Unsupported Cell Type. Double-Click to inspect/edit the content.

Objectives

This homework assignment is designed to introduce you to using GitHub and review concepts discussed from chapter 2 (Introduction to Statistical Learning).

[Due 13 September 2024, 11:59 pm Tucson Time (35 Points Total). Any use of generative AI for this assignment will result in a 0-point grade. No exceptions.]{style="color: red;"} Note that I may call upon you to talk through your code or create a pop quiz that covers these concepts. Please take the time to thoroughly understand the concepts presented in this assignment and lab.

Additional resources relevant to this assignment

- NumPy
- Pandas
- MatPlotLib
- Seaborn

Scores

(35 Points Total)

- **Written Answers** for Concept Questions: 15 points
- **Coding Exercises**: 15 points
- **Lab**: 5 points
- **Extra credit**: 2 points

Submission

Grades are **NOT exclusively based on your final answers** for this assignment. I'll be grading the overall structure and logic of your code. Feel free to use as many lines as you need to answer each of the questions. I also highly recommend and strongly encourage adding comments (#) to your code. Comments will certainly improve the reproducibility and readability of your submission. Commenting your code is also good coding practice. **Specifically for the course, you'll get better feedback if I understand your code in detail.**

Deliverables

Due 13 September 2024, 11:59 pm Tucson Time (35 Points Total). Any use of generative AI for this assignment will result in a 0-point grade. No exceptions. You may use other sources as long as you cite them clearly.

There are five deliverables:

1. Become Familiar with GitHub
2. ISLP Lab Notebook
3. Concept Questions*
4. Applied Programming Questions*
5. Your Questions*

Be sure to name your files accordingly:

- `lastname_homework_2.ipynb`
- `lastname_homework_2_rendered.pdf`
- `lastname_notebook_2.ipynb`
- `lastname_notebook_2_rendered.pdf`

Recall that you can also render your files into html and that you can use Quarto. Push all 4 of these files into your GitHub repository.

- are the questions that you'll answer in your lastname_homework_2.ipynb file.

---

1. Become Familiar with GitHub

Watch the video about GitHub if you've never used it before. Submissions will be accepted only through GitHub from this point forward. I will not accept assignments if they were pushed on D2L.

Remember to work on the main branch only. Do not work on any other branch, even if you merge. (I will demonstrate what this means in the video).

2. ISLP Lab Notebook

You can download the Chapter 2 Notebook. I strongly encourage you to type this lab notebook again, by hand, if you are not familiar with these programming concepts (as per our discussions in lectures). If you are familiar with these concepts, simply run the lab notebook, then push your notebook and its rendered version onto GitHub. Remember to rename your notebook accordingly.

3. Concept Questions

Answer the questions as thoroughly as possible.

4. Coding Questions

You may consult your text and other resources as long as you cite all resources used. Remember to document your code. If you code won't run or you have the incorrect answer, yet you have good documentation, I may read your comments and still give some points.

5. Your Questions

Remember to ask an inquisitive question, a critical question, and an applicative question.

Time Commitment

Do not wait until the last minute to start working on this HW. In most cases, working under pressure will certainly increase the time needed to answer each of these questions and I don't answer emails on the weekends. If you are struggling with an assignment, please post your question on D2L. If you come to office hours, be sure you can describe your previous efforts to me (see syllabus).

1. Please go over the relevant readings.\
2. If you're still struggling with any of the questions, do some independent research (e.g. stackoverflow is a wonderful resource).\
3. Don't forget that your classmates will also be working on the same questions - reach out for help. Post on D2L. Check under the Discussion forum for folks looking to interact with other students in this class or start your own thread).

Conceptual Questions

This homework is divided into *three* main parts: conceptual, application, and questions you will ask. Please note that several of these questions are modified from James *et al.* (2021).

Question 1

For each of four scenarios listed below, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

- The sample size (number of observations; n) is extremely large, and the number of predictors (features; p) is small.

  ANSWER: *I think that although the number of observations are high the fact that the number of predictors is small, Inflexible methods will be the best performer. Inflexible methods are more restrictive and estimate in a small range, the lower number of predictors creates a smaller range, therefore inflexible is a better approach in this scenario*

- The number of predictors (p) is extremely large, and the number of observations (n) is small.

  ANSWER: *Flexible methods would be better for this scenario becasue of the higher number of parameters, eventhough you have to worry about overfitting there are ways to workaround that with the flexible methods.*

- The relationship between the predictors and response is highly non-linear.

  ANSWER: *Flexible Methods, if the data was linear I would say to use inflexible methods because linear regression is an inflexible method. Since the the data has a bigger spread the flexible method would be better.*

- The variance of the error terms (i.e. $\sigma^2 = Var(\epsilon)$) is extremely high.

  ANSWER: *Flexible, same argument as the argument above. The higher the variance of the error shows a bigger spread of the data, therefore we want a method that can account for that big range of data. That method would be flexible not inflexible.*

Question 2

In a few sentences, please answer the following questions to the best of your knowledge. Feel free to conduct additional research and cite your sources.

