Regression

1. **Choose; Load; Inspect and Explore your Data [5].**

• **Pick a Domain and Dataset you are interested in:**For this assignment, you will need to find a dataset of your choosing (interest) and load into dataframe object with PANDAS library.

Perform a initial observation regarding the dataset, while doing that try to answer following question: **–** Detailed description about the dataset:

1. When and Who created the dataset?  
2. How did you get acess to the dataset?  
3. List out the attributes (columns) of a dataset.

**–** Guess some probable question that dataset could answer.

**–** Assess the basic fitness of the dataset.

The dataset containing information on nearly 54,000 diamonds was curated by SHIVAM AGRAWAL in 2017. This dataset encompasses various attributes of diamonds, making it a valuable resource for data analysis and visualization purposes.

Access to the dataset was facilitated through Kaggle.com, a prominent platform for sharing datasets and hosting data science competitions. Kaggle provides a diverse array of datasets contributed by individuals, organizations, and researchers worldwide, making it a convenient resource for data scientists and analysts to explore and utilize data for various projects and analyses.

The dataset includes several attributes or columns that provide valuable insights into the characteristics of each diamond. These attributes include:

* **Price**: The price of the diamond in US dollars, ranging from $326 to $18,823.
* **Carat**: The weight of the diamond, ranging from 0.2 to 5.01 carats.
* **Cut**: The quality of the cut, categorized as Fair, Good, Very Good, Premium, or Ideal.
* **Color**: The color of the diamond, graded from J (worst) to D (best).
* **Clarity**: A measurement of how clear the diamond is, with grades ranging from I1 (worst) to IF (best).
* **X, Y, Z**: Dimensions of the diamond, including length (x), width (y), and depth (z) measured in millimeters.
* **Depth**: The total depth percentage of the diamond, calculated as z divided by the mean of x and y, or 2 times z divided by the sum of x and y, ranging from 43 to 79.
* **Table**: The width of the top of the diamond relative to its widest point, expressed as a percentage and ranging from 43 to 95.

These attributes collectively provide a comprehensive overview of each diamond's characteristics, enabling thorough analysis and insights into factors influencing diamond pricing and quality in the market.

Some probable questions the dataset could help answer are:

1. How does the price of a diamond correlate with its carat weight?
2. What impact does the quality of the cut have on diamond prices?
3. Is there a relationship between diamond color grade and price?
4. How does clarity affect diamond pricing?
5. Are there trends between diamond dimensions and their prices?
6. Do depth percentage and table width influence diamond prices?

To assess the dataset's fitness:

* Check for completeness to ensure no missing values.
* Verify consistency and accuracy of attribute values.
* Ensure relevance of attributes to analysis goals.
* Consider dataset size for meaningful conclusions.
* Analyze data distribution to understand variability and representativeness.

By evaluating these factors, we can determine the dataset's suitability for analysis and decision-making.

If you're performing classification through logistic regression, you should focus on questions that involve predicting a categorical outcome based on the dataset's attributes. Here are a few questions that align well with logistic regression:

1. **Predicting Diamond Cut Quality**: Can logistic regression help predict the quality of a diamond cut (e.g., Fair, Good, Very Good, Premium, Ideal) based on its attributes like carat weight, color, clarity, and dimensions?

Please write a brief summary about the outcomes of your experiment. You can explain the following questions:

1. **What was your model performance in section 3.1.3.?**
2. **Did any of the methods you applied** {**Cross Validation and Feature selection**} **increased or decreased.**
3. **What did you learn and what could be the future direction?**

Classification

Please write a brief summary about the outcomes of your experiment. You can explain the following questions:

1. **What was your model performance in section 3.1.3.?**
2. **Did any of the methods you applied** {**Cross Validation and Feature selection**} **increased or decreased.**
3. **What did you learn and what could be the future direction?**