LOVELY PROFESSIONAL UNIVERSITY

upGrad Campus

INT 353 – Report

Name: Aryan Agnihotri

Reg. No.: 12015816

Roll No.: RK20CHB48

Course: INT353 - EDA Project

What is EDA?

Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Chosen dataset:

DIAMOND PRICES AND QUALITY

About the dataset:

There are 53,940 diamonds in the dataset with 10 features (carat, cut, color, clarity, depth, table, price, x, y, and z). Most variables are numeric in nature, but the variables cut, color, and clarity are ordered factor variables with the following levels. About the currency for the price column: it is Price ($) And About the columns x,y, and z they are diamond measurements as ((x: length in mm, y: width in mm,z: depth in mm))

The dataset is downloaded from Kaggle and is used to perform Exploratory Data Analysis. The original source for this dataset was kept confidential because of the sensitive nature of the data.

Attribute Information (Column Info):

1) Price : 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)

Why did I choose this dataset?

Diamonds are one of the most beautiful thing in the world. It has its own shine , cut and quality in it. I chose this dataset because currently I am into this business and also know the process how the diamond is made , how a stone is converted into a elegant and its most beautiful form by cutting it into 56 different cuts.

I personally know how the diamond is designed and how it is classified into different cuts and qualities of it with some natural and eye pleasing colors.

I hope analysing this dataset and publishing it in public domain later on will help people to know how the diamond is classified into different categories with respect its qualities and they would what it takes to shine out from a stone to a diamond.

Libraries to be used in analysis:

Here I have used following Python libraries to analyse this dataset:

1. NumPy
2. Pandas
3. Matplotlib
4. Seaborn

Technologies (software) used in the project:

I have used several technologies to complete the EDA on this dataset. The following technologies (software) were used:

- Jupyter Notebook
- PyCharm (to use Jupyter notebook server as plugin)
- Tableau (for better data visualisation)

Goals and Plans:

I plan to use univariate and multivariate data analysis techniques to find relations between the various fields that may otherwise look insignificant in the dataset.

I would use the following steps to complete my EDA on the chosen dataset:

1. importing dataset and the required libraries.
2. Interpretation of different columns.
3. Performing various techniques for missing value imputation
4. Standardising the data.
5. Finding outliers and removing duplicates and irrelevant data from the dataset
6. Performing univariate analysis
7. Performing bivariate analysis and multivariate analysis
8. Constructing hypothesis around the business problem and using the concepts of probability and statistics to work around the constructed hypothesis.

