exploratory data analysis of Diamonds

Piush Vaish

November 25, 2016

The aim of this report is to assess the quality and characteristics of the diamonds and gain insights about the sort of business the competitors have.

This report address few important questions such as:

1. What kind of diamonds are in stock?
2. Are diamonds priced correctly according to their carat weight?
3. Do other factors (such as cut, colour and clarity) affect the price?
4. Which factors affect the modelling of the price?

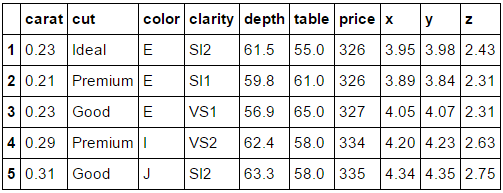
Analysis of the dataset shows that there 53940 number of diamonds within two different price bracket. These diamonds can be aimed at both cash-sensitive consumers and cash-rich luxury market. The variance in price depends on 4Cs namely Carat, Cut, Colour and Clarity along with width, length and depth. Also, multivariate linear regression is possible to predict the price using log10 scale of carat, cut, colour, clarity and volume (width \* length \* depth).

# General details about the dataset

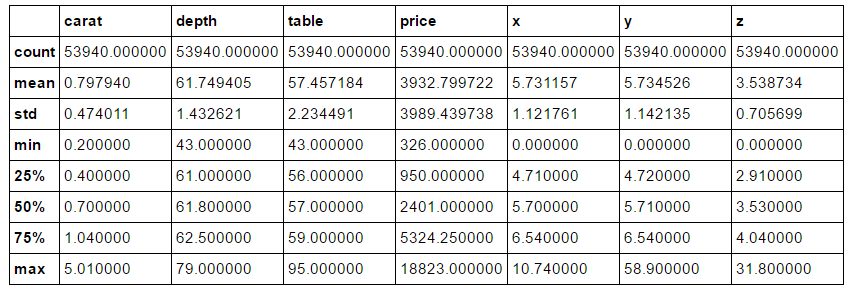
There are 53940 individual observations with 10 specific variables. There are no duplicates and missing values.

The different kind of variables are carat, cut, colour, clarity, depth, table, price, x, y, z. Cut, clarity and colour are categorical variables and others are continuous variables.

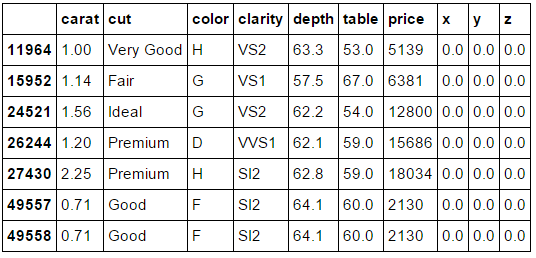
Cut, colour and clarity are ordinal variables with 5, 7 and 8 unique values respectively. Cut is the quality of the cut described as Fair, Good, Very Good, Premium, Ideal. Colour describes the diamond colour from J(worst) to D(best). Clarity is the measurement of how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best)).



Carat has 273 unique values ranging from 0.2 to 5.01 with mean of 0.797. Depth has 184 unique values ranging from 43.0 to 79.0 with mean of 61.74. Table has 127 unique values ranging from to 43.0 to 95.0 with mean of 57.45. Price is in US dollars with 11602 unique values ranging from $326 - $ 18,823 with mean of $ 3932.79.x is length in mm with 554 unique values ranging from 0 to 10.74 with mean of 5.73. y is the width in mm with 552 unique values ranging from 0 to 58.9 with mean of 5.73. z is the depth in mm with 375 unique values ranging from 0 to 31.8 with mean of 3.53.



One interesting observation from the summary of data shows that there are some values of x(length), y(width) and z(depth) which are 0. Further inspection determines that there are 8 lengths, 7 widths and 20 depths which are zero. These outliers are possible due to data entry errors, measurement errors, intentional errors or data processing errors. Also, depth percentage is calculated as z /mean (x, y) = 2\* z/ (x + y). Hence, there is inconsistency or lack of information while depth percentage was calculated.