- Briefly explain the “curse of dimensionality”, provide a hypothetical example illustrating the concept, and list at least one potential way that is generally used to handle it when using machine learning models.

  ANSWER: *The "curse of dimensionality", is where we are looking at data composed of high numbers of features and variables. higher dimensionality can cause a lot of problems like overfitting, mulitcolinearity and a multitude of other problems. An example of overfitting would be if we were using a flexible method that has a lot of variables and our flexible method fits the data too well. This will affect our accuracy and won't be a good predictor for new data that our model hasn't seen. Ways to combat overfitting and high dimensionality is to pinpoint the important features of our models using various techniques and drop the features that aren't important.*

  Source: https://www.smartsocs.com/the-curse-of-dimensionality-in-machine-learning/#:~:text=The%20Curse%20of%20Dimensionality%20arises,reliability%20of%20Machine%20Learning%20models.

- Explain why the relationship between model error and complexity differs when patterns are examined using training and test datasets.

  ANSWER: *When you are training your model the model error will be very minimal so as you increase the complexity of the model (add more features/variables), the error will stay very consistent. Depending on what kind of model you are using you could overfit your model with the increase of complexity. So as you move to a test dataset, the model error is going to increase drastically due to the overfitting. Overall, the relationship of the model error and the model complexity will be highly effected when switching between datasets based on what method you are using and if overfitting is prevelent in your model.*

- To the best of your knowledge, distinguish between training, test, and validation datasets. Briefly describe the importance of each in the context of Machine Learning.

  ANSWER: *Training datasets are a portion of the data that we tell our model what the data represents so it can familiarize itself with the patterns within the data. The Training dataset is where the initial learning is happening from us teaching it. The validation dataset is where we give the model unlabeled data and see how it performs and then assess and tweak our model to be more accurate. The Testing dataset is the final dataset which is also unlabeld data that is ran through the model to see how well we did in making the model.*

- Briefly discuss the consequences of overfitting and underfitting.

  ANSWER: *I've already talked a lot on how overfitting makes our model less accurate becasue it becomes too familiar with the data it already knows and does very poorly when it comes to testing. Underfitting is where the model is less accurate on the training data, this happens because we it doesn't see the realtionship between the inputs and the target values.*

Question 3

Explain whether each of the scenarios presented below is a classification or regression problem. Indicate whether the situation is mostly interested in conducting inference (explaining patterns) or prediction. Finally, indicate the number of observations (n) and features (p) associated with each of the .

- We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

  ANSWER: *This is a Regression problem that is interested in infering, explaining the patterns of CEO salary. There are 500 observations with 4 features 3 of which are used as predictors.*

- We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched by different companies. For each product we have recorded its type, whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

  ANSWER: *This is a classification problem that wants to classify if the product is gonna be either a success or failure. There are 20 observations with 16 parameters, I'm including the product name as a parameter itself, and success and failure as more of a binary classification parameter. There are a lot of parameters that may not be useful within the 10 other variables.*

- We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for a given year. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

  ANSWER: *This is a regression problem that wants to predict how exchange rates of certain currencies effect world stock markets. The number of observations is the number of weeks in the year that they are recording the variables. The number of parameters are 4 which are the combination of exchange rates.*

Question 4

Let's now revisit the bias-variance decomposition.

- Provide a sketch of typical (squared) bias, variance, training error, test error, and irreducible (sometimes called Bayes error) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. **There should be five main curves. Make sure to label each one**. Now, add two arrows (parallel to the X axis) indicating the direction of increase in over- and under-fitting, respectively. Finally, label the point, also in the X axis, where model complexity is optimal.

NOTE: Please either draw your sketch on a piece of paper and then scan, or draw it with matplotlib or seaborn. If you draw by hand and then scan, you'll need to insert the image using the relevant code AND submit the image file along with your homework.

```
import numpy as np
import matplotlib.pyplot as plt
```

- Explain what the conceptual importance of acknowledging the existence of irreducible error is in the context of Machine Learning.

  ANSWER:

- Finally, briefly explain why each of the **five curves** has the shape displayed.

  ANSWER:

Question 5

You will now think of some real-life applications for statistical learning.

- Describe three real-life applications in which classification might be useful. Describe the response variable, as well as the potential predictors. Is the goal of each application inference or prediction? Explain your answer.

