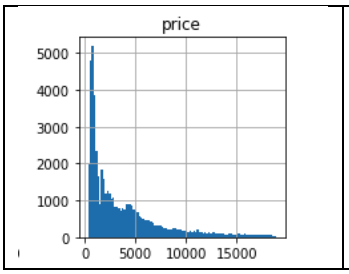
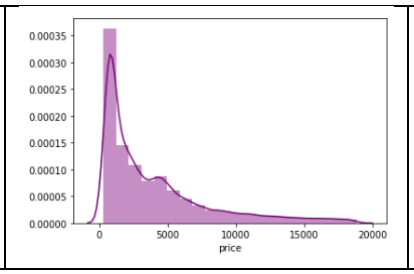
Scikit-Stack and Deep Learning - Predict diamond prices

The goal is to use the given 9 independent variables (6 numerical, 3 categorical) to predict the diamond price (target variable). The dataset contains 53,940 entries.

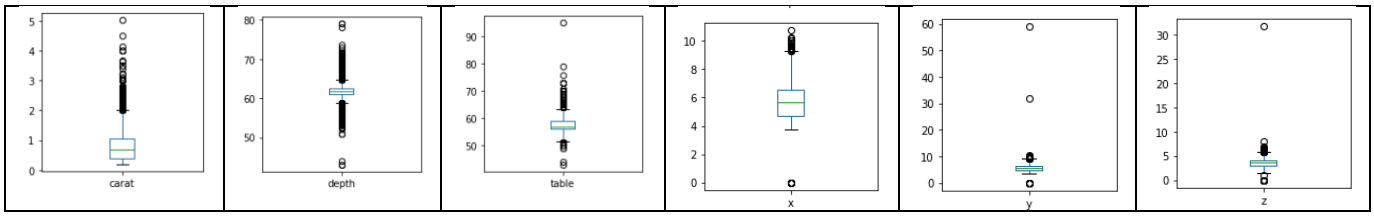
	count	mean	std	min	25%	50%	75%	max
carat	53940.0	0.797940	0.474011	0.2	0.40	0.70	1.04	5.01
depth	53940.0	61.749405	1.432621	43.0	61.00	61.80	62.50	79.00
table	53940.0	57.457184	2.234491	43.0	56.00	57.00	59.00	95.00
price	53940.0	3932.799722	3989.439738	326.0	950.00	2401.00	5324.25	18823.00
x	53940.0	5.731157	1.121761	0.0	4.71	5.70	6.54	10.74
y	53940.0	5.734526	1.142135	0.0	4.72	5.71	6.54	58.90
z	53940.0	3.538734	0.705699	0.0	2.91	3.53	4.04	31.80



