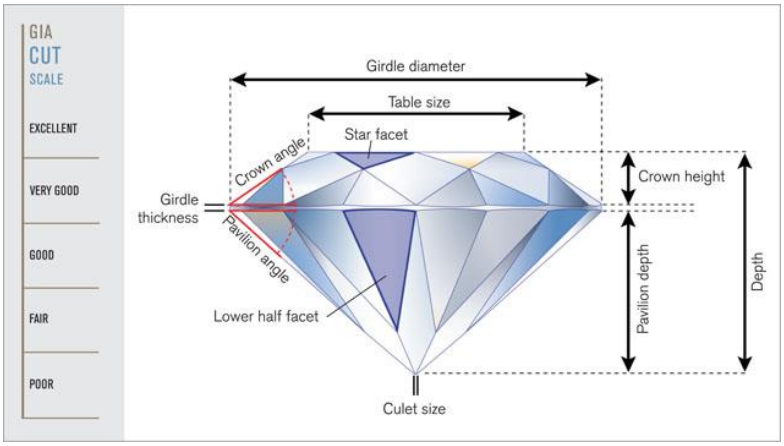
Diamonds Analysis



Team 5

Omar Gamal AbdelMageed AbdelGawad	20201701634	G2
Omar EmadEldeen Anwar Metwaly Gebal	20201701635	G2
Mariam Emad Roushdy Habashy Issak	20201701654	G3
Ahmed Ramy Ahmed Sayed	20201701602	G3

Table of Contents

Introduction

This report summarizes the statistical modeling and analysis results associated with a Diamonds dataset (1). This dataset is originally from Tiffany & Co's pricelist from 2017. The Diamonds dataset is composed of some features related to the diamonds like their cut, color, clarity, price, and other attributes.

Purposes And Methodology

The first purpose of this report is to document the diamonds' features using descriptive statistics. Another purpose of this report is to find the best regression model to predict the prices of diamonds, we'll compare between single variate linear and nonlinear regression, and multi variate linear and nonlinear regression. The third purpose is to predict the dimensions of diamonds depending on the price using single-variate linear regression. We also created a classification model to predict the clarity of diamonds based on their features using a neural network.

Objective

Our alternate hypothesis for the second purpose is that there is a strong positive relation between the carat of the diamond and the price. Whereas, if the diamond's weight in carat increases, the price increases. For the third purpose, the alternate hypothesis is that there is a strong positive relationship between the price and the x – length -, y – width - and z – depth - of the diamond. So, if the price paid for the diamond increases the size of the diamond also increases.

Report Structure

The structure of the report is in the same order as the purposes have been listed. Firstly, we will start with the descriptive statistics of the features, then we will dive into comparing the regression models of the prices. After that we will present the single variate linear regression of the dimensions. Lastly, we will present our classification model that predicts clarity.

Body

Data Cleaning

We started off by cleaning our dataset. Firstly, we replaced all the categorical values – cut, color, clarity - with numeric values that represent them. Then we checked for null values and found none. Lastly, we removed the outliers by checking if the z-score lies between 3 and -3. Which has resulted in a loss of 2,350 rows out of 53,940. Now our data lies at 51,590 rows.

Describing Our Dataset

Now that our dataset has 51,590 rows, we will start describing our variables.

Here is an overview of our dataset, these are its first 5 rows showing all the variables.

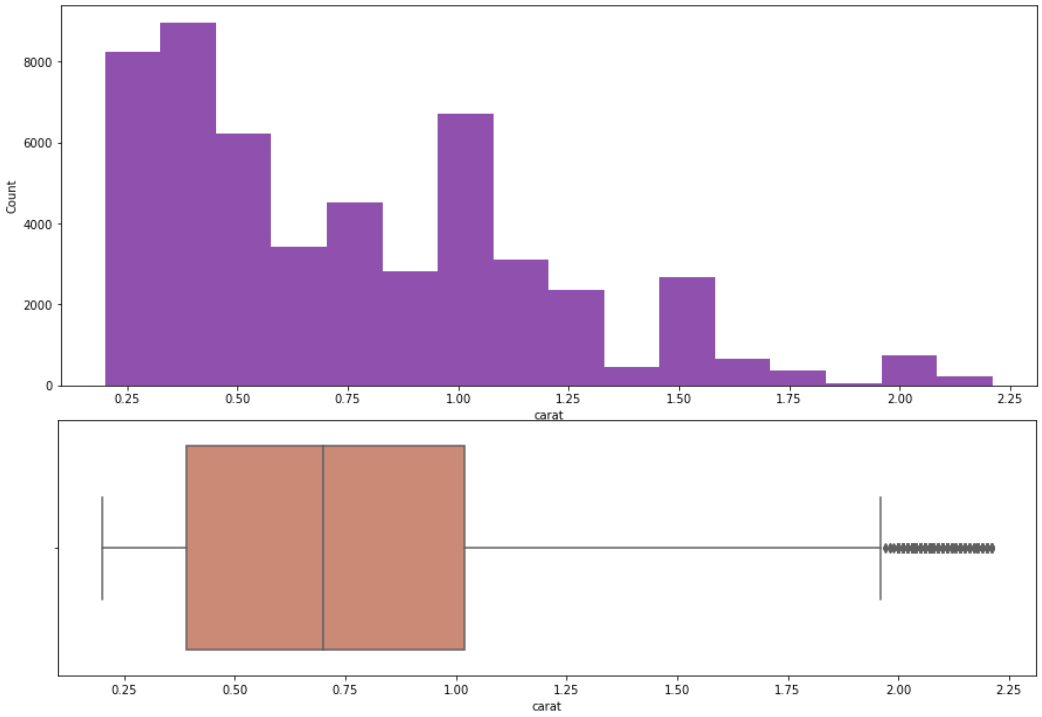
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

- In all our histograms we used 16 classes that came from 1 + 3.33 log (51,590).

1. Carat

The carat is the weight of the diamond.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' carat is a positively skewed distribution. That means that most of the diamonds are light in weight.

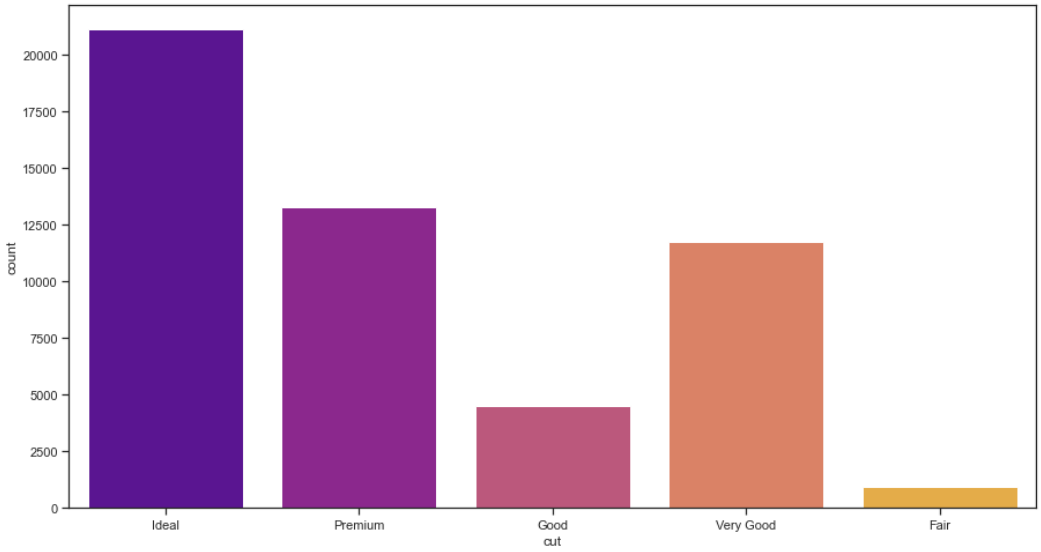
Proof:

Mean	0.7599 carat	The average weight of diamonds is 0.7599 carats, while the median is 0.7 carats and the mode is 0.3 carats, which confirms that the diamonds are positively skewed (mean > median > mode).
Median	0.7 carat	
Mode	0.3 carat	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{0.425}{0.7599} = 0.5592$	This means that carat values vary a lot (2).

2. Cut

It describes quality of the cut of the diamond. It varies from fair to ideal.

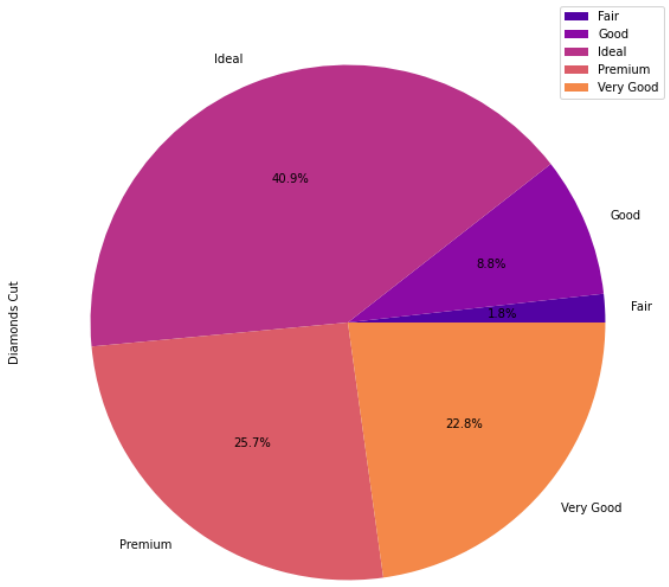
Since the cut is categorical ordinal data, we used a bar chart to visualize it.



Mode	Ideal	This shows that the diamonds are mostly cut of good quality. and that the cut that is done most frequently is an ideal cut.

Then we visualized it using a pie chart to compare the cut qualities percentages.