**Table: Rows containing width as zero**.

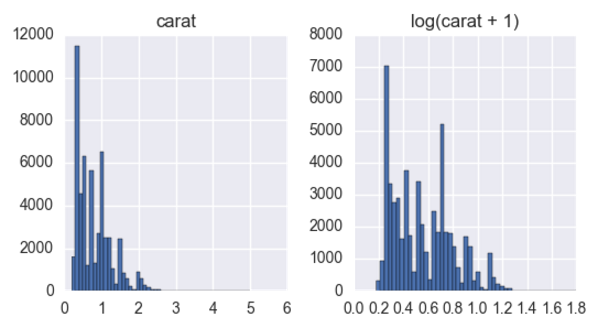
Upon further analysis when x, y, z is 0.0 reveals certain observations stand out due to the data not being consistent with the general perception. For example, observation number 27430 is premium cut, 2.25 carat weight, poor (H) colour, poor (S12) clarity but costs $ 18034. It means either the diamond’s price takes into account some other factors or it is an error. Therefore, it seems logical to remove these observations from the dataset entirely while building a model.

# Continuous Variable

The skewness of the distribution of the continuous variable is carat 1.116646, depth -0.082294, table 0.796896, price 1.618395, x 0.378676, y 2.434167 and z 1.522423 respectively. Therefore, depth is the only one with a slight negative skewness. Price and carat are very similarly skewed.

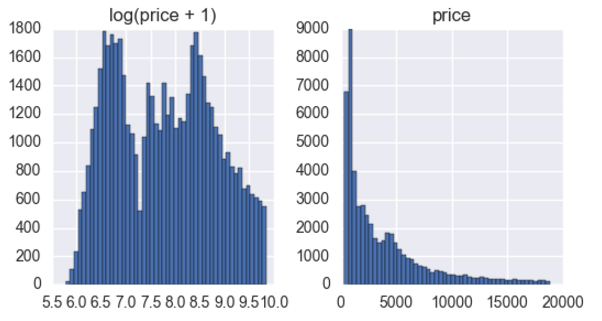
# Plotting Univariate Continuous Variables

**Carat**

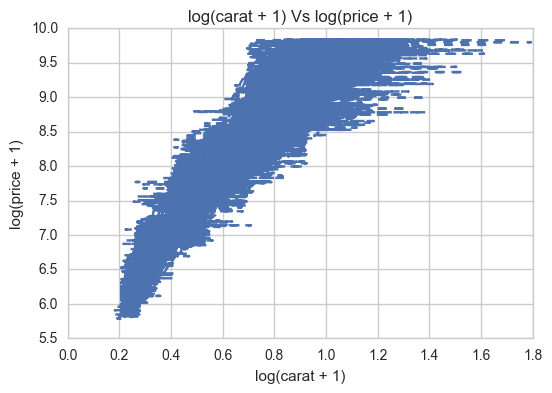
****

The carat is right skewed i.e. the observations are lower bound. Most of the diamonds are less than 1 carat. A log transformation to ensure a normal distribution indicates that there are two different kinds of diamonds in the dataset low carat diamonds and high carat. There are 9 diamonds bigger than 3.5 carats. Handful of diamonds at 4 carats and above is definitely far outside the norm. Limiting the X-axis to 3.5 implies that these 9 diamonds lay beyond the chart range. However, these outliers have very little bearing on the shape of the distribution.

