**A Project Report**

**On**

**Predicting the prices of diamonds using regression models**

By Team-12

M.Akhil - 317126510093

L.Kalyan Pavan - 317126510091

G.Deepika - 317126510073

CH.Surya Pratap - 317126510065

Under the supervision of

**Dr. Aruna Malapati**

**(June 2020**)

**ACKNOWLEDGMENTS**

We would like to express our sincere thanks to the people who have helped us most throughout our project. We are grateful to our trainer Dr.Aruna ma’am for her support to guide us at every obstacle. We also would like to thank the Goal street team for the nonstop support for the project. We wish to thank our parents for their personal support and attention who inspired us to go our way.

**ABSTRACT**

A new company, Intelligent Diamond Reseller (IDR), wants to get into the business of reselling diamonds. They want to innovate in the business, so they will use predictive modeling to estimate how much the market will pay for diamonds. Of course, to sell diamonds inthe market, first they have to buy them from the producers; this is where predictive modeling becomes useful.

Let’s say people at IDR know ahead of time that they will be able to sell a specific diamond in the market for USD 5,000. With that information, they know how much to pay when buying this diamond. If someone tries to sell that diamond to them for USD 2,750, then that would be a very good deal; likewise, it would be a bad deal to pay USD 6,000 for such a diamond.So, as you can see, for IDR, it would be very important to be able to predict the price the market will pay for diamonds accurately. They have been able to get a dataset (this is actually real-world data) containing the prices and key characteristics of about 54,000 diamonds.

The aim of this project is to forecast the price of a diamond using 10 features by implementing multiple regression algorithms.

**CONTENTS**

Title page……………………………………………………..………….….………1

Acknowledgements………………………………………………….…….….…….2

Abstract……………………………….……………………………………..……...3

1. Introduction…..………………………….……………...……….….…..........5
2. Data Collection..……………………………….……..………………...…....,6
3. Data Representation……………………………………………………...…...8
4. Models………………………………………………………………….…....16
   1. . Multiple Linear Regression……………………………………………..16
   2. . Ridge Regression…………………………………….………….............18
   3. . Lasso Regression……………………………………………….………..20
   4. . Elastic Net Regression……………………………….…………………..22
   5. T-test between lasso and ridge models………………………...................24
5. Conclusion……………………………………………....………….………...25

References…………………………………………….…………………….……….27

1. **INTRODUCTION**

Diamond is a [solid form of the element carbon](https://en.wikipedia.org/wiki/Allotropes_of_carbon) with its atoms arranged in a [crystal structure](https://en.wikipedia.org/wiki/Crystal_structure) called [diamond cubic](https://en.wikipedia.org/wiki/Diamond_cubic). At [room temperature and pressure](https://en.wikipedia.org/wiki/Standard_conditions_for_temperature_and_pressure), another solid form of carbon known as [graphite](https://en.wikipedia.org/wiki/Graphite) is the [chemically stable](https://en.wikipedia.org/wiki/Chemical_stability) form, but diamond almost never converts to it. Diamond has the highest [hardness](https://en.wikipedia.org/wiki/Scratch_hardness) and [thermal conductivity](https://en.wikipedia.org/wiki/Thermal_conductivity) of any natural material, properties that are utilized in major industrial applications such as cutting and polishing tools. They are also the reason that [diamond anvil cells](https://en.wikipedia.org/wiki/Diamond_anvil_cell) can subject materials to pressures found deep in the Earth.

Diamond forms under high temperature and pressure conditions that exist only about 100 miles beneath the earth’s surface. Diamond’s carbon atoms are bonded in essentially the same way in all directions. Another mineral, graphite, also contains only carbon, but its formation process and crystal structure are very different. Diamonds have been used as decorative items since ancient times; some of the earliest references can be traced back to 25,000–30,000 B.C.

Facts:

* Mineral: Diamond
* Chemistry: C
* Color: Colorless
* Refractive Index: 2.42
* Birefringence: None
* Specific Gravity: 3.52 (+/-0.01)
* Mohs Hardness: 10

Currently, gem production totals nearly 30 million carats (6.0 tonnes; 6.6 short tons) of cut and polished stones annually, and over 100 million carats (20 tonnes; 22 short tons) of mined diamonds are sold for industrial use each year, as are about 100 tonnes (110 short tons) of synthesized diamond. Diamonds are such a highly traded commodity that multiple organizations have been created for grading and certifying them based on the “four Cs”, which are color, cut, clarity, and carat.

1. **DATA COLLECTION**

The dataset is available on [Kaggle](https://www.kaggle.com/nsharan/h-1b-visa/data). It contains the prices and key characteristics of nearly 54,000 diamonds. The columns in the dataset include carat, cut, color, clarity, depth, table, price, x, y, z.

The following are the steps used to solve the problem.

● Read the Input dataset.

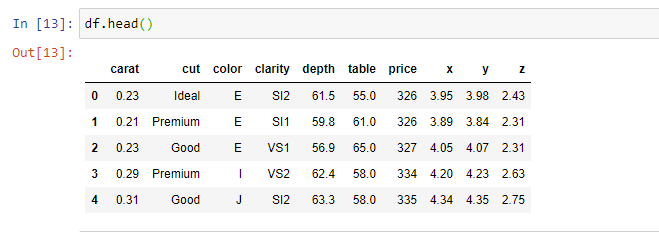
● Perform all necessary Data Normalization, Standardization processing to prepare the transformed format of given input dataset.

● Handle Missing values

● Perform Exploratory Data Analysis/Visualization and bring insights of the predictor variables.