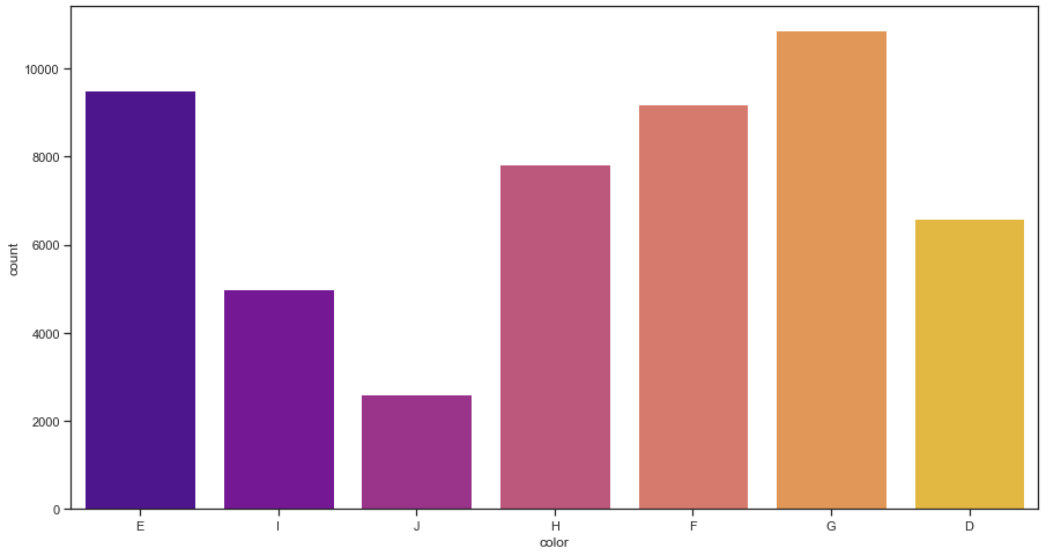
The pie chart confirms that most of the diamonds are of Ideal or Premium or Very Good cut.



3. Color

It describes the color of the diamond. It varies from J (worst) to D (best).

Since the color is categorical ordinal data, we used a bar chart to visualize it.

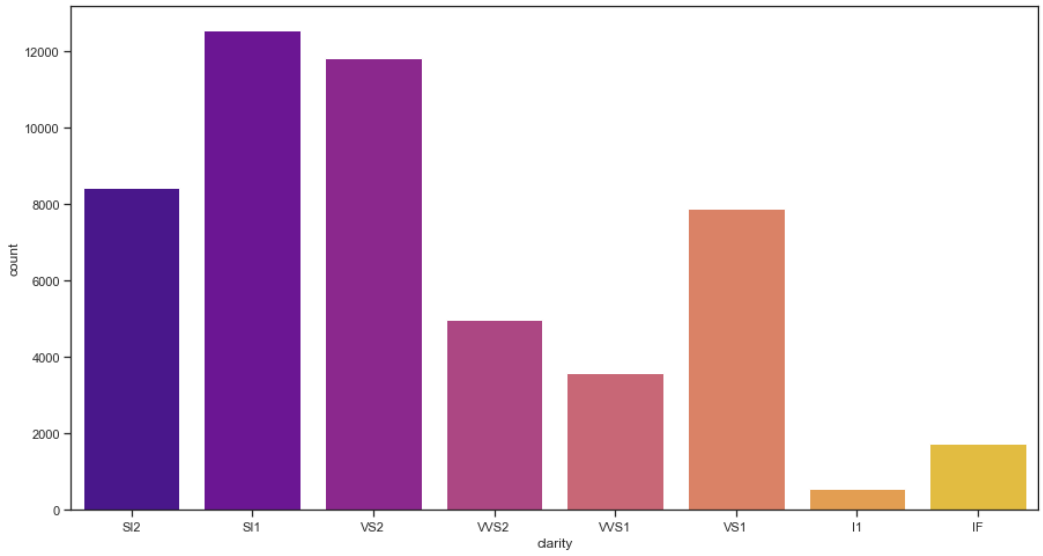


Mode	G	This shows that the color that appears most frequently in diamonds is G.

4. Clarity

Clarity - how obvious inclusions are within the diamond. A measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)).

Since the clarity is categorical ordinal data, we used a bar chart to visualize it.

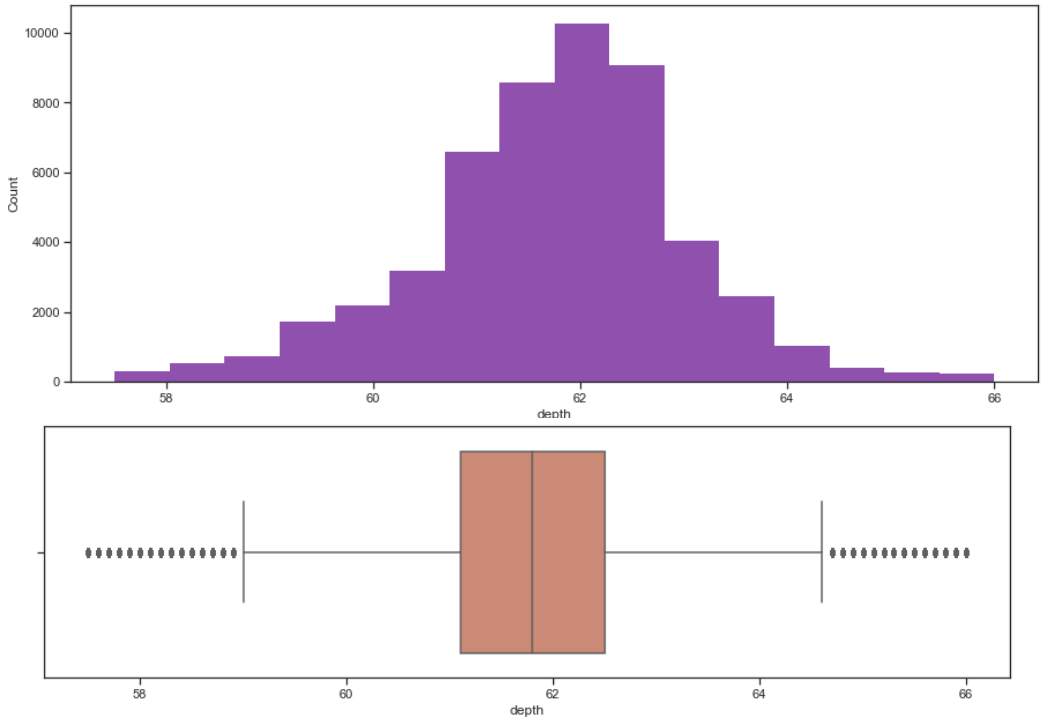


Mode	SI1	This shows that the most common clarity is SI1 which is the second worst clarity and that most of the diamonds are of bad clarity.

5. Depth Percentage

Depth percentage is the height of a diamond, measured from the culet to the table, divided by its average girdle diameter.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' depth percentage is almost normal distribution.

That means that most of the diamonds' depth percentage is approaching the mean.

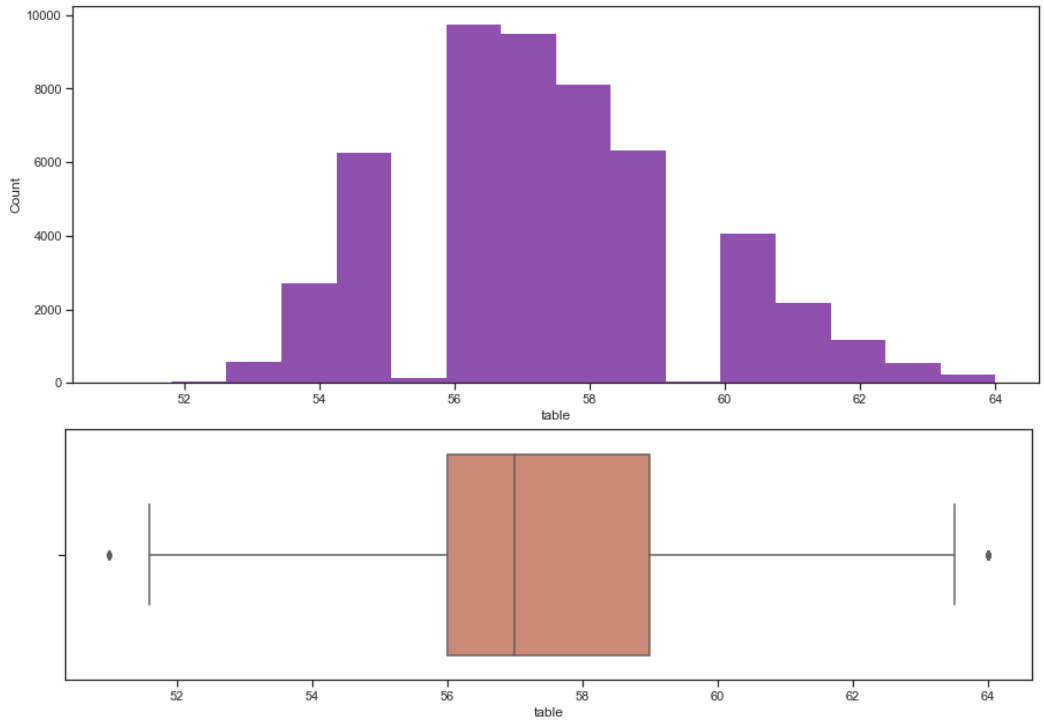
Proof:

Mean	61.7528%	The average depth percentage of diamonds is 61.7528%, while the median is 61.8% and the mode is 62%. Which means that mean, median and mode are approximately equal which confirms the diamond's depth percentage is distributed normally as stated before (mean = median = mode).
Median	61.8%	
Mode	62%	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{1.2963}{61.7528} = 0.0206$	The table percentage's coefficient of variation gives us a feeling of good method performance (2).

6. Table Percentage

Table percentage is the width of the diamond's table expressed as a percentage of its average diameter.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' table percentage looks as if it is almost normal distribution, however, it is positively skewed. This means that most of the diamonds' table percentages are low.

