CST 383 - Intro to Data Science

Dr. Glenn Bruns

# Lab: data exploration - two variables, part 2

Hints are at the end of the lab, but try hard to answer the questions before looking at the hints.

1. We will continue with the college data set.

<https://raw.githubusercontent.com/grbruns/cst383/master/College.csv>

1. Write code to load the data. For example:

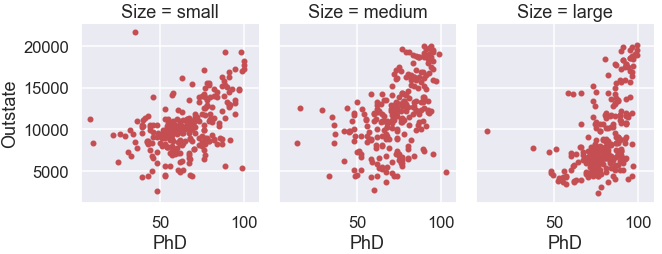
df = pd.read\_csv('...College.csv', index\_col=0)

Record your work on the following problems in your Python file.

1. Derive a new column, ‘Size’, from the F.Undergrad variable. The possible values of Size should be “small”, “medium”, or “large”. The value “small” should be assigned to the colleges in the “bottom 3rd” of F.Undergrad values, “medium” should be assigned to the “middle 3rd”, and “large” to the “top 3rd”. Use the Pandas ‘quantile’ function to find the corresponding F.Undergrad values. (If you're not sure how to do this, see the hints right away).

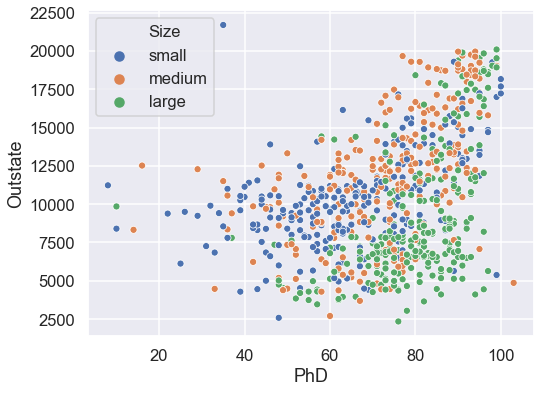
Create a barplot to make sure the number of small, medium, and large values are about the same. When you make a quick plot for yourself, no need for adding a title or axis labels. When you make a plot for a report or to share, it's important to use a good title and axis labels.

1. Use the faceting (also known as 'conditioning') idea to create three scatter plots, one for each value of your new variable size. The scatterplot should show PhD on the x axis and Outstate on the y axis. Try to make your plot look approximately like this:



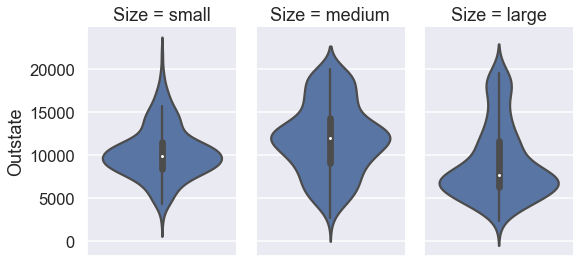
Look at the plots for a while and think their significance. Is the plot for large colleges different from the plot for small colleges? What does this say about large and small colleges. Also, do you see any interesting outliers?

1. Repeat problem 5, but this time show a single scatterplot, with color used to distinguish small, medium, and large schools. Your plot might look something like this:

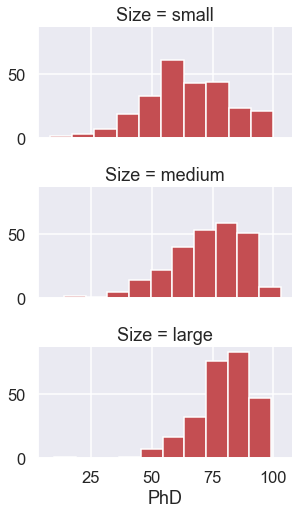


Look at the plot you created for this problem compared to the last problem. Which do you think is easier to interpret. Spend some time on this -- it's important.

1. Repeat the last plot, but this time use both shape and color to indicate college size. Do you think your new plot is easier to interpret than your plot of the previous problem?
2. Create three violin plots, showing the distribution of tuition at each of the three college size values. My plot looks like this:



1. Repeat the last plot, but now show the raw data on the plot.
2. If you still have time, repeat problem 4, but this time show 3 histograms of the PhD variable, one each for small, medium, and large schools. Show the three histograms in a single column of plots. Try to make your plot look similar to this:



1. If you still have time, freely explore some questions you have about the college data. For example, can you find out about selective schools with low graduation rates?

## 

## Hints:

1. -
2. -
3. My hints show just one way to answer the questions

breaks = df['F.Undergrad'].quantile([0,0.33, 0.66, 1.0])

df['Size'] = pd.cut(df['F.Undergrad'],

include\_lowest=True, bins=breaks,

labels=['small', 'medium', 'large'])

df.Size.value\_counts().plot.bar()

g = sns.FacetGrid(df, col='Size', height=4, aspect=0.8)

g.map(plt.scatter, 'PhD', 'Outstate', s=20, color="r")

sns.scatterplot(x='PhD', y='Outstate', hue='Size', data=df, s=50)

sns.scatterplot(x='PhD', y='Outstate', hue='Size', style='Size', data=df, s=55)

sns.catplot(y='Outstate', col='Size', data=df, kind='violin', height=4, aspect=0.7)

sns.catplot(y='Outstate', col='Size', data=df, kind='violin', inner='stick', height=4, aspect=0.7)

g = sns.FacetGrid(df, row='Size', height=2.5, aspect=1.8)

g.map(plt.hist, 'PhD', color="r")

1. -