Information of dataset

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53940 entries, 1 to 53940
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object
 2   color    53940 non-null  object
 3   clarity  53940 non-null  object
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.5+ MB
```

Glimpse Of Dataset

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.230000	Ideal	E	SI2	61.500000	55.000000	326	3.950000	3.980000	2.430000
2	0.210000	Premium	E	SI1	59.800000	61.000000	326	3.890000	3.840000	2.310000
3	0.230000	Good	E	VS1	56.900000	65.000000	327	4.050000	4.070000	2.310000
4	0.290000	Premium	I	VS2	62.400000	58.000000	334	4.200000	4.230000	2.630000
5	0.310000	Good	J	SI2	63.300000	58.000000	335	4.340000	4.350000	2.750000

Summary of Dataset

Summary Of The Dataset :

Out[91]:

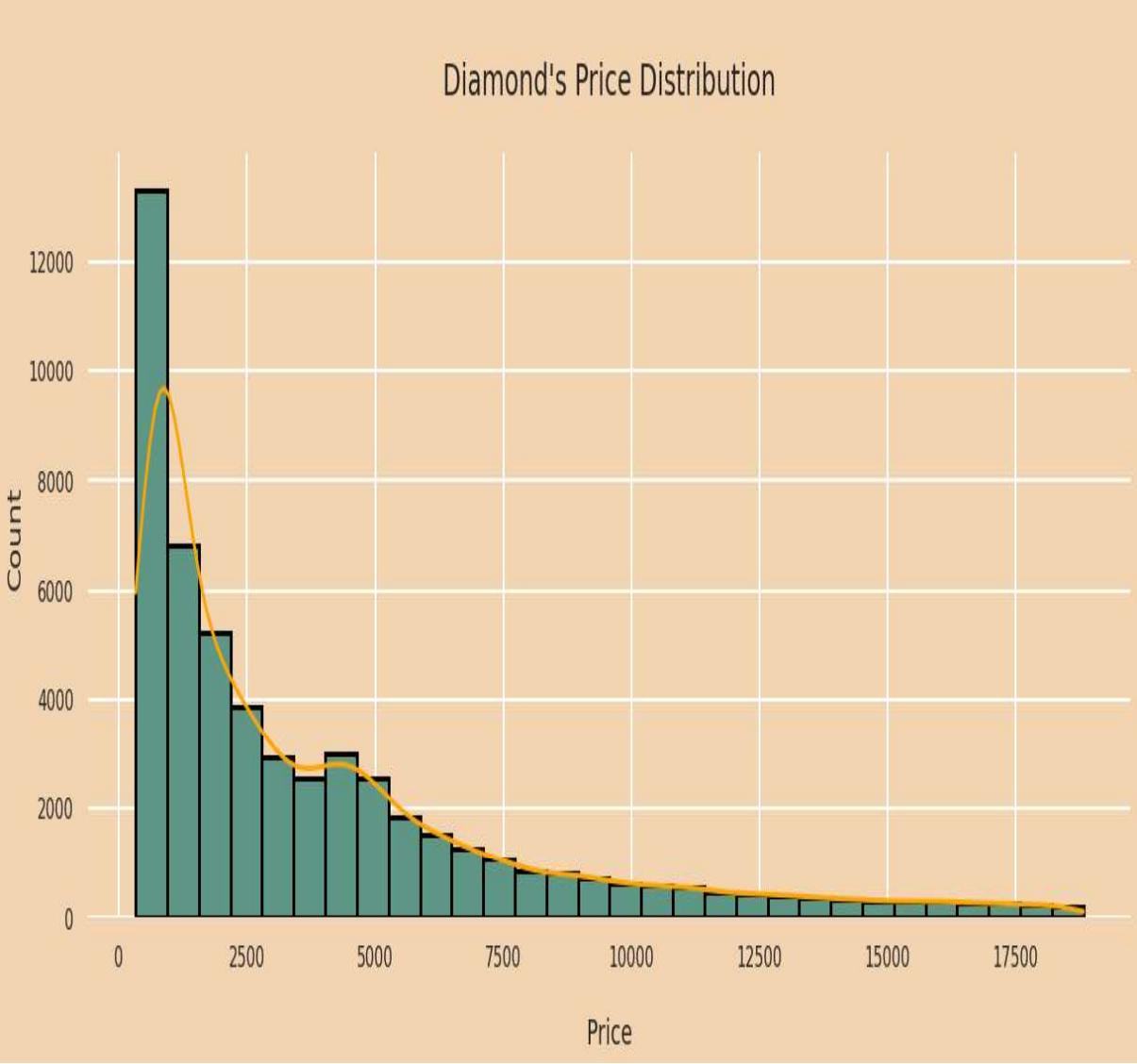
	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.00000
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

About the dataset:

There are 53,940 diamonds in the dataset with 10 features (carat, cut, color, clarity, depth, table, price, x, y, and z). Most variables are numeric in nature, but the variables cut, color, and clarity are ordered factor variables with the following levels. About the currency for the price column: it is Price ($) And About the columns x,y, and z they are diamond measurements as ((x: length in mm, y: width in mm,z: depth in mm))

The dataset is downloaded from Kaggle and is used to perform Exploratory Data Analysis. The original source for this dataset was kept confidential because of the sensitive nature of the data.