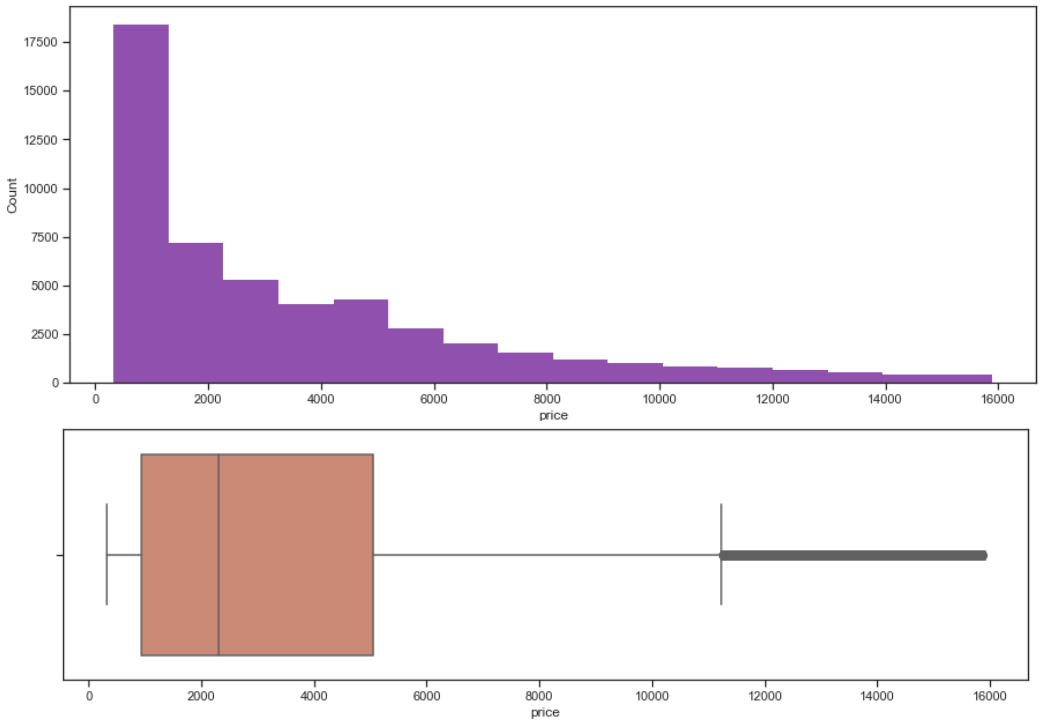
Proof:

Mean	57.3691%	The average table percentage of diamonds is 57.3691% which is greater that the median which is equal 57.0%, and both are greater than the mode which is equal to 56.0%. This confirms the diamond's table percentage distribution is positively skewed as stated above (mean > median > mode).
Median	57.0%	
Mode	56.0%	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{2.1}{57.3691} = 0.0366$	The table percentage's coefficient of variation gives us a feeling of good method performance (2).

7. Price

The diamonds' price in US dollars.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' price is a positively skewed distribution. That means that most of the diamonds are low priced.

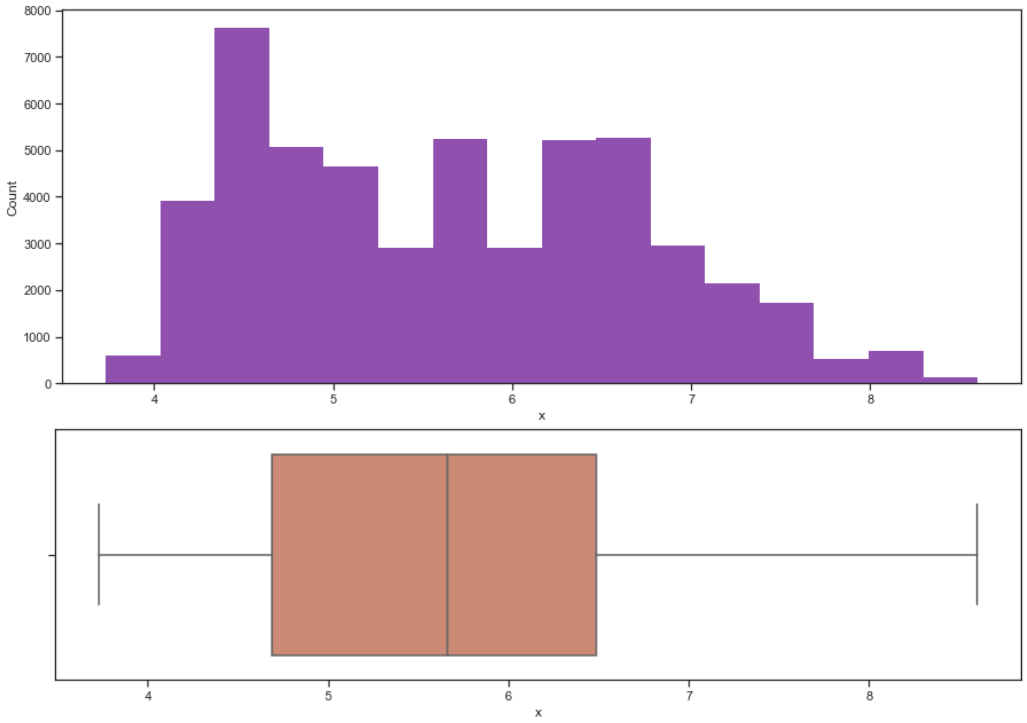
Proof:

Mean	3586.0318 USD	This shows that the average price of a diamond is 3586.0318 USD, while the median is 2303.5 USD, and the mode is 605. Which confirms that the diamonds' prices are positively skewed (mean > median > mode).
Median	2303.5 USD	
Mode	605 USD	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{3436.8742}{3586.0318} = 0.9584$	Which means that the price varies drastically (2).

8. X – Length in mm

The diamonds' x - length in mm.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' length is a positively skewed distribution. That means that most of the diamonds' lengths are small.

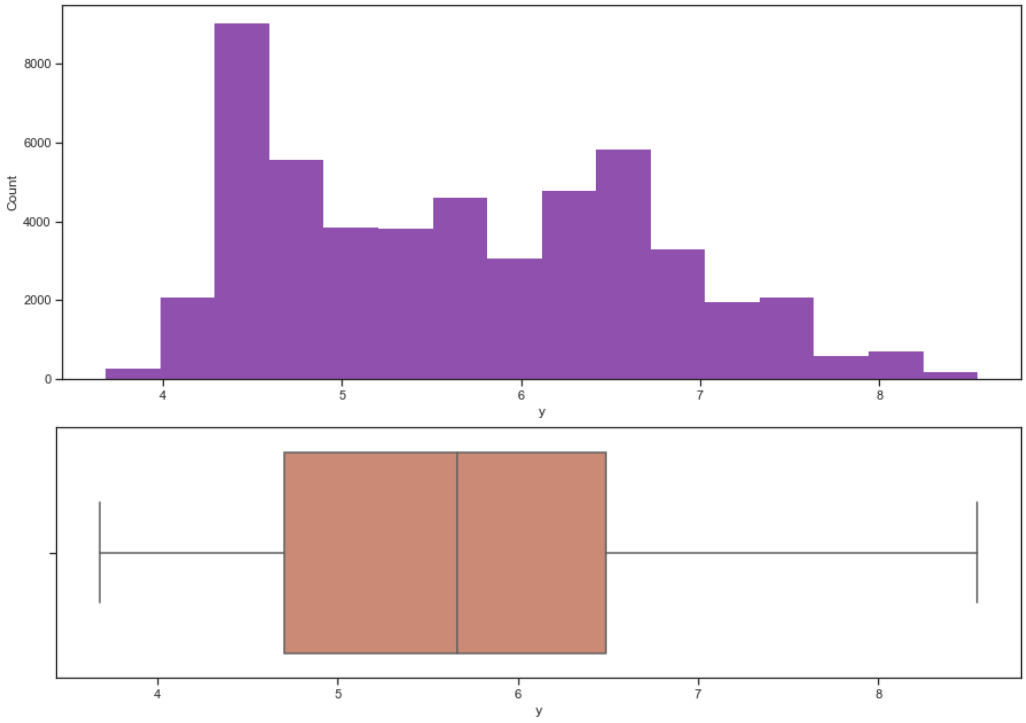
Proof:

Mean	5.6579 mm	This shows that the average length of a diamond in mm is 5.6579 mm which is approximately equal to the median which is 5.66 mm. However, the mode which is equal to 4.37 mm is smaller than both mean and the median. This confirms that the diamonds' lengths are positively skewed ((mean = median) > mode).
Median	5.66 mm	
Mode	4.37 mm	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{1.0574}{5.6579} = 0.1869$	Which means that the length varies a lot (2).

9. Y – Width in mm

The diamonds' y - width in mm.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' width is a positively skewed distribution. That means that most of the diamonds' widths are small.

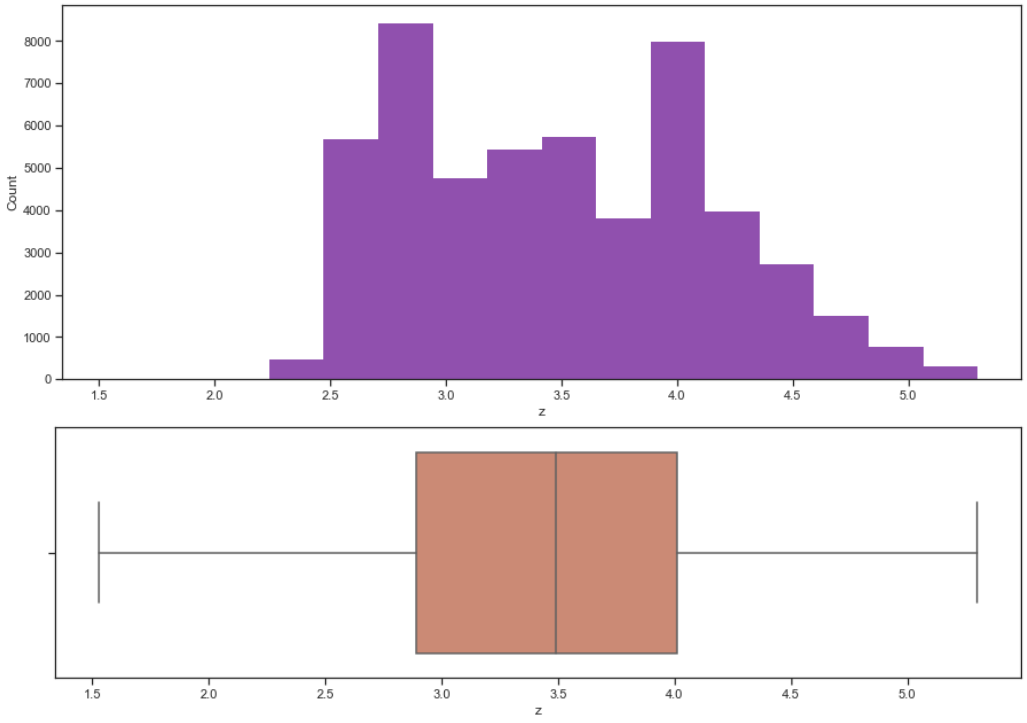
Proof:

Mean	5.6612 mm	This shows that the average width of a diamond in mm is 5.6612 mm which is approximately equal to the median which is 5.66 mm. However, the mode which is equal to 4.34 mm is smaller than both mean and the median. This confirms that the diamonds' widths are positively skewed ((mean = median) > mode).
Median	5.66 mm	
Mode	4.34 mm	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{1.0504}{5.6612} = 0.1855$	The width varies a lot [(2)]. Notice that its values are very close to the diamond's length which may give an indication that they are correlated.

