ECE 289 Final Project
Price Prediction for Diamonds

Te Sun
A53200185

I. Introduction

A database that contains features and prices of 53940 diamonds from *Kaggle* is used to analyze diamonds by their cut, color, clarity and other attributes. Price prediction for diamond is conducted based on this database by applying *linear regression* method. (Kaggle dataset: https://www.kaggle.com/shivam2503/diamonds/data)

A. Original Data

This data frame contains 53940 rows and 10 variables:

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 colour, 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 top of diamond relative to widest point (43–95)

The original data is divided evenly into two datasets as training data and testing data, respectivley.

B. Linear Regression Model and Features

Regression is one of the simplest supervised learning approaches to learn relationships between features and predictions. A linear regression in the following form is applied here to predict the price of diamonds:

$$Y = \theta_1 + \theta X \tag{1}$$

Where Y and X indicates the matrice of predictions and features and θ is the parameter vector obtained.

Feature 3, feature 4 and feature 5 are binary features with lengths of 5, 7 and 8, respectively. A "1" indicates that datum has this related characteristic. For example, [0, 1, 0, 0, 0] for feature 3 of datum means it's a "Good" cut. Therefore, each datum has 27 features.

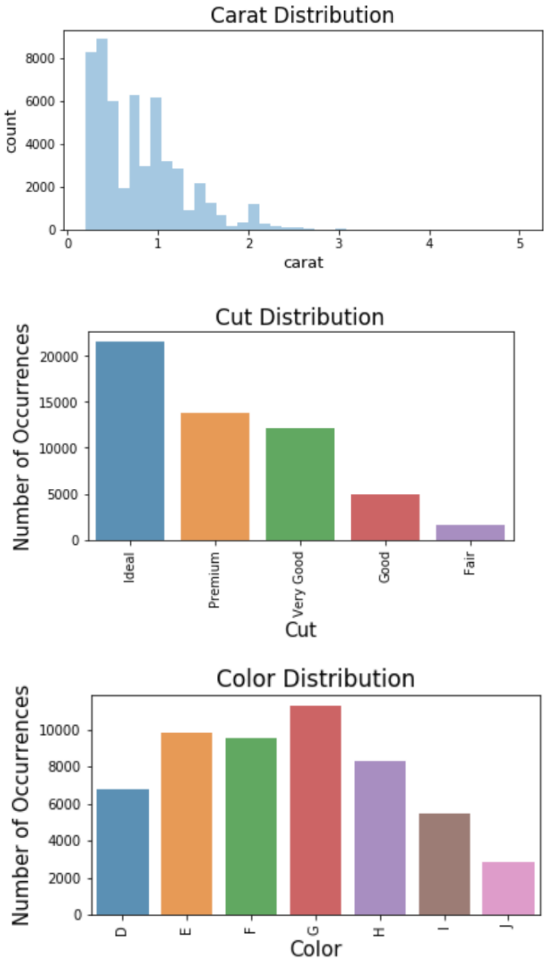

Fig. 1. Features Distribution

II. Data Analysis and Visualization

A. Features Distribution

The following figures show the distribution of carat, cut, color and clarity.

Table 1 lists the mean, variation and standard deviation value of each feature.

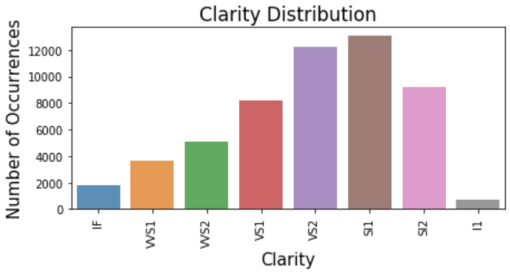

TABLE I. MEAN, VARIANCE AND STANDARD DEVIATION OF FEATURES

Features	Mean	Variance	Standard Deviation
carat	0.798	0.2247	0.4740
x	5.731	1.2583	1.1218
y	5.735	1.3045	1.1421
z	3.539	0.4980	0.7057
depth	61.749	2.0524	1.4326
table width	57.457	4.9929	2.2345

B. Correlations of Features

Correlations of each pair of features are calculated using *corr()* in *seaborn* package. Results are shown in Figure 2.

C. Data Normalization

According to the distribution of features, feature 6 and feature 10 can be normalized using feature scaling for each datum, which brings all values into the range [0,1]:

$$X_{normlized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Prices of diamond are also normalized during data training and denormalized after that.

III. RESULTS

Linear Regression is conducted using *linalg.lstsq* from *numpy* package which returns a least-squares solution to a linear matrix equation.

The matrix of parameter θ is obtained, shown in figure 3.

Root-mean-square error (RMSE) in floowing form is applied to evaluate the prediction result:

$$RMSE = \sqrt{\frac{\sum (y_{prediction\_i} - y_i)^2}{n}} \tag{3}$$

The RMSE of training set and test set are 1078.6044 and 1168.7067, respectively.

IV. CONCLUSION

The θ obtained in linear regression model represent how "important" and "effective" those features are. The distribution of cut is decreasing from "ideal" to "fair" and the diamonds with color "D" to "G" are more than "H" to "J", which indicates the high quality and price of diamonds.

Some features pairs show a high correlation such as x-y-z and carat, since a larger dimension indicates a larger size.

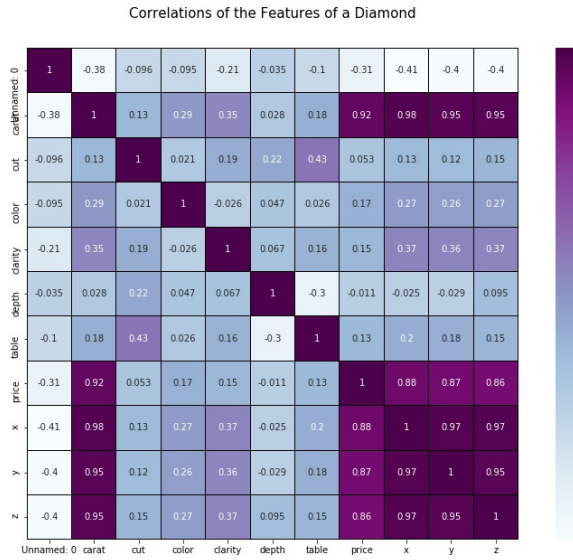

Fig. 2. Correlations of Features

```
theta =  [  2298.29115299  11323.70970458    416.90618626    -75.18764075
    566.96151757    719.30782985    670.30326006    942.80496889
   1169.47680748  -1201.19525308   -323.46613626    874.24152534
    671.76241604    164.66682458    125.45861547   -850.3643497
   1046.15394254   1402.82028048    737.8625345    1507.30936052
  -3455.34129167   1784.39206085    -31.75180511    -18.18903976
  -2076.12680837   1278.64978796   -408.3142044 ]
```

Fig. 3. Theta Value

Prices are also mostly correlated with x-y-z and carat since a larger size always leads to a higher price.

Linear regression models performs well for price prediction with RMSE of 1078.6 on training set and 1168.7 on test set, considering a price range of $326 to $18823.

There are still something to be improved. Standard deviations of features such as carat, depth and size are relatively small. Therefore it's reasonable to normalize them in linear form. However, The price distribution is not uniform so the data normalization in simple linear form might not fit very well. The model might performs better if price distribution are considered exponential distribution and then normalized them into [0,1].