  ANSWER: *Classification is a very broad subset of machine learning, for me I am somewhat familiar with image classification which is classification of images at a pixel by pixel basis. An example of this would be doing image segementation on Brain MRI scans to determine if the image contains a Tumor or not. To me this feels like both inference and prediction, the model looks at the images pixels and recognizes the patterns of the color values around the tumor and then does an over all prediction and segments out just the tumor. So there aren't just one varible as a predictor and a response it's more of values that run through a nueral network that are RGB or grayscale values and the weight and biases change. This is a basic overview but that is one example of classification I am familiar with that is more complex. Another more simplified example would be classifying if an animal is a cat or dog. The response variable would be cat or dog, the predictors would be color, height, weight, diameter of eyes, and lenght of ears. this would be a predictor model where you ask the model if it thinks if your dog is actually a dog based on the predictor variables you give it. The last classification example that I worked on a project like this once is predicting if a person would survive the titanic. The response variable would be survival or not, the predictors would be age, sex, race, your net worth, weight, height, and location of room. After training the model on the actual titanic guest list you can see if the model thinks you will survive or not, this is all fun and games but the algorithm could be implemented into cruise lines to determine where to place passenges in what rooms to have a higher survival rate if there were to be an emergency.*

- Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

  ANSWER: *The first example of a regression problem is predicting a new NBA player salaries, the response variable would be NBA salaries of current players and some of the predictor variables are minutes played, position, free throw percentage, 3 pt percentage, and field goal percentage. Another example of regression would be a more simple model which is predicting the spread of future disease based on airquality index in a given county and infection rates of the flu. The response variable would be infection rates and the predictor variables would be AQI index, zip code, state, population, and population density. The last example I have would be to use regression to look at the effect of higher populated areas on star visibility, this would be more of an inference problem to see if the visibility of stars is changed whether you are in a bigger city or not. This is based off of the notion that bigger cities have a higher light pollution, the response variable would be Star visibility and the predictor variables would be population, city, and population density. These are the three examples that I thought of for linear regression problems and two of them are projects that I have done for other classes.*

Question 6

Let's now review some ethical considerations associated with the development and implementation of Machine Learning algorithms. Please answer each question with as much detail as possible (one or two short paragraphs per question should suffice). Feel free to conduct additional research if you think it's necessary. I'm looking for well-supported arguments. Your grade won't be based on whether or not I agree with your position. Instead, do your best to provide a thoughtful and clear answer.

- From your perspective, describe how unbiased must an algorithm be before it can be deployed in the "real world"? Is that level commonly achieved, discussed, or examined?

  ANSWER: *In my opinion I feel that it is purely based on what the algorithm is used for, if it is used on humans or an algorithm that isn't used on people like the classifying dogs and cats example I provided. I only say this because I feel that sometimes there is a coded bias within the algorithms, an example I learned once in a class is that sometimes greenscreens don't fully work on people of color. This is an example of an algortith that I feel needs to be less biased before deployed into mass usage. If a algorithm is going to be used by a general population, for example the US population who have people of all different kind of backgrounds the algorithm shouldn't be coded with a specific bias for a select number of people. The algorithm should be more unbiased to fit in the other groups of people who use the algorithm as well. I also understand that their is a point where you can make the program unbiased to a point where it is still accurate, but the question is if we have taken the time to truly figure out that threshold before releasing these algorithms. I think we can still put in more work to make the algortihms more inclusive.*

- Research labs and companies generally invest the most in improving and developing ML algorithms. Do you think that society in general should also have immediate and equal access to these developments and to their benefits? How would you balance these two goals (e.g. profit and well-being)? Similarly, do you think ML already affects (or will affect) inequalities?

  ANSWER: *I think that not everyone should have equal access to all the ML algorithms out there, for example there are certain algorithms that military use that I don't think everyone should have access too. In terms of algorithms that are seen in day to day life I think that it is right for companies who develop them to sell them to be specific in their products like phones and other devices. They put their products out there to be bought but I feel the access is only limited to if you can afford them or not, there are always good alternatives out there as well. One thing I would like to discuss and point out is that if there is a ML algorith that can help solve a mass problem in society like say curing cancer, that company or person shouldn't gatekeep it to make a profit. That is making money off of peoples suffering and I think that is highly unethical and shouldn't take place. I think machine learning does affect inequalities and like I said above its been seen in greenscreens with race and I feel that with the high rise in gender fluidity a lot of programs don't take that into account and going forward we will see a change in how algorithms operate. *

- What do you see as a solution for problems associated with increasing automation and efficiency, both currently and in the future?

  ANSWER: *The solution for seeing an increase in automation and a lack of jobs would be to educate yourself in the field of automation and work on creating the programs yourself. Right now in the present I feel that this is the start of the automation trend in bussiness so the job market is becoming the issue. As we move into the future I think if people can't see the trend they will become apart of helping support it instead of seeing it as an issue. Kind of like what you did with your studies in History and Anthropology, if people follow suit and better implement and create with automation everything will be better. Robots and ML models didn't create themselves and there are new problems in the world everyday, so there will never be a shortage of implementation.*

- We discussed different trade-offs between models in class – one of these was related to model interpretability. Why do you think it is important for machine learning models to be interpretable? Provide at least two reasons. Among the reasons that you listed, which one do you believe is the most important? Explain.

  ANSWER: *I think that there are a lot of important reasons why models need to be interpretable, One of the main two reasons are that if the model is easily interpretable other people can improve and use your model better and another is to know how and why the model got the answer. The one reason I think is more important is the first reason becasue it is more universal and focuses on improving the model for future use. I also think that the second reason is very important becasue it helps you trust that your model is actually giving the right answer.*

- Now, between model performance and model interpretability: which of these two qualities do you think is more important. Explain.

