Assignment Dataprocessing HT24

This assignment is taken from the book ”Statistical Mahine Learning” that we will use in later courses. Hand in a jupyter notebook (.ipynb) containing the relevant markdown and code to complete each part. Use a separate cell for each part. The assignment relates to the College data set, which can be found in the file `College.csv` on the assignment 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
- Top25perc : New students from top 25
- 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

For a passing grade (G) parts (a)-(e) must be completed. To pass with distinction (VG) all parts must be completed.

(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 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. Overwrite your modified version to the data with the following: `college = college3`

(c) Use the `describe()` method of `college` 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 (yes/no) 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. Describe what you are looking for and what you find. For example, are there any correlations between personal spending, instructional spending and Elite status?