Distribution of Prices



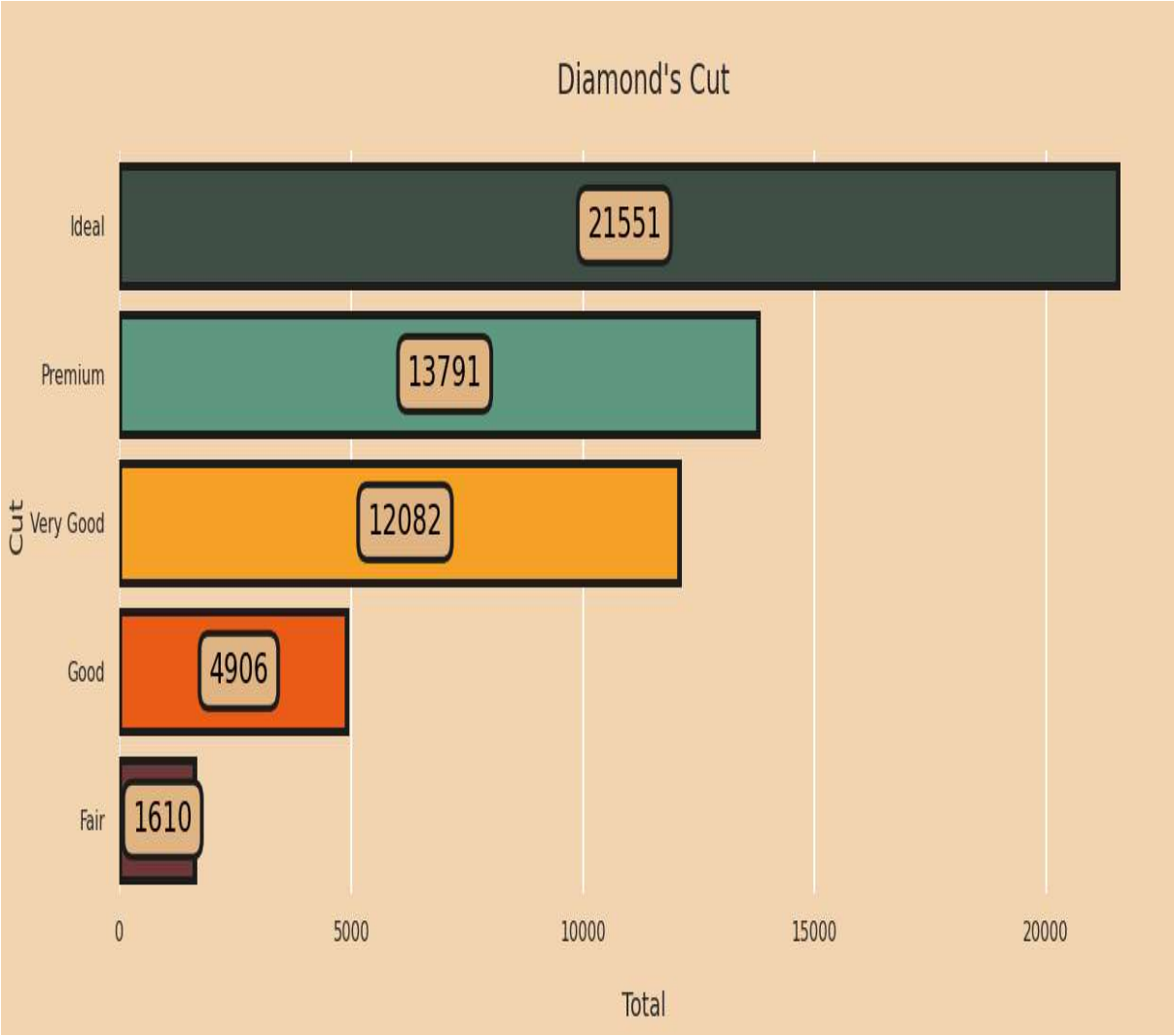
Insights:

- We can see the price distribution is right skewed. Most of the prices fall in between **326** to **2000**.
- Also there is a lot of high and medium prices which belongs to the rearest and nearly rear diamonds.

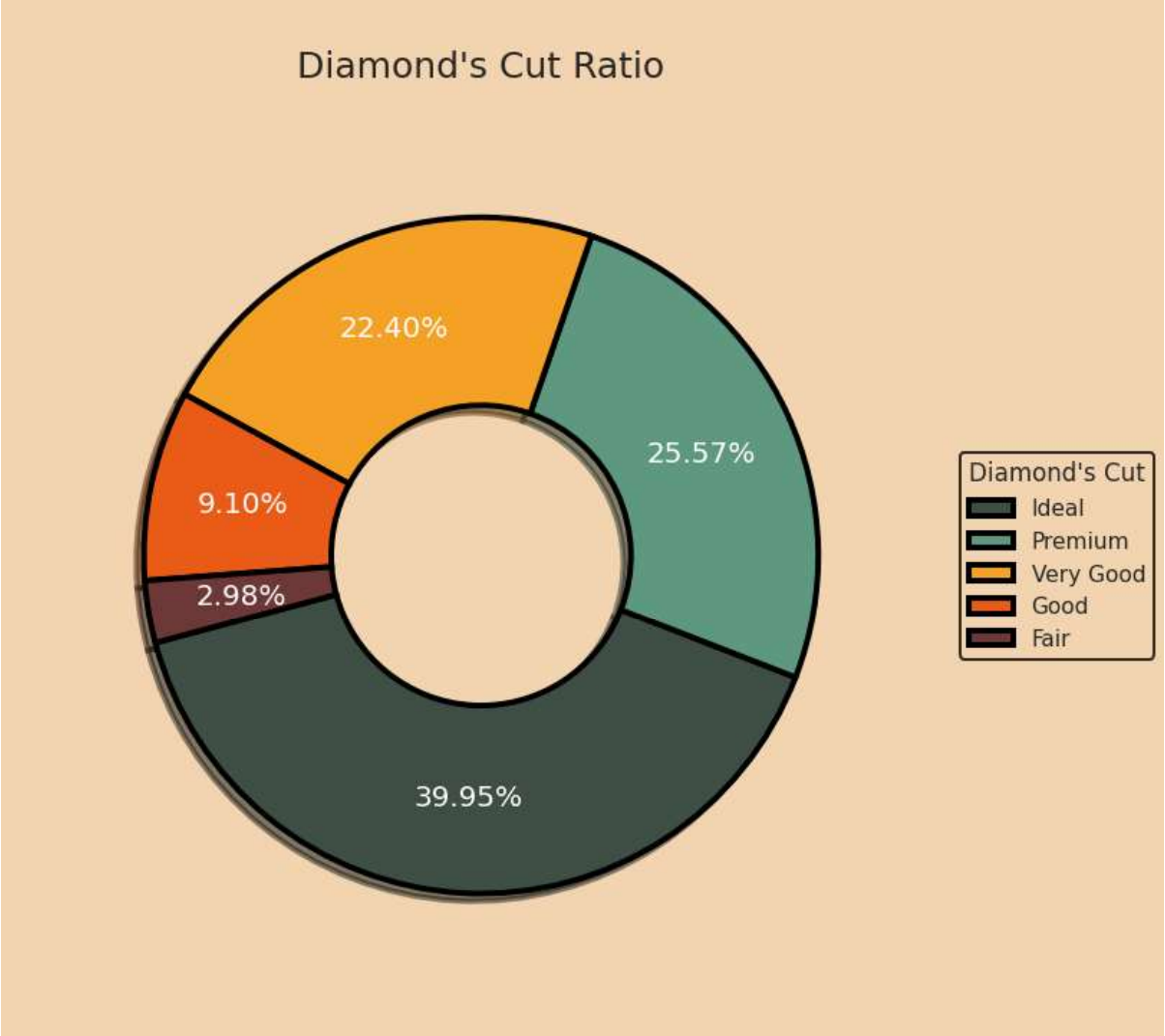
Univariate Analysis

There are 53,940 diamonds in the dataset with 10 features (carat, cut, color, clarity, depth, table, price, x, y, and z).So the analysis is done in the way that each dataset is figured with their internal proportion through various visualizations.

- ## Diamond’s Cut



- Diamond’s Cut Ratio

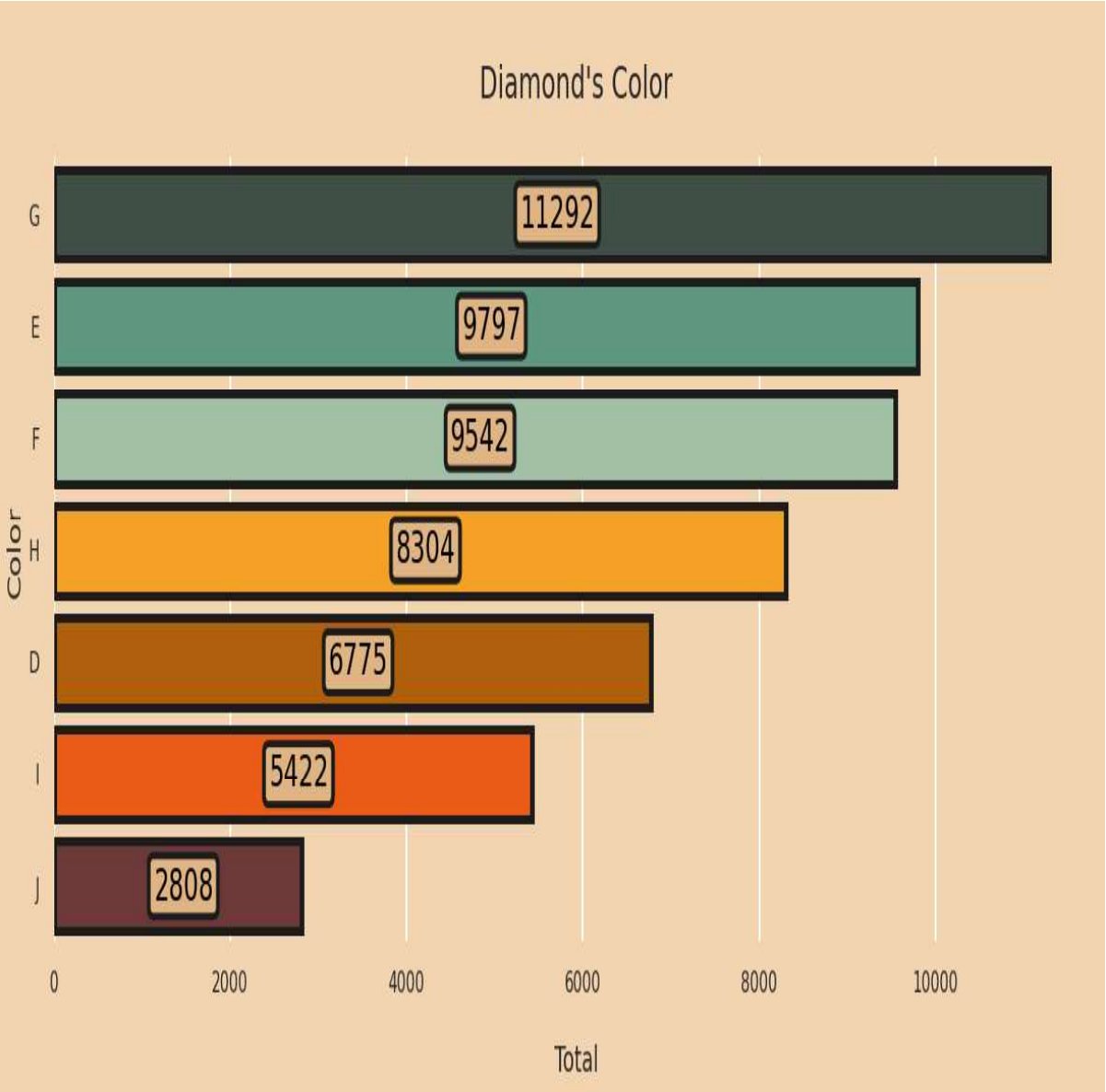


Insights:

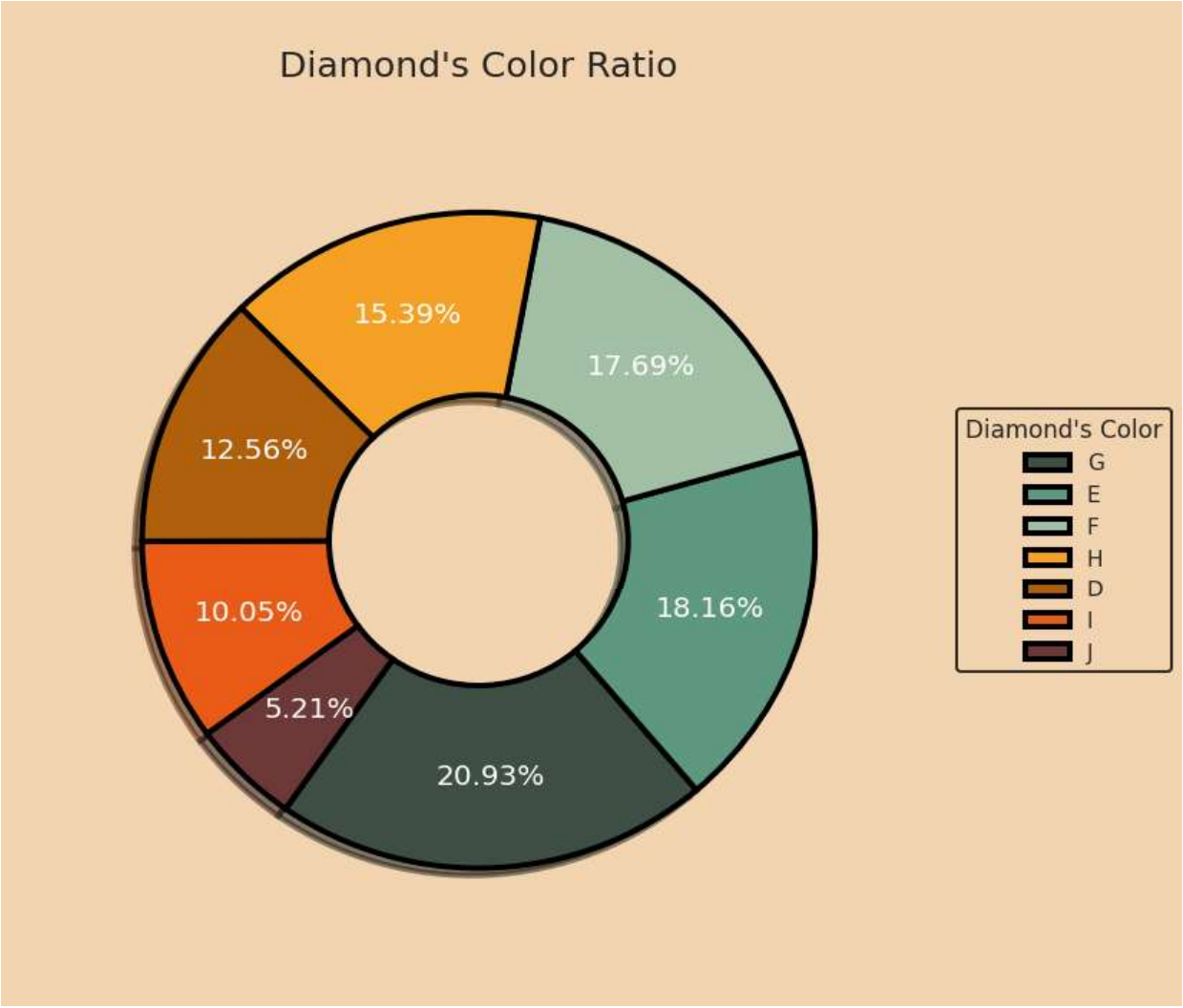
- Most of the diamonds have **Ideal Cut** with a ratio of **39.95%** followed by **Premium Cut** and **Very Good Cut**
- Only few have **Fair Cut** with a ratio of **2.98%**

- Diamond’s Color

Let's have a look on the diamond's color :



- Diamond’s Color Ratio

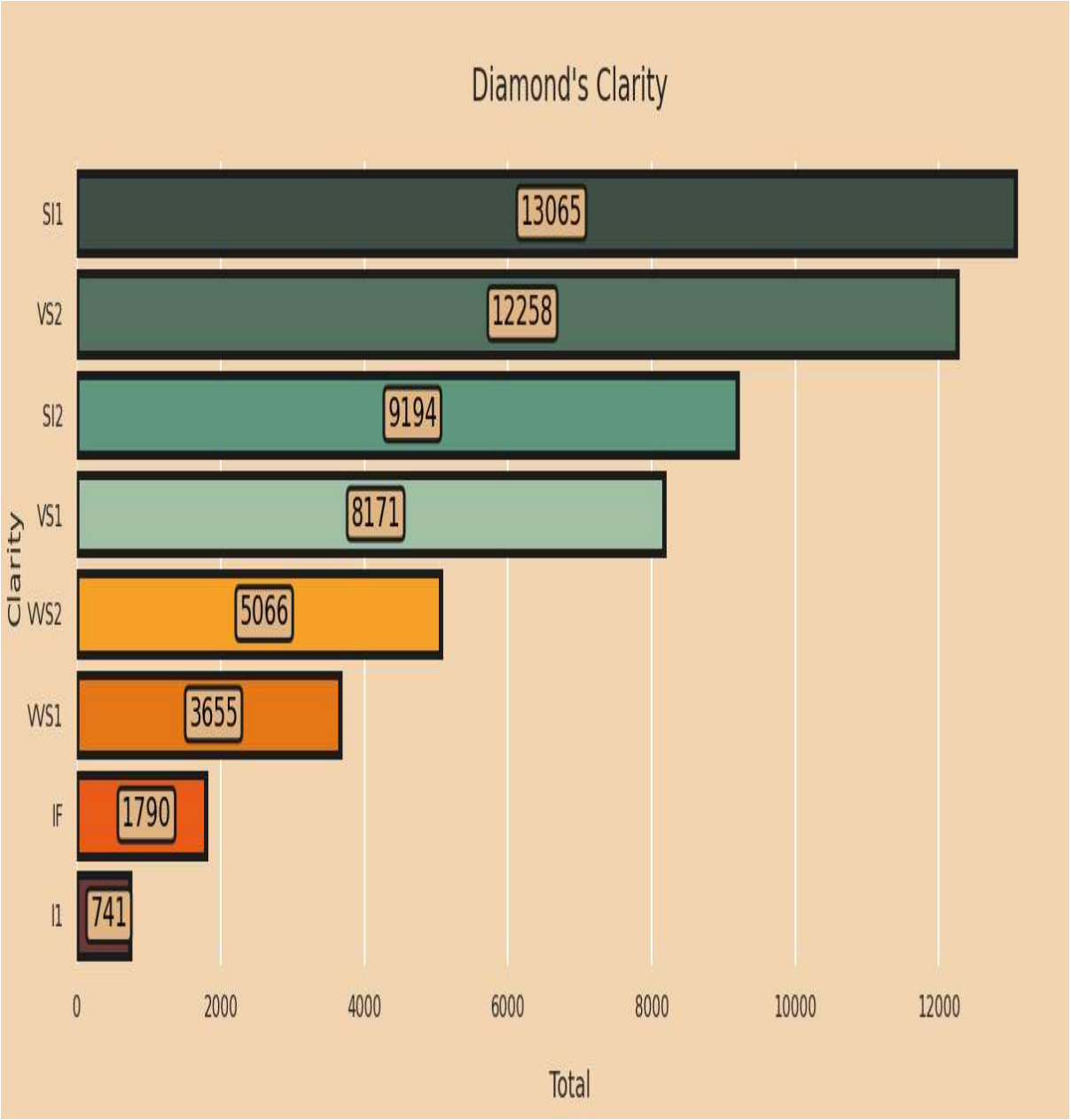


Insights:

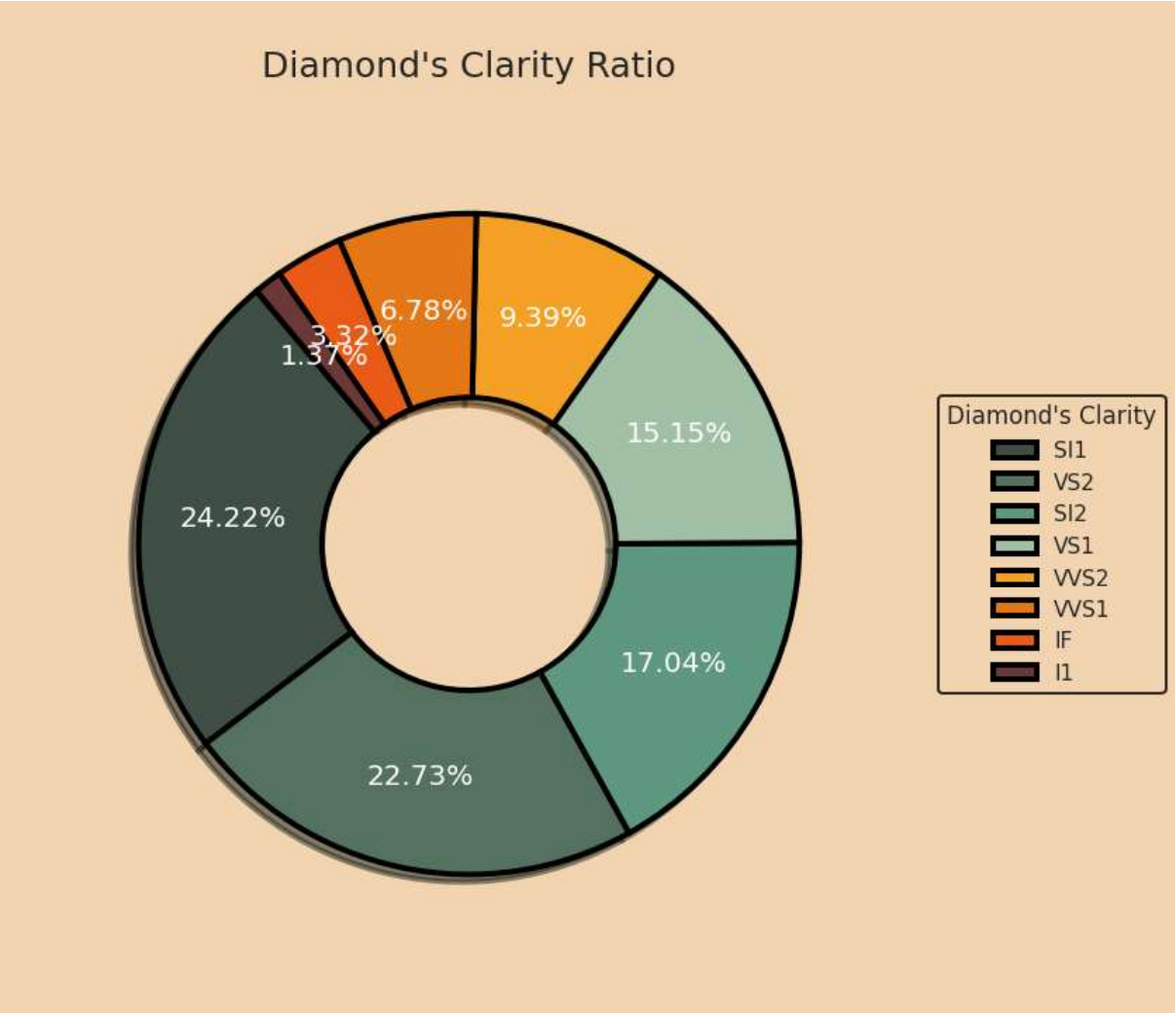
- Most of the diamonds have **G** color with a ratio of **20.93%** followed by **E** and **F**
- Only few have **J** color with a ratio of **5.21%**.

- Diamond’s Clarity

Let's have a look on the diamond's clarity :



- Diamond’s Clarity Ratio

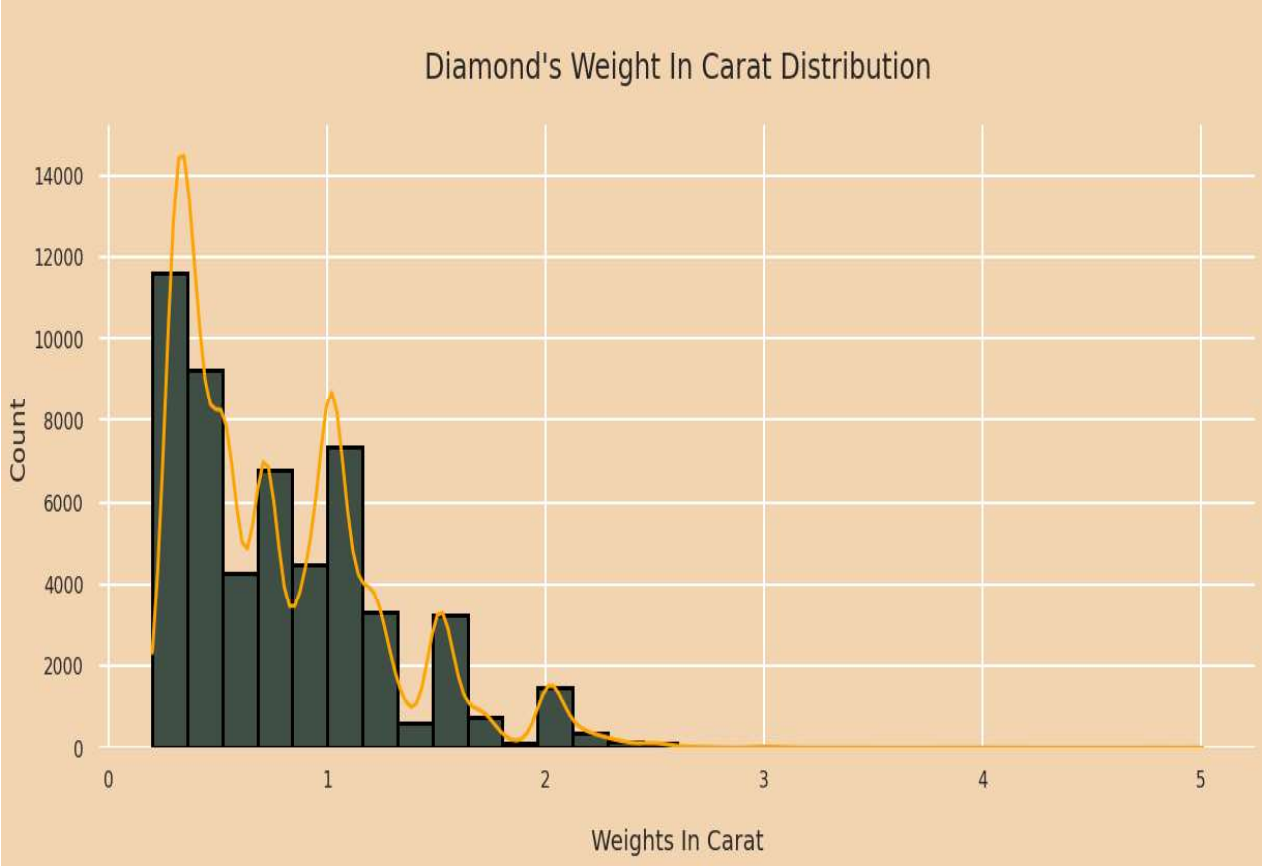


Insights:

- Most of the diamonds have **SI1** clarity with a ratio of **24.22%** followed by **VS2** and **SI2**
- Only few have **I1** clarity with a ratio of **1.37%**.

- Diamond’s Weight

Let's have a look on the distribution of weight in carat :

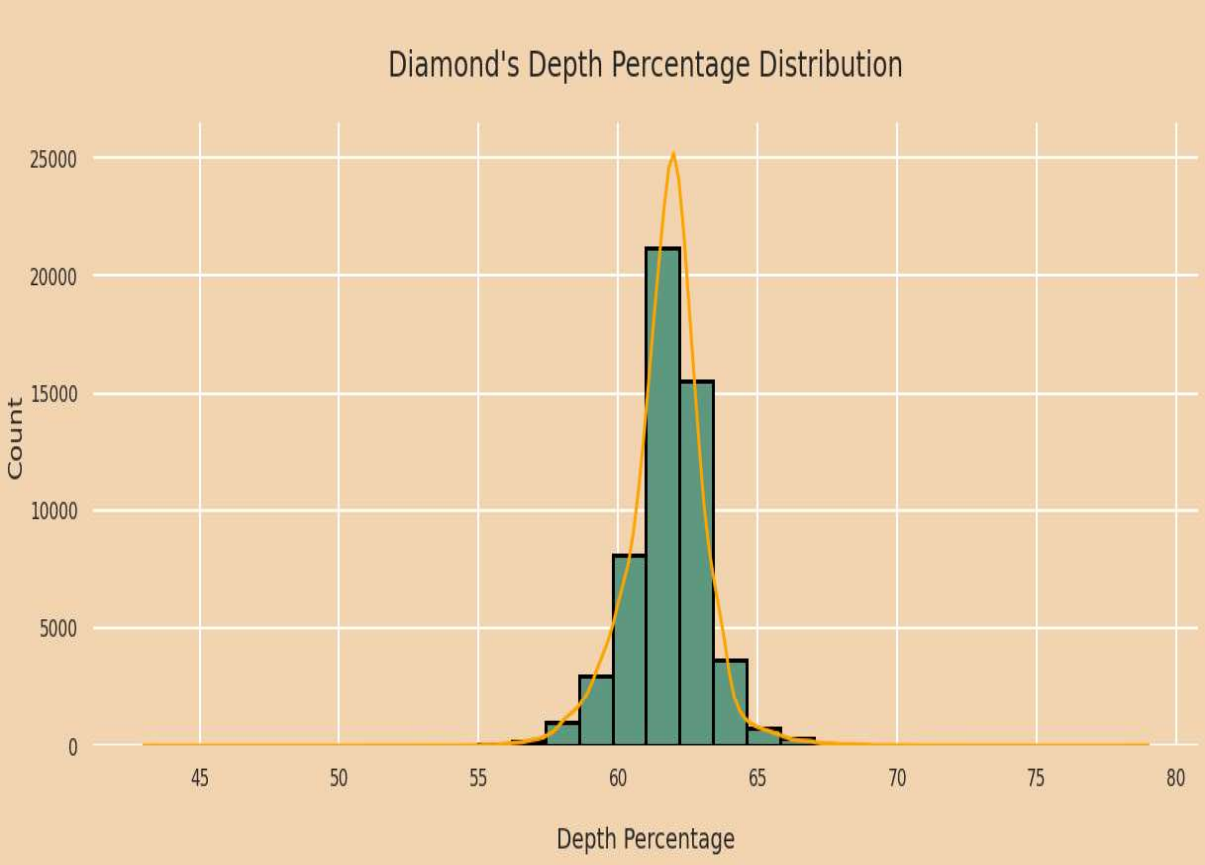


Insights:

- We can see the weight distribution is right skewed. Most of the weights fall in between **0.2 carat** to **1.2 carat**. And most of these cost in between **326** to **5000**.

- Diamond’s Depth Percentage

Let's have a look on the distribution of depth percentage :

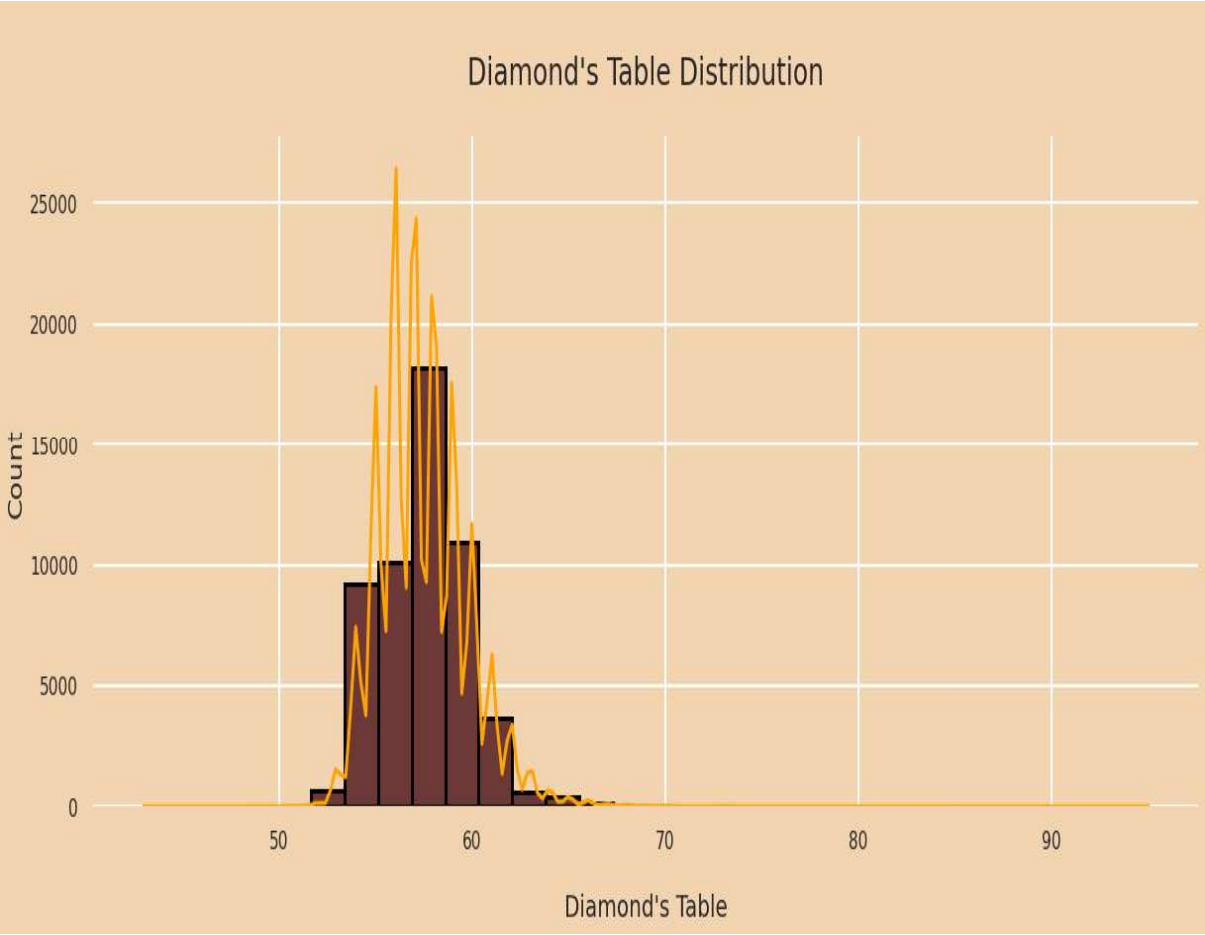


Insights:

- We can see the depth percentage distribution is normally distributed. Most of them fall in between **60** to **64**. And these cost in between **326** to **6200**.

- Diamond’s Table

Let's have a look on the distribution of diamond's table :



Insights:

- We can see most of the diamond's table fall in between **54** to **61**. And these cost in between **326** to **3800**.

Bivariate Analysis

- Diamond’s Price vs Cut

Insights:

- Most of the diamonds with **Ideal Cut** costs in between **326** to **2500**
- Most of the diamonds with **Premium Cut** costs in between **326** to **5000**
- Most of the diamonds with **Very Good Cut** costs in between **336** to **4800**
- Most of the diamonds with **Good Cut** costs in between **327** to **4700**
- Most of the diamonds with **Fair Cut** costs in between **337** to **5000**

- Diamond’s Price vs Color

Insights:

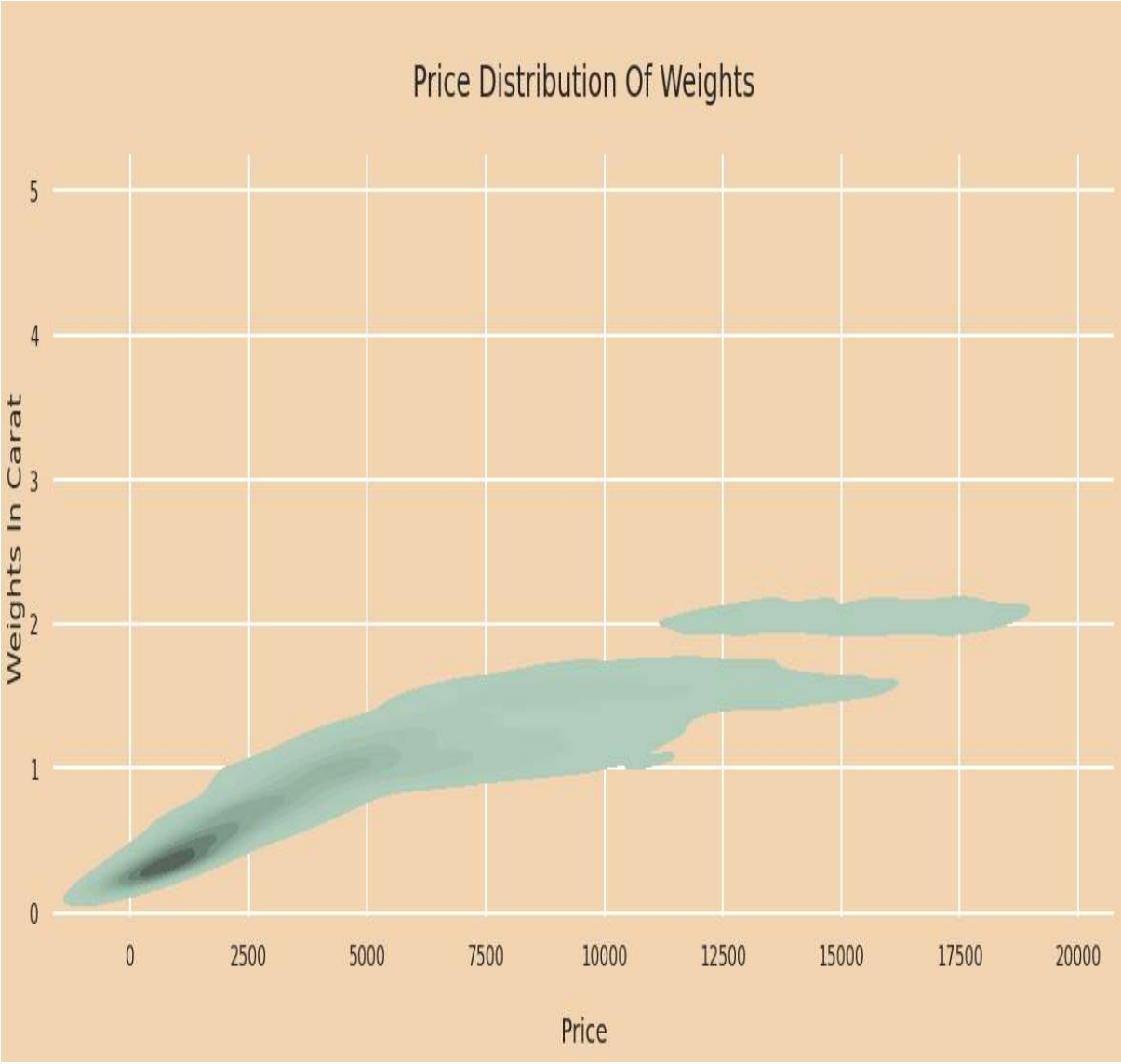
- Most of the diamonds with **G Color** costs in between **354** to **2500**
