INFO 6105
Data Science Engineering Methods and Tools

Northeastern University, Fall 2021

Problem Set 1, Due: Oct 16, 2019

Problem Set Rules:

1. Each student should hand in an individual problem set at the beginning of class.
2. Discussing problem sets with other students is permitted. Copying from another person or solution set is *not* permitted.
3. Late assignments will *not* be accepted. No exceptions.

1. (Total: 15 points)

   This exercise involves the Auto data set. Make sure that the missing values have been removed from the data.

   (a) (2 points) Which of the predictors are quantitative, and which are qualitative?

   quantitative - mpg, cylinders, displacement, horsepower, weight, acceleration, year

   qualitative - origin, name

   (b) (2 points) What is the range (e.g., minimum and maximum) of each quantitative predictor?

(c) (2 points) What is the mean and standard deviation of each quantitative predictor?

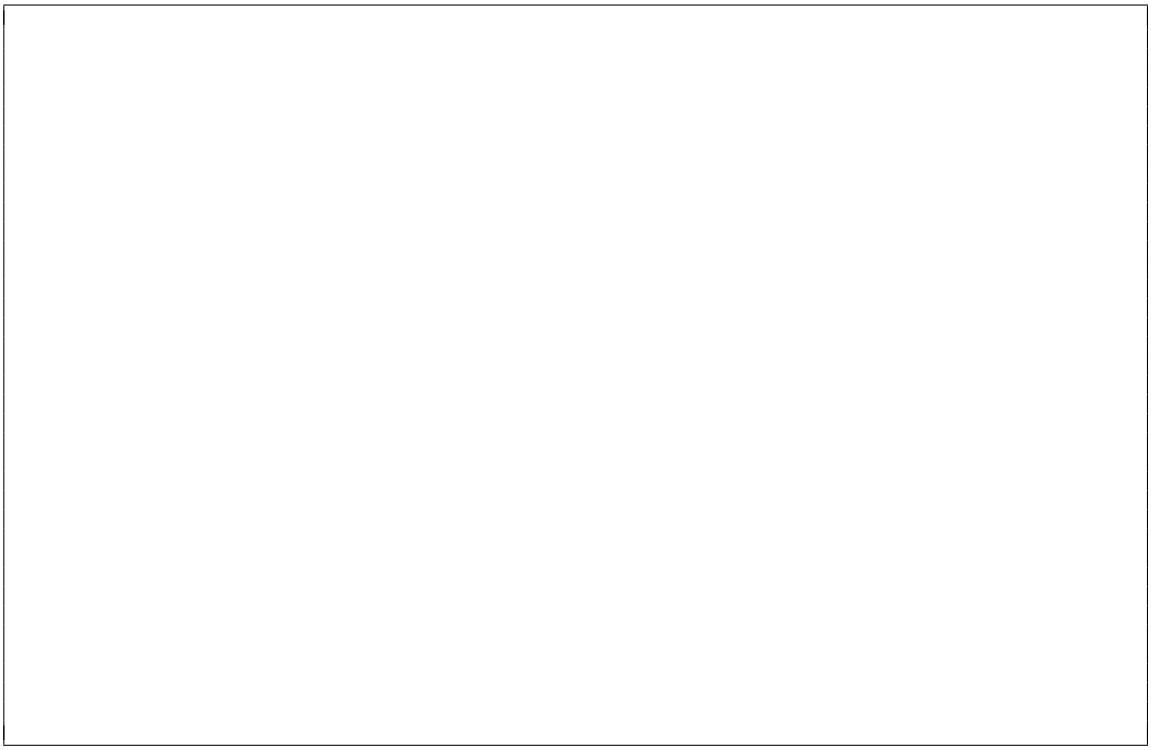
(d) (3 points) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) (3 points) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) (3 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

2. (Total: 20 points) This exercise involves the Boston housing data set.

(a) (2 points) How many rows are in this data set? How many columns? What do the rows and columns represent?