10. Z – Depth in mm

The diamonds' z - depth in mm.

It is quantitative continuous data; we used a histogram and a boxplot to visualize it.



This shows that the diamonds' width is a positively skewed distribution.

That means that most of the diamonds' widths are small.

Proof:

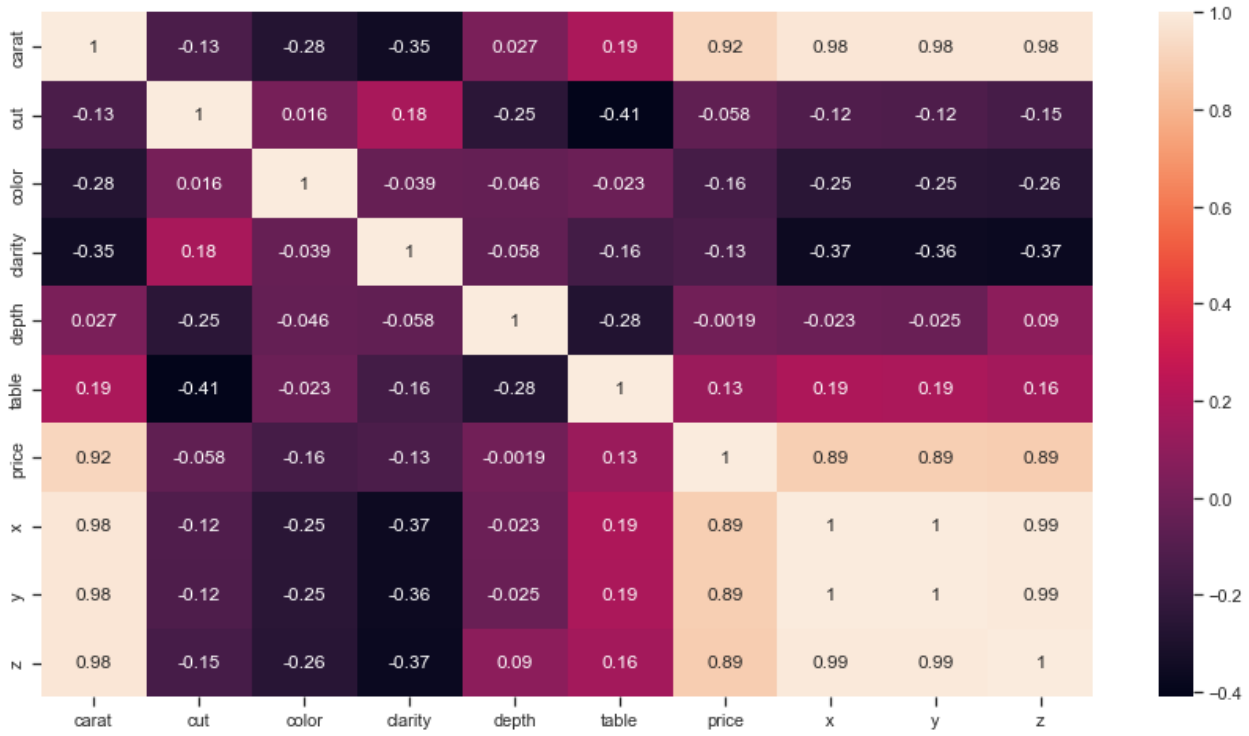
Mean	3.4946 mm	This shows that the average depth of a diamond in mm is 3.4946 mm, and the median is 3.49 mm, and the mode is equal to 2.7 mm. This confirms that the diamonds' depths are positively skewed (mean > median > mode).
Median	3.49 mm	
Mode	2.7 mm	
Coefficient of Variation	$\frac{\sigma}{\mu} = \frac{0.6531}{3.4946} = 0.1869$	The depth varies a lot (2).

A summary of all our features:

	count	mean	std	min	25%	50%	75%	max
carat	51590	0.759933	0.424983	0.20	0.39	0.70	1.02	2.21
cut	51590	2.952549	1.070666	0.00	2.00	3.00	4.00	4.00
color	51590	3.433553	1.694698	0.00	2.00	3.00	5.00	6.00
clarity	51590	3.086994	1.642585	0.00	2.00	3.00	4.00	7.00
depth	51590	61.752838	1.269255	57.50	61.10	61.80	62.50	66.00
table	51590	57.369137	2.100018	51.00	56.00	57.00	59.00	64.00
price	51590	3586.031847	3436.874161	326	926	2303.5	5047	15898
x	51590	5.657855	1.057420	3.73	4.69	5.66	6.49	8.60
y	51590	5.661272	1.050383	3.68	4.70	5.66	6.49	8.55
z	51590	3.494649	0.653051	1.53	2.89	3.49	4.01	5.30

We constructed a correlation matrix between all variables:

This matrix shows us the r's complement between all the variables.

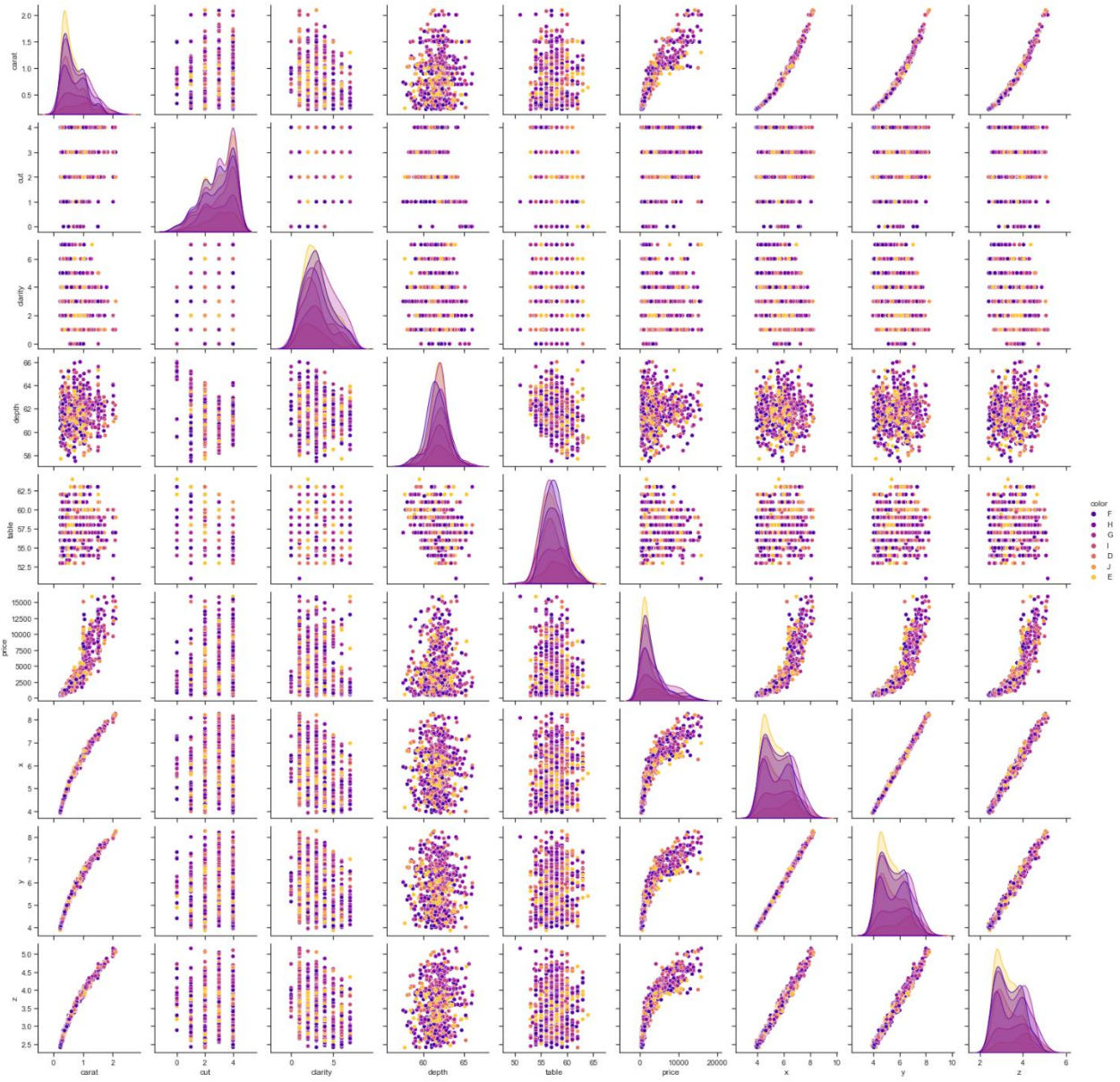


Since our goal is to predict the price of the diamonds, we will use the price as our target variable and calculate all the correlations related to it.

carat	cut	color	clarity	depth	table	price	x	y	z
0.9224	-0.0585	-0.1554	-0.1334	-0.0019	0.1317	1	0.8905	0.8918	0.8873

As we can see the strongest correlation is between the carat and the price. Right after it, comes the dimensions of the diamond (x, y, z) and the price.

