{"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/distributions).**\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 **histograms** and **density plots** to understand patterns in the data.\n\n## Scenario\n\nYou'll work with a real-world dataset containing information collected from microscopic images of breast cancer tumors, similar to the image below.\n\n![ex4_cancer_image](https://i.imgur.com/qUESsJe.png)\n\nEach tumor has been labeled as either [**benign**](https://en.wikipedia.org/wiki/Benign_tumor) (_noncancerous_) or **malignant** (_cancerous_).\n\nTo learn more about how this kind of data is used to create intelligent algorithms to classify tumors in medical settings, **watch the short video [at this link](https://www.youtube.com/watch?v=9Mz84cwVmS0)**!"},{"metadata":{},"cell_type":"markdown","source":"\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/cancer_b.csv\"):\n   os.symlink(\"../input/data-for-datavis/cancer_b.csv\", \"../input/cancer_b.csv\")\n   os.symlink(\"../input/data-for-datavis/cancer_m.csv\", \"../input/cancer_m.csv\")\nfrom learntools.core import binder\nbinder.bind(globals())\nfrom learntools.data_viz_to_coder.ex5 import *\nprint(\"Setup Complete\")","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"## Step 1: Load the data\n\nIn this step, you will load two data files.\n- Load the data file corresponding to **benign** tumors into a DataFrame called `cancer_b_data`. The corresponding filepath is `cancer_b_filepath`. Use the `\"Id\"` column to label the rows.\n- Load the data file corresponding to **malignant** tumors into a DataFrame called `cancer_m_data`. The corresponding filepath is `cancer_m_filepath`. Use the `\"Id\"` column to label the rows."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Paths of the files to read\n# Paths of the files to read\ncancer_b_filepath = \"../input/cancer_b.csv\"\ncancer_m_filepath = \"../input/cancer_m.csv\"\n\n# Fill in the line below to read the (benign) file into a variable cancer_b_data\ncancer_b_data = pd.read_csv(cancer_b_filepath,index_col='Id')\n\n# Fill in the line below to read the (malignant) file into a variable cancer_m_data\ncancer_m_data = pd.read_csv(cancer_m_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()\n","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 5 rows of the data for benign tumors."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Print the first five rows of the (benign) data\n____ # Your code here\ncancer_b_data.head()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"Use a Python command to print the first 5 rows of the data for malignant tumors."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Print the first five rows of the (malignant) data\n____ # Your code here\ncancer_m_data.head()","execution_count":null,"outputs":[]},{"metadata":{},"cell_type":"markdown","source":"In the datasets, each row corresponds to a different image. Each dataset has 31 different columns, corresponding to:\n- 1 column (`'Diagnosis'`) that classifies tumors as either benign (which appears in the dataset as **`B`**) or malignant (__`M`__), and\n- 30 columns containing different measurements collected from the images.\n\nUse the first 5 rows of the data (for benign and malignant tumors) to answer the questions below."},{"metadata":{"trusted":false},"cell_type":"code","source":"# Fill in the line below: In the first five rows of the data for benign tumors, what is the\n\n# largest value for 'Perimeter (mean)'?\nmax_perim = 87.46\n\n# Fill in the line below: What is the value for 'Radius (mean)' for the tumor with Id 842517?\nmean_radius = 20.57\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: Investigating differences"},{"metadata":{},"cell_type":"markdown","source":"#### Part A\n\nUse the code cell below to create two histograms that show the distribution in values for `'Area (mean)'` for both benign and malignant tumors. (_To permit easy comparison, create a single figure containing both histograms in the code cell below._)"},{"metadata":{"trusted":false},"cell_type":"code","source":"# Histograms for benign and maligant tumors\nsns.distplot(a=cancer_b_data['Area (mean)'], label=\"Benign\", kde=False)\nsns.distplot(a=cancer_m_data['Area (mean)'], label=\"Malignant\", kde=False)\nplt.legend()\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\nA researcher approaches you for help with identifying how the `'Area (mean)'` column can be used to understand the difference between benign and malignant tumors. Based on the histograms above, \n- Do malignant tumors have higher or lower values for `'Area (mean)'` (relative to benign tumors), on average?\n- Which tumor type seems to have a larger range of potential values?"},{"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: A very useful column\n\n#### Part A\n\nUse the code cell below to create two KDE plots that show the distribution in values for `'Radius (worst)'` for both benign and malignant tumors. (_To permit easy comparison, create a single figure containing both KDE plots in the code cell below._)"},{"metadata":{"trusted":false},"cell_type":"code","source":"# KDE plots for benign and malignant tumors\n____ # Your code here (benign tumors)\n____ # Your code here (malignant tumors)\nsns.kdeplot(data=cancer_b_data['Radius (worst)'], shade=True, label=\"Benign\")\nsns.kdeplot(data=cancer_m_data['Radius (worst)'], shade=True, label=\"Malignant\")\nplt.legend()\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\nA hospital has recently started using an algorithm that can diagnose tumors with high accuracy. Given a tumor with a value for `'Radius (worst)'` of 25, do you think the algorithm is more likely to classify the tumor as benign or malignant?"},{"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":"## Keep going\n\nReview all that you've learned and explore how to further customize your plots in the **[next tutorial](https://www.kaggle.com/alexisbcook/choosing-plot-types-and-custom-styles)**!"},{"metadata":{},"cell_type":"markdown","source":"---\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}