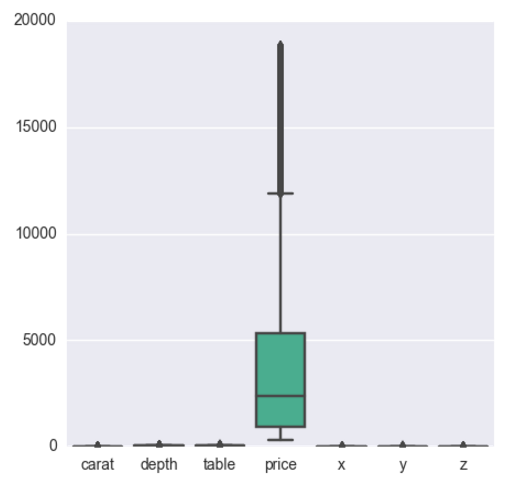
**Price**

****

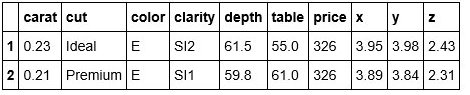
Price is a long tail distribution and right skewed. There is a high concentration of observations below the U$5,000 mark. Small amounts of very large prices are driving up the mean, while the median remains a more robust measure of the centre of the distribution. Scaling the price helps to make it a normal distribution and shows bimodality. So there are two different kinds of diamonds available for sale aimed at the luxury market and regular customers.



Applying the log transformation to carat and price variables shows a dip in the middle same as carat. Therefore, diamond’s price varies according to carat. Also, plotting log transformation of two variables helps to an almost-linear relationship. However, the increase is not completely linear because price certain higher priced diamonds does not correlate directly to the carat weight of the diamond. It means that there are other factors which influence the price. The log-transform of width, length and depth show a similar dip in the middle to carat. It confirms that there are two kinds of diamonds in the stock.

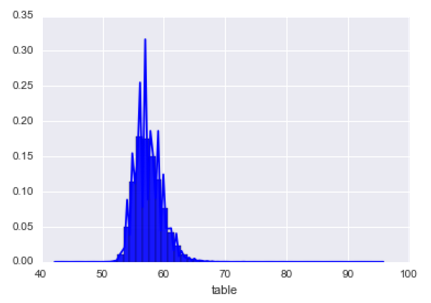
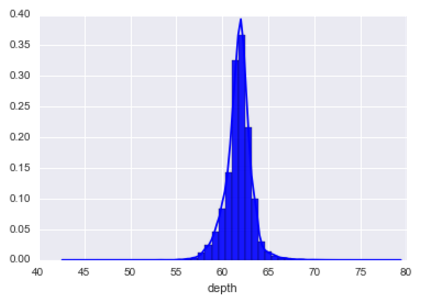


The boxplot confirms a huge variation in price. Less carat size diamonds are more sensitive to price than large diamonds. Taking log of the price will help the modelling.



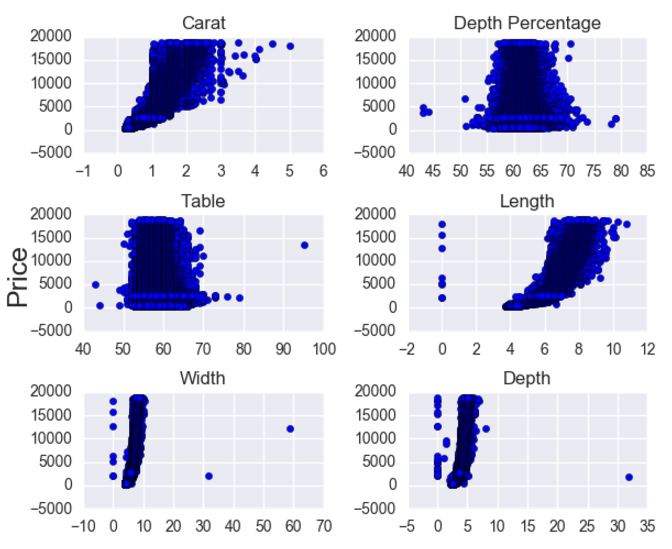
The diamond with maximum price is premium cut of 2.29 carat with value at U$18,823. There are units of diamonds with minimum price, both valued at U$326, one is 0.23 carat and Ideal cut, and another is 0.21 carats and Premium cut.

**Table and Depth**

** **

Both are fairly normally distributed with depth slightly negative skewed and mean of 61.74. Table is slightly right skewed with mean of 57.45.