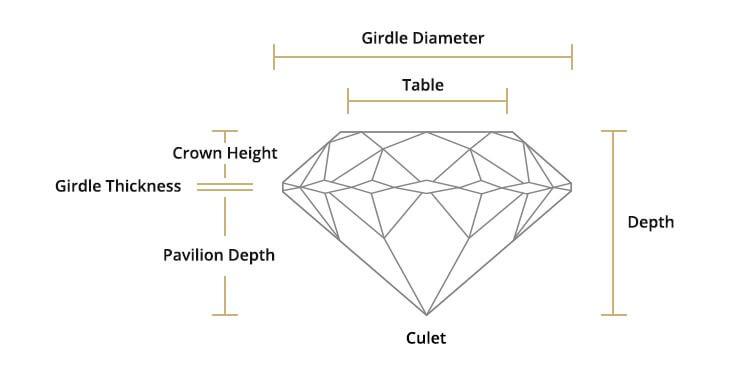
● Apply Regression models like multiple linear regression, ridge regression , lasso regression, elastic net models by splitting the data into train and test sets.

● Apply statistical tests to explain the goodness of the fit First five records of the dataset are as follows .



There are a total of 10 columns in our dataset each consisting of some specific records.The values which are stored and meaning of each column are explained below .

* CARAT : 0.2Kg - 5.01Kg
* CUT : Fair, Good, Very Good, Premium, Ideal
* COLOR : from J (Worst) to D (Best)
* CLARITY : I1 (Worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (Best)
* DEPTH : 50 to 70 %
* TABLE : 56 to 65 %

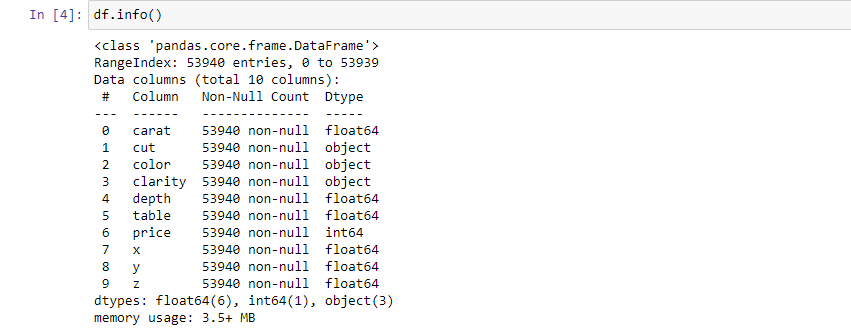


**3. DATA REPRESENTATION**

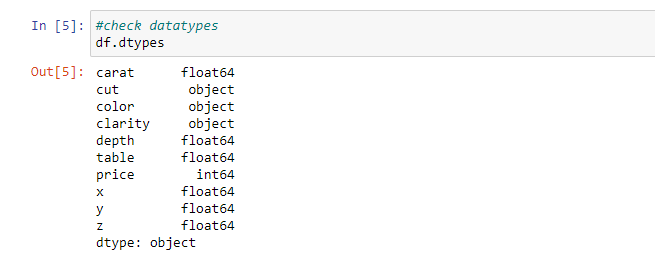
The dataset that we are using is the real world data which contains the prices and key characteristics of nearly 54,000 diamonds.

**3.1** **Dataset information:**

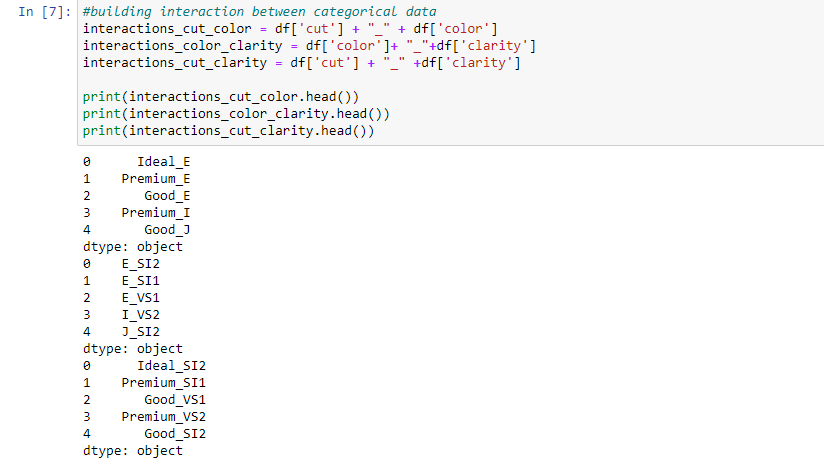
The necessary information of the dataset



The datatypes of the attributes in the dataset



There are 3 categorical features like cut, color, clarity whose data type is of object type. The interactions were built between these categorical data is as follows.



**3.2 Encoding:**

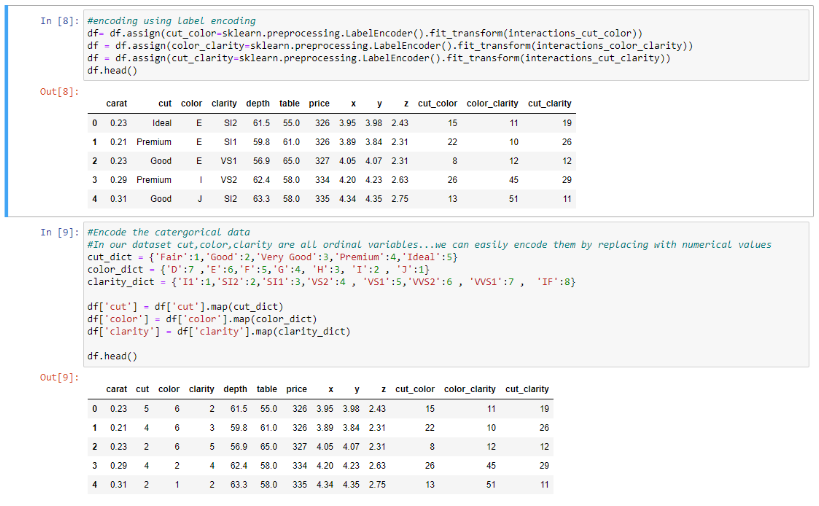
There is a need to convert non-numerical data into a numerical one and three of our features are non-numerical, namely “cut”, “color” and “clarity”. We can apply Label encoding to this problem and map those categories.