  ANSWER: *If we are doing more advanced stuff I think that model performance is more important and I think that the more advanced stuff like neural nets and CNNs have some very important uses. CNNs and other neural nets are highly noniterpretable but perform very well, they also have important application in the real world and do a lot of good in the medical field and other areas. So in the end I think becasue these model provide and are going to provide a lot more in society that performance outranks interprebility. People also still use and improve these models eventhough they aren't as interpretable, there are ways around low interprebility it seems which shows to me performance is better.*

- Particular examples in history show that different algorithms and datasets were originally developed/simulated/compiled with goals related to increasing systemic inequalities. For instance, let's take a quick look at the *iris* dataset (**Fisher 1936**). First, how many hits do you get after searching for *iris* dataset tutorial in Google (feel free to try any similar or more systematic search queries)? Based on this search, list at least two different modern uses of the dataset.

ANSWER: *When I looked up the tutorial I got atleast 100 results for the iris dataset. The dataset is used to teach the beginnings of classification problems of different iris in 3 main categories, this could translate to using them in classifiying what flowers are in peoples garden in your neighborhood. Another is to recognize if there are changes in the main iris over time, maybe as the years go on one of the iris types average petal lengths grew. This provides us insight on evolution of the iris flowers and could help us tweak the model to be more accurate in predicting.*

Applied Programming Questions

Question 7

This exercise relates to the `College` data set, which can be found in the file `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

1)Use the Pandas `read_csv()` function to read the data into `Python`. The dataset can be found on the ISLP website, which you downloaded in homework 1. Read dataset into `Python` and save it as `college.`

- 2) Note that the first column is just the name of each university. Although we are not going to use this column, it may be handy to have these names for later. First, assign the first column in college to the `index` of the `dataframe`. You can use the set_index() function in Pandas. . Second, delete the first column of college and overwrite the same object (`college`).
- 3) Now you should see that the first data column in `college` is `Private.` Note that another column labeled `Unnamed` now appears before the `Private` column. However, this is not a data column but rather the name that `Python` is giving to each row. Use the `describe()` function to produce a numerical summary of the variables in the data set.
- 4) Use the `pandas.plotting.scatter_matrix()` function to produce a scatterplot matrix of the first ten columns or variables of the data. You can reference the first ten columns of a dataframe using college.iloc[:, :10].

NOTE: iloc is for index and loc is for value. Be sure you've imported matplotlib.

- 5) Use the `boxplot()` function to produce side-by-side boxplots of `Outstate` versus `Private.`
- 6) Create a new binary variable called Elite, by binning the Top10perc variable to college. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.
- 7) Use the `describe()` or `value_counts()` function to see how many elite universities there are. Now use the `boxplot()` function to produce side-by-side boxplots of Outstate versus Elite.
- 8) Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the `plt.subplot()` function useful to create multiple plots in one figure. Create such a figure with four subplots.
- 9) Continue exploring the data, and provide a brief summary of what you discover. Some quick ideas: (1) explore graduation rates and briefly talk about what might be driving differences between universities. (2) Compare instructional expenditure per student and graduation rates.
- 10) Are there any other features, not taken into account in the dataset, that you think might be important for understanding whether a school is classified as elite or not?

```
#Imports for the coding part of the assignmnet
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#1
college = pd.read_csv('College.csv')
college.head()
```

	Unnamed: 0	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	T
0	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	
1	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	
2	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	
3	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	
4	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	

Next steps: Generate code with college | View recommended plots | New interactive sheet

```
#2
#I'm renaming the column becasue I don't like the name
college.rename(columns={'Unnamed: 0': 'Name'}, inplace=True)
#Now replacing the index to be indexed by name
college.set_index('Name', inplace=True)
#checking what the df looks like now
college.head()
```

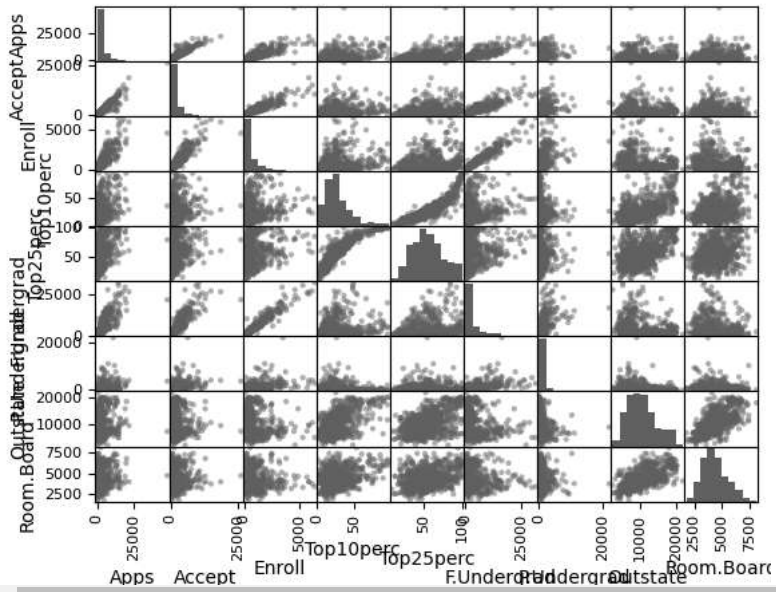
Name	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Ter
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	

