{"cells":[{"metadata":{},"cell_type":"markdown","source":"**This notebook is an exercise in the [Data Visualization](https://www.kaggle.com/learn/data-visualization) course. You can reference the tutorial at [this link](https://www.kaggle.com/alexisbcook/scatter-plots).**\n\n---\n"},{"metadata":{},"cell_type":"markdown","source":"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.\n\n## Scenario\n\nYou 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](https://fivethirtyeight.com/features/the-ultimate-halloween-candy-power-ranking/) containing results from a fun survey to crowdsource favorite candies.\n\n## Setup\n\nRun the next cell to import and configure the Python libraries that you need to complete the exercise."},{"metadata":{"trusted":false},"cell_type":"code","source":"import pandas as pd\npd.plotting.register_matplotlib_converters()\nimport matplotlib.pyplot as plt\n%matplotlib inline\nimport seaborn as sns\nprint(\"Setup Complete\")","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"The questions below will give you feedback on your work. Run the following cell to set up our feedback system."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Set up code checking\nimport os\nif not os.path.exists(\"../input/candy.csv\"):\n   os.symlink(\"../input/data-for-datavis/candy.csv\", \"../input/candy.csv\") \nfrom learntools.core import binder\nbinder.bind(globals())\nfrom learntools.data_viz_to_coder.ex4 import *\nprint(\"Setup Complete\")","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 1: Load the Data\n\nRead the candy data file into `candy_data`. Use the `\"id\"` column to label the rows."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Path of the file to read\ncandy_filepath = \"../input/candy.csv\"\n\n# Fill in the line below to read the file into a variable candy_data\ncandy_data = pd.read_csv(candy_filepath,index_col=\"id\")\n\n# Run the line below with no changes to check that you've loaded the data correctly\nstep_1.check()","execution_count":null,"outputs":[]},{"metadata":{"trusted":false},"cell_type":"code","source":"# Lines below will give you a hint or solution code\n#step_1.hint()\n#step_1.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 2: Review the data\n\nUse a Python command to print the first five rows of the data."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Print the first five rows of the data\n\ncandy_data.head() # Your code here","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"The dataset contains 83 rows, where each corresponds to a different candy bar. There are 13 columns:\n- `'competitorname'` contains the name of the candy bar. \n- 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).\n- `'sugarpercent'` provides some indication of the amount of sugar, where higher values signify higher sugar content.\n- `'pricepercent'` shows the price per unit, relative to the other candies in the dataset.\n- `'winpercent'` is calculated from the survey results; higher values indicate that the candy was more popular with survey respondents.\n\nUse the first five rows of the data to answer the questions below."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Fill in the line below: Which candy was more popular with survey respondents:\n# '3 Musketeers' or 'Almond Joy'? (Please enclose your answer in single quotes.)\nmore_popular = '3 Musketeers'\n\n# Fill in the line below: Which candy has higher sugar content: 'Air Heads'\n# or 'Baby Ruth'? (Please enclose your answer in single quotes.)\nmore_sugar = 'Air Heads'\n\n# Check your answers\nstep_2.check()","execution_count":null,"outputs":[]},{"metadata":{"trusted":false},"cell_type":"code","source":"# Lines below will give you a hint or solution code\n#step_2.hint()\n#step_2.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 3: The role of sugar\n\nDo people tend to prefer candies with higher sugar content? \n\n#### Part A\n\nCreate 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!_"},{"metadata":{"trusted":false},"cell_type":"code","source":"# Scatter plot showing the relationship between 'sugarpercent' and 'winpercent'\nsns.scatterplot(x=candy_data[\"sugarpercent\"],y=candy_data[\"winpercent\"]) # Your code here\n\n# Check your answer\nstep_3.a.check()","execution_count":null,"outputs":[]},{"metadata":{"trusted":false},"cell_type":"code","source":"# Lines

below will give you a hint or solution code\n#step_3.a.hint()\n#step_3.a.solution_plot()","execution_count":null,"outputs":[]},
{"metadata":{},"cell_type":"markdown","source":"#### Part B\n\nDoes 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?"},{"metadata":