**Label Encoding**:

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.

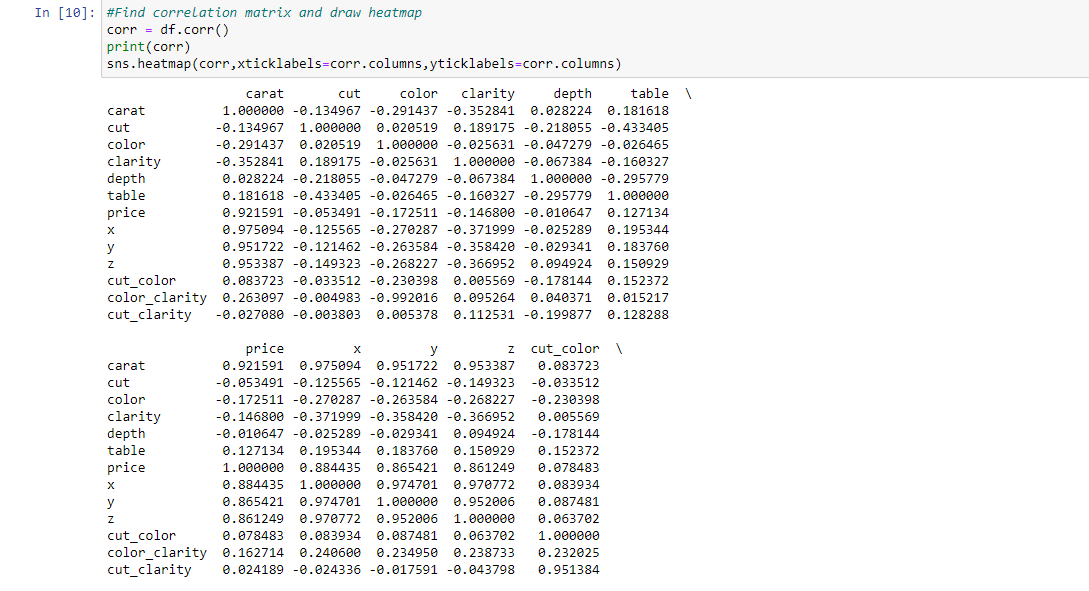
**Limitation of label Encoding:**

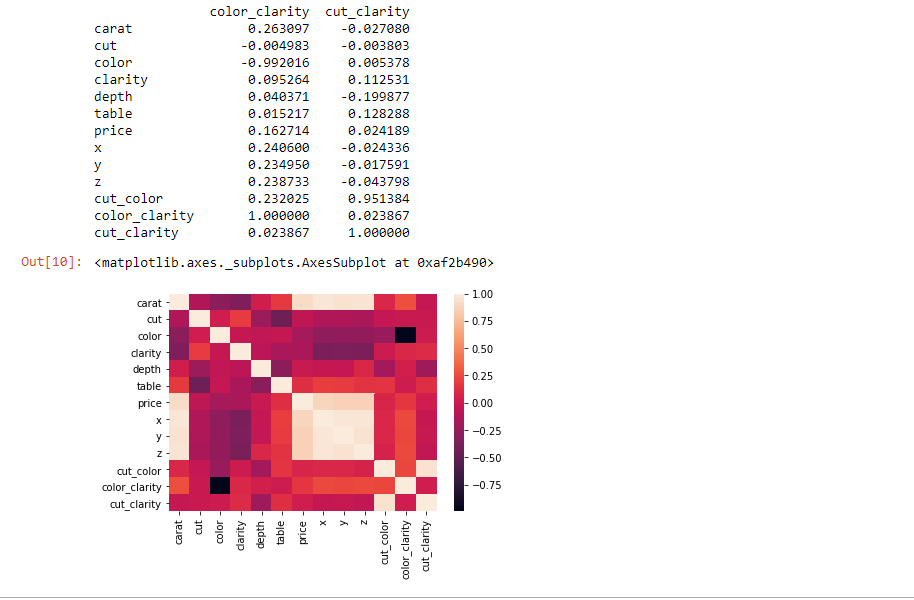
Label encoding converts the data in machine readable form, but it assigns a unique number(starting from 0) to each class of data. This may lead to the generation of priority issues in training of data sets. A label with high value may be considered to have high priority than a label having lower value.



**3.3. Feature selection using the correlation matrix:**

Let's view the [correlation matrix](https://www.displayr.com/what-is-a-correlation-matrix/) for our data. A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between the two variables. Higher the value more likely the data correlates! For this problem as we want to predict the price of a diamond, we will focus on the correlation between price vs all other columns.

