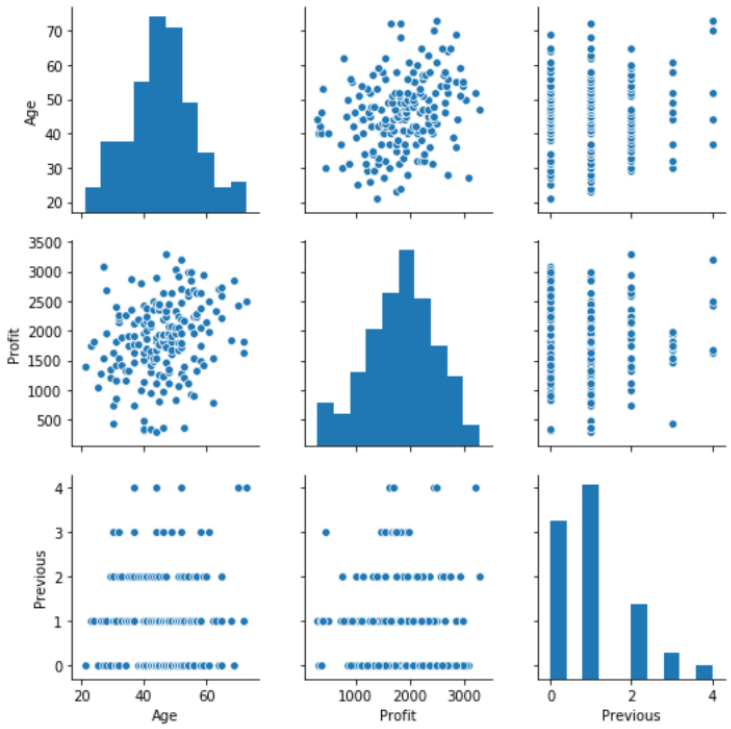
MIS 3335 – Data Analysis Using Python
Homework 3 – Plotting

50 points total

Your assigned task is to use the contents of these three datasets and your Python/pandas/Notebook skills to generate the following plots using seaborn. Pictures of the final plots are shown below for your reference. You do not need to add titles or other features unless specifically told to do so. ***Submit your notebook at the appropriate link in Blackboard***. Do not submit the data files.