- Most of the diamonds with **E Color** costs in between **326** to **3700**
- Most of the diamonds with **F Color** costs in between **342** to **4500**
- Most of the diamonds with **H Color** costs in between **337** to **5200**
- Most of the diamonds with **D Color** costs in between **357** to **2500**
- Most of the diamonds with **I Color** costs in between **334** to **6200**
- Most of the diamonds with **J Color** costs in between **335** to **6400**

- Diamond’s Price vs Clarity

Insights:

- Most of the diamonds with SI1 Clarity costs in between 326 to 5100
- Most of the diamonds with VS2 Clarity costs in between 334 to 2600
- Most of the diamonds with SI2 Clarity costs in between 326 to 5200
- Most of the diamonds with VS1 Clarity costs in between 327 to 3600
- Most of the diamonds with VVS2 Clarity costs in between 336 to 3500
- Most of the diamonds with VVS1 Clarity costs in between 336 to 3000
- Most of the diamonds with IF Clarity costs in between 369 to 2500
- Most of the diamonds with I1 Clarity costs in between 345 to 7500

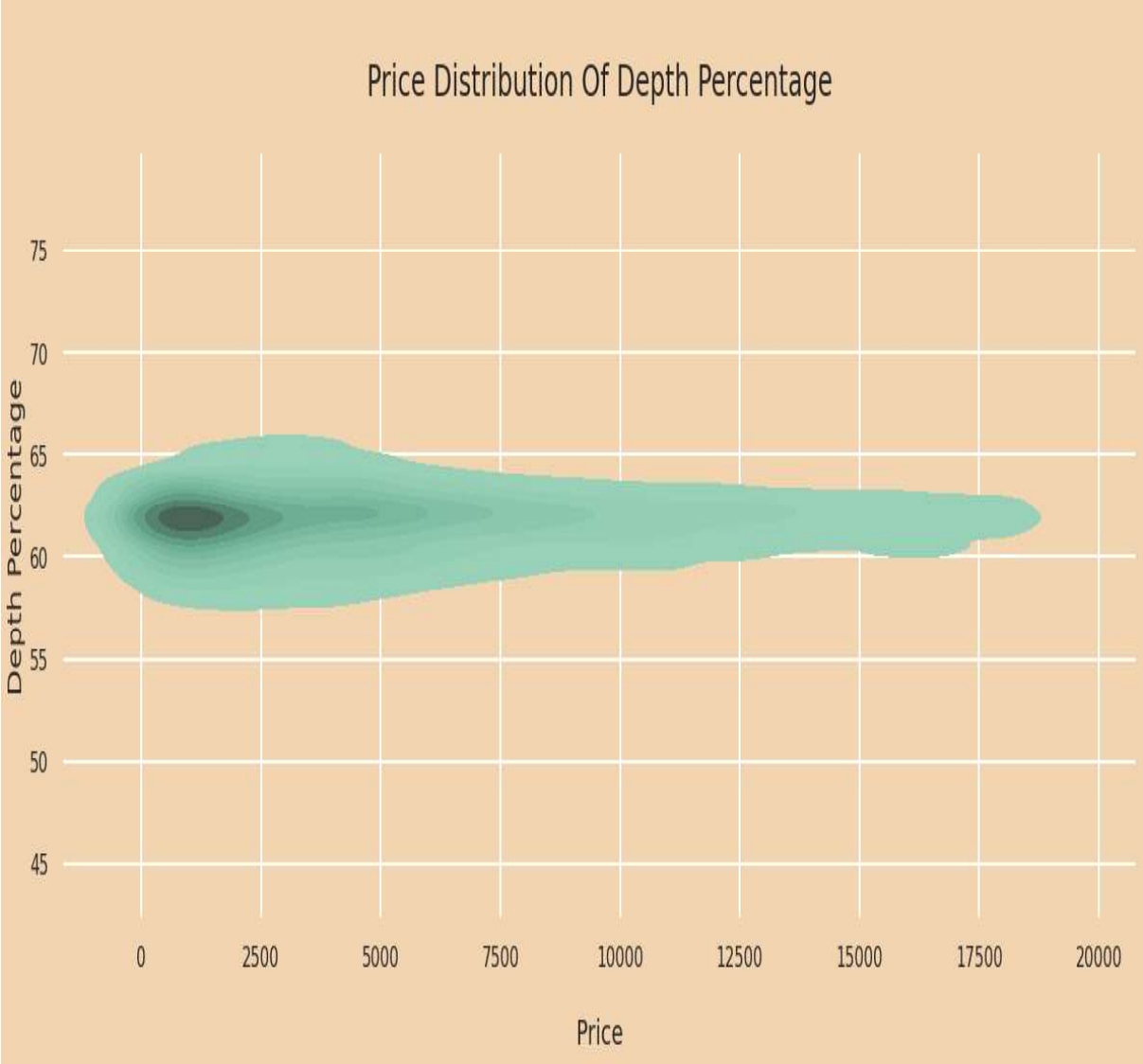
- Diamond's Price vs Weight



Insights:

- We can see the weight distribution is right skewed. Most of the weights fall in between **0.2 carat** to **1.2 carat**. And most of these cost in between **326** to **5000**.

- Diamond's Price vs Depth



Insights:

- We can see the depth percentage distribution is normally distributed. Most of them fall in between **60** to **64**. And these cost in between **326** to **6200**.

