**Course: DSC 530 Data Exploration and Analysis**

**Assignment: 12.2 - Term Project**

**File: Bgaggainpali\_DSC530\_FinalProject\_diamonds.ipynb**

**Name: Bhargava Gaggainpali**

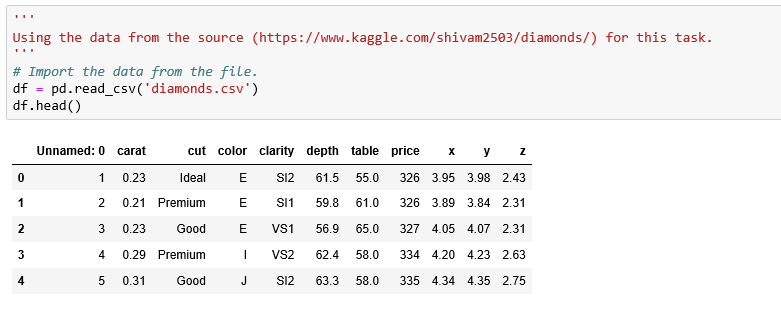
**Date: 05/30/2020**

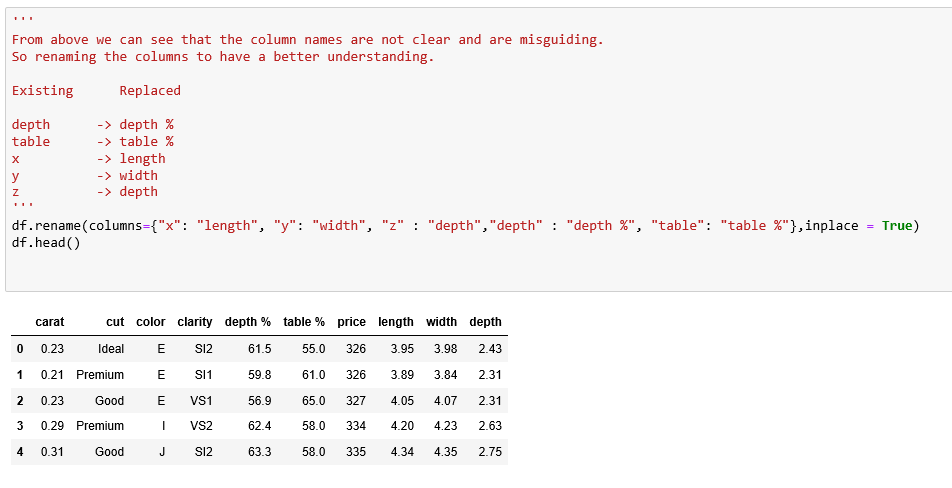
<https://github.com/bgaggainpali/bgaggainpali_DSC530>

-----------------------------------------------------------------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------

* A minimum of 5 variables in your dataset used during your analysis (for help with selecting, the author made his selection on page 6 of your book). Consider what you think could have an impact on your question – remember this is never perfect, so don’t be worried if you miss one (Chapter 1).





-----------------------------------------------------------------------------------------------------------------------------------------

* Describe what the 5 variables mean in the dataset (Chapter 1).

Five main features which are used to access the quality of a diamond, are listed down below.

Carat, Weight, Cut, Clarity & Color

Carat Weight : Carat is a metric that is used to measure the weight of a diamond. One carat is equivalent to 200mg. Diamond prices increase with diamond carat weight, which means bigger the diamond higher the price. If two diamonds weights are equal, then other features are used to determine the price.

Clarity : Diamonds are generated from sheer pressure and heat below the ground. Therefore, there will be some inclusion inside a diamond i.e., a mark or line pattern inside a diamond.

Cut: This feature is an important thing to notice in a diamond as it measures three crucial things, such as Brilliance, Fire, Scintillation.