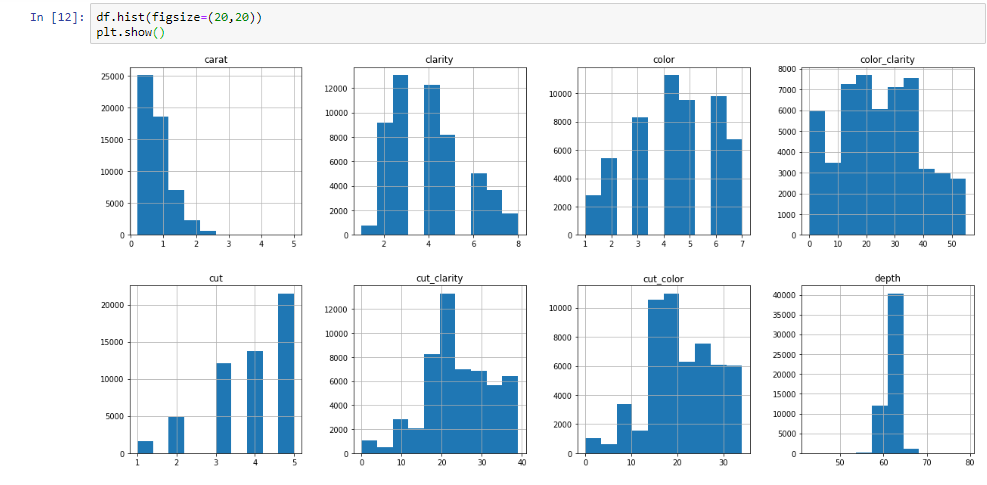




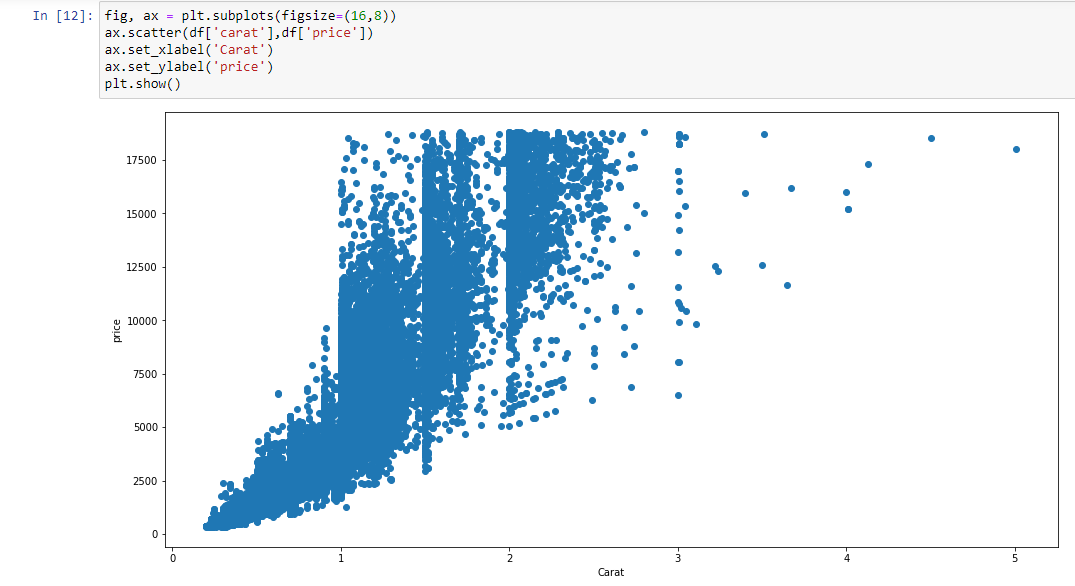
From the correlation matrix, it is observed that the features like carat, x, y, z play a crucial role in determining the price of a diamond.

