ESE 542: Statistics for Data Science

Project 1

In addition to answering the **bolded** questions on Coursera, also attach your notebook, both as .ipynb and .html .

In the following exercise, we will familiarize ourselves with Python by studying the College dataset, which can be found in the file `College.csv`. 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. First, import all the necessary modules using the import function. For this exercise, we will be mainly using `pandas`. In the future, we will also be using `numpy`,`seaborn`,and `matplotlib`. To learn more about these packages, you can read through the documentation:

   - https://pandas.pydata.org/
   - https://numpy.org/
   - https://seaborn.pydata.org/
   - https://matplotlib.org/

```
#Data Wrangling
import pandas as pd
import numpy as np

#Plotting
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

2. Load the College dataset using `pandas`.

3. Use the `head()` function to view the data.

4. Notice that there is a column 'Names' of each university's name. As we don't want to use these names as predictors, they are natural candidates to index our data. We can do this using the following code:

```
#college is the dataframe
college.set_index("Names", inplace = True)
```

5. Use the `head()` function again. You should now see that the indices have been replaced with the name of each university in the data set. This means that Python has given each row a name corresponding to the appropriate university. Python will not try to perform calculations on the row names.

6. Use the `info()` function to check and produce a numerical summary of your variables.

7. Examine if there are any duplicate rows and drop them if needed.

8. Replace any missing values in the 'Apps' column with 0. This dataframe will henceforth be our original dataframe.

9. Find the college with the least out-of-state tuition and name this variable `college_least_tuition`. The variable should return the name of a college, not its tuition. **Answer on Coursera: Which college in this dataset has the least amount of out-of-state tuition?**

10. From the original dataframe, select the 'PhD' column and name this new dataframe `phd_column`. Find the length of this dataframe and assign this value to the variable `phd_column_length`. **Enter the value of `phd_column_length` on Coursera.** Hint: Make sure to use double brackets to select a Pandas dataframe and single brackets to select a Pandas series. You will notice the difference in formatting: Pandas dataframes are neatly formatted, while Pandas series are not formatted. A Pandas dataframe gives extra functionality compared to a Pandas series, such as appending other dataframes and selecting multiple columns. A Pandas series is essentially a Numpy column.

11. From the original dataframe, select both the 'Private' and 'Top10perc' columns, and slice them such that only the rows with index 15 and 16 remain. Name this dataframe `private_top10`. From this dataframe, find the length of the filtered 'Private' column and name this variable `private_column_length`. **Enter the value of `private_column_length` on Coursera.**

12. From the original dataframe, select the row that only contains data about the "University of Pennsylvania". Name this dataframe `penn`.

13. From the original dataframe, select the rows that contain all colleges with the substring "Penn" in their names. The "P" should be capitalized. Name this dataframe `many_penns`. **Answer on Coursera: How many universities are there in this dataset with the name "Penn"?**

Continue to explore the dataset by using any of the skills you learned in Recitation 1. Keep this notebook in a location that is easily accessible, as we will continue with it for the next assignment.