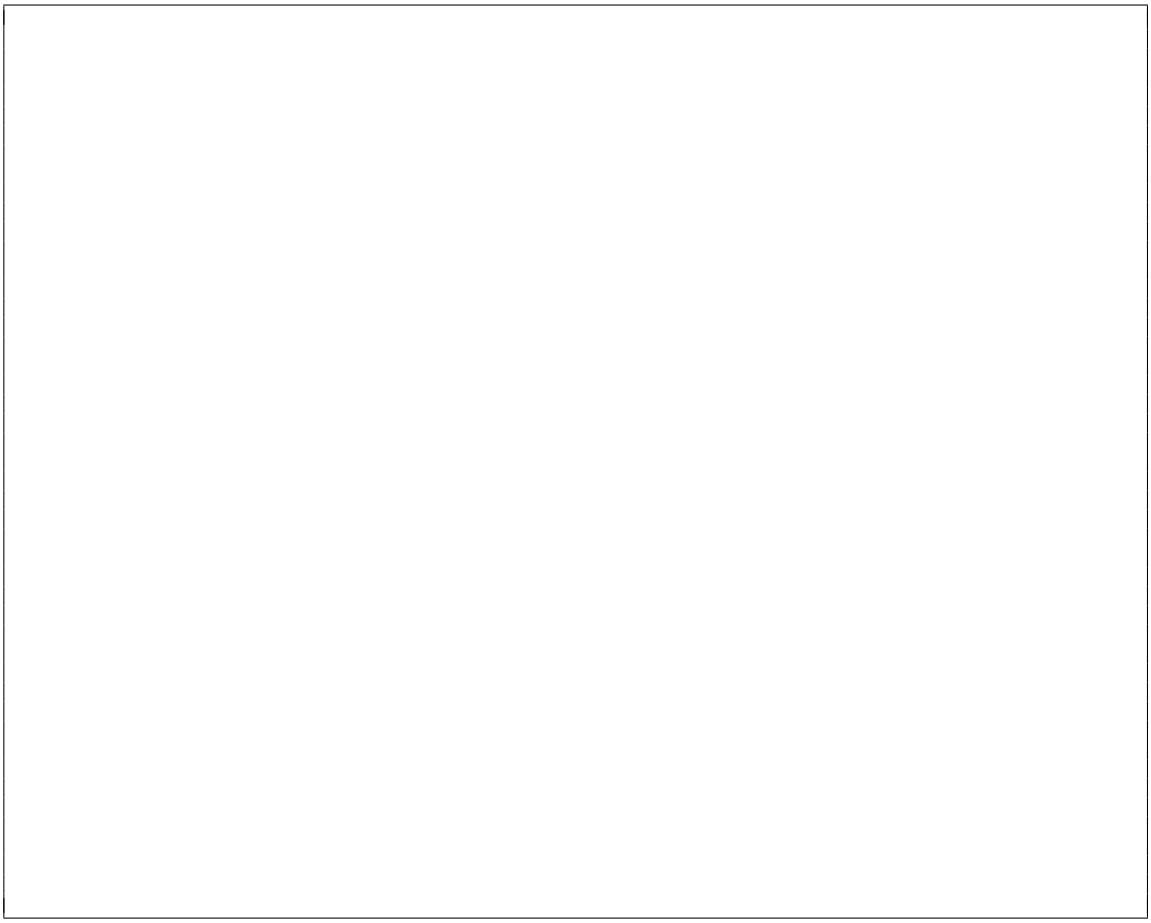


(b) (3 points) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

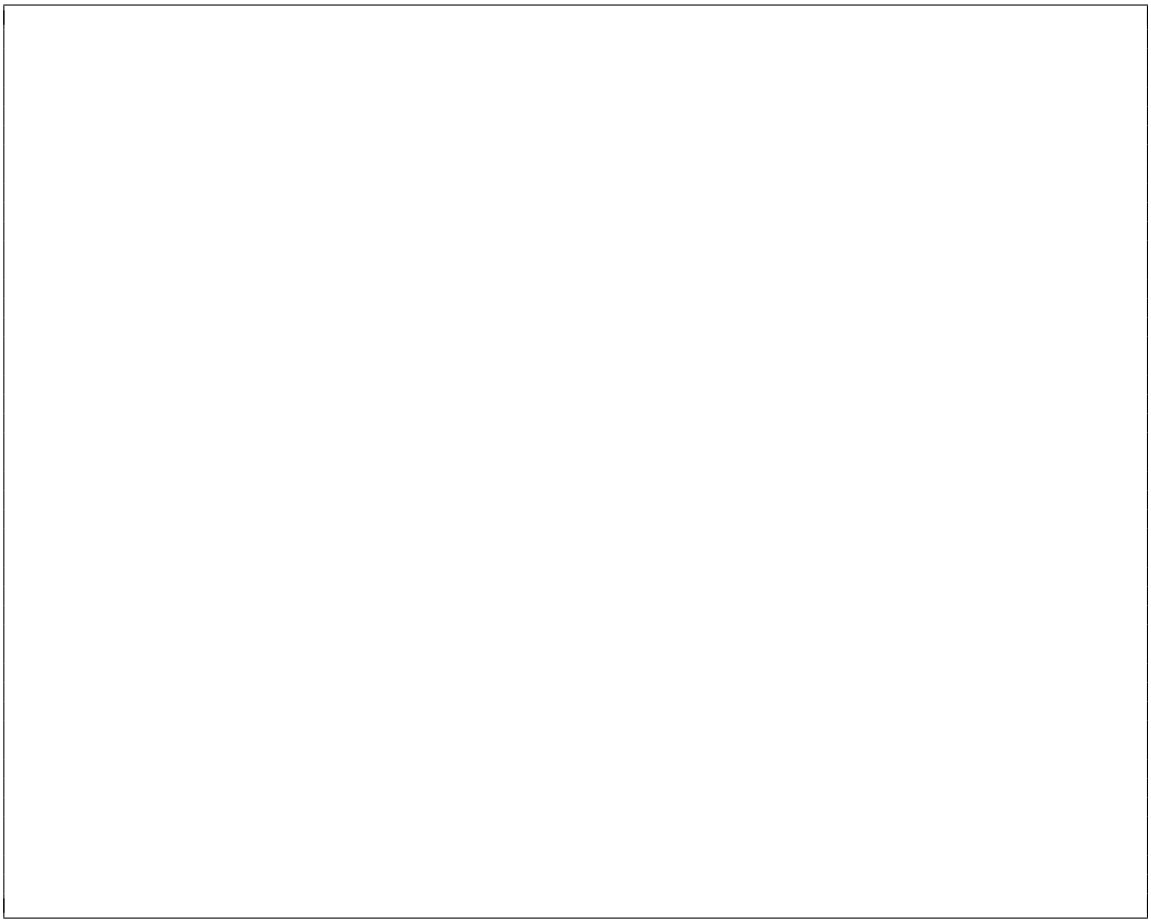
(c) (2 points) Are any of the predictors associated with per capita crime rate? If so, explain the