- Diamond’s Price vs Table

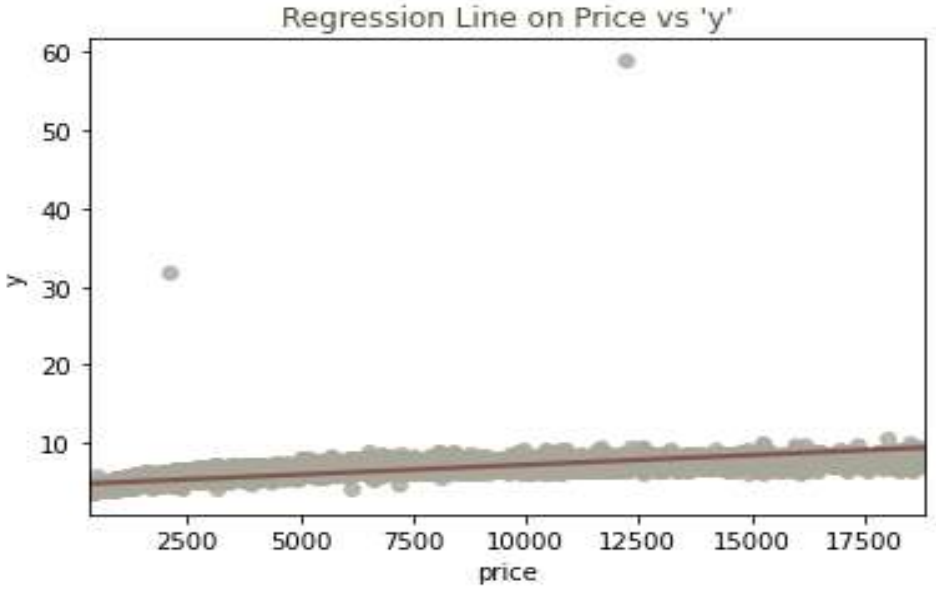
Insights:

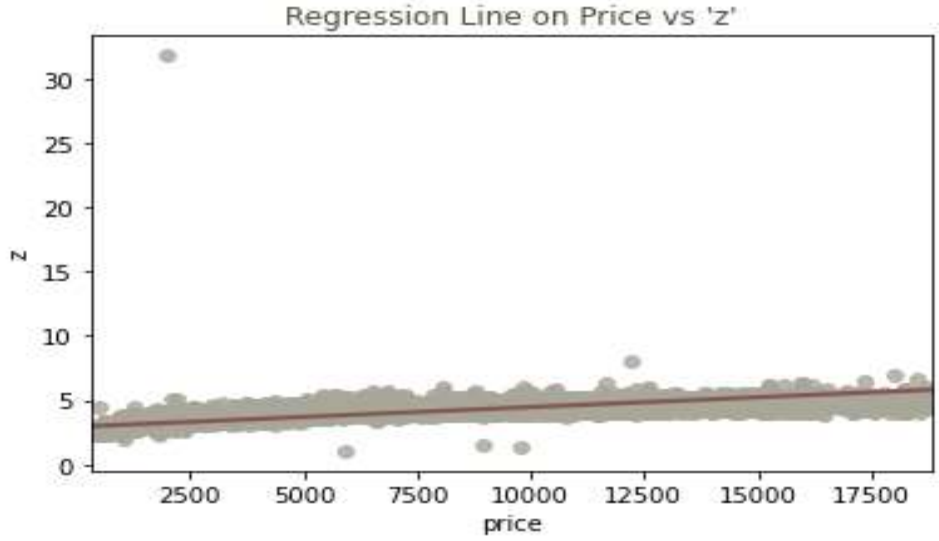
- We can see most of the diamond's table fall in between **54** to **61**. And these cost in between **326** to **3800**.

Multivariate Analysis

Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. Multivariate statistics concerns understanding the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other.

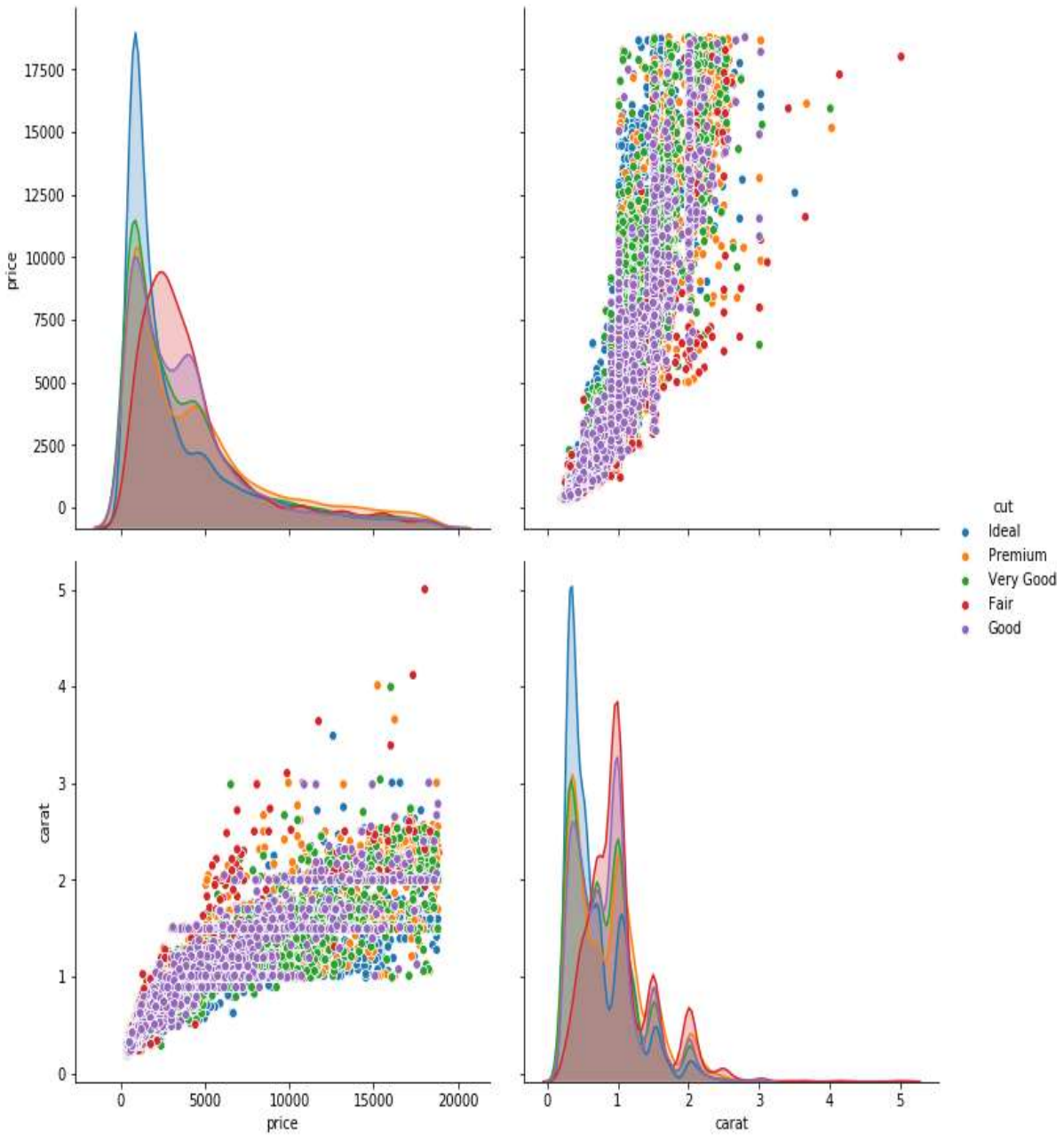
- Regression Graphs





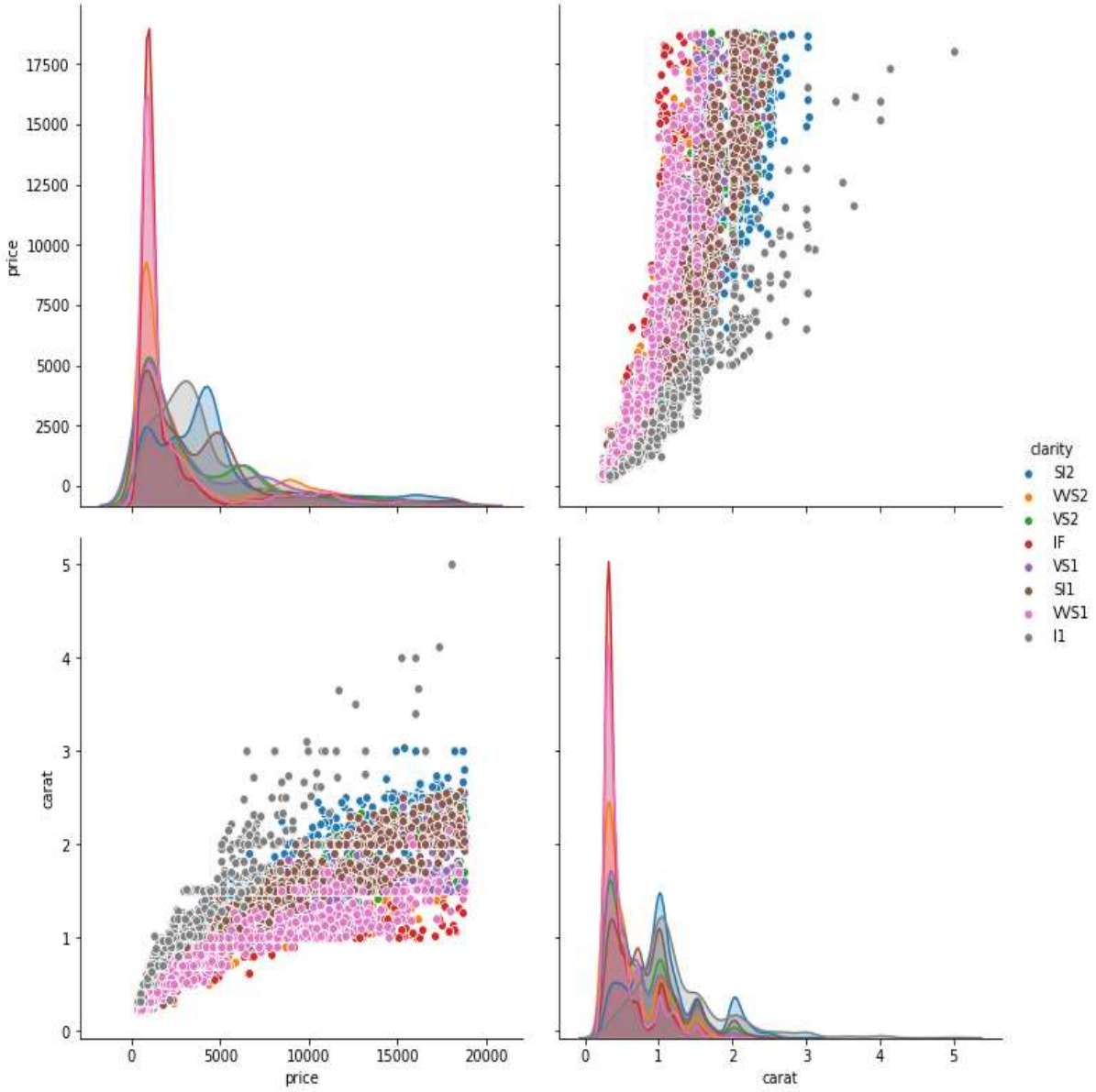
- "y" and "z" have some dimensional outlies in our dataset that needs to be eliminated.

- Price vs Cut types vs Carat



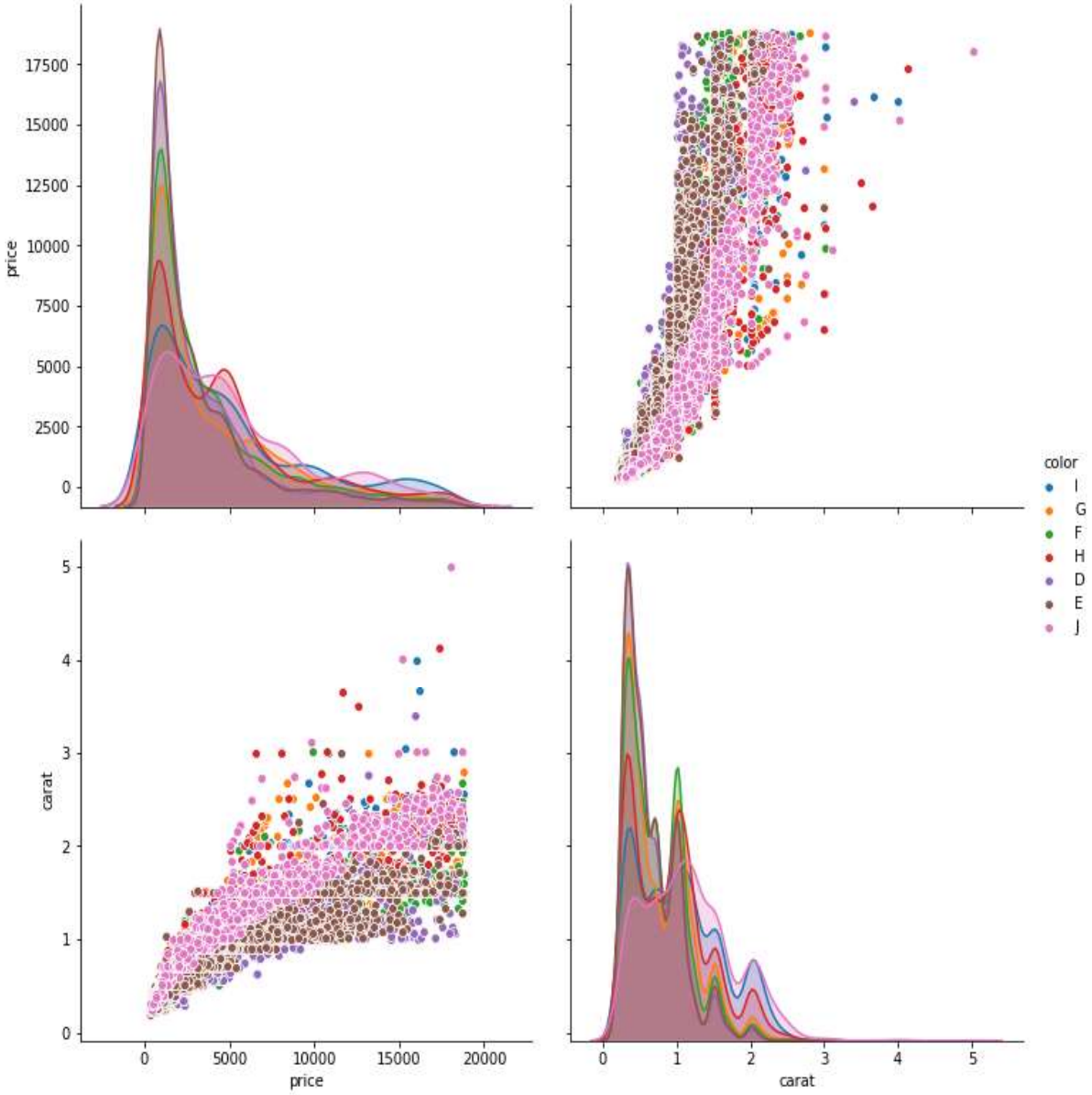
Fair cuts are most weighed, but they aren't the most expensive diamonds. Premium cuts weigh less than the fair and then cost more. Ideal cuts weigh way less and they are least expensive. The cut therefore is relatively considered while determining the price of the diamond.

- Price vs Clarity vs Carat



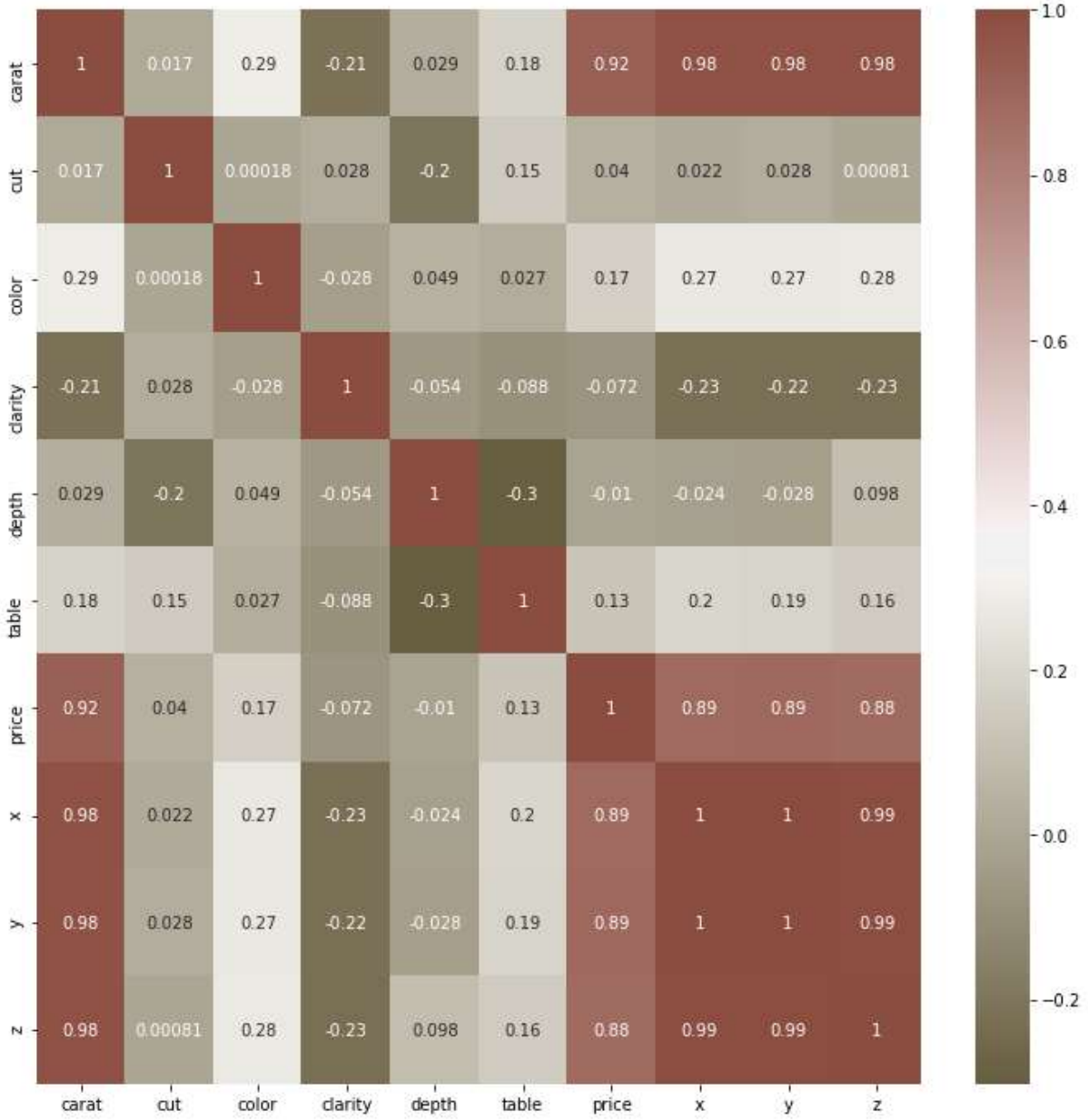
Here, we could see that I1 doesn't hold the highest clarity, even though it is the most priced. But there's something else: Apart from I1, if the rest stays, the price of a diamond could fairly be relative to its clarity, to some extent.

- Price vs Colour vs carat



Here, we could see that the color J which is the most weighed is also the most priced. The last 2 plots are very similar. We could see here that the color of the diamond is also very dependent on its price.

