Proposal for 6105
Final Project -- Shine bright like a diamond
Yuqing Wang

Problem Statement and Background

As we all know, the price of a diamond is not only up to the size, but also be affected by colors, clarity, depth, width and the quality of cut. But these factors affect the price of the diamond in a different way. So the purpose of my project is to figure out how exactly these factors affect the price of a diamond and try to find a perfect way to predict a diamond's price.

Objectives

1.display the distributes of all factors.
2. find out how these factors affect the price of diamonds.
3. figure out how these factors affected each other and the correlation among these factors.
4.find a way to predict prices by size, cut, clarity, width, depth, and color.
5. use some method to access the accuracy of different models.

Methods

1.display the distributes of all factors. (histogram)
2. find out how these factors affect the price of diamonds. (regression and dot graphs)
3. figure out how these factors affected each other and the correlation among these factors.(use heatmap show the correlations among all the factors)
4.find a way to predict prices by size, cut, clarity, width, depth, and color. (train_test_split, linear regression, lasso regression and ridge regression)
5. use some method to access the accuracy of different models. (validation_curve and accuracy score in metrics)

Dataset

1. price in US dollars ($326 - $18,823)
2. carat weight of the diamond (0.2--5.01)
3. cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. color diamond color, from J (worst) to D (best)
5. clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. x length in mm (0--10.74)
7. y width in mm (0--58.9)
8. z depth in mm (0--31.8)
9. depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
10. table width of the top of diamond relative to widest point (43--95)

Plans of Action

I’m really interested in this dataset because it is so classic and crystal clear. So I will use some graphs to show a rough distribution of the context. And then choose some obvious correlations to analyze. After I know perfectly about how different factors affect the price, I

will train a suitable model to predict the price by factors.

References

https://www.kaggle.com/shivam2503/diamonds