Assignment 2 4DV652

January 23, 2021

1 Assignment 2 (4Dv652 course)

This assignment conatins tasks taken from “An Introduction to Statistical Leanring” book (https://www.statlearning.com/)

2 Exercise 1

This exercise relates to the **College** data set, which can be found in the file URL: https://github.com/shifteight/R-lang/blob/master/ISLR/College.csv. 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

(a) Use the `pd.read_csv()` function to read the data into Pandas dataframe. Make sure that you have the directory set to the correct location for the data.

```
[ ]: #your code here
```

(b) Use the `describe()` function to produce a numerical summary of the variables in the data set.

[ ]:

```
#your code here
```

(c)Use the `seaborn.pairplot()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using `iloc(A[,1:10])` function.

[ ]:

```
#your code here
```

(d) Use the `seaborn.boxplot()` function to produce side-by-side boxplots of **Outstate** versus **Private** universities.

[ ]:

```
#your code here
```

(e) Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `plt.figure(figsize = (15,20))` useful.

[ ]:

```
#your code here
```

(f) Based on the plots above provide a brief summary of what you discover from the College dataset.

[ ]:

```
#your text here as markdown cell
```

3 Exercise 2

This exercise involves the Auto data set studied in the lab (https://github.com/shifteight/R-lang/blob/master/ISLR/Auto.csv). Make sure that the missing values have been removed from the data.

(a) Use the `pd.read_csv()` function to read the data into Pandas dataframe. Make sure you do not have Unnamed columns. You can use `drop(columns='Unnamed: 0')` function to remove such column.

[2]:

```
#your code here
```

(b) Display/named quantative predictors/variables, and which one are qualitative predictors/variables (categorical)?

[ ]:

```
#your code/text here
```

(c) What is the range of each quantitative predictor? What is the mean and standard deviation of each quantitative predictor? You can answer this using the `describe()` function.

[ ]:

```
#your code here
```

(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? The `drop(dataset.index[])` function may help.

[ ]:
```
#your code here
```

(e) Suppose that we wish to predict gas mileage (**mpg**) on the basis of the other variables. Which variables you suggest that might be useful in predicting mpg? You may use scatterplots or other tools of your choice for this task. Justify your answer.

[ ]:
```
#your code/text here
```

4 Exercise 3

This exercise involves the Boston housing data set (can be found at the following URL: https://github.com/vincentarelbundock/Rdatasets/blob/master/csv/MASS/Boston.csv)

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

(a) To begin, load in the Boston data set.

[ ]:
```
#your code
```

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

[3]:
```
#your code and text
```

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

[ ]:
```
#your answer
```

(d) 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.

[ ]:

```
#your answer
```

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

[ ]:

```
#your code and answer
```

(f) 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.

[ ]:

```
#your code and answer
```