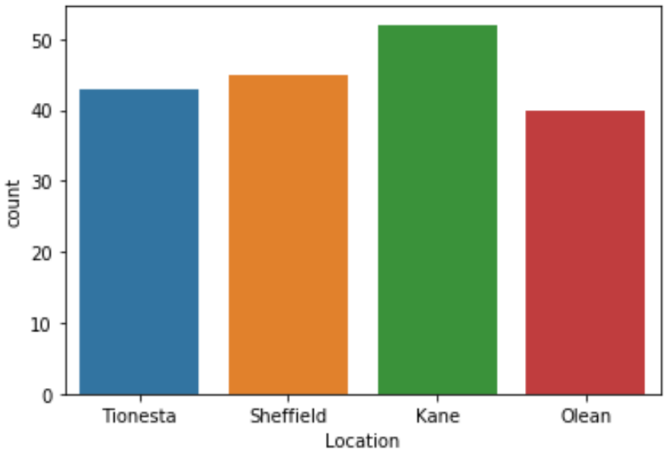
Dataset 1 – Applewood_2011.csv

1. Show all the pairwise plots.



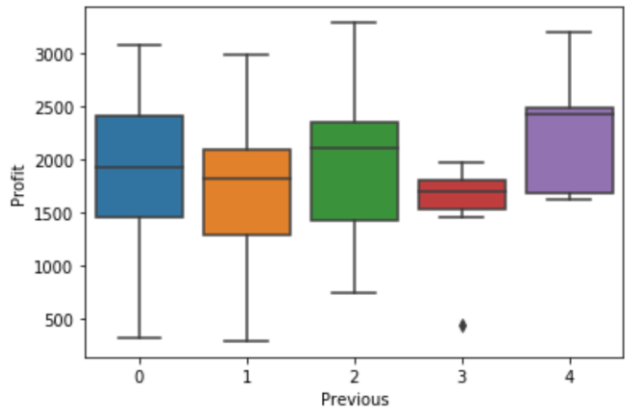
2 pts

2. Show a bar chart of the number of sales by Location.



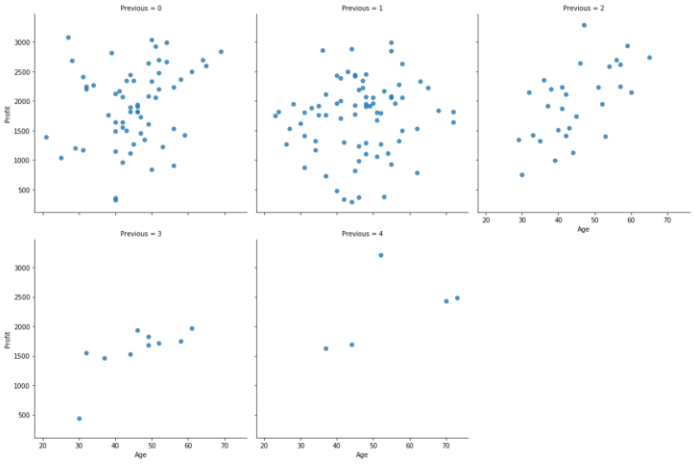
5 pts

3. Show box plots for profit with each value of “Previous” as a separate box.



5 pts

4. Scatter plots with each value of Previous on its own plot, three across.

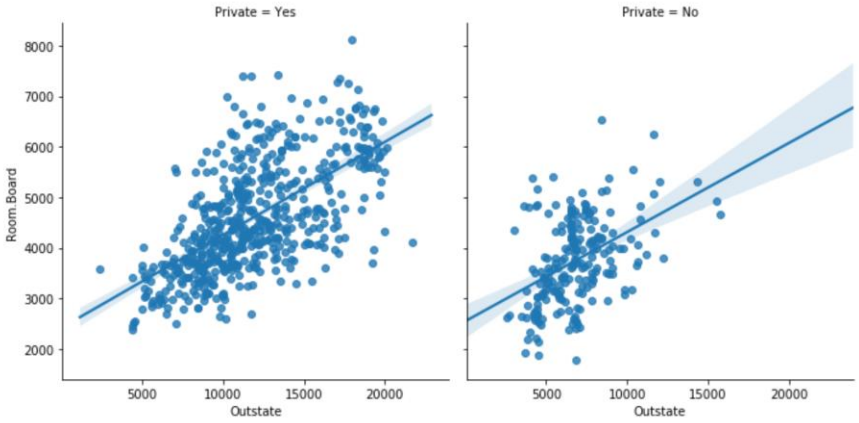


5 pts

Dataset 2 – College.csv

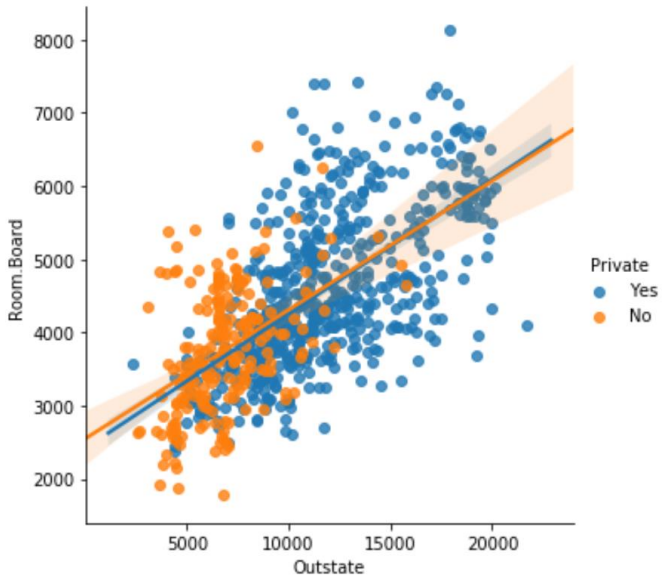
Scatter plot with regression line for Outstate versus Room.Board with these variations:

1. Each value of Private is in its own plot window.



5 pts

2. Each value of Private is a different color.



5 pts

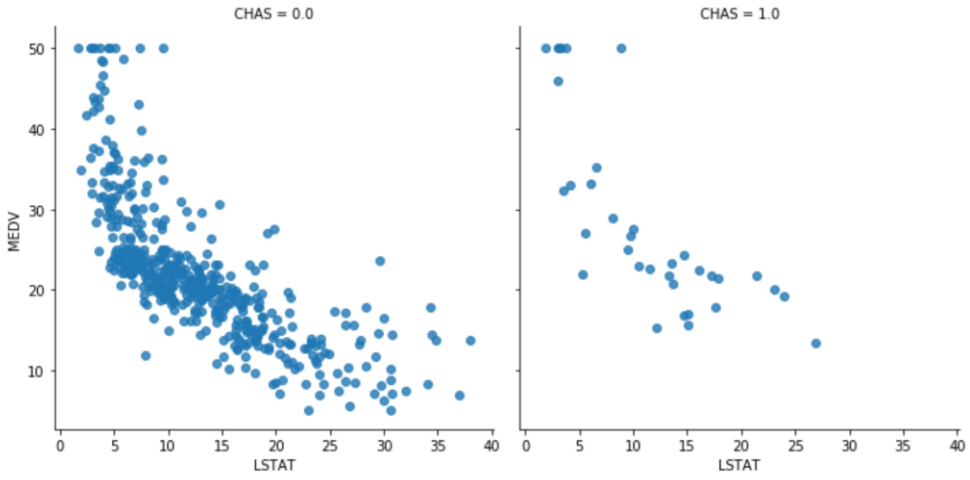
Dataset 3 – Boston_housing.csv

All the csv files to this point have been standard comma separated value files. The Boston_housing.csv file contains values separated by semi-colons (;). No problem; just use pd.read_csv the way you have been, then add the parameter to set the separator character (which defaults to a comma). That parameter is `sep=";"` which is added inside the parentheses at the end of the read_csv method and separated from the filename by a comma. Go to the "pandas.read_csv" help page if you need more clarification.