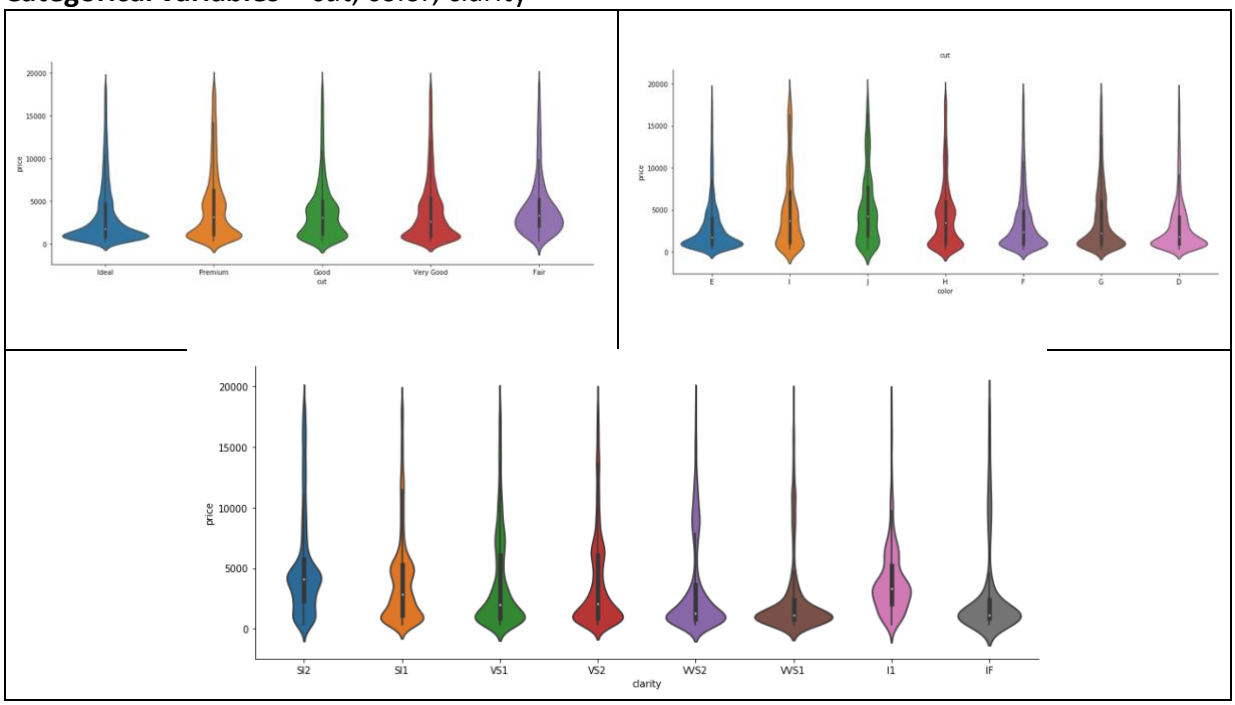


Price is Target Variable
Range: 326 ~ 18823
Mean: 3932.799722
Median: 2401
Mean > Median : right skew
High peak around 1,000 USD

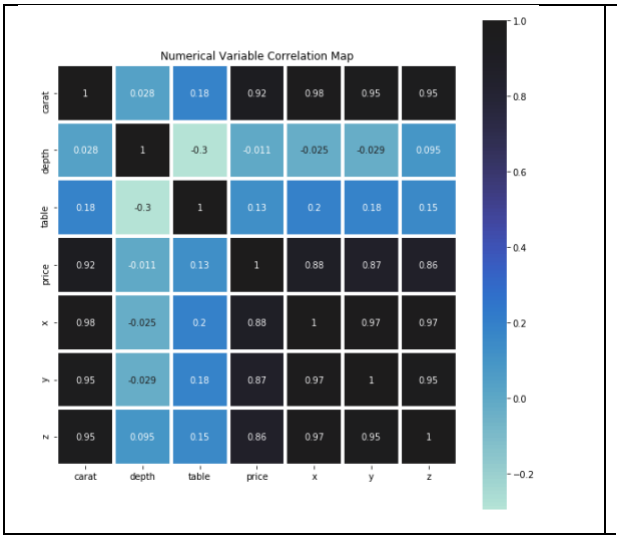
Numerical variables – carat, depth, table, x, y, z

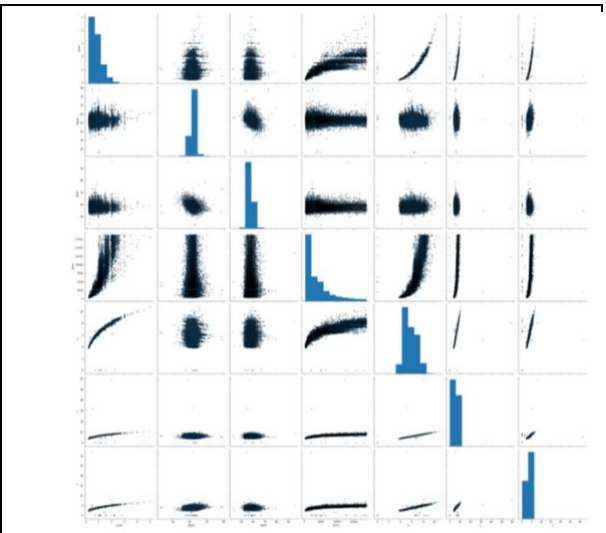


Categorical variables – cut, color, clarity



Among the independent variables, we can inspect that there some independent variables have very strong correlation with target variable (price) : carat (0.92), x(0.88), y(0.87), z(0.86). Also high correlation among independent variables (x, y, z and carat).





Missing values are when the dimension of the diamonds x = length, y = width, z = depth equal to 0 as it is not normal when it is 0, we treated it as missing values and remove them.

Results from 2 models: Multiple linear regression and Random Forest Regressor are being build. Also, due to the wide range of variables. We scale the variables

Model Summary is listed as below.

Model	**R2**	**Adjusted R2**	**MSE**	**MAE**
Linear Regression	0.9195	0.9193	1,290,368	726.29
Random Forest	0.9828	0.9827	274,684	263.92
Linear Regression *(standardized)*	0.9195	0.9193	1,290,368	726.29
Random Forest *(standardized)*	0.9830	0.9827	272,372	263.71

According to the result shown in the table, both models can explain more than 90% of the variability of the response data. As for the performance, the Random Forest has higher R2 and lower MSE and MSAE which indicated Random Forest models explain more of the response data and have smaller mean squared error.

Findings so far: Random forest regressor after standardization of numerical independent variables have the best performance among them.

Task 4: Deep Learning model using multiple linear Regression

We also build deep Learning sequential model to predict the price from diamonds dataset and the performance is also listed in below table.

Model	**R2**	**Adjusted R2**	**MSE**	**MAE**
Linear Regression	0.9195	0.9193	1,290,368	726.29
Random Forest	0.9828	0.9827	274,684	263.92
Linear Regression (standardized)	0.9195	0.9193	1,290,368	726.29
Random Forest (standardized)	0.9830	0.9827	272,372	263.71
Deep Learning	0.976	0.976	380,527	302.14

To sum up: the Random Forest with standardized independent variables are with the best result among these models.