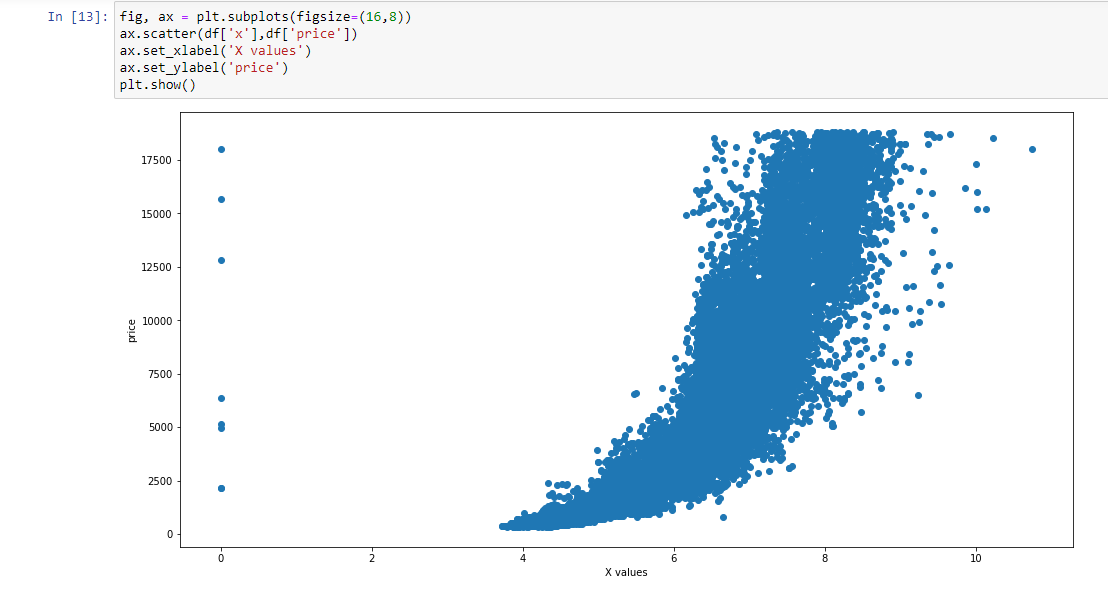
**3.4. Data Visualization:**

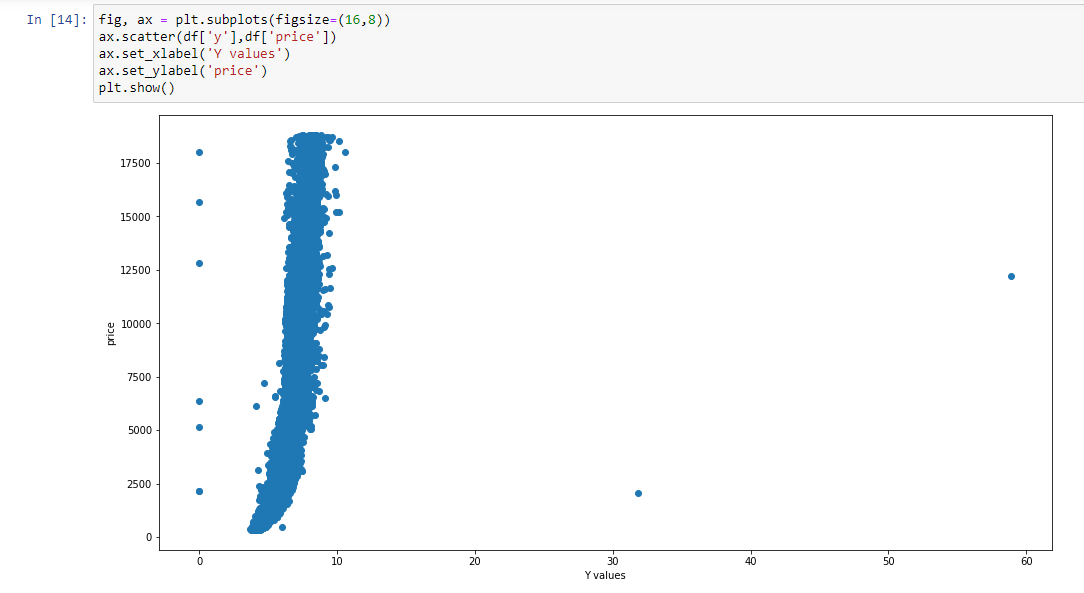
Then we performed data visualisation between the columns of the dataset and we plotted a histogram as follows.

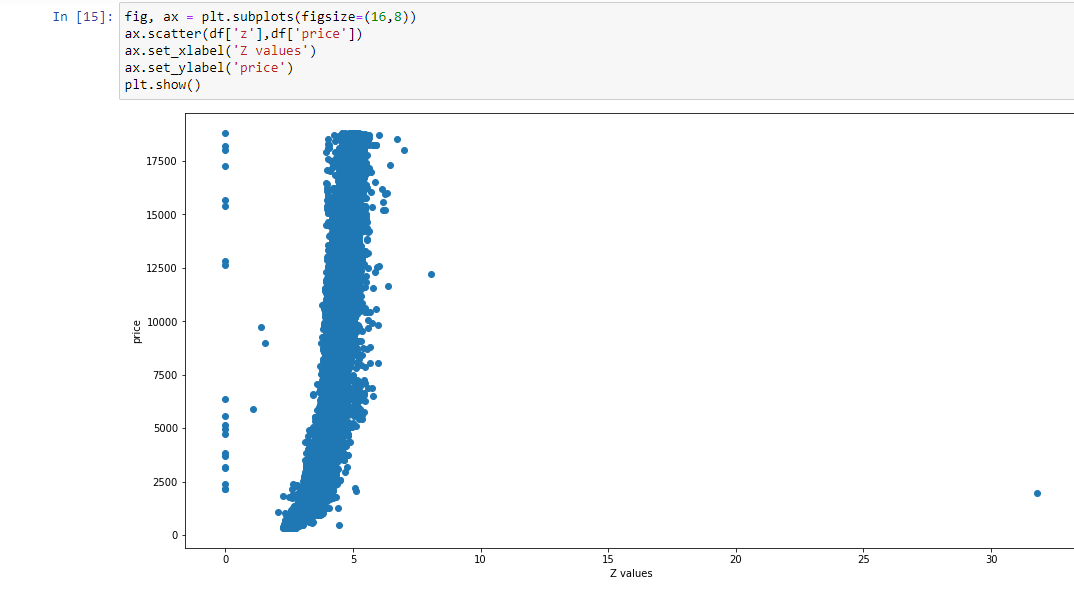
****

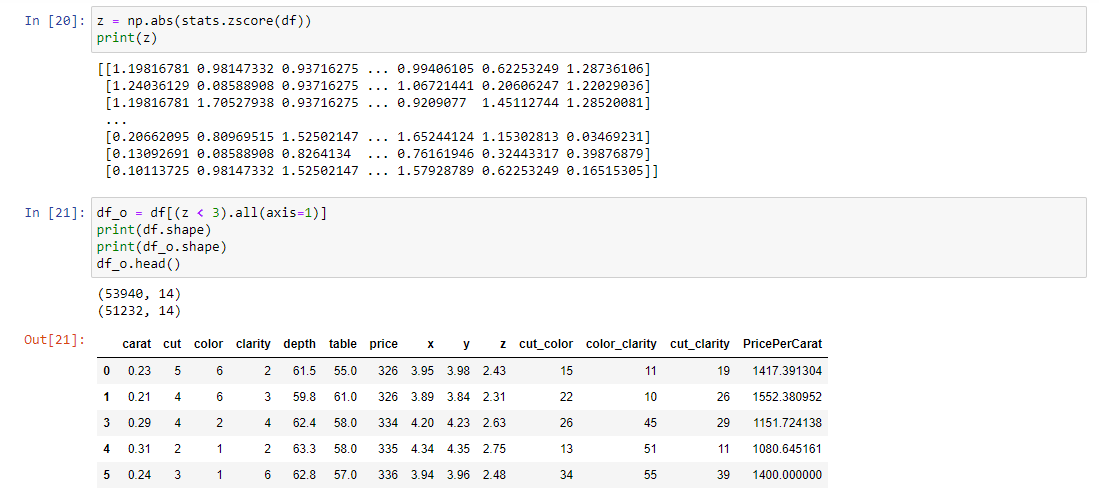
Then the scatter plots were drawn for the price and the highly correlated categoriical features like carat, x, y, z and outlayers were removed.

****

****

****

****

****

**4.MODELS**

* Multiple Linear Regression
* Ridge Regression
* Lasso Regression
* Elastic Net Regression

**Description:**

**4.1 Multiple Linear Regression:**

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) that involves more than one explanatory variable.