relationship.

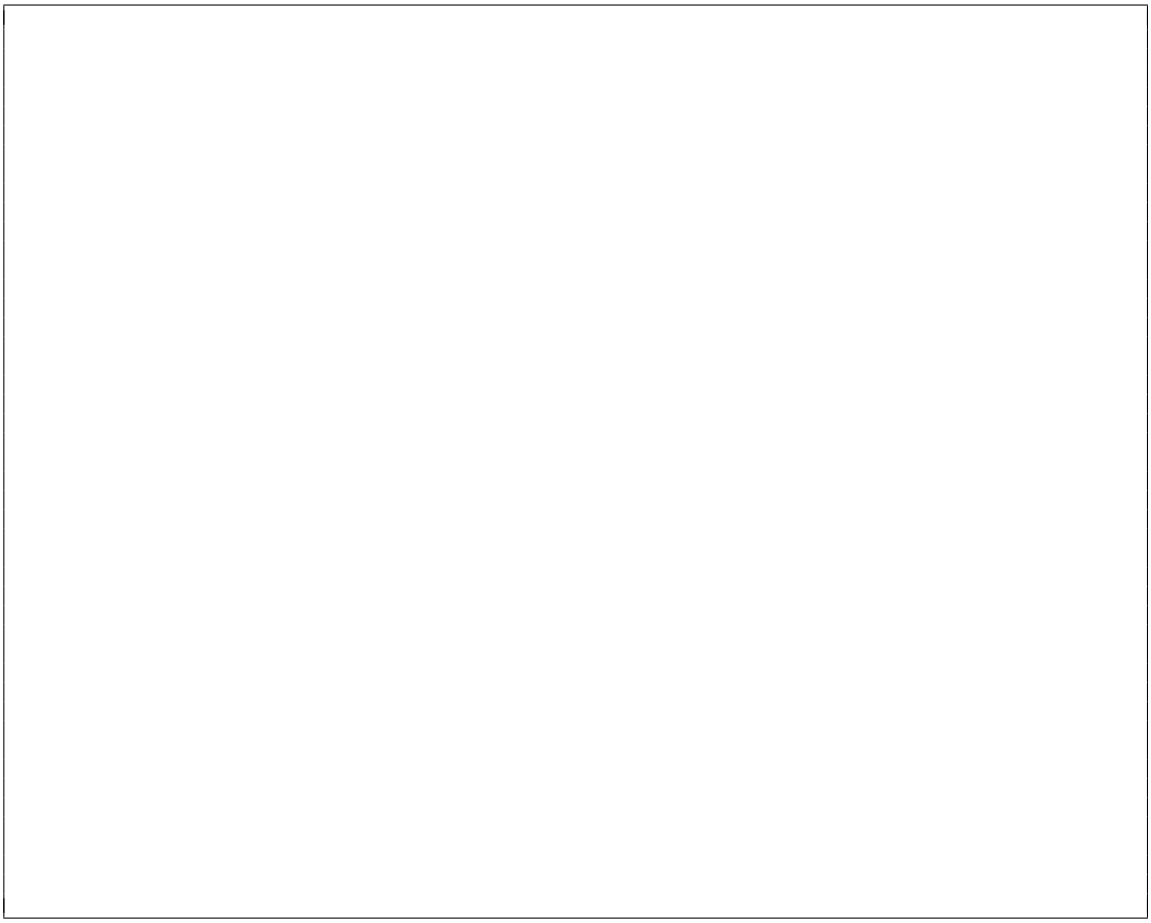
(d) (3 points) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



(e) (2 points) How many of the suburbs in this data set bound the Charles river?