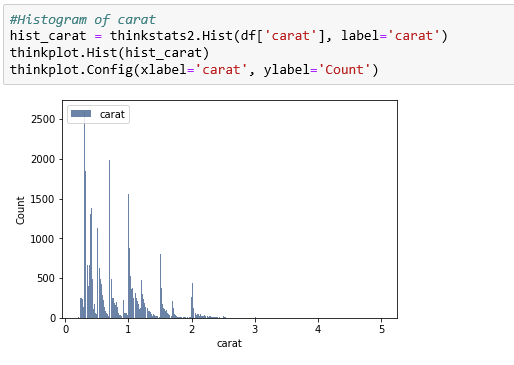
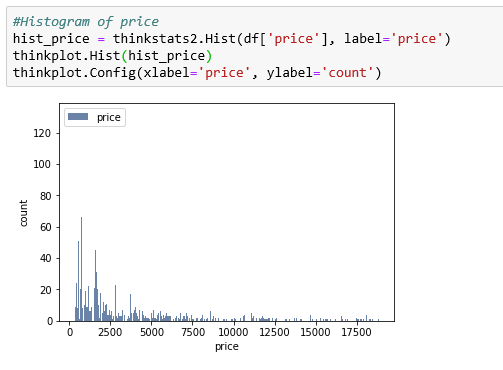
Depth %: Depth is the distance from a top surface i.e., table to a culet. The depth percentage is calculated by dividing the diamond depth by the overall width of a diamond. Lower the depth percentage the bigger the diamond looks from the below i.e., pavilion.

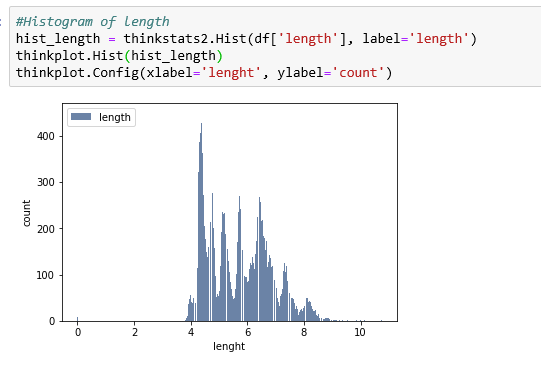
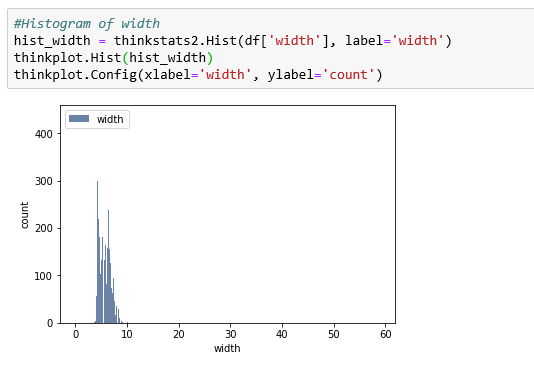
Table %: The table is the topmost surface of a diamond and the most significant facet of the round diamond. An appropriate width of a table will allow the light to enter and reflect on the appropriate direction.

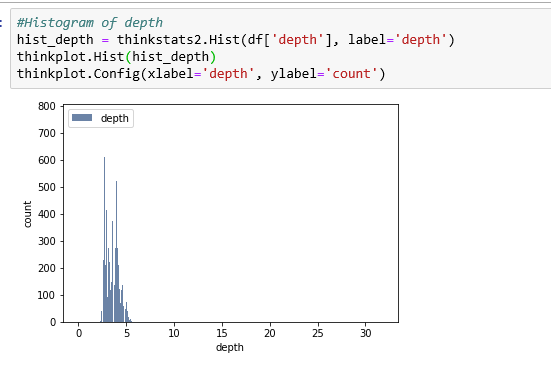
Color: Color measurement in diamond measures lacks color. If the diamond color is like a drop of water that is colorless, it will have a high value. As then only it can scatter the light without observing.

-----------------------------------------------------------------------------------------------------------------------------------------

* Include a histogram of each of the 5 variables – in your summary and analysis, identify any outliers and explain the reasoning for them being outliers and how you believe they should be handled (Chapter 2).