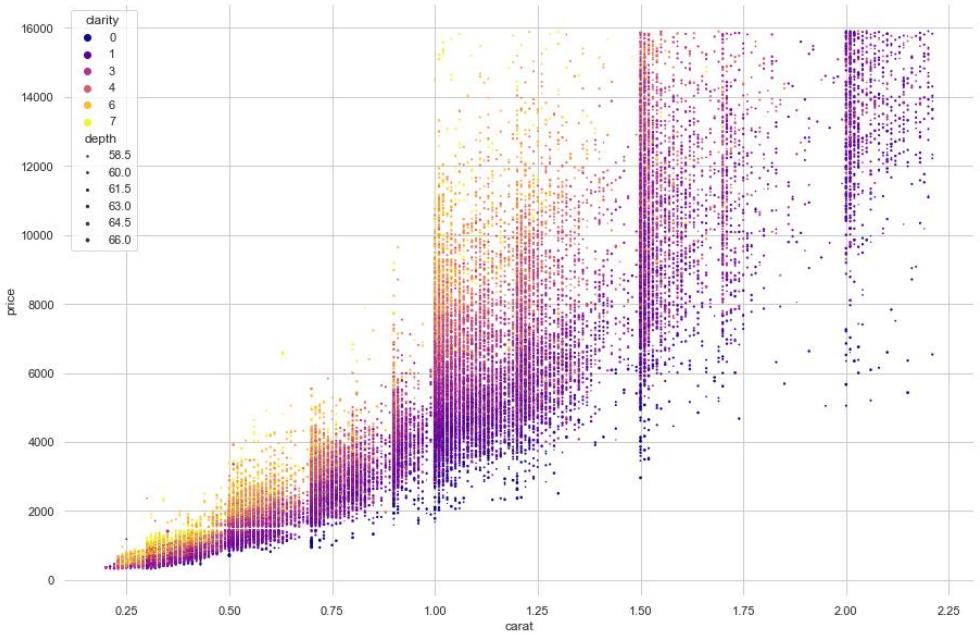
To help us decide on the type of regression and which data to correlate together more we will use a pair plot that does a scatter plot between all the variables and colors them according to the diamond color



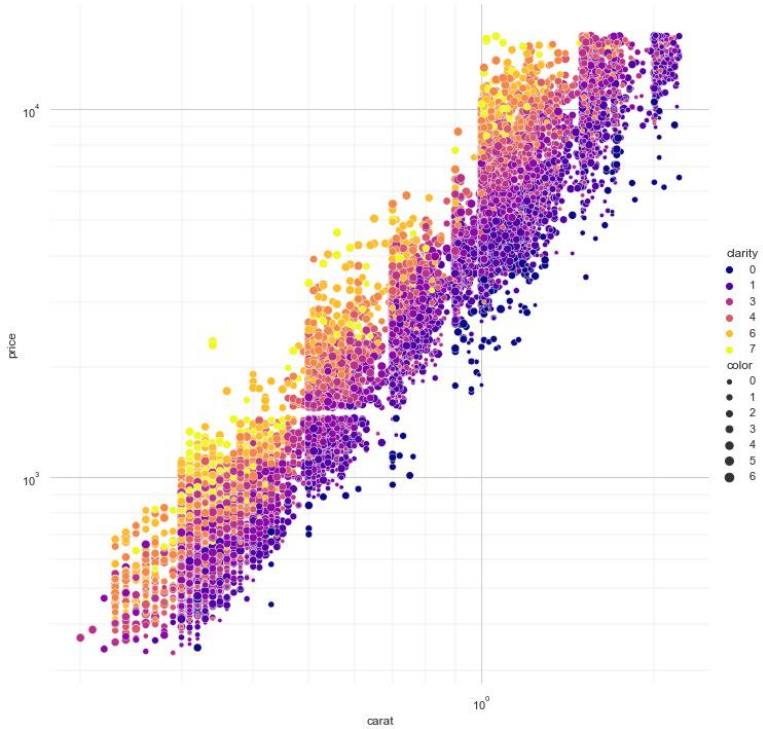
This shows us that the price linear regression can be done with the carat variable. This also shows us that we might get better results using nonlinear regression since the data seems to form a curve. Since there is a strong correlation between the price and other variables. We will try using multivariate linear and nonlinear regression.

This also shows that the dimensions (x, y, z) linear regression can be done with the price variable. We will use single-variate linear regression.

Here is a better look on two different scatter plots of the carat against the price.



Scatter Plot Between Carat and Price in Normal Scale



Scatter Plot Between Carat and Price in Logarithmic Scale

Single Variate Regression – Carat and Price

Linear Method 1 – Diamond Price Prediction from Carat Using Correlation Coefficient

We begin by getting the correlation coefficient (r) between the carat and the price:

$r = 0.92011$ **Strong Positive Relationship**

Then we calculate B_0

$$B_0 = r * \frac{\sigma_{price}}{\sigma_{carat}} = 0.92011 * \frac{3436.8742}{0.425} = 7410.7696$$

And B_1

$$B_1 = \mu_{price} - B_0 * \mu_{carat} = 3586.0318 - 7410.7696 * 0.7599 = -2066.7833$$

Finally, we calculate y_0

$$y_0 = B_0 x + B_1 = 7410.7696x - 2066.7833$$

We can now predict new price values

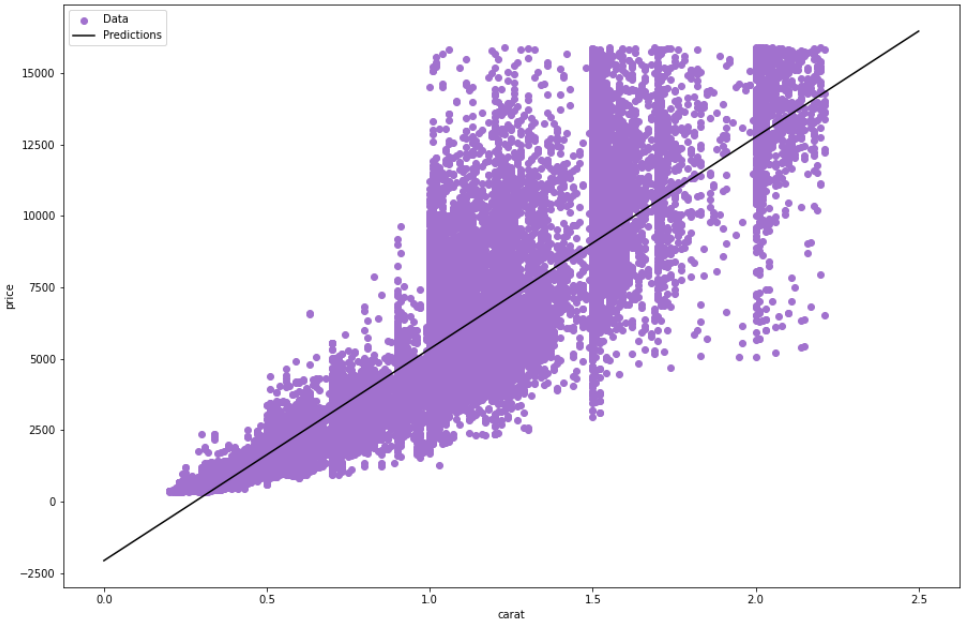
Example: $y(20) = 7410.7696(20) - 2066.7833 = 156.4476$

Whereas the real price value when $x = 20$ is 351.

$$Error = 351 - 156.4476 = 194.5524$$

This confirms our alternate hypothesis for purpose 2 that the carat and price are strong positively correlated.

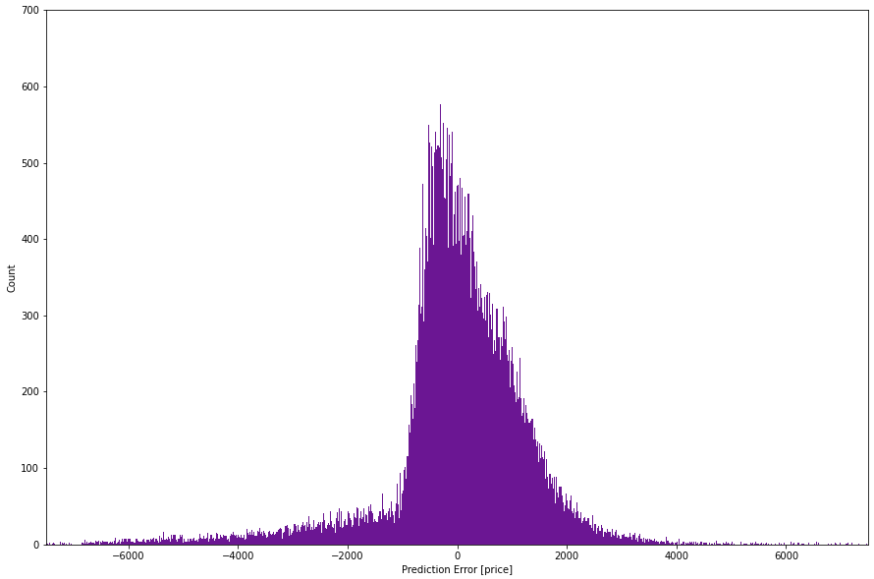
Since this is a single variable regression, it's easy to view the model's predictions as a function of the input:



To rate how well our regression line performs, we calculated the absolute mean error by calculating the absolute error of predictions out of all our records then getting their mean.

$$Absolute\ Mean\ Error = \sum_{y=0}^{n} |(y - \hat{y})| = 885.4745\ USD$$

Here is the error distribution



Linear Method 2 – Diamond Price Prediction from Carat Using Gradient Descent (One Neuron) Using TensorFlow

First, we split the dataset into a training set and a test set. We will use the test set in the final evaluation of the model.

The training set is 80% of the data and the test set is 20%.

In the table of statistics, it's easy to see that the features have different ranges. some range in the units, some in the hundreds, some in the thousands.

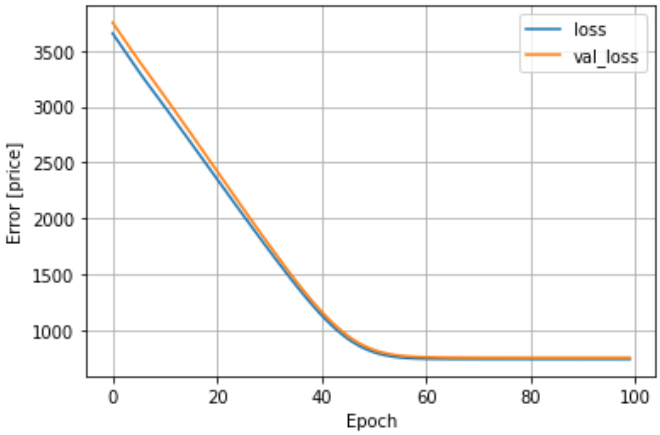
	mean	std
carat	0.785626	0.416910
depth	61.611885	1.096740
table	57.370539	2.002749
price	3715.782066	3427.572015
x	5.744671	1.032990
y	5.745005	1.026416
z	3.538966	0.633848

It is good practice to normalize features that have different ranges.

If we skip normalizing the features, the model will still converge (ie: reach the optimum loss. i.e. become able to predict and perform regression as best as it can), However, normalization makes training much more stable and faster in performance.

We applied a linear transformation y = mx + b to produce 1 output using a linear layer.

We built a linear regression model using gradient descent in a neural network.