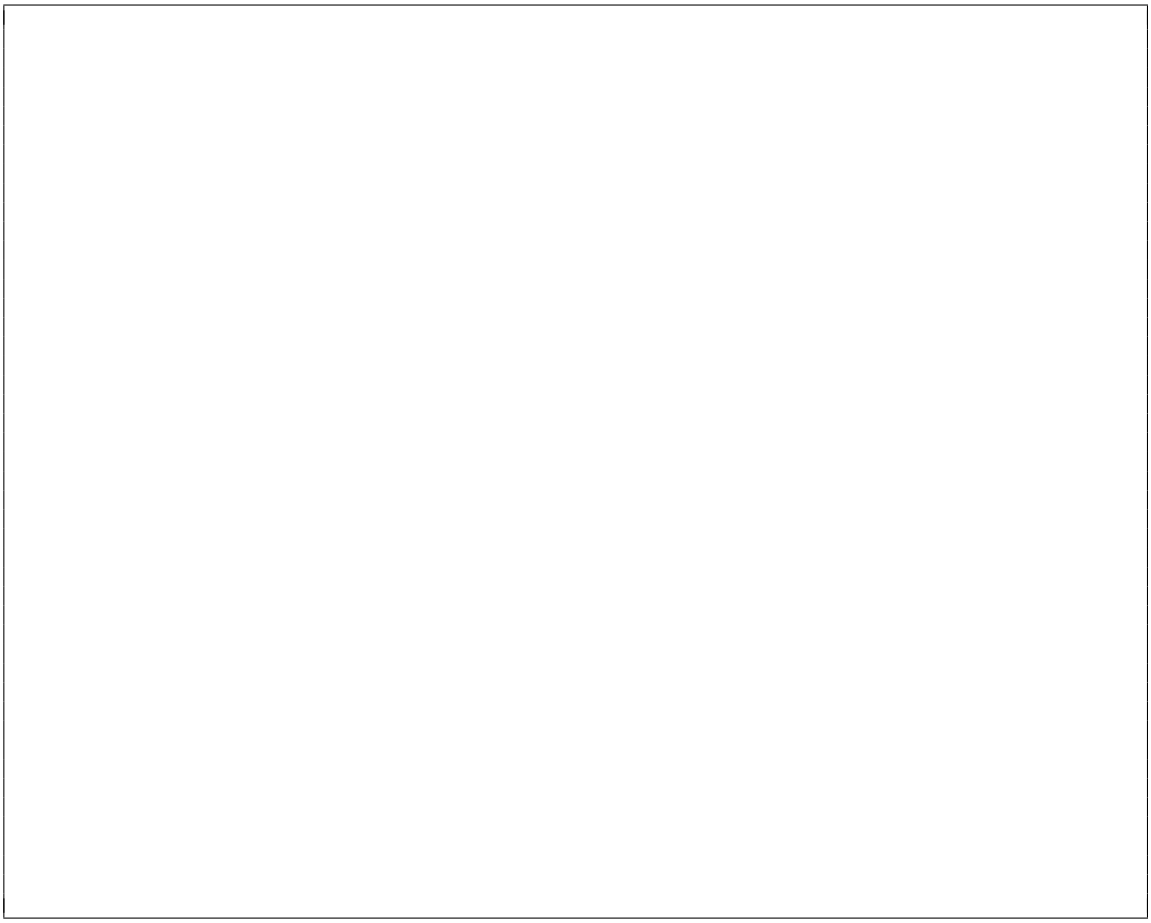


(f) (2 points) What is the median pupil-teacher ratio among the towns in this data set?



(g) (3 points) Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) (3 points) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

