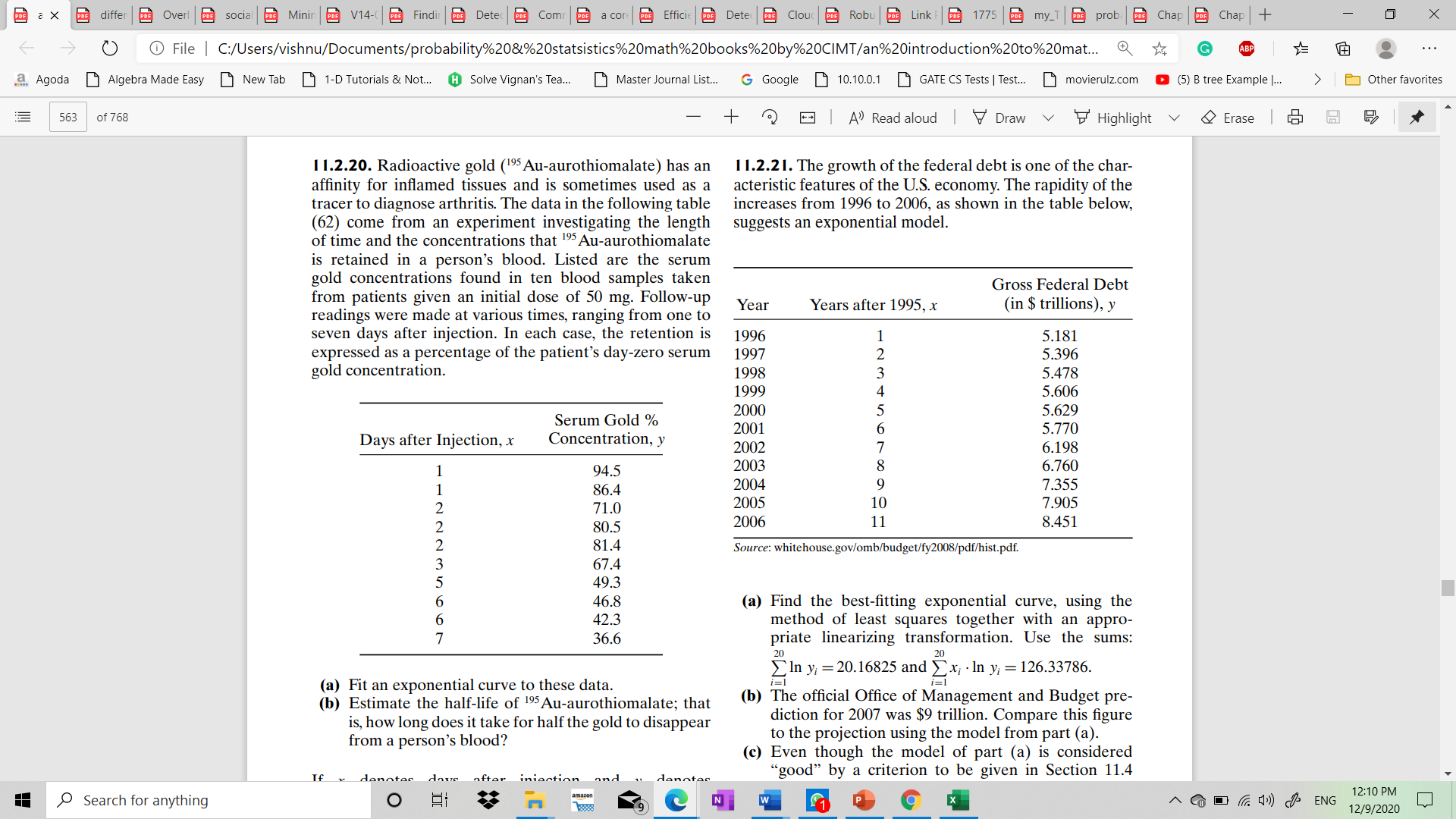
(a) Perform logistic regression on the training data in order to predict whether a given suburb has a crime rate above or below the median.