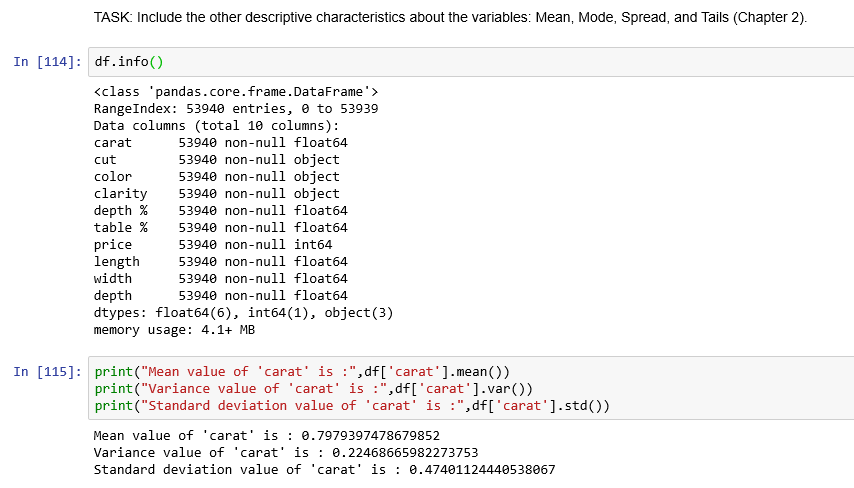
 

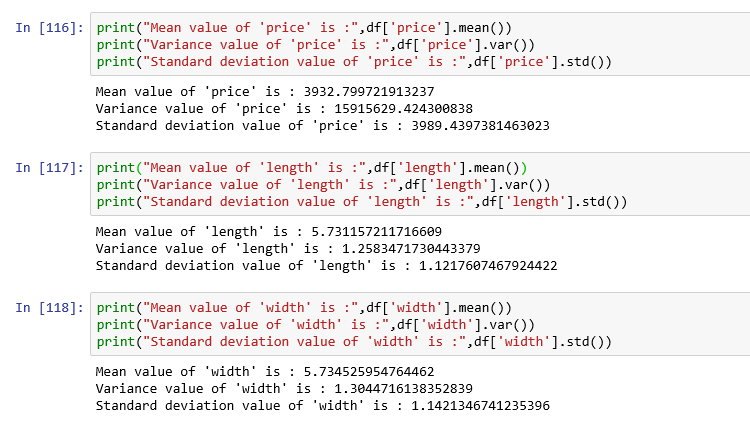
 

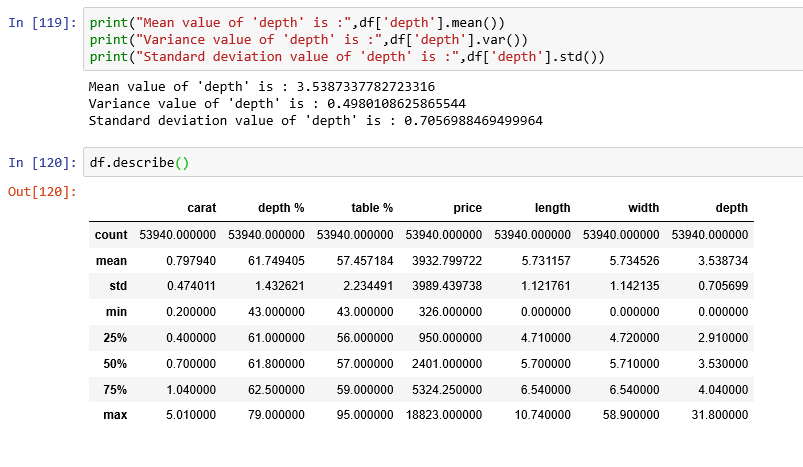


-----------------------------------------------------------------------------------------------------------------------------------------

* Include the other descriptive characteristics about the variables: Mean, Mode, Spread, and Tails (Chapter 2).

