The lapidarist problem

The lapidarist problem essentially is a problem to predict the value of some stolen diamonds. The approximation to the problem will be divided into three parts. Data cleaning, Variables analysis and data modeling in order to explain the relevant findings.

Data cleaning

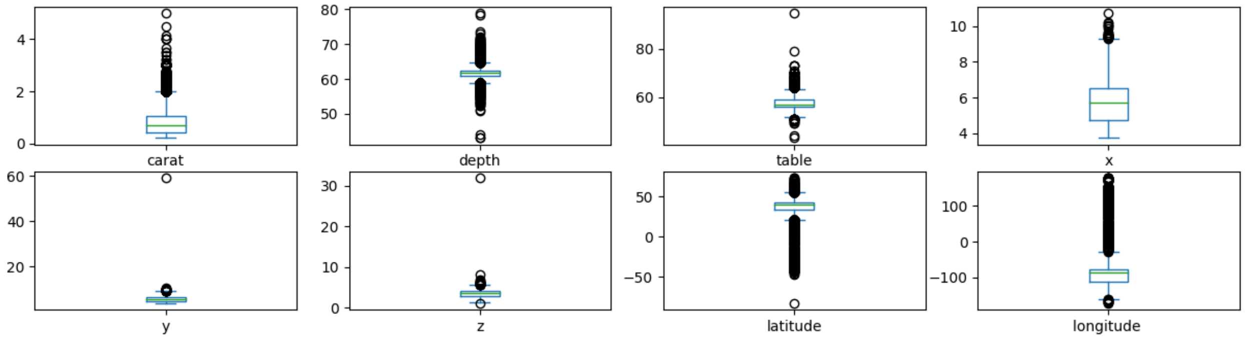
In this process the team has discovered some problems with the base dataset.

1. The presence of Nan's values. To solve this problem we have decided to eliminate the rows containing a nan value in any column.
2. In the spatial variables “x,y,z” were found some negative values which is impossible in real life, therefore, those rows with negative values were eliminated.
3. For all categorical variables had been eliminated all the symbols “*, ', #, ?, !, &” since they come in different classes as typos.

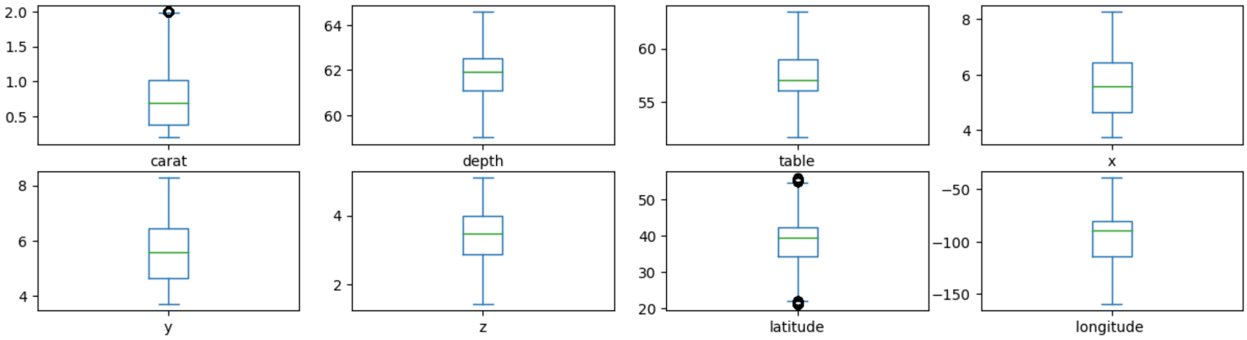
With this modification we consider the dataset is ready to use.

Variables analysis

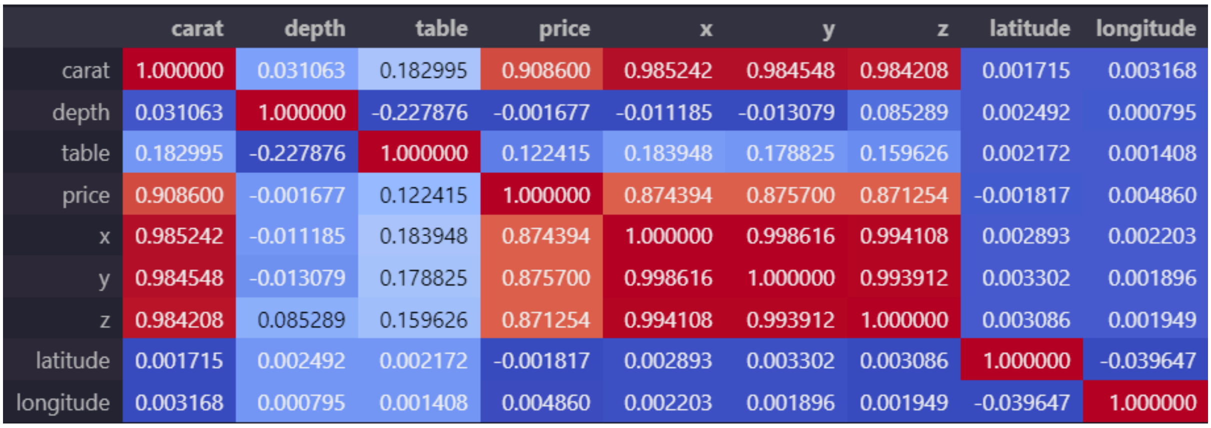
The numerical variables present numerous outliers as shown in the next figure



To solve this the Interquartile Range method to eliminate outliers was implemented, resulting in a better distribution of the data as shown in the next figure.