This plot shows how the error decreases with the number of iterations. This plot confirms that the model converges to find the least error value.

After training our neural networks concluded the following the weights.

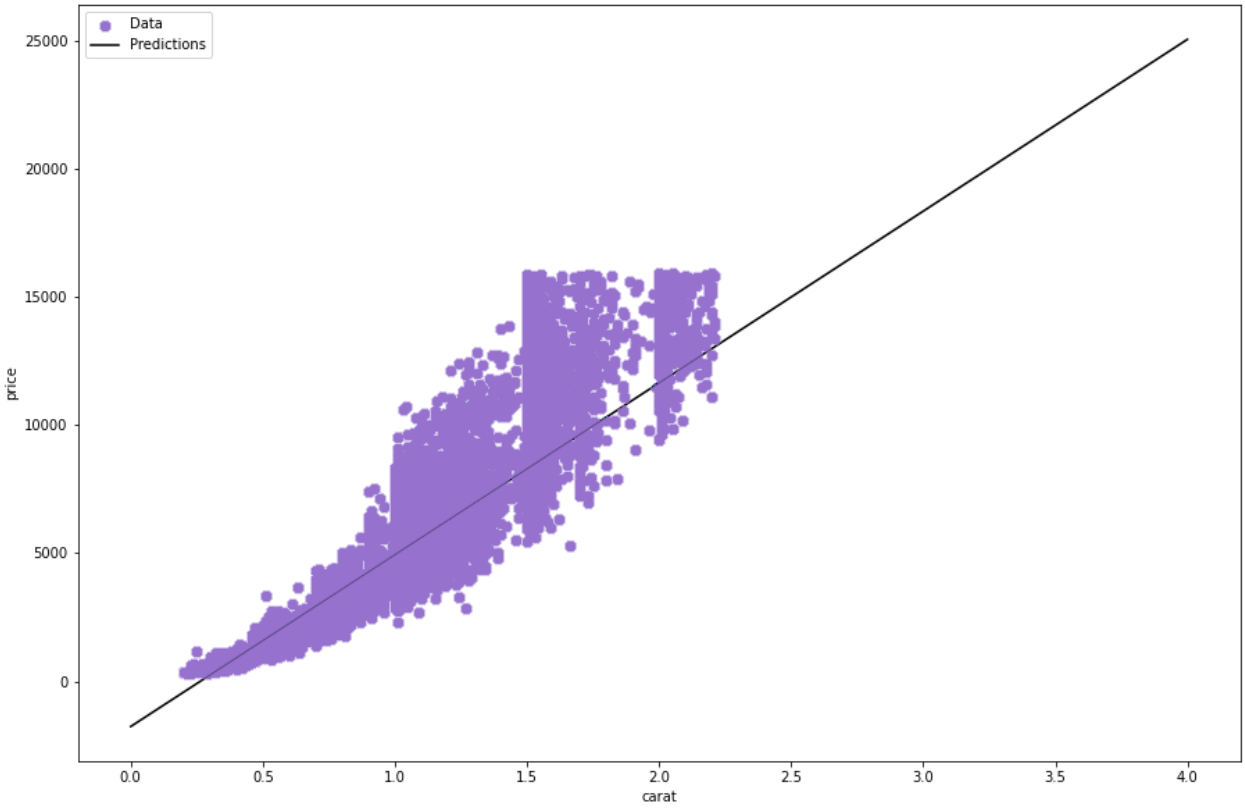
For m (slope) it will be 6662.417.

For b (y intercept) it will be -1734.53.

Our final equation will be $\hat{y} = 6662.417x - 1734.53$

Our neural network consists of 2 layers 1 for the normalization and the other for the output. Each layer contains only one neuron since this is single variate regression

Input Layer $\in \mathbb{R}^1$ Output Layer $\in \mathbb{R}^1$



To rate how well our regression line performs, we calculated the mean absolute error after the last epoch by calculating the absolute error of predictions out of all our records then getting their mean.

$$\sum_{y=0}^{n} |(y - \hat{y})| = \ 739.1205\ USD$$

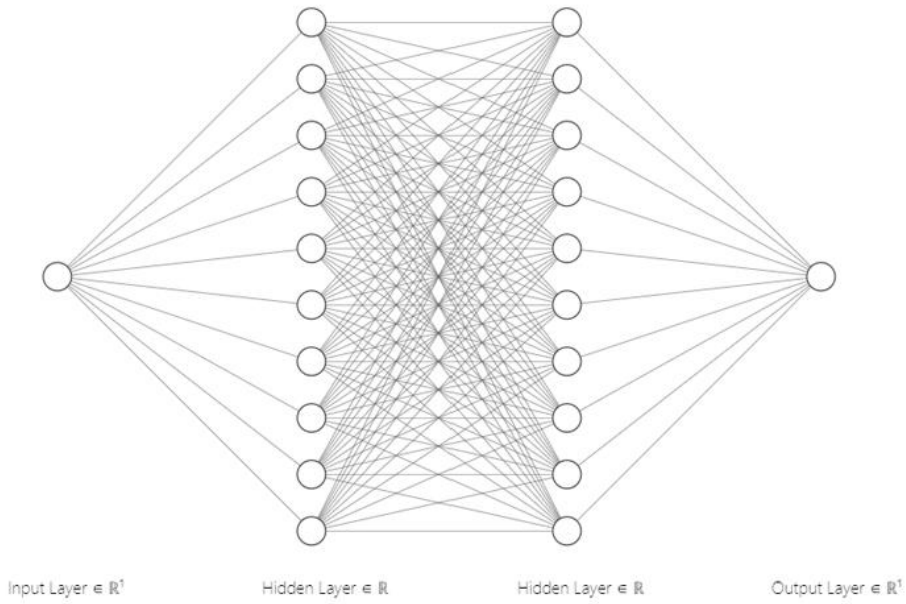
Nonlinear Method – Diamond Price Prediction from Carat Using Gradient Descent and a Neural Network with TensorFlow

First, we split the dataset into a training set and a test set. We will use the test set in the final evaluation of the model.

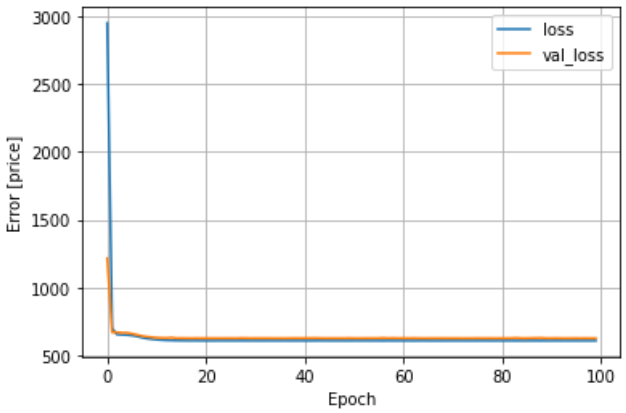
The training set is 80% of the data and the test set is 20%.

Then we applied normalization.

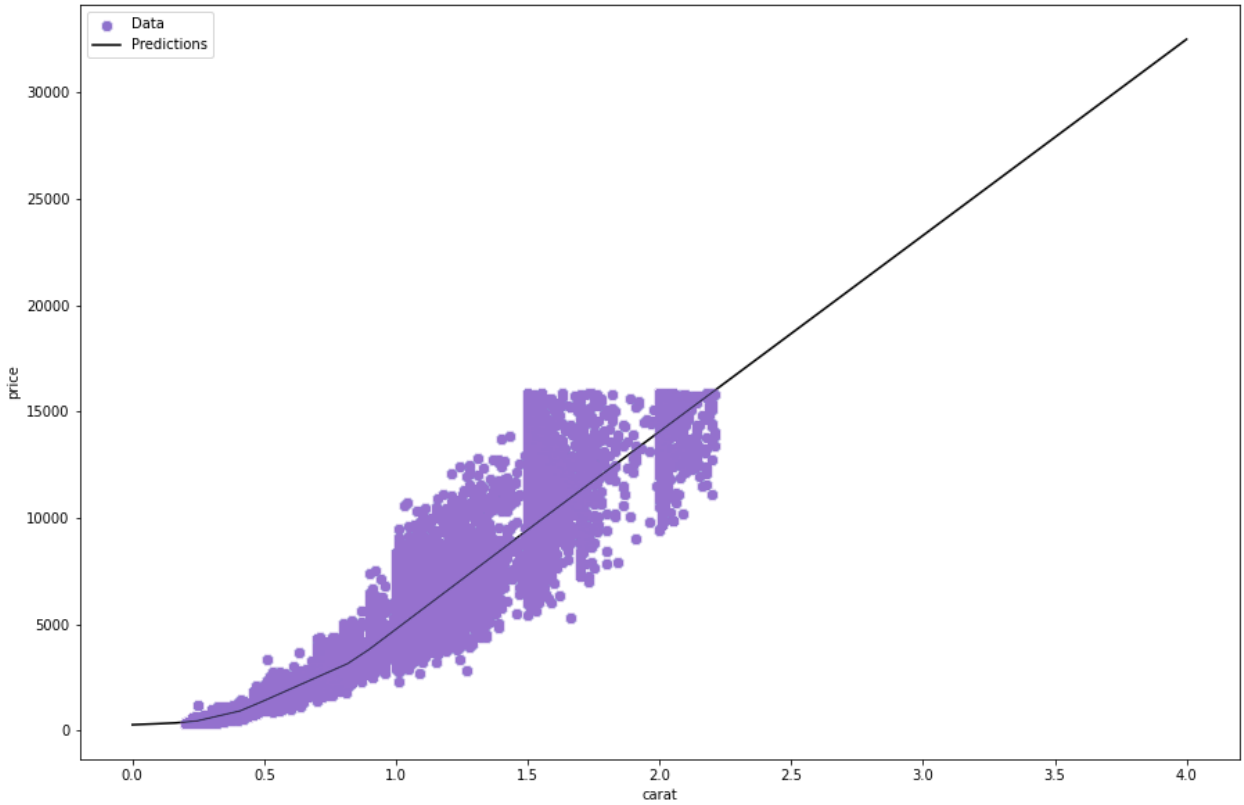
Our neural network consists of the normalization layer (one neuron) then 2 layers consisting of 64 neurons then a layer with 1 neuron (output layer).



The hidden layers here have only ten neurons for visualization purposes only



This plot shows how the error decreases with the number of iterations. This plot confirms that the model converges to find the least error value.