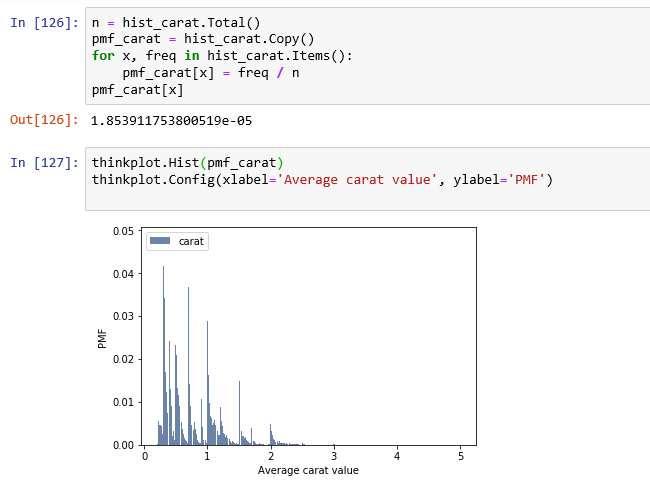


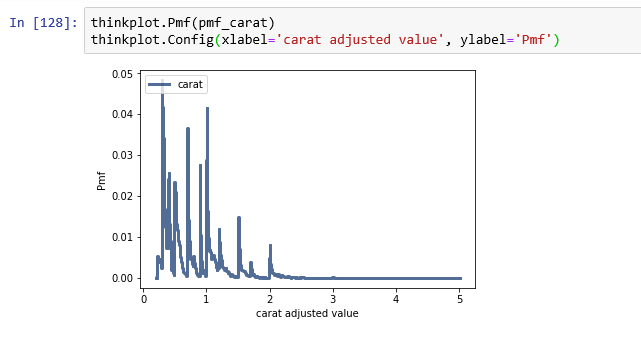


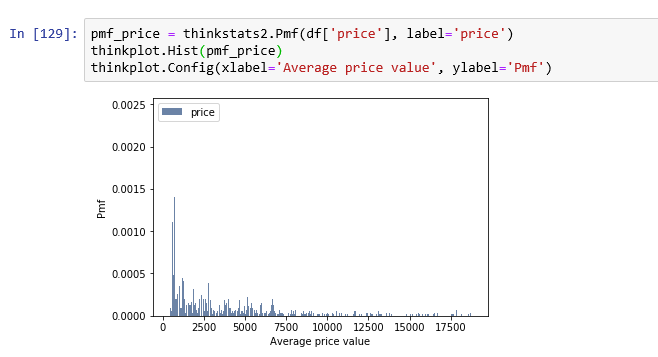


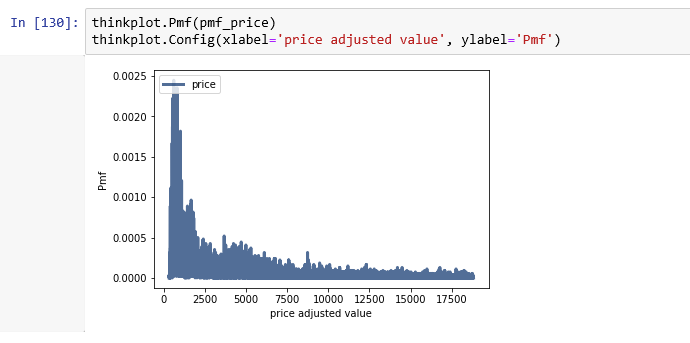
-----------------------------------------------------------------------------------------------------------------------------------------

* Using pg. 29 of your text as an example, compare two scenarios in your data using a PMF. Reminder, this isn’t comparing two variables against each other – it is the same variable, but a different scenario. Almost like a filter. The example in the book is first babies compared to all other babies, it is still the same variable, but breaking the data out based on criteria we are exploring (Chapter 3).



