Data Visualization: 4th lesson – Scatter Plots

In this tutorial, you’ll learn how to create advanced scatter plots.

***Set up the notebook***

As always, we begin by setting up the coding environment (this code is hidden, but you can un-hide it by clicking on the "Code" button immediately below this text, on the right).

import pandas as pd

pd.plotting.register\_matplotlib\_converters()

import matplotlib.pyplot as plt

%matplotlib inline

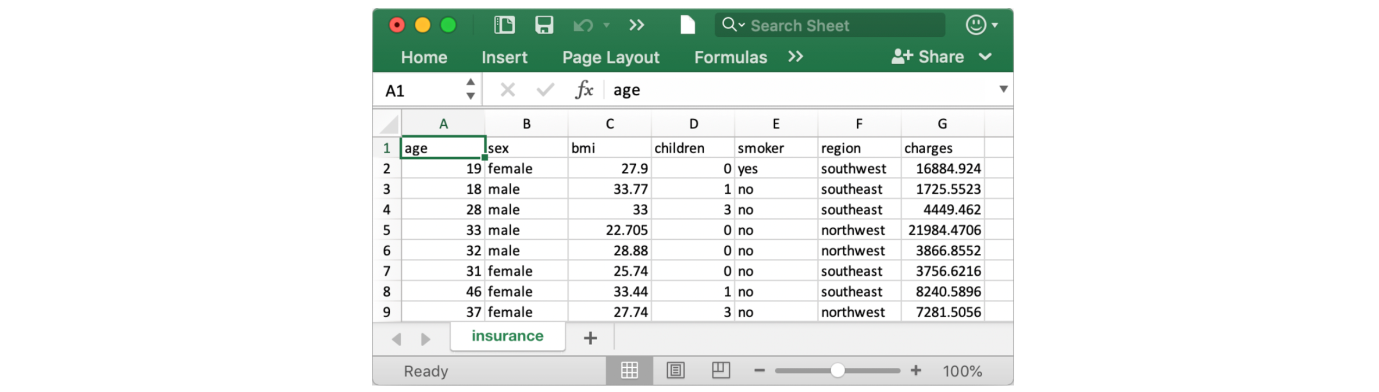
import seaborn as sns

print("Setup Complete")

Setup Complete

***Load and examine the data***

We'll work with a synthetic dataset of insurance charges, to see if we can understand why some customers pay more than others.



*# Path of the file to read*

insurance\_filepath = "../input/insurance.csv"

*# Read the file into a variable insurance\_data*

insurance\_data = pd.read\_csv(insurance\_filepath)

As always, we check the dataset loaded properly by printing the first five rows.

insurance\_data.head()

age sex bmi children smoker region charges

0 19 female 27.900 0 yes southwest 16884.92400

1 18 male 33.770 1 no southeast 1725.55230

2 28 male 33.000 3 no southeast 4449.46200

3 33 male 22.705 0 no northwest 21984.47061

4 32 male 28.880 0 no northwest 3866.85520

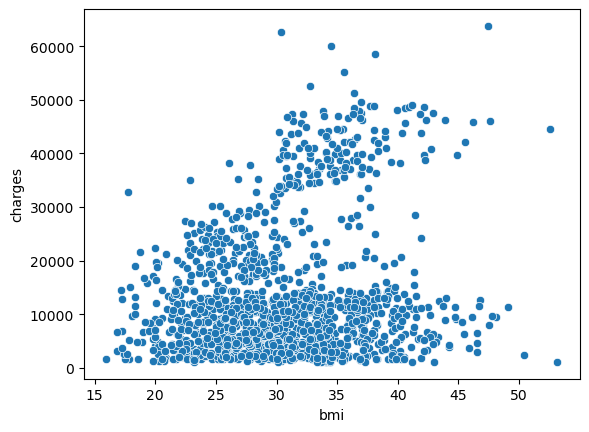
***Scatter plots***

To create a simple *scatter plot*, we use the sns.scatterplot command and specify the values for:

* The horizontal axis: x=insurance\_data['bmi']
* The vertical axis: y=insurance\_data['charges']

sns.scatterplot(x=insurance\_data['bmi'], y=insurance\_data['charges'])

<AxesSubplot:xlabel='bmi', ylabel='charges'>

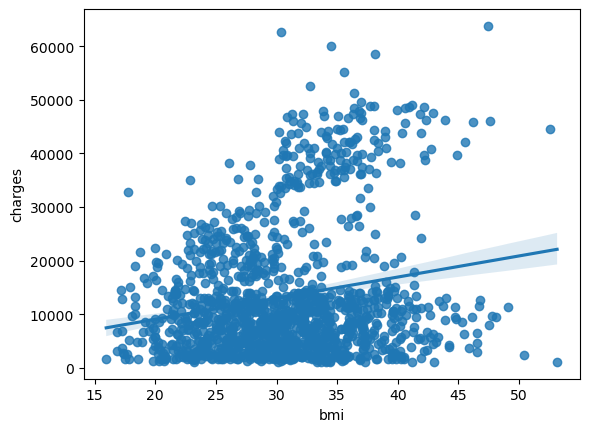


The scatterplot above suggests that body mass index (BMI) and insurance charges are positively correlated, where customers with higher BMI typically also tend to pay more in insurance costs (this pattern makes sense, since high BMI is typically associated with higher risk of chronic disease).