## The Formula for Multiple Linear Regression is:

*yi*=*β*0+*β*1*xi*1+*β*2*xi*2​+...+ *βpxip* +*ϵ*

where, for *i*=*n* observations:

*yi*=dependent variable

*xi=explanatory* variables

*β*0=y-intercept (constant term)

*βp*=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals)

## 

## Explaining Multiple Linear Regression:

A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

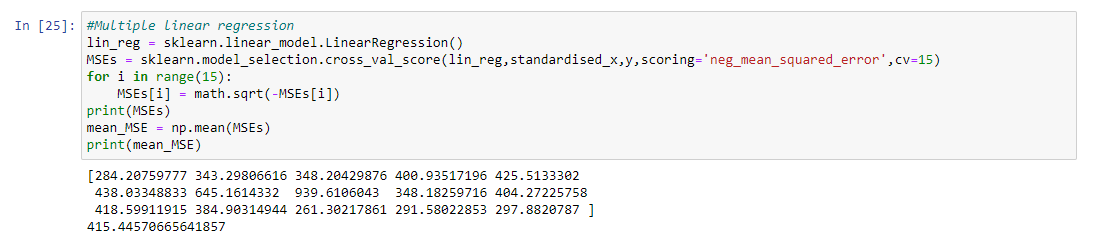
**The multiple regression model is based on the following assumptions:**

* There is a linear relationship between the dependent variables and the independent variables.
* The independent variables are not too highly correlated with each other.
* yi observations are selected independently and randomly from the population.
* Residuals should be normally distributed with a mean of 0 and variance *σ.*

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R2 by itself can't this be used to identify which predictors should be included in a model and which should be excluded. R2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.



**4.2 Ridge Regression:**

Ridge regression is a way to create a [parsimonious model](https://www.statisticshowto.com/parsimonious-model/) when the number of [predictor variables](https://www.statisticshowto.com/independent-variable-definition/#Predictor) in a set exceeds the number of observations, or when a data set has [multicollinearity](https://www.statisticshowto.com/multicollinearity/) (correlations between predictor variables).

## Ridge Regression vs. Least Squares:

Least squares regression isn’t defined at all when the number of predictors exceeds the number of observations; It doesn’t differentiate “important” from “less-important” predictors in a model, so it includes all of them. This leads to [overfitting](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/#overfitting) a model and failure to find unique solutions. Least squares also have issues dealing with multicollinearity in data. Ridge regression avoids all of these problems. It works in part because it doesn’t require [unbiased estimators](https://www.statisticshowto.com/unbiased/#UE); While least squares produces unbiased estimates, [variances](https://www.statisticshowto.com/probability-and-statistics/variance/) can be so large that they may be wholly inaccurate. Ridge regression adds just enough [bias](https://www.statisticshowto.com/what-is-bias/) to make the estimates reasonably [reliable](https://www.statisticshowto.com/reliability-validity-definitions-examples/) approximations to true population values.

## Shrinkage:

Ridge regression uses a type of [shrinkage estimator](https://www.statisticshowto.com/shrinkage-estimator/) called a *ridge estimator*. Shrinkage estimators theoretically produce new [estimators](https://www.statisticshowto.com/estimator/) that are shrunk closer to the “true” population parameters. The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

## 

## 

## Regularization:

Ridge regression belongs to a class of regression tools that use [L2 regularization](https://www.statisticshowto.com/regularization/). The other type of regularization, L1 regularization, limits the size of the coefficients by adding an *L1 penalty* equal to the [absolute value](https://www.statisticshowto.com/integer/#abs) of the magnitude of coefficients. This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will *not* result in sparse models.

A [tuning parameter](https://www.statisticshowto.com/tuning-parameter/) (λ) controls the strength of the penalty term. When λ = 0, ridge regression equals least squares regression. If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

## On Mathematics:

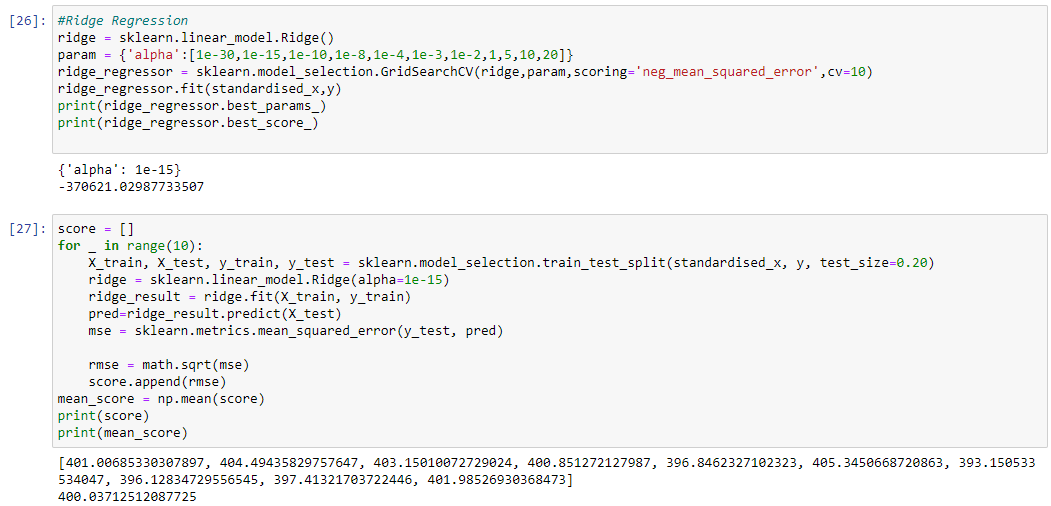
OLS regression uses the following formula to estimate coefficients:

ridge regression

If X is a centered and scaled matrix, the cross product matrix (X`X) is nearly [singular](https://www.statisticshowto.com/matrices-and-matrix-algebra/#SingularM) when the X-columns are highly correlated. Ridge regression adds a *ridge parameter* (k), of the [identity matrix](https://www.statisticshowto.com/matrices-and-matrix-algebra/#IdentityM) to the cross product matrix, forming a new matrix (X`X + kI). It’s called *ridge* regression because the diagonal of ones in the correlation matrix can be described as a ridge. The new formula is used to find the coefficients:



Choosing a value for *k* is not a simple task, which is perhaps one major reason why ridge regression isn’t used as much as least squares or [logistic regression](https://www.statisticshowto.com/logistic-regression/). You can read one way to find *k* in Dorugade and D. N. Kashid’s paper [Alternative Method for Choosing Ridge Parameter for Regression.](https://www.statisticshowto.com/wp-content/uploads/2017/07/dorugadeAMS9-12-2010.pdf).