To rate how well our regression line performs, we calculated the mean absolute error after the last epoch by calculating the absolute error of predictions out of all our records then getting their mean.

$$\sum_{y=0}^{n} |(y - \hat{y})| = \ 605.1967\ USD$$

Multi Variate Regression – Predicting the Price

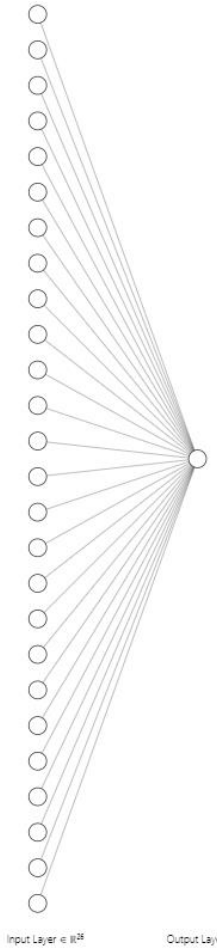
Linear Method - Diamond Price Prediction with Multiple Variables Using Gradient Descent and a Neural Network with TensorFlow

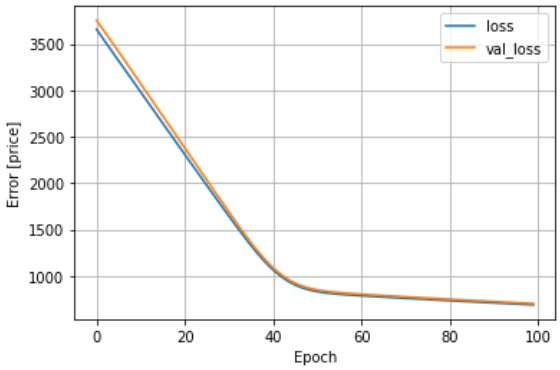
First, we split the dataset into a training set and a test set. We will use the test set in the final evaluation of the model.

The training set is 80% of the data and the test set is 20%.

Then we applied normalization.

Our neural network consists of the normalization layer (26 neurons) then a layer with 1 neuron (output layer).





This plot shows how the error decreases with the number of iterations. This plot confirms that the model converges to find the least error value.

To rate how well our regression line performs, we calculated the mean absolute error after the last epoch by calculating the absolute error of predictions out of all our records then getting their mean.

$$\sum_{y=0}^{n} |(y - \hat{y})| = \ 685.3607\ USD$$

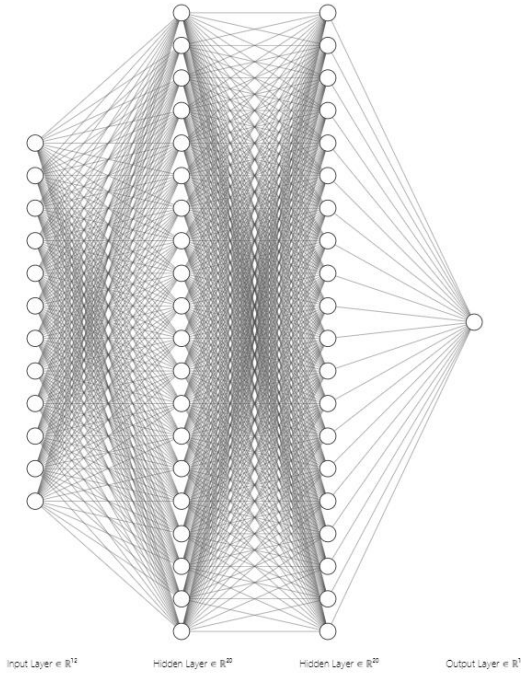
Nonlinear Method - Diamond Price Prediction with Multiple Variables Using Gradient Descent and a Neural Network with TensorFlow

First, we split the dataset into a training set and a test set. We will use the test set in the final evaluation of the model.

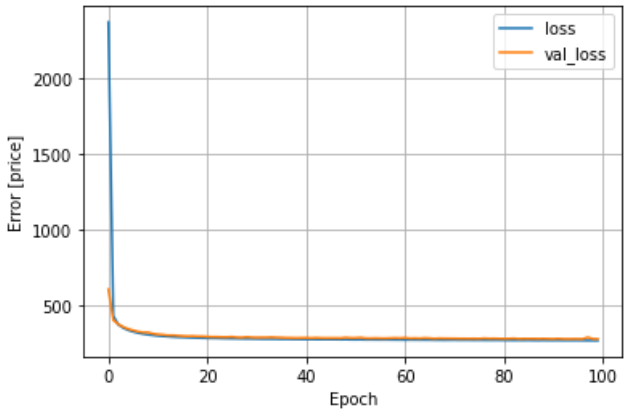
The training set is 80% of the data and the test set is 20%.

Then we applied normalization.

Our neural network consists of the normalization layer (26 neurons) then 2 hidden layers with 64 neurons then a layer with 1 neuron (output layer).



The diagram has a lower than actual number of neurons for visualization purposes



This plot shows how the error decreases with the number of iterations. This plot confirms that the model converges to find the least error value.

To rate how well our regression line performs, we calculated the mean absolute error after the last epoch by calculating the absolute error of predictions out of all our records then getting their mean.

$$\sum_{y=0}^{n} |(y - \hat{y})| = \ 261.0211\ USD$$

Here is a summary of the mean absolute error of all the regression methods we used:

Single Variate	Linear	R's Complement	885.4745 *USD*
		Neural Network	739.1205 *USD*
	Nonlinear	Neural Network	605.1967 *USD*
Multivariate	Linear	Neural Network	685.3607 *USD*
	Nonlinear	Neural Network	261.0211 *USD*

Now that we have tried single variate linear and nonlinear regression, and multi variate linear and nonlinear regression for the price feature, we now conclude that the best regression method for multi variate input, is the nonlinear method using gradient descent and a neural network. And the best method for the single variate is also the nonlinear method using gradient descent and a neural network.

Overall, nonlinear regression is giving better results.

Single Variate Regression – Price and Dimensions (x, y, z)

Predicting Dimensions (x, y, z) Using Price with Linear Regression Using the Correlation Coefficient

We will predict the dimensions of the length, width and depth of the diamonds based on the price.

We begin by getting the correlation coefficient (r) between the x – length - , y – width – and z -height- ,and the price:

$r_x = 0.8906$ **Strong Positive Relationship**

$r_y = 0.8922$ **Strong Positive Relationship**

$r_z = 0.8863$ **Strong Positive Relationship**

Then we calculate B_0

$$B_{0x} = r * \frac{\sigma_x}{\sigma_{price}} = 0.0002741$$

$$B_{0y} = r * \frac{\sigma_y}{\sigma_{price}} = 0.0002726$$

$$B_{0z} = r * \frac{\sigma_z}{\sigma_{price}} = 0.0001687$$

And B_1

$$B_{1x} = \mu_x - B_{0x} * \mu_{price} = 4.6833$$

$$B_{1y} = \mu_y - B_{0y} * \mu_{price} = 4.6910$$

$$B_{1z} = \mu_z - B_{0z} * \mu_{price} = 4.6815$$

Finally, we calculate y_0

$$y_{0x} = B_{0x}x + B_{1x} = 0.0002741x - 4.6833$$

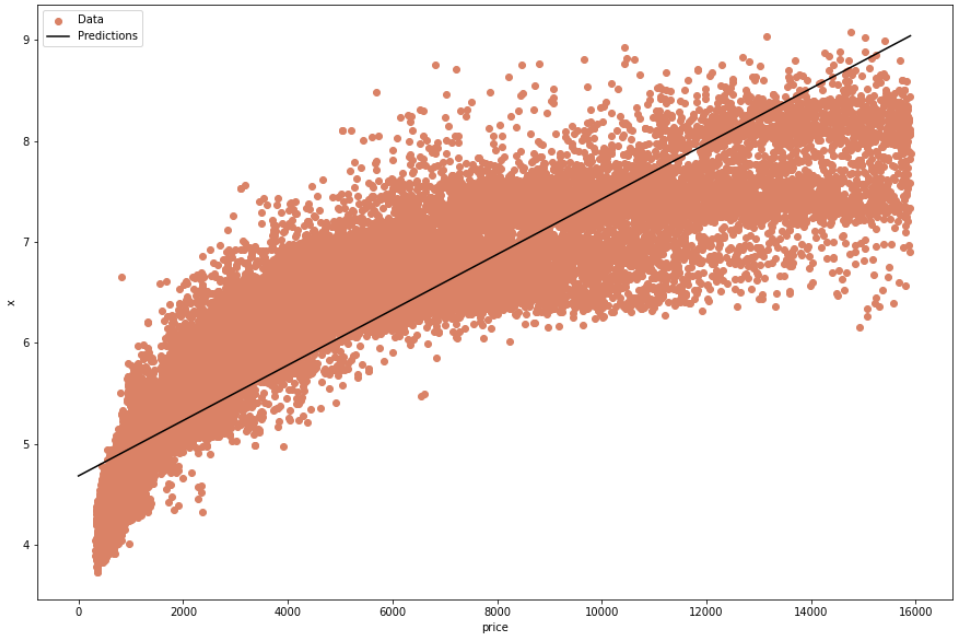
$$y_{0y} = B_{0y}x + B_{1y} = 0.0002726x - 4.6910$$

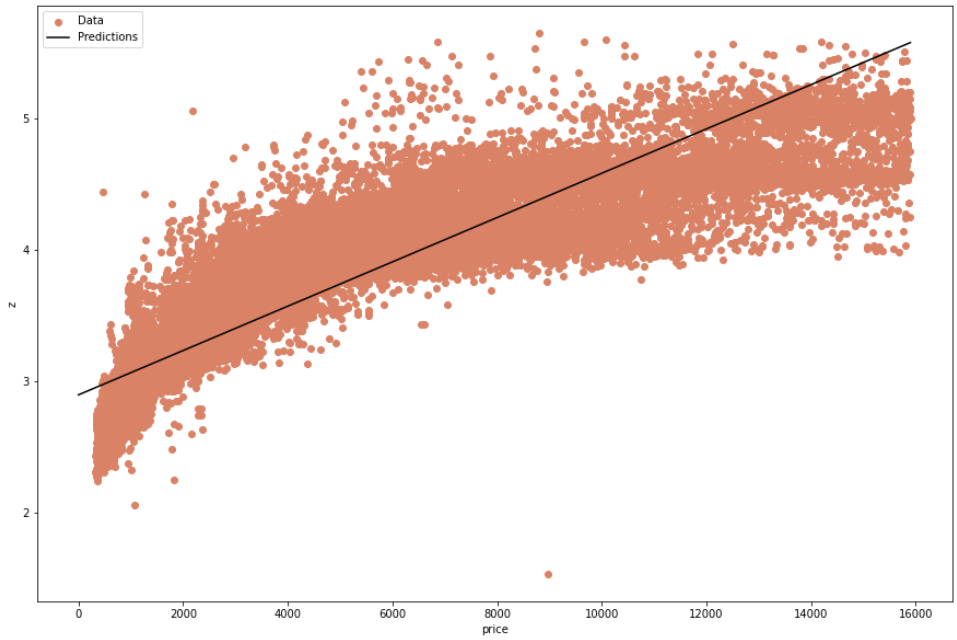
$$y_{0z} = B_{0z}x + B_{1z} = 0.0001687x - 4.6815$$