# Bivariate Plots of Continuous Variable

****

**Scatter plot of price vs various continous variables**

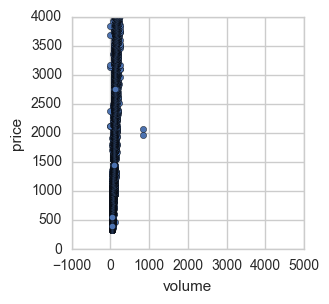
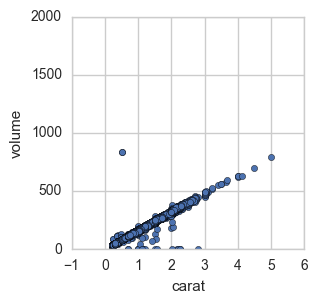
The above plot confirms that the price is linear with carat till 1 and after that the price is determined by other variables along with carat. It is a non-linear relationship and variance of the relationship increases as carat increases. There are outliers in the dataset even when 0 is removed from x, y and z. As the relationship is non-linear, running a linear model without any feature engineering will be a bad idea.

****

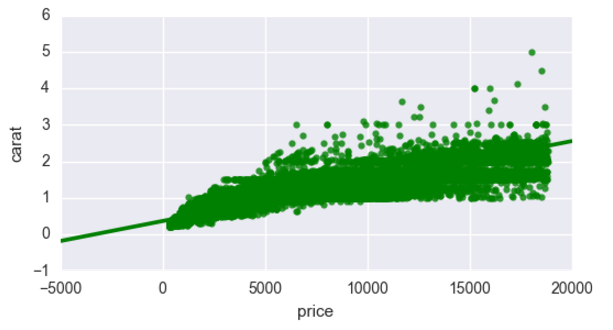
**Correlation Plot for numerical variables**

Carat along with x, y, z is positively correlated with price. Table and Depth have little correlation with price. Therefore, table and depth can be removed while modelling.

Calculation of the volume (x \* y \* z) shows that it is linear relationship with the carat size. Volume has a significant effect on the price. Therefore, we add volume during the modelling.



**Volume Vs carat Volume Vs Price**

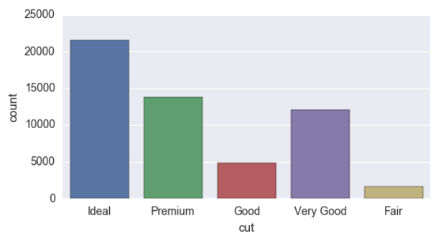
****

**Linear regression model fit of price and carat**

The above plot shows a 68% confidence interval, which corresponds with the standard error of the estimate.

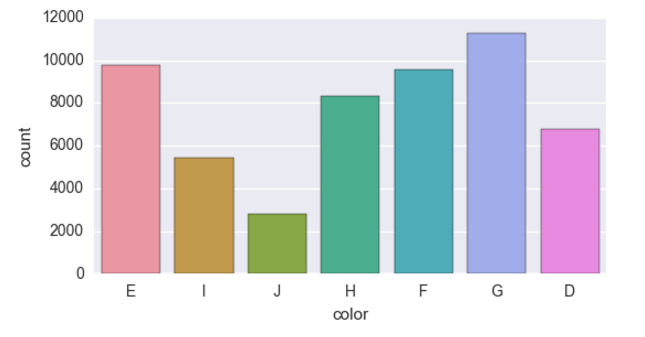
# Categorical Variables

**Number of observations of Cut Variable**



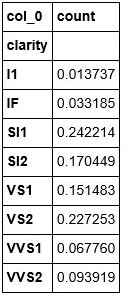
There are 21551 Ideal, 13791 Premium, 12082 Very Good, 4906 Good,1610 Fair diamonds. Therefore most of the diamonds are ideal and premium cut.