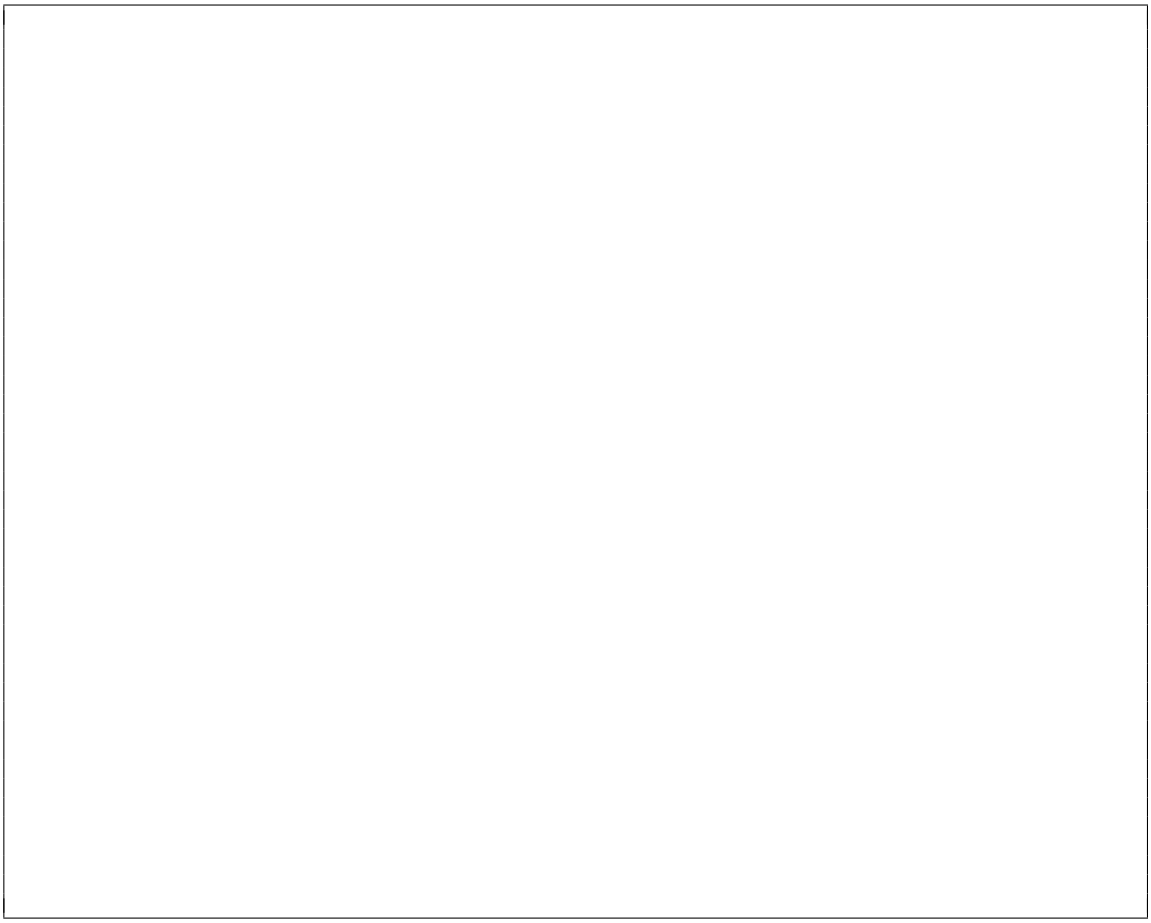


3. (Total: 24 points)

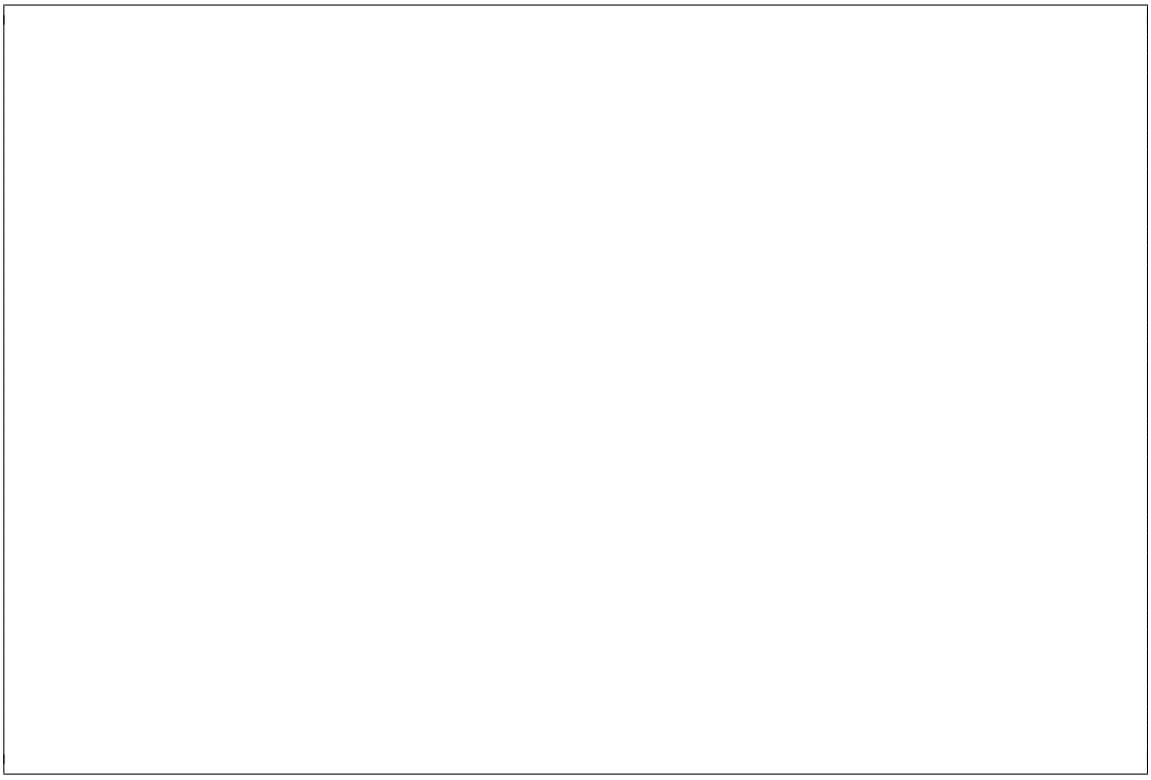
   In this question, you should use the Carseats data set to predict the sales in a new store with Price=$120, Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.