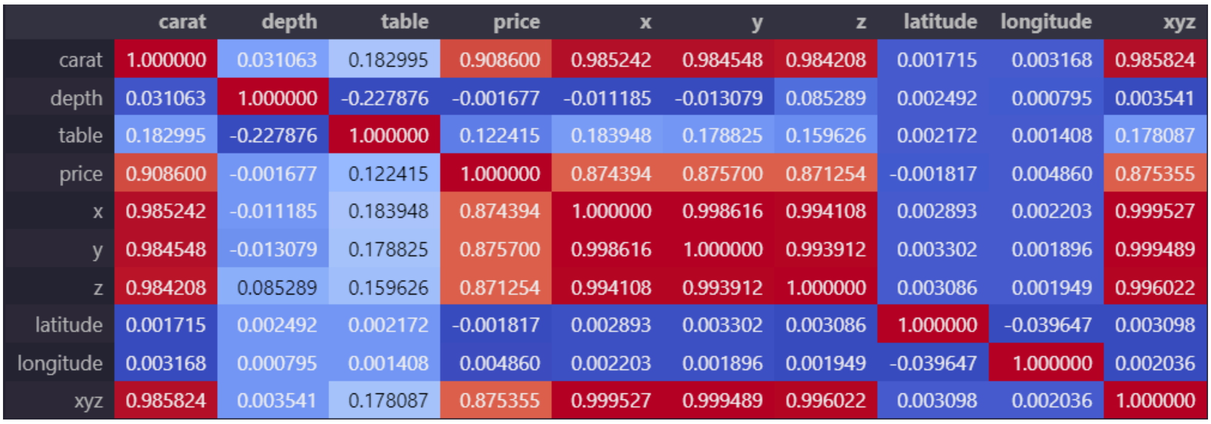


For more relevant information the correlation matrix was calculated see in the next figure



	carat	depth	table	price	x	y	z	latitude	longitude
carat	1.000000	0.031063	0.182995	0.908600	0.985242	0.984548	0.984208	0.001715	0.003168
depth	0.031063	1.000000	-0.227876	-0.001677	-0.011185	-0.013079	0.085289	0.002492	0.000795
table	0.182995	-0.227876	1.000000	0.122415	0.183948	0.178825	0.159626	0.002172	0.001408
price	0.908600	-0.001677	0.122415	1.000000	0.874394	0.875700	0.871254	-0.001817	0.004860
x	0.985242	-0.011185	0.183948	0.874394	1.000000	0.998616	0.994108	0.002893	0.002203
y	0.984548	-0.013079	0.178825	0.875700	0.998616	1.000000	0.993912	0.003302	0.001896
z	0.984208	0.085289	0.159626	0.871254	0.994108	0.993912	1.000000	0.003086	0.001949
latitude	0.001715	0.002492	0.002172	-0.001817	0.002893	0.003302	0.003086	1.000000	-0.039647
longitude	0.003168	0.000795	0.001408	0.004860	0.002203	0.001896	0.001949	-0.039647	1.000000

Since the coor between x, y, z is high (Probably because they explain the diamond size) PCA can be used to convert these variables into one. Carat is not included in this compression given its high correlation with the price. Resulting in the nex correlation matrix



	carat	depth	table	price	x	y	z	latitude	longitude	xyz
carat	1.000000	0.031063	0.182995	0.908600	0.985242	0.984548	0.984208	0.001715	0.003168	0.985824
depth	0.031063	1.000000	-0.227876	-0.001677	-0.011185	-0.013079	0.085289	0.002492	0.000795	0.003541
table	0.182995	-0.227876	1.000000	0.122415	0.183948	0.178825	0.159626	0.002172	0.001408	0.178087
price	0.908600	-0.001677	0.122415	1.000000	0.874394	0.875700	0.871254	-0.001817	0.004860	0.875355
x	0.985242	-0.011185	0.183948	0.874394	1.000000	0.998616	0.994108	0.002893	0.002203	0.999527
y	0.984548	-0.013079	0.178825	0.875700	0.998616	1.000000	0.993912	0.003302	0.001896	0.999489
z	0.984208	0.085289	0.159626	0.871254	0.994108	0.993912	1.000000	0.003086	0.001949	0.996022
latitude	0.001715	0.002492	0.002172	-0.001817	0.002893	0.003302	0.003086	1.000000	-0.039647	0.003098
longitude	0.003168	0.000795	0.001408	0.004860	0.002203	0.001896	0.001949	-0.039647	1.000000	0.002036
xyz	0.985824	0.003541	0.178087	0.875355	0.999527	0.999489	0.996022	0.003098	0.002036	1.000000

This table indicates that the variables latitude and longitude are not relevant for the price.

In the case of the cut variable, it was discovered that *Fair* and *Good* classes were not well balanced and it was decided to merge them in one new variable.

Modeling data

To make an exhaustive modeling of the data the next models were implemented:

1. Linear Regression
2. Decision Tree
3. Random Forest
4. Gradient Boost
5. Support Vector Machine

making a greed search in its parameters. The best result were given by using the variables *carat, depth, table,x,y and z* in a Gradient Boost with parameters:

- learning_rate: 0.1
- loss: 'squared_error'
- max_depth: 6
- n_estimators: 100

with metrics:

RMSE	MAE	R2
426.29	230.89	0.98

Conclusions

With the previously mentioned model the price of each diamond stolen, so **the value of the stolen diamonds is approximately $25965.03.**

In addition it can be concluded that the more relevant variables for estimating the diamond value are:

- Carat
- depth
- table
- x
- y
- z

with no modifications.