**Number of observations of Colour Variable**



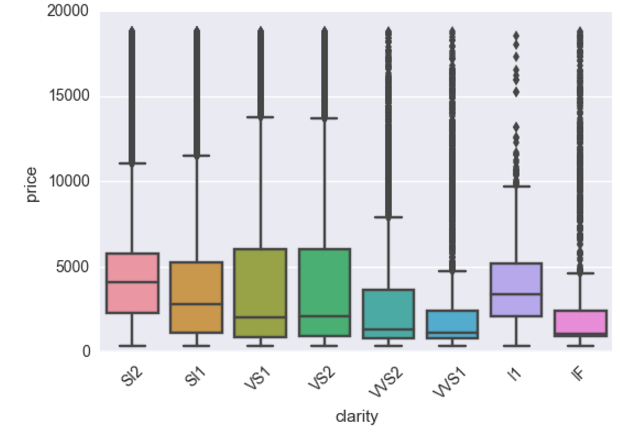
There are 11292 G, 9797 E, 9542 F, 8304 H, 6775 D, 5422 I, 2808 J diamonds. The colour of the diamond varies a lot and there are only 2808 diamonds of the best colour.

**Number of observations of Clarity Variable**

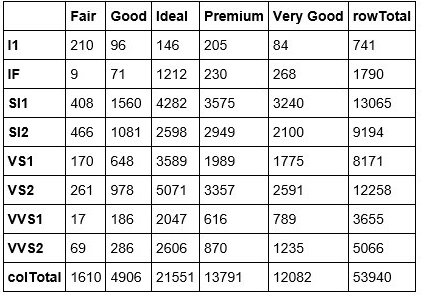


The number of diamonds according to their clarity are 13065 SI1,12258 VS2,9194 SI2,8171 VS1,5066 VVS2,3655 VVS1,1790 IF and 741 I1.The frequency table allows to extract the proportion of the data that belongs to each category. SI1 and VS2 have the frequency of 0.24 and .22. Hence, the clarity of most of the diamonds is on the lower to middle scale.