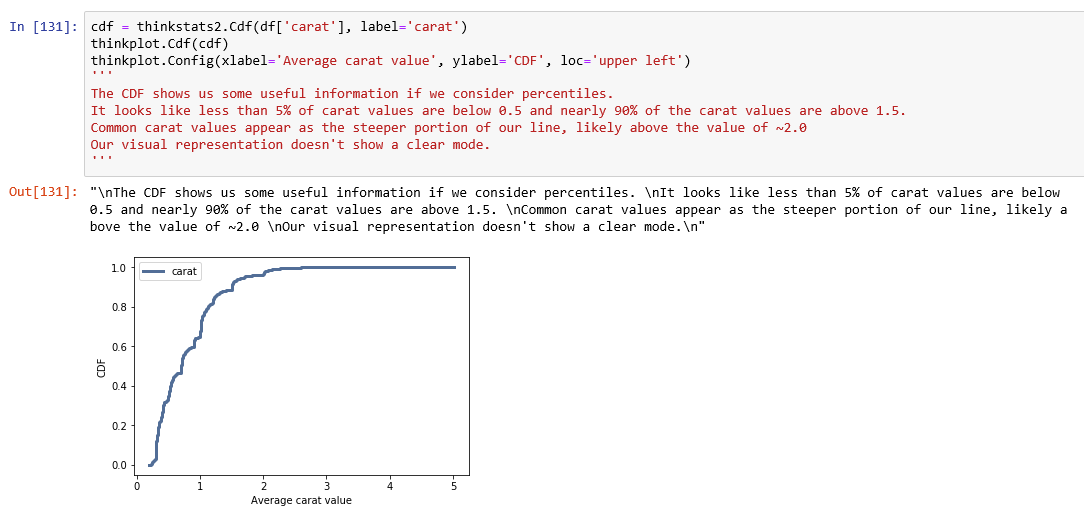






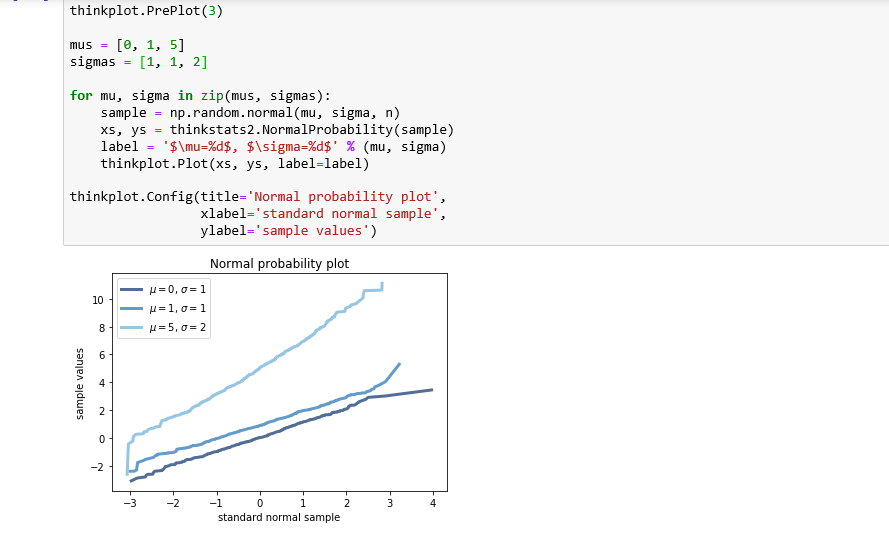
-----------------------------------------------------------------------------------------------------------------------------------------

* Create 1 CDF with one of your variables, using page 41-44 as your guide, what does this tell you about your variable and how does it address the question you are trying to answer (Chapter 4).



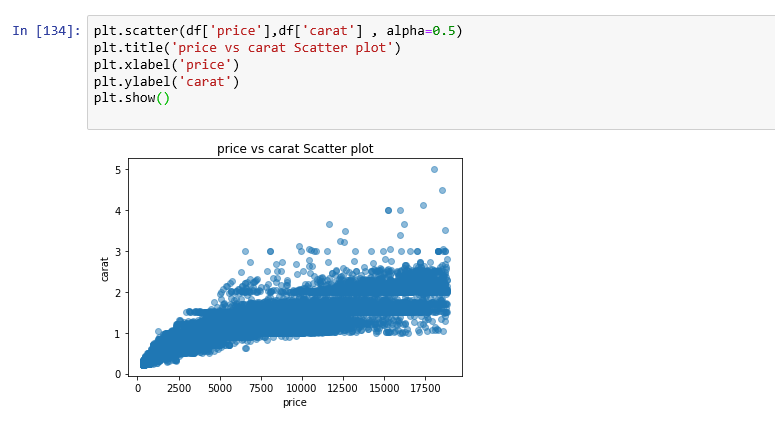
-----------------------------------------------------------------------------------------------------------------------------------------

