Project 4

Inesh Chakrabarti, Lawrence Liu, Nathan Wei

Introduction

For the first part of this project we will do regression analysis. The dataset we chose to use is one of diamond characteristics. We will conduct regressions to predict the price of a diamond given some features.

Dataset

Let us begin by understanding the dataset. The dataset consists of information about 53940 round-cut diamonds with ten features:

Feature	Description
`carat`	weight of the diamond (0.2–5.01)
`cut`	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
`color`	diamond colour, from J (worst) to D (best)
`clarity`	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
`x`	length in mm (0–10.74)
`y`	width in mm (0–58.9)
`z`	depth in mm (0–31.8)
`depth`	total depth percentage
`table`	width of top of diamond relative to widest point (43-95)
`price`	price in US dollars ($326-$18,823)

Question 1.1

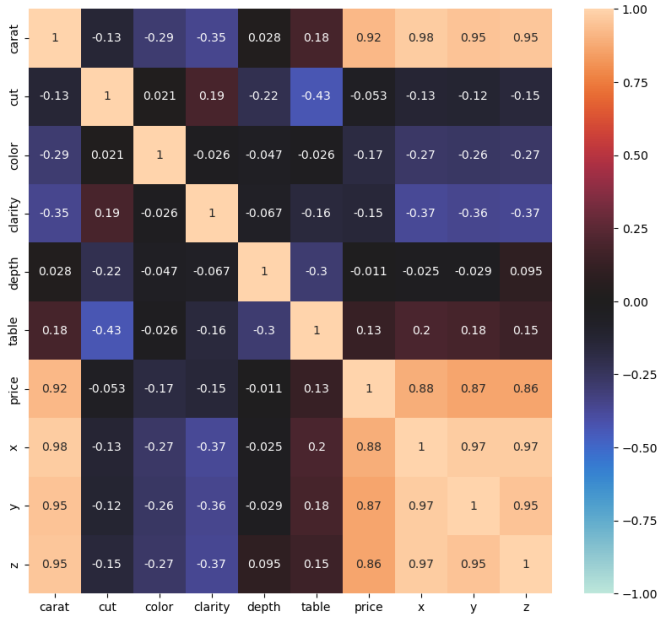

Figure 1: Feature Pearson Correlation Heatmap

We will be using the nine features to predict the `price`. We can begin by computing the Pearson correlation matrix heatmap for these features in the dataset in Figure 1. The values for correlation for `price` is given in Table 1(a). We see that, unsurprisingly, there is a massive collection of high correlation squares at the bottom right. These indicate high Pearson correlation coefficient between `price` and `x`, `y`, textttz. This suggests that

	Correlation
carat	0.9215914337868304
cut	-0.05349263851362828
color	-0.1725093772499559
clarity	-0.14680175361025616
depth	-0.010647725608533299
table	0.12713358133531918
x	0.8844357793744166
y	0.865421694764742
z	0.861250266123968

(a) Price

	Correlation
carat	0.7694571626172851
cut	0.00542011950342582
color	-0.011980043670033661
clarity	0.04512538515850012
depth	-0.03572374489729493
table	0.08458507638109278
x	0.7873455524189906
y	0.7717301198408058
z	0.7655421629234554

(b) Price per Carart

Table 1: Pearson Correlation Coefficients