

## Assignment (100 Points)

This assignment will allow you to demonstrate your ability to: (1) make use of Python's Pandas library; (2) perform basic exploratory data analysis on a provided data set; (3) create graphics using **Matplotlib** and **Seaborn** as part of your exploratory data analysis work; and (4) present your work in the form of a more "formal" research paper framework.

---

The data set you will be using contains prices and other attributes of nearly 54,000 diamonds. The data set is provided via a separate file ('diamonds.csv') which you will need to download from Canvas. A description of the attributes contained within the data set can be found here:

<https://ggplot2.tidyverse.org/reference/diamonds.html>

For this assignment, you will need to load the data file into your online GitHub repository and then read the data from your GitHub repository into a Pandas dataframe. You will then use your Python and Pandas skills to answer and complete the content required for the outline specified below.

Your deliverable **must** include the following:

**Part 1: Data Summary (5 Points)** – Explain how many use cases your data set provides; how many attributes are in each use case; what the data types are for each of the attributes; etc. Be sure to include any Python code used as part of your Data Summary work.

**Part 2: Exploratory Data Analysis (EDA) (40 Points)** – Provide summary statistics for each attribute; provide appropriate graphical analysis for each attribute using both Matplotlib and Seaborn. For example, if you believe it is appropriate to generate a histogram for a particular variable as part of your EDA, create it first using Matplotlib and then once again using Seaborn. Include a narrative describing your EDA findings. Be sure to include any Python code used as part of your EDA work.

**Part 3: Inferences (40 Points total)** – Perform whatever analysis is necessary to answer the following questions:

1. **(4 Points)** What proportion of diamonds are between .30 and 1.08 carats?
2. **(4 Points)** How many of the diamonds have equal x and y dimensions?
3. **(4 Points)** How many of the diamonds have a carat value that is less than the mean carat value?
4. **(4 Points)** How many diamonds have a Premium cut or better? Note that the possible values for the quality of a cut are ranked in ascending order as follows: **Fair / Good / Very Good / Premium / Ideal**
5. **(8 Points)** Which diamond has the highest price per carat? What is its value?
6. **(8 Points)** Using both Matplotlib and Seaborn, make boxplots of the diamond price for each distinct cut value and discuss any conclusions you can draw from the appearance of the boxplots.
7. **(8 Points)** Using both Matplotlib and Seaborn, make a scatter plot of price vs. carat. What can we say about the relationship between those two attributes?

Provide a short written narrative that explains your approach for each of these questions and tasks using formatted

Markdown cells in your Jupyter notebook. Be sure to include any Python code used as part of your work.

**Part 4: Conclusion (10 Points)** – A brief, concise narrative explaining your conclusions.

**References (5 Points)** - Be sure to include proper citations for any references you may have relied on as part of your work.

**Your Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Python code should include succinct explanatory comments.**

Save all of your work for this project within a single Notebook and upload / submit the link for it. Be sure to save your Notebook using the following nomenclature : **first initial\_last name\_Quiz1"** (e.g., J\_Smith\_Quiz1).