Next steps: Generate code with college | View recommended plots | New interactive sheet

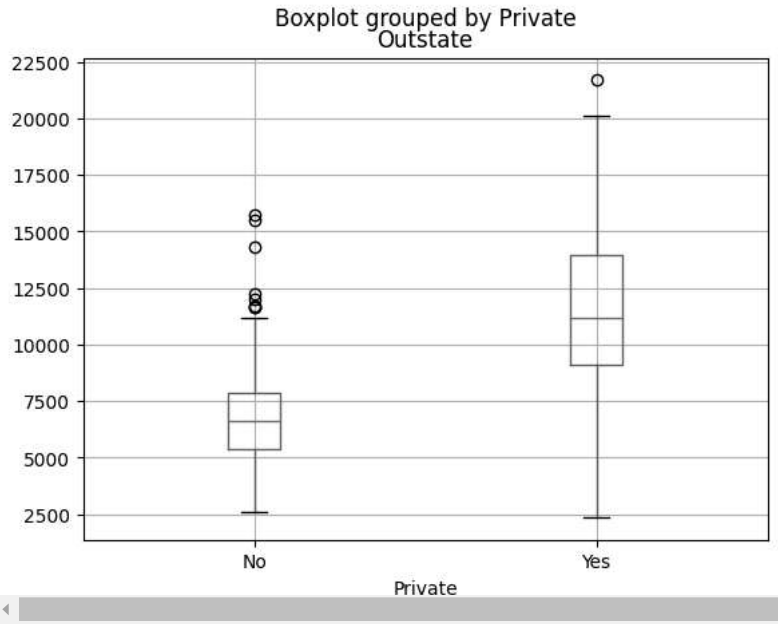
```
#3

college.describe()
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000

#4
```
pd.plotting.scatter_matrix(college.iloc[:, :10])
plt.show()
```



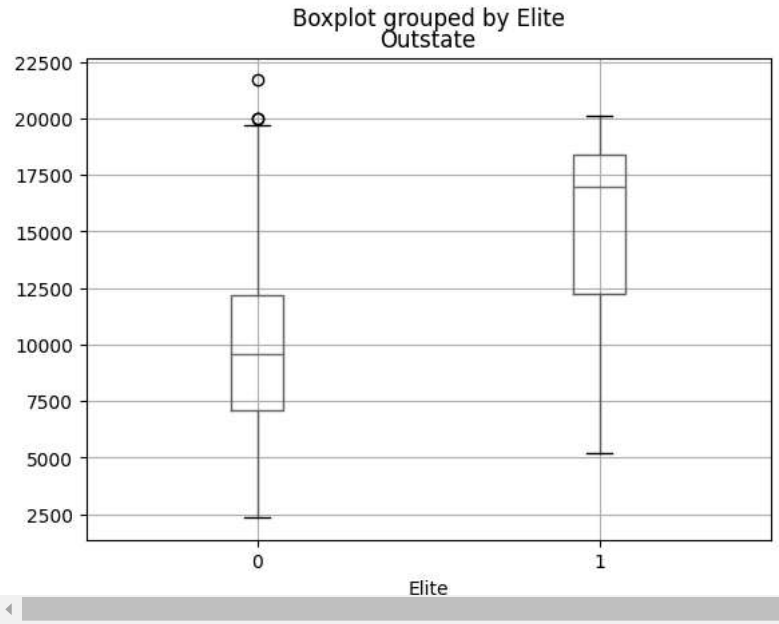
#5
```
college.boxplot(column='Outstate', by='Private')
plt.show()
```



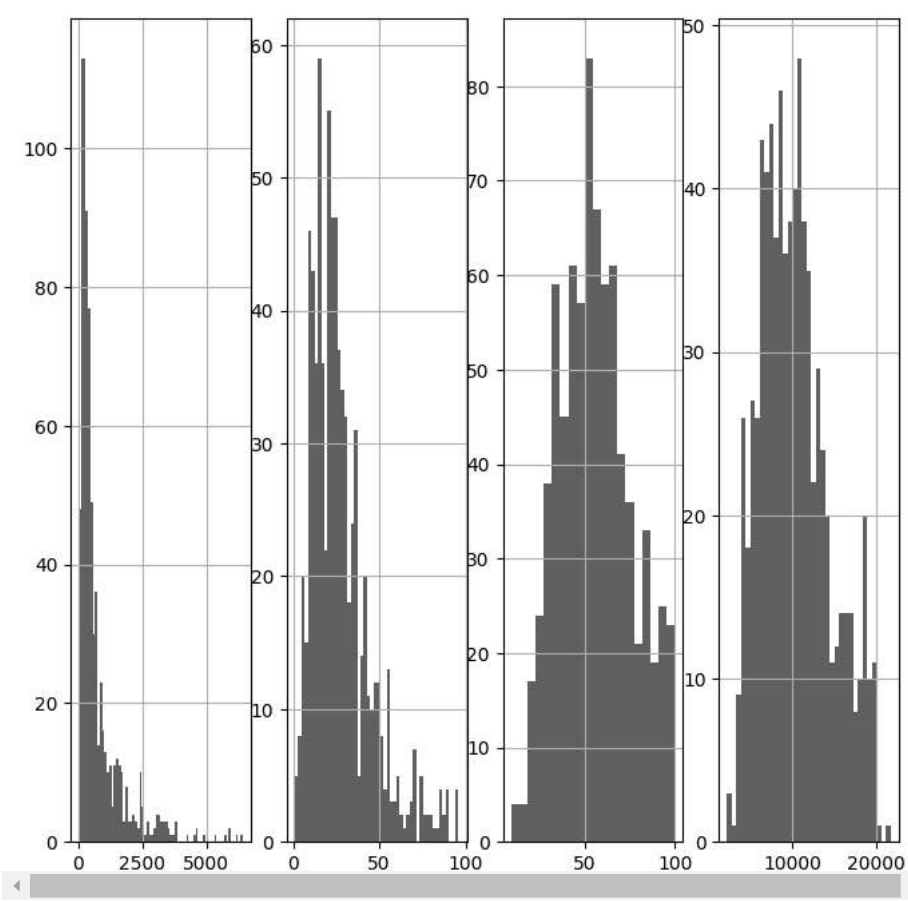
```
#6
#making a new column in the df that follows the guiudelines to be
#considered elite