   (a) (3 points) Fit a multiple regression model to predict Sales using Price, Advertising Urban, and US. Write out the model in equation form, being careful to handle the qualitative variables properly.

(b) (3 points) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!



(c) (4 points) Using the model from (a), predict sales in the new store and calculate 68% and 95% confidence intervals.



(d) (3 points) Using the model from (a), what is the probability that sales will be greater than 12000 units in the new store?

(e) (3 points) Using the model from (a), what is the probability that sales will be between 6000 and 10000 units in the new store?

(f) (2 points) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(g) (4 points) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Using this model, predict sales in the new store and calculate 68% and 95% confidence intervals.

(h) (2 points) How well do the models in (a) and (g) fit the data?

4. (Total: 27 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

   Price (in Dollars)

   Age (in months)

   Mileage

   FuelType Fuel Type (diesel, petrol, CNG)

   MetColor Metallic color (1=yes, 0=no)

   Automatic Automatic transmission (1=yes, 0=no)

   Displacement Engine displacement (in cu. inches)

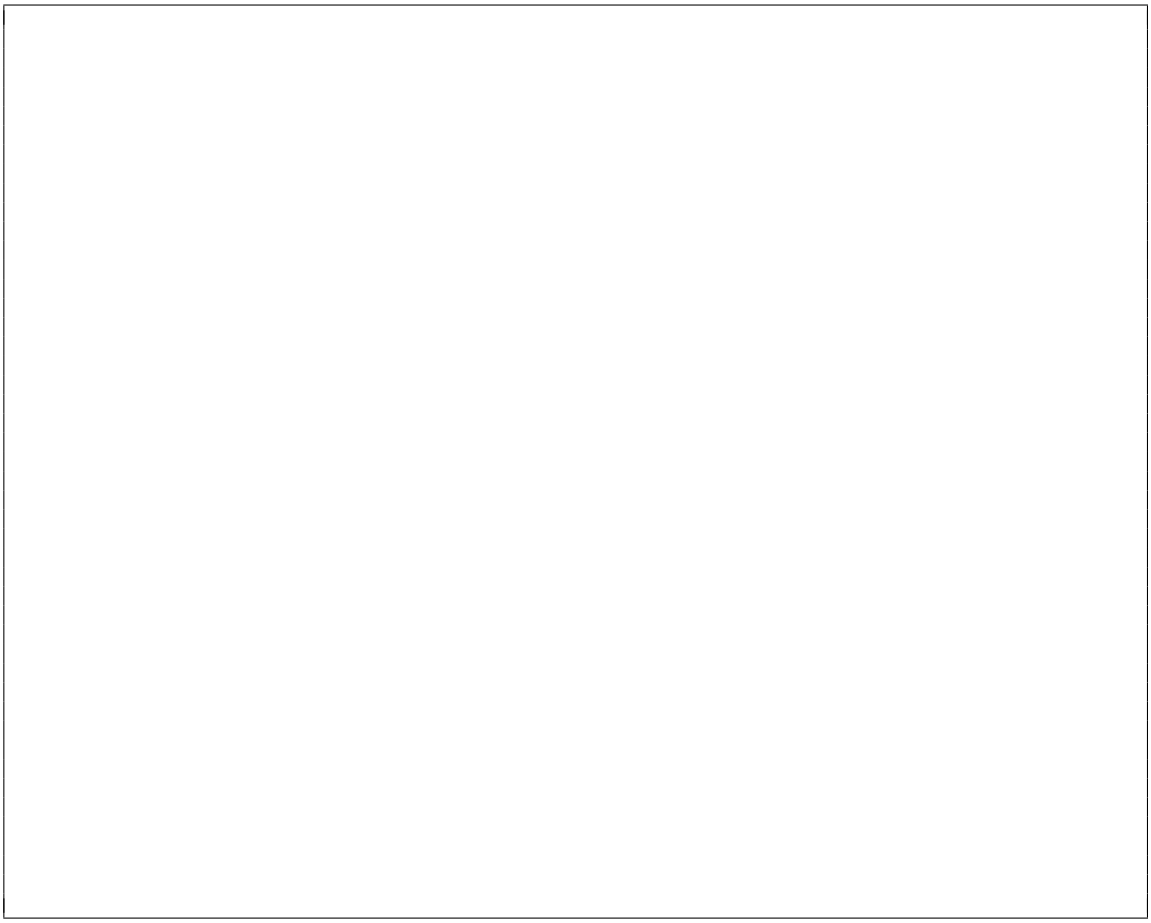
   Doors Number of doors

   Weight (in pounds)

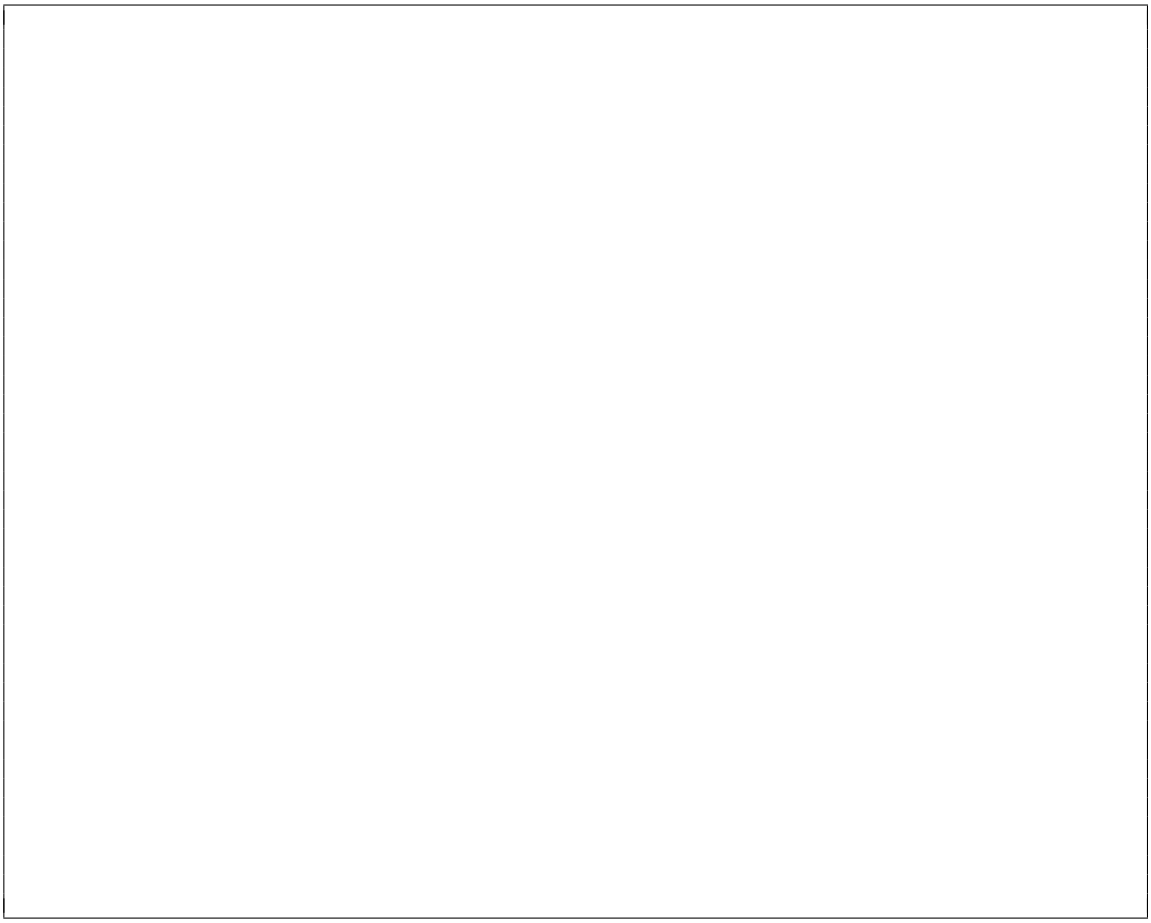
   Horsepower Engine horsepower

   (a) (2 points) Which of the predictors are quantitative, and which are qualitative?