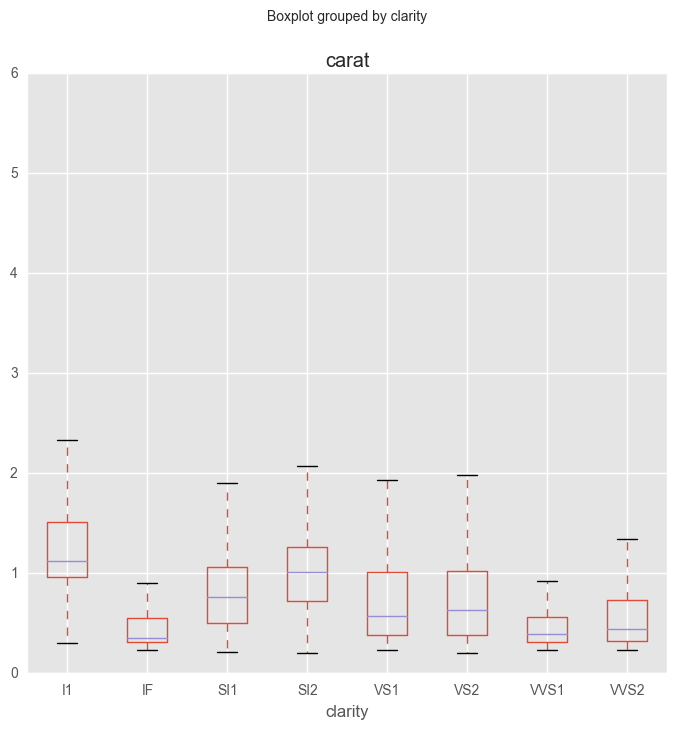
# Bivariate Analysis



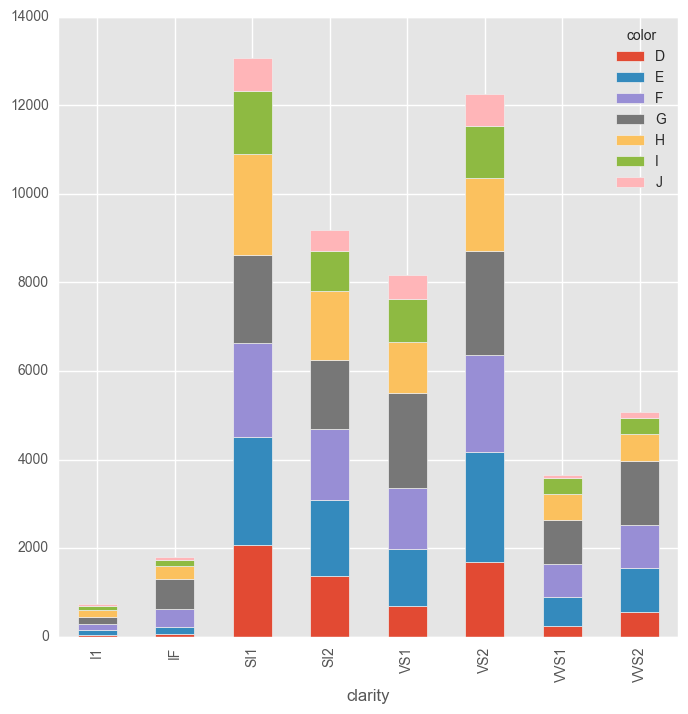
The boxplot shows that diamonds on the highest end of the clarity spectrum (IF = internally flawless) have a lower median prices than low clarity diamonds. It is unusual because diamonds with better clarity are expected to fetch higher prices. Hence there has to be other factors deciding the price of diamonds. The plot below displays relationship between cut and clarity.



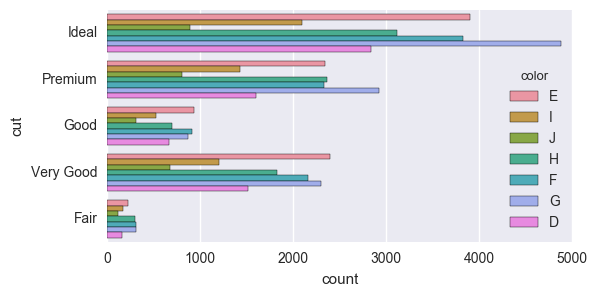
Even with low clarity, most of the diamonds have ideal and premium cut which can have an affect on the price.



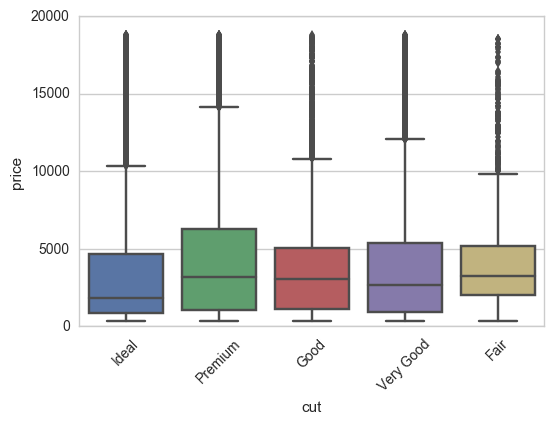
The box plot shows that diamonds with low clarity ratings also tend to be larger. Since size is an important factor in determining a diamond's value, it isn't too surprising that low clarity diamonds have higher median prices. Lighter diamonds are more expensive if they have a high clarity rating and conversely some of the heavier diamonds are not as expensive for having a low clarity rating.



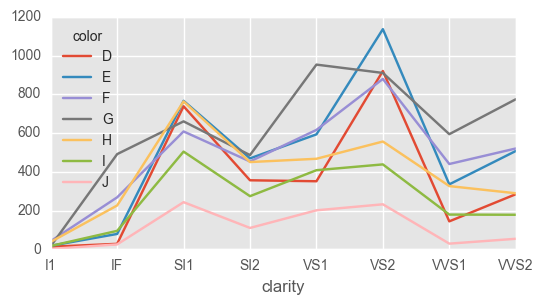
The above stacked bar chart shows that low clarity diamonds have good colour with the lowest clarity diamonds (I1) having no best colour (D). Also, there is no association between the clarity and colour. Even though the clarity increases colour quality varies e.g. High clarity quality (IF) diamonds contain very few diamonds with best colour.