**4.3 Lasso Regression:**

Lasso regression is a type of linear regression that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/mean/). The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of [multicollinearity](https://www.statisticshowto.com/multicollinearity/) or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

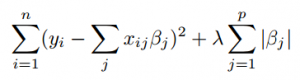
The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator.

## L1 Regularization:

Lasso regression performs L1 [regularization](https://www.statisticshowto.com/regularization/), which adds a penalty equal to the [absolute value](https://www.statisticshowto.com/integer/#abs) of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. [Ridge regression](https://www.statisticshowto.com/ridge-regression/)) *doesn’t* result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

## Performing the Regression:

Lasso solutions are quadratic programming problems, which are best solved with software (like [Matlab](https://www.mathworks.com/help/stats/lasso.html)). The goal of the algorithm is to minimize:

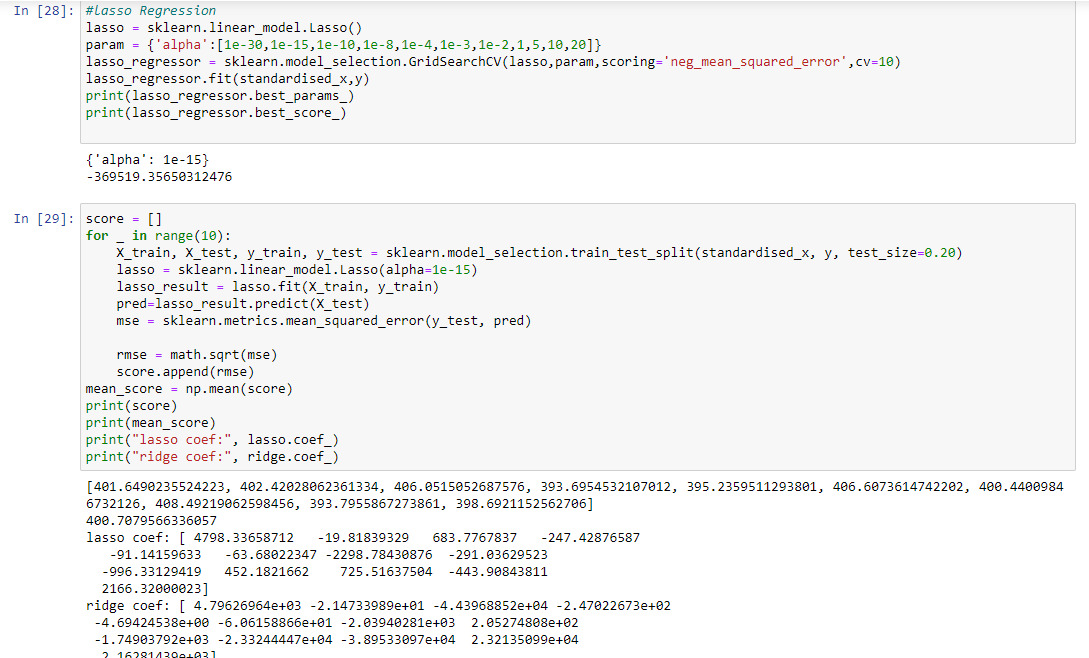


Which is the same as minimizing the [sum of squares](https://www.statisticshowto.com/residual-sum-squares/) with constraint Σ |Bj≤ s. Some of the βs are shrunk to exactly zero, resulting in a regression model that’s easier to interpret.

A [tuning parameter](https://www.statisticshowto.com/tuning-parameter/), λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

* When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
* As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, *all* coefficients are eliminated).
* As λ increases, [bias](https://www.statisticshowto.com/what-is-bias/) increases.
* As λ decreases, [variance](https://www.statisticshowto.com/probability-and-statistics/variance/) increases.

If an intercept is included in the model, it is usually left unchanged.



**4.4 Elastic Net Regression:**

Elastic net is basically a combination of both L1 and L2 regularization. So if you know elastic net, you can implement both Ridge and Lasso by tuning the parameters. So it uses both L1 and L2 penalty term, therefore its equation look like as follows:



So how do we adjust the lambdas in order to control the L1 and L2 penalty term? Let us understand by an example. You are trying to catch a fish from a pond. And you only have a net, then what would you do? Will you randomly throw your net? No, you will actually wait until you see one fish swimming around, then you would throw the net in that direction to basically collect the entire group of fishes. Therefore even if they are correlated, we still want to look at their entire group.

Elastic regression works in a similar way. Let's say, we have a bunch of correlated independent variables in a dataset, then the elastic net will simply form a group consisting of these correlated variables. Now if any one of the variable of this group is a strong predictor (meaning having a strong relationship with dependent variable), then we will include the entire group in the model building, because omitting other variables (like what we did in lasso) might result in losing some information in terms of interpretation ability, leading to a poor model performance.