{"trusted":false},"cell_type":"code","source":"#step_3.b.hint()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Check your answer (Run this code cell to receive
credit!)\nstep_3.b.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 4: Take
a closer look\n\n#### Part A\n\nCreate the same scatter plot you created in **Step 3**, but now with a regression line!"},
{"metadata":{"trusted":false},"cell_type":"code","source":"# Scatter plot w/ regression line showing the relationship between
'sugarpercent' and 'winpercent'\nsns.regplot(x=candy_data[\"sugarpercent\"],y=candy_data[\"winpercent\"]) # Your code here\n\n#
Check your answer\nstep_4.a.check()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Lines below will give you a hint or solution
code\n#step_4.a.hint()\n#step_4.a.solution_plot()","execution_count":null,"outputs":[]},{"metadata":
{},"cell_type":"markdown","source":"#### Part B\n\nAccording 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?"},{"metadata":
{"trusted":false},"cell_type":"code","source":"#step_4.b.hint()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Check your answer (Run this code cell to receive
credit!)\nstep_4.b.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 5:
Chocolate!\n\nIn 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!_"},{"metadata":{"trusted":false},"cell_type":"code","source":"#
Scatter plot showing the relationship between 'pricepercent', 'winpercent', and
'chocolate'\nsns.scatterplot(x=candy_data[\"pricepercent\"],y=candy_data[\"winpercent\"],hue=candy_data[\"chocolate\"]) # Your code
here\n\n# Check your answer\nstep_5.check()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Lines below will give you a hint or solution
code\n#step_5.hint()\n#step_5.solution_plot()","execution_count":null,"outputs":[]},{"metadata":
{},"cell_type":"markdown","source":"Can you see any interesting patterns in the scatter plot? We'll investigate this plot further
by adding regression lines in the next step!\n\n## Step 6: Investigate chocolate\n\n#### Part A\n\nCreate 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."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Color-coded scatter plot w/ regression
lines\nsns.lmplot(x=\"pricepercent\",y=\"winpercent\",hue=\"chocolate\",data=candy_data)# Your code here\n# Check your
answer\nstep_6.a.check()","execution_count":null,"outputs":[]},{"metadata":{"trusted":false},"cell_type":"code","source":"# Lines
below will give you a hint or solution code\n#step_6.a.hint()\n#step_6.a.solution_plot()","execution_count":null,"outputs":[]},
{"metadata":{},"cell_type":"markdown","source":"#### Part B\n\nUsing the regression lines, what conclusions can you draw about the
effects of chocolate and price on candy popularity?"},{"metadata":
{"trusted":false},"cell_type":"code","source":"#step_6.b.hint()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Check your answer (Run this code cell to receive
credit!)\nstep_6.b.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 7:
Everybody loves chocolate.\n\n#### Part A\n\nCreate 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."},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Scatter plot showing the relationship between 'chocolate' and
'winpercent'\nsns.swarmplot(x=candy_data[\"chocolate\"],y=candy_data[\"winpercent\"]) # Your code here\n\n# Check your
answer\nstep_7.a.check()","execution_count":null,"outputs":[]},{"metadata":{"trusted":false},"cell_type":"code","source":"# Lines
below will give you a hint or solution code\n#step_7.a.hint()\n#step_7.a.solution_plot()","execution_count":null,"outputs":[]},
{"metadata":{},"cell_type":"markdown","source":"#### Part B\n\nYou 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**?"},{"metadata":
{"trusted":false},"cell_type":"code","source":"#step_7.b.hint()","execution_count":null,"outputs":[]},{"metadata":
{"trusted":false},"cell_type":"code","source":"# Check your answer (Run this code cell to receive
credit!)\nstep_7.b.solution()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Keep
going\n\nExplore **[histograms and density plots](https://www.kaggle.com/alexisbcook/distributions)**."},{"metadata":
{},"cell_type":"markdown","source":"---\n\n\n\n\n*Have questions or comments? Visit the [Learn Discussion forum]
(https://www.kaggle.com/learn-forum/161291) to chat with other Learners.*"}],"metadata":{"kernelspec":
{"language":"python","display_name":"Python 3","name":"python3"},"language_info":
{"pygments_lexer":"ipython3","nbconvert_exporter":"python","version":"3.6.4","file_extension":".py","codemirror_mode":
{"name":"ipython","version":3},"name":"python","mimetype":"text/x-python"}},"nbformat":4,"nbformat_minor":4}