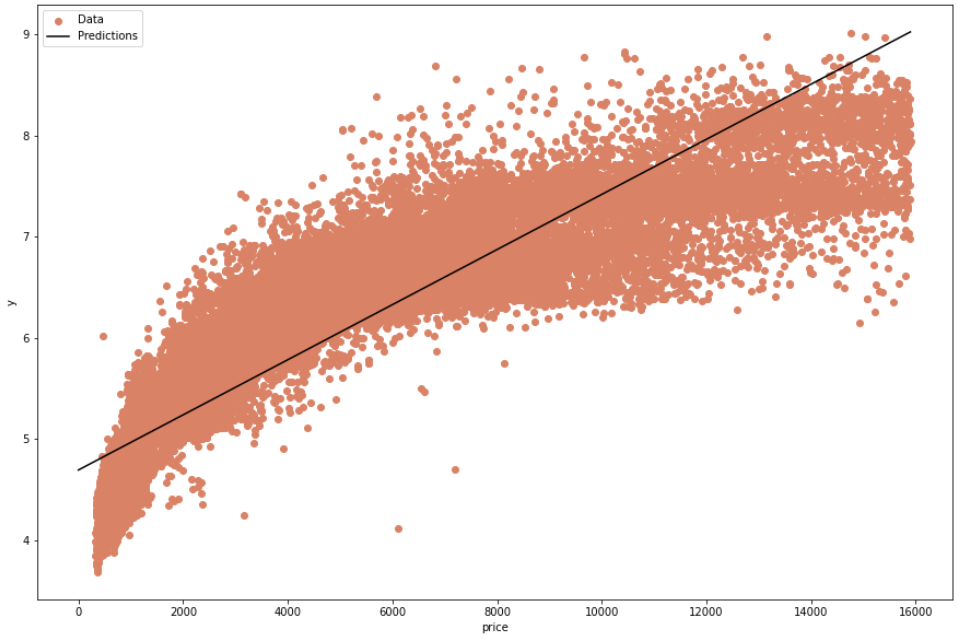
We can now predict new price values.

This confirms our alternate hypothesis for purpose 3 that the price and the length, width and height of the diamond are strong positively correlated.

Since this is a single variable regression, it's easy to view the model's predictions as a function of the input:







To rate how well our regression line performs, we calculated the absolute mean error by calculating the absolute error of predictions out of all our records then getting their mean.

For the x (length)

$$Absolute\ Mean\ Error = \sum_{y=0}^{n} |y - \hat{y}| = 0.4005\ mm$$

For the y (width)

$$Absolute\ Mean\ Error = \sum_{y=0}^{n} |y - \hat{y}| = 0.3964\ mm$$

For the z (depth)

$$Absolute\ Mean\ Error = \sum_{y=0}^{n} |y - \hat{y}| = 0.251\ mm$$

Then we created a function that calculates the estimated values for x, y and z all together when you enter a specific price.

Classification of The Clarity of The Diamonds Knowing Its Other Characteristics

First, we split the dataset into a training set and a test set. We will use the test set in the final evaluation of the model.

The training set is 80% of the data and the test set is 20%.

In the table of statistics, it's easy to see that the features have different ranges. some range in the units, some in the hundreds, some in the thousands.

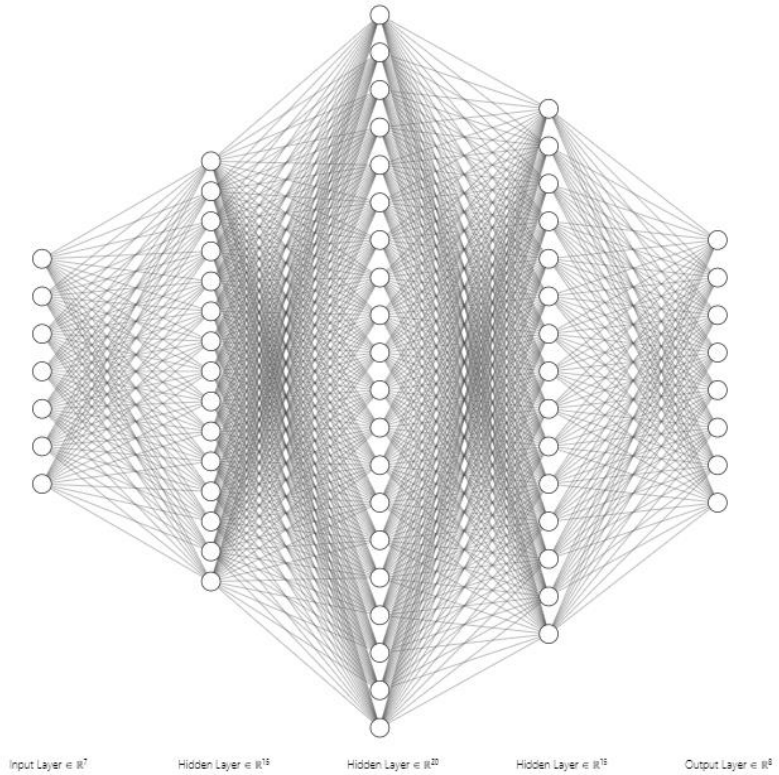
	mean	std
carat	0.731307	0.411764
depth	61.610507	1.062151
table	57.234395	1.996659
price	3498.515092	3437.424130
x	5.594549	1.045153
y	5.598036	1.037866
z	3.447391	0.641583

It is good practice to normalize features that have different ranges.

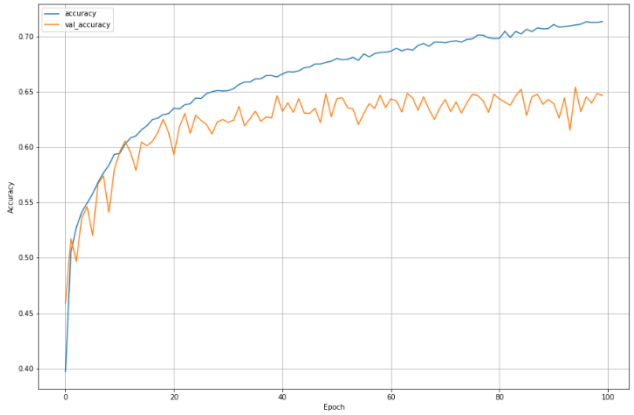
If we skip normalizing the features, the model will still converge (ie: reach the optimum loss. ie: become able to predict and perform regression as best as it can), However, normalization makes training much more stable.

We built a classification model using gradient descent in a neural network.

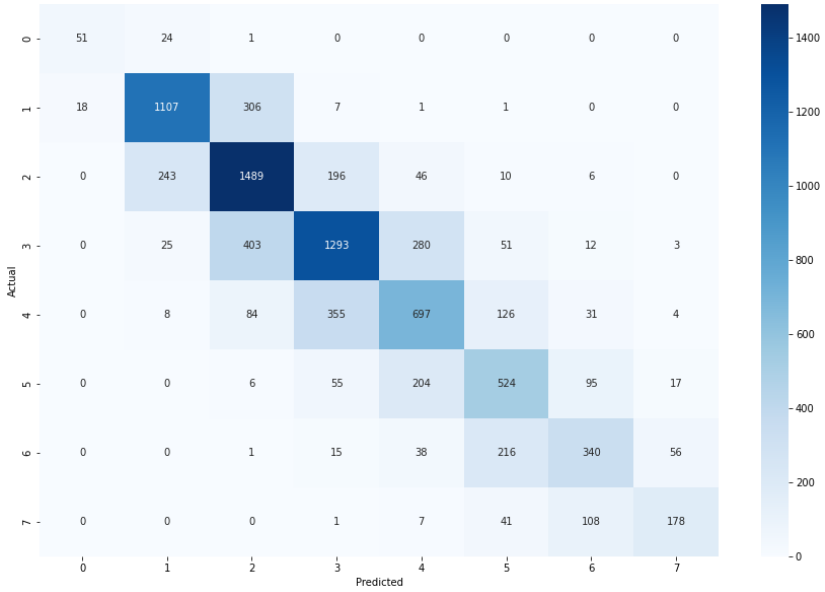
Our neural network consists of the normalization layer (19 neurons) then 3 hidden layers with (64,128,64) neurons respectively then a layer with 8 neurons (output layer).



The hidden layers in the diagram have fewer neurons for visualization purposes.



This plot shows how the accuracy increases with the number of iterations.



This confusion matrix shows that the best color we predict most accurately is 2 (SI1) and the worst color we predict is 0 (I1).

Our model can predict the clarity of diamonds with 64.68% accuracy. This accuracy is better than a random guess which is equal to 12.5%, however, it is not the best. Because there are no strong relationships with the clarity.

Conclusion

Now that we have confirmed and given evidence that supports all our alternative hypothesizes. Let's recap, firs we gave a brief description on every feature of our dataset. Then we proved that there is a strong positive relation between the carat of the diamond and the price. Whereas, if the diamond's weight in carat increases, the price increases. We also proved that there is a strong positive relationship between the price and the x – length -, y – width - and z – depth - of the diamond. So, if the price paid for the diamond increases the size of the diamond also increases.

Now our analysis procedures can be used by people to predict how much they're going to pay for a diamond with their desired features, mainly by weight. We also gave them the option to enter their price budget and get a prediction of how big or how small the diamond is going to be – the diamonds dimensions length, width and depth. We also created a classification to help people predict the diamond's clarity based on all the diamond's features.

References

(1) https://www.kaggle.com/shivam2503/diamonds
(2) https://www.westgard.com/lesson34.htm