So, if you look at the code below, we need to define alpha and l1\_ratio while defining the model. Alpha and l1\_ratio are the parameters which you can set accordingly if you wish to control the L1 and L2 penalty separately. Actually, we have

Alpha = a + b and l1\_ratio = a / (a+b)

where a and b weights are assigned to L1 and L2 terms respectively. So when we change the values of alpha and l1\_ratio, a and b are set accordingly such that they control trade off between L1 and L2 as:

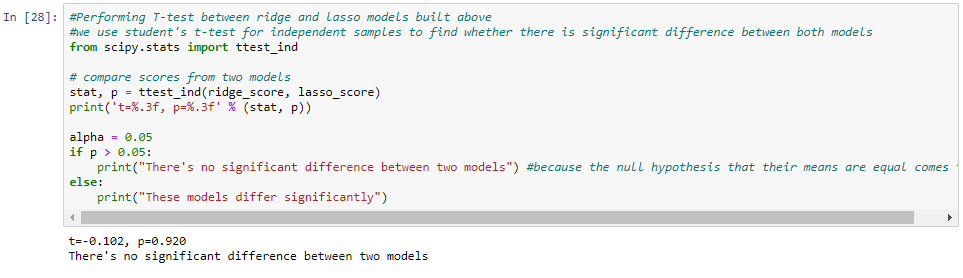
a \* (L1 term) + b\* (L2 term)

Let alpha (or a+b) = 1, and now consider the following cases:

* If l1\_ratio =1, therefore if we look at the formula of l1\_ratio, we can see that l1\_ratio can only be equal to 1 if a=1, which implies b=0. Therefore, it will be a lasso penalty.
* Similarly if l1\_ratio = 0, implies a=0. Then the penalty will be a ridge penalty.
* For the l1\_ratio between 0 and 1, the penalty is the combination of ridge and lasso.



**4.5 T-test between lasso and ridge models:**

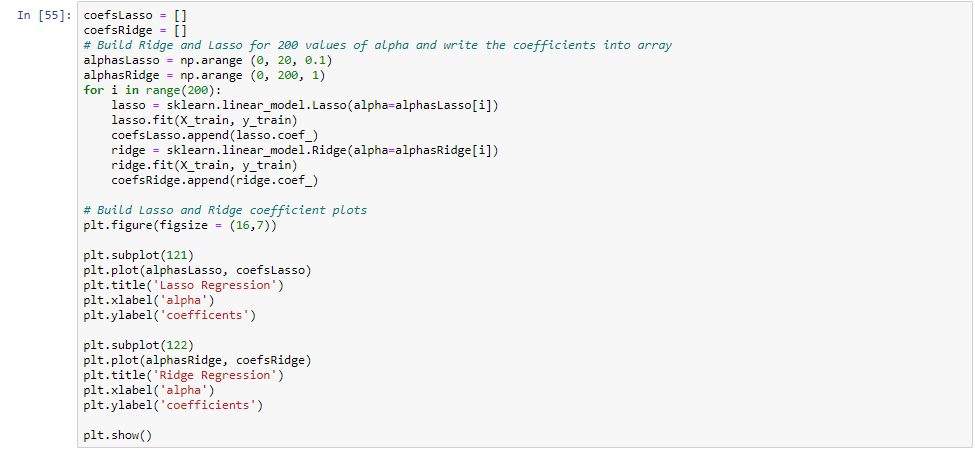


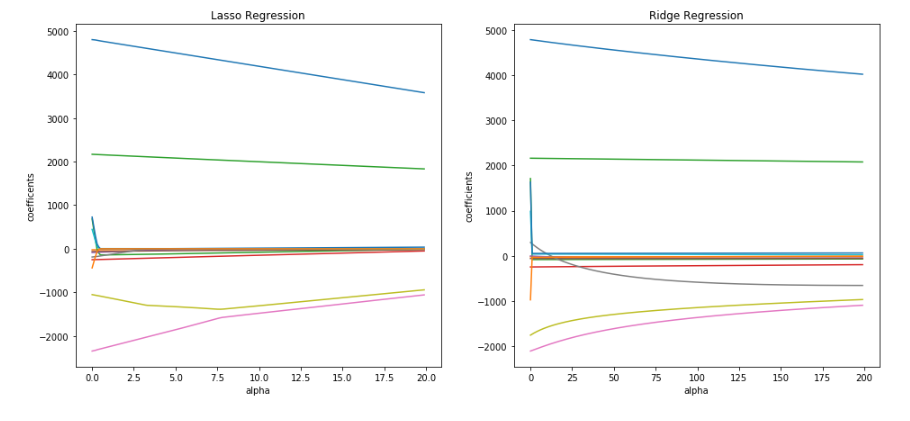
**5. CONCLUSION**

**Result visualization:**

The aim of this project was to test different data representation methods and train various models using this data. We collected our real data from sources like diamond price prediction dataset was acquired from Kaggle.The resulting transformed data was trained on the models mentioned above.

After performing the t-test between ridge and lasso regression models, it is observed that there is no significant difference between the two models. Now we are going to visualise the results of both ridge and lasso models for different alpha values.

****

****

From the graph, we can see that when we raise the alpha in Ridge regression, the coefficient's value decreases, but never becomes zero. It is the same in Lasso, that influences very less on the large coefficient's value, but the small values Lasso may reduce to zeroes. Therefore Lasso can also be used to determine which features are important to us and keeps the features that may influence the target element, while Ridge regression gives uniform penalties to all the features and in such way reduces the model complexity and prevents multi col-linearity.

**REFERENCES**

1. [*https://machinelearningmastery.com/evaluate-performance-machine-learning-algorithms-python-using-resampling/*](https://machinelearningmastery.com/evaluate-performance-machine-learning-algorithms-python-using-resampling/)
2. [*https://towardsdatascience.com/how-to-perform-lasso-and-ridge-regression-in-python-3b3b75541ad8*](https://towardsdatascience.com/how-to-perform-lasso-and-ridge-regression-in-python-3b3b75541ad8)
3. [*https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba*](https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba)
4. [*https://medium.com/@contactsunny/linear-regression-in-python-using-scikit-learn-f0f7b125a204*](https://medium.com/@contactsunny/linear-regression-in-python-using-scikit-learn-f0f7b125a204)
5. [*https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833*](https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833)
6. [*https://machinelearningmastery.com/k-fold-cross-validation/*](https://machinelearningmastery.com/k-fold-cross-validation/)