This notebook is an exercise in the Data Visualization course. You can reference the tutorial at this link.

In this exercise, you will use your new knowledge to propose a solution to a real-world scenario. To succeed, you will need to import data into Python, answer questions using the data, and generate **scatter plots** to understand patterns in the data.

Scenario

You work for a major candy producer, and your goal is to write a report that your company can use to guide the design of its next product. Soon after starting your research, you stumble across this very interesting dataset containing results from a fun survey to crowdsource favorite candies.

Setup

Run the next cell to import and configure the Python libraries that you need to complete the exercise.

In [ ]:

```python
import pandas as pd
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
print("Setup Complete")
```

The questions below will give you feedback on your work. Run the following cell to set up our feedback system.

In [ ]:
```
# Set up code checking
import os
if not os.path.exists("../input/candy.csv"):
    os.symlink("../input/data-for-datavis/candy.csv", "../input/candy.c
sv")
from learntools.core import binder
binder.bind(globals())
from learntools.data_viz_to_coder.ex4 import *
print("Setup Complete")
```

Step 1: Load the Data

Read the candy data file into `candy_data`. Use the `"id"` column to label the rows.

In [ ]:
```
# Path of the file to read
candy_filepath = "../input/candy.csv"

# Fill in the line below to read the file into a variable candy_data
candy_data = pd.read_csv(candy_filepath,index_col="id")

# Run the line below with no changes to check that you've loaded the da
ta correctly
step_1.check()
```

In [ ]:
```
# Lines below will give you a hint or solution code
#step_1.hint()
#step_1.solution()
```

Step 2: Review the data

Use a Python command to print the first five rows of the data.

In [ ]:
```
# Print the first five rows of the data

candy_data.head() # Your code here
```

The dataset contains 83 rows, where each corresponds to a different candy bar. There are 13 columns:

- `'competitorname'` contains the name of the candy bar.
- the next **9** columns (from `'chocolate'` to `'pluribus'`) describe the candy. For instance, rows with chocolate candies have `"Yes"` in the `'chocolate'` column (and candies without chocolate have `"No"` in the same column).
- `'sugarpercent'` provides some indication of the amount of sugar, where higher values signify higher sugar content.
- `'pricepercent'` shows the price per unit, relative to the other candies in the dataset.
- `'winpercent'` is calculated from the survey results; higher values indicate that the candy was more popular with survey respondents.

Use the first five rows of the data to answer the questions below.

In [ ]:

```
# Fill in the line below: Which candy was more popular with survey respondents:
# '3 Musketeers' or 'Almond Joy'?  (Please enclose your answer in single quotes.)
more_popular = '3 Musketeers'

# Fill in the line below: Which candy has higher sugar content: 'Air Heads'
# or 'Baby Ruth'? (Please enclose your answer in single quotes.)
more_sugar = 'Air Heads'

# Check your answers
step_2.check()
```

In [ ]:

```
# Lines below will give you a hint or solution code
#step_2.hint()
#step_2.solution()
```

Step 3: The role of sugar

Do people tend to prefer candies with higher sugar content?

Part A

Create a scatter plot that shows the relationship between `'sugarpercent'` (on the horizontal x-axis) and `'winpercent'` (on the vertical y-axis). *Don't add a regression line just yet -- you'll do that in the next step!*

In [ ]:

```python
# Scatter plot showing the relationship between 'sugarpercent' and 'winpercent'
sns.scatterplot(x=candy_data["sugarpercent"],y=candy_data["winpercent"]) # Your code here

# Check your answer
step_3.a.check()
```

In [ ]:

```python
# Lines below will give you a hint or solution code
#step_3.a.hint()
#step_3.a.solution_plot()
```

Part B

Does the scatter plot show a **strong** correlation between the two variables? If so, are candies with more sugar relatively more or less popular with the survey respondents?

In [ ]:

```python
#step_3.b.hint()
```

In [ ]:

```python
# Check your answer (Run this code cell to receive credit!)
step_3.b.solution()
```

Step 4: Take a closer look

Part A

Create the same scatter plot you created in **Step 3**, but now with a regression line!

In [ ]:

```
# Scatter plot w/ regression line showing the relationship between 'sugarpercent' and 'winpercent'
sns.regplot(x=candy_data["sugarpercent"],y=candy_data["winpercent"]) # Your code here

# Check your answer
step_4.a.check()
```

In [ ]:

```
# Lines below will give you a hint or solution code
#step_4.a.hint()
#step_4.a.solution_plot()
```

Part B

According to the plot above, is there a **slight** correlation between `'winpercent'` and `'sugarpercent'`? What does this tell you about the candy that people tend to prefer?

In [ ]:

```
#step_4.b.hint()
```

In [ ]:

```
# Check your answer (Run this code cell to receive credit!)
step_4.b.solution()
```

Step 5: Chocolate!

In the code cell below, create a scatter plot to show the relationship between `'pricepercent'` (on the horizontal x-axis) and `'winpercent'` (on the vertical y-axis). Use the `'chocolate'` column to color-code the points. *Don't add any regression lines just yet -- you'll do that in the next step!*

In [ ]:

```
#  Scatter plot showing the relationship between 'pricepercent', 'winpercent', and 'chocolate'
```

```
sns.scatterplot(x=candy_data["pricepercent"],y=candy_data["winpercent"],hue=candy_data["chocolate"]) # Your code here

# Check your answer
step_5.check()
```

In [ ]:

```
# Lines below will give you a hint or solution code
#step_5.hint()
#step_5.solution_plot()
```

Can you see any interesting patterns in the scatter plot? We'll investigate this plot further by adding regression lines in the next step!

Step 6: Investigate chocolate

Part A

Create the same scatter plot you created in **Step 5**, but now with two regression lines, corresponding to (1) chocolate candies and (2) candies without chocolate.

In [ ]:

```
# Color-coded scatter plot w/ regression lines
sns.lmplot(x="pricepercent",y="winpercent",hue="chocolate",data=candy_data)# Your code here
# Check your answer
step_6.a.check()
```

In [ ]:

```
# Lines below will give you a hint or solution code
#step_6.a.hint()
#step_6.a.solution_plot()
```

Part B

Using the regression lines, what conclusions can you draw about the effects of chocolate and price on candy popularity?

In [ ]:

```
#step_6.b.hint()
```

In [ ]:

```
# Check your answer (Run this code cell to receive credit!)
step_6.b.solution()
```

Step 7: Everybody loves chocolate.

Part A

Create a categorical scatter plot to highlight the relationship between `'chocolate'` and `'winpercent'`. Put `'chocolate'` on the (horizontal) x-axis, and `'winpercent'` on the (vertical) y-axis.

In [ ]:

```
# Scatter plot showing the relationship between 'chocolate' and 'winpercent'
sns.swarmplot(x=candy_data["chocolate"],y=candy_data["winpercent"]) # Your code here

# Check your answer
step_7.a.check()
```

In [ ]:

```
# Lines below will give you a hint or solution code
#step_7.a.hint()
#step_7.a.solution_plot()
```

Part B

You decide to dedicate a section of your report to the fact that chocolate candies tend to be more popular than candies without chocolate. Which plot is more appropriate to tell this story: the plot from **Step 6**, or the plot from **Step 7**?

In [ ]:

```
#step_7.b.hint()
```

In [ ]:

```
# Check your answer (Run this code cell to receive credit!)
step_7.b.solution()
```

Keep going

Explore **histograms and density plots**.

Have questions or comments? Visit the Learn Discussion forum to chat with other Learners.