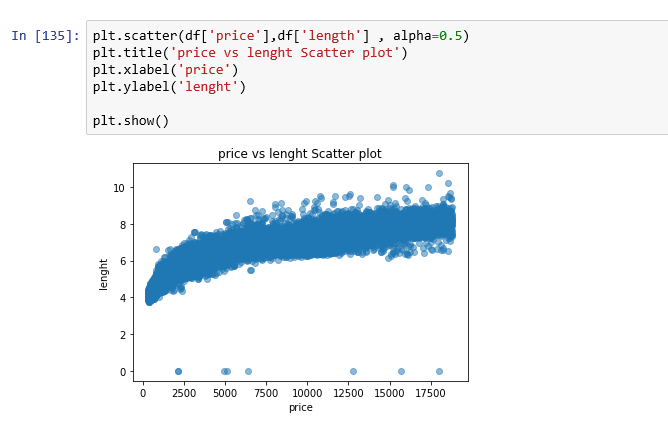
* Plot 1 analytical distribution and provide your analysis on how it applies to the dataset you have chosen (Chapter 5).

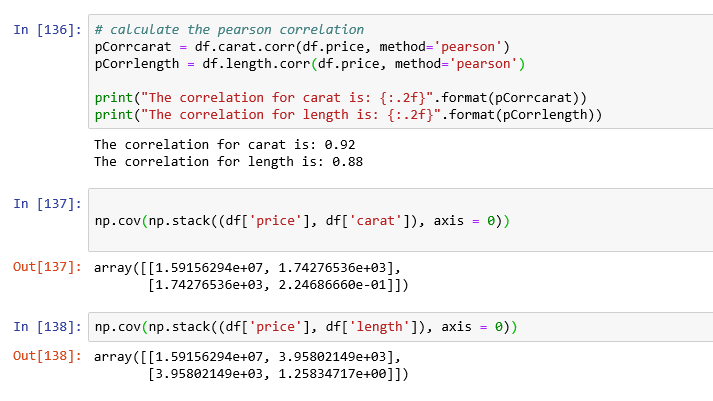


-----------------------------------------------------------------------------------------------------------------------------------------

* Create two scatter plots comparing two variables and provide your analysis on correlation and causation. Remember, covariance, Pearson’s correlation, and Non-Linear Relationships should also be considered during your analysis (Chapter 7).