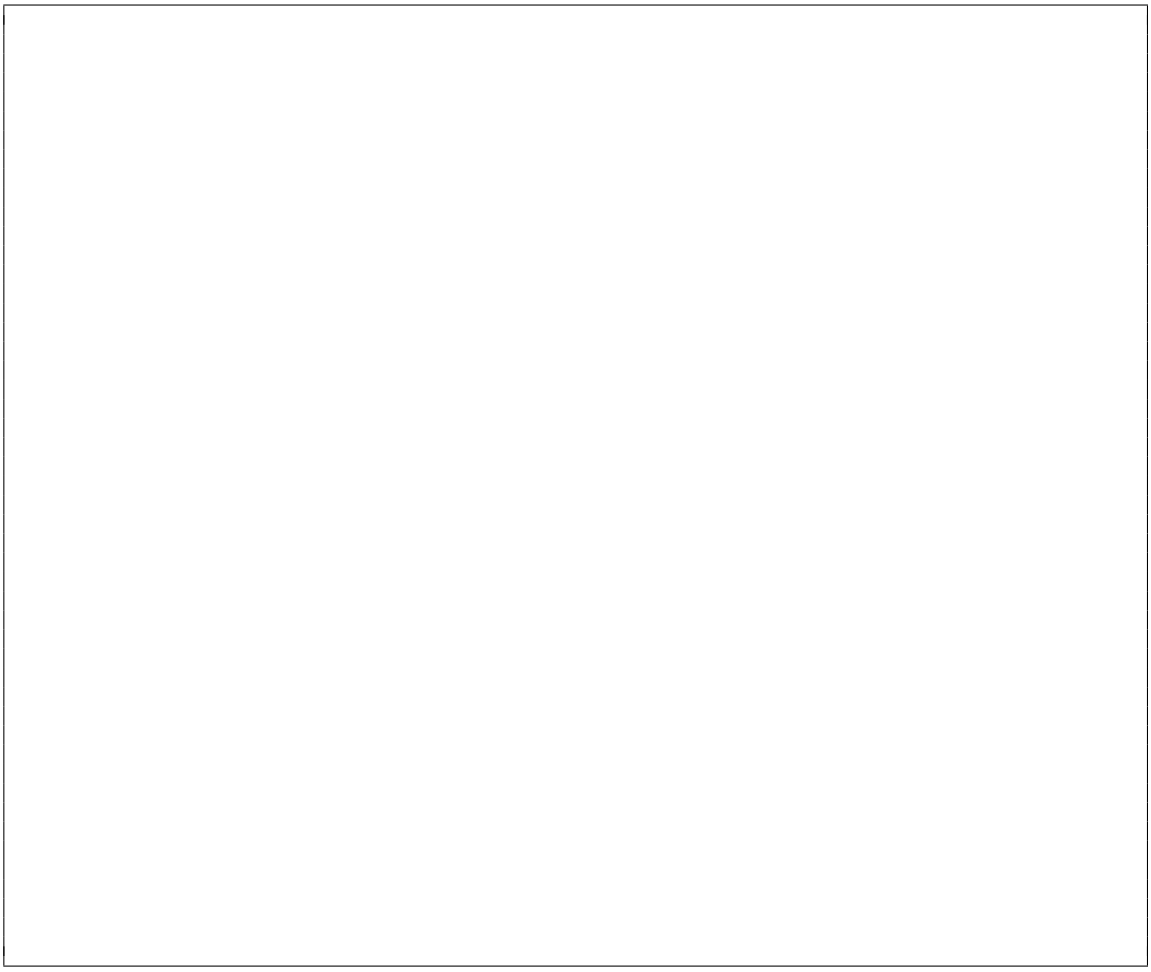
   (b) (2 points) What is the range (i.e., min and max) of each quantitative predictor?



(c) (2 points) What is the mean and standard deviation of each quantitative predictor?



(d) (4 points) Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.



(e) (4 points) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

(f) (4 points) Fit a simple linear regression with Price as the response and Age as the predictor.

(i) Is there a relationship between the predictor and the response?

(ii) How strong is the relationship between the predictor and the response?

(iii) What is the predicted price associated for a car with an age of 48 months? What are the associated 95% confidence intervals?

(g) (5 points) Fit a multiple linear regression with Price as the response and all other variables the predictors.

(i) Is there a relationship between the predictors and the response?

(ii) How strong is the relationship between the predictors and the response?

(iii) Which predictors appear to have a statistically significant relationship to the response?

(iv) What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

(v) What is the predicted price associated for a car with a mileage of 45000 miles, 48 months, diesel,

automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepower of 90, and non-metallic color? What are the associated 95% confidence intervals?

(h) (4 points) Which predictors predictors matter most for predicting the price for a car? (Find the first and the second most important variables)

5. (Total: 14 points)

This problem involves the Boston data set. We want to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) (3 points) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(b) (3 points) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) (4 points) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) (4 points) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor x, fit a model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$