college['Elite'] = np.where(college['Top10perc'] > 50, 1, 0)


#7
college['Elite'].value_counts()
```

Elite	count
0	699
1	78

```
#7 cont

college.boxplot(column='Outstate', by='Elite')
plt.show()
```




```
#8
fig, ax = plt.subplots(1, 4, figsize=(8, 8))
college['Enroll'].hist(bins=70, ax=ax[0])
college['Top10perc'].hist(bins=50, ax=ax[1])
college['Top25perc'].hist(bins=20, ax=ax[2])
college['Outstate'].hist(bins=35, ax=ax[3])

plt.show()
```



```
#9
college_filter = college[college['Grad.Rate'] > 60]
college_filter.describe()
```

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	
count	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000	474.000000	
mean	3398.985232	2210.002110	788.027426	33.101266	62.246835	3534.696203	607.109705	12008.310127	4682.497890	543.765823	1
std	4422.134747	2747.171183	979.507022	19.097344	19.135663	4943.071774	936.069711	3884.155533	1087.563554	156.336609	
min	141.000000	118.000000	46.000000	3.000000	19.000000	199.000000	1.000000	3040.000000	2190.000000	96.000000	
25%	935.250000	729.000000	284.000000	20.000000	49.000000	1090.500000	73.250000	9097.000000	3950.500000	450.000000	
50%	1750.500000	1313.500000	455.000000	29.000000	61.000000	1715.000000	283.500000	11652.000000	4517.000000	500.000000	1
75%	3840.500000	2440.500000	818.500000	42.000000	76.000000	3335.500000	733.500000	14500.000000	5513.750000	600.000000	1
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	9310.000000	20100.000000	8124.000000	2000.000000	6

I wanted to filter the data where the graduation rate is higher than 60% becasue I think that a good university should have a lot of grads. After filtering the data I was curious at what factors may increase the graduation rates. I used the describe() function on the filtered data so I can look at the original description and see how they compare. Some factors that are different are that there is a higher average of Expend, the books of higher graduation rates are lower on average and the enrollment is lower on average. These are just of the few things I noticed.

10)

I personally think that the dataset takes in account a lot of factors but they are all jumbled into one category of top 10 percent. If we were to do a deep dive of what it takes to be a top 10 percent college then we can go into more detail of what other factors we need to account for.

Question 8

This exercise involves the `auto` data.

- 1) Make sure that the missing values have been removed from the data. Which of the predictors are quantitative, and which are qualitative?
- 2) What is the range of each quantitative predictor? You can answer this using the `min()` `max()` `describe()` `apply()` `agg()` or other functions.
- 3) What is the mean and standard deviation of each quantitative predictor?
- 4) Now remove the 10th **through** 85th observations from `auto`. Save this as a new object called `auto2`. What is the range, mean, and standard deviation of each predictor in `auto2`?
- 5) Using the full data set (`auto`), investigate the predictors graphically, using scatter plots or other tools of your choice. Create some plots highlighting the relationships among the *predictors*. How could they be used for linear regression models?
- 6) Suppose that we were interested in predicting 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.