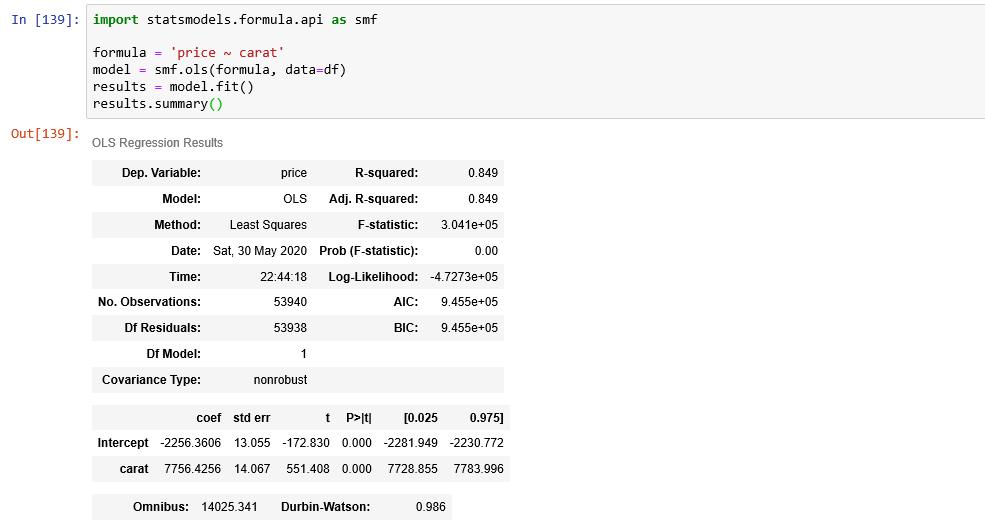


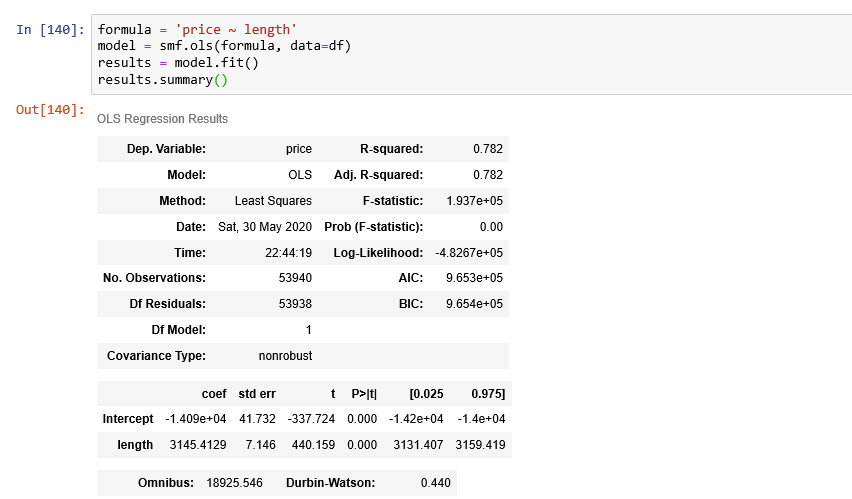




-----------------------------------------------------------------------------------------------------------------------------------------

* For this project, conduct a regression analysis on either one dependent and one explanatory variable, or multiple explanatory variables (Chapter 10 & 11).





-----------------------------------------------------------------------------------------------------------------------------------------

-----------------------------------------------------------------------------------------------------------------------------------------

# Introduction

With the Stockmarket fall because of the Corona virus situation, I was surprized to see how the gold and diamonds prices are rising. I guess people are looking for safe investments, which is causing the prices go high. This generally happens when the markets are uncertain.

I got intrested to do some data analysis regarding the diamonds. As I do not have much information about diamods, I search about the data and found a dataset provided by Shivam Agarwal in kaggle. This dataset containes huge amount of details and properties of the diamond like the price, shape, and other attributes. Looking at the data I choose this dataset to do my final project on EDA.

This dataset has the given variables with almost 52000 observations.

Variables:

carat

cut

color

clarity

depth

table

price

lenght

width

# Statistical/Hypothetical Question

1. what factors are influence the price of the diamonds.
2. which characteristics of the diamond influence the price.
3. How people prefer cut vs clarity which will influence the price.
4. How lower in the overall characteristics or quality of diamond will increase the outliners.
5. How very rare diamonds influence the price prediction.

# Outcome of your EDA

Through the analysis explored in Bgaggainpali\_DSC530\_FinalProject\_diamonds.ipynb I was able to answer the above questions.

# What do you feel was missed during the analysis?

The overall I see the price of the diamond was also depend on the regions as in some countries specific flavour of diamonds are most liked by people and are in more demand. So would like to extend the exploration in that areas.

# Were there any variables you felt could have helped in the analysis?

The Variable I was thinking would help was the structure of the diamond, which is the raw vs the polished levels of the shine and useage. The Region is another variable to consider as the price varies with different countries.

# Were there any assumptions made you felt were incorrect?

In the data set when I first looked at it, I was under the impression that the price of the diamond depends on the size of the diamond but as per the analysis its later observed that the higher the carat value is what makes the price higher.

# What challenges did you face, what did you not fully understand?

I have faced difficulty in understanding hypothesis concepts and overall challenges came from applying statistical methods and formulas as I do not have much prior experience in Stats.

# Submit a link to your repository to the assignment link during the final week of class.

All files associated with project can be found at:

<https://github.com/bgaggainpali/bgaggainpali_DSC530>