Applied

8. This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are

   - Private : Public/private indicator
   - Apps : Number of applications received
   - Accept : Number of applicants accepted
   - Enroll : Number of new students enrolled
   - Top10perc : New students from top 10 % of high school class
   - Top25perc : New students from top 25 % of high school class
   - F.Undergrad : Number of full-time undergraduates
   - P.Undergrad : Number of part-time undergraduates
   - Outstate : Out-of-state tuition
   - Room.Board : Room and board costs
   - Books : Estimated book costs
   - Personal : Estimated personal spending
   - PhD : Percent of faculty with Ph.D.s
   - Terminal : Percent of faculty with terminal degree
   - S.F.Ratio : Student/faculty ratio
   - perc.alumni : Percent of alumni who donate
   - Expend : Instructional expenditure per student
   - Grad.Rate : Graduation rate

   Before reading the data into Python, it can be viewed in Excel or a text editor.

   (a) Use the pd.read_csv() function to read the data into Python. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

   (b) Look at the data used in the notebook by creating and running a new cell with just the code college in it. You should notice that the first column is just the name of each university in a column named something like Unnamed: 0. We don't really want pandas to treat this as data. However, it may be handy to have these names for later. Try the following commands and similarly look at the resulting data frames:

```
college2 = pd.read_csv('College.csv', index_col=0)
college3 = college.rename({'Unnamed: 0': 'College'},
                          axis=1)
college3 = college3.set_index('College')
```

This has used the first column in the file as an index for the data frame. This means that pandas has given each row a name corresponding to the appropriate university. Now you should see that the first data column is Private. Note that the names of the colleges appear on the left of the table. We also introduced a new python object above: a *dictionary*, which is specified by (key, value) pairs. Keep your modified version of the data with the following:

dictionary

```
college = college3
```

(c) Use the describe() method of to produce a numerical summary of the variables in the data set.

(d) Use the pd.plotting.scatter_matrix() function to produce a scatterplot matrix of the first columns [Top10perc, Apps, Enroll]. Recall that you can reference a list C of columns of a data frame A using A[C].

(e) Use the boxplot() method of college to produce side-by-side boxplots of Outstate versus Private.

(f) Create a new qualitative variable, called Elite, by *binning* the Top10perc variable into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
college['Elite'] = pd.cut(college['Top10perc'],
                          [0,0.5,1],
                          labels=['No', 'Yes'])
```

Use the value_counts() method of college['Elite'] to see how many elite universities there are. Finally, use the boxplot() method again to produce side-by-side boxplots of Outstate versus Elite.

(g) Use the plot.hist() method of college to produce some histograms with differing numbers of bins for a few of the quantitative variables. The command plt.subplots(2, 2) may be useful: it will divide the plot window into four regions so that four plots can be made simultaneously. By changing the arguments you can divide the screen up in other combinations.

(h) Continue exploring the data, and provide a brief summary of what you discover.

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

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

(b) What is the *range* of each quantitative predictor? You can answer this using the min() and max() methods in numpy.

.min()
.max()

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

(d) 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) 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) 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.

10. This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set, which is part of the ISLP library.

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

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

(d) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(e) 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.

(f) How many of the suburbs in this data set bound the Charles river?

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

(h) 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.

(i) 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.