(b) Perform KNN on the training data, with several values of K, in order to predict whether a given suburb has a crime rate. Use only the variables. What test errors do you obtain? Which value of K seems to perform the best on this data set?

(c) Fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(d) Determine the confidence intervals for linear regression β0 and β1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

Q2. (i)



((ii)) Solve the following system of linear equations in five unknowns

x1 + 2x2 + 3x3 + 4x4 + 5x5 = 7

2x1 + x2 + 2x3 + 3x4 + 4x5 = −1

3x1 + 2x2 + x3 + 2x4 + 3x5 = −3

4x1 + 3x2 + 2x3 + x4 + 2x5 = 5

5x1 + 4x2 + 3x3 + 2x4 + x5 = 17

by considering an appropriate matrix equation Ax = y. Make use of the special form of the matrix A. The method used for the solution should easily generalise to a larger set of equations where the matrix A has the same structure; hence the solution should not involve typing in every number of A.

**M.Tech-I Year-I Sem FDS Mid Lab Exam-November 2020**

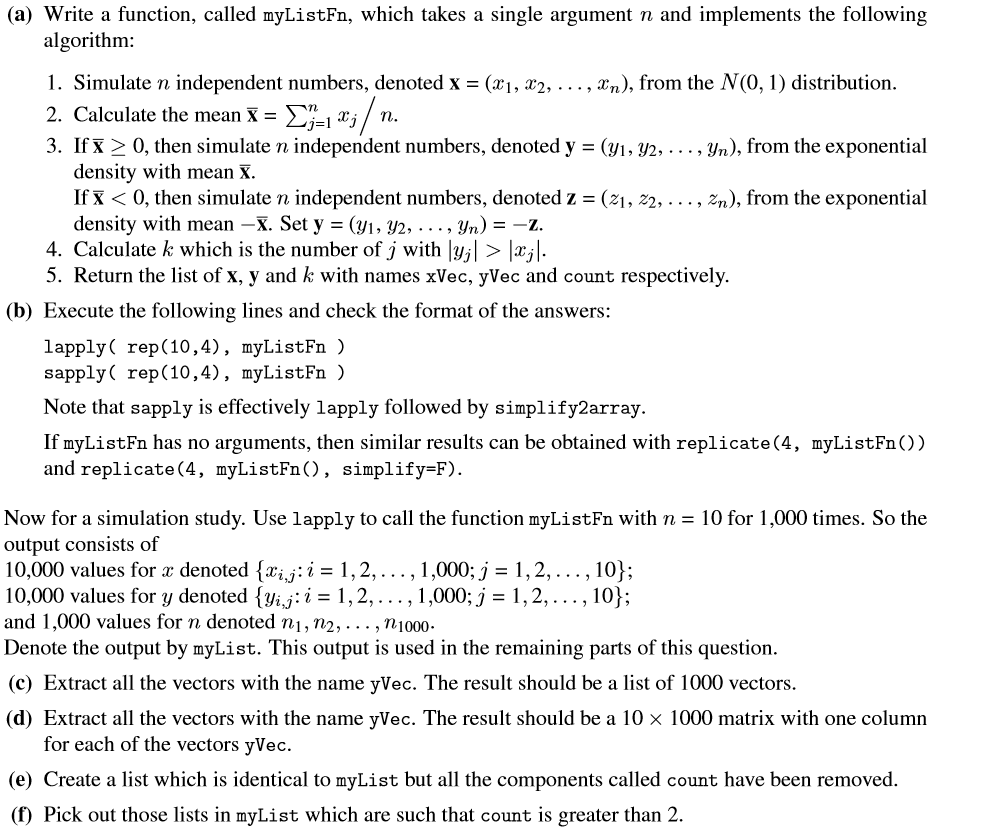
**Set3**

**Roll no. (7,10,13,16,17,18,21,22,25)**

Q1. This question should be answered using the Weekly data set, which is part of the ISLR package. It contains 1089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?
2. Use the full dataset to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
3. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. **Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).**
4. Experiment with different combinations of predictors, including possible transformations and interactions. Report the variables that provide the best results on the held out data.

Q2. (i)



Q2. (ii).