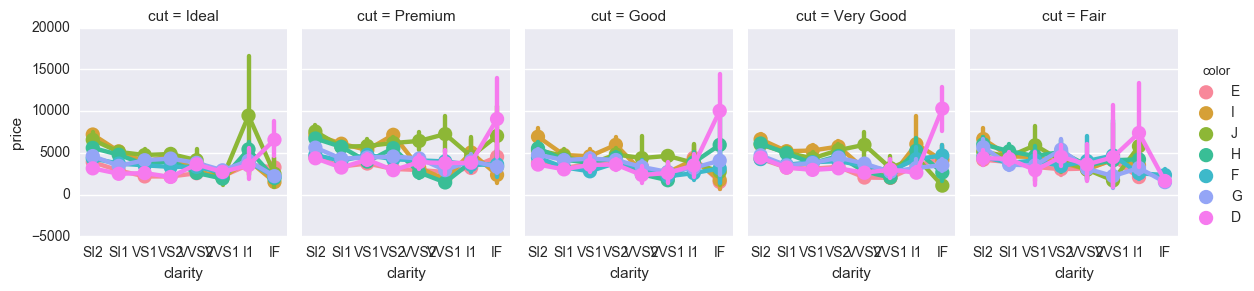
The most colour quality in both ideal and premium cut diamonds is G. G is on the poor side of the scale of the colour quality. Therefore, cut of the diamonds is not effected by the colour.



Even though ideal cut has a lot of variation in the price, it has a lower median price than any other cut. Premium, good, very good and fair have almost similar kind of median price.



The above plot at ideal cut shows that the most of the diamonds are at VS2 clarity with the best colour at E and D. The number of diamonds at ideal cut, colour E and VS2 clarity are 1136.



A mixed 3-way ANOVA compares the mean differences between clarity and price over cut and colour. The mixed ANOVA shows an interaction between these two factors on the dependent variable. Price is determined with clarity, colour and cut. For example, with I1 clarity and ideal cut, even with the worst grade of colour (J) the price is very high. Similarly, as the quality of cut decreases the clarity and good colour determines the price. Therefore, a combination of all three factors can have an effect on the price.

In conclusion, an analysis of the dataset states that the variance of price is determined with the 4Cs (Carat, Cut, Clarity and Colour) along with width, length and depth.

# Modelling

Bivariate plot analysis of continuous variables demonstrates that log10 scale of carat and volume (width \* length \* depth) have a linear relationship to the price. Price is independent of any dependent values and the distribution of price has almost the same variance with each value of dependent variable. Taking a log10 scale of price ensures a roughly normally distribution. Also, cut, clarity and colour have an effect on the price. As cut, clarity and colour are ordinal categorical variables, these three need to encoded into numerical variables depending on their order.

Therefore, it is possible to model the price of the diamonds using multivariate linear regression. Model will include cut, clarity, log10 scale of carat, depth and volume (width \* length \* depth) to determine the log10 scale of price of the diamonds.

# References

### Books

1. Sebastian Raschka, 2015. Python Machine Learning. 1 Edition. Packt Publishing.
2. Hector Cuesta, 2013. Practical Data Analysis. Edition. Packt Publishing.
3. Wes McKinney, 2012. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 1 Edition. O'Reilly Media.
4. Roger D. Peng, 2016. Exploratory Data Analysis with R. 1 Edition. Leanpub
5. John D. Kelleher, 2015. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (MIT Press). 1 Edition. The MIT Press.

### Websites

1. <http://scikit-learn.org/stable/>
2. <http://seaborn.pydata.org/>
3. <http://matplotlib.org/>