```
#1
auto = pd.read_csv('Auto.csv')
auto.head()
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

Next steps: Generate code with `auto` View recommended plots New interactive sheet

```
auto.describe()
```

	mpg	cylinders	displacement	weight	acceleration	year	origin
count	397.000000	397.000000	397.000000	397.000000	397.000000	397.000000	397.000000
mean	23.515869	5.458438	193.532746	2970.261965	15.555668	75.994962	1.574307
std	7.825804	1.701577	104.379583	847.904119	2.749995	3.690005	0.802549
min	9.000000	3.000000	68.000000	1613.000000	8.000000	70.000000	1.000000
25%	17.500000	4.000000	104.000000	2223.000000	13.800000	73.000000	1.000000
50%	23.000000	4.000000	146.000000	2800.000000	15.500000	76.000000	1.000000
75%	29.000000	8.000000	262.000000	3609.000000	17.100000	79.000000	2.000000
max	46.600000	8.000000	455.000000	5140.000000	24.800000	82.000000	3.000000

Question 2 and 3

Ranges

- mpg: min = 9 max = 46
- displacement: min = 68 max = 455
- weight: min = 1613 max = 5140
- acceleration: min = 8 max = 24

Sd

- mpg: sd = 7.826
- displacement: sd = 104.780
- weight: sd = 847.904
- acceleration: sd = 2.750

```
#4
auto2 = auto.drop(auto.index[9:84])
```

```
auto2.describe()
```

	mpg	cylinders	displacement	weight	acceleration	year	origin
count	320.000000	320.000000	320.000000	320.000000	320.000000	320.000000	320.000000
mean	24.474375	5.362500	186.540625	2930.318750	15.731563	77.165625	1.600000
std	7.894554	1.649499	99.372190	809.275266	2.680366	3.107389	0.816752
min	11.000000	3.000000	68.000000	1649.000000	8.500000	70.000000	1.000000
25%	18.000000	4.000000	99.500000	2213.750000	14.000000	75.000000	1.000000
50%	23.950000	4.000000	144.500000	2792.500000	15.550000	77.000000	1.000000
75%	30.750000	6.000000	250.000000	3474.750000	17.300000	80.000000	2.000000
max	46.600000	8.000000	455.000000	4997.000000	24.800000	82.000000	3.000000

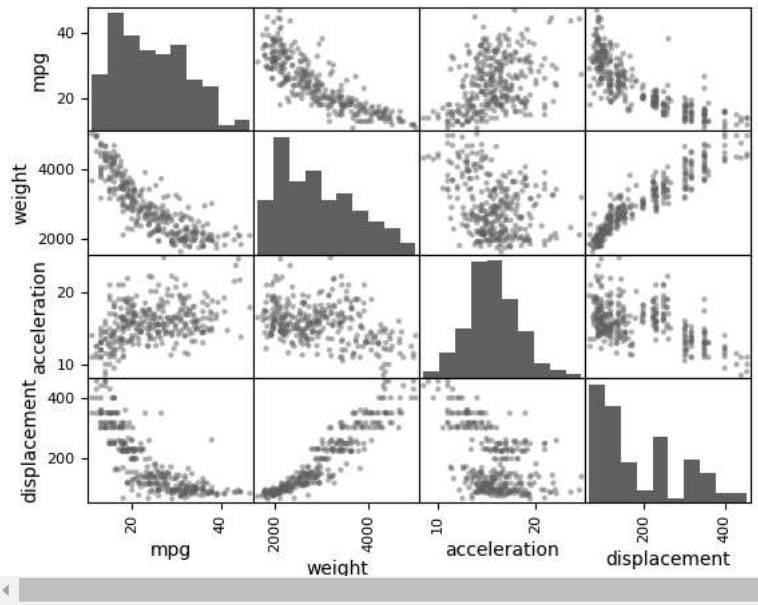
Ranges

- mpg: min = 11 max = 46
- displacement: min = 68 max = 455
- weight: min = 1649 max = 4997
- acceleration: min = 8.5 max = 24.8

Sd

- mpg: sd = 7.895
- displacement: sd = 99.327
- weight: sd = 809.275
- acceleration: sd = 2.680

```
#5
pd.plotting.scatter_matrix(auto2[['mpg', 'weight', 'acceleration', 'displacement']])
plt.show()
```



These plots can be used for linear regression models becasue they can see what variables have a linear relationship between one another and that is one of the first steps for SLR.

Question 6

The variables that seem to have the best predicting nature for mpg are weight and displacement. I can see that their graphs seem to resemble a linear relationship. If I wanted to dive deeper I could plot fitted vs residuals, residuals vs actual and QQ plots to check all the diagnostics for Regression.

```
# Your Questions

Use your scatterplot matrix visuals to **ask** three questions, 1
question per type:

-   Inquisitive (curious/ clarifying):

    QUESTION: *Based on my plot matrix, I see that the Acceleration vs
    Acceleration graph looks different from the others and resemples
    more of a bell curve rather than a linear plot, why is this the
    case?*

-   Critical (Verifying, proving, critique):

    QUESTION: *Would these relationships hold if we were to modify the
    study to include electric vehicles?*

-   Applicative:

    QUESTION: *If I were to be designing a car, why is it important to
    know that weight and displacement are correlated?*
```