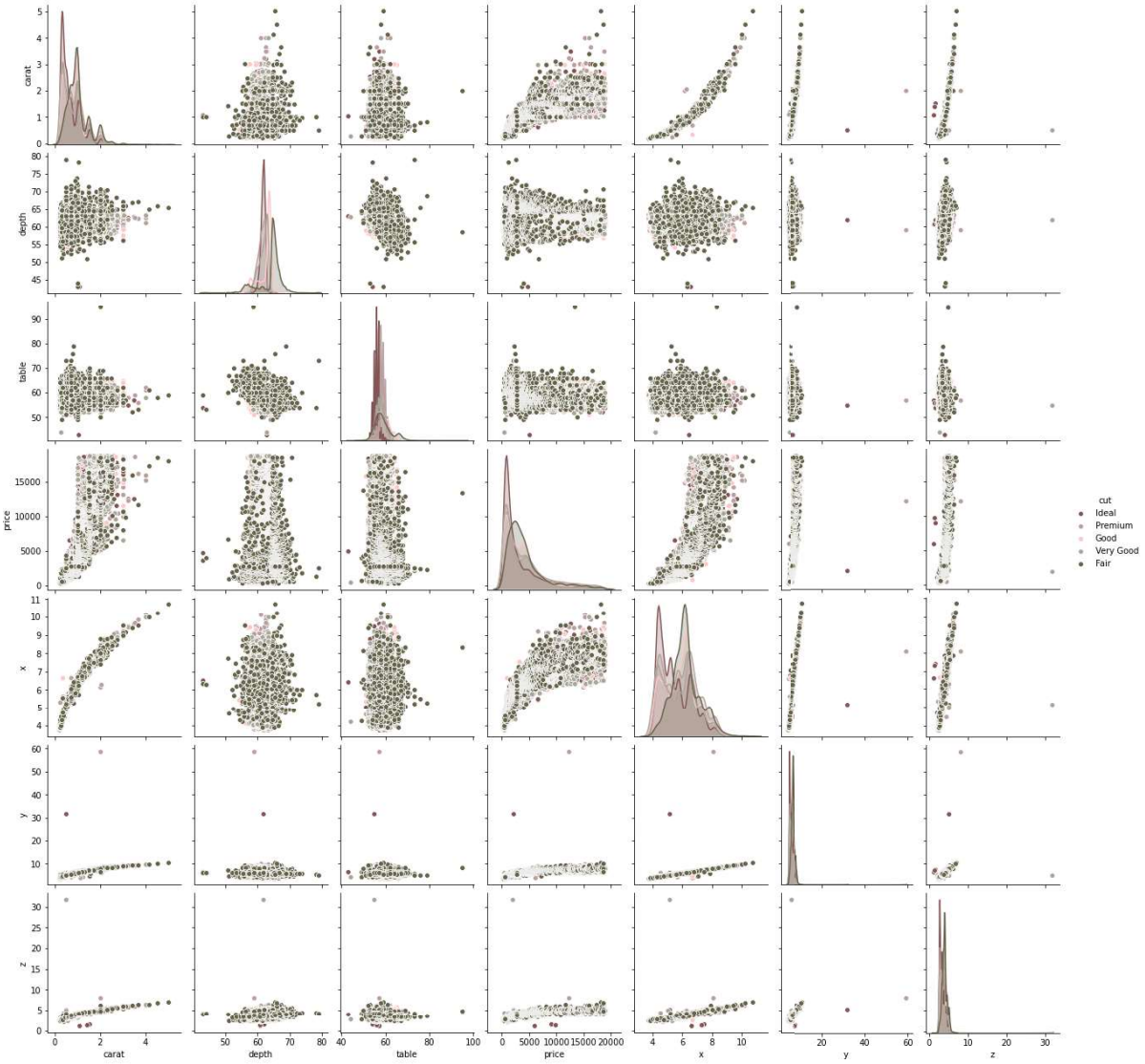
Heatmaps



Points to notice:

- "x", "y" and "z" show a high correlation to the target column.
- "depth", "cut" and "table" show low correlation. We could consider dropping but let's keep it.

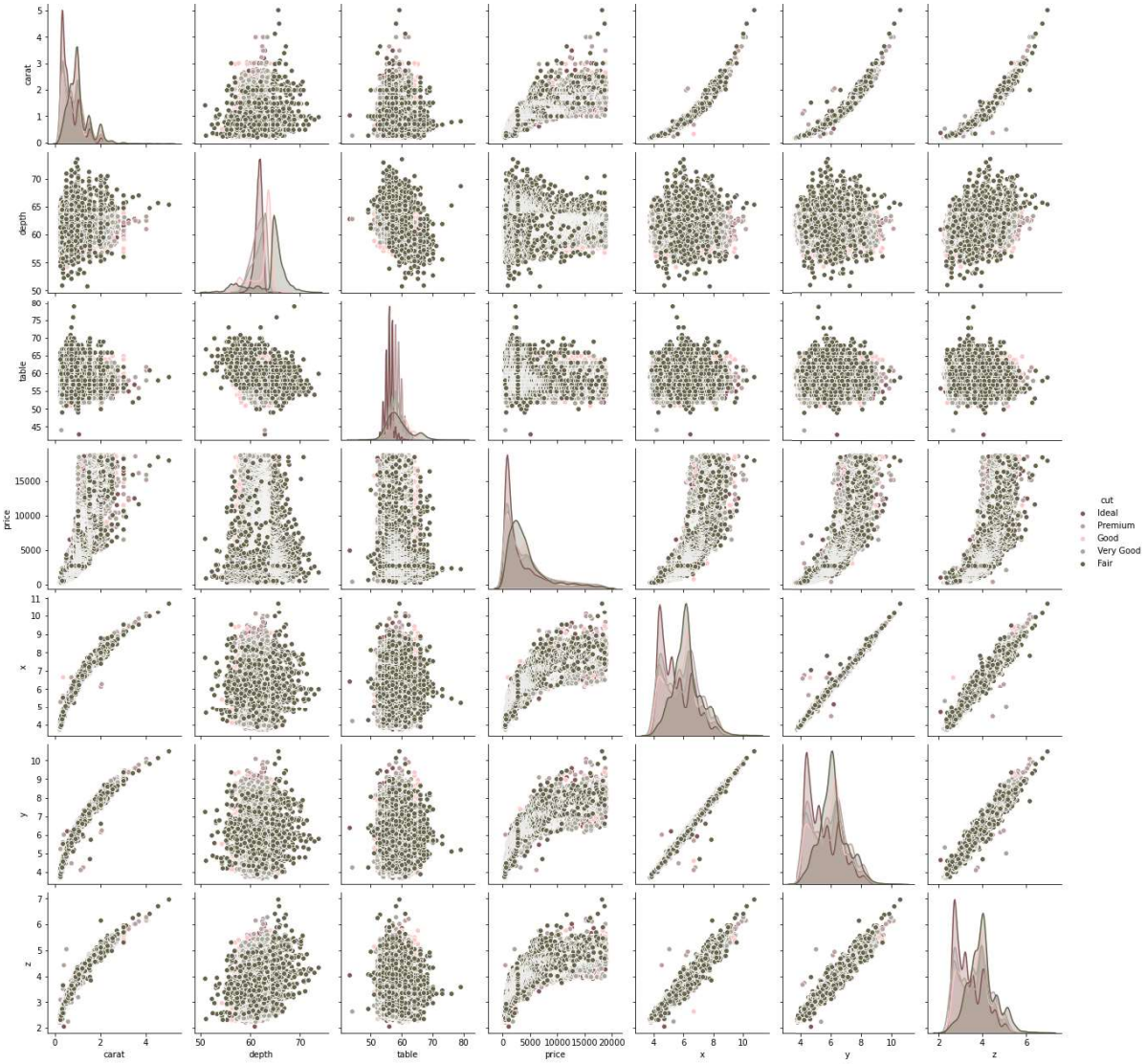
Pairplot of Dimensions



There are some features with datapoint that are far from the rest of the dataset which will affect the outcome of our regression model.

- "y" and "z" have some dimensional outlies in our dataset that needs to be eliminated.
- The "depth" should be capped but we must examine the regression line to be sure.
- The "table" featured should be capped too

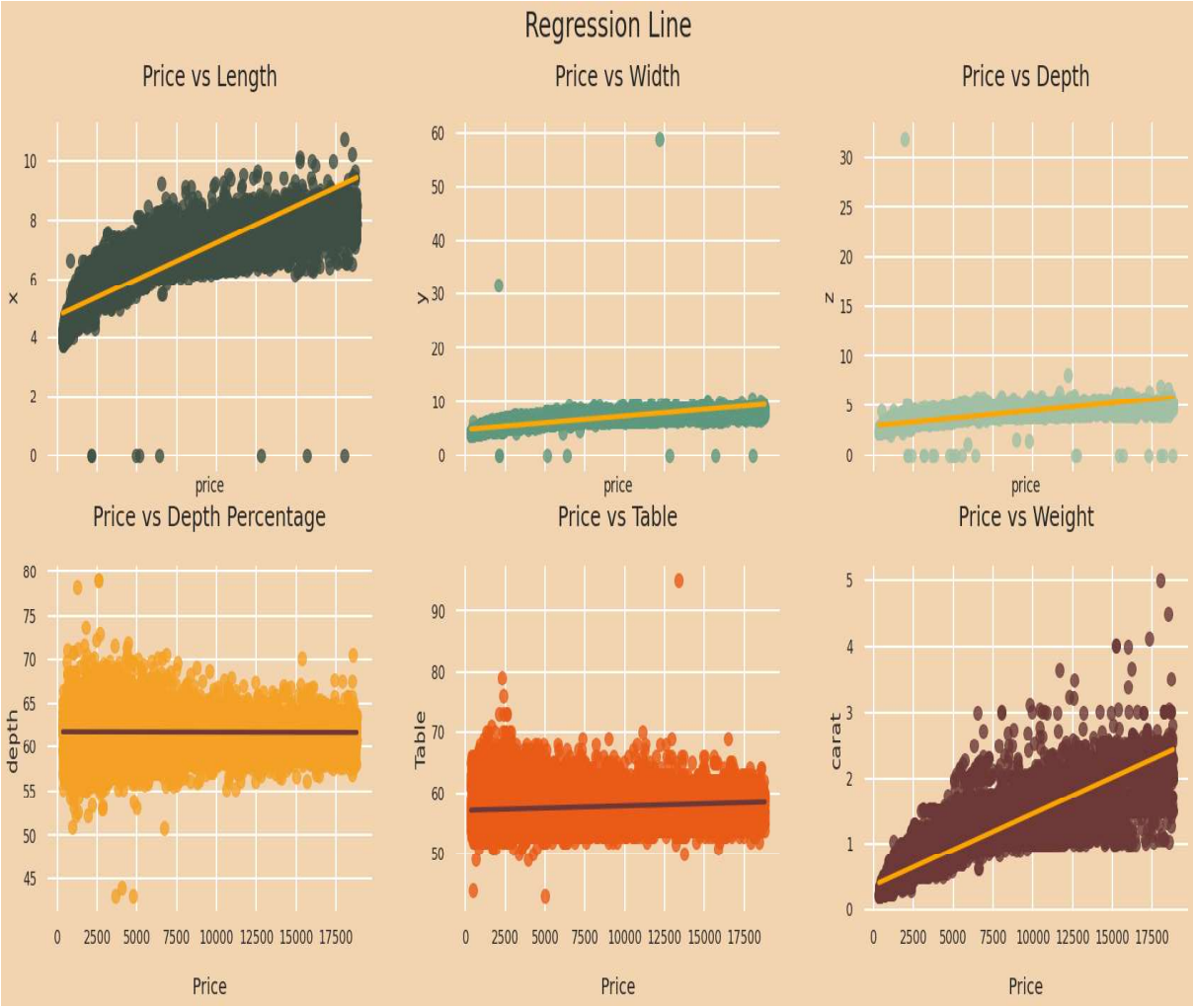
Pairplot of Cuts



Looked up the data in the pair wise form to get much clearer idea about the Cut

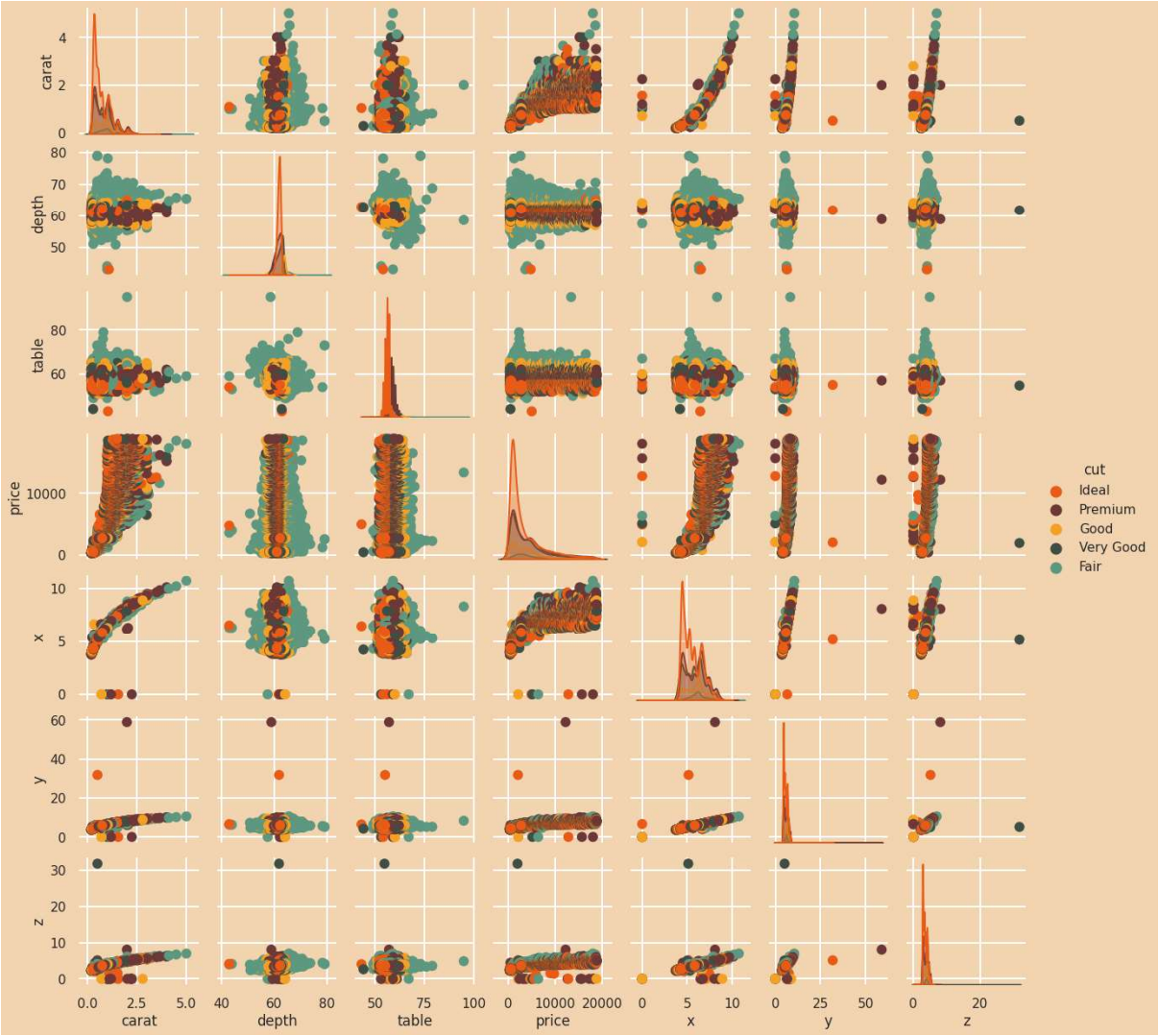
Outliers Detection

Let's have a look on the pairwise relationships :



Insights:

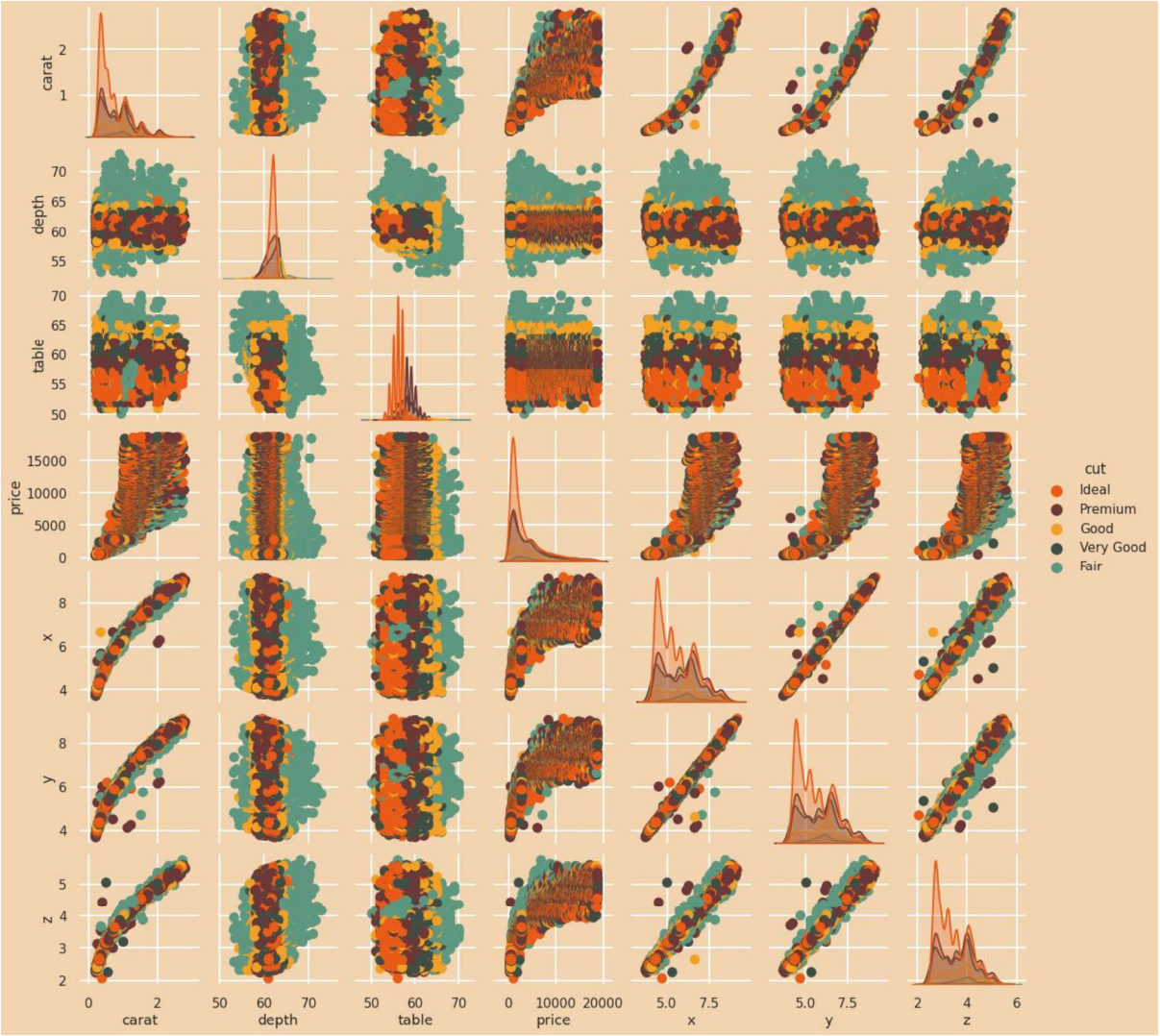
- We can clearly spot the outliers in these attributes. We will remove these outliers.



Insights:

- There are some features with datapoint that are far from the rest of the dataset which will affect the outcome of our regression model.
- **x**, **y** and **z** have some dimensional outlies in the dataset that needs to be eliminated.
- **depth**, **table** and **carat** should be capped before that we will examine the regression line for being sure.

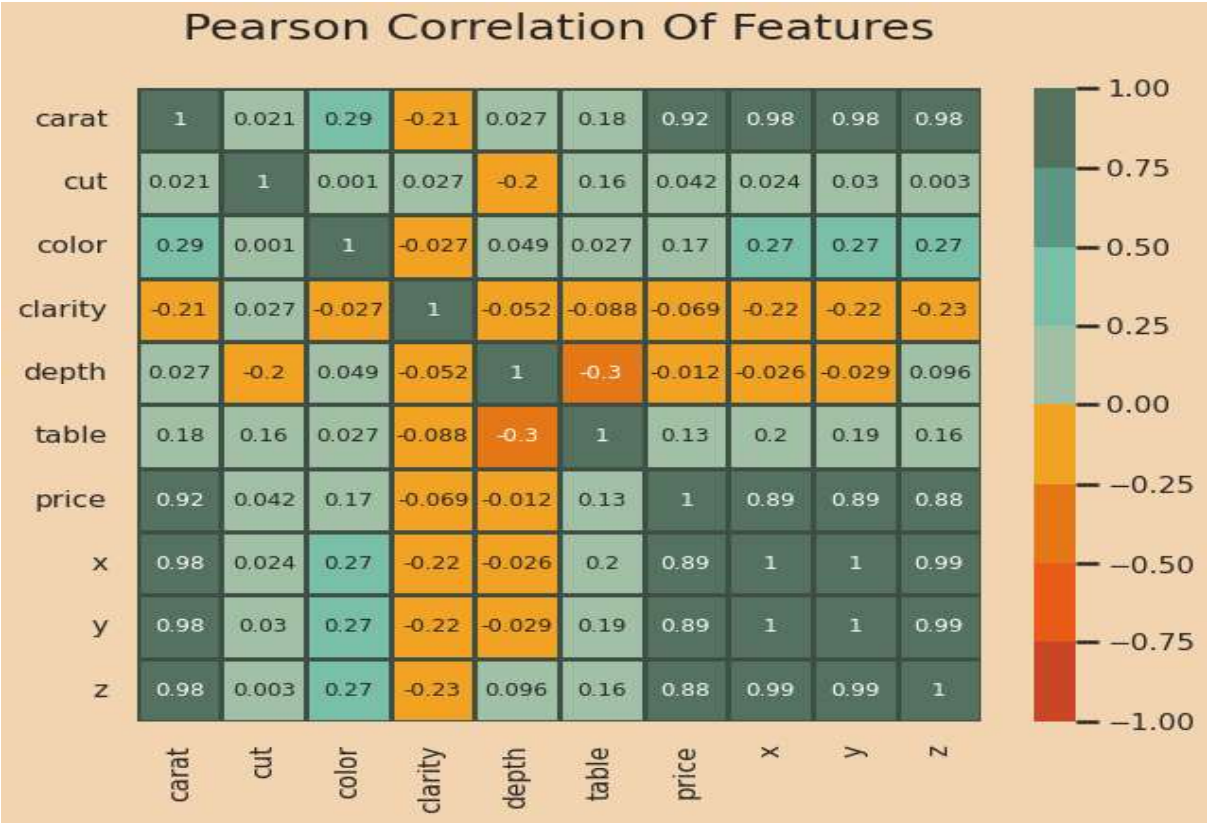