To double-check the strength of this relationship, you might like to add a *regression line*, or the line that best fits the data. We do this by changing the command to sns.regplot.

sns.regplot(x=insurance\_data['bmi'], y=insurance\_data['charges'])

<AxesSubplot:xlabel='bmi', ylabel='charges'>



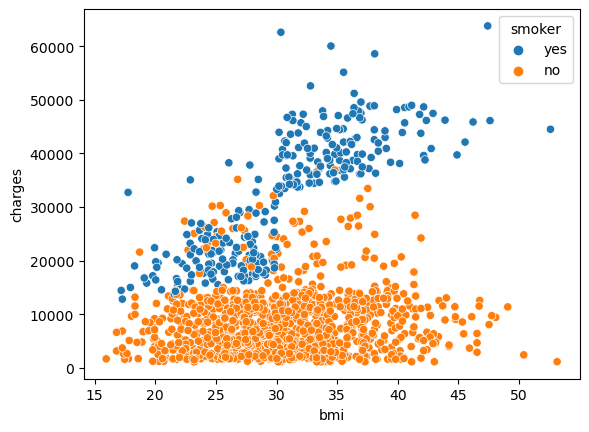
***Color-coded scatter plots***

We can use scatter plots to display the relationships between (not two, but...) three variables! One way of doing this is by color-coding the points.

For instance, to understand how smoking affects the relationship between BMI and insurance costs, we can color-code the points by 'smoker', and plot the other two columns ('bmi', 'charges') on the axes.

sns.scatterplot(x=insurance\_data['bmi'], y=insurance\_data['charges'], hue=insurance\_data['smoker'])

<AxesSubplot:xlabel='bmi', ylabel='charges'>

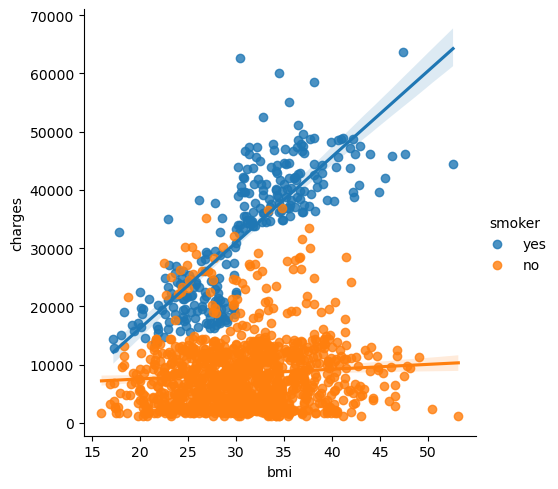


This scatter plot shows that while nonsmokers to tend to pay slightly more with increasing BMI, smokers pay much more.

To further emphasize this fact, we can use the sns.lmplot command to add two regression lines, corresponding to smokers and non-smokers (you'll notice that the regression line for smokers has a much steeper slope, relative to the line for non-smokers!)

sns.lmplot(x="bmi", y="charges", hue="smoker", data=insurance\_data)

<seaborn.axisgrid.FacetGrid at 0x7c1b45c4d510>



The sns.lmplot command above works slightly differently than the commands you have learned about so far:

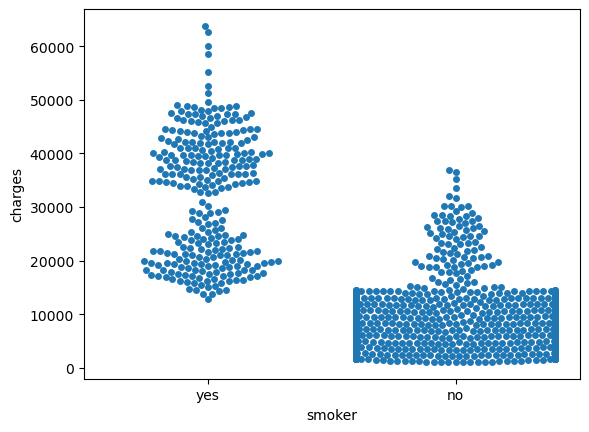
* Instead of setting x=insurance\_data['bmi'], to select the 'bmi'column in insurance\_data, we set x="bmi" to specify the name of the column only.
* Similarly, y="charges" and hue="smoker" also contain the names of columns.
* We specify the dataset with data=insurance\_data.

Finally, there's one more plot that you'll learn about, that might look slightly different from how you're used to seeing scatter plots. Usually, we use scatter plots to highlight the relationship between two continuous variables (like "bmi" and "charges"). However, we can adapt the design of the scatter plot to feature a categorical variable (like "smoker") on one of the main axes. We'll refer to this plot type as a *categorical scatter plot*, and we build it with the sns.swarmplot command.

sns.swarmplot(x=insurance\_data['smoker'],

y=insurance\_data['charges'])

<AxesSubplot:xlabel='smoker', ylabel='charges'>



Among other things, this plot shows us that:

* On average, non-smokers are charged less than smokers.
* The customers who pay the most are smokers; whereas the customers who pay the least are non-smokers.