Your Questions

Use your scatterplot matrix visuals to **ask** three questions, 1 question per type:

- Inquisitive (curious/ clarifying):

  QUESTION: *Based on my plot matrix, I see that the Acceleration vs Acceleration graph looks different from the others and resemples more of a bell curve rather than a linear plot, why is this the case?*

- Critical (Verifying, proving, critique):

  QUESTION: *Would these relationships hold if we were to modify the study to include electric vehicles?*

- Applicative:

  QUESTION: *If I were to be designing a car, why is it important to know that weight and displacement are correlated?*

Extra credit (2 points)

What are the most highly-correlated variables in the College dataset? Justify your answer.

```
#I'm going to use a correlation matrix to determine this
correlation_matrix = college.select_dtypes(include=np.number).corr()

print(correlation_matrix)
```

```
                Apps    Accept    Enroll  Top10perc  Top25perc  F.Undergrad  \
Apps        1.000000  0.943451  0.846822   0.338834   0.351640     0.814491
Accept      0.943451  1.000000  0.911637   0.192447   0.247476     0.874223
Enroll      0.846822  0.911637  1.000000   0.181294   0.226745     0.964640
Top10perc   0.338834  0.192447  0.181294   1.000000   0.891995     0.141289
Top25perc   0.351640  0.247476  0.226745   0.891995   1.000000     0.199445
F.Undergrad 0.814491  0.874223  0.964640   0.141289   0.199445     1.000000
P.Undergrad 0.398264  0.441271  0.513069  -0.105356  -0.053577     0.570512
Outstate    0.050159 -0.025755 -0.155477   0.562331   0.489394    -0.215742
Room.Board  0.164939  0.090899 -0.040232   0.371480   0.331490    -0.068890
Books       0.132559  0.113525  0.112711   0.118858   0.115527     0.115550
Personal    0.178731  0.200989  0.280929  -0.093316  -0.080810     0.317200
PhD         0.390697  0.355758  0.331469   0.531828   0.545862     0.318337
Terminal    0.369491  0.337583  0.308274   0.491135   0.524749     0.300019
S.F.Ratio   0.095633  0.176229  0.237271  -0.384875  -0.294629     0.279703
perc.alumni -0.090226 -0.159990 -0.180794   0.455485   0.417864    -0.229462
Expend      0.259592  0.124717  0.064169   0.660913   0.527447     0.018652
Grad.Rate   0.146755  0.067313 -0.022341   0.494989   0.477281    -0.078773
Elite       0.257285  0.113707  0.100996   0.759027   0.595890     0.060840

             P.Undergrad  Outstate  Room.Board     Books  Personal       PhD  \
Apps            0.398264  0.050159    0.164939  0.132559  0.178731  0.390697
Accept          0.441271 -0.025755    0.090899  0.113525  0.200989  0.355758
Enroll          0.513069 -0.155477   -0.040232  0.112711  0.280929  0.331469
Top10perc      -0.105356  0.562331    0.371480  0.118858 -0.093316  0.531828
Top25perc      -0.053577  0.489394    0.331490  0.115527 -0.080810  0.545862
F.Undergrad     0.570512 -0.215742   -0.068890  0.115550  0.317200  0.318337
P.Undergrad     1.000000 -0.253512   -0.061326  0.081200  0.319882  0.149114
Outstate       -0.253512  1.000000    0.654256  0.038855 -0.299087  0.382982
Room.Board     -0.061326  0.654256    1.000000  0.127963 -0.199428  0.329202
Books           0.081200  0.038855    0.127963  1.000000  0.179295  0.026906
Personal        0.319882 -0.299087   -0.199428  0.179295  1.000000 -0.010936
PhD             0.149114  0.382982    0.329202  0.026906 -0.010936  1.000000
Terminal        0.141904  0.407983    0.374540  0.099955 -0.030613  0.849587
S.F.Ratio       0.232531 -0.554821   -0.362628 -0.031929  0.136345 -0.130530
perc.alumni    -0.280792  0.566262    0.272363 -0.040208 -0.285968  0.249009
Expend         -0.083568  0.672779    0.501739  0.112409 -0.097892  0.432762
Grad.Rate      -0.257001  0.571290    0.424942  0.001061 -0.269344  0.305038
Elite          -0.116446  0.399477    0.298472  0.092176 -0.075269  0.341062

             Terminal  S.F.Ratio  perc.alumni    Expend  Grad.Rate     Elite
```

Apps	0.369491	0.095633	-0.090226	0.259592	0.146755	0.257285
Accept	0.337583	0.176229	-0.159990	0.124717	0.067313	0.113707
Enroll	0.308274	0.237271	-0.180794	0.064169	-0.022341	0.100996
Top10perc	0.491135	-0.384875	0.455485	0.660913	0.494989	0.759027
Top25perc	0.524749	-0.294629	0.417864	0.527447	0.477281	0.595890
F.Undergrad	0.300019	0.279703	-0.229462	0.018652	-0.078773	0.060840