After dropping outliers, let's have a look on the pairwise relationships :



Insights:

- We have dropped the outliers.
- This is now a much cleaner dataset.

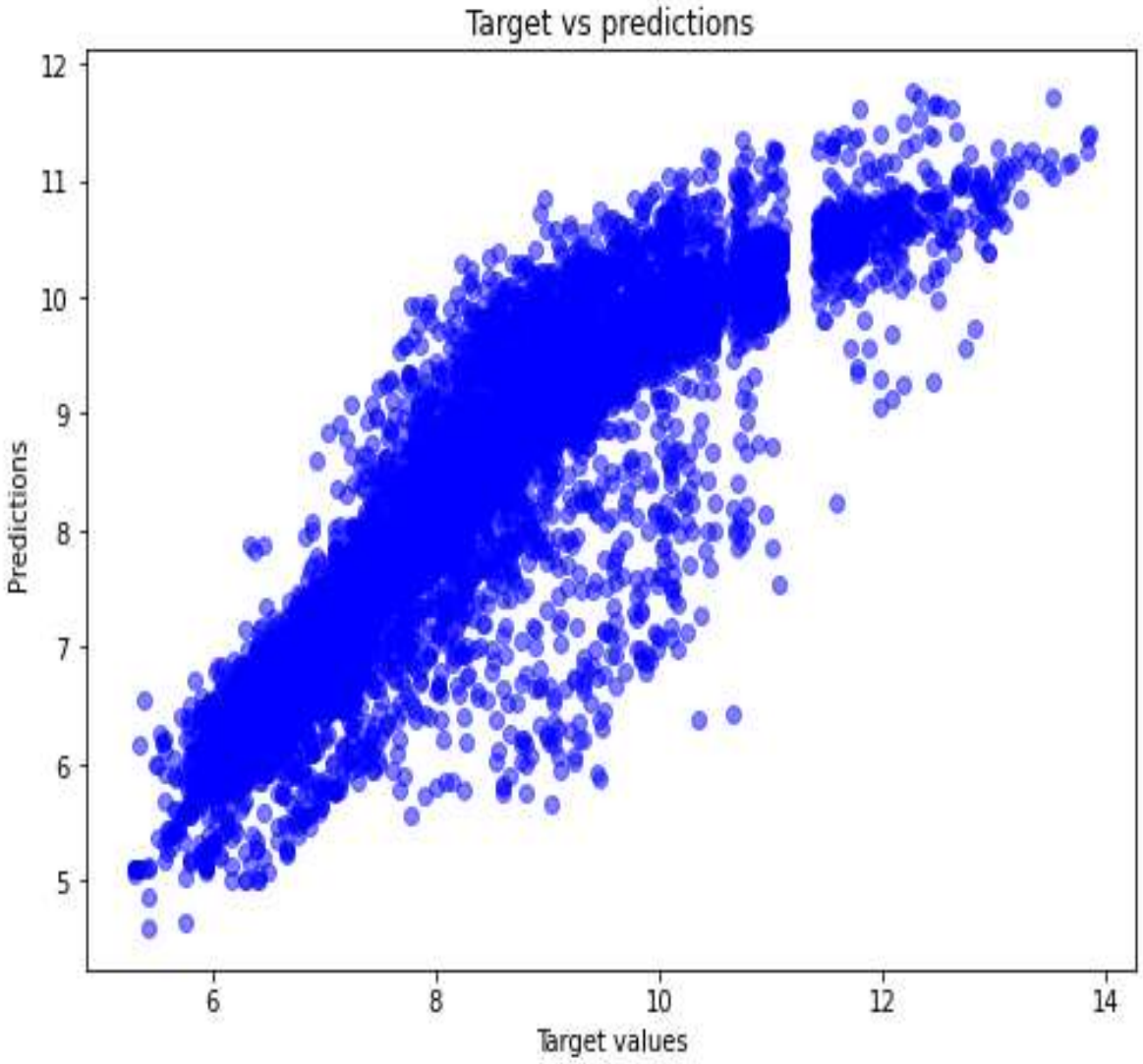
Correlation Map



Insights:

- High correlation price and length(x), price and width(y), price and depth(z).
- High correlation between carat and length(x), carat and width(y), carat and depth(z).
- length(x) is fully correlated with width(y). Length(x) is almost fully correlated with depth(z) as like width(y) and depth(z).
- between price and carat. Also high correlation between

Scope of Business in Future



In future perspective, Diamonds are going to be in great demand and the prices will hike as per the demand. The only thing which will not change in the future is that the value of a diamond will always be precious, elegant and rich.

Statistical Intuition about the Data and Overall Conclusion

The analysis of these datasets of Diamond prices gives us the indication that the prices for the best quality diamond is very expensive and on the same hand the prices of the weak quality diamonds are much lesser and these affects the markets as the bad ones will go more cheaper and thus demand will be less and the best ones will be in great demand which will make their prices touch the height of the sky. In future perspective, Diamonds are going to be in great demand and the prices will hike as per the demand. The only thing which will not change in the future is that the value of a diamond will always be precious, elegant and rich.

Thank You