2 pts

1. Scatter plot of LSTAT and MEDV, with the value of CHAS on separate plots.

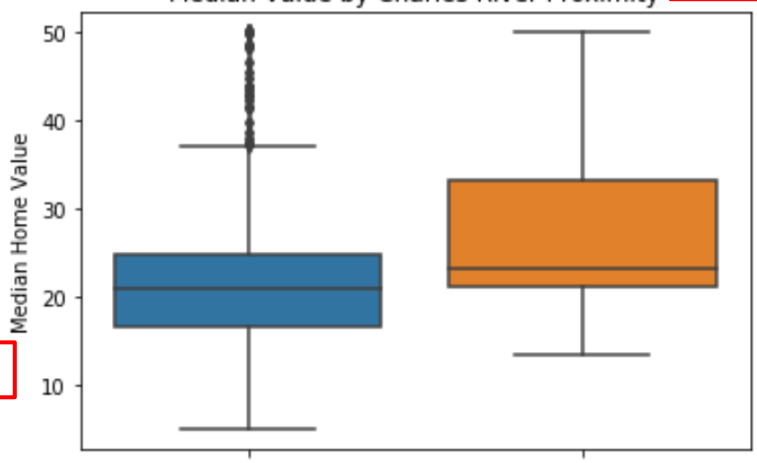
5 pts



NOTE: The following plots will need titles added.

4 pts

2. Box plot of median value with CHAS as the x variable.



1 pt

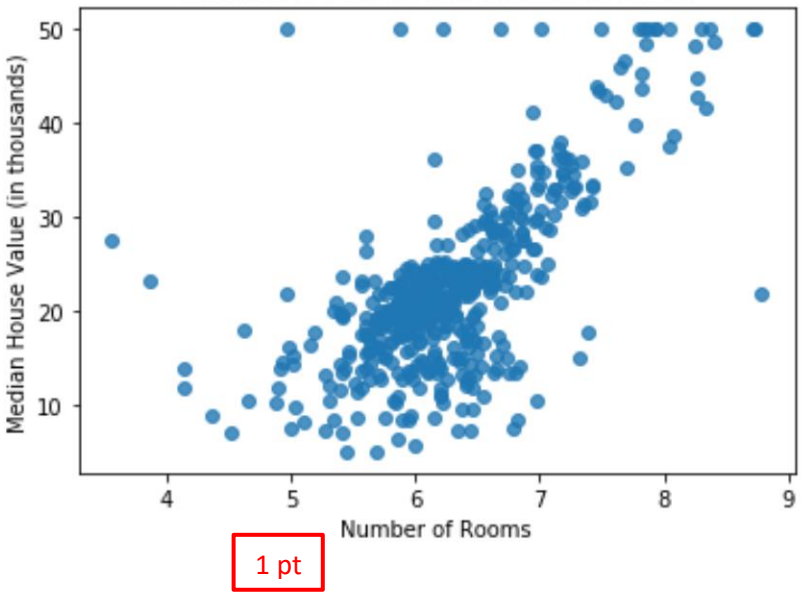
1 pt

1 pt

4 pts

3. Scatter plot without regression line of rooms versus median value

2 pts



1 pt

1 pt

1 pt

Appendix

More information about the datasets.

Applewood_2011.csv

You should already be familiar with this dataset.

College.csv

This dataset contains data from both private and public colleges and universities across the US.

Column Name	Description
School	Name of the school
Private	Is it a private school? (yes or no)
Apps	Number of applications received
Accept	Number of students accepted
Enroll	Number of students who enroll
Top10perc	Students in the top 10% of their HS class
Top25perc	Students in the top 25% of their HS class
F.Undergrad	Full-time undergrad students
P.Undergrad	Part-time undergrad students
Outstate	Out of state tuition
Room.Board	Cost of room and board
Books	Average cost of books
Personal	Average expenditure for student personal expenses
PhD	Number of PhD holders on faculty
Terminal	Number of terminal degree holding faculty
S.F.Ratio	Student to faculty ratio
perc.alumni	Percent of alumni active with school
Expend	In state expenditure per student
Grad.Rate	Graduation rate

Boston_housing.csv

Here is a description of the data attributes:

CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq. ft.
INDUS	Proportion of nonretail 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 distance to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	A factor reflecting racial composition by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000s