ESE 5410: Machine Learning for Data Science

Chapter 2 Project

In addition to answering the **questions** on Canvas, also submit your Project Notebook in the default location within Codio.

In the following exercise, we will perform exploratory data analysis (EDA) to extra insights using Python. Perform the following analyses by continuing from the previous assignment's notebook.

1 Part A

First, we return to the College dataset. To review, this dataset contains the following variables from 777 different universities and colleges in the US:

- 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

1. Use the `seaborn pairplot()` function on the original dataframe with all universities to produce a scatterplot matrix of the first ten columns of the data. Remember to use the `%matplotlib inline` command for plotting in Jupyter. Comment on your observations.

2. Use the `seaborn boxplot()` function on the original dataframe with all universities to produce side-by-side box plots of 'Outstate' versus 'Private'. **Answer on Canvas: On average, are private universities more expensive than public universities?**

3. Create a new qualitative variable named `large_university` by binning the 'Enroll' column. We are going to divide universities into two groups based on whether the number of new students enrolled exceeds the average (mean) of all new students enrolled.

4. Create a dataframe called `large_universities` that only includes large universities.

5. Use the `pandas describe(include="all")` function on `large_universities` to produce a numerical summary of each column. Within this dataframe, what is the 75 thpercentile for the column 'Enroll'? **Name this variable `enroll_answer` and enter your answer on Canvas.**

6. Use the `seaborn boxplot()` function on the `large_university` variable to determine whether larger universities are more expensive. **Answer on Canvas: On average, are large universities more expensive than small universities?**

7. To further examine the relationship between the size of a university against its out-of-state tuition, plot a scatter plot. **Answer on Canvas: What type of relationship does this show (weak, medium, strong)?**

8. Use `matplotlib.pyplot.hist()` or `seaborn.distplot()` to produce some histograms with different numbers of bins for a few of the quantitative variables. You may find the `matplotlib.pyplot.subplot()` function useful to plot multiple graphs under a single code-block. In particular, create a histogram to examine the number of applications received with a bin size of 200. **Answer on Canvas: Are there more universities that receive 0 to 10,000 applications or 10,000 to 20,000 applications?**

9. We are interested in the acceptance rate of these universities. To calculate the acceptance rate, we take the 'Accept' column and divide by the 'Apps' column. Add this column to a copied version of the original dataframe. Remember to copy the dataframe first before performing feature transformation, as we do not wish to alter or inadvertently create new columns in

our original data. **Answer on Canvas: Is there a positive or negative relationship between the university's acceptance rate and the percentage of students in the university who graduated from the top 10% from high school?**

Continue exploring the data and provide a brief summary of what you discover.

2 Part B

Next, let's turn to the Boston housing dataset, which contains the following variables from 506 different towns in Boston collected by the US Census Service:

- 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 : $1000(B_k - 0.63)^2$ where B_k is the proportion of African Americans by town
- LSTAT : % lower status of the population
- MEDV : Median value of owner-occupied homes in $1000's

1. To begin, load the Boston dataset. We can fetch this dataset by calling `sklearn`'s API:

```
    from sklearn.datasets import load_boston
    boston_dataset = load_boston()
    boston = pd.DataFrame(data = boston_dataset.data, columns = boston_dataset.feature_names)
    boston["MEDV"] = pd.Series(boston_dataset.target)
    boston.head()
```

2. Make pairwise scatterplots of some predictors (columns) in this dataset. Since this dataset includes many predictors, avoid using the `seaborn` `pairplot()` function with all the predictors to minimize run-time. Comment on your observations.

3. Are any of the predictors associated with per capita crime rate? If so, explain the relationship. **Answer on Canvas: Which predictor has the highest correlation with 'CRIM', besides 'CRIM' itself?**

4. 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. **Answer on Canvas: Which town number has the highest crime rate? Which town number has the lowest tax rate? Which town number has the highest pupil-teacher ratio?**

5. Count the number of towns in this dataset that are bound to the Charles river and name this number as the variable `charles`. **Enter your answer on Canvas.**

6. What is the median pupil-teacher ratio among the towns in this dataset? Name this number as the variable `med_pt`. **Enter your answer on Canvas.**

7. Which town of Boston has lowest median value of owner- occupied homes? Name this index as the variable `min_medv` and **enter your answer on Canvas**. 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 observations.

8. In this data set, how many of the suburbs average more than eight rooms per dwelling? Name this number as the variable `num_rooms` and **enter your answer on Canvas**. Comment on the suburbs that average more than eight rooms per dwelling.